Δημήτριος Φουσκάκης
Καθηγητής Ε.Μ.Π.

ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΜΕ ΧΡΗΣΗ ΤΗΣ R

2η έκδοση

Την οικογενειακή φίλη και φιλόλογο Αγγελική Τραχανά για τη συνεισφορά της στη φιλολογική επιμέλεια του βιβλίου, καθώς και την οικογένειά μου για την αμέριστη συμπαράστασή της κατά τη διάρκεια αυτού του συγγραφικού ταξιδιού. Επίσης θα ήθελα να ευχαριστήσω θερμά την υποψήφια διδάκτορα Ευστρατία Χαριτίδου της οποίας η συμβολή στην τελική διαμόρφωση του βιβλίου υπήρξε σημαντικότατη. Η κα. Χαριτίδου αρχικά αφιέρωσε πολύ χρόνο διαβάζοντας με μεγάλη προσοχή το βιβλίο, επισημαίνοντάς μου λάθη και παραλείψεις και εν συνεχεία βοήθησε ουσιαστικά, με περίσσια υπομονή και επιμονή, ώστε το βιβλίο να πάρει την τελική του μορφή. Επιπλέον με ζήλο επιμελήθηκε τις ασκήσεις στο τέλος κάθε κεφαλαίου καθώς και τη δημιουργία κάποιων διαγραμμάτων. Τέλος θα ήθελα να ευχαριστήσω τον Καθηγητή μου *Prof. David Draper*, διότι τα περισσότερα που έχω μάθει στη Στατιστική τα οφείλω σε αυτόν.

Η ιστοσελίδα του βιβλίου

`http://www.math.ntua.gr/~fouskakis/Rbook`

περιέχει συμπληρωματικό υλικό, π.χ. σύνολα δεδομένων που χρησιμοποιούνται στις ασκήσεις.

Αθήνα, Δεκέμβριος 2013

Δημήτριος Φουσκάκης

Πρόλογος $2^{ης}$ Έκδοσης

Η νέα έγχρωμη έκδοση του βιβλίου είναι ανανεωμένη, βελτιωμένη και επαυξημένη. Συγκεκριμένα, έχουν γίνει προσθήκες στο Κεφάλαιο 2 όπου πλέον γίνεται πλήρης αναφορά στις συναρτήσεις `apply()` καθώς και σε χρήση ελληνικών χαρακτήρων στην R. Το Κεφάλαιο 2 ολοκληρώνεται με μια νέα παράγραφο στην οποία παρουσιάζεται το `RStudio` και το `R Markdown`. Το `RStudio` είναι ένα ολοκληρωμένο περιβάλλον επιφάνειας εργασίας ανοιχτού κώδικα της R

το οποίο επιτρέπει στον χρήστη να τρέξει την R σε ένα πιο φιλικό περιβάλλον, ενώ το R Markdown παρέχει ένα πλαίσιο συγγραφής των αποτελεσμάτων της R. Στο Κεφάλαιο 7 έχει προστεθεί υλικό σχετικά με την επίδραση που έχουν στους συντελεστές του γραμμικού μοντέλου "συνήθεις" γραμμικοί μετασχηματισμοί (κεντράρισμα, τυποποίηση ή κανονικοποίηση) στις τιμές της μεταβλητής απόκρισης και/ή στις τιμές των ποσοτικών επεξηγηματικών μεταβλητών. Στο ίδιο κεφάλαιο αναπτύσσονται περαιτέρω τα πολλαπλασιαστικά μοντέλα, με ή χωρίς εικονικές μεταβλητές, ενώ επιπλέον έχει προστεθεί υλικό σχετικά με τον συντελεστή προσδιορισμού και τους κινδύνους που κρύβει η λανθασμένη χρήση του ως μέτρο καλής προσαρμογής. Το Κεφάλαιο 7 ολοκληρώνεται με μία νέα παράγραφο σχετικά με τα συνηθέστερα μέτρα σύγκρισης γραμμικών μοντέλων.

Στην παρούσα έκδοση έχουν προστεθεί επίσης δύο νέα κεφάλαια. Στο Κεφάλαιο 9 εισάγονται, περιγράφονται και αναλύονται οι δυνατότητες των βιβλιοθηκών ggplot2 και data.table της R, καθώς και της βιβλιοθήκης shiny του RStudio, οι οποίες χρησιμοποιούνται πολύ συχνά για τη διαγραμματική απεικόνιση, το χειρισμό δεδομένων μεγάλης κλίμακας και τη δημιουργία διαδραστικών διαδικτυακών εφαρμογών από την R, αντίστοιχα. Παρουσιάζονται οι συνηθέστεροι μηχανισμοί διαγραμματικής αναπαράστασης δεδομένων προερχόμενων από ποσοτικές ή κατηγορικές μεταβλητές, σε μία ή και περισσότερες διαστάσεις, καθώς επίσης και τα συνηθέστερα διαγράμματα που δημιουργούμε για να ανακαλύψουμε το είδος της εξάρτησης δεδομένων ιδίου ή διαφορετικού είδους, με χρήση της βιβλιοθήκης ggplot2. Επιπλέον, παρουσιάζονται, μέσω της βιβλιοθήκης data_table, οι συνηθέστεροι μηχανισμοί χειρισμού δεδομένων, όπως συνάθροιση, ομαδοποίηση, ταξινόμηση, επιλογή, κ.λπ. Τέλος, γίνεται αναφορά στη βασική δομή συγγραφής shiny εφαρμογών και παρουσιάζονται διάφορα εργαλεία διαμόρφωσης τους, μέσω παραδειγμάτων.

Στο Κεφάλαιο 10 παρουσιάζεται η μεθοδολογία των μοντέλων της δίτιμης λογιστικής παλινδρόμησης, οι τρόποι προσαρμογής αυτών με τη βοήθεια της R, ενώ έμφαση δίνεται στην ερμηνεία των συντελεστών καθώς και στον έλεγχο των προϋποθέσεων τους. Επιπλέον, παρουσιάζονται οι έλεγχοι καλής προσαρμογής, ενώ αναφορά γίνεται και στο πρόβλημα της ταξινόμησης, καθώς και

της χρήσης μεθόδων διασταυρωμένης επικύρωσης.

Κλείνοντας, θα ήθελα να ευχαριστήσω τον αγαπητό φίλο και συνάδελφο Γεώργιο Πετράκο καθώς και την υποψήφια διδάκτορα και φιλόλογο Βασιλική Παπαϊωάννου για την πολύτιμη βοήθειά τους καθώς και τα εύστοχα σχόλιά τους στη νέα αυτή έκδοση του βιβλίου.

Αθήνα, Ιούλιος 2021

Δημήτριος Φουσκάκης

μία υπό συνθήκη πρόταση.

Ο βρόχος `while` έχει την εξής δομή: `while(A) B`. Όσο η συνθήκη που καθορίζεται από το `A` είναι αληθής, εκτελείται η διαδικασία `B`. Σε περίπτωση που το `B` αποτελείται από παραπάνω από μία διαδικασίες πρέπει να τις εσωκλείσουμε σε `{ }`.

Για παράδειγμα ας υποθέσουμε ότι θέλουμε να εφαρμόσουμε την αριθμητική μέθοδο *Newton-Raphson* για την εύρεση ρίζας της εξίσωσης $f(x) = x^3+2x^2-7 = 0$. Θυμίζουμε ότι με βάση τη μέθοδο μπορούμε να επιλύσουμε μια εξίσωση της μορφής $f(x) = 0$ (υπό την προϋπόθεση ότι υπάρχει η παράγωγος $f'(x)$ ως προς x και είναι μη μηδενική) με βάση την επαναλαμβανόμενη σχέση:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{2.1}$$

Τότε η τιμή x_{n+1} αποτελεί ρίζα της εξίσωσης εφόσον $|f(x_{n+1})| \leq \epsilon$, όπου ϵ ένας μικρός αριθμός που εκφράζει την ***ανοχή*** (*tolerance*) ή σφάλμα της μεθόδου. Η αρχική τιμή x_0 που χρησιμοποιούμε είναι απλώς μια υπόθεση.

Εύκολα προκύπτει ότι $f'(x) = 3x^2 + 4x$. Ξεκινάμε από την αρχική τιμή $x_0 = 1$ και έστω ότι θέλουμε να επαναλάβουμε τη σχέση (2.1) μέχρις ότου $|f(x_{n+1})| \leq 0.000001$ ($tolerance = 0.000001$). Ο αριθμός των επαναλήψεων που θα απαιτηθεί για κάτι τέτοιο είναι άγνωστος *a priori*, οπότε δεν μπορούμε να χρησιμοποιήσουμε το βρόχο `for`. Χρησιμοποιούμε τον παρακάτω κώδικα:

```
> x<-1
> tolerance<-0.000001
> f<-x^3+2*x^2-7
> f.prime<-3*x^2+4*x
> while(abs(f)>tolerance)
 {
  x<-x-f/f.prime
  f<-x^3+2*x^2-7
  f.prime<-3*x^2+4*x
 }
> x
[1] 1.428818
> f
[1] 1.585059e-08
```

Η tapply() χρησιμοποιείται για να εφαρμόσει μία συνάρτηση σε υποσύνολα ενός διανύσματος. Ας υποθέσουμε ότι το παρακάτω διάνυσμα περιέχει τους βαθμούς (με άριστα το 20) δέκα μαθητών, όπου οι 5 πρώτοι είναι αγόρια και οι 5 τελευταίοι κορίτσια:

```
> x <- c( 12, 14, 16, 17, 14, 19, 18, 11, 17, 15)
> f <- gl(2,5)
> f
 [1] 1 1 1 1 1 2 2 2 2 2
Levels: 1 2
```

Με τη χρήση της συνάρτησης gl() δημιουργούμε τις δύο "κλάσεις" μαθητών (1: αγόρι, 2: κορίτσι). Με χρήση της εντολής tapply() μπορούμε να εφαρμόσουμε μια συνάρτηση ξεχωριστά στις κλάσεις του διανύσματος, π.χ.:

```
> tapply(x,f,sum)
 1  2
73 80
```

Προσθέτοντας το όρισμα simplify = FALSE το αποτέλεσμα είναι λίστα:

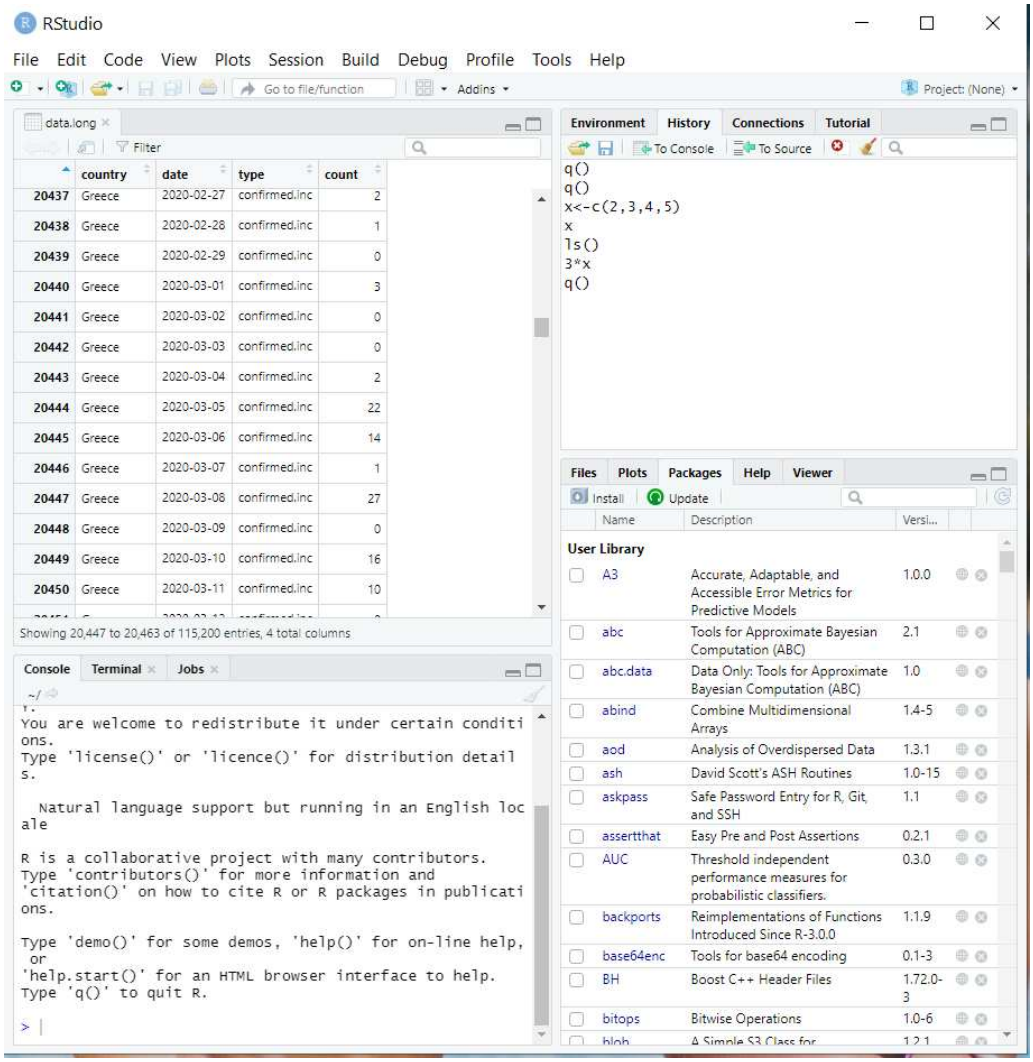
```
> tapply(x,f,sum,simplify=FALSE)
 $`1`
[1] 73

$`2`
[1] 80
```

Σε περιπτώσεις που η συνάρτηση που χρησιμοποιείται ως όρισμα στην tapply() επιστρέφει παραπάνω από δύο τιμές το αποτέλεσμα είναι πάντα λίστα:

```
> tapply(x,f,range)
$`1`
[1] 12 17

$`2`
[1] 11 19
```

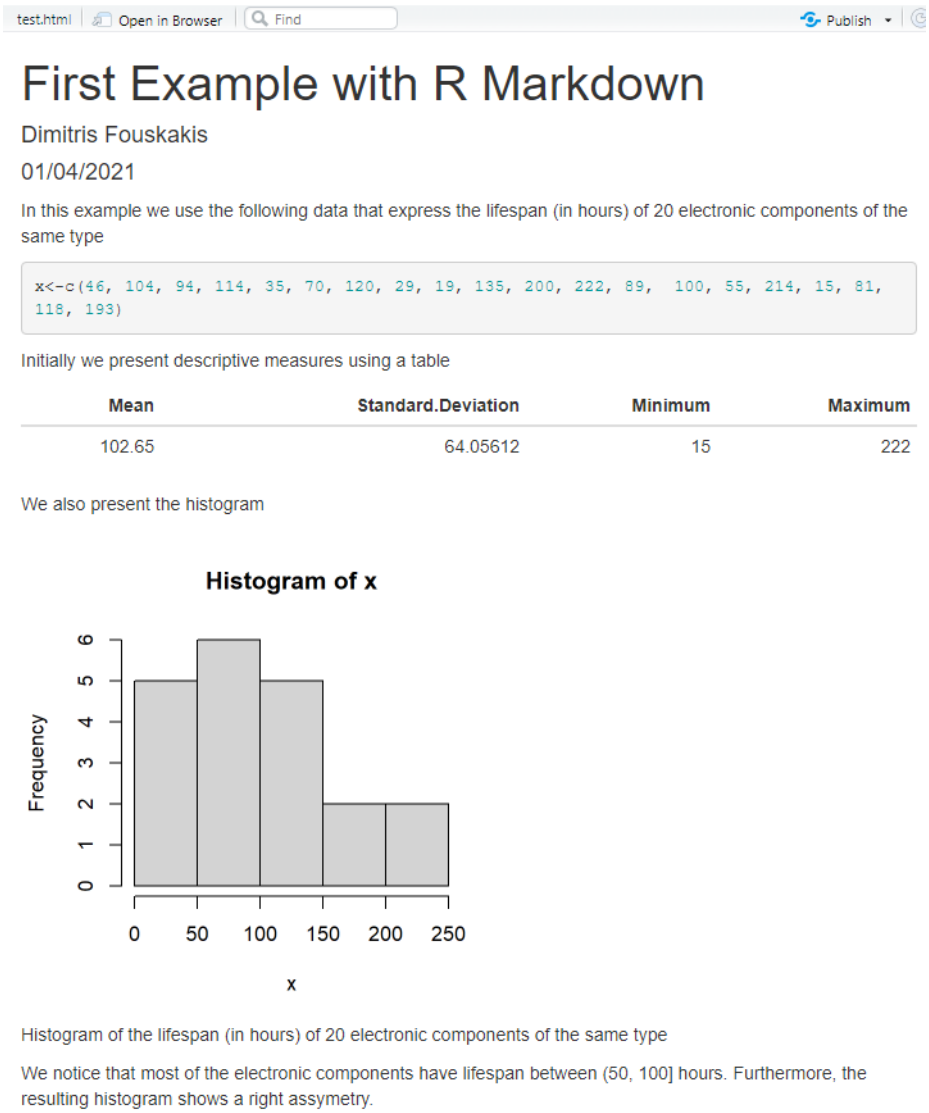
Ισοδύναμα, θα μπορούσαμε να είχαμε διαχωρίσει τους βαθμούς σε αυτούς των αγοριών και αυτούς των κοριτσιών, με χρήση της συνάρτησης split()

```
> split(x,f)
```



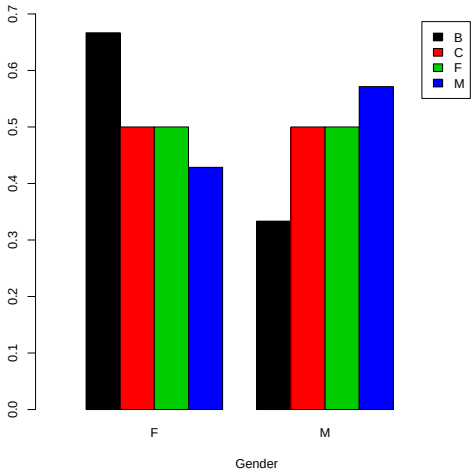
Διάγραμμα 2.3: Το περιβάλλον του RStudio

Το RStudio παρέχει επιπλέον στον χρήστη τη δυνατότητα να εκτελέσει τοπικές εργασίες. Μπορείτε να τις χρησιμοποιήσετε για να εκτελέσετε τις εντολές σας στο παρασκήνιο, ενώ συνεχίζετε κανονικά να χρησιμοποιείτε την κονσόλα. Η ***τοπική εργασία*** (*local job*) είναι ένας συντάκτης που εκτελείται σε μια ξεχωριστή, ειδική συνεδρία της R. Μπορείτε

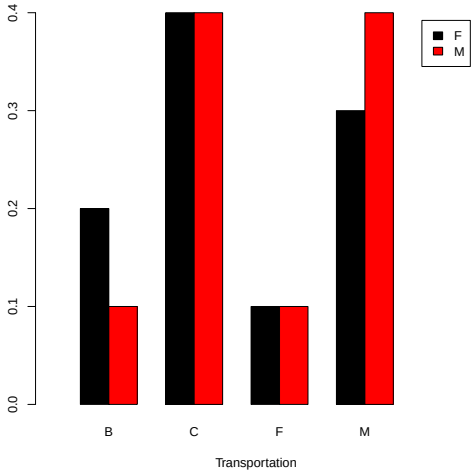


Διάγραμμα 2.4: Παράδειγμα εγγράφου με το R Markdown

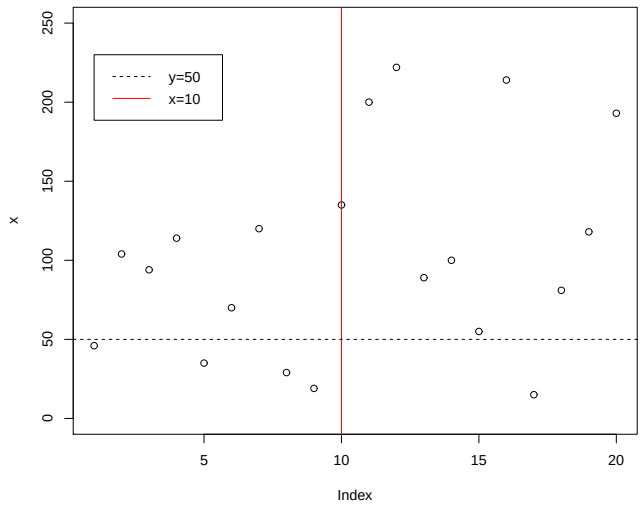
γαριασμό δωρεάν και να ανεβάσετε αρχεία (σας δίνεται όμως περιορισμένη χωρητικότητα). Με την επιλογή "Manage Accounts..." μπορείτε να διαχειριστείτε τον λογαριασμό σας.



