ECO 523 - Class 5

Nonparametric Econometrics

Carolina Caetano*

Contents

1 Introduction

So far we studied methods of estimating general multivariate models nonparametrically. Such models are in fact so general that the estimation has to acknowledge how little we know about the process. Because of this, rates of convergence (and therefore the size of the asymptotic standard errors) are large, and more so the larger the number of explanatory variables. However, you must realize that the slow rates of convergence are not a defect of the estimator, but rather a property of estimation with little information.

A general rule of estimation is that the more we know about a process, the better we can we estimate it. The trade-off is really of the risk-return kind. Imposing a given structure can yield the returns in the form of estimators better suited to the particular structure. Normally these methods have better rates of convergence, smaller asymptotic

*Special thanks to Umair Khalil for help with plots and editing.

variances and small sample bias. The risk is that if the assumed structure is wrong, the estimator will be biased, and that kind of bias does not decrease with the sample size.

In this class we will study some models that, though imposing structure, are still much more flexible than fully parameterized models. They are usually called "semi-parametric models," because they combine parametric and non-parametric aspects. As a general rule, the parametric coefficients of these models can be estimated at $\sqrt{n}$ rates.

2 Partially Linear Models

The partially linear model is one of the simplest semi-parametric models. The model is

$$Y = X^T\beta + g(Z) + u$$

that is, the model is fully parameterized in what concerns X, but the only structural assumption over the effect of Z is that is is separable from X. Assume that $\mathbb{E}(u|X,Z) = 0$. We will allow for heteroskedasticity, and denote $\sigma^2(X,Z) = \mathbb{V}ar(u|X,Z)$.

Pending more assumptions, it is possible to estimate β at the $\sqrt{n}$ rate, while g can be estimated at the rate corresponding to the dimension of Z, and the asymptotic variance of the estimation of Z is not affected by the estimation of β. Hence, this model presents advantages in the estimation of the effects of X as well as the effects of Z. In fact, the explanatory variables were divided in two groups, and from the asymptotic point of view, neither group affects the estimation of the other.

We will begin to study this model using Robinson's (1988) method, which is the most intuitive of all. By this method, both the identification and estimation of the model arise naturally.

We begin by conditioning on Z, then

$$\mathbb{E}(Y|Z) = \mathbb{E}(X|Z)^T\beta + g(Z)$$

and subtracting this equation from the model,

$$Y - \mathbb{E}(Y|Z) = [X - \mathbb{E}(X|Z)]^T\beta + u.$$

This procedure eliminates the nonparametric component of the original model, and what remains is identical to a regular linear model, except for the fact that both $\mathbb{E}(Y|Z)$ and $\mathbb{E}(X|Z)$ are not observable. They are, however, estimable, and we will deal with them later. Supposing that we can observe $Y - \mathbb{E}(Y|Z)$ and $X - \mathbb{E}(X|Z)$, the identification in

this model becomes trivial. Let,

$$\Phi = \mathbb{E}\{[X - \mathbb{E}(X|Z)][X - \mathbb{E}(X|Z)]^T\},$$

then Φ must be a positive definite matrix. For this, it is necessary (but not sufficient) that

1. X cannot contain a constant.

2. None of the components of X can be a deterministic function of Z.

If this were the case for, say X_j (the j-th component of the vector X), then $[X_j - \mathbb{E}(X_j|Z)] = 0$ for that component, rendering the identification condition false.

