Chapter 8

Bootstrap and Jackknife

Packages needed for these exercises include boot, bootstrap and DAAG.

8.1 *Compute a jackknife estimate of the bias and the standard error of the correlation statistic for the `law` data.*

```
    library(bootstrap)
    attach(law)
    n <- nrow(law)

    theta.hat <- cor(LSAT, GPA)

    #compute the jackknife replicates, leave-one-out estimates
    theta.jack <- numeric(n)

    for (i in 1:n)
        theta.jack[i] <- cor(LSAT[-i], GPA[-i])

    bias <- (n - 1) * (mean(theta.jack) - theta.hat)

    se <- sqrt((n-1) *
        mean((theta.jack - mean(theta.jack))^2))
    print(list(est=theta.hat, bias=bias, se=se))

## $est
## [1] 0.7763745
##
## $bias
## [1] -0.006473623
##
## $se
## [1] 0.1425186

    detach(law, unload=TRUE)
```

8.4 *Refer to the air-conditioning data set `aircondit` provided in the `boot` package. Assume that the times between failures follow an exponential model Exp(λ). Obtain the MLE of the hazard rate λ and use bootstrap to estimate the bias and standard error of the estimate.*

The MLE of λ is $1/\bar{x}$. The estimates of bias and standard error are printed in the summary of `boot` below.

```
    library(boot)
    x <- aircondit[1]
    rate <- function(x, i)
        return(1 / mean(as.matrix(x[i,])))
    boot(x, statistic=rate, R=2000)

##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = x, statistic = rate, R = 2000)
##
##
## Bootstrap Statistics :
##       original        bias    std. error
## t1* 0.00925212 0.001236637 0.004322421
```

8.5 *Refer to Exercise 8.4. Compute 95% bootstrap confidence intervals for the mean time between failures $1/\lambda$ by the standard normal, basic, percentile and BCa methods. Compare the intervals and explain why they may differ.*

The `aircondit` data is a data frame with 1 variable, so `aircondit[1]` extracts this variable.

```
    library(boot)
    x <- aircondit[1]
    meant <- function(x, i)
        return(mean(as.matrix(x[i,])))
    b <- boot(x, statistic=meant, R=2000)
    b

##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = x, statistic = meant, R = 2000)
##
##
## Bootstrap Statistics :
##     original    bias    std. error
```

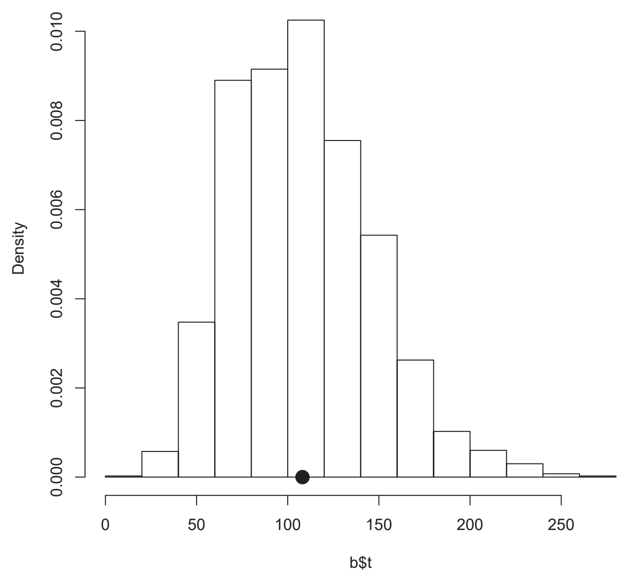
```
## t1* 108.0833 0.4974583    38.24609

    boot.ci(b, type=c("norm","perc","basic","bca"))

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = b, type = c("norm", "perc", "basic", "bca"))
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 32.6, 182.5 )   ( 23.0, 168.9 )
##
## Level     Percentile            BCa
## 95%   ( 47.2, 193.2 )   ( 57.5, 227.2 )
## Calculations and Intervals on Original Scale
## Some BCa intervals may be unstable
```

The replicates are not approximately normal, so the normal and percentile intervals differ. From the histogram of replicates, it appears that the distribution of the replicates is skewed - although we are estimating a mean, the sample size is too small for CLT to give a good approximation here. The BCa interval is a percentile type interval, but it adjusts for both skewness and bias.

```
    hist(b$t, prob=TRUE, main="")
    points(b$t0, 0, cex=2, pch=16)
```

