Nonparametric Methods in Economics and Finance
Lecture 5

Oliver Linton

May 10, 2006

Generalized Additive and Other Separable Models

- Consider now the generalized additive model in which there is some known transformation G for which
$$G\{m(x)\} = c + \sum_{\alpha=1}^{d} m_\alpha(x_\alpha),$$
where $m(x) = E(Y|X = x)$. This arises in the context of limited dependent variable models.

- For example, if Y_i is binary we might take G to be the inverse of a c.d.f. F, so that the model is
$$\begin{aligned} \Pr[Y_i = 1|X = x] &= F\left(c + \sum_{\alpha=1}^{d} m_\alpha(x_\alpha)\right) \\ &= m(x). \end{aligned}$$
In this example we have restrictions not just on the mean function $m(x)$ but on the entire distribution. For example, it follows that
$$\mathrm{var}(Y|X = x) = m(x)(1 - m(x)).$$

- An alternative is the transformation model

$$\Lambda(Y) = c + \sum_{\alpha=1}^{d} m_\alpha(X_\alpha) + \varepsilon,$$

where Λ is a strictly monotonic transformation and ε is independent of X. The transformation can be either parametric or nonparametric. This implies that

$$\mathrm{med}(Y|X = x) = \Lambda^{-1}\left(c_* + \sum_{\alpha=1}^{d} m_\alpha(X_\alpha)\right)$$

for some constant c_*, and

$$\begin{aligned} & E(Y|X = x) \\ = & \int \Lambda^{-1}\left(c + \sum_{\alpha=1}^{d} m_\alpha(X_\alpha) + e\right) f_\varepsilon(e) de \\ = & G\left(c + \sum_{\alpha=1}^{d} m_\alpha(X_\alpha)\right). \end{aligned}$$

- In some cases one is interested in models for both mean and variance. So for example

$$\begin{aligned} Y &= c + \sum_{\alpha=1}^{d} m_\alpha(X_\alpha) + \sigma(X)\varepsilon \\ \sigma^2(X) &= c_\sigma + \sum_{\alpha=1}^{d} v_\alpha(X_\alpha), \end{aligned}$$

where m_α and v_α are unknown functions. One can instead model some transformation of $\sigma^2(X)$, say $F(\sigma^2)$ as being additive.

- Rigby and Stasinopoulos (1995 etc.) consider the class of Mean and Dispersion Additive Models (MADAM's) in which

  $$\begin{aligned} E(Y|X &= x) = m(x) \\ \text{var}(Y|X &= x) = \phi(x)V(m(x)), \end{aligned}$$

  where

  $$\begin{aligned} G_m(m(x)) &= \sum_{\alpha=1}^{d} m_\alpha(x_\alpha) \\ G_\phi(\phi(x)) &= \sum_{\alpha=1}^{d} \phi_\alpha(x_\alpha) \end{aligned}$$

  for unknown functions m_α, ϕ_α and known functions V, G_m, and G_ϕ.

- Generalized additive separability is where there is some transformation G of m that is additively separable. There are a number of different forms of separability used in economics, see Leontieff (1947), Goldman and Uzawa (1964). Let $N_1, \ldots, N_s$ be a partition of $\{1, \ldots., d\}$.
  - Strong Separability: For functions m_α of $x^{(j)} = \{x_l; l \in N_j\}$ and scalar argument h

$$m(x) = h\{m_1(x^{(1)}) + \ldots + m_s(x^{(s)})\}$$

  - Weak Separability: For s-dim argument H,

$$m(x) = H\{m_1(x^{(1)}), \ldots, m_s(x^{(s)})\}$$

  - Pearce Separability: For additive functions m_α

$$m(x) = H\{m_1(x^{(1)}), \ldots, m_s(x^{(s)})\}.$$

- Pinkse (2001) discusses nonparametric estimation in the weakly separable case where both H and the m_α functions are unknown. Identifies only upto an unknown monotonic transformation. Horowitz (2001) discusses nonparametric estimation in the strongly separable case.

- Homothetic functions. Suppose that there exist functions h and g such that

$$m(x) = h[g(x)],$$

where g is linearly homogeneous

$$g(cx) = cg(x)$$

for scalar c, and h is strictly monotonic on its first element. Then we say that $m(x)$ is homothetic. Tripathi and Kim (2000) discuss estimation of homogenous functions.

