ΜΙΑ ΜΙΚΡΗ ΕΙΣΑΓΩΓΗ ΣΤΟ R

Βασίλειος Γιαγός
sas00012@sas.aegean.gr

Τμήμα Στατιστικής και Αναλογιστικής Επιστήμης
Πανεπιστήμιο Αιγαίου
Φεβρουάριος 2003, Καρλόβασι Σάμος

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

1.1 Τι είναι το R

Το R είναι μια γλώσσα προγραμματισμού συνοδευόμενη από ένα περιβάλλον για επεξεργασία δεδομένων, υπολογισμών και γραφημάτων. Αν και χρησιμοποιείται κυρίως στην στατιστική οι δημιουργοί του προτιμούν να το αποκαλούν εργαλείο για ανάλυση δεδομένων τονίζοντας ότι περιλαμβάνει και μοντέρνες και παλιές στατιστικές μεθοδολογίες. Τι είναι το R

Είναι ένα GNU [1] πρόγραμμα παρόμοιο με τη γλώσσα και το περιβάλλον S που αναπτύχθηκε στα εργαστήρια της Bell (στο παρελθόν AT&T , τώρα Lucent Technologies) από τον John Chambers και τους συνεργάτες του. Το R μπορεί να θεωρηθεί ως μια διαφορετική υλοποίηση του S . Υπάρχουν μερικές σημαντικές διαφορές, αλλά ο περισσότερος κώδικας που γράφεται για το S τρέχει αμετάβλητος και στο R (όμως το R και το S δεν είναι 100% συμβατά).Μοιραία το R συγκρίνεται με το S+ ή S-plus το εμπορικό πακέτο βασισμένο και αυτό στην γλώσσα S . Η κύρια και η πιο αξιοσημείωτη διαφορά είναι ότι το R **δεν έχει γραφικό περιβάλλον εργασίας** και απαιτεί αρκετή ενασχόληση για να γίνει οικείο σε κάποιον. R και S+

1.2 Αποκτώντας το R

Το R μπορεί κάποιος να το αποκτήσει δωρεάν στην ιστοσελίδα του: http://www.r-project.org ή σε ένα από τα πολλά mirrors του CRAN (Comprehensive R Archive) http://cran.r-project.org το οποίο είναι ένα δίκτυο διανομής του R σε πολλά μέρη του κόσμου μέσω του Internet . Υποστηρίζει πολλές πλατφόρμες και λειτουργικά όπως Linux, Windows και πληθώρα Unix λειτουργικών.

1.3 Βασικές Έννοιες

Ο σκοπός ενός στατιστικού προγράμματος είναι να επεξεργάζεται και να χειρίζεται δεδομένα. Το R χειρίζεται τα δεδομένα (π.χ. πίνακες, διανύσματα...) και τις δομές (π.χ. συναρτήσεις) σαν *αντικείμενα* (objects). Τα αντικείμενα ανάλογα με την δομή τους κατηγοριοποιούνται αυτόματα σε:

[1] "GNU's Not Unix!" Το project GNU άρχισε το 1984 με σκοπό τη δημιουργία ενός ολοκληρωμένου λειτουργικού συστήματος τύπου Unix το οποίο θα είναι δωρεάν και ανοιχτό (ως προς τον πηγαίο κώδικα) λογισμικό. Το R-project αναπτύσσεται κάτω από την αιγίδα του GNU .

Διάνυσμα *vectors* Ένα τυπικό διάνυσμα ή μονοδιάστατος πίνακας. Παίρνει αριθμιτικές τιμές.

Πίνακας *matrix* Πολυδιάστατοι πίνακες.

Παράγοντες *factors* Χρησιμοποιούνται για τον χειρισμό κατηγορικών μεταβλητών.

Λίστες *lists* Μια γενική μορφή διανύσματος που τα στοιχεία του δεν είναι απαραίτητα του ίδιου τύπου και μπορούν να εμπεριέχουν άλλα διανύσματα ή λίστες.

Πίνακες δεδομένων *data frames* Είναι δομημένοι όπως οι απλοί πίνακες αλλά σε κάθε γραμμή αντιστοιχεί μία παρατηρούμενη μονάδα. Οι γραμμές μπορούν να είναι διαφορετικών τύπων καθώς περιέχουν και κατηγορικές και ποσοτικές μεταβλητές.

Συναρτήσεις *fuctions* Θεωρούνται και αυτές αντικείμενα.

Τα αντικείμενα δημιουργούνται και αποθηκεύονται με βάση το όνομά τους. Επίσης το R λαμβάνει τον κάθε χαρακτήρα μοναδικά (*case sensitive*). Δηλαδή **δεν** είναι το ίδιο γράψουμε `A` και `a` όπως επίσης `OBJECT` `OBJECt` `object`.

Κεφάλαιο 2

Το περιβάλλον του R

2.1 Βασικές εντολές

Αρχίζοντας το R βλέπουμε μια γραμμή εκχώρησης εντολών:

```
>
```

Για να λάβουμε την αρχική βοήθεια:

```
> help.start()
updating HTML package listing
If nothing happens, you should open...
```

Αυτό έχει σαν αποτέλεσμα να δημιουργήσει ένα περιβάλλον βοήθειας που θα εκκινήσει τον τυπικό browser του συστήματος.

Σε οποιαδήποτε στιγμή ζητάμε βοήθεια με την εντολή:

```
> help("foo")
```

Όπου foo είναι αυτό που ζητάμε, ή ισοδύναμα:

```
> ?foo
```

Για να εκχωρήσουμε μία τιμή σε ένα αντικείμενο π.χ. στο zz την τιμή 6:

```
> zz <- 6
```

Είναι σημαντικό να διευκρινίσουμε ότι η εντολή εκχώρησης διαγράφει την προηγούμενη τιμή του αντικειμένου. Για παράδειγμα:

```
> zz <- 6
> zz <- 2
```

Τώρα πλέον το zz έχει την τιμή 2 και όχι 6.

