# Econometrics II

## Lecture 8
## Non- and Semiparametric Methods and Regression Discontinuity Design

Marie Paul

University of Duisburg-Essen
Ruhr Graduate School in Economics

Summer Semester 2020

# Outline of lecture

8. Non- and Semiparametric Methods
   - Kernel density estimation
   - Nonparametric regression
   - Semiparametric regression

9. Regression Discontinuity
   - Sharp RD
   - Fuzzy RD

# Self-study plan for Lecture 8

- This lecture introduces non- and semiparametric methods and regression discontinuity designs.
- For our next virtual meeting, please work through all slides of Lecture 9. Readings:
- Cameron and Trivedi (2005) Chapter 9 and Angrist and Pischke (2009) Chapter 6 as far as this is covered on the slides.
- Please make sure that you really understand how Kernels work (slides 7 and 8, well explained in Cameron and Trivedi).
- Please prepare answers to the "your own research" questions and collect questions and topics to discuss them in our virtual meeting.

# Non- and Semiparametric Methods

**Example 8.1 (Your own research)**

Discuss applications of such methods in your research field. Also consider application in the context of matching or RD.

# Outline

8. Non- and Semiparametric Methods
   - Kernel density estimation
   - Nonparametric regression
   - Semiparametric regression

# Motivation

**Nonparametric estimation methods** for data analysis makes **minimal assumptions** on the data generating process.

Nonparametric methods are essentially **local averaging methods** using "slices" of the data to estimate relationships over distributions.

1. **Kernel density estimation:** Smoother estimates of discrete distributions.
2. **Nonparametric regression:** More flexible regression estimation.

Can be used for detailed data description, exploratory data analysis and for fitting regression models more flexibly to "let the data speak for itself".

**Semi-parametric methods** are common in **multivariate** analyses due to problems of **sparseness** from a **curse of dimensionality** from additional variables.

Semi-parametric methods comprise a specified **parametric** part and an unspecified **nonparametric** part, reducing the dimensionality of the model.

# Histogram density estimator

A **kernel density estimator** smooths the intervals of the distribution of a discrete random variable according to some predetermined **weighting** scheme.

In contrast, a **histogram** splits the range of a random variable $x$ into **equally spaced intervals** containing the respective fraction of the sample.

To compare the approaches, consider estimation of the **density** $f(x_0)$ of a scalar $x$ evaluated at $x_0$

$$\begin{aligned} f(x_0) = \frac{dF(x_0)}{dx_0} &= \lim_{h \to 0} \frac{F(x_0 + h) - F(x_0 - h)}{2h} \\ &= \lim_{h \to 0} \frac{\Pr[x_0 - h < x < x_0 + h]}{2h}. \end{aligned} \tag{8.1}$$

The density can be estimated by the **sample analogue** of (8.1)

$$\widehat{f}_{\mathsf{H}}(x_0) = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbf{1}[x_0 - h < x_i < x_0 + h]}{2h}, \tag{8.2}$$

i.e. the **sample fraction** within $x_0 \pm h$ divided by the **bin width** $2h$.

# Kernel density estimator

Equation (8.2) can be rewritten as a **step function** with equal weights for all observations

$$\widehat{f}_{\mathsf{H}}(x_0) = \frac{1}{Nh} \sum_{i=1}^{N} \frac{1}{2} \times \mathbf{1}\left( \left| \frac{x_i - x_0}{h} \right| < 1 \right). \tag{8.3}$$

Evaluating $\widehat{f}_{\mathsf{H}}$ over the **full range** of $x$ according to (8.3) yields a histogram with bins equal to the number of intervals.

In contrast, the kernel density estimator **generalizes** the histogram density using a different weighting function

$$\widehat{f}(x_0) = \frac{1}{Nh} \sum_{i=1}^{N} K\left( \frac{x_i - x_0}{h} \right), \tag{8.4}$$

where $K(\cdot)$ is a weighting **kernel function**, $h$ is the **bandwidth**, $2h$ the **window width**, and $x$ is evaluated at each sample value, i.e. $x_1, x_2, \ldots x_N$.

# Example: Choice of kernel

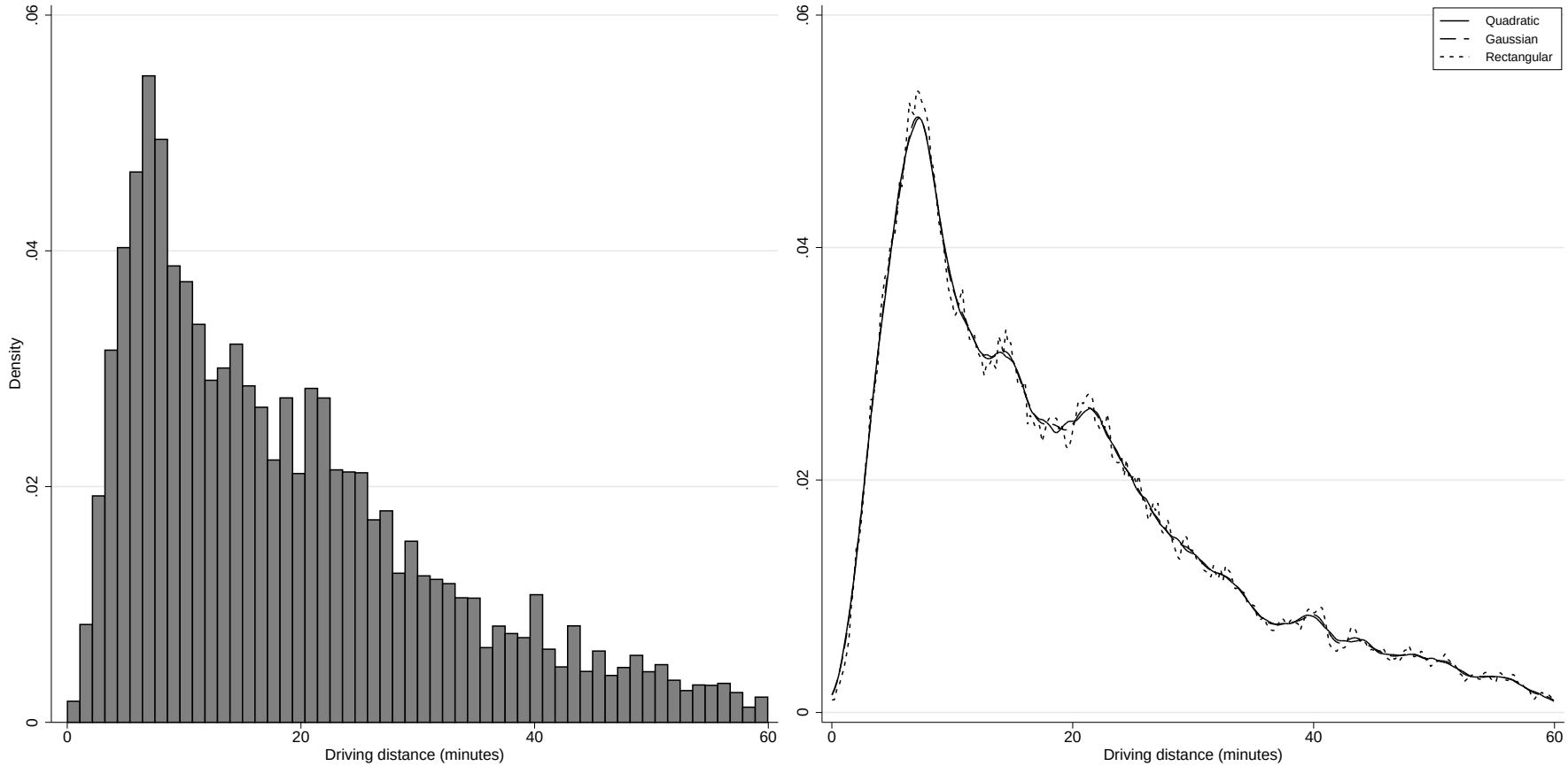



Figure 8.1. Driving distance to the closest hospital in Sweden.