Homothetic Separable

- Lewbel and Linton (2006). Let $R(v, z, w)$ be some function that can be nonparametrically estimated, for example, $R(v, z, w)$ could equal $E(Y \mid V = v, Z = z, W = w)$. More generally, $R(v, z, w)$ could be a density, distribution, quantile, or hazard function, or $R(v, z, w)$ could be a utility or cost function derived from a set of estimated product or factor demands. Assume there exist unknown functions h and g and known strictly monotonic functions B_1, B_2, and B_3 such that

$$R(v, z, w) = h[B_1(B_2(v)B_3(g(z)), w],$$

  where h is strictly monotonic on its first element.

- One leading example is when B_1 is the natural logarithm and B_2 and B_3 are exponentiation, which gives

$$R(v, z, w) = h[v + g(z), w].$$

- Another important example of equation is when B_1, B_2, and B_3 are the identity functions, which gives

$$R(v, z, w) = h[vg(z), w].$$

- A function $r(x, w)$ is defined to be homothetically separable in x if and only if

$$r(x, w) = h[s(x), w]$$

where h is strictly monotonic in s and s is linearly homogeneous. Let v be one element of x that never equals zero, and let z be the vector of all the other elements of x divided by v. Alternatively, rewrite x in polar coordinates as v, z, where v is length and z is direction. Either way, $s(x)$ is linearly homogeneous if and only if

$$s(x) = vg(z)$$

for some unrestricted function g, so a function r is homothetically separable in x if and only if it has the form of R in equation ?.

- In applications of homothetic separability, r may have multiple homogeneous components, that is,

$$r(x_0, x_1, \ldots, x_k) = h[s_1(x_1), \ldots, s_K(x_K), x_0]$$

for vectors $x_0, x_1, \ldots, x_K$. In this model, each homogeneous s_k function can be estimated separately by applying the method we propose to estimate g in equation ?, taking $x = x_k$ and w equal to the union of all the elements in $x_0, x_1, \ldots, x_K$ except x_k. Then, given estimates of each g_k function, the function h may be estimated by nonparametrically regressing r on $g_1, \ldots, g_K, x_0$. In the same way our estimator immediately extends to models like

$$R(v, z_0, z_1, \ldots, z_k) = h[v + \sum_k g_k(z_k), z_0],$$

where each g_k is estimated by taking $z = z_k$ and w equal to the union of all the elements in $z_0, z_1, \ldots, z_K$ except z_k.

- Homothetic and homothetically separable functions are commonly used in models of consumer preferences and firm production, e.g., $r(x, w)$ could be a utility or consumer cost function recovered from estimated consumer demand functions via revealed preference theory, or it could be a directly estimated production or producer cost function. See, e.g., Blackorby, Primont, and Russell (1978), Lewbel (1991), (1997), Matzkin (1994), Primont and Primont (1994), and Zellner and Ryu (1998).

- Matzkin (1992) provides a consistent estimator for the binary threshold crossing model

$$y = I[s(x) + \varepsilon \geq 0]$$

where $s(x)$ is linearly homogeneous and ε is independent of x. This threshold crossing model has

$$E(y|x) = h[s(x)]$$

where h is the distribution function of $-\varepsilon$, and so is equivalent to our framework with $r(x) =$

$E(y|x)$ and w empty. In an unpublished manuscript, Newey and Matzkin (1993) propose an estimator of Matzkin's (1992) model.

- Models satisfying equation without imposing homogeneity on s are called weakly separable. See Gorman (1959), Goldman and Uzawa (1964) and Blackorby, Primont, and Russell (1978). Pinkse (2001) provides a general nonparametric estimator of weakly separable models. Pinkse's estimator identifies $s(x)$ up to an arbitrary monotonic transformation, whereas our estimator provides the unique (up to scale) linear homogeneous $s(x) = vg(z)$ (or equivalently g up to location in equation) and exploits this structure of $s(x)$ to obtain a faster rate of convergence than Pinkse.

- Horowitz and Mammen (2005) who estimate generalized additively separable models where both h and s_k are unknown functions.

$$r(x) = h[\sum_k s_k(x_k)].$$

- Matzkin (2003) considers models of the form

$$y = m(x, \varepsilon)$$

with an unobserved scalar ε independent of x and, as one possible identifying assumption, m being linearly homogeneous in x and ε.

