8 Semiparametric Single Index Models

8.1 Index Models

A object of interest such as the conditional density $f(y \mid x)$ or conditional mean $\mathrm{E}(y \mid x)$ is a single index model when it only depends on the vector x through a single linear combination $x'\beta$.

Most parametric models are single index, including Normal regression, Logit, Probit, Tobit, and Poisson regression.

In a semiparametric single index model, the object of interest depends on x through the function $g(x'\beta)$ where $\beta \in \mathbb{R}^k$ and $g : \mathbb{R} \to \mathbb{R}$ are unknown. g is sometimes called a *link function.* In single index models, there is only one nonparametric dimension. These methods fall in the class of dimension reduction techniques.

The semiparametric single index regression model is

$$\mathrm{E}(y \mid x) = g\left(x'\beta\right) \tag{1}$$

where g is an unknown link function.

The semiparametric single index binary choice model is

$$P(y = 1 \mid x) = \mathrm{E}(y \mid x) = g\left(x'\beta\right) \tag{2}$$

where g is an unknown distribution function. We use g (rather than, say, F) to emphasize the connection with the regression model.

In both contexts, the function g includes any location and level shift, so the vector X_i cannot include an intercept. The level of β is not identified, so some normalization criterion for β is needed. It is typically easier to impose this on β than on g. One approach is to set $\beta'\beta = 1$. A second approach is to set one component of β to equal one. (This second approach requires that this variable correctly has a non-zero coefficient.)

The vector X_i must be dimension 2 or larger. If X_i is one-dimensional, then β is simply normalized to one, and the model is the one-dimensional nonparametric regression $\mathrm{E}(y \mid x) = g(x)$ with no semiparametric component.

Identification of β and g also requires that X_i contains at least one continuously distributed variable, and that this variable has a non-zero coefficient. If not, $X_i'\beta$ only takes a discrete set of values, and it would be impossible to identify a continuous function g on this discrete support.

8.2 Single Index Regression and Ichimura's Estimator

The semiparametric single index regression model is

$$\begin{aligned} y_i &= g\left(X_i'\beta\right) + e_i \\ \mathrm{E}(e_i \mid X_i) &= 0 \end{aligned}$$

This model generalizes the linear regression model (which sets $g(z)$ to be linear), and is a restriction of the nonparametric regression model.

The gain over full nonparametrics is that there is only one nonparametric dimension, so the curse of dimensionality is avoided.

Suppose g were known. Then you could estimate β by (nonlinear) least-squares. The LS criterion would be

$$S_n(\beta, g) = \sum_{i=1}^{n} \left(y_i - g\left(X_i'\beta\right)\right)^2.$$

We could think about replacing g with an estimate $\hat{g}$, but since $g(z)$ is the conditional mean of y_i given $X_i'\beta = z$, g depends on β, so a two-step estimator is likely to be inefficient.

In his PhD thesis, Ichimura proposed a semiparametric estimator, published later in the Journal of Econometrics (1993).

Ichimura suggested replacing g with the leave-one-out NW estimator

$$\hat{g}_{-i}\left(X_i'\beta\right) = \frac{\sum_{j\neq i} k\left(\dfrac{(X_j - X_i)'\beta}{h}\right) y_j}{\sum_{j\neq i} k\left(\dfrac{(X_j - X_i)'\beta}{h}\right)}.$$

The leave-one-out version is used since we are estimating the regression at the i'th observation i.

Since the NW estimator only converges uniformly over compact sets, Ichimura introduces trimming for the sum-of-squared errors. The criterion is then

$$S_n(\beta) = \sum_{i=1}^{n} \left(y_i - \hat{g}_{-i}\left(X_i'\beta\right)\right)^2 1_i(b)$$

He is not too specific about how to pick the trimming function, and it is likely that it is not important in applications.

The estimator of β is then

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, S_n(\beta).$$

The criterion is somewhat similar to cross-validation. Indeed, Hardle, Hall, and Ichimura (Annals of Statistics, 1993) suggest picking β and the bandwidth h jointly by minimization of $S_n(\beta)$.

In his paper, Ichimura claims that the $\hat{g}_{-i}(X_i'\beta)$ could be replaced by any other uniformly consistent estimator and the consistency of $\hat{\beta}$ would be maintained, but his asymptotic normality result would be lost. In particular, his proof rests on the asymptotic orthogonality of the derviative of $\hat{g}_{-i}(X_i'\beta)$ with e_i, which holds since the former is a leave-one-out estimator, and fails if it is a conventional NW estimator.

