Nonlinear Regression (Part 1)

Christof Seiler

Stanford University, Spring 2016, STATS 205

Overview

- Smoothing or estimating curves
  - Density estimation
  - Nonlinear regression
- Rank-based linear regression

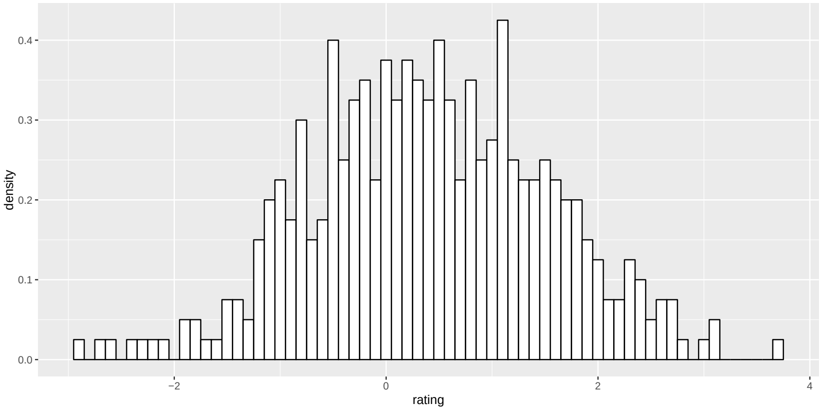
Curve Estimation

- A curve of interest can be a probability density function f
- In density estimation, we observe $X_1, \ldots, X_n$ from some unknown cdf F with density f

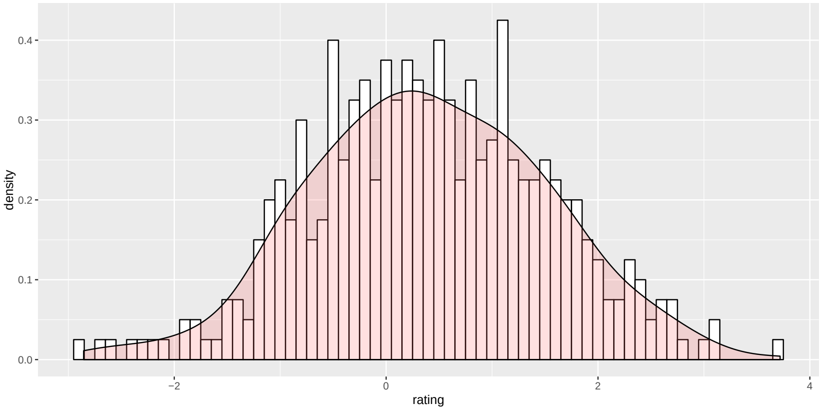
$$X_1, \ldots, X_n \sim f$$

- The goal is to estimate density f

Density Estimation



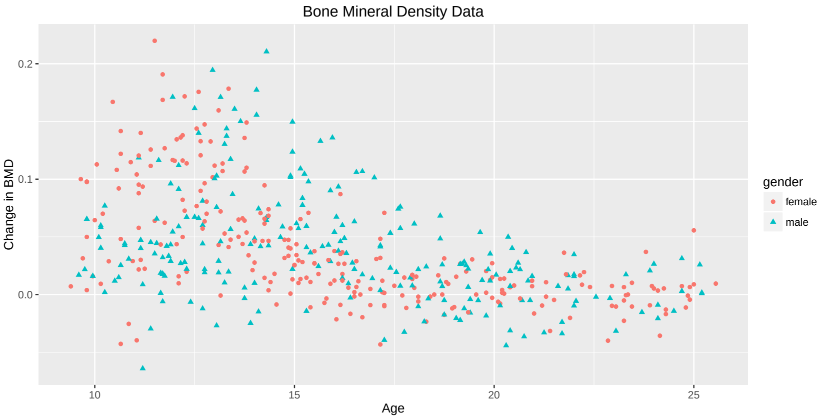
Density Estimation



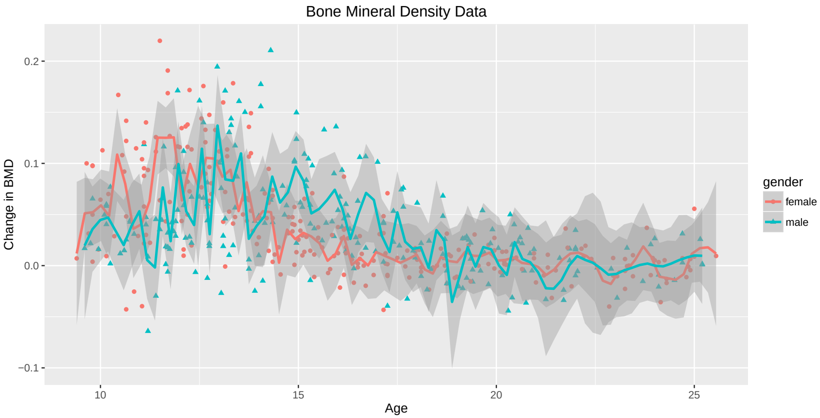
Nonlinear Regression

- A curve of interest can be a regression function r
- In regression, we observe pairs $(x_1, Y_1), \ldots, (x_n, Y_n)$ that are related as
$$Y_i = r(x_i) + \epsilon_i$$
with $E(\epsilon_i) = 0$
- The goal is to estimate the regression function r

