Today

1. Introduction to Chapter 7
2. Polynomial regressions
3. Step Functions
4. Regression splines

Announcement: Regression Tutorial from 1:00 - 3:00 in Dunn 304

1. Introduction to chapter 7
   - The truth is never linear!
   - But often the linearity assumption is good enough
   - When it is not
     - polynomials
     - step functions
     - splines
     - local regression, and
     - generalized additive models

   offer a lot of flexibility, without losing the ease and interpretability of linear models

2. polynomial regression

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \varepsilon_i$$

Degree 4 polynomial

- create new variables $X_1 = X$, $X_2 = X^2$, etc and then treat as multiple linear regression
- Not really interested in the coefficients; more interested in the fitted function values at any value x_0:

$$\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + \hat{\beta}_3 x_0^3 + \hat{\beta}_4 x_0^4$$

- Since $\hat{f}(x_0)$ is a linear function of the $\hat{\beta}_\ell$, we can get a simple expression for pointwise-variances $Var[\hat{f}(x_0)]$ at any value x_0

$\Rightarrow$ 95% CI for $\hat{f}(x_0)$ is

$$\hat{f}(x_0) \pm 2 \cdot se[\hat{f}(x_0)]$$

$\Uparrow$ the estimated pointwise standard error of $\hat{f}(x_0)$

Ex: Figure 7.1 "wage"

$\begin{cases} \text{high earners} & (>\$250,000) \\ \text{low earners} & (<\$250,000) \end{cases}$

- logistic Regression follows naturally. For instance, for figure 1, we model

$$Pr(y_i > 250,000 \mid x_i) = \frac{\exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_d x_i^d)}{1 + \exp(\beta_0 + \beta_1 x_i + \cdots + \beta_d x_i^d)}$$

Note: ① To get confidence intervals, compute upper and lower bounds on the logit scale, and then invert to get on probability scale.

② Can fit use $y \sim poly(x, degree = 3)$ in formula

3. Step Functions

Another way of creating transformations of a variable — cut the variable into distinct regions

$C_1(X) = I(X < 35)$, $C_2(X) = I(35 \le X < 50)$ ... $C_3 = I(50 \le X < 65)$

$C_4(X) = I(X \ge 65)$

Notes: ① Here $I(\cdot)$ is an indicator function that returns a 1 if the condition is true, and return 0 otherwise.

② For any value of X, $C_1(X) + \cdots + C_4(X) = 1$

③ $C_i(X)$'s are called dummy variables

Then use least squares to fit a linear model using $C_1(X) \cdots C_4(X)$ as predictors

General case

$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \cdots + \beta_K C_K(x_i) + \varepsilon_i$$

where β_0 can be interpreted as the mean value of Y for $X < c_1$

$\beta_0 + \beta_j = \beta_j$ represents the average increase in the response for X in $c_j < X < c_{j+1}$ relative $X < c_1$

($c_1 \cdots c_K$ are the cutpoints).

Advantages:

- Easy to work with. Creates a series of dummy variables representing each group
- useful way of creating interactions that are easy to interpret. For example, interaction effect of "Year" and "Age"

  I(Year < 2005) · age   I(Year ≥ 2005) · age

  would allow for different linear function in each age category.
- In R: I(year < 2005) or cut(age, c(18, 25, 40, 65, 90))

Disadvantage

- Choice of cutpoints can be problematic. For creating nonlinearities, smoother alternatives such as splines are available.

4. Regression splines

- a flexible class of basis functions that extends upon the polynomial regression and piecewise constant regression approaches

- Instead of a single polynomial in X over its whole domain, we can rather use different polynomials in regions defined by knots

$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \beta_{31} x_i^3 + \varepsilon_i & \text{if } x_i < c_j \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \varepsilon_i & \text{if } x_i > c_j \end{cases}$$

- Better to add constraints to the polynomials, eg. continuity

Eg: Figure 7.3

three constraints

$\begin{cases} \text{continuity of the fitted curve} \\ \text{continuity of the first derivatives} \\ \text{continuity of the second derivatives} \end{cases}$

⇒ Each constraint frees up one degree of freedom by reducing the complexity of the resulting piecewise polynomial fit.

