Political Methodology III: Model Based Inference

Justin Grimmer

Associate Professor
Department of Political Science
Stanford University

May 23rd, 2017

1) Task:
   - Estimate conditional expectation function (regression) $E[Y|\boldsymbol{X}]$
   - Estimate empirical distribution function $\hat{f}(x)$
   - Estimate conditional empirical distribution function $\hat{f}(x|\boldsymbol{Z})$
2) Objective Function
   - Mean Square Error $\rightsquigarrow$ Average (Predictive) Risk
   - Balance Bias-Variance Tradeoff
3) Optimization
   - Local Linear Regression
4) Validation
   - Bias/Variance Tradeoff in fitting data
   - LOOCV will be a workhorse tool

Nonparametric Regression

Motivating Example: Birth Weight

Motivating Example: Birth Weight

- Birth Weight is associated with a wide array of developmental outcomes

Motivating Example: Birth Weight

- Birth Weight is associated with a wide array of developmental outcomes
- Birth Weight varies significantly across socioeconomic and racial groups

Motivating Example: Birth Weight

- Birth Weight is associated with a wide array of developmental outcomes
- Birth Weight varies significantly across socioeconomic and racial groups
- Understand how Birth Weight varies with a variety of attributes

Motivating Example: Birth Weight

- Birth Weight is associated with a wide array of developmental outcomes
- Birth Weight varies significantly across socioeconomic and racial groups
- Understand how Birth Weight varies with a variety of attributes
- What are differences in birth weight across racial groups?

Motivating Example: Birth Weight

- Birth Weight is associated with a wide array of developmental outcomes
- Birth Weight varies significantly across socioeconomic and racial groups
- Understand how Birth Weight varies with a variety of attributes

- What are differences in birth weight across racial groups?
- Are differences larger (smaller) within sub-groups (smokers, obese, etc)?

Motivating Example: Birth Weight

- Birth Weight is associated with a wide array of developmental outcomes
- Birth Weight varies significantly across socioeconomic and racial groups
- Understand how Birth Weight varies with a variety of attributes

- What are differences in birth weight across racial groups?
- Are differences larger (smaller) within sub-groups (smokers, obese, etc)?

Suppose that we observe birth weight for a set of N observations of random variable birth weight,

Motivating Example: Birth Weight

- Birth Weight is associated with a wide array of developmental outcomes
- Birth Weight varies significantly across socioeconomic and racial groups
- Understand how Birth Weight varies with a variety of attributes

- What are differences in birth weight across racial groups?
- Are differences larger (smaller) within sub-groups (smokers, obese, etc)?

Suppose that we observe birth weight for a set of N observations of random variable birth weight,

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$$

Motivating Example: Birth Weight

- Birth Weight is associated with a wide array of developmental outcomes
- Birth Weight varies significantly across socioeconomic and racial groups
- Understand how Birth Weight varies with a variety of attributes

- What are differences in birth weight across racial groups?
- Are differences larger (smaller) within sub-groups (smokers, obese, etc)?

Suppose that we observe birth weight for a set of N observations of random variable birth weight,

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$$

We'll also observe a set of J covariates for each observation i

Motivating Example: Birth Weight

- Birth Weight is associated with a wide array of developmental outcomes
- Birth Weight varies significantly across socioeconomic and racial groups
- Understand how Birth Weight varies with a variety of attributes

- What are differences in birth weight across racial groups?
- Are differences larger (smaller) within sub-groups (smokers, obese, etc)?

Suppose that we observe birth weight for a set of N observations of random variable birth weight,

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$$

We'll also observe a set of J covariates for each observation i

$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$$

Conditional Expectation Functions and Regressions

Definition

For any two random variables Y and X define the regression of Y on X as

$$f(x) \equiv E[Y|X = x] = \int_{-\infty}^{\infty} y f(y|X = x) dy$$

Conditional Expectation Functions and Regressions

Definition

For any two random variables Y and X define the regression of Y on X as

$$f(x) \equiv E[Y|X=x] = \int_{-\infty}^{\infty} y f(y|X=x) dy$$

- Linear Regression is a special case with:

$$f(x) = \alpha + \beta x$$

- $E[Y|X]$ is called a conditional expectation function

Conditional Expectation Functions and Regressions

Definition

For any two random variables Y and X define the regression of Y on X as

$$f(x) \equiv E[Y|X = x] = \int_{-\infty}^{\infty} y f(y|X = x) dy$$

- Linear Regression is a special case with:

$$f(x) = \alpha + \beta x$$

- $E[Y|X]$ is called a conditional expectation function

We're going to consider more flexible functional forms for $f(x)$

Conditional Expectation Functions and Regressions

Definition

For any two random variables Y and X define the regression of Y on X as

$$f(x) \equiv E[Y|X = x] = \int_{-\infty}^{\infty} y f(y|X = x) dy$$

- Linear Regression is a special case with:

$$f(x) = \alpha + \beta x$$

- $E[Y|X]$ is called a conditional expectation function

We're going to consider more flexible functional forms for $f(x)$
Our estimator $\hat{f}(x)$ will estimate $f(x)$

Estimating $E[Y|X = x]$

Easy case: suppose X is discrete

Estimating $E[Y|X = x]$

Easy case: suppose X is discrete

Further suppose that $N \rightarrow \infty$. Infinite number of realizations of Y, $(y_1, \ldots, y_N)$

Estimating $E[Y|X = x]$

Easy case: suppose X is discrete
Further suppose that $N \rightarrow \infty$. Infinite number of realizations of Y, $(y_1, \ldots, y_N)$
And accompanying covariates $(x_1, \ldots, x_N)$, then

Estimating $E[Y|X = x]$

Easy case: suppose X is discrete
Further suppose that $N \rightarrow \infty$. Infinite number of realizations of Y, $(y_1, \ldots, y_N)$
And accompanying covariates $(x_1, \ldots, x_N)$, then

$$
\begin{aligned}
\widehat{f}(x) &= \text{average in each bin} \\
&= \frac{\sum_{i=1}^{N} y_i I(x_i = x)}{\sum_{i=1}^{N} I(x_i = x)}
\end{aligned}
$$

Bias-Variance Tradeoff

Bias-Variance Tradeoff

Hard Case: suppose X is continuous

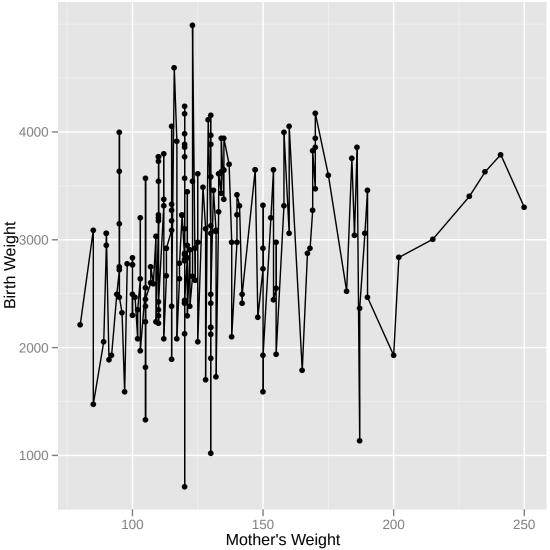
Bias-Variance Tradeoff

Hard Case: suppose X is continuous
Then all values of $x_1, x_2, \ldots, x_N$ are distinct.

Bias-Variance Tradeoff

Hard Case: suppose X is continuous
Then all values of $x_1, x_2, \ldots, x_N$ are distinct.
If we calculate conditional mean naively, we get the observed values of $y_1, \ldots, y_N$:

Bias-Variance Tradeoff



< R Code >

```
library(ggplot2)
qplot(lwt, bwt,
geom=c('point'),
xlab='Mothers Weight',
ylab='Birth Weight')
+geom_line()
```

Bias-Variance Tradeoff

Minimize variance:

Bias-Variance Tradeoff

Minimize variance: assume $\widehat{f}(x)$ is constant

Bias-Variance Tradeoff

Minimize variance: assume $\widehat{f}(x)$ is constant

$$\widehat{f}(x) = \frac{\sum_{i=1}^{N} Y_i}{N}$$

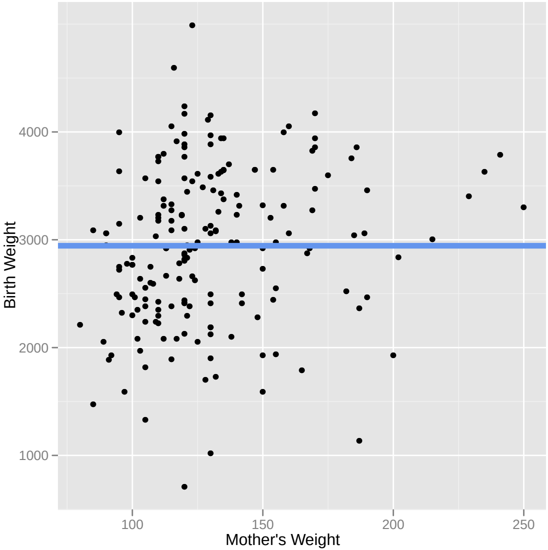
Bias-Variance Tradeoff

Minimize variance: assume $\widehat{f}(x)$ is constant

$$\widehat{f}(x) = \frac{\sum_{i=1}^{N} Y_i}{N}$$

This yields:

Bias-Variance Tradeoff




```
library(ggplot2)
qplot(lwt, bwt,
geom=c('point'),
xlab='Motherś Weight',
ylab='Birth Weight')
+geom_abline
(intercept=mean(bwt),
slope=0, lwd=2,
col='cornflowerblue')
```

Compromise

Curse of Dimensionality

- We're going to learn around a neighborhood of a point

Curse of Dimensionality

- We're going to learn around a neighborhood of a point
- Problem: high-dimensional space is sparse

Curse of Dimensionality

- We're going to learn around a neighborhood of a point
- Problem: high-dimensional space is sparse
- Suppose We have N draws of $X \sim \text{Uniform}(-1, 1)$

Curse of Dimensionality

- We're going to learn around a neighborhood of a point
- Problem: high-dimensional space is sparse
- Suppose We have N draws of $X \sim$ Uniform$(-1, 1)$
- How many points are between $[-0.1, 0.1]$?

