# Intro to non-parametric methods

# Introduction

Non-parametric estimation aims to estimate an unknown quantity while making as few assumptions as possible (about the data generating process)

We will look at the following

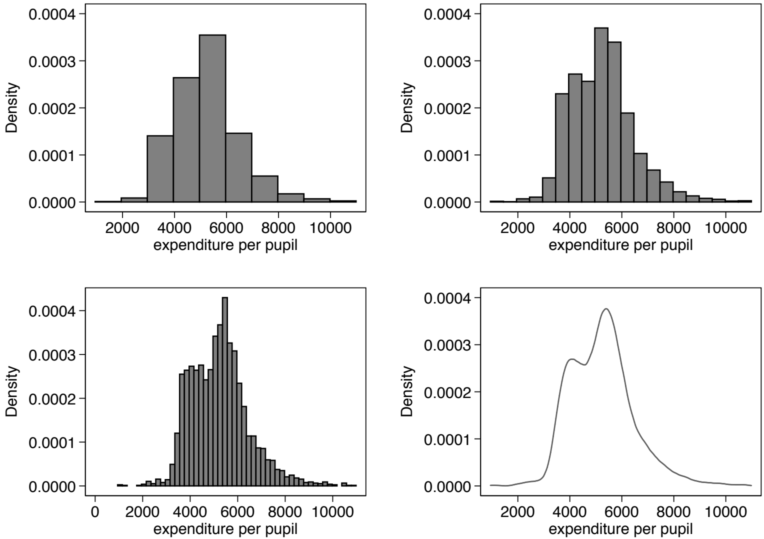
- density estimation
- regression
  - local constant
  - local linear

Non-parametric methods are local averaging methods

Key concern: how to define "local"

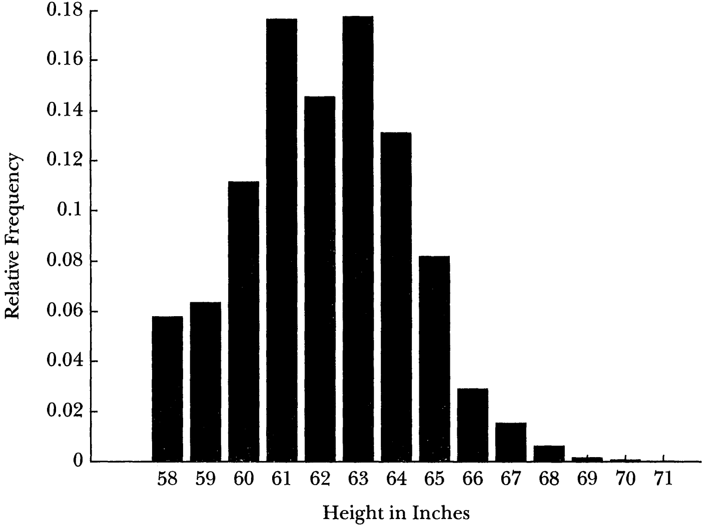
# Non-parametric density estimation

## Example



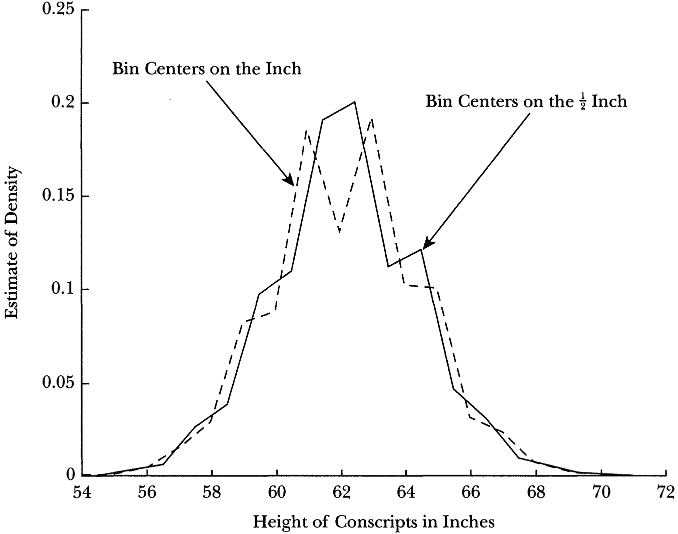
# Non-parametric density estimation

## Histogram



# Non-parametric density estimation

## Histogram



# Non-parametric density estimation

## Histogram

Remember $f(x) = dF(x)/dx$ so

$$
\begin{aligned}
f(x) &= \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h} \\
&= \lim_{h \to 0} \frac{\Pr(x - h < X < x + h)}{2h}
\end{aligned}
$$