Poisson Mixture Models

- Suppose that we observe bivariate count data (Y_1, Y_2) with associated covariates $(X_1, X_2) \in R^{d_1+d_2}$. There appears to be overdispersion and excess zeros in the data relative to the predictions of a pure Poisson model, which motivates the generalization to allow for heterogeneity. Conditional on a heterogeneity parameter ν and covariates $(X_1, X_2) = (x_1, x_2)$, Y_1 and Y_2 are independent Poisson with parameters

$$\begin{aligned}\lambda_1 &= \gamma_1(x_1) \cdot \nu \\ \lambda_2 &= \gamma_2(x_2) \cdot \nu,\end{aligned}$$

  where $\gamma_j(x_j)$ is some index function of the covariates.

- Suppose that the random effect ν is independent of the covariates and has density $g(\cdot)$, then the conditional frequency function $f(y_1, y_2|x_1, x_2)$ satisfies:

$$\begin{aligned} & f(y_1, y_2|x_1, x_2) \\ = & \int_0^\infty \left[\prod_{j=1}^{2} \frac{\exp(-\gamma_j(x_j) \cdot \nu)(\gamma_j(x_j) \cdot \nu)^{y_j}}{\Gamma(y_j + 1)} \right] g(\nu) d\nu. \end{aligned}$$

- The distribution of Y_1 is not independent of Y_2 conditional only on the observed covariates X_1, X_2. This captures the idea that there is a common component in the two outcome variables. Note that

$$\begin{aligned} \text{cov}[Y_1, Y_2|X_1 &= x_1, X_2 = x_2] \\ &= \gamma_1(x_1)\gamma_2(x_2)\sigma_\nu^2, \end{aligned}$$

which is strictly positive whenever $\sigma_\nu^2, \gamma_1(x_1), \gamma_2(x_2) > 0$. However, $\text{cov}(Y_1, Y_2)$ could be negative.

- The question we address is how to estimate efficiently the index functions γ_j and functionals of

f efficiently without specifying restrictions on g. Gurmu and Elder (2000) propose a method of doing some of this by parameterizing g and $\gamma_j(x_j)$. Hengartner (1997) has studied the question of estimation of functionals of g in a scalar Poisson mixture model when there are no covariates. Although g is nonparametrically identified, it can only be estimated at logarithmic rates even for g very smooth.

- Application to the study of fertility decision based on new data from Ireland at the turn of the century. Fertility is intimately connected with infant mortality and so we model these count variables jointly. The variables are children ever born CEB and children dead CD. By construction these are non-negative integer valued. In addition, there is the constraint that

$$CD \leq CEB.$$

This constraint is hard to impose directly in the usual models. We propose instead to define the variables $y_1 = CEB - CD$ and $y_2 = CD$ and to model these variables and then to go back to the implications for the raw variables.

- Suppose that x_j are scalar, different, and continuously distributed and that $\gamma_1(x_1)$ and $\gamma_2(x_2)$ are unknown but smooth functions. This model can be interpreted as a nonparametric double index model where there is a restriction on the way the two indexes $\gamma_1(x_1)$ and $\gamma_2(x_2)$ enter, specifically,

$$\begin{aligned} & f(y_1, y_2|x_1, x_2) \\ = \; & \left[\prod_{j=1}^{2} \frac{\gamma_j(x_j)^{y_j}}{\Gamma(y_j+1)}\right] \\ & \times E_v\left[\exp(-(\gamma_1(x_1)+\gamma_2(x_2))\nu)\cdot \nu^{y_1+y_2}\right] \\ = \; & \left[\prod_{j=1}^{2} \frac{\gamma_j(x_j)^{y_j}}{\Gamma(y_j+1)}\right] F(\gamma_1(x_1)+\gamma_2(x_2), y_1+y_2), \end{aligned}$$

where F is an unknown function of two scalar arguments, a continuous variable and an integer valued variable. Inside F is only a single index, the sum of the two indexes, but outside both indexes enter in a different way. Likewise inside F only the sum of the counts enters but outside both counts enter in a different way. This is something between an additive and multiplicative structure.

