ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΝΑΛΟΓΙΣΤΙΚΩΝ - ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΩΝ ΜΑΘΗΜΑΤΙΚΩΝ

«Πολυμεταβλητή Ανάλυση Κατηγορικών Δεδομένων: Εφαρμογές στη Βιοστατιστική»

Διπλωματική Εργασία για το Μεταπτυχιακό Πρόγραμμα Σπουδών

Η παρούσα Εργασία εκπονήθηκε
ως μερική ικανοποίηση των απαιτήσεων για την απόκτηση
του αντιστοίχου τίτλου σπουδών στην
Στατιστική και Αναλογιστικά-Χρηματοοικονομικά Μαθηματικά

Στελεκάτη Ιωάννα

30 Δεκεμβρίου 2022

ΣΑΜΟΣ

Στελεκάτη Ιωάννα

«Πολυμεταβλητή Ανάλυση Κατηγορικών Δεδομένων: Εφαρμογές στη Βιοστατιστική»

30 Δεκεμβρίου 2022

Διπλωματική Εργασία για το Μεταπτυχιακό Πρόγραμμα Σπουδών

Τμήμα Στατιστικής και Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών

Συγγραφέας: Στελεκάτη Ιωάννα, Α.Μ: 3332020025

Επιβλέπων: Ζήμερας Στέλιος, Αναπληρωτής Καθηγητής
Μέλος Επιτροπής: Καραγρηγορίου Αλέξανδρος, Καθηγητής
Μέλος Επιτροπής: Ρακιτζής Αθανάσιος, Επίκουρος Καθηγητής

ΣΑΜΟΣ

Ευχαριστίες

Στην έναρξη αυτής της εργασίας θα ήθελα να αποδώσω ευχαριστίες σε κάποια άτομα που χωρίς την βοήθειά και την υπομονή τους, δε θα ήταν δύσκολο να έρθει εις πέρας.

Αρχικά, θα ήθελα να ευχαριστήσω θερμά τον επιβλέπων καθηγητή της διπλωματικής μου εργασίας, τον κο. Ζήμερα Στέλιο. Οι συζητήσεις μας, η καθοδήγηση και οι συμβουλές του ήταν καθοριστικές για την επίτευξη της παρούσας διπλωματικής εργασίας.

Επιπρόσθετα, θα ήθελα να ευχαριστήσω τα υπόλοιπα μέλη της τριμελούς επιτροπής της διπλωματικής μου εργασίας, τον κο Καραγρηγορίου Αλέξανδρο, καθηγητή του Πανεπιστημίου Αιγαίου, του τμήματος Στατιστικής και Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών και τον κο Ρακιτζή Αθανάσιο, επίκουρο καθηγητή του Πανεπιστημίου Πειραιώς, του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης.

Τέλος, ένα μεγάλο ευχαριστώ στην οικογένειά μου, τους φίλους μου αλλά και τους συμφοιτητές μου για την υποστήριξη και τη συμπαράσταση που μου παρείχαν όλο αυτό το χρονικό διάστημα.

Περίληψη

Τα κατηγορικά δεδομένα διαδραματίζουν σημαντικό ρόλο σε διάφορους τομείς, όπως στη βιοϊατρική, από τις κοινωνικές έως και πολιτικές επιστήμες, στη διαφήμιση, ακόμη και στον έλεγχο ποιότητας.

Στην παρούσα εργασία, μελετώνται τρόποι ανάλυσης κατηγορικών τυχαίων μεταβλητών με δύο ή περισσότερες κατηγορίες. Αρχικά, παρουσιάζεται ο πίνακας συνάφειας για την εύρεση συσχέτισης για δύο κατηγορικές μεταβλητές και στη συνέχεια γίνεται ανάλυση πινάκων συνάφειας για τρεις η περισσότερες κατηγορικές μεταβλητές. Για την ανάλυση και την ερμηνεία κατηγορικών δεδομένων με πολλαπλές κατηγορίες απαιτούνται κάποιες στατιστικές τεχνικές, όπως ο έλεγχος ανεξαρτησίας. Επίσης, παρουσιάζονται και μέτρα συσχέτισης, τα οποία χαρακτηρίζουν την ισχύ της σχέσης, κάτι που είναι σημαντικό να εξετάζεται σε εφαρμογές με πραγματικά δεδομένα.

Στη συνέχεια, γίνεται αναφορά σε κάποια συγκεκριμένα μοντέλα που χρησιμοποιούνται για να ερμηνεύσουν την σχέση μεταξύ δύο ή περισσότερων μεταβλητών και ταυτόχρονα καθορίζονται και οι μορφές των μοντέλων.

Τέλος, παρουσιάζεται μια εφαρμογή ανάλυσης κατηγορικών δεδομένων στο σύνολο δεδομένων Framingham, η οποία ξεκίνησε το 1948 και έχει σκοπό να ερευνήσει τους παράγοντες που οδηγούν σε καρδιοπάθειες.

Λέξεις Κλειδιά: Κατηγορικά Δεδομένα, Πίνακας Συνάφειας, Μέτρα Συσχέτισης, Πολυμεταβλητή Ανάλυση, Δειγματοληψία Μοντέλου.

Abstract

Categorical data plays an important role in various fields, such as biomedicine, from social to political science, advertising, and even quality control.
In this thesis, ways of analyzing categorical random variables with two or more categories are studied.
Firstly, the contingency table is presented for two categorical variables and then correlation matrices are analyzed for three or more categorical variables.
Analyzing and interpreting categorical data with multiple categories requires some statistical techniques, such as independence testing. Correlation measures are also presented, which characterize the strength of the relationship, which is important to take this into consideration in applications with real data.
Then, a reference is made to some specific models that are used to interpret the relationship between two or more variables and at the same time the forms of the models are determined.
Finally, is presented an application of categorical data analysis to the Framingham dataset, which began in 1948 kai is intented to investigate factors leading to heart disease.

Key Words: Categorical Data, Contingency Table, Correlation Matrix, Multivariate Analysis, Model Sampling

Περιεχόμενα

Κατάλογος Γραφημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1. Εισαγωγή στα κατηγορικά δεδομένα

Οι μεταβλητές που μπορούν να λάβουν μόνο έναν περιορισμένο (και συνήθως σταθερό) αριθμό δυνατών τιμών αναφέρονται ως κατηγορικές μεταβλητές. Οι κατηγορικές μεταβλητές χαρακτηρίζονται από κλίμακες μέτρησης που χωρίζονται σε μια σειρά κατηγοριών. Οι μεταβλητές αυτές μπορούν να εφαρμοστούν σε πολλούς τομείς, όπως στις επιστήμες υγείας (για παράδειγμα εάν ένας ασθενής μπορεί να επιβιώσει από μια επέμβαση, -ναι ή όχι-), στις κοινωνικές επιστήμες (όπου αφορούν τη μέτρηση στάσεων και απόψεων), στις επιστήμες της συμπεριφοράς (σχετικά με την διάγνωση του τύπου ψυχικής ασθένειας, σχιζοφρένεια, κατάθλιψη, νεύρωση), στη δημόσια υγεία (εάν η αυξημένη ευαισθητοποίηση για το AIDS έχει οδηγήσει σε αυξημένη χρήση προφυλακτικών, -ναι ή όχι-), στη ζωολογία (αν για παράδειγμα η κύρια τροφή των αλιγατόρων, είναι τα ψάρια, τα ασπόνδυλα ή τα ερπετά) και στην εκπαίδευση (όπως στις μεθόδους εξετάσεων). Η χρήση κλιμάκων μέτρησης είναι επίσης κοινή σε ποσοτικούς τομείς, όπως οι επιστήμες της μηχανικής και ο βιομηχανικός ποιοτικός έλεγχος, όπου τα προϊόντα κατηγοριοποιούνται ανάλογα με το εάν συμμορφώνονται ή όχι με ορισμένους κανόνες ή προδιαγραφές.

Υπάρχουν δύο τύποι κατηγορικών μεταβλητών που χρησιμοποιούνται συχνά, οι ονομαστικές και οι διατακτικές κατηγορικές μεταβλητές. Οι ονομαστικές μεταβλητές αποτελούνται από ένα εύρος ασυμβίβαστων κατηγοριών που δεν χαρακτηρίζονται από την ιδιότητα της διάταξης. Επιπλέον, οι ονομαστικές αυτές μεταβλητές χωρίζονται σε δύο κατηγορίες, τις δυαδικές (διχοτομικές) και τις πολυωνυμικές μεταβλητές. Για παράδειγμα, στις περιπτώσεις των μεταβλητών φύλου (άνδρας ή γυναίκα) και ασθένειας (υγιής ή νοσούντας), η χρήση δυαδικών ονομαστικών μεταβλητών είναι η ιδανική. Οι πολυωνυμικές μεταβλητές, είναι ονομαστικές μεταβλητές των οποίων οι τιμές αντιστοιχούν σε περισσότερες από δύο κατηγορίες/κλάσεις και χρησιμοποιούνται για να περιγράψουν χαρακτηριστικά με τρία ή παραπάνω ενδεχόμενα. Παραδείγματα πολυωνυμικών μεταβλητών είναι η ομάδα αίματος (A, B, AB ή O), η μέθοδος διδασκαλίας (διάλεξη, χρήση διαφανειών, συζήτηση ή άλλο είδος), η αγαπημένη μουσική (κλασική, τζαζ, ροκ, κλπ), η οικογενειακή κατάσταση (ανύπαντρος/η, παντρεμένος/η, χήρος/α ή χωρισμένος/η), η προτίμηση του ροφήματος (καφές, τσάι, σοκολάτα ή άλλο) και οι κατηγορίες των πολιτικών κομμάτων.

Ένα ακόμη είδος μεταβλητών είναι οι διατακτικές κατηγορικές μεταβλητές, στις οποίες οι κατηγορίες είναι ταξινομημένες με συγκεκριμένη σειρά. Για παράδειγμα, στην περίπτωση των σταδίων ασθένειας (υγιής, ελαφριά και σοβαρά άρρωστος) και των τίτλων σπουδών (προπτυχιακό, μεταπτυχιακό ή διδακτορικό), η χρήση διατακτικών μεταβλητών μπορεί να περιγράψει αποτελεσματικά αυτή τη σχέση μεταξύ των κατηγοριών και των εξεταζόμενων χαρακτηριστικών. Από την άλλη πλευρά, οι διατακτικές μεταβλητές υποδηλώνουν συχνά ότι ορισμένες επιλογές είναι καλύτερες από άλλες, χωρίς όμως να μπορεί να προσδιοριστεί το πόσο καλύτερες είναι κάποιες επιλογές, καθώς τα διαστήματα μεταξύ των κατηγοριών δεν έχουν απαραίτητα την ίδια

απόκλιση. Στα πλαίσια της παρούσας εργασίας θα αναφερθούμε και στις τρείς παραπάνω κατηγορίες ονομαστικών μεταβλητών, δίνοντας μεγαλύτερη έμφαση στις ονομαστικές μεταβλητές που χαρακτηρίζονται από δύο ή περισσότερα ενδεχόμενα.

1.1 Πολυμεταβλητή ανάλυση

Τα δεδομένα που περιέχουν μεγάλο αριθμό μεταβλητών υποβάλλονται σε πολυπαραγοντικές τεχνικές ανάλυσης, όπου επιτρέπουν την ταυτόχρονη μελέτη πολλών μεταβλητών. Η ανάπτυξη εξειδικευμένων τεχνικών για την αξιολόγηση πολυμεταβλητών δεδομένων μπορεί να απαιτεί την παραβίαση ορισμένων θεωρητικών παραδοχών, αλλά μπορεί απλώς να περιλαμβάνει την κατηγοριοποίηση και την προσπάθεια εξεύρεσης τεράστιων όγκων δεδομένων, κάτι που δεν είναι ασυνήθιστο. Ως αποτέλεσμα, όταν υπάρχει μεγάλη ποσότητα πληροφοριών, υπάρχει συχνά πλεόνασμα από δεδομένα, πράγμα που σημαίνει ότι το ίδιο χαρακτηριστικό των παρατηρήσεων μετριέται και χαρακτηρίζεται από μεγάλο αριθμό μεταβλητών ταυτόχρονα.

Το αρχικό βήμα στις πολυμεταβλητές προσεγγίσεις είναι η κατηγοριοποίηση των δεδομένων προκειμένου να γίνει προσπάθεια ανακάλυψης μιας δομής μέσα σε αυτά. Εκτός από αυτό, οι πολυμεταβλητές προσεγγίσεις είναι αποτελεσματικές στη σύνοψη δεδομένων, ενώ αποκαλύπτουν επίσης βασικές διαστάσεις και σχέσεις στα δεδομένα. Στη βιβλιογραφία αναφέρονται μεταξύ άλλων ως στρατηγικές μείωσης και μέθοδοι κατηγοριοποίησης δεδομένων.

Με την εστίαση της στα φυτά και τα ζώα, η βιολογία είναι ένα εξαιρετικό παράδειγμα επιστήμης που βασίζεται στην ταξινόμηση. Με την χρήση της ταξινόμησης, γίνεται εύκολα κατανοητή η πολυπλοκότητα και η διαφοροποίηση των φαινομένων. Ως αποτέλεσμα, απαιτείται ταξινόμηση τόσο στην περιγραφή των φαινομένων όσο και στην ανάλυση τους. Πρέπει να σημειωθεί ότι η κατηγοριοποίηση δεν σημαίνει απλώς διαίρεση ενός ομοιογενούς συνόλου δεδομένων σε υποομάδες και καθορισμός των ομαδοποιήσεων που αντιστοιχούν στην πραγματικότητα.

Αντίθετα, η κατηγοριοποίηση συνεπάγεται στην ταξινόμηση ενός ετερογενούς συνόλου δεδομένων σε ομάδες που είναι αντιπροσωπευτικές της πραγματικότητας. Επομένως, όταν χρησιμοποιούνται πολλές μεταβλητές, είναι απαραίτητο να δημιουργηθεί μια υπόθεση για τη δομή των δεδομένων καθώς και μια κατάλληλη ερμηνεία για τα αποτελέσματα. Τα ψηφιακά δεδομένα γίνονται πιο προσιτά στην τρέχουσα εποχή και συχνά έχουν υψηλή γεωγραφική ανάλυση, δηλαδή αφορούν εξαιρετικά μικροσκοπικές γεωγραφικές ενότητες, όπως τα pixel, και επομένως είναι αρκετά χρήσιμα. Υπάρχει μια συνεχώς αυξανόμενη ανάγκη για προσεγγίσεις που μπορούν να χρησιμοποιηθούν για να περιγράψουν και να εξηγήσουν αυτά τα γεγονότα, με αποτέλεσμα να υπάρχει αυξανόμενη ζήτηση για αυτά.

Λόγω ότι οι υπολογισμοί είναι εξαιρετικά περίπλοκοι, η χρήση επαρκών προγραμμάτων υπολογιστή επιτρέπει την εμπειρική χρήση πολυμεταβλητών προσεγγίσεων στην εξέταση δεδομένων, κάτι που θα ήταν αδύνατο χωρίς τη χρήση κατάλληλων προγραμμάτων υπολογιστή στο παρελθόν. Παρά το γεγονός ότι οι μέθοδοι πολυμεταβλητής ταξινόμησης και ορισμένες από τις προσεγγίσεις αναπτύχθηκαν πριν από μερικές δεκαετίες, σήμερα θεωρούνται ως μέρος ενός

νέου πεδίου ανάλυσης δεδομένων που είναι γνωστό ως εξόρυξη δεδομένων, το οποίο βρίσκεται ακόμη στα πρώτα στάδια ανάπτυξης του.

Οι πολυμεταβλητές προσεγγίσεις περιέχουν μια ποικιλία στρατηγικών που ποικίλλουν ανάλογα με το ζήτημα που προσπαθούν να επιλύσουν, αλλά ο αριθμός των μεταβλητών που συνήθως χρησιμοποιούν δεν ξεπερνά τις 10 στις περισσότερες περιπτώσεις. Αν και οι πολυμεταβλητές προσεγγίσεις μπορεί να χρησιμοποιούν περισσότερες από 10 μεταβλητές, σε ορισμένες περιπτώσεις μπορούν να χρησιμοποιήσουν και περισσότερες από 100 μεταβλητές.

Ο (n x p) πίνακας δεδομένων $\boldsymbol{X}$ εμφανίζεται σαν ένα δείγμα από n παρατηρήσεις σε καθεμία από τις p τυχαίες μεταβλητές $X_1, X_2, \dots, X_p$. Επομένως, ο πίνακας $\boldsymbol{X}$ περιέχει τα p (n x 1) διανύσματα παρατήρησης $(\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_p})$ όπου

$$\boldsymbol{x_j} = \begin{pmatrix} x_{1,j} \\ x_{2,j} \\ \vdots \\ x_{n,j} \end{pmatrix}, \; j = 1,2, \dots, p \text{ και } \boldsymbol{X} = \left[\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_p}\right]$$

Έτσι, κάθε στήλη του $\boldsymbol{X}$ είναι ένα (n x 1) διάνυσμα παρατηρήσεων σε μία από τις μεταβλητές p. Κάθε γραμμή του $\boldsymbol{X}$ περιέχει παρατηρήσεις για τις p μεταβλητές $X_1, X_2, \dots, X_p$, που αντιστοιχεί σε ένα συγκεκριμένο άτομο ή αντικείμενο. Οι p τυχαίες μεταβλητές μαζί σχηματίζουν ένα (p x 1) διάνυσμα $\boldsymbol{x_i}$, όπου

$$\boldsymbol{x_i^T} = \begin{pmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,p-1} \\ x_{i,p} \end{pmatrix}^T$$

Επομένως, το διάνυσμα $\boldsymbol{x_k}$ θα χρησιμοποιηθεί για να υποδηλώσει ένα (n x 1) διάνυσμα παρατηρήσεων στη μεταβλητή X_k ή ένα (p x 1) διάνυσμα παρατηρήσεων στις μεταβλητές $X_1, X_2, \dots, X_p$, για κάθε k. Η επιλογή ανάμεσα σε αυτές τις δύο δυνατότητες συνήθως θα είναι ξεκάθαρη από το πλαίσιο αναφοράς.

1.2 Πολυμεταβλητές κατανομές και πολυμεταβλητές τυχαίες μεταβλητές

1.2.1 Από κοινού κατανομή

Η από κοινού συνάρτηση κατανομής για ένα τυχαίο διάνυσμα $\boldsymbol{X} = (X_1, X_2, \dots, X_p)$, στο σημείο $\boldsymbol{x} = (x_1, x_2, \dots, x_p)$, συμβολίζεται με $F_{X_1,X_2,\dots,X_p}(\boldsymbol{x})$, όπου

$$F_{X_1,X_2,\dots,X_p}(\boldsymbol{x}) = F_{X_1,X_2,\dots,X_p}(x_1, x_2, \dots, x_p) = P[X_1 \le x_1, X_2 \le x_2, \dots, X_p \le x_p\ ].$$

Η από κοινού πυκνότητα δηλώνεται από

$$f_{X_1,X_2,\dots,X_p}(\boldsymbol{x}) = f_{X_1,X_2,\dots,X_p}(x_1, x_2, \dots, x_p),$$

όπου $F_{X_1,X_2,\dots,X_p}(\boldsymbol{x})$=$\int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\dots\int_{-\infty}^{x_p} f_{X_1,X_2,\dots,X_p}(\boldsymbol{x})\, dx_1 dx_2 \dots dx_p$. Στο εφεξής με τον όρο πυκνότητα θα αναφερόμαστε είτε σε συναρτήσεις μάζας πιθανότητας ή πιθανότητες στην περίπτωση διακριτής τυχαίας μεταβλητής ή στην παράγωγο της συνάρτησης κατανομής στην περίπτωση συνεχών τυχαίων μεταβλητών.

1.2.2 Διαμερισμός της τυχαίας μεταβλητής

Μια διανυσματική τυχαία μεταβλητή $\boldsymbol{X}$ μπορεί να χωριστεί σε δύο αμοιβαία αποκλειόμενα υποσύνολα, όπου

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X_1} \\ \boldsymbol{X_2} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_q \\ x_{q+1} \\ x_{q+2} \\ \vdots \\ x_{q+s=p} \end{bmatrix},\ \boldsymbol{X_1} \text{ είναι } (q\ x\ 1),\ \boldsymbol{X_2} \text{ είναι } (s\ x\ 1) \text{ και } p = (q + s).$$

Έτσι, οι $\boldsymbol{X_1}$ και $\boldsymbol{X_2}$ είναι επίσης διανυσματικές τυχαίες μεταβλητές αλλά μικρότερης διάστασης από το $\boldsymbol{X}$. Η από κοινού συνάρτηση κατανομής $F_{\boldsymbol{X_1}}(\boldsymbol{x_1^*})$ για την μεταβλητή $\boldsymbol{X_1}$ μπορεί να υπολογιστεί από το $F_{\boldsymbol{X}}(\boldsymbol{x}^*)$ ενσωματώνοντας την πυκνότητα $f_{\boldsymbol{X}}(\cdot)$ σε ολόκληρο το εύρος των μεταβλητών στο $\boldsymbol{X_2}$. Δηλώνοντας την συνάρτηση της από κοινού κατανομής με $f_{\boldsymbol{X}}(\boldsymbol{x_1^*}, \boldsymbol{x_2^*})$, η $F_{\boldsymbol{X_1}}(\boldsymbol{x_1^*})$ δίνεται από τον τύπο:

$$F_{\boldsymbol{X_1}}(\boldsymbol{x_1^*}) = \int_{-\infty}^{x_1^*} \cdots \int_{-\infty}^{x_q^*} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(\boldsymbol{x_1}, \boldsymbol{x_2})\ dx_1 dx_2 \ldots dx_q dx_{q+1} \ldots dx_p$$
$$= \int_{-\infty}^{x_1^*} \cdots \int_{-\infty}^{x_q^*} f_{\boldsymbol{X_1}}(\boldsymbol{x_1})\, dx_1 dx_2 \ldots dx_q$$

όπου $f_{\boldsymbol{X_1}}(\boldsymbol{x_1})$ είναι η πυκνότητα του $\boldsymbol{X_1}$. Η περιθώρια πυκνότητα για το $\boldsymbol{X_1}$ λαμβάνεται από την από κοινού πυκνότητα για όλο το $\boldsymbol{X}$ ενσωματώνοντας το $f_{\boldsymbol{X}}(\cdot)$ στο εύρος των μεταβλητών στο $\boldsymbol{X_2}$.

$$f_{\boldsymbol{X_1}}(x_1, x_2, \ldots, x_q) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(\boldsymbol{x_1}, \boldsymbol{x_2})\, dx_{q+1} \ldots dx_p$$

Μια ειδική περίπτωση της κατανομής για το $\boldsymbol{X_1}$ εμφανίζεται όταν $q = 1$. Σε αυτή την περίπτωση το $\boldsymbol{X_1}$ είναι ισοδύναμο με τη τυχαία μεταβλητή X_1 και η κατανομή ονομάζεται περιθώρια κατανομή του X_1.

1.2.3 Δεσμευμένη κατανομή και ανεξαρτησία

Έστω X_1 και X_2 δυο μονοδιάστατες τυχαίες μεταβλητές. Η δεσμευμένη κατανομή για το X_2 δεδομένου X_1 ορίζεται ως:

$f_{X_2|X_1}(x_2|X_1 = x_1^*) = \frac{f_{X_1,X_2}(x_1^*,x_2)}{f_{X_1}(x_1^*)}$, όπου

$f_{X_1}(x_1^*)$ είναι η πυκνότητα για το X_1 που υπολογίζεται στο σημείο x_1^*. Οι δύο τυχαίες μεταβλητές X_1 και X_2 είναι ανεξάρτητες αν και μόνο αν

$$f_{X_2|X_1}(x_2|X_1 = x_1^*) = f_{X_2}(x_2) \text{ για όλα τα } x_1^* \text{ και } x_2$$

ή ισοδύναμα

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2) \text{ για κάθε } x_1, x_2.$$

1.2.4 Μέση τιμή και πίνακας συνδιακύμανσης

Το μέσο διάνυσμα $\boldsymbol{\mu}$ που αντιστοιχεί στην $(p\ x\ 1)$ τυχαία μεταβλητή $\boldsymbol{x}$ είναι το $(p\ x\ 1)$ διάνυσμα των στοιχείων $\mu_j = E[X_j]$, $\mathrm{j} = 1,2,\ldots,\mathrm{p}$ και γράφουμε $\boldsymbol{\mu} = E[\boldsymbol{x}]$. Ο πίνακας συνδιακύμανσης για το $\boldsymbol{x}$ είναι ο $(p\ x\ p)$ πίνακας $\boldsymbol{\Sigma}$ με διαγώνια στοιχεία $\sigma_j^2 = V[X_j]$, $\mathrm{j} = 1,2,\ldots,\mathrm{p}$ και εκτός των διαγώνιων στοιχείων τις συνδιακυμάνσεις $\sigma_{jk} = Cov(X_j, X_k),\ j \neq k,\ j,k = 1,2,\ldots,\mathrm{p}$. Το $\boldsymbol{\mu}$ διάνυσμα και ο πίνακας συνδιακυμάνσεων $\boldsymbol{\Sigma}$ δίνονται από

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix},\ \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \ldots & \sigma_p^2 \end{bmatrix}$$

Ο πίνακας συνδιακύμανσης $\boldsymbol{\Sigma}$ μπορεί επίσης να εκφραστεί ως

$$E[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})'] = \boldsymbol{\Sigma}$$

1.2.5. Πίνακας συσχέτισης

Ο πίνακας συσχέτισης $\boldsymbol{\rho}$ λαμβάνεται από τα στοιχεία του πίνακα συνδιακύμανσης $\boldsymbol{\Sigma}$ προσδιορίζοντας τα στοιχεία εκτός της διαγώνιου από

$$\rho_{jk} = \sigma_{jk} / \sqrt{\sigma_j^2 \sigma_k^2},\ j \neq k,\ j,k = 1,2,\ldots,p.$$

Ο πίνακας $\boldsymbol{\rho}$ δίνεται από

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & 1 & \vdots \\ \rho_{1p} & \rho_{2p} & \ldots & 1 \end{bmatrix}$$

Ο πίνακας συνδιακύμανσης $\boldsymbol{\Sigma}$ μπορεί επίσης να εκφραστεί ως

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1p}\sigma_1\sigma_p \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{2p}\sigma_2\sigma_p \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p}\sigma_1\sigma_p & \rho_{2p}\sigma_2\sigma_p & \ldots & \sigma_p^2 \end{bmatrix}$$

Κεφάλαιο 2^{ο} . Ανάλυση δυο κατηγορικών μεταβλητών

Η από κοινού κατανομή για ένα ζεύγος κατηγορικών τυχαίων μεταβλητών μπορεί να απεικονιστεί σε έναν δισδιάστατο πίνακα όπως φαίνεται στον Πίνακα 1. Η τυχαία μεταβλητή X θεωρείται ότι έχει ένα εύρος τιμών που αποτελείται από r κατηγορίες, ενώ η μεταβλητή Y θεωρείται ότι έχει c κατηγορίες. Η *πυκνότητα κελιών* για το κελί (i, j), δηλαδή η πιθανότητα μια παρατήρηση να αντιστοιχεί στο κελί (i, j), συμβολίζεται με f_{ij}, $i = 1,2, \dots, r$ $και$ $j = 1,2, \dots, c$, όπου όπως είναι γνωστό, ο πρώτος δείκτης αναφέρεται στη γραμμή και ο δεύτερος δείκτης στη στήλη. Οι οριακές πυκνότητες συμβολίζονται με $f_{i.}$ και $f_{.j}$ για τις μεταβλητές γραμμής και στήλης αντίστοιχα. Οι δεσμευμένες πυκνότητες για τις γραμμές που δίνονται στη στήλη j θα συμβολίζονται με $f_{i.}(i|j)$ και για τις στήλες που δίνονται στη γραμμή i με $f_{.j}(j|i)$.

Πίνακας 1. Αναπαράσταση πίνακα πυκνοτήτων δύο κατηγορικών μεταβλητών

	1	2	3	…	c	Σύνολο
1	f_{11}	f_{12}	f_{13}	…	f_{1c}	$f_{1}.$
2	f_{21}	f_{22}	f_{23}	…	f_{2c}	$f_{2}.$
3	f_{31}	f_{32}	f_{33}	…	f_{3c}	$f_{3}.$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
r	f_{r1}	f_{r2}	f_{r3}	…	f_{rc}	$f_{r}.$
Σύνολο	$f_{.1}$	$f_{.2}$	$f_{.3}$	…	$f_{.c}$	1

2.1 Ανεξαρτησία

Οι τυχαίες μεταβλητές X και Y είναι ανεξάρτητες εάν η πυκνότητα f_{ij} μπορεί να εκφραστεί ως το γινόμενο των αντίστοιχων οριακών πυκνοτήτων $f_{i.}$ και $f_{.j}$ για κάθε κελί (i, j). Η ανεξαρτησία μπορεί επίσης να οριστεί ως προς τις υπό συνθήκη πυκνότητες και τις περιθώριες πυκνότητες. Τα X και Y είναι ανεξάρτητα εάν η δεσμευμένη πυκνότητα για κάθε γραμμή είναι ίση με την περιθώρια πυκνότητα για το Y ή ισοδύναμα εάν η δεσμευμένη πυκνότητα για κάθε στήλη είναι ίση με την περιθώρια πυκνότητα για το X.

2.2 Παράδειγμα

Στην συνέχεια θα παρουσιάσουμε τα ευρήματα μιας μελέτης που εξέτασε την σχέση μεταξύ διατροφικών συνηθειών και διαφόρων παθήσεων (De Longerill et al., 1998). Στην μελέτη αυτή συγκρίθηκαν δυο διαφορετικές διατροφές, η Μεσογειακή διατροφή και μια διατροφή η οποία συστήνει χαμηλή κατανάλωση λιπαρών, και προτείνεται από την Αμερικανική Ένωση Καρδιολογίας (ΑΗΑ). Στην μελέτη αυτή, έλαβαν μέρος 605 ασθενείς με καρδιαγγειακές παθήσεις, οι οποίοι με τυχαίο τρόπο ακολούθησαν μια από τις δυο διατροφές. Τα αποτελέσματα της μελέτης παρουσιάζονται στον Πίνακα 2.

Μια σύγκριση των δεσμευμένων πυκνοτήτων $f_{.j}(j|i)$ για το αποτέλεσμα της υγείας δεδομένου του τρόπου διατροφής, αποκαλύπτει ότι το να είναι κάποιος υγιής $(j = 4)$ είναι το πιο συχνό αποτέλεσμα της Μεσογειακής διατροφής $(i = 2)$ με δεσμευμένη πυκνότητα ίση με 90.4%, η οποία είναι υψηλότερη από την αντίστοιχη για την διατροφή με χαμηλά λιπαρά $(i = 1)$ που αντιστοιχεί σε τιμή 78.8%. Επίσης, αναλύοντας τον Πίνακα 2 ανά στήλη με τις δεσμευμένες πυκνότητες, παρατηρούμε πως οι δεσμευμένες πυκνότητες των ασθενειών είναι μεγαλύτερες στην διατροφή χαμηλών λιπαρών συγκριτικά με την μεσογειακή διατροφή, ενώ το αποτέλεσμα «Υγιής» αντιστοιχεί σε μικρότερη πιθανότητα. Η μεταβλητότητα στις δεσμευμένες πυκνότητες υποδηλώνει μια αλληλεπίδραση μεταξύ των γραμμών και των στηλών. Οι υπό συνθήκη πυκνότητες $f_{.j}(j|i)$ αναφέρονται συχνά ως αναλογίες γραμμών και η περιθώρια πυκνότητα $f_{.j}$ ονομάζεται ολική αναλογία στήλης.

Πίνακας 2. Συχνότητες εμφάνισης και δεσμευμένες πυκνότητες μεταξύ διατροφών και αποτελέσματος υγείας

	Αποτέλεσμα				
	Καρκίνος	**Θάνατος**	**Μη θανατηφόρα ασθένεια**	**Υγιής**	Σύνολο
ΑΗΑ	15 (0.05/0.68)	24 (0.08/0.632)	25 (0.082/0.758)	239 (0.788/0.467)	303
Μεσογειακή	7 (0.023/0.32)	14 (0.046/0.368)	8 (0.027/0.242)	273 (0.904/0.533)	302
Σύνολο	22	38	33	512	605

[1] **Στην παρένθεση αναγράφονται οι δεσμευμένες πυκνότητες γραμμών / στηλών**

2.3 Λόγοι σχετικών πιθανοτήτων

Η από κοινού κατανομή μπορεί επίσης να μελετηθεί εξετάζοντας τους λόγους σχετικών πιθανοτήτων (odds ratio). Αρχικά θα ορίσουμε τους λόγους πιθανοτήτων. Ο λόγος f_{ij}/f_{is} αντιστοιχεί στην σχετική πιθανότητα να συμβεί το ενδεχόμενο που αντιστοιχεί στη στήλη j σε σχέση με τη στήλη s δεδομένου ότι συμβαίνει το ενδεχόμενο που αντιστοιχεί στην γραμμή i.

