BOOTSTRAP *RESAMPLING* OBSERVASI PADA ESTIMASI PARAMETER REGRESI MENGGUNAKAN *SOFTWARE* R

Joko Sungkono*

Abstrak : Pada tulisan ini, algoritma metode bootstrap *resampling* observasi dipaparkan secara detail dalam mengestimasi parameter regresi linear ganda. Berdasarkan algoritma tersebut dirancang suatu body program dalam bahasa pemrograman *software* R.

Untuk melihat keakuratan metode bootstrap *resampling* observasi diberikan suatu simulasi monte carlo dengan membangkitkan data dari variabel random berdistribusi tertentu.

Kata kunci: bootstrap, *resampling* observasi, *software* R, regresi linear,simulasi.

PENDAHULUAN

Bootstrap adalah suatu metode yang dapat bekerja tanpa membutuhkan asumsi distribusi karena sampel asli digunakan sebagai populasi. Dalam Sahinler dan Topuz [4], Efron menyatakan bahwa bootstrap adalah teknik *resampling* nonparametrik yang bertujuan untuk menentukan estimasi standar eror dan interval konfidensi dari parameter populasi seperti mean, rasio, median, proporsi, koefisien korelasi atau koefisien regresi tanpa menggunakan asumsi distribusi. Bootstrap dapat digunakan untuk mengatasi permasalahan dalam statistika baik masalah data yang sedikit, data yang menyimpang dari asumsinya maupun data yang tidak memiliki asumsi dalam distribusinya. Bootstrap adalah suatu metode yang berbasis komputer yang sangat potensial untuk dipergunakan pada masalah keakurasian, [3].

Bootstrap diperkenalkan pertama kali oleh Efron tahun 1979. Bootstrap adalah metode yang didasarkan pada simulasi data untuk keperluan inferensi statistik, [3]. Metode bootstrap digunakan untuk mencari distribusi sampling dari suatu estimator dengan prosedur *resampling* dengan pengembalian dari data asli. Metode bootstrap dilakukan dengan mengambil sampel dari sampel asli dengan ukuran sama dengan ukuran sampel asli dan dilakukan dengan pengembalian. Kedudukan sampel asli dalam metode bootstrap dipandang sebagai populasi. Metode peyampelan ini biasa disebut dengan *resampling* bootstrap. Bootstrap juga sering digunakan untuk mengestimasi standar eror estimator dan interval konfidensi dari suatu parameter populasi yang tidak diketahui. Pada dasarnya teknik estimasi dengan metode *resampling* bootstrap menggunakan semua kemungkinan sampel yaitu n^n. Akan tetapi hal ini sangat sulit untuk dilakukan untuk n>7. Untuk keperluan perhitungan biasanya digunakan pendekatan simulasi, sehingga disebut simulasi bootstrap. Misalkan dimiliki sampel random berukuran n yaitu $X_1, X_2, X_3, \ldots, X_n$ yang diambil dari suatu populasi dan statistik $\hat{\theta}$ adalah estimasi untuk parameter θ berdasar sampel asli.

Berdasarkan uraian metode *resampling* bootstrap menurut Efron dan Tibshirani [3], prosedur *resampling* bootstrap dapat dituliskan sebagai

Progdi Pend. Matematika, FKIP Universitas Widya Dharma Klaten

1. mengkonstruksi distribusi empiris $\hat{F}_n$ dari suatu sampel dengan memberikan probabilitas 1/n pada setiap X_i dimana i = 1, 2,..., n
2. mengambil sampel bootstrap berukuran n secara random dengan pengembalian dari distribusi empiris $\hat{F}_n$, sebut sebagai sampel bootstrap pertama X^{*1}
3. menghitung statistik $\hat{\theta}$ yang diinginkan dari sampel bootstrap X^{*1}, sebut sebagai $\hat{\theta}_1^*$
4. mengulangi langkah 2 dan 3 hingga B kali, diperoleh $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$
5. 1.mengkontruksi suatu distribusi probabilitas dari $\hat{\theta}_b^*$ dengan memberikan probabilitas 1/B pada setiap $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$. Distribusi tersebut merupakan estimator bootstrap untuk distribusi sampling $\hat{\theta}$ dan dinotasikan dengan $\widehat{F}^*$.
6. pendekatan estimasi bootstrap untuk θ adalah mean dari distribusi $\widehat{F}^*$ yaitu

$$\hat{\theta}^* = \sum_{b=1}^{B} \hat{\theta}_b^* \frac{1}{B}$$

Pendekatan bootstrap jika diulang lebih dari satu kali akan memberikan hasil yang berbeda, hal ini karena yang dilakukan adalah suatu simulasi. Jika dapat dilakukan menggunakan semua kemungkinan sampel yaitu n^n maka hasilnya akan sama.

