Κεφάλαιο 11

Τεχνικές Αναδειγματοληψίας

Ο στατιστικός πολύ συχνά ενδιαφέρεται να υπολογίσει μια εκτιμήτρια μαζί με το τυπικό της σφάλμα με σκοπό να κατασκευάσει διαστήματα εμπιστοσύνης για την πραγματική τιμή της παραμέτρου. Ωστόσο, αρκετές φορές είναι δύσκολο να βρεθεί μια ακριβής έκφραση για τη διακύμανση διαφόρων εκτιμητριών, και συνεπώς, είναι αδύνατο να υπολογιστεί το τυπικό τους σφάλμα. Βασικές μέθοδοι που οι στατιστικοί έχουν χρησιμοποιήσει είναι οι προσεγγίσεις ή οι μετασχηματισμοί για να πετύχουν κανονικότητα. Αυτό, όμως, μπορεί να είναι απαγορευτικό για ένα μεγάλο αριθμό προβλημάτων.

Σήμερα, η υπολογιστική δύναμη οδήγησε στις τεχνικές αναδειγματοληψίας, όπως είναι οι μέθοδοι jackknife και bootstrap. Σκοπός αυτού του κεφαλαίου είναι να παρουσιάσει τον τρόπο που μπορούν να εφαρμοστούν αυτές οι δυο μέθοδοι στην R, είτε ξεκινώντας από τις βασικές έννοιες, είτε χρησιμοποιώντας έτοιμες συναρτήσεις που υπάρχουν στις βιβλιοθήκες της.

11.1 Μέθοδος Jackknife

Η μέθοδος jackknife αποτελείται από δυο βήματα. Πρώτα, παράγονται τα jackknife δείγματα αφαιρώντας την x_i τιμή από το αρχικό δείγμα. Έπειτα, υπολογίζεται η προς εξέταση εκτιμήτρια για κάθε ένα από τα δείγματα jackknife, δηλαδή η

$$\hat{\theta}_i(x_1, \ldots, x_{i-1}, \ldots, x_n).$$

Στη συνέχεια ορίζεται η ψευδοτιμή

$$\hat{\theta}_i^* = n\hat{\theta} - (n-1)\hat{\theta}_i,$$

όπου $\hat{\theta}$ η εκτιμήτρια από το αρχικό δείγμα. Τέλος, η jackknife εκτιμήτρια είναι ίση με

$$J(\hat{\theta}) = \frac{1}{n}\sum \hat{\theta}_i^*$$

με τυπικό σφάλμα

$$\hat{\sigma}_{jack}(\hat{\theta}) = \left[\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\hat{\theta}_i^* - \hat{\theta^*}(.)\right)^2\right]^{1/2},$$

όπου

$$\hat{\theta^*}(.) = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i^*.$$

Το προσεγγιστικό 95% διάστημα εμπιστοσύνης για την προς εκτίμηση παράμετρο δίνεται από

$$J(\hat{\theta}) \pm t_{0.975,n-1}\cdot\hat{\sigma}_{jack}(\hat{\theta})$$

Θα εξεταστεί τώρα πως μπορεί να προγραμματιστούν τα πιο πάνω στην R για να γίνει η εκτίμηση του συντελεστή μεταβλητότητας $CV = \sqrt{Var(x)}/\bar{x}$ μαζί με το διάστημα εμπιστοσύνης του για ένα δείγμα με 25 παρατηρήσεις :

```
8.26    6.33    10.4    5.27    5.35    5.61    6.12    6.19
5.2     7.01    8.74    7.78    7.02    6       6.5     5.8
5.12    7.41    6.52    6.21    12.28   5.6     5.38    6.6
8.74
```

Αρχικά, εισάγονται τα δεδομένα στην R στη μορφή διανύσματος και μετά ορίζεται η συνάρτηση για τον υπολογισμό του συντελεστή μεταβλητότητας

```
> x <-c(8.26, 6.33, 10.4, 5.27, 5.35, 5.61, 6.12, 6.19, 5.2,
+ 7.01, 8.74, 7.78, 7.02, 6, 6.5, 5.8, 5.12, 7.41, 6.52, 6.21,
+ 12.28, 5.6, 5.38, 6.6, 8.74)
> CV<-function(x) {sqrt(var(x))/mean(x)}
> CV(x)
[1] 0.2524712
```

