Lecture 11
Introduction to Nonparametric Regression: Density Estimation

Non Parametric Regression: Introduction

- The goal of a regression analysis is to produce a reasonable analysis to the unknown response function \( f \), where for \( N \) data points \( (X_i, Y_i) \), the relationship can be modeled as

\[
y_i = m(x_i) + \varepsilon_i, \quad i = 1, \ldots, N
\]

- **Note:** \( m(.) = \text{E}[y | x] \) if \( \text{E}[\varepsilon | x] = 0 \) - i.e., \( \varepsilon \perp x \)

- We have different ways to model the conditional expectation function (CEF), \( m(.) \):
  - Parametric approach
  - Nonparametric approach
  - Semi-parametric approach.
Non Parametric Regression: Introduction

- Parametric approach: $m(.)$ is known and smooth. It is fully described by a finite set of parameters, to be estimated. Easy interpretation. For example, a linear model:
  \[ y_i = x_i' \beta + \epsilon_i, \quad i = 1, \cdots, N \]

- Nonparametric approach: $m(.)$ is smooth, flexible, but unknown. Let the data determine the shape of $m(.)$. Difficult interpretation.
  \[ y_i = m(x_i) + \epsilon_i, \quad i = 1, \cdots, N \]

- Semi-parametric approach: $m(.)$ have some parameters -to be estimated-, but some parts are determined by the data.
  \[ y_i = x_i' \beta + m(z_i) + \epsilon_i, \quad i = 1, \cdots, N \]
Non Parametric Regression: Introduction

• Parametric and non-parametric approaches use a weighted sum of the $y$'s to obtain the fitted values, $\hat{y}$. That is,

$$\hat{y}_i = \Sigma_i \omega_i y_i$$

• Instead of using equal weights as in OLS or weights proportional to the inverse of variance as often in GLS, a different rationale determines the choice of weights in nonparametric regression.

• In the single regressor case, the observations with the most information about $f(x_0)$ should be those at locations $x_i$ closest to $x_0$.

• Thus, a decreasing function of the distances of their locations $x_i$ from $x_0$ determine the weights assigned to $y_i$'s.

Non Parametric Regression: Introduction

• A decreasing function of the distances of their locations $x_i$ from $x_0$ determine the weights assigned to $y_i$'s.

• The points closest to $x_0$ receive more weight than those more remote from $x_0$. Often, points remote from $x_0$ receive little or no weight.
Density Estimation: Univariate Case

- We have a large number of observations on a RV X. We would like to “draw” the pdf of X.

- Simplest method: Use a histogram. That is, divide the range of X into a small number of intervals (bins), $h$, and count the number of times $X_i$ is observed in each interval:

  $$ p_i = \frac{n_i(h)}{N} $$

- Q: How wide should the bins be? Too small (too many bins) distribution looks jerky, too large (few bins), shape is not easy to visualize.

- Two questions: - Do we want the same bin-width everywhere?
  - Do we believe the density is zero for empty bins? 

Density Estimation – Bins: Example

Graph 1: Histogram of a Normally Distributed Variable with Mean=20, SD=5 (1000 obs., 7 bins)

Graph 3: Histogram of a Normally Distributed Variable with Mean=20, SD=5 (1000 obs., 50 bins)
Density Estimation: Problems with Histograms

• The histogram is close to, but not truly density estimation.

• It does not try to estimate \( f(x) \) at every \( x \). Rather, it partitions the sample space into bins, and only approximate the density at the center of each bin.

• Two problems with histograms:
  (1) For a given number of bins, moving their exact location (boundary points) can change the graph.

  (2) The density function produced is a step function and the derivative either equals zero or is not defined (when at the cutoff point for two bins).

  - This is a problem if we are trying to maximize a likelihood function that is defined in terms of the densities of the distributions.

