Στατιστική Μοντελοποίηση Δεδομένων στη γλώσσα R

Μπιτζίνης Γιάννης

Ιούνιος 2013

Πτυχιακή εργασία για το Τμήμα Ηλεκτρονικών Μηχανικών Τ.Ε. της Σχολής Εφαρμοσμένων Επιστημών

Τεχνολογικό Εκπαιδευτικό Ίδρυμα Κρήτης

Επιβλέπων Καθηγητής Ανδρουλάκης Γιώργος

Περίληψη

Η R είναι μία γλώσσα υψηλού επιπέδου και ένα περιβάλλον για την ανάλυση δεδομένων και γραφικών. Ο σχεδιασμός της R ήταν σε μεγάλο βαθμό επηρεασμένος από δύο υπάρχουσες γλώσσες: την S των Becker, Chambers και Wilks και την Scheme του Sussman. Η προκύπτουσα γλώσσα είναι παρόμοια στην εμφάνιση με την S, αλλά η υποκείμενη υλοποίηση και σημασιολογία προέρχονται από την Scheme.
Η ιδέα είναι να εισαγάγει τους χρήστες για τις υποθέσεις που βρίσκονται πίσω από τις δοκιμές, ενισχύοντας μία κριτική προσέγγιση στην στατιστική μοντελοποίηση, χρησιμοποιώντας ελάχιστη ή καθόλου στατιστική θεωρία και προϋποθέτοντας μηδενικό μαθηματικό ή στατιστικό υπόβαθρο.

Abstract

R is a high-level language and an environment for data analysis and graphics. The design of R was heavily influenced by two existing languages: Becker, Chambers and Wilks' S and Sussman's Scheme. The resulting language is very similar in appearance to S, but the under- lying implementation and semantics are derived from Scheme.
The idea is to introduce users to the assumptions that lie behind the tests, fostering a critical approach to statistical modelling, but involving little or no statistical theory and assuming no background in mathematics or statistics.

Ευχαριστίες

Η ολοκλήρωση αυτής της πτυχιακής υλοποιήθηκε με την υποστήριξη ανθρώπων στους οποίους θα ήθελα να εκφράσω τις θερμότερες ευχαριστίες μου. Πρώτα στον καθηγητή κ Ανδρουλάκη Γιώργο και στη συνέχεια ευχαριστώ την μητέρα μου και την αδελφή μου για την συνεχή συμπαράσταση και κατανόηση όλων αυτών των χρόνων.

Περιεχόμενα

Στατιστική Μοντελοποίηση

Το πιο δύσκολο μέρος οποιασδήποτε στατιστικής εργασίας είναι να ξεκινήσετε. Και ένα από τα δυσκολότερα πράγματα για να ξεκινήσετε είναι να επιλέξετε το σωστό είδος της στατιστικής ανάλυσης. Η επιλογή εξαρτάται από τη φύση των δεδομένων σας και στη συγκεκριμένη ερώτηση που προσπαθούν να απαντήσουν. Το κλειδί είναι να καταλάβετε τι είδους μεταβλητή *απόκρισης* έχετε, και να γνωρίζετε τη φύση των *επεξηγηματικών* μεταβλητών σας. Η μεταβλητή απόκρισης είναι αυτό που εργάζεστε: είναι η μεταβλητή της οποίας τη διαφορά προσπαθείτε να καταλάβετε. Αυτή είναι η μεταβλητή που πηγαίνει στον άξονα *y* του γραφήματος. Η επεξηγηματική μεταβλητή πηγαίνει στον άξονα *x* του γραφήματος: σας ενδιαφέρει στο βαθμό στον οποίο η διακύμανση στη μεταβλητή απόκρισης, σχετίζεται με την διακύμανση στην επεξηγηματική μεταβλητή. Θα πρέπει επίσης να εξετάσετε τον *τρόπο* που οι μεταβλητές στην ανάλυσή σας, μετρούν αυτό που φιλοδοξούν να μετρήσουν. Μια συνεχής μέτρηση είναι μια μεταβλητή, όπως το ύψος ή το βάρος που μπορεί να πάρει οποιαδήποτε πραγματική αριθμημένη τιμή. Μια κατηγορική μεταβλητή είναι ένας παράγοντας με δύο ή περισσότερα επίπεδα: το φύλο είναι ένας παράγοντας με δύο επίπεδα (αρσενικό και θηλυκό), και το χρώμα μπορεί να είναι ένας παράγοντας με επτά επίπεδα (κόκκινο, πορτοκαλί, κίτρινο, πράσινο, μπλε, έντονο μπλέ και το βιολετί).

Είναι απαραίτητο, επομένως, να μπορείτε να απαντήσετε στα ακόλουθα ερωτήματα:

• Ποιές από τις μεταβλητές σας είναι η μεταβλητή απόκρισης;

• Ποιές είναι οι επεξηγηματικές μεταβλητές;

• Είναι οι επεξηγηματικές μεταβλητές συνεχείς ή κατηγορηματικές, ή ένα μίγμα και των δύο;
• Τι είδους μεταβλητή απόκρισης έχετε: είναι μια συνεχής μέτρηση, μια αρίθμηση, ένα ποσοστό, ένα χρόνο μετά τον θάνατο ή μια κατηγορία;

Αυτά τα απλά στοιχεία-κλειδιά θα σας οδηγήσουν στην κατάλληλη στατιστική μέθοδο:

Οι επεξηγηματικές μεταβλητές

(α) Όλες οι επεξηγηματικές συνεχείς μεταβλητές	**Παλινδρόμηση**
(β) Όλες οι επεξηγηματικές κατηγορηματικές μεταβλητές	**Ανάλυση διακύμανσης(ANOVA)**
(γ) Επεξηγηματικές μεταβλητές και συνεχείς και κατηγορηματικές	**Ανάλυση συνδιακύμανσης (ANCOVA)**

Η μεταβλητή απόκρισης

(α) Συνεχής	**Κανονική παλινδρόμηση, ANOVA ή ANCOVA**
(β) Ποσοστιαία	**Λογιστική παλινδρόμηση**
(γ) Μέτρησης	**Λογιστικά-γραμμικά μοντέλα**
(δ) Δυαδική	**Δυαδική λογιστική ανάλυση**
(ε) Ώρας θανάτου	**Ανάλυση επιβίωσης**

Το αντικείμενο είναι ο προσδιορισμός των τιμών των παραμέτρων σε ένα συγκεκριμένο μοντέλο που οδηγούν στην *καλύτερη προσαρμογή του μοντέλου στα δεδομένα*. Τα δεδομένα είναι ιερά και απαραβίαστα, και μας λένε τι πραγματικά συνέβη κάτω από ένα δεδομένο σύνολο περιστάσεων. Είναι κοινό λάθος να πει κανείς 'τα στοιχεία προσαρμόστηκαν στο μοντέλο' σαν τα δεδομένα να ήταν ευέλικτα, και να είχαμε μια σαφή εικόνα της δομής του μοντέλου. Αντίθετα, αυτό που ψάχνουμε είναι το ελάχιστο επαρκές μοντέλο για να περιγράψει τα δεδομένα. Το μοντέλο είναι προσαρμοσμένο στα στοιχεία, όχι το αντίθετο. Το καλύτερο μοντέλο είναι το μοντέλο που παράγει τη λιγότερη ανεξήγητη μεταβολή (*η ελάχιστη εναπομένουσα αποκλίνουσα συμπεριφορά*), υπό τον περιορισμό ότι όλες οι παράμετροι του μοντέλου θα πρέπει να είναι στατιστικά σημαντικές.

Θα πρέπει να προσδιορίζεται το μοντέλο. Εμπεριέχει την μηχανιστική σας κατανόηση των σχετικών επεξηγηματικών μεταβλητών, καθώς και τον τρόπο που αυτές συνδέονται με τη μεταβλητή απόκρισης. Θέλετε το μοντέλο να είναι *απλό* λόγω της αρχής της φειδούς, καθώς και *επαρκές*, διότι δεν υπάρχει νόημα στη διατήρηση ενός ανεπαρκούς μοντέλου που δεν περιγράφει ένα σημαντικό μέρος της διακύμανσης των δεδομένων. Είναι πολύ σημαντικό να γίνει κατανοητό ότι *δεν υπάρχει ένα μοντέλο*: αυτό είναι ένα από τα κοινά σφάλματα που σιωπηρά εμπλέκονται στην παραδοσιακή παλινδρόμηση και την ανάλυση διακύμανσης (ANOVA), όπου χρησιμοποιούνται τα ίδια μοντέλα, συχνά αλόγιστα, ξανά και ξανά. Στις περισσότερες περιπτώσεις, θα υπάρξει ένας μεγάλος αριθμός διαφορετικών, περισσότερο ή λιγότερο αξιόπιστων μοντέλων που θα μπορούσαν να προσαρμοστούν σε οποιοδήποτε σύνολο δεδομένων που έχει δοθεί. Μέρος της εργασίας της ανάλυσης των δεδομένων είναι να καθοριστεί ποιά, εάν υπάρχουν, από τα πιθανά μοντέλα είναι κατάλληλα και, στη συνέχεια, από το σύνολο των κατάλληλων μοντέλων, ποιά είναι τα ελάχιστα επαρκή μοντέλα. Σε ορισμένες περιπτώσεις μπορεί να μην υπάρχει ενιαίο καλύτερο μοντέλο και ένα σύνολο από διαφορετικά μοντέλα, να μπορεί να περιγράψει όλα τα δεδομένα εξίσου καλά (ή εξίσου ανεπαρκώς, αν η διακύμανση είναι μεγάλη).

Μέγιστη Πιθανότητα

Τί, ακριβώς, εννοούμε όταν λέμε ότι οι τιμές των παραμέτρων θα πρέπει να δώσουν την 'καλύτερη προσαρμογή του μοντέλου στα δεδομένα'; Η σύμβαση που υιοθετούμε είναι ότι οι τεχνικές μας θα πρέπει να οδηγήσουν σε **αμερόληπτους, εκτιμητές της ελαχιστοποίησης της διακύμανσης**. Ορίζουμε 'καλύτερο' από την άποψη της **μέγιστης πιθανότητας**. Αυτή η έννοια μπορεί να είναι άγνωστη, γι' αυτό αξίζει να επενδύσει κανείς κάποιο χρόνο για να πάρει μια ιδέα για αυτήν. Δείτε πώς λειτουργεί:

• δίνοντας τα δεδομένα,

• και με δεδομένη την επιλογή μας για το μοντέλο,

• τι τιμές των παραμέτρων του μοντέλου κάνουν τα παρατηρούμενα δεδομένα πιο πιθανά;

Εμείς κρίνουμε το μοντέλο με βάση το πόσο σωστά θα ήταν τα πιθανά δεδομένα, αν ήταν και το μοντέλο.

Η Αρχή της Οικονομίας (Το ξυράφι του Occam)

Ένα από τα πιο σημαντικά θέματα που υπάρχει μέσα από αυτό το βιβλίο αφορά το μοντέλο της απλοποίησης. Η αρχή της οικονομίας αποδίδεται στις αρχές του 14ου αιώνα, στον Άγγλο φιλόσοφο νομιναλιστή, William του Occam, ο οποίος επέμεινε ότι, λαμβάνοντας υπόψη μια σειρά από εξίσου καλές εξηγήσεις για ένα συγκεκριμένο φαινόμενο, *η σωστή εξήγηση είναι η πιο απλή εξήγηση*. Λέγεται ξυράφι του Occam γιατί 'ξύρισε' τις επεξηγήσεις του στο ελάχιστο: η άποψή του ήταν ότι εξηγόντας κάτι, οι υποθέσεις δεν πρέπει να πολλαπλασιάζονται χωρίς λόγο. Ειδικότερα, για τους σκοπούς της εξήγησης, τα πράγματα που *δεν είναι γνωστά* ότι υπάρχουν δεν πρέπει, εκτός αν είναι απολύτως απαραίτητο, να υποτεθούν σαν υπάρχοντα. Για την στατιστική μοντελοποίηση, η αρχή της οικονομίας σημαίνει ότι:

• Τα μοντέλα πρέπει να έχουν όσο το δυνατόν λιγότερους παραμέτρους·

• γραμμικά μοντέλα θα πρέπει να προτιμώνται σε σχέση με μη-γραμμικά μοντέλα·

• πειράματα που στηρίζονται σε λίγες υποθέσεις θα πρέπει να προτιμώνται σε συστήματα που βασίζονται σε πολλά·

• Τα μοντέλα θα πρέπει να περικοπτόνται έως ότου είναι *ελάχιστα επαρκή*·

• απλές εξηγήσεις θα πρέπει να προτιμώνται αντί πολύπλοκων.

Η διαδικασία της απλούστευσης ενός μοντέλου αποτελεί αναπόσπαστο μέρος του ελέγχου υποθέσεων στην γλώσσα προγραμματισμού R. Σε γενικές γραμμές, μια μεταβλητή συγκρατείται στο μοντέλο μόνον *αν προκαλεί μια σημαντική αύξηση στην αποκλίνουσα συμπεριφορά όταν αφαιρείται από το τρέχον μοντέλο*. Επιδιώξτε την απλότητα, τότε λοιπόν μην την πιστεύετε.
Στο ζήλο μας για την απλοποίηση του μοντέλου, όμως, πρέπει να είμαστε προσεκτικοί για να μην πετάξουμε το μωρό μαζί με το νερό της μπανιέρας. Ο Αϊνστάιν έκανε μια χαρακτηριστική λεπτή τροποποίηση στο ξυράφι του Occam. Είπε: ' Ένα μοντέλο πρέπει να είναι όσο το δυνατόν απλό. Αλλά όχι απλούστερο.' Να θυμάστε, επίσης, τι είπε ο Oscar Wilde: 'η αλήθεια είναι σπάνια αγνή και ποτέ απλή.'

Τύποι στατιστικών μοντέλων

Η προσαρμογή των μοντέλων στα δεδομένα είναι η κεντρική λειτουργία της γλώσσας R. Η διαδικασία είναι ουσιαστικά υπό εξερεύνηση : δεν υπάρχουν σταθεροί και απόλυτα σε κανόνες. Το αντικείμενο είναι να καθοριστεί ένα ελάχιστο επαρκές μοντέλο (βλέπε Πίνακα 1) από το μεγάλο σύνολο των πιθανών μοντέλων, που θα μπορούσε να χρησιμοποιηθεί, για να περιγράψει το σύνολο δεδομένων που έχει δοθεί. Σε αυτό το βιβλίο θα συζητήσουμε πέντε τύπους του μοντέλου:

• το μηδενικό μοντέλο·

• το ελάχιστα επαρκές μοντέλο·

• το ισχύων μοντέλο·

• το μέγιστο μοντέλο· και

• το κορεσμένο μοντέλο.

Η σταδιακή εξέλιξη από το κορεσμένο μοντέλο (ή το μέγιστο μοντέλο, όποιο είναι κατάλληλο) μέσω μιας σειράς απλουστεύσεων, στο ελάχιστο επαρκές μοντέλο γίνεται στη βάση των **δοκιμών διαγραφής**. Αυτά είναι F-τεστ ή x^2 τεστ που αξιολογούν τη σημασία της αύξησης της αποκλίνουσας συμπεριφοράς που προκύπτει όταν ένας συγκεκριμένος όρος αφαιρείται από το ισχύων μοντέλο.

Πίνακας 1. Η στατιστική μοντελοποίηση περιλαμβάνει την επιλογή ενός ελάχιστα επαρκούς μοντέλου από μια δυνητικά μεγάλη σειρά από πιο πολύπλοκα μοντέλα, χρησιμοποιώντας την σταδιακή απλούστευση μοντέλου.

Μοντέλο	**Ερμηνεία**
Κορεσμένο μοντέλο	Μία παράμετρος για κάθε σημείο δεδομένων Προσαρμογή: τέλεια Βαθμοί ελευθερίας: κανένας Επεξηγηματική δύναμη του μοντέλου: καμία
Μέγιστο μοντέλο	Περιέχει όλους τους (p) παράγοντες, τις αλληλεπιδράσεις και τους συμπαράγοντες που μπορούν να είναι οποιουδήποτε ενδιαφέροντος. Πολλοί από τους όρους του μοντέλου είναι πιθανό να είναι ασήμαντοι Βαθμοί ελευθερίας: $n - p - 1$ Επεξηγηματική δύναμη του μοντέλου: εξαρτάται
Ελάχιστο επαρκές μοντέλο	Ένα απλοποιημένο μοντέλο με $0 \le p' \le p$ παραμέτρους

	Προσαρμογή: μικρότερη από το μέγιστο μοντέλο, αλλά όχι σημαντική Βαθμοί ελευθερίας: $n - p' - 1$ Επεξηγηματική δύναμη του μοντέλου: $r^2 = SSR/SSY$
Μηδενικό μοντέλο	Μόλις μία παράμετρο, η συνολική μέση τιμή είναι $\bar{y}$ Προσαρμογή: καμία: $SSE = SSY$ Βαθμοί ελευθερίας: $n - 1$ Επεξηγηματική δύναμη του μοντέλου: καμία

Τα μοντέλα είναι αναπαραστάσεις της πραγματικότητας που θα πρέπει να είναι το ίδιο ακριβή και βολικά. Ωστόσο, είναι αδύνατο να μεγιστοποιηθεί ο ρεαλισμός ενός μοντέλου, η γενικότητα και η ολότητα του ταυτόχρονα, και η αρχή της οικονομίας είναι ένα ζωτικής σημασίας εργαλείο που βοηθά να επιλέξετε ένα μοντέλο πάνω από ένα άλλο. Έτσι, θα περιλάβουμε μόνο μία επεξηγηματική μεταβλητή σε ένα μοντέλο, αν βελτίωνε σημαντικά την προσαρμογή του μοντέλου. Το γεγονός ότι πήγαμε στο πρόβλημα της μέτρησης δεν σημαίνει ότι πρέπει να το έχουμε στο μοντέλο μας. Η φιλαργυρία λέει ότι, με το να είναι ίσα άλλα πράγματα, εμείς προτιμούμε:

• ένα μοντέλο με $n - 1$ παραμέτρους από ένα μοντέλο με n παραμέτρους·

• ένα μοντέλο με $k - 1$ επεξηγηματικές μεταβλητές από ένα μοντέλο με k επεξηγηματικές μεταβλητές·

• ένα γραμμικό μοντέλο από ένα μοντέλο το οποίο είναι κυρτό·

• ένα μη κυρτό μοντέλο από ένα κυρτό·

• ένα μοντέλο χωρίς αλληλεπιδράσεις από ένα μοντέλο που περιέχει τις αλληλεπιδράσεις μεταξύ των παραγόντων.

Άλλες εκτιμήσεις περιλαμβάνουν την προτίμησή τους για τα μοντέλα που περιέχουν επεξηγηματικές μεταβλητές που είναι εύκολο να μετρηθούν πάνω από τις μεταβλητές που είναι δύσκολες ή ακριβές να μετρηθούν. Επίσης, προτιμάμε τα μοντέλα που βασίζονται στην καλή μηχανιστική κατανόηση της διαδικασίας, πάνω από καθαρά εμπειρικές συναρτήσεις.

Η οικονομία απαιτεί ότι το μοντέλο θα πρέπει να είναι όσο το δυνατόν απλούστερο. Αυτό σημαίνει ότι το μοντέλο δεν πρέπει να περιέχει καμία περιττή παραμέτρο ή επίπεδα παραγόντων. Αυτό το επιτυγχάνουμε με την προσαρμογή ενός μέγιστου μοντέλου και στη συνέχεια απλοποιώντας το, ακολουθώντας ένα ή περισσότερα από τα ακόλουθα βήματα:

• άρση των μη σημαντικών όρων αλληλεπίδρασης·

• αφαίρεση μη σημαντικών τετραγωνικών ή άλλων μη-γραμμικών όρων·

• την απομάκρυνση των μη σημαντικών επεξηγηματικών μεταβλητών·

• ομαδοποιώντας τα επίπεδα των παραγόντων που δεν διαφέρουν το ένα το άλλο·

• στην ανάλυση συνδιακύμανσης (ANCOVA), να μην θέσετε σημαντικές κλίσεις στη συνεχή επεξηγηματική μεταβλητή στο μηδέν.

Όλα τα παραπάνω υπόκεινται, βεβαίως, με τις επιφυλάξεις που οι απλουστεύσεις κάνουν καλή επιστημονική έννοια του όρου και δεν οδηγούν σε σημαντικές μειώσεις της επεξηγηματικής ισχύος.

Ακριβώς όπως δεν υπάρχει τέλειο μοντέλο, οπότε μπορεί να μην υπάρχει βέλτιστη κλίμακα μέτρησης για ένα μοντέλο. Ας υποθέσουμε, για παράδειγμα, ότι είχαμε μια διαδικασία που είχε Poisson λάθη με πολλαπλασιαστικά αποτελέσματα μεταξύ των επεξηγηματικών μεταβλητών. Στη συνέχεια, πρέπει κανείς να επιλέξει μεταξύ τριών διαφορετικών κλιμάκων, κάθε μια από τις οποίες βελτιστοποιεί μία από τις τρεις διαφορετικές ιδιότητες:

• η κλίμακα της $\sqrt{y}$ θα δώσει σταθερότητα της διακύμανσης·

• η κλίμακα της $y^{2/3}$ θα δώσει περίπου κανονικά λάθη·

• η κλίμακα της $\ln(y)$ θα δώσει προσθετικότητα.

Έτσι, οποιαδήποτε κλίμακα μέτρησης θα είναι πάντα ένας συμβιβασμός, και θα πρέπει να επιλέξετε την κλίμακα που δίνει την καλύτερη συνολική απόδοση του μοντέλου.

Βήματα που εμπλέκονται στην απλοποίηση μοντέλου

Δεν υπάρχουν δύσκολοι και γρήγοροι κανόνες, αλλά η διαδικασία που ορίζεται στον πίνακα 2 λειτουργεί καλά στην πράξη. Με τον μεγάλο αριθμό των επεξηγηματικών μεταβλητών, και με πολλές αλληλεπιδράσεις και μη γραμμικούς όρους, η διαδικασία της απλούστευσης μοντέλου μπορεί να πάρει ένα πολύ μεγάλο χρονικό διάστημα.

Πίνακας 2. Διαδικασία απλοποίησης μοντέλου.

Βήμα	Διαδικασία	Εξήγηση
1	Προσαρμόστε το μέγιστο μοντέλο	Προσαρμόστε όλους τους παράγοντες, τις αλληλεπιδράσεις και τους συμπαράγοντες ενδιαφέροντος. Σημειώστε την υπόλοιπη αποκλίνουσα συμπεριφορά. Εάν χρησιμοποιείτε Poisson ή διωνυμικά λάθη, ελέγξτε για υπερδιασπορά και

		αναπροσαρμόστε εάν είναι απαραίτητο.
2	Αρχίζοντας την απλοποίηση μοντέλου	Ελέγξτε την εκτιμώμενη παράμετρο χρησιμοποιώντας την συνάρτηση της R summary. Αφαιρέστε τους λιγότερο σημαντικούς όρους πρώτα, χρησιμοποιώντας update - , αρχίζοντας με τις αλληλεπιδράσεις με την υψηλότερη τάξη.
3	Εάν η διαγραφή προκαλέσει μια ασήμαντη αύξηση στην αποκλίνουσα συμπεριφορά	Αφήστε αυτόν τον όρο έξω από το υπόδειγμα. Ελέγξτε τις τιμές των παραμέτρων και πάλι. Αφαιρέστε το λιγότερο σημαντικό όρο που απομένει.
4	Εάν η διαγραφή προκαλέσει μια σημαντική αύξηση στην αποκλί-νουσα συμπεριφορά	Βάλτε τον όρο πίσω στο μοντέλο χρησιμοποιώντας **update +**. Αυτοί είναι οι στατιστικά σημαντικοί όροι όπως αξιολογήθηκαν με διαγραφή από το μέγιστο μοντέλο.
5	Κρατήστε την αφαίρεση όρων από το μοντέλο	Επαναλάβετε τα βήματα 3 ή 4 έως ότου το μοντέλο δεν περιέχει τίποτα, εκτός από τους σημαντικούς όρους. Αυτό είναι το ελάχιστο επαρκές μοντέλο. Εάν καμία από τις παραμέτρους είναι σημαντική, τότε το ελάχιστο επαρκές μοντέλο είναι το μηδενικό μοντέλο.

Αυτός είναι ο χρόνος που δαπανάται σωστά, διότι μειώνει τον κίνδυνο παραβλέποντας μια σημαντική πτυχή των δεδομένων. Είναι σημαντικό να συνειδητοποιήσουμε ότι δεν υπάρχει εγγυημένος τρόπος να βρεθούν όλες οι σημαντικές δομές σε ένα σύνθετο πλαίσιο δεδομένων.

Προειδοποιήσεις

Η απλοποίηση του μοντέλου είναι μια σημαντική διαδικασία, αλλά δεν θα πρέπει να λαμβάνεται στα άκρα. Για παράδειγμα, πρέπει να ληφθεί μέριμνα για την ερμηνεία της αποκλίνουσας συμπεριφοράς και των τυπικών σφαλμάτων που παράγονται με σταθερές παραμέτρους που έχουν εκτιμηθεί από τα δεδομένα. Και πάλι, η αναζήτηση για 'ωραίους αριθμούς' δεν θα πρέπει να επιδιώκεται άκριτα. Μερικές φορές υπάρχουν καλοί επιστημονικοι λόγοι για τη χρήση ενός συγκεκριμένου αριθμού (π.χ. η δύναμη 0.66 σε μία αλλομετρική σχέση μεταξύ της αναπνοής και της μάζας σώματος). Είναι πολύ πιο απλό, για παράδειγμα, να πούμε ότι η απόδοση αυξάνεται κατά 2 kg ανά εκτάριο για κάθε επιπλέον μονάδα λιπασμάτων, από το να πούμε ότι αυξάνει κατά 1.947 kg. Ομοίως, μπορεί να είναι προτιμότερο να πούμε ότι οι πιθανότητες μόλυνσης έχουν 10-πλάσια

αύξηση κάτω από μια δεδομένη θεραπεία, από το να πούμε ότι τα Logits αυξάνονται κατά 2.321· χωρίς απλοποίηση του μοντέλου αυτό είναι ισοδύναμο λέγοντας ότι υπάρχει μία 10.186-πλάσια αύξηση στην απόδοση. Θα ήταν παράλογο, ωστόσο, να οριστεί η εκτίμηση του 6 αντί 6.1 μόνο και μόνο επειδή το 6 είναι ένας ακέραιος αριθμός.

Παραγγελία της διαγραφής

Τα δεδομένα σε αυτό το βιβλίο εμπίπτουν σε δύο διακριτές κατηγορίες. Στην περίπτωση των προγραμματισμένων πειραμάτων, όλοι οι συνδυασμοί κατεργασίας εκπροσωπούνται ισότιμα και, εάν δεν υπάρξουν ατυχήματα, δεν υπάρχουν ελλειπείς τιμές. Τέτοια πειράματα λέγεται ότι είναι **ορθογώνια**. Στην περίπτωση των σπουδών παρατήρησης, όμως, δεν έχουμε κανέναν έλεγχο πάνω από τον αριθμό των ατόμων για τα οποία έχουμε στοιχεία, ή πάνω από τους συνδυασμούς των συνθηκών που παρατηρούνται. Πολλές από τις επεξηγηματικές μεταβλητές είναι πιθανόν να συσχετίζονται η μία με την άλλη, καθώς και με τη μεταβλητή απόκρισης. Με το να λείπουν οι συνδυασμοί κατεργασίας σαν κοινό σημείο ανάφορας, και τα δεδομένα που λέγεται ότι είναι **μη-ορθογώνια**. Αυτό κάνει μια σημαντική διαφορά στην στατιστική μοντελοποίηση μας, διότι, σε ορθογώνια σχέδια, η παραλλαγή που αποδίδεται σε έναν δεδομένο παράγοντα είναι σταθερή και δεν εξαρτάται από τη σειρά με την οποία οι παράγοντες απομακρύνονται από το μοντέλο. Σε αντίθεση, με μη ορθογώνια στοιχεία, διαπιστώνουμε ότι η διακύμανση που αναλογεί σε ένα δεδομένο παράγοντα *όντως* εξαρτάται από τη σειρά με την οποία οι παράγοντες απομακρύνονται από το μοντέλο. Θα πρέπει να προσέξουμε, ως εκ τούτου, για να κριθεί η σημασία των παραγόντων σε μη ορθογώνιες μελέτες, όταν αυτοί *απομακρύνονται από το μέγιστο μοντέλου* (δηλ. από το μοντέλο, συμπεριλαμβανομένων όλων των άλλων παραγόντων και των αλληλεπιδράσεων με τα οποία θα μπορούσαν να συγχέονται). Να θυμάστε ότι, *για μη-ορθογώνια δεδομένα, η σειρά έχει σημασία.*

Επίσης, αν οι επεξηγηματικές μεταβλητές σας συσχετίζονται μεταξύ τους, τότε η σημασία που αποδίδουν σε μια δεδομένη επεξηγηματική μεταβλητή θα εξαρτηθεί από το εάν την διαγράψετε από ένα μέγιστο μοντέλο ή να την προσθέσετε στο μηδενικό μοντέλο. Εάν πάντα κάνετε δοκιμές στην απλοποίηση μοντέλου, τότε δεν θα πέσετε σε αυτή την παγίδα.

Το γεγονός ότι έχετε εργαστεί πολύ και σκληρά για να συμπεριλάβετε μια συγκεκριμένη πειραματική κατεργασία, δεν δικαιολογεί τη διατήρηση του εν λόγω συντελεστή στο μοντέλο, αν η ανάλυση δείχνει ότι δεν έχει κάποια επεξηγηματική δύναμη. Οι πίνακες της ανάλυσης διακύμανσης (ANOVA) δημοσιεύονται συχνά περιέχοντας ένα μίγμα των σημαντικών και μη σημαντικών επιδράσεων. Αυτό δεν είναι ένα πρόβλημα σε ορθογώνια σχέδια, διότι το άθροισμα των τετραγώνων μπορεί κατηγορηματικά να αποδίδεται σε κάθε παράγοντα και σε όρους αλληλεπίδρασης. Αλλά μόλις υπάρξουν ελλειπείς τιμές ή άνιση βάρη, τότε είναι αδύνατο να πει κανείς, πώς οι εκτιμήσεις των παραμέτρων και τα πρότυπα σφάλματα των σημαντικών όρων θα έχουν αλλάξει, εάν οι μη σημαντικοί όροι είχαν διαγραφεί. Η καλύτερη πρακτική είναι η ακόλουθη:

• Πείτε αν τα δεδομένα σας είναι ορθογώνια ή όχι.

• Εξηγήστε τυχόν συσχετίσεις μεταξύ των επεξηγηματικών μεταβλητών σας.

• Παρουσιάστε ένα ελάχιστο επαρκές μοντέλο.

• Δώστε έναν κατάλογο των μη σημαντικών όρων που είχαν παραλειφθεί, και των αλλαγών της απόκλισης που προέκυψαν από τη διαγραφή τους.

Οι αναγνώστες μπορούν στη συνέχεια να κρίνουν για τον εαυτό τους το σχετικό μέγεθος των μη σημαντικών παραγόντων, καθώς και τη σημασία των συσχετίσεων μεταξύ των επεξηγηματικών μεταβλητών.

