Econometrics II

Lecture 2: Estimation Principles

David Schönholzer

Stockholm University

April 4, 2024

Plan for Today

Estimation Principles

- Last lecture: what can be learned?
- Today: how best to learn it?
- Goal is to find θ: *parameter*, *estimand*, *population estimator*
- We do so using $\widehat{\theta}$: *(sample) estimator*

- "Best" meaning:
  - Unbiased: $\mathbb{E}[\widehat{\theta}] = \theta$
  - Consistent: $\widehat{\theta} \xrightarrow{p} \theta$
  - Efficient: $Var(\widehat{\theta})$ as small as possible (but no smaller)

- Begin with *extremum estimators*
  - Covers large class of nonlinear estimators
  - Useful to illustrate general estimation principles

Table of Contents

Extremum Estimation

- Let $\mathbf{Z}_i$ be a matrix of data on i, e.g. $\mathbf{Z}_i = (Y_i, D_i, \mathbf{X}_i)$
- Want to maximize *population objective* $Q_0(\theta)$
- $\theta \in \Theta$ is parameter vector
- *Sample objective*: $\widehat{Q}_N(\theta, \mathbf{Z}_1, ..., \mathbf{Z}_N)$ with sample size N
- Define parameter of interest as:

  $$\theta_0 = \arg\max_{\theta \in \Theta} Q_0(\theta)$$

  where we assume the max is unique
- Extremum estimator maximize sample *criterion function*:

  $$\widehat{\theta} = \arg\max_{\theta \in \Theta} \widehat{Q}_N(\theta)$$

Examples of Extremum Estimation

- Example 1: OLS
  - $\mathbf{Z}_i = (Y_i, \mathbf{X}_i)$
  - θ is projection coefficient of Y_i on $\mathbf{X}_i$
  - $Q_0(\theta) = -\mathbb{E}\left[(Y_i - \mathbf{X}_i'\theta)^2\right]$ and $\theta_0 = \mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i'\right]^{-1}\mathbb{E}\left[\mathbf{X}_i Y_i'\right]$
  - $\widehat{Q}_N(\theta) = -\frac{1}{N}\sum_i^N (Y_i - \mathbf{X}_i'\theta)^2$

- Example 2: Nonlinear LS
  - Nonlinear parametric model $\mu(\mathbf{X}_i, \theta)$ for CEF
  - $Q_0(\theta) = -\mathbb{E}\left[(Y_i - \mu(\mathbf{X}_i, \theta))^2\right]$
  - $\widehat{Q}_N(\theta) = -\frac{1}{N}\sum_{i=1}^N (Y_i - \mu(\mathbf{X}_i, \theta))^2$

- But could be any estimator expressed with $Q_0(\theta)$

Consistency of Extremum Estimators

Definition (Uniform convergence in probability)

$\widehat{Q}_N(\theta)$ converges uniformly to $Q_0(\theta)$ if

$$\sup_{\theta \in \Theta} \left| \widehat{Q}_N(\theta) - Q_0(\theta) \right| \xrightarrow{p} 0.$$

Theorem (Consistency of Extremum Estimators)

If (i) $Q_0(\theta)$ is uniquely maximized at θ_0, (ii) Θ is compact, (iii) $Q_0(\theta)$ is continuous, and (iv) $\widehat{Q}_N(\theta)$ converges uniformly to $Q_0(\theta)$, then $\widehat{\theta} \xrightarrow{p} \theta_0$.

Table of Contents

Extremum Estimator 1: Classical Minimum Distance

- Sample objective:

$$\widehat{Q}_N(\theta) = -\left[\widehat{\boldsymbol{\pi}} - \mathbf{h}(\theta)\right]' \widehat{\mathbf{W}} \left[\widehat{\boldsymbol{\pi}} - \mathbf{h}(\theta)\right],$$

- where $\widehat{\boldsymbol{\pi}} \xrightarrow{p} \boldsymbol{\pi}$ is a vector of "reduced form" moments, e.g.
  - means of some variables of interest
  - covariances (recall variance component estimation)
  - other functions of the data
- $\mathbf{h}(\theta)$ is a *structural function* from model predictions
- $\widehat{\mathbf{W}} \xrightarrow{p} \mathbf{W}$ is a symmetric weighting matrix
  $\rightarrow$ e.g. $\mathbf{W} = \mathbf{I}$ or inverse variance weighting

- Hence, $\widehat{Q}_N(\theta)$: "squared distance" between data and model

Example CMD Estimation

- Example from behavioral economics: Laibson et al. (2007)
- They document two facts:
  1. Individuals borrow through credit cards with high interest
  2. Accumulate wealth by the time they retire
- What preferences can explain this behavior?

- Present bias β, discounting δ, risk aversion ρ
- Model yields, given $\theta = (\beta, \delta, \rho)$, predictions for moments:
  1. Share of 21-30 year olds with credit card: $h_1(\theta)$
  2. Share annual income borrowed with credit card: $h_2(\theta)$
  3. Wealth of 51-60 year olds: $h_3(\theta)$

- In the data, observe shares and wealth: $\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3$
- Optimal choice $\widehat{\theta}$ quantifies preference parameters

Extremum Estimator 2: Generalized MM

- Generalized MM sample criterion function:

$$\widehat{Q}_N(\theta) = -\widehat{\mathbf{g}}(\theta)'\widehat{\mathbf{W}}\widehat{\mathbf{g}}(\theta)$$

  where $\widehat{\mathbf{g}}(\theta) = \frac{1}{N}\sum_i f(\mathbf{Z}_i, \theta)$ and weights $\widehat{\mathbf{W}}$
- E.g. if $f(\mathbf{Z}_i, \theta) = (Y_i - \mathbf{X}_i'\beta)\,\mathbf{X}_i$ would be OLS
- Population *moment conditions*:

$$\mathbf{g}(\theta) = \mathbb{E}\left[f(\mathbf{Z}_i, \theta)\right] = 0$$

- Often originates from economic FOC
  - Euler condition in macro
  - Nash equilibrium in game

