Assignment 9 - Smoothing Methods

your name goes here

Due: Friday, April 6, 2018

Summary

Beyond the standard LS model, more flexible functional relationships can be built using local polynomial regression fitting. Generally, all the smoothing models covered use the mechanics of LS to fit the models and give prediction errors. However, the coefficients are not necessarily interpretable, because they do not extend over a range of X-values.

The homework in this assignment comes primarily from the alternative textbook, *An Introduction to Statistical Learning*, `http://www-bcf.usc.edu/~gareth/ISL/`.

Assignment

1. (n.b. You are welcome to do this problem with pencil.)
   It was mentioned in the chapter that a cubic regression spline with one knot at ξ can be obtained using a basis of the form $X, X^2, X^3, (X-\xi)^3_+$, where $(X-\xi)^3_+ = (X-\xi)^3$ if $X > \xi$ and equals 0 otherwise.

   We will now show that a function of the form

   $$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 (X-\xi)^3_+$$

   is indeed a cubic regression spline, regardless of the values of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$.

   (a) Find a cubic polynomial
   $$f_1(X) = a_1 + b_1 X + c_1 X^2 + d_1 X^3$$
   such that $f(X) = f_1(X)$ for all $X \le \xi$. Express a_1, b_1, c_1, d_1 in terms of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$.

   (b) Find a cubic polynomial
   $$f_2(X) = a_2 + b_2 X + c_2 X^2 + d_2 X^3$$
   such that $f(X) = f_2(X)$ for all $X > \xi$. Express a_2, b_2, c_2, d_2 in terms of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$. We have now established that $f(X)$ is a piecewise polynomial.

   (c) Show that $f_1(\xi) = f_2(\xi)$. That is, $f(X)$ is continuous at ξ.

   (d) Show that $f_1'(\xi) = f_2'(\xi)$. That is, $f'(X)$ is continuous at ξ.

   (e) Show that $f_1''(\xi) = f_2''(\xi)$. That is, $f''(X)$ is continuous at ξ.

2. Suppose we fit a curve with basis functions $bf_1(X) = I(0 \le X \le 2) - (X-1)I(1 \le X \le 2)$, $bf_2(X) = (X-3)I(3 \le X \le 4) + I(4 < X \le 5)$. We fit the linear regression model

   $$E[Y] = \beta_0 + \beta_1 bf_1(X) + \beta_2 bf_2(X) + \epsilon,$$

   and obtain coefficient estimates $b_0 = 1, b_1 = 1, b_2 = 3$. **Sketch the estimated curve** between $X = -2$ and $X = 2$. Note the intercepts, slopes, and other relevant information.

3. Use the `Boston` data with the variables `dis` (the weighted mean of distances to five Boston employment centers) and `nox` (nitrogen oxides concentration in parts per 10 million). Treat `dis` as the explanatory variable and `nox` as the response.

   (a) Use the `poly()` function to fit a cubic polynomial regression to predict `nox` using `dis`. Report the regression output, and plot the resulting data and polynomial fits.

   (b) Plot the polynomial fits for a range of different polynomial degrees (say, from 1 to 10), and report the associated residual sum of squares (SSE) on the full dataset. Which degree value seems best?

   (c) Using LOOCV, which polynomial degree is best?

   (d) Use the `bs()` function to fit a regression spline to predict `nox` using `dis`. Report the output for the fit using `df=4` in the function call. How did R choose the knots? What are they? Plot the resulting fit.

   (e) Now fit a regression spline for a range of degrees of freedom, and plot the resulting fits and report the resulting SSE on the full dataset. Describe the results obtained.

   (f) Using LOOCV, which `df` is best?

   (g) Fit a loess curve (local regrssion) for a few different options for the `span` parameter. What is your SSE on the full dataset? Which span value seems to fit the entire dataset best? Explain.

   (h) Using LOOCV, which `span` is best?