Επιπρόσθετα, ο λόγος f_{tj}/f_{ts} αντιστοιχεί στην σχετική πιθανότητα να συμβεί το ενδεχόμενο που αντιστοιχεί στη στήλη j έναντι της στήλης s δεδομένου ότι συμβαίνει το ενδεχόμενο που αντιστοιχεί στη γραμμή t. Ο λόγος σχετικών πιθανοτήτων (Odds Ratio - OR) είναι το πηλίκο των δύο σχετικών πιθανοτήτων και δίνεται από

$$\frac{f_{ij}}{f_{is}} / \frac{f_{tj}}{f_{ts}} = \frac{f_{ij} f_{ts}}{f_{is} f_{tj}}$$

Το OR είναι υποχρεωτικά 1 εάν ισχύει η υπόθεση της ανεξαρτησίας. Επίσης, υπό την ανεξαρτησία, οι πιθανότητες να συμβεί το ενδεχόμενο που αντιστοιχεί στη στήλη j σε σχέση με τη στήλη s δεν εξαρτώνται από την γραμμή.

2.4 Πίνακας συνάφειας δύο κατηγορικών μεταβλητών

Ένας πίνακας συνάφειας δύο κατηγορικών μεταβλητών παράγεται όταν ένα δείγμα από n παρατηρήσεις ταξινομείται ταυτόχρονα σε σχέση με δύο κατηγορικές τυχαίες μεταβλητές. Ο πίνακας συνάφειας (Πίνακας 3) είναι παρόμοιος με τον Πίνακα 1, με τη διαφορά ότι οι πυκνότητες των κελιών f_{ij} αντικαθίστανται από τις *παρατηρούμενες συχνότητες* ή *συχνότητες των κελιών* n_{ij} με $i = 1,2, \dots, r,\ j = 1,2 \dots, c$. Ένας πίνακας συνάφειας με r γραμμές και c στήλες ονομάζεται πίνακας με διάσταση $r \times c$ και παρέχει μια περίληψη της κατανομής συχνοτήτων του δείγματος. Διαιρώντας τις συχνότητες του δείγματος δια n προκύπτει ο πίνακας πυκνοτήτων. Τα σύνολα γραμμών και στηλών για τον πίνακα αντιπροσωπεύουν τις περιθώριες κατανομές του δείγματος για τις δύο τυχαίες μεταβλητές.

Πίνακας 3. Αναπαράσταση πίνακα συχνοτήτων δύο κατηγορικών μεταβλητών

	1	2	3	…	c	Σύνολο
1	n_{11}	n_{12}	n_{13}	…	n_{1c}	$n_{1.}$
2	n_{21}	n_{22}	n_{23}	…	n_{2c}	$n_{2.}$
3	n_{31}	n_{32}	n_{33}	…	n_{3c}	$n_{3.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
r	n_{r1}	n_{r2}	n_{r3}	…	n_{rc}	$n_{r.}$
Σύνολο	$n_{.1}$	$n_{.2}$	$n_{.3}$	…	$n_{.c}$	n

2.5 Δειγματοληψία μοντέλων για πίνακες συνάφειας

Υπάρχει μια ποικιλία μοντέλων δειγματοληψίας που μπορούν να χρησιμοποιηθούν για να περιγράψουν τη διαδικασία που συνδέεται με τον πίνακα συνάφειας $(r \times c)$ των n παρατηρήσεων. Οι πιο κοινές κατανομές είναι η Πολυωνυμική, η Υπεργεωμετρική, η Poisson και η από κοινού Πολυωνυμική. Η πιο προφανής επέκταση της απλής τυχαίας δειγματοληψίας για την διμεταβλητή ανάλυση είναι η Πολυωνυμική κατανομή.

2.5.1 Κατανομή Bernoulli

Έστω ένα πείραμα (δοκιμή) το οποίο έχει δυο δυνατά αποτελέσματα, την επιτυχία και την αποτυχία. Θεωρούμε ως Y την τυχαία μεταβλητή που σχετίζεται με την απόκριση του πειράματος, και έτσι η Y παίρνει την τιμή 1 αν το αποτέλεσμα είναι επιτυχία και την τιμή 0 αν το αποτέλεσμα είναι αποτυχία. Ορίζουμε την πιθανότητα επιτυχίας ως $P[Y = 1] = \pi$ και την πιθανότητα επιτυχίας ως $P[Y = 0] = 1 - \pi = q$. Η τυχαία μεταβλητή Y ορίζεται πως ακολουθεί κατανομή Bernoulli με παράμετρο π, και η κατανομή αυτή είναι από τις πιο απλές και βασικές κατανομές που εφαρμόζονται σε κατηγορικές τυχαίες μεταβλητές.

2.5.2 Διωνυμική κατανομή

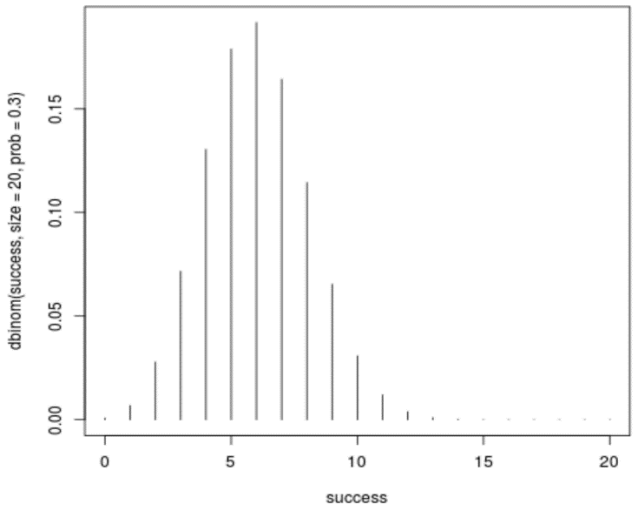
Πολλές εφαρμογές αναφέρονται σε έναν γνωστό αριθμό n δίτιμων παρατηρήσεων που ακολουθούν την κατανομή Bernoulli. Έστω $y_1, \dots y_n$, οι αποκρίσεις για n ανεξάρτητες και πανομοιότυπες δοκιμές, έτσι ώστε $P[Y_i = 1] = \pi$ και $P[Y_i = 0] = 1 - \pi = q$. Σημειώνουμε ότι με τον όρο «πανομοιότυπες» δοκιμές εννοούμε ότι η πιθανότητα της επιτυχίας π, είναι ίδια για κάθε δοκιμή και με τον όρο «ανεξάρτητες» δοκιμές εννοούμε ότι οι αποκρίσεις Y_i, είναι ανεξάρτητες, τυχαίες μεταβλητές. Ο συνολικός αριθμός των επιτυχιών, έστω $Y = \sum_{i=1}^{n} Y_i$ ακολουθεί την διωνυμική κατανομή με παραμέτρους n και π. Η συνάρτηση μάζας πιθανότητας για την πιθανή τιμή y της μεταβλητής Y , είναι

$$P[Y = y] = \binom{n}{y} \pi^y q^{n-y},$$

για $y = 0,1,2,\dots,n,$ όπου $\binom{n}{y} = \frac{n!}{y!(n-y)!}$.

Δεν υπάρχει εγγύηση ότι οι διαδοχικές διωνυμικές παρατηρήσεις είναι πάντα ανεξάρτητες και πανομοιότυπες. Έτσι κάποιες φορές χρησιμοποιούνται άλλες κατανομές. Μια τέτοια περίπτωση είναι όταν παίρνουμε διωνυμικό δείγμα από έναν γνωστό πληθυσμό χωρίς επανάθεση, για παράδειγμα, όταν παρατηρούμε την κατανομή του φύλου σε μια τάξη μαθητών, λαμβάνοντας

δείγμα 10 μαθητών από μια τάξη με σύνολο 20 μαθητών. Στην περίπτωση αυτή η υπεργεωμετρική κατανομή είναι καταλληλότερη, καθώς όταν συλλέγουμε το δείγμα αυτό επηρεάζει τις επόμενες επιλογές που έχουμε από μαθητές. Στο παρακάτω γράφημα παρουσιάζεται η Διωνυμική κατανομή για 20 δοκιμές, δηλαδή για παραμέτρους $n = 20$, και πιθανότητα επιτυχίας $\pi = 0.3$



Γράφημα 1-Αναπαράσταση Διωνυμικής Κατανομής (n = 20, π = 0.3).

2.5.3 Πολυωνυμική κατανομή

Για την πολυωνυμική κατανομή, θεωρούμε ότι επιλέγεται ένα τυχαίο δείγμα n παρατηρήσεων από έναν άπειρο πληθυσμό. Στη συνέχεια, οι παρατηρήσεις ταξινομούνται σε ένα από τα $r \times c$ κελιά του πίνακα και έστω οι τυχαίες μεταβλητές $n_{11}, n_{12}, \dots, n_{rc}, i = 1, \dots, r, j = 1, \dots, c$, που δηλώνουν την συχνότητα των κελιών. Η πιθανότητα για τις συχνότητες των κελιών δίνεται από

$$f(n_{11}, n_{12}, \dots, n_{rc}) = \frac{n!}{\prod_{i=1}^{r}\prod_{i=1}^{c} n_{ij}!}\prod_{i=1}^{r}\prod_{i=1}^{c} f_{ij}^{nij}$$

Όπου

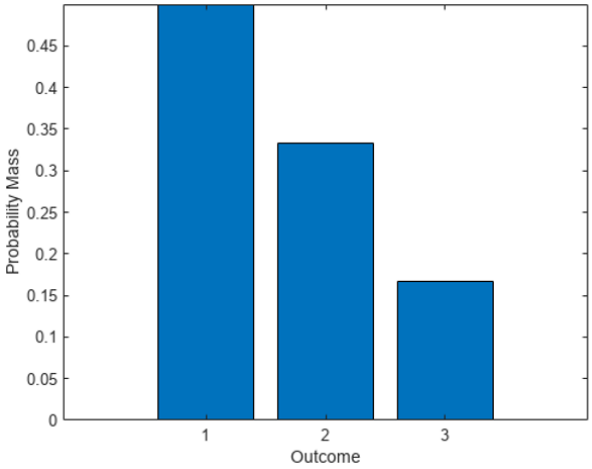
$$\sum_{i=1}^{r}\sum_{i=1}^{c} n_{ij} = n.$$

Οι μέσοι όροι, οι διακυμάνσεις και οι συνδιακυμάνσεις για το n_{ij} δίνονται από

$$E[n_{ij}] = nf_{ij}, \quad V[n_{ij}] = nf_{ij}(1- nf_{ij}) \text{ όπου } i = 1,2,\dots,r \text{ και όπου } j = 1,2,\dots,c$$

$$Cov[n_{ij}, n_{kl}] = -nf_{ij}f_{kl} \text{ με } i \neq k, j \neq l, i,k = 1,2,\dots,r, j,l = 1,2,\dots,c$$

Οι εκτιμητές μέγιστης πιθανοφάνειας των κελιών f_{ij} είναι οι αντίστοιχες αναλογίες του δείγματος n_{ij}/n. Μια χρήσιμη ιδιότητα της πολυωνυμικής κατανομής είναι ότι τα αθροίσματα πολυωνυμικών τυχαίων μεταβλητών κατανέμονται επίσης πολυωνυμικά. Οι παράμετροι αθροίζονται επίσης για να ληφθούν οι αντίστοιχες παράμετροι για την κατανομή των αθροισμάτων. Μια ειδική περίπτωση της πολυωνυμικής κατανομής είναι η διωνυμική όπου $c = 2$ και $r = 1$. Στην περίπτωση αυτή υπάρχουν μόνο δύο πιθανά κελιά. Στο Γράφημα 2 παρουσιάζεται ένα παράδειγμα πολυωνυμικής κατανομής για 3 πιθανά αποτελέσματα.



Γράφημα 2 - Πολυωνυμική κατανομή για τρία ενδεχόμενα

2.5.4 Υπεργεωμετρική κατανομή

Υποθέτουμε ότι ο πληθυσμός είναι πεπερασμένος και είναι γνωστές οι πληθυσμιακές συχνότητες των κελιών N_{ij}, οι οποίες κατανέμονται σε ένα πίνακα συχνοτήτων, όπου $i = 1,2,\dots,r$, $j = 1,2 \dots, c$. Τότε αν λάβουμε ένα τυχαίο δείγμα n παρατηρήσεων από τον πληθυσμό, οι πιθανότητες να παρατηρήσουμε στις κατηγορίες $n_{11}, n_{12}, \dots, n_{rc}$ δείγματα δίνονται σύμφωνα με την υπεργεωμετρική κατανομή από:

$$f(n_{11}, n_{12}, \dots, n_{rc}) = \prod_{i=1}^{r}\prod_{i=1}^{c} \frac{N_{ij}!}{n_{ij}!\,(N_{ij}! - n_{ij})!} \Big/ \frac{N!}{n!\,(N-n)!}$$

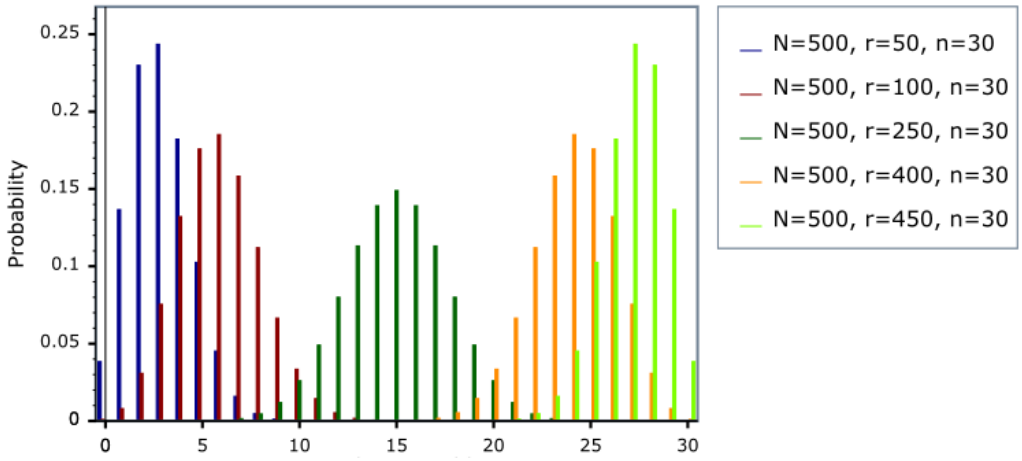
Οι μέσοι όροι, οι διακυμάνσεις και οι συνδιακυμάνσεις για το n_{ij} δίνονται από

$$E[n_{ij}] = nf_{ij},$$

$$V[n_{ij}] = (\frac{N-n}{n-1})nf_{ij}(1- f_{ij}) \text{ όπου } i = 1,2,\dots,r \text{ και όπου } j = 1,2,\dots,c$$

$$Cov[n_{ij}, n_{kl}] = -(\frac{N-n}{n-1})nf_{ij}f_{kl}\,,\ f_{ij} \neq \frac{N_{ij}}{N},\quad i \neq k, j \neq l,\ i,k = 1,2,\dots,r,\ j,l = 1,2,\dots,c$$

Στην περίπτωση μεγάλων πεπερασμένων πληθυσμών, η υπεργεωμετρική κατανομή μπορεί να προσεγγιστεί με την πολυωνυμική, με την προϋπόθεση ότι κάθε N_{ij} είναι μεγάλο. Στο Γράφημα 3 παρουσιάζεται η συνάρτηση μάζας πιθανότητας της υπεργεωμετρικής κατανομής για διάφορες παραμέτρους.



Γράφημα 3 - Αναπαράσταση της υπεργεωμετρικής κατανομής

.

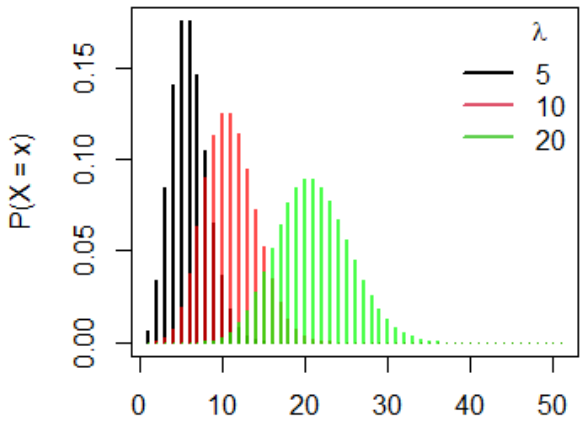
2.5.5 Κατανομή Poisson

Στις δειγματοληψίες με πολυωνυμικές και υπεργεωμετρικές κατανομές θεωρούμε ότι το συνολικό μέγεθος του δείγματος n είναι σταθερό. Μια εναλλακτική υπόθεση είναι να επιτρέψουμε στο n να είναι επίσης μια τυχαία μεταβλητή. Μια χρήσιμη κατανομή σε αυτή την περίπτωση είναι η κατανομή Poisson (Γράφημα 4). Οι κατανομές των συχνοτήτων των κελιών n_{ij} υποτίθεται ότι κατανέμονται αμοιβαία ανεξάρτητα σύμφωνα με την κατανομή Poisson με παραμέτρους $F_{ij} = E[n_{ij}]$. Σε αυτήν την περίπτωση, το συνολικό μέγεθος δείγματος n ακολουθεί επίσης την κατανομή Poisson με παράμετρο ίση με το άθροισμα των επιμέρους παραμέτρων των κελιών, $F_{..} =$

$E[n] = \sum_{i=1}^{r} \sum_{i=1}^{c} F_{ij}$. Η διακύμανση $V[n_{ij}]$. δίνεται επίσης από τον F_{ij} και η πιθανότητα για τις συχνότητες των κελιών $n_{11}, n_{12}, \dots, n_{rc}, i = 1, \dots, r, j = 1, \dots, c$, δίνεται από

$$f(n_{11}, n_{12}, \dots, n_{rc}) = \prod_{i=1}^{r} \prod_{i=1}^{c} F_{ij}^{n_{ij}} e^{-F_{ij}} / n_{ij}!$$

Δεδομένου ότι οι συχνότητες των κελιών θεωρούνται ότι είναι αμοιβαία ανεξάρτητες, η $Cov[n_{ij}, n_{kl}] = 0, \quad i \neq k, j \neq l, \quad i, k = 1,2, \dots, r, \quad j, l = 1,2, \dots, c.$ Οι μέγιστοι εκτιμητές πιθανοτήτων των παραμέτρων F_{ij} είναι οι συχνότητες δείγματος n_{ij}.



Γράφημα 4 - Κατανομή Poisson για διαφορετικές παραμέτρους.

2.5.6 Από κοινού πολυωνυμική κατανομή

Η από κοινού πολυωνυμική κατανομή προκύπτει από την κοινή κατανομή δύο ή περισσότερων ανεξάρτητων πολυωνυμικών κατανομών. Στον δισδιάστατο πίνακα συνάφειας, το μέγεθος δείγματος της γραμμής $n_{i.}$, $i = 1,2, \dots, r$, μπορεί να υπολογιστεί. Σε αυτή την περίπτωση, η πυκνότητα για τις συχνότητες κελιών σε κάθε γραμμή δίνεται από την πολυωνυμική κατανομή. Κάθε γραμμή του πίνακα αναφέρεται ως υποπληθυσμός. Η πυκνότητα για όλες τις r γραμμές δίνεται από το γινόμενο των μεμονωμένων πυκνοτήτων των γραμμών και ως εκ τούτου προκύπτει το γινόμενο των πολυωνυμικών κατανομών. Η από κοινού πολυωνυμική πυκνότητα γινομένου για έναν πίνακα συνάφειας $r\ x\ c$ δίνεται από το γινόμενο των πολυωνυμικών πυκνοτήτων r που αντιστοιχούν στις γραμμές και ως εκ τούτου

$$f(n_{11}, n_{12}, \dots, n_{rc)} = \prod_{i=1}^{r}\left[\frac{n_{i.}!}{\prod_{j=1}^{c} n_{i.}!}\prod_{j=1}^{c}\left[\frac{f_{ij}}{f_{i.}}^{n_{ij}}\right]\right]$$

Η από κοινού πολυωνυμική πυκνότητα μπορεί επομένως να προκύψει ελέγχοντας τα μεγέθη των γραμμών του δείγματος $n_{i.}$, όπως επίσης και με τον καθορισμό των μεγεθών των στηλών του δείγματος $n_{.j}$.

2.6 Έλεγχος Ανεξαρτησίας

Εφόσον οι μεταβλητές δεν εξαρτώνται η μία από την άλλη, οι αναμενόμενες συχνότητες για τα κελιά στον πίνακα συσχέτισης τους υπολογίζονται ως γινόμενο των αντίστοιχων περιθωρίων συχνοτήτων κάθε κελιού διαιρεμένο με τον συνολικό αριθμό των παρατηρήσεων του πίνακα. Στην περίπτωση που οι εξεταζόμενες μεταβλητές είναι ανεξάρτητες, οι παρατηρούμενες συχνότητες μπορεί να διαφέρουν από τις αναμενόμενες, είτε λόγω σε τυχαίων αιτίων ή λόγω της ύπαρξης κάποιας φύσεως συστηματικότητας, δηλαδή θα περιμέναμε ότι οι πυκνότητες του δείγματος $n_{ij} \,/\, n$ θα πρέπει να είναι παρόμοιες με το γινόμενο των περιθωρίων πυκνοτήτων του δείγματος $(n_{i.} / \, n) \times (n_{.j} \,/\, n)$ και επομένως οι εκτιμώμενες αναμενόμενες συχνότητες υπό την ανεξαρτησία είναι $(n_{i.} n_{.j} \,/\, n)$. Εάν οι μεταβλητές είναι εξαρτημένες, οι παρατηρούμενες συχνότητες μπορεί να διαφέρουν από τις αναμενόμενες κυρίως λόγω της ύπαρξης τυχαίων επιδράσεων. Έτσι, η ανεξαρτησία μεταξύ δύο κατηγοριών μεταβλητών ελέγχεται από μια συνάρτηση που βασίζεται στις διαφορές μεταξύ των παρατηρούμενων και αναμενόμενων συχνοτήτων των υπό εξέταση μεταβλητών. Στην περίπτωση ενός πίνακα με r γραμμές και c στήλες (συνολικές διάστασης $r \times c$), η ποσότητα που αντικατοπτρίζει τη διαφορά μεταξύ των προαναφερθέντων συχνοτήτων συμβολίζεται με X^2 ορίζεται ως

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij} - \frac{n_{i.}\, n_{.j}}{n}\right)^2}{\frac{n_{i.}\, n_{.j}}{n}}.$$

Οι προς διερεύνηση υποθέσεις είναι οι ακόλουθες

- H_0 = Οι μεταβλητές X και Y είναι ανεξάρτητες
- H_1 = Οι μεταβλητές X και Y δεν είναι ανεξάρτητες

Στη συνέχεια για να διαπιστώσουμε ποια από τις δύο παραπάνω υποθέσεις (H_0 ή H_1) ισχύει, εφαρμόζεται ο έλεγχος του X^2 με επίπεδο σημαντικότητας (ή αλλιώς στάθμη σημαντικότητας) της τάξεως του $\alpha = 0.05$ ή $\alpha = 0.01$. Η επιλογή του επιπέδου σημαντικότητας α αποτελεί αποκλειστικά επιλογή του ερευνητή και συχνά σχετίζεται με τη φύση του ερευνητικού πεδίου στο οποίο λαμβάνει χώρα η έρευνα. Παραδείγματος χάριν, σε έρευνες ψυχολογίας, η χρήση της τιμής $\alpha = 0.05$ ή $\alpha = 5\%$ είναι σχεδόν καθολική. Έτσι λοιπόν, με βάση την κατανομή X^2 και το επιλεγμένο επίπεδο σημαντικότητας α, η επιλογή της κατάλληλης υπόθεσης γίνεται ως:

— Αν το p-value για τον έλεγχο X^2 είναι μικρότερο ή ίσο από τη στάθμη σημαντικότητας $\alpha = 0.05$ (ή 0,01 αν το επίπεδο σημαντικότητας είναι αυτό) τότε απορρίπτεται η H_0 (αποδοχή της H_1) και συνεπώς οι μεταβλητές δεν είναι ανεξάρτητες.
— Αν το p-value για τον έλεγχο X^2 είναι μεγαλύτερο από τη στάθμη σημαντικότητας $\alpha = 0.05$ (ή 0,01 αν το επίπεδο σημαντικότητας είναι αυτό) τότε δεχόμαστε την H_0 και συνεπώς οι μεταβλητές μπορούν να θεωρηθούν ανεξάρτητες.

Οι προϋποθέσεις για την αξιοπιστία του ελέγχου X^2 είναι οι εξής :

- Καμία αναμενόμενη συχνότητα δε θα πρέπει να είναι μικρότερη του 1.
- Δεν πρέπει να υπάρχουν επικαλύψεις μεταξύ των συχνοτήτων μεταξύ των κελιών του πίνακα συνάφειας – ανεξαρτησία παρατηρήσεων.
- Το ποσοστό των αναμενόμενων συχνοτήτων που είναι μικρότερες από το 5, δε θα πρέπει να υπερβαίνουν το 20% των κελιών του πίνακα.

Έστω ότι πέντε κελιά σε έναν πίνακα διάστασης 3×4 έχουν αναμενόμενη συχνότητα μικρότερη του 5 και επομένως παραβιάζει η τρίτη προϋπόθεση για τη διενέργεια του ελέγχου X^2. Σε αυτή την περίπτωση, ο εκάστοτε χρήστης μπορεί να χρησιμοποιήσει τον έλεγχο του Fisher αντί για τον έλεγχο X^2 ώστε να διαπιστώσει το αν αυτή η παραβίαση της τρίτης προϋπόθεσης, επηρεάζει τα αποτελέσματα του ελέγχου. Θα διατυπώσουμε τον έλεγχο Fisher για την περίπτωση ενός πίνακα συνάφειας διάστασης 2x2 καθώς η γενίκευση του, αν και αυξάνει ιδιαίτερα την πολυπλοκότητα, μπορεί να εφαρμοσθεί απευθείας (Agresti, 2012). Ο έλεγχος αυτός βασίζεται στην υπεργεωμετρική κατανομή και υποθέτουμε σε ένα δείγμα μεγέθους n ότι οι παρατηρούμενες συχνότητες των κελιών είναι n_{11}, n_{12}, n_{21} και n_{22}, με αθροίσματα γραμμών $n_{1.}, n_{2.}$ και στηλών $n_{.1}, n_{.2}$. Σύμφωνα με την υπεργεωμετρική κατανομή, η πιθανότητα το κελί $(1,1)$ να έχει συχνότητα ίση με t δίνεται από

$$P(n_{11} = t) = \frac{\binom{n_{1.}}{t}\binom{n_{2.}}{n_{.1} - t}}{\binom{n}{n_{.1}}}.$$

Δεδομένων των γνωστών περιθωρίων συχνοτήτων, η ποσότητα n_{11} καθορίζει τις συχνότητες των υπόλοιπων τριών κελιών μονοσήμαντα. Υπό την εναλλακτική υπόθεση, δηλαδή ότι η μεταβλητή των γραμμών σχετίζεται με αυτή των στηλών, και γνωρίζοντας τις περιθώριες συχνότητες, θα αναμέναμε ότι αν το n_{11} παίρνει μεγαλύτερες τιμές, τότε αυτό θα είναι ένδειξη υπέρ της εναλλακτικής υπόθεσης. Πιο συγκεκριμένα, αν υποθέσουμε ότι η παρατηρούμενη συχνότητα στο κελί $(1,1)$ είναι ίση με t_0, αν $P(n_{11} \geq t_0) \leq 0.05$, τότε απορρίπτουμε τη μηδενική υπόθεση και δεχόμαστε ότι οι μεταβλητές είναι εξαρτημένες (Agresti, 2012). Παραδειγματικά, κατασκευάζουμε όλους τους 2x2 πίνακες με την προϋπόθεση ότι διατηρούνται οι περιθώριες συχνότητες γραμμών και στηλών και σύμφωνα με την παραπάνω σχέση αθροίζουμε όλες τις πιθανότητες, όπου $P(n_{11} \geq t_0)$. Αν αυτή η πιθανότητα είναι μικρότερη ή ίση του επιπέδου σημαντικότητας, τότε μπορούμε να απορρίψουμε τη μηδενική υπόθεση.

Το στατιστικό Pearson X^2 βασίζεται στην υπόθεση ενός πολυωνυμικού πληθυσμού με $r \times c$ κελιά. Σε μεγάλα δείγματα, οι αναλογίες δειγμάτων $(n_{i.}/\, n)$ και $(n_{.j}\, /\, n)$ θεωρούνται ότι είναι κανονικά κατανεμημένες. Το Pearson X^2 προκύπτει από την κατανομή των αθροισμάτων των τετραγώνων των τυποποιημένων κανονικών τυχαίων μεταβλητών. Μία εναλλακτική του στατιστικού X^2 μπορεί να ληφθεί χρησιμοποιώντας το λόγο πιθανοφανειών (likelihood ratio). Και πάλι υποθέτοντας πολυωνυμικό πληθυσμό, για μεγάλα δείγματα, το στατιστικό

$$H^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \ln(\frac{n_{ij}\, n}{n_{i.} n_{.j}})$$

ακολουθεί την κατανομή X^2 με $(r-1) \times (c-1)$ βαθμούς ελευθερίας. Για μεγάλα δείγματα τα δύο στατιστικά είναι συνήθως αρκετά συγκρίσιμα.

2.6.1 Τυποποιημένα κατάλοιπα

Ένας χρήσιμος τρόπος σύγκρισης των παρατηρούμενων και των αναμενόμενων συχνοτήτων είναι ο προσδιορισμός των τυποποιημένων καταλοίπων (standardized residuals) για κάθε κελί. Το στατιστικό Pearson X^2 παρέχει πληροφορίες σχετικά με τα κελιά που έχουν τη μεγαλύτερη συνεισφορά στο X^2. Οι τετραγωνικές ρίζες καθενός από τους όρους στο στατιστικό Pearson X^2 ονομάζονται συνήθως τυποποιημένα κατάλοιπα. Για το κελί (i, j) το τυποποιημένο υπόλοιπο δίνεται από

$$r_{ij}^{(1)} = (n_{ij} - n_{i.}n_{.j} / n)/\sqrt{n_{i.}n_{.j} / n}.$$

Μια εναλλακτική μέθοδος τυποποίησης των καταλοίπων είναι μέσω της χρήσης των αποκλίσεων Freeman-Tukey, $r_{ij}^{(2)}$, που δίνονται από τη σχέση (Jobson 1992)

$$r_{ij}^{(2)} = \sqrt{n_{ij}} + \sqrt{n_{ij} + 1} - \sqrt{4n_{i.}n_{.j} / n + 1}.$$

2.7 Υποθέσεις δειγματοληψίας μοντέλου

Η δοκιμή μας για ανεξαρτησία στους πίνακες συνάφειας που περιγράφονται παραπάνω υποθέτει ότι τα δεδομένα ελήφθησαν ως τυχαίο δείγμα μεγέθους n από έναν πολυωνυμικό πληθυσμό. Οι μονάδες πληθυσμού διαιρέθηκαν μεταξύ των rc κελιών με την πιθανότητα μια μονάδα να εμφανίζεται στο κελί (i, j) να συμβολίζεται με f_{ij} όπου $\sum_{i=1}^{r}\sum_{j=1}^{c} f_{ij} = 1$. Οι οριακές πυκνότητες για τις γραμμές δίνονται από το $f_{i.}$, όπου $f_{i.} = \sum_{j=1}^{c} f_{ij}$, $i = 1,2, \dots, r$. Ομοίως οι περιθώριες πυκνότητες για τις στήλες δίνονται από $f_{.j}$, όπου $f_{.j} = \sum_{j=1}^{c} f_{ij}$, $j = 1,2, \dots, c$. Οι εκτιμήσεις των δειγμάτων f_{ij}, $f_{i.}$ και $f_{.j}$ δίνονται από τα n_{ij} / n, $(n_{i.}/ n)$ και $(n_{.j} / n)$ αντίστοιχα. Αυτοί οι εκτιμητές είναι οι εκτιμητές μέγιστης πιθανοφάνειας κάτω από την υπόθεση της πολυωνυμικής δειγματοληψίας.