Extremum Estimator 3: Maximum Likelihood

- Call $\ell\left(\mathbf{Z}_i, \theta\right)$ the *log likelihood* of observing $\mathbf{Z}_i$ given θ
- Sample criterion:
$$\widehat{Q}_N(\theta) = \frac{1}{N}\sum_i \ell\left(\mathbf{Z}_i, \theta\right)$$
- Population criterion: $Q(\theta) = \mathbb{E}\left[\ell\left(\mathbf{Z}_i, \theta\right)\right]$
- Maximizing $\widehat{Q}_N(\theta)$ solves:
$$\frac{1}{N}\sum_i \mathbf{s}\left(\mathbf{Z}_i, \widehat{\theta}_{\mathsf{ML}}\right) = 0$$
  where $\mathbf{s}\left(\mathbf{Z}_i, \theta\right) \equiv \nabla_\theta \ell\left(\mathbf{Z}_i, \theta_0\right)$ is the score
- Key element in MLE: fully characterize $f\left(\mathbf{Z}_i, \theta\right)$
- More than just (mean) independence assumptions!

Extremum Estimator 4: OLS

- Population criterion: $Q_0(\theta) = -\mathbb{E}\left[(Y_i - \mathbf{X}_i'\theta)^2\right]$
- Sample criterion: $\widehat{Q}_N(\theta) = -\frac{1}{N}\sum_i^N (Y_i - \mathbf{X}_i'\theta)^2$
- Unlike general case, this criterion has explicit solution:

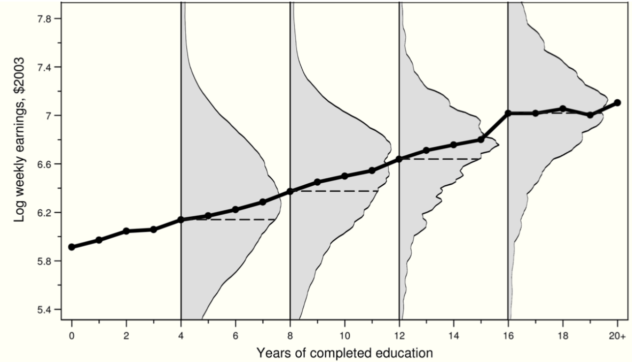
$$\begin{aligned}\widehat{\theta} &= \left(\sum_{i=1}^{N} \mathbf{X}_i\mathbf{X}_i'\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i Y_i\right) \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1} \left(\mathbf{X}'\mathbf{Y}\right)\end{aligned}$$

  for $\mathbf{X} = [\mathbf{X}_1, ..., \mathbf{X}_N]'$ and $\mathbf{Y} = [Y_1, ..., Y_N]'$
- Corresponds to model $Y_i = \mathbf{X}_i'\theta_0 + \varepsilon_i$ with restrictions
  - Specifically, $\mathbb{E}[X_i\varepsilon_i] = 0$ and $\dim(\mathbf{X}) = K$
- Under some conditions (Econometrics I), $\widehat{\theta} \xrightarrow{p} \theta_0$ (consistent)
- Side note: in general, $\mathbb{E}[\widehat{\theta}] \neq \theta_0$ (biased)
  - Unless either (a) CEF is linear or (b) $\mathbf{X}_i$ are fixed
  - This is not of great practical importance

Table of Contents

Reminder: CEF

- Central object to summarize data: $\mathbb{E}[Y_i|X_i]$
- Population average association of outcome Y_i with X_i
- Recall $\mathbb{E}[Y_i|X_i]$ is random but $\mathbb{E}[Y_i|X_i = x]$ is fixed



Why do economists love CEF and OLS? Many useful properties

The CEF Decomposition Property

Define $\varepsilon_i \equiv Y_i - \mathbb{E}[Y_i|X_i]$. Then:

Theorem (The CEF Decomposition Property)

If we write

$$Y_i = \mathbb{E}[Y_i|X_i] + \varepsilon_i$$

it holds by definition that

(a) $\mathbb{E}[\varepsilon_i|X_i] = 0$, *and therefore*

(b) $Cov(\varepsilon_i, X_i) = 0$

$\rightarrow$ Any Y_i can be decomposed into:

1. A piece "explained" by X_i: the CEF
2. A piece uncorrelated with (any function of) X_i

The CEF Prediction Property

Theorem (The CEF Prediction Property)

Let $m(X_i)$ be any function of X_i with finite second moment. The CEF solves:

$$\mathbb{E}[Y_i|X_i] = \arg\min_{m(X_i)} \mathbb{E}\left[(Y_i - m(X_i))^2\right],$$

so it minimizes MSE of prediction of Y_i given X_i

$\rightarrow$ CEF is the best function of X_i to predict Y_i

OLS Justification 1: Linear CEF Theorem

It turns out population OLS is a great estimator of the CEF

Recall population regression: $\beta_{\text{OLS}} \equiv \mathbb{E}[X_i X_i']^{-1}\mathbb{E}[X_i Y_i]$

- Defines linear projection $\mathbb{E}^*[Y_i|X_i] \equiv X_i'\beta_{\text{OLS}}$

Theorem (The Linear CEF Theorem)

Suppose the CEF is linear. Then

$$\mathbb{E}[Y_i|X_i] = \mathbb{E}^*[Y_i|X_i]$$

$\rightarrow$ OLS is great for linear CEF. But when is it linear?

- Multivariate Normal distributions
- Saturated models (see later today): one dummy for each possible value of CEF

OLS Justification 2: Best Linear Predictor

OLS is also good at predicting $Y_i|X_i$ directly:

Theorem (The Best-Linear-Predictor Theorem)

$\mathbb{E}^*[Y_i|X_i]$ *minimizes MSE of linear prediction of* Y_i *given* X_i

$\rightarrow$ CEF is best function predicting $Y_i|X_i$
$\rightarrow$ OLS is best *linear* function predicting $Y_i|X_i$

OLS Justification 3: Regression-CEF Relationship

Even when CEF is nonlinear, OLS is still good at predicting it:

Theorem (Regression-CEF Theorem)

$\mathbb{E}^*[Y_i|X_i]$ *minimizes MSE of any linear approximation of CEF, i.e.*

$$\beta_{OLS} = \arg\min_b \mathbb{E}\left[\left(\mathbb{E}[Y_i|X_i] - X_i'b\right)^2\right]$$