- When F, γ_1, γ_2 are unknown there is an identification issue, namely, let $\gamma_j^*(x_j) = c\gamma_j(x_j)$, $j = 1, 2$, and $F^*(\gamma, y) = F(\gamma/c, y)/c^y$ for any $c \neq 0$, then $(F^*, \gamma_1^*, \gamma_2^*)$ yields observational equivalence to (F, γ_1, γ_2). Therefore, we shall assume that $E[\nu] = 1$. Then

$$E\left[Y_j | X_j = x_j\right] = \gamma_j(x_j)$$

for $j = 1, 2$. Furthermore,

$$\text{var}[Y_j | X_j = x_j] = \gamma_j(x_j) + \sigma_\nu^2 \gamma_j^2(x_j)$$

for $j = 1, 2$, where $\sigma_\nu^2 = \text{var}[\nu]$. This means that the functions $\gamma_j(x_j)$ are identified from $E\left[Y_j | X_j = x_j\right]$. Given f one can obtain F.

Estimation

- Consider the model
$$G\left\{m(x)\right\} = c + \sum_{\alpha=1}^{d} m_\alpha(x_\alpha).$$
Estimation of $m_\alpha(x_\alpha)$ by marginal integration can be carried out in the analogous fashion, since
$$\begin{aligned} g_\alpha(x_\alpha) &= \int G(m(x))dQ_{-\alpha}(x_{-\alpha}) \\ &= c + m_\alpha(x_\alpha) + \sum_{\gamma\neq\alpha}^{d} \int m_\gamma(x_\gamma)dQ_{-\alpha}(x_{-\alpha}) \\ &\equiv m_\alpha(x_\alpha) + \mu_\alpha. \end{aligned}$$

- Let $\widehat{m}(x)$ be an estimator of $E(Y|X = x)$. Then let
$$\tilde{g}_\alpha(x_\alpha) = \int G(\widehat{m}(x))dQ_{-\alpha}(x_{-\alpha})$$
$$\widetilde{m}_\alpha(x_\alpha) = \tilde{g}_\alpha(x_\alpha) - \int \tilde{g}_\alpha(x_\alpha)dQ_\alpha(x_\alpha)$$

$$
\begin{aligned}
\widetilde{m}(x) &= F\left(\widehat{c} + \sum_{\alpha=1}^{d} \widetilde{m}_\alpha(x_\alpha)\right), \\
\widehat{c} &= \int G(\widehat{m}(x))dQ(x).
\end{aligned}
$$

- The instrumental variable approach can also be applied to generalized additive models. Under additivity we have

  $$m_\alpha(X_\alpha) = \frac{E\left[ZG(m\left(X\right))|X_\alpha\right]}{E\left[Z|X_\alpha\right]}$$

  for instrumental Z defined as

  $$Z(X) = \frac{f_\alpha(X_\alpha)f_{-\alpha}\left(X_{-\alpha}\right)}{f\left(X\right)}.$$

- Since $m\left(\cdot\right)$ is unknown, we need also consistent estimates of $m\left(X\right)$ in a preliminary step, and then the calculation in is feasible.

Homothetically Separable

- Consider

$$R(v, z, w) = H[vG(z), w],$$

where for identification

$$EG(Z) = 1.$$

- We start with a consistent estimator $\widehat{R}(\cdot)$ of the function $R(\cdot)$, and provide nonparametric estimators for G and H. Estimates of the original g and h can then be readily recovered from the estimates of G and H if desired.

- We first construct an initial consistent estimator of $G(z)$ by matching. For given values v, z, z', w suppose we can find a scalar u such that

$$R(v, z, w) = R(vu, z', w),$$

a match.

- Then

$$u = U(z, z') = G(z)/G(z').$$

- Then we have

$$G\left(z\right) = U(z, z')/E[U(Z, z')]$$

under the identification condition that $E[G(Z)] = 1$. The matching idea can be improved by averaging over the variables v, w.

- One advantage of a scale normalization like this over more simply normalizing at a point like $G(z_0) = 1$ is that the resulting limiting distributions at every point z will then not depend upon the distribution of $\widehat{R}(v, z_0, w)$.

- Given the function G, the function H can be defined as the conditional expectation

  $$H(\gamma, w) = E\left[R(V, Z, W) \mid VG(Z) = \gamma, W = w\right].$$

  Therefore, given an estimate $\widehat{G}$ of G, we can estimate the function H by a regression smooth of $\widehat{R}_i$ on $V_i\widehat{G}(Z_i), W_i$.