2.7.1 Κατανομή Poisson

Όπως έχει ήδη προαναφερθεί, δύο κοινά χρησιμοποιούμενα εναλλακτικά μοντέλα δειγματοληψίας οδηγούν στους ίδιους εκτιμητές μέγιστης πιθανότητας με το πολυωνυμικό μοντέλο. Εάν δεν υπάρχει περιορισμός στο συνολικό μέγεθος του δείγματος, οι συχνότητες των κελιών n_{ij} μπορούν να θεωρηθούν ως τυχαίες μεταβλητές με αναμενόμενη μέση τιμή

$$E[n_{ij}] = F_{ij} \ , \text{ όπου } F_{ij} \text{ η πυκνότητα δύο κατηγορικών μεταβλητών.}$$

Αν καθένα από τα n_{ij} θεωρείται ότι έχει μια ανεξάρτητη κατανομή Poisson, οι εκτιμητές μέγιστης πιθανοφάνειας για την αναμενόμενη τιμή $E[n_{ij}]$ δίνονται από το $(n_{i.}n_{.j} / n)$. Για την υπόθεση της κατανομής Poisson το συνολικό μέγεθος του δείγματος n δεν είναι σταθερό, αλλά είναι τυχαία μεταβλητή, δηλαδή εάν η δειγματοληψία πραγματοποιηθεί για μια καθορισμένη χρονική περίοδο και στη συνέχεια διακοπεί, το συνολικό μέγεθος δείγματος n που αποκτήθηκε μέχρι εκείνο το σημείο είναι μια τυχαία μεταβλητή. Η υπό συνθήκη κατανομή του n_{ij} δεδομένου

σταθερού n σε αυτήν την περίπτωση, είναι μια πολυωνυμική κατανομή. Ως εκ τούτου, οι παραπάνω διαδικασίες μπορούν να εφαρμοστούν στη δειγματοληψία Poisson. Η υπόθεση της ανεξαρτησίας στην περίπτωση Poisson υποδηλώνει ότι η πραγματική μέση τιμή του κελιού (i, j), $E[n_{ij}]$ ικανοποιεί την υπόθεση ανεξαρτησίας που δίνεται ως

$$F_{ij} = E[n_{ij}] = \frac{E[n_{i.}]E[n_{.j}]}{E[n]\,E[n]}E[n] = \frac{E[n_{i.}]E[n_{.j}]}{E[n]}$$

2.7.2 Από κοινού Πολυωνυμική Κατανομή

Μια δεύτερη εναλλακτική στον πολυωνυμικό πληθυσμό ονομάζεται από κοινού πολυωνυμική κατανομή. Στην περίπτωση του από κοινού πολυωνύμου, προστίθενται επιπλέον περιορισμοί στο δείγμα. Τα σύνολα των γραμμών $n_{i.}$ ή τα σύνολα των στηλών $n_{.j}$ έχουν καθοριστεί εκ των προτέρων. Έτσι λοιπόν, το δείγμα περιορίζεται στο να περιέχει συγκεκριμένο αριθμό παρατηρήσεων από κάθε κατηγορία μιας εκ των μεταβλητών. Οι εκτιμητές μέγιστης πιθανοφάνειας των μη περιορισμένων περιθωρίων κατανομών, $f_{.j}$ ή $f_{i.}$, δίνονται ως $n_{.j}/n$ ή $n_{i.}/n$ αντίστοιχα. Οι αναμενόμενες συχνότητες των κελιών υπό την υπόθεση της ανεξαρτησίας εκτιμώνται ως $n_{i.}n_{.j}/n$ όπως και στις δύο προηγούμενες περιπτώσεις.

Σε αυτή την περίπτωση ο έλεγχος αναφέρεται συχνά ως έλεχγος ομοιογένειας των αναλογιών των γραμμών ή στηλών. Αν οι περιθώριες τιμές $n_{i.}$ είναι σταθερές, τότε κάνουμε δειγματοληψία ανεξάρτητα από τους υποπληθυσμούς της γραμμής r. Σε αυτή την περίπτωση, η υπόθεση ανεξαρτησίας $f_{ij} = f_{i.}f_{.j}$ μπορεί να γραφεί ως $f_{ij}/f_{i.} = f_{.j}$, η οποία δηλώνει ότι οι υπό συνθήκη πυκνότητες για κάθε επίπεδο j, κάθε γραμμής i είναι ισοδύναμες με τις περιθώριες πυκνότητες σε κάθε επίπεδο του j.

Οι εκτιμώμενες αναμενόμενες αναλογίες των κελιών σε αυτό το μοντέλο λαμβάνονται γράφοντας την αναλογία $n_{ij} = n_{i.}n_{.j}/n$ υπό τη μορφή $(n_{ij}/n = n_{.j}/n)$. Οι εκτιμώμενες αναμενόμενες αναλογίες γραμμών $(n_{ij}/n_{i.})$ για κάθε επίπεδο j σε κάθε γραμμή, αναμένεται να είναι ομοιογενείς σε σχέση με τις r γραμμές. Παρόμοια είναι και η λογική, στην περίπτωση σταθερών περιθωρίων στηλών. Οι εκτιμώμενες αναμενόμενες αναλογίες στηλών θα πρέπει να είναι ομοιογενείς ως προς τις στήλες. Και τα τρία μοντέλα δειγματοληψίας για τον πίνακα συνάφειας $r \times c$ αποδίδουν ίδιες εκτιμήσεις, για τις αναμενόμενες συχνότητες υπό ανεξαρτησία. Ο λόγος πιθανοφανειών του στατιστικού H^2 που δίνεται παραπάνω, είναι πανομοιότυπος και για τα τρία δειγματοληπτικά μοντέλα.

2.8 Μοντέλα για πίνακες συνάφειας δυο διαστάσεων

Στη μελέτη μας της κοινής κατανομής για ποιοτικές διμεταβλητές τυχαίες μεταβλητές, η απόκλιση από την ανεξαρτησία χαρακτηριζόταν από την εξέταση της συμπεριφοράς των

αναλογιών γραμμών και στηλών. Εναλλακτικά βασιζόμασταν στη διερεύνηση της συσχέτισης μεταξύ γραμμών και στηλών. Αν και τέτοιες τεχνικές είναι συνήθως επαρκείς για τον χαρακτηρισμό της συμπεριφοράς σε πίνακες δύο διαστάσεων, οι πίνακες υψηλότερων διαστάσεων μελετώνται ευκολότερα χρησιμοποιώντας μοντέλα που συσχετίζουν τις συχνότητες των κελιών. Προτού περάσουμε στη ανάλυση της περίπτωσης των πολυδιάστατων πινάκων, θα κάνουμε μια εισαγωγή στα μοντέλα συχνότητας κελιών για πίνακες δύων διαστάσεων.

Η διμεταβλητή κατανομή μπορεί επίσης να χαρακτηριστεί από την άποψη των μοντέλων πιθανοτήτων που σχετίζουν τις πιθανότητες ή τις πυκνότητες των κελιών. Παραπάνω, το μοντέλο ανεξαρτησίας αξιολογήθηκε χρησιμοποιώντας τον έλεγχο X^2. Αυτό το μοντέλο δίνεται ως

$$f_{ij} = f_{i.}f_{.j}, \qquad i = 1,2,\dots,r \; και \; j = 1,2,\dots,c$$

Εκτός από το μοντέλο ανεξαρτησίας, υπάρχουν και πιο απλά μοντέλα που θα μπορούσαν επίσης να χρησιμοποιηθούν.

Α) Μοντέλο ίσης πιθανότητας κελιών

Το απλούστερο μοντέλο για έναν διδιάστατο πίνακα είναι το μοντέλο ίσης πιθανότητας κελιών

$$f_{ij} = \frac{1}{rc}, \qquad i = 1,2,\dots,r \; και \; j = 1,2,\dots,c$$

υποδεικνύοντας ότι όλα τα rc πιθανά γεγονότα είναι εξίσου πιθανά.

Β) Σταθερές πυκνότητες γραμμών/στηλών

Τα μοντέλα αυτά υποθέτουν σταθερές πυκνότητες στις γραμμές ή τις στήλες του πίνακα συνάφειας και εκφράζονται μέσω των σχέσεων

$f_{ij} = (1/c)f_{i.}$, όπου δηλώνει σταθερές πυκνότητες στηλών $f_{i.} = 1/c$

και

$f_{ij} = (1/r)f_{.j}$, όπου δηλώνει σταθερές πυκνότητες γραμμών $f_{i.} = 1/r$

Στο μοντέλο σταθερής πυκνότητας στήλης η οριακή πυκνότητα σε κάθε στήλη ισούται με $1/c$. Αντίστοιχα στο μοντέλο σταθερής πυκνότητας γραμμής η οριακή πυκνότητα σε κάθε γραμμή είναι ίση με $1/r$.

Σε κάθε περίπτωση οι γραμμές και οι στήλες είναι ανεξάρτητες ενώ κάποιες από τις οριακές συχνότητες είναι σταθερές. Στο μοντέλο σταθερής πυκνότητας στήλης οι υπό συνθήκη όροι των στηλών δίνονται ως $f_{i.}(i|j) = 1/c$. Ομοίως, για το μοντέλο σταθερής πυκνότητας γραμμής έχουμε $f_{.j}(j|i) = 1/r$.

Γ) Το κορεσμένο μοντέλο

Εάν το μοντέλο ανεξαρτησίας δεν ευσταθεί, η από κοινού πυκνότητα μπορεί να γραφεί ως γινόμενο των επιδράσεων και των καταλοίπων. Το μοντέλο δίνεται ως

$$f_{ij} = [1/rc][rf_{i.}][c/f_{.j}][f_{ij}/f_{i.}f_{.j}]$$

Ο τέταρτος όρος

$$f_{ij}/f_{i.}f_{.j}$$

αντιστοιχεί στον όρο των υπολοίπων. Ο όρος αυτός εγγυάται ότι η εξίσωση ισχύει για όλες τις από κοινού πυκνότητες. Το παραπάνω μοντέλο ονομάζεται κορεσμένο, επειδή προσαρμόζεται «πιστά» στον πίνακα. Αυτός ο τελευταίος όρος ονομάζεται και ως «όρος αλληλεπίδρασης», επειδή μετρά την αλληλεπίδραση μεταξύ γραμμών και στηλών. Η απόκλιση του όρου αλληλεπίδρασης από την τιμή 1 υποδεικνύει το μέγεθος και την κατεύθυνση της απόκλισης από την υπόθεση της ανεξαρτησίας.

Το κορεσμένο μοντέλο έχει μια ξεχωριστή παράμετρο για κάθε παρατήρηση και έχει τέλεια εφαρμογή. Παρόλο που αυτό ακούγεται καλό, δεν είναι ένα χρήσιμο μοντέλο, καθώς δεν εξομαλύνει τα δεδομένα, ούτε έχει τα πλεονεκτήματα που έχει ένα απλούστερο μοντέλο, όπως η παρρησία (parsimony) . Ωστόσο, χρησιμεύει ως βάση για άλλα μοντέλα, όπως για τον έλεγχο της προσαρμογής του μοντέλου. Ένα κορεσμένο μοντέλο εξηγεί όλες τις διακυμάνσεις από τη συστηματική συνιστώσα του μοντέλου. Για ένα γενικευμένο γραμμικό μοντέλο, έστω $\tilde{\theta}$ υποδηλώνει την εκτίμηση της θ για το κορεσμένο μοντέλο που αντιστοιχεί στην εκτιμώμενη μέση τιμή $\tilde{\mu}_i = \tilde{y}_i$, για κάθε i. Για ένα συγκεκριμένο ακόρεστο μοντέλο, οι αντίστοιχες εκτιμήσεις μέγιστης πιθανοφάνειας υποδηλώνονται από $\hat{\theta}$ και $\hat{\mu}_i$ αντίστοιχα. Για την εκτίμηση μέγιστης πιθανοφάνειας και $L(\hat{\mu}\,;y)$ γι αυτό το μοντέλο και εκτίμηση μέγιστης πιθανοφάνειας $L(y\,;y)$ στην περίπτωση του κορεσμένου μοντέλου, η ποσότητα

$$-2\log\frac{\text{εκτίμηση μέγιστης πιθανοφάνειας μοντέλου}}{\text{εκτίμηση μέγιστης πιθανοφάνειας κορεσμένου μοντέλου}} = -2[L(\hat{\mu}\,;y) - L(y\,;y)$$

περιγράφει την απόκλιση του προσαρμοσμένου μοντέλου από το κορεσμένο μοντέλο, δηλαδή την έλλειψη προσαρμογής στα δεδομένα.

Δ) Ένα λογαριθμογραμμικό μοντέλο (Loglinear Model) για την ανεξαρτησία

Κάτω από την παραδοχή του ανεξαρτησίας μπορούμε να γράψουμε ένα λογαριθμογραμμικό μοντέλο ως

$$lnf_{ij} = ln\tilde{f}_{..} + \left[ln\tilde{f}_{i.} - ln\tilde{f}_{..}\right] + \left[ln\tilde{f}_{.J} - ln\tilde{f}_{..}\right]$$

όπου $\tilde{f}$ είναι ο γεωμετρικός μέσος όρος, δεδομένου ότι προσπαθούμε να μοντελοποιήσουμε πυκνότητες που είναι αναλογίες, η χρήση γεωμετρικών μέσων αντί αριθμητικών μέσων μπορεί να είναι προτιμότερη, καταλήγοντας ότι

$$f_{ij} = \tilde{f}_{..}\left[\frac{\tilde{f}_{i.}}{\tilde{f}_{..}}\right]\left[\frac{\tilde{f}_{.J}}{\tilde{f}_{..}}\right]$$

όταν ισχύει η ανεξαρτησία. Επιπλέον,

$$ln\tilde{f}_{..} = \frac{1}{rc}\sum_{j=1}^{c}\sum_{i=1}^{r} lnf_{ij},$$

$$ln\tilde{f}_{i.} = \frac{1}{c}\sum_{j=1}^{c} lnf_{ij}$$

και

$$ln\tilde{f}_{.J} = \frac{1}{r}\sum_{i=1}^{r} lnf_{ij}.$$

Για άλλη μια φορά αυτό το γινόμενο τριών όρων μπορεί να είναι θεωρηθεί ως ένας μέσος όρος πολλαπλασιασμένος με τις επιδράσεις των εκάστοτε ζευγών γραμμής-στήλης. Η επίδραση της γραμμής προκύπτει από το πηλίκο του μέσου όρου της πυκνότητας της κάθε γραμμής $\tilde{f}_{i.}$, στο συνολικό μέσο όρο των πυκνοτήτων κελιών, $\tilde{f}_{..}$. Ομοίως, η επίδραση της στήλης προκύπτει από το πηλίκο του μέσου όρου της πυκνότητας της κάθε στήλης $\tilde{f}_{.J}$, στο συνολικό μέσο όρο των πυκνοτήτων κελιών $\tilde{f}_{..}$. Επομένως, οι επιδράσεις γραμμών και στηλών είναι πλέον αναλογίες των μέσων πυκνοτήτων των κελιών.

i) **Παράμετροι του λογαριθμογραμικού μοντέλου**

Από τον παραπάνω ορισμό του λογαριθμογραμμικού μοντέλου, οι όροι αυτού μπορούν να θεωρηθούν ως μέσες τιμές. Μπορούμε λοιπόν να ορίσουμε τις ποσότητες (Jobson 1992)

$$\mu = \frac{1}{rc}\sum_{j=1}^{c}\sum_{i=1}^{r} lnf_{ij} = ln\tilde{f}_{..}$$

$$\mu_{1(i)} = \frac{1}{c}\sum_{j=1}^{c} lnf_{ij} - ln\tilde{f}_{..} = [ln\tilde{f}_{i.} - ln\tilde{f}_{..}]$$

$$\mu_{2(i)} = \frac{1}{r}\sum_{i=1}^{r} lnf_{ij} - ln\tilde{f}_{..} = \left[ln\tilde{f}_{.j} - ln\tilde{f}_{..}\right].$$

Η παράμετρος μ, είναι επομένως ο λογάριθμος της συνολικής γεωμετρικής μέσης τιμής των πυκνοτήτων. Από την άλλη, η παράμετρος $\mu_{1(i)}$ αντιπροσωπεύει τον λογάριθμο του λόγου του γεωμετρικού μέσου όρου των πυκνοτήτων της γραμμής i προς το συνολικό γεωμετρικό μέσο όρο. Ομοίως, η παράμετρος $\mu_{2(i)}$ είναι ο λογάριθμος του λόγου του γεωμετρικού μέσου όρου των πυκνοτήτων της στήλης j προς το συνολικό γεωμετρικό μέσο όρο. Οι παράμετροι $\mu_{1(i)}$ και $\mu_{2(i)}$ έχουν τις ιδιότητες

$$\sum_{i=1}^{r}\mu_{1(i)} = 0 \text{ και } \sum_{j=1}^{c}\mu_{2(i)} = 0.$$

Αυτές οι παράμετροι επομένως είναι παρόμοιες με τις παραμέτρους επίδρασης που χρησιμοποιούνται στην ανάλυση διασποράς (ANOVA). Κάτω από την υπόθεση της ανεξαρτησίας, το μοντέλο γίνεται

$$lnf_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)}.$$

Το μοντέλο αυτό είναι ευρέως γνωστό ως λογαριθμογραμμικό μοντέλο για την ανεξαρτησία διδιάστατων πινάκων συνάφειας .

ii) Λογαριθμογραμμικό μοντέλο με αλληλεπιδράσεις

Εάν η υπόθεση της ανεξαρτησίας δεν ευσταθεί για τα δεδομένα, θα πρέπει να προστεθεί στο παραπάνω μοντέλου ένας όρος που να αντιπροσωπεύει την απόκλιση από την ανεξαρτησία. Σε αυτή την περίπτωση, η πυκνότητα μπορεί να εκφραστεί υπό την μορφή (Jobson 1992)

$$f_{ij} = \tilde{f}_{..}\left[\frac{\tilde{f}_{i.}}{\tilde{f}_{..}}\right]\left[\frac{\tilde{f}_{.j}}{\tilde{f}_{..}}\right]\left[\frac{f_{ij}\tilde{f}_{..}}{\tilde{f}_{i.}\tilde{f}_{.j}}\right].$$

Όπως στα προηγούμενα μοντέλα, ο όρος των αλληλεπιδράσεων $\left[\frac{f_{ij}\tilde{f}_{..}}{\tilde{f}_{i.}\tilde{f}_{.j}}\right]$ αποτελεί το λόγο της πραγματικής πυκνότητας, προς την πυκνότητα του μοντέλου υπό την υπόθεση της πλήρους ανεξαρτησίας.

Επιπλέον, το λογαριθμογραμμικό μοντέλο, θα έχει έναν επιπλέον όρο που θα αντιστοιχεί στις αλληλεπιδράσεις και θα δίνεται από τη σχέση

$$ln f_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{12(ij)}$$

όπου

$$\mu_{12(ij)} = ln f_{ij} + ln\tilde{f}_{..} - ln\tilde{f}_{i.} - ln\tilde{f}_{.j}.$$

Ο όρος των αλληλεπιδράσεων ακολουθεί επίσης τις ιδιότητες $\sum_{i=1}^{r}\mu_{12(ij)} = 0$ και $\sum_{j=1}^{c}\mu_{12(ij)} = 0$. Το μοντέλο αυτό συχνά αναφέρεται στη βιβλιογραφία ως κορεσμένο λογαριθμογραμμικό μοντέλο, καθώς περιγράφει τις πυκνότητες του πίνακα με ακρίβεια και χωρίς κανένα περιορισμό.

2.9 Μέτρα συσχέτισης

Υπάρχουν πολλοί συντελεστές που μπορούν να χρησιμοποιηθούν για την αξιολόγηση της της ισχύος της συσχέτισης μεταξύ δύο μεταβλητών. Ανάλογα με τον τύπο της μελέτης, αυτά τα μέτρα, γνωστά και ως *μέτρα συσχέτισης* ή *μέτρα συνάφειας*, και μπορούν να αναπαρασταθούν με διάφορους τρόπους. Είναι σημαντικό για έναν ερευνητή να γνωρίζει ορισμένες στατιστικές διαφορές προκειμένου να κατανοήσει καλύτερα τα μέτρα συσχέτισης που χρησιμοποιούνται στα πλαίσια των στατιστικών αναλύσεων. Ο ερευνητής πρέπει να γνωρίζει ότι τα μέτρα συσχέτισης δεν είναι τα ίδια με τα *μέτρα σημαντικότητας* (significance) στην επιστημονική κοινότητα. Με την απλή τυχαία δειγματοληψία, τα μέτρα σημαντικότητας σχετίζονται συνήθως με την μηδενική υπόθεση, η οποία δείχνει ότι δεν υπάρχει στατιστικά σημαντική διαφορά μεταξύ μιας παρατηρούμενης συσχέτισης και μιας αναμενόμενης συσχέτισης. Τόσο μια σχέση που εμφανίζει υψηλή συσχέτισης αλλά δεν είναι στατιστικά σημαντική, όσο και μια σχέση που δείχνει χαμηλή συσχέτιση αλλά είναι στατιστικά σημαντική, είναι και τα δύο πιθανά αποτελέσματα.

Τα περισσότερα μέτρα συσχέτισης είναι φραγμένα σε κάποιο διάστημα, και συνήθως χαμηλές κατ' απόλυτη τιμή εκτιμήσεις αυτών δηλώνουν και ασθενή συσχέτιση. Για παράδειγμα, ένας συντελεστής συσχέτισης του Pearson (r) με τιμή μηδέν υποδηλώνει ότι δεν υπάρχει στατιστική γραμμική συσχέτιση μεταξύ των υπό εξέταση μεταβλητών. Η τιμή του συντελεστή μπορεί να αλλάξει με βάση την έρευνα. Όταν ο συντελεστής συσχέτισης (r) έχει τιμή κοντά στο

1 σε μια ανάλυση συσχέτισης, υποδηλώνει ότι υπάρχει ισχυρή συσχέτιση μεταξύ των μεταβλητών. Όταν η εκτίμηση για το συντελεστή μιας ανεξάρτητης μεταβλητής (β) σε μια ανάλυση παλινδρόμησης έχει τιμή γύρω από το μηδέν, υποδηλώνει επίσης ότι δεν υπάρχει συσχέτιση μεταξύ της μεταβλητής αυτής και της εξαρτημένης. Κατ' αντιστοιχία, στις επόμενες ενότητες θα οριστούν και θα αναλυθούν κάποια γνωστά μέτρα συσχέτισης που αφορούν κατηγορικές μεταβλητές.

2.9.1 Ομοιογενείς αναλογίες

Μια εναλλακτική μέθοδος για να διατυπώσουμε το μοντέλο υπό την υπόθεση της ανεξαρτησίας είναι οι υπό συνθήκη αναλογίες, κάτι το οποίο είναι ιδιαίτερα κατάλληλο αν η μια μεταβλητή έχει το ρόλο της ανεξάρτητης και η άλλη μεταβλητή της εξαρτημένης σε ένα πίνακα συνάφειας. Αν υποθέσουμε ότι μηδενική υπόθεση είναι αληθής, τότε ισχύει ότι

$$\pi_{j|i} = \frac{f_{i.}f_{.j}}{f_{i.}f_{..}} = \frac{f_{.j}}{f_{..}} = \pi_{.j}$$

δηλαδή οι υπό συνθήκη αναλογίες $\pi_{j|i}$ που αντιστοιχούν στην γραμμή i και στη στήλη j είναι ομοιογενείς και ισούνται με όλες τις αναλογίες στην ίδια γραμμή. Το ίδιο ισχύει λόγω συμμετρίας αντίστοιχα, κάτω από την υπόθεση της ανεξαρτησίας, και για τις αναλογίες που αντιστοιχούν στις στήλες. Έτσι η απόκλιση των αναλογιών εντός γραμμών ή στηλών αποτελεί ένδειξη κατά της υπόθεσης της ανεξαρτησίας.

2.9.2 Λόγος σχετικών πιθανοτήτων – Odds ratio

Ο λόγος σχετικών πιθανοτήτων (OR) είναι ένα μέτρο που χρησιμοποιείται ευρέως στις ιατρικές μελέτες για να δείξει την πιθανότητα ενός ατόμου να πάσχει από μια ασθένεια. Όπως και τα υπόλοιπα μέτρα, εκφράζει τον βαθμό συνάφειας μεταξύ δυο κατηγορικών μεταβλητών, με έναν πιο σαφή τρόπο ερμηνείας ενός $2x2$ πίνακα συνάφειας. Ο τύπος υπολογισμού του είναι

$$OR = \frac{\frac{\pi_{11}}{\pi_{12}}}{\frac{\pi_{21}}{\pi_{22}}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

Όταν έχουμε $OR > 1$ τότε η πιθανότητα ενός γεγονότος να συμβεί είναι μεγαλύτερη του 0.5, ενώ όταν $OR < 1$ η πιθανότητα ενός γεγονότος να συμβεί είναι μικρότερη του 0.5, επομένως, όσο μεγαλύτερες είναι οι σχετικές πιθανότητες ενός γεγονότος να συμβεί, τόσο υψηλότερη η πιθανότητα ότι το γεγονός θα συμβεί και όσο μικρότερα είναι τα odds ενός γεγονότος να συμβεί, τόσο χαμηλότερη η πιθανότητα ότι το γεγονός θα συμβεί. Όταν $OR = 1$, τότε η πιθανότητα το

γεγονός να συμβεί ισούται με την πιθανότητα το γεγονός να μην συμβεί και όταν το $OR = 0$, τότε η πιθανότητα το γεγονός να συμβεί είναι 0. Επομένως, η ένταση της σχέσης μεταξύ δυο μεταβλητών μπορεί να εκφραστεί, με τον βαθμό με τον οποίο οι δυο σχετικές πιθανότητες διαφέρουν μέσω των OR.

2.9.3 Σχετικός κίνδυνος – Relative Risk

Για μια διχότομη μεταβλητή, μόνο μια αναλογία επαρκεί για να παρέχει όλη την πληροφορία, καθώς λόγω συμπληρωματικότητας η δεύτερη αναλογία μπορεί να υπολογιστεί ακριβώς και μονοσήμαντα. Αντίστοιχα, σε μια μεταβλητή με J κατηγορίες, απαιτούνται $J - 1$ αναλογίες για να έχουμε όλη την πληροφορία. Αν έχουμε δυο μεταβλητές, όπου X είναι η ανεξάρτητη μεταβλητή και Y είναι η εξαρτημένη μεταβλητή, με μόνο δυο κατηγορίες, τότε χρησιμοποιώντας τη μια κατηγορία, ως την κατηγορία αναφοράς – reference level, μπορούμε να υπολογίσουμε το πηλίκο δυο υπό συνθήκη αναλογιών:

$$RR = \frac{\pi_{j|2}}{\pi_{j|1}}$$

το οποίο ονομάζεται σχετικός κίνδυνος – relative risk (RR), καθώς συνήθως χρησιμοποιείται σε εφαρμογές Ιατρικής όπου η ανεξάρτητη μεταβλητή αντιστοιχεί σε κάποια θεραπεία και η εξαρτημένη μεταβλητή αντιστοιχεί στο αποτέλεσμα της θεραπείας. Η ποσότητα αυτή δηλώνει πόσες φορές πιο πιθανό είναι να συναντήσουμε την κατηγορία j στην μεταβλητή Y, αν έχουμε την κατηγορία 2 στην μεταβλητή X (αριθμητής) έναντι της κατηγορίας 1 στην μεταβλητή X (παρονομαστής). Για παράδειγμα, στον Πίνακα 4 παρουσιάζονται οι αναλογίες ανά γραμμή, όπου φαίνεται πως αν είσαι κάτοχος τίτλου τριτοβάθμιας εκπαίδευσης είναι $RR = \frac{\pi_{2|2}}{\pi_{2|1}} = \frac{0,694}{0,577} = 1.203$ φορές πιο πιθανό να είσαι καπνιστής σε σχέση με το να είσαι κάτοχος τίτλου δευτεροβάθμιας εκπαίδευσης.

Πίνακας 4. Σχέση μεταξύ εκπαίδευσης και καπνίσματος (αναλογίες ανά γραμμή)

	Κάπνισμα	
	Όχι	Ναι
Δευτεροβάθμια	0,423	0,577
Τριτοβάθμια	0,306	0,694

Ο σχετικός κίνδυνος είναι ένας μη αρνητικός αριθμός, ο οποίος δεν έχει άνω φράγμα. Όταν είναι ίσος με 1, τότε οι κατηγορικές μεταβλητές είναι ανεξάρτητες, δηλαδή δεν υπάρχει διαφορά στον κίνδυνο μεταξύ των ομάδων, ενώ όταν είναι διάφορος του 1, οι μεταβλητές είναι εξαρτημένες ή συσχετισμένες. Συγκεκριμένα, για τιμές μικρότερες του 1 το ενδεχόμενο είναι λιγότερο πιθανό να συμβεί στην ομάδα του αριθμητή από ότι στην ομάδα του παρονομαστή και επομένως έχουμε αρνητική συνάφεια, ενώ για τιμές μεγαλύτερες του 1 το ενδεχόμενο είναι περισσότερο πιθανό να συμβεί στην ομάδα του αριθμητή από ότι στην ομάδα του παρονομαστή και επομένως έχουμε θετική συνάφεια.

Η έννοια του σχετικού κινδύνου, χρησιμοποιείται ευρέως στις ιατρικές επιστήμες ως παράγοντας κινδύνου, όπως για παράδειγμα στην μελέτη της σχέσης μεταξύ μιας θεραπείας και μιας ασθένειας και επιτρέπει σε έναν ερευνητή να συγκρίνει τις σχετικές πιθανότητες επιβίωσης ενός ατόμου, εάν ανήκει σε μια ομάδα σε σχέση με την επιβίωση ενός ατόμου που ανήκει σε μια ομάδα χαμηλότερου κινδύνου. Όταν οι υπό συνθήκη αναλογίες είναι πολύ κοντά στα άκρα του διαστήματος $[0, 1]$, ο σχετικός κίνδυνος παρέχει καλύτερη πληροφόρηση από την σύγκριση αναλογιών. Για παράδειγμα, σε μελέτη που συγκρίνει δυο θεραπείες Α και Β με βάση την απόκριση στην εξαρτημένη μεταβλητή 1, δηλαδή το ποσοστό των ασθενών που πεθαίνουν, η διαφορά $\pi_{1|2} = 0.01$ και $\pi_{1|1} = 0.001$ είναι πολύ πιο σημαντική, από ότι η διαφορά $\pi_{1|2} = 0.51$ και $\pi_{1|1} = 0.501$, αν και στις δυο περιπτώσεις η διαφορά των αναλογιών είναι ίση με $\pi_{1|2} - \pi_{1|1} = 0.009$. Στην περίπτωση αυτή, ο σχετικός κίνδυνος είναι ίσος με $RR_1 = \frac{0.01}{0.001} = 10$ και $RR_1 = \frac{0.51}{0.501} = 1.02$ αντίστοιχα. Με άλλα λόγια στην πρώτη περίπτωση, το ποσοστό θανάτου όταν ένας ασθενής είναι στην ομάδα Α είναι 900% μεγαλύτερο από το ποσοστό θανάτου ενός ασθενή στην ομάδα Β, ενώ στην δεύτερη περίπτωση, το ποσοστό θανάτου για την ομάδα Α είναι μόλις 2% μεγαλύτερο του ποσοστού θανάτου της ομάδας Β.

2.9.4 Συντελεστής φ

Ο συντελεστής φ (Yule, 1912), εξαλείφει την επίδραση του μεγέθους του δείγματος, διαιρώντας το στατιστικό χ^2 με το μέγεθος του δείγματος n και στην συνέχεια υπολογίζοντας την τετραγωνική του ρίζα. Ο τύπος υπολογισμού του συντελεστή είναι:

$$\varphi = \sqrt{\frac{\chi^2}{n}}$$

Εναλλακτικά, στην περίπτωση ενός $2x2$ πίνακα συνάφειας, ο συντελεστής ορίζεται ως

$$\varphi = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1.}n_{2.}n_{.1}n_{.2}}}$$

και οι δυο σχέσεις είναι αλγεβρικά ισοδύναμες. Ο συντελεστής φ για 2x2 πίνακες κυμαίνεται στο διάστημα [-1, 1] αν και αυτό δεν συμβαίνει πάντα, καθώς το άνω και κάτω όριο του συντελεστή εξαρτάται από συγκεκριμένες συνθήκες. Η χρήση του συντελεστή φ περιορίζεται μόνο σε 2x2 πίνακες συνάφειας, που περιέχουν διχοτομημένες ονοματικές μεταβλητές, διότι για πίνακες μεγαλύτερης διάστασης η τιμή του συντελεστή μπορεί να ξεπεράσει την τιμή 1 . Οι Siegel και Castellan (1988) σημειώνουν, ότι όταν οι δυο μεταβλητές που συσχετίζονται είναι διατακτικές, τότε χρησιμοποιώντας τον συντελεστή φ χάνεται πληροφόρηση και εξ' αιτίας αυτού, κάτω από αυτές τις συνθήκες είναι προτιμότερο να χρησιμοποιούμε εναλλακτικά μέτρα συνάφειας, σχεδιασμένα για διατεταγμένους πίνακες. Οι Carroll (1961) και Guilford (1965) σημειώνουν, ότι αναγκαία συνθήκη για να ισούται ο συντελεστής φ με 1− ή 1 (τέλεια συνάφεια), σε έναν 2x2 πίνακα, είναι τα περιθώρια αθροίσματα κάθε γραμμής και στήλης να είναι ίσα. Επιπλέον, ο Liu (1980) σημειώνει ότι το ίδιο ισχύει όταν δυο συμμετρικά αντίθετα διαγώνια κελιά είναι 0.