OLS Justification 4: Law of Iterated Projections

Linear projections have equivalent property to LIE:

1. Long regression: $\mathbb{E}^*[Y_i|W_i, Z_i] = W_i\beta + Z_i\gamma$
2. Short regression: $\mathbb{E}^*[Y_i|W_i] = W_i\delta$
3. Auxiliary regression: $\mathbb{E}^*[Z_i|W_i] = W_i\pi$

Theorem (Law of Iterated Projections)

$\mathbb{E}^*[Y_i|W_i] = \mathbb{E}^*[\mathbb{E}^*[Y_i|W_i, Z_i]|W_i]$ *which implies* $\delta = \beta + \pi\gamma$

Proof of implication:

$$\begin{aligned}
\mathbb{E}^*[Y_i|W_i] &= \mathbb{E}^*[W_i\beta + Z_i\gamma|W_i] \\
&= \mathbb{E}^*[W_i|W_i]\beta + \mathbb{E}^*[Z_i|W_i]\gamma \\
&= W_i\beta + (W_i\pi)\,\gamma = W_i\,(\beta + \pi\gamma)
\end{aligned}$$

Illustration of LIP

```
clear
set seed 1234
set obs 1000
gen z = rnormal()
gen w = z + rnormal()
gen y = .5*w + .5*z + rnormal()
eststo lr:  reg y w z // long regression
local beta = _b[w]
local gamma = _b[z]
eststo sr:  reg y w // short regression
local delta = _b[w]
eststo ar:  reg z w // auxiliary regression
local pi = _b[w]
esttab lr sr ar, cells(b(fmt(a2)) se(par))
```

Results from LIP Simulation

	(1)	(2)	(3)
	y	y	z
w	0.44	0.76	0.53
	(0.033)	(0.024)	(0.016)
z	0.59		
	(0.044)		
_cons	0.022	0.037	0.025
	(0.031)	(0.034)	(0.022)
N	1000	1000	1000

- As predicted by the LIP:

$$0.76 = 0.44 + 0.53 \times 0.59$$
$$\widehat{\delta} = \widehat{\beta} + \widehat{\pi} \times \widehat{\gamma}$$

- Useful to think about omitted variable bias

OLS Justification 5: Frisch-Waugh-Lovell

- Recall the long regression: $\mathbb{E}^*[Y_i|W_i, Z_i] = W_i\beta + Z_i\gamma$
- Residuals: $\tilde{Y}_i \equiv Y_i - \mathbb{E}^*[Y_i|Z_i]$
- $\tilde{W}_i \equiv W_i - \mathbb{E}^*[W_i|Z_i]$

Theorem (Frisch-Waugh-Lovell)

$$\beta = \frac{\mathbb{E}[\tilde{W}_i \tilde{Y}_i]}{\mathbb{E}[\tilde{W}_i]^2}$$

$\rightarrow$ Recover β from long reg by running a residualized short reg
  - Extremely useful to visualize conditional relationships
- Multivariate versions of LIP and FWL also exist
- Both are mechanical results of OLS – work in every dataset!

Illustration of FWL

```
clear
set seed 1234
set obs 1000
gen z = rnormal()
gen w = z + rnormal()
gen y = .5*w + .5*z + rnormal()
eststo lr:  reg y w z // long regression
eststo far:  reg w z // flipped auxiliary regression
predict wres, res
eststo arr:  reg y z // other short regression
predict yres, res
eststo rr:  reg yres wres // residual regression
esttab lr far arr rr, cells(b(fmt(a2)) se(par))
```

Results from FWL Simulation

	(1)	(2)	(3)	(4)
	y	w	y	yres
w	0.44			
	(0.033)			
z	0.59	0.99	1.03	
	(0.044)	(0.030)	(0.033)	
wres				0.44
				(0.032)
_cons	0.022	-0.047	0.0015	5.4e-11
	(0.031)	(0.030)	(0.034)	(0.031)
N	1000	1000	1000	1000

$\rightarrow$ Useful when there are many controls

Application of FWL: Residualized Scatterplots

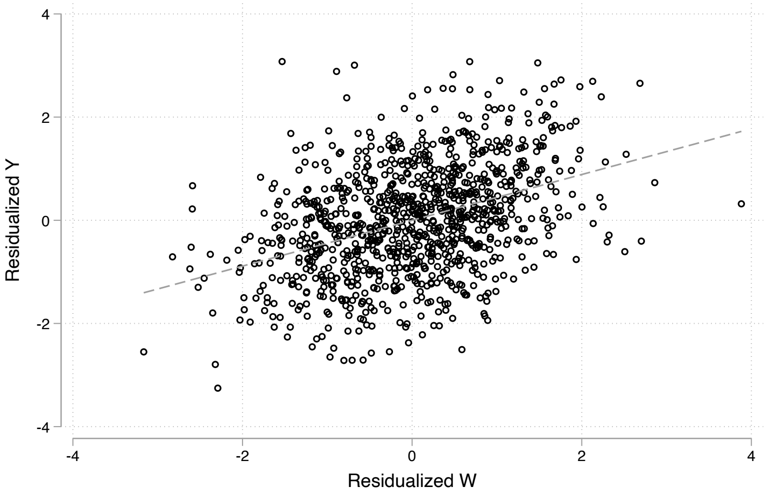


Table of Contents

OLS on Constant

- Important use of OLS: estimating means
- Simplest case: $Y_i = \mu + \varepsilon_i$
- Population OLS of this is $\beta_{\text{OLS}} = \mathbb{E}\left[Y_i\right]$
  - Convince yourself: $\mathbb{E}\left[X_i^2\right]^{-1}\mathbb{E}\left[X_i Y_i\right]$ with $X_i = 1$

- Sample OLS: $\widehat{\beta}_{\text{OLS}} = \frac{1}{N}\sum_{i=1}^{N} Y_i$
  - Again good exercise to evaluate $\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$
  - $N \times 1$ vectors $\mathbf{X} = (1, ..., 1)$ and $\mathbf{Y} = (Y_1, ..., Y_N)$