Στη συνέχεια, προχωρούμε με τον κώδικα υπολογισμού της jackknife εκτιμήτρια μαζί με το τυπικό της σφάλμα.

```
> jack <- numeric(length(x)-1)
> pseudo <- numeric(length(x))
> for (i in 1:length(x))
```

```
+ {
+ jack<-x[-i]
+ pseudo[i]<-length(x)*CV(x)-(length(x)-1)*CV(jack)
+ }
> jack.estim<-mean(pseudo)
> jack.estim
[1] 0.2617376
> jack.se<-sqrt(var(pseudo)/length(x))
> jack.se
[1] 0.05389943
```

Η πρώτη εντολή καθορίζει στην R ότι θα δημιουργηθούν τα δείγματα jackknife, `jack`, τα οποία περιέχουν $n-1$ παρατηρήσεις. Το δεύτερο διάνυσμα `pseudo` είναι αυτό που θα περιέχει τις n ψευδοτιμές. Με την εντολή `for` δημιουργείται ο βρόγχος με τον οποίο θα κατασκευαστούν οι ψευδοτιμές. Για κάθε i δημιουργείται το jackknife δείγμα αφαιρώντας την x_i παρατήρηση από το αρχικό δείγμα, και στη συνέχεια υπολογίζεται η i ψευδοτιμή. Με την εντολή `mean(pseudo)` υπολογίζεται η jackknife εκτιμήτρια με τυπικό σφάλμα το `jack.se`.

Το άνω φράγμα του προσεγγιστικού 95% διαστήματος εμπιστοσύνης για το συντελεστή μεταβλητότητας υπολογίζεται στην R από

```
> jack.estim+qt(0.975,length(x)-1)*jack.se
[1] 0.3729806
```

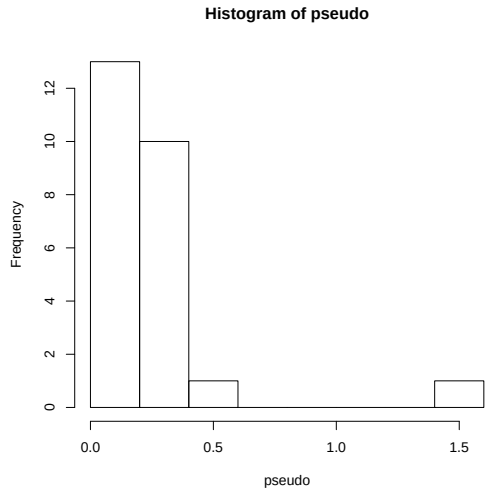
ενώ το κάτω φράγμα από

```
> jack.estim-qt(0.975,length(x)-1)*jack.se
[1] 0.1504947
```

Η μορφή των ψευδοτιμών από την μέθοδο jackknife φαίνεται από το ιστόγραμμα στο Σχήμα 11.1.

```
>hist(pseudo)
```

Ως τώρα παρουσιάστηκε η μέθοδος jackknife ξεκινώντας από τις βασικές της έννοιες. Το επόμενο παράδειγμα παρουσιάζει πως μπορεί να εφαρμοστεί η μέθοδος χρησιμοποιώντας την εντολή `jackknife`, η οποία βρίσκεται στη βιβλιοθήκη `bootstrap` της R, για τον υπολογισμό ενός διαστήματος εμπιστοσύνης για τη μέγιστη τιμή από έξι τιμές από την ομοιόμορφη κατανομή.



Σχήμα 11.1: Ιστόγραμμα ψευδοτιμών jackknife.

```
> library(bootstrap)
> x1<-runif(6)
> x1
[1] 0.3180501 0.6395107 0.2261756 0.2970479 0.4609984 0.8353474
> jack1<-jackknife(x1,max)
> jack1
$jack.se
[1] 0.1631973

$jack.bias
[1] -0.1631973

$jack.values
[1] 0.8353474 0.8353474 0.8353474 0.8353474 0.8353474 0.6395107

$call
jackknife(x = x1, theta = max)

> estim.jack<-mean(jack1$jack.values)
> estim.jack
```

```
[1] 0.802708
> bias<-jack1$jack.bias
> quantile(jack1$jack.values,c(0.025,0.05,0.95,0.975))
     2.5%        5%       95%     97.5%
0.6639903 0.6884699 0.8353474 0.8353474
> estim.jack+qt(0.975,length(x)-1)*jack1$jack.se
[1] 1.139531
> estim.jack-qt(0.975,length(x)-1)*jack1$jack.se
[1] 0.4658853
```

