Άσκηση 1 (bootstrap στο γραμμικό μοντέλο)**.** Τα δεδομένα του Πίνακα 1 περιέχουν $n = 14$ μετρήσεις του ποσοστού επιβίωσης αρουραίων (surv) αφού εκτέθηκαν σε διάφορα επίπεδα ραδιενέργειας (dose).

1. Υπολογίστε τις εκτιμήσεις ελαχίστων τετραγώνων για τις παραμέτρους του απλού γραμμικού μοντέλου
$$\log(\text{surv}) = \beta_0 + \beta_1 \times \text{dose}$$
και εξετάστε αν ισχύει η υπόθεση κανονικότητας των σφαλμάτων.

2. Προσομοιώστε $B = 10000$ bootstrap τιμές των $(\widehat{\beta}_0, \widehat{\beta}_1)$ και κατασκευάστε το διάγραμμα διασποράς και τα ιστογράμματα των προσομοιωμένων τιμών.

3. Να κατασκευαστούν 95% bootstrap διαστηματα εμπιστοσύνης (κανονικά, βασικά και ποσοστιαιών σημείων) για το β_1.

4. Απαντήστε στα 2 προηγούμενα ερωτήματα κάνοντας χρήση του πακέτου `boot`.

5. Σε επίπεδο σημαντικότητας $\alpha = 5\%$ η πιθανότητα επιβίωσης εξαρτάται από την δόση ;

	dose	surv
1	117.5	44.000
2	117.5	55.000
3	235.0	16.000
4	235.0	13.000
5	470.0	4.000
6	470.0	1.960
7	470.0	6.120
8	705.0	0.500
9	705.0	0.320
10	940.0	0.110
11	940.0	0.015
12	940.0	0.019
13	1410.0	0.700
14	1410.0	0.006

Πίνακας 1: Δεδομένα από: Efron, B. (1988). Computer-intensive methods in statistical regression. *SIAM Review*, **30**, 421-449. Διατίθενται στο `boot` (πακέτο της `R`).

Άσκηση 2 (bootstrap για έλεγχο μέσου)**.** Έστω

$$\boldsymbol{x} = (-0.89, -0.47, 0.05, 0.155, 0.279, 0.775, 1.0016, 1.23, 1.89, 1.96)$$

τυχαίο δείγμα 10 τιμών. Εκτιμήστε το p-value του ελέγχου υπόθεσης

$$H_0 : \mu = 1 \quad \text{έναντι} \quad H_1 : \mu \neq 1$$

χρησιμοποιώντας bootstrap, όπου μ η μέση τιμή του πληθυσμού.

Άσκηση 3 (bootstrap classification tree)**.** Θεωρήστε τα δεδομένα ιταλικών κρασιών που είναι διαθέσιμα στο πακέτο `pgmm`:

```
library('pgmm')
data(wine)
# this is just to get rid of weird characters in column names
cNames <- colnames(wine)
nCols <- length(cNames)
cNames <- gsub('-','.', cNames)
cNames <- gsub(' ','.', cNames)
cNames <- gsub('/','.', cNames)
cNames <- gsub("^([0-9])", "X\\1", cNames)
colnames(wine) <- cNames
```

Οι $n = 178$ παρατηρήσεις αποτελούνται από $p = 27$ μετρήσεις μεταβλητών που σχετίζονται με τη συγκέντρωση διάφορων συστατικών του εκάστοτε κρασιού. Επί πλέον, υπάρχει και άλλη μία μεταβλητή ($\texttt{Type} \in \{1, 2, 3\}$) που δίνει την κατάταξη κάθε κρασιού σε μία από τρεις διαφορετικές ποικιλίες:

'1' Barolo

'2' Grignolino

'3' Barbera.

Κατόπιν θεωρήστε ότι το παραπάνω σύνολο δεδομένων χωρίζεται σε δύο τμήματα train data (διάστασης 120×28) και test data (διάστασης 58×27) ως εξής:

```
set.seed(1)
n <- dim(wine)[1]
ind <- sample(n, 120, replace = FALSE)
train_data <- wine[ind,]
train_data$Type <- as.factor(train_data$Type)
test_data <- wine[-ind,-1]
test_labels <- wine[-ind,1]
```

Το ζητούμενο είναι να κατασκευάσουμε έναν κανόνα ταξινόμησης με βάση τα παρατηρήθεντα δεδομένα στο train dataset, με απώτερο σκοπό την πρόβλεψη της ποικιλίας κρασιού στο test dataset. Για τον σκοπό αυτό (α) θα εκτιμήσουμε ένα classification tree[1] και (β) θα εφαρμόσουμε bootstrap για να προσαρμόσουμε πολλά classification trees. Σε κάθε περίπτωση, θα αξιολογούμε την ακρίβεια πρόβλεψης με βάση την πραγματική ταξινόμηση του test dataset, πληροφορία η οποία δίνεται στο διάνυσμα `test_labels` που ορίστηκε παραπάνω[2].

[1] Τα δέντρα ταξινόμησης είναι μη παραμετρικοί classifiers, αλλά παρουσιάζουν ορισμένα μειονεκτήματα. Συχνά έχουν τάση για υπερπροσαρμογή (overfitting) και μπορεί να είναι ασταθή, καθώς μικρές αλλαγές στα δεδομένα μπορούν να οδηγήσουν σε διαφορετική δομή δέντρου. Αυτό σημαίνει ότι οι προκύπτουσες προβλέψεις σε νέα δεδομένα έχουν αρκετά μεγάλη διακύμανση.

[2] παρατηρήστε ότι αυτή η πληροφορία δεν χρησιμοποιείται από τη μέθοδο.

1. Προσαρμόστε ένα δέντρο ταξινόμησης και κατόπιν εκτιμήστε την ακρίβεια πρόβλεψης με βάση τις εξής εντολές

```
library(tree)
# fit tree to training dataset (Type is response variable)
fit1 <- tree(as.factor(Type)~., data = train_data)
# predict type of wine for the test_data
tr <- predict(fit1, newdata = test_data, type = 'class')
# tabulate with the true classification
table(test_labels, tr)
# prediction accuracy
paste0(
round(
100*sum(diag(table(test_labels, tr)))/dim(test_data)[1],
2), '%')
```

2. Κατόπιν, να επαναλάβετε την παραπάνω διαδικασία κάνοντας bootstrap $B = 1000$ δέντρων. Σε κάθε bootstrap δέντρο ταξινόμησης να εκτιμήσετε την προβλεψη στο `test_dataset` και κατόπιν να αναφέρετε την **τελική πρόβλεψη** που προκύπτει ως η πιο πιθανή κατάταξη από τα B δέντρα[3]. Τι παρατηρείτε;

[3] Η παραπάνω διαδικασία είναι γνωστή με τον όρο bootstrap aggregating (bagging) (δες L. Breiman (1996): Bagging Predictors. *Machine Learning*) και είναι ένα βασικό δομικό συστατικό αυτού που είναι γνωστό ως random forest (δες L. Breiman (2001): Random Forests. *Machine Learning*).