Nonparametrics

Introduction

Empirical distribution function

Density estimation

Conditional expectations

Introduction

Semi- and nonparametric methods

So far, we have seen models with a lot of structure

- OLS
- logit
- Manski
- Roy model

Nonparametric

Unknown regression function

Semiparametric BC: Klein and Spady

$$Y^* = X\beta + u$$

$$Y = \begin{cases} 1 & \text{if } Y^* > 0 \\ 0 & \text{if } Y^* \leq 0 \end{cases}$$

$$u|X \sim F_{u|X}$$

and $u \perp X$.

- How does it differ from Manski?

KS result

- Identification of $(\beta_0, F_{u|X})$
  - up to one scale normalization
  - $\sigma = 1$ or $\beta_1 = 1$
- $\sqrt{n}$-asymptotically Normal for β

How it works

Maximum likelihood.

$$\mathcal{L}_n(b, F) = 1/n \left(\sum_i Y_i log F(X_i b) + (1 - Y_i) log\,(1 - F(X_i b)) \right)$$

- Estimate β and F simultaneously
- Estimate F by a *sieve* or *kernel* approach

Regression discontinuity design

[as motivation]

Two approaches

- Global: sieves
- Local: kernels

We will focus on kernels.

Takeaway

Apply to local and global approaches:

- there are tools for estimating functions without making parametric assumptions.
- estimators converge more slowly than $\sqrt{n}$
- they involve choosing *tuning parameters*

Empirical distribution function

EDF

Let $X_i,\ i = 1, \cdots, n$ be a random sample from a distribution with cumulative distribution function $F(x) = P(X \leq x)$

Define the **empirical distribution function** (EDF) as

$$\hat{F}_n(x) = \frac{1}{n} \sum_i 1\{X_i \leq x\},$$

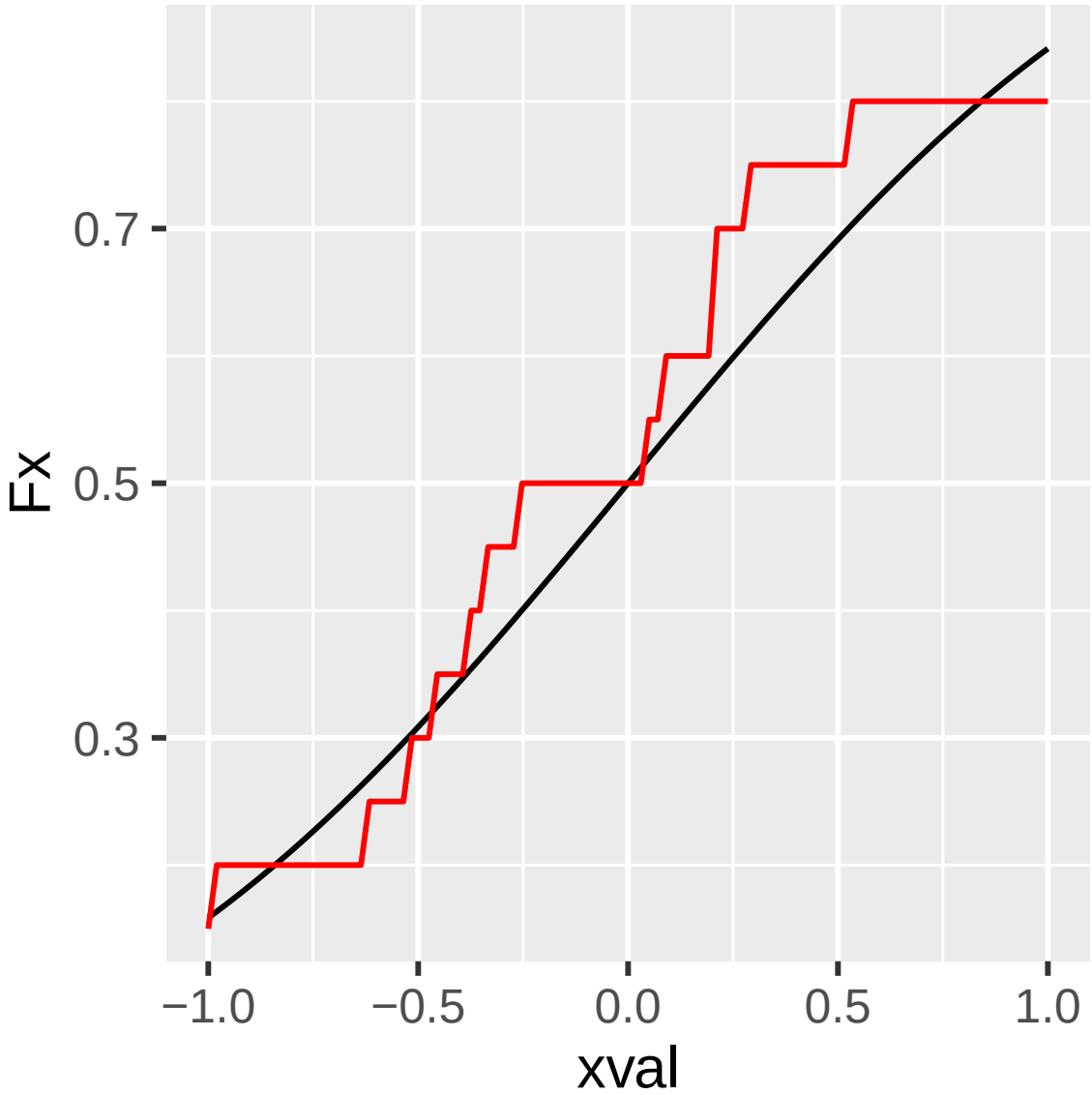
the sample analog at each value of x.