# Kernel functions

## Definition 8.2 (Kernel function)

The **kernel function** $K(\cdot)$ is a **continuous** and **bounded** function, **symmetric** around zero, that **integrates to unity** (Lee, 1996).

(i) $K(z)$ *is symmetric around 0 and continuous.*

(ii) $\int K(z)dz = 1$, $\int zK(z)dz = 0$, *and* $\int |K(z)|dz < \infty$.

(iii) *Either* (a) $K(z) = 0$ *if* $|z| \geqslant z_0$ *for some* $z_0$ *or* (b) $|z|K(z) \to 0$ *as* $|z| \to \infty$.

(iv) $\int z^2 K(z)dz = \kappa$, *where* $\kappa$ *is a constant*.

Given specification of $K(\cdot)$ and $h$, the **kernel density estimator** in (8.4) is straightforward to implement.

In most applications, both the kernel function and bandwidth can be chosen **optimally** to minimize the **mean squared error (MSE)**

$$\mathrm{MSE}[\theta] = \mathrm{E}\left[(\hat{\theta} - \theta)^2\right] = \underbrace{\mathrm{E}\left[\left(\hat{\theta} - \mathrm{E}[(\hat{\theta}]\right)^2\right]}_{\mathrm{V}(\hat{\theta})} + \underbrace{\left(\mathrm{E}[(\hat{\theta}] - \theta\right)^2}_{\mathrm{B}(\theta,\hat{\theta})^2}. \tag{8.5}$$

# Some common kernel functions

Table 8.1. Commonly used kernels

| **Kernel** | **Kernel function K(z)** | $\delta$ |
|---|---|---|
| Uniform | $\frac{1}{2} \times \mathbf{1}[\lvert z \rvert < 1]$ | 1.3510 |
| Triangular | $(1 - \lvert z \rvert) \times \mathbf{1}[\lvert z \rvert < 1]$ | - |
| Epanechnikov | $\frac{3}{4}(1 - z^2) \times \mathbf{1}[\lvert z \rvert < 1]$ | 1.7188 |
| Gaussian | $(2\pi)^{-1/2} \exp(-z^2/2)$ | 0.7764 |

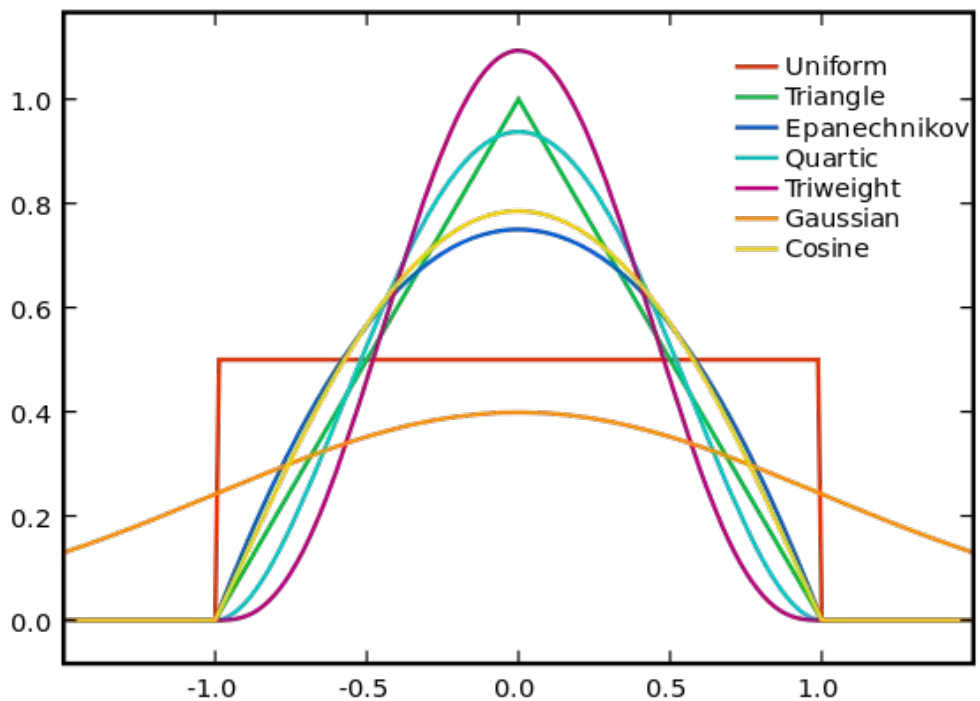



Figure 8.2. Weights of various kernels.

# Inference: Mean

The **mean** of the kernel density estimator given $K(\cdot)$ and bandwidth $h$ is

$$\begin{aligned} \mathsf{E}[\widehat{f}(x_0)] &= \mathsf{E}[\frac{1}{h}K(\frac{x-x_0}{h})] \\ &= \int K(z)f(x_0+hz)dz \end{aligned} \tag{8.6}$$

where $z = (x - x_0)/h$ and $x = x_0 + hz$.

A **second-order Taylor expansion** of $f(x_0 + hz)$ around $f(x_0)$ yields

$$\begin{aligned} \mathsf{E}[\widehat{f}(x_0)] &= \int K(z)\{f(x_0) + f'(x_0)hz + \frac{1}{2}f''(hz)^2\}dz \\ &= f(x_0)\int K(z)dz + hf'(x_0)\int zK(z)dz + \frac{1}{2}h^2 f''(x_0)\int z^2K(z)dz. \end{aligned} \tag{8.7}$$

Using **Definition 8.2** this reduces to the following and provides the formula for the bias.

$$\mathsf{E}[\widehat{f}(x_0)] - f(x_0) = \underbrace{\frac{1}{2}h^2 f''(x_0)\int z^2K(z)dz}_{\mathsf{B}(x_0)=O(h^2)}. \tag{8.8}$$

# Inference: Variance

Similarly, the **variance** of the kernel density estimator is

$$\mathsf{V}[\widehat{f}(x_0)] = \frac{1}{N}\mathsf{E}[(\frac{1}{h}K(\frac{x-x_0}{h}))^2] - \frac{1}{N}(\mathsf{E}[(\frac{1}{h}K(\frac{x-x_0}{h}))])^2. \tag{8.9}$$

The same change of variables and a first-order Taylor expansion yields

$$\mathsf{E}[(\frac{1}{h}K(\frac{x-x_0}{h}))^2] = \frac{1}{h}f(x_0)\int K(z)^2dz + f'(x_0)\int zK(z)^2dz. \tag{8.10}$$

Therefore,

$$\begin{aligned}\mathsf{V}[\widehat{f}(x_0)] = \frac{1}{Nh}f(x_0)\int K(z)^2dz + \frac{1}{N}f'(x_0)\int zK(z)^2dz \\ - \frac{1}{N}[f(x_0) + \frac{1}{2}h^2f''(x_0)\int z^2K(z)dz]^2,\end{aligned} \tag{8.11}$$

Or, (as $h \to 0$ and $N \to \infty$)

$$\mathsf{V}[\widehat{f}(x_0)] = \underbrace{\frac{1}{Nh}f(x_0)\int K(z)^2dz}_{\mathsf{V}(x_0)=O((Nh)^{-1})} + o(\frac{1}{Nh}). \tag{8.12}$$

# Optimal bandwidth and kernel

Choosing bandwidth is a **trade-off** between **precision** and **bias**, since a larger $h$ increases bias from (8.8) but reduces variance from (8.12) and vice versa.

