USC, Fall 2016, Economics 513

Lecture 12: Nonparametric regression and regression discontinuity I

Best fitting relations and conditional expectation

In the linear relation

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i$$

the regression coefficients are the partial effects if ε_i is unrelated to the independent variables $x_{i2}, \ldots, x_{iK}$. However is some cases we are not interested in partial effects, but only want the best fitting relation between y and $x_2, \ldots, x_K$.

We consider an example with one explanatory variable. The dependent variable y is the savings rate and the independent variable x is household income. The data are random sample $y_i, x_i, i = 1, \ldots, n$ of US households.

If y and x are discrete or in this case discretized by using income and savings rate brackets we can use the sample to estimate the joint population density of y, x. See Table 1.1.

From the estimate $\hat{f}(y, x)$ of the joint population density we can obtain an estimate of the conditional density of y given x

$$\hat{f}(y|x) = \frac{\hat{f}(y, x)}{\hat{f}(x)}$$

See Table 1.2

If there is an exact relation between y and x, every column has one 1 and rest 0's. If the relation is not exact, we can consider the average of y for every value of x. This is the conditional mean function. The population conditional mean function is

$$\mu(x) = E(y|x) = \sum_y y f(y|x)$$

and its estimator is

$$\hat{\mu}(x) = \sum_y y \hat{f}(y|x)$$

The estimator is plotted in Figure 1.1

This is an estimator and the roughness could be sampling variation around a smooth $\mu(x)$.

Table 1.1 Joint frequency distribution of X = income and Y = savings rate.

	X									
Y	0.5	1.5	2.5	3.5	4.5	5.5	6.7	8.8	12.5	17.5
.50	.001	.011	.007	.006	.005	.005	.008	.009	.014	.004
.40	.001	.002	.006	.007	.010	.007	.008	.009	.008	.007
.25	.002	.006	.004	.007	.010	.011	.020	.019	.013	.006
.15	.002	.009	.009	.012	.016	.020	.042	.054	.024	.020
.05	.010	.023	.033	.031	.041	.029	.047	.039	.042	.007
0	.013	.013	.000	.002	.001	.000	.000	.000	.000	.000
−.05	.001	.012	.011	.005	.012	.016	.017	.014	.004	.003
−.18	.002	.008	.013	.006	.009	.008	.008	.008	.006	.002
−.25	.009	.009	.010	.006	.009	.007	.005	.003	.002	.003
$p(x)$	.041	.093	.093	.082	.113	.103	.155	.155	.113	.052

Source: Adapted from R. Kosobud and J. N. Morgan, eds., *Consumer Behavior of Individual Families over Two and Three Years* (Ann Arbor: Institute for Social Research, The University of Michigan, 1964), Table 5–5.

Optimal prediction and projection

Why are we interested in $\mu(x) = \mathrm{E}(y|x)$? One reason is optimal prediction. Assume that we know the joint density $f(y, x)$ and that we obtain a random draw from the corresponding joint distribution. Only x is revealed and you must predict y. What is the best predictor $m(x)$?

Criterion is expected squared prediction error, so that the optimal predictor minimizes this

$$\mathrm{E}\left((y - m(x))^2\right) = \mathrm{E}\left(((y - m(x)) + (\mu(x) - m(x)))^2\right) =$$

$$= \mathrm{E}\left((y - \mu(x))^2\right) + 2\mathrm{E}\left((y - \mu(x))(\mu(x) - m(x))\right) +$$

$$+\mathrm{E}\left((\mu(x) - m(x))^2\right) \geq \mathrm{E}\left((y - \mu(x))^2\right)$$

The lower bound is achieved if $m(x) = \mu(x)$.

Conclusion: Optimal predictor is

$$m(x) = E(y|x) = \mu(x)$$

Table 1.2 Conditional frequency distributions of Y = savings rate for given values of X = income.

