Μοντέλα για την ανάλυση των επαναλαμβανόμενων κατηγορικών δεδομένων

ΒΑΣΙΛΕΙΟΣ Γ. Σ. ΒΑΣΔΕΚΗΣ
Πανεπιστήμιο Κρήτης

ΠΕΡΙΛΗΨΗ Η ανάλυση κατηγορικών δεδομένων είναι ιδιαίτερης σημασίας για την ψυχολογία μια και τέτοιου τύπου αποτελέσματα εμφανίζονται ουσιαστικά σε κάθε πείραμα. Ιδιαίτερης αντιμετώπισης χρήζουν τα επαναλαμβανόμενα κατηγορικά δεδομένα. Αυτά είναι πραγματοποιήσεις κατηγορικών μεταβλητών που λαμβάνονται επαναληπτικά σε διάφορα τακτά ή όχι χρονικά διαστήματα από ένα δείγμα υποκειμένων. Εξετάζεται η σημασία της ύπαρξης συσχετίσεων σε τέτοιου τύπου δεδομένα και ανασκοπώνται τρεις μέθοδοι ανάλυσης: η Cochran's Q, η Weighted Least Squares (WLS, Μέθοδος Σταθμισμένων Ελαχίστων Τετραγώνων) και η Generalized Estimation Equations (GEE, Μέθοδος Ισοτήτων Γενικευμένων Εκτιμητριών). Σύμφωνα με τα πλεονεκτήματα και τα μειονεκτήματα της κάθε μεθόδου φαίνεται ότι η WLS είναι λιγότερο κατάλληλη για μια αυτόματη μέθοδο ανάλυσης καθώς παρουσιάζει μερικά προβλήματα τα οποία επιλύονται από την GEE. Τα μειονεκτήματα της GEE δεν είναι τόσο μεγάλα ώστε να μη συνιστάται η χρήση της. Παρουσιάζεται επίσης και ένα παράδειγμα ανάλυσης δεδομένων με τη μέθοδο WLS όταν αυτά θεωρούνται συσχετισμένα και όταν θεωρούνται ασυσχέτιστα για να επιδειχθεί το αποτέλεσμα της λανθασμένης θεώρησης των κατηγορικών δεδομένων ως ανεξάρτητων.

Λέξεις κλειδιά: Επαναλαμβανόμενες μετρήσεις, κατηγορικά δεδομένα, μέθοδος Ισοτήτων Γενικευμένων Εκτιμητριών.

Εισαγωγή

Επαναλαμβανόμενα κατηγορικά δεδομένα εμφανίζονται όταν μια κατηγορική μεταβλητή μετράται σε μια σειρά από υποκείμενα επαναληπτικά πάνω στο χρόνο. Παραδοσιακά, η στατιστική θεωρεί αυτές τις μετρήσεις ως μια συλλογή μετρίων σε μέγεθος κατηγορικών χρονοσειρών. Τέτοια δεδομένα μπορούν να εμφανιστούν σε πολλές επιστήμες μεταξύ των οποίων και η ψυχολογία. Η εφαρμογή ενός ψυχολογικού τεστ με διακριτής κλίμακας απαντήσεις σε μια ομάδα ανθρώπων με κοινά χαρακτηριστικά, καθώς και η παρακολούθησή τους σε τακτά χρονικά διαστήματα, αποτελεί ένα κοινό παράδειγμα για την εμφάνιση τέτοιου τύπου δεδομένων. Ένα άλλο παράδειγμα προέρχεται από το Εργαστήριο μικροανάλυσης του Τμήματος Ψυχολογίας του Πανεπιστημίου της Κρήτης. Ένας αριθμός νηπίων εξετάστηκε στο Εργαστήριο ώστε να διαπιστωθούν οι δυνατότητες μίμησης κάτω από διάφορες εργαστηριακές συνθήκες. Τα αποτελέσματα μίμησης κωδικοποιήθηκαν με τη μορφή κατηγορικών μεταβλητών και περιέγραφαν κινήσεις των

Σημ.: Θα ήθελα να ευχαριστήσω τον κ. Ν. Παπαδόπουλο, Επίκουρο Καθηγητή του Τμήματος Ψυχολογίας του Πανεπιστημίου Κρήτης, για την ευκαιρία που μου έδωσε να παρουσιάσω την εργασία μου στο Συμπόσιο καθώς και την διδάκτορα κα Κ. Πετρουλάκη για τα χρήσιμα σχόλιά της πάνω στο κείμενο.

Διεύθυνση: Βασίλειος Γ. Σ. Βασδέκης, Τμήμα Ψυχολογίας, Πανεπιστήμιο Κρήτης, 741 00 Ρέθυμνο. Τηλ.: 01-9595850, 0831-25595, E-mail: vasdekis@fortezza.cc.ucr.gr

ματιών ή διάφορων μυών του προσώπου των μωρών που πρόδιδαν μίμηση σε ορισμένα ερεθίσματα. Τα μωρά παρατηρήθηκαν σε 9 χρονικές στιγμές και σκοπός της μελέτης μεταξύ άλλων είναι να μοντελοποιηθεί η ανάπτυξη της μίμησης.

