Gov 2002 - Section 7 - Kernels & Nonparametric Statistics

Connor Jerzak

Harvard University, Department of Government

October 26, 2016

Our goal

- **Setup**.
  - $X_n \sim P$. Data are coming in from the world, governed by some density, P.
  - We want to estimate $\hat{p}(x)$. This object will allow us to (a) understand the scientific process underlying the world, and (b) understand conditional relationships, since $\hat{\mathbb{E}}(y|x) = \int \hat{p}(y|x)y \ \ \mathrm{d}y$.

Parametric models

- **Parametric models**.
  - Allow us to find $\hat{p}(x)$ using a fixed number of parameters.
  - Example: Assume P is Normal. Once I know the mean, $\bar{x}$, and variance, $\hat{\sigma}_x^2$ of my data, x_n, I can derive all other features of this distribution.
  - **Key point**. Number of parameters is fixed. Increasing the data does not increase the number of parameters.

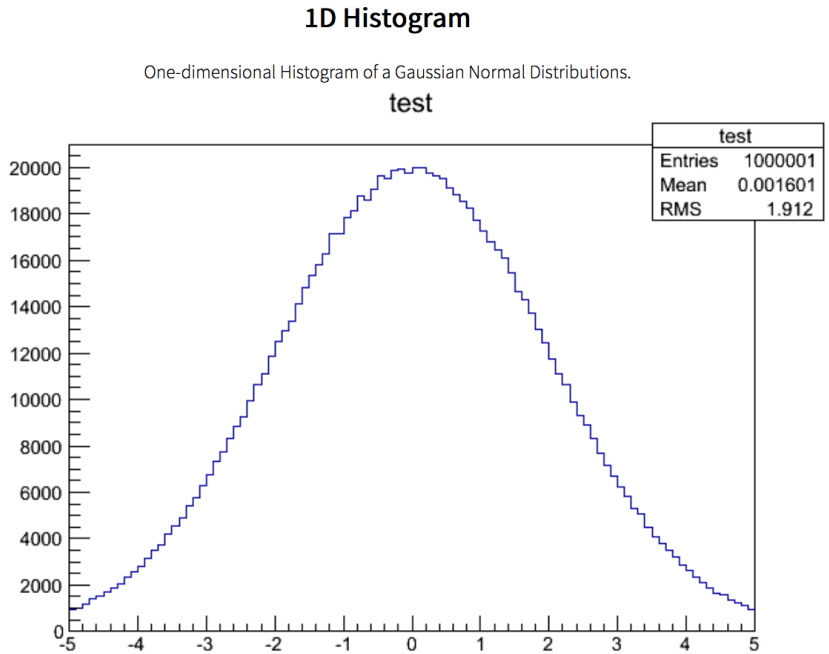
Non-parametric models

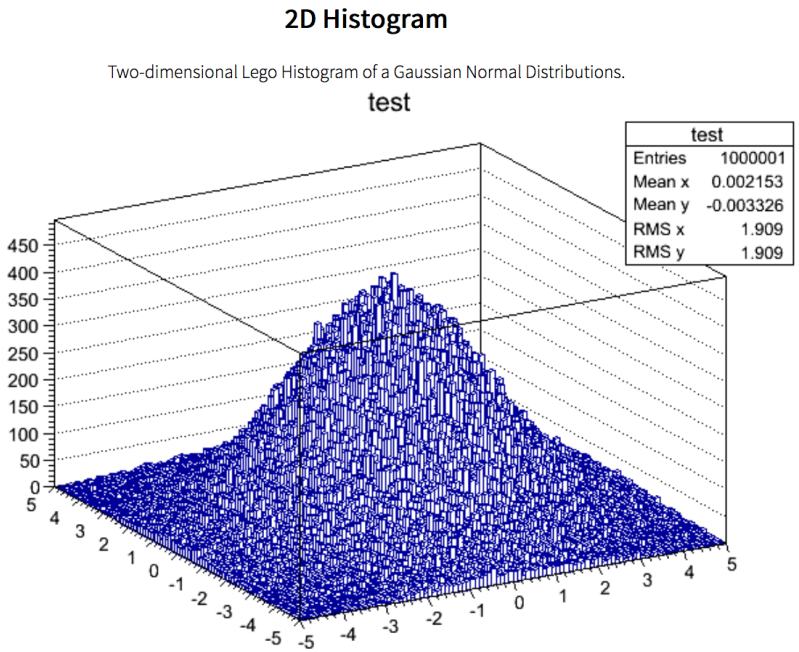
- **Non-parametric models**.
  - Allow us to find $\hat{p}(x)$ making fewer assumptions about the form of the underlying distribution, P.
  - Example: A histogram is one nonparametric estimate of a probability distribution. The number of bins grows with the number of datapoints. We will return to this fact.
  - **Key point**. Number of parameters is determined by the size of the training data, NOT the underlying statistical model. Increasing the data will increase the number of parameters.
  - Distribution-free? Hyperparametric?