Οι ψευδοτιμές παίρνονται με την εντολή `jack1$jack.values` και για να υπολογιστεί η jackknife εκτιμήτρια παίρνουμε τη μέση τους τιμή. Το τυπικό σφάλμα της εκτιμήτριας παίρνεται με την εντολή `jack1$jack.se` ενώ η εντολή `jack1$jack.bias` δίνει την μεροληψία της εκτιμήτριας. Με την εντολή `quantile` λαμβάνονται τα εμπειρικά ποσοστημόρια των ψευδοτιμών και με τις τελευταίες δυο εντολές υπολογίζεται το προσεγγιστικό 95% διάστημα εμπιστοσύνης.

11.2 Μέθοδος Bootstrap

Η μέθοδος bootstrap βασίζεται στη δημιουργία B νέων δειγμάτων με ίδιο μέγεθος με το αρχικό δείγμα. Τα δείγματα αυτά δημιουργούνται με δειγματοληψία με επανάθεση από το αρχικό δείγμα. Η εκτιμήτρια της παραμέτρου που μας ενδιαφέρει υπολογίζεται για το κάθε ένα από τα B δείγματα bootstrap και παράγουν την κατανομή bootstrap της εκτιμήτριας. Βασική προϋπόθεση είναι ότι οι παρατηρήσεις του αρχικού δείγματος απεικονίζουν όλον τον πληθυσμό.

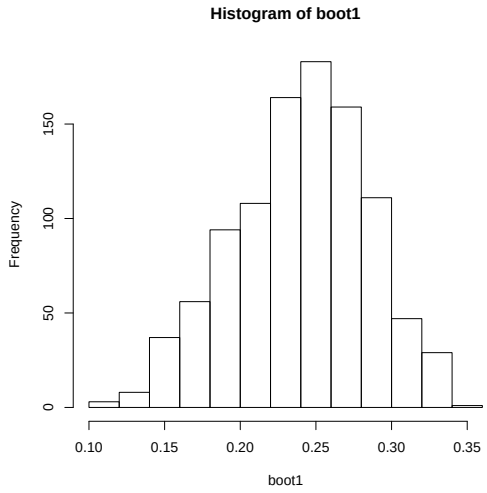
Στην R μπορούν να χρησιμοποιηθούν διάφορες εντολές για τον υπολογισμό των bootstrap εκτιμητριών όπως και το διάστημα εμπιστοσύνης για τη παράμετρο. Ο επόμενος κώδικας παρουσιάζει τον τρόπο εκτίμησης του συντελεστή μεταβλητότητας και την κατασκευή του διαστήματος εμπιστοσύνης με τη μέθοδο bootstrap χρησιμοποιώντας τα προηγούμενα δεδομένα.

```
> boot1 <-numeric(1000)
> for (i in 1:1000)
+ {
+ boot1[i] <- CV(sample(x,replace=T))
+ }
> boot.estim<-mean(boot1)
```

```
> boot.se<-sqrt(var(boot1))
> hist(boot1)
> quantile(boot1,0.975)
   97.5%
0.315552
> quantile(boot1,0.025)
     2.5%
0.1485921
> bias <- mean(boot1) - CV(x)
> CV(x) - bias
[1] 0.2646007
> CV(x) - bias - 1.96*boot.se
[1] 0.1772671
> CV(x) - bias + 1.96*boot.se
[1] 0.3519343
```

