

Σημειώσεις για την εφαρμογή Bootstrap

1. Βασική ιδέα του Bootstrap

Το **bootstrap** είναι μια μέθοδος επαναδειγματοληψίας που χρησιμοποιείται για να προσεγγίσουμε τη δειγματική κατανομή ενός εκτιμητή, χωρίς να χρειάζεται να γνωρίζουμε θεωρητικά την κατανομή του.

Η βασική ιδέα είναι ότι χρησιμοποιούμε το αρχικό δείγμα ως «πληθυσμό» και δημιουργούμε πολλά νέα δείγματα από αυτό, κάνοντας **δειγματοληψία με επανάθεση**.

Δηλαδή, κάθε bootstrap δείγμα:

- έχει το ίδιο μέγεθος με το αρχικό δείγμα,
- επιλέγεται από τις αρχικές παρατηρήσεις,
- επιτρέπει την επανάληψη της ίδιας παρατήρησης περισσότερες από μία φορές,
- μπορεί να μην περιλαμβάνει κάποιες αρχικές παρατηρήσεις.

2. Bootstrap για τον μέσο της μεταβλητής mpg

Στην πρώτη εφαρμογή χρησιμοποιείται το dataset mtcars και συγκεκριμένα η μεταβλητή mpg, η οποία μετρά τα μίλια ανά γαλόνι καυσίμου.

Αρχικά υπολογίζεται ο δειγματικός μέσος της mpg:

```
mean(mpg)
```

Στη συνέχεια ορίζεται ο αριθμός των bootstrap επαναλήψεων:

B = 1000

Αυτό σημαίνει ότι θα δημιουργηθούν 1000 bootstrap δείγματα. Για κάθε επανάληψη επιλέγονται τυχαία, με επανάθεση, n παρατηρήσεις από το αρχικό δείγμα:

```
temp = sample(1:n, n, replace = TRUE)  
boot_sample = mpg[temp]
```

Έπειτα υπολογίζεται ο μέσος του κάθε bootstrap δείγματος:

```
results[i] = mean(boot_sample)
```

Στο τέλος, το διάνυσμα results περιέχει 1000 bootstrap εκτιμήσεις του μέσου. Αυτές οι τιμές χρησιμοποιούνται για να προσεγγίσουμε τη δειγματική κατανομή του μέσου.

Το ιστόγραμμα:

```
hist(results, freq = FALSE)  
lines(density(results), col = "red")
```

δείχνει τη bootstrap κατανομή του μέσου της mpg.

3. Bootstrap για συντελεστές OLS

Στη δεύτερη εφαρμογή εφαρμόζεται bootstrap σε ένα απλό γραμμικό υπόδειγμα:

```
ols_boot = lm(mpg ~ wt, data = mtcars)
```

Το υπόδειγμα εξετάζει τη σχέση ανάμεσα στην κατανάλωση καυσίμου, mpg, και το βάρος του αυτοκινήτου, wt.

Η αρχική εκτίμηση δίνει δύο βασικούς συντελεστές:

```
beta0 = coefficients(ols_boot)[1]
```

```
beta1 = coefficients(ols_boot)[2]
```

όπου:

- beta0 είναι η εκτιμημένη σταθερά,
- beta1 είναι ο εκτιμημένος συντελεστής της μεταβλητής wt.

Στη συνέχεια δημιουργούνται 1000 bootstrap δείγματα, αυτή τη φορά όχι μόνο από μία μεταβλητή αλλά από ολόκληρες γραμμές του dataset:

```
smpl = sample(1:n, n, replace = TRUE)
```

```
temp_data = mtcars[smpl, ]
```

Για κάθε bootstrap dataset εκτιμάται ξανά το ίδιο υπόδειγμα:

```
temp_model = lm(mpg ~ wt, data = temp_data)
```

```
results[i, ] = coef(temp_model)
```

Έτσι, ο πίνακας results περιέχει 1000 ζεύγη εκτιμημένων συντελεστών: μία στήλη για τη σταθερά και μία για τον συντελεστή του wt.

4. Γραφική παρουσίαση των bootstrap συντελεστών

Τα δύο ιστογράμματα δείχνουν τις bootstrap κατανομές:

```
hist(results[, 1])
```

```
hist(results[, 2])
```

Το πρώτο ιστόγραμμα αφορά τη σταθερά του υποδείγματος και το δεύτερο τον συντελεστή της μεταβλητής wt.

Η εντολή:

```
abline(v = beta0, col = "green")
```

```
abline(v = beta1, col = "green")
```

προσθέτει σε κάθε γράφημα την αρχική OLS εκτίμηση. Με αυτόν τον τρόπο μπορούμε να συγκρίνουμε την αρχική εκτίμηση με το κέντρο της bootstrap κατανομής.

5. Bootstrap bias

Το bootstrap bias μετρά τη διαφορά ανάμεσα στον μέσο των bootstrap εκτιμήσεων και στην αρχική εκτίμηση του υποδείγματος.

Για τη σταθερά:

```
bias_beta0 = mean(results[, 1]) - beta0
```

Για τον συντελεστή του wt:

```
bias_beta1 = mean(results[, 2]) - beta1
```

Αν το bias είναι κοντά στο μηδέν, τότε οι bootstrap εκτιμήσεις είναι κατά μέσο όρο πολύ κοντά στην αρχική εκτίμηση. Αν το bias είναι μεγάλο, αυτό μπορεί να δείχνει ότι ο εκτιμητής παρουσιάζει συστηματική απόκλιση.