Section 11: Nonparametrics

Valentine Gilbert

December 2, 2021

Overview

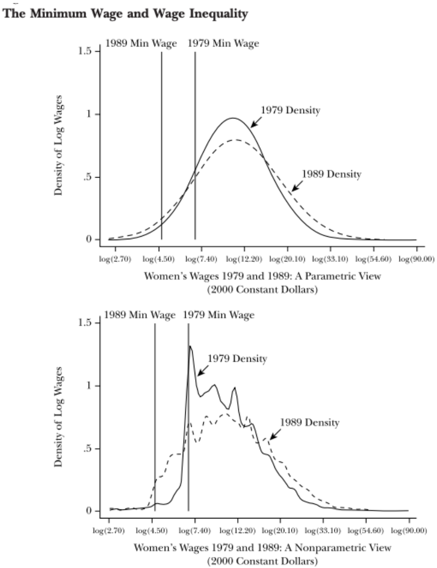
1. Overview

2. Nonparametric Density Estimation

3. Nonparameteric Regression

Motivating Example

- Between 1979 and 1989, wage inequality among women increased – the standard deviation of log wages rose by 25%
- Wages are approximately log-normal, so we could visualize this by assuming log-normality and estimating the parameters with MLE to generate the top figure
- Another approach is to estimate the distribution of wages nonparametrically to generate the bottom figure
- Which figure tells you more about the possible causes of increased inequality?


The Minimum Wage and Wage Inequality

Women's Wages 1979 and 1989: A Parametric View (2000 Constant Dollars)

Women's Wages 1979 and 1989: A Nonparametric View (2000 Constant Dollars)

Overview

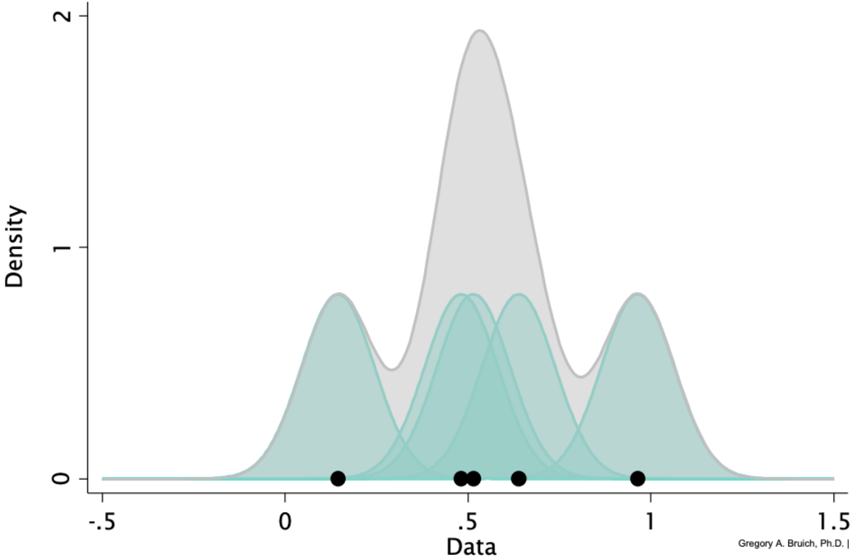
- Nonparametric techniques make minimal assumptions about the data generating process
- You've made frequent use of nonparametric estimation on problem sets – each time you created a binned scatter plot
- Nonparametric estimation is especially useful as a descriptive tool – nonparametric figures of the CEF or density convey far more information about the data than simple summary statistics
- The complexity of nonparametric models depends on the data – complexity is chosen to balance bias and variance

Overview

- Nonparametric density estimation techniques try to estimate the distribution of the data. You could nonparametrically estimate the joint density of two or more variables, or the marginal density of a single variable. Today, we'll focus on the marginal density, $f(x)$.
- The simplest way to nonparametrically estimate the density of a random variable is with a histogram
- Kernel density estimation models the distribution as a continuous function and smooths the estimate by making $\hat{f}(x)$ a weighted function of values within a rolling bandwidth
- The shape of the weighting function (the kernel) doesn't matter much, so the key choice is the bandwidth

