

1. Διδιάστατοι πίνακες συνάφειας χωρίς τη χρήση γενικευμένων γραμμικών μοντέλων

Έστω X, Y δύο κατηγορικές μεταβλητές απόκρισης με I και J στάθμες αντίστοιχα. Οι αποκρίσεις (X, Y) ενός τυχαία επιλεγμένου ατόμου από έναν πληθυσμό θεωρούνται ως τυχαίες μεταβλητές που ακολουθούν κάποια κατανομή. Συμβολίζουμε με π_{ij} την πιθανότητα

$$\pi_{ij} = P((X, Y) = (i, j)) = P(X = i, Y = j), \quad i = 1, 2, \dots, I, \quad j = 1, 2, \dots, J$$

Επίσης συμβολίζουμε με $\pi_{i\bullet} = P(X = i)$, $\pi_{\bullet j} = P(Y = j)$ τις περιθώριες κατανομές των X, Y . Είναι προφανές ότι

$$\pi_{i\bullet} = \sum_{j=1}^J \pi_{ij}, \quad i = 1, 2, \dots, I \quad \text{και} \quad \pi_{\bullet j} = \sum_{i=1}^I \pi_{ij}, \quad j = 1, 2, \dots, J.$$

Μπορούμε να παραστήσουμε τη διδιάστατη κατανομή των (X, Y) χρησιμοποιώντας έναν πίνακα με I γραμμές και J στήλες:

		Y				
		1	2	...	J	Total
X	1	π_{11}	π_{12}	...	π_{1J}	$\pi_{1\bullet}$
	2	π_{21}	π_{22}	...	π_{2J}	$\pi_{2\bullet}$
	...	...	...	...	...	...
	I	π_{I1}	π_{I2}	...	π_{IJ}	$\pi_{I\bullet}$
Total		$\pi_{\bullet 1}$	$\pi_{\bullet 2}$	...	$\pi_{\bullet J}$	1

Έστω τώρα ότι έχουμε N τυχαία επιλεγμένα άτομα (N είναι είτε σταθερά είτε τ.μ.) από τον πληθυσμό και N_{ij} το πλήθος των ατόμων (από τα N) τα οποία ταξινομούνται στο κελί (i, j) (κάθε άτομο ταξινομείται στην θέση (i, j) του πίνακα συνάφειας με πιθανότητα π_{ij}). Ο $I \times J$ πίνακας ο οποίος περιέχει τις παρατηρούμενες συχνότητες N_{ij} καλείται **πίνακας συνάφειας** (contingency table ή cross-classification table). Όμοια με παραπάνω συμβολίζουμε με $N_{i\bullet}$, $N_{\bullet j}$ τα αθροίσματα των γραμμών και των στηλών αντίστοιχα, δηλαδή,

$$N_{i\bullet} = \sum_{j=1}^J N_{ij}, \quad i = 1, 2, \dots, I \quad \text{και} \quad N_{\bullet j} = \sum_{i=1}^I N_{ij}, \quad j = 1, 2, \dots, J.$$

		Y				
		1	2	...	J	Total
X	1	N_{11}	N_{12}	...	N_{1J}	$N_{1\bullet}$
	2	N_{21}	N_{22}	...	N_{2J}	$N_{2\bullet}$
	...	...	...	...	...	...
	I	N_{I1}	N_{I2}	...	N_{IJ}	$N_{I\bullet}$
Total		$N_{\bullet 1}$	$N_{\bullet 2}$	...	$N_{\bullet J}$	$N = N_{\bullet\bullet}$

Σκοπός μας είναι η μελέτη διαφόρων χαρακτηριστικών της διδιάστατης κατανομής π_{ij} με βάση τις παρατηρούμενες συχνότητες N_{ij} . Η **δειγματική** διδιάστατη κατανομή θα συμβολίζεται με p_{ij} , και θα ισχύει ότι $p_{ij} = N_{ij} / N$.

1.1. Προεπιλεγμένα αθροίσματα γραμμών. Σε αρκετές περιπτώσεις πινάκων συνάφειας, η μία μεταβλητή (π.χ. η X) μπορεί να θεωρηθεί ως ερμηνευτική και η άλλη (π.χ. η Y) ως μεταβλητή απόκρισης. Συγκεκριμένα, επιλέγουμε τα άτομα από τον πληθυσμό για προκαθορισμένα αθροίσματα γραμμών $N_{1\bullet}, N_{2\bullet}, \dots, N_{I\bullet}$ (π.χ. επιλέγουμε $N_{1\bullet}$ άτομα με $X=1$, $N_{2\bullet}$ άτομα με $X=2$, κ.ο.κ.). Σε αυτές τις περιπτώσεις δεν έχει νόημα η διδιάστατη κατανομή π_{ij} που περιγράψαμε παραπάνω. Εδώ, η πιθανότητα να ταξινομηθεί ένα άτομο στην (i,j) θέση του πίνακα συνάφειας είναι

$$\pi_{j|i} = P(Y = j | X = i) = \frac{P(Y = j, X = i)}{P(X = i)} = \frac{\pi_{ij}}{\pi_{i\bullet}}.$$

Σε αυτές τις περιπτώσεις μας ενδιαφέρει να μελετήσουμε πως η κατανομή της Y αλλάζει όταν αλλάζει τιμές η X . Τέλος, η **δειγματική** κατανομή σε αυτή την περίπτωση θα συμβολίζεται με $p_{j|i}$, και ισχύει ότι $p_{j|i} = N_{ij} / N_{i\bullet}$.

1.2. Ανεξαρτησία. Οι μεταβλητές X, Y είναι ανεξάρτητες όταν

$$\pi_{ij} = P(X = i, Y = j) = P(X = i)P(Y = j) = \pi_{i\bullet} \cdot \pi_{\bullet j}$$

ή ισοδύναμα όταν

$$\pi_{j|i} = P(Y = j | X = i) = P(Y = j) = \pi_{\bullet j}$$

για κάθε $i = 1, 2, \dots, I, j = 1, 2, \dots, J$.

1.3. Δείκτες συνάφειας. Όταν οι μεταβλητές X, Y είναι δίτιμες (δύο στάθμες) τότε οι πίνακες συνάφειας είναι διαστάσεως 2×2 . Στους πίνακες αυτούς μπορούμε να ορίσουμε τις εξής δείκτες συνάφειας:

- *Relative risk*: π_{111} / π_{112}

- *Odds ratio* ή *cross product ratio*: $\theta = \frac{\pi_{111} / \pi_{211}}{\pi_{112} / \pi_{212}} = \frac{\pi_{111} / \pi_{12}}{\pi_{211} / \pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$

(τιμές τους κοντά στο 1 υποδηλώνουν ανεξαρτησία).

- *log odds ratio*: $\log \theta$

(τιμές του κοντά στο 0 υποδηλώνουν ανεξαρτησία). Το odds ratio και το log odds ratio μπορούν να χρησιμοποιηθούν και σε πίνακες $I \times J$, π.χ.

$$\theta_{ij} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}}.$$

1.4. Μοντέλα κατανομής συχνοτήτων κελιών. Έστω ότι $N_{11}, N_{12}, \dots, N_{IJ}$ άτομα ταξινομούνται στα IJ το πλήθος κελιά ενός πίνακα συνάφειας. Θεωρούμε ότι τα N_{ij} είναι τυχαίες μεταβλητές με μέση τιμή $m_{ij} = E(N_{ij})$ (αναμενόμενες συχνότητες).

