ΤΕΙ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΦΥΣΙΚΩΝ ΠΟΡΩΝ ΚΑΙ ΠΕΡΙΒΑΛΛΟΝΤΟΣ

«Ανάλυση διακύμανσης στη γλώσσα R»

Ελευθερία Τσίγκα Α.Μ. 242

Διπλωματική εργασία υπό την επίβλεψη του Δρ. Παπακώστα Ταξιάρχη

Περίληψη

Η παρούσα διπλωματική εργασία αποτελεί ένα εγχειρίδιο για την ανάλυση της διακύμανσης (analysisofvariance - ANOVA) στη γλώσσα R. Η τεχνική αυτή χρησιμοποιείται όταν όλες οι επεξηγηματικές μεταβλητές είναι κατηγορικές. Οι επεξηγηματικές μεταβλητές ονομάζονται παράγοντες, κάθε ένας εκ των οποίων έχει δύο ή περισσότερα επίπεδα. Στις περιπτώσεις κατά τις οποίες υπάρχει ένας μόνο παράγοντας με τρία ή περισσότερα επίπεδα χρησιμοποιούμε τη μέθοδο ANOVA ως προς ένα παράγοντα (μονοκατευθυντική ANOVA). Στην περίπτωση δύο ή περισσότερων παραγόντων, χρησιμοποιούμε ανάλυση διακύμανσης ANOVA δύο ή τριών κατευθύνσεων, ανάλογα με τον αριθμό των ερμηνευτικών μεταβλητών. Όταν υπάρχει επαναληψιμότητα σε κάθε συνδυασμό επιπέδων σε μια ανάλυση ANOVA πολλών κατευθύνσεων, το πείραμα ονομάζεται παραγοντικός σχεδιασμός. Σε αυτή την περίπτωση μας επιτρέπεται να μελετήσουμε τις αλληλεπιδράσεις μεταξύ των μεταβλητών, για τις οποίες ελέγχουμε αν η απόκριση σε έναν παράγοντα εξαρτάται από το επίπεδο κάποιου άλλου παράγοντα.

Abstract

This thesis is a manual of analysis of variance (ANOVA) in the R-language. The technique presented is used when all the explanatory variables are categorical.
The explanatory variables are called factors, and each factor has two or more levels. When there is a single factor with three or more levels we use one-way ANOVA. Where there are two or more factors, then we use two-way or three-way ANOVA, depending on the number of explanatory variables. When there is replication at each combination of levels in a multi-way ANOVA, the experiment is called a factorial design, and this allows us to study interactions between variables, in which we test whether the response to one factor depends on the level of another factor.

11. Ανάλυση Διακύμανσης

Η ανάλυση της διακύμανσης είναι η τεχνική που χρησιμοποιούμε όταν όλες οι επεξηγηματικές μεταβλητές είναι κατηγορικές. Οι επεξηγηματικές μεταβλητές ονομάζονται παράγοντες. Κάθε ένας από αυτούς τους παράγοντες έχει δύο ή περισσότερα επίπεδα. Στις περιπτώσεις κατά τις οποίες υπάρχει ένας μόνο παράγοντας με τρία ή περισσότερα επίπεδα χρησιμοποιούμε τη μέθοδο ANOVA ως προς ένα παράγοντα (μονοκατευθυντική ANOVA). Αν είχαμε ένα μόνο παράγοντα με μόλις δύο επίπεδα, θα χρησιμοποιούσαμε τις δοκιμασίες t Student (Student's t test, βλ. σελ.294), οι οποίες θα μας έδιναν ακριβώς την ίδια απάντηση που θα παίρναμε και από την μέθοδο ANOVA (θυμηθείτε τον κανόνα ότι $F=t^2$). Όταν υπάρχουν δύο ή περισσότεροι παράγοντες, τότε χρησιμοποιούμε ανάλυση διακύμανσης ANOVA δύο ή τριών κατευθύνσεων, ανάλογα με τον αριθμό των ερμηνευτικών μεταβλητών. Όταν υπάρχει επαναληψιμότητα σε κάθε συνδυασμό επιπέδων σε μια ανάλυση ANOVA πολλών κατευθύνσεων, το πείραμα ονομάζεται παραγοντικός σχεδιασμός. Σε αυτή την περίπτωση μας επιτρέπεται να μελετήσουμε τις αλληλεπιδράσεις μεταξύ των μεταβλητών, για τις οποίες ελέγχουμε, αν η απόκριση σε έναν παράγοντα εξαρτάται από το επίπεδο κάποιου άλλου παράγοντα.

Μονοκατευθυντική ANOVA

Υπάρχει ένα πραγματικό παράδοξο σχετικά με την ανάλυση της διακύμανσης το οποίο συχνά στέκεται εμπόδιο στη σαφή κατανόηση του τι ακριβώς συμβαίνει. Η ιδέα της ανάλυσης της διακύμανσης έγκειται στη σύγκριση δύο ή περισσότερων μέσων, αλλά αυτή η σύγκριση πραγματοποιείται συγκρίνοντας διακυμάνσεις. Με ποιο τρόπο μπορεί να λειτουργήσει αυτό;

Ο καλύτερος τρόπος για να δούμε τι συμβαίνει είναι να εργαστούμε μέσω ενός απλού παραδείγματος. Έχουμε ένα πείραμα στο οποίο οι ανά μονάδα επιφάνειας αποδόσεις καλλιεργειών μετρήθηκαν σε 10 τυχαία επιλεγμένα χωράφια για κάθε έναν από τρεις συγκεκριμένους τύπους εδάφους. Όλα τα χωράφια είχαν σπαρθεί με την ίδια ποικιλία σπόρων και τους είχαν παρασχεθεί τα ίδια λιπάσματα και οι ίδιοι έλεγχοι εισόδου παρασίτων. Το ερώτημα έγκειται στο αν ο τύπος του εδάφους επηρεάζει σημαντικά την απόδοση των καλλιεργειών, και, αν ναι, σε ποιο βαθμό.

```
results<-read.table("c:\\temp\\yields.txt",header
attach(results)
names(results)
[1] "sand"  "clay"  "loam"
```

Για να δείτε τα δεδομένα απλά πληκτρολογήστε την εντολή *results* και πατήστε το πλήκτρο *Return*:

```
   sand clay loam
1     6   17   13
2    10   15   16
3     8    3    9
4     6   11   12
5    14   14   15
6    17   12   16
7     9   12   17
8    11    8   13
9     7   10   18
10   11   13   14
```

Η εντολή *sapply* χρησιμοποιείται για να υπολογίσουμε τις μέσες αποδόσεις για τους τρεις διαφορετικούς τύπους εδάφους:

```
sapply(list(sand,clay,loam),mean)
[1]  9.9  11.5  14.3
```

Η μέση απόδοση ήταν υψηλότερη στο λασπώδες έδαφος (14.3) και χαμηλότερη στην άμμο (9.9).
Είναι χρήσιμο να έχουμε όλα τα δεδομένα απόδοσης σε ένα μοναδικό διάνυσμα που ονομάζεται y:

```
y<-c(sand,clay,loam)
```

και να έχουμε ένα ενιαίο διάνυσμα που ονομάζεται έδαφος, το οποίο να περιέχει τα επίπεδα του παράγοντα για τον τύπο εδάφους:

```
soil<-factor(rep(1:3,c(10,10,10)))
```

Πριν από την πραγματοποίηση της ανάλυσης της διακύμανσης, θα πρέπει να ελέγξουμε τη σταθερότητα αυτής (βλ. Κεφάλαιο 8) για τους τρεις διαφορετικούς τύπους εδάφους:

```
sapply(list(sand,clay,loam),var)
[1] 12.544444 15.388889  7.122222
```

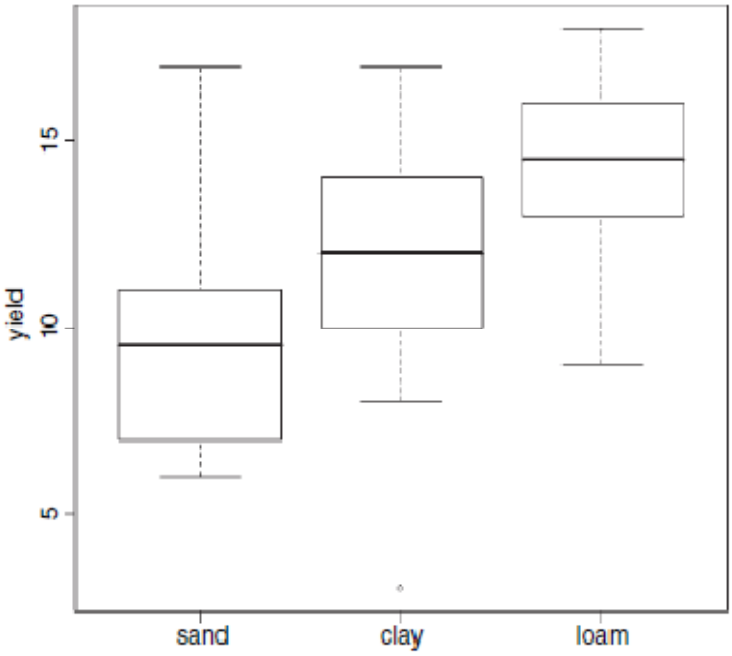
Οι διακυμάνσεις διαφέρουν κατά ένα συντελεστή μεγαλύτερο του 2. Αλλά είναι σημαντικό αυτό; Ελέγχουμε για ετεροσκεδαστικότητα χρησιμοποιώντας το Τεστ *Fligner-Killeen* της ομοιογένειας των διακυμάνσεων.

```
fligner.test(y~soil)

        Fligner-Killeen test of homogeneity of variances

data: y by soil
Fligner-Killeen:med chi-squared = 0.3651, df = 2, p-value = 0.8332
```

Θα μπορούσαμε να είχαμε χρησιμοποιήσει ένα *bartlett.test* (y ~ soil), το οποίο δίνει μια τιμή p = 0,5283 (αλλά αυτό είναι περισσότερο ένα τεστ μη κανονικότητας παρά ένα τεστ ισότητας των διακυμάνσεων). Είτε έτσι είτε αλλιώς, δεν υπάρχει καμία απόδειξη για οποιαδήποτε σημαντική διαφορά στη διακύμανση μεταξύ των τριών δειγμάτων. Συνεπώς είναι εύλογο να συνεχίσουμε με την ως προς ένα παράγοντα, ανάλυση της διακύμανσης.

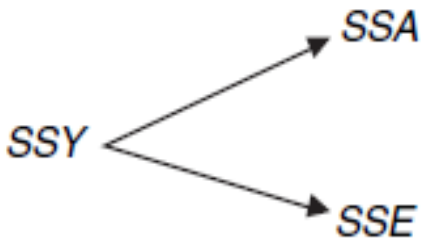
Επειδή η επεξηγηματική μεταβλητή είναι κατηγορική (τρία επίπεδα τύπου εδάφους), ο αρχικός έλεγχος των δεδομένων απαιτεί ένα διάγραμμα box-and-whisker του y έναντι του εδάφους, όπως αυτό:

```
plot(soil,y,names=c("sand","clay","loam"),ylab="yield")
```



Η μέση απόδοση είναι χαμηλότερη στην άμμο και υψηλότερη στο λασπώδες έδαφος, αλλά υπάρχει σημαντική απόκλιση από επανάληψη σε επανάληψη σε κάθε τύπο εδάφους (στο αργιλώδες έδαφος υπάρχει μέχρι και μια ακραία τιμή). Φαίνεται σαν η απόδοση στο λασπώδες έδαφος να αποδεικνύεται σημαντικά υψηλότερη από αυτή της άμμου (τα κουτιά τους στο διάγραμμα δεν επικαλύπτονται), αλλά δεν είναι σαφές το κατά πόσο η απόδοση στο αργιλώδες έδαφος είναι σημαντικά υψηλότερη από αυτή στην άμμο ή σημαντικά χαμηλότερη από ό, τι στο λασπώδες έδαφος. Η ανάλυση της διακύμανσης θα απαντήσει σε αυτά τα ερωτήματα.

Η ανάλυση της διακύμανσης περιλαμβάνει τον υπολογισμό της συνολικής απόκλισης της μεταβλητής απόκρισης (η απόδοση στην προκείμενη περίπτωση) και την κατανομή αυτής (αναλύοντάς την) σε πληροφοριακά συστατικά. Στην απλούστατη περίπτωση, κατανέμουμε τη συνολική απόκλιση σε μόλις δύο συνιστώσες: στην παραγοντική (παλινδρομική) και στην υπόλοιπη μεταβολή:



Παλινδρομική μεταβολή ονομάζεται το επεξεργασμένο άθροισμα των τετραγώνων των αποκλίσεων (SSA) ενώ υπόλοιπη ή μη παραγοντική μεταβολή καλείται το άθροισμα των τετραγώνων σφάλματος (SSE, επίσης γνωστό ως άθροισμα των τετραγώνων των καταλοίπων). Η υπόλοιπη μεταβολή ορίζεται ως το άθροισμα των τετραγώνων των διαφορών μεταξύ της μεμονωμένης τιμής y και του σχετικού μέσου μεταχείρισης:

$$SSE = \sum_{i=1}^{k} \sum (y - \bar{y}_i)^2.$$

Υπολογίζουμε εκ των προτέρων τη διάμεσο για το ν-οστό επίπεδο του παράγοντα, και στη συνέχεια προσθέτουμε τα τετράγωνα των διαφορών. Δεδομένου ότι εργαστήκαμε με αυτόν τον τρόπο, μπορείτε να δείτε πόσοι βαθμοί ελευθερίας θα πρέπει να συνδέονται με το SSE; Ας υποθέσουμε ότι υπήρχαν ν επαναλήψεις σε κάθε μεταχείριση (ν = 10 στο παράδειγμά μας). Και ας υποθέσουμε ότι υπάρχουν κ επίπεδα του παράγοντα (k = 3 στο παράδειγμά μας). Αν εκτιμήσετε κ παραμέτρους από τα δεδομένα πριν να υπολογίσετε το SSE, τότε, κατά τη διαδικασία, θα πρέπει να έχετε χάσει k βαθμούς ελευθερίας. Αφού καθένα από τα κ επίπεδα του παράγοντα έχει ν επαναλήψεις, θα πρέπει να υπάρχουν $k \times n$ αριθμοί σε όλο το πείραμα ($3*10 = 30$, στο παράδειγμά μας). Οπότε, οι βαθμοί ελευθερίας που συνδέονται με το SSE είναι $kn-k=k(n-1)$. Ένας άλλος τρόπος για να το δούμε αυτό είναι να πούμε ότι υπάρχουν n επαναλήψεις σε κάθε μεταχείριση, και ως εκ τούτου, $n-1$ βαθμοί ελευθερίας σφάλματος (επειδή 1 d.f. χάνεται στην εκτίμηση του κάθε μέσου μεταχείρισης). Υπάρχουν k μεταχειρίσεις (π.χ. k επίπεδα του παράγοντα) και ως εκ τούτου υπάρχει $k \times (n-1)$ d.f. για σφάλμα στο σύνολο του πειράματος.

Η συνιστώσα της μεταβολής που εξηγείται από τις διαφορές μεταξύ των μέσων μεταχείρισης, το άθροισμα των τετραγώνων των μεταχειρίσεων, παραδοσιακά συμβολίζεται ως SSA. Αυτό συμβαίνει επειδή στην ανάλυση της διακύμανσης δύο κατευθύνσεων, με δύο διαφορετικές κατηγορικές επεξηγηματικές μεταβλητές, το SSB

χρησιμοποιείται για να υποδηλώσει το άθροισμα των τετραγώνων που αναλογεί στις διαφορές μεταξύ των μέσων του δεύτερου παράγοντα, και το SSC για να υποδηλώσει το άθροισμα των τετραγώνων που αναλογεί στις διαφορές που υπάρχουν μεταξύ των μέσων του τρίτου παράγοντα, και ούτω καθεξής.

Τυπικά, υπολογίζουμε όλες, εκτός από μία, τις συνιστώσες της συνολικής διακύμανσης, και στη συνέχεια βρίσκουμε την τιμή της τελευταίας συνιστώσας αφαιρώντας τις υπόλοιπες από το SSY. Έχουμε ήδη ένα τύπο υπολογισμού του SSE, και έτσι θα μπορούσαμε να υπολογίσουμε το SSA από τη διαφορά: SSA = SSY-SSE. Στο πλαίσιο 11.1 φαίνεται περισσότερο λεπτομερώς ο τύπος υπολογισμού του SSA.

Box 11.1

Διορθωμένα αθροίσματα των τετραγώνων σε μονοκατευθυντική ANOVA

Ο ορισμός του συνολικού αθροίσματος των τετραγώνων, SSY, είναι το άθροισμα των τετραγώνων των διαφορών μεταξύ των σημείων δεδομένων, yij, και της συνολικής μέσης τιμής, $\bar{\bar{y}}$:

$$SSY = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{\bar{y}})^2,$$

όπου $\sum_{j=1}^{n} y_{ij}$ είναι το άθροισμα των n επαναλήψεων σε κάθε ένα από τα κ επίπεδα του παράγοντα. Το άθροισμα των τετραγώνων των σφαλμάτων, SSE , είναι το άθροισμα των τετραγώνων των διαφορών μεταξύ των σημείων δεδομένων, yij και των αντίστοιχων μέσων μεταχείρισης. $\bar{y}_i$:

$$SSE = \sum_{i-1}^{k} \sum_{j-1}^{n} (y_{ij} - \bar{y}_i)^2.$$

Το άθροισμα των τετραγώνων των μεταχειρίσεων, SSA, είναι το άθροισμα των τετραγώνων των διαφορών μεταξύ του αντίστοιχου μέσου μεταχείρισης , $\bar{y}_i$ και της συνολικής μέσης τιμής, $\bar{\bar{y}}$:

$$SSA = \sum_{i=1}^{k} \sum_{j=1}^{n} (\bar{y}_i - \bar{\bar{y}})^2 = n \sum_{i=1}^{k} (\bar{y}_i - \bar{\bar{y}})^2.$$

Τετραγωνίζοντας τον όρο μέσα στις αγκύλες ή εφαρμόζοντας άθροιση έχουμε:

$$\sum \bar{y}_i^2 - 2\bar{\bar{y}} \sum \bar{y}_i + k\bar{\bar{y}}^2.$$

Ας αφήσουμε το γενικό σύνολο όλων των τιμών της μεταβλητής απόκρισης $\sum_{i=1}^{k} \sum_{j=1}^{n} y_{ij}$ να εμφανίζεται ως $\sum y.$. Τώρα αντικαθιστούμε το $\bar{y}_i$ με το T_i/n (όπου Τ είναι η συμβατική ονομασία μας για τα k επιμέρους σύνολα μεταχειρίσης) και το $\bar{\bar{y}}$ με το $\sum y/kn$ για να πάρουμε

$$\frac{\sum_{i=1}^{k} T_i^2}{n^2} - 2\frac{\sum y \sum_{i=1}^{k} T_i}{nkn} + k\frac{\sum y \sum y}{knkn}.$$

Σημειώστε ότι $\sum_{i=1}^{k} T_i = \sum_{i=1}^{k} \sum_{j=1}^{n} y_{ij}$ οπότε οι προς τα δεξιά θετικοί και αρνητικοί όροι έχουν τη μορφή $(\sum y)^2/kn^2$.

Τέλος, πολλαπλασιάζοντας όλους τους όρους με το n έχουμε:

$$SSA = \frac{\sum T^2}{n} - \frac{(\sum y)^2}{kn}.$$

Ως άσκηση, θα πρέπει να αποδείξετε ότι SSY = SSA + SSE.

Ας εργαστούμε μέσα από τους αριθμούς που ανήκουν στην R. Από τον τύπο υπολογισμού του SSY, μπορούμε να εξαγάγουμε το καθολικό άθροισμα των τετραγώνων βρίσκοντας τις διαφορές μεταξύ των δεδομένων και του συνολικού μέσου:

```
sum((y-mean(y))^2)
[1] 414.7
```

Η υπόλοιπη μεταβολή, SSE, υπολογίζεται από τις διαφορές μεταξύ των αποδόσεων και των μέσων αποδόσεων για κάθε συγκεκριμένο τύπο εδάφους:

```
sand-mean(sand)
[1]  -3.9  0.1  -1.9  -3.9  4.1  7.1  -0.9  1.1  -2.9  1.1
clay-mean(clay)
[1]  5.5  3.5  -8.5  -0.5  2.5  0.5  0.5  -3.5  -1.5  1.5
loam-mean(loam)
[1]  -1.3  1.7  -5.3  -2.3  0.7  1.7  2.7  -1.3  3.7  -0.3
```

Χρειαζόμαστε τα αθροίσματα των τετραγώνων αυτών των διαφορών:

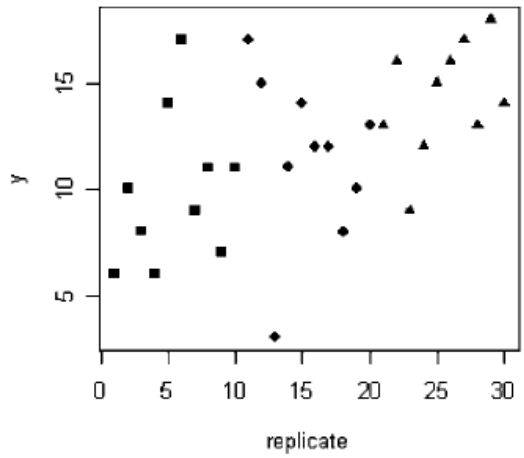
```
sum((sand-mean(sand))^2)
[1]  112.9
sum((clay-mean(clay))^2)
[1]  138.5
sum((loam-mean(loam))^2)
[1]  64.1
```

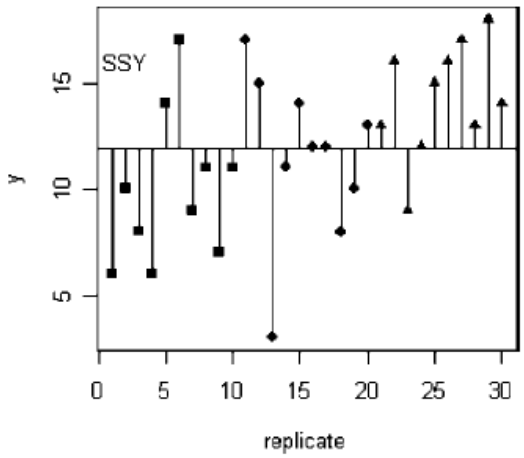
Για να πάρουμε το άθροισμα αυτών των συνόλων για όλους τους τύπους εδάφους, μπορούμε να χρησιμοποιήσουμε την εντολή *sapply* ως εξής:

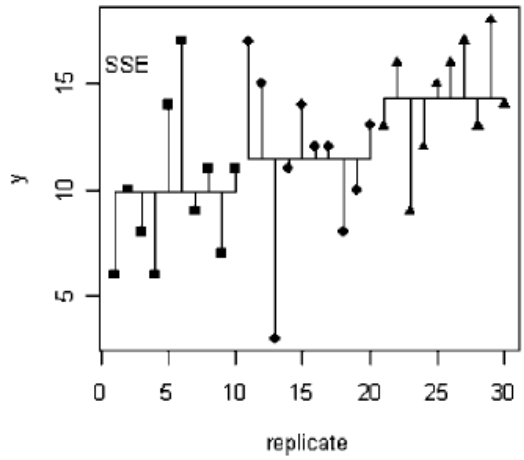
```
sum(sapply(list(sand,clay,loam),function (x) sum((x-mean(x))^2) ))
[1]  315.5
```

Οπότε το SSE, το ανερμήνευτο (ή υπόλοιπο ή λάθος) άθροισμα των τετραγώνων είναι 315,5.

Ο βαθμός κατά τον οποίο το SSE είναι μικρότερο από το SSY αποτελεί μια αντανάκλαση του μεγέθους των διαφορών μεταξύ των μέσων. Όσο μεγαλύτερη είναι η διαφορά μεταξύ των μέσων αποδόσεων στους διαφορετικούς τύπους εδάφους, τόσο μεγαλύτερη θα είναι η διαφορά μεταξύ των SSE και SSY. Αυτό είναι η βάση της ανάλυση της διακύμανσης. Μπορούμε να εξαγάγουμε συμπεράσματα σχετικά με τις διαφορές μεταξύ των μέσων εξετάζοντας τις διαφορές μεταξύ των διακυμάνσεων (ή για να είμαστε περισσότερο ακριβείς σε αυτό στάδιο, εξετάζοντας τις διαφορές μεταξύ των αθροισμάτων των τετραγώνων).







Πάνω αριστερά έχουμε ένα «ενδεικτικό γράφημα» των αποδόσεων χρησιμοποιώντας διαφορετικά σύμβολα για τους διαφορετικούς τύπους εδάφους: τετράγωνο = άμμος, διαμάντι = αργιλώδες έδαφος, τρίγωνο = λασπώδες έδαφος. Πάνω δεξιά είναι μια εικόνα του συνολικού αθροίσματος των τετραγώνων: το SSY είναι το άθροισμα των τετραγώνων των μηκών των γραμμών που συνδέουν κάθε δεδομένο με τη συνολική μέση τιμή, $\bar{\bar{y}}$. Κάτω αριστερά είναι μια εικόνα από το άθροισμα των τετραγώνων των σφαλμάτων: το SSE είναι το άθροισμα των τετραγώνων των μηκών των γραμμών που συνδέουν κάθε δεδομένο με το συγκεκριμένο μέσο μεταχείρισής του, $\bar{y}_i$. Ο βαθμός κατά τον οποίο οι γραμμές στο SSE είναι κοντύτερες από τις αντίστοιχες του SSY είναι ένα μέτρο της σπουδαιότητας της διαφοράς μεταξύ των μέσων αποδόσεων για τα διαφορετικά εδάφη. Στην ακραία περίπτωση κατά την οποία δεν υπήρχε μεταβολή μεταξύ των επαναλήψεων, το SSY είναι μεγάλο, αλλά το SSE είναι μηδενικό:

Αυτή η εικόνα δημιουργήθηκε χρησιμοποιώντας τον ακόλουθο κώδικα, όπου οι τιμές x, xvc, είναι

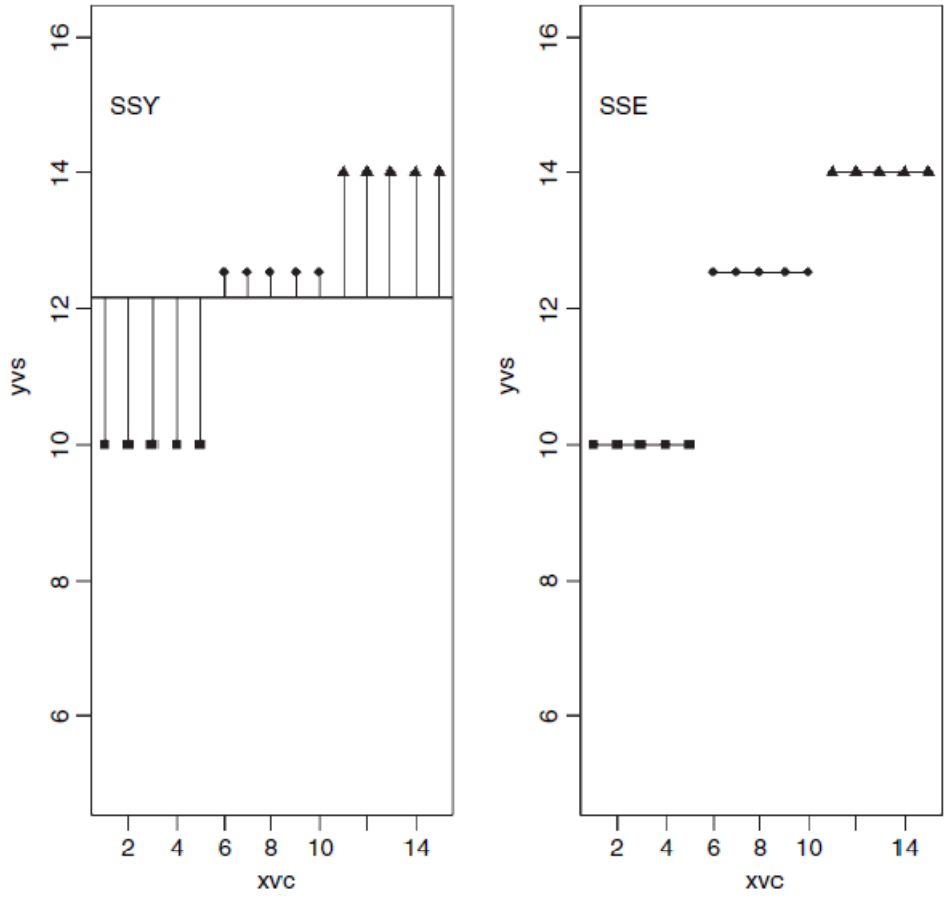
```
xvc<-1:15
```

και οι τιμές y, yvs είναι

```
yvs<-rep(c(10,12,14),each=5)
```

Για να παράγουμε διαδοχικά τα δύο γραφήματα, γράφουμε:

```
par(mfrow=c(1,2))
plot(xvc,yvs,ylim=c(5,16),pch=(15+(xvc>5)+(xvc>10)))
for (i in 1:15) lines(c(i,i),c(yvs[i],mean(yvs)))
```




```
abline(h=mean(yvs))
text(3,15,"SSY")
plot(xvc,yvs, ylim=c(5,16),pch=(15+(xvc(>)5)+(xvc(>)10)))
lines(c(1,5),c(10,10))
lines(c(6,10),c(12,12))
lines(c(11,15),c(14,14))
text(3,15,"SSE")
```

Η διαφορά μεταξύ των SSY και SSE ονομάζεται άθροισμα των τετραγώνων των μεταχειρίσεων και το SSA είναι το ποσό της μεταβολής στην απόδοση που εξηγείται από τις διαφορές μεταξύ των μέσων των μεταχειρίσεων.