	X									
Y	0.5	1.5	2.5	3.5	4.5	5.5	6.7	8.8	12.5	17.5
.50	.024	.118	.075	.073	.044	.049	.052	.058	.124	.077
.40	.024	.022	.064	.086	.088	.068	.052	.058	.071	.135
.25	.049	.064	.043	.086	.088	.107	.129	.123	.115	.115
.15	.049	.097	.097	.146	.142	.194	.271	.348	.212	.384
.05	.244	.247	.355	.378	.363	.281	.303	.252	.372	.135
0	.317	.140	.000	.024	.009	.000	.000	.000	.000	.000
−.05	.024	.129	.118	.061	.106	.155	.109	.090	.035	.058
−.18	.049	.086	.140	.073	.080	.078	.052	.052	.053	.038
−.25	.220	.097	.108	.073	.080	.068	.032	.019	.018	.058
Total	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$m_{Y\|X}$	−.012	.065	.048	.099	.079	.083	.112	.129	.154	.161

Another interpretation is that $\mu(x)$ is the best fitting (minimizes average squared deviation) relation between y and x. If we consider random variables y and $m(x)$ then $\mu(x)$ is the closest to y if we take the average squared deviation as a measure of distance. For this reason we call $\mu(x)$ the projection of y on the set of random variables x.

We could restrict the predictors to be linear

$$m(x) = a + bx$$

The best linear predictor is the solution to

$$\min_{a,b} E\left((y - a - bx)^2\right)$$

The first order conditions are

$$-2E(u) = 0 \Rightarrow E(y) = a + bE(x)$$

$$-2E(ux) = 0 \Rightarrow E(xy) = a + bE(x^2)$$

with

$$u = y - a - bx$$

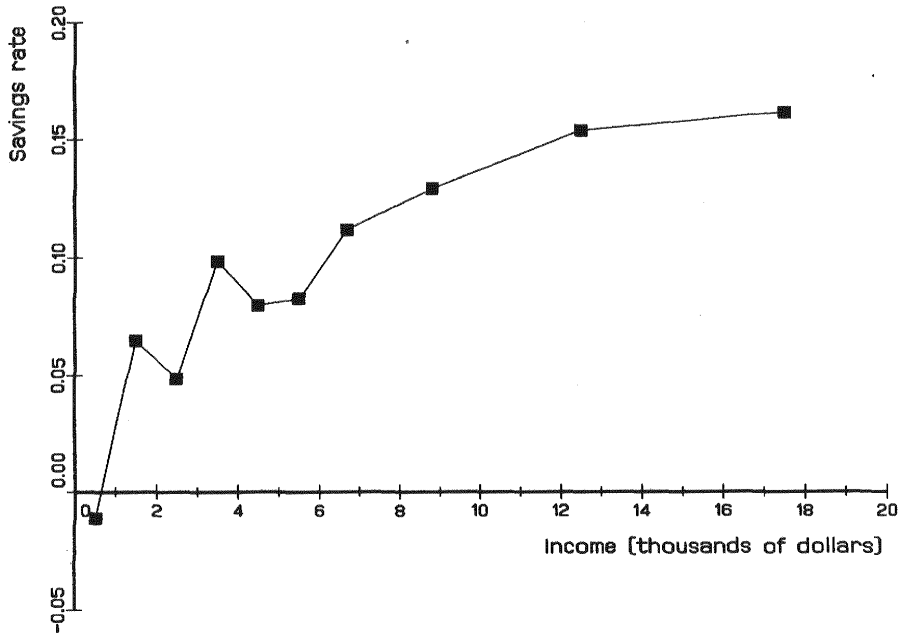


Figure 1.1 Conditional mean function: savings rate on income.

The solution is

$$b = \frac{\text{Cov}(x, y)}{\text{Var}(x)}$$

$$a = E(y) - bE(x)$$

If we estimate population by sample moments we have

$$\hat{b} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

This is the OLS solution!

In the example it is obvious how to estimate the conditional mean function with discrete or discretized data. How do we estimate this function if the variables are continuous?

Our application is to Regression Discontinuity (RD) in the next lecture. In RD we have in general a single independent variable so that x is a scalar.

The population conditional mean function is

$$\mu(x) = \mathrm{E}(y|x) = \frac{\int y f(y,x)\mathrm{d}y}{f(x)}$$

so what we need are estimators of $f(y,x)$ and $f(x)$, i.e. estimators of densities. Because we do not assume that these densities are in a parametric class, we call such estimators nonparametric.

Nonparametric estimation of densities

In this discussion I will use capital letters to distinguish random variables from their realization. An observation before we observe its value is X and the observed value is x.

Let $X_1, \ldots, X_n$ be a random sample from a continuous distribution with density f. We want to estimate $f(x)$ for x in the support of the distribution, i.e. for x wirg $f(x) > 0$.

