ECO 523 - Class 6

Nonparametric Econometrics

Carolina Caetano[*]

Contents

This class presents the nonparametric identification in a few models, followed by the direct application of the nonparametric estimation techniques learned in this course. We begin with nonparametric instrumental variables regression, which is a straight-forward application of series regression after a somewhat fancy identification strategy (which we do not study in depth). Then we move to the inverse probability weighted propensity score, which is an application of series regression after a smart identification strategy. The estimation may be done with kernels, but the results presented here use series estimators. Finally, we cover the regression discontinuity design, a technique that requires good boundary estimators. We cover the identification carefully, and then move to estimation, which is a straight-forward application of the local polynomial estimators.

1 Nonparametric instrumental variable regression

This is a recent and popular field of econometric research. Many of the estimators proposed are still in development. In fact, the asymptotic normality of the best known, and perhaps more intuitive among all the estimators has not yet been established.

[*] Special thanks to Lin Liu for help with plots and editing.

We will talk about the Newey-Powell series-based estimator. Consider the model

$$Y = g(X, Z_1) + u$$

However, $\mathbb{E}(u|X, Z_1) \neq 0$. Suppose that there exist instruments $Z = (Z_1, Z_2)$ such that $\mathbb{E}(u|Z_1, Z_2) = 0$. Pay special attention to the following restriction: Z is continuously distributed. Though this does happen for some applications, it rules out several instrumental variables, and certainly most instruments based on natural experiments.

If we were to follow an approach in the lines of the parametric case, we would look at the equation

$$\mathbb{E}(Y|Z = z) = \int g(x, z_1)\, F(dx|z).$$

Both $\mathbb{E}(Y|Z = z)$ and $F(dx|z)$ can be estimated. Ideally, g would be identifiable by the inverse of the integral equation above. Intuitively, if we wrote g as a function of $\mathbb{E}(Y|Z = z)$ and $F(dx|z)$, we could derive the identification, and naturally estimate g. For this we would need g to be a continuous functional of these terms, which would guarantee convergence by the continuous mapping theorem. However, it is possible to show that this is not true. Continuity fails in this case, and this is why this is called an "ill-posed inverse problem."

The solution found by Newey and Powell (2003) was to impose severe restrictions in g. Specifically, they imposed that g belongs to a compact set under the Sobolev norm. For example, for functions in the real line, the Sobolev norm is a norm in $\mathcal{C}^1[a, b]$ defined by

$$||f|| = \left\{ \int_a^b f(x)^2 dx + \int_a^b f'(x)^2 dx \right\}^{1/2}.$$

By that kind of restriction, the identification of g follows from theorems of the "intermediate value" kind.

In order to estimate g, they suppose that

$$g(x, z_1) = \sum_{k=1}^{\infty} p_k(x, z_1)\beta_k$$

where the series is absolutely convergent. Substituting into the identifying equation, we have:

$$\mathbb{E}(Y|Z = z) = \sum_{k=1}^{\infty} \mathbb{E}(p_k(X, Z_1)|Z = z)\beta_k$$

Estimation then follows in a two step procedure. First, we estimate $\mathbb{E}(Y_i|Z_i)$ and the $\mathbb{E}(p_k(X_i, Z_{1i})|Z_i)$ for $k = 1, \ldots, K$, and Newey and Powell use series regression for that.

Second, we regress the $\hat{\mathbb{E}}(Y_i|Z_i)$ onto the $\hat{\mathbb{E}}(p_k(X_i, Z_{1i})|Z_{=zi})$, $k = 1, \ldots, K$, in order to obtain $\hat{\beta}_1, \ldots, \hat{\beta}_K$. Then

$$\hat{g}(x, z_1) = \sum_{k=1}^{K} p_k(x, z_1)\hat{\beta}_k.$$

Newey and Powell showed the uniform consistency of this estimator, but the rates of convergence and asymptotic normality remain unknown in 2011.

2 Nonparametric Estimation of Average Treatment Effects

This section refers to the nonparametric estimation of treatment effects under selection on observables. That means that people may select to the treatment or to the control group according to the value of their explanatory variables, but conditional on the explanatory variables, assignment is random. In other words, it means that the covariates may influence selection to treatment, but nothing else.

There are many ways to estimate treatment effects under selection on observables. We will discuss one of the simplest methods, which is asymptotically efficient: the inverse probability weighted treatment effects estimator using the propensity score.