8.3 Asymptotic Distriubution of Ichimura's Estimator

Let β_0 denote the true value of β.

The tricky thing is that $\hat{g}_{-i}\left(X_i'\beta\right)$ is not estimating $g(X_i'\beta_0)$, rather it is estimating

$$G\left(X_i'\beta\right) = \mathrm{E}\left(y_i \mid X_i'\beta\right) = \mathrm{E}\left(g\left(X_i'\beta_0\right) \mid X_i'\beta\right)$$

the second equality since $y_i = g(X_i'\beta_0) + e_i$.

That is

$$G\left(z\right) = \mathrm{E}\left(y_i \mid X_i'\beta = z\right)$$

and $G\left(X_i'\beta\right)$ is then evaluated at $X_i'\beta$.

Note that

$$G\left(X_i'\beta_0\right) = g\left(X_i'\beta_0\right)$$

but for other values of β,

$$G\left(X_i'\beta\right) \neq g\left(X_i'\beta\right)$$

Hardle, Hall, and Ichimura (1993) show that the LS criterion is asymptotically equivalent to replacing $\hat{g}_{-i}\left(X_i'\beta\right)$ with $G\left(X_i'\beta\right)$, so

$$S_n\left(\beta\right) \simeq S_n^*\ \left(\beta\right) = \sum_{i=1}^{n}\left(y_i - G\left(X_i'\beta\right)\right)^2.$$

This approximation is essentially the same as Andrews' MINPIN argument, and relies on the estimator $\hat{g}_{-i}\left(X_i'\beta\right)$ being a leave-one-out estimator, so that it is orthogonal with the error e_i.

This means that $\hat{\beta}$ is asymptotically equivalent to the minimizer of $S_n^*\ \left(\beta\right)$, a NLLS problem. As we know from the Econ710, the asymptotic distribution of the NLLS estimator is identical to least-squares on

$$X_i^* = \frac{\partial}{\partial\beta}G\left(X_i'\beta\right).$$

This implies

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \to_d N\left(0, V\right)$$

$$\begin{aligned} V &= Q^{-1}\Omega Q^{-1} \\ Q &= \mathrm{E}\left(X_i^* X_i^{*\prime}\right) \\ \Omega &= \mathrm{E}\left(X_i^* X_i^{*\prime} e_i^2\right) \end{aligned}$$

To complete the derivation, we now find this X_i^*.

As $\hat{\beta}$ is $n^{-1/2}$ consistent, we can use a Taylor expansion of $g\left(X_i'\beta_0\right)$ to find

$$g\left(X_i'\beta_0\right) \simeq g\left(X_i'\beta\right) + g^{(1)}\left(X_i'\beta\right)X_i'\left(\beta_0 - \beta\right)$$

where

$$g^{(1)}\left(z\right) = \frac{d}{dz}g\left(z\right).$$

Then

$$\begin{aligned} G\left(X_i'\beta\right) &= \mathrm{E}\left(g\left(X_i'\beta_0\right) \mid X_i'\beta\right) \\ &\simeq \mathrm{E}\left(g\left(X_i'\beta\right) + g^{(1)}\left(X_i'\beta\right) X_i'\left(\beta_0 - \beta\right) \mid X_i'\beta\right) \\ &= g\left(X_i'\beta\right) - g^{(1)}\left(X_i'\beta\right) \mathrm{E}\left(X_i \mid X_i'\beta\right)' \left(\beta - \beta_0\right) \end{aligned}$$

since $g\left(X_i'\beta\right)$ and $g^{(1)}\left(X_i'\beta\right)$ are measureable with respect to $X_i'\beta$. Another Taylor expansion for$g\left(X_i'\beta\right)$ yields that this is approximately

$$\begin{aligned} G\left(X_i'\beta\right) &\simeq g\left(X_i'\beta_0\right) + g^{(1)}\left(X_i'\beta\right)\left(X_i - \mathrm{E}\left(X_i \mid X_i'\beta\right)\right)'\left(\beta - \beta_0\right) \\ &\simeq g\left(X_i'\beta_0\right) + g^{(1)}\left(X_i'\beta_0\right)\left(X_i - \mathrm{E}\left(X_i \mid X_i'\beta_0\right)\right)'\left(\beta - \beta_0\right) \end{aligned}$$

the final approximation for β in a $n^{-1/2}$ neighborhood of β_0. (The error is of smaller stochastic order.)

