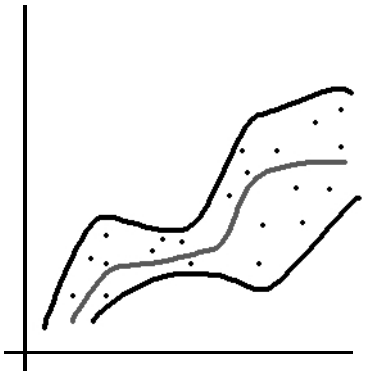
Econometrics 6027
Lecture 9
Nonparametric Estimation Methods



Say you have the data shown in this graph. This is highly nonlinear and the nonlinearity may not be captured by parameters in a conventional regression. Even quantiles will not capture the shape of this data. The average individual moves, as shown by the grey line.

An alternative way to analyze economic relationships is to use nonparametric regression analysis, which allows the data to determine the local shape of the conditional mean relationship. Highly nonlinear relationships are better estimated by nonparametric regressions because the latter better capture the shape of the relationship. Nonparametric models are of the form:

$$y = m(x) + u$$

where $E(u|x) = 0$ and $m(x)$ does not have a parametric representation.

In general, such models build a distribution by estimating the density to the width of a bandwidth around a point. We will focus on the kernel estimator, which assigns weights to data based on their distance from the point at the centre of the bandwidth.

1 Deriving the Kernel Estimator

We have $y = m(x) + u$, so:

$$\begin{aligned} E(y|x) &= E(m(x)) + E(u|x) \text{ but } E(u|x) = 0 \\ &= E(m(x)) \\ &= \int y * f(y|x)dy \\ f(y|x) &= \frac{f(y,x)}{f(x)} \end{aligned}$$

This is a population measure. We need a method to estimate these joint and marginal densities from a sample.

Let $X_1, X_2, ...X_n$ be iid random variables with continuous distribution function $F(x) = \Pr(X_i \leq x)$. Recall that $\Pr(X_i \leq x) = E\left[I\left(X_i \leq x\right)\right]$, where $I\left(.\right)$ is the indicator function. Replacing the population expectation by its sample analog gives the empirical function

$$\widehat{F}\left(x\right) = \frac{1}{n}\sum_{i=1}^{n} I\left(X_i \leq x\right),$$

which estimates the unknown distribution by counting the number of observations that are less than or equal to x. If X has a discrete distribution, then:

$$\widehat{F}\left(x\right) = \frac{1}{n}\sum_{i=1}^{n} I\left(X_i = x\right).$$

But how to estimate the pdf of x, $f\left(x\right)$? One natural estimator is the histogram estimator which computes the difference between two points of the empirical distribution function

$$\widehat{F}\left(x+h\right) - \widehat{F}\left(x-h\right) = \frac{1}{n}\sum_{i=1}^{n} I\left(x-h \leq X_i \leq x+h\right) \tag{1}$$

where h is the bandwidth. If we divide equation **??** by the width $2h$ and let $h \to 0$, we have the definition of the density

$$f(x) = \lim_{h\to 0} \frac{F\left(x+h\right) - F\left(x-h\right)}{2h}$$

The histogram density estimator is obtained by rewriting the above as

$$\widehat{f_h}\left(x\right) = \frac{1}{nh}\sum_{i=1}^{n} \frac{1}{2} I\left(\left|\frac{x-X_i}{h}\right| \leq 1\right)$$

Where the term within brackets is:

$$\left|\frac{x-X_i}{h}\right| = \begin{cases} \frac{x-X_i}{h} \leq 1 \text{ if } x - X_i \leq h \\ -\frac{x-X_i}{h} \leq 1 \text{ if } -x + X_i \leq h \end{cases}$$

This is a histogram estimator. You choose h, the bandwidth, which specifies the width of these rectangular histograms in the data. There are several drawbacks of this estimator. First, it is not smooth (differentiable) whereas the true density is smooth. Second, it weighs equally all points in the interval $[x-h, x+h]$ although we may prefer to assign more weight to points close to x.

A more appealing estimator is the **kernel density estimator** given by

$$\widehat{f_h}\left(x\right) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)$$

where h is the bandwidth parameter and $K(u)$ is the kernel (window) function. Note that the histogram estimator is also a kernel estimator with a kernel function $K(u) = \frac{1}{2} I(|u| \leq 1)$.

