

450A Political Methodology I

Lecture 3: What is Regression?

Stanford University

What is Regression Analysis?

Nonparametric Regression

Linear Regression

Combining Linear Regression with Nonparametric Regression

Conditioning on X

- Regression quantifies how an outcome variable Y varies as a function of one or more predictor variables X
- The basic idea: **conditioning** on X
- Goal is to characterize $f(Y|X)$, the conditional probability distribution of Y for different levels of X
- Instead of modeling the whole conditional density of Y given X , in regression we usually only model the **conditional mean** of Y given X : $E[Y|X = x]$
- Our key goal is to approximate the **conditional expectation function** $E[Y|X]$, which summarizes how the average of Y varies across all possible levels of X (also called the **population regression function**)
- Once we have estimated $E[Y|X]$, we can use it for **prediction** and/or **causal inference**, depending on what assumptions we are willing to make

Estimands, Estimators, and Estimates

- The conditional expectation function $E[Y|X]$ is the **estimand** (or **parameter**) we are interested in
- $\hat{E}[Y|X]$ is the **estimator** of this parameter of interest, which is a function of X
- For a given sample dataset, we obtain an **estimate** of $E[Y|X]$.
- Note: We will want to extend the regression idea to the case of multiple X variables, but we start with the simple bivariate case where we have a single X

What is Regression Analysis?

Nonparametric Regression

- Discrete X

- Continuous X

- Bias-Variance Tradeoff

Linear Regression

Combining Linear Regression with Nonparametric Regression

Nonparametric Regression with Discrete X

- Let's take a look at how people voted for president in 2024, using the 2024 Cooperative Election Study (CES), a large online survey
- We use two variables, for the 42,028 respondents who reported voting for Trump or Harris:
 - Y : vote (1 = Trump, 0 = Harris)
 - X : educational attainment
- Goal is to characterize the conditional expectation $E[Y|X]$, i.e. how the share voting Trump varies with education level

Source: Cooperative Election Study 2024 Common Content (Schaffner, Shih, Ansolabehere & Pope), doi:10.7910/DVN/X11EP6

educ: "What is the highest level of education you have completed?"

- 1. No high school
- 2. High school graduate
- 3. Some college
- 4. 2-year degree
- 5. 4-year degree
- 6. Postgraduate degree

Nonparametric Regression with Discrete X

trump: Respondent's vote for president, reported after the election:

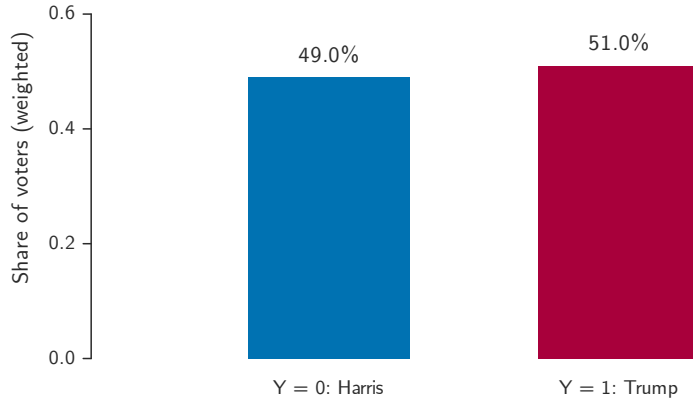
- 1. Donald Trump
- 0. Kamala Harris

Respondents who voted for another candidate or did not vote are left out. The vote is what respondents reported after the election.

Because Y is 0 or 1, its mean is a share:

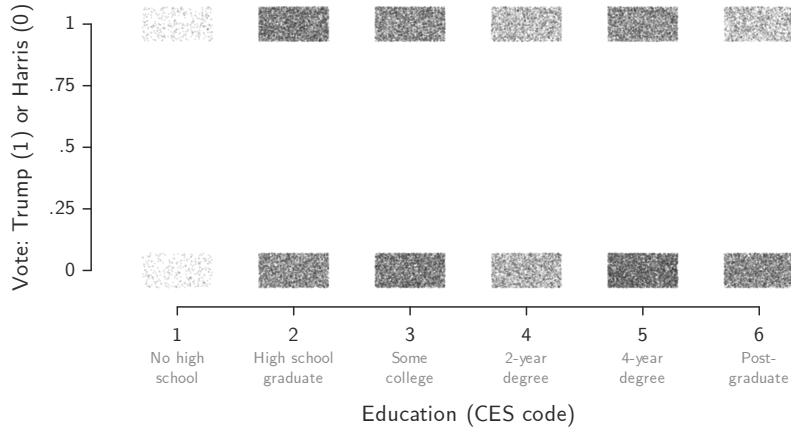
$$E[Y|X = x] = \Pr(Y = 1|X = x) = \text{the share voting Trump among respondents with } X = x$$

Marginal Distribution of Y



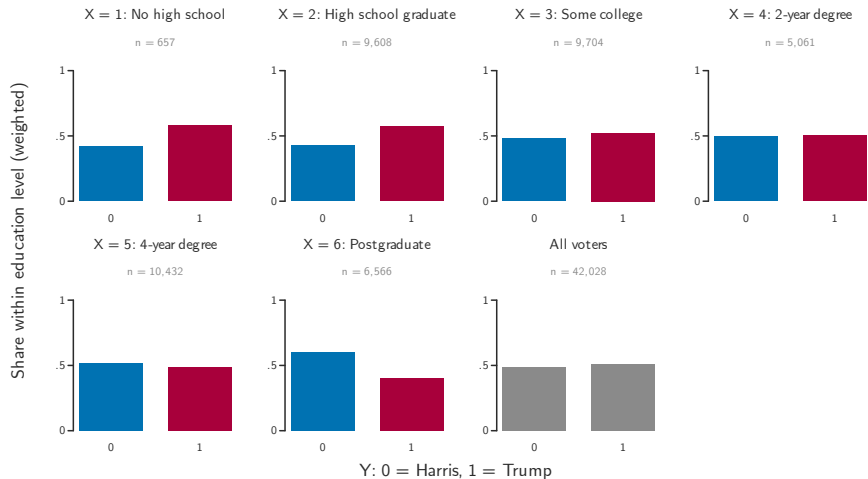
Source: Cooperative Election Study 2024: 42,028 respondents who reported voting for Trump or Harris; shares weighted by the survey weight

Vote and Education



Source: Cooperative Election Study 2024: 42,028 respondents who reported voting for Trump or Harris; shares weighted by the survey weight

Distribution of vote given education $p(y|x)$

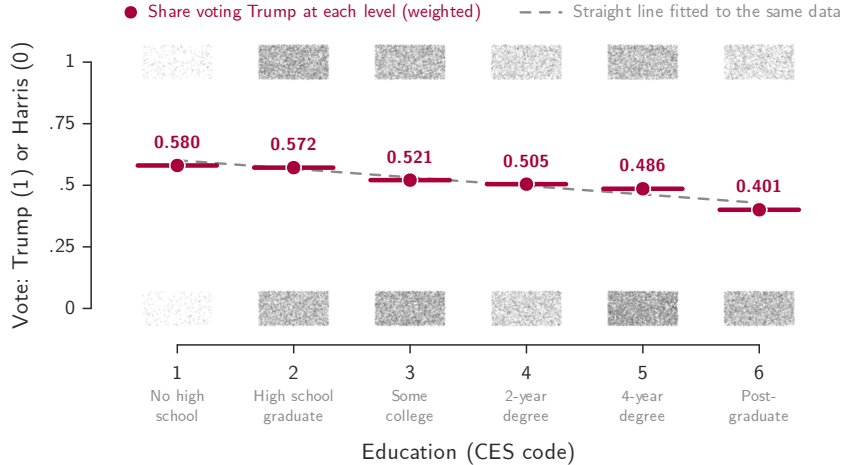


Source: Cooperative Election Study 2024: 42,028 respondents who reported voting for Trump or Harris; shares weighted by the survey weight