It is easier to estimate the distribution function. We can use the sample $X_1, \ldots, X_n$ to obtain the empirical c.d.f.

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

This is a step function. It is the c.d.f. of the empirical distribution that assigns probability $1/n$ to $X_1, \ldots, X_n$. We can show that

$$\mathrm{E}[F_n(x)] = F(x) \qquad \mathrm{Var}(F_n(x)) = \frac{F(x)(1-F(x))}{n} \qquad F_n(x) \xrightarrow{p} F(x)$$

Histogram estimator of a density

A density can be visualized/estimated using a histogram. If the support of the distribution is $[0,1]$ (this to simplify the notation) define bins

$$B_j = \left[\frac{j-1}{m}, \frac{j}{m}\right) \quad j = 1, \ldots, m$$

with width $h = \frac{1}{m}$.

The fraction of observations in bin j is

$$\hat{p}_j = \frac{1}{n}\sum_{i=1}^{n} I(X_i \in B_j)$$

The histogram estimator of the density is

$$\hat{f}_n(x) = \sum_{j=1}^{m} I(x \in B_j) \frac{\hat{p}_j}{h}$$

Define

$$p_j = \int_{B_j} f(x) \mathrm{d}x$$

then if $x \in B_j$

$$\mathrm{E}[\hat{f}_n(x)] = \frac{p_j}{h} = \frac{\int_{B_j} f(u) \mathrm{d}u}{h} \approx \frac{f(x)h}{h} = f(x)$$

where the area under the curve $\int_{B_j} f(u) \mathrm{d}u$ is approximated by $f(x)h$ and the approximation is better when h is smaller. For h small the histogram estimator is approximately unbiased for the density in x.

Histograms with different choices for h (or equivalently m) are given in the figure. The density is that of the distribution of the distance to galaxies that are encountered on a straight line from the earth's surface. It is particularly interesting whether there are clusters represented by modes of the density. Note that the number of modes is larger if h is smaller and that the histogram estimator is very 'wiggly' for small values of h.

Mean integrated squared error (MISE)

To evaluate an estimator $\hat{\theta}$ of θ we use the MSE

$$\mathrm{E}[(\hat{\theta} - \theta)^2]$$

For an estimator of a density we need to consider the MSE at each point of the support, i.e. each point x with $f(x) > 0$. We define the integrated squared error (ISE) (remember that the support was $[0, 1]$)

$$L(\hat{f}_n, f) = \int_0^1 (\hat{f}_n(x) - f(x))^2 \mathrm{d}x$$

and the mean integrated squared error (MISE)

$$R(\hat{f}_n, f) = \mathrm{E}[L(\hat{f}_n, f)]$$

Note that $L(\hat{f}_n, f)$ is a random variable.

The MISE can be written as

$$R(\hat{f}_n, f) = \int_0^1 b(x)^2 \mathrm{d}x + \int_0^1 v(x) \mathrm{d}x$$

with

$$b(x) = \mathrm{E}[\hat{f}_n(x)] - f(x)$$

and

$$v(x) = \mathrm{E}\left[(\hat{f}_n(x) - \mathrm{E}[\hat{f}_n(x)])^2\right]$$

are the bias and variance of $\hat{f}_n(x)$ that estimates $f(x)$. The MISE is the sum of the integral of the squared bias and that of the variance.

As a general rule the bias increases with the bin width h while the variance decreases with h.

For the bias we can show that

$$\int_0^1 b(x)^2 \approx \frac{h^2}{12} \int_0^1 f'(x)^2 \mathrm{d}x$$

i.e. the integrated squared bias is proportional to h^2 and the bias goes to 0 of h goes to 0.

For the variance we have

$$\int_0^1 v(x)\mathrm{d}x \approx \frac{1}{nh}$$

Note that the integrated variance decreases with h, do that the variance is larger if h is smaller (see the earlier figures).
Combining the results

$$R(\hat{f}_n, f) \approx \frac{h^2}{12} \int_0^1 f'(x)^2 \mathrm{d}x + \frac{1}{nh}$$

This expression is minimized if

$$h^* = \frac{1}{n^{\frac{1}{3}}} \left(\frac{6}{\int_0^1 f'(x)^2 \mathrm{d}x} \right)^{\frac{1}{3}}$$

and the minimized MISE is

$$R(\hat{f}_n, f) \approx \frac{1}{n^{\frac{2}{3}}} \left(\frac{3}{4}\right)^{\frac{2}{3}} \left(\int_0^1 f'(x)^2 \mathrm{d}x \right)^{\frac{1}{3}}$$

The MSE of an estimator $\hat{\theta}$ of a parameter θ is proportional to $\frac{1}{n}$. Consider for instance the unbiased sample mean. The slower rate at which the MISE converges to 0 is due to the fact that we estimate a density function and not a parameter. This means that we need more observations to get the same precision (as measured by the MSE or MISE). For instance if $n = 100$ the MSE for θ is proportional to .01 and for the same precision for $f(x)$ we need $n = 464$

observations.