Kernel Density Estimation for a 5 observation Data Set

Kernel is the Normal Density with $h = 0.1$



Bias-Variance Tradeoff: Loss Function

- The choice of bandwidth trades off bias and variance
- Wider bandwidths reduce variance by smoothing sampling variation; but they increase bias by estimating $f(x_0)$ with data farther from x_0
- Specifying an objective function allows us to characterize the *optimal bandwidth*
- Silverman (1986) characterizes the optimal bandwidth using the loss function:

$$ISE(h) = \int \left(\hat{f}(x) - f(x)\right)^2 dx$$

- The integrated squared error evaluates prediction error at all values of x, weighting all values equally

Bias-Variance Tradeoff: Optimal Bandwidth

- We'd like to choose a bandwidth that, *on average*, minimizes the integrated squared error
- In other words, we're interested in the expected integrated squared error, which we can decompose into bias and variance terms and rewrite with Taylor series approximations:

$$\begin{aligned} E\int\left(\hat{f}(x)-f(x)\right)^2 dx &= \int\left(E\left[\hat{f}(x)\right]-f(x)\right)^2 dx + \int var\left(\hat{f}(x)\right) dx \\ &= \int \text{bias}^2 dx + \int var\left(\hat{f}(x)\right) dx \\ &\approx \frac{1}{4}h^4 k_2^2 \int \left(f''(x)\right)^2 dx + \frac{1}{Nh}\int K(u)^2 du \end{aligned}$$

- The solution to the minimization problem is

$$h_{opt} = N^{-1/5}\, \underbrace{k_2^{-2/5}\left(\int K(u)^2 du\right)^{1/5}}_{constant} \underbrace{\left(\int \left(f''(x)\right)^2 dx\right)^{-1/5}}_{unknown}$$

Bias-Variance Tradeoff: Silverman's Rule

- To estimate the optimal bandwidth, we have to estimate the curvature of the unknown density function
- Silverman's approach is to proceed *as if* the unknown density $f(x)$ is Normal
- Then the optimal bandwidth simplifies considerably:

$$h_{opt} \approx \begin{cases} 1.06\sigma N^{-1/5} & \text{for the Normal kernel} \\ 1.05\sigma N^{-1/5} & \text{for the Epanechnikov kernel} \end{cases}$$

- The optimal bandwidth can be estimated using the sample standard deviation instead of σ

