STA 621: Nonparametric Statistics
Breheny

Assignment 2
Due: Tuesday, October 2

Mathematical concepts and derivations

1. What are the weights $\{w_i\}_{i=1}^n$ for the delete-d jackknife?

2. A rather interesting 1987 paper by DiCiccio and Tibshirani proposes a confidence interval procedure called ABC, for "approximate bootstrap confidence" intervals. Their idea is to use influence functions to approximate what the bootstrap confidence intervals would be, without actually performing any bootstrap resampling. The `boot` package has a function `abc.ci` to compute these intervals. To use it, however, the function that calculates your statistic must take two arguments: the data and a vector of weights. Why would the function require weights, as opposed to the vector of indices that the usual `boot` function requires?

3. Show that, for $\theta = \mathbb{E}(X)$,

$$-2 \log \mathcal{R}(\theta_0) \xrightarrow{\text{d}} \chi_1^2$$

Hint: do this in two parts. For the first part, take a Taylor series expansion of

$$\frac{1}{n} \sum_i \frac{x_i - \theta}{1 + \lambda(x_i - \theta)} = 0$$

about $\lambda = 0$ to show that $\lambda \approx (\bar{x} - \theta)/S$, where $S = n^{-1} \sum_i (x_i - \theta)^2$. In the second part, use the above approximation to show that $-2 \log \mathcal{R}(\theta_0) \approx n(\bar{x} - \theta_0)^2/S$ (Hint: use the fact that when $x \approx 0$, $\log(1 + x) \approx x - \frac{1}{2}x^2$).

Simulation

4. Compare three nonparametric methods for constructing confidence intervals for the variance of a random variable: the functional delta method, the bootstrap percentile interval, and the BC_a interval.

Conduct a simulation study to determine how the coverage probability and average interval width of these two intervals varies with the sample size n. For each of the distributions below, produce a plot of coverage probability versus sample size, with lines representing the various methods, as well as a corresponding plot for interval width.

(a) Carry out the above simulation with data generated from the standard normal distribution.

(b) Repeat using data generated from an exponential distribution with rate 1.

(c) Briefly, comment on the strengths and weaknesses of these three methods.

Implementation

5. Write an R function called jackknife which implements the jackknife. The function should accept two arguments: x (the data) and theta (a function which, when applied to x, produces the estimate). The function should return a named list with the following components:

   - bias: the jackknife estimate of bias
   - se: the jackknife estimate of standard error
   - values: the leave-one-out estimates $\{\hat{\theta}_{(i)}\}$

   Please submit the function via Dropbox and name your file Breheny-jack.R, with your last name replacing Breheny.

Application

6. The standardized test used by law schools is called the Law School Admission Test (LSAT), and it has a reasonably high correlation with undergraduate GPA. The course website contains data on the average LSAT score and average undergraduate GPA for the 1973 incoming class of 15 law schools. (See the notes for hints).

   (a) Use the jackknife to obtain an estimate of the bias and standard error of the correlation coefficient between GPA and LSAT scores. Comment on whether the estimate is biased upward or downward.

   (b) If x and y are drawn from a bivariate normal distribution, then $n\mathbb{V}(\hat{\rho}) \xrightarrow{\text{P}} (1-\rho^2)^2$. Use this to estimate the standard error of $\hat{\rho}$.

   (c) On page 21 of our textbook, the author gives the influence function for the correlation coefficient:

   $$L(x,y) = \tilde{x}\tilde{y} - \frac{1}{2}\theta(\tilde{x}^2 + \tilde{y}^2),$$

   where

   $$\tilde{x} = \frac{x - \mu_x}{\sqrt{\sigma_x^2}}$$

   and $\tilde{y}$ is defined similarly. Use this to estimate the standard error of $\hat{\rho}$.

   (d) Use the bootstrap to estimate the the standard error of $\hat{\rho}$.

   (e) Plot a histogram of your bootstrap replications $\{\hat{\rho}_b^*\}$. Does the sampling distribution appear to be normally distributed?

   (f) Compare the four estimates (a)-(d).

   (g) For each data point (x_i, y_i), make a plot of $\hat{\rho}_\epsilon$ vs. the mass at point i, as the point mass at i varies from 0 to 1/3 (and the rest of the mass is spread evenly on the rest of the observations).

   (h) Comment on why some plots slope upwards and others slope downwards.

   (i) Extra credit: Comment on the how the shape of the curve relates to the comparison between the delta method and jackknife estimates of the variance

7. The course website contains data consisting of test scores of 88 students in 5 subjects: Mechanics, Vectors, Algebra, Analysis, and Statistics. One natural question about this data is the extent to which these tests measure separate skills vs. general tests of quantitative ability. One way to quantify this is via the ratio of the largest eigenvalue of the correlation matrix to the sum of the eigenvalues:

$$\hat{\theta} = \hat{\lambda}_1 / \sum_{i=1}^{5} \hat{\lambda}_i,$$

where $\{\hat{\lambda}_i\}$ are the eigenvalues, sorted from largest to smallest.

(a) Use the bootstrap to estimate the standard error of $\hat{\theta}$.

(b) Plot a histogram of your bootstrap replications $\{\hat{\theta}_b^*\}$. Does the sampling distribution appear to be normally distributed?

(c) Compare the histogram in part (b) with the histogram from 5(e). Does one histogram appear "less normal" than the other? Why?