We see that

$$X_i^* = \frac{\partial}{\partial \beta} G\left(X_i'\beta\right) \simeq g^{(1)}\left(X_i'\beta_0\right)\left(X_i - \mathrm{E}\left(X_i \mid X_i'\beta_0\right)\right).$$

Ichimura rigorously establishes this result.

This asymptotic distribution is slightly different than that which would be obtained if the function g were known a priori. In this case, the asymptotic design depends on X_i, not $\mathrm{E}\left(X_i \mid X_i'\beta_0\right)$.

$$Q = \mathrm{E}\left(g^{(1)}\left(X_i'\beta_0\right)^2 X_i X_i'\right)$$

This is the cost of the semiparametric estimation.

Recall when we described identification that we required the dimension of X_i to be 2 or larger. Suppose that X_i is one-dimensional. Then $X_i - \mathrm{E}\left(X_i \mid X_i'\beta_0\right) = 0$ so $Q = 0$ and the above theory is vacuous (as it should be).

The Ichimura estimator achieves the semiparametric efficiency bound for estimation of β when the error is conditionally homoskedastic. Ichimura also considers a weighted least-squares estimator setting the weight to be the inverse of an estimate of the conditional variance function (as in Robinson's FGLS estimator). This weighted LS estimator is then semiparametrically efficient.

8.4 Klein and Spady's Binary Choice Estimator

Klein and Spady (Econometrica, 1993) proposed an estimator of the semiparametric single index binary choice model which has strong similarities with Ichimura's estimator.

The model is

$$y_i = 1\left(X_i'\beta \geq e_i\right)$$

where e_i is an error.

If e_i is independent of X_i and has distribution function g, then the data satisfy the single-index regression

$$\mathrm{E}\left(y \mid x\right) = g\left(x'\beta\right).$$

It follows that Ichimura's estimator can be directly applied to this model.

Klein and Spady suggest a semiparametric likelihood approach. Given g, the log-likelihood is

$$L_n(\beta, g) = \sum_{i=1}^{n} \left(y_i \ln g\left(X_i'\beta\right) + (1 - y_i) \ln\left(1 - g\left(X_i'\beta\right)\right)\right).$$

This is analogous to the sum–of-squared errors function $S_n(\beta, g)$ for the semiparametric regression model.

Similarly with Ichimura, Klein and Spady suggest replacing g with the leave-one-out NW estimator

$$\hat{g}_{-i}\left(X_i'\beta\right) = \frac{\sum_{j\neq i} k\left(\dfrac{\left(X_j - X_i\right)'\beta}{h}\right) y_j}{\sum_{j\neq i} k\left(\dfrac{\left(X_j - X_i\right)'\beta}{h}\right)}.$$

Making this substitution, and adding trimming function, this leads to the feasible likelihood criterion

$$L_n(\beta) = \sum_{i=1}^{n} \left(y_i \ln \hat{g}_{-i}\left(X_i'\beta\right) + (1 - y_i) \ln\left(1 - \hat{g}_{-i}\left(X_i'\beta\right)\right)\right) 1_i(b).$$

Klein and Spady emphasize that the trimming indicator should not be a function of β, but instead of a preliminary estimator. They suggest

$$1_i(b) = 1\left(\hat{f}_{X'\tilde{\beta}}\left(X_i'\tilde{\beta}\right) \geq b\right)$$

where $\tilde{\beta}$ is a preliminary estimator of β, and $\hat{f}$ is an estimate of the density of $X_i'\tilde{\beta}$. Klein and Spady observe that trimming does not seem to matter in their simulations.

The Klein-Spady estimator for β is the value $\hat{\beta}$ which maximizes $L_n(\beta)$.

In many respects the Ichimura and Klein-Spady estimators are quite similar.

Unlike Ichimura, Klein-Spady impose the assumption that the kernel k must be fourth-order (e.g. bias reducing). They also impose that the bandwidth h satisfy the rate $n^{-1/6} < h < n^{-1/8}$, which is smaller than the optimal $n^{-1/9}$ rate for a $4th$ order kernel. It is unclear to me if these are merely technical sufficient conditions, or if there a substantive difference with the semiparametric regression case.

Klein and Spady also have no discussion about how to select the bandwidth. Following the ideas of Hardle, Hall and Ichimura, it seems sensible that it could be selected jointly with β by minimization of $L_n(\beta)$, but this is just a conjecture.

