Applied Nonparametric Methods

Franco Peracchi
University of Rome Tor Vergata and EIEF

Spring 2014
## Contents

1 Nonparametric Density Estimators \hspace{1cm} 2
  1.1 Empirical densities ........................................ 6
  1.2 The kernel method ......................................... 14
  1.3 Statistical properties of the kernel method .............. 22
  1.4 Other methods for univariate density estimation ........ 40
  1.5 Multivariate density estimators ........................... 44
  1.6 Stata commands ........................................... 50

2 Nonparametric Regression Estimators \hspace{1cm} 53
  2.1 Polynomial regressions ..................................... 56
  2.2 Regression splines ......................................... 58
  2.3 The kernel method ......................................... 62
  2.4 The nearest neighbor method .............................. 66
  2.5 Cubic smoothing splines .................................... 72
  2.6 Statistical properties of linear smoothers ............... 75
  2.7 Average derivatives ....................................... 83
  2.8 Methods for high-dimensional data ....................... 85
  2.9 Stata commands ........................................... 98

3 Distribution function and quantile function estimators \hspace{1cm} 102
  3.1 Distribution functions and quantile functions ............ 102
  3.2 The empirical distribution function ...................... 116
  3.3 The empirical quantile function .......................... 126
  3.4 Conditional distribution and quantile functions ........ 143
  3.5 Estimating the conditional quantile function ............ 146
  3.6 Estimating the conditional distribution function .......... 160
  3.7 Relationships between the two approaches ................. 168
  3.8 Stata commands ........................................... 171
1 Nonparametric Density Estimators

Let the data $Z_1, \ldots, Z_n$ be a sample from the distribution of a random vector $Z \in \mathcal{Z} \subseteq \mathbb{R}^m$. We are interested in the general problem of estimating the (probability) distribution of $Z$ nonparametrically, that is, without restricting it to belong to some known parametric family.

We begin with the problem of how to estimate nonparametrically the density function of $Z$.

Although there are several other characterizations of a distribution (e.g., the distribution function, the quantile function, the characteristic function, the hazard function), there may be advantages in focusing on the density:

- The graph of a density may be easier to interpret if one is interested in aspects such as symmetry or multimodality.

- Estimates of certain population parameters, such as the mode, are more easily obtained from an estimate of the density.
Use of nonparametric density estimates

Nonparametric density estimates may be used for:

- **Exploratory data analysis**: What is the “shape” of the distribution of $Z$?
- **Estimating qualitative features** of a distribution such as unimodality, skewness, etc.
- **Specification and testing** of parametric models (e.g., do data provide support for a Gaussian model?)
- **Estimation of statistical functionals**: Many statistical problems involve estimation of a population parameter that can be represented as a statistical functional $\theta = T(f)$ of the population density $f$. In these cases, the **analogy principle** suggests estimating $\theta$ by $\hat{\theta} = T(\hat{f})$, where $\hat{f}$ is a reasonable estimate of $f$, possibly a nonparametric one.
Example 1 Given an estimate \( \hat{f} \) of \( f \), a natural estimate of the mode \( \theta = \arg\max_{z \in \mathbb{Z}} f(z) \) of \( f \) is the mode \( \hat{\theta} = \arg\max_{z \in \mathbb{Z}} \hat{f}(z) \) of \( \hat{f} \).

Example 2 Let

\[
F(z) = \int_{-\infty}^{z} f(u) \, du
\]

be the value of the distribution function (df) of a random variable (rv) \( Z \) at \( z \). Given an estimate \( \hat{f} \) of \( f \), a natural estimate of \( F(z) \) is

\[
\hat{F}(z) = \int_{-\infty}^{z} \hat{f}(u) \, du.
\]

Example 3 A characterization of the distribution of a continuous non-negative rv \( Z \) (e.g. the length of an unemployment spell) with df \( F \) and density function \( f \) is the hazard function

\[
\lambda(z) = \lim_{\delta \to 0^+} \frac{\Pr\{z < Z \leq z + \delta \mid Z > z\}}{\delta} = \lim_{\delta \to 0^+} \frac{F(z + \delta) - F(z)}{\delta[1 - F(z)]} = \frac{1}{1 - F(z)} \left[ \lim_{\delta \to 0^+} \frac{F(z + \delta) - F(z)}{\delta} \right] = \frac{f(z)}{1 - F(z)}, \quad z \geq 0.
\]

If \( \hat{f} \) is a nonparametric estimate of \( f \) and \( \hat{F} = \int_{-\infty}^{z} \hat{f}(u) \, du \) is the derived nonparametric estimate of \( F \), then a nonparametric estimate of \( \lambda(z) \) is

\[
\hat{\lambda}(z) = \frac{\hat{f}(z)}{1 - \hat{F}(z)}.
\]

The nonparametric estimate \( \hat{\lambda}(z) \) may then be compared with the benchmark parametric estimate based on the exponential distribution with constant hazard \( \lambda(z) = \lambda > 0 \).

\[\square\]
Example 4 Let $Z = (X, Y)$ be a random vector with two elements ($m = 2$) and consider the **conditional mean function** (CMF) or **mean regression function** of $Y$ given $X$, defined by

$$
\mu(x) = E(Y \mid X = x) = \int y \frac{f(x, y)}{f_X(x)} \, dy,
$$

where $f(x, y)$ is the joint density of $X$ and $Y$ and

$$
f_X(x) = \int f(x, y) \, dy
$$

is the marginal density of $X$. Given a nonparametric estimate $\hat{f}(x, y)$ of $f(x, y)$, a nonparametric estimate of $\mu(x)$ is

$$
\hat{\mu}(x) = \int y \frac{\hat{f}(x, y)}{\hat{f}_X(x)} \, dy,
$$

(1)

where $\hat{f}_X(x) = \int \hat{f}(x, y) \, dy$ is a nonparametric estimate of $f_X$. \hfill \Box
1.1 Empirical densities

We begin with the simpler problem of estimating univariate densities ($m = 1$), and discuss the problem of estimating multivariate densities ($m > 1$) later in Section 1.5.

Thus, let $Z_1, \ldots, Z_n$ be a sample from the distribution of a univariate rv $Z$ with df $F$ and density function $f$. The case when the data are a finite segment of a strictly stationary process is discussed in Pagan & Ullah (1999).

The choice of an appropriate method for estimating $f$ depends on the nature, discrete or continuous, of the rv $Z$. 
Discrete $Z$

If $Z$ is discrete with a distribution that assigns positive probability mass to the values $z_1, z_2, \ldots$, then the probability $f_j = \Pr\{Z = z_j\}$ may be estimated by the relative frequency

$$\hat{f}_j = n^{-1} \sum_{i=1}^{n} 1\{Z_i = z_j\}, \quad j = 1, 2, \ldots,$$

called the empirical probability that $Z = z_j$.

This is just the sample average of $n$ iid binary rv’s $X_{1j}, \ldots, X_{nj}$, where

$$X_{ij} = 1\{Z_i = z_j\}, \quad i = 1, \ldots, n,$$

has a Bernoulli distribution with mean $f_j$ and variance $f_j(1 - f_j)$. Thus, $\hat{f}_j$ is unbiased for $f_j$, that is, $\E \hat{f}_j = f_j$, and its sampling variance is

$$\text{Var} \hat{f}_j = \frac{f_j(1 - f_j)}{n}, \quad j = 1, 2, \ldots,$$

which is typically estimated by $\hat{\text{Var}} \hat{f}_j = n^{-1} \hat{f}_j(1 - \hat{f}_j)$.

It is easy to verify that $\hat{f}_j$ is consistent for $f_j$, that is, $\hat{f}_j \xrightarrow{p} f_j$. It is also $\sqrt{n}$-consistent and asymptotically normal, that is,

$$\sqrt{n} (\hat{f}_j - f_j) \Rightarrow \mathcal{N}(0, \text{AV}(\hat{f}_j)),$$

where $\text{AV}(\hat{f}_j) = f_j(1 - f_j)$ can be estimated consistently by $\hat{\text{AV}}(\hat{f}_j) = \hat{f}_j(1 - \hat{f}_j)$.

This justifies the use of a symmetric asymptotic confidence interval (CI) for $f_j$ of the form

$$\text{CI}_{1-2\alpha}(f_j) = \hat{f}_j \pm z(\alpha) \sqrt{\frac{\hat{f}_j(1 - \hat{f}_j)}{n}},$$

where $z(\alpha)$ is the upper $\alpha$th percentile of the $\mathcal{N}(0, 1)$ distribution (e.g. $z(.025) = 1.96$). Notice that this CI is problematic because it may contain values less than zero or greater than one.
Continuous $Z$

The previous approach breaks down when $Z$ is continuous because the relative frequency of a specific value $z$ in any sample is either zero or very small.

Notice however that, if $z$ is any value in a sufficiently small interval $(a, b]$ then, by the definition of probability density,

$$f(z) \approx \frac{\Pr\{a < Z \leq b\}}{b - a}.$$  

Thus, a natural estimate of $f(z)$ is the fraction of observations falling in the (small) interval $(a, b]$ divided by the length $b - a$ of such interval,

$$\hat{f}(z) = \frac{1}{n(b - a)} \sum_{i=1}^{n} 1\{a < Z_i \leq b\}, \quad a < z \leq b, \quad (2)$$

called the empirical density of $Z_i$ at $z$.

Notice that, if no sample value is repeated and the interval $(a, b]$ is small enough, then it contains at most one observation, so $\hat{f}(z)$ is equal either to zero or to $[n(b - a)]^{-1}$.  

8
Sampling properties

The numerator of (2) is the sum of $n$ iid rv’s $X_i = 1\{a < Z_i \leq b\}$, each having a Bernoulli distribution with mean

$$E X_i = \Pr\{a < Z \leq b\} = F(b) - F(a)$$

and variance

$$\text{Var} X_i = [F(b) - F(a)][1 - F(b) + F(a)].$$

Thus,

$$E \hat{f}(z) = \frac{F(b) - F(a)}{b - a},$$

so $\hat{f}(z)$ is generally biased for $f(z)$, although its bias is negligible if $b - a$ is sufficiently small. Notice that $\hat{f}(z)$ is unbiased for $f(z)$ if the df $F$ is linear on an interval containing $(a, b]$ or, equivalently, the density $f$ is constant on such interval.

Further

$$\text{Var} \hat{f}(z) = \frac{[F(b) - F(a)][1 - F(b) + F(a)]}{n(b - a)^2}$$

$$= \frac{1}{n(b - a)} \frac{F(b) - F(a)}{b - a} - \frac{1}{n} \left[ \frac{F(b) - F(a)}{b - a} \right]^2.$$

If $b - a$ is sufficiently small, then

$$\text{Var} \hat{f}(z) \approx \frac{1}{n(b - a)} f(z) - \frac{1}{n} f(z)^2,$$

which shows that letting $b - a \to 0$ reduces the bias of $\hat{f}(z)$, but also increases its sampling variance. Thus, letting $b - a \to 0$ as $n \to \infty$ is not enough for $\hat{f}(z)$ to be consistent for $f(z)$. This trade-off between bias (systematic error) and sampling variance (random error) is typical of nonparametric estimation problems with $n$ fixed.

For $\hat{f}(z) \overset{p}{\to} f(z)$ as $n \to \infty$, we need that both $b - a \to 0$ and $n(b - a) \to \infty$, that is, the rate at which $b - a \to 0$ must be slower than the rate at which $n \to \infty$. 
Histograms

Histograms are classical examples of empirical densities.

Construction of a histogram requires:

- The choice of an interval \((a_0, b_0]\) that contains the range of the data.
- A partition of \((a_0, b_0]\) into \(J\) bins \((c_{j-1}, c_j]\). The simplest case is when the bins are of constant width \(h = (b_0 - a_0)/J\), so

\[
c_j = c_{j-1} + h = a_0 + jh, \quad j = 1, \ldots, J - 1,
\]

with \(c_0 = a_0\) and \(c_J = b_0\).

The \(i\)th sample point \(Z_i\) falls in the \(j\)th bin if \(c_{j-1} < Z_i \leq c_j\). For any point \(z\) in the \(j\)th bin, a histogram estimate of \(f(z)\) is

\[
\hat{f}(z) = \frac{1}{nh_j} \sum_{i=1}^{n} 1\{c_{j-1} < Z_i \leq c_j\},
\]

the fraction of observations falling in the \(j\)th bin divided by the bin width \(h_j = c_j - c_{j-1}\). This is also the estimate of \(f(z')\) for any other \(z'\) falling in the same bin. It is easy to verify that the function \(\hat{f}\) is a proper density, that is, \(\hat{f}(z) \geq 0\) for all \(z\) and \(\int \hat{f}(z) \, dz = 1\).

In the case of equally spaced bins, where \(h_j = h\) for all bins, the histogram estimate becomes

\[
\tilde{f}(z) = \frac{1}{nh} \sum_{i=1}^{n} 1\{c_{j-1} < Z_i \leq c_j\}.
\]
Figure 1: Histogram estimates of density for different number of bins. The data are 200 observations from the Gaussian mixture \(0.6 \mathcal{N}(0, 1) + 0.4 \mathcal{N}(4, 4)\).
Sampling properties of the histogram estimator

Let \( Z_1, \ldots, Z_n \) be a sample from a continuous distribution on the interval \((0,1]\), with df \( F \) and density \( f \). Put \( a_0 = 0, b_0 = 1 \), and partition the interval \((0,1]\) into \( J \) bins of constant width \( h = 1/J \). If \( X_j \) denotes the number of observations falling in the \( j \)th bin, then the histogram estimator at any point \( z \) in the \( j \)th bin is \( \hat{f}(z) = X_j/(nh) \).

The random \( J \)-vector \((X_1, \ldots, X_J)\) has a multinomial distribution with index \( n \) and parameter \( \pi = (\pi_1, \ldots, \pi_J) \), where

\[
\pi_j = F(c_j) - F(c_{j-1}) = F(c_j) - F(c_j - h), \quad j = 1, \ldots, J.
\]

Hence

\[
\Pr\{X_1 = x_1, \ldots, X_J = x_J\} = \frac{n!}{x_1! \cdots x_J!} \prod_{j=1}^{J} \pi_j^{x_j},
\]

where \( x_j = 0, 1, \ldots, n \). By the properties of the multinomial distribution, \( X_j \) has mean \( n\pi_j \) and variance \( n\pi_j(1 - \pi_j) \). Hence

\[
E \hat{f}(z) = \pi_j h, \quad \text{Var} \hat{f}(z) = \frac{\pi_j(1 - \pi_j)}{nh^2}.
\]

Thus, the histogram estimator is biased for \( f(z) \) in general, and its bias is just the error made in approximating \( f(z) \) by \( \pi_j/h \).

Now let the number of bins \( J = J_n \) increase with the sample size \( n \) or, equivalently, let the bin width \( h_n = 1/J_n \) decrease with \( n \). Let \( \{(c_{jn-1}, c_{jn})\} \) denote the sequence of bins that contain the point \( z \) and let \( \pi_{jn} = F(c_{jn}) - F(c_{jn-1}) \). Then

\[
E \hat{f}(z) = \frac{\pi_{jn}}{h_n} = \frac{F(z) - F(z - h_n)}{h_n} \to f(z),
\]

provided that \( h_n \to 0 \) with \( n \to \infty \). Further

\[
\text{Var} \hat{f}(z) = \frac{1}{n h_n} \frac{\pi_{jn}}{h_n} - \frac{1}{n} \left( \frac{\pi_{jn}}{h_n} \right)^2.
\]

Thus, for \( \hat{f}(z) \overset{p}{\to} f(z) \), we need not only that \( h_n \to 0 \) as \( n \to \infty \), but also that \( nh_n \to \infty \) or, equivalently, that \( J_n/n \to 0 \) as \( n \to \infty \). That is, \( h_n \) must go to zero, or equivalently \( J \) must increase with \( n \), but not too fast.
Drawbacks of histograms

Histograms are useful tools for exploratory data analysis, but have several undesirable features, such as:

- Results depend on the partition of \((a_0, b_0]\) into bins, that is, on their number and position.
- They also depend on the choice of the range \((a_0, b_0]\).
- The histogram is a step function with jumps at the end of each bin. Thus, it is impossible to incorporate prior information on the degree of smoothness of a density.
- The method may also create difficulties when estimates of the derivatives of the density are needed as input to other statistical procedures.

In the case of equally spaced bins:

- The partition of \((a_0, b_0]\) depends only on the number \(J\) of bins or, equivalently, on the bin width \(h = (b_0 - a_0)/J\).
- Given the data, increasing \(J\) (reducing \(h\)) eventually produces a histogram that is only informative about the location of the distinct sample points.
- By contrast, reducing \(J\) (increasing \(h\)) eventually produces a completely uninformative rectangle. It is intuitively clear, however, that \(J\) may safely be increased if the sample size \(n\) also increases.
- The fact that the bin width \(h\) is kept fixed over the range of the data may lead to loss of detail where the data are most concentrated. If \(h\) is reduced to deal with this problem, then spurious noise may appear in regions where data are sparse.
1.2 The kernel method

We now present a related method that tries to overcome some of the drawbacks of histograms.

Consider the empirical density (2). Putting \( a = z - h \) and \( b = z + h \), where \( h \) is a small positive number, gives

\[
\tilde{f}(z) = \frac{1}{2nh} \sum_{i=1}^{n} 1\{z - h < Z_i \leq z + h\}.
\]

This is sometimes called a naive density estimate or, for reasons that will be clear soon, a uniform kernel density estimate.

Notice that \( \tilde{f}(z) \) is just the fraction of sample points falling in the interval \( (z - h, z + h] \) divided by the length \( 2h \) of this interval. Equivalently, \( \tilde{f}(z) \) is the sample average of a binary indicator which is equal to \((2h)^{-1}\) if \( Z_i \) is within \( h \) distance from \( z \) and is equal to zero otherwise.

Because \( z - h < Z_i \leq z + h \) if and only if \( Z_i - h \leq z < Z_i + h \), the estimate \( \tilde{f}(z) \) may also be constructed by the following two-step procedure:

1. Place a “box” of width equal to \( 2h \) and height equal to \((2nh)^{-1}\) around each sample observation.
2. Add up the heights of the “boxes” that contain the point \( z \); the result is equal to \( \tilde{f}(z) \).

If we constructed a histogram of constant bin width \( 2h \) having \( z \) at the center of one of its bins, then \( \tilde{f}(z) \) would coincide with the histogram estimate. An advantage over the histogram method is that there is no need to choose an interval \((a_0, b_0)\) and to partition this interval into bins. However, \( \tilde{f}(z) \) still has the following drawbacks:

- it is a step function with jumps at the points \( z = Z_i \pm h \), so it is not smooth;
- it depends on the choice of constant \( h \).
Figure 2: Uniform kernel density estimates for different bandwidth. The data are 200 observations from the Gaussian mixture $0.6 \cdot \mathcal{N}(0,1) + 0.4 \cdot \mathcal{N}(4,4)$. 

![Uniform kernel density estimates](image)
Smooth kernel density estimates

Because the event \( z-h < Z_i \leq z+h \) is equivalent to the event \( -h < Z_i - z \leq h \), which in turn is equivalent to the event \( -1 \leq (z - Z_i)/h < 1 \), the density estimate \( \hat{f}(z) \) may also be written

\[
\hat{f}(z) = \frac{1}{nh} \sum_{i=1}^{n} w \left( \frac{z - Z_i}{h} \right),
\]

where

\[
w(u) = \begin{cases} 
1/2, & \text{if } -1 \leq u < 1, \\
0, & \text{otherwise},
\end{cases}
\]

is a symmetric bounded non-negative function that integrates to one and corresponds to the density of a uniform distribution on the interval \([-1, 1]\). This estimate may be viewed as the average of \( n \) uniform densities with common variance equal to \( h^2 \), each centered about one of the sample observations.

It is now clear that \( \hat{f} \) is not smooth because it is a (weighted) sum of step functions. Since the sum of smooth functions is itself smooth, replacing \( w \) by a smooth function \( K \) gives an estimate of \( f \) which inherits all the continuity and differentiability properties of \( K \).

This leads to the class of estimates of \( f(z) \) of the form

\[
\hat{f}(z) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{z - Z_i}{h} \right),
\]

where \( K \) is a bounded function called the kernel function and \( h \) is a positive constant called the bandwidth. Notice that \( \hat{f}(z) \) is simply an average of transformed observations, the type of transformation depending on the kernel \( K \) and the bandwidth \( h \). If \( K \) is continuous, then so is \( \hat{f}(z) \), and if \( K \) is differentiable up to order \( r \), then so is \( \hat{f}(z) \).

An estimate in this class is called a Rosenblatt-Parzen density estimate (Rosenblatt 1955, Parzen 1962) or, simply, a kernel density estimate.
Remarks

- Notice that
  \[ \int_{-\infty}^{\infty} \hat{f}(z) \, dz = \frac{1}{nh} \sum_{i=1}^{n} \int_{-\infty}^{\infty} K \left( \frac{z - Z_i}{h} \right) \, dz. \]

  After a change of variables from \((z - Z_i)/h\) to \(u\), we have
  \[ \int_{-\infty}^{\infty} K \left( \frac{z - Z_i}{h} \right) \, dz = h \int_{-\infty}^{\infty} K(u) \, du. \]

  Thus, if \(K\) integrates to one, then so does \(\hat{f}\). If \(K\) is a proper density (i.e., is nonnegative and integrates to one), then so is \(\hat{f}\).

- The bandwidth \(h\) controls the degree of smoothness or regularity of a density estimate. Small values of \(h\) tend to produce estimates that are irregular, while large values of \(h\) correspond to very smooth estimates.

- If \(K\) corresponds to the density of a zero-mean rv \(U\), then \(h^{-1}K((v - Z_i)/h)\) is the density of the rv \(V_i = Z_i + hU\), generated from \(U\) through a location-scale transformation. Thus, \(\hat{f}(z)\) may be viewed as the average height at the point \(z\) of \(n\) densities with the same spread and the same shape, each centered around one of the sample observations (Figure 3). The smaller is \(h\), the more concentrated is each density, and therefore the smaller is the number of observations that contribute in an appreciable way to form \(\hat{f}(z)\) and the more irregular is the resulting estimate of \(f(z)\).

- The fact that \(h\) is independent of the point where the density is estimated is a nuisance, for it tends to produce spurious effects in regions where data are sparse. If a large enough \(h\) is chosen to eliminate this phenomenon, we may end up losing important detail in regions where the data are more concentrated.
Figure 3: Structure of a smooth kernel density estimate.
Example 5 If $K = \phi$, where $\phi$ denotes the density of the $\mathcal{N}(0, 1)$ distribution, then the kernel estimate of $f$ is continuous and has continuous derivates of every order.

Such an estimate may be viewed as the average of $n$ Gaussian densities with common variance equal to the squared bandwidth $h^2$, each centered about one of the observations.

Unlike the uniform kernel (3), which takes constant positive value in the interval $[Z_i - h, Z_i + h)$ and vanishes outside this interval, the Gaussian kernel is always positive, assumes its maximum when $z = Z_i$ and tends to zero as $|z - Z_i| \to \infty$.

Hence, while the uniform kernel estimate of $f(z)$ is based only on the observations that are within $h$ distance from the evaluation point $z$ and assigns them a constant weight, the Gaussian kernel estimate is based on all the observations but assigns them a weight that declines exponentially as their distance from the evaluation point increases.

$\square$
Figure 4: Gaussian kernel density estimates for different bandwidth. The data are 200 observations from the Gaussian mixture \(0.6 \cdot \mathcal{N}(0, 1) + 0.4 \cdot \mathcal{N}(4, 4)\).
Extensions

Let $Z_1, \ldots, Z_n$ be a sample from the distribution of $Z$. If $\hat{f}$ is a kernel estimate of the density $f$ of $Z$, then one may easily estimate other aspects of the distribution of $Z$.

**Example 6** A nonparametric estimate of its df is

$$\hat{F}(z) = \int_{-\infty}^{z} \hat{f}(u) \, du = n^{-1} \sum_{i=1}^{n} K \left( \frac{z - Z_i}{h} \right),$$

where

$$K(u) = \int_{-\infty}^{u} K(v) \, dv$$

is the integrated kernel. If $K$ corresponds to a proper density, then $K$ is the associated df, and so $\hat{F}$ is itself a proper df.

**Example 7** If $Z$ is a continuous non-negative rv $Z$, then a nonparametric estimate of its hazard function is

$$\hat{\lambda}(z) = \frac{\hat{f}(z)}{1 - \hat{F}(z)} = \frac{\sum_{i=1}^{n} K \left( \frac{z - Z_i}{h} \right)}{N - \sum_{i=1}^{n} K \left( \frac{z - Z_i}{h} \right)}.$$
1.3 Statistical properties of the kernel method

In evaluating the statistical accuracy of a kernel density estimator \( \hat{f} \), it is important to distinguish between its \textbf{local} and \textbf{global} properties:

- the first refer to the accuracy of \( \hat{f}(z) \) as an estimator of \( f(z) \), the value of \( f \) \textbf{at a given point} \( z \);

- the second refer to the degree of \textbf{statistical “closeness”} between the two \textbf{functions} \( \hat{f} \) and \( f \).

To stress its dependence on the bandwidth \( h \), a kernel density estimator will henceforth be denoted by \( \hat{f}_h \).
Local properties

The kernel density estimator of $f$ at a point $z$ may be written as the sample average

$$
\hat{f}_h(z) = n^{-1} \sum_{i=1}^{n} K_i(z),
$$

where

$$
K_i(z) = \frac{1}{h} K\left(\frac{z - Z_i}{h}\right)
$$

is a nonlinear transformation of $Z_i$. A natural measure of local accuracy of $\hat{f}_h(z)$ is therefore its mean squared error (MSE)

$$
\text{MSE}[\hat{f}_h(z)] = \mathbb{E}[\hat{f}_h(z) - f(z)]^2 = \text{Var} \hat{f}_h(z) + \text{Bias} \hat{f}_h(z))^2.
$$

If the data $Z_1, \ldots, Z_n$ are a sample from the distribution of a rv $Z$, then we have

$$
\text{Bias} \hat{f}_h(z) = \frac{1}{h} \mathbb{E} K\left(\frac{z - Z}{h}\right) - f(z),
$$

$$
\text{Var} \hat{f}_h(z) = \frac{1}{nh^2} \text{Var} K\left(\frac{z - Z}{h}\right). \tag{4}
$$

Thus, the estimator $\hat{f}_h(z)$ is biased for $f(z)$ in general. For $h$ fixed, its bias does not depend on the sample size $n$, whereas its variance tends to zero as $n$ increases.

By imposing the following additional assumptions on the density $f$ and the kernel $K$ we can study in more detail the sampling properties of $\hat{f}_h(z)$.

Assumption 1

(i) The density $f$ is twice continuously differentiable.