Στο παράδειγμά μας,

$$SSA = SSY - SSE = 414.7 - 315.5 = 99.2.$$

Τώρα μπορούμε να καταρτίσουμε τον πίνακα ANOVA. Υπάρχουν έξι στήλες, από αριστερά προς τα δεξιά, που δείχνουν την πηγή της μεταβολής, το άθροισμα των τετραγώνων που αποδίδεται στην εν λόγω πηγή, τους βαθμούς ελευθερίας αυτής της πηγής, τη διακύμανση της (παραδοσιακά ονομάζεται τετραγωνικός μέσος αντί για διακύμανση), τον λόγο F (εξετάζοντας την μηδενική υπόθεση ότι αυτή η πηγή μεταβολής δεν είναι σημαντικά

διαφορετική από το μηδέν) και την τιμή p που σχετίζεται με τον εν λόγω λόγο F (εάν p <0,05 τότε απορρίπτουμε την μηδενική υπόθεση). Μπορούμε να συμπληρώσουμε τα αθροίσματα των τετραγώνων απλά υπολογισμένα, και έπειτα να σκεφτούμε τους βαθμούς ελευθερίας:

Υπάρχουν συνολικά 30 σημεία δεδομένων, οπότε το σύνολο των βαθμών ελευθερίας είναι 30-1 = 29. Χάνουμε 1 d.f. επειδή κατά τον υπολογισμό του SSY έπρεπε εκ των προτέρων να εκτιμήσουμε μία παράμετρο από τα δεδομένα, τη συνολική μέση τιμή, $\bar{\bar{y}}$.

Source	Sum of squares	Degrees of freedom	Mean square	F ratio	p value
Soil type	99.2	2	49.6	4.24	0.025
Error	315.5	27	$s^2 = 11.685$		
Total	414.7	29			

Κάθε τύπος εδάφους έχει n = 10 επαναλήψεις, οπότε κάθε τύπος εδάφους έχει 10-1 = 9 d.f. σφάλμα, διότι εκτιμήσαμε μία παράμετρο από τα δεδομένα για κάθε τύπο εδάφους, δηλαδή τα μέσα μεταχείρισης $\bar{y}_i$ κατά τον υπολογισμό του SSE. Κατά συνέπεια, στο σύνολό του, το σφάλμα έχει 3 × 9 = 27 d.f. Υπήρχαν 3 τύποι εδάφους, οπότε υπάρχουν 3-1 = 2 df για κάθε έναν τύπο.

Τα μέσα τετράγωνα λαμβάνονται απλώς διαιρώντας κάθε άθροισμα των τετραγώνων με τους βαθμούς ελευθερίας που του αντιστοιχούν (στην ίδια γραμμή). Η διακύμανση του σφάλματος s^2, είναι το υπολειπόμενο μέσο τετράγωνο (το μέσο τετράγωνο για τη μη επεξηγηματική μεταβολή). Αυτό μερικές φορές καλείται «συγκεντρωτικό σφάλμα διακύμανσης», επειδή υπολογίζεται για όλες τις μεταχειρίσεις. Η εναλλακτική λύση θα ήταν να έχουμε τρεις ξεχωριστές διακυμάνσεις, μία για κάθε μεταχείριση:

```
sapply(list(sand,clay,loam),var)
[1]  12.544444  15.388889  7.122222
```

Θα δείτε ότι η συγκεντρωτική διακύμανση σφάλματος s^2=11,685 είναι απλά ο μέσος των τριών χωριστών διακυμάνσεων καθώς υπάρχουν ίσες επαναλήψεις σε κάθε τύπο εδάφους (n = 10):

```
mean(sapply(list(sand,clay,loam),var))
[1]  11.68519
```

Παραδοσιακά, δεν υπολογίζουμε τον συνολικό τετραγωνικό μέσο, για αυτό, το τελευταίο κελί της τέταρτης στήλης του πίνακα ANOVA είναι κενό. Ο λόγος F είναι η διακύμανση της μεταχείρισης διαιρούμενη με τη διακύμανση σφάλματος, ελέγχοντας τη μηδενική υπόθεση οι μέσοι των μεταχειρίσεων να είναι όλοι
ίδιοι. Αν απορρίψουμε αυτή την μηδενική υπόθεση, αποδεχόμαστε την εναλλακτική ότι τουλάχιστον ένας από τους μέσους είναι σημαντικά διαφορετικός από τους άλλους. Το

ερώτημα που τίθεται φυσικά σε αυτό το σημείο αφορά στο αν το 4,24 είναι ή όχι ένας μεγάλος αριθμός. Αν πρόκειται για ένα μεγάλο αριθμό τότε απορρίπτουμε τη μηδενική υπόθεση. Στην αντίθετη περίπτωση, την αποδεχόμαστε. Όπως πάντα, αποφασίζουμε αν η στατιστική δοκιμασία F=4,24 είναι μεγάλη ή μικρή συγκρίνοντάς την με την κρίσιμη τιμή F, δεδομένου ότι υπάρχουν 2 d.f. στον αριθμητή και 27 d.f. στον παρονομαστή. Οι κρίσιμες τιμές στην R βρέθηκαν από τη συνάρτηση *qf* η οποία μας δίνει τις ποσότητες της κατανομής F:

```
qf(.95,2,27)
[1] 3.354131
```

Το στατιστικό τεστ, που έχουμε υπολογίσει ίσο με 4,24, είναι μεγαλύτερο από την κρίσιμη τιμή των 3,35, οπότε απορρίπτουμε τη μηδενική υπόθεση. Τουλάχιστον ένα από τα εδάφη έχει μέση απόδοση που είναι σημαντικά διαφορετική από τις αντίστοιχες των άλλων. Η σύγχρονη προσέγγιση είναι να μην εργαστούμε καταναγκαστικά στο επίπεδο του 5%, αλλά να υπολογίσουμε την τιμή p που σχετίζεται με το στατιστικό τεστ του 4,24. Αντί να χρησιμοποιήσουμε τη συνάρτηση για τις ποσότητες της κατανομής F, χρησιμοποιούμε τη συνάρτηση *pf* για τις σωρευτικές πιθανότητες της κατανομής F ως ακολούθως:

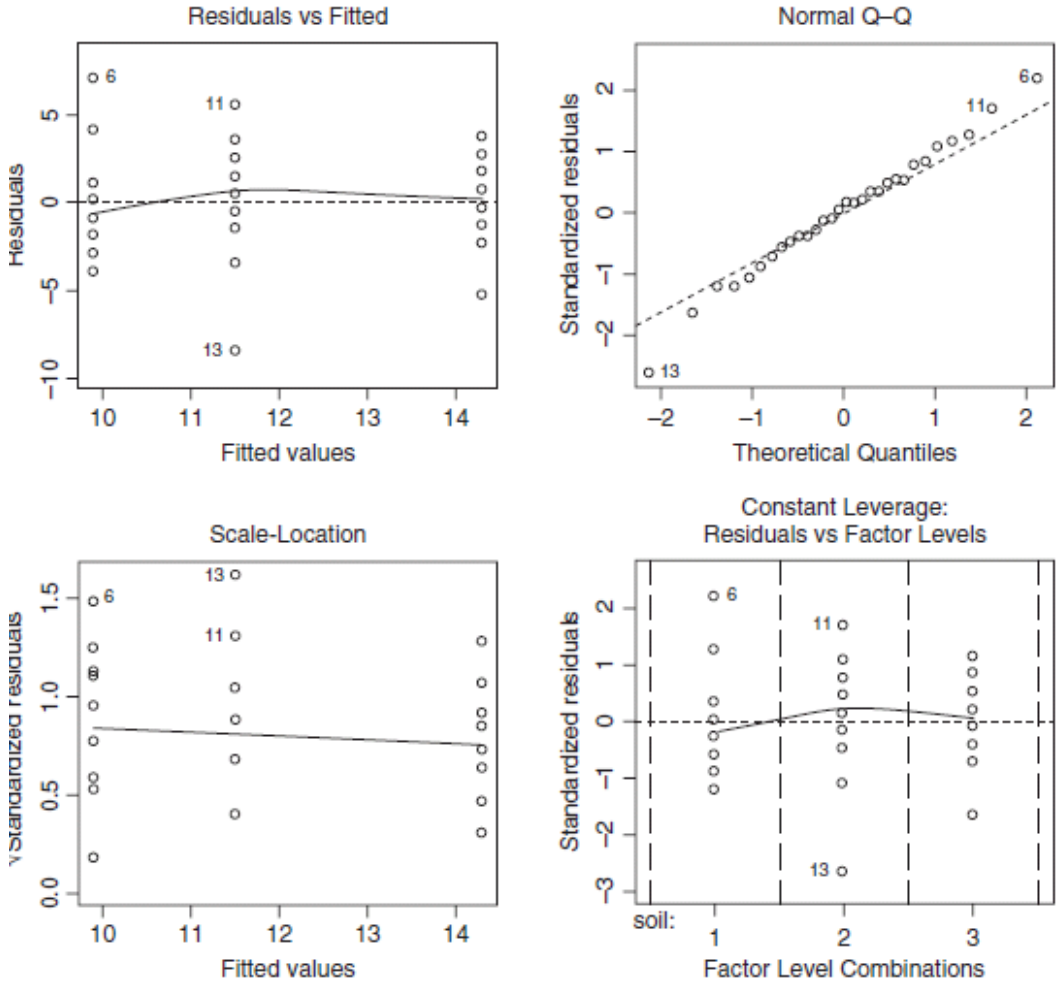
```
1-pf(4.24,2,27)
[1] 0.02503987
```

Η τιμή p είναι ίση 0,025, που σημαίνει ότι μια τιμή του F=4,24 ή μεγαλύτερη θα προκύψει τυχαία όταν η μηδενική υπόθεση είναι αληθινή σε αναλογία 25/1000. Αυτή είναι μία αρκούντως μικρή πιθανότητα (δηλαδή είναι μικρότερη από 5%) για να συμπεράνουμε ότι υπάρχει μια σημαντική διαφορά μεταξύ των μέσων αποδόσεων (δηλαδή απορρίπτουμε την μηδενική υπόθεση).

Αυτό ήταν πολλή δουλειά. Η R μπορεί να κάνει την όλη διαδικασία σε μια μόνο γραμμή:

```
summary(aov(y~soil))
            Df  Sum Sq Mean Sq F value  Pr(>F)
soil         2  99.200  49.600  4.2447 0.02495 *
Residuals   27 315.500  11.685
```

Εδώ βλέπετε όλες τις τιμές που υπολογίζονται. Η γραμμή των σφαλμάτων είναι επισημασμένη ως *Residuals*. Στη δεύτερη και στις επόμενες στήλες βλέπετε τους βαθμούς ελευθερίας για κάθε μεταχείριση και σφάλμα αντίστοιχα (2 και 27), τη μεταχείριση και το άθροισμα των τετραγώνων των σφαλμάτων (99,2 και 315,5), τον τετραγωνικό μέσο της μεταχείρισης που είναι 49,6, τη διακύμανση του σφάλματος s^2=11,685, το λόγο F και την τιμή p (με την επισήμανση ότι Pr (> F). Ο μονός αστερίσκος δίπλα στην τιμή p δείχνει ότι η διαφορά



μεταξύ των μέσων κάθε εδάφους είναι σημαντική στο ποσοστό του 5% (αλλά όχι στο 1%, στο οποίο θα είχαν δοθεί δύο αστερίσκοι). Παρατηρήστε ότι η R δεν εκτυπώνει την κάτω γραμμή του πίνακα ANOVA η οποία δείχνει το ολικό άθροισμα των τετραγώνων και το σύνολο των βαθμών ελευθερίας.
Το επόμενο πράγμα που θα κάναμε είναι να ελέγξουμε τις υποθέσεις του μοντέλου aov. Αυτό γίνεται χρησιμοποιώντας ένα γράφημα σαν αυτό (βλ. Κεφάλαιο 10):

```
plot(aov(y~soil))
```

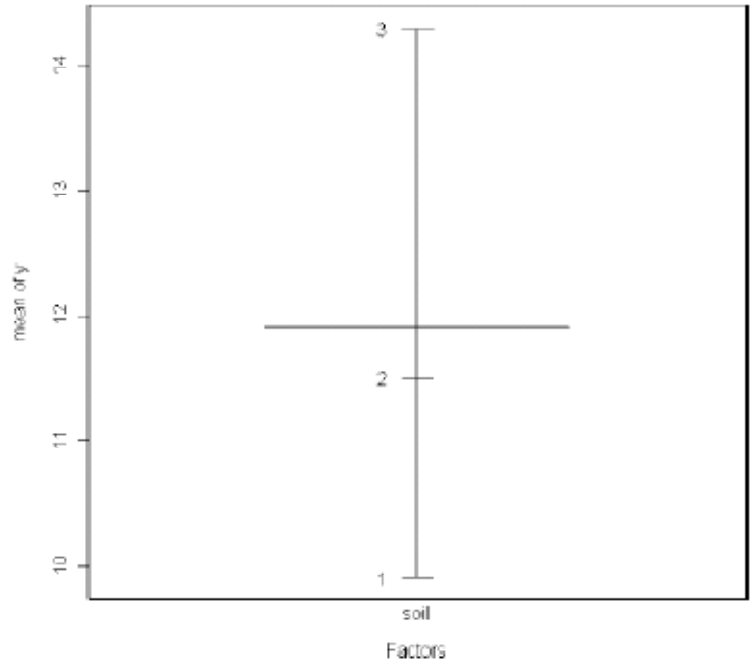
Το πρώτο γράφημα (πάνω αριστερά) ελέγχει την πιο σημαντική υπόθεση της σταθερότητας της διακύμανσης. Δεν πρέπει να υπάρχει μοτίβο στα κατάλοιπα έναντι των προσαρμοσμένων τιμών (τα τρία μέσα μεταχείρισης) και πράγματι, δεν υπάρχει. Το δεύτερο γράφημα (πάνω δεξιά) ελέγχει την υπόθεση της κανονικότητας των σφαλμάτων: θα πρέπει να υπάρχει μια ευθεία σχέση ανάμεσα στα τυποποιημένα μας υπολείμματα και στα θεωρητικά ποσοστημόρια που προκύπτουν από μια κανονική κατανομή. Τα σημεία 6,

11 και 13 βρίσκονται λίγο έξω από την ευθεία γραμμή, αλλά αυτό δεν είναι κάτι που θα πρέπει να μας ανησυχήσει (βλ. σελ. 339). Τα κατάλοιπα συμπεριφέρονται σωστά (κάτω αριστερά) ενώ δεν υπάρχουν ιδιαίτερα επιδραστικές τιμές που ενδεχομένως να στρεβλώσουν τις εκτιμήσεις των παραμέτρων (κάτω δεξιά).

Επίδραση μεγέθους

Ο καλύτερος τρόπος για να δείτε τις επιδράσεις μεγέθους είναι να χρησιμοποιήσετε τη συνάρτηση *plot.design* (η οποία χρησιμοποιεί ένα τύπο και όχι ένα μοντέλο αντικειμένου):

```
plot.design(y~soil)
```



Για περισσότερο περίπλοκα μοντέλα, ίσως να θέλετε να χρησιμοποιήσετε τη «βιβλιοθήκη» *effects* προκειμένου να πάρετε περισσότερο ελκυστικά γραφήματα (σελ. 178). Για να δείτε την επίδραση μεγέθους σε μορφή πίνακα χρησιμοποιήστε την *model.tables* (η οποία παίρνει ένα μοντέλο αντικειμένου ως επιχείρημά), ως ακολούθως:

```
model<-aov(y~soil);model.tables(model,se=T)
```

```
Tables of effects

  soil
soil
   1     2    3
-2.0  -0.4  2.4

Standard errors of effects
         soil
        1.081
replic.    10
```

Οι επιδράσεις εμφανίζονται ως αποκλίσεις από τον συνολικό μέσο: το έδαφος 1 (άμμος) έχει μέση απόδοση που είναι κατά 2,0 μικρότερη από τον συνολικό μέσο, και το έδαφος 3 (λασπώδες έδαφος) έχει μέση απόδοση 2,4 πάνω από τον συνολικό μέσο. Το τυπικό σφάλμα των επιδράσεων είναι 1,081 με μια επανάληψη n = 10 (αυτό είναι το πρότυπο σφάλμα ενός μέσου). Θα πρέπει να σημειώσετε ότι αυτό δεν είναι το κατάλληλο τυπικό σφάλμα για τη σύγκριση δύο μέσων (βλέπε παρακάτω). Αν συγκεκριμενοποιήσετε χρησιμοποιώντας την εντολή *means* έχετε:

```
model.tables(model,"means",se=T)

Tables of means
Grand mean

11.9

  soil
soil
   1     2     3
 9.9  11.5  14.3

Standard errors for differences of means
          soil
         1.529
replic.     10
```

Τώρα εκτυπώνονται οι τρεις μέσοι (και όχι οι επιδράσεις) και το τυπικό σφάλμα της διαφοράς των μέσων δίνεται (αυτό είναι ό, τι χρειάζεστε για να κάνετε ένα t-test ώστε να συγκρίνετε τους όποιους δυο μέσους).
Ένας άλλος τρόπος θεώρησης των επιδράσεων μεγέθους είναι να χρησιμοποιήσετε την *summary.lm* για να δείτε το μοντέλο, αντί για την *summary.aov* (όπως χρησιμοποιήσαμε προηγούμενα):

```
summary.lm(aov(y~soil))

Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    9.900       1.081    9.158  9.04e-10  ***
soil2          1.600       1.529    1.047   0.30456
soil3          4.400       1.529    2.878   0.00773  **
Residual standard error: 3.418 on 27 degrees of freedom
Multiple R-Squared: 0.2392,  Adjusted R-squared: 0.1829
F-statistic: 4.245 on 2 and 27 DF, p-value: 0.02495
```

Στην ανάλυση παλινδρόμησης (σελ. 399), το αποτέλεσμα της *summary.lm* ήταν εύκολα κατανοητό επειδή μας έδωσε το σημείο τομής και την κλίση (οι δύο παράμετροι που υπολογίστηκαν από το μοντέλο) και τα τυπικά τους σφάλματα. Ωστόσο, ο πίνακας αυτός έχει τρεις σειρές. Γιατί συμβαίνει αυτό; Τι είναι ένα σημείο τομής, στο πλαίσιο της ανάλυσης της διακύμανσης; Και γιατί τα τυπικά σφάλματα είναι διαφορετικά για το σημείο τομής και για τα εδάφη των τύπων 2 και 3; Καταλήγοντας, τι είναι το έδαφός 2 και τι το έδαφος 3;

Θα πάρει λίγο καιρό μέχρι να νιώσετε άνετα με τους πίνακες *summary.lm* για την ανάλυση της διακύμανσης. Οι λεπτομέρειες εξηγούνται στη σελίδα 365, αλλά το κεντρικό σημείο είναι ότι όλοι οι πίνακες *summary.lm* έχουν τόσες γραμμές όσες είναι και οι παράμετροι που υπολογίστηκαν από τα δεδομένα. Υπάρχουν τρεις σειρές σε αυτή την περίπτωση, επειδή το μοντέλο aov υπολογίζει τρεις παραμέτρους - μια μέση απόδοση - για κάθε έναν από τους τρεις τύπους εδάφους. Στο πλαίσιο του aov, ένα σημείο τομής είναι μια μέση τιμή. Σε αυτή την περίπτωση είναι η μέση απόδοση για την άμμο, γιατί δώσαμε σε αυτόν τον παράγοντα το επίπεδο 1 όταν υπολογίσαμε τις τιμές για τον παράγοντα «χώμα». Σε γενικές γραμμές, το σημείο τομής θα είναι το επίπεδο του παράγοντα του οποίου το όνομα ήρθε τελευταίο κατά αλφαβητική σειρά (βλ. σελ. 366). Οπότε, αν το σημείο τομής είναι η μέση απόδοση για την άμμο, ποιες είναι οι άλλες δύο σειρές με την ένδειξη soil2 και soil3. Αυτό, σε επίπεδο κατανόησης, είναι το πιο δύσκολο κομμάτι . Όλες οι άλλες σειρές στον πίνακα *summary.lm* του aov είναι διαφορές μεταξύ των μέσων. Έτσι η γραμμή 2, με την επισήμανση soil2, είναι η διαφορά μεταξύ των μέσων αποδόσεων σε άμμο και άργιλο, και η γραμμή 3, με την επισήμανση soil3, είναι η διαφορά μεταξύ των μέσων αποδόσεων ανάμεσα στο αμμώδες και στο αργιλώδες έδαφος:

```
tapply(y,soil,mean)-mean(sand)
  1   2   3
0.0 1.6 4.4
```

Η πρώτη γραμμή είναι ένας μέσος, έτσι ώστε η στήλη του τυπικού σφάλματος σε αυτήν τη γραμμή να περιέχει το τυπικό σφάλμα ενός μέσου. Οι γραμμές 2 και 3 είναι οι διαφορές μεταξύ των μέσων, οπότε οι αντίστοιχες στήλες τους τυπικού σφάλματος περιέχουν το τυπικό σφάλμα της διαφοράς μεταξύ δυο μέσων (και αυτό είναι ένας μεγαλύτερος αριθμός, βλ. σελ. 367). Το τυπικό σφάλμα ενός μέσου είναι:

$$se_{\text{mean}} = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{11.685}{10}} = 1.081,$$

λαμβάνοντας υπόψη ότι το τυπικό σφάλμα της διαφοράς μεταξύ δύο μέσων είναι

$$se_{\text{diff}} = \sqrt{2\frac{s^2}{n}} = \sqrt{2 \times \frac{11.685}{10}} = 1.529.$$

Ο πίνακας *summary.lm* δείχνει ότι το έδαφος τύπου 3 παράγει σημαντικά μεγαλύτερες αποδόσεις από το έδαφος τύπου 1 (το σημείο τομής) με τιμή p ίση με 0,00773. Η διαφορά μεταξύ των δύο μέσων ήταν 4,400 και το τυπικό σφάλμα της διαφοράς ήταν 1,529. Αυτή η διαφορά είναι σπουδαιότητας δύο αστέρων με την έννοια ότι 0,001<p<0,01. Αντιθέτως, το έδαφος τύπου 2, δεν παράγει κάποια σημαντικά μεγαλύτερη απόδοση από το έδαφος τύπου 1. Η διαφορά είναι 1,600 και το τυπικό σφάλμα της διαφοράς ήταν 1,529 (p= 0,30456). Το μόνο ζήτημα που απομένει σχετίζεται με το αν το έδαφος τύπου 2 είχε σημαντικά μικρότερη απόδοση από το έδαφος τύπου 3. Θα χρειαστούμε να κάνουμε κάποια νοερή αριθμητική για να το δούμε αυτό: η διαφορά μεταξύ των δυο αυτών μέσων ήταν 4,4-1,6 = 2,8 και έτσι η τιμή t είναι 2,8/1,529=1,83. Αυτή είναι μικρότερη από 2 (εμπειρικός κανόνας για το t), οπότε οι μέσες αποδόσεις των εδαφών τύπου 2 και 3 δεν είναι σημαντικά διαφορετικές. Για τον προσδιορισμό της ακριβούς τιμής με 10 επαναλήψεις, η κρίσιμη τιμή t δίδεται από τη συνάρτηση qt με 18 d.f.:

```
qt(0.975,18)
[1]  2.100922
```

Εναλλακτικά, μπορούμε να επεξεργαστούμε την τιμή p που σχετίζεται με το ήδη υπολογισμένο t =1,83:

```
2*(1 - pt(1.83, df = 18))
[1]  0.0838617
```

δεδομένου ότι p = 0,084. Πολλαπλασιάζουμε επί 2, επειδή αυτό είναι ένα two-tailed test (διαφέρει από το μηδέν)(βλ. σελ. 208). Δεν γνωρίζαμε εκ των προτέρων ότι, κάτω από τις ιδιαίτερες συνθήκες αυτού του πειράματος, το λασπώδες έδαφος ξεπερνάει σε απόδοση το αργιλώδες.

Το υπόλοιπο τυπικό σφάλμα στο αποτέλεσμα του *summary.lm* είναι η τετραγωνική ρίζα της διακύμανσης σφάλματος από τον πίνακα ANOVA: $\sqrt{11.685} = 3.418.$ Τα R-Squared και τα προσαρμοσμένα R-Squared εξηγούνται στη σελίδα 399. Η F- στατιστική και η τιμή p προέρχονται από τις δύο τελευταίες στήλες του πίνακα ANOVA.

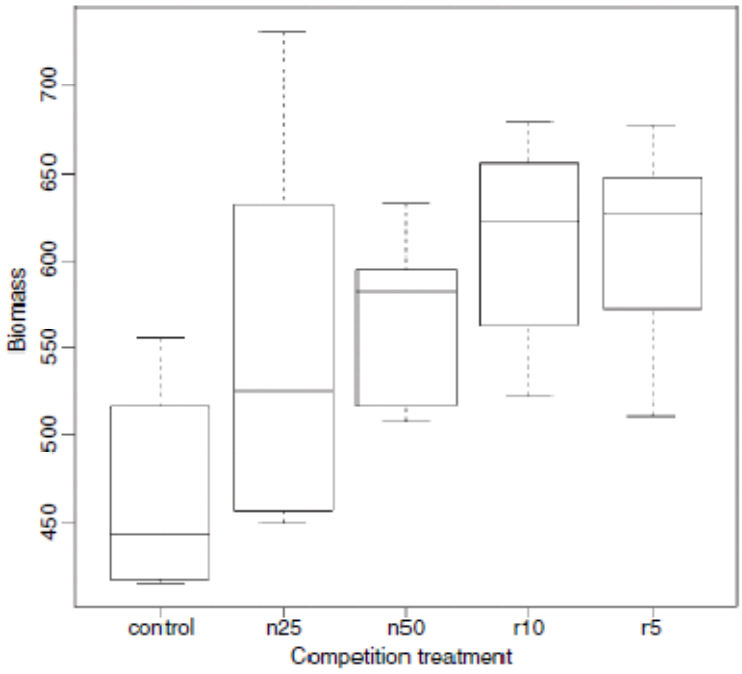
Οπότε, έτσι δουλεύει η ανάλυση της διακύμανσης. Όταν οι μέσοι είναι σημαντικά διαφορετικοί, τότε το άθροισμα των τετραγώνων που υπολογίζεται για κάθε επιμέρους μεταχείριση θα είναι σημαντικά μικρότερο από το άθροισμα των τετραγώνων που υπολογίζεται από τον συνολικό μέσο. Κρίνουμε τη σπουδαιότητα της διαφοράς μεταξύ των δύο αθροισμάτων των τετραγώνων χρησιμοποιώντας την ανάλυση της διακύμανσης.

Γραφήματα για την ερμηνεία της μονόδρομης ANOVA

Υπάρχουν δύο παραδοσιακοί τρόποι για να δείξουμε γραφικά τα αποτελέσματα της ANOVA:
• Γραφήματα box-and-whisker (κορμός και ουρές).
• Ραβδογραφήματα με ράβδους σφάλματος.

Ακολουθεί ένα παράδειγμα για να συγκρίνουμε τις δύο προσεγγίσεις. Έχουμε ένα πείραμα για τον ανταγωνισμό των φυτών με έναν παράγοντα και πέντε επίπεδα. Ο παράγοντας λέγεται ψαλίδισμα και τα πέντε επίπεδα συνίστανται στα ακόλουθα: έλεγχος (για παράδειγμα μη ψαλιδισμένα- unclipped), δύο εντάσεις κλαδέματος του φυτού και δύο εντάσεις κλαδέματος της ρίζας:

```
comp<-read.table("c:\\temp\\competition.txt",header=T);attach(comp);names(comp)
[1] "biomass"  "clipping"
plot(clipping,biomass,xlab="Competition treatment",ylab="Biomass")
```



Το γράφημα box-and-whisker είναι καλό στο να δείχνει τη φύση της μεταβολής σε κάθε τύπο μεταχείρισης, καθώς επίσης και το αν υπάρχει παραποίηση εντός οποιουδήποτε από αυτούς τους τύπους (π.χ. για τα γραφήματα ελέγχου, υπάρχει ένα ευρύτερο φάσμα τιμών μεταξύ της διάμεσου και του τρίτου τεταρτημόριου από ό, τι μεταξύ της διάμεσου και το πρώτου τεταρτημόριου). Δεν εμφανίζονται ακραίες τιμές πάνω από τις ουρές, οπότε οι κορυφές και οι πυθμένες των ράβδων αποτελούν τα μέγιστα και τα ελάχιστα σε κάθε μεταχείριση. Οι διάμεσοι για τις ανταγωνιστικές μεταχειρίσεις είναι όλες υψηλότερες από το τρίτο τεταρτημόριο των ελέγχων, γεγονός που υποδηλώνει ότι μπορεί να είναι

σημαντικά διαφορετικές από τους ελέγχους. Ωστόσο είναι δύσκολο να προταθεί ότι κάποια από τις ανταγωνιστικές μεταχειρίσεις είναι σημαντικά διαφορετική από οποιαδήποτε άλλη (δείτε παρακάτω για την ανάλυση). Θα μπορούσαμε να χρησιμοποιήσουμε την επιλογή *notch =T* για να αποκομίσουμε μια οπτική άποψη της σημασίας των διαφορών μεταξύ των μέσων. Όλες οι διάμεσοι των μεταχειρίσεων είναι έξω από την εγκοπή των ελέγχων, αλλά δεν υπάρχουν άλλες συγκρίσεις που να φαίνονται σημαντικές.