Secara teori, menurut Shao dan Tu [5], sifat asimtotis distribusi bootstrap mendekati distribusi sebenarnya. Pada penggunaannya, metode bootstrap harus dilakukan dengan bantuan komputer karena melibatkan perhitungan yang sangat banyak. Software – software statistik belum ada yang memberikan paket *resampling* bootstrap secara langsung, sehingga metode ini masih jarang digunakan oleh peneliti. Tulisan ini memberikan paket bootstrap *resampling* observasi secara detail dalam bahasa R yang disusun berdasarkan algoritma bootstrap *resampling* observasi pada kasus estimasi parameter regresi linear ganda.

ESTIMASI BOOTSTRAP UNTUK PARAMETER REGRESI LINEAR

Perlu diketahui bahwa metode bootstrap untuk regresi dapat dilakukan melalui *resampling* residual dan *resampling* observasi. Bootstrap *resampling* residual menggunakan software R telah dipaparkan secara detail dan berdasarkan simulasi memberikan pendekatan yang sangat akurat dalam estimasi parameter regresi [6]. Pada bagian ini akan dibahas metode bootstrap *resampling* observasi untuk estimasi parameter regresi linear. Tanpa mengurangi keumuman pembahasan diambil regresi linear ganda dengan satu variabel dependen dan dua variabel independen. Model regresi linear populasinya $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. Misalkan dimiliki sampel berpasangan antara variabel dependen dan independen yang dituliskan dalam bentuk matrik **Y** dan **X** dengan ukuran sampel n. Selanjutnya sampel ini disebut sampel asli. Pada dasarnya metode bootstrap *resampling* dapat dilakukan untuk semua kemungkinan sampel bootstrap yang dapat diambil yaitu n^n. Untuk n yang cukup besar misalkan n>7, maka perhitungan dengan semua kemungkinan sampel bootstrap sudah sulit dilakukan, sehingga digunakan pendekatan simulasi dengan B sampel bootstrap. Jika $w_i = (y_i, x_{1i}, x_{2i})$ menyatakan pasangan data sampel dmana i=1, 2, …, n, maka menurut Sahinler dan Topuz [4], prosedur bootstrap *resampling* observasi untuk estimasi parameter regresi dapat dituliskan sebagai berikut

1. Mengkonstruksikan distribusi empiris $\hat{F}_n$ dari $w_i, w_2, \ldots, w_n$.

2. Mengambil sampel berukuran n dengan pengembalian dari distribusi empiris $\hat{F}_n$, sebut $\boldsymbol{w}^{*1}$ sebagai sampel bootstrap pertama.
3. 1.Melakukan estimasi parameter regresi berdasarkan sampel bootstrap $\boldsymbol{w}^{*1}$

   $$\hat{\boldsymbol{\beta}}^{*1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}^{*1}$$
4. 1.Mengulangi langkah 2 dan 3 sebanyak B kali, sehingga diperoleh $\hat{\boldsymbol{\beta}}^{*1}, \hat{\boldsymbol{\beta}}^{*2}, \ldots, \hat{\boldsymbol{\beta}}^{*B}$.
5. 1.Mengkonstruksikan distribusi empiris dari $\hat{\boldsymbol{\beta}}^{*1}, \hat{\boldsymbol{\beta}}^{*2}, \ldots, \hat{\boldsymbol{\beta}}^{*B}$, yaitu $\hat{F}^*$
6. 1.pendekatan estimasi bootstrap untuk parameter regresi adalah mean dari distribusi empiris $\hat{F}^*$ yaitu

   $$\hat{\boldsymbol{\beta}}^* = \sum_{b=1}^{B} \hat{\boldsymbol{\beta}}^{*b} \frac{1}{B}$$

   Pada kasus ini karena terdapat dua variabel independen maka matrik $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$, $\hat{\boldsymbol{\beta}}^{*b} = (\hat{\beta}_0^{*b}, \hat{\beta}_1^{*b}, \hat{\beta}_2^{*b})$ dan $\hat{\boldsymbol{\beta}}^* = (\hat{\beta}_0^*, \hat{\beta}_1^*, \hat{\beta}_2^*)$.