Density Estimation: Definition of Histogram

• First, define the density function for a variable \( x \). For a particular value of \( x \), call it \( x_0 \), the density function is:

\[
f(x_0) = \lim_{h \to 0} \frac{F(x_0 + h) - F(x_0 - h)}{2h} = \lim_{h \to 0} \frac{\Pr \{x_0 - h < x < x_0 + h\}}{2h}
\]

• For a sample of data on \( x \) of size \( N \), a histogram with a column width of \( 2h \), centering the column around \( x_0 \) can be approximated by:

\[
\hat{f}_{Hist}(x_0) = \frac{1}{N} \sum_{i=1}^{N} I[x_0 - h < x_i < x_0 + h] = \frac{1}{Nh} \sum_{i=1}^{N} I[\left|\frac{x_i - x_0}{h}\right| < 1]
\]

• This function equals the fraction of the sample that lies within \( h \) of \( x_0 \), divided by the column width \((2h)\). We call this the naive estimator.

• \( x_0 \) is any value of \( X \), not necessarily equal to any \( x_i \)'s in the sample.
Density Estimation: Problems Revisited

• Dealing with the two problems:
  (1) Arbitrary location of the bin cutoff points
    Solution: Define a “moving” bin that is defined for every possible value of \( x \). Then, count how many actual \( x_i \)'s are within \( b/2 \) of the hypothetical point, and “normalizes” this count by the number of total observations (\( N \)) and the “bandwidth,” \( b \).

(2) Discontinuity in the function.
    Solution: Kernel estimation. It avoids the discontinuities in the estimated (empirical) density function. In terms of histogram formula, the kernel is everything to the right of the summation sign. The general formula for the kernel estimator (Parzen window):
    \[
    \hat{f}_{Hist}(x_0) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right)
    \]

Kernel Density Estimation (KDE): Properties

• Q: Is \( \hat{f}_{Hist}(x) \) a legitimate density function? It needs to satisfy:
  (1) nonnegative and (2) integrate to one.

• Easy to do: Require the Kernel (window) function, \( K(\cdot) \) to satisfy:
  (1) \( K(x) \geq 0 \)
  (2) \( \int K(u) \, du = 1 \)

Define the function: \( \delta_{\alpha}(x) = \frac{1}{h} K\left(\frac{x}{h}\right) \)

The \( \hat{f}_{Hist}(x) \) can be written as
    \[
    \hat{f}_{Hist}(x) = \frac{1}{N} \sum_{i=1}^{N} \delta(x_i - x_0)
    \]
KDE: Properties

• Check the properties of \( \hat{f}_{Hist}(x) \) and \( \delta_n(x) \):

\[
\int \delta_n(x - x_i) \, dx = \int \frac{1}{h} K \left( \frac{x - x_i}{h} \right) \, dx = \int K(u) \, du = 1
\]

\[
\int \hat{f}_{Hist}(x) = \int \frac{1}{N} \sum_i \delta_n(x - x_i) \, dx = \frac{1}{N} \sum_i \int \delta_n(x - x_i) \, dx = 1
\]

• The kernel function can be generalized.

Note: Any density function satisfies our requirements. For example, \( K(.) \) can be a normal density.

• \( \hat{f}_{Hist}(x) \) is a superposition of \( N \) density functions.

KDE: Kernels

• The kernel function \( K(.) \) is a continuous and bounded (usually symmetric around zero) real function which integrates to 1.

• \( h \) is a smoothing parameter (bandwidth). \( 2h \) is called the window width.

• The order of a kernel, \( \nu \), is defined as the order of the first non-zero moment, \( \kappa_\nu \). For example, if \( \kappa_1(K) = 0 \) and \( \kappa_2(K) > 0 \) then \( K \) is a 2\textsuperscript{nd} order kernel. The order of a symmetric kernel is always even.

• Symmetric non-negative kernels are second-order kernels. We will emphasize these kernels (\( \nu=2 \)).

