ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΑΝΑΛΥΣΗ ΕΠΑΝΑΛΑΜΒΑΝΟΜΕΝΩΝ ΔΙΑΤΑΞΙΜΩΝ ΔΙΑΚΡΙΤΩΝ ΔΕΔΟΜΕΝΩΝ

Δέσποινα Χ. Καρατίσογλου

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Δεκέμβριος 2005

Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΑΝΑΛΥΣΗ ΕΠΑΝΑΛΑΜΒΑΝΟΜΕΝΩΝ ΔΙΑΤΑΞΙΜΩΝ ΔΙΑΚΡΙΤΩΝ ΔΕΔΟΜΕΝΩΝ

Δέσποινα Χ. Καρατίσογλου

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Δεκέμβριος 2005

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ..... συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Μ. Κατέρη (Επιβλέπων)
- Μ. Κούτρας
- Γ. Πιτσέλης

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN APPLIED STATISTICS

ANALYSIS OF REPEATED ORDINAL DISCRETE DATA

By

Despina C. Karatisoglou

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfillment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
December 2005

Στους γονείς μου
Χρήστο και Μαρία
k *τον αδερφό μου*
Στάθη

Ευχαριστίες

Αισθάνομαι την ανάγκη να ευχαριστήσω θερμά όσους βοήθησαν στην ολοκλήρωση της συγκεκριμένης εργασίας. Πρώτα από όλους, θέλω εγκάρδια να ευχαριστήσω την κυρία Κατέρη Μαρία, Επίκουρη Καθηγήτρια του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς, για την πολύτιμη βοήθεια, την υπομονή που έδειξε και την ηθική συμπαράσταση της σε όλη τη διάρκεια όχι μόνο αυτής της εργασίας αλλά και των σπουδών μου στο εν λόγω τμήμα γενικότερα.

Ευχαριστώ τους κυρίους Κούτρα Μάρκο, Καθηγητή του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς, και Πιτσέλη Γεώργιο, Λέκτορα του του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς για την συμμετοχή τους στην επιτροπή και τις παρατηρήσεις τους.

Επίσης, θέλω να ευχαριστήσω την υποψήφια διδάκτορα Ελευθεράκη Αναστασία για τον χρόνο που μου αφιέρωσε, τις συμβουλές και την βοήθειά της σε οποιοδήποτε πρόβλημα προέκυψε. Ευχαριστώ πολύ τις συμφοιτήτριες Μπουγάτσα Ζαχαρούλα και Παπά Φραντζέσκα για την ηθική στήριξή τους καθώς επίσης και την φίλη μου Κάπαρη Κατερίνα. Τέλος, θέλω από καρδιάς να εκφράσω την ευγνωμοσύνη μου και την αγάπη μου προς την οικογένειά μου και ειδικότερα τους γονείς μου Χρήστο και Μαρία που προσφέρουν τα μέγιστα εδώ και χρόνια για να καταφέρω όσα ονειρεύτηκα.

Καρατίσογλου Δέσποινα

Δεκέμβριος 2005

Περίληψη

Οι επαναλαμβανόμενες μετρήσεις αποτελούν το κύριο είδος δεδομένων σε εφαρμογές πολλών επιστημών και ιδιαίτερα σε βιοϊατρικές έρευνες. Για το λόγο αυτό, οι μέθοδοι ανάλυσής τους βρίσκονται στο επίκεντρο του ενδιαφέροντος εδώ και πολλά χρόνια. Οι περισσότεροι ερευνητές, ωστόσο, έχουν ασχοληθεί κατά κύριο λόγο με την περίπτωση όπου η κύρια μεταβλητή απόκρισης που καταγράφεται και αποτελεί τον πυρήνα κάθε έρευνας είναι συνεχής. Κατ' επέκταση, έχουμε στη διάθεσή μας ένα πλήρες θεμελιωμένο θεωρητικό υπόβαθρο για τον συγκεκριμένο τύπο μεταβλητής.Τις δύο τελευταίες δεκαετίες γίνονται προσπάθειες προκειμένου η θεωρία να επεκταθεί και για διακριτές αποκρίσεις και ειδικότερα διατάξιμες.

Στην παρούσα εργασία ασχοληθήκαμε με την ανάλυση επαναλαμβανόμενων μετρήσεων διατάξιμων δεδομένων. Αρχικά, αναφερόμαστε στα χαρακτηριστικά αυτών, τα θετικά και τα αρνητικά τους σημεία και θέτουμε τα ερωτήματα και τους προβληματισμούς που τις διέπουν. Στη συνέχεια, παρουσιάζουμε τις δύο πιο δημοφιλείς μεθόδους για την ανάλυση των διαχρονικών δεδομένων, τα περιθώρια και τα μοντέλα τυχαίων επιδράσεων. Παρουσιάζονται, επίσης, οι επεκτάσεις που έχουν προταθεί για την αντιμετώπιση της διατάξιμης φύσης της απόκρισης. Καθώς, ένα από τα μείζοντα προβλήματα με αυτά τα δεδομένα εστιάζεται στην επιλογή του προγράμματος που θα διεκπεραιώσει την ανάλυση, κάνουμε μια προσπάθεια να καταγράψουμε το σύνολο των πιο ευέλικτων διαθέσιμων λογισμικών και στατιστικών προγραμμάτων. Για την πλήρη κατανόηση του τρόπου λειτουργίας μερικών από αυτών, προχωρήσαμε στην ανάλυση δύο συνόλων δεδομένων και στην ερμηνεία των αποτελεσμάτων τους.

– *Abstract* —

Repeated measures constitute the main type of data for many fields of science and especially for biomedical studies. This is why they gain great interest for many years. Most of the researchers, however, have studied mainly the case where the principal repeated variable of the study (the response variable) is continuous. Therefore, a totally valid theoretical background for this type of variables is available. The past two decades a lot of effort has been done in order to extend theory to discrete responses and mostly to ordinal ones.

In this project we review the analysis of longitudinal studies with repeated measures of ordinal data. At the beginning, we report their features, the advantages and disadvantages and we express the questions and problems that arise in this context. Next, we present the two more popular methods for the analysis of longitudinal data, namely the marginal and random-effects models. The extensions that have been proposed for handling the ordinal nature of the response are presented as well. One of the major problems with this type of data is focused on the choice of software that is going to perform the analysis. Thus we try to list the more flexible available software and statistical programs. Finally, in order to fully understand the way some of them work, we proceeded to the analysis of two datasets and the interpretation of their results.

e *Περιεχόμενα* f

Κεφάλαιο 1ο
Εισαγωγή

1.1 Διαχρονικές (longitudinal) μελέτες

Ο όρος "διαχρονική μελέτη" χρησιμοποιείται για τον χαρακτηρισμό μιας μελέτης κατά την οποία για κάθε πειραματική μονάδα η μεταβλητή που μας ενδιαφέρει, και στη συνέχεια θα αναφέρεται ως απόκριση ή εξαρτημένη μεταβλητή, παρατηρείται σε δύο ή περισσότερες περιπτώσεις. Κατά συνέπεια, το κύριο χαρακτηριστικό ενός διαχρονικού συνόλου δεδομένων είναι οι επαναλαμβανόμενες μετρήσεις για κάθε πειραματική μονάδα. Τέτοιου είδους μετρήσεις λαμβάνονται είτε χρονικά είτε υπό διαφορετικές πολλαπλές συνθήκες. Πολλοί συγγραφείς θεωρούν τα διαχρονικά δεδομένα ως ειδική κατηγορία των επαναλαμβανόμενων, όταν ο παράγοντας που τις καθορίζει είναι ο χρόνος. Δεν πρέπει να συγχέονται με τις χρονοσειρές, αφού οι διαχρονικές παρατηρήσεις καταγράφονται σε πιο βραχυπρόθεσμο χρονικό πλαίσιο από ότι οι χρονοσειρές.

1.2 Πεδία εφαρμογής

Οι διαχρονικές μελέτες έχουν κερδίσει σημαντικό έδαφος σε αρκετά ερευνητικά πεδία, όπως η γεωργία, οι κοινωνικές και φυσικές επιστήμες, η ιατρική κτλ. Λόγου χάρη:

- Στη γεωργία πραγματοποιούνται έρευνες για την ανάπτυξη των φυτών. Τα φυτά χωρίζονται σε ομάδες θεραπείας στην αρχή της αναπτυσσόμενης περιόδου και εν συνεχεία λαμβάνονται εβδομαδιαίες, συνήθως, παρατηρήσεις ενός συγκεκριμένου μέτρου ανάπτυξης.
- Ομοίως, σε μια κλινική δοκιμή σε άτομα πάσχοντα από κάποια ασθένεια ανατίθενται διαφορετικές αγωγές κατά την έναρξη. Ανά τακτά χρονικά διαστήματα λαμβάνονται μετρήσεις της μεταβλητής απόκρισης (π.χ. πίεση του αίματος σε υπερτασικούς ασθενείς) και έτσι, με την βοήθεια ενός τέτοιου σχεδιασμού, οι επιστήμονες είναι σε θέση να

προσδιορίσουν αν κάποια από τις αγωγές επιφέρει σημαντικές αλλαγές χρονικά στην κλινική κατάσταση των ασθενών. Σημειώνουμε ότι στην επιδημιολογία το σύνολο των ατόμων που παρακολουθείται αποκαλείται *cohort*.

[a] Στις κοινωνικές επιστήμες συχνά το ενδιαφέρον εστιάζεται γύρω από την καταγραφή της τάσης στη συμπεριφορά του πληθυσμού για κάποιο πολιτικό ή κοινωνικό θέμα. Στην αντίστοιχη ορολογία, το άτομο ή γενικότερα η μονάδα που παρακολουθείται αναφέρεται και ως *panel*.

1.3 Στόχοι μιας διαχρονικής μελέτης

Τα επιστημονικά ερωτήματα που μας ενδιαφέρουν αφορούν όχι μόνο τα συνήθη, όπως το κατά πόσο η απόκριση μπορεί να διαφέρει ανάμεσα στις διαφορετικές ομάδες (π.χ. θεραπείες), αλλά και ο εντοπισμός κάποιου προτύπου (*pattern*) στην συμπεριφορά της απόκρισης στο πέρασμα του χρόνου και η συμβολή ορισμένων σημαντικών ερμηνευτικών μεταβλητών στο εν λόγω πρότυπο. Επομένως, είναι απαραίτητο να χρησιμοποιήσουμε ένα αντιπροσωπευτικό μοντέλο, που να αναγνωρίζει τον τρόπο που συλλέγονται τα δεδομένα ώστε να πάρουμε κατάλληλες απαντήσεις.

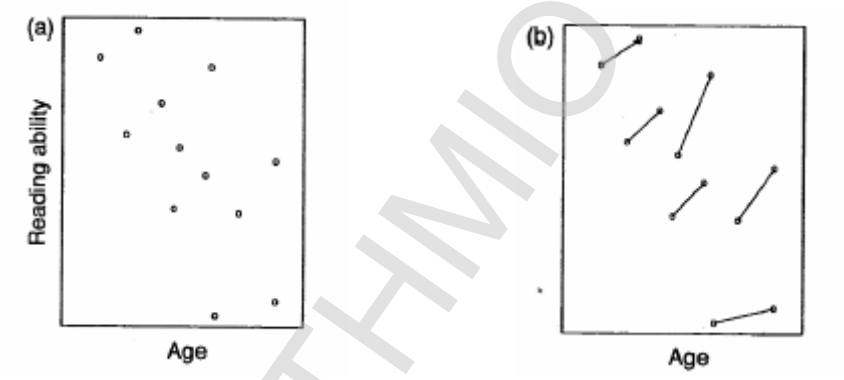
Αναφορικά με τις ερμηνευτικές μεταβλητές, υπάρχουν δύο είδη τέτοιων μεταβλητών. Αυτές που παραμένουν χρονικά σταθερές για κάθε μονάδα (π.χ το γένος και η φυλή) και αυτές που επηρεάζονται από το χρόνο (π.χ. η ηλικία, το βάρος κλπ). Όσον αφορά τις πρώτες, μετρώνται μόνο κατά την έναρξη της μελέτης (*baseline*) και η τιμή τους διαφοροποιείται μεταξύ (*between*) των μονάδων. Αντίθετα, οι άλλες μετρώνται συχνότερα και κατά συνέπεια διαφοροποιούνται εντός (*within*) των μονάδων.

1.4 Πλεονεκτήματα των διαχρονικών μελετών

Οι διαχρονικές μελέτες αντιπαρατίθενται με τις *cross-sectional*, όπου από κάθε μονάδα λαμβάνεται μόνο μία παρατήρηση. Τα ερωτήματα που πρέπει να απαντηθούν συχνά είναι τα ίδια και για τα δύο αυτά είδη μελέτης. Οι πρώτες, όμως, παρουσιάζουν μερικά σημαντικά πλεονεκτήματα. Κατ'αρχήν, απαιτείται μικρότερο δείγμα πειραματικών μονάδων. Για

παράδειγμα, όταν ερευνάται η επίδραση ενός φαρμάκου στο χρόνο, είναι συνήθως προτιμότερο να παρατηρούνται σταθερά τα ίδια άτομα παρά διαφορετικά σε κάθε συγκεκριμένη χρονική στιγμή.

Επίσης, οι διαχρονικές μελέτες επιτρέπουν στον ερευνητή να διαχωρίσει τις αλλαγές που καταγράφονται εντός των μονάδων με το πέρασμα του χρόνου (***ageing effects***) από αυτές που εμφανίζονται ανάμεσα στις πειραματικές μονάδες κατά την έναρξη (***cohort effects***). Οι ***cross-sectional*** μελέτες δεν μπορούν να διακρίνουν αυτή τη διαφοροποίηση. Προκειμένου να γίνει πιο κατανοητό το εν λόγω επιχείρημα, ας θεωρήσουμε μια υποθετική cross-sectional μελέτη (βλέπε Diggle et al., (2002)) στην οποία η ικανότητα ανάγνωσης παιδιών παραβάλλεται με την ηλικία τους (Γράφημα (a)). Στη συνέχεια θεωρούμε ότι τα ίδια δεδομένα λαμβάνονται από μια διαχρονική μελέτη όπου για τα παιδιά έγιναν δύο μετρήσεις (Γράφημα (b)).



Το συμπέρασμα από το πρώτο γράφημα είναι ότι η ικανότητα ανάγνωσης φαίνεται να είναι μικρότερη για παιδιά μεγαλύτερης ηλικίας. Αντίθετα από το δεύτερο, πέραν του παραπάνω συμπεράσματος, προκύπτει ακόμα ότι όλα τα παιδιά κατά την δεύτερη μέτρηση παρουσίασαν κάποια βελτίωση.

Ένα επιπλέον πλεονέκτημα είναι ότι σε μελέτες επαναλαμβανόμενων μετρήσεων οι μονάδες μπορούν να συμπεριφέρονται ως το προσωπικό τους ***control*** καθώς η μεταβλητή απόκρισης ενδεχομένως να μετράται τόσο υπό ***control*** όσο και υπό εναλλακτικές πειραματικές συνθήκες. Έτσι, η εσωτερική μεταβλητότητα που καταγράφεται εντός των μονάδων (***intra-subject variability***) μπορεί να σχετιστεί με τις αλλαγές της απόκρισης. Το αποτέλεσμα αυτού είναι η εξαίρεση της μεταξύ των μονάδων μεταβλητότητας από το σφάλμα μέτρησης και να προκύπτουν πιο αποτελεσματικοί εκτιμητές για τις σχετικές παραμέτρους από μια ***cross-sectional*** μελέτη με τον ίδιο αριθμό και το ίδιο ***pattern*** παρατηρήσεων.

1.5 Μειονεκτήματα των διαχρονικών μελετών

Παρά το γεγονός ότι τα διαχρονικού τύπου δεδομένα παρουσιάζουν σημαντικά οφέλη, δεν αποτελούν πανάκεια. Οι ιδιαιτερότητες που εμφανίζουν δημιουργούν ορισμένες δυσκολίες στην ανάλυσή τους. Αρχικά, αν και οι παρατηρήσεις μεταξύ των μονάδων θεωρούνται ανεξάρτητες, δεν μπορούμε να ισχυριστούμε το ίδιο για τις παρατηρήσεις εντός των μονάδων. Ανάμεσα σε αυτές υπάρχει κάποια μικρή ή μεγάλη συσχέτιση, όχι απόλυτη, που πρέπει να ληφθεί υπόψη, ώστε να εξάγουμε έγκυρα συμπεράσματα. Για την ανάλυση τέτοιων παρατηρήσεων απαιτούνται ειδικές στατιστικές μεθόδοι, οι οποίες είναι πολυπλοκότερες σε σχέση με τις κλασικές μεθόδους, όπου δεν παρατηρείται το φαινόμενο της συσχέτισης.

Ακόμα, τις περισσότερες φορές οι ερευνητές δεν μπορούν να ελέγχουν τις συνθήκες υπό τις οποίες λαμβάνονται οι παρατηρήσεις με συνέπεια συχνά τα δεδομένα να είναι μη-ισορροπημένα (***unbalanced***) ή μη-πλήρη (***incomplete***). Οι λόγοι για τους οποίους δεν καταγράφονται ορισμένες παρατηρήσεις διαφέρουν ανάλογα με το πεδίο της έρευνας. Στις κλινικές δοκιμές, για παράδειγμα, οι ελλιπείς μετρήσεις συνήθως οφείλονται σε αποχωρήσεις ασθενών πρίν το τέλος. Ένα τέτοιο φαινόμενο οδηγεί σε επιπλέον πολυπλοκότητα. Η πιο απλή αντιμετώπιση είναι να αναλύσουμε μόνο τις περιπτώσεις ασθενών με πλήρη δεδομένα. Σε περιπτώσεις όπου η κλινική δοκιμή συγκρίνει ένα νέο φάρμακο με το ***placebo***, είναι πιθανό μόνο οι ασθενείς που παρουσιάζουν βελτίωση να διαθέτουν πλήρη δεδομένα. Το αποτέλεσμα είναι να καταλήξουμε σε μια ανωτερότητα του φαρμάκου της οποίας το μέγεθος να είναι πλασματικό. Τίθεται, λοιπόν, το ερώτημα της σωστής αντιμετώπισης των ελλιπών δεδομένων.

1.6 Τρόποι ανάλυσης διαχρονικών δεδομένων - Προβλήματα

Στην συγκεκριμένη εργασία θα περιοριστούμε σε διαχρονικά δεδομένα που προκύπτουν από την μέτρηση κατηγορικής και ειδικότερα διατάξιμης μεταβλητής απόκρισης. Για την ανάλυση επαναλαμβανόμενων μετρήσεων έχουν αναπτυχθεί πολλές προσεγγίσεις στην περίπτωση που η απόκριση είναι συνεχής και ακολουθεί κανονική κατανομή. Η ανάπτυξη

αντίστοιχων μεθόδων για κατηγορικά δεδομένα δεν ήταν στο επίκεντρο του ενδιαφέροντος στο παρελθόν αλλά προσφάτως αποτελεί σημαντική και ενεργή περιοχή της έρευνας.

Για την προσαρμογή των δεδομένων υπάρχουν τρεις επεκτάσεις των γενικευμένων γραμμικών μοντέλων, οι οποίες αποτελούν τις πλέον δημοφιλείς επιλογές. Πρόκειται για τα μοντέλα μεταβάσεων (***transition models***), τα περιθώρια (***marginal models***) και τα μοντέλα τυχαίων επιδράσεων (***random effects models***). Στην παρούσα εργασία να ασχοληθούμε με τους δύο τελευταίους τύπους μοντέλων ενώ σε επόμενη ενότητα του κεφαλαίου αυτού ακολουθεί μια σύντομη αναφορά στα ***transition*** μοντέλα.

Στην περίπτωση των κατηγορικών αποκρίσεων συναντά κανείς αρκετές δυσκολίες κατά την εφαρμογή τόσο των περιθώριων όσο και των μοντέλων τυχαίων επιδράσεων. Σχετικά με τα πρώτα, όπως θα δούμε και στη συνέχεια, ένα από τα κρισιμότερα σημεία είναι η έλλειψη κοινής κατανομής για τις παρατηρήσεις της κάθε πειραματικής μονάδας και όταν, μάλιστα, πρόκειται για διατάξιμη απόκριση χρειάζεται επιπλέον επέκταση της υπάρχουσας θεωρίας εξαιτίας της ιδιαίτερης φύσης της. Στα μοντέλα τυχαίων επιδράσεων τα κυριότερα προβλήματα έχουν να κάνουν με την εκτίμηση των παραμέτρων του εκάστοτε μοντέλου. Πολύ συχνά, το υπολογιστικό μέρος είναι επίπονο καθώς απαιτούνται αριθμητικές ή *Monte-Carlo* μέθοδοι προσομοίωσης για την αποτίμηση της πιθανοφάνειας.