Διάγραμμα 3.12: Ομαδοποιημένο ραβδόγραμμα του φύλου δοθέντος του μεταφορικού μέσου για τα δεδομένα του Παραδείγματος 3.3.1



Διάγραμμα 3.13: Ομαδοποιημένο ραβδόγραμμα του μεταφορικού μέσου δοθέντος του φύλου για τα δεδομένα του Παραδείγματος 3.3.1



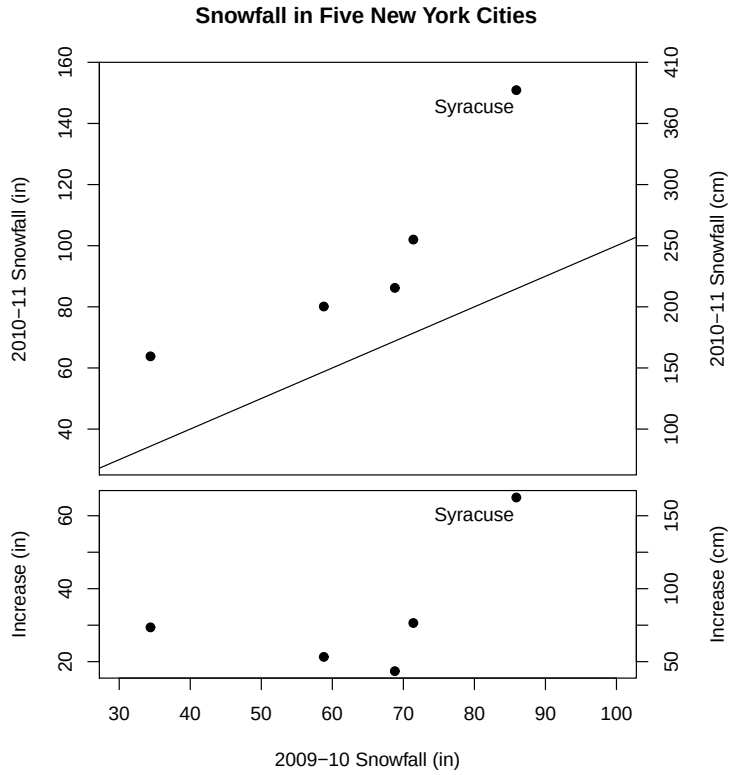
Διάγραμμα 4.23: Χρήση λεζάντας σε ένα διάγραμμα

```
> plot(x, ylim=c(0,250))
> abline(v=10, col="red")
> abline(h=50, lty=2)
> legend(1,230, lty=c(2,1), col=1:2, legend=c("y=50",
        "x=10"))
```

αναπαριστούμε τα σημεία `x` (του Παραδείγματος 3.2.1) με επιθυμητό εύρος τιμών για τον άξονα yy' το $(0, 250)$, εν συνεχεία προσθέτουμε στο παρόν διάγραμμα την ευθεία $x = 10$ (`v=10`) σε χρώμα κόκκινο (`col="red"`), την ευθεία $y = 50$ (`h=50`) με διακεκομμένες γραμμές (`lty=2`) και τέλος προσθέτουμε μία λεζάντα στο σημείο με συντεταγμένες $(1, 230)$, η οποία θα αποτελείται από μια ευθεία (`lty=1`) χρώματος κόκκινου (`col=2`) με συνοδευτικό κείμενο `y = 50` και μια διακεκομμένη ευθεία (`lty=2`) χρώματος μαύρου (`col=1`) με συνοδευτικό κείμενο `x = 10`. Προκύπτει τότε το Διάγραμμα 4.23.

Η θέση της λεζάντας μπορεί επίσης να καθοριστεί αντικαθιστώντας τις συντεταγμένες με τις ακόλουθες λέξεις (σε εισαγωγικά): `"bottomright"` (κάτω δεξιά), `"bottom"` (κάτω), `"bottomleft"` (κάτω αριστερά), `"left"` (αρι-

```
      pch=19, ylab="Increase (in)")
> text(80, 60, "Syracuse")
> tm<-par("yaxp")
> ticmarks<-seq(tm[1], tm[2], length=tm[3] + 1)
> axis(4, at=ticmarks, labels=as.character(round(2.54 *
   ticmarks, -1)))
> mtext("Increase (cm)", side=4, line=3)
```



Διάγραμμα 4.24: Γράφική παράσταση Παραδείγματος 4.4.1

Από το Διάγραμμα 4.24 συμπεραίνουμε ότι και οι πέντε πόλεις της Νέας Υόρκης είχαν περισσότερη χιονόπτωση τη χρονική περίοδο 2010-11 από τη χρονική περίοδο 2009-10. Από το κάτω διάγραμμα συμπεραίνουμε ότι στην πόλη *Syracuse* είχαμε τη μεγαλύτερη αύξηση στη χιονόπτωση (πάνω από 60 ίντσες ή 150 εκατοστά), ενώ στις άλλες 4 πόλεις η αύξηση ήταν περίπου 30

Ασκήσεις

4.1. Θεωρήστε τα δεδομένα της Άσκησης 3.1. Δώστε μια βελτιωμένη εικόνα του ιστογράμματος των βαθμών των φοιτητών δίνοντας χρώμα γκρι στα ορθογώνια, κατάλληλα ονόματα στους άξονες, επεξηγηματικό τίτλο και υπότιτλο. Επίσης προσθέστε δυο κάθετες ευθείες διαφορετικού χρώματος για τη μέση τιμή και τη διάμεσο του δείγματος.

4.2. Θεωρήστε τα δεδομένα της Άσκησης 3.3 για το ύψος των φυτών. Κατασκευάστε μια βελτιωμένη εικόνα των θηκοδιαγραμμάτων του ύψους των φυτών ξεχωριστά για βόρειες και νότιες περιοχές της Ελλάδας. Δώστε κατάλληλο όνομα και διαφορετικό χρώμα σε κάθε θηκοδιάγραμμα (`"turquoise"` και `"salmon"`) καθώς επίσης και κατάλληλο τίτλο στο γράφημα.

4.3. Θεωρήστε τις παρακάτω παρατηρήσεις μιας τ.μ. X που εκφράζει το ποσό των μηνιαίων εξόδων (σε χιλιάδες ευρώ) μιας τετραμελούς οικογένειας για τους μήνες από Ιανουάριο μέχρι Δεκέμβριο του τελευταίου έτους:
2.40 1.18 1.60 1.92 1.69 1.33 1.75 1.81 1.56 1.38 1.45 2.03

i. Αναπαραστήστε γραφικά τις τιμές των παρατηρήσεων του δείγματος μέσω της συνάρτησης `plot()` με τα σημεία του γραφήματος να αναπαρίστανται με αστερίσκους και να συνδέονται με γραμμές. Δώστε μπλε χρώμα στη γραφική παράσταση και επίσης δώστε κατάλληλο τίτλο (χρώματος γκρι) στο γράφημα και κατάλληλους τίτλους στους άξονες.

ii. Έστω ότι διαθέτουμε την ακόλουθη πληροφορία για τα μηνιαία έξοδα (σε χιλιάδες ευρώ) της οικογένειας κατά τους μήνες από Ιανουάριο μέχρι Δεκέμβριο του αμέσως προηγούμενου έτους:
2.22 1.88 2.67 1.46 1.90 1.26 1.44 1.70 1.87 2.12 1.62 1.69
Προσθέστε τη γραφική παράσταση των νέων παρατηρήσεων (σε πράσινο χρώμα) στο γράφημα του προηγούμενου ερωτήματος με χρήση της εντολής `lines()`. Τα νέα σημεία να αναπαρίστανται από κύβους και να ενώνονται με γραμμές μεγαλύτερου πάχους. Δώστε κατάλ-

τότε:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mu \text{ κατά πιθανότητα καθώς το } n \to \infty.$$

Μπορούμε γραφικά (και εμπειρικά) στην R να παρουσιάσουμε το αποτέλεσμα του παραπάνω θεωρήματος όπως φαίνεται στο παράδειγμα που έπεται.

Παράδειγμα 5.3.1.

Έστω $X_1, \ldots, X_n$ τυχαίο δείγμα από την κατανομή *Bernoulli* με παράμετρο p. Η μέση τιμή της εν λόγω κατανομής είναι πεπερασμένη και ισούται με p. Εφαρμόζοντας τότε τον A.N.M.A. έχουμε:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \to p \text{ κατά πιθανότητα καθώς το } n \to \infty.$$

Έστω $p = 0.3$ και $n = 5000$. Για την υλοποίηση του A.N.M.A. στην R πληκτρολογούμε τις εξής εντολές:

```
> x<-rbinom(5000, 1, 0.3)
> xbar<-cumsum(x)/(1:5000)
> plot(xbar, ylab="Cumulative Sample Mean", xlab="n")
> abline(h=0.3)
```

Με την πρώτη εντολή προσομοιώνουμε τυχαίο δείγμα μεγέθους 5000 από την κατανομή *Bernoulli*(0.3) (ή ισοδύναμα από τη Διωνυμική κατανομή με αριθμό επαναλήψεων 1 και πιθανότητα επιτυχίας 0.3). Εν συνεχεία θεωρούμε τη συνάρτηση του δειγματικού μέσου ως ακολουθία τιμών (`xbar<-cumsum(x)/(1:5000)`)[†] και απεικονίζουμε το διάγραμμα της εν λόγω ακολουθίας μαζί

[†]Η εντολή `cumsum(x)` παίρνει ως όρισμα ένα αριθμητικό διάνυσμα `x` και επιστρέφει το αθροιστικό άθροισμα (ένα διάνυσμα του οποίου το πρώτο στοιχείο συμπίπτει με το πρώτο στοιχείο του διανύσματος `x`, το δεύτερο στοιχείο ισούται με το άθροισμα των δύο πρώτων στοιχείων του διανύσματος `x`, κ.ο.κ, το τελευταίο στοιχείο συμπίπτει με το άθροισμα όλων των στοιχείων του διανύσματος `x`). Διαιρώντας τα στοιχεία του αθροιστικού αθροίσματος με τα στοιχεία του διανύσματος `1:5000`, προκύπτει ο αθροιστικός δειγματικός μέσος· ένα διάνυσμα με n στοιχεία, όπου το πρώτο στοιχείο συμπίπτει με το πρώτο στοιχείο του διανύσματος `x`, το δεύτερο στοιχείο ισούται με το δειγματικό μέσο των δύο πρώτων στοιχείων του διανύσματος `x`, κ.ο.κ, το τελευταίο στοιχείο συμπίπτει με το δειγματικό μέσο όλων των στοιχείων του διανύσματος `x`.

ολοκλήρωση του *Gauss* κλπ. Στο Κεφάλαιο 2 είδαμε π.χ. πώς μπορεί η R με τη βοήθεια της εντολής `integrate()` να μας δώσει μια εκτίμηση για την τιμή της παραπάνω ποσότητας I.

Η ***Monte Carlo ολοκλήρωση*** (*Monte Carlo integration*) είναι μία αρκετά απλή, αλλά συγχρόνως και αρκετά γενική μέθοδος εκτίμησης του ολοκληρώματος (5.2). Το ολοκλήρωμα (5.2) μπορεί να γραφεί εναλλακτικά ως:

$$I = \int_a^b w(x) f(x) dx, \tag{5.3}$$

όπου $w(x) = h(x)(b-a)$ και $f(x) = (b-a)^{-1}$. Παρατηρήστε ότι η συνάρτηση f είναι η σ.π.π. της Ομοιόμορφης κατανομής στο διάστημα (a, b). Επομένως αν η τ.μ. X ακολουθεί την Ομοιόμορφη κατανομή στο διάστημα (a, b), έχουμε ότι $I = \mathbb{E}_f[w(X)]$ (δηλαδή το ολοκλήρωμα (5.2) είναι η μέση τιμή της $w(X)$ με $X \sim f$). Με βάση τότε τον ***Ασθενή Νόμο των Μεγάλων Αριθμών* (A.N.M.A.)** αν $X_1, \ldots, X_n$ είναι τυχαίο δείγμα από την Ομοιόμορφη κατανομή στο διάστημα (a, b) τότε:

$$\hat{I} = \frac{1}{n} \sum_{i=1}^{n} w(X_i) \to \mathbb{E}_f[w(X)] = I \text{ κατά πιθανότητα καθώς το } n \to \infty.$$

Αυτή είναι η βασική μέθοδος *Monte Carlo* ολοκλήρωσης. Επιπλέον μπορούμε να εκτιμήσουμε το τυπικό σφάλμα του $\hat{I}$ χρησιμοποιώντας τον παρακάτω εκτιμητή:

$$\widehat{SE}\left(\hat{I}\right) = \frac{S}{\sqrt{n}}, \text{ όπου } S = \frac{\sum_{i=1}^{n} \left(w(X_i) - \hat{I}\right)^2}{n-1}.$$

Παράδειγμα 5.6.1.

Έστω $h(x) = x^3$, $a = 0$ και $b = 1$. Εύκολα προκύπτει ότι $I = 0.25$. Με χρήση της μεθόδου *Monte Carlo* ολοκλήρωσης στην R, για $n = 10000$ παρατηρήσεις έχουμε:

```
> x<-runif(10000)
> mean(x^3)
[1] 0.2496679
```

Όπως φαίνεται και παραπάνω το στατιστικό ελέγχου ακολουθεί την κατανομή *Student* με $(n-1)$ βαθμούς ελευθερίας. Απαραίτητη προϋπόθεση για αυτό είναι:

- ***Τα δεδομένα να προέρχονται από κανονικό πληθυσμό.***

  ή

- ***Το μέγεθος του δείγματος να είναι μεγάλο για να ισχύει το Κ.Ο.Θ.***

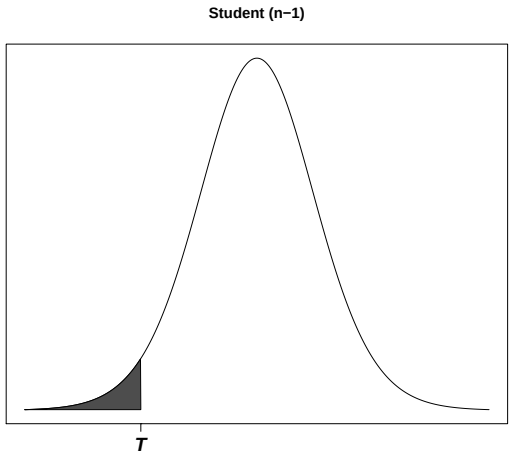
Στη δεύτερη περίπτωση που το μέγεθος του δείγματος είναι μεγάλο, η κατανομή *Student* προσεγγίζεται επίσης με αρκετά μεγάλη ακρίβεια από την Κανονική κατανομή, οπότε μπορούμε να θεωρήσουμε ότι το στατιστικό ελέγχου T ακολουθεί προσεγγιστικά την Τυποποιημένη Κανονική κατανομή. Με βάση λοιπόν τους πίνακες της κατανομής *Student* (ή της N(0,1) για μεγάλο μέγεθος δείγματος) και την εναλλακτική υπόθεση που έχουμε, μπορούμε να βρούμε την P-τιμή του παραπάνω ελέγχου (βλ. Διαγράμματα 6.7, 6.8 και 6.9).

- Αν $H_1 : \mu \neq \mu_0$ τότε η P-τιμή του αμφίπλευρου ελέγχου είναι 2 φορές η πιθανότητα δεξιά της παρατηρηθείσας τιμής του $|T|$ (ή ισοδύναμα 2 φορές η πιθανότητα αριστερά του $-|T|$).
- Αν $H_1 : \mu > \mu_0$ τότε η P-τιμή του μονόπλευρου ελέγχου είναι η πιθανότητα δεξιά της παρατηρηθείσας τιμής του T.
- Αν $H_1 : \mu < \mu_0$ τότε η P-τιμή του μονόπλευρου ελέγχου είναι η πιθανότητα αριστερά της παρατηρηθείσας τιμής του T.

Όλες οι παραπάνω πιθανότητες υπολογίζονται με χρήση της κατανομής *Student* με $(n-1)$ β.ε.

Επιπλέον μπορούμε να κατασκευάσαμε και το $(1-\alpha)\%$ δ.ε. για το μ και να ελέγξουμε αν η υποτιθέμενη τιμή μ_0 κάτω από τη μηδενική υπόθεση ανήκει στο εν λόγω διάστημα. Για τον αμφίπλευρο έλεγχο το εν λόγω διάστημα είναι το:

$$\left(\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}},\ \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \right),$$



Διάγραμμα 6.9: P-τιμή για τον έλεγχο της μηδενικής υπόθεσης $H_0 : \mu = \mu_0$ με εναλλακτική $H_1 : \mu < \mu_0$

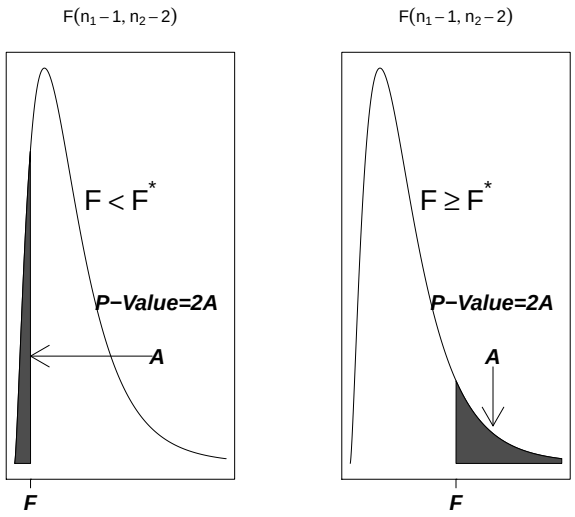
ενώ για τους μονόπλευρους ελέγχους με εναλλακτικές υποθέσεις $H_1 : \mu > \mu_0$ και $H_1 : \mu < \mu_0$ τα διαστήματα εμπιστοσύνης είναι:

$$\left(\bar{x} - t_{n-1,\alpha}\frac{s}{\sqrt{n}},\ +\infty\right) \text{ και } \left(-\infty,\ \bar{x} + t_{n-1,\alpha}\frac{s}{\sqrt{n}}\right)$$

αντίστοιχα. Στα παραπάνω $t_{n-1,\alpha}$ συμβολίζει το α-ποσοστιαίο σημείο της κατανομής *Student* με $n-1$ β.ε. (βλ. Διάγραμμα 6.10). Λόγω συμμετρίας της κατανομής *Student* γύρω από το μηδέν (όπως και στην Κανονική κατανομή) το $(1-\alpha/2)$-ποσοστιαίο σημείο ισούται με το αντίθετο του $\alpha/2$-ποσοστιαίου σημείου.

Ο παρακάτω κώδικας χρησιμοποιήθηκε για την κατασκευή του Διαγράμματος 6.10.

```
> x<-seq(-3, 3, length=1000)
> hx<-dt(x, 20)
> plot(x, hx, type="l", xlab=" ", ylab=" ", yaxt="n",
       main="Student(n-1)", xaxt="n")
> segments(1.96, -0.2, 1.96, dt(1.96,20))
```



Διάγραμμα 6.26: P-τιμή για τον έλεγχο της μηδενικής υπόθεσης $H_0 : \sigma_1^2 = \sigma_2^2$ με εναλλακτική την $H_1 : \sigma_1^2 \neq \sigma_2^2$ για κανονικούς πληθυσμούς. Το F^* δηλώνει τη διάμεσο της κατανομής

$$\left(\frac{1}{F_{n_1-1,n_2-2,\alpha/2}} \frac{s_1^2}{s_2^2}, \ \frac{1}{F_{n_1-1,n_2-2,1-\alpha/2}} \frac{s_1^2}{s_2^2} \right).$$

Στα παραπάνω, $F_{n_1-1,n_2-2,\alpha}$ συμβολίζει το α-ποσοστιαίο σημείο της κατανομής του *Snedecor* με παραμέτρους $(n_1 - 1)$ και $(n_2 - 1)$ (βλ. Διάγραμμα 6.27).