(ii) The kernel $K$ satisfies

$$
\int K(u) \, du = 1, \quad \int uK(u) \, du = 0, \quad 0 < \int u^2 K(u) \, du < \infty.
$$

If $K$ is a nonnegative function, then Assumption 1 requires $K$ to be the density of some probability distribution with zero mean and finite positive variance, such as the uniform and the Gaussian.
Bias

After a change of variable from $x$ to $u = (z - x)/h$, we have

$$
\text{Bias} \hat{f}_h(z) = \frac{1}{h} \int K \left( \frac{z - x}{h} \right) f(x) \, dx - f(z)
= \int K(u) [f(z - hu) - f(z)] \, du,
$$

Because $f$ is twice differentiable (Assumption 1), a 2nd order Taylor expansion of $f(z - hu)$ about $h = 0$ gives

$$
f(z - hu) - f(z) = -hu f'(z) + \frac{1}{2} h^2 u^2 f''(z) + O(h^3).
$$

If $h$ is sufficiently small, then

$$
\text{Bias} \hat{f}_h(z) \approx -hf'(z) \int uK(u) \, du + \frac{1}{2} \left( \int u^2 K(u) \, du \right) h^2 f''(z)
= \frac{1}{2} m_2^2 h^2 f''(z),
$$

where $m_2^2 = \int u^2 K(u) \, du$ and we used the fact that $K$ has mean zero.

Thus, for $h$ sufficiently small, the bias of $\hat{f}_h(z)$ is $O(h^2)$ and depends on:

- the degree of curvature $f''(z)$ of the density at the point $z$,
- the amount of smoothing of the data through the bandwidth $h$ and the spread $m_2$ of the kernel.

In particular:

- Bias $\hat{f}_h(z) \approx 0$ when $f''(z) = 0$, that is, when the density is linear in a neighborhood of $z$.
- Bias $\hat{f}_h(z) \to 0$ as $h \to 0$.
- If the bandwidth $h$ decreases with the sample size $n$, then the bias of $\hat{f}_h(z)$ vanishes as $n \to \infty$. 

24
Variance

First notice that

\[
\text{Var } \hat{f}_h(z) = \frac{1}{nh^2} \text{Var } K \left( \frac{z - Z}{h} \right) = \frac{1}{nh^2} \mathbb{E} K \left( \frac{z - Z}{h} \right)^2 - \frac{1}{nh^2} \left[ \mathbb{E} K \left( \frac{z - Z}{h} \right) \right]^2.
\]

Using (4) and the fact that \( \text{Bias } \hat{f}_h(z) = O(h^2) \), we have

\[
\mathbb{E} K \left( \frac{z - Z}{h} \right) = h [f(z) + \text{Bias } \hat{f}_h(z)] = h [f(z) + O(h^2)].
\]

Hence, after a change of variable from \( x \) to \( u = (z - x)/h \),

\[
\text{Var } \hat{f}_h(z) = \frac{1}{nh^2} \int K \left( \frac{z - x}{h} \right)^2 f(x) \, dx - \frac{1}{n} [f(z) + O(h^2)]^2
\]

\[
= \frac{1}{nh} \int K(u)^2 f(z - hu) \, du - \frac{1}{n} [f(z) + O(h^2)]^2.
\]

Taking a 1st order Taylor expansion of \( f(z - hu) \) around \( h = 0 \) gives

\[
\text{Var } \hat{f}_h(z) \approx \frac{1}{nh} \int K(u)^2 [f(z) - hu f'(z)] \, du + O(n^{-1})
\]

\[
= \frac{f(z)}{nh} \int K(u)^2 \, du - \frac{f'(z)}{n} \int uK(u)^2 \, du + O(n^{-1}).
\]

Separating the term of order \( O((nh)^{-1}) \) from that of order \( O(n^{-1}) \) we have that, for \( h \) sufficiently small,

\[
\text{Var } \hat{f}_h(z) \approx \frac{1}{nh} f(z) \int K(u)^2 \, du + O(n^{-1}).
\]
Remarks

- If the sample size is fixed, increasing the bandwidth reduces the variance of $\hat{f}_h(z)$ but, from (11), it also increases its bias.

- If the sample size increases and smaller bandwidths are chosen for larger $n$, then both the bias and the variance of $\hat{f}_h(z)$ may be reduced.

- For $\hat{f}_h(z) \xrightarrow{P} f(z)$ we need both $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

To conclude, in large samples and for $h$ sufficiently small,

$$
\text{MSE} \hat{f}_h(z) \approx \frac{f(z)}{nh} \int K(u)^2 du + \frac{1}{4} m_z^2 h^4 f''(z)^2. 
$$

(6)
Global properties

Common measures of distance between two functions $f$ and $g$ are:

- the $L_1$ distance
  \[
  d_1(f, g) = \int_{-\infty}^{\infty} |f(z) - g(z)| \, dz,
  \]

- the $L_2$ distance
  \[
  d_2(f, g) = \left[ \int_{-\infty}^{\infty} [f(z) - g(z)]^2 \, dz \right]^{1/2},
  \]

- the $L_\infty$ or uniform distance
  \[
  d_\infty(f, g) = \sup_{-\infty < z < \infty} |f(z) - g(z)|.
  \]

In the case of the $L_2$ distance, a global measure of performance of $\hat{f}_h$ as an estimator of $f$ is the mean integrated squared error (MISE)

\[
\text{MISE}(\hat{f}_h) = \mathbb{E} \left[ \int_{-\infty}^{\infty} [\hat{f}_h(z) - f(z)]^2 \, dz \right]
\]

where the expectation is with respect to joint distribution of $Z_1, \ldots, Z_n$.

Under appropriate regularity conditions

\[
\text{MISE}(\hat{f}_h) = \int_{-\infty}^{\infty} \mathbb{E} [\hat{f}_h(z) - f(z)]^2 \, dz
\]

that is, the MISE is equal to the integrated MSE. Hence, in large samples and for $h$ sufficiently small, one may approximate the MISE by integrating (6) under the further assumption that the second derivative of $f$ satisfies

\[
R(f) = \int f''(z)^2 \, dz < \infty.
\]

Our global measure of performance is therefore

\[
\text{MISE}(\hat{f}_h) \approx \frac{1}{nh} \int K(u)^2 \, du + \frac{1}{4} m_2^2 h^4 R(f).
\]
Optimal choice of bandwidth

For a fixed sample size $n$ and a fixed kernel function $K$, an optimal bandwidth may be chosen by minimizing the MISE($\hat{f}_h$) with respect to $h$. As an approximation to this problem, consider minimizing the right-hand side of (7) with respect to $h$

$$\min_{h>0} \frac{1}{nh} \int K(u)^2 \, du + \frac{1}{4} m_2^2 h^4 R(f).$$

The first-order condition for this problem is

$$0 = -\frac{1}{nh^2} \int K(u)^2 \, du + m_2^2 h^3 R(f).$$

Assuming that $0 < R(f) < \infty$ and solving for $h$ gives the optimal bandwidth

$$h^* = \left[ \frac{\int K(u)^2 \, du}{m_2^2 R(f)} \right]^{1/5} n^{-1/5} = O(n^{-1/5}). \quad (8)$$

Remarks:

• $h^*$ converges to zero as $n \to \infty$ but at the rather slow rate of $n^{-1/5}$.

• Smaller values of $h^*$ are appropriate for densities that are more wiggly, that is, such that $R(f)$ is high, or kernels that are more spread out, that is, such that $m_2$ is high.

Example 8 If $f$ is a $\mathcal{N}(\mu, \sigma^2)$ density and $K$ is a standard Gaussian kernel, then

$$R(f) = \sigma^{-5} \int \phi''(z)^2 \, dz = \frac{3}{8\sqrt{\pi}} \sigma^{-5}$$

and

$$\int K(u)^2 \, du = \frac{1}{2\sqrt{\pi}}.$$

The optimal bandwidth in this case is

$$h^* = \left[ \frac{\frac{1}{2\sqrt{\pi}}}{3} \frac{8\sqrt{\pi}}{} \sigma^5 \right]^{1/5}$$

$$= \left( \frac{4}{3} \right)^{1/5} \sigma n^{-1/5} \approx 1.059 \sigma n^{-1/5}. \quad \Box$$
Optimal choice of kernel function

Substituting the optimal bandwidth $h_*$ into (7) gives

$$\text{MISE}(\hat{f}_{h_*}) \approx \frac{5}{4} C(K) R(f)^{1/5} n^{-4/5},$$

where

$$C(K) = m_2^{2/5} \left[ \int K(u)^2 du \right]^{4/5}.$$

Given the optimal bandwidth, the MISE depends on the choice of kernel function only through the term $C(K)$. Thus, an optimal kernel may be obtained by minimizing $C(K)$ with respect to $K$. 
The Epanechnikov kernel

If we confine attention to kernels that correspond to densities of distributions with mean zero and unit variance, then an optimal kernel $K_*$ may be obtained by minimizing

$$C(K) = \left[ \int K(u)^2 \, du \right]^{4/5}$$

under the side conditions

$$K(u) \geq 0 \text{ for all } u, \quad \int K(u) \, du = 1,$$

$$\int u K(u) \, du = 0, \quad \int u^2 K(u) \, du = 1.$$ 

Solving this problem (see Pagan & Ullah 1999 for details), the optimal kernel is

$$K_*(u) = \frac{3}{4}(1 - u^2) \, 1\{|u| \leq 1\},$$

called the Epanechnikov kernel.

The efficiency loss from using suboptimal kernels, however, is modest. For example, using the uniform kernel $w$ only implies an efficiency loss of about 7% with respect to $K_*$. Thus, “it is perfectly legitimate, and indeed desirable, to base the choice of kernels on other considerations, for example the degree of differentiability or the computational effort involved” (Silverman 1986, p. 43).

It turns out that the crucial choice is not what kernel to use, but rather how much to smooth. This choice partly depends on the purpose of the analysis.
Figure 5: Uniform, Gaussian and Epanechnikov kernels.
Automatic bandwidth selection

It is often convenient to be able to rely on some procedure for choosing the bandwidth automatically rather than subjectively. This is especially important when a density estimate is used as an input to other statistical procedures.

- The simplest approach consists of choosing $h = 1.059\hat{\sigma}/n^{1/5}$, where $\hat{\sigma}$ is some estimate of the standard deviation of the data. This approach works reasonably well for Gaussian kernels and data that are not too far from Gaussian.

- A second approach is based on formula (8) and chooses

$$h = \left[ \frac{\int K(u)^2 du}{m_2 R(\tilde{f})} \right]^{1/5} n^{-1/5},$$

where $\tilde{f}$ is a preliminary kernel estimate of $f$ based on some initial bandwidth choice.

- A third approach starts from the observation that the MISE of a kernel density estimator $\hat{f}_h$, based on a given kernel $K$, may be decomposed as

$$\text{MISE}(\hat{f}_h) = \mathbb{E} \int_{-\infty}^{\infty} [\hat{f}_h(z)^2 + f(z)^2 - 2\hat{f}_h(z)f(z)] dz$$

$$= M(h) + \int f(z)^2 dz,$$

where

$$M(h) = \mathbb{E} \int \hat{f}_h(z)^2 dz - 2\mathbb{E} \int \hat{f}_h(z)f(z) dz.$$ 

Hence, the MISE of $\hat{f}_h$ depends on the choice of the bandwidth only through the term $M(h)$. Thus, minimizing the MISE with respect to $h$ is equivalent to minimizing the function $M$ with respect to $h$. Because such a function is unknown, this approach suggests minimizing with respect to $h$ an unbiased estimator of the function $M$. 

32
Cross-validation

To construct an unbiased estimator of $M(h)$, notice first that $\int \hat{f}_h(z)^2 \, dz$ is unbiased for $E \int \hat{f}_h(z)^2 \, dz$.

Consider next the kernel estimator of $f$ obtained by excluding the $i$th observation
\[
\hat{f}_{(i)}(z) = \frac{1}{(n-1)h} \sum_{j \neq i} K\left(\frac{z - Z_j}{h}\right), \quad i = 1, \ldots, n. \tag{9}
\]
If the data are a sample from the distribution of $Z$, then
\[
E_n \left[ \frac{1}{n} \sum_{i=1}^{n} \hat{f}_{(i)}(Z_i) \right] = E_n \hat{f}_{(i)}(Z_i)
\]
\[
= \int E_{(i)} \hat{f}_{(i)}(z) f(z) \, dz
\]
\[
= \int E_n \hat{f}_h(z) f(z) \, dz
\]
\[
= E_n \int \hat{f}_h(z) f(z) \, dz
\]
where $E_n$ denotes expectations with respect to the joint distribution of $Z_1, \ldots, Z_n$, $E_{(i)}$ denotes expectations with respect to the joint distribution of $Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n$, and we used the fact that $E_n \hat{f}_h(z) = h^{-1} E K((z - Z)/h)$ does not depend on $n$ and so it is equal to $E_{(i)} \hat{f}_{(i)}(z)$.

Thus, an unbiased estimator of $M(h)$ is the cross-validation criterion
\[
\hat{M}(h) = \int \hat{f}_h(z)^2 \, dz - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{(i)}(Z_i). \tag{10}
\]
The cross-validation procedure minimizes $\hat{M}(h)$ with respect to $h$.  

33
Asymptotic justification for cross-validation

Let

$$I_*(Z_1, \ldots, Z_n) = \min_h \int [\hat{f}_h(z) - f(z)]^2 dz$$

denote the integrated squared error of the density estimate obtained choosing $h$ optimally for the given sample, and let

$$I_{CV}(Z_1, \ldots, Z_n) = \int [\hat{f}_{h_*}(z) - f(z)]^2 dz$$

denote the integrated squared error obtained using the bandwidth $h_*$ that minimizes the cross-validation criterion $\hat{M}_h$.

Stone (1984) showed that, if $f$ is bounded and some other mild regularity conditions hold, then

$$\Pr \left\{ \lim_{n \to \infty} \frac{I_{CV}(Z_1, \ldots, Z_n)}{I_*(Z_1, \ldots, Z_n)} = 1 \right\} = 1.$$

Thus, cross-validation achieves the best possible choice of smoothing parameter, in the sense of minimizing the integrated squared error for a given sample.

The main drawback of cross-validation is that the resulting kernel density estimates tend to be highly variable and to undersmooth the data.
Asymptotic properties of kernel density estimators

Let \( \hat{f}_n \) denote a kernel density estimate corresponding to a random sample \( Z_1, \ldots, Z_n \) from a distribution with density \( f \) and a data-dependent bandwidth \( h_n \). We now provide sufficient conditions for the sequence \( \{\hat{f}_n(z)\} \) to be consistent for \( f(z) \) and asymptotically normal, where \( z \) is any point in the support of \( f \).

We again rely on the fact that, under our iid assumption, \( \hat{f}_n(z) \) may be represented as the sample average of \( n \) iid rv's,

\[
\hat{f}_n(z) = n^{-1} \sum_{i=1}^{n} K_{in}(z),
\]

where

\[
K_{in}(z) = \frac{1}{h_n} K \left( \frac{z - Z_i}{h_n} \right).
\]

Results are easily generalized to the case when, instead of a single point of evaluation \( z \), we are interested in a fixed set of points \( z_1, \ldots, z_J \).
Consistency

Convergence in probability follows immediately from the fact that if the kernel $K$ is the density of a distribution with zero mean and finite positive variance $m_2 = \int u^2 K(u) \, du$ then, from our previous results,

$$E K_{in}(z) = f(z) + \frac{1}{2} m_2 h_n^2 f''(z) + O(h_n^3)$$

and

$$\text{Var} K_{in}(z) = \frac{1}{h_n} \left[ f(z) \int K(u)^2 \, du + h_n f'(z) \int u K(u)^2 \, du \right] + O(h_n^4).$$

The only technical problem is the behavior of $h_n$ as a function of $n$.

**Theorem 1** Let $\{Z_i\}$ be a sequence of iid continuous rv with twice continuously differentiable density $f$, and suppose that the sequence $\{h_n\}$ of bandwidths and the kernel function $K$ satisfy:

(i) $h_n \to 0$;

(ii) $nh_n \to \infty$;

(iii) $\int K(u) \, du = 1$, $\int uK(u) \, du = 0$, and $m_2 = \int u^2 K(u) \, du$ is finite and positive.

Then $\hat{f}_n(z) \xrightarrow{p} f(z)$ for every $z$ in the support of $f$. 

36
Asymptotic normality

The next theorem could in principle be used to construct approximate symmetric confidence intervals for \( f(z) \) based on the normal distribution.

**Theorem 2** In addition to the assumptions of Theorem 1, suppose that \( \sqrt{nh_n^2} \to \lambda \), where \( 0 \leq \lambda < \infty \). Then

\[
\sqrt{nh_n} [\hat{f}_n(z) - f(z)] \Rightarrow N \left( \frac{1}{2} \lambda m_2 f''(z), f(z) \int K(u)^2 \, du \right)
\]

for every \( z \) in the support of \( f \). Further, \( \sqrt{nh_n} [\hat{f}_n(z) - f(z)] \) and \( \sqrt{nh_n} [\hat{f}_n(z') - f(z')] \) are asymptotically independent for \( z \neq z' \).
Remarks:

- Although consistent, \( \hat{f}_n(z) \) is generally **asymptotically biased** for \( f(z) \).

- There are three cases when **no asymptotic bias** arises:
  - \( h_n \) is chosen such that \( \lambda = 0 \);
  - \( m_2 = 0 \), in which case the kernel \( K \) may assume **negative** values (higher-order kernel), so the estimate \( \hat{f}_n \) may fail to be a proper density;
  - \( f''(z) = 0 \), that is, the density \( f \) is **linear** at the point \( z \).

- When \( \lambda > 0 \), the additional assumption in Theorem 2 implies that \( h_n = O(n^{-1/5}) \), as for the optimal bandwidth. In this case, letting \( h_n = cn^{-1/5} \), with \( c > 0 \), gives \( \sqrt{nh_n^2} = c^{5/2} = \lambda \) and \( \sqrt{nh_n} = c^{1/2}n^{2/5} \). Therefore

  \[
  n^{2/5} [\hat{f}_n(z) - f(z)] \Rightarrow \mathcal{N} \left( \frac{1}{2} c^2 m_2 f''(z), \frac{f(z)}{c} \int K(u)^2 du \right).
  \]

  Thus, \( \hat{f}_n(z) \) converges to its asymptotic distribution at the rate of \( n^{2/5} \), which is **slower** than the rate \( n^{1/2} \) achieved by standard parametric estimators (Figure 6).

- When \( \lambda = 0 \) (no asymptotic bias), the bandwidth tends to zero at a **faster** rate than \( O(n^{-1/5}) \) but, in this case, the convergence of \( \hat{f}_n(z) \) to its asymptotic normal distribution is **slower** than \( n^{2/5} \).

- An estimate of the asymptotic variance of \( \hat{f}_n(z) \) is \( f_n(z) \int K(u)^2 du \), so a confidence interval for \( f(z) \) based on the asymptotically normal distribution of \( \hat{f}_n \) is

  \[
  \text{CI}_{1-\alpha}(f(z)) = \hat{f}_n(z) \pm z(\alpha) \left[ \frac{f_n(z) \int K(u)^2 du}{nh_n} \right]^{1/2}.
  \]

  This symmetric confidence interval may contain negative values, so its use is problematic and using **bootstrap confidence interval** may be better.
Figure 6: Rates of convergence.
1.4 Other methods for univariate density estimation

The nearest neighbor method

One of the problems with the kernel method is the fact that the bandwidth is independent of the point at which the density is evaluated. This tends to produce too much smoothing in some regions of the data and too little in others.

For any point of evaluation \( z \), let \( d_1(z) \leq d_2(z) \leq \ldots \leq d_n(z) \) be the distances (arranged in increasing order) between \( z \) and each of the \( n \) data points. The \textit{kth nearest neighbor estimate} of \( f(z) \) is defined as

\[
\hat{f}(z) = \frac{k}{2n d_k(z)}, \quad k < n. \tag{13}
\]

The motivation for this estimate is the fact that, if \( h \) is sufficiently small, then we would expect a fraction of the observations approximately equal to \( 2hf(z) \) to fall in a small interval \([z-h, z+h]\) around \( z \). Since the interval \([z-d_k(z), z+d_k(z)]\) contains by definition exactly \( k \) observations, we have

\[
\frac{k}{n} = 2d_k(z)f(z).
\]

Solving for \( f(z) \) gives the density estimate (13).

Unlike the uniform kernel estimator, which is based on the number of observations falling in a box of fixed width centered at the point \( z \), the nearest neighbor estimator is inversely proportional to the width of the box needed to exactly contain the \( k \) observations nearest to \( z \). The smaller is the density of the data around \( z \), the larger is this width.

The number \( k \) regulates the \textit{degree of smoothness} of the estimator, with larger values of \( k \) corresponding to smoother estimates. The fraction \( \lambda = k/n \) of sample points in each neighborhood is called the \textit{span}. 

40
Problems with the nearest neighbor method

- $\hat{f}$ does not integrate to one and so it is not a proper density.

- Although continuous, $\hat{f}$ is not smooth because its derivatives are discontinuous at all points of the form $(Z_{[i]} + Z_{[i+k]})/2$.

One way of overcoming the second problem is to notice that

$$k = \sum_{i=1}^{n} w\left(\frac{z - Z_i}{d_k(z)}\right),$$

where $w$ is the uniform kernel, so $w((z - Z_i)/d_k(z))$ is equal to 1/2 if $Z_i$ is within distance $d_k(z)$ from $z$ and is equal to zero otherwise. So $\hat{f}(z)$ may alternatively be represented as

$$\hat{f}(z) = \frac{1}{nd_k(z)} \sum_{i=1}^{n} w\left(\frac{z - Z_i}{d_k(z)}\right).$$

Thus, the nearest neighbor estimator may be regarded as a uniform kernel density estimator with a varying bandwidth $d_k(z)$.

The lack of smoothness of the nearest neighbor estimator may therefore be overcome by considering a generalized nearest neighbor estimate of the form

$$\tilde{f}(z) = \frac{1}{nd_k(z)} \sum_{i=1}^{n} K\left(\frac{z - Z_i}{d_k(z)}\right).$$

If $K$ is a smooth kernel and the bandwidth $d_k(z)$ is continuously differentiable, then $\tilde{f}$ is a smooth estimator of $f$. 
Nonparametric ML

The log-likelihood of a sample $Z_1, \ldots, Z_n$ from a distribution with strictly positive density $f_0$ is defined as

$$L(f_0) = c + \sum_{i=1}^{n} \ln f_0(Z_i),$$

where $c$ is an arbitrary constant. One may then think of estimating $f_0$ by

$$\hat{f} = \arg\max_{f \in F} L(f),$$

where $F$ is some family of densities.

- When $F = \{f(z; \theta), \theta \in \Theta\}$ is a known parametric family of densities, one obtains the classical ML estimator

  $$\hat{f}(z) = f(z, \hat{\theta}),$$

  where $\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta)$ and $L(\theta) = c + \sum_{i=1}^{n} \ln f(Z_i, \theta)$.

- When $F$ includes all strictly positive densities on the real line, one can show that the nonparametric ML estimator is

  $$\hat{f}(z) = n^{-1} \sum_{i=1}^{n} \delta(z - Z_i),$$

  where

  $$\delta(u) = \begin{cases} \infty, & \text{if } u = 0, \\ 0, & \text{otherwise,} \end{cases}$$

  is the Dirac delta function. Thus, $\hat{f}$ is a function equal to infinity at each of the sample points and equal to zero otherwise (why is this unsurprising?).
Maximum penalized likelihood

Being infinitely irregular, \( \hat{f} \) provides an unsatisfactory solution to the problem of estimating \( f_0 \). If one does not want to make parametric assumptions, an alternative is to introduce a penalty for lack of smoothness and then maximize the log likelihood function subject to this constraint.

To quantify the degree of smoothness of a density \( f \) consider again the functional

\[
R(f) = \int f''(z)^2 \, dz.
\]

If \( f \) is wiggly, then \( R(f) \) is large. If the support of the distribution is an interval and \( f \) is linear or piecewise linear, then \( R(f) = 0 \).

One may then consider maximizing the penalized sample log-likelihood

\[
L_\lambda(f) = \sum_{i=1}^{n} \ln f(Z_i) - \lambda R(f),
\]

where \( \lambda > 0 \) represents the trade-off between smoothness and fidelity to the data. A maximum penalized likelihood density estimator \( \tilde{f} \) maximizes \( L_\lambda \) over the class of densities with continuous 1st derivative and square integrable 2nd derivative. The smaller is \( \lambda \), the rougher in terms of \( R(\hat{f}) \) is the maximum penalized likelihood estimator.

Advantages of this method:

- It makes explicit two conflicting goals in density estimation:
  - maximizing fidelity to the data, represented here by the term \( \sum_{i} \ln f(Z_i) \),
  - avoiding densities that are too wiggly, represented here by the term \( -\lambda R(f) \).

- The method places density estimation within the context of a unified approach to curve estimation.

- The method can be given a nice Bayesian interpretation.

Its main disadvantage is that the resulting estimate \( \tilde{f} \) is defined only implicitly, as the solution to a maximization problem.
1.5 Multivariate density estimators

Let $Z_1, \ldots, Z_n$ be a sample from the distribution of a random $m$-vector $Z$ with density function $f(z) = f(z_1, \ldots, z_m)$. How can we estimate $f$ nonparametrically?

Multivariate kernel density estimators

The natural starting point is the following multivariate generalization of the univariate kernel density estimator

$$
\hat{f}(z) = \hat{f}(z_1, \ldots, z_m) = \frac{1}{nh_1 \cdots h_m} \sum_{i=1}^{n} K_m \left( \frac{z_1 - Z_{i1}}{h_1}, \ldots, \frac{z_m - Z_{im}}{h_m} \right),
$$

where $K_m : \mathbb{R}^m \to \mathbb{R}$ is a multivariate kernel.

Example 9 An example of multivariate kernel is the product kernel

$$
K_m(u_1, \ldots, u_m) = \prod_{j=1}^{m} K(u_j),
$$

where $K$ is a univariate kernel. □

We now consider consistency and asymptotic normality of a special class of multivariate kernel density estimators of the form

$$
\hat{f}(z) = \frac{1}{nh_m} \sum_{i=1}^{n} K_m \left( \frac{z_1 - Z_{i1}}{h}, \ldots, \frac{z_m - Z_{im}}{h} \right).
$$

This type of estimators are a special case of (14) obtained by using the same bandwidth $h$ for each component of $Z$. They may be appropriate when the data have being previously rescaled using some preliminary estimate of scale.

Example 10 In the case of a product kernel, (15 becomes

$$
\hat{f}(z) = \frac{1}{nh_m} \sum_{i=1}^{n} \prod_{j=1}^{m} K \left( \frac{z_j - Z_{ij}}{h} \right),
$$

where $K$ is a univariate kernel. □
Asymptotic properties

**Theorem 3** Let \( \{Z_i\} \) be a sequence of iid continuous random \( m \)-vectors with a twice continuously differentiable density \( f \), and suppose that the sequence \( \{h_n\} \) of bandwidths and the multivariate kernel \( K_m : \mathbb{R}^m \to \mathbb{R} \) satisfy:

(i) \( h_n \to 0 \);

(ii) \( nh_n^m \to \infty \);

(iii) \( \int K_m(u) du = 1, \int uK_m(u) du = 0 \) and \( \int uu^\top K_m(u) du = M_2 \), a finite \( m \times m \) matrix.