Curse of Dimensionality

- We're going to learn around a neighborhood of a point
- Problem: high-dimensional space is sparse
- Suppose We have N draws of $X \sim \text{Uniform}(-1, 1)$
- How many points are between $[-0.1, 0.1]$?

$$N \times \Pr(X \in [-0.1, 0.1]) \quad = \quad N \times \frac{0.2}{2} = \frac{N}{10}$$

Curse of Dimensionality

- We're going to learn around a neighborhood of a point
- Problem: high-dimensional space is sparse
- Suppose We have N draws of $X \sim \text{Uniform}(-1, 1)$
- How many points are between $[-0.1, 0.1]$?

$$N \times \Pr(X \in [-0.1, 0.1]) \quad = \quad N \times \frac{0.2}{2} = \frac{N}{10}$$

- Suppose N draws $\boldsymbol{X} = (X_1, X_2, \ldots, X_{10})$ where

Curse of Dimensionality

- We're going to learn around a neighborhood of a point
- Problem: high-dimensional space is sparse
- Suppose We have N draws of $X \sim \text{Uniform}(-1, 1)$
- How many points are between $[-0.1, 0.1]$?

$$N \times \Pr(X \in [-0.1, 0.1]) \;=\; N \times \frac{0.2}{2} = \frac{N}{10}$$

- Suppose N draws $\boldsymbol{X} = (X_1, X_2, \ldots, X_{10})$ where

$$X_i \sim_{iid} \text{Uniform}(-1, 1)$$

Curse of Dimensionality

- We're going to learn around a neighborhood of a point
- Problem: high-dimensional space is sparse
- Suppose We have N draws of $X \sim \text{Uniform}(-1, 1)$
- How many points are between $[-0.1, 0.1]$?

$$N \times \Pr(X \in [-0.1, 0.1]) \quad = \quad N \times \frac{0.2}{2} = \frac{N}{10}$$

- Suppose N draws $\boldsymbol{X} = (X_1, X_2, \ldots, X_{10})$ where

$$X_i \sim_{iid} \text{Uniform}(-1, 1)$$

- How many points between -0.1 and 0.1 on all dimensions?

Curse of Dimensionality

- We're going to learn around a neighborhood of a point
- Problem: high-dimensional space is sparse
- Suppose We have N draws of $X \sim$ Uniform$(-1, 1)$
- How many points are between $[-0.1, 0.1]$?

$$N \times \Pr(X \in [-0.1, 0.1]) \quad = \quad N \times \frac{0.2}{2} = \frac{N}{10}$$

- Suppose N draws $\boldsymbol{X} = (X_1, X_2, \ldots, X_{10})$ where

$$X_i \sim_{iid} \text{Uniform}(-1, 1)$$

- How many points between -0.1 and 0.1 on all dimensions?

$$N \times \Pr(\boldsymbol{X} \in [-0.1, 0.1] \times [-0.1, 0.1] \times \ldots \times [-0.1, 0.1]) \quad =$$

Curse of Dimensionality

- We're going to learn around a neighborhood of a point
- Problem: high-dimensional space is sparse
- Suppose We have N draws of $X \sim$ Uniform$(-1,1)$
- How many points are between $[-0.1, 0.1]$?

$$N \times \text{Pr}(X \in [-0.1, 0.1]) \quad = \quad N \times \frac{0.2}{2} = \frac{N}{10}$$

- Suppose N draws $\boldsymbol{X} = (X_1, X_2, \ldots, X_{10})$ where

$$X_i \sim_{iid} \text{Uniform}(-1,1)$$

- How many points between -0.1 and 0.1 on all dimensions?

$$N \times \text{Pr}(\boldsymbol{X} \in [-0.1, 0.1] \times [-0.1, 0.1] \times \ldots \times [-0.1, 0.1]) \quad = \quad N \times \left(\frac{1}{10}\right)^{10} = \frac{N}{10{,}000{,}000{,}000}$$

Bias-Variance Tradeoff

Modeling the Relationship

$$Y = f(x_i) + \epsilon_i$$

Suppose $f(x_i) \equiv \beta_0 + \beta_{j=1}^{J} x_{ij}$ and $\boldsymbol{x} = (1, x_1, \ldots, x_J)$.

Modeling the Relationship

$$Y = f(x_i) + \epsilon_i$$

Suppose $f(x_i) \equiv \beta_0 + \beta_{j=1}^{J} x_{ij}$ and $\boldsymbol{x} = (1, x_1, \ldots, x_J)$.

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y}$$

Modeling the Relationship

$$Y = f(x_i) + \epsilon_i$$

Suppose $f(x_i) \equiv \beta_0 + \beta_{j=1}^{J} x_{ij}$ and $\boldsymbol{x} = (1, x_1, \ldots, x_J)$.

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y}$$

$$\hat{f}(\boldsymbol{X}) \equiv \widehat{\boldsymbol{Y}} = \underbrace{\boldsymbol{X}\left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}}_{\boldsymbol{H}}\boldsymbol{Y}$$

Modeling the Relationship

$$Y = f(x_i) + \epsilon_i$$

Suppose $f(x_i) \equiv \beta_0 + \beta_{j=1}^{J} x_{ij}$ and $\boldsymbol{x} = (1, x_1, \ldots, x_J)$.

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
\hat{f}(\boldsymbol{X}) \equiv \widehat{\boldsymbol{Y}} &= \underbrace{\boldsymbol{X}\left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}}_{\boldsymbol{H}}\boldsymbol{Y} \\
J + 1 &= \mathrm{tr}(\boldsymbol{H})
\end{aligned}$$

Modeling the Relationship

$$Y = f(x_i) + \epsilon_i$$

Suppose $f(x_i) \equiv \beta_0 + \beta_{j=1}^{J} x_{ij}$ and $\boldsymbol{x} = (1, x_1, \ldots, x_J)$.

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
\hat{f}(\boldsymbol{X}) \equiv \widehat{\boldsymbol{Y}} &= \underbrace{\boldsymbol{X}\left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}}_{\boldsymbol{H}}\boldsymbol{Y} \\
J + 1 &= \mathrm{tr}(\boldsymbol{H}) \\
\hat{f}(\boldsymbol{x}) &= \boldsymbol{x}^{'}\widehat{\boldsymbol{\beta}}
\end{aligned}$$

Modeling the Relationship

$$Y = f(x_i) + \epsilon_i$$

Suppose $f(x_i) \equiv \beta_0 + \beta_{j=1}^{J} x_{ij}$ and $\boldsymbol{x} = (1, x_1, \ldots, x_J)$.

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
\hat{f}(\boldsymbol{X}) \equiv \widehat{\boldsymbol{Y}} &= \underbrace{\boldsymbol{X}\left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}}_{\boldsymbol{H}}\boldsymbol{Y} \\
J+1 &= \mathrm{tr}(\boldsymbol{H}) \\
\hat{f}(\boldsymbol{x}) &= \boldsymbol{x}^{'}\widehat{\boldsymbol{\beta}} \\
&= \underbrace{\boldsymbol{x}^{'}}_{1\times J}\underbrace{\left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}}_{J\times J}\underbrace{\boldsymbol{X}^{'}}_{J\times N}\underbrace{\boldsymbol{Y}}_{N\times 1}
\end{aligned}$$

Modeling the Relationship

$$Y = f(x_i) + \epsilon_i$$

Suppose $f(x_i) \equiv \beta_0 + \beta_{j=1}^{J} x_{ij}$ and $\boldsymbol{x} = (1, x_1, \ldots, x_J)$.

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
\hat{f}(\boldsymbol{X}) \equiv \widehat{\boldsymbol{Y}} &= \underbrace{\boldsymbol{X}\left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}}_{\boldsymbol{H}}\boldsymbol{Y} \\
J+1 &= \mathsf{tr}(\boldsymbol{H}) \\
\hat{f}(\boldsymbol{x}) &= \boldsymbol{x}^{'}\widehat{\boldsymbol{\beta}} \\
&= \underbrace{\boldsymbol{x}^{'}}_{1\times J}\underbrace{\left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}}_{J\times J}\underbrace{\boldsymbol{X}^{'}}_{J\times N}\underbrace{\boldsymbol{Y}}_{N\times 1} \\
&= \sum_{i=1}^{N} h_i(\boldsymbol{x})Y_i
\end{aligned}$$

Modeling the Relationship

Definition

We will say $\hat{f}(\boldsymbol{x})$, an estimator of $f(\boldsymbol{x})$, is a linear smoother if for each $\boldsymbol{x}$ there exists a vector $\boldsymbol{h}(\boldsymbol{x}) = (h_1(\boldsymbol{x}), h_2(\boldsymbol{x}), \ldots, h_N(\boldsymbol{x}))$ such that

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{N} h_i(\boldsymbol{x}) Y_i$$

This implies that fitted values $\widehat{\boldsymbol{Y}} = (\hat{f}(\boldsymbol{x}_1), \hat{f}(\boldsymbol{x}_2), \ldots, \hat{f}(\boldsymbol{x}_N))$ and

$$\widehat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$$

where i^{th} row is $\boldsymbol{h}(\boldsymbol{x}_i) = (h_1(\boldsymbol{x}_1), h_2(\boldsymbol{x}_1), \ldots, h_N(\boldsymbol{x}_1))$

Modeling the Relationship

Definition

The matrix $\boldsymbol{H}$ is the smoothing matrix. The row $h(\boldsymbol{x}_i)$ is called the effective kernel for observation i, and the effective number of parameters $\nu =$ tr$(\boldsymbol{H})$.