Analysis of Variance

- R.A. Fisher: do means across groups differ?
- Suppose we have a sample of wages Y_i
- We also have $X_i = 1\,[\text{foreign}]$ and $W_i = 1\,[\text{female}]$
- Suppose we run

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + \beta_3 X_i \times W_i + \varepsilon_i$$

- This is called a *saturated model*
- Number of coefficients = number of possible RHS values

	$X = 0$	$X = 1$
$W = 0$	(domestic, male)	(foreign, male)
$W = 1$	(domestic, female)	(foreign, female)

Interpreting Group Indicator Coefficients

- How do we interpret $(\beta_0, \beta_1, \beta_2, \beta_3)$?
- CEF is necessarily linear, and thus OLS = CEF
- CEF: $\mathbb{E}[Y_i | X_i = x, W_i = w]$
- Specifically:

$$\begin{aligned}
\beta_0 &= \mathbb{E}\left[Y_i | X_i = 0, W_i = 0\right] \\
\beta_0 + \beta_1 &= \mathbb{E}\left[Y_i | X_i = 1, W_i = 0\right] \\
\beta_0 + \beta_2 &= \mathbb{E}\left[Y_i | X_i = 0, W_i = 1\right] \\
\beta_0 + \beta_1 + \beta_2 + \beta_3 &= \mathbb{E}\left[Y_i | X_i = 1, W_i = 1\right]
\end{aligned}$$

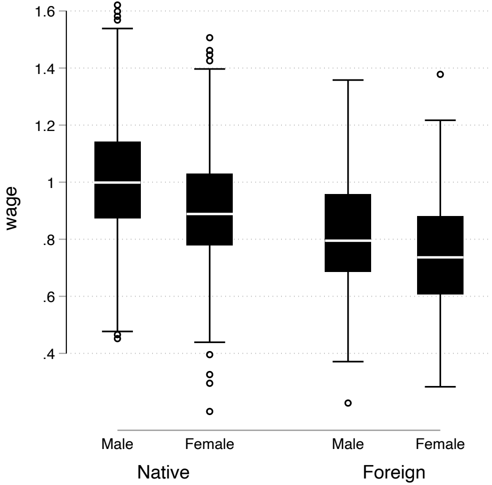
- There are other ways to parameterize same model, e.g.

$$Y_i = \gamma_0 X_i + \gamma_1 W_i + \gamma_2 (1 - X_i) \times W_i + \gamma_3 X_i \times W_i + \varepsilon_i$$

Illustration of Saturated Model

```
clear
set seed 123
set obs 1000
gen foreign = runiform() < .2
gen female = runiform() < .5
tab foreign female
gen wage = 1 - .1*female - .2*foreign + .05*foreign*female + .2*rnormal()
graph box wage, over(female, relabel(1 "Male" 2 "Female")) ///
      over(foreign, relabel(1 "Native" 2 "Foreign")) ///
      ylabel(.4(.2)1.6) xsize(4)
graph export figures/boxplot.pdf, replace
eststo sat:  reg wage foreign##female
esttab sat, cells(b(fmt(2)) se(par)) ///
      keep(1.foreign 1.female 1.foreign#1.female _cons) ///
      label
```

Results from Saturated Model Simulation



	(1) wage b/se
foreign=1	-0.20
	(0.02)
female=1	-0.11
	(0.01)
foreign=1 × female=1	0.05
	(0.03)
Constant	1.01
	(0.01)
Observations	1000

Many Means

- Consider now $X_i \in \{\xi_1, ..., \xi_J\}$ for large J (but $J < N$)
- ξ_j could be firm, or demographic group e.g. (foreign, female)
- All realizations of X_i: $\Pr\left(X_i = \xi_j\right) = \pi_j > 0$ and $\sum_j \pi_j = 1$
- We know that OLS = CEF if linear
- Thus OLS *is* $\mathbb{E}\left[Y_i | X_i = x\right]$ for all $x \in \{\xi_1, ..., \xi_J\}$

Method of Moments for Cell Means

- Can estimate using "cell means" (MM):

$$\widehat{\mathbb{E}}\left[Y_i|X_i=x\right]=\frac{\sum_i 1\left[X_i=x\right]Y_i}{\sum_i 1\left[X_i=x\right]}=\frac{\frac{1}{N}\sum_i 1\left[X_i=x\right]Y_i}{\frac{1}{N}\sum_i 1\left[X_i=x\right]}$$

- With a LLN:

$$\frac{1}{N}\sum_i 1\left[X_i=x\right]\xrightarrow{p}\mathbb{E}\left[1\left(X_i=x\right)\right]=\Pr\left(X_i=x\right)=\pi_j$$

$$\frac{1}{N}\sum_i 1\left[X_i=x\right]Y_i\xrightarrow{p}\mathbb{E}\left[Y_i\cdot 1\left(X_i=x\right)\right]=\mathbb{E}\left[Y_i|X_i=x\right]\pi_j$$

where the last step uses the LIE

OLS Estimates Cell Means

- So with continuity theorem

$$\frac{\frac{1}{N}\sum_i 1\left[X_i = x\right] Y_i}{\frac{1}{N}\sum_i 1\left[X_i = x\right]} \xrightarrow{p} \mathbb{E}\left[Y_i | X_i = x\right]$$

- Compare cell means to OLS of Y_i on $1\left[X_i = x\right]$ for all x:

$$\widehat{\beta}_{\text{OLS}} = \begin{bmatrix} \frac{\sum_i 1[X_i = \xi_1] Y_i}{\sum_i 1[X_i = \xi_1]} \\ \vdots \\ \frac{\sum_i 1[X_i = \xi_J] Y_i}{\sum_i 1[X_i = \xi_J]} \end{bmatrix}$$

- They are the same!
- So OLS estimates cell means for many groups

Table of Contents

Constructing Cells with a Window

- Consider scalar X_i but continuous with density $f(x)$
- Logic from before hard because $\Pr(X_i = x) = 0$
- So how can we approximate $\mathbb{E}[Y_i | X_i = x]$ best?
- We imitate the cell means logic
- Let's construct a small window $[x - h, x + h]$ for small $h > 0$
- h is called *bandwidth* or *window* – chosen/known by us

