PROBABILITY
AND
MATHEMATICAL STATISTICS
Vol. 0, Fasc. 0 (0000), pp. 000–000

AN APPLICATION OF BOOTSTRAP IN ESTIMATING THE BONE MINERAL DENSITY OF VIETNAMESE WOMEN*

BY

NG. V. THU (HCM City), NG. V. TUAN (Sydney) AND NG. D. PHUONG (HCM City)

Abstract. In this paper, we apply the bootstrap method to study the standard deviation for bone mineral density of Vietnameses women. This result is important in recognizing seriousness of the osteoporosis.

2000 AMS Mathematics Subject Classification: Primary: ; Secondary:.

Key words and phrases: Bootstrap, standard deviation, standard error, confidence interval, regression, osteoporosis, bmd.

1. AN INTRODUCTION

In order to diagnose osteoporosis for Vietnameses women the World Health Organization (WHO) established the following criteria for determining the T-score:

$$T - score = \frac{bmd - bmdp}{sd} \tag{1.1}$$

where $bmdp$ is the peak bone mineral density of Vietnameses women and sd is the standard deviation of the peak bone mineral density of Vietnameses women. The

* This work has been done jointly with the group of doctors: N. T. T. Huong , P. T. M. Duc, L. H. Quang, N. V. Dinh, N. B. Duc, N. H. Binh, N. T. Anh, L. T. Thanh, and Bo von Schoultz. We would like to express our sincere thanks to them and hope to develop more fruittful cooperations with them.

problem we study in this report is to determine the $bmdp$ and sd using the bootstrap method . Bradley Efron (1974) revolutionized the field of statistics with his invention of the bootstrap. The bootstrap broadly refers to a continually growing collection of methodologies in which data are resampled to incorporate into statistical inference the information contained in the data regarding their probability distribution. Conceptually simple yet computationally intense, the bootstrap owes much of its rise in popularity over the last 20 years to the advent of the personal computer over the same period. As computers become faster and more powerful, the bootstrap becomes a more practical and indispensible tool for the data analyst. It can solve many problems including the problem of osteoporosis for Vietmese women that we can't solve before.

2. BOOTSTRAP OVERVIEW

We begin the study with the following definitions of the bootstrap samples and their distributions.

DEFINITION 2.1 (Bootstrap sample). A bootstrap sample $x^{\#} = (x_1^{\#}, x_2^{\#}, \ldots, x_n^{\#})$ is a ramdom sample of size n where each $x_i^{\#}$ is abtained with probability $1/n$ by drawing with replacement from the original sample $x = (x_1, x_2, \ldots, x_n)$.

DEFINITION 2.2 (Bootstrap distribution). Let $\theta_n^{\#i} = \theta^{\#}\left(X_1^{\#i}, X_2^{\#i}, \ldots, X_n^{\#i}\right)$ denote a random bootstrap sample, $(i = 1, \ldots, B)$. The function $G^{\#}(t)$, $(-\infty < t < \infty)$, defined by

$$G^{\#}(t) = \mathbb{P}\left(\theta_n^{\#} < t\right) = \frac{\text{number of } \left\{\theta_n^{\#i} < t\right\}}{B} \tag{2.1}$$

is called the empirical bootstrap distribution.

2.1. Standard error. Let us collect many independent samples of the same size from the same population. For each sample we compute the value t_n of statistics

$\theta_n = \theta(X_1, X_2, \ldots, X_n)$. Then, the following question arises: If we take many samples, how do the values t_n change?

Specifically, if we take N samples from population, then we will have N values t_n^i, $(i = 1, \ldots, N)$. The standard deviation of these N values t_n^i is called *the standard error* and denoted by

$$se(\theta_n) = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} \left(t_n^i - \bar{t}_n\right)^2} \tag{2.2}$$

where $\bar{t}_n = \frac{1}{N} \sum_{i=1}^{N} t_n^i$. Therefore the standard error measures the magnitude of variability of t_n^i.

In many settings, we have no models for population. We then can't appeal to probability theory, and we also can't afford to actually take many samples. In applying the bootstrap method we first take one sample, then we have as thought it was the population and then, we take resamples from it to contructs the bootstrap distribution. The following steps are important:

Step 1: Generate B bootstrap samples $x^{\#1}, x^{\#2}, \cdots, x^{\#B}$.

Step 2: For each bootstrap sample , compute $t_n^{\#i} = \theta\left(x_1^{\#i}, \ldots, x_n^{\#i}\right)$.

Step 3: The bootstrap estimate of the standard error is

$$se^{\#}(\theta_n) = \sqrt{\frac{1}{B-1} \sum_{i=1}^{B} \left(t_n^{\#i} - \overline{t_n^{\#}}\right)^2} \tag{2.3}$$

where $\overline{t_n^{\#}} = \frac{1}{B} \sum_{i=1}^{B} t_n^{\#i}$.