Αν θέλουμε να εκχωρήσουμε ένα διάνυσμα:

```
> zz <- c(10.4, 5.6, 3.1, 6.4, 21.7)
```

Όπου εκχωρούμε τις τιμές του πίνακα χρησιμοποιώντας μία συνάρτηση c() η οποία παίρνει ως όρισμα τα στοιχεία του διανύσματος, κατασκευάζει ένα διάνυσμα και με την εντολή $<-$ το εκχωρεί στο αντικείμενο zz . Μια πιο πολύπλοκη διαδικασία:

```
> zz <- c(10.4, 5.6, 3.1, 6.4, 21.7)
> y <- c(zz,0,zz)
```

Τώρα δημιουργήσαμε ένα νέο διάνυσμα 11 στοιχείων [1] αντιγράφοντας το zz προσθέτοντας έκτο το 0 και αντιγράφοντας ξανά το zz . Ένας άλλος τρόπος για να τα καταχωρήσουμε είναι με την εντολή scan :

```
> zz <- scan()
1: 10.4 5.6 3.1 6.4 21.7
6:
Read 5 items
```

Η οποία απλά παίρνει τις τιμές από το πληκτρολόγιο ή από ένα αρχείο αν το ορίσουμε, θα το δούμε στην αποθήκευση δεδομένων, τερματίζεται με μια κενή γραμμή. Για να δούμε τις τιμές του y δίνουμε την εντολή:

```
> print(y)
 [1] 10.4  5.6  3.1  6.4 21.7  0.0 10.4  5.6  3.1  6.4 21.7
```

Η print εμφανίζει στην οθόνη τα περιεχόμενα του αντικειμένου π.χ.y . Το σύμβολο [1] μας υπενθυμίζει ότι έπεται το πρώτο στοιχείο του διανύσματος.

Μία άλλη πολύ χρήσιμη εντολή είναι η summary . Δημιούργεί μια "περίληψη" του αντικειμένου που παίρνει ως όρισμα και το είδος της περίληψης εξαρτάται από την κλάση του αντικειμένου (αν είναι π.χ. διάνυσμα, πίνακας, αλφαριθμητικό, κ.τ.λ.).

```
>summary(y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   4.350   6.400   8.582  10.400  21.700
```

Συγκεκριμένα για την μεταβλητή y και για κάθε ποσοτική, τα σύμβολα των αποτελεσμάτων σημαίνουν: [2]

Min.	ελάχιστο
1st Qu.	πρώτο τεταρτημόριο
Median	διάμεσος
3rd Qu.	τρίτο τεταρτημόριο
Max.	μέγιστο
NA's	ελλειπούσες τιμές

Υπολογίζουμε τις συσχετίσεις ανάμεσα σε δεδομένα με την ίδια λογική της summary χρησιμοποιώντας την εντολή cor από το correlation (συσχέτιση):

[1] 5 του zz , 0, ξανά 5 του zz

[2] Στο συγκεκριμένο παράδειγμα δεν υπάρχουν ελλειπούσες τιμές και δεν απεικονίζονται.

```
> cor(zz, use = "pairwise.complete.obs")
                MPG     ENGINE      HORSE     WEIGHT      ACCEL       YEAR
MPG       1.0000000 -0.7886070 -0.7712932 -0.8070043  0.4343043  0.4658041
ENGINE   -0.7886070  1.0000000  0.8970699  0.9333334 -0.5446237 -0.2022022
HORSE    -0.7712932  0.8970699  1.0000000  0.8594179 -0.7014089 -0.2827388
WEIGHT   -0.8070043  0.9333334  0.8594179  1.0000000 -0.4152115 -0.1233154
ACCEL     0.4343043 -0.5446237 -0.7014089 -0.4152115  1.0000000  0.3024866
YEAR      0.4658041 -0.2022022 -0.2827388 -0.1233154  0.3024866  1.0000000
ORIGIN           NA          NA          NA          NA          NA          NA
CYLINDER         NA          NA          NA          NA          NA          NA
FILTER..         NA          NA          NA          NA          NA          NA
         ORIGIN CYLINDER FILTER..
MPG          NA       NA       NA
ENGINE       NA       NA       NA
HORSE        NA       NA       NA
WEIGHT       NA       NA       NA
ACCEL        NA       NA       NA
YEAR         NA       NA       NA
ORIGIN       NA       NA       NA
CYLINDER     NA       NA       NA
FILTER..     NA       NA       NA
Warning message:
NAs introduced by coercion
```

Σχόλια: Όσες τιμές έχουν NA είναι διότι έχουμε συμπεριλάβει και τις ποιοτικές μεταβλητές των δεδομένων. Παρατηρούμε επίσης ότι στην διαγώνιο των αποτελεσμάτων έχουμε τον αριθμό 1 μια και συγκρίνονται με τον εαυτό τους. Οι μεταβλητές που χρησιμοποιούμε έχουν αγνοούμενες τιμές και ο υπολογισμός των συσχετίσεων θα τερματιζόταν με την πρώτη άγνωστη τιμή (missing value - NA) • ξεπερνάμε αυτό το πρόβλημα δίνοντας το όρισμα `use = "pairwise.complete.obs"` για να υπολογισθεί η συσχέτιση χρησιμοποιώντας μόνο τα ζευγάρια τιμών, από κάθε μεταβλητή, που δεν έχουν άγνωστες τιμές και αγνοεί ολόκληρο το ζευγάρι τιμών αν βρεθεί έστω και μία από το καθένα.

Μπορούμε να εκτελούμε και αριθμητικές διαδικασίες με την βοήθεια τελεστών όπως της γλώσσας προγραμματισμού c :

```
> x <- 2
> x <- x*x
> print(x)
[1] 4
```

Οι τελεστές είναι: $+ - */$ για πρόσθεση, αφαίρεση, πολλαπλασιασμό και διαίρεση αντίστοιχα.