the sample analog of which is

$$
\begin{aligned}
\hat{f}_{Hist}(x) &= \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbf{1}[x - h < X_i < x + h]}{2h} \\
&= \frac{1}{Nh} \sum_{i=1}^{N} \frac{1}{2} \times \mathbf{1}\left[ \left| \frac{X_i - x}{h} \right| < 1 \right]
\end{aligned}
$$

# Non-parametric density estimation

## Kernel density estimator

The histogram estimator can be generalized

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

where

- the weighting function $K(\cdot)$ is called a *kernel* function
- $h$ is a smoothing parameter called the *bandwidth*.

For $\hat{f} \to f$ we require that $Nh \to \infty$ and $h \to 0$.

# Non-parametric density estimation
## Kernel density estimator

It is usually assumed that $K(\cdot)$

- is symmetric $K(z) = K(-z)$
- integrates to 1
$$\int K(z)dz = 1$$
has zero mean
$$\int zK(z)dz = 0$$
and a finite second moment
$$\int z^2 K(z)dz = \kappa_2 < \infty$$

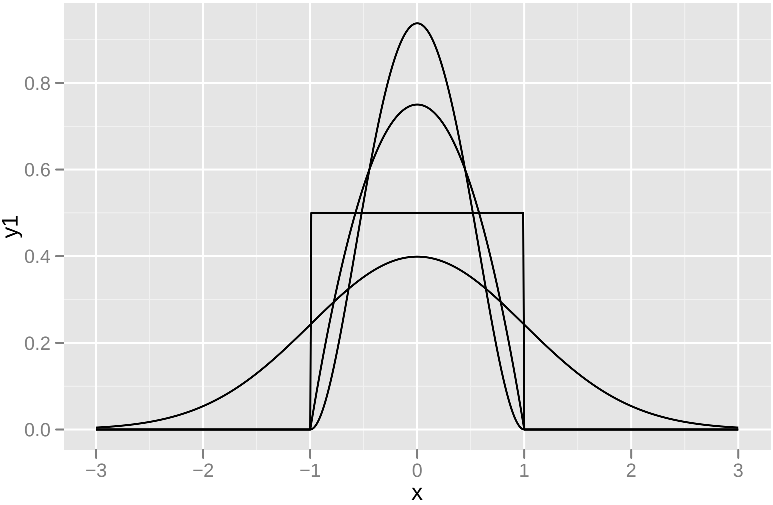
# Non-parametric density estimation

## Kernel density estimator – Some common kernels

| Kernel | $K(z)$ | $\delta$ |
|---|---|---|
| Uniform | $\frac{1}{2} \times 1[\lvert z \rvert < 1]$ | 1.3510 |
| Triangular | $(1 - \lvert z \rvert) \times 1[\lvert z \rvert < 1]$ | – |
| Epanechnikov | $\frac{3}{4}(1 - z^2) \times 1[\lvert z \rvert < 1]$ | 1.7188 |
| Biweight | $\frac{15}{16}(1 - z^2)^2 \times 1[\lvert z \rvert < 1]$ | 2.0362 |
| Gaussian | $(2\pi)^{-1/2} \exp(-z^2/2)$ | 0.7764 |

# Non-parametric density estimation

## Kernel density estimator – Some common kernels



# Non-parametric density estimation

## Kernel density estimator

We can show that

$$\int \hat{f}(x)dx = 1$$

$$\int x\hat{f}(x)dx = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$Var(\hat{f}(x)) = \hat{\sigma}^2 + h^2\kappa_2$$

where $\hat{\sigma}^2$ is the sample variance.