α) Ένα από τα απλούστερα μοντέλα που μπορούμε να θεωρήσουμε είναι οι N_{ij} να είναι ανεξάρτητες τ.μ. προερχόμενες από μία κατανομή Poisson. Σε αυτή την περίπτωση η από κοινού σ.π. των συχνοτήτων των κελιών θα είναι

$$f(\mathbf{n}; \mathbf{m}) = P(N_{ij} = n_{ij}, i = 1, 2, \dots, I, j = 1, 2, \dots, J) = \prod_{i=1}^I \prod_{j=1}^J e^{-m_{ij}} \frac{m_{ij}^{n_{ij}}}{n_{ij}!}$$

Εδώ, όλα τα αθροίσματα $N_{i\bullet}, N_{\bullet j}, N_{\bullet\bullet}$ είναι τυχαίες μεταβλητές και μάλιστα ακολουθούν κατανομή Poisson με μέσες τιμές $m_{i\bullet}, m_{\bullet j}, m_{\bullet\bullet}$ αντίστοιχα.

β) Μία άλλη περίπτωση που προκύπτει από την παραπάνω είναι να θεωρήσουμε (πράγμα αρκετά σύνηθες) ότι το συνολικό πλήθος του δείγματος $N_{\bullet\bullet}$ δεν είναι τυχαίο αλλά προκαθορισμένο. Σε αυτή την περίπτωση, αποδεικνύεται εύκολα ότι η από κοινού συνάρτηση πιθανότητας των συχνοτήτων των κελιών θα είναι η πολυωνυμική:

$$\begin{aligned} f(\mathbf{n}; \boldsymbol{\pi}) &= P(N_{ij} = n_{ij}, i = 1, 2, \dots, I, j = 1, 2, \dots, J \mid N_{\bullet\bullet} = n_{\bullet\bullet}) = \prod_{i=1}^I \prod_{j=1}^J e^{-m_{ij}} \frac{m_{ij}^{n_{ij}}}{n_{ij}!} \bigg/ e^{-m_{\bullet\bullet}} \frac{m_{\bullet\bullet}^{n_{\bullet\bullet}}}{n_{\bullet\bullet}!} \\ &= \frac{n_{\bullet\bullet}!}{\prod_{i,j} n_{ij}!} \prod_{i,j} \pi_{ij}^{n_{ij}}, \end{aligned}$$

όπου $\pi_{ij} = m_{ij} / m_{\bullet\bullet}$ και επομένως $\sum_{ij} \pi_{ij} = 1$.

γ) Μία άλλη περίπτωση προκύπτει όταν τα αθροίσματα των γραμμών $N_{i\bullet}$ δεν είναι τυχαία αλλά προκαθορισμένα. Σε αυτή την περίπτωση, η από κοινού συνάρτηση πιθανότητας των συχνοτήτων στην i -γραμμή ($N_{i1}, N_{i2}, \dots, N_{iJ}$) θα είναι η πολυωνυμική:

$$f(\mathbf{n}_i; \boldsymbol{\pi}_i) = P(N_{ij} = n_{ij}, j = 1, 2, \dots, J \mid N_{i\bullet} = n_{i\bullet}) = \prod_{j=1}^J e^{-m_{ij}} \frac{m_{ij}^{n_{ij}}}{n_{ij}!} \bigg/ e^{-m_{i\bullet}} \frac{m_{i\bullet}^{n_{i\bullet}}}{n_{i\bullet}!} = \frac{n_{i\bullet}!}{\prod_{j=1}^J n_{ij}!} \prod_{j=1}^J \pi_{j|i}^{n_{ij}}$$

όπου $\pi_{j|i} = m_{ij} / m_{i\bullet}$ και επομένως $\sum_j \pi_{j|i} = 1$. Επομένως η από κοινού σ.π. όλων των N_{ij} θα προκύπτει (υποθέτοντας ανεξαρτησία μεταξύ των γραμμών) από ένα γινόμενο πολυωνυμικών:

$$f(\mathbf{n}; \boldsymbol{\pi}) = \prod_{i=1}^I \frac{n_{i\bullet}!}{\prod_{j=1}^J n_{ij}!} \prod_{j=1}^J \pi_{j|i}^{n_{ij}}.$$

- Στις *prospective* μελέτες τα αθροίσματα των γραμμών $N_{i\bullet}$ είναι συνήθως προκαθορισμένα και θεωρούμε τις I το πλήθος J -διάστατες μεταβλητές κάθε γραμμής ($N_{i1}, N_{i2}, \dots, N_{iJ}$) ως ανεξάρτητες πολυωνυμικές.

- Στις *retrospective* μελέτες τα αθροίσματα των στηλών $N_{\bullet j}$ είναι συνήθως προκαθορισμένα και θεωρούμε τις J το πλήθος I -διάστατες μεταβλητές κάθε στήλης ($N_{1j}, N_{2j}, \dots, N_{Ij}$) ως ανεξάρτητες πολυωνυμικές.

- Στις *cross-sectional* μελέτες το συνολικό μέγεθος του δείγματος είναι προκαθορισμένο και θεωρούμε την I - J -διάστατη μεταβλητή ($N_{11}, N_{12}, \dots, N_{IJ}$) ως πολυωνυμική.

1.5. Εκτιμήτριες μέγιστης πιθανοφάνειας. Ας εξετάσουμε τις εκτιμήτριες μέγιστης πιθανοφάνειας (MLE's) των π_{ij} στην περίπτωση που έχουμε πολυωνυμική κατανομή. Η log-likelihood συνάρτηση είναι της μορφής ($\sum_{ij} \pi_{ij} = 1$):

$$\begin{aligned}
l(\boldsymbol{\pi}) &= \log L(\boldsymbol{\pi}) = \log \frac{n_{..}!}{\prod_{i,j} N_{ij}!} \prod_{i,j} \pi_{ij}^{N_{ij}} = c + \sum_{i,j} N_{ij} \log \pi_{ij} = c + \sum_{\substack{i,j \\ (i,j) \neq (1,1)}} N_{ij} \log \pi_{ij} + N_{11} \log \pi_{11} \\
&= c + \sum_{\substack{i,j \\ (i,j) \neq (1,1)}} N_{ij} \log \pi_{ij} + N_{11} \log(1 - \sum_{\substack{i,j \\ (i,j) \neq (1,1)}} \pi_{ij})
\end{aligned}$$

και θέτοντας

$$\frac{\partial}{\partial \pi_{hk}} l(\boldsymbol{\pi}) = \frac{N_{hk}}{\pi_{hk}} - \frac{N_{11}}{1 - \sum_{i,j: (i,j) \neq (1,1)} \pi_{ij}} = \frac{N_{hk}}{\pi_{hk}} - \frac{N_{11}}{\pi_{11}} = 0$$

(για $(h,k) \neq (1,1)$) προκύπτει ότι $\hat{\pi}_{hk} = \hat{\pi}_{11} N_{hk} / N_{11}$ και επομένως,

$$1 - \hat{\pi}_{11} = \sum_{i,j: (i,j) \neq (1,1)} \hat{\pi}_{ij} = \frac{\hat{\pi}_{11}}{N_{11}} \sum_{i,j: (i,j) \neq (1,1)} N_{ij} = \frac{\hat{\pi}_{11}}{N_{11}} (N_{..} - N_{11})$$

Άρα $\hat{\pi}_{11} = N_{11} / N_{..}$ και $\hat{\pi}_{hk} = N_{hk} / N_{..} = p_{ij}$. Αν τώρα υποθέσουμε ότι οι X, Y είναι ανεξάρτητες τότε όπως είδαμε παραπάνω $\pi_{ij} = \pi_{i\bullet} \cdot \pi_{\bullet j}$ και σε αυτή την περίπτωση οι MLE's όμοια είναι:

$$\hat{\pi}_{ij} = \frac{N_{i\bullet} N_{\bullet j}}{N_{..}^2} = p_{i\bullet} p_{\bullet j}$$

και επομένως $\hat{m}_{ij} = N_{..} \hat{\pi}_{ij} = N_{..} p_{i\bullet} p_{\bullet j} = N_{i\bullet} N_{\bullet j} / N_{..}$.