Nonparametric Regression with Discrete X

- With a 0/1 outcome, each panel comes down to one number, the share voting Trump
- That share is the mean of Y , so it is the summary we want: $E[Y|X]$
- Here our X variable education has a small number of levels (6) and many observations in each level (at least 657)
- In situations like this we can estimate $E[Y|X = x]$ as the sample mean of Y at each observed level x of X
- We use a weighted mean: each respondent counts in proportion to the CES survey weight, which corrects for groups the survey reached too often or too rarely

Nonparametric Regression with Discrete X



Source: Cooperative Election Study 2024: 42,028 respondents who reported voting for Trump or Harris; shares weighted by the survey weight

Nonparametric Regression

- The share voting Trump falls as education rises, but not along a straight line in the 1–6 code
 - Flat between the two lowest levels: 58.0% and 57.2%
 - A sharp drop at postgraduate degrees: 48.6% to 40.1%
 - “Some college” and “2-year degree” are not obviously ordered, another reason not to force a line through a code
 - No high school has the fewest respondents (657), so its share is the least precise
- Estimating $E[Y|X = x]$ level by level makes minimal assumptions
- This method does not impose any specific **functional form** on the relationship between Y and X (i.e. the shape of $E[Y|X]$)
- → It is called a **nonparametric regression**
- It works well as long as
 - X is discrete
 - there are a small number of values of X
 - a small number of X variables
 - a lot of observations at each X value
- But what do we do when X is continuous and has many values?

What is Regression Analysis?

Nonparametric Regression

- Discrete X

- Continuous X

- Bias-Variance Tradeoff

Linear Regression

Combining Linear Regression with Nonparametric Regression

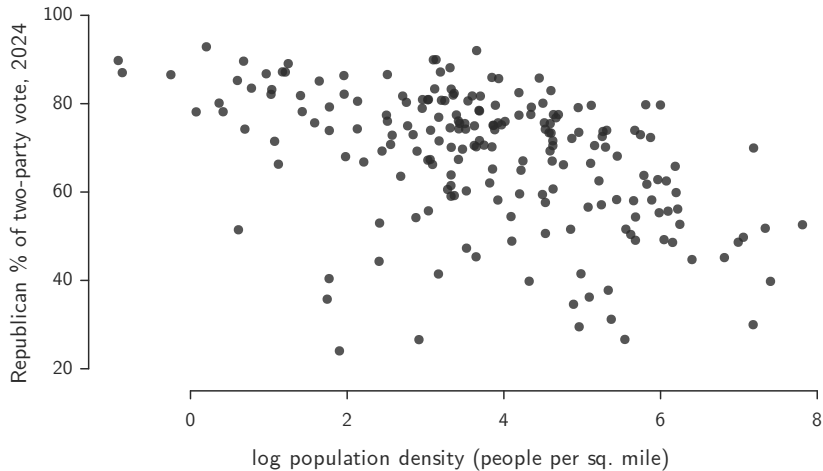
Nonparametric Regression with Continuous X

Consider the 2024 presidential election in U.S. counties:

- 3,034 counties, from under 0.1 people per square mile (rural Texas) to about 72,000 (Manhattan)
- We're interested in the relationship between:
 - Y : Republican (Trump) share of the two-party vote, in %
 - X : log population density, the log of people per square mile
($X = 0$: 1 person; $X = 4.6$: 100 people; $X = 6.9$: 1,000 people)
- For the next slides we use a random sample of 200 counties, so each point can be seen; we come back to all 3,034 at the end

Source: MIT Election Data and Science Lab; Census ACS 2019–23 and 2023 Gazetteer; $N = 3,034$ counties (Alaska, Connecticut, Louisiana and 7 other counties excluded)

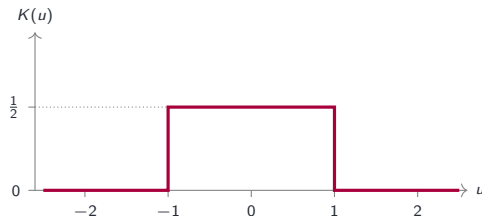
Nonparametric Regression with Continuous X



Source: A random sample of 200 counties (seed 450) from the 3,034 above. MIT Election Data and Science Lab; Census ACS 2019–23 and 2023 Gazetteer

Uniform Kernel Regression: Simple Local Averages

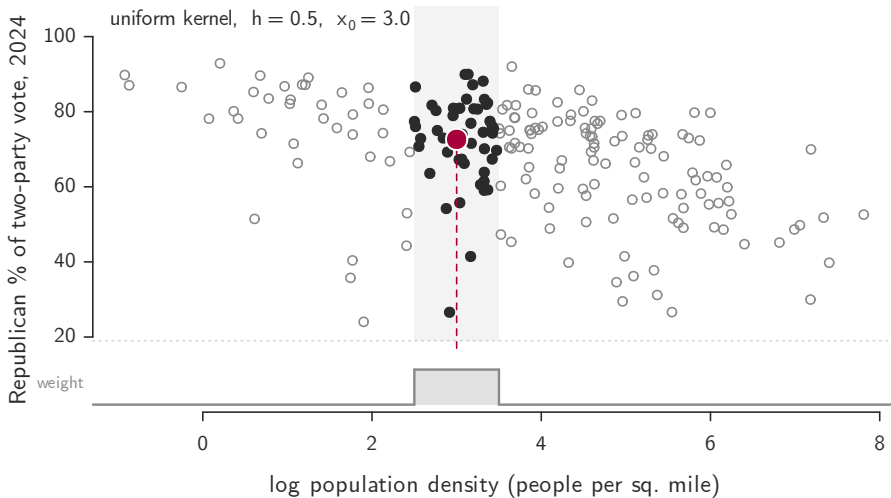
- One approach is to use a **moving local average** to estimate $E[Y|X]$
- Calculate the average of the observed y points that have x values in the interval $[x_0 - h, x_0 + h]$
- h = some positive number (called the **bandwidth**)
- **Uniform kernel**: every observation in the interval is equally weighted



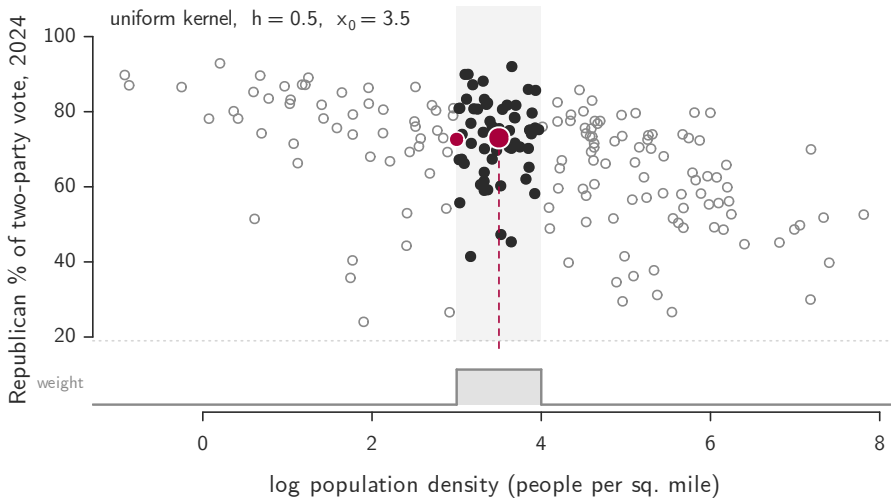
- This gives the **uniform kernel regression**:

$$\hat{E}[Y|X = x_0] = \frac{\sum_{i=1}^N K((X_i - x_0)/h) Y_i}{\sum_{i=1}^N K((X_i - x_0)/h)} \text{ where } K(u) = \frac{1}{2} \mathbf{1}_{\{|u| \leq 1\}}$$

