ECON 480-3
LECTURE 9: NON-PARAMETRIC REGRESSION

Ivan A. Canay
Northwestern University

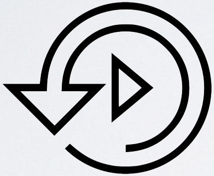
Past & Future

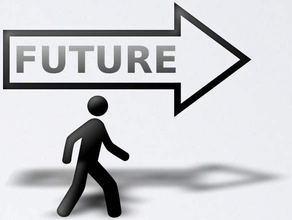
Part I

- Linear Regression
- Properties and Interpretation
- Endogeneity
- Panel Data

Part II: Topics

- Non-parametric Regression
- RDD and Matching
- CART and Random Forest
- LASSO





Setup

- Let (Y, X) be a random vector where Y and X take values in $\mathbf{R}$.
- Let P be the distribution of (Y, X).
- The case where $X \in \mathbf{R}^k$ will be discussed later.
- We are interested in the **conditional mean** of Y given X.

$$m(x) = E[Y|X = x] \; .$$

- Let $\{(Y_1, X_1), \ldots, (Y_n, X_n)\}$ be an i.i.d. sample from P.
- **Discrete case**: If X takes ℓ values $\{x_1, x_2, \ldots, x_\ell\}$, then

$$\hat{m}(x) = \frac{\sum_{i=1}^n I\{X_i = x\} Y_i}{\sum_{i=1}^n I\{X_i = x\}}$$

  is a natural estimator of $m(x)$ for $x \in \{x_1, x_2, \ldots, x_\ell\}$.
- **Straightforward**: $\hat{m}(x)$ is consistent and asymptotically normal if $E[Y^2] < \infty$.

Nearest Neighbor Estimator

- **Continuous X**: the event $\{X_i = x\}$ has zero probability.
- Affects the properties of the previous estimator.
- Assume $m(x)$ is continuous: take average of observations that are "close" to x.

Q-Nearest Neighbor Estimator

Let $J_q(x)$ be the set of indices in $\{1, \ldots, n\}$ associated with q closest-to-x values of $\{X_1, \ldots, X_n\}$. The q-**nearest neighbor** estimator is defined as

$$\hat{m}(x) = \frac{1}{q} \sum_{i \in J_q(x)} Y_i \, .$$

- $J_q(x)$ can be formally defined as follows:

  Let $d_i = |X_i - x|$ and denote by $d_{(1)}, \ldots, d_{(n)}$ the ordered statistics. Then

$$J_q(x) = \{i \in \{1, \ldots, n\} : d_i \leqslant d_{(q)}\} \, .$$

Nearest Neighbor Estimator

- **Step 1**: find the q observations with values of X_i closest to x.
- **Step 2**: average the outcomes of those observations.
- **Intuition**: is $m(x)$ is smooth, it should not change too much as x varies in a small neighborhood.
- $q = n$: we use all of the observations and $\hat{m}(x)$ just becomes $\bar{Y}_n$.

  ... produces a perfectly **flat** estimated function. Variance in very low. Bias is high for many values of x - unless $m(x)$ truly flat.
- $q = 1$: we use X_i very close to x, so the bias should be relatively small. Few obs. so variance is high.
- Could pick q using cross-validation (later).

Binned Estimator

- Flipped side of q-NN estimators.
- q-NN estimator takes an average according to the q observations closest to x.
- The number of "local" observations is always q.
- The distance of these observations is random. In particular,

$$h = \max_{i \in J_q(x)} |X_i - x|$$

  is **random**.
- Alternative approach: fix h and consider all observations with $|X_i - x| \leqslant h$.
- Flipped side: the number of local observations is now random.

Binned Estimator

Binned Estimator

Let $h > 0$ be given. The **binned estimator** is defined as

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} I\{|X_i - x| \leqslant h\} Y_i}{\sum_{i=1}^{n} I\{|X_i - x| \leqslant h\}}\ .$$

- It can be alternatively written as a weighted average estimator

$$\hat{m}(x) = \sum_{i=1}^{n} w_i(x) Y_i$$

with

$$w_i(x) = \frac{I\{|X_i - x| \leqslant h\}}{\sum_{i=1}^{n} I\{|X_i - x| \leqslant h\}}\ .$$

- Note that $\sum_{i=1}^{n} w_i(x) = 1$ so that $\hat{m}(x)$ is a weighted average of Y_i.
- For $x = 2$ and $h = \frac{1}{2}$: $\hat{m}(x)$ is the average of the Y_i for i such that $X_i \in [1.5 \leqslant x \leqslant 2.5]$.