Η στατιστική βιβλιογραφία έχει ασχοληθεί με τέτοιου τύπου προβλήματα χρησιμοποιώντας τυπικές μεθόδους όπως αυτή της Μέγιστης Πιθανοφάνειας (Maximum Likelihood, ML) και των Σταθμισμένων Ελαχίστων Τετραγώνων (Weighted Least Squares, WLS). Η φύση όμως των δεδομένων και κυρίως ο περιορισμός της υπολογιστικής ισχύος στη δεκαετία του '70 περιόρισαν την εφαρμογή των τεχνικών αυτών. Πρόσφατα, οι τεχνικές αυτές επαναξιολογήθηκαν και νέα εργαλεία ανάλυσης προτάθηκαν. Ανάμεσα σε πολλές δημοσιεύσεις αναφέρουμε αυτές των Liang και Zeger (1986), Lipsitz, Laird, και Harrington (1991), Liang, Zeger, και Qaqish (1992), Carey, Zeger, και Diggle (1993), οι οποίες εισήγαγαν στην ανάλυση των κατηγορικών δεδομένων τις Ισότητες Γενικευμένων Εκτιμητριών (Generalized Estimating Equations, GEE). Στην παρούσα δημοσίευση θα τεθεί το πρόβλημα της ανάλυσης των επαναλαμβανόμενων κατηγορικών δεδομένων δίνοντας έμφαση στη σημασία των συσχετίσεων που εμφανίζονται σε παρατηρήσεις που λαμβάνονται διαδοχικά πάνω στο χρόνο (Μέρος 2). Θα παρουσιασθούν συνοπτικά τρεις από τις προταθείσες μεθόδους ανάλυσής τους: Η Cochran's Q, η WLS και η GEE. Θα συζητηθούν επίσης τα σχετικά πλεονεκτήματα και μειονεκτήματα της καθεμιάς (Μέρος 3). Τέλος (Μέρος 4), θα δοθεί ένα παράδειγμα ανάλυσης επαναλαμβανόμενων κατηγορικών δεδομένων με τη μέθοδο WLS όταν τα δεδομένα θεωρηθούν εξαρτημένα και όταν θεωρηθούν ανεξάρτητα ώστε να μετρηθεί η επίπτωση της, λανθασμένα, μη θεώρησης συσχετίσεων σε συσχετισμένα δεδομένα.

Δομή του προβλήματος και εξάρτηση εναντίον ανεξαρτησίας

Ας υποθέσουμε ότι διαθέτουμε n υποκείμενα και ότι κάθε υποκείμενο έχει παρατηρήσεις σε q χρονικές στιγμές. Τα δεδομένα για το υποκείμενο i μπορούν να αναπαρασταθούν σαν ένα διάνυσμα όπου κάθε y_{it} είναι μια κατηγορική μεταβλητή με c

$$\mathbf{y}_i=(y_{i1},y_{i2},\ldots,y_{iq}) \qquad (1)$$

κατηγορίες. Ενδεχομένως, τα δεδομένα για κάθε υποκείμενο συνοδεύονται και από έναν πίνακα παρατηρήσεων $X_i=(\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iq})$ που αναπαριστούν επεξηγηματικές μεταβλητές. Αυτές θα χρησιμεύσουν στην κατασκευή ενός μοντέλου που θα προβλέπει κατά το δυνατόν τα δεδομένα. Οι επεξηγηματικές μεταβλητές μπορούν να είναι είτε "χρονικά αναλλοίωτες" όπως το φύλο ή η ηλικία κατά την οποία το άτομο έλαβε αρχικά μέρος στην έρευνα, είτε "χρονικά μεταβλητές" όπως διάφορα ψυχομετρικά χαρακτηριστικά. Πολλές φορές η έλλειψη επεξηγηματικών μεταβλητών οδηγεί τους ερευνητές στη χρήση συναρτήσεων που εξαρτώνται από το χρόνο με σκοπό να μοντελοποιηθεί η εξέλιξη της πιθανότητας επιτυχίας πάνω στο χρόνο. Τέτοιες συναρτήσεις είναι τα πολυώνυμα, τα οποία έχουν χρησιμοποιηθεί σε πολλές περιοχές της ανθρώπινης επιστήμης και αποτελούν πολύ καλές εναλλακτικές λύσεις ακόμα και σε περιπτώσεις όπου διατίθενται επεξηγηματικές μεταβλητές. Συνήθως πολυώνυμα πρώτου και δευτέρου βαθμού αρκούν για να εξηγήσουν ένα μεγάλο μέρος της διασποράς που παρουσιάζουν τα δεδομένα πάνω στο χρόνο.

Η περιθώρια κατανομή των y_{it}, $t=1,\ldots,q$ είναι πολυωνυμική με πιθανότητα επιτυχίας π_{itk} για $k=1,\ldots,c$ και $t=1,\ldots,q$. Οι επεξηγηματικές μεταβλητές συνδέονται με τις πιθανότητες επιτυχίας αφού αυτές οι τελευταίες μετασχηματιστούν. Ο λόγος είναι ότι μια προσπάθεια εξήγησης της πιθανότητας με ένα γραμμικό μοντέλο θα είχε σαν αποτέλεσμα για μερικές τιμές των επεξηγηματικών μεταβλητών να προβλέπεται πιθανότητα επιτυχίας μεγαλύτερη του ένα. Τέτοιοι μετασχηματισμοί είναι οι συναρτήσεις σύνδεσης (δες και McCullagh & Nelder, 1989) όπως η συνάρτηση logit (ή και λογιστικός μετασχηματισμός). Αν υποθέσουμε ότι οι y_{it} είναι δίτιμες μεταβλητές ($c=2$) τότε η πιθανότητα επιτυχίας γίνεται π_{it}. Χρησιμοποιώντας τότε το λογιστικό μετα-

σχηματισμό και για κάθε χρονική στιγμή t

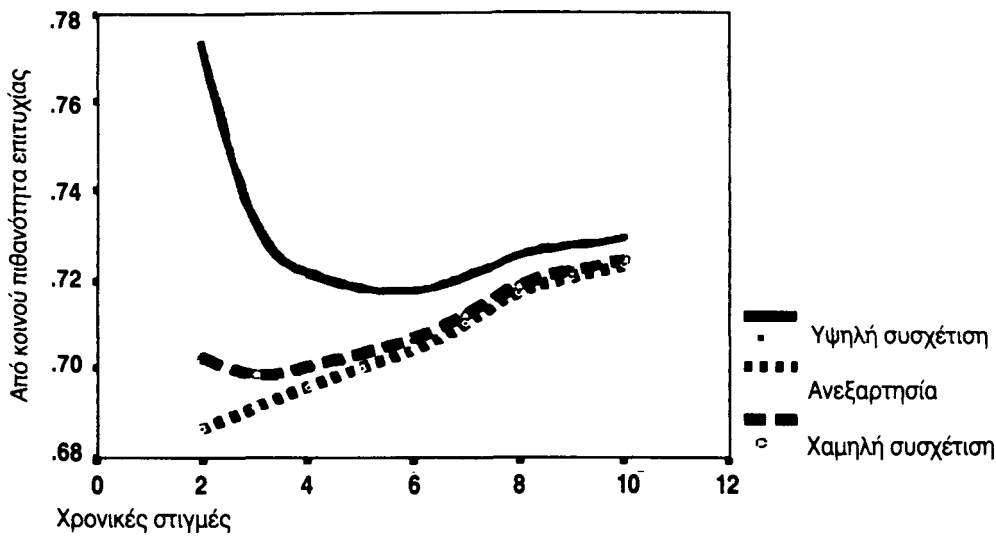
$$\log \frac{\pi_{it}}{1-\pi_{it}} = \mathbf{x}_{it}\boldsymbol{\beta} \qquad (2)$$