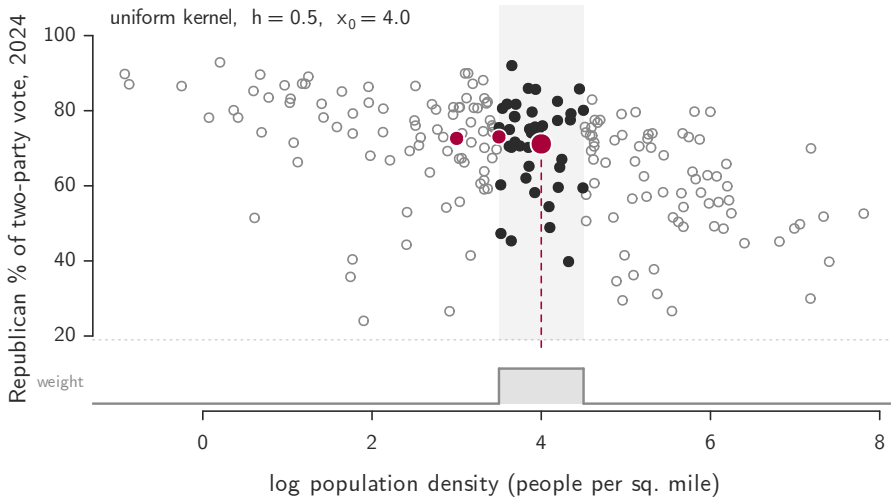
Kernel Regression: Simple Local Averages



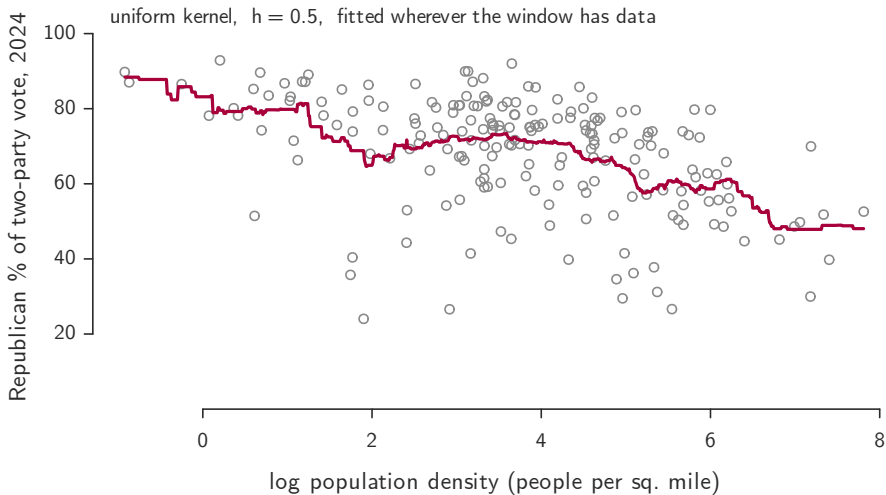
Kernel Regression: Simple Local Averages



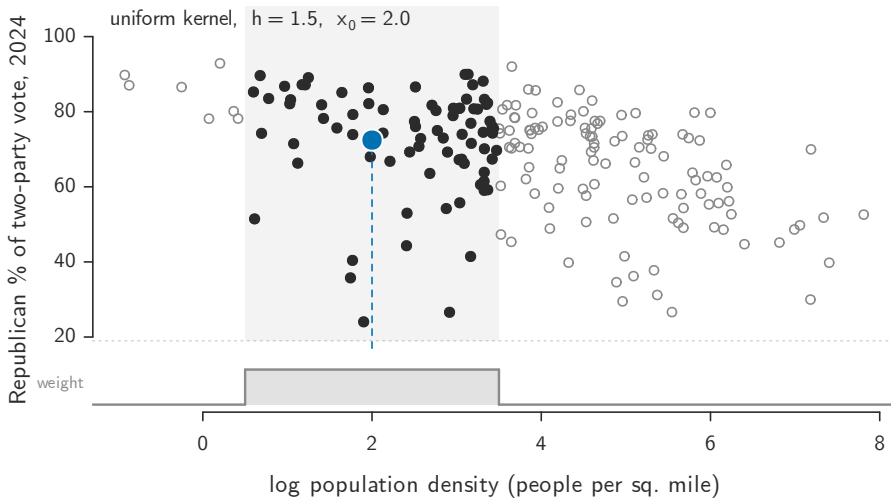
Kernel Regression: Simple Local Averages



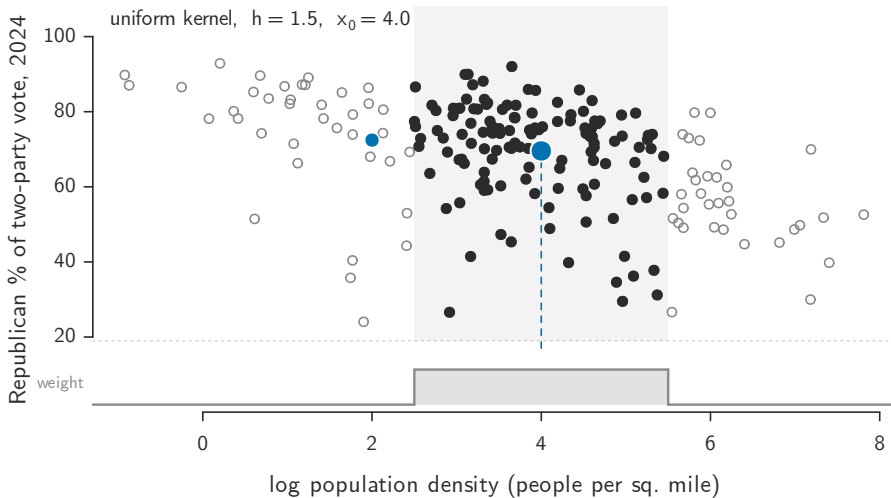
Kernel Regression: Simple Local Averages