2.2. The bootstrap t interval. Let θ is a parameter of interest and $\hat{\theta}$ is a plug in estimate of θ. In addition to point estimate $\hat{\theta}$, we may also be interested in constructing an interval to estimate θ with a desired confidence level. If α is a number between 0 and 1, typically it is taken as 0.01, 0.05, or 0.1,. A $(1 - \alpha) \times$

100% confidence interval can be as the following

$$\left(\hat{\theta} - z(1-\alpha/2)\cdot \hat{se}; \hat{\theta} + z(\alpha/2)\cdot \hat{se}\right) \tag{2.4}$$

where $\hat{se}$ can be either a bootstrap estimate or any other reasonable estimate of standard error of $\hat{\theta}$. And $z(\alpha/2)$ and $z(1-\alpha/2)$ are $100\cdot(\alpha/2)$ and $100\cdot(1-\alpha/2)$ percentiles, respectively, of the distribution of random variable $Z = (\hat{\theta}-\theta)/\hat{se}$. Note that the random variable Z used here may not necessarily have a standard normal distribution.

Whenever the normality holds, $z(\alpha/2)$ and $z(1-\alpha/2)$ values can be replaced by the standard scores from the standard normal table. For instance, $z(0.025) = -1.96$ and $z(0.975) = 1.96$. And thus, 95% confidence interval for θ will be constructed as $\left(\hat{\theta} - 1.96\cdot \hat{se}; \hat{\theta} + 1.96\cdot \hat{se}\right)$.

When Z can't be assumed to be standard normal or a t-distribution, the bootstrap can be used to obtain an accurate interval. Here is the produce:
Step 1: Generate B bootstrap sample $x^{\#1}, x^{\#2}, \cdots, x^{\#B}$.
Step 2: For each bootstrap sample i, compute $\hat{\theta}^{\#i} = \theta\left(x_1^{\#i}, \ldots, x_n^{\#i}\right)$ and the estimated standard error of $\hat{\theta}^{\#i}$ denoted by $\hat{se}^{\#i}$

$$Z^{\#i} = \frac{\hat{\theta}^{\#b} - \hat{\theta}}{\hat{se}^{\#i}} \tag{2.5}$$

Note that when $\hat{\theta}$ is not a sample mean but a more complicated statistics, bootstrap resampling may be used to estimate $\hat{se}^{\#i}$. for each bootstrap sample i. this resuals in a nested bootstrap resampling.
Step 3: The $\alpha/2$ quantile of $Z^{\#i}$ is estimated by the value $z(\alpha/2)$ such that

$$\frac{\#\left\{Z^{\#i} < z(\alpha/2)\right\}}{B} = \frac{\alpha}{2} \tag{2.6}$$

and the value $z(1-\alpha/2)$ such that

$$\frac{\#\left\{Z^{\#i} < z(1-\alpha/2)\right\}}{B} = 1 - \frac{\alpha}{2} \tag{2.7}$$

Step 4: Constructing the bootstrap t $(1-\alpha)\cdot 100\%$ confidence intervals:

$$\left(\hat{\theta} - z(1-\alpha/2)\cdot \hat{se}; \hat{\theta} - z(\alpha/2)\cdot \hat{se}\right) \tag{2.8}$$

2.3. The bootstrap percentiles. The interval between the $\alpha/2$ th and $(1-\alpha/2)$ th percentiles of the bootstrap distribution of a staitstics is a $(1-\alpha)\%$ bootstrap percentile confidence interval for corresponding parameter.

3. REGRESSION MODELS

Bootstrap resampling for regression models is a generalization of the bootstrap process described above. Rather than sample scalars, we sample a vector of value for each observation and compute the regression coefficient estimator for each bootstrap sample. Consider the regression model, $Y = X\beta + \varepsilon$, where X is an $n \times (p+1)$ matrix of the explanatory variables (including a column of one for the constant term), β is a $(p+1)\times 1$ vector of population regression coefficients, ε is an $n \times 1$ vector.

With standard method, if we want to make any confidence intervals or perform any hypothesis tests, we will need to assume distributional form for the errors ε. The usual assumption is that the errors are normally distributed and in practice this is often, although not always, a reasonable assumption. We can reduce this assumption by use bootstrap method. The bootstrap estimates of the standard deviations of the coefficient estimates are

$$se^{\#}\left(\hat{\beta}_j\right) = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\beta}_j^{\#i} - \bar{\hat{\beta}}_j^{\#}\right)^2}, \quad j = 0, \dots, p \tag{3.1}$$

where $\hat{\beta}_j^{\#i}$ the value of the bootstrap estimator for β_j in the i th sample, and $\bar{\hat{\beta}}_j^{\#}$ is mean of bootstrap estimates for B bootstrap sample. The bootstrap is also useful in forming confidence interval. The simplest nonparametric bootstrap confidence interval is known as the percentile interval.