Ένα πρόσθετο πρόβλημα είναι η απουσία επαρκών στατιστικών πακέτων και λογισμικών για την αντιμετώπιση τέτοιων προβλημάτων. Τα περισσότερα γνωστά στατιστικά πακέτα είτε δεν διαθέτουν σχετικές διαδικασίες είτε διαθέτουν για περιορισμένο τύπο δεδομένων. Έρευνα για τις δυνατότητες που υπάρχουν στην επιλογή του ανάλογου πακέτου, θα γίνει στο 4ο Κεφάλαιο.

1.7 Μοντέλα μεταβάσεων (Transition models)

Για την ανάλυση των επαναλαμβανόμενων δεδομένων στον συγκεκριμένο τύπο μοντέλων, οι παρατηρήσεις που λαμβάνονται από το *i*-υποκείμενο και θα συμβολίζονται με $Y_{i1}, Y_{i2}, ..., Y_{in_i}$, όπου n_i είναι το πλήθος αυτών, είναι συσχετισμένες διότι η *t*-παρατήρηση Y_{it} εξαρτάται από τις παρελθοντικές τιμές $Y_{i1}, Y_{i2}, ..., Y_{i,t-1}$. Οι τιμές αυτές αντιμετωπίζονται ως

πρόσθετες ερμηνευτικές μεταβλητές. Η δεσμευμένη κατανομή μιας τρέχουσας παρατήρησης, δοθέντων των προηγούμενων, περιγράφεται από ένα γενικευμένο γραμμικό μοντέλο.

Η γενική μορφή ενός τέτοιου μοντέλου θα είναι

$$g(E(Y_{it} \mid Y_{i1},...,Y_{i,t-1}) = \mathbf{x}_{it}^{'}\boldsymbol{\beta} + \sum_{r=1}^{s} f_r(Y_{i1},...,Y_{i,t-1};a_1,...,a_s) \ ,$$

όπου $g(\cdot)$ είναι η συνάρτηση σύνδεσης, $\mathbf{x}_{it}$ το διάνυσμα των ερμηνευτικών μεταβλητών και $\boldsymbol{\beta}$ το διάνυσμα των αντίστοιχων παραμέτρων. Επιπλέον, $f_1,...,f_s$ είναι συναρτήσεις των προηγούμενων παρατηρήσεων και πιθανόν ενός διανύσματος άγνωστων παραμέτρων $\boldsymbol{\alpha} = (a_1,...,a_s)^{'}$. Επιπλέον, η δεσμευμένη διακύμανση της Y_{it} είναι ανάλογη μιας γνωστής συνάρτησης του δεσμευμένου μέσου. Δηλαδή

$$\mathrm{Var}(Y_{it} \mid Y_{i1},...,Y_{i,t-1}) = f\,u(E(Y_{it} \mid Y_{i1},...,Y_{i,t-1})) \ ,$$

όπου $u(\cdot)$ είναι μια γνωστή συνάρτηση διασποράς και f άγνωστη παράμετρος κλίμακας.

Το πιο δημοφιλές στην κατηγορία αυτή είναι το μοντέλο ***Markov*** πρώτης-τάξης, κατά το οποίο η παρατήρηση Y_{it} εξαρτάται μόνο από την προηγούμενη παρατήρηση $Y_{i,t-1}$. Η διατάξιμη φύση της απόκρισης μπορεί να αξιοποιηθεί χρησιμοποιώντας κάποιο από τα σχετικά ***logit*** μοντέλα, στα οποία θα αναφερθούμε στο επόμενο κεφάλαιο. Η εκτίμησή τους γίνεται με τη βοήθεια της μέγιστης πιθανοφάνειας.

Για περισσότερες πληροφορίες γύρω από τον συγκεκριμένο τύπο μοντέλων μπορούν οι ενδιαφερόμενοι να αναζητήσουν στις εργασίες των Ekholm, Smith, and McDonald, (1995); Kosorok and Chao, (1996); Ekholm, Jokinen, McDonald and Smith, (2003).

1.8 Ελλιπή δεδομένα

Αναφέραμε πρωτύτερα ότι η παρουσία ελλιπών παρατηρήσεων είναι σύνηθες χαρακτηριστικό σε διαχρονικές έρευνες και ο τρόπος που αντιμετωπίζονται στην ανάλυση είναι ιδιαίτερης σημασίας. Οι Little and Rubin, (1987) έχουν διακρίνει τρεις μηχανισμούς που έχουν ως αποτέλεσμα αυτό το φαινόμενο και είναι οι ακόλουθοι:

1) Τα δεδομένα λείπουν εντελώς τυχαία (***missing completely at random, MCAR***). Αυτό σημαίνει ότι η πιθανότητα να καταγράψουμε μια παρατήρηση είναι ανεξάρτητη από αυτές που έχουμε καταγράψει ή όχι μέχρι τότε.
2) Τα δεδομένα λείπουν τυχαία (***missing at random, MAR***). Δηλαδή η πιθανότητα να καταγράψουμε μια παρατήρηση εξαρτάται από τις μέχρι στιγμής καταγεγραμμένες αλλά όχι από αυτές που δεν έχουν καταγραφεί.
3) Τα δεδομένα λείπουν με μη τυχαίο τρόπο (***informative/nonrandom/nonignorable***). Στην περίπτωση αυτή, η πιθανότητα να παρατηρήσουμε την μεταβλητή απόκρισης εξαρτάται από τις μη παρατηρούμενες τιμές αυτής.

Για την κατανόηση των παραπάνω, ας θεωρήσουμε μια μελέτη στην οποία η μεταβλητή απόκρισης μετράται για ένα καθορισμένο αριθμό περιπτώσεων (π.χ. επισκέψεων) για κάθε άτομο. Αν τα άτομα χάνουν τις επισκέψεις τους εντελώς τυχαία, τότε έχουμε να κάνουμε με τον πρώτο μηχανισμό. Αν η πιθανότητα να χάσουν μια επίσκεψη σχετίζεται με τις μετρήσεις κατά τις προηγούμενες επισκέψεις, τότε οδηγούμαστε στον δεύτερο μηχανισμό. Τέλος, ο τρίτος προκύπτει αν τα άτομα είναι λιγότερο ή περισσότερο πιθανό να χάσουν μια επίσκεψη βασιζόμενα στην μη καταγεγραμμένη παρατήρηση αυτής της επίσκεψης. Σημειώνουμε ότι τα χαρακτηριστικά αυτών των μηχανισμών αναφέρονται μόνο στην μεταβλητή απόκρισης και δεν απευθύνονται στις επιδράσεις των συμμεταβλητών.

Αν τα δεδομένα λείπουν με βάση τον πρώτο μηχανισμό, οι περισσότερες προσεγγίσεις της ανάλυσης είναι έγκυρες και το θέμα είναι πλέον η δυσκολία στην υλοποίησή της για μη-πλήρη δεδομένα. Όσον αφορά τον δεύτερο, τα συμπεράσματα που βασίζονται στην πιθανοφάνεια είναι επίσης έγκυρα. Δεν συμβαίνει, όμως, το ίδιο και για αναλύσεις που βασίζονται στις ροπές αφού εμφανίζεται μεροληψία. Το πρόβλημα γίνεται οξύτερο στην περίπτωση του τρίτου μηχανισμού, όπου οι βασιζόμενες στην πιθανοφάνεια και τις ροπές μέθοδοι είναι αμφότερες μεροληπτικές.

1.9 Ιστορική αναδρομή

Κλασικές αναφορές για την ανάλυση διαχρονικών δεδομένων αποτελούν τα βιβλία των Crowder and Hand, (1990); Pickles, (1990); Hagenaars, (1990); Girden, (1992); Davis, (2002). Όλα τα παραπάνω, σε μικρότερη ή μεγαλύτερη έκταση, αναφέρονται και σε

διαχρονικά κατηγορικά δεδομένα. Για την ανάλυση κατηγορικών επαναλαμβανόμενων δεδομένων οι ενδιαφέρομενοι μπορούν ακόμα να ανατρέξουν στα βιβλία των Agresti, (2002, Κεφάλαια 11 και 12) και Diggle et al., (1994, 2002).

Αναφέραμε ήδη ότι τα περιθώρια μοντέλα και η μέθοδος των γενικευμένων εξισώσεων εκτίμησης (*GEE* ή *GEE*1) που εισήχθη για την εκτίμησή τους, αποτελούν τον πιο βασικό τρόπο χειρισμού αυτού του τύπου των δεδομένων. Αποτέλεσαν το θέμα μελέτης αρκέτων βιβλίων και μεταξύ άλλων των Bergsma, (1997) και Hardin and Hilbe, (2003). Ακόμα, στις πρόσφατες εκδόσεις για μοντέλα τυχαίων επιδράσεων με διαχρονικά δεδομένα ανήκουν τα βιβλία των Bryk and Raudenbush, (1992); Longford, (1993); Goldstein, (1995); Hand and Crowder, (1996).

Η θεμελίωση της προσαρμογής των περιθώριων μοντέλων έγινε από τους Liang and Zeger, (1986), οι οποίοι προτείνουν μια επέκταση της θεωρίας των γενικευμένων γραμμικών μοντέλων για την ανάλυση διαχρονικών δεδομένων και εισάγουν τις εξισώσεις για την εκτίμηση των παραμέτρων τόσο για την περίπτωση της ανεξαρτησίας όσο και για περιπτώσεις πολυπλοκότερης δομής για τη συσχέτιση. Στη συνέχεια, ο Prentice, (1988) παρουσίασε ορισμένες μεθόδους για συσχετισμένα δίτιμα δεδομένα και μεταξύ αυτών την μέθοδο των Liang and Zeger.

Ο όγκος της σχετικής βιβλιογραφίας για δίτιμες μεταβλητές είναι σημαντικός. Ενδεικτικά προτείνουμε κάποιες από τις πιο χαρακτηριστικές αναφορές. Σημειώνουμε την συγκριτική παρουσίαση του Neuhaus, (1992) των διαφορετικών προσεγγίσεων για την ανάλυση διαχρονικών δίτιμων αποκρίσεων. Ο ίδιος επικεντρώνεται στην ερμηνεία των παραμέτρων, τα τυπικά σφάλματα και τα στατιστικά τεστ του Wald. Στην εργασία του Fitzmaurice, (1995) μελετώνται οι ιδιότητες των εκτιμητών αν υιοθετήσουμε την ανεξάρτητη δομή στις παρατηρήσεις και οι συνέπειες μιας τέτοιας επιλογής στην αποτελεσματικότητά τους. Οι Ekholm et al., (1995) αναφέρονται στην μοντελοποίηση της συσχέτισης με λόγους εξάρτησης (*dependence ratios*) που ορίζονται με την βοήθεια των από κοινού πιθανοτήτων επιτυχίας όλων των τάξεων. Επιπλέον, στην εργασία των Albert and Shane, (1995) προτείνεται μια προσέγγιση των *GEE* για την ανάλυση spatially συσχετισμένων δίτιμων δεδομένων, όταν υπάρχουν πολλές spatially τέτοιες παρατηρήσεις σε ένα μεσαίου μεγέθους πληθυσμό.

Η ανάλυση διατάξιμων επαναλαμβανόμενων μετρήσεων αποτελεί το αντικείμενο των Miller et al., (1993), οι οποίοι στην επέκταση της θεωρίας των Liang and Zeger, (1986) για αποκρίσεις. με τουλάχιστον δύο κατηγορίες (*polytomous*). Στη συνέχεια, συνδέουν και

συγκρίνουν τα αποτελέσματα με αυτά της μεθόδου των σταθμισμένων ελαχίστων τετραγώνων (***Weighted Least Squares, WLS***). Οι Stram et al., (1988) ασχολήθηκαν με τον ίδιο τύπο δεδομένων στην περίπτωση που εμφανίζονται ελλιπείς μετρήσεις και στο μοντέλο περιλαμβάνονται χρονο-εξαρτώμενες συμμεταβλητές. Επίσης, οι Kenward et al., (1994) παρουσιάζουν μια εφαρμογή με *MAR* παρατηρήσεις, όπου μέσα από το μοντέλο των αναλογικών λόγων πιθανοτήτων (***proportional odds***) συγκρίνουν την μέθοδο των GEE με αυτή της πιθανοφάνειας. Περεμφερές είναι και το θέμα της εργασίας των Mark and Gail, (1994), όπου πέραν της σύγκρισης προτείνονται προσαρμογές (***adaptations***) και για τις δύο μεθόδους ώστε να προκύπτουν έγκυρα αποτελέσματα κάτω από την ασθενέστερη υπόθεση των *MAR* παρατηρήσεων.

Οι Heagerty and Zeger, (1996) αναφέρονται σε στατιστικά μοντέλα για διατάξιμες αποκρίσεις, όπου η σχέση των εξαρτημένων παρατηρήσεων μοντελοποιείται μέσω κατάλληλα ορισμένων ***odds ratio***, καθώς επίσης και στην μέθοδο των *GEE*2 (επέκταση των *GEE*). Οι Toledano and Gatsonis, (1999) προτείνουν μεθόδους για την ανάλυση διατάξιμων αποκρίσεων όταν υφίσταται μη-μονότονος μηχανισμός ελλιπών δεδομένων και επεκτείνουν τη μεθοδολογία των *GEE* προς αυτή την κατεύθυνση. Επίσης, οι Huang et al., (2002) εφαρμόζουν την θεωρία των *GEE* στην περίπτωση που χρησιμοποιούνται πολλαπλές δείκτριες μεταβλητές για την αποτίμηση της υπό μελέτης απόκρισης, όπως συμβαίνει σε πολλές βιοϊατρικές και ψυχοκοινωνικές έρευνες. Επιπλέον, παρουσιάζονται γραφικές μέθοδοι για την διάγνωση της επάρκειας αυτών των μοντέλων.

Σχετικά, με τα μοντέλα τυχαίων επιδράσεων για διατάξιμα δεδομένα, αυτά προτάθηκαν αρχικά από τους Harville and Mee, (1984), των οποίων το αντικείμενο ήταν οι βέλτιστοι γραμμικοί αμερόληπτοι εκτιμητές (***Best Linear Unbiased Predictors, BLUP***) των παραμέτρων μιας λανθάνουσας (*latent*) κατανομής. Στη συνέχεια ο Jansen, (1990) πρότεινε ένα απλό μοντέλο τυχαίων επιδράσεων χρησιμοποιώντας μόνο μία τυχαία επίδραση και η εκτίμηση του γίνεται με τον *EM*-αλγόριθμο. Μια πιο γενική μέθοδος προτάθηκε από τους Hedeker and Gibbons, (1994), οι οποίοι χρησιμοποίησαν έναν γρηγορότερο ***Fisher-scoring*** αλγόριθμο για την εκτίμηση των παραμέτρων ενός μοντέλου με πολλαπλές τυχαίες επιδράσεις.

Προκειμένου να έχουμε συνεπείς εκτιμητές, οι Agresti and Lang, (1993) προτείνουν ταυτόχρονη προσαρμογή των Rasch μοντέλων, δεσμεύοντας ως προς επαρκή στατιστικά των παραμέτρων του μοντέλου, για όλους τους δυνατούς τρόπους με τους οποίους η απόκριση γίνεται δίτιμη. Οι Hedeker and Gibbons, (1997) χειρίζονται και αναλύουν την επιρροή των

ελλιπών δεδομένων σε διαχρονικές μελέτες μέσω μοντέλων τυχαίων επιδράσεων. Επίσης, οι Tutz and Hennevogl, (1995) παρουσιάζουν την γενική περίπτωση του ordinal αθροιστικού μοντέλου με τυχαίες επιδράσεις και τρεις εναλλακτικές μεθόδους εκτίμησης.

Στην εργασία του Crouchley, (1995) παρουσιάζεται ένα μοντέλο τυχαίων επιδράσεων για πολυμεταβλητές και ***grouped univariate*** διατάξιμες αποκρίσεις. Για την κατανομή των τυχαίων επιδράσεων υιοθετείται ένα ευρύ φάσμα εναλλακτικών επιλογών. Οι Hedeker and Mermelstein, (2000) ασχολούνται με την περιγραφή και την ανάλυση διατάξιμων μοντέλων τυχαίων επιδράσεων με μεταβαλλόμενες χρονικά επιδράσεις για τις συμμεταβλητές ενώ στην εργασία των Agresti et al., (2000) παρουσιάζεται η θεωρία των γενικευμένων γραμμικών μεικτών μοντέλων (***GLMM***) με επεκτάσεις για την περίπτωση των δίτιμων και των κατηγορικών επαναλαμβανόμενων αποκρίσεων.

Κλείνοντας, θα ήταν παράληψη να μην αναφερθούν βασικές εργασίες ανασκόπησης των διαθέσιμων προσεγγίσεων για την ανάλυση διαχρονικών δεδομένων τόσο με περιθώρια όσο και με μοντέλα τυχαίων επιδράσεων, όπως των Agresti, (1989); Pendergast et al., (1996) και Agresti and Natarajan, (2001).

1.10 Διάρθρωση της διπλωματικής

Τα Κεφάλαια 2 και 3 ασχολούνται με την θεωρία των περιθώριων και των μοντέλων τυχαίων επιδράσεων αντίστοιχα. Στο Κεφάλαιο 4 γίνεται αναφορά στα διαθέσιμα στατιστικά προγράμματα και τα λογισμικά που βοηθούν στην εκτίμηση των δύο παραπάνω τύπων μοντέλου και στο τελευταίο ακολουθεί η εφαρμογή της θεωρίας σε δύο διαφορετικά σύνολα δεδομένων που προσαρμόζουμε με κάποια από τα προγράμματα που αναφέρονται στο προηγούμενο κεφάλαιο.

Κεφάλαιο 2ο
Περιθώρια μοντέλα
(Marginal models)

2.1 Εισαγωγή

Στις περισσότερες έρευνες, και ειδικότερα όσες αφορούν μελέτη γύρω από την ανωτερότητα ή όχι ενός νέου φαρμακευτικού σκευάσματος, οι παρατηρήσεις που έχουμε τελικά στην διάθεσή μας δεν είναι κατ'ανάγκη ανεξάρτητες μεταξύ τους. Συνηθώς, υπάρχει κάποια συσχέτιση, η οποία προκύπτει από τον τρόπο που συλλέγονται οι παρατηρήσεις από τις πειραματικές μονάδες, έτσι όπως έχουν οριστεί από το πλαίσιο της μελέτης. Το συγκεκριμένο αυτό χαρακτηριστικό οφείλει να λαμβάνεται υπόψη κατά την ανάλυση των δεδομένων και για το λόγο αυτό τις τελευταίες δεκαετίες γίνεται προσπάθεια ώστε η θεωρία να καλύψει και το εν λόγω πεδίο.

Μια από τις πρώτες προσεγγίσεις αφορά στα περιθώρια (***marginal***) μοντέλα, που θα αναπτύξουμε στο κεφάλαιο αυτό. Ειδικότερα, στην ενότητα 2.2 θα αναφερθούμε σε γενικά στοιχεία των μοντέλων αυτών και στην επόμενη 2.3 θα δούμε την κλασική μέθοδο προσαρμογής αυτών μέσω της μέγιστης πιθανοφάνειας. Στην ενότητα 2.4 θα αναφερθούμε σε μια εναλλακτική μέθοδο προσαρμογής μέσω των γενικευμένων εκτιμητικών εξισώσεων Στις ενότητες 2.5 και 2.6 δίνεται ένα παράδειγμα και μια επέκταση της μεθόδου μέγιστης πιθανοφάνειας αντίστοιχα. Στην τελευταία ενότητα γίνεται λόγος για δύο τεστ καλής προσαρμογής όσον αφορά την συγκεκριμένη κατηγορία μοντέλων.