"Regressogram"

- Suppose we have one variable x_i
- Suppose we divide our data into K bins,$B_k = \{x : a < x < b\}$, $k = 1, 2, \ldots, K$ so that there are 3 observations in each bin $|B_k| = 3$ for all k
- We will say that the fitted value for x_i is then

$$\hat{f}(x_i) = \sum_{x_i \in B_k} \frac{Y_i}{3}$$

"Regressogram"

- In smoothing terms (sorting observations according to x_i values)

$$
\begin{aligned}
\boldsymbol{h}(x_i) &= (0,0,0,\frac{1}{3},\frac{1}{3},\frac{1}{3},0,0,0) \\
\boldsymbol{H} &= \begin{pmatrix}
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\
0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\
0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3}
\end{pmatrix}
\end{aligned}
$$

Local Averages

- Define $B_x = \{i : |x_i - x| < b\}$ and $|B_x| = n_x$.
- If $n_x > 0$,

$$\hat{f}(x) = \sum_{i \in B_x} \frac{Y_i}{n_x}$$

In smoothing terms

$$\begin{aligned} h_i(x) &= 0 \text{ if } |x_i - x| > b \\ h_i(x) &= \frac{1}{n_x} \text{ if } |x_i - x| < b \end{aligned}$$

More Sophisticated Weights $\rightsquigarrow$ Kernels

- Previous examples are binary (in/out)

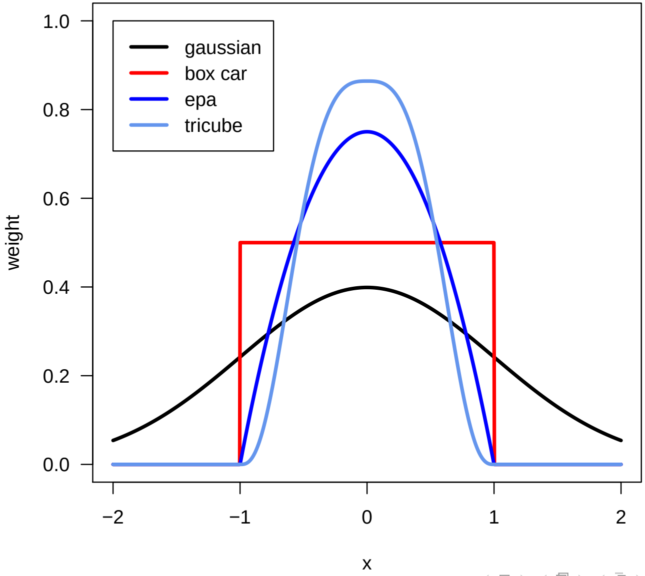
More Sophisticated Weights $\rightsquigarrow$ Kernels

- Previous examples are binary (in/out)
- Continuous weights via Kernel

More Sophisticated Weights $\rightsquigarrow$ Kernels

- Previous examples are binary (in/out)
- Continuous weights via Kernel
- Some famous examples (Define $I(x) \equiv 1$ if $|x| < 1,$ otherwise 0)

$$
\begin{aligned}
K(x) &= \frac{1}{2} I(x) \text{ (Box Car)} \\
K(x) &= \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right) \text{ (Gaussian)} \\
K(x) &= \frac{3}{4}(1 - x^2) I(x) \text{ (Epanechnikov)} \\
K(x) &= \frac{70}{81}(1 - |x|^3)^3 I(x) \text{ (Tri-cube)}
\end{aligned}
$$


gaussian
box car
epa
tricube
weight
x
1.0
0.8
0.6
0.4
0.2
0.0
−2
−1
0
1
2

Kernels and Bandwidth

Define $b > 0$ to be the bandwidth.

Kernels and Bandwidth

Define $b > 0$ to be the bandwidth.

$$K\left(\frac{x - x_i}{b}\right) \quad = \quad \text{Scales difference between } x \text{ and } x_i \text{ by } b$$

Kernels and Bandwidth

Define $b > 0$ to be the bandwidth.

$$K\left(\frac{x - x_i}{b}\right) \quad = \quad \text{Scales difference between } x \text{ and } x_i \text{ by } b$$

Consider the box-car kernel

Kernels and Bandwidth

Define $b > 0$ to be the bandwidth.

$$K\left(\frac{x - x_i}{b}\right) = \text{Scales difference between } x \text{ and } x_i \text{ by } b$$

Consider the box-car kernel

$$\frac{1}{2} I\left(\frac{x - x_i}{1}\right) = \frac{1}{2} \text{ if } |x - x_i| < 1$$

Kernels and Bandwidth

Define $b > 0$ to be the bandwidth.

$$K\left(\frac{x - x_i}{b}\right) \quad = \quad \text{Scales difference between } x \text{ and } x_i \text{ by } b$$

Consider the box-car kernel

$$\begin{aligned}
\frac{1}{2} I\left(\frac{x - x_i}{1}\right) &= \frac{1}{2} \text{ if } |x - x_i| < 1 \\
\frac{1}{2} I\left(\frac{x - x_i}{100}\right) &= \frac{1}{2} \text{ if } |x - x_i| < 100
\end{aligned}$$

Kernels and Bandwidth

Define $b > 0$ to be the bandwidth.

$$K\left(\frac{x - x_i}{b}\right) = \text{Scales difference between } x \text{ and } x_i \text{ by } b$$

Consider the box-car kernel

$$\frac{1}{2} I\left(\frac{x - x_i}{1}\right) = \frac{1}{2} \text{ if } |x - x_i| < 1$$
$$\frac{1}{2} I\left(\frac{x - x_i}{100}\right) = \frac{1}{2} \text{ if } |x - x_i| < 100$$
$$\frac{1}{2} I\left(\frac{x - x_i}{0.01}\right) = \frac{1}{2} \text{ if } |x - x_i| < 0.01$$

Kernels and Bandwidth

Define $b > 0$ to be the bandwidth.

$$K\left(\frac{x - x_i}{b}\right) = \text{Scales difference between } x \text{ and } x_i \text{ by } b$$

Consider the box-car kernel

$$\begin{aligned}
\frac{1}{2} I\left(\frac{x - x_i}{1}\right) &= \frac{1}{2} \text{ if } |x - x_i| < 1 \\
\frac{1}{2} I\left(\frac{x - x_i}{100}\right) &= \frac{1}{2} \text{ if } |x - x_i| < 100 \\
\frac{1}{2} I\left(\frac{x - x_i}{0.01}\right) &= \frac{1}{2} \text{ if } |x - x_i| < 0.01
\end{aligned}$$

$b \rightsquigarrow$ controls the sensitivity of Kernel

Local Regression

Definition

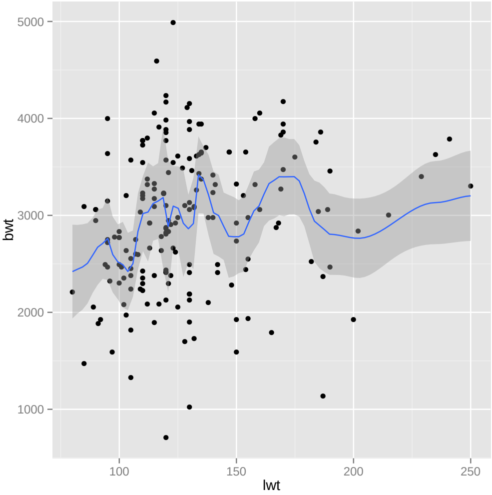
Suppose $b > 0$*. Define the local regression estimator* $\hat{f}(x)$ *of* $f(x)$ *as*

$$\hat{f}(x) = \sum_{i=1}^{N} h_i(x) Y_i$$

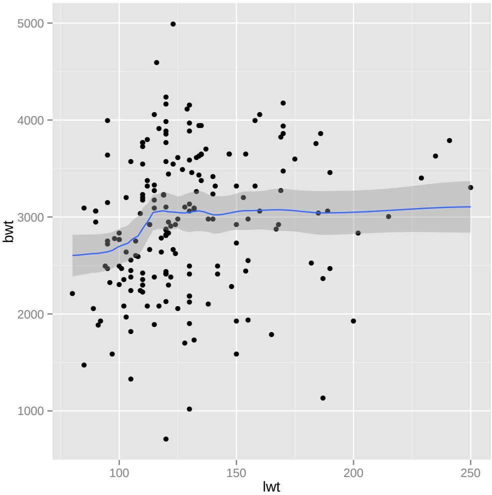
where

$$h_i(x) = \frac{K\left(\frac{x - x_i}{b}\right)}{\sum_{j=1}^{N} K\left(\frac{x - x_j}{b}\right)}$$

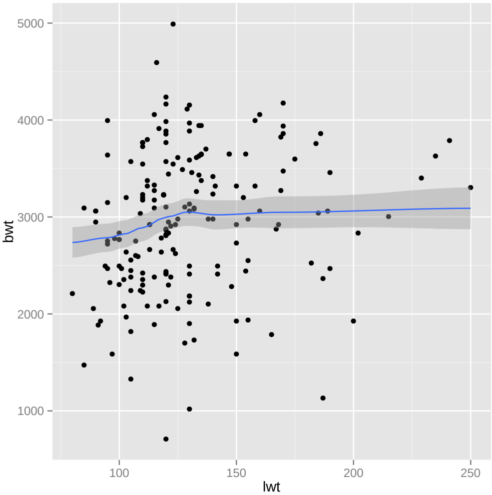
$\boldsymbol{H} = N \times N$ *matrix with* $H_{ij} = h_j(x_i)$