4. AN APPLICATION OF BOOTSTRAP METHODS IN ESTIMATING THE OSTEOPOROSIS FOR VIETNAMESE WOMEN

In this section we will discuss the application of bootstrap method in estimating the bone mineral density of Vietnameses women. A bone mineral density test (bmd) is the best way to determine bone health of women after menopause. People who have low bmd have hight risk of fracture. Every standard deviation decreases in bmd then risk of fracture increase from 2 to 3 times. Osteoporosis is most common in women after menopause, when it is called postmenopausal osteoporosis bone mineral density tests are performed to determine whether a patient has osteoporosis or osteopenia, a low bone mass that puts her at risk for osteoporosis. To make this determination, the technologist will calculate the patient's T-score. The World Health Organization (WHO) established the following criteria for determining the T-score:

$$T - score = \frac{bmd - bmdp}{sd} \tag{4.1}$$

where $bmdp$ is peak bone mineral density of Vietnameses women and sd is standard deviation of peak bone mineral density of Vietnameses women. The WHO's report defined diagnostic categories based on bmd measurements as follows:

- Normal: T-score above -1.
- Osteopenia: T-score between -1 and -2.5.
- Osteoporosis: T-score at or below -2.5.

Our main purpose is to estimate the $bmdp$ and sd using the data provided by a group of medical doctors. First, we explore the relationship between bmd and age of Vietnameses women. The model that we choose has the form

$$bmd_i = \beta_0 + \beta_1 age_i^1 + \beta_2 age_i^2 + \beta_3 age_i^3 + \varepsilon_i, \quad i = 1, \ldots, n \tag{4.2}$$

where n is the number of observations. Here we use **R** for statistical data analysis. The OLS estimator of the regression coefficients:

```
>reg3 <- lm(bmd ~ age + I(age^2) + I(age^3))
>summary(reg3)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.004e-01  7.323e-02   5.468 9.17e-08 ***
age          4.499e-02  7.691e-03   5.849 1.22e-08 ***
I(age^2)    -1.155e-03  2.327e-04  -4.963 1.13e-06 ***
I(age^3)     8.159e-06  2.128e-06   3.834 0.000152 ***
---
Signif. codes:0'***' 0.001'**' 0.01'*' 0.05'.' 0.1' ' 1
```

FIGURE 1. The relationship between bmd and age

Next , we use bootstrap method to estimate standard errors and interval confidence for the regression coefficients. The table 1 bellow gives a comparison of standard errors of the two methods.

TABLE 1. Comparison of standard errors

Coefficient	OLS method $se\left(\hat{\beta}_i\right)$	Bootstrap method $se^{\#}\left(\hat{\beta}_i\right)$
β_0	7.323e-02	7.227e-02
β_1	7.691e-03	7.337e-03
β_2	2.327e-04	2.341e-04
β_3	2.128e-06	2.063e-06

For each coefficient, the OLS standard errors are approximately equal to the bootstrap standard errors. This follows from the fact that the bootstrap histogram is of the standard normal shape. The Table 1 and Table 2 give the point estimate and interval confidence estimate for each coefficient from the two methods.

Classical confidence intervals are constructed under the assumption that the distribution of coefficient estimator is symmetic, i.e. the upper and lower bounds are of the same distance from the coefficient estimate. The bootstrap confidence interval can capture asymmetries in the distribution of the estimator so that the lower

TABLE 2. Compare point estimate

Coefficient	OLS method	Bootstrap method
β_0	4.00e-01	4.01e-01
β_1	4.49e-02	4.49e-01
β_2	-1.15e-03	-1.15e-03
β_3	8.15e-06	8.13e-06

TABLE 3. Comparison of interval confidence estimate

Coefficient	OLS method	Bootstrap method
β_0	(2.56e-01; 5.44e-01)	(2.74e-01; 5.30e-01)
β_1	(2.99e-02; 6.01e-02)	(3.00e-01; 5.97e-02)
β_2	(-1.61e-03; -6.97e-04)	(-1.58e-03; -7.02e-04)
β_3	(3.97e-06; 1.23e-05)	(4.06e-06; 1.21e-05)

bound can be further or closer to the coefficient estimate than the upper bound. In this case, the bootstrap confidence intervals are similar to the OLS confidence intervals because the distributions of bootstrap estimates have a normal shape.