2.2 Γενικά στοιχεία

Το σύνολο των συσχετισμένων μετρήσεων (***correlated measures***) δύναται να χωριστεί σε ομάδες ή συστάδες (***clusters***). Η κάθε ομάδα θα περιλαμβάνει το σύνολο των συσχετισμένων μεταξύ τους μετρήσεων ενώ οι ομάδες θεωρούνται ανεξάρτητες η μία από την άλλη. Αν *Y* είναι η μεταβλητή απόκρισης, τότε συμβολίζουμε με Y_{it} την τιμή της απόκρισης του *i*-υποκειμένου κατά την *t*-επανάληψη. Οι συσχετισμένες παρατηρήσεις ενδέχεται να είναι επαναλήψεις πάνω στο ίδιο άτομο που διαφοροποιούνται είτε χρονικά είτε λόγω διαφορετικών πειραματικών περιστάσεων (***repeated measures***). Πιθανόν, όμως, οι συγκεκριμένες παρατηρήσεις να αναφέρονται σε ένα σύνολο ατόμων που αναμένεται να παρουσιάζουν μεγαλύτερη ομοιογένεια ως προς το χαρακτηριστικό που μετράται σε σχέση με άλλα άτομα (π.χ. σε γενετικές έρευνες ως τέτοια μονάδα μπορεί να οριστεί μια οικογένεια) ή κοινό ιστορικό. Η εξάρτηση που υπάρχει ανάμεσα στις επαναλαμβανόμενες μετρήσεις είναι μεν δεδομένη αλλά δεν αποτελεί τον κύριο στόχο της μελέτης. Το ενδιαφέρον μας επικεντρώνεται στην συμπεριφορά που έχουν οι περιθώριες κατανομές (***marginal distributions***) των εν λόγω μετρήσεων. Για παράδειγμα, κατά την διάρκεια της θεραπείας χρόνιων παθήσεων, ως πρωτεύων στόχος θεωρείται η έρευνα προκειμένου να διαπιστωθεί αν η πιθανότητα επιτυχίας αυξάνει στο πέρασμα του χρόνου της περιόδου θεραπείας. Οι εν λόγω πιθανότητες είναι αυτές που αναφέρονται ως πρώτης τάξης περιθώριες κατανομές.

Για την μοντελοποίηση των περιθώριων κατανομών, εισήχθησαν τα περιθώρια (***marginal***) μοντέλα (Liang and Zeger, 1986). Η χρησιμοποίηση αυτών των μοντέλων αναφέρεται συχνά και ως "***population-averaged***" προσέγγιση και αυτό διότι οι παράμετροι αυτών ερμηνεύονται σε σχέση με την επίδραση των συμμεταβλητών στον μέσο όρο της απόκρισης (***average response***) για όλες τις ομάδες του πληθυσμού που μελετάται. Για διαχρονικά δεδομένα (***longitudinal data***), τα μοντέλα αυτά διαχωρίζουν την μοντελοποίηση των μεταξύ των ατόμων (***between-subject***) και εντός των ατόμων (***within-subject***) επιδράσεων. Οι πρώτες μοντελοποιούνται μέσα από την περιθώρια μέση τιμή $E(Y_{it})$ (***marginal mean***) ενώ οι άλλες μέσα από την δομή των συνδιακυμάνσεων $cov(Y_{ih}, Y_{ig})$.

Ας υποθέσουμε ότι το πλήθος των συσχετισμένων παρατηρήσεων σε κάθε ομάδα είναι *T*. Δηλαδή είτε έχουμε *T* επαναλήψεις για κάθε πειραματική μονάδα (π.χ. χορήγηση διαφορετικών δόσεων ενός συγκεκριμένου φαρμάκου για την αντιμετώπιση κάποιας

πάθησης) είτε κάθε ομάδα αποτελείται από T μονάδες με ομοιογένεια μεταξύ τους (π.χ. σύνολο παιδιών κάθε οικογένειας). Έτσι, σε κάθε ομάδα (π.χ. i, i=1,2,…,N) θα έχουμε τις συσχετισμένες αποκρίσεις $(Y_{i1}, Y_{i2}, ..., Y_{iT})$. Στη συνέχεια παρατίθενται ορισμένα μοντέλα ανάλογα με το είδος της μεταβλητής απόκρισης. Η επιλογή κάποιου από αυτά βασίζεται όχι τόσο στην προσαρμογή αλλά στο αν η ερμηνεία των παραμέτρων θα αφορά ξεχωριστές κατηγορίες της μεταβλητής απόκρισης ή κάποιο σύνολο αυτών. Διακρίνουμε, λοιπόν, τις παρακάτω περιπτώσεις:

(1) Δίτιμη απόκριση (***binary response***, π.χ. $Y_t = 1$ για την επιτυχία και $Y_t = 0$ για την αποτυχία)

Το κατάλληλο μοντέλο είναι :

$$\log it[P(Y_t = 1)] = a + b_t \ , \qquad (2.1)$$

$t = 1, 2, ..., T$ και $i = 1, 2, ..., N$, ενώ οι παράμετροι ικανοποιούν κάποιο περιορισμό προσδιορισιμότητας, όπως b_T =0. Υπενθυμίζουμε ότι το μοντέλο αναφέρεται στις περιθώριες πιθανότητες. Πιο συγκεκριμένα, αν

$$p(t_1, t_2, ..., t_T) = P(Y_1 = t_1, Y_2 = t_2, ..., Y_T = t_T)$$

με $t_1, t_2, ..., t_T \in \{0,1\}$, τότε $P(Y_{\mathbf{l}} = 1) = p(+, ..., +, 1, +, ..., +)$ με 1 στη θέση $\mathbf{l}$, όπου με + συμβολίζουμε το άθροισμα ως προς τις κατηγορίες των υπόλοιπων μεταβλητών. Αν $b_1 = b_2 = ... = b_T$, τότε έχουμε περιθώρια ομοιογένεια (***marginal homogeneity***).

(2) Πολύτιμη διατάξιμη απόκριση (***multicategory ordinal response***).

Για την περίπτωση που η μεταβλητή απόκρισης έχει περισσότερα από δύο επίπεδα και είναι διατάξιμη, υπάρχουν αρκετά εναλλακτικά μοντέλα. Έστω $\mathbf{x}$ το διάνυσμα των ερμηνευτικών μεταβλητών και ας υποθέσουμε ότι η μεταβλητή Y έχει I κατηγορίες. Κατά περίπτωση, θεωρούμε τα μοντέλα :

$$\log it[P(Y_t \le k \mid \mathbf{x}_t)] = a_k + \boldsymbol{\beta}' \mathbf{x}_t \ , \qquad (2.2)$$

όπου $k = 1, 2, ..., I-1$ και $t = 1, 2, ..., T$. Το μοντέλο αυτό υποθέτει την ίδια επίδραση $\boldsymbol{\beta}$ για τις ερμηνευτικές μεταβλητές για κάθε αθροιστική πιθανότητα (***cumulative probability***) και είναι γνωστό ως ***proportional odds*** μοντέλο (McCullagh, 1980). Αναφέρουμε ότι με τον όρο "***odds***" εννούμε τον λόγο της πιθανότητας "επιτυχίας" προς αυτή της "αποτυχίας", όπως ορίζουμε κάθε φορά αυτά τα ενδεχόμενα.

Το μοντέλο

$$\log[P(Y_t = k) / P(Y_t = I)] = a_k + \boldsymbol{\beta}' \mathbf{x}_t \,, \qquad (2.3)$$

όπου $k = 1, 2, ..., I-1$ και $t = 1, 2, ..., T$, αναφέρεται ως ***baseline-category logit*** μοντέλο. Το συγκεκριμένο μπορεί να χρησιμοποιηθεί και για μη-διατάξιμη μεταβλητή απόκρισης, δηλαδή ακόμα και αν είναι ονομαστική (***nominal***). Επίσης όταν τα ***odds*** ορίζονται σε διαδοχικές κατηγορίες της μεταβλητής απόκρισης, καταλήγουμε στο ***adjacent-categories logit*** μοντέλο

$$\log[P(Y_t = k) / P(Y_t = k+1)] = a_k + \boldsymbol{\beta}' \mathbf{x}_t \,, \qquad (2.4)$$

όπου $k = 1, 2, ..., I-1$ και $t = 1, 2, ..., T$. Τέλος, ορίζονται και τα ***continuation-ratio logit*** μοντέλα

$$\log[P(Y_{it} = k) / P(Y_{it} \geq k+1)] \text{ ή } \log[P(Y_t = k+1) / P(Y_t \leq k)] = a_k + \boldsymbol{\beta}' \mathbf{x}_t \,, \qquad (2.5)$$

όπου $k = 1, 2, ..., I-1$ και $t = 1, 2, ..., T$.

Το κάθε μοντέλο από τα προηγούμενα χρησιμοποιείται αναλόγως με το τι θέλουμε μέσα από κάθε έρευνα να μελετήσουμε ή να αναδείξουμε. Συνήθως, για την ερμηνεία των παραμέτρων χρησιμοποιούμε τον λόγο των σχετικών πιθανοτήτων (***odds ratio***). Αν θεωρήσουμε το μοντέλο (2.2), μπορούμε να συγκρίνουμε κάθε κατηγορία με την τελευταία (επίπεδο αναφοράς) δημιουργώντας έτσι *I*-1 ***odds ratios***. Ως κατηγορία αναφοράς θα μπορούσε να είχε οριστεί και η πρώτη. Όταν έχουμε διατάξιμη (***ordina***l) απόκριση, όπως στα μοντέλα (2.3) και (2.4), η σύγκριση μπορεί να αφορά γειτονικές κατηγορίες ή μια κατηγορία ως προς αυτές που είναι "μικρότερες" ή "μεγαλύτερες" από αυτή.

Στην περίπτωση που η προσαρμογή δεν είναι καλή, υπάρχουν ορισμένες μέθοδοι που ακολουθούνται για να πάρουμε καλύτερα αποτελέσματα: **1)** δοκιμάζουμε μια διαφορετική συνάρτηση σύνδεσης (***link function***), όπως ***log-log*** ή ***complementary log-log***, **2)** προσθέτουμε επιπλέον όρους, όπως αλληλεπιδράσεις, **3)** γενικεύουμε το μοντέλο προσθέτοντας παραμέτρους διασποράς (***dispersion parameters***, McCullagh, 1980; Cox, 1995) και **4)** επιτρέπουμε διαφορετικές επιδράσεις b_j για κάθε ***logit*** για κάποιες από τις ερμηνευτικές μεταβλητές (Peterson and Harrell, 1990).

Τέλος, σημειώνουμε ότι στην παρουσίαση των μοντέλων περιοριστήκαμε σε ισορροπημένα ***(balanced***) δεδομένα. Δηλαδή, κάθε ομάδα αποτελείται από *T* συσχετισμένες παρατηρήσεις

ενώ γενικότερα θα μπορούσε να είχαμε T_i παρατηρήσεις για την ***i***-ομάδα με αποτέλεσμα μη ισορροπημένα (***unbalanced***) δεδομένα.

2.3 Μέθοδος Μέγιστης Πιθανοφάνειας για την εκτίμηση παραμέτρων - Αδυναμίες

Ας θεωρήσουμε το αθροιστικό ***marginal*** μοντέλο για τις πρώτης τάξης περιθώριες πιθανότητες:

$$\log it[P(Y_t \le k \mid \mathbf{x}_t) = a_k + \boldsymbol{\beta}'\mathbf{x}_t, \text{ όπου } k = 1,2,...,I-1 \text{ και } t = 1,2,...,T .$$

Για την εκτίμηση των παραμέτρων b_i του διανύσματος $\boldsymbol{\beta}$, θα περίμενε κανείς να βρούμε την συνάρτηση πιθανοφάνειας (***likelihood function***) και στην συνέχεια να μεγιστοποιήσουμε με τις γνωστές μεθόδους τον λογάριθμο αυτής παίρνοντας τελικά τους επιθυμητούς εκτιμητές. Γενικώς, η συνάρτηση πιθανοφάνειας αναφέρεται στις από κοινού και όχι στις περιθώριες πιθανότητες και επομένως δεν μπορούμε να χρησιμοποιήσουμε τις πιθανότητες από το περιθώριο μοντέλο για να μεγιστοποιήσουμε τον λογάριθμό της. Επιπλέον, αντιλαμβάνεται κανείς ότι αν διαθέτουμε μεγάλο αριθμό παρατηρήσεων T για κάθε μονάδα ή ακόμα και πλήθος ερμηνευτικών μεταβλητών, η οποιαδήποτε προσέγγιση του προβλήματος εκτίμησης των παραμέτρων μέσω μέγιστης πιθανοφάνειας καθίσταται μη πρακτική.

Ένα επιπλέον πρόβλημα που τίθεται, είναι ότι προκειμένου να κατασκευάσει κανείς την συνάρτηση πιθανοφάνειας πρέπει να γνωρίζει την κατανομή πιθανότητας για τις τυχαίες μεταβλητές. Σε πολλές περιπτώσεις δεν υπάρχουν διαθέσιμες πληροφορίες για τον μηχανισμό παραγωγής των δεδομένων αλλά ακόμα και αν υπάρχουν στοιχεία για κάποια θεωρητική κατανομή, αυτή μπορεί να αποδειχθεί ανεπαρκής για τα παρατηρούμενα δεδομένα. Για παράδειγμά, ενδεχομένως η παρατηρούμενη διακύμανση της μεταβλητής απόκρισης να ξεπερνά αυτή που έχουμε υποθέσει εξαρχής (***overdispersion***), γεγονός που μπορεί να οφείλεται στην ετερογένεια ανάμεσα στις μονάδες της έρευνας. Τέλος, δεν μπορεί να αποκλειστεί το ενδεχόμενο το θεωρητικό μοντέλο που θα προκύψει να είναι ιδιαιτέρως πολύπλοκο για να κάνουμε εκτίμηση των παραμέτρων και κατ'επέκταση στατιστική συμπερασματολογία.

2.4 Εισαγωγή στη μέθοδο των GEE

Ένας τρόπος να προσπεράσουμε τα προβλήματα που δημιουργεί η διαδικασία μέγιστης πιθανοφάνειας (*ML*), είναι η χρησιμοποίηση των γενικευμένων εξισώσεων εκτίμησης παραμέτρων (*generalized estimating equations, GEE*), που αποτελούν επέκταση της *quasi-likelihood* (*QL*) στην ανάλυση διαχρονικών δεδομένων. Η *quasi-likelihood* μέθοδος προτάθηκε αρχικά από τον Wedderburn (1974) και στα πλαίσια αυτής δεν είναι πλέον απαραίτητος ο καθορισμός κάποιας συγκεκριμένης από κοινού κατανομής για το σύνολο των παρατηρήσεων. Αρκεί η υπόθεση ότι πρόκειται για μια κατανομή που ανήκει στην εκθετική οικογένεια και ο καθορισμός των δύο πρώτων ροπών (*moments*) αυτής, δηλαδή της μέσης τιμής και της διακύμανσης. Για το λόγο αυτό, η *GEE* μέθοδος είναι ημιπαραμετρική.

Κατ'αναλογία, καθορίζουμε την πιθανοφάνεια για τις περιθώριες κατανομές και έναν "χρησιμοποιούμενο" πίνακα συνδιακυμάνσεων ("*working*" *covariance matrix*) για το διάνυσμα των επαναλαμβανόμενων μετρήσεων από κάθε μονάδα. Οι Liang and Zeger (1986) πρότειναν την συγκεκριμένη μέθοδο για την μοντελοποίηση των περιθώριων πιθανοτήτων με την βοήθεια των *GLM* μοντέλων.

Η εφαρμογή της *GEE* μεθόδου προϋποθέτει να ενδιαφέρεται κάποιος για τις παραμέτρους **β** και όχι τόσο για τον πίνακα διακυμάνσεων-συνδιακυμάνσεων των επαναλαμβανόμενων μετρήσεων. Ειδικότερα, ο καθορισμός της δομής του αντιμετωπίζεται ως ένα πρόβλημα που πρέπει να ληφθεί υπόψη με κάποιον τρόπο έτσι, ώστε να προκύψουν λογικά στατιστικά τεστ για τις παραμέτρους του μοντέλου. Επομένως, δεν ενδείκνυται για καταστάσεις όπου το επιστημονικό ενδιαφέρον εστιάζεται γύρω από τη διακύμανση και/ή τις παραμέτρους συνδιακύμανσης.

Η συγκεκριμένη μέθοδος μοντελοποιεί χωριστά την επίδραση των συμμεταβλητών στην μεταβλητή απόκρισης (παλινδρόμηση) και την εντός-ομάδων εξάρτηση. Στη βιβλιογραφία αναφέρεται συχνά και ως *GEE*1 μέθοδος, προκειμένου για την διάκρισή της από άλλες παρεμφερείς, και είναι αυτή που εμφανίζεται συχνότερα στις εφαρμογές αρκετών στατιστικών λογισμικών. Ακολούθως της *GEE*1 μεθόδου αναπτύχθηκε η *GEE*2 (βλέπε Hardin and Hilbe, 2003), η οποία δεν κάνει τον προηγούμενο διαχωρισμό.

2.4.1 Γενικευμένα Γραμμικά Μοντέλα (GLM)

Τα περιθώρια μοντέλα αποκαλούνται και *GEE* μοντέλα, καθώς η συγκεκριμένη μέθοδος είναι η πιο δημοφιλής για την προσαρμογή και την εκτίμησή τους. Αποτελούν επέκταση των *GLM* στην περίπτωση των συσχετισμένων δεδομένων. Για το λόγο αυτό στη συνέχεια ακολουθεί μια σύντομη ανασκόπηση της θεωρίας τους. Τα *GLM* αντιπροσωπεύουν μια τάξη μοντέλων που χρησιμοποιούνται για την προσαρμογή μοντέλων παλινδρόμησης με σταθερές επιδράσεις (*fixed effects*) τόσο για κανονικά όσο και μη-κανονικά δεδομένα. Η εξαρτημένη μεταβλητή θεωρείται ότι ακολουθεί κατανομή, η οποία ανήκει στην λεγόμενη εκθετική οικογένεια και τα μοντέλα που ανήκουν στην κατηγορία των *GLM* διαφοροποιούνται ως προς τον τύπο της εξαρτημένης μεταβλητής σε σχέση πάντοτε με την κατανομή της. 'Ετσι, στην κατηγορία αυτή ανήκουν η γραμμική παλινδρόμηση για κανονικές μεταβλητές, η λογιστική παλινδρόμηση για δίτιμες και η παλινδρόμηση Poisson για μετρήσεις (*counts*).

Σε κάθε *GLM* υπάρχουν τρία θέματα που πρέπει να καθορίσουμε. Αρχικά, ο γραμμικός εκτιμητής (m_i) που είναι της μορφής

$$m_i = \mathbf{x}_i'\boldsymbol{\beta} .$$

όπου $\mathbf{x}_i$ είναι το διάνυσμα των ερμηνευτικών μεταβλητών για το *i*-υποκείμενο ενώ $\boldsymbol{\beta}$ είναι το διάνυσμα των σταθερών επιδράσεων. Το διάνυσμα $\mathbf{x}_i$ μπορεί να περιλαμβάνει συνεχείς μεταβλητές, δείκτριες ή ακόμα και αλληλεπιδράσεις. Στη συνέχεια καθορίζεται η συνάρτηση σύνδεσης (*link function*) $g(\cdot)$, η οποία σχετίζει την αναμενόμενη τιμή μ της εξαρτημένης μεταβλητής Y (δηλαδή $m_i = E(Y_i)$ με τον γραμμικό εκτιμητή.

$$g(m_i) = m_i \ .$$

Για παράδειγμα, στην κοινή πολλαπλή γραμμική παλινδρόμηση ως συνάρτηση $g(\cdot)$ χρησιμοποιούμε την ταυτοτική, δηλαδή $g(m_i) = m_i$ και επομένως $m_i = m_i$, ή

$$E(Y_i) = \mathbf{x}_i'\boldsymbol{\beta} \ .$$

Παρατηρούμε ότι κάτω από τον ταυτοτικό σύνδεσμο, η αναμενόμενη τιμή της εξαρτημένης μεταβλητής είναι μια γραμμική συνάρτηση των ερμηνευτικών μεταβλητών πολλαπλασιασμένες με τους αντίστοιχους συντελεστές.

