ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Διπλωματική Εργασία

Στατιστική Ανάλυση Πολυμεταβλητών Αστρονομικών, Βιολογικών και Αρχαιολογικών Δεδομένων

Αντωνία Ζήκου

Επιβλέπουσα Καθηγήτρια: Καρώνη Χρυσής

Επιτροπή Καθηγητών:

Χ. Καρώνη,	Ι. Βόντα,	Β. Παπανικολάου,
Καθηγήτρια, ΕΜΠ	Αναπληρώτρια Καθηγήτρια, ΕΜΠ	Καθηγητής, ΕΜΠ

Περιεχόμενα

Ευχαριστίες

Το χρονικό διάστημα κατά το οποίο διαμορφωνόταν η παρούσα διπλωματική εργασία συνεργάστηκαν και βοήθησαν πολλά άτομα, άμεσα και άλλοτε έμμεσα.

Θα ήθελα πρώτα από όλα να ευχαριστήσω μέσα από καρδιάς την καθηγήτρια του Ε.Μ.Π. και επιβλέπουσα στη συγκεκριμένη εργασία κα Χρυσηίδα Καρώνη, για την αμέριστη βοήθεια που μου προσέφερε. Ήταν δίπλα μου όποτε και αν την χρειάστηκα. Επίσης μου έδωσε την ευκαιρία να ασχοληθώ και να εργαστώ με ένα πολύ ενδιαφέρον και επίκαιρο θέμα. Η διπλωματική μου ήταν μια από τις καλύτερες εμπειρίες της σχολής μου χάρης εκείνη.

Ένα μεγάλο ευχαριστώ χρωστάω σε όλους τους δασκάλους, στην Ε και στους καθηγητές μου, καθώς χωρίς την δική τους επιμονή και υπομονή δεν θα είχα καταφέρει όσα έχω μέχρι σήμερα.

Οι γονείς μου και ο Βασίλης ήταν εκείνοι που στάθηκαν δίπλα μου και με στήριξαν όλα αυτά τα χρόνια, όποια επιλογή και αν έκανα. Δεν θα μπορούσα να μην τους ευχαριστήσω για την βοήθεια που μου έχουν προσφέρει όχι μόνο αυτούς τους τελευταίους μήνες, αλλά καθ' όλη την πορεία της εκπαίδευσής μου.

Τέλος θα ήθελα να στείλω ένα ιδιαίτερο ευχαριστώ στον Μιλτιάδη. Χωρίς εκείνον ίσως οι σπουδές μου να είχαν ολοκληρωθεί νωρίτερα, αλλά σίγουρα δεν θα ήμουν το ίδιο άτομο και η διπλωματική αυτή θα ήταν πολύ διαφορετική.

Αντωνία Ζήκου

Αθήνα, Μάιος 2018

ΠΕΡΙΛΗΨΗ

Η Λογιστική Παλινδρόμηση είναι μία στατιστική τεχνική ανάλυσης δεδομένων, η οποία βρίσκει ευρεία εφαρμογή στις μέρες μας. Σε αυτήν την εργασία ασχοληθήκαμε τόσο με το απλό όσο και το πολλαπλό μοντέλο λογιστικής παλινδρόμησης. Παράλληλα, αναπτύξαμε διάφορες μεθόδους ανάλυσης και συμπίεσης δεδομένων, προκειμένου να δειχθούν οι σχέσεις των μεταβλητών.

Στο πρώτο κεφάλαιο αναφέραμε περιληπτικά τα μοντέλα της λογιστικής παλινδρόμησης (απλό και πολλαπλό). Ασχοληθήκαμε κυρίως με την μορφή του μοντέλου, την προσαρμογή του και με τους ελέγχους που χρειάζεται να πραγματοποιηθούν.

Το δεύτερο κεφάλαιο αφορά την ανάλυση κυρίων συνιστωσών, μία μέθοδο που χρησιμοποιήθηκε σε αυτήν την εργασία κυρίως για την συμπίεση των δεδομένων μας και την απλοποίηση του μοντέλου μας.

Στο τρίτο κεφάλαιο κάνουμε μια αναφορά στην ανάλυση συστάδων και τα μέτρα ομοιότητας. Στο κεφάλαιο αυτό δίνεται και ένα μικρό παράδειγμα για την πλήρη κατανόηση των εννοιών.

Τέλος τα τέσσερα τελευταία κεφάλαια περιλαμβάνουν ένα ευρύ πεδίο εφαρμογών όλων των παραπάνω τεχνικών. Παρουσιάζονται και συζητούνται τα αποτελέσματα στους διάφορους τομείς της επιστήμης.

ABSTRACT

Logistic regression analysis is a statistical technique of data analysis, which nowadays is widely applied. In this thesis, we deal with both simple and multinomial logistic regression models. Furthermore, we develop various methods of data analysis and reduction in order to show the relationship among the variables.

In the first chapter we concisely discuss the models of logistic regression (simple and multinomial). We mainly deal with the form, the fitting of the model and the tests that have to be carried out.

The second chapter concerns principal components analysis. In this thesis, we employ this method chiefly for reducing the dimensionality of our data and simplifying our model.

In the third chapter we describe cluster analysis and similarity measures. This chapter also presents a small example for a complete understanding of the concepts.

Finally, the last four chapters include a wide range of applications of all the above techniques. The results, from various fields of science, are presented and discussed.

1 ΕΙΣΑΓΩΓΗ ΣΤΑ ΜΟΝΤΕΛΑ ΛΟΓΙΣΤΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

1.1 Διωνυμική λογιστική παλινδρόμηση

Οι μέθοδοι της παλινδρόμησης είναι ένα πολύτιμο εργαλείο όταν πρόκειται να μελετήσουμε τη σχέση μιας μεταβλητής απόκρισης (εξαρτημένη) με μια ή περισσότερες επεξηγηματικές μεταβλητές ανάλυσης δεδομένων. Αρκετά συχνά η μεταβλητή απόκρισης είναι διακριτή, λαμβάνοντας δύο ή περισσότερες πιθανές τιμές. Το μοντέλο της λογιστικής παλινδρόμησης είναι το πιο συχνά χρησιμοποιούμενο μοντέλο.

Πριν ξεκινήσουμε να περιγράψουμε εις βάθος το μοντέλο της λογιστικής παλινδρόμησης, είναι σημαντικό να κατανοήσουμε ότι ο στόχος της ανάλυσης με τη χρήση αυτού του μοντέλου είναι ίδια με εκείνη οποιουδήποτε άλλου μοντέλου παλινδρόμησης που χρησιμοποιείται στη στατιστική, δηλαδή, να βρει το βέλτιστο και με τις λιγότερο δυνατές μεταβλητές μοντέλο, προκειμένου να περιγράψουμε τη σχέση μεταξύ της εξαρτημένης μεταβλητής και ενός συνόλου ανεξάρτητων, κατηγορηματικών μεταβλητών. Οι ανεξάρτητες μεταβλητές συχνά καλούνται *συμμεταβλητές*. Ένα κοινό παράδειγμα μοντελοποίησης είναι το σύνηθες μοντέλο γραμμικής παλινδρόμησης, όπου η μεταβλητή απόκρισης θεωρείται ότι είναι συνεχής.

Αυτό που διακρίνει ένα μοντέλο λογιστικής παλινδρόμησης από το απλό γραμμικό μοντέλο είναι ότι η μεταβλητή απόκρισης στη λογιστική παλινδρόμηση είναι δυαδική ή δίτιμη. Αυτή η διαφορά μεταξύ λογιστικής και γραμμικής παλινδρόμησης αντικατοπτρίζεται τόσο από την ίδια τη μορφή του μοντέλου όσο και από τις υποθέσεις/παραδοχές του. Όταν η διαφορά αυτή κατανοηθεί, οι μέθοδοι που χρησιμοποιούνται σε μια ανάλυση λογιστικής παλινδρόμησης ακολουθούν, λίγο – πολύ, τις ίδιες γενικές αρχές που χρησιμοποιούνται και στη γραμμική παλινδρόμηση. Έτσι, οι τεχνικές που χρησιμοποιούνται στη γραμμική παλινδρόμηση θα μας βοηθήσουν να προσεγγίσουμε τη λογιστική παλινδρόμηση. Παρακάτω θα διευκρινίσουμε τις ομοιότητες και τις διαφορές της λογιστικής και της γραμμικής παλινδρόμησης με ένα παράδειγμα.

Παράδειγμα 1.

Προκειμένου να μελετηθούν οι παράγοντες κίνδυνου καρδιακής νόσου έχει καταμετρηθεί η ηλικία (AGE) και η ένδειξη παρουσίας ή απουσίας στεφανιαίας νόσου (CHD) σε 100 άτομα. Επίσης τα άτομα χωρίστηκαν σε ομάδες ανάλογα με την ηλικία τους (AGEGRP). Η εξαρτημένη μεταβλητή είναι η CHD, η οποία κωδικοποιείται με την τιμή «0» αν έχουμε απουσία στεφανιαίας νόσου στο άτομα και με την τιμή «1» αν έχουμε παρουσία. Γενικά θα μπορούσαμε να χρησιμοποιήσουμε οποιεσδήποτε δύο τιμές, αλλά για ευκολία χρησιμοποιούμε το μηδέν και τη μονάδα.

Είναι ενδιαφέρον να διερευνηθεί η σχέση μεταξύ της ηλικίας και της απουσίας ή παρουσίας στεφανιαίας νόσου σε αυτήν την ομάδα. Στην περίπτωση που η μεταβλητή απόκρισης ήταν συνεχής και όχι δυαδική, κατά πάσα πιθανότητα θα ήταν απαραίτητο να δημιουργήσουμε το διάγραμμα διασποράς του αποτελέσματος σε σχέση με την ανεξάρτητη μεταβλητή. Στο παράδειγμά μας το αντίστοιχο διάγραμμα θα εμφάνιζε δύο παράλληλες γραμμές στις τιμές του μηδενός και του ένα, το οποίο όμως δεν θα παρείχε μια σαφή εικόνα για τη σχέση της στεφανιαίας νόσου και της ηλικίας. Το πρόβλημα αυτό μπορεί να λυθεί εν μέρει με την ομαδοποίηση της ανεξάρτητης μεταβλητής (στο παράδειγμά μας AGEGRP) και τον υπολογισμό των μέσων τιμών σε κάθε υποομάδα. Χρησιμοποιώντας αυτή τη στρατηγική ομαδοποίησης της ηλικίας, υπολογίζεται για κάθε ηλικιακό γκρουπ η συχνότητα εμφάνισης, καθώς και η επί της εκατό παρουσία της CHD. Η βασική πληροφορία που μας παρέχεται είναι ότι όσο η ηλικία του ατόμου αυξάνεται, αυξάνεται αναλογικά και η πιθανότητα εμφάνισης της CHD. Η πληροφορία όμως αυτή είναι η ίδια που θα μπορούσαμε να προσκομίσουμε και αν πραγματοποιούσαμε μια αντίστοιχη ομαδοποίηση σε μια γραμμική παλινδρόμηση. Θα σημειώσουμε δύο σημαντικές διαφορές.

Η πρώτη διαφορά αφορά τη φύση της σχέσης μεταξύ του αποτελέσματος και των ανεξάρτητων μεταβλητών. Σε κάθε πρόβλημα παλινδρόμησης η βασική ποσότητα είναι η μέση τιμή της μεταβλητής απόκρισης, δεδομένης της αξίας της ανεξάρτητης μεταβλητής. Η ποσότητα αυτή ονομάζεται δεσμευμένη μέση τιμή και εκφράζεται ως E $(Y|x)$, όπου το Y υποδηλώνει την μεταβλητή απόκρισης, ενώ το x μια συγκεκριμένη τιμή της ανεξάρτητης μεταβλητής. Η ποσότητα E $(Y|x)$ διαβάζεται «η αναμενόμενη τιμή του Y, δεδομένης της τιμής του x». Στη γραμμική παλινδρόμηση υποθέτουμε ότι αυτή η μέση τιμή μπορεί να εκφραστεί ως μια γραμμική εξίσωση του x, με τη μορφή E $(Y|x)$ =β_0+$\beta_1 x$.

Η προηγούμενη έκφραση δηλώνει ότι η E $(Y|x)$ μπορεί να πάρει οποιαδήποτε τιμή, καθώς το x κυμαίνεται από το $-\infty$ στο $+\infty$. Σε μια δίτιμη μεταβλητή απόκρισης η δεσμευμένη μέση τιμή θα πρέπει να κυμαίνεται μεταξύ του μηδενός και του ένα, δηλαδή
(0 ≤ E$(Y|x)$ ≤ 1).

Πολλές συναρτήσεις κατανομής έχουν προταθεί ώστε να χρησιμοποιηθούν στην ανάλυση μιας δίτιμης μεταβλητής απόκρισης. Μερικές από αυτές έχουν αναλυθεί από τους Cox και Snell (Cox and Snell, 1989). Υπάρχουν δύο σημαντικοί λόγοι για να επιλέξει κάποιος τη λογιστική κατανομή. Αρχικά, από μαθηματικής απόψεως, είναι μια ευέλικτη και πολύ εύκολη στη χρήση συνάρτηση. Δεύτερον, το παραμετρικό μοντέλο της παρέχει τη βάση για σημαντικές εκτιμήσεις του αποτελέσματος.

Προκειμένου να απλοποιήσουμε τους τύπους, θα χρησιμοποιήσουμε την ισότητα $p(x)$ = E $(Y|x)$, για να αναπαραστήσουμε την δεσμευμένη μέση τιμή του Y δεδομένου του x, όταν χρησιμοποιείται η λογιστική κατανομή. Πιο συγκεκριμένα, το μοντέλο της λογιστικής παλινδρόμησης εκφράζεται μέσω της σχέσης:

$$p(x)= \frac{e^{\beta_0+\beta_1 x}}{1+ e^{\beta_0+\beta_1 x}} \qquad (1.1)$$

από την οποία είναι φανερό ότι ισχύει ο περιορισμός $0 \leq p(x) \leq 1$. Ένας μετασχηματισμός, που είναι καίριος για τη μελέτη μας πάνω στη λογιστική παλινδρόμηση είναι ο logit μετασχηματισμός. Ο μετασχηματισμός αυτός ορίζεται ως :

$$g(x) = \ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$

Η σημασία αυτού του μετασχηματισμού είναι ότι το $g(x)$ έχει πολλές από τις επιθυμητές ιδιότητες ενός μοντέλου γραμμικής παλινδρόμησης. Η λογαριθμική, $g(x)$

- Είναι γραμμική στις παραμέτρους της
- Μπορεί να είναι συνεχής
- Μπορεί να κυμαίνεται από -∞ έως +∞, ανάλογα με το εύρος του x.

Η δεύτερη σημαντική διαφορά μεταξύ των μοντέλων της λογιστικής και της γραμμικής παλινδρόμησης αφορά την δεσμευμένη κατανομή της μεταβλητής απόκρισης. Στο γραμμικό μοντέλο υποθέτουμε ότι η μεταβλητή απόκρισης μπορεί να εκφραστεί και ως $y = E(Y|x) + \varepsilon$. Η ποσότητα ε ονομάζεται *σφάλμα* και εκφράζει την απόκλιση της παρατήρησης από τον μέσο. Συνήθως, δεχόμαστε ότι το ε ακολουθεί κανονική κατανομή με μέση τιμή μηδέν και κάποια διακύμανση (διασπορά), που είναι σταθερή μεταξύ των επιπέδων της ανεξάρτητης μεταβλητής. Αυτό δεν συμβαίνει στην περίπτωση δίτιμης μεταβλητής απόκρισης. Σε αυτή την περίπτωση μπορούμε να εκφράσουμε την αξία της μεταβλητής ως $y = p(x) + \varepsilon$. Εδώ η ποσότητα ε μπορεί να πάρει δύο πιθανές τιμές. Αν $y = 1$ τότε $\varepsilon = 1 - p(x)$ με πιθανότητα $p(x)$, ενώ αν $y = 0$ τότε $\varepsilon = -p(x)$ με πιθανότητα $1 - p(x)$. Επομένως, το ε ακολουθεί κατανομή με μέση τιμή μηδέν και διακύμανση $p(x)\cdot[1 - p(x)]$. Συνεπώς, η μεταβλητή απόκρισης ακολουθεί διωνυμική κατανομή με πιθανότητα που εξαρτάται από το μέσο, $p(x)$. (Hosmer et al., 2013)

1.1.1 Η προσαρμογή του διωνυμικού μοντέλου

Ας υποθέσουμε ότι έχουμε ένα δείγμα n ανεξάρτητων παρατηρήσεων της μορφής (x_i, y_i), $i = 1, 2, \ldots, n$, όπου το y_i δηλώνει την τιμή μιας δίτιμης μεταβλητής απόκρισης και το x_i την τιμή της ανεξάρτητης μεταβλητής για την i-στη παρατήρηση. Επιπλέον, υποθέτουμε ότι η μεταβλητή απόκρισης έχει κωδικοποιηθεί με 0 και 1, που αντιπροσωπεύουν την απουσία ή την παρουσία του ενδεχομένου αντίστοιχα. Στο παράδειγμά μας,

$$Y = \begin{cases} 1 & \textit{αν το άτομο διαγνώστηκε με στεφανιαία νόσο} \\ 0 & \textit{διαφορετικά} \end{cases}$$

Η προσαρμογή του μοντέλου της λογιστικής παλινδρόμησης, στην εξίσωση (1.1), για ένα σύνολο δεδομένων απαιτεί να εκτιμήσουμε τις τιμές των άγνωστων παραμέτρων, β_0 και β_1. Στη γραμμική παλινδρόμηση, η μέθοδος που χρησιμοποιείται πιο συχνά για τον υπολογισμό άγνωστων παραμέτρων είναι η μέθοδος ελαχίστων τετραγώνων. Δυστυχώς, η

μέθοδος αυτή δεν μπορεί να εφαρμοστεί όταν πρόκειται για δίτιμη μεταβλητή απόκρισης. Η γενική μέθοδος εκτίμησης για το μοντέλο της γραμμικής παλινδρόμησης, ονομάζεται μέθοδος μέγιστης πιθανοφάνειας. Αυτή η μέθοδος θα μας δώσει και τη βάση για να προσεγγίσουμε την εκτίμηση στο μοντέλο της λογιστικής παλινδρόμησης. Σε γενικές γραμμές, η μέθοδος μέγιστης πιθανοφάνειας αποδίδει τιμές για τις άγνωστες παραμέτρους, οι οποίες μεγιστοποιούν την πιθανότητα να εξασφαλίσουμε τα δεδομένα των παρατηρήσεων. Προκειμένου να εφαρμοστεί αυτή η μέθοδος πρέπει πρώτα να κατασκευάσουμε μια συνάρτηση, που ονομάζεται συνάρτηση πιθανοφάνειας. Αυτή η συνάρτηση εκφράζει την πιθανότητα των παρατηρήσεων (δεδομένων) ως συνάρτηση των αγνώστων παραμέτρων. Οι εκτιμήτριες μέγιστης πιθανοφάνειας των παραμέτρων μεγιστοποιούν αυτή τη συνάρτηση. Δηλαδή οι εκτιμήτριες που προκύπτουν είναι αυτές που είναι περισσότερο σύμφωνες με τα παρατηρούμενα δεδομένα. Παρακάτω θα περιγράψουμε την μέθοδο εύρεσης των εκτιμητριών αυτών για το μοντέλο λογιστικής παλινδρόμησης.

Αν το Y είναι κωδικοποιημένο ως 0 και 1 τότε η εξίσωση (1.1) παρέχει την δεσμευμένη πιθανότητα το Y να είναι ίσο με 1, δεδομένου του x. Η πιθανότητα αυτή δηλώνεται ως $p(x)$ και είναι η πιθανότητα επιτυχίας της στατιστικής μονάδας i. Συνεπώς, η ποσότητα $1 - p(x)$ δίνει τη δεσμευμένη πιθανότητα το Y να είναι ίσο με 0, δεδομένου του x. Ένας βολικός τρόπος για να εκφράσουμε την συνάρτηση πιθανοφάνειας για τα ζεύγη (x_i, y_i), για κάθε παρατήρηση i, είναι ο τύπος

$$p(x_i)^{y_i} \cdot [1 - p(x_i)]^{1-y_i} \tag{1.2}$$

Καθώς οι παρατηρήσεις θεωρούνται ανεξάρτητες, η συνάρτηση πιθανοφάνειας L, που προκύπτει μέσω της σχέσης (1.2) γράφεται ως εξής:

$$\mathrm{L}(\beta) = \prod_{i=1}^{n} p(x_i)^{y_i} \cdot [1 - p(x_i)]^{1-y_i} \tag{1.3}$$

Η αρχή της μέγιστης πιθανοφάνειας δηλώνει ότι η εκτιμήτρια του β θα είναι εκείνη που μεγιστοποιεί την εξίσωση (1.3). Ωστόσο, από μαθηματικής απόψεως, είναι ευκολότερο να εργαστεί κανείς με τον λογάριθμο της έκφρασης (1.3). Έτσι η καινούρια έκφραση ορίζεται ως:

$$l = \ln[L(\beta)] = \sum_{i=1}^{n} \{y_i \cdot \ln p(x_i) + (1 - y_i) \cdot \ln[1 - p(x_i)]\} \tag{1.4}$$

Για να βρούμε την τιμή των β που μεγιστοποιεί το L(β), παραγωγίζουμε το L(β) ως προς β και ορίζουμε την έκφραση που προκύπτει ίση με μηδέν. Οι εξισώσεις αυτές, γνωστές και ως εξισώσεις πιθανοφάνειας είναι

$$\sum [y_i - p(x_i)] = 0 \tag{1.5}$$

και

$$\sum x_i [y_i - p(x_i)] = 0 \tag{1.6}$$

Στα αθροίσματα (1.5) και (1.6) είναι λογικό ότι το i κυμαίνεται από 1 έως n. Η αρίθμηση του i παραλείπεται όπου δεν είναι απαραίτητη.

Στη γραμμική παλινδρόμηση, οι εξισώσεις πιθανοφάνειας, που προκύπτουν από την παραγώγιση ως προς β, είναι γραμμικές και μπορούν εύκολα να λυθούν ως προς την άγνωστη παράμετρο. Στη λογιστική παλινδρόμηση έχουμε τις εξισώσεις (1.5) και (1.6), οι οποίες είναι μη γραμμικές στα β_0 και β_1 και απαιτούν ειδικές μεθόδους για την λύση τους. Οι μέθοδοι αυτοί έχουν επαναληπτικό χαρακτήρα και γι' αυτόν το λόγο έχουν δημιουργηθεί λογισμικά λογιστικής παλινδρόμησης (software).

Η τιμή του β που δίνεται από τη λύση στις εξισώσεις (1.5) και (1.6) καλείται εκτιμήτρια μέγιστης πιθανοφάνειας και συμβολίζεται με $\hat{\beta}$. Γενικά η χρήση του συμβόλου «^» δηλώνει την εκτιμήτρια μέγιστης πιθανοφάνειας της ποσότητας που εκτιμάμε. Για παράδειγμα η $\hat{p}(x_i)$ είναι η εκτιμήτρια μέγιστης πιθανοφάνειας της $p(x_i)$. Αυτή η ποσότητα παρέχει μια εκτίμηση της δεσμευμένης πιθανότητας το Y να είναι ίσο με 1, δεδομένου ότι το x είναι ίσο με το x_i. Ως εκ τούτου, αντιπροσωπεύει την προβλεπόμενη τιμή για το μοντέλο της λογιστικής παλινδρόμησης. Μια ενδιαφέρουσα συνέπια της εξίσωσης (1.5)
είναι η παρακάτω

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} p(x_i)$$

Δηλαδή, το άθροισμα των παρατηρούμενων τιμών του y είναι ίσο με το άθροισμα των προβλεπόμενων τιμών.

1.1.2 Εκτίμηση και ερμηνεία παραμέτρων

Μετά την εκτίμηση των συντελεστών, αυτό που κοιτάμε στο προσαρμοσμένο μοντέλο συνήθως αφορά την εκτίμηση της σημασίας των μεταβλητών στο μοντέλο. Αυτό συνήθως περιλαμβάνει τη διαμόρφωση και τον έλεγχο μιας στατιστικής υπόθεσης για τον προσδιορισμό αν οι ανεξάρτητες μεταβλητές στο μοντέλο σχετίζονται "σημαντικά" με την μεταβλητή απόκρισης. Η μέθοδος εκτέλεσης αυτού του ελέγχου είναι πολύ γενική και διαφέρει από μοντέλο σε μοντέλο μόνο σε μικρές λεπτομέρειες. Ξεκινώντας θα συζητήσουμε την γενική προσέγγιση για μια ενιαία ανεξάρτητη μεταβλητή.

Ο έλεγχος της σημασίας του συντελεστή μιας μεταβλητής σε κάθε μοντέλο σχετίζεται με την επόμενη ερώτηση. Το μοντέλο που περιλαμβάνει την εν λόγω μεταβλητή μπορεί να μας δώσει περισσότερες πληροφορίες, για την εξαρτημένη μεταβλητή, από ένα μοντέλο που δεν την περιέχει; Αυτή η ερώτηση απαντάται αν συγκρίνουμε τις παρατηρούμενες τιμές της μεταβλητής απόκρισης με εκείνες που προβλέπονται από τα δύο μοντέλα (το πρώτο με την μεταβλητή και το δεύτερο χωρίς). Η μαθηματική συνάρτηση που χρησιμοποιείται για τη σύγκριση των παρατηρούμενων και των προβλεπόμενων τιμών εξαρτάται από το επόμενο. Εάν οι προβλεπόμενες τιμές είναι καλύτερες ή ακριβέστερες όταν η μεταβλητή βρίσκεται στο μοντέλο, τότε θα πρέπει να θεωρήσουμε ότι η εν λόγω μεταβλητή είναι "σημαντική".

Είναι σημαντικό να σημειώσουμε ότι εμείς δεν εξετάζουμε αν οι προβλεπόμενες τιμές συμπίπτουν απόλυτα με τις παρατηρούμενες τιμές (αυτός είναι ο έλεγχος της καλής προσαρμογής). Αντί αυτού, η ερώτησή μας τίθεται με σχετική έννοια.

Η γενική μέθοδος για την εκτίμηση της σημασίας των μεταβλητών δείχνεται εύκολα για το μοντέλο γραμμικής παλινδρόμησης και η χρήση του αιτιολογεί την προσέγγιση για την λογιστική παλινδρόμηση. Η σύγκριση των δύο προσεγγίσεων υπογραμμίζει τις διαφορές μεταξύ της μοντελοποίησης μια συνεχούς και μιας δίτιμης μεταβλητής απόκρισης.

Στη γραμμική παλινδρόμηση, εξετάζεται η σχέση μεταξύ της εξαρτημένης με τις ανεξάρτητες μεταβλητές. Η διαδικασία αυτή ονομάζεται *ανάλυση διασποράς* και στην ουσία υπολογίζει το αν η μεταβλητότητα των τιμών της εξαρτημένης μεταβλητής εξηγείται από τη μεταβλητότητα των ανεξάρτητων μεταβλητών. Η ανάλυση αυτή χωρίζει το άθροισμα των τετραγώνων των αποκλίσεων (σφαλμάτων) των παρατηρήσεων σε δύο είδη:

1) Το άθροισμα τετραγώνων των σφαλμάτων σε σχέση με την παλινδρομική γραμμή (υπόλοιπα) SSE.
2) Το άθροισμα των τετραγώνων των προβλεπόμενων τιμών, με βάση το μοντέλο παλινδρόμησης SSR.

Στη γραμμική παλινδρόμηση η σύγκριση των παρατηρούμενων και προβλεπόμενων τιμών βασίζεται στο τετράγωνο της απόστασης μεταξύ των δύο. Εάν το y_i υποδηλώνει την παρατηρούμενη τιμή και το $\hat{y}_i$ την προβλεπόμενη τιμή για το i-άτομο στο μοντέλο, τότε το στοιχείο που χρειάζεται για να γίνει αυτή η σύγκριση είναι:

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Σύμφωνα με το μοντέλο που δεν περιέχει την ανεξάρτητη μεταβλητή, η οποία είναι υπό εξέταση, η μόνη παράμετρος είναι η β_0 και η $\hat{\beta}_0 = \bar{y}$ θα είναι ο μέσος όρος της μεταβλητής απόκρισης. Σε αυτή την περίπτωση, $y_i = y$ και το SSE ισούται με το συνολικό άθροισμα των τετραγώνων. Όταν συμπεριλαμβάνουμε την ανεξάρτητη μεταβλητή στο μοντέλο, οποιαδήποτε μείωση στο SSE οφείλεται στο γεγονός ότι ο συντελεστής κλίσης για την ανεξάρτητη μεταβλητή δεν είναι μηδέν. Η μεταβολή της αξίας του SSE οφείλεται στη μεταβλητότητα, όσον αφορά την παλινδρόμηση, δηλαδή στην SSR. Αυτή εκφράζεται ως:

$$\text{SSR} = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Στη γραμμική παλινδρόμηση, το ενδιαφέρον επικεντρώνεται στο μέγεθος της SSR. Μια μεγάλη τιμή της SSR μας δηλώνει ότι η ανεξάρτητη μεταβλητή είναι σημαντική, ενώ μια μικρή τιμή δείχνει ότι η ανεξάρτητη μεταβλητή δεν είναι χρήσιμη στη πρόβλεψη της απόκρισης.