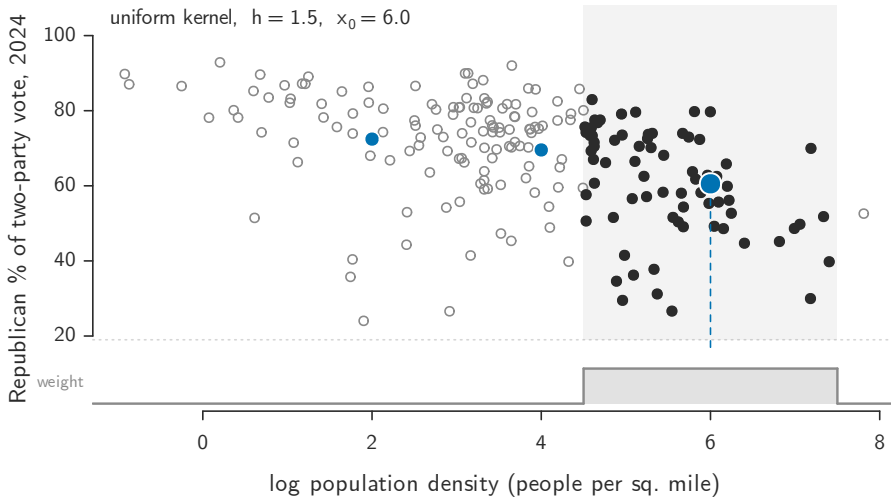
Changing the Bandwidth



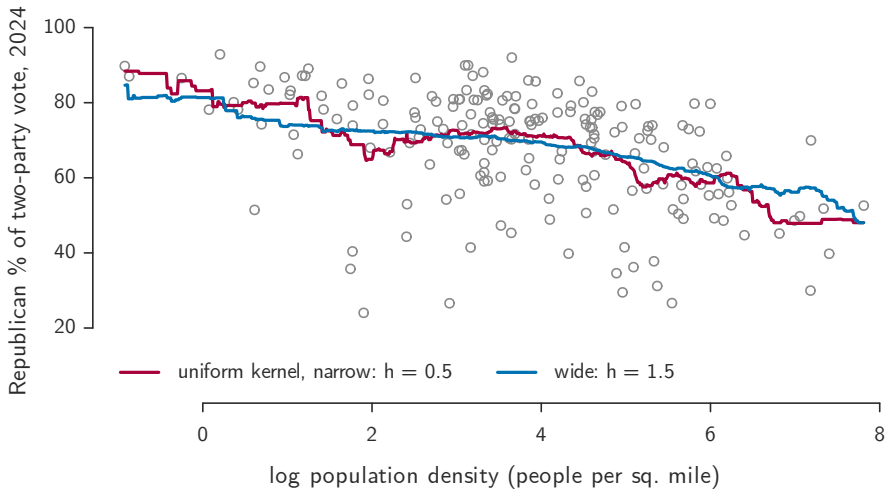
Changing the Bandwidth



Changing the Bandwidth



Narrow and Wide Bandwidth



Kernel Regression: Weighted Local Averages

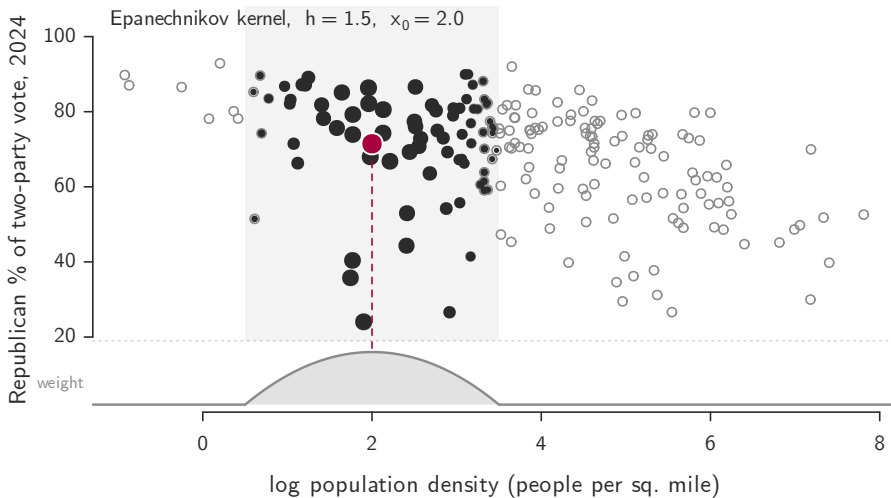
- Another approach is to construct **weighted local averages**
 - Data points that are closer to x_0 get more weight than points farther away
- ① decide on a symmetric, non-negative **kernel weight function** K that is zero outside $[-1, 1]$ and falls as $|u|$ grows (e.g. Epanechnikov)



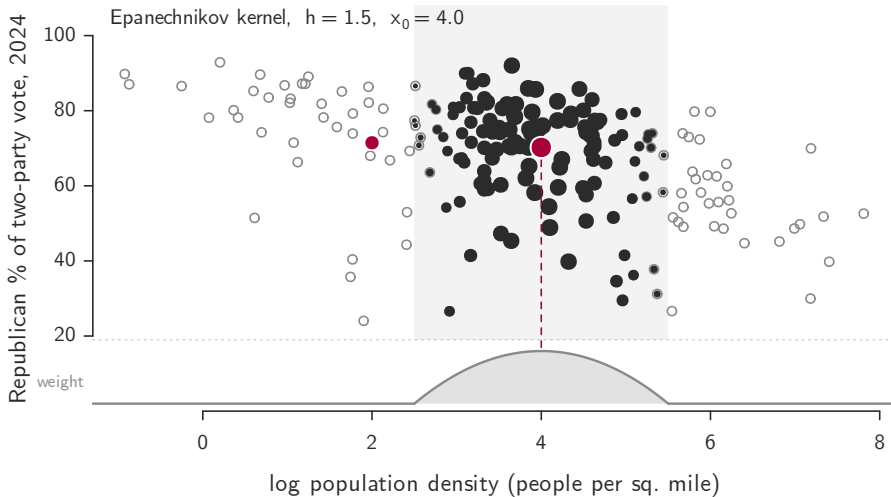
- ② compute weighted average of the observed y points that have x values in the bandwidth interval $[x_0 - h, x_0 + h]$ e.g.

$$\hat{E}[Y|X = x_0] = \frac{\sum_{i=1}^N K((X_i - x_0)/h) Y_i}{\sum_{i=1}^N K((X_i - x_0)/h)} \text{ where } K(u) = \frac{3}{4}(1 - u^2) \mathbf{1}_{\{|u| \leq 1\}}$$

