ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

Γενικευμένα Γραμμικά μοντέλα με χρήση της Γλώσσας προγραμματισμού R

Αθανάσιος Στεφ. Τάτσιος

ΕΡΓΑΣΙΑ
Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος
Ειδίκευσης στη Στατιστική
Μερικής Παρακολούθησης (Part-time)

Αθήνα
Οκτώβριος 2009

ΑΦΙΕΡΩΣΗ

Αφιερώνεται στη μνήμη του Πατέρα μου και στη Μητέρα μου.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω όλους τους καθηγητές και καθηγήτριες του μεταπτυχιακού προγράμματος Στατιστικής του Οικονομικού Πανεπιστημίου Αθηνών για τις ουσιαστικές γνώσεις που μας προσέφεραν.
Ιδιαίτερα ευχαριστώ την κυρία Δημάκη που με επέλεξε στο πρόγραμμα και τον κύριο Ντζιούφρα για τη βοήθεια που μου προσέφερε σαν επόπτης καθηγητής στη εργασία αυτή.
Τέλος ευχαριστώ τη Μαρρία που μου συμπαραστάθηκε όλο αυτό τον καιρό.

ΒΙΟΓΡΑΦΙΚΟ ΣΗΜΕΙΩΜΑ

Γεννήθηκα στον Πειραιά. Απέκτησα το πρώτο πανεπιστημιακό μου πτυχίο «Μαθηματικού» από το Καποδιστ ριακό Πανεπιστήμιο Αθηνών. Στη συνέχεια έκανα μεταπτυχιακές σπουδές στην Πληροφορική στο Τεχνικό Πανεπιστήμιο του Βερολίνου.

Το 2000 συμμετείχα σε ένα πρόγραμμα μεταπτυχιακών σπουδών που έγινε από το τμήμα μαθηματικών του ΤΕΙ Πειραιά σε συνεργασία με ένα Αγγλικό Πανεπιστήμιο με τίτλο «Πληροφορική και Διοίκηση».

Γνωρίζω να γράφω και να μιλάω πολύ καλά Αγγλικά και Γερμανικά.

Τα ενδιαφέροντα μου είναι αρκετά. Από τη λογοτεχνία μέχρι τον Αθλητισμό και από την μουσική μέχρι τον ιστορικό τουρισμό.

ABSTRACT

Athanasios Tatsios

Generalized Linear Models Using Programming Language R

October 2009

This paper is divided into two parts. The first part concerns the software package R and the second part is an introduction to basic theory of Generalized Linear Models using the programming language R.

The first part concerns the structure and the objects of the language R. We describe next the commands that are most frequently used in statistical inference in the context of the generalized linear models.

The next chapters concern the commands on statistical modeling with emphasis on commands relating to analysis of variance, linear and generalized linear models.

The second part concern implementation with R issues of the Generalized Linear Models that follow.

The next seven chapters concern the concepts and distributions used in generalized linear models. Then follows the estimation methods used in the context of generalized linear models. The next chapters concern the multiple regression and analysis of variance - covariance.

Finally, the last two chapters concern categorical data. We describe some methods using generalized linear models. We focus particularly on binary variables and logistic regression as well as contingency tables and Log Linear models.

ΠΕΡΙΛΗΨΗ

Αθανάσιος Τάτσιος

Γενικευμένα Γραμμικά Μοντέλα με χρήση της Γλώσσας Προγραμματισμού R

Οκτώβριος 2009

Η εργασία αυτή χωρίζεται σε δύο μέρη. Στο πρώτο μέρος πραγματευόμαστε το πακέτο λογισμικού R και στο δεύτερο μέρος γίνεται μία εισαγωγή στη βασική θεωρία των Γενικευμένων Γραμμικών Μοντέλων με χρήση της γλώσσας προγραμματισμού R.

Στο πρώτο μέρος γίνεται μία γενική εισαγωγή στη δομή της γλώσσας R και στα αντικείμενα που διαχειρίζεται. Στη συνέχεια περιγράφουμε τις εντολές που χρησιμοποιούνται πιο συχνά στη στατιστική συμπερασματολογία στα πλαίσια των Γενικευμένων Γραμμικών Μοντέλων.

Το επόμενο θέμα που περιγράφουμε είναι οι εντολές που αφορούν στατιστική μοντελοποίηση με έμφαση στις εντολές που αφορούν την ανάλυση διακύμανσης και τα γραμμικά και γενικευμένα γραμμικά μοντέλα

Στο δεύτερο μέρος γίνεται χρήση της R για την υλοποίηση των θεμάτων που ακολουθούν.

Ξεκινάμε με μία εισαγωγή στις έννοιες και κατανομές που χρησιμοποιούνται στα Γενικευμένα Γραμμικά Μοντέλα. Στη συνέχεια αναφερόμαστε στις μεθόδους εκτίμησης που χρησιμοποιούμε στα πλαίσια των Γενικευμένων γραμμικών μοντέλων. Το επόμενο θέμα είναι η πολλαπλή παλινδρόμηση και η ανάλυση διακύμανσης – συνδιακύμανσης.

Τέλος περιγράφουμε κάποιες μεθόδους από την αντιμετώπιση των κατηγορικών δεδομένων με τη βοήθεια Γενικευμένων Γραμμικών Μοντέλων. Εστιάζουμε ιδιαίτερα στις δίτιμες μεταβλητές και τη Λογιστική παλινδρόμηση καθώς επίσης στους πίνακες συνάφειας και τα Λογαριθμογραμμικά μοντέλα.

ΚΑΤΑΛΟΓΟΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΓΡΑΦΗΜΑΤΩΝ

1 ΕΙΣΑΓΩΓΗ

Η συνεχής ανάπτυξη της στατιστικής σε πεδία όπως η μοντελοποίηση, απαιτεί την αντιμετώπιση του προβλήματος του υπολογισμού διαφόρων παραμέτρων και στατιστικών στοιχείων. Ο όγκος όμως των απαιτούμενων υπολογισμών είναι τέτοιος που μόνο η χρήση υπολογιστή μπορεί να εξυπηρετήσει.

Σήμερα υπάρχουν αρκετά πακέτα λογισμικού τα οποία μπορούν και ανταποκρίνονται στις περισσότερες των απαιτήσεων που μπορεί να έχει μία εργασία που απαιτεί στατιστικούς υπολογισμούς.
Τα πακέτα αυτά προσφέρουν συνήθως ένα σύνολο τυποποιημένων εργασιών που μπορούν και υλοποιούν. Υπάρχουν λοιπόν περιπτώσεις που ένα τέτοιο πακέτο δεν μπορεί να ανταποκριθεί σε κάποια συγκεκριμένη εργασία.

Για να μπορεί ένα στατιστικό πακέτο να ανταποκριθεί στην υλοποίηση οποιονδήποτε εφαρμογών πρέπει να είναι αρκετά ευέλικτο και να μην περιορίζεται από ένα σύνολο σταθερών επιλογών που βρίσκονται σε ένα πίνακα επιλογών.

Το στατιστικό πακέτο R περιέχει υλοποιημένες τις περισσότερες μεθόδους και υπολογισμούς που απαιτούνται στη στατιστική υπό μορφή εντολών. Μας δίνει επίσης τη δυνατότητα, γράφοντας απλό κώδικα, να υλοποιήσουμε οποιονδήποτε εφικτό στατιστικό υπολογισμό επιθυμούμε. Το R διατίθεται ελεύθερα για χρήση και υποστήριξη μέσα από το Internet.

Τα γενικευμένα γραμμικά μοντέλα είναι ένας τομέας της στατιστικής που χρησιμοποιεί πολλές από τις μεθόδους της στατιστικής συμπερασματολογίας καθώς και επαναληπτικές υπολογιστικές τεχνικές. Το πρόβλημα που δημιουργούν τα γενικευμένα γραμμικά μοντέλα είναι ο όγκος και η πολυπλοκότητα των υπολογισμών. Από τη στιγμή που ο υπολογιστής αναλαμβάνει τους υπολογισμούς με ένα εργαλείο όπως το R τότε μπορούμε να χρησιμοποιήσουμε τις μεθόδους που επιθυμούμε και να προχωρήσουμε στις επεκτάσεις των γενικευμένων γραμμικών μοντέλων , δημιουργώντας νέες προσεγγιστικές μεθόδους υπολογισμού.

Η εργασία που ακολουθεί είναι χωρισμένη σε δύο μέρη. Το πρώτο μέρος είναι μία βασική εισαγωγή στο R με παραδείγματα, που εστιάζεται στις εντολές για γενικευμένα γραμμικά μοντέλα. Στο δεύτερο μέρος αναπτύσσεται η βασική θεωρία των γενικευμένων γραμμικών μοντέλων και γίνεται χρήση της R στα παραδείγματα που ακολουθούν τη θεωρία.

ΜΕΡΟΣ ΠΡΩΤΟ - ΣΤΟΙΧΕΙΑ της R

2 Εισαγωγή στην R

2.1 Τι είναι η R

Το R είναι ένα ολοκληρωμένο περιβάλλον εργασίας για στατιστικούς υπολογισμούς και γραφήματα. Το R μας εφοδιάζει μεταξύ των άλλων, με :

- Αποτελεσματικό χειρισμό και αποθήκευση δεδομένων .
- Χειρισμό πινάκων πολλών διαστάσεων.
- Μία απλή και αποτελεσματική γλώσσα προγραμματισμού (R), με διεπαφές με άλλες γλώσσες και δυνατότητες αποσφαλμάτωσης .
- Εργαλεία ανάλυσης δεδομένων και δημιουργίας γραφημάτων .

<u>*Η γλώσσα προγραμματισμού R έχει τα εξής χαρακτηριστικά :*</u>

Η R χρησιμοποιεί διερμηνευτή (interpreter) και όχι μεταφραστή (compiler).

Η R είναι “case sensitive” δηλαδή κάνει διαχωρισμό μεταξύ μικρών και κεφαλαίων χαρακτήρων.

Η R είναι μία διάλεκτος της S, η οποία χρησιμοποιείται ευρέως στη στατιστική κοινότητα.

Η σύνταξη της γλώσσας έχει μία επιφανειακή σχέση με την C, αλλά η σημασιολογία της είναι της FPL (Functional Programming Language = Συναρτησιακής Γλώσσας Προγραμματισμού), με μεγάλη συγγένεια με τις γλώσσες Lisp και APL.

Η γλώσσα R είναι ελεύθερα διαθέσιμη στο διαδίκτυο και η υποστήριξή της γίνεται με εθελοντική συνεισφορά.

Υπάρχουν πολλά διαθέσιμα πακέτα που υποστηρίζονται από την R και τα οποία υλοποιούν πολλές κλασσικές και μοντέρνες στατιστικές μεθόδους . <u>Πληροφορίες</u> στο : http://CRAN.R-project.org

Η R μπορεί να τρέξει σε διαφορετικές πλατφόρμες (πχ. Windows, Linux, Mac OS).
Η R έχει την ιδιότητα να επιτρέπει τον «υπολογισμό στη γλώσσα» , δηλαδή παρέχει την δυνατότητα να γράφουμε συναρτήσεις οι οποίες δέχονται εκφράσεις σαν είσοδο πράγμα πολύ χρήσιμο για στα τιστικά μοντέλα.

Περισσότερες πληροφορίες στην ιστοσελίδα : http://www.r-project.org

<u>Τρόπος χρήσης της R</u>

Η R λειτουργεί διαλογικά. Γράφουμε μετά το σήμα ετοιμότητας (Το σήμα ετοιμότητας στα Windows είναι το “>”, ενώ στο Unix “$”) την εντολή και πατάμε Enter. Μετά το πάτημα του πλήκτρου Enter ο κώδικας απασφαλματώνεται και εκτελείται και εμφανίζεται στη συνέχεια η απάντηση.
Σε περίπτωση που ο κώδικας είναι πάνω από μία γραμμή , μετά τη πρώτη γραμμή αλλάζει από “>” στο σύμβολο “+” για τις επόμενες γραμμές.

Η R μπορεί να χρησιμοποιηθεί με δύο τρόπους. Ο πρώτος τρόπος χρήσης είναι να εκτελούνται οι εκφράσεις από τη γραμμή εντολών (πολλοί χρήστες πιθανόν να μην απαιτούν κάτι παραπάνω από αυτό το επίπεδο) , ενώ ο δεύτερος τρόπος είναι μία μέθοδος κατά την οποία γράφουμε με κώδικα συναρτήσεις του χρήστη.

<u>Διαφορά της R με άλλα στατιστικά πακέτα.</u>

Η R λειτουργεί στη στατιστική ανάλυση με τη μορφή σειράς βημάτων. Σε κάθε βήμα αποθηκεύει τα αποτελέσματα σε αντικείμενα για περαιτέρω ανάλυση, ενώ τα άλλα στατιστικά πακέτα, όπως το SPSS, δίνουν ένα συνολικό τελικό αποτέλεσμα στην έξοδο.

2.2 Αντικείμενα στην R

Σε κάθε γλώσσα προγραμματισμού έχουμε τις μεταβλητές. Οι μεταβλητές είναι ένας τρόπος για να προσπελάσουμε δεδομένα που έχουν αποθηκευτεί στη μνήμη.

Οι οντότητες που δημιουργεί και χειρίζεται η R είναι γνωστές σαν «αντικείμενα». Τα «*αντικείμενα*» μπορεί να είναι μεταβλητές, διανύσματα, πίνακες, συναρτήσεις, strings (σειρές χαρακτήρων ή αλφαριθμητικά) , σύμβολα που παριστάνουν με τη σειρά τους αντικείμενα ή γενικότερα οποιεσδήποτε δομές.
Η R μας εφοδιάζει με ένα λεξικό μοντέλο όμοιο με της Pascal για να μπορούμε να χρησιμοποιούμε τα σύμβολα με ανάλογο τρόπο.

Παρατήρηση

Η ειδική συνάρτηση typeof μας επιστρέφει τον τύπο ενός αντικειμένου της R.

Ο παρακάτω πίνακας δείχνει τις πιο συνηθισμένες τιμές που επιστρέφονται από τη συνάρτηση *typeof*.

"NULL"	Δεν υπάρχουν στοιχεία (παριστάνει το «τίποτα») .
"symbol"	όνομα μεταβλητής.
"closure"	μία συνάρτηση.
"environment"	ένα περιβάλλον.
"language"	ένα κατασκεύασμα της γλώσσας R.
"char"	σειρά χαρακτήρων (string) (μόνο εσωτερικά) ***.
"logical"	διάνυσμα που περιέχει λογικές τιμές.
"integer"	διάνυσμα που περιέχει ακέραιες τιμές.
"double"	διάνυσμα που περιέχει πραγματικές τιμές.
"complex"	διάνυσμα που περιέχει μιγαδικές τιμές.
"character"	διάνυσμα που περιέχει τιμές χαρακτήρων.
"..."	ειδικό όρισμα . Εάν μία συνάρτηση έχει «...» σαν τυπικό

	όρισμα , τότε μπορεί να χρησιμοποιηθεί για να περάσει οποιοδήποτε όρισμα που δεν εκφράζεται με κάποιο τυπικό όρισμα.
"any"	ειδικός τύπος που ταιριάζει σε όλους τους τύπους.
"expression"	ένα αντικείμενο έκφρασης.
"list"	μία λίστα.
"raw"	διάνυσμα που περιέχει bytes.

Οι χρήστες δεν μπορούν εύκολα να χειριστούν αντικείμενα σημειωμένα με '***'.

Η συνάρτηση *storage.mode* επιστρέφει το *τύπο* με το οποίο έχει αποθηκευτεί το αντικείμενο και είναι χρήσιμο στην περίπτωση που χρησιμοποιούνται συναρτήσεις γραμμένες σε άλλη γλώσσα όπως C ή Fortran, για να εξασφαλίσει ότι τα αντικείμενα της R έχουν τον τύπο δεδομένων που περιμένει η καλούσα συνάρτηση.

Κατά τη διάρκεια των υπολογισμών τα αντικείμενα της R εξαναγκάζονται να μετατραπούν σε άλλους τύπους. Υπάρχουν πολλές συναρτήσεις διαθέσιμες για να κάνουν σαφή μετατροπή τύπων.

3 Βασικοί τύποι Δεδομένων στην R

Τα αντικείμενα δεδομένων τα οποία έχουν δεδομένα μόνο ενός τύπου ονομάζονται στην R ατομικά (atomic). Ατομικά είναι τα διανύσματα, οι πίνακες και τα Arrays

3.1 Διανύσματα

Τα διανύσματα είναι διατεταγμένα σύνολα δεδομένων. Τα δεδομένα ενό ς διανύσματος είναι υποχρεωτικά του ίδιου τύπου.

Η R έχει έξη βασικούς τύπους διανυσμάτων : logical, integer, real (double), complex, string (or character) and raw. Παρακάτω δίνονται τα typeof και storage.mode των διαφόρων τύπων διανυσμάτων.

typeof	*storage.mode*
logical	logical
integer	integer
double	double
complex	complex
character	character
raw	raw

Απλοί αριθμοί όπως το 4, 2 είναι διάνυσμα μήκους 1 (ένα). Υπάρχει δυνατότητα να υπάρχουν διανύσματα μήκους μηδέν.

Ο συμβολισμός n1:n2 παριστάνει τους αριθμούς από n1 μέχρι n2.
Πχ.

```
> 5:9
[1] 5 6 7 8 9
```

<u>Διανύσματα αριθμών</u>

συνάρτηση c() : Δημιουργεί ένα διάνυσμα αριθμών.

Πχ.

```
> a<-c(10, 7, 5, 67, 0, 1)
> a
[1] 10  7  5 67  0  1
> a<-c(1:10)
> a
 [1]  1  2  3  4  5  6  7  8  9 10
```

Τα στοιχεία διανύσματος μπορεί να προσπελαστούν με λειτουργίες δεικτών (πχ. αν a είναι διάνυσμα, με a[3] προσπελαύνουμε το 3^{ο} στοιχείο του διανύσματος).

```
> a[3]
[1] 5
```

Η συνάρτηση rep(ορισμα, φορες) καθορίζει πόσες **φορές** θα επαναληφθεί το **όρισμα.**

Πχ.

```
> x<-1:3
> x
[1] 1 2 3
> rep(x, 5)
 [1] 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3
```

Η συνάρτηση seq() : Για να δημιουργηθούν διανύσματα από αριθμούς μπορεί να χρησιμοποιηθεί η συνάρτηση ακολουθίας ***seq()***. Η seq έχει πέντε ορίσματα που είναι **from** (από), **to** (μέχρι), **by** (βήμα αύξησης), **length** (μήκος).

Πχ.

```
> s1<-seq(-5, 5, by=0.2)
> s1
```

```
 [1] -5.0 -4.8 -4.6 -4.4 -4.2 -4.0 -3.8 -3.6 -3.4 -3.2 -3.0 -2.8 -2.6 -2.4 -2.2
[16] -2.0 -1.8 -1.6 -1.4 -1.2 -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8
[31]  1.0  1.2  1.4  1.6  1.8  2.0  2.2  2.4  2.6  2.8  3.0  3.2  3.4  3.6  3.8
[46]  4.0  4.2  4.4  4.6  4.8  5.0

> s2<-seq(-5, 5)
> s2
 [1] -5 -4 -3 -2 -1  0  1  2  3  4  5

> s3<-seq(length=51, from=-5, by=0.5)
> s3
 [1] -5.0 -4.5 -4.0 -3.5 -3.0 -2.5 -2.0 -1.5 -1.0 -0.5  0.0  0.5  1.0  1.5  2.0
[16]  2.5  3.0  3.5  4.0  4.5  5.0  5.5  6.0  6.5  7.0  7.5  8.0  8.5  9.0  9.5
[31] 10.0 10.5 11.0 11.5 12.0 12.5 13.0 13.5 14.0 14.5 15.0 15.5 16.0 16.5
17.0
[46] 17.5 18.0 18.5 19.0 19.5 20.0
```

Διανύσματα χαρακτήρων

Τα διανύσματα χαρακτήρων μπορεί να δημιουργηθούν *και* με τη συνάρτηση **paste().** Η συνάρτηση αυτή παίρνει ένα αυθαίρετο αριθμό ορισμάτων και τα ενώνει ένα - ένα. Το προαιρετικό όρισμα **sep**=χαρακτήρας δηλώνει το χαρακτήρα που παρεμβάλλεται στην ένωση.

Πχ.

```
> temp<-paste("X", 1:10)
> temp
 [1] "X 1"  "X 2"  "X 3"  "X 4"  "X 5"  "X 6"  "X 7"  "X 8"  "X 9"  "X 10"
> cnt<-paste(c("X", "Y"), 1:10, sep="")
> cnt
 [1] "X1"  "Y2"  "X3"  "Y4"  "X5"  "Y6"  "X7"  "Y8"  "X9"  "Y10"
```

3.1.1 Πράξεις με διανύσματα

Η R έχει να κάνει με ολόκληρα διανύσματα δεδομένων κάθε φορά και οι περισσότερες πράξεις είναι διανυσματικές (δηλ στοιχείο προς στοιχείο). Αυτό σημαίνει για παράδειγμα ότι προσθέτοντας δύο διανύσματα του ίδιου μήκους θα προκύψει ένα διάνυσμα που θα περιέχει τα αθροίσματα των αντιστοίχων στοιχείων.

Τελεστές της R

-	Αφαίρεση και μοναδιαίος τελεστής.

+	Πρόσθεση και μοναδιαίος τελεστής.

!	Μοναδιαία άρνηση (not).

~	Περισπωμένη, χρησιμοποιείται στην δημιουργία τύπων στα μοντέλα.

?	Help (Βοήθεια).

:	Ακολουθία, binary (σε τύπους μοντέλων: αλληλεπίδραση).

*	Πολλαπλασιασμός.

/	Διαίρεση.

^	Δύναμη.

%x%	Ειδικοί δυαδικοί τελεστές., το x είναι σύμβολο.

%%	Υπόλοιπο διαίρεσης.

%/%	Ακέραια διαίρεση.

%*%	Γινόμενο πινάκων.

%o%	Εξωτερικό γινόμενο.

%x%	Kronecker γινόμενο.

%in%	Τελεστής ταιριάσματος (σε μοντέλα : φώλιασμα).

<	Μικρότερο.

>	Μεγαλύτερο.

==	Ίσο.

>=	Μεγαλύτερο ίσο.

<=	Μικρότερο ίσο.

&	Διανυσματικό Και (And).

&&	Και (And).

|	Διανυσματικο ή (Or).

||	‘Η (Or).
<-	Αριστερή ανάθεση.
->	Δεξιά ανάθεση.
$	Υποσύνολο λίστας.

Παρατηρήσεις :

Ο διανυσματικός πολλαπλασιασμός δύο πινάκων γίνεται με το σύμβολο «*» ενώ ο κλασσικός πολλαπλασιασμός πινάκων γίνεται με τον τελεστή «%*%» .

Εάν τα διανύσματα που μετέχουν σε μία διανυσματική πράξη δεν έχουν το ίδιο μήκος, τότε το διάνυσμα μικρότερου μήκους ανακυκλώνεται (επαναλαμβάνονται από την αρχή οι τιμές του).

Σε μερικές περιπτώσεις οι συνιστώσες ενός διανύσματος μπορεί να μην είναι τελείως γνωστές. Τότε έχουμε μία τιμή «Μη διαθέσιμη» (“not available” NA ή “missing value”).

Η συνάρτηση ελέγχου για μη διαθέσιμες τιμές είναι : **is.na(x)** όπου το **x** είναι το διάνυσμα.

Αριθμητικοί υπολογισμοί παράγουν κάποιες φορές απροσδιόριστα αποτελέσματα (πχ. 0/0). Τα απροσδιόριστα αποτελέσματα έχουν τη τιμή NAN. Για να ελέγξουμε την περίπτωση αυτή χρησιμοποιούμε τη συνάρτηση **is.nan**.

Υποσύνολα των διανυμάτων με διανύσματα δείκτες .

Μπορούμε να πάρουμε υποσύνολα ενός διανύσματος χρησιμοποιόντας διανύσματα που παίζουν το ρόλο δεικτών. Τα παρακάτω παραδείγματα εξηγούν τη μέθοδο.

Παράδειγματα

```
> x<-c(2, -13, 7, 0, 7, 32, 1, -6)
> x[2:5]
```

```
[1] -13   7   0   7
> x[-(2:5)]
[1]  2 32  1 -6
> y<-c(45, 7)
> x[y]
[1] NA  1
> y<-c(1, 3, 5, 9)
> x[y]
[1]  2  7  7 NA
```

<u>Παράδειγμα</u> με δείκτες που είναι λογικά διανύσματα

```
> y<-c(1, 3, 5, 9)
> z<-x[y]
> z
[1]  2  7  7 NA
> a<-z[!is.na(z)]
> a
[1] 2 7 7
> a<-(z+1)[!is.na(z) & z>3]
> a
[1] 8 8
```

3.2 Πίνακες και Arrays

Ένας πίνακας στην R είναι μία ορθογώνια διάταξη στοιχείων ίδιου τύπου σε δύο διαστάσεις (γραμμές και στήλες).

Στην R η γενίκευση των πινάκων σε περισσότερες διαστάσεις μας δίνει την δομή του Array (μπορεί να ονομαστεί πίνακας με διάσταση πάνω από 2).

Μπορούμε να δημιουργήσουμε ένα πίνακα από ένα διάνυσμα αποδίδοντας στο διάνυσμα διάσταση με την εντολή dim. Η διάσταση είναι με τη σειρά της ένα διάνυσμα (εδώ με δύο αριθμούς που παριστάνουν γραμμές και στήλες).

Πχ.

```
> tmp<-c(1:12)
> tmp
 [1]  1  2  3  4  5  6  7  8  9 10 11 12
> dim(tmp)<-c(3, 4)
> tmp
     [, 1] [, 2] [, 3] [, 4]
[1, ]    1    4    7   10
[2, ]    2    5    8   11
[3, ]    3    6    9   12
```

Δημιουργία πίνακα συνδέοντας διανύσματα σε γραμμές με τη συνάρτηση **rbind()** ή στήλες με την **cbind()**

πχ.

```
> pnx<-rbind(c(-1, 0, 2, 3), c(1, 0, 1, 0), c(1, 1, 2, 1))
> pnx
     [, 1] [, 2] [, 3] [, 4]
[1, ]   -1    0    2    3
[2, ]    1    0    1    0
[3, ]    1    1    2    1
> pnx1<-cbind(c(1, 2, 1), c(21, 90, 5))
> pnx1
     [, 1] [, 2]
[1, ]    1   21
[2, ]    2   90
[3, ]    1    5
```

Παρατήρηση : Σε περίπτωση διανυσμάτων με διαφορετικά μεγέθη έχουμε ανακύκλωση.

<u>Δημιουργία πίνακα με τη εντολή</u> **<u>matrix()</u>** όπου έχουμε σαν ορίσματα τα δεδομένα, τον αριθμό γραμμών και τον αριθμό στηλών.

Πχ.

```
> pnx2<-matrix(1:35, nrow=7, ncol=5)
> pnx2
     [, 1] [, 2] [, 3] [, 4] [, 5]
[1, ]    1    8   15   22   29
[2, ]    2    9   16   23   30
[3, ]    3   10   17   24   31
[4, ]    4   11   18   25   32
[5, ]    5   12   19   26   33
[6, ]    6   13   20   27   34
[7, ]    7   14   21   28   35
```

<u>Εντολές Δημιουργίας Array</u>

Για να δημιουργηθεί ένα Array χρησιμοποιείται η εντολή :

array(διάνυσμα_δεδομένων, dim=διάνυσμα_διαστάσεων)

Πχ.

```
> ar1<-array(1:12, dim=c(2, 3, 4))
> ar1
, , 1
     [, 1] [, 2] [, 3]
[1, ]    1    3    5
[2, ]    2    4    6

, , 2
     [, 1] [, 2] [, 3]
[1, ]    7    9   11
[2, ]    8   10   12
```

```
, ,  3
     [, 1] [, 2] [, 3]
[1, ]    1    3    5
[2, ]    2    4    6

, ,  4
     [, 1] [, 2] [, 3]
[1, ]    7    9   11
[2, ]    8   10   12
```

Πρώτα συμπληρώνεται η πρώτη διάσταση (δηλ. οι γραμμές) , τοποθετώντας τα στοιχεία ανά στήλες, οπότε δημιουργούνται και οι γραμμές. Τέλος για κάθε επίπεδο της τρίτης διάστασης δημιουργείται ένας πίνακας.

Δημιουργία Array με άλλο τρόπο

```
> z<-1:24
> dim(z)<-c(3, 4, 2)
> z
, ,  1
     [, 1] [, 2] [, 3] [, 4]
[1, ]    1    4    7   10
[2, ]    2    5    8   11
[3, ]    3    6    9   12

, ,  2
     [, 1] [, 2] [, 3] [, 4]
[1, ]   13   16   19   22
[2, ]   14   17   20   23
[3, ]   15   18   21   24
```

<u>Εντολές :</u> **length, dim**

Η εντολή **length** δίνει το πλήθος στοιχείων του Array ενώ η εντολή **dim** τις διαστάσεις του Array

```
> length(ar1)
[1] 24
> dim(ar1)
[1] 2 3 4
```

3.2.1 Πράξεις με πίνακες

Εκτός από τις πράξεις στοιχείο με στοιχείο σε πίνακες με ίσες διαστάσεις , έχουμε και το γινόμενο πινάκων που γίνεται με τον τελεστή %*% .

Πχ.

```
> a<-matrix(10:21, nrow=3, ncol=4)
> a
     [, 1] [, 2] [, 3] [, 4]
[1, ]   10   13   16   19
[2, ]   11   14   17   20
[3, ]   12   15   18   21
> b<-matrix(25:36, nrow=3, ncol=4)
> b
     [, 1] [, 2] [, 3] [, 4]
[1, ]   25   28   31   34
[2, ]   26   29   32   35
[3, ]   27   30   33   36
> a%*%b
Error in a %*% b : non-conformable arguments
```

(Δίνει μήνυμα λάθους διότι οι διαστάσεις δεν ικανοποιούν την συνθήκη που απαιτείται για τον πολλαπλασιασμό πινάκων).

```
> a*b
     [, 1] [, 2] [, 3] [, 4]
[1, ]  250  364  496  646
[2, ]  286  406  544  700
[3, ]  324  450  594  756
> c<-matrix(25:36, nrow=4, ncol=3)
> a%*%c
     [, 1] [, 2] [, 3]
[1, ] 1552 1784 2016
[2, ] 1658 1906 2154
[3, ] 1764 2028 2292
```

<u>Εντολή :</u> diag(x)

Εάν το**x** είναι πίνακας τότε το αποτέλεσμα δίνει ένα διάνυσμα με τα στοιχεία της διαγώνιου.

Εάν το **x** είναι διάνυσμα τότε το αποτέλεσμα είναι πίνακας με διαγώνια στοιχεία τα στοιχεία του διανύσματος.

Πχ

```
> a%*%c
     [, 1] [, 2] [, 3]
[1, ] 1552 1784 2016
[2, ] 1658 1906 2154
[3, ] 1764 2028 2292
> diag(a%*%c)
[1] 1552 1906 2292
> v<-2:6
> diag(v)
     [, 1] [, 2] [, 3] [, 4] [, 5]
[1, ]    2    0    0    0    0
[2, ]    0    3    0    0    0
[3, ]    0    0    4    0    0
[4, ]    0    0    0    5    0
[5, ]    0    0    0    0    6
```

Ανάστροφος πίνακας

Ο ανάστροφος πίνακας ενός πίνακα A δίνεται με την εντολή : t(A)

Πχ.

```
> A
    [, 1] [, 2] [, 3] [, 4]
[1, ]  25  28  31  34
[2, ]  26  29  32  35
[3, ]  27  30  33  36
> t(A)
    [, 1] [, 2] [, 3]
[1, ]  25  26  27
[2, ]  28  29  30
[3, ]  31  32  33
[4, ]  34  35  36
```

Αντίστροφος Πίνακας

Εάν υπάρχει ο αντίστροφος πίνακας A^{-1} ενός πίνακα A, τότε αυτός υπολογίζεται από την εντολή : **solve(A)**

Πχ.

```
> A<-c(2, 7, 8, 3, -9, -2, 3, 1, 7)
> dim(A)<-c(3, 3)
> solve(A)
          [, 1]      [, 2]      [, 3]
[1, ]  0.8591549  0.3802817 -0.4225352
[2, ]  0.5774648  0.1408451 -0.2676056
[3, ] -0.8169014 -0.3943662  0.5492958
```

Λύση γραμμικών συστημάτων

Ένα γραμμικό σύστημα μπορεί υπό μορφή πινάκων να γραφτεί στη μορφή :

Ax=b

Όπου **A** είναι ο πίνακας των συντελεστών, **x** το διάνυσμα των αγνώστων και **b** το διάνυσμα των σταθερών όρων.
Το σύστημα μπορεί (υπό προϋποθέσεις) να γραφτεί στη μορφή : $x=A^{-1}b$ που αποτελεί και τη λύση του συστήματος.

Στην **R** το διάνυσμα **x** δίνεται από την εντολή **solve(A,b)**.

Πχ.
Έστω ότι έχουμε να επιλύσουμε το παρακάτω σύστημα.

$$\left.\begin{aligned} 5x-3y &= 12 \\ x+2y &= 4 \end{aligned}\right\}$$

Το σύστημα γράφεται υπό μορφή πινάκων (αρκεί να υπάρχει ο αντίστροφος πίνακας) :

$$\begin{pmatrix} 5 & -3 \\ 1 & 2 \end{pmatrix} x = \begin{pmatrix} 12 \\ 4 \end{pmatrix} \Rightarrow x = \begin{pmatrix} 5 & -3 \\ 1 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 12 \\ 4 \end{pmatrix}$$

<u>Οπότε στην R έχουμε</u>

```
> A<-rbind(c(5, -3), c(1, 2))
> A
     [, 1] [, 2]
[1, ]    5   -3
[2, ]    1    2
> b<-c(12, 4)
> solve(A, b)
[1] 2.7692308  0.6153846
```

3.3 Λίστες

Οι Λίστες γενικεύουν τη δομή των διανυσμάτων.Τα στοιχεία μίας λίστας μπορεί να είναι διατεταγμένα στοιχεία οποιουδήποτε τύπου.

Τα στοιχεία μιας λίστας ονομάζονται συνιστώσες της λίστας.

Μία λίστα μπορεί να δημιουργηθεί
με την εντολή : **list(comp1, comp2, …compn)**.

Εναλλακτικά μπορούμε να δώσουμε ονόματα στις συνιστώσες με την εντολή **list(name1=comp1, name2=comp2, …, name=compn).**

Οι συνιστώσες αριθμούνται και με τον τρόπο αυτό μπορούμε να αναφερθούμε σε αυτές χρησιμοποιώντας με τον το [[ ]].
Δηλαδή, εάν το όνομα της λίστας είναι alst τότε με alst[[2]] αναφερόμαστε στο δεύτερο στοιχείο της λίστας.
Εάν η λίστα έχει ονόματα στις συνιστώσες της τότε μπορούμε να έχουμε πρόσβαση στη συνιστώσα με όνομα name2 με τον τελεστή $. Δηλαδή alst$name2

Ισχύει ότι : alst [[2]] και alst$name2 παριστάνουν το ίδιο πράγμα.

Πχ.

```
> l1<-list(1:23, 21:25, "An Example")
> l1
[[1]]
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

[[2]]
[1] 21 22 23 24 25

[[3]]
[1] "An Example"

> l2<-list(name="Maria", married="Yes", chilren.ages=c(5, 7, 10))
> l2
$name
[1] "Maria"
```

```
$married
[1] "Yes"

$chilren.ages
[1]  5  7 10

> l2[[3]]
[1]  5  7 10
> l2$chilren.ages
[1]  5  7 10
```

Μπορούμε να επεκτείνουμε τη λίστα δίνοντας νέες συνιστώσες. Εδώ θα δώσουμε σαν νέα συνιστώσα στη λίστα l2 τη λίστα l1.

```
> l2[4]<-list(l1)
> l2
$name
[1] "Maria"
$married
[1] "Yes"
$chilren.ages
[1]  5  7 10
[[4]]
[[4]][[1]]
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
[[4]][[2]]
[1] 21 22 23 24 25

[[4]][[3]]
[1] "An Example"
```

3.4 Πλαίσια δεδομένων (Data Frames)

Τα Data Frames (πλαίσια δεδομένων) είναι οι δομές της R που βρίσκονται πιο κοντά στα σύνολα δεδομένων (data set) του SPSS και SAS. Είναι δηλαδή πίνακες δεδομένων που περιέχουν στήλες με διαφορετικούς τύπους δεδομένων (αλλά στην ίδια στήλη έχουμε τον ίδιο τύπο) και «περιπτώσεις ανά μεταβλητή».

Ένα πλαίσιο δεδομένων είναι μία λίστα από διανύσματα, παράγοντες ή άλλα πλαίσια που όλα έχουν το ίδιο μήκος (στη περίπτωση των πινάκων οι γραμμές έχουν το ίδιο μήκος).

Στο πλαίσιο δεδομένων η θεώρηση γίνεται ανά στήλη που είναι και οι μεταβλητές του πλαισίου δεδομένων.

Ένα πλαίσιο δεδομένων έχει την ιδιότητα **names** για να δώσει ετικέτες στις μεταβλητές και ιδιότητα **row.names** για να δώσει ετικέτες στις περιπτώσεις (στις γραμμές).

Εντολές δημιουργίας πλαισίου δεδομένων

Η δημιουργία ενός πλαισίου γίνεται με την εντολή **data.frame** όταν ικανοποιούνται οι περιορισμοί για τοποθέτηση σε στήλες.

Στην περίπτωση που τα στοιχεία μιας λίστας μπορεί να ικανοποιήσουν τους περιορισμούς των πλαισίων τότε η λίστα μετατρέπεται σε πλαίσιο με την εντολή **as.data.frame.**

Για την εισαγωγή στοιχείων από εξωτερικό αρχείο και την τοποθέτηση τους σε πλαίσιο χρησιμοποιείται η εντολή **read.data.**

<u>Παράδειγμα</u>

```
> a1<-matrix(1:30)
> dim(a1)=c(10, 3)
> a1
      [, 1] [, 2] [, 3]
 [1, ]    1   11   21
 [2, ]    2   12   22
 [3, ]    3   13   23
 [4, ]    4   14   24
 [5, ]    5   15   25
 [6, ]    6   16   26
 [7, ]    7   17   27
 [8, ]    8   18   28
 [9, ]    9   19   29
[10, ]   10   20   30
> a2<-c("Yes", "No")
> a22<-rep(a2, 5)
> a22
 [1] "Yes" "No"  "Yes" "No"  "Yes" "No"  "Yes" "No"  "Yes" "No"
>  a3<-c("Thin",  "Thin",  "Thin",  "Thick",  "Thin",  "Thick",  "Thick",  "Thin",
+"Thick", "Thick")
> plaisio<-data.frame(a1, a22, a3)
> plaisio
         X1      X2      X3      a22     a3
1   1    11      21      Yes     Thin
2   2    12      22      No      Thin
3   3    13      23      Yes     Thin
4   4    14      24      No      Thick
5   5    15      25      Yes     Thin
6   6    16      26      No      Thick
7   7    17      27      Yes     Thick
8   8    18      28      No      Thin
9   9    19      29      Yes     Thick
1010     20       30     No      Thick
```

Σημαντικές εντολές : attach(dataframe) και detach(dataframe).
Οι εντολές αυτές τοποθετούν το dataframe πρώτο στη λίστα αναζήτησης μεταβλητών και αντίστοιχα το καταργούν. Οι μεταβλητές του dataframe δεν επηρεάζονται από οποιαδήποτε αλλαγή διότι χρησιμοποιείται ένα αντίγραφό τους.

3.5 Παράγοντες (Factors)

Οι παράγοντες χρησιμοποιούνται για να περιγράψουν όρους που έχουν πεπερασμένο πλήθος τιμών. Χρησιμοποιείται για την ομαδοποίηση άλλων διανυσμάτων ίδιου μήκους. Οι κατηγορικές μεταβλητές παρουσιάζονται στην R με το αντικείμενο δεδομένων παράγοντα (factor).

Ένας παράγοντας μπορεί να παριστάνει καθαρά ονομαστικά δεδομένα, αλλά μπορεί να παριστάνει και διατεταγμένες κατηγορίες. Στην τελευταία περίπτωση πρέπει να οριστεί και να έχει διάνυσμα κλάσης c(“ordered”, ”factor”).
Οι παράγοντες εκτελούνται χρησιμοποιώντας ένα array ακεραίων για να καθορίσουν τα ενεργά επίπεδα και ένα δεύτερο array ονομάτων που απεικονίζονται στους ακέραιους.

Ένας παράγοντας έχει την ιδιότητα επίπεδα (*levels*) και την κλάση “factor”. Προαιρετικά έχει την ιδιότητα *contrasts* για να ελέγξει τη παραμετροποίηση , όταν χρησιμοποιείται μέσα σε συναρτήσεις μοντελοποίησης.

Παράδειγμα
Έστω ότι έχουμε ένα δείγμα από 30 φορολογικούς λογαριασμούς από διαφορετικές περιοχές της Αυστραλίας που περιγράφονται με τρία γράμματα.

```
> state <- c("tas", "sa", "qld", "nsw", "nsw", "nt", "wa", "wa",
"qld", "vic", "nsw", "vic", "qld", "qld", "sa", "tas",
"sa", "nt", "wa", "vic", "qld", "nsw", "nsw", "wa",
"sa", "act", "nsw", "vic", "vic", "act")
```

Έχουμε τις συναρτήσεις **factor** για τη δημιουργία παραγόντων και **levels** για τα επίπεδα παραγόντων που δημιουργούνται.

Πχ.

```
> statef <- factor(state)
> statef
[1] tas sa qld nsw nsw nt wa wa qld vic nsw vic qld qld sa
[16] tas sa nt wa vic qld nsw nsw wa sa act nsw vic vic act
Levels: act nsw nt qld sa tas vic wa
> levels(statef)
[1] "act" "nsw" "nt" "qld" "sa" "tas" "vic" "wa"
```

Η συνάρτηση : tapply()

Χρησιμοποιείται για να εφαρμόσει μία συνάρτηση σε κάθε ομάδα συνιστωσών που ορίζονται από τα διάφορα επίπεδα της ποιοτικής μεταβλητής.

Πχ.

Συνεχίζοντας το προηγούμενο παράδειγμα θα βρούμε τους μέσους ανά ομάδα στο προηγούμενο δείγμα.

Έστω incomes οι τιμές των εισοδημάτων. Η έτοιμη συνάρτηση **mean** μας δίνει το μέσο.

```
> incomes <- c(60, 49, 40, 61, 64, 60, 59, 54, 62, 69, 70, 42, 56,
+ 61, 61, 61, 58, 51, 48, 65, 49, 49, 41, 48, 52, 46,
+ 59, 46, 58, 43)
> incmeans <- tapply(incomes, statef, mean)
> incmeans
   act      nsw      nt      qld      sa      tas      vic
wa
44.50000 57.33333 55.50000 53.60000 55.00000 60.50000 56.00000
52.25000
```

<u>Η συνάρτηση :</u> cut() και η συνάρτηση table()

Μπορούμε να κατασκευάσουμε παράγοντες από μία συνεχή μεταβλητή χρησιμοποιώντας τη συνάρτηση **cut()**.
Η συνάρτηση cut() παίρνει σαν όρισμα ένα διάνυσμα τιμών και ένα διάνυσμα με σημεία κοπής.

Πχ.
Έστω ότι θέλω ένα διάνυσμα τιμών (για παράδειγμα incomes) να το μοιράσω με σημεία κοπής 35, 45, 55, 65, 75 τότε έχουμε :

```
> facts<-cut(incomes, breaks=c(35, 45, 55, 65, 75))
> facts
 [1] (55, 65] (45, 55] (35, 45] (55, 65] (55, 65] (55, 65] (55, 65] (45, 55] (55, 65]
[10] (65, 75] (65, 75] (35, 45] (55, 65] (55, 65] (55, 65] (55, 65] (55, 65] (45, 55]
[19] (45, 55] (55, 65] (45, 55] (45, 55] (35, 45] (45, 55] (45, 55] (45, 55] (55, 65]
[28] (45, 55] (55, 65] (35, 45]

Levels: (35, 45] (45, 55] (55, 65] (65, 75]
```

Με τη βοήθεια της συνάρτησης **table()** υπολογίζουμε ένα πίνακα συχνοτήτων δύο εισόδων.

```
>table(facts, statef)
       statef
facts     act nsw nt qld sa tas vic wa
  (35, 45]   1   1  0   1  0   0   1  0
  (45, 55]   1   1  1   1  2   0   1  3
  (55, 65]   0   3  1   3  2   2   2  1
  (65, 75]   0   1  0   0  0   0   1  0
```

3.6 Δεικτοδότηση

Για να χρησιμοποιήσουμε στοιχεία από διάφορα σύνολα δεδομένων χρησιμοποιούμε δείχτες.

3.6.1 Απόδοση δεικτών σε Λίστες

Τα σύμβολα x[[i]], x[[i, j]], xa, x”a” χρησιμοποιούνται σε λίστες όπου στην περίπτωση του x[[i]] μπορούμε να κάνουμε πράξεις με δείκτες, ενώ με το συμβολισμό xa, x”a” έχουμε a που είναι μόνο κείμενο.

3.6.2 Απόδοση δεικτών σε διανύσματα

Έστω η παράσταση x[i]. Ανάλογα με τον τύπο του i έχουμε τα εξής :

i ακέραιος : Εάν είναι θετικός ακέραιος, το x[i] δίνει την αντίστοιχη συνιστώσα.. Εάν είναι αρνητικός δίνει όλα τα στοιχεία εκτός από το x[i].
i άλλος αριθμός : Μετατρέπεται σε ακέραιο μα αποκοπή.
i λογική τιμή : Μας δίνει τα στοιχεία για τα οποία το i είναι TRUE.
i χαρακτήρας : Τα string του i ταιριάζονται με με τα ονόματα των ορισμάτων του x.
Παράγοντας (Factor) : Το αποτέλεσμα είναι ίδιο με το αποτέλεσμα x[as.integer(i)].
Κενό : Η έκφραση x[] επιστρέφει x αλλά απορρίπτει «άσχετα» ορίσματα από το αποτέλεσμα.
NULL : φέρεται σαν να είναι integer(0).

3.6.3 Απόδοση δεικτών σε πίνακες και Arrays

Σε γενικές γραμμές η δεικτοδότηση πινάκων και Arrays είναι ίδια με αυτή των διανυσμάτων. Συνήθως ένα στοιχείο προσπελαύνεται με τη χρήση τόσων δεικτών όσος είναι ο αριθμός των διαστάσεών του.

Μία ιδιαιτερότητα είναι η ακόλουθη : Χρησιμοποιούμε ένα πίνακα ακεραίων σαν δείκτη.

Στο παρακάτω παράδειγμα ξεκαθαρίζεται η διαφορά να έχω στη πρώτη περίπτωση τη συνήθη δεικτοδότηση και μετά πρόσβαση με πίνακα δείχτη στα στοιχεία m[1, 1] και m[2, 2].

<u>Παράδειγμα</u>

```
> m<-matrix(1:4, 2)
> m
     [, 1] [, 2]
[1, ]    1    3
[2, ]   2    4
> m[1, 1]
[1] 1
> m[2, 2]
[1] 4
> i<-matrix(c(1, 1, 2, 2), 2, byrow=TRUE)
> i
     [, 1] [, 2]
[1, ]    1    1
[2, ]   2    2
> m[i]
[1] 1 4
```

3.7 Αντικείμενα της γλώσσας

Υπάρχουν τρία είδη αντικειμένων που απαρτίζουν τη γλώσσα R.Αυτά είναι : οι κλήσεις (*calls*), οι εκφράσεις (*expressions*), και τα ονόματα (*names*).
Τα αντικείμενα αυτά έχουν mode αντίστοιχα “call”, “expression”, “name”.

3.7.1 Το αντικείμενο NULL

Υπάρχει ένα ειδικό αντικείμενο που ονομάζεται NULL. Χρησιμοποιείται για να δηλώσει την απουσία ενός αντικειμένου. Δεν πρέπει να το συγχέουμε με ένα διάνυσμα ή λίστα μηδενικού μήκους.

Το αντικείμενο NULL δεν έχει τύπο και ούτε διαθέτει μετατρέψιμες ιδιότητες. Υπάρχει μόνο ένα NULL αντικείμενο στην R στο οποίο αναφέρονται όλες οι καταστάσεις.
Για ελέγξουμε εάν ένα αντικείμενο είναι NULL χρησιμοποιούμε is.null. Δεν μπορούμε να δώσουμε ορίσματα στο NULL.

3.7.2 Ιδιότητες αντικειμένων

Όλα τα αντικείμενα εκτός από το NULL έχουν ιδιότητες προσαρτημένες σε αυτά. Μία λίστα ιδιοτήτων μπορεί να προκύψει χρησιμοποιώντας την **attributes**. Μεμονωμένες ιδιότητες μπορεί να προσπελαστούν χρησιμοποιώντας την **attr**.

Οι πίνακες και τα Arrays είναι διανύσματα με τις ιδιότητες **dim** και **dimnames** προσαρτημένες.

Η ιδιότητα **names** δίνει ετικέτες στα αντικείμενα ενός διανύσματος ή μίας λίστας.
Η ιδιότητα **dim** χρησιμοποιείται για να υλοποιήσουμε **arrays**. Το περιεχόμενο ενός **array** αποθηκεύεται σε ένα διάνυσμα που είναι διατεταγμένο ως προς στήλες, και η ιδιότητα **dim** είναι ένα διάνυσμα αριθμών που δηλώνει τα μήκη των διαστάσεων του **array**.
Η R εξασφαλίζει ότι το μήκος του διανύσματος είναι το γινόμενο των μηκών των διαστάσεων. Το μήκος μιας ή περισσότερων διαστάσεων μπορεί να είναι μηδέν.

Το διάνυσμα δεν είναι το ίδιο με ένα μονοδιάστατο array διότι το δεύτερο έχει διάσταση ένα ενώ το πρώτο δεν έχει ιδιότητα dim.

Τα arrays μπορεί να δώσουν ονόματα σε κάθε διάσταση ξεχωριστά χρησιμοποιώντας την ιδιότητα dimnames που είναι μία λίστα από διανύσματα χαρακτήρων.

3.8 Αντικειμενοστραφής Προγραμματισμός

Το κεντρικό σημείο του αντικειμενοστραφούς προγραμματισμού είναι η έννοια της «κλάσης» (**class**) και της «μεθόδου» (**method**).
Μία κλάση αποτελείται από σχισμές (**slots**) που αποτελούν χώρους αποθήκευσης πληροφοριών της κλάσης. Ένα αντικείμενο είναι ένα στιγμιότυπο μιας κλάσης.

Οι υπολογισμοί γίνονται με τη βοήθεια μεθόδων. Οι μέθοδοι είναι συναρτήσεις που μπορούν να κάνουν ειδικούς υπολογισμούς πάνω σε κλάσεις.

Οι μέθοδοι είναι αυτοί που χαρακτηρίζουν μία γλώσσα «αντικειμενοστραφή». Μία άλλη βασική έννοια στον αντικειμενοστραφή προγραμματισμό είναι η «*κληρονομικότητα*».

Η μεγαλύτερη χρήση του αντικειμενοστραφούς προγραμματισμού στην R είναι μέσω των μεθόδων : plot, summary και print.

3.9 Προγραμματισμός στην R

Ο προγραμματισμός στην R συνίσταται στη δημιουργία συναρτήσεων του χρήστη. Για το σκοπό αυτό μας παρέχει τις γνωστές δομές ελέγχου και επανάληψης.

Οι εντολές μπορεί να ομαδοποιηθούν χρησιμοποιώντας άγκιστρα “{“ και “}”. Μία ομάδα εντολών ονομάζεται μπλοκ. Ένα μπλοκ υπολογίζεται μετά την εισαγωγή νέας γραμμής μετά το τέλος του μπλοκ.

Η μέθοδος με την οποία χειρίζεται τα ορίσματα η R όταν καλείται η συνάρτηση είναι «*κλήση με τιμή*». Συνεπώς τα ορίσματα συμπεριφέρονται σαν τοπικές μεταβλητές.

Η R έχει ένα <u>αργό</u> τρόπο απόδοσης τιμών στα ορίσματα των συναρτήσεων, δηλαδή τα ορίσματα δεν παίρνουν τιμές εάν αυτό δεν απαιτείται.

Πχ.

```
> {x<-12
+ x<-x+4
+ x<-x*2
+ }
>
> x
[1] 32
```

3.9.1 Δομές Ελέγχου

Η εντολή if

Η εντολή if είναι διαθέσιμη στην R σαν εντολή ελέγχου. Μπορούμε να έχουμε πολλές φωλιασμένες if.

Η σύνταξη της if είναι :

```
if ( statement1 )
        statement2
else
        statement3
```

<u>**Πχ.**</u>

```
> if( any(x <= 0) ) y <- log(1+x) else y <- log(x)
> y <- if( any(x <= 0) ) log(1+x) else log(x)
```

Τα φωλιασμένα if έχουν τη μορφή

```
if ( statement1 )
        statement2
else if ( statement3 )
        statement4
else if ( statement5 )
        statement6
else
        statement8
```

3.9.2 Εντολές Επανάληψης

Στην R διατίθενται οι τρεις βασικές εντολές επανάληψης **for, while, repeat**.

Η R είναι εφοδιασμένη επίσης με κάποιες επιπλέον συναρτήσεις για σιωπηρή επανάληψη, όπως **tapply, apply lapply**.

<u>Η R έχει επίσης τις εντολές :</u>

break για έξοδο από τον εσωτερικό βρόγχο.
next για επιστροφή του ελέγχου στην αρχή του βρόγχου.
return(A) για έξοδο από την τρέχουσα συνάρτηση και επιστροφή του A.

Η repeat επαναλαμβάνει το βρόγχο μέχρι να υπάρξει εντολή break.