Histograms, A Simple Non-parametric Estimator

- Divide domain of X_n into hypercubes (or bins), where each side has length h.
- $\hat{p}_h(x_i) = \hat{\theta}_j / c$ for $x_i \in B_j$
  - i is the unit index, j is the bandwidth index.
  - θ_j is the proportion of observations in the bin: $\hat{\theta}_j = \frac{1}{n} \sum_i^n I(x_i \in B_j)$.
  - c is a normalizing constant that ensures $\hat{p}_h(x)$ integrates to one: $\sum_j \hat{\theta}_j / c = 1$.

Histograms, visualized





Histograms, Problems

- The curse of dimensionality. If I divide the joint support of my data into hypercubes, there are an exponential number of hypercube bins. As a result,

$$\lim_{d\uparrow} \Pr(X_i \in B_j) \to 0, \tag{1}$$

  where d denotes the dimensions of my data.
- Histograms are not smooth.
- Histograms also depend on the end points of the bins, as well as the width of the bins.

Solution: Kernel density estimation. Key point: kernels are less locally-specific than histograms. That is, histogram bins don't influence each other. However, this property means that histograms throw away information that will allow us to better estimate $\hat{p}(x)$.

Kernels, Introduced

- *Kernel density estimation* uses the distances between my datapoints in order to assign probabilities and thus obtain $\hat{p}(x)$. In words, even if I don't have data points in some interval of the domain of x, I can still estimate the probability of obtaining such a point by using neighboring points.
- A kernel weights a sequence of datapoints, assigning more weight to some datapoints than to others based on the similarity to some target datapoint.
- *In sum*, kernels approximate the density of a point from P as an average of the density around this point, as well as the density of other points nearby, where the contribution of the other points is weighted by their similarity to the target observation.

