Lecture 1

Plan

1. Histograms and kernel density estimation
2. Nonparametric regression: local polynomial and spline estimators

Data example
Kenya demographic and health survey 2003:
$n = 4555$ observations on Kenyan children aged from 0 to 60 months (no same children)

$$\text{Z-score}_i = \frac{\text{height}_i - \text{med}(\text{height}_{RP})}{\sqrt{\text{var}(\text{height}_{RP})}},$$

with height_i as the height of the i-th child at a given age and $\text{med}(\text{height}_{RF})$ $(\text{var}(\text{height}_{RP}))$ as the median (variance) of the height of healthy children of the same age in a reference population. Value Z-score < -2 indicates that the child is stunted.

0 Some notations

For two deterministic series $\{a_n\}$, $\{b_n\}$

1. $a_n = \mathcal{O}(b_n)$, if $\exists C > 0$, such that $\sup_n |a_n/b_n| \le C$.
2. $a_n = \mathcal{o}(b_n)$, if $a_n/b_n \to 0$, $n \to \infty$.

Note that from $a_n = \mathcal{o}(b_n)$ follows $a_n = O(b_n)$, but from $a_n = O(b_n)$ does not follow $b_n = O(a_n)$.

The indicator function will be denoted by $\mathbb{I}(\cdot)$:

$$\mathbb{I}(A) = \begin{cases} 1, & \text{if } A \text{ is true} \\ 0, & \text{else} \end{cases}$$

1 Histograms and kernel density estimation

1.1 Empirical cumulative distribution function

Let $X_1, \dots, X_n \overset{i.i.d.}{\sim} P$, where P has c.d.f. F. The most well-known and studied nonparametric estimator of F is given by

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

Let us fix some $x \in \mathbb{R}$. Since $\mathrm{E}\{\mathbb{I}(X_1 \le x)\} = P(X_1 \le x) = F(x)$, it follows that $\mathbb{I}(X_1 \le x), \dots, \mathbb{I}(X_n \le x) \overset{i.i.d.}{\sim} B(1, F(x))$. In particular, it follows immediately that

$$\mathrm{MSE}\{F_n(x)\} = \mathrm{var}\{F_n(x)\} = \frac{1}{n}F(x)\{1 - F(x)\}.$$

Moreover, one can apply the strong law of large numbers to conclude that

$$F_n(x) \xrightarrow{a.s.} F(x), \;\; x \in \mathbb{R}, \; n \to \infty.$$

1.2 Histogram

Let $X_1, \dots, X_n \overset{i.i.d.}{\sim} F$, where F is an unknown c.d.f. with a p.d.f. $F^{'} = f$.
First, choose a starting point x_0, a binwidth $h > 0$ and define bins (or classes) $I_j = [x_0 + jh - h, x_0 + jh)$, $j \in \mathbb{Z}$. W.l.o.g. we set $x_0 = 0$. Since $f(x) = F^{'}(x)$ *a.e.*, a simple estimator for $f(x)$ at $x \in I_j$ would be

$$f_n(x; h) = \frac{F_n(jh) - F_n(jh - h)}{h} = \frac{1}{nh}\sum_{i=1}^{n} \mathbb{I}(jh - h < X_i \le jh) = \frac{1}{nh}\sum_{i=1}^{n} \mathbb{I}(X_i \in I_j).$$

This estimator is called the **regular histogram**. That is, the histogram estimates the density $f(x)$ for all $x \in I_j$ by the same value: the number of observations X_i in the bin I_j, scaled by the total number of observations n and binwidth h. In particular, the area under the histogram is 1.

As in the case of empirical c.d.f., we first study how good $f_n(x; h)$ estimates $f(x)$ at a fixed $x \in I_j$. We assume that f is Lipschitz continuous, that is, there exists a constant $L > 0$ such that for any $x, y \in \mathbb{R}$ it holds that $|f(x) - f(y)| \le L|x - y|$. Moreover, assume that $f(x) \le f_{max} < \infty$, for all x. The bin width $h \to 0$, such that $nh \to \infty$.

First, consider the expectation of the histogram:

$$\begin{aligned} \mathrm{E}\{f_n(x; h)\} &= \frac{1}{nh}\sum_{i=1}^{n} \mathrm{E}\{\mathbb{I}(X_i \in I_j)\} = \frac{1}{h}P(X_1 \in I_j) = \frac{1}{h}\int_{jh-h}^{jh} f(u)du \\ &= \frac{F(jh) - F(jh - h)}{h} = f(x^*), \end{aligned}$$

where $x^* \in I_j$. The last equality is due to the mean value theorem.
Due to Lipschitz continuity $|f(x^*) - f(x)| \leq L|x^* - x|$ and hence

$$|\text{bias}\{f_n(x;h)\}| = |\text{E}\{f_n(x;h)\} - f(x)| = |f(x^*) - f(x)| \leq L|x^* - x| \leq Lh,$$

since $x^*, x \in I_j$ and the width of I_j is h.