Η repeat έχει την παρακάτω σύνταξη :

repeat statement

Η while έχει την παρακάτω σύνταξη :

while (statement1) statement2

Η for έχει την παρακάτω σύνταξη :

for (name in διάνυσμα)

statement1

4 Αντικείμενα Συναρτήσεων

Στην R μπορούμε να χειριστούμε τις συναρτήσεις σαν όλα τα αντικείμενα. Οι συναρτήσεις αποτελούνται από τρία βασικά μέρη. Μία λίστα από **τυπικά ορίσματα**, το **σώμα** και το **περιβάλλον**.

Η λίστα ορισμάτων είναι μία λίστα χωριζόμενων από κόμμα ορισμάτων, τα οποία μπορεί να είναι ένα σύμβολο, ένα κατασκεύασμα της μορφής "*symbol=default*" ή το ειδικό όρισμα "**...**". (Το όρισμα "**...**" χρησιμοποιείται όταν ο αριθμός των ορισμάτων είναι άγνωστος ή θ έλουμε να περάσουμε ορίσματα σε μία άλλη συνάρτηση).

Το σώμα είναι εντολές της R.

Το περιβάλλον μιας συνάρτησης είναι το περιβάλλον που είναι ενεργό όταν δημιουργείται μία συνάρτηση. Κάθε σύμβολο δεσμευμένο με το περιβάλλον είναι διαθέσιμο για τη συνάρτηση.

Ο συνδυασμός του κώδικα μιας συνάρτησης και το δέσιμο με το περιβάλλον λέγεται «*κλείσιμο*» της συνάρτησης.

4.1 Ενσωματωμένες συναρτήσεις στην R

Στην R υπάρχει ένα σύνολο από ενσωματωμένες συναρτήσεις τις οποίες μπορούμε να καλέσουμε ανά πάσα χρονική στιγμή. Στον παρακάτω πίνακα δίνεται μία ομάδα από συναρτήσεις γενικής χρήσης που είναι ενσωματωμένες. Άλλες συναρτήσεις περιγράφονται στα κατάλληλα σημεία μέσα στο παρόν κείμενο.

Πίνακας 1 : Ενσωματωμένες συναρτήσεις της R

Όνομα συνάρτησης	**Λειτουργία**
abs()	Απόλυτη τιμή
sqrt()	Τετραγωνική ρίζα
sin()	Ημίτονο

cos()	Συνημίτονο
tan()	Εφαπτομένη
exp()	Εκθετική
log()	Λογάριθμος
log10()	Δεκαδικός Λογάριθμος
gamma()	Συνάρτηση γάμμα

4.2 *Κατανομές πιθανοτήτων*

Στην R υπάρχει διαθέσιμο ένα μεγάλο πλήθος συναρτήσεων για υπολογισμό της αθροιστικής συνάρτησης κατανομής $P(X \le x)$, της συνάρτησης πυκνότητας πιθανότητας, των ποσοστημορίων (αντίστροφη συνάρτησης κατανομής δηλ. δοθέντος q το μικρότερο x για το οποίο ισχύει $P(X \le x) > q$) και της προσομοίωσης (τυχαίοι αριθμοί από μία κατανομή).

Στον παρακάτω πίνακα δίνονται τα R-ονόματα των συναρτήσεων αυτών

Πίνακας 2 : R-ονόματα

Κατανομή	**R - όνομα**	**Ορίσματα**
Βήτα	beta	shape1, shape2, ncp
Διωνυμική	binom	size, prob
Cauchy	cauchy	location, scale
X^2 κατανομή	chisq	df, ncp
Εκθετική	exp	rate
F	f	df1, df2, ncp
Γάμμα	gamma	shape, scale
Γεωμετρική	geom	prob
Υπεργεωμετρική	hyper	m, n, k
Λογαριθμοκανονική κατανομή	lnorm	meanlog, sdlog
Λογιστική κατανομή	logis	location, scale
Αρνητική διωνυμική	nbinom	size, prob

Κανονική	norm	mean, sd
Κατανομή Poisson	pois	lambda
Student’s t	t	df, ncp
Ομοιόμορφη	unif	min, max
Weibull	weibull	shape, scale

Στο πακέτο της R μας δίνεται και η κατανομή της στατιστικής συνάρτησης του αθροίσματος βαθμών (rank sum) Wilcoxon που προκύπτει από δείγματα μεγέθους m και n αντίστοιχα.

Το R – όνομα της συνάρτησης Wilcoxon είναι : wilcox, με ορίσματα m ,n.

Για να γίνει όνομα μίας συνάρτησης ένα R–όνομα, ακολουθείται ο εξής κανόνας. Μπαίνει ένα πρόθεμα στο R-όνομα που δηλώνει τη συνάρτηση.

Τα προθέματα είναι

d για συνάρτηση πιθανότητας.

p για συνάρτηση κατανομής.

q για αντίστροφη συνάρτηση κατανομής (συνάρτηση ποσοστημορίων).

r για προσομοίωση (τυχαίοι αριθμοί).

Στις συναρτήσεις προσομοίωσης μπαίνει παράμετρος το πλήθος n των ζητούμενων αριθμών (σε κάποιες περιπτώσεις απαιτείται nn).

Το **ncp** είναι μία νέα παράμετρος που μπαίνει και είναι μη κεντρική παράμετρος (non centrality parameter).

Πχ.

dchisq(x,df, ncp=0) , όπου df=βαθμοί ελευθερίας.

Παραδείγματα

Θα υπολογίσουμε για το διάνυσμα x την πιθανότητα της τυποποιημένης κανονικής κατανομής και μετά της κανονικής κατανομής με μέσο 5 και τυπική απόκλιση 2. Μετά θα παράγουμε 30 τυχαίους αριθμούς από την κανονική κατανομή N(5,1). Τέλος θα βρούμε την πιθανότητα της ομοιόμορφης κατανομής για το διάνυσμα x.

```
x<-c(1, 4, 7, 10)
> pnorm(x)
[1] 0.8413447 0.9999683 1.0000000 1.0000000
> pnorm(x, mean=5, sd=2)
[1] 0.02275013 0.30853754 0.84134475 0.99379033
> rnorm(30, mean=2, sd=1)
 [1]  2.4271553  0.4877014  1.2503158  3.0117239  0.6548330  2.3621217
2.8124266
 [8]  3.4399093  2.3343802  1.8654787  2.3953171  2.1943066  1.1178424
0.1306050
[15]  1.4423840  1.0768016  3.5788187  3. 3658251  0.8893339  1.1768066
2.1564240
[22]  0.4657078  1.9558831  2.4969718  1.3854240  3.8294745  2.6434773
1.2623970
[29] 1.1245116 2.6176249
> punif(x, -5, 5)
[1] 0.6 0.9 1.0 1.0
```

Θα υπολογίσουμε το 95% σημείο για την κατανομή f με 1 και 18 βαθμούς ελευθερίας.

```
> qf(0.95,1,18)
[1] 4.413873
```

4.3 Γραφήματα

Οι δυνατότητες της R στο θέμα των γραφικών παραστάσεων είναι ένα από τα ισχυρά της συστατικά.

Οι εντολές που μας προσφέρει η R είναι χωρισμένες σε τρεις κατηγορίες :

- Υψηλού επιπέδου.
- Χαμηλού επιπέδου.

- Αλληλεπιδραστικές συναρτήσεις γραφημάτων .

Υψηλού επιπέδου εντολές

Συνάρτηση	**Λειτουργία**
boxplot(x)	Δημιουργεί το boxplot για το διάνυσμα x.
stem(x)	Δίνει το steam and leaf γράφημα για το x.
pairs(X)	Εάν το X είναι ένας αριθμητικός πίνακας ή πλαίσιο δεδομένων , τότε δημιουργεί όλα τα διαγράμματα κατανομών των μεταβλητών ανά δύο.
coplot(a~b|c)	Εάν τα a, b είναι αριθμητικά διανύσματα και το c αριθμητικό ή παράγοντας, τότε δημιουργείται το διάγραμμα διασποράς του a ως προς b για κάθε τιμή του c.
qqnorm(x)	Το διάνυσμα x είναι αριθμητικό, και δημιουργείται το διάγραμμα του x ως προς τα κανονικά scores.
qqline(x)	Προσθέτει γραμμές στο προηγούμενο διάγραμμα ενώνοντας τα τεταρτημόρια.
qqplot(x, y)	Δημιουργεί το διάγραμμα των τεταρτημορίων του x ως προς y για να συγκρίνει τις αναμενόμενες κατανομές.
hist(x)	Δημιουργεί ιστόγραμμα για το αριθμητικό διάνυσμα x.Το σύστημα εκτιμάει ένα δικό του αριθμό κλάσεων. Εναλλακτικά μπορούμε να προτείνουμε τον αριθμό των κλάσεων με **hist(x, nclass=n)** ή τα σημεία διακοπής αυτών με **hist(x, breaks=b, ...)**.
dotchart(x)	Δημιουργεί ένα διάγραμμα με κουκίδες για το διάνυσμα x.
ts.plot(x)	Δημιουργεί γράφημα χρονολογικής σειράς.

Μέθοδος plot

plot(x, y)	Εάν x, y είναι διανύσματα μας δίνει το διάγραμμα διασποράς του y ως προς x.
plot(x)	Εάν το x είναι χρονολογική σειρά μας δίνει το διάγραμμά της.
plot(f)	Εάν το f είναι παράγοντας μας δίνει το ραβδόγραμά του.

plot(f, y) Εάν το f είναι παράγοντας και το y αριθμητικό διάνυσμα μας δίνει τα boxplot του y για κάθε επίπεδο του f.

plot(df) Εάν το df είναι ένα Data frame τότε δημιουργεί τα διαγράμματα κατανομών του πλαισίου δεδομένων.

plot(~expr) Δημιουργεί τα διαγράμματα κατανομών των μεταβλητών του αντικειμένου expr.

plot(y~expr) Δημιουργεί τα διαγράμματα κατανομών των μεταβλητών του y ως προς κάθε μεταβλητή του αντικειμένου expr.

Μέθοδος summary

summary(object, ...) Είναι μία συνάρτηση που παράγει αποτελέσματα σύνοψης ανάλογα με την κλάση του ορίσματος object. Στα επόμενα θα περιγράψουμε κάποιες ειδικές περιπτώσεις.

Ορίσματα στις υψηλού επιπέδου εντολές

log="x" , log="y", log="xy" : Εξαναγκάζει τους άξονες x, y ή και τους δύο μαζί να είναι λογαριθμικοί. Δεν λειτουργεί με όλους τους τους τύπους γραφημάτων.

To type= argument ελέγχει τον τύπο του γραφήματος ως εξής :

type="p" Εκτυπώνει σημεία (the default).

type="l" Εκτυπώνει γραμμές.

type="b" Εκτυπώνει σημεία ενωμένα με γραμμές (both).

type="o" Εκτυπώνει σημεία επικαλυπτόμενα από γραμμές.

type="h" Εκτυπώνει κατακόρυφες γραμμές από τα σημεία στον οριζόντιο άξονα (high-density).

type="s"

type="S" Γράφημα βηματικής συνάρτησης. Στην πρώτη μορφή, η κορυφή του κατακόρυφου ορίζεται από το σημείο στη δεύτερη μορφή το σημείο ορίζει το βάθος.

type="n" Κανένα γράφημα.

xlab=string
ylab=string Ετικέτες για τους x και y άξονες.
main=string Τίτλος γραφήματος που τοποθετείται στην κορυφή .
sub=string Υπότιτλος που χρησιμοποιείται κάτω από τον x-άξονα.

<u>Χαμηλού επιπέδου εντολές</u>

Οι εντολές χαμηλού επιπέδου χρησιμοποιούνται για να δώσουν επιπλέον πληροφορία σε ένα τρέχον γράφημα.

points(x, y)
lines(x, y)
Προσθέτει σημεία ή συνδεδεμένες γραμμές στο τρέχον γράφημα.

text(x, y, labels, ...)
Προσθέτει κείμενο σε ένα γράφημα στα σημεία που δίνονται από τα x, y.

abline(a, b)
Προσθέτει μία γραμμή με slope b και intercept a στο τρέχον γράφημα .

polygon(x, y, ...)
Σχεδιάζει ένα πολύγωνο που ορίζεται από από τις διατεταγμένες συντεταγμένες (x, y) και (προαιρετικά) τη σκιάζει με γραμμές.

legend(x, y, legend, ...)
Προσθέτει υπόμνημα στην καθορισμένη θέση.

title(main, sub)
Προσθέτει τίτλο και υπότιτλο στο τρέχον γράφημα .

<u>Άλλες εντολές γραφημάτων</u>

locator(n, type)
Περιμένει για να επιλέξουμε σημεία με το αριστερό πάτημα του ποντικιού .

identify(x, y, labels)

Επιτρέπει στο χρήστη να φωτίσει οποιοδήποτε σημείο ορίζεται από τα x, y.

par()

Χρησιμοποιείται για να προσπελάσουμε και να αλλάξουμε τη λίστα των παραμέτρων των γραφημάτων.

mfcol=c(α, β)

mfrow=c(α, β)

Θέτει το μέγεθος ενός χώρου με πολλαπλές εικόνες. Το α είναι ο αριθμός στηλών και το β ο αριθμός γραμμών. Η μοναδική διαφορά των δύο συναρτήσεων είναι ότι η πρώτη γεμίζει πρώτα ανά στήλες τις εικόνες.

<u>Παραδείγματα</u>

Το πακέτο faithful της R περιέχει το διάνυσμα eruptions. Για αυτό το το αριθμητικό διάνυσμα έχουμε (δίνει 6 χαρακτηριστικά μεγέθη από την κατανομή της τα : Min., 1st Qu., Median, Mean, 3rd Qu., Max.)

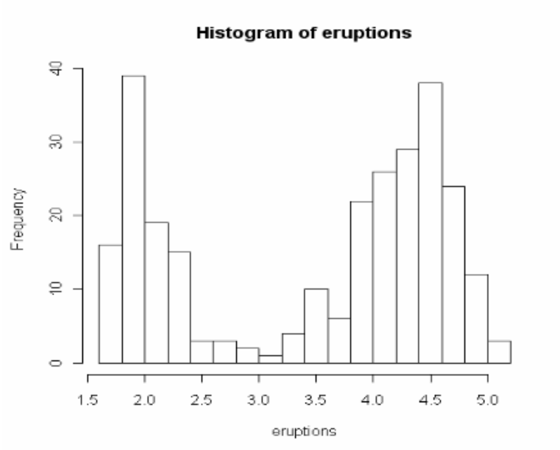
```
>summary(eruptions)
  Min.    1st Qu.    Median    Mean    3rd Qu.    Max.
 1.600     2.163     4.000     3.488     4.454     5.100

> stem(eruptions)

  The decimal point is 1 digit(s) to the left of the |

  16 | 070355555588
  18 | 000022233333335577777777888822335777888
  20 | 00002223378800035778
  22 | 0002335578023578
  24 | 00228
  26 | 23
  28 | 080
  30 | 7
```

32 | 2337
34 | 250077
36 | 0000823577
38 | 2333335582225577
40 | 0000000335778888800223355555577778
42 | 03335555778800233333555577778
44 | 0222223355557780000000023333357778888
46 | 0000233357700000023578
48 | 00000022335800333
50 | 0370

> hist(eruptions)

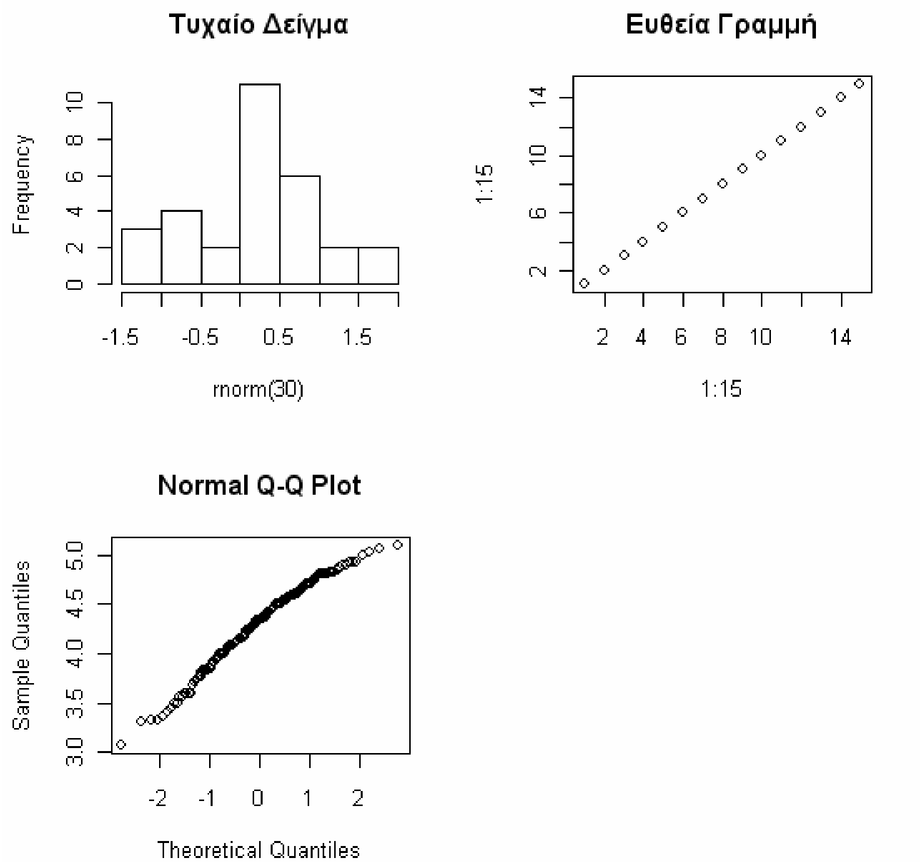
Γράφημα 1 : Ιστόγραμμα eruptions

> hist(eruptions, seq(1.6, 5.2, 0.2))



Γράφημα 2 : Ιστόγραμμα eruptions με καθορισμένα διαστήματα

```
> par(mfrow=c(2, 2))
> hist(rnorm(30), main="Τυχαίο Δείγμα")
> plot(1:15, 1:15, main="Ευθεία Γραμμή")
> qqnorm(long)
```



Γράφημα 3: Τυχαίο δείγμα, Q-Q Plot , ευθεία γραμμή

5 Συμπερασματολογία

Η R δίνει ένα πλήρες σετ εντολών για συμπερασματολογία. Εδώ θα αναφερθούν οι βασικότερες εντολές. Για περισσότερες πληροφορίες στο Manual της γλώσσας.

1. summary(x) , όπου το x είναι διάνυσμα

Δίνει 6 χαρακτηριστικά μεγέθη από την κατανομή της x (τα : Min., 1st Qu., Median, Mean, 3rd Qu., Max.)

<u>Πχ.</u>

```
>a
 [1] 12  3 65  7 90  8 23  6  5 78
> summary(a)
   Min.   1st Qu.    Median     Mean  3rd Qu.     Max.
   3.00     6.25   10.00      29.70     54.50     90.00
```

2. t-test για ένα δείγμα (H0: μ=μ0 vs H1: μ≠μ0)

t.test(x, mu=m0, conf.level=0.99)

Εάν το conf.level παραλείπεται, τότε εννοείται 0.95

<u>Παράδειγμα</u>

```
> a
 [1] 12  3 65  7 90  8 23  6  5 78 23  7 45  9 12
> t.test(a, mu=25)

        One Sample t-test

data:  a
t = 0.1601, df = 14, p-value = 0.875
alternative hypothesis: true mean is not equal to 25
```

95 percent confidence interval:
 10.12507   42.27493
sample estimates:
mean of x
     26.2

Δεν μπορούμε να απορρίψουμε την H_0 επειδή p-value = 0.875

3. t-test για δύο δείγματα (H0: μ1=μ2 vs H1: μ1≠μ2)

Για ασυσχέτιστα δείγματα : t.test(x, y, var.equal=F)
(το var.equal=F παραλείπεται)
Συσχετισμένα δείγματα δείγματα : t.test(x, y, var.equal=T)

Εάν το conf.level παραλείπεται, τότε εννοείται 0.95. Εδώ το mu= μ_1-μ_2.
Εάν το mu παραλείπεται εννοείται μηδέν.

<u>Παράδειγμα</u>
Κάνουμε έλεγχο υποθέτοντας διαφορετικές διακυμάνσεις.

```
> A
 [1] 79.98 80.04 80.02 80.04 80.03 80.03 80.04 79.97 80.05 80.03 80.02 80.00
[13] 80.02
> B
[1] 80.02 79.94 79.98 79.97 79.97 80.03 79.95 79.97
> t.test(A, B)

        Welch Two Sample t-test

data:  A and B
t = 3.2499, df = 12.027, p-value = 0.00694
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
```

```
 0.01385526 0.07018320
sample estimates:
mean of x mean of y
 80.02077  79.97875
```

Παρατηρούμε ότι η H_0 απορρίπτεται έχοντας υποθέσει ότι δεν είναι ίσες οι διακυμάνσεις στα δύο δείγματα.

Κάνουμε τον ίδιο έλεγχο υποθέτοντας ισότητα διακυμάνσεων.

```
> t.test(A, B, var.equal=TRUE)

        Two Sample t-test

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

4. Έλεγχος διακυμάνσεων δύο κανονικών πληθυσμών

var.test(x, y) **συγκρίνει τις δύο διακυμάνσεις**

Παράδειγμα

Συνεχίζοντας με τα δεδομένα του προηγούμενου παραδείγματος έχουμε :

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

        F test to compare two variances
```

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

Επειδή **p-value = 0.3938** δεν απορρίπτουμε την υπόθεση ισότητας των διασπορών.

5. Έλεγχος συσχέτισης δύο πληθυσμών

```
cor.test(x,  y,
            alternative = c("two.sided",  "less",  "greater"),
            method = c("pearson",  "kendall",  "spearman"),
            exact = NULL,  conf.level = 0.95, ...)
```

Οι πρώτες τιμές των ρυθμίσεων είναι οι προκαθορισμένες και παραλείπονται όταν είναι οι ζητούμενες.

<u>Παράδειγμα</u>

```
> x <- c(44.4,  45.9,  41.9,  53.3,  44.7,  44.1,  50.7,  45.2,  60.1)
> y <- c( 2.6,   3.1,   2.5,   5.0,   3.6,   4.0,   5.2,   2.8,   3.8)
> cor.test(x, y)

        Pearson's product-moment correlation

data:  x and y
t = 1.8411,  df = 7,  p-value = 0.1082
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
```

-0.1497426 0.8955795
sample estimates:
      cor
0.5711816

6. Ελεγχος καλής Προσαρμογής (Kolmogorov – Smirnov)

Στην περίπτωση σύγκρισης δείγματος με κατανομή είναι :

H_0 : το δείγμα ακολουθεί την κατανομή

Στην περίπτωση σύγκρισης δύο δειγμάτων είναι :

H_0 : τα δείγματα ακολουθούν την ιδία κατανομή

Η σύνταξη της εντολής είναι ανάλογα με τον έλεγχο :

ks.test(δείγμα, «κατανομή»)   ή   ks.test(δείγμα1, δείγμα2)

<u>παράδειγμα</u>

```
> ks.test(x, y)

        Two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.96,  p-value < 2.2e-16
alternative hypothesis: two-sided

> ks.test(x, "punif", 1, 20)

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.151,  p-value = 0.1847
alternative hypothesis: two-sided
```

7. Ακριβής διωνυμικός έλεγχος binom.test

Εκτελεί ένα έλεγχο με ***H_0 : πιθανότητα επιτυχιών = p***, σε ένα πείραμα Bernoulli.

binom.test(x, n, p = 0.5,
alternative = c("two.sided", "less", "greater"),
conf.level = 0.95)

Ορίσματα

x	Αριθμός επιτυχιών ή ένα διάνυσμα μεγέθους 2 που δίνει αριθμό επιτυχιών και αποτυχιών αντίστοιχα
n	Αριθμός προσπαθειών. Αγνοείται όταν το διάνυσμα x έχει μέγεθος 2.
p	Υποθετική πιθανότητα επιτυχιών.
alternative	Δείχνει ότι ο έλεγχος είναι "two.sided", "greater" or "less".
conf.level	Επίπεδο σημαντικότητας

<u>Πχ.</u>

```
> binom.test(682, 925, p=3/4)

        Exact binomial test

data:  682 and 925
number of successes = 682,  number of trials = 925,  p-value = 0.3825
alternative hypothesis: true probability of success is not equal to 0.75

95 percent confidence interval:
 0.7076683      0.7654066

sample estimates:
probability of success
            0.7372973
```

8. Τεστ ισότητας ποσοστών prop.test

Το prop.test ελέγχει την ισότητα των ποσοστών (πιθανότητες επιτυχίας) κάποιων ομάδων, ή αν τα ποσοστά των ομάδων είναι ίσια των τιμών ενός δοσμένου διανύσματος.

```
prop.test(x, n, p = NULL,
        alternative = c("two.sided", "less", "greater"),
        conf.level = 0.95, correct = TRUE)
```

Ορίσματα

x	Διάνυσμα πλήθους επιτυχιών, ή πίνακας με δύο στήλες που δίνει αριθμό επιτυχιών και αποτυχιών αντίστοιχα
n	Διάνυσμα πλήθους προσπαθειών. Παραλείπεται αν x είναι πίνακας
p	Διάνυσμα πιθανοτήτων επιτυχίας. Πρέπει να έχει τι ίδιο μήκος με τον αριθμό των ομάδων και οι τιμές του να είναι μεταξύ 0 και 1.
alternative	Δείχνει ότι ο έλεγχος είναι "two.sided", "greater" or "less"..
conf.level	Επίπεδο σημαντικότητας
correct	Μία λογική παράμετρος που δείχνει εάν η διόρθωση συνέχειας του Yates' θα πρέπει να εφαρμοστεί.

Πχ.

```
> smokers<-c(83, 90, 129, 70)
> patients<-c(86, 93, 136, 82)
> prop.test(smokers, patients)

        4-sample test for equality of proportions without continuity correction

data:  smokers out of patients
X-squared = 12.6004, df = 3, p-value = 0.005585
alternative hypothesis: two.sided
```

sample estimates:

```
   prop 1       prop 2       prop 3       prop 4
0.9651163    0.9677419    0.9485294    0.8536585
```

9. Πίνακες συνάφειας

Ο πίνακας συνάφειας δημιουργείται με την εντολή table ή με την εντολή as.table(πίνακας).

Ο έλεγχος Pearson X^2 δίνεται με την εντολή summary.

Η cross classification (διασταυρούμενη τοποθέτηση σε κλάσεις) δίνεται με την εντολή xtabs(~formula), όπου formula οι μεταβλητές χωριζόμενες με '+' Με την εντολή CrossTable της βιβλιοθήκης *(gmodels)* δίνονται τα αποτελέσματα σε πίνακα.

<u>Παράδειγμα</u>

Ο παρακάτω πίνακας συνάφειας που θα δημιουργηθεί παριστάνει τις συχνότητες των περιστατικών που κρυολόγησαν ή δεν κρυολόγησαν από την χορήγηση 4 διαφορετικών φαρμάκων.

Πίνακας 3 : Πίνακας συχνοτοτήτων περιστατικών

	med 1	med 2	med 3	med 4
get a cold	15	26	9	14
healthy	111	107	96	117

Με χρήση της R έχουμε

```
> pnx
   [, 1] [, 2] [, 3] [, 4]
l1   15   26    9   14
l2  111  107   96  117
> dimnames(pnx)<-list(c("get a cold", "healthy"), c("med 1", "med 2",  "med
3", "med 4"))
```

```
> pnx
           med 1 med 2 med 3 med 4
get a cold    15    26     9    14
healthy      111   107    96   117

> tbl<-as.table(pnx)

> summary(tbl)
Number of cases in table: 495
Number of factors: 2
Test for independence of all factors:
        Chisq = 7.651, df = 3, p-value = 0.05381
```

6 Στατιστικά μοντέλα στην R

6.1 Γραμμικά μοντέλα

Συνάρτηση : lm(τύπος, data=dataframe)

Η συνάρτηση **lm()** μας δίνει ένα αποδοτικό τρόπο υπολογισμού συνηθισμένων γραμμικών μοντέλων.

Η συνάρτηση **summary(αντικείμενο)** δίνει μία περιεκτική περίληψη των αποτελεσμάτων της ανάλυσης διακύμανσης.

Πχ.

```
> x<-c(12, 18, 24, 30, 36, 42, 48)
> y<-c(5.27, 5.68, 6.25, 7.21, 8.02, 8.71, 8.42)
```

Εκτίμηση ευθείας παλινδρόμησης

```
> gram1<-lm(y~x)
> gram1

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)           x
  3.9943        0.1029
```

Δηλαδή η γραμμή παλινδρόμησης είναι $\hat{y} = 3.9943 + 0.1029x$

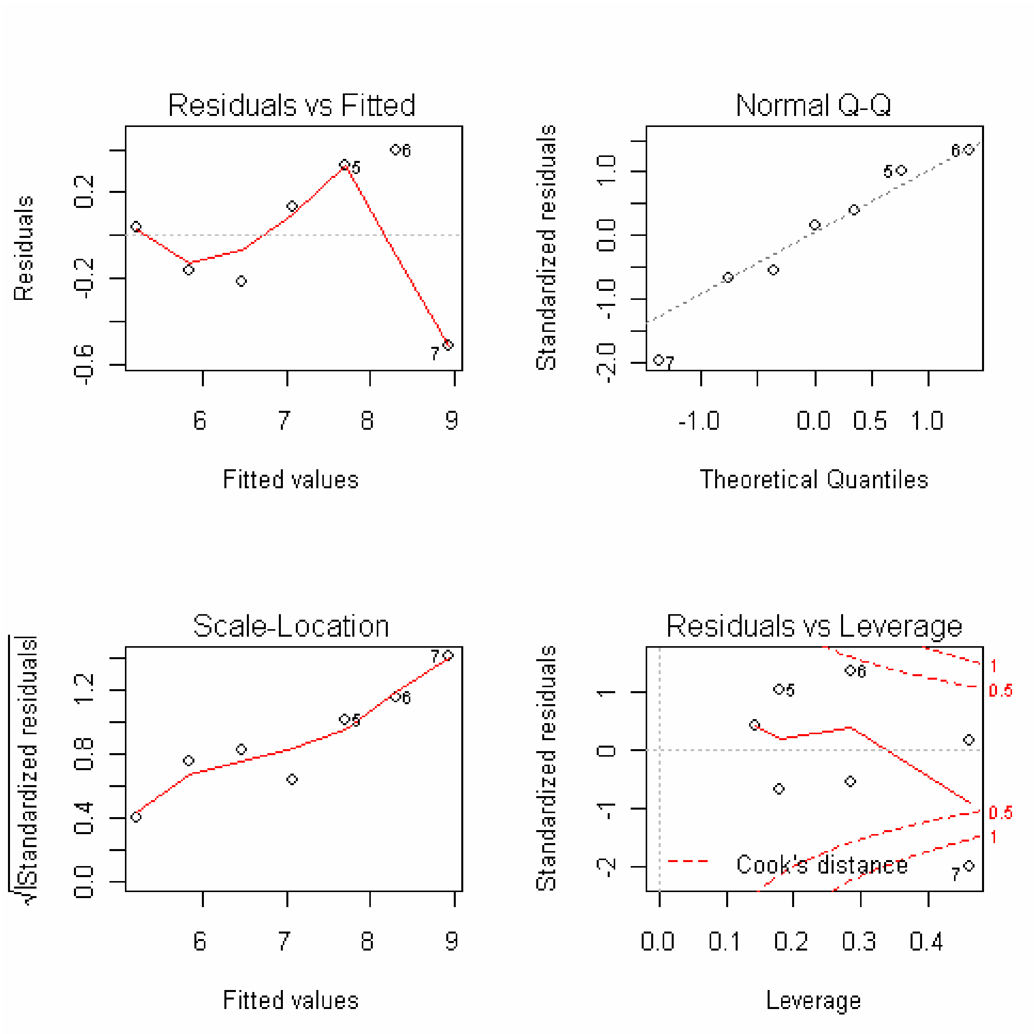
Αποτελέσματα της ανάλυσης διακύμανσης

```
> summary(gram1)

Call:
lm(formula = y ~ x)

Residuals:
       1          2          3          4          5          6          7
 0.04143   -0.16571  -0.21286    0.13000    0.32286    0.39571   -0.51143

Coefficients:
             Estimate   Std. Error   t value    Pr(>|t|)
(Intercept)  3.99429    0.35656      11.202     9.9e-05 ***
x            0.10286    0.01104      9.321      0.000239 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3504 on 5 degrees of freedom
Multiple R-squared: 0.9456,     Adjusted R-squared: 0.9347
F-statistic: 86.87 on 1 and 5 DF,  p-value: 0.0002393
```

Γραφήματα από ανάλυση διακύμανσης

Ορίζουμε να έχουμε στην οθόνη 2 x 2 γραφήματα και μετά εμφανίζουμε τα γραφήματα :

```
> par(mfrow=c(2, 2))
> plot(gram1)
```



Γράφημα 4 : Γραφήματα από ανάλυση διακύμανσης

6.2 Ανάλυση διακύμανσης ANOVA

Για την ανάλυση διακύμανσης χρησιμοποιούνται δύο συναρτήσεις. Η συνάρτηση **aov** που κάνει την ανάλυση και η συνάρτηση **summary** που εμφανίζει τη σύνοψη αποτελεσμάτων.

aov(formula, data = NULL, projections = FALSE, qr = TRUE,
contrasts = NULL, ...)

Προσαρμόζει ένα μοντέλο ανάλυσης διακύμανσης

Ορίσματα

formula Ένας τύπος που προσαρμόζει τα δεδομένα.

data Ένα data frame στο οποίο οι μεταβλητές του τύπου ψάχνουν για δεδομένα.

<u>Παράδειγμα</u>

Έστω ο πίνακας βαθμολογίας για πέντε διαφορετικές μεθόδους διδασκαλίας που θέλουμε να συγκρίνουμε.

Πίνακας 4 : Πίνακας βαθμολογίας

	Μέθοδοι				
	1	**2**	**3**	**4**	**5**
Βαθμοί	93	73	75	89	59
	97	77	84	81	64
	92	67	80	76	55
	85	76	70	75	67

```
> v1<-c(93, 97, 92, 85)
> v2<-c(73, 77, 67, 76)
> v3<-c(75, 84, 80, 70)
> v4<-c(89, 81, 76, 75)
> v5<-c(59, 64, 55, 67)
> v0<-c(v1, v2, v3, v4, v5)

> nov<-data.frame(block=gl(5, 4), v0=v0)

> nov.aov<-aov(v0~block, nov)

> nov.aov

Call:
```

```
  aov(formula = v0 ~ block,  data = nov)
Terms:
                block Residuals
Sum of Squares  1960.00    453.75
Deg. of Freedom       4        15

Residual standard error: 5.5
Estimated effects may be unbalanced

> summary(nov.aov)
            Df         Sum            Sq Mean       Sq              F value    Pr(>F)
block        4   1960.00        490.00          16.198          2.550e-05 ***
Residuals   15  453.75        30.25
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1
```

6.2.1 Η εντολή anova και η σύγκριση μοντέλων

Η εντολή **anova** μπορεί να χρησιμοποιηθεί για τη σύγκριση δύο μοντέλων. Μπορεί να χρησιμοποιηθεί επίσης και για τη δημιουργία πίνακα ανάλυσης διακύμανσης και πίνακα ανάλυσης απόκλισης (Deviance).

Στην περίπτωση που έχει τη σύνταξη :

anova(fitted.model1, fitted.model2, ...)

όπου τα ορίσματα αποτελούν προσαρμοσμένα μοντέλα , κάνει τη σύγκριση των μοντέλων.

Στην περίπτωση που έχει τη σύνταξη :

anova(fitted.model)

όπου υπάρχει μόνο ένα όρισμα που είναι ένα προσαρμοσμένο γραμμικό ή γενικευμένο γραμμικό μοντέλο, τότε δημιουργεί ένα πίνακα διακύμανσης ή πίνακα απόκλισης (Deviance) αντίστοιχα.

Ένα μοντέλο μπορεί να μεταβληθεί και να προκύψει νέο προσαρμοσμένο μοντέλο με τη βοήθεια της εντολής update.

new.model<-update(old.model, new.formula)

Το πλήρες γραμμικό μοντέλο δίνεται από το :

Fmfull <-lm(y~., data=dataset)

6.3 *Γενικευμένα γραμμικά μοντέλα*

Κάθε κατανομή απόκρισης αναγνωρίζει μία ποικιλία από συναρτήσεις σύνδεσης. Οι οικογένειες κατανομών με τις διαθέσιμες συναρτήσεις σύνδεσης εμφανίζονται στον παρακάτω πίνακα .

Ονομα Οικογένειας	**Συναρτήσεις σύνδεσης**
binomial	logit, probit, log, cloglog
gaussian	identity, log, inverse
Gamma	identity, inverse, log
inverse.gaussian	1/mu^2, identity, inverse, log
poisson	identity, log, sqrt
quasi	logit, probit, cloglog, identity, inverse ,log, 1/mu^2, sqrt

Στην R η συνάρτηση που προσαρμόζει ένα γενικευμένο γραμμικό μοντέλο είναι η:

glm(formula, family=family.generator, data=data.frame)

Παράδειγμα της κανονικής (Gaussian) οικογένειας

Στη περίπτωση αυτή η χρήση της συνάρτησης **lm** και της **glm** δίνουν τα ίδια αποτελέσματα.

Πχ.

```
> y<-c(166, 180, 73, 81, 229, 182, 233, 102, 190, 150, 221, 137, 173, 150, 92)
> x1<-c(10, 9, 10, 14, 8, 15, 6, 10, 7, 10, 11, 15, 8, 12, 10)
> x2<-c(20, 21, 12, 16, 24, 24, 23, 15, 20, 19, 25, 21, 19, 20, 14)
> flm<-lm(y~x1+x2)
> flm

Call:
lm(formula = y ~ x1 + x2)

Coefficients:
(Intercept)          x1          x2
     -2.765      -7.738      12.286

> fgm<-glm(y~x1+x2, family=gaussian)

> fgm

Call:  glm(formula = y ~ x1 + x2,  family = gaussian)

Coefficients:
(Intercept)          x1          x2
     -2.765      -7.738      12.286

Degrees of Freedom: 14 Total (i.e. Null);  12 Residual
Null Deviance:      38210
Residual Deviance: 164  AIC: 86.45

> summary(flm)

Call:
lm(formula = y ~ x1 + x2)
```

```
Residuals:
    Min         1Q          Median       3Q          Max
-5.6033     -2.1628     -0.1039     1.4675      5.9650

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.7648    6.3906    -0.433    0.673
x1           -7.7378    0.3638   -21.272   6.78e-11 ***
x2           12.2861    0.2566    47.889   4.50e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.697 on 12 degrees of freedom
Multiple R-squared: 0.9957,      Adjusted R-squared: 0.995
F-statistic:  1392 on 2 and 12 DF,   p-value: 6.253e-15

> summary(fgm)

Call:
glm(formula = y ~ x1 + x2,  family = gaussian)

Deviance Residuals:
    Min         1Q          Median       3Q          Max
-5.6033     -2.1628     -0.1039     1.4675      5.9650

Coefficients:
            Estimate    Std. Error   t value    Pr(>|t|)
(Intercept) -2.7648     6.3906       -0.433     0.673
x1          -7.7378     0.3638       -21.272    6.78e-11 ***
x2          12.2861     0.2566       47.889     4.50e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Dispersion parameter for gaussian family taken to be 13.66871)

    Null deviance: 38214.93 on 14 degrees of freedom
Residual deviance: 164.02 on 12 degrees of freedom
AIC: 86.448

Number of Fisher Scoring iterations: 2

<u>Παράδειγμα της διωνυμικής (binomial) οικογένειας</u>

Για να προσαρμόσουμε ένα διωνυμικό μοντέλο χρησιμοποιώντας glm() πρέπει να έχουμε την μεταβλητή απόκρισης σε μία από τις παρακάτω τρεις μορφές.

1. Η απόκριση είναι διάνυσμα που περιέχει δυαδικά στοιχεία (μ όνο 1 και 0).
2. Η απόκριση είναι πίνακας δύο διαστάσεων. Η πρώτη στήλη περιέχει τους αριθμούς επιτυχιών που προκύπτουν από τις δοκιμές και η δεύτερη τους αριθμούς αποτυχιών.
3. Εάν η απόκριση είναι παράγοντας τότε το πρώτο επίπεδο λαμβάνεται σαν αποτυχία (0) και το δεύτερο σαν επιτυχία (1).

Η προκαθορισμένη μορφή της συνάρτησης σύνδεσης είναι η **logit** (**λογιστική**).

<u>Παράδειγμα</u>
(εικονικά δεδομένα για χάρη του παραδείγματος)

Σε ένα ορεινό χωριό κάπου στην Ελλάδα , οι γηγενείς αρένες υποφέρουν από μία ασθένεια των οφθαλμών που με τη πάροδο του χρόνου προκαλεί τύφλωση. Στον παρακάτω πίνακα είναι δείγμα από άτομα που εξετάστηκαν για τύφλωση με τα αποτελέσματά τους.

Πίνακας 5 : Πίνακας εικονικών δεδομένων

Ηλικία :	20	35	45	55	70
Πλήθος Εξετασθέντων :	50	50	50	50	50
Πλήθος τυφλών :	6	17	26	37	44

Εάν y είναι ο αριθμός των τυφλών που βρίσκονται στην ηλικία x, και n είναι ο αριθμός των ατόμων που εξετάστηκαν τότε το μοντέλο ακολουθεί τη διωνυμική κατανομή, και έχει τη παρακάτω μορφή :

$$y \sim B\left(n, F\left(\beta_0 + \beta_1 x\right)\right)$$

Όπου F είναι η συνάρτηση σύνδεσης.

Θα εξετάσουμε το μοντέλο για δύο περιπτώσεις. Όταν η συνάρτηση σύνδεσης είναι η κανονική (περίπτωση probit) δηλ. $F(z) = \Phi(z)$ και στην περίπτωση που είναι η λογιστική (logit που είναι και η προκαθορισμένη περίπτωση) δηλαδή

$$F(z) = \frac{e^z}{(1+e^z)}$$

Θα εκτιμήσουμε και στα δύο μοντέλα το LD50 (δηλαδή την ηλικία στην οποία η πιθανότητα να είναι κάποιος τυφλός είναι 50%). Και στα δύο μοντέλα $LD50 = -\frac{\beta_0}{\beta}$ δηλ. το σημείο στο οποίο μηδενίζεται το όρισμα στη συνάρτηση κατανομής.

Έχουμε στην R

```
> xplace<-data.frame(x=c(20, 35, 45, 55, 70), n=rep(50, 5), y=c(6, 17, 26, 37, 44))
> xplace$Tmat<-cbind(xplace$y, xplace$n - xplace$y)
> fmp<-glm(Tmat~x, family=binomial(link=probit), data=xplace)
> fml<-glm(Tmat~x, family=binomial, data=xplace)
> fmp

Call: glm(formula = Tmat ~ x, family = binomial(link = probit), data = xplace)
```

```
Coefficients:
(Intercept)           x
   -2.10227      0.04815

Degrees of Freedom: 4 Total (i.e. Null);  3 Residual
Null Deviance:      82.14
Residual Deviance: 0.4547       AIC: 24.27
> 2.10227/0.04815
[1] 43.66085

> fml

Call:  glm(formula = Tmat ~ x,  family = binomial,  data = xplace)

Coefficients:
(Intercept)           x
   -3.53778      0.08114

Degrees of Freedom: 4 Total (i.e. Null);  3 Residual
Null Deviance:      82.14
Residual Deviance: 0.3171       AIC: 24.13

> summary(fmp)
Call:
glm(formula = Tmat ~ x,  family = binomial(link = probit),  data = xplace)

Deviance Residuals:
       1          2          3          4          5
-0.15582   0.02545   -0.08009   0.51246   -0.40097

Coefficients:
              Estimate     Std. Error     z value      Pr(>|z|)
(Intercept)   -2.102270     0.276287      -7.609       2.76e-14 ***
x              0.048147     0.005885       8.181       2.82e-16 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 82.14455  on 4  degrees of freedom
Residual deviance:  0.45473  on 3  degrees of freedom
AIC: 24.270

Number of Fisher Scoring iterations: 4

> summary(fml)

Call:
glm(formula = Tmat ~ x,  family = binomial,  data = xplace)

Deviance Residuals:
      1          2          3          4          5
-0.1797   0.1157  -0.1182   0.3791  -0.3372

Coefficients:
             Estimate   Std. Error   z value    Pr(>|z|)
(Intercept)  -3.53778   0.50232      -7.043     1.88e-12 ***
x            0.08114     0.01082     7.498  6.47e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 82.14455  on 4  degrees of freedom
Residual deviance:  0.31707  on 3  degrees of freedom
AIC: 24.132

Number of Fisher Scoring iterations: 4
```

Παράδειγμα της poisson οικογένειας

Η προκαθορισμένη μορφή της συνάρτησης σύνδεσης στην οικογένεια poisson είναι η **log**. Η χρήση της οικογένειας είναι για την προσαρμογή υποκατάστατων log linear μοντέλων για δεδομένα συχνοτήτων των οποίων η κατανομή είναι πολυωνυμική.

Πχ.

```
>fmod<-glm(y~A+B+x, family=poisson(link=sqrt), data=worm.counts)
```

Παράδειγμα από Quasi – likelihood (οιονεί πιθανοφάνειας)μοντέλα

Σε όλες τις οικογένειες, η διακύμανση της απόκρισης εξαρτάται από τον μέσο και θα έχει βαθμωτή παράμετρο σαν πολλαπλασιαστή. Η μορφή της εξάρτησης της διακύμανσης από τον μέσο είναι χαρακτηριστικό της κατανομής της απόκρισης. Για παράδειγμα στην κατανομή poisson είναι : $Var[y] = \mu$.

Στα οιωνεί πιθανοφάνειας μοντέλα (δηλαδή τα μοντέλα που μοιάζουν με μοντέλα πιθανοφάνειας), δεν μπορεί να καθοριστεί ακριβώς η κατανομή της απόκρισης για εκτίμηση ή συμπερασματολογία. Μπορούμε να έχουμε όμως μία συνάρτηση σύνδεσης και τη μορφή της συνάρτησης διακύμανσης που εξαρτάται από τον μέσο.

Επειδή η οιωνεί πιθανοφάνειας εκτίμηση χρησιμοποιεί ίδιες τεχνικές με αυτές της κανονικής κατανομής, μας εφοδιάζει με μεθόδους προσαρμογής μοντέλων κανονικής κατανομής με όχι – τυπική συνάρτηση σύνδεσης ή συνάρτηση διακύμανσης.

Πχ.

Θέλουμε να προσαρμόσουμε τη μη γραμμική συνάρτηση $y = \frac{\theta_1 z_1}{z_2 - \theta_2} + e$

Κάνοντας τους μετασχηματισμούς $x_1 = {z_2}/{z_1}, x_2 = -{1}/{z_1}, \beta_1 = {1}/{\theta_1}, \beta_2 = {\theta_2}/{\theta_1}$

μπορεί να γραφτεί σαν $y = \dfrac{1}{\beta_1 x_1 + \beta_2 x_2} + e$

Διαμορφώνοντας ένα κατάλληλο πλαίσιο δεδομένων μπορούμε να προσαρμόσουμε την παραπάνω μη γραμμική παλινδρόμηση με :

```
>nlfit<-glm(y~x1+x2-1. family=quasi(link=inverse, varianvce=constant), data=biochem)
```

6.4 Μη Γραμμικά μοντέλα ελαχίστων τετραγώνων και μέγιστης πιθανοφάνειας

Κάποιες μορφές μη γραμμικών μοντέλων μπορεί να προσαρμοστούν με γενικευμένα γραμμικά μοντέλα (glm). Η πλειονότητα των περιπτώσεων, όμως, απαιτεί να προσεγγίσουμε το πρόβλημα προσαρμογής της μη γραμμικής καμπύλης με μη γραμμική βελτιστοποίηση.

Οι ρουτίνες της R για μη γραμμική βελτιστοποίηση είναι οι optim() και nlm() καθώς επίσης και η nlminb().

Στη μη γραμμική βελτιστοποίηση υποθέτουμε αρχικές τιμές και το τελικό αποτέλεσμα εξαρτάται από την επιλογή αυτή. Δεν υπάρχει, επίσης, εγγύηση ότι η διαδικασία θα συγκλίνει σε ικανοποιητικές εκτιμήσεις.

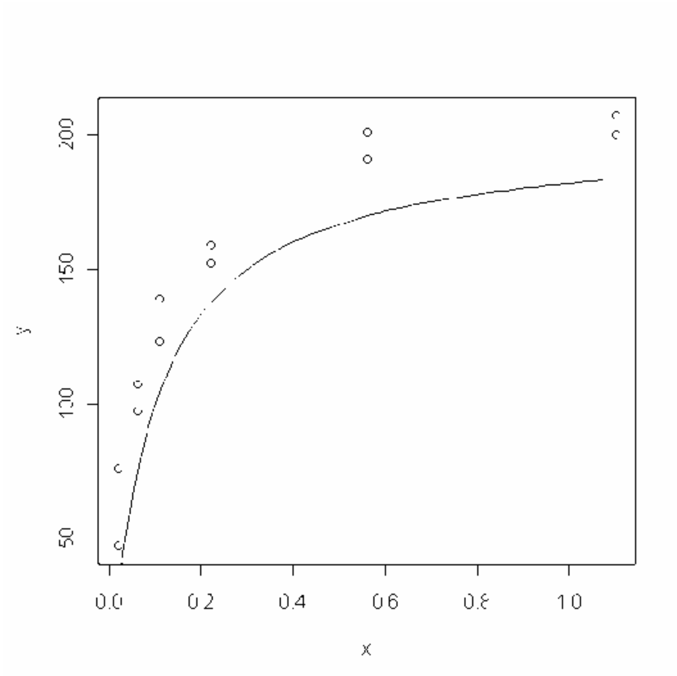
<u>Ελάχιστα τετράγωνα</u>

Η μέθοδος αυτή προσαρμόζει το μη γραμμικό μοντέλο με ελαχιστοποίηση του αθροίσματος των τετραγώνων των λαθών (SSE) ή καταλοίπων. Η μέθοδος αυτή έχει νόημα όταν τα παρατηρούμενα λάθη έχουν προκύψει αληθοφανώς από μία κανονική κατανομή.

Παράδειγμα

Θα υπολογίσουμε το αθροισμα τετραγώνων των διαφορών $y-\left(\frac{p1*x}{p2+x}\right)^{1/2}$ με τη βοήθεια της R.

```
> x<-c(0.02, 0.02, 0.06, 0.06, 0.11, 0.11, 0.22, 0.22, 0.56, 0.56, 1.10, 1.10)
> y<-c(76, 47, 97, 107, 123, 139, 159, 152, 191, 201, 207, 200)
> fn<-function(p) sum((y-(p[1]*x)/(p[2]+x))^2)
> fn
function(p) sum((y-(p[1]*x)/(p[2]+x))^2)
> xfit<-seq(0.02, 1.1, 0.05 )
> xfit
 [1] 0.02 0.07 0.12 0.17 0.22 0.27 0.32 0.37 0.42 0.47 0.52 0.57 0.62 0.67
0.72
[16] 0.77 0.82 0.87 0.92 0.97 1.02 1.07
> yfit<-200*xfit/(0.1+xfit)
> yfit
 [1]    33.33333    82.35294   109.09091   125.92593   137.50000   145.94595
152.38095
 [8]   157.44681   161.53846   164.91228   167.74194   170.14925   172.22222
174.02597
[15]   175.60976   177.01149   178.26087   179.38144   180.39216   181.30841
182.14286
[22] 182.90598
> plot(x, y)
> lines(spline(xfit, yfit))
```



Γράφημα 5 : Γράφημα προσαρμογής δεδομένων

```
> out<-nlm(fn, p=c(200, 0.1), hessian=TRUE)
> out
$minimum
[1] 1195.449

$estimate
[1] 212.68384222    0.06412146

$gradient
[1] -0.0001535005    0.0934207185

$hessian
            [, 1]             [, 2]
[1, ]    11.94725       -7661.319
[2, ] -7661.31875     8039421.153

$code
[1] 3
```

```
$iterations
[1] 26
```

Για να πάρουμε κατά προσέγγιση τα τυπικά σφάλματα κάνουμε το παρακάτω

```
> SEa<-sqrt(diag(2*out$minimum/(length(y)-2)*solve(out$hessian)))
> SEa
[1] 7.173465197      0.008744815
```

Μέγιστη πιθανοφάνεια

Η μέγιστη πιθανοφάνεια είναι μία μέθοδος για προσαρμογή μη γραμμικών μοντέλων η οποία εφαρμόζεται ακόμα και αν τα σφάλματα δεν είναι κανονικά. Η μέθοδος βρίσκει τις τιμές των παραμέτρων που μεγιστοποιούν την log likelihood ή ελαχιστοποιούν την αρνητική log-likelihood.

Πχ.

Για τις τιμές των διανυσμάτων x , y , n υπολογίζονται οι εκτιμητές μέγιστης πιθανοφάνειας που προσαρμόζει ένα λογιστικό μοντέλο που μπορεί να γίνει και με τη συνάρτηση glm().