The **mean-squared error (MSE)**, the sum of squared bias and variance, is used to choose bandwidth **optimally**, it is minimized.

In contrast, the **choice of kernel** does not matter much as long as the optimal bandwidth is used — which varies across kernels.

It can be shown that the **optimal kernel** is the **Epanechnikov** but the difference is negligible.

# Optimal bandwidth

Silverman (1986) has shown that (under some assumptions) the **optimal bandwidth** depends also on the **kernel** and on the **curvature** of the density and is given by

$$h^* = \delta \left( \int f''(x_0)^2 dx_0 \right)^{-0.2} N^{-0.2}, \tag{8.13}$$

where $\delta$ is defined by

$$\delta = \left( \frac{\int K(z)^2 dz}{(\int z^2 K(z) dz)^2} \right)^{0.2}. \tag{8.14}$$

This bandwidth minimizes mean integrated squared error (**MISE**). The MSE is a local measure at $x_0$ and the MISE is a **global measure** of performance.

# Optimal bandwidth: Plug-in bandwidth (Silverman's rule)

The last column of Table 8.1 shows the value of the parameter $\delta$ under the assumption that $f(x)$ is **normally distributed**.

Then $\int f''(x_0)^2 dx_0 = 3/(8\sqrt{\pi}\sigma^5)$ and

$$h^* = 1.3643\delta N^{-0.2}\sigma \tag{8.15}$$

from (8.13) where $\delta$ corresponds to the value obtained from evaluating (8.14) under normality of $f(x)$ and given kernel function $K(\cdot)$.

This is useful to generate a **rule of thumb (plug-in) bandwidth** measure. **Silverman's plug-in estimate** is defined by

$$h^* = 1.3643\delta N^{-0.2} \min(s, iqr/1.349) \tag{8.16}$$

where $s$ is the sample standard deviation and $iqr/1.349$ is used to guard against outliers when $s$ is very noisy. iqr is the sample interquartile range.

This is a **practical**, **well-functioning** and **easy to use** bandwidth estimate as it only requires knowledge of $N$, $s$, and iqr. Neverless one should check **alternative bandwidths** as the double and half of the plug-in bandwidth.

# Example: Choice of bandwidth

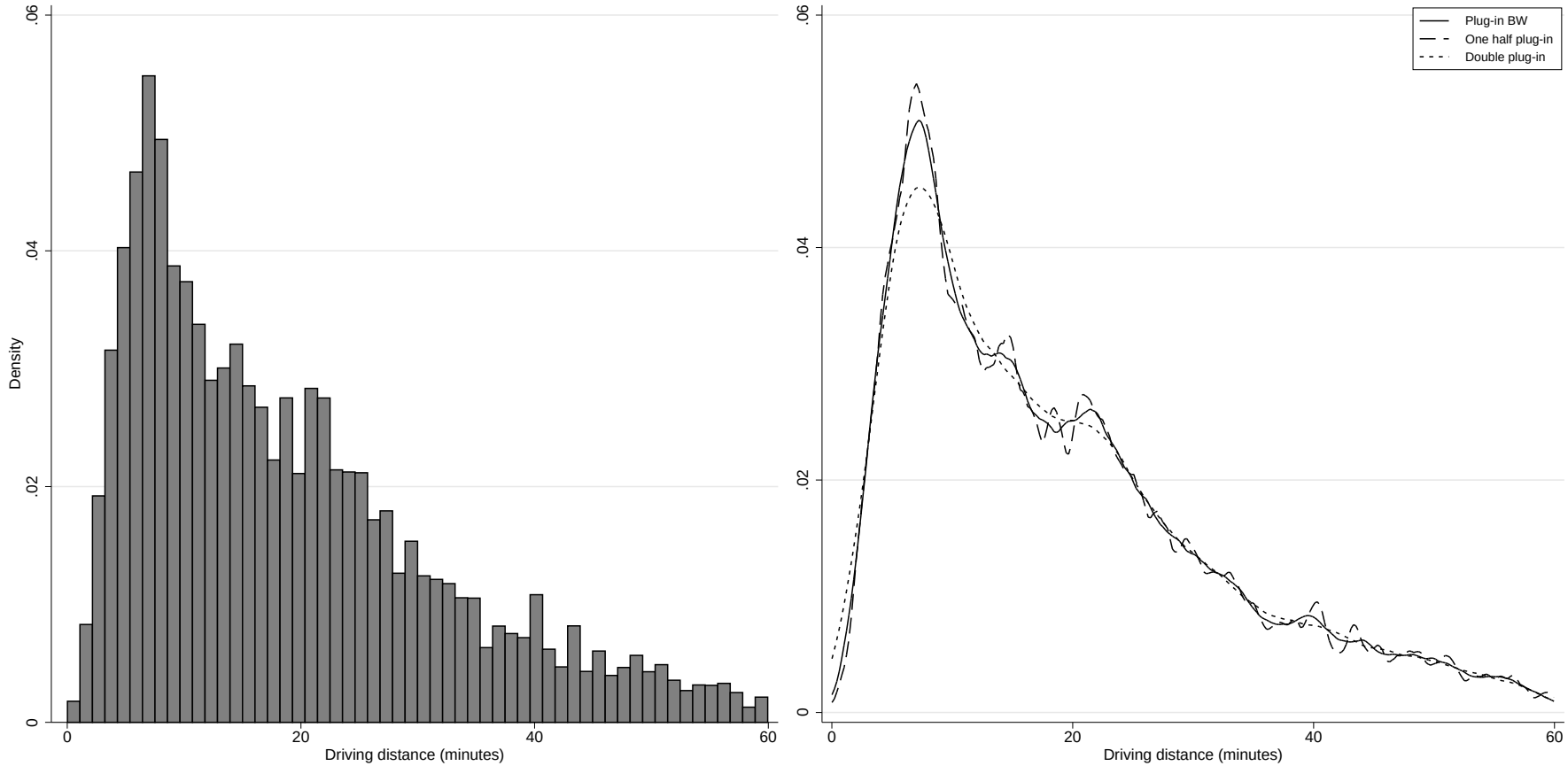



Figure 8.3. Distance to closest hospital in Sweden.

# Outline

8. Non- and Semiparametric Methods
   - Kernel density estimation
   - Nonparametric regression
   - Semiparametric regression

# Nonparametric local regression: Motivation

Now consider instead the **regression** case where we are interested in the relationship between a dependent variable $y$ and a **scalar** regressor $x$

$$\begin{aligned} y_i &= m(x_i) + \varepsilon_i, \quad i = 1, \ldots, N, \\ \varepsilon_i &\sim \text{iid}[0, \sigma_\varepsilon^2], \end{aligned} \tag{8.17}$$

where $m(\cdot)$ is an unknown function.

We can apply **nonparametric regression** by using **local weighted averages** to estimate the relationship.