If we observed $Y - \mathbb{E}(Y|Z)$ and $X - \mathbb{E}(X|Z)$, estimation would be trivial. In that case, let $\tilde{Y} = Y - \mathbb{E}(Y|Z)$, and $\tilde{X} = X - \mathbb{E}(X|Z)$, then the OLS estimator of β is:

$$\hat{\beta}_{inf} = \left[\sum_{i=1}^{n} \tilde{X}_i \tilde{X}_i^T\right]^{-1} \sum_{i=1}^{n} \tilde{X}_i \tilde{Y}_i,$$

where inf stands for "infeasible." This estimator has the following asymptotic behavior:

$$\sqrt{n}(\hat{\beta}_{inf} - \beta) \xrightarrow{d} N(0, \Phi^{-1}\Psi\Phi^{-1}),$$

where,

$$\Psi = \mathbb{E}\{\sigma^2(X, Z)[X - \mathbb{E}(X|Z)][X - \mathbb{E}(X|Z)]^T\}.$$

2.1 Estimating β

In order to estimate β, the terms $\mathbb{E}(Y|Z)$ and $\mathbb{E}(X|Z)$ can be substituted by nonparametric estimates $\hat{\mathbb{E}}(Y|Z)$ and $\hat{\mathbb{E}}(Y|Z)$. Hence,

$$\hat{\beta} = \left[\sum_{i=1}^{n} \hat{X}_i \hat{X}_i^T\right]^{-1} \sum_{i=1}^{n} \hat{X}_i \hat{Y}_i,$$

where $\hat{Y}_i = Y_i - \hat{\mathbb{E}}(Y_i|Z_i)$ and $\hat{X}_i = X_i - \hat{\mathbb{E}}(X_i|Z_i)$.

To see why the method works, consider the following decomposition:

$$\begin{aligned}
\hat{\beta} - \beta &= \left[\sum_{i=1}^{n} \hat{X}_i \hat{X}_i^T\right]^{-1} \sum_{i=1}^{n} \hat{X}_i \{[\tilde{Y}_i - \tilde{X}_i\beta] + [\hat{Y}_i - \tilde{Y}_i] + [\tilde{X}_i - \hat{X}_i^T]\beta\} \\
&= \left[\sum_{i=1}^{n} \hat{X}_i \hat{X}_i^T\right]^{-1} \sum_{i=1}^{n} \hat{X}_i \{u_i - [\hat{\mathbb{E}}(Y_i|Z_i) - \mathbb{E}(Y_i|Z_i)] + [\hat{\mathbb{E}}(X_i|Z_i) - \mathbb{E}(X_i|Z_i)]^T\beta\}.
\end{aligned}$$

Though the proof details depend heavily on the method chosen for the estimation of the unobservable terms, the essence of this approximation can be understood in the equation above. The decomposition shows u_i, which only affects the variance because it is mean independent of X and Z, and the residuals of the estimation of the nonparametric terms. We suppose that these estimators are undersmoothed, so that the bias is killed faster. The variance of these terms, which is large because these terms are estimated at slow nonparametric multivariate rates, is minimized because they are being averaged. The result is that even though we are using nonparametric estimators for the nuisance functions $\mathbb{E}(Y|Z)$ and $\mathbb{E}(X|Z)$, in general,

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Phi^{-1}\Psi\Phi^{-1}).$$

The theorem below supposes that $\mathbb{E}(Y|Z)$ and $\mathbb{E}(X|Z)$ are estimated by simple kernel regression. The results when the nuisance functions are estimated using local polynomials or series regressions are the same. The only changes are in the conditions required for the result to hold (see Linton (1995) when local polynomials are used, and Li (2000) when series estimators are used).

The estimator that uses simple kernel regression for the estimation of the nuisance functions was first proposed by Robinson (1988), but the conditions were later improved in Li (1996). For theoretical purposes, they assume that the observations with very low density are trimmed. Suppose that the estimator is

$$\hat{\beta} = \left[\sum_{i=1}^{n} \hat{X}_i \hat{X}_i^T\right]^{-1} \sum_{i=1}^{n} \hat{X}_i \hat{Y}_i \, \mathbf{1}(\hat{f}(Z_i) > b).$$

The promise is that as the sample size increases, $b \to 0$. This is a theoretical construct. In practice, you don't need to trim when you are estimating these models. If you have some outliers, then yes, choose b so you can get rid of them.