Ο πειρασμός να διατηρούνται οι όροι στο μοντέλο που είναι 'κοντά στην σημασία' θα πρέπει να αντισταθεί. Ο καλύτερος τρόπος για να προχωρήσουμε είναι ο εξής. Το αποτέλεσμα θα ήταν *σημαντικό*, αν ήταν στατιστικώς σημαντικό, τότε θα άξιζε την επανάληψη του πειράματος με υψηλότερη αντιγραφή και/ή πιο αποτελεσματικό μπλοκάρισμα, προκειμένου να αποδείξει τη σημασία του παράγοντα με ένα πειστικό και στατιστικώς αποδεκτό τρόπο.

Τύποι Μοντέλου στην R

Η δομή του μοντέλου ορίζεται στον τύπο μοντέλου όπως :

Μεταβλητή απόκρισης ~ επεξηγηματική μεταβλητή (ες)

όπου το σύμβολο της περισπωμένης ~ διαβάζει 'διαμορφώνεται ως συνάρτηση του' (βλ. Πίνακα 3 για παραδείγματα).

Πίνακας 3. Παραδείγματα τύπων μοντέλου στην R. Σε έναν τύπο μοντέλου, η συνάρτηση I (κεφαλαίο γράμμα i) αντιπροσωπεύει 'ως έχει' και χρησιμοποιείται για τη δημιουργία ακολουθιών I(1:10) ή για να υπολογίζει τετραγωνικούς όρους I(x^2).

Μοντέλο	**Τύπος Μοντέλου**	**Σχόλια**
Μηδενικό	y ~ 1	το 1 είναι το σημείο τομής των μοντέλων παλινδρόμησης, αλλά εδώ είναι η συνολική μέση τιμή *y*
Παλινδρόμηση	y ~ x	το *x* είναι μια συνεχής επεξηγηματική μεταβλητή
Παλινδρόμηση μέσω προέλευσης	y ~ x-1	Μην προσαρμόσετε ένα σημείο τομής
Μονόδρομη ANOVA	y ~ sex	το φύλο είναι κατηγορική μεταβλητή δύο επιπέδων
Μονόδρομη ANOVA	y ~ sex-1	όπως παραπάνω, αλλά δεν προσαρμόζουν ένα σημείο τομής (δίνει δύο μέσες τιμές και όχι μια και τη διαφορά)
Αμφίδρομη ANOVA	y ~ sex + γονότυπο	ο γονότυπος είναι μία τεσσάρων επιπέδων κατηγορική μεταβλητή
Παραγοντική ANOVA	y ~ N * P * K	*N*, *P* και *K* είναι δύο επιπέδων παράγοντες που πρέπει να

		προσαρμοστούν μαζί με όλες τις αλληλεπιδράσεις τους
Τριμερή ANOVA	y ~ N*P*K – N:P:K	Όπως και παραπάνω, αλλά μην προσαρμόσετε την τριμερή αλληλεπίδραση
Ανάλυση συνδιακύμανσης	y ~ x + sex	Μια κοινή κλίση για *y* κατά *x* αλλά με δύο σημεία τομής, ένα για κάθε φύλο
Ανάλυση συνδιακύμανσης	y ~ x * sex	Δύο κλίσεις και δύο παρακολουθήσεις
Φωλιασμένη ANOVA	y ~ a/b/c	Ο παράγοντας *c* φωλιασμένος μέσα στον παράγοντα *b* μέσα στον παράγοντα *a*
Split-plot ANOVA	y ~ a*b*c+Error(a/b/c)	Ένα παραγοντικό πείραμα, αλλά με τρία μεγέθη διαγράμματος και τρεις διαφορετικές διακυμάνσεις σφάλματος, ένα για κάθε μέγεθος του διαγράμματος
Πολλαπλή παλινδρόμηση	y ~ x + z	Δύο συνεχείς επεξηγηματικές μεταβλητές, κατ 'εφαρμογή επιφάνειας
Πολλαπλή παλινδρόμηση	y ~ x * z	Προσαρμόστε έναν όρο αλληλεπίδρασης, καθώς και (x + z + x: z)
Πολλαπλή παλινδρόμηση	y ~ x + I(x^2) + z + I(z^2)	Προσαρμόστε ένα τετραγωνικό όρο για τις δύο *x* και *z*
Πολλαπλή παλινδρόμηση	y <- poly(x,2) + z	Προσαρμόστε ένα τετραγωνικό πολυώνυμο για το *x* και γραμμικό *z*
Πολλαπλή παλινδρόμηση	y ~ (x + z + w)^2	Προσαρμόστε τρεις μεταβλητές καθώς και όλες τις αλληλεπιδράσεις τους, να είναι αμφίδρομες
Μη παραμετρικό μοντέλο	y ~ s(x) +s(z)	Η *y* είναι συνάρτηση της εξομαλυμένης *x* και η *z* σε ένα γενικευμένο πρόσθετο μοντέλο
Μεταμορφωμένη απόκριση και επεξηγηματικές μεταβλητές	log(y) ~ I(1/x) + sqrt(z)	Και οι τρεις μεταβλητές μετασχηματίζονται στο μοντέλο

Έτσι, μια απλή γραμμική παλινδρόμηση του y επί του x θα μπορούσε να γραφτεί ως

y ~ x

και μία μονόδρομη ανάλυση διακύμανσης ANOVA όπου το φύλο είναι παράγοντας δύο επιπέδων θα μπορούσε να γραφτεί ως

y ~ sex

Η δεξιά πλευρά του τύπου του μοντέλου δείχνει:

• τον αριθμό των επεξηγηματικών μεταβλητών και τις ταυτότητές τους - τις ιδιότητές τους (π.χ. συνεχής ή κατηγορηματική) συνήθως ορίζονται πριν από την προσαρμογή του μοντέλου·

• τις αλληλεπιδράσεις μεταξύ των επεξηγηματικών μεταβλητών (εάν υπάρχουν)·

• τους μη γραμμικούς όρους στις επεξηγηματικές μεταβλητές.

Στα δεξιά της περισπωμένης, ο αριθμός 1 έχει επίσης τη δυνατότητα να καθορίζει σε ορισμένες ειδικές περιπτώσεις τις μετατοπίσεις ή τους λάθους όρους. Όπως και με τη μεταβλητή απόκρισης, οι επεξηγηματικές μεταβλητές μπορεί να εμφανιστούν ως μετασχηματισμοί, είτε ως δυνάμεις ή πολυώνυμα.

Είναι πολύ σημαντικό να σημειωθεί ότι τα σύμβολα χρησιμοποιούνται με διαφορετικό τρόπο σε τύπους μοντέλου από ό, τι σε αριθμητικές εκφράσεις. Ειδικότερα:

+ δείχνει ένταξη της επεξηγηματικής μεταβλητής στο μοντέλο (όχι προσθήκη)·

- δείχνει τη διαγραφή της επεξηγηματικής μεταβλητής από το μοντέλο (όχι αφαίρεση)·

* δείχνει ένταξη των επεξηγηματικών μεταβλητών και των αλληλεπιδράσεων (όχι τον πολλαπλασιασμό)·

/ υποδηλώνει φώλιασμα των επεξηγηματικών μεταβλητών στο μοντέλο (όχι διαίρεση)·

| δηλώνει κατάσταση (όχι 'ή'), έτσι ώστε y ~ x | z διαβάζεται ως 'η y ως συνάρτηση της x δεδομένου της z'.

Υπάρχουν πολλά άλλα σύμβολα που έχουν ιδιαίτερη σημασία στους τύπους μοντέλου.
Η άνω και κάτω τελεία υποδηλώνει μια αλληλεπίδραση, έτσι ώστε A:B σημαίνει την αμφίδρομη αλληλεπίδραση μεταξύ των A και B, και N:P:K:Mg εννοείται η τεσσάρων κατευθύνσεων αλληλεπίδραση μεταξύ των N, P, K και Mg.

Κάποιοι όροι μπορούν να γραφτούν σε μια διευρυμένη μορφή. Έτσι:

A*B*C είναι το ίδιο με A+B+C+A:B+A:C+B:C+A:B:C
A/B/C είναι το ίδιο με A+B%in%A+C%in%B%in%A
(A+B+C)^3 είναι το ίδιο με A*B*C

(A+B+C)^2 είναι το ίδιο με A*B*C – A:B:C

Οι αλληλεπιδράσεις μεταξύ των επεξηγηματικών μεταβλητών

Οι αλληλεπιδράσεις μεταξύ δύο διεπίπεδων κατηγορικών μεταβλητών της μορφής A*B σημαίνει ότι αξιολογούνται δύο κύρια μέσα αποτέλεσματος και ένα μέσο αλληλεπίδρασης. Από την άλλη πλευρά, αν ο συντελεστής A έχει τρία επίπεδα και ο B έχει τέσσερα επίπεδα, τότε υπολογίστηκαν επτά παράμετροι για τις κύριες επιδράσεις (τρείς μέσες τιμές για τον A και τέσσερεις για τον B). Ο αριθμός των όρων αλληλεπίδρασης είναι $(a-1)(b-1)$, όπου a και b είναι οι αριθμοί των επιπέδων των παραγόντων A και B , αντίστοιχα. Έτσι, στην περίπτωση αυτή, η R θα εκτιμούσε $(3-1)(4-1) = 6$ παραμέτρους για την αλληλεπίδραση.

Οι αλληλεπιδράσεις μεταξύ δύο συνεχών μεταβλητών μπορούν να προσαρμοστούν με διαφορετικό τρόπο. Αν x και z είναι δύο συνεχείς επεξηγηματικές μεταβλητές, τότε η μέση τιμή x*z προσαρμόζεται x+z+x:z και ο όρος αλληλεπίδρασης x:z συμπεριφέρεται ως μια νέα μεταβλητή που είχε υπολογιστεί ότι ήταν το σημειακό γινόμενο των δύο διανυσμάτων x και z . Το ίδιο αποτέλεσμα θα μπορούσε να επιτευχθεί με τον υπολογισμό του γινομένου ρητά,

```
product.xz <- x * z
```

στη συνέχεια, χρησιμοποιώντας τον τύπο μοντέλου y ~ x + z + product.xz. Σημειώστε ότι η αναπαράσταση της αλληλεπίδρασης με το *γινόμενο* των δύο συνεχών μεταβλητών είναι μια υπόθεση, όχι ένα γεγονός. Η πραγματική αλληλεπίδραση μπορεί να είναι μια εντελώς διαφορετική λειτουργική μορφή (π.χ. x * z ^ 2).

Οι αλληλεπιδράσεις μεταξύ μιας κατηγορικής μεταβλητής και μιας συνεχούς μεταβλητής ερμηνεύεται ως μια ανάλυση της συνδιακύμανσης· μια ξεχωριστή κλίση και το σημείο τομής έχουν προσαρμοστεί για κάθε επίπεδο της κατηγορηματικής μεταβλητής. Έτσι, y ~ A * x θα προσαρμόσει τρεις εξισώσεις παλινδρόμησης, αν ο παράγοντας A είχε τρία επίπεδα· αυτό θα εκτιμήσει έξι παραμέτρους από τα δεδομένα - τρεις κλίσεις και τρεις παρακολουθήσεις.

Η κάθετος / ως χειριστής, χρησιμοποιείται για να υποδηλώσει φώλιασμα. Έτσι, με κατηγορικές μεταβλητές A και B ,

```
y ~ A/B
```

σημαίνει προσαρμογή ‘ A συν B μέσα στην A ’ . Αυτό θα μπορούσε να γραφτεί με δύο άλλους ισοδύναμους τρόπους:

```
y ~ A + A:B
```

```
y ~ A + B %in% A
```

δύο εκ των οποίων εναλλακτικών τονίζουν ότι δεν υπάρχει κανένα σημείο στην προσπάθεια να εκτιμηθεί μία κύρια επίδραση για την B (είναι πιθανώς μια ετικέτα παράγοντα όπως ‘αριθμός δέντρου 1’, που δεν έχει κανένα επιστημονικό ενδιαφέρον· βλ. σελ. 479).

Ορισμένες συναρτήσεις για τον καθορισμό των μη γραμμικών όρων και για αλληλεπιδράσεις υψηλότερης τάξης είναι χρήσιμες. Για να προσαρμόσουμε ένα πολυώνυμο παλινδρόμησης στην x και z , θα μπορούσαμε να γράψουμε

```
y ~ poly(x,3) + poly(z,2)
```

για να προσαρμόσουμε ένα κυβικό πολυώνυμο στην x και ένα τετραγωνικό πολυώνυμο στην z. Για να προσαρμόσουμε αλληλεπιδράσεις, αλλά μόνο μέχρι ένα ορισμένο επίπεδο, είναι χρήσιμος ο χειριστής ^. Ο τύπος

```
y ~ (A + B + C)^2
```

προσαρμόζει σε όλα τα κύρια αποτελέσματα και στις αμφίδρομες αλληλεπιδράσεις (δηλαδή αποκλείει την τριμερή αλληλεπίδραση ώστε A*B*C θα μπορούσε να συμπεριληφθεί).

Η συνάρτηση I (κεφαλαίο γράμμα i) αντιπροσωπεύει 'ως έχει'. Παρακάμπτει την ερμηνεία ενός συμβόλου μοντέλου ως χειριστή τύπου, όταν η πρόθεση είναι να χρησιμοποιηθεί ως αριθμητικός τελεστής. Ας υποθέσουμε ότι θέλετε να προσαρμόσετε το $1 / x$ ως επεξηγηματική μεταβλητή σε μια παλινδρόμηση. Θα μπορούσατε να δοκιμάσετε

```
y~1/x
```

αλλά αυτό κάνει πραγματικά κάτι πολύ περίεργο. Προσαρμόζει το φωλιασμένο x εντός της τομής! Όταν εμφανίζεται σε έναν τύπο μοντέλο, ο χειριστής της καθέτου υποτίθεται ότι συνεπάγεται φώλιασμα. Για να επιτευχθεί το αποτέλεσμα που θέλουμε, χρησιμοποιούμε την συνάρτηση I για να γράψουμε

```
y ~ I(1/x)
```

Πρέπει επίσης να χρησιμοποιήσουμε την I όταν θέλουμε το σύμβολο * να εκπροσωπεί τον πολλαπλασιασμό και το ^ να σημαίνει 'στη δύναμη' και όχι μια επέκταση μοντέλου αλληλεπίδρασης: έτσι για να προσαρμόσουμε το x και το x^2 σε μια τετραγωνική παλινδρόμηση θα γράφαμε

```
y ~ x+I(x^2)
```

Δημιουργία αντικειμένων τύπου

Μπορείτε να επιταχύνετε τη δημιουργία σύνθετων τύπων μοντέλου χρησιμοποιώντας paste για να δημιουργήσετε σειρές των ονομάτων των μεταβλητών και collapse για να ενταχθούν τα ονόματα των μεταβλητών ανά σύμβολα. Εδώ, για παράδειγμα, είναι μια πολλαπλή φόρμουλα παλινδρόμησης με 25 συνεχείς επεξηγηματικές μεταβλητές που δημιουργήθηκαν με τη χρήση της συνάρτησης as.formula:

```
xnames <- paste("x", 1:25, sep="")
(model.formula <- as.formula(paste("y ~ ", paste(xnames, collapse= "+"))))
```

```
y ~ x1 + x2+  x3+   x4+  x5+  x6+  x7+  x8+  x9+ x10+ x11+
        x12+ x13+  x14+ x15+ x16+ x17+ x18+ x19+ x20+ x21+
        x22+ x23+  x24+ x25
```

Πολλαπλοί όροι σφαλμάτων

Όταν υπάρχει φώλιασμα (π.χ. διάγραμμα τμημάτων-split plots σε ένα σχεδιασμένο πείραμα· Βλ. σελ. 470) ή προσωρινή ψευδοαναπαραγωγή (Βλ. σελ. 474), μπορείτε να συμπεριλάβετε μια συνάρτηση Error ως μέρος του τύπου μοντέλου. Ας υποθέσουμε ότι είχατε ένα παραγοντικό πείραμα τριών παραγόντων με κατηγορηματικές μεταβλητές A, B και C. Για να παραχθεί, κάθε επεξεργασία εφαρμόζεται σε διαγράμματα διαφόρων μεγεθών: η A εφαρμόζεται για να αναπαράγει ολόκληρα πεδία, η B εφαρμόζεται σε τυχαία μισά πεδία και η C εφαρμόζεται σε μικρότερα τμηματικά διαγράμματα (split-split plot) σε κάθε τομέα. Αυτό φαίνεται σε έναν τύπο μοντέλου όπως αυτό:

y ~ A*B*C + Error(A/B/C)

Σημειώστε ότι οι όροι στους τύπους μοντέλου χωρίζονται με αστερίσκους για να δείξουν ότι είναι ένα πλήρες παραγοντικό συμπεριλαμβανομένων όλων των όρων αλληλεπίδρασης, ενώ στην κατάσταση Error οι όροι χωρίζονται με καθέτους. Υπάρχουν, τόσοι πολλοί όροι στην κατάσταση Error, όσοι υπάρχουν σε διαφορετικά μεγέθη των διαγραμμάτων -τρεις σε αυτή την περίπτωση, αν και το μικρότερο μέγεθος του διαγράμματος (C σε αυτό το παράδειγμα) μπορεί να παραλειφθεί από τον κατάλογο - και οι όροι απαριθμούνται από αριστερά προς τα δεξιά από το μεγαλύτερο προς το μικρότερο διάγραμμα· βλ. σελ.. 469 για λεπτομέρειες και παραδείγματα.

Το σημείο τομής ως παράμετρος 1

Η απλή εντολή

y~1

προκαλεί το μηδενικό μοντέλο που θα προσαρμοστεί. Αυτό λειτουργεί στην μέγιστη μέση τιμή (ο συνολικός μέσος όρος) όλων των δεδομένων και στη συνολική απόκλιση (ή το συνολικό άθροισμα των τετραγώνων, SSY, σε μοντέλα με κανονικά λάθη και τη σύνδεση ταυτότητας). Σε ορισμένες περιπτώσεις, αυτό μπορεί να είναι το ελάχιστο επαρκές μοντέλο· είναι πιθανό ότι καμία από τις επεξηγηματικές μεταβλητές που έχουμε μετρήσει συνεισφέρουν ο,τιδήποτε σημαντικό στην κατανόηση της διακύμανσης στη μεταβλητή απόκρισης. Αυτό είναι συνήθως αυτό που δεν θέλετε να συμβεί στο τέλος του τριετούς ερευνητικού έργου σας.

Για να αφαιρέσετε το σημείο τομής (παράμετρος 1) από ένα μοντέλο παλινδρόμησης (δηλαδή να αναγκάσετε τη γραμμή παλινδρόμησης μέσω της προέλευσης) προσαρμόζετε το '−1', όπως εδώ:

y ~ x – 1

Δεν πρέπει να το κάνετε αυτό αν δεν ξέρετε ακριβώς τι κάνετε, και ακριβώς γιατί το κάνετε (βλ. σελ. 393 για λεπτομέρειες). Αφαιρόντας την τομή από ένα μοντέλο ANOVA, όπου όλες οι μεταβλητές είναι κατηγορηματικές έχει διαφορετική επίδραση:

y~sex – 1 Αυτό δίνει την μέση τιμή για τους άντρες και τη μέση για τις γυναίκες στο συνοπτικό πίνακα, αντί για τη μέση τιμή για τις γυναίκες και τη διαφορά στη μέση τιμή για τους άνδρες (βλ. σελ.. 366).

Η συνάρτηση update στην απλοποίηση μοντέλου

Στη συνάρτηση update που χρησιμοποιείται κατά την απλοποίηση μοντέλου, η τελεία '.' χρησιμοποιείται για να διευκρινίσει 'τι είναι ήδη εκεί' εκατέρωθεν της περισπωμένης. Έτσι, εάν το αρχικό μοντέλο σας είναι

```
model<-lm(y ~ A*B)
```

τότε η συνάρτηση update για να αφαιρέσει τον όρο αλληλεπίδρασης A:B θα μπορούσε να γραφτεί ως εξής:

```
model2<-update(model, ~ .- A:B)
```

Σημειώστε ότι δεν υπάρχει καμία ανάγκη να επαναλάβετε το όνομα της μεταβλητής απόκρισης, και τα σημεία στίξης 'περισπωμένη τελεία' σημαίνουν αναλάβετε το μοντέλο, όπως είναι, και αφαιρέσετε από αυτό, τον όρο αλληλεπίδρασης A:B.

Τύποι Μοντέλου για παλινδρόμηση

Το σημαντικό σημείο που πρέπει να κατανοήσετε είναι ότι οι τύποι μοντέλου μοιάζουν πολύ με εξισώσεις, αλλά υπάρχουν σημαντικές διαφορές. Η απλούστερη χρήσιμη εξίσωση μας μοιάζει με:

$$y = ax + b$$

Πρόκειται για ένα μοντέλο δύο παραμέτρων με μία παράμετρο για το σημείο τομής, a, και ένα άλλο για την κλίση, b, από τη γραφική παράσταση της συνεχούς μεταβλητής απόκρισης y εναντίον μίας συνεχούς επεξηγηματικής μεταβλητής x. Ο τύπος μοντέλου για την ίδια σχέση μοιάζει με:

```
y ~ x
```

Το ίσον αντικαθίσταται από μια περισπωμένη, καθώς και όλες οι παράμετροι έχουν μείνει εκτός. Άν είχαμε μια πολλαπλή παλινδρόμηση με δύο συνεχείς επεξηγηματικές μεταβλητές x και z, η εξίσωση θα ήταν:

```
y=a+bx+cz,
```

αλλά ο τύπος μοντέλου θα είναι

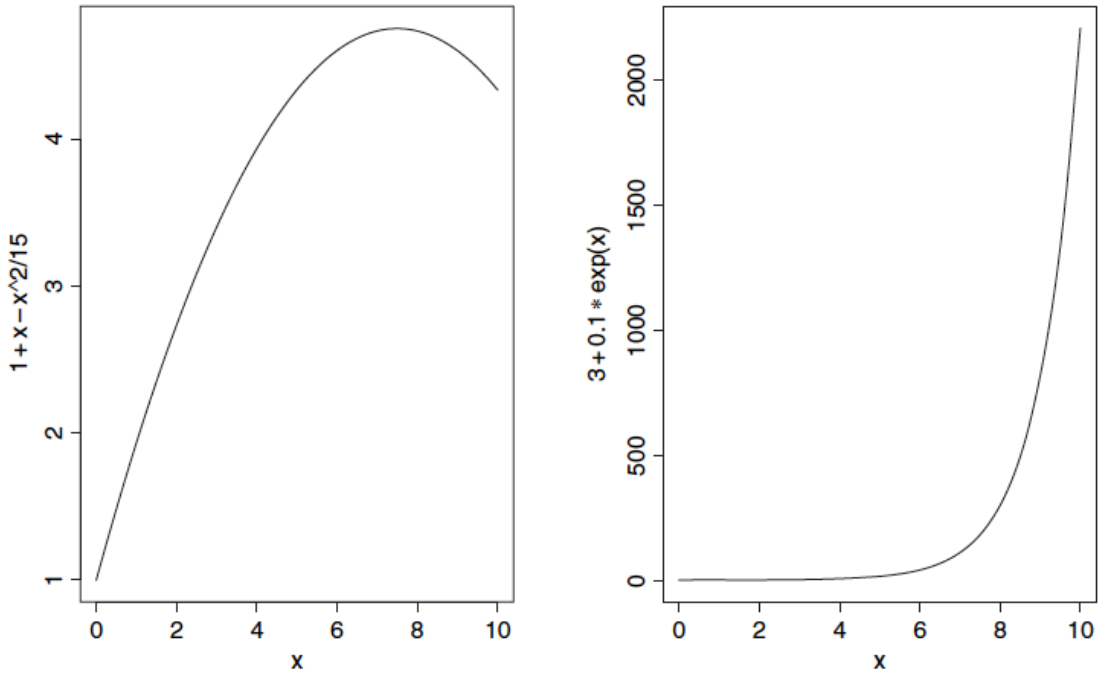
```
y ~ x + z
```

Είναι όλα υπέροχα απλά. Αλλά ένα λεπτό. Πώς η R ξέρει τι παραμέτρους θέλουμε για την εκτίμηση από τα δεδομένα; Έχουμε πει μόνο τα ονόματα των επεξηγηματικών μεταβλητών. Δεν έχουμε πει κάτι για το πώς αυτά θα προσαρμοστούν, ή τι είδους εξίσωση θέλουμε να προσαρμόσουμε στα δεδομένα. Το κλειδί σε αυτό είναι να καταλάβουμε τι είδους επεξηγηματική μεταβλητή μπορεί να προσαρμοστεί με τα

δεδομένα. Εάν η επεξηγηματική μεταβλητή x που καθορίζεται στα δεξιά της περισπωμένης είναι μια συνεχής μεταβλητή, τότε η R *υποθέτει* ότι θέλετε να κάνετε μια παλινδρόμηση, και ως εκ τούτου, ότι θέλουμε να εκτιμήσουμε δύο παραμέτρους σε μια γραμμική παλινδρόμηση των οποίων η εξίσωση είναι $y = a + bx$.

Μια κοινή παρανόηση είναι ότι τα γραμμικά μοντέλα περιλαμβάνουν μια ευθεία σχέση μεταξύ της μεταβλητής απόκρισης και των επεξηγηματικών μεταβλητών. Αυτή δεν είναι η περίπτωση, όπως μπορείτε να δείτε από αυτά τα δύο γραμμικά μοντέλα:

```
par(mfrow=c(1,2))
x<-seq(0,10,0.1)
plot(x,1+x-x^2/15,type="l")
plot(x,3+0.1*exp(x),type="l")
```



Ο ορισμός του γραμμικού μοντέλου είναι μια εξίσωση που περιέχει μαθηματικές μεταβλητές, παραμέτρους και τυχαίες μεταβλητές και ότι είναι γραμμική στις παραμέτρους και τις τυχαίες μεταβλητές. Αυτό που σημαίνει είναι ότι εάν a, b και c είναι παράμετροι τότε προφανώς

$$y = a + bx$$

είναι ένα γραμμικό μοντέλο, αλλά έτσι είναι

$$y = a + bx - cx^2$$

επειδή το x^2 μπορεί να αντικατασταθεί από το z το οποίο δίνει μια γραμμική σχέση

$$y = a + bx + cz,$$

και έτσι είναι

$$y = a + be^x$$

γιατί μπορούμε να δημιουργήσουμε μια νέα μεταβλητή $z = \exp(x)$, έτσι ώστε

$$y = a + bz.$$

Ορισμένα μοντέλα είναι μη γραμμικά, αλλά μπορούν εύκολα να γραμμικοποιηθούν με μετασχηματισμό. Για παράδειγμα:

$$y = \exp(a + bx)$$

είναι μη γραμμικό, αλλά λογαριθμίζοντας και τις δύο πλευρές, γίνεται

$$\ln(y) = a + bx$$

Αν η εξίσωση που θέλετε να προσαρμόσετε είναι πιο περίπλοκη από αυτήν, τότε θα πρέπει να καθορίσετε τη μορφή της εξίσωσης, και χρησιμοποιήστε μη γραμμικές μεθόδους (nls ή nlme) για να προσαρμόσετε το μοντέλο στα δεδομένα (βλ. σελ. 661).

Μετασχηματισμοί Box-Cox

Μερικές φορές δεν είναι σαφές από τη θεωρία ποια θα έπρεπε να είναι η βέλτιστη μετατροπή της μεταβλητής απόκρισης. Υπό αυτές τις συνθήκες, ο μετασχηματισμός Box-Cox προσφέρει μια απλή εμπειρική λύση. Η ιδέα είναι να βρούμε το μετασχηματισμό δύναμης, λ (λάμδα), ο οποίος μεγιστοποιεί την πιθανότητα όταν ένα καθορισμένο σύνολο των επεξηγηματικών μεταβλητών είναι προσαρμοσμένο σε

$$\frac{y^\lambda - 1}{\lambda}$$

ως απόκριση. Η τιμή του λάμδα μπορεί να είναι θετική ή αρνητική, αλλά δεν μπορεί να είναι μηδέν (θα παίρνατε ένα σφάλμα μηδενικού χάσματος όταν ο τύπος θα εφαρμοζόταν για τη μεταβλητή απόκρισης, y). Για την περίπτωση $\lambda = 0$, ο μετασχηματισμός Box-Cox ορίζεται ως $\log(y)$. Ας υποθέσουμε ότι $\lambda = -1$. Ο τύπος γίνεται τώρα

$$\frac{y^{-1} - 1}{-1} = \frac{1/y - 1}{-1} = 1 - \frac{1}{y},$$

και η ποσότητα αυτή παλινδρομείται έναντι των επεξηγηματικών μεταβλητών και του υπολογισμού της λογαριθμικής συνάρτησης πιθανότητας.