Binned Estimator

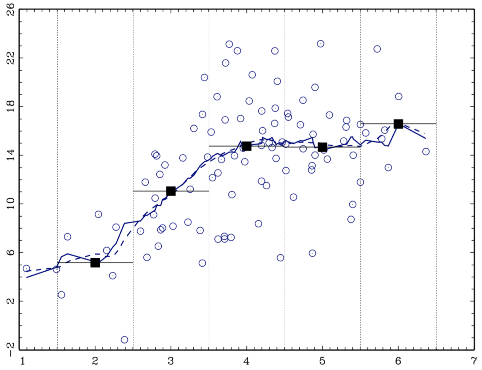


Figure 11.1: Scatter of (y_i, x_i) and Nadaraya-Watson regression

- Coarse grid for x: step-function approximation to $m(x)$. Squares in the figure.

  Introduces jumps in the estimated function at the edges of the partitions.

- Fine grid for x: Evaluate $\hat{m}(x)$ on a fine grid of values (smoother solid line).

QUESTIONS?

Kernel Estimator

- One deficiency with the binned estimator is it is discontinuous at $x = X_i \pm h$.
- Source of discontinuity: weights are based on indicator functions.
- **Idea**: continuous weights may lead to continuous estimators of $m(x)$.
- The family of weights typically used are called "kernels".

Definition (2nd order, non-negative, symmetric kernel)

A **second-order kernel** function $k(u) : \mathbf{R} \to \mathbf{R}$ satisfies

1. $\int_{-\infty}^{\infty} k(u)du = 1$ - definition of kernel
2. $0 \leqslant k(u) < \infty$ - makes the kernel non-negative
3. $k(u) = k(-u)$ - makes the kernel symmetric
4. $\kappa_2 = \int_{-\infty}^{\infty} u^2 k(u)du \in (0, \infty)$ - makes the kernel of order 2

Kernel Estimator

- **Note:** definition of the kernel does not involve continuity.
- **Indeed**: the binned estimator can be written in terms of a kernel function! Let

$$k_0(u) = \frac{1}{2} I\{|u| \leqslant 1\}$$

be the uniform density on $[-1, 1]$.

Note that

$$I\{|X_i - x| \leqslant h\} = I\left\{\frac{|X_i - x|}{h} \leqslant 1\right\} = 2k_0\left(\frac{X_i - x}{h}\right)$$

so that we can write $\hat{m}(x)$ as

$$\hat{m}(x) = \frac{\sum_{i=1}^n k_0\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^n k_0\left(\frac{X_i - x}{h}\right)} .$$

- This is a special case of the so-called **Nadaraya-Watson estimator**.

Nadaraya-Watson

Nadaraya-Watson Kernel Estimator

Let $k(u)$ be a second-order kernel and $h > 0$ be a bandwidth. Then, the **Nadaraya-Watson** estimator is defined as

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right)} .$$

- Also known as kernel regression estimator or local constant estimator.
- The bandwidth $h > 0$ plays the same role as before

  Large h: smoother estimates (but high bias): $h \to \infty \Rightarrow \hat{m}(x) \to \bar{Y}_n$.

  Small h: erratic estimates (but low bias): $h \to 0 \Rightarrow \hat{m}(X_i) \to Y_i$.
- Popular continuous kernels:

$$\textbf{Gaussian}: \quad k_g(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$$

$$\textbf{Epanechnikov}: \quad k_e(u) = \frac{3}{4}(1 - u^2) I\{|u| \leqslant 1\} .$$

Asymp. properties

We wish to show that for each x,

$$\sqrt{nh}(\hat{m}(x) - m(x)) = \sqrt{nh}\Delta_1(x) + \sqrt{nh}\Delta_2(x)$$

where

- $\sqrt{nh}\Delta_2(x)$ is asymp. normal centered at zero (as $nh \to \infty$)
- $\sqrt{nh}\Delta_1(x)$ determines asymptotic bias (as $nh \to \infty$)

- **Question:** why would nh be the rate of convergence?

  ... these are the “effective” number of observations we use.

- The second term adds to the **asymptotic variance**.
- The first term adds to the **asymptotic bias**.
- **Asymptotic framework**: $n \to \infty$, $h \to 0$, $nh \to \infty$.
- Go over sketch of the arguments.

