DATA130004: Homework 7

董晴园 14300680173

2017.12.8

1. Given i.i.d. random samples $X_1, ..., X_n$ with mean μ and variance σ^2, show that the sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

is an unbiased estimator of σ^2.

Proof:

$$\begin{aligned}
E(s^2) &= \frac{1}{n-1}\sum_{i=1}^{n} E\left(X_i - \overline{X}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[E\left(X_i^2\right) - 2E\left(X_i\overline{X}\right) + E\left(X^2\right)\right] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left[E\left(X_i^2\right) - \frac{2}{n}E\left(X_i^2 + \sum_{j\neq i} X_iX_j\right) + \frac{1}{n^2}E\left(\sum_{j=1}^{n} X_j\right)^2\right] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left[E\left(X_i^2\right) - \frac{2}{n}E\left(X_i^2\right) - \frac{2(n-1)}{n}(EX)^2 + \frac{1}{n}E\left(X_i^2\right) + \frac{n-1}{n}(EX)^2\right] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left[\frac{n-1}{n}E\left(X_i^2\right) - \frac{n-1}{n}(EX)^2\right] \\
&= E\left(X^2\right) - (EX)^2 = \sigma^2
\end{aligned}$$

2. Exercise 7.6: Efron and Tibshirani 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 $(x_{i1}, ..., x_{i5})$ 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).

Figure 1 is the scatter plots for each pair of test scores. The sample correlation matrix is as follows, and Figure 2 is the correlation plot.

```
> 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
```

Then we use bootstrap estimation to estimate the standard error of the correlation estimates. The codes to estimate $\hat{\rho}_{12}$ are as follows:

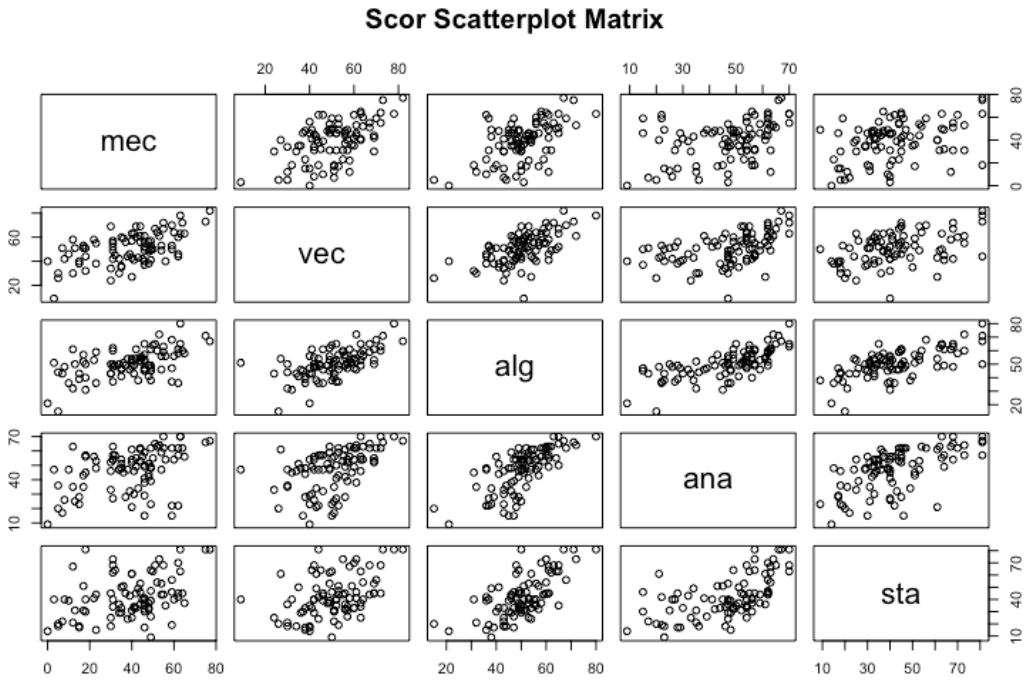

Figure 1: Scatter Plot Matrix of `scor`

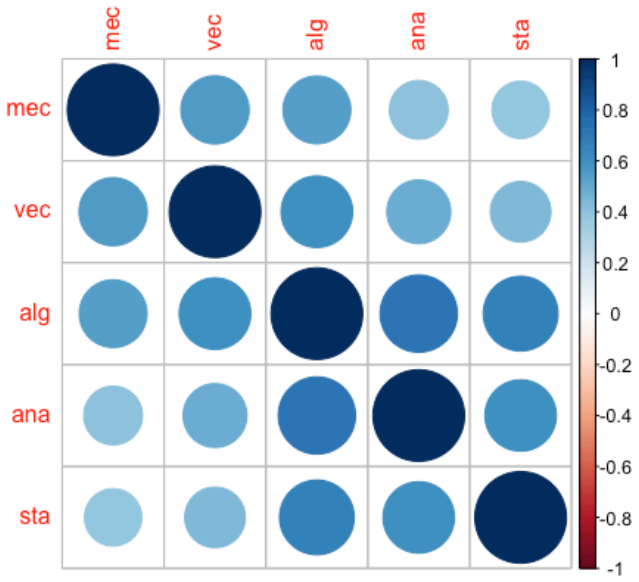

Figure 2: Correlation Plot of `scor`

```
library(bootstrap)    #for the law data
print(cor <- cor(scor$mec, scor$vec))
#set up the bootstrap
B <- 200            #number of replicates
n <- nrow(scor)      #sample size
R <- numeric(B)     #storage for replicates
#bootstrap estimate of standard error of R
for (b in 1:B) {
  #randomly select the indices
  i <- sample(1:n, size = n, replace = TRUE)
  mec <- scor$mec[i]       #i is a vector of indices
  vec <- scor$vec[i]
  R[b] <- cor(mec, vec)
}
#output
print(se.R <- sd(R))
[1] 0.08264175
```

Therefore, $s.e.(\hat{\rho}_{12}) \approx 0.083$. Using similar codes, we obtain $s.e.(\hat{\rho}_{34}) \approx 0.047$, $s.e.(\hat{\rho}_{35}) \approx 0.058$, $s.e.(\hat{\rho}_{45}) \approx 0.069$.

3. Exercise 7.7: Refer to Exercise 7.6. Efron and Tibshirani discuss the following example. The five-dimensional scores data have a 5×5 covariance matrix Σ, with positive eigenvalues $\lambda_1 > ... > \lambda_5$. In principal components analysis,

$$\theta = \frac{\lambda_1}{\Sigma_{j=1}^5 \lambda_j}$$

measures the proportion of variance explained by the first principal component. Let $\hat{\lambda}_1 > ... > \hat{\lambda}_5$ be the eigenvalues of $\hat{\Sigma}$, where $\hat{\Sigma}$ is the MLE of Σ. Compute the sample estimate of θ. Use bootstrap to estimate the bias and standard error of $\hat{\theta}$.

The sample estimate of θ is

$$\hat{\theta} = \frac{\hat{\lambda}_1}{\Sigma_{j=1}^5 \hat{\lambda}_j},$$

where $\hat{\lambda}_j$ is the jth eigenvalue of the covariance matrix Σ of the sample data.
Using bootstrap, we can estimate the bias and standard error of $\hat{\theta}$. And Figure 3 is the histogram of the bootstrap estimates $\hat{\theta}^b$, where the red vertical line represents the original sample estimate $\hat{\theta}$.

```
cov <- cov(scor)
lambda.hat <- eigen(cov)$values
theta.hat <- lambda.hat[1]/sum(lambda.hat)

#bootstrap estimate of bias
B <- 2000   #larger for estimating bias
n <- nrow(scor)
theta.b <- numeric(B)
for (b in 1:B) {
  i <- sample(1:n, size = n, replace = TRUE)
  scor.b <- matrix(c(scor$mec[i], scor$vec[i], scor$alg[i], scor$ana[i], scor$sta[i]),
                   ncol = 5)
  lambda.b <- eigen(cov(scor.b))$values
  theta.b[b] <- lambda.b[1]/sum(lambda.b)
}
bias <- mean(theta.b - theta.hat)
bias
[1] 0.001596521
```

```
#bootstrap estimate of standard error
print(sd(theta.b))
[1] 0.04683745

hist(theta.b, prob = TRUE)
abline(v=theta.hat, col="red")
```

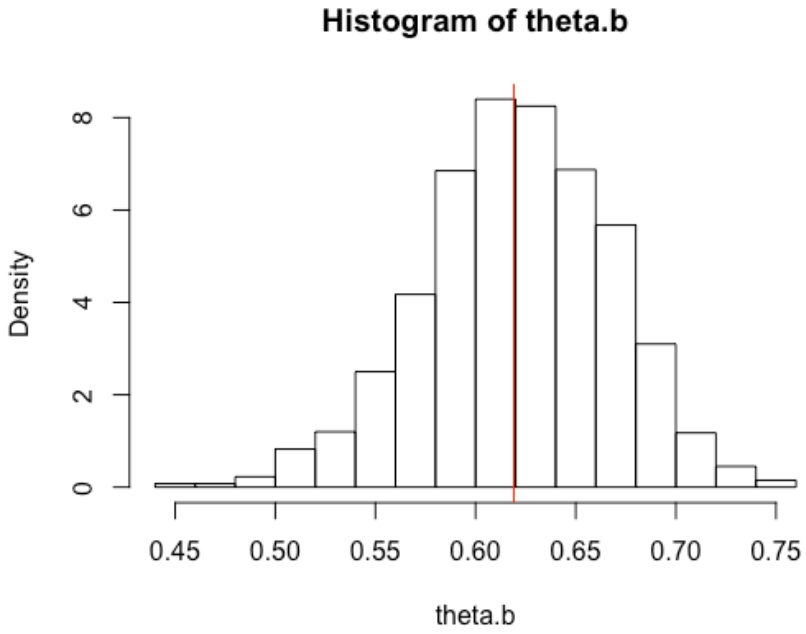

Figure 3: Histogram of the bootstrap sample $\hat{\theta}^b$