Suppose we have $N_0$ observations for $y$ for a distinct value $x_0$. Then $N_0^{-1} \sum y = \tilde{m}(x_0)$ is an estimator for $m(\cdot)$ with asymptotic distribution

$$\tilde{m}(x_0) \sim \left[ m(x_0), N_0^{-1} \sigma_\varepsilon^2 \right]. \tag{8.18}$$

The problem is that this estimator is unbiased but **possibly inconsistent** since $N_0 \to \infty$ need to be satisfied as $N \to \infty$, but this is not necessarily the case.

The problem of **sparseness** increases as the regressor becomes "more continuous" as fewer observations will end up inside $x_0$.

# Nonparametric local regression: Motivation

The solution to the sparseness problem is to average observations of $y$ in a **neighborhood** of $x_0$.

Note that $\tilde{m}(x_0)$ in (8.18) can be expressed as

$$\tilde{m}(x_0) = \sum_{i=1}^{N} w_{i0} y_i, \tag{8.19}$$

where **weights** are equal to

$$w_{i0} = \begin{cases} 1/N_0 & \text{if } x_i = x_0, \\ 0 & \text{if } x_i \neq x_0. \end{cases} \tag{8.20}$$

Consider instead the general **local weighted average estimator**

$$\begin{aligned} \hat{m}(x_0) &= \sum_{i=1}^{N} w_{i0,h} y_i, \\ w_{i0,h} &= w(x_i, x_0, h), \end{aligned} \tag{8.21}$$

where the weights sum to one, increasing as $x_i$ approaches $x_0$, and $h$ is a smoothing parameter defining the **window width**.

# Example: Locally weighted regressions

## Example 8.3 (K-nearest neighbors)

Consider the **unweighted** average of $y$ corresponding to the $x_0 \pm (k-1)/2$ **closest** observations of $x$.

Ordering observations by increasing $x$ and evaluating at $x_0 = x_i$ yields the **k-nearest neighbors estimator**

$$\hat{m}_k(x_i) = \frac{1}{k}(y_{i-(k-1)/2} + \cdots + y_{i+(k-1)/2}), \tag{8.22}$$

which is a special case of (8.21) with weights equal to

$$w_{i0,h} = \frac{1}{k} \times \mathbf{1}\left[|i-0| < \frac{k-1}{2}\right], \quad i = x_1 < x_2 < \cdots < x_0 < \cdots < x_N. \tag{8.23}$$

Note that this ignores **tied** values or observations close to the **endpoints**.

# Kernel regression

**Kernel regression** refers to nonparametric regression methods where the weights are obtained using a kernel function.

To estimate $m(x_0)$, consider instead the average of the $y_i$ observations for all $x_i$ observations within a distance of $x_0 \pm h$

$$\widehat{m}(x_0) = \frac{\sum_{i=1}^{N} \mathbf{1}\left[\left|\frac{x_i - x_0}{h}\right| < 1\right] y_i}{\sum_{i=1}^{N} \mathbf{1}\left[\left|\frac{x_i - x_0}{h}\right| < 1\right]}, \tag{8.24}$$

i.e. the **sum** of the $y_i$ values divided by the **number** of observations within the corresponding interval.

Replacing the indicator function by the kernel function defined in (8.4) yields the **kernel regression estimator**

$$\widehat{m}(x_0) = \frac{\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right) y_i}{\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right)}, \tag{8.25}$$

given choice of kernel $K(\cdot)$ and bandwidth $h$.

# Kernel regression

The kernel regression estimator is a special case of the weighted average in (8.21), with weights

$$w_{i0,h} = \frac{\frac{1}{Nh} K\left(\frac{x_i - x_0}{h}\right)}{\frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right)}. \tag{8.26}$$

Just as with the kernel density estimator, the regression estimator is **biased** and an **optimal bandwidth** balances the **trade-off** between increasing bias and decreasing variance using squared error loss.

There is an additional problem in this case however, as the optimal bandwidth will depend also on **unknown functions** (e.g., $m''(x)$).

Instead, a **cross-validation method** can be used to choose $\hat{h}^*$ to minimize estimated predict ion error

$$CV(h) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{m}_{-i}(x_i))^2, \tag{8.27}$$

where $\hat{m}_{-i}$ is a **leave-one-out** estimate of $m(x_i)$ such that $y_i$ is omitted in the estimation.

# Cross-validation

## Definition 8.4 (Cross-validation)

The **cross-validation** approach for finding the optimal bandwidth chooses $\hat{h}^*$ to **minimize** the MSE of the **prediction**

$$CV(h) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{m}_{-i}(x_i))^2, \tag{8.28}$$

where $\hat{m}_{-i}$ is a **leave-one-out** estimate

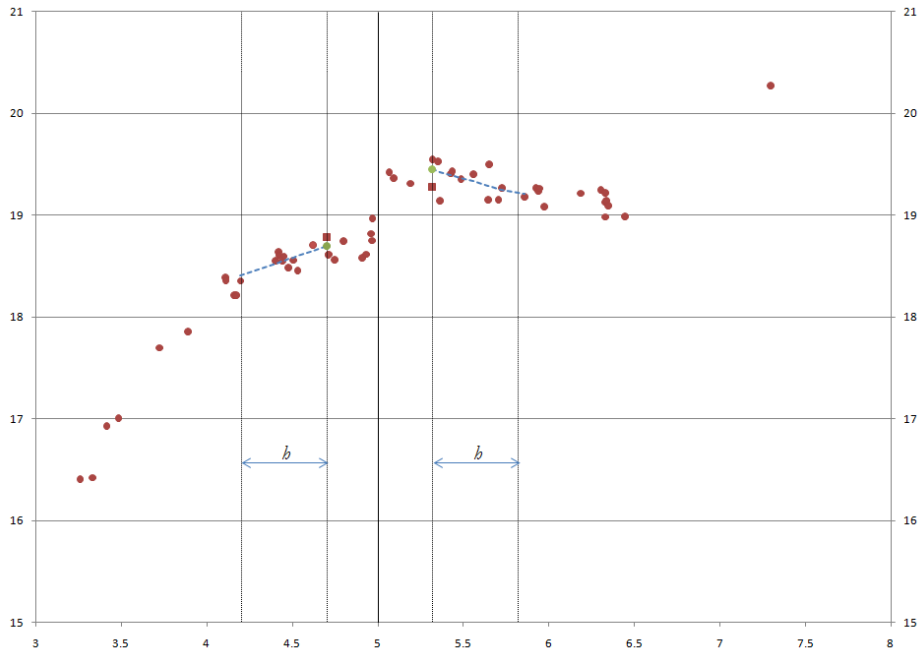
$$\hat{m}_{-i}(x_i) = \frac{\sum_{j\neq i} w_{ji,h} y_j}{\sum_{j\neq i} w_{ji,h}}, \tag{8.29}$$

obtained by leaving $y_i$ out from the kernel regression formula in (8.25).