Asymptotic Properties

- We first give the properties of the marginal integration estimators defined in GAM, where $\widehat{m}(x)$ is the local constant.

Theorem 0.1 (Linton and Härdle (1996)) *Suppose that A1-A6 hold for* $r = 2$ *and that* F, G *are twice continuously differentiable over the relevant compact interval. Then,*

$$\begin{aligned} n^{2/5}\left[\widetilde{m}_\alpha(x_\alpha) - m_\alpha(x_\alpha)\right] &\Longrightarrow N\left[b_\alpha(x_\alpha), v_\alpha(x_\alpha)\right] \\ n^{2/5}\left[\widetilde{m}(x) - m(x)\right] &\Longrightarrow N\left[b(x), v(x)\right] \end{aligned}$$

in distribution, where $b(x) = F'(c+\sum_{\alpha=1}^d m_\alpha(x_\alpha)) \sum_{\alpha=1}^d b_\alpha($ $v(x) = \{F'(c+\sum_{\alpha=1}^d m_\alpha(x_\alpha))\}^2 \sum_{\alpha=1}^d v_\alpha(x_\alpha)$ *and*

$$\begin{aligned} b_\alpha(x_\alpha) &= \frac{\delta^2}{2}\mu_2(K)\left[\int G'(m(x)) b_{NW}(x) q_{-\alpha}(x_{-\alpha}) dx_{-\alpha}\right] \\ v_\alpha(x_\alpha) &= \delta^{-1}\|K\|_2^2 \int G'(m(x))^2 \sigma^2(x) \frac{q_{-\alpha}^2(x_{-\alpha})}{f(x)} dx_{-\alpha}. \end{aligned}$$

- This result parallels the main result for the integration estimator in additive nonparametric regression.

- The variance has an additional factor due to $G'(m(x))^2$ that is equal to one in the additive model.

- Note that even though $G(m(x))$ is additive, the function m is not, which explains the form of the bias.

- We have for example,

$$\begin{aligned} \partial m(x)/\partial x_\alpha &= F'(G(m(x))m'(x_\alpha) \\ \partial^2 m(x)/\partial x_\alpha^2 &= F''(G(m(x))(m'(x_\alpha))^2 \\ &\quad +F''(G(m(x))m''(x_\alpha) \end{aligned}$$

  so one can simplify the bias formula a bit.

- These estimators are inefficient as was shown in Linton (1997).

- The argument is basically that

$$
\begin{aligned}
& G(\widehat{m}(x)) - G(m(x)) \\
\simeq \; & G'(m(x))\{\widehat{m}(x) - m(x)\} \\
& + \frac{1}{2} G''(m(x))\{\widehat{m}(x) - m(x)\}^2,
\end{aligned}
$$

where the second term is of order $h^4 + 1/nh^d$. The first term is just like the usual estimation error for additive nonparametric regression.

Oracle Performance

- Our purpose here is to define a standard by which to measure estimators of the components. The notion of efficiency in nonparametric models is not as clear and well understood as it is in parametric models. In particular, pointwise mean squared error comparisons do not provide a simple ranking between estimators like kernel, splines, and nearest neighbors. While minimax efficiencies can in principle serve this purpose, they are hard to work with and even harder to justify. Our approach is to measure our procedures against a given infeasible (oracle) procedures for estimating $m_\alpha(x_\alpha)$ based on knowledge of c and $m_{-\alpha}(\cdot)$. Linton (1996) already defined the oracle estimator when $G(\cdot)$ is the identity function, i.e., when we are in the additive regression setting. In this case, one smooths the partial errors $Y_i - c + m_{-\alpha}(X_{-\alpha i})$ on the direction of interest $X_{\alpha i}$. He showed that indeed the oracle estimate has mean squared error smaller than the comparable integration-type estimator. In the general case though, one cannot find simple transformations of Y_i and $c + m_{-\alpha}(X_{-\alpha i})$ to which one can apply one-dimensional smoothing and that result in a more efficient procedure than the integration-type estimators. In sum, it was not immediately clear to us how to even define oracle efficiency in these nonlinear models. We suggest the following solution $-$ impose our knowledge about $c+m_{-\alpha}(X_{-\alpha i})$ inside of a suitable criterion function.