Για δίτιμες μεταβλητές απόκρισης, η πλέον δημοφιλής επιλογή για της ανάλυση είναι η λογιστική παλινδρόμηση (βλέπε Hosmer and Lemeshow, 2000). Η μεταβλητή Y παίρνει τιμές 0 ή 1 και το μοντέλο είναι της μορφής

$$\log\left[\frac{P(Y_i=1)}{1-P(Y_i=1)}\right]=\mathbf{x}_i'\boldsymbol{\beta}\,.$$

Προκειμένου να κάνουμε την αντιστοιχία με τα προηγούμενα, διακρίνουμε ότι στην περίπτωση αυτή $P(Y_i=1)=E(Y_i)=m_i$ και επομένως είναι ο ***logit*** σύνδεσμος $g(m_i)=\log\left(\frac{m_i}{1-m_i}\right)$ ο οποίος σχετίζει την αναμενόμενη τιμή της απόκρισης με τον γραμμικό εκτιμητή.

Παρόμοια, τέλος, η παλινδρόμηση Poisson (Cameron and Trivedi, 1998), που χρησιμοποιείται για την μοντελοποίηση μετρήσεων (*count data*), είναι της μορφής

$$m_i=\exp(\mathbf{x}_i'\boldsymbol{\beta})\,,$$

ή διαφορετικά

$$\log(m_i)=\mathbf{x}_i'\boldsymbol{\beta}\,.$$

Η παραπάνω σχέση δείχνει ότι ο σύνδεσμος $g(m_i)=\log(m_i)$ ενώνει την αναμενόμενη τιμή της μεταβλητής Y με τον γραμμικό εκτιμητή.

Μέχρι στιγμής, έχουμε καθορίσει τον τρόπο με τον οποίο η μέση τιμή m_i σχετίζεται με τις συμμεταβλήτες. Σε ένα *GLM*, επιπρόσθετα χρειάζεται να προσδιορίσουμε τη μορφή της δεσμευμένης διακύμανσης της μετάβλητής Y, δοθέντων των ερμηνευτικών μεταβλητών. Αυτό γίνεται μέσω της σχέσης

$$Var(Y_i)=f\,u(m_i)\,,$$

όπου $u(\cdot)$ είναι μια γνωστή συνάρτηση διακύμανσης και f μια παράμετρος κλίμακας, η οποία είτε είναι γνωστή είτε πρόκειται να εκτιμηθεί. Παραδείγματος χάρην, στην κοινή γραμμική παλινδρόμηση είναι $u(m_i)=1$ και η παράμετρος f αντιπροσωπεύει την διακύμανση της δεσμευμένης κανονικής κατανομής της $Y\,|\,\mathbf{x}$. Στην περίπτωση της δίτιμης μεταβλητής απόκρισης, λόγω της κατανομής Bernoulli θα είναι $u(m_i)=m_i(1-m_i)$ και τυπικά $f=1$. Τέλος, όταν η εξαρτημένη μεταβλητή ακολουθεί κατανομή Poisson, τότε $u(m_i)=m_i$ και επίσης $f=1$. Συμπερασματικά γίνεται αντιληπτό, ότι τόσο η συνάρτηση σύνδεσης όσο και ο

καθορισμός της διακύμανσης συνήθως εξαρτώνται από την κατανομή της μεταβλητής απόκρισης. Σημειώνουμε ότι πέρα των προαναφερθέντων, υπάρχουν και πολλά άλλα είδη μοντέλων στην οικογένεια των *GLM*.

Στη συνέχεια γίνεται η εκτίμηση των παραμέτρων **β**, μέσα από τη λύση των εξισώσεων

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{N} (\frac{\partial m_i}{\partial \boldsymbol{\beta}})'(Var(Y_i))^{-1}(Y_i - m_i) = \mathbf{0},$$

όπου N είναι το σύνολο των υποκειμένων. Σύμφωνα με το μοντέλο που χρησιμοποιείται, οι παραπάνω εξισώσεις παίρνουν και την ανάλογη μορφή. Για παράδειγμα, στην περίπτωση της πολλαπλής παλινδρόμησης θα έχουμε

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{N} \mathbf{x}_i(Y_i - \mathbf{x}_i'\boldsymbol{\beta}) = \mathbf{0} \ .$$

Οι παραπάνω εξισώσεις, στην αρχική τους μορφή, εξαρτώνται μόνο από το μέσο και τη διακύμανση της μεταβλητής Y και συνεπώς η ακριβής κατανομή για την Y δεν είναι απαραίτητη για την εκτίμηση του διανύσματος **β**. Οπότε, οι εκτμητές που προκύπτουν καλούνται "*quasi-likelihood*" εκτιμητές (Wedderburn, 1974).

Τα *GLM* είναι μοντέλα σταθερών επιδράσεων, τα οποία υποθέτουν ότι όλες οι παρατηρήσεις είναι ανεξάρτητες μεταξύ τους. Έτσι, δεν είναι γενικώς κατάλληλα για την ανάλυση διαχρονικών δεδομένων. Εν τούτοις, μπορούν να επεκταθούν ώστε να λαμβάνουν υπόψη την υπάρχουσα συσχέτιση σε αυτό το είδος των δεδομένων. Ακριβώς αυτό κατάφεραν οι Liang and Zeger, (1986) με την ανάπτυξη των *GEE* μοντέλων.

2.4.2 Βασικές ιδέες της GEE μεθόδου

Καθώς τα *GEE* μοντέλα θεωρούνται επέκταση των *GLM* για συσχετισμένα δεδομένα, οι ρυθμίσεις (*specifications*) που πρέπει να γίνουν για τη μέθοδο που μελετάμε περιλαμβάνουν αυτές των *GLM* με μία επιπρόσθετη. Καταρχήν, έστω Y_{it} η απόκριση για το *i*-υποκείμενο στον χρόνο *t*, $\mathbf{x}_{it} = (x_{it1}, x_{it2}, ..., x_{itp})'$ το $p \times 1$ διάνυσμα των συμμεταβλητών και

$\boldsymbol{\beta} = (b_1, b_2, ..., b_p)'$ το $p \times 1$ διάνυσμα των άγνωστων παραμέτρων. Ο γραμμικός εκτιμητής καθορίζεται ως

$$m_{it} = \mathbf{x}_{it}^{'}\boldsymbol{\beta}\,.$$

Στη συνέχεια επιλέγουμε την συνάρτηση $g(\cdot)$, ώστε να συνδέσουμε την μέση τιμή $m_{it} = E(Y_{it})$ με τις συμμεταβλητές

$$g(m_{it}) = m_{it} = \mathbf{x}_{it}^{'}\boldsymbol{\beta}\,.$$

Στο επόμενο βήμα της προσέγγισης αυτής προσπαθούμε να περιγράψουμε τον τρόπο με τον οποίο η διακύμανση των Y_{it} εξαρτάται από την μέση τιμή:

$$Var(Y_{it}) = f u(m_{it}),$$

όπου με βάση την προηγούμενη ενότητα η $u(\cdot)$ είναι γνωστή συνάρτηση διασποράς και f παράμετρος κλίμακας.

Πέραν αυτών, σε ένα *GEE* μοντέλο οφείλουμε επιπλέον να καθορίσουμε τη δομή του "χρησιμοποιούμενου" πίνακα συσχέτισης ("*working*" *correlation matrix*) **R** των επαναλαμβανόμενων μετρήσεων. Ας υποθέσουμε ότι υπάρχουν T το πλήθος χρονικά σημεία στα οποία γίνονται οι μετρήσεις Δηλαδή, για το i-υποκείμενο θα είναι $\mathbf{Y}_{\mathrm{i}} = (Y_{i1}.Y_{i2}, ..., Y_{iT})'$. Τότε ο πίνακας θα είναι διάστασης $T \times T$, όπου το (s, v) στοιχείο του είναι η συσχέτιση των Y_{is} και Y_{iv}, η οποία είναι γνωστή ή εκτιμώμενη ή ακόμα και υποτιθέμενη. Δεν είναι, ωστόσο, απαραίτητο από όλα τα υποκείμενα που υπόκεινται στην έρευνα να έχουν ληφθεί και οι T μετρήσεις. Τότε για το i-υποκείμενο ο πίνακας συσχέτισης, έστω $\mathbf{R}_i$, θα είναι διάστασης $T_i \times T_i$ $(T_i < T)$. Επίσης, υποθέτουμε εξαρχής ο πίνακας **R**, και κατ'επέκταση ο $\mathbf{R}_i$, εξαρτώνται από ένα διάνυσμα παραμέτρων συσχέτισης **a** και το οποίο θα εκτιμηθεί από τα δεδομένα.

Στην πράξη, το μοντέλο που θα χρησιμοποιηθεί δεν είναι ποτέ ιδιαίτερα ακριβές και σωστό. Κανείς, εξάλλου, δεν θα μπορούσε να ισχυριστεί το αντίθετο, εφόσον από την αρχή δεν διαθέτουμε καμιά πληροφορία για τις συσχετίσεις ανάμεσα στις μετρήσεις και κάνουμε κάποιες αυθαίρετες υποθέσεις. Γενικά, συνιστάται να επιλέγουμε τη μορφή του πίνακα συσχέτισης ώστε να είναι συνεπής με τις παρατηρούμενες συσχετίσεις. Επειδή ενδέχεται η δομή που ορίζουμε να απέχει από αυτή που ισχύει στην πραγματικότητα, δικαιολογείται και ο όρος "χρησιμοποιούμενος" που του αποδίδουμε.

Εύλογα, λοιπόν, τίθεται το ερώτημα για ποιο λόγο να κάνουμε οποιαδήποτε υπόθεση για τον συγκεκριμένο πίνακα και να μην δεχτούμε εξ αρχής την πιο απλή περίπτωση που είναι η ανεξαρτησία. Η απάντηση έγκειται στο ότι ένα μοντέλο που προσεγγίζει ικανοποιητικά την πραγματική συσχέτιση (και επομένως τον πίνακα συνδιακύμανσης) μπορεί να βελτιώσει την αποτελεσματικότητα των εκτιμήσεων σε μεγάλο βαθμό σε σχέση με το μοντέλο της ανεξαρτησίας. Οι Liang and Zeger (1986) παρατήρησαν ότι στην περίπτωση της ανεξαρτησίας, οι εκτιμήσεις είναι αποτελεσματικές μόνο όταν οι πραγματικές συσχετίσεις είναι από ασθενείς μέχρι μέτριες.

Επιπλέον, οι εκτιμήσεις των παραμέτρων που προκύπτουν από τις λύσεις των εξισώσεων αυτών εξακολουθούν να είναι έγκυρες, ακόμα και όταν η δομή που έχουμε θέσει από την αρχή δεν είναι η σωστή. Και αυτό συμβαίνει διότι η συνέπεια των εκτιμητριών εξαρτάται μόνο από την πρώτη ροπή (δηλαδή από τη μέση τιμή) και όχι από την δεύτερη (δηλαδή την διακύμανση) (βλέπε Agresti, 2002 σελ. 467). Παρά το γεγονός, όμως, ότι η ποιότητα των εκτιμήσεων δεν επηρεάζεται, δεν ισχύει το ίδιο και με τα τυπικά σφάλματα αυτών καθώς η συσχέτιση συνδέεται άμεσα με αυτά. Τέλος, αγνοώντας την υπάρχουσα συσχέτιση οδηγούμαστε συχνά (αλλά όχι πάντοτε) σε υποεκτίμηση της διασποράς των εκτιμητών με αποτέλεσμα οι επιδράσεις να εμφανίζονται πιο σημαντικές από ότι είναι στην πραγματικότητα.

2.4.2.1 Μορφές του πίνακα συσχέτισης

Οι επιλογές που υπάρχουν όσον αφορά τη μορφή του πίνακα συσχέτισης είναι αρκετές. Το κίνητρο για να χρησιμοποιήσει κάποιος αυτή της ανεξαρτησίας είναι το γεγονός ότι στην περίπτωση που ο αριθμός των πειραματικών μονάδων είναι μεγάλος σε σχέση με τον αριθμό των επαναλήψεων για κάθε μία από αυτές, η επίδραση της συσχέτισης είναι μικρή. Ειδικότερα, η λύση των *GEE* οδηγεί στο ίδιο αποτέλεσμα που θα παίρναμε αν εφαρμόζαμε παλινδρόμηση για ανεξάρτητα δεδομένα. Η δομή αυτή, ιδιαίτερα για διαχρονικά δεδομένα, δεν αποτελεί λογική επιλογή. Για δίτιμες μεταβλητές απόκρισης μπορεί να οδηγήσει σε μεγάλη απώλεια αποτελεσματικότητας όταν στο μοντέλο υπάρχουν χρονοεξαρτώμενες μεταβλητές (Fitzmaurice, 1995).

Εναλλακτικά, θα μπορούσε να χρησιμοποιηθεί μία πιο ελαστική και ρεαλιστική επιλογή. Για παράδειγμα, μπορούμε να υποθέσουμε ότι οι συσχετίσεις $corr(Y_i, Y_j)$ είναι ίδιες για κάθε $i, j = 1, 2, ..., T$ (***exchangeable*** ή ***compound symmetry*** δομή). Ο πίνακας, τότε, έχει τη μορφή

$$\mathbf{R(a)} = \begin{pmatrix} 1 & a & \mathbf{L} & a \\ a & 1 & \mathbf{L} & a \\ \mathbf{M} & \mathbf{M} & \mathbf{O} & \mathbf{M} \\ a & a & \mathbf{L} & 1 \end{pmatrix},$$

όπου με *α* συμβολίζουμε τη τιμή της κοινής συσχέτισης. Η υπόθεση ότι η συσχέτιση ανάμεσα σε δύο οποιεσδήποτε μετρήσεις είναι η ίδια δεν μπορεί πάντοτε να δικαιολογηθεί σε μια διαχρονική μελέτη. Προτιμάται, όμως, όταν οι επαναλαμβανόμενες παρατηρήσεις δεν είναι χρονικές αλλά λαμβάνονται, για παράδειγμα από τα μέλη μιας οικογένειας ή τους μαθητές μιας τάξης.

Μια ακόμα πιο ρεαλιστική υπόθεση είναι οι συσχετίσεις να διαφέρουν για κάθε ζεύγος. Στην προκειμένη περίπτωση, αν *T* είναι ο αριθμός των μετρήσεων για κάθε μονάδα, τότε το μοντέλο θα έχει συνολικά $\frac{T(T-1)}{2}$ παραμέτρους. Εν τούτοις, ενδείκνυται μόνο όταν ο αριθμός *T* είναι μικρός, αφού διαφορετικά προστίθενται πολλές επιπλέον παράμετροι.

Μία άλλη επιλογή είναι το πρώτης τάξης αυτοπαλίνδρομο μοντέλο ***AR*-1**, όπου $corr(Y_i, Y_j) = a^{|i-j|}$. Εδώ η συσχέτιση μειώνεται όσο οι μετρήσεις απέχουν χρονικά μεταξύ τους. Επομένως, το συγκεκριμένο μοντέλο είναι μια λογική δομή για επαναλαμβανόμενες στο χρόνο παρατηρήσεις. Το μειονέκτημα του έχει να κάνει με το γεγονός ότι οι συσχετίσεις εξασθενούν πολύ γρήγορα καθώς το διάστημα που τις χωρίζει αυξάνεται. Ένας τέτοιος πίνακας θα είναι της μορφής

$$\mathbf{R(a)} = \begin{pmatrix} 1 & a & \mathbf{L} & a^{T-1} \\ a & 1 & \mathbf{L} & a^{T-2} \\ \mathbf{M} & \mathbf{M} & \mathbf{O} & \mathbf{M} \\ a^{T-1} & a^{T-2} & \mathbf{L} & 1 \end{pmatrix}.$$

Καταλήγωντας, συμπεραίνουμε ότι η επιλογή του πίνακα μπορεί να γίνει με βάση τη φύση των δεδομένων. Όταν οι πραγματικές συσχετίσεις είναι μέτριες, τότε οποιαδήποτε δομή και αν υποθέσουμε παίρνουμε παρόμοιες εκτιμήσεις για παραμέτρους και τυπικά σφάλματα.

2.4.3 Εκτίμηση των GEE μοντέλων

Το επόμενο βήμα αφορά την εκτίμηση του διανύσματος **β** αλλά και του πίνακα συνδιακυμάνσεων αυτού. Για το *i*-υποκείμενο, έστω $\mathbf{A}_i$ ο $T \times T$ διαγώνιος πίνακας με διαγώνια στοιχεία $u(m_{it})$, $t = 1, 2, ..., T$. Τότε ο "χρησιμοποιούμενος" πίνακας συνδιακυμάνσεων για το διάνυσμα $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{iT})'$ με τον οποίο θα εργαστούμε είναι:

$$\mathbf{V}_i(\mathbf{a}) = f\,\mathbf{A}_i^{1/2}\mathbf{R}_i(\mathbf{a})\mathbf{A}_i^{1/2}.$$

- Σημειώνουμε ότι στην περίπτωση που ο πίνακας **R** είναι ο πραγματικός πίνακας συσχετίσεων του $\mathbf{Y}_i$, τότε $\mathbf{V}_i(\mathbf{a}) = \mathrm{cov}(\mathbf{Y}_i)$.

Το διάνυσμα των παραμέτρων **β** προκύπτει από τη λύση των εξισώσεων :

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{N}\left(\frac{J\boldsymbol{\mu}_i}{J\boldsymbol{\beta}}\right)'[\mathbf{V}_i(\hat{\mathbf{a}})]^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}_\mathbf{p}\ , \qquad (2.6)$$

όπου $\boldsymbol{\mu}_i = (m_{i1}, m_{i2}, ..., m_{iT})'$, $\hat{\mathbf{a}}$ είναι μια συνεπής εκτίμηση του **a** και $\mathbf{0}_\mathbf{p}$ είναι το $p \times 1$ διάνυσμα $(0, 0, ..., 0)'$. Οι Liang and Zeger, (1986) πρότειναν την αντικατάσταση των f και **a** με αντίστοιχους συνεπείς εκτιμητές $\hat{f}(\boldsymbol{\beta})$ και $\hat{\mathbf{a}}(\boldsymbol{\beta}, f)$, οι οποίοι προκύπτουν χρησιμοποιώντας συναρτήσεις των τυποποιημένων καταλοίπων του Pearson (***standardized Pearson residuals***). Για την εκτίμηση, μάλιστα, του **a** προτάθηκαν από τους ίδιους, εκτιμητές για διάφορες επιλογές του πίνακα των συσχετίσεων εφόσον ο εκτιμητής για το **a** εξαρτάται από τον πίνακα αυτό**.**

Η λύση των εξισώσεων δίνεται μέσω επαναλήψεων ανάμεσα σε ***quasi-likelihood*** μεθόδους για την εκτίμηση του **β** και μιας ανθεκτικής ***(robust***) μεθόδου για την εκτίμηση του **a** ως μια συνάρτηση του **β**. Η διαδικασία έχει ως εξής:

1) Δοθέντων των εκτιμήσεων για τον πίνακα $\mathbf{R}_i(\mathbf{a})$ και την παράμετρο f, υπολογίζουμε μια ***updated*** εκτίμηση του **β** μέσω επαναληπτικά επανασταθμιζόμενων ελαχίστων τετραγώνων (*iteratively reweighted least squares*).

2) Δοθείσης της εκτίμησης για το **β** , υπολογίζουμε τα τυποποιημένα κατάλοιπα r_{ij} του Pearson μέσω της σχέσης :

$$r_{it} = \frac{Y_{it} - \hat{m}_{it}}{\sqrt{[\mathbf{V}_i]_{tt}}}.$$

3) Χρησιμοποιούμε τα παραπάνω κατάλοιπα για να πάρουμε συνεπείς εκτιμήσεις των **a** και φ

4) Επαναλαμβάνουμε τα βήματα 1,2 και 3 μέχρι να έχουμε σύγκλιση.

Οι αλγόριθμοι για τους *GEE* εκτιμητές δεν είναι απαραίτητο να συγκλίνουν καθώς συχνά από την πρώτη επανάληψη μπορούμε να έχουμε ικανοποιητικά αποτελέσματα (Lipsitz et al., 1991). Μια συνήθης προσέγγιση για τον πίνακα διακυμάνσεων-συνδιακυμάνσεων του **β** βασίζεται στον αντίστροφο του πίνακα πληροφορίας του Fisher και στη βιβλιογραφία αναφέρεται ως "*model-based*"

$$\hat{Var}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{n} \left(\frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\beta}} \right)' \hat{\mathbf{V}}_i^{-1} \left(\frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\beta}} \right) \right)^{-1},$$

όπου με $\hat{\mathbf{V}}_i$ δηλώνουμε τον πίνακα $\mathbf{V}_i(\hat{\mathbf{a}})$. Όπως έχει δειχθεί από τον Royall (1986), ο εν λόγω εκτιμητής δεν θα είναι συνεπής στην περίπτωση που το μοντέλο που έχουμε θεωρήσει δεν είναι σωστό.