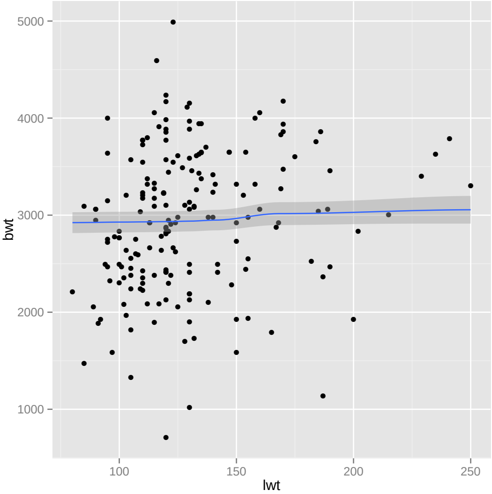

```
qplot(lwt, bwt,
geom='point') +
stat_smooth(degree = 0,
span = 0.1)
```




```
qplot(lwt, bwt,
geom='point') +
stat_smooth(degree = 0,
span = 0.4)
```




```
qplot(lwt, bwt,
geom='point') +
stat_smooth(degree = 0,
span = 0.7)
```




```
qplot(lwt, bwt,
geom='point') +
stat_smooth(degree = 0,
span = 1)
```

Local Regression

$$
\begin{aligned}
w_i(x) &= K\left(\frac{x_i - x}{b}\right) \\
\hat{a} &= \text{arg min}_a \sum_{i=1}^{N} w_i(x)(Y_i - a)^2 \\
\hat{a} &= \frac{\sum_{i=1}^{N} w_i(x) Y_i}{\sum_{i=1}^{N} w_i(x)} \\
\hat{a} &= \hat{f}(x)
\end{aligned}
$$

Local Polynomial regression

$$(\hat{a}, \hat{b}, \hat{c}) = \text{arg min}_{a,b,c} \sum_{i=1}^{N} w_i(x)(Y_i - a - bx_i - cx_i^2)^2$$

Define $\boldsymbol{X}_x$ to be an $N \times 3$ matrix:

$$\boldsymbol{X}_x = \begin{pmatrix} 1 & x_1 - x & (x_1 - x)^2 \\ 1 & x_2 - x & (x_2 - x)^2 \\ \vdots & \vdots & \vdots \\ 1 & x_N - x & (x_N - x)^2 \end{pmatrix}$$

$\boldsymbol{W}_x = N \times N$ matrix where $W_{ii} = w_i(x)$.

$$\hat{f}(x) = \left(\boldsymbol{X}_x^{'} \boldsymbol{W}_x \boldsymbol{X}_x\right)^{-1} \boldsymbol{X}_x^{'} \boldsymbol{W}_x \boldsymbol{Y}$$

Local Polynomial Regression

Definition

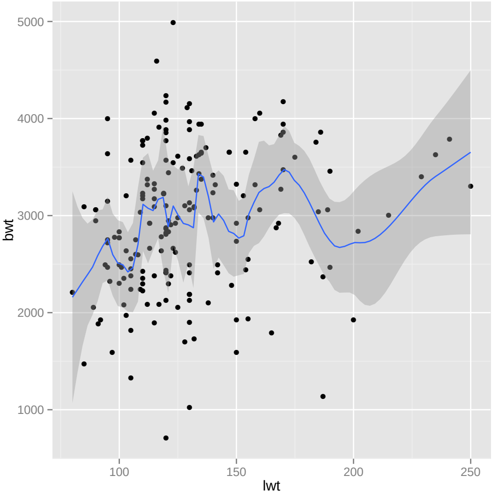
Suppose $b > 0$. Define the local polynomial regression estimator $\hat{f}(x)$ of $f(x)$ as

$$\hat{f}(x) \quad = \quad \left(\boldsymbol{X}_x^{'}\boldsymbol{W}_x\boldsymbol{X}_x\right)^{-1}\boldsymbol{X}_x^{'}\boldsymbol{W}_x\boldsymbol{Y}$$

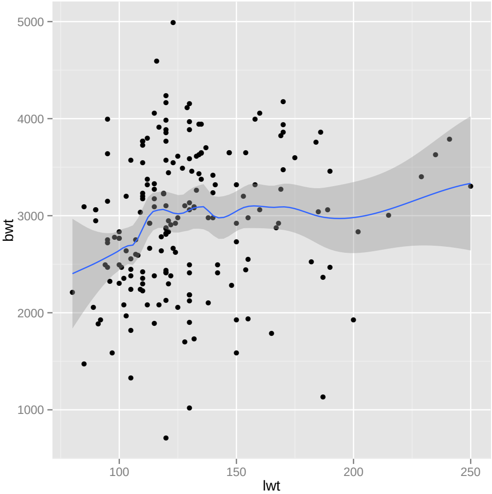
We can write this as $\hat{f}(x) = \sum_{i=1}^{N} h_i(x)Y_i$, where $\boldsymbol{h}(x) = (h_1(x), h_2(x), \ldots, h_N(x))$

$$h(x) \quad = \quad \boldsymbol{e}_1\left(\boldsymbol{X}_x^{'}\boldsymbol{W}_x\boldsymbol{X}_x\right)^{-1}\boldsymbol{X}_x^{'}\boldsymbol{W}_x$$

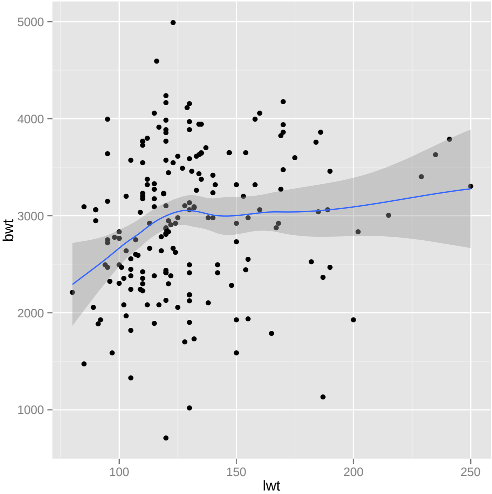
where $\boldsymbol{e}_1 = (1, 0, \ldots, 0)$.
$\boldsymbol{H}$ is an $N \times N$ matrix with typical entry $H_{ij} = h_j(x_i)$.



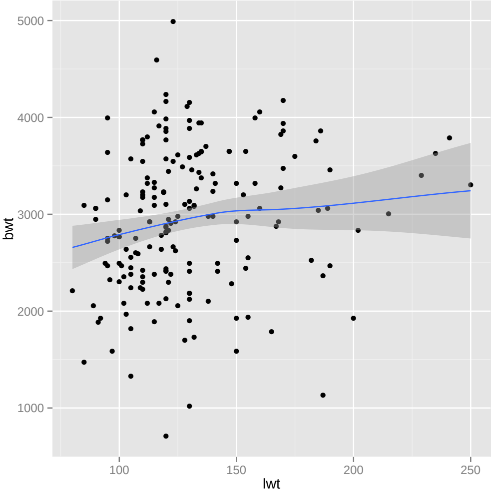

```
qplot(lwt, bwt,
geom='point') +
stat_smooth(degree = 1,
span = 0.1)
```



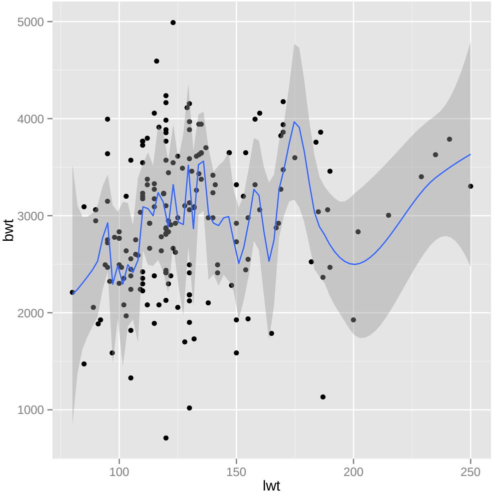

```
qplot(lwt, bwt,
geom='point') +
stat_smooth(degree = 1,
span = 0.4)
```



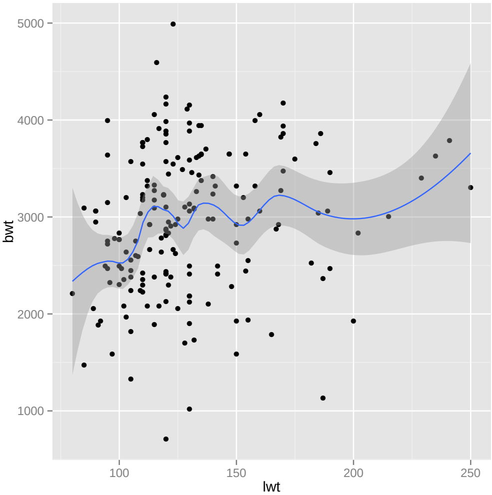

```
qplot(lwt, bwt,
geom='point') +
stat_smooth(degree = 1,
span = 0.7)
```



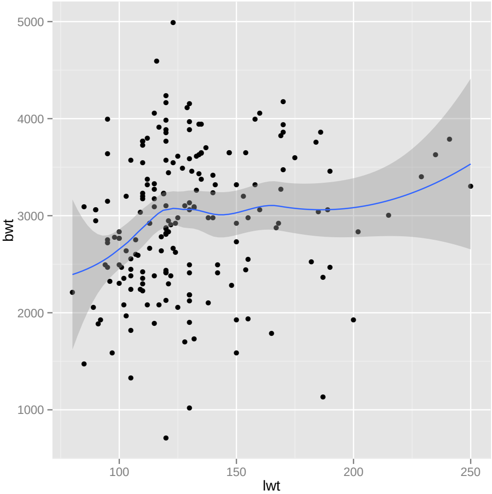

```
qplot(lwt, bwt,
geom='point') +
stat_smooth(degree = 1,
span = 1)
```