For this, we need to introduce some notation, which has become standard. We define causality in terms of potential outcomes. Define Y_1 to be the outcome under treatment, and the counterfactual Y_0 is the outcome under the control. Define

$$Y = Y_1 T + Y_0 (1 - T)$$

where $T = 1$ if treatment occurred, and $T = 0$ if the control occurred. The average treatment effect is therefore:

$$\theta = \mathbb{E}(Y_1 - Y_0)$$

but we only observe Y_1 or Y_0 for each observation, never both. Moreover, we cannot independently estimate $\mathbb{E}(Y_1)$ and $\mathbb{E}(Y_0)$, because the treatment and control groups are not comparable due to selection.

This is where the selection on observables assumption plays the fundamental role. We suppose that conditional on a set of observable variables X, selection to the treatment is random. This is expressed as:

$$T \perp\!\!\!\perp Y_0, Y_1 \mid X.$$

This means, for example, that conditional on X, the potential outcome under treatment is the same, independent of whether the person ended up actually taking the treatment or not. "Potential" here is the operative word.

Hence, conditional on X, the treatment and control groups are comparable, and therefore θ is identifiable by the following equation

$$\theta = \mathbb{E}(\mathbb{E}(Y_1|X) - \mathbb{E}(Y_0|X)) = \mathbb{E}(\mathbb{E}(Y|X, T = 1) - \mathbb{E}(Y|X, T = 0)).$$

It is important to notice in the identification strategy that by controlling for X before subtracting we are actually calculating the treatment effect for each possible X separately. Then, the treatment effect in the entire population is the "weighted averaged" treatment effect, where the weights are given by the distribution of X in the entire population.

Selection to the treatment or control group often does not guarantee that the subject will indeed take the treatment or control dosage. There are four possibilities:

A (Always Takers) would have taken the treatment dosage were they on the treatment or control group. For these observations, $Y = Y_1$ irrespective of T.

C (Compliers) would have taken the treatment dosage if selected to the treatment group, and the control dosage if selected to the control group. For these observations, $Y|T = 1 = Y_1$, and $Y|T = 0 = Y_0$.

D (Defiers) would have taken the treatment dosage were they on the control group, and the control dosage were they on the treatment group. For these observations, $Y|T = 1 = Y_0$, and $Y|T = 0 = Y_1$.

N (Never Takers) would have taken the control dosage were they on the treatment or control group. For these observations, $Y = Y_0$ irrespective of T.

The existence of these categories does not affect the identification condition, but it does affect the interpretation of θ. In fact, the interpretation of θ can become outright intractable because of the defiers. Their existence describes a world were basically anything goes, and no experiment (unless blind) would be valid. If one is willing to assume that defiers don't exist, then

$$\mathbb{E}(Y|X, T = 1) - \mathbb{E}(Y|X, T = 0) = \mathbb{E}(Y_1|X, C) - \mathbb{E}(Y_0|X, C)$$

and therefore θ is the expected treatment effect on the compliers group. This is what is referred to in the literature as "treatment on the treated," or "local average treatment effect (LATE)." Any experiment is in fact identifying the LATE, which can be both a good or a bad thing, depending on what is the intention of the study.

In order to estimate θ, observe that

$$\mathbb{E}(\mathbb{E}(Y|X,T=1)) = \mathbb{E}\left(\frac{\mathbb{E}(Y\,T|X)}{\mathbb{P}(T=1|X)}\right) = \mathbb{E}\left(\mathbb{E}\left(\frac{Y\,T}{\mathbb{P}(T=1|X)}\middle| X\right)\right) = \mathbb{E}\left(\frac{Y\,T}{\mathbb{P}(T=1|X)}\right)$$

The same can be done for $\mathbb{E}(\mathbb{E}(Y|X,T=0))$, yielding

$$\theta = \mathbb{E}\left(\frac{Y\,T}{\mathbb{P}(T=1|X)} - \frac{Y\,(1-T)}{1-\mathbb{P}(T=1|X)}\right).$$

The term $\mathbb{P}(T=1|X)$ is the propensity score, defined as the probability of being treated given the observables X. For it to be valid (and therefore to complete identification), it is necessary that $0 < \mathbb{P}(T=1|X) < 1$. The formula above, with the propensity score in the denominator is the inverse probability weighted treatment effect, and it is easy to estimate.