```
> x<-c(1.6907, 1.7242, 1.7552, 1.7842, 1.8113, 1.8369, 1.8610, 1.8839)
> y<-c(6, 13, 18, 28, 52, 53, 61, 60)
> n<-c(59, 60, 62, 56, 63, 59, 62, 60)
>fn<-function(p) sum(-(y*(p[1]+p[2]*x)-
                        n*log(1+exp(p[1]+p[2]*x))+log(choose(n, y))))
> out<-nlm(fn, p=c(-50, 20), hessian=TRUE)
Warning message:
In nlm(fn,  p = c(-50,  20),  hessian = TRUE) :
  NA/Inf replaced by maximum positive value
```

```
> out
$minimum
[1] 18.71513

$estimate
[1] -60.71727  34.27021

$gradient
[1] 1.462810e-08 2.259956e-08

$hessian
          [, 1]     [, 2]
[1, ]  58.48405 103.9787
[2, ] 103.97873 184.9662

$code
[1] 1

$iterations
[1] 21

> sqrt(diag(solve(out$hessian)))
[1] 5.554694 3.123437
```

6.5 Μη τυποποιημένα μοντέλα

Στην R έχουμε και άλλες δυνατότητες για ειδικής μορφής παλινδρόμηση και ανάλυση δεδομένων. Παραθέτουμε συνοπτικά κάποιες από αυτές.

Μικτά μοντέλα : Συνίσταται το πακέτο nlme το οποίο διαθέτει τις συναρτήσεις lme(), nlme() για γραμμικά και μη γραμμικά μοντέλα με συντελεστές που αντιστοιχούν σε τυχαίες επιδράσεις.

Τοπική προσεγγιστική παλινδρόμηση : Η συνάρτηση **loess()** προσαρμόζει μία μη παραμετρική παλινδρόμηση χρησιμοποιώντας μία τοπικά σταθμισμένη παλινδρόμηση. Η συνάρτηση loess() βρίσκεται στο πακέτο stat. Ανθεκτική παλινδρόμηση : Υπάρχουν αρκετές συναρτήσεις διαθέσιμες για παλινδρόμηση που αντιμετωπίζει την επίδραση των ακραίων τιμών. Για το σκοπό αυτό στη βιβλιοθήκη MASS υπάρχει η συνάρτηση lqs() για ισχυρή αντίσταση στην επίδραση, ενώ η συνάρτηση rlm() από το ίδιο πακέτο για λιγότερη αντίσταση αλλά στατιστικά πιο αποτελεσματική.

Προσθετικά Μοντέλα : Η μέθοδος αυτή σκοπεύει να δημιουργήσει μία συνάρτηση παλινδρόμησης από απλές προσθετικές συναρτήσεις με καθοριστικές μεταβλητές. Οι συναρτήσεις **avas** και **ace** στο πακέτο acepack και οι συναρτήσεις **bruto** και **mars** στο πακέτο mda μας εφοδιάζει με τέτοιες τεχνικές. Μία επέκταση είναι τα Γενικευμένα Προσθετικά Μοντέλα που βρίσκονται στα πακέτα gam και mgcv.

Μοντέλα βασισμένα σε δένδρα : Τα μοντέλα που βασίζονται σε δένδρα ψάχνουν για διακλαδώσεις των δεδομένων, ώστε σε τελευταία ανάλυση να διαμοιράσουν τα δεδομένα σε ομάδες όπου μέσα στην ομάδα να είναι όσο το δυνατόν πιο ομογενή και αντίθετα μεταξύ των ομάδων όσο το δυνατόν πιο ετερογενή. Τα μοντέλα καθορίζονται στη μορφή του γραμμικού μοντέλου. Η συνάρτηση προσαρμογής είναι η tree(). Μοντέλα βασισμένα σε δένδρα βρίσκονται στα πακέτα rpart και tree.

ΜΕΡΟΣ ΔΕΥΤΕΡΟ - Γενικευμένα Γραμμικά Μοντέλα

7 ΕΙΣΑΓΩΓΗ στα Γενικευμένα γραμμικά μοντέλα

7.1 Γενικά

Τα γενικευμένα γραμμικά μοντέλα (Generalized Linear Models = GLM) αποτελούν σχετικά καινούργιο τομέα της στατιστικής. Η θεματολογία των γενικευμένων γραμμικών μοντέλων στο μεγαλύτερό της μέρος δεν αποτελεί κάτι νέο στην στατιστική, αλλά ουσιαστικά ομαδοποιεί έννοιες και τεχνικές που προϋπάρχουν δημιουργώντας ένα ενοποιημένο θεωρητικό και εννοιολογικό πλαίσιο.

Η κοινή αντιμετώπιση των θεμάτων δημιούργησε προϋποθέσεις για τη δημιουργία νέων τεχνικών και μεθόδων αντιμετώπισης προβλημάτων. Δημιουργήθηκαν επίσης οι προϋποθέσεις για την διερεύνηση εφαρμογής των μεθόδων αυτών, έστω με τροποποίηση, και σε θέματα για τα οποία μέχρι στιγμής δεν υπήρχε μέθοδος αντιμετώπισής τους.

Στα GLM συμπεριλαμβάνονται μεταξύ των άλλων και τα εξής :

- Η απλή και πολλαπλή γραμμική παλινδρόμηση.
- Το t – test.
- Ανάλυση διακύμανσης και συνδιακύμανσης.
- Λογιστική παλινδρόμηση.
- Log- linear μοντέλα.
- Πολυωνυμικά μοντέλα.
- Πίνακες συνάφειας.

Και αρκετές άλλες μεθόδους.

Τα παραπάνω μοιράζονται κάποιες κοινές ιδιότητες και έχουν κοινή μέθοδο εκτίμησης παραμέτρων.

7.2 Βασικές Έννοιες

7.2.1 Μεταβλητές

Οι τυχαίες μεταβλητές αποδίδονται με τους όρους : **εξαρτημένη μεταβλητή** (depended variable), **απόκριση** (response).
Οι τυχαίες μεταβλητές μεταβάλλονται ελεύθερα σε σχέση με άλλες μεταβλητές οι οποίες παίρνουν συγκεκριμένες τιμές και λέγονται **ανεξάρτητες** (independed variables) ή **επεξηγηματικές** (explanatory).
Οι επεξηγηματικές μεταβλητές δεν είναι τυχαίες μεταβλητές .

Είδη μεταβλητών

Συνεχείς, όπου οι τιμές των μεταβλητών βρίσκονται μέσα σε κάποιο διάστημα πραγματικών αριθμών.
Διακριτές, όπου οι τιμές είναι μία ακολουθία τιμών.

Οι διακριτές μεταβλητές συνήθως παίρνουν σαν τιμές τις συχνότητες των παρατηρήσεων σε κάποιες μετρήσεις.

Μία άλλη κατηγοριοποίηση είναι η διάκριση σε **ποσοτικές** και **ποιοτικές** μεταβλητές.
Οι ποιοτικές μεταβλητές είναι πάντα διακριτές.

Μία επεξηγηματική ποιοτική μεταβλητή λέγεται **παράγοντας (factor)** και αποτελείται **από επίπεδα (levels).**

Μία επεξηγηματική ποσοτική μεταβλητή λέγεται συμεταβλητή αλλά θα αποδίδεται απλά με τον όρο **μεταβλητή (covariate).**

Συμβολισμοί

Έστω μία συνάρτηση πυκνότητας πιθανότητας f μιας τυχαίας συνεχούς μεταβλητής Y (ή κατανομή πιθανότητας αν η Y είναι διακριτή). Εάν με θ συμβολίσουμε τις παραμέτρους της κατανομής τότε η κατανομή θα συμβολίζεται με :

$$f(y;\theta)$$

Το σύμβολο «;» θα χωρίζει μεταβλητές και παραμέτρους .

Το σύμβολο «.» θα χρησιμοποιείται σαν δείκτης για συνολικά αθροίσματα

πχ. έστω y_{ij} με $i=1,2,...,N$ και ο $j=1,...,K$ τότε το $y_{.j}=\sum_{i=1}^{N} y_{ij}$

7.2.2 Μοντελοποίηση

Η διαδικασία της μοντελοποίησης περιλαμβάνει τα παρακάτω βήματα :

Καθορισμός ευλόγων εξισώσεων και κατανομών για να περιγράψουν τα κύρια χαρακτηριστικά της εξαρτημένης μεταβλητής (καθορισμός του μοντέλου).
Εκτίμηση των παραμέτρων που χρησιμοποιούνται στα μοντέλα.
Συμπερασματολογία (δηλαδή, έλεγχο υποθέσεων, καθορισμό διαστημάτων εμπιστοσύνης και τέλος εκτίμηση της καλής προσαρμογής του μοντέλου).

7.3 Κατανομές προερχόμενες από την κανονική

Οι κατανομές που θα ορίσουμε συνοπτικά είναι, η κανονική κατανομή, η X^2 κατανομή, η κατανομή t και η F- κατανομή.

7.3.1 Κανονική κατανομή

Η μεταβλητή Y έχει τη κανονική κατανομή με μέσο μ και διασπορά σ^2 και θα συμβολίζεται $Y \sim N(\mu, \sigma^2)$ όταν έχει συνάρτηση πυκνότητας πιθανότητας την :

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

Η κανονική κατανομή με $\mu=0$ και $\sigma^2=1$ καλείται <u>*τυποποιημένη κανονική*</u> <u>κατανομή.</u>

Έστω ότι οι Y_1, Y_2, ...Y_ν είναι κανονικά κατανεμημένες τυχαίες μεταβλητέ ς με
$Y\iota \sim N(\mu_\iota, \sigma^2_{\iota\iota})$ για $\iota=1, 2,\ldots, \nu$.

Με $Cov(Y_i, Y_j)=\rho_{ij}\sigma_i\sigma_j$ συμβολίζουμε τη συνδιακύμανση των Y_i, Y_j.

Η από κοινού κατανομή των Y_i είναι η πολυδιάστατη κανονική κατανομή με διάνυσμα μέσων $\mu=[\mu_1,\mu_2,\ldots,\mu_\nu]^T$ και πίνακα διακυμάνσεων – συνδιακυμάνσεων V του οποίου τα στοιχεία είναι $\rho_{ij}\sigma_i\sigma_j$.

Η από κοινού κατανομή των Y_i θα συμβολίζεται με $Y \sim N(\mu, V)$ όπου :
$Y = [Y_1, Y_2,\ldots, Y_\nu]^T$.

Εάν οι τυχαίες μεταβλητές Y1, Y2, ..., Yν είναι ανεξάρτητες και κανονικά κατανεμημένες και ισχύει : $Y\iota \sim N(\mu_\iota, \sigma^2_\iota)$ τότε για κάθε γραμμικό συνδυασμό τους $U = \alpha_1 Y_1 + \alpha_2 Y_2 + \ldots + \alpha_\nu Y_\nu$, όπου α_ι σταθερές , θα ισχύουν τα παρακάτω :

η U είναι κανονική τμ.

- $E(U) = \alpha_1\mu_1 + \alpha_2\mu_2 + \ldots + \alpha_\nu\mu_\nu$
- $Var(U) = \alpha^2_1\sigma^2_1 + \alpha^2_2\sigma^2_2 + \ldots + \alpha^2_\nu\sigma^2_\nu$

$$U = \sum_{i=1}^{\nu} \alpha_\iota \Upsilon_\iota \quad \sim \; N\left(\sum_{\iota=1}^{\nu} \alpha_\iota \mu_\iota , \sum_{\iota=1}^{\nu} \alpha_\iota^2 \sigma_\iota^2 \right)$$

7.3.2 Κατανομές χ- τετράγωνο (X^2)

Έστω $Z_i \sim N(0, 1)$ για i=1,2,…n. Η κεντρική χ – τετράγωνο κατανομή με n βαθμούς ελευθερίας X^2 ορίζεται ως :

$$X^2 = \sum_{i=1}^{n} Z_i^2$$

Συμβολίζουμε : $X^2 \sim \chi^2_n$

Με μορφή πινάκων το $X^2 = \sum_{i=1}^{n} Z_i^2$ γράφεται : $X^2 = z^T z$ όπου $z = [Z_1, \ldots, Z_n]^T$.

Εάν οι τυχαίες μεταβλητές $Y_1, Y_2, \ldots, Y_n$ είναι ανεξάρτητες και κανονικά κατανεμημένες με $Y_\iota \sim N(\mu_\iota, \sigma^2_\iota)$ τότε :

$$X^2 = \sum_{i=1}^{n} \left(\frac{Y_i - \mu_i}{\sigma_i} \right)^2 \sim \chi^2_n$$

Γενίκευση (μη κεντρική X^2 κατανομή).

Έστω ότι οι Y_ι δεν είναι αναγκαστικά ανεξάρτητες και ότι το διάνυσμα $y = [Y_1, Y_2, \ldots, Y_n]^T$, έχει την πολυδιάστατη κανονική κατανομή ,δηλαδή $y \sim N(\mu, V)$.

Έστω ότι ο πίνακας διακύμανσης – συνδιακύμανσης V δεν είναι μοναδιαίος, οπότε υπάρχει ο αντίστροφός του που είναι V^{-1}.

Έστω $X^2 = (y - \mu)^T V^{-1} (y - \mu)$ Τότε : $X^2 \sim \chi^2_n$

Μη-κεντρική χ-τετράγωνο κατανομή με **n** βαθμούς ελευθερίας και παραμέτρους μη-κεντρικότητας $\lambda = \mu^T V^{-1} \mu / 2$ ονομάζεται η κατανομή της τυχαίας μεταβλητής $\mathbf{y}^T \boldsymbol{V}^{-1} \mathbf{y}$

Συμβολισμός : $y^T V^{-1} y \sim \chi^2(n, \lambda)$

Εάν $X^2_1, X^2_2, \ldots, X^2_m$ είναι ανεξάρτητες τυχαίες μεταβλητές με $X^2_\iota \sim \chi^2(n_\iota, \lambda_\iota)$ με ι=1,2,…,m τότε :

$$\sum_{\iota=1}^{m} X_\iota^2 \sim x^2\left(\sum_{\iota=1}^{m} n_\iota, \sum_{\iota=1}^{m} \lambda_\iota\right)$$

Περίπτωση V να είναι μοναδιαίος πίνακας

Έστω $y \sim N(\mu, V)$, όπου ο πίνακας διακύμανσης – συνδιακύμανσης V είναι μοναδιαίος με τάξη $k < n$ ώστε να μην ορίζεται μονοσήμαντα ο αντίστροφός του V^{-1}.
Συμβολίζουμε με V^- τον γενικευμένο αντίστροφο του V.

Η τυχαία μεταβλητή $y^T V^- y$ έχει την χ – τετράγωνο κατανομή με k βαθμούς ελευθερίας και παράμετρο μη – κεντρικότητας $\lambda = \mu^T V^- \mu / 2$.

Βασική ιδιότητα

Εάν οι τυχαίες μεταβλητές X_1, X_2 ακολουθούν την χ-τετράγωνο κατανομή με n_1, n_2 βαθμούς ελευθερίας αντίστοιχα με $n_1 > n_2$ και είναι ανεξάρτητες, τότε η διαφορά τους ακολουθεί χ-τετράγωνο κατανομή με $n_1 - n_2$ βαθμούς ελευθερίας.

Δηλαδή : $X^2_1 - X^2_2 \sim \chi^2_{n1 - n2}$

7.3.3 t – Κατανομές

Η t – κατανομή ορίζεται σαν το πηλίκο *T* όπου :

Αριθμητής είναι μία τυποποιημένη κανονική μεταβλητή **Z.**

Παρανομαστής είναι η τετραγωνική ρίζα μίας (κεντρική) $\mathbf{X^2}$ κατανομής με n βαθμούς ελευθερίας διαιρεμένης με τους βαθμούς ελευθερίας της n.

Οι Z και $\mathbf{X^2}$ είναι ανεξάρτητες.

$$T = \frac{Z}{\left(X^2/n\right)^{1/2}}$$

συμβολίζουμε : $T \sim t_n$

7.3.4 F – Κατανομές

Η (κεντρική) F – κατανομή ορίζεται σαν το πηλίκο δύο τυχαίων ανεξάρτητων μεταβλητών (κεντρικών) X^2, διαιρεμένων κάθε μία με τους βαθμούς ελευθερίας τους ως εξής :

$$F = \frac{X_1^2/n}{X_2^2/m}$$

Όπου $X^2_1 \sim \chi^2_n$ και $X^2_2 \sim \chi^2_m$

Συμβολίζουμε : $F \sim F_{n,\, m}$

<u>Σχέση μεταξύ t και F κατανομής.</u>

Μπορεί να αποδειχθεί ότι ισχύει το εξής : $T = \dfrac{Z}{\sqrt{\dfrac{X^2}{n}}} \Rightarrow T^2 = \dfrac{Z^2/1}{X^2/n} \sim F_{1,\, n}$

Η F μη κεντρική κατανομή ορίζεται σαν το πηλίκο δύο ανεξάρτητων $\mathbf{X^2}$ κατανομών όπου στον αριθμητή η κατανομή είναι **μη κεντρική** και στον παρανομαστή **κεντρική.**

Δηλαδή έχουμε την ακόλουθη σχέση :

$$F = \frac{X_1^2 / n}{X_2^2 / m}$$

Όπου $X^2_1 \sim \chi^2(n, \lambda)$ και $X^2_2 \sim \chi^2_m$

$\mathbf{X^2_1}$, $\mathbf{X^2_2}$ ανεξάρτητες.

7.4 Το γραμμικό μοντέλο

Το κλασσικό γραμμικό μοντέλο έχει σε μορφή πινάκων τη μορφή :

$$\mathbf{Y = X\beta + e}$$

Όπου Y είναι το διάνυσμα των αποκρίσεων.
β είναι το διάνυσμα των παραμέτρων.
X πίνακας που περιέχει μόνο τιμές που είναι ένα, μηδέν ή τιμές ανεξάρτητων μεταβλητών.
e είναι το διάνυσμα των τυχαίων λαθών.

και τα στοιχεία *e* υποτίθεται ότι είναι ανεξάρτητα και ομοιόμορφα κατανεμημένα (independent and identically distributed) και ακολουθούν την κανονική κατανομή $N(0, \sigma^2)$.

Σε ποσοτικές επεξηγηματικές μεταβλητές, το X β είναι ο όρος όπου το *β* παριστάνει το ρυθμό μεταβολής της απόκρισης που αντιστοιχεί στις αλλαγές της ανεξάρτητης μεταβλητής *X*.
Στις ποιοτικές επεξηγηματικές μεταβλητές υπάρχει μια παράμετρος η οποία παριστάνει το κάθε επίπεδο (*level*) ενός παράγοντα (*Factor*).Τα αντίστοιχα στοιχεία επιλέγονται για να αποκλίσουν ή να συμπεριλάβουν τις κατάλληλες παραμέτρους για κάθε παρατήρηση και καλούνται ψευδομεταβλητές (*dummy* μεταβλητές).
Το υπόδειγμα αυτό αποτελεί τη βάση για τις περισσότερες αναλύσεις συνεχών δεδομένων.

Για παράδειγμα, η σύγκριση δύο μέσων και η σχέση μιας συνεχούς εξαρτημένης μεταβλητής και μίας συνεχούς ανεξάρτητης (covariate) σε δύο ομάδες χρησιμοποιεί τέτοιου είδους υπόδειγμα. Γενικεύσεις που υπάρχουν για τη σύγκριση περισσότερων των δύο μέσων (ανάλυση διακύμανσης) και η σχέση μεταξύ μιας συνεχούς εξαρτημένης και κάποιων επεξηγηματικών (πολλαπλή παλινδρόμηση) είναι επίσης αυτής της μορφής.

7.5 Η εκθετική οικογένεια κατανομών

Έστω μια τυχαία μεταβλητή *Y* της οποίας η συνάρτηση πιθανότητας (αν είναι διακριτή) ή η συνάρτηση πυκνότητας πιθανότητας (αν είναι συνεχής) , μπορεί να εκφραστεί ως :

$$f(y) = \exp\left\{\frac{y\theta - b(\theta)}{\alpha(\phi)} + c(y,\phi)\right\}$$

όπου *a, b, c* είναι γνωστές συναρτήσεις που καθορίζουν τη συγκεκριμένη κατανομή.

Θα λέμε ότι η *Y* ανήκει στην εκθετική οικογένεια κατανομών.

Το *θ* ονομάζεται κανονική παράμετρος και είναι παράμετρος κεντρικής τάσης . Το *φ* είναι παράμετρος κλίμακας (scale) και ονομάζεται παράμετρος διασποράς.

Για **σταθερό *φ*** έχουμε την εκθετική οικογένεια κατανομών μίας παραμέτρου που θα μας απασχολήσει στο μεγαλύτερο τμήμα της μελέτης αυτής.

Οι συναρτήσεις *a* και *c* είναι τέτοιες ώστε :

$$\alpha_i(\phi) = \frac{\phi}{w_i}$$

όπου w_i είναι το γνωστό βάρος για κάθε παρατήρηση . Συνήθως το *w* έχει τη τιμή 1 (ένα).

Η θεωρία για αυτό το είδος των κατανομών (της εκθετικής οικογένειας) δίνει εκφράσεις για το μέσο και τη διασπορά του Y. Οι εκφράσεις είναι:

$$E(Y) = \mu = b'(\theta)$$

$$\operatorname{var}(Y) = \frac{\phi}{w_i} b''(\theta)$$

Οι κατανομές πιθανότητας της απόκρισης Y παραμετροποιούνται συνήθως σε σχέση με τη μέση τιμή *μ* και τη διασπορά *φ*.

Παράδειγμα μέλους της εκθετικής οικογένειας αποτελεί η κανονική κατανομή.

Στην κανονική κατανομή ισχύουν :

$$f(y;\mu) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{-\frac{1}{2\sigma^2}(y-\mu)^2\}$$

Επειδή είναι : $(y-\mu)^2 = y^2 + \mu^2 - 2y\mu$

και $\frac{1}{(2\pi\sigma^2)^{1/2}} = (2\pi\sigma^2)^{-1/2} = -\frac{1}{2}\log\left(2\pi\sigma^2\right)$

η συνάρτηση της κανονικής κατανομής μπορεί να γραφτεί :

$$f(y) = \exp\left\{\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\lambda o\gamma\left(2\pi\sigma^2\right)\right\}$$

Οπότε : $\theta = \mu \quad \Rightarrow \quad b(\theta) = \frac{1}{2}\theta^2$

και τέλος $c(y,\phi) = -\frac{y^2}{2\sigma^2} - \frac{1}{2}\lambda o\gamma\left(2\pi\sigma^2\right)$ με $\phi = \sigma^2$

<u>Παρατήρηση</u>

Όμοια αποδυκνείεται και για άλλες κατανομές (όπως η διωνυμική , η Γάμμα ,η Poisson κλπ.) ότι ανήκουν στην εκθετική οικογένεια .

7.6 Η μονοπαραμετρική εκθετική οικογένεια κατανομών

Έστω μια απλή τυχαία μεταβλητή Y της οποίας η συνάρτηση πιθανότητας (αν είναι διακριτή) ή η συνάρτηση πυκνότητας πιθανότητας (αν είναι συνεχής) εξαρτάται από μία παράμετρο θ.

Ορίζουμε ότι η κατανομή ανήκει στην (μονοπαραμετρική) εκθετική οικογένεια αν μπορεί να γραφτεί στη μορφή :

$$\mathbf{f(y;\theta) = s(y)\, t(\theta)\, e^{a(y)\, b(\theta)}}$$

όπου a, b, s, t είναι γνωστές συναρτήσεις.

Η προηγούμενη συνάρτηση μπορεί να γραφτεί στη μορφή :

$$f(y;\theta) = \exp[a(y)\, b(\theta) + c(\theta) + d(y)]$$

όπου $s(y) = \exp[d(y)]$ και $t(\theta) = \exp[c(\theta)]$

Παρατηρούμε τη συμμετρία που υπάρχει μεταξύ y και θ

Εάν $a(y) = y$ τότε λέμε ότι η κατανομή είναι σε **κανονική μορφή** και το $b(\theta)$ ονομάζεται «η φυσική παράμετρος» της κατανομής.

Εάν υπάρχουν και άλλες παράμετροι, επιπρόσθετα της παραμέτρου θ, τότε αυτές θα θεωρούνται σαν τμήματα των συναρτήσεων a, b, c και d και τις οποίες θα μεταχειριζόμαστε σαν να είναι γνωστές.

Στην εκθετική οικογένεια ανήκουν πολλές γνωστές κατανομές. Η Poisson , η κανονική και η διωνυμική κατανομή ανήκουν στην εκθετική οικογένεια διότι μπορούν να γραφτούν στην κανονική μορφή όπως θα δείξουμε παρακάτω.

7.6.1 Κατανομή Poisson

Για τη διακριτή τυχαία μεταβλητή Y έχουμε :

$$f(y,\lambda) = \frac{\lambda^y e^{-\lambda}}{y!} \qquad \text{οπού } y=0,\ 1,\ 2,\ 3,\ \ldots$$

Η μορφή της f μπορεί να γραφτεί ξανά ως εξής :

f(y;λ) = exp [y logλ –λ – logy!]

η οποία είναι στην κανονική μορφή όπου η φυσική παράμετρος είναι η logλ.

7.6.2 Κανονική κατανομή

Η πυκνότητα πιθανότητας είναι η ακόλουθη :

$$f(y;\mu) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{-\frac{1}{2\sigma^2}(y-\mu)^2\}$$

Την παραπάνω συνάρτηση μπορούμε να την γράψουμε ως εξής (χρησιμοποιώντας την ιδιότητα $\alpha = e^{\log\alpha}$) :

$$f(y;\mu) = \exp\{-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\}$$

Όπου η παραπάνω μορφή είναι κανονική μορφή .

Η φυσική παράμετρος είναι $b(\mu) = \mu/\sigma^2$ και οι άλλοι όροι είναι :

$$c(\mu) = -\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) \quad \text{και} \qquad d(y) = -\frac{y^2}{2\sigma^2}$$

7.6.3 Η Διωνυμική κατανομή

‘Εστω ένα πείραμα στο οποίο μπορεί να έχουμε μόνο δύο δυνατά αποτελέσματα που είναι : «επιτυχία» και «αποτυχία».

Εάν υποθέσουμε ότι η τυχαία μεταβλητή Y παριστάνει τον αριθμό των «επιτυχιών» σε n ανεξάρτητες δοκιμές όπου η πιθανότητα επιτυχί ας είναι η ίδια σε όλες τις προσπάθειες και παραστήσουμε την πιθανότητα αυτή με π , τότε λέμε ότι η Y έχει την δυωνυμική κατανομή με συνάρτηση πιθανότητας :

$$f(y;\pi)=\binom{n}{y}\pi^{y}(1-\pi)^{n-y} \quad \text{όπου } y=0,\ 1,\ 2,\ \ldots,\ n$$

Συμβολικά γράφουμε *Y ~ b(n;π)*

Στον προηγούμενο συμβολισμό το **π** είναι η παράμετρος ενώ το **n** θεωρείται γνωστό.

Χρησιμοποιώντας τις ιδιότητες της λογαριθμικής και εκθετικής συνάρτησης μπορούμε να γράψουμε το παρακάτω:

$$f(y;\pi)=e^{\log[\binom{n}{y}\pi^{y}(1-\pi)^{n-y}]}=\exp\{y\log\pi-y\log(1-\pi)+n\log(1-\pi)+\log\binom{n}{y}\}=$$

$$=\exp\{y\log\frac{\pi}{1-\pi}+n\log\left(1-\pi\right)+\log\binom{n}{y}\}$$

Το οποίο είναι στην κανονική μορφή με φυσική παράμετρο το $\frac{\pi}{1-\pi}$.

7.6.4 Συμπεράσματα και επεκτάσεις

Για τις προηγούμενες τρεις κατανομές συνοψίζουμε τα αποτελέσματα στον παρακάτω πίνακα.

Κατανομή : Poisson

Έχει **Φυσική παράμετρο** $\log\lambda$ και **c**=-*λ* , **d** $-\log y!$

Κατανομή : Κανονική

Έχει **Φυσική παράμετρο** μ/σ^2 και **c**= $-\frac{1}{2}\frac{\mu^2}{\sigma^2}-\frac{1}{2}\log\left(2\pi\sigma^2\right)$, **d**= $-\frac{1}{2}\frac{y^2}{\sigma^2}$

Κατανομή : Διωνυμική

Έχει **Φυσική παράμετρο** $\log\left(\frac{\pi}{1-\pi}\right)$ και **c=** $n\log(1-\pi)$, **d=** $\lambda o\gamma\binom{n}{y}$

Υπάρχουν και άλλες κατανομές οι οποίες ανήκουν στην εκθετική οικογένεια . Δεν έχουν όλες την κανονική μορφή, αλλά παρ' όλα αυτά έχουν ενδιαφέρουσες ιδιότητες.

Το ζητούμενο είναι να βρούμε εκφράσεις οι οποίες θα υπολογίζουν την αναμενόμενη τιμή και τη διασπορά για το *α(y)* . Για να μπορέσουμε να προχωρήσουμε στους υπολογισμούς των εκφράσεων αυτών θα πρέπει να στηριχτούμε στους παρακάτω ορισμούς και θεωρήματα.

Θεώρημα 1

Έστω l η συνάρτηση της λογαριθμικής πιθανοφάνειας και U η πρώτη παράγωγος ως προς θ (Δηλαδή : $U = dl/d\theta$).

Τότε

Για κάθε κατανομή ισχύει ότι [1] :

(Χωρίς Απόδειξη)

$E[U] = 0$ και $Var(U)=E(U^2)= E(-U')$ **(1)** (όπου $U' = \mathbf{d^2 \mathit{l}/d\theta^2}$)

Ορισμοί

Το U που ορίστηκε ως $U = dl/d\theta$ ονομάζεται **score** και το Var(U) ονομάζεται **πληροφορία**.

Μπορούμε να εφαρμόσουμε τα αποτελέσματα αυτά για κατανομές στην εκθετική οικογένεια.

Από τη σχέση $f(y;\theta) = \exp[a(y)\, b(\theta) + c(\theta) + d(y)]$ έχουμε τις παρακάτω συνέπειες :

$l = \log f = a(y)b(\theta)+ c(\theta) +d(y)$

οπότε $U = dl/d\theta = a(y)\, b'(\theta) + c'(\theta)$

και $U' = d^2l/d\theta^2 = a(y)\, b''(\theta)+ c''(\theta)$

<u>Θεώρημα 2</u>

Εάν η τυχαία μεταβλητή Y ανήκει στην εκθετική οικογένεια τότε η αναμενόμενη τιμή και τη διασπορά για το α(y) είναι :

$E[a(Y)] = - c'(\theta)/b'(\theta)$

$Var[a(Y)] = [-b''(\theta)\ c'(\theta) - c''(\theta)\, b'(\theta)] / [b'(\theta)]^3$

Απόδειξη

Για την αναμενόμενη τιμή έχουμε :

$E(U) = E[a(Y)\ b'(\theta) + c'(\theta)] = b'(\theta)\ E[a(Y)] + c'(\theta)$

Και επειδή $E(U) = 0$ προκύπτει ότι $E[a(Y)] = - c'(\theta) / b'(\theta)$ **(2)**

Για τη διασπορά ισχύει

$Var(U) = [b'(\theta)]^2\ Var[a(Y)]$ και $E[- U'] = - b''(\theta)\ E[a(Y)] - c''(\theta)$

Οπότε λόγω της **(1)** προκύπτει :

$Var[a(Y)] = [-b''(\theta)\ E[a(Y)] - c''(\theta)] / [b'(\theta)]^2$ =>

$$\mathbf{Var[a(Y)] = [-b''(\theta)\ c'(\theta) - c''(\theta)\ b'(\theta)] / [b'(\theta)]^3} \quad \mathbf{(3)}$$

Μπορούμε να επιβεβαιώσουμε εύκολα ότι οι σχέσεις **(2), (3)** ισχύουν για την Poisson τη κανονική και τη διωνυμική κατανομή .

<u>Θεώρημα 3</u>

Όταν η τυχαία μεταβλητή Y ανήκει στην εκθετική (μονοπαραμετρική) οικογένεια και είναι στη τυπική της μορφή τότε ισχύουν:

$$\mathbf{E[Y] = - c'(\theta)}$$

$$\mathbf{Var[Y] = - c''(\theta)}$$

Η απόδειξη προκύπτει από το συνδυασμό των θεωρημάτων **1** και **2**.

7.6.5 Πολυδιάστατες κατανομές

Έστω Y_1, Y_2, …, Y_N ανεξάρτητες τυχαίες μεταβλητές οι οποίες έχουν την ίδια κατανομή που δίνεται από τη σχέση : f(y;θ) = exp[a(y) b(θ) +c(θ) + d(y).

Τότε η από κοινού συνάρτηση πυκνότητας πιθανότητας δίνεται από την παρακάτω συνάρτηση.

$$f(y_1, y_2, ..., y_N) = \prod_{i=1}^{N} \exp[b(\theta)a(y_i) + c(\theta) + d(y_i)] =$$

$$= \exp[b(\theta)\sum_{i=1}^{N} a(y_i) + Nc(\theta) + \sum_{i=1}^{N} d(y_i)]$$

Ο όρος $\sum_{i=1}^{N} a(y_i)$ θεωρείται μία επαρκής στατιστική συνάρτηση για το $b(\theta)$. Αυτό σημαίνει ότι, με κάποια βεβαιότητα το $\sum_{i=1}^{N} a(y_i)$ συνοψίζει όλη τη διαθέσιμη πληροφορία για τη παράμετρο θ, πράγμα που είναι πολύ σημαντικό για την εκτίμηση παραμέτρων [2].

Στη συνέχεια θα μελετήσουμε μια κλάση μοντέλων τα οποία στηρίζονται στην εκθετική οικογένεια κατανομών.

8 Γενικευμένα Γραμμικά Μοντέλα

Η πρόοδος στη στατιστική θεωρία μαζί με την ανάπτυξη των υπολογιστών μας επέτρεψαν να δημιουργήσουμε μεθόδους ανάλογους με αυτές που έχουν αναπτυχθεί για τα γραμμικά μοντέλα σε περιπτώσεις που οι αποκρίσεις ακολουθούν κατανομή διαφορετική από την κανονική, δεν είναι απαραίτητα συνεχείς (μπορεί να είναι κατηγορικές μεταβλητές) και δεν χρειάζεται να είναι στην απλή γραμμική μορφή : $y = X\beta + e$.

Μια σημαντική ανακάλυψη είναι ότι πολλές από τις χρήσιμες ιδιότητες της κανονικής κατανομής κατέχει η ομάδα κατανομών που ανήκει στην «εκθετική οικογένεια».

Τα πακέτα λογισμικού αποτελούν το βασικό εργαλείο για υπολογισμούς παραμέτρων. Η εκτίμηση παραμέτρων του γραμμικού μοντέλου $X\beta$ επεκτάθηκε στην εκτίμηση γραμμικών συνδυασμών του τύπου $g(X\beta)$.

Θεωρητικά οι διαδικασίες εκτίμησης είναι απλές. Στη πράξη απαιτούν ένα μεγάλο όγκο υπολογισμών οι οποίοι έγιναν εφικτοί μόνο μέσω υπολογιστών με τη βοήθεια αριθμητικών προσεγγίσεων μη γραμμικών συναρτήσεων.

Θα δούμε στη συνέχεια πως επεκτείνονται τα κλασσικά γραμμικά μοντέλα σε γενικευμένα γραμμικά μοντέλα.

Έστω ένα διάνυσμα $y=(y_1,y_2,\ldots,y_n)$ από n στοιχεία μιας τυχαίας μεταβλητής Y της οποίας οι συνιστώσες είναι ανεξάρτητα κατανεμημένες με μέσους $\mu=(\mu_1,\mu_2,\ldots,\mu_n)$.

Το συστηματικό μέρος του μοντέλου είναι ο καθορισμός του διανύσματος μ με ένα μικρό αριθμό άγνωστων παραμέτρων $\beta_1, \beta_2, \ldots, \beta_p$.

Στη περίπτωση των συνηθισμένων γραμμικών μοντέλων ο καθορισμός έχει τη μορφή $\eta = \sum_{i}^{\rho} x_i \beta_j$, όπου τα β_i είναι άγνωστοι παράμετροι που πρέπ ει να εκτιμηθούν από τα δεδομένα. Ολοκληρώσαμε τον καθορισμό του συστηματικού μέρους του μοντέλου.

Για το στοχαστικό μέρος υποθέτουμε ότι υπάρχει ανεξαρτησία και σταθερή διακύμανση λαθών. Θα πρέπει επίσης τα λάθη να ακολουθούν μία κανονική κατανομή. Εάν οι παρατηρήσεις συμβολιστούν με Y_i τότε μπορούμε να γράψουμε : $E(Y_i) = \mu_i$ όπου $i=1, \dots, n$

Στο γραμμικό μοντέλο τα *η* και *μ* συνδέονται με τη σχέση *μ=η*

Εάν τα x_{ij} είναι η τιμή της *j* μεταβλητής για την *i* παρατήρηση, τότε για τα γραμμικά μοντέλα έχουμε :

$$E(Y_i) = \mu_i = \sum_{i=1}^{p} x_{ij} \beta_\xi$$

Με συμβολισμό πινάκων έχουμε $\mu = X\beta$

όπου το *μ* είναι *nx1* το *X* είναι *nxp* και το *β* είναι *px1*.

<u>**Μπορούμε λοιπόν να θεωρήσουμε ότι το γραμμικό μοντέλο αποτελείται από τρία μέρη :**</u>

1. **Το στοχαστικό μέρος** : οι συνιστώσες του Y έχουν ανεξάρτητες κανονικές κατανομές με $E(Y) = \mu$ και σταθερή διασπορά σ^2.
2. **Το συστηματικό μέρος** : οι μεταβλητές $x_1, x_2, \dots, x_p$ δημιουργούν μία γραμμική πρόβλεψη του *η* που δίνεται από : $\eta = \sum_{1}^{p} x_j \beta_j$.
3. **Μία σύνδεση** μεταξύ του συστηματικού και του στοχαστικού μέρους *μ=η*.

Η γενίκευση μπορεί να γίνει σε δύο φάσεις. **Στην πρώτη φάση** εισάγουμε μία νέα θεώρηση του τρίτου μέρους. Θέτουμε :

$$\eta_i = g(\mu_i)$$

Όπου η συνάρτηση g ονομάζεται συνάρτηση σύνδεσης και είναι γνήσια μονότονη και διαφορίσιμη.

Σε δεύτερη φάση μπορούμε να έχουμε κατανομές *Y* διαφορετικές της κανονικής που προέρχονται την «εκθετική οικογένεια».

Το 1972 παρουσιάστηκε από τους Nelder & Wedderburn η ιδέα των γενικευμένων γραμμικών μοντέλων που στηρίχθηκε στις προηγούμενες ιδέες.

Ένα Γενικευμένο Γραμμικό Μοντέλο ορίζεται από ένα σύνολο ανεξάρτητων τυχαίων μεταβλητών $Y_1, \ldots, Y_N$ όπου η καθεμιά τους ακολουθεί μία κατανομή από την εκθετική οικογένεια και έχει τις παρακάτω ιδιότητες.

Η κατανομή για κάθε Y_ι είναι από την ίδια κατανομή που προέρχεται από την εκθετική οικογένεια, είναι στην κανονική μορφή της και εξαρτάται από μία μοναδική παράμετρο θ_ι (δεν απαιτείται τα θ_ι να είναι ίδια) δηλαδή :

$$f(y_i ; \theta_i) = \exp \{y_i b_i(\theta_i) + c_i(\theta_i) + d_i(y_i)\}$$

Υπάρχει συνάρτηση g και ένα μικρότερο σύνολο παραμέτρων $\beta_1, \beta_2, \ldots, \beta_\rho$ $(\rho<N)$ τέτοιο ώστε ένας γραμμικός συνδυασμός των β_ι να είναι ίσος με τη συνάρτηση της αναμενόμενης τιμής μ_ι των Y_ι, δηλαδή:

$$\mathbf{g(\mu_i) = x^T \beta} \quad \text{όπου}$$

g είναι μονότονη και διαφορίσιμη συνάρτηση που τη λέμε *συνάρτηση σύνδεσης* (link function).

$\mathbf{x_i}$ **px1** διάνυσμα από επεξηγηματικές μεταβλητές (covariates και dummy μεταβλητές για επίπεδα παραγόντων).

β είναι το **px1** διάνυσμα των παραμέτρων.

Παρατήρηση

Επειδή όλες οι μεταβλητές προέρχονται από την ίδια κατανομή, η από κοινού πυκνότητα πιθανότητας είναι :

$$f(y_1,...,y_N;\theta_1,...,\theta_N) = \exp\left\{\sum_{i=1}^{N} y_i b(\theta_\iota) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(y_i)\right\}$$

Οι παράμετροοι θ_ι δεν είναι του έμμεσου ενδιαφέροντος για την εξειδίκευση του μοντέλου.

Συνεπώς ένα γενικευμένο γραμμικό μοντέλο έχει τρεις συνιστώσες.

Μεταβλητές απόκρισης $Y_1, Y_2, \ldots, Y_N$ που υποθέτουμε ότι μοιράζονται την ίδια κατανομή από την εκθετική οικογένεια.

Ένα σύνολο από παραμέτρους β και επεξηγηματικές μεταβλητές X

$$\beta = \begin{pmatrix} \beta 1 \\ . \\ . \\ . \\ \beta_\rho \end{pmatrix} , \quad X = \begin{pmatrix} x_1^T \\ x_2^T \\ . \\ . \\ . \\ x_N^T \end{pmatrix}$$

Μία γνήσια μονότονη και διαφορίσιμη συνάρτηση σύνδεσης g τέτοια ώστε :

$$g(\mu_i)=x_i^T\boldsymbol{\beta}$$

Όπου : $\boldsymbol{\mu_i=E(Y_\iota)}$

Παρατήρηση

Μία ειδική περίπτωση του Γενικευμένου Γραμμικού Μοντέλου είναι το Γραμμικό μοντέλο :

$y = x\beta + e$ όπου τα e ακολουθούν την $N(0, \sigma^2)$.

Εδώ $y_i \sim N(\mu_\iota, \sigma^2)$ όπου $g(\mu_\iota)=\mu_\iota$, δηλαδή η g είναι η ταυτοτική.

8.1 ΕΚΤΙΜΗΤΙΚΗ

Για τις εκτιμήσεις παραμέτρων στη στατιστική χρησιμοποιούνται συνήθως δύο προσεγγίσεις που είναι η «μέθοδος της μέγιστης πιθανοφάνειας» και η «μέθοδος των ελαχίστων τετραγώνων» . Υπάρχει και μία παραλλαγή της μεθόδου των ελαχίστων τετραγώνων η οποία ονομάζεται «μέθοδος των σταθμισμένων ελαχίστων τετραγώνων».

Θα ξεκινήσουμε με μία επισκόπηση των αρχών που διέπουν καθεμιά από αυτές τις μεθόδους και θα συζητήσουμε κάποιες ιδιότητες των εκτιμητών. Οι εκτιμήσεις συνήθως υπολογίζονται αριθμητικά από επαναληπτικές διαδικασίες, οι οποίες είναι πολύ κοντά με τη μέθοδο εκτίμησης των «σταθμισμένων ελαχίστων τετραγώνων».

8.1.1 Μέθοδος Μέγιστης Πιθανοφάνειας

Έστω $Y_1, Y_2, \ldots, Y_N$ τυχαίες μεταβλητές με από κοινού συνάρτηση πυκνότητας πιθανότητας $f(y_1, y_2, \ldots y_N; \theta_1, \ldots, \theta_p)$ η οποία εξαρτάται από τις παραμέτρους $\theta_1, \theta_2, \ldots \theta_p$.

Εάν τώρα θεωρήσουμε ότι στη συνάρτηση f τα $y_1, y_2, \ldots, y_N$ είναι σταθερά και τα $\theta 1, \theta 2, \ldots, \theta p$ είναι μεταβλητές , τότε έχουμε τη συνάρτηση πιθανοφάνειας (likelihood) : $L(\theta_1, \ldots, \theta_p; y_1, y_2, \ldots y_N)$.

Με μορφή πινάκων η συνάρτηση πιθανοφάνειας γράφεται ως : $L(\theta;y)$,

$$\text{Όπου} \quad \begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_N \end{pmatrix} = y \quad \text{και} \quad \begin{pmatrix} \theta_1 \\ . \\ . \\ . \\ \theta_p \end{pmatrix} = \theta$$

Παρατηρούμε ότι η συνάρτηση πιθανοφάνειας L(θ;y) είναι αλγεβρικά ίδια με τη συνάρτηση f(y;θ). Η αλλαγή στον συμβολισμό γίνεται για να δοθεί έμφαση στην μετατόπιση από τη τυχαία μεταβλητή y με σταθερό το θ , στη μεταβλητή θ με σταθερό το y.

Έστω Ω το σύνολο των δυνατών τιμών του διανύσματος των παραμέτρων θ. Ο εκτιμητής της μέγιστης πιθανοφάνειας του θ είναι η τιμή $\hat{\theta}$ η οποία μεγιστοποιεί τη συνάρτηση πιθανοφάνειας.
Δηλαδή ισχύει :

$$L(\hat{\theta};y) \geq L(\theta;y) \text{ για κάθε } \theta \text{ στο } \Omega.$$

Επειδή η λογαριθμική συνάρτηση είναι μονότονη, μπορούμε ισοδύναμα να υπολογίσουμε την τιμή $\hat{\theta}$ σαν τη τιμή που μεγιστοποιεί τον λογάριθμο της συνάρτησης πιθανοφάνειας :

$$l(\theta;y) = logL(\theta;y).$$

Υπολογίζουμε λοιπόν το $\hat{\theta}$ σαν τη τιμή που ισχύει : $l(\hat{\theta};y) \geq l(\theta;y)$ για κάθε θ στο Ω.

Ο εκτιμητής $\hat{\theta}$ λαμβάνεται συνήθως με διαφόριση της λογαριθμικής συνάρτησης πιθανοφάνειας για καθένα από τα θ_j του διανύσματος $\boldsymbol{\theta}$ και στη συνέχεια λύνοντας το σύστημα εξισώσεων :

$$\frac{\partial l(\theta;y)}{\partial \theta_j} = 0 \qquad \text{για } j = 1, 2, \ldots, \rho$$

Για να αντιστοιχούν οι λύσεις πράγματι σε μέγιστες τιμές του $l(\theta;y)$ θα πρέπει ο πίνακας των δεύτερων παραγώγων $\frac{\partial^2 l(\theta;y)}{\partial \theta_j \partial \theta_k}$ να είναι αρνητικά ορισμένος για την τιμή $\theta = \hat{\theta}$.

<u>Παρατηρήσεις</u>

1) Εάν υπάρχει μόνο μία παράμετρος θ πρέπει να ελέγξουμε ότι για $\theta = \hat{\theta}$ η παράγωγος : $\dfrac{\partial^2 l(\theta;y)}{\partial\theta^2}$ είναι αρνητική.

2) Πρέπει να ελέγξουμε για την ύπαρξη τιμών της παραμέτρου θ στα άκρα του χώρου παραμέτρων Ω που να δίνουν τοπικά μέγιστα για τη συνάρτηση $l(\theta;y)$. Εάν υπάρχουν , η τιμή του $\hat{\theta}$ που αντιστοιχεί στο μεγαλύτερο από όλα τα μέγιστα , είναι ο εκτιμητής μέγιστης πίθανοφάνειας.

3) Οι εκτιμητές μέγιστης πιθανοφάνειας έχουν την ακόλουθη σημαντική ιδιότητα. *Εάν g(θ) είναι μία συνάρτηση της παραμέτρου θ , τότε ο εκτιμητής μέγιστης πιθανοφάνειας του g(θ) είναι* $g(\hat{\theta})$. Συνεπώς μπορούμε να εκτιμήσουμε τη μέγιστη πιθανοφάνεια με οποιαδήποτε συνάρτηση των παραμέτρων είναι πιο βολική και στη συνέχεια με τη χρήση της παραπάνω ιδιότητας να εξάγουμε εκτιμήσεις για τις ζητούμενες παρ αμέτρους.

4) Ιδιότητες που χαρακτηρίζουν τους εκτιμητές μέγιστης πιθανοφάνειας είναι μεταξύ των άλλων η **συνέπεια**, η **επάρκεια** και η **ασυμπτωτική ικανότητα**.

8.1.2 Μέθοδος των Ελαχίστων τετραγώνων

Η μέθοδος των ελαχίστων τετραγώνων συνίσταται στην εύρεση εκτ ιμητών $\hat{\beta}$ (που συμβολίζονται και με b) που ελαχιστοποιούν το άθροισμα των τετραγώνων των όρων σφαλμάτων $\boldsymbol{e_i}$.

Δηλαδή οι εκτιμητές ελαχίστων τετραγώνων προκύπτουν από την ελαχιστοποίηση της συνάρτησης : $S = \sum e_i^2$

Ας υποθέσουμε ότι οι τυχαίες μεταβλητές $Y_1, Y_2, \ldots, Y_N$ έχουν τιμές $E(Y_i) = \mu_i, \quad i = 1,2,\ldots,N$.

Έστω ότι και τα $\boldsymbol{\mu_i}$ είναι συναρτήσεις των παραμέτρων $\beta_1, \beta_2, \ldots, \beta_p$ $(p < N)$ που πρέπει να εκτιμηθούν.

Έστω ότι είναι : $\beta = \left(\beta_1,...,\beta_\rho\right)^T$

Θεωρούμε ότι ισχύει ο τύπος $Y_i = \mu_i + e_i$, για i=1,…,N. Θέλουμε να ελαχιστοποιήσουμε το *S* όπου :

$$S = \sum e_i^2 = \sum \left[Y_i - \mu_i(\beta)\right]^2$$

Το ανωτέρω γράφεται σε μορφή πινάκων ως εξής : $S = (y - \mu)^T(y - \mu)$

Όπου $y = \begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_N \end{pmatrix}$ και $\mu = \begin{pmatrix} \mu_1 \\ . \\ . \\ . \\ \mu_N \end{pmatrix}$

Οι εκτιμητές $\hat{\beta}$ λαμβάνονται από παραγώγιση του ***S*** ως προς κάθε στοιχείο β_j του $\boldsymbol{\beta}$ και λύνοντας στη συνέχεια το σύστημα εξισώσεων :

$$\frac{dS}{d\beta_j} = 0 \quad \text{για} \quad j = 1,\ldots,\rho.$$

Για να αντιστοιχούν οι λύσεις σε ελάχιστα θα πρέπει ο πίνακας των δεύτερων παραγώγων είναι θετικά ορισμένος. Εάν υπάρχουν και τοπικά ελάχιστα τότε η μικρότερη τιμή των ελαχίστων είναι η ζητούμενη.

8.1.3 Σταθμισμένα Ελάχιστα τετράγωνα

Πολλές φορές υπάρχει επιπρόσθετη πληροφορία για τα Y_i, όπως για παράδειγμα κάποιες παρατηρήσεις να είναι λιγότερο αξιόπιστες διότι έχουν μεγαλύτερη διακύμανση από άλλες. Για να εισάγουμε την επιπρόσθετη πληροφορία στους υπολογισμούς, αποδίδουμε βάρη στους όρους και ελαχιστοποιήσουμε το άθροισμα : $S_w = \sum w_i \left[Y_i - \mu_i(\beta)\right]^2$

όπου τα w_i παριστάνουν βάρη (πχ. $w_i=[var(Y_i)]^{-1}$).

Στη γενική περίπτωση τα Y_i, μπορεί να είναι συσχετισμένα. Εάν με ***V*** συμβολίζεται ο πίνακας διακύμανσης – συνδιακύμανσης , τότε οι εκτιμητές των σταθμισμένων ελαχίστων τετραγώνων προκύπτουν από την ελαχιστοποίηση του S_w όπου :

$$\mathbf{S_w = (y - \mu)^T} \boldsymbol{V}^{\mathbf{-1}} \mathbf{(y - \mu)}$$

Στην ειδική περίπτωση που οι όροι μ_i είναι γραμμικοί συνδυασμοί των παραμέτρων β_j (j=1,…,ρ με ρ<N) δηλαδή $\mu = X\beta$ για κάποιο $N \times p$ πίνακα X τότε $S_w = \left(y - X\beta\right)^T V^{-1}\left(y - X\beta\right)$

Οι παράγωγοι του S_w ως προς τα στοιχεία β_j του διανύσματος β είναι το διάνυσμα : $\frac{\partial S_w}{\partial \beta} = -2X^T V^{-1}(y - X\beta)$

Θέλουμε να ισχύει $\frac{\partial S_w}{\partial \beta} = 0$ οπότε $2X^T V^{-1}(y - X\beta) = 0$

συνεπώς : $X^T V^{-1} X b = X^T V^{-1} y$

Οπότε οι εκτιμητές των σταθμισμένων ελαχίστων τετραγώνων b του διανύσματος παραμέτρων β είναι η λύση των κανονικών εξισώσεων :

$$X^T V^{-1} X b = X^T V^{-1} y$$

(Υπό την προϋπόθεση ότι ο πίνακας των δευτέρων παραγώγων είναι θετικά ορισμένος).

Παρατηρήσεις

1) Σημαντική διαφορά μεταξύ των μεθόδων των ελαχίστων τετραγώνων και της μέγιστης πιθανοφάνειας είναι ότι τα ελάχιστα τετράγωνα μπορεί να χρησιμοποιηθούν χωρίς να κάνουμε υποθέσεις για τις κατανομές των

μεταβλητών απόκρισης Y_i. Αρκεί ο καθορισμός των αναμενόμενων τιμών και πιθανόν την δομή της διακύμανσης – συνδιακύμανσης.

2) Στη μέθοδο μέγιστης πιθανοφάνειας χρειάζεται να καθορίσουμε την από κοινού πυκνότητα πιθανότητας των Y_i.

3) Για να εξάγουμε την δειγματική κατανομή των εκτιμητών ελαχίστων τετραγώνων, χρειάζονται επιπρόσθετες προϋποθέσεις για τα Y_i.

8.2 Εκτιμήσεις για Γενικευμένα Γραμμικά Μοντέλα

Οι υπολογισμοί των εκτιμητών γίνεται με δύο αριθμητικές μεθόδους. Τη μέθοδο Newton – Raphson και τη μέθοδο των score.

Εκτίμηση με χρήση μεθόδου Newton – Raphson

Η μέθοδος που θα χρησιμοποιηθεί είναι η πολυδιάστατη έκδοση της μεθόδου Newton – Raphson για την εύρεση των λύσεων της εξίσωσης μίας μεταβλητής $f(x) = 0$ όπου για την περίπτωση αυτή οι προσεγγίσεις των λύσεων δίνονται από την αναδρομική σχέση :

$$x^{(m)} = x^{(m-1)} - \frac{f\left[x^{(m-1)}\right]}{f'\left[x^{(m-1)}\right]}$$

Θέλουμε να υπολογίσουμε τους εκτιμητές των παραμέτρων β για τα γενικευμένα γραμμικά μοντέλα που περιγράφονται από

$$f(y;\theta) = \exp\left\{\sum_{i=1}^{N} y_i b(\theta_\iota) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(y_i)\right\}$$

Η λογαριθμική συνάρτηση πιθανοφάνειας για ανεξάρτητες αποκρίσεις $Y_1,\ldots,Y_N$ είναι :

$$l(\theta; y) = \sum y_i b(\theta_i) + \sum c(\theta_i) + \sum d(y_i)$$

για τις οποίες ισχύουν επίσης :

$E(Y_i) = \mu_i = -\dfrac{c'(\theta_i)}{b'(\theta_i)}$ και $g(\mu_i) = x_i^T\beta = \eta_i$

όπου g είναι μονότονη και παραγωγίσιμη συνάρτηση.

Επειδή οι κατανομές ανήκουν στην εκθετική οικογέν εια, αποδεικνύεται ότι το ολικό μέγιστο της λογαριθμικής συνάρτησης πιθανοφάνειας $l(\theta;y)$ δίνεται μοναδικά από τη λύση του συστήματος εξισώσεων $\dfrac{\partial l}{\partial \theta} = 0$ ή ισοδύναμα από τη λύση των $\dfrac{\partial l}{\partial \beta} = 0$.

Μπορεί να αποδειχθεί [3] ότι : $\dfrac{\partial l}{\partial \beta_j} = U_j = \sum_{i=1}^{N} \dfrac{(y_i - \mu_i)x_{ij}}{Var(Y_i)} \left(\dfrac{\partial \mu_i}{\partial \eta_i} \right)$,

όπου τα x_{ij} είναι το *j – στοιχείο* του $x^T_{i.}$

Στη γενική τους περίπτωση οι εξισώσεις $U_j = 0$, j = 1, 2, …, p δεν είναι γραμμικές και συνεπώς πρέπει να λυθούν με μία επαναληπτική αριθμητική μέθοδο.

Μία επαναληπτική μέθοδος που υπολογίζει προσεγγιστικά τις λύσεις είναι η επέκταση για πολλές διαστάσεις της μεθόδου Newton – Raphson.

Η μέθοδος Newton – Raphson (για πολλές διαστάσεις) δίνει τη *m–οστή* προσέγγιση από τη σχέση :

$$b^{(m)} = b^{(m-1)} - \left[\frac{\partial^2 l}{\partial \beta_j \beta_k} \right]^{-1}_{\beta = b^{(m-1)}} U^{(m-1)}$$

όπου $\left[\dfrac{\partial^2 l}{\partial\beta_j \beta_k}\right]^{-1}_{\beta=b^{(m-1)}}$ είναι ο πίνακας των δευτέρων παραγώγων του l για την τιμή $\boldsymbol{\beta} = \boldsymbol{b}^{(m-1)}$ και $\mathbf{U}^{(m-1)}$ είναι το διάνυσμα των πρώτων παραγώγων $U_j = \partial l / \partial\beta_j$ για $\boldsymbol{\beta} = \boldsymbol{b}^{(m-1)}$.

<u>Εκτίμηση με χρήση της μεθόδου του score</u>

Μία εναλλακτική διαδικασία, η οποία κάποιες φορές είναι πιο απλή από τη μέθοδο Newton – Raphson , είναι η ονομαζόμενη «Μέθοδος (σκορ)».

Η «Μέθοδος του σκορ» προκύπτει αντικαθιστώντας τον πίνακα των δευτέρων παραγώγων με τον πίνακα των αναμενό μενων τιμών $E\left[\dfrac{\partial^2 l}{\partial\beta_j \partial\beta_\kappa}\right]$

Αποδεικνύεται ότι το $E\left[\dfrac{\partial^2 l}{\partial\beta_j \partial\beta_\kappa}\right]$ είναι ίσο με τον αρνητικό πίνακα διακυμάνσεων – συνδιακυμάνσεων των U_j , ο οποίος καλείται και <u>*πίνακας πληροφορίας*</u>.

Ο πίνακας πληροφορίας συμβολίζεται με $\boldsymbol{J} = \mathbf{E}[\mathbf{U}\mathbf{U}^{\mathbf{T}}]$ και έχει στοιχεία :

$$J_{jk} = E[\mathrm{U_j U_k}] = E\left[\frac{\partial l}{\partial\beta_j}\frac{\partial l}{\partial\beta_k}\right] = -\ E\left[\frac{\partial^2 l}{\partial\beta_j \partial\beta_\kappa}\right]$$

Από τα προηγούμενα προκύπτει ότι η εξίσωση :

$$b^{(m)} = b^{(m-1)} - \left[\frac{\partial^2 l}{\partial\beta_j \beta_k}\right]^{-1}_{\beta=b^{(m-1)}} U^{(m-1)}$$

μπορεί να γραφτεί ως εξής :

$$b^{(m)} = b^{(m-1)} + \left[j^{(m-1)}\right]^{-1} U^{(m-1)}$$

όπου ο όρος $J^{(m-1)}$ είναι ο πίνακας πληροφορίας που υπολογίζεται στο $b^{(m-1)}$.

Πολλαπλασιάζουμε τη προηγούμενη σχέση κατά μέλη με $J^{(m-1)}$ και λαμβάνουμε την επαναληπτική εξίσωση .

$$j^{(m-1)}b^{(m)} = j^{(m-1)}b^{(m-1)} + U^{(m-1)} \quad \textbf{(Σ)}$$

Αποδεικνύεται[4] ότι για τα γενικευμένα γραμμικά μοντέλα το *(j, k)* στοιχείο του *J* είναι :

$$j_{jk} = \sum_{i=1}^{N} \frac{x_{ij}x_{ik}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_\iota} \right)^2$$

Συνεπώς το J μπορεί να γραφτεί ως εξής : $\boldsymbol{J = X^T W X}$

Όπου το *W* είναι *NxN* διαγώνιος πίνακας με στοιχεία : $w_{ii} = \frac{1}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$

Η έκφραση στο δεξιό μέρος της **(Σ)** είναι διάνυσμα με στοιχεία :

$$\sum_k \sum_i \frac{x_{ij}x_{ik}}{\mathrm{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 b_k^{(m-1)} + \sum_i \frac{(y_i - \mu_i)x_{ij}}{\mathrm{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)$$

η οποία υπολογίζεται στο $b^{(m-1)}$.

Από τα παραπάνω προκύπτει ότι το δεξιό μέρος της **(Σ)** μπορεί να γραφτεί σαν $\boldsymbol{X^T W z}$ όπου το $\boldsymbol{z}$ παίρνει τις τιμές $z_i = \sum_k x_{ik} b_k^{(m-1)} + (y_i - \mu_i) \left(\frac{\partial \eta_i}{\partial \mu_i} \right)$ όπου τα $\boldsymbol{\mu_i}$ και $\frac{\partial \eta_i}{\partial \mu_i}$ υπολογίζονται στο $b^{(m-1)}$.

Συνεπώς η επαναληπτική εξίσωση για τη μέθοδο του ***score*** μπορεί να γραφεί ως εξής :

$$X^T W X b^{(m)} = X^T W z \quad \textbf{(Σ1)}$$

Η προηγούμενη εξίσωση έχει την ίδια μορφή με τις κανονικές εξισώσεις των γενικευμένων γραμμικών μοντέλων που προκύπτουν από τα σταθμισμένα ελάχιστα τετράγωνα με την διαφορά του ότι πρέπει να λυθούν με μία επαναληπτική μέθοδο επειδή τα ***z*** και ***W*** εξαρτώνται σε γενικές γραμμές από το ***b***.

Συμπεραίνουμε από τα παραπάνω ότι, οι εκτιμητές μέγιστης πιθανοφάνειας των γενικευμένων γραμμικών μοντέλων προκύπτουν από μία επαναληπτική διαδικασία σταθμισμένων ελαχίστων τετραγώνων.

Χρησιμοποιώντας υπολογιστή, θα πρέπει να παρατηρήσουμε ότι ο υπολογιστής χρειάζεται να λύσει συνήθως την **(Σ1)** (δηλαδή την $X^T W X b^{(m)} = X^T W z$).

Τα περισσότερα πακέτα λογισμικού που περιλαμβάνουν αναλύσεις βασισμένες σε γενικευμένα γραμμικά μοντέλα διαθέτουν αποτελεσματικά προγράμματα για να υπολογίσουν τις λύσεις.

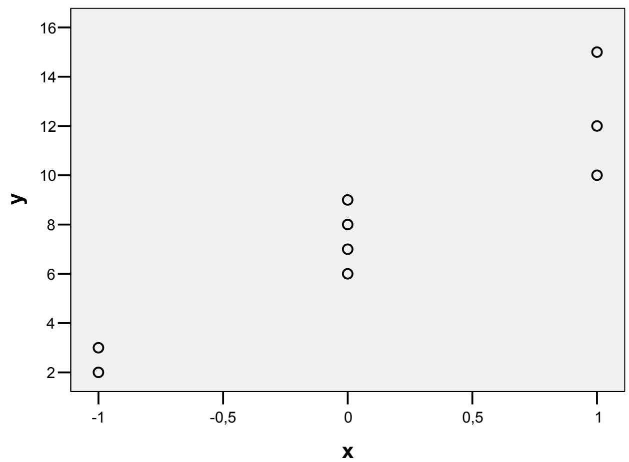
Τα προγράμματα αυτά ξεκινούν χρησιμοποιώντας μία αρχική προσέγγιση $b^{(0)}$ για να υπολογίσουν τα ***z*** και ***W***.
Στη συνέχεια η **(Σ1)** επιλύεται για να μας δώσει το $b^{(1)}$ το οποίο χρησιμοποιείται για να προκύψουν καλύτερες προσεγγίσεις για τα ***z*** και **W**.
Η διαδικασία αυτή επαναλαμβάνεται μέχρις ότου να επιτευχθεί ικανοποιητική σύγκλιση (προσέγγιση).
Όταν η διαφορά μεταξύ δύο διαδοχικών προσεγγίσεων $b^{(m)}$ και $b^{(m-1)}$ είναι επαρκώς μικρή τότε σαν εκτιμητής μέγιστης πιθανοφάνειας λαμβάνεται ο $b^{(m)}$.

Παράδειγμα 1

Τα δεδομένα του παρακάτω πίνακα είναι μετρήσεις y_i που παρατηρήθηκαν σε διάφορες τιμές μιας μεταβλητής x. Στη συνέχεια παρατίθεται το γράφημά τους.

Πίνακας 6 : Πίνακας τιμών (y_i ,x_i)

y_i	2	3	6	7	8	9	10	12	15
x_i	-1	-1	0	0	0	0	1	1	1



Γράφημα 6 : Γράφημα πίνακα τιμών (y_i ,x_i)

Ας υποθέσουμε ότι οι αποκρίσεις Y_i είναι τυχαίες μεταβλητές **Poisson**. Στην πράξη, μία τέτοια υπόθεση γίνεται συνήθως από την παρατήρηση ότι η μεταβλητότητα αυξάνει με το Y.

Στη κατανομή Poisson, η αναμενόμενη τιμή και διασπορά του Y_i είναι ίσες.

$$E(Y_i) = var(Y_i).$$

Αν υποθέσουμε ότι η σχέση μεταξύ των Y_i και x_i είναι γραμμική τότε :

$$\mathbf{E[Y_i] = \mu_i = \beta_1 + \beta_2 x_i = \chi_i^T \beta}$$

όπου i = 1, 2…, N (για το παράδειγμα που έχουμε είναι N = 9)

και $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$ και $\chi_i = \begin{pmatrix} 1 \\ x_i \end{pmatrix}$

Μπορούμε λοιπόν να πάρουμε σαν συνάρτηση σύνδεσης **$g(\mu_i)$** την ταυτοτική, οπότε έχουμε :

$$g(\mu_i) = \mu_i = \chi_i^T\beta = \eta_i \quad \Rightarrow \quad \frac{\partial \mu_i}{\partial \eta_i} = 1$$

Η σχέση αυτή απλοποιεί τις εξισώσεις διότι από τις σχέσεις :

$$w_{ii} = \frac{1}{\operatorname{var}(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \quad \text{και} \quad z_i = \sum_k x_{ik} b_k^{(m-1)} + \left(y_i - \mu_i\right)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)$$

προκύπτει ότι :

$$w_{ii} = \frac{1}{\mu_i} = \frac{1}{\beta_1 + \beta_2 x_i} \quad \text{και} \quad z_i = b_1 + b_2 x_i + \left(y_i - b_1 - b_2 x_i\right) = y_i$$

Οπότε χρησιμοποιώντας την εκτίμηση b για το β έχουμε :

$$J = X^T W X = \begin{pmatrix} \sum_{i=1}^{N} \frac{1}{b_1 + b_2 x_i} & \sum_{i=1}^{N} \frac{x_i}{b_1 + b_2 x_i} \\ \sum_{i=1}^{N} \frac{x_i}{b_1 + b_2 x_i} & \sum_{i=1}^{N} \frac{x_i^2}{b_1 + b_2 x_i} \end{pmatrix} \qquad \textbf{(T1)}$$

και $$X^T W z = \begin{pmatrix} \sum_{i=1}^{N} \frac{y_i}{b_1 + b_2 x_i} \\ \sum_{i=1}^{N} \frac{x_i y_i}{b_1 + b_2 x_i} \end{pmatrix} \qquad \text{(T2)}$$

Οι εκτιμητές μέγιστης πιθανοφάνειας προκύπτουν από τις επαναληπτικές εξισώσεις : $\left(X^T W X\right)^{(m-1)} b^{(m)} = \left(X^T W z\right)^{(m-1)}$ **(T3)**

Όπου ο εκθέτης (m-1) δηλώνει τον υπολογισμό στο $b^{(m-1)}$.

*Λύση του προβλήματος χρησιμοποιώντας **R** (πρώτος τρόπος)*

Για το παράδειγμά μας έχουμε αρχικές τιμές $b_1^{(0)} = 7$ και $b_2^{(0)} = 5$.

Για m = 3 με διαδοχικές προσεγγίσεις έχουμε την προσέγγιση $b_1^{(2)} = 7.4516$ και $b_2^{(2)} = 4.9353$. που προκύπτει από την εφαρμογή σε R εφαρμόζοντας τους τύπους **(T1)**, **(T2)**, **(T3).**

Η συνάρτηση prosegisi_b παίρνει σαν όρισμα τις τιμές b_1 ,b_2 υπολογίζει τους τύπους (T1), (T2) και δίνει αποτέλεσμα τις νέες τιμές των b_1 ,b_2. Θα χρησιμοποιηθεί για τον υπολογισμό των διαδοχικών προσεγγιστικών τιμών $b^{(m-1)}$.