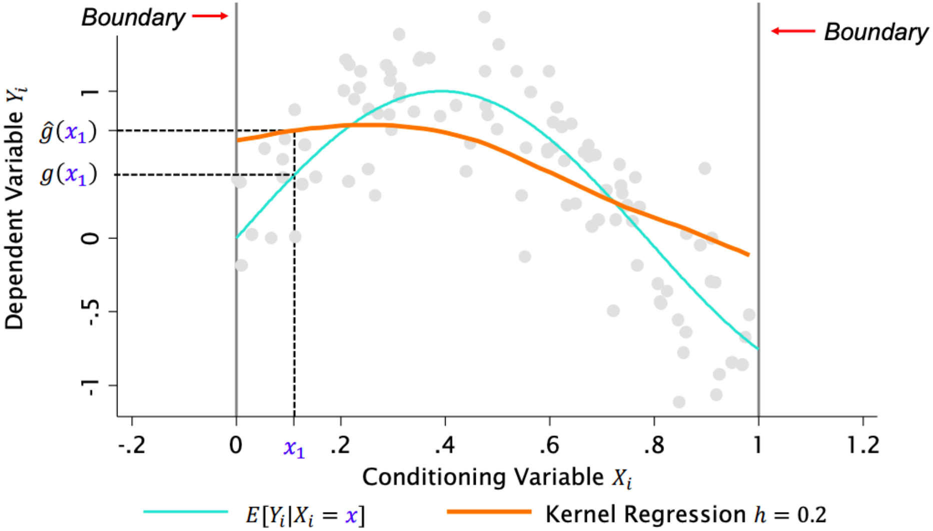
Kolmogorov-Smirnov Test

- We've talked about how to test hypotheses about averages from two different samples, as well as how to test sharp hypotheses using randomization inference
- The Kolmogorov-Smirnov test lets us test whether the distributions of two samples are the same using the empirical CDFs of the two samples ($\hat{F}_N(x)$ and $\hat{G}_N(x)$)
- The test statistic is simple: $D = \max_x |\hat{F}_N(x) - \hat{G}_N(x)|$
- We can also test for first order stochastic dominance of one distribution over the other with: $D_{\textit{1-sided}} = \max_x \hat{F}_N(x) - \hat{G}_N(x)$
- This gives additional information about treatment effects (and heterogeneity) that differences in means miss

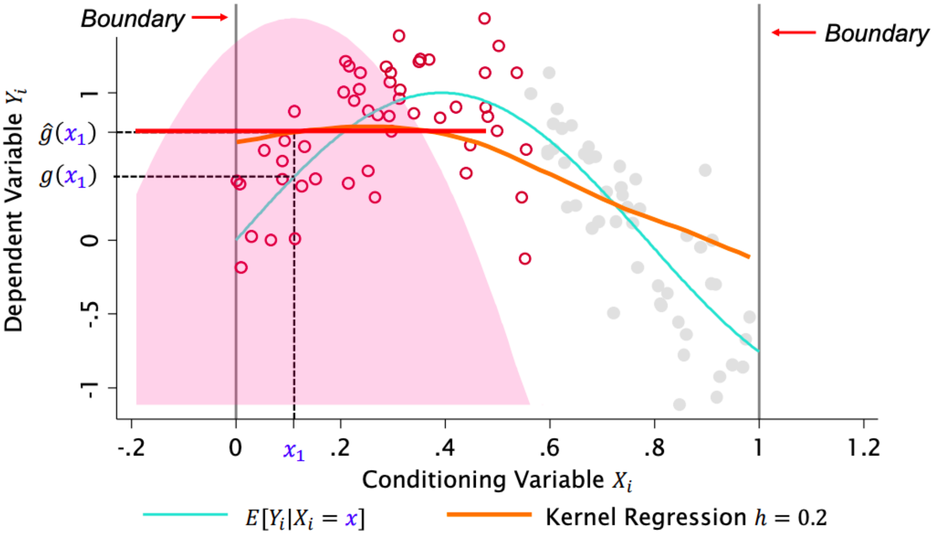
Overview

- Nonparametric regression estimates the conditional expectation function without any functional form assumptions
- Binned scatter plots are the simplest and most transparent nonparametric estimates of the CEF
- k-nearest neighbors and kernel regression estimate the CEF as a continuous function and smooth the estimates, but perform poorly at the boundaries of the data's support
- Local linear regression removes bias at the boundary by taking into account the local slope of the data
- Just like with nonparametric density estimation, the choice of bandwidth in nonparametric regression balances bias and variance

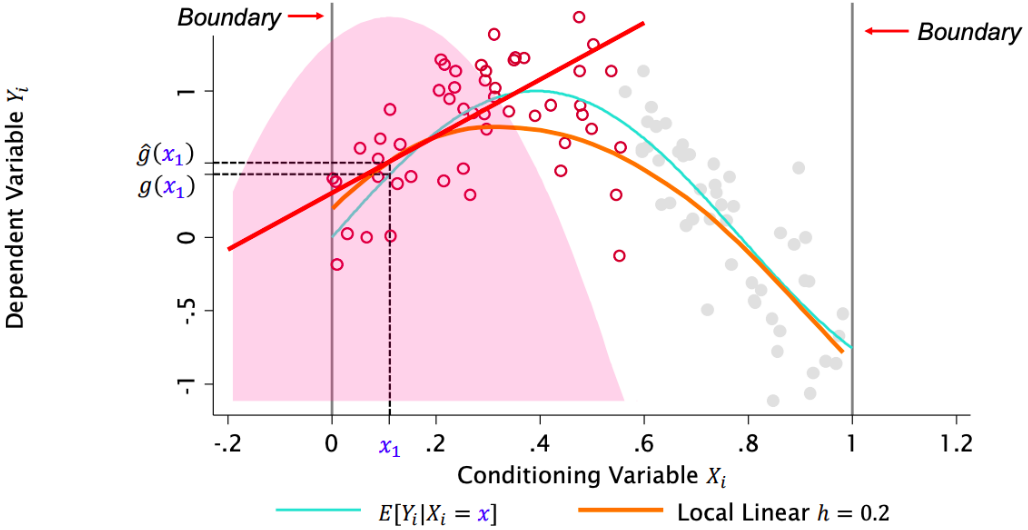
Bias of Kernel Regression at Boundary Points