The suggested estimator is simply

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i\,T_i}{\hat{\mathbb{P}}(T_i=1|X_i)} - \frac{Y_i\,(1-T_i)}{1-\hat{\mathbb{P}}(T_i=1|X_i)}\right] W_{n,i}$$

The weights $W_{n,i}$ are there to trim off observations for which the estimated propensity score $\hat{\mathbb{P}}(T_i=1|X_i)$ is too close to zero or to one. Though identification does not require trimming, estimation does. In fact, estimation requires also that the true $\mathbb{P}(T_i=1|X_i)$ are uniformly bounded away from zero and one. As the sample increases, it is possible to allow for less trimming. In practice, if the propensity score is estimated using series, trimming is used only to eliminate outliers.

In order to derive the estimators of the propensity score $\hat{\mathbb{P}}(T_i=1|X_i)$, we regress the T_i onto X_i using series. The results using kernel estimators are exactly the same (though you need to estimate the propensity score for each X_i at a time. You can find these results in Li et al. (2005).

2.1 Asymptotic results

For notation purposes, define

$$\theta_1(X_i) = \frac{\mathbb{E}(Y_i\,T_i|X_i)}{\mathbb{P}(T_i=1|X_i)} \quad \text{and} \quad \theta_0(X_i) = \frac{\mathbb{E}(Y_i\,(1-T_i)|X_i)}{1-\mathbb{P}(T_i=1|X_i)}$$
$$\implies \theta(X_i) = \theta_1(X_i) - \theta_0(X_i).$$

Also, let

$$
\begin{aligned}
u_{1i} &= Y_i\, T_i - \mathbb{P}(T_i = 1|X_i)\, \theta_1(X_i) \\
u_{2i} &= Y_i\, (1 - T_i) - (1 - \mathbb{P}(T_i = 1|X_i))\, \theta_0(X_i) \\
u_{3i} &= T_i - \mathbb{P}(T_i = 1|X_i)
\end{aligned}
$$

Theorem: Suppose that

1. X has compact support $\mathcal{S}$, and its density $f(x) > \delta$, for some $\delta > 0$, for all $x \in \mathcal{S}$.
2. $E(u_{ki}^2|X)$ is bounded for $k = 1, 2, 3$. Define $\sigma_1^2(X) = \mathbb{V}ar(Y_1|X)$ and $\sigma_0^2(X) = \mathbb{V}ar(Y_0|X)$.
3. For every K, P^K must have full rank. Moreover, the smallest eigenvalue of $E[P^K(X_i)P^K(X_i)^T]$ is bounded away from zero uniformly in K. Finally, there exists sequence of constants $\zeta_0(K)$ such that $\sup_{x\in\mathcal{S}} ||P^K(x)|| \leqslant \zeta_0(K)$, and $\zeta_0(K)^2/n \to 0$ as $n \to \infty$.
4. $\mathbb{P}(T_i = 1|X_i)$, $\theta_1(X_i)$ and $\theta_0(X_i)$ are all in $\mathcal{C}^\infty$.
5. The sample (Y_i, X_i, T_i), $i = 1, \ldots, n$ is i.i.d.
6. $K = n^\varepsilon$, for some $\varepsilon > 0$, and $K^7/n \to 0$.

Then,

$$
\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} N\left(0, \mathbb{E}\left[\frac{\sigma_1^2(X_i)}{\mathbb{P}(T_i = 1|X_i)} + \frac{\sigma_0^2(X_i)}{1 - \mathbb{P}(T_i = 1|X_i)} + (\theta(X_i) - \theta)^2\right]\right).
$$

***

The assumptions in this theorem may be relaxed. For example, condition 1 requires that the density be bounded away from zero, but it is possible to require inverse-lipschitz types of restrictions. The same way, we can allow K to grow faster, and the functions in assumption 4 need not be $\mathcal{C}^\infty$. However, these assumptions allow us to understand intuitively how the result is derived.

Variance estimation is not very pleasant, because it requires the nonparametric estimation of $\theta(X_i)$, as well as $\sigma_1^2(X)$ and $\sigma_0^2(X)$. It can be done naturally, however, by substituting these quantities by the direct intuitive estimated counterpart, and it should not present any difficulty. You may also use bootstrap methods to estimate the variance.

3 Regression Discontinuity Design

Regression discontinuity design (RDD) is in fact an identification strategy. Hence, it can be entirely parametric. However, since misspecification bias at the boundary tends to be a much more serious problem in parametric approaches (because it is equivalent to a low degree polynomial series regression), this is an area where the efforts to develop both the nonparametric theory and to apply the technology in the actual problems have been very consistent.