```
> prosegisi_b<-function(b1, b2){
+ a11<-sum(1/(b1+b2*x))
+  a12<-sum(x/(b1+b2*x))
+  a21<-a12
+  a22<-sum(x^2/(b1+b2*x))
+  c1<-sum(y/(b1+b2*x))
+  c2<-sum(x*y/(b1+b2*x))
+  J<-rbind(c(a11, a12), c(a21, a22))
+  K<-rbind(c1, c2)
+  b<-solve(J, K)
+  b1<-b[1]
+  b2<-b[2]
+  return(b)}
```

Δίνουμε αρχικές τιμές στα b_1 ,b_2 και υπολογίζουμε επαναληπτικά με τη βοήθεια της συνάρτησης τις νέες προσεγγίσεις των b_1 ,b_2.

```
>  b<-c(7, 5)
>  b1<-b[1]
>  b2<-b[2]
>  for(i in 1:3){
+ b<- prosegisi_b(b1, b2)
+  b1<-b[1]
+  b2<-b[2]  }
> b1
[1] 7.451633
> b2
[1] 4.9353
```

Λύση του προβλήματος με με R (δεύτερος τρόπος)

Για την εκτίμηση του ***b*** μπορούμε να χρησιμοποιήσουμε τις έτοιμες συναρτήσεις της R. Στην R μπορούμε να κάνουμε τις εκτιμήσεις με τη βοήθεια της συνάρτησης **glm**. Η χρήση της συνάρτησης glm μας δίνει επίσης επιπλέον αποτελέσματα.

```
> y<-c(2, 3, 6, 7, 8, 9, 10, 12, 15)
> x<-c(-1, -1, 0, 0, 0, 0, 1, 1, 1)
> dedos<-data.frame(y=y, x=x)
> preg<-glm(y~x, family=poisson(link=identity), data=dedos)
> preg

Call:  glm(formula = y ~ x,  family = poisson(link = identity),  data = dedos)

Coefficients:
(Intercept)          x
     7.452       4.935

Degrees of Freedom: 8 Total (i.e. Null);  7 Residual
Null Deviance:      18.42
Residual Deviance: 1.895        AIC: 40.01
```

9 Συμπερασματολογία

Θα περιγράψουμε τις δειγματικές κατανομές των εκτιμητών και των στατιστικών συναρτήσεων που μετράνε τη καλή προσαρμογή και θα δείξουμε πως χρησιμοποιούνται για να κάνουμε συμπερασματολογία.

Στη γενική περίπτωση το πρόβλημα της εύρεσης κατανομών με ακρίβεια είναι δυσεπίλυτο και βασιζόμαστε στα ασυμπτωτικά αποτελέσματα μεγάλων δειγμάτων. Η εφαρμογή των αποτελεσμάτων όμως, πρέπει να γίνεται με προσοχή και εφόσον πληρούνται κάποιες προϋποθέσεις.

Στην ειδική περίπτωση των γραμμικών μοντέλων που η απόκριση ακολουθεί κανονική κατανομή (κανονικά κατανεμημένη απόκριση), η δειγματική κατανομή μπορεί να καθοριστεί επακριβώς.

Για ανεξάρτητες παρατηρήσεις από κατανομές που ανήκουν στην εκθετική οικογένεια και ιδιαίτερα για γενικευμένα γραμμικά μοντέλα οι αναγκαίες προϋποθέσεις ικανοποιούνται. Θα μελετήσουμε τα βασικά σημεία που προκύπτουν εξετάζοντας τις δειγματικές κατανομές.

<u>Θεωρητικό υπόβαθρο</u>

Εάν $\hat{\theta}$ είναι μία **συνεπής** εκτιμήτρια της παραμέτρου θ και var($\hat{\theta}$) είναι η διακύμανση της εκτιμήτριας, τότε για μεγάλα δείγματα ισχύουν τουλάχιστον κατά προσέγγιση τα ακόλουθα αποτελέσματα :

Επειδή το $\hat{\theta}$ είναι συνεπής εκτιμήτρια, το E($\hat{\theta}$) προσεγγίζει το θ όσο μεγαλώνει το δείγμα. Συνεπώς το Το $\hat{\theta}$ είναι μία αμερόληπτη εκτιμήτρια.

Η στατιστική συνάρτηση $\frac{\hat{\theta}-\theta}{\sqrt{\operatorname{var}(\hat{\theta})}}$ ακολουθεί την κατανομή *N(0, 1).*

Το τετράγωνο του στατιστικού $\frac{\left(\hat{\theta}-\theta\right)^2}{\operatorname{var}(\hat{\theta})} \sim x_1^2$ (κατανομή x^2 με ένα βαθμό ελευθερίας).

Γενίκευση για p παραμέτρους.

Έστω $\boldsymbol{\theta}$ ένα διάνυσμα με ***p*** παραμέτρους και το $\hat{\theta}$ είναι συνεπής εκτιμήτρια του $\boldsymbol{\theta}$. Συμβολίζουμε με ***V*** τον πίνακα διακύμανσης – συνδιακύμανσης για το $\hat{\theta}$. Τότε το $\hat{\theta}$ είναι μία αμερόληπτη εκτιμήτρια του θ και εφ’ όσον ο πίνακας ***V*** είναι μη μοναδιαίος , η δειγματική κατανομή από τον ορισμό της κεντρικής X^2 κατανομής είναι : $\left(\hat{\theta}-\theta\right)^T V^{-1}\left(\hat{\theta}-\theta\right) \sim x_p^2$.

Εάν ο πίνακας των διακυμάνσεων – συνδιακυμάνσεων είναι μοναδιαίος τότε δεν έχει μοναδική αντίστροφο. Στην περίπτωση αυτή έχουμε δύο προσεγγίσεις για να δουλέψουμε.
(Υποθέτουμε ότι η τάξη του *V* είναι q < p).

Στη πρώτη προσέγγιση λαμβάνουμε μία γενικευμένη αντίστροφο V^- (δηλαδή ένα πίνακα με την ιδιότητα $VV^-V = V$). Στη περίπτωση αυτή μπορεί να δειχθεί ότι ασυμπτωτικά $\left(\hat{\theta}-\theta\right)V^-\left(\hat{\theta}-\theta\right) \sim x_q^2$.

Στη δεύτερη προσέγγιση είναι να εκφράσουμε το μοντέλο ξανά σε σχέση με ένα νέο διάνυσμα παραμέτρων *φ* μεγέθους q τέτοιο ώστε ο πίνακας διακυμάνσεων – συνδιακυμάνσεων του *φ* έστω *W* να είναι μοναδιαίος συνεπώς : $\left(\hat{\phi}-\phi\right)^T W^{-1}\left(\hat{\phi}-\phi\right) \sim x_q^2$.

Στη συνέχεια θα δημιουργήσουμε δειγματικές κατανομές για τις ακόλουθες στατιστικές συναρτήσεις :

- Το score $U_j = \dfrac{\partial l}{\partial \beta_j}$.
- τους εκτιμητές μέγιστης πιθανοφάνειας b_j .
- και για μια στατιστική συνάρτηση καλής προσαρμογής που προέρχεται από το τεστ του πηλίκου πιθανοφάνειας.

9.1 Δειγματική κατανομή για τη στατιστική συνάρτηση score

Στην παράγραφο 4.6.4 ορίσαμε τις έννοιες «*score*» και «*πληροφορία*». Η στατιστική συνάρτηση *score* που αντιστοιχεί σε μία παράμετρο β_j ορίζεται ως η παράγωγος της λογαριθμικής συνάρτησης πιθανοφάνειας ως προς β_j. Συνεπώς για το διάνυσμα β των p παραμέτρων τα *scores* είναι :

$$U_j = \frac{\partial l}{\partial \beta_j}, \quad \text{όπου } j = 1, 2, \ldots, p.$$

Στην ίδια παράγραφο δείξαμε ότι για τα γενικευμένα γραμμικά μοντέλα ισχύουν : $E(U_j) = 0$ για όλα τα j.

Γενικεύουμε την έννοια της πληροφορίας σε πολυδιάστατη κατανομή και έχουμε τον πίνακα πληροφορίας : $Var(U) = E(UU^T)$.

Από τους ορισμούς προκύπτει λοιπόν ότι ο πίνακας πληροφορίας J είναι ο πίνακας διακύμανσης – συνδιακύμανσης των U_j .

Δηλαδή $\boldsymbol{E(UU^T) = J}$, όπου $U = \left(U_1, U_2, \ldots, U_p\right)^T$.

Από το κεντρικό οριακό θεώρημα U έχει, τουλάχιστον ασυμπτωτικά, την πολυδιάστατη κανονική κατανομή $U \sim N(0,\ J)$ και συνεπώς θα έχουμε :

$$U^T J U \sim x_p^2 \qquad \textbf{(Σ2)}$$

με την προϋπόθεση να μην είναι η J μοναδιαία ώστε να υπάρχει J^{-1}.

Θα υπολογίσουμε την στατιστική συνάρτηση *score* για την κανονική και τη δυωνυμική κατανομή.

<u>Στατιστική συνάρτηση score για την κανονική κατανομή</u>

Έστω ότι οι τυχαίες μεταβλητές $Y_1, Y_2, \ldots, Y_N$ είναι ανεξάρτητες και ισόνομα κατανεμημένες και ακολουθούν την κανονική κατανομή $N(\mu, \sigma^2)$ όπου το σ^2 είναι γνωστό και σταθερό. Σε αυτή την περίπτωση στα γενικευμένα γραμμικά μοντέλα υπάρχει μόνο μία παράμετρος η οποία μας ενδιαφέρει που είναι το $\boldsymbol{\mu}$ και η συνάρτηση σύνδεσης είναι η **ταυτοτική**.

Η λογαριθμική συνάρτηση πιθανοφάνειας είναι :

$l(\mu; y_1, \ldots, y_N) = -\dfrac{1}{2\sigma^2}\sum (y_i - \mu)^2 - N\log\left[\sigma\sqrt{2\pi}\right]$ οπότε έχουμε

$$U = \frac{dl}{d\mu} = \frac{1}{\sigma^2}\sum (y_i - \mu)$$

άρα η στατιστική συνάρτηση score είναι : $U = \dfrac{1}{\sigma^2}\sum (y_i - \mu) = \dfrac{N}{\sigma^2}(\bar{y} - \mu)$

Είναι εύκολο να δούμε ότι ισχύει E(U) = 0 διότι E($\bar{Y}$) = μ.

Έχουμε ότι $J = \mathrm{var}(U) = \dfrac{N^2}{\sigma^4}\mathrm{var}(\bar{Y}) = \dfrac{N^2}{\sigma^4}\dfrac{\sigma^2}{N} = \dfrac{N}{\sigma^2}$ επειδή $\mathrm{var}(\bar{Y}) = \dfrac{\sigma^2}{N}$.

Συνεπώς η πληροφορία δίνεται από τον τύπο $J = \dfrac{N}{\sigma^2}$.

Για τη στατιστική συνάρτηση $U^T J U$ ισχύει (σύμφωνα με τα παραπάνω) :

$$U^T J U = \left[\frac{N(\bar{y} - \mu)}{\sigma^2}\right]^2 \frac{\sigma^2}{N} = \frac{(\bar{Y} - \mu)^2}{\sigma^2 / N}$$

Επειδή $\bar{Y} \sim N(\mu, \sigma^2 / N) \Rightarrow \dfrac{(\bar{Y} - \mu)}{\sigma / \sqrt{N}} \sim N(0,1) \Rightarrow \dfrac{(\bar{Y} - \mu)^2}{\sigma^2 / N} \sim x_1^2$

Δηλαδή $U^T J^{-1} U \sim x_1^2$

Τα ανωτέρω στατιστικά μπορούν να χρησιμοποιηθούν για συμπερασματολογία.

<u>Στατιστική συνάρτηση score για την διωνυμική κατανομή</u>

Έστω ότι η μεταβλητή Υ ακολουθεί την διωνυμική κατανομή $b(n,\pi)$. Η παράμετρος της κατανομής είναι το π.

Η λογαριθμική συνάρτηση πιθανοφάνειας είναι :

$$l(y;\pi) = y\log\pi + (n-y)\log(1-\pi) + \log\binom{n}{y}$$

Επειδή $\frac{dl}{d\pi} = \frac{y}{\pi} - \frac{n-y}{1-\pi} = \frac{y-n\pi}{\pi\left(1-\pi\right)}$ έχουμε ότι $U = \frac{y-n\pi}{\pi\left(1-\pi\right)}$

Οπότε η στατιστική συνάρτηση *score* είναι : $U = \frac{Y}{\pi} - \frac{n-Y}{1-\pi} = \frac{Y-n\pi}{\pi(1-\pi)}$

Ισχύει ότι $E(Y) = n\pi$ οπότε $E(U) = 0$.

Ισχύει επίσης $Var(Y) = n\pi(1-\pi)$

Συνεπώς $J = \mathrm{var}(U) = \frac{1}{\pi^2(1-\pi)^2}\mathrm{var}(Y) = \frac{n}{\pi(1-\pi)}$

Η πληροφορία είναι $J = \frac{n}{\pi(1-\pi)}$

Οπότε προκύπτει ότι

$$U^T J^{-1} U = \frac{(Y-n\pi)^2}{\pi^2(1-\pi)^2}\frac{\pi(1-\pi)}{n} = \frac{(Y-n\pi)^2}{n\pi(1-\pi)}$$

Αυτό το αποτέλεσμα είναι ισοδύναμο με τη χρήση της κανονικής κατανομής για τη προσέγγιση της δυωνυμικής.

Δηλαδή ισχύει προσεγγιστικά : $\frac{Y-n\pi}{\sqrt{\left[n\pi(1-\pi)\right]}} \sim N(0,1)$.

Αυτό μπορεί να χρησιμοποιηθεί για να γίνει συμπερασματολογία για τη παράμετρο π.

9.2 Στατιστική συνάρτηση Wald

Θα ορίσουμε τη στατιστική συνάρτηση Wald η οποία χρησιμοποιείται για να κάνουμε συμπερασματολογία για τους *εκτιμητές μέγιστης πιθανοφάνειας*.

Υποθέτουμε ότι η λογαριθμική συνάρτηση πιθανοφάνειας έχει μοναδικό μέγιστο στο **b** για να είναι η εκτιμήτρια στο **b** κοντά στη πραγματική τιμή ***β***.

Η προσέγγιση Taylor πρώτης τάξης για το διάνυσμα *score* $U(\beta)$ στο σημείο β=b δίνεται από :

$$U(\beta) \cong U(b) + H(b)(\beta - b)$$

όπου *H(b)* είναι ο πίνακας δευτέρων παραγώγων της λογαριθμικής συνάρτησης πιθανοφάνειας στο β=b .

Επειδή ο πίνακας *H* είναι συμπτωτικά ίσος με την αναμενόμενη τιμή του έχουμε : $H = E(H)$

Ισχύει επίσης ότι : $J = E(UU^T) = E(-H)$

Συνεπώς για μεγάλα δείγματα : $U(\beta) \cong U(b) - J(\beta - b)$

Το b είναι το σημείο στο οποίο η λογαριθμική συνάρτηση πιθανοφάνειας έχει μέγιστο άρα η παράγωγος είναι μηδέν οπότε ισχύει : $U(b) = 0$

Συνεπώς (εάν η *J* δεν είναι μη μοναδιαία) θα έχουμε προσεγγιστικά ότι :

$$(b - \beta) \cong J^{-1} U$$

Εάν η *J* θεωρηθεί ότι είναι σταθερά τότε θα είναι $E(b - \beta) \cong J^{-1} E(U) = 0$ (διότι $E(U) = 0$) και συνεπώς το ***b*** θα είναι τουλάχιστον ασυμπτωτικά αμερόληπτη εκτιμήτρια του ***β*** .

Ο πίνακας διακύμανσης – συνδιακύμανσης για το b είναι :

$$E\left[(b - \beta)(b - \beta)^T\right] \cong J^{-1} E\left(UU^T\right) J^{-1} = J^{-1}$$

Διότι $J = E(UU^T)$ και $\left(J^{-1}\right)^T = J^{-1}$ εφόσον J είναι συμμετρικός.

Για μεγάλα δείγματα , λοιπόν, $\left(b-\beta\right)^T J\left(b-\beta\right) \sim x_p^2$

ή ισοδύναμα $b-\beta \sim N\left(0, J^{-1}\right)$ ***(Σ4)***

Η στατιστική συνάρτηση $\left(b-\beta\right)^T J\left(b-\beta\right)$ ονομάζεται <u>στατιστική συνάρτηση Wald</u>. Χρησιμοποιείται για να κάνουμε συμπερασματολογία για το ***β***.

Για γραμμικά μοντέλα που οι μεταβλητές απόκρισης ακολουθούν κανονική κατανομή ισχύουν οι προηγούμενες σχέσεις ***(Σ4)***.

<u>Παράδειγμα</u>

Υποθέτουμε ότι οι μεταβλητές απόκρισης Y_1, Y_2, ..., Y_N είναι ανεξάρτητα κατανεμημένες με $Y_i \sim \left(x_i^T\beta, \sigma^2\right)$ όπου το σ^2 είναι γνωστή σταθερά.

Έστω ότι X είναι N *x* p πίνακας που αποτελείται από τις σειρές $x^T{}_i$ και υποθέτουμε ότι ο πίνακας X^TX δεν είναι μοναδιαίος.

Στη περίπτωση αυτή έχουμε $E\left(Y_i\right) = \mu_i = x_i^T\beta$

Η συνάρτηση σύνδεσης είναι η ταυτοτική οπότε σύμφωνα με τους συμβολισμούς που έχουμε υιοθετήσει στα προηγούμενα κεφάλαια είναι μ_i = η_i και συνεπώς $\dfrac{\partial\mu_i}{\partial\eta_i} = 1$.

Από την παράγραφο 6.4 έχουμε ότι : $j_{jk} = \sum_{i=1}^{N} \dfrac{x_{ij}x_{ik}}{Var(Y_i)}\left(\dfrac{\partial\mu_i}{\partial\eta_\iota}\right)^2$

Προκύπτουν λοιπόν τα παρακάτω αποτελέσματα

Τα στοιχεία του πίνακα J δίνονται από τη σχέση $J_{jk} = \dfrac{1}{\sigma^2}\sum_{i=1}^{N} x_{ij}x_{ik}$ δηλαδή ο πίνακας πληροφορίας μπορεί να παρασταθεί ως εξής : $J = \dfrac{1}{\sigma^2}X^TX$

Ο W είναι διαγώνιος πίνακας με στοιχεία ίσα με $1/\sigma^2$.

$Z = Xb + y - Xb = y$

Ο εκτιμητής μέγιστης πιθανοφάνειας b είναι η λύση του $X^T Xb = X^T y$ συνεπώς $b = \left(X^T X\right)^{-1} X^T y$.

Εφ’ όσον το *b* είναι γραμμικός συνδυασμός κανονικά κατανεμημένων τυχαίων μεταβλητών $Y_1, Y_2, \ldots, Y_N$ είναι και το ίδιο κανονικά κατανεμημένο.

Το *b* είναι επίσης αμερόληπτη εκτιμήτρια του β διότι :

$$E(b) = \left(X^T X\right)^{-1} X^T E(Y) = \left(X^T X\right)^{-1} X^T X \beta = \beta \quad \text{διότι} \quad E(y) = x\beta$$

Για να πάρουμε τον πίνακα διακύμανσης – συνδιακύμανσης του *b* χρησιμοποιούμε τη σχέση $b - \beta = \left(X^T X\right)^{-1} X^T y - \beta = \left(X^T X\right)^{-1} X^T (y - X\beta)$ οπότε :

$$E\left[(b-\beta)(b-\beta)^T\right] = \left(X^T X\right)^{-1} X^T E\left[(y - X\beta)(y - X\beta)^T\right] X \left(X^T X\right)^{-1} = \sigma^2 \left(X^T X\right)^{-1}$$

Διότι $E\left[(y - X\beta)(y - X\beta)^T\right]$ είναι διαγώνιος πίνακας με στοιχεία σ^2.

Επειδή όμως $X^T X / \sigma^2 = J^2$ τότε ο πίνακας διακύμανσης – συνδιακύμανσης για το b είναι J^{-1}.

Συμπεραίνουμε ότι η κατανομή για το b είναι $N\left(\beta, J^{-1}\right)$ ή ισοδύναμα : $(b-\beta)^T J (b-\beta) \sim x_p^2$.

9.3 Διαστήματα Εμπιστοσύνης για Παραμέτρους Μοντέλων

Για να είναι δυνατός ο υπολογισμός των διαστημάτων εμπιστοσύνης των παραμέτρων, θα πρέπει είτε τα μοντέλα να είναι γραμμικά με αποκρίσεις που ακολουθούν κανονικές κατανομές, ή στην γενική περίπτωση που είναι γενικευμένα γραμμικά μοντέλα να έχουμε μεγάλα δείγματα οπότε τα αποτελέσματα ισχύουν προσεγγιστικά.

Στην προηγούμενη παράγραφο δείξαμε ότι $b \sim N\left(\beta, J^{-1}\right)$. Αυτό μπορεί να γραφτεί ισοδύναμα ως εξής : $b - \beta \sim N\left(0, J^{-1}\right)$.

Μπορούμε λοιπόν να κάνουμε τους ακόλουθους υπολογισμούς με βάση τη δειγματική κατανομή της εκτιμήτριας μεγίστης πιθανοφάνειας b :

Αξιολογούμε την αξιοπιστία των εκτιμητριών *bj* από το μέγεθος των τυπικών τους λαθών : $SE(b_j) = \sqrt{u_{jj}}$

όπου u_{jj} ο j–στος όρος της διαγωνίου.

Μπορούμε να υπολογίσουμε τα διαστήματα εμπιστοσύνης για μεμονωμένες παραμέτρους. Για παράδειγμα ένα 95% διάστημα εμπιστοσύνης για τα β_j δίνεται από το : $b_j \pm 1.96\sqrt{u_{jj}}$

Μπορούμε να υπολογίσουμε τις συσχετίσεις μεταξύ των εκτιμητριών χρησιμοποιώντας τη σχέση : $corr\left(b_j, b_k\right) = \dfrac{u_{jk}}{\sqrt{u_{jj}}\sqrt{u_{kk}}}$

Στην περίπτωση που ο πίνακας πληροφορίας J εξαρτάται από τις παραμέτρους β , χρειάζεται στην πράξη να υπολογίσουμε τον πίνακα στο σημείο $\beta=b$ και κατά περίπτωση να χρησιμοποιούμε αντί του $J(b)$ τον πίνακα $-H(b)$ σαν μία εκτιμήτρια του $J(\beta)$.

Θα πρέπει να επισημάνουμε τέλος ότι, όπως αναφέρθηκα προηγουμένως, τα αποτελέσματα είναι περισσότερο προσεγγιστικά παρά ακριβείς εκτιμήσεις των ποσοτήτων που αναφέρθηκαν.

<u>Παράδειγμα 4</u>

Για τα δεδομένα του επόμενου υπολογίσουμε τον πίνακα πληροφορίας και τα διαστήματα εμπιστοσύνης.

Το μοντέλο παράγει αποκρίσεις Y_i που ακολουθούν κατανομή Poisson με $E(Y_i) = \beta_1 + \beta_2 x_i$

Πίνακας 7 : Δεδομένα Poisson Y_i

y_i	2	3	6	7	8	9	10	12	15
x_i	-1	-1	0	0	0	0	1	1	1

<u>Με χρήση της R έχουμε</u>

```
> y1<-c(2, 3, 6, 7, 8, 9, 10, 12, 15)
> x1<-c(-1, -1, 0, 0, 0, 0, 1, 1, 1)
> fitted.model<-glm(y1~x1, family=poisson(link=identity), data=ndf)
> fitted.model

Call:  glm(formula = y1 ~ x1,  family = poisson(link = identity),  data = ndf)

Coefficients:
(Intercept)          x1
      7.452       4.935

Degrees of Freedom: 8 Total (i.e. Null);  7 Residual
Null Deviance:      18.42
Residual Deviance: 1.895        AIC: 40.01
```

Οι εκτιμήσεις μεγίστης πιθανοφάνειας είναι b1 = 7.4516 και b2 = 4.9353.

<u>Θα υπολογίσουμε τώρα τον πίνακα πληροφορίας με τη βοήθεια της R.</u>

```
> a11<-function(y0, x0) sum(1/(7.4516+4.9353*x0))
> a11(y1, x1)
[1] 1.573807
> a12<-function(y0, x0) sum(x0/(7.4516+4.9353*x0))
> a12(y1, x1)
[1] -0.5526264
> a21<-function(y0, x0) sum(x0/(7.4516+4.9353*x0))
```

```
> a22<-function(y0, x0) sum(x0^2/(7.4516+4.9353*x0))
> pinakas<-cbind(c(a11(y1, x1), a21(y1, x1)), c(a12(y1, x1), a22(y1, x1)) )
> pinakas
          [, 1]       [, 2]
[1, ]  1.5738066 -0.5526264
[2, ] -0.5526264  1.0370091

> pin_1<-solve(pinakas)
> pin_1
              [, 1]             [, 2]
[1, ]    0.7816716      0.416556
[2, ]    0.4165560     1.186296
```

Ο αντίστροφος πίνακας πληροφορίας έχει υπολογιστεί στο b και είναι :

$$J^{-1} = \begin{pmatrix} 0.7817 & 0.4166 \\ 0.4166 & 1.1863 \end{pmatrix}$$

Ο πίνακας J^{-1} δείχνει ότι το b1 είναι κατά κάποιο τρόπο πιο αξιόπιστο από το b2 διότι το τυπικό σφάλμα $SE(b_1) = \sqrt{0.7817} = 0.88$ είναι μικρότερο από το $SE(b_2) = \sqrt{1.1863} = 1.09$.

Ο συντελεστής συσχέτισης για τα b_1 και b_2 είναι κατά προσέγγιση : $r = \frac{0.4166}{\sqrt{0.7817}\sqrt{1.1863}} \cong 0.43$

Ένα 95% διάστημα εμπιστοσύνης για το β_1 είναι :

$$7.4516 \pm 1.96\sqrt{0.7817} = (5.72, 9.18)$$

9.4 Επάρκεια του Μοντέλου

Η επάρκεια ενός μοντέλου είναι η ικανότητα του μοντέλου να περιγράψει τα δεδομένα που μελετάμε.

Ένα μοντέλο που περιλαμβάνει τόσες παραμέτρους όσα είναι τα δεδομένα (δηλαδή ίδιο πλήθος παραμέτρων και δεδομένων), περιγράφει τα δεδομένα επακριβώς άρα είναι ένα «επαρκές μοντέλο».

Ας υποθέσουμε ότι θέλουμε να εκτιμήσουμε την επάρκεια ενός μοντέλου που περιγράφει ένα σύνολο δεδομένων που μελετάμε. Η εκτίμηση γίνεται συγκρίνοντας την πιθανοφάνεια του μοντέλου που μελετάμε με την πιθανοφάνεια για το **πλήρες (maximal)** μοντέλο.

Το πλήρες (maximal) μοντέλο ορίζεται ως εξής :

- Το πλήρες μοντέλο είναι ένα γενικευμένο γραμμικό μοντέλο που χρησιμοποιεί την ίδια κατανομή με το μοντέλο που μας ενδιαφέρει. (πχ. και τα δύο κανονική ή και τα δύο διωνυμική κατανομή).
- Το πλήρες μοντέλο έχει την ίδια συνάρτηση σύνδεσης με το μοντέλο που μας ενδιαφέρει.
- Ο αριθμός των παραμέτρων στο πλήρες μοντέλο είναι ίσος με τον αριθμό των παρατηρήσεων N.

Συνεπώς το πλήρες μοντέλο μπορεί να θεωρηθεί ότι μας εφοδιάζει με μία πλήρη περιγραφή των δεδομένων και όπως αναφέρθηκε είναι ένα επαρκές μοντέλο. Με το πλήρες μοντέλο έχουμε τη μεγαλύτερη πολυπλοκότητα. Χρησιμοποιώντας ένα μοντέλο με ένα μικρότερο αριθμό παραμέτρων μπορούμε να επιτύχουμε την επιθυμητή προσαρμογή και συγχρόνως δεν συμπεριλαμβάνουμε παραμέτρους που δεν χρειαζόμαστε. Το πλήρες μοντέλο δεν έχει καλή προβλεπτική ικανότητα.

Οι συναρτήσεις πιθανοφάνειας του πλήρους μοντέλου με το μοντέλο που μελετάμε μπορεί να πάρουν τις αντίστοιχες τιμές εκτίμησης μέγιστης πιθανοφάνειας b_{max} και b και να προκύψουν αντίστοιχα οι τιμές $L(b_{max};y)$ και $L(b;y)$.

Εάν το μοντέλο που μελετάμε είναι φτωχό τότε η τιμή $L(b;y)$ θα είναι πολύ μικρότερη από την $L(b_{max};y)$, αλλιώς θα είναι προσεγγιστικά ίσες. Για το λόγο αυτό, προτείνεται η χρήση της στατιστικής συνάρτησης «γενικευμένου λόγου πιθανοφάνειας» $\lambda = \frac{L(b_{max};y)}{L(b;y)}$ σαν ένα μέτρο για τη καλή προσαρμογή.

Ισοδύναμα μπορεί να χρησιμοποιηθεί η διαφορά των λογαριθμικών συναρτήσεων πιθανοφάνειας :

$$\log\lambda = l(b_{max};y) - l(b;y)$$

Μεγάλες τιμές του λ υποδεικνύουν ότι το μοντέλο που μας ενδιαφέρει περιγράφει φτωχά τα δεδομένα. Για να καθορίσουμε κρίσιμη περιοχή για το logλ πρέπει να γνωρίζουμε τη δειγματική κατανομή.

9.4.1 Δειγματική Κατανομή για τη Στατιστική συνάρτηση «Λογαριθμική Πιθανοφάνεια»

Θέλουμε να εκτιμήσουμε ένα μοντέλο που έχει p παραμέτρους. Συμβολίζουμε με β το διάνυσμα των παραμέτρων.

Έστω ότι $l(\beta;y)$ είναι η λογαριθμική πιθανοφάνεια του μοντέλου. Για το $l(\beta;y)$ μία προσέγγιση με σειρά Taylor προκύπτει αναπτύσσοντας τη σειρά γύρω από την εκτιμήτρια μέγιστης πιθανοφάνειας b.

$$l(\beta;y) \cong l(b;y) + (\beta - b)^T U(b) + \frac{1}{2}(\beta - b)^T H(b)(\beta - b) \quad \textbf{(Σ5)}$$

όπου *U(b)* είναι το διάνυσμα των *scores* $\partial l / \partial \beta_j$ που έχει υπολογιστεί στο *b*, και *H(b)* είναι ο πίνακας των δευτέρων παραγώγων που έχει υπολογιστεί στο *b*.

Από τον ορισμό του *b* ισχύει ότι *U(b) = 0*.
Για μεγάλα δείγματα έχουμε ότι το *-H(b)* μπορεί να προσεγγιστεί από τον πίνακα πληροφορίας *J = E[-H]*.

Συνεπώς η προηγούμενη σχέση **(Σ5)** μπορεί να γραφτεί ως εξής :

$l(b;y) - l(\beta;y) = \frac{1}{2}(b-\beta)^T J (b-\beta)$ αλλά

επειδή $(b-\beta)^T J (b-\beta) \sim x_p^2$ προκύπτει $2[l(b;y) - l(\beta;y)] \sim x_p^2$

Χρησιμοποιούμε τη στατιστική συνάρτηση που βασίζεται σε αυτό το αποτέλεσμα για ένα test που :

- υπολογίζει την προσαρμογή ενός μοντέλου.
- και συγκρίνει εναλλακτικά μοντέλα.

9.5 Στατιστική συνάρτηση «Απόκλιση» (Deviance)

Ορίζουμε τη στατιστική συνάρτηση του «λογάριθμου λόγου πιθανοφανειών» *D* ως εξής :

$$D = 2\log\lambda = 2\left[l(b_{\max};y) - l(b;y)\right]$$

Οι Nelder και Wedderburn το 1972 ονόμασαν τη συνάρτηση *D* ως Deviance. Εμείς θα αποδίδουμε τον όρο Deviance με τον όρο Απόκλιση.

Θα μετασχηματίσουμε το D ως εξής :

$$D = 2\left\{\left[l(b_{\max};y) - l(\beta_{\max};y)\right] - \left[l(b;y) - l(\beta;y)\right] + \left[l(\beta_{\max};y) - l(\beta;y)\right]\right\} \quad \textbf{(Σ6)}$$

Το αποτέλεσμα αποτελείται από τρεις διαφορές που βρίσκονται σε αγκύλες. Η πρώτη αγκύλη ακολουθεί την x_N^2 κατανομή διότι το πλήρες μοντέλο έχει N παραμέτρους.

Η δεύτερη ακολουθεί τη x_p^2 κατανομή διότι έχει p παραμέτρους.

Η τρίτη αγκύλη είναι μία θετική σταθερά η οποία είναι κοντά στο μηδέν εάν το μοντέλο με p παραμέτρους περιγράφει τα δεδομένα σχεδόν τόσο καλά όσο τα περιγράφει το πλήρες μοντέλο.

Μπορούμε να πούμε λοιπόν ότι, εάν οι τυχαίες μεταβλητές που ορίζονται από τις δύο πρώτες διαφορές είναι ανεξάρτητες και ο τρίτος (σταθερός) όρος είναι κοντά στο μηδέν τότε :

Εάν το μοντέλο είναι καλό θα ισχύει : $D \sim x_{N-p}^2$.

Εάν το μοντέλο είναι φτωχό ο τρίτος όρος θα είναι μεγάλος και το D θα είναι μεγαλύτερο από αυτό που θέλει το x_{N-p}^2 (στην πραγματικότητα το D θα έχει μία μη-κεντρική X^2 κατανομή).

Σε γενικές γραμμές το $D \sim x_{N-p}^2$ δεν μας εφοδιάζει με μία πολύ καλή προσέγγιση για τη δειγματική κατανομή. Τα αποτελέσματα όμως έχουν καλή προσέγγιση για κανονικά μοντέλα.

9.5.1 Υπολογισμός Απόκλισης για κανονικά μοντέλα

Υποθέτουμε ότι οι τυχαίες μεταβλητές $Y_1, Y_1, \ldots, Y_N$ είναι ανεξάρτητες και ακολουθούν την κανονική κατανομή $N(\mu_i, \sigma^2)$, όπου τα μ_i μπορεί να διαφέρουν αλλά έχουν κοινό σ^2.

Η λογαριθμική συνάρτηση πιθανοφάνειας είναι :

$$l(\beta; y) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mu_\iota)^2 - \frac{1}{2}N\log\left(2\pi\sigma^2\right) \quad (\Sigma 7)$$

Για το πλήρες μοντέλο είναι : $E(Y_i)=\mu_i$, για κάθε $i=1,2, \ldots,N$ οπότε το $\boldsymbol{\beta}$ έχει στοιχεία $\mu_1,\mu_2,\ldots,\mu_N$.

Παραγωγίζοντας τη λογαριθμική συνάρτηση πιθανοφάνειας λαμ βάνουμε $\hat{\mu}_i=\bar{y}$.

Οπότε λόγω της **(Σ7)** θα ισχύει : $l(b_{\max};y)=-\frac{1}{2}N\log(2\pi\sigma^2)$

Θεωρούμε τώρα το μοντέλο στο οποίο όλα τα Yi έχουν τον ίδιο μέσο **μ**. Τότε $\hat{\mu}=\bar{y}$ και συνεπώς $l(b;y)=-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\bar{y})^2-\frac{1}{2}N\log(2\pi\sigma^2)$

Θα έχουμε : $D=2\left[l(b_{\max};y)-l(b;y)\right]=\frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i-\bar{y})^2$

Η στατιστική συνάρτηση D σχετίζεται με τη δειγματική διασπορά.

$S^2=\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{y})^2$ με τη σχέση $D=\frac{(N-1)S^2}{\sigma^2}$

Εάν το μοντέλο με τον κοινό μέσο **είναι σωστό**, τότε θα ισχύει :

$$\frac{(N-1)S^2}{\sigma^2}\sim x^2_{N-1} \quad .$$

<u>Παρατήρηση</u>

Από τα προηγούμενα έχουμε ότι το πλήρες μοντέλο για κανονικές ανεξάρτητες μεταβλητές με κοινή διασπορά είναι :

$$l(b_{\max};y)=-\frac{1}{2}N\log(2\pi\sigma^2)$$

Εάν υποθέσουμε ότι στο μοντέλο που μελετάμε οι μέσοι εξαρτώνται από p παραμέτρους $\boldsymbol{\beta}$ όπου $p<N$.

Έστω $\hat{\mu}_i$ δηλώνει την εκτίμηση του μ_i που υπολογίστηκε από την εκτίμηση μέγιστης πιθανοφάνειας του $\boldsymbol{b}$. Τότε :

$$l(b;y)=-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\hat{\mu}_i)^2-\frac{1}{2}N\log(2\pi\sigma^2)$$

Συνεπώς είναι : $D = \frac{1}{\sigma^2}\sum_{i=1}^{N}\left(y_i - \hat{\mu}_i\right)^2$

Εάν το μοντέλο είναι σωστό τότε $D \sim x^2_{N-p}$. Οι εκτιμήσεις $\hat{\mu}_i$ ονομάζονται προσαρμοσμένες τιμές για τα Y_i. Τα $(y_i - \hat{\mu}_i)$ ονομάζονται κατάλοιπα. Το D είναι το άθροισμα τετραγώνων των καταλοίπων διαιρεμένα με την παράμετρο σ^2.

9.5.2 Υπολογισμός Απόκλισης για μοντέλα Poisson

Υποθέτουμε ότι οι τυχαίες μεταβλητές $Y_1, Y_1, \ldots, Y_N$ είναι ανεξάρτητες και ακολουθούν την κατανομή Poisson με παραμέτρους λ_i.
Η λογαριθμική συνάρτηση πιθανοφάνειας είναι:

$$l\left(\beta ; y\right) = \sum y_i \log\lambda_i - \sum\lambda_i - \sum\log y_i!$$

Για το πλήρες μοντέλο οι εκτιμήτριες μέγιστης πιθανοφάνειας είναι $\hat{\lambda}_i = y_i$ συνεπώς : $l(b_{\max}; y) = \sum y_i \log y_i - \sum y_i - \sum \log y_i!$

Στο μοντέλο όπου όλα τα Y_i έχουν την ίδια παράμετρο λ οι εκτιμήτριες μέγιστης πιθανοφάνειας είναι $\hat{\lambda} = \sum \frac{y_i}{N} = \bar{y}$ οπότε : $l(b; y) = \sum y_i \log \bar{y} - N\bar{y} - \sum \log y_i!$

Έχουμε λοιπόν $D = 2\left[l(b_{\max}; y) - l(b; y)\right] = 2\left[\sum y_i \log y_i - \sum y_i \log \bar{y}\right] = 2\sum y_i \log\left(\frac{y_i}{\bar{y}}\right)$

Εάν το μοντέλο της μίας παραμέτρου είναι σωστό τότε προσεγγιστικά : $D \sim x^2_{N-1}$.

9.5.3 Αξιολόγιση «Καλής Προσαρμογής» ενός μοντέλου

Η δειγματική κατανομή της στατιστικής συνάρτησης Απόκλισης (Deviance) μπορεί να χρησιμοποιηθεί για να ερευνηθεί η επάρκεια ενός μοντέλου εκτιμώντας το D από τα δεδομένα και συγκρίνοντας τη τιμή με την κατάλληλη X^2 κατανομή.

Είναι γνωστό από τη θεωρία ότι η αναμενόμενη τιμή μιας τυχαίας μεταβλητής με x_m^2 κατανομή είναι m.

Εάν το μοντέλο είναι καλό, τότε θα πρέπει να περιμένουμε η τιμή του D να είναι κοντά στο μέσο της κατανομής. Εάν, λοιπόν, ένα μοντέλο με p παραμέτρους περιγράφει καλά ένα σύνολο δεδομένων από N παρατηρήσεις έτσι ώστε $D \sim x_{N-p}^2$ τότε θα πρέπει να περιμένουμε $D \cong N-p$. **(Σ7)**

Για κάποιες κατανομές, όπως η Poisson, η τιμή του D μπορεί να υπολογιστεί απ' ευθείας από τις προσαρμοσμένες τιμές και να συγκριθεί με τους βαθμούς ελευθερίας για να εκτιμηθεί η καλή προσαρμογή.

Για κάποιες άλλες κατανομές, όπως η κανονική κατανομή, η τιμή του D δεν μπορεί να υπολογιστεί απ' ευθείας λόγω της παρουσίας του σ^2. Γενικά, η **(Σ7)** δίνει ένα χοντροειδή τρόπο υπολογισμού της «καλής προσαρμογής» ενός μοντέλου.

Για να έχουμε καλύτερα αποτελέσματα στην εκτίμηση της καλής προσαρμογής χεησιμοποιούμε μεθόδους που στηρίζονται στα κατάλοιπα.

<u>Παράδειγμα 8</u>

Θεωρούμε το μοντέλο **Poisson** που προσαρμόζεται στα δεδομένα του προηγούμενου **Πίνακα 7** (επαναλαμβάνουμε τον πίνακα).

y_i	2	3	6	7	8	9	10	12	15
x_i	-1	-1	0	0	0	0	1	1	1

Όταν το μοντέλο $E(Y_i) = \beta_1 + \beta_2 x_i$ με $p = 2$ παραμέτρους προσαρμόζεται στις $N = 9$ παρατηρήσεις τότε έχουμε τις εκτιμήσεις β που υπολογίστηκαν στο **παράδειγμα 4** της παράγραφου 10.3. Το πρασαρμοσμένο μοντέλο το ονομάζουμε fitted.model.

<u>Με την εντολή summary της R έχουμε</u>