όπου είναι ένα διάνυσμα από άγνωστες παραμέτρους που πρέπει να υπολογισθούν. Τότε το μοντέλο που προκύπτει είναι αυτό της λογιστικής παλινδρόμησης. Εννοείται πως στην πράξη μπορεί να χρησιμοποιηθεί οποιαδήποτε συνάρτηση σύνδεσης η οποία ενδεχομένως είναι πιο κατάλληλη για κάποιο τύπο δεδομένων. Στην περίπτωση όπου $c>2$ τότε μια από τις κατηγορίες θεωρείται κατηγορία αναφοράς (ας υποθέσουμε η c) και όλοι οι λογιστικοί μετασχηματισμοί λαμβάνονται με βάση αυτή την κατηγορία, δηλαδή:

$$\log \frac{\pi_{itk}}{\pi_{itc}} = \mathbf{x}_{it}\boldsymbol{\beta}_k \quad k = 1,\ldots,c-1$$

Γίνεται φανερό ότι στην περίπτωση αυτή οι παράμετροι $\boldsymbol{\beta}_k$ προς εκτίμηση είναι διαφορετικές για κάθε κατηγορία k.

Το ενδιαφέρον στις επαναλαμβανόμενες μετρήσεις έρχεται από το γεγονός ότι εξαιτίας της επαναληπτικότητάς τους υπάρχουν συσχετίσεις μεταξύ των παρατηρήσεων κάθε υποκειμένου τις οποίες θα θέλαμε να λάβουμε υπόψη μας κατά τη δόμηση του μοντέλου. Στη στατιστική βιβλιογραφία εικάζεται ότι εαν δεν τις λάβουμε υπόψη, οι εκτιμώμενες διασπορές των παραμέτρων του μοντέλου θα βρίσκονται μακρυά από την πραγματικότητα με αποτέλεσμα τη λανθασμένη συμπερασματολογία. Το Σχήμα 1 δείχνει τι συμβαίνει όταν θεωρούμε τα κατηγορικά δεδομένα ανεξάρτητα πάνω στο χρόνο ενώ στην πραγματικότητα δεν είναι. Για μια δίτιμη κατηγορική μεταβλητή $(c=2)$ αναπαριστάται η από κοινού πιθανότητα επιτυχίας το χρόνο 1 και το χρόνο t, $t=2,\ldots,q$ όταν τα δεδομένα είναι εξαρτημένα και όταν είναι ανεξάρτητα. Τα δεδομένα είναι τεχνητά και προέρχονται από ένα θεωρητικό πείραμα στο οποίο η πιθανότητα επιτυχίας αυξάνεται γραμμικά στο χρόνο. Θεωρήθηκαν 10 χρονικές στιγμές $(q=10)$. Η καμπύλη που σχηματίζεται από τις τελείες αναπαριστά την από κοινού πιθανότητα επιτυχίας όταν τα δεδομένα είναι ανεξάρτητα πάνω στο χρόνο. Η διακεκομμένη και η μη διακεκομμένη καμπύλη αναπαριστούν την ίδια πιθανότητα, όταν υπάρχει μικρή και μεγάλη εξάρτηση μεταξύ των δεδομένων αντίστοιχα. Γίνεται φανερό, ιδίως στην περίπτωση της μεγάλης εξάρτησης, ότι η από κοινού πιθανότητα επιτυχίας στους χρόνους



Από κοινού πιθανότητα επιτυχίας στο χρόνο 1 και στο χρόνο t, $t=2,3,\ldots,10$ σε τρεις περιπτώσεις εξάρτησης μεταξύ των παρατηρήσεων.

1 και t, $t=2, \ldots, q$ ελαττώνεται αντί να αυξάνεται όπως στην περίπτωση της ανεξαρτησίας. Αυτό συμβαίνει μέχρι το χρόνο 6 και από εκεί και πέρα η καμπύλη γίνεται περίπου παράλληλη με αυτή της καμπύλης της περίπτωσης της ανεξαρτησίας. Αυτό συμβαίνει γιατί όσο οι παρατηρήσεις απομακρύνονται πολύ χρονικά μεταξύ τους αρχίζουν και συμπεριφέρονται ως ανεξάρτητες.

Ανασκόπηση μεθόδων ανάλυσης

Cochran's Q. Από τις πρώτες προσπάθειες για την ανάλυση επαναλαμβανόμενων κατηγορικών δεδομένων αποτελεί το Cochran's Q test. Αυτό είναι μια γενίκευση του McNemar's test, το οποίο αποτελεί το αντίστοιχο του ζευγαρωμένου t-test για κατηγορικά δεδομένα. Έτσι το Cochran's Q test είναι μια μονή ANOVA για επαναλαμβανόμενα κατηγορικά δεδομένα. Για κάθε κατηγορία της κατηγορικής μεταβλητής, ελέγχει την ισότητα των περιθωρίων πιθανοτήτων εμφάνισης αυτής της κατηγορίας πάνω στο χρόνο (Agresti, 1989). Εάν, για παράδειγμα, διαθέτουμε μετρήσεις μιας δίτιμης μεταβλητής σε 5 χρονικές στιγμές, τότε το Cochran's Q test ελέγχει την ισότητα των 5 πιθανοτήτων επιτυχίας.

