Semiparametric Estimation in Additive Partially Linear Models

David Shaw

3/27/2009

Table of Contents

Introduction

The model we will be analyzing is known as an additive partially linear model. It involves relating a response $Y \in \mathbb{R}^n$ to a set of covariates $(X, Z) \in \mathbb{R}^{nxp}$ semiparametrically. We have $X \in \mathbb{R}^{nxq}$, $Z \in \mathbb{R}^{nxr}$ and $p = q + r$ is the total number of covariates. The X will be modeled in the regular linear fashion while Z will be modeled using a nonparametric additive method.

Introduction

For each data point the model can be written as

$$Y_i = X_i\beta + \phi^*(Z_i) + \epsilon_i.$$

Here, $\phi^*(Z_i) = \sum_{j=1}^{r} \phi_j(Z_{ij})$, the ϵ_i have mean 0 and second moment $\sigma_i^2 = \sigma^2$, and X_i and Z_i denote the i^{th} row of X and Z respectively.
If we define $\phi_i^* = \phi^*(Z_i)$ then the model can then be written in matrix form

$$Y = X\beta + \Phi + e$$

with $\Phi = [\phi_1^* \ldots \phi_n^*]^T$.

Introduction

To complete our introduction, we will need to specify a type of function estimator that will be best suited for our needs. In other words, we need to find weights w_{nij} such that

$$\hat{\phi}^*(Z_j) = \sum_{i=1}^{n} w_{nij}\phi^*(Z_i)$$

in order to get a smooth estimate of each function ϕ^*. Many such smoothers exist.

Some examples of weights for simple or common data smooths are given in the next slide.

Introduction

1. k-Nearest Neighbor:

$$w_{nij} = \frac{1}{k} * I(i \in [j - k/2, j + k/2], i \neq j)$$

   with $k < n$ and $x_1 < \ldots < x_n$.

2. Kernel:

$$w_{nij} = K(\frac{x_i - x_j}{h}) / \sum_l K(\frac{x_l - x_j}{h})$$

   with $K(\cdot)$ a kernel function and h a bandwidth parameter.

3. Series:

$$w_{nij} = p^K(z_j)^T (P^T P)^- p^K(z_i)$$

   with $p^K(\cdot)$ a vector of estimation functions, P the matrix of all such vectors and $K = K(n)$.

For additive partially linear models, it makes the most sense to use the series estimator of the functions ϕ_i^*.

Estimation of the Parametric Component

Define H_A as the additive class of functions such that if $g, g' \in H_A$,

1. $g : \mathbb{R}^r \to \mathbb{R}$
2. $g(z) = \sum_{j=1}^r g_j(z_j)$ where $g_j(z_j)$ is continuous on a compact subset of $\mathbb{R}$
3. $\sum_{j=1}^r E[g_j(z_j)^2] < \infty$
4. $< g, g' >= E[gg']$
5. $g_j(0) = 0$ for $j = 2, \ldots, r$

The last requirement is needed so that estimation in our model will be identifiable. For use of kernel estimators, the identifiability condition is usually $Eg_j(z_j) = 0, j = 2, \ldots, r$. However, for series estimation the condition above makes more sense.

Note that H_A is an infinite-dimensional Hilbert space and is therefore not compact.

Estimation of the Parametric Component

Before estimating β we will need to come up with estimates for $\phi^*(z)$ for $z \in \mathbb{R}^r$. To estimate each $\phi_j(z_j)$ we will use a linear combination of functions $p_j^{K_j}(z_j) = [p_{1j}^{K_j}(z_j) \ldots p_{K_j j}^{K_j}(z_j)]^T$. Here, K_j is just the number of functions we are using to estimate $\phi_j(z_j)$ and each p term is a function of z_j. If we let $p^K(z)^T = [p_1^{K_1}(z_1)^T \ldots p_r^{K_r}(z_r)^T]$ then a linear combination of $p^K(z)^T$ estimates $\phi^*(z)$ for $K = \sum_j K_j$.

We know then that $p^K(z) \in H_A$ and that as $\min_j(K_j) \to \infty$, $\exists$ a linear combination of $p^K(z)$, say $\hat{p}^K(z)$, s.t. $\forall\ \epsilon > 0$, $E[(\hat{p}^K(Z) - g(Z))^2] < \epsilon,\ \forall\ g \in H_A$.