Σε αυτό το παράδειγμα, θέλουμε να βρεθεί η βέλτιστη μετατροπή της μεταβλητής απόκρισης, το οποίο είναι ο όγκος ξυλείας:

```
data<-read.delim("c:\\temp\\timber.txt") attach(data) names(data)
```

```
[1] "volume" "girth" "height"
```

Ξεκινάμε φορτώνοντας την βιβλιοθήκη MASS των Venables και Ripley:

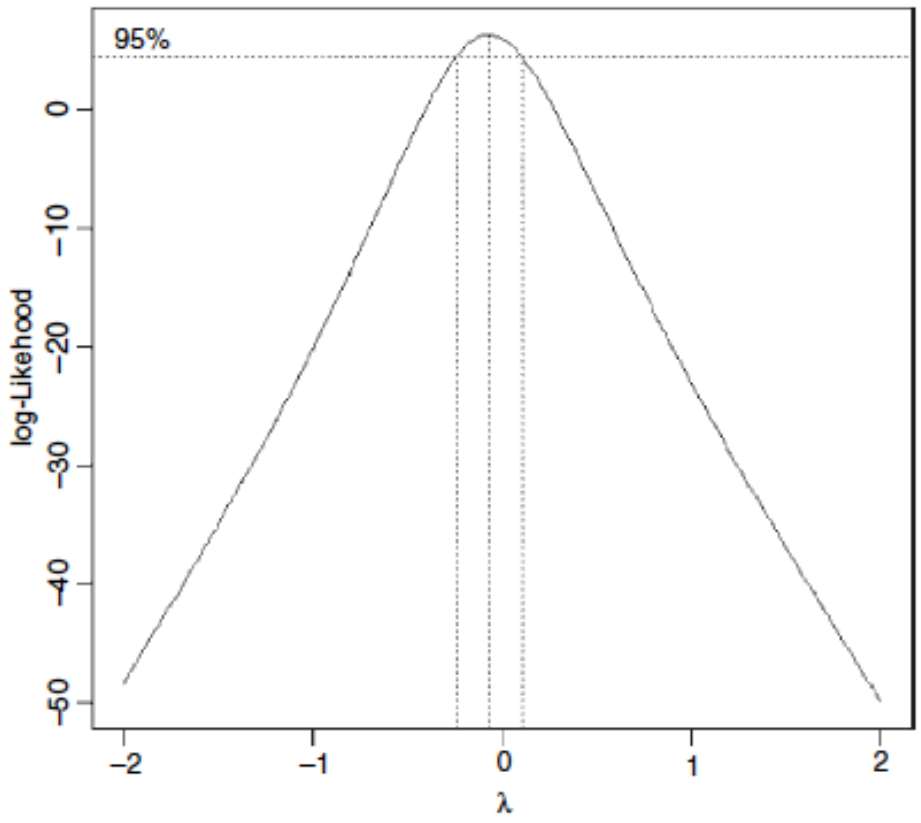
```
library(MASS)
```

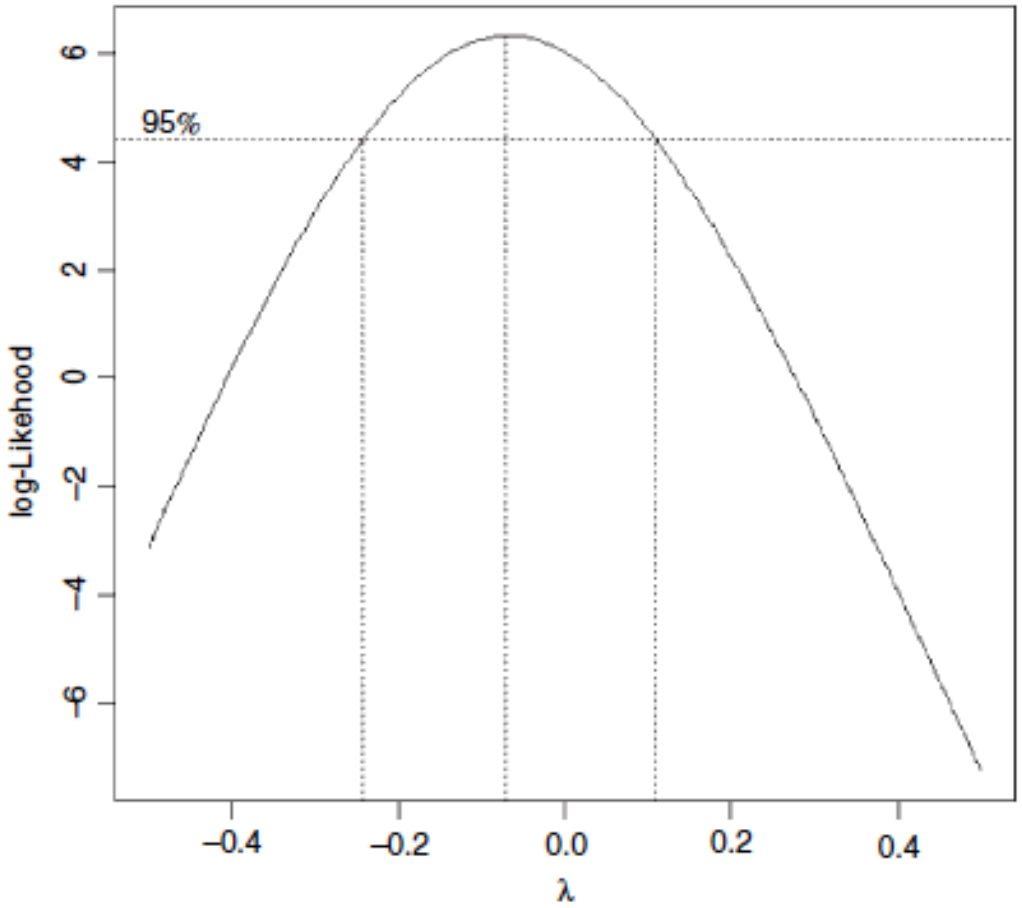
Η συνάρτηση boxcox είναι πολύ εύκολη στη χρήση: απλά προσδιορίστε τον τύπο μοντέλου, και οι προεπιλεγμένες ρυθμίσεις θα φροντίσουν για όλα τα υπόλοιπα.

```
boxcox(volume ~ log(girth)+log(height))
```

Είναι σαφές ότι η βέλτιστη τιμή του λάμδα είναι κοντά στο μηδέν (δηλ. το λογαριθμικό μετασχηματισμό). Μπορούμε να μεγεθύνουμε για να πάρουμε μια πιο ακριβή εκτίμηση καθορίζοντας το δικό μας, μη προεπιλεγμένο, εύρος τιμών του λάμδα. Φαίνεται σαν να ήταν λογικό να σχεδιάσετε από το -0.5 στο $+0.5$:

```
boxcox(volume ~ log(girth)+log(height),lambda=seq(-0.5,0.5,0.01))
```

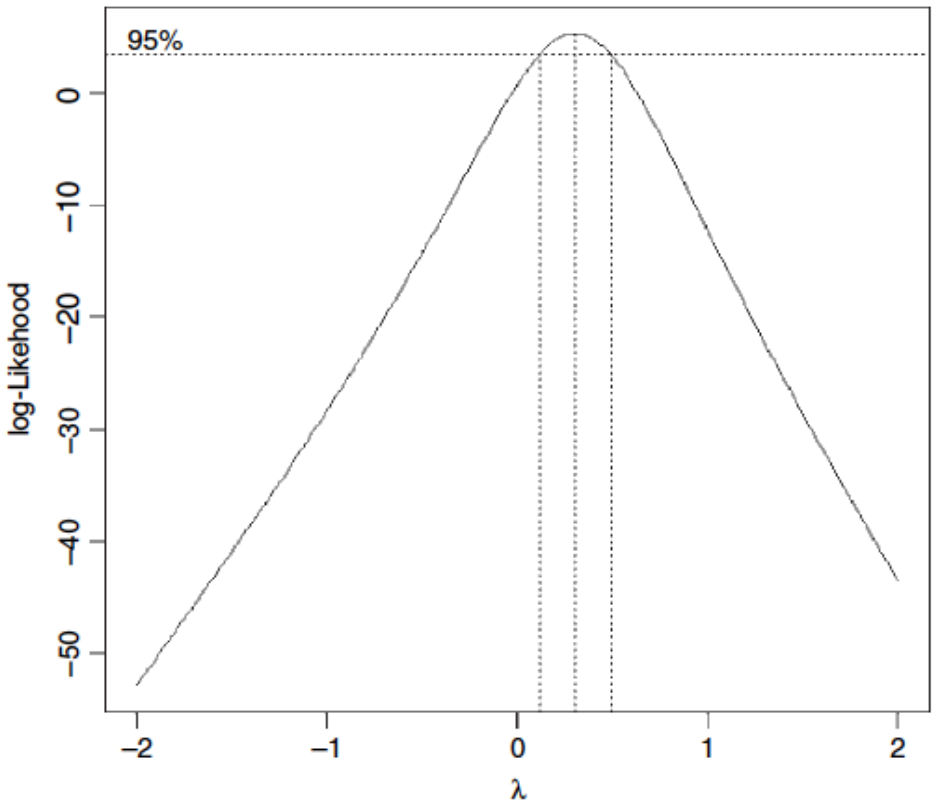




Η πιθανότητα μεγιστοποιείται όταν $\lambda \approx -0.08$, αλλά η λογαριθμική συνάρτηση πιθανότητας για $\lambda = 0$ είναι πολύ κοντά στο μέγιστο. Αυτό δίνει επίσης μια πολύ πιο απλή ερμηνεία, γι' αυτό θα πάμε με αυτό, και μοντελοποιώντας log(volume) ως συνάρτηση του log(girth) και log(height) (βλ. σελ. 518).

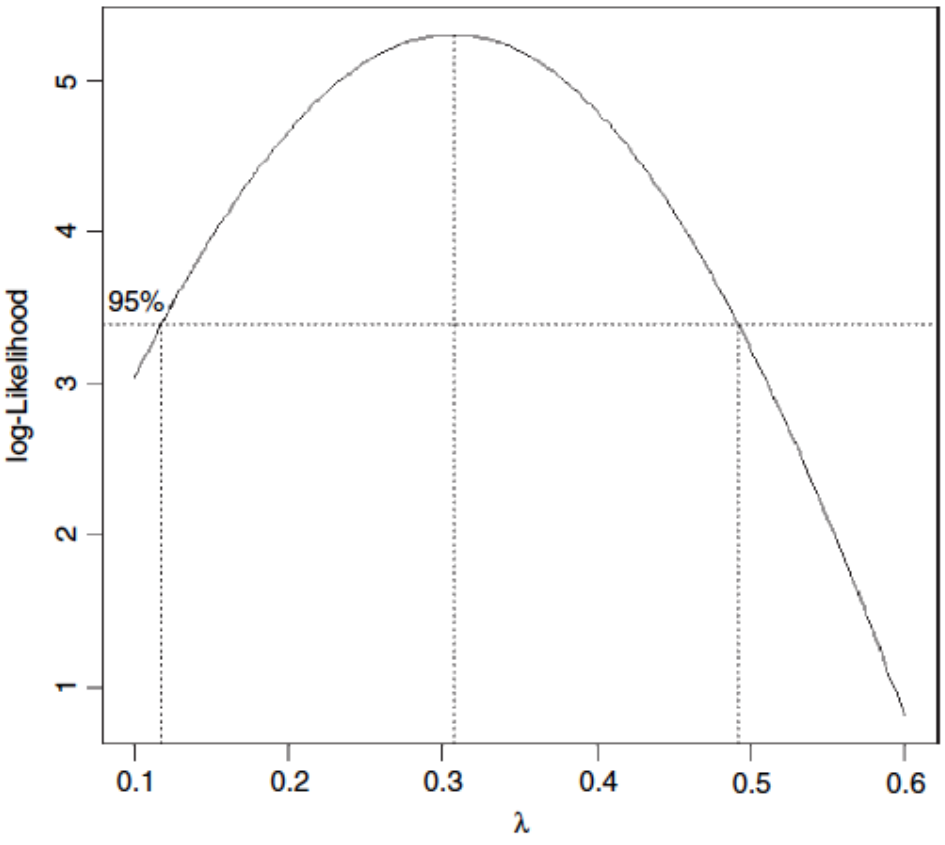
Τι θα συνέβαινε αν δεν είχαμε λογαριθμική μετατροπή των επεξηγηματικών μεταβλητών; Ποια θα ήταν η βέλτιστη μετατροπή του όγκου σε αυτή την περίπτωση; Για να το μάθουμε, επανάλαμβανουμε τη λειτουργία της συνάρτησης boxcox, απλά αλλάζοντας τον τύπο του μοντέλο σαν αυτόν:

```
boxcox(volume ~ girth+height)
```



Μπορούμε να μεγεθύνουμε από 0.1 έως 0.6 όπως παρακάτω:

```
boxcox(volume ~ girth+height,lambda=seq(0.1,0.6,0.01))
```



Αυτό υποδηλώνει ότι η μεταμόρφωση της κυβικής ρίζας θα ήταν καλύτερα (λ =1/3). Και πάλι, αυτό συμφωνεί με διαστατικά ορίσματα, δεδομένου ότι η απόκριση και οι επεξηγηματικές μεταβλητές θα έχουν όλες τις διαστάσεις του μήκους σε αυτή την περίπτωση.

Κριτική Μοντέλου

Υπάρχει ο πειρασμός να συνδεθείτε προσωπικά με ένα συγκεκριμένο μοντέλο. Οι στατιστικολόγοι το αποκαλούν 'ερωτεύοντας το μοντέλο σας'. Είναι καλό, να θυμάστε τις εξής αλήθειες για τα μοντέλα:

• Όλα τα μοντέλα είναι λάθος.

• Ορισμένα μοντέλα είναι καλύτερα από άλλα.

• Το σωστό μοντέλο δεν μπορεί ποτέ να είναι γνωστό με βεβαιότητα.

• Όσο απλούστερο το μοντέλο, τόσο καλύτερο είναι.

Υπάρχουν διάφοροι τρόποι με τους οποίους μπορούμε να βελτιώσουμε τα πράγματα, αν αποδειχτεί ότι παρόν μοντέλο μας είναι ανεπαρκές:

• Μεταμορφώστε τη μεταβλητή απόκρισης.

• Μετατρέψτε μία ή περισσότερες από τις επεξηγηματικές μεταβλητές.

• Δοκιμάστε να προσαρμόσετε διαφορετικές επεξηγηματικές μεταβλητές, αν έχετε.

• Χρησιμοποιήστε μια διαφορετική δομή σφάλματος.

• Χρησιμοποιήστε μη παραμετρικούς εξομαλυντές αντί παραμετρικές συναρτήσεις.

• Χρησιμοποιήστε διαφορετικά βάρη για διαφορετικές τιμές του y .

Όλα αυτά διερευνώνται στα επόμενα κεφάλαια. Στην ουσία, θα πρέπει να έχετε μια σειρά από εργαλεία για να διαπιστωθεί αν, και πώς, το μοντέλο σας είναι ανεπαρκές. Για παράδειγμα, το μοντέλο μπορεί:

• να προβλέπει κακώς, ορισμένες από τις τιμές του y ·

• να δείχνει μη σταθερή διακύμανση·

• να παρουσιάζει μη φυσιολογικά λάθη·

• να επηρεάζεται έντονα από ένα μικρό αριθμό σημείων δεδομένων επιρροής·

• να παρουσιάζει κάποιο είδος συστηματικού μοτίβου των υπολοίπων·

• να επιδεικνύει υπερδιασπορά

Ελέγχοντας το μοντέλο

Αφότου έχετε προσαρμόσει ένα μοντέλο στα δεδομένα θα πρέπει να διερευνηθεί πόσο καλά το μοντέλο περιγράφει τα δεδομένα. Ειδικότερα, θα πρέπει να εξετάσουμε για να δούμε αν υπάρχουν συστηματικές τάσεις για την καλή προσαρμογή. Για παράδειγμα, αυξάνεται η καλή προσαρμογή με τον αριθμό παρατήρησης, ή είναι μια συνάρτηση ενός ή περισσοτέρων επεξηγηματικών μεταβλητών; Μπορούμε να εργαστούμε με τα ανεπεξέργαστα υπόλοιπα:

$$\text{Υπόλοιπα} = y - \text{προσαρμοσμένες τιμές}$$

Για παράδειγμα, θα πρέπει να σχεδιάσουμε συστηματικά τα υπόλοιπα συναρτήσει:

• των προσαρμοσμένων τιμών (για να ψάξουν για ετεροσκεδαστικότητα)·

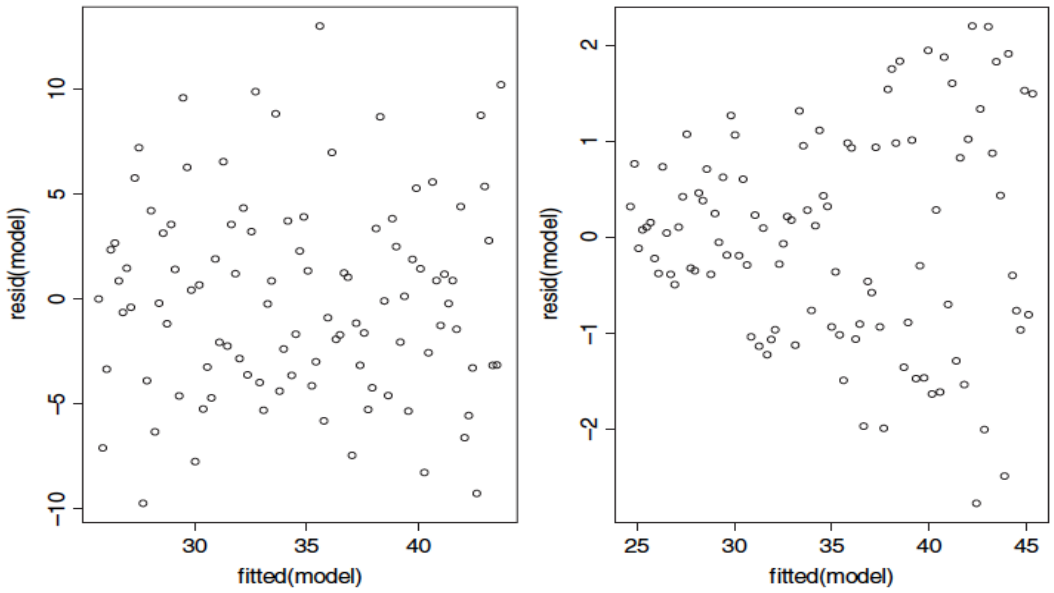
• των επεξηγηματικών μεταβλητών (για να ψάξουν για στοιχεία καμπυλότητας)·

• της ακολουθίας της συλλογής δεδομένων (για να ψάξουν για την χρονική συσχέτιση)·

• της κανονικής τυπικής απόκλισης (για να ψάξουν για μη κανονικότητα των σφαλμάτων).

Ετεροσκεδαστικότητα

Ένα καλό μοντέλο πρέπει επίσης να αντιπροσωπεύει την επαρκή σχέση διακύμανσης-μέσης τιμής και να παράγει αθροιστικές επιδράσεις στην κατάλληλη κλίμακα (όπως ορίζεται από τη συνάρτηση link). Το διάγραμμα τυποποιημένων υπολοίπων συναρτήσει προσαρμοσμένων τιμών θα πρέπει να μοιάζει με τον ουρανό τη νύχτα (σημεία διάσπαρτα τυχαία σε όλη την περιοχή σχεδίασης), χωρίς τάση στο μέγεθος ή το βαθμό διασποράς των υπολοίπων. Ένα κοινό πρόβλημα είναι ότι η διακύμανση αυξάνεται με τη μέση τιμή, έτσι ώστε παρατηρούμε ένα επεκτείνομενο, σε σχήμα βεντάλιας μοτίβο των υπολοίπων (πίνακας δεξιά).



Το διάγραμμα στα αριστερά είναι αυτό που θέλουμε να δούμε: χωρίς τάση των υπολοίπων με τις προσαρμοσμένες τιμές. Το διάγραμμα στα δεξιά είναι ένα πρόβλημα. Υπάρχει ένα σαφές σχέδιο αύξησης των υπολοίπων, καθώς οι προσαρμοσμένες τιμές μεγαλώνουν. Αυτή είναι μια εικόνα του πως μοιάζει η ετεροσκεδαστικότητα.

Μη κανονικότητα των σφαλμάτων

Τα σφάλματα μπορεί να είναι μη-κανονικά για διάφορους λόγους. Μπορούν να είναι μη συμμετρικά, με μακριές ουρές προς τα αριστερά ή προς τα δεξιά. Ή μπορουν να είναι κυρτά, με επίπεδη ή πιο αιχμηρή κορυφή στην κατανομή τους. Σε κάθε περίπτωση, η θεωρία αυτή βασίζεται στην υπόθεση των κανονικών σφαλμάτων, και αν τα σφάλματα αυτά *δεν* είναι κανονικά κατανεμημένα, τότε δεν θα έπρεπε να γνωρίζουμε πώς αυτό επηρεάζει την ερμηνεία των δεδομένων μας ή τις συνέπειες αυτής.

Χρειάζεται μεγάλη εμπειρία για την ερμηνεία διαγραμμάτων κανονικών σφαλμάτων. Εδώ έχουμε δημιουργήσει μια σειρά από σύνολα δεδομένων όπου εισάγουμε διαφορετικά, αλλά γνωστά είδη των μη κανονικών σφαλμάτων. Στη συνέχεια, τα σχεδιάσουμε χρησιμοποιώντας μία απλή ερασιτεχνική συνάρτηση που ονομάζεται mcheck (αναπτύχθηκε για πρώτη φορά από τον John Nelder στη γλώσσα του πρωτοτύπου GLIM· το όνομα αντιστοιχεί στον έλεγχο μοντέλου). Η ιδέα είναι να δούμε τι σχέδια δημιουργούνται στα κανονικά διαγράμματα από τα διάφορα είδη μη κανονικότητας. Σε πραγματικές εφαρμογές θα χρησιμοποιείτε το γενικό plot(model) και όχι το mcheck (βλ. παρακάτω). Πρώτα, γράφουμε την συνάρτηση mcheck. Η ιδέα είναι να παράγει δύο διαγραμμάτα, πλάι-πλάι: σε διάγραμμα των υπολοίπων συναρτήσει των προσαρμοσμένων τιμών στα αριστερά, και ένα διάγραμμα των παραγγελθέντων υπολοίπων συναρτήσει των ποσοστιαίων σημείων της κανονικής κατανομής στα δεξιά.

```
mcheck <-function (obj, . . . ) {
  rs<-obj$resid
  fv<-obj$fitted
  par(mfrow=c(1,2)
  plot(fv,rs,xlab="Fitted values",ylab="Residuals")
  abline(h=0, lty=2)
  qqnorm(rs,xlab="Normal scores",ylab="Ordered residuals",main="")
  qqline(rs,lty=2)
  par(mfrow=c(1,1))
  invisible(NULL)                }
```

Σημειώστε τη χρήση του $ (επιλογή συνιστώσας) για την εξαγωγή των υπολοίπων και των προσαρμοσμένων τιμών από το αντικείμενο μοντέλο το οποίο διοχετεύεται στη συνάρτηση ως obj (η έκφραση x$name είναι το όνομα της *συνιστώσας* του x). Οι συναρτήσεις qqnorm και qqline είναι ενσωματωμένες συναρτήσεις για να παράγουν διαγράμματα κανονικών πιθανοτήτων. Είναι καλή πρακτική προγραμματισμού να ρυθμίζετε τις παραμέτρους γραφικών στις προεπιλεγμένες ρυθμίσεις τους πριν από την έξοδο από την συνάρτηση.

Ο στόχος είναι να δημιουργηθεί ένας κατάλογος μερικών από των πιο κοινών προβλημάτων που προκύπτουν στον έλεγχο μοντέλου. Χρειαζόμαστε ένα διάνυσμα x τιμών για τα εξής μοντέλα παλινδρόμησης:

```
x<-0:30
```

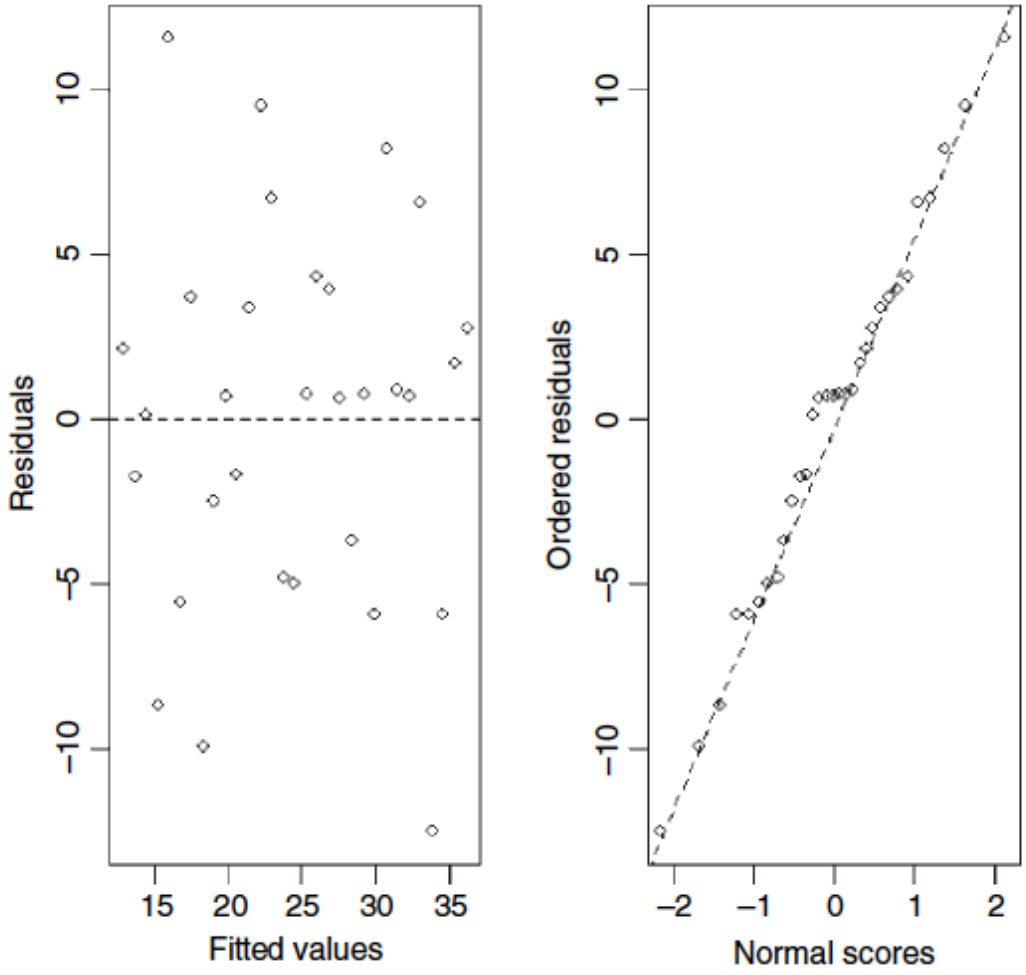
Τώρα έχουμε την κατασκευή των μεταβλητών απόκρισης σύμφωνα με την εξίσωση

$$y = 10 + x + \varepsilon$$

όπου τα σφάλματα, ε, έχουν μηδενική μέση, αλλά έχουν ληφθεί από διαφορετικές κατανομές πιθανότητας σε κάθε περίπτωση.

Κανονικά σφάλματα

```
e<-rnorm(31,mean=0,sd=5)
yn<-10+x+e
mn<-lm(yn ~ x)
mcheck(mn)
```



Δεν υπάρχει πρόταση της μη-σταθερής διακύμανσης (αριστερό διάγραμμα) και του κανονικού διαγράμματος (δεξιά) να είναι λογικά ευθεία. Η απόφαση ως προς το τι συνιστά μια σημαντική απόκλιση από την κανονικότητα χρειάζεται εμπειρία, και αυτός είναι ο λόγος για την εξέταση σε κάποιες ευδιάκριτες μη κανονικές, αλλά γνωστές, δομές σφάλματος παρακάτω.

Ομοιόμορφα σφάλματα

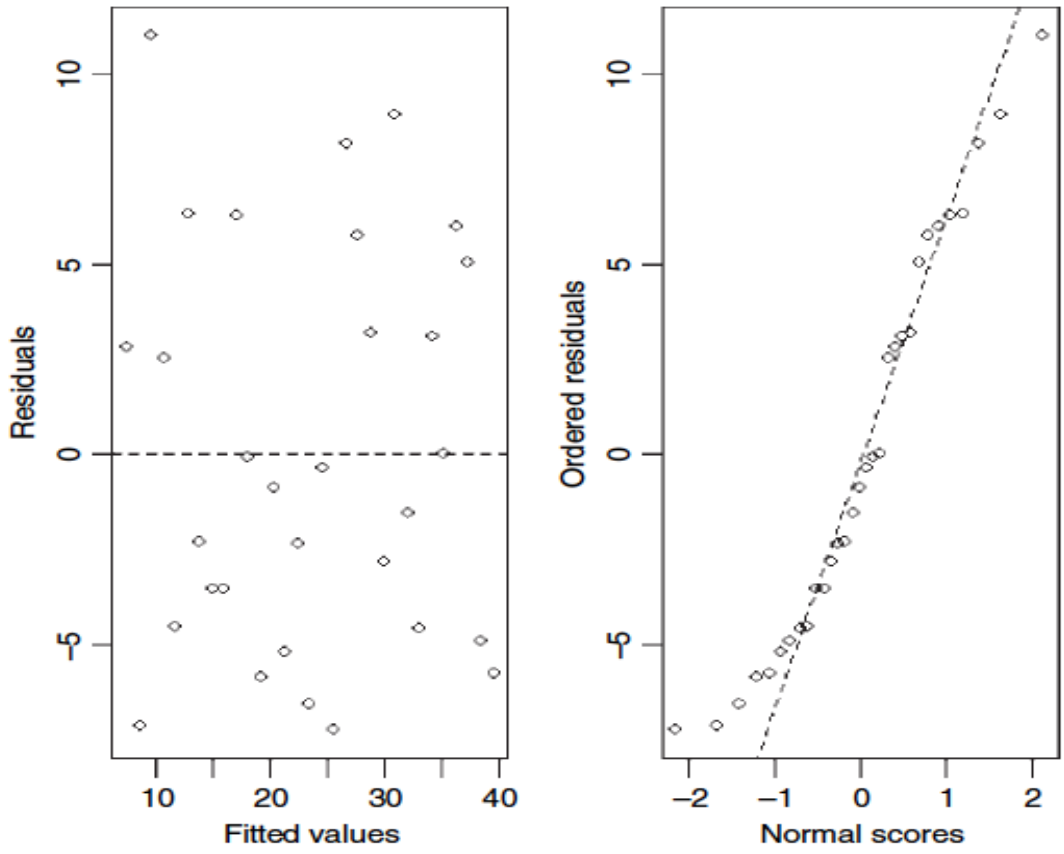
```
eu<-20*(runif(31)-0.5
yu<-10+x+eu
mu<-lm(yu ~ x)
mcheck(mu)
```

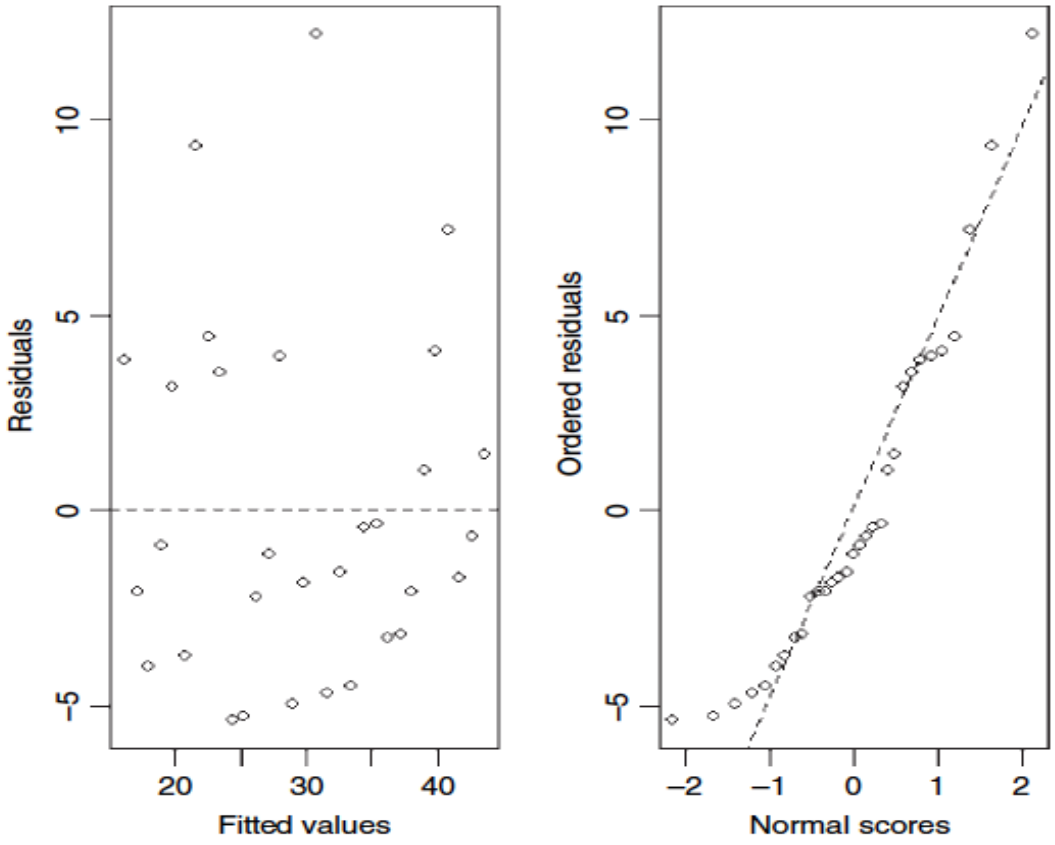
Τα ομοιόμορφα σφάλματα εμφανίζονται ως ευδιάκριτα σχέδια, σχήματος S στο διάγραμμα QQ στα δεξιά. Η προσαρμογή στο κέντρο είναι μια χαρά, αλλά το μεγαλύτερο και το μικρότερο των υπολείπων είναι πολύ μικρά (που περιορίζονται σε αυτό το παράδειγμα να είναι ± 10).