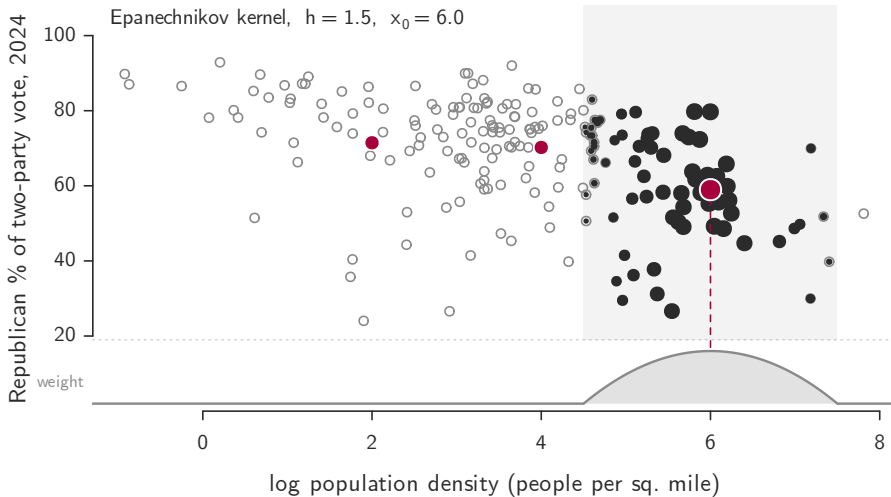
A Weight Function



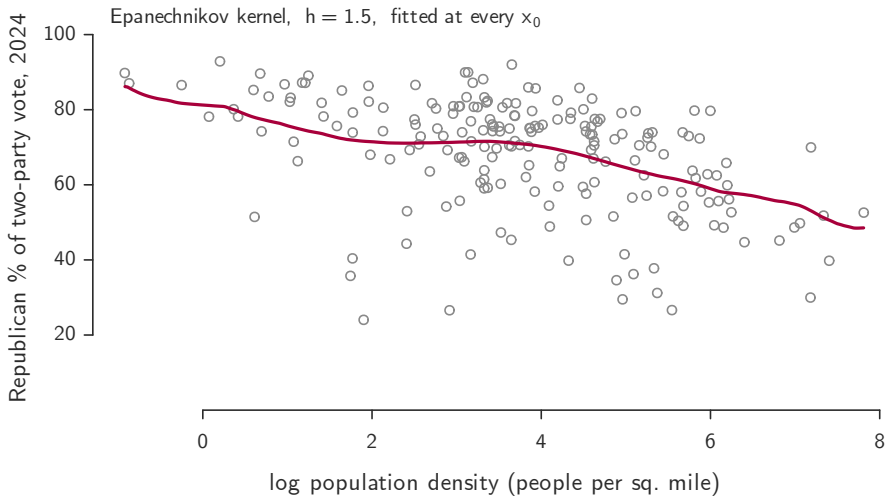
Weighted Local Averages



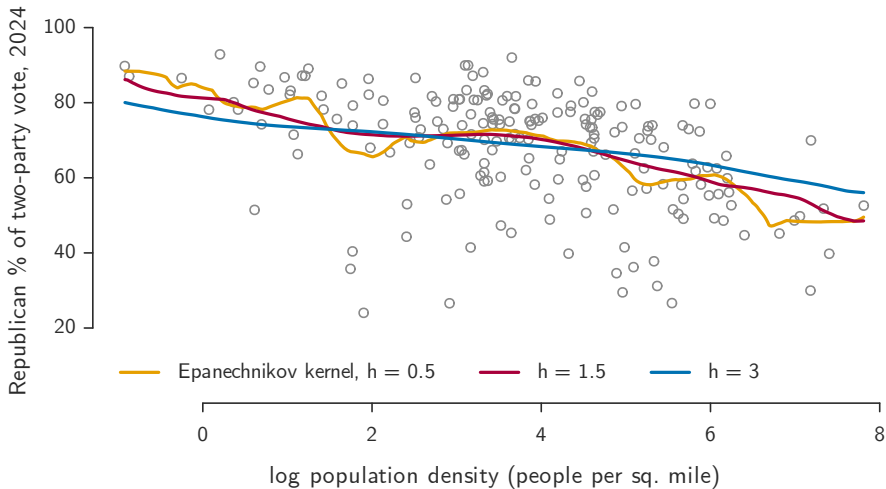
Weighted Local Averages



Kernel Regression: Weighted Local Averages



Changing the Bandwidth



What is Regression Analysis?

Nonparametric Regression

- Discrete X

- Continuous X

- Bias-Variance Tradeoff

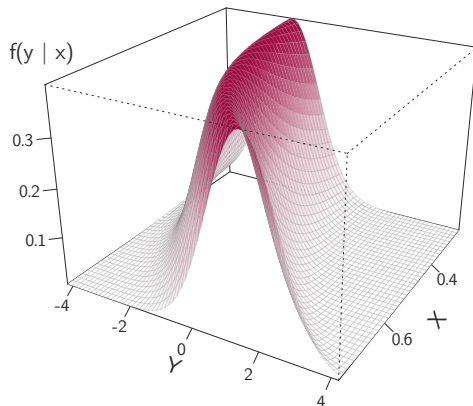
Linear Regression

Combining Linear Regression with Nonparametric Regression

- When choosing an estimator $\hat{E}[Y|X]$ for $E[Y|X]$, we face a **bias-variance tradeoff**
- Notice that we can choose 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)

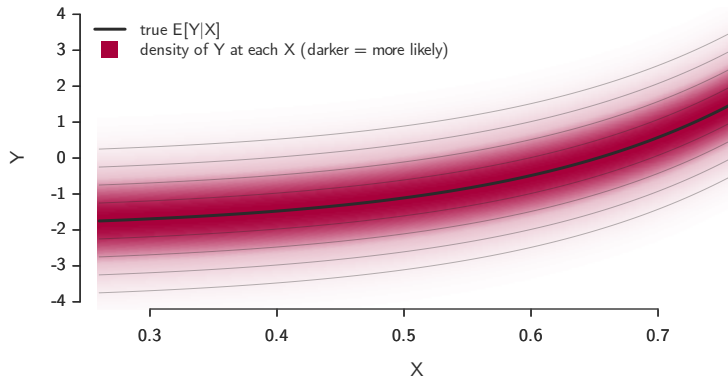
Hypothetical True Distribution

- Let's conduct a simulation experiment to actually see the tradeoff
- Suppose we have the following population distribution:



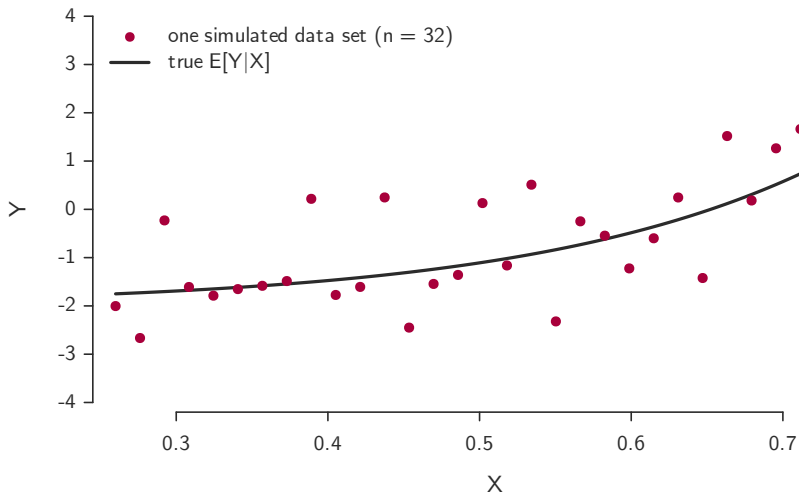
Hypothetical True Distribution

- Another way of representing the same population distribution:



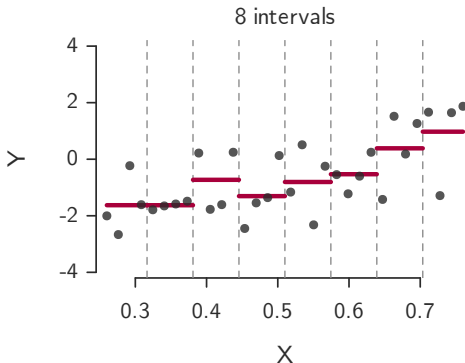
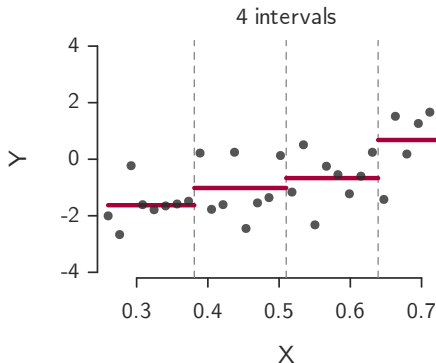
- From this distribution we draw 2,000 simulated data sets: X is held at 32 evenly spaced values from 0.26 to 0.76, and each Y is drawn from $f(y|x)$.