Partial Model Specification

- We suppose that the distribution $Y|X$ is completely unspecified apart from the restriction that

$$G(m(x)) = c + \sum_{\alpha=1}^{d} m_\alpha(x_\alpha).$$

- In this case, it is appropriate to look at least squares criteria that only imposes the mean information. Let $\theta = (\theta_0, \theta_1)$ and define the partial least squares criterion function

$$\begin{aligned} & Q_n(\theta) \\ = \; & \frac{1}{n} \sum_{i=1}^{n} K_h\left(x_\alpha - X_{\alpha i}\right) \\ & \times \left[Y_i - F\left\{c + m_{-\alpha}(X_{-\alpha i}) + \theta_0 + \theta_\alpha \cdot (X_{\alpha i} - x_c\right.\right. \end{aligned}$$

where $F = G^{-1}$. Let $\widehat{\theta} = (\widehat{\theta}_0, \widehat{\theta}_1)$ minimize $Q_n(\theta)$, and let $\widetilde{m}_\alpha(x_\alpha) = \widehat{\theta}_0(x_\alpha)$.

- We call $\widetilde{m}_\alpha(x_\alpha)$ an oracle estimator because its definition uses knowledge that only an oracle could have. A variety of smoothing paradigms could have been chosen here, and each will result in an 'oracle' estimate.

Theorem 0.2 *Suppose that conditions A given in the appendix hold. Then,*

$$n^{2/5}[\widetilde{m}_\alpha(x_\alpha) - m_\alpha(x_\alpha)] \Longrightarrow N\left[b_\alpha(x_\alpha), v_\alpha(x_\alpha)\right],$$

where

$$\begin{aligned} b_\alpha(x_\alpha) &= \frac{1}{2}\mu_2(K)m''_\alpha(x_\alpha) \\ v_\alpha(x_\alpha) &= ||K||_2^2 \frac{v_\alpha(x_\alpha)}{j_\alpha^2(x_\alpha)} \end{aligned}$$

$$\begin{aligned} j_\alpha(x_\alpha) &= \int F'\left[G\left\{m(x)\right\}\right]^2 f_X(x)dx_{-\alpha} \\ v_\alpha(x_\alpha) &= \int \sigma^2(x)F'\left[G\left\{m(x)\right\}\right]^2 f_X(x)dx_{-\alpha}, \end{aligned}$$

- When $Y|X$ is homoskedastic with constant variance σ^2, $v_\alpha(x_\alpha)$ is proportional to $i_\alpha(x_\alpha)$ and one obtains the simpler asymptotic variance

$$\frac{1}{nh}||K||_2^2 \frac{\sigma^2}{j_\alpha(x_\alpha)}.$$

- In this case, the asymptotic variance is less than that of the Linton and Härdle (1996) procedure. The relevant comparison is between

$$V_{LH} = \int \frac{1}{F' \left[G\left\{m(x)\right\}\right]^2} \frac{q^2(x_{-\alpha})}{f(x)} dx_{-\alpha},$$

[we have used the fact that $G'\left\{m(x)\right\} = 1\left/F'\left[G\left\{m(x\right.\right.\right.$

and

$$V_E = \frac{1}{\int F' \left[G\left\{m(x)\right\}\right]^2 f(x) dx_{-\alpha}}.$$

- Applying the Cauchy-Schwarz inequality, one obtains

$$V_E \leq V_{LH},$$

and the oracle estimator has lower variance than the integration estimator. In the heteroskedastic case, however, it is not possible to (uniformly) rank the two estimators unless the form of heteroskedasticity is restricted in some way, see the next section.

- The bias of $\widetilde{m}_\alpha(x_\alpha)$ is what you would expect if $c+m_{-\alpha}(\cdot)$ were known to be exactly zero. In the Linton and Härdle procedure there is an additional multiplicative factor to the bias

$$\int \frac{q_{-\alpha}(x_{-\alpha})}{F'\left[G\left\{m(x)\right\}\right]} dx_{-\alpha},$$

which can be either greater or less than one.

- Note that $\widetilde{m}_\alpha(x_\alpha)$ is not guaranteed to satisfy

$$\int \widetilde{m}_\alpha(x_\alpha) q_\alpha(x_\alpha) dx_\alpha = 0,$$

but the recentred estimate

$$\widetilde{m}^c_\alpha(x_\alpha) = \widetilde{m}_\alpha(x_\alpha) - \int \widetilde{m}_\alpha(x_\alpha) q_\alpha(x_\alpha) dx_\alpha$$

does have this property.