Μέθοδος WLS. Ο έλεγχος Cochran δεν είναι επαρκής γιατί συνήθως ο ερευνητής ενδιαφέρεται όχι μόνο για τη διάγνωση της ισότητας ή όχι των πιθανοτήτων επιτυχίας αλλά κυρίως για τη κατασκευή ενός μοντέλου που θα περιγράφει πώς αλλάζουν με το χρόνο αυτές οι πιθανότητες. Μέχρι τα τέλη της δεκαετίας του 1980 δύο μέθοδοι είχαν προταθεί για τη μοντελοποίηση. Η ML (Μέγιστης Πιθανοφάνειας) και η WLS (Σταθμισμένων Ελαχίστων Τετραγώνων). Η ML προσέγγιση ωστόσο ήταν δύσκολη στην εφαρμογή της, γιατί τα περιθώρια αθροίσματα σε κάθε χρονική στιγμή δεν ακολουθούν ανεξάρτητες μεταξύ τους κατανομές. Παρότι τελευταία αλγόριθμοι εφαρμογής της μεθόδου ML έχουν αρχίσει να εμφανίζονται (Balagtas, Becker, & Lang, 1995) η πολυπλοκότητά τους έστρεψε την έρευνα, εκείνα τα χρόνια, προς την WLS. Αυτή η μέθοδος προτάθηκε από τους Grizzle, Starmer, και Koch (1969). Οι ερευνητές θεώρησαν ότι τα δεδομένα μετατρέπονται από την αρχική μορφή τους (1) σε περιθώρια, δηλαδή κατανεμημένα σε έναν πίνακα συνάφειας του οποίου οι γραμμές διακρίνουν s ανεξάρτητους πληθυσμούς. Οι πληθυσμοί αντιστοιχούν σε όλους τους δυνατούς συνδυασμούς των επεξηγηματικών μεταβλητών (ή επίπεδα παραγόντων). Για παράδειγμα, μπορούμε να έχουμε διαφορετικές ηλικιακές ομάδες, ή φύλα, ή συνδυασμούς ηλικιακών ομάδων με το φύλο κ.ο.κ. Οι στήλες αναπαριστούν r δυνατές κατηγορίες μιας πολυωνυμικής κατανομής οι οποίες δομούνται από όλους τους δυνατούς συνδυασμούς των επιπέδων της κατηγορικής μεταβλητής με το χρόνο. Έτσι $r=c^q$ όπου c είναι ο αριθμός των επιπέδων της κατηγορικής μεταβλητής και q ο αριθμός των χρονικών στιγμών που γίνονται οι παρατηρήσεις. Τότε σε κάθε συνδυασμό i πληθυσμού και j κατηγορίας μεταβλητής έχουμε συχνότητες n_{ij} όπως φαίνεται και στον Πίνακα 1.

Συμβολίζουμε με $\mathbf{p}_i^T = (p_{i1},\ldots,p_{ir})$, $i = 1,\ldots,s$ τις πιθανότητες εμφάνισης καθεμιάς από τις r κατηγορίες για τους s ανεξάρτητους πληθυσμούς. Ομαδοποιούμε τις πιθανότητες αυτές σε έναν πίνακα $P=(\mathbf{p}_1,\ldots,\mathbf{p}_s)$. Θεωρούμε τότε ότι υπάρχει ένα διάνυσμα συναρτήσεων $\mathbf{F}^T(P)=(f_1(P),\ldots,f_u(P))$:

$$\mathbf{F}(P)=A\gamma \qquad (3)$$

όπου A $(u \times v)$ είναι ένας πίνακας σχεδιασμού και γ διάνυσμα $v \times 1$ αγνώστων παραμέτρων προς εκτίμηση. Ένα παράδειγμα τέτοιων συναρτήσεων μπορεί να είναι ο λογιστικός μετασχηματισμός και το μοντέλο η επέκταση της λογιστικής παλινδρόμησης σε επαναλαμβανόμενα κατηγορικά δεδομένα. Ελαχιστοποιώντας την τετραγωνική μορφή:

$$(\mathbf{F}(P) - A\gamma)^T S^{-1} (\mathbf{F}(P) - A\gamma) \qquad (4)$$

όπου S είναι ο ασυμπτωτικός πίνακας διασπορών της $\mathbf{F}(P)$, λαμβάνουμε εκτιμήτριες WLS για τους συντελεστές γ του μοντέλου. Έλεγχοι καλής προσαρμογής καθώς και διάφοροι έλεγχοι για το γ κατασκευάζονται εύκολα σύμφωνα με τη θεωρία των γραμμικών μοντέλων.

Πίνακας 1
Πρότυπο πίνακα εμφάνισης δεδομένων για την εφαρμογή της μεθόδου WLS

	Κατηγορίες Μεταβλητής				
Ανεξάρτητοι Πληθυσμοί	1	2	.....	r	Περιθώρια Αθροίσματα
1	n_{11}	n_{12}	.....	n_{1r}	$n_{1.}$
2	n_{21}	n_{22}	.....	n_{2r}	$n_{2.}$
.	.	.....	.	.	
.	.	.....	.	.	
.	.	.....	.	.	
s	n_{s1}	n_{s2}	.....	n_{sr}	$n_{s.}$

Παράδειγμα 1: Εναλλακτικός έλεγχος Cochran (Grizzle, Starmer, & Koch, 1969)

Ας υποθέσουμε, χάριν απλότητος, ότι $c=2$ και ότι διαθέτουμε μετρήσεις σε 3 χρονικές στιγμές. Τότε σύμφωνα με τα προηγούμενα υπάρχουν 8 συνδυασμοί ενδεχομένων $(r=8)$ οι οποίοι μαζί με τις πιθανότητες εμφάνισής τους δίνονται από τον Πίνακα 2.