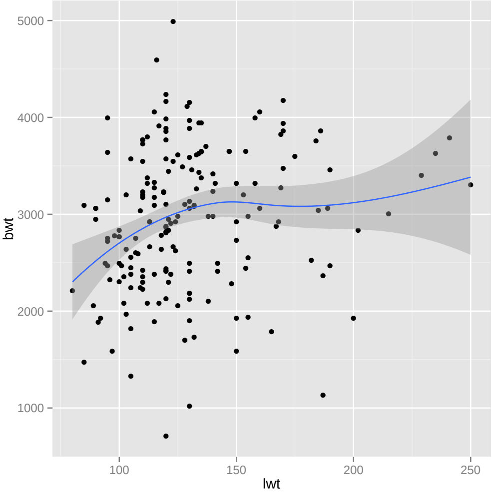

```
qplot(lwt, bwt,
geom='point') +
stat_smooth(degree = 2,
span = 0.1)
```



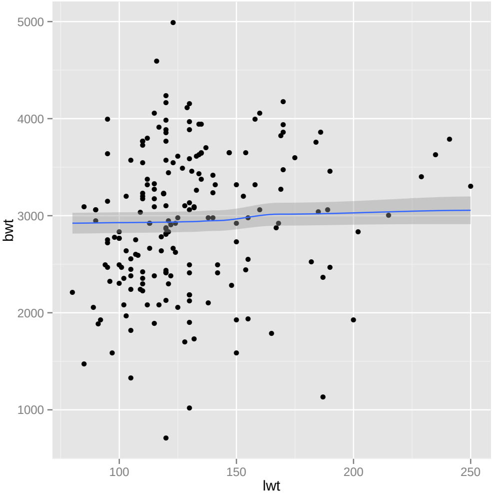

```
qplot(lwt, bwt,
geom='point') +
stat_smooth(degree = 2,
span = 0.4)
```



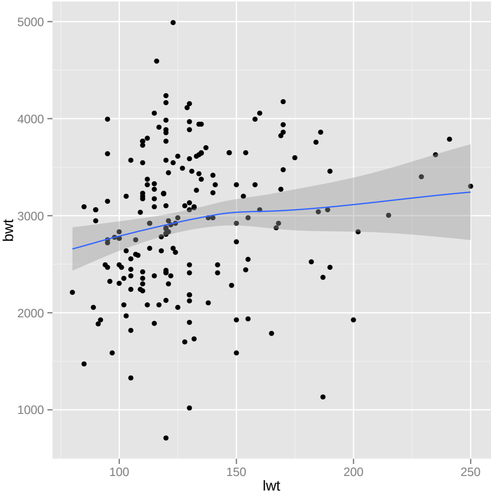

```
qplot(lwt, bwt,
geom='point') +
stat_smooth(degree = 2,
span = 0.7)
```



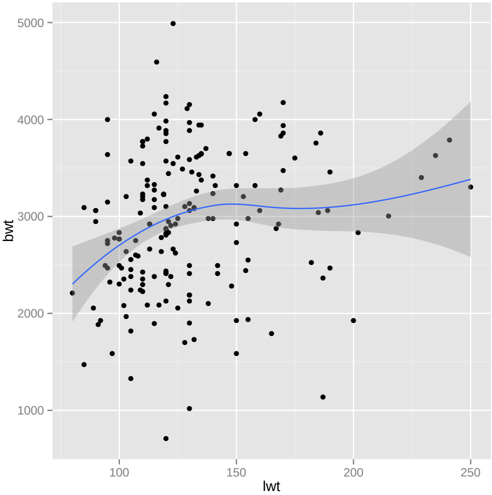

```
qplot(lwt, bwt,
geom='point') +
stat_smooth(degree = 2,
span = 1)
```




```
qplot(lwt, bwt,
geom='point') +
stat_smooth(degree = 0,
span = 1)
```




```
qplot(lwt, bwt,
geom='point') +
stat_smooth(degree = 1,
span = 1)
```




```
qplot(lwt, bwt,
geom='point') +
stat_smooth(degree = 1,
span = 1)
```

Bandwidth Selection

Leave One Out Cross Validation LOOCV

- Define $\hat{f}_{-i,b}(x)$ to be the linear smooth estimtator without observation i and bandwidth b.
- The LOOCV statistic will be

$$\text{CV(b)} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{f}_{-i,b}(x)\right)^2$$

Theorem

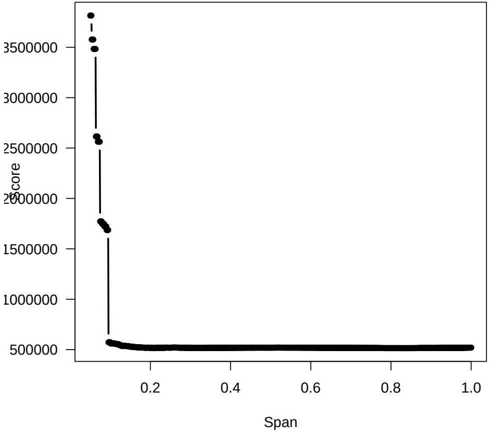
Suppose $\hat{f}(x)$ is a linear smoother. Then CV(b) can be written as

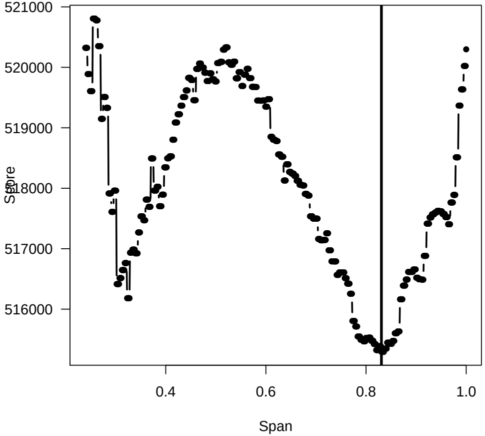
$$\text{CV(b)} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i - \hat{f}_b(x)}{1 - H_{ii}}\right)^2$$

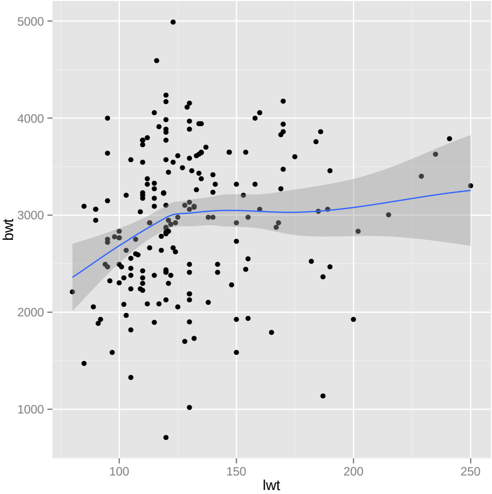
And is often approximated with the Generalized Cross Validation (GCV):

$$\text{GCV(b)} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i - \hat{f}_b(x)}{1 - \nu/N}\right)^2$$

where $\nu = tr(\boldsymbol{H})$


3500000
3000000
2500000
2000000
1500000
1000000
500000
Score
0.2
0.4
0.6
0.8
1.0
Span


521000
520000
519000
518000
517000
516000
Score
0.4
0.6
0.8
1.0
Span




```
qplot(lwt, bwt,
geom='point') +
stat_smooth(degree = 1,
span = 0.83)
```

Loess and Conditioning

Many models:

Loess and Conditioning

Many models: condition on many factors in order to learn relationship

Loess and Conditioning

Many models: condition on many factors in order to learn relationship
Possible to visualize relationships in lattice:

Loess and Conditioning

Many models: condition on many factors in order to learn relationship
Possible to visualize relationships in lattice:
Suppose we have continuous random variable X and discrete random variables Z_1 and Z_2.

Loess and Conditioning

Many models: condition on many factors in order to learn relationship
Possible to visualize relationships in lattice:
Suppose we have continuous random variable X and discrete random variables Z_1 and Z_2.
Then we might estimate

Loess and Conditioning

Many models: condition on many factors in order to learn relationship
Possible to visualize relationships in lattice:
Suppose we have continuous random variable X and discrete random variables Z_1 and Z_2.
Then we might estimate

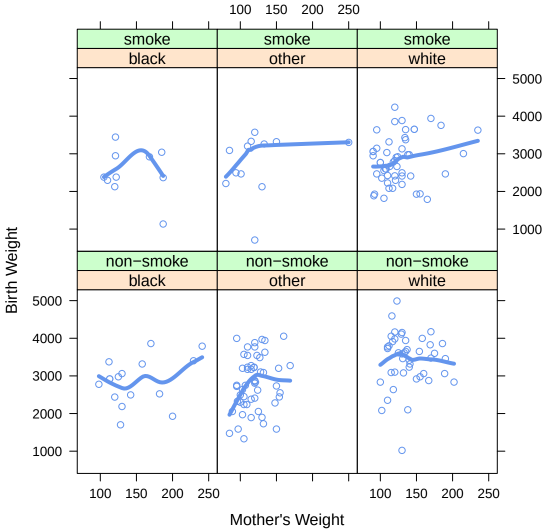
$$f(x, z_1, z_2) = E[Y|X = x, Z_1 = z_1, Z_2 = z_2]$$

Loess and Conditioning

Many models: condition on many factors in order to learn relationship
Possible to visualize relationships in lattice:
Suppose we have continuous random variable X and discrete random variables Z_1 and Z_2.
Then we might estimate

$$
\begin{aligned}
f(x, z_1, z_2) &= E[Y|X = x, Z_1 = z_1, Z_2 = z_2] \\
&= \int_{-\infty}^{\infty} y f(y|X = x, Z_1 = z_1, Z_2 = z_2) dy
\end{aligned}
$$