Then \( \hat{f}_n(z) \xrightarrow{p} f(z) \) for every \( z \) in the support of \( f \).

**Theorem 4** In addition to the assumptions of Theorem 3, suppose that \( h_n^2(nh_n^m)^{1/2} \to \lambda \), where \( 0 \leq \lambda < \infty \). Then

\[
(nh_n^m)^{1/2}[\hat{f}_n(z) - f(z)] \Rightarrow N\left( \frac{1}{2} \lambda b(z), f(z) \int K_m(u)^2 du \right)
\]

for every \( z \) in the support of \( f \), where \( b(z) = \text{tr}[f''(z)M_2] \). Further, \( (nh_n^m)^{1/2}[\hat{f}_n(z) - f(z)] \) and \( (nh_n^m)^{1/2}[\hat{f}_n(z') - f(z')] \) are asymptotically independent for \( z \neq z' \).

**Remarks:**

- The speed of convergence of \( \hat{f}_n(z) \) to its asymptotically normal distribution is inversely related to the dimension \( m \) of \( Z \) (a manifestation of the “curse-of-dimensionality” problem). When \( m = 1 \), we have the results in Theorem 2.

- When \( \lambda > 0 \), the additional assumption in Theorem 4 implies that \( h_n = O(n^{-1/(m+4)}) \). In this case, putting \( h_n = cn^{-1/(m+4)} \), with \( c > 0 \), gives \( \lambda = c^{(m+4)/2} \) and \( (nh_n^m)^{1/2} = c^{m/2}n^{2/(m+4)} \), and therefore

\[
n^{2/(m+4)} [\hat{f}_n(z) - f(z)] \Rightarrow N\left( \frac{1}{2} c^2 b(z), \frac{f(z)}{c^m} \int K_m(u)^2 du \right).
\]

- When \( \lambda = 0 \), the bandwidth tends to zero at a faster rate than \( O(n^{-1/(m+4)}) \) but the convergence of \( \hat{f}_n(z) \) to its asymptotically normal distribution is slower than the optimal rate.
The curse-of-dimensionality problem

Although conceptually straightforward, extending univariate nonparametric methods to multivariate settings is problematic for at least two reasons:

- While contour plots are enough for two dimensions, how to represent the results of a nonparametric analysis involving three or more variables is a neglected but practically important problem.

- When one-dimensional nonparametric methods are generalized to higher dimensions, their statistical properties deteriorate very rapidly because of the so-called curse-of-dimensionality problem, namely the fact that the volume of data required to maintain a tolerable degree of statistical precision grows much faster than the number of variables under examination.

For these reasons, simple generalizations of one-dimensional methods to the case of more than two or three variables tend to produce results that are difficult to represent and are too irregular, unless the size of the available data is very large.
Example 11 Consider a multivariate histogram constructed for a sample from the distribution of a random $m$-vector $Z$ with a uniform distribution on the $m$-dimensional unit hypercube, that is, whose components are iid as $\mathcal{U}(0, 1)$.

If we partition the unit hypercube into hypercubical cells of side equal to $h$, then each cell contains on average only $h^m$ percent of the data. Assume that at least 30 observations per cell are needed for a tolerably accurate histogram estimate. Then an adequate sample should have a number of observations at least equal to $n_* = 30h^{-m}$.

The table below shows some calculations for $m = 1, \ldots, 5$ and $h = .10$ and $h = .05$.

<table>
<thead>
<tr>
<th>Number $m$ of variables in $Z$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h = .10$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>number of cells</td>
<td>10</td>
<td>100</td>
<td>1,000</td>
<td>10,000</td>
<td>100,000</td>
</tr>
<tr>
<td>$n_*$</td>
<td>300</td>
<td>3,000</td>
<td>30,000</td>
<td>300,000</td>
<td>3,000,000</td>
</tr>
<tr>
<td>$h = .05$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>number of cells</td>
<td>20</td>
<td>400</td>
<td>8,000</td>
<td>160,000</td>
<td>3,200,000</td>
</tr>
<tr>
<td>$n_*$</td>
<td>600</td>
<td>12,000</td>
<td>240,000</td>
<td>4,800,000</td>
<td>96,000,000</td>
</tr>
</tbody>
</table>

Leaving aside the nontrivial problem of how it could be represented, a 5-dimensional histogram is likely to be estimated too imprecisely to be of practical use, unless the sample size is truly gigantic. □
Project pursuit

One method for nonparametric estimation of multivariate densities is projection pursuit (PP), introduced by Friedman, Stuetzle and Schroeder (1984).

This method assumes that the density of a random $m$-vector $Z$ may well be approximated by a density of the form

$$f_*(z) = f_0(z) \prod_{j=1}^{J} f_j(\alpha_j^\top z),$$

(16)

where $f_0$ is a known $m$-variate density, $\alpha_j$ is a vector with unit norm, $\alpha_j^\top z = \sum_{h=1}^{m} \alpha_{jh}z_h$ is a linear combination or “projection” of the variables in $Z$, and $f_1, \ldots, f_p$ are smooth univariate functions, called ridge functions.

The choice of $f_0$ is left to the investigator and should reflect prior information available about the problem.

The estimation algorithm determines the number $J$ of terms and the vectors $\alpha_j$ in (16), and constructs nonparametric estimates of the ridge functions $f_1, \ldots, f_J$ by minimizing a suitably chosen index of goodness-of-fit, or projection index. The curse-of-dimensionality problem is bypassed by using linear projections and nonparametric estimates of univariate ridge functions.

As a by-product of this method, the graphical information provided by the shape of the estimated ridge functions may be useful for exploring and interpreting the multivariate distribution of the data.

The PP method may be regarded as a generalization of the principal components method, where multidimensional data are projected linearly onto subspaces of much smaller dimension, chosen to minimize the variance of the projections.
Semi-nonparametric methods

Gallant and Nychka (1987) proposed to approximate the $m$-variate density $f(z_1, \ldots, z_m)$ by a **Hermite polynomial expansion**.

In the **bivariate** case ($m = 2$), their approximation is

$$f_*(z_1, z_2) = \frac{1}{\psi_K} \tau_K(z_1, z_2)^2 \phi(z_1) \phi(z_2),$$

where

$$\tau_K(z_1, z_2) = \sum_{h,k=0}^{K} \tau_{hk} z_1^h z_2^k$$

is a **polynomial** of order $K$ in $z_1$ and $z_2$, and

$$\psi_K = \int \int (z_1, z_2)^2 \phi(z_1) \phi(z_2) \, dz_1 \, dz_2$$

is a **normalization factor** to ensure that $f_*$ integrates to one. The order $K$ is typically chosen by minimizing some information criterion such as AIC or BIC. Because, **after the choice of $K$**, the model for $f$ is fully parametric, this is an example of **semi-nonparametric (SNP)** approach.

The class of densities that can be approximated in this way includes densities with arbitrary skewness and kurtosis, but excludes **violently oscillatory densities** or densities with **too fat or too thin tails**.

Since the polynomial expansion in (1.5) is **invariant** to multiplication of $\tau = (\tau_{00}, \tau_{01}, \ldots, \tau_{KK})$ by a scalar, some **normalization** is needed. A convenient normalization is $\tau_{00} = 1$. Under this normalization, expanding the square of the polynomial in (1.5) and rearranging terms gives

$$f_*(z_1, z_2) = \frac{1}{\psi_K} \left[ \sum_{h,k=0}^{2K} \gamma_{hk} z_1^h z_2^k \right] \phi(z_1) \phi(z_2),$$

where

$$\gamma_{hk} = \sum_{r=a_h}^{b_h} \sum_{s=a_k}^{b_k} \tau_{rs} \tau_{h-r, k-s},$$

with

$$a_h = \max(0, h - K), \quad b_h = \min(h, K),$$
$$a_k = \max(0, k - K), \quad b_k = \min(k, K).$$
1.6 Stata commands

We now briefly review the commands for histogram and kernel density estimation available in Stata, version 12.

These include the `histogram` and the `kdensity` commands. Both commands estimate univariate densities. The package `akdensity` in van Kerm (2012) allows estimating both the density and the distribution function by the kernel method.

Recently, some articles in the *Stata Journal* have been devoted to the use of nonparametric or semi-nonparametric methods for density estimation in a variety of statistical problems. Examples include De Luca (2008) and De Luca and Perotti (2011).
The `histogram` command

The basic syntax is:

\[
\text{histogram } \text{varname} [\text{if}] [\text{in}] [\text{weight}] [, [\text{continuous} \text{opts} | \text{discrete} \text{opts}] \text{options}]
\]

where:

- `varname` is the name of a continuous variable, unless the `discrete` option is specified,

- `continuous` opts includes:
  - `bin(#)`: sets the number of bins to #,
  - `width(#)`: sets the width of bins to #,
  - `start(#)`: sets the lower limit of first bin to # (the default is the observed minimum value of `varname`).

`bin()` and `width()` are alternatives. If neither is specified, results are the same as if `bin(k)` had been specified, with

\[
k = \min\{\sqrt{n}, 10 \cdot \ln n / \ln 10\},
\]

where \(n\) is the (weighted) number of observations.

- `discrete` opts includes:
  - `discrete`: specifies that the data are discrete and you want each unique value of `varname` to have its own bin (bar of histogram),

- `options` includes:
  - `density`: draws as density (the default),
  - `fraction`: draws as fractions,
  - `frequency`: draws as frequencies,
  - `percent`: draws as percentages,
  - `addlabels`: adds heights label to bars,
  - `normal`: adds a normal density to the graph,
  - `kdensity`: adds a kernel density estimate to the graph.
The **kdensity command**

The basic syntax is:

```
kdensity varname [if] [in] [weight] [, options]
```

where *options* includes:

- **kernel(kernel)**: specifies the kernel function. The available kernels include **epanechnikov** (default), **biweight**, **cosine**, **gaussian**, **parzen**, **rectangle**, and **triangle**.

- **bwidth(#)**: specifies half-width of kernel. If not specified, the half-width calculated and used corresponds to $h = 1.059 \sigma n^{-1/5}$, which would minimize the MISE if the data were Gaussian and a Gaussian kernel were used (there is a little inconsistency here, as Epanechnikov is the default kernel function).

- **generate(newvar_x newvar_d)**: stores the estimation points in *newvar_x* and the density estimate in *newvar_d*.

- **n(#)**: estimates density using # points. The default is min(n,50), where n is the number of observations in memory.

- **at(var_x)**: estimates the density using the values specified by *var_x*.

- **nograph**: suppresses graph.

- **normal**: adds a normal density to the graph.

- **student(#)**: adds a Student’s *t* density with # degrees of freedom to the graph.
2 Nonparametric Regression Estimators

We now assume that the data \((X_1, Y_1), \ldots, (X_n, Y_n)\) are a sample from the joint
distribution of \((X, Y)\), for which the conditional mean function (CMF)
\[
\mu(x) = \mathbb{E}(Y \mid X = x)
\]
is well defined, and consider the problem of estimating \(\mu(x)\) nonparamet-
rically. We focus on the case when \(X\) is continuous, because when \(X\) is
discrete with mass points at \(x_1, \ldots, x_J\), the CMF may simply be estimated by
\[
\hat{Y}_j = \frac{\sum_{i=1}^{n} 1\{X_i = x_j\} Y_i}{\sum_{i=1}^{n} 1\{X_i = x_j\}}, \quad j = 1, \ldots, J,
\]
the average of the sample values of \(Y\) for the cases when \(X = x_j\).

This is a more general problem than it may look at first.

**Example 12** The conditional probability function of a 0-1 rv \(Y\) satisfies
\[
\pi(x) = \Pr\{Y = 1 \mid X = x\} = \mathbb{E}(Y \mid X = x).
\]
Thus, the problem of estimating \(\pi(x)\) nonparametrically reduces to the prob-
lem of estimating the CMF of \(Y\) nonparametrically.

**Example 13** The conditional variance function (CVF) of a rv \(Y\) satisfies
\[
\sigma(x)^2 = \text{Var}(Y \mid X = x) = \mathbb{E}(Y^2 \mid X = x) - [\mathbb{E}(Y \mid X = x)]^2
\]
Thus, the problem of estimating \(\sigma(x)^2\) nonparametrically reduces to the prob-
lem of estimating the CMF’s of \(Y\) and \(Y^2\) nonparametrically.
Linear nonparametric regression estimators

We restrict attention to the class of nonparametric estimators of $\mu(x)$ that are linear, that is, of the form

$$\hat{\mu}(x) = \sum_{j=1}^{n} S_j(x) Y_j,$$

where the weight $S_j(x)$ assigned to $Y_j$ depends only on the $X_i$’s and the evaluation point $x$, **not** on the $Y_i$’s.

**Linearity** of this class of estimators is important and useful because:

- It lowers the computational burden relative to nonlinear estimators.
- It reduces the task of understanding the difference between alternative estimates of $\mu(x)$ to the task of understanding the differences in the weights $S_j(x)$.
- It simplifies considerably the task of evaluating the statistical properties of these estimators.
Regression smoothers

It is useful to represent a linear nonparametric regression estimator as a linear regression smoother.

Let $Y$ be the $n$-vector of observations on the outcome variable and let $X$ be the matrix of $n$ observations on the $k$ covariates. A regression smoother is a way of using $Y$ and $X$ to produce a vector $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)$ of fitted values that is less variable than $Y$ itself.

A regression smoother is linear if it can be represented as

$$\hat{\mu} = SY,$$

where $S = [s_{ij}]$ is an $n \times n$ smoother matrix that depends on $X$ but not on $Y$. Thus, the class of linear regression smoothers coincides with the class of linear predictors of $Y$.

The $i$th element of a linear regression smoother is a weighted average

$$\hat{\mu}_i = \sum_{j=1}^{n} s_{ij} Y_j$$

of the elements of $Y$ and $s_{ij}$ is the weight assigned to the $j$th element of $Y$ in the construction of $\hat{\mu}_i$. Linear nonparametric regression estimates are linear smoothers where $\hat{\mu}_i = \hat{\mu}(X_i)$ and $s_{ij} = S_j(X_i)$.

Example 14 A parametric example of linear smoother is the vector of OLS fitted values $\hat{\mu} = X\hat{\beta} = SY$, where $\hat{\beta} = (X^T X)^{-1} X^T Y$ and the smoother matrix $S = X(X^T X)^{-1} X^T$ is symmetric and idempotent.

A smoother matrix is not necessarily symmetric or idempotent. A matrix $S$ is said to preserve the constant if $S^T \iota = \iota$, where $\iota$ is a vector of ones. If a smoother matrix $S$ preserves the constant, then

$$n^{-1} \sum_{i=1}^{n} \hat{\mu}_i = n^{-1} \iota_n^T SY = n^{-1} \iota_n^T Y = \bar{Y},$$

where $\bar{Y}$ is the sample mean of $Y$. For example, the OLS smoother matrix preserves the constant if the design matrix $X$ contains a column of ones.

We now present other examples of linear regression smoothers. For ease of presentation, we assume that $X$ is a scalar rv.
2.1 Polynomial regressions

A polynomial regression represents a parsimonious and relatively flexible way of approximating an unspecified CMF.

A $k$-degree polynomial regression approximates $\mu(x)$ by a function of the form

$$m(x) = \alpha + \beta_1 x + \cdots + \beta_k x^k, \quad \beta_k \neq 0.$$

A polynomial regression estimated by OLS corresponds to a linear regression smoother defined by a symmetric idempotent smoother matrix.

Polynomials are frequently used because they can be evaluated, integrated, differentiated, etc., very easily. For example, if $m(x)$ is a $k$-degree polynomial, then

$$m'(x) = \beta_1 + 2\beta_2 x + \cdots + k\beta_k x^{k-1}$$

is a $(k - 1)$-degree polynomial, and

$$\int m(x) \, dx = \gamma + \alpha x + \frac{1}{2} \beta_1 x^2 + \cdots + \frac{1}{k + 1} \beta_k x^{k+1}$$

is a $(k + 1)$-degree polynomial.

However, if the CMF is very irregular, even on small regions of the approximation range, then a polynomial approximation tends to be poor everywhere (Figure 7).
Figure 7: Polynomial regression estimates. The sample consists of 200 observations from a Gaussian model with $\mu(x) = 1 - x + \exp[-50(x - .5)^2]$. 
2.2 Regression splines

One way of avoiding global dependence on local properties of the function $\mu(x)$ is to consider piecewise polynomial functions.

A regression spline is a smooth piecewise polynomial function and therefore represents a very flexible way of approximating $\mu(x)$. The simplest example is linear splines.
Linear splines

Select \( J \) distinct points or \textbf{knots}, \( c_1 < \cdots < c_J \), on the support of \( X \). These points define a partition of \( \mathbb{R} \) into \( J + 1 \) intervals. To simplify the notation, also define two \textbf{boundary knots} \( c_0 < c_1 \) and \( c_{J+1} > c_J \).

A \textbf{linear spline} is a \textbf{continuous piecewise linear function} of the form

\[
m(x) = \alpha_j + \beta_j x, \quad c_{j-1} < x \leq c_j, \quad j = 1, \ldots, J + 1.
\]

For \( m \) to be \textbf{continuous} at the first knot \( c_1 \), the model parameters must satisfy

\[
\alpha_1 + \beta_1 c_1 = \alpha_2 + \beta_2 c_1,
\]

which implies that \( \alpha_2 = \alpha_1 + (\beta_1 - \beta_2) c_1 \). On the interval \((c_0, c_2]\), we therefore have

\[
m(x) = \begin{cases} 
\alpha_1 + \beta_1 x, & \text{if } c_0 < x \leq c_1, \\
\alpha_1 + (\beta_1 - \beta_2) c_1 + \beta_2 x, & \text{if } c_1 < x \leq c_2.
\end{cases}
\]

A \textbf{more compact} representation of \( m \) on \((c_0, c_2]\) is

\[
m(x) = \alpha + \beta x + \gamma_1 (x - c_1)_+, \quad c_0 < x \leq c_2,
\]

where \( \alpha = \alpha_1, \beta = \beta_1, \gamma_1 = \beta_2 - \beta_1 \) and \( (x - c_1)_+ = \max(0, x - c_1) \).

Repeating this argument for all other knots, a \textbf{linear} spline may be represented as

\[
m(x) = \alpha + \beta x + \sum_{j=1}^{J} \gamma_j (x - c_j)_+,
\]

where \( \gamma_j = \beta_{j+1} - \beta_j \).

\textbf{Remarks}

- The number of free parameters in the model is only \( J + 2 \), less than the number \( 2(J + 1) \) of parameters of an \textbf{unrestricted} piecewise linear function. The difference \( 2(J + 1) - (J + 2) \) is equal to the \textbf{number of constraints} that must be imposed to ensure continuity.

- Although \textbf{continuous}, a linear spline is \textbf{not differentiable} for its derivative is a step function with jumps at \( c_1, \ldots, c_J \).

- This problem may be avoided by considering \textbf{smooth higher-order} (quadratic, cubic, etc.) piecewise polynomial functions.
Cubic splines

A cubic spline is a twice continuously differentiable piecewise cubic function. Given $J$ distinct knots $c_1 < \cdots < c_J$ in the support of $X$, it possesses the parametric representation

$$m(x) = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=1}^{J} \gamma_j (x - c_j)^3,$$

which contains $J + 4$ free parameters. It is easy to verify that $m$ satisfies all properties of a cubic spline, namely:

- it is a cubic polynomial in each subinterval $[c_j, c_{j+1}]$;
- it is twice continuously differentiable;
- its third derivative is a step function with jumps at $c_1, \ldots, c_J$.

Remarks:

- A natural cubic spline is a cubic spline that is forced to be linear outside the boundary knots $c_0$ and $c_{J+1}$. The number of free parameters in this case is just $J + 2$.
- The representation (18) lends itself directly to estimation by OLS.
- In general, given the sequence of knots and the degree of the polynomial, a regression spline may be estimated by an OLS regression of $Y$ on an appropriate set of vectors that represent a basis for the selected family of piecewise polynomial functions evaluated at the sample values of $X$.
- Regression splines estimated by OLS give linear regression smoothers defined by symmetric idempotent smoother matrices. The number and position of the knots control the flexibility and smoothness of the approximation.

Problems with regressions splines:

- How to select the number and the position of the knots.
- Considerable increase in the degree of complexity of the problem when $X$ is multivariate (Wahba 1990).
Figure 8: Splines (blue) vs. polynomial (black) regression estimates. The sample consists of 200 observations from a Gaussian model with $\mu(x) = 1 - x + \exp[-50(x - .5)^2]$. 

\[\mu(x) = 1 - x + \exp[-50(x - .5)^2]\]
2.3 The kernel method

Recall that the CMF may be written

\[ \mu(x) = \int y \frac{f(x, y)}{f_X(x)} \, dy = \frac{c(x)}{f_X(x)}, \]

where \( f_X(x) = \int f(x, y) \, dy \) is the marginal density of \( X \) and

\[ c(x) = \int y f(x, y) \, dy. \]

Thus, given nonparametric estimates \( \hat{c}(x) \) and \( \hat{f}_X(x) \) of \( c(x) \) and \( f_X(x) \), a nonparametric estimate of \( \mu(x) \) is

\[ \hat{\mu}(x) = \frac{\hat{c}(x)}{\hat{f}_X(x)}. \quad (19) \]

An advantage of the kernel method is that, under certain conditions, such an estimate may be computed directly without the need of estimating the joint density \( f(x, y) \).
The Nadaraya-Watson estimator

Consider the bivariate kernel density estimate

\[ \hat{f}(x, y) = \frac{1}{nh_X h_Y} \sum_{i=1}^{n} K_*(\frac{x - X_i}{h_X}, \frac{y - Y_i}{h_Y}), \]

where \( K_*: \mathbb{R}^2 \to \mathbb{R} \). If \( h_X = h_Y = h \) and \( \hat{f}_X(x) \) is the estimate of \( f_X(x) \) based on the kernel \( K(x) = \int K_*(x, y) \, dy \), then

\[ \int \hat{f}(x, y) \, dy = \hat{f}_X(x) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{x - X_i}{h} \right). \]

If, in addition, \( K_*(x, y) \) is such that \( \int y \, K_*(x, y) \, dy = 0 \), then

\[ \hat{c}(x) = \int y \, \hat{f}(x, y) \, dy = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{x - X_i}{h} \right) Y_i. \]

Substituting these two expressions into (19) gives the Nadaraya–Watson (NW) or kernel regression estimator

\[ \hat{\mu}(x) = \frac{\sum_{i=1}^{n} K \left( \frac{x - X_i}{h} \right) Y_i}{\sum_{i=1}^{n} K \left( \frac{x - X_i}{h} \right)}. \] (20)

Thus, \( \hat{\mu}(x) \) is a weighted average of the \( Y_j \) with nonnegative weights

\[ w_j(x) = \left[ \sum_{i=1}^{n} K \left( \frac{x - X_i}{h} \right) \right]^{-1} K \left( \frac{x - X_j}{h} \right), \quad j = 1, \ldots, n, \]

which add up to one. The NW estimator is therefore a linear smoother whose smoother matrix \( S = [w_j(X_i)] \) depends on the sample values \( X_i \), the selected kernel \( K \), and the bandwidth \( h \).

If the kernel corresponds to a unimodal density with mode at zero, then the closer \( X_j \) is to \( x \), the bigger is the weight assigned to \( Y_j \) in forming \( \hat{\mu}(x) \). If the kernel vanishes outside the interval \([-1, 1]\), then only values of \( Y \) for which \( |x - X_j| < h \) enter the summation. This may be exploited to speed up the computations.
Example 15 A family of kernels with **bounded support** which contains many kernels used in practice is

\[ K_\alpha(u) = c_\alpha (1 - u^2)^\alpha \mathbb{1}\{0 \leq |u| \leq 1\}, \quad \alpha = 0, 1, 2, \ldots, \]

where the constant \(c_\alpha\), chosen such that \(K_\alpha\) is a proper density, is

\[ c_\alpha = 2^{-2\alpha-1}\Gamma(2\alpha + 2) \Gamma(\alpha + 1)^{-2} \]

and \(\Gamma(z)\) is the gamma function.

- If \(\alpha = 0\), we obtain the **uniform kernel**

\[ K(u) = \begin{cases} 
1/2, & \text{if } -1 \leq u < 1, \\
0, & \text{otherwise}. 
\end{cases} \]

- If \(\alpha = 1\), we obtain the **Epanechnikov kernel**

\[ K(u) = \frac{3}{4}(1 - u^2) \mathbb{1}\{|u| \leq 1\}. \]

- If \(\alpha = 2\), we obtain the **quartic kernel**

\[ K(u) = \frac{15}{16}(1 - u^2)^2 \mathbb{1}\{|u| \leq 1\}. \]

- If \(\alpha \to \infty\), then \(K_\alpha\) converges to the **Gaussian kernel**.

\(\square\)
Figure 9: Gaussian kernel regression estimates. The sample consists of 200 observations from a Gaussian model with $\mu(x) = 1 - x + \exp[-50(x - .5)^2]$.
2.4 The nearest neighbor method

This method suggests considering the \(X_i\) that are closest to the evaluation point \(x\) and estimating \(\mu(x)\) by averaging the corresponding values of \(Y\).

A kth nearest neighborhood of \(x\), denoted by \(O_k(x)\), consists of the \(k\) points that are closest to \(x\). The corresponding nearest neighbor (NN) estimate of \(\mu(x)\) is

\[
\hat{\mu}(x) = \frac{1}{k} \sum_{j \in O_k(x)} Y_j = \sum_{j=1}^{n} w_j(x) Y_j,
\]

where

\[
w_j(x) = \begin{cases} 
1/k, & \text{if } j \in O_k(x), \\
0, & \text{otherwise}.
\end{cases}
\]

Thus, \(\hat{\mu}(x)\) is just a running mean or moving average of the \(Y_i\).

The NN estimate is clearly a linear smoother, and its degree of smoothness depends on the value of \(k\) or, equivalently, on the span \(\lambda = k/n\) of the neighborhood.

If \(X\) is a vector, then the choice of the metric is of some importance. Two possibilities are:

- the Euclidean distance
  \[
d(x, x') = [(x - x')^\top (x - x')]^{1/2},
\]

- the Mahalanobis distance
  \[
d(x, x') = [(x - x')^\top \hat{\Omega}^{-1} (x - x')]^{1/2},
\]
  where \(\hat{\Omega}\) is a pd estimate of the covariance matrix of the covariates.

Notice that both the NN estimator and the NW estimator solve the following weighted LS problem

\[
\min_{c \in \mathbb{R}} \sum_{i=1}^{n} w_i(x) (Y_i - c)^2.
\]

The two estimators correspond to different choices of weights.
Running lines

Using a running mean may give severe biases near the boundaries of the data, where neighborhoods tend to be highly asymmetric.

A way of reducing this bias is to replace the running mean by a running line

\[ \hat{\mu}(x) = \hat{\alpha}(x) + \hat{\beta}(x)x, \]

where \( \hat{\alpha}(x) \) and \( \hat{\beta}(x) \) are the intercept and the slope of the OLS line computed using only the \( k \) points (\( k > 2 \)) in the neighborhood \( O_k(x) \).