Ας υποθέσουμε ότι θέλουμε να ελέγξουμε εάν κατά μέσον όρο λαμβάνουμε τα ίδια αποτελέσματα σε κάθε χρονικό σημείο (ή αλλιώς ότι η πιθανότητα εμφάνισης επιτυχίας δεν αλλάζει με το χρόνο). Ο έλεγχος αυτός αποτελεί ανάλογο του ελέγχου Cochran's Q. Από τον πίνακα φαίνεται ότι η πιθανότητα επιτυχίας στα τρία χρονικά σημεία είναι

$$\pi_{i1} = p_1 + p_2 + p_3 + p_5$$
$$\pi_{i2} = p_1 + p_2 + p_4 + p_6$$
$$\pi_{i3} = p_1 + p_3 + p_4 + p_7$$

Ο έλεγχος τότε της ισότητας γίνεται μια συνάρτηση της μορφής που δόθηκε από τη συνάρτηση **F** παραπάνω.

$$p_1+p_2+p_3+p_5=p_1+p_2+p_4+p_6=p_1+p_3+p_4+p_7$$

ή αλλιώς

$$p_2-p_1=p_4-p_5=p_3-p_6$$

Μπορούμε τότε να εκφράσουμε την υπόθεση που μας ενδιαφέρει με δύο συναρτήσεις $f_1(P)$ και $f_2(P)$ οι οποίες είναι

$$f_1(P)=p_2-p_1-p_4+p_5,\ f_2(P)=p_2-p_1-p_3+p_6$$

Για αυτές τις συναρτήσεις το γραμμικό μοντέλο (3) γίνεται

$$\begin{pmatrix} f_1(P) \\ f_2(P) \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}$$

και ελέγχεται εάν

$$\gamma_1 = \gamma_2 = 0$$

Παράδειγμα 2. Λογιστική παλινδρόμηση για επαναλαμβανόμενα κατηγορικά δεδομένα

Όπως διατυπώθηκε και στο Μέρος 2, ισότητα (2), η λογιστική παλινδρόμηση χρησιμοποιεί το λογιστικό μετασχηματισμό είτε τα δεδομένα είναι εξαρτημένα είτε είναι ανεξάρτητα. Η διαφορά μεταξύ των δύο περιπτώσεων είναι ότι στην εξαρτημένη περίπτωση ο πίνακας S της ισότητας (4) έχει στοιχεία σε κάθε κελλί του, ενώ στην περίπτωση της ανεξαρτησίας υπάρχουν μόνο μονάδες στην κεντρική διαγώνιο και τα υπόλοιπα στοιχεία είναι μηδέν. Κατά τα άλλα και υποθέτοντας ότι μετράται μια δίτιμη κατηγορική μεταβλητή σε τρία χρονικά σημεία, όπως ακριβώς και

Πίνακας 2
Όλα τα δυνατά ενδεχόμενα στην περίπτωση μιας δίτιμης εξαρτημένης μεταβλητής (c=2) όταν αυτή μετράται σε τρεις χρονικές στιγμές (q=3)

Κατηγορίες (Ενδεχόμενα)			Πιθανότητα
1	1	1	p_1
1	1	0	p_2
1	0	1	p_3
0	1	1	p_4
1	0	0	p_5
0	1	0	p_6
0	0	1	p_7
0	0	0	p_8

στο προηγούμενο παράδειγμα, καθώς και ότι δεν έχουμε στη διάθεσή μας άλλες επεξηγηματικές μεταβλητές εκτός από το χρόνο (δηλαδή $s=1$ στον Πίνακα 1), έχουμε τους παρακάτω ορισμούς για τη συνάρτηση $\mathbf{F}(P)$:

$$\begin{pmatrix} f_1(P) \\ f_2(P) \\ f_3(P) \end{pmatrix} = A$$

όπου

$$f_1(P)=\log\frac{p_1+p_2+p_3+p_5}{p_4+p_6+p_7+p_8},$$
$$f_2(P)=\log\frac{p_1+p_2+p_4+p_6}{p_3+p_5+p_7+p_8},$$
$$f_3(P)=\log\frac{p_1+p_3+p_4+p_7}{p_2+p_5+p_6+p_8}$$
$$A=\begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} \quad \gamma=\begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}$$

Η μέθοδος WLS είναι εύκολη εννοιολογικά, όμως έχει κάποια μειονεκτήματα τα οποία την κάνουν ακατάλληλη ως μια αυτόματη μέθοδο.

α) Όταν οι επεξηγηματικές μεταβλητές είναι συνεχείς ή είναι πάρα πολλές τότε υπάρχει κίνδυνος κάποιοι συνδυασμοί τους να μην έχουν συχνότητες και έτσι η μέθοδος να είναι ασταθής. Τότε, ουσιαστικά, δεν εφαρμόζεται.

β) Οι μετρήσεις πρέπει όλες να βρίσκονται στα ίδια χρονικά σημεία, κάτι το οποίο δεν είναι εγγυημένο στις έρευνες των οποίων οι μετρήσεις είναι επαναλαμβανόμενες.

γ) Όταν ο στόχος είναι η μοντελοποίηση των πιθανοτήτων στο χρόνο και η μεταβλητή που μετράται έχει c κατηγορίες, τότε δημιουργούνται c^q κατηγορίες της πολυωνυμικής μεταβλητής και ο αριθμός αυτός γίνεται τεράστιος με προφανείς επιπτώσεις στην εφαρμογή της μεθόδου (π.χ. $3^5=243$).

δ) Η μέθοδος είναι περιθώρια, με την έννοια ότι εκτιμά και αξιοποιεί την περιθώρια κατανομή των δεδομένων σε κάθε χρονική στιγμή. Ωστόσο η από κοινού κατανομή των περιθωρίων αθροισμάτων είναι μάλλον πολύπλοκη και έτσι προκύπτουν θεωρητικές δυσκολίες στην εκτίμηση των παραμέτρων του μοντέλου (Balagtas et al., 1995).