Theorem: Suppose that

1. $g \in \mathcal{C}^{\nu}(\mathbb{R}^q)$, $\nu \geqslant 2$, and its partial derivatives (up to order ν satisfy the Lipschitz conditions of the form $|g(z) - g(z')| \leqslant H(z)||z - z'||$, where $H(z)$ is a continuous function with 4 finite moments.

2. $\mathbb{E}(||X||^4) < \infty$ and $\mathbb{E}(X|Z = z)$ is continuous in z. Z has a density function $f \in \mathcal{C}^{\nu-1}$ which is bounded.

3. $\mathbb{E}(u^4) < \infty$ and $\sigma^2(x, z)$ is continuous in z.

4. The kernel K is a product kernel, and k is a bounded ν-th order kernel (means that $\int v^m k(v) = 0$, $m = 1, \ldots, \nu - 1$, and $\int v^m k(v) \neq 0$), and $k(v)(1 + |v|)^{\nu+1}$ is bounded.

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

6. As $n \to \infty$ and $h \to 0$, $n(h_1, \ldots, h_q)^2 / b^4 \sum_{s=1}^{q} h_s^4 \to \infty$, and $nb^4 \sum_{s=1}^{q} h_s^{4\nu} \to 0$.

Then,

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Phi^{-1}\Psi\Phi^{-1}),$$

and

$$\hat{\Phi}^{-1}\hat{\Psi}\hat{\Phi}^{-1} \xrightarrow{p} \Phi^{-1}\Psi\Phi^{-1},$$

where

$$\hat{\Phi} = \frac{1}{n}\sum_{i=1}^{n}[X_i - \hat{\mathbb{E}}(X_i|Z_i)][X_i - \hat{\mathbb{E}}(X_i|Z_i)]^T,$$
$$\hat{\Psi} = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2[X_i - \hat{\mathbb{E}}(X_i|Z_i)][X_i - \hat{\mathbb{E}}(X_i|Z_i)]^T,$$

and

$$\hat{u}_i = Y_i - \hat{\mathbb{E}}(Y_i|Z_i) - [X_i - \hat{\mathbb{E}}(X_i|Z_i)]\hat{\beta}.$$

The assumptions of this theorem are not particularly new.

1. Assumptions 1, 2 and 3 are the usual regularity conditions. The moments existence conditions cannot be verified, and tough the continuity, smoothness and Lipschitz conditions can be tested, it is advisable that you do not thread this path. If you are not ready to assume smoothness, you should not be using this method.

2. Assumption 4 is new, in that it requires a higher order kernel (for a discussion on higher order kernels see section 2.2). These are kernels that assume negative values so that even moments cancel out. Though this is a serious theoretical restriction, it is not a restriction in practice. Any finite support kernel can be extended to a higher order and rescaled. The effect of such kernels is to reduce bias, hence if the sample is large, you can ignore this condition in practice. If your sample size is not particularly large, it is worthwhile to program a higher order kernel. There are many ways to build higher order kernels, and there are some that are already programmed into MATLAB. An example of a 4^{th} order kernel in the epanechnikov family is $k(v) = \frac{3}{8}(3 - 5v^2)\mathbf{1}(|\nu| < 1)$.

3. Condition 5 is the usual random sample assumption.

4. Condition 6 is the bandwidth assumption. It gives no clue as to the best choice of bandwidth in finite-samples. The recommendation is to follow the approaches learned for the method at hand (in this case for the Nadaraya-Watson) and undersmooth a little in order to kill the bias faster.

The theorem itself presents no difficulty for practitioners, because the asymptotic behavior of the estimator is the same as if we knew the values of the nuisance functions. hence, understanding this theorem is no more difficult than understanding the asymptotic behavior of the OLS (though the proofs are much more difficult).