• Higher-order kernels are obtained by multiplying a second-order kernel by an \((2\nu-1)\text{-th}\) order polynomial in \( z^2 \): See Hansen (2009).
KDE: Kernels

- Most common kernel functions:
  - Uniform kernel: \( K(z) = 0.5 \) for \( |z| \leq 1 \)
    \[ = 0 \] for \( |z| > 1 \)
  - Epanechnikov kernel: \( K(z) = 0.75(1-z^2) \) for \( |z| \leq 1 \)
    \[ = 0 \] for \( |z| > 1 \)
  - Gaussian (normal) kernel: \( K(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \)
  - Quartic (biweight) kernel: \( K(z) = \frac{15}{16} (1-z^2)^2 \) for \( |z| \leq 1 \)
    \[ = 0 \] for \( |z| > 1 \)
  - Triweight kernel: \( K(z) = \frac{35}{32} (1-z^2)^3 \) for \( |z| \leq 1 \)
    \[ = 0 \] for \( |z| > 1 \)

- Density graph: Plot \( \hat{f}_{Hist}(x) \) against \( x_0 \) and connect points.

KDE: Kernels - Examples

The Epanechnikov kernel \( K_e(z) = 0.75(1-z^2) I(|z| \leq 1) \) and the Gaussian kernel \( K_g(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \).

The Epanechnikov kernel is not differentiable at \( \pm 1 \).
KDE: Kernels - Examples

• In practice, the choice of the kernel does not matter very much in terms of getting a good approximation to the true density function.

Graph 5: Density of a Normally Distributed Variable, Mean=20, SD=5 (1000 obs., Gaussian kernel, bw=1)

Graph 6: Density of a Normally Distributed Variable, Mean=20, SD=5 (1000 obs., Quartic kernel, bw=2)

KDE: Statistical Inference

• Consistency

In general, for an i.i.d. sample of the RV $X$, for any value $x_0$, $\hat{f}(x_0)$ is a biased estimate of $f(x_0)$. Yet the bias goes to zero if $b \to 0$ as $N \to \infty$.

• The bias depends on $b$, the curvature of $f(\cdot)$, and $K(\cdot)$:

$$bias(\hat{f}(x_0)) = E[\hat{f}(x_0)] - f(x_0) = \frac{1}{2} h^2 f'''(x_0) \int_{-\infty}^{\infty} z^2 K(z) \, dz$$

=> The “size” of this bias is $O(h^2)$.

• Assuming that $b \to 0$ as $N \to \infty$, the variance of $\hat{f}(x_0)$ is:

$$\text{Var}[\hat{f}(x_0)] = \frac{1}{Nh^2} \int f(x_0) \left((K(z))^2 \, dz + o(1/Nh)$$

=> The variance depends on the $N$, $b$, $f(\cdot)$ and $K(\cdot)$. It will go to 0 as $Nh \to \infty$, so $b$ must converge to 0 at a slower rate than $N$ goes to $\infty$. 
The previous results were derived by approximating integrals by a Taylor expansion of $f(x+hu)$ in the argument $hu \rightarrow 0$.

The kernel estimator $f'(x_0)$ is pointwise consistent at any point $x_0$ if both the variance and bias disappear as $N \rightarrow \infty$, which requires that $h \rightarrow 0$ and $Nh \rightarrow \infty$.

The uniform convergence (stronger) property holds if $Nh/\ln(h) \rightarrow \infty$.

See Cameron and Trivedi’s (CT) textbook for formal details.

KDE: Statistical Inference

- Asymptotic normality

The kernel estimator is the sample average. A CLT can be applied. Using previous results:
- Given the order of the variance, the rate of convergence is $\sqrt{Nh}$, not $\sqrt{N}$ as in standard regression estimates.
- The estimator is biased, so we center $f'(x_0)$ around its expectation.

That is, by the CLT we get:

$$\sqrt{Nh} \ (f'(x_0) - E[f'(x_0)]) \rightarrow^d N(0, \ f(x_0) \int (K(z))^2 dz)$$

Note: Given the bias, $[f'(x_0) - f'(x_0)]$ is also asymptotically normally distributed, but with a non-zero mean.
**KDE: Bandwidth**

• As the previous formulas show, there is a genuine trade-off between avoiding bias and reducing the variance of the estimate at any given point $x$.

• In general, large $b$ reduce the variance by smoothing over a large number of points, but this is likely to lead to bias because the points are “averaged” in a mechanical way that does not account for the particular shape of the distribution.

• In contrast, small $b$ give higher variance but have less bias. In the limit, $b \to 0$, the kernel reproduced the data.