Bias of Kernel Regression at Boundary Points



Local Linear Regression Removes Bias at the Boundary



Optimal Bandwidth Selection

- The objective function that determines optimal bandwidth depends on our application
- If we care about getting the CEF right at a single point (like in an RD), then the loss function should just depend on prediction error at that point

$$h_{opt} = \arg\min E[\underbrace{(g(x_0) - \hat{g}(x_0))^2}_{\textit{loss function}}]$$

- If we care about fitting the entire CEF, we need a global measure of fit. Fan and Gijbels (1996) use the weighted integrated squared error loss function:

$$WISE(h) = \int (\hat{g}(x) - g(x))^2 \, w(x) dx$$

- $w(x) > 0$ is a weighting function – we might want to penalize prediction errors more for values of x where the data is dense

Optimal Bandwidth: Fan and Gijbels

- Just like with the optimal kernel density bandwidth, we can derive the optimal local linear regression bandwidth by decomposing the expected WISE into bias and variance terms, taking Taylor series approximations, and minimizing with respect to h
- The result is:

$$h_{opt} = N^{-1/5} \underbrace{k_2^{-2/5} \left(\int K(u)^2 du \right)^{1/5}}_{constant} \underbrace{\left(\int \left(g''(x)\right)^2 w(x) dx \right)^{-1/5}}_{unknown} \underbrace{\left(\int \frac{\sigma^2(x)}{f(x)} w(x) dx \right)^{1/5}}_{unknown}$$

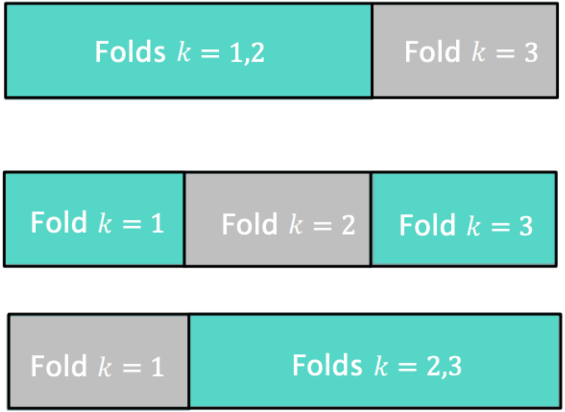
- How do we estimate the unknowns of this expression? Fan and Gijbels propose fitting a global 4th-order polynomial to calculate $\hat{\sigma}^2$ and $\hat{g}''(x)$. Choosing $w(x) = f(x)$ yields

$$\hat{h}_{opt} = N^{-1/5} C(K) \left(\frac{\hat{\sigma}^2 \times Range(X_i)}{\frac{1}{N} \sum_i (\hat{g}''(X_i))^2} \right)^{1/5}$$

Optimal Bandwidth: Cross Validation

- Another approach to choosing the optimal bandwidth is to use k-fold cross validation
- Idea is to estimate $g(x)$ using various bandwidths on subset of the data, then evaluate out-of-sample prediction error using data that did not contribute to the estimation
- Then choose bandwidth that minimizes this prediction error:

$$h_{opt} = \arg\min_{h} \sum_{i=1}^{N} (\hat{g}_{h,k}(x) - Y_i)^2$$



Gregory A. Bruich, Ph.D. | December 2, 2020 | Page 62