We cannot use h^* to obtain the optimal bin width. Instead we use the fact that for the ISE

$$L(\hat{f}_n, f) = \int_0^1 (\hat{f}_n(x) - f(x))^2 \mathrm{d}x = \int_0^1 \hat{f}_n(x)^2 \mathrm{d}x - 2\int_0^1 \hat{f}_n(x) f(x) \mathrm{d}x + \int_0^1 f(x)^2 \mathrm{d}x$$

where the last term does not depend on h to define

$$J(h) = \int_0^1 \hat{f}_n(x)^2 \mathrm{d}x - 2\int_0^1 \hat{f}_n(x) f(x) \mathrm{d}x$$

We estimate $J(h)$ with the cross-validation estimator

$$\hat{J}(h) = \int_0^1 \hat{f}_n(x)^2 \mathrm{d}x - \frac{2}{n}\sum_{i=1}^n \hat{f}_{n,-i}(X_i)$$

with $\hat{f}_{n,-i}(x)$ the histogram estimator of $f(x)$ that omits the i-th observation.

It can be shown

$$\hat{J}(h) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)h}\sum_{j=1}^m \hat{p}_j^2$$

In the figure the optimal choice is $m = 73$.

Kernel density estimator

Let K be the p.d.f. of a distribution with mean 0 and known variance. We call this density that does not depend on any parameters the *kernel*. Examples of kernels are the standard normal density (that has infinite support) and the Epanechnikov kernel with finite support

$$\begin{aligned} K(x) &= \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right) && |x| \leq \sqrt{5} \\ &= 0 && |x| > \sqrt{5} \end{aligned}$$

The kernel density estimator of f is

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^n \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

Because

$$\int \frac{1}{h} K\left(\frac{x - X_i}{h}\right) \mathrm{d}x = 1$$

the kernel density estimator puts a probability mass $1/n$ around X_i, not just in X_i as the empirical distribution does. How far the probability mass is spread

depends on h and h is called the *bandwidth* of the kernel density estimator. For the Epanechnikov kernel with bounded support the probability mass is spread over $[X_i - \sqrt{5}h, X_i + \sqrt{5}h]$. Therefore in x the density corresponding to X_i is

$$\frac{1}{nh} K\left(\frac{x - X_i}{h}\right)$$

and the kernel density estimator is the sum of these densities over the X_i.

Compare with the histogram estimator that spreads a probability mass $1/n$ in an interval of length h, so that in a particular bin the density corresponding to X_i is $\frac{1}{nh}$ if X_i is in the interval and is 0 if not. The histogram estimator is the sum of these densities over the X_i. Note that if we take the kernel in the density estimator as the density of the uniform distribution on $[-1/2, 1/2]$, then the density in x for an observation X_i is the same in the histogram and kernel density estimator. The bandwidth h in the kernel density estimator is like the bin width in the histogram estimator. The main difference is that in the kernel density estimator the 'bins' are overlapping, i.e. in x not only the densities of the X_i in the bin that contains x matter but potentially also densities of the X_i in neighboring 'bins'.

To obtain the MISE $R(\hat{f}_n, f)$ we compute the bias $b(x)$ and the variance $v(x)$. We find

$$\int_{-\infty}^{\infty} b(x)^2 \mathrm{d}x \approx \frac{1}{4}\sigma_K^4 \int_{-\infty}^{\infty} f''(x)\mathrm{d}x \cdot h^4$$

with σ_K^2 the variance of the kernel distribution. Compare this with the result for the histogram and we note that the bias for the kernel density estimator is of a smaller order.