Ο παρακάτω κώδικας χρησιμοποιήθηκε για την κατασκευή του Διαγράμματος 6.27.

```
> x<-seq(0, 4, length=1000)
> hx<-df(x, 5, 30)
> plot(x, hx, type="l", xlab=" ", ylab=" ", yaxt="n",
      main=expression(F(n[1]-1,n[2]-2)), xaxt="n")
> segments(2, -0.2, 2, df(2,5,30))
> text(2.6, 0.01, expression(alpha), cex=1.5, font=4)
> axis(1, at=2, labels=c(expression(F[list(n[1]-1,n[2]-1,
```

```
 NOTE: n is number in *each* group

> power.t.test(n=100, sd=2,  sig.level=0.05, power=0.95,
               type="one.sample", alternative="two.sided")

     One-sample t test power calculation

              n = 100
          delta = 0.7280472
             sd = 2
      sig.level = 0.05
          power = 0.95
    alternative = two.sided

> power.t.test(n=100, delta=1.5, sd=5,  sig.level=0.05,
               type="paired", alternative="one.sided")

     Paired t test power calculation

              n = 100
          delta = 1.5
             sd = 5
      sig.level = 0.05
          power = 0.9089875
    alternative = one.sided

 NOTE: n is number of *pairs*, sd is std.dev. of *differences
    * within pairs
```

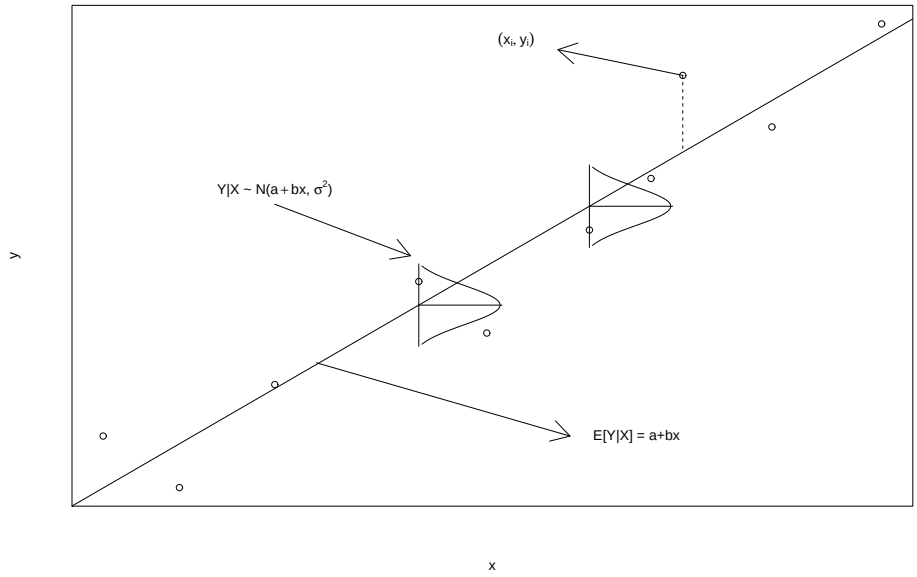
Παραπάνω παραθέτουμε κάποια παραδείγματα της χρήσης στην R της συνάρτησης `power.t.test()`. Στην πρώτη περίπτωση έχουμε ένα *two-sample t-test* όπου η τυπική απόκλιση του υπό μελέτη χαρακτηριστικού και στις δύο ομάδες είναι 2, η εναλλακτική υπόθεση είναι αμφίπλευρη, η ισχύς του ελέγχου θέλουμε να είναι 0.95, το επίπεδο σημαντικότητας να είναι 5%. Για να μπορέσουμε τότε να χαρακτηρίσουμε μια διαφορά μίας μονάδας ως στατιστικά σημαντική χρειαζόμαστε τουλάχιστον 105 παρατηρήσεις σε κάθε ομάδα. Στη

```
> power.prop.test(p1=.50, p2=.75, sig.level=0.05, power=0.95,
                  alternative="one.sided")

     Two-sample comparison of proportions power calculation

              n = 78.43743
             p1 = 0.5
             p2 = 0.75
      sig.level = 0.05
          power = 0.95
    alternative = one.sided

 NOTE: n is number in *each* group
```

Παραπάνω παραθέτουμε κάποια παραδείγματα της χρήσης στην R της συνάρτησης `power.prop.test()`. Στην πρώτη περίπτωση έχουμε έναν έλεγχο ποσοστών, σε επίπεδο σημαντικότητας 5%, με μονόπλευρη εναλλακτική υπόθεση, 50 παρατηρήσεις σε κάθε ομάδα, όπου τα ποσοστά επιτυχίας στις δύο ομάδες είναι 0.50 και 0.75. Τότε η ισχύς του ελέγχου είναι περίπου 0.84. Στη δεύτερη περίπτωση έχουμε έναν έλεγχο ποσοστών, σε επίπεδο σημαντικότητας 5%, με μονόπλευρη εναλλακτική υπόθεση, όπου τα ποσοστά επιτυχίας στις δύο ομάδες είναι 0.50 και 0.75. Για να είναι τότε η ισχύς του ελέγχου 0.95 χρειαζόμαστε 79 παρατηρήσεις σε κάθε ομάδα.

6.4.5 Έλεγχοι Καλής Προσαρμογής

Σε αρκετές περιπτώσεις σε προβλήματα στατιστικής ερχόμαστε αντιμέτωποι με το πρόβλημα επιλογής και προσδιορισμού ***μοντέλου***. Έστω π.χ. ότι έχουμε συλλέξει δεδομένα του επιπέδου ολικής χοληστερίνης σε 20 υγιείς ανθρώπους. Είναι ρεαλιστικό να υποθέσουμε ότι τα δεδομένα προέρχονται από κανονικό πληθυσμό; Αρκετοί από τους ελέγχους που είδαμε στις προηγούμενες παραγράφους (για μικρά μεγέθη δειγμάτων συνήθως) έγιναν υπό την προϋπόθεση ότι το τυχαίο δείγμα προερχόταν από την Κανονική κατανομή. Ο τρόπος που ελέγχαμε την εν λόγω προϋπόθεση ήταν με τη βοήθεια γραφικών παραστάσεων,



Διάγραμμα 7.2: Διάγραμμα απλής γραμμικής παλινδρόμησης

```
> plot(x, y, xaxt="n", yaxt="n")
> abline(lm(y~x))
> lr<-lm(y~x)
> yhat<-predict(lr)
> xx<-seq(-2, 2, .1)
> subplot(plot(dnorm(xx), xx, type="l"‘’, axes=F, xlab=" ",
          ylab=" "), x[5], yhat[5], hadj=0)
> segments(x[5], (yhat[5]-0.8),x[5], (yhat[5]+0.8))
> segments(x[5], yhat[5],(x[5]+1), yhat[5])
> subplot(plot(dnorm(xx), xx, type="l", axes=F, xlab=" ",
          ylab=" "), x[6], yhat[6], hadj=0)
> segments(x[6], (yhat[6]-0.8), x[6], (yhat[6]+0.8))
> segments(x[6], yhat[6], (x[6]+1), yhat[6])
> segments(x[9], y[9],x[9], yhat[9], lty=2)
> text(x[9]-2, y[9]+0.7, expression(group("(",list(x[i],
       y[i]),")")), cex=1, font=4)
> arrows(x[9], y[9], x[9]-1.5, y[9]+0.5)
```

ual) prediction interval) και αποτελεί ένα συμμετρικό $(1-\alpha)\%$ δ.ε. της τιμής, έστω y, της τ.μ. $Y = a + bx + \epsilon$.

Το διάστημα μέσης πρόβλεψης παρέχει πληροφορία για τον βαθμό αβεβαιότητας που έχουμε για την εκτίμηση της δεσμευμένης μέσης τιμής $\mathbb{E}[Y|X = x]$. Το διάστημα (ατομικής) πρόβλεψης παρέχει πληροφορία για τον βαθμό αβεβαιότητας που έχουμε για την τιμή που θα πάρει η τυχαία μεταβλητή Y όταν $X = x$. Το διάστημα (ατομικής) πρόβλεψης δηλαδή λαμβάνει επιπλέον υπόψη, πέραν της αβεβαιότητας που έχουμε από την εκτίμηση της δεσμευμένης μέσης τιμής $\mathbb{E}[Y|X = x]$, και τη μεταβλητότητα της δεσμευμένης κατανομής $Y|(X = x)$. Χρησιμοποιώντας δηλαδή το διάστημα μέσης πρόβλεψης γενικά υποεκτιμούμε την αβεβαιότητά μας για τη χρήση της ποσότητας $\hat{y}$ ως εκτιμητή της τιμής που θα πάρει η τυχαία μεταβλητή Y όταν $X = x$.

Το διάστημα μέσης πρόβλεψης θεωρείται κατάλληλο και χρησιμοποιείται όταν θέλουμε να κατασκευάσουμε διάστημα εμπιστοσύνης για την τιμή, έστω y, της μεταβλητής απόκρισης Y δοσμένης μίας εκ των ήδη παρατηρηθεισών τιμών της επεξηγηματικής μεταβλητής X, για αυτό και λέγεται επίσης και ***διάστημα εμπιστοσύνης προσαρμοσμένων*** (*fitted*) ***τιμών***. Αντιθέτως αν θέλουμε να χρησιμοποιήσουμε μια μελλοντική παρατήρηση, έστω x, της επεξηγηματικής μεταβλητής X τότε για την κατασκευή του διαστήματος εμπιστοσύνης της τιμής y της μεταβλητής απόκρισης Y χρησιμοποιούμε το ***διάστημα (ατομικής) πρόβλεψης***.

7.2.4 Προϋποθέσεις Απλού Γραμμικού Μοντέλου

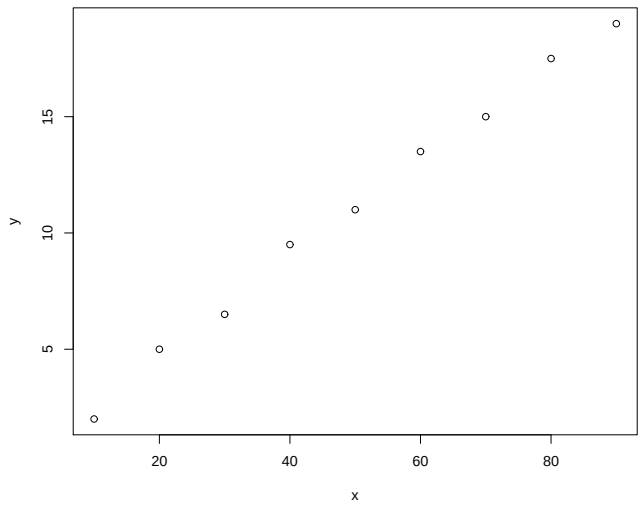
Η εκτίμηση των παραμέτρων ενός απλού γραμμικού μοντέλου με τη μέθοδο των ελαχίστων τετραγώνων προϋποθέτει την ικανοποίηση κάποιων βασικών προϋποθέσεων. Η παραβίαση κάποιας από αυτές τις προϋποθέσεις δημιουργεί πρόβλημα στην εγκυρότητα των αποτελεσμάτων. Στην ενότητα αυτή παρουσιάζουμε εν συντομία τις προϋποθέσεις αυτές καθώς και τρόπους ελέγχου αυτών στο στατιστικό πακέτο R. Η πρώτη προϋπόθεση, που είναι η γραμμικότητα, θα πρέπει να ελέγχεται αρχικώς για να δούμε αν όντως το κατάλληλο μοντέλο που περιγράφει τη σχέση μεταξύ της δεσμευμένης μέσης τιμής της

Πίνακας 7.2: Δεδομένα του Παραδείγματος 7.2.2

x_i	10	20	30	40	50	60	70	80	90
y_i	2.0	5.0	6.5	9.5	11.0	13.5	15.0	17.5	19.0

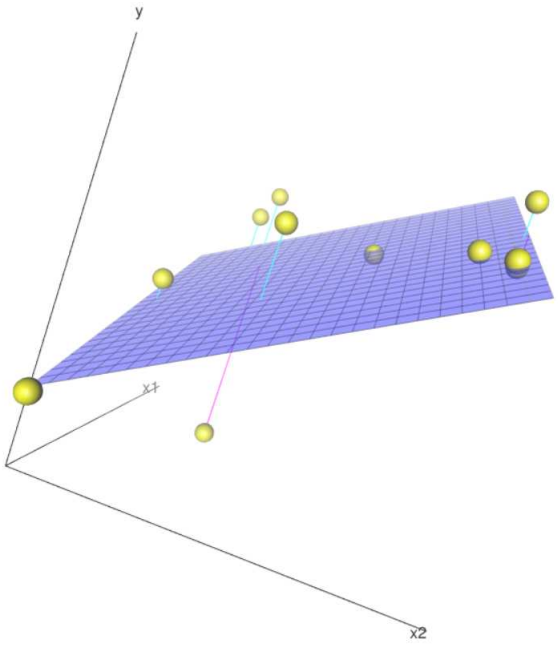
Έλεγχος Γραμμικότητας: Εισάγουμε τα δεδομένα στην R και με το διάγραμμα διασποράς των σημείων (x_i, y_i), $i = 1, \ldots, 9$ ελέγχουμε την υπόθεση της γραμμικότητας.

```
> x<-seq(10, 90, by=10)
> x
[1] 10 20 30 40 50 60 70 80 90
> y<-c(2, 5, 6.5, 9.5, 11, 13.5, 15, 17.5, 19)
> plot(x, y)
```



Διάγραμμα 7.16: Διάγραμμα διασποράς των δεδομένων του Παραδείγματος 7.2.2

Από το Διάγραμμα 7.16 παρατηρούμε ότι η υπόθεση της γραμμικότητας



Διάγραμμα 7.21: Διάγραμμα πολλαπλής γραμμικής παλινδρόμησης με δύο (ποσοτικές) επεξηγηματικές μεταβλητές

την κύλιση της ρόδας του ποντικιού (*scroll weel*) να το μεγεθύνετε. Το διάγραμμα εμφανίζετε στο `RGL device` παράθυρο της R και όχι στο συνηθισμένο γραφικό παράθυρο. Για να το αποθηκεύσετε (στη μορφή που το έχετε εν τέλει φέρει με χρήση του ποντικιού σας) χρησιμοποιείτε την εντολή `rgl.postscript()`. Με το όρισμα `fmt`, αλλάζετε τον τύπο του παραγόμενου αρχείου (μπορείτε να δημιουργήσετε αρχεία `ps`, `eps`, `tex`, `pdf`, `svg` και `pgf`). Με το όρισμα `axis.scales = FALSE` αποκρύπτονται οι τιμές των αξόνων. Τέλος με την παρακάτω εντολή

```
> scatter3d(y=data$y, x=data$x1, z=data$x2, surface=FALSE)
```

αφαιρούμε το επίπεδο παλινδρόμησης και αναπαριστούμε γραφικά μόνο τις τιμές των τριών ποσοτικών μεταβλητών (***τρισδιάστατο διάγραμμα διασποράς***). Για περισσότερες λειτουργίες και λεπτομέρειες της εντολής `scatter3d()`,

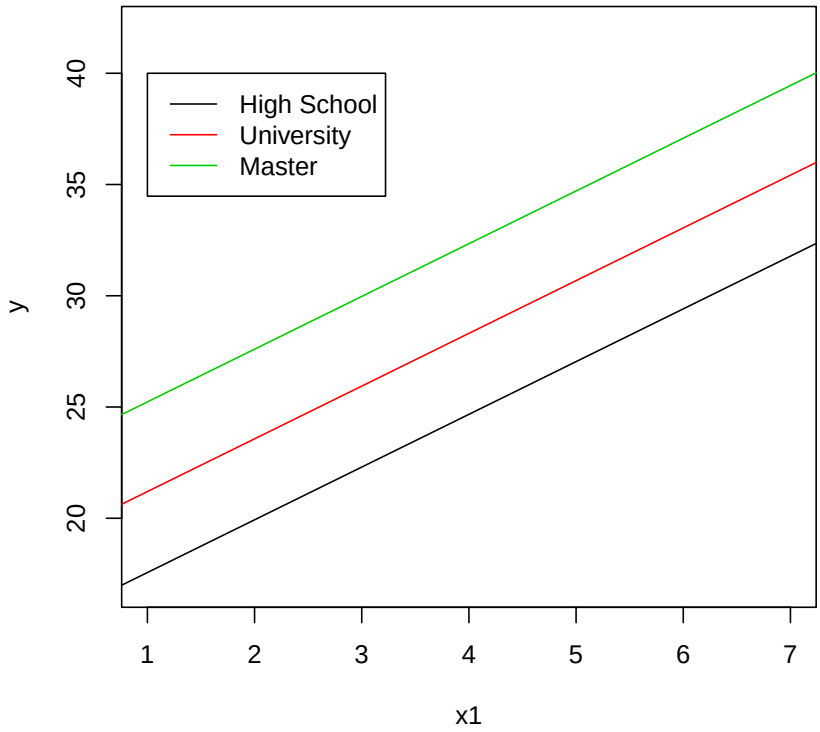
Το εν λόγω διάστημα καλείται ***διάστημα (ατομικής) πρόβλεψης*** (*(individual) prediction interval*) και αποτελεί ένα συμμετρικό $(1-\alpha)\%$ δ.ε. της τιμής, έστω y, της τ.μ. $Y = a + b_1 x_1 + \cdots + b_p x_p + \epsilon$.

Η διαφοροποίηση και η ερμηνεία των παραπάνω δύο διαστημάτων είναι ανάλογες με αυτή στο απλό γραμμικό μοντέλο.

7.3.3 Προϋποθέσεις Πολλαπλού Γραμμικού Μοντέλου

Όπως στο απλό, έτσι και στο πολλαπλό γραμμικό μοντέλο, η εκτίμηση των παραμέτρων με τη μέθοδο των ελαχίστων τετραγώνων προϋποθέτει την ικανοποίηση κάποιων βασικών προϋποθέσεων. Στην ενότητα αυτή παρουσιάζουμε εν συντομία τις προϋποθέσεις αυτές ενώ στο παράδειγμα που έπεται παρουσιάζονται τρόποι ελέγχου αυτών στην R. Η πρώτη προϋπόθεση, που είναι η γραμμικότητα, θα πρέπει να ελέγχεται αρχικώς για να δούμε αν όντως το κατάλληλο μοντέλο που περιγράφει τη σχέση μεταξύ της δεσμευμένης μέσης τιμής της μεταβλητής απόκρισης και των επεξηγηματικών μεταβλητών είναι το γραμμικό, ενώ οι υπόλοιπες προϋποθέσεις θα πρέπει να ελέγχονται σε περίπτωση που θέλουμε να προβούμε σε στατιστική συμπερασματολογία για τις παραμέτρους του μοντέλου ή τις μελλοντικές προβλέψεις, όταν π.χ. θέλουμε να κατασκευάσουμε διαστήματα εμπιστοσύνης ή να διεξάγουμε ελέγχους υποθέσεων.

