Lec 17: Semiparametric Estimation of Partially Linear Models

Weiping Zhang

December 21, 2020

Semiparametric Estimation of Partially Linear Models

Estimation of the Parametric Component
- Robinson's Estimator
- Li's estimator

Estimation of the Nonparametric Component

Andrew's MINPIN method

Semiparametric Efficiency Bound
- A Derivation of the Semiparametric Efficient Estimator
- A Feasible Semiparametric Efficient Estimator

Nonparametric Model Specification Tests
- Test for Correct Parametric Functional Form
- Nonparametric Test for Omitted Variables
- Specification Test in Partially Linear Models

Overview

- A model is called semiparametric if it is described by θ and τ where θ is finite-dimensional (e.g. parametric) and τ is infinite-dimensional (nonparametric).
- All moment condition models are semiparametric in the sense that the distribution of the data (τ) is unspecified and infinite dimensional. But the settings more typically called **semiparametric** are those where there is explicit estimation of τ.
- In many contexts the nonparametric part τ is a conditional mean, variance, density or distribution function.
- Often θ is the parameter of interest, and τ is a nuisance parameter, but this is not necessarily the case.
- In many semiparametric contexts, τ is estimated first, and then $\hat{\theta}$ is a two-step estimator. But in other contexts (θ, τ) are jointly estimated.

Semiparametric partially linear model

- A semiparametric partially linear model is given by

  $$Y_i = X_i'\beta_0 + g(Z_i) + u_i, i = 1, \ldots, n. \tag{1}$$

  where u_i is the random disturbance term, X_i and Z_i are $p \times 1$ and $q \times 1$ vectors of regressors, respectively, and $g(\cdot)$ is an unknown smooth function.
- The finite dimensional parameter β_0 is the parametric part of the model, and the unknown function $g(\cdot)$ is the nonparametric part of the model.
- For simplicity, we assume IID data with $E(u_i|X_i, Z_i) = 0$ and $E(u_i^2|X_i = x, Z_i = z) = \sigma^2(x, z)$. That is, we allow for a conditional heteroskedasticity of unknown form.

- It is well known that β_0 can be identified under some weak conditions. Due to the presence of the nonparametric part $g(\cdot)$, β_0 will not be identified if X_i contains a constant. Further, Z_i cannot contain a constant either because the function $g(\cdot)$ is unconstrained. As will be apparent later on, the identification condition for β_0 is that

  $$\Phi = E\{[X - E(X|Z)][X - E(X|Z)]'\}$$

  should be a positive definite matrix.
- This implies that none of the components of X can be a deterministic function of Z. Nevertheless, the identification may be possible even if X uniquely determines Z. For example, if $p = q = 1$ and $Z = X^2$ then

$$
\begin{aligned}
E(X|Z) &= \sqrt{Z}P(X > 0) + (-\sqrt{Z})P(X \leq 0) \\
&= \sqrt{Z}(1-c) - \sqrt{Z}c = \sqrt{Z}(1-2c)
\end{aligned}
$$

- and $\Phi = 4c(1-c)EX^2$, where $c = P(X \leq 0)$, so it is necessary and sufficient that X should be neither positive nor negative a.s.
- To proceed, it is worth mentioning that an early and important analysis of the model in (1) was that of Engle et al. (1986), who used it to study the impact of weather (Z_i here) on electricity demand. In his influential paper, Robinson (1988) demonstrates that β_0 can be estimated at the parametric $\sqrt{n}$-rate despite the presence of the nonparametric function g. His result parallels that of Speckman (1988) in the statistics literature.

Estimation of the Parametric Component

- We first present an infeasible estimator of β_0 in the model (1) to illustrate the core of the Robinson's (1988) method. Taking the conditional expectation of both sides of (1) given Z_i we have

$$E(Y_i|Z_i) = E(X_i'|Z_i)\beta_0 + g(Z_i) \tag{2}$$

- Subtracting (2) from (1) yields

$$Y_i - E(Y_i|Z_i) = [X_i - E(X_i|Z_i)]'\beta_0 + u_i. \tag{3}$$

- Let $\tilde{Y}_i = Y_i - E(Y_i|Z_i)$ and $\tilde{X}_i = X_i - E(X_i|Z_i)$. So (3) is linear regression model with dependent variable $\tilde{Y}_i$ and independent variable $\tilde{X}_i$. If $(\tilde{X}_i, \tilde{Y}_i)$ were observable, we can estimate β_0 by the ordinary least squares (OLS) procedure and obtain the OLS estimator

$$\tilde{\beta} = \Big(\frac{1}{n}\sum_{i=1}^{n} \tilde{X}_i \tilde{X}_i'\Big)^{-1} \frac{1}{n}\sum_{i=1}^{n} \tilde{X}_i \tilde{Y}_i. \tag{4}$$

- By the Lindeberg-Levy CLT for IID random vectors, we immediately have

$$\sqrt{n}(\tilde{\beta} - \beta_0) \rightsquigarrow N(0, \Phi^{-1}\Psi\Phi^{-1}) \tag{5}$$

where $\Psi = E[\sigma^2(X_i, Z_i)\tilde{X}_i\tilde{X}_i']$. Clearly (5) requires that Φ be positive definite.

- The basic idea underlying the above procedure is to first eliminate the unknown function $g(\cdot)$ by subtracting (2) from (1). Then we can estimate the parametric part by the standard LS procedure. Since $E(Y_i|Z_i)$ and $E(X_i|Z_i)$ are unknown in practice, $\tilde{\beta}$ is infeasible.
- If $g(Z_i)$ is linear in Z_i, then we can write (1) as

$$Y_i = X_i'\beta_0 + Z_i'\alpha_0 + u_i, i = 1, \ldots, n. \tag{6}$$

By partitioned regression, the OLS estimator $\tilde{\beta}_{ols}$ of β_0 can be obtained by regressing the residuals from the regression of Y_i on Z_i against the residuals from the regression of X_i on Z_i. That is,

$$\tilde{\beta}_{ols} = (X'M_Z X)^{-1} X' M_Z Y \tag{7}$$

where $X = (X_1, \ldots, X_n)'$, $Z = (Z_1, \ldots, Z_n)'$, $Y = (Y_1, \ldots, Y_n)'$ and

$$M_Z = I_n - Z(Z'Z)^{-1}Z'.$$

- So the infeasible estimator $\tilde{\beta}$ is a semiparametric analogue of the parametric estimator $\tilde{\beta}_{ols}$. As we will see later, $S_Z = I_n - M_Z$ can be regarded as a parametric smoothing operator. In the semiparametric or nonparametric context, we can replace S_Z by its nonparametric analog and obtain a feasible estimator for the parameters of interest.

Robinson's Estimator

- To obtain a feasible estimator for β_0, Robinson (1988) proposes to replace the unknown conditional expectations by their kernel estimates. Equivalently, we replace $\tilde{Y}_i = Y_i - E(Y_i|Z_i)$ and $\tilde{X}_i = X_i - E(X_i|Z_i)$ by $Y_i - \hat{Y}_i$ and $X_i - \hat{X}_i$, respectively, where

$$\hat{Y}_i = \hat{E}(Y_i|Z_i) = \frac{n^{-1}\sum_{j=1}^{n} Y_j \mathcal{K}_h(Z_j - Z_i)}{\hat{f}(Z_i)} \tag{8}$$

$$\hat{X}_i = \hat{E}(X_i|Z_i) = \frac{n^{-1}\sum_{j=1}^{n} X_j \mathcal{K}_h(Z_j - Z_i)}{\hat{f}(Z_i)} \tag{9}$$

and

$$\hat{f}(Z_i) = \frac{1}{n}\sum_{j=1}^{n} \mathcal{K}_h(Z_j - Z_i) \tag{10}$$

where $\mathcal{K}_h(z) = \prod_{s=1}^{q} K_h(z_{js} - z_{is})$.