Note: these are <u>numerical</u> moments, not <u>sampling</u> moments

# Non-parametric density estimation

## Kernel density estimator – Estimation bias

Expectations of kernel transformations

$$E[\frac{1}{h}K(\frac{X_i - x}{h})] = \int \frac{1}{h}K(\frac{z-x}{h})f(z)dz = \int K(u)f(x+hu)du$$

where $u = (z-x)/h$, so

$$\begin{aligned} E[\hat{f}(x)] &= E[\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{X_i - x}{h}\right)] \\ &= \frac{1}{N}\sum_{i=1}^{N} E[\frac{1}{h}K\left(\frac{X_i - x}{h}\right)] = \int K(u)f(x+hu)du \end{aligned}$$

which (typically) cannot be solved analytically

# Non-parametric density estimation

## Kernel density estimator – Estimation bias

Assume a 2nd order kernel: $\kappa_j = 0$ for $j < 2$

Substituting a 2nd order Taylor expansion

$$f(x + hu) \approx f(x) + f'(x)hu + \frac{1}{2}f''(x)h^2u^2$$

in $\int K(u)f(x + hu)du$ gives

$$E[\hat{f}(x)] = \int K(u)f(x + hu)du \approx f(x) + \frac{1}{2}f''(x)h^2\kappa_2$$

and the bias equals

$$Bias(\hat{f}(h)) = E[\hat{f}(h)] - f(x) \approx \frac{1}{2}f''(x)h^2\kappa_2$$

(higher order kernels have lower order bias)

For the variance we get

$$Var(\hat{f}(x)) = Var(\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{X_i - x}{h}\right)) = \frac{1}{Nh^2} Var(K\left(\frac{X_i - x}{h}\right))$$
$$= \frac{1}{Nh^2} E[K\left(\frac{X_i - x}{h}\right)^2] - \frac{1}{N}(\frac{1}{h} E[K\left(\frac{X_i - x}{h}\right)])^2$$

but since

$$\frac{1}{h} E[K\left(\frac{X_i - x}{h}\right)] \approx f(x)$$

the 2nd term above disappears as $N \to \infty$

# Non-parametric density estimation

## Kernel density estimator – Estimation bias

Now

$$\begin{aligned}
\frac{1}{h}E\left[K\left(\frac{X_i - x}{h}\right)^2\right] &= \frac{1}{h}\int K\left(\frac{z-x}{h}\right)^2 f(z)dz \\
&= \int K(u)^2 f(x+hu)du \\
&\approx \int K(u)^2 (f(x) + f'(x)hu + \frac{1}{2}f''(x)h^2u^2)du \\
&\approx f(x)R(K)
\end{aligned}$$

where $R(K) \equiv \int K(u)^2 du$ is the “roughness” of the kernel

So that

$$Var(\hat{f}(x)) \approx \frac{1}{Nh}f(x)R(K)$$

# Non-parametric density estimation
## Bandwidth

To get rid of bias: bandwidth should decrease as sample size is increasing

- in the limit (infinite sample size) the bandwidth should be zero (we know the density at each point)

To get rid of variance: bandwidth should decrease at a slower rate than the sample size is increasing

- the number of observations within the bandwidth increases with sample size (variance of our estimate goes to zero)

because we reduce the bandwidth we have slower than $\sqrt{N}$ convergence

# Non-parametric density estimation

## Mean Squared Error (MSE)

At a point $x$

$$MSE(x) \equiv E[\Big(f(x) - \hat{f}(x)\Big)^2] = Bias^2 + Var(\hat{f}(x))$$
$$\approx \left(\frac{1}{2} f''(x) h^2 \kappa_2\right)^2 + \frac{1}{Nh} f(x) R(K) \equiv AMSE(x)$$

Smoothing involves a trade-off between bias and variance:

- when the data are over-smoothed, the bias is large and the variance low
- when the data are under-smoothed, the bias is low and the variance high

optimal smoothing minimizes the risk (MSE)

# Non-parametric density estimation
## Bandwidth

A key question is how to choose $h$?