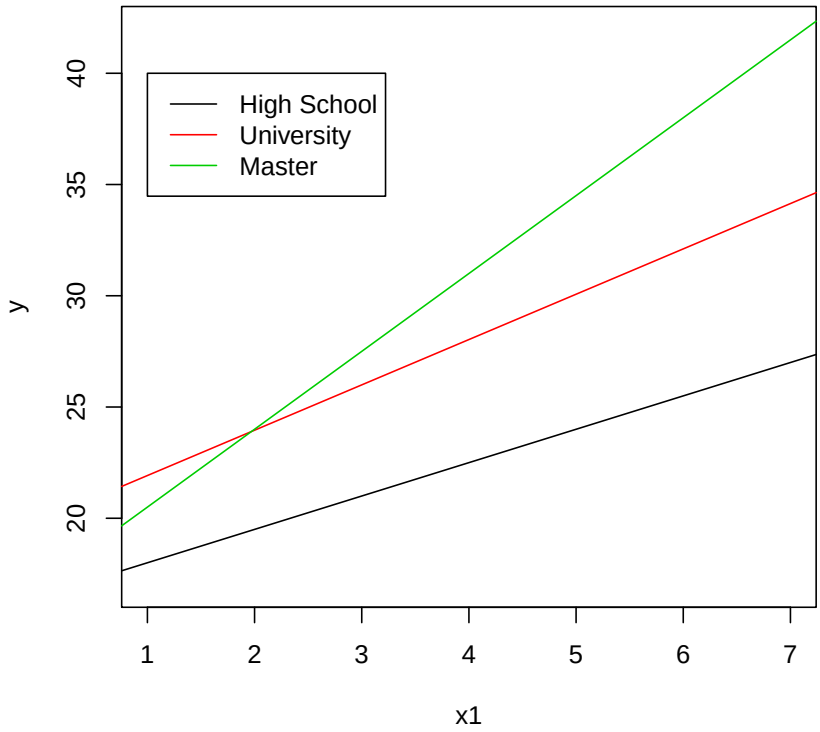
(α) **Γραμμικότητα**: Στο πολλαπλό γραμμικό μοντέλο (7.9) έχουμε υποθέσει ότι η δεσμευμένη μέση τιμή της μεταβλητής απόκρισης Y συνδέεται γραμμικά με το $\boldsymbol{x}$ ή ισοδύναμα ότι τα σφάλματα $\epsilon_i = Y_i - a - b_1 x_{i1} - \cdots - b_p x_{ip}$ $(i = 1, \ldots, n)$ έχουν μέση τιμή μηδέν. Στην περίπτωση της απλής γραμμικής παλινδρόμησης, με μία επεξηγηματική μεταβλητή, η υπόθεση αυτή όπως είδαμε μπορεί εύκολα να ελεγχθεί με το διάγραμμα διασποράς των σημείων (x_i, y_i) $(i = 1, \ldots, n)$. Στην πολλαπλή παλινδρόμηση με p επεξηγηματικές μεταβλητές θα μπορούσαμε να δημιουργήσουμε p διαφορετικά διαγράμματα διασποράς, ένα για κάθε επεξηγηματική μεταβλητή, και να εξετάσουμε την υπόθεση της γραμμικότητας. Βέβαια με τον τρόπο αυτόν δεν ελέγχουμε την εγκυρότητα της σχέσης (7.9), αλλά



Διάγραμμα 7.31: Οι τρεις εκτιμώμενες ευθείες παλινδρόμησης για τα δεδομένα του Παραδείγματος 7.5.1

του γραμμικού μοντέλου, και με τον τρόπο αυτόν μας παρουσιάζονται οι εικονικές μεταβλητές που δημιουργήθηκαν:

```
> model.matrix(results)
   (Intercept) x1 x22 x23
1            1  2   1   0
2            1  1   1   0
3            1  2   0   0
4            1  5   0   1
5            1  7   1   0
```



Διάγραμμα 7.32: Οι τρεις εκτιμώμενες ευθείες παλινδρόμησης με χρήση αλληλεπιδράσεων για τα δεδομένα του Παραδείγματος 7.5.1

τιμή (€3383) της παραμέτρου b_2 εκφράζει τη διαφορά μεταξύ του μέσου μισθού υπαλλήλων με απολυτήριο Λυκείου χωρίς προϋπηρεσία από τον μέσο μισθό υπαλλήλων που είναι απόφοιτοι ΑΕΙ χωρίς προϋπηρεσία, (γ) η εκτιμώμενη τιμή (€500) της παραμέτρου b_3 εκφράζει τη διαφορά μεταξύ του μέσου μισθού υπαλλήλων με απολυτήριο Λυκείου χωρίς προϋπηρεσία από τον μέσο μισθό υπαλλήλων με μεταπτυχιακό τίτλο χωρίς προϋπηρεσία, (δ) η εκτιμώμενη τιμή (€1500) της παραμέτρου b_1 εκφράζει την αναμενόμενη μεταβολή του μισθού ενός υπαλλήλου της εν λόγω εταιρείας με απολυτήριο Λυκείου όταν αυ-

v. Συγκρίνετε την τιμή του κριτηρίου AIC για το μοντέλο του τέταρτου ερωτήματος με και χωρίς αλληλεπίδραση.

7.9. Θεωρήστε τα δεδομένα `trees` της R τα οποία περιέχουν πληροφορία για 31 αγριοκερασιές: τη διάμετρο (`Girth`: σε ίντσες), το ύψος (`Height`: σε πόδια) και τον όγκο της παραγόμενης ξυλείας (`Volume`: σε κυβικά πόδια) (βλ. Άσκηση 3.5).

i. Κατασκευάστε τα διαγράμματα διασποράς των σημείων όλων των ανά δύο μεταβλητών με χρήση της εντολής `pairs()`. Τι συμπεραίνετε για τη σχέση των μεταβλητών;

ii. Δώστε τον πίνακα των τιμών των δειγματικών συντελεστών συσχέτισης για όλες τις ανά δύο μεταβλητές και σχολιάστε.

iii. Προσαρμόστε στα δεδομένα σας τα ακόλουθα γραμμικά μοντέλα:

$$\begin{aligned} \texttt{Volume} &= \alpha_1 + \alpha_2 \cdot \texttt{Height} + \epsilon_1, \quad \epsilon_1 \sim N(0, \sigma_1^2) \\ \texttt{Volume} &= \beta_1 + \beta_2 \cdot \texttt{Girth} + \epsilon_2, \quad \epsilon_2 \sim N(0, \sigma_2^2) \\ \texttt{Volume} &= \gamma_1 + \gamma_2 \cdot \texttt{Height} + \gamma_3 \cdot \texttt{Girth} + \epsilon_3, \quad \epsilon_3 \sim N(0, \sigma_3^2). \end{aligned}$$

Ποιο από τα παραπάνω μοντέλα φαίνεται να προσαρμόζεται καλύτερα στα δεδομένα και γιατί; Για την τεκμηρίωση της απάντησής σας υπολογίστε και την τιμή του κριτηρίου AIC για κάθε μοντέλο.

iv. Θεωρήστε το γραμμικό μοντέλο με μεταβλητή απόκρισης `Volume` και επεξηγηματικές μεταβλητές `Girth` και `Height`. Εφαρμόστε μια τμηματική μέθοδο (π.χ. *backward elimination*) για αυτόματη επιλογή επεξηγηματικών μεταβλητών και συγκρίνετε το αποτέλεσμα με την απάντηση που δώσατε στο προηγούμενο ερώτημα.

v. Προσθέστε στο τρίτο μοντέλο του τρίτου ερωτήματος το τετράγωνο της μεταβλητής `Girth`. Βελτιώνεται η προσαρμογή του μοντέλου στα δεδομένα ή όχι; Για την τεκμηρίωση της απάντησής σας υπολογίστε και την τιμή του κριτηρίου BIC για κάθε μοντέλο.

7.10. Τα δεδομένα που βρίσκονται στην ιστοσελίδα `http://www.math.ntua.gr/~fouskakis/Rbook/fl_crime.txt` αφορούν 67 κομητείες της Φλώ-

κλίσεις) που ακολουθούν Κανονική κατανομή $N(0, \sigma^2)$, με σ άγνωστη σταθερά. Το μοντέλο (8.1) καλείται και ***μοντέλο μέσων της αγωγής*** (*treatment means model*). Αν συμβολίσουμε με $\mathbf{Y} = (Y_1, \ldots, Y_{n_T})^T$ το τυχαίο δείγμα των αποκρίσεων (για όλες τις ομάδες), το παραπάνω μοντέλο μπορεί ισοδύναμα να πάρει τη μορφή:

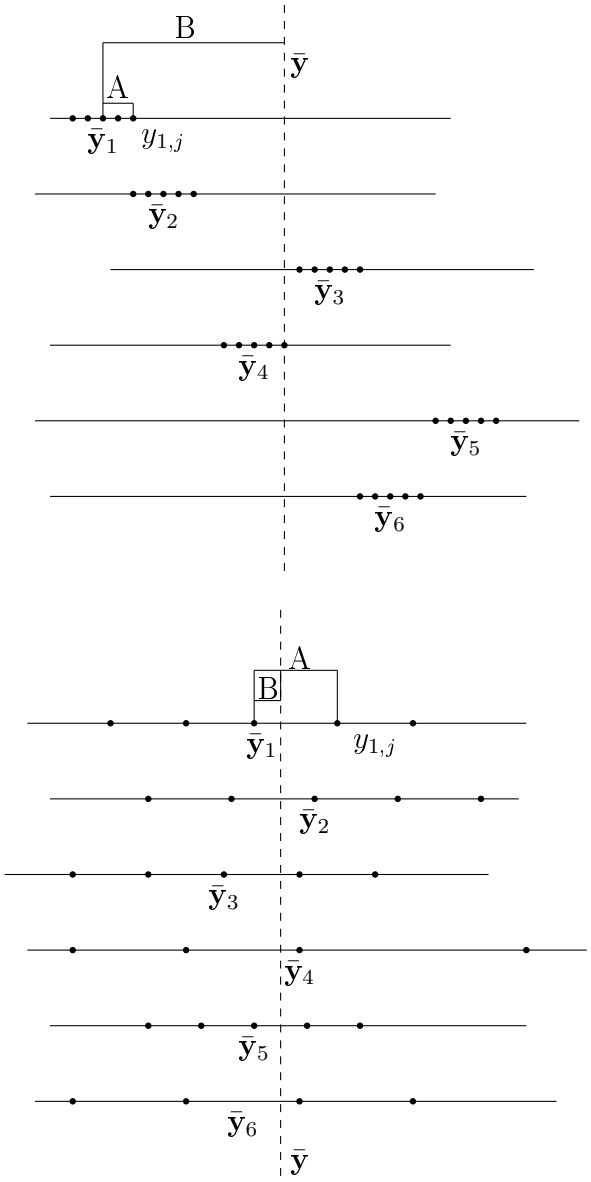
$$\mathbb{E}[Y_\ell | x_{\ell 1}, \ldots, x_{\ell k}] = \sum_{j=1}^{k} \mu_j x_{\ell j}, \quad \ell = 1, \ldots, n_T$$

όπου $x_{\ell j} = 1$ όταν η απόκριση Y_ℓ βρίσκεται στην ομάδα j $(j = 1, \ldots, k)$ και $x_{\ell j} = 0$ στις υπόλοιπες περιπτώσεις. Τέλος η σχέση (8.1) υπό μορφή πινάκων γράφεται:

$$\begin{array}{l} \\ \text{Ομάδα 1} \\ \\ \\ \\ \text{Ομάδα 2} \\ \\ \\ \\ \\ \text{Ομάδα } k\text{-1} \\ \\ \\ \text{Ομάδα } k \\ \\ \end{array}
\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{n_1 1} \\ \hline Y_{12} \\ \vdots \\ Y_{n_2 2} \\ \hline \vdots \\ \vdots \\ \hline Y_{1k-1} \\ \vdots \\ Y_{n_{k-1}k-1} \\ \hline Y_{1k} \\ \vdots \\ Y_{n_k k} \end{pmatrix}
= \left(\begin{array}{cccccc} 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ \hline 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \hline \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \hline 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ \hline 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{array}\right)
\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \vdots \\ \vdots \\ \mu_{k-1} \\ \mu_k \end{pmatrix}
+ \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{n_1 1} \\ \hline \epsilon_{12} \\ \vdots \\ \epsilon_{n_2 2} \\ \hline \vdots \\ \vdots \\ \hline \epsilon_{1k-1} \\ \vdots \\ \epsilon_{n_{k-1}k-1} \\ \hline \epsilon_{1k} \\ \vdots \\ \epsilon_{n_k k} \end{pmatrix}$$

$$\mathbf{Y} \quad = \quad \mathrm{X} \quad \boldsymbol{\mu} \quad + \quad \boldsymbol{\epsilon}.$$

Ο $(n_T \times k)$ πίνακας X είναι ο ***πίνακας σχεδιασμού*** του γραμμικού μοντέλου, και όπως φαίνεται μπορεί να διαμεριστεί με τη βοήθεια k μπλοκ πινάκων



Διάγραμμα 8.1: Σύγκριση μεταβλητότητας μεταξύ και εντός των ομάδων

σμα τετραγώνων μεταξύ των ομάδων (*sum of squares between*, SSB), που μετρά τη μεταβλητότητα των μέσων κάθε ομάδας από τον γενικό μέσο.

ότι για μεγάλα δείγματα, και κάτω από τη μηδενική υπόθεση ισότητας όλων των μέτρων θέσης (σε όλες τις ομάδες), το στατιστικό ελέγχου K ακολουθεί προσεγγιστικά την κατανομή X^2 με $(k-1)$ βαθμούς ελευθερίας, οπότε η P-τιμή του ελέγχου είναι η πιθανότητα (με βάση την κατανομή X^2 με $(k-1)$ βαθμούς ελευθερίας) δεξιά της παρατηρηθείσας τιμής του K. Με χρήση της εντολής `kruskal.test(Red_Cell_Folate, Group)` (ή ισοδύναμα με χρήση της εντολής `kruskal.test(Red_Cell_Folate~Group)`) μπορούμε να εφαρμόσουμε τον μη-παραμετρικό έλεγχο *Kruskal-Wallis* στην R.

```
> kruskal.test(Red_Cell_Folate~Group)

        Kruskal-Wallis rank sum test

data:  Red_Cell_Folate by Group
Kruskal-Wallis chi-squared = 4.1852, df = 2, p-value = 0.1234
```

Παρατηρούμε ότι η P-τιμή είναι $0.1234 > 0.05$, οπότε δεν έχουμε σοβαρές ενδείξεις για να απορρίψουμε την υπόθεση ότι τα "μέσα" επίπεδα φολικού στα ερυθρά αιμοσφαίρια δε διαφοροποιούνται μεταξύ των τριών ομάδων.

■

8.3 Ανάλυση Διασποράς με Δύο Παράγοντες

Ας υποθέσουμε ότι έχουμε μια συνεχή μεταβλητή απόκρισης Y και δύο κατηγορικές επεξηγηματικές μεταβλητές (παράγοντες) X_1 και X_2 με a και b στάθμες αντίστοιχα. Για παράδειγμα, έστω ότι η μεταβλητή Y εκφράζει τη συστολική πίεση ασθενών, η μεταβλητή X_1 την εφαρμοζόμενη θεραπευτική μέθοδο από $a = 3$ διαφορετικές μεθόδους και η μεταβλητή X_2 το φύλο των ασθενών ($b = 2$). Σκοπός μας είναι να δούμε αν υπάρχει διαφορά στις τιμές της μεταβλητής απόκρισης σε σχέση με τις διαφορετικές στάθμες κάθε παράγοντα χωριστά (***κύριες επιδράσεις των δύο παραγόντων***), καθώς επίσης και αν συγκεκριμένοι συνδυασμοί δύο σταθμών των δύο παραγόντων δημιουργούν διαφοροποιημένα αποτελέσματα, δηλαδή έχουμε ύπαρξη ***αλληλεπίδρασης*** (*interaction*), που σημαίνει ότι η κύρια επίδραση ενός παράγοντα δεν είναι η ίδια

Πίνακας 8.4: Δεδομένα και δειγματικοί μέσοι ενός ισορροπημένου παραγοντικού σχεδιασμού δύο παραγόντων

		Παράγοντας X_2				
		$j=1$	$j=2$	$\cdots$	$j=b$	Δειγματικοί Μέσοι
Παράγοντας X_1	$i=1$	$\begin{pmatrix} y_{111} \\ y_{112} \\ \vdots \\ y_{11n} \end{pmatrix} \bar{y}_{11.}$	$\begin{pmatrix} y_{121} \\ y_{122} \\ \vdots \\ y_{12n} \end{pmatrix} \bar{y}_{12.}$	$\cdots$	$\begin{pmatrix} y_{1b1} \\ y_{1b2} \\ \vdots \\ y_{1bn} \end{pmatrix} \bar{y}_{1b.}$	$\bar{y}_{1..}$
	$i=2$	$\begin{pmatrix} y_{211} \\ y_{212} \\ \vdots \\ y_{21n} \end{pmatrix} \bar{y}_{21.}$	$\begin{pmatrix} y_{221} \\ y_{222} \\ \vdots \\ y_{22n} \end{pmatrix} \bar{y}_{22.}$	$\cdots$	$\begin{pmatrix} y_{2b1} \\ y_{2b2} \\ \vdots \\ y_{1bn} \end{pmatrix} \bar{y}_{2b.}$	$\bar{y}_{2..}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$i=a$	$\begin{pmatrix} y_{a11} \\ y_{a12} \\ \vdots \\ y_{a1n} \end{pmatrix} \bar{y}_{a1.}$	$\begin{pmatrix} y_{a21} \\ y_{a22} \\ \vdots \\ y_{a2n} \end{pmatrix} \bar{y}_{a2.}$	$\cdots$	$\begin{pmatrix} y_{ab1} \\ y_{ab2} \\ \vdots \\ y_{abn} \end{pmatrix} \bar{y}_{ab.}$	$\bar{y}_{a..}$
Δειγματικοί Μέσοι		$\bar{y}_{.1.}$	$\bar{y}_{.2.}$	$\cdots$	$\bar{y}_{.b.}$	$\bar{y}_{...}$

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk},\ i = 1, \ldots, a,\ j = 1, \ldots, b \text{ και } k = 1, \ldots, n \tag{8.8}$$

όπου μ_{ij} είναι άγνωστες σταθερές (παράμετροι) και ϵ_{ijk} είναι ανεξάρτητες τ.μ. που ακολουθούν την Κανονική κατανομή με μέση τιμή 0 και άγνωστη διασπορά σ^2.

Παρατήρηση 8.3.1.

1. Η παρατήρηση y_{ijk} που αφορά την k τιμή της τ.μ. Y στην i στάθμη του παράγοντα X_1 και την j στάθμη του παράγοντα X_2, με βάση το μοντέλο (8.8), είναι το άθροισμα μιας σταθερής ποσότητας μ_{ij} και ενός τυχαίου σφάλματος ϵ_{ijk}.

2. Επειδή $\mathbb{E}[\epsilon_{ijk}] = 0$ έχουμε ότι $\mathbb{E}[Y_{ijk}] = \mu_{ij}$. Άρα η παράμετρος μ_{ij} δηλώνει την αναμενόμενη τιμή της τ.μ. Y για την i στάθμη του παράγοντα X_1 και τη j στάθμη του παράγοντα X_2.
3. Επειδή η παράμετρος μ_{ij} είναι σταθερή ποσότητα, έχουμε ότι $\mathbb{V}[Y_{ijk}] = \mathbb{V}[\epsilon_{ijk}] = \sigma^2$. Άρα η διασπορά της τ.μ. Y είναι σταθερή ανεξαρτήτως αγωγής.
4. Επειδή για τα τυχαία σφάλματα ισχύει ότι $\epsilon_{ijk} \sim N(0, \sigma^2)$ έπεται από την (8.8) ότι $Y_{ijk} \sim N(\mu_{ij}, \sigma^2)$. Επιπλέον λόγω της ανεξαρτησίας των ϵ_{ijk} έχουμε ότι οι τ.μ. Y_{ijk}, $i = 1, \ldots, a$, $j = 1, \ldots, b$ και $k = 1, \ldots, n$ είναι ανεξάρτητες. Αυτό σημαίνει ότι η τιμή που θα πάρει η τ.μ. Y σε κάποια αγωγή δεν εξαρτάται από την τιμή που έχει πάρει η ίδια τ.μ. στην ίδια ή σε κάποια άλλη αγωγή.

■

Από τις παραπάνω παρατηρήσεις συμπεραίνουμε ότι οι προϋποθέσεις που πρέπει να πληρούνται για ένα μοντέλο ανάλυσης διασποράς με δύο παράγοντες είναι:

(α) Η κατανομή που ακολουθεί η μεταβλητή απόκρισης Y στις διαφορετικές αγωγές είναι ***Κανονική***.

