Homework on the bootstrap. Econometrics GS63.2707 November 19, 2001, due on December 3, 2001

In this homework data written $y_1, y_2, \ldots, y_n$ and will be thought to have been drawn from iid random variables $Y_1, Y_2, \ldots, Y_n$ whose PDF and CDF shall be denoted $f$ and $F$ respectively. Our aim in these homeworks will be to use the bootstrap to make inferences about a population characteristic $\theta$, using a statistic $T$ that is an estimate for $\theta$ and has value $t$ in the sample. Generally speaking $t$ will be a symmetric function of the data, and therefore we will write

$$y(1), y(2), \ldots, y(n)$$

for a re-arrangement of the $y$’s in order from smallest to largest.

We will also be using the empirical distribution function (EDF) noted by $\hat{F}$ defined by

$$\hat{F}(y) = \frac{\#\{y_j \leq y\}}{n}$$

We will also write $Y^*$ to represent the random variable $Y$ as interpreted when drawing from the model $\hat{F}$. In addition we will write $E^*$ (respectively $\text{var}^*$ for the expected value and variance functional for the model $\hat{F}$). Lastly we will write $\hat{f}$ to denote the PDF associated with $\hat{F}$.

We will write $\bar{Y}$ for the random variable derived from the average of $Y_1, Y_2, \ldots Y_n$. Also we will write $\bar{y}$ for the sample average of the $y_i$’s.

1. Show that $E(\hat{F}(y)) = F(y)$ and that $\text{var}(\hat{F}(Y)) = F(y)(1 - F(y))/n$.

2. Show the following identities hold

$$E^*(Y^*) = \bar{y}$$

$$\text{var}^*(Y^*) = \frac{1}{n} \sum (y_j - \bar{y})^2$$

3. Let $T = \frac{1}{n} \sum(Y_j - \bar{Y})^2$ be a statistic, and let $t$ be the corresponding statistic derived from the data. Show that

$$E^*(T^*) = (n - 1)t/n$$

$$\text{var}^*(T^*) = (n-1)^2[m_4/n+(3-n)t^2/(n(n-1))]/n^2, \quad m_4 = n^{-1} \sum(y_j - \bar{y})^4$$

1
4. Let $T$ be the median of $Y$ and $t$ be the sample median.

(a) Show that $t = y_{(m+1)}$

(b) Show that $T^* > y_{(l)}$ if and only if fewer than $m + 1$ of the $Y_j^*$ are
less than or equal to $y_{(l)}$, and

(c) Show that

$$\text{Prob}^{*}(T^* > y_{(l)}) = \sum_{j=0}^{m} \left( \frac{n}{n} \right)^i (1 - \frac{i}{n})^{n-j}$$

(d) Use the previous formula to show that for $n = 11$ that

$$\text{Pr}^*(T^* \leq y_{(3)}) = \text{Pr}^*(T^* \geq y_{(9)}) = .051.$$  

(e) Confirm the previous result via simulation.

5. Design a simulation experiment as follows, specifying the details of your
parameter set as you go along.

(a) Set up a program to sample from the mixture of normals distri-
bution

$$Y = pN(0, \sigma_1) + (1 - p)N(0, \sigma_2)$$

where $p$ is a probability, assumed to be close to 1 and $\sigma_1 << \sigma_2$, so
that the distribution has a small probability of producing a large
draw.

Repeat the following steps for $p = .01$ and $p = .05$ with $\sigma_2/\sigma_1 = 5$:

- Draw a large sample from $Y$.
- Estimate the mean of $Y$ under the assumption that $f$ is nor-
mal (i.e., not a mixture distribution).
- Write down a (i) 95% confidence interval for the mean, (ii) the standard error for the estimate of the mean, both based
on the assumption that $Y$ is normal.
- Create a bootstrap estimate for the 95% confidence interval
and standard error. Compare the result to those arrived at
assuming that $Y$ was normal.

What are your general conclusions about the impact of contami-
nation on the standard error?
6. There are two major types of approximations involved in the bootstrap.
   The first if the approximation of \( f^* \) versus \( f \), i.e., the approximation of the EDF to the true distribution. The second is the approximation inherent in the Monte Carlo estimate in sampling from the EDF. Explain the differences in these two approximations in terms of the impact on the estimation method.

7. In this exercise we will use the bootstrap method to estimate the correlation coefficient of a sample. Recall that the correlation coefficient of bi-variate sample is written

\[
\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}
\]

where \( \sigma_{xy} \) is the covariance between \( x \) and \( y \) and \( \sigma_x \) and \( \sigma_y \) are the variance of \( x \) and \( y \) respectively.

Write a program to draw \( n \) samples from a bi-variate Gaussian distribution with a known coefficient (you can do this using a Cholesky decomposition) and bootstrap estimate \( \rho_{xy} \) and provide an estimate of the variance and standard error of \( \rho_{xy} \).