Η ίδια περίπου λογική ισχύει και στη λογιστική παλινδρόμηση: σύγκριση των παρατηρούμενων τιμών της μεταβλητής απόκρισης με τις προβλεπόμενες τιμές που λαμβάνονται από μοντέλα με και χωρίς την εν λόγω μεταβλητή. Στη λογιστική

παλινδρόμηση, η σύγκριση των παρατηρούμενων και προβλεπόμενων τιμών βασίζεται στη συνάρτηση πιθανοφάνειας, που ορίστηκε στη εξίσωση (1.4). Για να κατανοήσουμε καλύτερα αυτή τη σύγκριση, είναι χρήσιμο εννοιολογικά να σκεφτούμε μια παρατηρούμενη τιμή της μεταβλητής απόκρισης καθώς επίσης και μια προβλεπόμενη τιμή της από ένα κορεσμένο μοντέλο. Το κορεσμένο μοντέλο είναι εκείνο που περιέχει τόσες παραμέτρους όσα είναι και τα δεδομένα. Ένα απλό παράδειγμα κορεσμένου μοντέλου είναι το γραμμικό μοντέλο όταν υπάρχουν μόνο δύο σημεία δεδομένων, n=2.

Η σύγκριση των παρατηρούμενων και των προβλεπόμενων τιμών με βάση τη συνάρτηση πιθανοφάνειας, γίνεται με τη χρήση της ακόλουθης έκφρασης:

$$\text{D} = -2\ln\left[\frac{\pi\iota\theta\alpha\nu o\phi\acute{\alpha}\nu\varepsilon\iota\alpha\ \pi\rho o\sigma\alpha\rho\mu o\sigma\mu\acute{\varepsilon}\nu o\upsilon\ \mu o\nu\tau\acute{\varepsilon}\lambda o\upsilon}{\pi\iota\theta\alpha\nu o\phi\acute{\alpha}\nu\varepsilon\iota\alpha\ \kappa o\rho\varepsilon\sigma\mu\acute{\varepsilon}\nu o\upsilon\ \mu o\nu\tau\acute{\varepsilon}\lambda o\upsilon}\right] \quad (1.7)$$

Η ποσότητα μέσα στις μεγάλες αγκύλες στην παραπάνω έκφραση ονομάζεται πιθανότητα αναλογίας. Είναι απαραίτητο να πολλαπλασιάσουμε τον λογάριθμο με το μείον 2, έτσι ώστε να ληφθεί μια ποσότητα της οποίας η κατανομή να είναι γνωστή και που θα μπορεί επομένως να χρησιμοποιηθεί για έλεγχο υποθέσεων. Χρησιμοποιώντας την (1.4) η (1.7) γίνεται:

$$\text{D} = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{p}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{p}_i}{1-y_i}\right)\right] \quad (1.8)$$

Η ελεγχοσυνάρτηση D, στην εξίσωση (1.8) ονομάζεται deviance, και για την λογιστική παλινδρόμηση, παίζει τον ίδιο ρόλο που παίζει το άθροισμα τετραγώνων των σφαλμάτων στη γραμμική παλινδρόμηση. Στην πραγματικότητα, η απόκλιση όπως φαίνεται στην εξίσωση (1.8), όταν υπολογίζεται για γραμμική παλινδρόμηση, είναι ταυτόσημη με την SSE.

Επιπλέον, στην περίπτωση της δίτιμης μεταβλητής απόκρισης η πιθανοφάνεια του κορεσμένου μοντέλου είναι ίση είτε με 1 ή με 0. Ειδικότερα, σύμφωνα με τον ορισμό του κορεσμένου μοντέλου, $\hat{p}_i = y_i$ και η πιθανοφάνεια θα είναι:

$$L(\text{κορεσμένο μοντέλο}) = \prod_{i=1}^{n} {y_i}^{y_i} \times (1-y_i)^{(1-y_i)} = 1.0 \quad (1.9)$$

Έτσι από την εξίσωση (1.7) προκύπτει ότι η απόκλιση είναι:

$$\text{D} = -2\ln(\pi\iota\theta\alpha\nu o\phi\acute{\alpha}\nu\varepsilon\iota\alpha\ \pi\rho o\sigma\alpha\rho\mu o\sigma\mu\acute{\varepsilon}\nu o\upsilon\ \mu o\nu\tau\acute{\varepsilon}\lambda o\upsilon) \quad (1.10)$$

Συγκεκριμένα, για να εκτιμηθεί η σημασία μιας ανεξάρτητης μεταβλητής, συγκρίνουμε την τιμή του D με και χωρίς την ανεξάρτητη μεταβλητή στην εξίσωση. Η αλλαγή του D, λόγο της συμπερίληψης της μεταβλητής στο μοντέλο είναι:

$$G = D(\tau o\ \mu o\nu\tau\acute{\varepsilon}\lambda o\ \chi\omega\rho\acute{\iota}\varsigma\ \tau\eta\nu\ \mu\varepsilon\tau\alpha\beta\lambda\eta\tau\acute{\eta}) - D(\tau o\ \mu o\nu\tau\acute{\varepsilon}\lambda o\ \mu\varepsilon\ \tau\eta\nu\ \mu\varepsilon\tau\alpha\beta\lambda\eta\tau\acute{\eta}).$$

Το στοιχείο G έχει τον ίδιο ρόλο στην λογιστική παλινδρόμηση, όπως και το F-test στην γραμμική. Καθώς η πιθανοφάνει του κεκορεσμένου μοντέλου είναι πάντα κοινή και στις δύο τιμές του D, το G μπορεί να εκφραστεί ως:

$$G = -2\ln\left[\frac{\pi\iota\theta\alpha\nu o\phi\acute{\alpha}\nu\varepsilon\iota\alpha\ \chi\omega\rho\acute{\iota}\varsigma\ \tau\eta\nu\ \mu\varepsilon\tau\alpha\beta\lambda\eta\tau\acute{\eta}}{\pi\iota\theta\alpha\nu o\phi\acute{\alpha}\nu\varepsilon\iota\alpha\ \mu\varepsilon\ \tau\eta\nu\ \mu\varepsilon\tau\alpha\beta\lambda\eta\tau\acute{\eta}}\right]. \quad (1.11)$$

Στην ειδική περίπτωσή που έχουμε μία μόνο ανεξάρτητη μεταβλητή, είναι εύκολο να δείξουμε ότι όταν η μεταβλητή δεν είναι στο μοντέλο η εκτίμηση της μέγιστης πιθανοφάνειας από το β_0 είναι $\ln(n_1/n_0)$, όπου $n_1 = \sum y_i$ και $n_0 = \sum(1 - y_i)$, ενώ η προβλεπόμενη πιθανότητα για όλα τα αντικείμενα είναι σταθερή και ίση με n_1/n. Υπό αυτές τις προϋποθέσεις, η τιμή στο G είναι

$$G = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\hat{p}_i^{\,y_i}(1-\hat{p}_i)^{(1-y_i)}}\right], \quad (1.12)$$

ή

$$G = 2\left\{\sum_{i=1}^{n}[y_i\ln(\hat{p}_i) + (1-y_i)\ln(1-\hat{p}_i)] - [n_1\ln(n_1) + n_0\ln(n_0) - n\ln n]\right\}, \quad (1.13)$$

Κάτω από την υπόθεση ότι το β_1 είναι ίσο με μηδέν, το G ακολουθεί κατανομή X^2 με 1 βαθμό ελευθερίας.

Υπάρχουν δύο ακόμα ισοδύναμοι στατιστικοί έλεγχοι: ο έλεγχος Wald και ο έλεγχος των βαθμολογιών. Οι παραδοχές που απαιτούνται για τον καθένα είναι οι ίδιες με αυτές του ελέγχου πιθανοφάνειας στην εξίσωση (1.13). Μια πιο εμπλουτισμένη ανάλυση, για αυτούς τους τρεις ελέγχους και για τις παραδοχές, μπορούμε να βρούμε από τον Rao (1973). Ο έλεγχος Wald είναι ισοδύναμος με τον έλεγχο μέγιστης πιθανοφάνειας της παραμέτρου $\hat{\beta}_1$, με την εκτίμηση για το τυπικό σφάλμα. Ο τύπος που χρησιμοποιείται για τον έλεγχο Wald είναι

$$W = \frac{\hat{\beta}_1}{\widehat{SE}(\hat{\beta}_1)}.$$

Κάτω από την μηδενική υπόθεση και τις παραδοχές για το μέγεθος του δείγματος, η αναλογία αυτή ακολουθεί προσεγγιστικά την τυπική κανονική κατανομή. Οι εκτιμήσεις των τυπικών σφαλμάτων καθώς και εκείνες των παραμέτρων συνήθως λαμβάνονται από κάποιο λογισμικό του υπολογιστή.

Ορισμένα πακέτα λογισμικού χρησιμοποιούν την ελεγχοσυνάρτηση W^2, το οποίο ακολουθεί την X^2 κατανομή με 1 βαθμό ελευθερίας. Ο έλεγχος Wald συχνά συμπεριφέρεται με παρεκκλίνοντα τρόπο, αποτυγχάνοντας να απορρίψει την μηδενική υπόθεση, ενώ ο συντελεστής ήταν σημαντικός χρησιμοποιώντας τον έλεγχο πιθανοφανειών (Hauck and Donner, 1977). Γι' αυτόν το λόγο συχνά προτιμάται ο έλεγχος μέγιστης πιθανοφάνειας. Παρ' όλα αυτά, γενικά δεν υπάρχουν μεγάλες διαφορές μεταξύ των τιμών του G και του W^2. Στην πράξη η πιο ανησυχητική κατάσταση είναι όταν οι τιμές είναι κοντά και ο ένας έλεγχος δίνει $p < 0.05$, ενώ ό άλλος $p > 0.05$. Όταν συμβεί αυτό συμβουλευόμαστε την p – τιμή που μας δίνει ο έλεγχος πιθανοφανειών.

Το κύριο πλεονέκτημα του ελέγχου βαθμολογιών για τη σημασία μιας μεταβλητής είναι ότι δεν απαιτεί την εκτίμηση του συντελεστή. Η χρήση του ελέγχου αυτού περιορίζεται λόγω του ότι δεν είναι διαθέσιμος σε πολλά πακέτα. Ο έλεγχος βαθμολογιών βασίζεται στη θεωρία διανομής των παραγόντων του λόγου πιθανοφανειών. Στην ουσία πρόκειται για έναν έλεγχο πολλαπλών μεταβλητών που απαιτεί υπολογισμούς πινάκων.

Στην απλοποιημένη περίπτωση, ο έλεγχος αυτός βασίζεται στην υπό όρους κατανομή του παραγώγου της εξίσωσης (1.6), το οποίο (παράγωγο) λαμβάνεται από την εξίσωση (1.5). Στην απλοποιημένη περίπτωση έχουμε την δυνατότητα να δώσουμε μια έκφραση για τον έλεγχο βαθμολογιών. Ο έλεγχος χρησιμοποιεί την τιμή της εξίσωσης (1.6), που υπολογίζεται χρησιμοποιώντας τα $\beta_0 = \ln(n_1/n_0)$ και $\beta_1 = 0$. Όπως έχουμε σημειώσει νωρίτερα, υπό αυτές τις προϋποθέσεις λαμβάνουμε $\hat{p} = n_1/n = \bar{y}$, ενώ η αριστερή πλευρά της εξίσωσης (1.6) γίνεται $\sum x_i(y_i - \bar{y})$. Μπορεί να αποδειχθεί ότι η εκτιμώμενη διακύμανση είναι $\bar{y}(1-\bar{y})\sum(x_i - \bar{x})^2$. Η ελεγχοσυνάρτηση για τις βαθμολογίες (Score test) είναι

$$ST = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sqrt{\bar{y}(1-\bar{y})\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

(Hosmer et al., 2013).

1.2 Πολλαπλό μοντέλο λογιστικής παλινδρόμησης

Στην Παράγραφο 1.1 παρουσιάσαμε το μοντέλο λογιστικής παλινδρόμησης στο πλαίσιο ενός μοντέλου που περιέχει μία μόνο μεταβλητή. Όπως και στην περίπτωση της γραμμικής παλινδρόμησης, η ιδιαιτερότητα του μοντέλου λογιστικής παλινδρόμησης είναι η ικανότητα του να χειρίζεται πολλές μεταβλητές, οι οποίες μπορεί να έχουν και διαφορετικές μονάδες μέτρησης. Προκειμένου να εξετάσουμε το πολλαπλό μοντέλο λογιστικής παλινδρόμησης θα πρέπει να εκτιμήσουμε τις μεταβλητές καθώς και να κάνουμε ορισμένους ελέγχους για την σημαντικότητα τους. Χρησιμοποιούμε και εδώ την ίδια προσέγγιση που συζητήθηκε στο προηγούμενο κεφάλαιο για την προσαρμογή. Σε αυτήν την παράγραφο θα εισαχθεί μια εναλλακτική μέθοδος μοντελοποίησης, η οποία χρησιμοποιεί δείκτριες μεταβλητές για την

προσαρμογή των ανεξάρτητων μεταβλητών κατηγορικής ή ονομαστικής μορφής. Σε όλες τις περιπτώσεις, υποθέτουμε ότι υπάρχει μια προκαθορισμένη συλλογή μεταβλητών προς εξέταση.

1.2.1 Περιγραφή του μοντέλου

Θεωρούμε μια ομάδα από p ανεξάρτητες μεταβλητές καθώς και το διάνυσμα $x^{'} = (x_1, \dots, x_p)$. Προς το παρών υποθέτουμε ότι κάθε μια από αυτές τις μεταβλητές είναι κλιμακωτή. Θεωρούμε και εδώ ότι η δεσμευμένη πιθανότητα του αποτελέσματος δίνεται από τον τύπο P $(Y = 1|x) = p(x)$. Ο μετασχηματισμός logit στο πολλαπλό μοντέλο δίνεται από την εξίσωση

$$g(x) = \ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \qquad (1.14)$$

η οποία μπορεί ισοδύναμα να εκφραστεί ως

$$p(x) = \frac{e^{g(x)}}{1+ e^{g(x)}} \qquad (1.15)$$

Υπάρχουν περιπτώσεις που κάποιες από τις ανεξάρτητες μεταβλητές είναι διακριτές, μεταβλητές ονομαστικής κλίμακας. Τέτοιες μεταβλητές είναι η φυλή, το φύλο, η ομάδα θεραπείας και ούτω καθεξής. Όταν έχουμε τέτοιες μεταβλητές, δεν θα πρέπει φυσικά να τις συγχέουμε με τις συνεχείς (κλίμακα διαστήματος). Οι αριθμοί που χρησιμοποιούνται για να αναπαραστήσουν τις διάφορες τάξεις των μεταβλητών αυτών δεν έχουν καμία αριθμητική σημασία. Σε αυτήν την περίπτωση, η καλύτερη μέθοδος είναι να χρησιμοποιήσουμε δείκτριες ή εικονικές μεταβλητές. Ας υποθέσουμε για παράδειγμα, ότι μια από τις ανεξάρτητες μεταβλητές είναι η φυλή, η οποία έχει κωδικοποιηθεί ως “λευκό”, “μαύρο” και “άλλη”. Στην περίπτωση αυτή χρειάζονται δύο δείκτριες μεταβλητές, D1 και D2. Μια πιθανή κωδικοποίηση είναι ότι όταν ο ερωτώμενος είναι “λευκός” και οι δύο μεταβλητές θα είναι ίσες με το 0, όταν είναι “μαύρος”, η D1 θα είναι ίση με 1, ενώ η D2 ίση με 0 και όταν η απάντηση είναι “άλλη”, το D1 θα είναι ίσο με 0 και το D2 ίσο με 1. Ο Πίνακας 1.1 απεικονίζει την κωδικοποίηση αυτών των εικονικών μεταβλητών.

Πίνακας 1.1: Κωδικοποίηση κατηγορικής μεταβλητής

Φυλή	**D1**	**D2**
Λευκή	0	0
Μαύρη	1	0
Άλλη	0	1

Γενικά, εάν μια κατηγορική μεταβλητή έχει κ πιθανές τιμές, τότε απαιτούνται κ−1 δείκτριες μεταβλητές. Ο λόγος που χρησιμοποιούμε μια λιγότερη εικονική μεταβλητή είναι ότι τα μοντέλα μας θα περιέχουν και ένα σταθερό όρο. Για να κατανοηθεί καλύτερα ο συμβολισμός που χρησιμοποιείται, ας υποθέσουμε ότι j-οστή ανεξάρτητη μεταβλητή είναι η x_j, η οποία έχει κ_j κατηγορίες. Οι κ_j −1 δείκτριες μεταβλητές θα συμβολίζονται ως D_{jl}, ενώ οι συντελεστές ως β_{jl}, όπου l = 1, 2, … , κ_j −1. Έτσι ο μετασχηματισμός logit για ένα μοντέλο με p μεταβλητές, με την j-οστή να είναι κατηγορική, θα είναι

$$g(x) = \beta_0 + \beta_1 x_1 + \ldots + \sum_{l=1}^{\kappa_j - 1} \beta_{jl} D_{jl} + \ldots + \beta_p x_p$$

Με μικρές εξαιρέσεις, μπορούμε να παρακάμψουμε την άθροιση και τους διπλούς δείκτες, που χρησιμοποιούμε εδώ για την κατανόηση του πολλαπλού μοντέλου.

1.2.2 Προσαρμογή του πολλαπλού μοντέλου λογιστικής παλινδρόμησης

Ας υποθέσουμε ότι έχουμε ένα δείγμα n ανεξάρτητων παρατηρήσεων (x_i, y_i), όπου $i = 1, 2, \ldots, n$. Όπως και στην απλή περίπτωση, η προσαρμογή του μοντέλου απαιτεί την εκτίμηση του διανύσματος $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$. Η μέθοδος που χρησιμοποιείται στην περίπτωση του πολλαπλού μοντέλου είναι η ίδια που χρησιμοποιείται και στο απλό, μέθοδος μέγιστης πιθανοφάνειας. Η συνάρτηση πιθανοφάνειας είναι παρόμοια με αυτή που δίνεται στην εξίσωση (1.3), με την μόνη αλλαγή ότι το $p(x)$ ορίζεται τώρα όπως στην εξίσωση (1.14). Με την μέθοδο αυτή θα προκύψουν p+1 εξισώσεις πιθανοφάνειας, οι οποίες μπορούν να εκφραστούν ως εξής:

$$\sum_{i=1}^{n} [y_i - p(x_i)] = 0$$

Και

$$\sum_{i=1}^{n} x_{ij} [y_i - p(x_i)] = 0$$

όπου $j = 1, 2, \ldots, p$.

Όπως και στο απλό μοντέλο, απαιτείται η λύση των παραπάνω εξισώσεων, προκειμένου να υπολογιστεί μια εκτίμηση $\hat{\beta}$ των συντελεστών β. Έτσι οι προσαρμοσμένες τιμές για το πολλαπλό μοντέλο λογιστικής παλινδρόμησης θα είναι οι $\hat{p}(x_i)$ και η τιμή της έκφρασης στην εξίσωση (1.15) υπολογίζεται χρησιμοποιώντας τα $\hat{\beta}$ και τα x_i.

Στην προηγούμενη Παράγραφο έγινε μια σύντομη αναφορά στη μέθοδο εκτίμησης για τα τυπικά σφάλματα των εκτιμούμενων συντελεστών. Τώρα που το μοντέλο της λογιστικής

παλινδρόμησης έχει γενικευθεί στην περίπτωση πολλαπλών μεταβλητών, θα μελετήσουμε λεπτομερέστερα την εκτίμηση των τυπικών σφαλμάτων.

Η μέθοδος που χρησιμοποιείται για την εκτίμηση της διασποράς και τη συνδιακύμανση του εκτιμούμενου συντελεστή είναι η καλά-ανεπτυγμένη θεωρία της εκτίμησης της μέγιστης πιθανοφάνειας (βλ. Rao, 1973). Η θεωρία αυτή, στην ουσία, μας δηλώνει ότι τα τυπικά σφάλματα των εκτιμήσεων προέρχονται από τον πίνακα των δευτέρων μερικών παραγώγων της συνάρτησης πιθανοφάνειας. Αυτοί οι μερικοί παράγωγοι έχουν την ακόλουθη μορφή

$$\frac{\partial^2 l(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 p_i(1-p_i) \qquad (1.16)$$

και

$$\frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_h} = -\sum_{i=1}^{n} x_{ij}\, x_{il} p_i(1-p_i) \qquad (1.17)$$

όπου j,h = 0, 1, 2, …, p και αντί για $p(x_i)$ γράφουμε p_i. Ο πίνακας παρατηρήσεων, που περιέχει τους αρνητικούς όρους των εξισώσεων (1.16) και (1.17), είναι διαστάσεων (p+1)x(p+1) και συμβολίζεται $I(\beta)$. Οι διασπορές και οι συνδιακυμάνσεις λαμβάνονται από τον αντίστροφο πίνακα, ο οποίος συμβολίζεται με $Var(\beta) = I^{-1}(\beta)$. Εκτός από πολύ ειδικές περιπτώσεις, είναι πολύ δύσκολο να γραφεί η ακριβής μορφή των στοιχείων του πίνακα αυτού. Ως εκ τούτου, θα χρησιμοποιήσουμε το συμβολισμό $Var(\beta_j)$ για την διασπορά του $\hat{\beta}_j$, όπου στον πίνακα είναι το j-οστό στοιχείο της διαγωνίου και το $Cov(\beta_j, \beta_h)$ για την συνδιακύμνση των $\hat{\beta}_j$ και $\hat{\beta}_h$, το οποίο είναι ένα τυχαίο σημείο, εκτός διαγωνίου. Οι εκτιμήσεις των διασπορών και των συνδιακυμάνσεων, οι οποίες συμβολίζονται με $\widehat{Var}(\hat{\beta})$, δύνονται από την εκτίμηση του $Var(\beta)$ ως προς τα $\hat{\beta}$. Αντίστοιχα χρησιμοποιούμε τα $\widehat{Var}(\hat{\beta}_j)$ και $\widehat{Cov}(\hat{\beta}_j, \hat{\beta}_h)$ για να δηλώσουμε τα στοιχεία αυτού του πίνακα. Ως επί το πλείστον, χρησιμοποιούμε μόνο τα εκτιμημένα τυπικά σφάλματα των συντελεστών, τα οποία δίνονται από τον παρακάτω τύπο

$$\widehat{SE}(\hat{\beta}_j) = \left[\widehat{Var}(\hat{\beta}_j)\right]^{1/2} \qquad (1.18)$$

όπου $j = 0, 1, 2, \ldots, p$. Ο συμβολισμός αυτός χρησιμοποιείται στην ανάπτυξη μεθόδων για τον έλεγχο των συντελεστών και την εκτίμηση των διαστημάτων εμπιστοσύνης. Ο πίνακας πληροφοριών είναι χρήσιμος όταν μιλάμε για την προσαρμογή του μοντέλου και μια πιθανή τυποποίηση είναι η $\hat{I}(\hat{\beta}) = X'\hat{V}X$, όπου X είναι ένας n επί p+1 [nx(p+1)] πίνακας που περιέχει τα δεδομένα για κάθε υποκείμενο και V είναι ένας n επί n (nxn) διαγώνιος πίνακας με γενικό στοιχείο το $\hat{p}_i(1-\hat{p}_i)$. Έτσι ο πίνακας X έχει την μορφή

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

και ο V έχει την μορφή

$$V = \begin{bmatrix} \hat{p}_1(1-\hat{p}_1) & 0 & \cdots & 0 \\ 0 & \hat{p}_2(1-\hat{p}_2) & \dots & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \dots & 0 & \hat{p}_n(1-\hat{p}_n) \end{bmatrix},$$

όπου $\hat{p}_i = \hat{p}(x_i)$ είναι η τιμή της εξίσωσης (1.15), η οποία υπολογίζεται χρησιμοποιώντας την εκτιμημένη τιμή του β (δηλαδή το $\hat{\beta}$) και τις συνδιαλλαγές των i και x_i .

1.2.3 Έλεγχος της σημαντικότητας του μοντέλου

Αφού καταλήξουμε σε ένα συγκεκριμένο πολλαπλό (πολυμεταβλητό) μοντέλο λογιστικής παλινδρόμησης, θα πρέπει να μπούμε στην διαδικασία αξιολόγησής του. Όπως και στην περίπτωση που παρουσιάστηκε στην Παράγραφο 1.1, το πρώτο βήμα αυτής της διαδικασίας είναι συνήθως η αξιολόγηση της σημασίας των μεταβλητών στο μοντέλο. Ο έλεγχος της πιθανοφάνειας για την ολική σημασία των συντελεστών των ανεξάρτητων μεταβλητών του μοντέλου, εκτελείται ακριβώς με τον ίδιο τρόπο όπως και στην περίπτωση του απλού μοντέλου. Ο έλεγχος βασίζεται στην ελεγχοσυνάρτηση G που δίνεται στην εξίσωση (1.11). Η μόνη διαφορά είναι ότι οι εκτιμημένες τιμές $\hat{p}$ εξαρτώνται από το μοντέλο που περιέχει p+1 εκτιμημένες παραμέτρους, $\widehat{\boldsymbol{\beta}}$. Κάτω από την μηδενική υπόθεση ότι οι p συντελεστές «κλίσης» του μοντέλου είναι ίσοι με το μηδέν, το G ακολουθεί X^2 κατανομή με p βαθμούς ελευθερίας.

Καθώς ο στόχος μας είναι να επιτύχουμε το βέλτιστο μοντέλο, ελαχιστοποιώντας ταυτόχρονα και τον αριθμό των παραμέτρων, το επόμενο λογικό βήμα είναι να προσαρμόσουμε το νέο μοντέλο που περιέχει μόνο τις μεταβλητές που θεωρούμε στατιστικά σημαντικές και να το συγκρίνουμε με εκείνο που περιέχει όλες τις μεταβλητές.
Προτού καταλήξουμε στο συμπέρασμα ότι κάποιοι ή όλοι οι συντελεστές είναι μηδέν, μπορούμε να ερμηνεύσουμε τα αποτελέσματα του ελέγχου Wald.

$$W_j = \frac{\hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)}.$$

Ο αντίστοιχος έλεγχος Wald για το πολυμεταβλητό μοντέλο υπολογίζεται από την σχέση

$$W = \hat{\beta}' \left[\widehat{Var}(\hat{\beta})\right]^{-1} \hat{\beta} = \hat{\beta}' \left(X\widehat{V}X\right)\hat{\beta}$$

το οποίο ακολουθεί X^2 κατανομή με p+1 βαθμούς ελευθερίας, κάτω από την υπόθεση ότι κάθε ένας από τους p+1 συντελεστές είναι ίσος με το μηδέν. Ο έλεγχος Wald για τις πολλαπλές μεταβλητές, που είναι ισοδύναμος με τον έλεγχο της πιθανοφάνεις για τη σημασία του μοντέλου, βασίζεται μόνο στους p συντελεστές κλίσης, ενώ λαμβάνεται

εξαλείφοντας το $\hat{\beta}_0$ από τα $\hat{\beta}$ και την σχετική γραμμή (πρώτη ή τελευταία) και στήλη (πρώτη ή τελευταία) από το $X^{'}\hat{V}X$. Η αξιολόγηση αυτού του ελέγχου απαιτεί ένα επιπλέον βήμα τόσο για την εκτέλεση των πράξεων μεταξύ των διανυσμάτων-πίνακα, όσο και για τον υπολογισμό του $\hat{\beta}$, έτσι δεν υπάρχει κάποιο επιπλέων όφελος από τον έλεγχο πιθανοφανειών για τον προσδιορισμό της σημασίας του μοντέλου.

Ο αντίστοιχος έλεγχος βαθμολογιών, για την σημαντικότητα ενός πολλαπλού μοντέλου, βασίζεται στην κατανομή των p παραγώγων του $l(\beta)$ ως προς β. Η διεξαγωγή αυτού του ελέγχου παρουσιάζει τα ίδια μειονεκτήματα με αυτά του ελέγχου Wald. Προκειμένου να προσδιοριστεί λεπτομερώς θα απαιτούσε την εισαγωγή πρόσθετου συμβολισμού, ο οποίος όμως θα είχε ελάχιστη χρησιμότητα στην υπόλοιπη εργασία. Έτσι ο ενδιαφερόμενος αναγνώστης θα μπορούσε να ανατρέξει στους Cox και Hinkley (1974) ή στον Dobson (2002).