```
> summary(fitted.model)

Call:
glm(formula = y1 ~ x1,  family = poisson(link = identity),  data = ndf)

Deviance Residuals:
   Min         1Q      Median        3Q            Max
-0.7019        -0.3377         -0.1105        0.2958        0.7184

Coefficients:
               Estimate      Std. Error     z value       Pr(>|z|)
(Intercept)    7.4516        0.8841         8.428         < 2e-16 ***
x1             4.9353        1.0892         4.531  5.86e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 18.4206  on 8  degrees of freedom
```
Residual deviance: 1.8947 on 7 degrees of freedom
```
AIC: 40.008

Number of Fisher Scoring iterations: 3
```

Από την R υπολογίστηκε D = 1.8949 με 7 βαθμούς ελευθερίας. Αυτή η τιμή της Απόκλισης (deviance) είναι μικρή σε σχέση με τους βαθμούς ελευθερίας (στη πραγματικότητα είναι κάτω από το 5% της ουράς της κατανομής) πράγμα που δείχνει ότι το μοντέλο δεν προσαρμόζεται καλά στα δεδομένα.

9.6 Έλεγχος Υποθέσεων

<u>Έλεγχος υποθέσεων με χρήση δειγματικής κατανομής εκτιμητριών</u>

Για να κάνουμε ελέγχους υποθέσεων για την παράμετρο β μπορούμε να χρησιμοποιήσουμε την ασυμπτωτική δειγματική κατανομή των εκτιμητριών b.

Ασυμπτωτικά ισχύει : $b \sim N\left(\beta ; J^{-1}\right)$.

Ισοδύναμα μπορεί να χρησιμοποιηθεί <u>η στατιστική συνάρτηση Wald</u> που είναι $\left(b-\beta\right)^{T} J\left(b-\beta\right)$ και η οποία ακολουθεί την x_{p}^{2} κατανομή.

Σε κάποιες περιπτώσεις μπορεί να χρησιμοποιηθεί η στατιστική συνάρτηση score $U^{T} J^{-1} U$ το οποίο ακολουθεί επίσης την x_{p}^{2} κατανομή.

<u>Έλεγχος υποθέσεων με χρήση Απόκλισης (Devia nce)</u>

Καθορίζουμε ένα μοντέλο για κάθε υπόθεση και συγκρίνουμε τις στατιστικές συναρτήσεις της καλής προσαρμογής για τα συγκρινόμενα μοντέλα.

Τα μοντέλα που πρόκειται να συγκριθούν πρέπει να έχουν την ίδια κατανομή και την ίδια συνάρτηση σύνδεσης. Το μόνο που διαφέρει είναι ο αριθμός των παραμέτρων.

Θεωρούμε την μηδενική υπόθεση H_0 και την εναλλακτική υπόθεση H_1 όπου

$$H_0 : \beta = \beta_0 = \begin{pmatrix} \beta_1 \\ . \\ . \\ . \\ \beta_q \end{pmatrix} \quad \text{έναντι} \quad H_1 : \beta = \beta_1 = \begin{pmatrix} \beta_1 \\ . \\ . \\ . \\ \beta_p \end{pmatrix} \quad \text{όπου} \quad q < p < N$$

Κάνουμε τον έλεγχο H_0 έναντι H_1 χρησιμοποιώντας την διαφορά των αποκλίσεων:

$$\Delta D = D_0 - D_1 = 2\left[l\left(b_{\max}; y\right) - l\left(b_0; y\right)\right] - 2\left[l\left(b_{\max}; y\right) - l\left(b_1; y\right)\right] = 2\left[l\left(b_1; y\right) - l\left(b_0; y\right)\right]$$

Εάν τα δύο μοντέλα περιγράφουν καλά τα δεδομένα (με την προϋπόθεση να ισχύουν οι κανόνες ανεξαρτησίας) τότε $D_0 \sim x^2_{N-q}$ και $D_1 \sim x^2_{N-p}$ οπότε θα ισχύει $\Delta D \sim x^2_{p-q}$.

Εάν η τιμή της ΔD είναι συνεπής με την κατανομή x^2_{p-q} τότε θα επιλέγουμε το μοντέλο H_0 γιατί είναι απλούστερο.

Εάν η τιμή ΔD είναι στη κρίσιμη περιοχή (μεγαλύτερη δηλαδή από την πάνω ουρά στο σημείο 100 α% της x^2_{p-q} κατανομής) τότε απορρίπτουμε την H_0 και δεχόμαστε την H_1 με την αιτιολογία ότι το β_1 μας εφοδιάζει με μία σημαντικά καλύτερη περιγραφή των δεδομένων (έστω και αν δεν προσαρμόζεται καλά στα δεδομένα).

Η δειγματική κατανομή του ΔD προσεγγίζεται πολύ καλύτερα συνήθως από την X^2 κατανομή σε σχέση με τη προσέγγιση που γίνεται από τη δειγματική κατανομή μιας μεμονωμένης στατιστικής συνάρτησης απόκλισης D.

<u>Παρατήρηση:</u>

Σε μοντέλα που ακολουθούν τη κανονική κατανομή με κοινή διακύμανση σ^2 οι στατιστικές συναρτήσεις του πηλίκου λογαριθμικής πιθανοφάνειας εξαρτώνται από το σ^2 οπότε δεν μπορεί να υπολογιστούν απ΄ ευθείας από τις προσαρμοσμένες τιμές. Αυτή η δυσκολία ξεπερνιέται με τον ακόλουθο τρόπο.

Μέθοδος

Βρίσκουμε τις προσαρμοσμένες τιμές για τις αποκρίσεις Y_i.

Έστω $\hat{\mu}_i(0)$ ότι είναι οι προσαρμοσμένες τιμές για τις αποκρίσεις Y_i κάτω από την υπόθεση H_0.

Έστω $\hat{\mu}_i(1)$ ότι είναι οι προσαρμοσμένες τιμές για τις αποκρίσεις Y_i κάτω από την υπόθεση H_1.

Ισχύουν : $D_0 = \frac{1}{\sigma^2}\sum\left[y_i - \hat{\mu}_i(0)\right]^2$ και $D_1 = \frac{1}{\sigma^2}\sum\left[y_i - \hat{\mu}_i(1)\right]^2$.

Αν υποθέσουμε ότι η H_1 είναι σωστή. Τότε $D_1 \sim x^2_{N-p}$.

Εάν είναι και η H_0 επίσης σωστή τότε $D_0 \sim x^2_{N-q}$.

Οπότε $\Delta D = D_0 - D_1 \sim x^2_{p-q}$.

Εάν η H_0 δεν είναι σωστή τότε η ΔD θα ακολουθεί μη–κεντρική x^2 κατανομή.

Για να απαλείψουμε το σ^2 χρησιμοποιούμε το πηλίκο :

$$F = \frac{\dfrac{D_0 - D_1}{p-q}}{\dfrac{D_1}{N-p}} = \frac{\left\{\sum\left[y_i - \hat{\mu}(0)\right]^2 - \sum\left[y_i - \hat{\mu}_i(1)\right]^2\right\} \Big/ (p-q)}{\sum\left[y_i - \hat{\mu}_i(1)\right]^2 \Big/ N-p}$$

Το F μπορεί να υπολογιστεί απ’ ευθείας από τις προσαρμοσμένη τιμές.

Εάν ισχύει η H_0 τότε η F θα έχει, τουλάχιστον προσεγγιστικά, την κεντρική $F_{p-q,N-p}$ κατανομή.

Εάν η H_0 δεν είναι σωστή τότε η τιμή της F θα είναι μεγαλύτερη από τη τιμή που αναμένεται από την $F_{p-q,N-p}$.

Παράδειγμα 9

Έχουμε δει ότι χρησιμοποιείται μία μέθοδος προσαρμογής μοντέλων για τον έλεγχο της μηδενικής υπόθεσης : «Δεν υπάρχει διαφορά στο βάρος μεταξύ δύο φυτών που μεγαλώνουν κάτω από διαφορετικές συνθήκες».Τα δεδομένα είναι στον ακόλουθο πίνακα.

Πίνακας 8 : Βάρη φυτών κάτω από δύο συνθήκες

Ελέγχου	4.17	5.58	5.18	6.11	4.50	4.61	5.17	4.53	5.33	5.14
Θεραπείας	4.81	4.17	4.41	3.59	5.87	3.83	6.03	4.89	4.32	4.69

Έστω η μεταβλητή Y_{jk} η οποία δηλώνει το βάρος του k φυτού που μεγαλώνει στην j συνθήκη. Το j =1 για έλεγχο και j = 2 για θεραπεία και k = 1, 2, …, 10.

Τα βάρη θεωρούμε ότι ακολουθούν την κανονική κατανομή. Για τον έλεγχο υποθέσεων θα συγκρίνουμε τα μοντέλα.

Μοντέλο 1 : $Y_{jk} = \mu_j + e_{jk}$ όπου μj είναι ο μέσος για τη συνθήκη j.

Μοντέλο 0 : $Y_{jk} = \mu + e_{jk}$ όπου μ είναι ίδιος σε όλες τις περιπτώσεις.

Θα γίνει μελέτη με R

Θα υπολογιστεί το F απ' ευθείας από τις προσαρμοσμένη τιμές.

```
> y1<-c(4.17, 5.58, 5.18, 6.11, 4.50, 4.61, 5.17, 4.53, 5.33, 5.14)
> y2<-c(4.81, 4.17, 4.41, 3.59, 5.87, 3.83, 6.03, 4.89, 4.32, 4.69)
> m1<-mean(y1)
> m2<-mean(y2)
> y0<-c(y1, y2)
> m0<-mean(y0)
> s1<-sum((y1-m1)^2)
> s2<-sum((y2-m2)^2)
> s0<-sum((y0-m0)^2)
```

```
> F_pq<-((s0-s1-s2)/1)/((s1+s2)/18)
> F_pq
[1] 1.419101
```

Υπολογίζουμε το ποσοστημόριο για την τιμή $F_{1,\ 18}$ και επίπεδο σημαντικότητας α=5%

```
> qf(0.95,1,18)
[1] 4.413873
```

Η τιμή της F=1.419 είναι πολύ μικρότερη από την τιμή της συνάρτησης $F_{1,\ 18}$ οπότε και δεν απορρίπτεται η $\mathbf{H_0}$.

9.7 Κατάλοιπα

Η επάρκεια ενός μοντέλου εξετάζεται με τη βοήθεια μιας στατιστικής συνάρτησης «Καλής προσαρμογής». Οι συναρτήσεις όμως αυτές, δίνουν ένα **γενικό μέτρο** για την καλή προσαρμογή και δεν δίνουν πληροφορία για τη μορφή του μοντέλου. Η εξερεύνηση των ιδιαιτεροτήτων ενός μοντέλου γίνεται με τη χρήση των καταλοίπων.

Ας υποθέσουμε ότι έχουμε ένα μοντέλο που οι αποκρίσεις Y_i ακολουθούν κανονική κατανομή. Η μοντελοποίηση του μοντέλου γίνεται με την εξίσωση $Y_i = \mu_i + e_i$ όπου οι όροι σφαλμάτων e_i θεωρούνται ανεξάρτητοι και $e_i \sim N\left(0,\sigma^2\right)$. Οι αναμενόμενες τιμές μ_i είναι μία μονότονη συνάρτηση γραμμικών συνδυασμών των στοιχείων του διανύσματος παραμέτρων β.

Για αυτό το μοντέλο έχουμε : $\frac{\left(Y_i - \mu_i\right)}{\sigma} \sim N\left(0,1\right)$.

Τα κατάλοιπα που αντιστοιχούν στα Y_i ορίζονται ως : $(y_i - \hat{\mu}_i)$ όπου $\hat{\mu}_i$ είναι οι προσαρμοσμένες τιμές που υπολογίζονται με την εκτίμηση της μέγιστης πιθανοφάνειας b.

Εάν το $\hat{\sigma}$ είναι μία εκτίμηση του σ τότε ορίζονται τα τυποποιημένα κατάλοιπα r_i από τη σχέση : $r_i = \frac{(y_i - \hat{\mu}_i)}{\hat{\sigma}}$.

Θα πρέπει να εξεταστεί η κανονικότητα των τυποποιημένων καταλοίπων (χρησιμοποιώντας για παράδειγμα τα κατάλληλα γραφήματα) και η ύπαρξη συσχέτισης μεταξύ των r_i (πχ. με χρήση του τεστ Durbin Watson). Εάν τα κατάλοιπα είναι περίπου ασυσχέτιστα θα ακολουθούν κατά προσέγγιση την κατανομή $N(0,1)$.

Στα γενικευμένα γραμμικά μοντέλα ορίζονται τυποποιημένα κατάλοιπα με πολλούς διαφορετικούς τρόπους . Μία πρόταση είναι να γενικεύσουμε τον ορισμό των τυποποιημένων καταλοίπων για κανονικά μοντέλα και να προκύψει ο ορισμός των τυποποιημένων μοντέλων για γενικευμένα γραμμικά μοντέλα. Αυτό χρησιμοποιούμε στην τρέχουσα εργασία.

Έστω ότι $s_i = \sqrt{Var(\mu_i)}$ η εκτιμηθείσα τιμή για την τυπική απόκλιση των προσαρμοσμένων τιμών $\hat{\mu}_i$.

Τα **τυποποιημένα κατάλοιπα για γενικευμένα γραμμικά μοντέλα** ορίζονται ως : $r_i = \frac{(y_i - \hat{\mu}_i)}{s_i}$

<u>Παρατήρηση</u>

Άλλοι ορισμοί από αυτούς συζητούνται από τους McCullagh and Nelder (1989) καθώς και η χρήση ποικίλων μετασχηματισμών για να βελτιώσουν την κανονικότητα των καταλοίπων.

Υπάρχουν επίσης ορισμοί που στηρίζονται στις προσημασμέν ες τετραγωνικές ρίζες των όρων της στατιστικής συνάρτησης λογάριθμου λόγου πιθανοφανειών που προτάθηκαν από τους Pregibon (1981) και Pierce and Schafer (1986).

Κατάλοιπα Poisson

Για τη κατανομή Poisson έχουμε : $E(Y_i) = \text{var}(Y_i) = \lambda_i$ οπότε $r_i = \frac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}$

Τα κατάλοιπα Poisson μπορεί να θεωρηθούν σαν προσημασμένες τετραγωνικές ρίζες των όρων της στατιστικής συνάρτησης καλής προσαρμογής του Pearson : $\sum \frac{(o_i - e_i)^2}{e_i}$

όπου o_i είναι η παρατηρηθείσες τιμές και e_i είναι η προσαρμοσμένες τιμές $\hat{\lambda}_i$ που αναμένονται από το μοντέλο.

Χρήση τυποποιημένων καταλοίπων[5]

Τα τυποποιημένα κατάλοιπα μπορεί να χρησιμοποιηθούν για να καθοριστεί η επάρκεια ενός μοντέλου με τους ακόλουθους τρόπους.

Η σύγκρισή τους με την κανονική κατανομή προσδιορίζει την επάρκεια των δειγματικών προϋποθέσεων στο μοντέλο. Αυτό μπορεί να γίνει ελέγχοντας τη κατανομή συχνοτήτων και ψάχνοντας για τιμές πέρα από το εύρος πιθανοτήτων. Για παράδειγμα, δεν μπορεί πάνω από το 5% να είναι μικρότερο από -1.96 ή μεγαλύτερο από +1.96.

Γίνεται εκτίμηση της κανονικότητας με τη χρήση του Probability plot (είναι το plot των κανονικοποιημένων καταλοίπων ως προς τα normal scores). Τα σημεία θα πρέπει να βρίσκονται πάνω σε ευθεία γραμμή και συστηματικές αποκλίσεις η παρατηρήσεις που είναι έξω από τη γραμμή (και μακριά) δείχνουν απομάκρυνση από το μοντέλο. Οι τιμές αυτές θα πρέπει να ερευνηθούν στη συνέχεια.

Δημιουργούμε τα plot των τυποποιημένων καταλοίπων ως προς καθεμία από τις μεταβλητές για να δούμε εάν το μοντέλο περιγράφει επαρκώς την επίδραση της μεταβλητής. Εάν το μοντέλο είναι επαρκές δεν θα υπάρχει φανερό κάποιο σχήμα στο plot. Εάν είναι ανεπαρκές τότε τα σημεία θα σχηματίζουνε συστηματικά σχήματα , και αυτό θα σημαίνει ότι επιπρόσθετοι όροι πρέπει να συμπεριληφθούν στο μοντέλο για να περιγράψουν την επίδραση της μεταβλητής.

Μπορούμε να κάνουμε plot των καταλοίπων με τις προσαρμοσμένες τιμές. Εάν υπάρχει συστηματικό σχήμα στο plot αυτό σημαίνει ότι επιπρόσθετες μεταβλητές πρέπει να εισέλθουν στο μοντέλο.

Περισσότερες πληροφορίες για τα κατάλοιπα δίνεται από τους Drapper &Smith (1981), Belsey &Kuh &Welsch (1980), Cook & Weisberg (1982).

10 Πολλαπλή Παλινδρόμηση

10.1 Εισαγωγή

Στα προηγούμενα κεφάλαια αναφερθήκαμε σε μία ομάδα γενικευμένων γραμμικών μοντέλων που η απόκριση ακολουθεί κανονική κατανομή. Μία ειδική περίπτωση των μοντέλων αυτών είναι η πολλαπλή γραμμική παλινδρόμηση.

Η πολλαπλή γραμμική παλινδρόμηση προκύπτει από την εξίσωση :

$y = X\beta + e$ και περιγράφεται από την εξίσωση $E(y) = X\beta$

Όπου y είναι ένα $N x 1$ διάνυσμα αποκρίσεων.

X είναι ένας πίνακας $N x p$ από σταθερές.

β είναι ένα διάνυσμα παραμέτρων $p x 1$.

e είναι ένα $N x 1$ τυχαίο διάνυσμα του οποίου τα στοιχεία είναι ανεξάρτητα και ακολουθούν όλα την τυπική κανονική κατανομή $N\left(0,\sigma^2\right)$.

Η προηγούμενη εξίσωση περιγράφει ένα γενικευμένο γραμμικό μοντέλο , με τα εξής χαρακτηριστικά :

$E\left(Y_i\right) = \mu_i = x_i^T\beta$

Τα στοιχεία Y_i του y είναι ανεξάρτητα και κανονικά κατανεμημένα.

Η συνάρτηση σύνδεσης g(μ) είναι η ταυτοτική συνάρτηση διότι τα μ_i είναι ήδη γραμμικοί συνδυασμοί των παραμέτρων.

Τα x_i^T είναι η i–οστή γραμμή του πίνακα X.

Για τα μοντέλα πολλαπλής παλινδρόμησης θα πρέπει ο πίνακας των σταθερών X να έχει γραμμικά ανεξάρτητες στήλες ώστε ο πίνακας $X^T X$ να μην είναι μοναδιαίος.

Στα ακόλουθα θα δούμε κάποιες περιπτώσεις παλινδρόμησης. Θα ανακεφαλαιώσουμε και θα κριτικάρουμε τα θεωρητικά αποτελέσματα σε σχέση με σχετικά μοντέλα παλινδρόμησης[6].

10.1.1 Απλή γραμμική παλινδρόμηση

Η απλή γραμμική παλινδρόμηση μοντελοποιεί τη γραμμική σχέση μεταξύ μιας συνεχούς μεταβλητής απόκρισης η οποία είναι κανονικά κατανεμημένη, και μιας απλής επεξηγηματικής μεταβλητής.

Η εξίσωση της απλής γραμμικής παλινδρόμησης είναι :

$$E(Y_i) = \beta_0 + \beta_1 x_i \qquad \boldsymbol{i} = 1, 2, \ \dots \ \boldsymbol{N} \qquad \text{Όπου} \quad \mu_i = E(Y_i)$$

Το παραπάνω αντιστοιχεί στο μοντέλο $E(y) = X\beta$ όπου είναι :

$$y = \begin{pmatrix} Y_1 \\ . \\ . \\ . \\ Y_N \end{pmatrix} \quad , \quad X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix} \quad \text{και} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

10.1.2 Πολλαπλή γραμμική παλινδρόμηση

Ένα μοντέλο παλινδρόμησης που έχει τουλάχιστον δύο επεξηγηματικές (ανεξάρτητες) μεταβλητές ονομάζεται μοντέλο πολλαπλής παλινδρόμησης.

Εάν οι αποκρίσεις είναι γραμμικά συσχετισμένες με καθεμία από τις επεξηγηματικές μεταβλητές και $Y_i \sim N(\mu_i, \sigma^2)$ τότε ένα κατάλληλο μοντέλο είναι :

$$E(y) = X\beta \quad \text{όπου :}$$

$$y = \begin{pmatrix} Y_1 \\ . \\ . \\ . \\ Y_N \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots & \\ 1 & x_{N1} & \cdots & x_{Nk} \end{pmatrix} \quad \text{και} \quad \beta = \begin{pmatrix} \beta_0 \\ . \\ . \\ . \\ \beta_k \end{pmatrix}$$

10.1.3 Πολυωνυμική παλινδρόμηση

Η πολυωνυμική παλινδρόμιση χρησιμοποιείται για να προσαρμόσουμε τα δεδομένα σε μία καμπύλη γραμμή. Η καμπυλόγραμμη σχέση μεταξύ της μεταβλητής απόκρισης Y και μίας μοναδικής επεξηγηματικής μεταβλητής x μπορεί να μοντελοποιηθεί με ένα πολυώνυμο.

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_{p-1} x_i^{p-1} + e_i$$

Αυτό το πολυώνυμο είναι ειδική περίπτωση του $y = X\beta + e$ όπου οι δυνάμεις των x_i συμπεριφέρονται σαν ξεχωριστές μεταβλητές.

$$y = \begin{pmatrix} Y_1 \\ . \\ . \\ . \\ Y_N \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^{p-1} \\ \vdots & & \vdots & \ldots & \\ 1 & x_N & x_N^2 & \ldots & x_N^{p-1} \end{pmatrix} \quad \text{και} \quad \beta = \begin{pmatrix} \beta_0 \\ . \\ . \\ . \\ \beta_{p-1} \end{pmatrix}$$

Δεν πρέπει να χρησιμοποιούμε περισσότερες από τρεις ή τέσσερις όρους στο πολυωνυμικό μοντέλο (δηλαδή συμπεριλαμβάνουμε όρους μέχρι το πολύ τρίτου βαθμού) διότι :

- Εάν ο βαθμός του πολυώνυμου είναι πολύ μεγάλος τότε οι στήλες του X έχουν πολύ μεγάλη σχέση μεταξύ τους και ο $X^T X$ μπορεί να είναι σχεδόν μοναδιαίος.

- Υπάρχει κίνδυνος να δημιουργηθεί ένα μοντέλο το οποίο θα προσαρμόζεται πολύ καλά στα δεδομένα μέσα στη περιοχή των παρατηρήσεων αλλά θα είναι πολύ φτωχό για προβλέψεις έξω από την περιοχή αυτή.
- Μπορεί να δημιουργήσει την ψευδαίσθηση ότι ο μηχανισμός που συνδέει τις μεταβλητές περιγράφεται από ένα πολυώνυμο υψηλού βαθμού (που δεν είναι επιθυμητό λόγω της φτωχής προβλεψιμότητας) και δεν μας αφήνει να ερευνήσουμε για μια εναλλακτική εξίσωση.

10.1.4 Τριγωνομετρική παλινδρόμηση

Εάν η σχέση μεταξύ της απόκρισης Y και μιας ερμηνευτικής μεταβλητής x είναι κυκλική ή περιοδική, τότε ένα βολικό μοντέλο θα μπορούσε να είναι το παρακάτω :

$$Y_i = \beta_0 + \beta_1 \text{συν}\alpha_1 x_i + \beta_2 \text{ημ}\alpha_2 x_i + \beta_3 \text{συν}\alpha_3 x_i + \beta_4 \text{ημ}\alpha_4 x_i + \ldots + e_i$$

Όπου τα α_j είναι γνωστές σταθερές.

Η μορφή της παραπάνω παλινδρόμησης ονομάζεται τριγωνομετρική παλινδρόμηση και μπορεί να χρησιμοποιηθεί για να μοντελοποιήσει εποχικότητα σε οικονομικά δεδομένα, καρδιακούς ρυθμούς κλπ.

Στην εξίσωση $E(y) = X\beta$ είναι :

$$X = \begin{pmatrix} 1 & \text{συν}\alpha_1 x_1 & \text{ημ}\alpha_1 x_1 & \text{συν}\alpha_2 x_1 & \text{ημ}\alpha_2 x_1 & \ldots \\ \vdots & & \vdots & & & \\ 1 & \text{συν}\alpha_1 x_N & \text{ημ}\alpha_1 x_N & \text{συν}\alpha_2 x_N & \text{ημ}\alpha_2 x_N & \ldots \end{pmatrix} \quad \text{και} \quad \beta = \begin{pmatrix} \beta_0 \\ . \\ . \\ \beta_\kappa \\ . \end{pmatrix}$$

Για τους ίδιους λόγους με αυτούς που επισημάνθηκαν στην πολυωνυμική παλινδρόμηση δεν πρέπει να έχουμε πάρα πολλούς όρους στο δεύτερο μέλος της εξίσωσης .

10.2 Εκτίμηση Μέγιστης Πιθανοφάνειας

Θεωρούμε ότι αποκρίσεις Y_i είναι ανεξάρτητες και έχουν κατανομές $Y_i \sim N\left(\mu_i, \sigma^2\right)$ όπου $\mu_\iota = x_i^T \beta$.

Τότε η λογαριθμική συνάρτηση πιθανοφάνειας είναι :

$$l = \frac{-1}{2\sigma^2}\left(y - x\beta\right)^T\left(y - x\beta\right) - \frac{N}{2}\log\left(2\pi\sigma^2\right) \qquad \textbf{(Σ8)}$$

Όπου :

$$y = \begin{pmatrix} Y_1 \\ . \\ . \\ . \\ Y_N \end{pmatrix}, \quad X = \begin{pmatrix} x_1^T \\ . \\ . \\ . \\ x_N^T \end{pmatrix} \quad \text{και} \quad \beta = \begin{pmatrix} \beta_0 \\ . \\ . \\ . \\ \beta_p \end{pmatrix}$$

Από την **(Σ8)** έχουμε : $U = \frac{\partial l}{\partial \beta} = \frac{1}{\sigma^2} X^T\left(y - X\beta\right)$

Ο εκτιμητής μέγιστης πιθανοφάνειας b του *β* δίνεται από τη λύση του $U = \frac{\partial l}{\partial \beta} = 0$

Οπότε έχουμε $X^T\left(y - X\beta\right) = 0$ συνεπώς $X^T X b = X^T y$

Στην περίπτωση που ο $X^T X$ δεν είναι μοναδιαίος οι τιμές του b είναι :

$$b = \left(X^T X\right)^{-1} X^T y$$

Έχουμε αποδείξει ότι Ε(b) = β και $E\left[\left(b - \beta\right)\left(b - \beta\right)^T\right] = \sigma^2\left(X^T X\right)^{-1}$.

Επίσης το b είναι γραμμικός συνδυασμός στοιχείων των κανονικά κατανεμημένων αποκρίσεων Y_i, οπότε θα είναι $b \sim N\left(\beta, \sigma^2\left(X^T X\right)^{-1}\right)$.

Γενικότερα, εάν η διασπορά δεν είναι ίδια για όλες τις αποκρίσεις ,δηλαδή εάν είναι $Y_i \sim N\left(x_i^T \beta, \sigma_i^2\right)$, τότε οι εκτιμητές μέγιστης πιθανοφάνειας το β είναι η λύση του $X^T V^{-1} X b = X^T V^{-1} y$ όπου V είναι ο διαγώνιος πίνακας με στοιχεία $\upsilon_{ii} = \sigma_i^2$.

Στα γενικευμένα γραμμικά μοντέλα το σ^2 συμπεριφέρεται σαν σταθερά και δεν του γίνεται εκτίμηση, οπότε η παραπάνω εξίσωση δεν είναι πλήρως καθορισμένη. Συνήθως υπολογίζονται τα β και σ^2 συγχρόνως για να δώσουν τους εκτιμητές μέγιστης πιθανοφάνειας :

$$b = \left(X^T X\right)^{-1} X^T y \quad \text{και} \quad \tilde{\sigma}^2 = \frac{1}{N}\left(y - Xb\right)^T \left(y - Xb\right)$$

Μπορεί, όμως, να αποδειχθεί ότι ο εκτιμητής $\tilde{\sigma}^2$ δεν είναι αμερόληπτος. Μπορούμε να δείξουμε ότι $E\left(\tilde{\sigma}^2\right) = \dfrac{(N-p)\sigma^2}{N}$ οπότε ένας αμερόληπτος εκτιμητής του σ^2 μπορεί να οριστεί από την σχέση :

$$\tilde{\sigma}^2 = \frac{1}{N-p}\left(y - Xb\right)^T \left(y - Xb\right)$$

Με τους παραπάνω τύπους μπορούμε να κάνουμε έλεγχο υποθέσεων για το β καθώς και υπολογισμό διαστημάτων εμπιστοσύνης.

10.3 Εκτίμηση Ελαχίστων τετραγώνων

Εάν έχουμε το μοντέλο παλινδρόμησης $E(y) = X\beta$ και γνωρίζουμε το V, όπου $E\left[\left(y - X\beta\right)\left(y - X\beta\right)^T\right] = V$, τότε μπορούμε να εξάγουμε την εκτιμήτρια ελαχίστων τετραγώνων του β χωρίς να κάνουμε άλλες επιπλέον υποθέσεις για τη κατανομή της απόκρισης y.

Ελαχιστοποιούμε το : $S_w = \left(y - X\beta\right)^T V \left(y - X\beta\right)$

Έχουμε ότι $\frac{\partial s_w}{\partial \beta} = -2X^T V^{-1}(y - X\beta)$ και θέλουμε να ισχύει $\frac{\partial s_w}{\partial \beta} = 0$.

Η λύση είναι $b = \left(X^T V^{-1} X\right)^{-1} X^T V^{-1} y$

(Με την προϋπόθεση να υπάρχει ο αντίστροφος πίνακας.)

Στην ειδική περίπτωση που τα στοιχεία του y είναι ανεξάρτητα και έχουν κοινή διακύμανση, τότε :

$$b = \left(X^T X\right)^{-1} X^T y$$

Συμπερασματικά, μπορούμε να πούμε ότι, για μοντέλα παλινδρόμησης που έχουν σφάλματα που ακολουθούν κανονική κατανομή, οι εκτιμητές μέγιστης πιθανοφάνειας και ελαχίστων τετραγώνων συμπίπτουν.

10.4 Στατιστική συνάρτηση Απόκλισης

Ο υπολογισμός της απόκλισης (Deviance) γίνεται βάσει ορισμού με τη διαφορά $D = 2\left[l\left(b_{\max}; y\right) - l\left(b; y\right)\right]$.

Στο πλήρες μοντέλο έχουμε $\beta_{\max} = \begin{pmatrix} \beta_1 \\ . \\ . \\ . \\ \beta_N \end{pmatrix}$

Για απλούστερους υπολογισμούς μπορούμε να υποθέσουμε χωρίς βλάβη της γενικότητας ότι ο X στο πλήρες μοντέλο είναι μοναδιαίος πίνακας οπότε $b_{max} = y$.

Από την σχέση $l = \frac{-1}{2\sigma^2}(y - x\beta)^T (y - x\beta) - \frac{N}{2}\log\left(2\pi\sigma^2\right)$ με αντικατάσταση του $b_{max} = y$ έχουμε την εξίσωση :

$$l\left(b_{\max};y\right)=\frac{1}{2}N\log(2\pi\sigma^{2})$$

Για κάθε άλλο μοντέλο $E\left(y\right)=X\beta$ στο οποίο εμπλέκονται p παράμετροι όπου $\beta=\begin{pmatrix}\beta_1\\ .\\ .\\ .\\ \beta_p\end{pmatrix}$ με p < N, και με ***b*** να συμβολίζεται ο εκτιμητής μέγιστης πιθανοφάνειας έχουμε : $l=\frac{-1}{2\sigma^{2}}\left(y-x\beta\right)^{T}\left(y-x\beta\right)-\frac{N}{2}\log\left(2\pi\sigma^{2}\right)$

Οπότε η στατιστική συνάρτηση Απόκλισης είναι :

$$D=2\left[l\left(b_{\max};y\right)-l\left(b;y\right)\right]$$

$$=\frac{1}{\sigma^{2}}\left(y-Xb\right)^{T}\left(y-Xb\right)$$

$$=\frac{1}{\sigma^{2}}\left(y^{T}y-2b^{T}X^{T}y+b^{T}X^{T}Xb\right)\quad(\text{επειδή ισχύει } X^{T}Xb=X^{T}y)$$

$$=\frac{1}{\sigma^{2}}\left(y^{T}y-b^{T}X^{T}y\right)$$

Εάν το μοντέλο είναι σωστό τότε το $D\sim x_{N-p}^{2}$, ενώ στην αντίθετη περίπτωση το D ακολουθεί την μη–κεντρική χ τετράγωνο κατανομή με N-p βαθμούς ελευθερίας.

Όταν το σ^2 είναι άγνωστο η στατιστική συνάρτηση D δεν μπορεί να καθοριστεί πλήρως. Στον έλεγχο υποθέσεων προσπερνάμε αυτή τη δυσκολία χρησιμοποιώντας το πηλίκο των D.

Για να χρησιμοποιήσουμε μοντέλα παλινδρόμησης σε έλεγχο υποθέσεων θα πρέπει το πιο γενικό μοντέλο να περιγράφει καλά τα δεδομένα και συνεπώς το αντίστοιχο D στοιχείο έχει την κεντρική χ τετράγωνο κατανομή. Η προϋπόθεση αυτή θα πρέπει να ελεγχθεί (για παράδειγμα, εξετάζοντας τα κατάλοιπα).

Υποθέτουμε μία μηδενική υπόθεση H_0 και την εναλλακτική H_1. Αυτό μπορεί να εκφραστεί με όρους μοντέλων ως εξής :

$$H_0 : \beta = \beta_0 = \begin{pmatrix} \beta_1 \\ . \\ . \\ . \\ \beta q \end{pmatrix} \quad \text{και} \quad H_1 : \beta = \beta_1 = \begin{pmatrix} \beta_1 \\ . \\ . \\ . \\ \beta p \end{pmatrix} \quad \text{όπου } q < p < N .$$

Το μοντέλο που αντιστοιχεί στην H_1 είναι το πιο γενικό ($q < p$).

Έστω ότι X_0 και X_1 συμβολίζουν τους αντίστοιχους πίνακες σχεδιασμού , b_0, b_1 οι εκτιμητές μέγιστης πιθανοφάνειας, και D_0, D_1 οι στατιστικές συναρτήσεις απόκλισης.

Ο έλεγχος H_0 έναντι H_1 γίνεται χρησιμοποιώντας το ΔD.

$$\Delta D = D_0 - D_1 = \frac{1}{\sigma^2}\left[\left(y^T y - b_0^T X_0^T y\right) - \left(y^T y - b_1^T X_1^T y\right)\right]$$

$$= \frac{1}{\sigma^2}\left(b_1^T X_1^T y - b_0^T X_0^T y\right)$$

Εάν η η H_1 είναι σωστή οπότε $D_1 \sim x_{N-p}^2$. Εάν και η H_0 είναι επίσης σωστή τότε $D_0 \sim x_{N-q}^2$, αλλιώς η D_0 έχει μη-κεντρική χ-τετράγωνο κατανομή με N-q βαθμούς ελευθερίας.

Συνεπώς εάν η H_0 είναι σωστή τότε $D_0 - D_1 \sim x_{p-q}^2$ και συνεπώς

$$f = \frac{\dfrac{D_0 - D_1}{p-q}}{\dfrac{D_1}{N-p}} = \frac{\dfrac{b_1^T X_1^T y - b_0^T X_0^T y}{p-q}}{\dfrac{y^T y - b_1^T X_1^T y}{N-p}} \sim F_{p-q,N-p}$$

Εάν η H_0 δεν είναι σωστή τότε η f θα έχει μη-κεντρική κατανομή. Συνεπώς για τιμές της f που είναι μεγάλες σε σχέση με την κατανομή $F_{p-q,N-p}$ δεν μπορούμε να αποδεχθούμε την H_0.

Ο έλεγχος παρουσιάζεται στον παρακάτω πίνακα.

Πίνακας 9 : Πίνακας διασποράς

Ανάλυσης Διασποράς			
Πηγή Διακύμανσης	Βαθμοί Ελευθερίας	Άθροισμα Τετραγώνων	Μέσο Αθρ. Τετραγώνων
Μοντέλο με β_0	q	$b_0^T X_0^T y$	
Βελτίωση οφειλόμενη στο μοντέλο με β_1	p-q	$b_1^T X_1^T y - b_0^T X_0^T y$	$\dfrac{b_1^T X_1^T y - b_0^T X_0^T y}{p-q}$
Κατάλοιπα	N-p	$y^T y - b_1^T X_1^T y$	$\dfrac{y^T y - b_1^T X_1^T y}{N-p}$
Σύνολο	N	$y^T y$	

10.5 Συντελεστής Πολλαπλής συσχέτισης και R^2

Μια λιγότερο αυστηρή σύγκριση για την καλή προσαρμογή μεταξύ δύο μοντέλων μας εφοδιάζει το τετράγωνο του συντελεστή πολλαπλής συσχέτισης R^2.

Για την $y = X\beta + e$ με τις ιδιότητες :

1. τα στοιχεία e είναι ανεξάρτητα
2. $E(e_i) = 0$ και $var(e_i) = \sigma^2$ για όλα τα i.
3. το κριτήριο των ελαχίστων τετραγώνων είναι $S = \sum_{i=1}^{N} e_i^2 = e^T e = (y - X\beta)^T (y - X\beta)$

Η ελάχιστη τιμή του S για το μοντέλο είναι $\hat{S} = (y - Xb)^T (y - Xb) = y^T y - b^T X^T y$. Η παραπάνω τιμή μπορεί να χρησιμοποιηθεί για να μετρηθεί η προσαρμογή του μοντέλου .

Θα καθορίσουμε τον συντελεστή R^2 χρησιμοποιόντας το ελάχιστο μοντέλο. Η τιμή του S συγκρίνεται με τη τιμή του ελάχιστου μοντέλου.

Ελάχιστο μοντέλο είναι αυτό το οποίο ορίζεται ως : $E(Y_i)=\mu$ για όλα τα i.
Για το ελάχιστο μοντέλο ισχύει:

$E(y)=X\beta$ όπου $\beta=[\mu]$ και X=1,
όπου το 1 είναι ο Nx1 πίνακας με μονάδες.

Συνεπώς $X^T X = N$, $X^T y = \sum y_i$ και $b = \hat{\mu} = \bar{y}$.

Οι τιμές που αντιστοιχούν από το κριτήριο των ελαχίστων τετραγώνων είναι :
$\hat{S}_0 = y^T y - N\bar{y}^2 = \sum_{i=1}^{N} (y_i - \bar{y})^2$ όπου το $\hat{S}_0$ είναι ανάλογο με τη διακύμανση των παρατηρήσεων και θεωρείται σαν η χειρότερη δυνατή τιμή του S.

Κάθε άλλο μοντέλο με τιμή $\hat{S}$ υπολογίζεται σε σχέση με το ελάχιστο μοντέλο χρησιμοποιώντας το $\hat{S}_0$.
Η διαφορά $\hat{S}_0 - \hat{S} = b^T X^T y - N\bar{y}^2$ είναι η βελτίωση στη προσαρμογή που προκύπτει από το μοντέλο $E(y)=X\beta$.

Το R^2 ορίζεται ως το πηλίκο : $R^2 = \frac{\hat{S}_0 - \hat{S}}{\hat{S}_0} = \frac{b^T X^T y - N\bar{y}^2}{y^T y - N\bar{y}^2}$
και μεταφράζεται σαν το ποσοστό της συνολικής μεταβλητότητας στα δεδομένα που εξηγείται από το μοντέλο.

Το R^2, γενικά, παίρνει τιμές από 0 μέχρι 1 δηλ. $0 < R^2 < 1$. Η τετραγωνική ρίζα του R^2 ονομάζεται συντελεστής πολλαπλής συνδια κύμανσης.

Εάν το μοντέλο δεν περιγράφει τα δεδομένα καλύτερα από το ελάχιστο μοντέλο τότε θα είναι $\hat{S}_0 \cong \hat{S}$ οπότε $R^2 \cong 0$.

Εάν χρησιμοποιείται το πλήρες μοντέλο με Ν παραμέτρους, τότε ο X θα είναι ένας NxN μοναδιαίος πίνακας I, έτσι ώστε : $b=y$ και $b^T X^T y = y^T y$ και συνεπώς $R^2=1$ που αντιστοιχεί σε μία τέλεια προσαρμογή.

Μειονεκτήματα της χρήσης του R^2

Η αδυναμία για εύκολο προσδιορισμό της δειγματικής κατανομής του R^2.

Η τιμή του δεν μεταβάλεται με την αλλαγή του αριθμού των παραμέτρων που χρησιμοποιούνται στο προσαρμοσμένο μοντέλο.

Παράδειγμα 10

Στον παρακάτω πίνακα εμφανίζονται αποκρίσεις που είναι ποσοστά από συνολικές θερμίδες που προέρχονται από σύνθετους υδρογονάνθρακες για είκοσι άρρενες διαβητικούς εξαρτημένους από την ινσουλίνη, οι οποίοι έχουν εφαρμόσει για έξη μήνες δίαιτα υψηλών υδρογονανθράκων.

Η συμμόρφωση με τη δίαιτα νομίζεται ότι έχει σχέση με την ηλικία (σε χρόνια), σωματικό βάρος και από κάποια σαν πρωτεΐνη.

Πίνακας 10 : Δεδομένα δίαιτας υδρογονανθράκων

Α/Α	Υδρογονάνθρακες Y	Ηλικία x_1	Βάρος x_2	Πρωτεΐνη x_3
1	33	33	100	14
2	40	47	92	15
3	37	49	135	18
4	27	35	144	12
5	30	46	140	15
6	43	52	101	15
7	34	62	95	14
8	48	23	101	17

9	30	32	98	15
10	38	42	105	14
11	50	31	108	17
12	51	61	85	19
13	30	63	130	19
14	36	40	127	20
15	41	50	109	15
16	42	64	107	16
17	46	56	117	18
18	24	61	100	13
19	35	48	118	18
20	37	28	102	14

Θα προσαρμόσουμε το μοντέλο $E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$, στο οποίο οι υδρογονάνθρακες *Y* είναι γραμμικά σχετισμένοι με την ηλικία x_1, το βάρος x_2 και την πρωτεΐνη x_3. Εάν

$$y = \begin{pmatrix} Y_1 \\ . \\ . \\ . \\ Y_N \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ \vdots & & \vdots & \\ 1 & x_{N1} & x_{N2} & x_{N3} \end{pmatrix} \quad \text{και} \quad \beta = \begin{pmatrix} \beta_0 \\ . \\ . \\ . \\ \beta 3 \end{pmatrix}$$

Τότε για τα δεδομένα του **πίνακα 10** έχουμε :

$$X^T y = \begin{pmatrix} 752 \\ 34596 \\ 82270 \\ 12105 \end{pmatrix} \quad \text{και} \quad X^T X = \begin{pmatrix} 20 & 923 & 2214 & 318 \\ 923 & 45697 & 103003 & 14780 \\ 2214 & 102003 & 250346 & 35306 \\ 318 & 14780 & 35306 & 5150 \end{pmatrix}$$

Συνεπώς η λύση της εξίσωσης $X^T X b = X^T y$ είναι :

$$b = \begin{pmatrix} 36.9601 \\ -0.1137 \\ -0.2280 \\ 1.95770 \end{pmatrix} \quad \text{και} \quad \left(X^T X\right)^{-1} = \begin{pmatrix} 4.8158 & -0.0113 & -0.0188 & -0.1362 \\ -0.0113 & 0.0003 & 0.0000 & -0.0004 \\ -0.0188 & 0.0000 & 0.0002 & -0.0002 \\ -0.1362 & -0.0004 & -0.0002 & 0.0114 \end{pmatrix}$$

Όπου έχουμε προσέγγιση τεσσάρων δεκαδικών ψηφίων.

Έχουμε επίσης τους παρακάτω υπολογισμούς

$y^T y = 29368$, $b^T X^T y = 28800.337$ και $N\bar{y}^2 = 28275.2$, οπότε έχουμε $R^2 = 0.48$, δηλαδή το 48% της συνολικής μεταβλητότητας των δεδομένων εξηγείται από το μοντέλο.

Χρησιμοποιώντας την $\tilde{\sigma}^2 = \dfrac{1}{N-p}(y - Xb)^T (y - Xb)$ για να αποκτήσουμε ένα αμερόληπτο εκτιμητή του σ^2 , λαμβάνουμε $\hat{\sigma}^2 = 35.479$, και τα τυπικά σφάλματα για τα στοιχεία του b δείχνονται στον παρακάτω πίνακα.

Πίνακας 11 : Πίνακας ANOVA για τα δεδομένα της δίαιτας υδρογονανθράκων

Όρος	Εκτίμηση bj	Τυπικό σφάλμα (SE)
Σταθερά	36.960	13.071
Συντελεστής για Ηλικία	-0.114	0.109
Συντελεστής για βάρος	-0.228	0.083
Συντελεστής για πρωτεΐνη	1.958	0.635

Λύση του προβλήματος με R

```
> Y<-c(33, 40, 37, 27, 30, 43, 34, 48, 30, 38, 50, 51, 30, 36, 41, 42, 46, 24, 35, 37)
> X1<-c(33, 47, 49, 35, 46, 52, 62, 23, 32, 42, 31, 61, 63, 40, 50, 64, 56, 61, 48, 28)
> X2<-c(100, 92, 135, 144, 140, 101, 95, 101, 98, 105, 108, 85, 130, 127, 109, 107, 117, 100, 118, 102)
> X3<-c(14, 15, 18, 12, 15, 15, 14, 17, 15, 14, 17, 19, 19, 20, 15, 16, 18, 13, 18, 14)

> fit.model<-lm(Y~X1+X2+X3)
> fit.model

Call:
lm(formula = Y ~ X1 + X2 + X3)
```

```
Coefficients:
(Intercept)          X1          X2               X3
    36.9601      -0.1137      -0.2280       1.9577

> summary(fit.model)

Call:
lm(formula = Y ~ X1 + X2 + X3)

Residuals:
     Min       1Q       Median     3Q         Max
-10.3424  -4.8203   0.9897   3.8553   7.9087

Coefficients:
                Estimate       Std. Error      t       value Pr(>|t|)
(Intercept) 36.96006       13.07128     2.828      0.01213 *
X1              -0.11368        0.10933   -1.040      0.31389
X2              -0.22802        0.08329   -2.738      0.01460 *
X3               1.95771        0.63489     3.084       0.00712 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 5.956 on 16 degrees of freedom

Multiple R-squared: 0.4805,   Adjusted R-squared: 0.3831

F-statistic: 4.934 on 3 and 16 DF,   p-value: 0.01297

Παράδειγμα 11

Θα παρουσιάσουμε τη χρήση της στατιστικής συνάρτησης απόκλισης με χρήση των δεδομένων του προηγούμενου παραδείγματος 10 (Πίνακας 10) για τον παρακάτω έλεγχο υποθέσεων.

H_0 : Η απόκριση δεν εξαρτάται από την ηλικία (δηλ. $\beta_1 = 0$)

H_1 : Η απόκριση εξαρτάται από την ηλικία ($\beta_1 \neq 0$)

Το μοντέλο που αντιστοιχεί στη μηδενική υπόθεση είναι :

$E(Y_i) = \beta_0 + \beta_2 x_{i2} + \beta_3 x_{i3}$ (H_0)

Ο πίνακας X για το μοντέλο αυτό προκύπτει από τον πίνακα του προηγούμενου μοντέλου παραλείποντας τη δεύτερη στήλη , οπότε έχουμε τα παρακάτω αποτελέσματα.

$$X^T y = \begin{pmatrix} 752 \\ 82270 \\ 12195 \end{pmatrix}, \quad X^T X = \begin{pmatrix} 20 & 2214 & 318 \\ 2214 & 250346 & 35306 \\ 318 & 35306 & 5150 \end{pmatrix} \quad \text{και συνεπώς} \quad b = \begin{pmatrix} 33.130 \\ -0.222 \\ 1.824 \end{pmatrix}$$

Για το μοντέλο H_0 έχουμε λοιπόν $b^T X^T y = 28761.978$ και είναι $R^2 = 0.445$ δηλαδή το 44.5% της μεταβλητότητας μπορεί να εξηγηθεί από το μοντέλο. Ο έλεγχος για το H_0 συνοψίζεται τον παρακάτω πίνακα.

Πίνακας 12 : Σύνοψη του ελέγχου «Η απόκριση δεν εξαρτάται από την ηλικία»

Πηγή Διακύμανσης	Βαθμοί ελευθερίας	Άθροισμα τετραγώνων	Μέσο Άθροισμα τετραγώνων
Μοντέλο H_0	3	28761.978	
Βελτίωση οφειλόμενη στο μοντέλο H_1	1	38.359	38.36
Κατάλοιπα	16	567.663	35.48
Σύνολο	20	29368.000	

Η τιμή $f = 38.36 / 35.48 = 1.08$ δεν είναι στατιστικά σημαντική συγκρινόμενη με την $F_{1,16}$ κατανομή οπότε τα δεδομένα δεν μας δίνουν αποδείξεις εναντίον της H_0, οπότε δεν μπορούμε να απορρίψουμε την H_0 και συνεπώς φαίνεται ότι η απόκριση να είναι ανεξάρτητη από την ηλικία.

<u>Η σύγκριση των μοντέλων με χρήση της R γίνεται ως εξής</u>

```
>fit.model<-lm(Y~X1+X2+X3)
> fit.model1<-lm(Y~X2+X3)
```

```
> sygr<-anova(fit.model, fit.model1)
> sygr
Analysis of Variance Table

Model 1: Y ~ X1 + X2 + X3
Model 2: Y ~ X2 + X3

  Res.Df    RSS       Df     Sum of Sq      F        Pr(>F)
1     16      567.66
2     17      606.02        -1    -38.36          1.0812    0.3139
```

10.6 Γραφήματα καταλοίπων

Χρησιμοποιούμε τα γραφήματα των καταλοίπων καθώς και των τυποποιημένων καταλοίπων για να ελέγξουμε κανονικότητα καταλοίπων και πιθανό λανθασμένο καθορισμό του μοντέλου.

Για το μοντέλο παλινδρόμησης $y = X\beta + e$ έχουμε τις προϋποθέσεις ότι οι όροι λαθών $\boldsymbol{e_i}$ είναι ανεξάρτητοι και κανονικά κατανεμημένοι με μέσο **0** (μηδέν) και διακύμανση $\boldsymbol{\sigma^2}$ και δεν μεταβάλλονται συστηματικά με το $\boldsymbol{y}$ ή με όρους του $\boldsymbol{X}$**.**

Αυτές οι υποθέσεις μπορεί να ελεγχθούν μελετώντας τα κατάλοιπα

$$\hat{e}_i = y_i - x_i^T b$$

Εάν το μοντέλο είναι σωστό, τότε το διάνυσμα των καταλοίπων θα έχει τις παρακάτω ιδιότητες :

$E(\hat{e}) = 0$ και $E\left(\hat{e}\hat{e}^T\right) = E\left(yy^T\right) - XE\left(bb^T\right)X^T = \sigma^2\left[I - X\left(X^TX\right)^{-1}X^T\right]$ όπου ο I είναι ο μοναδιαίος πίνακας.

Τα τυποποιημένα κατάλοιπα ορίζονται από τη σχέση :

$r_i = \dfrac{\hat{e}_i}{\hat{\sigma}(1-p_{ii})^{1/2}}$ όπου το p_{ii} είναι το i–οστο στοιχείο της διαγωνίου της προβολής ή αλλοιώς *του πίνακα καπέλο* $P = X\left(X^T X\right)^{-1} X^T$.