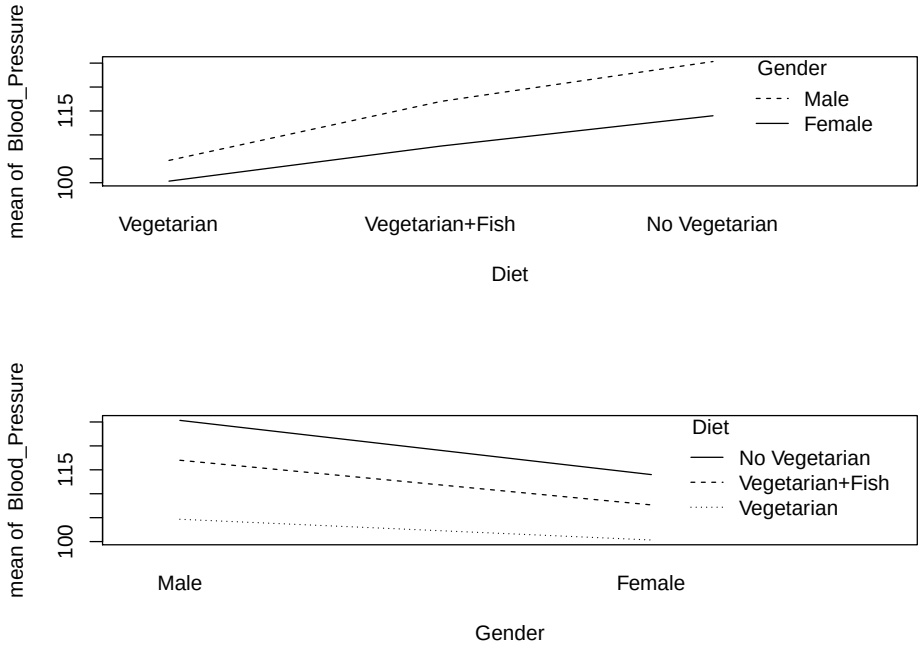
(β) Η διασπορά σ^2 της δεσμευμένης κατανομής της μεταβλητής απόκρισης Y δοθέντων των επεξηγηματικών μεταβλητών X_1 και X_2 είναι ίδια σε όλες τις αγωγές (***υπόθεση ομοσκεδαστικότητας***).

(γ) Οι αποκρίσεις μέσα σε κάθε αγωγή είναι ανεξάρτητες μεταξύ τους καθώς και ***ανεξάρτητες*** από αυτές των άλλων αγωγών.

Υποθέτοντας ότι:
(α) για τη στάθμη i $(i = 1, \ldots, a)$ του παράγοντα X_1 ο μέσος

$$\mu_{i.} = b^{-1} \sum_{j=1}^{b} \mu_{ij}$$

ραπάνω δύο διαγράμματα είναι περίπου παράλληλα μεταξύ τους, είναι αρκετά πιθανό οι αλληλεπιδράσεις να μην είναι στατιστικά σημαντικές. Επιπλέον, το γεγονός ότι και στα δύο διαγράμματα τα ίχνη μεταξύ τους διαφοροποιούνται ξεκάθαρα μας δηλώνει ότι πιθανότατα και οι δύο παράγοντες έχουν στατιστικά σημαντικές επιδράσεις. Με τη βοήθεια της εντολής `boxplot( )` μπορούμε επιπλέον να λάβουμε θηκοδιαγράμματα για κάθε ομάδα κάθε παράγοντα ξεχωριστά.



Διάγραμμα 8.10: Διαγράμματα αλληλεπιδράσεων για τα δεδομένα του Παραδείγματος 8.3.1

```
> par(mfrow=c(2,1))
> boxplot(Blood_Pressure~Diet)
> boxplot(Blood_Pressure~Gender)
```

Από το Διάγραμμα 8.11 παρατηρούμε ότι στο δείγμα οι μετρήσεις αρτηριακής πίεσης διαφοροποιούνται στις τρεις ομάδες του παράγοντα "Διατροφικές

```
Diet:Gender  2  39.00   19.50  0.4730 0.634243
Residuals   12 494.67   41.22
---
Signif. codes:
    0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> model.tables(A)
Tables of effects

 Diet
Diet
     Vegetarian Vegetarian+Fish   No Vegetarian
         -9.000           0.833           8.167

 Gender
Gender
  Male Female
 4.167 -4.167

 Diet:Gender
                 Gender
Diet              Male Female
  Vegetarian      -2.0  2.0
  Vegetarian+Fish  0.5 -0.5
  No Vegetarian    1.5 -1.5

> model.tables(A, type="means")
Tables of means
Grand mean

111.5

 Diet
Diet
     Vegetarian Vegetarian+Fish   No Vegetarian
         102.50          112.33          119.67

 Gender
```

σμό. Παραδείγματος χάρη δεν υπάρχει η ομάδα "Κλινική Α1" του παράγοντα X_2 για κάθε μία από τις ομάδες του παράγοντα X_1: "Νοσοκομείο Α", "Νοσοκομείο Β", "Νοσοκομείο Γ" και "Νοσοκομείο Δ". Αλλά ακόμα και να υπήρχε δε θα αφορούσε την ίδια ομάδα (άλλο είναι η "Κλινική Α1" του "Νοσοκομείου Α" και άλλο είναι η "Κλινική Α1" του "Νοσοκομείου Β"). Στους ιεραρχικούς σχεδιασμούς δε μας ενδιαφέρει να κάνουμε συγκρίσεις μεταξύ των σταθμών του παράγοντα X_2 που βρίσκονται σε διαφορετικές στάθμες του παράγοντα X_1 (δεν είναι λογικό να συγκρίνουμε κλινικές διαφορετικών νοσοκομείων). Αντίθετα θέλουμε να συγκρίνουμε τις στάθμες του παράγοντα X_1 καθώς και πιθανές διαφοροποιήσεις μεταξύ σταθμών του παράγοντα X_2 κάτω από συγκεκριμένη στάθμη του παράγοντα X_1 (σύγκριση νοσοκομείων καθώς και πιθανές διαφορές στις αποκρίσεις μεταξύ κλινικών που βρίσκονται στο ίδιο νοσοκομείο). Η παρουσίαση τέτοιων μοντέλων ξεφεύγει των σκοπών του παρόντος βιβλίου.

8.7 Ανάλυση Διασποράς με Επαναλαμβανόμενες Μετρήσεις

Αρκετές φορές στις στατιστικές μελέτες συναντάμε το φαινόμενο των εξαρτημένων δειγμάτων. Π.χ. (α) ας υποθέσουμε ότι έχουμε μετρήσεις ***της ίδιας ποσοτικής μεταβλητής*** Y (μεταβλητή απόκρισης) για ***τα ίδια άτομα*** για k διαφορετικές θεραπείες ή (β) ας υποθέσουμε ότι υπολογίζουμε την επίδοση ***των ίδιων ατόμων*** σε k διαφορετικά τεστ που έχουν την ***ίδια κλίμακα βαθμολόγησης*** ή (γ) ας υποθέσουμε ότι έχουμε μετρήσεις ***της ίδιας ποσοτικής μεταβλητής*** Y (μεταβλητή απόκρισης) για ***τα ίδια άτομα*** σε k διαφορετικές χρονικές στιγμές.

Στο τελευταίο παράδειγμα λόγου χάρη ας καλέσουμε Y_j τη μεταβλητή απόκρισης τη χρονική τιμή j $(j = 1, \ldots, k)$ και ας θεωρήσουμε ότι προέρχεται από πληθυσμό με άγνωστη μέση τιμή μ_j και άγνωστη τυπική απόκλιση σ_j. Έστω ότι διαθέτουμε τυχαίο δείγμα μεγέθους n και με y_{ij} καλούμε την τιμή της i-οστής παρατήρησης τη j χρονική στιγμή $(i = 1, 2, ..., n,$ και $j = 1, 2, ..., k)$.

Κεφάλαιο 9

Απεικονίσεις, Χειρισμοί και Διαδικτυακές Εφαρμογές

9.1 Εισαγωγή

Η σύγχρονη εποχή την οποία διανύουμε χαρακτηρίζεται ως εποχή των ***Μεγάλων Δεδομένων*** ή ***Δεδομένων Μεγάλης Κλίμακας*** (*Big Data*). Με τον όρο δεδομένα μεγάλης κλίμακας αναφερόμαστε σε δεδομένα με τεράστιο όγκο, ο οποίος μάλιστα σε αρκετές περιπτώσεις αυξάνεται εκθετικά με τον χρόνο. Ο όρος "όγκος" αναφέρεται συνήθως και σε πλήθος γραμμών (παρατηρήσεις) και σε πλήθος στηλών (μεταβλητές). Επομένως, τα δεδομένα μεγάλης κλίμακας διαθέτουν, συνήθως, πληροφορία από πολύ μεγάλο πλήθος μονάδων του πληθυσμού και για πολλά χαρακτηριστικά αυτών. Τα μεγάλης κλίμακας δεδομένα έχουν προκύψει από τις τελευταίες τεχνολογικές εξελίξεις κυρίως στον τομέα των επικοινωνιών και των ολοκληρωμένων κυκλωμάτων, οι οποίοι έχουν δώσει τη δυνατότητα να δημιουργηθούν μηχανισμοί παρακολούθησης των λειτουργιών ενός οργανισμού σε πολύ λεπτομερές επίπεδο. Επίσης, δεδομένα μεγάλης κλίμακας βρίσκουμε και σε μικρότερη κλίμακα οργάνωσης, στο επίπεδο του ατόμου. Οι περισσότεροι άνθρωποι διαθέτουν πλέον έναν ψηφιακό εαυτό, ως προβολή των δραστηριοτήτων τους στα κοινωνικά δίκτυα. Η *Google* εκτιμά ότι κάθε δύο μέρες το ψηφιακό υλικό που δημιουργείται από τους χρή-

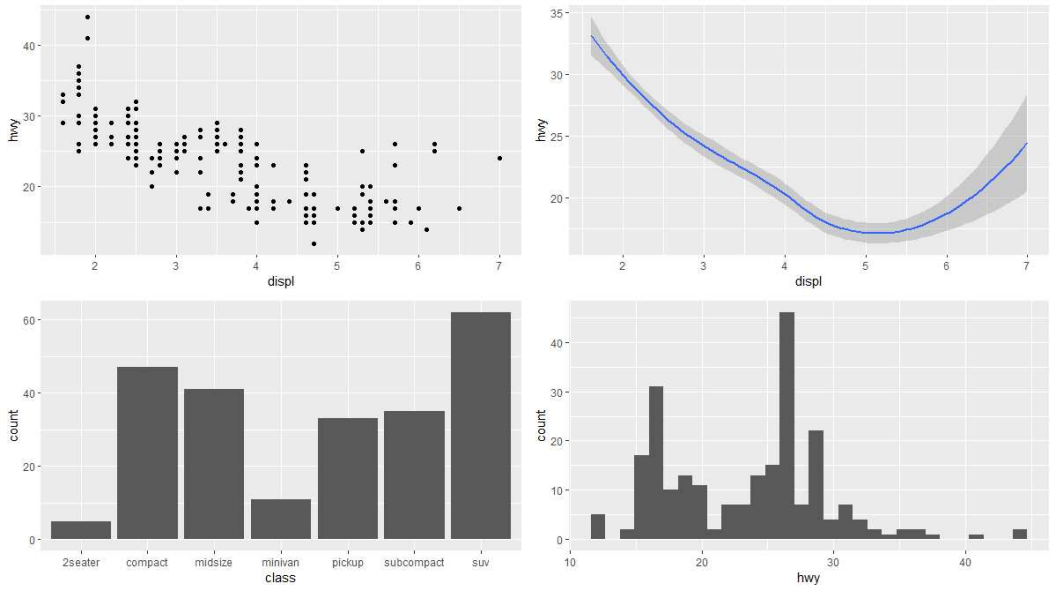
λεσματικούς μηχανισμούς χειρισμού δεδομένων, οι οποίοι μειώνουν αισθητά τη χρήση μνήμης του υπολογιστή. Είναι μια από τις βιβλιοθήκες με τις περισσότερες λήψεις στην R. Είναι πλέον ιδιαίτερα δημοφιλής σε Στατιστικούς και Επιστήμονες Δεδομένων, αφού με τη χρήση της ο ερευνητής πληκτρολογεί, με αποτελεσματικό τρόπο, λιγότερες γραμμές κώδικα και παίρνει ταχύτερα αποτελέσματα. Τέλος, θα αναφερθούμε στη βιβλιοθήκη `shiny`, η οποία μέσω του `RStudio` (βλ. Παράγραφο 2.13), σας επιτρέπει να δημιουργείτε ισχυρές διαδραστικές διαδικτυακές εφαρμογές από την R.

9.2 Η Βιβλιοθήκη ggplot2

Η δυνατότητα δημιουργίας διαγραμματικών απεικονίσεων των δεδομένων είναι ένα σημαντικό βήμα για την εξόρυξη πληροφορίας και ευρημάτων. Σε αυτήν την παράγραφο θα χρησιμοποιήσουμε τη βιβλιοθήκη `ggplot2` για την κατασκευή "κομψών" και σύνθετων διαγραμμάτων. Η βιβλιοθήκη `ggplot2` εφαρμόζει τη ***Γραμματική Γραφικών*** (*Grammar of Graphics*), γεγονός που την καθιστά ιδιαίτερα αποτελεσματική για την περιγραφή του τρόπου με τον οποίο οι απεικονίσεις πρέπει να αντιπροσωπεύουν τα δεδομένα και την έχει μετατρέψει σε μια κορυφαία βιβλιοθήκη σχεδίασης στην R. Η εκμάθηση της συγκεκριμένης βιβλιοθήκης επιτρέπει σχεδόν κάθε είδους οπτικοποίηση των δεδομένων, προσαρμοσμένη στις ακριβείς προδιαγραφές του χρήστη.

Ακριβώς όπως η γραμματική της γλώσσας μάς βοηθά να κατασκευάσουμε σημαντικές προτάσεις από λέξεις, η γραμματική των γραφικών μάς βοηθά να κατασκευάσουμε διαγράμματα από διαφορετικά "οπτικά στρώματα", όπως:

- Τα ***δεδομένα*** που θέλουμε να αναπαραστήσουμε.
- Τα ***γεωμετρικά αντικείμενα*** (κύκλοι, γραμμές, κ.λπ.) που θέλουμε να συμπεριλάβουμε.
- Το σύνολο ***αντιστοιχιών*** από τις μεταβλητές των δεδομένων στην ***αισθητική*** (εμφάνιση) των γεωμετρικών αντικειμένων που θα εφαρμόσουμε.



Διάγραμμα 9.4: Διαφορετικά διαγράμματα geom της ggplot()

τέσσερα αυτά διαγράμματα στο ίδιο γραφικό παράθυρο[†]:

```
> install.packages("cowplot")
> library(cowplot)

> p1<-ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_point()

> p2<-ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_smooth()

> p3<-ggplot(data = mpg, aes(x = class)) +
  geom_bar()

> p4<-ggplot(data = mpg, aes(x = hwy)) +
  geom_histogram()
```

[†]Οι εντολές par(mfrow=c(,)) και par(mfcol=c(,)) που είχαμε δει στο Κεφάλαιο 4, δεν δουλεύουν στην βιβλιοθήκη ggplot2.

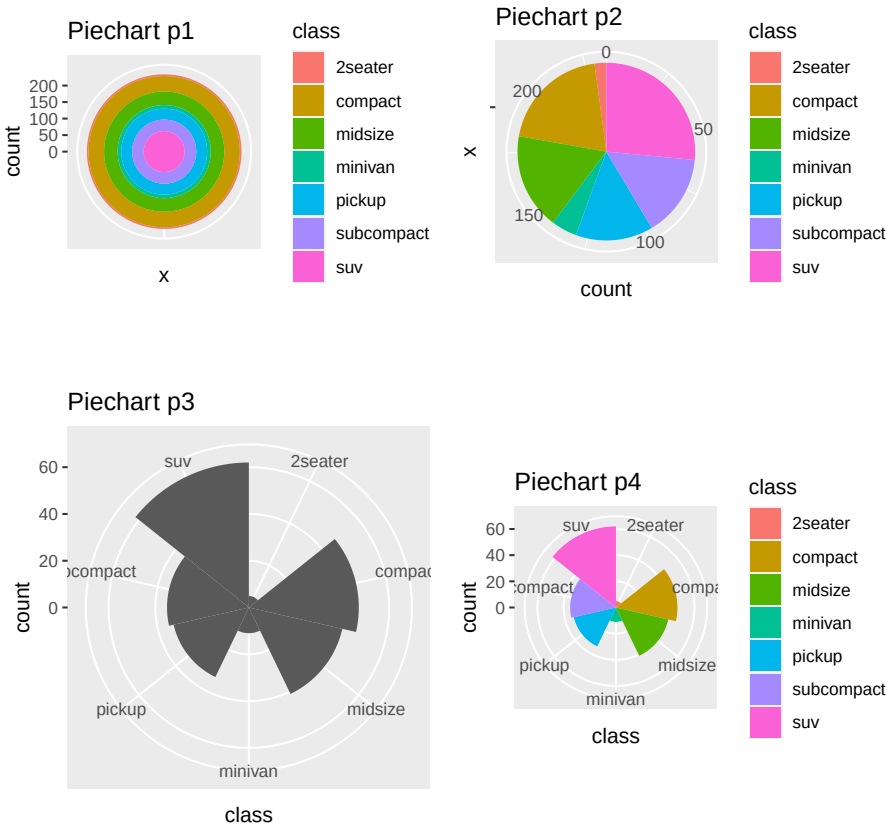
9.2.2 Συνήθεις Διαγραμματικές Απεικονίσεις

Στην παρούσα παράγραφο θα δούμε, μέσω εφαρμογών, τις συνηθέστερες διαγραμματικές απεικονίσεις που κατασκευάζουμε για να περιγράψουμε τις τιμές ποσοτικών και κατηγορικών μεταβλητών, καθώς επίσης και τα διαγράμματα που συνήθως δημιουργούμε για να ανακαλύψουμε πιθανές συσχετίσεις. Θα ξεκινήσουμε με απλές εφαρμογές, όπου επιθυμούμε να αναπαραστήσουμε και να περιγράψουμε γραφικά τις τιμές μιας μεταβλητής και εν συνεχεία θα προβούμε σε γραφήματα των τιμών περισσοτέρων από μίας μεταβλητής. Οι περισσότερες διαγραμματικές απεικονίσεις θα γίνουν με τη βοήθεια της βιβλιοθήκης `ggplot2`.

Θα χρησιμοποιήσουμε τα δεδομένα `mpg` (βλ. Πίνακα 9.1) καθώς και τα δεδομένα `midwest`, της βιβλιοθήκης `ggplot2`, που δίνουν δημογραφικά χαρακτηριστικά μεσοδυτικών κομητειών στις Η.Π.Α. (βλ. `?midwest`). Στο συγκεκριμένο σετ δεδομένων έχουμε πληροφορία για 28 χαρακτηριστικά (ποιοτικά και κατηγορικά) 437 κομητειών. Επίσης, θα χρησιμοποιήσουμε τα δεδομένα `mtcars` (βλ. `?mtcars`) της βιβλιοθήκης `ggplot2`, τα οποία περιλαμβάνουν την κατανάλωση καυσίμου και 10 πτυχές του σχεδιασμού και της απόδοσης 32 αυτοκίνητων (μοντέλα $1973-74$). Επιπλέον, θα χρησιμοποιήσουμε τα δεδομένα `gapminder` (βλ. `?gapminder`) της βιβλιοθήκης `gapminder` με δημογραφικά χαρακτηριστικά 142 χωρών από το 1952 μέχρι το 2007 (ανά 5 χρόνια). Τέλος, θα χρησιμοποιήσουμε τα οικονομικά δεδομένα χρονολογικής σειράς `economics` (βλ. `?economics`) 574 παρατηρήσεων σε 6 μεταβλητές της βιβλιοθήκης `ggplot2` και τα δεδομένα `AirPassengers` που μας δίνουν τον μηνιαίο αριθμό επιβατών διεθνών αεροπορικών εταιρειών από το 1949 μέχρι το 1960.