One important comment is that, though you may be tempted to build a more efficient estimator under heteroskedasticity by weighting the observations by the inverse of the variance of u, $\sigma^2(X,Z)$, this will not yield the desired results. This only works if in fact $\mathbb{V}ar(u|X,Z) = \sigma^2(Z)$. Efficiency in the partially linear model is a more complex matter, and requires the nonparametric estimation of $\sigma^2(X_i,Z_i)$, which is a nonparametric regression of $\dim X + q$ dimensions. Since the partially linear model is frequently used when the dimension of X is large, this is not necessarily a good idea. See Chamberlain (1992) for the derivation of the efficient estimator in the partially linear model.

2.2 Aside: Higher Order Kernels

Assumption 4 in the above theorem requires the use of higher order kernels. The main purpose of using them instead os using regular kernels is their ability to reduce the bias in prediction, this is illustrated in figure 1 which is taken from Marron (1994) along with some of the subsequent discussion. Panel (a) shows the regression curve to be estimated along with the estimate itself which uses a standard non-negative kernel. As can be seen at points around the peak there is a larger bias and the estimate is much worse than else where. This is because positive weights are given to points around the peak, but which are comparatively lower, and hence the prediction itself is lower. A simple approach for tackling this problem is to reduce the bandwidth, but that increases the variance, which is not what we want. Higher order kernels come in.

Panel (b) shows the original curve along with estimates from both nonnegative and higher order kernels, with the kernels themselves plotted as the two lower graphs for $x = 0.5$. As can be seen the estimate with the higher order kernel does a substantially better job compared to the nonnegative one and is much closer to the actual curve thus reducing the bias by a significant margin. This happens for two reasons, first, more weight is placed on observations near the center of the kernel (which directly reduces bias). Second, observations farther away receive negative weight that tend to *cancel* a bit of the excessive lower value of some of the positively weighted observations. It is precisely because of

this bias reduction feature that we use higher order kernels: it cancels the effect of some observations away from the point, while still using information on those variables.

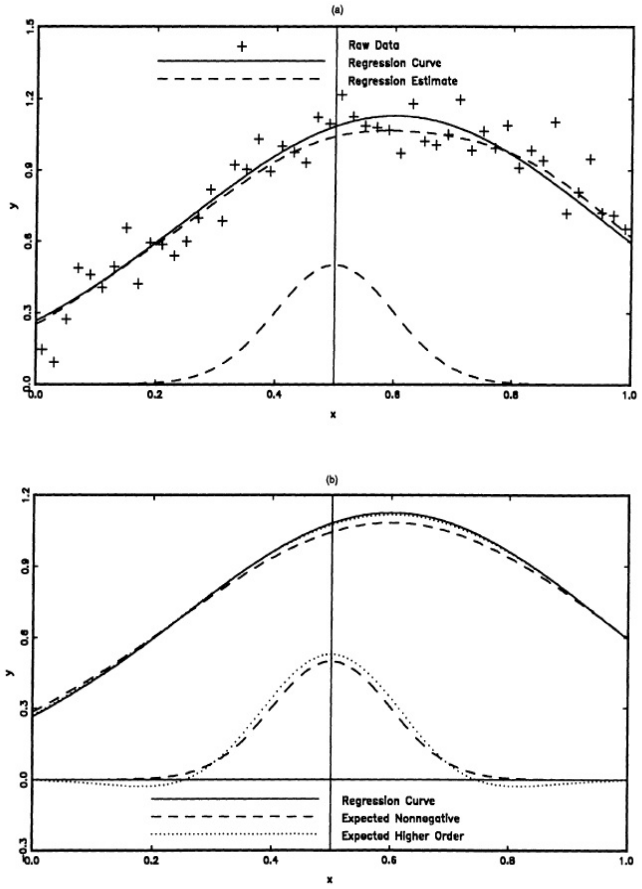

Figure 1. Example Illustrating the Visual Motivation for Higher-Order Kernels in Nonparametric Regression. (b) shows expected value curves for both nonnegative and also higher-order kernel estimates for the estimation setting shown in (a). For each kernel estimate the weights are represented by the curves at the bottom, centered at x = .5.