- In fact, the variance of $\widetilde{m}^c_\alpha(x_\alpha)$ and $\widetilde{m}_\alpha(x_\alpha)$ are the same to first order, while the bias of $\widetilde{m}^c_\alpha(x_\alpha)$ has $m''_\alpha(x_\alpha)$ replaced by

$$m''_\alpha(x_\alpha) - \int m''_\alpha(x_\alpha) q_\alpha(x_\alpha) dx_\alpha.$$

- According to integrated mean squared error, then, we are better off recentering because

$$
\begin{aligned}
& \int \left\{ m''_\alpha(x_\alpha) - \int m''_\alpha(x_\alpha) q_\alpha(x_\alpha) dx_\alpha \right\}^2 q_\alpha(x_\alpha) dx_\alpha \\
\leq \; & \int \left\{ m''_\alpha(x_\alpha) \right\}^2 q_\alpha(x_\alpha) dx_\alpha.
\end{aligned}
$$

Full Model Specification

- In many situations, the entire conditional distribution of $Y|X$ is completely specified by the mean equation , e.g., in one-parameter exponential families like probit. Thus, suppose that the conditional distribution of Y given $X = x$ belongs to the family

  $$f_{Y|X}(y|x) = \exp\{yS(x) - b(S(x)) + c(y)\}$$

  for known functions $b(\cdot)$ and $c(\cdot)$. Suppose also that holds and that G is the canonical link, i.e., $G = (b')^{-\alpha}$, so that $b'(S(x)) = m(x)$ and $\sigma^2(x) = b''(S(x))$, while

  $$S(x) = (G \circ b')^{-1}(G(m(x))) \equiv S_0\{G(m(x))\}.$$

  See Gourieroux, Monfort, and Trognon (1984a,b) for parametric theory and applications in economics.

- We can take account of the additional information contained in by employing the partial pseudo-likelihood criterion function

$$\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} K_h\left(x_\alpha - X_{\alpha i}\right)\left\{Y_i S_i(\theta) - b(S_i(\theta))\right\},$$

$$S_i(\theta) = S_0\left\{c + m_{-\alpha}(X_{-\alpha i}) + \theta_0 + \theta_1(X_{\alpha i} - x_\alpha)\right\}.$$

Let $\widetilde{\theta}$ minimize $\ell_n(\theta)$, and let $\ddot{m}_\alpha(x_\alpha) = \widetilde{\theta}_0(x_\alpha)$ be our infeasible estimate of $m_\alpha(x_\alpha)$.

Theorem 0.3 *Under the regularity conditions A given in the appendix, we have under the specification ,*

$$n^{2/5}[\ddot{m}_\alpha(x_\alpha) - m_\alpha(x_\alpha)] \Longrightarrow N\left[b_\alpha(x_\alpha), v_\alpha(x_\alpha)\right],$$

where

$$\begin{aligned} b_\alpha(x_\alpha) &= \frac{\mu_2(K)}{2} m''_\alpha(x_\alpha) \\ v_\alpha(x_\alpha) &= ||K||_2^2 \frac{1}{i_\alpha(x_\alpha)}, \end{aligned}$$

$$i_\alpha(x_\alpha) = \int b''\left[S_0\left\{G(m(x))\right\}\right] S_0'\left[G(m(x))\right]^2 f(x) dx_{-\alpha}.$$

- This estimator is more efficient than both the integration estimator and the two-step estimator based on the least squares criterion when is true. The bias is as in Theorem 1, and is design adaptive.

- We next discuss a number of leading examples and calculate the information quantity $i_\alpha(x_\alpha)$ for them.

Examples

Binomial

- Suppose that $Y_i \in \{0,1\}$, and that $m(x) = \Pr(Y = 1|X = x)$, where for some known G we have . Then, take $S = \ln F/(1-F)$ and $b' = F = G^{-1}$. Therefore, under the asymptotic variance of $\ddot{m}_\alpha(x_\alpha)$ is proportional to $1/i_\alpha(x_\alpha)$, where

$$i_\alpha(x_\alpha) = \int \frac{F'\left[G\left\{m(x)\right\}\right]^2}{m(x)\left\{1-m(x)\right\}} f(x)dx_{-\alpha},$$

while the variance of the Linton and Härdle (1996) procedure is proportional to

$$V_{LH}(x_\alpha) = \int \frac{m(x)\left\{1-m(x)\right\} q^2_{-\alpha}(x_{-\alpha})dx_{-\alpha}}{F'\left[G\left\{m(x)\right\}\right]^2 f(x)},$$

where for any joint distribution we have

$$\frac{1}{i_\alpha(x_\alpha)} \leq V_{LH}(x_\alpha),$$

by the Cauchy-Schwarz inequality.