With the histograms above we saw that as $h$ increased, the density

- became less “jumpy”
- but did a poorer job fitting the data

This highlights the trade-off between variance and bias

Methods for optimal bandwidth try to balance this using well defined criteria such as the integrated square error

# Non-parametric density estimation

## Mean Squared Error (MSE)

A global measure of fit is the asymptotic mean integrated square error:

$$\begin{aligned}AMISE = \int AMSE(x)dx &= \int \left( \left( \frac{1}{2} f''(x) h^2 \kappa_2 \right)^2 + \frac{1}{Nh} f(x) R(K) \right) dx \\ &= \frac{1}{4} R(f'') h^4 \kappa_2^2 + \frac{1}{Nh} R(K)\end{aligned}$$

and the bandwidth that minimizes it:

$$h_0 = R(f'')^{-1/5} (R(K)/\kappa_2^2)^{1/5} N^{-1/5}$$

Using this bandwidth we get

$$AMISE_0 = \frac{5}{4} (\kappa_2^2 R(K) R(f''))^{1/5} N^{-4/5}$$

(which converges at a slower than parametric rate $N^{-1}$)

# Non-parametric density estimation

## Bandwidth – Silverman's optimal bandwidth

If both the data ($f$) and the kernel are normal, then Silverman (1986) suggested the following bandwidth

$$h_{opt} \approx 1.06\sigma N^{-1/5}$$

it can also be adjusted by a factor $\delta$ (see table above) for different kernels

$$h_{opt} \approx 1.3643\delta\sigma N^{-1/5}$$

or

$$h_{opt} \approx 1.3643\delta N^{-1/5} \min(\sigma, IQR/1.349)$$

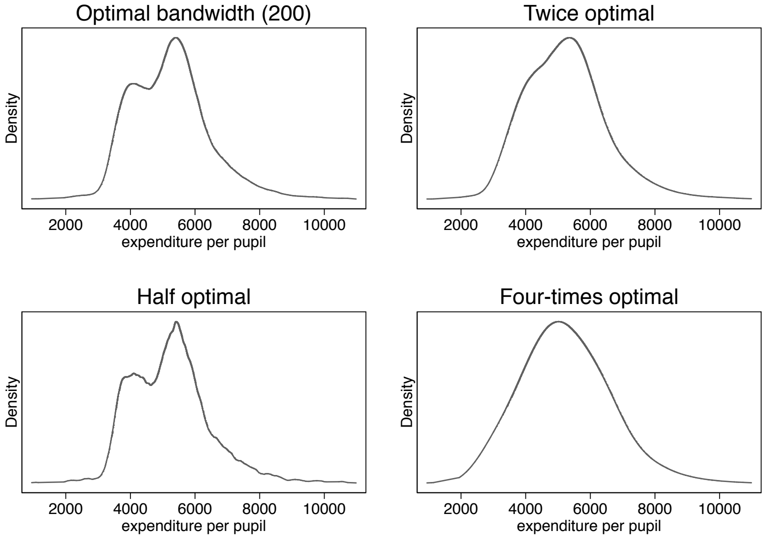
which is more robust against outliers

In practice this method works quite well, but

- there are other approaches such as cross validation, and methods that let the bandwidth vary
- do not forget to use your eyes: does it look reasonable?

# Non-parametric density estimation

## Expenditure per pupil



# Non-parametric density estimation

## Higher dimensions

The above method generalizes to higher dimensional cases, for example 2:

$$\hat{f}_{Hist}(x_1, x_2) = \frac{1}{Nh^2} \sum_{i=1}^{N} \frac{1}{4} \times \mathbf{1}\left[\left|\frac{X_{1i} - x}{h}\right| < 1\right] \times \mathbf{1}\left[\left|\frac{X_{2i} - x}{h}\right| < 1\right]$$

this generates a number of issues:

- same bandwidth in all dimensions?
- take correlation between $x_1, x_2$ into account (ellipsis)?

One solution is to transform the data data (equal variance and orthogonal) before the calculations, estimate the density and transform back

Suppose we have $n$ uniformly distributed data points on $[-1, 1]$

- How many points in $[-0.1, 0.1]$ ?

Suppose we have $n$ uniformly distributed data points on $[-1, 1]^k$

- How many points in $[-0.1, 0.1]^k$ ?