Αρνητικά διωνυμικά σφάλματα

```
enb<-rnbinom(31,2,.3)
ynb<-10+x+enb
mnb<-lm(ynb ~ x)
mcheck(mnb)
```

Τα μεγάλα αρνητικά υπόλοιπα είναι όλα πάνω από τη γραμμή, αλλά το πιο προφανές χαρακτηριστικό του διαγράμματος είναι το ενιαίο, πολύ μεγάλο θετικό υπόλοιπο (στην πάνω δεξιά γωνία). Σε γενικές γραμμές, αρνητικά διωνυμικά λάθη θα παράγουν ένα J-σχήματος στο διάγραμμα QQ. Τα μεγαλύτερα θετικά υπόλοιπα είναι υπερβολικά μεγάλα για να έχουν προέλθει από μια κανονική κατανομή. Οι τιμές αυτές μπορούν να αποδειχθούν να έχουν ιδιαίτερα μεγάλη επιρροή (βλ. παρακάτω).



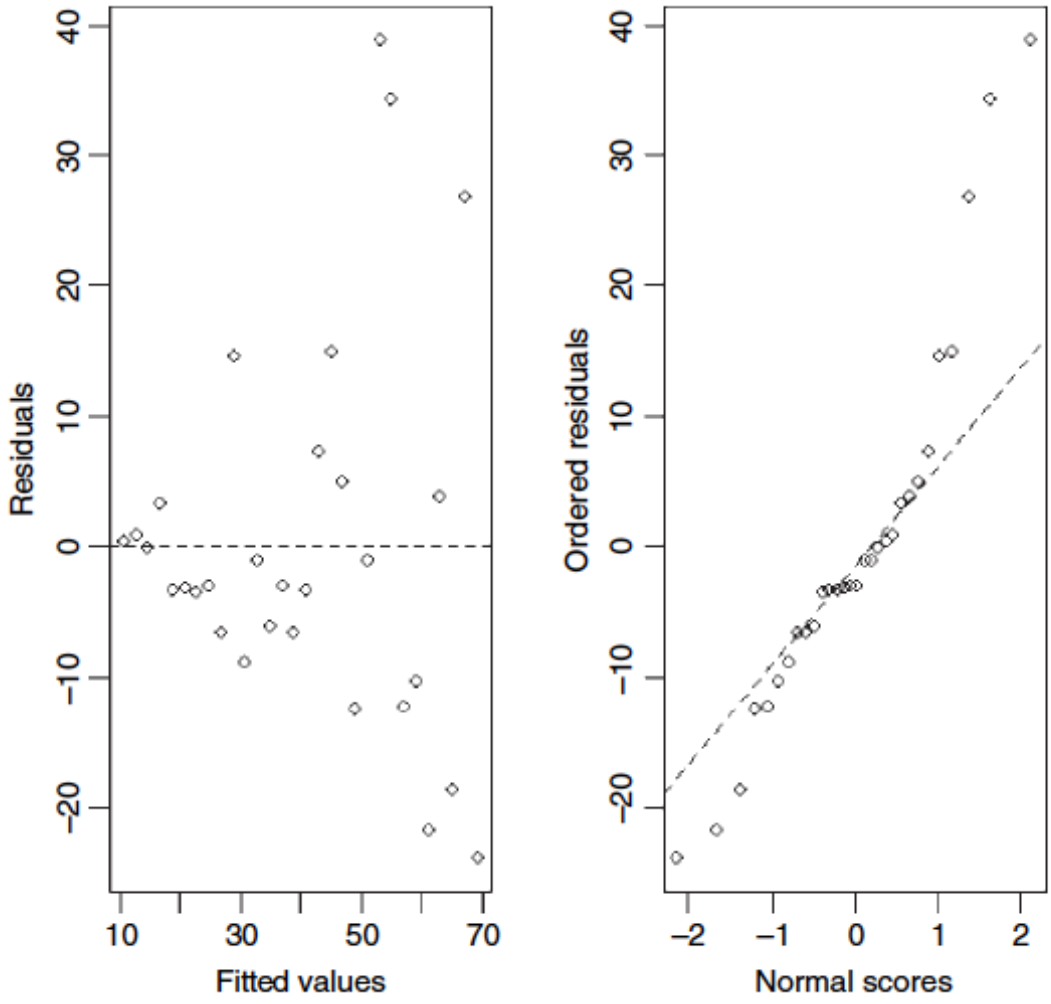


Σφάλματα Γάμμα και αυξάνοντας την διακύμανση

Εδώ η παράμετρος σχήματος έχει οριστεί σε 1 και η παράμετρος ρυθμού στο $1/x$, και η διακύμανση αυξάνεται με το τετράγωνο της μέσης τιμής.

```
eg<-rgamma(31,1,1/x)
yg<-10+x+eg
mg<-lm(yg ~ x)
mcheck(mg)
```

Το αριστερό διάγραμμα δείχνει τα υπόλοιπα να αυξάνονται απότομα με τις προσαρμοσμένες τιμές, και απεικονίζει μια ασυμμετρία μεταξύ του μεγέθους των θετικών και αρνητικών υπολοίπων. Το δεξιό διάγραμμα δείχνει την εξαιρετικά μη κανονική κατανομή των σφαλμάτων.



Επιρροή

Ένας από τους συνηθέστερους λόγους για την έλλειψη της προσαρμογής είναι μέσα από την ύπαρξη ακραίων τιμών στα δεδομένα. Είναι σημαντικό να κατανοήσουμε, ωστόσο, ότι ένα σημείο μπορεί να *φαίνεται* να είναι μια ακραία τιμή, λόγω της ατέλειας του μοντέλου, και όχι επειδή δεν υπάρχει κάτι λάθος με τα δεδομένα. Είναι σημαντικό να κατανοήσουμε ότι η ανάλυση των υπολοίπων είναι ένας πολύ φτωχός τρόπος αναζήτησης για επιρροή. Ακριβώς επειδή ένα σημείο έχει ιδιαίτερα μεγάλη επιρροή, αναγκάζει τη γραμμή παλινδρόμησης κοντά σε αυτό, και ως εκ τούτου το σημείο επιρροής μπορεί να έχει ένα πολύ μικρό υπόλοιπο.

Πάρτε αυτόν τον κύκλο των δεδομένων που δείχνει την μη σχέση μεταξύ y και x:

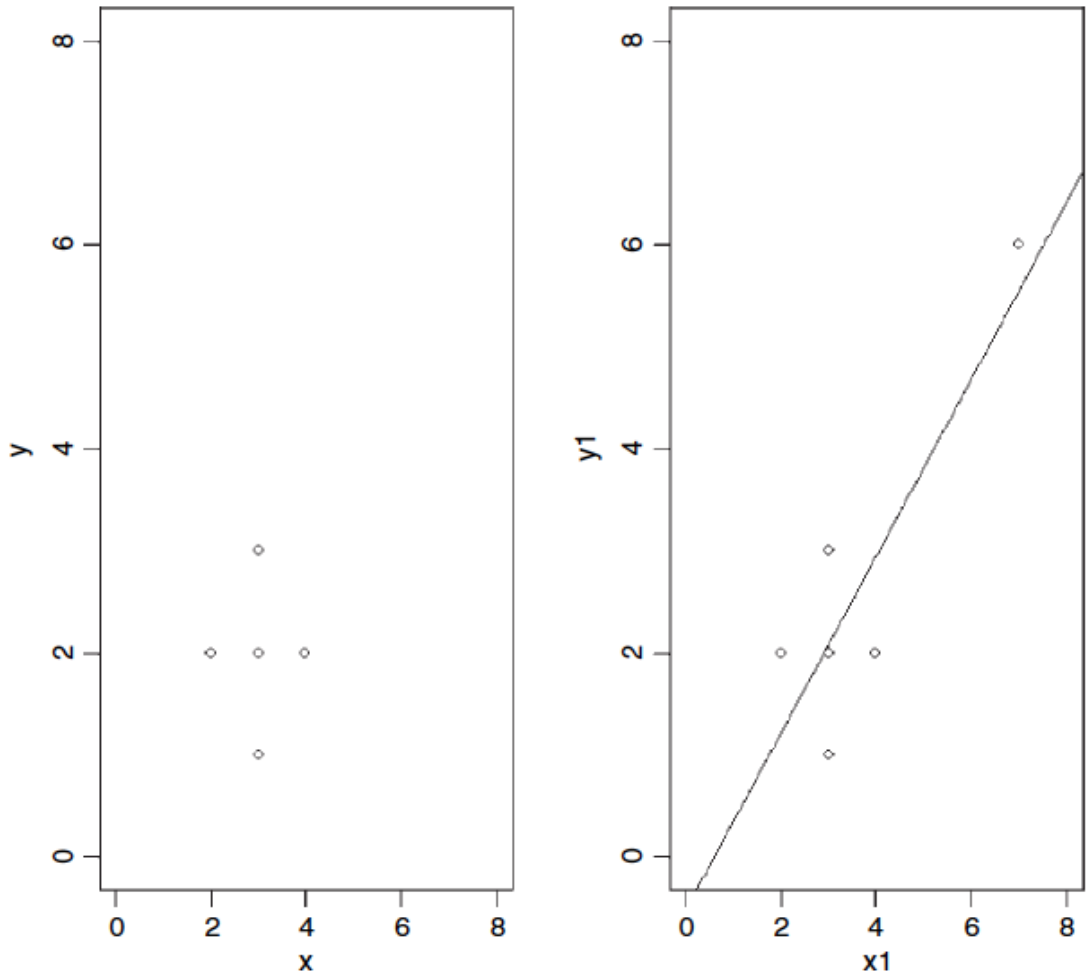
```
x<-c(2,3,3,3,4)
y<-c(2,3,2,1,2)
```

Θέλουμε να σχεδιάσουμε δύο γραφήματα δίπλα-δίπλα, και θέλουμε να έχουν τις ίδιες κλίμακες στους άξονες:

```
par(mfrow=c(1,2))
plot(x,y,xlim=c(0,8,),ylim=c(0,8))
```

Προφανώς, δεν υπάρχει καμία σχέση μεταξύ *y* και *x* στα αρχικά δεδομένα. Αλλά ας προσθέσουμε μια ακραία τιμή στο σημείο (7,6) χρησιμοποιώντας συνένωση και να δούμε τι θα συμβεί.

```
x1<-c(x,7)
y1<-c(y,6)
plot(x1,y1,xlim=c(0,8,),ylim=c(0,8))
abline(lm(y1 ~ x1))
```



Τώρα, υπάρχει μια σημαντική παλινδρόμηση του *y* επί του *x*. Η ακραία τιμή λέγεται ότι έχει μεγάλη επιρροή.

Για να μειωθεί η επιρροή των ακραίων τιμών, υπάρχουν μια σειρά από σύγχρονες τεχνικές που είναι γνωστές ως **ισχυρή παλινδρόμηση (robust regression).** Για να δείτε ένα από αυτά σε δράση, ας κάνουμε μια απλή γραμμική παλινδρόμηση αυτών των δεδομένων και εκτυπώστε την περίληψη:

```
reg<-lm(y1 ~ x1)
summary(reg)

Residuals:
      1       2       3       4       5       6
0.78261 0.91304 -0.08696 -1.08696 -0.95652 0.43478

Coefficients:
             Estimate  Std. Error     t value   Pr(>|t|)
(Intercept)   -0.5217      0.9876      -0.528     0.6253
x1             0.8696      0.2469       3.522     0.0244 *

Residual standard error: 0.9668 on 4 degrees of freedom
Multiple R-Squared: 0.7561, Adjusted R-squared: 0.6952
F-statistic: 12.4 on 1 and 4 DF, p-value: 0.02441
```

Οι τιμές των υπολοίπων κάνουν σημαντικό το σημείο ότι η ανάλυση των υπολοίπων είναι είναι ένας πολύ φτωχός τρόπος αναζήτησης για επιρροή. Ακριβώς επειδή το σημείο του αριθμού 6, έχει τόσο μεγάλη επιρροή, αναγκάζει τη γραμμή παλινδρόμησης να έρθει κοντά σε αυτό, και ως εκ τούτου το σημείο του αριθμού 6 έχει ένα μικρό υπόλοιπο (όπως μπορείτε να δείτε, 0.4348 είναι το δεύτερο μικρότερο από το σύνολο των υπολοίπων). Η κλίση της γραμμής παλινδρόμησης είναι 0.8696 με ένα πρότυπο σφάλμα του 0.2469, και αυτό είναι σημαντικά διαφορετικό από 0 ($p = 0.0224$) παρά το μικρό μέγεθος του δείγματος.

Συνεχίζοντας την ανάλυση της απλής γραμμικής παλινδρόμησης στη σελ. 267, εξετάζουμε τη συνάρτηση lm.influence. Αυτή παράγει τέσσερις συνιστώσες: $hat, $coefficients, $sigma και $wt.res (σταθμισμένα υπολείμματα):

```
lm.influence(lm(growth~tannin))


$hat
        1         2         3         4         5         6
0.3777778 0.2611111 0.1777778 0.1277778 0.1111111 0.1277778
        7         8         9
0.1777778 0.2611111 0.3777778



$coefficients
  (Intercept)            tannin
1   0.14841270    -2.619048e-02
2  -0.22690058     3.646617e-02
3  -0.39309309     5.360360e-02
4   0.58995046     5.530786e-02
5  -0.11111111    -2.794149e-18
6   0.06765747     2.537155e-02
7   0.06636637    -9.954955e-02
8   0.02873851    -1.616541e-02
9  -0.24444444     1.047619e-01
```

```
$sigma

       1        2        3        4        5        6
1.824655 1.811040 1.729448 1.320801 1.788078 1.734501
       7        8        9
1.457094 1.825513 1.757636


$wt.res
         1          2          3         4          5
0.2444444 -0.5388889 -1.3222222 2.8944444 -0.8888889
         6          7          8         9
1.3277778 -2.4555556 -0.2388889 0.9777778
```

Ας ρίξουμε μια ματιά σε κάθε ένα από αυτά τα στοιχεία με τη σειρά.

Η πρώτη συνιστώσα, $hat, είναι ένα διάνυσμα που περιέχει τη διαγώνιο του πίνακα hat. Αυτός είναι ο ορθογώνιος πίνακας προβολής στο χώρο μοντέλου (βλ. σελ. 273). Ο πίνακας X αποτελείται από μια στήλη από μονάδες (που αντιστοιχεί στο σημείο τομής) και μια στήλη των τιμών x (η επεξηγηματική μεταβλητή· βλ. σελ. 271):

```
X<-cbind(1,tannin)
```

Στη συνέχεια, ο πίνακας hat, H, δίνεται από $H = X(X'X)^{-1}X'$, όπου X' είναι η μεταφορά του X:

```
H<-X%*%ginv(t(X)%*%X)%*%t(X)
```

και θέλουμε τη διαγώνιο αυτού, που θα μπορούσαμε να πάρουμε πληκτρολογώντας:

```
diag(H)
```

Οι μεγάλες τιμές των στοιχείων αυτού του διάνυσματος σημαίνει ότι αλλάζοντας το y_i θα είχαμε μεγάλο αντίκτυπο στις προσαρμοσμένες τιμές· για παράδειγμα οι διαγώνιοι του hat είναι τα μέτρα της μόχλευσης του y_i.

Στη συνέχεια, $coefficients είναι ένας πίνακας του οποίου η i-στη γραμμή περιέχει την *αλλαγή* των εκτιμώμενων συντελεστών που προκύπτει όταν η i-στη υπόθεση πέσει από την παλινδρόμηση (αυτό είναι διαφορετικό από το S-PLUS, που δείχνει τους ίδιους τους *συντελεστές*). Τα δεδομένα στη σειρά 9 έχουν τη μεγαλύτερη επίδραση στην κλίση και τα δεδομένα στη γραμμή 4 έχουν τη μεγαλύτερη επίδραση στο σημείο τομής.

Η τρίτη συνιστώσα, $sigma, είναι ένα διάνυσμα του οποίου το i-στο στοιχείο περιέχει την εκτίμηση του υπολειμματικού πρότυπου σφάλματος που λαμβάνεται όταν η i-στη περίπτωση πέσει από την παλινδρόμηση· έτσι 1.824655 είναι το υπολειπόμενο πρότυπο σφάλμα, όταν το σημείο στο νούμερο 1 ρίπτεται lm(growth[- 1] ~ tannin[-1]), και η διακύμανση του σφάλματος στην περίπτωση αυτή είναι $1.824655^2 = 3.329$.

Τέλος, $wt.res είναι ένα διάνυσμα των σταθμισμένων υπολοίπων (ή υπολείμματα αποκλίνουσας συμπεριφοράς σε ένα γενικευμένο γραμμικό μοντέλο) ή ακατέργαστα υπόλοιπα, εάν τα βάρη δεν έχουν ρυθμιστεί (όπως σε αυτό το παράδειγμα).
Μια δέσμη συναρτήσεων είναι διαθέσιμη για να υπολογίσει μερικές από τις διαγνώσεις παλινδρόμησης (διαγραφή αφήνοντας ένα εκτός-leave-one-out) για γραμμικά και γενικευμένα γραμμικά μοντέλα:

```
influence.measures(lm(growth ~ tannin))
```

```
Influence measures of
                lm(formula =growth~tannin) :
     dfb.1_    dfb.tnnn   dffit  cov.r   cook.d    hat  inf
1    0.1323   -1.11e-01  0.1323  2.167  0.01017  0.378    *
2   -0.2038    1.56e-01 -0.2058  1.771  0.02422  0.261
3   -0.3698    2.40e-01 -0.3921  1.323  0.08016  0.178
4    0.7267   -3.24e-01  0.8981  0.424  0.24536  0.128
5   -0.1011   -1.21e-17 -0.1864  1.399  0.01937  0.111
6    0.0635    1.13e-01  0.3137  1.262  0.05163  0.128
7    0.0741   -5.29e-01 -0.8642  0.667  0.27648  0.178
8    0.0256   -6.86e-02 -0.0905  1.828  0.00476  0.261
9   -0.2263    4.62e-01  0.5495  1.865  0.16267  0.378    *
```

Τα ονόματα στήλης έχουν ως εξής: `dfb` = DFBETAS, `dffit` = DFFITS (οι δύο όροι εξηγούνται Cook και Weisberg, 1982, σελ. 124-125), `cov.r` = λόγος συνδιακύμανσης , `cook.d` = απόσταση Cook, `hat` = η διαγώνιος του πίνακα hat, και `inf` σηματοδοτεί τα σημεία επιρροής δεδομένων με αστερίσκο. Μπορείτε να εξαγάγετε τα σημεία επιρροής για τη σχεδίαση με $is.inf σαν αυτό:

```
modi<-influence.measures(lm(growth ~ tannin))
which(apply(modi$is.inf, 1, any))
```

```
1     9
```

```
1     9
```

```
growth[which(apply(modi$is.inf, 1, any))]
```

```
[1] 12  3tannin[which(apply(modi$is.inf, 1, any))]
```

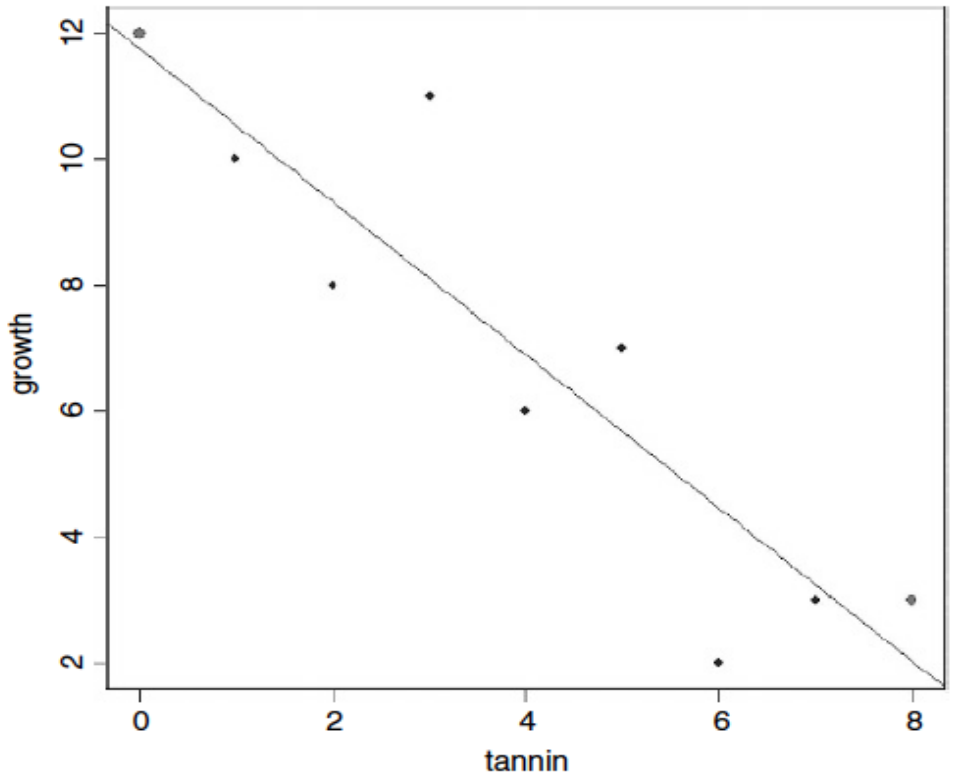
```
[1] 0 8
```

```
summary(modi)
```

```
Potentially influential observations of
                    lm(formula = growth~tannin) :

     dfb.1_   dfb.tnnn    dffit    cov.r    cook.d    hat
1      0.13      -0.11     0.13    2.17_*     0.01   0.38
9     -0.23       0.46     0.55    1.87_*     0.16   0.38
```

```
yp<-growth[which(apply(modi$is.inf, 1, any))]
xp<-tannin[which(apply(modi$is.inf, 1, any))]
plot(tannin,growth,pch=16) points(xp,yp,col="red",cex=1.3,pch=16)
abline(model)
```



Η κλίση θα είναι πολύ πιο απότομη αν δεν ήταν τα δύο σημεία πάνω αριστερά και κάτω δεξιά που δεν εμφανίζονται ως μαύρα διαμάντια - αυτά θα εμφανιστούν με κόκκινο χρώμα στην οθόνη σας:

```
coef(lm(growth ~ tannin))
```

```
(Intercept)      tannin
11.755556     -1.216667
```

```
coef(lm(growth ~ tannin,subset=(1:9 != 1 & 1:9 != 9)))
```

```
(Intercept)      tannin
12.000000     -1.321429
```

Περίληψη των Στατιστικών Μοντέλων στην R

Τα μοντέλα προσαρμόζονται με μία από τις ακόλουθες συναρτήσεις προσαρμογής μοντέλου:

lm προσαρμόζει ένα γραμμικό μοντέλο με κανονικά σφάλματα και σταθερή διακύμανση· γενικά αυτή χρησιμοποιείται για την ανάλυση παλινδρόμησης με τη χρήση συνεχών επεξηγηματικών μεταβλητών.

aov προσαρμόζει την ανάλυση της διακύμανσης με κανονικά σφάλματα, τη σταθερή διακύμανση και τη σύνδεση ταυτότητας· χρησιμοποιείται γενικά για τις κατηγορικές επεξηγηματικές μεταβλητές ή ANCOVA με ένα μείγμα από κατηγορηματικές και συνεχείς επεξηγηματικές μεταβλητές.

glm προσαρμόζει γενικευμένα γραμμικά μοντέλα για δεδομένα χρησιμοποιώντας κατηγορηματικές ή συνεχείς επεξηγηματικές μεταβλητές, καθορίζοντας μια δομή από την οικογένεια των **δομών σφάλματος (error structures)** (π.χ. Poisson για καταμέτρηση δεδομένων ή διωνυμική αναλογία για τα δεδομένα) και μια συγκεκριμένη **συνάρτηση σύνδεσης (link function).**

gam προσαρμόζει γενικευμένα προσθετικά μοντέλα των δεδομένων με μια δομή από την οικογένεια των δομών σφάλματος (π.χ. Poisson για καταμέτρηση δεδομένων ή διωνυμική αναλογία για τα δεδομένα), στην οποία οι συνεχείς επεξηγηματικές μεταβλητές μπορούν (προαιρετικά) να προσαρμοστούν ως αυθαίρετες ομαλοποιημένες συναρτήσεις χρησιμοποιώντας μη παραμετρικούς εξομαλυντές αντί των ειδικών παραμετρικών συναρτήσεων.

lme και lmer προσαρμόζει γραμμικά μικτής επίδρασης μοντέλα με καθορισμένα μίγματα σταθερών και τυχαίων επιδράσεων και επιτρέπει τον προσδιορισμό της δομής συσχέτισης μεταξύ των επεξηγηματικών μεταβλητών και της αυτοσυσχέτισης της μεταβλητής απόκρισης (π.χ. χρονοσειρές επιδράσεων με επαναλαμβανόμενες μετρήσεις). Η lmer επιτρέπει μη κανονικά σφάλματα και μη σταθερή διακύμανση με τις ίδιες οικογένειες σφάλματος όπως η GLM.

nls προσαρμόζει ένα μη-γραμμικό μοντέλο παλινδρόμησης μέσω ελαχίστων τετραγώνων, εκτιμώντας τις παραμέτρους μιας συγκεκριμένης μη γραμμικής συνάρτησης.

nlme προσαρμόζει μια καθορισμένη μη γραμμική συνάρτηση σε ένα μικτής επίδρασης μοντέλο, όπου οι παράμετροι της μη-γραμμικής συνάρτησης υποτίθεται ότι είναι τυχαίες επιδράσεις· επιτρέπει τον προσδιορισμό της δομής συσχέτισης μεταξύ των επεξηγηματικών μεταβλητών και της αυτοσυσχέτισης της μεταβλητής απόκρισης (π.χ. χρονοσειρές επιδράσεων με επαναλαμβανόμενες μετρήσεις).

loess προσαρμόζει ένα τοπικό μοντέλο παλινδρόμησης με μία ή περισσότερες συνεχείς επεξηγηματικές μεταβλητές χρησιμοποιώντας μη παραμετρικές τεχνικές για να παράγουν μία ομαλοποιημένη επιφάνεια μοντέλου.

tree προσαρμόζει ένα μοντέλο δέντρου παλινδρόμησης χρησιμοποιώντας δυαδική αναδρομική διαμέριση, οπότε τα δεδομένα χωρίζονται διαδοχικά κατά μήκος των αξόνων συντεταγμένων των επεξηγηματικών μεταβλητών, έτσι ώστε σε κάθε κόμβο, ο διαχωρισμός που επιλέγεται διακρίνει την μέγιστη μεταβλητή απόκρισης στους αριστερούς και δεξιούς κλάδους. Με μια κατηγορική μεταβλητή απόκρισης, το δέντρο ονομάζεται δέντρο ταξινόμησης, καθώς και το μοντέλο που χρησιμοποιήθηκε για την ταξινόμηση προϋποθέτει ότι η μεταβλητή απόκρισης ακολουθεί πολυωνυμική κατανομή.

Για τα περισσότερα από αυτά τα μοντέλα, μια σειρά από **γενικές συναρτήσεις** μπορούν να χρησιμοποιηθούν για την απόκτηση πληροφοριών σχετικά με το μοντέλο. Η σημαντικότερες και πιο συχνά χρησιμοποιούμενες έχουν ως εξής:

summary παράγει εκτιμήσεις των παραμέτρων και πρότυπα σφάλματα από την lm και πίνακες ANOVA από την aov· αυτό θα καθορίσει συχνά την επιλογή σας ανάμεσα σε lm και σε aov. Για κάθε lm ή aov μπορείτε να επιλέξετε summary.aov ή summary.lm για να πάρετε την εναλλακτική μορφή της εξόδου (έναν πίνακα ANOVA ή έναν πίνακα των εκτιμήσεων των παραμέτρων και πρότυπων σφαλμάτων· Βλ. σελ. 364.)

plot παράγει διαγνωστικά διαγράμματα για τον έλεγχο μοντέλου, συμπεριλαμβανομένων των υπολοίπων συναρτήσει των προσαρμοσμένων τιμών, ελέγχων επιρροής, κλπ.

anova είναι μια θαυμάσια χρήσιμη συνάρτηση για τη σύγκριση διαφορετικών μοντέλων και την παραγωγή πινάκων ANOVA.

update χρησιμοποιείται για να τροποποιήσει την τελευταία προσαρμογή του μοντέλου· σώζει τόσο την προσπάθεια δακτυλογράφησης όσο και χρόνο υπολογισμού.

Άλλες χρήσιμες γενικές συναρτήσεις περιλαμβάνουν τα ακόλουθα:

coef δίνει τους συντελεστές (εκτιμώμενες παραμέτροι) από το μοντέλο.

fitted δίνει τις προσαρμοσμένες τιμές, προβλέπεται από το μοντέλο για τις τιμές των επεξηγηματικών μεταβλητών που περιλαμβάνονται.

resid δίνει τα υπολοίπα (διαφορές μεταξύ μετρούμενων και προβλεπόμενων τιμών του y).

predict χρησιμοποιεί πληροφορίες από το προσαρμοσμένο μοντέλο για την παραγωγή ομαλών συναρτήσεων για τη χάραξη μιας γραμμής μέσω του γραφήματος σκέδασης των δεδομένων σας.

Προαιρετικά ορίσματα σε συναρτήσεις μοντέλου προσαρμογής

Εκτός αν υποστηρίζετε το αντίθετο, όλες οι σειρές του πλαισίου δεδομένων θα χρησιμοποιηθούν στην προσαρμογή μοντέλου, δεν θα υπάρχουν μετατοπίσεις, και σε όλες τις τιμές της μεταβλητής απόκρισης θα δοθεί ίδια βαρύτητα. Μεταβλητές ονομαζόμενες στον τύπο μοντέλου θα προέρχονται από το οριοθετημένο πλαίσιο δεδομένων (data=mydata), από τη συνάρτηση with (σελ. 18) ή από το attached πλαίσιο δεδομένων (εάν υπάρχει). Εδώ απεικονίζουμε τις παρακάτω επιλογές:

- subset
- weights
- data
- offset
- na.action

Θα εργαστούμε με ένα παράδειγμα που περιλαμβάνει την ανάλυση συνδιακύμανσης (σελ. 490 για λεπτομέρειες) όπου έχουμε μία μίξη από συνεχείς και από κατηγορηματικές επεξηγηματικές μεταβλητές:

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

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

Η απάντηση είναι η παραγωγή σπόρων (Fruit) με συνεχή επεξηγηματική μεταβλητή (Root diameter) και διεπίπεδο παράγοντα Grazing (Grazed and Ungrazed).

