STA 621: Nonparametric Statistics
Breheny

Assignment 5
Due: December 4

Mathematical concepts and derivations

1. Show that the linear weights $\{l_i(x_0)\}$ for local linear regression (defined on slide 13, 11-1 notes) satisfy
   (a) $\sum_i l_i(x_0) = 1$ for all x_0
   (b) $\sum_i l_i(x_0)(x_i - x_0) = 0$ for all x_0
   (c) If $K(x_i, x_0) = 0$, then $l_i(x_0) = 0$

2. Suppose $\hat{\mathbf{f}} = \mathbf{Ly}$ is a linear smoother. Show that

$$\frac{1}{n}\sum_i \left\{ y_i - \hat{f}_{(-i)}(x_i) \right\}^2 = \frac{1}{n}\sum_i \left(\frac{y_i - \hat{f}_i}{1 - l_{ii}} \right)^2 .$$

3. (a) Write the set of truncated spline basis functions for representing a cubic spline function with three knots.
   (b) For $x \in [0, 1]$ and equally spaced knots, plot the above basis functions (exclude the intercept term).
   (c) For the same range and knots as in (b), plot the B-spline basis functions, again excluding the intercept term. Note that we did not cover how to construct this basis; you may use the `splines` package to construct them for you.

4. Show that the objective function for penalized splines (defined on slide 30, 11-20 notes) is

$$(\mathbf{y} - \mathbf{N}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{N}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}'\boldsymbol{\Omega}\boldsymbol{\beta},$$

   where $\boldsymbol{\Omega}_{jk} = \int N_j''(t)N_k''(t)dt$. $N_j(\cdot)$ is defined on slide 35 of the same notes.

Simulation

5. Suppose $x_i \overset{\text{iid}}{\sim} Unif(-3, 3)$ for $i = 1, 2, \ldots, 100$ and that $y_i = f(x_i) + \epsilon$, where ϵ follows a standard normal distribution and $f(x) = -x^2$. Conduct a simulation study comparing three methods: polynomial regression (with linear and quadratic terms), smoothing splines, and local linear regression. NOTE: Some implementations refuse to predict outside the observed range. To avoid this, set two values of $\{x_i\}$ equal to -3 and 3 and let the other 98 be uniformly distributed.

(a) On average, how many degrees of freedom do splines and local linear regression need to represent f?

(b) At equally spaced points throughout the range of x, evaluate the bias, variance, and MSE of the three methods. Plot your results versus x and comment on what you see.

6. Repeat the above exercise, only change f to be the following piecewise function:

$$f(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x^2 & \text{if } x > 0 \end{cases}.$$

One interpretation of such an f is that it represents a risk factor for which low levels have no impact, but past a certain threshold, there is an increasingly severe risk. In addition to the three methods in the above exercise, add a fourth method: polynomial regression with terms up to x^5.

(a) How many degrees of freedom do splines and local linear regression need to represent f?

(b) At equally spaced points throughout the range of x, evaluate the bias, variance, and MSE of the three methods. Plot your results versus x and comment on what you see.

Application

NOTE: For the following three problems, use each of the following methods exactly once: local regression, regression splines, smoothing splines. The choice of which method to use for which problem is up to you.

7. The course website contains a data set `motorcycle` from an experiment on the efficacy of crash helmets. The response (`Acceleration`) is measured by an accelerometer placed inside a helmet. The change in this acceleration over time is of interest.

(a) Choose a criterion by which to select the smoothing parameter. Plot this criterion versus the smoothing parameter and choose the optimal value for use in (b) and (c).

(b) Plot a smooth curve estimating acceleration as a function of time, and include bands to indicate confidence regions.

(c) Prepare an ANOVA table testing the sequence of models: Null $\subset$ Linear $\subset$ Nonlinear.

8. The course website contains a data set `asthma` from a study of the relationship between childhood asthma and exposure to air pollution from concentrated animal feeding operations (CAFOs). For this problem, treat asthma (Yes/No) as following a binomial distribution given exposure, and use a smooth version of logistic regression for (a)-(c) below.

(a) Choose a criterion by which to select the smoothing parameter. Plot this criterion versus the smoothing parameter and choose the optimal value for use in (b) and (c).

(b) Plot a smooth curve estimating the relationship between exposure and the log-odds of developing asthma, with confidence bands.

(c) Plot a smooth curve estimating the relationship between exposure and the probability of developing asthma, with confidence bands.

(d) Prepare an ANOVA table (or rather, an analysis of deviance table) testing the sequence of models: Null $\subset$ Linear $\subset$ Nonlinear.

9. The standard R data set `airquality` contains daily measurements of air quality (in terms of ozone concentration) taken in New York during the summer of 1973. In this problem, let us investigate the relationship between temperature and ozone concentration.

    (a) Choose a criterion by which to select the smoothing parameter. Plot this criterion versus the smoothing parameter and choose the optimal value for use in (b) and (c).

    (b) Plot a smooth curve estimating average ozone concentration as a function of temperature, and include confidence bands to illustrate the uncertainty of the estimate.

    (c) Prepare an ANOVA table testing the sequence of models: Null $\subset$ Linear $\subset$ Nonlinear.