Bandwidth Estimation

- Let's estimate these "window cell means"

$$\widehat{\mathbb{E}}\left[Y_i|X_i = x\right] = \frac{\sum_i 1\left[x - h \leq X_i \leq x + h\right] \cdot Y_i}{\sum_i 1\left[x - h \leq X_i \leq x + h\right]}$$

- $\widehat{\mathbb{E}}\left[Y_i|X_i\right] \xrightarrow{p} \mathbb{E}\left[Y_i|X_i\right]$ as N gets large and h small
- But unless $\mathbb{E}\left[Y_i|X_i\right]$ constant in window, $\widehat{\mathbb{E}}\left[Y_i|X_i\right]$ biased
- On the other hand, variance increases as h shrinks
  - Intuitive: less observations in window
- Optimal h minimizing MSE infeasible: requires knowing $f(x)$
- Solution: use auxiliary density $K(x)$ (the "kernel")

Univariate Density Estimation

- Alternative approach for $\widehat{\mathbb{E}}\left[Y_i|X_i\right]$: for continuous Y_i and X_i

$$\mathbb{E}\left[Y_i|X_i = x\right] = \frac{\int y f_{X,Y}\left(x,y\right) dy}{\int f_{Y,X}\left(y,x\right) dy} = \frac{\int y f_{X,Y}\left(x,y\right) dy}{f\left(x\right)}$$

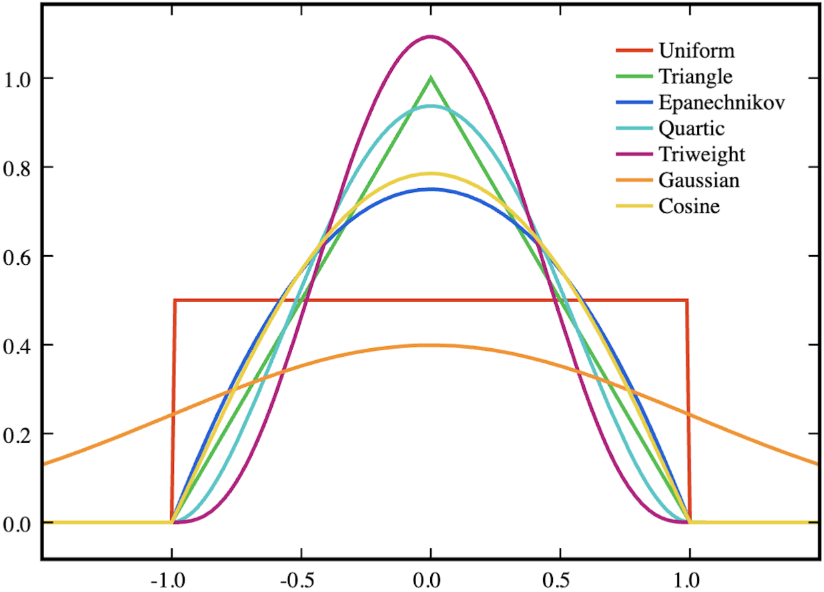
  so can estimate $f_{X,Y}\left(x,y\right)$ and $f\left(x\right)$ to get CEF too
- May also be interested in $f\left(x\right)$ in its own right
- CDF $F(x) = \Pr\left(X_i \leq x\right)$ and $\widehat{F}\left(x\right) = \frac{1}{N}\sum_{i=1}^{N} 1\left[X_i \leq x\right]$
- Definition of derivative: $f\left(x\right) = \lim_{h\to 0} \frac{F(x+h)-F(x)}{h}$
- Empirical equivalent: *histogram*

$$\widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{N} 1\left[x < X_i \leq x+h\right]$$

- Can use $K(\cdot)$ to construct continuous versions:

$$\widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{N} K\left(\frac{x - X_i}{h}\right)$$

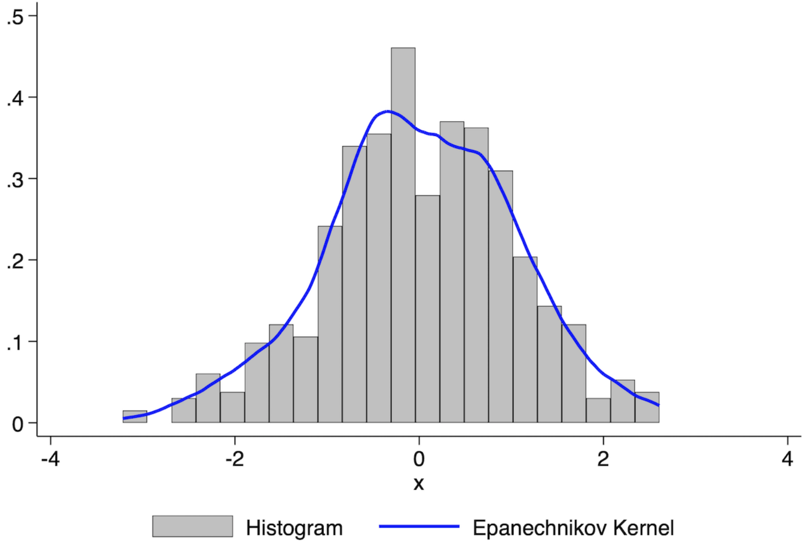
Many Choices for Smoothers $K(\cdot)$ (i.e. Kernels)



