20
Nonparametric Curve Estimation

In this Chapter we discuss nonparametric estimation of probability density functions and regression functions which we refer to as **curve estimation** or **smoothing**.

In Chapter 7 we saw that it is possible to consistently estimate a cumulative distribution function F without making any assumptions about F. If we want to estimate a probability density function $f(x)$ or a regression function $r(x) = \mathbb{E}(Y|X = x)$ the situation is different. We cannot estimate these functions consistently without making some smoothness assumptions. Correspondingly, we need to perform some sort of smoothing operation on the data.

An example of a density estimator is a **histogram**, which we discuss in detail in Section 20.2. To form a histogram estimator of a density f, we divide the real line to disjoint sets called **bins**. The histogram estimator is a piecewise constant function where the height of the function is proportional to number of observations in each bin; see Figure 20.3. The number of bins is an example of a **smoothing parameter**. If we smooth too much (large bins) we get a highly biased estimator while if we smooth too little (small bins) we get a highly variable estimator. Much of curve estimation is concerned with trying to optimally balance variance and bias.

L. Wasserman, *All of Statistics*

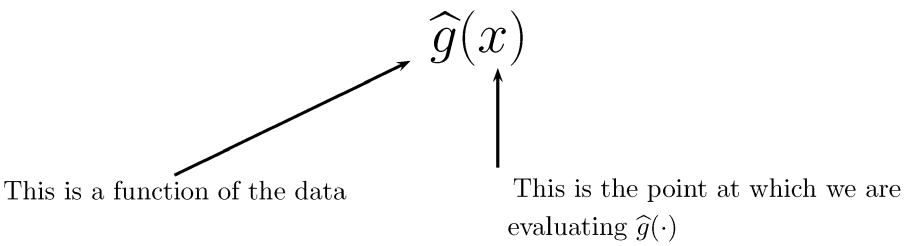

FIGURE 20.1. A curve estimate $\widehat{g}$ is random because it is a function of the data. The point x at which we evaluate $\widehat{g}$ is not a random variable.

20.1 The Bias-Variance Tradeoff

Let g denote an unknown function such as a density function or a regression function. Let $\widehat{g}_n$ denote an estimator of g. Bear in mind that $\widehat{g}_n(x)$ is a random function evaluated at a point x. The estimator is random because it depends on the data. See Figure 20.1.

As a loss function, we will use the **integrated squared error (ISE)**: [1]

$$L(g, \widehat{g}_n) = \int (g(u) - \widehat{g}_n(u))^2 \, du. \tag{20.1}$$

The **risk** or **mean integrated squared error (MISE)** with respect to squared error loss is

$$R(f, \widehat{f}) = \mathbb{E}\Big(L(g, \widehat{g}) \Big). \tag{20.2}$$

20.1 Lemma. *The risk can be written as*

$$R(g, \widehat{g}_n) = \int b^2(x) \, dx + \int v(x) \, dx \tag{20.3}$$

where

$$b(x) = \mathbb{E}(\widehat{g}_n(x)) - g(x) \tag{20.4}$$

is the bias of $\widehat{g}_n(x)$ at a fixed x and

$$v(x) = \mathbb{V}(\widehat{g}_n(x)) = \mathbb{E}\Big(\big(\widehat{g}_n(x) - \mathbb{E}\big(\widehat{g}_n(x)\big)^2\big) \Big) \tag{20.5}$$

is the variance of $\widehat{g}_n(x)$ at a fixed x.

[1] We could use other loss functions. The results are similar but the analysis is much more complicated.