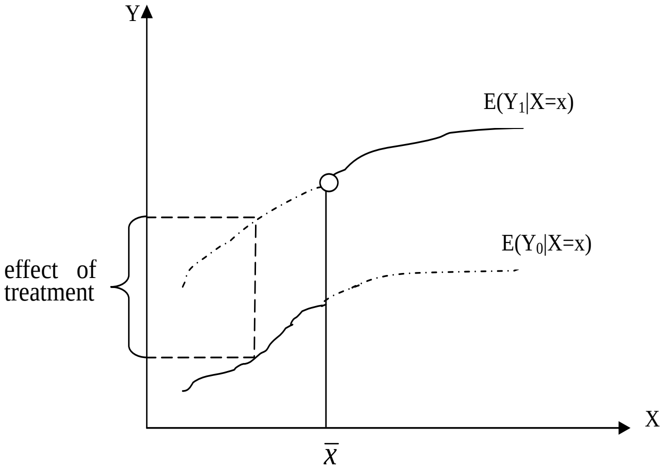
The idea of the method is simple. Suppose that treatment is given to a group based on being above a threshold $\bar{x}$ of a continuously distributed variable X. It $X_i > \bar{x}$, then that observation receives the treatment. Otherwise the treatment is denied. People can influence the value of X, but they do not have full control of it. The intuition behind the method is that the observations that are close to the threshold are comparable. People may control whether they have high or low X, but given that they have a value of X which is close to $\bar{x}$, they cannot control whether X_i will be above or below the threshold. Therefore, the group of observations directly above and directly below the threshold are comparable as if they had been obtained in a randomly assigned experiment.

If one is willing to assume random assignment in $(\bar{x} - \varepsilon, \bar{x} + \varepsilon)$, the estimation of the local effects of the treatments can be done by simple comparison between the treatment and control groups. We estimate the local effect in the sense that it represents the effect of treatment for the observations in $(\bar{x} - \varepsilon, \bar{x} + \varepsilon)$. This is a great source of criticism of the RDD: the estimated effects may be so local as to have little external validity for the rest of the population. The defense states that even though the estimated effects are local, they are local at the most interesting part of the population, that is, the part of the population for which governmental policies should be applied anyway.

The random assignment around the threshold is the intuition behind the RDD. This would give rise to questions about what is the acceptable size of ε. Identification, however, does not depend on random assignment. It depends on continuity arguments. In fact, if u is the unobservable, it is sufficient that the distribution of u conditional on X be continuous at $\bar{x}$, which is much weaker than random assigment on $(\bar{x} - \varepsilon, \bar{x} + \varepsilon)$.

Formally, define the potential outcomes as Y_1 for the treatment, and Y_0 for the control. Then

$$Y_0 = g_0(X) + u_0$$
$$Y_1 = g_1(X) + u_1$$



We are interested in the causal effect of treatment at $X = \bar{x}$, defined as

$$\theta = g_1(\bar{x}) - g_0(\bar{x})$$

Observe that exogeneity of u_0 and u_1 was never assumed, in the sense that u may be correlated with X, and therefore u may be correlated with the probability of treatment. Hence both $g_1(\bar{x})$ and $g_0(\bar{x})$ may not be identifiable. In fact, $g_1(\bar{x}) = \mathbb{E}(Y_1|X = \bar{x}) - \mathbb{E}(u_1|X = \bar{x})$ and $g_0(\bar{x}) = \mathbb{E}(Y_0|X = \bar{x}) - \mathbb{E}(u_0|X = \bar{x})$. The main identifying assumption is that

$$\mathbb{E}(u_1|X = \bar{x}) = \mathbb{E}(u_0|X = \bar{x}).$$

This implies that at $X = \bar{x}$, the unobservables are expected to be the same in the treatment and control group. Observe that we did not require that $\mathbb{E}(u_1|X = x) = \mathbb{E}(u_0|X = x)$ for all $x \in (\bar{x} - \varepsilon, \bar{x} + \varepsilon)$, as in the intuitive discussion in the introduction.

If X has a density around $\bar{x}$, this is not a strong assumption at all. We are asking for the treatment to be independent of u for a zero probability group. This is easy to assume in examples where people can control X, but not completely, as is the case with grades in exams, voting outcomes, dates of birth, etc.

Hence,

$$\theta = \mathbb{E}(Y_1|X = \bar{x}) - \mathbb{E}(Y_0|X = \bar{x}),$$

that is, we killed the endogeneity problem by restricting the search to the very local point where we can safely assume random assignment. θ is still not identifiable, because we can only observe treatment outcomes for $X > \bar{x}$. Hence $\mathbb{E}(Y_1|X = \bar{x})$ is not identifiable in general.