1.6. Έλεγχος καλής προσαρμογής. Έστω ότι επιθυμούμε να ελέγξουμε την υπόθεση

$$H_0: \pi_{ij} = \pi'_{ij}, i = 1, \dots, I, j = 1, \dots, J$$

(τα π_{ij} είναι ίσα με κάποιες προκαθορισμένες τιμές π'_{ij} με $\sum \pi'_{ij} = 1$). Αν ισχύει η H_0 τότε οι αναμενόμενες συχνότητες στα κελιά είναι $m_{ij} = N_{..} \pi'_{ij}$. Ο Pearson (1900) πρότεινε την εξής στατιστική συνάρτηση (X^2 του Pearson)

$$X^2 = \sum_{i,j} \frac{(N_{ij} - m_{ij})^2}{m_{ij}}$$

για την οποία απέδειξε ότι ασυμπτωτικά ακολουθεί χι-τετράγωνο κατανομή με $IJ-1$ βαθμούς ελευθερίας. Μεγάλες τιμές του X^2 υποδηλώνουν μεγάλες αποκλίσεις των παρατηρούμενων N_{ij} από τα αναμενόμενα m_{ij} και επομένως απορρίπτουμε σε ε.σ. α την H_0 όταν

$$X^2 = \sum_{i,j} \frac{(N_{ij} - m_{ij})^2}{m_{ij}} > \chi_{IJ-1; \alpha}^2.$$

Σε ορισμένες περιπτώσεις οι τιμές π'_{ij} δεν είναι πλήρως καθορισμένες αλλά εξαρτώνται από ένα μικρότερο σύνολο από t άγνωστες παραμέτρους οι οποίες εκτιμώνται από τα δεδομένα. Σε αυτή την περίπτωση η στατιστική συνάρτηση X^2 υπολογίζεται με βάση τα εκτιμώμενα π'_{ij} ($\hat{m}_{ij} = N_{..} \hat{\pi}'_{ij}$). Ο R.A. Fisher (1922) απέδειξε (παρά τον αντίθετο ισχυρισμό και τις σφοδρές αντιρρήσεις του Pearson) ότι χρησιμοποιώντας εκτιμώμενα m_{ij} αντί τα ακριβή, οι βαθμοί ελευθερίας της χι-τετράγωνο κατανομής του X^2 μειώνονται κατά t ($=$ πλήθος από εκτιμώμενες παραμέτρους που χρησιμοποιήθηκαν για την εκτίμηση των m_{ij}), και επομένως θα απορρίπτουμε την H_0 όταν

$$X^2 = \sum_{i,j} \frac{(N_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} > \chi_{IJ-1-t; \alpha}^2.$$

1.7. Έλεγχος ανεξαρτησίας. Στους διδιάστατους πίνακες συνάφειας με προκαθορισμένο μέγεθος δείγματος $N_{..}$ (πολυωνυμικό μοντέλο) ο έλεγχος της υπόθεσης της ανεξαρτησίας γράφεται ως εξής:

$$H_0 : \pi_{ij} = \pi_{i\bullet} \pi_{\bullet j} \text{ για κάθε } i, j.$$

Για τον έλεγχο της παραπάνω υπόθεσης μπορούμε να χρησιμοποιήσουμε είτε το X^2 του Pearson που περιγράψαμε σε προηγούμενη παράγραφο, είτε το λόγο των πιθανοφανειών.

α) Έλεγχος μέσω του X^2 του Pearson. Σε αυτή την περίπτωση έχουμε τον έλεγχο της υπόθεσης $H_0 : \pi_{ij} = \pi'_{ij}$ όπου $\pi'_{ij} = \pi_{i\bullet} \pi_{\bullet j}$. Επομένως τα π'_{ij} δεν είναι γνωστά αλλά όπως είδαμε παραπάνω εκτιμώνται από τα $\hat{\pi}_{ij} = p_{i\bullet} p_{\bullet j}$. Σύμφωνα και με σχόλια παραπάνω παραγράφου, απορρίπτουμε την H_0 όταν

$$X^2 = \sum_{i,j} \frac{(N_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} > \chi^2_{(I-1)(J-1); \alpha}$$

όπου $\hat{m}_{ij} = N_{..} \hat{\pi}_{ij} = N_{..} p_{i\bullet} p_{\bullet j} = N_{i\bullet} N_{\bullet j} / N_{..}$. Οι β.ε. της X^2 είναι $(I-1)(J-1)$ διότι για την εκτίμηση των π_{ij} (ή ισοδύναμα των m_{ij}) χρησιμοποιήσαμε τις $I-1$ εκτιμήτριες $p_{i\bullet}, i=1,2,\dots,I-1$ και τις $J-1$ εκτιμήτριες $p_{\bullet j}, j=1,2,\dots,J-1$ (σημειώνεται ότι $p_{I\bullet} = 1 - \sum_{i=1}^{I-1} p_{i\bullet}$, $p_{\bullet J} = 1 - \sum_{j=1}^{J-1} p_{\bullet j}$). Συνεπώς β.ε. = $IJ-1-(I-1)-(J-1) = (I-1)(J-1)$.

β) Έλεγχος μέσω του γενικευμένου λόγου πιθανοφανειών. Είναι γνωστό (Wilks (1935, 1938)) ότι κάτω από κατάλληλες συνθήκες ομαλότητας της πιθανοφάνειας $L(\boldsymbol{\pi})$ η στατιστική συνάρτηση

$$-2 \log \frac{\sup_{\boldsymbol{\pi} \in \Theta_0} L(\boldsymbol{\pi})}{\sup_{\boldsymbol{\pi} \in \Theta_1} L(\boldsymbol{\pi})} = 2 \left(\sup_{\boldsymbol{\pi} \in \Theta_1} l(\boldsymbol{\pi}) - \sup_{\boldsymbol{\pi} \in \Theta_0} l(\boldsymbol{\pi}) \right) \xrightarrow{n \rightarrow \infty} \chi^2_{\dim(\Theta_1) - \dim(\Theta_0)}$$