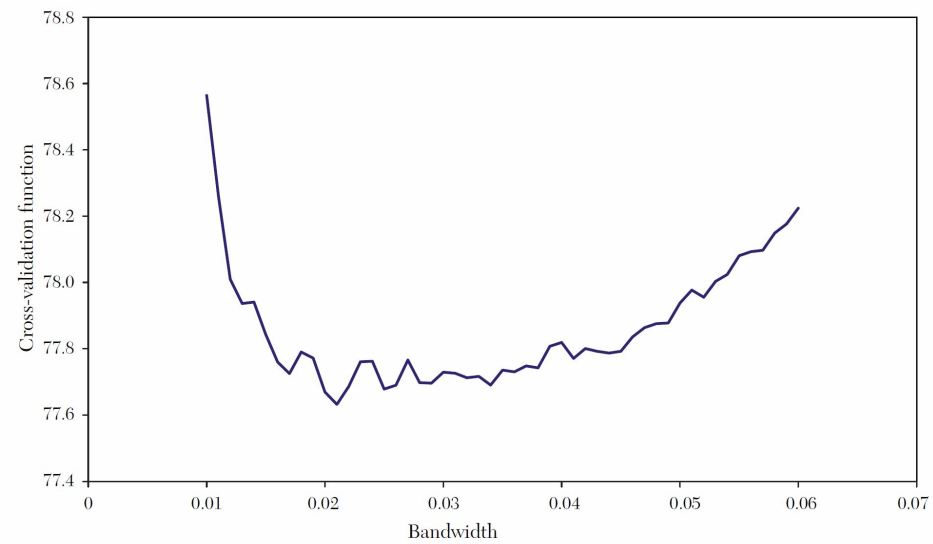
The intuition is to weigh **precision** against **bias** captured by predicting $\hat{m}_{-i}(x_i)$ for all observations given different bandwidths and evaluate

$$h^*_{CV} = \arg\min_h CV(h) \tag{8.30}$$

# Cross-validation illustration




(a) Predictions for bandwidth $h$




(b) MSE for various bandwidths

Figure 8.4. Graphical illustration of bandwidth choice using a cross-validation procedure and a local polynomial regression estimator

# Other common regression estimators

Other popular nonparametric regression estimators include (many more exist)

1. $k$-**Nearest Neighbors:** A kernel estimator with uniform weights but variable bandwidth $h_0 \simeq k/(2Nf(x_0))$.

$$\hat{m}_{k-\mathsf{NN}}(x_0) = \frac{1}{k}\sum_{i=1}^{N} \mathbf{1}[x_i \in N_k(x_0)]y_i. \tag{8.31}$$

2. **Local polynomial regression:** The kernel regression estimator is a **local constant estimator** as $m(x) = c$ in $x_0 \pm h$. Instead define

$$\begin{aligned}\underset{a_{0,s}}{\arg\min} \sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right)(y_i - a_{0,0} - a_{0,1}(x_i - x_0) \\ - \cdots - a_{0,p}\frac{(x_i - x_0)^p}{p!})^2,\end{aligned} \tag{8.32}$$

   i.e., **local polynomial estimator of degree** $p$, yielding the estimator

$$\hat{m}^{(s)}_{\mathsf{LPE}}(x_0) = \hat{a}_{0,s}. \tag{8.33}$$

# Example: Nonparametric regression

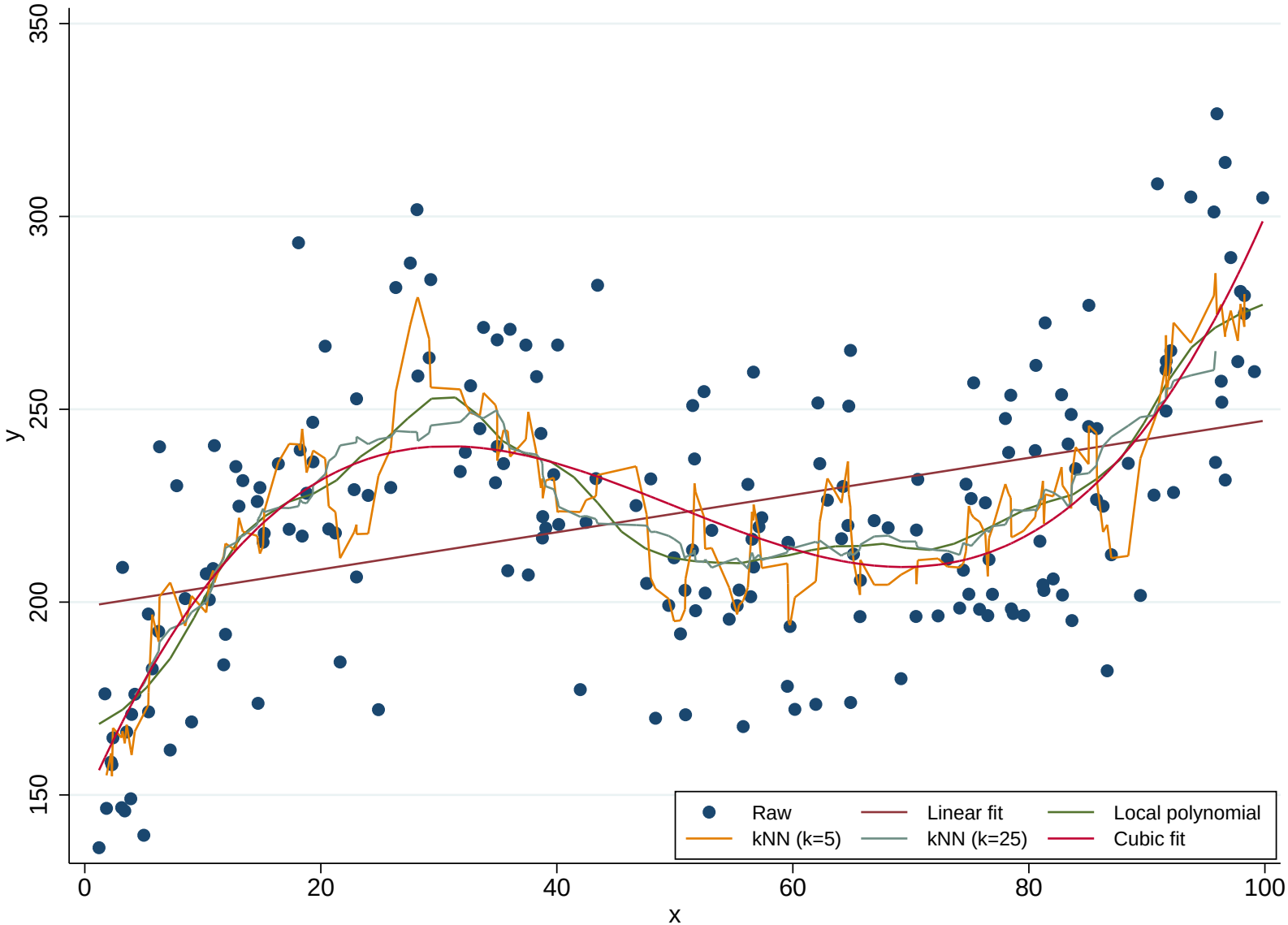



Figure 8.5. Various parametric and non-parametric estimates for the model $y = 150 + 6.5x - 0.15x^2 + 0.001x^3 + \varepsilon, \quad \varepsilon \sim [0, 25^2]$.

# Kernel regression: The multivariate case

Now consider the **multivariate case** of the regression of scalar $y$ on a $k$-dimensional vector $\mathbf{x}$, $y_i = m(\mathbf{x}_i) + \varepsilon_i = m(x_{1i}, \ldots, x_{ki}) + \varepsilon_i$.

The kernel regression estimator for $\widehat{m}(x_0)$ from (8.25) now becomes

$$\widehat{m}(x_0) = \frac{\frac{1}{Nh^k} \sum_{i=1}^{N} K\left(\frac{\mathbf{x}_i - \mathbf{x}_0}{h}\right) y_i}{\frac{1}{Nh^k} \sum_{i=1}^{N} K\left(\frac{\mathbf{x}_i - \mathbf{x}_0}{h}\right)}, \tag{8.34}$$

where $K(\cdot)$ is a **multivariate kernel**. The simplest case is the **product** of $k$ one dimensional kernels.