2.2 Αποθήκευση και επανάκτηση αντικειμένων

Το R παρέχει την δυνατότητα αποθήκευσης των αντικειμένων τα οποία αποθηκεύονται είτε σε ένα προκαθορισμένο πρότυπο που χρησιμοποιεί το R είτε σε ένα απλό αρχείο κειμένου. Η πλέον εύκολη αποθήκευση αντικειμένων γίνεται με την εντολή:

```
> save(zz, file="zz.Rdata")
```

Όπου το zz είναι το όνομα του αντικειμένου και στο "zz.Rdata" το όνομα του αρχείου που θα αποθηκευθεί. Για να μπορούμε να χρησιμοποιήσουμε το αποθηκευμένο αντικείμενο και σε άλλα στατιστικά πακέτα προσθέτουμε άλλη μια παράμετρο:

```
> save(zz, file="zz.Rdata",ascii = TRUE)
```

Που σημαίνει ότι θα το αποθηκεύσει ως ένα απλό αρχείο κειμένου.

Για να "φορτώσουμε", επανακτήσουμε, το προηγούμενο αντικείμενο δίνουμε:

```
> load("zz.Rdata")
```

Για επιπλέον πληροφορίες μπορεί κάποιος να ανατρέξει στο εγχειρίδιο R Data Import/Export που το συμπεριλαμβάνει κάθε διανομή του R .

2.3 Πακέτα

Το R αποτελείται από πακέτα που περιέχουν ρουτίνες για στατιστικούς ελέγχους, δεδομένα (data sets) , βιβλιοθήκες για εισαγωγή - εξαγωγή δεδομένων, εξειδικευμένες στατιστικές μεθόδους και γενικά επεκτείνουν τις δυνατότητες του R . Στην βασική του διανομή περιέχει 8 βασικά πακέτα. Μπορούμε να προσθέσουμε και άλλα χρησιμοποιώντας το CRAN . Για να δημιουργηθεί μια λίστα με μια μικρή περιγραφή των πακέτων που είναι εγκατεστημένα δίνουμε:

Παράδειγμα

```
>library()
Packages in library '/rw1061/library':

base                    The R base package
boot                    Bootstrap R (S-Plus) Functions (Canty)
class                   Functions for classification
cluster                 Functions for clustering (by Rousseeuw et al.)
ctest                   Classical Tests
.
.
.
ts                      Time series functions
```

Για να ενεργοποιήσουμε ένα πακέτο δίνουμε:

```
>library(foo)
```

Όπου foo είναι το όνομα του πακέτου που θέλουμε πχ `>library(ctest)` για το ctest πακέτο. Για να εγκαταστήσουμε ένα πακέτο αρκεί από το γραφικό περιβάλλον να δώσουμε: Packages - Install packages from και αναλόγως αν έχουμε το πακέτο στον τοπικό μας δίσκο ή θα το κατεβάσουμε μέσω διαδικτύου δίνουμε local zip file ή CRAN αντίστοιχα. Είναι δινατόν να το εγκαταστήσουμε και από την γραμμή εντολών του λειτουργικού μας δίνοντας την εντολή:

```
R CMD INSTALL name
```

Όπου name το όνομα του πακέτου.

2.3.1 Το πακέτο foreign

Στο R υπάρχει η δυνατότητα να εισάγει κανείς δεδομένα από άλλα στατιστικά πακέτα χρησιμοποιώντας το πακέτο foreign . Τα στατιστικά πακέτα που υποστηρίζονται είναι:

Spss	με την εντολή read.spss
S3	με τις εντολές data.restore, read.S, SModeNames
Stata	με την εντολή read.dta
SAS	με τις εντολές lookup.xport,read.xport,read.ssd
Epi Info	με την εντολή read.epiinfo

Παράδειγμα Εισάγουμε δεδομένα από το Spss . Συγκεκριμένα από το αρχείο Cars.sav το οποίο περιέχει στοιχεία για αυτοκίνητα. Έχουμε:

```
> zz <- read.spss("/Program files/spss/Cars.sav", to.data.frame = TRUE)
> summary(zz)
      MPG            ENGINE          HORSE           WEIGHT         ACCEL           YEAR
 Min.   : 9.00   Min.   :  4.0   Min.   : 46.00   Min.   : 732   Min.   : 8.00   Min.   : 0.00
 1st Qu.:17.50   1st Qu.:104.3   1st Qu.: 75.75   1st Qu.:2224   1st Qu.:13.63   1st Qu.:73.00
 Median :23.00   Median :148.5   Median : 95.00   Median :2811   Median :15.50   Median :76.00
 Mean   :23.51   Mean   :194.0   Mean   :104.83   Mean   :2970   Mean   :15.50   Mean   :75.75
 3rd Qu.:29.00   3rd Qu.:293.3   3rd Qu.:129.25   3rd Qu.:3612   3rd Qu.:17.07   3rd Qu.:79.00
 Max.   :46.60   Max.   :455.0   Max.   :230.00   Max.   :5140   Max.   :24.80   Max.   :82.00
 NA's   : 8.00                   NA's   :  6.00

   ORIGIN          CYLINDER          FILTER..
Japanese: 79   8 Cylinders:107   Selected    :291
European: 73   6 Cylinders: 84   Not Selected:107
American:253   5 Cylinders:  3   NA's        :  8
NA's    :  1   4 Cylinders:207
               3 Cylinders:  4
               NA's        :  1
```

Εδώ το zz είναι (ως προς την κλάση) ένας πίνακας δεδομένων data frame που περιέχει πολλές μεταβλητές. Όπως έχουμε αναφέρει στην summary (σελίδα 6) η "περίληψη" αυτή προσαρμόζεται ανάλογα με το είδος (κλάση) του αντικειμένου. Τα δεδομένα του αρχείου Cars.sav έχουν χρησιμοποιηθεί από το βασικό πακέτο του Spss και περιέχουν τις εξής μεταβλητές (με * είναι ποιοτικές:

MPG	Αριθμός μιλίων με ένα γαλόνι βενζίνης
ENGINE	Κυβικές ίντσες μηχανής
HORSE	Ιπποδύναμη
WEIGHT	Βάρος σε λίβρες
ACCEL	Επιτάχυνση από 0 σε 60 μίλια ανά ώρα
YEAR	Το πηλίκο της χρονιάς του μοντέλου δια 100
ORIGIN*	Χώρα προέλευσης
CYLINDER*	Αριθμός κυλίνδρων
FILTER*	Καταλυτικό (ναι ή όχι)