Estimation of the Parametric Component

Now if we let $p_j = [p_j^{K_j}(Z_{1j}) \dots p_j^{K_j}(Z_{nj})]^T \in \mathbb{R}^{nxK_j}$ and let $P = [p_1 \dots p_r] \in \mathbb{R}^{nxK}$ then P will be the matrix of all functions that will estimate each $\phi_j(Z_{ij})$ and thus each ϕ_i^*.
Recall our model in matrix form was

$$Y = X\beta + \Phi + e.$$

Then if we define $M = P(P^T P)^- P^T \in \mathbb{R}^{nxn}$ and $\tilde{A} = MA$ for any matrix or vector A with n rows, we know by the definition of our estimation functions in P that for large enough K, $\tilde{\Phi}$ is arbitrarily close to Φ. Therefore, premultiplying our model by M gives us

$$\tilde{Y} = \tilde{X}\beta + \tilde{\Phi} + \tilde{e} \approx \tilde{X}\beta + \Phi + \tilde{e}.$$

Estimation of the Parametric Component

Subtracting our original model from the one premultiplied by M, we get

$$Y - \tilde{Y} = (X - \tilde{X})\beta + \Phi - \tilde{\Phi} + e - \tilde{e} \approx (X - \tilde{X})\beta + e - \tilde{e}.$$

Thus, we can obtain a least squares estimate of β

$$\hat{\beta} = [(X - \tilde{X})^T (X - \tilde{X})]^- (X - \tilde{X})^T (Y - \tilde{Y})$$

where the generalized inverse becomes an inverse as $\min_j K_j \to \infty$ given certain conditions.

Estimation of the Nonparametric Component

Since $Y - X\hat{\beta} \approx \Phi$, $\forall\ z \in \mathbb{R}^r$ we can estimate $\phi^*(z)$ by

$$\hat{\phi}^*(z) = p^K(z)^T \hat{\gamma}$$

where $\hat{\gamma} = (P^T P)^{-} P^T (Y - X\hat{\beta})$.

Again, the generalized inverse $(P^T P)^{-}$ becomes an inverse asymptotically under certain assumptions.

Estimation of the Nonparametric Component

Assumptions:

1. $\forall\ i, (Y_i, X_i, Z_i)$ are iid as (Y, X, Z) and (X, Z) is defined on a compact subset of $\mathbb{R}^p$ with $\theta(z)$ a bounded function and $E\epsilon_i^2 < \infty$
2. $\forall\ K\ \exists$ a nonsingular B such that the smallest eigenvalue of $E[Bp^K(Z)p^K(Z)^TB^T]$ is bounded away from 0 uniformly in K and $\exists$ a sequence $\zeta_0(K)$ such that $sup_z||Bp^K(z)|| \leq \zeta_0(K)$ and $(\zeta_0(K))^2K/n \to 0$
3. $\exists\ \delta_j > 0$ such that $sup_z|f(z) - p^K(z)^TB^T\gamma_f| = O(\sum_j K_j^{-\delta_j})$ as $\min_j K_j \to \infty$ and $\sqrt{n}(\sum_j K_j^{-\delta_j}) \to 0$

The first assumption is standard. The second assumption ensures P^TP is asymptotically nonsingular. The third assumption ensures $\phi_j(z_j)$ is arbitrarily close to its estimator $p_j^{K_j}(z_j)^T\gamma_j$ as $K_j \to \infty$.

Estimation of the Nonparametric Component

Theorem

Under our assumptions

$$sup_{z \in \mathbb{R}^r} |\hat{\phi}^*(z) - \phi^*(z)| = O_p(\zeta_0(K))[\sqrt{\frac{K}{n}} + \sum_j K_j^{-\delta_j}].$$

Proof.

Li (2000). □

Efficiency of the Estimator

To prove efficiency, we will first need to prove some preliminary results.

Define $\Pi_A g$ as the projection of g onto H_A. If we let $\theta = E(X|Z)$ then $h = \Pi_A \theta \in H_A$ has the property that

$$E[(\theta - h)(\theta - h)^T] \\ = inf_{g \in H_A}\{E[(\theta - g)(\theta - g)^T]\}$$

Recall that the Hilbert space H_A is not compact, therefore we are unable to use Weierstrass's Theorem to say that there is an $h \in H_A$ that attains this infimum.

Efficiency of the Estimator

Lemma

For $\theta = E(X|Z)$, $\exists\ h \in H_A$ such that $h = \Pi_A \theta$ has the property

$$E[(\theta - h)(\theta - h)^T] = inf_{g \in H_A}\{E[(\theta - g)(\theta - g)^T]\}$$

Proof.

We will consider the case where X_i is univariate ($q = 1$) and there are only two nonparametric components ($r = 2$). That is,

$$Y_i = X_i\beta + \phi_1(Z_{i1}) + \phi_2(Z_{i2}) + \epsilon_i.$$

A more general proof will follow the same arguments.

Efficiency of the Estimator

If $\theta \in H_A$ then set $h = \theta$ and we are done.

We now consider the case where $\theta \notin H_A$.