Estimasi interval konfidensi bootstrap untuk parameter regresi diberikan dalam interval pendekatan normal dan interval persentil. Interval konfidensi bootstrap dengan pendekatan normal sebenarnya analog dengan interval konfidensi standar. Menurut Bennett [1], pemanfaatan metode bootstrap dalam mengkonstruksi interval ini adalah untuk menentukan standar eror dari estimator. Berdasarkan sampel bootstrap dengan replikasi B kali diperoleh $\hat{\boldsymbol{\beta}}^{*1}, \hat{\boldsymbol{\beta}}^{*2}, \ldots, \hat{\boldsymbol{\beta}}^{*B}$. Variansi estimator bootstrap $\hat{\beta}_k^*$ untuk k=0,1,2 diberikan oleh

$$\mathrm{V}(\hat{\beta}_k^*) = \sum_{b=1}^{B} (\hat{\beta}_k^{*b} - \hat{\beta}_k^*)^2/(B-1).$$

Standar eror bootstrap $Se_B(\hat{\beta}_k^*)$ diperoleh dari akar variansi. Interval konfidensi bootstrap pendekatan normal (1-α)x 100% untuk β_k diberikan oleh

$$\hat{\beta}_k + z_{\alpha/2} Se_B(\hat{\beta}_k^*) < \beta_k < \hat{\beta}_k + z_{1-\alpha/2} Se_B(\hat{\beta}_k^*).$$

Interval konfidensi bootstrap persentil didasarkan pada distribusi estimator bootstrap. Untuk setiap k, dibentuk distribusi empiris untuk $\hat{\beta}_k^{*1}, \hat{\beta}_k^{*2}, \ldots, \hat{\beta}_k^{*B}$ misalkan $\hat{F}^*$. Dari distribusi ini dapat dihitung nilai persentil yang merupakan ide dasar konstruksi interval konfidensi bootstrap persentil. Interval konfidensi bootstrap persentil (1-α)x 100% untuk β_k diberikan oleh

$$\left((\hat{F}^*)^{-1}(\alpha/2),\ (\hat{F}^*)^{-1}(1-\alpha/2)\right)$$

dengan $(\hat{F}^*)^{-1}(\alpha/2)$ adalah persentil ke 100x($\alpha/2$) dan $(\hat{F}^*)^{-1}(1-\alpha/2)$ merupakan persentil ke 100x($1-\alpha/2$) dari distribusi $\hat{F}^*$.

Algoritma bootstrap *resampling* observasi pada kasus estimasi parameter regresi dibuat dalam bahasa pemrograman R dengan nama fungsi "reg" dan diberikan sebagai berikut.

```
> fix(reg)
```

Kemudian pada fungsinya diisikan sebagai berikut.

```
function(y, x1, x2,B,a)
{
n<-length(y)
j<-seq(1:n)
b0 <- matrix(coef(lm(y ~ x1 + x2)), nrow = 1, ncol = 3)
b<-matrix(0,nrow=B,ncol=3)
is<-matrix(sample(j,n*B,replace=T),nrow=B, byrow=T)
ystar<-matrix(y[is],nrow=B)
x1star<-matrix(x1[is],nrow=B)
```

```
x2star<-matrix(x2[is],nrow=B)
for(i in 1:B){
b[i,]<-coef(lm(ystar[i,]~x1star[i,]+x2star[i,]))
}
bboot<-apply(b,2,mean)
l0<- b0[1] + qnorm(a/2) * sqrt(var(b[,1]))
l1<- b0[, 2] + qnorm(a/2) * sqrt(var(b[,2]))
l2 <- b0[, 3] + qnorm(a/2) * sqrt(var(b[,3]))
u0 <- b0[, 1] + qnorm(1-a/2) * sqrt(var(b[,1]))
u1 <- b0[, 2] + qnorm(1-a/2) * sqrt(var(b[,2]))
u2 <- b0[, 3] + qnorm(1-a/2) * sqrt(var(b[,3]))
int1 <- matrix(c(l0, u0, l1, u1, l2, u2), nrow = 3, byrow
= T)
lp0 <- quantile(b[,1], a/2)
lp1 <- quantile(b[,2], a/2)
lp2 <- quantile(b[,3], a/2)
up0 <- quantile(b[,1], 1-a/2)
up1 <- quantile(b[,2], 1-a/2)
up2 <- quantile(b[,3], 1-a/2)
int2 <- matrix(c(lp0, up0, lp1, up1, lp2, up2), nrow =
3, byrow= T)
list(est.boot=bboot,int.normal=int1,int.pct=int2)
}
```

SIMULASI

Untuk menjalankan program *resampling* bootstrap observasi di R ini diberikan simulasi dengan membangkitkan data melalui simulasi monte carlo.. Misalkan populasi terdiri dari 30 data dari untuk y, x_1 dan x_2 masing-masing sebagai berikut.

```
> x1<-rnorm(30,4,1.5)
> x2<-rnorm(30,5,1.5)
> e<-rnorm(30)
> y<-3+7*x1+x2+e
```