EDF

```
x <- rnorm(20,0,1)
xval <- seq(from=-1,to=1,length=100)
Fx <- pnorm(xval)
EDF <- function(x,data){mean(data <= x)}
Fhatx <- sapply(xval,EDF,data=x)
library(ggplot2)
```

(. . .)

```
qplot(x=xval,y=Fx,geom="line") + geom_line(y=Fhatx,colour="
```



EDF: Unbiasedness

Because

$$F(x) = P(X \leq x) = E(1\{X_i \leq x\})$$

it follows that

$$\begin{aligned} E(\hat{F}_n(x)) &= E\left(\frac{1}{n}\sum_i 1\{X_i \leq x\}\right) \\ &= \frac{1}{n}\sum_i E\left(1\{X_i \leq x\}\right) \\ &= F(x) \end{aligned}$$

so that $\hat{F}_n(x)$ is unbiased

- at each point x
- without imposing structure on $\mathcal{F} \ni F$ and

Glivenko-Cantelli

A stronger property holds:

For any $F(\cdot)$,

$$\sup |\hat{F}_n(x) - F(x)| \xrightarrow{a.s.} 0.$$

i.e. the EDF converges uniformly to the CDF.

- This uniform law of large numbers is called the **Glivenko-Cantelli theorem**
- It plays an important role in empirical process theory
- More details in e.g. Shaikh's notes