Τα ραβδογραφήματα με μπάρες σφάλματος προτιμώνται από πολλούς συντάκτες εφημερίδων, ενώ κάποιοι άνθρωποι θεωρούν ότι κάνουν τον έλεγχο της υπόθεσης ευκολότερο. Θα πρέπει να το δούμε αυτό. Σε αντίθεση με την S-PLUS, η R δεν έχει ενσωματωμένη εντολή που να ονομάζεται *error.bar* οπότε θα πρέπει να γράψουμε τη δική μας. Εδώ είναι μια πολύ απλή εκδοχή. Θα την ονομάσουμε *error.bars* ώστε να τη διακρίνουμε από την πολύ γενικότερη συνάρτηση S-PLUS.

```
error.bars<-function(yv,z,nn) {
xv<-
barplot(yv,ylim=c(0,(max(yv)+max(z))),names=nn,ylab=deparse(substitute(yv)
))
g=(max(xv)-min(xv))/50
for (i in 1:length(xv))  {
lines(c(xv[i],xv[i]),c(yv[i]+z[i],yv[i]-z[i]))
lines(c(xv[i]-g,xv[i]+g),c(yv[i]+z[i], yv[i]+z[i]))
lines(c(xv[i]-g,xv[i]+g),c(yv[i]-z[i], yv[i]-z[i]))
}}
```

Για να χρησιμοποιήσουμε αυτή τη συνάρτηση θα πρέπει να αποφασίσουμε τι είδους τιμές (z) θα χρησιμοποιήσουμε για τα μήκη των ράβδων. Ας χρησιμοποιήσουμε το τυπικό σφάλμα ενός μέσου με βάση τα σωρευμένα στοιχεία διακύμανσης των σφαλμάτων από την ANOVA, και στη συνέχεια θα επιστρέψουμε σε μια συζήτηση για τα πλεονεκτήματα και τα μειονεκτήματα των διαφορετικών ειδών ράβδων σφάλματος. Εδώ είναι η μονόδρομη ανάλυση διακύμανσης:

```
model<-aov(biomass~clipping)
summary(model)
```

```
            Df   Sum Sq  Mean Sq  F value    Pr(>F)
clipping     4    85356    21339   4.3015  0.008752  **
Residuals   25   124020     4961
```

Από τον πίνακα ANOVA μαθαίνουμε ότι η συγκεντρωτική διακύμανση σφάλματος είναι s^2=4961.0. Τώρα χρειάζεται να μάθουμε πόσοι αριθμοί χρησιμοποιήθηκαν για τον υπολογισμό αυτών των πέντε μέσων:

```
table(clipping)
```

```
clipping
control    n25   n50   r10   r5
      6      6     6     6    6
```

Υπήρξε ίση επανάληψη (γεγονός που καθιστά την ανάλυση ευκολότερη), και κάθε μέσος βασίστηκε σε έξι επαναλήψεις, οπότε το τυπικό σφάλμα του μέσου είναι $\sqrt{s^2/n} = \sqrt{4961/6} = 28.75.$ Θα πρέπει να σχεδιάσουμε μια ράβδο σφάλματος μέχρι το 28,75 από κάθε μέσο και προς τα κάτω κατά την ίδια απόσταση, γι 'αυτό χρειαζόμαστε 5 τιμές z, μια για κάθε ράβδο, της τάξης του 28,75:

```
se<-rep(28.75,5)
```

Θα πρέπει να επισημάνουμε τις πέντε διαφορετικές ράβδους: τα επίπεδα του παράγοντα πρέπει να είναι κατάλληλα γι 'αυτό:

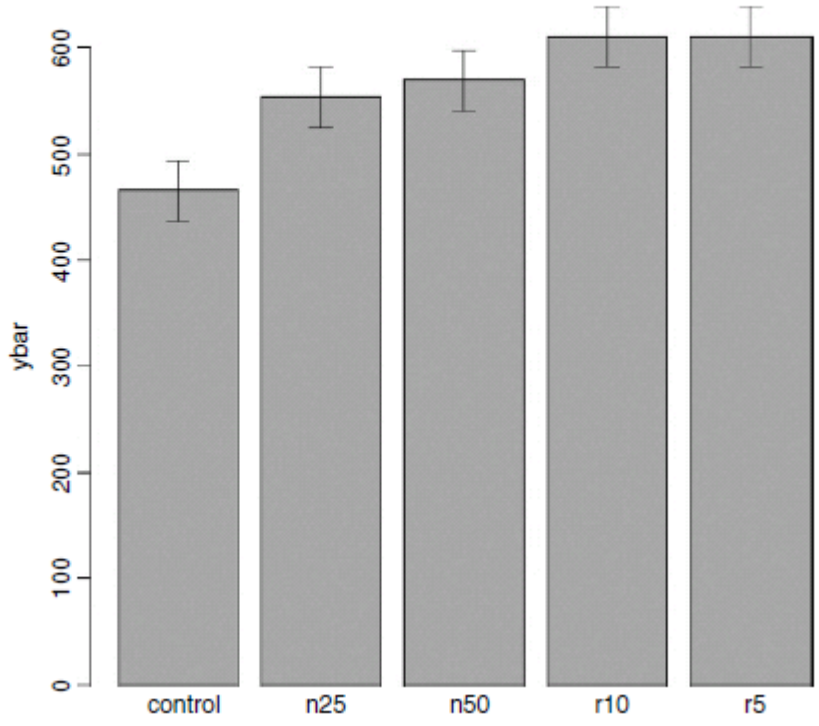
```
labels<-as.character(levels(clipping))
```

Τώρα επεξεργαζόμαστε τις τιμές των πέντε μέσων, οι οποίες θα είναι τα ύψη των ράβδων, και τις αποθηκεύσουμε ως ένα διάνυσμα που ονομάζεται *ybar*:

```
ybar<-as.vector(tapply(biomass,clipping,mean))
```

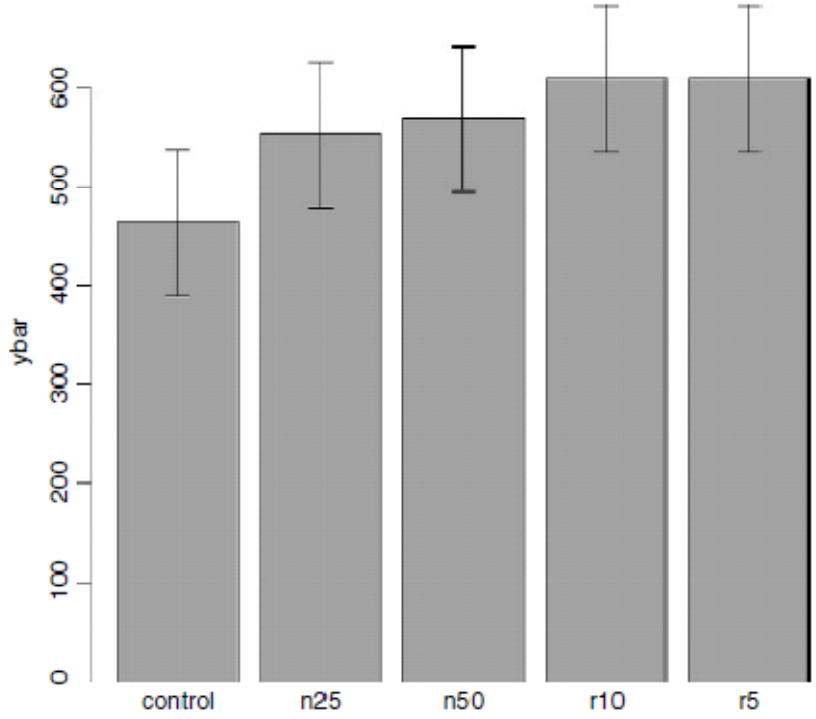
Τέλος, μπορούμε να δημιουργήσουμε το ραβδόγραμμα με μπάρες σφάλματος (η συνάρτηση ορίζεται στη σελ. 462):

```
error.bars(ybar,se,labels)
```



Δεν αποκομίζουμε την ίδια αίσθηση για την κατανομή των τιμών μέσα σε κάθε μεταχείριση, όπως αυτές είχαν ληφθεί από το γράφημα box-and-whisker, αλλά μπορούμε σίγουρα να δούμε καθαρά ποιοι μέσοι δεν είναι σημαντικά διαφορετικοί. Εάν, όπως εδώ, χρησιμοποιούμε ± 1 τυπικό σφάλμα ως το μήκος των ράβδων σφάλματος, τότε, όταν οι ράβδοι επικαλύπτονται, οι δύο μέσοι δεν είναι σημαντικά διαφορετικοί. Θυμηθείτε τον εμπειρικό κανόνα για το t: η σπουδαιότητα απαιτεί 2 ή περισσότερα τυπικά σφάλματα, και το ενδεχόμενο οι ράβδοι να αλληλεπικαλύπτονται συνεπάγεται ότι η διαφορά μεταξύ των μέσων είναι μικρότερη από 2 τυπικά σφάλματα. Υπάρχει επίσης και ένα άλλο ζήτημα. Για την σύγκριση μέσων, θα πρέπει να χρησιμοποιήσουμε το τυπικό σφάλμα της διαφοράς μεταξύ δύο μέσων (όχι το τυπικό σφάλμα ενός μέσου) στις δοκιμές μας (βλέπε σελ. 294). Αυτές οι ράβδοι θα είναι περίπου 1,4 φορές οι ράβδοι που έχουμε σχεδιάσει εδώ. Οπότε, ενώ μπορούμε να είμαστε βέβαιοι ότι οι δυο μεταχειρίσεις κλαδέματος ρίζας δεν διαφέρουν σημαντικά μεταξύ τους, και ότι οι δύο μεταχειρίσεις κλαδέματος κορμού δεν είναι επίσης σημαντικά διαφορετικές μεταξύ τους (επειδή οι ράβδοι τους επικαλύπτονται), δεν μπορούμε να συμπεράνουμε από αυτό το γράφημα ότι οι έλεγχοι έχουν σημαντικά χαμηλότερη βιομάζα από τους υπόλοιπους (επειδή οι ράβδοι σφάλματος δεν έχουν το σωστό μήκος για τον έλεγχο των διαφορών μεταξύ των μέσων).

Μια εναλλακτική γραφική μέθοδος είναι αντί για τα τυπικά σφάλματα των μέσων, να χρησιμοποιήσουμε 95% διαστήματα εμπιστοσύνης για τα μήκη των ράβδων. Αυτό είναι εύκολο να το κάνουμε: πολλαπλασιάζουμε τα τυπικά μας σφάλματα από το Student's t test, qt (.975,5) = 2,570 582, για να πάρουμε τα μήκη των διαστημάτων εμπιστοσύνης:



Τώρα, όλες οι ράβδοι σφάλματος επικαλύπτονται, υπονοώντας οπτικά ότι δεν υπάρχουν σημαντικές διαφορές μεταξύ των μέσων. Αλλά από την ανάλυση της διακύμανσης στην οποία απορρίψαμε τη μηδενική υπόθεση ότι όλοι οι μέσοι ήταν ίδιοι και ίσοι με p=0,00875

ξέρουμε ότι αυτό δεν είναι αληθινό. Αν είχαμε την περίπτωση κατά την οποία οι ράβδοι δεν επικαλύπτονταν όταν χρησιμοποιούμε διαστήματα εμπιστοσύνης (όπως εδώ), αυτό θα σήμαινε ότι οι μέσοι διέφεραν κατά περισσότερο από 4 τυπικά σφάλματα. Πρόκειται για μια πολύ μεγαλύτερη διαφορά από αυτήν που απλά απαιτείται για να συμπεράνουμε ότι οι μέσοι είναι σημαντικά διαφορετικοί. Οπότε, ούτε αυτό είναι τέλειο. Με τα τυπικά σφάλματα θα μπορούσαμε να είμαστε σίγουροι ότι οι μέσοι δεν ήταν σημαντικά διαφορετικοί όταν οι ράβδοι επικαλύπτονταν. Και με τα διαστήματα εμπιστοσύνης μπορούμε να είμαστε βέβαιοι ότι οι μέσοι είναι σημαντικά διαφορετικοί όταν οι ράβδοι δεν επικαλύπτονταν. Αλλά οι εναλλακτικές περιπτώσεις δεν είναι ξεκάθαρα αποκομμένες για κανένα από τα είδη ράβδων. Μπορούμε με κάποιο τρόπο να πάρουμε το καλύτερο και από τους δυο τρόπους, έτσι ώστε οι μέσοι να είναι σημαντικά διαφορετικοί, όταν οι ράβδοι δεν επικαλύπτονται, και όχι σημαντικά διαφορετικοί όταν οι ράβδοι επικαλύπτονται;

Η απάντηση είναι ναι, μπορούμε, αν χρησιμοποιήσουμε μπάρες ελάχιστης σημαντικής διαφοράς (LSD). Ας επανεξετάσουμε τον τύπο του Student's t test:

$$t = \frac{\text{a difference}}{\text{standard error of the difference}}$$

Λέμε ότι η διαφορά είναι σημαντική όταν t> 2 (σύμφωνα με τον εμπειρικό κανόνα, ή t> qt (0.975, df) αν θέλουμε να είμαστε πιο ακριβείς). Μπορούμε να αναδιατάξουμε τον τύπο για να βρούμε τη μικρότερη διαφορά που θα μπορούσαμε να θεωρήσουμε ως σημαντική. Μπορούμε να την ονομάσουμε ως τη λιγότερο σημαντική διαφορά:

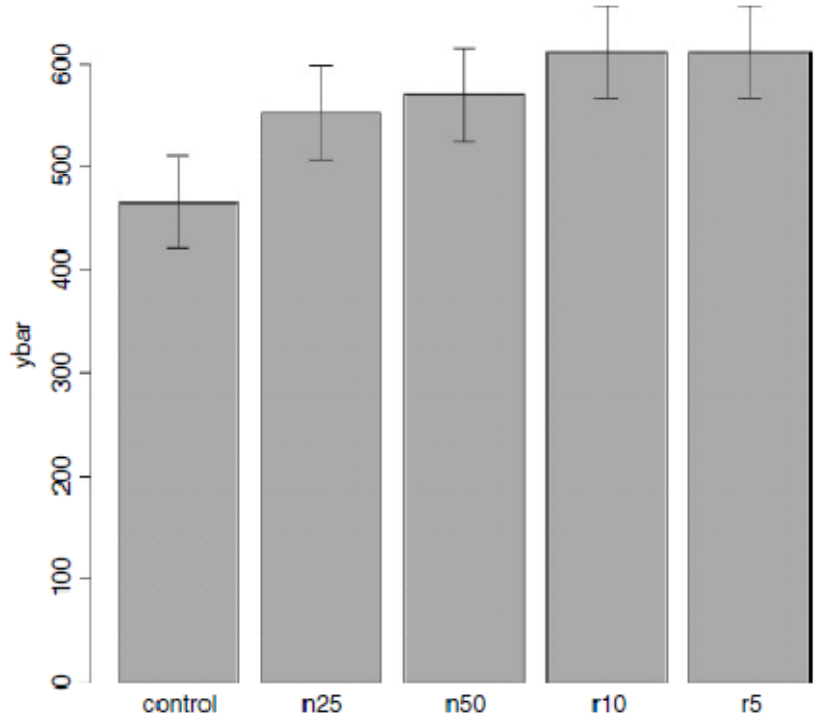
$$LSD = \text{qt}(0.975,\text{df}) \times \text{standard error of a difference} \approx 2 \times se_{\text{diff}}.$$

Στο παρόν παράδειγμα αυτή είναι:

```
qt(0.975,10)*sqrt(2*4961/6)
[1]  90.60794
```

επειδή μια διαφορά βασίζεται σε 12-2 = 10 βαθμούς ελευθερίας. Αυτό που λέμε είναι ότι οι δυο μέσοι θα ήταν σημαντικά διαφορετικοί εάν διέφεραν κατά 90,61 ή περισσότερο. Πώς μπορούμε να το δείξουμε αυτό διαγραμματικά; Θέλουμε τις επικαλυπτόμενες ράβδους να υποδεικνύουν μια διαφορά μικρότερη από 90,61, και τις μη επικαλυπτόμενες ράβδους να αντιπροσωπεύουν μια διαφορά μεγαλύτερη από 90,61. Με λίγη σκέψη θα συνειδητοποιήσετε ότι πρέπει να σχεδιάσουμε ράβδους που να έχουν μήκος LSD / 2, πάνω και κάτω από κάθε μέσο. Ας το δοκιμάσουμε αυτό στο παράδειγμά μας:

```
lsd<-qt(0.975,10)*sqrt(2*4961/6)
lsdbars<-rep(lsd,5)/2
error.bars(ybar,lsdbars,labels)
```



Τώρα μπορούμε να ερμηνεύσουμε τις σημαντικές διαφορές οπτικά. Η βιομάζα ελέγχου είναι σημαντικά χαμηλότερη από οποιαδήποτε από τις τέσσερις μεταχειρίσεις, αλλά καμία από τις τέσσερις μεταχειρίσεις δεν είναι σημαντικά διαφορετική από οποιαδήποτε άλλη. Η στατιστική ανάλυση αυτής της αντίθεσης εξηγήθηκε λεπτομερώς στο Κεφάλαιο 9. Δυστυχώς, οι περισσότεροι συντάκτες επιμένουν στις ράβδους σφάλματος ενός τυπικού σφάλματος. Είναι αλήθεια ότι υπάρχουν περίπλοκα ζητήματα σχετιζόμενα με τις ράβδους LSD (αν μη τι άλλο το επίμαχο ερώτημα των πολλαπλών συγκρίσεων, βλ. σελ. 483), αλλά τουλάχιστον αυτές κάνουν ό, τι προτίθετο να γίνει από το γράφημα σφάλματος (δηλαδή, επικαλυπτόμενες ράβδοι σημαίνουν μη-σπουδαιότητα και μη επικαλυπτόμενες ράβδοι καταδεικνύουν σπουδαιότητα). Ούτε τα τυπικά σφάλματα ούτε τα διαστήματα εμπιστοσύνης μπορούν να το πουν αυτό. Μια καλύτερη επιλογή θα μπορούσε να είναι να χρησιμοποιήσουμε γραφήματα *box-and-whisker* με την επιλογή *notch=T* να καταδεικνύει σπουδαιότητα (βλ. σελ. 159).

Παραγοντικά Πειράματα

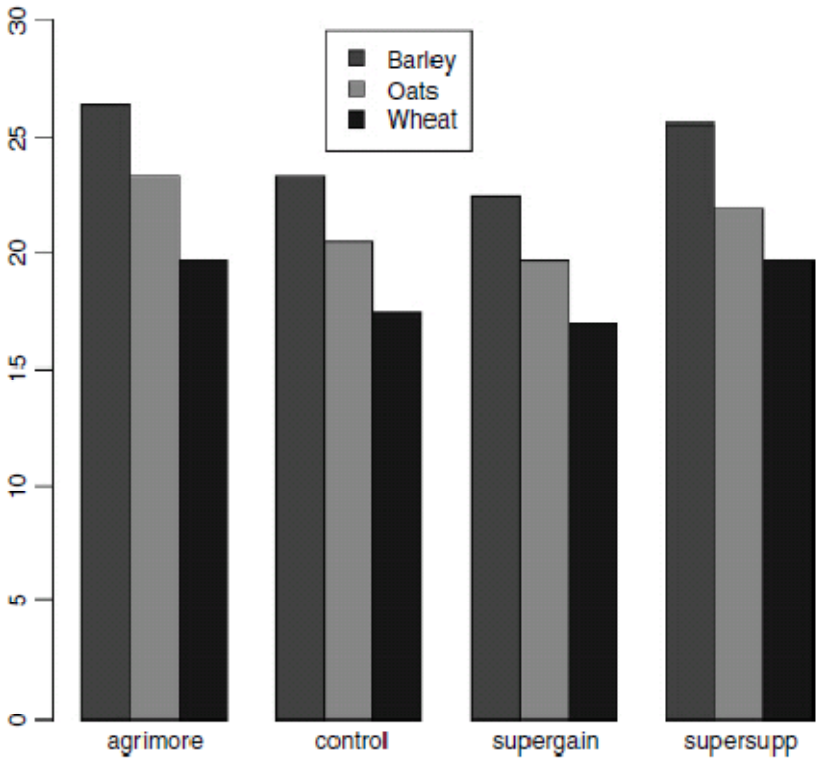
Ένα παραγοντικό πείραμα έχει δύο ή περισσότερους παράγοντες, καθένας με δύο ή περισσότερα επίπεδα, και επανάληψη για κάθε συνδυασμό των επιπέδων των παραγόντων. Αυτό σημαίνει ότι μπορούμε να διερευνήσουμε στατιστικές αλληλεπιδράσεις, στις οποίες η απόκριση σε ένα παράγοντα εξαρτάται από το επίπεδο κάποιου άλλου παράγοντα. Το παράδειγμά μας προέρχεται από μια δοκιμαστική κλίμακα γεωργικής εκμετάλλευσης της διατροφής των ζώων. Υπάρχουν δύο παράγοντες: η διατροφή και το συμπλήρωμα. Η διατροφή είναι ένας παράγοντας με τρία επίπεδα: το

κριθάρι, τη βρώμη και το σιτάρι. Το συμπλήρωμα είναι ένας παράγοντας με τέσσερα επίπεδα: agrimore, control, supergain και supersupp. Η μεταβλητή απόκρισης είναι η αύξηση του σωματικού βάρους μετά από 6 εβδομάδες.

```
weights<-read.table("c:\\temp\\growth.txt",header=T)
attach(weights)
```

Ο έλεγχος των δεδομένων διεξάγεται χρησιμοποιώντας το *barplot* (προσέξτε τη χρήση της *beside = T* για να πάρετε τις ράβδους σε παρακείμενες ομάδες και όχι σε κάθετες στοίβες):

```
barplot(tapply(gain,list(diet,supplement),mean),
                beside=T,ylim=c(0,30),col=rainbow(3))
```



Σημειώστε ότι ο δεύτερος παράγοντας στη λίστα (συμπλήρωμα) εμφανίζεται ως ομάδες ράβδων από τα αριστερά προς τα δεξιά σε αλφαβητική σειρά ανά επίπεδο παράγοντα, από το agrimore στο supersupp. Ο δεύτερος παράγοντας (διατροφή) εμφανίζεται σε τρία επίπεδα σε κάθε ομάδα ράβδων: κόκκινο = κριθάρι, πράσινο =βρώμη , μπλε =σιτάρι, πάλι κατά αλφαβητική σειρά ανά επίπεδο παράγοντα. Θα πρέπει να προσθέσουμε έναν μέσο για να εξηγήσουμε τα επίπεδα της διατροφής. Χρησιμοποιήστε την εντολή *locator(1)* για να βρείτε τις συντεταγμένες για την επάνω αριστερή γωνία του παραθύρου γύρω από τη

λεζάντα. Θα πρέπει να αυξήσετε τη δεδομένη κλίμακα επί του άξονα y για να κάνετε αρκετό χώρο για το κουτί της λεζάντας.

```
labs<-c("Barley","Oats","Wheat")
legend(locator(1),labs,fill=rainbow(3))
```

Επιθεωρούμε τις μέσες τιμές χρησιμοποιώντας ως συνήθως την εντολή *tapply*:

```
tapply(gain,list(diet,supplement),mean)
```

	agrimore	control	supergain	supersupp
barley	26.34848	23.29665	22.46612	25.57530
oats	23.29838	20.49366	19.66300	21.86023
wheat	19.63907	17.40552	17.01243	19.66834

Τώρα χρησιμοποιούμε aov ή lm για να προσαρμόσουμε μια παραγοντική ανάλυση διακύμανσης (η επιλογή αυτή επηρεάζει το αν παίρνουμε ένα πίνακα ANOVA ή μια λίστα από εκτιμήσεις παραμέτρων, όπως το προεπιλεγμένο αποτέλεσμα από την *summary*). Εκτιμούμε τις παραμέτρους για τις κύριες επιδράσεις κάθε επιπέδου διατροφής και συμπληρώματος, καθώς και τους - σχετιζόμενους με την αλληλεπίδραση μεταξύ διατροφής και συμπληρώματος - όρους. Οι βαθμοί ελευθερίας αλληλεπίδρασης είναι το προϊόν των βαθμών ελευθερίας των συστατικών όρων (3-1)x(4-1)=6. Το μοντέλο είναι το

`gain~diet + supplement + diet:supplement`, αλλά αυτό μπορεί να απλοποιηθεί χρησιμοποιώντας το συμβολισμό του αστερίσκου όπως παρακάτω:

```
model<-aov(gain~diet*supplement)
summary(model)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
diet	2	287.171	143.586	83.5201	2.998e-14	***
supplement	3	91.881	30.627	17.8150	2.952e-07	**
diet:supplement	6	3.406	0.568	0.3302	0.9166	
Residuals	36	61.890	1.719			

Ο πίνακας ANOVA δείχνει ότι δεν υπάρχει ένδειξη οποιασδήποτε αλληλεπίδρασης μεταξύ των δύο ερμηνευτικών μεταβλητών (p = 0,9166). Προφανώς οι επιδράσεις της διατροφής και του συμπληρώματος είναι αθροιστικές. Το μειονέκτημα του πίνακα ANOVA είναι ότι δεν μας δείχνει τα μεγέθη επίδρασης και δεν μας επιτρέπει να προσδιορίσουμε το πόσο πολλά επίπεδα από κάθε ένα από τους δυο παράγοντες είναι σημαντικά διαφορετικά. Ως προκαταρκτικό βήμα για την απλοποίηση του μοντέλου, η *summary.lm* είναι συχνά πιο χρήσιμη από ό, τι η *summary.aov:*

```
summary.aov:
summary.lm(model)

Coefficients:
                              Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)                    26.3485      0.6556   40.191   < 2e-16  ***
dietoats                       -3.0501      0.9271   -3.290  0.002248  **
dietwheat                      -6.7094      0.9271   -7.237  1.61e-08  ***
supplementcontrol              -3.0518      0.9271   -3.292  0.002237  **
supplementsupergain            -3.8824      0.9271   -4.187  0.000174  ***
supplementsupersupp            -0.7732      0.9271   -0.834  0.409816
dietoats:supplementcontrol      0.2471      1.3112    0.188  0.851571
dietwheat:supplementcontrol     0.8183      1.3112    0.624  0.536512
dietoats:supplementsupergain    0.2470      1.3112    0.188  0.851652
dietwheat:supplementsupergain   1.2557      1.3112    0.958  0.344601
dietoats:supplementsupersupp   -0.6650      1.3112   -0.507  0.615135
dietwheat:supplementsupersupp   0.8024      1.3112    0.612  0.544381

Residual standard error: 1.311 on 36 degrees of freedom
Multiple R-Squared: 0.8607, Adjusted R-squared: 0.8182
F-statistic: 20.22 on 11 and 36 DF,  p-value: 3.295e-012
```

Αυτό είναι ένα μάλλον πολύπλοκο μοντέλο, επειδή περιλαμβάνει 12 εκτιμώμενες παραμέτρους (ο αριθμός των γραμμών στον πίνακα): έξι κύριες επιδράσεις και έξι αλληλεπιδράσεις. Το αποτέλεσμα τονίζει εκ νέου ότι κανένας από τους όρους αλληλεπίδρασης δεν είναι σημαντικός, αλλά υποδηλώνει ότι το ελάχιστα επαρκές μοντέλο θα απαιτήσει πέντε παραμέτρους: ένα σημείο τομής, μια διαφορά που οφείλεται στη βρώμη, μια διαφορά που οφείλεται στο σιτάρι, μια διαφορά που οφείλεται στο control και μία διαφορά λόγω του supergain (αυτές είναι οι πέντε σειρές με αστερίσκους σπουδαιότητας). Αυτό εφιστά την προσοχή στο κύριο μειονέκτημα της χρήσης των προεπιλεγμένων αντίθετων μεταχειρίσεων. Αν κοιτάξετε προσεκτικά τον πίνακα, θα δείτε ότι η επίδραση μεγέθους των δύο από τα συμπληρώματα, το control και το supergain, δεν είναι μεταξύ τους σημαντικά διαφορετικά. Χρειάζεστε πολύ δουλειά για να κάνετε γρήγορα t-tests στο κεφάλι σας. Αγνοώντας τις ενδείξεις (επειδή αυτές είναι αρνητικές και για τα δύο), έχουμε 3,05 έναντι 3,88, μια διαφορά ίση με 0,83. Αλλά κοιτάξτε τα σχετιζόμενα τυπικά σφάλματα (και τα δυο είναι 0,927). Η διαφορά δηλαδή, είναι μικρότερη από 1 τυπικό σφάλμα μιας διαφοράς μεταξύ δύο μέσων. Για σπουδαιότητα, θα χρειαζόμαστε περίπου 2 τυπικά σφάλματα (θυμηθείτε τον εμπειρικό κανόνα, κατά τον οποίο όταν το $t \geq 2$, είναι σημαντικό, βλ. σελ. 228). Οι σειρές πρωταγωνιστούσαν στη στήλη της σπουδαιότητας, επειδή οι αντίθετες μεταχειρίσεις συγκρίνουν όλες τις κύριες επιδράσεις στις γραμμές με το σημείο τομής. Όταν, όπως εδώ, αρκετά επίπεδα των παραγόντων είναι διαφορετικά από το σημείο τομής, αλλά όχι διαφορετικά μεταξύ τους, επισημαίνονται όλα με αστερίσκους σπουδαιότητας. Αυτό σημαίνει ότι δεν μπορείτε να μετρήσετε τον αριθμό των γραμμών με αστερίσκους προκειμένου να προσδιορίσετε τον αριθμό των σημαντικά διαφορετικών επιπέδων παράγοντα.

Αρχικά απλοποιούμε το μοντέλο παραλείποντας τους όρους αλληλεπίδρασης:

```
model<-aov(gain~diet+supplement)
summary.lm(model)
```

```
Coefficients:
                   Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)         26.1230      0.4408   59.258   < 2e-16  ***
dietoats            -3.0928      0.4408   -7.016  1.38e-08  ***
dietwheat           -5.9903      0.4408  -13.589   < 2e-16  ***
supplementcontrol   -2.6967      0.5090   -5.298  4.03e-06  ***
supplementsupergain -3.3815      0.5090   -6.643  4.72e-08  ***
supplementsupersupp -0.7274      0.5090   -1.429     0.160
```

Είναι σαφές ότι πρέπει να διατηρήσουμε και τα τρία επίπεδα διατροφής (η βρώμη διαφέρει από το σιτάρι κατά 5,99- 3,09 = 2,90 με ένα τυπικό σφάλμα ίσο με 0,44). Ωστόσο, δεν είναι σαφές ότι χρειαζόμαστε τέσσερα επίπεδα συμπληρώματος. Το Supersupp δεν είναι προφανώς διαφορετικό από το agrimore (0,727 με τυπικό σφάλμα ίσο με 0,509). Ούτε το supergain είναι προφανώς διαφορετικό από τον χωρίς συμπλήρωμα έλεγχο των ζώων (3,38-2,70 = 0.68). Θα πρέπει να δοκιμάσουμε έναν νέο δύο επιπέδων παράγοντα για αντικαταστήσουμε τον παράγοντα συμπληρώματος των τεσσάρων επιπέδων, και να δούμε αν αυτό μειώνει σημαντικά την ερμηνευτική δύναμη του μοντέλου. Το Agrimore και το supersupp καταγράφονται ως τα καλύτερα ενώ ο έλεγχος και το supergain ως τα χειρότερα:

```
supp2<-factor(supplement)
levels(supp2)
```

```
[1] "agrimore" "control" "supergain" "supersupp"
```

```
levels(supp2)[c(1,4)]<-"best"
levels(supp2)[c(2,3)]<-"worst"
levels(supp2)
```

```
[1] "best"  "worst"
```

Τώρα μπορούμε να συγκρίνουμε τα δύο μοντέλα:

```
model2<-aov(gain~diet+supp2)
anova(model,model2)
```

```
Analysis of Variance Table

Model 1: gain ~ diet + supplement
Model 2: gain ~ diet + supp2
  Res.Df    RSS  Df  Sum of Sq      F  Pr((>)F)
1     42 65.296
2     44 71.284  -2    -5.988 1.9257    0.1584
```

Το απλούστερο δεύτερο μοντέλο, έχει εξοικονομήσει δύο βαθμούς ελευθερίας και δεν είναι πολύ χειρότερο από ό, τι το πιο σύνθετο μοντέλο (p = 0,158). Αυτό είναι το ελάχιστα

επαρκές μοντέλο: όλες οι παράμετροι είναι σημαντικά διαφορετικές από το μηδέν και σημαντικά διαφορετικές μεταξύ τους:

```
summary.lm(model2)
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   25.7593      0.3674   70.106    < 2e-16  ***
dietoats      -3.0928      0.4500   -6.873  1.76e-08  ***
dietwheat     -5.9903      0.4500  -13.311    < 2e-16  ***
supp2worst    -2.6754      0.3674   -7.281  4.43e-09  ***

Residual standard error: 1.273 on 44 degrees of freedom
Multiple R-Squared: 0.8396, Adjusted R-squared: 0.8286
F-statistic: 76.76 on 3 and 44 DF, p-value: 0
```

Η απλοποίηση του μοντέλου έχει μετατρέψει το αρχικό μας μοντέλο των 12 παραμέτρων σε ένα μοντέλο τεσσάρων παραμέτρων.

Ψευδοεπανάληψη: Εγκιβωτισμένα σχέδια και διαιρεμένα γραφήματα

Οι προσαρμοσμένες στο μοντέλο συναρτήσεις AOV και lmer έχουν τη δυνατότητα να αντιμετωπίζουν πολύπλοκες δομές σφαλμάτων, και είναι σημαντικό να μπορείτε να τα αναγνωρίζετε, και ως εκ τούτου να αποφεύγετε τις παγίδες της ψευδοεπανάληψης. Υπάρχουν δύο γενικές περιπτώσεις:

• Εγκιβωτισμένη δειγματοληψία, όταν επαναλαμβανόμενες μετρήσεις λαμβάνονται από το ίδιο άτομο, ή μελέτες παρατήρησης συντελούν σε πολλές διαφορετικές χωρικές κλίμακες (ως επί το πλείστον τυχαίες επιδράσεις).

• Ανάλυση διαιρεμένης απεικόνισης, όταν σχεδιασμένα πειράματα εφαρμόζουν διαφορετικές μεταχειρίσεις σε γραφήματα διαφορετικών μεγεθών (ως επί το πλείστον σταθερές επιδράσεις).

Πειράματα Split-plot

Σε ένα πείραμα split-plot οι διαφορετικές μεταχειρίσεις εφαρμόζονται σε γραφήματα διαφορετικών μεγεθών. Κάθε διαφορετικό μέγεθος γραφήματος συνδέεται με τη δική του διακύμανση σφάλματος, έτσι ώστε αντί να έχουμε μια διακύμανση σφάλματος (όπως σε όλους τους πίνακες ANOVA που έχουμε δει μέχρι τώρα), να έχουμε τόσους όρους σφάλματος όσα και τα διαφορετικά μεγέθη γραφημάτων. Η ανάλυση παρουσιάζεται ως μια σειρά από πίνακες ANOVA, ένα για κάθε μέγεθος γραφήματος, ιεραρχημένα από το μεγαλύτερο σε μέγεθος γράφημα με τη χαμηλότερη επανάληψη στην κορυφή, έως και το μικρότερο σε μέγεθος γράφημα με τη μεγαλύτερη επανάληψη στον πυθμένα.
Το ακόλουθο παράδειγμα αναφέρεται σε ένα σχεδιασμένο πείραμα πεδίου πάνω στην απόδοση των καλλιεργειών με τρεις μεταχειρίσεις: άρδευση (με δύο επίπεδα, αρδευόμενη ή μη), πυκνότητα σποράς (με τρία επίπεδα, χαμηλή, μέση και υψηλή), και λίπανση (με τρία επίπεδα, χαμηλή, μέση και υψηλή).

