PSY 503: Foundations of Psychological Methods

Lecture 15: Regression and Conditional Expectations

Robin Gomila

Princeton

October 28, 2020

What is regression?

- Regression quantifies how an outcome variable Y varies, **on average**, as a function of:
  - One variable (e.g., treatment assignment Z, treatment D, a predictor X)
  - Or more than one variables (e.g. a series of predictors X_1, ..., X_n)
- Let's focus on the bivariate case for now

OK! Wait! What is regression again?

- What is specific about regression in terms of quantifying the relationship between two variables? For example, how is regression different from correlation or covariance?
- Regression describes **conditional expectations**
  - Example:

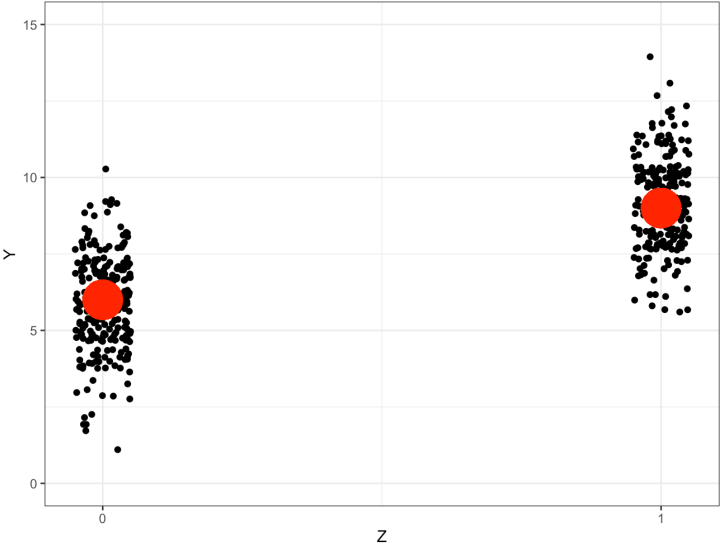
$$\mathbb{E}[Y_i|Z_i = z]$$

  - i.e., "the conditional expectation of Y_i given that Z_i equals the particular value z"
- Conditional expectations tell us how the average of one variable changes as we move the conditioning variable over the values this variable might take on
- The collection of all such averages is called the **conditional expectation function** (CEF)

CEF for binary treatments

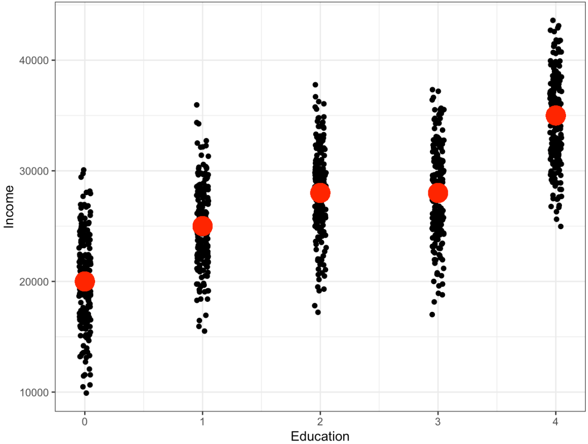
- Suppose participants were randomly assigned to an experimental condition Z_i, such that
  - $Z_i = 0$ if participants were assigned to the control condition
  - $Z_i = 1$ if participants were assigned to the treatment condition
- Suppose that we collected data for a dependent variable Y_i
- The CEF of Y given Z is the collection of two expectations:
  - $\mathbb{E}[Y_i|Z = 0]$
  - $\mathbb{E}[Y_i|Z = 1]$

Example of a CEF for a binary treatment



What if X is not binary?

