



Statistical Learning

Lecture 5: Non-linear Models

Yi He

January 23, 2023

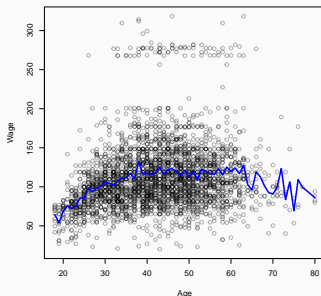
Plan for Today

1. Introduction
2. Polynomial and Step Regression
3. Spline Regression
4. Smoothing Splines
5. Extra Remarks

Introduction

Textbook Example: Wage Data Set

- Wage for males in the central Atlantic region of the US

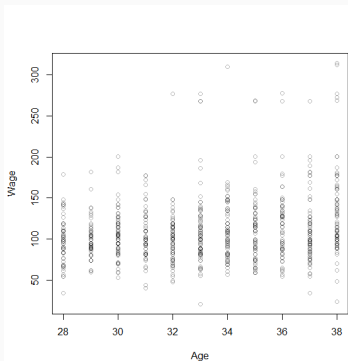


Average wage is **non-linear** and **non-smooth** in age

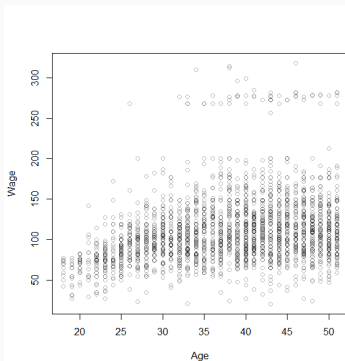
- Does it make sense to you that the *true* expected wage $\mathbb{E}[\text{wage}|\text{age}]$ to be non-smooth?

Nearest Neighbors

- The fraction $\alpha = k/n$ observations with the smallest *distance* to a given feature value x .



(a) $\alpha = 0.25$



(b) $\alpha = 0.75$

Nearest neighbors of Age=33

Local Linear Regression

- Target values y_i and **univariate** feature values $x_i \in \mathbb{R}$, $i = 1, \dots, n$
- Assign a weight $K(x_i, x)$ which is larger for x_i closer to x
- Fit a weighted least squares regression by minimizing

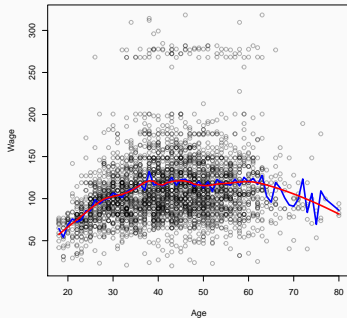
$$\sum_{i=1}^n K(x_i, x) (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i \in \mathcal{N}_\alpha(x)} K(x_i, x) (y_i - \beta_0 - \beta_1 x_i)^2,$$

where the equality holds when we assign zero weight $K(x_i, x) = 0$ for $i \notin \mathcal{N}_\alpha(x)$, and $\mathcal{N}_\alpha(x)$ denotes the index set of the nearest neighbors of x .

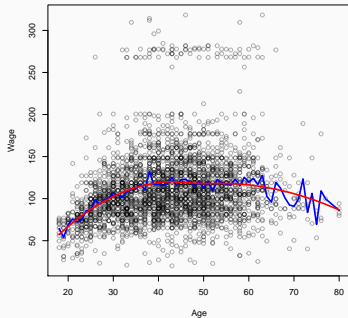
- The fitted value at x is given by $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$
- Both $\hat{\beta}_0 = \hat{\beta}_0(x)$ and $\hat{\beta}_1 = \hat{\beta}_1(x)$ depend on x .

Wage Data Set: Local Linear Regression

- Red line: fitted value with $\alpha = 0.25, 0.75$
- Estimated regression function gets 'smoother' for a larger α : bias-variance tradeoff



(a) $\alpha = 0.25$



(b) $\alpha = 0.75$

Polynomial and Step Regression

Let us begin with the case for uni-variate feature $X \in \mathbb{R}$. A smooth regression function $\mu(x) = \mathbb{E}[Y|X = x]$ often admits the expansion

$$\mu(x) = \underbrace{\beta_0 b_0(x) + \beta_1 b_1(x) + \dots + \beta_K b_K(x)}_{\text{To be fitted}} + \underbrace{\beta_{K+1} b_{K+1}(x) + \dots}_{\text{bias}}$$

where

- $\{b_i\}$ are *fixed, known* basis functions, often $b_0(x) = 1$
- Coefficients $\{\beta_i\}$'s do not depend on x
- the 'dimension' of the problem is represented by the number of basis functions required for a sufficiently good approximation