Density estimation

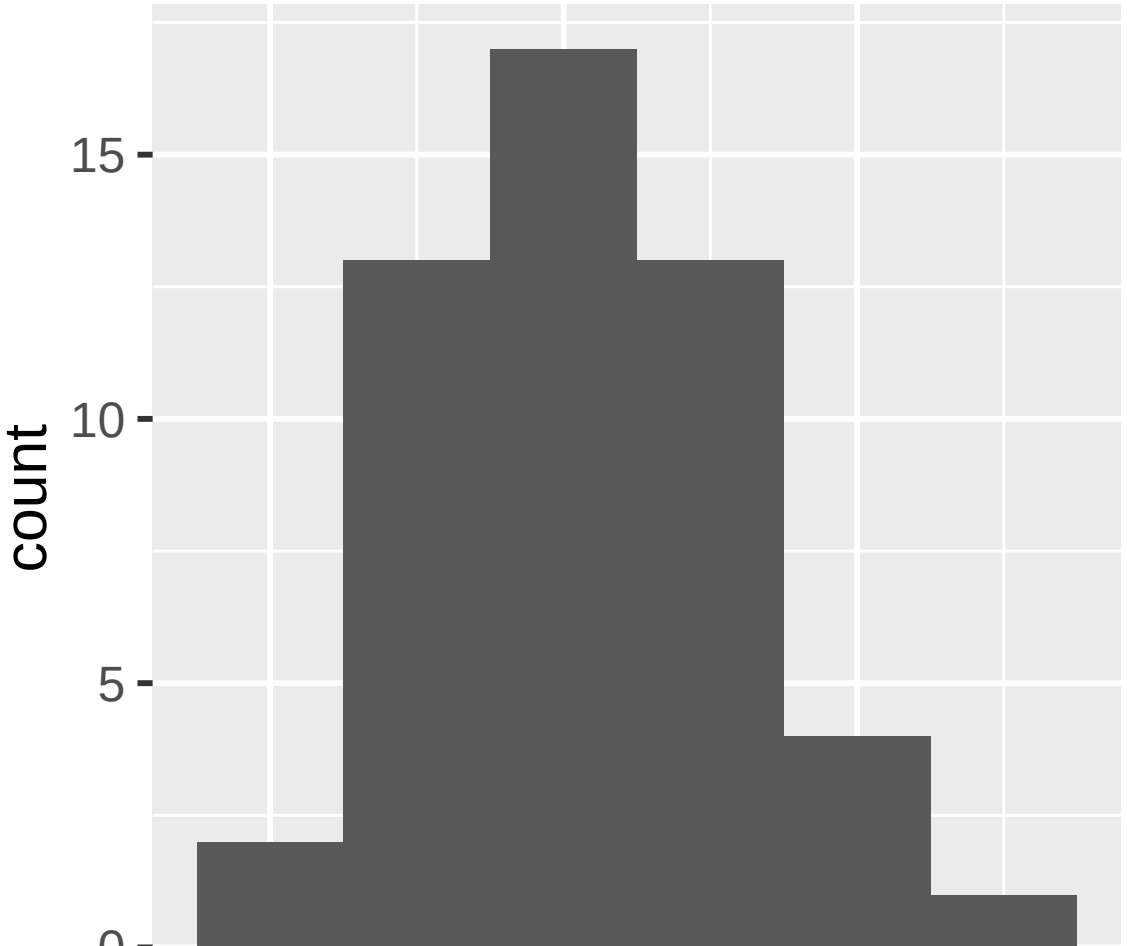
From cdf to pdf

- CDF: use EDF, with with great properties
- Reason: we can use all observations $X_i \leq x$ to estimate $F(x)$
- Assume continuous X and consider the pdf $f(x) = \frac{\partial F(x)}{\partial x}$
- We can only use data **at** x to estimate $f(x)$ unbiasedly

Histogram estimator

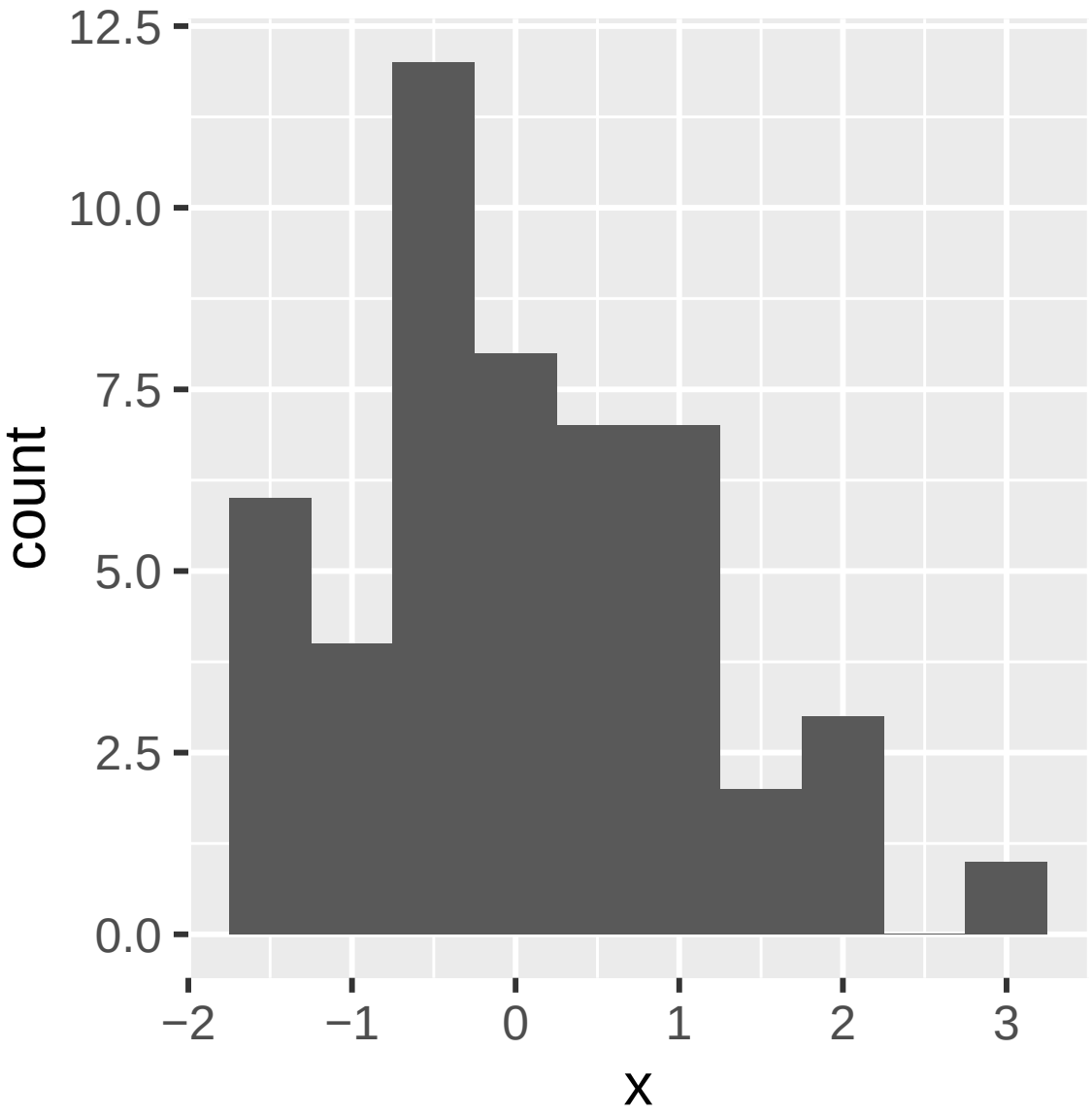
To estimate the function $f(x)$ at points $(x_0, x_1, \cdots, x_k)$, we can use a histogram.

```
df <- data.frame(x=rnorm(50,0,1))
ggplot(data=df,aes(x=x)) + geom_histogram(binwidth = 1.0,ce
```