2.9.5 Συντελεστής Cramer's V

Ο συντελεστής συνάφειας V που εισάχθηκε από τον Cramer (1946) είναι μια επέκταση του συντελεστή φ, για πίνακες συνάφειας με μεγαλύτερη διάσταση . Ο συντελεστής V κυμαίνεται στο διάστημα από 0 έως 1, με $V = 0$, μόνο όταν οι δυο μεταβλητές έχουν τα περιθώρια αθροίσματα στον πίνακα συνάφειας ίσα μεταξύ τους. Επομένως, όσο πιο άνισα τα περιθώρια αθροίσματα γραμμών και στηλών, τόσο ο συντελεστής θα είναι μεγαλύτερος του μηδέν. Ο τύπος υπολογισμού του συντελεστή είναι

$$Cramer's\ V = \sqrt{\frac{\varphi^2}{\min\{R, C\}}}$$

όπου R, C είναι το πλήθος των γραμμών και στηλών αντίστοιχα του πίνακα συνάφειας, ενώ η εκτίμηση για το φ αντιστοιχεί στην τιμή $\frac{\chi^2}{n}$.

2.9.6 Συντελεστής συνάφειας λ

Ο συντελεστής συνάφειας λ (Goodman & Kruskal, 1954) ανήκει στα μέτρα προγνωστικής συνάφειας και αποτελεί ένα μέτρο της αναλογικής μείωσης του σφάλματος, όταν οι τιμές της ανεξάρτητης μεταβλητής χρησιμοποιούνται για να προβλέψουν τις τιμές της εξαρτημένης μεταβλητής. Οι τιμές που λαμβάνει ο συντελεστής λ ανήκουν στο διάστημα [0,1], με $\lambda = 0$,όταν οι μεταβλητές είναι ανεξάρτητες, και $\lambda = 1$, όταν με βάση τη μια μεταβλητή μπορεί να γίνει τέλεια προγνώστη των τιμών της άλλης.

Αν υποθέσουμε ότι X είναι η ανεξάρτητη κατηγορική μεταβλητή και Y είναι η εξαρτημένη, τότε το λ δίνεται από

$$\lambda = \frac{\sum_i \max_j \pi_{ij} - \max_j \pi_{.j}}{1 - \max_j \pi_{.j}}$$

όπου $\max_j \pi_{ij}$ είναι η μέγιστη πιθανότητα για κάθε κατηγορία i της ανεξάρτητης μεταβλητή X και $\max_j \pi_{.j}$ η μέγιστη περιθώρια πιθανότητα της εξαρτημένης μεταβλητής Y. Η τιμή του συντελεστή λ κυμαίνεται στο διάστημα από 0 έως 1 και δεν ορίζεται, όταν όλες οι παρατηρήσεις συγκεντρώνονται μόνο σε μια από τις στήλες της εξαρτημένης μεταβλητής. Όταν $\lambda = 0$ τότε οι μεταβλητές είναι ανεξάρτητα κατανεμημένες, δηλαδή η γνώση της X δεν προσφέρει καμία πληροφόρηση για την Y, ενώ όταν $\lambda = 1$ η μεταβλητή X μπορεί να προβλέψει τέλεια την μεταβλητή Y.

Στον Πίνακα 5, παρουσιάζονται συνοπτικά κάποιες πληροφορίες για διάφορα μέτρα συσχέτισης κατηγορικών μεταβλητών.

Πίνακας 5. Μέτρα συνάφειας

Μέτρο συνάφειας	**Πεδίο Ορισμού**	**Συμμετρία**	**Είδος Δεδομένων**	**Διάσταση Πίνακα**	**Σχόλια**
Relative Risk(RR)	$[0, +\infty)$	Όχι	Ονοματικά και Διατακτικά	2×2	Χρησιμοποιείται ευρέως στις ιατρικές επιστήμες ως παράγοντας κινδύνου, για τη σύγκριση ομάδων. Όταν RR=1, τότε ανεξαρτησία
Odds Ratio(θ)	$[-\infty, +\infty)$	Ναι	Ονοματικά και Διατακτικά	2×2	Συνδέεται με τον σχετικό κίνδυνο μέσω σχέσης και μπορεί να χρησιμοποιηθεί και για πίνακες μεγαλύτερης διάστασης. Όταν $\theta = 1$ τότε ανεξαρτησία.
Yule's Q	$[-1, +1)$	Ναι	Ονοματικά και Διατακτικά	2×2	Βασίζεται στο odds ratio, αποτελεί ειδική περίπτωση του διατακτικού μέτρου gamma. Δεν συνιστάται η χρήση του όταν η συχνότητα των κελιών είναι πολύ μικρή.

					Έχει την τάση να αυξάνει τον βαθμό συνάφειας.
Yule's Y	$[-1,+1)$	Ναι	Ονοματικά και Διατακτικά	2×2	Γενικά είναι μικρότερος του Yule's Q, είναι λιγότερος ευαίσθητος από τις διαφορές των περιθώριων κατανομών απ' ότι ο Q. Ερμηνεύεται όπως ο συντελεστής συσχέτισης r του Pearson.
Yule's φ	$[-1,+1]$ ή $[0, sqrt(q-1)]$	Ναι	Ονοματικά και Διατακτικά	2×2 *ή* $I \times J$	Καθώς η τιμή του μπορεί να ξεπεράσει την τιμή 1, δεν θεωρείται το πιο κατάλληλο μέτρο. Είναι πολύ ευαίσθητος στις περιθώριες κατανομές και η σύγκριση του μεταξύ πινάκων μπορεί να είναι παραπλανητική. Αλγεβρικά ισούται με το συντελεστή συσχέτισης r του Pearson.
Tshuprow's T	$[0,1]$	Ναι	Ονοματικά και Διατακτικά	$I \times J$	Χρησιμοποιείται μόνο σε τετραγωνικούς πίνακες. T=φ για 2×2 πίνακες. Η χρήση του είναι περιοσμένη.
Cramer's V	$[0,1]$	Ναι	Ονοματικά και Διατακτικά	$I \times J$	$V = T = \varphi$ για 2×2. Θεωρείται το πιο διαδεδομένο, καθώς ανεξάρτητα από το μέγεθος του πίνακα κατανέμεται πιο κανονικά στο $[0,1]$. Εξαρτάται από την διάσταση του πίνακα και χρειάζεται προσοχή όταν συγκρίνουμε πίνακες. Επίσης $V = \tau^2$
Pearson's C	$[0,1]$ ή $[0,0.71]$	Ναι	Ονοματικά και Διατακτικά	$I \times J$	Για 2×2 πίνακες κυμαίνεται στο $[0,0.71]$. Ακόμα και όταν οι μεταβλητές είναι πλήρως εξαρτημένες, θα είναι < 1. Εξαρτάται από τη διάσταση του πίνακα και χρειάζεται προσοχή όταν συγκρίνουμε πίνακες. Προτείνεται να χρησιμοποιείται για πίνακες $> 5x5$.
Sakoda's C_{adj}	$[0,1]$	Ναι	Ονοματικά και Διατακτικά	$I \times J$	

Goodman-Kruskal's λ_α	[0,1]	Όχι	Ονοματικά και Διατακτικά	$I \times J$	Δεν ορίζεται όταν οι παρατηρήσεις συγκεντρώνονται σε μία από τις στήλες της εξαρτημένης μεταβλητής. Είναι ευαίσθητος στις ανισότητες των περιθώριων αθροισμάτων γραμμών και στηλών. Μπορεί $\lambda_\alpha = 0$ χωρίς να υπάρχει στατική ανεξαρτησία.
Goodman-Kruskal's λ_β	[0,1]	Όχι	Ονοματικά και Διατακτικά	$I \times J$	Δεν ορίζεται όταν οι παρατηρήσεις συγκεντρώνονται σε μία από τις στήλες της εξαρτημένης μεταβλητής. Είναι ευαίσθητος στις ανισότητες των περιθώριων αθροισμάτων γραμμών και στηλών. Μπορεί $\lambda_\beta = 0$ χωρίς να υπάρχει στατική ανεξαρτησία.
Goodman-Kruskal's λ	[0,1]	Ναι	Ονοματικά και Διατακτικά	$I \times J$	Δεν ορίζεται όταν οι παρατηρήσεις συγκεντρώνονται σε ένα μοναδικό κελί του πίνακα. Είναι ευαίσθητος στις ανισότητες των περιθώριων αθροισμάτων γραμμών και στηλών. Μπορεί $\lambda = 0$ χωρίς να υπάρχει στατική ανεξαρτησία
Goodman-Kruskal's τ_α	[0,1]	Όχι	Ονοματικά και Διατακτικά	$I \times J$	
Goodman-Kruskal's τ_β	[0,1]	Όχι	Ονοματικά και Διατακτικά	$I \times J$	
Goodman-Kruskal's τ	[0,1]	Όχι	Ονοματικά και Διατακτικά	$I \times J$	
Theil's $U(Y\|X)$	[0,1]	Όχι	Ονοματικά και Διατακτικά	$I \times J$	Η ερμηνεία των Theil's βασίζεται στη θεωρία της πληροφορίας και στην εντροπία των Shannon. Διαφοροποιούνται από τον συντελεστή λ, με την έννοια ότι λαμβάνει
Theil's $U(X\|Y)$	[0,1]	Όχι	Ονοματικά και Διατακτικά	$I \times J$	

					υπόψη του όλη την κατανομή της εξαρτημένης μεταβλητής. Ένα μειονέκτημα είναι ότι υπάρχει μια θετική σχέση μεταξύ του αριθμού των κατηγοριών και της βαρύτητας της διακύμανσης, εισάγοντας διφορούμενες ερμηνείες κατά την αξιολόγηση της διακύμανσης των μεταβλητών και της μεταξύ τους σχέσης.
Theil's U_{sym}	[0,1]	Ναι	Ονοματικά και Διατακτικά	$I \times J$	

Κεφάλαιο 3^{ο} . Ανάλυση πολυδιάσταστων πινάκων συνάφειας

Σε αυτή την ενότητα θα ασχοληθούμε με τη στατιστική συμπερασματολογία πολυδιάστατων πινάκων συνάφειας. Θα ξεκινήσουμε με την μια περίληψη τεχνικών για την ανάλυση ενός τρισδιάστατου πίνακα συνάφειας. Οι τεχνικές που θα παρουσιαστούν, μπορούν να επεκταθούν εύκολα και σε πίνακες υψηλότερων (τεσσάρων και πλέον) διαστάσεων.

3.1. Τρισδιάστατοι πίνακες συνάφειας

Ένας τρισδιάστατος πίνακας συνάφειας, προκύπτει από τη διαταξινόμηση (cross-classification) των κατηγοριών τριών ποιοτικών-κατηγορικών τυχαίων μεταβλητών. Γεωμετρικά ο πίνακας μπορεί να θεωρηθεί ότι αποτελείται από γραμμές, στήλες και επίπεδα (layers). Οι δείκτες για τις γραμμές, τις στήλες και τα επίπεδα θα συμβολίζονται με i, j και k αντίστοιχα. Αντίστοιχα, ο αριθμός των γραμμών, στηλών και επιπέδων θα συμβολίζεται με r, c και l, τα οποία εμφανώς πρέπει να είναι θετικοί ακέραιοι αριθμοί. Η πυκνότητα πιθανότητας για το κελί (i, j, k) αποτυπώνεται ως f_{ijk} και η θεωρητική συχνότητα των κελιών ορίζεται ως $F_{ijk} = nf_{ijk}$ για συνολική συχνότητα πίνακα ίση με n. Η κατανομή ενός δείγματος μεγέθους n στο σύνολο των $r \times c \times l$ κελιών παράγουν τις συχνότητες n_{ijk}. Ο παρακάτω πίνακας παρουσιάζει διαγραμματικά τις συχνότητες n_{ijk} για ένα δείγμα μεγέθους n.

Πίνακας 6. Τρισδιάστατος πίνακας συνάφειας

		Στήλες			
Επίπεδα	Γραμμές	1	2	⋯	c
1	1	n_{111}	n_{121}	⋯	n_{1c1}
	2	n_{211}	n_{221}	⋯	n_{2c1}
	⋮	⋮	⋮		⋮
	r	n_{r11}	n_{r21}	⋯	n_{rc1}
2	1	n_{112}	n_{122}	⋯	n_{1c2}
	2	n_{212}	n_{222}	⋯	n_{2c2}
	⋮	⋮	⋮		⋮
	r	n_{r12}	n_{r22}	⋯	n_{rc2}
l	1	n_{11l}	n_{12l}	⋯	n_{1cl}
	2	n_{21l}	n_{22l}	⋯	n_{2cl}
	⋮	⋮	⋮		⋮
	r	n_{r1l}	n_{r2l}	⋯	n_{rcl}

Αντιλαμβανόμαστε, ότι η διαγραμματική αναπαράσταση πινάκων τεσσάρων και πλέον διαστάσεων γίνεται αρκετά δύσκολη.

Όπως και στην περίπτωση του δισδιάστατου πίνακα συνάφειας, μπορούμε να ορίσουμε και τα αντίστοιχα σύνολα γραμμών, στηλών και επιπέδων. Για τους τρεις δυνατούς δισδιάστατους πίνακες συνάφειας, που προκύπτουν από τον Πίνακα 6, προκύπτουν τα περιθώρια σύνολα $n_{ij\cdot}, n_{i.k}$ και $n_{.jk}$. Τέλος, για τις τρεις εξεταζόμενες μεταβλητές, μπορούμε να πάρουμε τα μονομεταβλητά περιθώρια σύνολα $n_{i..}$, $n_{.j.}$ και $n_{..k}$.

3.2. Μοντέλα τριδιάστατων πινάκων συνάφειας

Τα μοντέλα που αναφέρθηκαν στο κεφάλαιο 2 για δισδιάστατους πίνακες συνάφειας, μπορούν να θεωρηθούν ως ειδικές περιπτώσεις του συνόλου των πιθανών μοντέλων ενός τρισδιάστατου πίνακα. Ξεκινάμε με το μοντέλο της ανεξαρτησίας για τρισδιάστατους πίνακες. Το

μοντέλο ανεξαρτησίας απαιτεί ότι η από κοινού πυκνότητα f_{ijk} του κελιού (i, j, k) είναι ίση με το γινόμενο των τριών μονομεταβλητών περιθωρίων πυκνοτήτων

$$f_{ijk} = f_{i..} f_{.j.} f_{..k}.$$

Η θεωρητική συχνότητα για μέγεθος δείγματος πλήθος n, δίνεται από τη σχέση

$$F_{ijk} = nf_{ijk} = F_{i..} F_{.j.} F_{..k} / n^2$$

όπου $F_{i..} = nf_{i..}$, $F_{.j.} = nf_{.j.}$ και $F_{..k} = nf_{..k}$.

3.2.1. Συμπερασματολογία για το μοντέλο ανεξαρτησίας

Δεδομένου μεγέθους δείγματος πλήθους n, οι εκτιμητές μέγιστης πιθανοφάνειας των εκτιμώμενων συχνοτήτων των κελιών κάτω από την υπόθεση της ανεξαρτησίας δίνονται ως

$$E_{ijk} = \frac{n_{i..} n_{.j.} n_{..k}}{n^2}, \text{ για } i = 1, 2, \dots, r, \ j = 1,2, \dots, c, \ k = 1,2, \dots, l.$$

Όπως και στην περίπτωση του δισδιάστατου πίνακα συνάφειας, οι συχνότητες των κελιών εξαρτώνται μόνο από τα περιθώρια σύνολα των γραμμών, στηλών και επιπέδων. Χρησιμοποιώντας τις εκτιμώμενες αναμενόμενες συχνότητες E_{ijk}, η δοκιμή καλής προσαρμογής X^2 για το μοντέλο ανεξαρτησίας, πραγματοποιείται χρησιμοποιώντας το στατιστικό

$$G^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{l} \frac{\left(E_{ijk} - n_{ilk}\right)^2}{E_{ijk}}$$

ή το στατιστικό

$$H^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{l} n_{ijk} (\ln n_{ijk} - \ln E_{ijk}).$$

Και τα δύο παραπάνω στατιστικά ακολουθούν κατανομή X^2 με $(rcl - r - l - c + 2)$ βαθμούς ελευθερίας λαμβάνοντας υπόψιν ότι ισχύει η υπόθεση της ανεξαρτησίας.

3.2.2. Άλλα μοντέλα τρισδιάστατων πινάκων

Για το υπόλοιπο αυτής της ενότητας, θεωρούμε ότι τα εξεταζόμενα μοντέλα δειγματοληψίας ακολουθούν πολυωνυμική ή Poisson κατανομή. Όπως και στην περίπτωση του

δισδιάστατου πίνακα, οι δύο κατανομές είναι ισοδύναμες όταν το μέγεθος του δείγματος n, είναι σταθερό. Επειδή το από κοινού πολυωνυμικό μοντέλο θέτει επιπρόσθετους περιορισμούς σε ορισμένες περιθώριες ποσότητες, πρέπει να τηρούνται πρόσθετες προϋποθέσεις για να επιτύχουμε εκτιμήσεις μέγιστης πιθανότητας. Εάν το μοντέλο της ανεξάρτητης δεν ταιριάζει στα δεδομένα του τρισδιάστατου πίνακα, είναι συχνά χρήσιμο να καθοριστεί ένα λιγότερο περιοριστικό μοντέλο. Τα παρακάτω μοντέλα, επιτρέπουν διάφορα επίπεδα εξάρτησης μεταξύ των τριών μεταβλητών που εμπεριέχονται στην ανάλυση.

Α) Μερική Ανεξαρτησία

Δεδομένου ότι υπάρχουν τρεις μεταβλητές στον πίνακα συνάφειας, είναι δυνατόν να υπάρχουν δύο μεταβλητές που σχετίζονται μεταξύ τους και παράλληλα να είναι ανεξάρτητες από την τρίτη εξεταζόμενη μεταβλητή. Αυτό το μοντέλο ονομάζεται μοντέλο μερικής ανεξαρτησίας, και δίνεται από την εξίσωση (Jobson 1992)

$$f_{ijk} = (f_{ij.})(f_{..k}).$$

Σε αυτή την περίπτωση, η τρίτη μεταβλητή με δείκτη k είναι ανεξάρτητη από τις υπόλοιπες δύο με δείκτες i και j. Η θεωρητική συχνότητα δίνεται από

$$F_{ijk} = \frac{F_{ij.}F_{..k}}{n}$$

και εκτιμάται από την ποσότητα

$$E_{ijk} = \frac{n_{ij.}n_{..k}}{n}.$$

Σε αυτή την περίπτωση, η δοκιμή X^2 καλής προσαρμογής χαρακτηρίζεται $(rc-1)(l-1)$ βαθμούς ελευθερίας (degrees of freedom).

Β) Υπό συνθήκη ανεξαρτησία

Ένα μοντέλο υπό συνθήκης ανεξαρτησίας, για ένα τρισδιάστατο πίνακα, θα πρέπει να επιτρέπει την ανεξαρτησία μεταξύ δύο μεταβλητών φιξάροντας/σταθεροποιώντας την Τρίτη μεταβλητή του πίνακα. Ένα παράδειγμα τέτοιου μοντέλου, μπορεί να διατυπωθεί μέσω της σχέσης

$$f_{ijk} = \frac{f_{i.k}f_{.jk}}{f_{..k}}$$

όπου οι μεταβλητές με δείκτες i και j είναι ανεξάρτητες σε κάθε επίπεδο της μεταβλητής με δείκτη k. Η θεωρητική συχνότητα δίνεται ως

$$F_{ijk} = \frac{F_{i.k}F_{.jk}}{F_{..k}}$$

ενώ ο εκτιμητής μέγιστης πιθανοφάνειας της ποσότητας αυτής, δίνεται από τη σχέση

$$E_{ijk} = \frac{n_{i.k}n_{.jk}}{n_{..k}}.$$

Σε αυτή την περίπτωση, η δοκιμή X^2 καλής προσαρμογής χαρακτηρίζεται από $l(r-1)(c-1)$ βαθμούς ελευθερίας.

Γ) Μη ύπαρξη αλληλεπιδράσεων τριών κατευθύνσεων

Το επόμενο βήμα για τη μετάβαση σε μοντέλα με πιο χαλαρές προϋποθέσεις, είναι να υποθέσουμε ότι το κάθε ζεύγος μεταβλητών σχετίζεται, αλλά ότι η σχέση μεταξύ οποιουδήποτε ζεύγους μεταβλητών δεν εξαρτάται από το επίπεδο της τρίτης μεταβλητής. Αυτό το μοντέλο συνήθως αναφέρεται ως μοντέλο δίχως ύπαρξη αλληλεπιδράσεων τριών κατευθύνσεων (no three-way interaction). Σε αυτή την περίπτωση μοντέλου, δεν είναι δυνατό να δοθεί έκφραση για τις ποσότητες f_{ijk} και F_{ijk} ώστε να τις εκτιμήσουμε απευθείας μέσω των εκτιμητών μέγιστης πιθανοφάνειας E_{ijk}. Για αυτό το μοντέλο οι ποσότητες E_{ijk} υπολογίζονται μέσω μιας επαναληπτικής διαδικασίας γνωστής ως επαναληπτική αναλογική προσαρμογή (iterative proportional fitting).

Εφόσον το προσαρμοσμένο μοντέλο προϋποθέτει ότι μεταξύ όλων των πιθανών ζευγών μεταβλητών υπάρχει συσχέτιση, αλλά χωρίς να υπάρχει αλληλεπίδραση μεταξύ και των τριών μεταβλητών, χρειαζόμαστε απλά ένα μοντέλο που διατηρεί τα τρία σύνολα δισδιάστατων οριακών συνόλων $n_{ij.}$, $n_{.jk}$ και $n_{i.k}$. Τα βήματα για την επαναληπτική αναλογική προσαρμογή είναι τα εξής (Jobson 1992):

Βήμα 1. Υπολογισμός των παρατηρούμενων περιθωρίων συνόλων $n_{ij.}$, $n_{.jk}$ και $n_{i.k}$.

Βήμα 2. Θέτουμε την αρχική τιμή 1, σε όλες τις εκτιμώμενες συχνότητες των κελιών, δηλαδή $E^0_{ijk} = 1$, για όλα τα i, j, k.

Βήμα 3. Υπολογισμός των νέων εκτιμήσεων των ποσοτήτων E_{ijk} έτσι ώστε να αθροίζουν στα περιθώρια σύνολα $n_{ij.}$, μέσω της σχέσης

$$E^1_{ijk} = E^0_{ijk}\left[\frac{n_{ij.}}{E^0_{ij.}}\right] \quad \gamma\iota\alpha\ \acute{o}\lambda\alpha\ \tau\alpha\ i,j,k.$$

Βήμα 4. Υπολογισμός των νέων εκτιμήσεων των ποσοτήτων E_{ijk} έτσι ώστε να αθροίζουν στα περιθώρια σύνολα $n_{i.k}$, μέσω της σχέσης

$$E^2_{ijk} = E^1_{ijk}\left[\frac{n_{i.k}}{E^1_{i.k}}\right] \quad \gamma\iota\alpha\ \acute{o}\lambda\alpha\ \tau\alpha\ i,j,k.$$

Βήμα 5. Υπολογισμός των νέων εκτιμήσεων των ποσοτήτων E_{ijk} έτσι ώστε να αθροίζουν στα περιθώρια σύνολα $n_{.jk}$, μέσω της σχέσης

$$E^3_{ijk} = E^2_{ijk}\left[\frac{n_{.jk}}{E^2_{.jk}}\right] \quad \gamma\iota\alpha\ \acute{o}\lambda\alpha\ \tau\alpha\ i,j,k.$$

Η διαδικασία που περιγράφεται από τα βήματα 3 έως 5 επαναλαμβάνονται έως ότου οι μεταβολές στις ποσότητες E_{ijk} να είναι μικρότερες από κάποια προκαθορισμένη τιμή που ορίζεται από το χρήστη.

Για το προσαρμοσμένο μοντέλο, οι τρεις δισδιάστατες εκτιμώμενες ποσότητες $E_{ij.}, E_{.jk}$ και $E_{i.k}$ είναι πολύ κοντά στις αντίστοιχες παρατηρούμενες ποσότητες $n_{ij.}, n_{.jk}$ και $n_{i.k}$. Το πλήθος των βαθμών ελευθερίας της δοκιμασίας X^2, για το παρόν μοντέλο, είναι της τάξεως του $(r-1)(c-1)(k-1)$. Το μοντέλο μη ύπαρξης αλληλεπιδράσεων τριών κατευθύνσεων υποδηλώνει ότι οποιαδήποτε αλληλεπίδραση μεταξύ δύο εκ των τριών μεταβλητών, δεν εξαρτάται από την τρίτη μεταβλητή.

3.3 Κορεσμένο Μοντέλο

Όπως και στην περίπτωση των δισδιάσταστων πινάκων συνάφειας, το πιο γενικό μοντέλο για την περίπτωση των πινάκων τριών ή περισσοτέρων διαστάσεων είναι το κορεσμένο μοντέλο που προσαρμόζεται «πιστά» στα δεδομένα του πίνακα, χωρίς την απαίτηση περαιτέρω περιορισμών. Το κορεσμένο μοντέλο για ένα πίνακα τριών διαστάσεων, περιλαμβάνει όρους αλληλεπίδρασης τριπλής κατεύθυνσης (three-way) και επιτρέπει στην αμφίδρομη (two-way) αλληλεπίδραση μεταξύ οποιουδήποτε ζεύγους μεταβλητών να ποικίλλει στο κάθε επίπεδο της τρίτης μεταβλητής. Αυτό το μοντέλο θα συζητηθεί περαιτέρω με την εισαγωγή του λογαριθμογραμμικού (loglinear) μοντέλου για τις περιπτώσεις πινάκων τριών και πάνω διαστάσεων.

3.4 Λογαριθμογραμμικό μοντέλο

Όταν πρόκειται για τη μελέτη των σχέσεων μεταξύ δύο ή περισσότερων κατηγορικών μεταβλητών, τα παραδοσιακά λογαριθμικά γραμμικά μοντέλα αποτελούν χρήσιμα εργαλεία. Οι πολυδιάστατοι πίνακες συχνοτήτων, χρησιμοποιούνται για τη δημιουργία αυτών των κατασκευών. Στο δείγμα, κάθε στήλη σε έναν τέτοιο πίνακα παρέχει τον αριθμό των περιπτώσεων στις οποίες συναντήθηκε ένας συγκεκριμένος συνδυασμός τιμών για τις μεταβλητές. Για κάθε κελί στον πολυδιάστατο πίνακα, υπάρχει μια πιθανότητα που σχετίζεται με αυτό, δηλαδή η πιθανότητα επιλογής μιας περίπτωσης με αυτόν τον συγκεκριμένο συνδυασμό τιμών στον πληθυσμό.

3.4.1. Επέκταση του λογαριθμογραμμικού μοντέλου στους πίνακες τριών διαστάσεων

Ξεκινάμε την εισαγωγή στα λογαριθμογραμμικά μοντέλα για τους πίνακες τριών κατευθύνσεων από τον επεκτεταμένο ορισμό των παραμέτρων μ που συναντήσαμε στην παράγραφο 2.8 για τους πίνακες συνάφειας δύο διαστάσεων. Το κορεσμένο μοντέλο για τον πίνακα τριών κατευθύνσεων δίνεται ως

$$lnF_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{13(ik)} + \mu_{23(jk)} + \mu_{123(ijk)}$$
$$i = 1,2,\dots,r\ ;\quad j = 1,2,\dots,c\ ; k = 1,2,\dots,l.$$

Οι παραπάνω ποσότητες που αποτελούν το δεξί μέρος του λογαριθμογραμμικού μοντέλου μπορούν να εκφραστούν συναρτήσει της ποσότητας F_{ijk} (Jobson 1992)

$$\mu = \frac{1}{rcl}\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l} lnF_{ijk},$$

$$\mu_{1(i)} = \frac{1}{cl}\sum_{j=1}^{c}\sum_{k=1}^{l} lnF_{ijk} - \mu,$$

$$\mu_{2(j)} = \frac{1}{rl}\sum_{i=1}^{r}\sum_{k=1}^{l} lnF_{ijk} - \mu,$$

$$\mu_{3(k)} = \frac{1}{rc}\sum_{i=1}^{r}\sum_{j=1}^{c} lnF_{ijk} - \mu,$$

$$\mu_{12(ij)} = \frac{1}{l}\sum_{k=1}^{l} lnF_{ijk} - \mu_{1(i)} - \mu_{2(j)} - \mu,$$

$$\mu_{13(ik)} = \frac{1}{c}\sum_{j=1}^{c} lnF_{ijk} - \mu_{1(i)} - \mu_{3(k)} - \mu,$$

$$\mu_{23(ij)} = \frac{1}{r}\sum_{i=1}^{r} lnF_{ijk} - \mu_{2(j)} - \mu_{3(k)} - \mu,$$

$$\mu_{123(ijk)} = lnF_{ijk} - \mu_{1(i)} - \mu_{2(j)} - \mu_{3(k)} - \mu_{12(ij)} - \mu_{13(ik)} - \mu_{23(ij)} - \mu.$$

Για τα αθροίσματα των παραπάνω ποσοτήτων, ισχύουν οι σχέσεις

$$\sum_{\iota=1}^{r} \mu_{1(i)} = \sum_{j=1}^{c} \mu_{2(j)} = \sum_{k=1}^{l} \mu_{3(k)} = 0,$$

$$\sum_{\iota=1}^{r}\sum_{j=1}^{c} \mu_{12(ij)} = \sum_{\iota=1}^{r}\sum_{k=1}^{l} \mu_{13(ik)} = \sum_{j=1}^{c}\sum_{k=1}^{l} \mu_{23(jk)} = 0,$$

$$\sum_{\iota=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l} \mu_{123(ijk)} = 0.$$

Σε σύγκριση με το κορεσμένο λογαριθμογραμμικό μοντέλο για πίνακες δύο διαστάσεων, το κορεσμένο μοντέλο για τριδιάστατους πίνακες περιέχει συνολικά τέσσερις όρους αλληλεπίδρασης. Τρεις από τους όρους αλληλεπίδρασης είναι αλληλεπιδράσεις δύο κατευθύνσεων (two-way), ενώ ο τελευταίος όρος είναι μια αλληλεπίδραση τριών κατευθύνσεων. Στην περίπτωση αμφίδρομης αλληλεπίδρασης, η αλληλεπίδραση είναι ανεξάρτητη από την τρίτη μεταβλητή. Ωστόσο, η αλληλεπίδραση τριών κατευθύνσεων ποικίλλει ανάλογα με τις κατηγορίες της τρίτης μεταβλητής. Όταν και αυτές οι δύο περιπτώσεις αλληλεπιδράσεων περιλαμβάνονται, η αμφίδρομη αλληλεπίδραση είναι ένας μέσος όρος των κατηγοριών της τρίτης μεταβλητής, ενώ η αλληλεπίδραση τριών κατευθύνσεων μετρά τις αποκλίσεις ή τα κατάλοιπα από τη μέση τιμή.

3.4.2. Ορισμός των παραμέτρων του μοντέλου συναρτήσει των συχνοτήτων εντός των κελιών

Οι παράμετροι μ είναι συναρτήσεις διαφόρων περιθωρίων συνόλων του πίνακα των λογαρίθμων των θεωρητικών συχνοτήτων lnF_{ijk}. Όπως και στην περίπτωση του δισδιάστατου πίνακα που χρησιμοποιήσαμε στην παράγραφο 2.8, οι παράμετροι μ, είναι επίσης συναρτήσεις των λογαρίθμων διαφόρων γεωμετρικών μέσων των συχνοτήτων. Οι εκφράσεις για τις παραμέτρους αυτές μπορούν συμπληρωματικά να γραφτούν ως

$$\mu = ln\tilde{F}_{...},$$

$$\mu_{1(i)} = ln\tilde{F}_{i..} - ln\tilde{F}_{...},$$

$$\mu_{2(j)} = ln\tilde{F}_{.j.} - ln\tilde{F}_{...},$$

$$\mu_{3(k)} = ln\tilde{F}_{..k} - ln\tilde{F}_{...},$$

$$\mu_{12(ij)} = ln\tilde{F}_{ij.} - ln\tilde{F}_{i..} - ln\tilde{F}_{.j.} + ln\tilde{F}_{...},$$

$$\mu_{13(ik)} = ln\tilde{F}_{i.k} - ln\tilde{F}_{i..} - ln\tilde{F}_{..k} + ln\tilde{F}_{...},$$

$$\mu_{23(jk)} = ln\tilde{F}_{.jk} - ln\tilde{F}_{.j.} - ln\tilde{F}_{..k} + ln\tilde{F}_{...},$$

$$\mu_{123(ijk)} = lnF_{ijk} - ln\tilde{F}_{ij.} - ln\tilde{F}_{i.k} - ln\tilde{F}_{.jk} + ln\tilde{F}_{i..} + ln\tilde{F}_{.j.}, + ln\tilde{F}_{..k} - ln\tilde{F}_{...},$$

όπου

$\tilde{F}_{...}$ είναι ο συνολικός γεωμετρικός μέσος όλων των συχνοτήτων F_{ijk},

$\tilde{F}_{i..}$ είναι ο γεωμετρικός μέσος των συχνοτήτων F_{ijk}, για σταθερό i,

$\tilde{F}_{.j.}$ είναι ο γεωμετρικός μέσος των συχνοτήτων F_{ijk}, για σταθερό j,

$\tilde{F}_{..k}$ είναι ο γεωμετρικός μέσος των συχνοτήτων F_{ijk}, για σταθερό k,

$\tilde{F}_{i..}$ είναι ο γεωμετρικός μέσος των συχνοτήτων F_{ijk}, για σταθερό i,

$\tilde{F}_{ij.}$ είναι ο γεωμετρικός μέσος των συχνοτήτων F_{ijk}, για σταθερά i και j,

$\tilde{F}_{.jk}$ είναι ο γεωμετρικός μέσος των συχνοτήτων F_{ijk}, για σταθερό j και k,

$\tilde{F}_{i.k}$ είναι ο γεωμετρικός μέσος των συχνοτήτων F_{ijk}, για σταθερό i και k.