Examples of Density Estimation

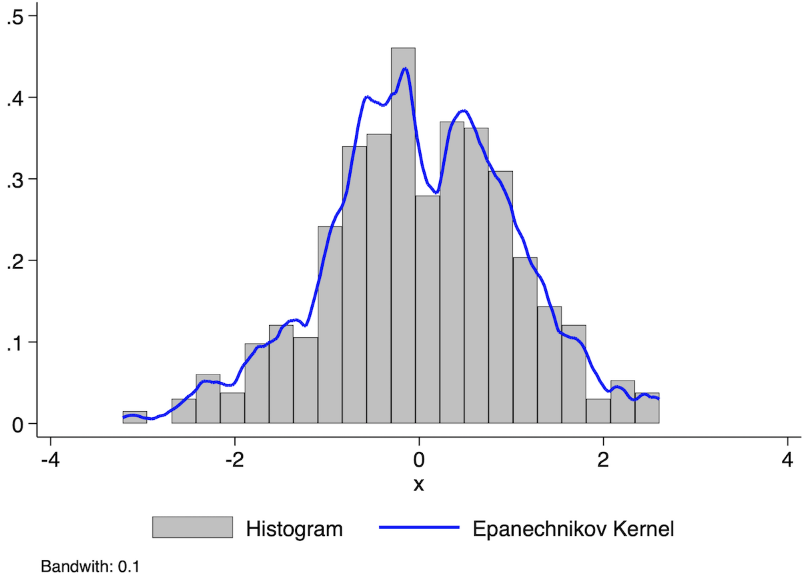
```
clear
set seed 1234
set obs 500
gen x = rnormal()
tw (histogram x, fc(gs12) lw(.1)) ///
(kdensity x, lc(blue) lw(.5)), ///
legend(label(1 "Histogram") label(2 "Epanechnikov Kernel")) ///
note(Bandwith:  optimal ( 0.25))
```

Optimal Bandwidth Kernel

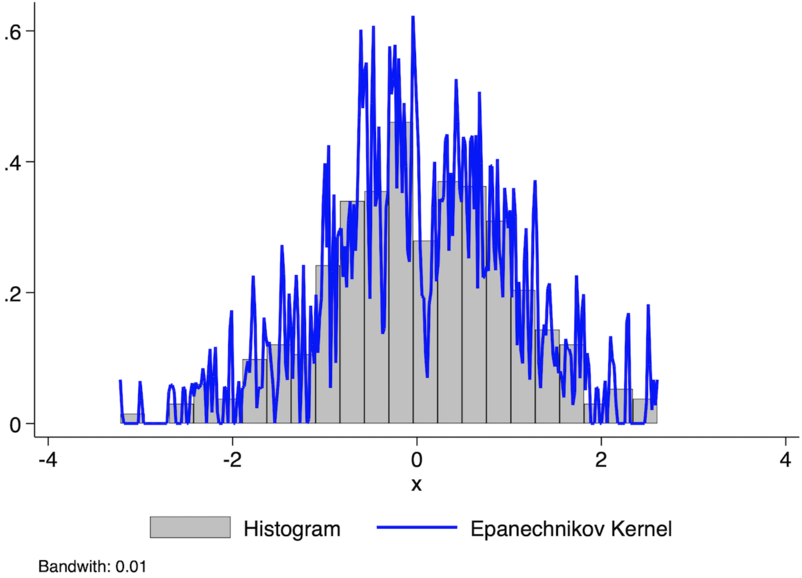


Bandwith: optimal (~0.25)

Smaller Bandwidth



Even Smaller



Large Bandwidth

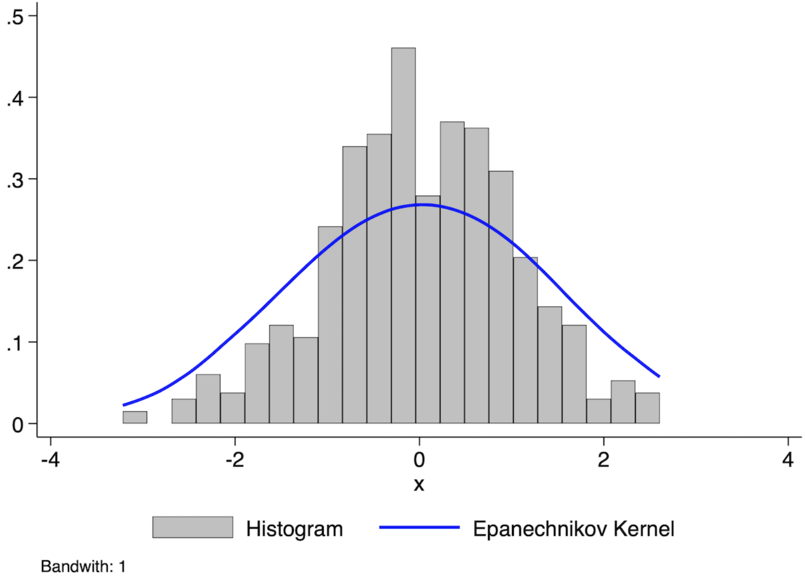
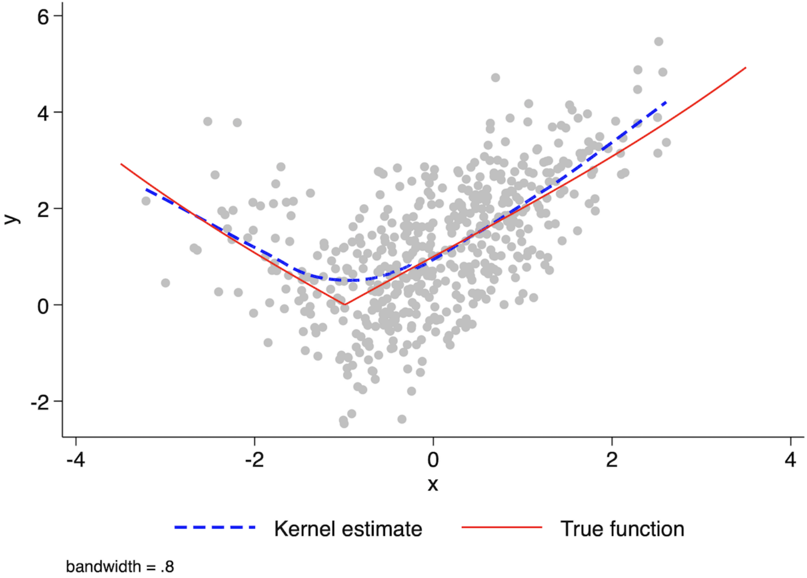