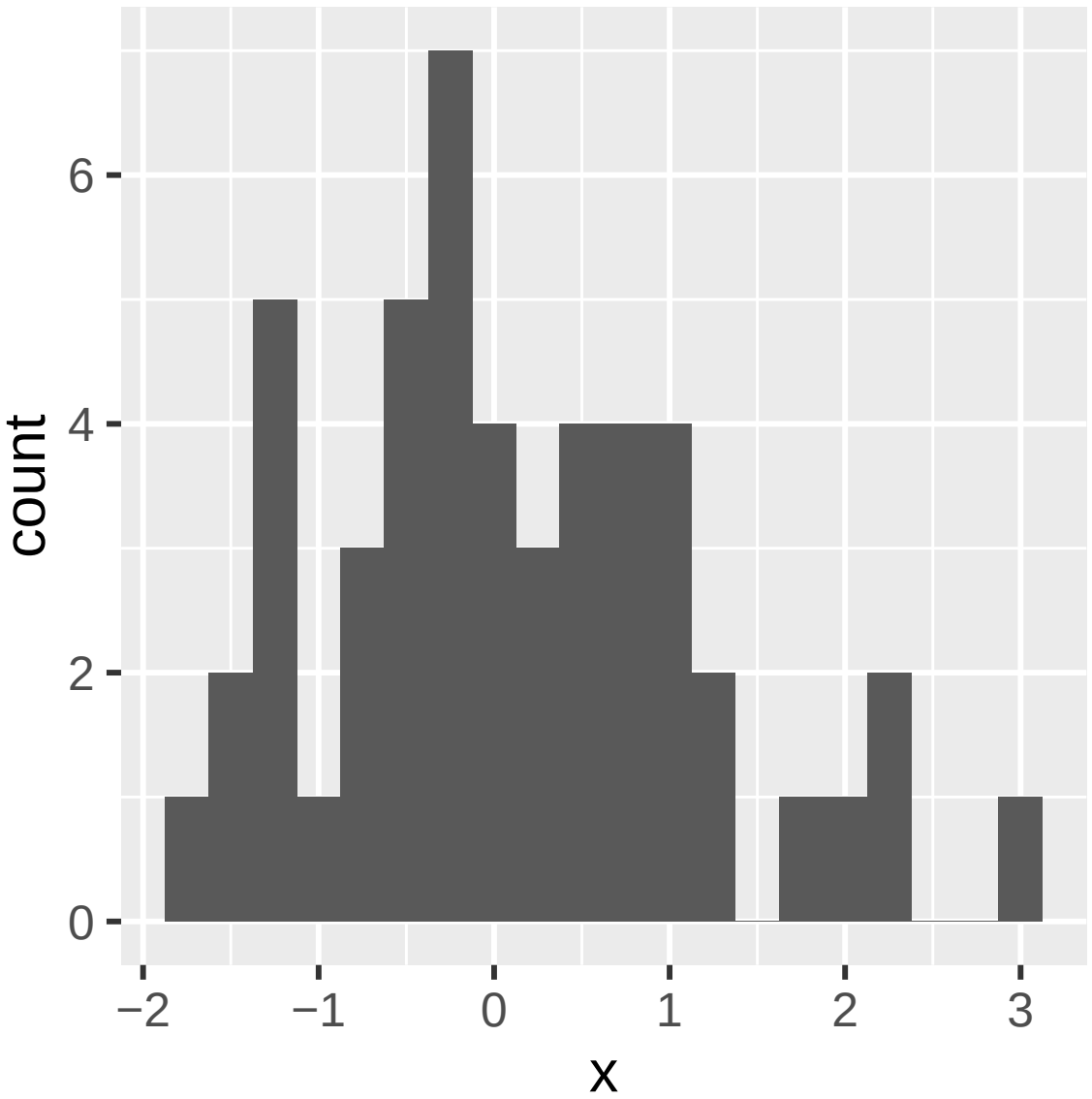
(...)

```
ggplot(data=df,aes(x=x)) + geom_histogram(binwidth = 0.5,ce
```