2.3.2 Το πακέτο ctest

Είναι μια συλλογή από κλασσικούς ελέγχους (τεστ) όπως:

binom.test	Ακριβείς διωνιμικός έλεγχος
chisq.test	Chi-squared έλεγχος του Pearson για μετρήσιμα δεδομένα
cor.test	Έλεγχος μηδενικής συσχέτισης
friedman.test	Friedman Rank Sum Test
kruskal.test	Kruskal-Wallis Rank Sum Test
ks.test	Kolmogorov-Smirnov Tests
mantelhaen.test	Cochran-Mantel-Haenszel Chi-Squared Test for Count Data
mcnemar.test	McNemar's Chi-squared Test for Count Data
oneway.test	Test for Equal Means in a One-Way Layout
pairwise.prop.test	Pairwise comparisons of proportions
pairwise.t.test	Pairwise t tests
pairwise.table	Tabulate p values for pairwise comparisons
pairwise.wilcox.test	Pairwise Wilcoxon rank sum tests
power.prop.test	Power calculations two sample test for of proportions
power.t.test	Power calculations for one and two sample t tests
prop.trend.test	Test for trend in proportions
t.test	Student's t-Test
var.test	F Test για να συγκριθούν οι διασπορές
wilcox.test	Wilcoxon Rank Sum and Signed Rank Tests

2.4 Κατανομές

Στο R μας δίνεται η δυνατότητα να χρησιμοποιήσουμε μία πληθώρα κατανομών:

Κατανομή	R εντολή	πρόσθετα ορίσματα
beta	beta	shape1, shape2, ncp
binomial	binom	size, prob
Cauchy	cauchy	location, scale
chi-squared	chisq	df, ncp
exponential	exp	rate
F	f	df1, df1, ncp
gamma	gamma	shape, scale
geometric	geom	prob
hypergeometric	hyper	m, n, k
log-normal	lnorm	meanlog, sdlog
logistic	logis	location, scale
negative	binomial	nbinom size, prob
normal	norm	mean, sd
Poisson	pois	lambda
Student"s t	t	df, ncp
uniform	unif	min, max
Weibull	weibull	shape, scale
Wilcoxon	wilcox	m, n

Βάζοντας ένα πρόθεμα μπροστά από κάθε όνομα κατανομής μπορούμε να υπολογίσουμε αντίστοιχα:

d για την συνάρτηση πυκνότητας πιθανότητας

p για την συνάρτηση κατανομής πιθανότητας

q για το ελάχιστο x ώστε δοθέντος ενός εκατοστημόριου (q) να υπολογίζει την $P((X \leq x) > q)$

r για την δημιουργία τυχαίων δεδομένων από κανονική κατανομή.

Μερικά παραδείγματα:

- Για να βρούμε την πιθανότητα 3 επιτυχιών σε 4 δοκιμές με πιθανότητα 0.5 στην διωνυμική:

```
> dbinom(3,4,0.5)
[1] 0.25
```

- Τώρα έχουμε $P(X > 2.45)$, $X \sim t$ με 5 β.ε.

```
> pt(2.45, df=5, lower.tail= FALSE)
[1] 0.0289682
```

Αν το lower.tail ήταν TRUE τότε θα υπολόγιζε: $P(X \leq 2.45)$

- Έστω ότι θέλουμε να υπολογίσουμε το 1% πάνω άκρο μιας $F(2, 7)$ κατανομής.

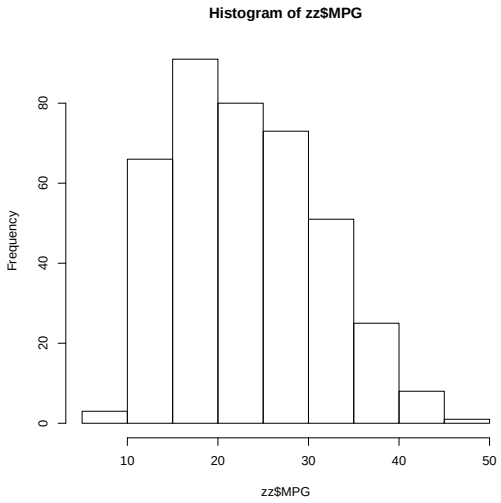
```
> qf(0.99,2,7)
[1] 9.546578
```

2.5 Γραφήματα στο R

Για τις ανάγκες των γραφημάτων θα χρησιμοποιήσουμε τα δεδομένα Cars.sav όπως τα περιγράψαμε παραπάνω.

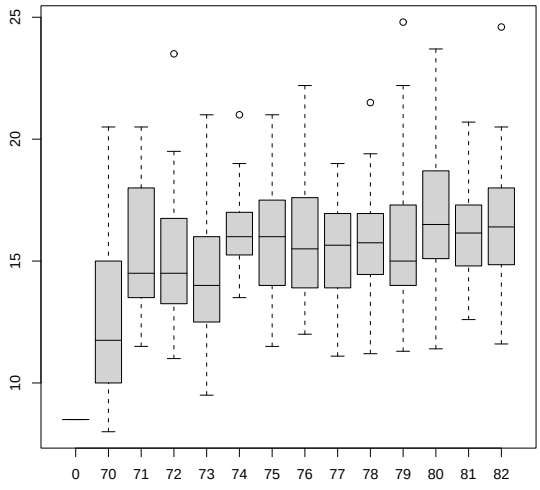
Ιστογράμματα Τα ιστογράμματα κατασκευάζονται με την εντολή:

```
> hist(zz$MPG)
```



Θηκογράμματα Τα θηκογράμματα ή διαγράμματα πλαισίων είναι ένας πολύ καλός τρόπος να αναπαραστήσουμε, διαγραμματικά, τα περιγραφικά μεγέθη ενός δείγματος ή πληθυσμού. Χρησιμεύουν επίσης στην γρήγορη σύγκριση ομάδων δεδομένων. Τα θηκογράμματα παίρνουν ορίσματα όπως τα ιστογράμματα μόνο που επιτρέπεται η προσθήκη και άλλων μεταβλητών:

```
> boxplot(ACCEL ~ YEAR, data = zz, col = "lightgray")
```