Regressors can either be **rescaled** to obtain a common optimal bandwidth or one can use **separate** bandwidths for each regressor.

A **curse of dimensionality** occurs because there are exponentially fewer observations within the bandwidth as the number of regressors increases.

This motivates the use of **semiparametric methods** where some structure is applied to the regression model to make estimation feasible.

# Outline

8. Non- and Semiparametric Methods
   - Kernel density estimation
   - Nonparametric regression
   - Semiparametric regression

# Motivation

**Semiparametric methods** are useful because they are

- Subject to **less restrictions** than fully parametric methods.
- Feasible in situations where a fully nonparametric analysis is **infeasible**.

Can retain consistency in situations where parametric estimators are **inconsistent** and nonparametric estimators are subject to **sparseness**.

A semiparametric specification is therefore more **robust** but potentially **less efficient** compared to fully parametric regression.

Semiparametric methods should optimally be **adaptive** (no efficiency loss) or attain the **semiparametric efficieny bound** (efficient given model)

$$\sqrt{N}(\beta(\hat{G}) - \beta) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathrm{V}_G]. \tag{8.35}$$

The two most commonly used semiparametric methods in econometrics are the **partially linear model** and the **single index model**.

# Partially linear model

The **partially linear model** specifies the conditional mean to be the linear function $(\mathbf{x}'\boldsymbol{\beta})$ plus an **unspecified nonlinear** scalar component $(\lambda(\mathbf{z}))$

$$\mathsf{E}[y|\mathbf{x},\mathbf{z}] = \mathbf{x}'\boldsymbol{\beta} + \lambda(\mathbf{z}). \tag{8.36}$$

Ignoring $\lambda(\mathbf{z})$ leads to inconsistent $\boldsymbol{\beta}$ unless $\mathsf{Cov}[\mathbf{x},\lambda(\mathbf{z})] = 0$.

The **Heckman two-step sample selection model** is a partial linear model where interest lies in $\boldsymbol{\beta}$ and $\lambda(\mathbf{z})$ is a control for sample selection.

However, the standard Heckman selection model is **fully parametric** while a semiparametric model do not specify $\lambda(\mathbf{z})$.

The model in (8.36) can also be estimated with fully nonparametric methods but it will be **inefficient**.

# Robinson difference estimator.

The partially linear model can be estimated using the **Robinson difference estimator**. Consider (8.36) in regression form

$$y = \mathbf{x}'\boldsymbol{\beta} + \lambda(\mathbf{z}) + u, \tag{8.37}$$

with $u = y - \mathrm{E}[y|\mathbf{x},\mathbf{z}]$. This implies that

$$\mathrm{E}[y|\mathbf{z}] = \mathrm{E}[\mathbf{x}|\mathbf{z}]'\boldsymbol{\beta} + \lambda(\mathbf{z}). \tag{8.38}$$

Subtracting (8.38) from (8.37) yields

$$y - \underbrace{\mathrm{E}[y|\mathbf{z}]}_{\hat{m}_{yi}} = (\mathbf{x} - \underbrace{\mathrm{E}[\mathbf{x}|\mathbf{z}]}_{\hat{\mathbf{m}}_{\mathbf{x}i}})'\boldsymbol{\beta} + u, \tag{8.39}$$

where the **conditional moments** can be nonparametrically estimated using appropriate estimators for $m_{yi}$ and $\mathbf{m}_{\mathbf{x}i}$.

Then OLS estimation of

$$y - \hat{m}_{yi} = (\mathbf{x} - \hat{\mathbf{m}}_{\mathbf{x}i})'\boldsymbol{\beta} + u, \tag{8.40}$$

yields $\sqrt{N}$-consistent and asymptotically normal estimates of $\boldsymbol{\beta}$ and $\lambda(\mathbf{z})$, where $\hat{\lambda}(\mathbf{z}) = \hat{m}_{yi} - \hat{\mathbf{m}}'_{\mathbf{x}i}\hat{\boldsymbol{\beta}}$.

# Single-index model

The **single index model** specifies the conditional mean to be an unknown scalar function of a **linear combination** of the regressors

$$\mathsf{E}[y|\mathbf{x}] = g(\mathbf{x}'\boldsymbol{\beta}). \tag{8.41}$$

The scalar function $g(\cdot)$ is now left unspecified in contrast to e.g. the **logit model** where $\mathsf{E}[y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})/[1 - \exp(\mathbf{x}'\boldsymbol{\beta})]$.

The model is only possible to nonparametrically identify up to **scale** since any $g^*(a + b(\mathbf{x}'\boldsymbol{\beta}))$ is **empirically equivalent** to $g(\mathbf{x}'\boldsymbol{\beta})$.

Any multiple of the regressors can thus be attributed to a different $g(\cdot)$. Instead we can **normalize**, say, $\beta_1 = 1$ and thereby interpret $\widehat{\beta}_j = \widehat{\beta}_j/\widehat{\beta}_1$.

$\boldsymbol{\beta}$ can then be estimated using e.g. the **average derivative (AD) estimator**.

# Average derivative estimator

Note that for the **linear index model** we have that $m(\mathbf{x}_i) = g(\mathbf{x}'\boldsymbol{\beta}))$ and thus

$$\boldsymbol{\delta} = \mathrm{E}\left[\frac{\partial m(\mathbf{x})}{\partial \mathbf{x}}\right] = \mathrm{E}[g'(\mathbf{x}'\boldsymbol{\beta})]\boldsymbol{\beta}, \tag{8.42}$$

where $\mathrm{E}[g'(\mathbf{x}'\boldsymbol{\beta})]$ is scalar so we can determine $\boldsymbol{\beta}$ **up to scale**.

The **generalized information matrix** states that for any function $h(\mathbf{x})$

$$\mathrm{E}[\partial h(\mathbf{x})/\partial \mathbf{x}] = -\mathrm{E}[h(\mathbf{x})s(\mathbf{x})], \tag{8.43}$$

where $s(\mathbf{x}) = \partial \ln f(\mathbf{x})/\partial \mathbf{x} = f'(\mathbf{x})/f(\mathbf{x})$. Thus,

$$\boldsymbol{\delta} = -\mathrm{E}[m(\mathbf{x})s(\mathbf{x})] = -\mathrm{E}[\mathrm{E}[y|\mathbf{x}]s(\mathbf{x})] = -\mathrm{E}[y]E[s(\mathbf{x})] \tag{8.44}$$

which sample analogue leads to the **average derivative (AD) estimator**

$$\widehat{\boldsymbol{\delta}}_{AD} = -\frac{1}{N}\sum_{i=1}^{N} y_i \widehat{s}(\mathbf{x}_i). \tag{8.45}$$

$\widehat{s}(\mathbf{x}) = \widehat{f}'(\mathbf{x})/\widehat{f}(\mathbf{x})$ is obtained by kernel density estimation on $\mathbf{x}$ and its derivative and $g(\cdot)$ by nonparametric regression of $y_i$ on $\mathbf{x}_i'\widehat{\boldsymbol{\delta}}$.

# Outline of lecture

8. Non- and Semiparametric Methods
   - Kernel density estimation
   - Nonparametric regression
   - Semiparametric regression

9. Regression Discontinuity
   - Sharp RD
   - Fuzzy RD

# Introduction

**Regression Discontinuity (RD)** methods exploit precise knowledge of the **rules** determining assignment to treatment around a **threshold** value of a variable.

Some thresholds are contextually **arbitrary** and therefore provide good natural experiments (e.g., credits, dates, quotas).

Estimation can be carried out using standard **regression** methods or using **non-parametric** techniques or a combination of both.

