Resampling Techniques

The statistician is frequently interested on obtaining an estimate together with its standard error so that he/she can construct confidence intervals for the true value of the parameters. However it is rather difficult to obtain an *exact* expression for the variance of several estimators and therefore it is impossible to calculate their standard error. Over the years statisticians have used approximations or transformations to achieve normality but this might be prohibitive for a large class of problems.

Nowadays, the cheap computing power turned the attention to intensive computing resampling methods as the bootstrap and the jackknife.

Bootsrap

The methodology of bootstrap is based upon creating B new samples each of the same size as the observed data which have been sampled *with replacement* from the observed data.The estimate (statistic) of interest is calculated for each bootstrap data set B times yielding to the bootstrap distribution of the statistic. A fundamental assumption is that the observed data represents the underlying population.

Correlation Coefficient

We consider the example of Efron and Tibshirani (1993) in which 82 law schools participated in a study of admission practice. For each of these schools 15 schools were randomly sampled and the correlation between the LSAT and GPA score was examined based on the 1973 entering class. To carry out a bootstrap analysis in S-Plus consider the following

```
> school <- 1:15
> lsat   <- c(576,635,558,578,666,580,555,661,651,605,653,575,545,572,594)
> gpa    <- c(3.39, 3.30, 2.81, 3.03, 3.44, 3.07, 3.00, 3.43, 3.36, 3.13,
  + 3.12, 2.74, 2.76, 2.88, 2.96)
> law.data <- data.frame(School=school, LSAT=lsat, GPA=gpa)
> boot.obj1  <- bootstrap(law.data, cor(LSAT, GPA), B=1000, seed=0)
Forming replications  1  to  100
Forming replications  101  to  200
Forming replications  201  to  300
Forming replications  301  to  400
Forming replications  401  to  500
Forming replications  501  to  600
Forming replications  601  to  700
Forming replications  701  to  800
Forming replications  801  to  900
Forming replications  901  to  1000
> summary(boot.obj1)
Call:
bootstrap(data = law.data, statistic = cor(LSAT, GPA), B = 1000)

Number of Replications: 1000

Summary Statistics:
      Observed      Bias   Mean     SE
Param   0.7764 -0.008768 0.7676 0.1322

Empirical Percentiles:
        2.5%    5%    95%  97.5%
Param 0.4673 0.523 0.9432 0.9593

BCa Confidence Limits:
        2.5%    5%    95%  97.5%
Param 0.3443 0.453 0.9255 0.9384
```

The function `bootstrap()` creates 1000 correlation coefficients by sampling with replacement from the 15 pairs. The summary statistic refer to the number of replications, the observed value of the parameter which equals to the mean of the bootstrap replication and the bootstrap estimates of bias and standard error. In addition we obtain the empirical percentiles together with bias–corrected and adjusted (BCa) percentiles. A 95% bootstrap confidence interval for the correlation coefficient of this example is given by $(0.3443, 0.9384)$, for example. Plotting the object (see Figure **??**) leads to a histogram of the resampling estimates together with a smooth density estimator. In this case, the histogram is not symmetric.

Regression Coefficients

The following examples shows how we can bootstrap regression coefficients by using the data in the previous section.

```
 boot.obj2 <- bootstrap(data=law.data, coef(lm(LSAT~GPA, data=law.data)), B=1000)
Forming replications  1  to  100
Forming replications  101  to  200
Forming replications  201  to  300
```

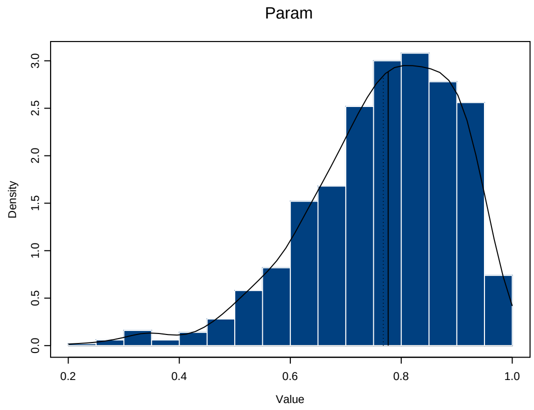

Figure 1: Histogram of 1000 bootstrap correlation coefficients.

```
Forming replications  301  to  400
Forming replications  401  to  500
Forming replications  501  to  600
Forming replications  601  to  700
Forming replications  701  to  800
Forming replications  801  to  900
Forming replications  901  to  1000

> summary(boot.obj2)
Call:
bootstrap(data = law.data, statistic = coef(lm(LSAT ~ GPA, data = law.data)), B = 1000)

Number of Replications: 1000

Summary Statistics:
            Observed   Bias  Mean    SE
(Intercept)    187.9 -5.373 182.5 90.34
        GPA    133.3  1.753 135.0 30.31

Empirical Percentiles:
             2.5%    5%   95% 97.5%
(Intercept) 38.12 49.42 335.7 393.5
        GPA 64.25 81.82 179.3 183.6

BCa Confidence Limits:
             2.5%    5%   95% 97.5%
(Intercept) 58.80 74.05 411.4 469.4
        GPA 35.27 55.51 170.6 175.8
```

```
Correlation of Replicates:
            (Intercept)     GPA
(Intercept)      1.0000 -0.9976
        GPA     -0.9976  1.0000
```

The above results show that the bootstrap estimate of the regression parameters are 187.9 and 133.3. In particular the slope is positive as it should be expected from the previous results (why?). Figure 2 shows the histogram of the bootstrap estimates.

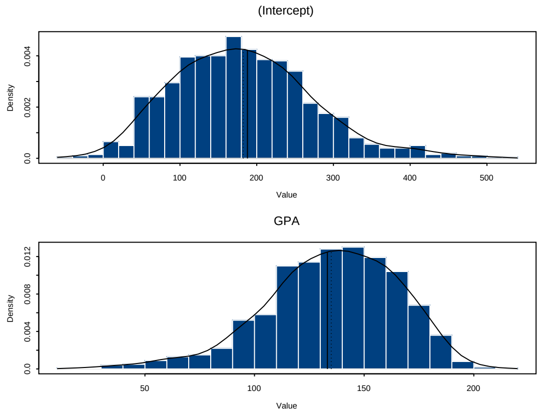

Figure 2: Histogram of 1000 bootstrap regression coefficients.

Jackknife

In jackknife, resampling a statistic amounts to its calculation for the n possible samples of size $n-1$, each with one observation left out. Here we use a simple example for illustration by generating 6 observations from the uniform between 0 and 1. It is well known that the maximum of the observations is the mle and it is biased. The `jackknife()` calculates the observed maximum and the jackknife estimate together with its bias and standard error.

```
> x <- runif(6)
> jack1 <- jackknife(x, statistic=max)
> summary(jack1)
Call:
jackknife(data = x, statistic = max)

Number of Replications: 6
```

```
Summary Statistics:
    Observed    Bias   Mean     SE
max   0.9777 -0.1184 0.9541 0.1184

Empirical Percentiles:
      2.5%     5%    95%  97.5%
max 0.8534 0.8712 0.9777 0.9777
```