4 Conditional Distribution Estimation

4.1 Estimators

The conditional distribution (CDF) of y_i given $X_i = x$ is

$$\begin{aligned} F(y \mid x) &= P(y_i \leq y \mid X_i = x) \\ &= E(1(y_i \leq y) \mid X_i = x). \end{aligned}$$

This is the conditional mean of the random variable $1(y_i \leq y)$. Thus the CDF is a regression, and can be estimated using regression methods.

One difference is that $1(y_i \leq y)$ is a function of the argument y, so CDF estimation is a set of regressions, one for each value of y.

Standard CDF estimators include the NW, LL, and WNW. The NW can be written as

$$\hat{F}(y \mid x) = \frac{\sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right) 1(y_i \leq y)}{\sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right)}$$

The NW and WNW estimators have the advantages that they are non-negative and non-decreasing in y, and are thus valid CDFs.

The LL estimator does not necessarily satisfy these properties. It can be negative, and need not be monotonic in y.

As we learned for regression estimation, the LL and WMW estimators both have "better" bias and boundary properties. Putting these two observations together, it seems reasonable to consider using the WNW estimator.

The estimator $\hat{F}(y \mid x)$ is smooth in x, but a step function in y. We discuss later estimators which are smooth in y.

4.2 Asymptotic Distribution

Recall that in the case of kernel regression, we had

$$\sqrt{n|H|}\left(\hat{g}(x) - g(x) - \kappa_2 \sum_{j=1}^{q} h_j^2 B_j(x)\right) \xrightarrow{d} N\left(0, \frac{R(k)^q \sigma^2(x)}{f(x)}\right)$$

where $\sigma^2(x)$ was the conditional variance of the regression, and the $B_j(x)$ equals (for NW)

$$B_j(x) = \frac{1}{2}\frac{\partial^2}{\partial x_j^2} g(x) + f(x)^{-1} \frac{\partial}{\partial x_j} g(x) \frac{\partial}{\partial x_j} f(x)$$

while for LL and WNW the bias term is just the first part.

Clearly, for any fixed y, the same theory applies. In the case of CDF estimation, the regression

equation is

$$1\left(y_i \leq y\right) = F\left(y \mid X_i\right) + e_i\left(y\right)$$

where $e_i(y)$ is conditionally mean zero and has conditional variance function

$$\sigma^2(x) = F\left(y \mid x\right)\left(1 - F\left(y \mid x\right)\right).$$

(We know the conditional variance takes this form because the dependent variable is binary.) I write the error as a function of y to emphasize that it is different for each y.

In the case of LL or NWW, the bias terms are

$$B_j\left(y \mid x\right) = \frac{1}{2}\frac{\partial^2}{\partial x_j^2} F\left(y \mid x\right)$$

the curvature in the CDF with respect to the conditioning variables.

We thus find for all (y, x)

$$\sqrt{n\left|H\right|}\left(\hat{F}\left(y \mid x\right) - F\left(y \mid x\right) - \kappa_2 \sum_{j=1}^{q} h_j^2 B_j\left(y \mid x\right)\right) \xrightarrow{d} N\left(0, \frac{R(k)^q F\left(y \mid x\right)\left(1 - F\left(y \mid x\right)\right)}{f(x)}\right)$$

and

$$AMSE\left(\hat{F}\left(y \mid x\right)\right) = \kappa_2^2 \left(\sum_{j=1}^{q} h_j^2 B_j\left(y \mid x\right)\right)^2 + \frac{R(k)^q F\left(y \mid x\right)\left(1 - F\left(y \mid x\right)\right)}{n\left|H\right| f(x)}$$

In the $q = 1$ case

$$AMSE\left(\hat{F}\left(y \mid x\right)\right) = h^4 \kappa_2^2 B\left(y \mid x\right)^2 + \frac{R(k) F\left(y \mid x\right)\left(1 - F\left(y \mid x\right)\right)}{nhf(x)}.$$

In the regression case we defined the WIMSE as the integral of the AMSE, weighting by $f(x)M(x)$. Here we also integrate over y. For $q = 1$

$$\begin{aligned} WIMSE &= \int\int AMSE\left(\hat{F}\left(y \mid x\right)\right) f(x) M(x)\left(dx\right) dy \\ &= h^4 \kappa_2^2 \int\int B\left(y \mid x\right)^2 dy f(x) M(x)\left(dx\right) + \frac{R(k) \int\int F\left(y \mid x\right)\left(1 - F\left(y \mid x\right)\right) dy M(x) dx}{nh} \end{aligned}$$

The integral over y does not need weighting since $F\left(y \mid x\right)\left(1 - F\left(y \mid x\right)\right)$ declines to zero as y tends to either limit.