Η πρώτη εντολή καθορίζει στην R το διάνυσμα στο οποίο θα φυλαχθούν οι εκτιμήτριες για το συντελεστή μεταβλητότητας για κάθε bootstrap δείγμα. Η δημιουργία κάθε bootstrap δείγματος γίνεται με την εντολή `sample(x,replace=T)`. Με το βρόγχο `for` υπολογίζεται η εκτιμήτρια του συντελεστή μεταβλητότητας για 1000 bootstrap δείγματα. Η bootstrap εκτιμήτρια δίνεται παίρνοντας τη μέση τιμή όλων των εκτιμητριών από τα bootstrap δείγματα, `mean(boot1)`, ενώ το τυπικό της σφάλμα υπολογίζεται από `sqrt(var(boot1))`. Στη συνέχεια δίνεται η εντολή για κατασκευή του ιστογράμματος των εκτιμητριών από τα bootstrap δείγματα για να παρατηρηθεί η κατανομή τους, η οποία δε φαίνεται να διαφέρει πολύ από την κανονική (βλέπε Σχήμα 11.2). Επίσης, δίνονται τα 2.5% και 97.5% εμπειρικά ποσοστημόριά τους, τα οποία ορίζουν και το εμπειρικό 95% διάστημα εμπιστοσύνης. Τέλος, υπολογίζεται η μεροληψία του συντελεστή μεταβλητότητας του αρχικού δείγματος πριν την εφαρμογή της μεθόδου bootstrap για να κατασκευαστεί στη συνέχεια το 95% προσεγγιστικό διάστημα εμπιστοσύνης, υποθέτοντας κανονικότητα.

Πιο κάτω θα παρουσιαστούν δυο παραδείγματα της μεθόδου bootstrap χρησιμοποιώντας τη βιβλιοθήκη `boot` της R. Το πρώτο παράδειγμα αναφέρεται στην εκτίμηση του συντελεστή συσχέτισης, ενώ το δεύτερο στην εκτίμηση των συντελεστών παλινδρόμησης.



Σχήμα 11.2: Ιστόγραμμα εκτιμητριών από τα bootstrap δείγματα.

11.3 Εκτίμηση Συντελεστή Συσχέτισης

Έστω το παράδειγμα από τους Efron και Tibshirani (1993) στο οποίο 82 σχολές νομικής συμμετείχαν σε μια μελέτη για την πρακτική εισδοχής των φοιτητών. Για κάθε μια από αυτές τις σχολές, 15 σχολεία επιλέγηκαν τυχαία και εξετάστηκε η συσχέτιση μεταξύ των αποτελεσμάτων της εξέτασης LSAT και του μέσου όρου (GPA) βάση της τάξης του 1973. Η bootstrap ανάλυση στην R έγινε με την εντολή `boot`, η οποία βρίσκεται στην ομώνυμη βιβλιοθήκη, όπως πιο κάτω

```
> library("boot")
> school<-1:15
> lsat<-c(576,635,558,578,666,580,555,661,651,605,653,575,545,572,594)
> gpa<-c(3.39,3.30,2.81,3.03,3.44,3.07,3.00,3.43,3.36,3.13,3.12,2.74,
+ 2.76,2.88,2.96)
> law.data <- data.frame(School=school, LSAT=lsat, GPA=gpa)
> correl<-function(data,indices)
+ {
+  data<-law.data[indices,]
+ cor(data[,2],data[,3])
+ }
> boot.obj1 <- boot(law.data, correl, 1000)
```

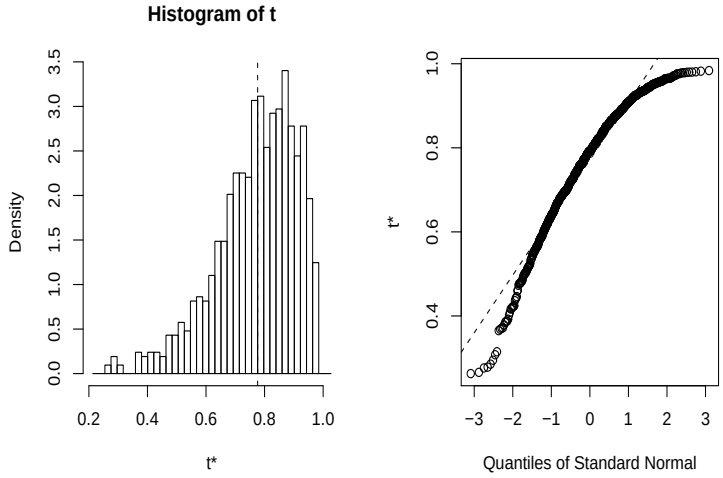
```
> boot.obj1

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = law.data, statistic = correl, R = 1000)


Bootstrap Statistics :
     original       bias    std. error
t1* 0.7763745 -0.005066455   0.1371331
> boot.ci(boot.obj1,type=c("norm","perc","bca"),conf=c(0.90,0.95))
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot.obj1, conf = c(0.9, 0.95), type = c("norm",
    "perc", "bca"))

Intervals :
Level      Normal             Percentile            BCa
90%   ( 0.5559,  1.0070 )   ( 0.5071,  0.9510 )   ( 0.3852,  0.9265 )
95%   ( 0.5127,  1.0502 )   ( 0.4245,  0.9644 )   ( 0.2788,  0.9407 )
Calculations and Intervals on Original Scale
Some BCa intervals may be unstable
> plot(boot.obj1)
```

