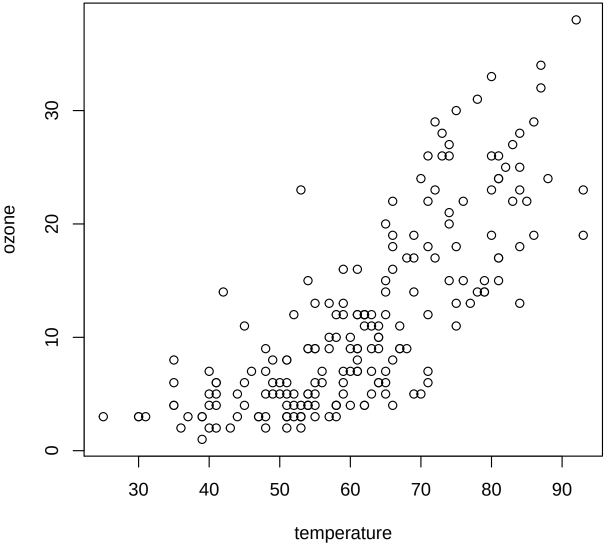
STATS 769

Moving Beyond Linearity

Yong Wang
Department of Statistics
The University of Auckland

2021

Nonlinear Relationship



Modelling Nonlinear Relationship

- When the relationship between and a predictor X and the response Y is apparently nonlinear, there are several approaches to handling it.
- With increasing sophistication:
  - Polynomial regression
  - Step functions
  - Regression splines
  - Smoothing splines
  - Local regression
  - Generalised additive models (GAM)
- The first five are typically used when X is univariate (one-dimensional), although extensions to the multivariate case are possible.
- The GAM is specifically designed to deal with a multivariate X.

Polynomial Regression

- Instead of using the simple linear regression model

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

we can include higher-order terms of X in the model, i.e.,

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_d X^d + \epsilon.$$

- This is known as polynomial regression.
- Although it describes a nonlinear relationship between X and Y, this is still a linear regression model.
- Good to use for d not large: Nonlinear and smooth.
- Has numerical problems for a large d (> 3, say).

Step Functions

- Denote the indicator function by

$$I(Z) = \begin{cases} 1, & \text{if } Z = \texttt{TRUE}; \\ 0, & \text{if } Z = \texttt{FALSE}. \end{cases}$$

- We partition the range of X into intervals at cut points $c_1, c_2, \dots, c_m$.
- Then we can use a step function to model the relationship between Y and X, as follows:

$$\begin{aligned} Y = {} & \alpha_0 I(X < c_1) + \alpha_1 I(c_1 \le X < c_2) + \cdots \\ & + \alpha_{p-1} I(c_{m-1} \le X < c_m) + \alpha_m I(X \ge c_m) + \epsilon. \end{aligned}$$

- Again, this is a linear regression model (without the intercept).
- The least square estimate α_j ($j = 0, \dots, m$) is simply the mean of y_i's for those x_i in the jth interval.

Step Functions II

- Fully flexible and adaptive
- Performance affected by the number of cutpoints and their locations.
- Not continuous, let alone "smooth" (not differentiable everywhere).
- Prediction accuracy not very high.
- Another representation:

$$\begin{aligned} Y &= \beta_0 + \beta_1 I(X \geq c_1) + \cdots + \beta_m I(X \geq c_m) + \epsilon \\ &= \beta_0 + \sum_{j=1}^{m} \beta_j I(X \geq c_j) + \epsilon, \end{aligned} \tag{1}$$

where $\beta_0 = \alpha_0$ and $\beta_j = \alpha_j - \alpha_{j-1}$ $(j = 1, \ldots, m)$ (why?).