Observe that the converge rate is the same as in regression. The optimal bandwidths are the same rates as in regression.

4.3 Bandwidth Selection

I do not believe that bandwidth choice for nonparametric CDF estimation is widely studied. Li-Racine suggest using a CV method based on conditional density estimation.

It should also be possible to directly apply CV methods to CDF estimation.

The leave-one-out residuals are

$$\hat{e}_{i,i-1}(y) = 1(y_i \leq y) - \hat{F}_{-i}(y \mid X_i)$$

So the CV criterion for any fixed y is

$$\begin{aligned} CV(y,h) &= \frac{1}{n}\sum_{i=1}^{n} \hat{e}_{i,i-1}(y)^2 M(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \left(1(y_i \leq y) - \hat{F}_{-i}(y \mid X_i)\right)^2 M(X_i) \end{aligned}$$

If you wanted to estimate the CDF at a single value of y you could pick h to minimize this criterion.

For estimation of the entire function, we want to integrate over the values of y. One method is

$$\begin{aligned} CV(h) &= \int CV(y,h)dy \\ &\simeq \delta \sum_{j=1}^{N} CV(y_j^*, h) \end{aligned}$$

where y_j^* is a grid of values over the support of y_i such that $y_j - y_{j=1} = \delta$. To calculate this quantity, it involves N times the number of calculations as for regression, as the leave-one-out computations are done for each y_j^* on the grid. My guess is that the grid over the y values could be coarse, e.g. one could set $N = 10$.

4.4 Smoothed Distribution Estimators - Unconditional Case

The CDF estimators introduced above are not smooth, but are discontinuous step functions. For some applications this may be inconvenient. It may be desireable to have a smooth CDF estimate as an input for a semiparametric estimator. It is also the case that smoothing will improve high-order estimation efficiency. To see this, we need to return to the case of univariate data.

Recall that the univariate DF estimator for iid data y_i is

$$\hat{F}(y) = \frac{1}{n}\sum_{i=1}^{n} 1(y_i \leq y)$$

It is easy to see that this estimator is unbiased and has variance $F(y)(1 - F(y))/n$.

Now consider a smoothed estimator

$$\tilde{F}(y) = \frac{1}{n}\sum_{i=1}^{n} G\left(\frac{y-y_i}{h}\right)$$

where $G(x) = \int_{-\infty}^{x} k(u)du$ is a kernel distribution function (the integral of a univariate kernel function). Thus $\tilde{F}(y) = \int_{-\infty}^{y} \hat{f}(x)dx$ where $\hat{f}(x)$ is the kernel density estimate.

To calculate its expectation

$$\begin{aligned}
\mathrm{E}\tilde{F}(y) &= \mathrm{E}G\left(\frac{y-y_i}{h}\right) \\
&= \int G\left(\frac{y-x}{h}\right) f(x)dx \\
&= h\int G(u)\, f(y-hu)\, du
\end{aligned}$$

the last using the change of variables $u = (y-x)/h$ or $x = y - hu$ with Jacobian h.

Next, do not expand $f(y-hu)$ in a Taylor expansion, because the moments of G do not exist. Instead, first use integration by parts. The integral of f is F and that of $hf(y-hu)$ is $-F(y-hu)$, and the derivative of $G(u)$ is $k(u)$. Thus the above equals

$$\int k(u)\, F(y-hu)\, du$$

which can now be expanded using Taylor's expansion, yielding

$$\mathrm{E}\tilde{F}(y) = F(y) + \frac{1}{2}\kappa_2 h^2 f^{(1)}(y) + o\left(h^2\right)$$

Just as in other estimation contexts, we see that the bias of $\tilde{F}(y)$ is of order h^2, and is proportional to the second derivative of what we are estimating, as $F^{(2)}(y) = f^{(1)}(y)$

Thus smoothing introduces estimation bias.