($\dim(\Theta) = \text{διάσταση του χώρου } \Theta$). Αν λοιπόν θέλουμε να ελέγξουμε την υπόθεση της ανεξαρτησίας $H_0 : \pi_{ij} = \pi_{i\bullet} \pi_{\bullet j}$ στο πολυωνυμικό μοντέλο των πινάκων συνάφειας, αρκεί να υπολογίσουμε τη συνάρτηση (likelihood-ratio chi-square statistic)

$$\begin{aligned} 2 \left(\sup_{H_0 \cup H_1} l(\boldsymbol{\pi}) - \sup_{H_0} l(\boldsymbol{\pi}) \right) &= 2 \left(\sup_{H_0 \cup H_1} \sum_{i,j} N_{ij} \log \pi_{ij} - \sup_{H_0} \sum_{i,j} N_{ij} \log \pi_{ij} \right) \\ &= 2 \left(\sum_{i,j} N_{ij} \log \frac{N_{ij}}{N_{..}} - \sum_{i,j} N_{ij} \log \frac{N_{i\bullet} N_{\bullet j}}{N_{..}^2} \right) = 2 \sum_{i,j} N_{ij} \log \frac{N_{ij}}{\hat{m}_{ij}} \end{aligned}$$

και συνεπώς απορρίπτουμε την H_0 όταν

$$G^2 := 2 \sum_{i,j} N_{ij} \log \frac{N_{ij}}{\hat{m}_{ij}} > \chi^2_{(I-1)(J-1)}$$

διότι β.ε. = $\dim(\Theta_1) - \dim(\Theta_0) = (IJ-1) - (I-1+J-1) = (I-1)(J-1)$ όπου Θ_0 και Θ_1 εδώ είναι οι παραμετρικοί χώροι υπό την H_0 και $H_0 \cup H_1$ αντίστοιχα.

Σημειώνεται ότι το X^2 και το G^2 είναι ίσα ασυμπτωτικά. Το X^2 συνήθως συγκλίνει πιο γρήγορα στην χι-τετράγωνο κατανομή από ότι το G^2 . Η χι-τετράγωνο προσέγγιση της κατανομής του G^2 δεν είναι καλή όταν $N_{..} / IJ < 5$. Αντίθετα, το X^2 προσεγγίζεται καλύτερα από την χ^2 ακόμη και για $N_{..} / IJ = 1$ με την προϋπόθεση ο πίνακας να μην περιέχει πολύ μικρές ή σχετικά μεγάλες αναμενόμενες συχνότητες.

2. Πίνακες συνάφειας χρησιμοποιώντας γενικευμένα γραμμικά μοντέλα

Παραπάνω είδαμε πως μπορούμε να μελετήσουμε διδιάστατους πίνακες συνάφειας χωρίς να χρειαστεί να καταφύγουμε σε αποτελέσματα της θεωρίας των γενικευμένων γραμμικών μοντέλων (GLM). Εάν όμως επιθυμούμε να μελετήσουμε πίνακες συνάφειας με περισσότερες από δύο διαστάσεις τότε ο υπολογισμός των MLE's δεν είναι πάντα τόσο απλός και θα πρέπει αναγκαστικά να βασιστούμε στη γενική θεωρία των γενικευμένων γραμμικών μοντέλων. Επιπλέον η θεωρία των GLM προσφέρει ένα θεωρητικό πλαίσιο για την ενιαία μελέτη των πινάκων συνάφειας όλων των διαστάσεων. Ας δούμε πως μπορεί να ενταχθεί η μελέτη των πινάκων συνάφειας στα GLM ξεκινώντας από την απλούστερη περίπτωση των διδιάστατων πινάκων συνάφειας.

2.1. Μελέτη διδιάστατων πινάκων χρησιμοποιώντας γενικευμένα γραμμικά μοντέλα.

Όπως είδαμε στην προηγούμενη παράγραφο, στους διδιάστατους πίνακες μπορούμε να θεωρήσουμε τα εξής μοντέλα:

1) *Μοντέλο Poisson*. Ισχύει ότι $E(N_{ij}) = m_{ij} = m_{..}\pi_{ij}$ και η υπόθεση της ανεξαρτησίας (γραμμών και στηλών) μεταφράζεται ως εξής: $E(N_{ij}) = m_{..}\pi_{i.}\pi_{.j}$.

2) *Πολυωνυμικό μοντέλο*. Ισχύει ότι $E(N_{ij}) = n\pi_{ij}$ και η υπόθεση της ανεξαρτησίας μεταφράζεται ως εξής: $E(N_{ij}) = n\pi_{i.}\pi_{.j}$.

3) *Μοντέλο γινομένου πολυωνυμικών*. Ισχύει ότι $E(N_{ij}) = n_{i.}\pi_{j|i} = n_{i.}\pi_{ij} / \pi_{i.}$ και η υπόθεση της ομοιογένειας (ίδια κατανομή των $(N_{i1}, N_{i2}, \dots, N_{iJ})$ σε κάθε γραμμή) μεταφράζεται ως εξής: $E(N_{ij}) = n_{i.}\pi_{.j}$.

2.1.1. Παραμετροποίηση του μοντέλου. Παρατηρούμε ότι οι συνήθεις υποθέσεις (ανεξαρτησία, ίδια κατανομή στις γραμμές κτλ) μπορούν να εκφραστούν ως πολλαπλασιαστικά μοντέλα στα οποία οι αναμενόμενες συχνότητες είναι ίσες με γινόμενα περιθωρίων πιθανοτήτων και προκαθορισμένων αθροισμάτων συχνοτήτων. Συνεπώς, κάτω από αυτές τις υποθέσεις ο λογάριθμος των αναμενόμενων συχνοτήτων μπορεί να γραφεί στη (προσθετική τώρα) μορφή

$$\eta_{ij} = \log m_{ij} = \log E(N_{ij}) = \mathbf{x}_i^T \boldsymbol{\beta}$$

για κατάλληλες παραμέτρους $\boldsymbol{\beta}$. Μία αρκετά βολική παραμετροποίηση του μοντέλου είναι η ακόλουθη

$$\mu = \bar{\eta}_{..}, \quad \lambda_i^X = \bar{\eta}_{i.} - \bar{\eta}_{..}, \quad \lambda_j^Y = \bar{\eta}_{.j} - \bar{\eta}_{..}, \quad \lambda_{ij}^{XY} = \eta_{ij} - \bar{\eta}_{i.} - \bar{\eta}_{.j} + \bar{\eta}_{..}$$

όπου $\bar{\eta}_{i.} = \frac{1}{J} \sum_{j=1}^J \eta_{ij}$, $\bar{\eta}_{.j} = \frac{1}{I} \sum_{i=1}^I \eta_{ij}$, $\bar{\eta}_{..} = \frac{1}{IJ} \sum_{i=1}^I \sum_{j=1}^J \eta_{ij}$, και επομένως

$$\eta_{ij} = \log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}.$$

Το μοντέλο αυτό περιγράφει επακριβώς οποιοδήποτε σύνολο από αναμενόμενες συχνότητες (πλήρες ή κορεσμένο μοντέλο, saturated model). Οι παράμετροι λ_i^X, λ_j^Y θεωρούνται ως αποκλίσεις από το γενικό μέσο μ , και είναι εύκολο να δούμε ότι αθροίζουν στο 0. Επομένως υπάρχουν $I-1$ ανεξάρτητες παράμετροι γραμμών και $J-1$ ανεξάρτητες παράμετροι στηλών. Αντίστοιχα, οι αλληλεπιδράσεις λ_{ij}^{XY} ικανοποιούν

$$\sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY} = 0$$

Επομένως $(I-1)(J-1)$ από αυτές τις παραμέτρους είναι γραμμικά ανεξάρτητες.