Στην αρχή κατασκευάζεται η στατιστική συνάρτηση (συντελεστής συσχέτισης) με τέτοιο τρόπο έτσι ώστε να μπορεί να χρησιμοποιηθεί στην εντολή `boot`. Η παράμετρος `data` της συνάρτησης καθορίζει το πλαίσιο δεδομένων (δείγμα), ενώ η παράμετρος `indices` θα επιτρέψει στην εντολή `boot` να διαλέξει το δείγμα bootstrap από το αρχικό δείγμα με δειγματοληψία με επανάθεση. Η εντολή `boot` δημιουργεί 1000 συντελεστές συσχέτισης για τα δεδομένα `law.data`. Τα αποτελέσματα της εντολής `boot` δίνουν την αρχική εκτιμήτρια για τον συντελεστή συσχέτισης (πριν την εφαρμογή της μεθόδου) μαζί με την μεροληψία και το τυπικό της σφάλμα. Σημειώνεται εδώ ότι η bootstrap εκτιμήτρια υπολογίζεται αφαιρώντας

τη μεροληψία από την αρχική εκτιμήτρια. Η εντολή `boot.ci` δίνει τα διαστήματα εμπιστοσύνης για το συντελεστή συσχέτισης. Στο παράδειγμα επιλέγηκαν τα ακόλουθα διαστήματα εμιστοσύνης με επίπεδα εμπιστοσύνης 90% και 95%:

1. το προσεγγιστικό διάστημα εμπιστοσύνης με την κανονική (Normal),
2. το εμπειρικό διάστημα εμπιστοσύνης χρησιμοποιώντας ποσοστημόρια (Percentile),
3. το διάστημα εμπιστοσύνης χρησιμοποιώντας τα προσαρμοσμένα ποσοστημόρια λαμβάνοντας υπόψην τη διόρθωση της μεροληψίας (BCa).

Παρατηρώντας το ιστόγραμμα και το QQ-γράφημα (Σχήμα 11.3), τα οποία κατασκευάζονται με την εντολή `plot` και όρισμα το αντικείμενο `boot`, φαίνεται ότι οι bootstrap εκτιμήτριες δεν ακολουθούν την κανονική κατανομή. Συνεπώς, είναι καλύτερο να χρησιμοποιηθούν τα εμπειρικά διαστήματα εμπιστοσύνης, παρά το προσεγγιστικό με τη βοήθεια της κανονικής.



Σχήμα 11.3: Ιστόγραμμα και QQ γράφημα των 1000 εκτιμητριών bootstrap για τον συντελεστή συσχέτισης.

11.4 Συντελεστές Παλινδρόμησης

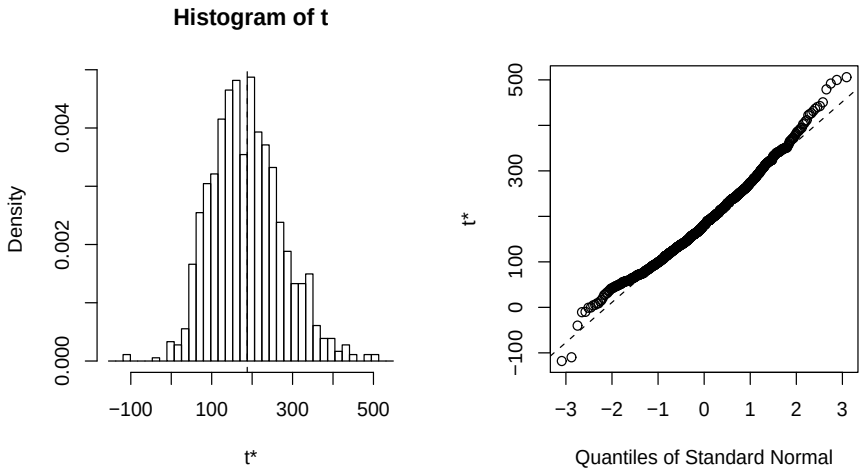
Το ακόλουθο παράδειγμα προβάλλει τον τρόπο εκτίμησης των συντελεστών παλινδρόμησης με την μέθοδο bootstrap χρησιμοποιώντας τα δεδομένα από την

προηγούμενη ενότητα.

