

# 1 Non-Parametric Regression

Last Updated: April 28, 2008

## 1.1 Introduction

Consider the problem of estimating an unknown density for a random variable,  $X$ , which we denote by  $f(x)$ . One way of approaching this problem would be to assume that the density has a particular form. For example if we assumed that the data were Normally distributed then we would have that  $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$  and we could estimate the distribution by estimating the vector  $(\mu, \sigma^2)$ . In other words, our parametric assumption reduces the dimensionality of the problem from an infinite dimensional problem to a finite dimensional problem. However, assumptions always come with the cost that they may not be correct. Accordingly, it is important to devise methods that allow us to estimate  $f(x)$  without making any assumptions.

This problem can be extended to the estimation of a regression function  $E[y|x]$ . One way of estimating the regression model, which we have already discussed, would be to assume that the conditional expectation takes on a particular functional form. For example, if we make a linearity assumption we will have that  $E[y|x] = x\beta$ . More generally, we can assume that  $E[y|x] = f(x; \beta)$  where  $f(\cdot; \cdot)$  is a known function that depends on an unknown parameter  $\beta$ . Of course, our ability to understand the regression function is only as good as our assumptions and, thus, it will be useful to be able to estimate  $E[y|x]$  with minimal assumptions.

## 1.2 Non-Parametric Density Estimation

We begin by choosing a function  $K(\cdot)$  or a **Kernel** such that  $\int K(\psi) d\psi = 1$ . We are going to argue that the function

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K(\psi_i)$$

for  $\psi = \frac{x_i - x}{h}$  constitutes a “good” estimator of  $f(x)$ . Contrast this estimator with

$$\frac{1}{nh} \sum_{i=1}^n 1 \left( -\frac{1}{2} \leq \psi_i \leq \frac{1}{2} \right).$$

This estimator sorts through the data and places a weight of  $h^{-1}$  on all observations that lie in the interval  $(x - \frac{h}{2}, x + \frac{h}{2})$ . The idea of this estimator is that it places positive weight on all observations that are within  $h$  of the point  $x$ , but no weight on the other observations. So, the idea is to have  $K(\cdot)$  behave like the indicator function in the sense that it places greater emphasis on the observations that are closest to  $x$ . Two examples of  $K(\cdot)$  are (i)  $K(\psi) = (2\pi)^{-1/2} \exp(-\frac{1}{2}\psi^2)$  and (ii)  $K(\psi) = (2c)^{-1}$  for  $\psi \in [-c, c]$ . Both of these Kernels place greater weight on the observations  $x_i$  that are closest to  $x$ .

Before we derive some of the properties of  $\hat{f}(x)$ , we will need to make a few assumptions:

(A1) *iid* observations of  $x_i$ ; (A2)  $K(\cdot)$  is symmetric around 0 and satisfies (i)  $\int K(\psi) d\psi = 1$ , (ii)  $\int \psi^2 K(\psi) d\psi = \mu_2 \neq 0$ , (iii)  $\int K^2(\psi) < \infty$ ; (A3)  $f_{(2)}$  is bounded; (A4)  $h \rightarrow 0$  as  $n \rightarrow \infty$  and (A5)  $nh \rightarrow \infty$ . Assumption A1 just says that the sample is random and can easily be relaxed. Assumption A2 basically (aside from some technicalities) says that the Kernel needs to behave in the way that we described above. Assumption A3 is a technicality. Assumption A4 says that as the sample size grows we will place increasing weight only on the observations that are close

to  $x$ . Finally, assumption A5 tells us that  $h$  or the **bandwidth** will tend to zero at a rate that is slower than the rate at which the sample size grows.

We now make some useful calculations that will help us to determine the bias and variance of  $\widehat{f}(x)$ . First, we note that

$$E \left[ \widehat{f}(x) \right] = E [\omega_1]$$

where  $\omega_1 \equiv \frac{1}{h} K \left( \frac{x_1 - x}{h} \right)$ . Next, note that we will have that

$$\begin{aligned} E [\omega_1] &= h^{-1} E \left[ K \left( \frac{x_1 - x}{h} \right) \right] \\ &= h^{-1} \int K \left( \frac{x_1 - x}{h} \right) f(x_1) dx_1 \\ &= \int K(\psi) f(x + h\psi) d\psi. \end{aligned}$$

Thus, the bias of  $\widehat{f}(x)$  will be given by

$$bias \widehat{f}(x) = \int K(\psi) [f(x + h\psi) - f(x)] d\psi.$$

We also note that

$$V \left( \widehat{f}(x) \right) = n^{-1} V(\omega_1)$$

and that

$$\begin{aligned} E [\omega_1^2] &= h^{-2} E \left( K^2 \left( \frac{x_1 - x}{h} \right) \right) \\ &= h^{-2} \int K^2 \left( \frac{x_1 - x}{h} \right) f(x_1) dx_1 \\ &= h^{-1} \int K^2(\psi) f(x + h\psi) d\psi. \end{aligned}$$

Thus, we will have that

$$V\left(\widehat{f}(x)\right) = (nh)^{-1} \int K^2(\psi) f(x+h\psi) d\psi - n^{-1} \left[ \int K(\psi) f(x+h\psi) d\psi \right]^2.$$