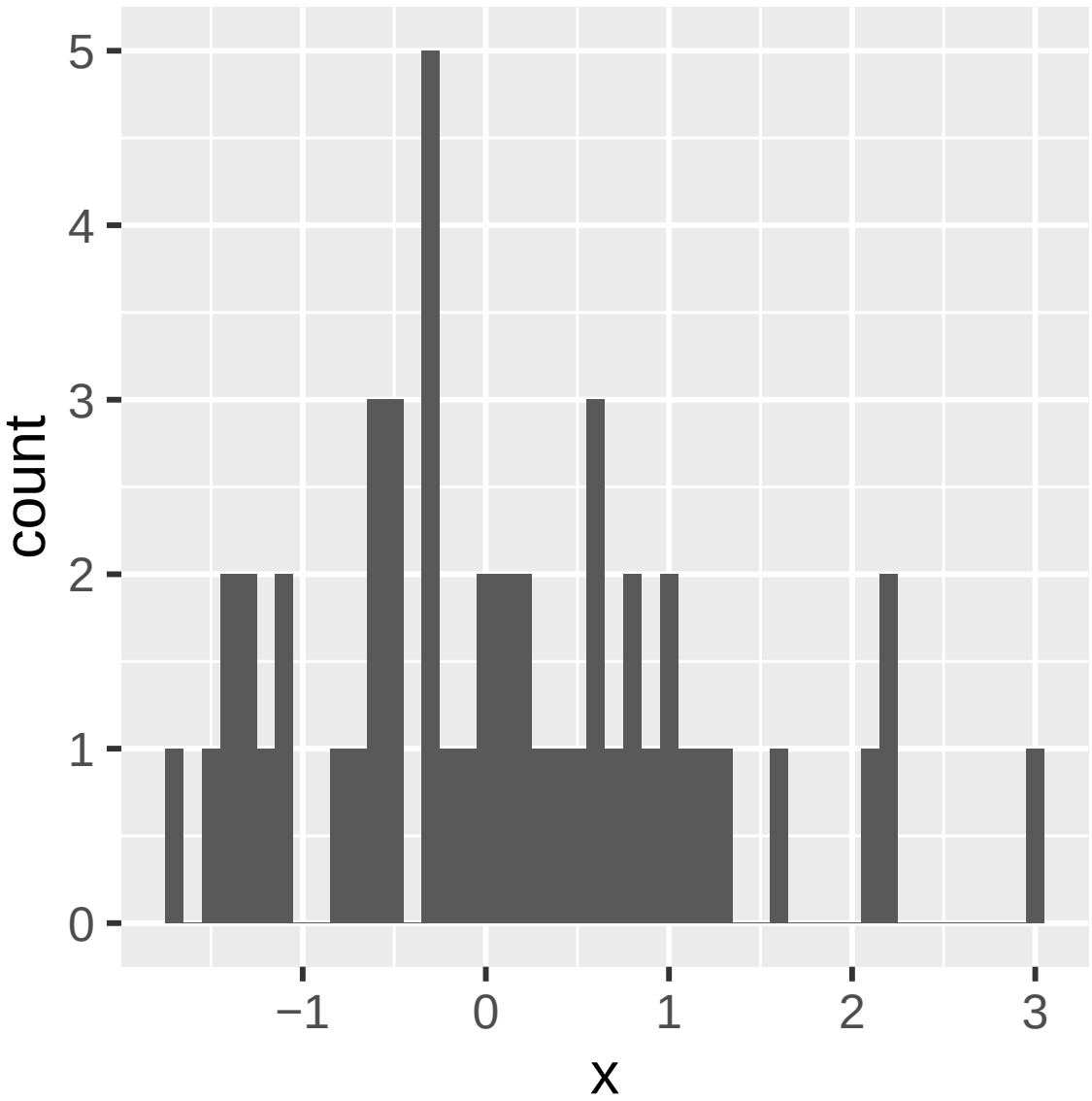
(...)

```
ggplot(data=df,aes(x=x)) + geom_histogram(binwidth = 0.25,c
```