1.3 **Πολυωνυμική λογιστική παλινδρόμηση**

Προκειμένου να ορίσουμε το πολυωνυμικό μοντέλο λογιστικής παλινδρόμησης, θα θέσουμε C τον αριθμό των κατηγοριών της μεταβλητή απόκρισης Y. Για παράδειγμα αν πάρουμε τα δεδομένα των γενικών εκλογών του 2002/3 για το Ηνωμένο Βασίλειο (Jowell et al., 2003), τότε το Y αντιπροσώπευε τα κόμματα που πήραν μέρος (Εργατικό κόμμα, Συντηρητικοί, Φιλελεύθεροι δημοκρατικοί) και το C θα είναι ίσο με 3. Ο αριθμός των κατηγοριών του Y μπορεί να είναι 0, 1, …, C-1, όπου το 0 υποδηλώνει την κατηγορία που θα θεωρηθεί ως κατηγορία αναφοράς. Τόσο η εκχώρηση αριθμών σε κατηγορίες (συμπεριλαμβανομένης και της επιλογής της κατηγορίας αναφοράς), όσο και το σύνολο των αριθμών που χρησιμοποιούνται είναι αυθαίρετα και οποιαδήποτε αλλαγή σε αυτά δεν επηρεάζει και ολόκληρο το μοντέλο. Για παράδειγμα, οι κατηγορίες του Y μπορούν συχνά να κωδικοποιηθούν σε ένα σύνολο δεδομένων ως 1, 2, …, C αντί να ξεκινήσουμε από το 0, χωρίς αυτό να επηρεάσει το μοντέλο μας.

Μια κατάλληλη κατανομή πιθανότητας για τέτοιες κατηγορικές μεταβλητές είναι η πολυωνυμική κατανομή. Οι παράμετροί του είναι οι πιθανότητες κατηγορίας

$$p^{(j)} = P(Y = j)$$

για $j = 0, 1, 2, \ldots, C-1$.

Επομένως στο παράδειγμα των εκλογών, αυτές είναι οι πιθανότητες να ψηφίσουμε το Εργατικό κόμμα, τους Συντηρητικούς ή τους Φιλελεύθερους. Επειδή το άθροισμα όλων των πιθανοτήτων πρέπει να είναι 1, είναι αναγκαίο να γνωρίζουμε μόνο τις C-1 πιθανότητες και με αυτές τις πληροφορίες είμαστε σε θέση να καθορίσουμε τα πάντα. Για παράδειγμα αν έχουμε C = 3, μπορούμε να υπολογίσουμε το $p^{(0)} = 1 - p^{(1)} - p^{(2)}$, αν τα $p^{(1)}$ και $p^{(2)}$ είναι γνωστά. Στην περίπτωση που C = 2, η πολυωνυμική κατανομή γίνεται η διωνυμική,

που έχει αναλυθεί. Εκεί χρησιμοποιούμε απλούστερους συμβολισμούς όπου $p^{(1)} = p$ και $p^{(0)} = 1 - p$. Στην περίπτωση που έχουμε παραπάνω από δύο κατηγορίες είναι χρήσιμο να φανερώνεται η κατηγορία j στον συμβολισμό.

Ας υποθέσουμε ότι έχουμε δεδομένα σε n σύνολα παρατηρήσεων $(Y_i, X_{1i}, \dots, X_{ki})$, $i = 1, \dots, n$, και $k = 1$, …, p όπου το Y είναι μια μεταβλητή απόκρισης με C κατηγορίες και $X_1, \dots, X_k$ είναι επεξηγηματικές μεταβλητές. Το πολυωνυμικό μοντέλο λογιστικής παλινδρόμησης ορίζεται από τις ακόλουθες παραδοχές:

1. Οι παρατηρήσεις Y_i είναι στατιστικά ανεξάρτητες μεταξύ τους.

2. Οι παρατηρήσεις Y_i είναι ένα τυχαίο δείγμα από έναν πληθυσμό όπου το Y_i ακολουθεί πολυωνυμική κατανομή με παραμέτρους πιθανότητας.

3. Ο λόγος των πιθανοτήτων κάθε μη-κατηγορίας αναφοράς $j = 1, \dots, C - 1$, ως προς την πιθανότητα της κατηγορίας αναφοράς 0 εξαρτάται από τις επεξηγηματικές μεταβλητές μέσω της σχέσης

$$\log\left(\frac{p_i^{(j)}}{p_i^{(0)}}\right) = {\beta_0}^{(j)} + \beta_1^{(j)} X_{1i} + \cdots + \beta_k^{(j)} X_{ki} \tag{1.19}$$

για κάθε $j = 1, \dots, C - 1$, όπου ${\beta_0}^{(j)}$ και $\beta_1^{(j)}, \dots, \beta_k^{(j)}$ είναι άγνωστοι παράμετροι του πληθυσμού.

Μπορούμε να δούμε ότι το πολυωνυμικό λογιστικό μοντέλο μοιάζει με ένα σύνολο από $C-1$ διωνυμικά λογιστικά μοντέλα, ένα για κάθε κατηγορία μη αναφοράς της μεταβλητής απόκρισης έναντι της κατηγορίας αναφοράς, με την διαφορά ότι $p_i^{(1)} + p_i^{(0)} \neq 1$. Για παράδειγμα, όταν έχουμε C = 3, η σχέση (1.19) καθορίζει τα δύο υπομοντέλα

$$\log\left(\frac{p_i^{(1)}}{p_i^{(0)}}\right) = a^{(1)} + \beta_1^{(1)} X_{1i} + \cdots + \beta_k^{(1)} X_{ki} \tag{1.20}$$

$$\log\left(\frac{p_i^{(2)}}{p_i^{(0)}}\right) = a^{(2)} + \beta_1^{(2)} X_{1i} + \cdots + \beta_k^{(2)} X_{ki} \tag{1.21}$$

Η προσαρμογή των μοντέλων γίνεται με την μέθοδο μέγιστης πιθανοφάνειας, ενώ οι συντελεστές αυτών ελέγχονται με τις μεθόδους που περιγράφηκαν παραπάνω. Η προδιαγραφή (1.19) προϋποθέτει ότι οι ίδιες επεξηγηματικές μεταβλητές $X_1, \dots, X_\kappa$ συμπεριλαμβάνονται σε κάθε ένα από τα $C-1$ υπομοντέλα.

2 ΑΝΑΛΥΣΗ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ

2.1 Εισαγωγή

Ο κύριος στόχος της ανάλυσης κυρίων συνιστωσών είναι να αντικαταστήσει p μετρικές συσχετισμένες μεταβλητές με λίγες μη συσχετισμένες μεταβλητές που περιέχουν τις περισσότερες από τις πληροφορίες του αρχικού συνόλου. Αυτό απλοποιεί σε μεγάλο βαθμό την εργασία της κατανόησης της δομής των δεδομένων, επειδή είναι πολύ ευκολότερο να ερμηνεύσει κανείς δύο ή τρεις μη συσχετισμένες μεταβλητές παρά 20 ή 30, που εμφανίζουν ένα περίπλοκο μοτίβο αλληλοσυσχετίσεων. Για να μετουσιώσουμε τον στόχο αυτό σε μια πρακτική μέθοδο, πρέπει να εξηγήσουμε με μεγαλύτερη ακρίβεια τι εννοούμε όταν λέμε ότι πρέπει να διατηρηθεί το «μεγαλύτερο μέρος των πληροφοριών».

Η κεντρική ιδέα βασίζεται στην έννοια του ποσοστού της συνολικής διακύμανσης (το άθροισμα των διακυμάνσεων των p αρχικών μεταβλητών) που δικαιολογείται για καθεμία από τις νέες μεταβλητές. Η ανάλυση κυρίων συνιστωσών μετασχηματίζει το σύνολο των συσχετισμένων μεταβλητών $(x_1, \dots, x_p)$ σε ένα σύνολο μη συσχετισμένων μεταβλητών $(y_1, \dots, y_p)$ που ονομάζονται κύριες συνιστώσες, με τέτοιο τρόπο ώστε το y_1 να εξηγεί τη μέγιστη δυνατή συνολική διακύμανση, το y_2 τη μέγιστη δυνατή υπόλοιπη διακύμανση, κ.ο.κ. Το πλήρες σύνολο των p κυρίων συνιστωσών εξηγεί τη συνολική διακύμανση:

$$\sum_{j=1}^{p} var(y_j) = \sum_{i=1}^{p} var(x_i)$$

Ωστόσο, αν αποδειχθεί ότι οι πρώτες λίγες κύριες συνιστώσες δικαιολογούν αρκετά μεγάλο μέρος της συνολικής διακύμανσης, ενώ το μεγαλύτερο μέρος της διακύμανσης των x εξηγείται από τα πρώτα λίγα y, τότε μπορούμε να αγνοήσουμε τις υπόλοιπες κύριες συνιστώσες χωρίς μεγάλη απώλεια πληροφοριών. Είναι συνηθισμένο να τυποποιούμε τα x σε μοναδιαία διακύμανση πριν εφαρμόσουμε την ανάλυση κύριων συνιστωσών, έτσι ώστε κάθε μεταβλητή x να έχει την ίδια συνεισφορά στη συνολική διακύμανση, και άρα

$$\sum_{i=1}^{p} var(x_i) = p \ .$$

Άλλος ένας στόχος της ανάλυσης κυρίων συνιστωσών είναι να ερμηνεύσει την υποκείμενη δομή των δεδομένων σε σχέση με τις πιο σημαντικές κύριες συνιστώσες. Πολλές φορές, μπορούμε να αναγνωρίσουμε τις κύριες συνιστώσες με κάποια ποσότητα που έχει ουσιαστικό ενδιαφέρον. Για παράδειγμα, βλέπουμε συχνά ότι η πρώτη κύρια συνιστώσα συσχετίζεται θετικά με κάθε ένα από τα x, και έτσι μπορεί να ερμηνευθεί ως ένας σταθμισμένος μέσος όλων των μεταβλητών. Ας υποθέσουμε ότι έχουμε μετρήσεις για το ύψος, το μέγεθος ποδιού, το άνοιγμα χεριών, το βάρος, τη μέση και τους γοφούς ενήλικων ανδρών. Επειδή αυτές οι έξι μεταβλητές συσχετίζονται όλες θετικά μεταξύ τους, η πρώτη συνιστώσα, που είναι ένα μέτρο μεγέθους, θα συσχετίζεται θετικά με όλες τους. Μερικές

φορές, οι συνιστώσες διαχωρίζουν ένα υποσύνολο των μεταβλητών από κάποιο άλλο. Μπορούμε να ερμηνεύσουμε τέτοιες κύριες συνιστώσες αν σκεφτούμε τα κοινά χαρακτηριστικά που έχει κάθε υποσύνολο των μεταβλητών. Έτσι, για τις μετρήσεις σώματος που αναφέραμε προηγουμένως, η δεύτερη συνιστώσα θα μπορούσε να δείχνει την αντίθεση μεταξύ του υποσυνόλου (ύψος, μέγεθος ποδιού, άνοιγμα χεριών) και του υποσυνόλου (βάρος, μέση, γοφοί). Η πρώτη συνιστώσα διαχωρίζει τους άντρες ανάλογα με το μέγεθος, με τους μικρόσωμους άνδρες στο ένα άκρο και τους μεγαλόσωμους στο άλλο άκρο της κλίμακας. Η δεύτερη συνιστώσα περιέχει άνδρες που είναι σχετικά υπέρβαροι για το ύψος τους στο ένα άκρο και άνδρες που είναι σχετικά λεπτοί στο άλλο.

2.2 **Μερικές πιθανές εφαρμογές**

1. Ας υποθέσουμε ότι έχουμε βαθμολογίες εξετάσεων σε διάφορα μαθήματα για ένα σύνολο ατόμων. Θέλουμε να τις συνδυάσουμε με τέτοιο τρόπο ώστε να πάρουμε ένα συνολικό μέτρο της ακαδημαϊκής ικανότητας. Με μια ανάλυση κύριων συνιστωσών των δεδομένων, βρίσκουμε ότι η πρώτη συνιστώσα συσχετίζεται θετικά με όλες τις βαθμολογίες. Ερμηνεύουμε αυτή τη συνιστώσα ως ένα μέτρο της γενικής ικανότητας. Ωστόσο, η δεύτερη συνιστώσα εξηγεί επίσης ένα μεγάλο ποσοστό της διακύμανσης στις βαθμολογίες των εξετάσεων και διαχωρίζει τα μαθήματα των θετικών επιστημών από εκείνα των ανθρωπιστικών επιστημών. Από αυτό συμπεραίνουμε ότι η ικανότητα δεν μπορεί να αποτυπωθεί ικανοποιητικά με μία μόνο μεταβλητή. Υπάρχει διαφορά μεταξύ της ικανότητας στα μαθήματα των θετικών και των ανθρωπιστικών επιστημών.

2. Ας υποθέσουμε ότι ενδιαφερόμαστε να δημιουργήσουμε ένα μέτρο στέρησης. Έχουμε αρκετούς δείκτες που θα μπορούσαν να θεωρηθούν μέτρα στέρησης. Καθένας από αυτούς μετράει κάποια διαφορετική πτυχή της στέρησης, και θέλουμε με κάποιο τρόπο να εξαγάγουμε ό,τι είναι κοινό μεταξύ τους για να φτάσουμε στην καρδιά της έννοιας. Τι πρέπει να κάνουμε; Μία προσέγγιση θα ήταν να εφαρμόσουμε μα ανάλυση κύριων συνιστωσών στο σύνολο των δεικτών που θα παίξουν τον ρόλο των μεταβλητών x. Αν η πρώτη συνιστώσα εξηγεί ένα μεγάλο ποσοστό της συνολικής διακύμανσης των αρχικών μέτρων στέρησης, θα μπορούσαμε να την χρησιμοποιήσουμε στη θέση των αρχικών μεταβλητών ως μέτρο στέρησης.

3. Μπορεί να θέλουμε να απλοποιήσουμε τη δομή των δεδομένων μας πριν προχωρήσουμε σε περαιτέρω αναλύσεις – όπως η ανάλυση συστάδων ή η πολλαπλή παλινδρόμηση (multiple regression) – μειώνοντας τις πολλές μεταβλητές, που εμφανίζουν έντονη συσχέτιση, σε λίγες ανεξάρτητες κύριες συνιστώσες (θυσιάζοντας όσο το δυνατόν λιγότερη πληροφορία).

2.3 Γενική περιγραφή της ανάλυσης κυρίων συνιστωσών

Η ανάλυση κυρίων συνιστωσών μετασχηματίζει ένα σύνολο συσχετισμένων μεταβλητών (x) σε ένα σύνολο μη συσχετισμένων συνιστωσών (y). Οι κύριες συνιστώσες είναι γραμμικοί συνδυασμοί των x, τους οποίους γράφουμε ως εξής:

$$\begin{aligned} y_1 &= \alpha_{11}x_1 + \alpha_{21}x_2 + \cdots + \alpha_{p1}x_p \\ y_2 &= \alpha_{12}x_1 + \alpha_{22}x_2 + \cdots + \alpha_{p2}x_p \\ &\vdots \\ y_p &= \alpha_{1p}x_1 + \alpha_{2p}x_2 + \cdots + \alpha_{pp}x_p \end{aligned}$$

Κάθε συνιστώσα είναι ένα σταθμισμένο άθροισμα των x, όπου τα α_{ij} είναι οι *συντελεστές στάθμισης* (weights) ή *συντελεστές* (coefficients).

Είναι σαφές ότι πρέπει να υπάρχουν κάποιοι περιορισμοί στα α_{ij}. Διαφορετικά, θα μπορούσαμε να μεγαλώσουμε τη διακύμανση οποιουδήποτε y απλώς μεγαλώνοντας αρκετά τα α_{ij}. Για να δούμε τι απαιτείται, πρέπει να επιστρέψουμε στην περίπτωση των δύο μεταβλητών. Καταλήξαμε στις κύριες συνιστώσες περιστρέφοντας τους άξονες ενώ ταυτόχρονα τους διατηρούσαμε ορθογώνιους μεταξύ τους. Στη γενική περίπτωση, πρέπει να βρούμε τον ισοδύναμο αλγεβρικό τύπο για την ορθογώνια περιστροφή. Αποδεικνύεται ότι αυτό απαιτεί να ικανοποιούν τα α_{ij} τις παρακάτω συνθήκες:

$$\sum_{i=1}^{p} a_{ij}^2 = 1 \qquad (j = 1,2,\ldots,p),$$

$$\sum_{i=1}^{p} a_{ij}\, a_{ik} = 0 \qquad (j \neq k,\ j=1,\ldots,p,\ k=1,\ldots,p)$$

Ένας άλλος τρόπος να περιγράψουμε το αποτέλεσμα των παραπάνω συνθηκών είναι να πούμε ότι δεν αλλάζουν τις σχετικές θέσεις ή τη διάταξη των σημείων.

Μια σημαντική συνέπεια που έχει η συνθήκη της ορθογωνικότητας είναι ότι η συνολική διακύμανση των y είναι ίση με τη συνολική διακύμανση των x, δηλαδή

$$\sum_{j=1}^{p} var(y_i) = \sum_{i=1}^{p} var(x_i)$$

Αυτό σημαίνει ότι η συνολική διακύμανση δεν αλλάζει, αντί γι' αυτό, η διακύμανση ανακατανέμεται μεταξύ των μεταβλητών. Τα y υπολογίζονται με φθίνουσα σειρά σπουδαιότητας, έτσι ώστε το y_1 να έχει τη μέγιστη διακύμανση και άρα να εξηγεί το μεγαλύτερο ποσοστό της συνολικής διακύμανσης. Επομένως, η πρώτη κύρια συνιστώσα μπορεί να θεωρηθεί ως η καλύτερη μονοδιάστατη σύνοψη των δεδομένων. Η δεύτερη συνιστώσα, το y_2, υπολογίζεται έτσι ώστε να έχει τη δεύτερη μεγαλύτερη διακύμανση, με βάση τους περιορισμούς

$$\sum_{i=1}^{p} a_{i2}^2 = 1 \quad \text{και} \quad \sum_{i=1}^{p} a_{i1}a_{i2} = 0,$$

για να είναι ορθογώνια (μη συσχετισμένη) με το y_1. Οι πρώτες δύο συνιστώσες προσφέρουν την καλύτερη διδιάστατη σύνοψη των δεδομένων. Οι επόμενες συνιστώσες υπολογίζονται με σειρά φθίνουσας διακύμανσης, και κάθε συνιστώσα δεν είναι συσχετισμένη με τις προηγούμενες.

Άρα, το μαθηματικό πρόβλημα που θέτει αυτή η διαδικασία είναι η εύρεση μιας μεθόδου προσδιορισμού των α_{ij} με τέτοιο τρόπο ώστε οι συνιστώσες να έχουν τις απαιτούμενες ιδιότητες. Παρόλο που αυτό φαίνεται να είναι ένα αξεπέραστο πρόβλημα, λύνεται εύκολα επειδή είναι ισοδύναμο με ένα πολύ γνωστό πρόβλημα της άλγεβρας μητρών που ασχολείται με την εύρεση των αποκαλούμενων ιδιοτιμών και ιδιοδιανυσμάτων ενός πίνακα – ο πίνακας στην περίπτωση αυτή είναι είτε ο πίνακας συνδιακύμανσης είτε, πιο συχνά, ο πίνακας συσχέτισης.

Ενδεικτικά θα αναφέρουμε ότι το διάνυσμα $\boldsymbol{x} \in K^n - \{0\}$ ονομάζεται ιδιοδιάνυσμα ενός πίνακα A, αν ικανοποιεί την εξίσωση $A\boldsymbol{x} = \lambda\boldsymbol{x}$, για κάποιο $\lambda \in K$, το οποίο τότε ονομάζεται ιδιοτιμή του πίνακα A, όπου $K = \mathbb{R}$ ή $\mathbb{C}$.

Υπάρχουν p τυπικοί αλγόριθμοι που προσδιορίζουν τους συντελεστές στάθμισης α_{ij} και τις διακυμάνσεις των κυρίων συνιστωσών. Οι τελευταίες συνήθως συμβολίζονται με $(\lambda_1, \lambda_2, \dots, \lambda_p)$ και παρατίθενται από τη μεγαλύτερη προς τη μικρότερη.

2.3.1 Επιλογή πλήθους κύριων συνιστωσών

Όπως συμβαίνει στην ανάλυση συσχέτισης και στην πολυδιάστατη προσαρμογή κλίμακας, ένας από τους στόχους της ανάλυσης κύριων συνιστωσών είναι να είμαστε σε θέση να σχεδιάσουμε ένα γράφημα των δεδομένων σε μια, δύο ή τρεις διαστάσεις χωρίς να χάσουμε πολλές πληροφορίες. Όταν επιλέγουμε το πλήθος των συνιστωσών, σκοπός μας είναι να διατηρήσουμε ένα όσο το δυνατόν μικρότερο σύνολο, αλλά, ταυτόχρονα, να έχουμε ένα ικανοποιητικό πλήθος (συνιστωσών) που να επιτρέπει μια καλή αναπαράσταση των αρχικών δεδομένων. Η διακύμανση της συνιστώσας j είναι η ιδιοτιμή λ_i. Επειδή οι συνιστώσες υπολογίζονται με βάση τη σειρά διακύμανσης, ισχύει ότι $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Αν οι μεταβλητές x είναι τυποποιημένες έτσι ώστε να αναλύεται η μήτρα συσχέτισης, το άθροισμα των διακυμάνσεων των x θα είναι ίσο με p. Άρα το άθροισμα των ιδιοτιμών, δηλαδή η συνολική διακύμανση των y, θα είναι επίσης ίση με p.

Το ποσοστό της συνολικής διακύμανσης που εξηγείται από τη συνιστώσα j είναι

$$\frac{\lambda_j}{\lambda_1+\lambda_2+\cdots+\lambda_p} .$$

Το ποσοστό που εξηγείται από τις πρώτες k συνιστώσες μαζί είναι

$$\frac{\lambda_1+\lambda_2+\cdots+\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_p} .$$

Στην πράξη, τα ποσοστά αυτά εκφράζονται συχνά με τη μορφή ποσοστών επί τοις εκατό.

Υπάρχουν αρκετά κριτήρια που μπορούμε να χρησιμοποιήσουμε για να αποφασίσουμε πόσες συνιστώσες πρέπει να διατηρήσουμε:

i. Διατηρούμε τις πρώτες k συνιστώσες που εξηγούν ένα «μεγάλο» ποσοστό της συνολικής διακύμανσης, έστω 70-80%.
ii. Αν η μήτρα συσχέτισης αναλύεται, διατηρούμε μόνο τις συνιστώσες με ιδιοτιμές μεγαλύτερες από 1. Η λογική πίσω από αυτόν τον εμπειρικό κανόνα είναι ότι μια συνιστώσα με ιδιοτιμή 1 εξηγεί την ίδια ποσότητα διακύμανσης με μια από τις αρχικές μεταβλητές x. Ωστόσο, ο Jolliffe (1972) προτείνει ότι είναι καλύτερο να διατηρούμε συνιστώσες με ιδιοτιμή μεγαλύτερη από 0,7 σε σύγκριση με την αποκοπή στην τιμή 1.
iii. Εξετάζουμε ένα γράφημα παραγόντων (screeplot). Αυτό είναι ένα γράφημα των ιδιοτιμών ως προς το πλήθος των συνιστωσών. Η ιδέα είναι να αναζητήσουμε την «καμπή» που αντιστοιχεί στο σημείο μετά από το οποίο οι ιδιοτιμές μειώνονται με πιο αργό ρυθμό. Η προσθήκη συνιστωσών μετά από το συγκεκριμένο σημείο εξηγεί ένα σχετικά μικρό επιπλέον ποσοστό της διακύμανσης.
iv. Εξετάζουμε αν η συνιστώσα έχει μια λογική και χρήσιμη ερμηνεία.

2.3.2 Ερμηνεία

Ο συντελεστής στάθμισης που δίνεται στη μεταβλητή i στη συνιστώσα j είναι το α_{ij}. Τα σχετικά μεγέθη των α_{ij} αντανακλούν τη σχετική συνεισφορά κάθε μεταβλητής στη συνιστώσα. Για να ερμηνεύσουμε μια συνιστώσα, εξετάζουμε το μοτίβο των τιμών των α_{ij} για την συγκεκριμένη συνιστώσα. Συχνά, προσαρμόζουμε την κλίμακα των συντελεστών έτσι ώστε οι συντελεστές των πιο σημαντικών συνιστωσών (δηλαδή αυτών που εξηγούν το μεγαλύτερο ποσοστό της διακύμανσης) να είναι μεγαλύτεροι από αυτούς που αντιστοιχούν σε λιγότερο σημαντικές συνιστώσες. Αυτοί οι προσαρμοσμένοι (ως προς την κλίμακα) συντελεστές, που ονομάζονται *επιβαρύνσεις συνιστωσών,* είναι οι συντελεστές για την ανακατασκευή των x από τα y και υπολογίζεται ως εξής:

$$a_{ij}^{*} = \sqrt{\lambda_i} a_{ij} \qquad , (i = 1,\ldots,p \ , j = 1,\ldots,p).$$

Όταν αναλύσουμε τη μήτρα συσχετίσεων των x, θα μπορέσουμε να ερμηνεύσουμε το a_{ij}^{*} ως τον συντελεστή συσχέτισης μεταξύ της μεταβλητής i και της συνιστώσας j. Αυτό είναι ιδιαίτερα χρήσιμο για την ερμηνεία.

2.4 Βαθμολογίες συνιστωσών

Ας υποθέσουμε ότι θέλουμε να υπολογίσουμε τη βαθμολογία ενός ατόμου σε μια συγκεκριμένη συνιστώσα από την ανάλυση κύριων συνιστωσών των τυποποιημένων δεδομένων. Έχουμε

$$y_j = a_{1j}x_1 + a_{2j}x_2 + \cdots + a_{pj}x_p \ ,$$

όπου οι μεταβλητές $x_1,\ x_2, \ldots, x_p$ είναι όλες τυποποιημένες με μέση τιμή 0 και διακύμανση 1, ενώ το y_j έχει διακύμανση λ_j.

Ωστόσο είναι πιο συνηθισμένο να τυποποιούμε τις βαθμολογίες των συνιστωσών με μοναδιαία διακύμανση, έτσι ώστε το $\tilde{y}_j = y_j / \sqrt{\lambda_j}$ να έχει διακύμανση ίση με 1. Άρα,

$$\tilde{y}_j = \tilde{a}_{1j}x_1 + \tilde{a}_{2j}x_2 + \cdots + \tilde{a}_{pj}x_p \ ,$$

όπου

$$\tilde{\alpha}_{ij} = \frac{a_{ij}}{\sqrt{\lambda_j}} = \frac{a^*_{ij}}{\lambda_j}$$

Τα $\tilde{\alpha}_{ij}$ ονομάζονται *συντελεστές βαθμολογίας συνιστωσών*.

2.5 Αντικατάσταση των αρχικών μεταβλητών με τις βαθμολογίες των κύριων συνιστωσών

Μία χρήση της ανάλυσης κύριων συνιστωσών είναι η αντικατάσταση ενός μεγαλύτερου συνόλου p μεταβλητών από ένα μικρότερο σύνολο q κύριων συνιστωσών. Η πρώτη συνιστώσα y_1 μπορεί να χρησιμοποιηθεί μόνη της ως μια μονομεταβλητή (q=1) σύνοψη των αρχικών μεταβλητών $x_1, \ldots, x_p$, είτε για χρήση σε περαιτέρω αναλύσεις, είτε ως δείκτης. Πράγματι, οι συντελεστές των βαθμολογιών των συνιστωσών χρησιμοποιούνται μερικές φορές στη βαθμολόγηση νέων ατόμων σε έναν τέτοιο δείκτη. Οι πρώτες δύο συνιστώσες μπορούν να χρησιμοποιηθούν για την δημιουργία του γραφήματος των δεδομένων – είτε σε τέτοια κλίμακα ώστε να ισχύει $Var(y_1) = \lambda_1$, $Var(y_2) = \lambda_2$ ή με τις συνιστώσες $\tilde{y}_1$ και $\tilde{y}_2$, τυποποιημένες ώστε να έχουν μοναδιαία διακύμανση.

Το ερώτημα είναι πόσες πληροφορίες χάνονται από την αντικατάσταση των p μεταβλητών x με τις πρώτες q συνιστώσες τους ή, πιο συγκεκριμένα, πόσο καλά μπορεί να ανακατασκευαστεί η μεταβλητή x_i από τα $\tilde{y}_1, \ldots, \tilde{y}_q$ (για $i = 1, \ldots, p$).