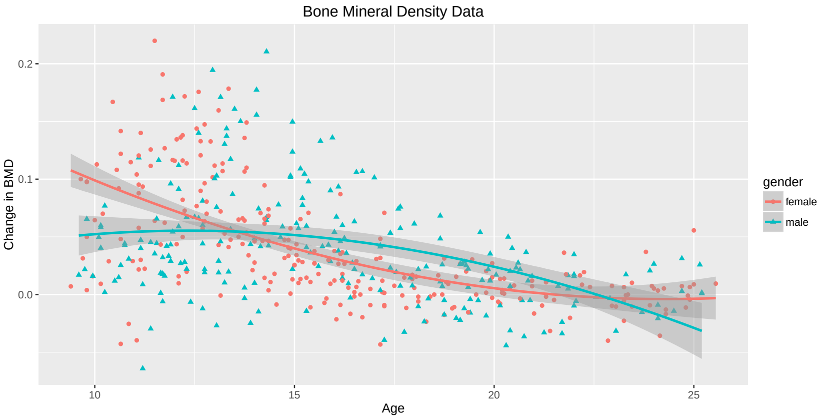
Nonlinear Regression



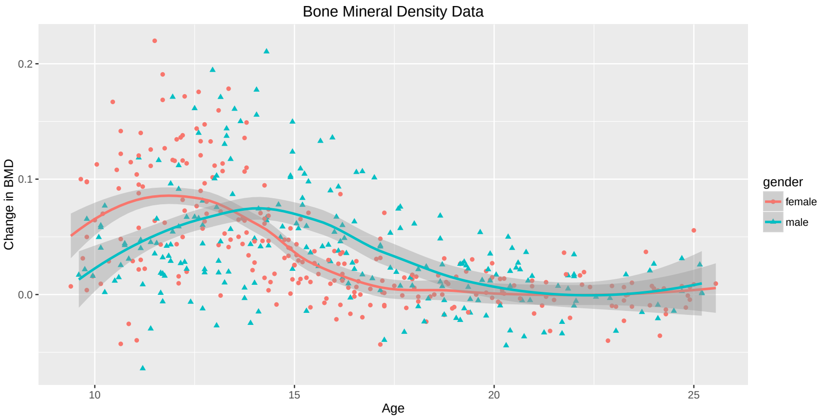
Nonlinear Regression



Nonlinear Regression



Nonlinear Regression



The Bias–Variance Tradeoff

- Let $\widehat{f}_n(x)$ be an estimate of a function $f(x)$
- Define the **squared error** loss function as

$$\text{Loss} = L(f(x), \widehat{f}_n(x)) = (f(x) - \widehat{f}_n(x))^2$$

- Define average of this loss as **risk** or **Mean Squared Error** (MSE)

$$\text{MSE} = R(f(x), \widehat{f}_n(x)) = \text{E}(\text{Loss})$$

- The expectation is taken with respect to $\widehat{f}_n$ which is random
- The MSE can be decomposed into a bias and variance term

$$\text{MSE} = \text{Bias}^2 + \text{Var}$$

- The decomposition is easy to show

The Bias–Variance Tradeoff

- Expand

$$\mathrm{E}((f-\widehat{f})^2) = \mathrm{E}(f^2 + \widehat{f}^2 + 2f\widehat{f}) = \mathrm{E}(f^2) + \mathrm{E}(\widehat{f}^2) - \mathrm{E}(2f\widehat{f})$$

- Use $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2$

$$\mathrm{E}((f-\widehat{f})^2) = \mathrm{Var}(f) + \mathrm{E}(f)^2 + \mathrm{Var}(\widehat{f}) + \mathrm{E}(\widehat{f})^2 - \mathrm{E}(2f\widehat{f})$$

- Use $\mathrm{E}(f) = f$ and $\mathrm{Var}(f) = 0$

$$\mathrm{E}((f-\widehat{f})^2) = f^2 + \mathrm{Var}(\widehat{f}) + \mathrm{E}(\widehat{f})^2 - 2f\,\mathrm{E}(\widehat{f})$$