Two alternative estimators

- The presence of the random denominator $\hat{f}(Z_i)$ causes some technical difficulties for the derivation of the asymptotic properties of the resulting feasible estimator of β_0. Several approaches have been proposed to handle this issue in the literature. We now discuss two of them.
- The first approach is to use the trimming technique as in Robinson (1988). Define a feasible estimator of β_0 by

$$\hat{\beta} = \Big[\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{X}_i)(X_i - \hat{X}_i)'\mathbf{1}_i\Big]^{-1}\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{X}_i)(Y_i - \hat{Y}_i)\mathbf{1}_i \qquad (11)$$

where $\mathbf{1}_i = \mathbf{1}(\hat{f}(Z_i) \geq b)$, the trimming parameter $b = b(n)$ satisfies $b \to 0$ as $n \to \infty$.

Asymptotics

- To derive the asymptotic distribution of $\hat{\beta}$, we make a definition that is adapted from Robinson (1988).

Definition

Let $\alpha > 0$ and $\mu \geq 2$ be an integer. $\mathcal{G}_{\mu}^{\alpha}$ is the class of smooth functions $g : \mathbb{R}^q \mapsto \mathbb{R}$ satisfying:
(1) g is μ-time partially differentiable
(2) g and its partial derivative functions up to order μ all satisfy the Lipschitz conditions of the type: $|g(z') - g(z)| \leq G(z)\|z' - z\|$ for all z, where $G(z)$ is a continuous function having finite αth moment, and $\| \cdot \|$ denotes the usual Euclidean norm.

- We make the following assumptions that parallel to those of Robinson (1988).

Assumptions

- A1. $\{(Y_i, X_i, Z_i)\}_{i=1}^{n}$ are IID, Z_i admits a Lebesgue density function $f_Z \in \mathcal{G}_{\nu-1}^{\infty}$ (i.e., f_Z is uniformly bounded on its support), $g \in \mathcal{G}_{\nu}^{4}$, where $\nu \geq 2$ is an integer.
- A2. $E(u_i|X_i, Z_i) = 0$ a.s., $E(u_i^2|X_i = x, Z_i = z) = \sigma^2(x, z)$ is continuous in z. $Eu_i^4 < \infty$ and $EX_{is}^4 < \infty$ for $s = 1, \ldots, p$.
- A3. $\Phi = E[X_i - E(X_i|Z_i)][X_i - E(X_i|Z_i)]'$ is positive definite.
- A4. $\mathcal{K}$ is a product of a univariate kernel K that is a bounded νth order kernel. Also, $K(u) = O((1 + |u|^{\nu+1+\epsilon})^{-1})$ for some $\epsilon > 0$.
- A5. As $n \to \infty, b \to 0$, $n(h_1 \ldots h_q)^2 b^4 \to \infty$ and $nb^{-4} \sum_{s=1}^{q} h_s^{4\nu} \to 0$.

Remarks

- Note that the assumption on the trimming parameter b is embedded in Assumption A5 and one can let it converge to zero at a extremely slow rate. For this reason, we may ignore the presence of b and make the assumptions on the bandwidth more transparent to us: Assumption A5 essentially requires that as $n \to \infty$,

$$\sqrt{n}\Big[\sum_{s=1}^{q} h_s^{2\nu} + \frac{1}{nh_1 \dots h_q}\Big] \to 0. \tag{12}$$

That is, the asymptotic MSE of $Y_i - \hat{Y}_i$ and $X_i - \hat{X}_i$ in estimating $\tilde{Y}_i = Y_i - E(Y_i|Z_i)$ and $\tilde{X}_i = X_i - E(X_i|Z_i)$ is of order $o_p(n^{-1/2})$.

- When this condition is satisfied, the difference between the feasible estimator $\hat{\beta}$ and the infeasible estimator $\tilde{\beta}$ in (4) is asymptotically negligible.
- As we will see, the requirement in (12) is typical in the literature of semiparametric estimation where nonparametric objects are estimated in the first stage and then one obtains the estimator for the finite dimensional parameters in the second stage.
- If one uses the second order kernel, then $\nu = 2$ in the above assumptions. Assume that $h_1, \ldots, h_q$ are of the same order of magnitude as h. Assumption A5 requires $nh^{2q}b^4 \to \infty$ and $nb^{-4}h^{4\nu} \to 0$ so that $q \leq 3$. In the case where $q \geq 4$ Assumption A5 requires the use of higher order kernel. Nevertheless, Li (1996) shows that Condition A5 can be replaced by a weaker condition:

- A5*: As $n \to \infty, b \to 0$, $nb^4 \sum_{s=1}^{q} h_s^{4\nu} \to 0$, $nh_1 \dots h_q b^{-4} \to \infty$, and $nb^{-4}(h_1 \dots h_q)^2 / \sum_{s=1}^{q} h_s^4 \to \infty$.
- The above assumption is counter-intuitive. Li (1996) shows that

$$\hat{\beta} - \tilde{\beta} = O_p\Big(\sum_{s=1}^{q} h_s^{2\nu} + \frac{n^{-1/2}}{nh_1 \dots h_q} + \sum_{s=1}^{q} h_s^2 (nh_1 \dots h_q)^{-1} \Big). \tag{13}$$

  $\hat{\beta}$ will be asymptotically equivalent to $\tilde{\beta}$ provided the last expression is of order $o_p(n^{-1/2})$. For detailed explanation as to why the estimation error has the order of the form (13) rather than (12), see Li (1996).

- With the weaker assumption in A5*, the use of second order requires that $q \leq 5$ implying that higher order kernels are required only when $q \geq 6$. Due to the curse of dimensionality, we won't expect $q \geq 6$ in practice, so A5* implies that a nonnegative second order kernel will be able do most the work in practice.
- we have the following theorem:

Theorem

Under Assumptions A1-A5 or Assumptions A1-A4 and A5, we have*

$$\sqrt{n}(\hat{\beta} - \beta_0) \rightsquigarrow N(0, \Phi^{-1}\Psi\Phi^{-1}).$$

Proof. See Robinson (1988) and Li (1996).

- The above Theorem says that the feasible estimator $\hat{\beta}$ is asymptotically equivalent to the infeasible estimator $\tilde{\beta}$ in (4). To obtain the standard error or confidence interval for β_0, we need to estimate the asymptotic variance of $\hat{\beta}_0$ consistently.
- We ask the reader to verify a consistent estimator is given by $\Phi^{-1}\Psi\Phi^{-1}$, where

$$\hat{\Phi} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{X}_i)(X_i - \hat{X}_i)'\mathbf{1}_i \tag{14}$$

$$\hat{\Psi} = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2(X_i - \hat{X}_i)(X_i - \hat{X}_i)'\mathbf{1}_i \tag{15}$$

and

$$\hat{u}_i = (Y_i - \hat{Y}_i) - (X_i - \hat{X}_i)'\hat{\beta}.$$

- There are two problems associated with the results in above Theorem. First, the Monte Carlo evidence in Stock (1989) suggests that the first-order asymptotic distribution may provide poor approximations to the behavior of semiparametric estimators in small samples.
- Second, the asymptotic distribution of $\hat{\beta}$ does not depend on the bandwidth h and hence does not provide a way of choosing h in practice.
- In view of these problems, Linton (1995) derives a second-order asymptotic expansion of the MSE of $\sqrt{n}(\hat{\beta}-\beta_0)$ to the order $O(n^{-2\lambda})$, where $0<\lambda<1/2$ and obtains the optimal h by minimizing the approximate MSE as was done for the nonparametric estimators.

- In the case of conditional homoskedasticity ($\sigma^2(x,z) = \sigma^2$) and $h_1 = \cdots = h_q = h$. Linton shows that the asymptotic expansion has the form

$$\begin{aligned} & MSE(\sqrt{n}(\hat{\beta} - \beta_0) \\ & \qquad = Var(\sqrt{n}(\hat{\beta} - \beta_0) + Bias(\sqrt{n}(\hat{\beta} - \beta_0)Bias(\sqrt{n}(\hat{\beta} - \beta_0)' \\ & \qquad \cong \sigma^2 \Phi^{-1} + \frac{V}{nh^q} + nh^8 B, \end{aligned}$$

where the matrix V and B are free of n and h. Clearly, the last two terms in the above expression form a correction to the first-order asymptotic MSE of $\sqrt{n}(\hat{\beta} - \beta_0)$. So the optimal h that minimizes the above MSE is $h = O(n^{-2/(8+q)})$.