Computation of (21) for the \( n \) sample points may be based on the recursive formulae for OLS and only requires a number of operations of order \( O(n) \).

Notice that \( \hat{\alpha}(x) \) and \( \hat{\beta}(x) \) in (21) solve the locally weighted LS problem

\[ \min_{(\alpha, \beta) \in \mathbb{R}^2} \sum_{i=1}^{n} w_i(x)(Y_i - \alpha - \beta X_i)^2, \]

where the weight \( w_i(x) \) is equal to one if the \( i \)th point belongs to \( O_k(x) \) and to zero otherwise.

Running lines tend to produce ragged curves but the method can be improved replacing OLS by locally WLS, with weights that decline as the distance of \( X_i \) from \( x \) increases.
Lowess

An example of locally weighted LS estimator is the lowess (LOcally WEighted Scatterplot Smoother) estimator introduced by Cleveland (1979) and computed as follows:

**Algorithm 1**

1. **Identify the kth nearest neighborhood** $O_k(x)$ of $x$.

2. **Compute the distance** $\Delta(x) = \max_{i \in O_k(x)} d(X_i, x)$ of the point in $O_k(x)$ that is farthest from $x$.

3. **Assign weights** $w_i(x)$ to each point in $O_k(x)$ according to the tricube function

   $$w_i(x) = W \left( \frac{d(X_i, x)}{\Delta(x)} \right),$$

   where

   $$W(u) = \begin{cases} (1 - u^3)^3, & \text{if } 0 \leq u < 1, \\ 0, & \text{otherwise}. \end{cases}$$

4. **Compute** $\hat{\mu}(x)$ as the predicted value of $Y$ corresponding to $x$ from a WLS regression that uses the observations in $O_k(x)$ and the weights defined in step (3).

Notice that the weight $w_i(x)$ is maximum when $X_i = x$, decreases as the distance of $X_i$ from $x$ increases, and becomes zero if $X_i$ is the $k$th nearest neighbor of $x$. 
Figure 10: Lowess estimates. The sample consists of 200 observations from a Gaussian model with \( \mu(x) = 1 - x + \exp[-50(x - .5)^2] \).
Local polynomial fitting

Another straightforward generalization is **local polynomial fitting** (Fan & Gijbels 1996), which replaces the running line by a **running polynomial** of order $k \geq 1$,

$$
\hat{\mu}(x) = \hat{\alpha}(x) + \hat{\beta}_1(x)x + \cdots + \hat{\beta}_k(x)x^k,
$$

where the coefficients $\hat{\alpha}, \hat{\beta}_1(x), \ldots, \hat{\beta}_k(x)$ solve the problem

$$
\min_{(\alpha, \beta_1, \ldots, \beta_k) \in \mathbb{R}^2} \sum_{i=1}^{n} K_i(x)(Y_i - \alpha - \beta_1X_i - \cdots - \beta_kX_i^k)^2,
$$

where the $K_i(x)$ are general kernel weights (Figure 11).
Figure 11: Locally linear fitting with Gaussian kernel weights. The sample consists of 200 observations from a Gaussian model with $\mu(x) = 1 - x + \exp[-50(x - .5)^2]$. 
2.5 Cubic smoothing splines

The regression analogue of the maximum penalized likelihood problem is the problem of finding a smooth function that best interpolates the observations on $Y$ without fluctuating too wildly.

Let $[a, b]$ be a closed interval that contains the observed values of $X$. Then this problem may be formalized as follows

$$
\min_{m \in C^2[a,b]} Q(m) = \sum_{i=1}^{n} [Y_i - m(X_i)]^2 + \lambda \int_{a}^{b} m''(u)^2 \, du,
$$

(22)

where $C^2[a,b]$ is the class of functions defined on $[a,b]$ that have continuous 1st derivative and integrable 2nd derivative, and $\lambda \geq 0$ is a fixed parameter. The residual sum of squares in the functional $Q$ measures fidelity to the data, whereas the second term penalizes for excessive fluctuations of the estimated CMF.

A solution $\hat{\mu}$ to problem (22) exists, is unique, and can be represented as a natural cubic spline with at most $J = n - 2$ interior knots corresponding to distinct sample values of $X$ (Reinsch 1967). Such a solution is called a cubic smoothing spline.

Larger values of $\lambda$ tend to produce solutions that are smoother, whereas smaller values tend to produce solutions that are more wiggly. In particular:

- If $\lambda \to \infty$, problem (22) reduces to the OLS problem.
- If $\lambda = 0$, there is no penalty for curvature and the solution $\hat{\mu}$ is a twice differentiable function that exactly interpolates the data.
Smoothing splines and penalized LS

Knowledge of the form of the solution to (22) makes it possible to define its smoother matrix. If we consider the representation of \( \hat{\mu} \) as a cubic spline, then the required number of basis functions is \( J + 4 = n - 2 + 4 = n + 2 \). The solution to problem (22) may therefore be written

\[
\hat{\mu}(x) = \sum_{j=1}^{n+2} \alpha_j B_j(x),
\]

where \( \alpha_1, \ldots, \alpha_{n+2} \) are coefficients to be determined and \( B_1(x), \ldots, B_{n+2}(x) \) is a set of twice differentiable basis functions.

Define the \( n \times (n + 2) \) matrix \( \mathbf{B} = [B_{ij}] \), where \( B_{ij} = B_j(X_i) \) is the value of the \( j \)th basis function at the point \( X_i \). Notice that the matrix \( \mathbf{B} \) has at most rank \( n \), so it does not have full column rank. The residual sum of squares in (22) may then be written

\[
\sum_{i=1}^{n} [Y_i - \hat{\mu}(X_i)]^2 = \sum_{i=1}^{n} \left[ Y_i - \sum_{j=1}^{n+2} \alpha_j B_j(X_i) \right]^2 = (\mathbf{Y} - \mathbf{B}\alpha)^\top(\mathbf{Y} - \mathbf{B}\alpha),
\]

where \( \alpha = (\alpha_1, \ldots, \alpha_{n+2}) \), while the penalty term in (22) may be written

\[
\int_{a}^{b} \hat{\mu}''(u)^2 \, du = \int_{a}^{b} \left[ \sum_{j=1}^{n+2} \alpha_j B_j''(u) \right]^2 \, du
= \int_{a}^{b} \left[ \sum_{r=1}^{n+2} \sum_{s=1}^{n+2} \alpha_r \alpha_s B_r''(u)B_s''(u) \right] \, du
= \sum_{r=1}^{n+2} \sum_{s=1}^{n+2} \alpha_r \alpha_s \Omega_{rs} = \alpha^\top \Omega \alpha,
\]

where \( \Omega \) is the \((n+2)\times(n+2)\) matrix with generic element \( \Omega_{rs} = \int_{a}^{b} B_r''(u)B_s''(u) \, du \).

Problem (22) is therefore equivalent to the penalized LS problem

\[
\min_{\alpha} (\mathbf{Y} - \mathbf{B}\alpha)^\top(\mathbf{Y} - \mathbf{B}\alpha) + \lambda \alpha^\top \Omega \alpha. \tag{23}
\]

Thus, the original problem (22) is reduced to the much simpler problem of minimizing a quadratic form in the \((n + 2)\)-dimensional vector \( \alpha \).
The penalized LS estimator

The solution to the penalized LS problem (23) exists and must satisfy the normal equation

\[ 0 = -2B^\top (Y - B\tilde{\alpha}) + 2\lambda\Omega \tilde{\alpha}. \]

If \( B^\top B + \lambda\Omega \) is a nonsingular matrix, then

\[ \tilde{\alpha} = (B^\top B + \lambda\Omega)^{-1}B^\top Y. \]

Because \( \tilde{\alpha} \) formally coincides with a ridge-regression estimate, it also has a nice Bayesian interpretation.

Since \( \hat{\mu} = B\tilde{\alpha} \), the smoother matrix of a cubic smoothing spline is

\[ S = B(B^\top B + \lambda\Omega)^{-1}B^\top, \]

and is symmetric but not idempotent.

See Murphy and Welch (1992) for an independent derivation of the same result.
2.6 Statistical properties of linear smoothers

The smoother matrix of a linear smoother typically depends on the sample values of \(X\) and a parameter (or set of parameters) which regulates the amount of smoothing of the data. To emphasize this dependence, we denote the smoother matrix by \(S_\lambda = [s_{ij}]\), where \(\lambda\) is the \textbf{smoothing parameter}. We adopt the convention that larger values of \(\lambda\) correspond to more smoothing.

Let the data be a sample from the distribution of \((X, Y)\), suppose that the CMF \(\mu(x) = \mathbb{E}(Y \mid X = x)\) and the CVF \(\sigma^2(x) = \text{Var}(Y \mid X = x)\) of \(Y\) are both well defined, and let \(\mu\) and \(\Sigma\) denote, respectively, the \(n\)-vector with generic element \(\mu_i = \mu(X_i)\) and the diagonal \(n \times n\) matrix with generic element \(\sigma_i^2 = \sigma^2(X_i)\).

For a linear smoother \(\hat{\mu} = S_\lambda Y\), with generic element \(\hat{\mu}_i = \hat{\mu}(X_i)\), we have \(\mathbb{E}(\hat{\mu} \mid X) = S_\lambda \mu\). The \textbf{bias} of \(\hat{\mu}\) is therefore

\[
\text{Bias}(\hat{\mu} \mid X) = \mathbb{E}(\hat{\mu} \mid X) - \mu = (S_\lambda - I_n)\mu.
\]

A case when \(\hat{\mu}\) is \textbf{unbiased} is when \(\mu = X\beta\) and \(S_\lambda\) is such that \(S_\lambda X = X\).

The sampling variance of \(\hat{\mu}\) is

\[
\text{Var}(\hat{\mu} \mid X) = S_\lambda \Sigma S_\lambda^T.
\]

Given an estimate \(\hat{\Sigma}\) of the matrix \(\Sigma\), \(\text{Var}(\hat{\mu} \mid X)\) may be estimated by \(S_\lambda \hat{\Sigma} S_\lambda^T\). The result may then be used to construct \textbf{pointwise standard error bands} for the regression estimates. If the smoother is \textbf{approximately unbiased}, then these standard error bands also represent \textbf{pointwise confidence intervals}. An alternative way of constructing confidence intervals is the \textbf{bootstrap}.

If the data are \textbf{conditionally homoskedastic}, that is, \(\sigma^2(x) = \sigma^2\) for all \(x\), then \(\Sigma = \sigma^2 I_n\) may be estimated by \(\hat{\sigma}^2 I_n\), where \(\hat{\sigma}^2 = n^{-1} \sum (Y_i - \hat{\mu}_i)^2\). The sampling variance of \(\hat{\mu}_i\) may then be estimated by \(\hat{\sigma}^2(\sum_{j=1}^n s^2_{ij})\).
Choosing the amount of smoothing

One way of choosing the smoothing parameter optimally is to maximize a global measure of accuracy, conditionally on $X$.

As a global measure of accuracy, consider the average conditional MSE (AMSE)

$$\text{AMSE}(\lambda) = n^{-1} \sum_{i=1}^{n} \mathbb{E}(\hat{\mu}_i - \mu_i)^2$$

$$= n^{-1} \left[ \text{tr}(S_{\lambda} \Sigma S_{\lambda}^\top) + \mu^\top (S_{\lambda} - I_n)^\top (S_{\lambda} - I_n) \mu \right].$$

When the data are conditionally homoskedastic, this reduces to

$$\text{AMSE}(\lambda) = n^{-1} [\sigma_i^2 \text{tr}(S_{\lambda} S_{\lambda}^\top) + \mu^\top (S_{\lambda} - I_n)^\top (S_{\lambda} - I_n) \mu].$$

In general, increasing the amount of smoothing tends to increase the bias and to reduce the sampling variance of a smoother. The AMSE provides a summary of this trade-off.

The following relationship links the AMSE to the average conditional mean squared error of prediction or average predictive risk (APR)

$$n^{-1} \sum_{i=1}^{n} \mathbb{E}(Y_i - \hat{\mu}_i)^2 = n^{-1} \sum_{i=1}^{n} \sigma_i^2 + \text{AMSE}(\lambda).$$

Thus, minimizing the AMSE is equivalent to minimizing the APR.

Since the APR depends on unknown parameters, one way of choosing the smoothing parameter optimally is to minimize an unbiased estimate of the APR.
Cross-validation

An approximately unbiased estimator of the APR is the cross-validation criterion

\[
CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} [Y_i - \hat{\mu}_{(i)}(X_i)]^2,
\]

where \(\hat{\mu}_{(i)}(X_i)\) is the value of the smoother at the point \(X_i\), computed by excluding the \(i\)th observation, and \(Y_i - \hat{\mu}_{(i)}(X_i)\) is the \(i\)th predicted residual. Minimizing \(CV(\lambda)\) represents an automatic method for choosing the smoothing parameter \(\lambda\).

**Example 16** In the OLS case, \(\hat{\mu}_{(i)}(X_i) = \hat{\beta}_n^T X_i\), where

\[
\hat{\beta}_n = (X^T X)^{-1} X^T Y = X_i \hat{U}_i /
\]

is the OLS coefficient computed by excluding the \(i\)th observation, \(\hat{\beta}_n\) is the OLS coefficient computed for the full sample, \(\hat{U}_i = Y_i - \hat{\beta}_n^T X_i\) is the OLS residual and \(h_{ii}\) is the \(i\)th diagonal element of the matrix \(H = X(X^T X)^{-1} X^T\). Since \(Y_i - \hat{\beta}_n^T X_i = \hat{U}_i/(1 - h_{ii})\), the cross-validation criterion is

\[
CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} [Y_i - \hat{\beta}_n^T X_i]^2 = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{\hat{U}_i}{1 - h_{ii}} \right)^2.
\]

Cross-validation requires computing the sequence \(\hat{\mu}_{(1)}(X_1), \ldots, \hat{\mu}_{(n)}(X_n)\). This is very simple if the smoother matrix preserves the constant, that is, \(S_{\lambda}^T i = \iota\). Because \(\sum_{i=1}^{n} s_{ij} = 1\) in this case, \(\hat{\mu}_{(i)}(X_i)\) may be computed by setting the weight of the \(i\)th observation equal to zero and then dividing all the other weights by \(1 - s_{ii}\) in order for them to add up to one. This gives

\[
\hat{\mu}_{(i)}(X_i) = \sum_{j \neq i} s_{ij} Y_j / 1 - s_{ii} = \sum_{j=1}^{n} s_{ij} Y_j / s_{ii} - s_{ii} Y_i / 1 - s_{ii}.
\]

Hence, the \(i\)th predicted residual is \(Y_i - \hat{\mu}_{(i)}(X_i) = \hat{U}_i/(1 - s_{ii})\), where \(\hat{U}_i = Y_i - \hat{\mu}_i\), and the cross-validation criterion becomes

\[
CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{\hat{U}_i}{1 - s_{ii}} \right)^2.
\]
Equivalent kernels and equivalent degrees of freedom

Using the representation (17), it is easy to compare linear smoothers obtained from different methods.

The set of weights \( S_1(x), \ldots, S_n(x) \) assigned to the sample values of \( Y \) in the construction of \( \hat{\mu}(x) \) is called the equivalent kernel evaluated at the point \( x \). The \( i \)th row of the smoother matrix \( S_\lambda \) gives the equivalent kernel evaluated at the point \( X_i \). A comparison of equivalent kernels is therefore a simple and effective way of comparing different linear smoothers.

**Example 17** Given two different linear smoothers, one may compare the different weights,

\[
\{s_{i1}^{(1)}, \ldots, s_{in}^{(1)}\}, \quad \{s_{i1}^{(2)}, \ldots, s_{in}^{(2)}\},
\]

that they give to the sample values of \( Y \) in constructing the estimate of the value \( \mu(X_i) \) of \( \mu(x) \) at a given point \( X_i \). Alternatively, given a smoother and two different points \( X_i \) and \( X_j \), one may compare the different weights,

\[
\{s_{i1}, \ldots, s_{in}\}, \quad \{s_{j1}, \ldots, s_{jn}\},
\]

that the smoother gives to the sample values of \( Y \) in constructing the estimates of \( \mu(X_i) \) and \( \mu(X_j) \).

A synthetic index of the amount of smoothing of the data is the effective number of parameters defined, by analogy with OLS, as

\[
k_* = \text{tr} \, S_\lambda.
\]

The number \( df = n - k_* \) is called equivalent degrees of freedom.

The effective number of parameters is equal to the sum of the eigenvalues of the matrix \( S_\lambda \). If \( S_\lambda \) is symmetric and idempotent (as in the case of regression splines), then

\[
k_* = \text{rank} \, S_\lambda.
\]

The effective number of parameters depends mainly on the value of the smoothing parameter \( \lambda \), while the configuration of the predictors tends to have a much smaller effect. As \( \lambda \) increases, we usually observe a decrease in \( k_* \) and a corresponding increase in \( df \).
Asymptotic properties

We now present some asymptotic properties of regression smoothers. For simplicity, we confine our attention to estimators based on the kernel method.

Given a sample of size $n$ from the distribution of $(X, Y)$, where $X$ is a random $k$-vector and $Y$ is a scalar rv, a NW estimator of the CMF of $Y$ is

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

**Theorem 5** Suppose that the sequence $\{h_n\}$ of bandwidths and the kernel $K: \mathbb{R}^k \to \mathbb{R}$ satisfy:

(i) $h_n \to 0$;

(ii) $nh_n^k \to \infty$;

(iii) $\int K(u) du = 1$, $\int uK(u) du = 0$, and $\int uu^\top K(u) du$ is a finite $k \times k$ matrix.

Then $\hat{\mu}_n(x) \xrightarrow{p} \mu(x)$ for every $x$ in the support of $X$.

**Theorem 6** In addition to the assumptions of Theorem 5, suppose that

$$h_n^2(nh_n^k)^{1/2} \to \lambda,$

where $0 \leq \lambda < \infty$. If the CMF $\mu(x)$ and the CVF $\sigma^2(x)$ are smooth and $X$ has a twice continuously differentiable density $f_X(x)$ then

$$(nh_n^k)^{1/2} [\hat{\mu}_n(x) - \mu(x)] \Rightarrow \mathcal{N} \left( \frac{b(x)}{f_X(x)}, \frac{\sigma^2(x)}{f_X(x)} \int K(u)^2 du \right)$$

for every $x$ in the support of $X$, where

$$b(x) = \lim_{n \to \infty} h_n^{-(k+2)} \mathbb{E} \left\{ [\mu(X_i) - \mu(x)] K \left( \frac{x - X_i}{h_n} \right) \right\}.$$

Further, $(nh_n^k)^{1/2}[\hat{\mu}_n(x) - \mu(x)]$ and $(nh_n^k)^{1/2}[\hat{\mu}_n(x') - \mu(x')]$ are asymptotically independent for $x \neq x'$. 

79
Remarks

- $\hat{\mu}_n(x)$ is \textit{asymptotically biased} for $\mu(x)$, unless $h_n$ is chosen such that $\lambda = 0$. In this case, however, the convergence of $\hat{\mu}_n(x)$ to its limiting distribution is \textit{slower} than in the case when $\lambda > 0$.

- The convergence rate of $\hat{\mu}_n(x)$ is inversely related to the number $k$ of covariates, reflecting the \textit{curse-of-dimensionality} problem. For $k$ fixed, the maximal rate is achieved when $h_n = cn^{-1/(k+4)}$, for some $c > 0$. In this case, $\lambda = c^{(k+4)/2}$ and therefore

$$
n^{2/(k+4)} [\hat{\mu}_n(x) - \mu(x)] \Rightarrow \mathcal{N} \left( \frac{c^2 b(x)}{f(x)}, \frac{\sigma^2(x)}{c^k f(x)} \int K(u)^2 du \right).$$

Even when there is a single covariate ($k = 1$), the \textit{fastest} convergence rate is only equal to $n^{-2/5}$ and is \textit{lower} than the rate of $n^{-1/2}$ achieved by a regular estimator of $\mu(x)$ in a parametric context.

These problems have generated two active lines of research:

- One seeks ways of \textit{eliminating the asymptotic bias} of $\hat{\mu}_n(x)$ while retaining the maximal convergence rate of $n^{-2/(k+4)}$. One possibility (Schucany & Sommers 1977, Bierens 1987) is to consider a weighted average of two estimators with different bandwidths chosen such that the resulting estimator is asymptotically centered at $\mu(x)$.

- The other seeks ways of \textit{improving the speed of convergence} of $\hat{\mu}_n(x)$. For example, if the kernel $K$ is chosen such that $\int uu^\top K(u) du = 0$ (allowing therefore for \textit{negative} values of $K$), then the optimal convergence rate can be shown to be $n^{-3/(k+6)}$. More generally, if one chooses a kernel with \textit{zero moments up to order} $m$, then the \textit{optimal bandwidth} becomes $h_n = cn^{-1/(k+2m)}$, and the \textit{maximal convergence rate} becomes $n^{-m/(k+2m)}$ which, as $m \to \infty$, tends to the rate $n^{-1/2}$ typical of a parametric estimator.
Tests of parametric models

Nonparametric regression may be used to test the validity of a parametric model. For this purpose one may employ both informal methods, especially graphical ones, and more formal tests.

To illustrate, let \([X, Y]\) be a sample of size \(n\) from the distribution of \((X, Y)\), let \(\mu(x)\) denote the CMF of \(Y\), and consider testing for the goodness-of-fit of a linear regression model. The null hypothesis \(H_0: \mu(x) = \alpha + \beta x\) specifies the \(\mu(x)\) as linear in \(x\), whereas the alternative hypothesis specifies \(\mu(x)\) as a smooth nonlinear function of \(x\).

Let \(\hat{U} = (I_n - H) Y\), with \(H = X (X^\top X)^{-1} X^\top\), be the OLS residual vector and let \(\tilde{U} = (I_n - S_\lambda) Y\) be the residual vector associated with some other linear smoother \(\hat{\mu} = S_\lambda Y\). Because \(H_0\) is nested within the alternative, an \(F\)-type test rejects \(H_0\) for large values of the statistic

\[
  F = \frac{\hat{U}^\top \hat{U} - \tilde{U}^\top \tilde{U}}{\tilde{U}^\top \tilde{U}} = \frac{Y^\top M_* Y - Y^\top M_* Y}{Y^\top M_* Y},
\]

where \(M_* = (I_n - S_\lambda)^\top (I_n - S_\lambda)\).

For general types of linear smoother, the distribution of \(F\) under \(H_0\) is difficult to derive for it depends on the unknown parameters \(\alpha\) and \(\beta\).
Connections with the Durbin-Watson test

Azzalini and Bowman (1993) suggest replacing $H_0$ by the hypothesis that $\mathbb{E} \hat{U} = 0$, conditional on $X$. The alternative is now that $\mathbb{E} \hat{U}$ is a smooth function of the covariates. This leads to a test that rejects $H_0$ for large values of the statistic

$$F_* = \frac{\hat{U}^\top \hat{U} - \hat{U}^\top M_* \hat{U}}{\hat{U}^\top M_* \hat{U}}.$$  

It can be shown that $F_*$ depends only on the data and the value of the smoothing parameter $\lambda$. Rejecting $H_0$ for large values of $F_*$ is equivalent to rejecting for small values of the statistic

$$\xi = \frac{1}{F_* + 1} = \frac{\hat{U}^\top M_* \hat{U}}{\hat{U}^\top \hat{U}},$$

which has the same form as the Durbin–Watson test statistic.

In order to determine the critical value of a test based on $F_*$, one needs to compute probabilities such as

$$\Pr\{F_* > c\} = \Pr\{\hat{U}^\top \hat{U} - \hat{U}^\top M_* \hat{U} > c, \hat{U}^\top M_* \hat{U}\} = \Pr\{\hat{U}^\top A \hat{U} > 0\},$$

where $A = I_n - (1 + c)M_*$. Hence, the problem reduces to calculating the distribution of the quadratic form $\hat{U}^\top A \hat{U}$ under the null hypothesis. Azzalini and Bowman (1993) offer some suggestions.
2.7 Average derivatives

Suppose that the CMF is sufficiently smooth and write it as \( \mu(x) = c(x)/f_X(x) \), where \( c(x) = \int y f(x,y) \, dy \). Then its slope is

\[
\mu'(x) = \frac{c'(x)f_X(x) - c(x)f_X'(x)}{f_X(x)^2} = \frac{c'(x) - \mu(x)f_X'(x)}{f_X(x)},
\]

and its average slope or average derivative is

\[
\delta = \mathbb{E} \mu'(X) = \int \mu'(x) f_X(x) \, dx,
\]

(24)

where the expectation is with respect to the marginal distribution of \( X \). The average derivative provides a convenient summary of the CMF through a finite dimensional parameter (a scalar parameter if \( X \) is unidimensional).

Notice that, if \( \mu(x) \) is linear, that is, \( \mu(x) = \alpha + \beta^T x \), then

\[
\mu'(x) = \beta = \mathbb{E} \mu'(X).
\]

This is no longer true if \( \mu(x) \) is nonlinear.

Under regularity conditions, integrating by parts gives the following equivalent representation of the average derivative

\[
\delta = \mathbb{E} s_X(X) \mu(X),
\]

(25)

where \( s_X(x) = -f_X'(x)/f_X(x) \) is the score of the marginal density of \( X \).

A third representation follows from the fact that, by the Law of Iterated Expectations,

\[
\delta = \mathbb{E} s_X(X) \mu(X) = \mathbb{E}[s_X(X) \mathbb{E}(Y \mid X)] = \mathbb{E} s_X(X) Y,
\]

(26)

where the last expectation is with respect to the joint distribution of \( (X,Y) \).

These three equivalent representations provide alternative approaches to estimating the average derivative.
Estimation of the average derivative

If the kernel $K$ is differentiable, an analog estimators of the average derivative $\delta$ based on (24) is

$$\hat{\delta}_1 = n^{-1} \sum_{i=1}^{n} D_i \hat{\mu}'(X_i),$$

where $D_i = 1\{\hat{f}_X(X_i) > c_n\}$ for some trimming constant $c_n$ that goes to zero as $n \to \infty$. The lower bound on $\hat{f}_X$ is introduced to avoid erratic behavior where $X$ is sparse, so the value of $\hat{f}_X(x)$ is very small.

An analog estimator based on (25) is

$$\hat{\delta}_2 = n^{-1} \sum_{i=1}^{n} D_i \hat{s}(X_i) \hat{\mu}(X_i),$$

where $\hat{s}_X(x) = -\hat{f}'_X(x)/\hat{f}_X(x)$ is the estimated score.

Another analog estimator, based on (26), is

$$\hat{\delta}_3 = n^{-1} \sum_{i=1}^{n} D_i \hat{s}_X(X_i) Y_i.$$
2.8 Methods for high-dimensional data

We now review a few nonparametric approaches that try to overcome some of the problems encountered in trying to extend univariate nonparametric methods to multivariate situations.