Οι Liang and Zeger (1986) συνιστούν την παρακάτω εκτίμηση για τον ίδιο πίνακα, η οποία προτάθηκε από τον Royall (1986) και είναι ανθεκτική (*robust*). Μάλιστα, αποτελεί συνεπή εκτιμήτρια του πίνακα $\hat{Var}(\hat{\boldsymbol{\beta}})$ ακόμα και αν η δομή που έχουμε θέσει για τον πίνακα συσχέτισης των $\mathbf{Y}_i$ δεν είναι σωστή

$$\hat{Var}(\hat{\boldsymbol{\beta}}) = M_0^{-1} M_1 M_0^{-1},$$

όπου $M_0 = \sum_{i=1}^{N} \left(\frac{\partial \hat{m}_i}{\partial \boldsymbol{\beta}} \right)' \hat{\mathbf{V}}_i^{-1} \left(\frac{\partial \hat{m}_i}{\partial \boldsymbol{\beta}} \right)$ και $M_1 = \sum_{i=1}^{N} \left(\frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\beta}} \right)' \hat{\mathbf{V}}_i^{-1} (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)' \hat{\mathbf{V}}_i^{-1} \left(\frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\beta}} \right)$.

Έχοντας εκτιμήσει το διάνυσμα των παραμέτρων **β**, μπορούμε να προχωρήσουμε σε ελέγχους υποθέσεων. Ας θεωρήσουμε τη μηδενική υπόθεση $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$, όπου $\mathbf{C}$ ένας $c \times p$ πίνακας σταθερών και **d** ένα $p \times 1$ διάνυσμα σταθερών αριθμών. Έτσι, η H_0 αντιστοιχεί σε c το πλήθος περιορισμούς για τα στοιχεία του **β**. Το τεστ που χρησιμοποιείται είναι αυτό του Wald, το οποίο παίρνει τη μορφή:

$$Q_C = (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})' [\mathbf{C} \hat{Var}(\hat{\boldsymbol{\beta}}) \mathbf{C}']^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}).$$

Καθώς το διάνυσμα $\hat{\boldsymbol{\beta}}$ ακολουθεί ασυμπτωτικά κανονική κατανομή, το Q_C ακολουθεί X_c^2 κατανομή κάτω από την H_0. Το συγκεκριμένο στατιστικό τεστ μας επιτρέπει να κάνουμε ελέγχους για κάποιο υποσύνολο ή ακόμα και για μία συνιστώσα του $\hat{\boldsymbol{\beta}}$. Για μηδενικές υποθέσεις της μορφής $H_0 : \boldsymbol{\beta}^* = \mathbf{0}$, όπου $\boldsymbol{\beta}^*$ είναι ένα διάνυσμα $p'\times 1$ υποσύνολο του $\boldsymbol{\beta}$, και $\mathrm{H}_0 : b_k = 0$ για k=1,2,…,p το τεστ του Wald παίρνει τη μορφή $\hat{\boldsymbol{\beta}}^{*\prime} Var(\hat{\boldsymbol{\beta}}^*)\hat{\boldsymbol{\beta}}^* \sim X_{p'}^2$ και $\dfrac{\hat{b}_k^2}{\mathrm{var}(\hat{b}_k^2)} \sim X_1^2$ αντίστοιχα.

2.4.4 *Ιδιότητες των GEE εκτιμητών – Πλεονεκτήματα και αδυναμίες*

Οι *QL* εκτιμητές έχουν παρόμοιες ιδιότητες με τους αντίστοιχους *ML* εκτιμητές (McCullagh, 1983). Ο *QL* εκτιμητής $\hat{\boldsymbol{\beta}}$ ακολουθεί ασυμπτωτικά κανονική κατανομή με πίνακα διακυμάνσεων $V = \left(\sum_i \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)' [u(m_i)]^{-1} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right) \right)^{-1}$. Επιπλέον, είναι συνεπής (*consistent*) εκτιμητής του $\boldsymbol{\beta}$ ($\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$), ακόμα και όταν η συνάρτηση διακύμανσης δεν έχει καθοριστεί σωστά. Χρειάζεται, όμως, να έχει χρησιμοποιηθεί η σωστή συνάρτηση σύνδεσης (*link function*) και ο σωστός γραμμικός συνδυασμός των συμμεταβλητών (*linear predictor*). Δηλαδή, υποθέτωντας ότι το μοντέλο $g(\boldsymbol{\mu}_i) = \sum_t b_t x_{it}$ είναι σωστό, τότε η συνέπεια του $\hat{\boldsymbol{\beta}}$ διατηρείται και στην περίπτωση που η συνάρτηση διακύμανσης δεν είναι $u(m_i)$ (Agresti, 2002).

Όσον αφορά κατηγορικά δεδομένα, η *GEE* προσέγγιση προτιμάται από την *ML* καθώς στο υπολογιστικό κομμάτι είναι πιο απλή. Μερικά από τα πλεονεκτήματά της είναι ότι δεν απαιτείται η γνώση μιας πολυδιάστατης κατανομής αλλά και η συνέπεια των εκτιμητών, ακόμα και όταν δεν έχουμε προσδιορίσει σωστά τον πίνακα των συσχετίσεων. Μολαταύτα, παρουσιάζει και ορισμένες αδυναμίες. Καθώς δεν προσδιορίζει την από κοινού κατανομή των μετρήσεων, δεν έχουμε στη διάθεσή μας κάποια συνάρτηση πιθανοφάνειας. Κατά συνέπεια, δεν μπορούμε να χρησιμοποιήσουμε τεστ που βασίζονται σε αυτήν για έλεγχους καλής

προσαρμογής, σύγκριση μοντέλων και εξαγωγή συμπερασμάτων για τις παραμέτρους του μοντέλου. Αντί αυτών, χρησιμοποιούνται τα στατιστικά τεστ του Wald που κατασκευάζονται με την ασυμπτωτική κανονικότητα των εκτιμητών αλλά και τον εκτιμώμενο πίνακα συνδιακύμανσής τους.

Τα εν λόγω τεστ παρουσιάζουν, όμως, ορισμένα μειονεκτήματα. Αν το μέγεθος του δείγματος δεν είναι αρκετά μεγάλο, τα εμπειρικά τυπικά σφάλματα (***empirically based standard errors***) τείνουν να υποεκτιμούν τα πραγματικά (π.χ. Firth, 1993b). Επίσης, τα τεστ για τους ελέγχους υποθέσεων που αφορούν τις παραμέτρους εξαρτώνται από την κλίμακα μέτρησης (***measurement scale***) και επηρεάζονται από τους μετασχηματισμούς. Τέλος, απαιτούν την εκτίμηση του πίνακα συνδιακυμάσεων των εκτιμώμενων παραμέτρων. Το πρόβλημα που προκύπτει στο σημείο αυτό είναι ότι αν το μέγεθος του δείγματος (δηλ. οι πειραματικές μονάδες) είναι μικρό ή ο αριθμός των επαναλαμβανόμενων μετρήσεων ανά μονάδα είναι μεγάλος, τότε οι εκτιμήσεις των διακυμάνσεων και των συνδιακυμάνσεων ίσως να είναι ασταθείς (***unstable***).

Κατά την εφαρμογή των ***GEE*** οφείλουμε να είμαστε προσεκτικοί σε κάποια σημεία. Σε πολλές μελέτες επαναλαμβανόμενων μετρήσεων συμβαίνει συχνά να υπάρχουν σε μία ή περισσότερες ομάδες ελλιπή δεδομένα. Αναλύοντας μόνο τα παρατηρούμενα δεδομένα, το πιθανότερο είναι να καταλήξουμε σε μεροληπτικούς εκτιμητές. Αναφορικά με μεθόδους που δεν βασίζονται στην πιθανοφάνεια, όπως οι ***GEE***, δεν χρειάζεται να ληφθούν υπ'όψη τα συγκεκριμένα δεδομένα μόνο στην περίπτωση που λείπουν εντελώς τυχαία (***MCAR***). Υπάρχει, ωστόσο το ενδεχόμενο τα δεδομένα να λείπουν απλώς τυχαία (***MAR***). Στην περίπτωση αυτή, αποδεικνύεται η ακαταλληλότητα των ***GEE*** εκτιμητών (Kenward et al., 1994). Επιπλέον, αν στο μοντέλο υπάρχουν χρονοεξαρτώμενες (***time-dependen***t) συμμεταβλητές, ο εκτιμητής του διανύσματος **β** ίσως να μην είναι συνεπής (Pepe and Anderson, 1994). Στην περίπτωση αυτή, χρησιμοποιούμε διαγώνιο πίνακα συνδιακυμάνσεων ή επαληθεύουμε ότι το αποτέλεσμα σε κάποιο συγκεκριμένο χρονικό σημείο δεν εξαρτάται από τις τιμές των συμμεταβλητών σε άλλα χρονικά σημεία, έχοντας δεσμεύσει ως προς τις τιμές τους για το σημείο αυτό.

2.4.5 GEE για μεταβλητές απόκρισης με περισσότερα από δύο επίπεδα (polytomous)

Στην εργασία του Prentice, (1988) δίνεται μια επέκταση των *GEE*, σύμφωνα με την οποία οι ανά δύο συσχετίσεις και οι περιθώριες πιθανότητες της απόκρισης μοντελοποιούνται από κοινού. Οι Miller et al., (1993) περιγράφουν τον τρόπο που η μέθοδος αυτή εφαρμόζεται σε διαχρονικά *polytomous* δεδομένα.

Όπως και πριν, έστω Y_{it} η απόκριση για το *i*-άτομο στην *t*-χρονική στιγμή με $i = 1, 2, ..., N$ και $t = 1, 2, ..., T$ ενώ τα επίπεδα της απόκρισης είναι *K* (*k* = 1,2,...,*K*). Τότε, θα είναι

$$\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{iT})' ,$$

όπου $Y_{it} = k$, αν το *i*-άτομο ανήκει στο *k*-επίπεδο την *t*-χρονική στιγμή. Ομοίως με την μέθοδο των Stram et al. (1988), δημιουργείται το $KT \times 1$ διάνυσμα των δείκτριων μεταβλητών ,

$$\mathbf{Y}_i^{**} = (Y_{i11}, ..., Y_{i1K}, Y_{i21}, ..., Y_{i2K}, ..., Y_{iTK})',$$

όπου $Y_{itk} = 1$ αν $Y_{it} = k$ και $Y_{itk} = 0$ διαφορετικά. Εφόσον $\sum_k Y_{itk} = 1$ για κάθε *i* και *t*, η *K*-οστή δείκτρια είναι περιττή (***redundant***). Επομένως, αγνοώντας τη προκύπτει το $(K-1)T \times 1$ διάνυσμα

$$\mathbf{Y}_i^{*} = (Y_{i11}, ..., Y_{i1,K-1}, Y_{i21}, ..., Y_{i2,K-1}, ..., Y_{iT,K-1})' .$$

Η περιθώρια προσδοκία για την Y_i^* μπορεί να εκφραστεί ως

$$\boldsymbol{\pi}_i^{*} = (p_{i11}, ..., p_{i1,K-1}, p_{i21}, ..., p_{i2,K-1}, ..., p_{iT,K-1})' ,$$

όπου $\sum_k p_{itk} = 1$ για κάθε *i* και *t*.

Για την περιγραφή των δεδομένων μπορούμε να χρησιμοποιήσουμε κάποιον από τους συνδέσμους που περιγράψαμε στην ενότητα 2.2. Λόγου χάρη, με τον αθροιστικό ***logit*** παίρνουμε το μοντέλο

$$\log\text{it}[P(Y_{it} \le k)] = \log \frac{p_{it1} + p_{it2} + ... + p_{itk}}{p_{it,k+1} + p_{it,k+2} + ... + p_{itK-1}} = \mathbf{x}_{itk}'\boldsymbol{\beta} .$$

Ο εκτιμητής για το $p\times 1$ διάνυσμα $\boldsymbol{\beta}$ των παραμέτρων του μοντέλου προκύπτει από τη λύση των εξισώσεων

$$\mathbf{U}(\boldsymbol{\beta})=\sum_{i=1}^{N}\mathbf{D}_i^{'}\mathbf{V}_i^{-1}(\mathbf{Y}_i^{*}-\boldsymbol{\pi}_i^{*})=\mathbf{0}\ , \tag{2.7}$$

όπου $\mathbf{D}_i=\partial\boldsymbol{\pi}_i^{*}/\partial\boldsymbol{\beta}$ και $\mathbf{V}_i=\mathrm{cov}(\mathbf{Y}_i^{*})$. Ο πίνακας $\mathbf{V}_i$ έχει στοιχεία

$$\mathrm{cov}(Y_{itk},Y_{it'k'})=\begin{cases}p_{itk}(1-p_{itk}),\ \text{αν}\ \ t=t',k=k'\\ -p_{itk}p_{itk'},\ \ \text{αν}\ \ t=t',k\neq k'\\ \dfrac{\mathrm{corr}(Y_{itk},Y_{it'k'})}{[p_{itk}(1-p_{itk})p_{it'k'}(1-p_{it'k'})]^{1/2}},\ \ \text{αν}\ \ t\neq t'\ \text{και}\ \forall\ k,k'.\end{cases}$$

Είδαμε ότι υπάρχουν διάφορες επιλογές για τη δομή των συσχετίσεων $corr(Y_{itk},Y_{it'k'})$, οι οποίες εξαρτώνται από ένα $q\times 1$ διάνυσμα παραμέτρων $\mathbf{a}$. Έτσι, ο "χρησιμοποιούμενος" πίνακας συνδιακύμανσης $\mathbf{V}_i$ εξαρτάται τόσο από το $\boldsymbol{\beta}$ όσο και από το $\mathbf{a}$.

Για την μοντελοποίηση των συσχετίσεων μεταξύ των χρονικών στιγμών, ορίζουμε $\mathbf{Z}_i=\mathbf{Z}_i(\boldsymbol{\beta})$ ως το $[(K-1)^2T(T-1)]/2\times 1$ διάνυσμα που περιέχει τις συσχετίσεις ανάμεσα στις δείκτριες μεταβλητές του διανύσματος $\mathbf{Y}_i^{*}$

$$\mathbf{Z}_i(\boldsymbol{\beta})=(Z_{i(11)(21)}(\boldsymbol{\beta}),Z_{i(11)(22)}(\boldsymbol{\beta}),...,Z_{i(11)(2[K-1])}(\boldsymbol{\beta}),...,Z_{i([T-1][K-1])(T[K-1])}(\boldsymbol{\beta}))^{'}\ ,$$

όπου

$$Z_{i(tk)(t'k')}=Z_{i(tk)(t'k')}(\boldsymbol{\beta})=\frac{(Y_{itk}-p_{itk}(\boldsymbol{\beta}))(Y_{it'k'}-p_{it'k'}(\boldsymbol{\beta}))}{[p_{itk}(\boldsymbol{\beta})(1-p_{itk}(\boldsymbol{\beta}))p_{it'k'}(\boldsymbol{\beta})(1-p_{it'k'}(\boldsymbol{\beta}))]^{1/2}}$$

για $t\neq t'$ και για κάθε k ή k'. Ισχύει ότι $E[Z_{i(tk)(t'k')}]=\mathrm{corr}(Y_{itk},Y_{it'k'})=h_{i(tk)(t'k')}(\mathbf{a})=d_{i(tk)(t'k')}$.

Για την εκτίμηση του $\mathbf{a}$ εισάγουμε ένα δεύτερο σύστημα εξισώσεων

$$\mathbf{U}(\mathbf{a})=\sum_{i=1}^{N}\mathbf{E}_i^{'}\mathbf{W}_i^{-1}(\mathbf{Z}_i-\boldsymbol{\eta}_i(\mathbf{a}))=\mathbf{0}\ , \tag{2.8}$$

όπου $\mathbf{E}_i=\partial\boldsymbol{\eta}_i(\mathbf{a})/\partial\mathbf{a}$ και

$$\boldsymbol{\eta}_i(\mathbf{a})=(h_{i(11)(21)}(\mathbf{a}),h_{i(11)(22)}(\mathbf{a}),...,h_{i(11)(2[K-1])}(\mathbf{a}),...,h_{i([T-1][K-1])(T[K-1])}(\mathbf{a}))^{'}.$$

Ο πίνακας $\mathbf{W}_i = \mathrm{cov}(\mathbf{Z}_i)$ είναι ο αντίστοιχος πίνακας συνδιακύμανσης, ο οποίος είναι μπλοκ-διαγώνιος με στοιχεία

$$\mathrm{var}(Z_{i(tk)(t'k')}) = 1 + \frac{(1-2p_{itk})(1-2p_{it'k'})}{(p_{itk}(1-p_{itk})p_{it'k'}(1-p_{it'k'}))^{1/2}} d_{i(tk)(t'k')} - d^{2}_{i(tk)(t'k')}$$

και

$$\mathrm{cov}(Z_{i(tk)(t'k')}, Z_{i(tk)(t'h)}) = -(1-2p_{itk})\left[\frac{p_{it'k}d_{i(tk)(t'h)}}{(p_{itk}(1-p_{itk})p_{it'k'}(1-p_{it'k'}))^{1/2}} + \frac{p_{it'h}d_{i(tk)(t'k')}}{(p_{itk}(1-p_{itk})p_{it'h}(1-p_{it'h}))^{1/2}}\right]$$
$$- \frac{p_{it'k}p_{it'h}}{(p_{it'k'}(1-p_{it'k'})p_{it'h}(1-p_{it'h}))^{1/2}} - d_{i(tk)(t'k')}d_{i(tk)(t'h)}$$

για $t \neq t'$, $k' \neq h$ και για οποιαδήποτε k, k'.

Η λύση των εξισώσεων (2.7) και (2.8) προκύπτουν με τη χρήση επαναληπτικών επανασταθμιζόμενων ελαχίστων τετραγώνων (*iteratively reweighted least squares*).

2.4.6 Εναλλακτική μέθοδος για την ανάλυση διατάξιμων δεδομένων

2.4.6.1 Γενικά στοιχεία

Στους Stram et al., (1988) οφείλεται η πρώτη προσέγγιση για την ανάλυση των επαναλαμβανόμενων παρατηρήσεων όταν η απόκριση είναι διατάξιμη. Εφαρμόζεται μόνο στην περίπτωση όπου για τις πειραματικές μονάδες οι παρατηρήσεις λαμβάνονται στις ίδιες χρονικές στιγμές. Σε κάθε τέτοια στιγμή υιοθετείται το μοντέλο των αναλογικών λόγων πιθανότητας (*odds*) για την περιθώρια κατανομή της απόκρισης ενώ οι παράμετροι κάθε τέτοιου μοντέλου εκτιμώνται από την αντίστοιχη συνάρτηση πιθανοφάνειας. Σημειώνουμε ότι αυτές οι εκτιμήσεις αποδίδονται χωρίς να υποθέσουμε κάποια συγκεκριμένη δομή στη συσχέτιση των παρατηρήσεων.

2.4.6.2 Μεθοδολογία

Έστω Y_{it} η απόκριση για την *i*-πειραματική μονάδα στην *t*-χρονική στιγμή (ή μέτρηση) με $i = 1, 2, ..., N$ και $t = 1, 2, ..., T$. Επίσης, ας συμβολίσουμε με *K* τον αριθμό των επιπέδων της διατάξιμης απόκρισης. Θεωρούμε τη μεταβλητή

$$Y^*_{itk} = \begin{cases} 1, & \text{αν } Y_{it} = k \\ 0, & \text{διαφορετικά} \end{cases} \quad \text{, όπου } k = 1, 2, ..., K .$$

Αντί, λοιπόν, για τις τιμές Y_{it} μπορούμε να θεωρήσουμε τα αντίστοιχα διανύσματα των παραπάνω δείκτριων μεταβλητών $\mathbf{Y}^*_{it} = (Y^*_{it1}, Y^*_{it2}, ..., Y^*_{itK})'$.