- Use $(\mathrm{E}(\widehat{f}) - f)^2 = f^2 + \mathrm{E}(\widehat{f})^2 - 2f\,\mathrm{E}(\widehat{f})$

$$\mathrm{E}((f-\widehat{f})^2) = (\mathrm{E}(\widehat{f}) - f)^2 + \mathrm{Var}(\widehat{f}) = \mathrm{Bias}^2 + \mathrm{Var}$$

The Bias–Variance Tradeoff

- This described the risk at one point
- To summarize the risk, for density problems, we need to integrate

$$R(f, \widehat{f}_n) = \int R(f(x), \widehat{f}_n(x))dx$$

- For regression problems, we sum over all

$$R(r, \widehat{r}_n) = \sum_{i=1}^{n} R(r(x_i), \widehat{r}_n(x_i))$$

The Bias–Variance Tradeoff

- Consider the regression model

$$Y_i = r(x_i) + \epsilon_i$$

- Suppose we draw new observation $Y_i^* = r(x_i) + \epsilon_i^*$ for each x_i
- If we predict Y_i^* with $\widehat{r}_n(x_i)$ then the **squared prediction error** is

$$(Y_i^* - \widehat{r}_n(x_i))^2 = (r(x_i) + \epsilon_i^* - \widehat{r}_n(x_i))^2$$

- Define *predictive risk* as

$$\mathsf{E}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i^* - \widehat{r}_n(x_i))^2\right)$$

The Bias–Variance Tradeoff

- Up to a constant, the average risk and the predictive risk are the same

$$\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i^* - \widehat{r}_n(x_i))^2\right) = R(r, \widehat{r}_n) + \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}((\epsilon_i^*)^2)$$

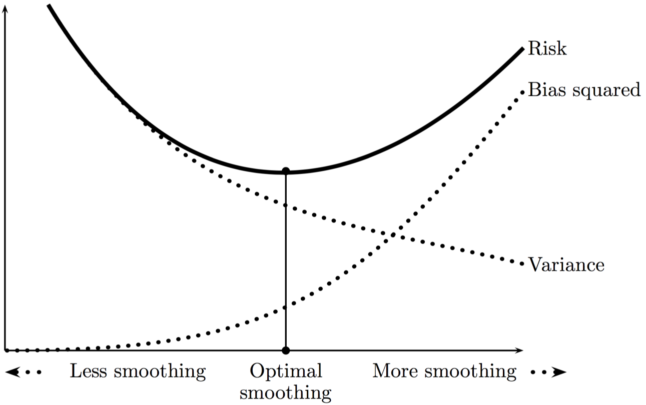
- and in particular, if error ϵ_i has variance σ^2, then

$$\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i^* - \widehat{r}_n(x_i))^2\right) = R(r, \widehat{r}_n) + \sigma^2$$

The Bias–Variance Tradeoff

- Challenge in smoothing is to determine how much smoothing to do
- When the data are oversmoothed, the bias term is large and the variance is small
- When the data are undersmoothed the opposite is true
- This is called the bias–variance tradeoff
- Minimizing risk corresponds to balancing bias and variance

The Bias–Variance Tradeoff



Source: Wassermann (2006)

The Bias–Variance Tradeoff (Example)

- Let f be a pdf
- Consider estimating $f(0)$
- Let h be a small and positive number
- Define

$$p_h := \mathrm{P}\left(-\frac{h}{2} < X < \frac{h}{2}\right) = \int_{-h/2}^{h/2} f(x)dx \approx hf(0)$$

- Hence

$$f(0) \approx \frac{p_h}{h}$$

The Bias–Variance Tradeoff (Example)

- Let X be the number of observations in the interval $(-h/2, h/2)$
- Then $X \sim \text{Binom}(n, p_h)$
- An estimate of p_h is $\widehat{p_h} = X/n$ and estimate of $f(0)$ is

$$\widehat{f}_n(0) = \frac{\widehat{p_h}}{h} = \frac{X}{nh}$$

- We now show that the MSE of $\widehat{f}_n(0)$ is (for some constants A and B)

$$\text{MSE} = Ah^4 + \frac{B}{nh} = \text{Bias}^2 + \text{Variance}$$

The Bias–Variance Tradeoff (Example)

- Taylor expand around 0

$$f(x) \approx f(0) + xf'(0) + \frac{x^2}{2} f''(0)$$

- Plugin

$$p_h = \int_{-h/2}^{h/2} f(x)dx \approx \int_{-h/2}^{h/2} \left(f(0) + xf'(0) + \frac{x^2}{2} f''(0) \right) dx$$

$$= hf(0) + \frac{f''(0)h^3}{24}$$

The Bias–Variance Tradeoff (Example)