Asymp. properties: expansion

- Write $Y_i = m(X_i) + U_i$ so that $E[U_i|X_i] = 0$ and let $\sigma^2(x) = \text{Var}[U_i|X_i = x]$.
- Fix $x \in \mathbf{R}$ and write
$$Y_i = m(x) + (m(X_i) - m(x)) + U_i \ .$$
- Study the **numerator** of $\hat{m}(x)$:

Asymp. properties: Δ_2

$$\hat{m}(x) - m(x) = \frac{\hat{\Delta}_1(x)}{\hat{f}(x)} + \frac{\hat{\Delta}_2(x)}{\hat{f}(x)} \quad \text{with} \quad \hat{\Delta}_2(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) U_i$$

Find $E[\hat{\Delta}_2(x)]$ and Var $\left[\hat{\Delta}_2(x)\right]$.

Asymp. properties: Δ_2

$$\text{Var}\left[\hat{\Delta}_2(x)\right] = \frac{1}{nh^2}\int_{-\infty}^{\infty} k\left(\frac{z-x}{h}\right)^2 \sigma^2(z)f(z)dz \quad \text{change of vars} \quad u = \frac{1}{h}(z-x)\ .$$

Asymp. properties: Δ_2

The term

$$R(k) = \int_{-\infty}^{\infty} k\,(u)^2\, du$$

denote the roughness of the kernel. Our derivations then lead to

Term $\hat{\Delta}_2(x)$

$$\text{Var}\left[\hat{\Delta}_2(x)\right] = \frac{\sigma^2(x) f(x) R(k)}{nh} + o\left(\frac{1}{nh}\right),$$

and so by the CLT,

$$\sqrt{nh}\hat{\Delta}_2(x) \xrightarrow{d} N\Big(0, \sigma^2(x) f(x) R(k)\Big).$$

Asymp. properties: Δ_1

$$\hat{\Delta}_1 = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) \left(m(X_i) - m(x)\right).$$

Find $E[\hat{\Delta}_1(x)]$

Asymp. properties: Δ_1

$$\int_{-\infty}^{\infty} k(u)\left(m'(x)hu + \frac{h^2u^2}{2}m''(x)\right)(f(x) + uhf'(x))\,du + o(h^2) =$$

$$= \left(\int_{-\infty}^{\infty} uk(u)\,du\right) m'(x)f(x)h + h^2\left(\int_{-\infty}^{\infty} u^2k(u)\,du\right)\left(\frac{1}{2}m''(x)f(x) + m'(x)f'(x)\right) + o(h^2)\,.$$

$$= h^2\kappa_2 f(x)B(x) + o(h^2)$$

- **Notation**:
$$\kappa_2 = \int_{-\infty}^{\infty} u^2k(u)\,du \quad \text{and} \quad B(x) = \left(\frac{1}{2}m''(x) + f^{-1}(x)m'(x)f'(x)\right)\,.$$

- **Variance of $\hat{\Delta}_1(x)$** A similar expansion shows that Var $\left[\hat{\Delta}_1(x)\right] = O\left(\frac{h^2}{nh}\right) = o\left(\frac{1}{nh}\right)$.

Term $\hat{\Delta}_1(x)$

By a triangular array CLT (and some conditions on h),

$$\sqrt{nh}(\hat{\Delta}_1(x) - h^2\kappa_2 f(x)B(x)) \xrightarrow{d} 0\,.$$

Asymptotic Normality

Final step: put all the pieces together and use $\hat{f}(x) \xrightarrow{P} f(x)$.

Asymptotic Normality

Suppose that

1. $f(x)$ is continuously differentiable at the interior point x with $f(x) > 0$.
2. $m(x)$ is twice continuously differentiable at x.
3. $\sigma^2(x) > 0$ is continuous at x.
4. $k(x)$ is a non-negative, symmetric, 2nd order kernel.
5. $E[|Y|^{2+\delta}] < \infty$ for some $\delta > 0$.
6. $n \to \infty$, $h \to 0$, $nh \to \infty$, and $h = O(n^{-1/5})$.

It follows that

$$\sqrt{nh}\left(\hat{m}(x) - m(x) - h^2\kappa_2 B(x)\right) \xrightarrow{d} N\left(0, \frac{\sigma^2(x)R(k)}{f(x)}\right).$$

QUESTIONS?

Asymptotic MSE

Asymptotic MSE

The asymptotic mean squared error of the NW estimator is

$$MSE(x) = h^4 \kappa_2^2 B^2(x) + \frac{\sigma^2(x) R(k)}{nhf(x)} .$$

- Optimal rate h: $Cn^{-1/5}$ with MSE convergence $O(n^{-4/5})$.
- Same rate as in density estimation.
- The constant C is a function of $(\kappa_2, B(x), \sigma^2(x), R(k), f(x))$.
- Plug-in approach possible but cumbersome.
- Other methods, like Cross Validation, may be easier - See Econ 481.