⇒ Cubic spline (for K knots, use in total 4 + K degrees of freedom)

$$4(K+1) - 3K = K + 4$$

⇒ linear spline : fitting a line in each region of the predictor space defined by the knots, requiring continuity at each knot.

* The spline basis representation

① A linear spline with knots at ξ_k, $k = 1, \ldots K$ is a piecewise linear polynomial continuous at each knot.

⇒ we can represent this model as

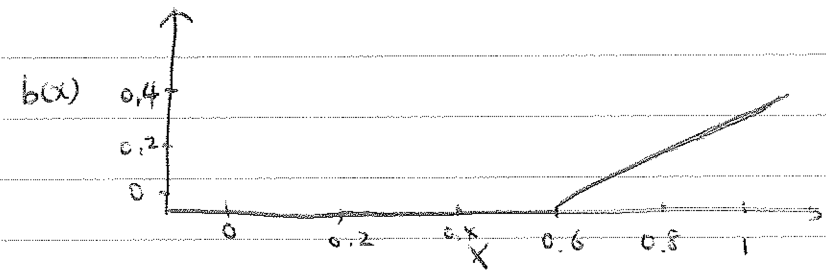
$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_K b_K(x_i) + \beta_{K+1} b_{K+1}(x_i) \boxed{+ \cdots + \beta_{K+K} b_{K+K}(x_i)} + \varepsilon_i$$

where b_k's are basis functions

$$\left.\begin{array}{l} b_1(x_i) = x_i \\ b_{k+1}(x_i) = (x_i - \xi_k)_+, \quad k = 1, \ldots K \end{array}\right\} \begin{array}{l} x_i,\ (x_i - \xi_1)_+ \\ (x_i - \xi_2)_+ \cdots + (x_i - \xi_K)_+ \end{array}$$

(Note : here $(\ )_+$ mean positive part, i.e

$$(x_i - \xi_k)_+ = \begin{cases} x_i - \xi_k & \text{if } x_i > \xi_k \\ 0 & \text{otherwise} \end{cases}$$)



② A cubic spline with knots at ξ_k, $k = 1, \ldots K$ is a piecewise cubic polynomial with continuous derivatives up to order 2 at each knot.

⇒ we can represent this model with truncated power basis functions

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+3} b_{K+3}(x_i) + \varepsilon_i$$

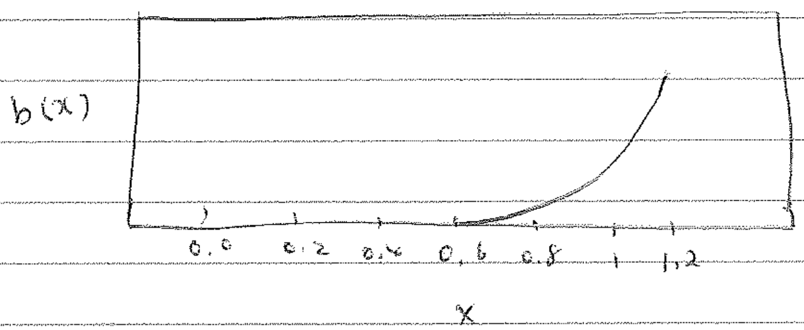
$$b_1(x_i) = x_i$$

$$b_2(x_i) = x_i^2$$

$$b_3(x_i) = x_i^3$$

$$b_{k+3}(x_i) = (x_i - \xi_k)_+^3 \quad k=1, \cdots K$$

where $(x_i - \xi_k)_+^3 = \begin{cases} (x_i - \xi_k)^3 & \text{if } x_i > \xi_k \\ 0 & \text{otherwise} \end{cases}$



③ A natural Cubic Spline.

- A natural Cubic Spline extrapolates linearly beyond the boundary knots
- This adds $4 = 2 \times 2$ extra constraints
- Allow us to put more internal knots for the same degrees of freedom as a regular cubic spline

* Knot placement

① Where should we place the knots?

② How many knots should we use?

⇒ One strategy is to decide K, the number of knots, and then place them at appropriate quantiles of the observed X.

Answer to ②

- use Cross-validation to choose the value of K giving the smallest RSS.

Answer to ①

- Specify the degrees of freedom, and then place the knots at uniform quantiles of the data.