They establish the asymptotic distribution for their estimator. Similarly as in Ichimura, letting

g denote the distribution of e_i, define the function

$$G\left(X_i'\beta\right) = \mathrm{E}\left(g\left(X_i'\beta_0\right) \mid X_i'\beta\right).$$

Then

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \to_d N\left(0, H^{-1}\right)$$

$$H = \mathrm{E}\left(\frac{\partial}{\partial\beta}G\left(X_i'\beta\right)\frac{\partial}{\partial\beta}G\left(X_i'\beta\right)'\frac{1}{g\left(X_i'\beta_0\right)\left(1 - g\left(X_i'\beta_0\right)\right)}\right)$$

They are not specific about the derivative component, but if I understand it correctly it is the same as in Ichimura, so

$$\frac{\partial}{\partial\beta}G\left(X_i'\beta\right) \simeq g^{(1)}\left(X_i'\beta_0\right)\left(X_i - \mathrm{E}\left(X_i \mid X_i'\beta_0\right)\right).$$

The Klein-Spady estimator achieves the semiparametric efficiency bound for the single-index binary choice model.

Thus in the context of binary choice, it is preferable to use Klein-Spady over Ichimura. Ichimura's LS estimator is inefficient (as the regression model is heteroskedastic), and it is much easier and cleaner to use the Klein-Spady estimator rather than a two-step weighted LS estimator.

8.5 Average Derivative Estimator

Let the conditional mean be

$$\mathrm{E}\left(y \mid x\right) = \mu\left(x\right)$$

Then the derivative is

$$\mu^{(1)}(x) = \frac{\partial}{\partial x}\mu(x)$$

and a weighted average is

$$\mathrm{E}\left(\mu^{(1)}(X)w(X)\right)$$

where $w(x)$ is a weight function. It is particularly convenient to set $w(x) = f(x)$, the marginal density of X. Thus Powell, Stock and Stoker (Econometrica, 1989) define this as the average derivative

$$\delta = \mathrm{E}\left(\mu^{(1)}(X)f(X)\right).$$

This is a measure of the average effect of X on y. It is a simple vector, and therefore easier to report than a full nonparametric estimator.

There is a connection with the single index model, where

$$\mu\left(x\right) = g\left(x'\beta\right)$$

for then

$$\mu^{(1)}(x) = \beta g^{(1)}(x'\beta)$$

$$\delta = c\beta$$

where

$$c = \mathrm{E}\left(g^{(1)}(x'\beta) f(X)\right).$$

Since β is identified only up to scale, the constant c doesn't matter. That is, a (normalized) estimate of δ is an estimate of normalized β.

PSS observe that by integration by parts

$$\begin{aligned}
\delta &= \mathrm{E}\left(\mu^{(1)}(X) f(X)\right) \\
&= \int \mu^{(1)}(x) f(x)^2 dx \\
&= -2\int \mu(x) f(x) f^{(1)}(x) dx \\
&= -2\mathrm{E}\left(\mu(X) f^{(1)}(X)\right) \\
&= -2\mathrm{E}\left(y f^{(1)}(X)\right)
\end{aligned}$$

By the reasoning in CV, an estimator of this is

$$\hat{\delta} = -\frac{2}{n-1}\sum_{i=1}^{n} y_i \hat{f}_{(-i)}^{(1)}(X_i)$$

where $\hat{f}_{(-i)}(X_i)$ is the leave-one-out density estimator, and $\hat{f}_{(-i)}^{(1)}(X_i)$ is its first derivative.

This is a convenient estimator. There is no denominator messing with uniform convergence. There is only a density estimator, no conditional mean needed.

PSS show that $\hat{\delta}$ is $n^{-1/2}$ consistent and asy. normal, with a convenient covariance matrix.

The asymptotic bias is a bit complicated.

Let $q = \dim(X)$. Set $p = ((q+4)/2$ if q is even and $p = (q+3)/2)$ if q is odd. e.g. $p = 2$ for $q = 1$, $p = 3$ for $q = 2$ or $q = 3$ and $p = 4$ for $q = 4$.

PSS require that the kernel for estimation of f be of order at least p. Thus a second-order kernel for $q = 1$, a fourth order for $q = 2$, 3, or 4.

PSS then show that the asymptotic bias is

$$n^{1/2}\left(\mathrm{E}\hat{\delta} - \delta\right) = O\left(n^{1/2} h^p\right)$$

which is $o(1)$ if the bandwidth is selected so that $nh^{2p} \to 0$. This is violated (too big) if h is selected to be optimal for estimation of $\hat{f}$ or $\hat{f}^{(1)}$. This requirement needs the bandwidth to undersmooth to reduce the bias. This type of result is commonly seen in semiparametric methods. Unfortunately, it does not lead to a practical rule for bandwidth selection.