Loess and Conditioning



< R Code >

```
xyplot(bwt~lwt|race*smoke
lwd=4,
col='cornflowerblue',
xlab='Mother's Weight',
ylab='Birth Weight',
panel= function(x, y,
...){
panel.xyplot(x, y,
...);
panel.loess(x,y,span=.8,
degree=1, ...)  } )
```

Densities

Histogram, Probability Mass Function

Simplest case: probability mass functions

Histogram, Probability Mass Function

Simplest case: probability mass functions

- Suppose we observe $x_1, x_2, \ldots, x_N$

Histogram, Probability Mass Function

Simplest case: probability mass functions

- Suppose we observe $x_1, x_2, \ldots, x_N$
- If X's takes on few unique values, say $K << N$, then an estimator for the pmf is,

Histogram, Probability Mass Function

Simplest case: probability mass functions

- Suppose we observe $x_1, x_2, \ldots, x_N$
- If X's takes on few unique values, say $K << N$, then an estimator for the pmf is,

$$\widehat{p}(X = x) \equiv \widehat{p}(x) \quad = \quad \frac{\sum_{i=1}^{N} I(x_i = x)}{N}$$

Histogram, Probability Mass Function

Simplest case: probability mass functions

- Suppose we observe $x_1, x_2, \ldots, x_N$
- If X's takes on few unique values, say $K << N$, then an estimator for the pmf is,

$$\widehat{p}(X = x) \equiv \widehat{p}(x) \quad = \quad \frac{\sum_{i=1}^{N} I(x_i = x)}{N}$$

Will converge on the probability mass function (by weak law of large numbers)

Histogram, Probability Density Function

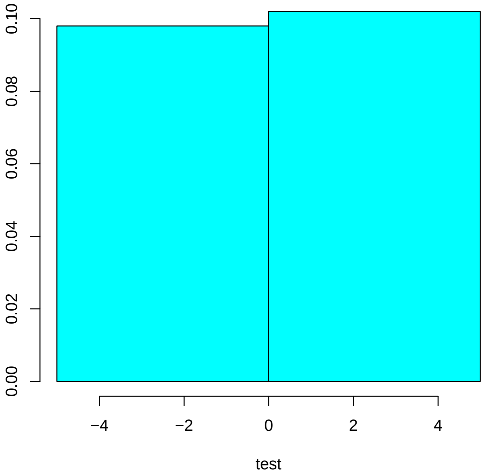
Harder case: probability density functions

- Suppose we observe $x_1, x_2, \ldots, x_N$, with X's continuous
- Goal: estimate the pdf $f(x)$.
- Define a set of bins $b_1, b_2, \ldots, b_M$ such that
  - $b_1 \leq x_1$ and $b_M \geq x_N$
- Then, we can define an estimator for the pdf $f(x)$ as

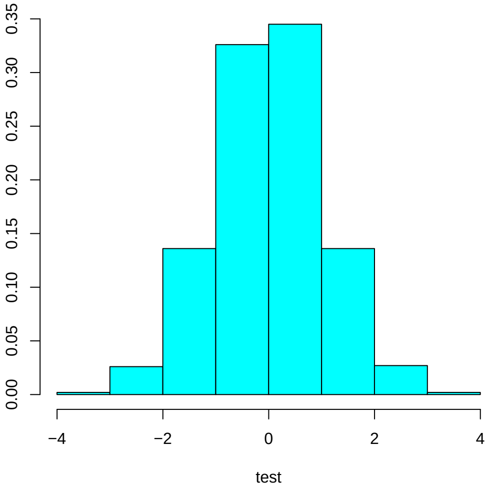
$$\widehat{f(x)} = \frac{\sum_{i=1}^{N} I(x_i > b_z \text{ and } x_i \leq b_{z+1})}{N} \times I(x \in (b_z, b_{z+1}])$$

In words: $\widehat{f(x)}$ is equal to proportion of observations in the bandwidth that x resides in

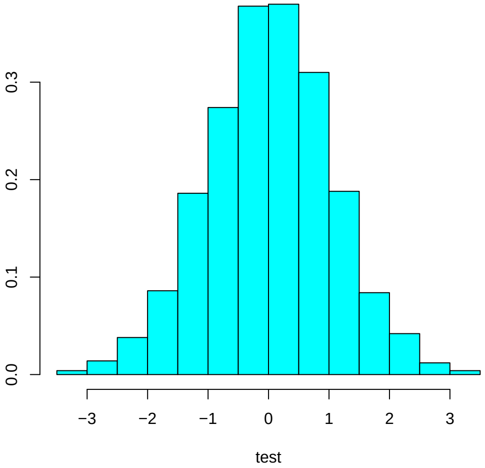
Bias Variance Tradeoff on Bin Size



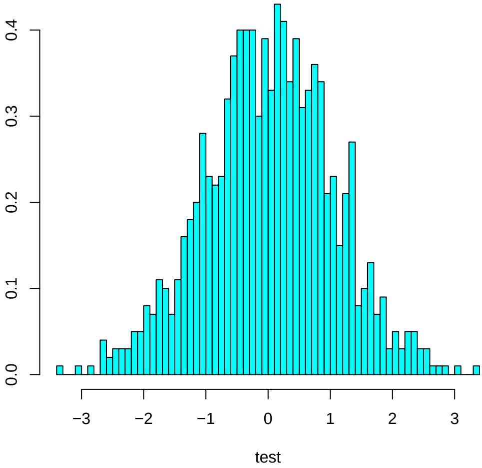
Bias Variance Tradeoff on Bin Size



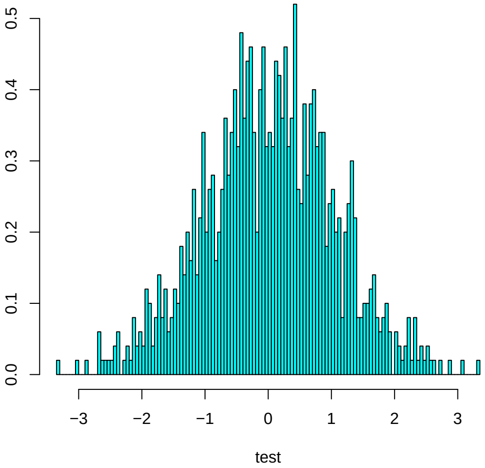
Bias Variance Tradeoff on Bin Size



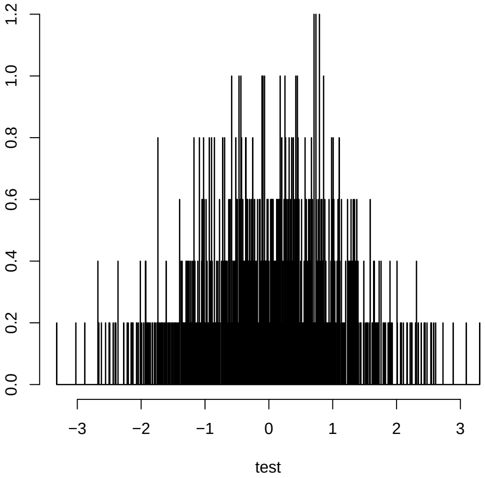
Bias Variance Tradeoff on Bin Size



Bias Variance Tradeoff on Bin Size



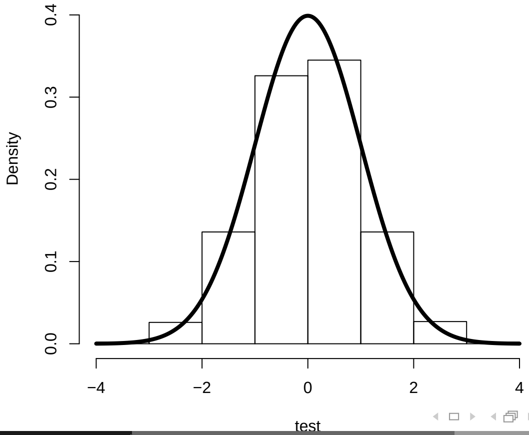
Bias Variance Tradeoff on Bin Size



Bias-Variance Tradeoff/Guidance

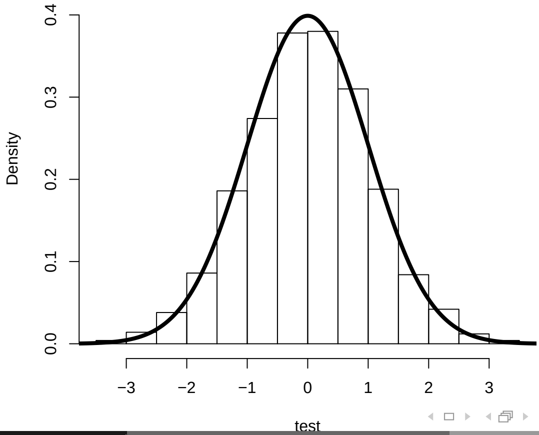
Classic problem:

Bias-Variance Tradeoff



Bias-Variance Tradeoff/Guidance

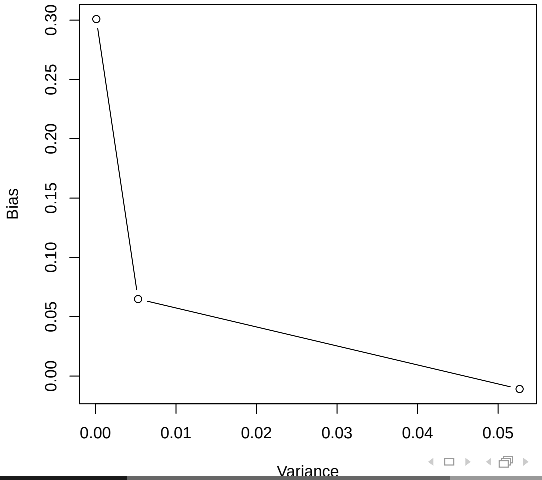
Classic problem:
Bias-Variance Tradeoff



Bias-Variance Tradeoff/Guidance

Classic problem:

Bias-Variance Tradeoff



Navigating the Bias Variance Tradeoff

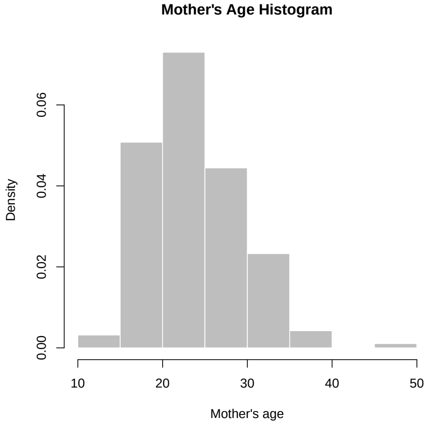
Two formulas for navigating bias-variance tradeoff for data x

1) width= $\frac{3.5 \times \mathsf{sd}(x)}{n^{1/3}}$ (Scott 1979)
2) width= $\frac{2(Q_3 - Q_1)}{n^{1/3}}$ (Freedman and Diaconis (1981))

Histogram Creation

< R Code >

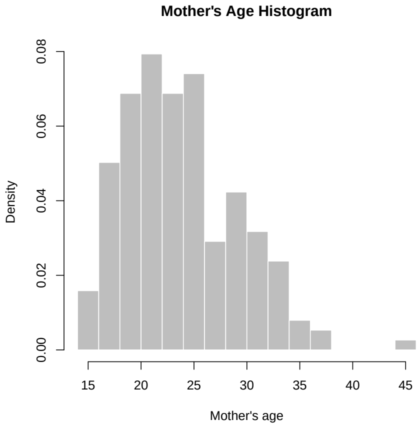
```
    truehist(birthwt$age, xlab='Mothers age',
ylab='Density', main='Mother's Age Histogram', col='gray',
border='white')
```



Histogram Creation

< R Code >

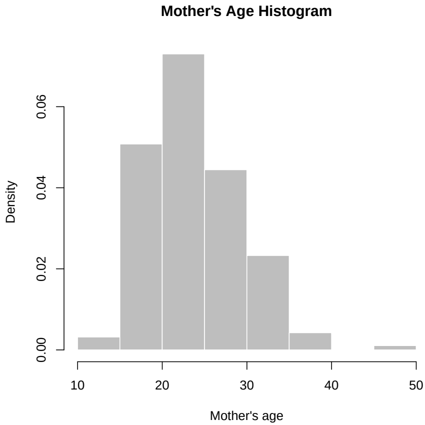
```
    truehist(birthwt$age, xlab='Mothers´ age',
ylab='Density', main='Mother's Age Histogram', col='gray',
border='white', nbins='fd')
```



Histogram Creation

< R Code >

```
truehist(birthwt$age, xlab='Mothers´ age',
ylab='Density', main='Mother's Age Histogram', col='gray',
border='white', nbins='scott')
```



Density Estimation

Histograms make continuous data discrete

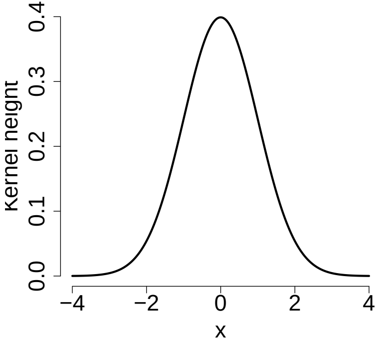
Density estimators "smooth" out histograms.

We will use Kernels again

Gaussian Kernel

Define

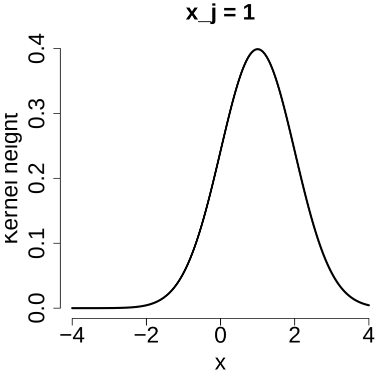
$$K(x) = \frac{\exp[-x^2]}{\sqrt{2\pi}}$$



Gaussian Kernel

Define

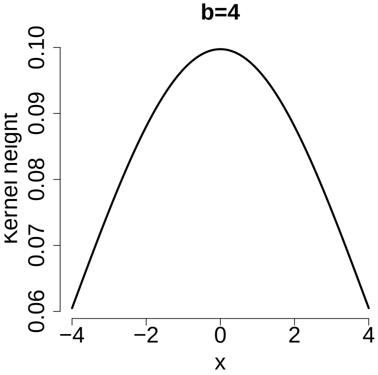
$$K\left(\frac{x - x_j}{b}\right) = \frac{\exp[-\left(\frac{x-x_j}{b}\right)^2]}{\sqrt{2\pi}}$$



Gaussian Kernel

Define

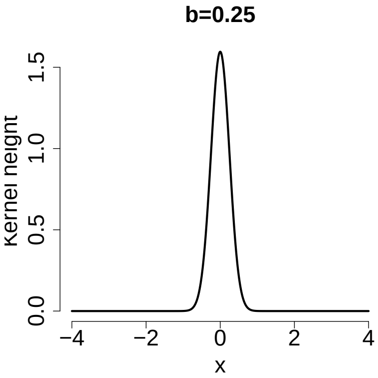
$$K(\frac{x - x_j}{b}) = \frac{\exp[-\left(\frac{x-x_j}{b}\right)^2]}{\sqrt{2\pi}}$$



Gaussian Kernel

Define

$$K\left(\frac{x - x_j}{b}\right) = \frac{\exp[-\left(\frac{x-x_j}{b}\right)^2]}{\sqrt{2\pi}}$$



Density Estimation

For all values $x \in \Re$ define the kernel density estimator as,

$$\widehat{f}(x) \quad = \quad \frac{1}{nb}\sum_{j=1}^{n} K(\frac{x - x_j}{b})$$

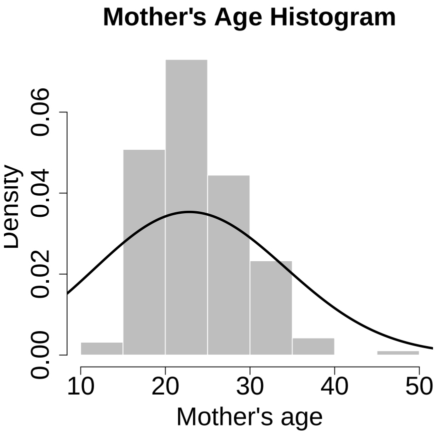
In words: at each point x, calculate weighted average of points, where weight is given by kernel

- As b increases, weight on nearby points is more evenly distributed (redistributed away from x)
- As b decreases, weight on nearby points is more concentrated (redistributed towards x)

Another bias variance tradeoff!

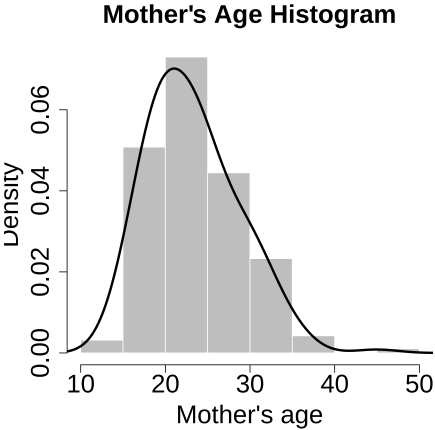
Density Estimation

```
truehist(birthwt$age, xlab='Mothers age',
ylab='Density', main='Mothers Age Histogram', col='gray',
border='white', nbins='scott')
lines(density(birthwt$age, width=40), lwd=3)
```



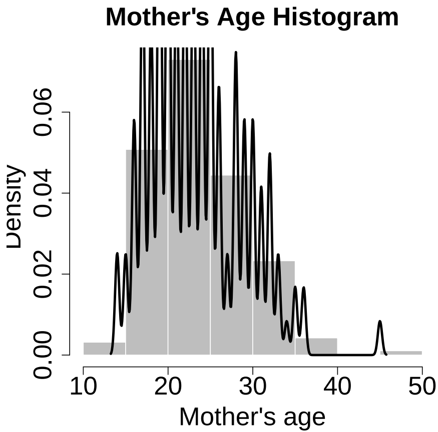
Density Estimation

```
truehist(birthwt$age, xlab='Mothers age',
ylab='Density', main='Mothers Age Histogram', col='gray',
border='white', nbins='scott')
lines(density(birthwt$age, width=10), lwd=3)
```



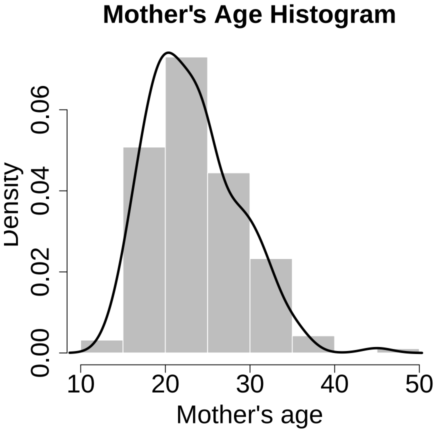
Density Estimation

```
truehist(birthwt$age, xlab='Mothers age',
ylab='Density', main='Mothers Age Histogram', col='gray',
border='white', nbins='scott')
lines(density(birthwt$age, width=1), lwd=3)
```