Μέθοδος Ισοτήτων Γενικευμένων Εκτιμητριών

Μια εναλλακτική μέθοδος άρχισε να εφαρμόζεται από τα τέλη της προηγούμενης δεκαετίας. Αυτή βασίζεται στην απευθείας μοντελοποίηση της *συσχέτισης* μεταξύ των παρατηρήσεων με κάποια συνάρτηση και στη χρησιμοποίηση κάποιου κριτηρίου για την εκτίμηση των παραμέτρων

του μοντέλου. Επειδή η μέθοδος αυτή παραβλέπει την πιθανοφάνεια των παρατηρήσεων δεν είναι απόλυτα παραμετρική. Βασίζεται πάνω σε αποτελέσματα που προέκυψαν από τη μελέτη των γενικευμένων γραμμικών μοντέλων την περασμένη δεκαετία και ειδικότερα από τη μελέτη των ημι-πιθανοφανιακών συναρτήσεων. Οι συναρτήσεις αυτές δίνουν εκτιμήτριες οι οποίες έχουν ασυμπτωτικές ιδιότητες κοντινές προς αυτές των εκτιμητριών πιθανοφάνειας και το βασικότερο πλεονέκτημά τους είναι ότι δε χρειάζεται να έχει κανείς ακριβή ιδέα για την κατανομή των παρατηρήσεων. Τέτοιες συναρτήσεις χρησιμοποιήθηκαν από τους Liang και Zeger (1986) καθώς και Prentice (1988). Η μεθοδολογία ονομάστηκε GEE (Generalized Estimating Equations, Ισότητες Γενικευμένων Εκτιμητριών).

Η εφαρμογή της τεχνικής αυτής στα επαναλαμβανόμενα κατηγορικά δεδομένα απαιτεί τη μοντελοποίηση της συσχέτισης μεταξύ των παρατηρήσεων. Κλειδί για τη μοντελοποίηση αυτή είναι η έννοια που θα χρησιμοποιηθεί για να εκφραστεί αυτή η συσχέτιση. Όταν οι παρατηρήσεις είναι συνεχείς, η κατά Pearson συσχέτιση είναι επαρκής. Στην περίπτωση όμως των ονομαστικών κατηγορικών δεδομένων η κατά Pearson συσχέτιση δεν έχει νόημα εκτός εάν και οι δύο κατηγορικές μεταβλητές είναι δίτιμες. Έτσι χρειάζονται άλλες έννοιες με τις οποίες θα εκφραστεί η συσχέτιση των δεδομένων. Μια από αυτές, που είναι και η επικρατέστερη στη βιβλιογραφία (Agresti, 1989) είναι ο λόγος των πιθανοτήτων μεταξύ των παρατηρήσεων στις χρονικές στιγμές s και t.

Ας θεωρήσουμε, για ευκολία, ότι έχουμε παρατηρήσεις από δίτιμες μεταβλητές ($c=2$). Θυμίζουμε τότε ότι οι παρατηρήσεις έχουν τη μορφή που δίνεται από την (1) και οι πιθανότητες επιτυχίας σε κάθε χρονική στιγμή δίνονται από ένα διάνυσμα $\boldsymbol{\pi}_i$ για κάθε υποκείμενο i, $i=1, \ldots, n$ ως εξής:

$$\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{iq}) \qquad (5)$$

Τότε, ο λόγος αυτός δίνεται από την (6)

$$\psi_{ist} = \frac{\pi_{ist}(1 - \pi_{is} - \pi_{it} + \pi_{ist})}{(\pi_{is} - \pi_{ist})(\pi_{it} - \pi_{ist})} \qquad (6)$$

όπου π_{ist} είναι η από κοινού πιθανότητα επιτυχίας και στις δύο χρονικές στιγμές για το υποκείμενο i. Όπως φαίνεται και από τον Πίνακα 3, η ποσότητα ψ_{ist} σχηματίζεται από το γινόμενο των από κοινού πιθανοτήτων επιτυχίας και αποτυχίας και από το γινόμενο των υπόλοιπων πιθανοτήτων του πίνακα συνάφειας. Εάν υπάρχει ανεξαρτησία μεταξύ των δύο χρονικών στιγμών, τότε οι πιθανότητες επιτυχίας στο χρόνο t είτε υπήρξε επιτυχία το χρόνο s είτε δεν υπήρξε (π_{ist} και $\pi_{is} - \pi_{ist}$ αντίστοιχα) θα είναι ίσες. Το ίδιο θα συμβαίνει και με τις πιθανότητες αποτυχίας του χρόνου t. Εύκολα μπορεί κανείς να δει ότι στην περίπτωση αυτή $\psi_{ist} = 1$. Έτσι απόκλιση της τιμής του ψ_{ist} από την τιμή 1 δίνει και ένα μέτρο της συσχέτισης μεταξύ των δεδομένων.

Είναι λογικό να υποθέσουμε πως αυτός ο λόγος μεταβάλλεται σύμφωνα με κάποια συναρτησιακή μορφή έτσι ώστε καθώς απομακρύνονται τα χρονικά σημεία να έχουμε και λιγότερο εξαρτημένες παρατηρήσεις. Μια τέτοια συναρτησιακή μορφή θα μπορούσε να εξαρτάται από τις χρονικές στιγμές s και t κατά πολλούς τρόπους, όπως, για παράδειγμα, από τη διαφορά τους, s-t καθώς δείχνει και η παρακάτω ισότητα:

$$\psi_{ist} = \alpha^{|s-t|} \qquad (7)$$

(Fitzmaurice & Lipsitz, 1995), όπου α είναι μια παράμετρος εκτιμήσιμη από τα δεδομένα. Είναι εύκολα παρατηρήσιμο ότι όσο απομακρύνονται οι δύο χρονικές στιγμές τότε το ψ_{ist} τείνει στο 1 και οι παρατηρήσεις είναι ανεξάρτητες.