9.2.2.1 Μία Ποσοτική Μεταβλητή

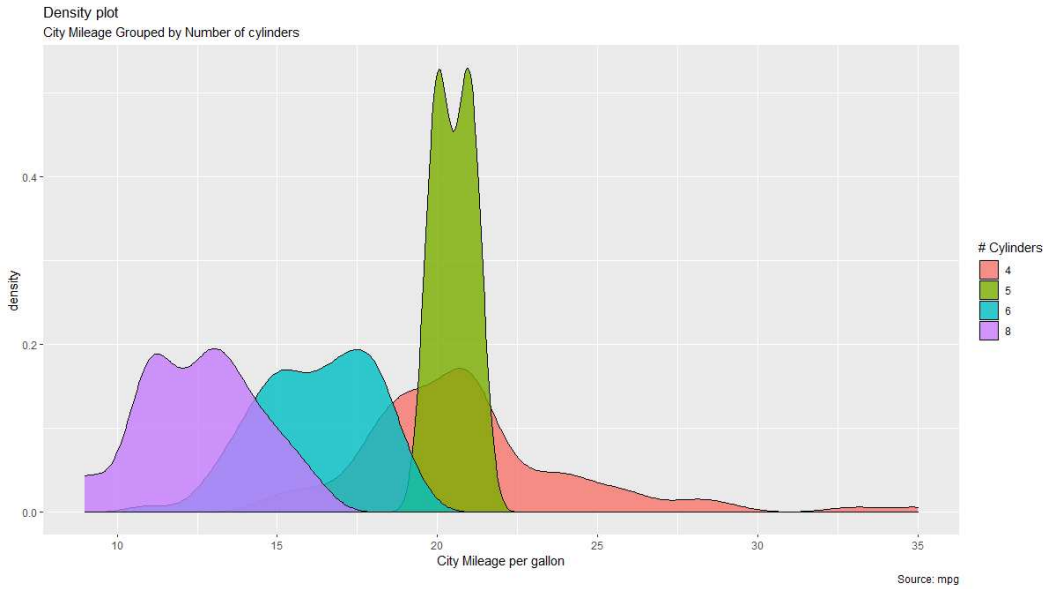
Τις τιμές μιας ποσοτικής μεταβλητής συνήθως τις αναπαριστούμε με τη βοήθεια ενός ***ιστογράμματος*** ή ενός ***θηκοδιαγράμματος***, όπως ήδη έχουμε δει. Όπως θα δούμε στη συνέχεια, τα θηκοδιαγράμματα αποτελούν εξαιρετικά εργαλεία για να μελετήσουμε την κατανομή των τιμών μιας ποσοτικής μεταβλητής σε διάφορες στάθμες μιας κατηγορικής μεταβλητής. Η κάτω δεξιά



Διάγραμμα 9.29: Διαφορετικά τομεογράμματα της βιβλιοθήκης `ggplot2`

9.2.2.3 Δύο Ποσοτικές Μεταβλητές

Όταν έχουμε τιμές από δύο ποσοτικές μεταβλητές θέλουμε, συνήθως, να εξετάσουμε αν υπάρχει κάποιο είδος εξάρτησης και για τον λόγο αυτόν κατασκευάζουμε ένα ***διάγραμμα διασποράς*** (βλ. Διάγραμμα 9.1). Στο διάγραμμα αυτό, αν το επιθυμούμε, μπορούμε να προσθέσουμε μια καμπύλη εξομάλυνσης (`geom_smooth()`) μαζί με ένα 95% διάστημα εμπιστοσύνης γύρω από αυτήν την καμπύλη. Στη συνάρτηση `geom_smooth()`, με το όρισμα `method`, δηλώνουμε τη μέθοδο εξομάλυνσης που θα χρησιμοποιηθεί και με το όρισμα `se` δηλώνουμε



Διάγραμμα 9.37: Διαγράμματα πυκνοτήτων μέσω της βιβλιοθήκης ggplot2

και στη λεζάντα.

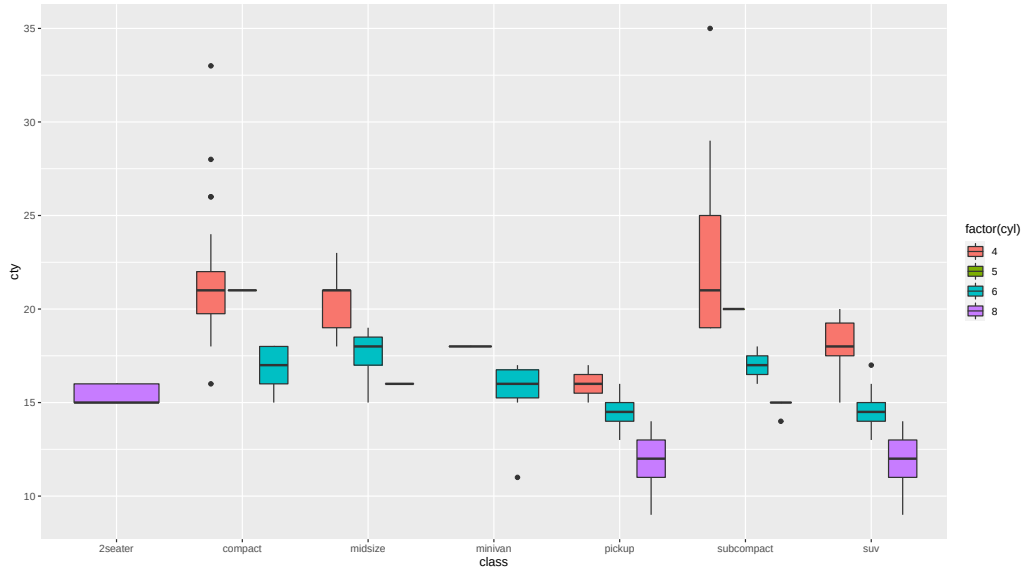
Το ***διάγραμμα βιολιού*** (*violin plot*) μας παρουσιάζει την καμπύλη εκτίμησης της συνάρτησης πυκνότητας πιθανότητας της ποσοτικής μεταβλητής σε κάθε στάθμη της κατηγορικής μεταβλητής. Για λόγους καλαισθησίας, η συγκεκριμένη καμπύλη σχεδιάζεται δύο φορές (συμμετρικά), δεξιά και αριστερά από κάθε ετικέτα στον xx' άξονα, δηλώνοντας το όνομα κάθε στάθμης της κατηγορικής μεταβλητής. Είναι χρήσιμο το διάγραμμα βιολιού να συνοδεύεται από το *Tufte boxplot*. Με τον παρακάτω κώδικα κατασκευάζουμε το Διάγραμμα 9.38:

```
> library(ggthemes)
> ggplot(mpg, aes(x = class, y = cty)) +
  geom_violin(color="red") +
  geom_tufteboxplot()
```

9.2.2.7 Μία Ποσοτική και δύο Κατηγορικές Μεταβλητές

Μπορούμε να κατασκευάσουμε ***ομαδοποιημένα θηκοδιαγράμματα*** σε περιπτώσεις που διαθέτουμε πληροφορία από μία ποσοτική και δύο κατηγορικές μεταβλητές. Για να το πετύχουμε αυτό, προσθέτουμε στη συνάρτηση `geom_boxplot()`, του δεύτερου επιπέδου, μία αισθητική αντιστοίχιση `fill`. Για παράδειγμα, με τον παρακάτω κώδικα βλέπουμε, στα δεδομένα `mpg`, την κατανομή των τιμών της μεταβλητής `cty` για αυτοκίνητα με διαφορετικό αριθμό κυλίνδρων και διαφορετική κατηγορία (βλ. Διάγραμμα 9.39):

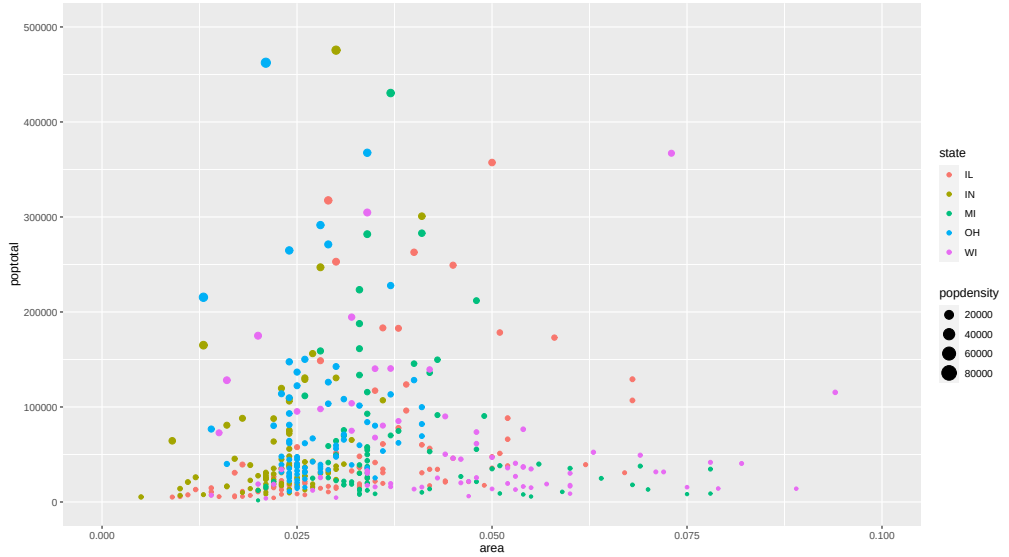
```
> ggplot(mpg, aes(x=class, y=cty)) +
  geom_boxplot(aes(fill=factor(cyl)))
```



Διάγραμμα 9.39: Ομαδοποιημένα θηκοδιαγράμματα χρησιμοποιώντας τη βιβλιοθήκη `ggplot2`

9.2.2.8 Τρεις Ποσοτικές Μεταβλητές

Υπάρχουν δύο τρόποι για να αναπαραστήσουμε γραφικά τις τιμές τριών ποσοτικών μεταβλητών. Ο ένας είναι να κατασκευάσουμε ένα ***τρισδιάστατο διά-***



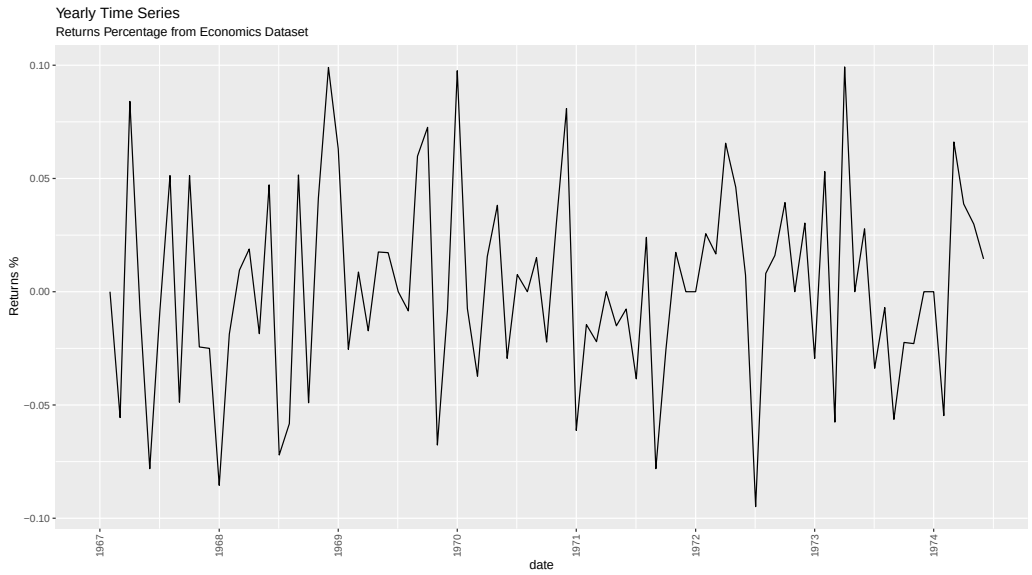
Διάγραμμα 9.43: Διάγραμμα φυσαλίδων σε διαφορετικές στάθμες μίας κατηγορικής μεταβλητής μέσω της βιβλιοθήκης `ggplot2`

9.2.2.11 Περισσότερες Διαστάσεις

Στην Παράγραφο 9.2.2.10 είδαμε πως μπορούμε με τη βοήθεια ενός διαγράμματος φυσαλίδων να αναπαραστήσουμε γραφικά τις παρατηρήσεις από τρεις ποσοτικές μεταβλητές και μια κατηγορική. Επίσης, στην Παράγραφο 9.2.2.9 χρησιμοποιήσαμε ένα διάγραμμα διασποράς για την αναπαράσταση παρατηρήσεων από δύο ποσοτικές μεταβλητές και δύο κατηγορικές.

Το ***κινούμενο διάγραμμα φυσαλίδων*** (*animated bubble plot*) μπορεί να χρησιμοποιηθεί για να δείξουμε πως οι παρατηρήσεις τριών ποσοτικών και μίας κατηγορικής μεταβλητής αλλάζουν ως προς μία πέμπτη διάσταση που συνήθως είναι ο χρόνος. Για την κατασκευή του χρησιμοποιούμε τη βιβλιοθήκη `gganimate`. Ας θεωρήσουμε τα δεδομένα `gapminder` της βιβλιοθήκης `gapminder`. Αρχικά, δημιουργούμε ένα διάγραμμα φυσαλίδων όπου στον xx' άξονα τοποθετούμε τις τιμές της μεταβλητής `gdpPercap` και στον yy' άξονα τοποθετούμε τις τιμές της μεταβλητής `lifeExp`. Χρησιμοποιούμε ***λογαριθ-***

```
  y="Returns %") +  # title and caption
  scale_x_date(labels = lbls,
  breaks = brks) +  # change to monthly ticks and labels
  theme(axis.text.x = element_text(angle = 90, vjust=0.5))
  # rotate x axis text
```



Διάγραμμα 9.49: Ετήσιο διάγραμμα χρονοσειράς, χρησιμοποιώντας τη βιβλιοθήκη ggplot2

9.3 Η Βιβλιοθήκη data.table

Στο Κεφάλαιο 2 χρησιμοποιήσαμε πλαίσια δεδομένων (data.frames) για να καταχωρήσουμε στην R τις παρατηρήσεις που έχουμε συλλέξει από ένα δείγμα διαφόρων ειδών μεταβλητών. Η βιβλιοθήκη data.table παρέχει μια εναλλακτική δομή δεδομένων (λίστα), με το όνομα data.table, την οποία μπορούμε να χρησιμοποιήσουμε για την καταχώρηση στην R ενός σετ δεδομένων. Ειδικά για δεδομένα μεγάλης κλίμακας, η χρήση της συγκεκριμένης δομής μειώνει αισθητά τη χρήση μνήμης του υπολογιστή, ενώ γίνεται χρήση ταχύτερων και

```
> head(ans)
   ori des    N
1: JFK LAX 3387
2: LGA PBI  245
3: EWR LAX   62
4: JFK MIA 1876
5: JFK SEA  298
6: EWR MIA  848
```

Ισοδύναμα μπορούμε, όπως και προηγουμένως, να γράψουμε `by = c("ori", "des")`. Επιπλέον, παρατηρήστε πως έχει διατηρηθεί η αρχική σειρά εμφάνισης των στάθμεων των μεταβλητών `ori` και `des` (βλ. `head(flights)`).

Σε όλα τα παραπάνω παραδείγματα συναθροίσεων που πραγματοποιήσαμε, υπολογίσαμε πλήθος πτήσεων. Φυσικά, μας δίνεται η δυνατότητα να χρησιμοποιήσουμε και άλλες συναρτήσεις. Για παράδειγμα, με τον κώδικα που ακολουθεί, υπολογίζουμε τη μέση καθυστέρηση απογείωσης και προσγείωσης μόνο για τον αερομεταφορέα "AA," για κάθε αεροδρόμιο προέλευσης και προορισμού καθώς και για κάθε μήνα:

```
> ans <- flights[ca == "AA",
      .(mean(ad), mean(dd)),
       by = .(ori, des, m)]
> head(ans)
   ori des m         V1         V2
1: JFK LAX 1  6.590361 14.2289157
2: LGA PBI 1 -7.758621  0.3103448
3: EWR LAX 1  1.366667  7.5000000
4: JFK MIA 1 15.720670 18.7430168
5: JFK SEA 1 14.357143 30.7500000
6: EWR MIA 1 11.011236 12.1235955
```

Από το παραπάνω αποτέλεσμα παρατηρούμε ότι αφού δε δώσαμε ονόματα στις δύο στήλες που κατασκευάζονται, δόθηκαν αυτόματα τα ονόματα `V1` και `V2`. Εύκολα αυτά μπορούν να αλλαχθούν με τη χρήση του ακόλουθου κώδικα:

```
> ans <- flights[ca == "AA",
      .(mean_arrival_delay = mean(ad),
        mean_departure_delay = mean(dd)),
       by = .(ori, des, m)]
```

```
   dep_delay arr_delay      N
1:      TRUE      TRUE  72836
2:     FALSE      TRUE  34583
3:     FALSE     FALSE 119304
4:      TRUE     FALSE  26593
```

9.3.1 Αναδιαμόρφωση, Στοίβαξη και Διαχωρισμός Δεδομένων

Κλείνουμε αυτήν την εισαγωγή στη βιβλιοθήκη `data.table` με μια γρήγορη αναφορά σε **αναδιαμορφώσεις** (*reshaping*), **στοιβάξεις** (*stacking*) και **διαχωρισμούς** (*splitting*) δεδομένων.

Στο Κεφάλαιο 8 είδαμε την αναγκαιότητα μετατροπής, σε κάποιες περιπτώσεις των δεδομένων από **ευρεία** σε **μακρά μορφή** και το αντίστροφο. Η βιβλιοθήκη `data.table` έχει συναρτήσεις που με εύκολο τρόπο μπορούν να **αναδιαμορφώσουν** τα δεδομένα.

Η συνάρτηση `melt()` χρησιμοποιείται για να μετατρέψει τα δεδομένα **από ευρεία σε μακρά μορφή**. Για παράδειγμα, τα παρακάτω δεδομένα

```
> DT = data.table(ID = letters[1:3], Age = 20:22, OB_A
    = 1:3, OB_B = 4:6, OB_C = 7:9)
> DT
   ID Age OB_A OB_B OB_C
1:  a  20    1    4    7
2:  b  21    2    5    8
3:  c  22    3    6    9
```

με τη χρήση της συνάρτησης `melt()` μετατρέπονται σε ευρεία μορφή:

```
> melt(DT, id.vars = c("ID","Age"))
   ID Age variable value
1:  a  20     OB_A     1
2:  b  21     OB_A     2
3:  c  22     OB_A     3
4:  a  20     OB_B     4
5:  b  21     OB_B     5
6:  c  22     OB_B     6
```

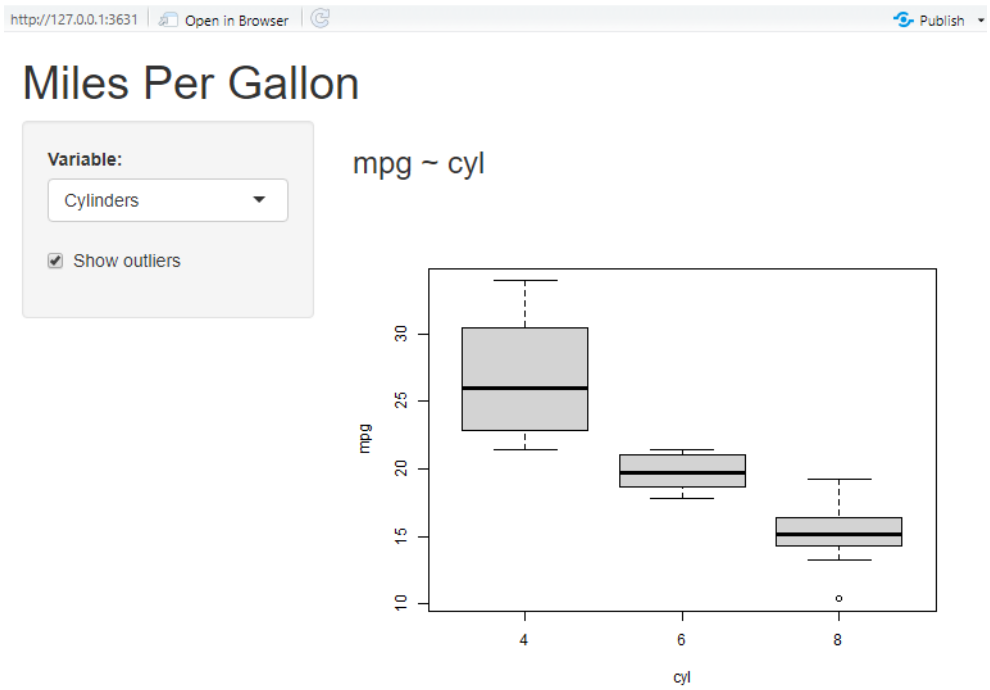
```
> split(DT, f = DT$src, keep.by=FALSE)
$D1
   src id v
1:  D1  1 u
2:  D1  2 p

$D2
   src id v
1:  D2  1 r
2:  D2  2 q

$D3
   src id v
1:  D3  1 x
2:  D3  2 t
```

Με το όρισμα `keep.by = FALSE` αφαιρούμε τη στήλη διαχωρισμού του πίνακα.