An Example of a Simulated Data Set

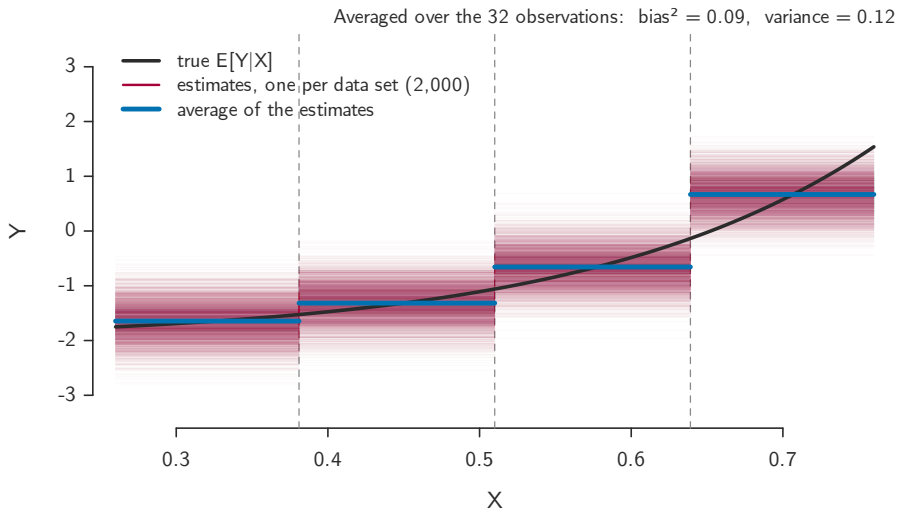


Two Estimators

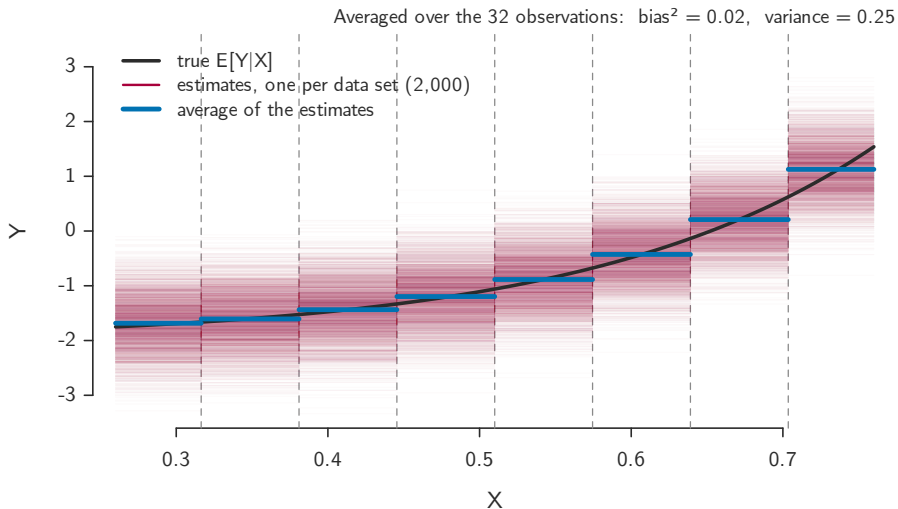
- For each simulated data set, we apply two simple estimators of $E[Y|X]$:
 - Divide X into **4** ranges and take the mean for each
 - Divide X into **8** ranges and take the mean for each
- We then evaluate how well these estimators do in terms of bias and variance.



Simulated Distribution of Estimates: 4 Intervals



Simulated Distribution of Estimates: 8 Intervals



Bias-Variance Tradeoff

- In this comparison, the less “flexible” estimator (4 ranges) has more bias
- and the more “flexible” estimator (8 ranges) has more variance
- So choosing the flexibility means balancing bias against variance
- With more data (or fewer variables) we can “afford” a more flexible estimator, and both can shrink

What is Regression Analysis?

Nonparametric Regression

Linear Regression

Combining Linear Regression with Nonparametric Regression

Parametric Approach: Linear Regression

- Linear regression works by assuming a linear parametric form for the conditional expectation function:

$$E[Y|X] = \beta_0 + X \beta_1$$

- Conditional expectation defined by only two **coefficients** which are estimated from the data:
 - β_0 is called the **intercept** or **constant**
 - β_1 is called the **slope coefficient**
- Notice that the linear functional form imposes a constant slope
- Assumption: Change in $E[Y|X]$ is the same at all values of X
- Geometrically, the linear regression function will look like:
 - A **line** in cases with a single X variable
 - A **plane** in cases with two X variables
 - A **hyperplane** in cases with more than two X variables

Parametric Approach: Linear Regression in R

R Code

```
> counties <- read.csv("county_pres.csv")
> counties$rep24 <- 100 * counties$rep2p_2024 # Republican % of two-party vote
> counties <- subset(counties, !is.na(rep24)) # counties with 2024 returns
> head(counties[, c("county", "state", "rep24", "pct_ba", "log_density")])
```

	county	state	rep24	pct_ba	log_density
1	Autauga County	AL	73.3589	28.2827	4.602467
2	Baldwin County	AL	79.3476	32.7976	5.016762
3	Barbour County	AL	57.4150	11.4647	3.331267
4	Bibb County	AL	82.3849	11.4682	3.571988
5	Blount County	AL	90.7769	15.5790	4.521148
6	Bullock County	AL	26.9589	8.9997	2.791669

pct_ba: % of adults 25 and older with a BA or higher. We now regress the Republican share on pct_ba.

Source: MIT Election Data and Science Lab; Census ACS 2019–23; N = 3,034 counties (Alaska, Connecticut, Louisiana and 7 other counties excluded)

Parametric Approach: Linear Regression in R

R Code

```
> lm.out <- lm(rep24 ~ pct_ba, data = counties)
> summary(lm.out)

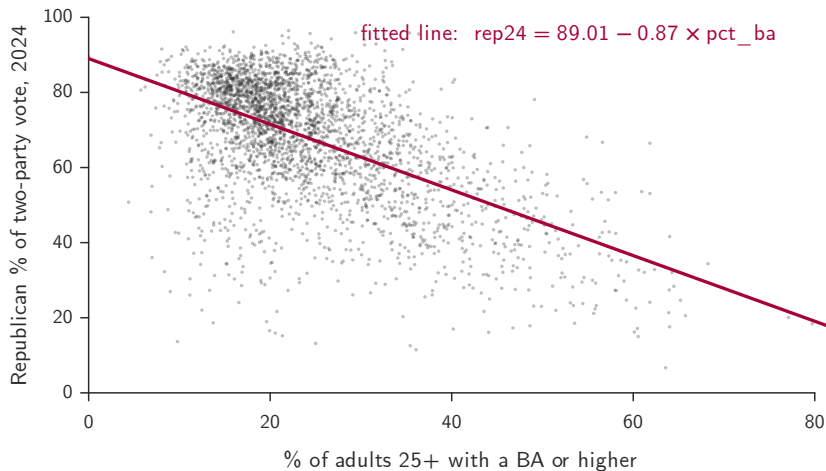
Call:
lm(formula = rep24 ~ pct_ba, data = counties)

Residuals:
    Min       1Q   Median       3Q      Max
-66.800  -6.879   1.956   8.586  40.102

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  89.00934    0.61309   145.18  <2e-16 ***
pct_ba       -0.87431    0.02342   -37.33  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.12 on 3032 degrees of freedom
Multiple R-squared:  0.3149,    Adjusted R-squared:  0.3147
F-statistic: 1393 on 1 and 3032 DF,  p-value: < 2.2e-16
```

Parametric Approach: Linear Regression



Source: MIT Election Data and Science Lab; Census ACS 2019–23; N = 3,034 counties (Alaska, Connecticut, Louisiana and 7 other counties excluded)

Linear Regression: Interpretation

Suppose we have computed coefficient estimates for our regression:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$\widehat{\text{rep24}} = 89.0 - 0.87 \text{ pct_ba}$$

Q: What does $\hat{\beta}_0$ tell us?

Substituting 0 in for x we see that $E[Y|X = 0] = \beta_0$, so $\hat{\beta}_0$ estimates the average of Y where $X = 0$.