The kernel function is a weight function that puts different weights on different points. Typically, it puts more weight on points near x and the weights decline as X_i gets farther away from x. "Near" and "far" from x is determined by the bandwidth parameter h. If h is large, we choose a large neighbourhood around x and the density estimate could be too smooth. If h is small, we take only a small neighbouhood around x and this may result in an estimate that is too variable. We will discuss the optimal choice of h and its effect on the bias and the variance of $\widehat{f_h}(x)$ later in the lecture.

1.1 Common Kernel Functions

Some commonly used kernels are:

1. Gaussian

$$K(u) = (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{u^2}{2}\right)$$

2. Epanechnikov

$$K(u) = \frac{3}{4}\left(1 - u^2\right) I(|u| \leq 1)$$

   Note that the Epanechnikov kernel is quadratic between -1 and 1.

3. Quartic

$$K(u) = \frac{15}{16}\left(1 - u^2\right)^2 I(|u| \leq 1)$$

   The quartic kernel is a 4th order kernel and is square of the Epanechnikov kernel (up to a constant).

1.2 Properties of the Kernel Estimator

1. continuous
   $K(u)$ is continuous

2. symmetric
   $K(u) = K(-u)$ around 0

3. behaves like a density
   $\int_{-\infty}^{\infty} K(u)du = 1$

4. variance is finite
   - $0 < \int u^2 K(u)du < \infty$.

The order of a kernel is the smallest positive integer p such that $\int_{-\infty}^{\infty} K(u)du \neq 0$. If $K(u)$ is a proper pdf, then $p = 2$. The standard kernels are second-order kernels. Higher-order kernels ($p > 2$) require $K(u) < 0$ for some u which is not possible with density functions.

2 Choosing the Bandwidth

The choice of h requires a trade-off between *bias* and *variance*. If we choose a very small h, then bias is small but variance is large. The density estimate is not very smooth. If we choose a very large h, then the bias is large but variance is small. We get an oversmoothed density that distorts the shape of the true density.

There are three ways to choose h:

1. minimizing the Integrated Mean Square Error (IMSE),
2. the Silverman Rule, and
3. Cross-Validation.

We now examine each in turn.

2.1 Minimize Integrated Mean Square Error (IMSE)

One way to determine the optimal choice of h is by minimizing the integrated mean squared error (IMSE). If the Mean Square Error (MSE) is:

$$MSE = (Bias(\hat{f}))^2 + Var(\hat{f})$$

Then IMSE is:

$$IMSE = \int [(Bias(\hat{f}))^2 + Var(\hat{f})^2]dx$$

The bias and variance are defined as follows:

$$Bias(\hat{f}) = \frac{h^2}{2}\frac{\partial^2 f}{\partial x^2}\int u^2 K(u)du$$

$$(Bias(\hat{f}))^2 = \frac{h^4}{4}(\frac{\partial^2 f}{\partial x^2})^2(\int u^2 K(u)du)^2$$

$$Var(\hat{f}_h(x)) = \frac{1}{nh}f(x)\int K^2(u)du$$

assuming that $f(x)$ is twice continuously differentiable. Using these expressions for the bias and the variance, we can find IMSE by integrating across x:

$$\begin{aligned} IMSE &= \frac{h^4}{4}\left[\int u^2 K(u)du\right]^2 \int \left[\frac{\partial^2 f(X)}{\partial X^2}\right]^2 dx + \frac{1}{nh}\int K^2(u)du \int f(x)dx \\ &= \frac{h^4}{4}\left[\int u^2 K(u)du\right]^2 \int \left[\frac{\partial^2 f(X)}{\partial X^2}\right]^2 dx + \frac{1}{nh}\int K^2(u)du \end{aligned}$$

Since $\int f(x)dx = 1$.

Minimizing IMSE with respect to h, the first order condition is:

$$FOC(IMSE) = \frac{\partial IMSE}{\partial h} = 0 \rightarrow h$$

$$\frac{\partial IMSE}{\partial h} = \frac{4h^3}{4}\left(\int u^2 K(u)du\right)^2 \int \left(\frac{\partial^2 f}{\partial x^2}\right)^2 dx - \frac{1}{nh^2}\int K^2(u)du = 0$$

$$h^5 = \frac{1}{n}\left[\frac{\int K^2(u)du}{\left(\int u^2 K(u)du\right)^2 \int \left(\frac{\partial^2 f}{\partial X^2}\right)^2 dx}\right]$$