```
yields<-read.table("c:\\temp\\splityield.txt",header=T)
attach(yields)
names(yields)
```

```
[1] "yield"  "block"  "irrigation"  "density"  "fertilizer"
```

Τα μεγαλύτερα χωράφια ήταν τα τέσσερα ολόκληρα κομμάτια (μπλοκ), καθένα από τα οποία χωρίστηκε στο μισό, και η άρδευση κατανεμήθηκε τυχαία στο ένα μισό του χωραφιού. Κάθε αρδευόμενο χωράφι χωρίστηκε στα τρία και μία από τις τρεις διαφορετικές πυκνότητες σποράς (χαμηλή, μεσαία ή υψηλή) τοποθετήθηκε τυχαία (ανεξάρτητα για κάθε επίπεδο άρδευσης και για κάθε κομμάτι). Τέλος, κάθε κομμάτι πυκνότητας διαιρέθηκε στα τρία, και μία από τις τρεις μεταχειρίσεις θρεπτικών λιπασμάτων (N, P, ή N και P μαζί) διατέθηκε τυχαία. Ο τύπος του μοντέλου ορίζεται ως παραγοντικός, χρησιμοποιώντας τη σημειογραφία των αστερίσκων. Η δομή σφάλματος ορίζεται στον όρο *Error*, με τα μεγέθη χωραφιών να παρατίθενται από αριστερά προς τα δεξιά, από το μεγαλύτερο προς το μικρότερο, και με κάθε μεταβλητή χωρισμένη από τον φορέα /. Σημειώστε ότι το μικρότερο σε μέγεθος κομμάτι, το λίπασμα, δεν χρειάζεται να εμφανίζεται στον όρο *Error*:

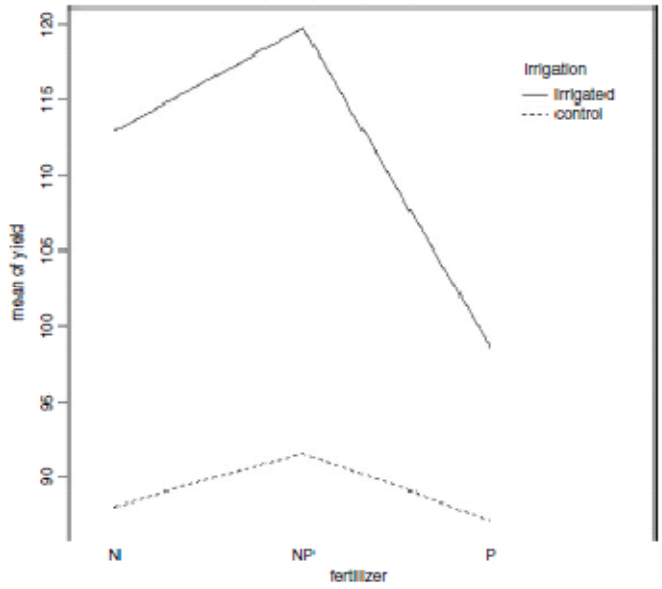
```
model<-aov(yield~irrigation*density*fertilizer+Error(block/irrigation/density))
summary(model)
```

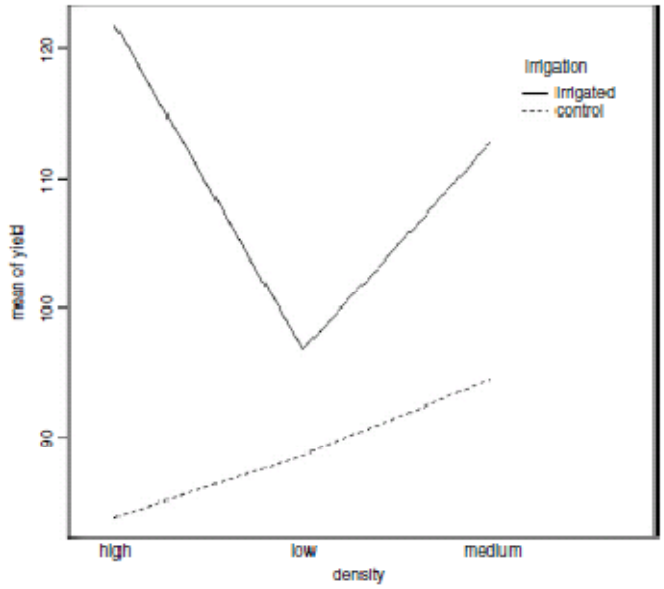
```
Error: block
          Df  Sum Sq Mean Sq F value Pr(>F)
Residuals  3 194.444  64.815

Error: block:irrigation
           Df  Sum Sq Mean Sq F value  Pr(>F)
irrigation  1 8277.6  8277.6  17.590 0.02473 *
Residuals   3 1411.8   470.6

Error: block:irrigation:density
                  Df  Sum Sq Mean Sq F value  Pr(>F)
density            2 1758.36  879.18  3.7842 0.05318 .
irrigation:density 2 2747.03 1373.51  5.9119 0.01633 *
Residuals         12 2787.94  232.33

Error: Within
                             Df  Sum Sq Mean Sq F value    Pr(>F)
fertilizer                    2 1977.44  988.72 11.4493 0.0001418 ***
irrigation:fertilizer         2  953.44  476.72  5.5204 0.0081078 **
density:fertilizer            4  304.89   76.22  0.8826 0.4840526
irrigation:density:fertilizer 4  234.72   58.68  0.6795 0.6106672
Residuals                    36 3108.83   86.36
```

Εδώ βλέπετε τους τέσσερις πίνακες ANOVA, έναν για κάθε μέγεθος χωραφιού: τα blocks είναι τα μεγαλύτερα κομμάτια, τα μισά κομμάτια παίρνουν από μια μεταχείριση άρδευσης, το ένα τρίτο του κάθε μισού κομματιού παίρνει από μια μεταχείριση πυκνότητας σποράς, και το ένα τρίτο κάθε κομματιού διαφορετικής μεταχείρισης πυκνότητας σποράς παίρνει μια μεταχείριση λιπάσματος. Σημειώστε ότι η μη σημαντική κύρια επίδραση για την πυκνότητα (p= 0,053) δεν σημαίνει ότι η πυκνότητα είναι ασήμαντη, γιατί η πυκνότητα εμφανίζεται να έχει μια σημαντική αλληλεπίδραση με την άρδευση (οι όροι πυκνότητας,

ακυρώνουν, όταν αντισταθμιστούν, πάνω από δύο μεταχειρίσεις άρδευσης, βλέπε παρακάτω). Ο καλύτερος τρόπος για να κατανοήσουμε τους δύο σημαντικούς όρους αλληλεπίδρασης είναι να τους σχεδιάσουμε χρησιμοποιώντας ένα γράφημα σαν αυτό:

```
interaction.plot(fertilizer,irrigation,yield)
```





Η άρδευση αυξάνει την απόδοση αναλογικά περισσότερο στα N-λιπασμένα κομμάτια από ό, τι στα P-λιπασμένα. Η αλληλεπίδραση μεταξύ πυκνότητας και άρδευσης είναι πιο περίπλοκη:

```
interaction.plot(density,irrigation,yield)
```

Στα αρδευόμενα χωράφια, η απόδοση είναι ελάχιστη στα χαμηλής πυκνότητας, αλλά στα χωράφια ελέγχου απόδοσης είναι ελάχιστη στα υψηλής πυκνότητας. Εναλλακτικά, θα μπορούσατε να χρησιμοποιήσετε το πακέτο *effects* που παίρνει ένα μοντέλο αντικειμένου (ένα γραμμικό μοντέλο ή ένα γενικευμένο γραμμικό μοντέλο) και προσφέρει ελκυστικά γραφήματα συγκεκριμένων επιδράσεων αλληλεπίδρασης (σελ. 178).

Οι τιμές που λείπουν σε ένα γράφημα split-plot

Όταν υπάρχουν τιμές που λείπουν, τότε οι παράγοντες έχουν επιπτώσεις σε περισσότερα από ένα στρώματα και η ίδια κύρια επίδραση καταλήγει σε περισσότερους από έναν πίνακες ANOVA. Ας υποθέσουμε ότι έλειπε η τιμή της 69ης απόδοσης :

```
yield[69]<-NA
```

Τώρα ο συγκεντρωτικός πίνακας δείχνει πολύ διαφορετικός:

```
model<-aov(yield~irrigation*density*fertilizer+Error(block/irrigation/density))
summary(model)

Error: block
           Df  Sum Sq Mean Sq F value Pr(>F)
irrigation  1   0.075   0.075   9e-04 0.9788
Residuals   2 167.704  83.852

Error: block:irrigation
           Df  Sum Sq Mean Sq F value  Pr(>F)
irrigation  1 7829.9  7829.9 21.9075 0.04274 *
density     1  564.4   564.4  1.5792 0.33576
Residuals   2  714.8   357.4

Error: block:irrigation:density
                   Df  Sum Sq Mean Sq F value  Pr(>F)
density             2 1696.47  848.24  3.4044 0.07066 .
fertilizer          1    0.01    0.01 2.774e-05 0.99589
irrigation:density  2 2786.75 1393.37  5.5924 0.02110 *
Residuals          11 2740.72  249.16

Error: Within
                           Df  Sum Sq Mean Sq F value    Pr(>F)
fertilizer                  2 1959.36  979.68 11.1171 0.0001829 ***
irrigation:fertilizer       2  993.59  496.79  5.6375 0.0075447 **
density:fertilizer          4  273.56   68.39  0.7761 0.5482571
irrigation:density:fertilizer 4 244.49  61.12  0.6936 0.6014280
Residuals                  35 3084.33  88.12
```

Παρατηρήστε ότι με μία μόνο τιμή να λείπει, κάθε κύρια επίδραση εμφανίζεται σε δύο πίνακες (όχι σε ένα όπως παραπάνω). Συνιστάται, στις περιπτώσεις κατά τις οποίες σε ένα πείραμα split-plot λείπουν τιμές, να χρησιμοποιείτε lmer ή LME αντί για aov, προκειμένου για να προσαρμόσετε το μοντέλο.

Τυχαίες επιδράσεις και εγκιβωτισμένα σχέδια

Τα μοντέλα μικτών επιδράσεων ονομάζονται έτσι επειδή σε αυτά οι ερμηνευτικές μεταβλητές είναι ένα μίγμα σταθερών και τυχαίων επιδράσεων:

•Οι σταθερές επιδράσεις επηρεάζουν μόνο τη μέση τιμή του y.
• Οι τυχαίες επιδράσεις επηρεάζουν μόνο τη διακύμανση του y.

Μία τυχαία επίδραση πρέπει να θεωρηθεί ότι προέρχεται από έναν πληθυσμό αποτελεσμάτων: η ύπαρξη αυτού του πληθυσμού είναι μια επιπλέον υπόθεση. Μιλάμε για πρόβλεψη τυχαίων αποτελεσμάτων, αντί για εκτίμηση: εκτιμούμε σταθερές επιδράσεις από τα δεδομένα, αλλά έχουμε την πρόθεση να κάνουμε προβλέψεις σχετικά με τον πληθυσμό από τον οποίο δειγματοποιήθηκαν τα τυχαία μας αποτελέσματα. Οι σταθερές επιδράσεις είναι άγνωστες σταθερές που θα πρέπει να εκτιμηθούν από τα δεδομένα και διέπουν τη δομή της διακύμανσης – συνδιακύμανσης της μεταβλητής απόκρισης. Επίσης είναι συχνά πειραματικές μεταχειρίσεις που εφαρμόστηκαν υπό την καθοδήγησή μας, και τα τυχαία αποτελέσματα είναι είτε κατηγορικές είτε συνεχείς μεταβλητές που διακρίνονται από το γεγονός ότι τυπικά δεν ενδιαφερόμαστε για τις τιμές των παραμέτρων, αλλά μόνο για τη διακύμανση την οποία αυτές εξηγούν.

Μία ή περισσότερες από τις ερμηνευτικές μεταβλητές αντιπροσωπεύουν χρονική ή χωρική ομαδοποίηση. Τα τυχαία αποτελέσματα που προέρχονται από την ίδια ομάδα θα συσχετίζονται. Το γεγονός αυτό παραβαίνει κάποια από τις θεμελιώδεις παραδοχές των προτύπων στατιστικών μοντέλων: την ανεξαρτησία των σφαλμάτων. Τα μοντέλα μικτών επιδράσεων τακτοποιούν αυτή τη μη-ανεξαρτησία των σφαλμάτων μέσω της μοντελοποίησης της δομής της διακύμανσης που εισήχθη από την ομαδοποίηση των δεδομένων. Ένα σημαντικό πλεονέκτημα των μοντέλων τυχαίων επιδράσεων είναι ότι προσφέρουν λιτότητα σχετικά με τον αριθμό των βαθμών ελευθερίας που χρησιμοποιούνται από τα επίπεδα του παράγοντα. Αντί να εκτιμά μια μέση τιμή για κάθε επίπεδο παράγοντα, το μοντέλο τυχαίων επιδράσεων εκτιμά την κατανομή των μέσων (συνήθως ως την τυπική απόκλιση των διαφορών των μέσων τιμών για κάθε επίπεδο παράγοντα γύρω από μια συνολική μέση τιμή). Τα μοντέλα μικτών αποτελεσμάτων είναι ιδιαίτερα χρήσιμα σε περιπτώσεις όπου υπάρχει χρονική ψευδοεπανάληψη (επαναλαμβανόμενες μετρήσεις) ή /και χωρική ψευδοεπανάληψη (π.χ. εγκιβωτισμένα σχέδια ή πειράματα split-plot). Τα μοντέλα αυτά μπορούν να επιτρέπουν την:

• χωρική αυτοσυσχέτιση μεταξύ των γειτονικών επαναλήψεων.
• χρονική αυτοσυσχέτιση σε επαναλαμβανόμενες μετρήσεις των ίδιων ατόμων.
• διαφορές στη μέση απόκριση μεταξύ των κομματιών στα οποία είναι χωρισμένα τα χωράφια, σε ένα πείραμα με χωράφια.
• διαφορές μεταξύ των ατόμων σε μια ιατρική μελέτη που περιλαμβάνει επαναλαμβανόμενες μετρήσεις..

Το θέμα είναι ότι πραγματικά δεν θέλουμε να σπαταλάμε πολύτιμους βαθμούς ελευθερίας στην εκτίμηση των παραμέτρων για κάθε ένα από τα ξεχωριστά επίπεδα των τυχαίων κατηγορικών μεταβλητών. Από την άλλη πλευρά, θέλουμε να χρησιμοποιούμε όλες τις μετρήσεις που έχουμε λάβει, αλλά λόγω της ψευδοεπανάληψης θέλουμε να λαμβάνουμε υπόψη μας τα παρακάτω δύο:
• τη δομή συσχέτισης που χρησιμοποιείται για να μοντελοποιήσουμε τη συσχέτιση στο εσωτερικό των ομάδων που σχετίζεται με χρονικές και χωρικές εξαρτήσεις, και

• τη συνάρτηση διακύμανσης, που χρησιμοποιείται για την μοντελοποίηση της μη σταθερής διακύμανσης των σφαλμάτων στο εσωτερικό των ομάδων χρησιμοποιώντας βάρη.

Σταθερά ή τυχαία αποτελέσματα;

Είναι δύσκολο χωρίς μεγάλη εμπειρία να γνωρίζουμε πότε να χρησιμοποιήσουμε κατηγορικές επεξηγηματικές μεταβλητές ως σταθερές επιδράσεις και πότε ως τυχαίες. Μερικές κατευθυντήριες γραμμές δίνονται παρακάτω.
• Ενδιαφέρομαι για την επίδραση μεγέθους; Το ναι σημαίνει σταθερά αποτελέσματα.
• Είναι λογικό να υποθέσουμε ότι τα επίπεδα του παράγοντα προέρχονται από έναν πληθυσμό επιπέδων; Το ναι σημαίνει τυχαία αποτελέσματα.
• Υπάρχουν αρκετά επίπεδα του παράγοντα στο πλαίσιο δεδομένων στα οποία να μπορεί να βασιστεί μια εκτίμηση της διακύμανσης του πληθυσμού των αποτελεσμάτων; Το όχι σημαίνει σταθερές επιδράσεις.
• Είναι τα επίπεδα του παράγοντα πληροφοριακά; Το ναι σημαίνει σταθερές επιδράσεις
• Είναι τα επίπεδα του παράγοντα μόνο αριθμητικές τιμές; Το ναι σημαίνει τυχαίες επιδράσεις.
• Ενδιαφέρομαι ως επί το πλείστον να εξάγω συμπεράσματα σχετικά με την κατανομή των επιδράσεων, με βάση το τυχαίο δείγμα των επιδράσεων που εκπροσωπούνται στο πλαίσιο δεδομένων; Το ναι σημαίνει τυχαίες επιδράσεις.
• Υπάρχει ιεραρχική δομή; Το ναι σημαίνει ότι θα πρέπει να αναρωτηθούμε αν τα δεδομένα είναι πειραματικά ή παρατηρήσεις.
• Είναι ένα ιεραρχικό πείραμα, όπου τα επίπεδα του παράγοντα είναι πειραματικοί χειρισμοί; Το ναι σημαίνει σταθερές επιδράσεις σε ένα σχέδιο split-plot (βλ. σελ.. 469)
• Είναι μια ιεραρχική μελέτη παρατήρησης; Το ναι σημαίνει τυχαίες επιδράσεις, ίσως σε μια ανάλυση συνιστωσών διακύμανσης (βλέπε σελ.. 475).
• Όταν το μοντέλο σας περιέχει τόσο σταθερές όσο και τυχαίες επιδράσεις χρησιμοποιήστε μοντέλα μικτών επιδράσεων.
• Εάν η δομή του μοντέλου σας είναι γραμμική, χρησιμοποιήστε γραμμικά μικτά αποτελέσματα, *lmer*.
• Σε αντίθετη περίπτωση, καθορίστε την εξίσωση του μοντέλου και χρησιμοποιήστε μη γραμμικές μικτές επιδράσεις, *nlme*.

Απομακρύνοντας την ψευδοεπανάληψη

Η ακραία απάντηση στην ψευδοεπανάληψη σε ένα σύνολο δεδομένων είναι απλά η εξάλειψή της. Η χωρική ψευδοεπανάληψη μπορεί να αντισταθμιστεί και η χρονική ψευδοεπανάληψη μπορεί να αντιμετωπιστεί διεξάγοντας χωριστούς πίνακες ANOVA, έναν κάθε φορά. Αυτή η προσέγγιση έχει δύο βασικές αδυναμίες:
• Δεν μπορεί να θέσει ερωτήσεις σχετικά με τις επιδράσεις των μεταχειρίσεων που σχετίζονται με τη μακροχρόνια ανάπτυξη των μέσων προφίλ απόκρισης (π.χ. διαφορές στους ρυθμούς ανάπτυξης μεταξύ διαδοχικών μετρήσεων).

• Τα συμπεράσματα που εξάγονται από κάθε μία από τις ξεχωριστές αναλύσεις δεν είναι ανεξάρτητα μεταξύ τους, ενώ δεν είναι πάντα σαφές το πώς θα πρέπει να συνδυάζονται.

Ανάλυση διαχρονικών δεδομένων

Το βασικό χαρακτηριστικό των διαχρονικών δεδομένων είναι ότι τα ίδια άτομα μετρούνται επανειλημμένως μέσα στο χρόνο. Αυτό θα μπορούσε να αποτελέσει χρονική ψευδοεπανάληψη εάν τα δεδομένα χρησιμοποιήθηκαν άκριτα στην παλινδρόμηση ή στην ANOVA. Το σύνολο των παρατηρήσεων σε ένα επιμέρους θέμα θα τείνει να συσχετίζεται θετικά, και αυτή η συσχέτιση πρέπει να ληφθεί υπόψη κατά τη διεξαγωγή της ανάλυσης. Η εναλλακτική λύση είναι μια σύγχρονη μελέτη, με όλα τα δεδομένα να συλλέγονται σε μια συγκεκριμένη χρονική στιγμή, κατά την οποία κάθε άτομο συνεισφέρει με ένα μόνο σημείο δεδομένων. Το πλεονέκτημα των διαχρονικών μελετών έγκειται στο ότι είναι σε θέση να διαχωρίζουν τις ηλικιακές από τις ομαδικές επιδράσεις. Στις σύγχρονες μελέτες, αυτές οι επιδράσεις είναι άρρηκτα συνδεδεμένες μεταξύ τους. Αυτό είναι ιδιαίτερα σημαντικό όταν οι ηλικιακές διαφορές σημαίνουν ότι οι ομάδες που κατάγονται από διαφορετικές χρονικές στιγμές αντιμετωπίζουν διαφορετικές συνθήκες, έτσι ώστε τα άτομα της ίδιας ηλικίας σε διαφορετικές πληθυσμιακές ομάδες να αναμένονται να είναι διαφορετικά.

Υπάρχουν δύο ακραίες περιπτώσεις στις διαχρονικές μελέτες:
• Μερικές μόνο μετρήσεις σε ένα μεγάλο αριθμό ατόμων.
• Ένας μεγάλος αριθμός μετρήσεων σε μερικά μόνο άτομα.

Στην πρώτη περίπτωση είναι δύσκολο να προσαρμόσουμε ένα ακριβές μοντέλο για τις αλλαγές των ατόμων, αλλά οι επιδράσεις των μεταχειρίσεων είναι πιθανόν να ελεγχθούν αποτελεσματικά. Στη δεύτερη περίπτωση, είναι δυνατόν να έχουμε ένα ακριβές μοντέλο του τρόπου με τον οποίο τα άτομα αλλάζουν μέσα στο χρόνο, αλλά υπάρχει μικρότερη δύναμη για τη δοκιμή της σημασίας των επιπτώσεων των μεταχειρίσεων, ειδικά εάν η μεταβολή από άτομο σε άτομο είναι μεγάλη. Στην πρώτη περίπτωση, λιγότερη προσοχή θα πρέπει να δοθεί στην εκτίμηση της δομής συσχέτισης, ενώ στη δεύτερη περίπτωση η προσοχή θα εστιαστεί κυρίως στο μοντέλο διακύμανσης.

Οι στόχοι είναι οι εξής:
• να εκτιμηθεί η μέση χρονική πορεία μιας διαδικασίας.
• να χαρακτηριστεί ο βαθμός της ετερογένειας από άτομο σε άτομο στα πλαίσια της διαδικασίας.
• να προσδιοριστούν οι παράγοντες που σχετίζονται με τα δύο παραπάνω, συμπεριλαμβανομένων των πιθανών ομαδικών επιδράσεων.

Η απόκριση δεν είναι η ατομική μέτρηση, αλλά η σειρά των μετρήσεων σε ένα επιμέρους θέμα. Αυτό μας δίνει τη δυνατότητα να κάνουμε διαχωρισμό μεταξύ των ηλικιακών και των χρονικών επιδράσεων. Για λεπτομέρειες βλ. Diggle et al. (1994).

Ανάλυση προερχόμενης μεταβλητής

Η ιδέα εδώ είναι να απαλλαγούμε από την ψευδοεπανάληψη μειώνοντας τις επαναληπτικές μετρήσεις σε ένα σύνολο συνοπτικών στατιστικών στοιχείων, και στη συνέχεια να αναλύσουμε αυτά τα στοιχεία χρησιμοποιώντας πρότυπες στατιστικές παραμετρικές τεχνικές όπως η ANOVA ή η παλινδρόμηση. Η τεχνική αυτή είναι αδύναμη, όταν οι τιμές των ερμηνευτικών μεταβλητών αλλάζουν στο πέρασμα του χρόνου. Η ανάλυση των μεταβλητών έχει περισσότερο νόημα, όταν βασίζεται στις παραμέτρους των επιστημονικά ερμηνεύσιμων μη γραμμικών μοντέλων από κάθε χρονική ακολουθία. Ωστόσο, το θεωρητικά καλύτερο μοντέλο μπορεί να μην είναι το καλύτερο μοντέλο από στατιστικής άποψης.

Υπάρχουν τρεις ποιοτικά διαφορετικές πηγές τυχαίας μεταβολής:
• τυχαίες επιδράσεις, όπου οι πειραματικές μονάδες διαφέρουν (π.χ. γονότυπος, ιστορία, μέγεθος, κατάσταση φυσιολογίας) με αποτέλεσμα να υπάρχουν εγγενώς τόσο υψηλοί όσο και χαμηλοί ανταποκριτές.
• σειριακή συσχέτιση, όπου μπορεί να υπάρχει διαφορετική χρονικά στοχαστική μεταβολή μέσα σε μια μονάδα (π.χ. δυνάμεις της αγοράς, φυσιολογία, οικολογική διαδοχή, ασυλία), έτσι ώστε η συσχέτιση να εξαρτάται από τον χρόνο διαχωρισμού των ζευγών των μετρήσεων για το ίδιο άτομο, και να εξασθενεί με την πάροδο του χρόνου.
• σφάλμα μέτρησης, όπου η τεχνική ανάλυση ίσως εισάγει ένα στοιχείο συσχέτισης (π.χ. κοινός βιοπροσδιορισμός πυκνών δειγμάτων ή διαφορετικός βιοπροσδιορισμός υστερόχρονων δειγμάτων).

Ανάλυση συνιστωσών διακύμανσης

Για τις τυχαίες επιδράσεις συχνά ενδιαφερόμαστε περισσότερο για το ζήτημα του τι μέρος της μεταβολής της μεταβλητής απόκρισης μπορεί να αποδοθεί σε έναν δεδομένο παράγοντα, από ότι για την εκτίμηση των μέσων ή την αξιολόγηση της σπουδαιότητας των διαφορών μεταξύ των μέσων. Αυτή η διαδικασία ονομάζεται ανάλυση συνιστωσών διακύμανσης.
Ακολουθεί ένα κλασικό παράδειγμα ψευδοεπανάληψης (Snedecor Cochran, 1980):

```
rats<-read.table("c:\\temp\\rats.txt",header=T)
attach(rats)
names(rats)