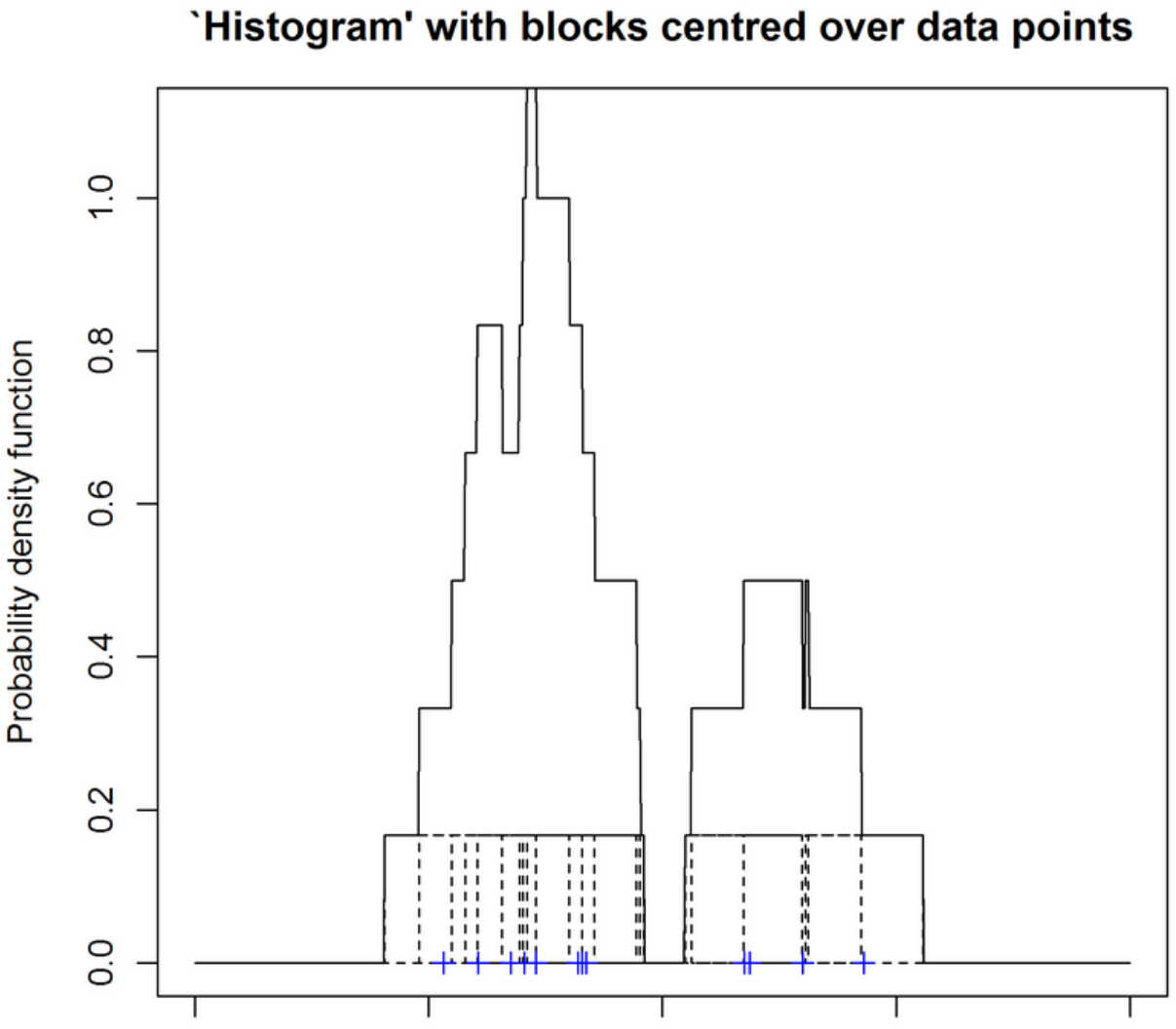
Kernels, Definition

- If $\{x_1, x_2, ..., x_n\} \sim P$, we can write $\hat{p}_h(x)$ as:

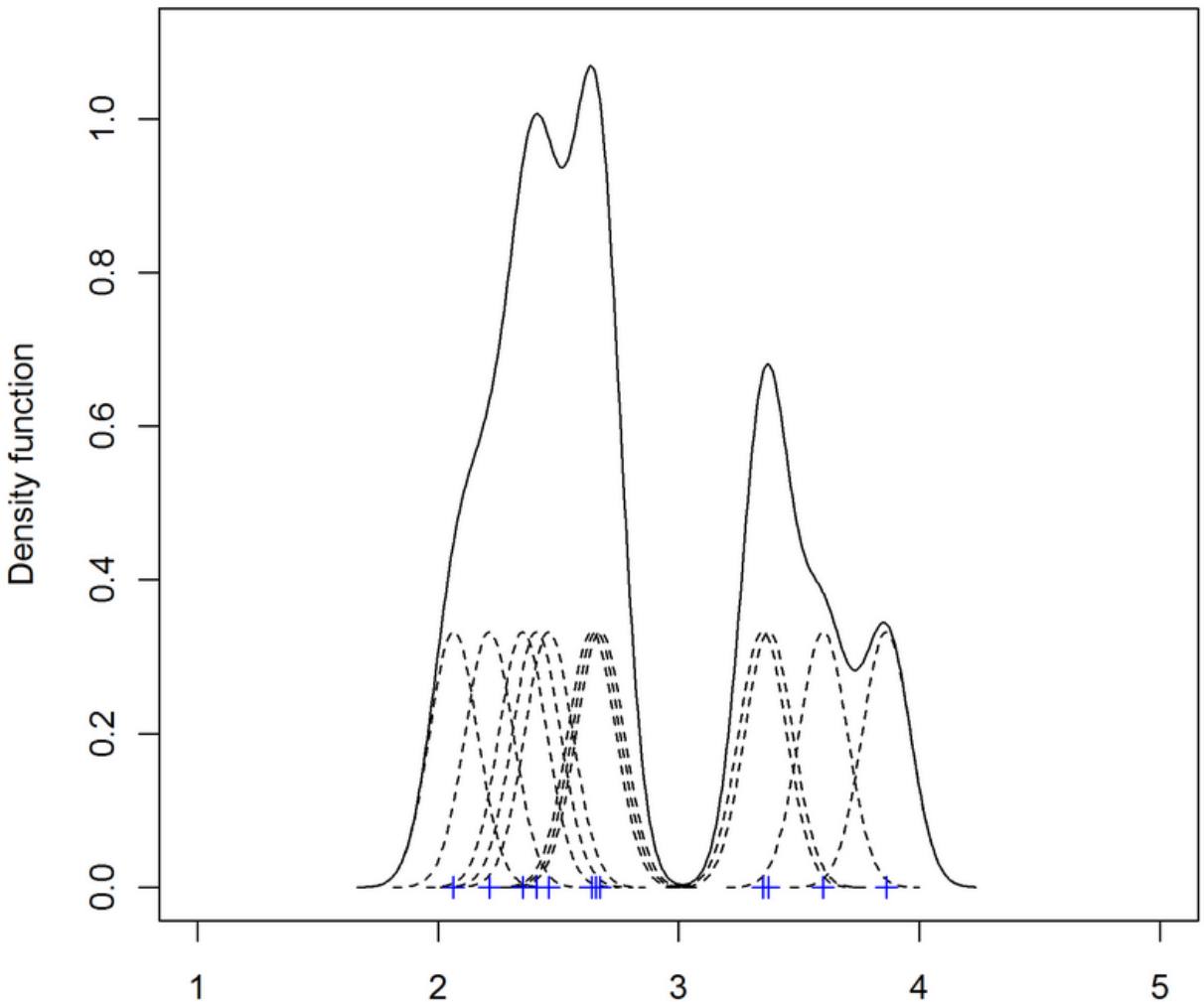
$$\hat{p}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x; x_i), \tag{2}$$