Στην Ενότητα 2.5, οι (τυποποιημένες) κύριες συνιστώσες δίνονται ως γραμμικές συναρτήσεις των (τυποποιημένων) αρχικών μεταβλητών.

$$\tilde{y}_j = \tilde{a}_{1j}\, x_1 + \tilde{a}_{2j}\, x_2 + \cdots + \tilde{a}_{pj}\, x_p \qquad (j = 1, \ldots, p),$$

Οι εξισώσεις αυτές μπορούν να αντιστραφούν και να δώσουν

$$x_i = a^*_{i1}\tilde{y}_1 + \cdots + a^*_{ip}\tilde{y}_p \qquad (i = 1, \ldots, p).$$

όπου $a^*_{ij} = \lambda_j \tilde{a}_{ij} = \sqrt{\lambda_j}\, a_{ij}$ είναι η επιβάρυνση της συνιστώσας που παρουσιάσαμε στην Ενότητα 2.2. Υπενθυμίζουμε ότι η επιβάρυνση αυτή είναι η συσχέτιση μεταξύ των x_i και y_i. Τώρα, ας υποθέσουμε ότι προσπαθούμε να ανακατασκευάσουμε τη μεταβλητή x_i χρησιμοποιώντας μόνο τις πρώτες δύο συνιστώσες. Η ανακατασκευασμένη τιμή είναι

$$\hat{x}_i = a^*_{i1}\tilde{y}_1 + a^*_{i2}\tilde{y}_2 \,.$$

Θα είναι κοντά στο x_i , αν οι υπόλοιπες συσχετίσεις ή επιβαρύνσεις $a^*_{i3}, \ldots, a^*_{ip}$ έχουν τιμή κοντά στο μηδέν. Ισοδύναμα, μπορούμε να κρίνουμε πόσο καλά αναπαράγεται κάθε x_i από τις πρώτες q συνιστώσες ελέγχοντας πόσο κοντά στην τιμή 1 βρίσκεται η *κοινοτικότητα* (communality), όπου η κοινοτικότητα είναι το άθροισμα των τετραγώνων των πρώτων q επιβαρύνσεων, έτσι ώστε η κοινοτικότητα για τη μεταβλητή x_i να είναι ίση με

$$a^{*2}_{i1} + \cdots + a^{*2}_{iq} \qquad (i = 1, \ldots, p).$$

Είναι το τετράγωνο του συντελεστή πολλαπλής συσχέτισης μεταξύ των μεταβλητών x_i και των $y_1, \ldots, y_q$. (Bartholomew et al., 2008)

3 ΑΝΑΛΥΣΗ ΣΥΣΤΑΔΩΝ

3.1 Τι είναι η ανάλυση συστάδων

Η ανάλυση συστάδων είναι η τεχνική κατά την οποία οι παρατηρήσεις συνδυάζονται σε ομάδες ή συστάδες έτσι ώστε:

1. Κάθε ομάδα να είναι ομοιογενής ή συμπαγής διατηρώντας ορισμένα χαρακτηριστικά. Αυτό σημαίνει ότι οι παρατηρήσεις σε κάθε ομάδα είναι παρόμοιες μεταξύ τους.
2. Κάθε ομάδα θα πρέπει να διαφέρει από όλες τις υπόλοιπες όσον αφορά τα κοινά χαρακτηριστικά. Αυτό σημαίνει ότι οι παρατηρήσεις μιας ομάδας διαφέρουν από τις παρατηρήσεις των υπολοίπων.

Ο ορισμός της ομοιότητας ή της ομοιογένειας διαφέρει από ανάλυση σε ανάλυση και εξαρτάται από το αντικείμενο της μελέτης.

Ας πάρουμε για παράδειγμα μια τράπουλα. Τα 52 φύλλα μπορούν να χωριστούν με διάφορους τρόπους. Ένας τρόπος είναι να βάλουμε τα κόκκινα φύλλα σε μια ομάδα και όλα τα μαύρα στην άλλη. Οι παίκτες του blackjack θα έβαζαν όλες τις φιγούρες σε μια ομάδα και όλα τα υπόλοιπα φύλλα σε μια άλλη. Όμοια στο παιχνίδι με τις κούπες έχει νόημα να γίνει η εξής ομαδοποίηση: (1) όλες οι κούπες, (2) η ντάμα μπαστούνι, (3) οι υπόλοιπες κάρτες. Είναι φανερό ότι μπορούμε να δημιουργήσουμε πολλά διαφορετικά «σχήματα» ομάδων, που το καθένα εξαρτάται από τον στόχο και τον σκοπό του παιχνιδιού.

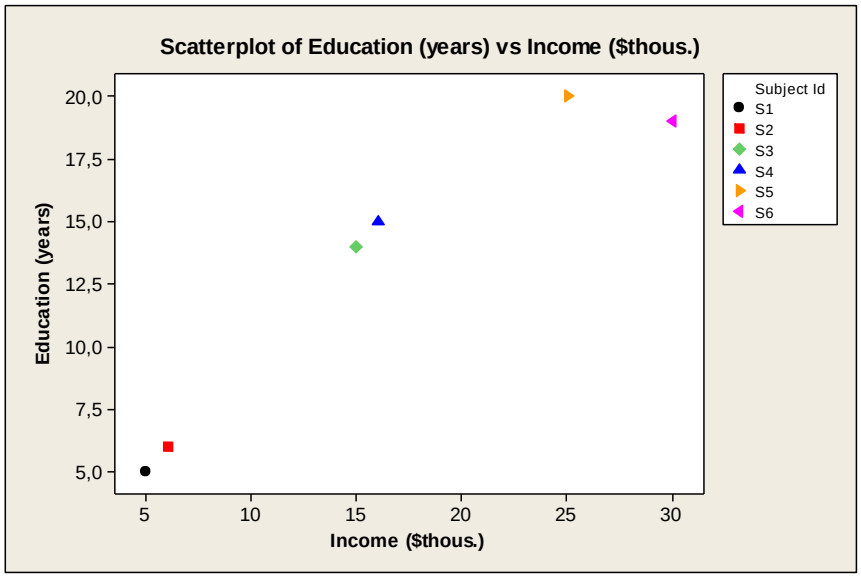
3.2 Γεωμετρική περιγραφή της ανάλυσης συστάδων

Γεωμετρικά, η ιδέα της ανάλυσης συστάδων είναι πολύ απλή. Θα πάρουμε ως παράδειγμα τα δεδομένα που δίνονται στον Πίνακα 3.1. Ο πίνακας περιέχει το εισόδημα και την εκπαίδευση σε χρόνια έξι υποθετικών ατόμων.

Πίνακας 3.1 Υποθετικά δεδομένα

Υποκείμενα ID	Εισόδημα ($ χιλιάδες)	Εκπαίδευση (χρόνια)
S1	5	5
S2	6	6
S3	15	14
S4	16	15
S5	25	20
S6	30	19

Όπως φαίνεται στην Εικόνα 3.1, κάθε παρατήρηση μπορεί να εκπροσωπηθεί από ένα σημείο στον δυσδιάστατο χώρο. Γενικά, κάθε παρατήρηση μπορεί να αναπαρασταθεί ως ένα σημείο στον p-διάστατο χώρο, όπου p είναι ο αριθμός των μεταβλητών ή των χαρακτηριστικών που χρησιμοποιούνται για να περιγράψουν το αντικείμενο/υποκείμενο. Τώρα ας υποθέσουμε ότι εμείς θέλουμε να σχηματίσουμε τρεις ομοιογενείς ομάδες. Εξετάζοντας τον πίνακα παρατηρούμε ότι το S1 και το S2 μπορούν να σχηματίσουν μια ομάδα, το S3 και το S4 μια άλλη και τα υποκείμενα S5 και S6 θα ενωθούν σε μια τρίτη ομάδα.



Εικόνα 3.1 : Γράφημα υποθετικών δεδομένων.

Όπως μπορούμε να δούμε, η ανάλυση συστάδων ομαδοποιεί τις παρατηρήσεις, έτσι ώστε οι παρατηρήσεις σε κάθε ομάδα να είναι παρόμοιες σε ότι αφορά τις μεταβλητές της ομαδοποίησης. Είναι επίσης πιθανό να ομαδοποιήσουμε τις μεταβλητές, έτσι ώστε οι μεταβλητές σε κάθε ομάδα να ομοιάζουν σε σχέση με τις παρατηρήσεις που περιγράφουν. Γεωμετρικά, αυτό είναι ισοδύναμο με το να αναπαραστήσουμε τα δεδομένα σε έναν n-διάστατο χώρο παρατηρήσεων, και θεωρώντας ομάδες μεταβλητών. Ο σκοπός της ανάλυσης συστάδων φαίνεται παρόμοιος με αυτόν της ανάλυσης παραγόντων.

Ανακαλώντας αυτό στην ανάλυση παραγόντων προσπαθούμε να αναγνωρίσουμε συστάδες μεταβλητών έτσι ώστε οι μεταβλητές σε κάθε συστάδα να έχουν κάτι κοινό, για παράδειγμα να μετρούν τον ίδιο λανθάνον παράγοντα. Είναι επομένως δυνατό να χρησιμοποιούμε την ανάλυση παραγόντων για την ομαδοποίηση των παρατηρήσεων και την ανάλυση συστάδων για την ομαδοποίηση των μεταβλητών. Η τεχνική της ανάλυσης συστάδων που συνηθίζεται στην ομαδοποίηση των παρατηρήσεων είναι γνωστή ως ανάλυση Q-παραγόντων. Παρ' όλα αυτά δεν προτείνεται η χρήση της ανάλυσης Q-παραγόντων για την ομαδοποίηση παρατηρήσεων καθώς παρουσιάζει πρόσθετα προβλήματα. Συμφωνούμε με την φιλοσοφία ότι : (1) εάν κάποιος ενδιαφέρεται για τον εντοπισμό λανθάνων παραγόντων και των δεικτών τους, τότε θα πρέπει να χρησιμοποιήσει την ανάλυση παραγόντων, αφού η τεχνική αυτή έχει αναπτυχθεί για αυτόν τον σκοπό και (2) εάν κάποιος ενδιαφέρεται να ομαδοποιήσει τις παρατηρήσεις, τότε θα ήταν προτιμότερο να χρησιμοποιήσει ανάλυση συστάδων, καθώς αντίστοιχα με πριν, η τεχνική αυτή έχει εξελιχθεί ειδικά γι' αυτόν τον σκοπό.

Στην περίπτωση πολλών παρατηρήσεων ή όταν έχουμε περισσότερες από τρεις μεταβλητές/χαρακτηριστικά είναι πιθανό οι γραφικές παραστάσεις να μην βοηθάνε στον εντοπισμό των συστάδων. Σε αυτήν την περίπτωση, θα μπορούσαμε να κάνουμε μια αναλυτική τεχνική για τον εντοπισμό των συστάδων ή ομάδων σημείων σε έναν δοσμένο διαστατό χώρο.

3.3 **Σκοπός της ανάλυσης συστάδων**

Ο σκοπός της ανάλυσης συστάδων είναι η ομαδοποίηση των παρατηρήσεων σε συστάδες, έτσι ώστε κάθε συστάδα να είναι όσο γίνεται πιο ομοιογενής όσο αναφορά τις εκάστοτε μεταβλητές. Το πρώτο βήμα στην ανάλυση συστάδων είναι η επιλογή του μέτρου ομοιότητας. Έπειτα, θα πρέπει να επιλέξουμε την τεχνική ομαδοποίησης που θα χρησιμοποιήσουμε (π.χ. ιεραρχικά ή μη ιεραρχικά). Τρίτο βήμα είναι η επιλογή της μεθόδου ομαδοποίησης για την επιλεγμένη τεχνική (π.χ. μέθοδος κέντρου βάρους στην ιεραρχική τεχνική ομαδοποίησης). Τέλος, το τέταρτο βήμα είναι η απόφαση σχετικά με τον αριθμό των συστάδων που θα δημιουργηθούν.

3.4 **Μέτρα ομοιότητας**

Στην γεωμετρική προσέγγιση της ομαδοποίησης, συνδυάσαμε οπτικά τα αντικείμενα S1 και S2 σε μια ομάδα ή συστάδα, καθώς αυτά τα δύο αντικείμενα φαίνεται να είναι κοντά το ένα στο άλλο στον δισδιάστατο χώρο. Με άλλα λόγια, χρησιμοποιήσαμε έμμεσα την απόσταση μεταξύ των δύο σημείων ως μέτρο ομοιότητας. Μπορούν να χρησιμοποιηθούν διάφορα

μέτρα ομοιότητας. Ως εκ τούτου, ένα από τα θέματα που αντιμετωπίζει ο ερευνητής είναι η επιλογή του καταλληλότερου μέτρου ομοιότητας. Προς το παρόν, ας υποθέσουμε ότι έχουμε επιλέξει ως μέτρο ομοιότητας την ευκλείδεια τετραγωνική απόσταση μεταξύ δύο σημείων. Η ευκλείδεια τετραγωνική απόσταση μεταξύ των στοιχείων S1 και S2 δίνεται από τον τύπο

$$D_{12}^2 = (5-6)^2 + (5-6)^2 = 2$$

Όπου D_{12}^2 είναι η ευκλείδεια τετραγωνική απόσταση μεταξύ των στοιχείων S1 και S2. Όσο μεγαλύτερη είναι η ομοιότητα των στοιχείων, τόσο μικρότερη θα είναι η απόσταση μεταξύ αυτών και αντίστροφα. Ο τύπος για τον υπολογισμό των ευκλείδειων τετραγωνικών αποστάσεων για p μεταβλητές είναι

$$D_{ij}^2 = \sum_{k=1}^{p}(x_{ik} - x_{jk})^2 , \qquad (3.1)$$

όπου D_{ij}^2 είναι η τετραγωνική απόσταση μεταξύ των στοιχείων i και j, x_{ik} η τιμή της k μεταβλητής για το i-οστό στοιχείο, x_{jk} η τιμή της k μεταβλητής του j-οστού στοιχείου, ενώ p είναι το πλήθος των μεταβλητών. Ο Πίνακας 3.2 δίνει τις ομοιότητες, όπως μετρήθηκαν από τις ευκλείδειες τετραγωνικές αποστάσεις, μεταξύ των έξι στοιχείων.

Πίνακας 3.2 : Ομοιότητες ευκλείδειων αποστάσεων.

	S1	S2	S3	S4	S5	S6
S1	0.00	2.00	181.00	221.00	625.00	821.00
S2	2.00	0.00	145.00	181.00	557.00	745.00
S3	181.00	145.00	0.00	2.00	136.00	250.00
S4	221.00	181.00	2.00	0.00	106.00	212.00
S5	625.00	557.00	136.00	106.00	0.00	26.00
S6	821.00	745.00	250.00	212.00	26.00	0.00

Όλοι οι αλγόριθμοι της ανάλυσης συστάδων απαιτούν έναν τύπο μέτρου προκειμένου να αξιολογήσουν την ομοιότητα ενός ζευγαριού παρατηρήσεων ή συστάδων. Τα μέτρα ομοιότητας μπορούν να κατηγοριοποιηθούν στους εξής τρεις τύπους : (1) μέτρα απόστασης, (2) συντελεστές σχέσης και (3) συντελεστής συσχέτισης. Παρακάτω θα αναλύσουμε αυτούς τους τρεις τύπους.

3.4.1 Μέτρα απόστασης

Θα ελέγξουμε εν συντομία την χρησιμότητα των μέτρων απόστασης στην ανάλυση συστάδων. Γενικά, η ευκλείδεια απόσταση μεταξύ των σημείων i και j στις p διαστάσεις δίνεται από τον τύπο

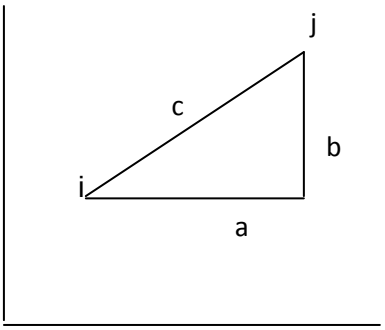
$$D_{ij} = \left(\sum_{k=1}^{p}(x_{ik} - x_{jk})^2\right)^{1/2},$$

όπου D_{ij} είναι η απόσταση μεταξύ των παρατηρήσεων i και j, ενώ p το πλήθος των μεταβλητών. Η ευκλείδεια απόσταση είναι ειδική περίπτωση μίας πιο γενικής μετρικής, της μετρικής Minkowski και δίνεται από τον τύπο

$$D_{ij} = \left(\sum_{k=1}^{p}\left(|x_{ik} - x_{jk}|\right)^n\right)^{1/n}, \quad (3.4)$$

όπου D_{ij} είναι η απόσταση Minkowski μεταξύ των παρατηρήσεων i και j, ενώ p είναι το πλήθος των μεταβλητών και n = 1, 2, …, +∞. Όπως είδαμε και προηγουμένως, τοποθετώντας στο n την τιμή 2 παίρνουμε την ευκλείδεια απόσταση και με την τιμή 1 τα αποτελέσματα που παίρνουμε ονομάζονται αποστάσεις Manhattan (ή οικοδομικό τετράγωνο). Για παράδειγμα στην Εικόνα 3.2 η απόσταση Manhattan μεταξύ των σημείων i και j δίνεται από το a+b. Όπως προϊδεάζει το όνομα «οικοδομικό τετράγωνο», η απόσταση αυτή είναι ο δρόμος που χρειάζεται κάποιος να πάρει φυσιολογικά σε μια πόλη αν θέλει να πάει από το σημείο i στο σημείο j. Γενικά, η απόσταση οικοδομικού τετραγώνου δίνεται από τον τύπο

$$D_{ij} = \sum_{k=1}^{p}|x_{ik} - x_{jk}|.$$



Εικόνα 3.2 : Απόσταση Manhattan

Για άλλους τύπους αποστάσεων παίρνουμε αποτέλεσμα με την κατάλληλη επιλογή του n, παρ' όλα αυτά δεν χρησιμοποιούνται συχνά. Όπως αναφέραμε, η ευκλείδεια απόσταση είναι το πιο κοινό μέτρο ομοιότητας, παρ' όλα αυτά η κλίμακα δεν είναι αμετάβλητη. Αυτό σημαίνει ότι οι αποστάσεις μεταξύ των παρατηρήσεων μπορούν να αλλάξουν αν αλλάξει και η κλίμακα. Για παράδειγμα, θα πάρουμε τα δεδομένα του Πίνακα 3.1. Ας υποθέσουμε ότι τα έσοδα δίνονται σε δολάρια, αντί για χιλιάδες δολάρια. Η τετραγωνική ευκλείδεια απόσταση μεταξύ των παρατηρήσεων 1 και 2 είναι

$$D_{12}^2 = (5000 - 6000)^2 + (5 - 6)^2 = 100000 + 1 = 100001.$$

Όπως βλέπουμε, η μεταβλητή των εσόδων επηρεάζει τον υπολογισμό της απόστασης. Έτσι, η κλίμακα χρησιμοποιείται συχνότερα στον υπολογισμό παρατηρήσεων που έχουν

σημαντική επίδραση στην απόσταση. Είναι σημαντικό να έχουμε μεταβλητές σε παρόμοια κλίμακα μέτρησης. Παρ' όλα αυτά, αν δεν μπορούν να μετρηθούν σε μια παρεμφερή κλίμακα, μπορούμε να χρησιμοποιήσουμε την στατιστική απόσταση, η οποία έχει το πλεονέκτημα ότι η κλίμακα μεταβάλλεται.

Ευκλείδεια Απόσταση για Τυποποιημένα Δεδομένα

Ας υπολογίσουμε την τετραγωνική ευκλείδεια απόσταση μεταξύ των παρατηρήσεων S1 και S2 για τα υποθετικά δεδομένα στον Πίνακα 3.1, αφού πρώτα έχουν τυποποιηθεί. Έτσι έχουμε,

$$SD_{12}^2 = \left[\left(\frac{5-16.167}{9.988}\right)+\left(\frac{6-16.167}{9.988}\right)\right]^2 + \left[\left(\frac{5-13.167}{6.369}\right)+\left(\frac{6-13.167}{6.369}\right)\right]^2$$

$$= \left(\frac{5-6}{9.988}\right)^2 + \left(\frac{5-6}{6.369}\right)^2$$

$$= 0.010 + 0.025 = 0.035$$

Σε αντίθεση με τα μη τυποποιημένα δεδομένα, αφού κάθε μεταβλητή x_i αντικαθιστάται από τα x_i/s_i, η τετραγωνική ευκλείδεια απόσταση για τα τυποποιημένα δεδομένα υπολογίζεται από το $1/\hat{s}_i{}^2$, όπου $\hat{s}_i$ είναι η τυπική απόκλιση της μεταβλητής i. Με άλλα λόγια, μια μεταβλητή με μεγάλη διακύμανση έχει μικρότερη βαρύτητα από μια μεταβλητή με μικρότερη διακύμανση. Αυτό σημαίνει ότι όσο μεγαλύτερη η διακύμανση τόσο μικρότερη η βαρύτητα και αντίστροφα. Τότε η κρίσιμη ερώτηση είναι : γιατί η διακύμανση θα πρέπει να είναι ένας παράγοντας για τον προσδιορισμό της σημασίας μιας δεδομένης μεταβλητής για τον προσδιορισμό της ευκλείδειας απόστασης; Κάποιος μπορεί να χρησιμοποιήσει τυποποιημένα δεδομένα, εάν αυτό είναι απαραίτητο. Αυτό που είναι σημαντικό είναι ότι η τυποποίηση επηρεάζει το αποτέλεσμα των συστάδων.

Μία χρήσιμη ιδιότητα της ευκλείδειας απόστασης για τα τυποποιημένα δεδομένα είναι ότι η κλίμακα είναι αμετάβλητη. Για παράδειγμα, ας υποθέσουμε ότι αλλάζουμε την κλίμακα για τα έσοδα σε δολάρια, αντί για χιλιάδες δολάρια. Μία αλλαγή στην κλίμακα αλλάζει επίσης και την κανονική απόκλιση της μεταβλητής των εσόδων σε 9.988×1000. Η ευκλείδεια απόσταση μεταξύ των παρατηρήσεων 1 και 2 για τα τυποποιημένα δεδομένα, όταν η κλίμακα της μεταβλητής των εσόδων έχει αλλάξει, είναι

$$SD_{12}^2 = \left(\frac{5000-6000}{9.988\times 1000}\right)^2 + \left(\frac{5-6}{6.369}\right)^2$$

$$= \left(\frac{5-6}{9.988}\right)^2 + \left(\frac{5-6}{6.369}\right)^2$$

$$= 0.010 + 0.025 = 0.035,$$

το οποίο είναι ίδιο με πριν.

Απόσταση Mahalanobis

Το δεύτερο μέτρο των στατιστικών αποστάσεων είναι η Mahalanobis απόσταση. Είναι σχεδιασμένη να λαμβάνει υπόψη της την συσχέτιση μεταξύ των μεταβλητών, καθώς είναι επίσης και αμετάβλητη κλίμακα. Για ασυσχέτιστες μεταβλητές η απόσταση Mahalanobis υποβαθμίζεται στην ευκλείδεια απόσταση για μη τυποποιημένα δεδομένα. Η ευκλείδεια απόσταση για τυποποιημένα δεδομένα είναι μία ειδική περίπτωση της απόστασής Mahalanobis. Στην περίπτωση δύο μεταβλητών, η απόσταση Mahalanobis μεταξύ των παρατηρήσεων i και j δίνεται από τον τύπο

$$MD_{ij}^2 = \frac{1}{1-r^2}\left[\frac{\left(x_{i1}-x_{j1}\right)^2}{{s_1}^2} + \frac{\left(x_{i2}-x_{j2}\right)^2}{{s_2}^2} - \frac{2r\left(x_{i1}-x_{j1}\right)\left(x_{i2}-x_{j2}\right)}{s_1 s_2}\right]$$

όπου ${s_1}^2$, ${s_2}^2$ είναι οι διακυμάνσεις των μεταβλητών 1 και 2 αντίστοιχα και r είναι ο συντελεστής συσχέτισης των δύο μεταβλητών. Είναι φανερό ότι όταν οι μεταβλητές είναι ασυσχέτιστες (r=0) η απόσταση μετατρέπεται σε στατιστική απόσταση, ενώ όταν οι διακυμάνσεις των μεταβλητών είναι ίσες με 1 και οι μεταβλητές είναι ασυσχέτιστες τότε η απόσταση Mahalanobis μετατρέπεται σε ευκλείδεια απόσταση. Έτσι η ευκλείδεια και η στατιστική απόσταση είναι ειδικές περιπτώσεις της απόστασης Mahalanobis. Τα πακέτα SAS και SPSS δεν έχουν την δυνατότητα να χρησιμοποιήσουν την απόσταση Mahalanobis.

3.4.2 Συντελεστές Σχέσης

Αυτός ο τύπος μέτρου χρησιμοποιείται συνήθως για δίτιμες μεταβλητές. Στην περίπτωση των δυαδικών δεδομένων, κάποιος μπορεί να χρησιμοποιήσει αυτό το μέτρο είτε ως πολυμορφική συσχέτιση, είτε ως απλό συντελεστή ισοδυναμίας ή θα μπορούσε οι διακυμάνσεις να αντιπροσωπεύουν την ομοιότητα μεταξύ των παρατηρήσεων. Με βάση τα προηγούμενα, ο 2×2 πίνακας δύο δυαδικών μεταβλητών είναι:

	1	0
1	a	b
0	c	d

όπου τα a,b,c, και d είναι οι συχνότητες των συμβάντων. Η ομοιότητα μεταξύ δύο μεταβλητών δίνεται από τον τύπο

$$\frac{a+d}{a+b+c+d}$$

Υπάρχουν και άλλες παραλλαγές του μέτρου αυτού, όπως για παράδειγμα ο συντελεστής Jaccard. Για περισσότερες λεπτομέρειες μπορεί κάποιος να ανατρέξει στο εγχειρίδιο των Sneath και Sokal (1973) και του Hartigan (1975).

3.4.3 Συντελεστής Συσχέτισης

Ένα διαφορετικό μέτρο ομοιότητας είναι ο συντελεστής συσχέτισης των ροπών Pearson. Για την ακρίβεια, οι συντελεστές συσχέτισης και οι συντελεστές σχέσης θα λέγαμε ότι είναι μέτρα ομοιότητας, όπου μια υψηλή τιμή αναπαριστά ομοιότητα και αντίστροφα. Οι συντελεστές συσχέτισης μπορούν εύκολα να μετατραπούν σε μέτρα ομοιότητας υπολογίζοντας την διαφορά τους από την μονάδα, ωστόσο , δεν ικανοποιούνται όλες οι ιδιότητες μιας πραγματικής μετρικής.

Θα πρέπει να σημειωθεί ότι αυτά δεν είναι τα μόνα μέτρα που μπορούμε να χρησιμοποιήσουμε για την ομαδοποίηση των παρατηρήσεων. Γενικά, μπορούμε να χρησιμοποιήσουμε οποιοδήποτε μέτρο ομοιότητας μεταξύ δύο αντικειμένων, αρκεί να έχει νόημα στον ερευνητή. Για παράδειγμα, στον εντοπισμό των τραπεζικών καταστημάτων, ένα σημαντικό μέτρο ομοιότητας μεταξύ δύο πιθανών τοποθεσιών θα μπορούσε να είναι ο χρόνος οδήγησης και όχι η απόσταση σε μίλια ή χιλιόμετρα. Ακόμα, στην περίπτωση της έρευνας που σχετίζεται με εικόνα, οι αντιληπτές αποστάσεις ή οι ομοιότητες ίσως έχουν μεγαλύτερο νόημα από τις ευκλείδειες αποστάσεις. Κλείνοντας, μπορούμε να επιλέξουμε ένα μέτρο ομοιότητας, το οποίο να είναι σε συμφωνία με το αντικείμενο μελέτης. Αρκεί να πούμε βέβαια ότι διαφορετικά μέτρα απόστασης θα οδηγήσουν και σε διαφορετικά αποτελέσματα για την ρύθμιση των ομάδων.

3.5 Αξιοπιστία και εξωτερική εγκυρότητα των συστάδων

Η ανάλυση συστάδων είναι ευρετική τεχνική, έτσι είναι πιθανό να βρεθεί μια λύση ομαδοποίησης ακόμα και όταν δεν υπάρχει καμία φυσική ομάδα στα δεδομένα μας. Γι’ αυτόν τον λόγο η εξακρίβωση της αξιοπιστίας και της εξωτερικής ισχύς της λύσης είναι απαραίτητη.

3.5.1 Αξιοπιστία

Η αξιοπιστία μπορεί να διαπιστωθεί με την διαδικασία διασταυρωμένης επικύρωσης που προτάθηκε από τους McIntyre και Blashfield (1980). Αρχικά τα δεδομένα μας χωρίζονται στα δύο. Πραγματοποιούμε την ανάλυση συστάδων αρχικά στο πρώτο μισό δείγμα και διευκρινίζουμε τα κέντρα βάρους των συστάδων. Έπειτα, αντιστοιχίζουμε τις παρατηρήσεις του δεύτερου μισού του δείγματος στα κέντρα βάρους των συστάδων που είχαν την μικρότερη ευκλείδεια απόσταση. Το επίπεδο αξιοπιστίας εξαρτάται από τον βαθμό συμφωνίας μεταξύ των αναγραφόμενων παρατηρήσεων και της ανάλυσης συστάδων. Η διαδικασία μπορεί να επαναληφθεί πραγματοποιώντας ανάλυση συστάδων για το δεύτερο μισό του δείγματος και αντιστοιχίζοντας τις παρατηρήσεις του πρώτου.