For the variance

$$v(x) \approx \frac{1}{nh} \int_{-\infty}^{\infty} K(u)^2 \mathrm{d}u f(x)$$

Combining these results we have that the MISE is

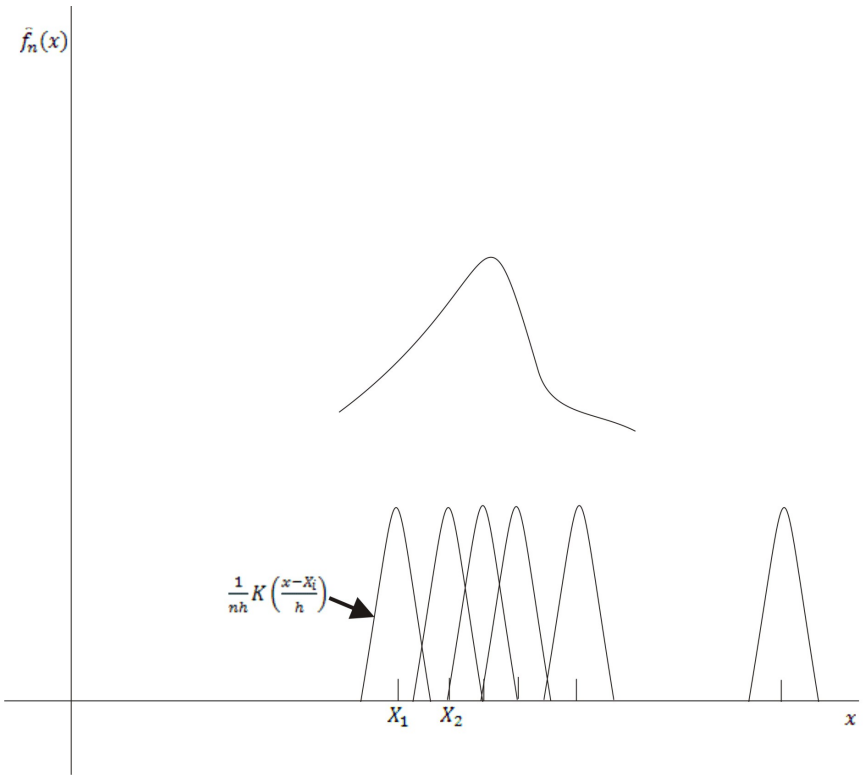
$$R(\hat{f}_n, f) = \frac{1}{4}\sigma_K^4 \int_{-\infty}^{\infty} f''(x)^2 \mathrm{d}x h^4 + \frac{1}{nh} \int_{-\infty}^{\infty} K(u)^2 \mathrm{d}u$$

The bias increases with h and the variance decreases with h. If we minimize over h we obtain the optimal bandwidth

$$h^* = \frac{(\sigma_K^2)^{-2/5} \left(\int_{-\infty}^{\infty} K(u)^2 \mathrm{d}u\right)^{1/5} \left(\int_{-\infty}^{\infty} f''(x)^2 \mathrm{d}x\right)^{-1/5}}{n^{1/5}}$$

We can take

$$\int_{-\infty}^{\infty} f''(x)^2 \mathrm{d}x$$



as a measure of smoothness of $f(x)$ and it is seen that unsmooth densities have a larger optimal bandwidth.

If we substitute h^* in the MISE and keep the term that converges to 0 at the slowest pace we have

$$R(\hat{f}_n, f) \approx Cn^{-4/5}$$

Note that this converges slower than the parametric rate $1/n$, but faster than the rate for the histogram estimator where the MISE goes to 0 as $n^{-2/3}$ which is slower. The required number of observations for .01 precision is now $n = 316$.

The optimal h derived above depends on unknown functions. We can again obtain h^* by cross-validation. Minimizing the MISE is equivalent to minimizing $\mathrm{E}(J(h))$ with

$$J(h) = \int_{-\infty}^{\infty} \hat{f}_n(x)^2 \mathrm{d}x - 2\int_{-\infty}^{\infty} \hat{f}_n(x) f(x) \mathrm{d}x$$

and we estimate $J(h)$ by the cross-validation estimator

$$\hat{J}(h) = \int_{-\infty}^{\infty} \hat{f}_n(x)^2 \mathrm{d}x - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{n,-i}(X_i)$$

Multivariate density estimation and the curse of dimensionality

We have a random sample $X_1, \ldots, X_n$ with X_i a d-dimensional random vector from a d-dimensional distribution with density f

The kernel density estimator of f is

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} \prod_{j=1}^{d} K\left(\frac{x_j - X_{ij}}{h}\right)$$

The MISE is proportional to $n^{-\frac{4}{4+d}}$ which converges slowly to 0 if d is large.