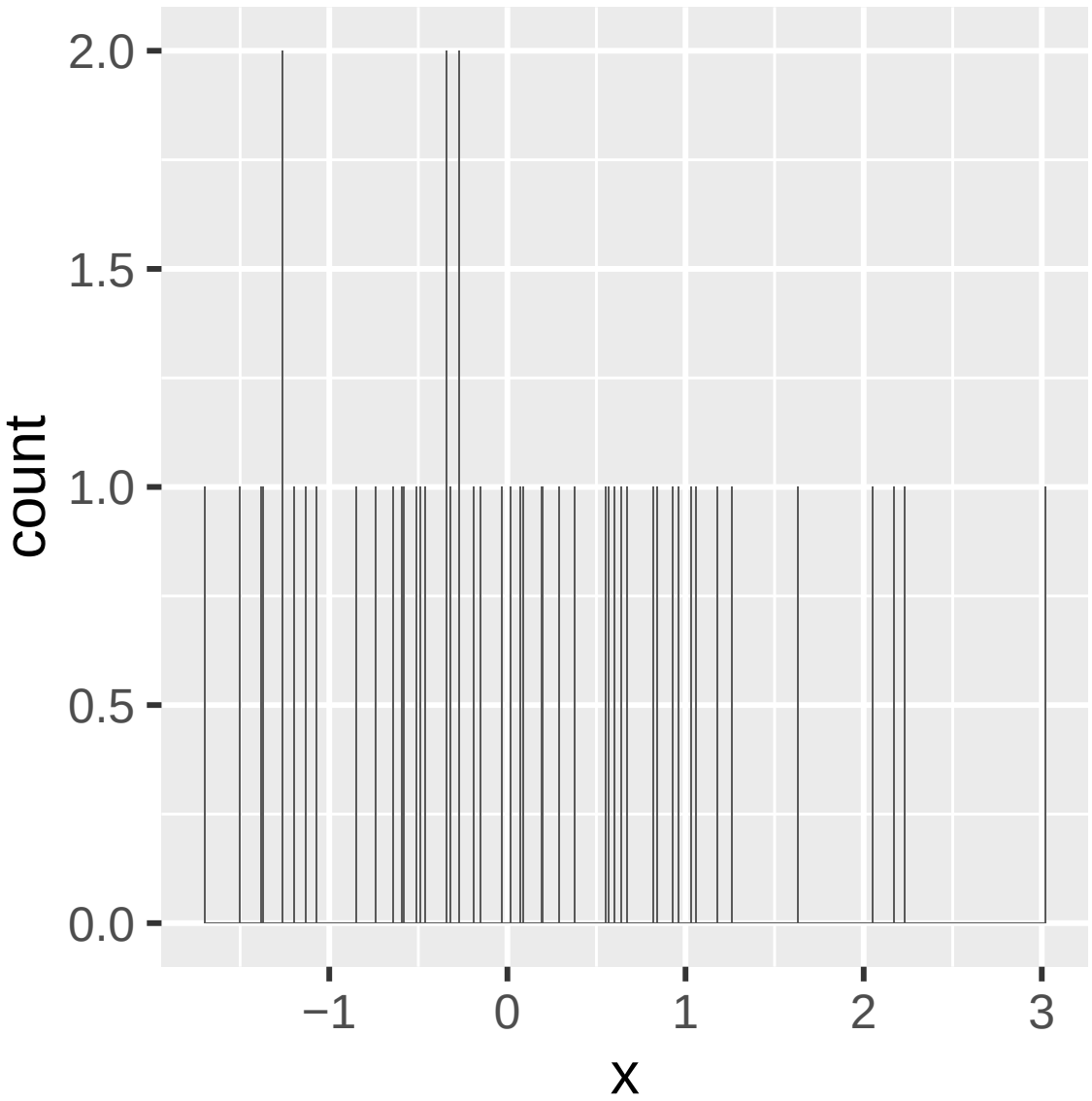
(...)

```
ggplot(data=df,aes(x=x)) + geom_histogram(binwidth = 0.1,ce
```



(...)

```
ggplot(data=df,aes(x=x)) + geom_histogram(binwidth = 0.01,c
```



Histogram approach

Graphs suggest the importance of the **bandwidth**

Kernel density estimator

The **kernel density estimator** for f at a point x is given by

$$\hat{f}_h(x) = \frac{1}{nh} \sum_i K\left(\frac{X_i - x}{h}\right)$$

where:

- uniform **kernel**:
$$K(u) = \frac{1}{2} 1\{|u| \leq 1\}$$
- Normalized count of X_is in the range $[x - h, x + h]$
- When is $\hat{f}_h(x) = 0$? 1?

Tuning parameters

- The **kernel** function K
- The **bandwidth** h or *smoothing parameter*

Bandwidth

The bandwidth h controls the smoothness of the density estimator by considering a narrow v wide range around x for estimating $f(x)$.

Bandwidth (2)

Experiment.

```
library(manipulate)
n <- 100
z <- runif(n,0,1)<=0.5
x <- z*rnorm(n,0,1) + (1-z)*rnorm(n,5,1)
manipulate(
  plot(density(x, bw)),
  bw = slider(0, 2, step = 0.01, initial = 0.5))
```

Kernel function

The kernel function determines how much weight is given to observations around x:

- uniform kernel
- Gaussian kernel, $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2)$
- Epanechnikov: $K(u) = \frac{3}{4}(1 - u^2)1\{|u| \leq 1\}$

Kernel functions (2)

We will assume that kernels

1. integrate to one, $\int K(u)du = 1$
2. are symmetric, $K(u) = K(-u)$
3. have bounded, non-zero variance $0 < \int u^2 K(u)du < \infty$
4. are bounded

Kernel functions (3)

```
manipulate(
  plot(density(x, 0.1, kernel=kernel)),
  kernel = picker("gaussian","epanechnikov","triangular")
)
```

Tuning parameters

- Bandwidth choice is more important than the kernel
- Pick a kernel that is smooth: $\hat{f}_h(x)$ inherits its properties!