Subset (Υποσύνολα)

Ίσως η πιο συνηθισμένη επιλογή μοντελοποίησης είναι να προσαρμοστεί το μοντέλο σε ένα υποσύνολο των δεδομένων (π.χ. προσαρμογή του μοντέλου στα δεδομένα από τα φυτά που μόλις χρησιμοποιούνται για βοσκή-grazed plants). Θα μπορούσατε να το κάνετε αυτό χρησιμοποιώντας δείκτες για τη μεταβλητή απόκρισης και όλες τις επεξηγηματικές μεταβλητές:

```
model<-lm(Fruit[Grazing=="Grazed"] ~ Root[Grazing=="Grazed"])
```

αλλά είναι πολύ πιο απλό να χρησιμοποιήθει το όρισμα subset, ειδικά όταν υπάρχουν πολλές επεξηγηματικές μεταβλητές:

```
model<-lm(Fruit ~ Root,subset=(Grazing=="Grazed"))
```

Η απάντηση, φυσικά, είναι η ίδια και στις δύο περιπτώσεις, αλλά οι πίνακες summary.lm και summary.aov είναι πιο τακτοποιημένοι με την subset. Σημειώστε τις παρενθέσεις που χρησιμοποιήθηκαν με την επιλογή subset (όχι τις αγκύλες που χρησιμοποιούνται με τους δείκτες στο πρώτο παράδειγμα).

Weights (Συντελεστές στάθμισης)

Η προεπιλογή είναι για όλες τις τιμές της απόκρισης να έχουν ίσα βάρη (όλοι ίσοι ενός)

```
weights = rep(1, nobs)
```

Υπάρχουν δύο είδη στάθμισης στη στατιστική μοντελοποίηση, και θα πρέπει να είστε σε θέση να τα διακρίνετε μεταξύ τους:

- περίπτωση στάθμης να δίνει τη σχετική σημασία της υπόθεσης, έτσι ώστε το βάρος των 2 σημαίνει ότι υπάρχουν δύο τέτοιες περιπτώσεις·
- αντίστροφο της διακύμανσης, στην οποία τα βάρη υποβαθμίζουν τα εξαιρετικά μεταβαλλόμενα δεδομένα.

Αντί να χρησιμοποιηθεί το αρχικό μέγεθος ρίζας ως συμμεταβλητή (όπως παραπάνω), θα μπορούσατε να χρησιμοποιήσετε Root ως στάθμιση στην προσαρμογή ενός μοντέλου Grazing ως η μόνη κατηγορηματική επεξηγηματική μεταβλητή:

```
model<-lm(Fruit ~ Grazing,weights=Root)
summary(model)
```

```
Call:
lm(formula = Fruit~Grazing, weights = Root)
Coefficients:
                Estimate  Std. Error  t value    Pr(>|t|)
(Intercept)      70.725       4.849    14.59      <2e-16 ***
GrazingUngrazed -16.953       7.469    -2.27       0.029 *

Residual standard error: 62.51 on 38 degrees of freedom
Multiple R-Squared: 0.1194, Adjusted R-squared: 0.0962
F-statistic: 5.151 on 1 and 38 DF, p-value: 0.02899
```

Περιττό να πούμε ότι, η χρήση των σταθμίσεων μεταβάλλει τις εκτιμήσεις των παραμέτρων και των πρότυπων σφαλμάτων τους:

```
model<-lm(Fruit ~ Grazing)
summary(model)
```

```
Coefficients:
               Estimate  Std. Error  t value    Pr(>|t|)
(Intercept)      67.941       5.236   12.976    1.54e-15 ***
GrazingUngrazed -17.060       7.404   -2.304      0.0268 *

Residual standard error: 23.41 on 38 degrees of freedom
Multiple R-Squared: 0.1226, Adjusted R-squared: 0.09949
F-statistic: 5.309 on 1 and 38 DF, p-value: 0.02678
```

Όταν οι συντελεστές στάθμισης (w) καθορίζονται, το μοντέλο προσαρμόζεται με χρήση σταθμισμένων ελαχίστων τετραγώνων, στα οποία η ποσότητα που πρέπει να ελαχιστοποιηθεί είναι $\sum w \times d^2$ (αντί $\sum d^2$), όπου d είναι η διαφορά μεταξύ της μεταβλητής απόκρισης και των προσαρμοσμένων τιμών που έχει προβλέψει το μοντέλο.

Ελλιπείς τιμές

Ένα σημαντικό ζήτημα είναι τι να κάνετε για τις τιμές που λείπουν στο πλαίσιο δεδομένων (σελ. 120). Στην ιδανική περίπτωση, βέβαια, δεν υπάρχουν τιμές που λείπουν, έτσι ώστε να μην χρειάζεται να ανησυχείτε για το τι μέτρα πρέπει να ληφθούν (na.action). Αν υπάρχουν ελλιπείς τιμές, έχετε δύο επιλογές:

- να αφήσετε έξω κάθε γραμμή του πλαισίου δεδομένων στην οποία μία ή περισσότερες μεταβλητές λείπουν, τότε na.action = na.omit
- να αποτύχει η διαδικασία προσαρμογής, έτσι ώστε na.action = na.fail

Σε περίπτωση αμφιβολίας, θα πρέπει να καθορίσετε na.action = na.fail για να μην έχετε δυσάρεστες εκπλήξεις όταν ανύποπτο NAs στο πλαίσιο δεδομένων προκαλέσει περίεργη (αλλά απροειδοποίητη) συμπεριφορά στο μοντέλο. Ας εισαγάγουμε μια τιμή που λείπει στην αρχική στάθμιση root:

```
Root[37]<-NA
model<-lm(Fruit ~ Grazing*Root)
```

Το μοντέλο προσαρμόζεται χωρίς σχόλια, και το μόνο πράγμα που μπορεί να παρατηρήσετε είναι ότι οι υπολειπόμενοι βαθμοί ελευθερίας μειώνονται από 36 σε 35. Αν θέλετε να ενημερώνεστε για τις τιμές που λείπουν, στη συνέχεια, χρησιμοποιήστε την επιλογή na.action.

```
model<-lm(Fruit ~ Grazing*Root,na.action=na.fail)
```

```
Error in na.fail.default(list(Fruit = c(59.77, 60.98, 14.73,
19.28, 34.25, : missing values in object
```

Αν διενεργείτε παλινδρόμηση με χρονοσειρές δεδομένων που περιλαμβάνουν τις τιμές που λείπουν, τότε θα πρέπει να χρησιμοποιήσετε na.action = NULL, έτσι ώστε τα υπόλοιπα και οι προσαρμοσμένες τιμές να είναι επίσης χρονοσειρές (αν οι τιμές που λείπουν είχαν παραλειφθεί, τότε το προκύπτον διάνυσμα δεν θα ήταν χρονοσειρά του σωστού μήκους).

Offsets (Μετατοπίσεις)

Δεν θα χρησιμοποιήσετε μετατοπίσεις με ένα γραμμικό μοντέλο (μπορείτε να αφαιρέσετε απλά την μετατόπιση από την τιμή της μεταβλητής απόκρισης, και να εργαστείτε με τις τροποποιημένες τιμές). Αλλά με γενικευμένα γραμμικά μοντέλα μπορεί να θέλετε να καθορίσετε μέρος της μεταβολής της απόκρισης χρησιμοποιώντας μετατόπιση (βλ. σελ. 518 για λεπτομέρειες και παραδείγματα).

Πλαίσια δεδομένων που περιέχουν τα ίδια ονόματα μεταβλητών

Αν έχετε πολλά διαφορετικά πλαίσια δεδομένων που περιέχουν τα ίδια ονόματα μεταβλητών (δηλαδή x και y), τότε ο απλούστερος τρόπος για να εξασφαλιστεί ότι οι σωστές μεταβλητές που χρησιμοποιούνται για την μοντελοποίηση είναι να ονομάσετε το πλαίσιο δεδομένων στην κλήση της συνάρτησης:

```
model<-lm(y ~ x,data=correct.frame)
```

Η εναλλακτική λύση είναι πολύ πιο δύσκολη να πληκτρολογηθεί:

```
model<-lm(correct.frame$y ~ correct.frame$x)
```

Κριτήριο πληροφοριών του Akaike

Το κριτήριο πληροφοριών του Akaike (AIC) είναι γνωστό στο στατιστικό εμπόριο ως **κύρωση λογαριθμικής πιθανότητας**. Εάν έχετε ένα μοντέλο για το οποίο μπορεί να ληφθεί μία τιμή λογαριθμικής πιθανότητας, τότε

$$AIC = -2 \times \log\text{-}likehood + 2(p+1),$$

όπου p είναι ο αριθμός των παραμέτρων του μοντέλου, και προστίθεται 1 για την εκτιμώμενη διακύμανση (μπορείτε, αν θέλατε, να το ονομάσετε μια άλλη παράμετρο). Για να απομυθοποιήσει το AIC ας το υπολογίσουμε με το χέρι. Θα επανεξετάσουμε τα δεδομένα παλινδρόμησης για τα οποία υπολογίζεται η λογαριθμική πιθανότητα με το χέρι στην σ. 217.

```
attach(regression)
names(regression)
```

```
[1] "growth" "tannin"
```

```
growth
```

```
[1] 12 10 8 11 6 7 2 3 3
```

Υπάρχουν εννέα τιμές της μεταβλητής απόκρισης, growth, και υπολογίσαμε νωρίτερα τη λογαριθμική πιθανότητα ως –23.98941. Υπήρχε μόνο μια παράμετρος εκτιμώμενη από τα δεδομένα για τους υπολογισμούς αυτούς (η μέση τιμή του y), έτσι $p = 1$. Αυτό σημαίνει ότι το AIC θα πρέπει να είναι

$$AIC = -2 \times -23.98941 + 2 \times (1+1) = 51.97882 .$$

Ευτυχώς, δεν χρειάζεται να πραγματοποιήσουμε τους υπολογισμούς αυτούς, επειδή υπάρχει μια ενσωματωμένη συνάρτηση για τον υπολογισμό του AIC. Η οποία παίρνει ένα μοντέλο αντικειμένου, σάν το όρισμα της, γι 'αυτό πρέπει να προσαρμόσουμε ένα μοντέλο μίας παραμέτρου με τα δεδομένα της growth, όπως αυτό:

```
model<-lm(growth~1)
```

Στη συνέχεια, μπορούμε να πάρουμε το AIC άμεσα:

```
AIC(model)
```

```
[1] 51.97882
```

Το AIC ως μέτρο της προσαρμογής του μοντέλου

Όσο περισσότερες παράμετροι υπάρχουν στο μοντέλο, τόσο καλύτερη είναι η προσαρμογή. Θα μπορούσατε να αποκτήσετε μια τέλεια προσαρμογή, αν είχατε μια ξεχωριστή παράμετρο για κάθε σημείο δεδομένων, αλλά αυτό το μοντέλο θα έχει απολύτως καμία επεξηγηματική δύναμη. Πάντα θα υπάρχει ένας συμβιβασμός μεταξύ της καλής προσαρμογής και του αριθμού των παραμέτρων που απαιτούνται από την οικονομία. Το AIC είναι χρήσιμο διότι κυρώνει ρητά κάθε περιττές παραμέτρους του μοντέλου, με την προσθήκη $2(p+1)$ στην αποκλίνουσα συμπεριφορά.

Κατά τη σύγκριση των δύο μοντέλων, όσο μικρότερο είναι το AIC, τόσο καλύτερη είναι η προσαρμογή. Αυτή είναι η βάση της αυτοματοποιημένης απλούστευσης μοντέλου χρησιμοποιώντας την step.

Μπορείτε να χρησιμοποιήσετε την συνάρτηση του AIC για να συγκρίνετε δύο μοντέλα, με τον ίδιο ακριβώς τρόπο, όπως μπορείτε να χρησιμοποιήσετε την anova (όπως εξηγείται στη σελ. 325). Εδώ έχουμε αναπτύξει μια ανάλυση της συνδιακύμανσης που εισάγεται στη σελίδα 490.

```
model.1<-lm(Fruit ~ Grazing*Root)
model.2<-lm(Fruit ~ Grazing+Root)
```

```
AIC(model.1, model.2)
```

```
         df         AIC
model.1   5    273.0135
model.2   4    271.1279
```

Επειδή το model.2 έχει το χαμηλότερο AIC, εμείς το προτιμούμε από το model.l. Η λογαριθμική πιθανότητα κυρώνεται με $2\times(4+1)=10$ στο μοντέλο 1, επειδή αυτό το μοντέλο περιείχε 4 παραμέτρους (2 κλίσεις και 2 τομές) και $2\times(3+1)=8$ στο model.2 γιατί αυτό το μοντέλο είχε 3 παραμέτρους (δύο τομές και μια κοινή κλίση). Μπορείτε να δείτε που οι δύο τιμές των AIC προέρχονται από υπολογισμό:

```
-2*logLik(model.1)+2*(4+1)
[1] 273.0135
-2*logLik(model.2)+2*(3+1)
[1] 271.1279
```

Leverage (Μόχλευση)

Τα σημεία αυξάνουν την επιρροή στο βαθμό που βρίσκονται μόνα τους, μακριά από τη μέση τιμή του x (είτε προς τα αριστερά ή προς τα δεξιά). Για να το εξηγήσουμε αυτό, τα μέτρα της μόχλευσης για ένα δεδομένο σημείο y δεδομένων είναι ανάλογα με $(x-\bar{x})^2$. Το συνηθέστερο μέτρο της μόχλευσης είναι

$$h_i = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum(x_j-\bar{x})^2}$$

όπου ο παρονομαστής είναι SSX. Ένας καλός εμπειρικός κανόνας είναι ότι ένα σημείο έχει ιδιαίτερα μεγάλη επιρροή στην περίπτωση που

$$h_i > \frac{2p}{n},$$

όπου p είναι ο αριθμός των παραμέτρων στο μοντέλο. Θα μπορούσαμε εύκολα να υπολογίσουμε την τιμή μόχλευσης κάθε σημείου του διανύσματος μας x_1 όπως δημιουργήθηκε στην σ. 345. Είναι πιο αποτελεσματικό, ίσως, να γράψουμε μια γενική συνάρτηση που θα μπορούσε να πραγματοποιήσει τον υπολογισμό των h τιμών για κάθε διάνυσμα x τιμών,

```
leverage<-function(x){1/length(x)+(x-mean(x))^2/sum((x-mean(x))^2)}
```

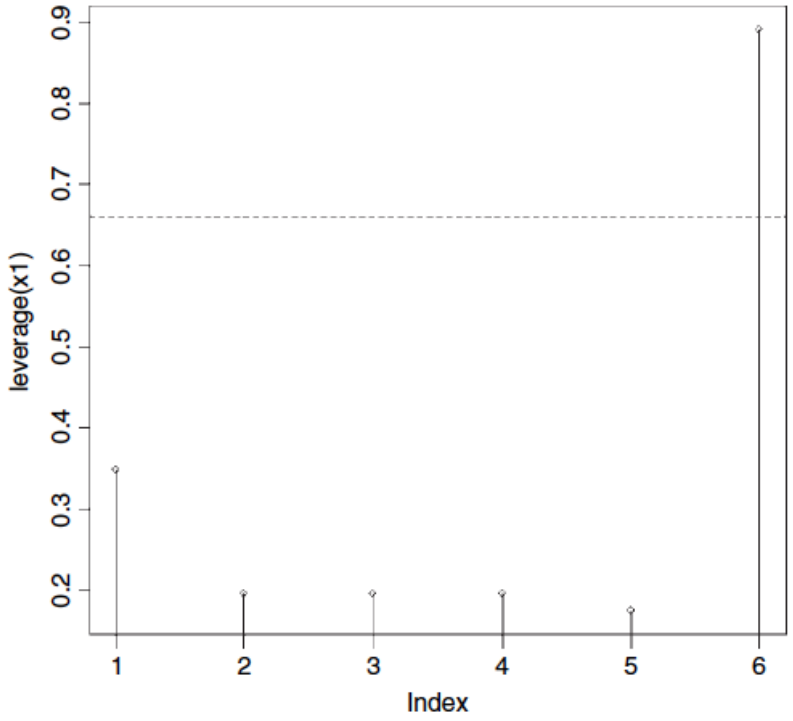
Στη συνέχεια, χρησιμοποιήστε τη συνάρτηση που ονομάζεται leverage στο x_1 :

```
[1] 0.3478261 0.1956522 0.1956522 0.1956522 0.1739130
[6] 0.8913043
```

Αυτή εφιστά την προσοχή αμέσως στην έκτη τιμή του x: Η h τιμή του είναι πάνω από το διπλάσιο, το επόμενο μεγαλύτερο. Το αποτέλεσμα είναι ακόμη σαφέστερο αν η σχεδιάσουμε τις τιμές μόχλευσης

```
plot(leverage(x1),type="h")
abline(0.66,0,lty=2)
points(leverage(x1))
```

Σημειώστε ότι αν η οδηγία σχεδίασης έχει ένα μόνο όρισμα (όπως εδώ), τότε οι x τιμές για το διάγραμμα λαμβάνονται ως η *σειρά* των αριθμών στο διάνυσμα που πρέπει να απεικονιστεί (ονομάζεται Δείκτης και λαμβάνοντας την ακολουθία 1:6 στην προκειμένη περίπτωση) . Θα ήταν χρήσιμο να σχεδιάσετε την τιμή του εμπειρικού κανόνα του τι αποτελεί ένα σημείο επιρροής. Στην περίπτωση αυτή, p = 2 και n (ο αριθμός των σημείων στο γράφημα) = 6, έτσι, ένα σημείο έχει επιρροή εάν $h_i > 0.66$.



Αυτό είναι αρκετό για να μας προειδοποιήσει ότι το σημείο (7,6) θα μπορούσε να έχει σημαντική επίδραση στις εκτιμήσεις των παραμέτρων του μοντέλου μας. Μπορούμε να δούμε αν αυτό είναι αληθές, επαναλαμβάνοντας την παλινδρόμηση χωρίς το σημείο (7,6). Υπάρχουν διάφοροι τρόποι για να γίνει αυτό. Αν γνωρίζουμε το δείκτη του σημείου, [6] σε αυτό το παράδειγμα, μπορούμε να αφήσουμε αυτό το σημείο ρητώς, χρησιμοποιώντας τη σύμβαση του αρνητικού δείκτη (βλ. σελ. 24.).

```
reg2<-lm(y1[-6] ~ x1[-6])
summary(reg2)
```

```
Residuals:
        1           2           3           4           5
1.955e-16   1.000e+00   4.572e-18  -1.000e+00  -9.890e-17
Coefficients:
             Estimate   Std. Error     t value     Pr(>|t|)
(Intercept) 2.000e+00    1.770e+00        1.13        0.341
x1[-6]     -2.587e-17    5.774e-01   -4.48e-17        1.000

Residual standard error: 0.8165 on 3 degrees of freedom
Multiple R-Squared: 2.465e-32, Adjusted R-squared: -0.3333
F-statistic: 7.396e-32 on 1 and 3 DF, p-value: 1
```

Το σημείο (7,6) είχε πράγματι πολύ μεγάλη επιρροή, διότι χωρίς αυτό, η κλίση της καμπύλης είναι μηδέν. Παρατηρήστε ότι τα υπόλοιπα από τα σημεία που δημιουργούνται δεν είναι ακριβώς μηδέν· είναι διάφοροι αριθμοί πολλαπλάσια του 10^{-17}.

Εναλλακτικά, θα μπορούσαμε να χρησιμοποιήσουμε τις στάθμες για να 'ζυγίσουμε' το σημείο (7,6), την επιρροή του οποίου θέλουμε να ελέγξουμε. Πρέπει να δημιουργήσουμε ένα διάνυσμα των συντελεστών στάθμισης: μονάδες για τα δεδομένα που θέλουμε να συμπεριλάβουμε και μηδενικά για τα δεδομένα που θέλουμε να αφήσουμε έξω. Στην απλή αυτή περίπτωση θα μπορούσαμε να πληκτρολογήσουμε τους συντελεστές στάθμισης άμεσα όπως εδώ:

```
w<-c(1,1,1,1,1,0)
```

αλλά σε γενικές γραμμές, θα θέλουμε να τα υπολογίσουμε με βάση κάποιο λογικό κριτήριο. Μια κατάλληλη προϋπόθεση για την ένταξη εδώ θα είναι $x_1 < 6$:

```
(w<-(x1<6))
[1] TRUE TRUE TRUE TRUE TRUE FALSE
```

Σημειώστε ότι όταν υπολογίζουμε το διάνυσμα βάρους με τον τρόπο αυτό, θα έχουμε TRUE και FALSE αντί 1 και 0, αλλά αυτό λειτουργεί εξίσου καλά. Το νέο μοντέλο μοιάζει με αυτό:

```
reg3<-lm(y1 ~ x1,weights=w)
summary(reg3)
```

Τέλος, θα μπορούσαμε να χρησιμοποιήσουμε την subset ώστε να αφήσουμε εκτός το σημείο ή τα σημεία που θα θέλαμε να εξαιρεθούν από την προσαρμογή του μοντέλου. Από όλες τις επιλογές, αυτή είναι η πιο γενική, και η ευκολότερη στη χρήση. Όπως με τους συντελεστές στάθμισης, το υποσύνολο δηλώνεται ως μέρος των προδιαγραφών του μοντέλου. Μας λέει ποια σημεία να συμπεριλάβει αντί να αποκλείσει, έτσι ώστε η λογική να συμπεριλάβει οποιαδήποτε σημεία για τα οποία $x_1 < 6$ (ας πούμε) :

```
reg4<-lm(y1 ~ x1,subset=(x1<6))
summary(reg4)
```

Οι έξοδοι του reg4 και του reg3 είναι ακριβώς οι ίδιες όπως στον reg2 χρησιμοποιώντας δείκτες.

Ατελώς Προσδιορισμένο Μοντέλο

Το μοντέλο μπορεί να έχει τους λάθος όρους σε αυτό, ή οι όροι μπορούν να συμπεριληφθούν στο μοντέλο με λάθος τρόπο. Ασχολούμαστε με την επιλογή των όρων για ένταξη στο ελάχιστο επαρκές μοντέλο στο κεφάλαιο 9. Εδώ απλώς σημειώνουμε ότι **ο μετασχηματισμός των επεξηγηματικών μεταβλητών** συχνά παράγει βελτιώσεις στην απόδοση του μοντέλου. Οι πιο συχνά χρησιμοποιούμενοι μετασχηματισμοί είναι οι λογαριθμικοί, των δυνάμεων και οι αντίστροφοι.

Όταν τόσο η κατανομή λάθους όσο και συναρτησιακή μορφή της σχέσης είναι άγνωστη, δεν υπάρχει ενιαίο συγκεκριμένο σκεπτικό για την επιλογή κάθε δεδομένου μετασχηματισμού σε προτίμηση ενός άλλου. Ο στόχος είναι ρεαλιστικός, δηλαδή να βρούμε ένα μετασχηματισμό που να δίνει:

• σταθερή διακύμανση λάθους·

• περίπου κανονικά σφάλματα·

• αθροιστικότητα·

• μια γραμμική σχέση μεταξύ των μεταβλητών απόκρισης και των επεξηγηματικών μεταβλητών·

• απλή επιστημονική ερμηνεία.

Η επιλογή είναι βέβαιο ότι θα είναι ένας συμβιβασμός και, ως εκ τούτου, είναι καλύτερο να επιλυθεί με ποσοτική σύγκριση της απόκλισης που παράγεται υπό διαφορετικές μορφές μοντέλου. Και πάλι, στις δοκιμές για τη μη-γραμμικότητα στη σχέση μεταξύ y και x μπορούμε να προσθέσουμε έναν όρο x^2 στο μοντέλο· μια σημαντική παράμετρος στον όρο x^2 υποδηλώνει καμπυλότητα στη σχέση μεταξύ y και x.

Ένα επιπλέον στοιχείο ατελούς προσδιορισμού μπορεί να προκύψει λόγω της **διαρθρωτικής μη γραμμικότητας**. Ας υποθέσουμε, για παράδειγμα, ότι προσαρμόζαμε ένα μοντέλο της μορφής

$$y = a + \frac{b}{x},$$

αλλά η υποκείμενη διαδικασία ήταν πραγματικά της μορφής

$$y = a + \frac{b}{c + x}$$

τότε η προσαρμογή θα είναι κακή. Φυσικά, αν *γνωρίζαμε* ότι η δομή του μοντέλου ήταν αυτής της μορφής, τότε θα μπορούσε να προσαρμοστεί ως ένα μη γραμμικό μοντέλο (σελ. 663) ή ως μη-γραμμικό μικτής επίδρασης μοντέλο (σελ. 671), αλλά στην πράξη αυτό συμβαίνει σπάνια.

Ελέγχοντας το Μοντέλο στην R

Τα δεδομένα που εξετάζουμε σε αυτή την ενότητα είναι για την αποσύνθεση των βιοδιασπώμενων πλαστικών στο έδαφος: η απόκριση, y , είναι η μάζα των πλαστικών που απομένει και η επεξηγηματική μεταβλητή, x , είναι η διάρκεια της ταφής:

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

```
[1] "time" "amount"
```

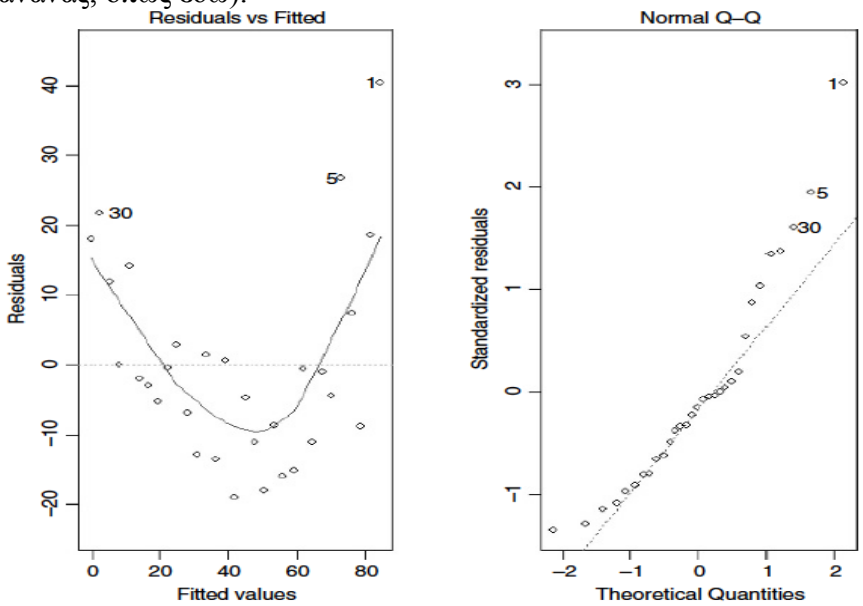
Για τους σκοπούς της απεικόνισης θα προσαρμόσουμε μια γραμμική παλινδρόμηση σε αυτά τα δεδομένα και στη συνέχεια θα χρησιμοποιήσουμε διαγράμματα ελέγχου μοντέλου για να διερευνηθεί η επάρκεια του εν λόγω μοντέλου:

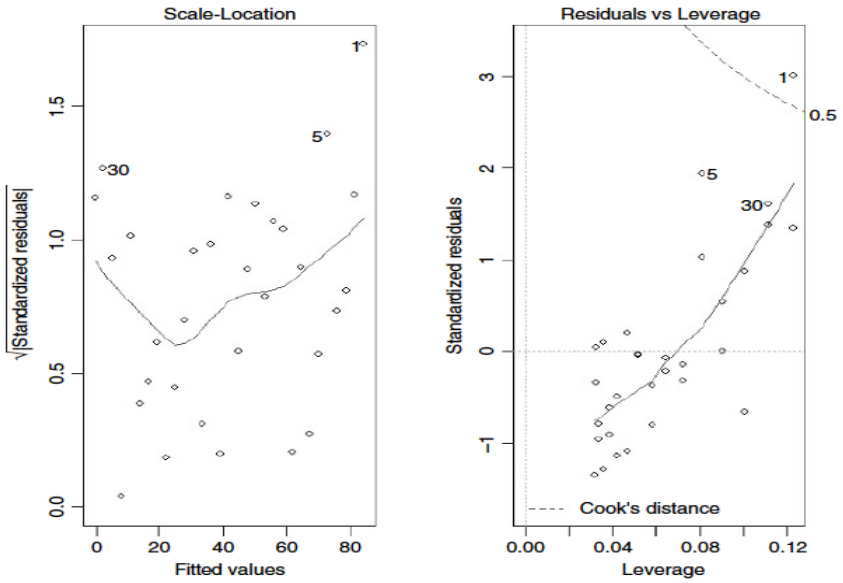
```
model<-lm(amount ~ time)
```

Το βασικό μοντέλο ελέγχου δεν θα μπορούσε είναι απλούστερο:

```
plot(model)
```

Αυτή η μία εντολή παράγει μια σειρά από γραφικές παραστάσεις, κατανεμημένες σε τέσσερις σελίδες. Τα πρώτα δύο γραφήματα είναι τα πιο σημαντικά. Κατ 'αρχάς, παίρνετε ένα διάγραμμα των υπολοίπων συναρτήσει των προσαρμοσμένων τιμών (αριστερό διάγραμμα), το οποίο παρουσιάζει πολύ έντονη καμπυλότητα· τα περισσότερα από τα υπόλοιπα για τις ενδιάμεσες προσαρμοσμένες τιμές είναι αρνητικά και τα θετικά υπόλοιπα συγκεντρώνονται στις μικρότερες και τις μεγαλύτερες προσαρμοσένες τιμές. Θυμηθείτε, αυτό το διάγραμμα πρέπει να μοιάζει με τον νυχτερινό ουρανό, χωρίς κανένα σχέδιο οποιουδήποτε είδους. Αυτό υποδηλώνει συστηματική ανεπάρκεια στη δομή του μοντέλου. Ίσως η σχέση μεταξύ y και x να είναι μη-γραμμική αντί γραμμικής, όπως υποθέτουμε εδώ· δεύτερον, μπορείτε να πάρετε ένα διάγραμμα QQ (σελ. 341), το οποίο δείχνει έντονη μη κανονικότητα των υπολοίπων (η γραμμή πρέπει να είναι ίσια, όχι σε σχήμα μπανάνας, όπως εδώ).





Το τρίτο γράφημα είναι σαν μία θετικά-εκτιμημένη έκδοση του πρώτου γραφήματος· Είναι καλό για την ανίχνευση μη σταθερότητας της διακύμανσης (ετεροσκεδαστικότητα), η οποία εμφανίζεται ως μια τριγωνική σκέδαση (όπως μια φέτα τυρί). Το τέταρτο γράφημα δείχνει έντονο μοτίβο στα τυποποιημένα υπόλοιπα ως συνάρτηση της μόχλευσης. Το γράφημα δείχνει επίσης την απόσταση του Cook, τονίζοντας την ταυτότητα των σημείων δεδομένων με ιδιαίτερη επιρροή.