3.5.2 Εξωτερική Εγκυρότητα

Η εξωτερική εγκυρότητα αποκτάται συγκρίνοντας τα αποτελέσματα της ανάλυσης συστάδων με τα εξωτερικά κριτήρια. Για παράδειγμα, ας υποθέσουμε ότι ομαδοποιούμε εταιρίες με βάση χρηματοοικονομικούς δείκτες και έτσι καταλήγουμε σε δύο συστάδες : τις εταιρίες που είναι οικονομικά υγιείς και εκείνες που δεν είναι. Τότε η εξωτερική εγκυρότητα μπορεί να εξακριβωθεί αν σχετίσουμε τα αποτελέσματα της ανάλυσης συστάδων με την ταξινόμηση που λαμβάνουμε από ανεξάρτητους εκτιμητές (π.χ. ακροατές, οικονομικοί αναλυτές, χρηματοοικονομικούς αναλυτές κλπ). Για μεγαλύτερη εμβάθυνση πάνω στο θέμα της ανάλυσης συστάδων, ο ενδιαφερόμενος μπορεί να ανατρέξει στο βιβλίο του Subhash Sharma (1996).

4 ΣΤΑΤΙΣΤΙΚΗ ΜΕΛΕΤΗ ΣΕ ΒΙΟΛΟΓΙΚΑ ΔΕΔΟΜΕΝΑ ΟΙΚΟΣΥΣΤΗΜΑΤΟΣ

4.1 Παρουσίαση δείγματος και μεταβλητών

Το δείγμα μας αποτελείται από 58 πτηνά, από τα οποία κάποια αποτελούν το αρχικό μας δείγμα, ενώ τα υπόλοιπα συλλέχτηκαν από διαφορετικό μέρος προκειμένου να μελετηθούν. Και τα δύο δείγματα συλλέχθηκαν από μέρη της Σκωτίας. Στο αρχικό δείγμα, μας δίνονταν οι διαστάσεις του ράμφους καθώς και το φύλο του πτηνού.

Σκοπός μας είναι η κατασκευή ενός εργαλείου, με το οποίο θα μπορούμε να προσδιορίζουμε το φύλο άλλων πτηνών, μελλοντικά, βασιζόμενοι μόνο στις απλές μετρήσεις του ράμφους και χωρίς περισσότερη εξέταση. Κατά συνέπεια η εξαρτημένη μεταβλητή ως προς την οποία εξετάσαμε το δείγμα μας είναι το φύλο. Η κωδικοποίηση που υιοθετήθηκε φαίνεται στον πίνακα 4.1.

Πίνακας 4.1 : *Κωδικοποίηση εξαρτημένης μεταβλητής.*

Φύλο	Αρσενικό	Θηλυκό
Κωδικοποίηση	1	0

Οι ανεξάρτητες ή επεξηγηματικές μεταβλητές είναι και οι δύο συνεχείς και είναι το μήκος και το πλάτος του ράμφους των πτηνών. Για συντομία θα χρησιμοποιήσουμε παρακάτω τον συμβολισμό

- Μήκος = x_1
- Πλάτος = x_2

Μερικά χρήσιμα χαρακτηριστικά των μεταβλητών αυτών βρίσκονται στον πίνακα 4.2.

Πίνακας 4.2: *Ιδιότητες μεταβλητών.*

Μεταβλητές	**Μέση τιμή**	**Διάμεσος**	**Μέγιστη τιμή**	**Ελάχιστη τιμή**
Μήκος	39.01	38.50	43.80	34.60
Πλάτος	17.11	17.00	19.20	15.00

Ένα μικρό μέρος του δείγματος που χρησιμοποιήθηκε παρουσιάζεται στον Πίνακα 4.3.

Πίνακας 4.3 : Δείγμα

Πτηνό	Δείγμα	Μήκος	Πλάτος	Φύλο
1	1	36.0	15.9	0
2	1	36.3	16.9	0
3	1	37.2	15.6	0
4	1	38.1	16.9	1
5	1	38.4	17.8	1

4.2 Εφαρμογή του μοντέλου λογιστικής παλινδρόμησης

Στο σημείο αυτό πρόκειται να προσαρμόσουμε τα δεδομένα μας με χρήση της λογιστικής παλινδρόμησης του στατιστικού πακέτου της R, με σκοπό την επιλογή του καταλληλότερου στατιστικού μοντέλου για το συγκεκριμένο πρόβλημα. Ειδικότερα στην περίπτωση αυτή έχουμε να κάνουμε με ένα γενικευμένο γραμμικό μοντέλο, με δυαδικά ή διωνυμικά δεδομένα.

Έχοντας ορίσει τις κατηγορικές και επεξηγηματικές μεταβλητές μπορούμε να προχωρήσουμε στη διερεύνηση των παραμέτρων. Προσαρμόζοντας το μοντέλο μας, λαμβάνουμε από την R τα αποτελέσματα στον Πίνακα 4.4:

Πίνακας 4.4 : *Αποτελέσματα*

	Estimate	**Std. Error**	**z-value**	**Pr(> |z|)**
Μήκος	1.636	1.070	1.529	0.1262
Πλάτος	3.636	2.048	1.775	0.0758
Likelihood ratio test = 71.2 on 2 df, AIC = 14.156				

Από τα Αποτελέσματα του προσαρμοσμένου μοντέλου, παρατηρούμε από τις p-τιμές των ελέγχων Wald ότι ενώ το πλάτος του ράμφους φαίνεται να είναι μια αρκετά στατιστικά σημαντική μεταβλητή, δεν μπορούμε να πούμε, με σιγουριά, το ίδιο και για το μήκος, αφού η p-τιμή του είναι ίση με 0.1262.

Σε αυτό το σημείο θα χρειαστεί να ελέγξουμε την σημαντικότητα της μεταβλητής του μήκους και κατά πόσο μας δίνει απαραίτητες πληροφορίες για την κατηγορική μεταβλητή. Για τον σκοπό αυτό θα προσαρμόσουμε ξανά το μοντέλο αφαιρώντας αρχικά την μεταβλητή του μήκους, η οποία είναι και η λιγότερο σημαντική (stepwise) και με την βοήθεια των ελέγχων Deviance και AIC θα κάνουμε την σύγκριση των μοντέλων.

Πριν προχωρήσουμε στα αποτελέσματα, είναι σημαντικό να διευκρινίσουμε ότι λόγω της μεταβλητής του μήκους, δεν περιμένουμε η deviance να μας δώσει αξιόπιστα

αποτελέσματα. Ωστόσο η διαφορά των deviance, που στην ουσία αποτελεί τον έλεγχο του λόγου πιθανοφανειών, παίζει σπουδαίο ρόλο στην κατάληξη των αποτελεσμάτων.

Με τη βοήθεια του στατιστικού πακέτου της R παίρνουμε τα αποτελέσματα που παρουσιάζονται στον Πίνακα 4.5.

Πίνακας 4.5 : *Αποτελέσματα*

Models	**df**	**Resid. Deviance**	**AIC**
Mod1($x_1 + x_2$)		8.1558	14.156
Mod2(x_1)	1	13.9817	17.982
Mod3(x_2)	1	14.5062	18.506

Αρχικά επιλέγουμε να ξεκινήσουμε τη σύγκριση των μοντέλων μας με τη βοήθεια του κριτήριου AIC. Με βάση αυτό το κριτήριο το βέλτιστο μοντέλο είναι αυτό με την μικρότερη τιμή AIC. Από τα αποτελέσματα στον πίνακα 4.5 βλέπουμε ότι η αφαίρεση οποιασδήποτε μεταβλητής είχε ως αποτέλεσμα την αύξηση της τιμής AIC. Ως εκ τούτου, με βάση το κριτήριο AIC, καταλήγουμε στο συμπέρασμα ότι καλύτερο μοντέλο είναι αυτό που περιέχει και τις δύο επεξηγηματικές μεταβλητές.

Αν τώρα βασιστούμε, για τη σύγκριση των μοντέλων, στην ανάλυση της deviance, είμαστε αναγκασμένοι να απορρίψουμε τα δύο πιο απλά μοντέλα (με μια μεταβλητή), καθώς το μοντέλο που περιέχει και τις δύο μεταβλητές (μήκος και πλάτος) είναι στατιστικά καλύτερο, αφού η διαφορά των deviance μειώνεται αισθητά.
Συνεπώς οι δύο έλεγχοι συμφωνούν και έτσι καταλήγουμε στο ότι το μοντέλο με όλες τις μεταβλητές είναι το βέλτιστο.

Οι παραπάνω έλεγχοι μας οδηγούν στην σκέψη ότι οι δύο μεταβλητές ίσως να έχουν κάποια συσχέτιση μεταξύ τους. Υπολογίζοντας λοιπόν τον συντελεστή συσχέτισης (correlation test in R) του πλάτους και του μήκους καταλήγουμε ότι είναι περίπου 0,75. Βλέπουμε λοιπόν ότι οι δύο μεταβλητές είναι όντως αρκετά συσχετισμένες, καθώς ο συντελεστής τους είναι κοντά στη μονάδα.

Η προσαρμοσμένη εξίσωση παλινδρόμησης θα είναι

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = -125.936 + 1.636\, x_1 + 3.636\, x_2 \ ,$$

όπου $\hat{p}$ η εκτιμώμενη τιμή της πιθανότητας αρσενικού φύλου. Η προηγούμενη σχέση μπορεί να γραφεί ισοδύναμα ως

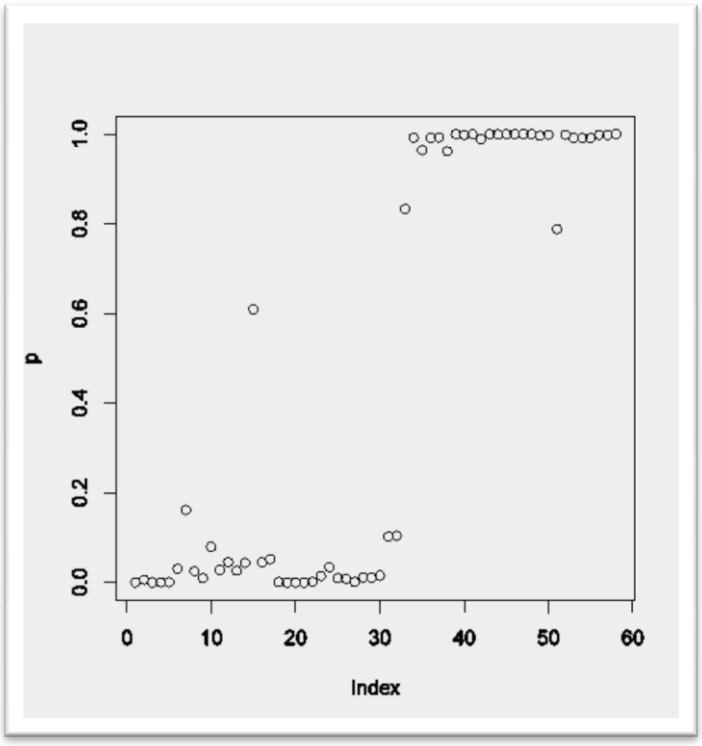
$$\hat{p} = \frac{\exp(-125.936+1.636\ x1+3.636\ x2)}{1+\exp(-125.936+1.636\ x1+3.636\ x2)}$$

Στον πίνακα 4.6 έχουν υπολογιστεί, με τη βοήθεια του στατιστικού πακέτου του Minitab, οι εκτιμημένες πιθανότητες της μεταβλητής απόκρισης, για το αρχικό μας δείγμα.

Πίνακα 4.6 : *Εκτιμημένες πιθανότητες*

Αριθμός δείγματος	$\widehat{\boldsymbol{p}}$	**Y**
1	0.00010	0
2	0.00605	0
3	0.00024	0
4	0.00028	0
5	0.00027	0
6	0.03033	0
7	0.16154	0
8	0.02497	0
9	0.00968	0
10	0.07967	0
11	0.02728	0
12	0.04536	0
13	0.02633	0
14	0.04381	0
15	0.60912	0
16	0.04457	0
17	0.05207	0
18	0.10377	1
19	0.83302	1
20	0.99232	1
21	0.96435	1
22	0.99144	1
23	0.99272	1
24	0.96176	1
25	0.99998	1
26	0.99826	1
27	0.99965	1
28	0.98858	1
29	0.99993	1
30	0.99923	1
31	0.99973	1
32	0.99987	1
33	1.00000	1
34	0.99997	1

Παράλληλα προσαρμόζοντας τα $\hat{p}$ στα δεδομένα του μοντέλου μας παίρνουμε το γράφημα στην Εικόνα 4.1.



Εικόνα 4.1 : *Γράφημα εκτιμημένων πιθανοτήτων*

Όταν το $\hat{p}$ βρίσκεται κοντά στη μονάδα ($\hat{p}$ ≥0.5), τότε το $y = 1$, δηλαδή το πτηνό είναι φύλου αρσενικού, ενώ για τα $\hat{p}$ με τιμή κοντά στο 0 $(\hat{p} < 0.5)$, το $y = 0$ και άρα πρόκειται για θηλυκό πτηνό.

Στον πίνακα θα δούμε την ποσότητα των αρσενικών πτηνών όταν έχουμε $\hat{p} \geq 0.5$ και την ποσότητα των θηλυκών όταν $\hat{p} < 0.5$, καθώς επίσης και τα σφάλματα. Για να πούμε ότι το μοντέλο μας έχει καλή προβλεπτική ικανότητα θα πρέπει οι αριθμοί αυτοί να είναι ικανοποιητικά μεγάλοι (ενώ τα σφάλματα μικρά).

Πίνακας 4.7 : *Αριθμός εκτιμημένων πιθανοτήτων*

	Y=1	**Y=0**
$\boldsymbol{\hat{p} \geq 0.5}$	16 (94%)	1
$\boldsymbol{\hat{p}} < 0.5$	1	16 (94%)

Εφόσον τα ποσοστά σφαλμάτων είναι πολύ χαμηλά, μπορούμε να υποθέσουμε ότι το μοντέλο μας έχει καλή προβλεπτική ικανότητα. Παρακάτω θα ακολουθήσει και περαιτέρω έλεγχος σχετικά με την προβλεπτική ικανότητα του μοντέλου.

Καθώς το μοντέλο έχει καλή προβλεπτική ικανότητα μπορούμε να προσαρμόσουμε σε αυτό τα δεδομένα του δεύτερου δείγματος, για το οποίο δεν γνωρίζουμε το φύλο των πτηνών και να υπολογίσουμε τις εκτιμημένες πιθανότητες. Με το ίδιο σκεπτικό όπως και παραπάνω

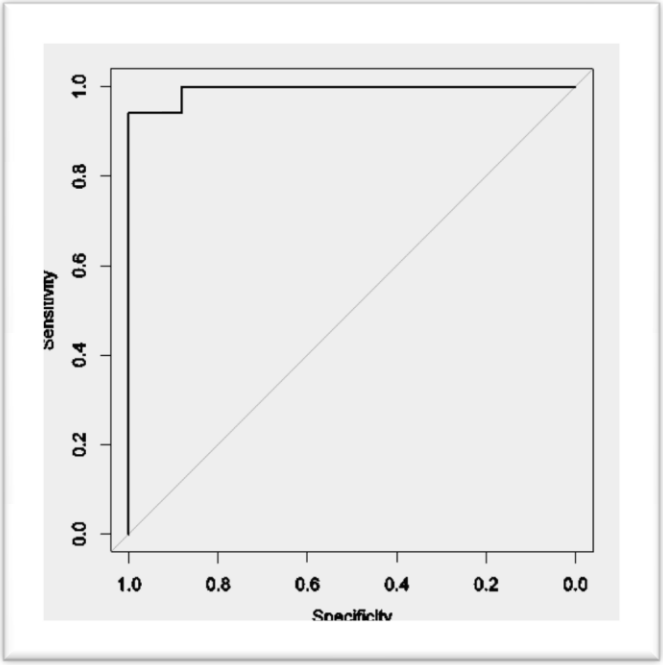
μπορούμε να υποθέσουμε με μεγάλη σιγουριά (λόγω της καλής πρόβλεψης) το φύλο των πτηνών (Πίνακας 4.8).

Πίνακας 4.8 : *Εκτιμημένες πιθανότητες για το δεύτερο δείγμα.*

Αριθμός δείγματος	$\widehat{\boldsymbol{p}}$	**Y**
1	0.001092	0
2	0.000001	0
3	0.000184	0
4	0.000016	0
5	0.001460	0
6	0.013989	0
7	0.033420	0
8	0.009421	0
9	0.007726	0
10	0.000847	0
11	0.010685	0
12	0.010685	0
13	0.014759	0
14	0.099625	0
15	0.996102	1
16	0.998425	1
17	0.783978	1
18	0.998822	1
19	0.991207	1
20	0.990884	1
21	0.990884	1
22	0.998638	1
23	0.998337	1
24	0.999921	1

Στη συνέχεια θα εφαρμόσουμε την καμπύλη του ROC (Receiver Operating Characteristic) προκειμένου να εξετάσουμε καλύτερα την προβλεπτική ικανότητα του μοντέλου. Ο κάθετος άξονας μετράει την ευαισθησία, ενώ ο οριζόντιος την ειδικότητα (specificity). Για να θεωρηθεί καλή μια καμπύλη ROC θα πρέπει να ισχύουν δύο πράγματα :

- να έχουμε υψηλή ευαισθησία και χαμηλή ειδικότητα
- το εμβαδόν (AUC), που σχηματίζεται κάτω από κάθε καμπύλη να είναι αρκετά μεγάλο. Όσο μεγαλύτερη η τιμή του, τόσο καλύτερη είναι η πρόβλεψη.



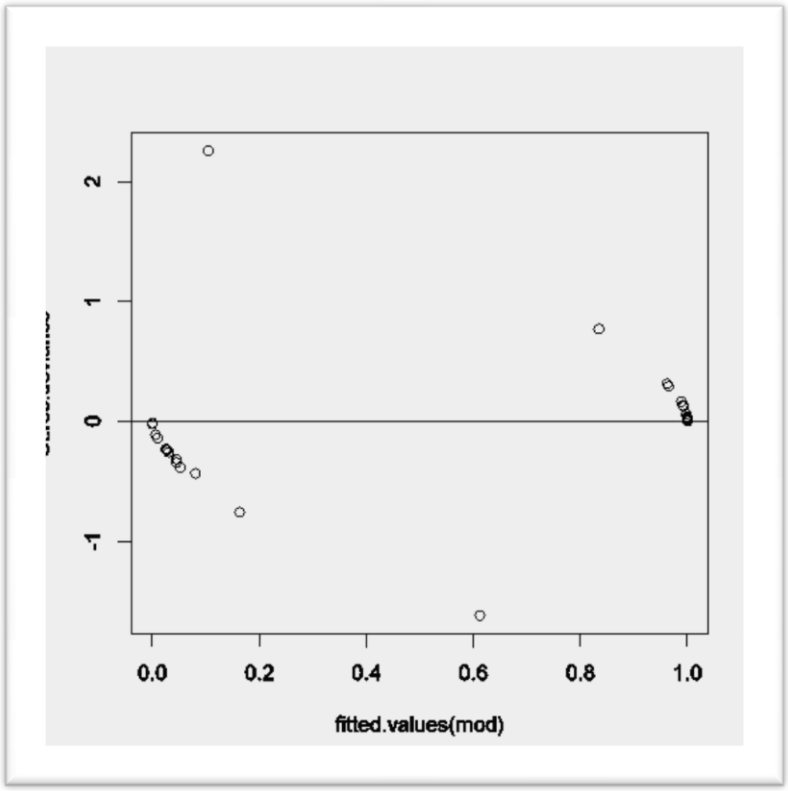
Εικόνα 4.2 : *Καμπύλη ROC*

Από το παραπάνω διάγραμμα παρατηρούμε ότι έχουμε μια πολύ καλή καμπύλη ROC. Αυτό φαίνεται από το γεγονός ότι η ευαισθησία είναι πολύ υψηλή, ενώ παράλληλα η τιμή του εμβαδού κάτω από την καμπύλη είναι AUC=0.9931. Μπορούμε λοιπόν να συμπεράνουμε, τόσο από την καμπύλη, όσο και από το εμβαδόν της ότι η προβλεπτική ικανότητα του μοντέλου είναι πολύ καλή.

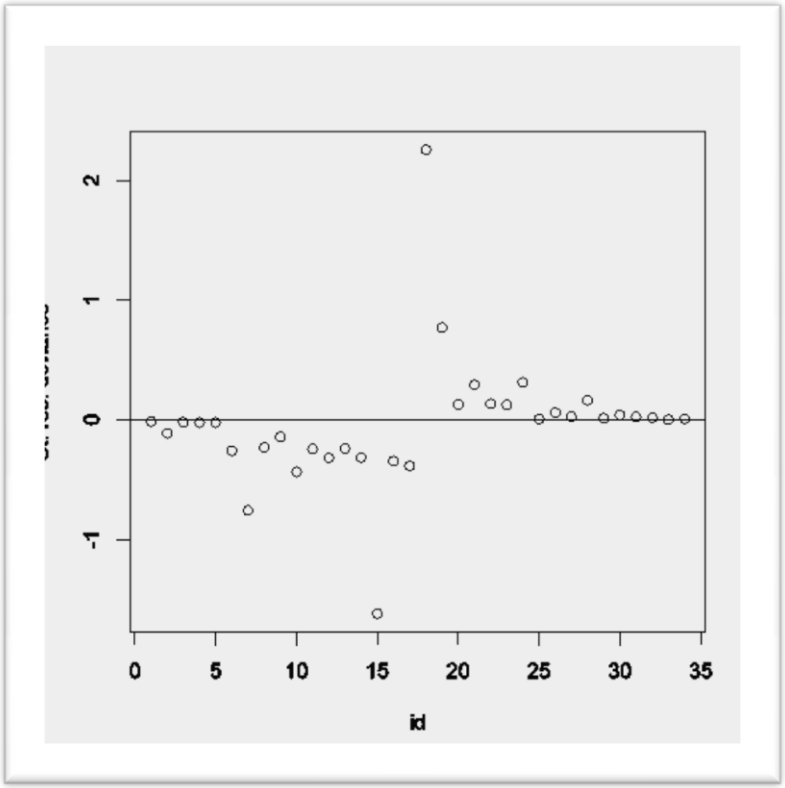
Ολοκληρώνοντας το κεφάλαιο, θα εφαρμόσουμε μερικούς διαγνωστικούς ελέγχους για το βέλτιστο μοντέλο, οι οποίοι βασίζονται στα υπόλοιπα και στα μέτρα επιρροής. Αρχικά θα κατασκευάσουμε τα γραφήματα των τυποποιημένων υπολοίπων deviance σε σχέση με τις εκτιμώμενες τιμές και με βάση τη σειρά των δεδομένων (id) αντίστοιχα.

Με τις δύο αυτές γραφικές παραστάσεις μπορούμε

- Να ελέγξουμε την υπόθεση της ανεξαρτησίας των παρατηρήσεων
- Να εντοπίσουμε πιθανές παρατηρήσεις που αποκλείουν σημαντικά από τις υπόλοιπες.



Εικόνα 4.3 *: Τυποποιημένα υπόλοιπα deviance ως προς τις προσαρμοσμένες τιμές.*

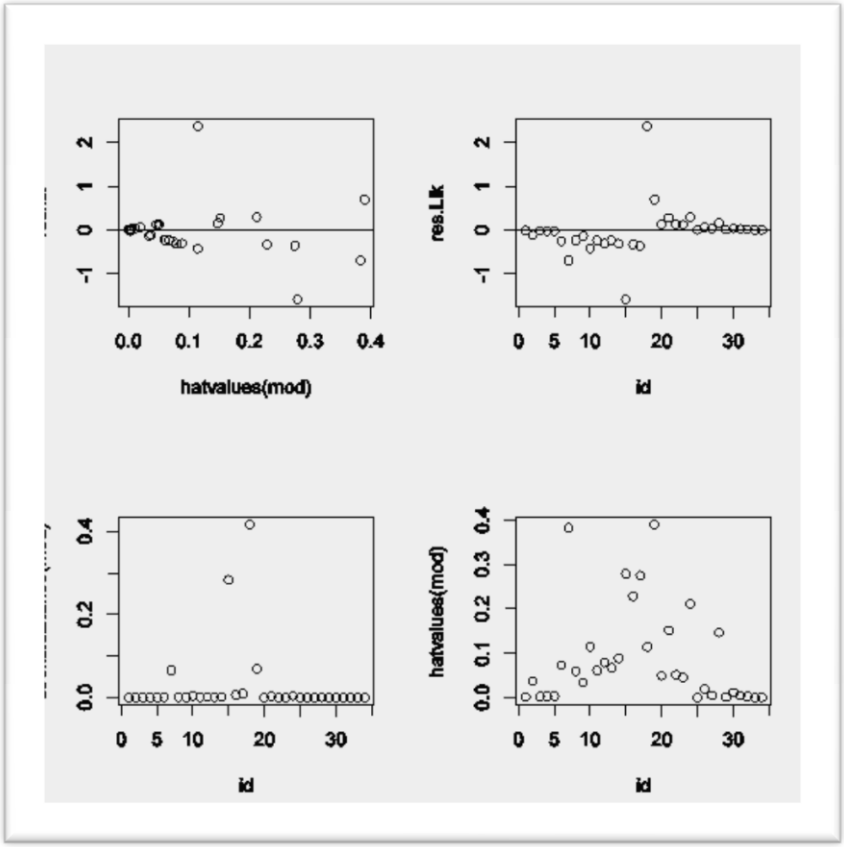


Εικόνα 4.4 : *Τυποποιημένα υπόλοιπα deviance σε σχέση με τον αριθμό της παρατήρησης.*

Από τα παραπάνω σχήματα μας είναι δύσκολο να απορρίψουμε την υπόθεση της ανεξαρτησίας των υπολοίπων, αφού τα υπόλοιπα δεν κατανέμονται τυχαία γύρο από την περιοχή του μηδενός. Επίσης αν παρατηρήσουμε θα δούμε ότι η 15^{η} και η 18^{η} παρατήρηση έχουν σημαντικά μεγαλύτερη απόλυτη τιμή για τα τυποποιημένα υπόλοιπα σε σχέση με τις υπόλοιπες. Οι συγκεκριμένες παρατηρήσεις αντιστοιχούν σε πτηνά, των οποίων τα ράμφη τους έχουν ιδιαίτερα δυσανάλογες διαστάσεις σε σχέση με το φύλο τους, γεγονός που τις διαφοροποιεί από όλες τις άλλες.

Για τον εντοπισμό πιθανών σημείων επιρροής μπορούμε να κατασκευάσουμε το διάγραμμα των υπολοίπων πιθανοφάνειας ως προς τα h_{ii} καθώς και τα γραφήματα δείκτη (index plots) των υπολοίπων πιθανοφάνειας, των h_{ii}, των αποστάσεων του Cook.

Όλα τα προαναφερθέντα γραφήματα παρουσιάζονται στην Εικόνα 4.5, το οποίο κατασκευάστηκε από την R.



Εικόνα 4.5 : *Διάφορα γραφήματα ελέγχου της R.*

Από την εικόνα των γραφικών αυτών παραστάσεων είναι φανερό ότι δεν υπάρχει κανένα χαρακτηριστικό που να δημιουργεί σοβαρές αμφιβολίες ως προς την ορθότητα του μοντέλου.

5 ΣΤΑΤΙΣΤΙΚΗ ΜΕΛΕΤΗ ΣΕ ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ

5.1 Παρουσίαση δείγματος και μεταβλητών

Το δείγμα μας αποτελείται από 115 συνολικά άτομα, στα οποία έγιαναν 17 διαφορετικές εξετάσεις (μετρήσεις), προκειμένου να δούμε αν αυτά έχουν εκτεθεί ή όχι στην χημική ουσία. Το δείγμα μας χωρίζεται σε δύο ομάδες. Γνωρίζουμε ότι από τα 115 άτομα, τα 52 έχουν εκτεθεί στην χημική ουσία, ενώ τα 27 δεν έχουν. Η δεύτερη ομάδα αποτελείται από άτομα, τα οποία δεν γνωρίζουμε αν είναι εκτεθειμένα ή όχι.

Σκοπός μας είναι να ερευνήσουμε την ύπαρξη της εξάρτησης της πιθανότητας έκθεσης ενός ατόμου στην ουσία με τα αποτελέσματα των εξετάσεων. Επομένως η εξαρτημένη μεταβλητή, σε αυτή την περίπτωση θα είναι η έκθεση ή όχι του ατόμου στην χημική ουσία. Η κωδικοποίηση που χρησιμοποιήθηκε φαίνεται στον Πίνακα 5.1.