2.1.2. Έλεγχοι ανεξαρτησίας. Η παραπάνω παραμετροποίηση είναι αρκετά βολική στην περίπτωση που θέλουμε να κάνουμε έλεγχο της ανεξαρτησίας των X, Y . Για παράδειγμα, στο πολυωνυμικό μοντέλο η υπόθεση της ανεξαρτησίας μπορεί να εκφραστεί ως εξής:

$$\eta_{ij} = \log E(N_{ij}) = \log n + \log \pi_{i\bullet} + \log \pi_{\bullet j} = \mu + \lambda_i^X + \lambda_j^Y$$

διότι αν $\eta_{ij} = \log n + \log \pi_{i\bullet} + \log \pi_{\bullet j}$ τότε είναι εύκολο να δούμε ότι

$$\lambda_i^X = \log \pi_{i\bullet} - \frac{1}{I} \sum_h \log \pi_{h\bullet}, \quad \lambda_j^Y = \log \pi_{\bullet j} - \frac{1}{J} \sum_h \log \pi_{\bullet h}, \quad \mu = \log n + \frac{1}{I} \sum_h \log \pi_{h\bullet} + \frac{1}{J} \sum_h \log \pi_{\bullet h}$$

Επομένως, το μοντέλο ανεξαρτησίας των X, Y είναι ειδική περίπτωση του παραπάνω πλήρους μοντέλου θεωρώντας ότι $\lambda_{ij}^{XY} = 0$ (οι παράμετροι λ_{ij}^{XY} αντικατοπτρίζουν απόκλιση από την ανεξαρτησία). Το πλήρες μοντέλο έχει $1+(I-1)+(J-1)+(I-1)(J-1) = IJ$ ανεξάρτητες παραμέτρους, ενώ το μοντέλο ανεξαρτησίας έχει $1+(I-1)+(J-1) = I+J-1$ ανεξάρτητες παραμέτρους. Ο έλεγχος ανεξαρτησίας γίνεται εξετάζοντας αν το μοντέλο $\eta_{ij} = \mu + \lambda_i^X + \lambda_j^Y$ είναι αποδεκτό.

2.1.3. Μοντέλο κατανομής συχνοτήτων κελιών. Ας θεωρήσουμε τώρα το απλούστερο μοντέλο για την κατανομή των συχνοτήτων των κελιών που είναι το μοντέλο Poisson (οι συχνότητες είναι ανεξάρτητες τ.μ. που ακολουθούν την κατανομή Poisson με μέση τιμή m_{ij}). Η από κοινού συνάρτηση πιθανότητας των συχνοτήτων N_{ij} των κελιών είναι της μορφής

$$f(\mathbf{n}; \mathbf{m}) = \prod_{i=1}^I \prod_{j=1}^J e^{-m_{ij}} \frac{m_{ij}^{n_{ij}}}{n_{ij}!}$$

όπου οι μέσες τιμές m_{ij} εξαρτώνται από κάποιες παραμέτρους $\boldsymbol{\beta}$ μέσω μιας σχέσης της μορφής

$$\ln m_{ij} = \ln E(N_{ij}) = \mathbf{x}_i^T \boldsymbol{\beta}$$

ανάλογα με το μοντέλο. Παρατηρούμε ότι η παραπάνω κατανομή $f(\mathbf{n}; \mathbf{m})$ ανήκει στην εκθετική οικογένεια κατανομών και μάλιστα τα m_{ij} εξαρτώνται από ένα μικρότερο σύνολο παραμέτρων (π.χ. κάποιες από τις $\mu, \lambda_i^X, \lambda_j^Y, \lambda_{ij}^{XY}$) με λογαριθμική συνάρτηση σύνδεσης (δηλαδή η σχέση των m_{ij} με τα $\mu, \lambda_i^X, \lambda_j^Y, \lambda_{ij}^{XY}$ είναι «λογαριθμογραμμική», loglinear). Επομένως έχουμε το κλασσικό γενικευμένο γραμμικό μοντέλο και οι εκτιμήτριες μέγιστης πιθανοφάνειας μπορούν να βρεθούν χρησιμοποιώντας τις γνωστές επαναληπτικές μεθόδους (Fisher's scoring method).

Ο Birch (1963) απέδειξε ότι οι εκτιμήτριες μέγιστης πιθανοφάνειας στα loglinear μοντέλα είναι ίδιες είτε πρόκειται για μοντέλο Poisson είτε για πολυωνυμικό μοντέλο είτε για μοντέλο που είναι γινόμενο πολυωνυμικών. Η μόνη συνθήκη για να ισχύει αυτό είναι να περιέχονται στο μοντέλο οι παράμετροι οι οποίοι αντιστοιχούν σε προκαθορισμένα αθροίσματα. Για παράδειγμα, αν σε διδιάστατο πίνακα συνάφειας τα $N_{i\bullet} = n_{i\bullet}$ είναι προκαθορισμένα, τότε μπορούμε για την εκτίμηση των παραμέτρων να χρησιμοποιήσουμε το μοντέλο Poisson αρκεί να συμπεριλάβουμε και την παράμετρο που αντιστοιχεί στο άθροισμα $n_{i\bullet}$, δηλαδή το λ_i^X , ώστε οι εκτιμημένες αναμενόμενες τιμές ικανοποιούν ότι $\hat{m}_{i\bullet} = n_{i\bullet}$ για κάθε i . Επομένως θα θεωρούμε πάντα το μοντέλο Poisson εξασφαλίζοντας όμως ότι ικανοποιείται η παραπάνω συνθήκη.

2.1.4. Έλεγχος καλής προσαρμογής. Για διδιάστατους πίνακες συνάφειας, η στατιστική συνάρτηση που προκύπτει από το γενικευμένο λόγο πιθανοφάνειών (likelihood ratio statistic ή Deviance) και το χι-τετράγωνο του Pearson αντίστοιχα είναι

$$G^2 = 2 \sum_{i,j} N_{ij} \log \frac{N_{ij}}{\hat{m}_{ij}}, \quad X^2 = \sum_{i,j} \frac{(N_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}.$$

Όταν το μοντέλο είναι σωστό, οι παραπάνω στατιστικές συναρτήσεις ακολουθούν ασυμπτωτικά χι-τετράγωνο κατανομή. Οι β.ε. είναι ίσοι με το πλήθος των παραμέτρων στο πλήρες μοντέλο μείον το πλήθος των (ανεξάρτητων) παραμέτρων στο μοντέλο που εξετάζουμε. Άρα απορρίπτουμε το εκάστοτε μοντέλο σε ε.σ. α αν

$$G^2 > \chi^2_{df;a} \text{ ή } X^2 > \chi^2_{df;a}$$

Τα δ.ε. και οι έλεγχοι για τις παραμέτρους του μοντέλου γίνονται σύμφωνα με τη γενική θεωρία GLM.

2.2. Μελέτη τριδιάστατων πινάκων χρησιμοποιώντας γενικευμένα γραμμικά μοντέλα.