**KDE: Bandwidth - Examples**
Q: How do we deal with the trade-off between bias and variance?
A: A “natural” approach is to minimize the MSE:

\[
\text{MSE}(f^*(x_0)) = \text{Var}[f^*(x_0)] + [\text{bias}(f^*(x_0))]^2
\]

\[\Rightarrow \text{optimal bandwidth}\]
KDE: Bandwidth - Selection

• A “natural” approach is to minimize the MSE:
  \[ \text{MSE}(\hat{f}(x_0)) = \text{Var}[\hat{f}(x_0)] + [\text{bias}(\hat{f}(x_0))]^2 \]

• As shown in previous formulas, the bias is \( O(h^2) \) and the variance is \( O(1/Nh^2) \). Intuitively, \( h \) should be chosen to make the (bias)^2 and the variance are of the same order.

• The square of the bias is \( O(h^4) \) => \( h^4 = 1/Nh \), => \( h = (1/N)^{1/5} \).
  That is, \( b = O(N^{-0.2}) \) and \( \sqrt{Nh} = O(N^{0.4}) \).

• A more formal derivation is given on C&T.

• Note: Since the MSE is approximated using asymptotic expansion, it is called AMSE (asymptotic MSE).

KDE: Bandwidth and MISE

• Rosenblatt (1956) developed the idea of a global measure of accuracy for \( \hat{f}(x_0) \): Minimizing the SSE at a very large number of hypothetical points. As the number of points goes to infinity, this amounts to minimizing the mean of the integrated squared error (MISE). If the previous asymptotic approximations are used, the MISE becomes AMISE.

• That is, an optimal bandwidth minimizes
  \[ \text{MISE}(b) = E[ \int (\hat{f}(x_0) - f(x_0))^2 \, dx_0 ] = E[ \int \text{MSE}(\hat{f}(x_0)) \, dx_0 ] \]

• Differentiating AMISE(b) w.r.t. \( b \) yields the optimal bandwidth:
  \[ b^* = \delta \left[ \int (\hat{f}(x_0))^2 \, dx_0 \right]^{-0.2} N^{-0.2} \]
  where \( \delta \) depends on the kernel function used:
  \[ \delta = \left[ \int (K(z))^2 \, dz \right]^{0.2} \left[ \int z^2 K(z) \, dz \right]^{0.4} \]

• Note: \( \int (K(z))^2 \, dz \) is called the roughness of \( K(.) \).
KDE: Optimal Bandwidth

- The optimal bandwidth, \( h^* \):
  \[
  h^* = \delta \left[ \int (f''(x_0)^2 \, dx_0) \right]^{0.2} N^{-0.2}.
  \]

- The optimal bandwidth decreases (very slowly) as \( N \) increases. Then, \( h^* \to 0 \) as \( N \to \infty \) (as required for consistency).

- \( h^* \) depends on \( \delta \), which is a function of the kernel \( K() \). For example, if \( K() \) is Gaussian,
  \[
  \delta = \left[ \frac{1}{2} \int (K(z))^2 \, dz \right]^{0.2} \left[ \int z^2 K(z) \, dz \right]^{0.4} = \left[ \frac{1}{2} \left( \frac{1}{\sqrt{\pi}} \right) \right]^{0.2} \left[ \frac{1}{\sigma^2} = 1 \right]^{-0.4}
  \]
  \[
  = \left( \frac{1}{2} \times \frac{1}{\sqrt{\pi}} \right)^{0.2} \times 0.776388
  \]
  \[
  = \left( \frac{1}{2} \times \frac{1}{\sqrt{\pi}} \right)^{0.2} \times 0.776388
  \]

- Values for \( \delta \) are given in Table 9.1 in C&T.

- This result also shows that if the true density function has a lot of curvature \( (f''(x) \text{ is large}) \), the bandwidth should be smaller.

KDE: Optimality

- The optimal \( h^* \) is unknown – we do not know \( f(x_0) \) or \( f''(x_0) \). Approximations methods are required. In practice, a normal density is commonly used instead of \( f(x_0) \).

- As seen in the graphs, the choice of the kernel matters very little. More formally, \( \text{MISE}(h^*) \) varies little across the different kernels.