Πίνακας 5.1: *Κωδικοποίηση εξαρτημένης μεταβλητής.*

Έκθεση	Ναι	Όχι
Κωδικοποίηση	1	0

Και σε αυτή την περίπτωση οι ανεξάρτητες/επεξηγηματικές μεταβλητές είναι όλες συνεχείς. Οι μεταβλητές αυτές είναι μετρήσεις που λάβαμε, μέσω των εξετάσεων που πραγματοποιήθηκαν στο δείγμα μας.

Στον Πίνακα 5.2 παρουσιάζουμε ένα μικρό μέρος του δείγματος που χρησιμοποιήθηκε.

Πίνακας 5.2: *Δείγμα*

C1	**C2**	**C3**	**C4**	**C5**	**C6**	**C7…**	**…C13**	**C14**	**C15**	**C16**	**C17**	**Έκθεση**
4.1	6	0.35	6.8	1.23	2.34	45	0.93	272	114	2.12	2.91	1
6.3	8	0.33	5.7	1.08	2.29	44	1.80	232	102	1.02	5.6	1
6.9	9	0.42	6.7	0.86	2.37	44	1.17	352	118	1.02	6.57	0
5	12	0.33	5.6	1	2.34	43	0.93	344	100	1.04	5.38	0
10.2	11	0.35	6.6	0.95	2.33	44	1.93	455	99	1.57	4.20	0

5.2 Εφαρμογή του μοντέλου λογιστικής παλινδρόμησης

Αρχικά θα προσαρμόσουμε το μοντέλο, με όλα τα δεδομένα, με τη βοήθεια του στατιστικού πακέτου της R. Και σε αυτή την περίπτωση σκοπός μας είναι εν τέλει να καταλήξουμε στο καταλληλότερο στατιστικά μοντέλο. Μετά την προσαρμογή του μοντέλου, η R μας δίνει τα αποτελέσματα που φαίνονται στον Πίνακα 5.3.

Πίνακας 5.3 : *Αποτελέσματα*

Μετρήσεις	Coef	SE coef	z-value	P(> |z|)
C1	0.454021	0.228609	1.99	0.047
C2	-0.0921234	0.170204	-0.54	0.588
C3	-17.5697	8.86137	-1.98	0.047
C4	0.553136	0.897255	0.62	0.538
C5	1.61846	2.69260	0.60	0.548
C6	-34.4002	10.5028	-3.28	0.001
C7	0.245205	0.244073	1.00	0.315
C8	0.161909	0.136211	1.19	0.235
C9	-0.801239	0.842821	-0.95	0.342
C10	0.0818266	0.0398471	2.05	0.040
C11	0.0220476	0.0679672	0.32	0.746
C12	0.085805	0.051733	1.66	0.097
C13	0.727097	0.53589	1.36	0.175
C14	-0.000898	0.005542	-0.16	0.871
C15	0.070042	0.051537	1.36	0.174
C16	-1.53301	2.5132	-0.61	0.542
C17	0.200235	0.50768	0.39	0.693
AIC = 85.405 **Residual deviance = 49.405 on 61 df**				

Από τα αποτελέσματα του προσαρμοσμένου μοντέλου με όλες τις μεταβλητές και ειδικότερα από τις p-τιμές των ελέγχων Wald, γίνεται γρήγορα αντιληπτό ότι το μοντέλο αυτό δεν είναι στατιστικά πολύ καλό και ενδεχομένως να μπορούμε να προβούμε σε κάποιες βελτιώσεις. Πιο συγκεκριμένα μπορούμε να δούμε ότι πολλές μετρήσεις που έχουν ληφθεί έχουν p-value >> 0,5, όπως παραδείγματος χάριν η C11, C14 και λοιπά. Βλέπουμε επίσης ότι η πιο σημαντική μεταβλητή του μοντέλου είναι η μεταβλητή C6 με p-τιμή 0,001 και ακολουθούν οι λιγότερο σημαντικές C10 και μαζί οι C1 και C3.

Προκειμένου να βελτιώσουμε το αρχικό μοντέλο, θα χρησιμοποιήσουμε την διαδικασία διαδοχικής αφαίρεσης (backward elimination), η οποία ξεκινάει εισάγοντας όλες τις μεταβλητές, όπως έχουμε ήδη κάνει και αφαιρεί μια-μια τις μεταβλητές αρχίζοντας από την

λιγότερο σημαντική. Με βάση τον έλεγχο Wald και το κριτήριο AIC θα βρούμε το καταλληλότερο, για τα δεδομένα μας, μοντέλο.

Καταλήγουμε έτσι στο μοντέλο με τα αποτελέσματα στον Πίνακα 5.4.

Πίνακας 5.4 : *Αποτελέσματα*

Μεταβλητές	**Coef**	**Se coef**	**z-value**	**Pr(> |z|)**
C1	0.33156	0.17201	1.93	0.054
C3	-19.1186	7.67830	-2.49	0.013
C6	-25.9297	7.5429	-3.44	0.001
C8	0.22627	0.10884	2.08	0.038
C10	0.06781	0.03123	2.17	0.030
C12	0.06468	0.04566	1.42	0.157
C13	0.71168	0.46271	1.54	0.124
C15	0.08377	0.04376	1.91	0.056
AIC = 71.863 Residual deviance = 53.863 on 70 df				

Από τα p-value των αποτελεσμάτων βλέπουμε ότι η σημαντικότερη μεταβλητή είναι εκείνη των μετρήσεων C6 και ακολουθούν οι C3 και C10. Εφαρμόζοντας το κριτήριο AIC βλέπουμε ότι το μοντέλο μας βελτιώθηκε αρκετά, με AIC = 71.825, έναντι του αρχικού με AIC = 85.38. Παρατηρούμε όμως ότι για δύο μεταβλητές (C12, C13) η p-value είναι οριακή. Παρακάτω θα ελέγξουμε, με βάση τις deviance, αν το μοντέλο μας είναι καλύτερο από αυτό που δεν περιλαμβάνει τις δύο μεταβλητές.

Ένας άλλος τρόπος σύγκρισης των μοντέλων είναι με την σύγκριση των τιμών της ελεγχοσυνάρτησης deviance από την κατανομή X^2. Η μέθοδος αυτή δεν δίνει πάντα ακριβή αποτελέσματα, όμως εδώ ίσως μας φανεί χρήσιμη. Με την βοήθεια της R λαμβάνουμε τα αποτελέσματα, στα οποία παρουσιάζεται ο πίνακας ανάλυσης της deviance για τα δύο μοντέλα.

Πίνακας 5.5 : *Υπόλοιπα deviance*

	Df	**Resid. deviance**	**Pr(>Chi)**	**AIC**
Model 1	70	53.863		71.863
Model 2	72	60.111	0.044	74.111

Συγκρίνοντας τώρα τα μοντέλα με την μέθοδο ανάλυσης της deviance βλέπουμε, εν τέλει, ότι το βέλτιστο μοντέλο είναι εκείνο που περιέχει τις μεταβλητές C12 και C13.

Η προσαρμοσμένη εξίσωση παλινδρόμησης του μοντέλου είναι η

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = 35.77 + 0.33\ C1 - 19.12\ C3 - 25.93\ \ C6 + 0.23\ C8 + 0.068\ C10 + 0.065\ C12 + 0.7\ C13 + 0.084\ C15$$

Η προηγούμενη σχέση γράφεται ισοδύναμα, ως προς $\hat{p}$

$$\hat{p} = \frac{\exp(35.77 + 0.33\ C1 - 19.12\ C3 - 25.93\ C6 + 0.23\ C8 + 0.068\ C10 + 0.065\ C12 + 0.7\ C13 + 0.084\ C15)}{1+\exp(35.77 + 0.33C1 - 19.12C3 - 25.93C6 + 0.23C8 + 0.068C10 + 0.065C12 + 0.7C13 + 0.084C15)}$$

Ένα δείγμα από τις εκτιμημένες πιθανότητες των τιμών της μεταβλητής απόκρισης, της πρώτης ομάδας ατόμων, φαίνεται στον πίνακα 5.6. Οι τιμές αυτές υπολογίστηκαν μέσω του Minitab.

Πίνακας 5.6 : *Εκτιμημένες πιθανότητες*

#Άτομο	$\hat{p}$	**Y**
1	0.94	1
2	0.53	1
3	0.31	1
4	0.78	1
5	0.81	1
6	0.95	1
7	0.59	1
53	0.06	0
54	0.39	0
55	0.71	0
56	0.05	0
57	0.32	0
58	0.4	0
59	0.94	0

Στον Πίνακα 5.7 βλέπουμε τον αριθμό των εκτιμημένων πιθανοτήτων. Αυτό που αναμένουμε είναι για $\hat{p} \geq 0.5$ να έχουμε Y= 1, δηλαδή το άτομο έχει εκτεθεί στη χημική ουσία, ενώ για $\hat{p} < 0.5$, Y=0 και δεν υπάρχει έκθεση του ατόμου.

Πίνακα 5.7 : *Αριθμός εκτιμημένων πιθανοτήτων*

	# Y=1	**# Y=0**
$\hat{p} \geq 0.5$	46 (88.5%)	6
$\hat{p} < 0.5$	9	18 (66.7%)

Παρατηρούμε ότι η προβλεπτική ικανότητα του μοντέλου είναι πολύ καλή, ειδικά για τις περιπτώσεις των ατόμων που έχουν εκτεθεί στην χημική ουσία.

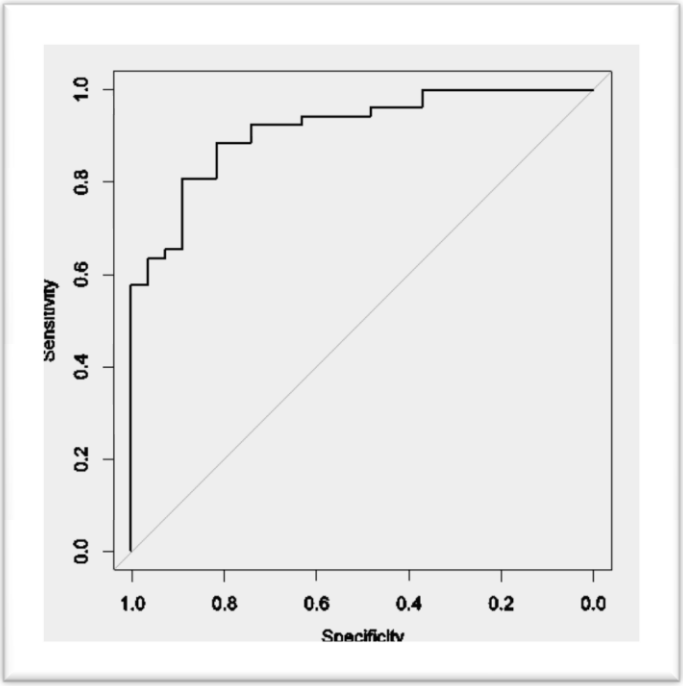
Αρχικός σκοπός μας ήταν να δημιουργήσουμε ένα μοντέλο, το οποίο να μπορεί να προβλέπει με καλή πιθανότητα τις τιμές της μεταβλητής απόκρισης. Στην συνέχεια λοιπόν θα προσαρμόσουμε στο μοντέλο μας τα δεδομένα της δεύτερης ομάδας ατόμων (εκείνων που δεν γνωρίζουμε αν εκτέθηκαν ή όχι) και θα υπολογίσουμε την εκτιμημένη πιθανότητα. Όπως και προηγουμένως, στην περίπτωση που έχουμε $\hat{p} \geq 0.5$ θα θεωρούμε ότι το άτομο έχει εκτεθεί, ενώ όταν προκύπτει $\hat{p} < 0.5$ θα θεωρούμε ότι πιθανότατα το άτομο δεν εκτέθηκε. Οι τιμές της $\hat{p}$ υπολογίστηκαν και εδώ με το στατιστικό πακέτο Minitab.

Πίνακας 5.8 : *Εκτιμημένες πιθανότητες*

#Άτομο	$\hat{p}$	Y
1	0.970617	1
2	0.963649	1
3	0.909175	1
4	0.165136	0
5	0.993842	1
6	0.984768	1
7	0.808656	1
8	0.896191	1
9	0.996537	1
10	0.084656	0
12	0.995952	1
13	0.983673	1
14	0.999319	1
15	0.993117	1
16	0.279952	0
17	0.970599	1
18	0.931375	1
19	0.007463	0
20	0.927331	1
21	0.929942	1
22	0.896553	1
23	0.986674	1
24	0.998934	1
25	0.979492	1
26	0.999047	1
27	0.999869	1
28	0.992727	1
29	0.998133	1
30	0.902322	1
31	0.998410	1
32	0.451552	0
33	0.853772	1
34	0.999778	1
35	0.989591	1

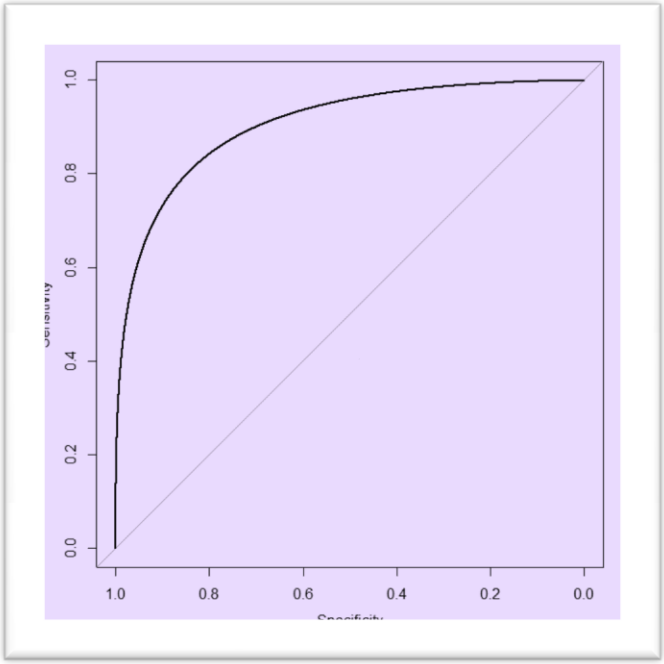
Με βάση τα αποτελέσματα στους Πίνακες 5.8 και 5.9, μπορούμε να πούμε ότι τα άτομα που εμφάνισαν Y=1, είναι πολύ πιθανό, να έχουν έρθει σε επαφή με τη χημική ουσία. Εκείνοι που εμφάνισαν Y=0, φαίνεται να μην διατρέχουν κάποιον ιδιαίτερο κίνδυνο.

Σε αυτό το σημείο θα εφαρμόσουμε και θα εξετάσουμε την καμπύλη του ROC (Receiver Operating Characteristic). Όπως έχουμε προαναφέρει, η καμπύλη ROC είναι ένα καλό εργαλείο όταν πρόκειται να εξετάσουμε την προβλεπτική ικανότητα του μοντέλου μας. Με την βοήθεια της R λαμβάνουμε το διάγραμμα που εμφανίζεται στην Εικόνα 5.1.



Εικόνα 5.1 : *Καμπύλη ROC*

Μπορούμε να δούμε και μια άλλη εκδοχή της καμπύλης ROC στην Εικόνα 5.2. Στο παρακάτω διάγραμμα έχουμε εξομαλύνει τα σημεία της γραμμής δημιουργώντας μία μονοκόμματη καμπύλη. Η διόρθωση αυτή διευκολύνει την ανάλυση της καμπύλης και τη διευκρίνιση του εμβαδού που σχηματίζεται κάτω από αυτήν.

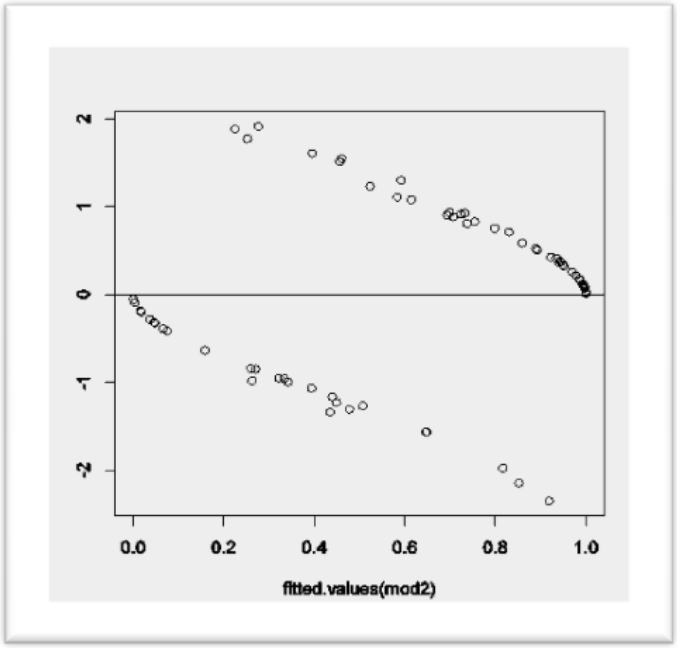


Εικόνα 5.2 : *Εξομαλυμένη καμπύλη ROC*

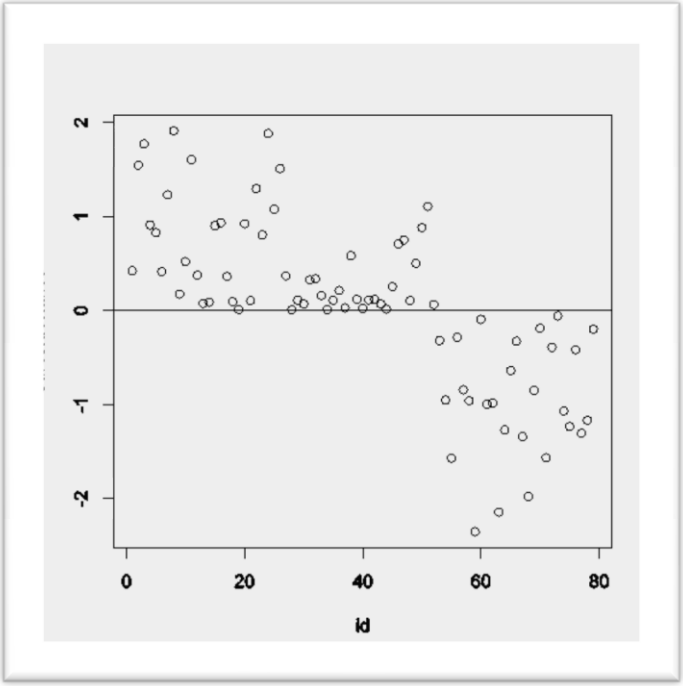
Από τα σχήματα στις Εικόνες 5.1 και 5.2 βλέπουμε ότι η ευαισθησία του μοντέλου είναι πολύ υψηλή, ενώ από την R παίρνουμε την τιμή του ποσοστού που καλύπτει το εμβαδόν κάτω από την καμπύλη, η οποία είναι AUC=0.91. Αυτό σημαίνει ότι η προβλεπτική ικανότητα του μοντέλου μας είναι πολύ καλή.

Τέλος θα κάνουμε μερικούς από τους διαγνωστικούς ελέγχους προκειμένου να ελέγξουμε την ανεξαρτησία των παρατηρήσεων, καθώς και την ύπαρξη σημαντικών αποκλίσεων, που ίσως να μας δημιουργούν πρόβλημα στα συμπεράσματά μας.

Από το στατιστικό πακέτο της R παίρνουμε τα γραφήματα τυποποιημένων υπολοίπων, Εικόνα 5.3 και Εικόνα 5.4.



Εικόνα 5.3 : *Τυποποιημένα υπόλοιπα deviance ως προς τις προσαρμοσμένες τιμές.*



Εικόνα 5.4 : *Τυποποιημένα υπόλοιπα deviance σε σχέση με τον αριθμό της παρατήρησης.*

Στα τυποποιημένα υπόλοιπα δεν υπάρχουν ενδείξεις που θα πρέπει να μας προβληματίσουν σχετικά με τις παρατηρήσεις.

6 ΣΤΑΤΙΣΤΙΚΗ ΜΕΛΕΤΗ ΣΕ ΑΡΧΑΙΟΛΟΓΙΚΑ ΕΥΡΗΜΑΤΑ

6.1 Προϊστορικά ορυκτά μεταλλεύματα και ο προσδιορισμός τους

Στην Δυτική Ευρώπη, τα συνηθέστερα αρχαιολογικά ευρήματα και συνήθως τα μοναδικά που διασώζονται, για το μεγαλύτερο μέρος της ύπαρξης του ανθρώπου, είναι εκείνα που κατασκευάζονταν από πέτρα. Τα αντικείμενα αυτά χρησιμοποιήθηκαν ευρέως από το 3000 π.χ (ή λίγο νωρίτερα) έως το 1500 π.χ. Αναφερόμαστε στην Νεολιθική ή πρώιμη Εποχή του Χαλκού, η οποία χαρακτηρίζεται από την εμφάνιση ή την ευρεία εφαρμογή των πρώιμων μεθόδων καλλιέργειας. Ο προσδιορισμός των πηγών για τις πρώτες ύλες αυτών των εργαλείων είναι ένα γεωλογικό πρόβλημα. Ενιαία εργαλεία μικρού μεγέθους, όπως ξύστρα ή μαχαίρια, μπορεί να έχουν κατασκευαστεί από τυχαία ευρήματα πυρόλιθου όπως, βότσαλα από ένα ρέμα ή κάποια χαλίκια, όπου και στις δύο περιπτώσεις είναι πολύ πιθανό τα ευρήματα να βρεθούν σε μεγάλη απόσταση από τον γονικό βράχο. Το πρόβλημα αυτό στο παρελθόν ήταν ακόμα δυσκολότερο επειδή, αν και υπάρχουν πολλές περιγραφές ως προς τις αλλαγές του υλικού, τόσο στην όψη όσο και στον χρωματισμό (Rankine 1952, Hurst και Kelly 1966), δεν υπάρχει καμία περιγραφή, μέχρι και σήμερα, στην πετρολογία για τον διαχωρισμό των πυρόλιθων από διαφορετικές πηγές.

Η μεταβολή του πετρογραφικού σχεδίου σε παρόμοιους τύπους πετρωμάτων είναι γνωστή. Οι καλά μελετημένες περιοχές, με μεγάλη γεωλογική ποικιλία, μας επιτρέπουν τον ευκολότερο προσδιορισμό του δείγματος ως προς την πηγή του πετρώματος (ή τουλάχιστον τον περιορισμό των εναλλακτικών επιλογών), μέσω της εξέτασης του ορυκτολογικού σχεδίου και της σύγκρισης της σύστασης μιας μικρής επιφάνειας του δείγματος, με τα γνωστά πετρώματα. Τα πλεονεκτήματα αυτής της μεθόδου έχουν χρησιμοποιηθεί, τόσο στην Βρετανία όσο και αλλού , για τον προσδιορισμό της πηγής πέτρινων στηλών, που κατασκευάστηκαν από σκληρό βράχο (Shotton 1969).

Ο πυρόλιθος έχει γενικά θεωρηθεί ως χαλκηδόνιος λίθος, ομοιόμορφης σύνθεσης, που αποτελείται σχεδόν εξ ολοκλήρου από πυρίτιο και νερό. Η δομή του είναι λιγότερο ομοιόμορφη και περιλαμβάνει και λεπτό, άμορφο και κρυσταλλικό διοξείδιο του πυριτίου, με ενσωματωμένα απολιθώματα ανθρακικού ασβεστίου (κοινή ονομασία, κιμωλία). Ωστόσο, υπάρχει σημαντική μεταβολή της δομής μεταξύ πυρόλιθων της ίδιας πηγής, ενώ τα απολιθώματα δεν είναι σε γενικές γραμμές καλά διατηρημένα. Έχουν γίνει αρκετές προσπάθειες, για των εντοπισμό των πηγών του πυρόλιθου, με την χρήση των απολιθωμάτων, χωρίς όμως κάποια αξιοσημείωτη επιτυχία (Bigod 1950, Wetzel1941, Laming 1959). Ο λόγος που η μέθοδος αυτή δεν είχε επιτυχία είναι επειδή, σε αρκετές

περιπτώσεις, παρατηρήθηκε χρήση διαφορετικών πηγών πυρόλιθου στην ίδια περιοχή, χωρίς όμως να μπορούμε να εντοπίσουμε την τοποθεσία των πηγών αυτών (Deflandre 1966).

Η ύπαρξη ιχνοστοιχείων στον πυρόλιθο μας οδηγεί σε μια άλλη, πολλά υποσχόμενη προσέγγιση. Το Βρετανικό Μουσείο Εργαστηριακών ερευνών πρόσφατα κατέδειξε ότι μέσω φασματοσκοπικής ανάλυσης, μπορούν να εντοπιστούν στον πυρόλιθο, μέχρι και είκοσι διαφορετικά ιχνοστοιχεία. Σε ένα παλαιότερο άρθρο (Sieveking et al., 1970), οι συγγραφείς είχαν επισημάνει ότι υπάρχουν διαφοροποιήσεις στα ιχνοστοιχεία, που συνθέτουν τον πυρόλιθο, ως προς την γεωγραφική τους τοποθεσία. Εξαιτίας της έλλειψης άλλων καταλληλότερων μεθόδων, καταλήξαμε ότι η μέθοδος, που περιγράφηκε, μπορεί να χρησιμοποιηθεί, τουλάχιστον σε ορισμένες περιπτώσεις, για τον προσδιορισμό των πηγών των πρώτων υλών, που προορίζονταν για την κατασκευή εργαλείων από πυρόλιθο. Κάτι τέτοιο έχει ιδιαίτερη σημασία για την αρχαιολογία.

6.2 **Παρουσίαση δείγματος και μεταβλητών**

Προκειμένου να ελεγχθεί η εγκυρότητα αυτής της μεθόδου, επιλέχθηκε μια ομάδα δειγμάτων από διάφορες περιοχές της Βρετανίας και της υπόλοιπης Ευρώπης, οι οποίες ήταν γνωστές για την χρήση του πυρόλιθου στην προϊστορική εποχή. Οι περιοχές που επιλέχθηκαν ήταν εκείνες που χρησιμοποιήθηκαν, από τον άνθρωπο, κυρίως στη Νεολιθική και στην Πρώιμη Εποχή του Χαλκού για την εξαγωγή πετρωμάτων. Οι περιοχές αποτελούνταν κυρίως από ορυχεία πυρόλιθου , με μία ή δύο εξαιρέσεις. Τέτοιες τοποθεσίες είναι, για παράδειγμα, οι περιοχές Grimes Graves, Cissbury και Easton Down στην Μεγάλη Βρετανία, καθώς και η Spiennes στο Βέλγιο.

Για την αρχική μας μελέτη επιλέχθηκε μια ομάδα επτά ορυχείων/ πηγών πυρόλιθου, πέντε από την Μ. Βρετανία και δύο από την Δυτική Ευρώπη. Περίπου είκοσι δείγματα από κάθε θέση εξετάστηκαν αρχικά με φασματοσκοπία εκπομπής και έπειτα με φασματοσκοπία ατομικής απορρόφησης, προκειμένου να αναλυθούν σε οκτώ κύρια ιχνοστοιχεία, τον σίδηρο (Fe), το αργίλιο (Al), το ασβέστιο (Ca), το μαγνήσιο (Mg), το νάτριο (Na), το λίθιο (Li), το κάλιο (K) και τέλος τον φώσφορο (P). Τα αποτελέσματα αυτών των αναλύσεων έδειξαν έγκυρες στατιστικά διαφορές των ιχνοστοιχείων που υπάρχουν στις περιοχές που ερευνήθηκαν, με εξαίρεση δύο από αυτές. Οι δύο τοποθεσίες (Blackpatch και Cissbury) βρίσκονται σε απόσταση πέντε μιλίων μεταξύ τους στο Sussex Downs (Sieveking et al., 1970).

Για την διευκόλυνσή μας έχουμε κωδικοποιήσει την εξαρτημένη μας μεταβλητή, δηλαδή τις τοποθεσίες, από όπου συλλέχτηκε το δείγμα μας, με τον τρόπο που βλέπουμε στον Πίνακα 6.1.

Πίνακας 6.1 : *Η κωδικοποίηση των τοποθεσιών*

Grimes Graves	Easton Down	Black Patch	Cissbury	Peppard
1	2	3	4	5

Beer	Clanfield	Le Grand Pressigny	Spienners	St. Gertrude
6	7	8	9	10

Στον πίνακα 6.2 παραθέτουμε ένα μικρό μέρος του δείγματός μας, ένα εύρημα από κάθε τοποθεσία. Όλες οι επεξηγηματικές μεταβλητές είναι συνεχείς.