KDE: bias

See e.g. Li and Racine (p.11):

$$
\begin{aligned}
E\left(\hat{f}_h(x)\right) &= E\left(\frac{1}{nh}\sum_i K\left(\frac{X_i - x}{h}\right)\right) \\
&= \frac{1}{h}E\left(K\left(\frac{X_i - x}{h}\right)\right) \\
&= \frac{1}{h}\int K\left(\frac{x_i - x}{h}\right) f(x_i)dx_i \\
&= \int K(u)\, f(x + uh)du
\end{aligned}
$$

- second equality from iidness
- fourth equality from change of variables $x_i = x + uh$

KDE: bias (2)

Proceed with a Taylor expansion of f around x, assuming the derivatives exist:

$$\begin{aligned}\int K(u) f(x+uh)du &= \int K(u)\left[f(x) + uhf'(x) + \frac{u^2h^2}{2}f''(x) + O(h^3)\right] \\ &= f(x)\int K(u)du + hf'(x)\int uK(u)du \\ &+ h^2 f''(x)\int u^2 K(u)du \\ &= f(x) + h^2 b(x) + O(h^3)\end{aligned}$$

where

$$b(x) = \frac{f''(x)}{2}\int u^2 K(u)du$$

KDE: bias (3)

1. It is biased!
2. Bias depends on h
3. Bias depends on properties of the kernel
4. Bias depends on the smoothness of f

NW: Variance

Similarly, we can show (Li and Racine, p. 12)

$$Var\left(\hat{f}_h(x)\right) = \frac{1}{nh}\left(v(x) + O(h)\right)$$

with

$$v(x) = f(x)\int (K(u))^2 du$$

1. Variance decreases as $h \to \infty$
2. Variance depends on the roughness of the kernel

MSE

As $n \to \infty$, $h \to 0$, $nh \to \infty$

$$\begin{aligned} MSE\left(\hat{f}_h(x)\right) &= bias^2 + Var \\ &= h^4 b^2(x) + \frac{1}{nh} v(x) + o(h^4 + (nh)^{-1}) \\ &= O(h^4 + (nh)^{-1}) \end{aligned}$$

Choice of bandwidth

- consistency requires, for some c, α,

$$h = cn^{-1/\alpha}$$

- optimal h has $c(x) = v(x)/(4b^2(x))$ and $\alpha = 5$
- in practice: cross-validation

Rate of convergence

A standard CLT is trouble:

$$\sqrt{n}(\hat{f}(x) - f(x) - h^2 b(x)) \overset{"d"}{\to} \mathcal{N}\left(0, \frac{1}{h} v(x)\right).$$

The limit is well-defined for fixed h, but not for our $h_n^* \to 0$.

ROC (2)

Remember: $h_n^* \sim n^{-1/5}$,

$$\sqrt{nh}(\hat{f}(x) - f(x) - h^2 b(x)) \xrightarrow{d} \mathcal{N}(0, v(x)).$$

so the rate of convergence is $\sqrt{nh} \sim n^{2/5} < n^{1/2}$

- What does that mean?

Cross-validation

In practice, we choose a bandwidth based on minimizing the mean integrated squared error **MISE**

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

Terms:

1. Convolution kernel
2. Leave-one-out
3. Does not depend on h

CV: second term

In the second term,

$$\int \hat{f}_h(x) f(x) dx = E\left(\hat{f}_h(X)\right)$$

Estimating this by

$$\frac{1}{n} \sum_i \hat{f}_h(X_i)$$

does not work because $f(h)$ and X_i are dependent.

CV: second term (2)

Instead, use the leave-one-out estimator

$$\hat{f}_{-i}(X_i) = \frac{1}{h(n-1)} \sum_{j \neq i} K\left(\frac{X_j - X_i}{h}\right)$$

to estimate $E\left(\hat{f}_h(X)\right)$ by

$$\frac{1}{n} \sum_i \hat{f}_{-i}(X_i)$$

CV: third term

The third term can be written

$$\int \hat{f}_h(x)^2 dx = \frac{1}{n^2 h^2} \sum_i \sum_j \int K\left(\frac{X_i - x}{h}\right) K\left(\frac{X_j - x}{h}\right) dx$$
$$\equiv \frac{1}{n^2 h} \sum_i \sum_j \bar{K}\left(\frac{X_i - X_j}{h}\right)$$

where $\bar{K}$ is the *convolution kernel*, and can be obtained from K

CV: implementation

To minimize the MISE, CV chooses:

$$\hat{h} = \text{argmin} CV_f(h)$$

where

$$CV_f(h) = \frac{1}{n^2 h} \sum_i \sum_j \bar{K}\left(\frac{X_i - X_j}{h}\right) - \frac{2}{hn(n-1)} \sum_i \sum_{j \neq i} K\left(\frac{X_j - X_i}{h}\right)$$

It can be shown that $\hat{h}$ converges to h^*.