The interesting part comes in the analysis of variance.

$$\begin{aligned}
var\left(\tilde{F}(y)\right) &= \frac{1}{n} var\left(G\left(\frac{y-y_i}{h}\right)\right) \\
&= \frac{1}{n}\left(\mathrm{E}G\left(\frac{y-y_i}{h}\right)^2 - \left(\mathrm{E}G\left(\frac{y-y_i}{h}\right)\right)^2\right) \\
&\simeq \frac{1}{n}\left(\int G\left(\frac{y-x}{h}\right)^2 f(x)dx - F(y)^2\right)
\end{aligned}$$

Let's calculate this integral. By a change of variables

$$\int G\left(\frac{y-x}{h}\right)^2 f(x)dx = h\int G(u)^2 f(y-hu)du$$

Once again we cannot direct apply a Taylor expansion, but need to first do integration-by-parts. Again the integral of $hf(y-hu)$ is $-F(y-hu)$. The derivative of $G(u)^2$ is $2G(u)k(u)$. So the above is

$$2\int G(u)\,k(u)F(y-hu)\,du$$

and then applying a Taylor expansion, we obtain

$$F(y)\left(2\int G(u)\,k(u)du\right)-f(y)\,h\left(2\int G(u)\,k(u)udu\right)+o(h)$$

since $F^{(1)}(y)=f(y)$.

Now since the derivative of $G(u)^2$ is $2G(u)k(u)$, it follows that the integral of $2G(u)k(u)$ is $G(u)^2$, and thus the first integral over $(-\infty,\infty)$ is $G(\infty)^2-G(-\infty)^2=1-0=1$ since $G(u)$ is a distribution function. Thus the first part is simply $F(y)$. Define

$$\alpha(k)=2\int G(u)\,k(u)udu>0$$

For any symmetric kernel k, $\alpha(k)>0$. This is because for $u>0$, $G(u)>G(-u)$, thus

$$\int_0^\infty G(u)\,k(u)udu>\int_0^\infty G(-u)\,k(u)udu=-\int_{-\infty}^0 G(u)\,k(u)udu$$

and so the integral over $(-\infty,\infty)$ is positive. Integrated kernels and the value $\alpha(k)$ are given in the following table.

Kernel	**Integrated Kernel**	$\alpha(k)$
Epanechnikov	$G_1(u)=\frac{1}{4}\left(2+3u-u^3\right)1(\lvert u\rvert\leq 1)$	$9/35$
Biweight	$G_2(u)=16\left(8+15u-10u^3+3u^5\right)1(\lvert u\rvert\leq 1)$	$50/231$
Triweight	$G_3(u)=32\left(16+35u-35u^2+21u^5-5u^7\right)1(\lvert u\rvert\leq 1)$	$245/1287$
Gaussian	$G_\phi(u)=\Phi(u)$	$1/\sqrt{\pi}$

Together, we have

$$\begin{aligned}
var\left(\tilde{F}(y)\right) &\simeq \frac{1}{n}\left(\int G\left(\frac{y-x}{h}\right)^2 f(x)dx-F(y)^2\right)\\
&= \frac{1}{n}\left(F(y)-F(y)^2-\alpha(k)\,f(y)\,h+o(h)\right)\\
&= \frac{F(y)(1-F(y))}{n}-\alpha(k)\,f(y)\,\frac{h}{n}+o\left(\frac{h}{n}\right)
\end{aligned}$$

The first part is the variance of $\hat{F}(x)$, the unsmoothed estimator. Smoothing reduces the variance by $\alpha_0 f(y)\,\frac{h}{n}$.