Η απόσταση του Cook είναι μια προσπάθεια να συνδυάσει τη μόχλευση και τα υπόλοιπα σε ένα ενιαίο μέτρο. Οι απόλυτες τιμές των υπολοίπων διαγραφής $\left|r_i^*\right|$ σταθμίζονται ως εξής:

$$C_i = \left|r_i^*\right|\left(\frac{n-p}{p}.\frac{h_i}{1-h_i}\right)^{1/2}$$

Τα δεδομένα σημεία 1, 5 και 30 ξεχωρίζουν όντας επιρροής, ιδιαίτερα το σημείο με τον αριθμό 1. Άν θα ήμασταν πιο ευτυχισμένοι με άλλες πτυχές του μοντέλου, θα επαναλαμβάναμε την μοντελοποίηση, αφήνοντας εκτός, κάθε ένα από τα σημεία αυτά με τη σειρά τους. Εναλλακτικά, θα μπορούσαμε να αξιολογήσουμε τα δεδομένα με τη μεθοδο jackknife (βλέπε σελ. 422), η οποία περιλαμβάνει να αφήνουμε κάθε σημείο δεδομένων έξω, ένα κάθε φορά, με τη σειρά. Σε κάθε περίπτωση, αυτό είναι σαφώς ότι *δεν* είναι ένα καλό μοντέλο για τα δεδομένα αυτά. Η ανάλυση ολοκληρώνεται στην σελ. 407.

Εξαγωγή πληροφοριών από τα αντικείμενα μοντέλου

Μπορείτε συχνά να θέλετε να εξαγάγετε υλικό από προσαρμοσμένα μοντέλα (π.χ. κλίσεις, υπόλοιπα ή p τιμές) και υπάρχουν τρεις διαφορετικοί τρόποι για να γίνει αυτό:

• με βάση το όνομα, π.χ. coef(model)

• με λίστα από δείκτες, π.χ. summary(model)[[3]]

• χρησιμοποιώντας $ για να αναφέρουμε το στοιχείο, π.χ. model$resid

Το μοντέλο αντικειμένου που χρησιμοποιούμε για να επιδειχθούν αυτές οι τεχνικές είναι η απλή γραμμική παλινδρόμηση που αναλύθηκε πλήρως με το χέρι στην σελ. 388.

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

```
[1] "growth" "tannin"
```

```
model<-lm(growth ~ tannin)
summary(model)
```

```
Call:
lm(formula = growth~tannin)


Residuals:
     Min       1Q   Median       3Q      Max
 -2.4556  -0.8889  -0.2389   0.9778   2.8944
Coefficients:
             Estimate  Std.Error   t value    Pr(>|t|)
(Intercept)   11.7556     1.0408    11.295    9.54e-06 ***
tannin        -1.2167     0.2186    -5.565    0.000846 ***


Residual standard error: 1.693 on 7 degrees of freedom
Multiple R-Squared: 0.8157, Adjusted R-squared: 0.7893
F-statistic: 30.97 on 1 and 7 DF, p-value: 0.000846
```

Κατ 'όνομα

Μπορείτε να εξαγάγετε τους συντελεστές του μοντέλου, των προσαρμοσμένων τιμών, τα υπόλοιπα, τα μεγέθη επίδρασης και τους πίνακες διακύμανσης-συνδιακύμανσης με βάση το όνομα, ως εξής:

```
coef(model)
```

```
(Intercept)      tannin
  11.755556   -1.216667
```

δίνει τις εκτιμήσεις παραμέτρων (‘συντελεστές’) για το σημείο τομής (a) και κλίσης (b) ·

```
fitted(model)
```

```
        1         2         3         4         5         6
11.755556 10.538889  9.322222  8.105556  6.888889  5.672222
        7         8         9
 4.455556  3.238889  2.022222
```

δίνει τις εννέα προσαρμοσμένες τιμές ($\hat{y} = a + bx$) που χρησιμοποιούνται για τον υπολογισμό των υπολοίπων·

resid(model)

```
        1          2          3         4          5
0.2444444 -0.5388889 -1.3222222 2.8944444 -0.8888889
        6          7          8         9
1.3277778 -2.4555556 -0.2388889 0.9777778
```

δίνει τα υπόλοιπα (y - προσαρμοσμένες τιμές) για τα εννέα σημεία δεδομένων·

effects(model)

```
(Intercept)     tannin
-20.6666667 -9.4242595 -1.3217694 2.8333333 -1.0115 1.143538
 -2.7013585 -0.5462557  0.6088470

attr(,"assign")
[1] 0 1
attr(,"class")
[1] "coef"
```

Για ένα γραμμικό μοντέλο προσαρμοσμένο από την lm ή την aov, οι effects είναι οι ασύνδετες τιμές με μοναδιαίους βαθμούς ελευθερίας που λαμβάνονται με την προβολή των δεδομένων επί των διαδοχικών ορθογωνίων υποδιαστημάτων, τα οποία δημιουργούνται από την διάσπαση QR κατά την διαδικασία προσαρμογής. Το πρώτο r (ίσον με 2 στην περίπτωση αυτή· είναι η κατάταξη του μοντέλου) που σχετίζεται με τους συντελεστές και με το χρονικό υπόλοιπο του διάστηματος των υπολοίπων, αλλά δεν συνδέεται με συγκεκριμένα υπόλοιπα. Παράγει ένα αριθμητικό διάνυσμα ίδιου μήκους με των υπολοίπων της κατηγορίας coef. Στις δύο πρώτες σειρές επισημαίνονται οι αντίστοιχοι συντελεστές (τομής και κλίσης), και οι υπόλοιπες 7 σειρές δεν επισημαίνονται.

vcov(model)

```
             (Intercept)       tannin
(Intercept)    1.0832628  -0.19116402
tannin        -0.1911640   0.04779101
```

Αυτό εξάγει τον πίνακα διακύμανσης-συνδιακύμανσης των παραμέτρων του μοντέλου.

Με λίστα δεικτών

Το αντικείμενο μοντέλου είναι μια λίστα με πολλές συνιστώσες. Εδώ η κάθε μία από αυτές εξηγείται με τη σειρά. Η πρώτη είναι ο τύπος μοντέλου (ή 'Call') που δείχνει την μεταβλητή απόκρισης (growth) και την επεξηγηματική μεταβλητή (ες) (tannin):

summary(model)[[1]]

```
lm(formula = growth~tannin)
```

Η δεύτερη περιγράφει τα χαρακτηριστικά του αντικειμένου που ονομάζεται summary(model):

summary(model)[[2]]

```
growth ~ tannin
attr(,"variables")
list(growth, tannin)
attr(,"factors")
        tannin
growth       0
tannin       1
attr(,"term.labels")
[1] "tannin"
attr(,"order")
[1] 1 attr(,"intercept")
[1] 1
attr(,"response")
[1] 1
attr(,".Environment")
<environment: R_GlobalEnv>
attr(,"predvars")
list(growth, tannin)
attr(,"dataClasses")
  growth    tannin
"numeric" "numeric"
```

Η τρίτη δίνει τα υπόλοιπα για τα εννέα σημεία δεδομένων:

summary(model)[[3]]

```
         1          2          3          4          5
0.2444444 -0.5388889 -1.3222222  2.8944444 -0.8888889
         6          7          8          9
1.3277778 -2.4555556 -0.2388889  0.9777778
```

Η τέταρτη δίνει τον πίνακα παραμέτρων, συμπεριλαμβανομένων των τυπικών σφαλμάτων των παραμέτρων, t τιμών και p αξιών:

summary(model)[[4]]

```
            Estimate  Std. Error    t value          Pr(>|t|)
(Intercept)11.755556   1.0407991  11.294740      9.537315e-06
tannin     -1.216667   0.2186115  -5.565427      8.460738e-04
```

Η πέμπτη ασχολείται με το αν οι αντίστοιχες συνιστώσες της προσαρμογής (το πλαίσιο μοντέλου, ο πίνακας μοντέλου, η απόκριση ή η διάσπαση QR) θα πρέπει να επιστραφούν. Η προεπιλογή είναι FALSE:

summary(model)[[5]]

```
(Intercept)    tannin
      FALSE     FALSE
```

Η έκτη είναι η υπολειμματική απόκλιση σφάλματος: η τετραγωνική ρίζα της διακύμανσης σφάλματος από τον πίνακα summary.aov (το οποίο δεν εμφανίζεται εδώ: $s^2 = 2.867 \cdot$βλ. σελ. 396):

summary(model)[[6]]

```
[1] 1.693358
```

Η έβδομη δείχνει τον αριθμό των γραμμών στον πίνακα summary.lm (που δείχνει δύο παραμέτρους που έχουν εκτιμηθεί από τα δεδομένα με αυτό το μοντέλο, και τους υπολειμματικούς βαθμούς ελευθερίας (df = 7):

summary(model)[[7]]

```
[1] 2 7 2
```

Η όγδοη είναι $r^2 = SSR / SST$, το κλάσμα της συνολικής διακύμανσης στη μεταβλητή απόκρισης που εξηγείται από το μοντέλο (βλέπε σελ. 399 για λεπτομέρειες):

```
summary(model)[[8]]
[1] 0.8156633
```

Η ένατη είναι το προσαρμοσμένο R^2 , εξηγείται στη σελ. 399, αλλά σπάνια χρησιμοποιείται στην πράξη:

```
summary(model)[[9]]
[1] 0.7893294
```

Η δέκατη δίνει τον λόγο F των πληροφοριών: οι τρεις τιμές που δίνονται εδώ είναι ο λόγος F (30.97398), ο αριθμός των βαθμών ελευθερίας του μοντέλου (δηλαδή στον αριθμητή: numdf) και οι υπολειμματικοί βαθμοί ελευθερίας (δηλαδή στον παρονομαστή: dendf):

```
summary(model)[[10]]
   value    numdf    dendf
30.97398 1.00000 7.00000
```

Η ενδέκατη συνιστώσα είναι ο πίνακας συσχέτισης των εκτιμήσεων των παραμέτρων:

```
summary(model)[[11]]
              (Intercept)      tannin
(Intercept)   0.37777778   -0.06666667
tannin       -0.06666667    0.01666667
```

Συχνά θα θέλετε να εξάγετε στοιχεία από τον πίνακα παραμέτρων που ήταν το τέταρτο αντικείμενο παραπάνω. Το πρώτο από αυτά είναι το σημείο τομής (a , η τιμή της growth στην tannin= 0):

```
summary(model)[[4]][[1]]
[1] 11.75556
```

Το δεύτερο είναι η κλίση (b , η μεταβολή στην growth ανά μονάδα μεταβολής στην tannin):

```
summary(model)[[4]][[2]]
[1] -1.216667
```

Το τρίτο είναι το πρότυπο σφάλμα της τομής, se_a :

```
summary(model)[[4]][[3]]
[1] 1.040799
```

Το τέταρτο είναι το πρότυπο σφάλμα της κλίσης, se_b:

```
summary(model)[[4]][[4]]
[1] 0.2186115
```

Το πέμπτο είναι η τιμή του t του Student για το σημείο τομής $= a / se_a$:

```
summary(model)[[4]][[5]]
[1] 11.29474
```

Το έκτο είναι η τιμή του t του Student για την κλίση $= b / se_b$:

```
summary(model)[[4]][[6]]
[1] -5.565427
```

Το έβδομο είναι η τιμή p για το σημείο τομής: η πιθανότητα παρατήρησης μίας t τιμής αυτής μεγάλης ή μεγαλύτερης, εάν η μηδενική υπόθεση (H_0: τομή = 0) είναι αληθής:

```
summary(model)[[4]][[7]]
[1] 9.537315e-06
```

Απορρίπτουμε την H_0 επειδή $p < 0.05$. Το όγδοο είναι η τιμή p για την κλίση: η πιθανότητα παρατήρησης μίας t τιμής αυτής μεγάλης ή μεγαλύτερης, εάν η μηδενική υπόθεση (H_0: τομή = 0) είναι αληθής:

```
summary(model)[[4]][[8]]
[1] 0.0008460738
```

Απορρίπτουμε την H_0 επειδή $p < 0.05$. Για να αποθηκεύσετε τα δύο πρότυπα σφάλματα (1.0407991 και 0.2186115) γράψτε

```
sea<-summary(model)[[4]][[3]]
seb<-summary(model)[[4]][[4]]
```

Εξαγωγή συνιστωσών του μοντέλου χρησιμοποιώντας $

Ένας άλλος τρόπος για την εξαγωγή συνιστωσών του μοντέλου είναι να χρησιμοποιήσετε το σύμβολο $. Για να πάρετε το σημείο τομής (a) και την κλίση (b) της παλινδρόμησης, πληκτρολογήστε

```
model$coef
```

```
(Intercept)      tannin
 11.755556   -1.216667
```

Για να πάρετε τις προσαρμοσμένες τιμές ($\hat{y} = a + bx$) που χρησιμοποιούνται για τον υπολογισμό των υπολοίπων, πληκτρολογήστε

```
model$fitted
```

```
        1         2         3         4         5         6
11.755556 10.538889  9.322222  8.105556  6.888889  5.672222
        7         8         9
 4.455556  3.238889  2.022222
```

Για να να πάρετε τα ίδια τα υπόλοιπα, πληκτρολογήστε

```
model$resid
```

```
        1          2          3          4          5
0.2444444 -0.5388889 -1.3222222  2.8944444 -0.8888889
        6          7          8          9
1.3277778 -2.4555556 -0.2388889  0.9777778
```

Τέλος, οι υπολειμματικοί βαθμοί ελευθερίας (9 σημεία - 2 εκτιμώμενες παράμετροι = 7 d.f) είναι

```
model$df
```

```
[1] 7
```

Εξαγωγή συνιστωσών από τον πίνακα summary.aov

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

```
            Df    Sum Sq   Mean Sq    F value      Pr(>F)
tannin       1    88.817    88.817     30.974    0.000846 ***
Residuals    7    20.072     2.867
```

Μπορείτε να πάρετε τους βαθμούς ελευθερίας, τα αθροίσματα των τετραγώνων, την μέση τιμή των τετραγώνων, τον λόγο F και τις p τιμές από τον πίνακα ANOVA για ένα μοντέλο σαν αυτό:

```
summary.aov(model)[[1]][[1]]
[1] 1 7
summary.aov(model)[[1]][[2]]
[1] 88.81667 20.07222
summary.aov(model)[[1]][[3]]
[1] 88.816667 2.867460
summary.aov(model)[[1]][[4]]
[1] 30.97398 NA
summary.aov(model)[[1]][[5]]
[1] 0.0008460738 NA
```

Θα πρέπει να πειραματιστείτε για να δείτε τις συνιστώσες του αντικειμένου του ίδιου του μοντέλου (π.χ. model[[3]]).

Ο πίνακας summary.lm για συνεχείς και κατηγορηματικές επεξηγηματικές μεταβλητές

Είναι σημαντικό να κατανοήσετε τη διαφορά μεταξύ summary.lm και summary.aov για το ίδιο μοντέλο. Εδώ είναι μια μονόδρομη ανάλυση διακύμανσης του πειράματος ανταγωνισμού φύτευσης (σελ. 155):

```
comp<-read.table("c:\\temp\\competition.txt",header=T)
attach(comp)
names(comp)
[1] "biomass" "clipping"
```

Η κατηγορηματική επεξηγηματική μεταβλητή είναι η clipping και έχει πέντε επίπεδα ως εξής:

```
levels(clipping)
[1] "control" "n25" "n50" "r10" "r5"
```

Η ανάλυση του μοντέλου διακύμανσης έχει προσαρμοστεί σαν αυτό:

```
model<-lm(biomass ~ clipping)
```

και οι δύο διαφορετικές περιλήψεις είναι:

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

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

summary.lm(model)

```
Coefficients:
             Estimate   Std. Error   t value  Pr(>|t|)
(Intercept)    465.17        28.75    16.177   9.4e-15 ***
clippingn25     88.17        40.66     2.168   0.03987   *
clippingn50    104.17        40.66     2.562   0.01683   *
clippingr10    145.50        40.66     3.578   0.00145  **
clippingr5     145.33        40.66     3.574   0.00147  **

Residual standard error: 70.43 on 25 degrees of freedom
Multiple R-Squared: 0.4077, Adjusted R-squared: 0.3129
F-statistic: 4.302 on 4 and 25 DF, p-value: 0.008752
```

Η τελευταία, summary.lm, είναι πολύ πιο κατατοπιστική. Δείχνει τα μεγέθη επίδρασης και τα πρότυπα σφάλματα τους. Τα μεγέθη επίδρασης εμφανίζονται υπό τη μορφή των αντιθέσεων (όπως εξηγείται λεπτομερώς στη σελ. 370). Τα μόνα ενδιαφέροντα πράγματα στην summary.aov είναι η διακύμανση σφάλματος ($s^2 = 4961$) και ο λόγος F (4,3015) που δείχνουν ότι υπάρχουν σημαντικές διαφορές να εξηγηθούν. Η πρώτη γραμμή του πίνακα summary.lm περιέχει μία μέση τιμή και όλες οι άλλες σειρές παρουσιάζουν διαφορές μεταξύ των μέσων τιμών. Έτσι, το πρότυπο σφάλμα στην πρώτη γραμμή (με την ένδειξη (`intercept`)) είναι το τυπικό σφάλμα της μέσης $se_{mean} = \sqrt{s^2/(k \times n)}$, ενώ τα πρότυπα σφάλματα για τις άλλες σειρές είναι τα πρότυπα σφάλματα της διαφοράς μεταξύ των δύο μέσων τιμών $se_{diff} = \sqrt{2 \times s^2/n}$, όπου υπάρχουν επίπεδα συντελεστών k το καθένα με αντιγραφή ίσον με n .

Έτσι από πού προέρχονται τα μεγέθη επίδρασης; Τι είναι το 465.17 και τι το 88.17; Για να κατανοήσουμε τις απαντήσεις στα ερωτήματα αυτά, πρέπει να γνωρίζουμε πώς η εξίσωση για τις επεξηγηματικές μεταβλητές είναι δομημένη σε ένα γραμμικό μοντέλο, όταν η ερμηνευτική μεταβλητή, όπως εν προκειμένω, είναι κατηγορηματική. Για να ανακεφαλαιώσουμε, το γραμμικό μοντέλο παλινδρόμησης γράφεται ως

lm(y ~ x)

όπου η R ερμηνεύει ως διπαραμετρική γραμμική εξίσωση. Η R το γνωρίζει αυτό, επειδή το x είναι μια συνεχής μεταβλητή,

$$y = a + bx$$

στην οποία οι τιμές των παραμέτρων a και b θα πρέπει να εκτιμηθούν από τα δεδομένα. Αλλά τι γίνεται με την ανάλυσή της διακύμανσης; Έχουμε μία επεξηγηματική

μεταβλητή, x = clipping, αλλά είναι μια κατηγορική μεταβλητή με πέντε επίπεδα, control, n25, n50, r10 και r5. Το μοντέλο aov είναι ακριβώς ανάλογο με το μοντέλο παλινδρόμησης

```
aov(y ~ x)
```

αλλά ποια είναι η σχετική εξίσωση; Ας δούμε την εξίσωση πρώτα, και στη συνέχεια, θα προσπαθήσουμε να το καταλάβουμε:

$$y = a + bx_1 + cx_2 + dx_3 + ex_4 + fx_5$$

Αυτό μοιάζει με μια πολλαπλή παλινδρόμηση, με πέντε επεξηγηματικές μεταβλητές $x_1,...,x_5$. Το σημείο-κλειδί για να κατανοήσουμε είναι ότι $x_1,...,x_5$ είναι τα επίπεδα του παράγοντα που ονομάζεται x. Το σημείο τομής, a, είναι η συνολική (ή μέγιστη) μέση τιμή για το σύνολο του πειράματος. Οι παράμετροι $b,...,f$ είναι οι διαφορές μεταξύ της μέγιστης μέσης τιμής και της μέσης τιμής για ένα δεδομένο επίπεδο παράγοντα. Θα πρέπει να επικεντρωθούμε για να το καταλάβουμε αυτό.

Με μια κατηγορηματική επεξηγηματική μεταβλητή, όλες οι μεταβλητές κωδικοποιούνται ως $x = 0$ εκτός από το επίπεδο του παράγοντα που σχετίζεται με την εν λόγω τιμή y, όταν το x είναι κωδικοποιημένο ως $x = 1$. Θα το βρείτε δύσκολο να το καταλάβετε αυτό χωρίς μια καλή πρακτική. Ας δούμε την πρώτη γραμμή των δεδομένων στο πλαίσιο μας:

```
comp[1,]
```

```
     biomass clipping
1        551      n25
```

Έτσι, η πρώτη τιμή biomass (551) στο πλαίσιο δεδομένων προέρχεται από την επεξεργασία της n25 clipping η οποία, από όλα τα επίπεδα του παράγοντα (παραπάνω), καταλαμβάνει τη δεύτερη θέση στο αλφάβητο. Αυτό σημαίνει ότι για αυτήν την σειρά του πλαισίου δεδομένων $x_1 = 0$, $x_2 = 1$, $x_3 = 0$, $x_4 = 0$, $x_5 = 0$. Η εξίσωση ως εκ τούτου για την πρώτη σειρά, φαίνεται σαν αυτή:

$$y = a + b \times 0 + c \times 1 + d \times 0 + e \times 0 + f \times 0,$$

έτσι ώστε το μοντέλο για την προσαρμοσμένη τιμή στην n25 είναι

$$\hat{y} = a + c \cdot$$

και ομοίως για τα άλλα επίπεδα παράγοντα. Η προσαρμοσμένη τιμή $\hat{y}$ είναι το άθροισμα των δύο παραμέτρων, a και c . Η εξίσωση προφανώς δεν περιέχει μία επεξηγηματική μεταβλητή (δεν υπάρχει x στην εξίσωση, όπως θα υπήρχε σε μια εξίσωση παλινδρόμησης, παραπάνω). Σημειώστε, επίσης, πως το πλήρες μοντέλο περιλαμβάνει πολλές παραμέτρους: αντιπροσωπεύεται από τα γράμματα a έως f και υπάρχουν έξι από αυτές. Αλλά μπορούμε να εκτιμήσουμε μόνο πέντε παραμέτρους σε αυτό το πείραμα (μία μέση τιμή για κάθε ένα από τα πέντε επίπεδα παράγοντα). Το μοντέλο μας περιέχει μια περιττή παράμετρο, και πρέπει να το αντιμετωπίσουμε αυτό. Υπάρχουν διάφοροι

λογικοί τρόποι για να το κάνουμε αυτό, και οι άνθρωποι διαφέρουν σχετικά με το ποιός είναι ο καλύτερος τρόπος. Οι συγγραφείς της R συμφωνούν ότι οι *αντιθέσεις επεξεργασίας* αποτελούν την καλύτερη λύση. Η μέθοδος αυτή έχει να κάνει με μια παράμετρο a , τη συνολική μέση τιμή. Η μέση τιμή του επιπέδου παράγοντα που έρχεται πρώτη (control, στο παράδειγμά μας) προωθείται σε πλεονεκτική θέση, και οι άλλες επιδράσεις εμφανίζονται ως διαφορές (αντιθέσεις) μεταξύ αυτής της μέσης τιμής και των άλλων τεσσάρων των επιπέδων παράγοντα.

Ένα παράδειγμα θα μπορούσε να βοηθήσει να γίνει σαφέστερο. Εδώ είναι πέντε μέσες τιμές:

```
means<-tapply(biomass,clipping,mean)
means
```

```
 control       n25       n50       r10       r5
465.1667  553.3333  569.3333  610.6667  610.5000
```

Η ιδέα είναι ότι η μέση τιμή control (465.1667) γίνεται η πρώτη παράμετρος του μοντέλου (γνωστό ως το σημείο τομής). Η δεύτερη παράμετρος είναι η διαφορά μεταξύ της δεύτερης μέσης τιμής (n25 = 553.333) και του σημείου τομής:

```
means[2]-means[1]
```

```
     n25
88.16667
```

Η τρίτη παράμετρος είναι η διαφορά μεταξύ της τρίτης μέσης τιμής (n50 = 569.333) και του σημείου τομής:

```
means[3]-means[1]
```

```
     n50
104.1667
```

Η τέταρτη παράμετρος είναι η διαφορά μεταξύ της τέταρτης μέσης τιμής (r10 = 610.6667) και του σημείου τομής:

```
means[4]-means[1]
```

```
  r10
145.5
```

Η πέμπτη παράμετρος είναι η διαφορά μεταξύ της πέμπτης μέσης τιμής (r5 = 610.5) και του σημείου τομής:

```
means[5]-means[1]
```

```
      r5
145.3333
```

Ενώ ασχολούμαστε τόσο πολύ για τα μεγέθη επίδρασης. Τι γίνεται με τα πρότυπα σφάλματα τους; Η πρώτη σειρά είναι μια μέση τιμή, γι 'αυτό χρειαζόμαστε το πρότυπο σφάλμα μίας μέσης τιμής ενός μονοεπίπεδου παράγοντα. Αυτή η μέση τιμή βασίζεται σε έξι αριθμούς σε αυτό το παράδειγμα, έτσι ώστε το πρότυπο σφάλμα της είναι $\sqrt{s^2/n}$, όπου η διακύμανση του σφάλματος s^2 = 4961 λαμβάνεται από την summary.aov (model) όπως παραπάνω:

```
sqrt(4961/6)
```

```
[1] 28.75471
```

Όλες οι άλλες γραμμές έχουν το ίδιο πρότυπο σφάλμα, αλλά είναι μεγαλύτερο από αυτό. Αυτό συμβαίνει επειδή οι επιδράσεις στην 2η και στις επόμενες σειρές *δεν* είναι οι μέσες τιμές, αλλά *οι διαφορές μεταξύ αυτών*. Τούτο σημαίνει ότι το κατάλληλο πρότυπο σφάλμα δεν είναι το τυπικό σφάλμα μιας μέσης τιμής, αλλά μάλλον το τυπικό σφάλμα της διαφοράς μεταξύ των δύο μέσων. Όταν δύο δείγματα είναι ανεξάρτητα, η διακύμανση της διαφοράς τους είναι το άθροισμα των δύο διακυμάνσεων τους. Έτσι, ο τύπος για το τυπικό σφάλμα της διαφοράς μεταξύ των δύο μέσων είναι

$$se_{diff} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} .$$

Όταν οι δύο διακυμάνσεις και τα δύο μεγέθη των δειγμάτων είναι ίδια (όπως είναι εδώ, επειδή το σχέδιό μας είναι ισορροπημένο και χρησιμοποιούμε την συγκεντρωτική διακύμανση σφαλμάτων (4961) από τον πίνακα summary.aov), ο τύπος απλοποιείται σε $\sqrt{2 \times s^2/n}$:

```
sqrt(2*4961/6)
```

```
[1] 40.6653
```

Με λίγη πρακτική, θα πρέπει να απομυθοποιείται η προέλευση των αριθμών στον πίνακα summary.lm. Αλλά χρειάζεται πολύ δουλειά, και οι άνθρωποι το βρίσκουν πολύ δύσκολο αυτό στην αρχή, οπότε μην αισθάνεστε άσχημα γι 'αυτό.

Αντιθέσεις

Οι αντιθέσεις είναι η ουσία του ελέγχου υποθέσεων και της απλοποίησης του μοντέλου στην ανάλυση της διακύμανσης και στην ανάλυση της συνδιακύμανσης. Χρησιμοποιούνται για να συγκρίνουν μέση τιμή ή ομάδες από αυτές με άλλες αντίστοιχα ή ομάδες αυτών, σε ό, τι είναι γνωστό ως συγκρίσεις μοναδιαίου βαθμού ελευθερίας. Υπάρχουν δύο είδη αντιθέσεων που θα μπορούσανε να μας ενδιαφέρουν:

- αντιθέσεις που είχαμε προγραμματίσει να εξετάσουμε σε πειραματικό στάδιο του σχεδιασμού (αυτές αναφέρονται ως αντιθέσεις ***εκ των προτέρων-a priori***)·

- αντιθέσεις που φαίνονται ενδιαφέρουσες αφού έχουμε δει τα αποτελέσματα (αυτές αναφέρονται ως αντιθέσεις ***εκ των υστέρων-a posteriori***).

Μερικοί άνθρωποι είναι πολύ φαντασμένοι σχετικά με τις *a posteriori* αντιθέσεις, με το σκεπτικό ότι ήταν απρογραμμάτιστες. Υποτίθεται ότι δεν πρέπει να αποφασίσετε τι συγκρίσεις να κάνετε *αφού* έχετε δει την ανάλυση, αλλά οι επιστήμονες αυτό κάνουν όλη την ώρα. Το βασικό σημείο είναι ότι θα πρέπει να κάνετε αντιθέσεις μόνο *αφού* η ANOVA απέδειξε ότι πραγματικά υπάρχουν σημαντικές διαφορές που πρέπει να διερευνηθούν. Δεν είναι καλή πρακτική να πραγματοποιούνται δοκιμές για να συγκρίνετε τη μεγαλύτερη μέση τιμή με τη μικρότερη , αν η ANOVA απέτυχε να απορρίψει τη μηδενική υπόθεση (όσο δελεαστικό αν και αυτό μπορεί να είναι).

Υπάρχουν δύο σημαντικά σημεία για να καταλάβετε σχετικά με τις αντιθέσεις :

- υπάρχει ένας τεράστιος αριθμός *πιθανών* αντιθέσεων, και

- υπάρχουν μόνο k - 1 *ορθογώνιες* αντιθέσεις ,

όπου k είναι ο αριθμός των επιπέδων παράγοντα. Δύο αντιθέσεις λέγεται ότι είναι **ορθογώνιες** μεταξύ τους, εάν οι συγκρίσεις είναι στατιστικά ανεξάρτητες. Τεχνικά, δύο αντιθέσεις είναι ορθογώνιες αν το γινόμενο του αθροίσματος των συντελεστών της αντίθεσης τους είναι μηδέν (θα δούμε τι σημαίνει αυτό σε λίγο).

Ας πάρουμε ένα απλό παράδειγμα. Ας υποθέσουμε ότι έχουμε έναν παράγοντα με πέντε επίπεδα και τα επίπεδα του παράγοντα ονομάζονται a , b , c , d , και e . Ας αρχίσουμε γράφοντας τις πιθανές αντιθέσεις. Προφανώς θα μπορούσαμε να συγκρίνουμε κάθε μέση τιμή χωριστά με κάθε άλλη:

a vs. b , a vs. c , a vs. d , a vs e , b vs. c , b vs. d , b vs. e , c vs. d , c vs. e , e vs. d

Αλλά θα μπορούσαμε επίσης να συγκρίνουμε τα ζεύγη των μέσων:

$\{a,b\}$ vs. $\{c,d\}$, $\{a,b\}$ vs. $\{c,e\}$, $\{a,b\}$ vs. $\{d,e\}$, $\{a,c\}$ vs. $\{b,d\}$, $\{a,c\}$ vs. $\{b,e\}$,...