Επιπλέον, σε κάθε χρονική στιγμή *t* και για κάθε μονάδα *i* καταγράφεται ένα *q*-διάστατο διάνυσμα συμμεταβλητών $\mathbf{x}_{it} = (\mathbf{x}_{it1}, \mathbf{x}_{it2}, ..., \mathbf{x}_{itq})'$. Στη συνέχεια θα το συμβολίζουμε απλώς με $\mathbf{x}$ προς χάρην απλότητας. Τέλος, ορίζουμε $z_{tk}(\mathbf{x}) = P(Y^*_{itk} = 1)$ και $g_{tk}(\mathbf{x}) = \sum_{\mathbf{l}=1}^{k} z_{t\mathbf{l}}(\mathbf{x})$ για κάθε *t* και *k*. Δηλαδή, η ποσότητα $g_{tk}(\mathbf{x})$ είναι η αθροιστική πιθανότητα $P(Y_{it} \le k)$ για κάθε *i*. Σύμφωνα με το μοντέλο των αναλογικών *odds*, θα είναι

$$\log \frac{g_{tk}(\mathbf{x})}{1 - g_{tk}(\mathbf{x})} = a_{tk} + \mathbf{x}'\boldsymbol{\beta}_t \ , \qquad (2.9)$$

όπου $\boldsymbol{\beta}_t$ είναι το *q*-διάστατο διάνυσμα των άγνωστων παραμέτρων, που μπορεί να εξαρτάται από την χρονική στιγμή *t*. Καθώς, λείπει ο δείκτης *k*, η επίδραση κάθε συμμεταβλητής στο λογάριθμο του *odds* του ενδεχομένου $Y_{it} \le k$ δεν εξαρτάται από την κατηγορία (επίπεδο) της απόκρισης.

Στην περίπτωση που υπάρχουν ελλιπή δεδομένα ορίζουμε

$$d_{it} = \begin{cases} 1, & \text{όταν } \mathbf{x}_{it} \text{ και } Y_{it} \text{ παρατηρούνται} \\ 0, & \text{διαφορετικά} . \end{cases}$$

Οι υποθέσεις που ακολουθούν αυτές τις δείκτριες είναι:

1) Για κάθε *i* και *t*, οι μεταβλητές d_{it} μπορεί να εξαρτώνται από το διάνυσμα $\mathbf{x}_{it}$.

2) Είναι ανεξάρτητες από τις $Y_{it} \mid \mathbf{x}_{it}$.

3) Δοθέντος του $\mathbf{x}_{it}$, θεωρούνται ανεξάρτητες των a_{tk} και $\boldsymbol{\beta}_t$.

4) Για κάθε $t = 1, 2, ..., T$, τα διανύσματα $(Y_{it}^*, \mathbf{x}_{it}, d_{it})$ θεωρούνται ανεξάρτητα και ακολουθούν την ίδια κατανομή για όλες τις i μονάδες, $i = 1, 2, ..., N$.

Η συγκεκριμένη προσέγγιση θεωρεί ότι τα ελλιπή δεδομένα προέρχονται με βάση τον πρώτο μηχανισμό (*MCAR*) και επομένως αγνοούνται. Έτσι, οι παράμετροι $\boldsymbol{\beta}_t$ και $\boldsymbol{\alpha}_t = (a_{t1}, a_{t2}, ..., a_{tK})'$ προκύπτουν από τη μεγιστοποίηση του λογαρίθμου της πιθανοφάνειας στην *t*-χρονική στιγμή. Η μεγιστοποίηση είναι ισοδύναμη με αυτή της ποσότητας

$$\sum_{i=1}^{N} d_{it} \sum_{k=1}^{K+1} Y_{itk} \left\{ \log[g_{tk}(\mathbf{x}) - g_{t,k-1}(\mathbf{x})] \right\} ,$$

όπου $g_{t0} = 0, g_{t,K+1} = 1$ και το g_{tk} ικανοποιεί το μοντέλο (2.9). Το διάνυσμα $(\hat{\boldsymbol{\alpha}}_t', \hat{\boldsymbol{\beta}}_t')'$ ασυμπτωτικά ακολουθεί κανονική κατανομή με μέση τιμή $(\boldsymbol{\alpha}_t', \boldsymbol{\beta}_t')'$ και πίνακα συνδιακύμανσης του οποίου η εκτίμηση είναι συνεπής και δίνεται από τον τύπο (Α.2) στη σελίδα 636 της εργασίας των Stram et al., (1988).

2.4.6.3 Συμπερασματολογία

Οι υποθέσεις που αφορούν τις παραμέτρους για κάθε συγκεκριμένη χρονική στιγμή και είναι της μορφής

$$H_0 : \mathbf{C}_t \boldsymbol{\beta}_t = \mathbf{0}_c ,$$

όπου $\mathbf{C}_t$ είναι ένας $c \times q$ πίνακας σταθερών $(c \le q)$, μπορούν να ελεγχθούν με το στατιστικό τεστ του Wald

$$W_t = (\mathbf{C}_t \hat{\boldsymbol{\beta}}_t)' (\mathbf{C}_t \hat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_t) \mathbf{C}_t')^{-1} (\mathbf{C}_t \hat{\boldsymbol{\beta}}_t) .$$

Υπό την H_0, ασυμπτωτικά ακολουθεί την X^2 κατανομή με c βαθμούς ελευθερίας. Ανάλογους ελέγχους μπορούμε να κάνουμε και για τις παραμέρους κάθε συμμεταβλητής κατά μήκος του χρόνου.

Επιπρόσθετα, συνδυάζοντας τις εκτιμήσεις των παραμέτρων μιας συμμεταβλητής για κάθε χρονική στιγμή παίρνουμε *pooled* εκτιμήσεις. Έτσι, για την *m*-μεταβλητή, $m = 1, 2, ..., q$ θα έχουμε

$$\hat{b}_m = \sum_{t=1}^{T} w_t \hat{b}_{tm} ,$$

όπου $\mathbf{w} = (w_1, w_2, ..., w_T)'$ οποιοδήποτε διάνυσμα με βάρη που αθροίζουν στη μονάδα. Ειδικότερα, όταν $\mathbf{w} = (\mathbf{1}_T' \hat{\mathrm{V}}\mathrm{ar}(\hat{b}_m)\mathbf{1}_T)^{-1}(\hat{\mathrm{V}}\mathrm{ar}(\hat{b}_m))^{-1}\mathbf{1}_T$, όπου $\mathbf{1}_T$ το $T \times 1$ διάνυσμα (1,1,…,1)', τότε ο εκτιμητής $\hat{b}_m$ που προκύπτει έχει τη μικρότερη ασυμπτωτική διακύμανση μεταξύ των γραμμικών εκτιμητών.

2.5 Παράδειγμα

Το παράδειγμα που ακολουθεί αναφέρεται σε μια διαχρονική έρευνα (***longitudinal study***), που διεξήχθη από την Biometric Society και αναφέρεται στην σύγκριση ενός νέου φαρμάκου σε σχέση με κάποιο ήδη καθιερωμένο για την αντιμετώπιση της διανοητικής κατάθλιψης (Koch et al., 1977). Οι 340 ασθενείς κατατάσσονται σε δύο κατηγορίες ανάλογα με το αν η αρχική διάγνωση έδειχνε "ήπια" ή "σοβαρή" κατάσταση της πάθησης. Σε κάθε κατηγορία, στα άτομα χορηγείται με τυχαίο τρόπο η νέα ή η συνήθης θεραπεία. Τα άτομα παρακολουθούνται για 3 εβδομάδες και κάθε φορά καταγράφεται αν η κατάστασή τους χαρακτηρίζεται φυσιολογική (*normal*) ή όχι (***abnormal***). Ενδεικτικά, έχουμε τον ακόλουθο πίνακα:

Διάγνωση	Θεραπεία	Καταγραφή της πάθησης σε τρεις χρονικές στιγμές							
		NNN	*NNA*	*NAN*	*NAA*	*ANN*	*ANA*	*AAN*	*AAA*
Ήπια	Συνήθης	16	13	9	3	14	4	15	6
	Καινούρια	31	0	6	0	22	2	9	0
Σοβαρή	Συνήθης	2	2	8	9	9	15	27	28
	Καινούρια	7	2	5	2	31	5	32	6

N: ***Normal*** *A*: ***Abnorma***

Εν συνεχεία, ορίζουμε τις ακόλουθες μεταβλητές. ‘Εστω, X_1 η αρχική διάγνωση με $X_1 = 1$ για την σοβαρή και $X_1 = 0$ για την ήπια κατάσταση. Επίσης, ας συμβολίσουμε με X_2 το φάρμακο που λαμβάνει ο ασθενής με $X_2 = 1$ για το καινούριο και $X_2 = 0$ για το σύνηθες. Τέλος, συμβολίζουμε με X_3 τις χρονικές στιγμές της μέτρησης. Στην συγκεκριμένη μεταβλητή μπορούμε να αναθέσουμε αυθαίρετα τιμές (*scores*). Στο παράδειγμα, χρησιμοποιούμε τις τιμές {0, 1, 2}, δηλαδή τους λογαρίθμους με βάση 2 των αριθμών των εβδομάδων κατά τις οποίες έγιναν οι μετρήσεις (1, 2 και 4). Η μεταβλητή απόκρισης είναι η Y_t, η οποία παίρνει τιμές 1 (*normal*) και 0 (*abnormal*) στην *t* χρονική στιγμή.

Επομένως, για το *i*-άτομο με $i = 1,2,\ldots,340$ ένα ρεαλιστικό μοντέλο που επιτρέπει την αλληλεπίδραση του φαρμάκου με το χρόνο είναι το ακόλουθο :

$$\log\text{it}[P(Y_t = 1)] = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_2 X_3 \quad .$$

Στην παράγραφο 2.4.2 είδαμε τα βήματα της μεθόδου των GEE και κατ΄αναλογία εδώ έχουμε :

- $g(m_{it}) = \log(\frac{m_{it}}{1-m_{it}}) = \mathbf{X}_{it}^{'}\boldsymbol{\beta}$, όπου ο δείκτης *i* αναφέρεται στο υποκείμενο και ο δείκτης *j* στην χρονική στιγμή.
- $\text{Var}(\text{Y}_{it}) = u(m_{it})f$ με $u(m_{it}) = m_{it}(1-m_{it})$ και επομένως f =1, καθώς η μεταβλητή απόκρισης είναι δίτιμη.
- Ως πίνακα συσχέτισης μπορούμε να χρησιμοποιήσουμε την ανεξάρτητη δομή για κάθε υποκείμενο. Ο πίνακας θα είναι διάστασης 3×3, αφού για κάθε άτομο υπάρχουν 3 μετρήσεις, και είναι της μορφής

$$\mathbf{R}_i(\mathbf{a}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Υπενθυμίζουμε ότι ο "χρησιμοποιούμενος" πίνακας συνδιακύμανσης δίνεται από τον τύπο

$$\mathbf{V_i}(\mathbf{a}) = j\ \mathbf{A}_i^{1/2}\mathbf{R}_i(\mathbf{a})\mathbf{A}_i^{1/2}.$$

όπου ο πίνακας $\mathbf{A}_i$ θα είναι διαγώνιος διάστασης 3×3, όπου το *t*-διαγώνιο στοιχείο θα είναι η τιμή $u(m_{it}) = m_{it}(1-m_{it})$. Τότε οι *GEE* θα έχουν τη μορφή

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{340} \left(\frac{J\boldsymbol{\mu}_i}{J\boldsymbol{\beta}} \right)' [\mathbf{V}_i(\hat{\mathbf{a}})]^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}_4 ,$$

όπου $\boldsymbol{\beta} = (b_1, b_2, b_3, b_4)'$, $\boldsymbol{\mu}_i = (m_{i1}, m_{i2}, m_{i3})'$ για κάθε $i = 1, 2, ..., 340$, $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3})'$ και $\mathbf{0}_4 = (0, 0, 0, 0)'$.

Τα αποτελέσματα που παίρνουμε (Agresti, 2002) είναι τα εξής :

GEE Parameter Estimates
Empirical Std Error Estimates

Parameter	*Estimate*	*Std Error (SE)*
Intercept (α)	-0.0280	0.1742
β_1	-1.3139	0.1460
β_2	-0.0596	0.2285
β_3	0.4824	0.1199
β_4	1.0174	0.1877

Από τα παραπάνω αποτελέσματα εξάγουμε τα ακόλουθα αποτελέσματα:

- Για $X_2 = 0$, δηλαδή ο ασθενής λαμβάνει το σύνηθες φάρμακο, η εκτίμηση της επίδρασης του χρόνου είναι $\hat{b}_3 \approx 0.48$ ενώ για το νέο φάρμακο, όπου $X_2 = 1$, είναι $\hat{b}_3 + \hat{b}_4 \approx 1.49$. Στη δεύτερη περίπτωση, η κλίση είναι μεγαλύτερη από την πρώτη (εξάλλου η εκτίμηση του συντελεστή της αλληλεπίδρασης b_4 είναι θετική) και επομένως, υπάρχει σαφής ένδειξη ότι με το νέο φάρμακο παρουσιάζεται ταχύτερη βελτίωση της κατάστασης του ασθενή.
- Η εκτίμηση της επίδρασης της αρχικής διάγνωσης είναι $\hat{b}_1 \approx -1.31$. Αυτό σημαίνει ότι αν κρατήσουμε σταθερές τις μεταβλητές X_2 και X_3, τότε ο εκτιμώμενος λόγος πιθανοτήτων (*odds*) να είναι φυσιολογική (*normal*) η κατάσταση ενός ασθενή με σοβαρή αρχική διάγνωση ως προς το να μην είναι φυσιολογική είναι ίσο με $e^{-1.31} \approx 0.27$ φορές τον αντίστοιχο λόγο πιθανοτήτων για ασθενή με ήπια αρχική διάγνωση. Έτσι, καταλήγουμε ότι είναι ευνοϊκότερο στη συνέχεια να παρουσιάσει ομαλή πορεία στην υγεία του ένας ασθενής με ήπια παρά σοβαρή αρχική διάγνωση.

- Η εκτίμηση για την επίδραση του φαρμάκου είναι $\hat{b}_2 \approx -0.06$ με τυπικό σφάλμα $\text{SE} \approx 0.23$. Διαπιστώνεται, λοιπόν, ότι μετά την πρώτη εβδομάδα, όπου $X_3 = 0$, δεν υπάρχει αξιοσημείωτη διαφορά ανάμεσα στις δύο αγωγές. Στις επόμενες χρονικές στιγμές όπου γίνονται οι μετρήσεις η εικόνα αυτή διαφοροποιείται. Εδικότερα, μετά την δεύτερη εβδομάδα, όπου $X_3 = 1$, ο εκτιμώμενος λόγος πιθανοτήτων να είναι φυσιολογική η κατάσταση ασθενή που λαμβάνει το νέο φάρμακο ως προς το να μην είναι φυσιολογική είναι $e^{\hat{b}_2+\hat{b}_4} = e^{-0.06+1.01} = e^{0.95} \approx 2.59$ φορές τον αντίστοιχο λόγο πιθανοτήτων για ασθενή που λαμβάνει το σύνηθες φάρμακο. Η διαφορά μεταξύ των δύο αυτών ποσοστών μεγαλώνει μετά το πέρας της τέταρτης εβδομάδας, όπου $X_3 = 2$. Στην συγκεκριμένη περίπτωση, ο λόγος των δύο ποσοστών θα είναι $e^{\hat{b}_2+2\hat{b}_4} = e^{-0.06+2*1.01} = e^{1.96} \approx 7.10$. Το γεγονός αυτό επιβεβαιώνει και την αρχική ένδειξη που είχαμε για την επίδραση του χρόνου.

2.6 Επέκταση της ML μεθόδου

Στην ενότητα 2.3 αναφερθήκαμε στην *ML* μέθοδο και τις δυσκολίες που αντιμετωπίζουμε σε γενικότερο πλαίσιο με την εφαρμογή της. Υπάρχει, όμως, μία διαφορετική προσέγγιση αυτής και η οποία θεωρεί το περιθώριο μοντέλο ως ένα σύνολο περιορισμών και χρησιμοποιεί μεθόδους μεγιστοποίησης της πιθανοφάνειας βασιζόμενες σε γινόμενο πολυωνυμικών (Haber, 1985; Lang and Agresti, 1994; Lang, 2004). Η μέθοδος συνδυάζει την επαναληπτική χρήση των μη-ορισμένων πολλαπλασιαστών Lagrange (***undetermined Lagrangian multipliers***) με την μέθοδο Newton-Raphson. Ειδικότερα, καθορίζουμε τις ***Lagrangian*** εξισώσεις πιθανοφάνειας της μορφής $h(\boldsymbol{\theta})=0$, όπου $\boldsymbol{\theta}$ είναι το διάνυσμα που περιέχει το σύνολο των πολυωνυμικών πιθανοτήτων και τους πολλαπλασιαστές Lagrange και για τη λύση αυτών επιστρατεύουμε την μέθοδο Newton-Raphson:

$$q^{(t+1)} = q^{(t)} - \left(\frac{\partial h(q^{(t)})}{\partial q}\right)^{-1} h(q^{(t)}).$$

Παρόλ'αυτά, οι υπολογισμοί είναι χρονικά ασύμφοροι κυρίως όταν υπάρχουν αρκετές παρατηρήσεις στις ομάδες (***clusters***) ή ακόμα όταν υπάρχουν πολλές ερμηνευτικές

μεταβλητές και ειδικότερα αν κάποιες είναι συνεχείς. Η δυσκολία έγκειται περισσότερο στο μέρος που αφορά την μέθοδο Newton-Raphson, αφού θα χρειαστεί για τη λύση των εξισώσεων να αντιστραφεί ένας πίνακας με διάσταση μεγαλύτερη από τον αριθμό των κελιών του πίνακα συνάφειας των δεδομένων. Για να ξεπεραστεί το εν λόγω πρόβλημα, εναλλακτικά χρησιμοποιείται μια ασυμπτωτική προσέγγιση του πίνακα που έχει πιο απλή μορφή και απαιτείται η αντιστροφή μόνο ενός διαγώνιου και ενός συμμετρικού θετικά ορισμένου πίνακα (Lang, 1996a; Lang and Agresti, 1994).

2.7 Καλή προσαρμογή (Goodness-of-fit)

Έχουμε επισημάνει μέχρι τώρα ότι τα περιθώρια μοντέλα αποτελούν την πλέον διαδεδομένη επιλογή για την ανάλυση επαναλαμβανόμενων κατηγορικών δεδομένων. Όπως και με τα κοινά μοντέλα λογιστικής παλινδρόμησης, είναι ιδιαίτερα σημαντικό να ελέγξουμε την καλή προσαρμογή (*goodness-of-fit*) αυτών. Εντούτοις, δεν έχουν αναπτυχθεί αρκετές μέθοδοι για την συγκεκριμένη αξιολόγηση.

Πρόσφατα, έχουν προταθεί δύο μέθοδοι από τους Barnhart and Williamson, (1998) και Horton et al., (1999) αντίστοιχα για συσχετισμένα δίτιμα δεδομένα. Τα δύο αυτά τεστ μπορούν να θεωρηθούν αντίστοιχα ως επέκταση των ελέγχων καλής προσαρμογής των Tsiatis, (1980) και Hosmer and Lemeshow, (1980) για τη συνηθισμένη λογιστική παλινδρόμηση (όπου τα δίτιμα δεδομένα είναι ανεξάρτητα) στην περιθώρια λογιστική παλινδρόμηση (όπου τα δίτιμα δεδομένα είναι συσχετισμένα). Ειδικότερα, τα στατιστικά συμπεράσματα βασίζονται στην X^2 κατανομή.