Σε πίνακες συνάφειας με τρεις διαστάσεις (X, Y, Z κατηγορικές μεταβλητές με I, J, K στάθμες αντίστοιχα.) μπορούμε και πάλι να θεωρήσουμε τα μοντέλα:

- 1) *Poisson*: Οι συχνότητες N_{ijk} σε κάθε κελί είναι ανεξάρτητες και ακολουθούν κατανομή Poisson με μέση τιμή m_{ijk} .
- 2) *Πολυωνυμικό*: Η από κοινού κατανομή των συχνοτήτων N_{ijk} ακολουθεί πολυωνυμική κατανομή με παραμέτρους n και π_{ijk} (το γενικό άθροισμα είναι προκαθορισμένο, $N_{\dots} = n$)
- 3) *Γινόμενο Πολυωνυμικών*: Αν κάποια από τα αθροίσματα (π.χ. των γραμμών σε κάθε στρώμα ή των στρωμάτων) είναι προκαθορισμένα τότε η από κοινού κατανομή των συχνοτήτων N_{ijk} είναι ένα γινόμενο κατάλληλων πολυωνυμικών κατανομών.

Τα παραπάνω μοντέλα γενικεύονται και σε περισσότερες διαστάσεις.

2.2.1. Παραμετροποίηση του μοντέλου. Και σε πίνακες συνάφειας με περισσότερες από δύο διαστάσεις οι συνήθεις υποθέσεις (π.χ. ανεξαρτησία μεταξύ X, Y, Z) μπορούν να εκφραστούν ως πολλαπλασιαστικά μοντέλα στα οποία οι αναμενόμενες συχνότητες είναι ίσες με γινόμενα περιθωρίων πιθανοτήτων και προκαθορισμένων αθροισμάτων συχνοτήτων. Συνεπώς, και στις τρεις διαστάσεις υπό τις συνήθεις υποθέσεις ο λογάριθμος των αναμενόμενων συχνοτήτων γράφεται στη μορφή

$$\eta_{ijk} = \log m_{ijk} = \log E(N_{ijk}) = \mathbf{x}_i^T \boldsymbol{\beta}$$

για κατάλληλες παραμέτρους $\boldsymbol{\beta}$. Η ανάλογη με τη διδιάστατη περίπτωση παραμετροποίηση εδώ είναι

$$\mu = \bar{\eta}_{\dots}, \quad \lambda_i^X = \bar{\eta}_{i..} - \bar{\eta}_{\dots}, \quad \lambda_j^Y = \bar{\eta}_{.j.} - \bar{\eta}_{\dots}, \quad \lambda_k^Z = \bar{\eta}_{\dots k} - \bar{\eta}_{\dots}$$

$$\lambda_{ij}^{XY} = \bar{\eta}_{ij.} - \bar{\eta}_{i..} - \bar{\eta}_{.j.} + \bar{\eta}_{\dots}, \quad \lambda_{ik}^{XZ} = \bar{\eta}_{i.k} - \bar{\eta}_{i..} - \bar{\eta}_{\dots k} + \bar{\eta}_{\dots}, \quad \lambda_{jk}^{YZ} = \bar{\eta}_{.jk} - \bar{\eta}_{.j.} - \bar{\eta}_{\dots k} + \bar{\eta}_{\dots}$$

$$\lambda_{ijk}^{XYZ} = \eta_{ijk} - \bar{\eta}_{ij.} - \bar{\eta}_{i.k} - \bar{\eta}_{.jk} + \bar{\eta}_{i..} + \bar{\eta}_{.j.} + \bar{\eta}_{\dots k} - \bar{\eta}_{\dots}$$

και αντικαθιστώντας προκύπτει το πλήρες μοντέλο (saturated model)

$$\eta_{ijk} = \log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$$

Τα αθροίσματα των παραμέτρων ως προς οποιοδήποτε δείκτη είναι ίσο με 0:

$$\sum_i \lambda_i^X = \sum_j \lambda_j^Y = \sum_k \lambda_k^Z = \sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY} = \dots = \sum_k \lambda_{ijk}^{XYZ} = 0$$

Το πλήρες μοντέλο προφανώς έχει IJK ανεξάρτητες παραμέτρους (ίσες με το πλήθος κελιών του πίνακα).

Στο εξής θα θεωρήσουμε ιεραρχικά (hierarchical) μοντέλα. Στα μοντέλα αυτά όταν υπάρχει η αλληλεπίδραση μεταξύ κάποιων μεταβλητών τότε στο μοντέλο θα υπάρχουν και οι μικρότερες αλλη-

λεπιδράσεις καθώς και οι κύριες επιδράσεις μεταξύ των συγκεκριμένων μεταβλητών. Για παράδειγμα, αν στο μοντέλο υπάρχει η λ_{ik}^{XZ} τότε θα υπάρχουν και οι λ_i^X, λ_k^Z ενώ π.χ. αν υπάρχει η λ_{ijk}^{XYZ} τότε θα υπάρχουν όλες οι ανά δύο αλληλεπιδράσεις των X, Y, Z καθώς και οι κύριες επιδράσεις τους, δηλαδή θα έχουμε το πλήρες μοντέλο. Μερικά από τα ιεραρχικά μοντέλα που μπορούν να θεωρηθούν σε έναν τριδιάστατο πίνακα συνάφειας δίνονται στον παρακάτω πίνακα:

Loglinear Model	Συμβολισμός,
$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$	(X, Y, Z)
$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$	(XY, Z)
$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$	(XY, YZ)
$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ}$	(XY, YZ, XZ)
$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$	(XYZ)

Παρατήρηση. Το στατιστικό πακέτο SPSS χρησιμοποιεί μια διαφορετική παραμετροποίηση η οποία δεν βασίζεται στον περιορισμό ότι τα αθροίσματα των παραμέτρων ως προς οποιοδήποτε δείκτη είναι ίσο με 0 (sum-to-zero constraints). Συγκεκριμένα, στο SPSS η παραμετροποίηση γίνεται έτσι ώστε να ικανοποιούνται οι περιορισμοί (χρησιμοποιούμε β αντί λ)

$$\beta_i^X = \beta_j^Y = \beta_k^Z = \beta_{ij}^{XY} = \beta_{ik}^{XZ} = \dots = \beta_{ijk}^{XYZ} = 0$$

(corner-point constraints). Παρόλα αυτά, η σύνδεση των τύπων ανεξαρτησίας και των όρων αλληλεπίδρασης που περιέχονται στο μοντέλο παραμένει η ίδια με παραπάνω. Για παράδειγμα σε έναν 3×3 πίνακα συνάφειας ο πίνακας σχεδιασμού με τις παραμετροποιήσεις $\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y$ και $\log m_{ij} = \alpha + \beta_i^X + \beta_j^Y$ αντίστοιχα είναι (χρησιμοποιούμε μόνο τις ανεξάρτητες παραμέτρους)

$$\begin{bmatrix} \log m_{11} \\ \log m_{12} \\ \log m_{13} \\ \log m_{21} \\ \log m_{22} \\ \log m_{23} \\ \log m_{31} \\ \log m_{32} \\ \log m_{33} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & -1 & -1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 0 \\ 1 & -1 & -1 & 0 & 1 \\ 1 & -1 & -1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \lambda_1^X \\ \lambda_2^X \\ \lambda_1^Y \\ \lambda_2^Y \end{bmatrix}, \quad \begin{bmatrix} \log m_{11} \\ \log m_{12} \\ \log m_{13} \\ \log m_{21} \\ \log m_{22} \\ \log m_{23} \\ \log m_{31} \\ \log m_{32} \\ \log m_{33} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1^X \\ \beta_2^X \\ \beta_1^Y \\ \beta_2^Y \end{bmatrix}$$

2.2.2. Τύποι ανεξαρτησίας σε τριδιάστατους πίνακες συνάφειας. Έστω π_{ijk} οι πιθανότητες των ijk κελιών με $\sum_{i,j,k} \pi_{ijk} = 1$ (πολυωνυμικό μοντέλο, οι X, Y, Z είναι μεταβλητές απόκρισης).