Τα γραφήματα πιθανοτήτων των καταλοίπων μπορεί να χρησιμοποιηθούν για να εκλεχθεί η προϋπόθεση κανονικότητας. Τα τυποποιημένα κατάλοι πα θα πρέπει να είναι οριακά μη συσχετισμένα. Οπότε εάν παρατηρηθούν σημαντικές σειριακές συσχετίσεις μεταξύ τους, είναι πολύ πιθανόν να υποδεικνύουν λανθασμένο καθορισμό του μοντέλου. Κάνουμε συχνά έλεγχο για σειριακές συσχετίσεις (χρησιμοποιώντας πχ. το τεστ του Darbin – Watson).

Πρέπει να δημιουργούμε γραφήματα των τυποποιημένων κατάλοιπων με τις προσαρμοσμένες τιμές $\hat{y}_i = x_i^T b$ και με κάθε μία από τις επεξηγηματικές μεταβλητές. Μορφές σχημάτων σε αυτά τα γραφήματα μπορεί να υποδηλώνουν λανθασμένο καθορισμό του μοντέλου[7].

10.7 Ορθογωνιότητα (Orthogonality)

Όταν προσαρμόζουμε τα δεδομένα στα μοντέλα, οι εκτιμήσεις των παραμέτρων εξαρτώνται από τους όρους που συμπεριλαμβάνουμε στο μοντέλο. Για παράδειγμα οι εκτιμήσεις των β_0, β_2 διαφέρουν στα παρακάτω μοντέλα :

$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3}$ και $E(Y_i) = \beta_0 + \beta_3 x_{i3}$ (που δεν συμπεριλαμβάνουν $\beta_2 x_{i2}$)

Ο έλεγχος για την υπόθεση β_1=0 επίσης, εξαρτάται από τους όρους που έχουμε συμπεριλάβει στο μοντέλο.

Οι εκτιμήσεις, τα διαστήματα εμπιστοσύνης και οι έλεγχοι υποθέσεων εξαρτώνται συνήθως από το ποιες μεταβλητές συμπεριλαμβάνονται στο μοντέλο.
Έχουμε όμως μία περίπτωση που δεν συμβαίνει αυτό. Η εξαίρεση αυτή συμβαίνει όταν ο πίνακας *X είναι ορθογώνιος*.

Ένας πίνακας είναι ορθογώνιος όταν μπορεί να διαμεριστεί σε συνιστώσες $X_1, X_2, \ldots, X_m$ που αντιστοιχούν σε υπομοντέλα.

Δηλαδή $X = \left[X_1, X_2, \ldots, X_m\right]$ όπου $m \le p$,

με την ιδιότητα $X_j^T X_k = 0$, όπου 0 είναι πίνακας με μηδενικά για κάθε j≠k.

Έστω ότι ο $\beta = \begin{pmatrix} \beta_0 \\ . \\ . \\ . \\ \beta_m \end{pmatrix}$ είναι η διαμέριση των παραμέτρων. Τότε

$E(y) = X\beta = X_1\beta_1 + \ldots + X_m\beta_m$ και ο $X^T X$ είναι ο διαγώνιος πίνακας

$$X^T X = \begin{pmatrix} X_1^T X_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & X_m^T X_m \end{pmatrix} \text{ και συνεπώς } X^T y = \begin{pmatrix} X_1^T y \\ . \\ . \\ . \\ X_m^T y \end{pmatrix}$$

όπου το 0 χρησιμοποιείται για να δείξει ότι τα υπόλοιπα στοιχεία του πίνακα είναι μηδενικά.

Έχουμε : $b^T X^T y = b_1^T X_1^T y + \ldots + b_m^T X_m^T y$

Συνεπώς οι εκτιμήσεις $b_j = \left(X_j^T X_j\right)^{-1} X_j^T y$ παραμένουν αμετάβλητες με την παράλειψη ή προσθήκη συνιστωσών στο μοντέλο.

Συνεπώς οι υποθέσεις H_1: $\beta_1=0, \ldots,$ H_m: $\beta_m=0$ μπορεί να ελεγχθούν ανεξάρτητα.

Πίνακας 13: Έλεγχος υποθέσεων όταν ο πίνακας σχεδιασμού X είναι ορθογώνιος.

Πηγή Διακύμανσης	Βαθμοί Ελευθερίας	Άθροισμα Τετραγώνων
Μοντέλο που αντιστοιχεί στο H_1	p_1	$b_1^T X_1^T y$
….	…	….
Μοντέλο που αντιστοιχεί στο H_m	p_m	$b_m^T X_m^T y$
Κατάλοιπα	$N - \sum_{j=1}^{m} p_j$	$y^T y - b^T X^T y$
Σύνολο	N	$y^Y y$

<u>Παρατήρηση</u>

Τα πλεονεκτήματα όμως της καθετότητας, μπορούμε να εκμεταλλευτούμε μόνο εάν ο πίνακας X έχει σχεδιαστεί για να έχει αυτή την ιδιότητα. Αυτό μπορεί να πραγματοποιηθεί εάν τα στοιχεία του X είναι ψευδομεταβλητές που αντιστοιχούν σε επίπεδα παραγόντων.

10.8 Συγγραμμικότητα

Εάν οι επεξηγηματικές μεταβλητές είναι σχεδόν συσχετισμένες μεταξύ τους τότε οι στήλες του πίνακα X μπορεί να είναι περίπου γραμμικά εξαρτημένες, οπότε ο πίνακας $X^T X$ θα είναι σχεδόν μοναδιαίος. Σε αυτή την περίπτωση η εξίσωση $X^T X b = X^T y$ θα δίνει λύση $\boldsymbol{b}$ που θα είναι ασταθής, με την έννοια ότι μικρές μεταβολές στα δεδομένα μπορεί να προκαλέσουν μεγάλες μεταβολές στη λύση $\boldsymbol{b}$. Μερικά τουλάχιστον από τα στοιχεία του $\sigma^2 \left(X^T X \right)^{-1}$ επίσης, θα έχουν μεγάλες αποκρίσεις σε εκτιμήσεις του $\boldsymbol{b}$ με μεγάλες διακυμάνσεις και συνδιακυμάνσεις. Συνεπώς προσεκτική έρευνα του πίνακα $\left(X^T X \right)^{-1}$ μπορεί να εμφανίσει την ύπαρξη συγγραμμικότητας.

Η ανάλυση του προβλήματος είναι περισσότερο δύσκολη από την ανάλυση που κάναμε στην παραπάνω παράγραφο. Μπορεί να απαιτείται επιπλέον πληροφορία από τη περιοχή από την οποία προέρχονται τα δεδομένα , ένας εναλλακτικός καθορισμός του μοντέλου ή κάποιες άλλες μη υπολογιστικές προσεγγίσεις.

Το πρόβλημα με τη συγγραμμικότητα εμφανίζεται ιδιαίτερα όταν θέλουμε να επιλέξουμε μερικά υποσύνολα από επεξηγηματικές μεταβλητές οι οποίες περιγράφουν «καλύτερα» τα δεδομένα. Εάν δύο μεταβλητές είναι πολύ συσχετισμένες, είναι πολύ δύσκολο χρησιμοποιώντας μόνο στατιστικούς όρους να καθορίσουμε ποια πρέπει να συμπεριληφθεί στο μοντέλο.

10.9 Μη Γραμμική Παλινδρόμηση

Με τον όρο «μη–γραμμική παλινδρόμηση» εννοούμε τη χρήση επαναληπτικών μεθόδων που είναι ανάλογες με αυτές που χρησιμοποιούνται στα γενικευμένα γραμμικά μοντέλ α

Η μη–γραμμική παλινδρόμηση χρησιμοποιείται για δύο τύπους μοντέλων. Ο πρώτος τύπος είναι τα μοντέλα της μορφής $E(Y)=g\left(x^T\beta\right)$ τα οποία είναι γενικευμένα γραμμικά μοντέλα υπό τον όρο η κατανομή του Y να είναι μέλος της εκθετικής οικογένειας και η συνάρτηση σύνδεσης είναι μονότονη.

Σαν παράδειγμα αναφέρουμε την συνάρτηση που δίνει την απόδοση [8] (σοδειά) των φυτών:

$$E(Y)=\frac{1}{\beta_0+\beta_1 x+\beta_2 x^2}$$

όπου Y είναι η σοδειά ανά ανά φυτό και x είναι ένα μέτρο για την πυκνότητα των φυτών. Εάν η Y είναι κανονικά κατανεμημένη τότε μπορεί να χρησιμοποιηθούν οι μέθοδοι των γενικευμένων γραμμικών μοντέλων για εκτίμηση και συμπερασματολογία.

Ο δεύτερος τύπος μοντέλων μη–γραμμικής παλινδρόμησης, είναι της μορφής $E(Y) = g(x, \beta)$ όπου η g είναι εσωτερικά μη–γραμμική στις παραμέτρους. Για παράδειγμα το λογιστικό μοντέλο ανάπτυξης :

$$E(Y) = \frac{\beta_0}{1 + \beta_1 \exp(\beta_2 x)}$$

Τα αποτελέσματα, όμως, για τις κατανομές που ισχύουν για τα γενικευμένα γραμμικά μοντέλα δεν εφαρμόζονται Για παράδειγμα, οι δειγματικές κατανομές των εκτιμητριών μπορεί να μην είναι κανονικές.

11 Ανάλυση Διακύμανσης και Συνδιακύμανσης

11.1 Εισαγωγή

Η ανάλυση διακύμανσης είναι μία μέθοδος που ασχολείται με τον έλεγχο της υπόθεσης αν «πολλές ομάδες δεομένων έχουν τον ίδιο μέσο» .
Στην μεθοδολογία της ανάλυσης διακύμανσης αναπτύσσεται μία μέθοδος που ανιχνεύει τις διαφορές μεταξύ των μέσων των ομάδων. Συνεπώς μελετάμε τις πηγές των αποκλίσεων που παρατηρούνται στα δεδομένα.

Η μεθοδολογία αυτή μπορεί να εκφραστεί με ένα γραμμικό μ οντέλο το οποίο έχει τη μορφή $y = X\beta + e$ όπου είναι $e \sim N\left(0, \sigma^2 I\right)$ όπου $\boldsymbol{y}$, $\boldsymbol{e}$ είναι τυχαία διανύσματα διαστάσεως N και X είναι Nxp πίνακας σταθερών, $\boldsymbol{\beta}$ είναι διάνυσμα από $\boldsymbol{p}$ παραμέτρους και $\boldsymbol{I}$ είναι ο μοναδιαίος πίνακας.

Τα μοντέλα της ανάλυσης διακύμανσης (ANOVA) διαφέρουν από τα μοντέλα παλινδρόμησης που αναφέρθηκαν προηγουμένως στο ότι ο πίνακας σχεδιασμού X αποτελείται αποκλειστικά από ψευδομεταβλητές. Στην ανάλυση συνδιακύμανσης (ANCOVA) ο πίνακας σχεδιασμού X αποτελείται από ψευδομεταβλητές και μεταβλητές.

Επειδή η επιλογή των ψευδομεταβλητών είναι αρκετά αυθαίρετη , ένα σημαντικό θέμα είναι η καλύτερη επιλογή του X.
Τα κύρια ερωτήματα που τίθενται με την ανάλυση διακύμανσης και συνδιακύμανσης συνεπάγονται συγκρίσεις μέσων.

Σε αυτή τη μελέτη θεωρούμε μοντέλα σταθερών αποτελεσμάτων στα οποία τα επίπεδα των παραγόντων θεωρούνται σταθερά, οπότε το $\boldsymbol{\beta}$ είναι ένα διάνυσμα από σταθερές. Δεν μελετάμε μοντέλα με τυχαία αποτελέσματα όπου τα επίπεδα των παραγόντων θεωρούνται σαν μία τυχαία επιλογή από τον πληθυσμό των δυνατών επιπέδων και το $\boldsymbol{\beta}$ το χειριζόμαστε σαν ένα διάνυσμα τυχαίων μεταβλητών.

Το πρόβλημα της εκτίμησης διακυμάνσεων για τα στοιχεία του $\boldsymbol{\beta}$ στα μοντέλα με τυχαία αποτελέσματα (που ονομάζονται και μοντέλα με μεταβλητές συνιστώσες) συζητείται στα βιβλία των McCullagh & Nelder (1989) στα πλαίσια των γενικευμένων γραμμικών μοντέλων..

Τα στοιχεία του διανύσματος αποκρίσεων $\boldsymbol{y}$ επίσης, υποτίθεται ότι είναι ανεξάρτητα και συνεπώς δεν υποθέτουμε περιπτώσεις που εμπλέκουν επαναλαμβανόμενες μετρήσεις.

<u>Θεωρητικό υπόβαθρο</u>

Στα ANOVA και ANCOVA μοντέλα υποθέτουμε ότι η τυχαία συνιστώσα $\boldsymbol{e}$ ακολουθεί την κανονική κατανομή. Η υπόθεση αυτή μας δίνει κάποιους υπολογισμούς που ήδη έχουμε κάνει και θα χρησιμοποιήσουμε στη συνέχεια.

1. Η συνάρτηση λογαριθμικής πιθανοφάνειας είναι :

$$l = -\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) - \frac{N}{2}\log\left(2\pi\sigma^2\right)$$

Συνεπώς η εκτιμήτρια μέγιστης πιθανοφάνειας ή των ελάχιστων τετράγωνων είναι η λύση των κανονικών εξισώσεων :

$$X^T Xb = X^T y \qquad \textbf{(Σ8)}$$

Τα μοντέλα ANOVA περιέχουν συνήθως περισσότερες παραμέτρους από ανεξάρτητες εξισώσεις. Συνεπώς στο $E(y) = X\beta$ το $X^T X$ είναι μοναδιαίο και δεν υπάρχει μοναδική λύση για το **(Σ8)**. Στην περίπτωση αυτή λέμε ότι το $\boldsymbol{\beta}$ δεν είναι εκτιμήσιμο.

Για να μπορέσουμε να πάρουμε μία ειδική λύση χρησιμοποιούνται επιπλέον εξισώσεις, έτσι ώστε το b να είναι η λύση του συστήματος :

$$X^T Xb = X^T y \quad \text{και} \quad Wb = 0 .$$

Οι εξισώσεις $W\beta = 0$ ονομάζονται εξισώσεις περιορισμών.

Η ελάχιστη τιμή του όρου $\left(y - X\beta\right)^T\left(y - X\beta\right)$ όμως είναι μοναδική και προκύπτει χρησιμοποιώντας κάθε λύση του **(Σ8)** οπότε η τιμή του $\left(y - Xb\right)^T\left(y - Xb\right)$ δεν εξαρτάται από την επιλογή των εξισώσεων περιορισμών. Άλλες, όμως, ιδιότητες του ***b*** εξαρτώνται από την επιλογή του *W*.

2. Αποδείξαμε προηγουμένως ότι για το πλήρες μοντέλο με *N* παραμέτρους, η εκτιμήτρια μέγιστης πιθανοφάνειας είναι $b_{\max} = y$ και συνεπώς $l\left(b_{\max}; y\right) = -\frac{N}{2}\log\left(2\pi\sigma^2\right)$.

Για κάθε άλλο μοντέλο με p παραμέτρους και εκτιμήτρια μέγιστης πιθανοφάνειας *b*, η στατιστική συνάρτηση απόκλισης είναι :

$$D = 2\left[l\left(b_{\max}; y\right) - l\left(b; y\right)\right] = \frac{1}{\sigma^2}\left(y - Xb\right)^T\left(y - Xb\right) = \frac{1}{\sigma^2}\left(y^T y - b^T X^T y\right)$$

Εάν το μοντέλο είναι σωστό τότε $D \sim x^2_{N-p}$, αλλιώς το D ακολουθεί μία μη-κεντρική χ τετράγωνο κατανομή. Επειδή το *D* δεν καθορίζεται επ’ ακριβώς όταν το σ^2 είναι άγνωστο ο έλεγχος υποθέσεων γίνεται συγκρίνοντας τα κατάλληλα πηλίκα των αποκλίσεων (Deviance) και χρησιμοποιώντας την *F* - κατανομή.

11.2 Απλή ANOVA (Ενός Παράγοντα)

Έστω ότι οι πειραματικές μονάδες κατανέμονται τυχαία σε *J* επίπεδα ενός παράγοντα, τότε αυτό ονομάζεται «τελείως τυχαίος πειραματικός σχεδιασμός». Τα δεδομένα τοποθετούνται όπως στον επόμενο πίνακα.

Πίνακας 14 : Δομή πίνακα ANOVA ενός παράγοντα

Factor Level		Responses	Totals
A_1	Y_{11}	$Y_{12}...Y_{1n_1}$	$Y_{1.}$
A_2	Y_{21}	$Y_{22}...Y_{2n_2}$	$Y_{2.}$
…			
A_j	Y_{j1}	$Y_{j2}...Y_{jn_j}$	$Y_{j.}$

Οι αποκρίσεις μπορεί να γραφτούν σαν διάνυσμα στήλη ως ε ξής :

$$y = \left[Y_{11}, ..., Y_{1n_1}, Y_{21}, ..., Y_{2n_2}, ..., Y_{jn_j} \right]^T$$

με μήκος $N = \sum_{j=1}^{J} n_{j.}$

Οι μεταβλητές $Y_{11}, Y_{12}, ..., Y_{1n_1}$ έχουν την ίδια αναμενόμενη τιμή οπότε είναι αντίγραφα. Γενικότερα, όλες οι αποκρίσεις με το ίδιο j είναι αντίγραφα. Για να απλοποιήσουμε τη συζήτηση, θα θεωρήσουμε την περίπτωση που όλα τα δείγματα έχουν το ίδιο μέγεθος, δηλαδή $n_j = K$ οπότε $N = JK$.

Εξετάζουμε (θεωρούμε) τρεις διαφορετικές διατυπώσεις του μοντέλου που αντιστοιχούν στην υπόθεση ότι οι μέσοι των αποκρίσεων διαφέρουν για διαφορετικά επίπεδα του παράγοντα.

<u>Η απλούστερη έκδοση του μοντέλου είναι</u> : $E\left(Y_{jk}\right) = \mu_j$ όπου $j = 1, ..., J$

Με όρους στοιχείων Y_i του διανύσματος y, το προηγούμενο μπορεί να γραφτεί ως $E\left(Y_i\right) = \sum_{j=1}^{J} x_{ij} \mu_j$ με i = 1, 2, …, N , όπου x_{ij} = 1 όταν η απόκριση Y_i αντιστοιχεί στο επίπεδο A_j ενώ x_{ij} =0 στις υπόλοιπες περιπτώσεις.

Συνεπώς

$$E\left(y\right) = X\beta \quad \text{με} \quad \beta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ . \\ \mu_j \end{pmatrix} \quad \text{και} \quad X = \begin{pmatrix} 1 & 0 & ... & 0 \\ 0 & 1 & ... & 0 \\ . & . & O & . \\ . & O & . & 0 \\ 0 & . & . & 1 \end{pmatrix}$$

όπου 0 και 1 είναι διανύσματα μήκους K από μηδενικά και μονάδες αντίστοιχα.

Το O δείχνει ότι όλοι οι υπόλοιποι όροι του πίνακα είναι όλα μηδενικά.

Ο $X^T X$ είναι ο JxJ διαγώνιος πίνακας.

$$X^T X = \begin{pmatrix} K & . & . & . & . \\ . & . & . & O & . \\ . & . & K & . & . \\ . & O & . & . & . \\ . & . & . & . & K \end{pmatrix} \quad \text{και} \quad X^T y = \begin{pmatrix} Y_{1.} \\ Y_{2.} \\ . \\ . \\ . \\ Y_{j.} \end{pmatrix}$$

Οπότε θα έχουμε

$$b = \frac{1}{K}\begin{pmatrix} Y_{1.} \\ Y_{2.} \\ . \\ . \\ . \\ Y_{j.} \end{pmatrix} = \begin{pmatrix} \overline{y}_1 \\ \overline{y}_2 \\ . \\ . \\ . \\ \overline{y}_j \end{pmatrix} \quad \text{και} \quad b^T X^T y = \frac{1}{K}\sum_{j=1}^{J} Y_{j.}^2$$

Οι προσαρμοσμένες τιμές είναι $\hat{y} = \left[\overline{y}_1, \overline{y}_1, ..., \overline{y}_1, \overline{y}_2, ...\overline{y}_j\right]^T$

Το μειονέκτημα με αυτή την απλή μορφή είναι ότι δεν μπορεί να επεκταθεί σε περισσότερους από ένα παράγοντες.

<u>Η δεύτερη έκδοση του μοντέλου</u> έχει την μορφή : $E\left(Y_{jk}\right) = \mu + \alpha_j$

όπου $j = 1, 2, \ldots, J$

Το μ παριστάνει τη μέση επίδραση (αποτέλεσμα) για όλα τα επίπεδα . Το α_j είναι μία επιπρόσθετη επίδραση λόγω του επιπέδου Aj. Υπάρχουν J+1 παράμετροι.

$$\beta = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ . \\ . \\ . \\ \alpha_j \end{pmatrix} \quad , \qquad X = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & & \\ \vdots & & & O & \\ \vdots & O & & & \\ 1 & & & & 1 \end{pmatrix}$$

όπου τα 1 και 0 είναι διανύσματα μήκους Κ.

Συνεπώς $$X^T y = \begin{pmatrix} Y_{..} \\ Y_{1.} \\ \vdots \\ Y_{J.} \end{pmatrix} \quad \text{και} \quad X^T X = \begin{pmatrix} N & & K & \dots & K \\ K & & K & & \\ \vdots & & & O & \\ \vdots & O & & & \\ K & & & & K \end{pmatrix}$$

Η πρώτη γραμμή του (J+1) x (J+1) πίνακα $X^T X$ είναι το άθροισμα των υπόλοιπων γραμμών, οπότε ο πίνακας $X^T X$ είναι μοναδιαίος και δεν υπάρχει μοναδική λύση για τις κανονικές εξισώσεις $X^T X b = X^T y$. Η γενική λύση μπορεί να γραφτεί ως :

$$b = \begin{pmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_J \end{pmatrix} = \frac{1}{K} \begin{pmatrix} 0 \\ Y_{1.} \\ \vdots \\ Y_{J.} \end{pmatrix} - \lambda \begin{pmatrix} -1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

Όπου το λ είναι μία αυθαίρετη σταθερά.

Παραδοσιακά απαιτούμε να ισχύει ο περιορισμός του αθροίσματος στο μηδέν $\sum_{j=1}^{J} \alpha_j = 0$, οπότε $\frac{1}{K}\sum_{j=1}^{J} Y_{j.} - J\lambda = 0$ και συνεπώς θα είναι $\lambda = \frac{1}{JK}\sum_{j=1}^{J} Y_{j.} = \frac{Y_{..}}{N}$

Αυτά δίνουν τη λύση $\hat{\mu} = \frac{Y_{..}}{N}$ και $\hat{\alpha}_j = \frac{Y_{j.}}{K} - \frac{Y_{..}}{N}$ για j = 1, …, J.

Οπότε έχουμε $b^T X^T y = \frac{Y_{..}^2}{N} + \sum_{j=1}^{J} Y_{j.}\left(\frac{Y_{j.}}{K} - \frac{Y_{..}}{N}\right) = \frac{1}{K}\sum_{j=1}^{J} Y_{j.}^2$

Το οποίο είναι ίδιο με τη πρώτη μορφή του μοντέλου και οι προσαρμοσμένες τιμές $\hat{y} = \left[\bar{y}_1, \bar{y}_1, ..., \bar{y}_1, \bar{y}_2, ... \bar{y}_j\right]^T$ είναι επίσης οι ίδιες. Οι περιορισμοί αθροίσματος στο μηδέν χρησιμοποιούνται στα περισσότερα στατιστικά πακέτα υπολογιστών.

<u>Η τρίτη έκδοση του μοντέλου είναι</u> : $E(Y_{jk}) = \mu + \alpha_j$ με τον περιορισμό $\alpha_1 = 0$. Εδώ το μ παριστάνει την επίδραση του πρώτου επιπέδου και το α_j μετράει τη διαφορά μεταξύ του πρώτου και του j –οστού επιπέδου του παράγοντα. Το $\alpha_1 = 0$ λέγεται γωνιακή παραμετροποίηση.

Για αυτή την έκδοση έχουμε J παραμέτρους.

$$\beta = \begin{pmatrix} \mu \\ \alpha_2 \\ \vdots \\ \alpha_j \end{pmatrix} \quad \text{και} \quad X = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & 1 & & \\ \vdots & . & . & O \\ \vdots & O & . & . \\ 1 & & & 1 \end{pmatrix}$$

Οπότε $$X^T y = \begin{pmatrix} Y_{..} \\ Y_{2.} \\ \vdots \\ Y_{j.} \end{pmatrix} \quad \text{και} \quad X^T X = \begin{pmatrix} N & K & \dots & K \\ K & K & \dots & O \\ \vdots & \dots & \cdots & \vdots \\ K & O & \dots & K \end{pmatrix}$$

Ο JxJ πίνακας $X^T X$ δεν είναι μοναδιαίος, συνεπώς υπάρχει μία μοναδική λύση : $$b = \frac{1}{K} \begin{pmatrix} Y_{1.} \\ Y_{2.} - Y_{1.} \\ \vdots \\ Y_{j.} - Y_{1.} \end{pmatrix}$$

Συνεπώς έχουμε $$b^T X^T y = \frac{1}{K}\left[Y_{..} Y_{1.} + \sum_{j=2}^{J} Y_{j.}\left(Y_{j.} - Y_{1.}\right) \right] = \frac{1}{K} \sum_{j=1}^{J} Y_{j.}^2$$

Οι προσαρμοσμένες τιμές είναι $\hat{y} = \left[\bar{y}_1, \bar{y}_1, \ldots, \bar{y}_1, \bar{y}_2, \ldots \bar{y}_j \right]^T$, δηλαδή είναι ίδιες με τις προηγούμενες προσαρμοσμένες τιμές.

Συμπέρασμα

Οι τρεις καθορισμοί των μοντέλων διαφέρουν αλλά οι τιμές του $b^T X^T y$ και του $D_1 = \frac{1}{\sigma^2}\left(y^T y - b^T X^T y\right) = \frac{1}{\sigma^2}\left[\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{jk}^2 - \frac{1}{K}\sum_{j=1}^{J} Y_{j.}^2\right]$ είναι οι ίδιες για όλες τις περιπτώσεις.

Έλεγχοι υποθέσεων

Οι προηγούμενες τρεις εκδόσεις του μοντέλου αντιστοιχούν όλες στην υπόθεση H_1 ότι οι μέσοι των αποκρίσεων για κάθε επίπεδο μπορεί να διαφέρουν για να συγκρίνουμε αυτό με τη μηδενική υπόθεση H_0 ότι όλοι οι μέσοι είναι ίσοι πρέπει να δημιουργήσουμε το αντίστοιχο μοντέλο.

Αντιστοιχούμε στην $\mathbf{H_0}$ το μοντέλο : $E\left(Y_{jk}\right) = \mu$ όπου $\beta = [\mu]$ και το X είναι ένα διάνυσμα από N μονάδες (1).

Τότε θα ισχύουν : $X^T X = N$, $X^T y = Y_{..}$ και συνεπώς :

$$b = \hat{\mu} = Y_{..}/N \quad \text{και} \quad b^T X^T y = Y_{..}^2/N \quad \text{και} \quad D_0 = \frac{1}{\sigma^2}\left[\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{jk}^2 - \frac{Y_{..}^2}{N}\right]$$

Για να ελέγξουμε $\boldsymbol{H_0}$ έναντι της $\boldsymbol{H_1}$, θεωρούμε ότι η $\boldsymbol{H_1}$ είναι σωστή οπότε $D_1 \sim x_{N-J}^2$.

Εάν επιπλέον $\boldsymbol{H_0}$ είναι σωστή τότε $D_0 \sim x_{N-1}^2$, αλλιώς το D_0 ακολουθεί μη-κεντρική χι-τετράγωνο κατανομή.

Συνεπώς αν H_0 είναι σωστή τότε :

$$D_0 - D_1 = \frac{1}{\sigma^2}\left[\frac{1}{K}\sum_{j=1}^{J} Y_{j.}^2 - \frac{1}{N}Y_{..}^2\right] \sim x_{J-1}^2 \quad \text{και συνεπώς} \quad f = \frac{\frac{D_0 - D_1}{J-1}}{\frac{D_1}{N-j}} \sim F_{J-1,N-J}$$

Εάν η $\boldsymbol{H_0}$ είναι σωστή, τότε η f είναι πιθανό να είναι μεγαλύτερη από αυτή τη τιμή που αναμένεται από την κατανομή $F_{J-1,N-J}$. Τα αποτελέσματα του ελέγχου τοποθετούνται σε ένα πίνακα ANOVA.

Θα δούμε τώρα τα αριθμητικά αποτελέσματα χρησιμοποιώντας το παράδειγμα στο οποίο έχουμε σαν δεδομένα τα βάρη των φυτών .

Παράδειγμα 12

Ένα πείραμα διεξάγεται για να συγκρίνει τα βάρη αποξηραμένων φυτών, που λαμβάνονται σε συνθήκες ελέγχου και δύο θεραπειών.

Συνεπώς η απόκριση, δηλ. το βάρος των φυτών, εξαρτάται από ένα παράγοντα με τρία επίπεδα τα οποία είναι : Επίπεδο ελέγχου, επίπεδο θεραπείας Α και επίπεδο θεραπείας Β.

Μας ενδιαφέρει να βρούμε εάν υπάρχουν διαφορές μεταξύ των μέσων των τριών ομάδων.

Πίνακας 15 : Βάρη αποξηραμένων φυτών, σε συνθήκες ελέγχου και δύο θεραπειών

Έλεγχος	4.17	5.58	5.18	6.11	4.50	4.61	5.17	4.53	5.33	5.14
Θεραπεία Α	4.81	4.17	4.41	3.59	5.87	3.83	6.03	4.89	4.32	4.69
Θεραπεία Β	6.31	5.12	5.54	5.50	5.37	5.29	4.92	6.15	5.80	5.26

Εδώ έχουμε N=30 , J=3 , K=10

$\frac{Y_{..}^2}{N} = 772.0599$ (εδώ N = 30) και $\frac{1}{K}\sum_{j=1}^{J} Y_{j.}^2 = 775.8262$ (εδώ K=10, J=3)

Οπότε : $D_0 - D_1 = 3.7663/\sigma^2$ και $\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{jk}^2 = 786.3138$ συνεπώς

$D_1 = 10.4921/\sigma^2$

Πίνακας 16 : Πίνακας ANOVA για τα βάρη φυτών

Πηγή Μεταβλητότητας	Βαθμοί Ελευθερίας	Άθροισμα Τετραγώνων	Μέσο Άθροισμα	f
Μέσος	1	772.0599		
Μεταξύ Θεραπειών	2	3.7663	1.883	4.85
Κατάλοιπα	27	10.4921	0.389	
Σύνολο	30	786.3138		

Εφ' όσον η τιμή f=4.85 είναι σημαντική στο επίπεδο 5% συγκρινόμενη με την τιμή $F_{2,\ 27}$ = 3, 3541 της κατανομής, βγάζουμε το συμπέρασμα ότι οι μέσοι των ομάδων διαφέρουν.

Για να ερευνήσουμε περαιτέρω το αποτέλεσμα αυτό, είναι βολικό να χρησιμοποιήσουμε τη πρώτη μορφή του μοντέλου $E\left(Y_{jk}\right) = \mu_j$. Οι εκτιμηθέντες μέσοι είναι : $b = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \hat{\mu}_3 \end{pmatrix} = \begin{pmatrix} 5.032 \\ 4.661 \\ 5.526 \end{pmatrix}$

Εάν χρησιμοποιήσουμε τα $\hat{\sigma}^2 = \dfrac{1}{N-J}\left(y - Xb\right)^T \left(y - Xb\right) = \dfrac{1}{N-J}\left(y^T y - b^T X^T y\right)$ λαμβάνουμε $\hat{\sigma}^2 = 10.4921 / 27 = 0.389$ (δηλαδή το μέσο τετράγωνο των καταλοίπων του πίνακα ANOVA). Ο πίνακα διακύμανσης – συνδιακύμανσης του b είναι $\hat{\sigma}^2\left(X^T X\right)^{-1}$ όπου : $X^T X = \begin{pmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{pmatrix}$

Συνεπώς το τυπικό σφάλμα για κάθε στοιχείο του b είναι $\sqrt{0.389/10} = 0.197$

Τώρα μπορεί κανείς να δει ότι η στατιστική σημαντικότητα οφείλεται στην επίδραση του μέσου της θεραπείας Β. Το $\hat{\mu}_3 = 5.526$ είναι μεγαλύτερο από την άποψη στατιστικής σημαντικότητας από τους άλλους δύο μέσους.

Λύση με χρήση R

Στην αρχή θα ορίσουμε επίπεδα παραγόντων L . Μετά θα δημιουργήσουμε το αντίστοιχο γραμμικό μοντέλο και θα το εκτιμήσουμε με την εντολή lm.
Τέλος θα έχουμε τα συνοπτικά αποτελέσματα με τις εντολές summary και anova.

```
> L<-gl(3, 10,30, labels = c("trA", "trB","Control"))
> L
```

```
 [1] trA   trA   trA   trA   tr A   trA   trA   trA   trA   trA   trB   trB   trB
trB   trB
[16] trB   trB   trB   trB   trB   Control Control Control Control Control Control
Control Control Control Control
Levels: trA trB Control

> Dedos<-c(4.81,4.17,4.41,3.59,5.87,3.83,6.03,4.89,4.32,4.69,
 6.31,5.12,5.54,5.50,5.37,5.29,4.92,6.15,5.80,5.26,
4.17,5.58,5.18,6.11,4.50,4.61,5.17,4.53,5.33,5.14)

> treat_lm<-lm(Dedos~L)
> anova(treat_lm)
Analysis of Variance Table

Response: Dedos
            Df    Sum Sq    Mean Sq    F value    Pr(>F)
L            2    3.7663     1.8832     4.8461   0.01591 *
Residuals   27   10.4921     0.3886
---
Signif. codes:  0 ‘***’ 0.001 ‘** ’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(treat_lm)

Call:
lm(formula = Dedos ~ L)

Residuals:
   Min       1Q      Median     3Q      Max
-1.0710  -0.4180   -0.0060   0.2627   1.3690

Coefficients:
             Estimate   Std. Error    t      value Pr(>|t|)
(Intercept)   4.6610    0.1971     23.644    < 2e-16 ***
LtrB          0.8650    0.2788      3.103    0.00446 **
LControl      0.3710    0.2788      1.331    0.19439
```

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6234 on 27 degrees of freedom

Multiple R-squared: 0.2641,   Adjusted R-squared: 0.2096

F-statistic: 4.846 on 2 and 27 DF, p-value: 0.01591

11.3 Πίνακας ANOVA δύο παραγόντων με Επαναλήψεις

Εστω ότι τα δεδομένα μας επηρεάζονται από δύο παράγοντες Α και Β. Ο παράγοντας Α με J επίπεδα και ο παράγοντας Β με K επίπεδα διασταυρώνονται για να προκύψουν οι JK υποομάδες που δημιουργούνται από όλους τους συνδυασμούς των επιπέδων Α και Β. Σε κάθε υποομάδα μπορεί να υπάρχουν πολλαπλές παααρτηρήσεις (ή *επαναλήψεις*). Υποθέτουμε ότι έχουμε σταθερό αριθμό επαναλήψεων L.

Πίνακας δεδομένων που επηρεάζονται από δύο παράγοντες και έχουν ίδιο αριθμό παρατηρήσεων σε κάθε υποομάδα.

Πίνακας 17 : Πίνακας ANOVA δύο παραγόντων

	Επίπεδα του παράγοντα Β			
Επίπεδα του παράγοντα Α	B_1	…	B_K	Σύνολο
A_1	$y_{111},\ldots,y_{11L}$	…	$y_{1K1},\ldots,y_{1KL}$	$y_{1..}$
	…	…		
A_J	$y_{J11},\ldots,y_{J1L}$	…	$y_{JK1},\ldots,y_{JKL}$	$y_{J..}$
Σύνολο	$y_{.1.}$	…	$y_{.K.}$	$y_{...}$

Οι κύριες υποθέσεις είναι :

H_O : δεν υπάρχουν αποτελέσματα αλληλεπίδρασης, δηλαδή τα αποτελέσματα των A και B είναι προσθετικά.

H_A : δεν υπάρχουν διαφορές στην απόκριση που σχετίζεται με τα διαφορετικά επίπεδα του παράγοντα A.

H_B : δεν υπάρχουν διαφορές στην απόκριση που σχετίζονται με τη τα διαφορετικά επίπεδα του παράγοντα B.

Εμείς χρειάζεται να εξετάσουμε το πλήρες μοντέλο και τρία μειωμένα (απλοποιημένα) μοντέλα τα οποία προκύπτουν παραλείποντας διάφορους όρους από το πλήρες μοντέλο.

Το πλήρες μοντέλο είναι :

$$E\left(Y_{jkl}\right)=\mu+\alpha_j+\beta_k+\left(\alpha\beta\right)_{jk}$$

Όπου οι όροι $\left(\alpha\beta\right)_{jk}$ αντιστοιχούν στα αποτελέσματα της αλληλεπίδρασης και τα α_j και β_k στην κύρια επίδραση των παραγόντων.

Το προσθετικό μοντέλο είναι :

$$E\left(Y_{jkl}\right)=\mu+\alpha_j+\beta_k$$

Το μοντέλο αυτό συγκρίνεται με το πλήρες μοντέλο για να ελέγξουμε την υπόθεση H_O.

Το μοντέλο το οποίο προκύπτει παραλείποντας επιδράσεις που οφείλονται στο B είναι :

$$E\left(Y_{jkl}\right)=\mu+\alpha_j$$

Το μοντέλο αυτό συγκρίνεται με το προσθετικό μοντέλο για να ελέγξουμε την υπόθεση H_B.

Το μοντέλο το οποίο προκύπτει παραλείποντας επιδράσεις που οφείλονται στο A είναι :

$$E\left(Y_{jkl}\right)=\mu+\beta_k$$

Το μοντέλο αυτό συγκρίνεται με το προσθετικό μοντέλο για να ελέγξουμε την υπόθεση H_A.

Επειδή οι επαναλήψεις μέσα στην ίδια υποομάδα έχουν τις ίδιες αναμενόμενες τιμές, τα μοντέλα μπορεί να έχουν μέχρι JK ανεξάρτητες αναμενόμενες τιμές τη στιγμή που το κορεσμένο μοντέλο έχει $1+J+K+JK=(J+1)(K+1)$ παραμέτρους.

Δημιουργείται το πρόβλημα να είναι ο πίνακας $X^T X$ μοναδιαίος. Για να το λύσουμε πρέπει μπορούμε να εισάγουμε επιπλέον περιορισμούς.

$$\alpha_1+\alpha_2+\alpha_3=0 \quad , \quad \beta_1+\beta_2=0$$

$$(\alpha\beta)_{11}+(\alpha\beta)_{12}=0 \ , \quad (\alpha\beta)_{21}+(\alpha\beta)_{22}=0, \quad (\alpha\beta)_{31}+(\alpha\beta)_{32}=0$$

$$(\alpha\beta)_{11}+(\alpha\beta)_{21}+(\alpha\beta)_{31}=0$$

(Οι υπόλοιπες συνθήκες $(\alpha\beta)_{12}+(\alpha\beta)_{22}+(\alpha\beta)32=0$ προκύπτει από τις προηγούμενες εξισώσεις των περιορισμών).

Αυτές οι εξισώσεις είναι οι συμβατικοί περιορισμοί για την ANOVA.

Εναλλακτικά μπορούμε να πάρουμε τους γωνιακούς περιορισμούς .

$$\alpha_1=\beta_1=(\alpha\beta)_{11}=(\alpha\beta)_{12}=(\alpha\beta)_{21}=(\alpha\beta)_{31}=0$$

Σε κάθε μία περίπτωση ο αριθμός των ανεξάρτητων (γραμμικά) παραμέτρων είναι : 1 για το μ, $J-1$ για τα α_j, $K-1$ για τα β_k, και $(J-1)(K-1)$ για τα $(\alpha\beta)_{jk}$..

Για τα προηγούμενα μοντέλα οι εκτιμήτριες b εξαρτώνται από την επιλογή των περιορισμών και τις ψευδομεταβλητές. Οι προσαρμοσμένες τιμές όμως $\hat{y}=Xb$ είναι ίδιες για όλες τις μορφές των μοντέλων και συνεπώς το ίδιο ισχύει και για τις τιμές $b^T X^T y$ και $\sigma^2 D=y^T y-b^T X^T y$.

<u>Παράδειγμα 13</u>

Υποθέτουμε τα πλασματικά δεδομένα του παρακάτω πίνακα , στα οποία ο παράγοντας Α (με $J=3$ επίπεδα) και ο παράγοντας Β (με $K=2$ επίπεδα). Σε κάθε υποομάδα υπάρχουν $L=2$ παρατηρήσεις ή *επαναλήψεις*.

Πίνακας 18 : Πλασματικά δεδομένα για άσκηση ANOVA

	Επίπεδα του παράγοντα B		
Επίπεδα του παράγοντα A	B1	B2	Σύνολο
A1	6.8, 6.6	5.3, 6.1	24.8
A2	7.5, 7.4	7.2, 6.5	28.6
A3	7.8, 9.1	8.8, 9.1	34.8
Σύνολο	45.2	43.0	88.2

Για τα δεδομένα του προηγούμενου πίνακα έχουμε τις τιμές $y^T y = 664.1$ και τα υπόλοιπα αποτελέσματα τα συνοψίζουμε στον παρακάτω πίνακα.
Θα χρησιμοποιήσουμε τους δείκτες S, I, A, B, M αντίστοιχα για το κορεσμένο μοντέλο, τα μοντέλα που αντιστοιχούν στις υποθέσεις H_O, H_B, H_A και για το μοντέλο που αποτελείται από το γενικό μέσο.

Πίνακας 19 : Σύνοψη αποτελεσμάτων για εικονικά δεδομένα

Όροι στο μοντέλο	Αριθμός παραμέτρων	Βαθμοί ελευθερίας	$b^T X^T y$	Απόκλιση (Deviance)
$\mu+\alpha_j+\beta_k+(\alpha\beta)_{jk}$	6	6	662.6200	$\sigma^2 D_s = 1.4800$
$\mu+\alpha_j+\beta_k$	4	8	661.4133	$\sigma^2 D_I = 2.6867$
$\mu+\alpha_j$	3	9	661.0100	$\sigma^2 D_B = 3.0900$
$\mu+\beta_k$	2	10	648.6733	$\sigma^2 D_A = 15.4267$
μ	1	11	648.2700	$\sigma^2 D_M = 15.8300$

Για να ελέγξουμε την H_O, παίρνουμε σαν δεδομένο ότι η το πλήρες μοντέλο είναι σωστό, έτσι ώστε $D_S \sim x_6^2$.

($D_S \sim x_6^2$ υπάρχουν 6 βαθμοί ελευθερίας διότι είναι N= 12 παρατηρήσεις και το μοντέλο έχει JK=6 ανεξάρτητες παραμέτρους).

Εάν η H_I είναι επίσης σωστή τότε $D_I \sim x_8^2$ οπότε θα είναι $D_I - D_S \sim x_2^2$ και

$$f = \frac{\frac{D_I - D_S}{2}}{\frac{D_S}{6}} \sim F_{2,6}$$

Η τιμή $f = \frac{\frac{2.6867 - 1.48}{2\sigma^2}}{\frac{1.48}{6\sigma^2}} = 2.45$ δεν είναι στατιστικά σημαντική συνεπώς τα δεδομένα δεν μας εφοδιάζουν με αποδείξεις εν αντίον της H_I .
Από τη στιγμή που δεν απορρίπτουμε την H_I συνεχίζουμε με τον έλεγχο H_A και H_B.

Για την υπόθεση H_B θεωρούμε τη διαφορά στην προσαρμογή μεταξύ των μοντέλων $E\left(Y_{jkl}\right) = \mu + \alpha_j$ και $E\left(Y_{jkl}\right) = \mu + \alpha_j + \beta_k$. Η διαφορά $D_B - D_I$ συγκρίνεται με το D_S χρησιμοποιώντας το f :

$$f = \frac{\frac{D_B - D_I}{1}}{\frac{D_S}{6}} = \frac{\frac{3.09 - 2.6867}{\sigma^2}}{\frac{1.48}{6\sigma^2}} = 1.63$$

το οποίο δεν είναι στατιστικά σημαντικό συγκρινόμενο με την $F_{1,6}$ κατανομή, πράγμα που δηλώνει ότι δεν υπάρχουν διαφορές που οφείλονται στα επίπεδα του παράγοντα Β.
Ο αντίστοιχος έλεγχος για την υπόθεση H_A δίνει $f = 25.82$ το οποίο συγκρινόμενο με τη κατανομή $F_{2,6}$ είναι στατιστικά σημαντικό. Συμπεραίνουμε λοιπόν ότι οι μέσοι των αποκρίσεων επηρεάζονται από διαφορές στα επίπεδα του παράγοντα Α.

Για τις προηγούμενες αναλύσεις έχουμε θέσει την προϋπόθεση ότι το κορεσμένο μοντέλο μας δίνει μία καλή περιγραφή για τα δεδομένα και συνεπώς το D_S έχει κεντρική χ-τετράγωνο κατανομή. Επομένως το D_S χρησιμοποιήθηκε στον παρανομαστή για όλα τα F-τεστ. Αυτό αντιστοιχεί στην συμβατική προσέγγιση ANOVA.
Για τα προηγούμενα δεδομένα όμως, θα μπορούσε να συμφωνηθεί ότι, αφού δεν απορρίπτουμε την H_O και το προσθετικό μοντέλο τα περιγράφει σχεδόν

τόσο καλά όσο και το κορεσμένο μοντέλο και επιπλέον είναι απλούστερο τότε να πρέπει να χρησιμοποιείται το $D_I \sim x_8^2$ στα F-τεστ για H_A , H_B. Η χρήση του D_I είναι πιο συνεπής με την προσέγγιση της προσαρμογής – μοντέλου και συμφωνεί με την ανάλυση δεδομένων.

Πίνακας 20 : Πίνακας ANOVA για τα εικονικά δεδομένα

Πηγή Μεταβλητότητας	Βαθμοί Ελευθερίας	Άθροισμα Τετραγώνων	Μέσο Άθροισμα	f
Μέσος	1	648.2700		
Επίπεδα του Α	2	12.7400	6.3700	25.82
Επίπεδα του Β	1	0.4033	0.4033	1.63
Αλληλεπιδράσεις	2	1.2067	0.6033	2.45
Κατάλοιπα	6	1.4800	0.2467	
Σύνολο	12	664.1		

Ένα άλλο χαρακτηριστικό της προηγούμενης ανάλυσης είναι ότι οι έλεγχοι υποθέσεων είναι κατά κάποια έννοια ανεξάρτητοι. Δηλαδή τα αποτελέσματα δεν επηρεάζονται από ποιους επιπλέον όρους έχουμε συμπεριλάβει στο μοντέλο, εκτός αυτών που έχουν σχέση με την ερωτούμενη υπόθεση .

<u>Παράδειγμα 14</u>

Η υπόθεση ότι δεν υπάρχουν διαφορές λόγω του παράγοντα Β. Δηλαδή η υπόθεση $H_\mathrm{B} : \beta_k = 0$ για όλα τα k, θα μπορούσε το ίδιο καλά να ελεγχθεί χρησιμοποιώντας :

είτε τα μοντέλα (1) : $E\left(Y_{jkl}\right) = \mu + \alpha_j + \beta_k$ και $E\left(Y_{jkl}\right) = \mu + \alpha_j$ και συνεπώς προκύπτει $\sigma^2 D_B - \sigma^2 D_I = 3.0900 - 2.6867 = 0.4033$

Είτε τα μοντέλα (2) : $E\left(Y_{jkl}\right) = \mu + \beta_k$ και $E(Y_{jkl}) = \mu$ και συνεπώς προκύπτει

$\sigma^2 D_M - \sigma^2 D_A = 15.8300 - 15.4267 = 0.4033$

Η αιτία για αυτό το χαρακτηριστικό είναι ότι τα δεδομένα είναι ισορροπημένα, που σημαίνει ότι, υπάρχει ίδιος αριθμός παρατηρήσεων σε κάθε υποομάδα. Σαν αποτέλεσμα, είναι δυνατό να καθορίσουμε τον πίνακα σχεδιασμού X με τέτοιο τρόπο ώστε να υπάρχουν ορθογώνιες συνιστώσες που να αντιστοιχούν σε καθένα από τα μοντέλα που μας ενδιαφέρουν και να είναι οι έλεγχοι υποθέσεων ανεξάρτητοι.

Στον παραπάνω πίνακα συνοψίζονται οι έλεγχοι υποθέσεων για το συμβατικό ANOVA. Ο πρώτος αριθμός στη στήλη «Άθροισμα τετραγώνων» είναι η τιμή $b^T X^T y$ που αντιστοιχεί στο μοντέλο $E(Y_{jkl}) = \mu$. Ο δεύτερος αριθμός είναι οι διαφορές των τιμών της $b^T X^T y$ για τα μοντέλα $E\left(Y_{jkl}\right) = \mu + \alpha_j + \beta_k$ και $E\left(Y_{jkl}\right) = \mu + \beta_k$ και ανάλογα το ίδιο ισχύει για τον τρίτο αριθμό. Ο τέταρτος αριθμός είναι οι διαφορές των τιμών $b^T X^T y$ για το πλήρες μοντέλο και το προσθετικό μοντέλο. Το «άθροισμα τετραγώνων των καταλοίπων» είναι $\sigma^2 D_S$, δηλαδή «Το συνολικό άθροισμα τετραγώνων» $y^T y$ μείον τη τιμή $b^T X^T y$ για το πλήρες μοντέλο. Οι βαθμοί ελευθερίας προκύπτουν ανάλογα.

11.4 Διασταυρωμένοι και Φωλιασμένοι Παράγοντες

Λέμε ότι οι παράγοντες A και B είναι διασταυρωμένοι όταν υπάρχει μία υποομάδα που αντιστοιχεί σε κάθε συνδυασμό επιπέδων A_j και B_k και όλες οι συγκρίσεις αντιπροσωπεύονται από τους όρους α_j και β_k και $(\alpha\beta)_{jk}$ στο πλήρες μοντέλο.

$E\left(Y_{jkl}\right) = \mu + \alpha_j + \beta_k + \left(\alpha\beta\right)_{jk}$, j = 1, …, J, k = 1, …, K

Στην προηγούμενη παράγραφο το παράδειγμα και η ανάλυση της διακύμανσης αναπτύχθηκαν με διασταυρωμένους πα ράγοντες.
Στις περιπτώσεις που **δεν υπάρχει** για κάθε συνδυασμό επιπέδων A_j και B_k μία αντίστοιχη υποομάδα λέμε ότι έχουμε φωλιασμένους παράγοντες.
Ο παρακάτω πίνακας μας δείχνει ένα παράδειγμα με φωλιασμένους παράγοντες.

Πίνακας 21 : Πείραμα δύο φωλιασμένων παραγόντων

	Μέθοδος Α1			Μέθοδος Α2	
Σχολεία	Β1	Β2	Β3	Β4	Β5
Αποκρίσεις	Y_{111}	Y_{121}	Y_{131}	Y_{241}	Y_{251}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	Y_{11n_1}	Y_{12n_2}	Y_{13n_3}	Y_{24n_4}	Y_{25n_5}

Ο παραπάνω πίνακας αντιπροσωπεύει ένα πείραμα για τη σύγκριση δύο εκπαιδευτικών μεθόδων (A_1 και A_2) εκ των οποίων η πρώτη μέθοδος δοκιμάζεται σε τρία σχολεία (B_1, B_2 και B_3) και η άλλη μέθοδος σε δύο άλλα σχολεία (B_4 και B_5). Θέλουμε να συγκρίνουμε τα αποτελέσματα των δύο εκπαιδευτικών μεθόδων και πιθανές διαφορές στις αποκρίσεις μεταξύ των σχολείων που χρησιμοποιούν την ίδια μέθοδο. Δεν είναι λογικό να κάνουμε συγκρίσεις μεταξύ σχολείων που χρησιμοποιούν διαφορετικές μεθόδους.

Το κορεσμένο μοντέλο εμπλέκει τις παραμέτρους μ, α_1, α_2, $(\alpha\beta)_{11}$, $(\alpha\beta)_{12}$, $(\alpha\beta)_{13}$, $(\alpha\beta)_{24}$ και $(\alpha\beta)_{25}$.

Οι συμβατικοί περιορισμοί του αθροίσματος στο μηδέν είναι :

$\alpha_1+\alpha_2=0$, $(\alpha\beta)_{11}+(\alpha\beta)_{12}+(\alpha\beta)_{13}=0$ και . $(\alpha\beta)_{24}+(\alpha\beta)_{25}=0$,

Οι γωνιακοί περιορισμοί είναι $\alpha_1=0$, $(\alpha\beta)_{11}=0$ και $(\alpha\beta)_{24}=0$.

Για να ελέγξουμε την υπόθεση ότι δεν υπάρχουν διαφορές μεταξύ των εκπαιδευτικών μεθόδων (αλλά επιτρέπονται διαφορές μεταξύ των σχολείων που χρησιμοποιούν την ίδια μέθοδο) το κορεσμένο μοντέλο συγκρίνεται με ένα μοντέλο που έχει παραμέτρους μ, β_1, β_2, β_3, β_4, και β_5 όπου τα β_ι είναι αποτελέσματα σχολείων και υπακούουν στους περιορισμούς $\beta_1+\beta_2+\beta_3=0$ και $\beta_4+\beta_5=0$ ή $\beta_1=0$ και $\beta_4=0$.