9.4 Η Βιβλιοθήκη shiny

Η `shiny` είναι μια βιβλιοθήκη του `RStudio` (βλ. Παράγραφο 2.13), με τη βοήθεια της οποίας μπορούμε να δημιουργήσουμε διαδραστικές διαδικτυακές εφαρμογές με την R. Στην παρούσα παράγραφο θα αναφερθούμε εν συντομία στη βιβλιοθήκη. Για περισσότερες πληροφορίες μπορείτε να επισκεφτείτε την ιστοσελίδα `https://shiny.rstudio.com`. Τη βιβλιοθήκη μπορούμε να τη χρησιμοποιήσουμε, αφού πρώτα την εγκαταστήσουμε και τη φορτώσουμε, αποκλειστικά μέσω του `Rstudio`. Αρχικά, θα αναφερθούμε στη βασική δομή συγγραφής μιας `shiny` εφαρμογής και εν συνεχεία θα παραθέσουμε ένα απλό παράδειγμα. Στην Παράγραφο 9.4.1 θα αναφέρουμε κάποια επιπλέον εργαλεία διαμόρφωσης R `shiny` εφαρμογών.

Κάθε `shiny` εφαρμογή είναι ένας φάκελος που περιέχει το αρχείο `server.R` και το αρχείο `ui.R`, καθώς και ορισμένα επιπρόσθετα προαιρετικά αρχεία, όπως δεδομένα, χρήσιμους συντάκτες, κ.τ.λ. Το όνομα του φακέλου είναι και το όνομα της εφαρμογής. Το αρχείο `ui.R` περιέχει τον κώδικα σχεδιασμού της ιστοσελίδας που προβάλλει την εφαρμογή. Συνήθως χωρίζεται σε δύο ***πλαί-***



Διάγραμμα 9.50: R `shiny` εφαρμογή

`Working Directory` και εν συνεχεία "`Choose Directory...`"). Στη συνέχεια πληκτρολογούμε

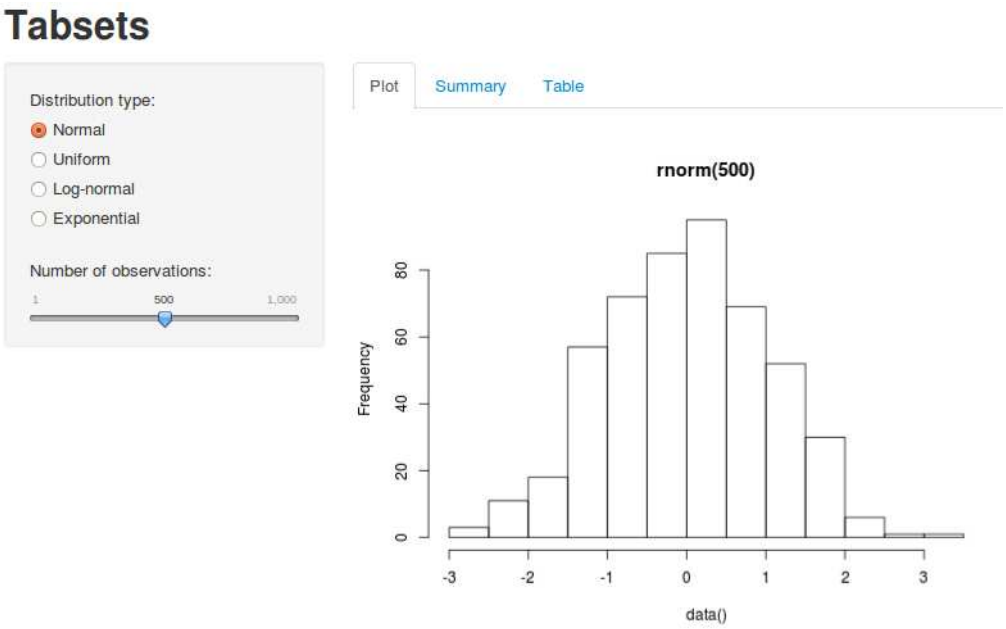
```
> library(shiny)
> runApp("Miles per Gallon")
```

όπου ως όρισμα στη συνάρτηση `runApp` δηλώνουμε, σε `" "`, το όνομα του φακέλου που έχουμε αποθηκεύσει τα αρχεία `server.R` και `ui.R`. Σε ένα νέο παράθυρο θα εμφανιστεί η εφαρμογή (βλ. Διάγραμμα 9.50).

Το παραγόμενο *HTML* αρχείο μπορείτε να το δημοσιεύσετε χρησιμοποιώντας την επιλογή "`Publish...`" στο άνω δεξί μέρος του αρχείου. Για την δημοσίευση του χρειάζεται να έχετε δημιουργήσει λογαριασμό στην ιστοσε-

```
  # Generate an HTML table view of the data
  output$table <- renderTable({
    data.frame(x=data())
  })
})
```

Το αποτέλεσμα αναπαρίσταται στο Διάγραμμα 9.52.



Διάγραμμα 9.52: Καρτέλες σε μια R shiny εφαρμογή

Με τη συνάρτηση helpText μπορείτε να προσθέσετε ***διευκρινιστικό κείμενο*** στην R shiny εφαρμογή. Με τη συνάρτηση submitButton δίνεται στον χρήστη η δυνατότητα να ***ενημερώσει***, όποτε επιθυμεί, τα αποτελέσματα με βάση τις επιλογές που έχει κάνει. Για παράδειγμα, θεωρήστε το παρακάτω αρχείο ui.R. Ο χρήστης επιλέγει ένα σετ δεδομένων από μια λίστα (συνάρτηση selectInput) και πληκτρολογεί, ***σε ένα κενό πλαίσιο για εισαγωγή αριθμητικών τιμών*** (συνάρτηση numericInput), τον αριθμό παρατηρήσεων που θέλει να προβάλλει. Με τη συνάρτηση helpText προστίθεται διευκρινιστικό

Κεφάλαιο 10

Εισαγωγή στη Δίτιμη Λογιστική Παλινδρόμηση

10.1 Εισαγωγή

Ας επανέλθουμε στο πρόβλημα της ***πρόβλεψης*** της μέσης τιμής μιας μεταβλητής (***μεταβλητή απόκρισης***), έστω Y, όταν γνωρίζουμε τις τιμές κάποιας ή κάποιων άλλων μεταβλητών (***επεξηγηματικές μεταβλητές***), έστω $X_1, \ldots, X_p$, συμβολιζόμενες εν συντομία με $\mathbf{X}$. Στο Κεφάλαιο 7, κάτω από την υπόθεση πως η μεταβλητή απόκρισης ήταν ποσοτική, αντιμετωπίσαμε το εν λόγω πρόβλημα χρησιμοποιώντας το πολλαπλό γραμμικό (ή πολλαπλασιαστικό) μοντέλο παλινδρόμησης.

Σε πολλές εφαρμογές η μεταβλητή απόκρισης παίρνει μόνο τις τιμές 0 και 1, οι οποίες αντιστοιχούν σε δύο συμπληρωματικά ενδεχόμενα, τα δύο πιθανά αποτελέσματα μιας διαδικασίας ή ενός τυχαίου πειράματος. Για παράδειγμα, αν ο ασθενής έχει καρκίνο ή όχι, αν ο άνεργος βρίσκει δουλειά ή όχι, αν θα πετύχει ένας φοιτητής στην εξέταση ενός μαθήματος ή όχι, αν θα εγκριθεί το αίτημα ενός μισθωτού από την τράπεζα για χορήγηση δανείου ή όχι, αν η ομάδα σου θα κερδίσει τον επόμενο αγώνα ή όχι, αν θα προβεί ένας οικογενειάρχης σε αγορά νέου αυτοκινήτου ή όχι, κ.λπ. Οι τιμές 0 και 1 της μεταβλητής απόκρισης αποτελούν τη συνηθέστερη κωδικοποίηση των δύο συμπληρωματι-

Επιπλέον, μπορούμε να χρησιμοποιήσουμε ως συνάρτηση σύνδεσης την *log-log*, με τύπο:

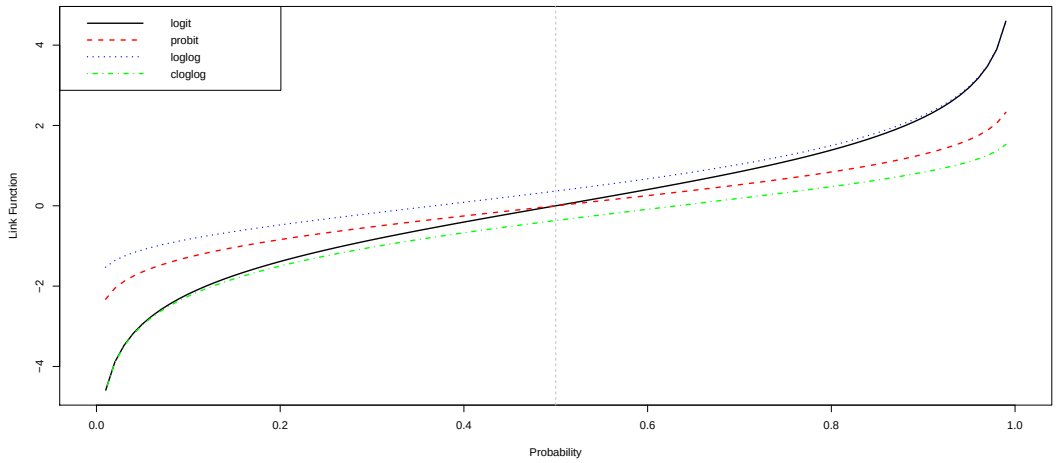
$$g^{-1}(\pi) = -\log\left(-\log(\pi)\right).$$

Το *log-log* μοντέλο χρησιμοποιείται κυρίως σε περιπτώσεις όπου η πιθανότητα να συμβεί η "επιτυχία" δεν είναι πολύ μικρή.

Τέλος, μπορούμε να χρησιμοποιήσουμε ως συνάρτηση σύνδεσης την *complementary log-log* με τον τύπο:

$$g^{-1}(\pi) = \log\left(-\log(1-\pi)\right).$$

Το *complementary log-log* μοντέλο χρησιμοποιείται κυρίως σε περιπτώσεις όπου η πιθανότητα να συμβεί η "επιτυχία" δεν είναι πολύ μεγάλη.



Διάγραμμα 10.2: Σύγκριση των συναρτήσεων σύνδεσης σε σχέση με τις τιμές που μπορεί να πάρει η πιθανότητα επιτυχίας

Στο Διάγραμμα 10.2 παρουσιάζουμε τις τέσσερις συνδετικές συναρτήσεις που αναφέραμε παραπάνω. Παρατηρούμε ότι όλες οι συναρτήσεις σύνδεσης είναι συνεχείς και αύξουσες. Οι συναρτήσεις *logit* και *probit* φαίνεται να έχουν αρκετές ομοιότητες, μία από τις οποίες είναι η συμμετρία που παρουσιάζουν

$$\log\left(Odds[Y=1|X=x, X_1=x_1, X_2=x_2, X_2=x_2]\right) = a + bx + b_1x_1 + b_2x_2 + b_3x_3.$$

Σε αυτήν την περίπτωση έχουμε τις παρακάτω ερμηνείες:

- Η τιμή $\exp(a)$ εκφράζει τη σχετική πιθανότητα εμφάνισης καρκίνου του πνεύμονα στους λατινοαμερικάνους άνδρες ηλικίας μηδέν ετών.
- Η τιμή $\exp(b)$ εκφράζει τη μεταβολή του λόγου σχετικών πιθανοτήτων (*odds ratio*) εμφάνισης καρκίνου του πνεύμονα, όταν η ηλικία αυξηθεί κατά ένα έτος και διατηρώντας κοινή τη φυλή και το φύλο.
- Η τιμή $\exp(b_1)$ εκφράζει την τιμή του λόγου σχετικών πιθανοτήτων (*odds ratio*) εμφάνισης καρκίνου του πνεύμονα των λευκών ως προς τους λατινοαμερικάνους υπό την προϋπόθεση πως έχουν κοινή ηλικία και φύλο.
- Η τιμή $\exp(b_2)$ εκφράζει την τιμή του λόγου σχετικών πιθανοτήτων (*odds ratio*) εμφάνισης καρκίνου του πνεύμονα των αφροαμερικάνων ως προς τους λατινοαμερικάνους υπό την προϋπόθεση πως έχουν κοινή ηλικία και φύλο.
- Η τιμή $\exp(b_3)$ εκφράζει την τιμή του λόγου σχετικών πιθανοτήτων (*odds ratio*) εμφάνισης καρκίνου του πνεύμονα των γυναικών ως προς τους άνδρες υπό την προϋπόθεση πως έχουν κοινή ηλικία και φυλή.

3. Ας υποθέσουμε πως θέλουμε στο μοντέλο που μελετάμε να συμπεριλάβουμε ως επεξηγηματικές μεταβλητές και την ηλικία και το φύλο, καθώς και την ***αλληλεπίδραση*** των δύο μεταβλητών (βλ. Παράγραφο 7.5). Ας ονομάσουμε X_1 την ηλικία και X_2 το φύλο (= 0: άνδρας, 1: γυναίκα). Έστω το κάτωθι μοντέλο λογιστικής παλινδρόμησης:

$$\log\left(Odds[Y=1|X_1=x_1, X_2=x_2]\right) = a + b_1x_1 + b_2x_2 + b_3x_1x_2.$$

Σε αυτήν την περίπτωση έχουμε τις παρακάτω ερμηνείες:

- Η τιμή $\exp(a)$ εκφράζει τη σχετική πιθανότητα εμφάνισης καρκίνου του πνεύμονα στους άνδρες ηλικίας μηδέν ετών.
- Η τιμή $\exp(a+b_2)$ εκφράζει τη σχετική πιθανότητα εμφάνισης καρκίνου του πνεύμονα στις γυναίκες ηλικίας μηδέν ετών.
- Η τιμή $\exp(b_1)$ εκφράζει τη μεταβολή του λόγου σχετικών πιθανοτήτων (*odds ratio*) εμφάνισης καρκίνου του πνεύμονα, όταν η ηλικία αυξηθεί κατά ένα έτος στους άνδρες.
- Η τιμή $\exp(b_1 + b_3)$ εκφράζει τη μεταβολή του λόγου σχετικών πιθανοτήτων (*odds ratio*) εμφάνισης καρκίνου του πνεύμονα, όταν η ηλικία αυξηθεί κατά ένα έτος στις γυναίκες.

■

10.2.2 Στατιστική Συμπερασματολογία

Ας θεωρήσουμε ότι διαθέτουμε p επεξηγηματικές μεταβλητές (οι οποίες μπορεί να είναι και ποσοτικές και κατηγορικές), τις οποίες τις συμβολίζουμε με τη βοήθεια του τυχαίου διανύσματος $\boldsymbol{X} = (X_1, ..., X_p)^T$, τις οποίες θέλουμε να χρησιμοποιήσουμε για να εκτιμήσουμε την πιθανότητα επιτυχίας μιας δίτιμης μεταβλητής απόκρισης Y. Έστω ότι διαθέτουμε πληροφορία από ένα τυχαίο δείγμα $(Y_1, X_{11}, \ldots, X_{1p})^T, \ldots, (Y_n, X_{n1}, \ldots, X_{np})^T$, μεγέθους n και συμβολίζουμε με $(y_1, x_{11}, \ldots, x_{1p})^T, \ldots, (y_n, x_{n1}, \ldots, x_{np})^T$ τα δεδομένα αυτού.

Επιπλέον, ας συμβολίσουμε με $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$ το διάνυσμα των επεξηγηματικών μεταβλητών για κάθε μονάδα i του δείγματος και ας θεωρήσουμε ότι έχει γνωστή τιμή $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ $(i = 1, \ldots, n)$. Η δεσμευμένη μέση τιμή της τυχαίας μεταβλητής Y_i δοθέντος του $\boldsymbol{X}_i = \boldsymbol{x}_i$ με βάση το λογιστικό μοντέλο παλινδρόμησης, είναι:

$$\mathbb{E}[Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i] = \mathbb{P}[Y_i = 1|\boldsymbol{X}_i = \boldsymbol{x}_i] \equiv \pi_i = \frac{\exp(a + b_1 x_{i1} + \cdots + b_p x_{ip})}{1 + \exp(a + b_1 x_{i1} + \cdots + b_p x_{ip})}.$$

Χρησιμοποιώντας την παραπάνω σχέση και τη σ.μ.π. της κατανομής *Bernoulli*,

κοντά στο μηδέν δηλώνουν πως η χρήση των p επεξηγηματικών μεταβλητών δεν βελτίωσε την προσαρμογή του μοντέλου, σε σχέση με αυτή που είχε χωρίς καμία επεξηγηματική μεταβλητή.

Το μειονέκτημα του παραπάνω δείκτη είναι πως δεν φτάνει ποτέ την τιμή ένα. Για τον λόγο αυτό ο *Nagelkerke* (1991) πρότεινε την ακόλουθη μικρή αλλαγή στον παραπάνω τύπο:

$$R_N^2 = \frac{R_{CS}^2}{1 - \exp\left(-\frac{D(null)}{n}\right)}.$$

Κριτήρια πληροφορίας: Στην Παράγραφο 7.4 είχαμε ορίσει τα κριτήρια ποινικοποιημένης πιθανοφάνειας AIC και BIC. Τα κριτήρια αυτά μπορούν και σ' αυτήν την περίπτωση να εφαρμοστούν. Χρησιμοποιώντας τους συμβολισμούς αυτού του κεφαλαίου, το κριτήριο AIC, για ένα μοντέλο λογιστικής παλινδρόμησης M, με d_M άγνωστες παραμέτρους, και για μέγεθος δείγματος n, δίνεται από τον τύπο

$$AIC(M) = -2LL + 2d_M,$$

ενώ το BIC από τον τύπο:

$$BIC(M) = -2LL + d_M \log(n).$$

Τα εν λόγω κριτήρια, μπορούν να χρησιμοποιηθούν για σύγκριση δύο οποιονδήποτε μοντέλων, όχι κατά ανάγκη εμφωλευμένων. Το μοντέλο με τη μικρότερη τιμή στο κριτήριο AIC (ή BIC) είναι αυτό που επιλέγεται. Τα κριτήρια λαμβάνουν υπόψη τους και την προσαρμογή των δύο μοντέλων, αλλά και την διάστασή (πολυπλοκότητά) τους. Άρα, αν ένα μοντέλο έχει λίγο καλύτερη προσαρμογή από ένα άλλο, αλλά πολύ μεγαλύτερη διάσταση (αριθμό επεξηγηματικών μεταβλητών), τα κριτήρια μπορεί να επιλέξουν το δεύτερο ως "πιο κατάλληλο", επειδή είναι πιο φειδωλό, χωρίς να αποκλίνει πολύ συγκριτικά με το πρώτο ως προς την προσαρμογή του. Όπως είχαμε δει και στην Παράγραφο 7.4, η διαφορά των δύο κριτηρίων είναι ότι το BIC ποινικοποιεί περισσότερο, σε σχέση με το AIC, μοντέλα μεγαλύτερης διάστασης (δηλαδή με περισσότερες επεξηγηματικές μεταβλητές), όταν το μέγεθος του δείγματος (n) είναι μεγαλύτερο του 7.39 ($n > e^2 \approx 7.39$).

k έναν αριθμό μεγαλύτερο του 25, που είναι ο προκαθορισμένος μέγιστος αριθμός επαναλήψεων που πραγματοποιεί η R. Αν το πρόβλημα παραμείνει τότε πιθανότατα αντιμετωπίζετε ένα από τα ζητήματα που αναφέραμε προηγουμένως (πχ. πολυσυγγραμικότητα ή πλήρης διαχωρισμός), το οποίο πρέπει πρώτα να λύσετε και εν συνεχεία να προσαρμόσετε εκ νέου το μοντέλο.

10.2.7 Ταξινόμηση στη Λογιστική Παλινδρόμηση

Το μοντέλο λογιστικής παλινδρόμησης μπορεί, όπως έχουμε αναφέρει, να χρησιμοποιηθεί και ως ένα εργαλείο ***δυαδικής ταξινόμησης*** ή ισοδύναμα για να προβλέψουμε, από τις παρατηρήσεις που διαθέτουμε, τις τιμές της μεταβλητής απόκρισης Y. Χρησιμοποιούμε τότε δισδιάστατους πίνακες ταξινόμησης (σύγχυσης) για να σχολιάσουμε την ***προβλεπτική ικανότητα*** του μοντέλου. Οι συγκεκριμένοι πίνακες, πρέπει να επισημάνουμε, πως χρησιμοποιούνται αποκλειστικά, όταν ο ερευνητικός σκοπός είναι να προβούμε σε προβλέψεις και όχι να ανακαλύψουμε αιτιώδεις σχέσεις. Ο λόγος είναι πως οι πίνακες αυτοί ποινικοποιούν μόνο τις εσφαλμένες προβλέψεις και όχι τη φτωχή προσαρμογή του μοντέλου όπως κάνουν, για παράδειγμα, οι μέθοδοι που είδαμε στην Παράγραφο 10.2.4.