- With such a choice of h the correction to the asymptotic MSE is of order $O(n^{-2\lambda})$, where $\lambda = (8-q)/(2(8+q))$. It is clear that the optimal bandwidth h for estimating β_0 is different than the order of the optimal bandwidth in estimating $E(Y|Z=z)$ or $E(X|Z=z)$.
- Another problem with the Robinson's (1988) estimator is how to choose b. Unfortunately, there is no practical guideline that can easily be followed in practice. Depending on the values of q (dimension of Z_i) and n, Robinson (1988) chooses different sequences of b.

Li's estimator

- One undesirable feature of the Robinson's (1988) estimator is the use of a trimming technique which requires the researcher to choose a nuisance trimming parameter b. Noticing this, Li (1996) proposes to use the density weighted approach to avoid a random denominator issue.
- Multiplying (3) by $f_i = f(Z_i)$, we have

$$[Y_i - E(Y_i|Z_i)]f_i = f_i[X_i - E(X_i|Z_i)]'\beta_0 + u_i f_i. \tag{16}$$

Now one can estimate the unknown finite dimensional parameter β_0 by regressing $[Y_i - E(Y_i|Z_i)]f_i$ on $f_i[X_i - E(X_i|Z_i)]$ to obtain

$$\tilde{\beta}_f = \Big(\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\tilde{X}_i' f_i^2\Big)^{-1}\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\tilde{Y}_i f_i^2. \tag{17}$$

- is easy to show that $\tilde{\beta}_f$ is asymptotically normally distributed, i.e.,

$$\sqrt{n}(\tilde{\beta}_f - \beta_0) \rightsquigarrow N(0, \Phi_f^{-1}\Psi_f\Phi_f^{-1}), \tag{18}$$

where

$$\Phi_f = E[\tilde{X}_i\tilde{X}_i' f_i^2], \Psi_f = E[\sigma^2(X_i, Z_i)\tilde{X}_i\tilde{X}_i' f_i^4]. \tag{19}$$

- As before, $\tilde{\beta}_f$ is not feasible. A feasible estimator of β_0 can be obtained by replacing $[Y_i - E(Y_i|Z_i)]f_i$ and $f_i[X_i - E(X_i|Z_i)]$ with $(Y_i - \hat{Y}_i)\hat{f}_i$ and $(X_i - \hat{X}_i)\hat{f}_i$, where $\hat{Y}_i, \hat{X}_i$ and $\hat{f}_i$ as defined before. This leads to the feasible density-weighed estimator for β_0:

$$\hat{\beta}_f = \Big[\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{X}_i)(X_i - \hat{X}_i)'\hat{f}_i^2\Big]^{-1}\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{X}_i)(Y_i - \hat{Y}_i)\hat{f}_i^2 \tag{20}$$

- Since no trimming technique is needed here, one can let $b = 1$ in Assumptions A5 and A5*. Also, the asymptotic variance of $\hat{\beta}_f$ will be different from that of $\hat{\beta}$, so Assumption A3 is replaced by
  A3*: $\Phi_f = E\{[X_i - E(X_i|Z_i)][X_i - E(X_i|Z_i)]' f_i^2\}$ is positive definite.
- We have the following theorem:

Theorem

Under Assumptions A1,A2, A3, A4 and A5 or A5* with* $b = 1$, *we have*

$$\sqrt{n}(\hat{\beta}_f - \beta_0) \rightsquigarrow N(0, \Phi_f^{-1} \Psi_f \Phi_f^{-1}).$$

- This theorem says that the feasible estimator $\hat{\beta}_f$ is asymptotically equivalent to the infeasible estimator $\tilde{\beta}_f$ in (17). To obtain the standard error or confidence interval for β_0, we need to estimate the asymptotic variance of $\hat{\beta}_f$ consistently. It is easy to verify a consistent estimator given by

$$\hat{\Phi}_f = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{X}_i)(X_i - \hat{X}_i)'\hat{f}_i^2 \tag{21}$$

$$\hat{\Psi}_f = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_{f,i}^2(X_i - \hat{X}_i)(X_i - \hat{X}_i)'\hat{f}_i^2 \tag{22}$$

and

$$\hat{u}_{f,i} = (Y_i - \hat{Y}_i) - (X_i - \hat{X}_i)'\hat{\beta}_f.$$

- Even though the density estimator $\hat{\beta}_f$ avoids the random denominator issue and it does not require the use of the trimming parameter, it is not based on any efficiency argument. In fact, it is well known that if the error process is conditionally homoskedastic, the unweighted Robinson's (1988) estimator $\hat{\beta}$ is semiparametrically efficient.
- When the error is conditionally heteroscedastic, $E(u_i^2|X_i = x, Z_i = z) = \sigma^2(x, z)$, say, one might conjecture that an efficient estimator of β_0 could be obtained by choosing weight function $w_i = 1/\sigma^2(x, z)$ as in the parametric setup. Unfortunately, this conjecture is usually incorrect. It turns out that this approach will not lead to efficient estimator of β_0 except in the special case for which the conditional variance is only a function of $Z_i = z$. That is, if $E(u_i^2|X_i = x, Z_i = z) = \sigma^2(z)$, the choice of weight function $w_i = 1/\sigma^2(z)$ will lead efficient estimation of β_0. Efficient estimation of β_0 in the general case is more complex. See Ai and Chen (2005) for details.

Estimation of the Nonparametric Component

- From (2) we have

  $$g(z) = E(Y_i - X_i'\beta_0 | Z_i = z) \tag{23}$$

  After obtaining a $\sqrt{n}$-consistent estimator $\hat{\beta}$ of β_0, we can estimate $g(z)$ consistently by

  $$\hat{g}(z) = \frac{n^{-1}\sum_{j=1}^{n}(Y_j - X_j'\hat{\beta})\mathcal{K}_h(Z_j - z)}{\hat{f}(z)}, \tag{24}$$

  where $\mathcal{K}_h(Z_i - z) = \prod_{s=1}^{q} K_{h_s}(Z_{is} - z_s)$, $\hat{f}(z) = \frac{1}{n}\sum_{j=1}^{n}\mathcal{K}_h(Z_j - z)$, and the choice of bandwidth and kernel can be quite different from those for estimating β_0.

- In order to obtain a $\sqrt{n}$-consistent estimator of β_0, a higher order kernel is required for $q \geq 6$. This is not necessary for estimating $g(z)$ regardless the value of q. One can always use a second order kernel to estimate $g(z)$ consistently and could choose the smoothing parameter h based upon the least squares cross-validation principle.
- Since the nonparametric kernel estimator has a slower convergence rate than the parametric $\sqrt{n}$-rate, it is easy to show $\hat{g}(z)$ has the same asymptotic distribution as

$$\tilde{g}(z) = \frac{n^{-1}\sum_{j=1}^{n}(Y_j - X_j'\beta_0)\mathcal{K}_h(Z_j - z)}{\hat{f}(z)}. \qquad (25)$$

The study of the asymptotic property of $\tilde{g}(z)$ is standard and we leave it as an exercise.

Another Interpretation of Robinson's Estimator

- One can easily obtain Robinson's estimator of β_0 through many other ways. As a matter of fact, Robinson's estimator can be regarded as a profile estimator in the semiparametric literature. To see this, consider a local constant approximation of $g(Z_i)$ by $\alpha = g(z)$ in the neighborhood of z. We can estimate both the finite and infinite dimensional parameters by minimizing the following objective function with respect to β and α :

$$\sum_{i=1}^{n}(Y_i - X_i'\beta - \alpha)^2 \mathbf{1}_i \mathcal{K}_h(Z_i - z), \tag{26}$$

or equivalently, in matrix notation,

$$(Y - X\beta - \alpha 1)' W_z (Y - X\beta - \alpha 1), \tag{27}$$

where $W_z = diag(\mathbf{1}_1 \mathcal{K}_h(Z_1 - z), \ldots, \mathbf{1}_n \mathcal{K}_h(Z_n - z))$, and $\mathbf{1}_i, i = 1, \ldots, n$ is as defined below (11).