Project pursuit regression estimation

The projection pursuit (PP) method was originally introduced in the regression context by Friedman and Stuetzle (1981). Specifically, given a rv $Y$ and a random $k$-vector $X$, they proposed to approximate the CMF of $Y$ by a function of the form

$$m(x) = \alpha + \sum_{j=1}^{p} m_j(\beta_j^T x),$$

(27)

where $\beta_j = (\beta_{1j}, \ldots, \beta_{kj})$ is a vector with unit norm that specifies a particular direction in $\mathbb{R}^k$, $\beta_j^T x = \sum_{i=1}^{k} \beta_{ij} x_i$ is a linear combination or projection of the variables in $X$, and $m_1, \ldots, m_p$ are unknown smooth univariate “ridge functions” to be determined empirically.

Remarks:

- By representing the CMF as a sum of arbitrary functions of linear combinations of the elements of $x$, PP regression allows for quite complex patterns of interaction among the covariates.

- Linear regression models may be viewed as special cases corresponding to $p = 1$ and $m_1(u) = u$, single-index models as special cases corresponding to $p = 1$, and sample selection models as special cases corresponding to $p = 2$.

- The curse-of-dimensionality problem is bypassed by using linear projections and nonparametric estimates of univariate ridge functions.
The PP algorithm

Algorithm 2

1. Given estimates \( \hat{\beta}_j \) of the projection directions and estimates \( \hat{m}_j \) of the first \( h-1 \) ridge functions in (27), compute the approximation errors

\[
 r_i = \tilde{Y}_i - \sum_{j=1}^{h-1} \hat{m}_j(\hat{\beta}_j^\top X_i), \quad i = 1, \ldots, n,
\]

where \( \tilde{Y}_i = Y_i - \bar{Y} \).

2. Given a vector \( b \in \mathbb{R}^k \) such that \( ||b|| = 1 \), construct a linear smoother \( \hat{m}(b^\top X_i) \) based on the errors \( r_i \) and compute the RSS

\[
 S(b) = \sum_{i=1}^{n} \left[ r_i - \hat{m}(b^\top X_i) \right]^2.
\]

3. Determine \( b_* \) and the function \( \hat{m}_* \) for which \( S(b) \) is minimized.

4. Insert \( b_* \) and the function \( \hat{m}_* \) as the \( h \)th terms into (27).

5. Iterate (1)–(4) until the decrease in the RSS becomes negligible.

As stressed by Huber (1985), PP regression “emerges as the most powerful method yet invented to lift one-dimensional statistical techniques to higher dimensions”. However, it is not without problems:

- **Interpretability** of the individual terms of (27) is difficult when \( p > 1 \).
- PP regression has problems in dealing with **highly nonlinear structures**. In fact, there exist functions that cannot be represented as a finite sum of ridge functions (e.g. \( \mu(x) = \exp(x_1 x_2) \)).
- The **sampling theory** of PP regression is still lacking.
- The **choice of the amount of smoothing** in constructing the nonparametric estimates of the ridge functions is delicate.
- PP regression tends to be **computationally expensive**.

We now discuss an approach that may be regarded as a simplification of PP regression.
Additive regression

An important property of the linear regression model

\[ m(x) = \alpha + \sum_{j=1}^{k} \beta_j x_j \]

is additivity. If there is no functional relation between the elements of \( x \), this property makes it possible to “separate” the effect of the different covariates and to interpret \( \beta_j \) as the (constant) partial derivative of \( \mu(x) \) with respect to the \( j \)th covariate.

An additive regression model approximates the CMF of \( Y \) by a function of the form

\[ m(x) = \alpha + \sum_{j=1}^{k} m_j(x_j), \]

where \( m_1, \ldots, m_k \) are univariate smooth functions, one for each covariate. The linear regression model corresponds to \( m_j(u) = \beta_j u \) for all \( j \) or, equivalently, \( m'_j(u) = \beta_j \).

Remarks:

- The additive model (28) is a special case of the PP regression model (27) corresponding to \( p = k \) and \( \beta_j = e_j \), the \( j \)th unit vector.
- Because \( m'_j(x_j) = \partial m(x)/\partial x_j \) if there is no functional relationship between the elements of \( X \), additive regression retains the interpretability of the effect of the individual covariates.
- Interactions between categorical and continuous ones may be modeled by estimating separate versions of the model for each value of the categorical variable.
- Interactions between continuous covariates may be modelled by creating compound variables, such as products of pairs.
Characterizing additive regression models

An additive regression model may be characterized as the solution to an approximation problem. In turn, this leads to a general iterative procedure for estimating this class of models.

This way of proceeding is very similar in spirit to obtaining OLS as the sample analog of the normal equations that characterize the best linear approximation to an arbitrary CMF.

Before stating the main result, it is worth recalling a few definitions and results. We refer to Luenberger (1969) for details.
Banach and Hilbert spaces

A normed vector space $\mathcal{X}$ is a vector space on which is defined a function $\|x\|: \mathcal{X} \to \mathbb{R}_+$, called the norm of $x$, that satisfies the following properties:

(i) $\|x\| = 0$ for all $x = 0$;

(ii) (triangle inequality) $\|x + y\| \leq \|x\| + \|y\|$ for all $y, x \in \mathcal{X}$;

(iii) $\|\alpha x\| = |\alpha|\|x\|$ for all real numbers $\alpha$.

In a normed vector space, an infinite sequence of vectors $\{x_n\}$ is said to converge to a vector $x$, written $x_n \to x$, if the sequence $\{\|x - x_n\|\}$ of real numbers converges to zero. An infinite sequence $\{x_n\}$ is called a Cauchy sequence if $\|x_n - x_m\| \to 0$ for $n, m \to \infty$. A normed vector space $\mathcal{X}$ is said to be complete if every Cauchy sequence in $\mathcal{X}$ has a limit in $\mathcal{X}$. A complete vector space is also called a Banach space.

Given a vector space $\mathcal{X}$, an inner product is a function $\langle x | y \rangle: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that satisfies the following properties:

(i) $\langle x | y \rangle = \langle y | x \rangle$;

(ii) $\langle x + y | z \rangle = \langle x | z \rangle + \langle y | z \rangle$;

(iii) $\langle \alpha x | z \rangle = \alpha \langle x | z \rangle$ for all scalar $\alpha$;

(iv) $\langle x | x \rangle \geq 0$, and $\langle x | x \rangle = 0$ if and only if $x = 0$.

Two vectors $x, y$ in a vector space $\mathcal{X}$ with inner product $\langle \cdot | \cdot \rangle$ are said to be orthogonal, written $x \perp y$, if $\langle x | y \rangle = 0$. A vector $x \in \mathcal{X}$ is said to be orthogonal to a subset $\mathcal{M}$ of $\mathcal{X}$, written $x \perp \mathcal{M}$, if $x \perp m$ for all $m \in \mathcal{M}$.

On a vector space with inner product $\langle \cdot | \cdot \rangle$, the function $\|x\| = \sqrt{\langle x | x \rangle}$ is a norm. A complete vector space on which an inner product is defined is called a Hilbert space. Hence, a Hilbert space is a Banach space equipped with an inner product which induces a norm.
Example 18 An important example of Hilbert space is the space $\mathcal{H}$ of all random variables with finite variance defined on a probability space $(\Omega, \mathcal{A}, P)$. In this case, given two random variables $Y, X \in \mathcal{H}$, we define the inner product

$$\langle X \mid Y \rangle = \text{Cov}(X, Y).$$

Viewed as elements of $\mathcal{H}$, two uncorrelated random variables are therefore orthogonal. The inner product thus defined induces the norm

$$\|X\| = [\text{Cov}(X, X)]^{1/2} = [\text{Var}(X)]^{1/2}.$$
The Classical Projection Theorem

**Theorem 7** Let $\mathcal{M}$ be a closed subspace of a Hilbert space $\mathcal{H}$. For any vector $x \in \mathcal{H}$, there exists a unique vector $m_* \in \mathcal{M}$ such that $\|x - m_*\| \leq \|x - m\|$ for all $m \in \mathcal{M}$. Further, a necessary and sufficient condition for $m_* \in \mathcal{M}$ to be such a vector is that $x - m_* \perp \mathcal{M}$.

The vector $m_* \in \mathcal{M}$ such that $x - m_* \perp \mathcal{M}$ is called the **orthogonal projection of** $x$ **onto** $\mathcal{M}$.

The next result is just a restatement of the Classical Projection Theorem.

**Theorem 8** Let $\mathcal{M}$ be a closed subspace of a Hilbert space $\mathcal{H}$. Let $x \in \mathcal{H}$ and let $\mathcal{V}$ be the affine subspace $x + \mathcal{M}$. Then, there exists a unique vector $x_*$ in $\mathcal{V}$ with minimum norm. Further, $x_* \perp \mathcal{M}$.
The main result

**Theorem 9** Let $Z = (X_1, \ldots, X_k, Y)$ be a random $(k + 1)$-vector with finite second moments, let $\mathcal{H}$ be the Hilbert space of zero-mean functions of $Z$, with the inner product defined as $\langle \psi | \phi \rangle = E \psi(Z)\phi(Z)$, and let $\mu(x) = \mu(x_1, \ldots, x_n) = E(Y \mid X_1 = x_1, \ldots, X_k = x_k)$. For any $j = 1, \ldots, k$, let $M_j$ be the subspace of zero-mean square integrable functions of $X_j$ and let $M^a = M_1 + \cdots + M_k$ be the subspace of zero-mean square integrable additive functions of $X = (X_1, \ldots, X_k)$. Then a solution to the following problem

$$\min_{m \in M^a} E [\mu(X) - m(X)]^2$$

(29)

exists, is unique, and has the form

$$m_*(X) = \sum_{j=1}^k m_j(X_j),$$

where $m_1, \ldots, m_k$ are univariate functions such that

$$m_j(X_j) = E \left[ Y - \sum_{h \neq j} m_h(X_h) \mid X_j \right], \quad j = 1, \ldots, k.$$ (30)
Proof

We only sketch the proof. The details can be found in Stone (1985). If we endow the Hilbert space $\mathcal{H}$ with the norm $\|\psi\| = \left[ E \psi(Z)^2 \right]^{1/2}$, then problem (29) is equivalent to the minimum norm problem

$$\min_{m \in \mathcal{M}^a} \|Y - m\|^2.$$

Under some technical conditions, $\mathcal{M}^a$ and $\mathcal{M}_1, \ldots, \mathcal{M}_k$ are closed subspaces of $\mathcal{H}$. Thus, by the Classical Projection Theorem, there exists a unique vector $m_* \in \mathcal{M}^a$ that solves (29). Moreover, $m_*$ is characterized by the orthogonality condition $Y - m_* \perp \mathcal{M}^a$ or equivalently, since $\mathcal{M}^a = \mathcal{M}_1 + \cdots + \mathcal{M}_k$, by the set of orthogonality conditions

$$Y - m_* \perp \mathcal{M}_j, \quad j = 1, \ldots, k.$$

Because $m_* \in \mathcal{M}^a$, it must be of the form $m_*(X) = \sum_{j=1}^k m_j(X_j)$. Further, because the conditional expectation $E(\cdot | X_j)$ is an orthogonal projection onto $\mathcal{M}_j$, we must have

$$E \left[ Y - \sum_{j=1}^k m_j(X_j) \mid X_j \right] = 0, \quad j = 1, \ldots, k,$$

that is,

$$m_j(X_j) = E \left[ Y - \sum_{l \neq j} m_h(X_h) \mid X_j \right], \quad j = 1, \ldots, k.$$
Estimating additive regression models

Let $S_j$ denote a smoother matrix for univariate smoothing on the $j$th covariate. Because the sample analogue of the conditional mean $\mu_j(x) = \mathbb{E}(Y | X_j = x)$ is the univariate smoother $S_jY$, the analogy principle suggests the following equation system as the sample counterpart of (30)

$$m_j = S_j \left( \tilde{Y} - \sum_{h \neq j} m_h \right), \quad j = 1, \ldots, k,$$

where $m_j$ is the $n$-vector with generic element $m_j(X_{ij})$ and $\tilde{Y}$ is the $n$-vector with generic element $Y_i - \bar{Y}$.

This results in the following system of $nk \times nk$ linear equations

$$\begin{bmatrix}
I_n & S_1 & S_1 & \cdots & S_1 \\
S_2 & I_n & S_2 & \cdots & S_2 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
S_k & S_k & S_k & \cdots & I_n
\end{bmatrix}
\begin{bmatrix}
m_1 \\
m_2 \\
\vdots \\
m_k
\end{bmatrix}
= \begin{bmatrix}
S_1 \tilde{Y} \\
S_2 \tilde{Y} \\
\vdots \\
S_k \tilde{Y}
\end{bmatrix}.$$  \hfill (31)

If $(\tilde{m}_1, \ldots, \tilde{m}_k)$ is a solution to the above system and we put $\hat{\alpha} = \bar{Y}$, then an estimate of $\mu = (\mu(X_1), \ldots, \mu(X_n))$ is

$$\hat{\mu} = \hat{\alpha} \iota_n + \sum_{j=1}^{k} \tilde{m}_j.$$
The backfitting algorithm

One way of solving system (31) is the Gauss–Seidel method. This method solves iteratively for each vector $m_j$ from the relationship

$$m_j = S_j \left( \bar{Y} - \sum_{h \neq j} m_h \right),$$

using the latest values of $\{m_h, h \neq j\}$ at each step. The process is repeated for $j = 1, \ldots, k, 1, \ldots, k, \ldots$, until convergence. The result is the backfitting algorithm:

**Algorithm 3**

1. **Compute the univariate smoother matrices** $S_1, \ldots, S_k$.
2. **Initialize the algorithm** by setting $m_j = m_j^{(0)}$, $j = 1, \ldots, k$.
3. **Cycle for** $j = 1, \ldots, k$, $m_j = S_j(\bar{Y} - \sum_{h \neq j} m_h)$, where the smoother matrix $S_j$ for univariate smoothing on the $j$th covariate is applied to the residual vector $\bar{Y} - \sum_{h \neq j} m_h$ obtained from the previous step.
4. **Iterate step (3)** until the changes in the individual functions become negligible.

The smoother matrices $S_j$ can be computed using any univariate linear smoother and do not change through the backfitting algorithm.

To ensure that $\bar{Y} - \sum_{h \neq j} m_h$ has mean zero at every stage, $S_j$ may be replaced by the centered smoother matrix $S_j(I_n - n^{-1}t_n t_n')$.

A possible starting point for the algorithm is the fitted value of $Y$ from an OLS regression of $Y$ on $X$. In this case $m_j^{(0)} = X_j \hat{\beta}_j$, where $X_j$ and $\hat{\beta}_j$ denote respectively the $j$th column of $X$ and the $j$th element of the OLS coefficient vector.

It can be shown that convergence of the backfitting algorithm is guaranteed in the case of smoothing splines or when the chosen linear smoother is a projection operator, that is, defined by a symmetric idempotent smoother matrix.
Partially linear models

A partially linear model is of the form

\[ \mathbb{E}(Y \mid X, W) = m_1(X) + m_2(W), \]

where \( m_1(x) = \beta^\top x \) is an unknown linear function and \( m_2(w) \) is an unknown univariate smooth function.

For simplicity (and without loss of generality), let \( X \) be a scalar rv's and write the model as

\[ Y = \beta X + m_2(W) + U, \quad (32) \]

where \( \mathbb{E}(U \mid X, W) = 0 \).

If \( \mathbb{E}Xm_2(W) = 0 \), then \( \beta \) may be estimated consistently and efficiently by an OLS regression of \( Y \) on \( X \). This is typically the case when \( X \) has mean zero and is independent of \( W \). In general, however, \( X \) and \( m_2(W) \) are correlated, so a regression of \( Y \) on \( X \) gives inconsistent estimates of \( \beta \).

Notice however that

\[
\mathbb{E}(Y \mid W) = \mathbb{E}[\mathbb{E}(Y \mid X, W) \mid W] \\
= \mathbb{E}[\beta X + m_2(W) \mid W] \\
= \beta \mathbb{E}(X \mid W) + m_2(W). 
\]

Subtracting this expression from (32) gives

\[ Y - \mathbb{E}(Y \mid W) = \beta [X - \mathbb{E}(X \mid W)] + U. \]

If \( \hat{Y} = P Y \) is a linear regression smoother of \( Y \) on \( W \) and \( \hat{X} = S X \) is a linear regression smoother of \( X \) on \( W \), then \( \beta \) may simply be estimated by an OLS regression of \( Y - \hat{Y} = (I_n - P) Y \) on \( X - \hat{X} = (I_n - S) X \). The resulting estimate of \( \beta \) is

\[
\hat{\beta} = [X^\top (I_n - S)^\top (I_n - S) X ]^{-1} X^\top (I_n - S)^\top (I_n - P) Y. 
\]

This method, proposed by Robinson (1988), may be regarded as a generalization of the double residual regression (or Frisch-Waugh-Lovell) method for partitioned linear regressions.
The backfitting algorithm for partially linear models

Now consider applying the backfitting algorithm. Because in this case $m_1$ is linear, two different smoothers may be employed:

- An **OLS smoother**, with smoother matrix $S_1 = X(X^\top X)^{-1}X^\top$, which produces estimates of the form $\hat{m}_1 = X\hat{\beta}$, where $\hat{\beta}$ is the desired estimate of $\beta$.

- A **nonparametric smoother** with smoother matrix $S_2$. Applied to the vector $W = (W_1, \ldots, W_n)$, this smoother produces estimates of the vector $m_2 = (m_2(W_1), \ldots, m_2(W_n))$ of the form $\hat{m}_2 = S_2 Y$.

In this case, the steps of the backfitting algorithm are

\[
\hat{m}_1 = S_1(Y - \hat{m}_2), \\
\hat{m}_2 = S_2(Y - \hat{m}_1).
\]

Premultiplying the first of these two equations by $X^\top$ and substituting the expression for $\hat{m}_2$ from the second equation gives

\[
X^\top \hat{m}_1 = X^\top (Y - \hat{m}_2) = X^\top (I_n - S_2) Y + S_2 \hat{m}_1,
\]

where we used the fact that $X^\top S_1 = X^\top$. Because $\hat{m}_1 = X\hat{\beta}$, we have

\[
X^\top \hat{X}\hat{\beta} = X^\top (I_n - S_2) Y + X^\top S_2 X\hat{\beta}
\]

or, equivalently,

\[
X^\top (I_n - S_2) Y = X^\top (I_n - S_2) X\hat{\beta}.
\]

Solving for $\hat{\beta}$ we then get

\[
\hat{\beta} = [X^\top (I_n - S_2) X]^{-1} X^\top (I_n - S_2) Y.
\]

In this case, no iteration is necessary.

This method may be regarded as another way of generalizing the **double-residual regression method** for partitioned linear regressions. In fact, if $S_2 = W(W^\top W)^{-1}W^\top$, then the two methods coincide.

The method proposed by Robinson (1988) coincides with the backfitting algorithm if $S_2 = P = S$ is a **projection matrix** (i.e., symmetric and idempotent).
2.9 Stata commands

We now briefly review the commands for linear nonparametric regression estimation available in Stata, version 12.

These include the \texttt{lowess} command for lowess, the \texttt{lpoly} command for kernel-weighted local polynomial smoothing, and the \texttt{mkspline} command for constructing linear or natural cubic splines. All three commands consider the case of a single regressor.

The package \texttt{bspline} in Newson (2012) allows one to estimate univariate and multivariate splines.
The `lowess` command

This is a computationally intensive command. For example, running `lowess` on 1,000 observations require performing 1,000 regressions each involving a number of observations equal to $n$ times the span. Since Stata does not take advantage of the recursive formulae for OLS or WLS, this may take a long time on a slow computer.

The basic syntax is:

```
lowess yvar xvar [if] [in] [weight] [, options]
```

where `options` includes:

- **mean**: running-mean smooth (default is running-line least squares).
- **noweight**: suppresses weighted regressions (default is tricube weighting function).
- **bwidth(#)**: span of the data (default is 0.80). Centered subsets containing a fraction `bwidth()` of the observations are used for calculating smoothed values for each point in the data except for end points, where smaller uncentered subsets are used.
- **logit**: transforms dependent variable to logits and adjust smoothed mean to equal mean of `yvar`.
- **adjust**: adjusts smoothed mean to equal mean of `yvar`.
- **generate(newvar)**: creates `newvar` containing smoothed values of `yvar`.
- **nograph**: suppresses graph. This option is often used with the `generate()` option.
The `lpoly` command

The basic syntax is:

```
lpoly yvar xvar [if] [in] [weight] [, options]
```

where `options` includes:

- **kernel(expression)**: specifies kernel function. The available kernels include `epanechnikov` (default), `biweight`, `cosine`, `gaussian`, `parzen`, `rectangle`, and `triangle`.

- **bwidth(#)**: specifies the half-width of the kernel. If `bwidth()` is not specified, a “rule-of-thumb” bandwidth is calculated and used. A local variable bandwidth may be specified in `varname`, in conjunction with an explicit smoothing grid using the `at()` option.

- **degree(#)**: specifies the degree of the polynomial smooth. The default is `degree(0)` corresponding to a running mean. This is equivalent to a NW estimator.

- **generate([newvar] x newvar_s)**: stores smoothing grid in `newvar_x` and smoothed points in `newvar_s`. If `at()` is not specified, then both `newvar_x` and `newvar_s` must be specified. Otherwise, only `newvar_s` is to be specified.

- **n(#)**: obtain the smooth at # points. The default is `min(n, 50)`.

- **at(varname)**: specifies a variable that contains the values at which the smooth should be calculated. By default, the smoothing is done on an equally spaced grid but one can use `at()` to instead perform the smoothing at the observed $X$, for example.

- **nograph**: suppresses graph. This option is often used with the `generate()` option.

- **noscatter**: suppresses scatterplot only. This option is useful when the number of resulting points would be so large as to clutter the graph.

- **pwidth(#)**: specifies pilot bandwidth for standard error calculation.
The mkspline command

In the first syntax, it creates a set of \( k \) variables containing a linear spline of \( oldvar \) with knots at the specified \( k - 1 \) knots:

\[
\text{mkspline } \text{newvar}_1 \ #1 \ \text{newvar}_2 \ #2 \ [\ldots] \ \text{newvar}_k = oldvar \ [if] \ [in] \\
[, \ \text{marginal displayknots}]
\]

where

- \text{marginal} specifies that the new variables be constructed so that, when used in estimation, the coefficients represent the change in the slope from the preceding interval. The default is to construct the variables so that, when used in estimation, the coefficients measure the slopes for the interval.

- \text{displayknots} displays the values of the knots that were used in creating the linear or restricted cubic spline.

In the second syntax, \text{mkspline} creates \# variables named \text{stubname}1, \ldots, \text{stubname}\#, containing a linear spline of \( oldvar \):

\[
\text{mkspline } \text{stubname } # = oldvar \ [if] \ [in] \ [weight] \ [, \ \text{marginal pctile displayknots}]
\]

where \text{pctile} specifies that the knots be placed at percentiles of \( oldvar \). The default are equally spaced knots over the range of \( oldvar \).

In the third syntax, \text{mkspline} creates variables containing a natural cubic spline of \( oldvar \):

\[
\text{mkspline } \text{stubname} = oldvar \ [if] \ [in] \ [weight], \ \text{cubic} \ [\text{nknots(\#)} \ \text{knots(numlist)} \ \text{displayknots}]
\]

where \text{nknots(\#)} specifies the number of knots that are to be used for the natural cubic spline. This number must be between 3 and 7 unless the knot locations are specified using \text{knots(\)}. The default number of knots is 5.
3 Distribution function and quantile function estimators

The distribution function and the quantile function are equivalent ways of characterizing the probability distribution of a univariate rv Z.

3.1 Distribution functions and quantile functions

The distribution function (df) of Z is a function $F: \mathbb{R} \to [0, 1]$ defined by

$$F(z) = \Pr\{Z \leq z\}, \quad z \in \mathbb{R}.$$

Sometimes, instead of working with the df, it is convenient to work with the function

$$S(z) = 1 - F(z) = \Pr\{Z > z\}, \quad z \in \mathbb{R},$$

called the survivor function.

A df $F$ is called absolutely continuous if it is differentiable almost everywhere and satisfies

$$F(z) = \int_{-\infty}^{z} F'(t) \, dt, \quad -\infty < z < \infty.$$

Any function $f$ such that $F(z) = \int_{-\infty}^{z} f(t) \, dt$ is called a density for $F$. Any such density must agree with $F'$ except possibly on a set with measure zero. If $f$ is continuous at $z_0$, then $f(z_0) = F'(z_0)$.

Sometimes the df is defined in terms of its density. For example, the density of the $\mathcal{N}(0, 1)$ distribution is

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right).$$

The associated df is defined by the integral

$$\Phi(z) = \int_{-\infty}^{z} \phi(u) \, du,$$

but this integral does not have a closed form expression.
Properties of the distribution function

- $F$ is **non decreasing, right-continuous**, and satisfies $F(-\infty) = 0$ and $F(\infty) = 1$.

- For any $z$,
  \[ \Pr\{Z = z\} = F(z+) - F(z-). \]
  where $F(z+) = \lim_{h \to 0} F(z + h)$ and $F(z-) = \lim_{h \to 0} F(z - h)$.

- If $F$ is **continuous at** $z$, then
  \[ F(z+) = F(z-) = F(z), \]
  and so $\Pr\{Z = z\} = 0$.

- The rv $F(Z)$ is distributed as $U(0, 1)$.

- Given $c < d$, the rv $X = \min(d, \max(c, Z))$ is called a **censored** version of $Z$. The df’s of $Z$ and $X$ agree over the interval $(c, d)$. 


Quantiles

For any $p \in (0, 1)$, a $p$th quantile of a rv $Z$ is a number $Q_p \in \mathbb{R}$ such that

$$\Pr\{Z < Q_p\} \leq p \quad \text{and} \quad \Pr\{Z > Q_p\} \leq 1 - p, \quad (33)$$

or equivalently

$$\Pr\{Z \leq Q_p\} \geq p \quad \text{and} \quad \Pr\{Z \geq Q_p\} \geq 1 - p.$$

Thus, a $p$th quantile satisfies

$$\Pr\{Z < Q_p\} \leq p \leq \Pr\{Z \leq Q_p\},$$

or equivalently

$$\Pr\{Z > Q_p\} \leq 1 - p \leq \Pr\{Z \geq Q_p\}.$$

The set of $p$th quantiles of $Z$ is a closed interval of the real line. A $p$th quantile is unique if the df of $Z$ is strictly increasing at $Q_p$, in which case $F(Q_p) = p$.

A quantile corresponding to $p = .50$ is called a median of $Z$ and will be denoted by $\zeta$. Quantiles corresponding to $p = .10, .25, .75$ and $.90$ are called, respectively, lower deciles, lower quartiles, upper quartiles and upper deciles.
Quantiles as solutions to a minimization problem

Recall that, if a rv $Z$ has finite mean, then its mean solves the problem
\[ \min_{c \in \mathbb{R}} E(Z - c)^2. \]
A $p$th quantile of $Z$ may also be defined implicitly as a solution to a particular minimization problem.