Two types:

1. **Sharp** RD: Related to Randomized Controlled Trial.
2. **Fuzzy** RD: Related to IV setting.

# Outline

# Sharp Regression Discontinuity

- Sharp research design: used when treatment status is **deterministic** and a **discontinuous** function of covariate $x_i$.
- For example, suppose

$$D_i = \begin{cases} 1 & \text{if } x_i \geqslant c \\ 0 & \text{if } x_i < c \end{cases} \qquad (9.46)$$

- Assignment is **deterministic**: if we know $x_i$, we know $D_i$.
- Treatment is a **discontinuous** function: it jumps at $x_i = c$.

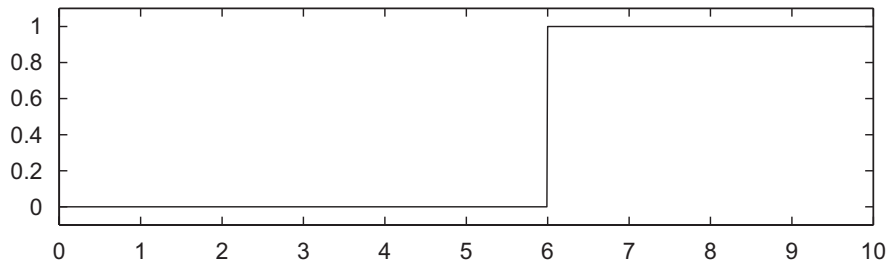



Figure 9.6. Assignment Probabilities.

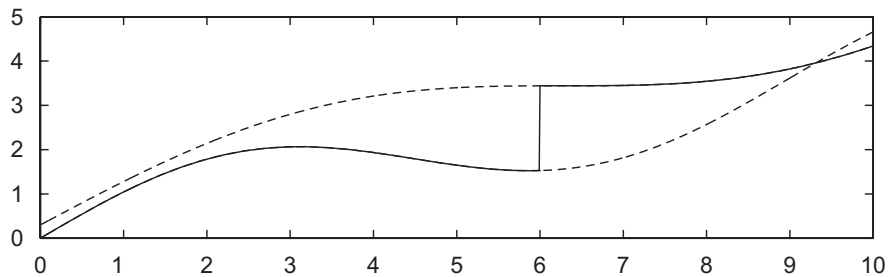



Figure 9.7. Potential and Observed Outcomes.

# Regression Discontinuity

**Example 9.1 (Your own research)**

Give an example of RD-papers in your research area or discuss what kind of rule could be exploited.

# Estimation

Sharp RD compares individuals just **above** and just **below** the cutoff point $c$.

There are no values of $x_i$ for which **both** treated and non-treated individuals are observed.

Instead, validity of RD relies on our willingness to **extrapolate** in a **neighborhood** of $c$.

Then we may consider the regression equation

$$Y_i = \alpha + \beta x_i + \tau D_i + \eta_i, \tag{9.47}$$

where $\tau$ is the causal effect of interest.

Thus, $Y_i$ includes **two functions** of $x_i$: the smooth $\beta x_i$ and the discontinuous

$$D_i = \mathbb{1}(x_i \geqslant c). \tag{9.48}$$

# Estimation II

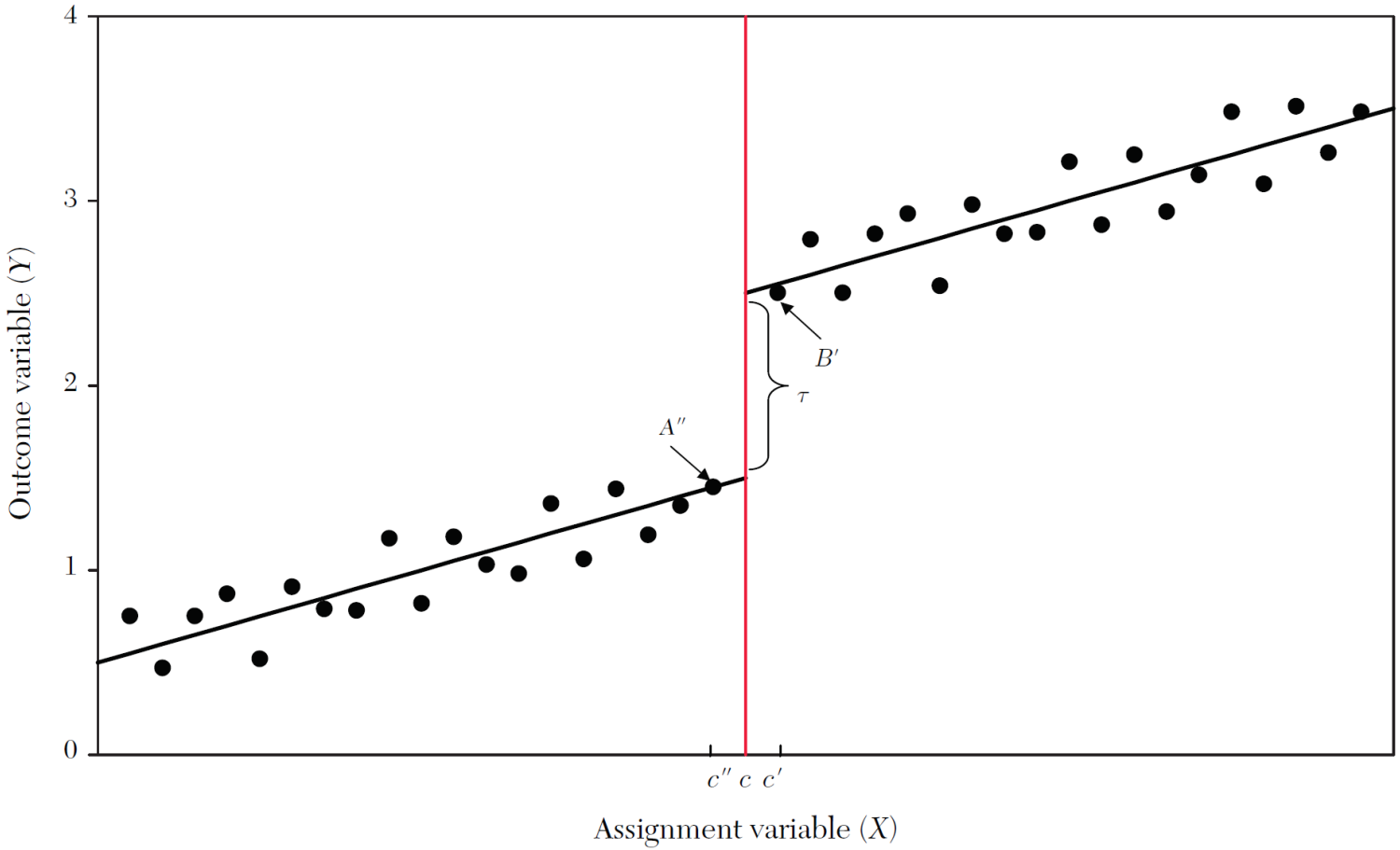



Figure 9.8. Linear RD setup

# Estimation III

Often, we cannot simply assume $Y_i$ to be a linear function of $x_i$. Instead, we may consider

$$Y_i = f(x_i) + \tau D_i + \eta_i. \tag{9.49}$$

As long as $f(x_i)$ is **continuous** in a neighborhood of $c$, it should be possible to estimate this model. Use **polynomials** or **local linear regression**.

