Bootstrap for regression lab

Shaoran Sun

May 3, 2016

The bootstrap for simple linear regression

1

The estimated slope is the relationship between SAT and college GPA: for every unit of increasement of SAT, there is 0.0018 unit of increasement of college GPA.

2

When the index is in the form of a decimal point like: "x.y, or 77.2", then the value is repeated more than once.

3

The 95% bootstrap interval for the slope of the population regression line is **(0.001076773, 0.002550616)**

Duality between confidence intervals and hypothesis tests

4

Less than 0.05, because 0 is not within the interval.

So we can reject the null hypothesis, and conlcude that $\beta_1 \neq 0$

5

The 99% bootstrap interval for the slope of the population regression line is **(0.0008904364, 0.002740267)**

Less than 0.01, because 0 is not within the interval.

So we can reject the null hypothesis, and conlcude that $\beta_1 \neq 0$

6

When theta is one of the bounds of the interval.

The bootstrap for multiple regression

7

CollGPA = 0.00123*SAT+0.376*HSGPA+0.0227*Rec

8

For SAT: (0.0005905682, 0.001961265)

For HSGPA: (0.1399587, 0.6003998)

For Rec: (-0.09318261, 0.1284393)

9

p-values for both thetas of SAT and HSGPA are less than 0.05, because 0 is not within the interval. So we can reject the null hypothesis, and conlcude that

$$\beta_1 \neq 0$$

$$\beta_2 \neq 0$$

p-value for theta of Rec, however, is greater than 0.05, because 0 is within the interval. So we can fail to reject the null hypothesis, and conlcude that

$$\beta_3 = 0$$

10

For $\beta_1 \& \beta_2$, no matter how wide/loose the bootstrap interval is, 0 is never within the interval.

For β_3, no matter how narrow/strict the bootstrap interval is, 0 is always within the interval.

Lab Summary

We have the following result initially:

CollGPA = 0.00123*SAT+0.376*HSGPA+0.0227*Rec

After the bootstrap interval test:

p-values for both thetas of SAT and HSGPA are less than 0.05. So we can conlcude that $\beta_1 \neq 0 \& \beta_2 \neq 0$

p-value for theta of Rec, however, is greater than 0.05. So we can conlcude that $\beta_3 = 0$

Hence, we can conlcude that variable *Rec* is not significant, and SAT and HSGPA are significant. We have very strong evidence that SAT is significant, and relatively strong evidence that HSGPA is significant.