ή σε τριάδες :

$\{a,b,c\}$ vs. d , $\{a,b,c\}$ vs. e , $\{a,b,d\}$ vs. c , $\{a,b,d\}$ vs. e , $\{a,c,d\}$ vs. b ,…

ή ομάδες των τεσσάρων μέσων:

$\{a,b,c,d\}$ vs. e , $\{a,b,c,e\}$ vs. d , $\{a,b,d,e\}$ vs. c , $\{a,c,d,e\}$ vs. b , $\{b,c,d,e\}$ vs. a

Μπορείτε να πάρετε αναμφίβολα την ιδέα. Υπάρχουν απόλυτα μάζες πιθανών αντιθέσεων. Στην πράξη, όμως, θα πρέπει να συγκρίνουμε τα πράγματα μόνο μία φορά, είτε άμεσα είτε έμμεσα. Έτσι, οι δύο αντιθέσεις a vs. b και a vs. c εμμέσως έρχονται σε αντίθεση με b vs. c . Αυτό σημαίνει ότι αν έχουμε πραγματοποιήσει τις δύο αντιθέσεις a vs. b και a vs. c τότε η τρίτη αντίθεση b vs. c *δεν* είναι μια ορθογώνια αντίθεση, γιατί έχει πραγματοποιηθεί ήδη , έμμεσα. Ποιες συγκεκριμένες αντιθέσεις είναι ορθογώνιες

εξαρτάται κατά πολύ από την επιλογή της πρώτης αντίθεσης που θα κάνετε. Ας υποθέσουμε ότι υπήρχαν σοβαροί λόγοι για τη σύγκριση των $\{a,b,c,e\}$ vs. d. Για παράδειγμα, το d μπορεί να είναι το εικονικό φάρμακο και οι άλλες τέσσερις θα μπορούσαν να είναι διαφορετικά είδη φαρμακευτικής αγωγής, έτσι κάνουμε αυτή την πρώτη αντίθεση μας. Επειδή $k-1=4$ έχουμε μόνο τρεις πιθανές αντιθέσεις που είναι ορθογώνιες προς αυτή. Μπορεί να υπάρχουν *a priori* λόγοι για την ομάδα των $\{a,b\}$ και των $\{c,e\}$ έτσι ώστε να την κάνουμε τη δεύτερη ορθογώνια αντίθεση. Αυτό σημαίνει ότι δεν έχουμε βαθμούς ελευθερίας στην επιλογή των τελευταίων δύο ορθογωνίων αντιθέσεων: πρέπει να είναι a vs. b και c vs. e. Απλά να θυμάστε ότι με τις *ορθογώνιες αντιθέσεις συγκρίνετε τα πράγματα μόνο μία φορά.*

Συντελεστές αντίθεσης

Οι συντελεστές αντίθεσης είναι ένας αριθμητικός τρόπος που ενσωματώνει την υπόθεση που θέλετε να δοκιμάσετε. Οι κανόνες για την κατασκευή συντελεστών αντίθεσης είναι απλοί:

- Οι επεξεργασίες που πρέπει να εξεταστούν από κοινού παίρνουν το ίδιο πρόσημο (συν ή πλην).
- Οι ομάδες μέσων τιμών που πρέπει να αντιτεθούν παίρνουν αντίθετο πρόσημο.
- Τα επίπεδα παράγοντα που πρέπει να αποκλειστούν παίρνουν ένα συντελεστή αντίθεσης 0.
- Οι συντελεστές αντίθεσης, c, πρέπει να δίνουν άθροισμα 0.

Ας υποθέσουμε ότι με τον πέντε επιπέδων παράγοντα μας $\{a,b,c,d,e\}$ θέλουμε να ξεκινήσουμε συγκρίνοντας τα τέσσερα επίπεδα $\{a,b,c,e\}$ με το μονοεπίπεδο d. Όλα τα επίπεδα εισέρχονται στην αντίθεση, έτσι ώστε κανένας από τους συντελεστές να είναι μηδέν. Οι τέσσερις όροι $\{a,b,c,e\}$ ομαδοποιούνται έτσι ώστε να πάρουν όλοι το ίδιο πρόσημο (μείον, για παράδειγμα, αν και δεν υπάρχει καμία διαφορά ποιο πρόσημο θα επιλεγεί). Είναι που πρέπει να σύγκριθουν με το d, έτσι ώστε να πάρει το αντίθετο πρόσημο (συν, στην περίπτωση αυτή). Η επιλογή του τι αριθμητικές τιμές να δώσετε στους συντελεστές αντίθεσης είναι εξ ολοκλήρου από σας. Οι περισσότεροι άνθρωποι χρησιμοποιούν ακέραιους αριθμούς και όχι κλάσματα, αλλά αυτό πραγματικά δεν έχει σημασία. Αυτό που έχει σημασία είναι ότι το άθροισμα του c είναι 0. Οι θετικοί και οι αρνητικοί συντελεστές πρέπει να αθροίζουν την ίδια τιμή. Στο παράδειγμά μας, συγκρίνοντας τέσσερις μέσες τιμές με μία μέση τιμή, μια φυσική επιλογή των συντελεστών θα ήταν -1 για κάθε έναν από τα $\{a,b,c,e\}$ και $+4$ για το d. Εναλλακτικά, μπορούσαμε να επιλέξουμε $+0.25$ για κάθε έναν από τα $\{a,b,c,e\}$ και -1 για το d.

Factor level:	a	b	c	d	e
contrast 1 coefficients:	-1	-1	-1	4	-1

Ας υποθέσουμε ότι η δεύτερη αντίθεση είναι για να συγκρίνει τα $\{a,b\}$ με τα $\{c,e\}$. Επειδή αυτή η αντίθεση αποκλείει το d, θέτουμε το συντελεστή αντίθεση της σε 0. Τα $\{a,b\}$ παίρνουν το ίδιο πρόσημο (δηλαδή, συν) και τα $\{c,e\}$ παίρνουν το αντίθετο πρόσημο. Επειδή ο αριθμός των επιπέδων σε κάθε πλευρά της αντίθεσης είναι ίσος (2 και στις δύο περιπτώσεις) μπορούμε να χρησιμοποιήσουμε το όνομα της αριθμητικής τιμής για όλους τους συντελεστές. Η τιμή 1 είναι πλέον η προφανής επιλογή (αλλά μπορείτε να χρησιμοποιήσετε 13.7, αν θέλατε να εναντιωθείτε):

Factor level:	a	b	c	d	e
Contrast 2 coefficients:	1	1	−1	0	−1

Υπάρχουν μόνο δύο δυνατότητες για τις υπόλοιπες ορθογώνιες αντιθέσεις: a vs. b και c vs. e:

Factor level:	a	b	c	d	e
Contrast 3 coefficients:	1	−1	0	0	0
Contrast 4 coefficients:	0	0	1	0	−1

Η διακύμανση y που αποδίδεται σε μία συγκεκριμένη αντίθεση ονομάζεται **άθροισμα των τετραγώνων αντίθεσης**, SSC. Τα αθροίσματα των τετραγώνων των $k-1$ ορθογώνιων αντιθέσεων δίνουν άθροισμα στο συνολικό άθροισμα των τετραγώνων επεξεργασίας, SSA ($\sum_{i=1}^{k-1} SSC_i = SSA$ · βλ. σελ. 451.). Το άθροισμα των τετραγώνων αντίθεσης υπολογίζεται ως εξής:

$$SSC_i = \frac{(\sum(c_i T_i / n_i))^2}{\sum(c_i^2 / n_i)},$$

όπου το c_i είναι οι συντελεστές αντίθεσης (παραπάνω), n_i είναι το μέγεθος του δείγματος σε κάθε επίπεδο παράγοντα και T_i είναι τα σύνολα των y τιμών μέσα σε κάθε επίπεδο παράγοντα (που συχνά αποκαλούνται τα σύνολα επεξεργασίας). Η σπουδαιότητα μίας αντίθεσης κρίνεται από μια δοκιμή F, διαιρώντας το άθροισμα των τετραγώνων αντίθεσης με την διακύμανση του σφάλματος. Η δοκιμή F έχει 1 βαθμό ελευθερίας στον αριθμητή (γιατί η αντίθεση είναι μια σύγκριση των δύο μέσων, και $2-1=1$) και $k(n-1)$ βαθμούς ελευθερίας στον παρονομαστή (οι βαθμοί ελευθερίας των σφαλμάτων διακύμανσης) .

Ένα παράδειγμα αντιθέσεων στην R

Το ακόλουθο παράδειγμα προέρχεται από το πείραμα ανταγωνισμού που αναλύσαμε στη σελ. 155, στο οποίο η βιομάζα (biomass) των φυτών ελέγχου σε σύγκριση με τη βιομάζα (biomass) των φυτών που καλλιεργούνται σε συνθήκες όπου ο ανταγωνισμός μειώθηκε σε ένα, από τέσσερις διαφορετικούς τρόπους. Υπάρχουν δύο επεξεργασίες στις οποίες οι ρίζες των γειτονικών φυτών κόπηκαν (έως 5 cm ή 10 cm βάθος) και δύο επεξεργασίες στις οποίες οι βλαστοί των γειτονικών φυτών ψαλιδίστηκαν (clipping) (25% ή 50% των γειτονικών κόπηκαν στο επίπεδο του εδάφους) .

```
comp<-read.table("c:\\temp\\competition.txt",header=T)
attach(comp)
names(comp)
[1] "biomass" "clipping"
```

Ξεκινάμε με την μονόδρομη ανάλυση της διακύμανσης:

```
model1<-aov(biomass ~ clipping)
summary(model1)
```

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

Η επεξεργασία ψαλίδισμα (clipping) έχει μια πολύ σημαντική επίδραση στη βιομάζα (biomass). Αλλά έχουμε κατανοήσει πλήρως το αποτέλεσμα αυτού του πειράματος; Μάλλον όχι. Για παράδειγμα, ποια επίπεδα του παράγοντα είχαν τη μεγαλύτερη επίδραση στη βιομάζα, και ήταν όλα σημαντικά διαφορετικά από τους ελέγχους επεξεργασίας του ανταγωνισμού; Για να απαντήσουμε στα ερωτήματα αυτά, θα πρέπει να χρησιμοποιήσουμε την summary.lm:

```
summary.lm(model1)
```

```
Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)    465.17        28.75    16.177   9.33e-15 ***
clippingn25     88.17        40.66     2.168    0.03987   *
clippingn50    104.17        40.66     2.562    0.01683   *
clippingr10    145.50        40.66     3.578    0.00145  **
clippingr5     145.33        40.66     3.574    0.00147  **

Residual standard error: 70.43 on 25 degrees of freedom
Multiple R-Squared: 0.4077, Adjusted R-squared: 0.3129
F-statistic: 4.302 on 4 and 25 DF, p-value: 0.008752
```

Αυτό μοιάζει σαν να πρέπει να κρατήσουμε και τις πέντε παραμέτρους, γιατί και οι πέντε σειρές του συνοπτικού πίνακα έχουν ένα ή περισσότερα σημαντικά αστέρια. Εάν αληθεύει, αυτό δεν είναι η περίπτωση που μας ενδιαφέρει. Το παράδειγμα αυτό αναδεικνύει την σημαντική αδυναμία των αντιθέσεων επεξεργασίας: οι οποίες δεν δείχνουν πόσα σημαντικά επίπεδα του παράγοντα πρέπει να διατηρήσουμε στο ελάχιστο επαρκές μοντέλο, επειδή όλες οι γραμμές συγκρίνονται με το σημείο τομής (με το

controls σε αυτή την περίπτωση, απλά επειδή το όνομα του επιπέδου του παράγοντα για 'controls' προωθείται σε πλεονεκτική θέση):

```
levels(clipping)
```

```
[1] "control" "n25" "n50" "r10" "r5"
```

***Εκ των προτέρων* αντιθέσεις (*a priori*)**

Σε αυτό το πείραμα, υπάρχουν αρκετές προγραμματισμένες συγκρίσεις που θα θέλαμε να κάνουμε. Η προφανής θέση για να αρχίσουμε είναι με τη σύγκριση των φυτών ελέγχου, που εκτίθενται στην πλήρη αυστηρότητα του ανταγωνισμού, με όλες τις άλλες επεξεργασίες. Δηλαδή, θέλουμε να αντιπαραβάλλουμε το πρώτο επίπεδο ψαλίδισματος με τα άλλα τέσσερα επίπεδα. Οι συντελεστές αντίθεσης, ως εκ τούτου, θα είναι $4,-1,-1,-1,-1$. Η επόμενη προγραμματισμένη σύγκριση μπορεί να αντιπαραβάλλει τις επεξεργασίες κλαδέματος των βλαστών (n25 και n50) με τις επεξεργασίες κλαδέματος των ριζών (r10 και r5). Κατάλληλοι συντελεστές αντίθεσης για αυτήν θα ήταν $0,1,1,-1,-1$ (γιατί αγνοούμε τον έλεγχο (control) σε αυτή την αντίθεση). Μια τρίτη αντίθεση θα μπορούσε να συγκρίνει τα δύο διαφορετικά βάθη του κλαδέματος της ρίζας· $0,0,0,1,-1$. Η τελευταία ορθογώνια αντίθεση συνεπώς, θα πρέπει να συγκρίνει τις δύο εντάσεις του ψαλίδισματος (clipping) των βλαστών: $0,1,-1,0,0$. Επειδή ο παράγοντας που ονομάζεται ψαλίδισμα (clipping) έχει πέντε επίπεδα υπάρχουν μόνο $5-1=4$ ορθογώνιες αντιθέσεις.

Η R είναι εξαιρετικά καλή στο να χειρίζεται τις αντιθέσεις, και μπορούμε να συσχετίσουμε αυτές τις πέντε a priori αντιθέσεις που ορίζονται από το χρήστη με την κατηγορική μεταβλητή που ονομάζεται ψαλίδισμα (clipping) σαν αυτό:

```
contrasts(clipping)<-cbind(c(4,-1,-1,-1,-1),c(0,1,1,-1,-1),c(0,0,0,1,-1),c(0,1,-1,0,0))
```

Μπορούμε να ελέγξουμε ότι έχει κάνει όπως θέλαμε, πληκτρολογώντας

```
contrasts(clipping)
```

```
        [,1] [,2] [,3] [,4]
control    4    0    0    0
n25       -1    1    0    1
n50       -1    1    0   -1
r10       -1   -1    1    0
r5        -1   -1   -1    0
```

η οποία παράγει τον πίνακα των συντελεστών αντίθεσης που ορίσαμε. Σημειώστε ότι όλες οι στήλες αθροίζουν στο μηδέν (δηλαδή κάθε σύνολο συντελεστών αντίθεσης έχει καθοριστεί σωστά). Σημειώστε, επίσης, ότι τα γινόμενα οποιωνδήποτε από τις δύο στήλες αθροίζουν στο μηδέν (αυτό δείχνει ότι όλες οι αντιθέσεις είναι ορθογώνιες, όπως

προβλεπόταν): για παράδειγμα, συγκρίνοντας τις αντιθέσεις 1 και 2, δίνουν γινόμενα $0+(-1)+(-1)+1+1=0$.

Τώρα μπορούμε να επαναπροσαρμόσουμε το μοντέλο και να ελέγξουμε τα αποτελέσματα των αντιθέσεων που ορίσαμε , παρά την επεξεργασία των εξ ορισμού αντιθέσεων:

```
model2<-aov(biomass ~ clipping)
summary.lm(model2)
```

```
Coefficients:
             Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)  561.80000     12.85926   43.688     <2e-16 ***
clipping1    -24.15833      6.42963   -3.757   0.000921 ***
clipping2    -24.62500     14.37708   -1.713   0.099128   .
clipping3      0.08333     20. 33227   0.004   0.996762
clipping4     -8.00000     20.33227   -0.393   0.697313

Residual standard error: 70.43 on 25 degrees of freedom
Multiple R-Squared: 0.4077, Adjusted R-squared: 0.3129
F-statistic: 4.302 on 4 and 25 DF, p-value: 0.008752
```

Αντί να απαιτούνται πέντε παράμετροι (όπως προτείνεται από τις αρχικές αντιθέσεις επεξεργασίας), αυτή η ανάλυση δείχνει ότι χρειαζόμαστε μόνο δύο παραμέτρους: την συνολική μέση τιμή (561.8) και την αντίθεση μεταξύ των ελέγχων και των τεσσάρων επεξεργασιών ανταγωνισμού ($p=0.000921$). Όλες οι άλλες αντιθέσεις είναι μη σημαντικές.

Όταν ορίζουμε τις αντιθέσεις, το σημείο τομής είναι η συνολική (μέγιστη) μέση τιμή:

```
mean(biomass)
```

```
[1] 561.8
```

Η δεύτερη σειρά, με επισήμανση clipping1, εκτιμάει, όπως και όλες οι αντιθέσεις, τη διαφορά μεταξύ των δύο μέσων. Αλλά ποιές δύο μέσες τιμές, ακριβώς; Οι μέσες τιμές για τα διαφορετικά επίπεδα του παράγοντα είναι:

```
tapply(biomass,clipping,mean)
```

```
 control       n25       n50       r10       r5
465.1667  553.3333  569.3333  610.6667  610.5000
```

Η πρώτη αντίθεση συγκρίνει το controls (μέση τιμή = 465.1667, όπως παραπάνω) με τη μέση τιμή των άλλων τεσσάρων επεξεργασιών. Ο απλούστερος τρόπος για να πάρετε την άλλη μέση τιμή είναι να δημιουργηθεί ένας νέος παράγοντας, c_1 που έχει τιμή 1 για το control και 2 για τα υπόλοιπα:

```
c1<-factor(1+(clipping!="control"))
tapply(biomass,c1,mean)
       1        2
465.1667 585.9583
```

Η εκτίμηση στην πρώτη σειρά, αντανακλώντας την αντίθεση 1, είναι η διαφορά μεταξύ της συνολικής
μέσης τιμής (561.8) και η μέση τιμή των τεσσάρων επεξεργασιών μη-ελέγχου (585.9583):

```
mean(biomass)-tapply(biomass,c1,mean)
       1         2
96.63333 -24.15833
```

και θα δείτε την εκτίμηση στη γραμμή 2, όπως −24.15833. Τι γίνεται με τη δεύτερη αντίθεση στη γραμμή 3; Αυτό συγκρίνει τις επεξεργασίες κλαδέματος της ρίζας και του βλαστού και ο c_2 είναι ένας παράγοντας που ομαδοποιεί επεξεργασίες της ρίζας και τις δύο του βλαστού.

```
c2<factor(2*(clipping=="n25")+2*(clipping=="n50")+(clipping=="r10")+(clipping=="
r5"))
```

Μπορούμε να υπολογίσουμε τη μέση βιομάζα για τις δύο επεξεργασίες χρησιμοποιώντας tapply, μετά αφαιρούμε τις μέσες τιμές της μίας από την άλλη χρησιμοποιώντας τη συνάρτηση diff και στη συνέχεια μειώνουμε κατα το ήμισυ τις διαφορές:

```
diff(tapply(biomass,c2,mean))/2
       1         2
72.70833 -24.62500
```

Έτσι, η δεύτερη αντίθεση στην σειρά 3 (−24.625) είναι *το ήμισυ της διαφοράς* μεταξύ της επεξεργασίας κλαδέματος ρίζας και βλαστού . Τι γίνεται με την τρίτη σειρά; Ο αριθμός αντίθεσης 3 είναι μεταξύ των δύο επεξεργασιών κλαδέματος ρίζας. Γνωρίζουμε τις τιμές τους ήδη από την tapply, παραπάνω:

```
    r10       r5
610.6667 610.5000
```

Οι δύο μέσες τιμές διαφέρουν κατά 0.166666 έτσι ώστε η τρίτη αντίθεση είναι το ήμισυ της διαφοράς μεταξύ των δύο μέσων τιμών:

```
(610.666666-610.5)/2
[1] 0.083333
```

Η τελική αντίθεση συγκρίνει τις δύο επεξεργασίες κλαδέματος ρίζας, και η αντίθεση είναι το ήμισυ της διαφοράς μεταξύ των δύο αυτών μέσων τιμών:

```
(553.3333-569.3333)/2
[1] -8
```

Για να ανακεφαλαιώσουμε: η πρώτη αντίθεση συγκρίνει τη συνολική μέση τιμή με τη μέση τιμή των τεσσάρων επεξεργασιών μη ελέγχου, η δεύτερη αντίθεση είναι κατά το ήμισυ η διαφορά μεταξύ των μέσων τιμών επεξεργασίας του κλαδέματος της ρίζας και του βλαστού , η τρίτη αντίθεση είναι το ήμισυ της διαφοράς μεταξύ των δύο επεξεργασιών κλαδέματος της ρίζας, και η τέταρτη αντίθεση είναι το ήμισυ της διαφοράς μεταξύ των δύο επεξεργασιών κλαδέματος του βλαστού.

Είναι σημαντικό να σημειωθεί ότι τα πρώτα τέσσερα πρότυπα σφάλματα στον πίνακα summary.lm είναι όλα διαφορετικά το ένα από το άλλο. Όπως είδαμε, η εκτίμηση στην πρώτη γραμμή του πίνακα είναι μια μέση τιμή, ενώ όλες οι άλλες γραμμές περιέχουν εκτιμήσεις ότι υπάρχουν *διαφορές μεταξύ των μέσων τιμών*. Η συνολική μέση τιμή της επάνω σειράς βασίζεται σε 30 αριθμούς έτσι ώστε το πρότυπο σφάλμα της μέσης τιμής είναι $se = \sqrt{s^2 / 30}$, όπου το s^2 προέρχεται από τον πίνακα ANOVA:

```
sqrt(4961/30)
[1] 12.85950
```

Η μικρή διαφορά στο τέταρτο δεκαδικό ψηφίο οφείλεται σε σφάλματα στρογγυλοποίησης στην κλήση της διακύμανσης 4961.0. Η επόμενη σειρά συγκρίνει δύο μέσες τιμές γι 'αυτό χρειαζόμαστε το πρότυπο σφάλμα της διαφοράς μεταξύ των δύο μέσων τιμών. Η πολυπλοκότητα προέρχεται από το γεγονός ότι για τις δύο μέσες τιμές, κάθε μία βασίζεται σε διαφορετικούς αριθμούς. Η συνολική μέση τιμή βασίζεται σε όλα τα πέντε επίπεδα του παράγοντα (30 αριθμοί), ενώ η μη-ελέγχου μέση τιμή με την οποία συγκρίνεται βασίζεται σε τέσσερεις μέσες τιμές (24 αριθμοί). Κάθε επίπεδο έχει παράγοντα n = 6 επαναλήψεων, έτσι ώστε ο παρονομαστής στον πρότυπο τύπο σφάλματος είναι 5 × 4 × 6 = 120. Ετσι, το πρότυπο σφάλμα της διαφοράς μεταξύ των δύο αυτών μέσων τιμών είναι $se = \sqrt{s^2 / (5 \times 4 \times 6)}$:

```
sqrt(4961/(5*4*6))
[1] 6.429749
```

Για τη δεύτερη αντίθεση στην σειρά 3, κάθε μία από τις μέσες τιμές βασίζονται σε 12 αριθμούς, ώστε το πρότυπο σφάλμα είναι $se\sqrt{2 \times (s^2 / 12)}$ έτσι ώστε το πρότυπο σφάλμα είναι το ήμισυ της διαφοράς είναι

```
sqrt(2*(4961/12))/2
[1] 14.37735
```

Οι τελευταίες δύο αντιθέσεις είναι και οι δύο μέσες τιμές που βασίζονται σε έξι αριθμούς, έτσι ώστε το πρότυπο σφάλμα

```
sqrt(2*(4961/6))/2
[1] 20.33265
```

Η πολυπλοκότητα των υπολογισμών αυτών είναι ένας άλλος λόγος που προτιμούν την επεξεργασία αντίθεσης, παρά τις αντιθέσεις που καθορίζονται από το χρήστη ως προεπιλογή. Το πλεονέκτημα των ορθογώνιων αντιθέσεων, ωστόσο, είναι ότι ο πίνακας summary.lm μας δίνει μια πολύ καλύτερη ιδέα για τον αριθμό των παραμέτρων που απαιτούνται στο ελάχιστο επαρκές μοντέλο (2 στην περίπτωση αυτή). Οι αντιθέσεις επεξεργασίας έχουν πέντε σημαντικά αστέρια σε όλες τις σειρές (βλέπε παρακάτω), επειδή όλες οι μη ελέγχου επεξεργασίες συγκρίνονται με την ομάδα ελέγχου (το σημείο τομής).

Απλοποίηση μοντέλου με σταδιακή διαγραφή

Μια εναλλακτική λύση για τον καθορισμό των αντιθέσεων από μας (όπως παραπάνω) είναι να συναθροίσετε τα μη σημαντικά επίπεδα παράγοντα σε μια *a posteriori* σταδιακή διαδικασία . Για να αποδειχθεί αυτό, επανερχόμαστε στις αντιθέσεις της επεξεργασίας. Πρώτον, θα απενεργοποιήσουμε τις οριζόμενες από το χρήστη αντιθέσεις :

```
contrasts(clipping)<-NULL
options(contrasts=c("contr.treatment","contr.poly"))
```

Τώρα μπορούμε να προσαρμόσουμε το μοντέλο με τα πέντε επίπεδα του παράγοντα ως σημείο εκκίνησης:

```
model3<-aov(biomass ~ clipping)
summary.lm(model3)
```

```
Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)    465.17       28.75   16.177   9.33e-15  ***
clippingn25     88.17       40.66    2.168    0.03987   *
clippingn50    104.17       40.66    2.562    0.01683   *
clippingr10    145.50       40.66    3.578    0.00145   **
clippingr5     145.33       40.66    3.574    0.00147   **
```

Κοιτάζοντας προς τα κάτω τον κατάλογο των εκτιμήσεων των παραμέτρων, βλέπουμε ότι οι πιο όμοιοι είναι τα αποτελέσματα του κλαδέματος ρίζας σε 10 cm και 5 cm (145.5 έναντι 145.33). Θα ξεκινήσουμε με την απλοποίηση αυτών σε μία μόνο επεξεργασία κλαδέματος ρίζας που ονομάζεται root. Το κόλπο είναι να χρησιμοποιήσετε το βέλος ‘gets’ για να αλλάξετε τα ονόματα των κατάλληλων επιπέδων του παράγοντα. Ξεκινήστε με την αντιγραφή του αρχικού ονόματος του παράγοντα:

```
clip2<-clipping
```

Ελέγξτε τώρα τους αριθμούς επιπέδου των διαφόρων ονομάτων του επιπέδου παράγοντα:

```
levels(clip2)
[1] "control" "n25" "n50" "r10" "r5"
```

Το σχέδιο είναι να συναθροιστούν μαζί r10 και r5 με το ίδιο όνομα, 'root'. Πρόκειται για το τέταρτο και πέμπτο επίπεδο clip2, έτσι γράφουμε:

```
levels(clip2)[4:5]<-"root"
```

Αν πληκτρολογήσουμε

```
levels(clip2)
[1] "control" "n25" "n50" "root"
```

βλέπουμε ότι το r10 και r5 έχουν όντως αντικατασταθεί από το 'root'.

Το επόμενο βήμα είναι να προσαρμοστεί ένα νέο μοντέλο με clip2 στη θέση του clipping, και να εξετάσει αν το νέο απλούστερο μοντέλο είναι σημαντικά χειρότερο ως περιγραφή των δεδομένων με τη χρήση anova:

```
model4<-aov(biomass ~ clip2)
anova(model3,model4)

Analysis of Variance Table
Model 1: biomass~clipping
Model 2: biomass~clip2
     Res.Df      RSS     Df   Sum of Sq           F  Pr(>F)
1        25    124020
2        26    124020     -1   -0.083333 0.0000168    0.9968
```

Όπως αναμενόταν, η απλούστευση αυτού του μοντέλου είναι απολύτως δικαιολογημένη.

Το επόμενο βήμα είναι να διερευνήθούν οι επιδράσεις που χρησιμοποιούν την summary.lm:

```
summary.lm(model4)
```

```
Coefficients:
             Estimate   Std. Error   t value    Pr(>|t|)
(Intercept)    465.17        28.20    16.498    2.66e-15 ***
clip2n25        88.17        39.87     2.211     0.03602   *
clip2n50       104.17        39.87     2.612    0.014744   *
clip2root      145.42        34.53     4.211    0.000269 ***
```

Φαίνεται σαν να μην είναι σημαντικά διαφορετικές μεταξύ τους οι δύο επεξεργασίες ψαλιδίσματος του βλαστού (n25 και n50) (διαφέρουν κατά μόλις 16.0 με ένα πρότυπο σφάλμα του 39.87). Μπορούμε να τις ομαδοποιήσουμε αυτές μαζί, σε μία επεξεργασία κλαδέματος βλαστού ως εξής:

```
clip3<-clip2
levels(clip3)[2:3]<-"shoot"
levels(clip3)
[1] "control" "shoot" "root"
```

Στη συνέχεια, προσαρμόζονται σε ένα νέο μοντέλο με clip3 στη θέση του clip2:

```
model5<-aov(biomass ~ clip3)
anova(model4,model5)

Analysis of Variance Table
Model 1: biomass~clip2
Model 2: biomass~clip3
   Res.Df     RSS    Df   Sum of Sq      F Pr(>F)
1      26  124020
2      27  124788    -1        -768  0.161 0.6915
```

Και πάλι, η απλούστευση αυτή ήταν απόλυτα δικαιολογημένη. Μήπως διαφέρουν οι επεξεργασίες ανταγωνισμού ρίζας και βλαστού;

```
clip4<-clip3
levels(clip4)[2:3]<-"pruned"
levels(clip4)
[1] "control" "pruned"
```

Τώρα προσαρμόζονται σε ένα νέο μοντέλο με clip4 στη θέση του clip3:

```
model6<-aov(biomass ~ clip4)
anova(model5,model6)
```

```
Analysis of Variance Table

Model 1: biomass~clip3
Model 2: biomass~clip4
    Res.Df     RSS    Df   Sum of Sq         F Pr(>F)
1       27  124788
2       28  139342    -1      -14553  3.1489  0.08726.
```

Αυτή η απλοποίηση ήταν σημαντικά κοντά, αλλά είμαστε αδίστακτοι (p >0.005) , και έτσι δεχόμαστε την απλοποίηση. Τώρα έχουμε το ελάχιστο επαρκές μοντέλο:

summary.lm(model6)

```
Coefficients:
            Estimate  Std. Error   t value   Pr(>|t|)
(Intercept)    465.2        28.8    16.152   1.11e-15 ***
clip4pruned    120.8        32.2     3.751   0.000815 ***
```

Έχει μόνο δύο παραμέτρους: ο μέσος όρος για τους ελέγχους (465.2) και η διαφορά μεταξύ της μέσης τιμής ελέγχου και της 4ης επεξεργασίας μέσης τιμής (465.2 + 120.8 = 586.0):

tapply(biomass,clip4,mean)

```
 control     pruned
465.1667   585.9583
```

Γνωρίζουμε ότι αυτές οι δύο μέσες τιμές είναι σημαντικά διαφορετικές από την τιμή p 0.000815, αλλά για να δούμε πώς γίνεται, μπορούμε να κάνουμε ένα τελικό model7 που δεν έχει καθόλου καμία επεξηγηματική μεταβλητή (χωράει μόνο τη συνολική μέση τιμή). Αυτό επιτυγχάνεται με το γράψιμο y ~ 1 στον τύπο μοντέλου:

```
model7<-aov(biomass~1)
anova(model6,model7)
Analysis of Variance Table

Model 1: biomass~clip4
Model 2: biomass~1
    Res.Df     RSS  Df   Sum of Sq         F Pr(>F)
1       28  139342
2       29  209377  -1       -70035 14.073  0.000815***
```

Σημειώστε ότι η τιμή *p* είναι ακριβώς η ίδια όπως στο model6. Οι τιμές *p* στην R υπολογίζονται έτσι ώστε να αποφευχθεί η ανάγκη για αυτό το τελικό στάδιο στην απλοποίηση μοντέλου: είναι ‘ *p* για τη διαγραφή’ τιμών.