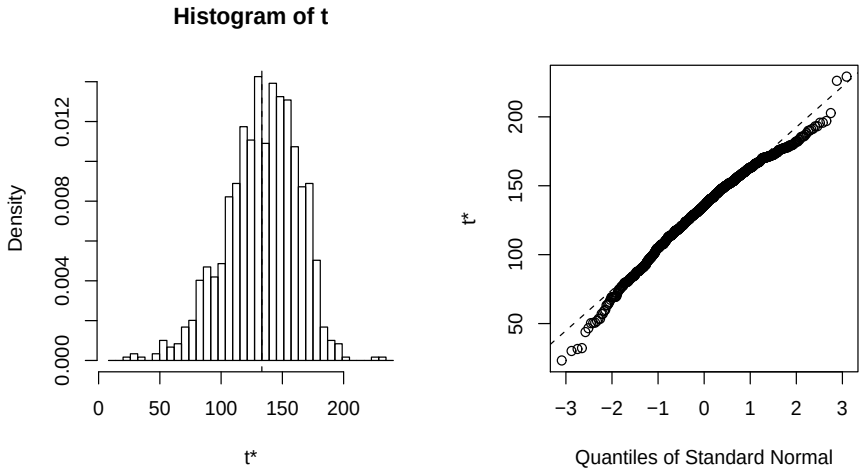
```
> regcoef<-function(data,indices)
+ {
+  data<-law.data[indices,]
+ mod<-lm(LSAT~GPA,data)
+ coef(mod)
+ }
> boot.obj2 <- boot(law.data,regcoef,1000)
> boot.obj2

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = law.data, statistic = regcoef, R = 1000)


Bootstrap Statistics :
    original     bias    std. error
t1* 187.8996 -0.5759099    88.19863
t2* 133.2509  0.2066162    29.52671
> boot.ci(boot.obj2,index=1,type=c("norm","perc","bca"),
+ conf=c(0.90,0.95))
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot.obj2, conf = c(0.9, 0.95), type = c("norm",
    "perc", "bca"), index = 1)

Intervals :
Level      Normal             Percentile            BCa
90%   ( 43.4, 333.5 )   ( 58.3, 343.6 )   ( 74.8, 387.7 )
95%   ( 15.6, 361.3 )   ( 43.8, 376.4 )   ( 62.3, 441.7 )
Calculations and Intervals on Original Scale
Some BCa intervals may be unstable
```

```
> boot.ci(boot.obj2,index=2,type=c("norm","perc","bca"),
+ conf=c(0.90,0.95))
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot.obj2, conf = c(0.9, 0.95), type = c("norm",
    "perc", "bca"), index = 2)

Intervals :
Level      Normal             Percentile            BCa
90%   ( 84.5, 181.6 )   ( 81.0, 176.0 )   ( 63.5, 170.3 )
95%   ( 75.2, 190.9 )   ( 69.3, 180.9 )   ( 45.0, 173.8 )
Calculations and Intervals on Original Scale
Some BCa intervals may be unstable
> plot(boot.obj2,index=1)
> plot(boot.obj2,index=2)
```

Τα πιο πάνω αποτελέσματα δείχνουν ότι η bootstrap εκτιμήτριες για το σταθερό όρο και την κλίση να είναι ίσες με 133.3 και 187.9, αντίστοιχα. Συγκεκριμένα, η κλίση είναι θετική όπως αναμενόταν από τα προηγούμενα αποτελέσματα (γιατι;). Για τα διαστήματα εμπιστοσύνης για τον καθένα συντελεστή παλινδρόμησης, τίθεται στην εντολή `boot.ci` το όρισμα `index` να είναι ίσο με 1 και 2, αντίστοιχα. Το ίδιο όρισμα δίνεται και για την κατασκευή των γραφημάτων της μεθόδου (Σχήματα 11.4 και 11.5).



Σχήμα 11.4: Ιστόγραμμα και QQ γράφημα των 1000 εκτιμητριών bootstrap για τον σταθερό όρο.



Σχήμα 11.5: Ιστόγραμμα και QQ γράφημα των 1000 εκτιμητριών bootstrap για την κλίση.