Define the convex function
\[ \ell_p(v) = [p - 1\{v < 0\}] v = \begin{cases} (1 - p)|v|, & \text{if } v < 0, \\ pv, & \text{if } v \geq 0, \end{cases} \]
called the asymmetric absolute loss function. Also define the expected loss $L(c) = E \ell_p(Z - c)$, where
\[ \ell_p(Z - c) = [p - 1\{Z < c\}] (Z - c) = \begin{cases} (1 - p)(c - Z), & \text{if } Z < c, \\ p(Z - c), & \text{if } Z \geq c. \end{cases} \]

Then we have the following:

**Theorem 10** A $p$th quantile of $Z$ solves the problem
\[ \min_{c \in \mathbb{R}} L(c). \] (34)

**Proof.** Consider for simplicity the case when $Z$ is a continuous rv with density $f$ and finite mean $\mu$. Since $L(c)$ is convex, it is enough to show that $Q_p$ is a root of the equation $L'(c) = 0$. First notice that
\[
L(c) = p \int_{-\infty}^{\infty} (z - c) f(z) \, dz - \int_{-\infty}^{c} (z - c) f(z) \, dz \\
= p(\mu - c) - \int_{-\infty}^{c} z f(z) \, dz + cF(c).
\]
Hence,
\[
L'(c) = -p - cf(c) + F(c) + cf(c) = -p + F(c).
\]
Since $Q_p$ satisfies $F(Q_p) = p$, it is a root of $L'(Q_p) = 0$. \qed

If $p = .50$, then $\ell_p(v) = |v|/2$ is symmetric and is proportional to the familiar symmetric absolute loss function. Further, the corresponding minimum expected loss $(E |Z - \zeta|)/2$ is proportional to the mean absolute deviation of $Z$ from its median.
Figure 12: The asymmetric absolute loss function $\ell_p(v)$. 

![Graph showing the asymmetric absolute loss function for different values of p (0.25, 0.50, 0.75).]
The quantile function

The quantile function ($qf$) of a rv $Z$ is a function $Q:(0, 1) \to \mathbb{R}$ defined by

$$Q(p) = \inf \{ z \in \mathbb{R}: F(z) \geq p \}. \quad (35)$$

If $Q_p$ is unique, then $Q(p) = Q_p$. While the df $F$ maps $\mathbb{R}$ into $[0, 1]$, the qf $Q$ maps $(0, 1)$ into $\mathbb{R}$.

Properties of the quantile function

- $Q$ is non decreasing and left-continuous.
- $Q(F(z)) \leq z$ and $F(Q(p)) \geq p$. Thus,

$$F(z) \geq p \iff z \geq Q(p).$$

This justifies calling $Q$ the inverse of $F$ and using the notation $Q(p) = F^{-1}(p)$. Also,

$$F(z) = \sup \{ p \in [0, 1]: Q(p) \leq z \}. \quad (36)$$

- $Q$ is continuous iff $F$ is strictly increasing.
- $Q$ is strictly increasing iff $F$ is continuous, in which case

$$Q(p) = \inf \{ z \in \mathbb{R}: F(z) = p \}, \quad F(Q(p)) = p.$$

Further, if $F$ is continuous, then $Z$ has the same distribution as $Q(U)$, where $U \sim \mathcal{U}(0, 1)$ (quantile transformation).
Derivation of the qf from the df

If the df $F$ is **continuous** and can be expressed in closed form, then the qf is easily obtained by solving the equation $F(z) = p$.

**Example 19** If $Z \sim \mathcal{U}(0, 1)$, then $F(z) = z$, so $Q(p) = p$. □

**Example 20** If $Z \sim \mathcal{E}(\lambda)$, $\lambda > 0$, then

$$F(z) = 1 - \exp(-\lambda z),$$

so

$$Q(p) = -\frac{\ln(1 - p)}{\lambda}.$$ □
Table 1: Distribution function $F(z)$ and quantile function $Q(p)$ of selected distributions.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$F(z)$</th>
<th>$Q(p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cauchy</td>
<td>$\frac{1}{2} + \frac{1}{\pi} \arctan z$</td>
<td>$\tan[\pi(p - \frac{1}{2})]$</td>
</tr>
<tr>
<td>Chi-square(1)</td>
<td>$2\Phi(\sqrt{z}) - 1$</td>
<td>$[\Phi^{-1}((p + 1)/2)]^2$</td>
</tr>
<tr>
<td>Exponential</td>
<td>$1 - e^{-z}$</td>
<td>$-\ln(1 - p)$</td>
</tr>
<tr>
<td>Gaussian</td>
<td>$\Phi(z)$</td>
<td>$\Phi^{-1}(p)$</td>
</tr>
<tr>
<td>Gumbel</td>
<td>$1 - \exp(-e^z)$</td>
<td>$\ln[-\ln(1 - p)]$</td>
</tr>
<tr>
<td>Laplace</td>
<td>$\frac{1}{2}e^z, z &lt; 0$</td>
<td>$\ln 2p, p &lt; \frac{1}{2}$</td>
</tr>
<tr>
<td></td>
<td>$1 - \frac{1}{2}e^z, z \geq 0$</td>
<td>$-\ln 2(1 - p), p \geq \frac{1}{2}$</td>
</tr>
<tr>
<td>Logistic</td>
<td>$\frac{e^z}{1 + e^z}$</td>
<td>$\ln \frac{p}{1 - p}$</td>
</tr>
<tr>
<td>Log-normal</td>
<td>$\Phi(\ln z)$</td>
<td>$\exp[\Phi^{-1}(p)]$</td>
</tr>
<tr>
<td>Pareto</td>
<td>$1 - (\alpha/z)^\beta$</td>
<td>$\alpha(1 - p)^{-1/\beta}$</td>
</tr>
<tr>
<td>Uniform</td>
<td>$z$</td>
<td>$p$</td>
</tr>
<tr>
<td>Weibull</td>
<td>$1 - \exp(-\gamma z^\alpha)$</td>
<td>$[-\frac{1}{\alpha} \ln(1 - p)]^{1/\alpha}$</td>
</tr>
</tbody>
</table>
Figure 13: Quantile functions of selected distributions.
Other properties of the quantile function

- If the first two moments of $Z$ exist, then
  \[
  
  \begin{align*}
  E Z &= \int_0^1 Q(p) \, dp, \\
  \text{Var } Z &= \int_0^1 Q(p)^2 \, dp - \left[ \int_0^1 Q(p) \, dp \right]^2.
  \end{align*}
  \]

- If $Z$ has a continuous and strictly positive density $f$ in a neighborhood of $Q(p)$, then differentiating the identity $F(Q(p)) = p$ gives $f(Q(p)) Q'(p) = 1$. Hence,
  \[
  Q'(p) = \frac{1}{f(Q(p))}.
  \]
  The slope $Q'$ is known as the sparsity function or quantile-density function, whereas the composition $f(Q(\cdot))$ is known as the density-quantile function. The sparsity function is well defined whenever the density is strictly positive, and is strictly positive whenever the density is bounded.

- Differentiating the identity $\ln Q'(p) = -\ln f(Q(p))$ gives
  \[
  \frac{Q''(p)}{Q'(p)} = -\frac{f'(Q(p))}{f(Q(p)) Q'(p)}.
  \]
  Hence, the score function $s = -f'/f$ satisfies
  \[
  s(Q(p)) = \frac{Q''(p)}{[Q'(p)]^2} = -\frac{d}{dp} \frac{1}{Q'(p)} = -\frac{d}{dp} f(Q(p)).
  \]

- If the function $g$ is a monotonically increasing and left-continuous, then the qf of the rv $g(Z)$ is equal to $g(Q(\cdot))$. In particular, the qf of $g(Z) = \alpha + \beta Z$, with $\beta \geq 0$, is equal to $g(Q(p)) = \alpha + \beta Q(p)$.

- Given $c < d$, let
  \[
  X = \min(d, \max(c, Z))
  \]
  be a censored version of $Z$. It then follows from the previous result that the qf’s of $Z$ and $X$ agree at all $p$ such that $F(c) < p < F(d)$. In particular, $Z$ and $X$ have the same median provided that $F(c) < 1/2 < F(d)$. This is generally not true for the mean.
Moment-based summaries of a distribution

Center, spread and symmetry of a distribution are intuitive but somewhat vague concepts. Kurtosis is not even intuitive.

If $Z$ is a rv with finite 4th moment, then it is customary to define:

- **Center** in terms of the mean $\mu = E(Z)$.
- **Spread** in terms of the variance $\sigma^2 = E(Z - \mu)^2$ or the standard deviation $\sigma = [E(Z - \mu)^2]^{1/2}$.
- **Symmetry** in terms of the 3rd moment $E(Z - \mu)^3$.
- **Kurtosis** in terms of the 4th moment $E(Z - \mu)^4$.

If $Z \sim \mathcal{N}(\mu, \sigma^2)$, then $E(Z - \mu)^3 = 0$ and $E(Z - \mu)^4 = 3\sigma^4$.

The drawbacks with these measures are:

- they may not exist (e.g., for all four measures to exist, a $t$ distribution needs at least 5 degrees of freedom);
- their interpretation is not easy (especially for kurtosis);
- they lack robustness.
Quantile-based summaries of a distribution

An alternative is to use quantile-based summaries of a distribution. Thus, we may alternatively define:

- **Center** in terms of the median $\zeta = Q(.50)$.
- **Spread** in terms of either the **interquartile range**
  \[
  \text{IQR} = Q(.75) - Q(.25),
  \]
  or the **interdecile range**
  \[
  \text{IDR} = Q(.90) - Q(.10).
  \]
  The IQR (IDR) is the interval of values which contains the central 50% (80%) of the probability mass of $Z$.
- **Symmetry** in terms of the **difference**
  \[
  [Q(.90) - Q(.50)] - [Q(.50) - Q(.10)],
  \]
  or the **ratios**
  \[
  \frac{Q(.90) - Q(.50)}{Q(.50) - Q(.10)}
  \]
  or
  \[
  \frac{[Q(.90) - Q(.50)] - [Q(.50) - Q(.10)]}{Q(.90) - Q(.10)},
  \]
  also known as **Kelley’s measure**. One may also consider analogous measures with $Q(.90)$ and $Q(.10)$ replaced by $Q(.75)$ and $Q(.25)$.
- **Kurtosis**, or more precisely **tail weight**, in terms of either the **difference** IDR - IQR or the **ratio** IDR/IQR.
Functionals of the distribution function or the quantile function

Sometimes, the parameter of interest is a particular functional of either the df or the qf. We illustrate this with two examples.

**Example 21** Given a rv $Z$ with finite mean, the $\alpha$-level tail conditional mean of $Z$ is formally defined as

$$
\tau(\alpha) = E(Z \mid Z \leq Q(\alpha)) = \frac{1}{\alpha} \int_{-\infty}^{Q(\alpha)} z \, dF(z), \quad 0 < \alpha < 1.
$$

In financial applications, $Z$ represents the return on a given asset during a period, while $Q(\alpha)$ and $\tau(\alpha)$ are **measures of financial risk**:

- The quantile $Q(\alpha)$, also called the **Value-at-Risk (VaR)** at level $\alpha$, gives a lower bound on the loss made in the worst $\alpha$ percent of the periods.

- The tail conditional mean $\tau(\alpha)$, also called the $\alpha$-level expected shortfall, represents the expected value of a loss (negative return) that exceeds the VaR.

Typical values of $\alpha$ are .01, .05 or .10.

If $F$ is continuous and strictly increasing, a change of variable from $F(z)$ to $p$ gives the equivalent representation

$$
\tau(\alpha) = \frac{1}{\alpha} \int_{F(-\infty)}^{F(Q(\alpha))} F^{-1}(p) \, dp = \frac{1}{\alpha} \int_{0}^{\alpha} Q(p) \, dp. \tag{37}
$$

This second representation is particularly convenient when the quantiles of $Z$ have a **closed form expression**. □
Example 22 Let $Z$ be a non-negative rv with finite nonzero mean $\mu$. The Lorenz curve of $Z$ is formally defined as

$$L(\alpha) = \frac{1}{\mu} \int_0^\alpha Q(p) \, dp, \quad 0 < \alpha < 1.$$ 

The Lorenz curve is commonly used in economics to describe the distribution of income and is associated with the Gini inequality index.

From (37), the following relationship links the Lorenz curve and the tail conditional mean of $Z$

$$L(\alpha) = \frac{\alpha}{\mu} \tau(\alpha).$$

The generalized Lorenz curve (Shorrocks 1983) is the Lorenz curve scaled up by the mean, and is equal to

$$GL(\alpha) = \int_0^\alpha Q(p) \, dp = \alpha \tau(\alpha), \quad 0 < \alpha < 1.$$ 

If the non-negative rv $Z$ represents individual income, then $GL(\alpha)$ simply cumulates individual incomes up to the $\alpha$th quantile. \(\square\)
3.2 The empirical distribution function

Let $Z_1, \ldots, Z_n$ be a sample from the distribution of a rv $Z$, and consider the problem of estimating the df $F$ of $Z$ when $F$ is not restricted to a known parametric family.

The key is the fact that, for any $z$,

$$F(z) = \Pr\{Z \leq z\} = \text{E}1\{Z \leq z\},$$

that is, $F(z)$ is just the mean of the Bernoulli rv $1\{Z \leq z\}$. This suggests estimating $F(z)$ by its sample counterpart

$$\hat{F}(z) = n^{-1} \sum_{i=1}^{n} 1\{Z_i \leq z\},$$

(38)

namely the fraction of sample points such that $Z_i \leq z$. Viewed as a function defined on $\mathbb{R}$, $\hat{F}$ is called the empirical distribution function (edf).

In fact, $\hat{F}$ is the df of a discrete probability measure, called the empirical measure, which assigns to a set $A$ a probability equal to the fraction of sample points contained in $A$. In particular, each distinct sample point receives probability equal to $1/n$, whereas each sample point repeated $m \leq n$ times receives probability equal to $m/n$.

Suppose that all observations are distinct, and let

$$Z_{[1]} < Z_{[2]} < \cdots < Z_{[n]}$$

be the sample order statistics. Then the edf may also be written

$$\hat{F}(z) = \begin{cases} 0, & \text{if } z < Z_{[1]}, \\ i/n, & \text{if } Z_{[i]} \leq z < Z_{[i+1]}, \\ 1 & \text{if } z \geq Z_{[n]}. \end{cases} \quad i = 1, \ldots, n - 1,$$

This shows that the edf contains all the information carried by the sample, except the order in which the observations have been drawn or are arranged.
Figure 14: Empirical distribution function $\hat{F}$ for a sample of 100 observations from a $\mathcal{N}(0, 1)$ distribution.
The Riemann-Stieltjes integral

In what follows we shall sometimes use the notation of the Riemann-Stieltjes integral (see e.g. Apostol 1974, Chpt. 7).

Given a df $F$ and an integrable function $g$, define

$$ E_F g(Z) = \int g(z) dF(z) = \int g(z) f(z) \, dz $$

if $F$ is continuous with probability density function $f(z)$, and

$$ E_F g(Z) = \int g(z) dF(z) = \sum_j g(z_j) f(z_j) $$

if $F$ is discrete with probability mass function $f(z_j)$.

Because the edf $\hat{F}$ is discrete, with probability mass function that assigns probability mass $1/n$ to each distinct sample point, we have

$$ E_{\hat{F}} g(Z) = \int g(z) d\hat{F}(z) = n^{-1} \sum_{i=1}^n g(Z_i). $$
The empirical process

Let \( Z_1, \ldots, Z_n \) be a sample from the distribution of a rv \( Z \) with df \( F \), let

\[
\theta = T(F) = \int g(z) \, dF(z)
\]

be the population parameter of interest, let \( \hat{F} \) be the edf, and let

\[
\hat{\theta} = T(\hat{F}) = \int g(z) \, d\hat{F}(z) = n^{-1} \sum_{i=1}^{n} g(Z_i)
\]

be a “plug-in” or bootstrap estimator of \( \theta \).

The estimation error associated with \( \hat{\theta} \) is

\[
\hat{\theta} - \theta = \int g(z) \, d\hat{F}(z) - \int g(z) \, dF(z) = \int g(z) \, d[\hat{F}(z) - F(z)].
\]

Thus, the estimation error ultimately depends on the difference between the edf \( \hat{F} \) and the population df \( F \).

In asymptotic analysis, attention typically focuses on the rescaled estimation error

\[
\sqrt{n} (\hat{\theta} - \theta) = \int g(z) \, dp_n(z),
\]

where

\[
p_n(z) = \sqrt{n} [\hat{F}(z) - F(z)].
\]

Notice that \( p_n(z) \) is a function defined on \( \mathbb{R} \) for any given sample, but is a rv for any given \( z \). The collection

\[
p_n = \{p_n(z), \, z \in \mathbb{R}\}
\]

of all such rv’s is therefore a stochastic process with index set \( \mathbb{R} \). This process is called the empirical process.
Figure 15: Realization of the empirical process $p_n$ for a sample of $n = 100$ observations from a $\mathcal{N}(0, 1)$ distribution.
Finite sample properties

The key is again the fact that $\hat{F}(z)$ is the average of $n$ iid rv’s $1\{Z_i \leq z\}$ that have a common Bernoulli distribution with parameter $F(z)$.

**Theorem 11** If $Z_1, \ldots, Z_n$ is a sample from a distribution with df $F$, then for any $z \in \mathbb{R}$:

(i) $n\hat{F}(z) \sim \text{Bi}(n, F(z));$

(ii) $E \hat{F}(z) = F(z);$

(iii) $\text{Var} \hat{F}(z) = n^{-1}F(z)[1 - F(z)];$

(iv) $\text{Cov}[\hat{F}(z), \hat{F}(z')] = n^{-1}F(z)[1 - F(z')]$ for any $z' \geq z$.

**Remarks:**

- For any $z$, $\hat{F}(z)$ is an unbiased for $F(z)$.
- Because of the correlation between $\hat{F}(z)$ and $\hat{F}(z')$, care is needed in drawing inference about the shape of $F$ from small samples.
- As $n \to \infty$, the sampling variance of $\hat{F}(z)$ vanishes, so $\hat{F}(z)$ is consistent for $F(z)$. Further, as $n \to \infty$, the correlation between $\hat{F}(z)$ and $\hat{F}(z')$ also vanishes.
- Being the average of $n$ iid rv’s with finite variance, $\hat{F}(z)$ also satisfies the standard CLT and is therefore asymptotically normal.
Asymptotic properties

We now consider the properties of a sequence \( \{\hat{F}_n(z)\} \) of edf’s indexed by the sample size \( n \). As in the case of density estimation, we distinguish between local properties (valid at a finite set of \( z \) values) and global properties.

Given \( J \) distinct points

\[-\infty < z_1 < \cdots < z_J < \infty,\]

let \( F \) be the \( J \)-vector consisting of \( F_1, \ldots, F_J \), with \( F_j = F(z_j) \), and let \( \hat{F}_n = (\hat{F}_{n1}, \ldots, \hat{F}_{nJ}) \) be the sample analogue of \( F \), with \( \hat{F}_{nj} = \hat{F}_n(z_j) \). The sampling distribution of the \( J \)-vector

\[ p_n = \sqrt{n}(\hat{F}_n - F) \]

corresponds to the \( J \)-dimensional distribution of the empirical process \( p_n \).

As \( n \to \infty \) we have:

- \( \hat{F}_n \xrightarrow{as} F \),
- \( p_n \Rightarrow \mathcal{N}_J(0, \Sigma) \), where \( \Sigma \) is the \( J \times J \) matrix with generic element

\[ \sigma_{jk} = \min(F_j, F_k) - F_j F_k, \quad j, k = 1, \ldots, J. \]

These two properties may be viewed as special cases of the following two results.

- **Dvoretzky-Kiefer-Wolfowitz (DKW) inequality**: For all \( n \) and any \( \epsilon > 0 \),

\[ \Pr \left\{ \sup_{z \in \mathbb{R}} |\hat{F}(z) - F(z)| \geq \epsilon \right\} \leq 2e^{-2n\epsilon^2}. \]  
  (39)

- The empirical process \( \{p_n(Q(u)), u \in (0, 1)\} \) converges weakly to the **Brownian bridge** or tied-down Brownian motion, that is, the unique Gaussian process \( U \) with continuous sample paths on \([0, 1] \) such that \( \mathbb{E} U(t) = 0 \) and \( \text{Cov}[U(s), U(t)] = \min(s, t) - st \), with \( 0 < s, t < 1 \).
The Glivenko-Cantelli theorem

As a measure of distance between the edf $\hat{F}_n$ and the population df $F$ consider the Kolmogorov-Smirnov statistic

$$D_n = \sup_{-\infty < z < \infty} |\hat{F}_n(z) - F(z)|.$$

Since $\hat{F}_n$ depends on the data, $D_n$ is itself a rv.

It would be nice if, as $n \to \infty$, $D_n \to 0$ in some appropriate sense. In fact, one can shown that $D_n \xrightarrow{as} 0$ as $n \to \infty$, that is,

$$\Pr\{ \lim_{n \to \infty} D_n = 0 \} = 1$$

or, equivalently, the event that $D_n$ does not converge to zero as $n \to \infty$ occurs with zero probability under repeated sampling.

This fundamental result, known as the Glivenko-Cantelli theorem, implies that the entire probabilistic structure of $Z$ can almost certainly be uncovered from the sample data provided that $n$ is large enough. Vapnik (1995) calls this “the most important result in the foundation of statistics”.

Figure 16: Empirical distribution function $\hat{F}$ for a sample of 1,000 observations from a $\mathcal{N}(0, 1)$ distribution.
Multivariate and conditional edf

The definition of edf and its sampling properties are easily generalized to the multivariate case.

If \( Z_1, \ldots, Z_n \) is a sample from the distribution of a random \( m \)-vector \( Z \), then the \( m \)-variate edf is defined as

\[
\hat{F}(z_1, \ldots, z_m) = n^{-1} \sum_{i=1}^{n} 1\{Z_{i1} \leq z_1, \ldots, Z_{im} \leq z_m\}
\]

\[
= n^{-1} \sum_{i=1}^{n} 1\{Z_{i1} \leq z_1\} \cdots 1\{Z_{im} \leq z_m\}.
\]

**Example 23** If \( m = 2 \) and \( Z_i = (X_i, Y_i) \), then the bivariate edf is defined as

\[
\hat{F}(x, y) = n^{-1} \sum_{i=1}^{n} 1\{X_i \leq x, Y_i \leq y\}
\]

\[
= n^{-1} \sum_{i=1}^{n} 1\{X_i \leq x\} 1\{Y_i \leq y\}.
\]

Although conceptually straightforward, multivariate df estimates suffer of the curse of dimensionality problem and are difficult to represent graphically unless \( m \) is small, say equal to 2 or 3.
3.3 The empirical quantile function

We now consider the problem of estimating the qf of a rv $Z$ given a sample $Z_1, \ldots, Z_n$ from its distribution.

Sample quantiles

A sample analogue of (33) is called a sample quantile. Thus, for any $p \in (0, 1)$, a $p$th sample quantile is a number $\hat{Q}_p \in \mathbb{R}$ such that

\[
-n^{-1} \sum_{i=1}^{n} 1 \{ Z_i < \hat{Q}_p \} \leq p \quad \text{and} \quad n^{-1} \sum_{i=1}^{n} 1 \{ Z_i > \hat{Q}_p \} \leq 1 - p,
\]

that is, about $np$ of the observations are smaller than $\hat{Q}_p$ and about $n(1 - p)$ are greater. Equivalently, we have the condition

\[
\sum_{i=1}^{n} 1 \{ Z_i < \hat{Q}_p \} \leq np \leq \sum_{i=1}^{n} 1 \{ Z_i \leq \hat{Q}_p \},
\]

that is, $np$ cannot be smaller than the number of observations strictly less than $\hat{Q}_p$ and cannot be greater than the number of observations less than or equal to $\hat{Q}_p$.

A $p$th sample quantile is unique if $np$ is not an integer, and lies in a closed interval defined by two adjacent order statistics if $np$ is an integer.

Example 24 A sample quantile corresponding to $p = 1/2$ is called a sample median and will be denoted by $\hat{\zeta}$. In fact, if $Z_{[1]} \leq \cdots \leq Z_{[n]}$ are the sample order statistics, then:

- when $n$ is odd ($n/2$ is not an integer), $\hat{\zeta} = Z_{[k]}$, where $k = (n + 1)/2$,
- when $n$ is even ($n/2$ is an integer), $\hat{\zeta}$ is any point in the closed interval $[Z_{[k]}, Z_{[k+1]}]$, where $k = n/2$.
Asymmetric LAD

It can be shown that a $p$th sample quantile is also a solution to the sample analogue of (34), namely

$$\min_{c \in \mathbb{R}} \mathbb{E}_F \ell_p(Z - c) = n^{-1} \sum_{i=1}^n \ell_p(Z_i - c),$$

(41)

where $\ell_p(v) = [p - 1\{v < 0\}] v$ is the asymmetric absolute loss function. Problem (41) is called an asymmetric least absolute deviations (ALAD) problem. Notice that solving problem (41) avoids the need of sorting.

Since the function $\ell_p(v)$ is convex but not differentiable at $v = 0$, the ALAD objective function $L(c) = n^{-1} \sum_{i=1}^n \ell_p(Z_i - c)$ is itself convex but not differentiable at all $z$ that correspond to a sample value. However, we can always define its subgradients

$$L'(c-) = \lim_{h \to 0} \frac{L(c - h) - L(c)}{h} = p - \frac{1}{n} \sum_{i=1}^n 1\{Z_i < c\}$$

(actually its gradient when $c$ is not equal to a sample value) and

$$L'(c+) = \lim_{h \to 0} \frac{L(c + h) - L(c)}{h} = \frac{1}{n} \sum_{i=1}^n 1\{Z_i \leq c\} - p,$$

with $L'(c-) \approx -L'(c+)$ in large samples. We can then characterize $\hat{Q}_p$ as a solution to (41) by the fact that both $L'(\hat{Q}_p-) \geq 0$ and $L'(\hat{Q}_p+) \geq 0$ must hold, that is,

$$\frac{1}{n} \sum_{i=1}^n 1\{Z_i < \hat{Q}_p\} \leq p \leq \frac{1}{n} \sum_{i=1}^n 1\{Z_i \leq \hat{Q}_p\},$$

which is just a restatement of condition (40).
Example 25 A sample median \( \hat{\zeta} \) is a solution to problem (41) for \( p = 1/2 \) and minimizes the least absolute deviations (LAD) objective function

\[
L(c) = \frac{1}{2n} \sum_{i=1}^{n} |Z_i - c|.
\]

Notice that minimizing \( L(c) \) over \( \mathbb{R} \) is equivalent to solving

\[
\min_{c \in \mathbb{R}} \sum_{i=1}^{n} |Z_i - c|.
\]

In this case, the subgradients are

\[
L'(c-) = \lim_{h \to 0} \frac{L(c-h) - L(c)}{h} = \frac{1}{2} - \frac{1}{n} \sum_{i=1}^{n} 1\{Z_i < c\}
\]

and

\[
L'(c+) = \lim_{h \to 0} \frac{L(c+h) - L(c)}{h} = \frac{1}{n} \sum_{i=1}^{n} 1\{Z_i \leq c\} - \frac{1}{2},
\]

so the following condition must hold

\[
\frac{1}{n} \sum_{i=1}^{n} 1\{Z_i < \hat{\zeta}\} \leq \frac{1}{2} \leq \frac{1}{n} \sum_{i=1}^{n} 1\{Z_i \leq \hat{\zeta}\},
\]

that is, at most half of the observations are less than \( \hat{\zeta} \) and at least half are greater or equal to \( \hat{\zeta} \). \( \square \)
Figure 17: LAD objective function for different sample sizes.
Figure 18: Subgradients $L'(c^-)$ and $L'(c^+)$ of the LAD objective function for $n = 4$. 
Figure 19: Subgradients $L'(c-)$ and $L'(c+)$ of the LAD objective function for $n = 5$. 
The empirical quantile function

The empirical quantile function (eqf) is a real-valued function defined on (0, 1) by

\[ \hat{Q}(p) = \inf \{ z \in \mathbb{R} : \hat{F}(z) \geq p \} . \]

Thus, the eqf is just the sample analogue of problem (35) and satisfies all properties of a qf. In particular, it is the left-continuous inverse of the right-continuous edf \( \hat{F} \).