Ας θεωρήσουμε το περιθώριο μοντέλο λογιστικής παλινδρόμησης

$$\log \mathrm{it}(p_{it}) = \mathbf{x}_{it}^{'}\boldsymbol{\beta}$$

με $p_{it} = E(Y_{it} \mid \mathbf{x}_{it})$ και $\mathrm{var}(Y_{it} \mid \mathbf{x}_{it}) = p_{it}(1-p_{it})$, όπου ο δείκτης i $(i = 1, 2, ..., N)$ αναφέρεται στο υποκείμενο και ο δείκτης t $(t = 1, 2, ..., T)$ στην χρονική περίσταση κατά την οποία λαμβάνεται η παρατήρηση. Η εκτίμηση του διανύσματος $\boldsymbol{\beta}$ των παραμέτρων του μοντέλου δίνεται από τη λύση των εξισώσεων (2.6) της ενότητας 2.4.3. Το X^2 τεστ του Pearson δίνεται από τον τύπο

$$G=\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{(Y_{it}-\hat{p}_{it})^2}{\hat{p}_{it}(1-\hat{p}_{it})}=NT+(1-2\hat{\boldsymbol{\pi}})'\hat{\mathbf{A}}^{-1}\hat{\mathbf{e}}\approx NT+(1-2\hat{\boldsymbol{\pi}})'\hat{\mathbf{A}}^{-1}(1-\mathbf{H})\mathbf{e}\ ,$$

όπου $\hat{p}_{it}=\text{logit}^{-1}(\mathbf{x}_{it}'\hat{\boldsymbol{\beta}})$ και $\hat{\boldsymbol{\pi}}=(\hat{\boldsymbol{\pi}}_{\mathbf{1}},\hat{\boldsymbol{\pi}}_{\mathbf{2}},...,\hat{\boldsymbol{\pi}}_{\mathbf{N}})'$ με $\hat{\boldsymbol{\pi}}_i=(p_{i1},p_{i2},...,p_{iT})'$. Επίσης, ο πίνακας $\mathbf{A}$ είναι αυτός που είδαμε στην ενότητα 2.4.3 ενώ με $\mathbf{e}$ συμβολίζουμε τα υπόλοιπα $\mathbf{Y}-\boldsymbol{\pi}$, όπου $\mathbf{Y}=(\mathbf{Y}_1',\mathbf{Y}_2',...,\mathbf{Y}_N')'$. Ο πίνακας $\mathbf{H}$ αντιστοιχεί στο γινόμενο $\mathbf{AX(X'AV^{-1}AX)^{-1}X'AV^{-1}}$, όπου $\mathbf{X}$ είναι ο πίνακας $(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_N)'$ ενώ ο πίνακας $\mathbf{V}$ είναι ο ίδιος με αυτόν που συναντήσαμε στην ενότητα 2.4.3. Αντιμετωπίζοντας την ποσότητα $(1-2\hat{\boldsymbol{\pi}})'\hat{\mathbf{A}}^{-1}$ ως σταθερή, το παραπάνω τεστ έχει κατά προσέγγιση μέση τιμή $\hat{E}(G)=NT$ και διακύμανση

$$\hat{\text{var}}(G)=(1-2\hat{\boldsymbol{\pi}})'\hat{\mathbf{A}}^{-1}(\mathbf{I}-\hat{\mathbf{H}})\hat{\text{cov}}(\mathbf{Y})(\mathbf{I}-\hat{\mathbf{H}}')\hat{\mathbf{A}}^{-1}(1-2\hat{\boldsymbol{\pi}})\,.$$

Εναλλακτικά, έχει προταθεί το επόμενο στατιστικό τεστ βασίζεται σε ένα μη σταθμισμένο άθροισμα των τετραγώνων των υπολοίπων (Hosmer et al., 1997). Ορίζουμε, λοιπόν, ως

$$U=\sum_{i=1}^{N}\sum_{t=1}^{T}(Y_{it}-\hat{p}_{it})^2=\hat{\boldsymbol{\pi}}'(1-\hat{\boldsymbol{\pi}}')+(1-2\hat{\boldsymbol{\pi}})'\hat{\mathbf{e}}\approx\hat{\boldsymbol{\pi}}'(1-\hat{\boldsymbol{\pi}}')+(1-2\hat{\boldsymbol{\pi}})'(1-\mathbf{H})\mathbf{e}$$

με ασυμπτωτική μέση τιμή $\hat{E}(U)=\hat{\boldsymbol{\pi}}'(1-\hat{\boldsymbol{\pi}}')$ και διακύμανση

$$\hat{\text{var}}(U)=(1-2\hat{\boldsymbol{\pi}})'(\mathbf{I}-\hat{\mathbf{H}})\hat{\text{cov}}(\mathbf{Y})(\mathbf{I}-\hat{\mathbf{H}}')(1-2\hat{\boldsymbol{\pi}})\,.$$

Και τα δύο παραπάνω τεστ ακολουθούν ασυμπτωτικά την κανονική κατανομή, αν και για το πρώτο έχουν διατυπωθεί διαφωνίες (Osius and Rojek, 1992). Περισσότερες πληροφορίες για το θεωρητικό υπόβαθρο αυτών αλλά και διαφορετικές προσεγγίσεις μπορούν οι ενδιαφερόμενοι να αναζητήσουν την εργασία του Pan, (2002).

Κεφάλαιο 3ο
Μοντέλα τυχαίων επιδράσεων
(Random Effects Models)

3.1 Εισαγωγή

Στο προηγούμενο κεφάλαιο ασχοληθήκαμε με την ανάπτυξη και εφαρμογή των περιθώριων μοντέλων για την ανάλυση επαναλαμβανόμενων μετρήσεων. Υπάρχει, όμως, και μια διαφορετική προσέγγιση του συγκεκριμένου θέματος, με τα μοντέλα τυχαίων επιδράσεων (***random effects models***). Υπενθυμίζουμε ότι στα συνήθη γραμμικά μοντέλα οι παράμετροι που εισάγονται αποκαλούνται σταθερές επιδράσεις (***fixed effects***) και εξαντλούν όλα τα επίπεδα ή κατηγορίες του παράγοντα που μας ενδιαφέρει (π.χ. φύλο, ομάδα θεραπείας). Αντιθέτως, στα μοντέλα τυχαίων επιδράσεων εισάγουμε παραμέτρους που απευθύνονται μόνο σε ένα δείγμα των δυνατών επιπέδων ή κατηγοριών.

Η ανάπτυξη των μοντέλων τυχαίων επιδράσεων όσον αφορά κατηγορικές αποκρίσεις είναι μεταγενέστερη των συνεχών αποκρίσεων, όπου υποθέτουμε κανονικότητα. Στην επόμενη ενότητα αναφερόμαστε σε γενικά στοιχεία της κατηγορίας αυτής των μοντέλων και στην 3.3 γίνεται λόγος για τις μεθόδους πρόβλεψης και προσαρμογής. Η ενότητα 3.4 αφορά στην συμπερασματολογία και η επόμενη στην πρόβλεψη των τυχαίων επιδράσεων. Στις δύο τελευταίες ενότητες δίνεται ένα παράδειγμα και η σύνδεση των μοντέλων τυχαίων επιδράσεων με τα περιθώρια αντίστοιχα.

3.2 Γενικά στοιχεία

Έχει αναφερθεί νωρίτερα ότι οι παρατηρήσεις που λαμβάνονται κατ'επανάληψη από κάποιο άτομο εμφανίζουν συσχέτιση, κατά κύριο λόγο θετική. Γνωρίζουμε ότι για συνεχείς μεταβλητές, η πολυδιάστατη κανονική κατανομή παρέχει αξιοσημείωτη ευχέρεια για την περιγραφή της εξάρτησης των δεδομένων. Αλλά στην περίπτωση των κατηγορικών μεταβλητών, δεν υπάρχει ανάλογη κατανομή και έτσι ο καθορισμός ενός επαρκούς μοντέλου είναι πιο σύνθετος. Ένας τρόπος για την αντιμετώπιση του προβλήματος αυτού είναι η εισαγωγή "***cluster-level***" όρων στο μοντέλο. Οι συγκεκριμένοι όροι παίρνουν την ίδια τιμή για κάθε παρατήρηση μέσα σε κάποια ομάδα (***cluste***r) αλλά διαφορετικές τιμές για καθεμία από τις ομάδες. Σε γενικές γραμμές οι όροι αυτοί δεν παρατηρούνται (***unobserved***) και επειδή αντιστοιχούν σε ομάδες που επιλέγονται με τυχαίο τρόπο δικαιολογείται να αποκαλούνται όροι τυχαίων επιδράσεων. Με την εισαγωγή τους σε ένα γενικευμένο γραμμικό μοντέλο (***GLM***) προκύπτει μια νέα κατηγορία μοντέλων, τα γενικευμένα γραμμικά μεικτά μοντέλα (***generalized linear mixed models-GLMM***), όπου συνυπάρχουν όροι σταθερών και τυχαίων επιδράσεων.

Το μέρος των τυχαίων επιδράσεων, που αποτελεί και το κύριο χαρακτηριστικό της κατηγορίας αυτής των μοντέλων, βοηθά ώστε να περιγράψουμε την επιπλέον μεταβλητότητα που εμφανίζεται σε αρκετές περιπτώσεις και οφείλεται σε παράγοντες όπως η ετερογένεια μεταξύ των ατόμων, συμμεταβλητές που δεν έχουν ληφθεί υπόψη και το τυχαίο σφάλμα μέτρησης (***random measurement error***) των επεξηγηματικών μεταβλητών (Follmann and Lambert, 1989). Με πιο απλά λόγια, μας επιτρέπει να αναλύσουμε την επιρροή των υποκειμένων στις επαναλαμβανόμενες μετρήσεις τους. Επιπλέον, παρέχει ένα μηχανισμό για την ερμηνεία της υπερμεταβλητότητας (***overdispersion***), δηλαδή της παρουσίας μεγαλύτερης μεταβλητότητας στα δεδομένα από αυτή που προβλέπει το μοντέλο (Breslow and Clayton, 1993). Σε αντίθεση με τις σταθερές επιδράσεις, επιτρέπει στο φαινόμενο της συσχέτισης να ενσωματώνεται στις εκτιμήσεις των παραμέτρων, στα τυπικά σφάλματα αλλά και τους ελέγχους υποθέσεων.

Για να γίνει αντιληπτό, λόγου χάρη το θέμα της ετερογένειας, δίνεται το ακόλουθο παράδειγμα. Σε άρθρο του Randall, (1989) παρουσιάζονται τα αποτελέσματα ενός παραγοντικού πειράματος, όπου καθένας από 9 κριτές δοκιμάζει 8 μπουκάλια κρασιού. Τα

μπουκάλια χαρακτηρίζονται από τους παράγοντες επαφή και θερμοκρασία. Ο πρώτος, με επίπεδα "ναι" και "όχι", καθορίζει αν είχε λάβει χώρα επαφή του χυμού με τον φλοιό όταν τα σταφύλια συνθλίβονταν. Αντίστοιχα, η θερμοκρασία διακρίνεται σε "υψηλή" και "χαμηλή". Την δοκιμασία των κρασιών από τους κριτές ακολουθούν δηλώσεις όσον αφορά την πικρότητά του σε μια κλίμακα 5 βαθμών που κυμαίνεται από ελάχιστη έως υπερβολική. Καθώς ο κάθε κριτής δοκιμάζει και τα 8 μπουκάλια, γίνεται αντιληπτό ότι οι παρατηρήσεις δεν μπορούν να είναι ανεξάρτητες. Επιπλέον, η απόφαση περί της πικρότητας για το οποιοδήποτε κρασί κρίνεται υποκειμενική, εφόσον κάθε κριτής διαθέτει συγκεκριμένη ευαισθησία στο θέμα αυτό. Άρα, στο μοντέλο για την περιγραφή του εν λόγω πειράματος πρέπει να ληφθεί υπόψη η μεταβαλλόμενη ευαισθησία στο σύνολο των κριτών.

Υπάρχουν αρκετά χαρακτηριστικά των μοντέλων τυχαίων επιδράσεων που τα καθιστούν ιδιαίτερα χρήσιμα σε διαχρονικές έρευνες. Αρχικά, δεν είναι απαραίτητο τα άτομα που συμμετέχουν να έχουν τον ίδιο αριθμό μετρήσεων. ‘Ετσι, στην ανάλυση μπορούν να συμπεριλαμβάνονται και περιπτώσεις ατόμων με ελλιπή δεδομένα. Το γεγονός αυτό αποτελεί σημαντικό πλεονέκτημα έναντι μεθόδων που απαιτούν πλήρη δεδομένα για δύο κυρίως λόγους. Πρώτον, διότι έχοντας συμπεριλάβει στη μελέτη όλα τα δεδομένα, η ανάλυση έχει αυξημένη στατιστική ισχύ και δεύτερον, διότι αν ληφθούν υπόψη μόνο τα πλήρη δεδομένα, ενδεχομένως να οδηγηθούμε σε μεροληπτικά αποτελέσματα καθώς το δείγμα δεν θα είναι αντιπροσωπευτικό του πληθυσμού.

Επίσης, εφόσον ο χρόνος αντιμετωπίζεται συχνά ως συνεχή μεταβλητή, δεν είναι υποχρεωτικό να λαμβάνονται μετρήσεις από κάθε άτομο στις ίδιες χρονικές στιγμές με τα υπόλοιπα. Η αντιμετώπιση αυτή είναι χρήσιμη σε διαχρονικές μελέτες όπου, για παράδειγμα, οι ***follow-up*** χρόνοι δεν είναι ομοιόμορφοι κατά μήκος των ατόμων. Ακόμα, μπορούν στο μοντέλο να εισαχθούν συμμεταβλητές σταθερές αλλά και μεταβαλλόμενες στο χρόνο και συνεπώς να μελετήσουμε τυχόν αλλαγές στην εξαρτημένη μεταβλητή που οφείλονται σε σταθερά χαρακτηριστικά των ατόμων (π.χ. το φύλο) ή σε χαρακτηριστικά που διαφοροποιούνται χρονικά. Τέλος, σε αντίθεση με την μέθοδο των ***GEE***, όπου οι εκτιμήσεις έχουν να κάνουν με τις κατά μέσο όρο (***average***) μεταβολές σε έναν πληθυσμό, τα ***GLMM*** μας δίνουν εκτιμήσεις για κάθε άτομο. Με τον τρόπο αυτό, μπορούν να γίνουν αντιληπτές συμπεριφορές που χρονικά τείνουν να αποκλίνουν από τη μέση τάση. Λόγου χάρη, ενώ στο

πέρασμα του χρόνου συνήθως παρατηρείται αύξηση στο κάπνισμα, πιθανόν να υπάρχει ένα ποσοστό άτομων που δεν ακολουθεί αυτή τη συμπεριφορά.

3.3 Προσαρμογή και πρόβλεψη

Η προσαρμογή των υπό μελέτη μοντέλων είναι πιο σύνθετη σε σχέση με τα περιθώρια. Η δυσκολία έγκειται στο γεγονός ότι η συνάρτηση πιθανοφάνειας αυτών περιέχει ένα ολοκλήρωμα που ο υπολογισμός του απαιτεί τη χρήση αριθμητικών μεθόδων. Στην περίπτωση όπου στο μοντέλο έχουν εισαχθεί πολυμεταβλητές τυχαίες επιδράσεις γίνεται αντιληπτό ότι το υπολογιστικό μέρος είναι επίπονο.

3.3.1 Εισαγωγή στα GLMM

Έστω Y_{it} η t-οστή παρατήρηση του i-υποκειμένου με $i = 1, 2, ..., N$ και t =1,2,3,…,T. Επίσης, ας συμβολίσουμε με $\mathbf{x}_{it}$ το διάνυσμα στήλη με τις τιμές των ερμηνευτικών μεταβλητών και με $\boldsymbol{\beta}$ το αντίστοιχο διάνυσμα των σταθερών επιδράσεων του μοντέλου. Ανάλογα, ορίζουμε ως $\mathbf{u}_i$ το διάνυσμα των τυχαίων επιδράσεων για την i-ομάδα (δηλαδή το i-υποκείμενο), το οποίο παραμένει σταθερό για όλες τις παρατηρήσεις εντός αυτής. Με $\mathbf{z}_{it}$ δηλώνουμε το διάνυσμα στήλη με τις ερμηνευτικές μεταβλητές που αντιστοιχούν στις τυχαίες επιδράσεις. Δεσμεύοντας ως προς το διάνυσμα $\mathbf{u}_i$, ένα *GLMM* μοντέλο έχει ομοιότητες με ένα κοινό *GLM*. Έστω $m_{it} = E(Y_{it} | \mathbf{u}_i)$ η μέση τιμή της δεσμευμένης κατανομής της μεταβλητής Y_{it} δοθέντος του διανύσματος $\mathbf{u}_i$. Επίσης, ας συμβολίσουμε με $Var(Y_{it} | \mathbf{u}_i) = f_{it} u(m_{it})$ την διακύμανση της εν λόγω κατανομής, όπου τυπικά $f_{it} = f / w_{it}$ με w_{it} να είναι γνωστά βάρη και f άγνωστη παράμετρος διασποράς. Η συνάρτηση $u(\cdot)$ καλείται συνάρτηση διασποράς, κατ'αναλογία με την θεωρία των *GLM*. Γενικά ένα *GLMM* μοντέλο θα είναι της μορφής

$$g(m_{it}) = \mathbf{x}_{it}^{'}\boldsymbol{\beta} + \mathbf{z}_{it}^{'}\mathbf{u}_i \quad , \tag{3.1}$$

όπου $g(\cdot)$ είναι η συνάρτηση σύνδεσης για την οποία έγινε λόγος στο προηγούμενο κεφάλαιο. Το διάνυσμα $\mathbf{u}_i$ θεωρούμε ότι ακολουθεί πολυδιάστατη κανονική κατανομή $N(\mathbf{0}, \boldsymbol{\Sigma})$ με τον πίνακα $\boldsymbol{\Sigma}$ να εξαρτάται από άγνωστες συνιστώσες διασποράς και ενδεχομένως παραμέτρους συσχέτισης.

3.3.2 Επιλογές για το μοντέλο

Τα μοντέλα αυτά για κατηγορικές αποκρίσεις γενικά υιοθετούν τους συνδέσμους ***probit*** ή ***logit*** και η θεωρία τους περιλαμβάνει διάφορες μεθόδους για την ενσωμάτωση και εκτίμηση των τυχαίων επιδράσεων. Στη βιβλιογραφία η πιο δημοφιλής επιλογή αφορά στον σύνδεσμο ***logit*** και για την περιγραφή των δεδομένων μπορούμε να προσαρμόσουμε κάποιο από τα μοντέλα που θεωρήσαμε στην ενότητα 2.2 αρκεί να προσθέσουμε όρους τυχαίων επιδράσεων. Καθώς η πλέον συνήθης επιλογή για την ανάλυση διατάξιμων δεδομένων είναι το μοντέλο (2.2) με την ιδιότητα των αναλογικών λόγων πιθανότητας (***proportional odds***), το οποίο στη συγκεκριμένη περίπτωση θα είναι της μορφής

$$\log it[P(Y_{it} \le k \mid \mathbf{x}_{it}, \mathbf{z}_{it})] = a_k + \boldsymbol{\beta}^{'}\mathbf{x}_{it} + \mathbf{u}_i^{'}\mathbf{z}_{it}.$$

Πολλά από τα μεικτά μοντέλα για την αντιμετώπιση τέτοιων δεδομένων αποτελούν γενικεύσεις αυτού. Στη συνέχεια ακολουθεί μια σύντομη περιγραφή της θεωρίας του εν λόγω μοντέλου και των βασικών ιδιοτήτων του.

Ας υποθέσουμε ότι η μεταβλητή απόκρισης Y έχει K διατάξιμες κατηγορίες. Οι αθροιστικές πιθανότητες για το i-υποκείμενο θα είναι

$$p_{itk} = P(Y_{it} \le k) = \sum_{z=1}^{k} P(Y_{it} = z) \quad , \quad k = 1, 2, ..., K-1 \ .$$

Τότε το αθροιστικό logit ορίζεται ως

$$\log it[P(Y_{it} \le k)] = \log\left[\frac{P(Y_{it} \le k)}{1 - P(Y_{it} \le k)}\right] = \log\left[\frac{P(Y_{it} \le k)}{P(Y_{it} \ge k)}\right] \ , \ k = 1, 2, ..., K-1 \ . \tag{3.2}$$

Παρατηρούμε ότι σε κάθε αθροιστικό *logit* χρησιμοποιούνται όλες οι κατηγορίες της μεταβλητής *Y*. Επίσης, είναι όμοιο με ένα κοινό *logit* για δίτιμη απόκριση αν συνδυάσουμε τις κατηγορίες από 1 μέχρι *z* ως την πρώτη κατηγορία της δίτιμης και τις υπόλοιπες, από *z*+1 μέχρι *K*, ως τη δεύτερη.