Define C such that the expression in $[.] = C^5$. Then,

$$h = \left(\frac{1}{n}\right)^{1/5} C$$
$$= Cn^{-1/5}$$

2.2 Silverman Rule

$h = \hat{\sigma} C_v(k) n^{-1/5}$ if $X \sim N(\mu, \sigma^2)$

$K(u)$ Gaussian

$h \approx 1.059\sigma n^{-1/5}$

If asymmetric, $h \approx 0.9An^{-1/5}$

If $f(x)$ is the normal density $N\left(\mu, \sigma^2\right)$ and $K\left(u\right)$ is the Gaussian kernel, direct calculation shows that the optimal bandwidth is given by $h^* \approx 1.059\sigma n^{-\frac{1}{5}}$. This is the "plug-in" rule typically used in empirical work. If the density is bimodal or very skewed, two alternative choices of h are $h = .9Rn^{-\frac{1}{5}}$, where R is the interquartile range or $h = .9An^{-\frac{1}{5}}, A = \min\left(\sigma, \frac{R}{1.34}\right)$. Similarly, by minimizing $IMSE$ with respect to $K\left(u\right)$ yields the Epanechnikov kernel, suggesting that it is the optimal kernel.

2.3 Cross-Validation

The value of the bandwidth parameter can be selected through cross validation. Let $\hat{m_{h_j}}(X)$ denote the kernel estimate obtained from leaving out observation j and then fitting a kernel of Y on X using all the remaining $n-1$ points. The cross-validation function is:

$$CV(h) = \frac{1}{n}\sum_{j=1}^{n}[Y_j - \hat{m_{h_j}}(X)]^2$$

Then we choose h that minimizes $CV(h)$ over a grid of different values for h.

It has been found that in practice the choice of the kernel does not affect the results substantially whereas the choice of the bandwidth is crucial. If we choose a very small h, then the bias is small but the variance is large and since there are no enough points for averaging or smoothing, we get an undersmoothed density estimate. By contrast, if we choose a very large h, then the bias is large but the variance is small, and we get an oversmoothed density estimate that may lead to significant distortions in the shape of the true density (e.g., a bimodal density can be estimated as unimodal).

3 Asymptotic Properties

Recall that the order of a kernel is the smallest positive integer p such that $\int_{-\infty}^{\infty} K(u)du \neq 0$. From the equations for IMSE, we see that $Bias(\hat{f}_h) = O(h^2)$ and $Var = O((nh)^{-1})$. For convergence in mean square, we require that $Bias \to 0$ and $Var \to 0$ as $n \to \infty$.

Consistency
We wish to know if the kernel density estimator $\hat{f}_h(x)$ is consistent. We find that:

$$\text{If } h \to 0 \text{ and } nh \to \infty, \text{ then } \hat{f}_h(x) \xrightarrow{p} f(x)$$

That is, the kernel estimator is consistent under these conditions. This can be derived as follows, given that $(\hat{f}_h(x) - f(x))$ is the bias of the kernel estimator.

Say we normalize by multiplying by $\sqrt{nh}$:

$$\begin{aligned}
\sqrt{nh}E(\hat{f}_h(x) - f(x)) &\approx \sqrt{nh}Bias(\hat{f}_h(x)) \text{ where} \\
Bias(\hat{f}_h(x)) &= \frac{h^2}{2}\frac{\partial^2 f}{\partial x^2}\int u^2 K(u)du \text{ since} \\
h^2 &= \sqrt{h^4} = (h^4)^{1/2}, \text{ so} \\
\sqrt{nh}E(\hat{f}_h(x) - f(x)) &\approx (nh)^{1/2}\frac{(h^4)^{1/2}}{2}\frac{\partial^2 f}{\partial x^2}\int u^2 K(u)du \\
&\approx \sqrt{n}h^{\frac{(4+1)}{2}}\frac{1}{2}\frac{\partial^2 f}{\partial x^2}\int u^2 K(u)du \\
&\approx \sqrt{n}h^{\frac{5}{2}}\frac{1}{2}\frac{\partial^2 f}{\partial x^2}\int u^2 K(u)du \\
&\approx \sqrt{nh^5}\frac{1}{2}\frac{\partial^2 f}{\partial x^2}\int u^2 K(u)du \\
&= O^{rder}\left((nh^5)^{1/2}\right) \\
&= O\left(\sqrt{nh^5}\right) \\
&= O(a)
\end{aligned}$$

Thus, the order of the bias of the kernel estimator is bounded at a constant $a = \sqrt{nh^5}$. This bias is a constant, but we want it to go to zero; otherwise, we could have say that $n \to \infty$ and $h \to 0$ and the bias could explode. If we add the condition that $nh^5 \to 0$, the bias will go to zero. Thus $O^{rder}\left((nh^5)^{1/2}\right)$ goes to zero, on the condition that $nh^5 \to 0$.