- Example of the CEF $\mathbb{E}[Y_i|X_i]$ for a non-binary predictor variable X



Estimating the population CEF

- How do we estimate the population CEF $\mathbb{E}[Y_i|X_i]$?
- Run regression on a random sample. What does this mean?
- Concretely: find a way to estimate $\widehat{\mathbb{E}}[Y_i|X_i]$
    - Estimate all possible conditional averages
    - In experiments (where X is Z), two possible conditional averages: $\hat{\mu}_T$ and $\hat{\mu}_C$

Regression: parameter, estimator, estimand

- In regression, the CEF $\mathbb{E}[Y_i|X_i]$ is (generally) the **parameter** that we are interested in
- For a given sample dataset, we obtain an **estimate** $\widehat{\mathbb{E}}[Y_i|X_i]$ of the parameter $\mathbb{E}[Y_i|X_i]$

A note on CEFs involing more than one conditioning variables

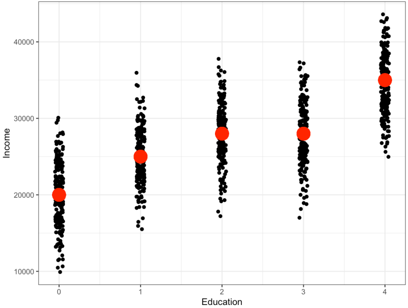
- When the CEF involves k conditioning variables, we write:

$$\mathbb{E}[Y_i | X_{1i},\ ...,\ X_{ki}]$$

- These CEFs are more difficult to plot / visualize
- Idea is the same, instead of looking at average value of Y conditioning on treatment or income, we condition on treatment/income and other variables such as gender, education, etc.
- Let's stick with the bivariate case for now

The CEF is the best predictor of Y!

- Suppose that we knew the full joint cumulative distribution (CDF) of X and Y and then someone gave us a randomly drawn value of X



- The CEF minimizes the **Mean Squared Error** (MSE)
  - MSE: average of the squares of the errors. i.e., the average squared difference between the estimated values and the actual values

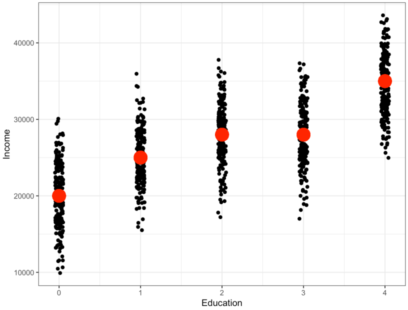
The CEF is the best predictor of Y!

- No better way (in terms of MSE) to approximate Y given X than the CEF
  - True independently of the distribution of Y and X
- This make the CEF a natural target of inquiry:
  - If the CEF is known, much is known about how X relates to Y

Nonparametric regression

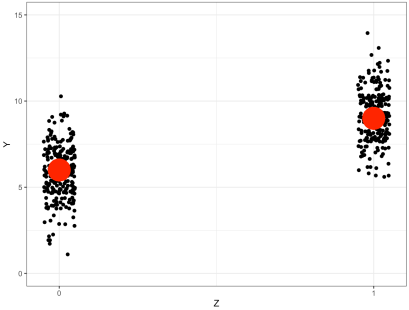
Nonparametric regression

- Nonparametric strategies make minimal assumption about the functional form of the data generating process
- Nonparametric regression does not impose a **functional form** on the relationship between Y and X
  - i.e., Does not impose a shape on the CEF $\mathbb{E}[Y|X]$



Nonparametric regression

- Nonparametric strategies make minimal assumption about the functional form of the data generating process
- Nonparametric regression does not impose a **functional form** on the relationship between Y and X
  - i.e., Does not impose a shape on the CEF $\mathbb{E}[Y|X]$