If the joint density is the product of standard normals and we consider the relative MSE in $x = 0$, i.e. the ratio of the MSE and $f(0)$, we can compute the sample size that makes the relative MSE less than .1. For $d = 1$ we find that $n = 4$ for $d = 5$, $n = 768$ and for $d = 10$, $n = 842000$. This is the curse of dimensionality.

High-dimensional nonparametric density estimation requires extremely large data sets.

Nonparametric kernel regression

The Nadaraya-Watson kernel estimator of $\mu(x) = \mathrm{E}(Y|X = x)$ is

$$\hat{\mu}_n(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)}$$

This estimator can also be written as

$$\hat{\mu}_n(x) = \sum_{i=1}^{n} w_i(x) Y_i$$

with

$$w_i(x) = \frac{K\left(\frac{x - X_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x - X_i}{h}\right)}$$

i.e. the estimator is a weighted average of the Y_i with weights that are larger if X_i is close to x.

Because

$$\mu(x) = \frac{\int_{-\infty}^{\infty} y f(y,x) \mathrm{d}y}{f(x)}$$

we can replace the densities in the numerator and denominator by kernel density estimators. This reproduces the Nadaraya-Watson estimator.
The MISE of the estimator is

$$R(\hat{\mu}_n, \mu) = \mathrm{E}\left[\int_{-\infty}^{\infty} (\hat{\mu}_n(x) - \mu(x))^2 \mathrm{d}x\right]$$

and we find that the integrated squared bias is proprtional to h^4 and the integrated variance is proportional to $1/nh$ (same as for kernel density estimator). The minimal MISE decreases as $n^{-4/5}$ which is the same as for the kernel density estimator.

The optimal bandwidth can again be estimated using cross-validation, i.e. we minimize

$$\hat{J}(h) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_{n,-i}(X_i))^2$$

with $\hat{\mu}_{n,-i}(x)$ the NW estimator with observation i left out.

Local linear regression

The kernel regression estimator is very simple for the case that the kernel K is the density of a uniform distribution, i.e.

$$\begin{aligned} K(u) &= \frac{1}{2} \quad -1 \leq u \leq 1 \\ &= \qquad\quad \text{otherwise} \end{aligned}$$

The weights are in that case

$$w_i(x) = \frac{I(x-h \leq X_i \leq x+h)}{\sum_{j=1}^{n} I(x-h \leq X_j \leq x+h)}$$

i.e. we take the average of the Y_i over the observations with $x-h \leq X_i \leq x+h$.

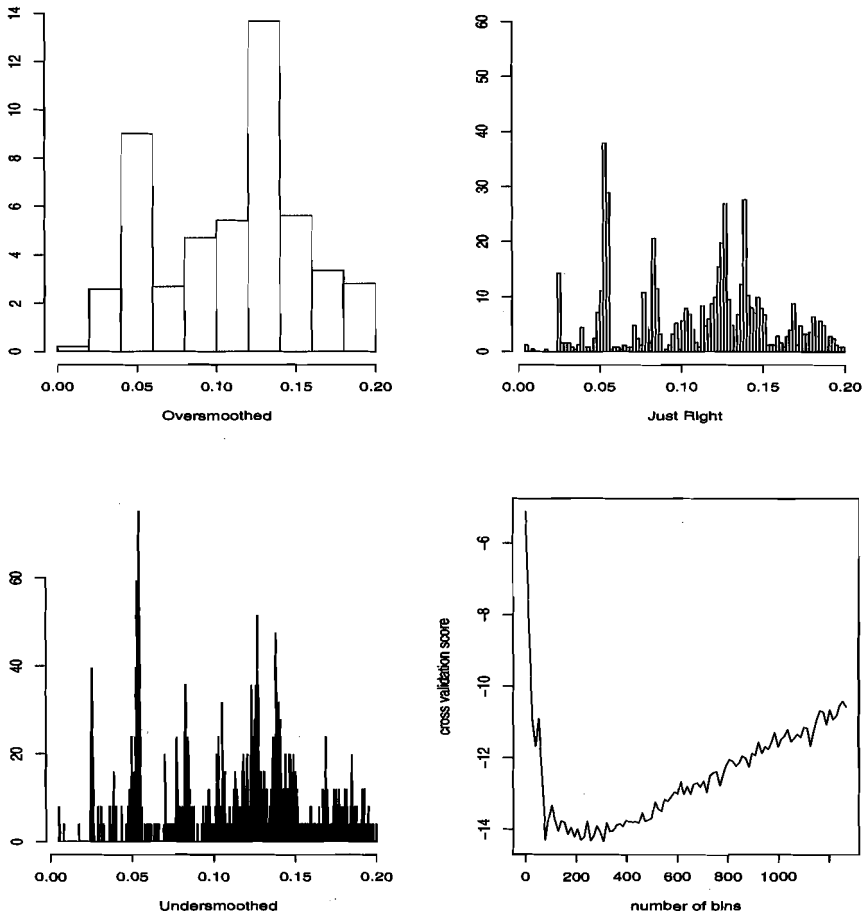
Instead of taking the average we could also use these observations to estimate a linear regression model, i.e. solve