Our estimate of the average Republican share is 89.0% for a county where no adult has a BA. Only one county is at 0%: Loving County, TX, population 54. The next lowest is at 4.4%. So this is an extrapolation beyond nearly all of the data.

Suppose we have computed coefficient estimates for our regression:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$\widehat{\text{rep24}} = 89.0 - 0.87 \text{ pct_ba}$$

Q: What does $\hat{\beta}_1$ tell us?

Sign: The sign of the slope estimate $\hat{\beta}_1$ tells us if the fitted $\hat{E}[Y|X]$ increases or decreases with a change in X

$\hat{\beta}_1 > 0$: $\hat{E}[Y|X]$ increases with X

$\hat{\beta}_1 < 0$: $\hat{E}[Y|X]$ decreases with X

$\hat{\beta}_1 = 0$: $\hat{E}[Y|X]$ is flat in X

Suppose we have computed coefficient estimates for our regression:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$\widehat{\text{rep24}} = 89.0 - 0.87 \text{ pct_ba}$$

Q: What does $\hat{\beta}_1$ tell us?

Magnitude: The slope estimate $\hat{\beta}_1$ tells us **how fast** $\hat{E}[Y|X]$ increases or decreases with a one unit change in X

A one unit increase in X is associated with a $\hat{\beta}_1$ unit increase in the fitted average of Y , $\hat{E}[Y|X]$.

Here a county with 1 point more of its adults holding a BA has a fitted Republican share 0.87 points lower. A county with 10 points more has a share 8.7 points lower. These are observational data, so the slope describes a difference between counties. It is not an effect of education.

Linear Regression: Justification

- Linear regression imposes a **strong** assumption on $E[Y|X]$
- Why would we ever want to do this?
 - Theoretical reason to assume linearity
 - Ease of interpretation
 - Bias-variance tradeoff
 - Analytical derivation of sampling distributions (next few weeks)
 - We can make the model more flexible, even in a linear framework (e.g. we can add polynomials, use log transformations, etc.)

What is Regression Analysis?

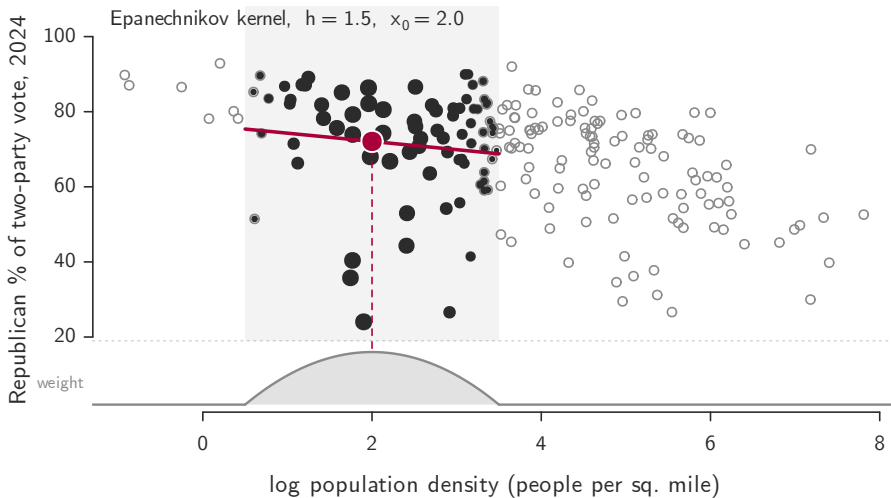
Nonparametric Regression

Linear Regression

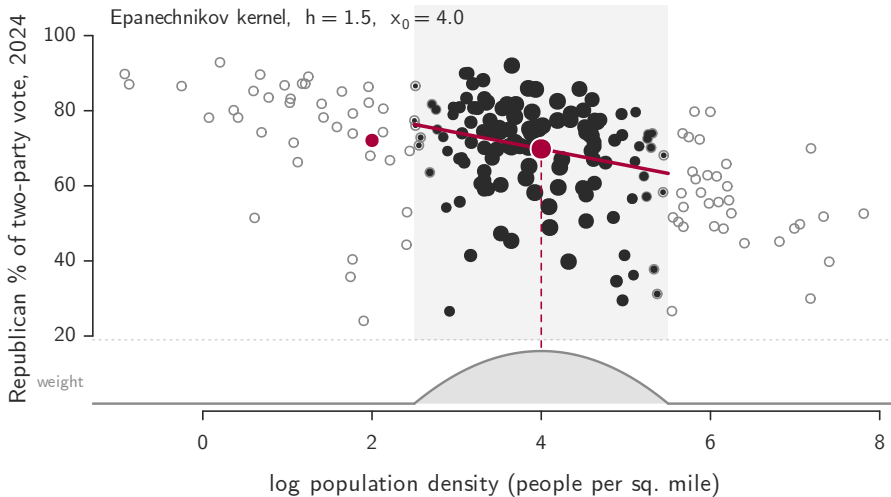
Combining Linear Regression with Nonparametric Regression

- The moving local average methods for nonparametric regression didn't take into account how y was *changing* with x
- We can incorporate this information by fitting a linear regression within each band
- Locally weighted scatterplot smoothing (**LOWESS** or **LOESS**):
 - ① Pick a subset of the data that falls in the interval $[x_0 - h, x_0 + h]$
 - ② Fit a line to this subset of the data (= **locally linear regression**), weighting the points by their distance to x_0 using a kernel function
 - ③ Use the fitted regression line to predict the expected value of $E[Y|X = x_0]$
- R's `loess()` sets the neighborhood by the nearest fraction `span` of the data (default 0.75), with tricube weights
- In R, `loess(Y ~ X, degree = 1)` fits local lines; the default, `loess(Y ~ X)`, fits local quadratics