# Non-parametric regression

The standard tool to estimate the relationship between an outcome $y$ and an explanatory variable $x$ is linear regression

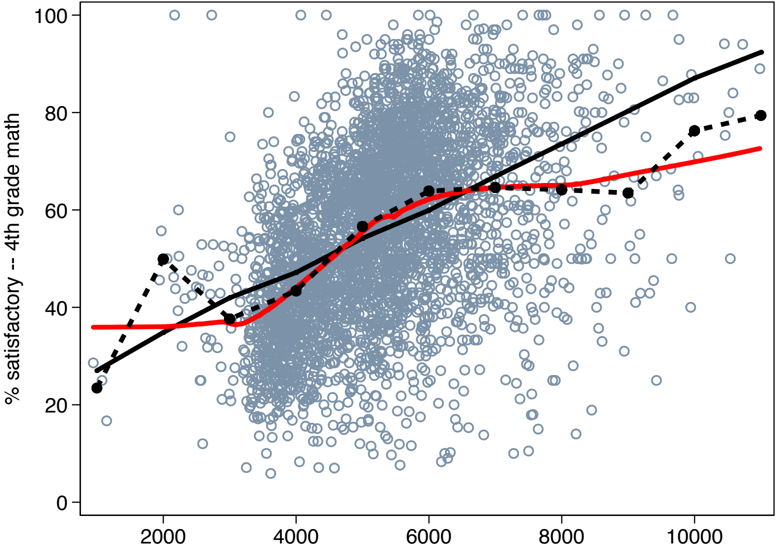
$$y_i = x_i\beta + \epsilon_i$$

this does not need to be linear in $x_i$:

- polynomials
- splines
- non-parametric regression (the Nadaraya-Watson estimator, or local constant estimator)

# Non-parametric regression

## Example



# Non-parametric regression

## Kernel regression

The same methods for nonparametric density estimation can be used to estimate a regression function

$$E[Y|X=x] = \int y f(y|X=x) dy$$

since

$$\hat{E}[Y|X=x] = \int y \hat{f}_{Y|X}(y|x) dy = \int y \frac{\hat{f}_{YX}(y,x)}{\hat{f}_X(x)} dy$$

We know how to estimate

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

so what is left is

$$\int y \hat{f}_{YX}(y,x) dy$$

# Non-parametric regression

## Kernel regression

Take the following bivariate kernel $K(u,\, v) = K_1(u)K_2(v)$ then

$$\hat{f}_{YX}(y,x) = \frac{1}{Nh^2}\sum_{i=1}^{N} K\left(\frac{X_i - x}{h}, \frac{Y_i - y}{h}\right)$$
$$= \frac{1}{Nh^2}\sum_{i=1}^{N} K_1\left(\frac{X_i - x}{h}\right) K_2\left(\frac{Y_i - y}{h}\right)$$

so that

$$\int y\hat{f}_{YX}(y,x)dy = \frac{1}{Nh^2}\int y\sum_{i=1}^{N} K_1\left(\frac{X_i - x}{h}\right) K_2\left(\frac{Y_i - y}{h}\right) dy$$
$$= \frac{1}{Nh}\sum_{i=1}^{N} K_1\left(\frac{X_i - x}{h}\right) \int y\frac{1}{h}K_2\left(\frac{Y_i - y}{h}\right) dy$$
$$= \frac{1}{Nh}\sum_{i=1}^{N} K_1\left(\frac{X_i - x}{h}\right) Y_i$$

We can now write

$$\hat{g}(x) = \frac{\int y \hat{f}_{YX}(y,x)dy}{\hat{f}_X(x)} = \frac{\frac{1}{Nh}\sum_{i=1}^{N} Y_i K_1\left(\frac{X_i - x}{h}\right)}{\frac{1}{Nh}\sum_{i=1}^{N} K_1\left(\frac{X_i - x}{h}\right)} = \sum_{i=1}^{N} \omega_h(X_i,\, x) Y_i$$

when $K(x) = \frac{1}{2} \cdot 1[x - h < X_i < x + h]$ then $\hat{g}(x)$ is the average $Y$ for observations within a window $h$ of $x$

This is the Nadaraya-Watson kernel regression estimator

Bandwidth is sometimes chosen using cross validation

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

where $\hat{g}_{h,(-i)}$ is the regression estimate leaving out observation $i$

# Non-parametric regression

## Local constant regression

The standard Nadaraya-Watson kernel regression estimator can also be seen as fitting a constant function

$$g(x) = \alpha$$

where

$$\hat{\alpha} = \arg\min_a \sum_{i=1}^{N} K\left(\frac{X_i - x}{h}\right)(Y_i - a)^2$$