Comments

Kernel Choice

- Asymptotic distribution depends on the kernel through $R(k)$ and κ_2
- Optimal kernel minimizes $R(k)$: same as for density estimation.

  ... Epanechnikov family is also optimal for regression.

On bandwidth choice

- The constant C for the optimal bandwidth depends on the first and second derivatives of the mean function $m(x)$.

  ... when the derivative function $B(x)$ is large, the optimal bandwidths is small.

  ... when the derivative is small, the optimal bandwidth is large.
- For nonparametric regression, reference bandwidths (e.g., Silverman) are not natural.

  ... no natural reference $m(x)$ which dictates the first and second derivative.

Further Comments

Bias Term: needs to be estimated to obtain valid confidence intervals.

- $B(x)$ depends on $m'(x), m''(x), f'(x)$ and $f(x)$. Estimating these objects is arguably more complicated than the problem we started out with.
- Could use a (proper) residual bootstrap. See Econ 481.

Undersmoothing: researchers ignore the bias (arguing it is small)

- To justify this, h should be smaller than optimal.
- Undersmoothing is about choosing h such that

$$\sqrt{nh}h^2 \to 0 \, ,$$

  which makes the bias small, i.e.,

$$\sqrt{nh}h^2 \kappa_2 B(x) \approx 0 \, .$$

- Optimal choice (i.e., Cross validation) are incompatible with the above restriction as

$$nhh^4 \to C > 0 \, .$$

- Undersmoothing does not work well in finite samples. Better methods exist.

Curse of dimensionality

- Let $d_x > 1$ be the dimension of X (we don't use k here to avoid confusion with the kernel).
- NW implementation similar to the one we just described

  ... but requires multivariate kernel and d_x bandwidths.
- The rate of convergence of the NW estimator becomes

$$\sqrt{nh_1 \ldots h_{d_X}} \quad \text{or} \quad \sqrt{nh^{d_x}}\ .$$

- Curse of dimensionality: The higher d_x, the slower the rate.
- Makes sense: it gets harder to find "effective" observations.
- Optimal bandwidths and MSE are

$$h = O(n^{\frac{-1}{4+d_x}}) \text{ and } MSE = O(n^{\frac{-4}{4+d_x}})\ .$$

Limitations of NW

Linear conditional mean

- Suppose $m(x) = \beta_0 + \beta_1 x$. The NW estimator may not perform well here.
- In fact, take $Y_i = \beta_0 + \beta_1 X_i$. (no error)
- NW performs poorly if the marginal distribution of X_i is not roughly uniform.
- NW estimator applied to purely linear data yields a nonlinear output.
- Larger h does not help: makes $\hat{m}(x)$ flatter but not linear.

Boundaries of the support

- The NW estimator may not perform well: bias of order $O(h)$.
- Change of variable argument no longer applies.
- For x s.t. $x \leqslant \min\{X_1, \ldots, X_n\}$; the NW estimator is an average only of Y_i values for observations to the right of x.
- If $m(x)$ is positively sloped, the NW estimator will be upward biased.

  ... the estimator is inconsistent at the boundary.

QUESTIONS?

Local Linear Estimator

- The Nadaraya-Watson estimator is often called a **local constant estimator**.
- It locally (about x) approximates the CEF $m(x)$ as a **constant function**.
- **Interpretation**: $\hat{m}(x)$ solves the minimization problem

$$\hat{m}(x) = \underset{c}{\text{argmin}} \sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right)(Y_i - c)^2 . \tag{1}$$

- This is a weighted regression of Y_i on an intercept only.
- Without the weights, this estimation problem reduces to the sample mean. The NW estimator generalizes this to a "local" mean.
- **Insight:** we may construct alternative nonparametric estimators of $m(x)$ by alternative local approximations.
- A popular choice is the **local linear (LL) approximation**.

Local Linear Estimator

- **Idea:** Instead of approximating $m(x)$ locally as a constant, the local linear approximation approximates $m(x)$ locally by a linear function.
- **Estimation**: use locally weighted least squares.

Local Linear (LL) Estimator

For each x solve the following minimization problem,

$$\{\hat{\beta}_0(x), \hat{\beta}_1(x)\} = \underset{(b_0, b_1)}{\text{argmin}} \sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right)(Y_i - b_0 - b_1(X_i - x))^2 \,. \tag{2}$$

The **local linear estimator** of $m(x)$ is the local **intercept**: $\hat{\beta}_0(x)$.