Appendix

1

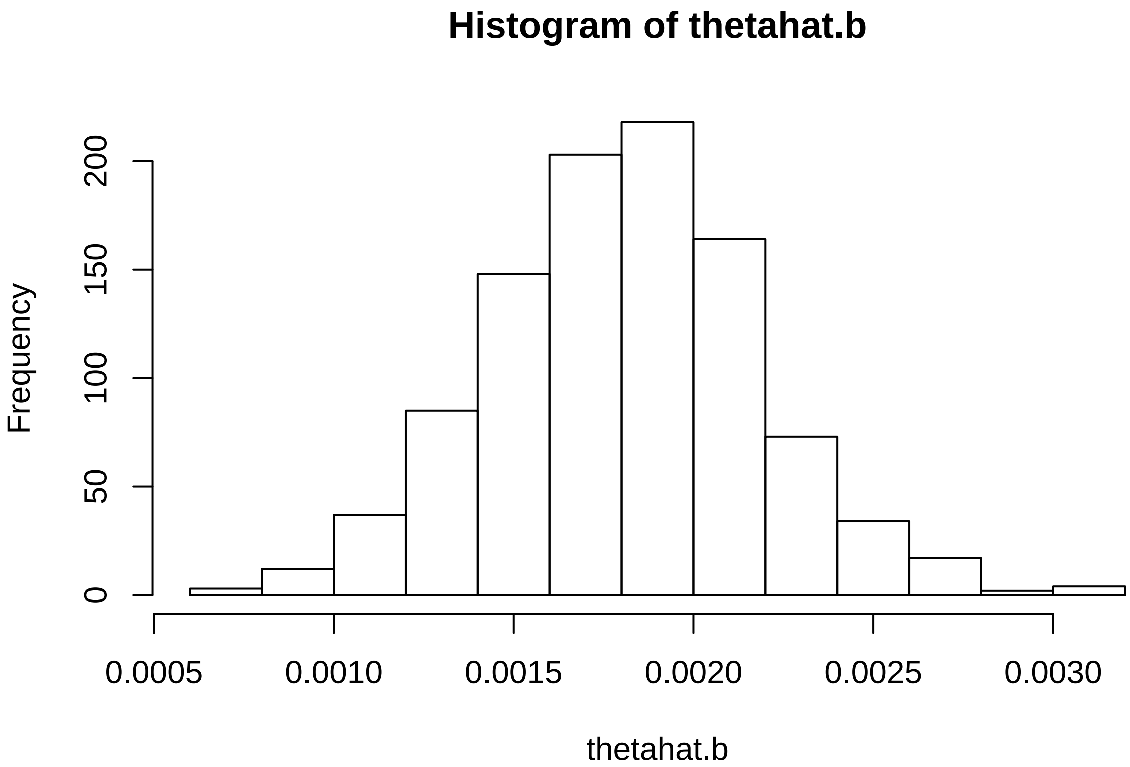
```r
### create our data and calculate thetahat, the slope of the regression line
GPA <- read.delim("~/Desktop/3480 Lab/GPA.txt")
oursample = GPA
regr=lm(CollGPA ~ SAT, data=oursample)
summary(regr)
```

```
##
## Call:
## lm(formula = CollGPA ~ SAT, data = oursample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.46421 -0.47381 -0.00147  0.29138  1.69570
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.1518923  0.3079925   0.493    0.623
## SAT         0.0018020  0.0002968   6.071 2.42e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.642 on 98 degrees of freedom
## Multiple R-squared:  0.2733, Adjusted R-squared:  0.2659
## F-statistic: 36.85 on 1 and 98 DF,  p-value: 2.417e-08
```

```r
thetahat = lm(CollGPA ~ SAT, data=oursample)$coeff[2]
thetahat
```

```
##        SAT
## 0.00180201
```

```r
thetahat.b = rep(NA,1000)
for (i in 1:1000) {
### draw the bootstrap sample and calculate thetahat.b
index = 1:100
bootindex = sample(index, 100, replace=T)
bootsample = oursample[bootindex,]
thetahat.b[i] = lm(CollGPA ~ SAT, data=bootsample)$coeff[2]
}
hist(thetahat.b)
```




```
quantile(thetahat.b, .025); quantile(thetahat.b, .975)
```

```
##        2.5%
## 0.001089233
```

```
##       97.5%
## 0.002566345
```

2

```
### draw the bootstrap sample
index = 1:100
bootindex = sample(index, 100, replace=T)
bootsample = oursample[bootindex,]
bootsample
```

```
##       CollGPA  SAT
## 17       1.01  954
## 7        1.38 1058
## 47       1.99 1126
## 30       2.03 1170
## 58       1.29  778
## 85       0.95  948
## 24       2.48 1150
## 17.1     1.01  954
## 83       0.22  704
```

```
## 89       3.02 1374
## 11       1.50  896
## 75       2.39 1084
## 24.1     2.48 1150
## 41       1.68 1256
## 4        1.10  706
## 55       2.35  604
## 19       2.00 1046
## 36       1.57 1038
## 13       1.90  958
## 96       1.80  772
## 61       1.90  950
## 37       2.46 1090
## 18       3.66 1500
## 9        1.38 1104
## 95       2.39 1134
## 96.1     1.80  772
## 90       1.85 1014
## 63       3.30  956
## 87       1.86 1000
## 36.1     1.57 1038
## 9.1      1.38 1104
## 58.1     1.29  778
## 34       1.48 1180
## 56       2.82  854
## 55.1     2.35  604
## 40       2.04 1114
## 4.1      1.10  706
## 19.1     2.00 1046
## 50       2.31 1214
## 47.1     1.99 1126
## 47.2     1.99 1126
## 36.2     1.57 1038
## 14       3.11 1246
## 21       2.60 1198
## 31       3.50 1034
## 41.1     1.68 1256
## 11.1     1.50  896
## 35       1.54  952
## 96.2     1.80  772
## 15       1.92 1106
## 19.2     2.00 1046
## 22       2.55  940
## 44       2.10 1222
## 92       2.15  400
## 30.1     2.03 1170
## 56.1     2.82  854
## 7.1      1.38 1058
## 66       1.68 1168
## 66.1     1.68 1168
## 86       1.99 1182
## 41.2     1.68 1256
## 39       2.11 1096
## 93       1.46  998
```

```
## 90.1    1.85 1014
## 84      2.31 1222
## 53      2.50 1116
## 20      2.05 1054
## 22.1    2.55  940
## 37.1    2.46 1090
## 72      3.09 1084
## 57      1.80  814
## 61.1    1.90  950
## 67      1.94  970
## 94      2.29  776
## 100     0.89  864
## 86.1    1.99 1182
## 15.1    1.92 1106
## 70      1.31 1232
## 34.1    1.48 1180
## 48      2.24 1158
## 2       2.56 1254
## 8       1.50 1008
## 71      1.68 1140
## 44.1    2.10 1222
## 14.1    3.11 1246
## 64      1.80 1352
## 12      1.29  848
## 94.1    2.29  776
## 83.1    0.22  704
## 24.2    2.48 1150
## 46      2.03  886
## 12.1    1.29  848
## 29      1.62  852
## 96.3    1.80  772
## 86.2    1.99 1182
## 83.2    0.22  704
## 50.1    2.31 1214
## 90.2    1.85 1014
## 36.3    1.57 1038
## 33      2.39 1018
```