- Given β, we estimate α by

$$\alpha_\beta(z) = \arg\min_{\alpha \in \mathbb{R}} (Y - X\beta - \alpha 1)' W_z (Y - X\beta - \alpha 1). \quad (28)$$

It is easy to show that

$$\alpha_\beta(z) = [1' W_z 1]^{-1} 1' W_z (Y - X\beta) = s(z)'(Y - X\beta), \quad (29)$$

where $s(z) = [1' W_z 1]^{-1} 1' W_z$ is a smoothing operator. The finite dimensional parameter β is then estimated by minimizing the following profile least squares:

$$\begin{aligned} &(Y - X\beta - \alpha_\beta(Z))' W_z (Y - X\beta - \alpha_\beta(Z)) \\ &= [(Y - SY) - (X - SX)\beta]'[(Y - SY) - (X - SX)\beta], \end{aligned}$$

where $\alpha_\beta(Z) = (\alpha_\beta(Z_1), \ldots, \alpha_\beta(Z_n))'$ and $S = (s(Z_1), \ldots, s(Z_n))'$.

- The solution to the last minimization is

$$\hat{\beta} = [(X - SX)'(X - SX)]^{-1}(X - SX)'(Y - SY). \quad (30)$$

  which is exactly equal to that defined in (11) but in different format.

- The profile likelihood estimator for $g(z)$ is given by

$$\begin{aligned} g^{+}(z) &= \alpha_{\hat{\beta}}(z) = s(z)'(Y - X\hat{\beta}) \\ &= \frac{n^{-1}\sum_{i=1}^{n}(Y_i - X_i'\hat{\beta})\mathcal{K}_h(Z_j - z)\mathbf{1}_i}{n^{-1}\sum_{i=1}^{n}\mathcal{K}(Z_i - z)\mathbf{1}_i}. \end{aligned} \quad (31)$$

  The last expression is almost same as that in (24). The only difference lies in the appearance of $\mathbf{1}_i$ in (31). One can easily show that the two estimators are asymptotically equivalent for any z in the interior of the support of Z_i.

A Profile Likelihood Estimator Based upon the Local Linear Procedure

- When Robinson (1988) derives the two-step estimator for the finite dimensional parameters, he uses the local constant procedure to obtain the preliminary estimator for the nonparametric objects. It is easy to show that the local constant procedure can be replaced by the local linear (or polynomial) procedure. Compared with the local constant procedure, the local linear procedure does not require the use of trimming parameter and usually imposes that Z_i is compactly supported.

- Assuming that g is second order differentiable and the density $f(z)$ of Z_i is strictly positive, we can use the local linear principle to estimate both the finite dimensional parameter β_0 and the infinite dimensional parameter $g(\cdot)$. Denote the first derivative of $g(z)$ by $g^{(1)}(z)$. When $\tilde{z}$ lies in the neighborhood of z, we have $g(\tilde{z}) \simeq g(z) + g^{(1)}(z)(\tilde{z} - z) = a_0 + a_1'(\tilde{z} - z)$. For notational simplicity, we denote $a = (a_0, a_1')'$ and suppress its dependence on z frequently. We can choose β and α to minimize:

$$\frac{1}{n}\sum_{i=1}^{n}[Y_i - X_i'\beta - Z_i(z)'\alpha]^2\mathcal{K}_h(Z_i - z). \qquad (32)$$

- Let $Z_i(z) = [1, (Z_i - z)']'$, and $\tilde{Z}_z = (Z_1(z)', \ldots, Z_n(z)')'$. Given β, we estimate $\alpha = \alpha(z) = (g(z), (g^{(1)}(z))')'$ by

  $$\alpha_\beta(z) = \arg \min_{\alpha \in \mathbb{R}^{q+1}} (Y - X\beta - \tilde{Z}_z\alpha)' W_z (Y - X\beta - \tilde{Z}_z\alpha). \quad (33)$$

  Define the smoothing operator by $S(z) = [\tilde{Z}_z W_z \tilde{Z}_z]^{-1} \tilde{Z}_z' W_z$. Then

  $$\alpha_\beta(z) = S(z)(Y - X\beta). \quad (34)$$

  In particular, the estimator for $g(z)$ is given by

  $$g_\beta(z) = s(z)'(Y - X\beta), \quad (35)$$

  where $s(z)' = e_1' S(z)$, and $e_1 = (1, 0, \ldots, 0)'$ is $(q+1) \times 1$ vector.

- The parameter β is then estimated by the **profile likelihood** method (more precisely, it is a **profile least squares** method in the current context):

$$\begin{aligned}\hat{\beta}_{ll} &= \arg\min_{\beta}[Y - X\beta - g_{\beta}(Z)]'[Y - X\beta - g_{\beta}(Z)] \\ &= \arg\min_{\beta}[(Y - S(z)Y) - (X - S(z)X)\beta]' \\ &\quad \cdot [(Y - S(z)Y) - (X - S(z)X)\beta],\end{aligned}$$

  where $g_{\beta}(Z) = (g_{\beta}(Z_1), \ldots, g_{\beta}(Z_n))'$ and $S = (s(Z_1), \ldots, s(Z_n))'$. That is,

$$\hat{\beta}_{ll} = [(X - SX)'(X - SX)]^{-1}(X - SX)'(Y - SY). \quad (36)$$

- The profile likelihood estimator for $\alpha(z)$ is given by

  $$\hat{\alpha}_{ll}(z) = \alpha_{\hat{\beta}}(z) = S(z)(Y - X\hat{\beta}_{ll}). \tag{37}$$

  In particular, the profile likelihood estimator for $g(z)$ is

  $$\hat{g}_{ll}(z) = g_{\hat{\beta}_{ll}}(z) = s(z)'(Y - X\hat{\beta}_{ll}). \tag{38}$$

  The study of the asymptotic distributions of $\hat{\beta}_{ll}$ and $\hat{\alpha}_{ll}(z)$ are straightforward.

Generated regressors

- The general model is

  $$y_i = \theta_\tau(X_i) + e_i, \quad E(e_i|X_i) = 0$$

  where θ is finite dimensional but τ is an unknown function. Suppose τ is identified by another equation so that we have a consistent estimate of $\hat{\tau}(x)$ for $\tau(x)$.
- Then we could estimate θ by least-squares of y_i on $\hat{\tau}(Z_i)$. This problem is called generated regressors, as the regressor is a (consistent) estimate of a infeasible regressor.
- In general, $\hat{\theta}$ is consistent. But what is its distribution?

Andrew's MINPIN method

- Andrews (1994) provides a general framework for proving the $\sqrt{n}$-consistency and asymptotic normality of a wide variety of semiparametric estimators. He names the estimators MINPIN because they are estimators that MINimize a criterion function that may depend on *Preliminary Infinite dimensional Nuisance parameter* estimators. His method can be used to derive the asymptotic distribution of various semiparametric estimators, including an estimator of β_0 in the partially linear model.

- Let $\theta \in \Theta \subset \mathbb{R}$ denote a finite dimensional parameter, and τ denote some infinite dimensional parameter. Further, let $\hat{\tau}$ be some preliminary nonparametric estimator for $\tau \in \mathcal{H}$, where $\mathcal{H}$ is a class of smooth functions. We shall use θ_0 and τ_0 to denote the true parameters corresponding to θ and τ. Suppose that $\hat{\theta}$ is a consistent estimator of θ_0 that solves a minimization problem with the following first order condition (FOC):

$$\sqrt{n}\bar{m}_n(\theta, \hat{\tau}) = 0, \tag{39}$$

  where $\bar{m}_n(\theta, \hat{\tau}) = n^{-1} \sum_{i=1}^{n} m(W_i, \theta, \hat{\tau})$. As in Andrews (1994), one can allow the function m to depend on i, in which case we can write $m(W_i, \theta, \hat{\tau})$ simply as $m_i(\theta, \hat{\tau})$.

Example

- For the partially linear model: $Y_i = X_i'\beta_0 + g(Z_i) + u_i$, we can choose $W_i = (Y_i, X_i', Z_i')'$, $\theta = \beta$ and $\hat{\tau}$ involves kernel estimators of the conditional mean functions and density. To see this more clearly, we focus on the Robinson's (1988) estimator which minimizes

$$Q_n(\theta, \hat{\tau}) = \frac{1}{n}\sum_{i=1}^{n}[(Y_i - \hat{Y}_i) - (X_i - \hat{X}_i)'\theta]^2 \mathbf{1}_i, \tag{40}$$

where $\mathbf{1}_i = \mathbf{1}(\hat{f}(Z_i) \geq b)$, $\hat{Y}_i, \hat{X}_i$ and $\hat{f}_i$ are defined in (8) through (10). Let $\hat{\tau} = \{(\hat{Y}_i, \hat{X}_i, \hat{f}_i), i = 1, \ldots, n\}$. Then the first order condition is given by

$$\sqrt{n}\bar{m}_n(\theta, \hat{\tau}) = n^{-1/2}\sum_{i=1}^{n} m(W_i, \theta, \hat{\tau}) = 0, \tag{41}$$

where $m(W_i, \theta, \hat{\tau}) = \mathbf{1}_i(X_i - \hat{X}_i)[(Y_i - \hat{Y}_i) - (X_i - \hat{X}_i)'\theta]$.