Από τα παραπάνω γίνεται κατανοητό ότι το μοντέλο που υιοθετείται προς ανάλυση δίνεται όσον αφορά τις παραμέτρους που θα αφορούν τις πιθανότητες επιτυχίας από την (2) και όσον αφορά τις παραμέτρους που αφορούν τη συσχέτιση μεταξύ των παρατηρήσεων στο χρόνο από συναρτησιακές μορφές όπως η (7). Άρα ο αριθμός των παραμέτρων προς εκτίμηση ισούται προς το μέγεθος του διανύσματος $\boldsymbol{\beta}$ της ισότητας (2) αυξημένου κατά 1 ([επειδή θεωρήσαμε

Πίνακας 3
Υπόδειγμα από κοινού πιθανοτήτων των δυνατών ενδεχομένων μιας δίτιμης μεταβλητής (c=2) μεταξύ δύο χρονικών στιγμών

	Χρόνος t		
Χρόνος s	1	0	
1	π_{st}	π_s-π_{st}	π_s
0	π_t-π_{st} π_t+π_{st}	1-π_s-	1-π_s
	π_t	1-π_t	1

στην ισότητα (7) ότι η συσχέτιση των δεδομένων περιγράφεται από την παράμετρο *α*]). Η μέθοδος GEE χρησιμοποιεί δύο πολυμεταβλητές ισότητες. Η πρώτη ισότητα εξαρτάται από το **β** και τις παραμέτρους της συσχέτισης και μπορεί να εκτιμήσει τόσες παραμέτρους όσες είναι και οι παράμετροι **β**. Αυτό συμβαίνει γιατί οι πιθανότητες επιτυχίας εξαρτώνται από τις παραμέτρους **β** σύμφωνα με το μοντέλο (2). Η δεύτερη ισότητα εξαρτάται από τις ίδιες παραμέτρους από τις οποίες εξαρτάται και η πρώτη ισότητα και μπορεί να εκτιμήσει τόσες παραμέτρους όσες είναι και οι παράμετροι της συσχέτισης.

Ξεκινώντας με αρχικές τιμές για τις παραμέτρους της συσχέτισης εκτιμάται το **β** από την πρώτη ισότητα και με τη σειρά, εκτιμάται η παράμετρος της συσχέτισης από τη δεύτερη ισότητα. Η νέα αυτή εκτίμηση της παραμέτρου της συσχέτισης χρησιμοποιείται για μια νέα εκτίμηση των παραμέτρων **β** κ.ο.κ. Η διαδικασία αυτή συνεχίζεται μέχρι τη σύγκλιση. Οι εκτιμήτριες που προκύπτουν έχουν καλές ασυμπτωτικές ιδιότητες.

Ενδιαφέρον έχει η επέκταση του μοντέλου σε κατηγορικά δεδομένα με πολλές κατηγορίες. Στην περίπτωση αυτή θεωρούμε μια από τις κατηγορίες της μεταβλητής ως "κατηγορία αναφοράς" και παίρνουμε όλους τους λόγους πιθανοτήτων (6) με βάση αυτή την κατηγορία. Μπορεί να δειχθεί ότι οι λόγοι πιθανοτήτων που είναι απαραίτητοι για να προσδιορίσουν τη συσχέτιση των δεδομένων είναι $(c-1)^2$ (Agresti, 1990), όπου *c* είναι ο αριθμός των κατηγοριών της κατηγορικής μεταβλητής. Στην περίπτωση αυτή ισχύουν όλα τα παραπάνω μόνο που ο αριθμός των ισοτήτων είναι μεγαλύτερος. Η μέθοδος GEE έχει πολλά πλεονεκτήματα και κυριότερα είναι:

α) Αντιμετωπίζει επιτυχώς όλα τα μειονεκτήματα της WLS. Οι επεξηγηματικές μεταβλητές μπορούν να είναι είτε συνεχείς είτε κατηγορικές και οι μετρήσεις δεν είναι αναγκαίο να βρίσκονται στα ίδια χρονικά σημεία, κάτι το οποίο θεωρείται δύσκολα πραγματοποιήσιμο στην πράξη.

β) Έχει καλές ασυμπτωτικές ιδιότητες.

Ένα από τα μειονεκτήματά της είναι τεχνικής φύσης. Η επιλογή της δεύτερης ισότητας που θα δώσει εκτιμήσεις για την παράμετρο που χαρακτηρίζει τη συσχέτιση μεταξύ των παρατηρήσεων παρουσιάζει κάποιες θεωρητικές δυσκολίες μια και χρησιμοποιούνται ποσότητες οι οποίες δεν είναι ανεξάρτητες μεταξύ τους ενώ εκλαμβάνονται ως τέτοιες από τη μέχρι πρόσφατα εφαρμογή της μεθόδου (Fitzmaurice & Lipsitz, 1995). Ανοικτό παραμένει ακόμα το ερώτημα του προσδιορισμού των συνεπειών αυτής της πρακτικής στη συμπερασματολογία.

Παράδειγμα ανάλυσης επαναλαμβανόμενων κατηγορικών δεδομένων

Τα παρακάτω δεδομένα αναλύθηκαν από τους Koch et al. (1977). Αναφέρονται σε μια διαχρονική μελέτη η οποία συγκρίνει ένα νέο και ένα τυπικό φάρμακο για τη θεραπεία ασθενών που πάσχουν από κατάθλιψη. Τα υποκείμενα χωρί-

σθηκαν σε δύο ομάδες διάγνωσης ανάλογα με το βαθμό της κατάθλιψης (ήπια, βαριά) από την οποία υπέφεραν. Σε κάθε διαγνωστική ομάδα, τα υποκείμενα χωρίσθηκαν κατά τυχαίο τρόπο σε δύο ομάδες περίθαλψης (τυπικό και νέο φάρμακο). Οι ασθενείς παρακολουθήθηκαν σε τρεις χρονικές στιγμές (1, 2 και 4 εβδομάδες από την αρχή της έρευνας). Κάθε χρονική στιγμή γινόταν και μια μέτρηση η οποία χαρακτήριζε την κατάσταση του ασθενούς ως κανονική ή μη κανονική.