FIGURE 2. Bootstrap distribution of regression coefficients

With each value $A = age$ we compute the value of $B = bmd$ as below

$$B = \hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 A^2 + \hat{\beta}_3 A^3$$

age have peak bone mineral density is

$$A_{\max} = \frac{-\hat{\beta}_2 - \sqrt{\hat{\beta}_2^2 - 3\hat{\beta}_1\hat{\beta}_3}}{3\hat{\beta}_3} \tag{4.3}$$

and peak mineral density is

$$B_{\max} = \hat{\beta}_0 + \hat{\beta}_1 A_{\max} + \hat{\beta}_2 A_{\max}^2 + \hat{\beta}_3 A_{\max}^3 \tag{4.4}$$

Recall that, we want estimate standard deviation of $A_{\max}$. Bootstrap estimator for standard deviation of $A_{\max}$ is

$$sd = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left(A_{\max}^{\#i} - \bar{A}_{\max}^{\#}\right)^2} \tag{4.5}$$

where $A_{\max}^{\#i}$ is the value of the bootstrap estimator for $A_{\max}$ in the i th sample. Here is **R** code for estimator sd.

```
>setwd("C:/")
>data<-read.table("data.txt",header=T,na.strings=".")
>attach(data)
#determine sample size
>n<- length ( age)
>B < -100000 #Number  bootstrap
#create new object to store  coefficients
>beta0 <- numeric (B)
>beta1 <- numeric (B)
>beta2 <- numeric (B)
>beta3 <- numeric (B)
# resampling
>for (i in 1:B)
{ Resample <- Data[ sample (1:n, n, replace =T), ]
y <- Resample [, " bmd "]
x <- Resample [, " age "]

fix <- lm(y ~ x+I(x ^2)+I(x ^3))
beta0 [i] <- fix$coefficients[1]
beta1 [i] <- fix$coefficients[2]
beta2 [i] <- fix$coefficients[3]
beta3 [i] <- fix$coefficients[4]
}
>A.max<- (-beta2-sqrt(beta2^2 - 3*beta3*beta1))
        /(3*beta3)
>B.max <- beta0 + beta1*A.max + beta2*A.max^2
        + beta3*A.max^3
>sd(B.max) #The result that we need is
[1] 0.01299935
>mean(B.max)
[2] 0.933978
```

so the $T - scpre$ can be compute by

$$T - score = \frac{bmd - 0.9339}{0.013} \tag{4.6}$$

5. A CONCLUSION.

In applied statistics, the estimation methods are usually that as maximum likelihood estimation, non- parametric estimation,. . . . The advantage of the bootstrap method we explore here does not need any additional assumption of distributions and it can be used in solving problems that in the past were regarded as unsolvability.

REFERENCES

[1] ichaelr. Chernick. (2007). *Bootstrap Methods:A Guide for Practitioners and Researchers*. A John Wiley & Sons, Inc., Publication.

[1] Bradley Efron. (1994). *An Introduction to the Bootstrap*. Chapman & Hall/CRC, Inc., Publication.

[2] Phillip Good. (2004). *Permutation, Parametric and Bootstrap Tests of Hypotheses*. Springer Publication.

[3] F.M. Dekking and C. Kraikamp. (2007). *A Modern Introduction to Probability and Statistics*. Springer Publication.

[4] John Bibby and Helge Toutenburg. (1977). *Prediction and Improved Estimation In Linear Models*. A John Wiley & Sons, Inc., Publication.

[5] Roger W. Johnson. (2001). *An Introduction To The Bootstrap*. Teaching Statistics. 2001; 23: 49 - 54.

[6] Chris Ricketts and John Berry. *Teaching Statistics Through Resampling*. Center for Teaching Mathematics, University of Plymouth, UK.

[7] Jason S Haukoos and Roger J Lewis. (2005). *Advanced Statistics: Bootstrapping Confidence Intervals For statistics with "Difficult" Distributions*. Academic Emergency Medicine. Apr 2005; 12, 4: 360 - 365; ProQuest Medical Library.

[8] James Carpenter and John Bithell. (2000). *Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians*. Statist. Med; 19:1141 - 1164.

[9] Kenneth A. Bollen and Robert Stine. (1990). *Direct and Inderect Effects: Classical and Bootstrap Estimates of Varuability*. Sociological Methodology; 20: 115 - 140.

Nguyen Van Thu
Department of Mathematics, International University, HCM City
E-mail: nvthu@hcmiu.edu.vn

Nguyen Van Tuan
Garvan Institute of Medical Research Sydney, Australia
E-mail: t.nguyen@garvan.org.au

Nguyen Duc Phuong
Department of Mathematics, University of Natural sciences, HCM City
E-mail: nducphuong@gmail.com

Received on April - 04 - 08;
last revised version on xx.xx.xxxx