- From (41) one can solve for θ and denote the resulting solution as $\hat{\theta}$. We consider the case where $m(W_i, \theta, \tau)$ is differentiable with respect to θ. Since τ is infinite dimensional, a mean value expansion in (θ, τ) is not available. Andrews (1994) suggests expanding $\sqrt{n}\bar{m}_n(\hat{\theta}, \hat{\tau})$ about θ_0 only and using the high level concept of *stochastic equicontinuity* to handle $\hat{\tau}$.
- **Definition of Stochastic Equicontinuity** Define $v_n(\tau) = n^{-1/2}\sum_{i=1}^{n}[m(W_i, \tau) - Em(W_i, \tau)]$, then $\{v_n(\cdot), n \geq 1\}$ is stochastic equicontinuous at τ_0 if, for all $\epsilon > 0$ and $\eta > 0$, there exits a $\delta > 0$ such that

  $$\lim_{n\to\infty} P\Big(\sup_{\tau\in\mathcal{H},\rho(\tau,\tau_0)<\delta} |v_n(\tau) - v_n(\tau_0)| > \eta\Big) < \epsilon,$$

  where $\mathcal{H}$ is a class of smooth functions, and $\rho(\cdot)$ is a pseudo-metric (i.e., a metric except that $\rho(\tau_1, \tau_2) = 0$ does not necessarily imply that $\tau_1 = \tau_2$).

- By the mean value expansion in $\hat{\theta}$ we have

$$0 = \sqrt{n}\bar{m}_n(\hat{\theta}, \hat{\tau}) = \sqrt{n}\bar{m}_n(\theta_0, \hat{\tau}) + \frac{\partial}{\partial \theta'}\bar{m}_n(\theta^*, \hat{\tau})\sqrt{n}(\hat{\theta} - \theta_0), \tag{42}$$

where θ^* is an "intermediate value" between $\hat{\theta}$ and θ_0. Under certain conditions, we can guarantee that

$$\begin{aligned} \frac{\partial}{\partial \theta'}\bar{m}(W_i, \theta^*, \hat{\tau}) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \theta'}m(W_i, \theta^*, \hat{\tau}) \\ &\overset{P}{\to} E[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \theta'}m(W_i, \theta_0, \tau_0)] \\ &= E[\frac{\partial}{\partial \theta'}m(W_i, \theta_0, \tau_0)] = M \end{aligned} \tag{43}$$

where M is nonsingular. Then we have

$$\begin{aligned}
\sqrt{n}(\hat{\theta}-\theta_0) &= -\Big[\frac{\partial}{\partial\theta'}\bar{m}_n(\theta^*,\hat{\tau})\Big]^{-1} n^{-1/2}\sum_{i=1}^{n} m(W_i,\theta_0,\hat{\tau}) \\
&= -[M^{-1}+o_p(1)]n^{-1/2}\sum_{i=1}^{n} m(W_i,\theta_0,\hat{\tau}) \\
&= -M^{-1}n^{-1/2}\sum_{i=1}^{n} m(W_i,\theta_0,\tau_0)+o_p(1) \\
&\rightsquigarrow N(0,M^{-1}SM^{-1}) \qquad (44)
\end{aligned}$$

provided that

$$n^{-1/2}\sum_{i=1}^{n}[m(W_i,\theta_0,\hat{\tau})-m(W_i,\theta_0,\tau_0)]=o_p(1), \qquad (45)$$

where $S=Var(m(W_i,\theta_0,\tau_0))$.

- In practice, (45) can be difficult to verify. Andrews (1994) suggests using the concept of stochastic equicontinuity to establish it. Let $v_n = \sqrt{n}\bar{m}_n(\theta_0, \tau)$, if $v_n(\cdot)$ is stochastically equicontinuous, Andrews (1994) shows that

$$|v_n(\hat{\tau}) - v_n(\tau_0)| \xrightarrow{P} 0$$

provided that $\rho(\hat{\tau}, \tau_0) \xrightarrow{P} 0$, where $\rho(\cdot)$ is pseudo-metric.

- The following assumptions are adapted from Andrews (1994).

Assumptions

A1. $\hat{\theta} \xrightarrow{P} \theta_0 \in \Theta \subset \mathbb{R}^r$, where θ_0 is in the interior of Θ.
A2. $P(\hat{\tau} \in \mathcal{H}) \to 1$, and $\hat{\tau} \xrightarrow{P} \tau_0 \in \mathcal{H}$.
A3. $v_n(\tau_0) \rightsquigarrow N(0, S)$
A4. $\{v_n(\cdot)\}$ is stochastically equicontinuous at τ_0.
A5. $m(\theta, \tau)$ is twice continuously differentiable in $\theta \in \Theta$, $n^{-1}\sum_{i=1}^{n} m(W_i, \theta, \tau) \xrightarrow{P} E[m(W_i, \theta, \tau)]$, and $n^{-1}\sum_{i=1}^{n} \frac{\partial}{\partial\theta'} m(W_i, \theta, \tau) \xrightarrow{P} E[\frac{\partial}{\partial\theta'} m(W_i, \theta, \tau)]$ uniformly over $\Theta \times \mathcal{H}$.

- Andrews (1994) also gives conditions that imply A1. In practice, A2 and A3 can be easily verified. The most difficult part to verify is the stochastic equicontinuity A4. This is especially true for some highly nonlinear semiparametric models. Assumption A5 says that uniform weak law of large numbers hold for $m(W_i, \theta, \tau)$ and its derivatives with respect to θ.
- **Theorem**: Under Assumptions A1-A5,

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightsquigarrow N(0, M^{-1}SM^{-1}).$$

Proof. Under A4, (45) holds. Then (44) holds by A3 and Slutsky theorem. We are left to show (43). Let $M(\theta, \tau) = E[\frac{\partial}{\partial \theta'} m(W_i, \theta, \tau)]$. Under A5 and A1-A2, with probability approaching 1 we have

$$
\begin{aligned}
&\|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta'}m(W_i,\theta^*,\hat{\tau})-M(\theta_0,\tau_0)\| \\
&=\|\Big\{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta'}m(W_i,\theta^*,\hat{\tau})-M(\theta^*,\hat{\tau})\Big\} \\
&\qquad\qquad\qquad +\{M(\theta^*,\hat{\tau})-M(\theta_0,\tau_0)\}\| \\
&\leq\|\Big\{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta'}m(W_i,\theta^*,\hat{\tau})-M(\theta^*,\hat{\tau})\Big\}\| \\
&\qquad\qquad\qquad +\|\{M(\theta^*,\hat{\tau})-M(\theta_0,\tau_0)\}\| \\
&\leq\sup_{\theta\Theta}\sup_{\tau\in\mathcal{H}}\|\Big\{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta'}m(W_i,\theta,\tau)-M(\theta,\tau)\Big\}\| \\
&\qquad\qquad\qquad +\|\{M(\theta^*,\hat{\tau})-M(\theta_0,\tau_0)\}\| \\
&=o_p(1)+o_p(1)=o_p(1).
\end{aligned}
$$

Hence $\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta'}m(W_i,\theta^*,\hat{\tau}) = M(\theta_0,\tau_0)+o_p(1)$ and (43) follows. ■

- For a detailed proof of the above theorem in a more general framework, see Theorem 1 in Andrews (1994). In fact, Andrews (1994) does not assume IID data; the above result holds for weakly dependent time-series data and for independent but non-identically distributed data. In either latter case, the assumptions need to be modified to reflect dependence fact or the fact that the random variables need not have the same expectation at every point. In addition, the verification of stochastic equicontinuity may be difficult and Andrews restricts Z_i to have compact support.