[1] "Glycogen" "Treatment" "Rat" "Liver"
```

Τρεις πειραματικές μεταχειρίσεις χορηγήθηκαν σε αρουραίους και ως μεταβλητή απόκρισης αναλύθηκε η περιεκτικότητα σε γλυκογόνο των συκωτιών τους. Υπήρχαν δύο αρουραίοι ανά μεταχείριση, έτσι το συνολικό δείγμα ήταν n=3*2=6. Το δύσκολο κομμάτι ήταν ότι αφότου κάθε αρουραίος σκοτώθηκε, το συκώτι του κόπηκε σε τρία κομμάτια: ένα αριστερό κομμάτι, ένα κεντρικό κομμάτι και ένα δεξιό κομμάτι. Έτσι προκύπτουν έξι αρουραίοι κάθε ένας από τους οποίους παράγει τρία κομμάτια ήπατος, δημιουργώντας

ένα σύνολο 6*3 = 18 αριθμών. Τέλος, έγιναν δύο ξεχωριστά παρασκευάσματα από κάθε εμβρεγμένο κομμάτι συκωτιού, για να εκτιμηθεί το σφάλμα μέτρησης που σχετίζεται με τον αναλυτικό μηχανισμό. Σε αυτό το σημείο, υπάρχουν 2*18 = 36 συνολικά αριθμοί στο πλαίσιο δεδομένων. Τα επίπεδα των παραγόντων είναι αριθμοί, οπότε πριν ξεκινήσουμε πρέπει να δηλώσουμε τις ερμηνευτικές μεταβλητές ως κατηγορικές:

```
Treatment<-factor(Treatment)
Rat<-factor(Rat)
Liver<-factor(Liver)
```

Εδώ είναι η ανάλυση πραγματοποιημένη με λάθος τρόπο:

```
model<-aov(Glycogen~Treatment)
summary(model)
```

```
            Df   Sum Sq  Mean Sq  F value     Pr(>F)
Treatment    2  1557.56   778.78   14.498  3.031e-05 ***
Residuals   33  1772.67    53.72
```

Η μεταχείριση έχει πολύ σημαντική επίδραση στην περιεκτικότητα του συκωτιού σε γλυκογόνο $(p = 0.000\,03)$. Αυτό είναι λάθος! Έχουμε κάνει ένα κλασικό σφάλμα ψευδοεπανάληψης. Κοιτάξτε τη γραμμή σφάλματος στον πίνακα ANOVA: λέει ότι τα κατάλοιπα έχουν 33 βαθμούς ελευθερίας. Υπήρχαν όμως μόνο 6 αρουραίοι σε ολόκληρο το πείραμα, οπότε το σφάλμα d.f. πρέπει να είναι 6-1-2 = 3 (όχι 33)!

Εδώ είναι η ανάλυση της διακύμανσης πραγματοποιημένη σωστά, απομακρύνοντας την ψευδοεπανάληψη:

```
tt<-as.numeric(Treatment)
yv<-tapply(Glycogen,list(Treatment,Rat),mean)

tv<-tapply(tt,list(Treatment,Rat),mean)
model<-aov(as.vector(yv)~factor(as.vector(tv)))
summary(model)
```

```
                      Df    Sum Sq   Mean Sq  F value  Pr(>F)
factor(as.vector(tv))  2   259.593   129.796    2.929  0.1971
Residuals              3   132.944    44.315
```

Τώρα, οι βαθμοί ελευθερίας σφάλματος είναι σωστοί (d.f.= 3 και όχι 33), και η ερμηνεία είναι εντελώς διαφορετική: δεν υπάρχουν σημαντικές διαφορές στο γλυκογόνο του συκωτιού στο πλαίσιο των τριών πειραματικών μεταχειρίσεων (p = 0,1971).

Υπάρχουν δύο διαφορετικοί τρόποι για να γίνει σωστά η ανάλυση στην R: ANOVA με πολλαπλούς όρους σφάλματος (aov) ή γραμμικό μοντέλο μικτών επιδράσεων (lmer). Το πρόβλημα είναι ότι τα κομμάτια του ίδιου ήπατος είναι ψευδοεπαναλήψεις επειδή είναι

χωρικά συσχετισμένα (προέρχονται από τον ίδιο αρουραίο), άρα δεν είναι ανεξάρτητα, όπως απαιτείται για να είναι αληθινές επαναλήψεις. Ομοίως, τα δύο σκευάσματα από κάθε κομμάτι ήπατος είναι πολύ υψηλά συσχετισμένα (τα συκώτια είχαν εμβαπτιστεί πριν ληφθούν τα παρασκευάσματα, έτσι ουσιαστικά είναι το ίδιο δείγμα (σίγουρα όχι ανεξάρτητες επαναλήψεις των πειραματικών μεταχειρίσεων).

Εδώ είναι η σωστή ανάλυση χρησιμοποιώντας AOV με πολλαπλούς όρους σφάλματος. Στον όρο Error αρχίζουμε με τη μεγαλύτερη κλίμακα (μεταχείριση), έπειτα είναι οι υπό μεταχείριση αρουραίοι και έπειτα τα κομμάτια ήπατος αυτών των υπό μεταχείριση αρουραίων. Τέλος, υπήρχαν επαναλαμβανόμενες μετρήσεις (δύο παρασκευάσματα) φτιαγμένα για κάθε κομμάτι συκωτιού.

```
Error: Treatment
           Df  Sum Sq  Mean Sq
Treatment   2 1557.56  778.78
Error: Treatment:Rat
           Df  Sum Sq  Mean Sq  F value  Pr(>F)
Residuals   3  797.67  265.89
Error: Treatment:Rat:Liver
           Df  Sum Sq  Mean Sq  F value  Pr(>F)
Residuals  12   594.0     49.5
Error: Within
           Df  Sum Sq  Mean Sq  F value  Pr(>F)
Residuals  18  381.00   21.17
```

Μπορείτε να κάνετε τη σωστή- χωρίς ψευδοεπανάληψη- ανάλυση της διακύμανσης με αυτήν τη διαδικασία (Πλαίσιο 11.2).

Πλαίσιο 11.2 Αθροίσματα των τετραγώνων σε ιεραρχικά σχέδια

Το τέχνασμα για την κατανόηση αυτών των αθροισμάτων των τετραγώνων είναι να εκτιμήσετε ότι με ένθετες κατηγορικές ερμηνευτικές μεταβλητές (τυχαίες επιδράσεις), ο διορθωτικός συντελεστής, ο οποίος αφαιρείται από το άθροισμα των τετραγώνων των μερικών αθροισμάτων, δεν είναι ο συμβατικός $(\sum y)^2/kn.$. Αντ 'αυτού, ο διορθωτικός συντελεστής είναι το μη διορθωμένο άθροισμα των τετραγώνων των μερικών αθροισμάτων από το επίπεδο στην σειρά που βρίσκεται ακριβώς πάνω από το υπό ερώτηση επίπεδο. Αυτό είναι πολύ δύσκολο να εντοπιστεί χωρίς αρκετή πρακτική. Το ολικό άθροισμα των τετραγώνων, SSY, και το άθροισμα των τετραγώνων των μεταχειρίσεων, SSA, υπολογίζονται κατά τον συνήθη τρόπο (βλ. Πλαίσιο 11.1):

$$SSY = \sum y^2 - \frac{(\sum y)^2}{n},$$

$$SSA = \frac{\sum_{i=1}^{k} C_i^2}{n} - \frac{(\sum y)^2}{kn}.$$

Η ανάλυση είναι πιο εύκολο να κατανοηθεί στο πλαίσιο ενός παραδείγματος. Για τα δεδομένα των αρουραίων τα σύνολα μεταχείρισης βασίστηκαν σε 12 αριθμούς (δύο αρουραίοι, τρία κομμάτια ήπατος ανά αρουραίο και δύο σκευάσματα για κάθε κομμάτι ήπατος). Σε αυτή την περίπτωση, στον παραπάνω τύπο υπολογισμού του SSA, έχουμε n = 12 και kn = 36. Πρέπει να υπολογίσουμε τα αθροίσματα των τετραγώνων για τους υπό μεταχείριση αρουραίους, SS_{Rats}, τα κομμάτια ήπατος των υπό μεταχείριση αρουραίων , $SS_{\text{Liverbits}}$ και τα σκευάσματα που έχουν παρασχεθεί σε αυτούς τους αρουραίους, $SS_{\text{Preparations}}$:

$$SS_{\text{Rats}} = \frac{\sum R^2}{6} - \frac{\sum C^2}{12},$$

$$SS_{\text{Liverbits}} = \frac{\sum L^2}{2} - \frac{\sum R^2}{6},$$

$$SS_{\text{Preparations}} = \frac{\sum y^2}{1} - \frac{\sum L^2}{2}.$$

Ο διορθωτικός συντελεστής σε οποιοδήποτε επίπεδο είναι το μη διορθωμένο άθροισμα των τετραγώνων του προηγούμενου επιπέδου. Το τελευταίο άθροισμα των τετραγώνων θα μπορούσε να έχει υπολογιστεί από τη διαφορά:

$$SS_{\text{Preparations}} = SSY - SSA - SS_{\text{Rats}} - SS_{\text{Liverbits}}.$$

Η δοκιμασία F για την ισότητα των μέσων μεταχείρισης είναι η διακύμανση της μεταχείρισης διαιρεμένη με τη διακύμανση των υπό μεταχείριση αρουραίων από την αμέσως κάτω γραμμή: F = 778,78/265,89 =2.928.956, με 2 d.f. στον αριθμητή και 3 d.f. στον παρονομαστή (όπως βρήκαμε από την ανωτέρω ANOVA).
Για να το μετατρέψουμε αυτό σε μια συστατική ανάλυση διακύμανσης χρειαζόμαστε λίγη δουλειά. Οι τετραγωνικοί μέσοι μετατρέπονται σε συνιστώσες διακύμανσης ως ακολούθως:

Υπολείμματα = παρασκευάσματα μέσα στα κομμάτια ήπατος: αμετάβλητο = 21,17,
Κομμάτια ήπατος μέσα στους υπό μεταχείριση αρουραίους: (49,5-21,17)/2=14,165
Υπό μεταχείριση αρουραίοι (265,89-49,5)/6=36,065

Διαιρείτε τη διαφορά στην διακύμανση με το σύνολο των αριθμών στο από κάτω επίπεδο (δηλαδή σε αυτήν την περίπτωση, δύο σκευάσματα ανά κομμάτι ήπατος, και έξι σκευάσματα ανά αρουραίο).
Η ανάλυση των δεδομένων των αρουραίων χρησιμοποιώντας lmer εξηγείται στη σελ. 648.

Ποια είναι η διαφορά μεταξύ το split-plot και των ιεραρχικών δειγμάτων;

Τα Split-plot πειράματα έχουν πληροφοριακά επίπεδα παράγοντα ενώ τα ιεραρχικά δείγματα έχουν μη πληροφοριακά. Αυτή είναι η διάκριση. Στο πείραμα άρδευσης, τα επίπεδα του παράγοντα ήταν ως εξής:

```
evels(density)
[1] "high"  "low"  "medium"
evels(fertilizer)
[1] "N"  "NP"  "P"
```

Δείχνουν την πυκνότητα των σπόρων που έχουν σπαρθεί, και το είδος των λιπάσματος που χρησιμοποιήθηκε: είναι πληροφοριακά. Εδώ είναι τα επίπεδα του παράγοντα από το πείραμα με τους αρουραίους:

```
levels(Rat)
[1] "1"  "2"
levels(Liver)
[1] "1"  "2"  "3"
```

Αυτά τα επίπεδα παράγοντα είναι μη πληροφοριακά, επειδή ο υπ'αριθμ. 2 υπό της μεταχείρισης 1 αρουραίος δεν έχει τίποτα κοινό με τον υπ'αριθμ. 2 υπό της μεταχείρισης 2 αρουραίο, ή με τον υπ'αριθμ. 2 υπό της μεταχείρισης 3 αρουραίο. Τα κομμάτια ήπατος με τον αριθμό 3 από τον αρουραίο 1 δεν έχουν τίποτα κοινό με τα κομμάτια ήπατος με τον αριθμό 3 από τον αρουραίο 2. Σημειώστε, ωστόσο, ότι τα αριθμημένα επίπεδα του παράγοντα δεν είναι πάντα μη πληροφοριακά: τα επίπεδα μεταχείρισης 1, 2 και 3 είναι πληροφοριακά: το 1 είναι ο έλεγχος, το 2 είναι ένα συμπλήρωμα διατροφής και το 3 είναι ένας συνδυασμός δύο συμπληρωμάτων διατροφής.

Όταν τα επίπεδα του παράγοντα είναι πληροφοριακά, η μεταβλητή είναι γνωστή ως μια σταθερή επίδραση. Όταν τα επίπεδα του παράγοντα είναι μη πληροφοριακά , η μεταβλητή είναι γνωστή ως μια τυχαία επίδραση. Σε γενικές γραμμές, ενδιαφερόμαστε για τις σταθερές επιδράσεις που επηρεάζουν τον μέσο, και για τις τυχαίες επιδράσεις καθώς αυτές επηρεάζουν τη διακύμανση. Έχουμε την τάση να μην αναφερόμαστε στην επίδραση μεγέθους που αναλογεί στις τυχαίες επιδράσεις αλλά να εστιάζουμε κυρίως στην επίδραση μεγέθους και στα τυπικά σφάλματα της, όταν έχουμε σταθερές επιδράσεις. Έτσι, η άρδευση, η πυκνότητα και το λίπασμα είναι σταθερές επιδράσεις, και οι αρουραίοι και τα κομμάτια ήπατος είναι τυχαίες επιδράσεις.

ANOVA με aov ή lm

Η διαφορά μεταξύ των lm και aov αφορά κυρίως στη μορφή του αποτελέσματος: ο πίνακας summary aov περιλαμβάνεται στην παραδοσιακή μορφή ανάλυσης της διακύμανσης, με μία σειρά για κάθε κατηγορική μεταβλητή και για κάθε όρο αλληλεπίδρασης. Από την άλλη πλευρά, ο περιληπτικός πίνακας lm παράγει μία γραμμή ανά εκτιμώμενη παράμετρο (δηλαδή μία γραμμή για κάθε επίπεδο παράγοντα και μία γραμμή για κάθε επίπεδο αλληλεπίδρασης). Εάν έχετε πολλαπλούς όρους σφάλματος, θα πρέπει να χρησιμοποιήσετε aov επειδή η lm δεν υποστηρίζει τον όρο *Error*. Εδώ είναι η ίδια ανάλυση διακύμανσης κατά δύο παράγοντες προσαρμοσμένη, χρησιμοποιώντας aov αρχικά και στη συνέχεια, lm:

```
daphnia<-read.table("c:\\temp\\Daphnia.txt",header=T)
attach(daphnia)
names(daphnia)
[1] "Growth.rate" "Water" "Detergent" "Daphnia"
model1<-aov(Growth.rate~Water*Detergent*Daphnia)
summary(model1)
                       Df  Sum Sq Mean Sq F value     Pr(>F)
Water                   1   1.985   1.985  2.8501  0.0978380    .
Detergent               3   2.212   0.737  1.0586  0.3754783
Daphnia                 2  39.178  19.589 28.1283  8.228e-09  ***
Water:Detergent         3   0.175   0.058  0.0837  0.9686075
Water:Daphnia           2  13.732   6.866  9.8591  0.0002587  ***
Detergent:Daphnia       6  20.601   3.433  4.9302  0.0005323  ***
Water:Detergent:Daphnia 6   5.040   0.975  1.3995  0.2343235
Residuals              48  33.428   0.696
model2<-lm(Growth.rate~Water*Detergent*Daphnia)
summary(model2)
```

```
Coefficients:
                                          Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)                                2.81126      0.48181     5.835   4.48e-07
WaterWear                                 -0.15808      0.68138    -0.232    0.81753
DetergentBrandB                           -0.03536      0.68138    -0.052    0.95883
DetergentBrandC                            0.47626      0.68138     0.699    0.48794
DetergentBrandD                           -0.21407      0.68138    -0.314    0.75475
DaphniaClone2                              0.49637      0.68138     0.728    0.46986
DaphniaClone3                              2.05526      0.68138     3.016    0.00408
WaterWear:DetergentBrandB                  0.46455      0.96361     0.482    0.63193
WaterWear:DetergentBrandC                 -0.27431      0.96361    -0.285    0.77712
WaterWear:DetergentBrandD                  0.21729      0.96361     0.225    0.82255
WaterWear:DaphniaClone2                    1.38081      0.96361     1.433    0.15835
WaterWear:DaphniaClone3                    0.43156      0.96361     0.448    0.65627
DetergentBrandB:DaphniaClone2              0.91892      0.96361     0.954    0.34506
DetergentBrandC:DaphniaClone2             -0.16337      0.96361    -0.170    0.86609
DetergentBrandD:DaphniaClone2              1.01209      0.96361     1.050    0.29884
DetergentBrandB:DaphniaClone3             -0.06490      0.96361    -0.067    0.94658
DetergentBrandC:DaphniaClone3             -0.80789      0.96361    -0.838    0.40597
DetergentBrandD:DaphniaClone3             -1.28669      0.96361    -1.335    0.18809
WaterWear:DetergentBrandB:DaphniaClone2   -1.26380      1.36275    -0.927    0.35837
WaterWear:DetergentBrandC:DaphniaClone2    1.35612      1.36275     0.995    0.32466
WaterWear:DetergentBrandD:DaphniaClone2    0.77616      1.36275     0.570    0.57164
WaterWear:DetergentBrandB:DaphniaClone3   -0.87443      1.36275    -0.642    0.52414
WaterWear:DetergentBrandC:DaphniaClone3   -1.03019      1.36275    -0.756    0.45337
WaterWear:DetergentBrandD:DaphniaClone3   -1.55400      1.36275    -1.140    0.25980

Residual standard error: 0.8345 on 48 degrees of freedom
Multiple R-Squared: 0.7147, Adjusted R-squared: 0.578
F-statistic: 5.227 on 23 and 48 DF, p-value: 7.019e-07
```

Σημειώστε ότι δύο σημαντικές αλληλεπιδράσεις, οι Water–Daphnia και Detergent–Daphnia, εμφανίζονται στον πίνακα aov αλλά όχι στην περιληπτική lm (αυτό συχνά οφείλεται στο γεγονός ότι η lm summary δείχνει περισσότερο τις συγκρίσεις των μεταχειρίσεων παρά τις συγκρίσεις Helmert). Αυτό εφιστά την προσοχή περισσότερο στη σημασία της απλούστευσης του μοντέλου από ότι στα ανά σειρά t-tests (δηλ. αφαιρώντας τον μη σημαντικό τριπλό όρο αλληλεπίδρασης, στην περίπτωση αυτή).

Η κύρια διαφορά είναι ότι υπάρχουν οκτώ σειρές στο συνοπτικό πίνακα aov (τρεις κύριες επιδράσεις, τρεις αμφίδρομες αλληλεπιδράσεις, μια τριπλή αλληλεπίδραση και ένας όρος σφάλματος), αλλά υπάρχουν 24 γραμμές στο συνοπτικό πίνακα lm (τέσσερα επίπεδα detergent και τρία επίπεδα Daphnia clone με δύο επίπεδα νερού). Μπορείτε να δείτε εύκολα το αποτέλεσμα της model1 σε διάταξη γραμμικού μοντέλου, ή της model2 ως έναν πίνακα ANOVA χρησιμοποιώντας επιλογές summary όπως lm ή aov:

```
summary.lm(model1)
summary.aov(model2)
```

Μεγέθη επίδρασης

Σε περίπλοκα σχεδιασμένα πειράματα, είναι πιο εύκολο να συνοψίσουμε τα μεγέθη αποτελεσμάτων με την συνάρτηση *model.tables*. Αυτή παίρνει το όνομα του προσαρμοσμένου μοντέλου αντικειμένου ως το πρώτο της επιχείρημα, και μπορείτε να καθορίσετε αν θέλετε τα τυπικά σφάλματα (όπως θα κάνατε συνήθως):

```
model.tables(model1, "means", se = TRUE)
```

```
Tables of means
Grand mean

3.851905

  Water

Water
 Tyne   Wear
3.686  4.018

  Detergent

Detergent
BrandA  BrandB  BrandC  BrandD
 3.885   4.010   3.955   3.558

  Daphnia

Daphnia
Clone1  Clone2  Clone3
 2.840   4.577   4.139

Water:Detergent

      Detergent
Water  BrandA  BrandB  BrandC  BrandD
 Tyne   3.662   3.911   3.814   3.356
 Wear   4.108   4.109   4.095   3.760

Water:Daphnia

      Daphnia
Water  Clone1  Clone2  Clone3
 Tyne   2.868   3.806   4.383
 Wear   2.812   5.348   3.894

Detergent:Daphnia

Daphnia
Detergent  Clone1  Clone2  Clone3
   BrandA   2.732   3.919   5.003
   BrandB   2.929   4.403   4.698
   BrandC   3.071   4.773   4.019
   BrandD   2.627   5.214   2.834

Water:Detergent:Daphnia
, , Daphnia = Clone1

Detergent
Water  BrandA  BrandB  BrandC  BrandD
 Tyne   2.811   2.776   3.288   2.597
 Wear   2.653   3.082   2.855   2.656

, , Daphnia = Clone2

Detergent
Water  BrandA  BrandB  BrandC  BrandD
 Tyne   3.308   4.191   3.621   4.106
 Wear   4.530   4.615   5.925   6.322

, , Daphnia = Clone3
```

Σημειώστε ότι τα τυπικά σφάλματα είναι τυπικά σφάλματα διαφορών, και είναι διαφορετικά σε καθένα από τα διαφορετικά στρώματα, επειδή η επανάληψη διαφέρει. Όλα τα τυπικά σφάλματα χρησιμοποιούν την ίδια συγκεντρωτική διακύμανση σφαλμάτων $s^2 = 0.696$ 0(βλ. ανωτέρω). Για παράδειγμα, οι τριπλές αλληλεπιδράσεις έχουν

$se = \sqrt{2 \times 0.696/3} = 0.681$ και οι κύριες επιδράσεις Daphnia έχουν $se = \sqrt{2 \times 0.696/24} = 0.2409.$

Ελκυστικά γραφήματα των μεγεθών των αποτελεσμάτων μπορεί να επιτευχθούν χρησιμοποιώντας τη «βιβλιοθήκη» *effects* (σελ. 178).

Επαναλήψεις

Η συνάρτηση *replications* σας επιτρέπει να ελέγχετε τον αριθμό των επαναλήψεων σε κάθε επίπεδο σε ένα πειραματικό σχεδιασμό:

```
replications(Growth.rate~Daphnia*Water*Detergent,daphnia)
                                Daphnia                   Water
Detergent
                                     24                      36
18
                          Daphnia:Water     Daphnia:Detergent
Water:Detergent
                                     12                       6
9
Daphnia:Water:Detergent
                                      3
```

Υπάρχουν τρεις επαναλήψεις για την τριμερή αλληλεπίδραση και για όλες τις αμφίδρομες αλληλεπιδράσεις (για να το δείτε αυτό θα πρέπει να θυμάστε τον αριθμό των επιπέδων για κάθε παράγοντα: υπάρχουν δύο τύποι νερού, τρεις κλώνοι Daphnia και τέσσερα απολυμαντικά (βλέπε παραπάνω).

Πολλαπλές Συγκρίσεις

Κατά τη σύγκριση των μέσων για τα επίπεδα ενός παράγοντα σε μια ανάλυση διακύμανσης, μια απλή σύγκριση με τη χρήση πολλαπλών t tests θα διογκώσει την πιθανότητα της δήλωσης μιας σημαντικής διαφοράς, ενώ στην πραγματικότητα δεν υπάρχει καμία. Αυτό συμβαίνει διότι τα διαστήματα υπολογίζονται με μια δεδομένη πιθανότητα κάλυψης για κάθε χρονικό διάστημα, αλλά η ερμηνεία της κάλυψης γίνεται συνήθως σε σχέση με όλη την «οικογένεια-ομάδα» διαστημάτων (δηλαδή για τον παράγοντα ως σύνολο).

Αν ακολουθήσετε το πρωτόκολλο της απλοποίησης του μοντέλου που προτείνεται σε αυτό το βιβλίο, τότε τα ζητήματα από τις πολλαπλές συγκρίσεις δεν θα προκύπτουν πολύ συχνά. Ένα περιστασιακό σημαντικό t test μεταξύ μιας δέσμης μη σημαντικών όρων αλληλεπίδρασης δεν είναι πιθανό να επιβιώσει ενός ελέγχου διαγραφής (βλέπε σελ. 325). Επίσης, αν έχετε παράγοντες με μεγάλο αριθμό επιπέδων θα μπορούσατε να εξετάσετε τη χρήση μικτών μοντέλων αποτελεσμάτων αντί της ANOVA (δηλαδή αντιμετωπίζοντας τους παράγοντες, ως τυχαίες και όχι ως σταθερές επιδράσεις, βλ. σελ. 627).

Ο John Tukey παρουσιάζει διαστήματα βασισμένα στο φάσμα του δείγματος των μέσων και όχι σε επιμέρους διαφορές. Στις μέρες μας, αυτές ονομάζονται ειλικρινείς σημαντικές διαφορές Tukey. Τα διαστήματα που εξάγονται από της συνάρτηση *TukeyHSD*

βασίζονται σε στατιστικά Studentized. Πρακτικά, τα διαστήματα που κατασκευάστηκαν με αυτόν τον τρόπο θα μπορούσαν να εφαρμοστούν μόνο σε σχέδια όπου σε κάθε επίπεδο του παράγοντα γίνεται ο ίδιος αριθμός παρατηρήσεων. Αυτή η συνάρτηση ενσωματώνει μια προσαρμογή για το μέγεθος του δείγματος η οποία παράγει λογικά διαστήματα για ηπίως ασύμμετρα σχέδια.

Το παρακάτω παράδειγμα αφορά στην απόδοση των μυκήτων που συγκεντρώθηκαν σε 16 διαφορετικούς βιότοπους:

```
data<-read.table("c:\\temp\\Fungi.txt",header=T)
attach(data)
names(data)
```

Πρώτα διαπιστώνουμε αν υπάρχει οποιαδήποτε μεταβολή στην απόδοση των μυκήτων η οποία θα πρέπει να εξηγηθεί:

```
model<-aov(Fungus.yield~Habitat)
summary(model)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Habitat	15	7527.4	501.8	72.141	< 2.2e-16	***
Residuals	144	1001.7	7.0			

Έτσι έχουμε $(p < 0.000\,001)$. Αλλά αυτό δεν έχει πραγματικό ενδιαφέρον, επειδή απλά δείχνει ότι ορισμένοι βιότοποι παράγουν περισσότερους μύκητες από άλλους. Είναι πιθανό να ενδιαφερόμαστε σχετικά με το σε ποιους βιότοπους παράγονται σημαντικά περισσότεροι μύκητες από άλλους. Οι πολλαπλές συγκρίσεις εδώ, είναι ένα ζήτημα, επειδή υπάρχουν 16 βιότοποι με συνέπεια να υπάρχουν $(16 \times 15)/2 = 120$ πιθανές συγκρίσεις ζευγών.

Υπάρχουν δύο επιλογές:

• εφαρμόζουμε τη συνάρτηση *TukeyHSD* στο μοντέλο για να πάρουμε τις ειλικρινείς σημαντικές διαφορές Tukey.

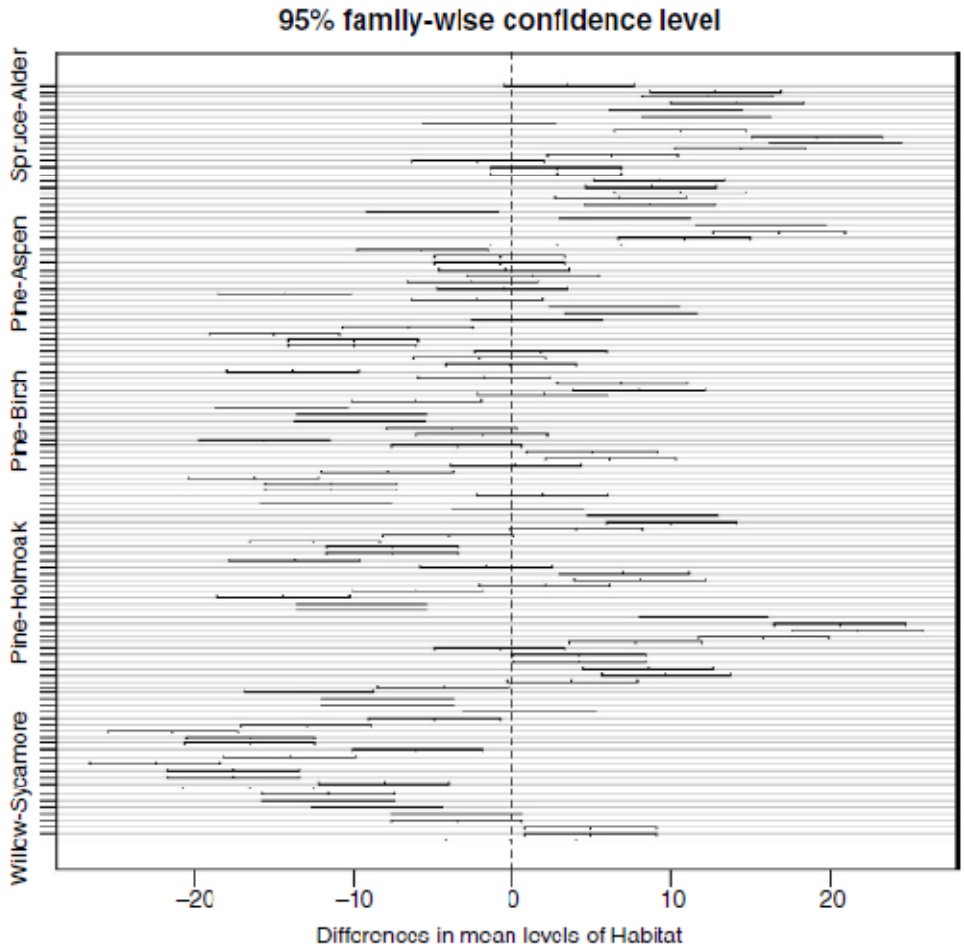
• Χρησιμοποιούμε τη συνάρτηση *pairwise.t.test* για να πάρουμε τις προσαρμοσμένες τιμές p για όλες τις συγκρίσεις.