5

```r
quantile(thetahat.b, .005); quantile(thetahat.b, .995)
```

```
##         0.5%
## 0.0009040265
```

```
##       99.5%
## 0.002897891
```

7

```r
GPAfull <- read.delim("~/Desktop/3480 Lab/GPAfull.txt")
full.reg=lm(CollGPA ~ SAT + HSGPA + Rec, data=GPAfull)
summary(full.reg)
```

```
##
## Call:
## lm(formula = CollGPA ~ SAT + HSGPA + Rec, data = GPAfull)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0979 -0.4407 -0.0094  0.3859  1.7606
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.1532639  0.3229381  -0.475 0.636156
## SAT          0.0012269  0.0003032   4.046 0.000105 ***
## HSGPA        0.3763511  0.1142615   3.294 0.001385 **
## Rec          0.0226843  0.0509817   0.445 0.657358
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5895 on 96 degrees of freedom
## Multiple R-squared:  0.3997, Adjusted R-squared:  0.381
## F-statistic: 21.31 on 3 and 96 DF,  p-value: 1.16e-10
```

8

```r
### create our data
oursample = GPAfull
SATthetahat = lm(CollGPA ~ SAT + HSGPA + Rec, data=oursample)$coeff[2]
HSGPAthetahat = lm(CollGPA ~ SAT + HSGPA + Rec, data=oursample)$coeff[3]
Recthetahat = lm(CollGPA ~ SAT + HSGPA + Rec, data=oursample)$coeff[4]
SATthetahat; HSGPAthetahat; Recthetahat
```