Ανάλογα με τις ιδιότητες που χαρακτηρίζουν τις συχνότητες F_{ijk} αλλάζει η αντίστοιχη μορφή του λογαριθμογραμμικού μοντέλου. Αυτό συμβαίνει καθώς κάποιες από τις ποσότητες του μοντέλου μηδενίζονται. Η παραπάνω αναπαράσταση αποτελεί το κορεσμένο μοντέλο, που προσαρμόζεται πιστά στα δεδομένα του τριδιάστατου πίνακα συνάφειας.

Α) Μοντέλο ανεξαρτησίας

Σε αυτή την περίπτωση, οι συχνότητες των κελιών γράφονται ως γινόμενο τριών περιθωρίων συχνοτήτων

$$F_{ijk} = \frac{F_{i..}F_{.j.}F_{..k}}{n^2}.$$

Αυτή η θεώρηση οδηγεί στο λογαριθμογραμμικό μοντέλο της μορφής

$$lnF_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)}$$

καθώς όλοι οι υπόλοιπες παράμετροι του μοντέλου μηδενίζονται.

Β) Μοντέλο μερικής ανεξαρτησίας

Στην περίπτωση του μοντέλου μερικής ανεξαρτησίας μπορούμε να καταλήξουμε σε συνολικά τρεις διαφορετικές μορφές μοντέλου, ανάλογα με το ποιες από τις τρεις εξεταζόμενες μεταβλητές παρουσιάζουν αλληλεπίδραση. Αν θεωρήσουμε την ύπαρξη αλληλεπίδρασης δύο κατευθύνσεων μεταξύ των μεταβλητών με δείκτες i και j, η παράμετρος $\mu_{12(ij)}$ είναι μη μηδενική. Όλες οι υπόλοιπες αλληλεπιδράσεις μηδενίζονται. Έτσι λοιπόν, το λογαριθμογραμμικό μοντέλο θα πάρει τη μορφή

$$lnF_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)}.$$

Με ανάλογη λογική μπορούμε να εκφράσουμε την παραπάνω σχέση, όταν υπάρχει αλληλεπίδραση μεταξύ των μεταβλητών με δείκτες i, k και j, k.

Γ) Υπό συνθήκη ανεξαρτησία

Στα μοντέλα υπό συνθήκης ανεξαρτησίας, η σχέση μεταξύ i, k αναπαρίσταται από την ποσότητα $\mu_{13(ik)}$, ενώ η αλληλεπίδραση μεταξύ i, j από την ποσότητα $\mu_{12(ij)}$. Εφόσον τα i, k είναι ανεξάρτητα, η ποσότητα $\mu_{23(jk)}$ ισούται με το 0. Το μοντέλο μας θα δίνεται από τη σχέση

$$lnF_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{13(ik)}.$$

Αναλόγως με το ποια από τις τρεις αλληλεπιδράσεις μπορεί να θεωρηθεί μηδενική, αλλάζει και η μορφή του παραπάνω γραμμικού μοντέλου.

Δ) Μη ύπαρξη αλληλεπίδρασης τριών κατευθύνσεων

Σε αυτή την περίπτωση, η μοναδική αλληλεπίδραση που μπορεί να θεωρηθεί αμελητέα, είναι αυτή μεταξύ και των τριών μεταβλητών. Έτσι, μόνο ο όρος $\mu_{123(ijk)}$ είναι μηδενικός. Το λογαριθμογραμμικό μοντέλο δίνεται από τον τύπο

$$lnF_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{13(ik)} + \mu_{23(jk)}.$$

3.5. Πίνακες συνάφειας τεσσάρων διαστάσεων

Για έναν τετραδιάστατο πίνακα συνάφειας, το λογαριθμογραμμικό μοντέλο δεδομένης της πλήρους ανεξαρτησίας μεταξύ των 4 κατηγορικών μεταβλητών, δίνεται από τη σχέση

$$lnF_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{4(h)}.$$

Η κορεσμένη μορφή του μοντέλου δίνεται από τον τύπο

$$lnF_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{4(h)} + \mu_{12(ij)} + \mu_{13(ik)} + \mu_{14(ih)} + \mu_{23(jk)} + \mu_{24(jh)} + \mu_{34(kh)} + \mu_{123(ijk)} + \mu_{234(jkh)} + \mu_{124(ijh)} + \mu_{134(ikh)} + \mu_{1234(ijkh)}.$$

Ανάμεσα σε αυτά τα δύο μοντέλα υπάρχουν πάνω από 100 άλλα πιθανά μοντέλα ανάλογα με τη σχέσεις εξάρτησης των τεσσάρων μεταβλητών. Η προσαρμογή όλων των δυνατών μοντέλων για τον προσδιορισμό του απλούστερου μοντέλου που περιγράφει επαρκώς τα δεδομένα, μπορεί να αποδειχθεί υπολογιστικά ακριβή. Για πλήθος διαστάσεων μεγαλύτερο του τέσσερα, ο αριθμός των πιθανών μοντέλων αυξάνεται ραγδαία. Για πίνακες τεσσάρων διαστάσεων ή και μεγαλύτερων

διαστάσεων επομένως, θα πρέπει να στραφούμε σε επαναληπτικές διαδικασίες αναζήτησης κατάλληλου μοντέλου.

3.6 Επιλογή Μοντέλου

Η επιλογή του μοντέλου ξεκινά με μια σειρά εξέτασης εναλλακτικών μοντέλων, όπου στη συνέχεια επιλέγεται αυτό που ταιριάζει καλύτερα στα δεδομένα. Η επιλογή του μοντέλου μπορεί να είναι ιδιαίτερα περίπλοκη. Ένα σημαντικό πλεονέκτημα αυτής της μεθόδου έναντι άλλων είναι ότι δεν υποθέτει απλώς το μοντέλο που ταιριάζει καλύτερα στα δεδομένα, αλλά αντιθέτως θεωρεί αυτό το στοιχείο της διαδικασίας μοντελοποίησης ως βασικό συστατικό της συνολικής έρευνας.

Συγκεκριμένα, οι περισσότερες στατιστικές αναλύσεις φτάνουν μέχρι το σημείο όπου υποθέτουν ένα γραμμικό μοντέλο για τα δεδομένα και στη συνέχεια απλώς προσαρμόζουν αυτό το μεμονωμένο μοντέλο στα δεδομένα, δοκιμάζοντας συνήθως κάποια μορφή μηδενικής υπόθεσης με στόχο να προκύψουν χρήσιμα συμπεράσματα. Τα συγκρίσιμα μοντέλα μπορεί να είναι εξαιρετικά από αρχιτεκτονικής άποψης και διαφορετικά μεταξύ τους, έχοντας ποικίλο αριθμό παραμέτρων, εμφανίζοντας διαφορετικά χαρακτηριστικά.

Για παράδειγμα, ένα μοντέλο μπορεί να είναι γραμμικό, ενώ ένα άλλο θα είναι λογαριθμικό και ένα άλλο μπορεί να είναι εκθετικό. Τα μοντέλα μπορεί να είναι από συμβατικά έως πολύ περίπλοκα και μπορούν να πραγματοποιήσουν ένα μεγάλο εύρος από διαφορετικές υποθέσεις διανομής σχετικά με τα δεδομένα στα οποία βασίζονται.

3.6.1 AIC

Όταν η ποιότητα ενός μοντέλου αξιολογείται αποκλειστικά ως προς την προσαρμογή του στα δεδομένα, οδηγούμαστε στην επιλογή του μοντέλο με τις περισσότερες επεξηγηματικές (ανεξάρτητες) μεταβλητές. Ωστόσο, μια τέτοια φιλοσοφία επιλογής μοντέλου, μπορεί να οδηγήσει συχνά σε υπερπροσαρμογή που οφείλεται στον υπερβολικό αριθμό παραμέτρων. Το κριτήριο πληροφορίας του Akaike (AIC) εμπεριέχει έναν επιπλέον όρο με στόχο να ξεπεραστεί το πρόβλημα της υπερπροσαρμογής (overfitting). Λαμβάνει υπόψη την εκτίμηση της παραγόμενης πιθανότητας, καθώς και τον αριθμό των ανεξάρτητων μεταβλητών που συμμετέχουν στην παλινδρόμηση. Η εξίσωση που εκφράζει το παραπάνω κριτήριο είναι η

$$AIC = 2k - 2ln(\hat{L})$$

Το $\hat{L}$ είναι η μέγιστη τιμή της εκτιμώμενης πιθανότητας του μοντέλου και το k αντιστοιχεί στον αριθμό των παραμέτρων του μοντέλου. Αυτή η «βαθμολογία» αντιπροσωπεύει μια εκτίμηση της προσαρμογής του μοντέλου (όπως υπολογίζεται από το log-likelihood), λαμβάνοντας επίσης

υπόψη τον αριθμό των μεταβλητών που περιλαμβάνονται στην εκτίμηση του μοντέλου (Stoica 2004).

Λόγω του ότι αποτελεί ένα δείκτη σχετικής μέτρησης, η χρησιμότητα του μπορεί να γίνει εμφανής μόνο όταν συγκρίνονται δύο ή περισσότερα μοντέλα. Μικρότερες τιμές του κριτηρίου AIC, δηλώνουν ιδανικότερα μοντέλα. Αυτή η μέτρηση είναι επομένως χρήσιμη μόνο σε περιπτώσεις όπου είναι απαραίτητο να συγκριθούν δύο ή περισσότερα μοντέλα. Όσο αυξάνεται ο αριθμός των παραμέτρων, αυξάνεται η τιμή του κριτηρίου, ενώ μεγαλύτερες τιμές πιθανοφάνειας οδηγούν σε μείωση αυτού. Έτσι, αντιλαμβανόμαστε ότι το κριτήριο αυτό δίνει αντιπροσωπευτικές τιμές όσον αφορά την αποδοτική επιλογή μοντέλου, καθώς λαμβάνει υπόψη τόσο την προσαρμοστική ικανότητα του μοντέλου, όσο και το φαινόμενο της υπερπροσαρμογής.

3.6.2. BIC

Ένα σημαντικό εργαλείο της Στατιστικής είναι το μπεϋζιανό κριτήριο πληροφορίας BIC (Bayesian Information Criterion) το οποίο παρουσιάστηκε από τον Akaike και τον Schwarz (1977-1978) ως βελτίωση του κριτηρίου AIC. Το BIC όπως και το AIC χρησιμοποιείται προκειμένου να επιλεχθεί το κατάλληλο μοντέλο ανάμεσα από ένα πλήθος υποψήφιων μοντέλων και το μοντέλο με τη μικρότερη τιμή του κριτηρίου θεωρείται το καλύτερο, εκείνο δηλαδή που προσαρμόζεται καλύτερα στα δεδομένα. Η γενική μορφή του BIC διαφέρει με αυτή του AIC ως προς τον όρο ποινικοποίησης. Η εξίσωση του κριτηρίου BIC, δίνεται από τη σχέση (Findley 1991)

$$BIC = klog(n) - 2ln(\hat{L}),$$

Με $\hat{L}$ συμβολίζουμε την εκτιμώμενη πιθανοφάνεια, με k το πλήθος των επεξηγηματικών μεταβλητών του μοντέλου και με n το πλήθος των στοιχείων του συνόλου δεδομένων που χρησιμοποιούνται για την προσαρμογή του εκάστοτε μοντέλου (π.χ. μοντέλου λογιστικής παλινδρόμησης). Σε σχέση με το κριτήριο AIC, το κριτήριο BIC είναι αυστηρότερο. Δηλαδή, το κριτήριο BIC προτείνει την επιλογή μοντέλων με ίσο ή μικρότερο αριθμό παραμέτρων σε σχέση με το AIC, για μεγάλα σύνολα δεδομένων καθώς $\log(n) > 2$ για $n > 100$.

3.6.3 Σταδιακή επιλογή μοντέλου

Οι στρατηγικές επιλογής προς τα εμπρός και προς τα πίσω χρησιμοποιούνται συνήθως σε συνδυασμό για τη προσαρμογή γραμμικών μοντέλων παλινδρόμησης. Η σταδιακή παλινδρόμηση εξετάζει όλες τις πιθανές μεταβλητές στο μοντέλο μετά από κάθε βήμα στο οποίο περιλαμβάνεται μια νέα μεταβλητή, ελέγχοντας εάν η συνεισφορά των ανεξαρτήτων μεταβλητών έχει μειωθεί κάτω από ένα προκαθορισμένο όριο. Το όριο αυτό είναι καθορισμένο από το χρήστη και τις

ανάγκες του εξεταζόμενου φαινομένου. Είναι απαραίτητο να διαγραφεί μια μεταβλητή από το μοντέλο εάν διαπιστωθεί ότι δεν είναι στατιστικά σημαντική. Κατά τη χρήση της επαναληπτικής αξιολόγησης και επιλογής μοντέλου, υπάρχουν δύο επίπεδα σημαντικότητας που πρέπει να ληφθούν υπόψη. Το ένα αφορά την εισαγωγή μεταβλητών και το άλλο είναι για την εξάλειψη μεταβλητών από την εξίσωση του μοντέλου.

Όταν η στατιστική σημαντικότητα μιας μεταβλητής έχει τιμή (p-value) που είναι μικρότερη από το επιλεγμένο από το χρήστη όριο της στάθμης σημαντικότητας α, η προσθήκη αυτής θεωρείται απαραίτητη και ευεργετική για το προσαρμοσμένο μοντέλο. Αντίθετα, όταν η στατιστική σημαντικότητα μιας μεταβλητής έχει τιμή (p-value) μεγαλύτερη από την επιλεγμένη από το χρήστη στάθμη σημαντικότητας α, η προσθήκη αυτής θεωρείται μη αναγκαία, καθώς δεν προσφέρει βελτίωση της προβλεπτικής ή προσαρμοστικής ικανότητας του προτεινόμενου μοντέλου στα δεδομένα. Ως στάθμη στατιστικής σημαντικότητας, επιλέγεται συνήθως η τιμή $\alpha = 0.05$, χωρίς όμως να απορρίπτονται άλλες (κατά κύριο λόγο μικρότερες) πιθανές τιμές.

Η σταδιακή προσέγγιση για την επιλογή ενός μοντέλου ξεκινά με ένα συγκεκριμένο μοντέλο, στο οποίο είτε προσθέτουμε όρους (προς τα εμπρός) είτε διαγράφουμε όρους (προς τα πίσω). Οι όροι προστίθενται/αφαιρούνται κατά ένας τη φορά μέχρι να ληφθεί το απλούστερο μοντέλο που ταιριάζει στα δεδομένα. Μιας και η σταδιακή αυτή διαδικασία ξεκινά με ένα συγκεκριμένο μοντέλο, αυτό το μοντέλο πρέπει να επιλεγεί με κατάλληλο τρόπο (Jobson 1992, p. 70). Μια κοινή προσέγγιση για την επιλογή ενός σημείου εκκίνησης της διαδικασίας, είναι η προσαρμογή όλων των μοντέλων ίδιας τάξης. Οι ομάδες μοντέλων ίδιας τάξης για ένα πίνακα συνάφειας τεσσάρων διαστάσεων δίνονται ως:

1) Τάξη 1: [1] [2] [3] [4],
2) Τάξη 2: [12] [13] [14] [23] [24] [34],
3) Τάξη 3: [123] [134] [234] [124],
4) Τάξη 4: [1234].

Οι αριθμοί εντός των αγκύλων, δηλώνουν το ποιες από τις 4 μεταβλητές του πίνακα τεσσάρων διαστάσεων συμμετέχουν στη διαμόρφωση του μοντέλου.

Το απλούστερο μοντέλο ίδιας τάξης που ταιριάζει περισσότερο στα δεδομένα του πίνακα θα αποτελεί ένα άνω όριο. Συνήθως, εάν ένα συγκεκριμένο μοντέλο ίδιας τάξης ταιριάζει επαρκώς στα δεδομένα, όλα τα μοντέλα υψηλότερης τάξης θα ταιριάζουν επίσης στα δεδομένα του πίνακα. Από την άλλη πλευρά, ένα κατώτερο όριο της σταδιακής διαδικασίας θα ήταν το μοντέλο υψηλότερης τάξης που δεν περιγράφει ικανοποιητικά τα δεδομένα του πίνακα. Συνήθως τα άνω και κάτω όρια διαφέρουν μόνο κατά μία ή δύο τάξεις.

Για να περιγράψουμε τις διαδικασίες, συμβολίζουμε με q την τάξη του μοντέλου που αποτελεί το άνω όριο και με r την τάξη του αντίστοιχου κάτω ορίου. Η προς τα εμπρός (forward stepwise procedure) διαδικασία ξεκινά με το μοντέλο τάξης r (κάτω όριο) και προσθέτει όρους, έναν κάθε φορά, με τέτοιο τρόπο ώστε η μεταβολή της πιθανοφάνειας του μοντέλου να

μεγιστοποιηθεί. Προστίθενται όροι μόνο εφόσον η αύξηση της πιθανοφάνειας είναι σημαντική. Για την αντίστροφη διαδικασία (backward stepwise procedure), ξεκινάμε από ένα μοντέλο τάξης q (άνω όριο). Σε κάθε βήμα, ο όρος που αφαιρείται είναι αυτός που παράγει τη μικρότερη μεταβολή στην πιθανοφάνεια του μοντέλου. Οι όροι καταργούνται σε περίπτωση που δε μπορούν στατιστικά σημαντικοί. Όπως αναφέραμε και παραπάνω, ο έλεγχος αυτός βασίζεται στις παραγόμενες τιμές στατιστικής σημαντικότητας (p-value) της εκάστοτε μεταβλητής. Οι μεταβλητές αυτές παρέχονται από τα περισσότερα στατιστικά πακέτα λογισμικών ανοικτού κώδικα όπως η R. Η δυνατότητα αυτή διευκολύνει δραματικά τη διαδικασία επιλογής μοντέλου, μειώνοντας παράλληλα τον απαιτούμενο χρόνο διεξαγωγής της μεθόδου.

Κεφάλαιο 4ο. Γενικευμένα Γραμμικά Μοντέλα

Τα γενικευμένα γραμμικά μοντέλα (generalized linear models) διευρύνουν τη θεωρία των κλασικών μοντέλων γραμμικής παλινδρόμησης με στόχο να συμπεριλάβουν στη μεθοδολογία, περιπτώσεις όπου η μεταβλητή απόκρισης Y δεν ακολουθεί κανονική κατανομή. Αποτελούνται από τρία χαρακτηριστικά:

1) Τον τυχαίο παράγοντα (random component), που αφορά τη μεταβλητή απόκρισης Y και την κατανομή πιθανότητας αυτής.

2) Τον συστηματικό παράγοντα (systematic component), που προσδιορίζει τις επεξηγηματικές μεταβλητές X_i που θα συμπεριληφθούν σε μια γραμμική συνάρτηση πρόβλεψης.

3) Τη συνδετική συνάρτηση (link function) που συνδέει τη συνάρτηση $E(Y)$ που θα πρέπει να προβλεφθεί, με τους γραμμικούς όρους.

Οι Nelder and Wedderburn (1972) παρουσίασαν την κλάση των γενικευμένων γραμμικών μοντέλων (GLM), ωστόσο τα σημαντικότερα μοντέλα της κλάσης αυτής είχαν καθιερωθεί πριν το 1972.

Ο τυχαίος παράγοντας ενός γενικευμένου γραμμικού μοντέλου αποτελείται από τη μεταβλητή απόκρισης Y, η οποία χαρακτηρίζεται από ανεξάρτητες παρατηρήσεις $(y_1, \dots, y_N)$ από μια κατανομή που ανήκει στην οικογένεια των εκθετικών κατανομών. Το N αποτελεί το πλήθος των αντικειμένων που περιλαμβάνονται στο σύνολο δεδομένων. Στη συγκεκριμένη οικογένεια κατανομών, οι συναρτήσεις πυκνότητας πιθανότητας έχουν τη μορφή

$$f(y_i;\ \theta_i) = a(\theta_i)b(y_i)\exp[\ y_i\ Q(\theta_i)\ ].$$

Στην οικογένεια αυτή, ανήκουν πολλές από τις ευρέως γνωστές κατανομές, όπως η Διωνυμική, η Poisson, οι κατανομές Γάμμα και Βήτα κ.α. Η τιμή της παραμέτρου θ_i ποικίλει για $i = 1, \dots, n$ ως συνάρτηση των επεξηγηματικών μεταβλητών X_i του μοντέλου. Η παράμετρος $Q(\theta)$ καλείται *φυσική παράμετρος* (natural parameter). Στην παράγραφο Χ παρουσιάζουμε μια πιο γενική μορφή

για την συνάρτηση f, η οποία επιτρέπει την συμπερίληψη παραμέτρων που αντιστοιχούν σε μέτρα διασποράς.

Ο συστηματικός παράγοντας ενός GLM, αντιστοιχεί ένα διάνυσμα παραμέτρων $(\eta_1, \dots \eta_N)$ στο σύνολο των επεξηγηματικών μεταβλητών, μέσω μιας γραμμικής συνάρτησης. Έστω ότι το x_{ij} αντιστοιχεί την τιμή της επεξηγηματικής μεταβλητής j για το αντικείμενο i του συνόλου δεδομένων. Τότε

$$\eta_i = \sum_j \beta_j x_{ij} , \quad i = 1, \dots, N .$$

Αυτός ο γραμμικός συνδυασμός επεξηγηματικών μεταβλητών ονομάζεται *γραμμικός εκτιμητής*. Συνήθως, θεωρούμε ότι

$$x_{i0} = 1 , \qquad i = 1, \dots, N$$

που αποτελεί τον σταθερό όρο που αντιστοιχεί στην παράμετρο β_0 του μοντέλου.

Ο τρίτος παράγοντας ενός GLM (συνδετική συνάρτηση), στοχεύει στο να ενώσει τις επιδράσεις του τυχαίου και του συστηματικού παράγοντα. Έστω

$$\mu_i = E(Y_i), \qquad i = 1, \dots, N,$$

τότε το μοντέλο ενώνει τη μέση τιμή μ_i με τις μεταβλητές η_i ως $\eta_i = g(\mu_i)$, όπου η g είναι μια μονότονη και παραγωγίσιμη συνάρτηση. Τότε,

$$g(\mu_i) = \sum_j \beta_j x_{ij} , \quad i = 1, \dots, N.$$

Η συνδετική συνάρτηση $g(\mu) = \mu$, καλείται ταυτοτικός σύνδεσμος, όπου $\eta_i = \mu_i$. Η συγκεκριμένη συνάρτηση αντιστοιχεί στην περίπτωση της συνήθους γραμμικής παλινδρόμησης όπου η μεταβλητή απόκρισης Y ακολουθεί κανονική κατανομή. Η συνάρτηση που μετασχηματίζει τη μέση τιμή, στην φυσική παράμετρο, καλείται κανονικός σύνδεσμος (canonical link). Σε αυτή την περίπτωση,

$$g(\mu_i) = Q(\theta_i), \quad i = 1, \dots, N$$

και

$$Q(\theta_i) = \sum_j \beta_j x_{ij} .$$

Στη συνέχεια, παρουσιάζουμε τα δύο σημαντικότερα γενικευμένα γραμμικά μοντέλα που κάνουν χρήση κατηγορικών μεταβλητών.

4.1. Διωνυμικά μοντέλα Logit για δυαδικά δεδομένα

Πολλές μεταβλητές απόκρισης είναι δυαδικές. Συμβολίζουμε με 1 και 0, τις περιπτώσεις όπου η μεταβλητή απόκρισης αντιστοιχεί σε «επιτυχία» ή «αποτυχία», όπου

$$P(Y = 1) = p \quad \text{και} \quad P(Y = 0) = 1 - p\,,$$

Καταλήγοντας στο ότι $E(Y) = p$. Η περίπτωση αυτή αντιστοιχεί στη διωνυμική κατανομή με $n = 1$. Η αντίστοιχη συνάρτηση πιθανότητας μπορεί να γραφεί ως

$$f(y\,; p) = p^{y}\,(1-p)^{1-y} = (1-p)\left(\frac{p}{1-p}\right)^{y}$$
$$= (1-p)\exp[\,y\,(\log\left(\frac{p}{1-p}\right)\,],$$

όπου $y = 0$ ή 1. Λαμβάνοντας υπόψιν τη γενική μορφή των συναρτήσεων της οικογένειας εκθετικών κατανομών $f(y_i;\ \theta_i) = a(\theta_i)b(y_i)\exp[\,y_i\ Q(\theta_i)\,]$, και αντικαθιστώντας όπου θ την παράμετρο p, έχουμε

$$\alpha(p) = 1 - p,\ b(y) = \ 1 \ \text{και}\ Q(p) = \ \log\left(\tfrac{p}{1-p}\right)\,.$$

Έτσι, μπορούμε να πούμε ότι η φυσική παράμετρος $\log\left(\frac{p}{1-p}\right)$ είναι η λογαριθμική πιθανότητα του αποτελέσματος 1 (επιτυχία) ή αλλιώς το logit του p. Στην περίπτωση όπου κάνουμε χρήση της κανονικής συνδετικής συνάρτησης για την περίπτωση της διωνυμικής κατανομής, καταλήγουμε στα μοντέλα λογιστικής παλινδρόμησης που θα δούμε στη συνέχεια.

4.2. Poisson λογαριθμογραμμικά μοντέλα για διακριτά δεδομένα

Ορισμένες μεταβλητές απόκρισης έχουν μετρήσεις ως πιθανά αποτελέσματά τους. Σε μια έρευνα υγείας, η παρατήρηση μπορεί να είναι ο αριθμός των ασθενειών στο παρελθόν για τις οποίες το άτομο επισκέφτηκε ένας γιατρός. Οι μετρήσεις εμφανίζονται επίσης ως καταχωρήσεις σε πίνακες συνάφειας (contingency tables).

Η απλούστερη περίπτωση κατανομής τέτοιων δεδομένων είναι η κατανομή Poisson. Η συνάρτηση πιθανότητας για τη μεταβλητή Y, μπορεί να γραφεί ως

$$f(y\,;\mu) = \frac{e^{-\mu}\mu^{y}}{y!} = \exp(-\mu)\left(\frac{1}{y!}\right)\exp(y[log\mu]),\ \ y = 0, 1, 2, \dots$$

Όπως και στην προηγούμενη παράγραφο για τη διωνυμική κατανομή, έτσι και εδώ θέτουμε $\theta = \mu$ και έχουμε

$$a(\mu) = \exp(-\mu),\quad b(y) = \frac{1}{y!}\ \text{και}\quad Q(\mu) = \log(\mu).$$

Η φυσική παράμετρος θα είναι η $\log(\mu)$, οπότε στην περίπτωση της κανονικής συνδετικής συνάρτησης θα ισχύει $\eta = \log(\mu)$. Το γενικευμένο γραμμικό μοντέλο που κάνει χρήση αυτής της συνάρτησης, θα είναι της μορφής

$$\log(\mu_i) = \sum_j \beta_j x_{ij}\,,\quad i = 1, \dots, N.$$

Το μοντέλο αυτό, μπορεί να το συναντήσουμε στη βιβλιογραφία ως Poisson λογαριθμογραμμικό μοντέλο.

4.3. Ροπές και πιθανοφάνεια στην περίπτωση των γενικευμένων γραμμικών μοντέλων

Έχοντας παρουσιάσει τα GLM για δυαδικά δεδομένα και δεδομένα μέτρησης, στρέφουμε τώρα την προσοχή μας στις εξισώσεις πιθανοφάνειας και στις μεθόδους προσαρμογής των GLM. Το υπόλοιπο αυτού του κεφαλαίου παρέχει γενικά αποτελέσματα που εφαρμόζονται κατά τη διαδικασία εκτίμησης των παραμέτρων των προαναφερθέντων γραμμικών μοντέλων.

4.3.1. Εκθετική Οικογένεια Κατανομών

Σε αντίθεση με την προηγούμενη παράγραφο, θα ασχοληθούμε με περιπτώσεις κατανομών που χαρακτηρίζονται από περισσότερες από μια εκτιμώμενες παραμέτρους. Όπως και προηγουμένως, θεωρούμε πως ο τυχαίος παράγοντας ενός γενικευμένου γραμμικού μοντέλου, καθορίζει ότι οι N το πλήθος παρατηρήσεις του συνόλου δεδομένων $(y_1, \dots, y_N)$ που αντιστοιχούν στην μεταβλητή απόκρισης Y, είναι ανεξάρτητες. Επιπλέον, η συνάρτηση πιθανότητας ή η συνάρτηση πυκνότητας πιθανότητας των y_i είναι της μορφής

$$f(y_i\,;\theta_i,\varphi)=\exp\left(\frac{[y_i\theta_i-b(\theta_i)]}{a(\varphi)}+c(y_i,\varphi)\right).$$

Οι συναρτήσεις κατανομής αυτής της μορφής, ανήκουν στην εκθετική οικογένεια κατανομών διασποράς, ενώ η παράμετρος φ ονομάζεται παράμετρος διασποράς. Από την άλλη πλευρά, οι παράμετροι θ_i ονομάζονται φυσικές παράμετροι.

Στην περίπτωση που η παράμετρος φ δεν είναι απαραίτητη, όπως σε περίπτωση κατανομής με μια παράμετρο, η παραπάνω συνάρτηση απλοποιείται στη μορφή

$$f(y_i;\ \theta_i)=a(\theta_i)b(y_i)\exp[\ y_i\ Q(\theta_i)\ ].$$

Προφανώς, η γενικευμένη συνάρτηση $f(y_i\,;\theta_i,\varphi)$ δεν είναι αναγκαία στην περίπτωση μονοπαραμετρικών κατανομών όπως η διωνυμική και η Poisson που είδαμε παραπάνω. Συνήθως, η ποσότητα $\alpha(\varphi)$ παίρνει τη μορφή $\alpha(\varphi)=\frac{\varphi}{\omega_i}$ όπου το ω_i είναι γνωστό.

4.3.2. Συναρτήσεις μέσης τιμής και διασποράς για τον τυχαίο παράγοντα

Οι συναρτήσεις $E(Y_i)$ και $Var(Y_i)$ μπορούν να γραφούν συναρτήσει των ποσοτήτων της συνάρτησης $f(y_i\,;\theta_i,\varphi)$. Έστω ότι με

$$L_i=\log f(y_i\,;\theta_i,\varphi)$$

συμβολίζουμε τη συνεισφορά της μεταβλητής y_i στην λογαριθμική πιθανοφάνεια, οπότε η συνολική λογαριθμική πιθανοφάνεια είναι της μορφής $L=\sum_j L_i$. Τότε,

$$L_i=\frac{[y_i\theta_i-b(\theta_i)]}{a(\varphi)}+c(y_i,\varphi).$$

Παραγωγίζοντας ως προς τις φυσικές παραμέτρους, παίρνουμε

$$\frac{\partial L_i}{\partial\theta_i}=\frac{[y_i-b'(\theta_i)]}{a(\varphi)}$$

και

$$\frac{\partial^2 L_i}{\partial\theta_i^2}=\frac{[-b''(\theta_i)]}{a(\varphi)}$$

όπου τα $b'(\theta_i)$ και $-b''(\theta_i)$ συμβολίζουν τις πρώτες δύο παραμέτρους της συνάρτησης $b(\cdot)$ υπολογισμένη στην παράμετρο θ_i . Με βάση τις ιδιότητες

$$E\left(\frac{\partial L}{\partial \theta}\right) = 0$$

και

$$E\left(\frac{\partial L}{\partial \theta}\right)^2 = -E(\frac{\partial^2 L}{\partial \theta^2})$$

που ισχύουν για τις εκθετικές οικογένειες κατανομών (Cox and Hinkley 1974), παίρνουμε ότι

$$E\left(\frac{\partial L}{\partial \theta}\right) = \frac{E[Y_i - b'(\theta_i)]}{a(\varphi)} = 0$$

ή

$$\mu_i = E(Y_i) = b'(\theta_i).$$

Παράλληλα, εφόσον

$$E\left(\frac{\partial^2 L}{\partial \theta^2}\right) = -\frac{b''(\theta_i)}{\alpha(\varphi)}$$

παίρνουμε

$$\frac{b''(\theta_i)}{\alpha(\varphi)} = \left(\frac{E[Y_i - b'(\theta_i)]}{a(\varphi)}\right)^2 = \frac{var(Y_i)}{[a(\varphi)]^2}.$$

Έτσι λοιπόν, μπορούμε να εκφράσουμε τη συνάρτηση διασποράς της μεταβλητής απόκρισης Y_i συναρτήσει των ποσοτήτων της συνάρτησης $f(y_i\,;\theta_i, \varphi)$, ως

$$var(Y_i) = b''(\theta_i)a(\varphi).$$

Συνοψίζοντας, η συνάρτηση $b(\cdot)$ καθορίζει τις ροπές της μεταβλητής Y_i.