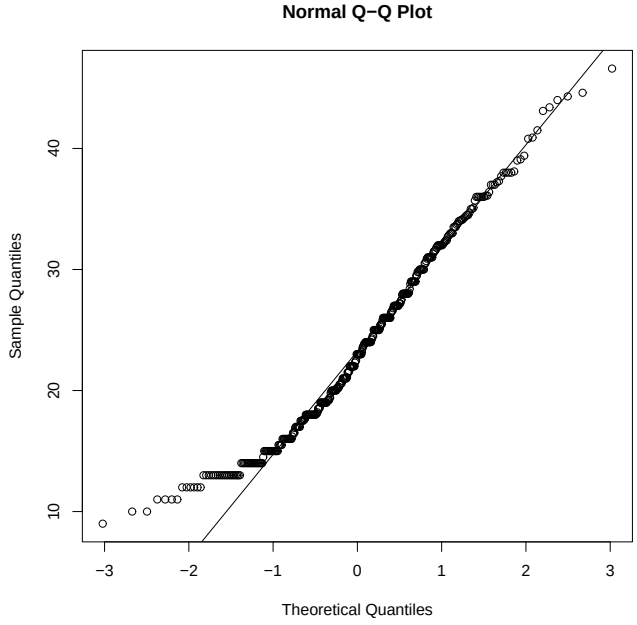


Σχήμα 2.1: Εδώ έχουμε ένα θηκόγραμμα με την επιτάχυνση των αυτοκινήτων ανά έτος κατασκευής τους, χρωματίζοντας και τα πλαίσια.

Πιθανοθεωρητικά διαγράμματα κανονικής κατανομής (normal plots) : Τα διαγράμματα απευθύνονται σε ένα δείγμα και χρησιμεύουν στο να προσδιορίσουμε πόσο κοντά στην κανονική κατανομή είναι οι παρατηρούμενες τιμές σε σχέση με αυτές της κανονικής κατανομής. Υπάρχουν δύο ειδών: τα p-p και τα q-q normal plots . Η διαφορά τους είναι στο ότι τα p-p συγκρίνουν

την αθροιστική συχνότητα με την υποτιθέμενη κανονική κατανομή ενώ τα q-q τα εκατοστημόρια τη της παρατηρούμενης τιμής ως προς την αναμενόμενη κανονική κατανομή. Η εντολή `qqnorm` σχεδιάζει το αντικείμενο που έχει ως όρισμα και η `qqline` απλά προσθέτει την γραμμή που αναπαριστά την κανονική κατανομή.

```
> qqnorm(zz$MPG) ; qqline(zz$MPG)
```



Σχήμα 2.2: Πιθανοθεωρητικό διάγραμμα q-q plot του MPG

2.5.1 Η εντολή Plot

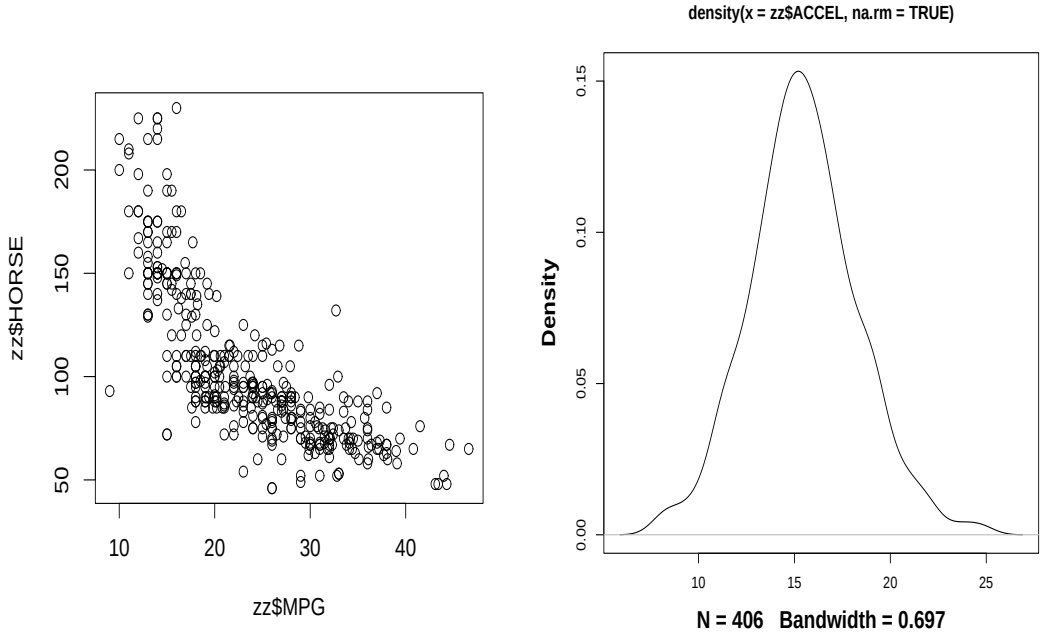
Η πιο διαδεδομένη συνάρτηση στα γραφήματα είναι η plot που είναι μια γενική συνάρτηση. Ο τύπος του γραφήματος εξαρτάται από τον τύπο ή την κατηγορία του πρώτου αντικειμένου που μπαίνει ως όρισμα. Για να κάνουμε ένα scatterplot με δύο μεταβλητές αρκεί να δώσουμε δύο ορίσματα (σχήμα 2.3):

```
> plot(zz$MPG,zz$HORSE)
```