Or allow the two potential outcomes to be **different** functions of $x_i$:

$$\begin{aligned} \mathbb{E}[Y_{0i}|x_i] &= \alpha + \beta_{01} x_i + \beta_{02} x_i^2 + \dots \\ \mathbb{E}[Y_{1i}|x_i] &= \alpha + \tau + \beta_{11} x_i + \beta_{12} x_i^2 + \dots \end{aligned} \tag{9.50}$$

...which may be better: we don't use observations from one group to estimate parameters for the other group. Use interaction term (difference of expectations with $D$) to implement this.

# Nonlinearity mistaken for discontinuity

## Example 9.2

Angrist and Pischke (2009), p. 254

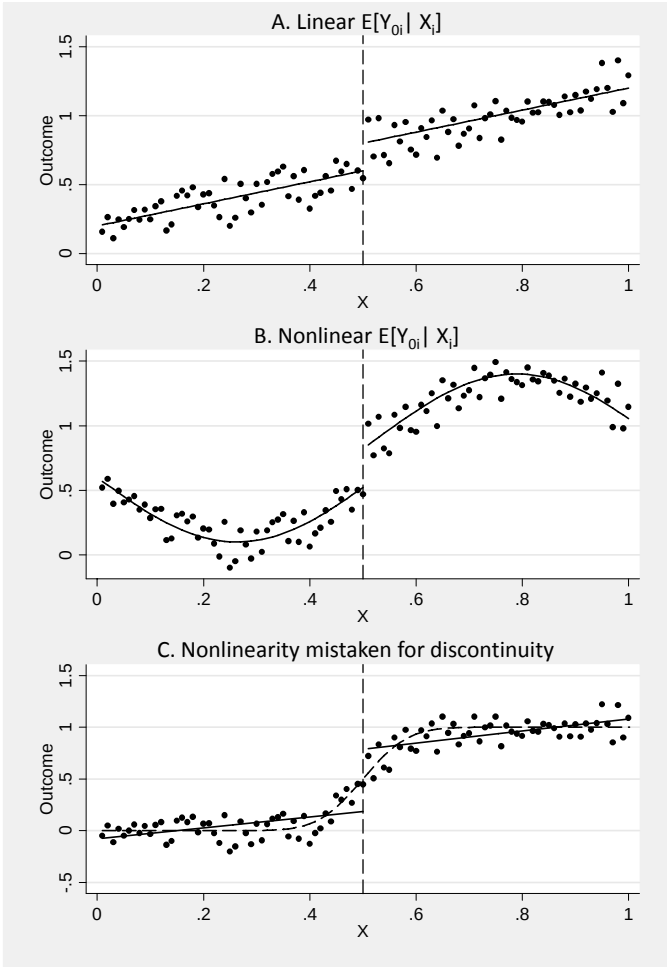



Figure 6.1.1: The sharp regression discontinuity design

The challenge in RD is to separate a jump from the continuous function!

# Outline

# Fuzzy Regression Discontinuity

Fuzzy RD exploits discontinuities in the **probability** or **expected value** of treatment conditional on a covariate.

As a result, the discontinuity becomes an **IV for treatment**.

Suppose

$$\Pr\left(D_i = 1 \mid x_i\right) = \begin{cases} g_1\left(x_i\right) & \text{if } x_i \geqslant c \\ g_0\left(x_i\right) & \text{if } x_i < c \end{cases}. \tag{9.51}$$

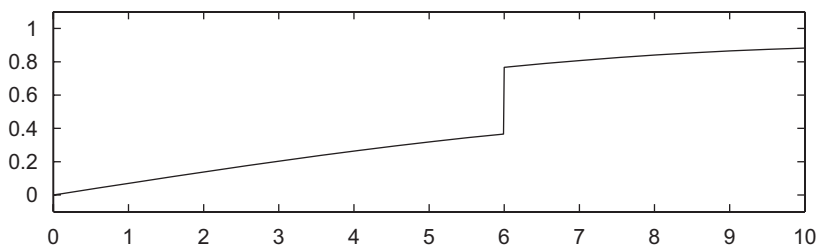



Figure 9.9. Assignment Probabilities.

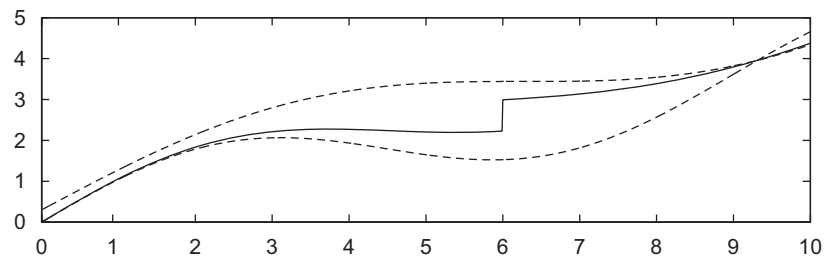



Figure 9.10. Potential and Observed Outcomes.

# Fuzzy Regression Discontinuity II

Thus

$$\mathbb{E}\left[D_i|x_i\right] = g_0\left(x_i\right) + \left[g_1\left(x_i\right) - g_0\left(x_i\right)\right] T_i, \tag{9.52}$$

where

$$T_i = \mathbb{1}\left(x_i \geqslant c\right). \tag{9.53}$$

The fuzzy RD model may be estimated using **2SLS**.

We can model $g_0\left(x_i\right)$ and $g_1\left(x_i\right)$ as $p$th order **polynomials** of $x_i$. Or we may use a non-parametric version.

$T_i$ and its interactions with $x_i$ (if treatment effect changes with x) and polynomials can be used as instruments for $D_i$.

The validity of estimates in fuzzy RD depends on our ability to distinguish the **continuous relationship** between $Y_i$ and $x_i$ from the **discrete jump** at $c$.

# Simple Model

RD estimates can be constructed from the **regression**:

$$Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + ... + \rho D_i + \eta_i \tag{9.54}$$

Use only $T_i$ as an **instrument**. The first stage is:

$$D_i = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2 + ... + \pi T_i + \zeta_{1i} \tag{9.55}$$

Substitute to obtain the fuzzy RD **reduced form**:

$$Y_i = \mu + \kappa_1 x_i + \kappa_2 x_i^2 + ... + \rho\pi T_i + \zeta_{2i} \tag{9.56}$$

By dividing you can then obtain an IV (Wald) estimate which is to be interpreted as a local-local (effect on compliers, effect relating to those with xses near the cut-off) effect of D and y.

# Fuzzy RD

## Example 9.3 (Angrist and Lavy, 1999)

Estimate the effect of class-size on children's test scores.
In Israel classes are capped at at most 40 children. So the approach uses jumps in class sizes (not probabilities) and enrollement in the school as the running variable.
The approach is fuzzy, because the rule does not perfectly predict average class sizes as sometimes classes are split before size 40.

## Example 9.4 (von Ehrlich, Seidel, 2018, AEJ: EP)

- From 1971 to 1994 West German geographical areas near to the Iron Curtain benefited from large scale subsidies. All districts with either 50 percent of their area or population within a distance of 40 km from the inner-German and Czechoslovakian border became part of the Zonenrandgebiet.
- The authors apply a spatial RD based on municipalities and grid cells in a close neighborhood on either side of the treatment border. Location is the running variable. If other relevant factors vary continuously at this border, a discontinuity in economic outcomes can be interpreted as the causal effect of the place-based policy.
- Because administrative borders may not be drawn randomly, the authors also exploit the jump at the distance of 40 km from the Iron Curtain directly in a fuzzy RD.