Berdasarkan data ini kita sudah mengetahui nilai parameter populasinya $\beta_0, \beta_1, \beta_2$ berturut-turut 3, 7 dan 1. Selanjutnya untuk melihat keakuratan metode bootstrap akan dilakukan estimasi berdasarkan data sampel dengan replikasi 1000 dan 2000. Estimasi parameter regresi menggunakan *resampling* bootstrap dengan replikasi B=1000 dan tingkat kepercayaan interval 95% diberikan oleh

```
> reg(y,x1,x2,1000,0.05)
est.boot
[1] 3.0006930    7.0196876    0.9618854
int.normal
        [,1]         [,2]
[1,] 1.0808995    4.729977
[2,] 6.7714915    7.280731
[3,] 0.6586942    1.287724
int.pct
        [,1]         [,2]
[1,] 1.1815665    4.901583
[2,] 6.7644155    7.260796
[3,] 0.6249278    1.269918
```

Estimasi parameter regresi menggunakan *resampling* bootstrap dengan replikasi B=2000 dan tingkat kepercayaan interval 95% diberikan oleh

```
> reg(y,x1,x2,2000,0.05)
est.boot
[1] 2.9525762   7.0170010   0.9711647
int.normal
        [,1]         [,2]
[1,] 1.1332022   4.677674
[2,] 6.7748625   7.277360
[3,] 0.6714893   1.274929
int.pct
        [,1]         [,2]
[1,] 1.234168    4.803348
[2,] 6.761918    7.237729
[3,] 0.668859    1.266203
```

Berdasarkan output R yang diperoleh dapat diberikan ringkasan hasil dalam empat tempat desimal sebagaimana dalam Tabel 1 dan Tabel 2 berikut.

Tabel 1. Simulasi Bootstrap dengan Replikasi 1000

Parameter	Estimasi	Int. Pendekatan Normal		Int. Persentil	
		BB	BA	BB	BA
β_0	3,0007	1,0809	4,7299	1,1816	4,9016
β_1	7,0197	6,7715	7,2807	6,7644	7,2608
β_2	0,9619	0,6587	1,2877	0,6249	1,2699

Tabel 2. Simulasi Bootstrap dengan Replikasi 2000

Parameter	Estimasi	Int. Pendekatan Normal		Int. Persentil	
		BB	BA	BB	BA
β_0	2,9526	1,1332	4,6777	1,2342	4,8033
β_1	7,0170	6,7749	7,2774	6,7619	7,2377
β_2	0,9712	0,6715	1,2749	0,6689	1,2662

Berdasarkan Tabel 1 dan Tabel 2, estimasi bootstrap untuk parameter regresi β_0, β_1, β_2 berdasarkan *resampling* bootstrap observasi cukup dekat dengan parameter populasinya. Sedangkan estimasi interval konfidensi baik interval bootstrap pendekatan normal maupun interval bootstrap persentil memberikan hasil yang hampir sama dan keduanya memuat parameter populasi dengan range yang cukup sempit.

SIMPULAN

Berdasarkan uraian pembahasan di atas terdapat beberapa poin penting yang dapat disimpulkan. Program bootstrap *resampling* observasi menggunakan R disusun berdasarkan algoritma bootstrap *resampling* observasi yang telah diuraikan. Berdasarkan studi simulasi, metode bootstrap *resampling* observasi dapat digunakan sebagai metode alternatif yang memberikan hasil estimasi parameter regresi yang sangat dekat dengan parameter populasi. Estimasi interval juga memberikan interval konfidensi yang memuat parameter populasi dan dengan range interval yang cukup sempit. Hal ini menunjukkan bahwa metode bootstrap *resampling* observasi memiliki keakuratan yang tinggi.

DAFTAR RUJUKAN

Bennett, P. J., 2009, *Introduction to the Bootstrap and Robust Statistics*. Winter Term, PSY711/712.

Bickel, P. J. and Freedman, D. A., 1981, *Some Asymtotic Theory for the Bootstrap*, Ann. Statist., no. 6, 9, 1196–1217.

Efron, B. and Tibshirani, R. J., 1993, *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Sahinler, S. and Topuz, D., 2007, *Bootstrap and Jackknife Resampling Algorithms for Estimation of Regression Parameters*, JAQM, no. 2, **2**, 188-199.

Shao, J. and Tu, D., 1995, *The jacknife and bootstrap*, Springer Verlag Inc., New York.

Sungkono, J., 2013, *Resampling bootstrap pada R*, Magistra, No.84 TH. XXV, ISSN 0215-9511.