  where $K_h(\cdot) \geq 0$, $K(u; q) = K(-u; q)$, and $\int_{-\infty}^{\infty} K(u; q) \mathrm{d}u = 1$.
- These requirements ensure that the $\hat{p}_h(x)$ results in a proper PDF, and that the average value of the kernel around x is 0.
- In other words, you can think if a kernel function, $K_h(x; x_i)$ as similar to a scaled probability density centered around x_i.

A Uniform (Box) Kernel



A Smooth Kernel



Bandwidth Selection

- Small values of h lead to spiky estimates (high variance) while larger h values lead to oversmoothing (high bias).
  - h controls the breadth of the kernel function, i.e. how much my estimate, $\hat{p}(x = x_i)$, is influenced by the other units.
- Some researchers select h by finding the h that minimizes approximations of the Mean Integrated Squared Error (MISE). This is not usually tractable.

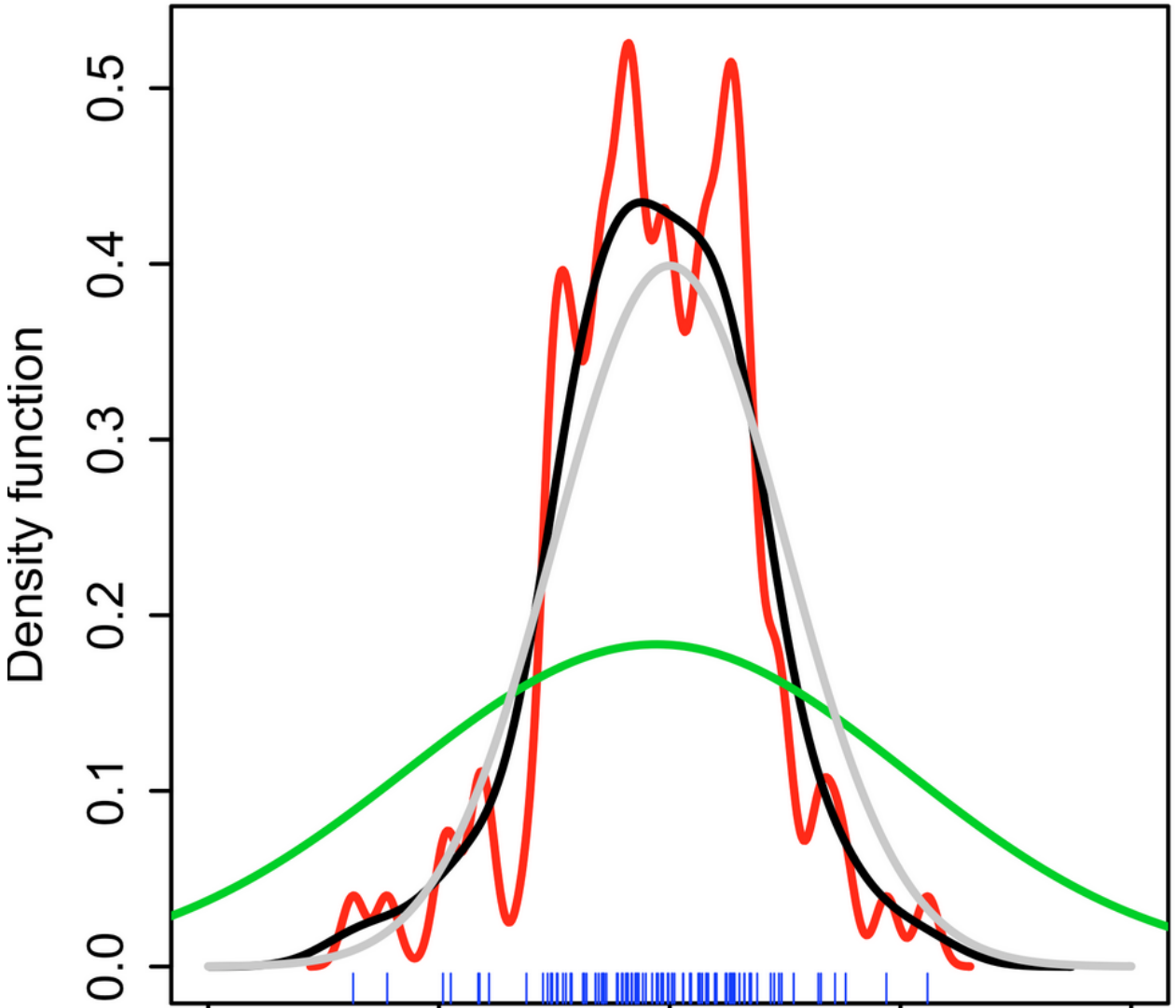
$$\text{MISE}(h) = \mathbb{E}\left[\int (\hat{p}_h(x) - P(x))^2\right].$$