Next, bound the variance

$$\begin{aligned} \text{var}\{f_n(x;h)\} &= \frac{1}{nh^2}\text{var}\left\{\mathbb{I}(X_1 \in I_j)\right\} = \frac{1}{nh^2}P(X_1 \in I_j)\{1 - P(X_1 \in I_j)\} \\ &= \frac{1}{nh}f(x^*)\{1 - hf(x^*)\} \leq \frac{f_{max}}{nh} - \frac{f_{max}^2}{n} \leq \frac{f_{max}}{nh}. \end{aligned}$$

Putting bias and variance bounds together, one obtains the bound on the mean squared error of the histogram $f_n(x;h)$ for all x

$$\text{MSE}\{f_n(x;h)\} = \text{bias}\{f_n(x;h)\}^2 + \text{var}\{f_n(x;h)\} \leq L^2h^2 + \frac{f_{max}}{nh}.$$

The binwidth h that minimizes the right-hand side of the last inequality is given by

$$h_{MSE} = \left(\frac{f_{\max}}{2L^2 n}\right)^{1/3}.$$

Plugging-in this value to $MSE\{f_n(x;h)\}$ leads to $MSE\{f_n(x;h_{MSE})\} = \mathcal{O}(n^{-2/3})$.
From these results we can conclude:

1. A histogram $f_n(x;h)$ is a (L_2) consistent point estimator for $f(x)$, $x \in I_j$, if h is of order $\mathcal{O}(n^{-1/3})$, since in this case $MSE\{f_n(x;h)\} = \mathcal{O}(n^{-2/3})$ and $f_n(x;h) \xrightarrow{L_2} f(x)$.

2. The order of $MSE\{f_n(x;h_{MSE})\} = \mathcal{O}(n^{-2/3})$ is larger (=one needs more data) than the parametric rate n^{-1}.

3. Large value of h corresponds to a small variance and large bias and vice versa: smaller values of h imply a small bias, but large variance. Such an effect is called *bias-variance trade-off* and h_{MSE} balances bias and variance of $f_n(x;h)$.

1.3 Kernel density estimators

Can we find another estimator of f, that would have a smaller MSE at each x and if yes, which assumptions are needed?

In a histogram one fixes classes I_j and finds the number of observations that fall into each class, that is, $f_n(x;h) = h^{-1}\{F_n(jh) - F_n(jh-h)\}$. Recall again that

$$f(x) = F'(x) \approx \frac{F(x+h) - F(x-h)}{2h}$$

for some sufficiently small $h > 0$ and consider another approximation

$$\begin{aligned}
\widehat{f}(x;h) &= \frac{F_n(x+h) - F_n(x-h)}{2h} = \frac{1}{2hn}\sum_{i=1}^{n}\{\mathbb{I}(X_i \le x+h) - \mathbb{I}(X_i \le x-h)\} \\
&= \frac{1}{2hn}\sum_{i=1}^{n}\mathbb{I}(x-h < X_i \le x+h) = \frac{1}{2hn}\sum_{i=1}^{n}\mathbb{I}\left(\frac{|X_i - x|}{h} \le 1\right) \\
&=: \frac{1}{nh}\sum_{i=1}^{n} K_u\left(\frac{X_i - x}{h}\right),
\end{aligned}$$

where $h > 0$ and $h \to 0$ and

$$K_u(x) = \begin{cases} 1/2, & |x| \le 1 \\ 0, & |x| > 1 \end{cases}$$

is the density of continuous uniform distribution on $[-1, 1]$.

Compared to a histogram, in $\widehat{f}(x;h)$ not the classes are fixed, but an interval of length $2h$ around x. Estimator $\widehat{f}(x;h)$ with K_u is known as the average shifted histogram or just the Rosenblatt estimator.

The Rosenblatt estimator, as well as a regular histogram, is a piecewise constant (=not smooth), which is a clear drawback. A simple way out is to replace K_u with an appropriate smooth function K. Such a more general estimator is known as the Parzen-Rosenblatt kernel density estimator or just kernel density estimator.

Definition 1.1. Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$ with a given density $F' = f$. A **kernel density estimator** for f is defined via

$$\widehat{f}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),\ x \in \mathbb{R},\ h > 0.$$

Thereby $K : \mathbb{R} \to \mathbb{R}$, such that $\int_{-\infty}^{\infty} K(x)dx = 1$ is known as **kernel** and $h > 0$ is called **bandwidth**. A **jth moment** of a kernel K is defined as $\mu_j = \int_{-\infty}^{\infty} x^j K(x)dx$.