Εδώ είναι ο έλεγχος Tukey στην πράξη: παράγει έναν προεπιλεγμένο πίνακα τιμών p:

```
TukeyHSD(model)

Tukey multiple comparisons of means
  95% family-wise confidence level

Fit: aov(formula = Fungus.yield ~ Habitat)

$Habitat
                      diff         lwr         upr       p adj
Ash-Alder         3.53292777  -0.5808096   7.6466651  0.1844088
Aspen-Alder      12.78574402   8.6720067  16.8994814  0.0000000
Beech-Alder      12.32365349   8.2099161  16.4373908  0.0000000
Birch-Alder      14.11348150   9.9997441  18.2272189  0.0000000
Cherry-Alder     10.29508769   6.1813503  14.4088250  0.0000000
Chestnut-Alder   12.24107899   8.1273416  16.3548163  0.0000000
Holmoak-Alder    -1.44360558  -5.5573429   2.6701318  0.9975654
Hornbeam-Alder   10.60271044   6.4889731  14.7164478  0.0000000
Lime-Alder       19.19458205  15.0808447  23.3083194  0.0000000
Oak-Alder        20.29457340  16.1808360  24.4083108  0.0000000
Pine-Alder       14.34084715  10.2271098  18.4545845  0.0000000


Rowan-Alder       6.29495226   2.1812149  10.4086896  0.0000410
Spruce-Alder     -2.15119456  -6.2649319   1.9625428  0.9036592
Sycamore-Alder    2.80900108  -1.3047363   6.9227384  0.5644643
...
Spruce-Rowan     -8.44614681 -12.5598842  -4.3324095  0.0000000
Sycamore-Rowan   -3.48595118  -7.5996885   0.6277862  0.2019434
Willow-Rowan     -3.51860059  -7.6323379   0.5951368  0.1896363
Sycamore-Spruce   4.96019563   0.8464583   9.0739330  0.0044944
Willow-Spruce     4.92754623   0.8138089   9.0412836  0.0049788
Willow-Sycamore  -0.03264941  -4.1463868   4.0810879  1.0000000
```

Μπορείτε αν προτιμάτε να σχεδιάσετε τα διαστήματα εμπιστοσύνης (ή ασφαλώς να κάνετε και τα δύο)

```
plot(TukeyHSD(model))
```



Οι βιότοποι που βρίσκονται σε αντίθετες πλευρές της διακεκομμένης γραμμής και δεν την καλύπτουν είναι σημαντικά διαφορετικοί μεταξύ τους.

Εναλλακτικά, μπορείτε να χρησιμοποιήσετε τη συνάρτηση *pairwise.t.test* στην οποία καθορίζετε πρωτίστως την απαντητική μεταβλητή, και στη συνέχεια την κατηγορική ερμηνευτική μεταβλητή που περιέχει τα επίπεδα του παράγοντα που θέλετε να συγκρίνετε, χωρισμένα με ένα κόμμα (περισπωμένη):

```
pairwise.t.test(Fungus.yield,Habitat)

        Pairwise comparisons using t tests with pooled SD

data: Fungus.yield and Habitat
         Alder    Ash      Aspen    Beech    Birch    Cherry   Chestnut Holmoak
Ash      0.10011  -        -        -        -        -        -        -
Aspen    < 2e-16  6.3e-11  -        -        -        -        -        -
Beech    < 2e-16  5.4e-10  1.00000  -        -        -        -        -
Birch    < 2e-16  1.2e-13  1.00000  1.00000  -        -        -        -
Cherry   4.7e-13  2.9e-06  0.87474  1.00000  0.04943  -        -        -
Chestnut < 2e-16  7.8e-10  1.00000  1.00000  1.00000  1.00000  -        -
Holmoak  1.00000  0.00181  < 2e-16  < 2e-16  < 2e-16  3.9e-16  < 2e-16  -
Hornbeam 1.1e-13  8.6e-07  1.00000  1.00000  0.10057  1.00000  1.00000  < 2e-16
Lime     < 2e-16  < 2e-16  1.1e-05  1.9e-06  0.00131  3.3e-10  1.4e-06  < 2e-16
Oak      < 2e-16  < 2e-16  1.4e-07  2.0e-08  2.7e-05  1.9e-12  1.5e-08  < 2e-16
Pine     < 2e-16  3.9e-14  1.00000  1.00000  1.00000  0.02757  1.00000  < 2e-16
Rowan    1.8e-05  0.51826  8.5e-06  4.7e-05  3.9e-08  0.03053  6.2e-05  5.3e-08
Spruce   1.00000  0.00016  < 2e-16  < 2e-16  < 2e-16  < 2e-16  < 2e-16  1.00000
Sycamore 0.50084  1.00000  2.1e-12  1.9e-11  3.3e-15  1.5e-07  2.7e-11  0.01586
Willow   0.51826  1.00000  1.9e-12  1.6e-11  2.8e-15  1.4e-07  2.4e-11  0.01702
         Hornbeam Lime     Oak      Pine     Rowan    Spruce   Sycamore
Ash      -        -        -        -        -        -        -
Aspen    -        -        -        -        -        -        -
Beech    -        -        -        -        -        -        -
Birch    -        -        -        -        -        -        -
Cherry   -        -        -        -        -        -        -
Chestnut -        -        -        -        -        -        -
Holmoak  -        -        -        -        -        -        -
Hornbeam -        -        -        -        -        -        -
Lime     1.3e-09  -        -        -        -        -        -
Oak      8.4e-12  1.00000  -        -        -        -        -
Pine     0.05975  0.00253  6.1e-05  -        -        -        -
Rowan    0.01380  < 2e-16  < 2e-16  1.5e-08  -        -        -
Spruce   < 2e-16  < 2e-16  < 2e-16  < 2e-16  2.5e-09  -        -
Sycamore 4.2e-08  < 2e-16  < 2e-16  1.1e-15  0.10218  0.00187  -
Willow   3.8e-08  < 2e-16  < 2e-16  9.3e-16  0.10057  0.00203  1.00000

P value adjustment method: holm
```

Όπως βλέπετε, η προεπιλεγμένη μέθοδος προσαρμογής των τιμών p είναι η Holm, αλλά άλλες μέθοδοι προσαρμογής περιλαμβάνουν τις Hochberg, Hommel, Bonferroni, BH, BY, fdr και none. Χωρίς προσαρμογή των τιμών p, η σύγκριση rowan–willow φαίνεται πολύ σημαντική $(p = 0.003\,35)$, , όπως μπορείτε να δείτε:

```
pairwise.t.test(Fungus.yield,Habitat,p.adjust.method="none")
```

Μου αρέσει η *TukeyHSD* επειδή είναι συντηρητική αλλά όχι γελοιωδώς, (σε αντίθεση με την Bonferroni). Για παράδειγμα, ο Tukey δίνει τη σύγκριση birch–cherry, ως μη σημαντική $(p = 0.101\,1027)$, ενώ ο Holm κάνει αυτή τη διαφορά σημαντική $(p = 0.049\,43).$. Ο Tukey είχε τη Willow-Holm Oak ως σημαντική $(p = 0.038\,091\,0),$ ενώ ο Bonferroni «ρίχνει έξω το μωρό μαζί με τα απόνερα» $(p = 0.056\,72).$ Θα πρέπει να αποφασίσετε πόσο ενημερωμένοι θέλετε να είστε στο πλαίσιο της συγκεκριμένης ερώτησή σας.
Υπάρχει ένα χρήσιμο πακέτο για πολλαπλές συγκρίσεις που ονομάζεται *multcomp*:

```
install.packages("multcomp")
```

Μπορείτε να δείτε αμέσως το πόσο αμφιλεγόμενο είναι το ζήτημα των πολλαπλών συγκρίσεων, απλά κοιτώντας το μήκος του καταλόγου των διαφορετικών μεθόδων πολλαπλών συγκρίσεων που υποστηρίζονται στο παρόν πακέτο.

- the many-to-one comparisons of Dunnett
- the all-pairwise comparisons of Tukey
- Sequen
- AVE
- changepoint
- Williams
- Marcus
- McDermott
- Tetrade
- Bonferroni correction
- Holm
- Hochberg
- Hommel
- Benjamini–Hochberg
- Benjamini–Yekutieli

Η παλιομοδίτικη διόρθωση *Bonferroni* είναι ιδιαίτερα συντηρητική, διότι οι τιμές p πολλαπλασιάζονται με τον αριθμό των συγκρίσεων. Αντί να χρησιμοποιήσουν τις συνήθεις διαδικασίες Bonferroni και Holm, οι μέθοδοι προσαρμογής περιλαμβάνουν λιγότερο συντηρητικές διορθώσεις που λαμβάνουν υπόψη τους τις ακριβείς συσχετίσεις μεταξύ των στατιστικών ελέγχων χρησιμοποιώντας την πολυμεταβλητή κατανομή t. Επομένως, οι προκύπτουσες διαδικασίες είναι ουσιαστικά περισσότερο ισχυρές (οι προσαρμοσμένες τιμές p των Bonferroni και Holm καταγράφονται για αναφορά). Φαίνεται ότι δεν υπάρχει λόγος να χρησιμοποιήσετε τη μη τροποποιημένη διόρθωση Bonferroni επειδή κυριαρχείται από τη μέθοδο Holm, η οποία είναι έγκυρη κάτω από αυθαίρετες υποθέσεις. Οι έλεγχοι έχουν σχεδιαστεί για να ταιριάζουν σε πολλαπλές συγκρίσεις μέσα στο γενικό γραμμικό μοντέλο. Αυτό σημαίνει ότι επιτρέπουν τις συμμεταβλητές, τα ένθετα αποτελέσματα, τους συσχετισμένους μέσους και τις τιμές που λείπουν. Οι τέσσερις πρώτες μέθοδοι έχουν σχεδιαστεί για να δώσουν ισχυρό έλεγχο στα ομαδικά ποσοστά σφάλματος. Οι μέθοδοι Benjamini, Hochberg, και Yekutieli ελέγχουν το ποσοστό ψεύτικων ευρημάτων, το οποίο είναι το αναμενόμενο ποσοστό των ψευδών ευρημάτων μεταξύ των απορριφθέντων υποθέσεων. Το ποσοστό των ψεύτικων ευρημάτων είναι μια λιγότερο αυστηρή συνθήκη από ότι το ομαδικό ποσοστό σφάλματος, με αποτέλεσμα αυτές οι μέθοδοι είναι πιο ισχυρές
από τις άλλες.

Προβολές Μοντέλων

Αν θέλετε να δείτε πώς τα διαφορετικά επίπεδα παράγοντα συμβάλλουν με τα αθροιστικά αποτελέσματα τους σε κάθε μία από τις παρατηρούμενες τιμές απόκρισης, χρησιμοποιήστε τη συνάρτηση proj ως ακολούθως:

```
library(help="multcomp")
```

```
              (Intercept)       Water   Detergent     Daphnia  Water:Detergent
Water:Daphnia
1                3.851905  -0.1660431  0.03292724  -1.0120302      -0.05698158
0.1941404
2                3.851905  -0.1660431  0.03292724  -1.0120302      -0.05698158
0.1941404
3                3.851905  -0.1660431  0.03292724  -1.0120302      -0.05698158
0.1941404
...
```

Το όνομα *proj* προέρχεται από το γεγονός ότι η συνάρτηση επιστρέφει μία μήτρα ή μια λίστα μητρών δίνοντας τις «προβολές των δεδομένων σε όρους ενός γραμμικού μοντέλου».

Πολυμεταβλητή ανάλυση διακύμανσης

Ενίοτε, δύο ή περισσότερες απαντητικές μεταβλητές μετρώνται στο ίδιο πείραμα. Ασφαλώς, ο τυπικός τρόπος για να προχωρήσετε περιλαμβάνει την ανάλυση κάθε μεταβλητής απόκρισης ξεχωριστά. Αλλά υπάρχουν περιπτώσεις κατά τις οποίες θέλετε να αντιμετωπίσετε την ομάδα των μεταβλητών απόκρισης, ως μια πολυμεταβλητή απόκριση. Η συνάρτηση για να το κάνετε αυτό είναι η manova, η πολυμεταβλητή ανάλυση διακύμανσης. Σημειώστε ότι η manova, δεν υποστηρίζει πολυστρωματική ανάλυση διακύμανσης, οπότε ο τύπος δεν πρέπει να περιλαμβάνει κάποιον όρο Error.

```
data<-read.table("c:\\temp\\manova.txt",header=T)
attach(data)
names(data)
```

```
[1] "tear" "gloss" "opacity" "rate" "additive"
```

Πρώτα, δημιουργήστε μια απαντητική πολυμεταβλητή, Y, ενώνοντας τις τρεις χωριστές απαντητικές μεταβλητές (tear, gloss και opacity), ως ακολούθως:

```
Y <- cbind(tear, gloss, opacity)
```

Στη συνέχεια, προσαρμόστε την πολυμεταβλητή ανάλυση διακύμανσης χρησιμοποιώντας τη συνάρτηση manova:

```
model<-manova(Y~rate*additive)
```

Υπάρχουν δύο τρόποι για να ελέγξετε το αποτέλεσμα. Πρώτον, ως μια πολυμεταβλητή ανάλυση διακύμανσης:

```
summary(model)
```

```
               Df  Pillai  approx F  num Df  den Df    Pr(>F)
rate            1  0.6181    7.5543       3      14  0.003034  **
additive        1  0.4770    4.2556       3      14  0.024745   *
rate:additive   1  0.2229    1.3385       3      14  0.301782
Residuals      16
```

Αυτή δείχνει σημαντικές κύριες επιδράσεις τόσο για την κύρια τιμή όσο και για την πρόσθετη, αλλά καμία μεταξύ τους αλληλεπίδραση. Σημειώστε ότι οι δοκιμασίες F βασίζονται σε 3 και 14 βαθμούς ελευθερίας (όχι σε 1 και 16). Η προεπιλεγμένη μέθοδος summary.manova είναι η στατιστική Pillai-Bartlett. Άλλες επιλογές περιλαμβάνουν τις μεθόδους Wilks Hotelling-Lawley και Roy. Δεύτερον, θα θελήσετε να εξετάσετε κάθε μια από τις τρεις μεταβλητές χωριστά:

```
summary.aov(model)
```

```
Response tear :
               Df   Sum Sq  Mean Sq  F value   Pr(>F)
rate            1  1.74050  1.74050  15.7868  0.001092  **
additive        1  0.76050  0.76050   6.8980  0.018330   *
rate:additive   1  0.00050  0.00050   0.0045  0.947143
Residuals      16  1.76400  0.11025
Response gloss :
               Df   Sum Sq  Mean Sq  F value   Pr(>F)
rate            1  1.30050  1.30050   7.9178  0.01248  *
additive        1  0.61250  0.61250   3.7291  0.07139  .
rate:additive   1  0.54450  0.54450   3.3151  0.08740  .
Residuals      16  2.62800  0.16425
Response opacity :
               Df  Sum Sq  Mean Sq  F value  Pr(>F)
rate            1   0.421    0.421   0.1036  0.7517
additive        1   4.901    4.901   1.2077  0.2881
rate:additive   1   3.961    3.961   0.9760  0.3379
Residuals      16  64.924    4.058
```

Παρατηρήστε ότι μία από τις τρεις μεταβλητές απόκρισης, το opacity, δεν είναι σημαντικά συσχετισμένη με καμία από τις ερμηνευτικές μεταβλητές.

12.
Ανάλυση Συνδιακύμανσης

Η ανάλυση της συνδιακύμανσης (ANCOVA) συνδυάζει στοιχεία από την παλινδρόμηση και την ανάλυση της διακύμανσης. Η μεταβλητή απόκρισης είναι συνεχής, και υπάρχει τουλάχιστον μια συνεχής επεξηγηματική μεταβλητή και τουλάχιστον μία κατηγορική επεξηγηματική μεταβλητή. Η διαδικασία λειτουργεί ως ακολούθως:

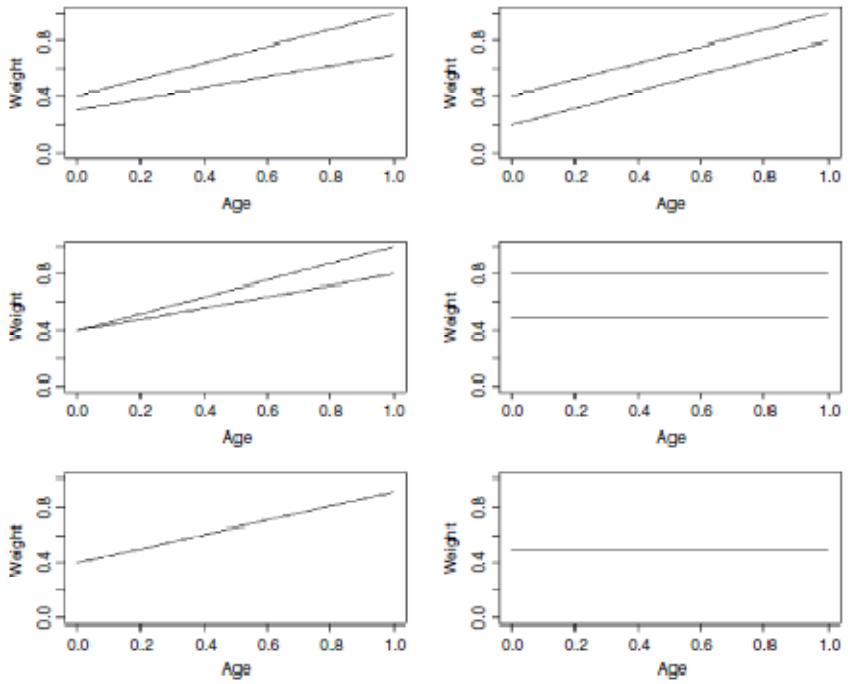
• Προσαρμόζουμε δύο ή περισσότερες γραμμικές παλινδρομήσεις των y έναντι των x (μία για κάθε επίπεδο του παράγοντα).
• Εκτιμούμε διαφορετικές κλίσεις και τομές για κάθε επίπεδο.
• Χρησιμοποιούμε την απλοποίηση του μοντέλου (δοκιμές διαγραφής) για να εξαλείψουμε μη απαραίτητες παραμέτρους.

Για παράδειγμα, θα μπορούσαμε να χρησιμοποιήσουμε τη μέθοδο ANCOVA σε ένα ιατρικό πείραμα, όπου η μεταβλητή απόκρισης ήταν «ημέρες για ανάρρωση» και οι επεξηγηματικές μεταβλητές ήταν «καπνιστής ή μη καπνιστής» (κατηγορική) και «καταμέτρηση κυττάρων του αίματος» (συνεχής). Στα οικονομικά, το ποσοστό τοπικής ανεργίας θα μπορούσε να μοντελοποιηθεί ως συνάρτηση της χώρας (κατηγορική μεταβλητή) και του μεγέθους του τοπικού πληθυσμού (συνεχής μεταβλητή). Ας υποθέσουμε ότι μοντελοποιούμε το βάρος (μεταβλητή απόκρισης) ως συνάρτηση του φύλου και της ηλικίας. Το φύλο είναι ένας παράγοντας με δύο επίπεδα (αρσενικό και θηλυκό) και η ηλικία είναι μια συνεχής μεταβλητή. Ως εκ τούτου, το μέγιστο μοντέλο έχει τέσσερις παραμέτρους: δύο κλίσεις (μια κλίση για τους άνδρες και μια για τις γυναίκες) και δύο σημεία τομής (ένα για τους άνδρες και ένα για τις γυναίκες), ως ακολούθως:

$$weight_{\text{male}} = a_{\text{male}} + b_{\text{male}} \times age,$$
$$weight_{\text{female}} = a_{\text{female}} + b_{\text{female}} \times age.$$

Το μέγιστο μοντέλο εμφανίζεται στην πάνω αριστερή πλευρά του πίνακα. Η απλοποίηση του μοντέλου είναι ένα ουσιαστικό μέρος της ανάλυσης της συνδιακύμανσης, επειδή η αρχή της οικονομίας απαιτεί να κρατάμε στο μοντέλο όσο το δυνατό λιγότερες παραμέτρους.

Υπάρχουν έξι πιθανά μοντέλα σε αυτή την περίπτωση, και η διαδικασία της απλοποίησης του μοντέλου ξεκινά ρωτώντας κατά πόσον χρειαζόμαστε και τις τέσσερις παραμέτρους (πάνω αριστερά). Ίσως θα μπορούσαμε να το κάνουμε με 2 τομές και μια κοινή κλίση (επάνω δεξιά), ή με ένα κοινό σημείο τομής και δύο διαφορετικές κλίσεις (κέντρο αριστερά). Στο παράδειγμα μας, η ηλικία μπορεί να μην έχει σημαντική επίδραση στην απόκριση, και έτσι να χρειαζόμαστε μόνο δύο παραμέτρους για να περιγράψουμε τις κύριες επιδράσεις του φύλου πάνω στο βάρος. Αυτό θα μπορούσε να μας δώσει δύο ξεχωριστές, οριζόντιες γραμμές στο γράφημα (ένα μέσο βάρος για κάθε φύλο, κέντρο δεξιά). Εναλλακτικά, μπορεί να μην υπάρχει καθόλου επίδραση του φύλου. Σε αυτή την περίπτωση χρειαζόμαστε μόνο δύο παραμέτρους (μία κλίση και ένα σημείο τομής) για να περιγράψουμε την επίδραση της ηλικίας στο βάρος (κάτω αριστερά). Οριακά, μπορεί ούτε η συνεχής ούτε η κατηγορική επεξηγηματική μεταβλητή να έχουν οποιαδήποτε σημαντική επίδραση στην απόκριση. Σε αυτήν την περίπτωση η απλοποίηση του μοντέλου θα οδηγήσει σε ένα μονοπαραμετρικό μηδενικό μοντέλο $\hat{y} = \bar{y}$ (μία ενιαία οριζόντια γραμμή, κάτω δεξιά).



Ανάλυση συνδιακύμανσης στην R

Θα μπορούσαμε να χρησιμοποιήσουμε τις εντολές lm ή aov. Η επιλογή επηρεάζει μόνο τη μορφή του συνοπτικού πίνακα. Θα πρέπει να χρησιμοποιήσουμε και τις δύο μεθόδους και να συγκρίνουμε τα αποτελέσματά τους. Το υπό μελέτη παράδειγμά μας αφορά ένα πείραμα σχετικά με την επίδραση της βόσκησης στην παραγωγή σπόρων ενός διετούς φυτού. Σαράντα φυτά καταμερίστηκαν σε δύο μεταχειρίσεις, βόσκηση και μη βόσκηση, και τα υπό βόσκηση φυτά εκτέθηκαν σε κουνέλια κατά τη διάρκεια των δύο πρώτων εβδομάδων της επιμήκυνσης των μίσχων τους. Στη συνέχεια τα φυτά αυτά προστατεύτηκαν από μεταγενέστερη βόσκηση με την ανέγερση ενός φράχτη και τους δόθηκε η δυνατότητα να αναγεννηθούν. Καθώς το αρχικό μέγεθος των φυτών θεωρήθηκε πιθανόν να επηρεάσει την παραγωγή φρούτων, η διάμετρος της κορυφής της ρίζας μετρήθηκε πριν από τη μεταφύτευση κάθε φυτού. Στο τέλος της καλλιεργητικής περιόδου, η παραγωγή φρούτων (ξηρό βάρος σε χιλιοστόγραμμα) καταγράφηκε για καθένα από τα 40 φυτά. Το μέγεθος αυτό αποτελεί την μεταβλητή απόκρισης στην ακόλουθη ανάλυση.

```
regrowth<-read.table("c:\\temp\\ipomopsis.txt",header=T)
attach(regrowth)
names(regrowth)
```

```
[1] "Root"  "Fruit"  "Grazing"
```

Ο σκοπός της άσκησης είναι να εκτιμήσουμε τις παραμέτρους του ελάχιστα επαρκούς μοντέλου για αυτά τα δεδομένα. Ξεκινάμε ελέγχοντας τα δεδομένα με ένα γράφημα της παραγωγής φρούτων σε σχέση με το μέγεθος της ρίζας για κάθε μία από τις δύο μεταχειρίσεις χωριστά: τα διαμάντια είναι τα φυτά που δεν έχουν βοσκηθεί και τα τρίγωνα είναι τα φυτά που έχουν βοσκηθεί,

```
plot(Root,Fruit,
    pch=16+as.numeric(Grazing),col=c("blue","red")[as.numeric(Grazing)])
```

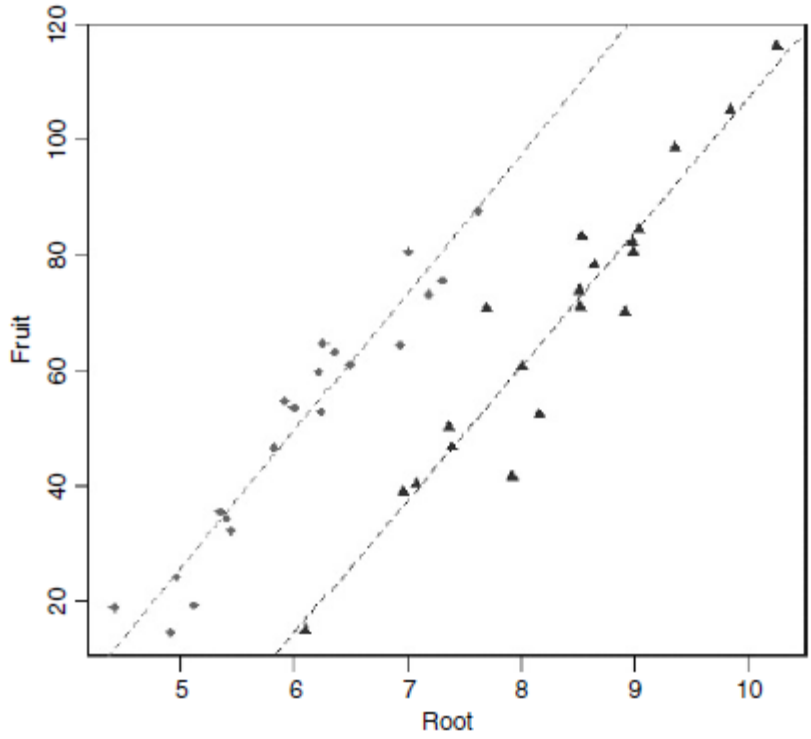
όπου τα κόκκινα διαμάντια αντιπροσωπεύουν φυτά που δεν έχουν βοσκηθεί και τα μπλε τρίγωνα αντιπροσωπεύουν τα φυτά που έχουν βοσκηθεί. Σημειώστε τη χρήση της συνάρτησης *as.numeric* για να επιλέξετε τα διαγραμματικά σύμβολα και τα χρώματα. Πώς οι μεταχειρίσεις βόσκησης αντανακλώνται στα επίπεδα του παράγοντα;

```
levels(Grazing)
[1] "Grazed"   "Ungrazed"
```

Τώρα μπορούμε να χρησιμοποιήσουμε λογικούς δείκτες (σελ. 21) για να σχεδιάσουμε ευθείες γραμμικής παλινδρόμησης για τις δύο μεταχειρίσεις βόσκησης ξεχωριστά, χρησιμοποιώντας την εντολή *abline* (θα μπορούσαμε αντ'αυτής, να είχαμε χρησιμοποιήσει την εντολή *subset*):

```
abline(lm(Fruit[Grazing=="Grazed"]~Root[Grazing=="Grazed"]),lty=2,col="blue")
abline(lm(Fruit[Grazing=="Ungrazed"]~Root[Grazing=="Ungrazed"]),lty=2,col="red")
```

Σημειώστε τη χρήση της *as.numeric* για να επιλέξευτε τα σύμβολα και τα χρώματα, καθώς και τη χρήση των δεικτών στο πλαίσιο της συνάρτησης *abline* για να προσαρμόσετε γραμμικά μοντέλα παλινδρόμησης ξεχωριστά για κάθε επίπεδο της μεταχείρισης βόσκησης (θα μπορούσαμε αντ 'αυτού να έχουμε χρησιμοποιήσει την εντολή subset).



Το περίεργο σχετικά με αυτά τα δεδομένα είναι ότι η βόσκηση φαίνεται να αυξάνει την παραγωγή φρούτων, αποτέλεσμα που είναι ένα ιδιαίτερα αντι-διαισθητικό:

,

```
tapply(Fruit,Grazing, mean)
 Grazed   Ungrazed
67.9405    50.8805
```

Αν κάνετε ένα t test θα δείτε ότι αυτή η διαφορά είναι στατιστικώς σημαντική (p=0,027) (αν και όπως εξηγήθηκε προηγούμενα, σε αυτή την περίπτωση είναι λάθος να το κάνετε):

```
t.test(Fruit~Grazing)

        Welch Two Sample t-test

data: Fruit by Grazing
t = 2.304, df = 37.306, p-value = 0.02689
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
2.061464  32.058536
sample estimates:
mean in group Grazed mean in group Ungrazed
             67.9405                50.8805
```

Πολλά σημαντικά σημεία είναι άμεσα εμφανή από αυτή την αρχική ανάλυση:

- Φυτά διαφορετικού μεγέθους διατίθενται στις δύο θεραπείες.
- Τα φυτά που χρησιμοποιούνται για βόσκηση ήταν αρχικά μεγαλύτερα.

• Η γραμμή παλινδρόμησης για τα φυτά που δε βοσκήθηκαν είναι πάνω από την αντίστοιχη αυτών που βοσκήθηκαν.
• Οι γραμμές παλινδρόμησης είναι σχεδόν παράλληλες.
• Τα σημεία τομής(δεν φαίνονται προς τα αριστερά) είναι πιθανό να είναι σημαντικά διαφορετικά.

Κάθε ένα από αυτά τα σημεία θα εξεταστεί λεπτομερώς.