In order to identify $\mathbb{E}(Y_1|X = \bar{x})$, we will suppose that $\mathbb{E}(Y_1|X = x)$ is continuous at $\bar{x}$. Then,

$$\theta = \lim_{x \downarrow \bar{x}} \mathbb{E}(Y_1|X = x) - \mathbb{E}(Y_0|X = \bar{x}),$$

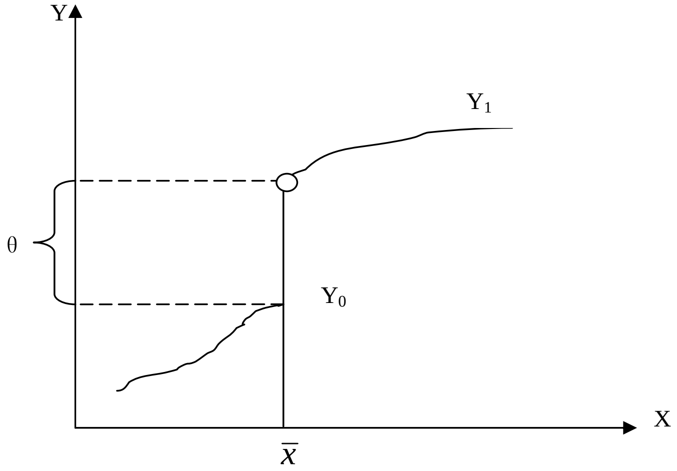
is identifiable. However, since $\mathbb{P}(X = \bar{x}) = 0$, $\mathbb{E}(Y_0|X = \bar{x})$ cannot be estimated by averaging observations such that $X_i = \bar{x}$. We will estimate this by also assuming continuity of $\mathbb{E}(Y_0|X = x)$ at $\bar{x}$, and therefore

$$\theta = \lim_{x \downarrow \bar{x}} \mathbb{E}(Y_1|X = x) - \lim_{x \uparrow \bar{x}} \mathbb{E}(Y_0|X = x).$$

Since we can only observe Y_1 for $X > \bar{x}$ and Y_0 for $X \leq \bar{x}$, we have can do

$$\theta = \lim_{x \downarrow \bar{x}} \mathbb{E}(Y|X = x) - \lim_{x \uparrow \bar{x}} \mathbb{E}(Y|X = x).$$

The assumption that both $\mathbb{E}(Y_1|X = x)$ and $\mathbb{E}(Y_0|X = x)$ are continuous is ad hoc. In general, all the assumptions above could be substituted by the requirement that both g_1 and g_0 are continuous at $\bar{x}$ and also that $\mathbb{E}(u_1|X = x)$ and $\mathbb{E}(u_0|X = x)$ are continuous at $\bar{x}$. As promised, the requirements are not of exogeneity per se, but rather of continuity.



Estimation of θ should be done by the estimation of the limits above. This is equivalent to estimation at the boundary (that is, using observations only above the threshold to estimate $\lim_{x \downarrow \bar{x}} \mathbb{E}(Y_1|X = x)$, and below the threshold to estimate $\lim_{x \uparrow \bar{x}} \mathbb{E}(Y_1|X = x)$, and the local polynomial is the most adequate among the options studied in this course.

3.1 Asymptotic behavior of the local polynomial RDD

Let

$$\hat{\theta} = b(\bar{x})^{+} - b(\bar{x})^{-},$$

where $b(\bar{x})^{+}$ is the local polynomial estimator of $\lim_{x \downarrow \bar{x}} \mathbb{E}(Y_1|X = x)$ and $b(\bar{x})^{-}$ is the local polynomial estimator of $\lim_{x \uparrow \bar{x}} \mathbb{E}(Y_0|X = x)$. We will suppose that the estimation was done by dividing the sample above and below the threshold and estimating the boundaries. Moreover, we will assume that the polynomial chosen is of odd degree.

Define $Y = T\,Y_1 + (1 - T)\,Y_0$, where $T = \mathbf{1}(X > \bar{x})$. Write

$$Y = m(X) + \varepsilon$$

where $m(X) = \mathbb{E}(Y|X)$ and $\varepsilon = Y - \mathbb{E}(Y|X)$. Observe that, for $X_i > \bar{x}$, $m(X_i) = \mathbb{E}(Y_1|X_i)$, and for $X_i \leqslant \bar{x}$, $m(X_i) = \mathbb{E}(Y_0|X_i)$. Then $b(\bar{x})^{+}$ is the local polynomial estimator of $m(\bar{x})$ using only observations such that $X_i > \bar{x}$, and $b(\bar{x})^{-}$ is the local polynomial estimator of $m(\bar{x})$ using only observations such that $X_i \leqslant \bar{x}$. This notation will make writing the assumptions below easier. We assume that the identifying assumptions hold. The following result is concerned exclusively with estimation.