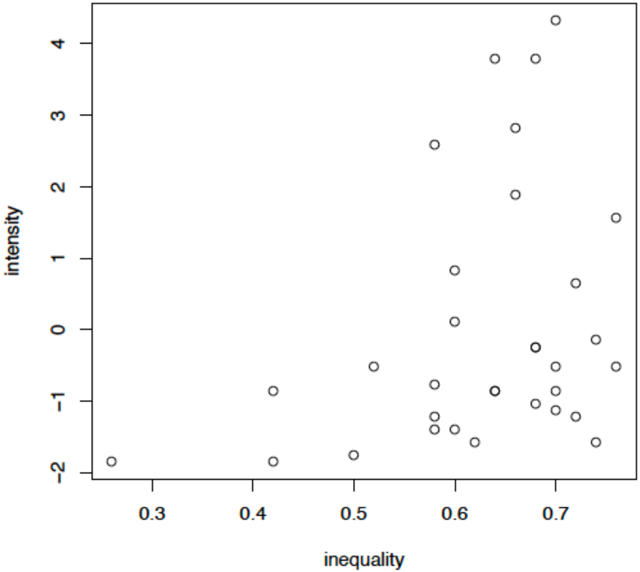
Nonparametric regression

- Works well as long as:
  - X is discrete
  - Small number of values of X
  - Small number of X variables
- What if X is continuous?

Nonparametric regression with continuous X

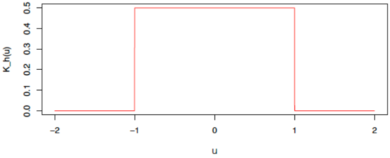
- Let's consider the data from a sociology paper:
  - Chirot, D. & Ragin, C. (1975). The market, tradition and peasant rebellion: The case of Romania. **American Sociological Review** 40, 428-444
  - Peasant rebellions in Romanian counties in 1907
  - Peasants made up 80% of the population
  - About 60% of them owned no land, which was mostly concentrated among large landowners
- We're interested in the CEF $\mathbb{E}[Y|X]$ in which
  - Y: intensity of the peasant rebellion
  - X: inequality of land tenure

Nonparametric regression with continuous X



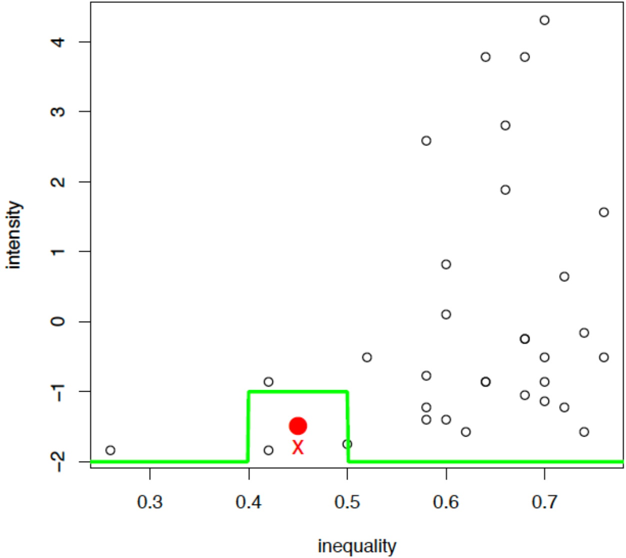
Uniform Kernel Regression: Simple Local Averages

- One approach is to use a moving local average to estimate $E[Y|X]$
- Calculate the average of the observed y points that have x values in the interval $[x_0 - h,\ x_0 + h]$
- $h =$ some positive number (called the **bandwidth**)
- **Uniform kernel:** every observation in the interval is equally weighted