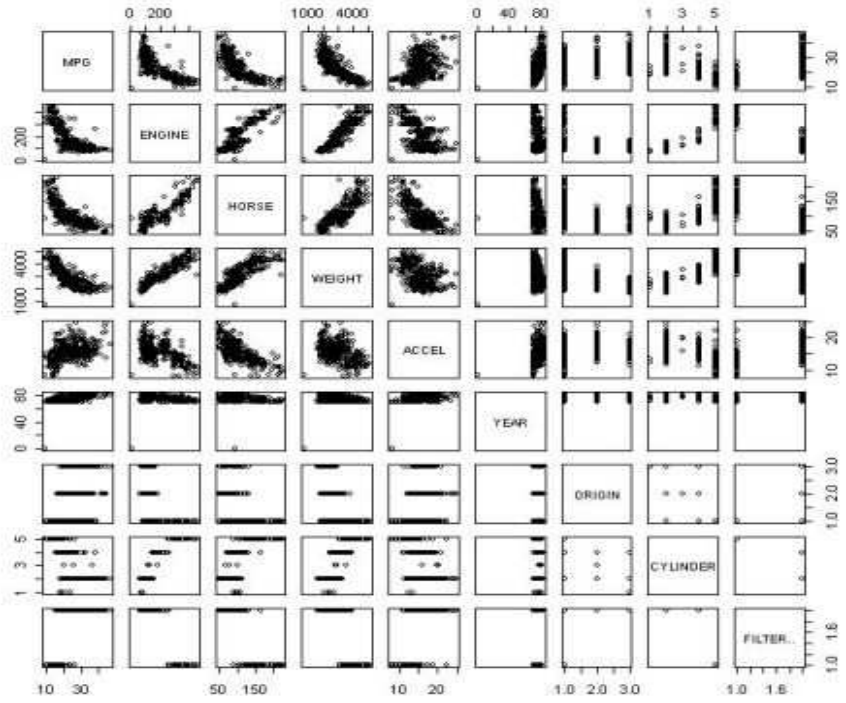
Για ένα πιο εντυπωσιακό γράφημα μπορούμε να δώσουμε το αντικείμενο χωρίς όρισμα και αυτόματα να δημιουργήσει ένα πίνακα με scatterplots (π.χ. `> plot(zz)` σχήμα 2.4). Μία άλλη επιλογή είναι να χρησιμοποιήσουμε το plot σε συνδυασμό με μια άλλη συνάρτηση όπως μια εκτίμηση του ιστογράμματος kernell (kernell density estimate):

```
>plot(density(zz$ACCEL,na.rm=TRUE))
```

Παρατηρούμε ότι η γραφική παράσταση είναι ένα είδος "λειασμένου" ιστογράμματος (smoothing) .



Σχήμα 2.3: Το πρώτο σχήμα είναι το scatterplot των zz$MPG και zz$HORSE ενώ το δεύτερο είναι η kernell density estimate .



Σχήμα 2.4: Εδώ απεικονίζεται ο πίνακας των scatterplots όλων των μεταβλητών, ποσοτικών και ποιοτικών για τα δεδομένα του αρχείου `Cars.sav` .

Κεφάλαιο 3

Έλεγχοι

3.1 Γενικά για τους έλεγχους - tests

Μία πολύ συνηθισμένη λειτουργία των στατιστικών πακέτων είναι να συγκρίνουν διάφορα μεγέθη δειγμάτων όπως η μέση τιμή και η διασπορά. Όλοι οι "κλασσικοί" έλεγχοι (τεστ) εμπεριέχονται στο πακέτο **ctest** το οποίο πρέπει να ενεργοποιηθεί πριν το χρησιμοποιήσουμε. Το πακέτο όπως περιγράψαμε περιέχει πλήθος ελέγχων, που δεν είναι δυνατή ολόκληρη η παρουσίασή τους, άλλα μια μικρή επίδειξη του.

3.2 Ένα παράδειγμα απλών ελέγχων

Τα δεδομένα, που τα χρησιμοποιούμε σε όσους έλεγχους θα παρουσιάσουμε, αναφέρονται στην λανθάνουσα θερμότητα της τήξης του πάγου [1]. Τα καταχωρούμε στα αντίστοιχα αντικείμενα A,B:

```
A <- scan()
79.98 80.04 80.02 80.04 80.03 80.03 80.04 79.97
80.05 80.03 80.02 80.00 80.02

B <- scan()
80.02 79.94 79.98 79.97 79.97 80.03 79.95 79.97
```

Ελέγχουμε την ισότητα των μέσων τιμών των δύο δειγμάτων χρησιμοποιώντας ένα μη-ζευγαρωτό t -έλεγχο (unpaired t-test) . Έστω οι υποθέσεις:

$$H_0 : \mu_A = \mu_B \Rightarrow \mu_A - \mu_B = 0$$

$$H_1 : \mu_A \neq \mu_B \Rightarrow \mu_A - \mu_B \neq 0$$

Στο R έχουμε:

```
> library(ctest)
> t.test(A, B)

        Welch Two Sample t-test
```

[1] Από το "'An introduction to R"' σελ. 38

```
data:  A and B
t = 3.2499, df = 12.027, p-value = 0.00694
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01385526 0.07018320
sample estimates:
mean of x mean of y
 80.02077  79.97875
```

Ο έλεγχος μας παρουσιάζει μια σημαντική διαφορά($p-value = 0.00694 < 0.05$), υποθέτοντας κανονικότητα. Είναι προκαθορισμένο στο R να μην δέχεται ως υπόθεση την ισότητα των διασπορών στα δύο δείγματα (σε αντίθεση με το S-plus). Μπορούμε να χρησιμοποιήσουμε ένα F -έλεγχο (F-test) για να εξετάσουμε την ισότητα των διασπορών με την βοήθεια των λόγων διασπορών, προϋποθέτοντας ότι τα δύο δείγματα προέρχονται από κανονικούς πληθυσμούς, δηλαδή:

$$H_0 : \sigma_a^2 = \sigma_b^2 \Rightarrow \frac{\widehat{S}_a^2}{\widehat{S}_b^2} = 1$$

$$H_1 : \sigma_a^2 \neq \sigma_b^2 \Rightarrow \frac{\widehat{S}_a^2}{\widehat{S}_b^2} \neq 1$$

```
> var.test(A, B)

        F test to compare two variances

data:  A and B
F = 0.5837, num df = 12, denom df =  7, p-value = 0.3938
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1251097 2.1052687
sample estimates:
ratio of variances
         0.5837405
```

Ο οποίος μας δείχνει ανυπαρξία σημαντικής διαφοράς ($p-value = 0.3938 > 0.05$), οπότε μπορούμε να χρησιμοποιήσουμε τον κλασσικό t έλεγχο που προϋποθέτει ισότητα διασπορών.

```
> t.test(A, B, var.equal=TRUE)
Two Sample t-test data:  A and B
t = 3.4722, df = 19, p-value = 0.002551
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01669058 0.06734788
sample estimates:
mean of x mean of y
 80.02077  79.97875
```

Αυτή τη φορά μας παρουσιάζει μια ακόμα πιο σημαντική διαφορά ($p - value = 0.002551 < 0.00694 < 0.05$) σε σχέση με τον προηγούμενο έλεγχο.