Από τα παραπάνω γίνεται φανερό ότι $c=2$ και $q=3$. Οι επεξηγηματικές μεταβλητές που χρησιμοποιήθηκαν ήταν η αρχική διάγνωση κατάθλιψης (κατηγορική), η περίθαλψη (κατηγορική) και οι χρονικές περίοδοι (συνεχής). Οι Koch et al. (1977) χρησιμοποίησαν τη μέθοδο WLS. Το βέλτιστο μοντέλο μαζί με τους συντελεστές και τις τυπικές αποκλίσεις τους δίδονται στο πρώτο μέρος του Πίνακα 4.

Τα δεδομένα αναλύθηκαν και ως να ήταν ανεξάρτητα θέτοντας δηλαδή στην ισότητα (4) τον πίνακα S ίσο με τον ταυτοτικό (τα στοιχεία της διαγωνίου ίσα με 1 και αυτά εκτός διαγωνίου ίσα με 0). Η ανάλυση αυτή παρουσιάζεται στο δεύτερο μέρος του Πίνακα 4. Τα αποτελέσματα δείχνουν ότι ο χρόνος συμβάλλει θετικά στην αύξηση της πιθανότητας κανονικής κατάστασης της κατάθλιψης και ότι υπάρχει στατιστικώς σημαντική αλληλεπίδραση μεταξύ του χρόνου και των επιπέδων περίθαλψης. Το αρνητικό πρόσημο σημαίνει ότι με το νέο φάρμακο είχαμε μεγαλύτερη αύξηση της πιθανότητας για κανονική κατάσταση της κατάθλιψης. Η ύπαρξη στατιστικώς σημαντικής αλληλεπίδρασης εξηγεί και τη φαινομενική μη σημαντικότητα του παράγοντα περίθαλψη.

Η ανάλυση που έγινε όταν τα δεδομένα θεωρήθηκαν ανεξάρτητα μεταξύ τους έδωσε παρόμοια αποτελέσματα για τους συντελεστές του μοντέλου (με την έννοια ότι οι διαφορές δεν είναι τόσο συστηματικές) και μικρές αλλά συστηματικές διαφορές μεταξύ των εκτιμώμενων τυπικών αποκλίσεων των συντελεστών. Το φαινόμενο είναι ιδιαίτερα έντονο στο συντελεστή του χρόνου. Συμπερασματικά, το αποτέλεσμα της λανθασμένης θεώρησης των επαναλαμβανόμενων κατηγορικών δεδομένων ως ανεξάρτητων, φαίνεται να είναι η εκτίμηση μικρών, αριθμητικά, τυπικών αποκλίσεων των παρα-

Πίνακας 4
Αποτελέσματα ανάλυσης δεδομένων όταν αυτά θεωρήθηκαν εξαρτημένα (Koch et al., 1977) και όταν θεωρήθηκαν ανεξάρτητα ("ανεξάρτητη" ανάλυση)

Αποτελέσματα από Koch et al. (1977)	Συντελεστής	Τυπ. απόκλιση
Σταθερά	-1.382	0.185
Αρχική διάγνωση κατάθλιψης	1.282	0.146
Περίθαλψη	0.052	0.228
Χρόνος	1.474	0.154
Περίθαλψη x Χρόνος	-0.994	0.192

Αποτελέσματα "ανεξάρτητης" ανάλυσης	Συντελεστής	Τυπ. απόκλιση
Σταθερά	-1.993	0.197
Αρχική διάγνωση κατάθλιψης	0.845	0.143
Περίθαλψη	0.484	0.226
Χρόνος	1.081	0.097
Περίθαλψη x Χρόνος	-1.252	0.192

μέτρων του μοντέλου έτσι ώστε να δίνεται μια σχετικώς ψευδής εικόνα ακρίβειας της εκτίμησης.

Βιβλιογραφία

Agresti, A. (1989). *Categorical data analysis.* New York: Wiley.

Balagtas, C. C., Becker, M. P., & Lang, J. B. (1995). Marginal modelling of categorical data from crossover experiments. *Applied Statistics, 44,* 63-78.

Carey, V., Zeger, S. L., & Diggle, P. J. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika, 80,* 517-526.

Fitzmaurice, G. M., & Lipsitz, S. R. (1995). A model for binary time series data with serial odds ratio patterns. *Applied Statistics, 44,* 51-61.

Grizzle, J. E., Starmer, C. F., & Koch, G. G. (1969). Analysis of categorical data by linear models. *Biometrics, 25,* 489-504.

Koch, C. G., Landis, J. R., Freeman, J. L., Freeman, D. H., & Lehnen, R. G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics, 33,* 133-158.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73,* 13-22.

Liang, K. Y., Zeger, S. L., & Qaqish, B. (1992). Multivariate regression analyses for categorical data (with discussion). *Journal of the Royal Statistical Society B, 54,* 3-40.

Lipsitz, S. R., Laird, N. M., & Harrington, D. P. (1991). Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika, 78,* 153-160.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models.* (2nd ed). London: Chapman and Hall.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics, 44,* 1033-1048.

Models for the analysis of repeated categorical measurements

VASSILIOS G. S. VASDEKIS
University of Crete, Greece

ABSTRACT The analysis of categorical data is of special interest for psychology since such data appear frequently in research. Repeated categorical data need special treatment since these are realizations of categorical variables taken repeatedly in various time intervals on a sample of subjects. Measurements of this type have correlations and the meaning of the existence of correlations is examined. Three methods (two old and one recently introduced) are reviewed: Cochran's Q, Weighted Least Squares (WLS) and Generalized Estimation Equations (GEE). According to each method's properties, WLS seems to be a less appropriate automatic method for the analysis since some of the problems appearing in the application of the method are resolved by GEE. Finally, an example of data analysis using WLS is presented. Data are analysed as being correlated and uncorrelated. Differences in the estimation of standard errors of the coefficients of the model are pointed out.

Keywords: Kategorical data, Generalized Estimation Equations Method, repeated measurements.

Address: Vassilios G. S. Vasdekis, Department of Psychology, University of Crete, 741 00 Rethymno, Greece. Tel.: *30-1-9595850, *30-831-25595, E-mail: vasdekis@fortezza.cc.ucr.gr