Semiparametric Efficiency Bound

- In the parametric literature, we know how to judge whether a parametric estimator is asymptotically efficient. For example, the Cramer Rao lower bound (CRLB) is frequently invoked to do this. Similarly, we can judge whether a semiparametric estimator is asymptotically efficient by looking at a semiparametric analog of the CRLB, namely, the semiparametric efficiency bound (SEB). We delay general discussion of SEB to later chapter. The derivation of the SEB for a partially linear model can be found in Chamberlain (1992), whereas Ai and Chen (2003) consider efficient estimation for general semiparametric models which include the partially linear model as a special case. We base on Ai and Chen (2003) to discuss the SEB for the estimators in partially linear models.

- Consider the following partially linear model

$$Y_i = X_i'\beta + g(Z_i) + u_i, i = 1, \ldots, n, \tag{46}$$

where $E(u_i|X_i, Z_i) = 0$ and $E(u_i^2|X_i, Z_i) = \sigma^2(X_i, Z_i)$.

A Derivation of the Semiparametric Efficient Estimator

- For the moment, we pretend that $\sigma^2(\cdot)$ is known. Ai and Chen (2003) show that the efficient estimation of β_0 can be obtained by minimizing the following objective function

$$E\{[Y_i - X_i'\beta - g(Z_i)]^2/\sigma^2(X_i, Z_i)\} \tag{47}$$

with respect to β and g where $\beta \in \mathcal{B}$, a compact set in $\mathbb{R}^p$ and $g \in \mathcal{G}$, a class of smooth functions defined on $\mathbb{R}^q$. In practice, we work with the sample analog of (47) by minimizing

$$\frac{1}{n}\sum_{i=1}^{n}[Y_i - X_i'\beta - g(Z_i)]^2/\sigma^2(X_i, Z_i) \tag{48}$$

with respect to β and g.

- To obtain the estimator of β and g, we can concentrate out the unknown function g. To do so, we first treat β as fixed and apply calculus of variations to (47) to obtain

  $$2E[E\{[Y_i - X_i'\beta - g(Z_i)]/\sigma^2(X_i, Z_i)|Z_i\}a(Z_i)] = 0, \quad (49)$$

  where $a(Z_i)$ is an arbitrary function of Z_i that has second moments. Intuitively, the last expression can be obtained by differentiating

  $$E\{[Y_i - X_i'\beta - g(Z_i) - \epsilon a(Z_i)]^2/\sigma^2(X_i, Z_i)\} \quad (50)$$

  with respect to ϵ and then evaluating at $\epsilon = 0$. So $a(Z_i)$ in (49) indicates the directional change of $g(Z_i)$.

- Since (49) has to hold for all $a(Z_i)$ that has second moments, it implies that

$$E\{[Y_i - X_i'\beta - g(Z_i)]/\sigma^2(X_i, Z_i)|Z_i\} = 0 \tag{51}$$

Solving for $g(Z_i)$ in (51), we have

$$g_\beta(Z_i) = \frac{1}{E(\frac{1}{\sigma^2(X_i,Z_i)}|Z_i)} E(\frac{Y_i - X_i'\beta}{\sigma^2(X_i, Z_i)}|Z_i) \tag{52}$$

Plugging (52) into (48), we concentrate out the infinite dimensional parameter $g(\cdot)$ and obtain

$$\frac{1}{n}\sum_{i=1}^{n}[Y_i^* - X_i^{*\prime}\beta]^2/\sigma^2(X_i, Z_i), \tag{53}$$

where

$$Y_i^* = Y_i - E(\frac{Y_i}{\sigma^2(X_i, Z_i)}|Z_i)/E(\frac{1}{\sigma^2(X_i, Z_i)}|Z_i)$$
$$X_i^* = X_i - E(\frac{X_i}{\sigma^2(X_i, Z_i)}|Z_i)/E(\frac{1}{\sigma^2(X_i, Z_i)}|Z_i)$$

- The solution to the minimization of (53) is given by

$$\tilde{\beta}_{eff} = \Big[\frac{1}{n}\sum_{i=1}^{n} X_i^{*\prime} X_i^* \sigma^{-2}(X_i, Z_i)\Big]^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i^{*\prime} Y_i^* \sigma^{-2}(X_i, Z_i). \quad (54)$$

- By the Lindeberg CLT, we have

  $$\sqrt{n}(\tilde{\beta}_{eff} - \beta_0) \rightsquigarrow N(0, V_0^{-1}), \tag{55}$$

  where $V_0 = E[X_i^{*\prime} X_i^* \sigma^{-2}(X_i, Z_i)]$. V_0^{-1} is the SEB for β_0. Using a different method, Chamberlain (1992) also obtains the above SEB. If $\sigma^2(X_i, Z_i) = \sigma^2(Z_i)$, the formula for the SEB can be greatly simplified to get

  $$V_0 = E\{[X_i - E(X_i|Z_i)][X_i - E(X_i|Z_i)]' \sigma^{-2}(Z_i). \tag{56}$$

  If $\sigma^2(X_i, Z_i) = \sigma^2$, a constant function, then

  $$V_0 = \sigma^2 E\{[X_i - E(X_i|Z_i)][X_i - E(X_i|Z_i)]'\}.$$

  which implies that the Robinson's (1988) estimator for β_0 reaches the SEB in the special case of conditional homoskedasticity.

A Feasible Semiparametric Efficient Estimator

- The above estimator $\tilde{\beta}_{eff}$ is infeasible in practice. To derive a feasible estimator of β_0 we can replace the unknown quantities in $\tilde{\beta}_{eff}$ by their nonparametric kernel estimators. Since $\sigma^2(X_i, Z_i)$ is unknown, we estimate it by

$$\tilde{\sigma}^2(X_i, Z_i) = \frac{\sum_{j=1}^{n} \tilde{u}_j^2 \mathcal{K}_{h_x}(X_i - X_j)\mathcal{K}_{h_z}(Z_i - Z_j)}{\sum_{j=1}^{n} \mathcal{K}_{h_x}(X_i - X_j)\mathcal{K}_{h_z}(Z_i - Z_j)}$$

  where $\tilde{u}_i = \hat{E}(Y_i|Z_i) - \hat{E}(X_i'|Z_i)\tilde{\beta}$ is a consistent estimator of u_i based a preliminary consistent estimator $\tilde{\beta}$ of β_0.

- Thus, we can estimate Y_i^* and X_i^* respectively by

$$Y_i^* = Y_i - \hat{E}(\frac{Y_i}{\sigma^2(X_i, Z_i)}|Z_i)/\hat{E}(\frac{1}{\sigma^2(X_i, Z_i)}|Z_i)$$
$$X_i^* = X_i - \hat{E}(\frac{X_i}{\sigma^2(X_i, Z_i)}|Z_i)/\hat{E}(\frac{1}{\sigma^2(X_i, Z_i)}|Z_i)$$

- Depending on whether the density $f(x, z)$ of (X_i, Z_i) is compactly supported, different technicalities have to be dealt with via the use of various techniques, e.g., the trimming technique or the technique to handle the boundary bias issue. For brevity, we omit the technical details here.

Nonparametric Model Specification Tests

- In practice, we must confront the fact that all models are potentially misspecified. Many popular parametric model specification tests require the specification of the set of parametric alternatives for which one will reject the null. If there exist some alternative models which the test cannot detect, then the test is said to be an "inconsistent test" since it lacks power in certain directions. One famous example is the Jarque-Bera test for normality. Since this test can only detect deviations from normality by the third and fourth moments, it does not have power in detecting distributions that differ from normal distributions only through higher order moments.
- A popular application of nonparametric methods turns out to be model specification test which is usually a consistent test in that it has power in detecting all kinds of deviations from the null at certain rates.