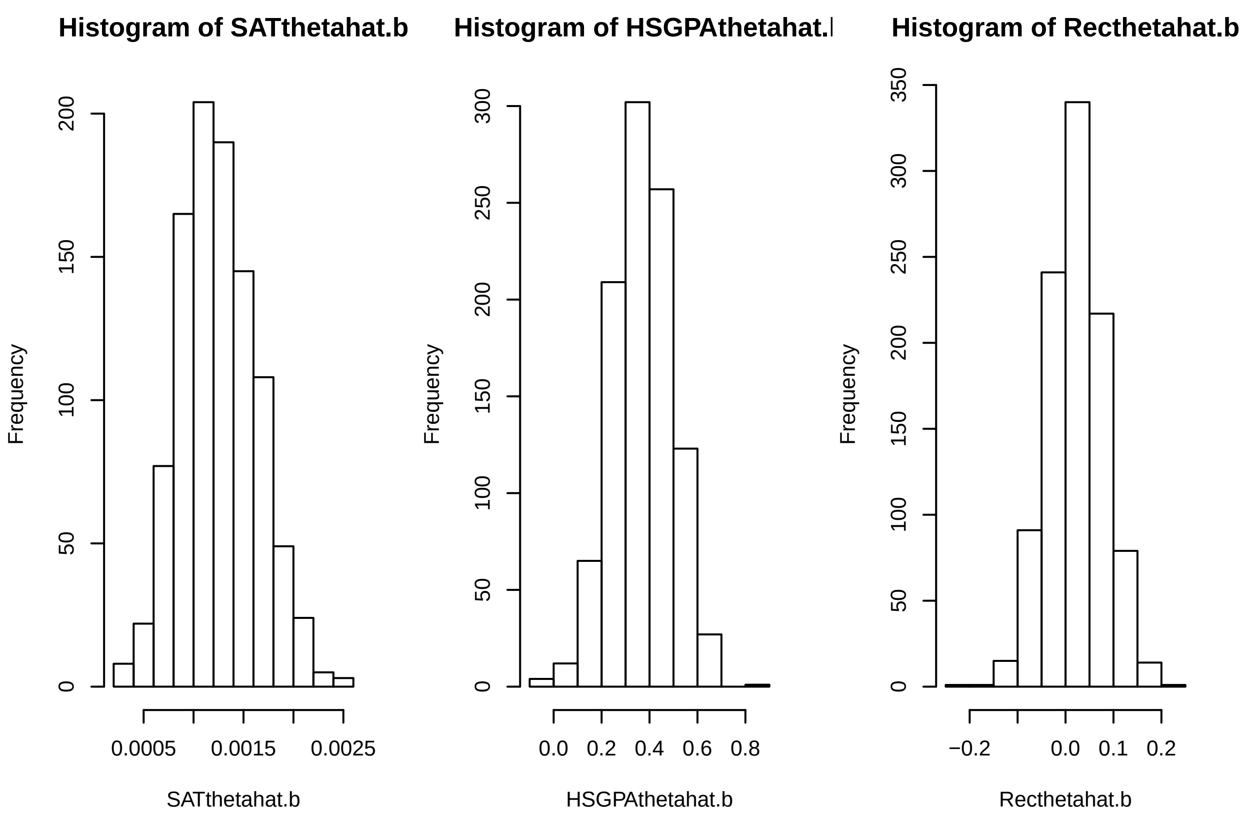
```
##        SAT
## 0.00122693

##     HSGPA
## 0.3763511

##        Rec
## 0.02268425
```

```r
SATthetahat.b = rep(NA,1000); HSGPAthetahat.b = rep(NA,1000); Recthetahat.b = rep(NA,1000)
for (i in 1:1000) {
### draw the bootstrap sample and calculate thetahat.b
index = 1:100
```

```
bootindex = sample(index, 100, replace=T)
bootsample = oursample[bootindex,]
SATthetahat.b[i] = lm(CollGPA ~ SAT + HSGPA + Rec, data=bootsample)$coeff[2]
HSGPAthetahat.b[i] = lm(CollGPA ~ SAT + HSGPA + Rec, data=bootsample)$coeff[3]
Recthetahat.b[i] = lm(CollGPA ~ SAT + HSGPA + Rec, data=bootsample)$coeff[4]
}
par(mfrow=c(1,3))
hist(SATthetahat.b); hist(HSGPAthetahat.b); hist(Recthetahat.b)
```




```
quantile(SATthetahat.b, .025); quantile(SATthetahat.b, .975)
```

```
##         2.5%
## 0.0005734486
```

```
##       97.5%
## 0.002041753
```

```
quantile(HSGPAthetahat.b, .025); quantile(HSGPAthetahat.b, .975)
```

```
##      2.5%
## 0.1268043
```

```
##     97.5%
## 0.6021838
```

```r
quantile(Recthetahat.b, .025); quantile(Recthetahat.b, .975)
```

```
##        2.5%
## -0.09245781
```

```
##     97.5%
## 0.1290182
```

10

```r
quantile(SATthetahat.b, .005); quantile(SATthetahat.b, .995)
```

```
##         0.5%
## 0.0003652806
```

```
##      99.5%
## 0.00233658
```

```r
quantile(HSGPAthetahat.b, .005); quantile(HSGPAthetahat.b, .995)
```

```
##       0.5%
## 0.04327649
```

```
##     99.5%
## 0.6465829
```

```r
quantile(Recthetahat.b, .05); quantile(Recthetahat.b, .95)
```

```
##          5%
## -0.07287221
```

```
##       95%
## 0.1203809
```