4.3.3. Συναρτήσεις μέσης τιμής και διασποράς για διωνυμικά και Poisson γενικευμένα γραμμικά μοντέλα

Έστω ότι η μεταβλητή Y_i ακολουθεί την κατανομή Poisson. Τότε

$$f(y_i\,;\mu_i) = \frac{e^{-\mu_i}\,\mu_i^{y_i}}{y_i!} = \exp(y_i \log \mu_i - \mu_i - \log y_i!)$$

$$= \exp[y_i\,\theta_i - \exp(\theta_i) - \log y_i!]$$

όπου $\theta_i = log\ \mu_i$. Η συνάρτηση αυτή, ανήκει στην εκθετική οικογένεια κατανομών διασποράς με

$$b(\theta_i) = \exp(\theta_i)\,,\ a(\varphi) = 1 \ \text{ και } c(y_i\,,\varphi) = -\log y_i!\,.$$

Η φυσική παράμετρος θα είναι η $\theta_i = log\ \mu_i$, ενώ από τις εξισώσεις που προέκυψαν στην παράγραφο Χ, για τη μέση τιμή και τη διασπορά, θα έχουμε

$$E(Y_i) = b'(\theta_i) = \exp(\theta_i) = \mu_i\,,$$

$$Var(Y_i) = b''(\theta_i) = \exp(\theta_i) = \mu_i\,.$$

Στη συνέχεια, θεωρούμε ότι για την ποσότητα $n_i Y_i$ ισχύει

$$n_i Y_i \ \sim\ B\ (n_i, p_i),$$

δηλαδή ακολουθεί τη διωνυμική κατανομή με παραμέτρους n_i και p_i. Στην περίπτωσή μας, το y_i δηλώνει το ποσοστό των επιτυχιών, οπότε $E(Y_i) = p_i$ που είναι ανεξάρτητο του n_i. ;Εστω ότι $\theta_i = \log\left(\frac{p_i}{1-p_i}\right)$, άρα

$$p_i = \frac{\exp(\theta_i)}{[1 + \exp(\theta_i)]}$$

και

$$\log(1 - p_i) = -\log[1 + \exp(\theta_i)].$$

Τότε, για τη συνάρτηση της εκθετικής οικογένειας κατανομών διασποράς, έχουμε

$$f(y_i; p_i, n_i) = \binom{n_i}{n_i y_i} p_i^{n_i y_i}(1 - p_i)^{n_i - n_i y_i}$$

$$= \exp[\frac{y_i\theta_i - \log[1+\exp(\theta_i)]}{\frac{1}{n_i}} + log\binom{n_i}{n_i y_i}$$

Αντιλαμβανόμαστε, ότι η παραπάνω συνάρτηση ανήκει στην εκθετική οικογένεια κατανομών διασποράς με

$$b(\theta_i) = \log[1+\exp(\theta_i)]\,, a(\varphi) = 1/n_i \text{ και } c(y_i\,, \varphi) = log\binom{n_i}{n_i y_i}\,.$$

Η φυσική παράμετρος είναι η $\theta_i = \log\left(\frac{p_i}{1-p_i}\right)$, ενώ για τις συναρτήσεις μέσης τιμής και διασποράς θα ισχύει

$$E(Y_i) = b'(\theta_i) = \exp(\theta_i)/[1+\exp(\theta_i)] = p_i\,,$$

$$Var(Y_i) = b''(\theta_i)a(\varphi) = \exp(\theta_i)\,/\{[1+\exp(\theta_i)]^2 n_i\} = \frac{p_i(1-p_i)}{n_i}\,.$$

4.4. Εξισώσεις πιθανοφάνειας για γενικευμένα γραμμικά μοντέλα

Για N ανεξάρτητες παρατηρήσεις και σύμφωνα με τη σχέση $L_i = \frac{[y_i\theta_i - b(\theta_i)]}{a(\varphi)} + c(y_i, \varphi)$, η ολική συνάρτηση λογαριθμικής πιθανοφάνειας $L(\boldsymbol{\beta})$ είναι

$$L(\boldsymbol{\beta}) = \sum_i L_i = \sum_i \log f(y_i\,; \theta_i, \varphi) = \sum_i \frac{y_i\theta_i - b(\theta_i)}{a(\varphi)} + \sum_i c(y_i, \varphi)\,.$$

Στη συνέχεια, παραγωγίζουμε τη συνάρτηση πιθανοφάνειας $L(\boldsymbol{\beta})$ ως προς την παράμετρο β_j και εξισώνουμε με το 0, καταλήγοντας στη σχέση

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_i \frac{\partial L_i}{\partial \beta_j} = 0$$

για όλα τα j. Για την παραγώγιση της συνάρτησης $L(\boldsymbol{\beta})$ κάνουμε χρήση του κανόνα της αλυσίδας ως

$$\frac{\partial L_i}{\partial \beta_j} = \frac{\partial L_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}\,.$$

Εφόσον $\frac{\partial L_i}{\partial \theta_i} = [y_i - b'(\theta_i)]/a(\varphi)$, ενώ καθώς $\mu_i = b'(\theta_i)$ και $var(Y_i) = b''(\theta_i)a(\varphi)$,καταλήγουμε στο ότι

$$\frac{\partial L_i}{\partial \theta_i} = (y_i - \mu_i)/a(\varphi)$$

και

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = var(Y_i)/a(\varphi).$$

Επίσης, καθώς $\eta_i = \sum_j \beta_j x_{ij}$, έχουμε

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}.$$

Τελικά, καθώς $\eta_i = g(\mu_i)$, η ποσότητα $\frac{\partial \mu_i}{\partial \eta_i}$ θα εξαρτάται από τη συνδετική συνάρτηση του μοντέλου. Συνοψίζοντας, αντικαθιστώντας τις προαναφερθείσες ποσότητες, στη σχέση $\frac{\partial L_i}{\partial \beta_j} = \frac{\partial L_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}$, παίρνουμε

$$\frac{\partial L_i}{\partial \beta_j} = \frac{y_i - \mu_i}{a(\varphi)} \frac{\alpha(\varphi)}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} = \frac{(y_i - \mu_i)x_{ij}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}.$$

Αθροίζοντας τις παραπάνω λογαριθμικές πιθανοφάνειες για όλα τα N το πλήθος, μέλη του συνόλου δεδομένων, παίρνουμε

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0\ ,\ \ j = 0,1,2, \dots$$

Το διάνυσμα των παραμέτρων $\boldsymbol{\beta}$ του μοντέλου μπορεί να μην εμφανίζεται άμεσα στις παραπάνω εξισώσεις, ωστόσο εμπεριέχεται στην ποσότητα μ_i, καθώς

$$\mu_i = g^{-1}(\sum_j \beta_j x_{ij})\ .$$

Διαφορετικές συνδετικές συναρτήσεις $g(\cdot)$ οδηγούν σε διαφορετικά σύνολα εξισώσεων.

4.4.1. Εκτιμητές πιθανοφάνειας για τα διωνυμικά γενικευμένα γραμμικά μοντέλα

Υποθέτουμε ότι η ποσότητα $n_i Y_i$ ακολουθεί διωνυμική κατανομή $B(n_i, p_i)$. Θυμίζουμε ότι στη συγκεκριμένη περίπτωση, η ποσότητα y_i είναι το ποσοστό των επιτυχιών του δείγματος σε n_i το πλήθος δοκιμές. Το διωνυμικό γραμμικό μοντέλο με περισσότερες από δύο επεξηγηματικές μεταβλητές X_i μπορεί να γραφεί ως

$$p_i = \Phi(\sum_j \beta_j x_{ij}) \,,$$

όπου με Φ συμβολίζουμε την αθροιστική κατανομή κάποιας συνεχούς κατανομής πιθανοτήτων. Εφόσον $p_i = \mu_i = \Phi(\eta_i)$ και $\eta_i = \sum_j \beta_j x_{ij}$, έχουμε

$$\frac{\partial \mu_i}{\partial \eta_i} = \varphi(\eta_i) = \varphi(\sum_j \beta_j x_{ij}) \,,$$

όπου $\varphi(u) = \frac{d\Phi(u)}{du}$. Εφόσον παραπάνω έχουμε δείξει ότι $var(Y_i) = p_i(1-p_i)/n_i$, τότε η οι εξισώσεις πιθανοφάνειας μετά από αντικατάσταση, μπορούν να γραφούν ως

$$\sum_{i=1}^{N} \frac{n_i(y_i - p_i)x_{ij}}{p_i\,(1-p_i)}\ \varphi(\sum_j \beta_j x_{ij}) = 0 \,,\ \ j = 0,1,2,\dots$$

όπου $p_i = \Phi(\sum_j \beta_j x_{ij})$.

Κάνοντας χρήση της logit συνδετικής συνάρτησης, έχουμε $\eta_i = \log[\frac{p_i}{1-p_i}]$, οπότε

$$\frac{\partial \eta_i}{\partial p_i} = \frac{1}{[p_i(1-p_i)]}$$

και

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial p_i}{\partial \eta_i} = p_i(1-p_i).$$

Τότε, η λογαριθμική συνάρτηση πιθανοφάνειας μπορεί να απλοποιηθεί και να λάβει τη μορφή

$$\sum_{i=1}^{N} n_i(y_i - p_i)x_{ij} = 0\ ,\ \ j = 0,1,2,\dots$$

όπου $p_i = \Phi(\sum_j \beta_j x_{ij})$, όπου με Φ συμβολίζουμε την αθροιστική λογαριθμική κατανομή.

4.4.2. Εκτιμητές πιθανοφάνειας για τα Poisson γενικευμένα γραμμικά μοντέλα

Το γενικό Poisson λογαριθμογραμμικό μοντέλο μπορεί να γραφεί με τη μορφή πινάκων ως

$$\log \boldsymbol{\mu} = \boldsymbol{X}\,\boldsymbol{\beta}.$$

Κάνοντας χρήση της λογαριθμικής συνδετικής συνάρτησης παίρνουμε $\eta_i = \log \mu_i$, οπότε $\mu_i = \exp(\eta_i)$ και $\frac{\partial \mu_i}{\partial \eta_i} = \exp(\eta_i) = \mu_i$. Τότε, για $var(Y_i) = \mu_i$ (λόγω της ιδιότητας της Poisson κατανομής η οποία έχει κοινή διασπορά και μέση τιμή), οι εξισώσεις λογαριθμικής πιθανοφάνειας παίρνουν τη μορφή

$$\sum_{i=1}^{N} (y_i - \mu_i)x_{ij} = 0\ ,\quad j = 0, 1, 2,$$

Το παραπάνω σύνολο σχέσεων, εξισώνει το στατιστικό $\sum_j y_i x_{ij}$ του διανύσματος παραμέτρων $\boldsymbol{\beta}$ με την αντίστοιχη μέση τιμή του μοντέλου μ_i.

Κεφάλαιο 5ο . Μοντέλα λογιστικής παλινδρόμησης (logit)

Η λογιστική παλινδρόμηση είναι μία από τις κυριότερες μεθόδους μηχανικής μάθησης, η οποία προσπαθεί να εξηγήσει/προβλέψει την πιθανότητα μια ποιοτική εξαρτημένη μεταβλητή Y, να ανήκει σε μία από τις πιθανές κλάσεις του προβλήματος δεδομένης της πληροφορίας που παρέχεται από το σύνολο των ανεξάρτητων μεταβλητών $\boldsymbol{X} = (X_1, \dots, X_n)$. Βασίζεται στη χρήση λογαριθμικών συναρτήσεων, οι οποίες βοηθούν ώστε να ερμηνεύσουμε τη σχέση ανάμεσα στην ποιοτική μεταβλητή και τις ανεξάρτητες μεταβλητές, προβλέποντας τις πιθανότητες ενός γεγονότος. Οι συναρτήσεις αυτές, γνωστές επίσης και ως σιγμοειδείς μετατρέπουν τις πιθανότητες σε δίτιμες τιμές, ώστε αργότερα να χρησιμοποιηθούν στις προβλέψεις μας.

Οι 2 συνηθέστερες περιπτώσεις λογιστικής παλινδρόμησης είναι οι εξής

1) **Διωνυμική λογιστική παλινδρόμηση.** Η εξαρτημένη μεταβλητή έχει μόνο δύο πιθανά αποτελέσματα (π.χ. 0: ήττα 1: νίκη)

2) **Πολυωνυμική λογιστική παλινδρόμηση.** Η εξαρτημένη μας μεταβλητή έχει περισσότερα από 2 αποτελέσματα (π.χ. η πρόβλεψη της ποιότητας ενός φαγητού 0: κακό 1: καλό 2: άριστο)

Μια άλλη περίπτωση λογιστικής παλινδρόμησης που υπάρχει στη βιβλιογραφία, είναι η περίπτωση της διατακτικής λογιστικής παλινδρόμησης, κατά την οποία η εξαρτημένη μεταβλητή λαμβάνει περισσότερες από δύο τιμές που μεταξύ τους υπάρχει η ιδιότητα της διάταξης. Ωστόσο, στα πλαίσια της παρούσας εργασίας θα ασχοληθούμε εκτενώς με τις δύο πρώτες περιπτώσεις. Σε επίπεδο πληθυσμού, ο τύπος της απλής γραμμικής παλινδρόμησης (με μία ανεξάρτητη μεταβλητή x), είναι ο

$$y = a + \beta_1 x,$$

ενώ αντίστοιχα ο τύπος που περιγράφει την πολλαπλή γραμμική παλινδρόμηση (2 ή περισσότερες ανεξάρτητες μεταβλητές), είναι ο

$$y = a + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p,$$

όπου με $\beta_1, \beta_2, \dots, \beta_p$ θεωρούμε τις πραγματικές επιδράσεις των μεταβλητών $x_1, x_2, \dots, x_p$στην μεταβλητή y.

Εφαρμόζοντας την σιγμοειδή συνάρτηση παίρνουμε

$$p(x) = \frac{1}{1 + e^{-y}}$$

ή

$$1 + e^{-y} = \frac{1}{p(x)}$$

ή

$$e^{-y} = \frac{1 - p(x)}{p(x)}$$

ή

$$lne^{-y} = \ln\left(\frac{1 - p(x)}{p(x)}\right)$$

ή

$$-y = \ln\left(\frac{1 - p(x)}{p(x)}\right)$$

ή

$$y = \ln\left(\frac{p(x)}{1-p(x)}\right).$$

Οπότε αν αντικαταστήσουμε την παραπάνω σχέση στην εξίσωση της πολλαπλή γραμμική παλινδρόμηση, προκύπτει

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

Οι εκτιμήσεις των παραμέτρων $\beta_1, \beta_2, \dots, \beta_p$ του μοντέλου της λογιστικής παλινδρόμησης, έστω κατ' αντιστοιχία $b_0, b_1, \dots, b_p$ εκτιμώνται χρησιμοποιώντας τη μέθοδο της μέγιστης πιθανοφάνειας, κάνοντας χρήση τόσο των τιμών των ανεξάρτητων μεταβλητών για κάθε παρατήρηση $\boldsymbol{x} = (x_1, \dots, x_p)$ όσο και την αντίστοιχη τιμή της εξαρτημένης μεταβλητής y_i, καθώς αναφερόμαστε σε αλγόριθμο μάθησης με επίβλεψη.

Παρόλο που η λογιστική παλινδρόμηση ανήκει στην κατηγορία των γραμμικών μοντέλων, δεν έχει κοινές προϋποθέσεις με τα μοντέλα απλής γραμμικής παλινδρόμησης. Παραδείγματος χάριν, δεν απαιτεί γραμμική σχέση μεταξύ εξαρτημένης και ανεξάρτητων μεταβλητών. Δεν υπάρχει τυχαίο σφάλμα στο λογιστικό μοντέλο όπως στην γραμμική παλινδρόμηση και εφαρμόζεται ανεξάρτητα από το αν οι ανεξάρτητες μεταβλητές ακολουθούν την κανονική κατανομή.

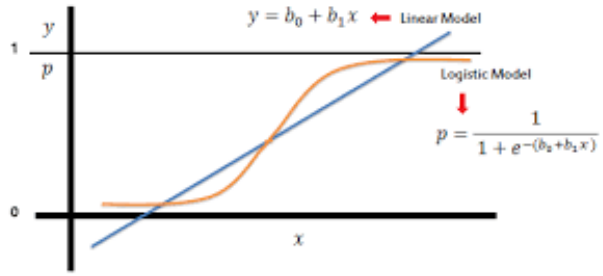
Ωστόσο, για τη διαδικασία προσαρμογής της θεωρούμε ότι ακολουθούνται κάποιες άλλες προϋποθέσεις/περιορισμοί. Συγκεκριμένα,

1) Θεωρούμε τη μη ύπαρξη πολυσυγγραμικότητας μεταξύ των ανεξάρτητων μεταβλητών.
2) Θεωρούμε ότι οι ανεξάρτητες μεταβλητές σχετίζονται γραμμικά με τον λογάριθμο των πιθανοτήτων.
3) Προϋποθέτουμε ένα μεγάλο δείγμα για καλή πρόβλεψη.
4) Προϋποθέτουμε ότι οι παρατηρήσεις είναι ανεξάρτητες η μία από την άλλη. Δεν υπάρχουν τιμές που να επηρεάζουν την ανεξάρτητες μεταβλητές.
5) Η λογιστική παλινδρόμηση με 2 κλάσεις απαιτεί η εξαρτημένη μεταβλητή είναι δυαδική, ενώ π.χ. η διατακτική λογιστική παλινδρόμηση απαιτεί να υπάρχει διάταξη στην εξαρτημένη μεταβλητή

Η επιλογή μεταξύ της χρήσης γραμμικής ή λογιστικής παλινδρόμησης βασίζεται στις τιμές της εξαρτημένης μεταβλητής Y. Η γραμμική παλινδρόμηση προβλέπει συνεχείς μεταβλητές όπως πχ την αξία ενός σπιτιού και το αποτέλεσμα που θα προκύψει κυμαίνεται από το -∞ έως +∞. Από την στιγμή που οι προβλεπόμενες τιμές δεν είναι τιμές πιθανότητας, αλλά συνεχείς τιμές για τις κατηγορίες, θα ήταν δύσκολο να βρεθεί το κατάλληλο όριο, το οποίο θα μας βοηθήσει ώστε να διαχωρίζουμε τις κλάσεις ταξινόμησης.

Όσον αφορά τη γραφική παράσταση της λογιστικής παλινδρόμησης, αντικαθιστώντας την εξίσωση της απλής γραμμικής παλινδρόμησης στην εξίσωση $p(x) = \frac{1}{1+e^{-y}}$, παίρνουμε

$$p(x) = \frac{1}{1 + e^{-(a+\beta * x_1)}}.$$



Γράφημα 5 - Σύγκριση γραμμικής και λογιστικής παλινδρόμησης

Η συνάρτηση αυτή δίνει το παραπάνω σιγμοειδές γράφημα για μια λογιστική παλινδρόμηση με μία ανεξάρτητη μεταβλητή, ενώ παρουσιάζεται και η διαγραμματική διαφοροποίηση μεταξύ γραμμικής και λογιστικής παλινδρόμησης. Από το σχήμα φαίνεται ότι μεγάλες τιμές του Y οδηγούν σε μεγάλες πιθανότητες, ενώ μικρές σε πιθανότητες κοντά στο 0.

Όσα αναφέρθηκαν παραπάνω, αποτελούν μια εισαγωγή/συνοπτική παρουσίαση των όσων θα ακολουθήσουν στη συνέχεια του κεφαλαίου. Στις επόμενες παραγράφους θα ασχοληθούμε με την περιγραφή της κατασκευής των μοντέλων διωνυμικής και πολυωνυμικής λογιστικής παλινδρόμησης, για μία ή περισσότερες ανεξάρτητες μεταβλητές X. Παράλληλα, θα ασχοληθούμε με τη μεθοδολογία της μέγιστης πιθανοφάνειας, η οποία οδηγεί στην εκτίμηση των παραμέτρων των εξεταζόμενων μοντέλων. Τέλος, θα αναφερθούμε σε μέτρα διερεύνησης της αποδοτικής προσαρμογής των μοντέλων λογιστικής παλινδρόμησης στα δεδομένα, καθώς και για το πώς πραγματοποιείται η επιλογή των στατιστικά σημαντικών μεταβλητών του μοντέλου.

5.1 Διωνυμική λογιστική παλινδρόμηση

Αυτού του είδους λογιστικής παλινδρόμησης περιλαμβάνει διακριτές/ποιοτικές εξαρτημένες μεταβλητές που παίρνουν αποκλειστικά δύο τιμές. Οι δυαδικές αυτές μεταβλητές κωδικοποιούνται συχνά με το 1 που δηλώνει "επιτυχία" και το 0 που δηλώνει "αποτυχία" ανάλογα με το φαινόμενο που μας ενδιαφέρει. Ο μέσος όρος μιας διχοτομικής μεταβλητής με κωδικούς 1

και 0 είναι ίσος με το ποσοστό των περιπτώσεων με κωδικό 1 στο συνολικό δείγμα και συχνά λαμβάνει την έννοια της πιθανότητας.

5.1.1 Απλή διωνυμική λογιστική παλινδρόμηση

Για μια δυαδική εξαρτημένη μεταβλητή (response variable) Y και μία ανεξάρτητη μεταβλητή X (predictor), ισχύει

$$p(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x).$$

Σε αυτή την περίπτωση το μοντέλο απλής λογιστικής παλινδρόμησης παίρνει τη μορφή

$$p(x) = \frac{\exp(a + \beta x)}{1 + \exp(a + \beta x)}.$$

Ισοδύναμα, το μοντέλο logit (log odds) οδηγεί σε γραμμική σχέση με την ανεξάρτητη μεταβλητή (predictor), καθώς

$$logit\big(p(x)\big) = log\frac{p(x)}{1 - p(x)} = a + \beta x.$$

5.1.2 Πολλαπλή διωνυμική λογιστική παλινδρόμηση

Όπως και στην περίπτωση της πολλαπλής γραμμικής παλινδρόμησης, μπορούμε να ορίσουμε με παρόμοιο τρόπο την πολλαπλή λογιστική παλινδρόμηση, η οποία χαρακτηρίζεται από δύο και άνω ανεξάρτητες μεταβλητές. Το μοντέλο λογιστικής παλινδρόμησης, για $p(\boldsymbol{x}) = P(Y = 1)$ και συγκεκριμένες τιμές για τις p το πλήθος ανεξάρτητες μεταβλητές $(\boldsymbol{x} = (x_1, x_2, \dots, x_p))$ είναι

$$logit\big(p(\boldsymbol{x})\big) = log\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})} = a + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

Έτσι, η εναλλακτική φόρμουλα για τον απευθείας υπολογισμό της πιθανότητας $p(\boldsymbol{x})$ θα δίνεται από τη σχέση

$$p(\boldsymbol{x}) = \frac{\exp(a + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}{1 + \exp(a + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}.$$

Η παράμετρος β_j αναφέρεται στην επίδραση της μεταβλητής x_j στα log odds ($logit(p(\boldsymbol{x}))$) όταν $Y = 1$, θεωρώντας της υπόλοιπες μεταβλητές x_k σταθερές. Παραδείγματος χάριν, η ποσότητα $\exp(\beta_j)$ αποτελεί την πολλαπλασιαστική επίδραση στην πιθανότητα $p(\boldsymbol{x})$ για μια αύξηση στην τιμή της μεταβλητής x_j της τάξεως της μίας μονάδας, όταν διατηρούμε σταθερές τις τιμές των υπολοίπων μεταβλητών x_k.

5.2 Στατιστική σημαντικότητα των παραμέτρων του μοντέλου

Για το λογιστικό μοντέλο με μία ανεξάρτητη μεταβλητή X,

$$logit(p(x)) = a + \beta x + \varepsilon,$$

ο έλεγχος στατιστικής σημαντικότητας εστιάζει στη διερεύνηση της υπόθεσης ανεξαρτησίας $H_0 : \beta = 0$. Ο έλεγχος του Wald αξιοποιεί το λογάριθμο της πιθανοφάνειας του εκτιμητή b, κάνοντας χρήση του στατιστικού $z = \frac{b}{SE}$ ή του τετραγώνου αυτού. Με SE συμβολίζουμε το τυπικό σφάλμα (standard error) της εκτίμησης της παραμέτρου β. Το τυπικό σφάλμα ισούται με $SE = \frac{1}{\sqrt{I(b)}}$, όπου $I(b) = -E\left[\frac{d^2L(\beta)}{d\beta^2}\right]$, δηλαδή με τον πίνακα πληροφορίας για τον εκτιμητή b (Agresti, 2012). Υπό τη μηδενική υπόθεση H_0, το στατιστικό z^2 ακολουθεί την κατανομή X_1^2 (Χ-τετράγωνο με 1 βαθμό ελευθερίας). Ο έλεγχος των λόγων των πιθανοφανειών (log-likelihood) χρησιμοποιεί το διπλάσιο της διαφοράς μεταξύ της μέγιστη λογαριθμικής πιθανοφάνειας εκτίμηση b και στο σημείο $\beta = 0$, ενώ επίσης ακολουθεί κατανομή X_1^2. Ένα διάστημα εμπιστοσύνης για το β προκύπτει από την αντιστροφή του ελέγχου $H_0 : \beta = 0$. Το διάστημα είναι επί της ουσίας ένα σύνολο τιμών για τις οποίες το στατιστικό X^2 δε γίνεται μεγαλύτερο από την τιμή $X_1^2(\alpha) = z_{\frac{\alpha}{2}}^2$. Οπότε για την προσέγγιση του Wald, το αντίστοιχο διάστημα εμπιστοσύνης δίνεται από τον τύπο

$$\left[\frac{\hat{\beta}}{SE}\right]^2 \leq z_{\frac{\alpha}{2}}^2 \quad \Leftrightarrow \quad \hat{\beta} \pm z_{\frac{\alpha}{2}}^2\, SE.$$

Πέρα όμως από τον έλεγχο στατιστικής σημαντικότητας του συντελεστή β, περαιτέρω ενδιαφέρον έχει η κατασκευή ενός διαστήματος εμπιστοσύνης των εκτιμήσεων/προβλέψεων του μοντέλου της απλής λογιστικής παλινδρόμησης για διάφορες τιμές της μεταβλητής x. Για σταθερό $x = x_0$, παίρνουμε $logit(\hat{p}(x_0)) = \hat{a} + \hat{\beta}x_0$. Το τυπικό σφάλμα αυτής της εκτίμησης προκύπτει από τη ρίζα της διασποράς

$$var(\hat{a} + \hat{\beta}x_0) = var(\hat{a}) + x_0^2 var(\hat{\beta}) + 2x_0 cov(\hat{a}, \hat{\beta}).$$

Έτσι λοιπόν, το διάστημα εμπιστοσύνης για την εκτίμηση $logit\big(p(x_0)\big)$ είναι το

$$\hat{a} + \hat{\beta}x_0 \pm 1.96 \cdot SE.$$

Παίρνοντας τον αντίστροφο μετασχηματισμό για κάθε άκρο του παραπάνω διαστήματος εμπιστοσύνης, μπορούμε να βρούμε το διάστημα εμπιστοσύνης της πιθανότητας

$$p(x_0) = \frac{\exp(\hat{a} + \hat{\beta}x_0)}{1 + \exp(\hat{a} + \hat{\beta}x_0)}.$$

5.3 Μέγιστη Πιθανοφάνεια

Έστω ότι έχουμε δείγμα μεγέθους n, και θεωρούμε τις n μεταβλητές απόκρισης $\boldsymbol{Y}_i$ (μία για κάθε στοιχείο του συνόλου δεδομένων) ανεξάρτητες μεταξύ τους. Επίσης, έστω ότι συμβολίζουμε με $\boldsymbol{x}_i = (x_{i0}, x_{i1}, \dots, x_{ip})$ τις τιμές των p το πλήθος ανεξάρτητων μεταβλητών (predictors) για το i στοιχείο του δείγματος. Για την πρώτη συνιστώσα, ισχύει $x_{i0} = 1$, θεωρώντας την ως σταθερό όρο $\forall\ i\ =\ 1, \dots, n$. Το μοντέλο λογιστικής παλινδρόμησης, που αντιμετωπίζει τον σταθερό όρο β_0 ως παράμετρο μιας ανεξάρτητης/επεξηγηματικής μεταβλητής με τιμή 1, είναι το

$$p(\boldsymbol{x}_i) = \frac{\exp(\sum_{j=0}^{p} \beta_j x_{ij})}{1 + \exp(\sum_{j=0}^{p} \beta_j x_{ij})}.$$

Επίσης, συμβολίζουμε με y_i το πλήθος των μεταβλητών απόκρισης $Y_i\ \forall\ i\ =\ 1, \dots, N$ του δείγματος που έχουν την τιμή 1 και με n_i το πλήθος των φορών που συναντάμε τις ίδιες τιμές για τη μεταβλητή $\boldsymbol{x}_i$ στο δείγμα. Εφόσον θεωρήσαμε ότι οι μεταβλητές απόκρισης $\{Y_1, Y_2, \dots, Y_N\}$ είναι ανεξάρτητες μεταξύ τους και ακολουθούν τη διωνυμική κατανομή, ισχύει ότι

$$E(Y_i) = n_i p(\boldsymbol{x}_i)$$

όπου $n_1 + n_2 + \cdots + n_N = n$. Η από κοινού συνάρτηση πυκνότητας πιθανότητας είναι ανάλογη του γινομένου των N διωνυμικών συναρτήσεων,

$$\begin{aligned}\prod_{i=1}^{N} p(\boldsymbol{x}_i)^{y_i}\big(1 - p(\boldsymbol{x}_i)\big)^{n-y_i} &= \{\prod_{i=1}^{N} \exp\left[\log\left(\frac{p(\boldsymbol{x}_i)}{1 - p(\boldsymbol{x}_i)}\right)^{y_i}\right]\}\{\prod_{i=1}^{N}\big(1 - p(\boldsymbol{x}_i)\big)^{n_i}\} \\ &= \{\exp[\textstyle\sum_{i=1}^{N} y_i log\, \frac{p(x_i)}{1-p(x_i)}]\}\{\textstyle\prod_{i=1}^{N}\big(1 - p(\boldsymbol{x}_i)\big)^{n_i}\}.\end{aligned}$$

Το δεξί μέρος της εξίσωσης λογιστικής παλινδρόμησης για το i στοιχείο του δείγματος, είναι $(\sum_{j=0}^{p} \beta_j x_{ij})$. Έτσι ο αντίστοιχος εκθετικός όρος μπορεί να γραφεί ως

$$\exp[\sum_{i=1}^{N} y_i log \frac{p(\boldsymbol{x}_i)}{1-p(\boldsymbol{x}_i)}] = \exp[\sum\nolimits_i y_i (\sum\nolimits_j \beta_j x_{ij})] = \exp[\sum\nolimits_j (\sum\nolimits_i y_j x_{ij}) \beta_j].$$

Επιπλέον, καθώς $[1-p(\boldsymbol{x}_i)] = [1+\exp(\sum_{j=0}^{p} \beta_j x_{ij})]^{-1}$, η λογαριθμική συνάρτηση πιθανοφάνειας δίνεται ως

$$\begin{aligned} L(\boldsymbol{\beta}) &= \log\{\exp[\sum_{i=1}^{N} y_i log \frac{p(\boldsymbol{x}_i)}{1-p(\boldsymbol{x}_i)}]\} \{\prod_{i=1}^{N} (1-p(\boldsymbol{x}_i))^{n_i}\}] \\ &= \log \{\exp[\sum_{i=1}^{N} y_i log \frac{p(\boldsymbol{x}_i)}{1-p(\boldsymbol{x}_i)}]\} \{\prod_{i=1}^{N} \left(\left[1+\exp(\sum\nolimits_{j=0}^{p} \beta_j x_{ij})\right]^{-1}\right)^{n_i}\} \\ &= \sum\nolimits_j (\sum\nolimits_i y_j x_{ij}) \beta_j + \log \{\prod_{i=1}^{N} \left(\left[1+\exp(\sum\nolimits_{j=0}^{p} \beta_j x_{ij})\right]^{-1}\right)^{n_i}\} \\ &= \sum\nolimits_j (\sum\nolimits_i y_j x_{ij}) \beta_j + \log \{\prod_{i=1}^{N} \left(1+\exp(\sum\nolimits_{j=0}^{p} \beta_j x_{ij})\right)^{-n_i}\} \\ &= \sum\nolimits_j (\sum\nolimits_i y_j x_{ij}) \beta_j - \sum_i n_i \log\left[1+\exp\left(\sum_j \beta_j x_{ij}\right)\right]. \end{aligned}$$

Οι εκτιμητές μέγιστης πιθανοφάνειας προκύπτουν από την εξίσωσης της παραγώγου της λογαριθμικής πιθανοφάνειας ως προς τις παραμέτρους του μοντέλου με το 0

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta} = 0.$$

Για παράδειγμα, παραγωγίζοντας τη συνάρτηση λογαριθμικής πιθανοφάνειας ως προς β_j, παίρνουμε (Agresti 2013)

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_i y_j x_{ij} - \sum_i n_i x_{ij} \frac{\exp(\sum_k \beta_k x_{ik})}{1 + \exp(\sum_k \beta_k x_{ik})}.$$

Εξισώνοντας με το μηδέν καταλήγουμε στη σχέση

$$\sum_i y_j x_{ij} - \sum_i n_i x_{ij} \hat{p}(\boldsymbol{x}_i) = 0, \quad j = 0,1, \dots, p\ ,$$

όπου το

$$\hat{p}(\boldsymbol{x}_i) = \frac{\exp(\sum_k \hat{\beta}_k x_{ik})}{1 + \exp(\sum_k \hat{\beta}_k x_{ik})}$$

είναι ο εκτιμητής μέγιστης πιθανοφάνειας της πιθανότητας $p(\boldsymbol{x}_i)$. Εφόσον η παραπάνω εξίσωση είναι μη γραμμική, ο υπολογισμός των εκτιμήσεων μέγιστης πιθανοφάνειας βασίζεται σε επαναληπτικές διαδικασίες, όπως η μέθοδος Newton-Raphson. (Agresti 2013)

5.4 Μέτρα καλής προσαρμογή μοντέλου στα δεδομένα

Μια διαφορά μεταξύ της αναμενόμενης συνάρτησης κατανομής και της εμπειρικής συνάρτησης κατανομής χρησιμοποιείται συχνά για την εκτέλεση δοκιμών καλής προσαρμογής. Η θεωρία της εμπειρικής συνάρτησης κατανομής είναι καλά εδραιωμένη για ολόκληρα δείγματα, ανεξάρτητα από το εάν οι παράμετροι είναι γνωστές ή άγνωστες. Προκειμένου να αξιολογηθούν τα μοντέλα, πρέπει να ληφθούν υπόψη μια σειρά από διαφοροποιημένα και αμφιλεγόμενα προβλήματα, τα οποία έχουν ωθήσει ορισμένους ακαδημαϊκούς να αναρωτηθούν εάν οι μετρήσεις καλής προσαρμογής πρέπει να χρησιμοποιούνται εξαρχής. Ωστόσο, οι ποσοτικές προβλέψεις συνεχίζουν να αποτελούν βασικό στοιχείο των χρησιμοποιούμενων μοντέλων και οι μετρήσεις καλής προσαρμογής με τη μία ή την άλλη μορφή εξακολουθούν να είναι η τυπική μέθοδος για την αξιολόγηση ποσοτικών (π.χ. γραμμική παλινδρόμηση) ή ποιοτικών προβλέψεων (π.χ. λογιστική παλινδρόμηση).