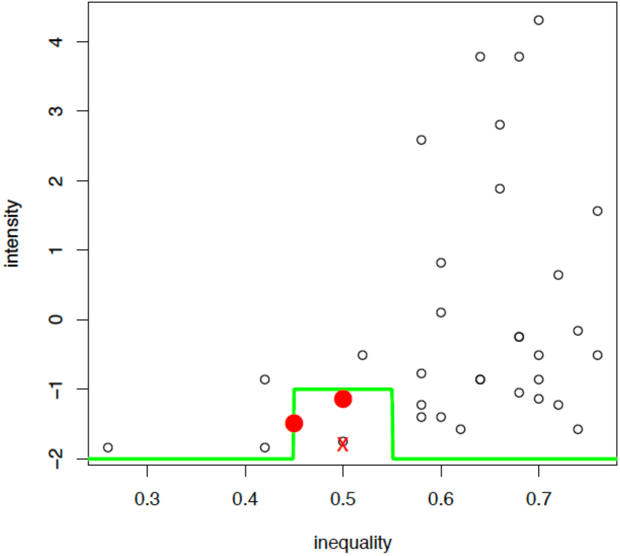


- **Uniform kernel regression**: $\mathbb{E}[Y|X = x_0]$

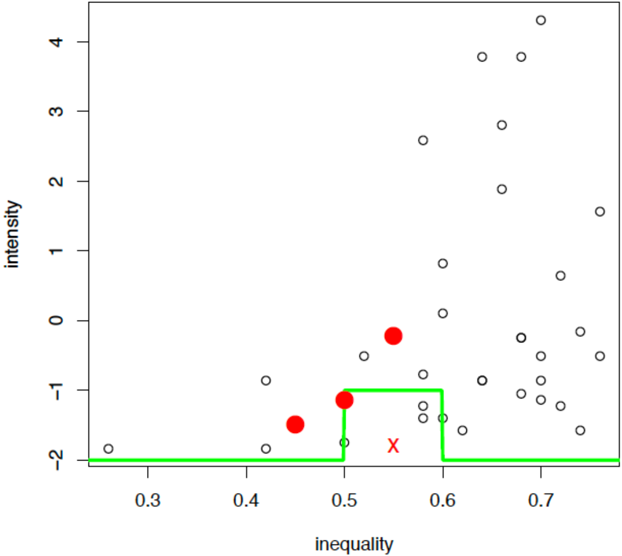
Kernel Regression: Simple Local Averages



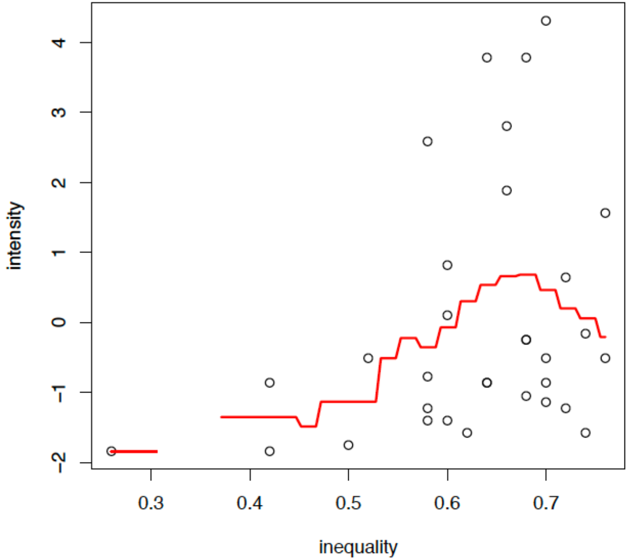
Kernel Regression: Simple Local Averages



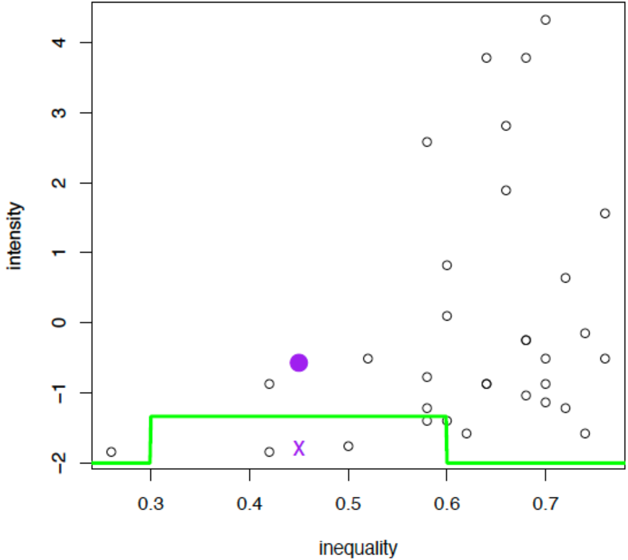
Kernel Regression: Simple Local Averages