Για να κατανοήσουμε το αποτέλεσμα της ανάλυσης της συνδιακύμανσης είναι χρήσιμο να εργαστούμε μέσω χειροκίνητων υπολογισμών. Ξεκινάμε επεξεργαζόμενοι τα αθροίσματα, τα αθροίσματα των τετραγώνων και τα αθροίσματα των προϊόντων για το σύνολο των συνδυαζόμενων δεδομένων (40 ζεύγη αριθμών), και στη συνέχεια, για κάθε επεξεργασία χωριστά (20 ζεύγη αριθμών). Θα πρέπει να συμπληρώσουμε ένα πίνακα συνόλων, επειδή αυτό μας βοηθά να είμαστε πολύ καλά οργανωμένοι για τους υπολογισμούς αυτούς. Ελέγξτε για να δείτε πού (και γιατί), τα αθροίσματα και τα αθροίσματα των τετραγώνων των διαμέτρων των ριζών (οι τιμές x) και οι αποδόσεις σε φρούτα (οι τιμές y) έχουν καταχωριστεί στον πίνακα: Πρώτα, θα πρέπει να επεξεργαστούμε τα συνολικά ποσά βασιζόμενοι στο σύνολο των 40 σημεία δεδομένων.

```
sum(Root);sum(Root^2)

[1] 287.246
[1] 2148.172
sum(Fruit);sum(Fruit^2)

[1] 2376.42
[1] 164928.1
sum(Root*Fruit)

[1] 18263.16
```

Αυτά είναι τα περίφημα πέντε, που θα πρέπει να χρησιμοποιήσουμε άμεσα, ώστε να ολοκληρώσουμε το συνολική σύνοψη των δεδομένων. Τώρα επιλέγουμε τις διαμέτρους των ριζών τόσο για τα υπό βόσκηση όσο και για τα μη υπό βόσκηση φυτά και έπειτα τις αποδόσεις των δύο αυτών κατηγοριών σε φρούτα:

```
sum(Root[Grazing=="Grazed"]);sum(Root[Grazing=="Grazed"]^2)
[1] 166.188
[1] 1400.834
sum(Root[Grazing=="Ungrazed"]);sum(Root[Grazing=="Ungrazed"]^2)
[1] 121.058
[1] 747.3387
sum(Fruit[Grazing=="Grazed"]);sum(Fruit[Grazing=="Grazed"]^2)
[1] 1358.81
[1] 104156.0
sum(Fruit[Grazing=="Ungrazed"]);sum(Fruit[Grazing=="Ungrazed"]^2)
[1] 1017.61
[1] 60772.11
```

Τέλος, θέλουμε τα αθροίσματα των προϊόντων: πρώτα για τα φυτά που βοσκήθηκαν και στη συνέχεια για τα φυτά που δε βοσκήθηκαν:

```
sum(Root[Grazing=="Grazed"]*Fruit[Grazing=="Grazed"])
[1] 11753.64
sum(Root[Grazing=="Ungrazed"]*Fruit[Grazing=="Ungrazed"])
[1] 6509.522
```

Και εδώ είναι ο πίνακάς μας:

	Sums	Squares and products
x ungrazed	121.058	747.3387
y ungrazed	1017.61	60772.11
xy ungrazed		6509.522
x grazed	166.188	1400.834
y grazed	1358.81	104156.0
xy grazed		11753.64
x overall	287.246	2148.172
y overall	2376.42	164928.1
xy overall		18263.16

Τώρα έχουμε όλες τις απαραίτητες πληροφορίες για την εκτέλεση των υπολογισμών των διορθωμένων αθροισμάτων των τετραγώνων και των προϊόντων, τα SSY, SSX και SSXY, για το σύνολο των δεδομένων (n = 40) και για τις δύο χωριστές μεταχειρίσεις (με 20

επαναλήψεις σε κάθε μία). Για να πάρετε τη σωστή απάντηση θα πρέπει να είστε εξαιρετικά μεθοδικοί, αλλά δεν υπάρχει τίποτα το μυστηριώδες ή δύσκολο σε αυτή τη διαδικασία. Πρώτα, υπολογίστε τις στατιστικές παλινδρόμησης για το σύνολο του πειράματος, αγνοώντας τη μεταχείριση της βόσκησης, χρησιμοποιώντας τα περίφημα πέντε που μόλις υπολογίσαμε:

$$SSY = 164\,928.1 - \frac{2376.42^2}{40} = 23\,743.84,$$

$$SSX = 2148.172 - \frac{287.246^2}{40} = 85.4158,$$

$$SSXY = 18\,263.16 - \frac{287.246 \times 2376.42}{40} = 1197.731,$$

$$SSR = \frac{1197.731^2}{85.4158} = 16\,795,$$

$$SSE = 23\,743.84 - 16\,795 = 6948.835.$$

Η επίδραση των διαφορών μεταξύ των δύο μεταχειρίσεων βόσκησης, το SSA, είναι:

$$SSA = \frac{1358.81^2 + 1017.61^2}{20} - \frac{2376.42^2}{40} = 2910.436.$$

Στη συνέχεια υπολογίζουμε τα στατιστικά στοιχεία παλινδρόμησης για καθεμία από τις μεταχειρίσεις βόσκησης χωριστά. Πρώτον για τα φυτά που υπόκεινται σε βόσκηση:

$$SSY_g = 104\,156 - \frac{1358.81^2}{20} = 11\,837.79,$$

$$SSX_g = 1400.834 - \frac{166.188^2}{20} = 19.9111,$$

$$SSXY_g = 11\,753.64 - \frac{1358.81 \times 166.188}{20} = 462.7415,$$

$$SSR_g = \frac{462.7415^2}{19.9111} = 10\,754.29,$$

$$SSE_g = 11837.79 - 10754.29 = 1083.509,$$

οπότε η κλίση της γραφικής παράστασης των φρούτων σε σχέση με την ρίζα για τα φυτά που υπόκεινται σε βόσκηση δίνεται από τον τύπο:

$$b_g = \frac{SSXY_g}{SSX_g} = \frac{462.7415}{19.9111} = 23.240.$$

Τώρα, για τα φυτά που δε βοσκήθηκαν:

$$SSY_u = 60\,772.11 - \frac{1017.61^2}{20} = 8995.606,$$
$$SSX_u = 747.3387 - \frac{121.058^2}{20} = 14.58677,$$
$$SSXY_u = 6509.522 - \frac{121.058 \times 1017.61}{20} = 350.0302,$$
$$SSR_u = \frac{350.0302^2}{14.58677} = 8399.466,$$
$$SSE_u = 8995.606 - 8399.466 = 596.1403,$$

οπότε η κλίση της γραφικής παράστασης των φρούτων έναντι της ρίζας για τα φυτά που δε βοσκήθηκαν δίνεται από:

$$b_u = \frac{SSXY_u}{SSX_u} = \frac{350.0302}{14.58677} = 23.996$$

Τώρα προσθέστε τα στατιστικά στοιχεία παλινδρόμησης σε όλα τα επίπεδα του παράγοντα (βόσκηση και μη βόσκηση):

$$SSY_{g+u} = 11\,837.79 + 8995.606 = 20\,833.4,$$
$$SSX_{g+u} = 19.9111 + 14.58677 = 34.49788,$$
$$SSXY_{g+u} = 462.7415 + 350.0302 = 812.7717,$$
$$SSR_{g+u} = 10\,754.29 + 8399.436 = 19\,153.75,$$
$$SSE_{g+u} = 1083.509 + 596.1403 = 1684.461.$$

Το SSR για ένα μοντέλο με μια ενιαία κοινή κλίση δίνεται από τον τύπο:

$$SSR_c = \frac{(SSXY_{g+u})^2}{SSX_{g+u}} = \frac{812.7717^2}{34.49788} = 19\,148.94,$$

και η τιμή της ενιαίας κοινής κλίσης είναι:

$$b = \frac{SSXY_{g+u}}{SSX_{g+u}} = \frac{812.7717}{34.49788} = 23.560$$

Η διαφορά μεταξύ των δύο εκτιμήσεων του SSR $(SSR_{\text{diff}} = SSR_{\text{g+u}} - SSR_{\text{c}} = 19153.75 - 19148.94 = 4.81)$ είναι ένα μέτρο της σπουδαιότητας της διαφοράς μεταξύ των δύο τομών υπολογισμένη ξεχωριστά για κάθε επίπεδο παράγοντα. Τέλος, το SSE υπολογίζεται από τη διαφορά:

$$SSE = SSY - SSA - SSR_{\text{c}} - SSR_{\text{diff}}$$
$$= 23743.84 - 2910.44 - 19148.94 - 4.81 = 1679.65.$$

Τώρα μπορούμε να ολοκληρώσουμε τον πίνακα ANOVA για το πλήρες μοντέλο

Source	SS	d.f.	MS	*F*
Grazing	2910.44	1		
Root	19148.94	1		
Different slopes	4.81	1	4.81	n.s.
Error	1679.65	36	46.66	
Total	23743.84	39		

Οι βαθμοί ελευθερίας σφάλματος είναι 40-4 = 36 επειδή έχουμε υπολογίσει τέσσερις παραμέτρους από τα δεδομένα: δύο κλίσεις και δύο τομές. Έτσι, η διακύμανση του σφάλματος είναι 46,66 (= SSE/36). Καθώς η διαφορά ανάμεσα στις κλίσεις είναι ξεκάθαρα μη σημαντική $(F = 4.81/46.66 = 0.10)$ μπορούμε να προσαρμόσουμε ένα απλούστερο μοντέλο με μια κοινή κλίση της τάξης του 23,56. Το άθροισμα των τετραγώνων των διαφορών ανάμεσα στις κλίσεις (4,81) γίνεται τώρα μέρος του αθροίσματος των σφαλμάτων των τετραγώνων:

Source	SS	d.f.	MS	*F*
Grazing	2910.44	1	2910.44	63.9291
Root	19148.94	1	19148.94	420.6156
Error	1684.46	37	45.526	
Total	23743.84	39		

Αυτό είναι το ελάχιστα επαρκές μοντέλο. Και οι δύο όροι είναι πολύ σημαντικοί και δεν υπάρχουν περιττά επίπεδα παράγοντα.
Το επόμενο βήμα είναι να υπολογίσουμε τις κλίσεις για τις δύο παράλληλες γραμμές παλινδρόμησης. Αυτό επιτυγχάνεται ακριβώς όπως και πριν, με την αναδιάταξη της εξίσωσης της ευθείας γραμμής για να πάρουμε $a = y - bx.$. Για κάθε γραμμή μπορούμε να χρησιμοποιήσουμε τις μέσες τιμές των x και y, με την κοινή κλίση σε κάθε περίπτωση.

Έτσι,

$$a_1 = \overline{Y_1} - b\overline{X_1} = 50.88 - 23.56 \times 6.0529 = -91.7261,$$
$$a_2 = \overline{Y_2} - b\overline{X_2} = 67.94 - 23.56 \times 8.309 = -127.8294.$$

Αυτό αποδεικνύει ότι τα υπό βόσκηση φυτά παράγουν, κατά μέσο όρο, 36,1 mg φρούτων, δηλαδή λιγότερα από τα φυτά που δε βοσκήθηκαν (127,83-91,73).
Τέλος, θα πρέπει να υπολογίσουμε τα τυπικά σφάλματα για την κοινή κλίση παλινδρόμησης και για τη διαφορά της μέσης γονιμότητας μεταξύ των μεταχειρίσεων, με βάση την διακύμανση σφάλματος στο ελάχιστα επαρκές μοντέλο, όπως αυτά αναφέρονται στο παραπάνω πίνακα:

$$s^2 = \frac{1684.46}{37} = 45.526$$

Τα τυπικά σφάλματα λαμβάνονται ως εξής: Το τυπικό σφάλμα της κοινής κλίσης βρέθηκε με τον συνήθη τρόπο

$$se_b = \sqrt{\frac{s^2}{SSX_{g+u}}} = \sqrt{\frac{45.526}{19.9111 + 14.45667}} = 1.149.$$

Και το τυπικό σφάλμα της τομής της παλινδρόμησης για τη μεταχείριση βόσκησης έχει επίσης βρεθεί με τον συνήθη τρόπο:

$$se_a = \sqrt{s^2\left[\frac{1}{n} + \frac{(0-\overline{x})^2}{SSX_{g+u}}\right]} = \sqrt{45.526\left[\frac{1}{20} + \frac{8.3094^2}{34.498}\right]} = 9.664.$$

Είναι σαφές ότι το σημείο τομής των -127,829 είναι σημαντικά πολύ μικρότερο από το μηδέν (t=127.829/9.664=13.2), γεγονός που υποδηλώνει ότι υπάρχει ένα όριο στα μεγέθη του ριζώματος πριν να μπορεί να αρχίσει η αναπαραγωγή. Τέλος, το τυπικό σφάλμα της διαφοράς μεταξύ του ύψους των δύο γραμμών (η επίδραση της βόσκησης) δίνεται από τον τύπο:

$$se_{\hat{y}_u - \hat{y}_g} = \sqrt{s^2\left[\frac{2}{n} + \frac{(\overline{x}_1 - \overline{x}_2)^2}{SSX_{g+u}}\right]}$$

ο οποίος, υποκαθιστώντας τις τιμές της διακύμανσης σφάλματος και τα μέσα μεγέθη ριζώματος των φυτών στις δύο μεταχειρίσεις, γίνεται:

$$se_{\bar{y}_u - \bar{y}_g} = \sqrt{45.526\left[\frac{2}{20} + \frac{(6.0529 - 8.3094)^2}{34.498}\right]} = 3.357.$$

Αυτό υποδηλώνει ότι οι όποιες ευθείες διαφέρουν στο ύψος κατά περισσότερο από περίπου 2 x 3,357 = 6,66 mg βάρους θα πρέπει να θεωρηθούν ως σημαντικά διαφορετικές. Έτσι, η παρούσα διαφορά του 36,09 σαφώς αντιπροσωπεύει μια πολύ σημαντική μείωση της γονιμότητας που προκαλείται από τη βόσκηση (t = 10,83).

Οι υπολογισμοί με το χέρι ήταν μπερδεμένοι, αλλά στην R, η ανάλυση της συνδιακύμανσης είναι εξαιρετικά απλή χρησιμοποιώντας την εντολή lm. Η μεταβλητή απόκρισης είναι η γονιμότητα, και υπάρχει ένας πειραματικός παράγοντας (βόσκηση) με δύο επίπεδα (βόσκηση και μη βόσκηση) και μία συμμεταβλητή (αρχική διάμετρος ρίζας). Υπάρχουν 40 τιμές για κάθε μία από αυτές τις μεταβλητές. Όπως είδαμε νωρίτερα, τα μεγαλύτερα φυτά διατέθηκαν σε διαδικασία βόσκησης, αλλά για μια δεδομένη διάμετρο ριζώματος (ας πούμε 7 mm), και το γράφημα σκέδασης δείχνει ότι τα υπό βόσκηση φυτά παρήγαγαν λιγότερα φρούτα από τα φυτά που δε βοσκήθηκαν (δεν προτείνεται κάτι περισσότερο από μια απλή σύγκριση των μέσων). Πρόκειται για ένα εξαιρετικό παράδειγμα της αξίας της ανάλυση της συνδιακύμανσης. Εδώ, η σωστή ανάλυση με τη χρήση της ANCOVA αντιστρέφει πλήρως την ερμηνεία μας για τα δεδομένα.
Η ανάλυση προχωρά με τον ακόλουθο τρόπο. Αρχικά, προσαρμόζουμε το πιο πολύπλοκο μοντέλο, στη συνέχεια το απλοποιούμε αίροντας τους μη σημαντικούς όρους μέχρι να μείνουμε με ένα ελάχιστα επαρκές μοντέλο, στο οποίο όλες οι παράμετροι είναι σημαντικά διαφορετικές από το μηδέν. Για την ANCOVA, το πιο πολύπλοκο μοντέλο έχει διαφορετικές κλίσεις και τομές για κάθε επίπεδο παράγοντα. Εδώ έχουμε ένα παράγοντα δύο επιπέδων (βόσκηση και μη βόσκηση) και προσαρμόζουμε ένα γραμμικό μοντέλο με δύο παραμέτρους (y=a+bx) οπότε, το πιο περίπλοκο μοντέλο έχει τέσσερις παραμέτρους (δύο κλίσεις και δύο τομές). Για να προσαρμόσουμε τις διαφορετικές κλίσεις και τομές χρησιμοποιούμε το συμβολισμό του αστερίσκου *:

```
ancova <- lm(Fruit~Grazing*Root)
```

Θα πρέπει να συνειδητοποιήσετε ότι η σειρά έχει σημασία: θα παίρναμε διαφορετικά αποτελέσματα, αν το μοντέλο είχε γραφτεί ως Fruit ~ Root * Grazing (περισσότερα σχετικά στη σελ. 507).

```
summary(ancova)

Coefficients:
                    Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)         −125.173      12.811   −9.771  1.15e-11  ***
GrazingUngrazed       30.806      16.842    1.829    0.0757  .
Root                  23.240       1.531   15.182    < 2e-16  ***
GrazingUngrazed:Root   0.756       2.354    0.321    0.7500
```

Αυτό δείχνει ότι το αρχικό μέγεθος της ρίζας έχει μια τεράστια επίδραση στην παραγωγή φρούτων (t = 15,182), αλλά δεν υπάρχει καμία ένδειξη οποιασδήποτε διαφοράς στην κλίση αυτής της σχέσης μεταξύ των δύο μεταχειρίσεων βόσκησης (αυτή είναι η αλληλεπίδραση της βόσκησης και της ρίζας με t = 0,321, p>>0,05). Ο πίνακας ANOVA για το μέγιστο μοντέλο μοιάζει με αυτόν:

```
anova(ancova)

Analysis of Variance Table

Response: Fruit
             Df  Sum Sq  Mean Sq  F value     Pr(>F)
Grazing       1  2910.4   2910.4  62.3795  2.262e-09 ***
Root          1 19148.9  19148.9 410.4201   < 2.2e-16 ***
Grazing:Root  1     4.8      4.8   0.1031       0.75
Residuals    36  1679.6     46.7
```

Το επόμενο βήμα είναι να διαγράψουμε τον όρο της μη σημαντικής αλληλεπίδρασης από το μοντέλο. Αυτό μπορούμε να το κάνουμε χειροκίνητα ή αυτόματα: εδώ θα κάνουμε και τα δύο για τους σκοπούς της παρουσίασης. Η συνάρτηση για τη χειροκίνητη απλούστευση του μοντέλου είναι εκσυγχρονισμένη. Προσαρμόζουμε το τρέχον μοντέλο (εδώ ονομάζεται ancova) διαγράφοντας από αυτό όρους. Η σύνταξη είναι σημαντική: η στίξη διαβάζει «κόμμα περισπωμένη τελεία μείον». Ορίζουμε ένα νέο όνομα για το απλοποιημένο μοντέλο:

```
ancova2<-update(ancova, ~ . - Grazing:Root)
```

Τώρα συγκρίνουμε το απλοποιημένο μοντέλο με τις μόλις τρεις παραμέτρους (μία κλίση και δύο τομές) με το μέγιστο μοντέλο χρησιμοποιώντας την anova ως ακολούθως:

```
anova(ancova,ancova2)

Analysis of Variance Table

Model 1: Fruit ~ Grazing * Root
Model 2: Fruit ~ Grazing + Root
  Res.Df     RSS  Df  Sum of Sq      F  Pr(>F)
1     36 1679.65
2     37 1684.46  −1      −4.81 0.1031    0.75
```

Αυτό μας λέει ότι η απλοποίηση του μοντέλου ήταν δικαιολογημένη, επειδή προκάλεσε μια αμελητέα μείωση στην επεξηγηματική του ισχύ (p =0,75). Για να διατηρήσουμε τον όρο αλληλεπίδρασης στο μοντέλο θα πρέπει να έχουμε (p<0,05).
Το επόμενο βήμα στην απλοποίηση του μοντέλου περιλαμβάνει τον έλεγχο του κατά πόσον η βόσκηση είχε μια σημαντική επίδραση στην παραγωγή φρούτων αφού έχουμε τον έλεγχο του αρχικού μεγέθους της ρίζας. Η διαδικασία είναι παρόμοια: ορίζουμε ένα νέο μοντέλο και χρησιμοποιούμε την εντολή *update* για να αφαιρέσουμε τη βόσκηση από την ancova2 ως ακολούθως:

```
ancova3<-update(ancova2, ~ . - Grazing)
```

Τώρα συγκρίνουμε τα δύο μοντέλα χρησιμοποιώντας την *anova*

```
anova(ancova2,ancova3)
Analysis of Variance Table

Model 1: Fruit ~ Grazing + Root
Model 2: Fruit ~ Root
  Res.Df     RSS  Df  Sum of Sq      F     Pr(>F)
1     37  1684.5
2     38  6948.8  -1   -5264.4 115.63 6.107e-13  ***
```

Η απλοποίηση του μοντέλου είναι ακόμα πάρα πολύ μακριά. Η αφαίρεση του όρου της βόσκησης προκαλεί μια τεράστια μείωση στην ερμηνευτική ισχύ του μοντέλου, με μια τιμή F ίση με 115,63 και με μία εξουδετερωτικά μικρή τιμή p. Η επίδραση της βόσκησης στη μείωση της παραγωγής φρούτων είναι εξαιρετικά σημαντική και για αυτό θα πρέπει να παραμείνει στο μοντέλο. Έτσι η *ancova2* είναι το ελάχιστα επαρκές μοντέλο μας, και θα πρέπει να εξετάσουμε το συνοπτικό του πίνακα για να συγκρίνουμε με τους προηγούμενους υπολογισμούς που κάναμε το χέρι:

```
summary(ancova2)
Coefficients:
                Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)     -127.829       9.664  -13.23   1.35e-15 ***
GrazingUngrazed   36.103       3.357   10.75   6.11e-13 ***
Root              23.560       1.149   20.51    < 2e-16 ***

Residual standard error: 6.747 on 37 degrees of freedom
Multiple R-Squared: 0.9291, Adjusted R-squared: 0.9252
F-statistic: 242.3 on 2 and 37 DF, p-value: < 2.2e-16
```

Ξέρετε ότι έχετε το ελάχιστα επαρκές μοντέλο, επειδή κάθε γραμμή του πίνακα συντελεστών του έχει ένα ή περισσότερα αστέρια σημαντικότητας (τρία σε αυτή την περίπτωση, επειδή οι επιδράσεις είναι όλες τόσο ισχυρές). Σε αντίθεση με την αρχική μας ερμηνεία που βασιζόταν στη μέση παραγωγή φρούτων, η βόσκηση συσχετίζεται με μία μείωση της τάξης των 36,103 mg στην παραγωγή φρούτων.

```
anova(ancova2)
```

```
Analysis of Variance Table

Response: Fruit
           Df   Sum Sq  Mean Sq  F value     Pr(>F)
Grazing     1   2910.4   2910.4   63.929  1.397e-09  ***
Root        1  19148.9  19148.9  420.616  < 2.2e-16  ***
Residuals  37   1684.5     45.5
```

Αυτές είναι οι τιμές που αποκομίσαμε από τον «μακρύ» δρόμο στη σελ. 495.
Τώρα επαναλαμβάνουμε την απλοποίηση του μοντέλου χρησιμοποιώντας τη συνάρτηση αυτοματοποιημένης απλοποίησης που ονομάζεται *step*. Δεν θα μπορούσε να είναι πιο εύκολη στη χρήση. Το πλήρες μοντέλο ονομάζεται ancova:

```
step(ancova)
```

Η συνάρτηση αυτή προκαλεί τον έλεγχο όλων των όρων για να διαπιστωθεί εάν αυτοί χρειάζονται στο ελάχιστα επαρκές μοντέλο. Το κριτήριο που χρησιμοποιείται είναι το AIC, το κριτήριο πληροφοριών Akaike (σελ. 353). Στην ορολογία, αυτό είναι μία «τιμωρημένη πιθανότητα». Με απλά λόγια αυτό σημαίνει ότι ζυγίζει το αναπόφευκτο trade-off μεταξύ των βαθμών ελευθερίας. Μπορείτε να έχετε μια τέλεια εφαρμογή του μοντέλου, αν έχετε μια παράμετρο για κάθε σημείο δεδομένων, αλλά αυτό το μοντέλο έχει μηδενική επεξηγηματική ισχύ. Έτσι η απόκλιση μειώνεται όσο οι βαθμοί ελευθερίας στο μοντέλο αυξάνονται. Το AIC προσθέτει δύο φορές τον αριθμό των παραμέτρων του μοντέλου στην απόκλιση (για να το τιμωρήσει). Η απόκλιση θα θυμάστε, είναι διπλάσια από την πιθανοφάνεια του τρέχοντος μοντέλου. Τέλος πάντων, το AIC είναι ένα μέτρο της έλλειψης προσαρμογής. Μεγάλο AIC είναι κακό ενώ μικρό AIC είναι καλό. Το πλήρες μοντέλο (τέσσερις παράμετροι: δύο κλίσεις και δύο τομές) έχει τοποθετηθεί πρώτο, και το AIC υπολογίζεται ίσο με 157,5:

```
Start:  AIC = 157.5
Fruit ~ Grazing * Root

               Df  Sum of Sq      RSS     AIC
- Grazing: Root   1      4.81  1684.46  155.61
<none>                        1679.65  157.50

Step:  AIC = 155.61
Fruit ~ Grazing + Root

          Df  Sum of Sq      RSS    AIC
<none>                   1684.5  155.6
- Grazing   1    5264.4   6948.8  210.3
- Root      1   19148.9  20833.4  254.2
```

```
Call:
lm(formula = Fruit ~ Grazing + Root)

Coefficients:
(Intercept)  GrazingUngrazed          Root
    -127.83           36.10         23.56
```

Στη συνέχεια, η συνάρτηση *step* προσπαθεί να αφαιρέσει τον πιο περίπλοκο όρο (την αλληλεπίδραση της βόσκησης με τη ρίζα). Αυτό μειώνει το AIC σε 155,61 (μια βελτίωση, έτσι ώστε η απλοποίηση να είναι δικαιολογημένη). Δεν είναι δυνατή η οποιαδήποτε περαιτέρω απλοποίηση (όπως είδαμε όταν χρησιμοποιήσαμε την εντολή *update* για να απομακρύνουμε τον όρο της βόσκησης από το μοντέλο), επειδή το AIC ανεβαίνει στο 210,3 όταν η βόσκηση απομακρύνεται και φτάνει μέχρι 254,2 εάν απομακρυνθεί το μέγεθος της ρίζας. Έτσι, η εντολή *step* έχει βρει το ελάχιστα επαρκές μοντέλο (όπως θα δούμε αργότερα αυτό δεν συμβαίνει πάντα, η εντολή είναι καλή, αλλά όχι τέλεια).

ANCOVA και Πειραματικός Σχεδιασμός

Υπάρχει ένα εξαιρετικά σημαντικό γενικό μήνυμα σε αυτό το παράδειγμα αναφορικά με τον σχεδιασμό του πειράματος. Ανεξάρτητα από το πόσο προσεκτικά τυχαιοποιούμε αρχικά, οι πειραματικές ομάδες μας, είναι πιθανό να είναι ετερογενείς. Μερικές φορές, όπως στην προκειμένη περίπτωση, μπορεί να έχουμε κάνει αρχικές μετρήσεις τις οποίες στη συνέχεια μπορούμε να χρησιμοποιήσουμε ως συμμεταβλητές, αλλά αυτό δε συμβαίνει πάντα. Είναι βέβαιο ότι θα υπάρχουν σημαντικοί παράγοντες που δε μετρήσαμε. Αν δεν είχαμε μετρήσει το αρχικό μέγεθος της ρίζας σε αυτό το παράδειγμα, θα είχαμε καταλήξει σε ένα εντελώς λανθασμένο συμπέρασμα σχετικά με την επίδραση της βόσκησης στην απόδοση των φυτών.

Μια πολύ καλύτερη σχεδίαση του πειράματος θα περιλάμβανε τη μέτρηση των διαμέτρων των ριζωμάτων για όλα τα φυτά κατά την έναρξη του πειράματος (όπως έγινε εδώ), αλλά στη συνέχεια θα έπρεπε να τοποθετήσουμε τα φυτά σε ομοιογενή ζεύγη με ριζώματα παρόμοιου μεγέθους. Στη συνέχεια, θα επιλέγαμε τυχαία ένα από τα φυτά και θα το κατανέμαμε σε μία από τις δύο μεταχειρίσεις βόσκησης (π.χ. πετώντας ένα νόμισμα). Το άλλο φυτό του ζεύγους θα έμπαινε στην μεταχείριση που απομένει. Σύμφωνα με αυτό το σύστημα, οι κατηγορίες μεγέθους των δύο μεταχειρίσεων θα επικαλύπτονταν, και η ανάλυση της συνδιακύμανσης θα ήταν περιττή.