Li and Wang's Test for Correct Parametric Functional Form

- To test whether a parametric model is correctly specified, we usually need to compare an estimate of the parametric model with its consistent nonparametric estimate. The null of interest is

  $$H_0 : P[E(Y_i|X_i) = m(X_i;\theta)] = 1 \quad for\ some\ \theta_0 \in \Theta \subset \mathbb{R}^p$$

  where $m(x;\theta)$ is a known function, X_i is a $q \times 1$ vector of regressors, θ is a $p \times 1$ vector of unknown parameters. Θ is the parameter space on $\mathbb{R}^p$. The alternative hypothesis is

  $$H_1 : P[E(Y_i|X_i) = m(X_i;\theta)] < 1 \quad for\ all\ \theta \in \Theta \subset \mathbb{R}^p$$

- Define $u_i = Y_i - m(X_i; \theta_0)$. The null is equivalent to

$$H_0 : E(u_i|X_i) = 0, a.s. \tag{57}$$

Noticing that

$$\begin{aligned} E[u_i E(u_i|X_i) f(X_i)] &= E\{E[u_i|X_i) f(X_i)|X_i]\} \\ &= E\{[E(u_i|X_i)]^2 f(X_i)\} = 0 \end{aligned}$$

under the null hypothesis and > 0 under the alternative hypothesis, where f is the pdf of X_i, we will construct a consistent model specification test based upon this observation under (57). Density weighting is used here simply to avoid a random denominator that would otherwise appear in the kernel estimator and make it hard to establish the asymptotic theory for the test.

- Clearly, the sample analogue of $E[u_i E(u_i|X_i) f(X_i)]$ is

$$\frac{1}{n}\sum_{i=1}^{n} u_i E(u_i|X_i) f(X_i),$$

which is infeasible since we don't observe u_i and don't know the functional form of f. To construct a feasible test statistic, we need to replace u_i and f by their consistent estimates. Let $\hat{u}_i = Y_i - m(X_i; \hat{\theta})$, where $\hat{\theta}$ is a $\sqrt{n}$-consistent estimator of θ based on the null model. Let $\mathcal{K}_{h,ij} = \prod_{s=1}^{q} h_s^{-1} K((X_{is} - X_{js})/h_s)$. We can estimate $E(u_i|X_i) f(X_i) = E[u_i f(X_i)|X_i]$ by the leave-one-out kernel esitmator

$$\frac{1}{n-1}\sum_{j=1, j\neq i}^{n} \hat{u}_j \mathcal{K}_{h,ij}$$

- Our test statistic is based upon

$$I_n = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i \Big\{ \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} \hat{u}_j \mathcal{K}_{h,ij} \Big\} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \hat{u}_i \hat{u}_j \mathcal{K}_{h,ij}$$

The next theorem states the asymptotic null distribution of I_n.

Theorem

Under certain regularity conditions and H_0,

$$\frac{n(h_1 \cdots h_q)^{1/2} I_n}{\hat{\sigma}} \rightsquigarrow N(0,1),$$

where $\hat{\sigma}^2 = 2n^{-2} h_1 \dots h_q \sum_{i=1}^{n} \sum_{j \neq i}^{n} \hat{u}_i^2 \hat{u}_j^2 \mathcal{K}_{h,ij}^2$ is a consistent estimator of $\sigma_e^2 = 2\kappa_{02}^q E[\sigma^4(X_i) f(X_i)]$ and $\sigma^2(x) = E(u_i^2 | X_i = x)$.

- Since the test is one-sided, we will reject the null when $T_n > z_\alpha$ at the significance level α. It is easy to show that the above test T_n is consistent. It has power in detecting any deviations from the null at the nonparametric rate $n^{-1/2}(h_1 \cdots h_q)^{-1/4}$.
- Li and Wang (1998) also proposed a **wild Bootstrap** test for finite sample. Fan and Linton (2003) further analyze the accuracy of the bootstrap test statistic.

Härdle and Mammen's Test

- Härdle and Mammen (1993) also consider testing

  $$H_0 : P[E(Y_i|X_i) = m(X_i;\theta)] = 1 \quad for\ some\ \theta_0 \in \Theta \subset \mathbb{R}^p$$

  where $m(x;\theta)$ is a known function, X_i is a $q \times 1$ vector of regressors, θ is a $p \times 1$ vector of unknown parameters. Θ is the parameter space on $\mathbb{R}^p$. The alternative hypothesis is the negation of H_0. Let $m(x) = E(Y|X = x)$ and $\hat{m}(x)$ be its nonparametric kernel estimate. Härdle and Mammen (1993) propose a consistent test for parametric functional form based upon

  $$I_n = \int [\hat{m}(x) - m(x;\hat{\theta})]^2 w(x) dx,$$

  where $w(x)$ is a nonnegative weight function, and $\hat{\theta}$ is a $\sqrt{n}$-consistent estimator for θ.

- If one choose $w(x) = f^2(x)$ and use $\hat{f}(x) = n^{-1}\sum_{i=1}^{n} \mathcal{K}_h(X_i - x)$ to replace $f(x)$, then the above test statistic turns to be

$$
\begin{aligned}
I_n \approx^* & \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}[Y_i - m(X_i,\hat{\theta})][Y_i - m(X_j,\hat{\theta})] \\
& \cdot \int \mathcal{K}_h(X_i - x)\mathcal{K}_h(X_j - x)dx \\
= & \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \hat{u}_i\hat{u}_j\bar{\mathcal{K}}_h(X_i, X_j)
\end{aligned}
$$

  where $\hat{u}_i = Y_i - m(X_i,\hat{\theta})$, $\bar{\mathcal{K}}_h(X_i, X_j) = \prod_{s=1}^{q} h_s^{-1}\bar{K}((X_{is} - X_{js})/h_s)$, and $\bar{K}(v) = \int K(u)K(v-u)du$ is the convolution kernel for K.

* The small terms are ignored in above derivation.

- The above test statistic has similar form to the test of Li and Wang (1998). By the latter result, we can prove the following theorem

Theorem

Under certain regularity conditions and H_0,

$$T_n = \frac{n(h_1 \cdots h_q)^{1/2}[I_n - c(n)]}{\hat{\sigma}} \rightsquigarrow N(0,1),$$

where $\hat{\sigma}^2 = 2n^{-2}(h_1 \dots h_q)^{-1} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \hat{u}_i^2 \hat{u}_j^2 \bar{\mathcal{K}}_h^2(X_i, X_j)$ *and* $c(n) = (nh_1 \dots h_q)^{-1} \bar{\kappa}^q(0) \sum_{i=1}^{q} \hat{u}_i^2$.

- If one chooses $w(x) = f(x)$, this can yields another version of the Härdle and Mammen's (1993) test.

Nonparametric Test for Omitted Variables

- Nonparametric tests for omitted variables have also been widely studied in the literature. Let $X \in \mathbb{R}^q$ be a $q \times 1$ vector of continuous random variables, and partition $X = (W, Z) \in \mathbb{R}^p \times \mathbb{R}^{q-p}$, where $1 \leq p < q$. The null hypothesis is that the conditional mean of Y does not depend on Z i.e.,

  $$H_0 : P[E(Y|W, Z) = E(Y|W)] = 1.$$

  The alternative is the negation of H_0.

- Let $u_i = Y_i - E(Y_i|W_i)$, then $E(u_i|X_i) = 0$ under the null hypothesis. So we can construct a test statistic based on

  $$E[u_i f_w(W_i) E\{u_i f_w(W_i)|X_i\} f(X_i)],$$

  where f_w and f are the density functions of W_i and X_i, respectively. Let $\hat{f}_{w_i}$ and $\hat{Y}_i$ be the leave-one-out kernel estimators of $f_w(W_i)$ and $E(Y_i|W_i)$, respectively. That is

  $$\hat{f}_{w_i} = \frac{1}{n-1} \sum_{j \neq i} \mathcal{K}_{h_w}(W_j - W_i),$$

  $$\hat{Y}_i = \frac{1}{(n-1)\hat{f}_{w_i}} \sum_{j \neq i} Y_i \mathcal{K}_{h_w}(W_j - W_i),$$

  where $\mathcal{K}$ is the product kernel with bandwidth $h_w = (h_{w,1}, \ldots, h_{w,p})$.

- A feasible test statistic is given by

$$I_n = \frac{1}{n(n-1)} \sum_i \sum_{j \neq i} (Y_i - \hat{Y}_i) \hat{f}_{w_i} (Y_j - \hat{Y}_j) \hat{f}_{w_j} \mathcal{K}_h(X_i - X_j),$$

  where $\mathcal{K}_h(X_i - X_j) = \prod_{s=1}^{q} K_{h_s}(X_{is} - X_{js})$.