Linear Regression on Transformed Data

- Suppose we use the approximation

$$\mu(x) \approx \beta_0 \underbrace{b_0(x)}_{\text{known}} + \beta_1 \underbrace{b_1(x)}_{\text{known}} + \dots + \beta_K \underbrace{b_K(x)}_{\text{known}} =: f(x),$$

where K is a hyperparameter.

- Least-squares regression for y_i on transformed features $b_0(x_i), b_1(x_i), \dots, b_K(x_i)$
- Larger K , smaller bias but larger variance.

Naive Examples of the Basis functions

Polynomial functions

(Auto Example in Lecture 2)

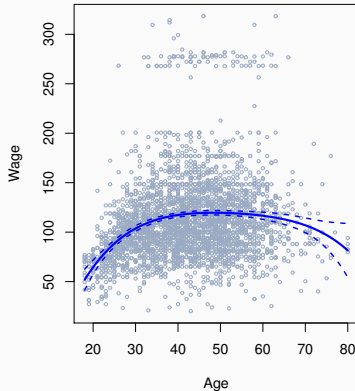
- Taylor expansion: $b_i(x) = x^i$
- Not suitable for dummy variables: $0^i = 0$ and $1^i = 1$
- Sensitive to outliers

Step functions

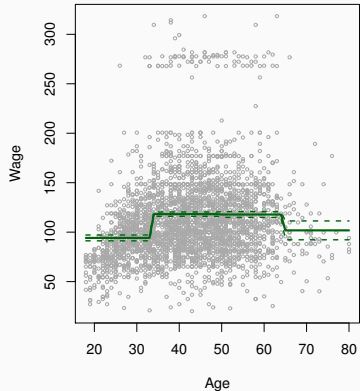
- Cutpoints c_1, \dots, c_K in the range of X
- $K + 1$ interval indicators:
 $b_0(x) = \mathbb{1}(x < c_1), b_1(x) = \mathbb{1}(c_1 \leq x < c_2), \dots, b_{K-1}(x) = \mathbb{1}(c_{K-1} \leq x < c_K), b_K(x) = \mathbb{1}(x \geq c_K)$
- fitted function is discontinuous at cutpoints

Polynomial and Step Regression

- May choose K using cross-validation
- Cutpoints are usually specified based on interpretive convenience in practice



(c) Polynomial regression $K = 4$



(d) Step regression $K = 2$

Spline Regression

Cubic Spline

- Knots (cutpoints): $\xi_1 < \xi_2 < \dots < \xi_M$
- A cubic polynomial in any sub interval $[\xi_j, \xi_{j+1}]$

$$f(x) = \delta_{0,j} + \delta_{1,j}x + \delta_{2,j}x^2 + \delta_{3,j}x^3, \quad j = 0, \dots, M,$$

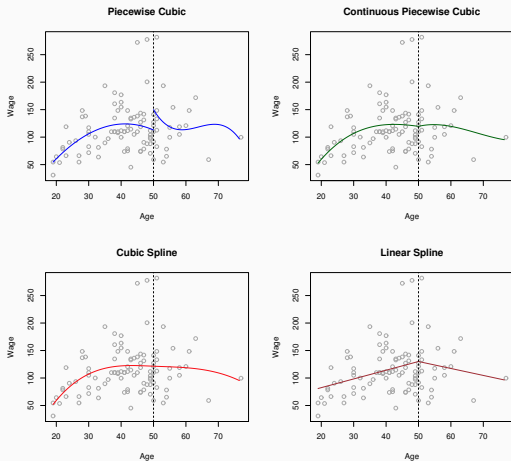
where $\xi_0 = \min(X)$ and $\xi_{M+1} = \max(X)$

- Twice continuously differentiable (**at knots**):

$$\lim_{x \uparrow \xi_j} f(x) = \lim_{x \downarrow \xi_j} f(x), \quad \lim_{x \uparrow \xi_j} f'(x) = \lim_{x \downarrow \xi_j} f'(x), \quad \lim_{x \uparrow \xi_j} f''(x) = \lim_{x \downarrow \xi_j} f''(x)$$

- $4 + M(4 - 3) = 4 + M$ free coefficients: 4 in $[\xi_0, \xi_1]$, and add $4 - 3 = 1$ at each knot

Constraints at Knots: Illustration



ISLR Figure 7.3

Go to www.menti.com and use the code 4850 9341



Does it make sense to you that the true regression function is smooth?