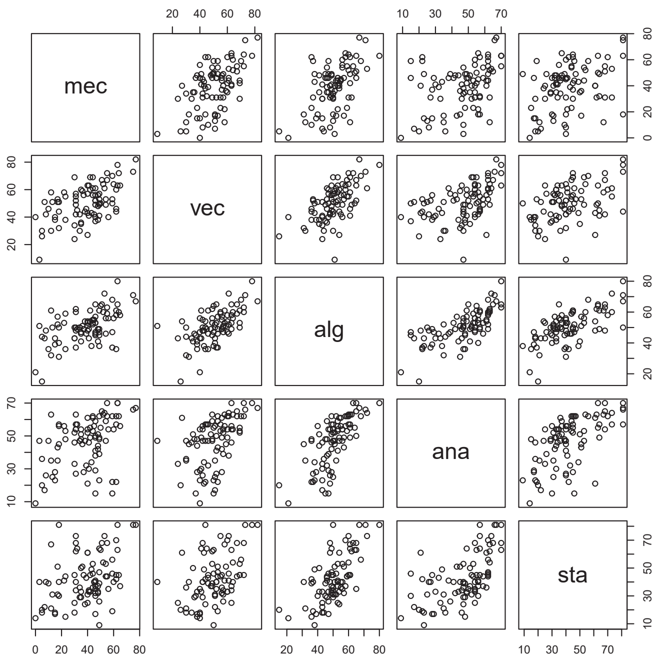


8.6 *Efron and Tibshirani (1993) discuss the* `scor (bootstrap)` *test score data on 88 students who took examinations in five subjects. The first two tests (mechanics, vectors) were closed book and the last three tests (algebra, analysis, statistics) were open book. Each row of the data frame is a set of scores* $(x_{i1}, \ldots, x_{i5})$ *for the* i^{th} *student. Use a panel display to display the scatter plots for each pair of test scores. Compare the plot with the sample correlation matrix. Obtain bootstrap estimates of the standard errors for each of the following estimates:* $\hat{\rho}_{12} = \hat{\rho}(mec,\ vec)$, $\hat{\rho}_{34} = \hat{\rho}(alg,\ ana)$, $\hat{\rho}_{35} = \hat{\rho}(alg,\ sta)$, $\hat{\rho}_{45} = \hat{\rho}(ana,\ sta)$.

```r
library(bootstrap)
attach(scor)
cor(scor)
```

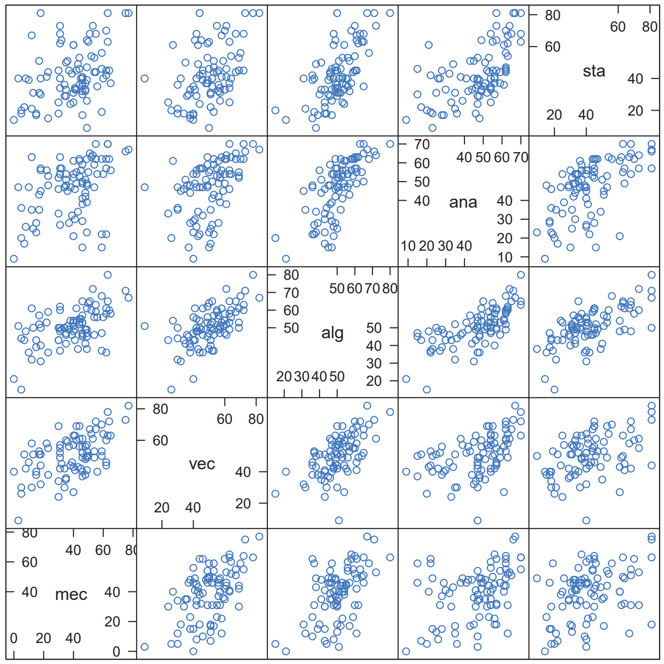
```
##           mec       vec       alg       ana       sta
## mec 1.0000000 0.5534052 0.5467511 0.4093920 0.3890993
## vec 0.5534052 1.0000000 0.6096447 0.4850813 0.4364487
## alg 0.5467511 0.6096447 1.0000000 0.7108059 0.6647357
## ana 0.4093920 0.4850813 0.7108059 1.0000000 0.6071743
## sta 0.3890993 0.4364487 0.6647357 0.6071743 1.0000000
```

```
pairs(scor)
```



Alternately, the *lattice* version:

```
lattice::splom(scor)
```



Scatter Plot Matrix

From the plots and correlation matrix, it appears that open book scores have higher correlations than closed book. All test scores have positive sample correlations. The approximate standard errors of the estimates are given by the output from the `boot` function below.

```
library(boot)
cor.stat <- function(x, i=1:NROW(x)) {
    cor(x[i,1], x[i,2])
}

x <- as.matrix(scor)
```

Bootstrap estimate of $se(\hat{\rho})_{12} = \hat{\rho}(mec, vec)$:

```
boot(x[,1:2], statistic=cor.stat, R=2000)

##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
```

```
## boot(data = x[, 1:2], statistic = cor.stat, R = 2000)
##
##
## Bootstrap Statistics :
##      original       bias    std. error
## t1* 0.5534052 -0.004728317  0.07659473
```