Πίνακας 6.2: *Δείγμα για τα αρχαιολογικά ευρήματα.*

No.	Al	Fe	Mg	K	Na	Li	P	Ca	Τοποθεσία
1	636	100	29	240	254	6	87	472	1
99	266	108	29	202	238	8	135	3525	2
139	325	59	22	204	158	14	64	2490	3
6	294	140	23	195	174	4	42	1035	4
79	343	110	21	275	246	4	72	448	5
275	1220	290	53	481	394	11	258	680	6
394	195	22	34	145	130	4	75	2250	7
59	656	514	27	256	190	8	3	133	8
119	314	149	73	215	328	44	840	18100	9
159	308	100	35	196	221	5	240	1500	10

6.3 Εφαρμογή ανάλυσης κυρίων συνιστωσών και ανάλυση συστάδων

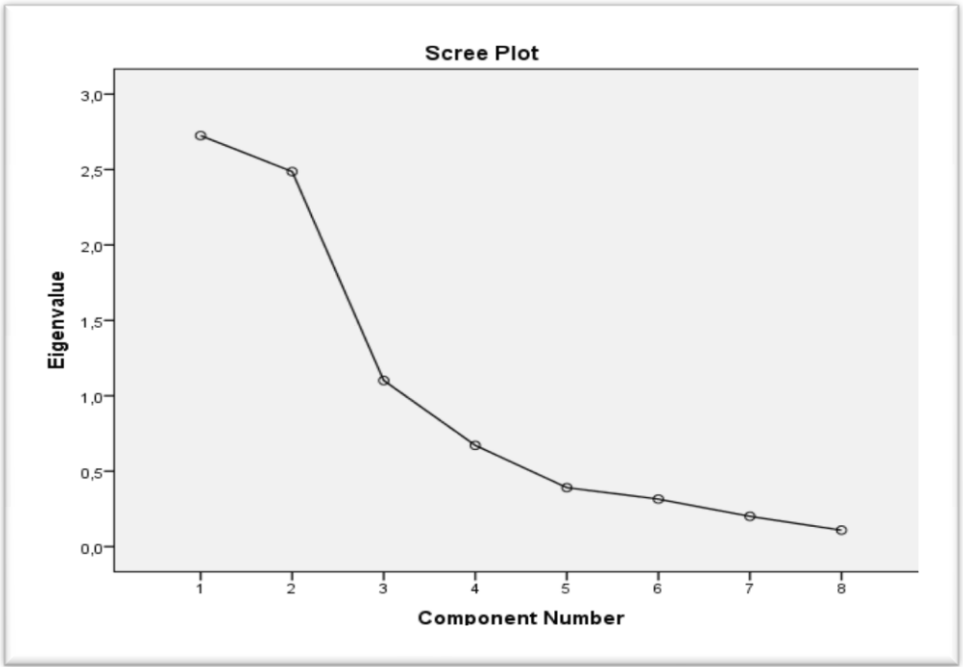
Πριν προχωρήσουμε σε περαιτέρω αναλύσεις, είναι απαραίτητο να απλοποιήσουμε τη δομή των δεδομένων μας, προκειμένου να μειώσουμε τις πολλές μεταβλητές, που εμφανίζουν έντονη συσχέτιση, σε λίγες ανεξάρτητες κύριες συνιστώσες, χωρίς όμως να θυσιάσουμε πολύτιμη πληροφορία. Για τον σκοπό αυτό θα χρησιμοποιήσουμε αρχικά την ανάλυση κυρίων συνιστωσών.

Η ανάλυση κυρίων συνιστωσών, που εφαρμόστηκε στα δεδομένα μας, με βάση το πακέτο SPSS, οδήγησε στις ιδιοτιμές που φαίνονται στον Πίνακα 6.3. Παρατηρούμε ότι μόνο οι πρώτες τρεις ιδιοτιμές είναι μεγαλύτερες από 1. Επίσης, μπορούμε να δούμε το ποσοστό της διακύμανσης που εξηγείται από κάθε συνιστώσα, καθώς και το αθροιστικό ποσοστό της διακύμανσης που εξηγείται από τις πρώτες συνιστώσες. Για παράδειγμα οι πρώτες δύο συνιστώσες εξηγούν μόλις λίγο περισσότερο από το 65% της συνολικής διακύμανσης.

Πίνακας 6.3 : *Διακύμανση που εξηγείται από κάθε κύρια συνιστώσα, δεδομένου των ιχνοστοιχείων.*

Συνιστώσα	Διακύμανση (Ιδιοτιμή)	%	Αθροιστικό %
1	2.725	34.062	34.062
2	2.486	31.078	65.140
3	1.101	13.768	78.908
4	0.671	8.393	87.301
5	0.391	4.890	92.191
6	0.315	3.937	96.128
7	0.201	2.510	98.639
8	0.109	1.361	100.000

Στην Εικόνα 6.1 φαίνεται το γράφημα παραγόντων (screeplot). Στην τέταρτη συνιστώσα παρατηρούμε ότι υπάρχει μια καμπύλη. Η τέταρτη και οι επόμενες συνιστώσες έχουν παρόμοιες ιδιοτιμές (Πίνακας 6.3), που σημαίνει ότι καθεμία από αυτές εξηγεί ένα παρόμοιο αλλά μικρό ποσοστό της συνολικής διακύμανσης.



Εικόνα 6.1 : *Γράφημα παραγόντων των ιδιοτιμών ως προς τον αριθμό της κύριας συνιστώσας.*

Από τον Πίνακα 6.3 και την Εικόνα 6.1 μπορούμε να συμπεράνουμε ότι θα χρειαστούμε και τις τρεις πρώτες συνιστώσες, ώστε να μας παρέχουν μια ικανοποιητική αναπαράσταση των μεταβλητών.

Στον Πίνακα 6.4 μπορούμε να δούμε τη συσχέτιση που έχουν οι πρώτες τρεις κύριες συνιστώσες με τις επεξηγηματικές μεταβλητές. Η πρώτη κύρια συνιστώσα συσχετίζεται θετικά με όλες τις μεταβλητές, άλλα έχει πολύ υψηλή συσχέτιση με το Al, το K, το Na και λιγότερο με τον P. Δεδομένων αυτών συμπεραίνουμε ότι η πρώτη κύρια συνιστώσα ξεχωρίζει τα δείγματα ανάλογα με την περιεκτικότητά τους στα παραπάνω ιχνοστοιχεία. Η δεύτερη συνιστώσα συσχετίζεται περισσότερο με το Mg, το Li και το Ca, ενώ έχει και μια καλή συσχέτιση και με το Al και το K, με αρνητικό όμως πρόσημο. Αντίστοιχα η δεύτερη συνιστώσα ξεχωρίζει τα δείγματα που περιέχουν σε μεγάλες ποσότητες τα 3 πρώτα στοιχεία και σε μικρότερες το Al και το K, από εκείνα που έχουν την αντίθετη περιεκτικότητα. Τέλος είναι φανερό ότι η τρίτη συνιστώσα ξεχωρίζει τα δείγματα ανάλογα με την περιεκτικότητά τους σε Fe.

Πίνακας 6.4 : *Συσχέτιση για τις τρεις πρώτες κύριες συνιστώσες*.

	α_1	α_2	α_3
Al	0.746	-0.528	0.143
Fe	0.128	-0.206	0.882
Mg	0.398	0.682	0.336
K	0.839	-0.445	-0.040
Na	0.851	-0.197	-0.140
Li	0.442	0.768	0.013
P	0.564	0.267	-0.401
Ca	0.224	0.896	0.090

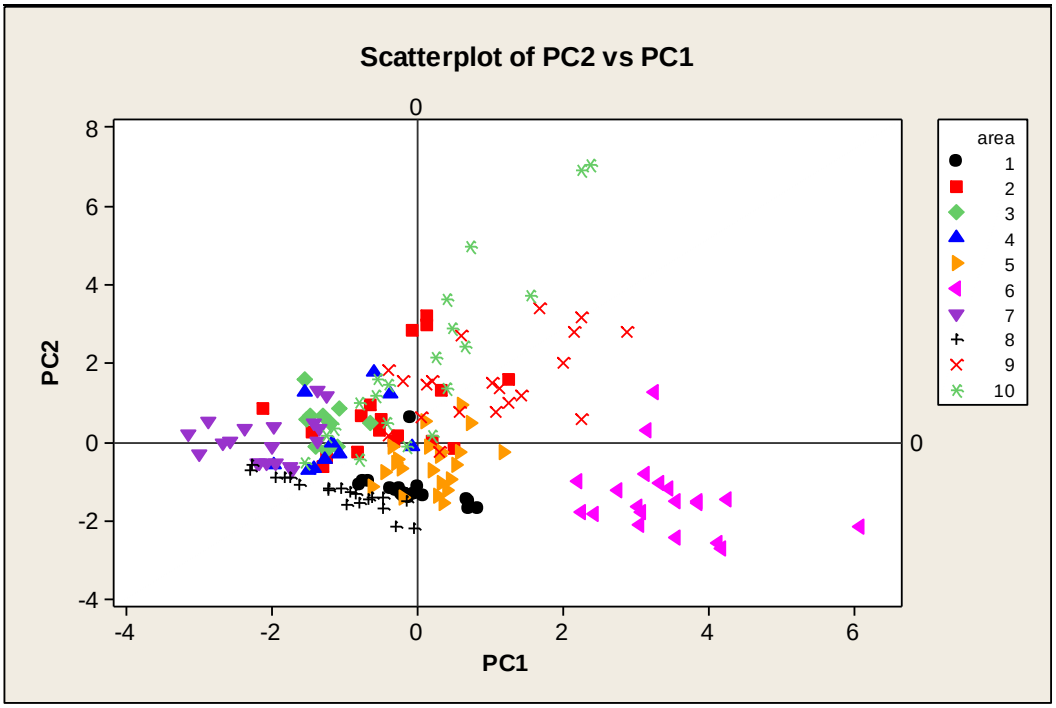
Αυτή η ομαδοποίηση φαίνεται και από την μήτρα ή τον πίνακα συσχετίσεων (Πίνακας 6.5). Εύκολα παρατηρούμε τις έντονες συσχετίσεις μεταξύ των Al, K, Na και μεταξύ των στοιχείων στην ομάδα Mg, Li, Ca. Υπάρχουν χαμηλές και πολλές φορές και αρνητικές συσχετίσεις μεταξύ των δύο ομάδων, καθώς και ιδιαίτερα χαμηλή συσχέτιση του Fe με όλα τα υπόλοιπα στοιχεία. Τέλος παρατηρούμε ότι το P δεν μπορεί να τοποθετηθεί ξεκάθαρα σε μια ομάδα, καθώς έχει μέτριες συσχετίσεις με πολλά στοιχεία από διάφορες ομάδες.

Πίνακας 6.5: *Μήτρα συσχετίσεων για τα δεδομένα των ιχνοστοιχείων.*

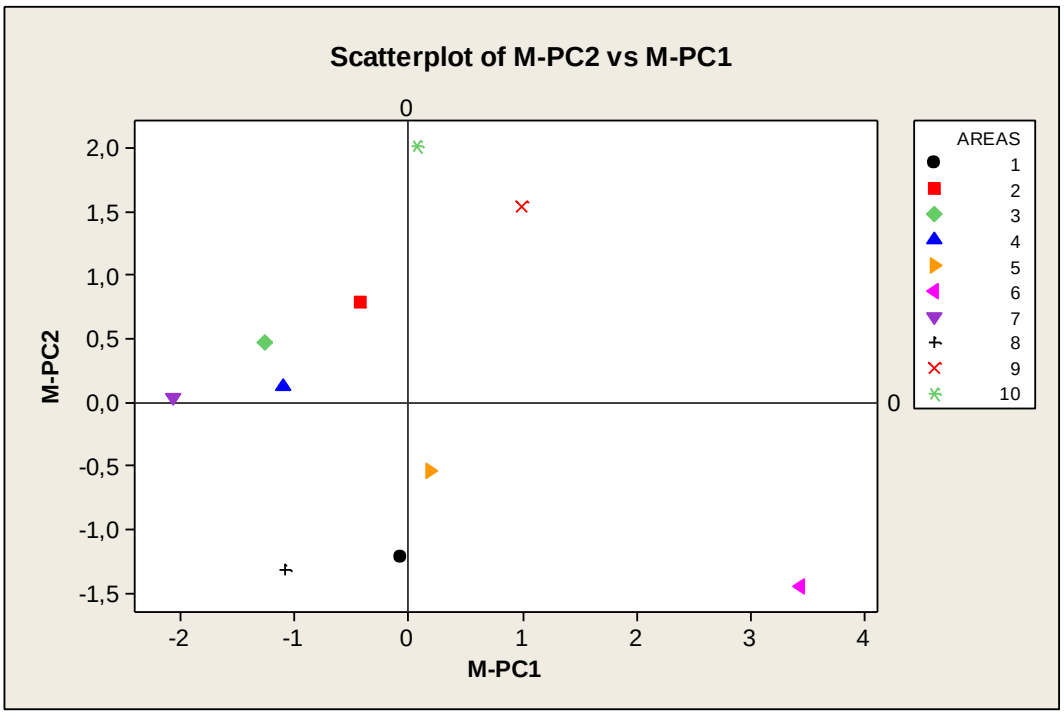
	Al	**K**	**Na**	**P**	**Mg**	**Li**	**Ca**	**Fe**
Al	1.00							
K	0.845	1.00						
Na	0.619	0.768	1.00					
P	0.173	0.289	0.362	1.00				
Mg	0.022	0.012	0.165	0.193	1.00			
Li	-0.54	0.044	0.194	0.385	0.570	1.00		
Ca	-0.285	-0.177	0.023	0.267	0.658	0.741	1.00	
Fe	0.254	0.124	0.015	-0.110	0.074	-0.074	-0.088	1.00

Χρησιμοποιώντας το πακέτο Minitab μπορούμε να πάρουμε επιπλέων γραφήματα για τις δύο πρώτες συνιστώσες. Έτσι στην Εικόνα 6.2 αναπαριστάται η συσχέτιση των μεταβλητών

των περιοχών με τις βαθμολογίες των δύο πρώτων συνιστωσών, ενώ στην Εικόνα 6.3 έχουμε την αντίστοιχη γραφική παράσταση με τους μέσους κάθε τοποθεσίας. Τέλος στα γραφήματα μπορούμε να παρατηρήσουμε και μια σχετική ομαδοποίηση που υπάρχει μεταξύ των μεταβλητών. Παρακάτω θα αναφερθούμε αναλυτικότερα στην ομαδοποίηση των μεταβλητών.

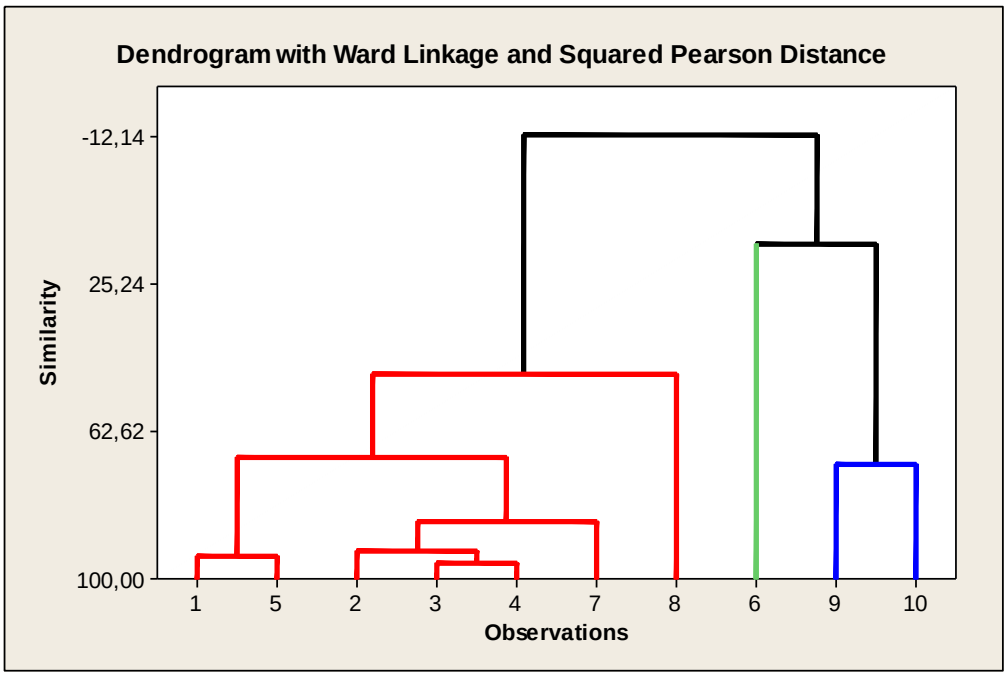


Εικόνα 6.2 : *Γράφημα συσχετίσεων για τις βαθμολογίες των δύο πρώτων κυρίων συνιστωσών.*



Εικόνα 6.3 : *Γράφημα συσχετίσεων για τους μέσους των περιοχών.*

Στη συνέχεια, προκειμένου να κάνουμε μια περεταίρω ομαδοποίηση των δεδομένων μας, θα δουλέψουμε με την ανάλυση κατά συστάδες. Το Minitab μας δίνει το δενδρόγραμμα που βλέπουμε στην Εικόνα 6.3. Από το δενδρόγραμμα μπορούμε να διακρίνουμε τρεις βασικές ομάδες, στις οποίες μπορούμε να χωρίσουμε τις περιοχές.



Εικόνα 6.3 : *Δενδρόγραμμα για την ανάλυση συστάδων.*

Από την Εικόνα 6.2 και την Εικόνα 6.3 καταλήγουμε στο συμπέρασμα ότι οι περιοχές θα πρέπει να χωριστούν στις ακόλουθες κατηγορίες :

- Η τοποθεσία 6
- Οι τοποθεσίες 9 και 10 μαζί και
- Όλες οι υπόλοιπες τοποθεσίες μαζί.

6.4 Προσαρμογή πολυωνυμικού μοντέλου λογιστικής παλινδρόμησης

Όπως περιγράφηκε στην Ενότητα 6.3, προκειμένου να απλοποιήσουμε την μορφή του αρχικού μοντέλου μας θα αντικαταστήσουμε τις αρχικές μας μεταβλητές με εκείνες που προκύπτουν από την ανάλυση κυρίων συνιστωσών.

Επιπλέων θα ομαδοποιήσουμε τις εξαρτημένες μεταβλητές όπως μελετήθηκε στο δενδρόγραμμα (Εικόνα 6.3). Έτσι για το νέο μας μοντέλο οι εξαρτημένες μεταβλητές θα είναι αυτές που φαίνονται στον Πίνακα 6.6 με την εξής κωδικοποίηση

Πίνακας 6.6 : *Κωδικοποίηση περιοχών*

Περιοχή	**Κωδικοποίηση**
Τοποθεσία 6	1
Τοποθεσία 9 κ' 10	2
Λοιπές τοποθεσίες	3

Οι νέες επεξηγηματικές μεταβλητές υπολογίστηκαν με την μεθοδολογία που περιγράφηκε αναλυτικά στην Ενότητα 2.6. Ένα μικρό μέρος του δείγματος που θα χρησιμοποιήσουμε στην μετέπειτα ανάλυση παρουσιάζεται στον Πίνακα 6.7.

Πίνακας 6.7 : *Δείγμα*

A_1	A_2	A_3	**Περιοχή**
2009.881	-272.803	882.333	1
5289.491	16168.681	16100.06	2
1082.583	1121.809	1359.049	2
1073.76	-42.719	533.723	3
977.408	-459.075	637.989	3

Είμαστε τώρα έτοιμοι να προσαρμόσουμε το μοντέλο μας. Εισάγοντας τα δεδομένα μας και ακολουθώντας τη διαδικασία που διαθέτει η R, λαμβάνουμε τα δεδομένα στον Πίνακα 6.8. Λόγω του ότι το μοντέλο μας είναι πολυωνυμικό, η εξαρτημένη μεταβλητή περιέχει 3 διαφορετικές κατηγορίες, έτσι περιμένουμε η R να δημιουργήσει δύο μοντέλα για τα δεδομένα μας. Είναι σημαντικό να σημειωθεί ότι έχουμε επιλέξει ως κατηγορία αναφοράς την 3, δηλαδή τις Λοιπές Τοποθεσίες.

Πίνακας 6.8: *Αποτελέσματα από R*

	Μοντέλο	**Intercepts**	A_1	A_2	A_3
Coefficients	1	-134.73	0.06597	-0.02997	0.01580
	2	-4.02	0.00579	0.00771	-0.00894
Std. Errors	1	1.338e-06	0.003608	0.004795	0.005392
	2	3.901e-06	0.00091	0.001937	0.00213
Residual Deviance: 97.45822 AIC: 113.4582					

Ο δείκτης AIC των μοντέλων μας είναι αρκετά μεγάλος (AIC = 113.46). Είναι οπότε απαραίτητο να κάνουμε τον έλεγχο για την καλή προσαρμογή των μοντέλων. Για τον σκοπό αυτό θα χρησιμοποιήσουμε το Wald test καθώς και τον έλεγχο των p-values. Η R με τις κατάλληλες εντολές μας παρέχει τα δεδομένα που βρίσκονται στους Πίνακες 6.9 και 6.10 για το πρώτο και το δεύτερο μοντέλο αντίστοιχα.

Πίνακας 6.9: *Αποτελέσματα ελέγχου για το πρώτο μοντέλο.*

	Coefficient	**Std. Errors**	**Z stat**	**P value**	**Rate ratios**
Σταθερά	-134.73	1.338 e-06	-1.007 e+08	0.0	3.077 e-59
A_1	0.066	3.608 e-03	1.828 e+01	0.0	1.068 e+00
A_2	-0.03	4.795 e-03	-6.25 e+00	4.11 e-10	9.705 e-01
A_3	0.0158	5.392 e-03	2.93 e+00	3.38 e-03	1.016 e+00

Πίνακας 6.10: *Αποτελέσματα ελέγχου για το δεύτερο μοντέλο.*

	Coefficient	**Std. Errors**	**Z stat**	**P value**	**Rate ratios**
Σταθερά	-4.02	3.901 e-06	-1.03 e+06	0.0	0.018
A_1	0.0058	9.104 e-04	6.35 e+00	2.13 e-10	1.0058
A_2	0.0077	1.937 e-03	3.98 e+00	6.93 e-05	1.0077
A_3	-0.0089	2.13 e-03	-4.2 e+00	2.67 e-05	0.991

Από τα παραπάνω αποτελέσματα βλέπουμε ότι όλες οι p- τιμές των μεταβλητών είναι πολύ μικρές και κοντά στο 0 και για τα δύο μοντέλα. Συνεπώς και οι τρεις επεξηγηματικές μεταβλητές είναι στατιστικά σημαντικές. Στο ίδιο συμπέρασμα καταλήγουμε και από τα αποτελέσματα του ελέγχου Wald.

Λόγω των πολλαπλών εξαρτημένων μεταβλητών και των αποτελεσμάτων στους Πίνακες 6.9 και 6.10, οι προσαρμοσμένες εξισώσεις παλινδρόμησης είναι οι

$$ln\left(\frac{\hat{p}_1}{\hat{p}_3}\right) = -134.73 + 0.066A_1 - 0.03A_2 + 0.0158A_3 = R_1$$

και

$$ln\left(\frac{\hat{p}_2}{\hat{p}_3}\right) = -4.02 + 0.0058A_1 + 0.0077A_2 - 0.0089A_3 = R_2$$

Είναι γνωστό ότι το άθροισμα όλων των πιθανοτήτων θα πρέπει να κάνει 1. Συνεπώς

$$\hat{p}_1 + \hat{p}_2 + \hat{p}_3 = 1 \tag{6.1}$$

και επειδή $\hat{p}_1 = e^{R_1}\hat{p}_3$ και $\hat{p}_2 = e^{R_2}\hat{p}_3$ καταλήγουμε σε μια νέα σχέση των τριών πιθανοτήτων

$$\hat{p}_1 + \hat{p}_2 + \hat{p}_3 = (e^{R_1} + e^{R_2} + 1)\hat{p}_3 \tag{6.2}$$

Συνδυάζοντας τις σχέσεις (6.1) και (6.2) μπορούμε εύκολα να υπολογίσουμε το $\hat{p}_3$ από την σχέση

$$\hat{p}_3 = \frac{1}{e^{R_1} + e^{R_2} + 1} \tag{6.3}$$

Αφού υπολογίσουμε το $\hat{p}_3$ μπορούμε να υπολογίσουμε και τα $\hat{p}_1$ και $\hat{p}_2$.

Στον Πίνακα 6.11 μπορούμε να δούμε τις υπολογισμένες τιμές των πιθανοτήτων και ύστερα να μελετήσουμε τις τιμές της εξαρτημένης μεταβλητής. Ο τρόπος που κατατάσσουμε ένα εύρημα σε μια περιοχή είναι παρόμοιος με αυτόν που χρησιμοποιήσαμε στο απλό λογιστικό μοντέλο. Η μόνη διαφορά είναι ότι εδώ δεν έχουμε μια δίτιμη μεταβλητή και έτσι θα χρειαστεί να μελετήσουμε περισσότερες πιθανότητες (τόσες σε αριθμό, όσες και οι κατηγορίες του Y). Το σκεπτικό μας και εδώ είναι παρόμοιο με τα προηγούμενα κεφάλαια. Όταν η εκάστοτε πιθανότητα πάρει τιμή μεγαλύτερη του 0.5, τότε κατατάσσουμε το εύρημα στην αντίστοιχη περιοχή.

Λόγω του μεγάλου μεγέθους του δείγματος, δεν είναι εφικτό να δώσουμε τα αποτελέσματα ολόκληρου του πληθυσμού. Έτσι στον πίνακα έχουμε ένα μικρό μόνο μέρος αυτού. Το υπόλοιπο δείγμα λειτουργεί με τον ίδιο τρόπο που περιγράψαμε.

Πίνακας 6.11: *Εκτιμημένες πιθανότητες.*

Δείγμα	$\widehat{\boldsymbol{p}}_1$	$\widehat{\boldsymbol{p}}_2$	$\widehat{\boldsymbol{p}}_3$	**Y (πρόβλεψη)**	**Y (πραγμ.τιμή)**
6	9.19 e-24	7.46 e-02	9.25 e-01	3	3
11	4.82 e-23	7.41 e-02	9.26 e-01	3	3
46	7.5 e-34	4.17 e-02	9.58 e-01	3	3
81	1	1.94 e-10	1.97 e-09	1	1
82	1	4.6 e-36	5.08 e-37	1	1
86	1	2.04 e-09	1.25 e-09	1	1
100	1	1.26 e-19	4.45 e-19	1	1
101	1.21 e-42	9.44 e-02	9.06 e-01	3	3
120	7.29 e-21	5.18 e-04	9.995 e-01	3	3
139	1.96 e-43	1.06 e-02	9.895 e-01	3	3
140	3.18 e-11	9.996 e-01	3.83 e-04	2	2
151	5.44 e-31	9.38 e-01	6.192 e-02	2	2
152	9.28 e-28	9.13 e-01	8.67 e-02	2	2
176	4.39 e-23	4.02 e-01	5.98 e-01	3	2
179	4.3 e-29	3.4 e-01	6.6 e-01	3	2

Από τα αποτελέσματα και τον Πίνακα 6.11 μπορούμε να συμπεράνουμε ότι το μοντέλο μας έχει αρκετά καλή προβλεπτική ικανότητα. Η μόνη περίπτωση που το μοντέλο μας μπορεί να

«υποθέσει» λάθος είναι όταν το εύρημα ανήκει στην περιοχή 10 (δείγματα 176, 179). Σε αυτήν την περίπτωση καλό θα ήταν να κάνουμε κάποιον περεταίρω έλεγχο.

Η ποσότητα Rate ratios, που παίρνουμε από τα αποτελέσματα στην R, είναι ο παράγοντας exp($\hat{\beta}_j$) επί τον οποίο πολλαπλασιάζεται η σχετική πιθανότητα ένα δείγμα να ανήκει στην ομάδα 1 ή 2, όταν η εκάστοτε ανεξάρτητη μεταβλητή A_j αυξηθεί κατά μία μονάδα, με δεδομένο ότι οι υπόλοιπες μεταβλητές παραμένουν σταθερές. Αν το $\hat{\beta}_j$ είναι θετικό, ο παράγοντας exp($\hat{\beta}_j$) είναι μεγαλύτερος από την μονάδα και η σχετική πιθανότητα αυξάνεται, ενώ αν το $\hat{\beta}_j$ είναι αρνητικό, ο παράγοντας exp($\hat{\beta}_j$) είναι μικρότερος της μονάδας και η σχετική πιθανότητα μειώνεται.

Κατασκευάζουμε 97.5% διαστήματα εμπιστοσύνης για της παραμέτρους του μοντέλου και των exp($\hat{\beta}_j$). Τα διαστήματα αυτά παρουσιάζονται στον Πίνακα 6.12.