- Technically speaking we can select the best kernel. The one that minimizes the AMISE. It is a calculus of variation problem.

- The Epanechnikov (1969) kernel is “optimal,” but the advantage is tiny.

- Since the Epanechnikov kernel is the best, it is used to judge the efficiency of a kernel.
KDE in Practice

• As mentioned above to calculate \( h^* \) we need the unknown \( f''(x_0) \). Approximations methods are required. In practice, a normal density is commonly used instead of \( f(x) \).

1. If \( X \) is normally distributed, then we get

\[
\int (f''(x_0)^2 \, dx_0)^{0.2} = 1.3643
\]

\[
b^* = \frac{\delta}{\{ \int (f''(x_0)^2 \, dx_0 \}^{0.2} N^{0.2}} = 1.3643 \, \delta \, N^{0.2}, \quad \sigma = \text{SD}(x).
\]

If in addition, \( K(.) \) is normal (\( \delta = 0.776388 \)) => \( b^* = 1.059 N^{0.2} \).

If in addition, \( K(.) \) is the Epanechnikov => \( b^* = 2.34 N^{0.2} \).

2. A refinement of the formula in (1), to account for outliers, is

\[
b^* = 1.3643 \, \delta \, N^{0.2} \min[\sigma, \text{iqr}/1.349],
\]

where “iqr” is the (sample) interquartile range.

• These rules for selecting \( h^* \) are generally called rules of thimb.

KDE in Practice

3. These formulas seem to work well even if \( X \) is not normally distributed, especially if \( X \) is symmetric and unimodal.

When \( f(x) \) is multimodal or skewed, the above \( b^* \) will oversmooth.

4. There is also a “cross-validation” (CV) approach. CV attempts to make a direct estimate of the squared error, and pick the \( b \) which minimizes this estimate.

As \( \text{MISE}(b) \) is unknown, CV replaces it with an estimate. The goal is to find an estimate of \( \text{MISE}(b) \), and find the \( b \) which minimizes this estimate. See C&T.

• Try different \( b \)’s and, if anything lean toward a smaller \( b \) (more “jerky” estimates of the density function of \( x \)).
KDE: Confidence Intervals

• We can obtain confidence intervals for estimates of \( f(x_0) \) for any point \( x_0 \). Use the variance formula above to get the conventional C.I.:

\[
f(x_0) \in f'(x_0) - \text{bias}(x_0) \pm z_{\alpha/2} \sqrt{\frac{1}{Nh} f'(x_0) \int (K(z))^2 \, dz}}
\]

where bias(\( x_0 \)) is given above and we have assumed that \( f'(x_0) \) is asymptotically normal.

• Problem: It can contain negative values.
Solution: Consider constructing the C.I. by inverting a test statistic.

\[
C(x) = \{ f : |f'(x_0) - \text{bias}(x_0)| / \sqrt{\frac{1}{Nh} f'(x_0) \int (K(z))^2 \, dz}} \leq z_{\alpha/2}} \}
\]

This set must be found numerically.

• In practice, it is hard to calculate the bias, and, there may not be a reason to calculate the C.I. for \( f'(x_0) \).

Estimating the Derivative of a Density

• Sometimes we need to estimate \( f'(x) \) or even \( f''(x) \) –like in a C.I.

• One approach for estimating \( f'(x) \) is straightforward:

\[
f'(x) = \frac{[f'(x_0 + \Delta) - f'(x_0 - \Delta)]}{2\Delta},
\]

for some small \( \Delta \).

• Alternatively, differentiate the expression \( \hat{f}(x_0) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right) \) with respect to \( x_0 \):

\[
\hat{f}'(x_0) = \frac{1}{Nh^2} \sum_{i=1}^{N} K'(\frac{x_i - x_0}{h})
\]

• We can extend this approach to get the \( r \)-th derivative:

\[
\hat{f}^{(r)}(x_0) = \frac{1}{Nh^{r+1}} \sum_{i=1}^{N} K^{(r)}\left(\frac{x_i - x_0}{h}\right)
\]
Estimating the Derivative of a Density

- Since the Gaussian kernel has derivatives of all orders this is a common choice for derivative estimation.