Υπάρχουν δύο τρόποι αξιολόγησης της καλής προσαρμογής ενός μοντέλου στα δεδομένα. Ο πρώτος τρόπος επαφίεται στη χρήση μεθόδων οπτικής διερεύνησης/παρουσίασης, όπου επιτρέπουν μια οπτική σύγκριση των ομοιοτήτων και των διαφορών μεταξύ των προβλέψεων των μοντέλων και των παρατηρούμενων δεδομένων. Ο δεύτερος τρόπος είναι ο τρόπος μέσω της χρήσης αριθμητικών μετρήσεων, όπου παρέχουν συνοπτικά μέτρα της συνολικής ακρίβειας των προβλέψεων.

Σύμφωνα με τη δεύτερη οπτική, ένα μέτρο καλής προσαρμογής δίνεται από τη σχέση

$$R^2 = 1 - \frac{H}{H_0}.$$

Στην παραπάνω σχέση, η ποσότητα H_0^2 αποτελεί το λογάριθμο της πιθανοφάνειας ενός μοντέλου που χαρακτηρίζεται από πλήρη ανεξαρτησία μεταξύ της μεταβλητής απόκρισης Y και των ανεξάρτητων επεξηγηματικών μεταβλητών $\boldsymbol{X}$. Από την άλλη πλευρά, η ποσότητα H^2 αποτελεί το λογάριθμο της πιθανοφάνειας του εξεταζόμενου μοντέλου. Αντίστοιχα, ένας δείκτης καλής προσαρμογής που λαμβάνει υπόψη και την πολυπλοκότητα του προσαρμοσμένου μοντέλου, είναι το «προσαρμοσμένο» R^2 που δίνεται από τη σχέση

$$R^2 = 1 - \frac{\frac{H}{q-r}}{\frac{H_0}{q_0 - r_0}},$$

όπου με q συμβολίζουμε το πλήθος των κελιών του πίνακα συνάφειας, ενώ με r και r_0 τους βαθμούς ελευθερίας των μοντέλων με πιθανοφάνεια H^2 και H_0^2 αντίστοιχα. Και οι δύο παραπάνω δείκτες παίρνουν τιμές μεταξύ 0 και 1, ενώ τιμές πολύ κοντά στη μονάδα αντιστοιχούν σε πολύ ικανοποιητικά μοντέλα.

5.5. Πολυωνυμική λογιστική παλινδρόμηση

Όπως αναφέραμε και στην εισαγωγή του παρόντος κεφαλαίου, μια εξίσου σημαντική κατηγορία μοντέλων λογιστικής παλινδρόμησης είναι η πολυωνυμική λογιστική παλινδρόμηση (multinomial logistic regression). Στόχος είναι η αντιμετώπιση προβλημάτων όπου η εξαρτημένη μεταβλητή απόκρισης, λαμβάνει περισσότερες από δύο τιμές. Όπως και η διωνυμική λογιστική παλινδρόμηση, έτσι και η πολυωνυμική χρησιμοποιεί εκτιμήσεις μέγιστης πιθανότητας για τον προσδιορισμό της πιθανότητας κατηγοριοποίησης $p(\boldsymbol{x}_i)$.

Ωστόσο, απαιτείται προσεκτική αξιολόγηση του μεγέθους του δείγματος καθώς και σωστή διαχείριση των ακραίων τιμών (outliers) όταν χρησιμοποιείται η πολυωνυμική λογιστική παλινδρόμηση. Επιπλέον, η πολυωνυμική λογιστική παλινδρόμηση θεωρείται συχνά ως μια ελκυστική εναλλακτική, καθώς δεν απαιτεί υποθέσεις κανονικότητας, γραμμικότητας ή ομοσκεδαστικότητας. Απαραίτητο ωστόσο, είναι να τηρούνται υποθέσεις, όπως αυτή της ανεξαρτησίας μεταξύ των δυνατών επιλογών της εξαρτημένης μεταβλητής απόκρισης. Η βασική προϋπόθεση αυτής της υπόθεσης είναι ότι η επιλογή ή η ιδιότητα μέλους σε μια ομάδα δεν έχει καμία επίδραση στην επιλογή ή στην συμμετοχή σε μια άλλη κατηγορία/τιμή της εξαρτημένης μεταβλητής (Engel 1988).

Όσον αφορά την προσαρμογή του μοντέλου αυτού, θεωρούμε αρχικά ότι η εξαρτημένη μεταβλητή απόκρισης μπορεί να πάρει K το πλήθος δυνατές τιμές. Σε αυτή την περίπτωση δημιουργούμε $K - 1$ μοντέλα διωνυμικής λογιστικής παλινδρόμησης, καθώς η μία από τις K

κατηγορίες της πολυωνυμικής παλινδρόμησης θα λειτουργήσει ως οδηγός (pivot). Συγκεκριμένα, επιλέγοντας ως οδηγό, την τελευταία κατηγορία της μεταβλητής απόκρισης, έχουμε

$$ln\frac{P(Y_i = 1)}{P(Y_i = K)} = \boldsymbol{\beta_1} \cdot \boldsymbol{X_i}$$

$$ln\frac{P(Y_i = 2)}{P(Y_i = K)} = \boldsymbol{\beta_2} \cdot \boldsymbol{X_i}$$

...

$$ln\frac{P(Y_i = K-1)}{P(Y_i = K)} = \boldsymbol{\beta_{K-1}} \cdot \boldsymbol{X_i}$$

Τα $\boldsymbol{\beta_i}$ αποτελούν διανύσματα παραμέτρων, που χρησιμοποιούνται για την προσαρμογή μοντέλων διωνυμικής παλινδρόμησης και μπορούν να εκτιμηθούν μέσω των μεθοδολογιών που παρουσιάστηκαν στις παραγράφους 4.2 και 4.3. Έτσι λοιπόν, έχουμε δημιουργήσει ένα σύνολο $K-1$ μοντέλων που το καθένα δίνει την πιθανότητα συμπερίληψης (της εκάστοτε παρατήρησης του συνόλου δεδομένων), σε κάθε ζεύγος εξεταζόμενων μεταβλητών. Συνεπώς, το μοντέλο $ln\frac{P(Y_i=1)}{P(Y_i=K)} = \boldsymbol{\beta_1} \cdot \boldsymbol{X_i}$ δίνει την πιθανότητα, δεδομένων των τιμών ανεξάρτητων μεταβλητών $\boldsymbol{x_i} = (x_{i0}, x_{i1}, \dots, x_{ip})$, το στοιχείο i να ανήκει στην κατηγορία 1, και όχι στην K.

Εφαρμόζοντας την εκθετική συνάρτηση και στα δύο μέλη των παραπάνω $K-1$ μοντέλων, παίρνουμε

$$P(Y_i = 1) = P(Y_i = K)e^{\boldsymbol{\beta}_1 \boldsymbol{X}_i}$$

$$P(Y_i = 2) = P(Y_i = K)e^{\boldsymbol{\beta}_2 \boldsymbol{X}_i}$$

...

$$P(Y_i = K-1) = P(Y_i = K)e^{\boldsymbol{\beta}_{K-1} \boldsymbol{X}_i}.$$

Δεδομένου ότι οι K παραγόμενες πιθανότητες πρέπει να αθροίζουν στη μονάδα, βρίσκουμε

$$P(Y_i = K) = 1 - \sum_{k=1}^{K-1} P(Y_i = k) = 1 - \sum_{k=1}^{K-1} P(Y_i = K)e^{\boldsymbol{\beta}_k \boldsymbol{X}_i},$$

ή

$$P(Y_i = K) = \frac{1}{1 + \sum_{k=1}^{K-1} P(Y_i = K)e^{\boldsymbol{\beta}_k \boldsymbol{X}_i}}.$$

Μπορούμε πλέον εύκολα να χρησιμοποιήσουμε την παραπάνω σχέση για να υπολογίσουμε τις πιθανότητες

$$P(Y_i = 1) = \frac{e^{\boldsymbol{\beta}_1 \boldsymbol{X}_i}}{1 + \sum_{k=1}^{K-1} P(Y_i = K) e^{\boldsymbol{\beta}_1 \boldsymbol{X}_i}}$$

$$P(Y_i = 2) = \frac{e^{\boldsymbol{\beta}_2 \boldsymbol{X}_i}}{1 + \sum_{k=1}^{K-1} P(Y_i = K) e^{\boldsymbol{\beta}_k \boldsymbol{X}_i}}$$

...

$$P(Y_i = K - 1) = \frac{e^{\boldsymbol{\beta}_{K-1} \boldsymbol{X}_i}}{1 + \sum_{k=1}^{K-1} P(Y_i = K) e^{\boldsymbol{\beta}_k \boldsymbol{X}_i}}.$$

Αντιλαμβανόμαστε, ότι το i στοιχείο του συνόλου δεδομένων θα ταξινομηθεί/κατηγοριοποιηθεί στην κλάση/κατηγορία που χαρακτηρίζεται από τη μεγαλύτερη πιθανότητα. Η δυνατότητα κατασκευής πολλαπλών διωνυμικών λογιστικών παλινδρομήσεων επαφίεται στην υπόθεση της ανεξαρτησίας μεταξύ των δυνατών εναλλακτικών/τιμών της εξαρτημένης μεταβλητής.

Κεφάλαιο 6ο. Εφαρμογές στην βιοστατιστική

Ως επιδημιολογία ορίζεται η μελέτη της κατανομής διαφόρων νοσημάτων στον ανθρώπινο πληθυσμό και των παραγόντων που διαμορφώνουν την κατανομή αυτή (Τριχόπουλος 1982). Η καρδιαγγειακή επιδημιολογία ξεκίνησε τη δεκαετία του 1930 ως συνέπεια των παρατηρούμενων αλλαγών στα αίτια της θνησιμότητας. Το 1948, ξεκίνησε η Καρδιολογική Μελέτη Framingham από την Υπηρεσία Δημόσιας Υγείας των ΗΠΑ για τη μελέτη της επιδημιολογίας και των παραγόντων κινδύνου για καρδιαγγειακή νόσο. Την ίδια χρονιά το Εθνικό Ινστιτούτο Υγείας επεκτάθηκε για να συμπεριλάβει πολλά ινστιτούτα, το καθένα αφιερωμένο στη μελέτη συγκεκριμένων ασθενειών (Last, 1995). Η συγκεκριμένη μελέτη μεταφέρθηκε στο Εθνικό Ινστιτούτο Καρδιολογίας που ιδρύθηκε το 1949, τώρα γνωστό ως Εθνικό Ινστιτούτο Καρδιολογίας, Πνευμονολογίας και Αιματολογίας, και παραμένει υπό τη σημερινή διεύθυνση. Τέσσερα χρόνια μετά την έναρξη Καρδιολογική Μελέτη Framingham, οι ερευνητές εντόπισαν τα υψηλά επίπεδα χοληστερόλης και την υψηλή αρτηριακή πίεση ως σημαντικούς παράγοντες κινδύνου για την ανάπτυξη των καρδιαγγειακών παθήσεων.

Σήμερα, ένας παράγοντας κινδύνου ορίζεται ως ένα μετρήσιμο χαρακτηριστικό που σχετίζεται αιτιολογικά με την αυξημένη συχνότητα της νόσου και που αποτελεί σημαντικό ανεξάρτητο προγνωστικό παράγοντα αυξημένου κινδύνου εμφάνισης της νόσου. Αυτή η ευρεία επισκόπηση περιγράφει μερικές από τις πιο σημαντικές γνώσεις σχετικά με τα αίτια των καρδιαγγειακών παθήσεων που προέκυψαν από την Καρδιολογική Μελέτη Framingham. Η έμφαση δίνεται στον εντοπισμό των παραγόντων κινδύνου και στην αξιολόγηση της προγνωστικής τους ικανότητας και των επιπτώσεων τους στην πρόληψη της νόσου.

Από το 1970 η Καρδιολογική Μελέτη Framingham συνδέεται επίσης στενά με το Πανεπιστήμιο της Βοστώνης. Η πόλη Framingham, που βρίσκεται 32 χιλιόμετρα δυτικά της Βοστώνης της Μασαχουσέτης, επιλέχθηκε επειδή ήταν ο τόπος μιας επιτυχημένης κοινοτικής μελέτης για τη φυματίωση που πραγματοποιήθηκε το 1918 και λόγω της γειτνίασης της με τα μεγάλα ιατρικά κέντρα της Βοστώνης, η παρουσία πολλών μεγάλων εργοδοτών είχε ως αποτέλεσμα την υποστήριξη μιας καλά ενημερωμένης και άκρως συνεργάσιμης ιατρικής και πολιτικής κοινότητας. Η πρώτη ομάδα περιελάβανε 5209 υγιείς κατοίκους ηλικίας μεταξύ 30 και 60 ετών που εγγράφηκαν το 1948 για διετείς εξετάσεις. Το 1971, 5124 γιοι και κόρες (και οι σύζυγοι τους) των αρχικών ατόμων της ομάδας στρατολογήθηκαν για τη Μελέτη Απογόνων. Τέλος, το 2002, 4095 συμμετέχοντες συμπεριλήφθηκαν στην ομάδα τρίτης γενιάς της μελέτης (Last, 1995). Το πιο σημαντικό, η Καρδιολογική Μελέτη Framingham και άλλες επιδημιολογικές μελέτες, έδειξαν ότι η συστολική και διαστολική αρτηριακή πίεση έχει μια συνεχή, ανεξάρτητη, διαβαθμισμένη και θετική συσχέτιση με τα καρδιαγγειακά αποτελέσματα.

Ακόμη και οι υψηλές φυσιολογικές τιμές της αρτηριακής πίεσης συνδέονται με αυξημένο κίνδυνο καρδιαγγειακής νόσου. Υπό το φως αυτών των μελετών, η έκθεση Joint National Committee VII ανέπτυξε μια νέα ταξινόμηση της αρτηριακής πίεσης για ενήλικες από 18 ετών και άνω, συμπεριλαμβανομένης μιας νέας κατηγορίας που ονομάζεται προυπέρταση, καθώς αυτά τα άτομα διατρέχουν αυξημένο κίνδυνο εξέλιξης σε υπέρταση και παρουσιάζουν αυξημένες πιθανότητες καρδιαγγειακής νόσου (Last, 1995).

Για άτομα ηλικίας 40 έως 70 ετών, κάθε αύξηση 20 mm υδραργύρου στη συστολική αρτηριακή πίεση ή 10 mm υδραργύρου στη διαστολική αρτηριακή πίεση διπλασιάζει τον κίνδυνο καρδιαγγειακής νόσου σε όλο το εύρος της αρτηριακής πίεσης από 115/75 έως 185/115 mm υδραργύρου. Σε κλινικές δοκιμές, η αντιυπερτασική θεραπεία έχει συσχετιστεί με 35% έως 40% μείωση της επίπτωσης εγκεφαλικού επεισοδίου, το 20% έως 25% έχει συσχετιστεί με έμφραγμα του μυοκαρδίου και πάνω από 50% έχει συσχετιστεί με καρδιακή ανεπάρκεια. Στην Ισπανία η συχνότητα της υπέρτασης είναι υψηλός και υπολογίζεται ότι είναι περίπου 34% στον ενήλικο πληθυσμό. Στον πληθυσμό της Girona, οι τάσεις στην ευαισθητοποίηση, τη θεραπεία και τον έλεγχο έχουν βελτιωθεί τα τελευταία 10 χρόνια, αν και το ποσοστό ελεγχόμενης υπέρτασης απέχει ακόμη πολύ από το ιδανικό (Last, 1995).

6.1. Εφαρμογή σε πραγματικά δεδομένα

Για την παρούσα εφαρμογή θα αξιοποιήσουμε τα δεδομένα της καρδιολογικής μελέτης του Framingham για τη διερεύνηση της υπέρτασης με βάση το κάπνισμα.

(https://rdrr.io/cran/LocalControl/man/framingham.html) . Τα δεδομένα που συλλέχθηκαν σε μια μελέτη 24 ετών, κατάλληλη για την ανάλυση κινδύνων που μπορούν να οδηγήσουν στην εμφάνιση υπέρτασης ή θανάτου σε συναρτήσει του καπνίσματος. Το σύνολο δεδομένων περιέχει τις παρατηρήσεις 2316 ασθενών και 11 στήλες που αφορούν δημογραφικά στοιχεία αυτών αλλά και στοιχεία της υγείας τους. Συγκεκριμένα, τα δημογραφικά στοιχεία περιέχουν μεταβλητές όπως

1) το φύλο και
2) η ηλικία των ασθενών,

ενώ οι υπόλοιπες μεταβλητές αναφέρονται σε χαρακτηριστικά όπως

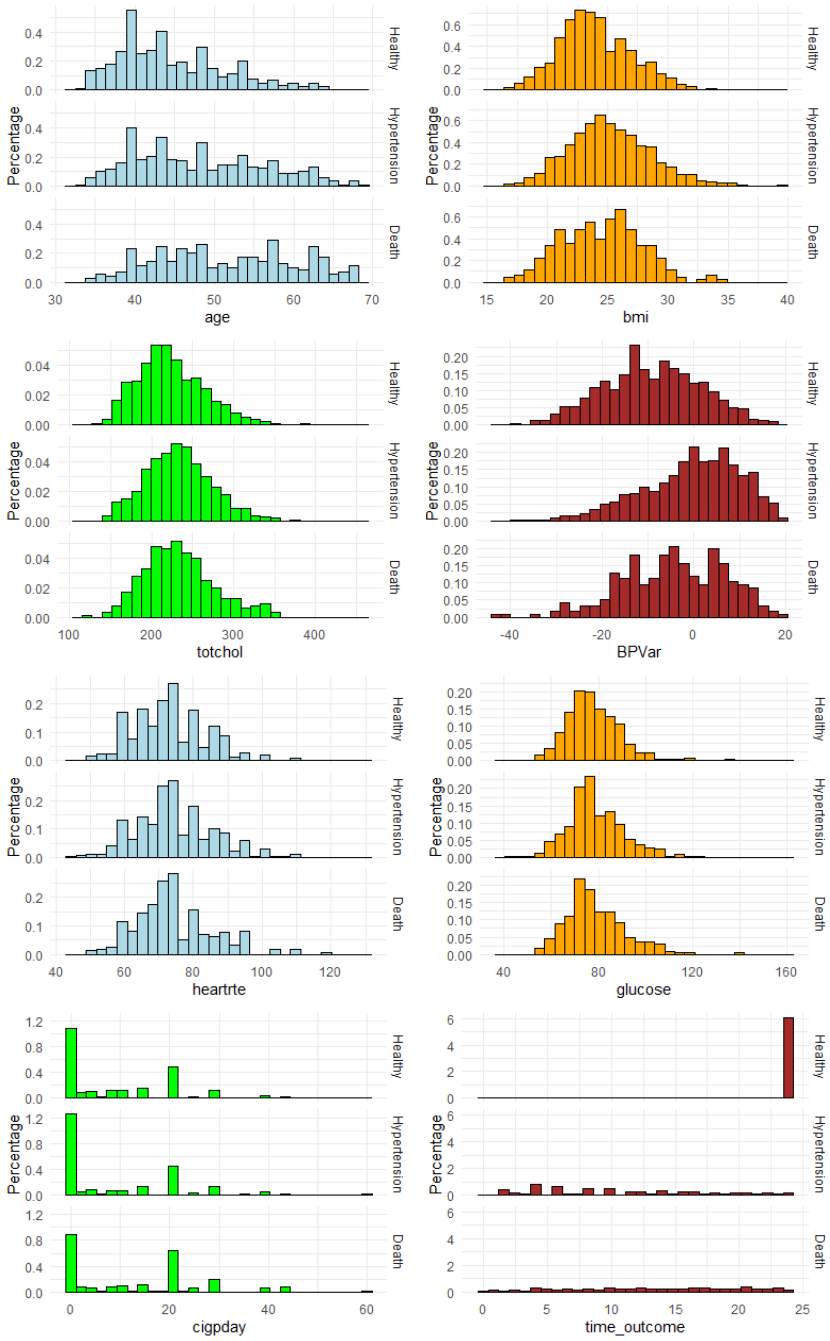
3) η συχνότητα καπνίσματος,
4) ο δείκτης μάζας σώματος,
5) τα επίπεδα γλυκόζης στο αίμα,
6) η μέση αρτηριακή πίεση,
7) τα επίπεδα χοληστερόλης του ασθενή,
8) το αν είναι καπνιστής κατά την περίοδο διεξαγωγής της έρευνας,
9) ο καρδιακός ρυθμός κατά τη διάρκεια της μέτρησης
10) ο χρόνος εμφάνισης του τελικού αποτελέσματος
11) το τελικό αποτέλεσμα : (α) θάνατος, (β) περιστατικό υπέρτασης, (γ) έξοδος από την έρευνα χωρίς ο ασθενής να εμφανίσει πρόβλημα υγείας

Οι μεταβλητές αυτές (μεταβλητές 1 – 10), θα χρησιμοποιηθούν ως ανεξάρτητες μεταβλητές, τροφοδοτώντας ένα μοντέλο πολυωνυμικής λογιστικής παλινδρόμησης με στόχο την αποδοτικότερη πρόβλεψη της κατηγορικής μεταβλητής του συνόλου δεδομένων (μεταβλητή 11). Η κατηγορική μεταβλητή χωρίζεται σε τρεις κλάσεις ανάλογα με την κατάσταση της υγείας των συμμετεχόντων στην έρευνα. Η πρώτη κλάση αναφέρεται στους αποθανόντες κατά τη διάρκεια της έρευνας, η δεύτερη σε άτομα τα οποία εμφάνισαν υπέρταση και η τρίτη σε άτομα που δεν ανήκαν σε καμία από τις παραπάνω περιπτώσεις. Στόχος είναι μέσω του μοντέλου της λογιστικής παλινδρόμησης να επιτύχουμε την υψηλότερη δυνατή διαγνωστική ακρίβεια εξετάζοντας τη συνεισφορά των μεταβλητών στο προτεινόμενο μοντέλο ταξινόμησης. Για την πραγματοποίηση της εφαρμογής θα στραφούμε στη χρήση του λογισμικού R.

Τέλος, κατά τη διάρκεια της ανάλυσης και πριν την παρουσίαση των τελικών επιπέδων ακρίβειας, θα ασχοληθούμε με την περιγραφή των κατανομών των ανεξάρτητων και εξαρτημένων μεταβλητών του προβλήματος, μέσω της κατασκευής ιστογραμμάτων. Επιπλέον, μέσω της χρήση του κριτηρίου πληροφορίας AIC θα γίνει επιλογή του κατάλληλου υποσυνόλου ανεξάρτητων μεταβλητών που ελαχιστοποιεί την τιμή του παραπάνω κριτηρίου και θα συγκριθεί η ακρίβεια αυτού του (βέλτιστου) μοντέλου συγκριτικά με το πλήρες.

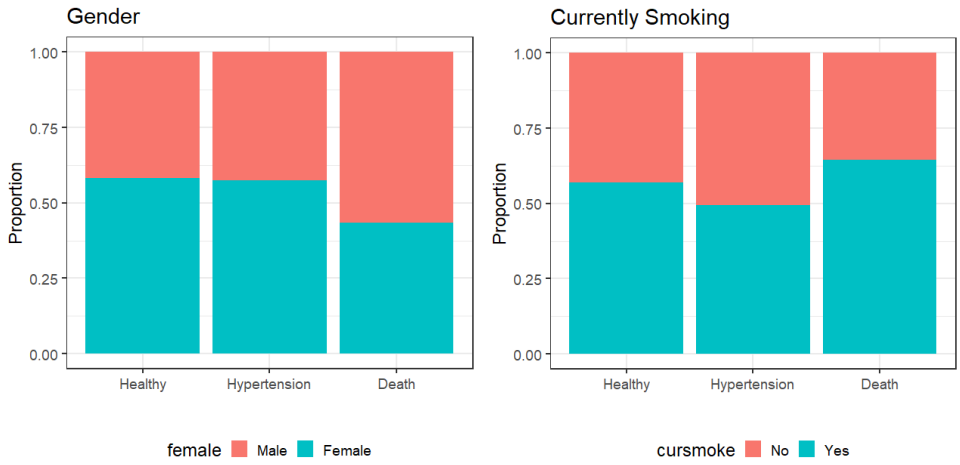
Τα ιστογράμματα του Γραφήματος 6, παρουσιάζουν την κατανομή των τιμών των ανεξάρτητων (επεξηγηματικών) μεταβλητών του εξεταζόμενου προβλήματος ταξινόμησης για κάθε επίπεδο της μεταβλητής απόκρισης. Από το ιστογράμματα, παρατηρούμε ότι ο δείκτης μάζας σώματος και η αλλαγή στην αρτηριακή πίεση παρουσιάζονται να έχουν αυξημένες τιμές τόσο στους υπερτασικούς όσο και στους ασθενείς που απεβίωσαν. Η πλειονότητα των

παρουσιαζόμενων κατανομών φαίνεται να απέχει από την κανονική κατανομή, καθώς σύμφωνα με τα διαγράμματα παρατηρούμε θετική ή αρνητική λοξότητα, όμως το χαρακτηριστικό αυτό δεν επηρεάζει τις προϋποθέσεις εφαρμογής του μοντέλου λογιστικής παλινδρόμησης. Ωστόσο, εκτός από την περίπτωση του αριθμού τσιγάρων που καταναλώνονται ημερησίως από τους ασθενείς, οι κατανομές των μεταβλητών είναι μονοκόρυφες. Από την άλλη πλευρά, η κατανομή του ημερήσιου αριθμού τσιγάρων είναι πολυκόρυφη.



Γράφημα 6 -. Ιστογράμματα των συνεχών ανεξάρτητων μεταβλητών του προβλήματος ταξινόμησης

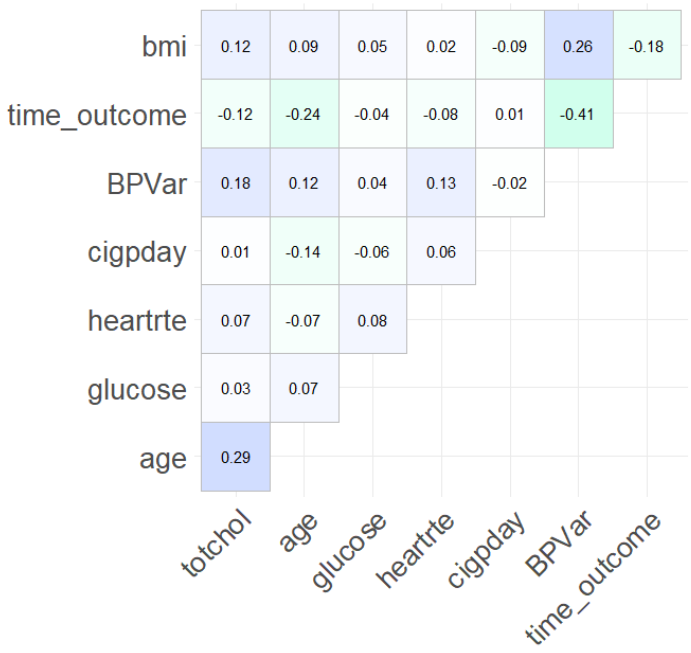
Η πρώτη και ψηλότερη κορυφή αντιστοιχεί στα άτομα που δεν είναι καπνιστές, ενώ από τα άτομα που καπνίζουν, η πλειονότητα αυτών φαίνεται να καταναλώνει περίπου 20 τσιγάρα την ημέρα. Τέλος, συνολικά για το δείγμα, η διάμεση ηλικία των ασθενών είναι τα 40 χρόνια, τα μέσα επίπεδα γλυκόζης προσεγγίζουν τις 80 μονάδες, ενώ η διάμεση τιμή επιπέδων χοληστερόλης των συμμετεχόντων στην έρευνα είναι 227 μονάδες. Στη συνέχεια, παρουσιάζουμε τα ραβδογράμματα των κατηγορικών μεταβλητών της ανάλυσης (Γράφημα 7). Σύμφωνα με το ραβδόγραμμα που αναφέρεται στο φύλο των συμμετεχόντων, το δείγμα της έρευνας αποτελείται από 1019 άντρες και 1297 γυναίκες, με τους άνδρες να έχουν υψηλότερη αναλογία στην κατηγορία των αποβιωσάντων. Επιπλέον, κατά τη διάρκεια διεξαγωγής της έρευνας, 1238 άτομα ήταν καπνιστές, ενώ 1078 όχι, και σύμφωνα με το Γράφημα 7, οι καπνιστές εμφανίζουν το μεγαλύτερο ποσοστό στην κατηγορία των αποβιωσάντων. Τέλος, αναφορικά με το σύνολο του δείγματος, 702 άτομα που έλαβαν μέρος στην έρευνα, δεν παρουσίασαν κάποιο σημαντικό πρόβλημα υγείας μέχρι το τέλος αυτής, 1346 παρουσίασαν υπέρταση και 268 απεβίωσαν.