Density Estimation

```
   truehist(birthwt$age, xlab='Mothers age',
ylab='Density', main='Mothers Age Histogram', col='gray',
border='white', nbins='scott')
    lines(density(birthwt$age, width='SJ-dpi'), lwd=3)
```



Back to the Birth Weight

An initial inference:

Back to the Birth Weight

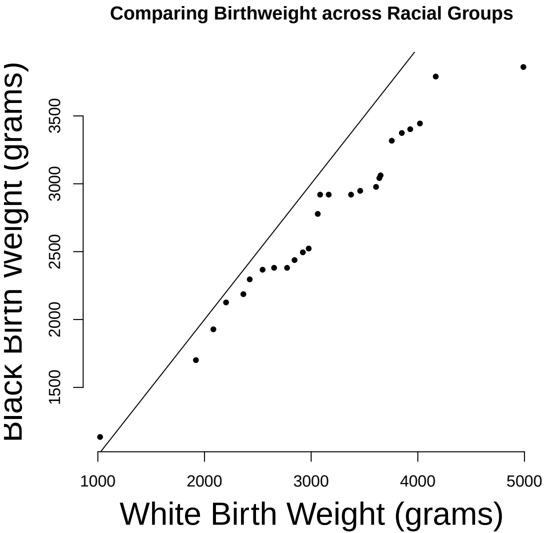
An initial inference:
Compare (contrast):

Back to the Birth Weight

An initial inference:
Compare (contrast):

$$\widehat{p}(y|\text{white}) \quad \text{and} \quad \widehat{p}(y|\text{black})$$

Back to the Birth Weight




```
qqplot(
bwt[which(race2=='white')],
bwt[which(race2=='black')],
xlab='White Birth Weight
(grams)', ylab='Black Birth
Weight (grams)',
frame.plot=F,
main='Comparing Birth
Weight across Racial Groups',
pch=20)
arrows(0,0, 1e7, 1e7)
```

Box Plot

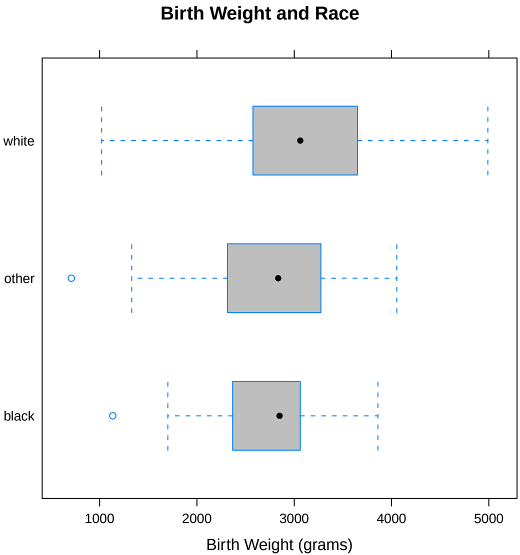
Box Plot

- QQ-plots compare two distributions directly

Box Plot

- QQ-plots compare two distributions directly
- Often want to compare many distributions (and readers are uncomfortable with qq-plots)

Box Plot



< R Code >

```
library(lattice)
bwplot(race2 ~bwt,
xlab='Birth Weight
(grams)',
frame.plot=F,main =
'Birthweight and Race',
panel=function(...){
panel.bwplot(...,
fill='gray', lty=1,
lwd=2, pch=20) } )
```

Violin Plot

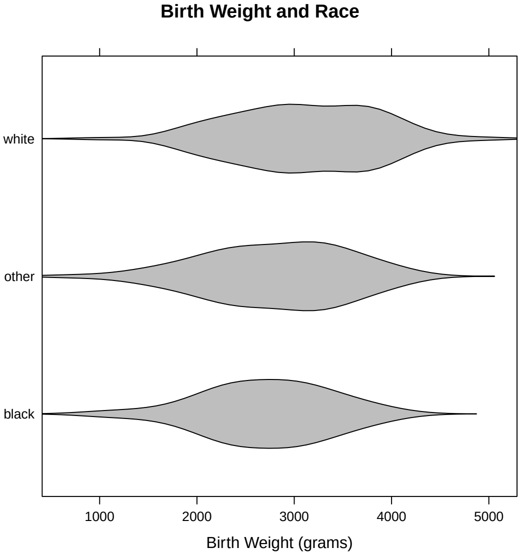
Violin Plot

- Box plots can obscure variation in distributions [histogram]

Violin Plot

- Box plots can obscure variation in distributions [histogram]
- Violin plots allow smooth estimates of data distribution [density]

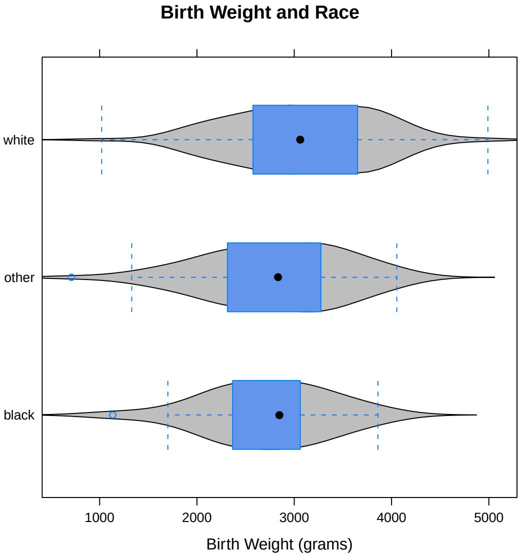
Violin Plot



< R Code >

```
library(lattice)
bwplot(race2~bwt,
xlab='Birth Weight
(grams)', frame.plot=F,
main='Birth Weight and
Race',
panel=function(...){
panel.violin(...,
col='gray',
bw='SJ-dpi') } )
```

Violin Plot



< R Code >

```
library(lattice)
bwplot(race2~bwt,
xlab='Birth Weight
(grams)', frame.plot=F,
main='Birth Weight and
Race',
panel=function(...){
panel.violin(...,
col='gray',
bw='SJ-dpi') } )
```

Box (Violin) Plot with Conditioning

- Comparing distributions within strata (buckets)

Box (Violin) Plot with Conditioning

- Comparing distributions within strata (buckets)
  - Here: compare distribution of birth weights across race, by smoker and non-smoker

Box (Violin) Plot with Conditioning

- Comparing distributions within strata (buckets)
  - Here: compare distribution of birth weights across race, by smoker and non-smoker
- Formally:

Box (Violin) Plot with Conditioning

- Comparing distributions within strata (buckets)
  - Here: compare distribution of birth weights across race, by smoker and non-smoker
- Formally:

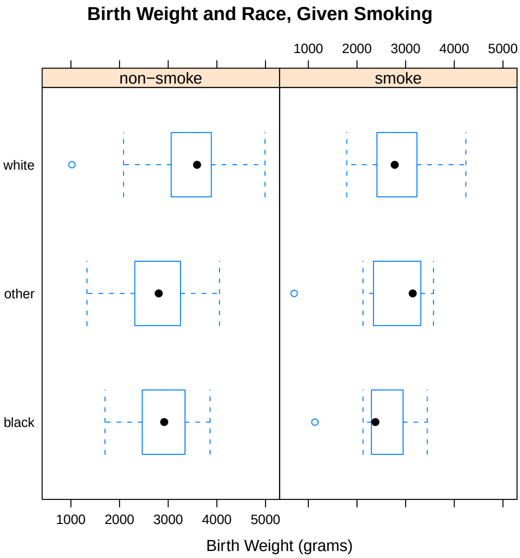
$$\widehat{p}(y|\text{black, smoke})$$
$$\widehat{p}(y|\text{white, smoke})$$
$$\widehat{p}(y|\text{other, smoke})$$

Box (Violin) Plot with Conditioning

- Comparing distributions within strata (buckets)
  - Here: compare distribution of birth weights across race, by smoker and non-smoker
- Formally:

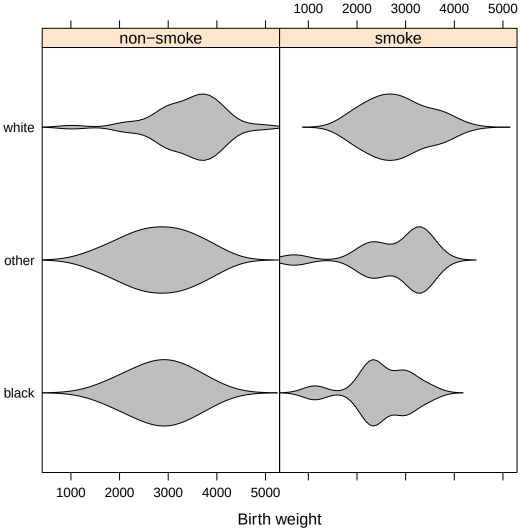
$$\begin{aligned}
\widehat{p}(y|\text{black, smoke}) \quad &; \quad \widehat{p}(y|\text{black, non-smoke}) \\
\widehat{p}(y|\text{white, smoke}) \quad &; \quad \widehat{p}(y|\text{white, non-smoke}) \\
\widehat{p}(y|\text{other, smoke}) \quad &; \quad \widehat{p}(y|\text{other, non-smoke})
\end{aligned}$$

Box (Violin) Plot with Conditioning




```
library(lattice)
bwplot(race2~bwt|smoke,
xlab='Birth Weight
(grams)',
frame.plot=F,main =
'Birth Weight and Race,
Given Smoking')
```

Box (Violin) Plot with Conditioning




```
 library(lattice)
bwplot(race2~bwt|smoke,
xlab='Birth weight',
panel=function(...){
panel.violin(...,
col='gray',bw='SJ-dpi')
} )
```

Nonparametrics: Bias/Variance Tradeoff