Similarly, for consistency, we need the variance to be bounded. We have

$$\begin{aligned}
&Var[\sqrt{nh}(\hat{f}_h(x) - f(x))] \\
&= nhVar(\hat{f}_h(x) - f(x)) \\
&= nhVar(\hat{f}_h(x)) \\
&= nh\left[\frac{1}{nh}f(X)\int K^2(u)du\right] \\
&= f(X)\int K^2(u)du \\
&= O(c) \to \text{ Bounded}
\end{aligned}$$

Thus the variance converges to a constant c given $n \to \infty$, $h \to 0$, $nh \to \infty$, $nh^5 \to 0$).

Asymptotic Normality

We have shown that $\sqrt{nh}(\hat{f}_n(X) - f(X))$ converges in distribution. If $nh^5 \to 0$ then as $h \to 0$ and $n \to \infty$,

$$\sqrt{nh}(\hat{f}_h(X) - f(X)) \to N(0, f(X)fK(u)^2du)$$

Does the non-parametric kernel estimator converge faster or slower than parametric estimators? Parametric estimators converge at a rate of $n^{\frac{1}{2}}$. Recall from minimizing IMSE we found that the optimal bandwidth was $h = cn^{\frac{-1}{5}}$. The kernel converges at a rate of $\sqrt{nh} = \sqrt{n * cn^{\frac{-1}{5}}}$ and since we can eliminate the constant which doesn't change over time, we have the kernel convergence rate as $\sqrt{n^{\frac{5-1}{5}}} = \sqrt{n^{\frac{4}{5}}} = n^{\frac{2}{5}}$.

The kernel rate of convergence is slower than the parametric rate of convergence: $n^{\frac{2}{5}} < n^{\frac{1}{2}}$. There is thus a decrease in asymptotic normality.

4 The Multivariate Kernel Estimator

We can extend the approach outlined above to estimating the *joint* PDF:

$$\hat{f}_h = \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) K\left(\frac{y - Y_i}{h}\right).$$

You can show that $Bias\left(\hat{f}_h(x,y)\right) = O(h^2)$ as before but $Var(\hat{f}_h(x,y)) = O\left((nh^j)^{-1}\right)$, where j is the number of variables ($j = 2$ for bivariate density as in this example).

This is the well-known **Curse of Dimensionality** that plagues nonparametric multivariate estimators. Recall that in parametric models, if we double the sample size we reduce the variance of the estimators by a half regardless of the dimensionality (number of variables). In the nonparametric estimation, this is not the case.

The general form for the kernel density estimator of a $P-m$ dimensional variable x is:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_H(x_i - x)$$

where $K_H = \det(H)^{-1}K(H^{-1}x)$ for some multivariate kernel function $K(x)$ and for a given $P \times P$ matrix of bandwidths H. The simplest multivariate kernel is a product of univariate kernels of the form $K(x) = \prod_{p-1}^{P} k(x_p)$, and a typical matrix of bandwidths would either be diagonal or related to the sample covariance S of the variable x such that $H = h \cdot S^{\frac{1}{2}}$ for some positive scalar h.

When x is univariate, the kernel density estimator reduces to:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x)$$

where $K_h(x) = h^{-1}k(\frac{x}{h})$ for a scalar bandwidth h^2.

5 Kernel Regression

Suppose we have the regression model:

$$Y_i = m(x) + u$$

with $E(u|x) = 0$ but u could be heteroskedastic. We wish to estimate $m(X)$ without parameterizing it. We derive the regression function as:

$$m(X) = E(Y|X) = \int Y f(Y|X)dy = \int \frac{Y f(X,Y)}{f(X)}dY$$

where the numerator $Y f(X,Y)$ is the joint density of X and Y and the denominator $f(X)$ is the marginal density of X. This is equal to:

$$m(x) = \frac{1}{f(X)}\int Y f(X,Y)dY$$

The sample version is:

$$\hat{m}(X) = \frac{1}{\hat{f}_h(X)} \int Y \hat{f}_h(X,Y) dY$$

but we know from the last section that the kernel estimator for a joint PDF is:

$$\widehat{f_h}(X,Y) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) K\left(\frac{y - Y_i}{h}\right).$$

and we found at the beginning of the lecture that the kernel estimator for a univariate PDF was:

$$\widehat{f_h}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

Substituting these expressions into the regression function, we obtain:

$$\begin{aligned}
\hat{m}(X) &= \frac{\frac{1}{nh^2} \int Y \sum K\left(\frac{X-X_i}{h}\right) K\left(\frac{Y-Y_i}{h}\right) dy}{\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X-X_i}{h}\right)} \\
&= \frac{\frac{1}{nh} \sum_i K\left(\frac{X-X_i}{h}\right) \int \frac{1}{h} Y K\left(\frac{Y-Y_i}{h}\right) dY}{\frac{1}{nh} \sum_i K\left(\frac{X-X_i}{h}\right)}
\end{aligned}$$