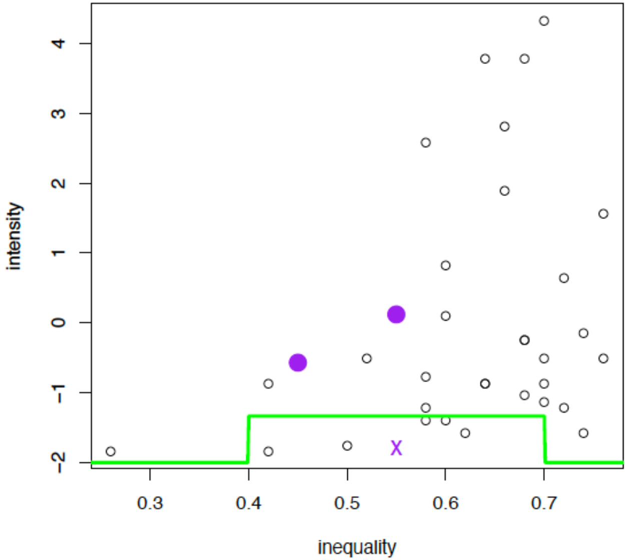
Kernel Regression: Simple Local Averages



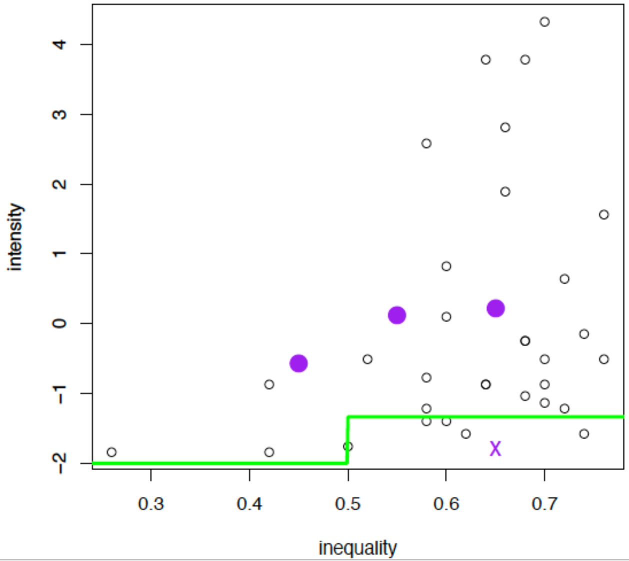
Changing the bandwidth



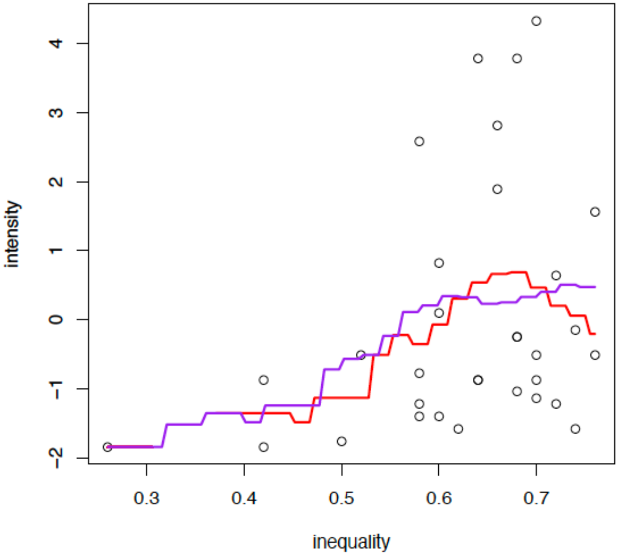
Changing the bandwidth



Changing the bandwidth



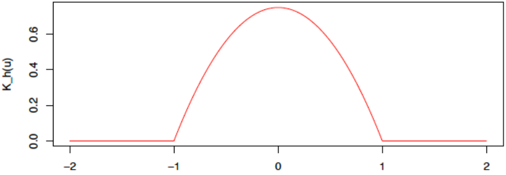
Changing the bandwidth



Kernel Regression: Weighted Local Averages

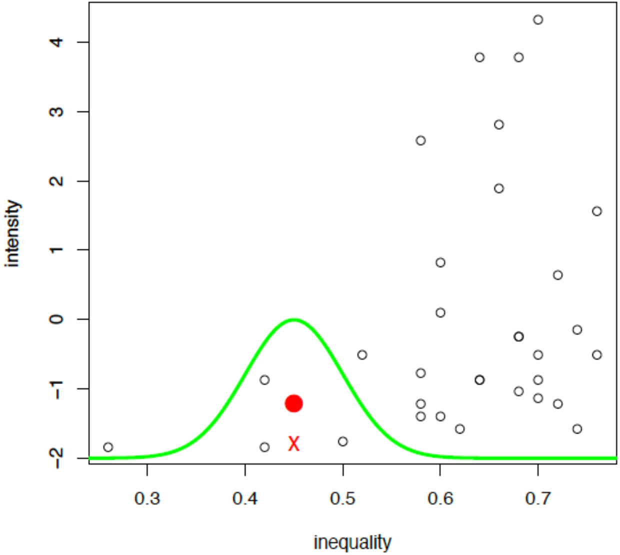
- Another approach is to construct weighted local averages