Πίνακας 6.12: *Διαστήματα εμπιστοσύνης.*

	$\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}}$ **(2.5%)**		**exp(**$\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}}$**) (97.5%)**	
	Μοντέλο 1	Μοντέλο 2	Μοντέλο 1	Μοντέλο 2
Σταθερά	3.077 e-59	0.018	3.0766 e-59	0.018
$\boldsymbol{A_1}$	1.0607 e+00	1.004	1.076 e+00	1.0076
$\boldsymbol{A_2}$	9.614 e-01	1.004	9.796 e-01	1.0116
$\boldsymbol{A_3}$	1.005 e+00	0.987	1.0267 e+00	0.9952

7 ΣΤΑΤΙΣΤΙΚΗ ΜΕΛΕΤΗ ΣΕ ΕΞΩΠΛΑΝΗΤΕΣ

7.1 Παρουσίαση δείγματος και μεταβλητών

Οι εξωπλανήτες είναι πλανήτες εκτός του ηλιακού μας συστήματος. Ο πρώτος εξωπλανήτης ανακαλύφθηκε το 1995 από τον Mayor και τον Queloz. Ο πλανήτης ήταν παρόμοιος σε μάζα με τον Δία και βρέθηκε σε τροχιά γύρω από το αστέρι 51 Πήγασος. Στην περίοδο που μεσολάβησε ανακαλύφθηκαν πάνω από εκατό εξωπλανήτες. Σχεδόν όλοι ανιχνεύτηκαν χρησιμοποιώντας την βαρυτική επίδραση που ασκούν στα κεντρικά τους αστέρια. Ένας εκπληκτικός αριθμός εξωπλανητών και ανακαλύψεις γύρω από αυτούς δίνεται από τους Mayor και Frei (2003).

Από τις ιδιότητες των εξωπλανητών που βρέθηκαν μέχρι τώρα φαίνεται ότι η θεωρία της πλανητικής ανάπτυξης που έχει κατασκευαστεί για τους πλανήτες του Ηλιακού συστήματος μπορεί να χρειάζεται αναδιατύπωση. Οι εξωπλανήτες δεν ομοιάζουν με τους γνωστούς μας εννιά πλανήτες. Το πρώτο βήμα κατανόησης των εξωπλανητών είναι να προσπαθήσουμε να τους κατατάξουμε σε ομάδες ανάλογα με τις γνωστές τους ιδιότητες.

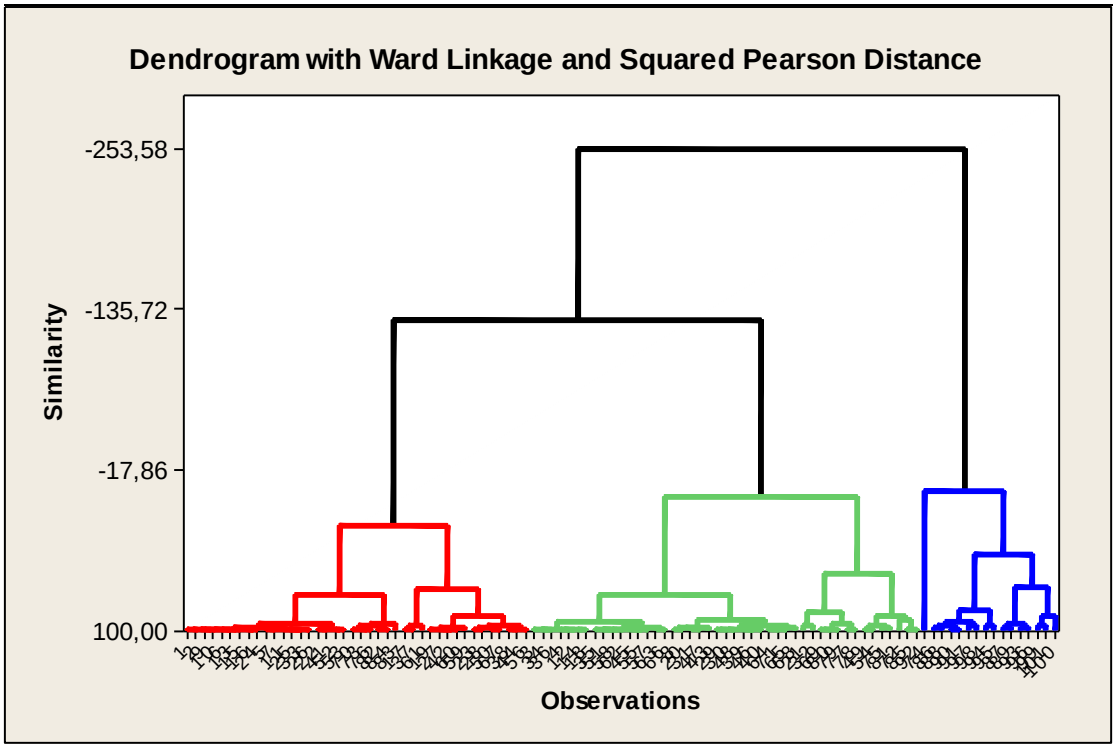
Θα διερευνήσουμε τη δομή των εξωπλανητικών δεδομένων χρησιμοποιώντας διάφορες μεθόδους ανάλυσης. Ένα μικρό δείγμα των δεδομένων δίνεται στον Πίνακα 7.1. Στα δεδομένα δίνεται η μάζα (σε σχέση με εκείνη του Δία), η περίοδος (σε σχέση με εκείνη της Γης) και η εκκεντρικότητα των εξωπλανητών, που ανακαλύφθηκαν μέχρι και τον Οκτώβριο του 2002. Όλες οι επεξηγηματικές μεταβλητές μας είναι συνεχείς.

Πίνακας 7.1: *Δείγμα εξωπλανητών.*

Id	Μάζα	Περίοδος	Εκκεντρικότητα
1	0.120	4.950.000	0.00
2	0.197	3.971.000	0.00
3	0.210	44.280.000	0.34
4	0.220	75.800.000	0.28
5	0.230	6.403.000	0.08

7.2 Εφαρμογή ανάλυσης συστάδων

Στην περίπτωση των εξωπλανητών σκοπός μας δεν είναι η απλοποίηση του μοντέλου μας. Θα χρησιμοποιήσουμε, όμως, την ανάλυση συστάδων γιατί θέλουμε να συνδυάσουμε τα δεδομένα μας με τέτοιο τρόπο ώστε να πάρουμε είδη/ομάδες πλανητών ανάλογα με τις ιδιότητές τους.
Το δενδρόγραμμα για τους πλανήτες φαίνεται αρκετά χρήσιμο στην κατανόηση των συστάδων. Στην Εικόνα 7.1 μπορούμε να διακρίνουμε ξεκάθαρα τις ομάδες που σχηματίζονται. Η διαδικασία που ακολουθεί το Minitab είναι η σύγκριση των πλανητών ανά δύο, μέχρις ότου να καταλήξει, εν τέλει, σε μία και μόνο συστάδα.



Εικόνα 7.1 : *Δενδρόγραμμα συστάδων για τους εξωπλανήτες.*

Η απόφαση για τον βέλτιστο αριθμό ομάδων είναι αρκετά δύσκολη και δεν υπάρχει καμία μέθοδος που να μπορεί να συνιστάται σε όλες τις περιπτώσεις (βλέπε Everitt et al., 2001). Η επιλογή των τριών ομάδων έγινε για λόγους διευκόλυνσης, κυρίως για την επεξήγηση των αποτελεσμάτων. Θα μπορούσαμε βέβαια να κάνουμε μια αναλυτικότερη ομαδοποίηση (5 ομάδων), αλλά τότε θα έπρεπε να έχουμε πολύ καλές γνώσεις πάνω στην αστρονομία, προκειμένου να είμαστε σε θέση να εξηγήσουμε με επιτυχία τα αποτελέσματα.

Ένας τρόπος διασύνδεσης των στοιχείων είναι ο k-means, κατά τον οποίο η ομαδοποίηση γίνεται με έναν συγκεκριμένο αριθμό k συστάδων. Ο αριθμός των ομάδων είναι γνωστός από πριν. Με έναν επαναληπτικό αλγόριθμο μοιράζουμε τις παρατηρήσεις στις ομάδες ανάλογα με το ποια ομάδα ταιριάζει πιο καλά στην παρατήρηση (Aldenderfer και Blashfield, 1984). Βασιζόμενοι στο δενδρόγραμμα (Εικόνα 7.1) και στην τακτική διασύνδεσης των

στοιχείων k-means, δημιουργούμε 3 συστάδες. Προσαρμόζοντας τα δεδομένα μας στο Minitab λαμβάνουμε πληροφορίες για τις συστάδες και μια αρχική εικόνα για τον τρόπο διαχωρισμό των πλανητών (Πίνακας 7.2).

Πίνακας 7.2 : *Αποτελέσματα Minitab.*

	Αριθμός παρατηρήσεων	**Άθροισμα τετραγώνων**	**Μέση τιμή απόστασης από το κ. β.**	**Μέγιστή τιμή απόστασης από το κ. β.**
Cluster 1	40	20931249	585.40	2099.48
Cluster 2	45	13057099	384.25	2010.01
Cluster 3	16	29386312	1007.72	3867.55

Όπου κ.β. είναι το κέντρο βάρους. Στον Πίνακα 7.2 βλέπουμε ότι το Minitab, βασιζόμενο στα στοιχεία που του δώσαμε, δημιουργεί τρεις διαφορετικές ομάδες. Οι δύο πρώτες είναι αρκετά μεγάλες πληθυσμιακά (40 και 45 πλανήτες αντίστοιχα), ενώ η τρίτη ομάδα είναι αρκετά μικρότερη από τις άλλες δύο (16 πλανήτες). Στον Πίνακα 7.3 παρουσιάζονται αναλυτικά οι τρεις ομάδες.

Πίνακας 7.3 : *Συσχετίσεις συστάδων και μεταβλητών*

Μεταβλητές	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Grand centroid**
Μάζα	1.665	2.364	10.19	3.327
Περίοδος	514.518	507.994	1492.45	666.531
Εκκεντρότητα	0.082	0.424	0.38	0.282

Βλέπουμε ότι στην πρώτη ομάδα μπορούμε να κατατάξουμε όλους τους πλανήτες που έχουν περίπου την ίδια μάζα με εκείνη του Δία, σχετικά μικρή περίοδο και πολύ μικρή εκκεντρότητα. Στην δεύτερη ομάδα κατατάσσουμε μεγαλύτερους πλανήτες με μικρή περίοδο και μεγάλη εκκεντρότητα, ενώ στην μικρότερη σε πληθυσμό ομάδα βρίσκουμε ογκώδεις πλανήτες με ακραίες περιόδους και μέτρια εκκεντρότητα. Την διαφορά που έχει η τρίτη ομάδα από τις άλλες δύο την βλέπουμε και στον Πίνακα 7.4.

Πίνακας 7.4 : *Αποστάσεις μεταξύ των ομάδων.*

	Cluster 1	**Cluster 2**	**Cluster 3**
Cluster 1	0.000	6.571	977.971
Cluster 2	6.571	0.000	984.489
Cluster 3	977.971	984.489	0.000

7.3 Προσαρμογή πολυωνυμικού μοντέλου λογιστικής παλινδρόμησης

Στο σημείο αυτό θα προσπαθήσουμε να δημιουργήσουμε το καλύτερο στατιστικά μοντέλο, το οποίο θα κατατάσσει κάθε νέο εξωπλανήτη σε μια από τις τρεις ομάδες, ανάλογα πάντα με την μάζα, την περίοδο και την εκκεντρότητά του. Την διαδικασία αυτή θα την πραγματοποιήσουμε με την βοήθεια του στατιστικού πακέτου της R. Υπενθυμίζουμε ότι το μοντέλο που θα δημιουργήσουμε βασίζεται πάνω στις συστάδες που αναπτύξαμε στο προηγούμενο κεφάλαιο. Έτσι κάθε δοσμένος εξωπλανήτης έχει κατηγοριοποιηθεί, όπως περιγράψαμε, σε μια συστάδα (cluster).

Λόγω του ότι στο αρχικό μας μοντέλο η εξαρτημένη μεταβλητή έχει 3 κατηγορίες (3 συστάδες), είναι αναμενόμενο η R να δημιουργήσει δύο διαφορετικά μοντέλα για τα δεδομένα μας. Παίρνοντας ως κατηγορία αναφοράς την Ομάδα 1, τα δύο μοντέλα θα έχουν ως εξαρτημένη μεταβλητή την Ομάδα 2 και 3 αντίστοιχα.

Πίνακας 7.5 : *Αποτελέσματα R*.

	Μοντέλο	**Intercepts**	**Μάζα**	**Περίοδος**	**Εκκεντρότητα**
Coefficients	1 (Cluster 2)	-20.48595	0.2802999	0.005674665	51.68424
	2 (Cluster 3)	-22.20479	2.1114371	0.007504992	26.08303
Std. Errors	1 (Cluster 2)	1.110198	0.4156757	0.0009357108	0.36926171
	2 (Cluster 3)	0.1069995	0.3015832	0.0006268853	0.09532173
Residual Deviance: 24.20916 AIC: 40.20916					

Στα αποτελέσματα του Πίνακα 7.5 βλέπουμε έναν αρκετά χαμηλό δείκτη AIC (AIC = 40.21). Αυτό σημαίνει ότι και τα δύο μοντέλα, εκ πρώτης όψεως φαίνεται να είναι στατιστικά καλά. Παρ' όλα αυτά θα κάνουμε κάποιους περεταίρω ελέγχους, προκειμένου να επιβεβαιώσουμε την ορθότητα των μοντέλων μας.

Τα αποτελέσματα των ελέγχων Wald test και p-value, που έγιναν μέσω της R, για τα δύο μοντέλα βρίσκονται στους Πίνακες 7.6 και 7.7.

Πίνακας 7.6: *Αποτελέσματα ελέγχου για το πρώτο μοντέλο.*

	Coefficient	**Std. Errors**	**Z stat**	**P value**	**Rate ratios**
Σταθερά	-20.486	1.11019	-18.4526	0.00	1.2678 e-09
Μάζα	0.2803	0.41568	0.67432	0.5	1.3235 e+00
Περίοδος	0.00567	0.00094	6.06455	1.32e-09	1.0057 e+00
Εκκεντρότητα	51.684	0.36926	139.966	0.00	2.7937 e+22

Πίνακας 7.7: *Αποτελέσματα ελέγχου για το δεύτερο μοντέλο.*

	Coefficient	**Std. Errors**	**Z stat**	**P value**	**Rate ratios**
Σταθερά	-22.2048	0.107	-207.522	0.00	2.27 e-10
Μάζα	2.1114	0.3016	7.001	2.538 e-12	8.26
Περίοδος	0.0075	0.00063	11.9719	0.00	1.008
Εκκεντρότητα	26.0830	0.0953	273.631	0.00	2.13 e+11

Από τα παραπάνω αποτελέσματα βλέπουμε ότι σχεδόν όλες οι μεταβλητές μας έχουν πολύ χαμηλή p-τιμή (σχεδόν μηδέν). Η μεταβλητή της μάζας, στο πρώτο μόνο μοντέλο (Πίνακας 7.6), φαίνεται να είναι στατιστικά μη σημαντική. Παρ' όλα αυτά, μπορούμε να την συμπεριλάβουμε στο μοντέλο μας λόγω της σημαντικότητας της στο δεύτερο μοντέλο (Πίνακας 7.7) και της καλής τιμής του δείκτη AIC, που βρήκαμε παραπάνω.

Με βάση τα αποτελέσματα στον Πίνακα 7.6 και στον Πίνακα 7.7 οι προσαρμοσμένες εξισώσεις παλινδρόμησης είναι

$$ln\left(\frac{\hat{p}_2}{\hat{p}_1}\right) = -20.486 + 0.28 * \mu\acute{\alpha}\zeta\alpha + 0.0057 * \pi\varepsilon\rho\acute{\iota}o\delta o\varsigma + 51.684 * \varepsilon\kappa\kappa\varepsilon\nu\tau\rho\acute{o}\tau\eta\tau\alpha = R_1$$

και

$$ln\left(\frac{\hat{p}_3}{\hat{p}_1}\right) = -22.205 + 2.112 * \mu\acute{\alpha}\zeta\alpha + 0.0075 * \pi\varepsilon\rho\acute{\iota}o\delta o + 26.083 * \varepsilon\kappa\kappa\varepsilon\nu\tau\rho\acute{o}\tau\eta\tau\alpha = R_2$$

Χρησιμοποιώντας τις σχέσεις (6.1) – (6.3) μπορούμε να υπολογίσουμε τις πιθανότητες $\hat{p}_1$, $\hat{p}_2$ και $\hat{p}_3$ καθώς και την εξαρτημένη μεταβλητή Y. Σε αυτήν την περίπτωση έχουμε χρησιμοποιήσει ως κατηγορία αναφοράς την 1 και έτσι στους τύπους την θέση του $\hat{p}_3$ παίρνει το $\hat{p}_1$. Επομένως οι σχέσεις που χρησιμοποιούμε εδώ έχουν την μορφή

$$\hat{p}_1 = \frac{1}{e^{R_1} + e^{R_2} + 1}$$

και

$$\hat{p}_2 = e^{R_1}\hat{p}_1 \quad \text{ή} \quad \hat{p}_3 = e^{R_1}\hat{p}_1$$

Όπως έχει γίνει γνωστό, όταν η πιθανότητα έχει τιμή μεγαλύτερη από το 0.5, τότε ο εξωπλανήτης ανήκει στην κατηγορία που υποδηλώνει η πιθανότητα, ενώ όταν έχει μικρότερη τιμή από 0.5, στρεφόμαστε στις άλλες δύο πιθανότητες. Δηλαδή όταν έχουμε $\hat{p}_i > 0.5$, τότε $Y = i$, οπότε στατιστικά είναι πολύ πιθανό ο εξωπλανήτης να ταιριάζει στην κατηγορία i.

Τα αποτελέσματα των πιθανοτήτων καθώς και της εξαρτημένης μεταβλητής δίνονται στον πίνακα 7.8. Οι υπολογισμοί έγιναν με την χρήση της R.

Πίνακας 7.8: *Εκτιμημένες πιθανότητες.*

Δείγμα	$\widehat{\boldsymbol{p}}_1$	$\widehat{\boldsymbol{p}}_2$	$\widehat{\boldsymbol{p}}_3$	**Y (πρόβλεψη)**	**Y (πραγμ.τιμή)**
1	1	1.35 e-09	3.04 e-10	1	1
8	4.38 e-03	9.96 e-01	1.1 e-06	2	2
11	1	1.96 e-08	2.37 e-09	1	1
16	1	1.57 e-09	9.91 e-10	1	1
17	1.65 e-01	1.2 e-01	7.14 e-01	3	3
20	1.32 e-04	1	1.33 e-06	2	2
21	1.23 e-11	1	1.78 e-05	2	2
36	9.63 e-01	3.68 e-02	2.46 e-05	1	1
37	5.41 e-01	2.8 e-03	4.56 e-01	1 ή 3	3
46	2.42 e-02	9.76 e-01	2.13 e-04	2	2
51	9.99 e-01	8.67 e-04	6.01 e-05	1	1
60	1	2.05 e-06	2.57 e-05	1	1
61	1.26 e-02	9.18 e-02	8.96 e-01	3	3
100	7.83 e-15	2.99 e-12	1	3	3
101	7.91 e-13	2.47 e-09	1	3	3

Το μοντέλο που κατασκευάσαμε δείχνει να έχει πολύ καλή προβλεπτική ικανότητα, αφού κατατάσσει τους πλανήτες στην καταλληλότερη κατηγορία με πιθανότητα περίπου 93%.

Με την ίδια διαδικασία που ακολουθήσαμε και στην Ενότητα 6.4 κατασκευάζουμε 97.5% διαστήματα εμπιστοσύνης για της παραμέτρους του μοντέλου ($\hat{\beta}_j$) και των exp($\hat{\beta}_j$). Τα διαστήματα αυτά παρουσιάζονται στον Πίνακα 7.9.

Πίνακας 7.9: *Διαστήματα εμπιστοσύνης.*

	$\widehat{\boldsymbol{\beta}}_j$ **(2.5%)**		exp($\widehat{\boldsymbol{\beta}}_j$) **(97.5%)**	
	Μοντέλο 1	Μοντέλο 2	Μοντέλο 1	Μοντέλο 2
Σταθερά	1.439 e-10	1.843 e-10	1.117 e-08	2.81 e-10
Μάζα	5.860 e-01	4.574 e+00	2.99 e+00	1.492 e+01
Περίοδος	1.004 e+00	1.006 e+00	1.008 e+00	1.009 e+00
Εκκεντρότητα	1.355 e+22	1.764 e+11	5.76 e+22	2.564 e+11

Παράρτημα 1:
Εντολές για το MINITAB

Στο Minitab η ανάλυση ενός μοντέλου λογιστικής παλινδρόμησης εκτελείται από την γραμμή εργαλείων με την εντολή ***Stat→Regression→Binary Logistic Regression***.

Στη συνέχεια, για δυαδικά δεδομένα, η εξαρτημένη μεταβλητή ($y = 0$ ή 1) εισάγεται στο πλαίσιο ***Response***. Για διωνυμικά δεδομένα, ο αριθμός επιτυχιών (γεγονότων) εισάγεται στο πλαίσιο ***Success*** και ο αριθμός δοκιμών στο πλαίσιο ***Trial.*** Οι επεξηγηματικές μεταβλητές εισάγονται στο πλαίσιο ***Model***. (Χ. Καρώνη, 2010).

Για τα μοντέλα πολυωνυμικής λογιστικής παλινδρόμησης οι αντίστοιχες εντολές είναι ***Stat→Regression→Nominal Logistic Regression***.

Η ανάλυση κυρίων συνιστωσών εκτελείται από την γραμμή εργαλείων με τις εντολές ***Stat→Multivariate→Principal Components Analysis***.

Στη συνέχεια εισάγουμε στο πλαίσιο ***Variables*** τις επεξηγηματικές μεταβλητές και επιλέγουμε τον αριθμό των συνιστωσών που επιθυμούμε. Για τις γραφικές παραστάσεις πηγαίνουμε στο ***Principal Components Analysis – Graphs*** και επιλέγουμε τα ***Scree plot*** και ***Score plot for first 2 components***.

Παράρτημα 2:
Ο κώδικας που χρησιμοποιήθηκε στην R

Εντολές για τα μοντέλα λογιστικής παλινδρόμησης

Οι αναγραφόμενες εντολές έδωσαν τα αποτελέσματα του κεφαλαίου 5 για τα ιατρικά δεδομένα. Οι περισσότερες από τις εντολές αυτές χρησιμοποιήθηκαν και στο Κεφάλαιο 4 με μικρές αλλαγές, κυρίως στα ονόματα.

❖ Προετοιμασία των δεδομένων

```
> cc<-raed.table("dioxin.txt", header=TRUE)

> attach(cc)

> cc
```

❖ Προσαρμογή του μοντέλου

```
> mod1<-glm(y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14+x15+x16+x17,
family=binomial)

> summary(mod1)
```

❖ Διαδικασία διαδοχικής αφαίρεσης (backward elimination)

```
> step(mod1,direction="backward", test="Chisq")
```

❖ Σύγκριση των τιμών της ελεγχοσυνάρτησης deviance

```
> mod2<-glm(y~x1+x3+x6+x8+x10+x12+x13+x15, family=binomial)

> summary(mod2)
```

```
>mod4<-glm(y~x1+x3+x6+x8+x10+x15, family=binomial)

> summary(mod4)

> anova(mod2,mod4, test="Chisq")
```

❖ Καμπύλη ROC

```
> library(pROC)

> roc(y, fitted.values(mod2), plot=TRUE)
```

Εντολές για τα μοντέλα της πολυωνυμικής λογιστικής παλινδρόμησης

Οι εντολές που αναφέρουμε μας έδωσαν τα αποτελέσματα του κεφαλαίου 6. Οι ίδιες εντολές, με τις απαραίτητες τροποποιήσεις χρησιμοποιήθηκαν και στο Κεφάλαιο 7.

❖ Προετοιμασία των δεδομένων

```
> library(nnet)

> flints<-read.table("flints-3.txt",header=TRUE)

> attach(flints)

> flints
```

❖ Δήλωση κατηγορικών μεταβλητών και κατηγορία αναφοράς

```
>area<-factor(area,c("3","1","2"))
```

❖ Προσαρμογή του μοντέλου

```
> mod<-multinom(area~A1+A2+A3)
```

```
> summary(mod)

> output<-summary(mod)
```

❖ Wald test και p-values

```
> z <- output$coefficients/output$standard.errors

> p <- (1 - pnorm(abs(z), 0, 1))*2

> model1<-cbind(output$coefficients[1,],output$standard.errors[1,],z[1,],p[1,],
exp(output$coefficients[1,]))

> colnames(model1) <- c("Coefficient","Std. Errors","z stat","p value", "Rate ratios")

> model1

> model2<-cbind(output$coefficients[2,],output$standard.errors[2,],z[2,],p[2,],
exp(output$coefficients[2,]))

> colnames(model2) <- c("Coefficient","Std. Errors","z stat","p value", "Rate ratios")

> model2
```

❖ Διαστήματα εμπιστοσύνης

```
>exp(confint(mod))
```

Βιβλιογραφία

Aldenderfer M.S. & Blashfield R.K. (1984), Cluster Analysis. *Sage Publications*, Newbury Park

Bartholomew D. J., Steele F., Moustaki I. and Galbraith I. J. (2008)*, Analysis of Multivariate Social Science Data*, Taylor & Francis Group LLC

Bigod, A. (1950), Origine des Silex des Stations Prehistoriques de Sonmont et d'Olendon (Calvados), *Bull. de la SOC*, Linndenne de Normandie, 9 serie, VI, 62-68.

Cox, D. R. and Hinkley D. V. (1974), *Theoretical Statistics*, Chapman & Hall, London

Cox D. R. and Snell E. J. (1989), *Analysis of Binary Data*, Second Edition, Chapman & Hall, London

Deflandre G. (1966), etude Micropaleontologique des Silex du Site de Pincevent, *Gallia Prehistoire IX*, 380-381

Dobson A. (2002), *An Introduction to Generalized Linear Models*, SecondEdition, Chapman & Hall, London

Everitt B. S. (2001), *Statistics for Psychologists*, Mahwah, New Jersey, USA: Lawrence Erlbaum.

Everitt B. S., Landau S., and Leese M. (2001), *Cluster Analysis*, 4th edition,
London, UK: Arnold.

Everitt S. B. and Hothorn T. (2010), *A Handbook of Statistical Analyses Using R*, Taylor & Francis

Hartigan J. (1975), *Clustering Algorithms,* Wiley, New York.

Hauck W. W. and Donner A. (1977), Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, **72**, 851–853.

Hosmer W. D. Jr., Lemeshow S. and Sturdivant X. R. (2013), *Applied Logistic Regression*, John Wiley & Sons

Hurst V. J. and Kelly A. R. (1966), Patination of Cultural Flints. In Caldwell, J. R. (ed.), *New Roads to Yesterday*, London, esp. 51 8-520.

Jolliffe I. T. (1972), Discarding Variables in a Principal Component Analysis. I: Artificial Data, *Journal of the Royal Statistic Society, **21**, 160-173*

Jowell R. and the Central Co-ordinating Team (2003), *European Social Survey 2002/2003: Technical Report*, Centre for Comparative Social Surveys, City University

Laming A. (1952), Les Microorganisms des Silex. In Laming, A. (ed.), Paris, *La DPcouverte du Pass* &, 263-7.

McIntyre R. M. and Blashfield (1980), "A Nearest-Centroid Technique for Evaluating the Minimum-Variance Clustering Procedure", *Multivariate Behavioral Research*, 15, 225-238.

Mayor M. and Frei P. (2003), The Discovery of Exoplanets, *New Worlds in the Cosmos*, Cambridge, UK: Cambridge University Press

Mayor M. and Queloz D. (1995), "A Jupiter-mass companion to a solar-type star," *Nature*, 378, 355

Rao C. R. (1973), *Linear Statistical Inference and its Application*, Second Edition, Wiley Inc., New York.

Rankine W. F. (1952), Implements of coloured flint in Britain, *Archaeological News Letter 4* (lo), 145-149.

Sharma S. (1996), *Applied Multivariate Techniques*, John Wiley & Sons

Sneath P. and Sokal R. (1973), *Numerical Taxonomy,* Freeman, San Francisco.

Shotton F. W. (1969), Petrological Examination, *Science and Archaeology*, Brothwell, D. and Higgs, E. S. (eds.), rev. edn., London, pp. 571-511

Sieveking G. de G., Bush P., Ferguson J., Craddock T. P., Hughes J. M., Cowell R. M. (1972), Prehistoric Flint Mines and Their Identification as Sources of Raw Material, *Archaeometry 14(2),* 151-176

Sieveking G. de G., Craddock P., Hughes M. J., Bush P. and Ferguson J. (1970), Characterization of flint mine products, *Nature* 228, 251-254.

Wetzel O. (1941), Mikropalaontologische Untersuchlingem an baltischen Feuerstein, *Quartiir* 111, 127-1 31.

Καρώνη Χ. και Οικονόμου Π. (2017), *Στατιστικά Μοντέλα Παλινδρόμησης*, Εκδόσεις Συμεών