Table of Contents

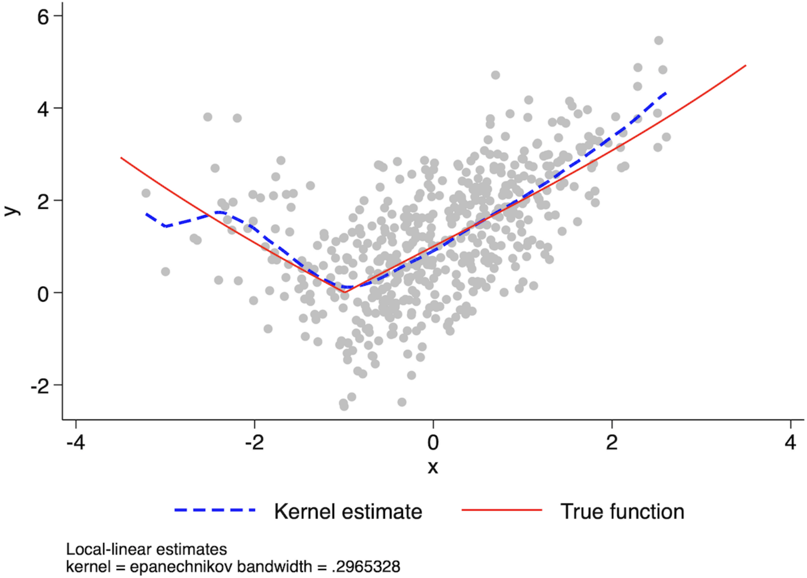
Simulating a Nonlinear CEF

```
clear
set seed 1234
set obs 500
gen x = rnormal()
gen y = abs(1 + x + .01*x^3) + rnormal()
* traditional LOWESS
lowess y x, ///
m(o) mc(gs12) lineopts(lc(blue) lw(.5)) ///
addplot(function y = abs(1 + x + .01*x^3), range(-3.5 3.5) lc(red))
///
legend(order(2 3) label(2 Kernel estimate) label(3 True function))
///
title("")
```

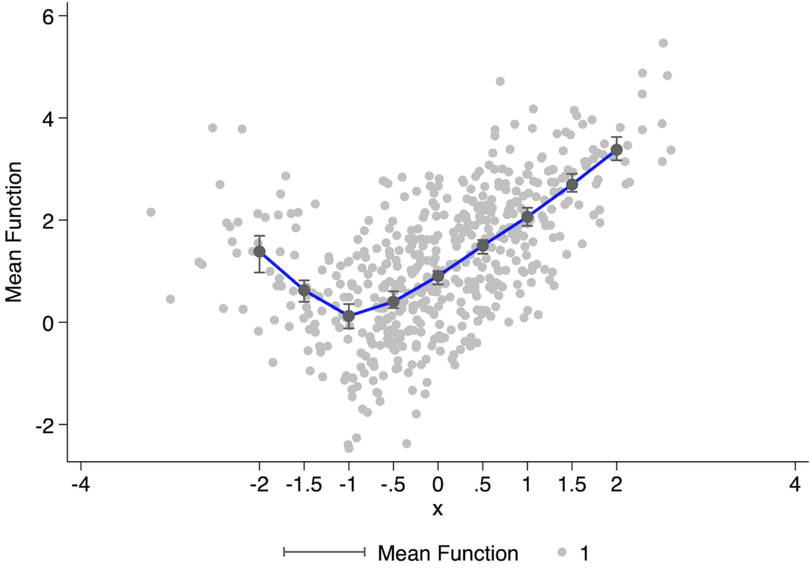
Locally Weighted Scatterplot Smoothing (LOWESS)



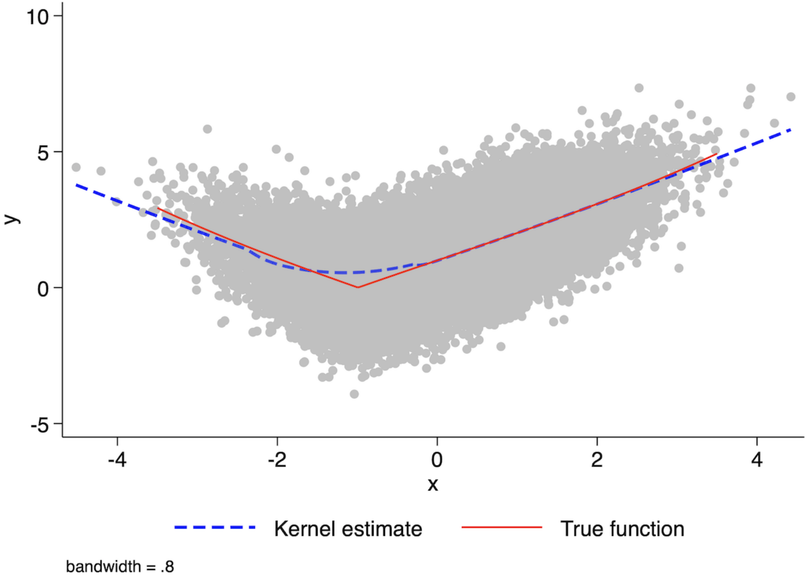
Modern Cell Means Smoother: `npregress`



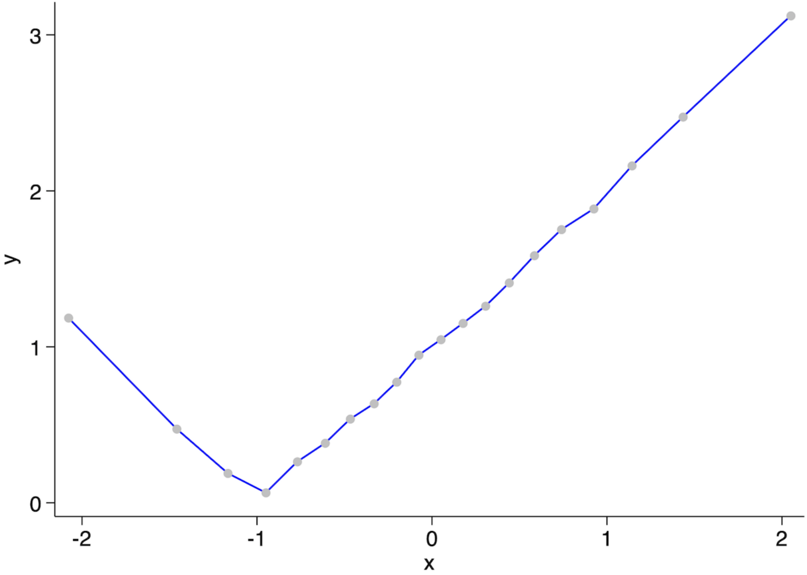
`npregress` Also Estimates Confidence Intervals