Conditional expectations

Multivariate density

Bivariate extension of NW estimates the joint density $f(x, y)$ using products of kernel:

$$\hat{f}_h(x, y) = \frac{1}{nh^2} \sum_i K\left(\frac{y - Y_i}{h}\right) K\left(\frac{x - X_i}{h}\right)$$

- Picture: boxes in $\mathbb{R}^2$

Multivariate

General formulation:

$$\hat{f}_h(x, y) = \frac{1}{nh_1 \cdots h_K} \sum_i \prod_{k=1}^{K} K\left(\frac{x_k - X_{ik}}{h_k}\right)$$

- Bias $\sim \sum_k h_k^2$
- Variance $\sim \frac{1}{nh_1 \cdots h_K}$ (Why?)
- Optimal bandwidth $\sim n^{-1/(K+4)}$
- Rate of convergence $\sim n^{2/(K+4)}$

Multivariate: CLT

For details, see Li and Racine (p. 29):

$$\sqrt{nh_1 \cdots h_K} \left(\hat{f}(x) - f(x) - \text{bias} \right) \xrightarrow{d} \mathcal{N}$$

Conditional expectation

- our primary interest is in the conditional expectation function

$$m(x) \equiv E(Y|X = x) = \int yf(y|X = x)dy$$
$$= \int y\frac{f(y,x)}{f(x)}dy$$

- we can use univariate and multivariate density estimators

Nadaraya-Watson

Assume one bandwidth for x, y:

$$\begin{aligned}
\hat{m}(x) &= \int y \frac{\hat{f}_h(y,x)}{\hat{f}_h(x)} dy \\
&= \int y \frac{\frac{1}{nh^2} \sum_i K\left(\frac{X_i - x}{h}\right) K\left(\frac{Y_i - y}{h}\right)}{\frac{1}{nh} \sum_i K\left(\frac{X_i - x}{h}\right)} dy \\
&= \frac{\frac{1}{nh} \sum_i K\left(\frac{X_i - x}{h}\right) \int \frac{1}{h} y K\left(\frac{Y_i - y}{h}\right) dy}{\frac{1}{nh} \sum_i K\left(\frac{X_i - x}{h}\right)}
\end{aligned}$$

m: integral

Use symmetry of K,

$$\int \frac{1}{h} yK\left(\frac{Y_i - y}{h}\right) dy = \int \frac{1}{h} yK\left(\frac{y - Y_i}{h}\right) dy$$

and use the change of variables

$$\begin{aligned} u &= (y - Yi)/h \\ y &= Y_i + hu \\ dy &= hdu \end{aligned}$$

so that the integral above becomes

$$\begin{aligned} \int (Y_i + hu) K(u)\, du &= Y_i \int K(u)\, du + h \int uK(u)\, du \\ &= Y_i \end{aligned}$$

NW

Plugging in

$$\int \frac{1}{h} y K\left(\frac{Y_i - y}{h}\right) dy = Y_i$$

leads to

$$\hat{m}_h(x) = \frac{\frac{1}{nh}\sum_i K\left(\frac{X_i - x}{h}\right) Y_i}{\frac{1}{nh}\sum_i K\left(\frac{X_i - x}{h}\right)}$$

NW (2)

Picture: average Y in a small neighbourhood around x

- What happens when $h \to \infty$?
- What happens when $h \to 0$?

Bias, variance, optimal bandwidth, and rates of convergence are as for the density estimates. Relevant dimension is the dimension of the regressors X.

NW (3)

- Alternative way to write the estimator:

$$\hat{m}_h(x) = \text{argmin} \sum_i w_i \left(Y_i - b_0\right)^2,$$

  with $w_i = K\left(\frac{X_i - x}{h}\right)$.
- Locally weighted regression on a constant
- Solving the first order conditions yields:

$$\hat{m}_h(x) = \frac{\sum_i w_i Y_i}{\sum_i w_i}$$

Cross-validation

Works, with

$$CV(h) = \frac{1}{n} \left(Y_i - \hat{m}_{-i}(X_i)\right)^2 M(X_i)$$

where M is some weight function.

- the h that minimizes CV is asymptotically equivalent to the optimal bandwidth
- no analytics required!
- proof: involved

CV: implementation

(in two slides)

Local polynomial

Solve

$$\min \sum_i w_i (Y_i - b_0 - \sum_{p=1}^{P} b_p (X_i - x)^p)$$

- Fit a $p-$th order regression, locally.
- Picture: whiteboard
- Better convergence rates
- Better at the boundary

Demonstration

see “cross-validation.R”