Notice that \( \hat{Q}(p) = \hat{Q}_p \) when \( np \) is not an integer. Further, there is a close association between the values of the eqf and the sample order statistics, for

\[ \hat{Q}(p) = Z_{[i]}, \quad \text{for} \quad \frac{i - 1}{n} < p \leq \frac{i}{n}, \quad i = 1, \ldots, n. \]

The rescaled difference

\[ q_n(p) = \sqrt{n} [\hat{Q}(p) - Q(p)], \]

is a function on (0, 1) for any given sample, and is a rv for any given \( p \). The collection

\[ q_n = \{ q_n(p), 0 < p < 1 \} \]

of all such rv’s is therefore a stochastic process with index set (0, 1). This process is called the sample quantile process.
Figure 20: Empirical quantile function $\hat{Q}$ for a sample of 100 observations from a $\mathcal{N}(0,1)$ distribution.
Figure 21: Empirical quantile function $\hat{Q}$ for a sample of 1000 observations from a $\mathcal{N}(0,1)$ distribution.
Figure 22: Realization of the sample quantile process $q_n$ for a sample of $n = 100$ observations from a $\mathcal{N}(0, 1)$ distribution.
Finite sample properties

Because sample quantiles essentially coincide with sample order statistics, we present a result on the df and the density of the sample order statistic \( Z[i] \), corresponding to the sample quantile \( \hat{Q}(i/n) \).

**Theorem 12** Let \( Z_1, \ldots, Z_n \) be a sample from a distribution with df \( F \). Then the df of \( Z[i] \), \( 1 \leq i \leq n \), is

\[
G(z) = \Pr\{Z[i] \leq z\} = \sum_{k=i}^{n} \binom{n}{k} F(z)^k [1 - F(z)]^{n-k}.
\]

If \( F \) has a density \( f = F' \), then the density of \( Z[i] \) is

\[
g(z) = G'(z) = n \binom{n-1}{i-1} f(z) F(z)^{i-1} [1 - F(z)]^{n-i}.
\]

Interest often centers not just on a single quantile, but on a finite number of them. This is the case when we seek a detailed description of the shape of a probability distribution, or we are interested in constructing some L-estimate, that is, linear combinations of sample quantiles, such as the sample IQR, the sample IDR, or a trimmed mean.

Sample quantiles corresponding to different values of \( p \) are dependent. Hence, from a practical point of view, what matters is their joint distribution. The next result presents the joint density of two sample order statistics, \( Z[i] \) and \( Z[j] \), corresponding respectively to the sample quantiles \( \hat{Q}(i/n) \) and \( \hat{Q}(j/n) \).

**Theorem 13** Let \( Z_1, \ldots, Z_n \) be a sample from a continuous distribution with df \( F \) and density \( f \). Then the joint density of \( Z[i] \) and \( Z[j] \), \( 1 \leq i < j \leq n \), is

\[
g(u, v) = \frac{n!}{(i-1)!(j-1-i)!(n-j)!} f(u) f(v) F(u)^{i-1} \times [F(v) - F(u)]^{j-1-i} [1 - F(v)]^{n-j}.
\]
Asymptotic properties

Since the exact sampling properties of a finite set of sample quantiles are somewhat complicated, we now consider their asymptotic properties.

Let \( \hat{Q}_n \) be a solution to (41) (to simplify notation we henceforth drop the reference to \( p \)) and again assume that \( Z_1, \ldots, Z_n \) is a sample from a continuous distribution with df \( F \) and finite strictly positive density \( f \). Under this assumption, \( Q = F^{-1}(p) \) is the unique solution to (34) for any \( 0 < p < 1 \).

Further, the ALAD objective function

\[
L_n(c) = n^{-1} \sum_{i=1}^{n} \ell_p(Y_i - c)
\]

can be shown to converge almost surely as \( n \to \infty \), uniformly on \( \mathbb{R} \), to the population objective function

\[
L(c) = \mathbb{E} \ell_p(Y - c).
\]

Thus, as \( n \to \infty \), \( \hat{Q}_n \xrightarrow{as} Q \) for any \( 0 < p < 1 \).
Asymptotic normality of a sample quantile

Because the ALAD objective function \( L_n(c) = n^{-1} \sum_{i=1}^{n} \ell_p(Z_i - c) \) is not differentiable, we cannot derive asymptotic normality of a sample quantile \( \hat{Q}_n \) (where the index \( p \) has been dropped for simplicity) by taking a 2nd-order Taylor expansion of \( L_n \).

There are various approaches to the problem. A simple approach is based on the fact that the subgradient of \( L_n \),

\[
L'_n(c) = n^{-1} \sum_{i=1}^{n} 1\{Z_i \leq c\} - p,
\]

is increasing in \( c \) (Figures 18–19). Since \( L'_n(\hat{Q}_n) \geq 0 \), this implies that \( \hat{Q}_n \leq c \) iff \( L'_n(c) > 0 \).

Next notice that, for any \( t \in \mathbb{R} \), \( T_n = \sqrt{n} (\hat{Q}_n - Q) \leq t \) iff \( \hat{Q}_n \leq Q + t/\sqrt{n} \). Thus, putting \( c = Q + t/\sqrt{n} \), the df of the rescaled and recentered difference \( T_n \) is

\[
\Pr\{T_n \leq t\} = \Pr\left\{ L'_n \left( Q + \frac{t}{\sqrt{n}} \right) > 0 \right\} = \Pr\{\bar{W}_n > 0\},
\]

with \( \bar{W}_n = n^{-1} \sum_{i=1}^{n} W_i \), where \( W_i = 1\{Z_i \leq Q + t/\sqrt{n}\} - p \) is a binary rv that takes values \( 1 - p \) and \( -p \) with probabilities \( F(Q + t/\sqrt{n}) \) and \( 1 - F(Q + t/\sqrt{n}) \) respectively. Thus, \( \bar{W}_n \) is an average of iid rv’s with finite mean and variance. As \( n \to \infty \),

\[
\text{E} W_i = F \left( Q + \frac{t}{\sqrt{n}} \right) - p = F \left( Q + \frac{t}{\sqrt{n}} \right) - F(Q) \to \frac{f(Q) t}{\sqrt{n}},
\]

while

\[
\text{Var} W_i = F \left( Q + \frac{t}{\sqrt{n}} \right) \left[ 1 - F \left( Q + \frac{t}{\sqrt{n}} \right) \right] \to p(1 - p).
\]

If \( f(Q) > 0 \) then, by the De Moivre-Laplace Central Limit Theorem,

\[
\Pr\{\bar{W}_n > 0\} = \Pr\left\{ \frac{W_n - f(Q) t/\sqrt{n}}{\sqrt{p(1 - p) / n}} > -\frac{t}{\omega} \right\} \to \Phi \left( \frac{t}{\omega} \right)
\]

as \( n \to \infty \), where \( \omega = \sqrt{p(1 - p) / f(Q)} \). Hence, \( T_n \Rightarrow \mathcal{N}(0, \omega^2) \) as \( n \to \infty \).
Asymptotic joint normality of sample quantiles

Given \( J \) distinct values

\[
0 < p_1 < \ldots < p_J < 1,
\]

let \( Q \) be the \( J \times 1 \) vector consisting of \( Q_1, \ldots, Q_J \), with \( Q_j = Q(p_j) \), and let \( \hat{Q}_n = (\hat{Q}_{n1}, \ldots, \hat{Q}_{nJ}) \), with \( \hat{Q}_{nj} = \hat{Q}_n(p_j) \), be the sample analogue of \( Q \). The sampling distribution of the vector

\[
q_n = \sqrt{n} (\hat{Q}_n - Q)
\]
corresponds to the \textit{J-dimensional distribution} of the sample quantile process \( q_n \).

Extending the previous argument one can show that, if \( F \) possesses a continuous density \( f \) which is positive and finite at \( Q_1, \ldots, Q_J \), then

\[
q_n \Rightarrow \mathcal{N}_{J}(0, \Omega)
\]

as \( n \to \infty \), where \( \Omega \) is the \( J \times J \) matrix with generic element

\[
\omega_{rs} = \frac{\min(p_r, p_s) - p_r p_s}{f(Q_r) f(Q_s)}, \quad r, s = 1, \ldots, J.
\]  (42)

Csörgő (1983) and Shorack and Wellner (1986) discuss an analogue of the DKW inequality and weak convergence of the sample quantile process to the Brownian bridge with finite dimensional distributions equal to the asymptotic distribution of \( q_n \).
Bahadur representation

Bahadur (1966) and Kiefer (1967) established the close link between the sample quantile process \( q_n(p) = \sqrt{n} [\hat{Q}_n(p) - Q(p)] \) and the empirical process \( p_n(z) = \sqrt{n} [\hat{F}_n(z) - F(z)] \).

They showed that, under regularity conditions, with probability tending to one as \( n \to \infty \),

\[
\hat{Q}_n(p) - Q(p) = \frac{1}{n} \sum_{i=1}^{n} \frac{p - 1\{Z_i \leq Q(p)\}}{f(Q(p))} + R_n,
\]

uniformly for \( p \in [\epsilon, 1 - \epsilon] \) and some \( \epsilon > 0 \), where the remainder \( R_n \) is \( O_p(n^{-3/4}(\ln \ln n)^{3/4}) \). Multiplying both sides by \( \sqrt{n} \) gives

\[
\sqrt{n} [\hat{Q}_n(p) - Q(p)] = \frac{\sqrt{n} [F(Q(p)) - \hat{F}_n(Q(p))]}{f(Q(p))} + R'_n,
\]

where the remainder \( R'_n \) is \( O_p(n^{-1/4}(\ln \ln n)^{3/4}) \). Thus,

\[
q_n(p) = -\frac{p_n(Q(p))}{f(Q(p))} + o_p(1).
\]

Multivariate normality of sample quantiles follows immediately from this representation.
Estimating the asymptotic variance of sample quantiles

From (42), a \( p \)th sample quantile has asymptotic variance

\[
\text{AV}(\hat{Q}_n(p)) = \frac{p(1-p)}{f(Q(p))^2}.
\]

In particular, a sample median \( \hat{\zeta}_n \) has asymptotic variance

\[
\text{AV}(\hat{\zeta}_n) = \frac{1}{4f(\zeta)^2},
\]

where \( \zeta = Q(.50) \).

In practice, the asymptotic variance of sample quantiles has to be estimated from the data. If \( \hat{f} \) is a nonparametric estimate of \( f \), an estimate of the asymptotic variance of \( \hat{Q}_n(p) \) is

\[
\hat{\text{AV}}(\hat{Q}_n(p)) = \frac{p(1-p)}{\hat{f}(\hat{Q}_n(p))^2}.
\]

As a simple consistent estimator of \( f(Q(p))^{-1} \), Cox and Hinkley (1974) suggest

\[
\frac{Z([np]+h_n) - Z([np]-h_n)}{2h_n/n},
\]

where \([x]\) denotes the integer part of \( x \) and \( h_n \) is a bandwidth that goes to zero as \( n \to \infty \).
Asymptotic relative efficiency of the median to the mean

We may use result (42) to compare the asymptotic properties of the sample mean $\bar{Z}_n$ and the sample median $\hat{\zeta}_n$ for a random sample from a rv $Z$ with a unimodal distribution that is symmetric about $\mu$ with variance $0 < \sigma^2 < \infty$ and density $f$ that is strictly positive at $\mu$.

Under these assumptions, as $n \to \infty$,

$$\sqrt{n} (\bar{Z}_n - \mu) \Rightarrow \mathcal{N}(0, \sigma^2),$$

$$\sqrt{n} (\hat{\zeta}_n - \mu) \Rightarrow \mathcal{N} \left( 0, \frac{1}{4f(\mu)^2} \right).$$

Because the two estimators are asymptotically normal with the same asymptotic mean, their comparison may be based on the ratio of their asymptotic variances

$$\text{ARE}(\hat{\zeta}_n, \bar{Z}_n) = \frac{\text{AV}(\bar{Z}_n)}{\text{AV}(\hat{\zeta}_n)} = 4\sigma^2 f(\mu)^2,$$

called the asymptotic relative efficiency (ARE) of the sample median to the sample mean.

Because $\text{Var} \bar{Z}_n \approx n^{-1} \text{AV}(\bar{Z}_n)$ and $\text{Var} \hat{\zeta}_n \approx n^{-1} \text{AV}(\hat{\zeta})$, the ARE is equal to the ratio of the sample sizes that the sample mean and the sample median respectively need to attain the same level of precision (inverse of the sampling variance). Thus, if 100 observations are needed for the sample median to attain a given level of precision, then the sample mean would need about $100 \times \text{ARE}$ observations.

The ARE of the sample median increases with the peakedness $f(\mu)$ of the density $f$ at $\mu$. It is easy to verify that

$$\text{ARE}(\hat{\zeta}_n, \bar{Z}_n) = \begin{cases} 
2/\pi \approx .64, & \text{if } Z \text{ Gaussian}, \\
\pi^2/12 \approx .82, & \text{if } Z \text{ logistic}, \\
2, & \text{if } Z \text{ double exponential}.
\end{cases}$$

The table below shows the ARE of the sample median for $t$ distributions with $m \geq 3$ dof.

<table>
<thead>
<tr>
<th>$m$</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>8</th>
<th>$\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARE</td>
<td>1.62</td>
<td>1.12</td>
<td>.96</td>
<td>.80</td>
<td>.64</td>
</tr>
</tbody>
</table>

142
3.4 Conditional distribution and quantile functions

Consider a random vector \((X, Y)\), where \(X\) is \(\mathbb{R}^k\)-valued and \(Y\) is real-valued. The conditional distribution of \(Y\) given \(X = x\) may **equivalently** be characterized through either:

- The **conditional df** (or **distributional regression function**), a real-valued function defined on \(\mathbb{R} \times \mathbb{R}^k\) by
  \[
  F(y \mid x) = \Pr\{Y \leq y \mid X = x\}.
  \]

- The **conditional qf** (or **quantile regression function**), a real-valued function defined on \((0, 1) \times \mathbb{R}^k\) by
  \[
  Q(p \mid x) = \inf\{y \in \mathbb{R} : F(y \mid x) \geq p\}.
  \]

For any fixed \(x\), \(F(y \mid x)\) and \(Q(p \mid x)\) satisfy all properties of a df and a qf, respectively. In particular,

\[
F(y \mid x) = \sup\{p \in (0, 1) : Q(p \mid x) \leq y\}.
\]

To stress the relationship between the conditional qf and the conditional df, the notations \(Q(p \mid x) = F^{-1}(p \mid x)\) and \(F(y \mid x) = Q^{-1}(y \mid x)\) will also be used.

Further, for any fixed \(x\), the following **monotonicity** or **no-crossing** properties must be satisfied:

- if \(y' > y\), then \(F(y' \mid x) \geq F(y \mid x)\),
- if \(p' > p\), then \(Q(p' \mid x) \geq Q(p \mid x)\).
Conditional and marginal df’s and qf’s

The Law of Iterated Expectations implies that

$$EY = E \mu(X) = \int \mu(x) \, dH(x),$$

where $H$ is the df of $X$. Thus, given two sub-populations with mean $\mu_1$ and $\mu_2$ respectively, the difference in their means may be decomposed as

$$\mu_1 - \mu_2 = \int [\mu_1(x) - \mu_2(x)] \, dH_1(x) + \int \mu_2(x) \, [dH_1(x) - dH_2(x)],$$

where the 1st term on the rhs reflects differences in the CMF of the two sub-populations, while the 2nd term reflects differences in their composition.

When the two CMF’s are linear, that is, $\mu_j(x) = \beta_j^T x$, $j = 1, 2$, we have the so-called **Blinder-Oaxaca decomposition**

$$\mu_1 - \mu_2 = (\beta_1 - \beta_2)^T \mu_{X,1} + \beta_2^T (\mu_{X,1} - \mu_{X,2}),$$

where $\mu_{X,j}$ denotes the mean of $X$ in group $j = 1, 2$.

In the case of a df we have

$$F(y) = \int F(y \mid x) \, dH(x),$$

so differences in the marginal distribution of $Y$ for two sub-populations may be decomposed as follows

$$F_1(y) - F_2(y) = \int [F_1(y \mid x) - F_2(y \mid x)] \, dH_1(x) +$$

$$+ \int F_2(y \mid x) \, [dH_1(x) - dH_2(x)],$$

where the 1st term on the rhs reflects differences in the cdf of the two sub-populations, while the 2nd term reflects differences in their composition.

Simple decomposition of this kind are not available for quantiles. Machado and Mata (2005) and Melly (2005) propose methods essentially based on inversion of (43). Recently, Firpo, Fortin and Lemieux (2011) have proposed an approximate method based on the recentered influence function. See also Fortin, Lemieux and Firpo (2011).
Estimation when \( X \) is discrete

The methods reviewed so far can easily be extended to estimation of the conditional df and qf when the covariate vector \( X \) has a discrete distribution.

All is needed is partitioning the data according to the values of \( X \). The conditional df and the regression qf corresponding to each of these values may then be estimated by treating the relevant subset of observations as if they were a separate sample.

**Example 26** When \( X \) is a discrete rv and \( x \) is one of its possible values, the conditional df of \( Y \) given \( X = x \) is defined as

\[
F(y \mid x) = \Pr\{Y \leq y \mid X = x\} = \frac{\Pr\{Y \leq y, X = x\}}{\Pr\{X = x\}}.
\]

If \( O(x) = \{i: X_i = x\} \) is the set of sample points such that \( X_i = x \) and \( n(x) \) is their number, then the sample counterpart of \( \Pr\{X = x\} \) is the fraction \( n(x)/n \) of sample points such that \( X_i = x \). Hence, if \( n(x) > 0 \), a reasonable estimate of \( F(y \mid x) \) is the fraction of sample points in \( O(x) \) such that \( Y_i \leq y \)

\[
\hat{F}(y \mid x) = n(x)^{-1} \sum_{i=1}^{n} 1\{Y_i \leq y, X_i = x\} = n(x)^{-1} \sum_{i \in O(x)} 1\{Y_i \leq y\}.
\]

(44)

Viewed as a function of \( y \) for \( x \) fixed, \( \hat{F}(y \mid x) \) is called the conditional edf of \( Y_i \) given \( X_i = x \).

For these estimates to make sense, the number of data points corresponding to each of the possible values of \( X \) ought to be sufficiently large. This approach breaks down, therefore, when either \( X \) is continuous or \( X \) is discrete but its support contains too many points relative to the available sample size.

In what follows, we begin with the problem of estimating the conditional qf \( Q(p \mid x) \) when \( X \) is a continuous random \( k \)-vector. In the last two decades, this problem has received considerable attention in econometrics.

145
3.5 Estimating the conditional quantile function

Let \((X_1, Y_1), \ldots, (X_n, Y_n)\) be a sample from the joint distribution of \((X, Y)\). The conditional qf \(Q(p \mid x)\), viewed as a function of \(x\) for a given \(p\), may be obtained by solving the problem

\[
\min_{c \in C} E \ell_p(Y - c(X)), \quad 0 < p < 1,
\]

(45)

where \(C\) is the class of real valued functions defined on \(\mathbb{R}^k\).

For a given \(p \in (0, 1)\), this suggests estimating \(Q(p \mid x)\) by choosing a function of \(x\), out of a suitable family \(C^* \subset C\), as to solve the sample analogue of (45)

\[
\min_{c \in C^*} n^{-1} \sum_{i=1}^n \ell_p(Y_i - c(X_i)), \quad 0 < p < 1.
\]

This is called the quantile regression approach.

To recover \(Q(p \mid x)\), now viewed as a function of both \(p\) and \(x\), it is customary to select \(J\) distinct values

\[0 < p_1 < \ldots < p_J < 1,
\]

and then estimate \(J\) distinct functions \(Q_1, \ldots, Q_J\), each defined on \(\mathbb{R}^k\), with \(Q_j(x) = Q(p_j \mid x)\). There may be as many such values as one wishes. By suitably choosing their number and position, one may get a reasonably accurate description of \(Q(p \mid x)\).

Given an estimate \(\hat{Q}(p \mid x)\) of \(Q(p \mid x)\), \(F(y \mid x)\) may be estimated by inversion

\[
\hat{F}(y \mid x) = \sup \{p \in (0, 1): \hat{Q}(p \mid x) \leq y\}.
\]

This is a proper df iff \(\hat{Q}(p \mid x)\) is a proper qf.
Linear quantile regression

In the original approach of Koenker and Bassett (1978), $C^*$ is the class of linear functions of $x$, that is, $c(x) = b^\top x$ (unless stated otherwise, the model always includes a constant).

In this case, the sample counterpart of problem (45) is the **ALAD problem**

$$\min_{b \in \mathbb{R}^k} n^{-1} \sum_{i=1}^{n} \ell_p(Y_i - b^\top X_i), \quad 0 < p < 1. \quad (46)$$

A solution $\hat{\beta}(p)$ to (46) is called a linear quantile regression estimate. Given a solution $\hat{\beta}(p)$ to (46), an estimate of $Q(p \mid x)$ is

$$\hat{Q}(p \mid x) = \hat{\beta}(p)^\top x.$$  

If $X_i$ only contains a constant term, that is, $X_i = 1$, $i = 1, \ldots, n$, then problem (46) is equivalent to problem (41), so $\hat{\beta}(p) = \hat{Q}(p)$ in this case.

By suitably redefining the elements of the covariate vector $X_i$, one may easily generalize problem (46) to cases in which $C^*$ is a class of functions that depend linearly on a finite dimensional parameter vector, such as the class of polynomial functions of $X$ of a given degree.
Computational aspects of linear quantile regression

The lack of smoothness of the ALAD objective function implies that gradient methods cannot be employed to solve (46).

An ALAD estimate may however be computed efficiently using linear programming methods (Barrodale & Roberts 1973, Koenker & d’Orey 1987). This is because the ALAD problem (46) may be reformulated as the following linear program

$$\min_{\beta \in \mathbb{R}^k} p\beta^T U^+ + (1 - p)\ell^T U^-$$
subject to $Y = X\beta + U^+ - U^-$
$$U^+ \geq 0, U^- \geq 0,$$

where $\ell$ is the $n$-vector of ones, and $U^+$ and $U^-$ are $n$-vectors with generic elements equal to $U^+_i = \max(0, Y_i - \beta^T X_i)$ and $U^-_i = -\min(0, Y_i - \beta^T X_i)$ respectively.

Simpler but cruder algorithms based on iterative WLS may also be employed.

The dual of the above linear program is

$$\max_{\delta \in [p-1,p]^n} Y^T \delta$$

subject to $X^T \delta = 0$

or, setting $\alpha = \delta + (1 - p)\ell_n$,

$$\max_{\alpha \in [0,1]^n} Y^T \alpha$$

subject to $X^T \alpha = (1 - p)X^T \ell_n$.

The solution $\hat{\alpha}_p$ to the dual problem connects quantile regression to regression rank scores, the regression generalization of classical rank scores proposed by Gutenbrunner and Jurečková (1992).
Asymptotics for linear quantile regression estimators

Let \( \hat{\beta}_n = (\hat{\beta}_n(p_1), \ldots, \hat{\beta}_n(p_J)) \) be a \( kJ \)-vector of linear quantile regression estimators, and consider the following assumptions:

**Assumption 2** \((X_1, Y_1), \ldots, (X_n, Y_n)\) is a random sample from the joint distribution of \((X, Y)\), where the first element of \(X\) is the constant term.

**Assumption 3** \( EXX^\top = P \), a finite \( pd \times k \) matrix.

**Assumption 4** The conditional df of \(Y\) given \(X = x\) satisfies

\[
Pr\{Y \leq y \mid X = x\} = F(y - \beta^\top x),
\]

where \(F\) has a continuous and strictly positive density \(f\) and a continuous and strictly increasing qf \(Q\) such that \(Q(p) = 0\) for some \(0 < p < 1\).

Assumption 4 is equivalent to the assumption that

\[
Y_i = \beta^\top X_i + U_i, \quad i = 1, \ldots, n,
\]

where the \(U_i\) are distributed independently of the \(X_i\) with df \(F\) (pure location shift model). This assumption implies that

\[
Q(p \mid x) = \beta^\top x + Q(p) = \beta(p)^\top x,
\]

where the first element of \(\beta(p)\) is equal to \(\beta_1 + Q(p)\), with \(\beta_1\) the first element of \(\beta\). If \(p\) is such that \(Q(p) = 0\), then \(\beta(p) = \beta\).
Consistency and asymptotic normality

Under Assumptions 2–4,

\[ \hat{\beta}_n \xrightarrow{as} \beta, \]

where \( \beta = (\beta(p_1), \ldots, \beta(p_J)) \). Further,

\[ \sqrt{n}(\hat{\beta}_n - \beta) \Rightarrow \mathcal{N}_{k,J}(0, \Omega \otimes P^{-1}), \quad (47) \]

where \( \Omega \) is the \( J \times J \) matrix with generic element

\[ \omega_{jk} = \frac{\min(p_r, p_s) - p_r p_s}{f(Q(p_r)) f(Q(p_s))}, \quad r, s = 1, \ldots, J. \]

It follows from this result that, if \( X \) only contains a constant term, then \( P = 1 \), \( \hat{\beta}_n = \bar{Q}_n \), and the asymptotic variance of \( \hat{\beta}_n \) is equal to \( \Omega \).

Suppose, in addition, that \( F \) has finite variance \( \sigma^2 \). Then, under the above two assumptions, result (47) implies that the ARE of a linear quantile regression estimator \( \hat{\beta}_n(p) \) relative to the OLS estimator \( \tilde{\beta}_n \) is equal to

\[ \text{ARE}(\hat{\beta}_n(p), \tilde{\beta}_n) = \frac{\sigma^2 f(Q(p))^2}{p(1 - p)}. \]

In practice, to construct estimates of the asymptotic variance of linear quantile regression estimators, one needs estimates of the density of the quantile regression errors \( U_i = Y_i - X_i^\top \beta \).
Extensions

Koenker and Portnoy (1987) provide a **uniform Bahadur representation** for linear quantile regression estimators.