Let $\{a_j\}_1^\infty$, $\{b_j\}_1^\infty$ be complete base functions that can expand ϕ_1 and ϕ_2 respectively. Then $\{c_j\}_1^\infty = (a_1, b_1, a_2, b_2, \ldots)$ is a complete base function that can expand any $\phi^* = \phi_1 + \phi_2 \in H_A$. WLOG assume $\{c_j\}$ is orthonormal. In other words, $E[c_i c_j] = \delta_{ij}$.

Now let $h = h_1 + h_2 = \sum_j c_j \alpha_j$ where $\alpha_j = E[\theta c_j]$ and let $\eta = \theta - h$.

Efficiency of the Estimator

Multiplying $\theta = h + \eta$ by c_j and taking expectation we have

$$\alpha_j = E[\theta c_j] = E[c_j \sum_i c_i \alpha_i + c_j \eta] = \sum_i \delta_{ij} \alpha_i + E[c_j \eta] = \alpha_j + E[c_j \eta]$$

Therefore $\forall\ j$,

$$E[c_j \eta] = 0 \implies \eta \perp \{c_j\} \implies \eta \perp f\ \forall\ f \in H_A$$

Thus if $h = \sum_j c_j \alpha_j < \infty$, then the fact that $\theta = h + \eta$ for $\eta \perp H_A$ implies $\exists\ h \in H_A$ that has the property

$$E[(\theta - h)^2] = inf_{g \in H_A} E[(\theta - g)^2].$$

But since $E\theta^2 < \infty$, we have $E[(h + \eta)^2] = Eh^2 + E\eta^2$.
$= \sum_j \alpha_j^2 + E\eta^2 < \infty$ which implies $h = \sum_j c_j \alpha_j < \infty$. $\square$

Efficiency of the Estimator

Theorem

Define $u_i = X_i - \Pi_A X_i$ *and assume* $U = E[u_i u_i^T]$ *is positive definite. Then under our assumptions,*

$$\sqrt{n}(\hat{\beta} - \beta) \to N(0, \sigma^2 U^{-1})$$

in distribution and

$$\hat{\sigma}^2 \hat{U}^{-1} \to \sigma^2 U^{-1}$$

in probability where $\hat{\sigma}^2 = n^{-1} \sum_i (y_i - X_i\hat{\beta} - \hat{\phi}^*(Z_i))^2$ *and* $\hat{U} = n^{-1} \sum_i (X_i - \tilde{X}_i)(X_i - \tilde{X}_i)^T$.

Efficiency of the Estimator

Outline of Proof.
(i) $\sqrt{n}(\hat{\beta} - \beta) \to N(0, \sigma^2 U^{-1})$

Recall that $X_i = \theta_i + v_i = \eta_i + h_i + v_i$ where $E[v_i | Z_i] = 0$, $h_i = \Pi_A \theta_i \in H_A$ and $\eta_i \perp H_A$.

Also recall that $\hat{\beta} = [(X - \tilde{X})^T (X - \tilde{X})]^{-1} (X - \tilde{X})^T (Y - \tilde{Y})$ asymptotically.

We write $X - \tilde{X} = \eta + v + (h - \tilde{h}) - \tilde{\eta} - \tilde{v}$ and define $S_{A,B} = n^{-1} \sum_i AB^T$ and $S_A = S_{A,A}$. Then if $S_{X-\tilde{X}}^{-1}$ exists, since $S_{A,B+C} = S_{A,B} + S_{A,C}$ we have

$$\sqrt{n}(\hat{\beta} - \beta) = S_{X-\tilde{X}}^{-1} [\sqrt{n} S_{X-\tilde{X}, \phi^* - \tilde{\phi}^* + \epsilon - \tilde{\epsilon}}]$$

Efficiency of the Estimator

It can be shown with a little bit of detail that

1. $S_{X-\tilde{X}} = U + o_p(1)$
2. $S_{X-\tilde{X},\phi^*-\tilde{\phi}^*} = o_p(n^{-1/2})$
3. $S_{X-\tilde{X},\tilde{\epsilon}} = o_p(n^{-1/2})$
4. $\sqrt{n}S_{X-\tilde{X},\epsilon} \to N(0,\sigma^2 U)$ in distribution

The first item can be proven taking

$$X - \tilde{X} = \eta + v + (h - \tilde{h}) - \tilde{v} - \tilde{\eta}$$

and showing that $(h - \tilde{h}) = o_p(1)$, $\tilde{v} = o_p(1)$ and $\tilde{\eta} = o_p(1)$. Thus, since $\eta + v = u$ then $X - \tilde{X} = u + o_p(1)$ and by LLN, $S_{X-\tilde{X}} \to E[uu^T] = U$.