Σύγκριση των τριών τύπων των αντιθέσεων

Προκειμένου να δείτε τις διαφορές μεταξύ της επεξεργασίας, Helmert και του αθροίσματος των αντιθέσεων, θα αναλύσουμε εκ νέου το πείραμα ανταγωνισμού χρησιμοποιώντας την κάθε μία με τη σειρά.

Επεξεργασία αντιθέσεων

Αυτή είναι προεπιλογμένη στην R. Αυτές είναι οι αντιθέσεις που παίρνετε, εκτός και αν επιλέξετε ρητά το αντίθετο.

```
options(contrasts=c("contr.treatment","contr.poly"))
```

Εδώ είναι οι συντελεστές αντίθεσης, όπως ορίζονται σύμφωνα με τις αντιθέσεις επεξεργασίας:

```
contrasts(clipping)
```

```
        n25 n50 r10 r5
control   0   0   0  0
n25       1   0   0  0
n50       0   1   0  0
r10       0   0   1  0
r5        0   0   0  1
```

Παρατηρήστε ότι οι αντιθέσεις δεν είναι ορθογώνιες (τα γινόμενα των συντελεστών δεν αθροίζουν στο μηδέν)

```
output.treatment<-lm(biomass ~ clipping)
summary(output.treatment)
```

```
Coefficients:
             Estimate    Std. Error    t value  Pr(>|t|)
(Intercept)    465.17         28.75     16.177  9.33e-15 ***
clippingn25     88.17         40.66      2.168  0.03987    *
clippingn50    104.17         40.66      2.562  0.01683    *
clippingr10    145.50         40.66      3.578  0.00145   **
clippingr5     145.33         40.66      3.574  0.00147   **
```

Με τις αντιθέσεις επεξεργασίας, τα επίπεδα του παράγοντα είναι τοποθετημένα σε αλφαβητική σειρά και το επίπεδο που έρχεται πρώτο γίνεται το σημείο τομής. Στο παράδειγμά μας, αυτό είναι το control, έτσι ώστε να μπορέσουμε να διαβάσουμε την μέση τιμή του control ως 465.17, και το πρότυπο σφάλμα της μέσης τιμής ως 28.75. Οι υπόλοιπες τέσσερις σειρές είναι οι διαφορές μεταξύ των μέσων τιμών, καθώς και τα πρότυπα σφάλματα είναι τα πρότυπα σφάλματα των διαφορών. Έτσι, ψαλιδίζοντας γειτονικά πίσω σε 25 εκατοστά αυξάνεται η βιομάζα (biomass) από 88.17 πάνω από τους ελέγχους(controls) και αυτή η διαφορά είναι σημαντική σε $p = 0.03987$. Και ούτω καθεξής. Το μειονέκτημα των αντιθέσεων επεξεργασίας είναι ότι όλες οι γραμμές φαίνεται να είναι σημαντικές, παρά το γεγονός ότι οι γραμμές 2 ως 5 δεν είναι στην πραγματικότητα σημαντικά διαφορετικές μεταξύ τους, όπως είδαμε νωρίτερα.

Αντιθέσεις Helmert

Αυτή είναι η προεπιλογή στην S-PLUS, οπότε προσέξτε εάν εναλλάσσεστε μεταξύ των δύο γλωσσών.

```
options(contrasts=c("contr.helmert","contr.poly"))
contrasts(clipping)
```

```
         [,1] [,2] [,3] [,4]
control    -1   -1   -1   -1
n25         1   -1   -1   -1
n50         0    2   -1   -1
r10         0    0    3   -1
r5          0    0    0    4
```

Παρατηρήστε ότι οι αντιθέσεις είναι ορθογώνιες (το άθροισμα των γινομένων είναι μηδέν) και οι συντελεστές τους αθροίζουν στο μηδέν, σε αντίθεση με τις αντιθέσεις επεξεργασίας, απο πάνω.

```
output.helmert<-lm(biomass ~ clipping) summary(output.helmert)
```

```
Coefficients:
            Estimate  Std. Error   t value   Pr(>|t|)
(Intercept)  561.800      12.859    43.688     <2e-16 ***
clipping1     44.083      20.332     2.168     0.0399   *
clipping2     20.028      11.739     1.706     0.1004
clipping3     20.347       8.301     2.451     0.0216   *
clipping4     12.175       6.430     1.894     0.0699   .
```

Με τις αντιθέσεις Helmert, το σημείο τομής είναι η συνολική μέση τιμή (561.8). Η πρώτη αντίθεση (αντίθεση 1 στη σειρά 2, επισημαίνεται ως clipping1) συγκρίνει την πρώτη μέση τιμή σε αλφαβητική αλληλουχία, με το μέσο όρο του πρώτου και του δεύτερου παράγοντα επιπέδων σε αλφαβητική σειρά (control και n25· βλέπε παραπάνω): η τιμή της παραμέτρου του είναι ο μέσος όρος των δύο πρώτων επιπέδων του παράγοντα, μείον τη μέση τιμή του πρώτου επιπέδου παράγοντα:

```
(465.16667+553.33333)/2-465.166667
```

```
[1] 44.08332
```

Η τρίτη σειρά περιέχει την αντίθεση ανάμεσα στο τρίτο επίπεδο παράγοντα (N50) και των δύο επιπέδων που έχουν ήδη συγκριθεί (control και n25): η τιμή του είναι η διαφορά μεταξύ του μέσου όρου των 3 πρώτων επιπέδων του παράγοντα και του μέσου όρου των δύο πρώτων επιπέδων παράγοντα:

```
(465.16667+553.33333+569.333333)/3-(465.166667+553.3333)/2
```

```
[1] 20.02779
```

Η τέταρτη σειρά περιέχει την αντίθεση ανάμεσα στο τέταρτο επίπεδο παράγοντα (r10) και τα τρία επίπεδα που έχουν ήδη συγκριθεί (control, n25 και n50): η τιμή του είναι η διαφορά μεταξύ του μέσου όρου των τεσσάρων πρώτων επιπέδων του παράγοντα και του μέσου όρου των τριών πρώτων επιπέδων παράγοντα:

```
(465.16667+553.33333+569.333333+610.66667)/
4-(553.3333+465.166667+569.3333)/3
```

```
[1] 20.34725
```

Η πέμπτη και τελευταία σειρά περιέχει την αντίθεση μεταξύ του πέμπτου επιπέδου παράγοντα (r5) και των τεσσάρων επιπέδων που έχουν ήδη συγκριθεί (control, n25, n50 και r10): η τιμή του είναι η διαφορά μεταξύ του μέσου όρου των πρώτων πέντε επιπέδων παράγοντα (τη μέγιστη μέση τιμή), και του μέσου όρου των τεσσάρων πρώτων επιπέδων παράγοντα:

```
mean(biomass)-(465.16667+553.33333+569.333333+610.66667)/4
```

```
[1] 12.175
```

Τόσα πολλά για τις εκτιμήσεις των παραμέτρων. Τώρα κοιτάξτε τα πρότυπα σφάλματα. Έχουμε δει μάλλον λίγες από τις τιμές τους σε οποιαδήποτε από τις αναλύσεις που έχουμε κάνει μέχρι τώρα. Το τυπικό σφάλμα στη γραμμή 1 είναι το τυπικό σφάλμα της συνολικής μέσης τιμής, με το s^2 να λαμβάνεται από το συνολικό πίνακα ANOVA: $\sqrt{s^2/2n}$.

```
sqrt(4961/30)
```

```
[1] 12.85950
```

Το τυπικό σφάλμα στη γραμμή 2 είναι μια σύγκριση της ομάδας των δύο μέσων τιμών με μία μέση τιμή (2 × 1 = 2). Ετσι το 2 πολλαπλασιάζεται με το μέγεθος του δείγματος *n* στον παρονομαστή: $\sqrt{s^2 / 2n}$.

```
sqrt(4961/(2*6))
[1] 20.33265
```

Το τυπικό σφάλμα στη γραμμή 3 είναι μια σύγκριση της ομάδας των τριών μέσων τιμών με μια ομάδα των δύο μέσων τιμών (έτσι 3 × 2 = 6 στον παρονομαστή): $\sqrt{s^2 / 6n}$.

```
sqrt(4961/(3*2*6))
[1] 11.73906
```

Το τυπικό σφάλμα στη γραμμή 4 είναι μια σύγκριση της ομάδας των τεσσάρων μέσων τιμών με μια ομάδα των τριών μέσων τιμών (έτσι 4 × 3 = 12 στον παρονομαστή): $\sqrt{s^2 / 12n}$.

```
sqrt(4961/(4*3*6))
[1] 8.30077
```

Το τυπικό σφάλμα στη γραμμή 5 είναι μια σύγκριση της ομάδας των πέντε μέσων τιμών με μια ομάδα των τεσσάρων μέσων τιμών (έτσι 5 × 4 = 20 στον παρονομαστή): $\sqrt{s^2 / 20n}$.

```
sqrt(4961/(5*4*6))
[1] 6.429749
```

Είναι αλήθεια ότι οι εκτιμήσεις των παραμέτρων και τα πρότυπα σφάλματα τους είναι πολύ πιο δύσκολο να κατανοηθούν με τον Helmert από ό, τι με τις αντιθέσεις επεξεργασίας. Αλλά το πλεονέκτημα των αντιθέσεων Helmert είναι ότι σας δίνουν τις σωστές ορθογώνιες αντιθέσεις, και ως εκ τούτου δίνουν μια πιο σαφή εικόνα ποια επίπεδα του παράγοντα πρέπει να διατηρηθούν στο ελάχιστο επαρκές μοντέλο. Ωστόσο, δεν εξαλείφουν την ανάγκη για προσεκτική απλοποίηση μοντέλου. Όπως είδαμε νωρίτερα, το παράδειγμα αυτό απαιτεί μόνο δύο παραμέτρους στο ελάχιστο επαρκές μοντέλο, αλλά οι Helmert αντιθέσεις (όπως παραπάνω) υποδεικνύουν την ανάγκη για τρείς (ωστόσο μόνο οριακά σημαντικές) παραμέτρους.

Άθροισμα αντιθέσεων

Τα αθροίσματα αντιθέσεων είναι η τρίτη εναλλακτική λύση:

```
options(contrasts=c("contr.sum","contr.poly"))
output.sum<-lm(biomass ~ clipping)
summary(output.sum)
```

```
Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  561.800      12.859   43.688    <2e-16 ***
clipping1    -96.633      25.719   -3.757  0.000921 ***
clipping2     -8.467      25.719   -0.329  0.744743
clipping3      7.533      25.719    0.293  0.772005
clipping4     48.867      25.719    1.900  0.069019   .
```

Όπως με τις Helmert αντιθέσεις, η πρώτη γραμμή περιέχει τη συνολική μέση τιμή και το πρότυπο σφάλμα του συνολικού μέσου όρου. Οι υπόλοιπες τέσσερις σειρές είναι διαφορετικές: είναι οι διαφορές μεταξύ της μέγιστης μέσης τιμής και των τεσσάρων πρώτων μέσων τιμών του παράγοντα (control, n25, n50, r10 και r5):

```
tapply(biomass,clipping,mean) - 561.8
   control        n25        n50        r10         r5
-96.633333  -8.466667   7.533333  48.866667  48.700000
```

Τα πρότυπα σφάλματα είναι όλα τα ίδια (25.719) για όλες τις τέσσερις αντιθέσεις. Οι αντιθέσεις συγκρίνουν τη μέγιστη μέση τιμή (με βάση 30 αριθμούς) με μία μόνο επεξεργασία μέσης τιμής

```
sqrt(4961/30+4961/10)
[1] 25.71899
```

Δημιουργώντας Ψευδώνυμα (aliasing)

Η δυνατότητα ψευδωνύμων συμβαίνει όταν δεν υπάρχουν διαθέσιμες πληροφορίες σχετικά με τις οποίες να μπορεί να βασιστεί μια εκτίμηση της αξίας παραμέτρου. Οι παράμετροι μπορούν να γίνουν ψευδώνυμα για έναν από τους δύο λόγους:

- δεν υπάρχουν στοιχεία στο πλαίσιο των δεδομένων από το οποίο να εκτιμηθεί η παράμετρος (π.χ. τιμές που λείπουν, τμηματικά σχέδια ή συσχετισμός μεταξύ των ερμηνευτικών μεταβλητών), ή

- το μοντέλο είναι δομημένο κατά τέτοιο τρόπο ώστε η τιμή της παραμέτρου δεν μπορεί να εκτιμηθεί (π.χ. υπερπροσδιορισμένα μοντέλα με περισσότερες παραμέτρους από ό, τι είναι απαραίτητο)

Ενδογενής aliasing εμφανίζεται λόγω της *δομής του μοντέλου*. **Εξωγενής aliasing** εμφανίζεται λόγω της *φύσης των δεδομένων*.

Ας υποθέσουμε ότι σε ένα παραγοντικό πείραμα όλα τα ζώα που έλαβαν το επίπεδο 2 της διατροφής (παράγοντας A) και το επίπεδο 3, της θερμοκρασίας (παράγοντας B) πέθαναν κατά λάθος ως αποτέλεσμα της επίθεσης από έναν παθογόνο μύκητα. Αυτός ο συγκεκριμένος συνδυασμός δίαιτας και θερμοκρασίας δεν συνεισφέρει κάποιο δεδομένο στην μεταβλητή απόκρισης, έτσι ο όρος αλληλεπίδρασης $A(2)$: $B(3)$ δεν μπορεί να εκτιμηθεί. Είναι **εξωγενώς δημιουργία ψευδωνύμου**, και η εκτίμηση των παραμέτρων του έχουν οριστεί σε μηδέν.

Αν μία συνεχής μεταβλητή έχει απόλυτα συσχετιστεί με μια άλλη μεταβλητή που έχει ήδη προσαρμοστεί στα δεδομένα (ίσως επειδή είναι ένα σταθερό πολλαπλάσιο της πρώτης μεταβλητής), τότε ο δεύτερος όρος έχει δημιουργηθεί ως ψευδώνυμο και δεν προσθέτει τίποτα στο μοντέλο. Ας υποθέσουμε ότι $x_2 = 0.5x_1$ στη συνέχεια η προσαρμογή ενός μοντέλου με $x_1 + x_2$ θα οδηγήσει σε x_2 που ενδογενώς έχει δημιουργηθεί ψευδώνυμο και θα δώσει μηδενική εκτίμηση των παραμέτρων.

Αν όλες οι τιμές μιας συγκεκριμένης ερμηνευτικής μεταβλητής μηδενίζονται για ένα δεδομένο επίπεδο ενός συγκεκριμένου παράγοντα, τότε γι'αυτό το επίπεδο σκόπιμα έχει δημιουργηθεί ψευδώνυμο. Αυτό το είδος των ψευδωνύμων είναι ένα χρήσιμο τέχνασμα προγραμματισμού της ANCOVA όταν θέλουμε μια συμεταβλητή να προσαρμοστεί σε κάποια επίπεδα του παράγοντα, αλλά όχι σε άλλα.

Ορθογώνιες πολυωνυμικές αντιθέσεις: contr.poly

Εδώ είναι τα δεδομένα από ένα τυχαιοποιημένο πείραμα με τέσσερα επίπεδα συμπληρώματος διατροφής:

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

```
[1] "treatment" "response"
```

Αρχίζουμε σημειώνοντας ότι τα επίπεδα του παράγοντα είναι με αλφαβητική σειρά (όχι σε αλληλουχία κατάταξης - none, low, medium, high - όπως θα προτιμούσαμε):

```
tapply(response,treatment,mean)
```

```
high   low   medium   none
4.50  5.25     7.00   2.50
```

Ο πίνακας summary.lm από τη μονόδρομη ανάλυση της διακύμανσης μοιάζει με

```
model<-lm(response ~ treatment)
summary(model)

Call:

lm(formula = response~treatment)

Residuals :

       Min            1Q       Median           3Q          Max

-1.250e+00   -5.000e-01    1.388e-16   5.000e-01   1.000e+00

Coefficients:

              Estimate  Std. Error  t value  Pr(>|t|)

(Intercept)     4.5000      0.3750   12.000  4.84e-08 ***

treatmentlow    0.7500      0.5303    1.414  0.182717

treatmentmedium 2.5000      0.5303    4.714  0.000502 ***

treatmentnone  -2.0000      0.5303   -3.771  0.002666 **


Residual standard error: 0.75 on 12 degrees of freedom
Multiple R-Squared: 0.8606, Adjusted R-squared: 0.8258
F-statistic: 24.7 on 3 and 12 DF, p-value: 2.015e-05
```

Ο πίνακας summary.aov μοιάζει με:

```
summary.aov(model)

            Df  Sum Sq   Mean Sq    F value      Pr(>F)

treatment    3  41.687    13.896     24.704   2.015e-05 ***

Residuals   12   6.750     0.563
```

Μπορούμε να δούμε ότι η επεξεργασία είναι ένας παράγοντας, αλλά δεν διατάσσεται:

```
is.factor(treatment)

[1] TRUE

is.ordered(treatment)

[1] FALSE
```

Για να την μετατρέψετε σε ένα διατεταγμένο παράγοντα, χρησιμοποιήστε την συνάρτηση ordered όπως :

```
treatment<-ordered(treatment,levels=c("none","low","medium","high"))
levels(treatment)

[1] "none" "low" "medium" "high"
```

Τώρα, τα επίπεδα του παράγοντα εμφανίζονται στην διατεταγμένη ακολουθία τους, όχι όμως με αλφαβητική σειρά. Προσαρμόζοντας τον διατεταγμένο παράγοντα δεν κάνει κάποια διαφορά στον πίνακα summary.aov:

```
model2<-lm(response ~ treatment)
summary.aov(model2)
```

```
            Df    Sum Sq    Mean Sq    F value     Pr(>F)
treatment    3    41.687     13.896     24.704  2.015e-05 ***
Residuals   12     6.750      0.562
```

αλλά ο πίνακας summary.lm είναι θεμελιωδώς διαφορετικός όταν διατάσσονται οι παράγοντες. Τώρα οι αντιθέσεις δεν είναι contr.treatment αλλά contr.poly (που σημαίνει ορθογώνιες πολυωνυμικές αντιθέσεις):

```
summary(model2)
```

```
Call:lm(formula = response~treatment)


Residuals:
       Min           1Q       Median           3Q          Max
-1.250e+00  -5.000e-01  -1.596e-16  5.000e-01  1.000e+00
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    4.8125      0.1875   25.667  7.45e-12 ***
treatment.L    1.7330      0.3750    4.621  0.000589 ***
treatment.Q   -2.6250      0.3750   -7.000  1.43e-05 ***
treatment.C   -0.7267      0.3750   -1.938  0.076520   .

Residual standard error: 0.75 on 12 degrees of freedom
Multiple R-Squared: 0.8606, Adjusted R-squared: 0.8258
F-statistic: 24.7 on 3 and 12 DF, p-value: 2.015e-05
```

Τα επίπεδα του παράγοντα που ονομάζονται επεξεργασίες δεν επισημαίνονται πλέον low, medium, none, όπως με τις αντιθέσεις επεξεργασίας (παραπάνω). Αντ' αυτού, είναι χαρακτηρισμένα L, Q και C, που ξεχωρίζουν για τους γραμμικούς, τετραγωνικούς και κυβικούς πολυωνυμικούς όρους, αντίστοιχα. Αλλά τι είναι οι συντελεστές, και γιατί είναι τόσο δύσκολο να ερμηνευθούν; Το πρώτο πράγμα που θα παρατηρήσετε είναι ότι το σημείο τομής 4.8125 δεν είναι πλέον ένα από τις επεξεργασίες μέσης τιμής:

```
tapply(response,treatment, mean)
```

```
none  low   medium  high
2.50  5.25    7.00  4.50
```

Θα μπορούσατε να προσαρμόσετε ένα πολυωνυμικό μοντέλο παλινδρόμησης στις μέσες τιμές της απόκρισης με τα τέσσερα επίπεδα επεξεργασίας που αντιπροσωπεύεται από μία συνεχή (dummy) επεξηγηματική μεταβλητή (δηλαδή, x <-c (1, 2, 3, 4)), και στη συνέχεια να προσαρμόσετε τους όρους για το x, x^2 και x^3 ανεξάρτητα. Αυτό θα έμοιαζε κάπως έτσι:

```
yv<-as.vector(tapply(response,treatment,mean))
x<-1:4 model<-lm(yv ~ x+1(x^2)+I(x^3)) summary(model)
Call:
lm(formula = yv~x + I(x^2) + I(x^3))

Residuals:ALL 4 residuals are 0: no residual degrees of
freedom!

Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   2.0000          NA       NA        NA
x            -1.7083          NA       NA        NA
I(x^2)        2.7500          NA       NA        NA
I(x^3)       -0.5417          NA       NA        NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-Squared: 1, Adjusted R-squared: NaN
F-statistic: NaN on 3 and 0 DF, p-value: NA Call:
lm(formula = yv~xv + I(xv^2) + I(xv^3))
```

Έτσι η εξίσωση του y ως συνάρτηση επεξεργασίας του (x) θα μπορούσε να γραφτεί

$$y = 2 - 1.7083x + 2.75x^2 - 0.5417x^3 .$$

Σημειώστε ότι το σημείο τομής *δεν* είναι ένα από τις μέσες τιμές του παράγοντα επιπέδου (η μέση τιμή του επιπέδου παράγοντα 1 (καμία) είναι η εξίσωση που αξιολογήθηκε για $x = 1$ (δηλαδή $2 - 1.7083 + 2.75 - 0.5417 = 2.5$) Γιατί λοιπόν η R δεν το κάνει με αυτόν τον τρόπο; Υπάρχουν δύο βασικοί λόγοι: ορθογωνικότητα και υπολογιστική ακρίβεια. Εάν οι γραμμικές, τετραγωνικές και κυβικές αντιθέσεις είναι ορθογώνιες και προσαρμοσμένες σταδιακά, τότε μπορούμε να δούμε εάν η προσθήκη ενός επιπλέον όρου παράγει σημαντικά βελτιωμένη επεξηγηματική δύναμη στο μοντέλο. Σε αυτή την περίπτωση, για παράδειγμα, δεν υπάρχει καμία δικαιολογία για τη διατήρηση του κύβικου όρου ($p = 0.07652$). Η υπολογιστική ακρίβεια μπορεί να γίνει ένα σημαντικό πρόβλημα κατά την προσαρμογή πολλών πολυωνυμικών όρων, γιατί αυτοί οι όροι συσχετίζονται απαραιτήτως πολύ:

```
x<-1:4
x2<-x^2
x3<-x^3 cor(cbind(x,x2,x3))
```

```
           x         x2        x3
x  1.0000000  0.9843740 0.9513699
x2 0.9843740  1.0000000 0.9905329
x3 0.9513699  0.9905329 1.0000000
```

Οι ορθογώνιες πολυωνυμικές αντιθέσεις διορθώνουν και τα δυο προβλήματα αυτά ταυτόχρονα. Εδώ είναι ένας τρόπος για την απόκτηση ορθογώνιων πολυωνυμικών αντιθέσεων για έναν παράγοντα με τέσσερα επίπεδα. Οι αντιθέσεις θα ανέλθουν σε πολυώνυμα του βαθμού $k-1=4-1=3$.

όρος	x_1	x_2	x_3	x_4
γραμμικός	−3	−1	1	3
τετραγωνικός	1	−1	−1	1
κυβικός	−1	3	−3	1

Σημειώστε ότι οι γραμμικοί όροι x ισαπέχουν, και έχουν μια μέση τιμή μηδέν (δηλαδή κάθε σημείο του άξονα x διαχωρίζεται με 2). Επίσης, σημειώστε ότι όλες οι σειρές αθροίζουν στο μηδέν. Το βασικό σημείο είναι ότι τα σημειακά γινόμενα των όρων σε οποιαδήποτε απο τις δύο σειρές, επίσης αθροίζουν στο μηδέν: έτσι για τους γραμμικούς και τετραγωνικούς όρους έχουμε γινόμενα των $(-3,\ 1,\ -1,\ 3)$, για γραμμικούς και κυβικούς όρους $(3,\ -3\ ,\ -3,\ 3)$ και για τετραγωνικούς και κυβικούς όρους $(1,\ -3,\ 3,\ 1)$. Στην R, οι ορθογώνιες πολυωνυμικές αντιθέσεις έχουν διαφορετικές αριθμητικές τιμές, αλλά τις ίδιες ιδιότητες:

```
t(contrasts(treatment))
```

```
           none        low     medium       high
.L   -0.6708204 -0.2236068  0.2236068  0.6708204
.Q    0.5000000 -0.5000000 -0.5000000  0.5000000
.C   -0.2236068  0.6708204 -0.6708204  0.2236068
```

Αν θέλατε να είστε ιδιαίτερα παράλογοι, θα μπορούσατε να ανακατασκευάσετε τις τέσσερις εκτιμώμενες μέσες τιμές από αυτές τις πολυωνυμικές αντιθέσεις και τα αποτελέσματα της επεξεργασίας θα φαίνοταν στο summary.lm

```
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    4.8125      0.1875   25.667  7.45e-12 ***
treatment.L    1.7330      0.3750    4.621  0.000589 ***
treatment.Q   -2.6250      0.3750   -7.000  1.43e-05 ***
treatment.C   -0.7267      0.3750   -1.938  0.076520   .
```

φροντίζοντας τα σύμβολα τόσο στις αντιθέσεις όσο και στους συντελεστές. Οι μέσες τιμές για none, low, medium και high είναι αντίστοιχα

```
4.8125 - 0.6708204*1.733 - 0.5*2.6250 + 0.2236068*0.7267
[1] 2.499963
4.8125 - 0.2236068*1.733+0.5*2.6250 - 0.6708204*0.7267
[1] 5.250004
4.8125 + 0.2236068*1.733 + 0.5*2.6250 + 0.6708204*0.7267
[1] 6.999996
4.8125 + 0.6708204*1.733 - 0.5*2.6250 - 0.2236068*0.7267
[1] 4.500037
```

σε συμφωνία (με 3 δεκαδικά ψηφία) με τις τέσσερις μέσες τιμές

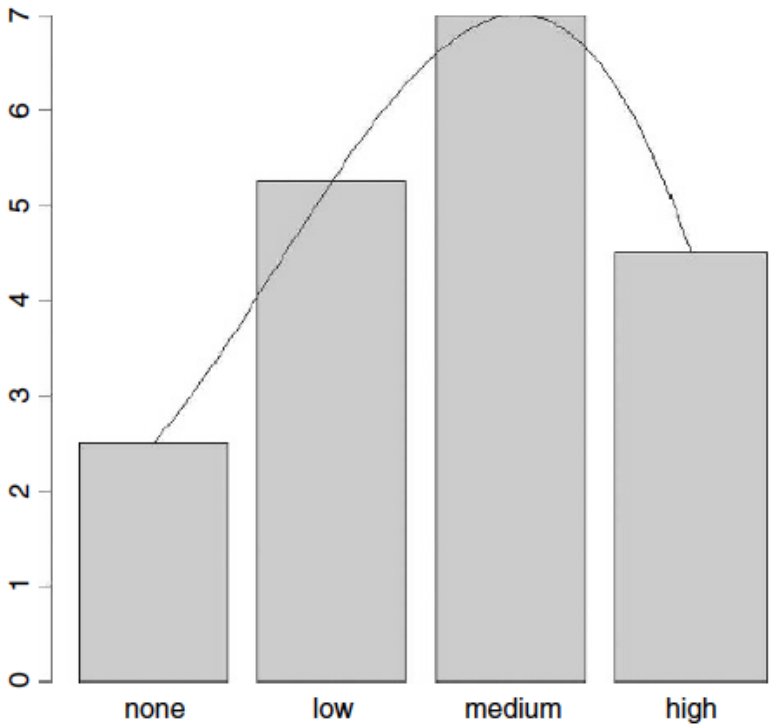
```
tapply(response,treatment,mean)
none   low  medium  high
2.50  5.25    7.00  4.50
```

Έτσι, οι παράμετροι μπορούν να ερμηνευθούν ως συντελεστές σε ένα πολυωνυμικό μοντέλο βαθμού 3 ($= k-1$, επειδή υπάρχουν $k = 4$ επίπεδα του παράγοντα που ονομάζονται επεξεργασία), αλλά μόνο εφ' όσον τα επίπεδα του παράγοντα ισαπέχουν (και δεν ξέρουμε αν αυτό είναι αλήθεια από τις πληροφορίες στο τρέχον πλαίσιο δεδομένων, γιατί γνωρίζουμε μόνο τη κατάταξη) και τα μεγέθη των τάξεων είναι ίσα (αυτό ισχύει στην παρούσα περίπτωση, όπου n = 4).

Επειδή έχουμε τέσσερα σημεία δεδομένων (η επεξεργασία μέσων τιμών) και τέσσερις παραμέτρους, η προσαρμογή του μοντέλου στα δεδομένα είναι τέλεια (δεν υπάρχουν υπόλοιπα βαθμών ελευθερίας και δεν υπάρχει ανεξήγητη διακύμανση). Μπορούμε να δούμε πώς η πολυωνυμική συνάρτηση μοιάζει σχεδιάζοντας την ομαλή καμπύλη στην κορυφή ενός ραβδογράμματος για τις μέσες τιμες:

```
y<-as.vector(tapply(response,treatment,mean))
model<-lm(y ~ poly(x,3))
model
```

```
Call:lm(formula = y~poly(x, 3))
Coefficients:
(Intercept)  poly(x, 3)1  poly(x, 3)2  poly(x, 3)3
     4.8125       1.7330      -2.6250      -0.7267
```



Τώρα μπορούμε να δημιουργήσουμε μια ομαλή σειρά x τιμών μεταξύ 1 και 4 από την οποία μπορούμε να προβλέψουμε την ομαλή πολυωνυμική συνάρτηση:

```
xv<-seq(1,4,0.1)
yv<-predict(model,list(x=xv))
```

Η μόνη μικρή δυσκολία είναι ότι οι τιμές του άξονα x στο ραβδόγραμμα δεν κλιμακώνονται ακριβώς μία-προς-μια με τις x τιμές μας, και πρέπει να ρυθμίσουμε τη x-θέση της ομαλής γραμμής μας από xv σε $xs = -0.5 + 1.2xv$. Οι παράμετροι -0.5 και 1.2 προέρχονται σημειώνοντας ότι είναι τα κέντρα των τεσσάρων ράβδων στα

```
(bar.x<-barplot(y))
```

```
     [,1]
[1,] 0.7
[2,] 1.9
[3,] 3.1
[4,] 4.3
```

```
barplot(y,names=levels(treatment))
xs<--0.5+1.2*xv
lines(xs,yv)
```