Its MSE is

$$MSE\left(\tilde{F}(y)\right) = \frac{F(y)\left(1-F(y)\right)}{n} - \alpha\left(k\right) f\left(y\right) \frac{h}{n} + \frac{\kappa_2^2 h^4}{4} f^{(1)}(y)^2$$

The integrated MSE is

$$\begin{aligned} MISE\left(\tilde{F}(y)\right) &= \int MSE\left(\tilde{F}(y)\right) dy \\ &= \frac{\int F(y)\left(1-F(y)\right) dy}{n} - \alpha\left(k\right) \frac{h}{n} + \frac{\kappa_2^2 h^4 R\left(f^{(1)}\right)}{4} \end{aligned}$$

where

$$R\left(f^{(1)}\right) = \int f^{(1)}(y)^2 dy$$

The first term is independent of the smoothing parameter h (and corresponds to the integrated variance of the unsmoothed EDF estimator). To find the optimal bandwidth, take the FOC:

$$\frac{d}{dh} MISE\left(\hat{F}(y)\right) = -\frac{\alpha\left(k\right)}{n} + \kappa_2^2 h^3 R\left(f^{(1)}\right) = 0$$

and solve to find

$$h_0 = \left(\frac{\alpha\left(k\right)}{\kappa_2^2 R\left(f^{(1)}\right)}\right)^{1/3} n^{-1/3}$$

The optimal bandwidth converges to zero at the fast $n^{-1/3}$ rate.

Does smoothing help? The unsmoothed estimator has MISE of order n^{-1}, and the smoothed estimator (with optimal bandwidth) is of order $n^{-1} - n^{-4/3}$. We can thus think of the gain in the scaled MISE as being of order $n^{-4/3}$, which is of smaller order than the original n^{-1} rate.

It is important that the bandwidth not be too large. Suppose you set $h \propto n^{-1/5}$ as for density estimation. Then the square bias term is of order $h^4 \propto n^{-4/5}$ which is larger than the leading term. In this case the smoothed estimator has larger MSE than the usual estimator! Indeed, you need h to be of smaller order than $n^{-1/4}$ for the MSE to be no worse than the unusual case.

For practical bandwidth selection, Li-Racine and Bowman et. al. (1998) recommend a CV method. For fixed y the criterion is

$$CV(h,y) = \frac{1}{n}\sum_{i=1}^{n}\left(1\left(y_i \leq y\right) - \tilde{F}_{-i}\left(y\right)\right)^2$$

which is the sum of squared leave-one-out residuals. For a global estimate the criterion is

$$CV(h) = \int CV(h,y) dy$$

and this can be approximated by a summation over a grid of values for y.

This is essentially the same as the CV criterion we introduced above in the conditional case.

4.5 Smoothed Conditional Distribution Estimators

The smoothed versions of the CDF estimators replace the indicator functions $1(y_i \leq y)$ with the integrated kernel $G\left(\frac{y-y_i}{h_0}\right)$ where we will use h_0 to denote the bandwidth smoothing in the y direction.

The NW version is

$$\tilde{F}(y \mid x) = \frac{\sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right) G\left(\frac{y-y_i}{h_0}\right)}{\sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right)}$$

with $H = \{h_1, ...h_q\}$. The LL is obtained by a local linear regression of $G\left(\frac{y-y_i}{h_0}\right)$ on $X_i - x$ with bandwidths H. And similarly the WNW.

What is its distribution? It is essentially that of $\hat{F}(y \mid x)$, plus an additional bias term, minus a variance term.

First take bias. Recall

$$Bias\left(\hat{F}(y \mid x)\right) \simeq \kappa_2 \sum_{j=1}^{q} h_j^2 B_j(y \mid x)$$

where for LL and WNW

$$B_j(y \mid x) = \frac{1}{2}\frac{\partial^2}{\partial x_j^2} F(y \mid x).$$

And for smoothed DF estimation, the bias term is

$$\kappa_2 h^2 \frac{1}{2}\frac{\partial^2}{\partial y^2} F(y)$$

If you work out the bias of the smoothed CDF, you find it is the sum of these two, that is $\tilde{F}(y \mid x)$

$$Bias\left(\tilde{F}(y \mid x)\right) \simeq \kappa_2 \sum_{j=0}^{q} h_j^2 B_j(y \mid x)$$

where for $j \geq 1$ the $B_j(y \mid x)$ are the same as before, and for $j = 0$,

$$B_0(y \mid x) = \frac{1}{2}\frac{\partial^2}{\partial y^2} F(y \mid x).$$

For variance, recall

$$var\left(\hat{F}(y \mid x)\right) = \frac{R(k)^q F(y \mid x)(1 - F(y \mid x))}{f(x) n |H|}$$

and for smoothed DF estimation, the variance was reduced by the term $-\alpha_0 f(y)\frac{h}{n}$. In the CDF

case it turns out to be similarly adjusted:

$$var\left(\tilde{F}\left(y \mid x\right)\right) = \frac{R(k)^q \left[F\left(y \mid x\right)\left(1 - F\left(y \mid x\right)\right) - h_0 \alpha\left(k\right) f\left(y \mid x\right)\right]}{f(x) n \left|H\right|}$$