- Data points that are closer to x_0 get more weight than points farther away

1. Decide on a symmetric kernel weight function K_h (e.g. Epanechnikov)

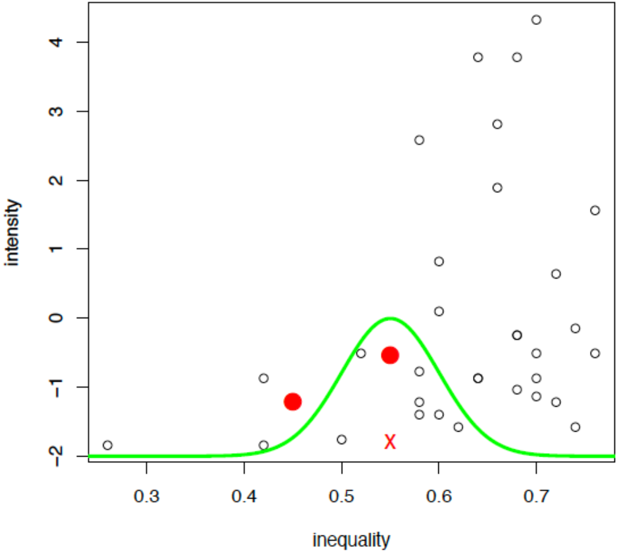


2. Compute weighted average of the observed y points that have x values in the bandwidth interval $[x_0 - h,\, x_0 + h]$

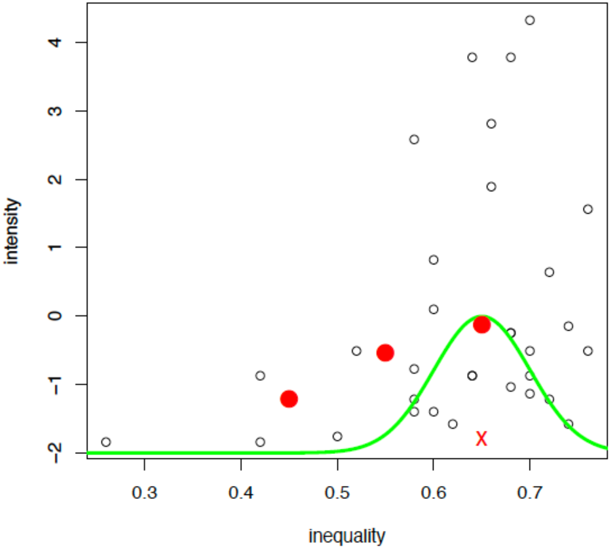
Kernel Regression: Weighted Local Averages