Poisson

- Suppose that $Y_i \in \{0, 1, 2, \ldots\}$ with conditional distribution
$$\Pr(Y = k | X = x) = \frac{m(x)^k}{k!} \exp[-m(x)]$$
for some function $m(x) = E(Y|X = x)$ that satisfies with $G = \log$. Poisson regression models with flexible form have been considered in Hausman, Hall, and Griliches (1984).

- The variance of the Linton and Härdle (1996) procedure is proportional to
$$\begin{aligned} & V_{LH}(x_\alpha) \\ = \; & \exp(-c - m_\alpha(x_\alpha)) \\ & \times \int \frac{\exp(-m_{-\alpha}(x_{-\alpha}))q^2_{-\alpha}(x_{-\alpha})dx_{-\alpha}}{f(x)}, \end{aligned}$$
while under
$$\frac{1}{i_\alpha(x_\alpha)} = \frac{\exp(-c - m_\alpha(x_\alpha))}{\int \exp(m_{-\alpha}(x_{-\alpha}))f(x)dx_{-\alpha}}.$$

Variance Models (ARCH)

- Suppose that with probability one $E(Y|X=x) = 0$ and

  $$\text{var}(Y_i|X_i = x) = \sigma^2(x) = F_\sigma \left[c_\sigma + \sigma_\alpha(x_\alpha) + \sigma_{-\alpha}(x_{-\alpha}\right.$$

  for some positive function F_σ. When $F_\sigma = \exp$ and $X_i = (Y_{i-1}, \ldots, Y_{i-d})$ we have the multiplicative volatility model of Yang and Härdle (1997).

- The partial least squares criterion would be

  $$Q_n(\theta) = \frac{-1}{2n} \sum_{i=1}^{n} K_h\left(x_\alpha - X_{\alpha i}\right) \left\{Y_i^2 - \sigma_i^2(\theta)\right\},$$

  $$\sigma_i^2(\theta) = F_\sigma \left[c_\sigma + \sigma_{-\alpha}(x_{-\alpha}) + \theta_0 + \theta_1(X_{\alpha i} - x_\alpha)\right],$$

  which is of the form with Y replaced by Y^2, while the partial pseudo-likelihood function [for which one posits that $Y|X=x$ is $N(0, \sigma^2(x))$] is

  $$\ell_n(\theta) = \frac{-1}{2n} \sum_{i=1}^{n} K_h\left(x_\alpha - X_{\alpha i}\right) \left\{\log \sigma_i^2(\theta) + \frac{Y_i^2}{\sigma_i^2(\theta)}\right\},$$

which is of the form with Y replaced by Y^2.

- In this case,

$$v_\alpha(x_\alpha) \\ = \int \left(\frac{\kappa_4(x)+2}{4}\right)\left[\frac{F'_\sigma}{F_\sigma}\{c_\sigma + \sigma_\alpha(x_\alpha) + \sigma_{-\alpha}(x_{-\alpha})\}\right]$$

$$j_\alpha(x_\alpha) = i_\alpha(x_\alpha) \\ = \frac{1}{2}\int\left[\frac{F'_\sigma}{F_\sigma}\{c_\sigma + \sigma_\alpha(x_\alpha) + \sigma_{-\alpha}(x_{-\alpha})\}\right]^2 f(x$$

where $\kappa_4(x)$ is the conditional fourth cumulant of $Y|X = x$. When $F_\sigma = \exp$, the information is constant, in fact $i_\alpha(x_\alpha) = 1/2$. When F_σ is the identity,

$$i_\alpha(x_\alpha) = \int F_\sigma^{-2}\{c + \sigma_\alpha(x_\alpha) + \sigma_{-\alpha}(x_{-\alpha})\}\, f(x)dx\,/2\,.$$