Κάθε έλεγχος του πακέτου ctest (2.3.2 σελ. 10) έχει τα δικά του ορίσματα και μπορούμε να πάρουμε περισσότερες πληροφορίες δίνοντας: `help( , package ="ctest")` ή προσθέτωντας το όνομα του ελέγχου π.χ.: `help(t.test , package ="ctest")`

Κεφάλαιο 4

Γραμμικά μοντέλα

Το R χειρίζεται, όπως έχουμε αναφέρει, τα μοντέλα ως **αντικείμενα**. Τα μοντέλα μοιάζουν με τις συναρτήσεις. Η κύρια διαφορά είναι ότι αποτελούνται αριθμητικές μεταβλητές, παράγοντες και φυσικά εξαρτημένες μεταβλητές που συγκρίνονται αργότερα με παρατηρούμενες αριθμητικές τιμές.

4.1 Ορισμός ενός γραμμικού μοντέλου

Για να ορίσουμε ένα μοντέλο, γενικά, ορίζουμε την εξαρτημένη μεταβλητή ακολουθεί το σύμβολο "$\sim$" και τέλος οι παράγοντες:

$$y \sim x$$

Εδώ ορίσαμε την y ως εξαρτημένη και την χ ως παράγοντα - ανεξάρτητη. Προσθέτοντας τελεστές μπορούμε να ορίσουμε τις σχέσεις των παραγόντων:

```
y ~ x +z^2 -5
```

Όπου το `y ~ x +z^2 -5` μας δηλώνει: $y = x + y^2 - 5$. Για να προσαρμόσουμε ένα γραμμικό μοντέλο δίνουμε το όρισμα:

$$object <- \mathrm{lm}(formula, \mathrm{data} = data.frame)$$

Όπου *object* είναι το αντικείμενο που θα περιέχει το μοντέλο, *formula* ο τύπος του μοντέλου και *data.frame* τα δεδομένα αν περιέχονται σε ένα μόνο αντικείμενο, για παράδειγμα:

```
> fm <- lm(MPG ~ ENGINE + HORSE , data = zz)
```

Το R αυτόματα κάνει μια ανάλυση γραμμικής παλινδρόμησης στο μοντέλο που του έχουμε εκχωρήσει τα δεδομένα τα πήραμε από το αρχείο Cars.sav (2.3.1 σελ. 9) και μπορούμε να την δούμε με:

```
> summary(fm)

Call:
lm(formula = MPG ~ ENGINE + HORSE, data = zz)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-22.2185  -3.2771  -0.3892   2.3535  16.5639

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.533786   0.755398  49.687  < 2e-16 ***
ENGINE      -0.037064   0.005067  -7.315 1.48e-12 ***
HORSE       -0.066312   0.013906  -4.769 2.63e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.67 on 389 degrees of freedom
Multiple R-Squared: 0.6439,     Adjusted R-squared: 0.6421
F-statistic: 351.7 on 2 and 389 DF,  p-value: < 2.2e-16
```

Σχετικά με την παλινδρόμηση, παρατηρούμε ότι όλες οι τιμές των F, t test είναι μικρές δηλαδή σημαντικές παράμετροι και υπάρχει σχέση μεταξύ εκτιμώμενων και παρατηρούμενων τιμών, και το R^2 μεγάλο οπότε δεχόμαστε το μοντέλο. Ουσιαστικά υποθέσαμε ότι η κατανάλωση βενζίνης σχετίζεται γραμμικά με τον κυβισμό και την ιπποδύναμη. Το μοντέλο μας είναι σύμφωνα με την γραμμική παλινδρόμηση:

$$y = -0.037064\, ENGINE - 0.066312\, HORSE + 37.533786$$

4.2 Συναρτήσεις για πιο λεπτομερειακές αναλύσεις

Στην γενική κλάση των γραμμικών μοντέλων (lm) υπάρχουν και επιπλέον συναρτήσεις για περαιτέρω εξαγωγή πληροφοριών, σχεδιασμού διαγραμμάτων κ.λ.π. Μια λίστα με αυτές είναι:

anova($object_1, object_2$) Συγκρίνει ένα υπο-μοντέλο με ένα άλλο υπαρκτό και παράγει ένα πίνακα ανάλυσης διακύμανσης.

coefficients(object) ή απλά coef(object) Παράγει έναν πίνακα συντελεστών της παλινδρόμησης.

deviance(object) Το άθροισμα των τετραγώνων των υπολοίπων (residuals).

formula(object) Παράγει τον τελικό τύπο του μοντέλου.

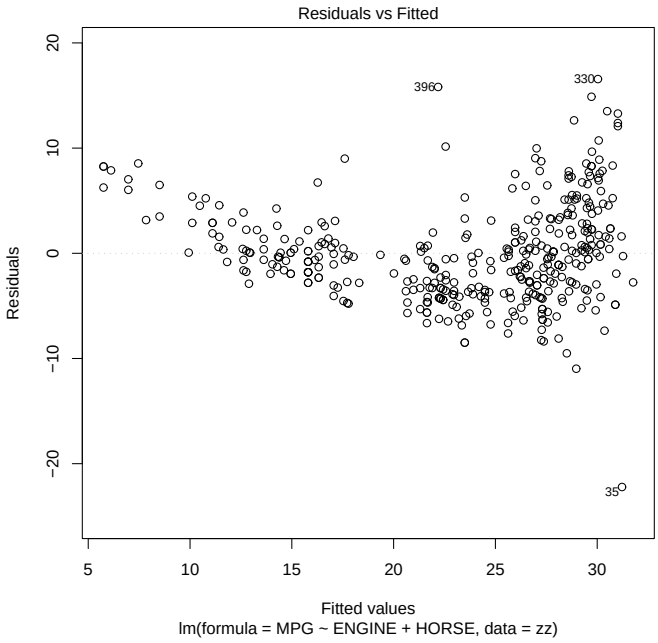
plot(object) Παράγει 4 διαγράμματα: για τα υπόλοιπα, τις προβλεπόμενες τιμές και μερικά διαγνωστικά.