α) Οι X, Y, Z είναι αμοιβαία ανεξάρτητες (mutually independent) όταν

$$\pi_{ijk} = \pi_{i..} \pi_{.j.} \pi_{..k} \text{ για κάθε } i, j, k \text{ ή ισοδύναμα όταν } \log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$$

β) Η μεταβλητή Y είναι από κοινού ανεξάρτητη (jointly independent) από τις X, Z όταν

$$\pi_{ijk} = \pi_{i.k} \pi_{.j.} \text{ για κάθε } i, j, k \text{ ή ισοδύναμα όταν } \log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}$$

(ανάλογες σχέσεις ισχύουν και όταν η X είναι ανεξάρτητη από τις Y, Z ή η Z είναι ανεξάρτητη από τις X, Y).

γ) Οι X, Y είναι δεσμευμένα ανεξάρτητες (conditionally independent) δεδομένης της Z όταν

$$\pi_{ijk} = \pi_{i\bullet k} \pi_{\bullet jk} / \pi_{\bullet\bullet k} \text{ για κάθε } i, j, k \text{ ή ισοδύναμα όταν } \log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

(ανάλογες σχέσεις ισχύουν και όταν οι X, Z είναι δεσμευμένα ανεξάρτητες από την Y ή οι Y, Z είναι δεσμευμένα ανεξάρτητες από την X).

Είναι εύκολο να διαπιστώσουμε ότι η αμοιβαία ανεξαρτησία αποδεικνύει από κοινού ανεξαρτησία π.χ. της Y από τις X, Z η οποία με τη σειρά της αποδεικνύει δεσμευμένη ανεξαρτησία των X, Y δεδομένης της Z .

2.2.3. Μοντέλο κατανομής συχνοτήτων κελιών. Ας θεωρήσουμε αρχικά και εδώ το μοντέλο Poisson (οι συχνότητες σε κάθε κελί είναι ανεξάρτητες τ.μ. που ακολουθούν την κατανομή Poisson με μέση τιμή m_{ijk}). Η από κοινού σ.π. των συχνοτήτων N_{ijk} των κελιών σε τριδιάστατους πίνακες συνάφειας είναι της μορφής

$$f(\mathbf{n}; \mathbf{m}) = \prod_{i=1}^I \prod_{j=1}^J \prod_{k=1}^K e^{-m_{ijk}} \frac{m_{ijk}^{n_{ijk}}}{n_{ijk}!}$$

όπου οι μέσες τιμές m_{ijk} εξαρτώνται από κάποιες από τις παραμέτρους

$$\mu, \lambda_i^X, \lambda_j^Y, \lambda_k^Z, \lambda_{ij}^{XY}, \lambda_{ik}^{XZ}, \lambda_{jk}^{YZ}, \lambda_{ijk}^{XYZ}$$

ανάλογα με το μοντέλο (π.χ. $\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$ κάτω από την υπόθεση της αμοιβαίας ανεξαρτησίας). Αν και σε ορισμένες περιπτώσεις (π.χ. υποθέσεις αμοιβαίας ανεξαρτησίας, από κοινού ανεξαρτησίας, δεσμευμένης ανεξαρτησίας) μπορούμε να βρούμε κλειστές εκφράσεις για τις εκτιμήτριες ML όσων από τις παραπάνω παραμέτρους περιέχονται στο μοντέλο μεγιστοποιώντας την συνάρτηση πιθανοφάνειας, είναι γενικά προτιμότερο (και αυτό γίνεται και από τα περισσότερα στατιστικά πακέτα) να χρησιμοποιήσουμε τη θεωρία των γενικευμένων γραμμικών μοντέλων.

Παρατηρούμε ότι η παραπάνω κατανομή $f(\mathbf{n}; \mathbf{m})$ ανήκει στην εκθετική οικογένεια κατανομών και μάλιστα τα m_{ijk} , εξαρτώνται από ένα μικρότερο σύνολο παραμέτρων (π.χ. κάποιες από τις $\mu, \lambda_i^X, \lambda_j^Y, \lambda_k^Z, \lambda_{ij}^{XY}, \lambda_{ik}^{XZ}, \lambda_{jk}^{YZ}, \lambda_{ijk}^{XYZ}$), με λογαριθμική συνάρτηση σύνδεσης (loglinear model). Επομένως και εδώ έχουμε το κλασσικό γενικευμένο γραμμικό μοντέλο και οι εκτιμήτριες μέγιστης πιθανοφάνειας μπορούν να βρεθούν χρησιμοποιώντας τις γνωστές επαναληπτικές μεθόδους. Για παράδειγμα, υποθέτοντας αμοιβαία ανεξαρτησία των X, Y, Z λαμβάνουμε το μοντέλο

$$\log m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$$

το οποίο περιέχει $1+(I-1)+(J-1)+(K-1)$ ανεξάρτητες παραμέτρους:

$$\beta = (\mu, \lambda_1^X, \dots, \lambda_{I-1}^X, \lambda_1^Y, \dots, \lambda_{J-1}^Y, \lambda_1^Z, \dots, \lambda_{K-1}^Z).$$

Τα αποτελέσματα του Birch (1963) προφανώς ισχύουν και σε περισσότερες από δύο διαστάσεις. Επομένως, οι εκτιμήτριες μέγιστης πιθανοφάνειας στα loglinear μοντέλα είναι ίδιες είτε πρόκειται για οποιοδήποτε μοντέλο (Poisson, πολωνυμικό, γινόμενο πολωνυμικών) αρκεί στο μοντέλο να περιέχονται οι παράμετροι οι οποίοι αντιστοιχούν στα προκαθορισμένα αθροίσματα. Π.χ. αν σε τριδιάστατο πίνακα συνάφειας τα $N_{ij\bullet} = n_{ij\bullet}$ είναι προκαθορισμένα, τότε μπορούμε για την εκτίμηση των παραμέτρων να χρησιμοποιήσουμε το μοντέλο Poisson αρκεί να συμπεριλάβουμε και την παράμετρο που αντιστοιχεί στο άθροισμα $n_{ij\bullet}$, δηλαδή το λ_{ij}^{XY} , ώστε οι εκτιμημένες αναμενόμενες τιμές ικανοποιούν ότι $\hat{m}_{ij\bullet} = n_{ij\bullet}$ για κάθε i και j .

