Homework 8

Advanced Methods for Data Analysis (36-402)

Due Friday March 29, 2019, at 3:00 PM

See the syllabus for general instructions regarding homework. Note that you should **always show all your work** and submit **exactly two files** (Rmd or R file as *code*, and knitted or scanned/merged pdf or html as *writeup*). Make sure that everything you submit is readable.

Problem 1: Degrees of freedom for linear smoothers

Suppose that $Y_i = \mu(x_i) + \epsilon_i$, where ϵ_i are uncorrelated and have mean 0, but each row has its own variance σ_i^2. Consider modifying the definition of degrees of freedom to $\sum_{i=1}^{n} \frac{Cov[Y_i, \hat{Y}_i]}{\sigma_i^2}$. Show that this still equals to $tr(W)$ for a linear smoother with influence matrix W.

Problem 2: Degrees of freedom for kernels and splines (housing data revisited)

For this problem, we will use the `housing` data that you have used in HWs 2 and 5. Merge the traning and test data sets into a single data set.

We will be fitting models in which we try to predict median house value from mean household income. We will use both splines and normal-kernel regressions with a variety of tuning parameters while we attempt to use "effective degrees-of-freedom" to make the comparisons on a common scale.

Part a

Plot `Median_house_value` versus `Mean_household_income`.

Part b

First, fit several spline modes with different smoothing parameters. We will use `smooth.spline` with the `df` version of the smoothing parameter. (This is the same as the `effective degrees-of-freedom` that was defined in lecture.)

Use `df` parameter values from 2 to 27 in steps of 1. For each `df`, run five-fold cross-validation on the full data set, computing the estimated MSE of prediction on each held-out fold. Compute the sample average of the five estimated MSE of prediction values for each `df`. Also, for each `df`, compute an estimated standard error for the average by computing the sample standard deviation of the five estimated MSE of prediction values divided by $\sqrt{5}$. Plot the sample average of the five estimated MSE of prediction values against `df`, and find the `df` value that provides the lowest average. Compare the differences between the plotted averages to the estimated standard errors of the averages. Say why this comparison either does or does not inspire confidence that we have really found the `df` that provides the best out-of-sample prediction.

Part c

Next, perform a similar analysis for kernel regression with a normal (Gaussian) kernel. This time, there is no `df` option for the smoothing parameter. Instead, use the `bws` parameter to set the bandwidth to each of the values from 4000 to 7000 in steps of 300. (That should be 11 different values, for a sanity check.) Make sure that you use the same five folds for five-fold cross-validation as you used in part (b). The comparisons will not be fair if you use different folds. Use the `npreg` function in `library(np)` to fit the kernel regressions. In the instructions from part (b), replace `df` by `bws` and repeat all of the analyses for the kernel regression models. (This calculation is time-consuming.)

Part d

For each bandwidth h, the effective degrees-of-freedom for a kernel regression is

$$df(h) = K(0) \sum_{i=1} \frac{1}{\sum_{j=1}^{n} K([x_i - x_j]/h)}.$$

(This calculation is time-consuming.) For each bandwidth (`bws`) in part (c), compute the corresponding effective degrees-of-freedom. Draw a single plot that contains the pairs

(effective degrees-of-freedom,estimated MSE of prediction)

for both the spline fits and the kernel fits, labeled well enough to be able to tell which are which. Comparing the differences between the estimated MSE's for splines and kernels to the estimated standard errors, say why this comparison does or does not inspire confidence that we can tell which method provides better out-of-sample prediction.

Part e

Finally redraw the plot from part (a) and add

(i) a line for the spline fit corresponding to the best smoothing parameter value, and

(ii) a line for the kernel regression fit with the best smothing parameter value.

For the two lines, use a common set of 201 predictor values that is equally-spaced over the observed range of `Mean_household_income` and extending 5% at each end. Comment on how well or badly each of the two fitted curves seems to fit the data. Also offer a reason for why the two fitted curves behave so differently at and beyond the extremes of the predictor.