Theorem: Suppose that

1. Let $\mathcal{M}$ denote the class of Borel measurable functions that have a finite second moment. Assume that $m(\cdot) = E(Y \mid X = \cdot)$ belongs to $\mathcal{M}$.

2. $E(Y^2) < \infty$.

3. $\mathbb{E}(|\varepsilon|^{2+\zeta}|X = x)$ is uniformly bounded for some $\zeta > 0$. Define $\sigma^2(x) := Var(\varepsilon \mid X = x)$. Then $\sigma^2(x)$ is continuous for $x \neq \bar{x}$, and the right and left limits of $\sigma^2(x)$ at $\bar{x}$ exist. Denote them as $\sigma^2(\bar{x})^{+}$ and $\sigma^2(\bar{x})^{-}$ respectively.

4. X is a random variable with a density function $f_X(x) = f(x)$. Let $\bar{x}$ be such that $f(\bar{x}) > 0$ and f is continuously differentiable.

5. m is $p+1$ times continuously differentiable in $x \neq \bar{x}$, and the right and left limits of m and its derivatives (up to order $p+1$) at $\bar{x}$ exist. Define $m^{(p+1)}(\bar{x})^{+} = \lim_{x \downarrow \bar{x}} m^{(p+1)}(x)$ and $m^{(p+1)}(\bar{x})^{-} = \lim_{x \uparrow \bar{x}} m^{(p+1)}(x)$.

6. The kernel $k(\cdot)$ has a compact support, is bounded, symmetric, integrates to 1, and has a second moment which is different from zero (kernel of order $s \geqslant 2$). Moreover, the kernel is a Lipschitz function.

7. The sample (Y_i, X_i), $i = 1, \ldots, n$ is i.i.d.

8. As $n \to \infty$, $h \to 0$, $nh \to \infty$ and $\sqrt{nh}\, h^{p+1}$ is bounded.

Then,

$$\sqrt{nh}\left(\hat{\theta} - \theta - B(\bar{x})\right) \xrightarrow{d} N\left(0, \frac{\sigma^2(\bar{x})^+ + \sigma^2(\bar{x})^-}{f(\bar{x})}\, e_1^T \Gamma^T \Delta \Gamma e_1\right)$$

where

$$B(\bar{x}) = c_{1,p}\, h^{p+1} \left[\frac{m^{(p+1)}(\bar{x})^+ - m^{(p+1)}(\bar{x})^-}{(p+1)!}\right] e_1^T \Gamma^{-1} \begin{bmatrix} \gamma_{p+1} \\ \vdots \\ \gamma_{2p+1} \end{bmatrix}.$$

with Γ and the other variables defined as in the local polynomial asymptotic results.

***

The assumptions of this theorem are slight variations of the assumptions required for the proof of the local polynomial. The differences are the requirement for existence of right and left limits for some of the quantities and Lipschitz kernels.

The result is also not different from the local polynomial results. It is simply the asymtptotic result for the subtraction of two independent sequences of random variables $b(\bar{x})^+$ and $b(\bar{x})^-$. The sequences are independent because the data used in $b(\bar{x})^+$ (observations such that $X_i > \bar{x}$) is entirely different from the data used in $b(\bar{x})^-$ (observations such that $X_i \leqslant \bar{x}$) and the data is i.i.d.

Estimation of the variance can be done directly, or indirectly through some bootstrap method. Directly, $\sigma^2(\bar{x})^+$ and $\sigma^2(\bar{x})^+$ need to be non-parametrically estimated. One possibility is to estimate $\sigma^2(\bar{x})^+$ by regressing (using local polynomial) the squared residuals $\hat{\varepsilon}_i^2 = [Y_i - \hat{m}(X_i)]^2$ onto the X_i at $\bar{x}$ using only observations such that $X_i > \bar{x}$. Conversely, $\sigma^2(\bar{x})^-$ can be estimated by regressing (using local polynomial) the squared residuals $\hat{\varepsilon}_i^2 = [Y_i - \hat{m}(X_i)]^2$ onto the X_i at $\bar{x}$ using only observations such that $X_i \leqslant \bar{x}$.