- Cross validation is also often used:

$$\hat{p}_{h,i}(x_i) = \frac{1}{(n-1)} \sum_{j \neq i} K(X_j, X_i).$$

Select the h that gives the best pseudo-likelihood: $\prod_{i=1}^{n} \hat{p}_{h,i}(x_i)$.

Bandwidth Implications



Application 1- Nonparametric Regression

- I want to model my outcome, y, as a function of my input data:

$$Y_i = m(X_i) + \epsilon_i, \quad i = 1, 2, .., N,$$

where $m(X) = \mathbb{E}[Y|X = x]$ and $\mathbb{E}[\epsilon_i|X = x] = 0$.

- The problem: $m(x)$ is unknown.
- In OLS, we impose a strong but interpretable functional form on this conditional relationship:

$$Y_i = X_i; \beta + \epsilon_i, \quad i = 1, 2, .., N,$$

Nonparametric Regression

- In nonparametric regression, I let the data flexibly determine the shape of $m(\cdot)$:

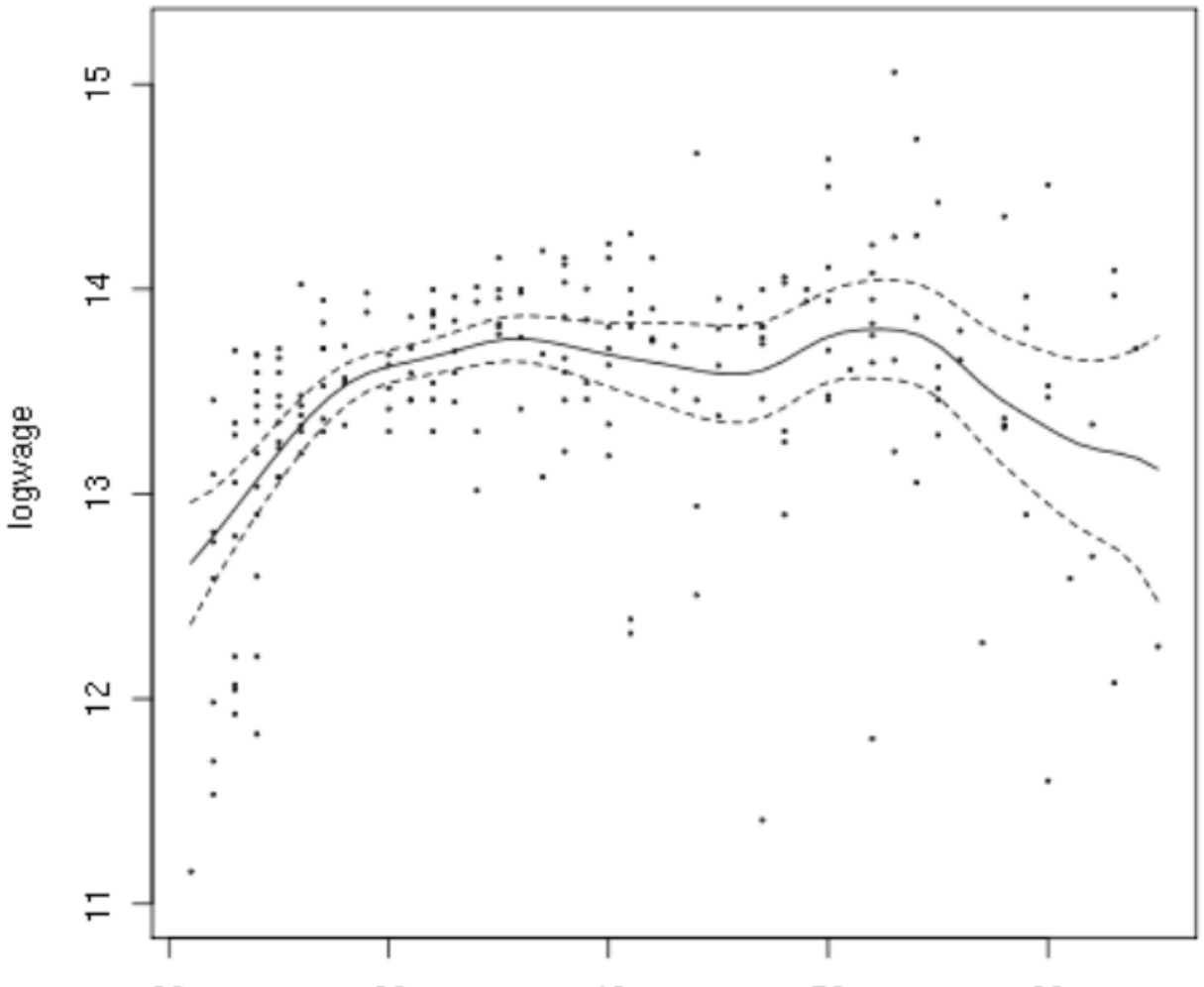
$$Y_i = m(X_i) + \epsilon_i, \quad i = 1, 2, .., N,$$

where

$$\hat{m}_h(x) = \frac{\sum_{i=1}^{n} K_h(x; x_i) \cdot y_i}{\sum_{i=1}^{n} K_h(x; x_i)}.$$

- Notice that $\hat{m}_h(x)$ is just a weighted average, where the weights are determined by the form of $K_h(\cdot;\cdot)$ and h.
- This is a straightforward application of kernel density estimation. Once we have $\hat{p}(x, y)$, we can use it to estimate any quantity of interest (for example, $\hat{\mathbb{E}}[Y|X = x]$).