11.5 Περισσότερο πολύπλοκα μοντέλα

Μπορούμε να επεκτείνουμε τα μοντέλα της ανάλυσης διακύμανσης για πιο πολύπλοκες υποθέσεις και για να μελετηθούν σχεδιασμοί με περισσότερους από δύο παράγοντες. Οι παράγοντες μπορεί να είναι διασταυρωμένοι ή φωλιασμένοι ή ένα μείγμα από αυτές τις μορφές. Τα μοντέλα μπορεί να συμπεριλαμβάνουν από κύριες επιδράσεις μέχρι επιδράσεις υψηλότερου βαθμού όπως για παράδειγμα τον όρο $(\alpha\beta\gamma)_{jkl}$ καθώς και τον όρο $(\alpha\beta)_{jk}$. Αυτές οι επεκτάσεις δεν παράγουν θεμελιώδεις διαφορές από τα μοντέλα που έχουν ήδη συζητηθεί.

Σε όλα τα προηγούμενα παραδείγματα μελετήσαμε μόνο υποθέσεις όπου από το πλήρες μοντέλο παραλείπονται κάποιες παράμετροι για να προκύψει το περιορισμένο μοντέλο. Για παράδειγμα το πλήρες μοντέλο $E\left(Y_{jk}\right)=\mu+\alpha_j$ συγκρίνεται με το περιορισμένο μοντέλο $E\left(Y_{jk}\right)=\mu$.

Για να ελέγξουμε πιο πολύπλοκες υποθέσεις, για παράδειγμα ότι η κατάσταση ελέγχου και η θεραπεία στο πείραμα βαρών των φυτών έχουν την ίδια επίδραση, αλλά διαφέρουν από τη θεραπεία Β δηλαδή $\alpha_1=\alpha_2$ αλλά το α_3 δεν είναι κατ’ ανάγκη ίδιο, επιλέγουμε κατάλληλα παραμέτρους και ψευδομεταβλητές.

Οι πολλαπλοί έλεγχοι δεν είναι γενικά ανεξάρτητοι. Για να έχουμε οπωσδήποτε ανεξάρτητους πολλαπλούς ελέγχους πρέπει να υπάρχει ένας πίνακας σχεδιασμού με ορθογώνιες συνιστώσες οπότε το συνολικό άθροισμα τετραγώνων μπορεί να διαμεριστεί σε ασύνδετους όρους που αντιστοιχούν στις υποθέσεις. Συνήθως αυτό είναι δυνατό μόνο όταν οι υποθέσεις είναι ιδιαίτερα απλές και εάν υπάρχει ίδιος αριθμός παρατηρήσεων σε κάθε υποομάδα.

11.6 Επιλογή Περιορισμών και Ψευδομεταβλητών

Ο καθορισμός ενός μοντέλου ANOVA γίνεται με τέτοιο τρόπο ώστε να μπορούμε να ερμηνεύσουμε τις παραμέτρους που ορίζουμε σαν επιδράσεις που οφείλονται στα επίπεδα των παραγόντων και αλληλεπιδράσεις.

Ο αριθμός των παραμέτρων είναι συνήθως μεγαλύτερος από τον αριθμό των ανεξάρτητων κανονικών εξισώσεων. Για να αντιμετωπιστεί το πρόβλημα εισάγονται επιπλέον εξισώσεις στην μορφή του αθροίσματος στο μηδέν.
Εάν ο σχεδιασμός δεν είναι ισορροπημένος (δηλαδή δεν υπάρχει ίδιος αριθμός παρατηρήσεων σε κάθε υποομάδα), υπάρχει προβληματισμός για την πλέον κατάλληλη επιλογή των εξισώσεων.

Στο πλαίσιο των γενικευμένων γραμμικών μοντέλων, αυτό σημαίνει ότι οι εξισώσεις $X^T Xb = X^T y$ και $Wb = 0$ δεν είναι οι κανονικές εξισώσεις που προέρχονται από τη μέθοδο της μέγιστης πιθανοφάνειας ή των ελαχίστων τετραγώνων. Οπότε οι τυποποιημένες διαδικασίες δεν μπορεί να χρησιμοποιηθούν. Οι όροι του β επίσης δεν είναι σε γενικές γραμμές προσδιορίσιμοι, και μοναδικές αμερόληπτες σημειακές εκτιμήσεις και διαστήματα εμπιστοσύνης μπορεί να προκύψουν μόνο για κάποιους γραμμικούς συνδυασμούς των παραμέτρων, που καλούνται εκτιμήσιμες συναρτήσεις.
Εάν ο κύριος σκοπός της ανάλυσης δεδομένων είναι ο έλεγχος υποθέσεων, η χρήση του περιορισμού αθροίσματος-στο-μηδέν είναι τελείως κατάλληλη και βολική αρκεί να υποστηρίζεται από κατάλληλο λογισμικό. Τα περισσότερα πακέτα λογισμικού χρησιμοποιούν αυτή τη μέθοδο.

Εάν χρησιμοποιούνται οι γωνιακοί περιορισμοί, τα στοιχεία του β καθώς και οι αντίστοιχες στήλες του X τακτοποιούνται ως : $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$ και $X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}$

έτσι ώστε ο $X_1^T X_1$ δεν είναι μοναδιαίος και το β_2 τίθεται ίσο με μηδέν.

Συνεπώς : $E(y) = X\beta = X_1\beta_1$

Στην περίπτωση αυτή μπορεί να επιλυθούν οι κανονικές εξισώσεις $X_1^T X_1 b_1 = X_1^T y$ χρησιμοποιώντας την τυπική πολλαπλή παλινδρόμηση ή προγράμματα γενικευμένων γραμμικών μοντέλων και οι εκτιμητές να έχουν ποικίλες επιθυμητές ιδιότητες (πχ. b_1 είναι αμερόληπτη και έχει πίνακα διακύμανσης-συνδιακύμανσης $\sigma^2\left(X_1^T X_1\right)^{-1}$.

Οι παράμετροι που υπόκεινται σε γωνιακούς περιορισμούς δεν ερμηνεύονται τόσο καλά όσο στην περίπτωση των περιορισμών αθροίσματος - στο-μηδέν που η ερμηνεία είναι πιο ξεκάθαρη. Οι υπολογισμοί επίσης θα πρέπει να επαναληφθούν για κάθε νέο μοντέλο που π ρέπει να προσαρμοστεί.

Για απλά, καλά σχεδιασμένα πειράματα όπου είναι δυνατή η επιλογή των ψευδομεταβλητών με τέτοιο τρόπο ώστε ο πίνακας σχεδιασμού X να έχει ορθογώνιες συνιστώσες που αντιστοιχούν σε καθεμία από τις υποθέσεις που θέλουμε να ελέγξουμε υπάρχουν υπολογιστικά πλεονεκτήματα (πχ. οι εκτιμήσεις των παραμέτρων είναι ίδιες για όλα τα μοντέλα) και πλεονεκτήματα στη ερμηνεία (πχ. ανεξαρτησία των ελέγχων υποθέσεων).

Σε ένα μη ισορροπημένο πειραματικό σχεδιασμό όμως και για μοντέλα που εμπλέκουν πολύπλοκες υποθέσεις, είναι απίθανο να υπάρχει ορθογώνια μορφή.

<u>Συμπέρασμα</u>

Κάθε λύση b των κανονικών εξισώσεων $X^T X b = X^T y$ και $Wb = 0$ αντιστοιχεί σε ένα μοναδικό ελάχιστο του $(y - X\beta)^T (y - X\beta)$. Συνεπώς η στατιστική συνάρτηση $\sigma^2 D = y^T y - b^T X^T y$ <u>είναι ίδια, ανεξάρτητα</u> από το τρόπο με τον οποίο καθορίστηκαν τα μοντέλα.

Συνεπώς για οποιαδήποτε ομάδα από μοντέλα η επιλογή των περιορισμών και των ψευδομεταβλητών επ ηρεάζουν τις διαδικασίες υπολογισμού καθώς και τις εκτιμήσεις των παραμέτρων. Στα ίδια μοντέλα όμως, οι προηγούμενες επιλογές δεν επηρεάζουν τους ελέγχους υποθέσεων.

11.7 Ανάλυση Συνδιακύμανσης (ANCOVA)

Με την ανάλυση διακύμανσης μελετάμε μοντέλα στα οποία οι επεξηγηματικές μεταβλητές είναι μόνο ψευδομεταβλητές και αντιπροσωπεύουν επίπεδα παραγόντων.

Στην *Ανάλυση Συνδιακύμανσης* (ANCOVA) μελετάμε μικτά μοντέλα στα οποία κάποιες από τις επεξηγηματικές μεταβλητές είναι ψευδομεταβλητές που αντιπροσωπεύουν επίπεδα παραγόντων, και άλλες είναι συνεχείς και καλούνται «συμεταβλητές» (covariates).

Τις «συμεταβλητές» (covariates) θα καλούμε απλά μεταβλητές και έτσι θα ξεχωρίζουν από τις ψευδομεταβλητές.

Με την ανάλυση συνδιακύμανσης (ANCOVA) ενδιαφερόμαστε να συγκρίνουμε μέσους από υποομάδες που ορίζονται από επίπεδα παραγόντων για τους οποίους υπάρχει το ενδεχόμενο να επηρεάζονται από κάποιες μεταβλητές. Για το λόγο αυτό συγκρίνουμε τους μέσους αφού προηγουμένους λάβουμε υπόψη μας τις επιδράσεις των μεταβλητών.

<u>Παράδειγμα 15</u>

Ένα τυπικό παράδειγμα ορίζεται από τα δεδομένα του επόμενου πίνακα. Οι αποκρίσεις Y_{jk} είναι αποτελέσματα (scores) που έχουν επιτευχθεί και έχουν μετρηθεί σε τρία επίπεδα ενός παράγοντα που μετράει τρεις διαφορετικές εκπαιδευτικές μεθόδους, και οι μεταβλητές x_{jk} είναι μετρήσεις ικανότητας πριν από την έναρξη της εκπαίδευσης.
Θέλουμε να συγκρίνουμε εκπαιδευτικές μεθόδους λαμβάνοντας υπόψη διαφορές στην αρχική ικανότητα μεταξύ των τριών ομάδων των αντικειμένων που μελετάμε.

Πίνακας 22 : Πίνακας Δεδομένων [9]

Εκπαιδευτική Μέθοδος	M_1		M_2		M_3	
	y	x	y	x	y	x
	8	4	8	3	4	1
	4	1	6	5	7	2
	8	5	7	5	7	2
	3	1	9	4	7	3
	4	2	8	3	7	4
	3	1	5	1	5	1
	6	4	7	2	7	4

Για να ελέγξουμε την υπόθεση :

H_0 : Δεν υπάρχουν διαφορές στους μέσους των αποτελεσμάτων που επιτεύχθηκαν μεταξύ των τριών εκπαιδευτικών μεθόδων, μετά από ρύθμιση των αρχικών ικανοτήτων.

Συγκρίνουμε το πλήρες μοντέλο $E\left(Y_{jk}\right)=\mu_j+\gamma x_{jk}$ με το απλοποιημένο μοντέλο $E\left(Y_{jk}\right)=\mu+\gamma x_{jk}$ όπου j = 1, 2, 3 και k = 1, 2, …, 7.

Έστω $y_1=\begin{pmatrix} Y_{j1} \\ \vdots \\ Y_{j7} \end{pmatrix}$ και $x_j=\begin{pmatrix} x_{j1} \\ \vdots \\ x_{j7} \end{pmatrix}$

Οπότε με τη μορφή πίνακα το πλήρες μοντέλο είναι $E(y)=X\beta$ με $y=\begin{pmatrix} y_1 \\ y_2 \\ y3 \end{pmatrix}$

και $\beta=\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \gamma \end{pmatrix}$ και $X=\begin{pmatrix} 1 & 0 & 0 & x_1 \\ 0 & 1 & 0 & x_2 \\ 0 & 0 & 1 & x_3 \end{pmatrix}$ όπου το 0 και το 1 είναι διανύσματα διαστάσεως 7.

Υπολογίζουμε τα : $X^T X$, $X^T y$ και συνεπώς το b

Στη συνέχεια υπολογίζουμε τα $y^T y$ και $b^T X^T y$ συνεπώς για το πλήρες μοντέλο είναι : $\sigma^2 D_1 = y^T y - b^T X^T y$.

Για το απλοποιημένο μοντέλο έχουμε :

$\beta = \begin{pmatrix} \mu \\ \gamma \end{pmatrix}$ $X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{pmatrix}$ οπότε υπολογίζουμε τα : $X^T X$ και $X^T y$ και

$b^T X^T y$ και τέλος το $\sigma^2 D_0$

Εάν υποθέσουμε ότι το πλήρες μοντέλο είναι σωστό , τότε $D_1 \sim x_{17}^2$.

Εάν η μηδενική υπόθεση που αντιστοιχεί στο μοντέλο $E\left(Y_{jk}\right) = \mu + \gamma x_{jk}$ είναι σωστή τότε $D_0 \sim x_{19}^2$ οπότε $f = \dfrac{\dfrac{D_0 - D_1}{2\sigma^2}}{\dfrac{D_1}{17\sigma^2}} \sim F_{3,17}$.

<u>Θα αντιμετωπίσουμε το πρόβλημα με την R</u>

Αρχίζουμε με τον ορισμό των διανυσμάτων Y , x και L (για επίπεδα παραγόντων).

```
> y1<-c(8,4,8,3,4,3,6)
> y2<-c(8,6,7,9,8,5,7)
> y3<-c(4,7,7,7,7,5,7)
> x1<-c(4,1,5,1,2,1,4)
> x2<-c(3,5,5,4,3,1,2)
> x3<-c(1,2,2,3,4,1,4)
> Y<-c(y1,y2,y3)
> x<-c(x1,x2,x3)
> L<-gl(3,7,21,labels=c("M1","M2","M3"))
```

Ορίζουμε το μοντέλο που θα μελετηθεί και το προσαρμόζουμε γραμμικά.

```
> model_lm<-lm(Y~L+x)
> model_lm
```

Call:
lm(formula = Y ~ L + x)

Coefficients:

(Intercept)	LM2	LM3	x
3.1109	1.4356	1.2557	0.7902

> summary(model_lm)

Call:
lm(formula = Y ~ L + x)

Residuals:

Min	1Q	Median	3Q	Max
-2.4975	-0.6913	-0.1568	1.0529	1.7283

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.1109	0.6508	4.780	0.000174 ***
LM2	1.4356	0.6438	2.230	0.039536 *
LM3	1.2557	0.6307	1.991	0.062818 .
x	0.7902	0.1844	4.284	0.000502 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.179 on 17 degrees of freedom
Multiple R-squared: 0.6263, Adjusted R-squared: 0.5604
F-statistic: 9.498 on 3 and 17 DF, p-value: 0.0006505

> aov(model_lm)
Call:
aov(formula = model_lm)

Terms:

	L	x	Residuals
Sum of Squares	14.09524	25.51249	23.63037
Deg. of Freedom	2	1	17

Residual standard error: 1.178992
Estimated effects may be unbalanced

Υπολογίζουμε την τιμή του ποσοστημόριου που αντιστοιχεί στην στατιστική συνάρτηση $F_{3,17}$ για επίπεδο 0.05.

```
> f_stat<-qf(0.95,3,17)
> f_stat
[1] 3.196777
```

Για τα προηγούμενα δεδομένα έχουμε $f = 9.498$ που δείχνει ότι υπάρχει στατιστικά σημαντική διαφορά στα αποτελέσματα (scores) που έχουν επιτευχθεί με τις εκπαιδευτικές μεθόδους μετά από τη προσαρμογή για τις αρχικές διαφορές στην ικανότητα.

12 Δίτιμες Μεταβλητές και Λογιστική Παλινδρόμηση

12.1 Κατανομές Πιθανοτήτων

Θεωρούμε ένα πείραμα στο οποίο η απόκριση Y μπορεί να πάρει μόνο μία από δύο δυνατές τιμές. Για παράδειγμα, οι αποκρίσεις μπορεί να είναι «παρών» ή «απών» . Οι όροι που χρησιμοποιούνται για αυτές τις δύο κατηγορίες είναι «επιτυχία» και «αποτυχία». Από σύμβαση η «επιτυχία» συμβολίζεται με 1 και η «αποτυχία» με 0.

Η Y ονομάζεται δίτιμη μεταβλητή. Ορίζουμε την δίτιμη τυχαία μεταβλητή ως εξής :

$$Y = \begin{cases} 1 & \text{εάν το αποτέλεσμα είναι επιτυχία} \\ 0 & \text{εάν το αποτέλεσμα είναι αποτυχία} \end{cases}$$

Εάν π είναι η πιθανότητα επιτυχίας τότε : $P(Y=1) = \pi$ και $P(Y=0) = 1-\pi$.

Η συνάρτηση πιθανότητας της τυχαία μεταβλητή Y που ορίσαμε παραπάνω είναι : $f(y;\pi) = P(Y=y) = \pi^{y}(1-\pi)^{1-y}$.

Το π είναι η παράμετρος της συνάρτησης.

Θεωρούμε n τυχαίες δίτιμες μεταβλητές, $Y_1, Y_2, \ldots, Y_n$ οι οποίες είναι ανεξάρτητες με $P(Y_i=1) = \pi_i$.

Η από κοινού πιθανότητα είναι :

$$f(\pi;y) = \prod_{j=1}^{n} \pi_j^{y_j}(1-\pi_j)^{1-y_j} = \exp\left[\sum_{j=1}^{n} y_j \log\left(\frac{\pi_j}{1-\pi_j}\right) + \sum_{j=1}^{n} \log(1-\pi_j)\right] \qquad \textbf{(A)}$$

Όπου $\pi = [\pi_1, \ldots, \pi_n]^T$ και $y = [y_1 \ldots y_n]^T$, και ανήκει στην εκθετική οικογένεια κατανομών.

Στην περίπτωση που τα π_j είναι όλα ίσα μεταξύ τους μπορούμε να ορίσουμε τη συνάρτηση $Y=\sum_{j=1}^{n} Y_j$ στην οποία το Y παριστάνει τον αριθμό των «επιτυχιών» σε n ανεξάρτητες προσπάθειες και λέμε ότι το Y ακολουθεί τη διωνυμική κατανομή.

Εάν η τυχαία μεταβλητή Y έχει τη *διωνυμική κατανομή* τότε συμβολίζουμε με :

$Y \sim b(n,\pi)$ και έχει συνάρτηση πιθανότητας :

$$P(Y=y)=\binom{n}{y}\pi^{y}(1-\pi)^{n-y} \quad \text{όπου} \quad y = 0, 1, \ldots, n$$

Υποθέτουμε, τέλος, τη γενική περίπτωση των N ανεξάρτητων τυχαίων μεταβλητών $Y_1, Y_2, \ldots, Y_N$ που ακολουθούν δυωνυμική.

Τότε εάν $Y_i \sim b(n_i,\pi_i)$ τότε η λογαριθμική συνάρτηση πιθανοφάνειας είναι :

$$l(\pi_1,\ldots,\pi_N;y_1,\ldots,y_N)=\sum_{i=1}^{N}\left\{y_i \log\left(\frac{\pi_i}{1-\pi_i}\right)+n_i\log(1-\pi_i)+\log\binom{n_i}{y_i}\right\} \quad \textbf{(B)}$$

Η παραπάνω κατανομή δεν αντιστοιχεί απ' ευθείας με την εξίσωση κανονικής μορφής για την εκθετική οικογένεια, διότι τα n_i μπορεί να μην είναι όλα τα ίδια.

Παρ' όλα αυτά, εάν η από κοινού κατανομή των Y_i γράφεται με όρους δυαδικών μεταβλητών Z_i τότε από τις εξισώσεις **(A)** και **(B)** συνάγεται ότι ανήκει στην εκθετική οικογένεια.

12.2 Γενικευμένα Γραμμικά Μοντέλα με δίτιμες μεταβλητές

Εστω N ανεξάρτητες μεταβλητές Y_i για τις οποίες ισχύει : $Y_i \sim b(n_i, \pi_i)$.

Με $n_1, n_2, \ldots, n_N$ συμβολίζουμε τον αριθμό δοκιμών που υπάρχουν σε κάθε ομάδα που περιγράφεται από μία μεταβλητή Y_i.

Τα μοντέλα με δίτιμες μεταβλητές μπορεί να χρησιμοποιηθούν για να περιγράψουμε την αναλογία των «επιτυχιών» $P_i = {y_i}/{n_i}$ σε κάθε ομάδα με όρους «επιπέδων ενός παράγοντα» και άλλες επεξηγηματικές μεταβλητές που χαρακτηρίζουν την ομάδα. Στην περίπτωση αυτή θα ισχύει $0 \le \frac{y_i}{n_i} \le 1$.

Θέλουμε να μελετήσουμε τη σχέση μεταξύ της απόκρισης π που είναι η πιθανότητα «επιτυχιών» και των επεξηγηματικών μεταβλητών.

Μοντελοποιούμε τις πιθανότητες π_i ως : $g(\pi_i) = \eta_i = x_i^T \beta$

Όπου το x_i είναι το διάνυσμα επεξηγηματικών μεταβλητών (αποτελείται από ψευδομεταβλητές για τα επίπεδα παραγόντων καθώς και μεταβλητές για τιμές που έχουν παρατηρηθεί), το β είναι το διάνυσμα παραμέτρων και g είναι η συνάρτηση σύνδεσης.

Το μοντέλο έχει ένα μεγάλο μειονέκτημα. Τα π_i είναι πιθανότητες οπότε θα πρέπει να υπάρξουν περιορισμοί πάνω στις τιμές του β για να έχουμε τιμές του π_i στο διάστημα [0, 1].

Η πιο απλή περίπτωση είναι το γραμμικό μοντέλο : $\pi = x^T \beta$

Αυτό χρησιμοποιείται σε κάποιες πρακτικές εφαρμογές αλλά έχει το μειονέκτημα ότι αν και το π είναι μία πιθανότητα οι προσαρμοσμένες τιμές $x^T b$ μπορεί να είναι έξω από το διάστημα [0 , 1].

Αυτό το γραμμικό μοντέλο είναι ισοδύναμο με την χρήση της ταυτοτικής συνάρτησης ως συνάρτηση σύνδεσης g, με περιορισμούς πάνω στα x, β_1 και β_2 ώστε να περιορίζεται το π στο διάστημα [0, 1].

Για να εξουδετερώσουμε το μειονέκτημα που προκύπτει από τη μοντελοποίηση του π επιλέγουμε κατάλληλα τη συνάρτηση σύνδεσης ώστε να δίνει τιμές για εκτιμήσεις του π μέσα στο διάστημα [0, 1].

Για τη δυωνυμική κατανομή, επιλέγουμε τις παρακάτω <u>συναρτήσεις σύνδεσης</u> :

Λογιστική (Logit) : $g(\pi_i) = \ln \frac{\pi_i}{1-\pi_i}$

Οπότε θα είναι $\pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$

Probit : $g(\pi_i) = \Phi^{-1}(\pi_i)$ όπου με Φ συμβολίζεται η κανονική κατανομή.

Οπότε θα είναι $\pi_i = \Phi(\eta_i)$

Ακραίων τιμών (C-log-log) : $g(\pi_i) = \ln(-\ln(1-\pi_i))$

Οπότε θα είναι $\pi_i = 1 - \exp(-e^{\eta_i})$

<u>Παρατήρηση :</u>

Το μοντέλο $g(\pi_i) = \ln \frac{\pi_i}{1-\pi_i}$ χρησιμοποιείται ευρέως για ανάλυση πολυμεταβλητών δεδομένων που εμπλέκουν δυαδικές αποκρίσεις. Μας εφοδιάζει με ισχυρές τεχνικές ανάλογες με την πολλαπλή παλινδρόμηση και την ANOVA για συνεχείς αποκρίσεις.

12.3 Μοντέλα Απόκρισης σε Δόση (φαρμάκων)

Ιστορικά, μία από τις πρώτες χρήσεις των παλινδρομικών μοντέλων για δυωνυμικά δεδομένα ήταν για τη μελέτη των αποτελεσμάτων βιο-αναλύσεων.

Οι αποκρίσεις ήταν οι αναλογίες ή τα ποσοστά των «επιτυχιών». Για παράδειγμα η αναλογία των πειραματόζωων που σκοτώθηκαν από διάφορα επίπεδα δόσεων ενός τοξικού συστατικού. Τα δεδομένα αυτής της μορφής ονομάζονται ποσοστιαίες αποκρίσεις.

Ο σκοπός είναι να περιγράψουμε την πιθανότητα των «επιτυχιών» π σαν μία συνάρτηση της δόσης x. Για παράδειγμα $g(\pi) = \beta_1 + \beta_2 x$.

Αυτές οι επιπλέον συνθήκες δηλώνουν ότι, οι τυπικές μέθοδοι για την εκτίμηση των β_1 και β_2 για Γενικευμένα Γραμμικά μοντέλα δεν μπορούν να εφαρμοστούν απ' ευθείας. Αυτό το μοντέλο δεν χρησιμοποιείται ευρέως στην πράξη.

Ένα από τα αρχικά μοντέλα που χρησιμοποιήθηκαν για δεδομένα βιο-αναλύσεων είναι το λεγόμενο ***probit*** μοντέλο $g(\pi_i) = \Phi^{-1}(\pi_i)$.

όπου το Φ παριστάνει την αθροιστική συνάρτηση πιθανότητας της τυπικής κανονικής κατανομής N(0, 1).

Συνεπώς : $\Phi^{-1}(\pi) = \beta_1 + \beta_2 x$ και η συνάρτηση σύνδεσης g είναι η αντίστροφη συνάρτηση της αθροιστικής κανονικής κατανομής Φ^{-1}.

Τα probit μοντέλα χρησιμοποιούνται σε διάφορες περιοχές της βιολογίας και των κοινωνικών επιστημών στις οποίες υπάρχει φυσική ερμηνεία του μοντέλου. Για παράδειγμα η τιμή $\boldsymbol{x=\mu}$ ονομάζεται η «μέση θανατηφόρα δόση» LD(50) (Lethal Dose 50%) διότι αντιστοιχεί στη δόση που απαιτείται για να σκοτώσει τα μισά πειραματόζωα.

Ένα άλλο μοντέλο που χρησιμοποιείται είναι το λογιστικό η logit μοντέλο. Το μοντέλο αυτό δίνει αριθμητικά αποτελέσματα τα οποία μοιάζουν πολύ με αυτά του **probit** μοντέλου αλλά είναι υπολογιστικά ευκολότερα.

Επειδή $g(\pi_i) = \ln \frac{\pi_i}{1-\pi_i}$ έχουμε $\log\left(\frac{\pi}{1-\pi}\right) = \beta_1 + \beta_2 x$

Η $\log\left(\frac{\pi}{1-\pi}\right)$ έχει φυσική ερμηνεία ότι είναι ο λογάριθμος των odds (που εκφράζουν το λόγο επιτυχιών προς αποτυχίες).

Από το μοντέλο αυτό προκύπτει ότι : $\pi_i = \dfrac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)}$

Το λογιστικό μοντέλο χρησιμοποιείται ευρέως για δυωνυμικά δεδομένα και έχουν υλοποιηθεί σε πολλά στατιστικά προγράμματα. Τα σχήματα των διαγραμμάτων είναι όμοια με αυτά των probit μοντέλων εκτός από τις ουρές των κατανομών.

Για τα δεδομένα απόκρισης Δόσεων χρησιμοποιούνται επίσης αρκετά άλλα μοντέλα. Για παράδειγμα χρησιμοποιούμε τη συνάρτηση ακραίων τιμών ή συμπληρωματική *log log συνάρτηση* **(c- log log)**.

Δηλαδή : $\log\left[-\log(1-\pi)\right] = \beta_1 + \beta_2 x$ Οπότε $\pi = 1 - \exp\left[-\exp(\beta_1 + \beta_2 x)\right]$.

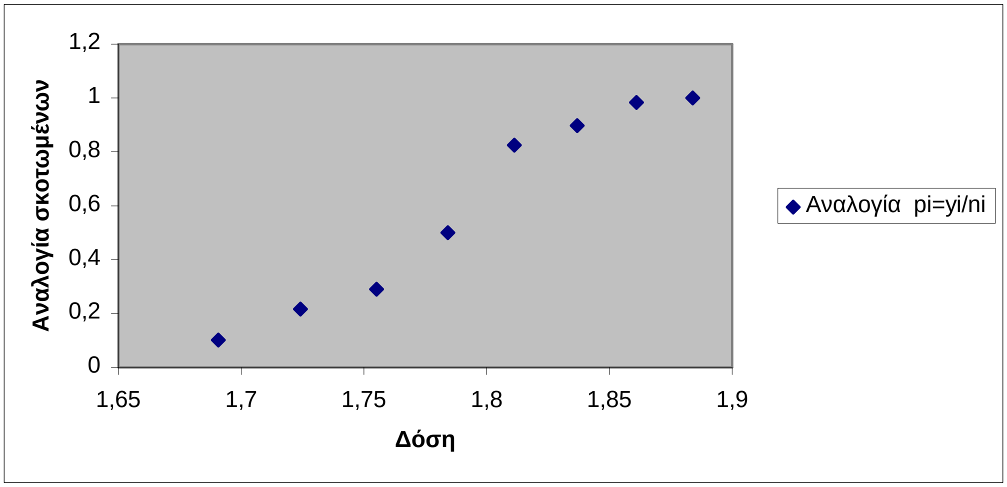
Αυτό το μοντέλο είναι παρόμοιο με το λογιστικό και το probit μοντέλο για τιμές του π κοντά στο 0, 5 αλλά διαφέρουν για τιμές του π κοντά στο 0 ή το 1.

12.4 Παράδειγμα μοντέλων απόκρισης σε Δόση

Ο παρακάτω πίνακας δείχνει τον αριθμό εντόμων που πέθαναν μετά από έκθεση πέντε ωρών σε αεριώδη ανθρακικό ντιζουλφίδιο σε ποικίλες συγκεντρώσεις[10].

Πίνακας 23 : Δεδομένα θνησιμότητας των Σκαθαριών

Δόση x_i **($\log_{10}CS_2 mgl^{-1}$)**	**Αριθμός των** Εντόμων n_i	**Αριθμός που** Σκοτώθηκαν y_i	**Αναλογία** $p_i = y_i / n_i$
1, 6907	59	06	0, 102
1, 7242	60	13	0, 217
1, 7552	62	18	0, 290
1, 7842	56	28	0, 500
1, 8113	63	52	0, 825
1, 8369	59	53	0, 898
1, 8610	62	61	0, 984
1, 8839	60	60	1, 000



Γράφημα 7 : Αναλογιών $p_i=y_i/n_i$ ως προς τις δόσεις x_i

Αρχίζουμε με την προσαρμογή του λογιστικού μοντέλου

$$\pi_i = \frac{\exp(\beta_1+\beta_2 x_i)}{1+\exp(\beta_1+\beta_2 x_i)} \qquad \text{οπότε}$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1+\beta_2 x_i \qquad \text{και} \qquad \log(1-\pi_i) = -\log\left[1+\exp(\beta_1+\beta_2 x_i)\right]$$

Από τη σχέση **(Β)** προκύπτει η ακόλουθη λογαριθμική συνάρτηση πιθανοφάνειας.

$$l = \sum_{i=1}^{N}\left[y_i(\beta_1+\beta_2 x_i) - n_i\log\left[1+\exp(\beta_1+\beta_2 x_i)\right] + \log\binom{n_i}{y_i}\right]$$

Και τα score ως προς β_1 και β_2 θα είναι :

$$U_1 = \frac{\partial l}{\partial \beta_1} = \sum\left\{y_i - n_i\left[\frac{\exp(\beta_1+\beta_2 x_i)}{1+\exp(\beta_1+\beta_2 x_i)}\right]\right\} = \sum(y_i - n_i\pi_i) \qquad \text{και}$$

$$U_2 = \frac{\partial l}{\partial \beta_2} = \sum\left\{y_i x_i - n_i x_i\left[\frac{\exp(\beta_1+\beta_2 x_i)}{1+\exp(\beta_1+\beta_2 x_i)}\right]\right\} = \sum x_i(y_i - n_i\pi_i)$$

Ο πίνακας πληροφορίας είναι :

$$J = \begin{pmatrix} \sum n_i\pi_i(1-\pi_i) & \sum n_i x_i\pi_i(1-\pi_i) \\ \sum n_i x_i\pi_i(1-\pi_i) & \sum n_i x_i^2\pi_i(1-\pi_i) \end{pmatrix}$$

Οι **εκτιμητές μέγιστης πιθανοφάνειας** προκύπτουν από τη λύση της επαναληπτικής εξίσωσης :

$$j^{(m-1)}b^{(m)} = j^{(m-1)}b^{(m-1)} + U^{(m-1)}$$

όπου ο δείκτης (m) δείχνει τη (m)–οστή προσέγγιση και ***b*** είναι το διάνυσμα των εκτιμήσεων.

Ξεκινώντας από τις τιμές $b_1^{(0)} = 0$ και $b_2^{(0)} = 0$ δείχνονται διαδοχικές προσεγγίσεις στον παρακάτω πίνακα μαζί με τις προσαρμοσμένες τιμές $\hat{y}_i = n_i \hat{\pi}_i$.

Ο εκτιμηθείς πίνακας διακύμανσης – συνδιακύμανσης για το b είναι $\left[J(b)\right]^{-1}$

Η στατιστική συνάρτηση πηλίκου λογαριθμικής πιθανοφάνειας είναι :

$$D = \sum_{i=1}^{N}\left[y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n - y_i}{n - \hat{y}_i}\right)\right]$$

όπου το $\hat{y}_i$ δηλώνει τις προσαρμοσμένες τιμές.

Εφαρμόζοντας τις τιμές του **πίνακα 23** στους παραπάνω τύπους έχουμε τα αποτελέσματα .

Πίνακας 24: Προσαρμόζοντας το λογιστικό μοντέλο στα δεδομένα θανάτου σκαθαριών

	Αρχική Εκτίμηση	Πρώτη Προσέγγι-ση	Δεύτερη Προσέγγι-ση	Τέταρτη Προσέγγι-ση	Δέκατη Προσέγγι-ση
b_1	0	-37, 849	-53, 851	-60, 700	-60, 717
b_2	0	21, 334	30, 382	34, 261	34, 270

$$\left[J(b)\right]^{-1} = \begin{pmatrix} 26,802 & 15,061 \\ 15,061 & 8,469 \end{pmatrix}, \quad D = 11,\ 23$$

Οι εκτιμήσεις είναι b_1= -60,72 και b_2 =34,27 και τα τυπικά τους σφάλματα είναι αντίστοιχα $\sqrt{26,802} = 5,18$ και $\sqrt{8,469} = 2,91$.

Εάν το λογιστικό μοντέλο μας δίνει μία καλή περιγραφή των δεδομένων τότε η στατιστική συνάρτηση πηλίκου λογαριθμικής πιθανοφάνειας **D** έχει κατά προσέγγιση μία x_6^2 κατανομή, διότι υπάρχουν N=8 παρατηρήσεις και p=2 παραμέτρους.

Επειδή το πάνω 5% σημείο της x_6^2 κατανομής είναι 12, 59 συμπεραίνουμε ότι το μοντέλο δεν προσαρμόζεται ιδιαίτερα καλά στα δεδομένα.

<u>Θα αντιμετωπίσουμε το πρόβλημα με την R</u>

```
> x <-c(1.6907, 1.7242, 1.7552, 1.7842, 1.8113, 1.8369, 1.8610, 1.8839)
> n<-c(59, 60, 62, 56, 63, 59, 62, 60)
> y<-c(06, 13, 18, 28, 52, 53, 61, 60)
```

Δημιουργούμε το πλαίσιο δεδομένων dedos για να το χρησιμοποιήσουμε στην δημιουργία του γενικευμένου γραμμικού μοντέλου dosi.

```
> dedos<-data.frame(x=x,y=y,n=n)
> dedos$Y<-cbind(dedos$y,dedos$n-dedos$y)
> dosi<-glm(Y~x,family=binomial,data=dedos)
> dosi
```

Call: glm(formula = Y ~ x, family = binomial, data = dedos)

Coefficients:

(Intercept)	x
-60.72	**34.27**

Degrees of Freedom: 7 Total (i.e. Null); 6 Residual

Null Deviance: 284.2

Residual Deviance: 11.23 AIC: 41.43

<u>Παρατήρηση</u>

Επειδή χρησιμοποιούμε κατανομή από την δυωνυμική οικογένεια, η προκαθορισμένη συνάρτηση σύνδεσης είναι η λογιστική (logit).

Χρησιμοποιούμε για τα δεδομένα θνησιμότητας σκαθαριών διαφορετικά μοντέλα και παρακάτω δείχνουμε τα αποτελέσματα χρήσης των διαφορετικώς μοντέλων.

<u>Χρήση της συνάρτησης σύνδεσης probit</u>

> dosi<-glm(Y~x,family=binomial(link=probit),data=dedos)
> dosi

Call: glm(formula = Y ~ x, family = binomial(link = probit), data = dedos)

Coefficients:

(Intercept) x
-34.94 19.73

Degrees of Freedom: 7 Total (i.e. Null); 6 Residual
Null Deviance: 284.2
Residual Deviance: 10.12 AIC: 40.32

<u>Χρήση της συνάρτησης σύνδεσης ακραίων τιμών (Extreme value)</u>

> dosi<-glm(Y~x,family=binomial(link=cloglog),data=dedos)
> dosi

Call: glm(formula = Y ~ x, family = binomial(link = cloglog), data = dedos)

Coefficients:

(Intercept) x
-39.57 22.04

Degrees of Freedom: 7 Total (i.e. Null); 6 Residual
Null Deviance: 284.2
Residual Deviance: 3.446 AIC: 33.64

Πίνακας 25: Διαφορετικά μοντέλα για δεδομένα θνησιμότητας σκαθαριών

Απόκλιση (Deviance)	Λογιστικό Μοντέλο	Probit Μοντέλο	Μοντέλο Ακραίων τιμών (Extreme value)
D	**11, 23**	**10, 12**	**3, 45**

Από τις τιμές του D που υπολογίστηκε για τα διάφορα μοντέλα, συμπεραίνεται ότι το μοντέλο Ακραίων τιμών (Extreme value) δίνει την καλύτερη περιγραφή των δεδομένων.

12.5 Εκτίμηση Μέγιστης Πιθανοφάνειας και η Στατιστική Συνάρτηση Απόκλισης

Λογιστική παλινδρόμηση

Για το γενικό λογιστικό μοντέλο, οι εκτιμήσεις μέγιστης πιθανοφάνει ας των παραμέτρων β και συνεπώς των πιθανοτήτων $\pi_i = g^{-1}\left(x_i^T\beta\right)$ προκύπτουν μεγιστοποιώντας τη συνάρτηση λογαριθμικής πιθανοφάνειας $l(\pi;y)=\sum_{i=1}^{N}\left[y_i\log\pi_i+(n_i-y_i)\log(1-\pi_i)+\log\binom{n_i}{y_i}\right]$.

Οι εκτιμήσεις μεγίστης πιθανοφάνειας είναι δυνατές ακόμα και αν n_i =1 ή /και y_i=0 σε αντίθεση με κάποιες μεθόδους ελαχίστων τετραγώνων.

Για να μετρήσουμε την «καλή προσαρμογή» του μοντέλου, χρησιμοποιούμε τη στατιστική συνάρτηση απόκλισης :

$$D=2\left[l(\hat{\pi}_{\max};y)-l(\hat{\pi};y)\right]$$

όπου $\hat{\pi}_{\max}$ είναι το διάνυσμα των εκτιμήσεων μέγιστης πιθανοφάνειας που αντιστοιχεί στο πλήρες μοντέλο και $\hat{\pi}$ είναι το διάνυσμα των εκτιμήσεων μέγιστης πιθανοφάνειας του μοντέλου που μας ενδιαφέρει.

Χωρίς βλάβη της γενικότητας, για το πλήρες μοντέλο παίρνουμε τα π_i ως τις παραμέτρους που θέλουμε να εκτιμήσουμε. Τότε :

$$\frac{\partial l}{\partial \pi_i} = \frac{y_i}{\pi_i} - \frac{n_i - y_i}{1 - \pi_i}$$

Οπότε το i-οστό στοιχείο του $\hat{\pi}_{\max}$ είναι η λύση της εξίσωσης $\partial l / \partial \pi_i = 0$ που είναι y_i / n_i (δηλαδή το ποσοστό των επιτυχιών στη i-οστή υποομάδα).

οπότε :

$$l(\hat{\pi}_{\max}; y) = \sum_{i=1}^{N} \left[y_i \log\left(\frac{y_i}{n_i}\right) + (n_i - y_i) \log\left(1 - \frac{y_i}{n_i}\right) + \log\binom{n_i}{y_i} \right]$$

και συνεπώς

$$D = 2\sum_{i=1}^{N} \left[y_i \log\left(\frac{y_i}{n_i \hat{\pi}_i}\right) + (n_i - y_i) \log\left(\frac{n_i - y_i}{n_i - n_i \hat{\pi}_i}\right) \right]$$

Οπότε το D θα έχει τη μορφή : $D = 2\sum \mathrm{o} \log \frac{\mathrm{o}}{e}$

Όπου το ***o*** παριστάνει τις παρατηρηθείσες συχνότητες $\boldsymbol{y_i}$ και $\boldsymbol{(n_i - y_i)}$ και το ***e*** παριστάνει τις αντίστοιχες εκτιμηθείσες αναμενόμενες συχνότητες ή προσαρμοσμένες τιμές $n_i \hat{\pi}_i$ και $(n_i - n_i \hat{\pi}_i)$.

Επειδή το D αντίθετα με τα κανονικά δεδομένα απόκρισης δεν περιέχει το σ^2, μπορεί να χρησιμοποιηθεί απ' ευθείας για τον έλεγχο της «καλής προσαρμογής» καθώς επίσης και για έλεγχο υποθέσεων χρησιμοποιώντας την προσέγγιση $D \sim x^2_{N-p}$ όπου p είναι ο αριθμός των παραμέτρων β που εκτιμήθηκαν.

12.6 Εναλλακτικά κριτήρια «Καλής Προσαρμογής»

Μία εναλλακτική μέθοδος από την εκτίμηση της μέγιστης πιθανοφάνειας είναι να γίνει εκτίμηση με το σταθμισμένο άθροισμα τετραγώνων :

$$S_w = \sum_{i=1}^{N} \frac{\left(y_i - n_i \pi_i\right)^2}{n_i \pi_i \left(1-\pi_i\right)} \quad \text{με} \quad E\left(Y_i\right) = n_i \pi_i \quad \text{και} \quad \operatorname{var}\left(Y_i\right) = n_i \pi_i \left(1-\pi_i\right)$$

Αυτό είναι ισοδύναμο με το να ελαχιστοποιήσουμε την στατιστική συνάρτηση χ-τετράγωνο του Pearson :

$$X^2 = \sum \frac{\left(o-e\right)^2}{e}$$

όπου *o* είναι οι παρατηρηθείσες συχνότητες και *e* παριστάνει τις αναμενόμενες συχνότητες που προέρχονται από το μοντέλο και τις αθροίσεις πάνω από όλα τα $2 \times N$ κελιά του πίνακα.

Αυτό εξηγείται από το παρακάτω :

$$X^2 = \sum_{i=1}^{N} \frac{\left(y_i - n_i \pi_i\right)^2}{n_i \pi_\iota} + \sum_{i=1}^{N} \frac{\left[\left(n_i - y_i\right) - n_i\left(1-\pi_i\right)\right]^2}{n_i\left(1-\pi_i\right)} = \sum_{i=1}^{N} \frac{\left(y_i - n_i \pi_i\right)^2}{n_i \pi_i \left(1-\pi_i\right)} \left(1 - \pi_i + \pi_i\right) = S_w$$

Όταν το X^2 παίρνει τιμές στις εκτιμηθείσες αναμενόμενες συχνότητες , τότε η στατιστική συνάρτηση είναι : $X^2 = \sum_{i=1}^{N} \frac{\left(y_i - n_i \hat{\pi}_\iota\right)^2}{n_i \hat{\pi}_\iota \left(1 - n\, \hat{\pi}_i\right)}$

η οποία είναι ασυμπτωτικά ισοδύναμη με τη στατιστική συνάρτηση απόκλισης : $D = 2\sum_{i=1}^{N}\left[y_i \log\left(\frac{y_i}{n_i \hat{\pi}_i}\right) + \left(n_i - y_i\right)\log\left(\frac{n_i - y_i}{n_i - n_i \hat{\pi}_i}\right)\right]$

Η απόδειξη χρησιμοποιεί το ανάπτυγμα της σειράς Taylor του *slog(s/t)* στο s=t, (δηλαδή : $s \log\frac{s}{t} = \left(s-t\right) + \frac{1}{2}\frac{\left(s-t\right)^2}{t} + \cdots$)

Έχουμε

$$D = 2\sum_{i=1}^{N}\left\{\left(y_i - n\ \hat{\pi}_i\right) + \frac{1}{2}\frac{\left(y_i - n_i\hat{\pi}_i\right)^2}{n_i\hat{\pi}_i} + \left[\left(n_i - y_i\right) - \left(n_i - n_i\hat{\pi}_i\right)\right] + \frac{1}{2}\frac{\left[\left(n_i - y_i\right) - \left(n_i - n_i\hat{\pi}_i\right)\right]^2}{n_i - n_i\hat{\pi}_i} + \cdots\right.$$

$$\cong \sum_{i=1}^{N}\frac{\left(y_i - n_i\hat{\pi}_i\right)^2}{n_i\hat{\pi}_i\left(1-\hat{\pi}_i\right)} = X^2$$

Κάτω από την υπόθεση ότι το μοντέλο είναι σωστό, η κατανομή μεγάλων δειγμάτων του D είναι $D \sim x^2_{N-p}$, και ως εκ τούτου προσεγγιστικά $X^2 \sim x^2_{N-p}$.

Η επιλογή μεταξύ των D, X^2 εξαρτάται από την επάρκεια της προσέγγισης στην x^2_{N-p} κατανομή. Υπάρχουν αποδείξεις για να προτείνουμε ότι η X^2 είναι συχνά καλύτερη από το D διότι το D επηρεάζεται ανάρμοστα από πολύ μικρές συχνότητες. Όλες οι προσεγγίσεις είναι πολύ φτωχές εάν οι αναμενόμενες συχνότητες είναι μικρότερες από 1.

12.7 Παρατηρήσεις

Στα δίτιμα μοντέλα έχουμε όμοιο τρόπο αντιμετώπισης θεμάτων με την πολλαπλή παλινδρόμηση. Οι έλεγχοι για την εισαγωγή ή εξαγωγή κάποιων όρων δεν είναι ανεξάρτητοι και συνεπώς πρέπει να είμαστε πολύ προσεχτικοί με τους όρους που εισάγονται στο μοντέλο σε κάθε φάση.

Εάν υπάρχουν πολλές επεξηγηματικές μεταβλητές, χρησιμοποιούνται βηματικές μέθοδοι επιλογής για να προσδιοριστεί το καλύτερο υποσύνολο μεταβλητών που θα εισαχθεί στο μοντέλο.

Για την εκτίμηση της επάρκειας του προτεινόμενου μοντέλου μπορεί να χρησιμοποιηθούν τα τυποποιημένα κατάλοιπα. Ένας απλός ορισμός για τα τυποποιημένα κατάλοιπα είναι $r_i = \frac{p_i - \hat{\pi}_i}{\sqrt{\left[\frac{\hat{\pi}_i(1-\hat{\pi}_i)}{n_i}\right]}}$ όπου το $p_i = y_i / n_i$ είναι το ποσοστό που παρατηρήθηκε και $\hat{\pi}_i$ είναι το ποσοστό που εκτιμήθηκε με βάση το μοντέλο.

Τα r_i έχουν προσεγγιστικά μέσο μηδέν (0) και τυπική απόκλιση ένα (1). Είναι οι προσημασμένες τετραγωνικές ρίζες που συνεισφέρουν στη στατιστική συνάρτηση X^2.

Όταν παριστάνονται γραφικά ως προς τους παράγοντες των επιπέδων και τις μεταβλητές δεν πρέπει να εμφανίζουν συστηματικά σχήματα. Η κατανομή πιθανοτήτων τους όμως, μπορεί να απέχει πολύ από την κανονική.

Πιο πολύπλοκα κατάλοιπα , τα οποία είναι κοντά στην κανονική κατανομή, περιγράφονται από τους Cox και Snell (1968).

Πιο πρόσφατα οι Pierce και Schafer (1986) έδειξαν ότι οι προσημασμένες τετραγωνικές ρίζες που είναι τμήμα της στατιστικής συνάρτησης D $d_i = \pm\sqrt{2}\left[y_i \log\left(\frac{y_i}{n\hat{\pi}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n\hat{\pi}_i}\right)\right]$ είναι προσεγγιστικά κανονικά κατανεμημένες εάν το μοντέλο είναι σωστό και μας εφοδιάζουν με κατάλληλα κατάλοιπα για διαγνωστικούς σκοπούς.

13 Πίνακες Συνάφειας και Λογαριθμο-Γραμμικά μοντέλα (Log – Linear)

13.1 Εισαγωγή

Θα μελετήσουμε δεδομένα για τα οποία η απόκριση και οι επεξηγηματικές μεταβλητές είναι όλες κατηγορικές. Τα δεδομένα δεν είναι στη μορφή ποσοστών όπως στα δίτιμα μοντέλα, αλλά είναι δεδομένα μετρήσεων.

Οι παρατηρήσεις αποτελούνται από πλήθος ή συχνότητες στα κελιά ενός πίνακα συνάφειας που έχει διαμορφωθεί με μία διασταυρούμενη κατάταξη κάποιων μεταβλητών.

Θα ασχοληθούμε στη συνέχεια με Γενικευμένα γραμμικά μοντέλα στα οποία ο πίνακας συνάφειας έχει απλή μορφή. Πιο πολύπλοκες καταστάσεις όπως για παράδειγμα όταν οι αποκρίσεις μπορεί να θεωρηθούν σαν επαναλαμβανόμενες μετρήσεις στα ίδια άτομα δεν θα αντιμετωπιστούν.

Για πολύπλοκες περιπτώσεις πινάκων συνάφειας σας παραπέμπουμε στα βιβλία Bishop, Fienberg and Holland (1975), Everitt (1977), Finberg (1980) και Freeman (1987).

<u>Παράδειγμα 16</u> : Τυχαιοποιημένη ελεγχόμενη μελέτη εμβολίων κατά της γρίπης.

Για μία μελέτη ενός ανασχεδιασμένου εμβολίου κατά της γρίπης, οι ασθενείς τοποθετήθηκαν τυχαία σε δύο ομάδες.

Στη πρώτη ομάδα δόθηκε το εμβόλιο ενώ στη δεύτερη ένα πλασματικό φάρμακο (placebo).

Οι αποκρίσεις είναι επίπεδα από «αντισώματα αναστολής αιμοσυγκολλητίνης» (HIA) που βρέθηκαν στο αίμα έξη εβδομάδες μετά τον εμβολιασμό.

Για τη μελέτη αυτή υπάρχει μόνο μία επεξηγηματική μεταβλητή που είναι η θεραπεία (εμβόλιο ή εικονικό φάρμακο) που είναι ονομαστική και μία μεταβλητή απόκρισης (επίπεδο ΗΙΑ) που είναι αριθμός διάταξης. Θα θεωρούνται εδώ και οι δύο μεταβλητές σαν κατηγορικές.

Οι συχνότητες στα κελιά των γραμμών θα πρέπει να αθροίζονται στον αριθμό των ατόμων της κάθε ομάδας (35 και 38 αντίστοιχα). Θέλουμε να μάθουμε εάν το σχέδιο των αποκρίσεων είναι το ίδιο και στις δύο ομάδες.

Πίνακας 26 : Δοκιμές εμβολίων κατά της γρίπης (Εικονικά δεδομένα)

	Απόκριση			
	Μικρή	Μέτρια	Μεγάλη	Σύνολο
πλασματικό φάρμακο	25	9	4	38
εμβόλιο	5	18	12	35

13.2 Κατανομές Πιθανοτήτων

Θεωρούμε την μεταβλητή Y η οποία ακολουθεί την κατανομή Poisson Η συνάρτηση πιθανότητας είναι:

$$P(Y=y)=\frac{\lambda^{y}e^{-\lambda}}{y!} \qquad \text{με} \qquad y=0,1,\ldots$$

Γνωρίζουμε από τη θεωρία ότι : $E(Y)=Var(Y)=\lambda$

Θα χρησιμοποιήσουμε την κατανομή Poisson σε δεδομένα που περιγράφονται με πίνακες συνάφειας.

Οι πίνακες συχνοτήτων που θα χρησιμοποιηθούν θα έχουν τους εξής συμβολισμούς. Κατ' αρχήν, για πίνακες δύο διαστάσεων με J κατηγορίες για τη μεταβλητή A και K κατηγορίες για τη μεταβλητή B, θα συμβολίζουμε με Y_{jk} τη συχνότητα για το (j, k) στοιχείο του πίνακα. Με $Y_{j.}$ και $Y_{.k}$ συμβολίζουμε τα σύνολα των γραμμών και των στηλών αντίστοιχα και τέλος με n το γενικό σύνολο. Οι συχνότητες των κελιών Y_{jk} είναι οι εξαρτημένες μεταβλητές που επιθυμούμε για το μοντέλο.

Πίνακας 27 : πίνακας δύο διαστάσεων

	B_1	B_2	$\cdots$	B_k	Σύνολο
A_1	Y_{11}	Y_{12}	$\cdots$	Y_{1K}	$Y_{1.}$
A_2	Y_{21}				$Y_{2.}$
$\vdots$	$\vdots$				
A_j	Y_{j1}			Y_{JK}	$Y_{J.}$
Σύνολο	$Y_{.1}$	$Y_{.2}$		$Y_{.K}$	$n = Y_{..}$

Γενικεύοντας τώρα σε περισσότερες διαστάσεις, για ένα $J \times K \times \ldots \times L$ πίνακα οι συχνότητες συμβολίζονται με $Y_{jk\ldots l}$ και τοποθετούνται σε ένα διάνυσμα y με στοιχεία τα οποία έχουν δείχτες $i = 1, 2, \ldots, N$.