0

Yes, it is hard to explain why small change in x gives large change in y

0

No, there could be significant changes at particular feature values

0

Depends on the problems



Regression Splines

A cubic spline with knots ξ_1, \dots, ξ_M can be represented by a basis function

$$f(x) = \beta_0 + \sum_{i=1}^{M+3} b_i(x)\beta_i$$

where

- b_i are the B-spline basis functions (not to be discussed), depending on the knots
- $M + 3$ features, excluding the intercept
- generate b_i 's using `bs(...,df=M+3)` function in `splines` package
- Linear regression: y_i on $b_1(x_i), b_2(x_i), \dots, b_{M+3}(x_i)$, including an intercept

Natural Boundary Conditions

As it is easy to overfit around/beyond data boundary, we often impose that the so-called *natural boundary conditions*:

$f(x)$ is a linear function when $x < \xi_0$ or $x \geq \xi_{M+1}$.

That is,

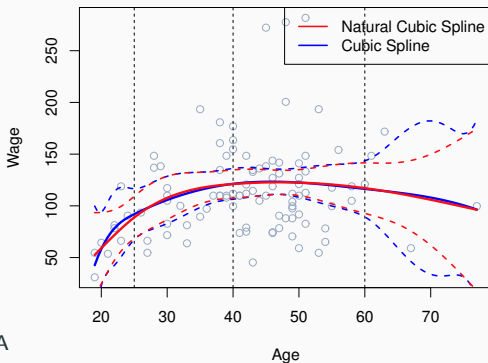
$$f(x) = \begin{cases} \delta_{0,0} + \delta_{1,0}x + \cancel{\delta_{2,0}x^2} + \cancel{\delta_{3,0}x^3} & -\infty < x < \xi_0 \\ \delta_{0,1} + \delta_{1,1}x + \delta_{2,1}x^2 + \delta_{3,1}x^3 & \xi_0 \leq x < \xi_1 \\ \delta_{0,2} + \delta_{1,2}x + \delta_{2,2}x^2 + \delta_{3,2}x^3 & \xi_1 \leq x < \xi_2 \\ \vdots & \\ \delta_{0,M+2} + \delta_{1,M+2}x + \cancel{\delta_{2,M+2}x^2} + \cancel{\delta_{3,M+2}x^3} & \xi_{M+1} \leq x < \infty \end{cases}$$

- $4 + (M + 2) - 2 \times 2 = M + 2$ free coefficients
- $M + 1$ degrees of freedom, excluding an intercept

Natural Cubic Splines

R Package: `splines`

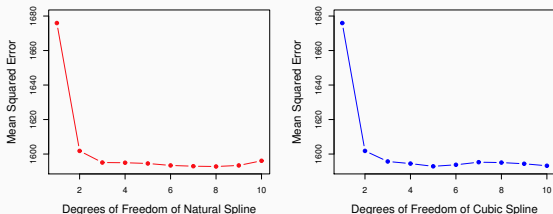
- `bs(x, ...)` for any degree splines
- `ns(x, ...)` for natural cubic splines: boundary knots are the smallest and largest observations



ISLR Figure 7.4

Knot Placement

- Knots may be specified based on convenient interpretation in real-life problems
- Otherwise, we often place the knots at appropriate quantiles of the observed feature values.
- To decide the number of knots, one may use cross validation.



ISLR Figure 7.6: the x -axis is the number of basis functions (K)

Smoothing Splines

Roughness

Consider any twice continuously differentiable function $g(x)$.

We measure roughness (non-linearity) of the function $g(x)$ at point x through the magnitude of its second derivative $g''(x)$.

No roughness $g''(x) \equiv 0$ everywhere implies that g is linear.