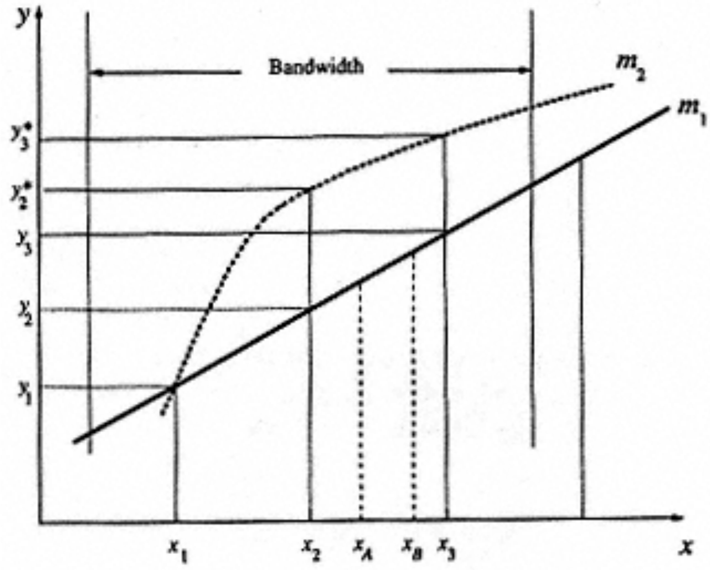
But kernel regression does not perform well at the boundaries

If the regression function is flat there is no problem, but otherwise we

- over (under) estimate if the regression function is convex (concave)
- over (under) estimate the regression function at the left (right) boundary
- get bias if density has no zero derivative (data not equally spaced)

# Non-parametric regression

## Bias of kernel regression



# Non-parametric regression

## Local linear regression

Local linear regression fits

$$g(x) = \alpha + \beta(z - x)$$

so

$$(\hat{\alpha},\, \hat{\beta}) = \arg\min_{a,\, b} \sum_{i=1}^{N} K\left(\frac{X_i - x}{h}\right)(Y_i - a - b(X_i - x))^2$$

and

$$\hat{g}(x) = \hat{\alpha} + \hat{\beta}(x - x) = \hat{\alpha}$$

# Non-parametric regression

## Partial linear regression (Robinson, Econometrica 1988)

We can write

$$y_i = X_i\beta + g(Z_i) + e_i$$

so that

$$E[y_i|Z_i] = E[X_i|Z_i]\beta + g(Z_i)$$

and taking the difference

$$\underbrace{y_i - E[y_i|Z_i]}_{e_{yi}} = (\underbrace{X_i - E[X_i|Z_i]}_{e_{xi}})\beta + e_i$$

we can now estimate $E[y_i|Z_i]$ and $E[X_i|Z_i]$ using kernel regression
estimate $\beta$ from a regression of $\hat{e}_{yi}$ on $\hat{e}_{xi}$
and then estimate $g(z)$ using a kernel regression of $(y_i - X_i\hat{\beta})$ on $Z_i$

# Non-parametric regression

## Partial linear regression

Alternatively

1. Sort the data by $z$
2. First difference $y$ and $X$ and estimate $\beta$ using OLS on the first differenced data
3. Calculate $\hat{e} = y - X\hat{\beta}$
4. Estimate $g(z)$ using a LLR of $\hat{e}$ on $z$

See Yatchew (JEL 36(2), 1998) for more details and more efficient estimators

# Non-parametric regression

## Confidence intervals

Statistical packages often implement asymptotic CIs, or use the bootstrap

```
pctile _x = expp, nquant(100)
gen _w = .

set seed 32423
lpoly math4 expp, degree(1) at(_x) gen(b0)


forv r=1/199 {
   di . _c
   if (mod(`r', 50) == 0) di " `= 50 * int(`r' / 50)'"
   bsample, weight(_w)
   lpoly math4 expp [fw=_w], degree(1) at(_x) gen(_b`r') nograph
}

egen ci_lower = rowpctile(_b*), p(5)
egen ci_upper = rowpctile(_b*), p(95)

sort _x
twoway (rarea ci* _x if _x < .) (line b0 _x if _x < .)
```