- The estimator of $f^{(r)}(x)$ is biased – same order as for estimation of $f(x)$. But the variance is of a much larger order.

- We can derive an optimal bandwidth, we can optimize $\text{MISE}(b)$, as before.

- In practice, for either method, you should use a larger bandwidth than you use for estimating $f'(x_0)$.

- We can also ask the question of which kernel function is optimal. Muller (1984) found that the Biweight class is the optimal for the first derivative and for a second derivative the Triweight class.

Density Estimation: $k$-Nearest Neighbor

- In K-Nearest Neighbor ($k$-NN), instead of fixing bin width $b$ and counting the number of instances, we fix the instances (neighbors) $k$ and check bin width.

- The neighborhood is defined through those $X$-variables which are among the $k$-nearest neighbors of $x$. The $k$-NN estimator is given by:

$$\hat{f}_{\text{Hist}}(x_0) = \frac{k}{2Nd_k(x_0)}$$

where $d_k(x)$ is the ordered distance to $k$-th closest instance to $x$. $d_k(x)$ is usually the Euclidian distance (others, OK: Minkowski, Manhattan).

- Intuitively, we allow the bandwidth to vary depending on the density of the function. At areas of low density, we use a higher bandwidth to average over a larger number of (dispersed) points.
Density Estimation: $k$-Nearest Neighbor

**Figure 4.12.** Several $k$-nearest-neighbor estimates of two unidimensional densities: a Gaussian and a bimodal distribution. Notice how the finite $n$ estimates can be quite “spiky.” From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.
Density Estimation: Kernel or $k$-NN?

- The asymptotic analysis of the $k$-NN estimator are complicated by the fact that $d_k(x)$ is random. The solution is to condition on $d_k(x)$, which is similar to treating it as fixed.

- Then, the conditional bias and variance are identical to those of the standard kernel estimator.

- For the unconditional bias, we need moments of $d_k(x)$, under Euclidian distance, given by the $k$-th order statistics. It turns out that the MSE behaves similarly to the kernel estimator’s MSE.

Q: Which one is better?

Not clear. In the tails, the Kernel estimator has smaller bias, but larger variance (the $k$-NN tends to be smoother in the tails).

Density Estimation: Multivariate Case

- Now suppose that $X$ is a $d$-vector and we want to estimate its density $f(x) = f(x_1;..., x_d)$. Easy to extend the idea to this multivariate cases. Computations and interpretation get complicated once we move beyond 3 dimensions.

- Multivariate Kernel density estimator

$$
\hat{f}(x) = \frac{1}{Nh^d} \sum_{t=1}^{N} K\left(\frac{x - x'}{h}\right)
$$

- Multivariate Gaussian kernel
  - spheric
  $$
  K(u) = \left(\frac{1}{\sqrt{2\pi}}\right)^d \exp\left[-\frac{\|u\|^2}{2}\right]
  $$
  - ellipsoid
  $$
  K(u) = \frac{1}{(2\pi)^{d/2} |S|^{1/2}} \exp\left[-\frac{1}{2} u^T S^{-1} u\right]
  $$
Density Estimation: Multivariate Case

• We can estimate an optimal bandwidth, as before, by minimizing MISE(\(h\)). The optimal bandwidth for the \(j\)-th variable is:

\[
h_j^* = \delta \left[ \int (f''(x_0))^2 \, dx_0 \right]^{-0.2} \, N^{-0.2} = C_v(K,d) \, N^{-1/(2r+d)} \, s,
\]

• The optimal bandwidths will all be of order \(N^{1/(2r+d)}\) and the optimal MISE of order \(N^{-2r/(2r+d)}\). This rates are slower than the univariate \((d = 1)\) case.

• The fact that dimension has an adverse effect on convergence rates is called the curse of dimensionality.

• Rules of thumb can be derived for the constant \(C_v(K,d)\).

For example, for the Epanechnikov kernel, \(C_v=2(K,d)\), is for \(d=2\), 2.20, for \(d=3\), 2.12; for \(d=3\), 2.07.

Readings


• Hardle (1990), *Applied Nonparametric Regression*