2.2.4. Έλεγχος καλής προσαρμογής. Για τριδιάστατους πίνακες συνάφειας, η στατιστική συνάρτηση που προκύπτει από το γενικευμένο λόγο πιθανοφάνειών (likelihood ratio statistic ή Deviance) και το χι-τετράγωνο του Pearson αντίστοιχα είναι

$$G^2 = 2 \sum_{i,j,k} N_{ijk} \log \frac{N_{ijk}}{\hat{m}_{ijk}},$$

$$X^2 = \sum_{i,j,k} \frac{(N_{ijk} - \hat{m}_{ijk})^2}{\hat{m}_{ijk}}$$

Όταν το μοντέλο είναι σωστό, οι παραπάνω στατιστικές συναρτήσεις ακολουθούν ασυμπτωτικά χι-τετράγωνο κατανομή. Οι β.ε. της κατανομής αυτής είναι ίσοι με το πλήθος των παραμέτρων στο πλήρες μοντέλο μείον το πλήθος των (ανεξάρτητων) παραμέτρων στο μοντέλο που εξετάζουμε:

Μοντέλο	β.ε.	Μοντέλο	β.ε.
(X, Y, Z)	$IJK - I - J - K + 2$	(XY, YZ)	$J(I-1)(K-1)$
(XY, Z)	$(K-1)(IJ-1)$	(XZ, YZ)	$K(I-1)(J-1)$
(XZ, Y)	$(J-1)(IK-1)$	(XY, XZ)	$I(J-1)(K-1)$
(YZ, X)	$(I-1)(JK-1)$	(XY, XZ, YZ)	$(I-1)(J-1)(K-1)$
		(XYZ)	0

(οι παραπάνω β.ε. ισχύουν είτε έχουμε Poisson είτε πολυωνυμικό μοντέλο). Άρα απορρίπτουμε το εκάστοτε μοντέλο σε ε.σ. α αν

$$G^2 > \chi^2_{df;a} \text{ ή } X^2 > \chi^2_{df;a}$$

Τα δ.ε. και οι έλεγχοι για τις παραμέτρους του μοντέλου γίνονται και εδώ σύμφωνα με τη γενική θεωρία GLM.

2.2.5. Σύγκριση μοντέλων. Έστω δύο παραμετρικά μοντέλα M_1 και M_2 τέτοια ώστε το M_2 περιέχεται στο M_1 . Αν επιθυμούμε να ελέγξουμε την υπόθεση της μορφής: H_0 : «το M_2 είναι σωστό» έναντι του μεγαλύτερου μοντέλου H_1 : «το M_1 είναι σωστό», τότε σύμφωνα με το κριτήριο του γ.λ.π. απορρίπτουμε σε ε.σ. α την H_0 όταν

$$G^2(M_2 | M_1) = G^2(M_2) - G^2(M_1) > \chi^2_{v_2 - v_1; \alpha}$$

Στους διδιάστατους και τριδιάστατους πίνακες συνάφειας έχουμε τα ακόλουθα μοντέλα:

1. Μοντέλα σε διδιάστατους πίνακες συνάφειας

α) X, Y : μεταβλητές απόκρισης (το γενικό άθροισμα $N_{..} = n$ είναι προκαθορισμένο)

Ερμηνεία	Μοντέλο	β.ε.
i) X, Y ανεξάρτητες	$m_{ij} = n\pi_{i.}\pi_{.j}$	$IJ - I - J + 1$
ii) Πλήρες μοντέλο	$m_{ij} = n\pi_{ij}$	0

β) X : μεταβλητή απόκρισης, Y ερμηνευτική μεταβλητή (τα αθροίσματα $N_{.j} = n_{.j}$ είναι προκαθορισμένα). Στα μοντέλα πρέπει (από την συνθήκη του Birch) να περιλαμβάνεται ο όρος $\mu + \lambda_j^Y$.

Ερμηνεία	Μοντέλο	β.ε.
i) $H X$ έχει ίδια κατανομή σε όλες τις στάθμες της Y	$m_{ij} = n_{.j}\pi_{i.}$	$IJ - I - J + 1$
ii) Πλήρες μοντέλο	$m_{ijk} = n_{.j}\pi_{ij}$	0

2. Μοντέλα σε τριδιάστατους πίνακες συνάφειας

α) X, Y, Z : μεταβλητές απόκρισης, (το γενικό άθροισμα $N_{...} = n$ είναι προκαθορισμένο)

Ερμηνεία	Μοντέλο	β.ε.
i) X, Y, Z αμοιβαία ανεξάρτητες	$m_{ijk} = n\pi_{i..}\pi_{.j.}\pi_{..k}$	$IJK - I - J - K + 2$
ii) Z ανεξάρτητη από τις X, Y	$m_{ijk} = n\pi_{ij.}\pi_{..k}$	$(K-1)(IJ-1)$
iii) X, Z ανεξάρτητες δεδομένης της Y (ανεξάρτητες σε όλες τις στάθμες της Y)	$m_{ijk} = n\pi_{ij.}\pi_{.jk}\pi_{..k}$	$J(I-1)(K-1)$
iv) Ανά δύο εξαρτημένες (η σχέση μεταξύ δύο μεταβλητών είναι η ίδια σε όλα τα επίπεδα της τρίτης μεταβλητής)	$m_{ijk} = n\pi_{ij.}\pi_{.jk}\pi_{i.k}$	$(I-1)(J-1)(K-1)$
v) Πλήρες μοντέλο	$m_{ijk} = n\pi_{ijk}$	0

β) X, Y : μεταβλητές απόκρισης, Z : ερμηνευτική μεταβλητή (τα αθροίσματα $N_{..k} = n_{..k}$ είναι προκαθορισμένα). Στα μοντέλα πρέπει (συνθήκη Birch) να περιλαμβάνεται ο όρος $\mu + \lambda_k^Z$ (από την συνθήκη του Birch)..

Ερμηνεία	Μοντέλο	β.ε.
i) Οι X, Y έχουν ίδια κατανομή σε όλες τις στάθμες της Z	$m_{ijk} = n_{..k}\pi_{ij.}$	$(K-1)(IJ-1)$
ii) Οι X, Y ανεξάρτητες σε κάθε στάθμη της ερμηνευτικής μεταβλητής Z	$m_{ijk} = n_{..k}\pi_{i.k}\pi_{.jk}$	$K(I-1)(J-1)$
iii) Πλήρες μοντέλο	$m_{ijk} = n_{..k}\pi_{ijk}$	0

γ) X : μεταβλητή απόκρισης, Y, Z : ερμηνευτικές μεταβλητές (τα αθροίσματα $N_{.jk} = n_{.jk}$ είναι προκαθορισμένα). Στα μοντέλα πρέπει (συνθήκη Birch) να περιλαμβάνεται ο όρος $\mu + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}$

Ερμηνεία	Μοντέλο	β.ε.
i) $H X$ έχει ίδια κατανομή σε όλες τις στάθμες της Y (για δεδομένη στάθμη της Z)	$m_{ijk} = n_{.jk}\pi_{i.k}$	$K(I-1)(J-1)$
ii) $H X$ έχει ίδια κατανομή σε όλες τις στάθμες των Y, Z	$m_{ijk} = n_{.jk}\pi_{i..}$	$(I-1)(JK-1)$
iii) Πλήρες μοντέλο	$m_{ijk} = n_{.jk}\pi_{ijk}$	0

Στο SPSS τα παραπάνω μοντέλα μπορούν να εξετασθούν χρησιμοποιώντας τη διαδικασία Analyze / loglinear / General (αφού πρώτα θέσουμε τις παρατηρούμενες συχνότητες των κελιών ως βάρη: Data/ Weight cases). Π.χ. το μοντέλο (XY, YZ) καθορίζεται από την επιλογή Analyze/loglinear/General /model: X, Y, Z (main effects), $X*Y, Y*Z$ (interactions).