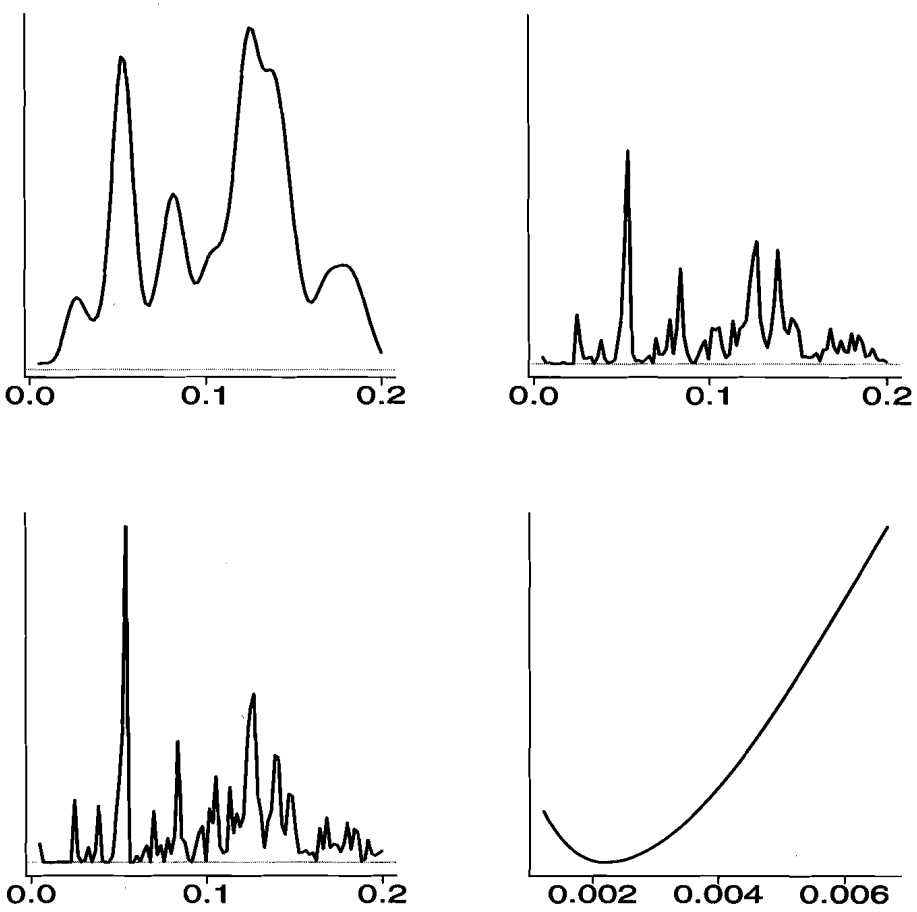
$$\min_{\alpha,\beta} \sum_{i=1}^{n} I(x-h \leq X_i \leq x+h)(Y_i - \alpha - \beta(X_i - x))^2$$

If $\beta = 0$ we obtain the estimator above, but if not, then

$$\hat{\mu}_n(x) = \hat{\alpha}$$

This is the *local linear regression* estimator of μ. Local linear estimators have a smaller bias than NW kernel estimators, a property that is important for the RD estimator.


Oversmoothed
Just Right
Undersmoothed
cross validation score
number of bins


0.0
0.1
0.2
0.0
0.1
0.2
0.0
0.1
0.2
0.002
0.004
0.006