- Since X is binomial, we have $\mathrm{E}(X) = np_h$
- Use Taylor approximation $p_h \approx hf(0) + \frac{f''(0)h^3}{24}$ and combine

$$\mathrm{E}(\widehat{f}_n(0)) = \frac{\mathrm{E}(X)}{nh} = \frac{p_h}{h} \approx f(0) + \frac{f''(0)h^2}{24}$$

- After rearranging, the bias is

$$\text{Bias} = \mathrm{E}(\widehat{f}_n(0)) - f(0) \approx \frac{f''(0)h^2}{24}$$

The Bias–Variance Tradeoff (Example)

- For the variance term, note that $\text{Var}(X) = np_h(1 - p_h)$, then

$$\text{Var}(\widehat{f}_n(0)) = \frac{\text{Var}(X)}{n^2 h^2} = \frac{p_h(1 - p_h)}{nh^2}$$

- Use $1 - p_h \approx 1$ since h is small

$$\text{Var}(\widehat{f}_n(0)) \approx \frac{p_h}{nh^2}$$

- Combine with Taylor expansion

$$\text{Var}(\widehat{f}_n(0)) \approx \frac{hf(0) + \frac{f''(0)h^3}{24}}{nh^2} = \frac{f(0)}{nh} + \frac{f''(0)h}{24n} \approx \frac{f(0)}{nh}$$

The Bias–Variance Tradeoff (Example)

- And combinding both terms

$$\text{MSE} = \text{Bias}^2 + \text{Var}(\widehat{f}_n(0)) = \frac{(f''(0))^2 h^4}{576} + \frac{f(0)}{nh} \equiv Ah^4 + \frac{B}{nh}$$

- As we smooth more (increase h),
  the bias term increases and the variance term decreases
- As we smooth less (decrease h),
  the bias term decreases and the variance term increases
- This is a typical bias–variance analysis

The Curse of Dimensionality

- Problem with smoothing is the curse of dimensionality in high dimensions
- Estimation gets harder as the dimensions of the observations increases
- **Computational:** Computational burden increases exponentially with dimension, and
- **Statistical:** If data have dimension d then we need sample size n to grow exponentially with d
- The MSE of any nonparametric estimator of a smooth curve has form (for $c > 0$)

$$\text{MSE} \approx \frac{c}{n^{4/(4+d)}}$$

- If we want to have a fixed $\text{MSE} = \delta$ equal to some small number δ, then solving for n

$$n \propto \left(\frac{c}{\delta}\right)^{d/4}$$

The Curse of Dimensionality

- We see that $n \propto \left(\frac{c}{\delta}\right)^{d/4}$ grows exponentially with dimension d
- The reason for this is that smoothing involves estimating a function in a local neighborhood
- But in high-dimensional problems the data are very sparse, so local neighborhood contain very few points

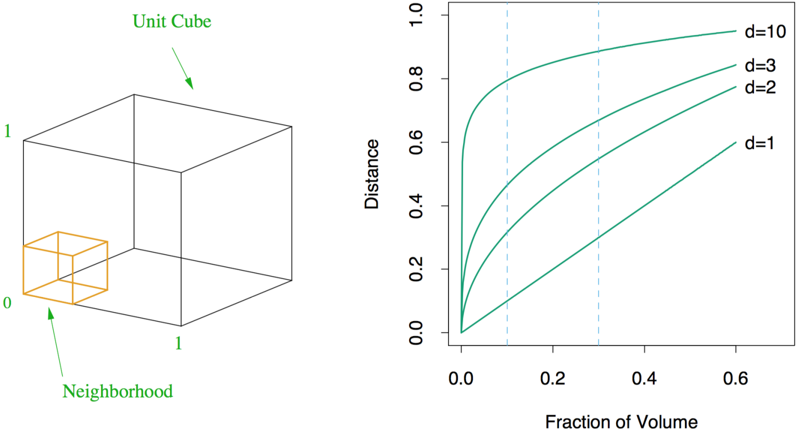
The Curse of Dimensionality (Example)

- Suppose n data points uniformly distributed on the interval $[0, 1]$
- How many data points will we find in the interval $[0, 0.1]$?
- The answer: about $n/10$ points
- Now suppose n point in 10 dimensional unit cube $[0, 1]^{10}$
- How many data points in $[0, 0.1]^{10}$?
- The answer: about

$$n \times \left(\frac{0.1}{1}\right)^{10} = \frac{n}{10,000,000,000}$$

- Thus, n has to be huge to ensure that small neighborhoods have any data in them
- Smoothing methods can in principle be used in high-dimensions problems
- But estimator won't be accurate, confidence interval around the estimate will be distressingly large

The Curse of Dimensionality (Example)



Source: Hastie, Tibshirani, and Friedman (2009)

- In ten dimensions 80% of range to capture 10% of the data

References

- Wassermann (2006). All of Nonparametric Statistics
- Hastie, Tibshirani, Friedman (2009). The Elements of Statistical Learning