Γράφημα 7 - Ραβδογράμματα των κατηγορικών μεταβλητών του προβλήματος παλινδρόμησης με βάση την κατάσταση της υγείας

Προτού περάσουμε στην προσαρμογή του μοντέλου παλινδρόμησης, θα εξετάσουμε τις ανά δύο γραμμικές συσχετίσεις μεταξύ των ανεξάρτητων μεταβλητών συνεχούς κλίμακας. Στο Γράφημα 8 παρουσιάζονται οι τιμές αυτών των συσχετίσεων. Παρατηρούμε ότι η μεταβλητή χρόνος εμφάνισης του τελικού αποτελέσματος, έχει μέτρια αρνητική συσχέτιση με τις τιμές της αρτηριακής πίεσης ($r = -0.41$) και παρουσιάζει ασθενή αρνητική συσχέτιση και με τη μεταβλητή ηλικία ($r = -0.24$). Επίσης, η μεταβλητή αυτή είναι σταθερή για όλους τους υγιείς του δείγματος καθώς συνέχισαν ως το τέλος της μελέτης, επομένως δεν μπορεί να προσφέρει διαχωρισμό μεταξύ υγιών και ασθενών ατόμων, καθώς δεν είναι εφικτό να γνωρίζουμε το χρόνο επιβίωσης, θανάτου ή εμφάνισης υπέρτασης πριν από την μέτρηση της τελικής έκβασης κάθε συμμετέχοντα. Έτσι η μεταβλητή χρόνος εμφάνισης του αποτελέσματος θα πρέπει να αφαιρεθεί

από την προσαρμογή του μοντέλου λογιστικής παλινδρόμησης. Την χαμηλή συσχέτιση των μεταβλητών επιβεβαίωσε και ο δείκτης VIF, με μέγιστη τιμή ίση με 2.6 για την μεταβλητή του αριθμού των τσιγάρων που καταναλώνει κάθε συμμετέχων.



Γράφημα 8 - Γραμμικές συσχετίσεις μεταξύ των επεξηγηματικών μεταβλητών

Αρχικά εφαρμόσθηκε μοντέλο πολυωνυμικής παλινδρόμησης για να εξεταστεί η σχέση των προβλεπτικών παραγόντων με την κατάσταση της υγείας και τα αποτελέσματα παρουσιάζονται στον παρακάτω Πίνακα 7. Ας θεωρήσουμε τις πιθανότητες π_{ij}, με $i = 1,2,\dots,2316$ που δηλώνει το συμμετέχοντα και $j = 1,2$, όπου με $j = 1$ δηλώνουμε τη υπέρταση και με $j = 2$ δηλώνουμε το ενδεχόμενο του θανάτου. Το επίπεδο αναφοράς ορίζεται να είναι η μη παθολογική – υγιής κατάσταση με το σύμβολο J. Το μοντέλο στην γενική μορφή του έχει εξίσωση

$$\log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = \alpha_j + x_i' \cdot \beta_j$$

όπου α_j είναι διάνυσμα σταθερών όρων και β_j είναι το διάνυσμα των προς εκτίμηση συντελεστών των ανεξάρτητων μεταβλητών και παρουσίασε τιμή AIC ίση με 3829.532 και Deviance ίσο με 3789.532. Παρατηρήθηκε πως οι υπερτασικοί ασθενείς είχαν αυξημένο ΔΜΣ, αρτηριακή πίεση,

ηλικία, και κάπνιζαν περισσότερο σε σύγκριση με τους υγιείς. Οι γυναίκες είχαν 1.54 φορές αυξημένο κίνδυνο εμφάνισης υπέρτασης σε σύγκριση με τους άνδρες. Οι ασθενείς που απεβίωσαν είχαν υψηλότερη αρτηριακή πίεση, είχαν μεγαλύτερη ηλικία και κάπνιζαν περισσότερο σε σύγκριση με τους υγιείς. Συνολικά, το μοντέλο λογιστική παλινδρόμησης είχε δείκτη προσαρμογής R^2 ίσο με 0.117 και ταξινόμησε τους συμμετέχοντες με ακρίβεια 63.43% παρουσιάζοντας υψηλή ευαισθησία στην υπέρταση και ειδικότητα στους υγιείς. Ωστόσο το μοντέλο είχε χαμηλή ευαισθησία στο να προβλέψει το θάνατο (Πίνακες 8 και 9).

Πίνακας 7. Αποτελέσματα πολλαπλής πολυωνυμικής λογιστικής παλινδρόμησης

Παράγοντας	Εκτίμηση	Στατιστικό	*p*	*OR*
Υπέρταση				
Σταθερός όρος	-3.454	-4.830	0.000	
Γλυκόζη	0.003	0.702	0.483	
Φύλο (Άνδρες)	**0.432**	**3.752**	**0.000**	**1.54**
Κάπνισμα	-0.314	-1.940	0.052	
Δείκτης μάζας σώματος	**0.062**	**3.737**	**0.000**	**1.06**
Αρτηριακή πίεση	**0.068**	**13.636**	**0.000**	**1.07**
Καρδιακός ρυθμός	0.002	0.363	0.717	
Ηλικία	**0.055**	**7.605**	**0.000**	**1.05**
Χοληστερόλη	-0.001	-1.132	0.258	
Αριθμός τσιγάρων	**0.024**	**3.218**	**0.001**	**1.02**
Θάνατος				
Σταθερός όρος	-7.502	-6.986	0.000	
Γλυκόζη	0.001	0.196	0.845	
Φύλο (Άνδρες)	-0.047	-0.276	0.782	
Κάπνισμα	0.175	0.734	0.463	
Δείκτης μάζας σώματος	0.016	0.644	0.519	
Αρτηριακή πίεση	**0.031**	**4.393**	**0.000**	**1.03**
Καρδιακός ρυθμός	0.010	1.394	0.163	
Ηλικία	**0.116**	**11.473**	**0.000**	**1.12**
Χοληστερόλη	-0.002	-1.033	0.302	
Αριθμός τσιγάρων	**0.031**	**3.170**	**0.002**	**1.03**

Πίνακας 8. Πίνακας συνάφειας μεταξύ προβλέψεων του μοντέλου λογιστικής παλινδρόμησης και των πραγματικών δεδομένων				
		Πραγματική τιμή		
		υγιείς	υπέρταση	θάνατος
Πρόβλεψη	υγιείς	294	178	45
	υπέρταση	402	1164	212
	θάνατος	6	4	11

Πίνακας 9. Ευαισθησία και Ειδικότητα του μοντέλου λογιστικής παλινδρόμησης.			
	Υγιείς	Υπερτασικοί	Αποβιώσαντες
Ευαισθησία	41.88%	86.48%	04.10%
Ειδικότητα	86.18%	36.70%	99.51%

Στη συνέχεια, εξετάστηκε αν το εφαρμοσμένο μοντέλο μπορεί να έχει ικανοποιητική προβλεπτική ικανότητα σε άγνωστα για αυτό δεδομένα. Για το σκοπό αυτό, χρησιμοποιήθηκε η μέθοδος της ενδοεπικύρωσης (cross-validation), όπου το δείγμα χωρίστηκε σε πέντε υποσύνολα, και το μοντέλο πολυωνυμικής παλινδρόμησης εφαρμόσθηκε κάθε φορά στα τέσσερα από τα πέντε, αφήνοντας το πέμπτο υποσύνολο ως σύνολο ελέγχου με 463 παρατηρήσεις. Η διαδικασία αυτή επαναλήφθηκε πέντε φορές με τυχαίο διαχωρισμό των πέντε υποσυνόλων, ώστε να μειωθεί η μεροληψία της επιλογής του δείγματος. Τα αποτελέσματα παρουσιάζονται στον Πίνακα Χ, όπου φαίνεται πως το μοντέλο στα σύνολα ελέγχου παρουσίασε ικανοποιητική προσαρμογή στα άγνωστα δεδομένα (Ακρίβεια = 62.97%) με το αντίστοιχο μοντέλο που προσαρμόστηκε σε ολόκληρο το σύνολο των δεδομένων.

Πίνακας 10. Πίνακας συνάφειας και μέση εκτιμώμενη ακρίβεια του μοντέλου πολυωνυμικής λογιστικής παλινδρόμησης με χρήση cross-validation.

Σύνολο ελέγχου		**Πραγματική τιμή**		
		υγιείς	υπέρταση	θάνατος
Πρόβλεψη	υγιείς	87	55	13
	υπέρταση	121	346	64
	θάνατος	2	2	3
Ακρίβεια: 62.97%				

Στον Πίνακα 11 παρουσιάζονται οι τιμές του AIC, του BIC και του Deviance, για διαφορετικά μοντέλα που έχουν διαφορετικές ανεξάρτητες μεταβλητές. Παρατηρούμε πως ο συνδυασμός των μεταβλητών που ελαχιστοποιούν το κριτήριο AIC είναι στην τελευταία γραμμή και το συγκεκριμένο μοντέλο, με τις ανεξάρτητες μεταβλητές Φύλο, Κάπνισμα, BMI, Πίεση, Ηλικία, και Πλήθος τσιγάρων, αντιστοιχεί επίσης και σε μικρότερο BIC. Παρά την καλύτερη προσαρμογή του μοντέλου βάσει του AIC, δεν παρατηρήθηκε στατιστικά σημαντική διαφορά από το πλήρες

μοντέλο, δηλαδή εκείνο με όλες τις μεταβλητές ως ανεξάρτητες, $\chi^2 = 4.05$, $p = .669$, επομένως το μειωμένο μοντέλο παρέχει την ίδια προβλεπτική ισχύ χρησιμοποιώντας λιγότερες μεταβλητές.

Πίνακας 10. Αποτελέσματα πολλαπλής πολυωνυμικής λογιστικής παλινδρόμησης				
Ανεξάρτητες μεταβλητές	**AIC**	**BIC**	**Deviance**	**R^2**
Γλυκόζη	4297.02	4320.01	4289.02	0.001
Φύλο	4280.98	4303.97	4272.98	0.005
Κάπνισμα	4275.52	4298.51	4267.52	0.006
Δείκτης μάζας σώματος	4237.37	4260.36	4229.37	0.015
Αρτηριακή πίεση	4294.59	4317.58	4286.59	0.067
Καρδιακός ρυθμός	4014.15	4037.14	4006.15	0.001
Ηλικία	4130.84	4153.83	4122.84	0.040
Χοληστερόλη	4277.06	4300.05	4269.06	0.006
Αριθμός τσιγάρων	4281.79	4304.78	4273.79	0.004
Γλυκόζη, Φύλο, Κάπνισμα, Δείκτης μάζας σώματος, Αρτηριακή πίεση, Καρδιακός ρυθμός, Ηλικία, Χοληστερόλη, Αριθμός τσιγάρων	3829.53	3944.48	3789.53	0.117
Φύλο, Κάπνισμα, BMI, Πίεση, Ηλικία, Πλήθος τσιγάρων	3821.59	3902.05	3793.59	0.116

Συμπεράσματα

Στην παρούσα εργασία, ερευνήθηκαν οι τρόποι ανάλυσης κατηγορικών τυχαίων μεταβλητών με δυο ή περισσότερες κατηγορίες. Αρχικά δόθηκαν οι απαραίτητοι ορισμοί από τη βιβλιογραφία και στην συνέχεια ορίστηκαν τα κύρια χαρακτηριστικά της ανάλυσης δυο κατηγορικών μεταβλητών με δυο κατηγορίες η κάθε μία. Το βασικό εργαλείο της ανάλυσης για την εύρεση συσχέτισης μεταξύ των μεταβλητών αυτών είναι ο πίνακας συνάφειας, ο οποίος παρουσιάζει τις διαφορετικές αναλογίες και ποσοστά που προκύπτουν για τα διαφορετικά επίπεδα των μεταβλητών. Σαν επέκταση, οι κατηγορικοί πίνακες συνάφειας με πολλαπλές κατηγορίες επιτρέπουν να αναλύσουμε τη σχέση μεταξύ τριών ή περισσότερων κατηγορικών μεταβλητών. Αυτοί οι πίνακες είναι παρόμοιοι με τους πίνακες συνάφειας με δύο κατηγορίες, αλλά έχουν πρόσθετες γραμμές και στήλες για να αντιπροσωπεύουν τις πρόσθετες μεταβλητές.

Για την δημιουργία ενός πίνακα συνάφειας με πολλές κατηγορίες, προσδιορίζουμε τις τρεις ή περισσότερες μεταβλητές προς ανάλυση, οι οποίες μπορεί να είναι οποιαδήποτε χαρακτηριστικά ή ιδιότητες που μπορούν να χωριστούν σε κατηγορίες ή ομάδες. Αφού προσδιοριστούν οι μεταβλητές, δημιουργούμε έναν πίνακα με γραμμές και στήλες που αντιπροσωπεύουν καθεμία από τις μεταβλητές. Τα κελιά του πίνακα περιέχουν τις μετρήσεις ή τα ποσοστά των ατόμων που εμπίπτουν σε κάθε συνδυασμό κατηγοριών. Για παράδειγμα, ένας κατηγορικός πίνακας συνάφειας

με τρεις κατηγορίες μπορεί να δείξει ότι το 60% των ανδρών ερωτηθέντων μεταξύ 18 και 34 ετών είναι συνδεδεμένοι με την ομάδα Α, ενώ το 40% είναι συνδεδεμένοι με την ομάδα Β. Η ανάλυση κατηγοριών δεδομένων με πολλαπλές κατηγορίες περιλαμβάνει τη χρήση στατιστικών τεχνικών για την ερμηνεία και ανάλυση των δεδομένων από τις πληροφορίες που περιέχονται στον πίνακα συνάφειας. Αυτές οι τεχνικές μπορούν να περιλαμβάνουν ελέγχους ανεξαρτησίας, που αξιολογούν εάν υπάρχει σημαντική σχέση μεταξύ των μεταβλητών, όπως και τον έλεγχο χ^2, που συγκρίνει διαφορές στις αναλογίες μεταξύ των κατηγοριών. Επίσης, πέρα από τους ελέγχους υποθέσεων και τη στατιστική σημαντικότητα, παρουσιάστηκαν και μέτρα συσχέτισης μεταξύ δυο κατηγορικών μεταβλητών, τα οποία χαρακτηρίζουν την ισχύ της σχέσης, κάτι που είναι πολύ σημαντικό σε εφαρμογές με πραγματικά δεδομένα. Συνολικά, και λόγω της ευελιξίας τους, οι κατηγορικοί πίνακες συνάφειας με πολλαπλές κατηγορίες αποδεικνύονται χρήσιμα εργαλεία για την κατανόηση και την ανάλυση πολύπλοκων σχέσεων μεταξύ κατηγορικών μεταβλητών και για τη λήψη τεκμηριωμένων αποφάσεων με βάση αυτά τα δεδομένα.

Στην συνέχεια εισάχθηκαν συγκεκριμένα μοντέλα που χρησιμοποιούνται και χαρακτηρίζουν και ερμηνεύουν την σχέση μεταξύ δυο ή περισσοτέρων μεταβλητών. Για παράδειγμα, τα λογαριθμογραμμικά μοντέλα, τα οποία χρησιμοποιούν το λογάριθμο των σχετικών πιθανοτήτων, δίνουν την δυνατότητα να βρεθεί η επίδραση πολλών κατηγορικών ή συνεχών μεταβλητών πάνω σε μια εξαρτημένη κατηγορική μεταβλητή, χρησιμοποιώντας γραμμικές ή σχέσεις ανώτερης τάξης. Για να αναπτύξουμε ένα λογαριθμογραμμικό μοντέλο, πρέπει πρώτα να καθορίσουμε τη μορφή του μοντέλου και τις μεταβλητές ή επιδράσεις που θα συμπεριληφθούν. Η πιο κοινή μορφή ενός λογαριθμογραμμικού μοντέλου είναι το κορεσμένο μοντέλο, το οποίο περιλαμβάνει όλους τους πιθανούς συνδυασμούς των μεταβλητών πρόβλεψης. Για παράδειγμα, εάν έχουμε δύο μεταβλητές πρόβλεψης, την Α και τη Β, το κορεσμένο μοντέλο θα περιλαμβάνει τις κύριες επιδράσεις των Α και Β, καθώς και την αλληλεπίδραση μεταξύ των Α και Β. Αφού καθοριστεί η μορφή του μοντέλου, μπορούν να υπολογιστούν οι λογάριθμοι των σχετικών πιθανοτήτων για κάθε μεταβλητή πρόβλεψης. Ο λόγος log-odds για μια μεταβλητή πρόβλεψης μετρά τη σχετική πιθανότητα να συμβεί ένα γεγονός σε μια κατηγορία σε σύγκριση με μια άλλη. Για παράδειγμα, ο λόγος log-odds για τη μεταβλητή Α μπορεί να μετρήσει τη σχετική πιθανότητα να συμβεί ένα ενδεχόμενο στην κατηγορία Α1 σε σύγκριση με την κατηγορία Α2. Οι λόγοι σχετικών πιθανοτήτων προσδίδουν μια ικανοποιητική ερμηνεία της σχέσης κατηγορικών μεταβλητών και βρίσκουν εφαρμογή σε ποικίλους τομείς. Για τον υπολογισμό των log-odds, συνήθως χρησιμοποιείται η εκτίμηση μέγιστης πιθανοφάνειας (MLE), η οποία είναι μια στατιστική τεχνική που βρίσκει τις τιμές των παραμέτρων που μεγιστοποιούν την πιθανότητα των δεδομένων που δίνονται στο μοντέλο. Οι λόγοι log-odds χρησιμοποιούνται στη συνέχεια για την πρόβλεψη των πιθανοτήτων της μεταβλητής απόκρισης που λαμβάνει κάθε μία από τις Κ πιθανές κατηγορίες.

Τέλος, παρουσιάστηκε μια εφαρμογή ανάλυσης κατηγορικών δεδομένων στο σύνολο δεδομένων Framingham, από μια ανοικτή μακροχρόνια μελέτη, η οποία ξεκίνησε το 1948 και έχει σκοπό να ερευνήσει τους παράγοντες που οδηγούν σε καρδιοπάθειες. Με βάση την υπέρταση, ως μεταβλητή απόκρισης, χρησιμοποιήθηκαν μοντέλα πολλαπλής πολυωνυμικής παλινδρόμησης, ώστε να εντοπιστούν οι παράγοντες που φαίνεται να έχουν μεγαλύτερη επίδραση στην καρδιοπάθεια. Το τελικό μοντέλο που επιλέχθηκε, είχε ικανοποιητική ακρίβεια πρόβλεψης των

υπερτασικών ασθενών από τους υγιείς και καθόρισε τους προβλεπτικούς παράγοντες. Οι μέθοδοι ανάλυσης που παρουσιάστηκαν, όπως η πολυωνυμική λογιστική παλινδρόμηση, αποτελούν ένα ισχυρό εργαλείο για την ανάλυση κατηγορικών δεδομένων και την κατανόηση των σχέσεων μεταξύ πολλαπλών κατηγορικών μεταβλητών και δικαίως χρησιμοποιούνται ευρέως σε τομείς όπως η ψυχολογία, η κοινωνιολογία και η ιατρική, για να οδηγήσουν σε προβλέψεις σχετικά με τα μελλοντικά αποτελέσματα και να ενημερώσουν τη λήψη αποφάσεων με βάση τα πραγματικά δεδομένα.

Βιβλιογραφία

Agresti A. (2013). *Categorical Data Analysis*, Wiley Series in Probability and Statistics, 3rd edition, New Jersey.

Altonji, J. G., T. E. Elder, and C. R. Taber (2005): "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools," *Journal of Political Economy*, 113, 151–184.

Altonji, J. G. and R. L. Matzkin (2005): "Cross section and panel data estimators for nonseparable models with endogenous regressors," *Econometrica*, 73, 1053–1102.

Athey, S. and G. W. Imbens (2006): "Identification and inference in nonlinear difference-indifferences models," *Econometrica*, 74, 431–497.

Birch, M. W. (1964). The detection of partial association, I: the 2 x 2 case. Journal of the Royal Statistical Society, B 26, 313—324.

Birch, M. W. (1965). The detection of partial association, Il: the general case. Journal of the Royal Statistical Society, B 27, 111—124.

Breslow, N.E. (1981). Odds ratio estimators when data are sparse. Biometrika 68, 73-84.

Breslow, N.E., and Day, N.E. (1980). Statistical Methods in Cancer Research I: The Analysis of Case-Control Studies. International Agency for Research on Cancer, Lyon.

Breslow, N.E. and Liang, K. Y. (1982). The variance of the Mantel-Haenszel estimator. Biometrics 38, 943—952.

Berry, K. J., & Mielke, P. W. (1992). A family of multivariate measures of association for nominal independent variables. *Educational and Psychological Measurement, 52*(1), 41-55.

Canay, I. A. and A. M. Shaikh (2016): "Practical and theoretical advances for inference in partially identified models," *Working paper*.

Chamberlain, G. and E. E. Leamer (1976): "Matrix weighted averages and posterior bounds," *Journal of the Royal Statistical Society, Series B*, 73–84.

Chernozhukov, V. and C. Hansen (2005): "An IV model of quantile treatment effects," *Econometrica*, 73, 245–261.

Chesher, A. (2003): “Identification in nonseparable models,” *Econometrica*, 71, 1405–1441.

Chesher, A. and A. Rosen (2015): “Generalized instrumental variable models,” *Working paper*.

Cochran, William G. (1952). "The Chi-square Test of Goodness of Fit". The Annals of Mathematical Statistics. 23 (3): 315–345. doi:10.1214/aoms/1177729380. JSTOR 2236678.

Cohen, J., & Nee, J. C. (1984). Estimators for two measures of association for set correlation. *Educational and Psychological Measurement, 44*(4), 907-917.

Conley, T. G., C. B. Hansen, and P. E. Rossi (2012): “Plausibly exogenous,” *Review of Economics and Statistics*, 94, 260–272.

Cross, P. J. and C. F. Manski (2002): “Regressions, short and long,” *Econometrica*, 70, 357–368.

De Longerill, M., Salen, P., Martin, J., Monjaud, I., Boucher, P., Mamelle, N. (1998). Mediterranean Dietary pattern in a Randomized Trial. *Archives of Internal Medicine*, 158, 1181-1187.

Gibbons, J. A. (1985). Shrinkage formulas for two nominal level measures of association. *Educational and Psychological Measurement, 45*(3), 551-566.

Gibbons, J. D. (1993). *Nonparametric measures of association*. Thousand Oaks, CA: Sage Publications.

Cochran, W.G. (1954). Some methods for strengthening the common X^2 test. Biometrics 10, 417—451.

Connett, J. , Ejigou, A. , McHugh, R. , and Breslow, N. (1982). The precision of the Mantel-Haenszel estimator in case-control studies with multiple matching. American Journal of Epidemiology 116, 875—877.

Donald, A. , and Donner, A. (1987). Adjustments to the Mantel-Haenszel chisquare statistic and odds ratio variance estimator when the data are clustered. Statistics in Medicine 6, 491—499.

Engel, J. (1988). "Polytomous logistic regression". Statistica Neerlandica. 42 (4): 233–252. doi:10.1111/j.1467-9574.1988.tb01238.x.

Findley, D. F. (1991). "Counterexamples to parsimony and BIC". Annals of the Institute of Statistical Mathematics. 43 (3): 505–514. doi:10.1007/BF00053369. S2CID 58910242.

Jobson J.D. (1992). *Applied multivariate data analysis.* Springer Verlag, Volume 2, New York.

Kateri, M. (2014). Contingency Table Analysis, Springer.

Keown, L. L., & Hakstian, A. R. (1973). Measures of association for the component analysis of Likert scale data. *Journal of Experimental Education, 41*(3), 22-27.

Kim, S., & Olejnik, S. (2005). Bias and precision of measures of association for a fixed-effect multivariate analysis of variance model. *Multivariate Behavioral Research, 40*(4), 401-421.

Kline, P. and A. Santos (2013): "Sensitivity to missing data assumptions: Theory and an evaluation of the U.S. wage structure," *Quantitative Economics*, 4, 231–267.

Koenker, R. and G. Bassett (1978): "Regression quantiles," *Econometrica*, 33–50.

Leamer, E. E. (1978): *Specification searches: Ad hoc inference with nonexperimental data*, vol. 53, John Wiley & Sons Incorporated.

Kraemer, H. C. (2000). Measures of association. In *Encyclopedia of psychology* (Vol. 5, pp. 135-139). Washington, DC: American Psychological Association.

Krieger, A. M., & Green, P. E. (1993). Generalized measures of association for ranked data with an application to prediction accuracy. *Journal of Classification, 10*(1), 93-114.

Liebetrau, A. M. (1983). *Measures of association*. Newbury Park, CA: Sage Publications.

Siegel, S. (1956). *Nonparametric Statistics For The Behavioral Sciences*. New York: McGraw-Hill.

Stevens, J. P. (1972). Global measures of association in multivariate analysis of variance. *Multivariate Behavioral Research, 7*(3), 373-378.

Stoica, P.; Selen, Y. (2004), "Model-order selection: a review of information criterion rules", IEEE Signal Processing Magazine (July): 36–47, doi:10.1109/MSP.2004.1311138, S2CID 17338979

Wilcox, R. R. (2007). Local measures of association: Estimating the derivative of the regression line. *British Journal of Mathematical and Statistical Psychology, 60*, 107-117.

Παράρτημα 1

Κώδικας εφαρμογής

```
library(LocalControl);library(ggplot2);
library(tidyverse)
library(ggcorrplot)
library(lattice)
library(psych)
library(DataExplorer)
library(car)
library(caret)
library(scales)
library(caTools)
library(modelr)
library(broom)
library(cowplot)
library(pROC)
library(e1071)
library(MASS)
data(framingham)
df=framingham


# ----  Outcome
ggplot(df, aes(outcome)) +
 geom_bar(stat = "count") + scale_fill_manual(values=c('grey70', 'grey20')) +   theme_bw(base_size = 18) +
 theme(legend.position="bottom")+
 labs(title = "Outcome")


# ----  Gender   ------
```

```
 ggplot(df, aes(female, fill = `bmi`)) +
 geom_bar(stat = "count", position = "dodge") +
 scale_fill_manual(values=c('grey70', 'grey20')) +
 labs(title = "Gender", x = "") +  theme_bw(base_size = 18) +
 theme(legend.position="bottom")
# ----  Currently Smoking   ------
ggplot(df, aes(cursmoke, fill = `bmi`)) +
 geom_bar(stat = "count", position = "dodge") +
 scale_fill_manual(values=c('grey70', 'grey20')) +
 labs(title = "Currently Smoking", x = "") +  theme_bw(base_size = 18) +
 theme(legend.position="bottom")
# ----- Age -----
ggplot(df,aes(age, , fill='female')) +
geom_density(lwd = 3, show.legend = T, alpha = 0.7) +
labs(title = "Age") +
scale_fill_manual(values=c('grey70', 'grey20')) +
scale_color_manual(values=c('grey50', 'black')) +
theme_bw(base_size = 18) + theme(legend.position="bottom")
# ----- BPvar -----
ggplot(df,aes(BPVar, , fill='female')) +
 geom_density(lwd = 3, show.legend = T, alpha = 0.7) +
 labs(title = "Systolic Pressure") +
 scale_fill_manual(values=c('grey70', 'grey20')) +
 scale_color_manual(values=c('grey50', 'black')) +
 theme_bw(base_size = 18) + theme(legend.position="bottom")
# ----- Heart Rate -----
ggplot(df,aes(heartrte, , fill='female')) +
 geom_density(lwd = 3, show.legend = T, alpha = 0.7) +
 labs(title = "Heart Rate") +
 scale_fill_manual(values=c('grey70', 'grey20')) +
```

```
  scale_color_manual(values=c('grey50', 'black')) +
  theme_bw(base_size = 18) + theme(legend.position="bottom")
# ----- Cholesterol -----
ggplot(df,aes(totchol, , fill='female')) +
  geom_density(lwd = 3, show.legend = T, alpha = 0.7) +
  labs(title = "Cholesterol Level") +
  scale_fill_manual(values=c('grey70', 'grey20')) +
  scale_color_manual(values=c('grey50', 'black')) +
  theme_bw(base_size = 18) + theme(legend.position="bottom")
# ---- Cigarettes per day ----
ggplot(df,aes(cigpday, color = 'female', fill='female')) +
geom_density(lwd = 3, show.legend = T, alpha = 0.7) +
labs(title = "Cigarettes per day", x = "Cigarettes per day") +
scale_fill_manual(values=c('grey70', 'grey20')) +
scale_color_manual(values=c('grey50', 'black')) +
theme_bw(base_size = 18) + theme(legend.position="bottom")
# ---- Cigarettes per day ----
ggplot(df,aes(BPVar, color = 'female', fill='female')) +
geom_density(lwd = 3, show.legend = T, alpha = 0.7) +
labs(title = "Average Blood Pressure", x = "Average Blood Pressure") +
scale_fill_manual(values=c('grey70', 'grey20')) +
scale_color_manual(values=c('grey50', 'black')) +
theme_bw(base_size = 18) + theme(legend.position="bottom")
# ---- Glucose levels -----
ggplot(df,aes(glucose, color = 'female', fill='female')) +
  geom_density(lwd = 3, show.legend = T, alpha = 0.7) +
  labs(title = "Glucose levels", x = "Glucose") +
  scale_fill_manual(values=c('grey70', 'grey20')) +
  scale_color_manual(values=c('grey50', 'black')) +
  theme_bw(base_size = 18) + theme(legend.position="bottom")
```

```
#  ----- Correlations ----
df2 <- df %>% select("totchol",
                     "age",
                     "glucose",
                     "heartrte",
                     "cigpday",
                     "BPVar",
                     "time_outcome",
                     'bmi')


corr <- round(cor(df2, use="complete.obs"), 2)
options(repr.plot.width=10, repr.plot.height=8)
ggcorrplot(corr, lab = TRUE, colors = c("aquamarine", "white", "dodgerblue"),
           show.legend = F, outline.color = "gray", type = "upper", #hc.order = T,
           tl.cex = 20, lab_size = 10, sig.level = .2) +
  labs(fill = "Correlation")
# ---- Setting Factors
df$outcome<-factor(df$outcome, c(0,1,2))
df$female<-factor(df$female, c(0,1))
df$cursmoke<-factor(df$cursmoke, c(0,1))
# ---- Splitting
df1 <- df[1:1621,];df2<-df[1622:2316,]
# ---- Classification (3 classes) -----
library(nnet)
glm.fit=multinom(outcome~glucose+female+cursmoke+bmi+BPVar+heartrte+age+totchol+cigpday, data=df1)
summary(glm.fit)
stepAIC(glm.fit)
glm.fit2=multinom(outcome~female+cursmoke+bmi+BPVar+age+totchol+cigpday, data=df1)
summary(glm.fit2)
```

```
#Prediction
probs<-predict(glm.fit, df2, "probs")
predictions <- c();
for (i in 1:length(probs[,1])){
  predictions[i] <- which.max(probs[i,])
}
observed.classes<-df2$outcome
accuracy=length(which(predictions==observed.classes))/length(predictions);accuracy
# ---- Classification (2 classes) ------
# Merging classes 1 and 2
for (i in 1:length(df$outcome)){
  if (df$outcome[i]==2){
    df$outcome[i]=1
  }
}
df$outcome
df$outcome <- factor(df$outcome, c(0,1))

df1 <- df[1:1621,];df2<-df[1622:2316,]
# ----- Fit the full model
model<-glm(outcome~glucose+female+cursmoke+bmi+BPVar+heartrte+age+totchol+cigpday,data=df,family = binomial)
summary(model)
stepAIC(model,trace=FALSE)
# ----- Fit the best model
model2<-glm(outcome~female+bmi+BPVar+age+cigpday,data=df1,family = binomial)
summary(model2)
# ----- Best Model -----
probs<-predict(model2, df2)
predictions1 <- c();
for (i in 1:length(probs)){
```

```
 if(probs[i]>0){
   predictions1[i] <- 1
 }else{
   predictions1[i] <- 0
 }
}
observed.classes<-df2$outcome
which(predictions1==observed.classes);length(which(predictions1==observed.classes))/length(probs)
# ----- Full Model -----
probs<-predict(model, df2)
predictions2 <- c();
for (i in 1:length(probs)){
 if(probs[i]>0){
   predictions2[i] <- 1
 }else{
   predictions2[i] <- 0
 }
}
observed.classes<-df2$outcome
which(predictions2==observed.classes);length(which(predictions2==observed.classes))/length(probs)
# ----- Confusion Matrices -----
expected_value <- factor(observed.classes)
predicted_value <- factor(predictions)
#Creating confusion matrix
example <- confusionMatrix(data=predicted_value, reference = expected_value);example
# ----- Plotting the Confusion Matrix ------
expected_value <- factor(c(0, 0, 1, 1))
predicted_value <- factor(c(0, 1, 0, 1))
Y     <- c(56, 60, 148, 431)
confM <- data.frame(expected_value, predicted_value, Y)
```

```
ggplot(data = confM, mapping = aes(x = expected_value, y = predicted_value)) +
  geom_tile(aes(fill = Y), colour = "black") +
  geom_text(aes(label = sprintf("%1.0f", Y)), vjust = 1) +
  scale_fill_gradient(low = "lightblue", high = "dodgerblue") +
  theme_bw() + theme(legend.position = "none")
train_control <- trainControl(method='repeatedcv',number=5,repeats = 5,verboseIter = TRUE)
model_cv <-
caret::train(outcome~glucose+female+cursmoke+bmi+BPVar+heartrte+age+totchol+cigpday,data=df,trControl=train_control,
                    method='multinom')
```