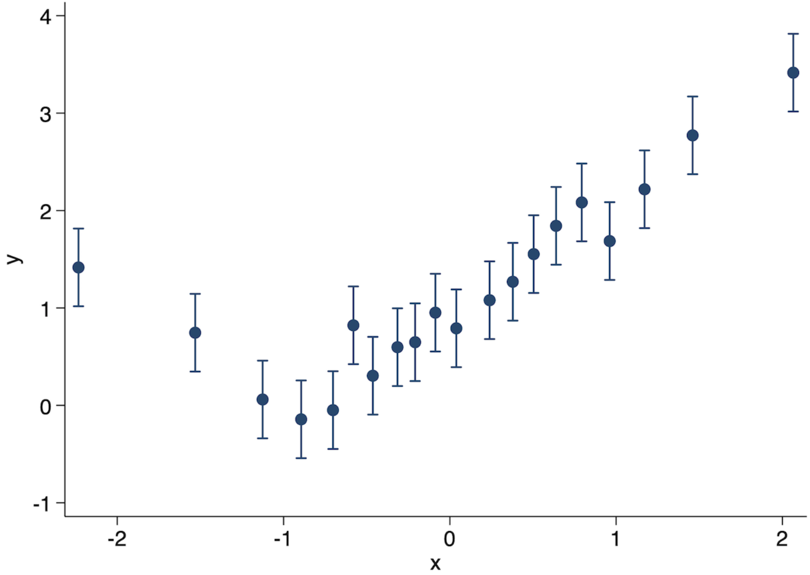
Big Data: Most Smoothers Are Slow (here: LOWESS)



Most Important Technique: `binscatter`



Cutting Edge: `binsreg`



Appendix: Semi-parametric Efficiency of OLS

Efficiency of Cell Means Estimation

- How efficient is OLS?
- Recall BLUE (Gauss-Markov Theorem):
  - OLS is most efficient (i.e. lowest variance)...
  - ... among all linear unbiased estimators...
  - ... assuming $\mathbb{E}\left[\varepsilon_i | \mathbf{X}_i\right] = 0$ and $\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}_i\right] = \sigma^2 \boldsymbol{I}$
- But what about general, nonlinear estimators?
  - MLE reaches Cramér-Rao Lower Bound (minimal variance)
  - Can OLS compete?

- Side note: recent work shows OLS is actually BUE... (Hansen 2022, ECMA)

Semi-Parametric Efficiency of OLS

- It turns out the answer is yes (Chamberlain 1987)
- OLS is semi-parametrically efficient
  - We do not need errors to be homoskedastic
  - Using cell-means logic can show OLS = MLE
  - So OLS reaches Cramér-Rao Lower Bound as well
- Suppose i.i.d. random sample $\mathbf{Z}_i = (Y_i, \mathbf{X}_i')'$
- Because it is a sample, Y_i and $\mathbf{X}_i$ are discrete
- Take on values $z_j = \left(y_j, \mathbf{x}_j'\right)'$ for $j = 1, ..., J$ with

$$\mathbb{E}\left[1\left(\mathbf{Z}_i = z_j\right)\right] = \Pr\left(\mathbf{Z}_i = z_j\right) = \pi_j$$

Population OLS of Cell Means

- Population OLS:

$$
\begin{aligned}
\beta_{\text{OLS}} &= \mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i'\right]^{-1}\mathbb{E}\left[\mathbf{X}_i Y_i\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{J} 1\left[\mathbf{Z}_i = z_j\right]\mathbf{x}_j\mathbf{x}_j'\right]^{-1}\mathbb{E}\left[\sum_{j=1}^{J} 1\left[\mathbf{Z}_i = z_j\right]\mathbf{x}_j y_j\right] \\
&= \left[\sum_{j=1}^{J}\pi_j\mathbf{x}_j\mathbf{x}_j'\right]^{-1}\left[\sum_{j=1}^{J}\pi_j\mathbf{x}_j y_j\right]
\end{aligned}
$$

- Unknown parameters: $\pi = (\pi_1, ..., \pi_J)'$

Log Likelihood of Cell Means

- Fact: $\mathbf{Z}_i \sim \text{Multinomial}\left(\pi_1, \ldots, \pi_J\right)$
- Hence, log likelihood of data (dropping constant):

$$\log f\left(\mathbf{Z}_1, \ldots, \mathbf{Z}_N, \pi\right) = \sum_{i=1}^{N} \sum_{j=1}^{J} 1\left[\mathbf{Z}_i = z_j\right] \log \pi_j$$

- Maximize this subject to $\pi_j \geq 0$ and $\sum_j \pi_j = 1$ yields

$$\widehat{\pi}_{\text{MLE}} = \begin{bmatrix} \frac{1}{N} \sum_{i=1}^{N} 1\left[\mathbf{Z}_i = z_1\right] \\ \vdots \\ \frac{1}{N} \sum_{i=1}^{N} 1\left[\mathbf{Z}_i = z_J\right] \end{bmatrix}$$

Cell Means OLS is MLE

- *Invariance Property of MLE*: For any $\mu = f(\theta)$, the MLE is

$$\widehat{\mu}_{\text{MLE}} = f\left(\widehat{\theta}_{\text{MLE}}\right)$$

- Plugging MLE into population OLS:

$$\begin{aligned}
\widehat{\beta}_{\text{MLE}} &= \left[\sum_j \widehat{\pi}_j \mathbf{x}_j \mathbf{x}_j'\right]^{-1} \left[\sum_j \widehat{\pi}_j \mathbf{x}_j y_j\right] \\
&= \left[\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{N} 1\left[\mathbf{Z}_i = z_j\right] \mathbf{x}_j \mathbf{x}_j'\right]^{-1} \left[\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{N} 1\left[\mathbf{Z}_i = z_j\right] \mathbf{x}_j y_j\right] \\
&= \left[\frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i \mathbf{X}_i'\right]^{-1} \left[\frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i Y_i\right]
\end{aligned}$$

- Hence, OLS is MLE! MLE reaches CRLB, and so does OLS