They show that, under Assumptions 2–4, with probability tending to one as $n \to \infty$,

$$
\sqrt{n} [\hat{\beta}_n(p) - \beta(p)] = P^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{p - 1 \{U_i \leq Q(p)\}}{f(Q(p))} X_i + o_p(1),
$$

uniformly for $p \in [\epsilon, 1 - \epsilon]$ and some $\epsilon > 0$.

Among other things, the asymptotically linear representation (48) helps explain both the **asymptotic normality** and the **good robustness properties** of linear quantile regression estimators with respect to outliers in the $Y$-space (although not to outliers in the $X$-space).

Both properties follow from the fact that the **influence function** of $\hat{\beta}(p)$ is equal to

$$
P^{-1} \frac{p - 1 \{U_i \leq Q(p)\}}{f(Q(p))} X_i,
$$

which is a **bounded** function of $U_i$ (although not of $X_i$).
Drawbacks of linear quantile regression estimators

Although increasingly used in empirical work to describe the conditional distribution of an outcome of interest, linear quantile regression estimators have several drawbacks.

Some have to do with general properties of quantiles, especially:

- the difficulty of imposing the no-crossing condition;
- the complicated relationship between marginal and conditional quantiles;
- the difficulty of generalizing quantiles to the case when $Y$ is vector valued.

Others have to do with their specific properties, especially their behavior under heteroskedasticity. The relevant issues are:

- the validity of the linearity assumption;
- the form of the asymptotic variance of linear quantile regression estimators when the linearity assumption does not hold;
- how to consistently estimate this asymptotic variance.
Behavior under heteroskedasticity

To illustrate the problem, let $X$ be a scalar rv and suppose first that $Y = \mu(X) + U$, where the regression error $U$ is independent (not just mean independent) of $Y$ with continuous strictly increasing df $F$. Then $F(y | x) = F(y - \mu(x))$. By definition, the $p$th conditional quantile of $Y$ satisfies

$$F(Q(p | x) | x) = p.$$ 

Hence

$$Q(p | x) = Q(p) + \mu(x),$$

where $Q(p) = F^{-1}(p)$ is the $p$th quantile of $F$. Thus, for any $p \neq p'$,

$$Q(p' | x) - Q(p | x) = Q(p') - Q(p)$$

for all $x$, that is, the distance between any pair of conditional quantiles is independent of $x$. If $\mu(x) = \alpha + \beta x$, then $Q(p | x) = [\alpha + Q(p)] + \beta x$, that is, the conditional quantiles of $Y$ are a family of parallel lines with common slope $\beta$.

Now suppose that $\mu(x) = \alpha + \beta x$ but $Y$ is conditionally heteroskedastic, that is, $Y = \alpha + \beta X + \sigma(X) U$, where the function $\sigma(x)$ is strictly positive. The homoskedastic model is a special case where $\sigma(x) = 1$ for all $x$. Now

$$F(y | x) = F\left(\frac{y - \alpha - \beta x}{\sigma(x)}\right),$$

so

$$Q(p | x) = \alpha + \beta x + \sigma(x)Q(p).$$

In this case:

- although $\mu(x)$ is linear in $x$, conditional quantiles need not be;
- the distance between any pair of conditional quantiles depends on $x$,

$$Q(p' | x) - Q(p | x) = \sigma(x)[Q(p') - Q(p)].$$

Partial exceptions are the cases when:

- $\mu(x)$ is linear and $F$ is symmetric about zero, implying that the conditional median $Q(.5 | x)$ (but not other quantiles) is linear in $x$;
- $\sigma(x)$ is linear in $x$, implying that conditional quantiles are linear in $x$, although no longer with a common slope.
Figure 23: Quantiles of $Y \mid X = x \sim \mathcal{N}(\mu(x), \sigma^2(x))$ when $\mu(x) = 1 + x$ and either $\sigma^2(x) = 1$ (homoskedasticity) or $\sigma^2(x) = 1 + (2x + .5)^2$ (heteroskedasticity).
Implications

- Linear quantile regressions may be a poor approximation to population quantiles when data are conditionally heteroskedastic and the square root of the conditional variance function is far from being linear in \( x \).

- In the general case when \( \mu(x) \) is an arbitrary function, the conditional quantiles of \( Y \) are of the form

\[
Q(p \mid x) = \mu(x) + \sigma(x)Q(p).
\]

In the absence of prior information, it is impossible to determine whether nonlinearity of \( Q(p \mid x) \) reflects nonlinearity of \( \mu(x) \), heteroskedasticity, or both.

**Example 27** If \( \mu(x) = \beta_0 + \beta_1 x \) and \( \sigma(x) = \delta_0 + \delta_1 x \), then

\[
Q(p \mid x) = \alpha(p) + \beta(p)x,
\]

where \( \alpha(p) = \beta_0 + \delta_0 Q(p) \) and \( \beta(p) = \beta_1 + \delta_1 Q(p) \).

If instead \( \mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2 \) and \( \sigma(x) = \delta_0 + \delta_1 x \), then

\[
Q(p \mid x) = \alpha(p) + \beta(p)x + \gamma x^2,
\]

where \( \alpha(p) = \beta_0 + \delta_0 Q(p) \) and \( \beta(p) = \beta_1 + \delta_1 Q(p) \) depend on \( p \), but \( \gamma = \beta_2 \) does not.

Finally, if \( \mu(x) = \beta_0 + \beta_1 x \) and \( \sigma(x) = (\delta_0 + \delta_1 x)^2 \), then

\[
Q(p \mid x) = \alpha(p) + \beta(p)x + \gamma(p)x^2,
\]

where \( \alpha(p) = \beta_0 + \delta_0^2 Q(p) \), \( \beta(p) = \beta_1 + 2\delta_0 \delta_1 Q(p) \) and \( \gamma(p) = \delta_0^2 Q(p) \) all depend on \( p \).
Interpreting linear quantile regressions

Consider the ordinary (mean) regression model $Y = \beta^\top X + U$, where $U$ and $X$ are uncorrelated. In this model, $\beta^\top X$ may equivalently be interpreted as the best linear predictor of $Y$ given $X$, or as the best linear approximation to the conditional mean $\mu(X) = \mathbb{E}(Y \mid X)$, that is,

$$
\beta = \arg \min_{b \in \mathbb{R}^k} \mathbb{E}(Y - b^\top X)^2
= \arg \min_{b \in \mathbb{R}^k} \mathbb{E}_X[\mu(X) - b^\top X]^2,
$$

where the first expectation is with respect to the joint distribution of $X$ and $Y$, while the second is with respect to the marginal distribution of $X$.

Angrist, Chernozhukov and Fernandez-Val (2006) show that a similar interpretation is available for linear quantile regression. More precisely, they show that if $\beta(p)$ is the slope of a linear quantile regression, then

$$
\beta(p) = \arg \min_{b \in \mathbb{R}^k} \mathbb{E}_p(Y - b^\top X)
= \arg \min_{b \in \mathbb{R}^k} \mathbb{E}_X w(X, b) [Q(p \mid X) - b^\top X]^2,
$$

with

$$
w(x, b) = \int_0^1 (1 - u) f (Q(p \mid x) + u \Delta(x, b) \mid x) \, du,
$$

where $\Delta(x, b) = b^\top x - Q(p \mid x)$ and $f(y \mid x)$ is the conditional density of $Y$ given $X = x$. Thus, linear quantile regression minimizes a weighted quadratic measure of discrepancy between the population conditional qf and its best linear approximation. The weighting function, however, is not easily interpretable.

For an alternative proof of this result, see Theorem 2.7 in Koenker (2005).
Inference under heteroskedasticity

From the practical point of view, heteroskedasticity implies that estimated linear quantile regressions may cross each other, thus violating a fundamental property of quantiles and complicating the interpretation of the results of a statistical analysis.

It also implies that the asymptotic variance matrix of $\hat{\beta}$ is no longer given by (47). The block corresponding to the asymptotic covariance between $\hat{\beta}_j$ and $\hat{\beta}_k$ is instead equal to

$$[\min(p_j, p_k) - p_j p_k] B_j^{-1} P B_k^{-1},$$

where

$$B_j = \mathbb{E}_X \left[ f(X^T \beta_j \mid X) XX^T \right]$$

and $f(y \mid x)$ denotes the conditional density of $Y$ given $X = x$.

An analog estimator of $B_j$ is not easy to obtain, as it requires estimating the conditional density $f(y \mid x)$, so bootstrap methods are typically used to carry out inference in this case.
Figure 24: Scatterplot of the data and estimated quantiles for a random sample of 200 observations from the model $Y \mid X \sim \mathcal{N}(\mu(X), \sigma^2(X))$, with $X \sim \mathcal{N}(0, 1)$, $\mu(X) = 1 + X$ and either $\sigma^2(X) = 1$ (homoskedasticity) or $\sigma^2(X) = 1 + (2X + .5)^2$ (heteroskedasticity).
Nonparametric quantile regression estimators

To overcome the problems arising with the linearity assumption, several nonparametric quantile regression estimators have been proposed. These include:

- **kernel and nearest neighbor** methods (Chauduri 1991),
- **regression splines** with a fixed number of knots (Hendricks & Koenker 1992),
- **smoothing splines** and **penalized likelihood** (Koenker, Ng & Portnoy 1994),
- **locally linear fitting** (Yu & Jones 1998), that is, \( \hat{Q}(p | x) = \hat{\beta}(p; x)^	op x, \)
with

\[
\hat{\beta}(p; x) = \arg\min_{b \in \mathbb{R}^k} n^{-1} \sum_{i=1}^{n} \ell_p(Y_i - b^	op X_i) W_i(x), \quad 0 < p < 1,
\]

where the \( W_i(x) \) are nonnegative weights that add up to one and, typically, give more weight to \( Y \)-values for which \( X_i \) is closer to \( x \).

In all these cases, the family of functions \( \mathcal{C}^* \) is left essentially unrestricted, except for smoothness.

**Drawbacks**

- **Curse-of-dimensionality problem:** It is not clear how to generalize these estimators to cases when there are more than two or three covariates.

- Because all these estimators are nonlinear, it is hard to represent them in ways that facilitate comparisons. For example, it is **not clear** how to generalize the concepts of equivalent kernel and equivalent degrees of freedom that prove so useful for linear smoothers.
3.6 Estimating the conditional distribution function

If the interest is not merely in a few quantiles but in the whole conditional distribution of $Y$ given a random $k$-vector $X$, why not estimating the conditional df $F(y \mid x)$ directly?

Given a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the joint distribution of $(X, Y)$, Stone (1977) suggested estimating $F(y \mid x)$ nonparametrically by a nearest-neighbor estimator of the form

$$\hat{F}(y \mid x) = n^{-1} \sum_{i=1}^{n} W_i(x) 1\{Y_i \leq y\}, \quad -\infty < y < \infty,$$

where $W_i(x) = W_i(x; X_1, \ldots, X_n)$ gives more weight to $Y$-values for which $X_i$ is closer to $x$.

Viewed as a function of $y$ for $x$ given, $\hat{F}(y \mid x)$ is a proper df if:

- $W_i(x) \geq 0$, $i = 1, \ldots, n$,
- $\sum_{i=1}^{n} W_i(x) = 1$.

Stute (1986) studied the properties of two estimators of this type:

- The first is based on weights of the form
  $$W_i(x) = \frac{1}{h_n} K \left( \frac{\hat{H}(x) - \hat{H}(X_i)}{h_n} \right),$$
  where $\hat{H}(x)$ is the edf of the $X$, $K$ is a smooth kernel with bounded support and $h_n$ is the bandwidth. For any $x$, $\hat{F}(y \mid x)$ is not a proper df, as the kernel weights do not add up to one.

- The second, denoted by $\hat{F}^*(y \mid x)$ is a proper df for it chooses normalized weights $W_i^*(x) = W_i(x)/\sum_j W_j(x)$.

Estimators of this kind tend to do rather poorly when data are sparse, which is typically the case $k$ is greater than 2 or 3.
**Avoiding the curse of dimensionality**

We now describe a simple **semi-nonparametric method** that appears to perform well even when $k$ is **large** relative to the available data. The basic idea is to partition the range of $Y$ into $J + 1$ intervals defined by a set of “knots”

$$-\infty < y_1 < \ldots < y_J < \infty,$$

and then estimate the $J$ functions $F_1(x), \ldots, F_J(x)$, where $F_j(x) = F(y_j | x) = \Pr(Y \leq y_j | X = x)$.

If the conditional distribution of $Y$ is **continuous** with support on the whole real line then, at any $x$ in the support of $X$, the sequence of functions $\{F_j(x)\}$ must satisfy the following conditions:

$$0 < F_j(x) < 1, \quad j = 1, \ldots, J, \quad (49)$$

$$0 < F_1(x) < \cdots < F_J(x) < 1. \quad (50)$$

One way of automatically imposing the **nonnegativity** condition (49) is to model not $F_j(x)$ directly, but rather the conditional log odds

$$\eta_j(x) = \ln \frac{F_j(x)}{1 - F_j(x)}.$$

Given an estimate $\hat{\eta}_j(x)$ of $\eta_j(x)$, one may then estimate $F_j(x)$ by

$$\hat{F}_j(x) = \frac{\exp \hat{\eta}_j(x)}{1 + \exp \hat{\eta}_j(x)}.$$
Estimation

By the analogy principle, $\eta_j$ may be estimated by maximizing the sample log likelihood

$$L(\eta) = \sum_{i=1}^{n} [1\{Y_i \leq y_j\} \eta(X_i) - \ln(1 + \exp(\eta(X_i)))]$$

over a suitable family $\mathcal{H}^*$ of functions of $x$.

This approach, which may be called the distributional regression approach, entails fitting $J$ separate logistic regression, one for each threshold value $y_j$. Boundedness of $1\{Y_i \leq y_j\}$ ensures good robustness properties with respect to outliers in the $Y$-space.

Alternative specifications of $\mathcal{H}^*$ correspond to alternative estimation methods. As for quantile regressions, one may distinguish between:

- **parametric methods** (e.g. logit, probit);
- **nonparametric methods** with $\mathcal{H}^*$ unrestricted, except for smoothness;
- **nonparametric methods** with $\mathcal{H}^*$ restricted (such as projection pursuit, additive or locally-linear modeling).
Example 28 The simplest case is when \( \eta(y \mid x) = \theta(y)^\top x \).

For \( j = 1, \ldots, J \), let \( \hat{\theta}_{jn} \) be the logit estimator of \( \theta_j = \theta(y_j) \) and let \( \hat{\eta}_{jn}(x) = \hat{\theta}_{jn}^\top x \) be the implied estimator of \( \eta_j(x) \). Also let

\[
F(x) = \begin{pmatrix} F_1(x) \\ \vdots \\ F_J(x) \end{pmatrix}, \quad \theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_J \end{pmatrix}, \quad \eta(x) = \begin{pmatrix} \eta_1(x) \\ \vdots \\ \eta_J(x) \end{pmatrix} = (I_J \otimes x^\top) \theta,
\]

where \( I_J \) is the \( J \times J \) unit matrix. Finally, let \( \hat{F}_n(x), \hat{\theta}_n \) and \( \hat{\eta}_n(x) = (I_J \otimes x^\top) \hat{\theta}_n \) be the estimators of \( F(x), \theta \) and \( \eta(x) \) respectively.

Under mild regularity conditions, as \( n \to \infty \),

\[
\sqrt{n} (\hat{\theta}_n - \theta) \Rightarrow \mathcal{N}_{kJ}(0, \mathcal{I}^{-1}),
\]

where

\[
\mathcal{I} = \mathbb{E}[V(X) \otimes XX^\top]
\]

and \( V(x) \) is the \( J \times J \) matrix with generic element

\[
V_{jk}(x) = \min(F_j(x), F_k(x)) - F_j(x) F_k(x).
\]

Hence, as \( n \to \infty \),

\[
\sqrt{n} [\hat{\eta}_n(x) - \eta(x)] \Rightarrow \mathcal{N}_J(0, \mathcal{A}(x)),
\]

where

\[
\mathcal{A}(x) = (I_J \otimes x^\top) \mathcal{I}^{-1} (I_J \otimes x).
\]

Therefore, as \( n \to \infty \),

\[
\sqrt{n} [\hat{F}_n(x) - F(x)] \Rightarrow \mathcal{N}_{J}(0, \Sigma(x)),
\]

where

\[
\Sigma(x) = V(x) \mathcal{A}(x) V(x)^\top.
\]

\[\square\]
Figure 25: Estimated conditional df’s of log monthly earnings (Peracchi 2002).
Imposing monotonicity

While our approach automatically imposes the nonnegativity condition (49), it does **not** guarantee that the monotonicity condition (50) also holds. Since $\eta_j(x)$ is strictly increasing in $F_j(x)$, monotonicity is **equivalent** to the condition that

$$-\infty < \eta_1(x) < \ldots < \eta_J(x) < \infty$$

for all $x$ in the support of $X$.

If $\eta_j(x) = \gamma_j + \mu_j(x)$, then monotonicity holds if

$$\gamma_j > \gamma_{j-1}, \quad \mu_j(x) \geq \mu_{j-1}(x).$$

One case where these two conditions are satisfied is the **ordered logit model**. This model is **restrictive**, however, for it implies that changes in the covariate vector $X$ affect the conditional distribution of $Y$ only through a **location shift**.

An alternative is to model $F_1(x)$ and the conditional probabilities or **discrete hazards**

$$\lambda_j(x) = \Pr\{Y \leq y_j \mid Y > y_{j-1}, X = x\}$$

$$= \frac{S_{j-1}(x) - S_j(x)}{S_{j-1}(x)}, \quad j = 2, \ldots, J,$$

where $S_j(x) = 1 - F_j(x)$ is the **survivor function** evaluated at $y_j$.

Using the recursion

$$S_j(x) = [1 - \lambda_j(x)] S_{j-1}(x), \quad j = 2, \ldots, J,$$

we get

$$S_k(x) = S_1(x) \prod_{j=2}^k [1 - \lambda_j(x)], \quad k = 2, \ldots, J,$$

that is,

$$F_k(x) = 1 - [1 - F_1(x)] \prod_{j=2}^k [1 - \lambda_j(x)], \quad k = 2, \ldots, J.$$

If $F_1(x)$ and the $\lambda_j(x)$ are modeled to guarantee that $0 < F_1(x) < 1$ and $0 < \lambda_j(x) < 1$, then **both** monotonicity and the constraint (49) are automatically satisfied.
Figure 26: Estimated conditional df’s of log monthly earnings imposing monotonicity (Peracchi 2002).
Extensions: autoregressive models

Consider the discrete-time univariate AR(1) process

\[ Y_t = \rho Y_{t-1} + \sigma U_t, \]

where \(|\rho| < 1\), \(\sigma > 0\) and the \(\{U_t\}\) are iid with zero mean and marginal df \(G\). The conditional df of \(Y_t\) given \(Y_{t-1} = x\) is

\[ F(y \mid x) = \Pr\{Y_t \leq y \mid Y_{t-1} = x\} = G\left(\frac{y - \rho x}{\sigma}\right). \quad (51) \]

By the stationarity assumption, this is also the conditional df of \(Y_{t+h}\) given \(Y_{t+h-1} = x\) for any \(|h| = 0, 1, 2, \ldots\).

The assumptions implicit in (51) are strong. As an alternative, one may retain the assumption that \(F(y \mid x)\) is time-invariant and apply the results of the previous section by letting \(X_t = Y_{t-1}\).

Extensions to a univariate \(p\)th autoregression is straightforward. In this case, what is assumed to be time-invariant is

\[ F(y \mid x) = \Pr\{Y_t \leq y \mid X_t = x\}, \]

where \(X_t = (Y_{t-1}, \ldots, Y_{t-p})\) and \(x = (x_1, \ldots, x_p)\).
3.7 Relationships between the two approaches

Koenker, Leorato and Peracchi (2013) ask the general question: How does the distributional regression (DR) approach outlined in the previous section relate to quantile regression (QR) approach?

Here we focus on two specific questions:

- What restrictions on the family of conditional quantiles are implied by specific assumptions on the family of conditional log odds?

- What restrictions on the family of conditional log odds are implied by specific assumptions on the family of conditional quantiles?
From log odds to quantiles

If $F(y \mid x)$ is continuous, the fact that $F(Q(p \mid x) \mid x) = p$ implies

$$\eta(Q(p \mid x) \mid x) = \ln \frac{p}{1 - p}, \quad p \in (0, 1).$$

Suppose that $\eta(y \mid x)$ is some known continuously differentiable function of $(x, y)$ and $\eta_y(y \mid x)$ is strictly positive for every $(x, y)$, with subscripts denoting partial derivatives. By the implicit function theorem applied to

$$\eta(y \mid x) = \ln \frac{p}{1 - p},$$

the conditional qf is unique and continuously differentiable in $x$ with derivative

$$Q_x(p \mid x) = -\frac{\eta_x(y \mid x)}{\eta_y(y \mid x)} \bigg|_{y=Q(p \mid x)} = -\frac{F_x(y \mid x)}{f(y \mid x)} \bigg|_{y=Q(p \mid x)}.$$

Thus, $Q_x(p \mid x)$ and $\eta_x(Q(p \mid x))$ have opposite sign.

Next notice that conditional quantiles are linear in $x$ iff the partial derivative $Q_x(p \mid x)$ does not depend on $x$. If the conditional log odds are linear in $x$, that is, $\eta(y \mid x) = \gamma(y) + \delta(y)x$, then a quantile regression is linear in $x$ iff

$$Q_x(p \mid x) = -\frac{\delta(y)}{\gamma'(y) + \delta'(y)x} \bigg|_{y=Q(p \mid x)}$$

does not depend on $x$. Sufficient conditions are: (i) $\gamma(y)$ is linear in $y$, and (ii) $\delta(y)$ does not depend on $y$ (as with the ordered logit model).
From quantiles to log odds

If \( Q(p \mid x) \) is a known continuously differentiable function of \((p, x)\) such that \( Q(p \mid x) = y \), where \( y \) is a fixed number, then

\[
F(Q(p \mid x)) = F(y \mid x).
\]

By the chain rule,

\[
F_x(y \mid x) = -Q_x(p \mid x) \frac{f(y \mid x)}{p=F(y \mid x)}
\]
or, equivalently,

\[
\eta_x(y \mid x) = -Q_x(p \mid x) \frac{\eta_y(y \mid x)}{p=F(y \mid x)}.
\]

Clearly, conditional log odds are linear iff \( \eta_x(y \mid x) \) does not depend on \( x \). If conditional quantiles are linear, that is \( Q(p \mid x) = \alpha(p) + \beta(p) x \), then conditional log odds are linear iff

\[
\eta_x(y \mid x) = -\beta(p) \frac{\eta_y(y \mid x)}{p=F(y \mid x)}
\]
does not depend on \( x \). Sufficient conditions are: (i) \( \beta(p) \) does not depend on \( p \), and (ii) \( \eta_y(y \mid x) \) does not depend on \( x \).

Example 29 Let \( Y = \alpha + \beta X + U \), where \( U \) has a logistic distribution with mean zero and qf \( Q \). Then \( Q(p \mid x) = \alpha(p) + \beta x \), where \( \alpha(p) = \alpha + Q(p) \), whereas

\[
F(y \mid x) = \frac{\exp(y - \alpha - \beta x)}{1 + \exp(y - \alpha - \beta x)}.
\]

Hence \( \eta(y \mid x) = \gamma(y) + \delta(y)x \), where \( \delta(y) = -\beta \) and \( \gamma(y) = y - \alpha \), is also linear in \( x \). Notice that, except for the sign, \( Q(p \mid x) \) and \( \eta(y \mid x) \), viewed as functions of \( x \), have the same slope. \( \square \)
3.8 Stata commands

We now briefly review the commands available in Stata, version 12.

These include the `qreg` command for linear conditional quantile estimation, the associated `iqreg`, `sqreg` and `bsqreg` commands, and the post-estimation tools in `qreg postestimation`.

At the moment, Stata only offers the `cumul` for estimating univariate distribution functions and has no command for estimating cdf’s.
The `qreg` command

The basic syntax is:

```
qreg depvar indepvars [if] [in] [weight] [, qreg_options]
```

where `qreg_options` includes:

- `quantile(#)`: specifies the quantile to be estimated and should be a number between 0 and 1, exclusive. Numbers larger than 1 are interpreted as percentages. The default value of 0.5 corresponds to the median.

- `level(#)`: sets the confidence level. The default is `level(95)` (95 percent).

- `wlsiter(#)`: specifies the number of WLS iterations that will be attempted before the linear programming iterations are started. The default value is 1. If there are convergence problems, increasing this number should help.

Notice that the standard errors produced by `qreg` are based on the homoskedasticity assumption and should not be trusted.
Other commands

The iqreg, sqreg and bsqreg commands all assume linearity of conditional quantiles but estimate the variance matrix of the estimators (VCE) via the bootstrap. Their syntax similar to that of qreg. For example,

iqreg depvar indepvars [if] [in] [weight] [, iqreg_options]

Remarks:

• iqreg estimates IQR regressions, i.e. regressions of the difference in quantiles. The available options include:
  
  – quantiles(# #): specifies the quantiles to be compared. The first number must be less than the second, and both should be between 0 and 1, exclusive. Numbers larger than 1 are interpreted as percentages. Not specifying this option is equivalent to specifying quantiles(.25 .75), meaning the IQR.
  
  – reps(#): specifies the number of bootstrap replications to be used to obtain an estimate of the VCE. The default is reps(20), arguably a little small.

• sqreg estimates simultaneous-quantile regression essentially using the algorithm in Koenker and d’Orey (1987). It produces the same coefficients as qreg for each quantile. Bootstrap estimation of the VCE includes between-quantile blocks. Thus, one can test and construct confidence intervals comparing coefficients describing different quantiles. The available options include:
  
  – quantiles(# [# [# ...]]) specifies the quantiles to be estimated and should contain numbers between 0 and 1, exclusive. Numbers larger than 1 are interpreted as percentages. The default value of 0.5 corresponds to the median.
  
  – reps(#): same as above.

• bsqreg is equivalent to sqreg with one quantile.
Post-estimation tools

The following postestimation commands are available for \texttt{qreg}, \texttt{iqreg}, \texttt{sqreg} and \texttt{bsqreg}:

- \texttt{estat}: variance matrix and estimation sample summary,
- \texttt{estimates}: cataloging estimation results,
- \texttt{lincom}: point estimates, standard errors, testing, and inference for linear combinations of the coefficients,
- \texttt{linktest}: link test for model specification,
- \texttt{margins}: marginal means, predictive margins, marginal effects, and average marginal effects,
- \texttt{nlcom}: point estimates, standard errors, testing, and inference for nonlinear combinations of coefficients,
- \texttt{predict}: predictions, residuals, influence statistics, and other diagnostic measures,
- \texttt{predictnl}: point estimates, standard errors, testing, and inference for generalized predictions,
- \texttt{test}: Wald tests of simple and composite linear hypotheses,
- \texttt{testnl}: Wald tests of nonlinear hypotheses.
References


178