To help us characterize the bias and variance of  $\widehat{f}(x)$  as a function of  $h$ , we make the following expansion

$$f(h\psi + x) = f(x) + h\psi f_{(1)}(x) + \frac{h^2}{2}\psi^2 f_{(2)}(x) + \dots$$

This gives us that

$$\begin{aligned} \text{bias}\widehat{f}(x) &= \int K(\psi) \left[ h\psi f_{(1)}(x) + \frac{h^2}{2}\psi^2 f_{(2)}(x) + \dots \right] d\psi \\ &= \int K(\psi) \left[ \frac{h^2}{2}\psi^2 f_{(2)}(x) \right] d\psi \\ &= \frac{h^2}{2}\mu_2 f_{(2)}(x) \end{aligned}$$

where  $\mu_2 = \int \psi^2 K(\psi) d\psi$ . Note that the first term on the second line goes away because the Kernel is symmetric. In addition, this approximation is valid up to  $O(h^2)$ . Similar calculations yields an approximation up to  $O((nh)^{-1})$  of

$$V\left(\widehat{f}(x)\right) = (nh)^{-1} f(x) \int K^2(\psi) d\psi.$$

These calculations show us that the bandwidth choice will affect the bias and the variance in opposite ways. A small bandwidth will result in a low bias and a high variance and a large bandwidth will do the opposite.

There is a literature on the optimal bandwidth choice. The basic idea is that you want to

choose  $h$  so as to achieve a balance between bias and variance. One way of doing this is to choose  $h$  that solves

$$\min_h \int \left[ \text{bias} \hat{f}(x)^2 + V(\hat{f}(x)) \right] dx.$$

This choice of  $h$  minimizes the Mean Integrated Square Error (MISE). Generally, solving this problem is tedious and not terribly interesting. A heuristic solution to the problem can be arrived at by noting that the bias will be  $O(h^2)$  and the variance will be  $O((nh)^{-1})$  then we will have that

$$MISE = \max \{ O(h^4), O((nh)^{-1}) \}$$

which suggests that  $h$  should be chosen to equate the bias and variance. Hence, we will have  $h_{opt} = cn^{-1/5}$  where  $c$  is a constant that depends on the unknown density. It is important to note that  $c$  still depends on the density that we are trying to estimate and, hence, the optimal bandwidth is still a guess. Some researchers have calculated  $h_{opt}$  for different densities. For example, if the underlying density is normal then  $h_{opt} = 1.06\sigma n^{-1/5}$ . In practice, these procedure of choosing the optimal bandwidths are not used with tremendous frequency among applied researchers. Instead, many applied researchers such as Angus Deaton and Chris Paxson use more intuitive procedures like “visual inspection.”

### 1.3 Non-Parametric Regression Estimation

How do we use what we know about density estimation to estimate a conditional expectation or regression function? Recall that the definition of the regression function is

$$E[y|x] = \int \frac{yf(y,x)}{f_1(x)} dy = m(x).$$

In the earlier units, we were making the parametric assumption on  $m(x)$  that it is a linear function. This would obtain if, for example, the joint distribution of  $(y, x)$  were normal. Now, we are going to allow  $m(x)$  to remain unrestricted.

To do this, we are simply going to replace the densities in the above conditional expectations with their Kernel estimates. This gives us

$$\hat{m}(x) = \int y \frac{(nh^2)^{-1} \sum K\left(\frac{y_i - y}{h}, \frac{x_i - x}{h}\right)}{(nh)^{-1} \sum K_1\left(\frac{x_i - x}{h}\right)} dy.$$

Next, we define  $v_i = \frac{y_i - y}{h}$  and note that the numerator can be written as

$$\begin{aligned} \int (nh^2)^{-1} \sum (y_i - hv) K\left(v, \frac{x_i - x}{h}\right) h dv &= n^{-1} \sum y_i \int K\left(v, \frac{x_i - x}{h}\right) h^{-1} dv \\ &\quad - n^{-1} \sum \int v K\left(v, \frac{x_i - x}{h}\right) dv \\ &= n^{-1} \sum y_i \int K\left(v, \frac{x_i - x}{h}\right) h^{-1} dv \\ &= (nh)^{-1} \sum y_i K_1\left(\frac{x_i - x}{h}\right). \end{aligned}$$

Note that the third equality followed because the Kernel is symmetric around zero. Thus, we

will have that

$$\hat{m}(x) = \frac{\sum y_i K_1\left(\frac{x_i-x}{h}\right)}{\sum K_1\left(\frac{x_i-x}{h}\right)} = \sum y_i \omega_i(x)$$

where  $\omega_i(x) \equiv \frac{K_1\left(\frac{x_i-x}{h}\right)}{\sum K_1\left(\frac{x_i-x}{h}\right)}$ .

This estimator which is also known as the **Nardaya-Watson Estimator** can be interpreted as a weighted least squares estimator. To see this, consider the objective function

$$\sum \omega_i(x) (y_i - m(x))^2.$$

This is a least squares problem where the solution is the Nardaya-Watson Estimator. In fact, the solution can be interpreted as a weighted average of  $y$ . We can extend this logic and, instead, solve the problem

$$\sum \omega_i(x) (y_i - m(x) - \beta(x)(x_i - x))^2.$$

The solution to this problem is the Local Linear Regression. Note that the Nardaya-Watson Estimator is a special case of this estimator.