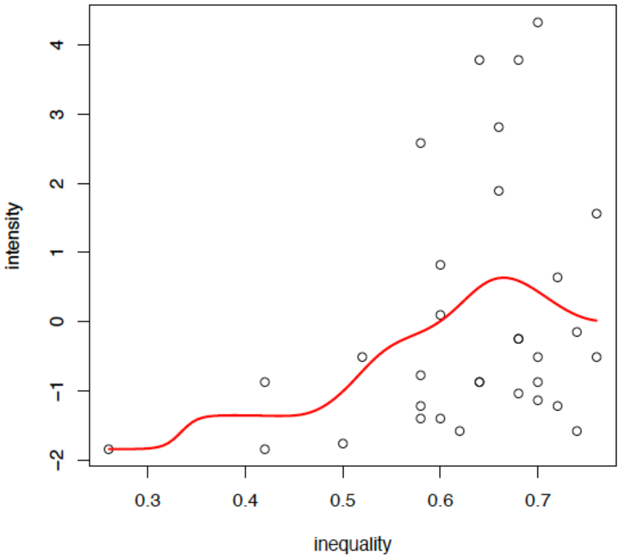
Kernel Regression: Weighted Local Averages



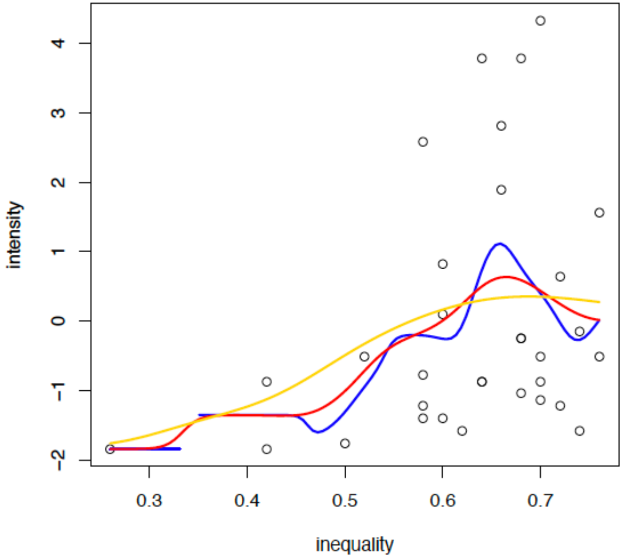
Kernel Regression: Weighted Local Averages



Kernel Regression: Weighted Local Averages



Kernel Regression: Weighted Local Averages



Bias - Variance tradeoff

- When choosing an estimator $\widehat{\mathbb{E}}[Y|X]$ for $\mathbb{E}[Y|X]$, we face a **bias-variance tradeoff**
- Notice that we can chose models with various levels of flexibility:
    - A very flexible estimator allows the shape of the function to vary (e.g. a kernel regression with a small bandwidth)
    - A very inflexible estimator restricts the shape of the function to a particular form (e.g. a kernel regression with a very wide bandwidth)

Bias - Variance tradeoff

- A less "flexible" estimator leads to more bias
- A more "flexible" estimator leads to more variance
- As the name suggests, this problem cannot be fixed
- With lots of data points, we can "afford" to use a more flexible estimator

Regression and causality

- Regression describes the CEF
- Regression does not have magic powers to tell you whether X or Z causes Y
  - That would be amazing
  - That is not the case
- Under very specific assumptions, regression allows you to identify causal relationships. These assumptions can be about:
  - The design of the study (e.g., experimental, random assignment)
  - The statistical model (more soon!)
  - ... or both (more soon!)