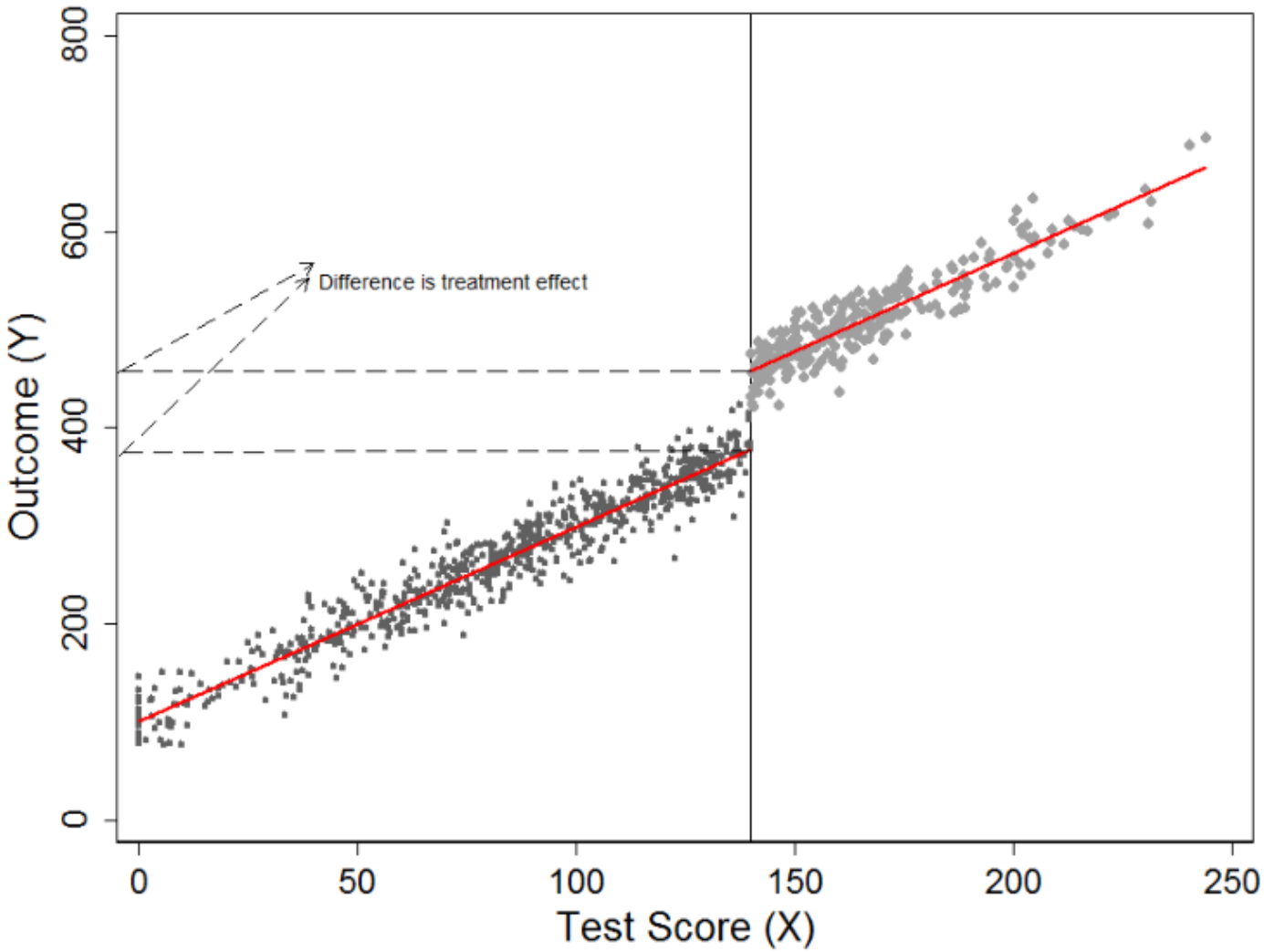
Kernel Regression



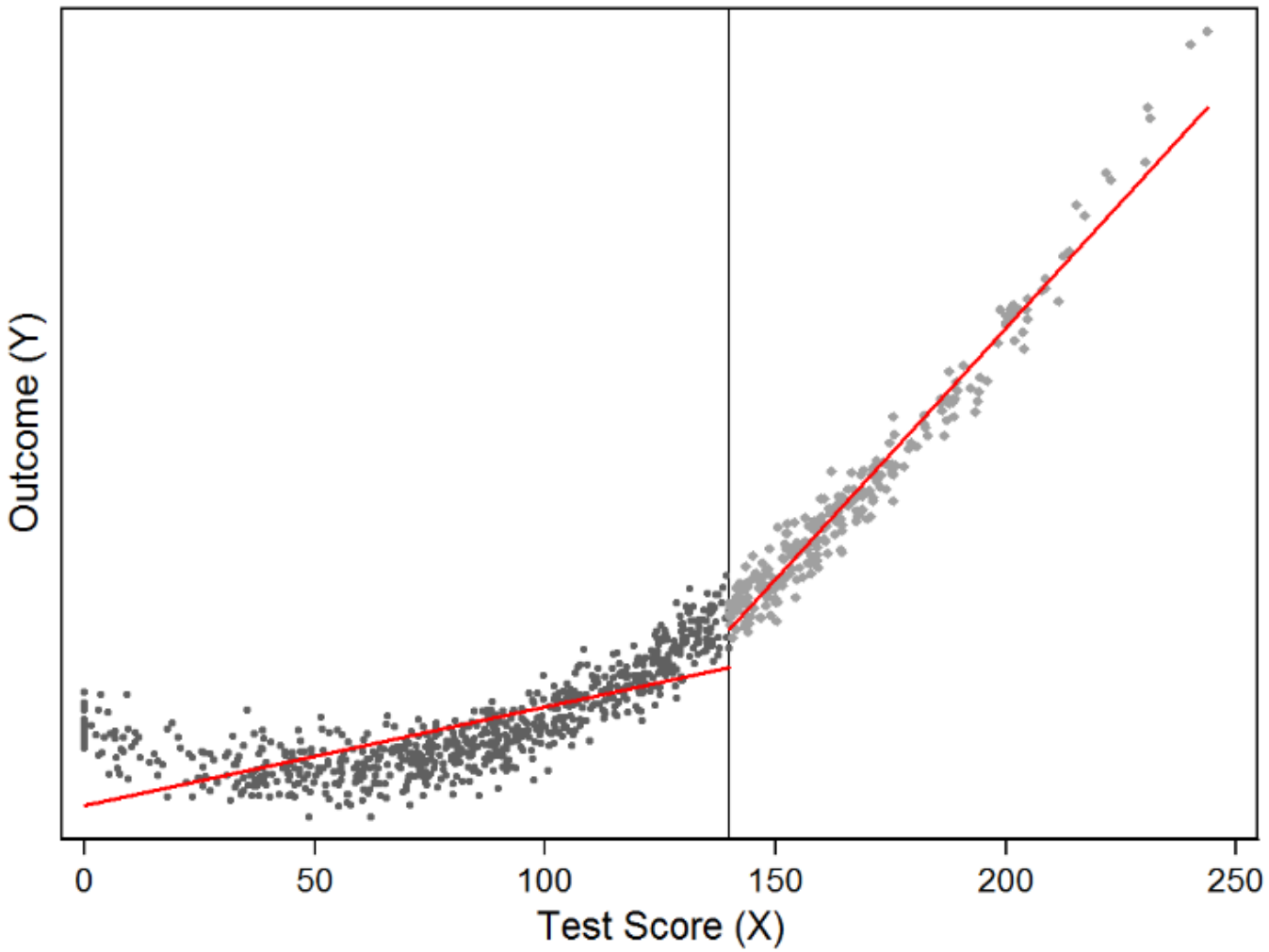
Application 2 - RDD

- **Regression discontinuity design (RDD)** is an observational data technique that attempts to capture the causal effect of an intervention. It does so by exploiting a cutoff or discontinuity above/below which some intervention is assigned.
  - By comparing units just below and just above the threshold, we can estimate the local average treatment effect of the intervention.
- **Examples**.
  - Scholarship thresholds.
  - Bare winners in US elections.
  - Age determines pension availability.

RDD, depicted



RDD, depicted 2



Local-linear & Semi-parametric RDD

- Local linear regression is often used to estimate the local average treatment effect in an RD design. In particular, we assume that the treatment, T is given based on a cutoff value of Z_i.

$$Y_i = \beta_0 + \tau \cdot T_i + \beta_1 \cdot (Z_i - c) + \epsilon_i. \tag{3}$$

or

$$Y_i = \beta_0 + \tau \cdot T_i + m(Z_i - c) + \epsilon_i. \tag{4}$$

Notice:

  - T_i is an indicator of whether i received the treatment (was just above/below threshold).
  - c denotes the cutoff in the variable Z that leads to treatment.
  - τ denotes the local average treatment effect.
  - $Z_i - c$ denotes the distance of unit i from the threshold. We control for the distance of each unit the cutoff, c.
  - Notice that this is estimated only for the units within a given bandwidth, h: $c - h \leq X_i \leq c + h$.

Nonparametric RDD, Practical Considerations

- In practice, researchers often assume higher-order terms and interactions as well:

$$\begin{aligned} Y_i = \beta_0 + \tau \cdot T_i + \beta_1 \cdot (Z_i - c) \\ + \beta_2 \cdot (Z_i - c)^2 + \cdots + \beta_1 \cdot (Z_i - c)^p \\ + \beta_1 \cdot (Z_i - c)^p \cdot T_i + \cdots + \epsilon_i. \end{aligned} \tag{5}$$

- Often, units will be imbalanced on background covariates across the threshold ("sorting at the threshold").
- Choice of bandwidth is arbitrary. In your research, show how your results are robust to reasonable range of the bandwidths.

Acknowledgements

- These slides build from material presented by Pearl, Rubin, Jordan, Jackman, and others.