Regression Splines

- Regression splines can be considered as extensions to step functions.
- We want to introduce higher-order terms into (1).
- Hence we replace β_0 with $\alpha_0 + \alpha_1 X + \cdots + \alpha_d X^d$, and each $I(X \geq c_j)$ with

$$(X - c_j)_+^d = \begin{cases} (X - c_j)^d, & \text{if } X \geq c_j; \\ 0, & \text{if } X < c_j. \end{cases}$$

- The dth-degree regression spline is thus given by

$$Y = \sum_{j=0}^{d} \alpha_j X^d + \sum_{j=1}^{m} \beta_j (X - c_j)_+^d.$$

- The step function is in fact the 0th-degree regression spline.
- It is still a linear regression model.

Regression Splines II

- A dth-degree regression spline is continuous (if $d \geq 1$).
- It has a continuous $(d-1)$th-order derivative (if $d \geq 2$) and hence is smooth (of order $d-1$).
- Large d can cause numerical problems.
- It is quite common to use cubic (third-degree) regression splines in practice.
- Fully flexible and adaptive.
- Good prediction accuracy, if cut points (better known as knots) are well chosen.

Smoothing Splines

- Smoothing splines take an approach that penalises the RSS.
- In particular, it minimises

$$\mathrm{RSS}^{\mathrm{sspline}}(f;\lambda) = \sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \int [f''(x)]^2 \,\mathrm{d}x$$

  over all possible functions f, where $\lambda \geq 0$ is a tuning parameter.
- This is a regularisation method.
- Theory shows that the solution $\widehat{f}$:
  1. is a piecewise cubic polynomial with knots at the unique values of $x_1, \ldots, x_n$ inside the region between the two extrema (minimum and maximum) of the x_i's.
  2. is linear outside of this region;
  3. has continuous first and second derivatives at the knots (and everywhere).
- This is known as a natural cubic spline.

Smoothing Splines II

- A smoothing spline does not need to choose knots.
  - All unique values of $x_1, \ldots, x_n$ are taken as knots.
- It is thus a memory-based method — all unique x_i-values must be remembered for prediction.
- It needs to choose an appropriate value for λ. Some good methods exist for this purpose.
- The smoothness of the spline is controlled by the value of λ, which corresponds uniquely to a value of the effective degree of freedom df_λ.
- Fully flexible and adaptive.
- Good prediction accuracy.

Local Regression

- Local regression computes the fit at each target point x_0, by assigning different weights to observations:
  - Higher weights for those close to x_0.
  - Lower or 0 weights for those far away.
- Denote the weight function by $w(x_i; x_0)$.
  - For example, using a truncated normal density with mean x_0
  - User can specify span for the fraction of training points closest to x_0 to have positive weights, which also helps provide a standard deviation (for the normal density).
- Minimise

$$\sum_{i=1}^{n} w(x_i; x_0)(y_i - \beta_0 - \beta_1 x_i)^2$$

  to find $\widehat{\beta}_0$ and $\widehat{\beta}_1$.
- The prediction is just

$$\widehat{f}(x_0) = \widehat{\beta}_0 + \widehat{\beta}_1 x_0.$$

- Repeat the above process for another x_0.

Local Regression II

- It is also a memory-based method — all observations must be remembered for prediction.
- Good prediction accuracy.
- One can further replace the simple linear regression model with a (higher-order) polynomial to improve prediction accuracy.
- Computationally costly for making many predictions

Generalised Additive Models

- A GAM is an extension to multiple linear regression and can be written as

$$Y = \beta_0 + f_1(X_1) + \cdots + f_p(X_p) + \epsilon,$$

  where f_j can be a nonlinear function of X_j, e.g., a smoothing spline.
- To fit a GAM to the data is to find $\widehat{\beta}_0, \widehat{f}_1, \ldots, \widehat{f}_p$ that minimise the RSS.
- With GAMS, transforming a variable does not need to be conducted beforehand and is now just part of the model-fitting process (by a computing method known as backfitting).
- It is also possible to include functions such as $f_{jk}(X_j, X_k)$ for nonlinear interactions.

Recommended Readings

ISLv2 (basics):

- Sections 7.1–7.6, 7.7.1
- Labs: Section 7.8

ESL (advanced):

- Sections 5.2, 5.4, 9.1