Μία πιο σύνθετη ANCOVA: Δύο παράγοντες και μία συνεχής συμμεταβλητή

Το ακόλουθο πείραμα, έχει ως μεταβλητή απόκριση το βάρος, ως δύο κατηγορικές επεξηγηματικές μεταβλητές το φύλο, και ως μια συνεχή συμμεταβλητή τον γονότυπο και την ηλικία. Υπάρχουν έξι επίπεδα για τον γονότυπο και δύο επίπεδα για το φύλο.

```
Gain <-read.table("c:\\temp\\Gain.txt",header=T)
attach(Gain)
names(Gain)
```

```
[1] "Weight"  "Sex"  "Age"  "Genotype"  "Score"
```

Ξεκινάμε από την προσαρμογή του μέγιστου μοντέλου με τις 24 παραμέτρους του: διαφορετικές κλίσεις και σημεία τομής για κάθε συνδυασμό φύλου και γονότυπου.

```
m1<-lm(Weight~Sex*Age*Genotype)
summary(m1)
```

```
Coefficients:
                              Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)                    7.80053     0.24941   31.276    < 2e-16  ***
Sexmale                       -0.51966     0.35272   -1.473    0.14936
Age                            0.34950     0.07520    4.648   4.39e-05  ***
GenotypecloneB                 1.19870     0.35272    3.398    0.00167  **
GenotypecloneC                -0.41751     0.35272   -1.184    0.24429
GenotypecloneD                 0.95600     0.35272    2.710    0.01023  *
GenotypecloneE                -0.81604     0.35272   -2.314    0.02651  *
GenotypecloneF                 1.66851     0.35272    4.730   3.41e-05  ***
Sexmale:Age                   -0.11283     0.10635   -1.061    0.29579
Sexmale:GenotypecloneB        -0.31716     0.49882   -0.636    0.52891
Sexmale:GenotypecloneC        -1.06234     0.49882   -2.130    0.04010  *
Sexmale:GenotypecloneD        -0.73547     0.49882   -1.474    0.14906
Sexmale:GenotypecloneE        -0.28533     0.49882   -0.572    0.57087
Sexmale:GenotypecloneF        -0.19839     0.49882   -0.398    0.69319
Age:GenotypecloneB            -0.10146     0.10635   -0.954    0.34643
Age:GenotypecloneC            -0.20825     0.10635   -1.958    0.05799  .
Age:GenotypecloneD            -0.01757     0.10635   -0.165    0.86970
Age:GenotypecloneE            -0.03825     0.10635   -0.360    0.72123
Age:GenotypecloneF            -0.05512     0.10635   -0.518    0.60743
Sexmale:Age:GenotypecloneB     0.15469     0.15040    1.029    0.31055
Sexmale:Age:GenotypecloneC     0.35322     0.15040    2.349    0.02446  *
Sexmale:Age:GenotypecloneD     0.19227     0.15040    1.278    0.20929
Sexmale:Age:GenotypecloneE     0.13203     0.15040    0.878    0.38585
Sexmale:Age:GenotypecloneF     0.08709     0.15040    0.579    0.56616

Residual standard error: 0.2378 on 36 degrees of freedom
Multiple R-Squared: 0.9742, Adjusted R-squared: 0.9577
F-statistic: 59.06 on 23 and 36 DF, p-value: < 2.2e-16
```

Υπάρχουν μια ή δύο σημαντικές παράμετροι, αλλά δεν γίνεται καθόλου σαφές το αν οι τριπλές ή οι αμφίδρομες αλληλεπιδράσεις χρειάζεται να παραμείνουν στο μοντέλο. Σε ένα πρώτο στάδιο, ας χρησιμοποιήσουμε την εντολή *step* για να δούμε που μπορούμε να φτάσουμε αναφορικά με την απλοποίηση του μοντέλου:

m2<-step(m1)

```
Start: AIC= -155.01

Weight ~ Sex * Age * Genotype

                     Df  Sum of Sq    RSS      AIC
- Sex:Age:Genotype    5     0.349  2.385  -155.511
<none>                             2.036  -155.007

Step:  AIC= -155.51
Weight ~ Sex + Age + Genotype + Sex:Age + Sex:Genotype +
Age:Genotype

                 Df  Sum of Sq    RSS      AIC
- Sex:Genotype    5     0.147  2.532  -161.924
- Age:Genotype    5     0.168  2.553  -161.423
- Sex:Age         1     0.049  2.434  -156.292
<none>                        2.385  -155.511


Step:  AIC= -161.92
Weight ~ Sex + Age + Genotype + Sex:Age + Age:Genotype

                 Df  Sum of Sq    RSS      AIC
- Age:Genotype    5     0.168  2.700  -168.066
- Sex:Age         1     0.049  2.581  -162.776
<none>                        2.532  -161.924

Step:  AIC= -168.07
Weight ~ Sex + Age + Genotype + Sex:Age

             Df  Sum of Sq     RSS      AIC
- Sex:Age     1     0.049   2.749  -168.989
<none>                     2.700  -168.066
- Genotype    5    54.958  57.658     5.612

Step:  AIC= -168.99
Weight ~ Sex + Age + Genotype

             Df  Sum of Sq     RSS      AIC
<none>                     2.749  -168.989
- Sex         1    10.374  13.122   -77.201
- Age         1    10.770  13.519   -75.415
- Genotype    5    54.958  57.707     3.662

Call:
lm(formula = Weight ~ Sex + Age + Genotype)

Coefficients:
   (Intercept)  Sexmale     Age  GenotypeCloneB
GenotypeCloneC
        7.9370  -0.8316  0.2996          0.9678
 -1.0436
GenotypeCloneD  GenotypeCloneE  GenotypeCloneF
        0.8240         -0.8754          1.5346
```

Call:

Σίγουρα δεν χρειαζόμαστε την τριμερή αλληλεπίδραση, παρά την επίδραση της `Sexmale:Age:GenotypeCloneC` η οποία έδωσε από μόνη της ένα σημαντικό t test. Τι συμβαίνει όμως με τις διμερείς αλληλεπιδράσεις; Η *step* αφήνει έξω το φύλο ανάλογα με το γονότυπο και στη συνέχεια αξιολογεί τις άλλες δύο. Δε χρειάζεται το φύλο ανάλογα με το γονότυπο. Δοκιμάστε να αφαιρέσετε το φύλο ανάλογα με την ηλικία. Δε συμβαίνει τίποτα. Τι γίνεται με τις κύριες επιδράσεις; Αυτές είναι όλες εξαιρετικά σημαντικές. Αυτή είναι η ιδέα της R για το ελάχιστα επαρκές μοντέλο: τρεις κύριες επιδράσεις, αλλά καμία αλληλεπίδραση. Δηλαδή, η κλίση της καμπύλης της αύξησης του σωματικού βάρους σε σχέση με την ηλικία δεν διαφέρει ανάλογα με το φύλο ή τον γονότυπο, αλλά τα αντίστοιχα σημεία τομής διαφέρουν. Θα ήταν μια καλή ιδέα να κοιτάξουμε τον πίνακα *summary.lm* για αυτό το μοντέλο:

```
summary(m2)

Coefficients:
              Estimate   Std. Error    t value   Pr(>|t|)
(Intercept)    7.93701      0.10066     78.851    < 2e-16 ***
Sexmale       −0.83161      0.05937   −14.008     < 2e-16 ***
Age            0.29958      0.02099     14.273    < 2e-16 ***
GenotypeCloneB 0.96778      0.10282      9.412   8.07e-13 ***
GenotypeCloneC −1.04361     0.10282    −10.149   6.21e-14 ***
GenotypeCloneD 0.82396      0.10282      8.013   1.21e-10 ***
GenotypeCloneE −0.87540     0.10282     −8.514   1.98e-11 ***
GenotypeCloneF 1.53460      0.10282     14.925    < 2e-16 ***


Residual standard error: 0.2299 on 52 degrees of freedom
Multiple R-Squared: 0.9651, Adjusted R-squared: 0.9604
F-statistic: 205.7 on 7 and 52 DF, p-value: < 2.2e-16
```

Σε αυτήν την περίπτωση οι συγκρίσεις Helmert θα ήταν πραγματικά πρακτικές (βλ. σελ. 378). Όλα είναι διαφορετικά σε επίπεδο τριών αστεριών από τον γονότυπο. Γονότυπος [1] Φύλο [1], αλλά δεν είναι προφανές ότι τα σημεία τομής για τους γονότυπους B και D χρειάζονται διαφορετικές τιμές (+0,96 και +0,82 πάνω από τον γονότυπο A με $se_{\text{diff}} = 0.1028$), ούτε είναι προφανές ότι οι γονότυποι C και E έχουν διαφορετικές κλίσεις (-1,043 και -0,875).Θα μπορούσαμε να μειώσουμε τον αριθμό των επιπέδων του γονότυπου από έξι που έχουμε τώρα σε τέσσερις, χωρίς καμία απώλεια της επεξηγηματικής μας ισχύος;
Δημιουργούμε μια νέα κατηγορική μεταβλητή που ονομάζεται *newGenotype* με διαφορετικά επίπεδα για τους κλώνους A και F, B και D και C και E συνδυαστικά.

```
newGenotype<-Genotype
levels(newGenotype)
[1] "CloneA" "CloneB" "CloneC" "CloneD" "CloneE" "CloneF"
levels(newGenotype)[c(3,5)]<-"ClonesCandE"
levels(newGenotype)[c(2,4)]<-"ClonesBandD"
levels(newGenotype)
[1] "CloneA" "ClonesBandD" "ClonesCandE" "CloneF"
```

Στη συνέχεια, επαναλαμβάνουμε τη μοντελοποίηση χρησιμοποιώντας την εντολή *newGenotype* (4 επίπεδα) αντί της εντολής *Genotype* (6 επίπεδα):

```
m3<-lm(Weight~Sex+Age+newGenotype)
```

και ελέγχουμε ότι η απλοποίηση ήταν δικαιολογημένη

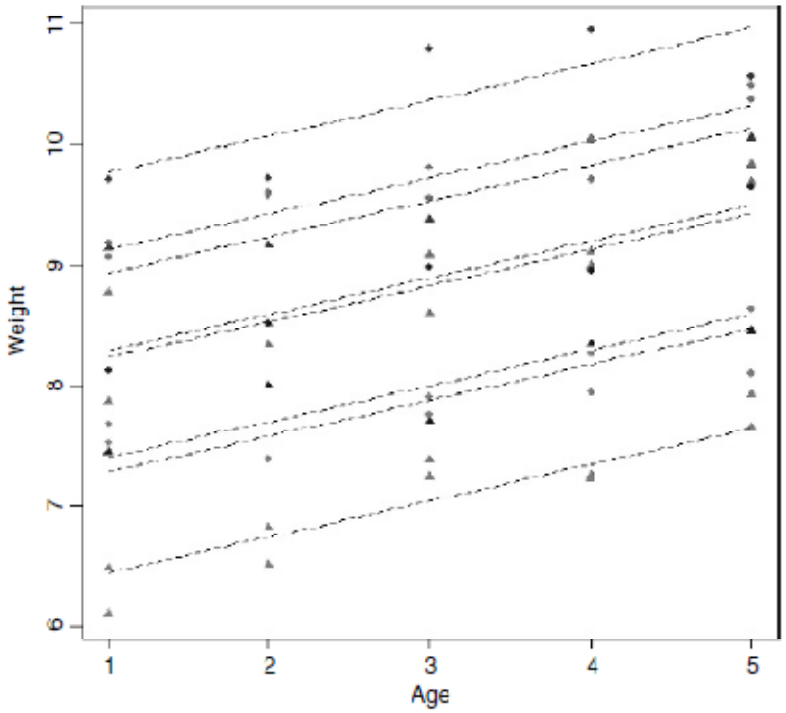
```
anova(m2,m3)
Analysis of Variance Table
Model 1: Weight ~ Sex + Age + Genotype
Model 2: Weight ~ Sex + Age + newGenotype
  Res.Df     RSS  Df  Sum of Sq      F  Pr(>F)
1     52 2.74890
2     54 2.99379  -2  -0.24489 2.3163 0.1087
```

Πράγματι, ήταν. Η τιμή p ήταν 0,1087 οπότε δεχόμαστε το απλούστερο μοντέλο m3:

```
Coefficients:
                        Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)              7.93701     0.10308   76.996   < 2e-16 ***
Sexmale                 -0.83161     0.06080  -13.679   < 2e-16 ***
Age                      0.29958     0.02149   13.938   < 2e-16 ***
newGenotypeClonesBandD   0.89587     0.09119    9.824  1.28e-13 ***
newGenotypeClonesCandE  -0.95950     0.09119  -10.522  1.10e-14 ***
newGenotypeCloneF        1.53460     0.10530   14.574   < 2e-16 ***

Residual standard error: 0.2355 on 54 degrees of freedom
Multiple R-Squared: 0.962, Adjusted R-squared: 0.9585
F-statistic: 273.7 on 5 and 54 DF, p-value: < 2.2e-16
```

Μετά από μια ανάλυση συνδιακύμανσης, είναι χρήσιμο να σχεδιάσουμε τις προσαρμοσμένες ευθείες σε μια γραφική παράσταση με κάθε επίπεδο του παράγοντα να αντιπροσωπεύεται από διαφορετικά σύμβολα σχεδίασης και διαφορετικούς τύπους ευθειών (βλ. σελ. 167.):

```
plot(Age,Weight,col=as.numeric(newGenotype),pch=(15+as.numeric(Sex)))
xv<-c(1,5)
for (i in 1:2) {
for (j in 1:4){
a<-coef(m3)[1]+(i>1)* coef(m3)[2]+(j>1)*coef(m3)[j+2];b<-coef(m3)[3]
yv<-a+b*xv
lines(xv,yv,lty=2)
} }
```



Σημειώστε τη χρήση του χρώματος για να αντιπροσωπεύσουμε τους τέσσερις τύπους γονότυπου `col=as.numeric(newGenotype)` και τα σύμβολα σχεδίασης για να αντιπροσωπεύσουν τα δύο φύλα `pch=(15+as.numeric(Sex))`. Μπορείτε να δείτε ότι τα αρσενικά (κύκλοι) είναι βαρύτερα από τα θηλυκά (τρίγωνα) σε όλους τους γονότυπους. Άλλες εντολές που πρέπει να λαμβάνουμε υπόψη μας στην αποτύπωση των αποτελεσμάτων της ANCOVA είναι οι split και augPred σε γραφικά πλέγματα.

Οι αντιθέσεις και οι παράμετροι των μοντέλων ANCOVA

Στην ανάλυση της συνδιακύμανσης, εκτιμούμε μια κλίση και ένα σημείο τομής για κάθε επίπεδο ενός ή περισσοτέρων παραγόντων. Ας υποθέσουμε ότι μοντελοποιούμε το βάρος (η μεταβλητή απόκρισης) ως συνάρτηση του φύλου και της ηλικίας, όπως απεικονίστηκε στη σ. 490. Η δυσκολία οφείλεται στο γεγονός ότι υπάρχουν αρκετοί διαφορετικοί τρόποι για να εκφράσουμε τις τιμές των τεσσάρων παραμέτρων στον πίνακα summary.lm:

• δύο κλίσεις και δύο σημεία τομής (όπως στις εξισώσεις στη σελ. 490).
• μια κλίση και μια διαφορά μεταξύ των σημείων τομής, και ένα σημείο τομής και μία διαφορά μεταξύ των κλίσεων.

• μια συνολική μέση κλίση, μια συνολική μέση τομή, μια διαφορά μεταξύ των κλίσεων και μία διαφορά μεταξύ των σημείων τομής.

Στη δεύτερη περίπτωση (δύο εκτιμήσεις και δύο διαφορές) πρέπει να πάρουμε μια απόφαση σχετικά με το πιο επίπεδο παράγοντα να συνδέσουμε με την εκτίμηση αυτή και ποιο επίπεδο της διαφοράς (π.χ πρέπει τα αρσενικά να εκφράζονται ως το σημείο τομής και τα θηλυκά ως η διαφορά μεταξύ των κλίσεων ή το αντίστροφο); Όταν τα επίπεδα του παράγοντα δεν είναι διαταγμένα (η τυπική περίπτωση), τότε η R παίρνει το επίπεδο του παράγοντα που έρχεται πρώτο στο αλφάβητο ως εκτίμηση και τα άλλα εκφράζονται ως διαφορές. Στο παράδειγμά μας, οι εκτιμήσεις των παραμέτρων θα ήταν τα θηλυκά, και οι παράμετροι για τα αρσενικά θα εκφράζονταν ως οι διαφορές από τις τιμές για τα θηλυκά, επειδή το «f» είναι πριν από το «m» στο αλφάβητο. Αυτό πρέπει να καταστεί σαφές με ένα παράδειγμα:

```
Ancovacontrasts <-read.table("c:\\temp\\Ancovacontrasts.txt",header=T)
attach(Ancovacontrasts)
names(Ancovacontrasts)
[1] "weight" "sex" "age"
```

Πρώτα επεξεργαζόμαστε τις δύο παλινδρομήσεις ξεχωριστά, ώστε να γνωρίζουμε τις τιμές των δύο κλίσεων και των δύο σημείων τομής:

```
lm(weight[sex=="male"]~age[sex=="male"])

Coefficients:
(Intercept)  age[sex == "male"]
      3.115                1.561

lm(weight[sex=="female"]~age[sex=="female"])

Coefficients:
(Intercept)  age[sex == "female"]
     1.9663                0.9962
```

Οπότε, το σημείο τομής για τα αρσενικά είναι 3,115 και το αντίστοιχο για τα θηλυκά είναι 1,966. Η διαφορά μεταξύ του πρώτου (θηλυκά) και του δεύτερου (αρσενικά) σημείου τομής ως εκ τούτου είναι

$$3.115 - 1.9266 = +1.1884.$$

Τώρα μπορούμε να κάνουμε μια συνολική παλινδρόμηση, αγνοώντας το φύλο:

```
lm(weight~age)

Coefficients:
(Intercept)     age
      2.541   1.279
```

Αυτό μας λέει ότι το μέσο σημείο τομής είναι 2,541 και η μέση κλίση είναι 1,279.
Στη συνέχεια μπορούμε να προχωρήσουμε σε μια ανάλυση της συνδιακύμανσης και να συγκρίνουμε τα αποτελέσματα που παράγονται από κάθε μία από τις τρεις διαφορετικές επιλογές που επιτρέπονται από την R: μεταχείριση (η προεπιλογή στην R και στη Glim), Helmert (η προεπιλογή στην S-PLUS), και άθροισμα. Πρώτα παρουσιάζεται η ανάλυση χρησιμοποιώντας αντίθετες μεταχειρίσεις, όπως χρησιμοποιήθηκε από τις R και Glim:

```
options(contrasts=c("contr.treatment", "contr.poly"))
model1<-lm(weight~age*sex)
summary(model1)

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    1.9663      0.6268    3.137   0.00636  ***
age            0.9962      0.1010    9.862  3.33e-08  ***
sexmale        1.1489      0.8864    1.296   0.21331
age:sexmale    0.5646      0.1429    3.952   0.00114  ***
```

Το σημείο τομής (1,9663) είναι το σημείο τομής για τα θηλυκά (επειδή το f είναι πριν από το m στο αλφάβητο). Η παράμετρος της ηλικίας (0,9962) είναι η κλίση της γραφικής παράστασης του βάρους σε σχέση με την ηλικία για τα θηλυκά. Η παράμετρος του φύλου (1,1489) είναι η διαφορά μεταξύ των σημείων τομής των θηλυκών και των αρσενικών (1.9663+1,1489=3,1152).Η αλληλεπίδραση μεταξύ της ηλικίας και του φύλου είναι η διαφορά μεταξύ των κλίσεων στα γραφήματα των θηλυκών και των αρσενικών (0,9962+0,5646=1,5608). Έτσι, με τις αντίθετες μεταχειρίσεις, οι παράμετροι (από το 1 έως το 4) αποτελούν ένα σημείο τομής, μια κλίση, μια διαφορά μεταξύ δύο σημείων τομής και μια διαφορά μεταξύ δύο κλίσεων. Στη στήλη του τυπικού σφάλματος βλέπουμε, από τη γραμμή 1 και προς τα κάτω, το τυπικό σφάλμα ενός σημείου τομής για μια παλινδρόμηση με τα θηλυκά μόνο (0,6268 με n=10, $\sum x^2 = 385$ and $SSX = 82.5$) το τυπικό σφάλμα της κλίσης μόνο για τα θηλυκά (0,1010, με SSX = 82, 5), το τυπικό σφάλμα της διαφοράς μεταξύ δύο κλίσεων καθεμία βασισμένη σε n = 10 σημεία δεδομένων $(\sqrt{2 \times 0.6268^2} = 0.8864)$ και το τυπικό σφάλμα της διαφορά μεταξύ δύο κλίσεων, καθεμία βασισμένη σε n = 10 σημεία δεδομένων;
$(\sqrt{2 \times 0.1010^2} = 0.1429).$
Οι τύποι υπολογισμού για αυτά τα τυπικά σφάλματα βρίσκονται στη σελ. 496. Πολλοί άνθρωποι είναι πιο άνετοι με αυτή τη μέθοδο παρουσίασης από ότι με τη μέθοδο των συγκρίσεων Helmert ή με το άθροισμα των συγκρίσεων.
Πάμε τώρα στην ανάλυση χρησιμοποιώντας τις συγκρίσεις Helmert:

```
options(contrasts=c("contr.helmert", "contr.poly"))
model2<-lm(weight~age*sex)
summary(model2)

Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  2.54073     0.44319    5.733  3.08e-05  ***
age          1.27851     0.07143   17.899  5.26e-12  ***
sex1         0.57445     0.44319    1.296   0.21331
age:sex1     0.28230     0.07143    3.952   0.00114  ***
```

Ας δούμε αν μπορούμε να επεξεργαστούμε το τι αντιπροσωπεύουν οι τέσσερις τιμές των παραμέτρων. Η πρώτη παράμετρος, 2,540 73 (επισημασμένη ως Intercept), είναι το σημείο τομής της συνολικής παλινδρόμησης, αγνοώντας το φύλο (βλ. παραπάνω). Η παράμετρος που επισημαίνεται ως ηλικία (1,27851) είναι μια κλίση επειδή η ηλικία είναι η συνεχής επεξηγηματική μεταβλητή μας. Και πάλι, θα δείτε ότι είναι η κλίση για την παλινδρόμηση του βάρους κατά ηλικία, αγνοώντας το φύλο. Η τρίτη παράμετρος, το φύλο (0,57445), πρέπει να έχει κάτι να κάνει με κλίσεις, επειδή το φύλο είναι η κατηγορική μας μεταβλητή. Αν θέλουμε να ανακατασκευάσουμε το δεύτερο σημείο τομής (για τα αρσενικά) θα πρέπει να προσθέσουμε 0,5744 στη συνολική τομή: 2,54073+0,57445=3,11518. Για να πάρουμε το σημείο τομής για τα θηλυκά θα πρέπει να το αφαιρέσουμε: 2,54073-0,57445=1.96628. Η τέταρτη παράμετρος (0,28230), είναι η ηλικία: το φύλο, είναι η διαφορά μεταξύ της συνολικής μέσης κλίσης (1,279) και της κλίσης των αρσενικών: 1,27851+0,28230=1,56081.. Για να πάρουμε την κλίση του βάρους σε σχέση με την ηλικία για τα θηλυκά θα πρέπει να αφαιρέσουμε τον όρο αλληλεπίδρασης από το όρος της ηλικίας: 1,27851-0,28230=0,99621.

Στη στήλη του τυπικού σφάλματος, από πάνω προς τα κάτω, μπορείτε να δείτε το τυπικό σφάλμα μιας τομής βασισμένο σε μια παλινδρόμηση με όλα τα 20 σημεία (η συνολική παλινδρόμηση, αγνοώντας το φύλο, που είναι ίση με 0,44319) και το τυπικό σφάλμα της κλίσης που βασίζεται σε μια παλινδρόμηση πάλι με για το σύνολο των 20 σημείων (0,071 43). Τα τυπικά σφάλματα των διαφορών (τόσο των τομών όσο και των κλίσεων) περιλαμβάνουν το ήμισυ της διαφοράς μεταξύ των τιμών των αρσενικών και των θηλυκών, γιατί με τις συγκρίσεις Helmert η διαφορά είναι μεταξύ της τιμής των αρσενικών και της συνολικής τιμής, και όχι μεταξύ των τιμών των αρσενικών και των θηλυκών. Έτσι, η τρίτη γραμμή έχει το τυπικό σφάλμα μιας διαφοράς μεταξύ της συνολικής τομής και της τομής για τα αρσενικά που βασίζεται σε μια παλινδρόμηση με 10 σημεία (0,44319=0,8864/2) και η κάτω σειρά έχει το τυπικό σφάλμα της διαφοράς μεταξύ της συνολικής κλίσης και της κλίσης για τα αρσενικά, με βάση μια παλινδρόμηση με 10 σημεία (0.1429/2=0.07143). Έτσι, οι τιμές στις δύο κάτω σειρές του πίνακα Helmert είναι απλά το ήμισυ των τιμών στις ίδιες γραμμές του πίνακα μεταχειρίσεων.

Το πλεονέκτημα των συγκρίσεων Helmert έγκειται στον έλεγχο των υποθέσεων σε πιο περίπλοκα μοντέλα από αυτό, επειδή είναι εύκολο να δούμε ποιους όρους πρέπει να διατηρήσουμε σε ένα απλοποιημένο μοντέλο ελέγχοντας τα επίπεδα της σπουδαιότητάς τους σε ένα πίνακα summary.lm. Το μειονέκτημα τους έγκειται στο ότι είναι πολύ πιο

δύσκολο να ανακατασκευάσουμε τις κλίσεις και τα σημεία τομής από τις τιμές των εκτιμώμενων παραμέτρων (βλ. επίσης σελ.. 378).

Τέλος, θα εξετάσουμε την τρίτη επιλογή που είναι το άθροισμα των συγκρίσεων:

```
options(contrasts=c("contr.sum", "contr.poly"))
model3<-lm(weight~age*sex)
summary(model3)

Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  2.54073     0.44319    5.733  3.08e-05  ***
age          1.27851     0.07143   17.899  5.26e-12  ***
sex1        -0.57445     0.44319   -1.296   0.21331
age:sex1    -0.28230     0.07143   -3.952   0.00114  ***
```

Οι δύο πρώτες εκτιμήσεις είναι οι ίδιες με εκείνες που παράγονται από τις συγκρίσεις Helmert: η συνολική κλίση και το συνολικό σημείο τομής του γραφήματος που σχετίζονται με το βάρος και με την ηλικία, αγνοώντας το φύλο. Η παράμετρος του φύλου (-0,57445) έχει αντίθετο πρόσημο σε σχέση με την αντίστοιχη του Helmert: αυτό δείχνει πώς να υπολογίσουμε το σημείο τομής για τα θηλυκά (το πρώτο) από το συνολικό σημείο τομής 2,54073-0,57445=1,96628. Ο όρος αλληλεπίδρασης έχει επίσης αντίθετο πρόσημο: για να πάρουμε την κλίση για τα θηλυκά, προσθέτουμε τον όρο αλληλεπίδρασης στην κλίση για την ηλικία: 1,27851-0,28230=0,99621. .

Τα τέσσερα τυπικά σφάλματα για το άθροισμα των συγκρίσεων είναι ακριβώς τα ίδια με εκείνα που προκύπτουν από τις συγκρίσεις Helmert (όπως εξηγήθηκε ανωτέρω).

Στη summary.aov η διάταξη έχει σημασία

Οι άνθρωποι συχνά μπερδεύονται με τον πίνακα ANOVA που παράγεται από τη summary.aov στην ανάλυση της συνδιακύμανσης. Συγκρίνουμε τους πίνακες που παράγονται από αυτά τα δύο μοντέλα:

```
summary.aov(lm(weight~sex*age))
```

```
            Df    Sum Sq   Mean Sq   F value      Pr(>F)
sex          1    90.492    90.492   107.498   1.657e-08   ***
age          1   269.705   269.705   320.389   5.257e-12   ***
sex:age      1    13.150    13.150    15.621    0.001141   ***
Residuals   16    13.469     0.842
```

summary.aov(lm(weight~age*sex))

```
            Df    Sum Sq   Mean Sq   F value      Pr(>F)
age          1   269.705   269.705   320.389   5.257e-12   ***
sex          1    90.492    90.492   107.498   1.657e-08   ***
age:sex      1    13.150    13.150    15.621    0.001141   ***
Residuals   16    13.469     0.842
```

Ακριβώς τα ίδια αθροίσματα τετραγώνων και ακριβώς οι ίδιες τιμές p. Κανένα πρόβλημα. Αλλά κοιτάξτε αυτά τα δύο μοντέλα από την σκοπιά του παραδείγματος των φυτών που αναλύθηκε λεπτομερώς νωρίτερα (σελ. 490):

summary.aov(lm(Fruit~Grazing*Root))

```
              Df    Sum Sq   Mean Sq    F value      Pr(>F)
Grazing        1    2910.4    2910.4    62.3795   2.262e-09   ***
Root           1   19148.9   19148.9   410.4201    <2.2e-16   ***
Grazing:Root   1       4.8       4.8     0.1031        0.75
Residuals     36    1679.6      46.7
```

summary.aov(lm(Fruit~Root*Grazing))

```
              Df    Sum Sq   Mean Sq    F value      Pr(>F)
Root           1   16795.0   16795.0   359.9681   < 2.2e-16   ***
Grazing        1    5264.4    5264.4   112.8316   1.209e-12   ***
Root:Grazing   1       4.8       4.8     0.1031        0.75
Residuals     36    1679.6      46.7
```

Στην περίπτωση αυτή η σειρά των μεταβλητών εντός του τύπου του μοντέλου έχει μια τεράστια επίδραση: αλλάζει το άθροισμα των τετραγώνων που συνδέεται με τις δύο κύριες επιδράσεις (το μέγεθος των ριζών είναι η συνεχής και η βόσκηση είναι η κατηγορική, βόσκηση ή μη βόσκηση) και επίσης αλλάζει τις τιμές p. Ο όρος της αλληλεπίδρασης, το εναπομείναν άθροισμα των τετραγώνων και η διακύμανση του σφάλματος παραμένουν αμετάβλητες. Οπότε, ποια είναι η διαφορά μεταξύ των δύο περιπτώσεων;

Στο πρώτο παράδειγμα, όπου η διάταξη ήταν άνευ σημασίας, οι τιμές χ της συνεχούς μεταβλητής (ηλικία) ήταν ταυτόσημες για τα δύο φύλα (υπάρχει μια τιμή για τα αρσενικά και μια τιμή για τα θηλυκά σε καθεμία από τις δέκα πειραματικά ελεγχόμενες ηλικίες):

table(sex,age)

```
        age
sex      1  2  3  4  5  6  7  8  9  10
 female  1  1  1  1  1  1  1  1  1   1
 male    1  1  1  1  1  1  1  1  1   1
```

Στο δεύτερο παράδειγμα, οι τιμές x (μέγεθος ριζών) ήταν διαφορετικές στις δύο μεταχειρίσεις και το μέσο μέγεθος ρίζας ήταν μεγαλύτερο για τα φυτά που χρησιμοποιήθηκαν για βόσκηση από ότι για τα φυτά που δεν χρησιμοποιήθηκαν τοιουτοτρόπως.

```
tapply(Root,Grazing, mean)
Grazed   Ungrazed
8.3094     6.0529
```

Κάθε φορά που οι τιμές χ είναι διαφορετικές σε διαφορετικά επίπεδα παράγοντα, και / ή όταν υπάρχει διαφορετική επαναληψιμότητα σε διαφορετικά επίπεδα παράγοντα, τότε τα SSX και η SSXY θα ποικίλουν από επίπεδο σε επίπεδο, γεγονός που θα επηρεάσει τον τρόπο με τον οποίο το άθροισμα των τετραγώνων κατανέμεται στις κύριες επιδράσεις. Δεν υπάρχει καμία συνέπεια για εσάς όσον αφορά στην ερμηνεία του μοντέλου, επειδή τα επίδραση μεγέθους και τα τυπικά σφάλματα στον πίνακα summary.lm μένουν ανεπηρέαστα:

```
summary(lm(Fruit~Root*Grazing))

Coefficients:
                    Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)         −125.173      12.811   −9.771  1.15e-11  ***
Root                  23.240       1.531   15.182    < 2e-16  ***
GrazingUngrazed       30.806      16.842    1.829    0.0757  .
Root:GrazingUngrazed   0.756       2.354    0.321    0.7500

summary(lm(Fruit~Grazing*Root))

Coefficients:
                    Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)         −125.173      12.811   −9.771  1.15e-11  ***
GrazingUngrazed       30.806      16.842    1.829    0.0757  .
Root                  23.240       1.531   15.182    < 2e-16  ***
GrazingUngrazed:Root   0.756       2.354    0.321    0.7500
```