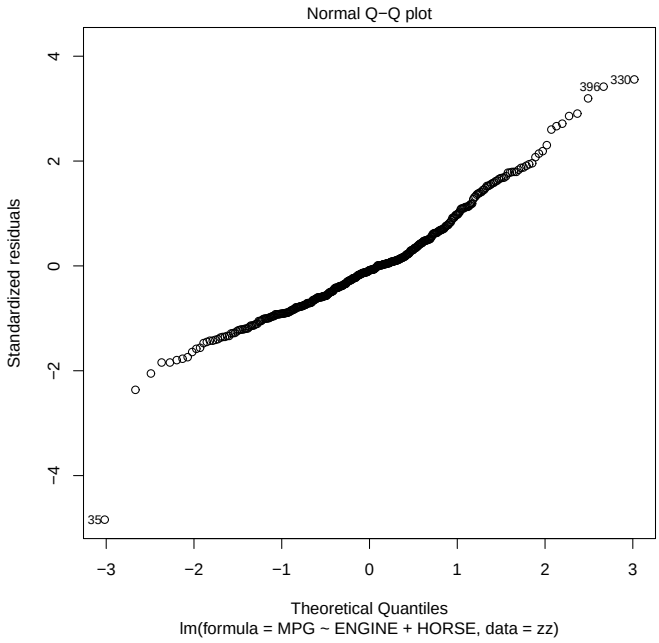
residuals(object) ή resid(object) . Ο πίνακας των υπολοίπων (residuals), σταθμισμένος κατάλληλα.

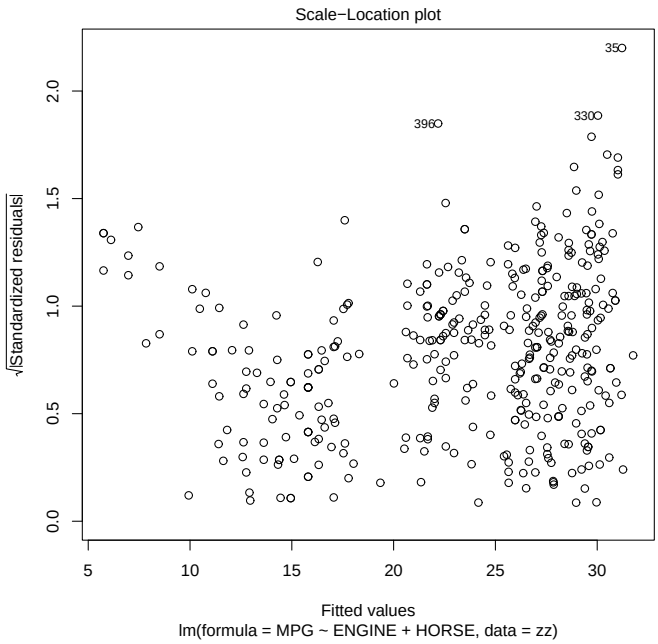
step(object) Διαλέγει ένα βέλτιστο μοντέλο προσθέτοντας ή αφαιρώντας παράγοντες διατηρώντας την ιεραρχία, σύμφωνα με τη μεγαλύτερη τιμή του AIC (Akaike's An Information Criterion) από την βηματική stepwise παλινδρόμηση.

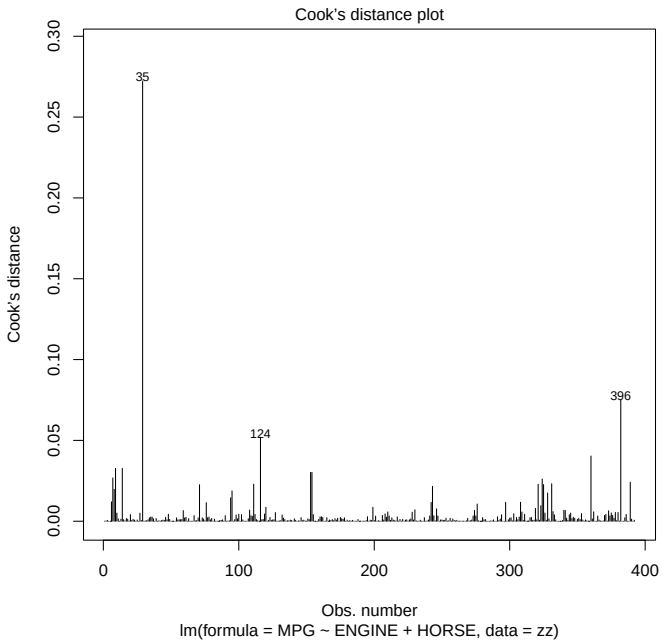
summary(object) Δίνει μια αναλυτική περίληψη των αποτελεσμάτων της ανάλυσης παλινδρόμισης.

Τέλος μερικά διαγράμματα από το μοντέλο `plot(fm)` :


Residuals vs Fitted
Residuals
Fitted values
lm(formula = MPG ~ ENGINE + HORSE, data = zz)
396
330
35


Normal Q−Q plot
Standardized residuals
Theoretical Quantiles
lm(formula = MPG ~ ENGINE + HORSE, data = zz)
396
330
35


Scale−Location plot
√|Standardized residuals|
Fitted values
lm(formula = MPG ~ ENGINE + HORSE, data = zz)
396
330
35


Cook's distance plot
Cook's distance
Obs. number
lm(formula = MPG ~ ENGINE + HORSE, data = zz)
35
124
396

Βιβλιογραφία

An itroduction to R, Bill Venables, David M. Smith and R Developent Core team, 1.6.1 (2002-11-01) ISBN 3-901167-55-2

R Data Import/Export, 1.6.1 (2002-11-01) ISBN 3-901167-53-6

Θα τα βρείτε στο www.r-project.org

Practical Regression and Anova using R, Julian Faraway, July 2002
Θα το βρείτε στον διακτυακό τόπο: http://www.stat.lsa.umich.edu/∼faraway/book
Και τέλος η εσωτερική βοήθεια-αναφορά του R .

Ευρετήριο