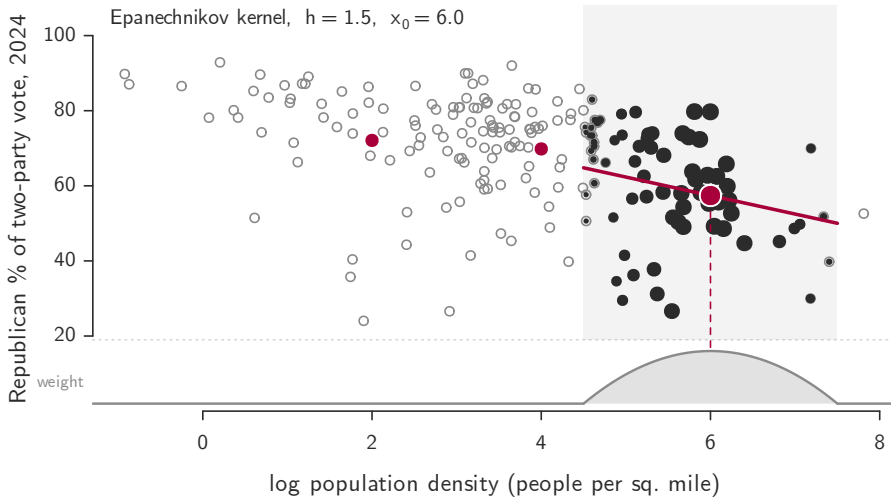
Weighted Local Linear Regressions



Weighted Local Linear Regressions



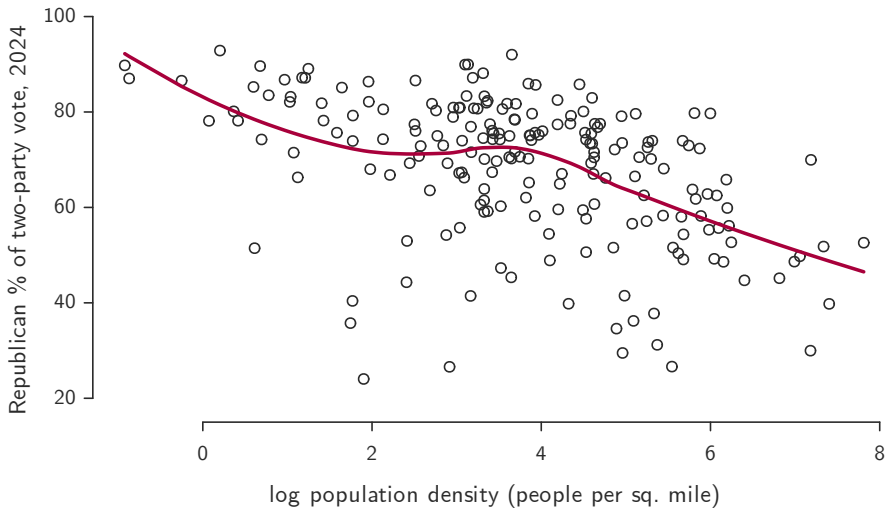
Weighted Local Linear Regressions



R Code

```
set.seed(450)
cty200 <- counties[sample(nrow(counties), 200), ] # a random sample of 200
cty200 <- cty200[order(cty200$log_density), ]
lo.out <- loess(rep24 ~ log_density, data = cty200)
plot(rep24 ~ log_density, data = cty200)
lines(lo.out$fitted, x = lo.out$x, col = "red")
```

LOESS Characterization of $E[Y|X]$

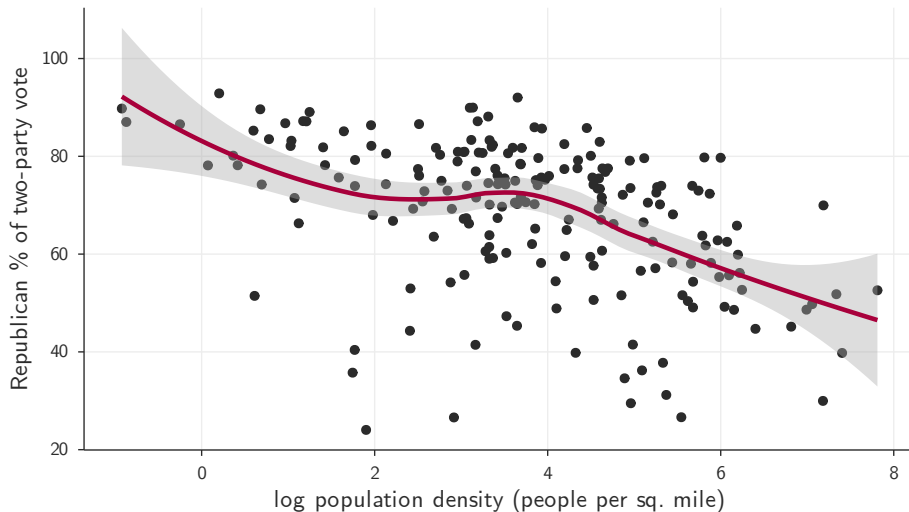


R Code

```
library(ggplot2)

ggplot(cty200, aes(x = log_density, y = rep24)) +
  geom_point() +
  geom_smooth(method = "loess")
```

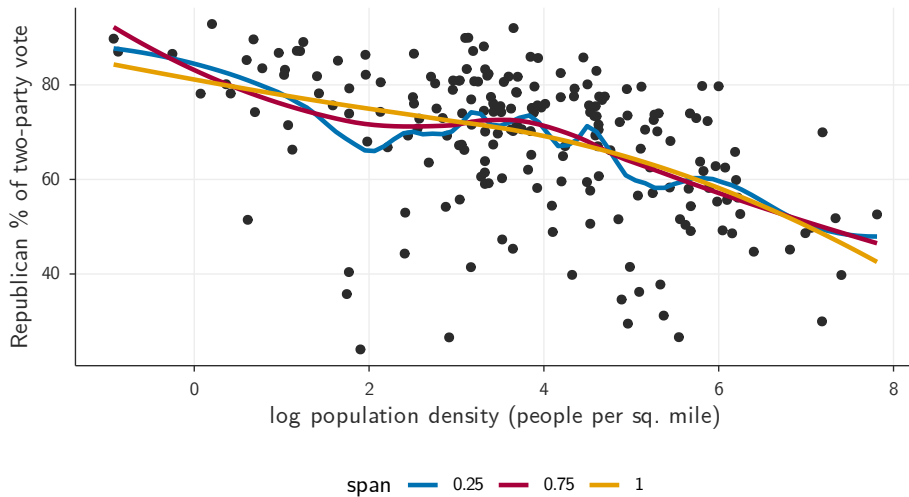
LOESS Characterization of $E[Y|X]$



R Code

```
ggplot(cty200, aes(x = log_density, y = rep24)) +  
  geom_point() +  
  geom_smooth(method = "loess", span = 0.25, se = FALSE, aes(colour = "0.25")) +  
  geom_smooth(method = "loess", span = 0.75, se = FALSE, aes(colour = "0.75")) +  
  geom_smooth(method = "loess", span = 1, se = FALSE, aes(colour = "1")) +  
  labs(colour = "span")
```

LOESS Characterization of $E[Y|X]$



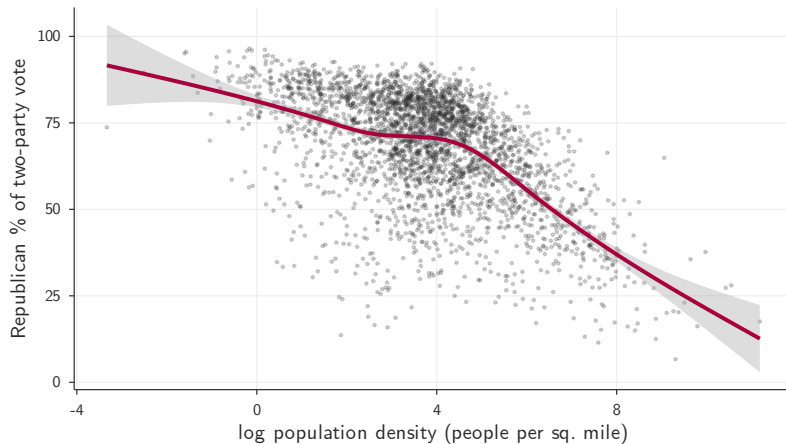
R Code

```
ggplot(counties, aes(x = log_density, y = rep24)) +  
  geom_point(size = 0.5, alpha = 0.2) +  
  geom_smooth(method = "gam", formula = y ~ s(x, bs = "cs"))
```

$s(x)$ is a smooth function of x . A GAM chooses how wiggly it is from the data. Now we use all 3,034 counties. The GAM picks a curve with 5.9 effective degrees of freedom (a straight line has 1). The curve falls about 3 points per unit of log density up to 4.5 (90 people per square mile), then about 9.

With many more data points, the bandwidth choice matters less. With all 3,034 counties, LOESS curves with span 0.25 and span 0.75 differ by at most 2.2 points. In the sample of 200 they differ by up to 5.7 points (both over the middle 96% of densities).

GAM Characterization of $E[Y|X]$



Source: MIT Election Data and Science Lab; Census ACS 2019–23 and 2023 Gazetteer; N = 3,034 counties (Alaska, Connecticut, Louisiana and 7 other counties excluded)