Bootstrap estimate of $se(\hat{\rho})_{34} = \hat{\rho}(alg, ana)$:

```
    boot(x[,3:4], statistic=cor.stat, R=2000)

##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = x[, 3:4], statistic = cor.stat, R = 2000)
##
##
## Bootstrap Statistics :
##      original        bias    std. error
## t1* 0.7108059 -0.0006840475  0.04908036
```

Bootstrap estimate of $se(\hat{\rho})_{35} = \hat{\rho}(alg, sta)$:

```
    boot(x[,c(3,5)], statistic=cor.stat, R=2000)

##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = x[, c(3, 5)], statistic = cor.stat, R = 2000)
##
##
## Bootstrap Statistics :
##      original       bias    std. error
## t1* 0.6647357 -0.001299509  0.06016103
```

Bootstrap estimate of $se(\hat{\rho})_{45} = \hat{\rho}(ana, sta)$:

```
    boot(x[,4:5], statistic=cor.stat, R=2000)

##
```

```
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = x[, 4:5], statistic = cor.stat, R = 2000)
##
##
## Bootstrap Statistics :
##      original       bias    std. error
## t1* 0.6071743 -0.002470251  0.06942418

    detach(scor, unload=TRUE)
```

8.7 *Refer to Exercise 8.6. Efron and Tibshirani (1993) discuss the following example. The five-dimensional scores data have a 5×5 covariance matrix Σ, with positive eigenvalues $\lambda_1 > \cdots > \lambda_5$. Let $\hat{\lambda}_1 > \cdots > \hat{\lambda}_5$ be the eigenvalues of $\hat{\Sigma}$, where $\hat{\Sigma}$ is the MLE of Σ. Compute the sample estimate $\hat{\theta} = \frac{\hat{\lambda}_1}{\sum_{j=1}^5 \hat{\lambda}_j}$ of $\theta = \frac{\lambda_1}{\sum_{j=1}^5 \lambda_j}$. Use bootstrap to estimate the bias and standard error of $\hat{\theta}$.*

```
    library(boot)
    library(bootstrap)
    attach(scor)
    x <- cov(as.matrix(scor))
    e <- eigen(x)
    lambda <- e$values
    lambda

## [1] 686.98981 202.11107 103.74731  84.63044  32.15329

    lambda / sum(lambda)

## [1] 0.61911504 0.18214244 0.09349705 0.07626893 0.02897653

    th <- function(x, i) {
        y <- as.matrix(x[i, ])
        s <- cov(y)
        e <- eigen(s)
        lambda <- e$values
        max(lambda / sum(lambda))
        }
    boot(scor, statistic=th, R=2000)
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = scor, statistic = th, R = 2000)
##
##
## Bootstrap Statistics :
##     original       bias    std. error
## t1* 0.619115 0.0002187777  0.04674975

    detach(scor, unload=TRUE)
```

The estimates are $\hat{\lambda}_1 = 686.99$ and $\hat{\theta} \doteq 0.619$, with bias and std. error of $\hat{\theta}$ equal to 0.00074 and 0.047.

8.8 *Refer to Exercise 8.7. Obtain the jackknife estimates of bias and standard error of $\hat{\theta}$.*

```
    library(bootstrap)
    attach(scor)
    x <- as.matrix(scor)
    n <- nrow(x)
    theta.jack <- numeric(n)
    lambda <- eigen(cov(x))$values
    theta.hat <- max(lambda / sum(lambda))

    for (i in 1:n) {
        y <- x[-i, ]
        s <- cov(y)
        lambda <- eigen(s)$values
        theta.jack[i] <- max(lambda / sum(lambda))
        }

    bias.jack <- (n-1) * (mean(theta.jack) - theta.hat)
    se.jack <- sqrt((n-1)/n * sum((theta.jack -
                     mean(theta.jack))^2))

    c(theta.hat, bias.jack, se.jack)

## [1] 0.619115038 0.001069139 0.049552307

    list(est=theta.hat, bias=bias.jack, se=se.jack)
```

```
## $est
## [1] 0.619115
##
## $bias
## [1] 0.001069139
##
## $se
## [1] 0.04955231
```

```r
detach(scor, unload=TRUE)
```

The jackknife estimate of bias of $\hat{\theta}$ is approximately 0.001 and the jackknife estimate of se is approximately 0.05. These estimates are not very different from the bootstrap estimates above.

8.9 *Refer to Exercise 8.7. Compute 95% percentile and BCa confidence intervals for* $\hat{\theta}$.

```r
library(bootstrap)
attach(scor)
library(boot)
b <- boot(scor, statistic=th, R=2000)
boot.ci(b, type=c("perc", "bca"))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = b, type = c("perc", "bca"))
##
## Intervals :
## Level     Percentile            BCa
## 95%   ( 0.5281,  0.7102 )   ( 0.5249,  0.7065 )
## Calculations and Intervals on Original Scale
```

```r
detach(scor, unload=TRUE)
```