The overall roughness over the domain D is given by

$$\int_D (g''(t))^2 dt$$

Following our textbook, from now on we omit the domain D and write this definite integral in short as

$$\int (g''(t))^2 dt.$$

Penalized Least Squares

Like in Lecture 3, imposing an inequality constraint $\int (g''(t))^2 dt \leq b$ is equivalent to solving the unconstrained optimization problem:

$$\underset{g \in \mathcal{C}^2}{\text{minimize}} \sum_{i=1}^n (y_i - g(x_i))^2 + \underbrace{\lambda \int (g''(t))^2 dt}_{\text{penalty}}, \quad \lambda > 0,$$

where \mathcal{C}^2 denotes the class of twice continuously differentiable functions.

- $g(x_i) = \text{Average}\{y_j : x_j = x_i\}$ if $\lambda = 0$: non-smoothed solution
- $g(x) \rightarrow$ linear least-squares estimates, as $\lambda \rightarrow \infty$
- λ controls the smoothness of the estimated function

Optional Material: Finding the Smoothing Splines

- Suppose $n \geq 2$ and $a < \xi_1 < \dots < \xi_n < b$. Given any values z_1, \dots, z_n , there is a unique natural cubic spline \hat{g} with knots at the points ξ_n satisfying $\hat{g}(\xi_n) = z_i$.
- For any twice continuously differentiable function g on $[a, b]$ with $g(\xi_i) = z_i$ for all $i = 1, \dots, n$,

$$\int (g''(t))^2 dt \geq \int (\hat{g}''(t))^2 dt$$

\Rightarrow the solution must be a natural cubic spline, with knots at x_1, \dots, x_n !

Otherwise we could choose a natural spline with same values $g(x_i)$ but smaller penalty, a contradiction.

Ridge Regression

Consider only the *natural spline* g with knots at x_1, \dots, x_n .

- It can be shown that

$$\int (g''(t))^2 dt = \mathbf{g}^T W \mathbf{g}, \quad \mathbf{g} = (g(x_1), \dots, g(x_n))^T.$$

where W is a closed-form symmetric positive-definite matrix.

The optimization problem becomes

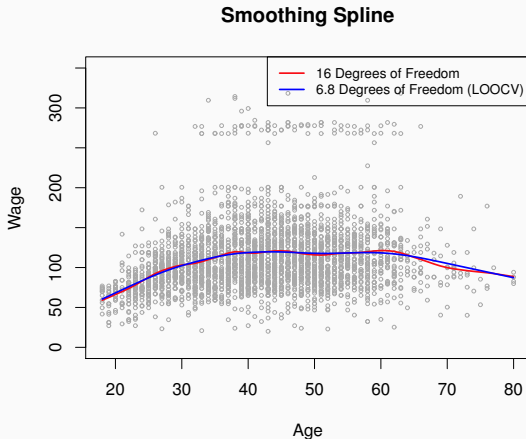
$$\text{minimize} \quad \sum_{i=1}^n (y_i - g(x_i))^2 + \underbrace{\lambda \mathbf{g}^T W \mathbf{g}}_{\text{quadratic penalty}}.$$

Tute Q2: the solution is

$$\mathbf{g} = (I + \lambda W)^{-1} \mathbf{y}, \quad \mathbf{y} = (y_1, \dots, y_n)^T.$$

Finally, we solve the (unique) natural spline function g .

- Degrees of freedom $df = \text{tr}(I + \lambda W)^{-1} \in (0, n]$
- The corresponding λ can be calculated
- One may choose λ by using cross-validation in practice



Extra Remarks

For multivariate feature $X = (X_1, \dots, X_p)^T$,

- The local linear regression can still work for small p
- Measure the distance via vector length $\|x_i - x_j\|$
- For (very) large p , the Euclidean norm becomes a meaningless measure (not to be shown) and the local regression performs badly in general: beyond our course

Additive Models

For multivariate feature $X = (X_1, \dots, X_p)^T$, we may consider an *additive* model:

$$f(X_1, \dots, X_p) = f_1(X_1) + \dots + f_p(X_p).$$

For example,

$$\text{wage} = \beta_0 + f_1(\text{year}) + f_2(\text{age}) + f_3(\text{education}) + \epsilon$$

- restrictive model: bias-variance tradeoff
- expand unknown $f_i(x) = \sum_{k=1}^{K_i} b_k^{(i)}(x)\beta_{k,i}$
- ISLR Section 7.7.1: beyond our course