2.3 Estimation of g

Estimation of g is straight-forward. Observe that the model can be written as

$$Y - X^T\beta = g(Z) + u$$

This suggests a nonparametric regression of $Y - X^T\beta$ onto Z. The only difficulty is that β is not observed. However, since it can be estimated at the rate $\sqrt{n}$, while g can only be estimated at much slower rates, the use of $\hat{\beta}$ instead of β is of no consequence.

For example, if the estimator used is linear in the dependent variable (which is the case

with the kernel and series estimators), then

$$\begin{aligned}\hat{g}(x) &= \sum_{i=1}^{n} w_i(Y_i - X_i^T\hat{\beta}) \\ &= \sum_{i=1}^{n} w_i(Y_i - X_i^T\beta) - \left[\sum_{i=1}^{n} w_i X_i\right]^T (\hat{\beta} - \beta).\end{aligned}$$

The first element has the same behavior as any of the estimators we studied in the previous classes. It converges at the nonparametric rate corresponding to q, the dimension of Z. The term $\sum_{i=1}^{n} w_i X_i$ can be bounded by assumption (or trimmed), and the term $(\hat{\beta} - \beta)$ converges at the $\sqrt{n}$ rate. It is therefore negligible in comparison to the first. We have,

$$\sqrt{n}(\hat{g}(x) - g(x) - B) \xrightarrow{d} N(0, V),$$

where B and V are the bias and variance of $\hat{g}_{inf}(x) = \sum_{i=1}^{n} w_i(Y_i - X_i^T\hat{\beta})$, the nonparametric regression using $(Y_i - X_i^T\beta)$.

Hence, the asymptotic results for this estimator, no matter what technique was used, are exactly the same as in the purely nonparametric case, provided the conditions for the $\sqrt{n}$ convergence of $\hat{\beta}$ are satisfied, and the variables satisfy the conditions of the convergence of the nonparametric estimator used.

2.4 Programming the partially linear estimator

In a platform such as Stata, $\hat{\beta}$ can be obtained by the following step procedure:

1. Regress the Y_i onto the Z_i, and X_i onto the Z_i using the methods studied in the previous classes.
2. Build the variables $\hat{Y}_i = Y_i - \hat{\mathbb{E}}(Y_i|Z_i)$ and $\hat{X}_i = X_i - \hat{\mathbb{E}}(X_i|Z_i)$.
3. Regress $\hat{Y}_i$ onto $\hat{X}_i$ by simple OLS regression. Use a command such as "robust" to yield the correct standard errors.

The estimates, standard errors and test statistics calculated using this procedure will yield the correct values according to the theorem above.

For the estimation of g,

1. Build the variable $Y_i^{new} = Y_i - X_i^T\hat{\beta}$.
2. Regress Y_i^{new} onto Z_i using the approaches discussed in previous classes.

3. For variance estimation, bootstrap, cross-validation etc. simply use Y_i^{new} exactly in the same way you previously used Y_i.

For a matrix-based platform, let

$$Y^{new} = \begin{bmatrix} Y_1 - \hat{\mathbb{E}}(Y_1|Z_1) \\ \vdots \\ Y_n - \hat{\mathbb{E}}(Y_n|Z_n) \end{bmatrix}, \; X^{new} = \begin{bmatrix} X_{11} - \hat{\mathbb{E}}(X_{11}|Z_1) & \dots & X_{1K} - \hat{\mathbb{E}}(X_{1K}|Z_1) \\ \vdots & & \vdots \\ X_{n1} - \hat{\mathbb{E}}(X_{n1}|Z_n) & \dots & X_{nK} - \hat{\mathbb{E}}(X_{nK}|Z_n) \end{bmatrix}$$

then

$$\hat{\beta} = (X^{new\,T} X^{new})^{-1} X^{new\,T} Y^{new}$$

and to estimate g, use the same strategies as in the previous classes, only substitute Y^{new} where we previously had Y.