Efficiency of the Estimator

(ii) $\hat{\sigma}^2 \hat{U}^{-1} \to \sigma^2 U^{-1}$

In the previous part, it was stated that we can prove with a bit of detail that $\hat{U} = S_{X-\tilde{X}} = U + o_p(1)$.

Since $\hat{\beta} - \beta = O_p(n^{-1/2})$ and $\hat{\phi}^*(Z_i) - \phi^*(Z_i) = o_p(1)$ we see that

$$\hat{\epsilon}_i - \epsilon_i = X_i(\beta - \hat{\beta}) + (\phi^*(Z_i) - \hat{\phi}^*(Z_i)) = o_p(1)$$

$\therefore\ \hat{\epsilon}_i = \epsilon_i + o_p(1)$. By LLN, we see that $n^{-1} \sum_i \hat{\epsilon}_i^2 \to E[\epsilon_i^2] = \sigma^2$, thus $\hat{\sigma}^2 \hat{U}^{-1} \to \sigma^2 U^{-1}$.

Efficiency of the Estimator

We now calculate the information in the parametric component.

In our problem we have

$$p(y_i, X_i, Z_i; \beta, \phi^*) = y_i - X_i\beta - \phi^*(Z_i)$$

Recall we have the definitions $\dot{\ell}_\beta = \nabla_\beta \log[p(y, X, Z; \beta, \phi^*)]$ and $\tilde{\ell}_\beta = \dot{\ell}_\beta - \Pi_{\phi^*}\dot{\ell}_\beta$. We will use the relation

$$I(\beta) = \int \tilde{\ell}_\beta \tilde{\ell}_\beta^T dP$$

for the information in β.
Note that the projection operator Π_{ϕ^*} is the same as the previous projection operator used Π_A since we are projecting our variable into the space of additive functions.

Efficiency of the Estimator

It can be seen through differentiation that $\dot{\ell}_\beta = \epsilon_i^{-1} X_i$ and thus $\tilde{\ell}_\beta = \epsilon_i^{-1}(X_i - \Pi_A X_i) = \epsilon_i^{-1} u_i$.
Therefore the information in the parametric component is

$$
\begin{aligned}
I(\beta) &= \int \tilde{\ell}_\beta \tilde{\ell}_\beta^T dP \\
&= \int \epsilon_i^{-2} u_i u_i^T dP \\
&= E[\epsilon_i^{-2}] E[u_i u_i^T] \geq [E\epsilon_i^2]^{-1} E[u_i u_i^T] \qquad (1)
\end{aligned}
$$

Where the inequality in (1) follows from Jensen's inequality.

Efficiency of the Estimator

Recall we defined $E[u_i u_i^T] = U$ and $E\epsilon_i^2 = \sigma^2$ meaning the information in β is

$$I(\beta) \geq \sigma^{-2} U = [Var(\sqrt{n}(\hat{\beta} - \beta))]^{-1}$$

This is just the usual Fisher information inequality. However, due to a result by Chamberlain (1992) it can be shown that for $\Sigma = E[p(y_i, X_i, Z_i; \beta, \phi^*)^2]$ and $g_{\phi^*} = E[\partial p(y_i, X_i, Z_i; \beta, \phi^*)/\partial \phi^*]$ that

$$I(\beta) \leq E[(\dot{\ell}_\beta - g_{\phi^*} h)\Sigma^{-1}(\dot{\ell}_\beta - g_{\phi^*} h)^T]$$

$\forall\ h \in H_A$. Thus letting $h = \Pi_A X_i$ and using the facts that $g_{\phi^*} = 1$ and $\Sigma^{-1} = \sigma^{-2}$ we see that

$$I(\beta) \leq \sigma^{-2} E[u_i u_i^T]$$

meaning $I(\beta) = \sigma^{-2} U = [Var(\sqrt{n}(\hat{\beta} - \beta))]^{-1}$. Therefore $\hat{\beta}$ is an asymptotically efficient estimator of β.

Applications

A popular application of these partially linear additive models is in partially linear time series, particularly additive stochastic regression mentioned in Härdle, et al (2000). In it we assume

$$Y_t = m(U_t) + \epsilon_t = \sum_j \phi_j(X_{tj}) + g(Z_t) + \epsilon_t$$

with $Z_t \in \mathbb{R}^r$. The partially linear additive model discussed in this talk arises when we assume that some of the ϕ_j's are related to the X_{tj}'s linearly. Usually the function g is not assumed to be additive.

Models of this type have important applications in economics.

References

- Chamberlain, G.. "Efficiency Bounds for Semiparametric Regression," *Econometrica* 60 (1992), 567-96.
- Härdle, Liang & Gao. *Partially Linear Models*. (Heidelberg : Physica-Verlag, 2000).
- Li, Qi. "Efficient Estimation of Additive Partially Linear Models," *International Economic Review* 41 (2000), 1073-92.