In sum, the MSE is

$$MSE\left(\tilde{F}\left(y \mid x\right)\right) = \kappa_2^2 \left(\sum_{j=0}^{q} h_j^2 B_j\left(y \mid x\right)\right)^2 + \frac{R(k)^q \left[F\left(y \mid x\right)\left(1 - F\left(y \mid x\right)\right) - h_0 \alpha\left(k\right) f\left(y \mid x\right)\right]}{f(x) n \left|H\right|}$$

The WIMSE, $q = 1$ case, is

$$\begin{aligned} WIMSE &= \int\int AMSE\left(\tilde{F}\left(y \mid x\right)\right) f(x) M(x)\left(dx\right) dy \\ &= \kappa_2^2 \int\int \left(h_0^2 B_0\left(y \mid x\right) + h_1^2 B_1\left(y \mid x\right)\right)^2 dy f(x) M(x)\left(dx\right) \\ &\quad + \frac{R(k)\left[\int\int F\left(y \mid x\right)\left(1 - F\left(y \mid x\right)\right) dy M(x) dx - h_0 \alpha(k) \int M(x) dx\right]}{n h_1} \end{aligned}$$

4.6 Bandwidth Choice

First, consider the optimal bandwidth rates.

As smoothing in the y direction only affects the higher-order asymptotic distribution, it should be clear that the optimal rates for $h_1, ..., h_q$ is unchanged from the unsmoothed case, and is therefore equal to the regression setting. Thus the optimal bandwidth rates are $h_j \sim n^{-1/(4+q)}$ for $j \geq 1$.

Substituting these rates into the MSE equation, and ignoring constants, we have

$$MSE\left(\tilde{F}\left(y \mid x\right)\right) \sim \left(h_0^2 + n^{-2/(4+q)}\right)^2 + \frac{1}{n^{4/(4+q)}} - \frac{h_0}{n^{4/(4+q)}}$$

Differentiating with respect to h_0

$$0 = 4\left(h_0^2 + n^{-2/(4+q)}\right) h_0 - \frac{1}{n^{4/(4+q)}}$$

and since h_0 will be of smaller order than $n^{-1/(4+q)}$, we can ignore the h_0^3 term, and then solving the remainder we obtain $h_0 \sim n^{-2/(4+q)}$. E.g. for $q = 1$ then the optimal rate is $h_0 \sim n^{-2/5}$.

What is the gain from smoothing? With optimal bandwidth, the MISE is reduced by a term of order $n^{-6/(4+q)}$. This is $n^{-6/5}$ for $q = 1$ and n^{-1} for $q = 2$. This gain increases as q increases. Thus the gain in efficiency (from smoothing) is increased when X is of higher dimension. Intuitively, increasing X is equivalent to reducing the effective sample size, increasing the gain from smoothing.

How should the bandwidth be selected?

Li-Racine recommend picking the bandwidths by using a CV method for conditional density estimation, and then rescaling.

As an alternative, we can use CV directly for the CDF estimate. That is, define the CV criterion

$$\begin{aligned} CV(y,h) &= \frac{1}{n}\sum_{i=1}^{n}\left(1\left(y_i \leq y\right) - \tilde{F}_{-i}\left(y \mid X_i\right)\right)^2 M\left(X_i\right) \\ CV(h) &= \int CV(h,y)dy \end{aligned}$$

where $h = (h_0, h_1, ..., h_q)$ includes smoothing in both the y and x directions. The estimator $\tilde{F}_{-i}$ is the smooth leave-one-out estimator of F. This formulae allows includes NW, LL and WNW estimation.

The second integral can be approximated using a grid.

To my knowledge, this procedure has not been formally investigated.