Using the change of variable technique, we define $v = \frac{Y-Y_i}{h}$ such that $Y = Y_i + hv$ and $dY = hdv$ and substitute. By substitution, we know:

$$YK\left(\frac{Y - Y_i}{h}\right) = (Y_i + hv) * K\left(\frac{Y - Y_i}{h}\right) = (Y_i + hv) * K(v)$$

so:

$$\begin{aligned}
\hat{m}(X) &= \frac{\sum K\left(\frac{X-X_i}{h}\right) \int \frac{1}{h} (Y_i + hv) K(v) h dv}{\sum K\left(\frac{X-X_i}{h}\right)} \\
&= \frac{\sum K\left(\frac{X-X_i}{h}\right) \int (Y_i + hv) K(v) dv}{\sum K\left(\frac{X-X_i}{h}\right)} \\
&= \frac{\sum K\left(\frac{X-X_i}{h}\right) Y_i \int K(v) dv + \sum_i K\left(\frac{X-X_i}{h}\right) h \int v K(v) dv}{\sum K\left(\frac{X-X_i}{h}\right)}
\end{aligned}$$

But $\int K(v)dv = 1$ and $vK(v)dv = 0$. Therefore,

$$\hat{m}(X) = \frac{\sum_i Y_i K\left(\frac{X-X_i}{h}\right)}{\sum_i K\left(\frac{X-X_i}{h}\right)}$$

5.1 Analysis of the Kernel Estimator

This is the Nadaraya-Watson nonparametric Kernel estimator. Note that you still need to choose the h. The more variables we have, the slower convergence we

will have. There are good packages in R and also by Jeff Racine for programming non-parametrics. You can use splines instead of bandwidths for calculation. The CRS program gives you the marginal effect program.

It is interesting to note that if $h = 0$, then

$$\hat{m_h}(X) = \begin{cases} Y_i, \text{ if } & X = X_i \\ 0, \text{ if } & X \neq X_i \end{cases}$$

Also, if $h \to \infty$, then $\hat{m_h} = \overline{Y}$ for all X. Hence, a bigger value of h will produce a smoother (or flatter) $\hat{m_h}(X)$. Therefore, the sensible value of h should lie between these two extremes.

It is also interesting to note that we can use an alternate notation according to which the estimator is rewritten as $\hat{m_h}(X) = \sum_{i=1}^{n} w_i Y_i$ which is a weighted average with weights $w_i = \frac{K\left(\frac{X-X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X-X_i}{h}\right)}$. The motivation behind this form of the estimator is that if $m(X)$ is sufficiently smooth, then in a small neighbourhood of X, $m(X)$ is nearly constant and can be estimated by a weighted average of Y_i's which correspond to those X_i's that are near X (for a given h: if $h = 0$, $M = Y$ for $X = X_i$, 0 otherwise. $h = \infty$, $m = \overline{Y} \epsilon X$).

5.2 Asymptotic Properties

Consistency

If $h \to 0$ and $nh \to \infty$, then $\hat{m_h}(X) \to^p m(X)$.

Asymptotic Normality

If $hn^{\frac{1}{5}} \to 0$ then:

$$\sqrt{nh}(\hat{m_h}(X) - m(X)) \to^d N\left(0, \frac{E(u^2|X)}{f(X)} \int_{-\infty}^{\infty} K(u)^2 du\right)$$

5.3 Multivariate Kernel Regression

Let $m(X, Z) = E(Y|X, Z)$. Then the kernel estimator is given by:

$$\hat{m_h}(X, Z) = \frac{\sum_i Y_i K\left(\frac{X-X_i}{h}\right) K\left(\frac{Z-Z_i}{h}\right)}{K\left(\frac{X-X_i}{h}\right) K\left(\frac{Z-Z_i}{h}\right)}$$

The derivation of this estimator is similar to the derivation of the univariate estimator. Substitute the kernel estimators of the joint densities $f(Y, X, Z)$ and $f(X, Z)$ into $E(Y|X, Z) = \frac{\int Y f(Y,X,Z) fY}{f(X,Z)}$. Analogously to the nonparametric estimation of multivariate densities, this estimator also suffers from the curse of dimensionality and its rate of convergences is given $(nh^j)^{\frac{1}{2}}$, where j is the number of regressors.