Για να προβούμε στις προβλέψεις των τιμών της μεταβλητής απόκρισης θα πρέπει να χρησιμοποιήσουμε ένα ***σημείο διαχωρισμού*** (*cutoff point*) p^*. Αν η πιθανότητα πρόβλεψης από το μοντέλο λογιστικής παλινδρόμησης είναι μικρότερη του p^*, εκτιμούμε την τιμή της τ.μ. Y ως μηδέν, αλλιώς ως ένα. Τότε μπορούμε να δημιουργήσουμε τον ***πίνακα σύγχυσης*** (*confussion matrix*), ο οποίος είναι ένας 2×2 πίνακας, όπου οι γραμμές αντιστοιχούν στις πραγματικές παρατηρούμενες τιμές κάθε μίας κλάσης της Y και οι στήλες στις προβλέψεις με βάση το μοντέλο παλινδρόμησης. Αν καλέσουμε ως θετική κλάση την "επιτυχία (1)" και ως αρνητική κλάση την "αποτυχία (0)", στα κελιά του πίνακα αναγράφονται οι συχνότητες από τις ***ορθές θετικές*** (*true positive - TP*), τις ***ορθές αρνητικές*** (*true negative - TN*), τις ***ψευδείς θετικές*** (*false positive - FP*) και τις ***ψευδείς αρνητικές*** (*false negative - FN*)

λογιστικής παλινδρόμησης, τόσο καλύτερος ταξινομητής είναι. Στην πραγματικότητα, η τιμή AUC εκφράζει την πιθανότητα η προβλεπτική πιθανότητα επιτυχίας που παίρνουμε από το μοντέλο λογιστικής παλινδρόμησης όταν η παρατηρούμενη τιμή $y = 1$, να είναι μεγαλύτερη από την αντίστοιχη πιθανότητα όταν η παρατηρούμενη τιμή $y = 0$.

10.2.8 Μέθοδος Διασταυρωμένης Επικύρωσης

Για την υλοποίηση όλων των τεχνικών που αναφέραμε στην Παράγραφο 10.2.7, μπορούμε να χρησιμοποιήσουμε όλα τα δεδομένα που διαθέτουμε. Σε αυτήν την περίπτωση η μέθοδος που εφαρμόσαμε πραγματοποίησε ***εντός του δείγματος προβλέψεις*** (*in-sample predictions*). Το μειονέκτημα αυτής της μεθόδου είναι ότι χρησιμοποιούμε τα δεδομένα δύο φορές: μία φορά για την προσαρμογή του μοντέλου και μία φορά για τον υπολογισμό προβλέψεων. Τότε είναι πιθανό, αν στα δεδομένα μας υπάρχουν ακραίες τιμές, το εκτιμώμενο μοντέλο να προσαρμόζεται και ως προς αυτές, δίνοντάς μας ίσως μια εσφαλμένη εικόνα. Το φαινόμενο αυτό ονομάζεται ***υπερπροσαρμογή*** (*overfitting*), καθώς τα κριτήρια καλής προσαρμογής που χρησιμοποιούμε για να πάρουμε πληροφορία για την "ποιότητα" του μοντέλου στηρίζονται σε όλες τις παρατηρήσεις και επομένως έχουν "βελτιστοποιηθεί" ως προς όλα τα δεδομένα. Πιθανότατα όμως σε ένα νέο σετ δεδομένων, στο οποίο θέλουμε να χρησιμοποιήσουμε το μοντέλο για να προβούμε σε προβλέψεις, η συμπεριφορά του μοντέλου να είναι πολύ χειρότερη[†].

Για την αποφυγή των παραπάνω προβλημάτων συνήθως εφαρμόζουμε τη μέθοδο της ***διασταυρωμένης επικύρωσης*** (*cross-validation*). Η μέθοδος είναι πολύ απλή και μπορεί να εφαρμοστεί σε οποιοδήποτε μοντέλο, όχι μόνο στη λογιστική παλινδρόμηση. Η γενική της ιδέα είναι να διαμερίσουμε τα δεδομένα σε δύο υποσύνολα (ξένα μεταξύ τους) και να χρησιμοποιήσουμε το ένα υποσύνολο για την προσαρμογή του μοντέλου και το άλλο για τον υπολογισμό των

[†]Πέραν του φαινομένου της υπερπροσαρμογής, υπάρχει και το φαινόμενο της ***υποπροσαρμογής*** (*underfitting*), κατά το οποίο το μοντέλο που χρησιμοποιούμε δεν έχει καλή προσαρμογή, όντας πιθανότατα μη κατάλληλο για να περιγράψει τη σχέση των δεδομένων, με αποτέλεσμα επίσης να καταλήγουμε σε προβλέψεις με μεγάλα σφάλματα.

νέκτημα αυτής της τεχνικής, συγκριτικά με την προηγούμενη που αναφέραμε, είναι ότι κάθε παρατήρηση εγγυημένα χρησιμοποιείται τόσο για εκπαίδευση, όσο και για εξέταση (μάλιστα για εξέταση χρησιμοποιείται ακριβώς μια φορά). Το μειονέκτημα από την άλλη πλευρά αυτής της τεχνικής είναι ότι η αναλογία δεδομένων εκπαίδευσης–εξέτασης εξαρτάται από τον αριθμό των υποσυνόλων k. Αν το n δεν είναι τόσο μεγάλο και επιλέξουμε $k = 10$, μπορεί το υποσύνολο εξέτασης να περιέχει πολύ μικρό πλήθος δεδομένων.

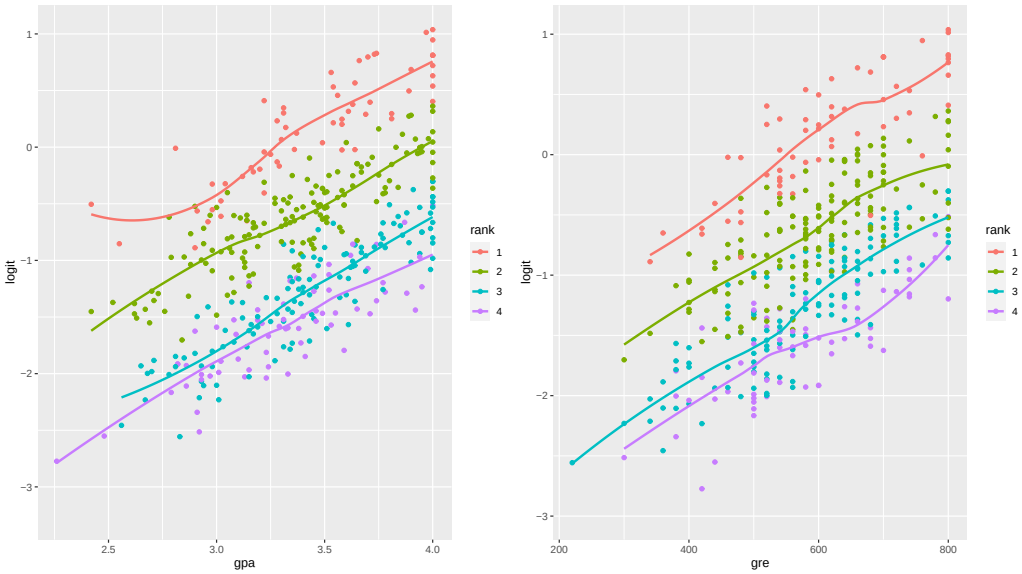
Παράδειγμα 10.2.1.

Έστω ότι ενδιαφερόμαστε να δούμε πώς η επίδοση φοιτητών στο ειδικό τεστ αξιολόγησης των γνώσεων τους πάνω σε ένα συγκεκριμένο γνωστικό αντικείμενο (`gre` - *graduate record exam score*), η μέση βαθμολογία τους στις προπτυχιακές σπουδές (`gpa` - *grade point average*) και το κύρος του ιδρύματος που έκαναν τις προπτυχιακές τους σπουδές (`rank`) επηρεάζουν την εισαγωγή τους (`admit`) σε μεταπτυχιακά προγράμματα ενός Πανεπιστημίου.

Έστω ότι για τις παραπάνω μεταβλητές διαθέτουμε δεδομένα για 400 φοιτητές. Τα δεδομένα βρίσκονται στον ακόλουθο σύνδεσμο `http://www.math.ntua.gr/~fouskakis/Rbook/binary.csv`. Οι μεταβλητές `gre` και `gpa` είναι ποσοτικές, ενώ η μεταβλητή `rank` είναι διάταξης (`1`: πολύ υψηλό κύρος, `2`: υψηλό κύρος, `3`: μέτριο κύρος, 4: χαμηλό κύρος). Τέλος, η μεταβλητή ενδιαφέροντος (`admit`) είναι δίτιμη (`1`: εισήχθησαν στο μεταπτυχιακό πρόγραμμα, `2`: δεν εισήχθησαν στο μεταπτυχιακό πρόγραμμα).

Δεδομένα: Αρχικά καταχωρούμε τα ***δεδομένα*** σε ένα πλαίσιο δεδομένων το οποίο το ονομάζουμε `mydata`. Την ανάκτηση των δεδομένων την κάνουμε απευθείας από τον σύνδεσμο που δίνεται, χρησιμοποιώντας την εντολή `read.csv()`, ενώ παρατηρούμε στο αρχείο πως τα δεδομένα στην πρώτη τους γραμμή έχουν τα ονόματα των μεταβλητών:

```
> mydata <- read.csv("http://www.math.ntua.gr/~fouskakis/
   Rbook/binary.csv", header=T)
> head(mydata)
  admit gre  gpa rank
1     0 380 3.61    3
2     1 660 3.67    3
```



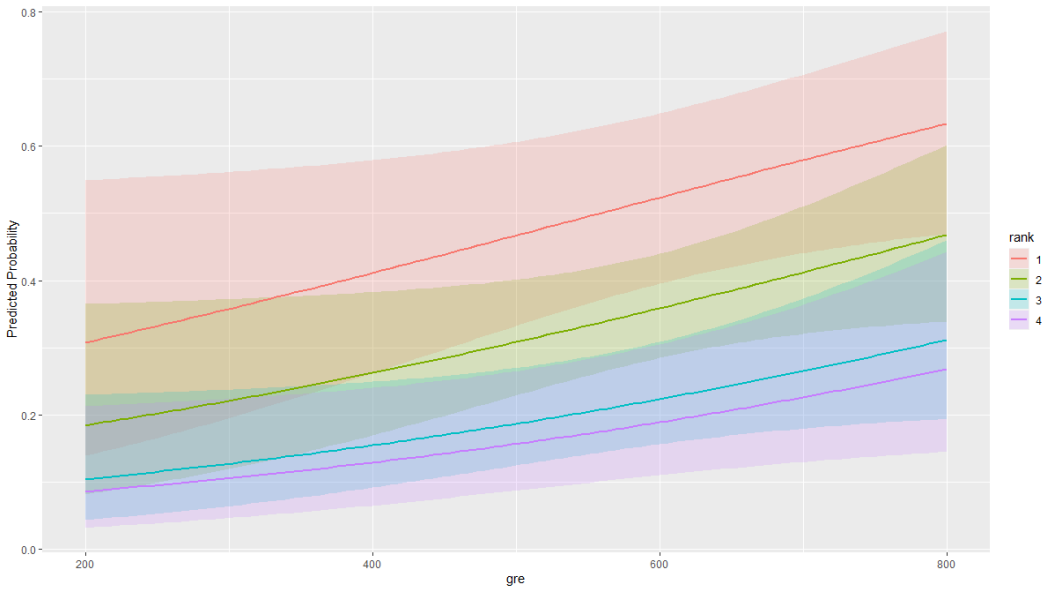
Διάγραμμα 10.4: Γραφικός έλεγχος γραμμικότητας στο μοντέλο λογιστικής παλινδρόμησης

```
> interaction_gre<-log(mydata$gre)*mydata$gre
> dttest<-data.frame(mydata, interaction_gpa,
                     interaction_gre)
> summary(glm(admit ~ gre + gpa + rank + interaction_gpa
          + interaction_gre, data = dttest,
          family = "binomial"))

Call:
glm(formula = admit ~ gre + gpa + rank + interaction_gpa
  + interaction_gre, family = "binomial", data = dttest)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6154  -0.8695  -0.6394   1.1288   2.1172

Coefficients:
                 Estimate Std. Error z value Pr(>z)
(Intercept)      0.080409  15.926413   0.005 0.995972
```

Στη συνέχεια με τη βοήθεια της βιβλιοθήκης ggplot2, αναπαριστούμε γραφικά την παραπάνω πληροφορία:

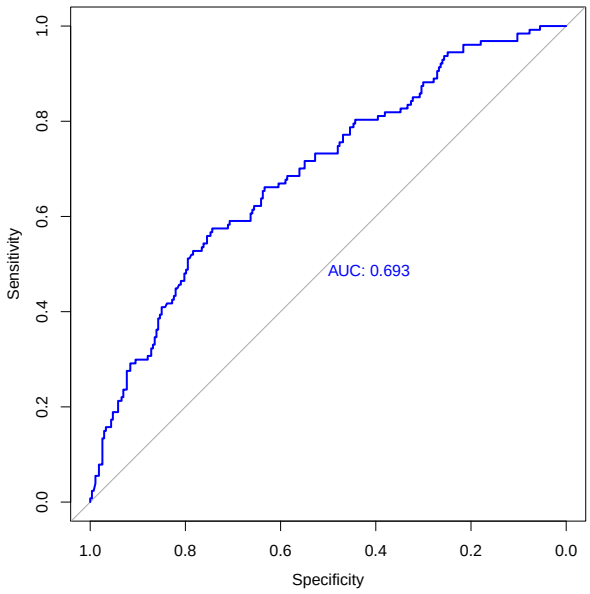
```
> library(ggplot2)
> ggplot(newdata3, aes(x = gre, y = PredictedProb)) +
  geom_ribbon(aes(ymin = LL,
  ymax = UL, fill = rank), alpha = 0.2) +
  geom_line(aes(colour = rank), size = 1) +
  ylab("Predicted Probability")
```

Από το Διάγραμμα 10.6 παρατηρούμε πως όσο υψηλότερη βαθμολογία έχει ένας φοιτητής (με μέσο βαθμό πτυχίου ίσο με τον μέσο όρο) στο τεστ gre, τόσο υψηλότερη είναι η πιθανότητα να γίνει αποδεκτός στο μεταπτυχιακό πρόγραμμα. Η εν λόγω επίδραση είναι υψηλότερη όταν το ίδρυμα του φοιτητή που κάνει την αίτηση έχει πολύ ψηλό κύρος και μειώνεται όσο μειώνεται το κύρος του προπτυχιακού ιδρύματος του φοιτητή. Τέλος, δε φαίνεται να υπάρχει ένδειξη για αλληλεπίδραση των δύο επεξηγηματικών μεταβλητών.



Διάγραμμα 10.6: Προβλέψεις στη λογιστική παλινδρόμηση

Ταξινόμηση: Ας εστιάσουμε τώρα στην προβλεπτική ικανότητα του μοντέ-



Διάγραμμα 10.7: Η καμπύλη ROC για τη λογιστική παλινδρόμηση του Παραδείγματος 10.2.1, σε εντός του δείγματος προβλέψεις

```
> accuracy <- sum(diag(confusion_matrix))/sum(
   confusion_matrix)
> accuracy
[1] 0.69
> sensitivity<-confusion_matrix[2,2]/(confusion_matrix[2,2]+
   confusion_matrix[1,2])
> sensitivity
[1] 0.5748031
> specificity<-confusion_matrix[1,1]/(confusion_matrix[1,1]+
   confusion_matrix[2,1])
> specificity
[1] 0.7435897
> youden<-sensitivity+specificity-1
> youden
[1] 0.3183929
```

Η ακρίβεια του ταξινομητή είναι 0.69, οπότε και πάλι καταλήγουμε στο

Ασκήσεις

10.1. Θεωρήστε τα δεδομένα του ακόλουθου συνδέσμου `http://www.math.ntua.gr/~fouskakis/Rbook/Salaries.csv` που παρουσιάζουν το μισθό μελών του διδακτικού και ερευνητικού προσωπικού ενός Πανεπιστημίου στις Η.Π.Α. σε 9 μήνες, κατά το έτος 2008-09. Τα στοιχεία που έχουν συλλεχθεί αφορούν τις παρακάτω μεταβλητές: `rank` = βαθμίδα (`AssocProf`: Αναπληρωτής Καθηγητής, `AsstProf`: Επίκουρος Καθηγητής, `Prof`: Καθηγητής), `discipline` = Είδος Σχολής στην οποία ανήκει το μέλος (`theoretical`: θεωρητική, `applied`: εφαρμοσμένη), `yrs.since.phd` = έτη από την απόκτηση του διδακτορικού διπλώματος του μέλους, `yrs.service` = έτη απασχόλησης του μέλους, `sex` = φύλο (`female`: γυναίκα, `male`: άνδρας), `salary` = μισθός στους 9 μήνες σε δολάρια. Χρησιμοποιώντας την R:

i. Δημιουργήστε μια νέα κατηγορική μεταβλητή με το όνομα `salary_cat` με τιμή `0` (χαμηλός μισθός), όταν η μεταβλητή `salary < 100000` δολάρια ή τιμή `1` διαφορετικά.

ii. Προσαρμόστε το μοντέλο λογιστικής παλινδρόμησης με μεταβλητή απόκρισης τη `salary_cat` και επεξηγηματικές τις `rank`, `discipline`, `yrs.since.phd`, `yrs.service` και `sex`.

iii. Ελέγξτε στο παραπάνω μοντέλο αν υπάρχει πρόβλημα πολυσυγγραμικότητας. Σε περίπτωση που υπάρχει αντιμετωπίστε το αφαιρώντας την "προβληματική" μεταβλητή.

iv. Ελέγξτε αν στο μοντέλο που καταλήξατε υπάρχει πρόβλημα γραμμικότητας.

10.2. Θεωρήστε τα δεδομένα της Άσκησης 10.1. Χρησιμοποιώντας την R:

i. Προσαρμόστε το μοντέλο λογιστικής παλινδρόμησης με μεταβλητή απόκρισης τη `sex` και επεξηγηματικές μεταβλητές τις `rank`, `discipline`, `yrs.service` και `salary`.

ii. Ερμηνεύστε τους συντελεστές της λογιστικής παλινδρόμησης χρησιμοποιώντας λόγους σχετικών πιθανοτήτων και σχολιάστε την επί-

Agresti, A. and Franklin, C. (2007). *Statistics. The Art and Science of Learning from Data*. Prentice Hall. New Jersey.

Aitchison, J. and Dunsmore, I.R. (1975). *Statistical Prediction Analysis*. Cambridge University Press. Cambridge.

Albert, J. (2007). *Bayesian Computation with R*. Springer. New York.

Albert, J. and Rizzo, M. (2011). *R by Example*. Springer. New York.

Bain, L.J. and Engelhardt, M. (1992). *Introduction to Probability and Mathematical Statistics*. 2nd Edition. Duxbury Press. Pacific Grove.

Benjamin, J.R. and Cornell, C.A. (1970). *Probability, Statistics, and Decision for Civil Engineers*. McGraw - Hill. New York.

Bertsimas, D. and Freund, R.M. (2000). *Data, Models and Decisions: The Fundamentals of Management Science*. South-Western College Publishing. Ohio.

Bower, A.H. and Lieberman, G.J. (1972). *Engineering Statistics*. Prentice Hall. New York.

Braun, W.J. and Murdoch, D.J. (2007). *A First Course in Statistical Programming with R*. Cambridge University Press. Cambdridge.

Casella, G. and Berger, O. (2002). *Statistical Inference*. 2nd Edition. Duxbury Press. Pacific Grove.

Chambers, J.M. (2008). *Software for Data Analysis: Programming with R*. Springer. New York.

Chambers, J.M. and Hastie, T.J. (1991). *Statistical Models in S*. Chapman and Hall. London.

Chatfield, C. (1995). *Problem Solving. A Statistician's Guide*. Chapman and Hall. London.

Ευρετήριο