It is in fact possible to reduce computer burden in the partially linear estimator, particularly for matrix-based platforms, but this has to be done in a case-by-case basis, depending on the specific choices of the non-parametric estimators of the nuisance functions and g.

3 Other models

This section considers models capable of reducing the dimensionality curse. We will not go into the asymptotic results of such models. They follow in general the same structure as the partially linear model, and it is easy to find results for them should you need them in the future. For the purposes of this course, it is best that you know about their existence and general setup.

3.1 Additive models

Additive models are neither more general, nor more restrictive than partially linear models. They are simply another option for reduction of the dimensionality curse. The model is

$$Y = \beta_0 + g_1(X_1) + g_2(X_2) + \cdots + g_q(X_q) + u$$

Observe that it is impossible to identify β_0 independently of the other terms. Because of this, it is convenient to define $\beta_0 = \mathbb{E}(Y)$ and impose that $\mathbb{E}(g_l(X_l)) = 0$ for $l = 1, 2, ...q$.

There are many ways to estimate this model. The easiest, and most intuitive method of estimating additive models is with series regression. It suffices to consider the expansion of each term g_l separately. Observe that the definition of the variables $X_1, \dots, X_q$ is arbitrary. Each of them could be multivalued, and even contain different combinations of the same

terms, in order to incorporate variable interactions. What is fundamental is that the model is separable into a finite number of terms, and the variables in each term are known. The general result is that each term g_l can be estimated at the nonparametric rate corresponding to the number of variables that compose X_l, and not be affected by the estimation of all the other terms.

To simplify matters, we will consider the separability with two groups of variables. Generalization for more groups is trivial. The model is

$$Y = \beta_0 + g_1(X) + g_2(Z) + u$$

and suppose that

$$g_1(x) = \sum_{k=1}^{\infty} p_{1,k}(x)\beta_{1k} \quad \text{and} \quad g_2(z) = \sum_{k=1}^{\infty} p_{2,k}(z)\beta_{2k},$$

which means that g_1 has a series expansion, and g_2 has another series expansion. These expansions may be different, and come from entirely different classes of bases.

The idea is to perform a series regression, using K_1 elements of the first basis and K_2 elements of the second basis. The procedure is therefore equivalent to the regression of the Y_i onto $1, p_{1,1}(X_i), \ldots, p_{1,K_1}(X_i), p_{2,1}(Z_i), \ldots, p_{1,K_2}(Z_i)$.

The estimates are:

$$\begin{aligned}
\hat{g}_1(x) &= \sum_{k=1}^{K_1} p_{1,k}(x)\hat{\beta}_{1k} \\
\hat{g}_2(z) &= \sum_{k=1}^{K_2} p_{2,k}(z)\hat{\beta}_{2k}, \\
\hat{Y}(x,z) &= \hat{\beta}_0 + \hat{g}_1(x) + \hat{g}_2(z).
\end{aligned}$$

The asymptotic results are identical to the general series case, but the $\hat{Y}(x,z)$ converges at the rate of the slowest of $\hat{g}_1(x)$ and $\hat{g}_2(z)$. Hence, in a fully separable model, where each g_l has only one variable, the rate of convergence is that of a univariate nonparametric regression.

3.2 Varying coefficients model

This is a very popular model. Suppose that

$$Y = Z^T\beta(X) + u.$$

where $\beta(X)$ is a $p \times 1$ vector.

This model is an excellent candidate for series regression. If

$$\beta(x) = \sum_{k=1}^{\infty} p_k(x) b_k,$$

where the $p_k(x)$ are $p \times 1$ vectors, then

$$Y = \sum_{k=1}^{\infty} Z^T p_k(x) b_k + u.$$

We can approximate the true process using K elements of this basis, by regressing the Y_i onto $Z_i^T p_1(X_i), \ldots, Z_i^T p_K(X_i)$ in order to obtain $\hat{b}_1, \ldots, \hat{b}_K$.