Theorem

Under certain regularity conditions and H_0,

$$\frac{n(h_1 \cdots h_q)^{1/2} I_n}{\hat{\sigma}} \rightsquigarrow N(0,1),$$

where $\hat{\sigma}^2 = 2n^{-2} h_1 \ldots h_q \sum_{i=1}^{n} \sum_{j \neq i}^{n} \hat{u}_i^2 \hat{u}_j^2 \hat{f}_{w_i}^2 \hat{f}_{w_j}^2 \mathcal{K}_h^2(X_i, X_j)$ and $\hat{u}_i = Y_i - \hat{Y}_i$.

- Even though the above test statistic has asymptotic normal distribution under the null, simulations in Li (1999) and Lavergne and Vuong (2000) reveal that the asymptotic normal approximation does not work well for small to moderate samples. In practice, one can use the wild bootstrap principle to construct a wild bootstrap test statistic.

Consistent Test for Partially Linear Models

- For a semiparametric model, we can follow Fan and Li (1996) and test whether its specification is correct or not.
- To test whether a partially linear model is correctly specified, we follow Fan and Li (1996) and construct a nonparametric test. The null of interest is

$$H_0 : P(E(Y|X,Z) = X'\beta_0 + g_0(Z)) = 1 \tag{58}$$

for some $\beta_0 \in \mathcal{B} \subset \mathbb{R}^p$ and some $g_0 \in \mathcal{H}$, a certain space of smooth functions that are υth differentiable. The alternative hypothesis is

$$H_0 : P(E(Y|X,Z) = X'\beta_0 + g_0(Z)) < 1 \tag{59}$$

for all $\beta_0 \in \mathcal{B} \subset \mathbb{R}^p$ and all $g_0 \in \mathcal{H}$.

- For notational simplicity, denote $W_i = (X_i', Z_i')'$. Let $f_Z(\cdot)$ and $f_W(\cdot)$ denote the density of Z_i and W_i, respectively. Then $E[u_i E(u_i|W_i) f_W(W_i)] \geq 0$ and the equality holds if and only H_0 holds. One can base a test on the sample analogue of $E[u_i E(u_i|W_i) f_W(W_i)]$. Instead, noting that

$$E[u_i f_Z(Z_i) E(u_i f_Z(Z_i)|W_i) f_W(W_i)] \geq 0 \tag{60}$$

and the equality holds if and only H_0 holds, we propose a test that is based on an estimator of

$$\frac{1}{n}\sum_{i=1}^{n} u_i f_Z(Z_i) E(u_i f_Z(Z_i)|W_i) f_W(W_i)$$

to overcome the random denominator problem.

- We first estimate the finite dimensional parameter β_0 by the Li's (1996) method

$$\hat{\beta} = \Big[\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{X}_i)(X_j - \hat{X}_j)'\hat{f}_{iZ}^2(Z_i)\Big]^{-1} \times \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{X}_i)(Y_j - \hat{Y}_j)'\hat{f}_{iZ}^2(Z_i), \qquad (61)$$

where

$$\hat{Y}_i = \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} Y_j \mathcal{K}_a^z(Z_j - Z_i)/\hat{f}(Z_i),$$

$$\hat{X}_i = \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} X_j \mathcal{K}_a^z(Z_j - Z_i)/\hat{f}(Z_i),$$

$$\hat{f}_{iZ}(Z_i) = \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} \mathcal{K}_a^z(Z_j - Z_i),$$

- and $\mathcal{K}_a^z(Z_j - Z_i) = \prod_{s=1}^{q} a_s^{-1} K^z((Z_{js} - Z_{is})/a_s)$. That is, $\hat{Y}_i$ and $\hat{X}_i$ are the leave-one-out estimators of $E(Y_i|Z_i)$, $E(X_i|Z_i)$ and $f_Z(Z_i)$, respectively.
- Let $\hat{u}_i = (Y_i - \hat{Y}_i) - (X_i - \hat{X}_i)'\hat{\beta}$. Then we can estimate the density-weighted error $u_i f_Z(Z_i)$ by $\hat{u}_i \hat{f}_Z(Z_i)$. Our test statistic is based upon

$$I_n = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i \hat{f}_{iZ}(Z_i)\Big\{\frac{1}{n-1}\sum_{j=1, j\neq i}^{n} \hat{u}_j \hat{f}_{jZ}(Z_j)\mathcal{K}_h(W_j - W_i)\Big\},$$

where $\mathcal{K}_h(W_j - W_i) = [\prod_{s=1}^{q} h_s^{-1} K((X_{js} - X_{is})/h_s)] \times [\prod_{s=p+1}^{p+q} h_s^{-1} K((Z_{js} - Z_{is})/h_s)]$.

- Under some regularity conditions and H_0, we have

$$T_n = \frac{n(h_1 \dots h_{p+q})^{1/2} I_n}{\hat{\sigma}} \rightsquigarrow N(0,1)$$

  where
  $\hat{\sigma}^2 = \frac{2h_1 \dots h_{p+q}}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n \hat{u}_i^2 \hat{f}_Z(Z_i) \hat{u}_j^2 \hat{f}_Z^2(Z_j) \mathcal{K}_h^2(W_j - W_i)$.
- Since the test is one-sided, we will reject the null when $T_n > Z_\alpha$ at the significance level α. One can follow Fan and Li (1996) and show that the test is consistent. It has power in detecting any deviations from the null at the nonparametric rate $n^{-1/2}(h_1 \dots h_{p+q})^{-1/4}$.

A modified test for PLM

- We now propose a test that applies the idea of Fan and Li (1996) but avoids some drawbacks of their test. Noting that under the null of correct specification, we have

  $$E(u_i|X_i'\beta_0, Z_i) = 0 a.s. \tag{62}$$

  Let $V_i = (X_i'\beta_0, Z_i')'$ and f_V be its density. Then

  $$E\{u_i f_Z(Z_i) E[u_i f_Z(Z_i)|V_i] f_V(V_i)\} \geq 0, \tag{63}$$

  and the equality holds if and only if H_0 holds, we propose a test that is based on an estimator of

  $$\frac{1}{n}\sum_{i=1}^{n} u_i f_Z(Z_i) E[u_i f_Z(Z_i)|V_i] f_V(V_i)$$

  to overcome the random denominator problem.

- We first estimate the finite dimensional parameter β_0 by $\hat{\beta}$ as defined in (58). Since $X_i'\beta_0$ is not observable, we replace it by $\bar{V}_{1i} = X_i'\hat{\beta}$ and denote $\bar{V}_i = (\bar{V}_{1i}, Z_i')'$. Let $\hat{u}_i = (Y_i - \hat{Y}_i) - (X_i - \hat{X}_i)'\hat{\beta}$. Then we can estimate the density-weighted error $u_f f_Z(Z_i)$ by $\hat{u}_i \hat{f}_{iZ}(Z_i)$. Our test statistic is based upon

  $$I_n = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i \hat{f}_{iZ}(Z_i)\Big\{\frac{1}{n-1}\sum_{i=1, j\neq i}^{n} \hat{u}_j \hat{f}_{iZ}(Z_j)\mathcal{K}_h(\bar{V}_j - \bar{V}_i)\Big\},$$

  where

  $$\mathcal{K}_h(\bar{V}_j - \bar{V}_i) = h_1^{-1}K((\bar{V}_{1j} - \bar{V}_{1i})/h_1) \times [\prod_{s=2}^{q+1} h_s^{-1}K((Z_{js} - Z_{is})/h_s)].$$

- Define

$$T_n = \frac{n(h_1 \ldots h_{q+1})^{1/2} I_n}{\hat{\sigma}},$$

where

$$\hat{\sigma}^2 = \frac{2h_1 \ldots h_{q+1}}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \hat{u}_i^2 \hat{f}_{iZ}^2(Z_i) \hat{u}_j^2 \hat{f}_{iZ}(Z_j) \mathcal{K}_h(\bar{V}_j - \bar{V}_i).$$

- Under some regularity conditions and H_0, we have

$$T_n \rightsquigarrow N(0,1).$$

One can follow Fan and Li (1996) and show that the test is consistent. It has power in detecting any deviations from the null at the nonparametric rate $n^{-1/2}(h_1 \ldots h_{q+1})^{-1/4}$.