Ξεκινάμε με μοντέλα πιθανοτήτων για πίνακες δύο διαστάσεων. Ο πιο απλός λαμβάνεται υποθέτοντας ότι οι τυχαίες μεταβλητές Y_{jk} είναι ανεξάρτητες και καθεμία έχει την κατανομή *Poisson* με παραμέτρους $\lambda_{jk} > 0$. Η από κοινού κατανομή είναι απλά το γινόμενο των επί μέρους κατανομών *Poisson*.

$$f(y;\lambda) = \prod_{j=1}^{J} \prod_{k=1}^{K} \frac{\lambda_{jk}^{y_{jk}} e^{-\lambda_{jk}}}{y_{jk}!}$$

Στις πιο συνηθισμένες περιπτώσεις έχουμε περιορισμούς πάνω στα Y_{jk}, για παράδειγμα, η συνολική συχνότητα n είναι σταθερή από τον σχεδιασμό της μελέτης. Στην περίπτωση αυτή, λόγω της προσθετικής ιδιότητας των ανεξάρτητων τυχαίων μεταβλητών με την κατανομή Poisson, το άθροισμά τους δοθέντος του n έχει επίσης την Poisson κατανομή με παράμετρο $\lambda_{..} = \sum\sum \lambda_{jk}$. Συνεπώς η υπό συνθήκη κατανομή των Y_{jk} δοθέντος του n είναι :

$$f(y|n) = \prod_{j=1}^{J}\prod_{k=1}^{K} \frac{\lambda_{jk}^{y_{jk}} e^{-\lambda_{jk}}}{y_{jk}!} \Big/ \frac{\lambda_{..}^{n} e^{-\lambda_{..}}}{n!} = n! \prod_{j=1}^{J}\prod_{k=1}^{K} \frac{\theta_{jk}^{y_{jk}}}{y_{jk}!} \quad \text{όπου} \quad \theta_{jk} = \frac{\lambda_{jk}}{\lambda_{..}}$$

διότι είναι $\lambda_{..}^{n} = \prod\prod \lambda_{..}^{y_{jk}}$ και $e^{-\lambda_{..}} = \prod\prod e^{-\lambda_{jk}}$.

Η προηγούμενη κατανομή ονομάζεται πολυωνυμική κατανομή. Από τον ορισμό προκύπτει ότι $0 \le \theta_{jk} \le 1$ και $\sum_{j}\sum_{k} \theta_{jk} = 1$, και όντως οι όροι θ_{jk} αντιπροσωπεύουν τις πιθανότητες των κελιών.

Μία άλλη μορφή περιορισμών εφαρμόζεται στην περίπτωση που δεν έχουμε σταθερό το συνολικό άθροισμα, αλλά τα αθροίσματα των γραμμών ή των στηλών. Στην περίπτωση αυτή η κατανομή πιθανοτήτων για κάθε γραμμή (αντίστοιχα στήλη) είναι πολυωνυμική. Παραδείγματος χάριν, εάν το άθροισμα $y_{j.}$ της j γραμμής είναι σταθερό τότε η κατανομή για τη j γραμμή είναι :

$$f\left(y_{j1},\ldots,y_{jK} \mid y_{j.}\right) = y_{j.}! \prod_{k=1}^{K} \theta_{jk}^{y_{jk}} \Big/ y_{jk}! \quad \text{όπου} \quad \sum_{k} \theta_{kj} = 1$$

Οι γραμμές (ή στήλες) θεωρούνται ανεξάρτητες συνεπώς (για το παραπάνω παράδειγμα), εάν όλα τα αθροίσματα των γραμμών είναι σταθερά, τότε η από κοινού συνάρτηση κατανομής για όλα τα Y_{jk} είναι :

$$f\left(y \mid y_{j.}, j = 1,2,\ldots,J\right) = \prod_{j=1}^{J} y_{j.}! \prod_{k=1}^{K} \theta_{jk}^{y_{jk}} \Big/ y_{jk}! \quad \text{όπου} \quad \sum_{k} \theta_{kj} = 1 \ \text{ για κάθε } j \text{ γραμμή.}$$

Αυτή είναι το γινόμενο της πολυωνυμικής κατανομής και είναι ένα βολικό μοντέλο για τα τυχαιοποιημένα δεδομένα δοκιμών που αφορά την τυχαιοποιημένη ελεγχόμενη μελέτη εμβολίων κατά της γρίπης.

Για πίνακες συνάφειας με περισσότερες από δύο διαστάσεις, εάν συμβολίσουμε με Y_i, $i=1, \ldots N$ την συχνότητα, τότε οι δύο κύριες κατανομές πιθανοτήτων είναι οι ακόλουθες.

Κατανομή Poisson

$$f(y;\lambda)=\prod_{i=1}^{N}\lambda_i^{y_i}e^{-\lambda_i}\Big/y_i!$$

Χωρίς περιορισμούς για τις συχνότητες y_i ή για τις παραμέτρους λ_i.

Πολυωνυμική Κατανομή

$$f(y;\theta\,|\,n)=n!\prod_{i=1}^{N}\theta_i^{y_i}\Big/y_i! \quad \text{όπου είναι} \quad n=\sum_{i=1}^{N}y_i \quad \text{και} \quad \sum_{i=1}^{N}\theta_i=1.$$

13.3 Λογαριθμογραμμικά (Log –Linear) μοντέλα

Για την κατανομή Poisson με τύπο $f(y;\lambda)=\prod_{i=1}^{N}\lambda_i^{y_i}e^{-\lambda_i}\Big/y_i!$ και συχνότητες κελιών $Y_1,\ldots,Y_N$ και παραμέτρους $\lambda_1,\ldots,\lambda_N$ οι αναμενόμενες συχνότητες των κελιών δίνονται από το $E(Y_i)=\lambda_i$ για $i=1, \ldots, N$.

Για κάθε πολυωνυμική κατανομή με συχνότητες κελιών $Y_1,\ldots,Y_N$ και πιθανότητες κελιών $\theta_1,\ldots,\theta_N$ με $n=\sum_{i=1}^{N}Y_i$ και $\sum_{i=1}^{N}\theta_i=1$ μπορεί να αποδειχθεί[11] ότι :

$$E(Y_i)=n\theta_i \text{ για } i=1, \ldots, N.$$

Για τους πινάκες συνάφειας δύο διαστάσεων μπορούν να διατυπωθούν όλες οι συνηθισμένες υποθέσεις για τις αναμενόμενες συχνότητες των κελιών.

Για παράδειγμα, εάν η υπόθεση είναι : οι μεταβλητές γραμμών και στηλών είναι ανεξάρτητες τότε $\theta_{jk} = \theta_{j.}\theta_{.k}$ όπου τα $\theta_{j.}$ και $\theta_{.k}$ παριστάνουν τις περιθώριες πιθανότητες των μεταβλητών γραμμών και στηλών και $\sum_j \theta_{j.} = 1$ και $\sum_k \theta_{.k} = 1$. Για αυτό το λόγο, οι αναμενόμενες συχνότητες για την πολυωνυμική κατανομή είναι $E\left(Y_{jk}\right) = n\theta_{j.}\theta_{.k}$.

Με όμοιο τρόπο αντιμετωπίζονται πίνακες υψηλότερης διάστασης. Οι πιο κοινές υποθέσεις μπορεί να εκφραστούν σαν πολλαπλασιαστικά μοντέλα στα οποία οι αναμενόμενες συχνότητες των κελιών δίνονται σαν γινόμενα των περιθωρίων πιθανοτήτων και των σταθερών περιθωρίων συνολικών συχνοτήτων.

Η μορφή της συνάρτησης πιθανότητας και της αναμενόμενης τιμής των προηγούμενων κατανομών μας οδηγεί να λάβουμε σαν συνάρτηση σύνδεσης για γενικευμένα γραμμικά μοντέλα, τον λογάριθμο.

$$\eta_i = \log E\left(Y_i\right) = x_i^T \beta \quad \text{για} \quad i = 1, \ldots, N$$

Απ' όπου προκύπτει το όνομα log – linear μοντέλο.

<u>Παράδειγματα :</u>

1. Η σχέση $E\left(Y_{jk}\right) = n\theta_{j.}\theta_{.k}$ μπορεί να εκφραστεί ως :

$$\eta_{jk} = \log E\left(Y_{jk}\right) = \mu + \alpha_j + \beta_k$$

2. Σε αναλογία με την ανάλυση διακύμανσης, το αντίστοιχο πλήρες μοντέλο $E\left(Y_{jk}\right) = n\theta_{jk}$ μπορεί να γραφτεί σαν :

$$\eta_{jk} = \log E\left(Y_{jk}\right) = \mu + \alpha_j + \beta_k + \left(\alpha\beta\right)_{jk}$$

Έτσι ώστε η υπόθεση ανεξαρτησίας $\theta_{jk} = \theta_{j.}\theta_{.k}$ για όλα τα j και k είναι ισοδύναμη με την υπόθεση «μη αλληλεπίδρασης» $(\alpha\beta)_{jk} = 0$ τα j και k.

Οι όροι υψηλότερης τάξης στα log-linear μοντέλα , ορίζονται συνήθως σαν αποκλίσεις από τους όρους χαμηλότερης τάξης.

3. Στην εξίσωση $\eta_{jk} = \log E(Y_{jk}) = \mu + \alpha_j + \beta_k$ το α_j παριστάνει την διαφορά της επίδρασης της γραμμής j ως προς τη μέση επίδραση μ.

<u>Παρατήρηση</u>

Τα μοντέλα αυτά είναι επίσης ιεραρχικά με την έννοια του ότι δεν συμπεριλαμβάνονται όροι υψηλότερης τάξης σε ένα μοντέλο εάν δεν συμπεριληφθούν πρώτα όλοι οι σχετικοί χαμηλότερης τάξης όροι.

Όμοια με τα ANOVA μοντέλα, τα log-linear μοντέλα έχουν πάρα πολλές παραμέτρους και συνεπώς απαιτούνται περιορισμοί είτε γωνιακοί είτε αθροίσματος στο μηδέν. Σε γενικές γραμμές, για τις κύριες επιδράσεις α_j όπου $j=1, \ldots, J$ υπάρχουν $(J-1)$ ανεξάρτητες παράμετροι. Για τις αλληλεπιδράσεις πρώτης τάξεως $(\alpha\beta)_{jk}$ όπου $j=1, \ldots, J$ *και* $k=1, \ldots, K$ υπάρχουν $(J-1)(K-1)$ ανεξάρτητες παράμετροι κλπ.

Στην ανάλυση δεδομένων ενός πίνακα συνάφειας, τα περισσότερα ερωτήματα έχουν σχεδόν πάντα να κάνουν με τις σχέσεις μεταξύ των μεταβλητών. Συνεπώς στα log-linear μοντέλα αυτό που κατ’ αρχήν ενδιαφέρει είναι οι αλληλεπιδράσεις που αναπτύσσονται μεταξύ δύο ή περισσότερων μεταβλητών. Επειδή τα μοντέλα είναι ιεραρχικά , στα μοντέλα που χρησιμοποιούνται για έλεγχο υποθέσεων αναπτύσσονται όροι αλληλεπίδρασης και έχουν όλους τους αντίστοιχους όρους των κύριων επιδράσεων.

Στις εκφράσεις για αναμενόμενες συχνότητες κελιών για πολυωνυμικές κάποιοι όροι είναι καθορισμένες σταθερές. Για παράδειγμα καθορισμένες σταθερές είναι οι όροι **n** στην $E(Y_i)=n\theta_i$ και το $y_{j.l}$ στην $E(Y_{jkl})=y_{j.l}\theta_{jkl}$. Αυτό σημαίνει ότι οι αντίστοιχες παράμετροι πρέπει πάντα να συμπεριλαμβάνονται στα log-linear μοντέλα .

Για παράδειγμα, το πλήρες μοντέλο που αντιστοιχεί στο $E(Y_{jkl})=y_{j.l}\theta_{jkl}$ είναι $\eta_{jkl}=\mu+\alpha_j+\beta_k+\gamma_l+(\alpha\beta)_{jk}+(\alpha\gamma)_{jl}+(\beta\gamma)_{kl}+(\alpha\beta\gamma)_{jkl}$ στην οποία η έκφραση $\mu+\alpha_j+\gamma_l+(\alpha\gamma)_{jl}$ αντιστοιχεί στο σταθερό περιθώρια άθροισμα $y_{j.l}$ και το υπόλοιπο μέρος $\beta_k+(\alpha\beta)_{jk}+(\beta\gamma)_{kl}+(\alpha\beta\gamma)_{jkl}$ αντιστοιχεί στην πιθανότητα του κελιού θ_{jkl}.

Από τα παραπάνω προκύπτει ότι οποιαδήποτε υπόθεση για τη δομή της πιθανότητας κελιού μπορεί να διατυπωθεί παραλείποντας όρους στην $\beta_k+(\alpha\beta)_{jk}+(\beta\gamma)_{kl}+(\alpha\beta\gamma)_{jkl}$ καθ’ όσον η έκφραση του σταθερού περιθώριου αθροίσματος είναι απαραίτητο μέρος οποιουδήποτε μοντέλου.

Τα ίδια μοντέλα εφαρμόζονται για τις δύο προηγούμενες κατανομές πιθανοτήτων αν και υπάρχουν διαφορές στους όρους που που πρέπει να συμπεριληφθούν στα μοντέλα καθώς και στην ερμηνεία των όρων «αλληλεπίδρασης».

13.4 Εκτίμηση Μέγιστης Πιθανοφάνειας

Για τη κατανομή Poisson η λογαριθμική συνάρτηση μέγιστης πιθανοφάνειας είναι $l=\sum_i(y_i\log\lambda_i-\lambda_i-\log y_i!)$ όπου $\lambda_i=E(Y_i)$.

Για την πολυωνυμική κατανομή η λογαριθμική συνάρτηση μέγιστης πιθανοφάνειας είναι $l = \log n! + \sum_i \left(y_i \log \theta_i - \log y_i ! \right)$ η οποία μπορεί να γραφτεί στη μορφή $l = constant + \sum_i \log E\left(Y_i\right)$.

Διότι $E\left(Y_i\right) = n\theta_i$ (υποκείμενο στον περιορισμό $\sum \theta_i = 1$).

Όπου τα $E\left(Y_i\right)$ υπόκεινται σε διάφορους περιορισμούς. Συμπερασματικά λοιπόν, για τις προηγούμενες κατανομές οι συναρτήσεις λογαριθμικής πιθανοφάνειας εξαρτώνται μόνο από τις παρατηρούμενες συχνότητες των κελιών y και τις αναμενόμενες συχνότητές τους $E\left(y\right)$.

Υπάρχουν δύο προσεγγίσεις για τον υπολογισμό των αναμενό μενων συχνοτήτων. Η πρώτη μέθοδος υπολογίζει τις αναμενόμενες συχνότητες από την μεγιστοποίηση της λογαριθμικής συνάρτησης πιθανοφάνειας που υπόκειται στους σχετικούς περιορισμούς.

Η άλλη μέθοδος , που είναι σύμφωνη με την συνήθη προσέγγιση των γενικευμένων γραμμικών μοντέλων, χρησιμοποιεί το log-linear μοντέλο $\eta_i = \log E\left(Y_i\right) = X_i^T \beta$.

Κατ’ αρχήν εκτιμάται το β και στη συνέχεια χρησιμοποιείται αυτή η εκτίμηση για να υπολογιστούν τα η_i και από εδώ οι προσαρμοσμένες τιμές $\exp\left(\eta_i\right)$. Λόγω των ιδιοτήτων των εκτιμητών μέγιστης πιθανοφάνειας, αυτές οι προσαρμοσμένες τιμές θα είναι οι εκτιμητές μέγιστης πιθανοφάνειας των αναμενόμενων συχνοτήτων $E\left(Y_i\right)$.

Ο Birch το 1963 έδειξε ότι για κάθε log-linear μοντέλο οι εκτιμητές μέγιστης πιθανοφάνειας είναι οι ίδιοι για τις κατανομές πιθανοτήτων που συζητήθηκαν υπό τον όρο να συμπεριλαμβάνονται πάντα στο μοντέλο οι όροι που αντιστοιχούν στα σταθερά περιθώρια αθροίσματα. Αυτό σημαίνει ότι , για λόγους εκτίμησης μπορεί να θεωρηθεί πάντα η κατανομή Poisson. Επειδή η κατανομή Poisson ανήκει στην εκθετική οικογένεια, και οι περιορισμοί των παραμέτρων μπορεί να εξυπηρετηθούν με βολική επιλογή των στοιχείων του β, μπορεί να εφαρμοστούν όλα τα τυπικά αποτελέσματα των γενικευμένω ν γραμμικών μοντέλων. Πιο συγκεκριμένα μπορεί να εφαρμοστεί η μέθοδος του Newton – Raphson για την εκτίμηση του β. Αυτή η μέθοδος έχει υποστηριχθεί από τον Nelder (1974).

Η εναλλακτική προσέγγιση που βασίζεται στην απ' ευθείας εκτίμηση των αναμενόμενων συχνοτήτων των κελιών $E(Y_i)$, είναι να βρούμε τους εκτιμητές μέγιστης πιθανοφάνειας με ένα τέτοιο τρόπο, ώστε να ενσωματωθούν όλοι οι περιορισμοί στις πιθανότητες. Συνήθως δεν υπάρχουν κλειστής φόρμας λύσεις για τις αναμενόμενες συχνό τητες των κελιών από τοις εξισώσεις της περιορισμένης μέγιστης πιθανοφάνειας, συνεπώς πρέπει να χρησιμοποιηθούν αριθμητικές μέθοδοι για τον υπολογισμό προσεγγιστικών λύσεων. Μία επαναληπτική μέθοδος χρησιμοποιείται για να προσαρμόσει τις εκτιμώμενες αναμενόμενες συχνότητες των κελιών μέχρι να φθάσουν προσθετικά στα απαιτούμενα περιθώρια αθροίσματα (τουλάχιστον μέχρι κάποια προκαθορισμένη ακρίβεια). Αυτή η διαδικασία ονομάζεται «επαναληπτική αναλογική προσαρμογή [12].

13.5 Έλεγχος Υποθέσεων και Καλή Προσαρμογή

Μπορούμε να υποθέσουμε ότι οι N παράμετροι που υπάρχουν στο πλήρες μοντέλο είναι προσεγγιστικά οι αναμενόμενες συχνότητες. Μπορούν να εκτιμηθούν από τις αντίστοιχες παρατηρηθείσες συχνότητες $y_1,\dots,y_N$. Συνεπώς οι δύο κατανομές πιθανοτήτων έχουν για λογαριθμική συνάρτηση πιθανοφάνειας την εξίσωση της μορφής :

$$l(b_{\max};y) = constant + \sum_i y_i \log y_i$$

Για κάθε άλλο μοντέλο, έστω ότι e_i συμβολίζει τις αναμενόμενες συχνότητες των κελιών που εκτιμήθηκαν. Τότε η λογαριθμική συνάρτηση πιθανοφάνειας έχει τη μορφή :

$$l(b;y) = constant + \sum_i y_i \log e_i$$ όπου οι σταθερές είναι ίδιες.

Οπότε η στατιστική συνάρτηση του πηλίκου λογαριθμικής πιθανοφάνειας θα είναι :

$$D = 2\left[l(b_{\max};y) - l(b;y) \right] = 2\sum_{i=1}^{N} y_i \log \frac{y_i}{e_i}$$

η οποία έχει τη μορφή $D = 2\sum o \log \frac{o}{e}$ όπου o και e συμβολίζουν αντίστοιχα τις παρατηρηθείσες και τις αναμενόμενες που εκτιμήθηκαν (δηλ. προσαρμοσμένες) συχνότητες των κελιών. Το άθροισμα είναι πάνω σε όλα τα κελιά του πίνακα.

Εάν το μοντέλο προσαρμόζει καλά τα δεδομένα, τότε για μεγάλα δείγματα το D έχει κεντρική *χ-τετράγωνο* κατανομή με βαθμούς ελευθερίας που δίνονται από τον αριθμό των κελιών με μη μηδενικές παρατηρηθείσες συχνότητες (δηλαδή N εάν $y_i > 0$ για όλα τα i) μείον τον αριθμό των ανεξάρτητων μη μηδενικών παραμέτρων στο μοντέλο.

Στους πίνακες συνάφειας χρησιμοποιείται από την D πιο συχνά η στατιστική συνάρτηση χ-τετράγωνο : $X^2 = \sum \frac{(o-e)^2}{e}$.

Έχουμε δει σε προηγούμενα κεφάλαια, ότι οι δύο αυτές στατιστικές συναρτήσεις είναι ασυμπτωτικά ισοδύναμες και συνεπώς για μεγάλα δείγματα το X^2 έχει τη *χ-τετράγωνο* κατανομή με βαθμούς ελευθερίας που δίνονται όπως και παραπάνω.

Η μορφή που έχει η *χ-τετράγωνο* στατιστική συνάρτηση, προτείνει ότι το τυποποιημένο κατάλοιπο για κάθε κελί να ορίζεται ως : $r_i = \frac{o_i - e_i}{\sqrt{e_i}}$

Αυτός ο ορισμός των καταλοίπων προκύπτει φυσικά από το μοντέλο *Poisson* διότι $E(Y_i) = \text{var}(Y_i)$ και συνεπώς e_i είναι επίσης μία εκτίμηση της διασποράς για την συχνότητα του κελιού.

Παρέκκλιση από το μοντέλο μπορεί να ανιχνευθεί εξετάζοντας τα κατάλοιπα. Τιμές που είναι πολύ μακριά από το μηδέν σε οποιαδήποτε κατεύθυνση (το $|r_i| > 3$ αντιστοιχεί χονδρικά στο 1% των ουρών της τυπικής κανονικής κατανομής) ή σχέδια στα κατάλοιπα από κάποια τμήματα του πίνακα μπορεί να προτείνει άλλα περισσότερο κατάλληλα μοντέλα.

13.5.1 Υπερδιασπορά (Overdispersion)

Υπερδιασπορά ονομάζεται το φαινόμενο στο οποίο η μεταβλητότητα είναι μεγαλύτερη από αυτή που αναμένεται.

Αυτό συμβαίνει κάποιες φορές που έχουμε απόκριση που ακολουθεί κατανομή Poisson. Όταν τα δεδομένα μετράνε πλήθος (απαριθμήσεις) , τότε αυτά μεταβάλλονται συνήθως πολύ περισσότερο από ότι θα αναμέναμε , εάν η κατανομή της απόκρισης είναι πραγματικά η Poisson.

Γνωρίζουμε ότι στην Poisson ο μέσος και η διακύμανση συμπίπτουν. Οπότε έχουμε υπερδιασπορά (Overdispersion) στην περίπτωση που η διακύμανση είναι μεγαλύτερη από την αναμενόμενη τιμή του μέσου που προκύπτει από ένα γενικευμένο γραμμικό μοντέλο.
Δηλαδή : Εάν η Y ακολουθεί Poisson τότε έχουμε υπερδιασπορά εάν : Var(Y) > E(Y) .

Η υπερδιασπορά είναι ένα φαινόμενο που δεν υπάρχει στα συνηθισμένα γραμμικά μοντέλα που έχουμε κανονικότητα, διότι στην κανονική κατανομή έχουμε μία ξεχωριστή παράμετρο από το μέσο για να περιγράψουμε μεταβλητότητα, τη σ^2, ενώ στη Poisson $\mu=\sigma^2$.

Η υπερδιασπορά είναι λοιπόν συνήθης στην εφαρμογή Poisson γενικευμένων γραμμικών μοντέλων σε μεταβλητές που παριστάνουν απαριθμήσεις.

<u>Αρνητική διωνυμική Παλινδρόμηση</u>

Μια εναλλακτική προσέγγιση για την μοντελοποίηση της υπερβολικής διασποράς είναι να ξεκινήσουμε από ένα τυποποιημένο μοντέλο Poisson παλινδρόμησης και να προσθέσουμε ένα τυχαίο *i* αποτέλεσμα που να εκπροσωπεί απαρατήρητα μη παρατηρηθείσα ετερογένεια.

Ας υποθέσουμε λοιπόν, ότι η δεσμευμένη κατανομή της απόκρισης Y_i δοθέντος θ_i είναι πράγματι Poisson με μέσο (και διακύμανση) $\theta_i\mu_i$.
Η ιδέα είναι ότι αν παρατηρηθεί θ_i τα δεδομένα θα μπορούσε να είναι Poisson. Δυστυχώς, εμείς δεν παρατηρούμε θ_i. Αντ 'αυτού, θεωρούμε ότι έχει μια δοθείσα κατανομή. Θα καταλήξει να είναι βολικό να υποθέσουμε ότι έχει κατανομή γάμμα με παραμέτρους με $\alpha = \beta = 1/\sigma^2$ όπου σ^2 αντιπροσωπεύει η διακύμανση του μη παρατηρήσιμου.

Με αυτές τις πληροφορίες μπορεί να υπολογιστεί η μη δεσμευμένη κατ ανομή της απόκρισης, που συμβαίνει να είναι αρνητική διωνυμική κατανομή με πυκνότητα :

$$P(Y \mid y) = \frac{\Gamma(\alpha + y)}{y!\Gamma(\alpha)} \frac{\beta^{\alpha}\mu^{y}}{(\mu + \beta)^{\alpha + y}} \qquad \text{Όπου } \alpha = \beta = 1/\sigma^2$$

Η αρνητική διωνυμική κατανομή είναι γνωστή ως η κατανομή του αριθμού των αποτυχιών πριν από k επιτυχίες σε μια σειρά δοκιμών Bernoulli με κοινή πιθανότητα επιτυχίας π.
Η προκύπτουσα πυκνότητα μπορεί να ληφθεί από την ανωτέρω έκφραση θέτοντας $a = k$ και $\pi = \beta(\mu + \beta)$.

Η αρνητική διωνυμική κατανομή με : $\alpha = \beta = 1/\sigma^2$ έχει :

Μέσο E(Y)=μ και διακύμανση $Var(Y) = \mu(1 + \sigma^2\mu)$.

Εάν το σ^2 τείνει στο μηδέν έχουμε αποκτήσει διακύμανση που τείνει στη διακύμανση Poisson. Εάν $\sigma^2 > 0$ τότε η διακύμανση είναι μεγαλύτερο από το μέσο. Έτσι, η αρνητική διωνυμική κατανομή έχει υπερδιασπορά σε σχέση με την Poisson.

Είναι ενδιαφέρον ότι μπορεί κανείς να αντλήσει τον ίδιο μέσο και διασπορά χωρίς την παραδοχή ότι οι μη παρατηρήσιμες έχει μια διανομή γάμμα. Μπορούμε να προχωρήσουμε στη εκτίμηση των παραμέτρων που επηρεάζουν το μ για μια σταθερή τιμή της σ^2 χρησιμοποιώντας μέγιστη οιονεί πιθανοφάνεια. Η στρατηγική αυτή έχει εφαρμοστεί σε διαδικασίες διαφόρων στατιστικών πακέτων (πχ. GLM Stata), αλλά αυτό δεν λύνει το πρόβλημα της εκτίμησηςτου ίδιου του σ^2.

Η εναλλακτική λύση είναι η χρήση μέγιστης πιθανοφάνειας, η οποία απαιτεί την παραδοχή κατανομής γάμμα για τις μη παρατηρήσιμες, έτσι ώστε το αποτέλεσμα να έχει αρνητική διωνυμική κατανομή. Αυτή η στρατηγική, η οποία θεσπίζει ισχυρότερες παραδοχές αλλά αποδίδει εκτιμήσεις για το σ^2 και τις παραμέτρους που επηρεάζουν το μ, έχει υλοποιηθεί με σε κάποια στατιστικά πακέτα (όπως με την διαδικασία nbreg του Stata).

Επειδή το μοντέλο Poisson είναι μια ειδική περίπτωση της αρνητικής διωνυμικής, (δηλαδή η περίπτωση $\sigma^2 = 0$), μπορεί κανείς να χρησιμοποιήσει ένα τυπικό έλεγχο πηλίκου πιθανοφάνειας για τη σύγκρισή τους. Υπάρχει, ωστόσο, μια μικρή δυσκολία.

Επειδή το μοντέλο Poisson είναι σε ένα όριο του χώρου παραμέτρων, η στατιστική συνάρτηση ελέγχου δεν έχει την τυπική χ^2 κατανομή με ένα βαθμό ελευθερίας.

Ορισμένες έρευνες δείχνουν ότι η κατανομή προσεγγίζεται καλύτερα ως μείγμα 50:50 από το μηδέν και μία κατανομή chi-τετράγωνο με ένα βαθμό ελευθερίας (και αυτό είναι που κάνει το στατιστικό πακέτο Stata).

Κλείνοντας σημειώνουμε ότι υπάρχουν εναλλακτικές διατυπώσεις του αρνητικού διωνυμικού μοντέλου που οδηγούν σε ελαφρώς διαφορετικά μοντέλα, μεταξύ των οποίων και ένα που οδηγεί στην υπερβολική διασπορά

Poisson της προηγούμενης ενότητας. Η διατύπωση όμως που δίνεται εδώ, είναι η μοναδική σε κοινή χρήση.

13.6 Αριθμητικά Παραδείγματα

Παράδειγμα 17: Διατμηματική μελέτη του κακοήθους μελανώματος [13]

Τα δεδομένα που υπάρχουν στον παρακάτω πίνακα είναι από μία διατμηματική μελέτη ασθενών οι οποίοι πάσχουν από μία μορφή καρκίνου που ονομάζεται «κακοήθες μελάνωμα». Για ένα δείγμα από n=400 ασθενών παρατίθενται στον ακόλουθο πίνακα ο αριθμός των ασθενών σε σχέση με τη θέση του όγκου και τον ιστολογικό του τύπο.

Πίνακας 28 : Δεδομένα κακοήθους μελανώματος

	Θέση			
Τύπος Όγκου	Κεφάλι και Λαιμός	Κορμός	Άκρα	Σύνολο
Μελανωματικά στίγματα του Hutchinson	22	2	10	34
Επιφανειακά εξαπλούμενο μελάνωμα	16	54	115	185
Κομβώδες	19	33	73	125
Ακαθόριστο	11	17	28	56
Σύνολο	68	106	226	400

Στο παράδειγμα υπάρχουν δύο μεταβλητές απόκρισης, η «θέση» και ο «τύπος όγκου».

Οι συχνότητες στα κελιά θεωρούνται ότι είναι τυχαίες μεταβλητές με τον περιορισμό ότι πρέπει να αθροίζονται στο *n*.

Η ερώτηση που μας ενδιαφέρει είναι το εάν υπάρχει κ άποια συσχέτιση μεταξύ των δύο μεταβλητών απόκρισης. Στον πίνακα που ακολουθεί εμφανίζονται ποσοστά των συνόλων στις γραμμές και τις στήλες. Εμφανίζεται ότι τα «Μελανωματικά στίγματα του Hutchinson» είναι πιο συνηθισμένα στο κεφάλι και το λαιμό, αλλά υπά ρχουν ελάχιστες αποδείξεις για σχέση μεταξύ άλλου τύπου όγκου και θέσης.

Πίνακας 29 : Κακοήθες μελάνωμα – ποσοστά ως προς τις γραμμές και στήλες για τύπο όγκου και Θέση

	Θέση			
Τύπος Όγκου	Κεφάλι και Λαιμός	Κορμός	Άκρα	Όλες οι Θέσεις
Ποσοστά Γραμμών				
Μελανωματικά στίγματα του Hutchinson	64, 7	5, 9	29, 4	100
Επιφανειακά εξαπλούμενο μελάνωμα	8, 6	29, 2	62, 2	100
Κομβώδες	15, 2	26, 4	58, 4	100
Ακαθόριστο	19, 6	30, 4	50, 0	100
Όλοι οι Τύποι	17, 0	26, 5	56, 5	100
Ποσοστά Στηλών				
Μελανωματικά στίγματα του Hutchinson	32, 4	1, 9	4, 4	8, 50
Επιφανειακά εξαπλούμενο μελάνωμα	23, 5	50, 9	50, 9	46, 25
Κομβώδες	27, 9	31, 1	32, 3	31, 25
Ακαθόριστο	16, 2	16, 0	12, 4	14, 00
Όλοι οι Τύποι	100, 0	99, 9	100	100, 0

Θέλουμε να εξετάσουμε εάν υπάρχει οποιαδήποτε σχέση μεταξύ τύπου όγκου και θέσης. Αυτό γίνεται εξετάζοντας την μηδενική υπόθεση H_0: *οι δύο μεταβλητές είναι ανεξάρτητες*.

$H_0: E\left(Y_{jk}\right) = n\theta_{j.}\theta_{.k}$

Όπου $\sum \theta_{j.} = 1$ και $\sum \theta_{.k} = 1$.

Το αντίστοιχο log-linear μοντέλο είναι $\eta_{jk} = \log E\left(Y_{jk}\right) = \mu + \alpha_j + \beta_k$ υποκείμενο σε περιορισμούς όπως $\sum \alpha_j = 0$ και $\sum \beta_k = 0$ (περιορισμοί αθροίσματος στο μηδέν) ή

$a_1 = 0$ και $\beta_1 = 0$ (γωνιακοί περιορισμοί).

Επειδή υπάρχουν J=4 τύποι όγκων και K=3 θέσεις, υπάρχουν $1+\left(J-1\right)+\left(K-1\right) = J+K-1 = 6$ παράμετροι που πρέπει να εκτιμηθούν.

Το μοντέλο αυτό συγκρίνεται (σιωπηρά) με το πλήρες μοντέλο $\eta_{jk} = \log E\left(Y_{jk}\right) = \mu + \alpha_j + \beta_k + \left(\alpha\beta\right)_{jk}$ που υπόκειται στους κατάλληλους περιορισμούς. Το πλήρες μοντέλο έχει 12 παραμέτρους έτσι ώστε $\hat{\eta}_{jk} = \log y_{jk}$ και D=0.

Εάν η H_0 είναι σωστή τότε η στατιστική συνάρτηση D ή το X^2 έχουν την κατανομή χ^2_{N-p} όπου N=12 και p =6.

Χρησιμοποιείται η κατανομή Poisson με συνάρτηση σύνδεσης την λογαριθμική. Οι παράμετροι αντιστοιχούν σε γωνιακούς περιορισμούς.

Οι προσαρμοσμένες τιμές προκύπτουν από από τις εκτιμήσεις όπως φαίνεται στα ακόλουθα παραδείγματα :

$\hat{\eta}_{11} = 1.754$ οπότε $e_{11} = e^{1.754} = 5.78$

$\hat{\eta}_{43} = 1.754 + 0.499 + 1.201 = 3.454$ οπότε $e_{43} = e^{3.454} = 31.64$

Η στατιστική συνάρτηση πηλίκου λογαριθμικής πιθανοφάνειας είναι $D = 2\sum o \log\left(o/e\right) = 51.759$.

Η στατιστική συνάρτηση $X^2 = \sum\left[\left(o-e\right)^2/e\right] = 65.813$

Και από τις δύο ανωτέρω στατιστικές συναρτήσεις φαίνεται ότι το μοντέλο προσαρμόζεται φτωχά επειδή $\Pr\left(\chi_6^2 > 50\right) < 0.001$

Συνεπώς απορρίπτουμε τη μηδενική υπόθεση H_0 και συμπεραίνουμε ότι υπάρχει σχέση μεταξύ τύπου όγκου και θέσης.

Τα κατάλοιπα δίνονται από τον τύπο $r_i = \frac{o_i - e_i}{\sqrt{e_i}}$.

Το μεγαλύτερο κατάλοιπο, παρατήρηση 1 για το κελί (1, 1), μετράει πολύ για την έλλειψη της προσαρμογής, επιβεβαιώνοντας ότι το κύριο σήμα στα δεδομένα είναι η σχέση των στιγμάτων μελανώματος του Hutchinson με το κεφάλι και τον λαιμό.

Λύση του προβλήματος με χρήση R

Δημιουργούμε τον πίνακα **pnx** των συχνοτήτων και στη συνέχεια τη δομή πίνακα **tbl** με την εντολή **as.table** για να δουλέψουμε με τις εντολές πινάκων. Για ευκολία συμβολίζουμε με A1 ,A2, A3, A4 για τις γραμμές και B1, B2, B3 για τις στήλες.

```
> r1<-c(22,2,10)
> r2<-c(16,54,115)
> r3<-c(19,33,73)
> r4<-c(11,17,28)
> pnx<-rbind(r1,r2,r3,r4)
> pnx
   [,1] [,2] [,3]
r1   22    2   10
r2   16   54  115
r3   19   33   73
r4   11   17   28

> dimnames(pnx)<-list(c("A1","A2","A3","A4"),c("B1","B2","B3") )
> tbl<-as.table(pnx)
```

```
> summary(tbl)
Number of cases in table: 400
Number of factors: 2
Test for independence of all factors:
        Chisq = 65.81, df = 6, p-value = 2.943e-12
```

Από το **Chisq = 65.81** και **p-value = 2.943e-12** συμπεραίνουμε ότι απορρίπτεται η $\mathbf{H_0}$ και υπάρχει σχέση μεταξύ τύπου όγκου και θέσης.

Στη συνέχεια φορτώνουμε τη βιβλιοθήκη **gmodels** και χρησιμοποιούμε τη συνάρτηση **CrossTable** για τη δημιουργία του πίνακα συνάφειας.

(Αν δεν υπάρχει η βιβλιοθήκη τότε :

- την κατεβάζουμε με τη βοήθεια της εντολής του περιβάλλοντος της R Packages -> Install Packages.
- Επιλέγουμε χώρα προέλευσης USA
- Επιλέγουμε gmodels
- Αφού κατέβη το πακέτο το φορτώνουμε)

```
> CrossTable(tbl)

   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  400
```

	B1	B2	B3	Row Total
A1	22	2	1 0	34
	45.517	5.454	4.416	
	0.647	0.059	0.294	0.085
	0.324	0.019	0.044	
	0.055	0.005	0.025	
A2	16	54	115	185
	7.590	0.505	1.050	
	0.086	0.292	0.622	0.463
	0.235	0.509	0.509	
	0.040	0.135	0.287	
A3	19	33	73	125
	0.238	0.000	0.080	
	0.152	0.264	0.584	0.312
	0.279	0.311	0.323	
	0.048	0.082	0.182	
A4	11	17	28	56
	0.230	0 .314	0.419	
	0.196	0.304	0.500	0.140
	0.162	0.160	0.124	
	0.028	0.042	0.070	
Column Total	68	106	226	400
	0.170	0.265	0.565	

<u>Παράδειγμα 18</u> : Μελέτη ελέγχου περιπτώσεων για γαστρικό έλκος και έλκος δωδεκαδακτύλου λόγω χρήσης ασπιρίνης

Σε αυτή την μελέτη ελέγχου περιπτώσεων, συγκεντρώθηκε μία ομάδα από ασθενείς με έλκος και μία ομάδα από ασθενείς ελέγχου που δεν είναι γνωστό αν έχουν έλκος. Οι δύο ομάδες έχουν παρόμοια χαρακτηριστικά ως προς την ηλικία, το φύλο και την κοινωνικό – οικονομική κατάσταση.
Οι ασθενείς με έλκος χωρίστηκαν σε δύο κλάσεις σε σχέση με το είδος του έλκους (γαστρικού ή δωδεκαδακτύλου). Χρήση ασπιρίνης εξετάστηκε σε όλους τους ασθενείς. Τα αποτελέσματα δείχνονται στον παρακάτω πίνακα.

Πίνακας 30 : Γαστρικό έλκος[14] και έλκος δωδεκαδακτύλου και χρήση ασπιρίνης. Συχνότητες

	Χρήση Ασπιρίνης		
	Μη - Χρήστες	Χρήστες	Σύνολο
Γαστρικό Έλκος			
Περιπτώσεις	39	25	64
Ελέγχου	62	6	68
Δωδεκαδακτυλικό Έλκος			
Περιπτώσεις	49	8	57
Ελέγχου	53	8	61

Αυτός είναι ένας $2\times2\times2$ πίνακας συνάφειας με μία μεταβλητή απόκρισης (Χρήση Ασπιρίνης) και δύο επεξηγηματικές μεταβλητές. Στους υποπίνακες για κάθε είδος έλκους τα σύνολα των γραμμών λαμβάνονται σταθερά.

Οι σχετικές ερωτήσεις είναι :
Έχει σχέση το γαστρικό έλκος με τη χρήση της ασπιρίνης
Έχει σχέση το έλκος δωδεκαδακτύλου με τη χρήση ασπιρίνης

Είναι ίδια οποιαδήποτε σχέση που αφορά τη χρήση ασπιρίνης και για τις δύο περιπτώσεις έλκους.

Όταν τα δεδομένα παρουσιάζονται με μορφή ποσοστών ως προς τα σύνολα των γραμμών, φαίνεται ότι η χρήση ασπιρίνης είναι πιο συχνή στους ασθενείς με έλκος από αυτούς που αποτελούν την ομάδα ελέγχου στην περίπτωση του γαστρικού έλκους, αλλά όχι για το έλκος του δωδεκαδακτύλου.

Πίνακας 1 : Γαστρικό έλκος και έλκος Δωδεκαδακτύλου και χρήση ασπιρίνης

	Χρήση Ασπιρίνης		
	Μη - Χρήστες	Χρήστες	Σύνολο
Γαστρικό Έλκος			
Περιπτώσεις	61	39	100
Ελέγχου	91	9	100
Δωδεκαδακτυλικό Έλκος			
Περιπτώσεις	86	14	100
Ελέγχου	87	13	100

Για να εξετάσουμε πότε το γαστρικό έλκος έχει σχέση με τη χρήση ασπιρίνης , ελέγχουμε τη μηδενική υπόθεση ότι η πιθανότητα η χρήση ασπιρίνης να είναι ανεξάρτητη από το καθεστώς της νόσου (περιπτώσεων ή ελέγχου) για την ομάδα του γαστρικού έλκους ($\boldsymbol{l}$=**1**) ως προς την εναλλακτική υπόθεση της μη-ανεξαρτησίας .

H0 : $E\left(Y_{jkl}\right) = y_{j.l}\theta_{.kl}$

H1 : $E\left(Y_{jkl}\right) = y_{j.l}\theta_{jkl}$

(Τα αθροίσματα των γραμμών λαμβάνονται σαν σταθερά).

Ανάλογες υποθέσεις μπορεί να τεθούν και για το έλκος δωδεκαδακτύλου (χρησιμοποιώντας l=2).

Οπότε η από κοινού υπόθεση ότι δεν υπάρχει συσχέτιση μεταξύ ασπιρίνης και ασθένειας, για τις δύο περιπτώσεις έλκους, αντιστοιχεί στο log-linear μοντέλο :

$$\eta_{jkl} = \log E\left(Y_{jkl}\right) = \mu + \alpha_j + \gamma_l + \left(\alpha\gamma\right)_{jl} + \beta_k + \left(\beta\gamma\right)_{kl} \qquad \textbf{(M1)}$$

όπου οι τέσσερις πρώτοι όροι αντιστοιχούν στα σταθερά αθροίσματα των γραμμών $y_{j.l}$ και οι δύο τελευταίοι όροι καλύπτουν την επίδραση της ασπιρίνης, επιτρέποντας διαφορετικά δυνατά επίπεδα χρήσης ασπιρίνης για κάθε θέση έλκους.

Αυτό το μοντέλο συγκρίνεται με το πλήρες μοντέλο για να ελέγξει την υπόθεση «μη συσχέτισης» μεταξύ της χρήσης ασπιρίνης και έλκους.
Εάν αυτή η υπόθεση απορριφθεί, μπορούμε να ελέγξουμε την υπόθεση ότι ο βαθμός της σχέσης είναι ο ίδιος για τις δύο θέσεις έλκους χρησιμοποιώντας ένα όρο επιπλέον πολλαπλασιαστικής πιθανότητας $\theta_{jk.}$ ή ισοδύναμα το log-linear μοντέλο :

$$\eta_{jkl} = \mu + \alpha_j + \gamma_l + \left(\alpha\gamma\right)_{jl} + \beta_k + \left(\beta\gamma\right)_{kl} + \left(\alpha\beta\right)_{jk} \qquad \textbf{(M2)}$$

Για το μοντέλο (M1) η στατιστική συνάρτηση πηλίκου λογαριθμικής πιθανοφάνειας δίνει D=17.697 με 2 βαθμούς ελευθερίας. Αυτό δείχνει φτωχή προσαρμογή, συνεπώς θα πρέπει να απορρίψουμε την υπόθεση της «μη συσχέτισης» μεταξύ χρήσης ασπιρίνης και γαστρικού ή δωδεκαδακτυλικού έλκους.

Το μοντέλο (M2) είναι σημαντικά καλύτερο. Δίνει ΔD=11.41 με 1 βαθμό ελευθερίας επιβεβαιώνοντας την ύπαρξη σχέσης. Ωστόσο είναι ακόμα φτωχό , με D=6.283 με 1 βαθμό ελευθερίας οπότε πρέπει να απορρίψουμε την υπόθεση ότι η συσχέτιση με τη χρήση ασπιρίνης είναι ίδια για γαστρικό και δωδεκαδακτυλικό έλκος. Αυτό είναι συνεπές με την παρατήρηση ότι τα δεδομένα του πίνακα δείχνουν ότι η χρήση ασπιρίνης έχει σχέση με το γαστρικό έλκος αλλά όχι με το δωδεκαδακτυλικό.

Λύση του προβλήματος με χρήση R

Για να εξετάσουμε πότε το γαστρικό έλκος έχει σχέση με τη χρήση ασπιρίνης χρησιμοποιούμε μόνο τον υποπίνακα που έχει σχέση με το γαστρικό έλκος.

```
> c1<-c(61,39)
> c2<-c(91,9)
> c3<-c(86,14)
> c4<-c(87,13)
> pnx12<-rbind(c1,c2)
> tbl12<-as.table(pnx12)
> tbl12
   A  B
c1 61 39
c2 91  9
> summary(tbl12)
Number of cases in table: 200
Number of factors: 2
Test for independence of all factors:
        Chisq = 24.671, df = 1, p-value = 6.8e-07
```

Θα εξετάσουμε τώρα την υπόθεση «μη συσχέτισης» μεταξύ της χρήσης ασπιρίνης και έλκους. Αυτό το μοντέλο συγκρίνεται με το πλήρες μοντέλο . Με A , B συμβολίζουμε τους παράγοντες.

```
>A<-gl(4, 2, labels = c("Gastric cases","Gastric Control", "Dod Cases", "Dod Control"))
> B<-gl(2, 1,8, labels = c("Use", "Not User"))

> print(d.AD<-data.frame(A,B,times))
            A             B          times
1    Gastric cases     Use          61
2    Gastric cases     Not User     39
```

3	Gastric Control	Use	91
4	Gastric Control	Not User	9
5	Dod Cases	Use	86
6	Dod Cases Not	User	14
7	Dod Control	Use	87
8	Dod Control	Not User	13

> glm_aspirin <- glm(times ~ A + B, family=poisson())

> anova(glm_aspirin)

Analysis of Deviance Table
Model: poisson, link: log

Response: times

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev
NULL			7	201.991
A	3	0.000	4	201.991
B	1	168.456	3	33.535

> summary(glm_aspiri n)

Call:
glm(formula = times ~ A + B, family = poisson())

Deviance Residuals:

1	2	3	4	5	6	7	8
-2.3511	4.0773	1.0610	-2.5077	0.5220	-1.1490	0.6306	-1.4063

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	4.398e+00	1.028e-01	42.76	<2e-16 ***
AGastric Control	-1.117e-15	1.414e-01	-7.90e-15	1
ADod Cases	-1.558e-15	1.414e-01	-1.10e-14	1
ADod Control	-1.173e-15	1.414e-01	-8.30e-15	1
BNot User	-1.466e+00	1.281e-01	-11.45	<2e-16 ***

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 201.991  on 7  degrees of freedom
Residual deviance:  33.535  on 3  degrees of freedom
AIC: 86.9

Number of Fisher Scoring iterations: 5
```

13.7 Παρατηρήσεις

Τα προηγούμενα παραδείγματα είναι ιδιαίτερα απλά οπότε οι υπολογισμοί εύκολοι και η ερμηνεία των αποτελεσμάτων απλή. Για πίνακες συνάφειας που εμπλέκουν πάνω από τρεις μεταβλητές, η επιλογή των μοντέλων καθώς και η ερμηνεία γίνεται πολύ πιο πολύπλοκη[15].

Μία εναλλακτική προσέγγιση από τις μεθόδους πιθανοφάνειας που μελετήσαμε στα προηγούμενα βασίζεται σε μοντελοποίηση συναρτήσεων των πολυωνυμικών πιθανοτήτων θ σαν γραμμικών συνδυασμών των παραμέτρων[16], δηλαδή: $F(\theta) = X\beta$ και χρησιμοποιώντας το κριτήριο των σταθμισμένων ελαχίστων τετραγώνων $S_w = \left[F(p) - X\beta\right]^T V^{-1}\left[F(p) - X\beta\right]$

για εκτίμηση και έλεγχο υποθέσεων (όπου το p αντιπροσωπεύει τις εκτιμηθείσες πιθανότητες και V είναι ο πίνακας διακύμανσης-συνδιακύμανσης για το F(p)).

Ένα από τα πλεονεκτήματα της μεθόδου αυτής είναι ότι μπορεί να χρησιμοποιηθεί σε γραμμικά και μη-γραμμικά (συμπεριλαμβανομένων log-linear) μοντέλα.

Η μέθοδος αυτή είναι υπολογιστικά πιο πολύπλοκη από τις μεθόδους πιθανοφάνειας και χρησιμοποιείται λιγότερο. Για μια εισαγωγή σε αυτή την προσέγγιση βλέπε Freeman (1987).

Οι μέθοδοι των πινάκων συνάφειας, συμπεριλαμβανομένων των log-linear μοντέλων, σχεδιάστηκαν αρχικά για την ανάλυση ονομαστικών κατηγορικών δεδομένων. Στην πράξη χρησιμοποιούνται επίσης για διατεταγμένες κατηγορίες είτε αγνοώντας τη διάταξη ή αναθέτοντας τιμές μεταβλητών στις κατηγορίες, βλέπε[17] .

Ο McCullagh (1980) έδειξε ότι τα γενικευμένα γραμμικά μοντέλα μπορεί να επεκταθούν ώστε να δώσουν μεθόδους που έχουν την ίδια μορφή με μεθόδους παλινδρόμησης για διατεταγμένα δεδομένα[18].

14 ΒΙΒΛΙΟΓΡΑΦΙΑ

Βιβλιογραφία που χρησιμοποιήθηκε

Γενικευμένα γραμμικά μοντέλα

Alan Agresti : An Introduction to Categorical Data Analysis (Wiley 2007)

P.McCullagh, J.A. Nelder : Generalized Linear Models (Chapma & Hall 1990)

Dobson A. : An Introductionn to Generalized Linear Models (Chapma & Hall 1994)

Γιάννης Ντζούφρας : Σημειώσεις παραδώσεων (Οικονομικό Πανεπιστήμιο Αθηνών)

Γλώσσα προγραμματισμού R

W.N.Venables, D.M. Smith and the R Development Core Team.: An Introduction to R (http://www.r-project.org)

R Development Core TeamR : Language Definition (http://www.r-project.org)

Επιπλέον Βιβλιογραφία

Γενικευμένα γραμμικά μοντέλα

McCulloch, C.E. & Searle, S.R. : Generilized, linear and mixed models (Wiley 2001)

Fahrmeir L., Tutz G. : Multivariate statistical mode ling based on generalized linear models(Springer Verlag 1980)

Hosmer D.W. and Lemeshov S. : Applied logistic Regression (Wiley 2000)

Collet D : Modeling Binary Data, Decond Edition (Chapman Hall 2003)

Aldrich J.H. and F.D. Nelson : Linear probability, Lo git and Probit Models (SAGE University paper 1984)

Cox D.R. and E.J. Snell : Analysis of binary data. Second edition (Chapman Hall 1989)

Fienberg S.E. : The Analysis of Cross-Classified Categorical Data .Second edition (Springer Verlag 1980)

Γλώσσα προγραμματισμού R

Ιστοσελίδα : http://www.r-project.org

Παραπομπές

[1] Για την απόδειξη μπορείτε να δείτε το “An Introduction to Generalized Linear Models” A. Dobson Appendix A (Chapman Hall 1994)

[2] Cox and Hinkley 1974, κεφάλαιο 2.(Chapman Hall)

[3] Για την απόδειξη” A. Dobson Appendix B .(Chapman Hall 1994)

[4] Απόδειξη A. Dobson (1994) Appendix B

[5] Για επιπλέον πληροφορίες για τα κατάλοιπα στα γενικευμένα γραμμικά μοντέλα McCullagh & Nelder (.Chapman Hall 1989) .

[6] Περισσότερες πληροφορίες sto “An Introduction to Statistical Concepts” των Kleinbaum, Kupper and Muller (Duxbury Resource center 1988)

[7] Περισσότερα στο βιβλίο των Draper & Smith «Applied regression Analysis” (Wiley 1981)

[8] Δίνεται από τον Holliday στο αρθρο “Plant Population and crop Yield.(1960)

[9] Δεδομένα από “Statistical Principles in Experimental Design” Winer 1971, σελίδα 776

[10] Τα δεδομένα προέρχονται από το “Biomedical Journal” Bliss 1935.

[11] Απόδειξη στο βιβλίο “Discrete Multivariate Analysis : Theory and practice” των Bishop, Fienberg and Holland 1975, παρ 13.4 (M.I.T.)

[12] Η μέθοδος αυτή περιγράφεται στο βιβλίο “Discrete Multivariate Analysis : Theory and practice” των Bishop, Fienberg and Holland 1975.

[13] Κακοήθες μελάνωμα, συχνότητες για τύπο όγκου και θέση . Αρθρο Roberts (1981) “Maligant Melanoma of the Agina and Locoregional Control”..

[14] Αρθρο Duncan 1986

[15] Για συστηματική προσαρμογή πολύπλοκων log -linear μοντέλων δίνεται από τους Bishop, Fienberg και Holland (1975 κεφάλαια 4 και 9).

[16] Έχει προταθεί από τους Grizzle, Starmer και Koch (1969)στο άρθρο τους “Analysis of Categorical data by Linear models”.

[17] Στο άρθρο “Coding Ordinal Independent Variables in Multiple regression Analysis” , των Walter, Feinstein και Wells (1987)

[18] Βιβλίο αναφοράς αποτελεί το βιβλίο των McCullagh και Nelder (1989).