- The LL estimator of the derivative of $m(x)$ is the estimated slope coefficient:

$$\hat{m}'(x) = \hat{\beta}_1(x) \,.$$

- Note: If we write the local model

$$Y_i = \beta_0 + \beta_1(X_i - x) + U_i \quad \text{with} \quad E[U|X = x] = 0 \,,$$

then using the regressor $X_i - x$ rather than X_i makes the intercept equal to $m(x) = E[Y|X = x]$.

LL estimator: example

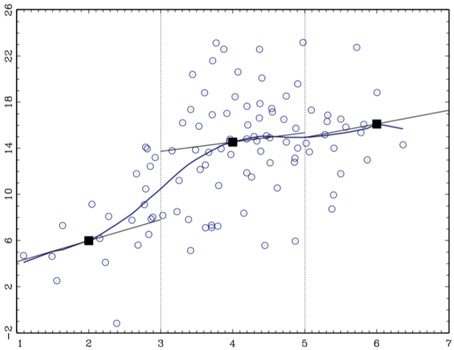


Figure 11.2: Scatter of (y_i, x_i) and Local Linear fitted regression

- Scatter plot divided into three regions depending on the regressor x.
- Linear regression fit in each region, with the obs. weighted by the Epanechnikov kernel with $h = 1$.
- **solid line**: Then $\hat{m}(x)$ is evaluated not only at $x \in \{2, 4, 6\}$ but at a fine grid.

Least Squares formula

- For each x set
$$Z_i(x) = \left(1, X_i - x\right)'$$
and
$$k_i(x) = k\left(\frac{X_i - x}{h}\right) .$$
- Then
$$\begin{pmatrix} \hat{\beta}_0(x) \\ \hat{\beta}_1(x) \end{pmatrix} = \left(\sum_{i=1}^{n} k_i(x) Z_i(x) Z_i(x)'\right)^{-1} \sum_{i=1}^{n} k_i(x) Z_i(x) Y_i .$$
- For each x, the estimator is just weighted least squares of Y in $Z(x)$.
- In fact: as $h \to \infty$, the LL estimator approaches the full-sample linear least-squares estimator
$$\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x .$$
- As $h \to \infty$ all observations receive equal weight regardless of x.
- The LL estimator is a flexible generalization of the linear OLS estimator.

Asymptotic Normality

- Deriving the asymp. distribution of the LL estimator involves similar tools to those used with the NW estimator (but more involved). Skip here.

Asymptotic Normality

Let $\hat{m}(x)$ be the LL estimator as previously defined. Under conditions 1-6 in the NW theorem,

$$\sqrt{nh}\left(\hat{m}(x) - m(x) - h^2\kappa_2 \frac{1}{2} m''(x)\right) \xrightarrow{d} N\left(0, \frac{\sigma^2(x)R(k)}{f(x)}\right) .$$

- Relative to the Bias of the NW estimator,

$$B(x) = \left(\frac{1}{2} m''(x) + f^{-1}(x) m'(x) f'(x)\right)$$

  the second term is no longer present. Simplified bias suggests reduced bias.

- Bias of LL does not depend of $f(x)$: design adaptive.
- In theory, bias could be larger as opposing terms could cancel out.
- Indeed, weaker condition 1: only continuity of $f(x)$ is required - no diff.

NW vs LL: comments

- The LL estimator preserves linear data (in contrast to NW).

  If $Y_i = \beta_0 + \beta_1 X_i$, then for any sub-sample, a local linear regression fits exactly, so $\hat{m}(x) = m(x)$.

- The distribution of the LL estimator is **invariant** to the first derivative of m: it has zero bias when the true regression is linear.

- LL estimator has better properties at the **boundary** than the NW estimator.

  Intuition: even if x is at the boundary, as the LL estimator fits a (weighted) LS line through data near the boundary, if the true relationship is linear this estimator will be unbiased.

  For the LL estimator the order of the bias is $O(h^2)$ at all x (vs $O(h)$ for NW).

- Extensions that allow for discontinuities in $m(x), f(x)$ and $\sigma(x)$ exist.

Further Comments

- LL estimators is perhaps the most popular judged by journal article counts.
- Particularly useful in **RDD settings** (next class).
- The LL estimator does not always beat the NW estimator in simulations.

  If $m(x)$ is quite flat $\Rightarrow$ NW estimator does better.

  If $m(x)$ is steeper and curvier $\Rightarrow$ LL estimator tends to do better.
- **Explanation**: in finite samples the NW estimator tends to have a smaller variance.

  Gives it an advantage when bias is low $\approx m(x)$ is flat.
- **Extension**: LL extends to Local Polynomial estimator. Topic of 481.

THE END!