The predicted coefficients are therefore:

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

In order to understand the asymptotic behavior of this estimator, let $P(x) = [p_1(x) \ \ldots \ p_K(x)]$,

$$P_Z = \begin{bmatrix} Z_1^T p_1(X_1) & \ldots & Z_1^T p_K(X_1) \\ \vdots & & \vdots \\ Z_n^T p_1(X_n) & \ldots & Z_n^T p_K(X_n) \end{bmatrix},$$

then

$$\hat{\beta}(x) = P(x)\hat{b} = P(x)(P_Z^T P_Z)^{-1} P_Z^T Y$$

The regularity conditions need to be altered in order to incorporate the Z_i as elements of the basis, but it is easy to see that the resulting asymptotic behavior is exactly similar to the results we saw for series estimators. In fact, pending conditions,

$$\sqrt{n} V(x)^{-1/2} (\hat{\beta}(x) - \beta(x)) \xrightarrow{d} N(0, I)$$

where

$$V(x) = \frac{1}{n} P(x)(P_Z^T P_Z)^{-1} P_Z^T \hat{\Sigma} P_Z (P_Z^T P_Z)^{-1} P(x)^T,$$

with $\hat{\Sigma} = Diag\{\hat{u}_1^2, \ldots, \hat{u}_n^2\}$.

3.3 Single Index Models

There is a large number of papers in this subject, but we will not treat these models in depth here. The model is

$$Y = g(X^T\beta) + u.$$

The identification of β requires that g cannot be constant, and that X does not contain a constant term. This is a so called **location normalization**. It is necessary to impose further restrictions for identification. A common choice is to make $||\beta|| = 1$, known as a **scale normalization**. There are more requirements for the identification of the single-index model, but we will focus on its estimation.

There are many approaches for the estimation of β for a wide class of functions g. Here we will treat g as belonging to as general a class as possible. Ichimura (1993) proposed the estimator we will study. Suppose that we knew β. Then we would be able to estimate g by Nadaraya-Watson regression, or any other method. Hence, suppose that

$$\hat{g}(x^T\beta) = \left[\sum_{i=1}^{n} k\left(\frac{X_i^T\beta - x^T\beta}{h}\right)\right]^{-1} \sum_{i=1}^{n} k\left(\frac{X_i^T\beta - x^T\beta}{h}\right) Y_i$$

Then β could be estimated by non-linear least squares regression, by solving the problem

$$\min_{\beta} \sum_{i=1}^{n} [Y_i - \hat{g}(x^T\beta)]^2 w(X_i)\mathbf{1}(X_i \in A_n)$$

for some weight function w. The function $\mathbf{1}(X_i \in A_n)$ trims the optimization process to include only observations with higher densities. This assures that the denominator of $\hat{g}$ does not get dangerously close to zero. It can be done without fear, because of the parametric nature of the estimand β. This is an optimization problem that cannot be solved analytically, but that can be approximated algorithmically using computers.

Asymptotic results for this estimator can be obtained using Taylor expansions, though the formalities are quite technical. Under the usual regularity conditions, especially the differentiability of the function g,

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V^{-1}\Sigma V^{-1})$$

where

$$\Sigma = \mathbb{E}\left\{\sigma^2(X_i)w(X_i)g'(X_i^T\beta)^2[X_i - \mathbb{E}(X_i|X_i^T\beta)][X_i - \mathbb{E}(X_i|X_i^T\beta)]^T\right\},$$
$$V = \mathbb{E}\left\{w(X_i)g'(X_i^T\beta)^2[X_i - \mathbb{E}(X_i|X_i^T\beta)][X_i - \mathbb{E}(X_i|X_i^T\beta)]^T\right\}.$$

These terms can be estimated by substituting all unobserved parts with estimators.