Το *proportional odds* μοντέλο είναι

$$\text{logit}[P(Y_{it} \le k)] = a_k + \boldsymbol{\beta}' \mathbf{x}_{it} \text{ με } i = 1, 2, ..., N,\ t = 1, 2, ..., T \text{ και } k = 1, 2, ..., K-1.$$

Σε αυτή την έκφραση, μια θετική τιμή για την παράμετρο b_j του διανύσματος $\boldsymbol{\beta}$ υποδηλώνει αρνητική σχέση ανάμεσα στην μεταβλητή *Y* και την αντίστοιχη συμμεταβλητή x_j. Δηλαδή, όταν η τελευταία έχει μεγάλη τιμή, τότε είναι λιγότερο πιθανό να πάρει και η απόκριση ομοίως μεγάλη τιμή.

Καθώς το διάνυσμα των παραμέτρων δεν φέρει τον δείκτη *k*, η επίδραση κάθε συμμεταβλητής θεωρείται ομοιογενής κατά μήκος των *K*-1 αθροιστικών *logits*. Πιο συγκεκριμένα, το μοντέλο υποδηλώνει ότι οι λόγοι των *odds* για την περιγραφή της επίδρασης των ερμηνευτικών μεταβλητών στην μεταβλητή απόκρισης είναι ίδιοι με όποιον τρόπο και αν μετατρέψουμε την απόκριση από διατάξιμη σε δίτιμη ($\le k$, $> k$).

Γενικά, εφόσον

$$\text{logit}[P(Y \le k \mid \mathbf{x}_1)] - \text{logit}[P(Y \le k \mid \mathbf{x}_2)] = \log\left[\frac{P(Y \le k \mid \mathbf{x}_1)/P(Y > k \mid \mathbf{x}_1)}{P(Y \le k \mid \mathbf{x}_2)/P(Y > k \mid \mathbf{x}_2)}\right] = \boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2)$$

γίνεται κατανοητό ότι ο λογάριθμος του λόγου των *odds* (*log odds ratio*) είναι ανάλογος της απόστασης ανάμεσα στα διανύσματα $\mathbf{x}_1$ και $\mathbf{x}_2$. Με τον τρόπο αυτό δικαιολογείται και η ιδιότητα που έχουμε αποδώσει στο εν λόγω μοντέλο. Στην περίπτωση που από το σύνολο των ερμηνευτικών μεταβλητών κάποια από αυτές, έστω η x_j, μεταβληθεί κατά μία μονάδα και οι υπόλοιπες παραμείνουν σταθερές, ο παραπάνω λόγος θα ισούται με $\exp(b_j)$.

Είναι άξιο να αναφερθεί ότι τα μοντέλα παλινδρόμησης για διατάξιμες αποκρίσεις συχνά υποκινούνται και περιγράφονται χρησιμοποιώντας την ιδέα των "ορίων" (*thresholds*), που εισήχθη από τον Bock, (1975). Με βάση το συγκεκριμένο σκεπτικό, δεν χρειάζεται να αναθέσουμε σκορ (*scores*) στις κατηγορίες της *Y* καθώς υποτίθεται ότι πίσω από την διατάξιμη μεταβλητή απόκρισης που παρατηρούμε υπάρχει μια συνεχής κρυμμένη (*latent*) μεταβλητή (Anderson and Phillips, 1981), πιθανότατα μη παρατηρήσιμη, και από την οποία

προκύπτει η πρώτη μέσω των σημείων αποκοπής (*cutoff points*) $a_1, a_2, ..., a_{K-1}$. Έτσι, αν υποθέσουμε ότι η συνεχής μεταβλητή είναι η *S*, τότε οι κατηγορίες της *Y* αντιστοιχούν σε διαδοχικά διαστήματα της συνεχούς κλίμακας της *S* ενώ τα παραπάνω σημεία θεωρούνται άγνωστα. Δηλαδή, θα είναι

$$Y = \begin{cases} 1, & \text{αν } S \le a_1 \\ 2, & \text{αν } a_1 < S \le a_2 \\ \vdots & \vdots \\ K-1, & \text{αν } a_{K-2} < S \le a_{K-1} \\ K, & \text{αν } S > a_{K-1}. \end{cases}$$

Στη συνέχεια, υποθέτοντας ότι η μεταβλητή *S* ακολουθεί την τυπική λογιστική κατανομή (*standard logistic distribution*) προκύπτει το μοντέλο (3.2), καθώς πλέον

$$P(Y \le k) = P(S \le k-1) = \frac{e^{a_{k-1}}}{1+e^{a_{k-1}}}.$$

Ωστόσο, όπως παρατήρησαν οι McCullagh and Nelder, (1989), η υπόθεση μιας κρυμμένης συνεχούς μεταβλητής δεν είναι αυστήρη αξίωση για την χρησιμοποίηση μοντέλων παλινδρόμησης για διατάξιμες αποκρίσεις.

Η επιλογή του μοντέλου των αναλογικών *odds* δεν είναι πάντοτε λογική και δεν είναι δύσκολο να αναζητήσει κανείς παραδείγματα που παραβιάζουν την αναλογική υπόθεση για την οποία γίνεται λόγος (Peterson and Harrell, 1990). Συνεπώς, το μοντέλο δεν εφαρμόζεται πάντοτε όταν πρόκειται για διατάξιμη απόκριση.

3.3.3 Εκτίμηση

Ο καθορισμός ενός *GLMM* γίνεται σε δύο στάδια. Αρχικά, δοθέντων των τυχαίων επιδράσεων u_i, έχουμε υποθέσει ότι τα δεδομένα προέρχονται από κατανομή που ανήκει στην εκθετική οικογένεια. Εξάλλου, έχει προαναφερθεί ότι δεσμεύοντας ως προς το διάνυσμα **u**, ένα *GLMM* καταλήγει να μοιάζει σε ένα κοινό *GLM*. Στο δεύτερο στάδιο χρειαζόμαστε μια υπόθεση για την κατανομή των u_i. Τυπικά, θεωρούμε ότι είναι ανεξάρτητα και ακολουθούν πολυμεταβλητή κατανομή $N(\mathbf{0}, \mathbf{\Sigma})$. Ας θεωρήσουμε τα διανύσματα

$\mathbf{y} = (Y_{11}, Y_{12}, ..., Y_{1T}, ..., Y_{N1}, Y_{N2}, ..., Y_{NT})'$ και $\mathbf{u} = (u_1, u_2, ..., u_N)'$. Έστω, λοιπόν, $f(\mathbf{y}|\mathbf{u};\boldsymbol{\beta})$ η δεσμευμένη συνάρτηση πιθανότητας του διανύσματος **y** δοθέντος του **u** και $f(\mathbf{u};\boldsymbol{\Sigma})$ η κανονική συνάρτηση πιθανότητας για το διάνυσμα **u**.

Έχουμε αναφέρει νωρίτερα ότι τα u_i είναι μη παρατηρήσιμα. Οπότε για να πάρουμε την συνάρτηση πιθανοφάνειας, υπολογίζουμε το γινόμενο των πολυωνυμικών (*multinomials*) ως συνήθως και εν συνεχεία ολοκληρώνουμε ως προς αυτά. Δοθέντων των δεδομένων, η συνάρτηση πιθανοφάνειας εξαρτάται από τις παραμέτρους σταθερών επιδράσεων (*fixed effects parameters*) αλλά και από τις παραμέτρους της κατανομής των τυχαίων επιδράσεων. Με άλλα λόγια, η *GLMM* συνάρτηση πιθανοφάνειας είναι η συνάρτηση πιθανότητας $f(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma})$ του διανύσματος **y** ως συνάρτηση των **β** και **Σ**. Δηλαδή,

$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma}; \mathbf{y}) = f(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \int f(\mathbf{y} \mid \mathbf{u}; \boldsymbol{\beta}) f(\mathbf{u}; \boldsymbol{\Sigma}) d\mathbf{u} .$$

Στη βιβλιογραφία συχνά αναφέρεται και ως περιθώρια πιθανοφάνεια (*marginal likelihood*). Υπολογίζεται αριθμητικά και ακολούθως μεγιστοποιείται ως προς τις παραμέτρους **β** και **Σ**. Στις επόμενες υποενότητες ακολουθεί περιγραφή των μεθόδων για τον υπολογισμό του παραπάνω ολοκληρώματος και κατά συνέπεια της συνάρτησης πιθανοφάνειας.

3.3.3.1 Μέθοδος των "Gauss-Hermite Quadrature"

Το ολοκλήρωμα που καθορίζει την συνάρτηση πιθανοφάνειας έχει διάσταση που εξαρτάται από τη δομή των τυχαίων επιδράσεων. Αν, για παράδειγμα, θεωρήσουμε το *logistic-normal* μοντέλο

$$\text{logit}[P(Y_{it} = 1 \mid u_i)] = x_{it}'\boldsymbol{\beta} + u_i ,$$

όπου $u_i \sim N(0, s^2)$, τότε το ολοκλήρωμα είναι μονοδιάστατο και για τον προσδιορισμό της πιθανοφάνειας απαιτούνται οι τυπικές μέθοδοι ολοκλήρωσης.

Η Gauss-Hermite *quadrature* μέθοδος χρησιμοποιείται για τον υπολογισμό του ολοκληρώματος ενός γινομένου συνάρτησης $f(\cdot)$ με μια έτερη, η οποία έχει το σχήμα κανονικής κατανομής. Η προσέγγιση αυτού γίνεται με ένα ορισμένο και σταθμισμένο

άθροισμα όρων της συνάρτησης αυτής σε συγκεκριμένα σημεία. Στην περίπτωση των μονοδιάστατων κανονικών τυχαίων επιδράσεων, η προσέγγιση θα είναι της μορφής

$$\int_{-\infty}^{+\infty} f(u)\exp(-u^2) = \sum_{k=1}^{q} c_k f(s_k),$$

όπου $\{c_k\}$ και $\{s_k\}$ είναι τα βάρη και τα *quadrature* σημεία αντίστοιχα.

Στη συνέχεια γίνεται η μεγιστοποίηση της πιθανοφάνειας με κλασικούς αλγορίθμους, όπως αυτός των Newton-Raphson, και έτσι προκύπτουν *ML* εκτιμητές των $\boldsymbol{\beta}$ και $\boldsymbol{\Sigma}$. Η ακρίβεια της προσέγγισης βελτιώνεται όσο αυξάνεται ο αριθμός q των όρων του αθροίσματος.

3.3.3.2 Μέθοδος Monte-Carlo

Ο επαρκής προσδιορισμός της συνάρτησης πιθανοφάνειας γίνεται πιο δύσκολος όταν η διάσταση του ολοκληρώματος είναι μεγαλύτερη του 5. Στην περίπτωση αυτή προτιμώνται οι μέθοδοι Monte-Carlo καθώς υπολογιστικά είναι πιο εύχρηστες. Έχουν αναπτυχθεί αρκετές προσεγγίσεις του δεδομένου προβλήματος (π.χ. McCulloch, 1997) και μεταξύ άλλων σε αυτές περιλαμβάνεται και ο συνδυασμός Monte-Carlo με τον *EM* αλγόριθμο (Dempster, Laird and Rubin, 1977) με την συντομογραφία *MCEM* (Wei and Tanner, 1990). Ο αλγόριθμος *EM* είναι χρήσιμη επαναληπτική μέθοδος για την εύρεση *ML* εκτιμητών, όταν υπάρχουν ελλιπή δεδομένα. Η εφαρμογή τους στα *GLMM* δικαιολογείται αν θεωρήσουμε τις τυχαίες επιδράσεις ως τέτοια.

Σε κάθε κύκλο του, το *E*-βήμα υπολογίζει την συνάρτηση πιθανοφάνειας κάνοντας μια πρόβλεψη για τα δεδομένα που λείπουν και ακολούθως το *M*-βήμα την μεγιστοποιεί παίρνοντας τις αντίστοιχες εκτιμήσεις των παραμέτρων. Ειδικότερα, το *E*-βήμα στην *r*-επανάληψη υπολογίζει τη μέση τιμή

$$E\{\log h(\mathbf{y},\mathbf{u};\boldsymbol{\beta},\boldsymbol{\Sigma}) \mid \mathbf{y};\boldsymbol{\beta}^{(r)},\boldsymbol{\Sigma}^{(r)}\},$$

όπου $h(\mathbf{y},\mathbf{u};\boldsymbol{\beta},\boldsymbol{\Sigma})$ είναι η κοινή κατανομή των πλήρη δεδομένων, δηλαδή $h(\mathbf{y},\mathbf{u};\boldsymbol{\beta},\boldsymbol{\Sigma}) = f(\mathbf{y}\mid\mathbf{u};\boldsymbol{\beta})\,f(\mathbf{u};\boldsymbol{\Sigma})$. Στη θέση των παραμέτρων $\boldsymbol{\beta}$ και $\boldsymbol{\Sigma}$ χρησιμοποιούνται οι εκτιμήσεις $\boldsymbol{\beta}^{(r)}$ και $\boldsymbol{\Sigma}^{(r)}$ του συγκεκριμένου βήματος ενώ στη συνέχεια το *M*-βήμα μεγιστοποιεί το αποτέλεσμα δίνοντας νέες εκτιμήσεις $\boldsymbol{\beta}^{(r+1)}$ και $\boldsymbol{\Sigma}^{(r+1)}$.

Δεν είναι εφικτή η αναλυτική αποτίμηση της παραπάνω μέσης τιμής και έτσι, ο αλγόριθμος *MCEM* την προσεγγίζει χρησιμοποιώντας Monte-Carlo μεθόδους. Πιθανοί τρόποι για να γίνει αυτό περιλαμβάνουν την χρησιμοποίηση ανεξάρτητων προσομοιώσεων από την κατανομή $\mathbf{u}|\mathbf{y}$, για τις τρέχουσες εκτιμήσεις των παραμέτρων και τη χρήση Μαρκοβιανών αλυσίδων Monte-Carlo (*MCMC*).

3.3.3.3 Penalized Quasi-likelihood (PQL) προσέγγιση

Οι μέθοδοι ολοκλήρωσης Gauss-Hermite και Monte-Carlo αποδίδουν συναρτήσεις πιθανοφάνειας των οποίων η μεγιστοποίηση δίνει εκτιμητές που συγκλίνουν στους ML εκτιμητές όσο αυξάνεται ο αριθμός των ***quadrature*** σημείων στην πρώτη ή το μέγεθος του δείγματος στη δεύτερη. Εκτός από αυτές, υπάρχουν και άλλες μέθοδοι λιγότερο σύνθετες σε σχέση με τις προηγούμενες και η μεγιστοποίηση αφορά μια αναλυτική προσέγγιση της συνάρτησης πιθανοφάνειας. Η προσέγγιση βασίζεται στο ανάπτυγμα Taylor πρώτης τάξης της πιθανοφάνειας (Agresti et al., 2000).

Κατά καιρούς έχουν αναπτυχθεί διάφοροι αλγόριθμοι για το πρόβλημα αυτό, όπως των Breslow and Clayton (1993) που βασίζεται στην ιδέα της ψευδο-πιθανοφάνειας (***pseudo-likelihood***). Η μέθοδος αυτή αφορά την επαναληπτική εφαρμογή της θεωρίας των γενικευμένων γραμμικών μοντέλων και δεν περιλαμβάνει αριθμητική ολοκλήρωση ή Monte Carlo προσεγγίσεις και κατ'επέκταση είναι πιο εύκολος στην εφαρμογή από άλλες ακριβείς (***exact***) ***ML*** μεθόδους.

Το κύριο πλεονέκτημα της ***PQL*** είναι η απλότητα στην εφαρμογή της εφόσον δεν απαιτείται αριθμητική ολοκλήρωση και υπολογιστικά μπορεί να εφαρμοστεί για μεγάλο σύνολο δεδομένων και πολύπλοκα μοντέλα. Παρόλα αυτά, δεν αποδίδει ***ML*** εκτιμητές και αποδεικνύεται ότι σε σχέση με την μέγιστη πιθανοφάνεια η παρουσίασή της είναι μάλλον ανεπαρκής (McCulloch, 1997). Επίσης, η προσέγγιση που προκύπτει χειροτερεύει στην περίπτωση που τα δεδομένα έχουν μεγάλη απόκλιση από την κανονική κατανομή. Τα συμπεράσματα αυτά δικαιολογούν και τον όρο ***penalized***, αφού προς χάρην της απλότητας στην εφαρμογή δεν παίρνουμε τελικώς αποτελέσματα που να θεωρούνται ικανοποιητικά. Εν

τούτοις, η *PQL* μπορεί να δώσει μια καλή προσέγγιση της *ML* όταν οι διακυμάνσεις των τυχαίων επιδράσεων είναι σχετικά μικρές ή η μεταβλητή απόκρισης είναι περίπου κανονική.

3.4 Συμπερασματολογία των παραμέτρων του μοντέλου

Μετά την προσαρμογή ενδιαφερόμαστε για την συμπερασματολογία γύρω από τις παραμέτρους του μοντέλου. Αναφορικά με την παράμετρο **β**, μπορούμε χρησιμοποιώντας την ασυμπτωτική κανονικότητα του αντίστοιχου *ML* εκτιμητή να κατασκευάσουμε προσεγγιστικό διάστημα εμπιστοσύνης (McCullagh and Nelder, 1989). Επιπλέον, μπορούμε να κάνουμε ελέγχους υποθέσεων βασιζόμενοι στο τεστ του λόγου πιθανοφανειών $-2\log l$ (*likelihood ratio test*), που ακολουθεί ασυμπτωτικά την X^2 κατανομή κάτω από την μηδενική υπόθεση. Στην περίπτωση που η προσαρμογή του μοντέλου γίνεται με τη βοήθεια του *MCEM* αλγόριθμου, θα χρειαστεί πρόσθετος προγραμματισμός για την εκτίμηση των παραγώγων της πιθανοφάνειας για την τιμή του εκτιμητή του **β** (Booth and Hobert, 1999, sec. 6).

Στην περίπτωση των συνιστωσών της διασποράς το πρόβλημα είναι περισσότερο σύνθετο, καθώς η ποσότητα $-2\log l$ δεν ακολουθεί απαραίτητα ασυμπτωτικά την X^2 κατανομή. Η φύση των δεδομένων δεν έχει να κάνει με την ύπαρξη της συγκεκριμένης δυσκολίας αφού ανάλογη είναι και η περίπτωση των κανονικών δεδομένων (Miller, 1977). Σε γενικές γραμμές ο υπολογισμός της πραγματικής κατανομής της παραπάνω ποσότητας είναι αρκετά επίπονος. Εν τούτοις, δεν αποκλείεται ορισμένες φορές να είναι γνωστή. Ειδικότερα, αν στο μοντέλο υπάρχει μονοδιάστατη συνιστώσα διασποράς s^2 και ενδιαφερόμαστε για τον έλεγχο $H_0 : s^2 = 0$ έναντι $H_1 : s^2 > 0$, τότε αποδεικνύεται ότι η κατανομή του $-2\log l$ είναι 50:50 μίξη των X_0^2 και X_1^2 τυχαίων μεταβλητών (Self and Liang, 1987). Οπότε, όταν $\sigma>0$ και $t = -2\log l > 0$, το *p-value* για το τεστ θα είναι $\frac{1}{2}P(X_1^2 > t)$ (Agresti et al., 2000).

3.5 *Πρόβλεψη των τυχαίων επιδράσεων*

Εκτός από τις σταθερές επιδράσεις και την διασπορά, προβλέψεις μπορούμε να έχουμε και για τις τυχαίες επιδράσεις. Μετά τη συλλογή των δεδομένων, μπορούμε να πάρουμε πληροφορίες για τα u_i από την δεσμευμένη κατανομή $\mathbf{u}|\mathbf{y}$. Η συγκεκριμένη κατανομή καθορίζεται από το *GLMM* που έχουμε υποθέσει μέσω της σχέσης $f(\mathbf{u}\,|\,\mathbf{y}) \propto f(\mathbf{y},\mathbf{u})$. Για παράδειγμα, μια σημειακή εκτίμηση για το διάνυσμα $\mathbf{u}$ δίνεται από τη μέση τιμή $E(\mathbf{u}\,|\,\mathbf{y})$ (*posterior mean*). Είναι μια καλή εκτίμηση με την έννοια ότι το μέσο τετραγωνικό σφάλμα είναι μικρότερο από αυτό άλλων εκτιμητών (Searle et al., 1992).

Παρόλα αυτά, προκύπτουν δύο σημαντικά ζητήματα. Το πρώτο έχει να κάνει με το γεγονός ότι η ποσότητα $E(\mathbf{u}\,|\,\mathbf{y})$ εξαρτάται από τις παραμέτρους $\boldsymbol{\beta}$ και $\boldsymbol{\Sigma}$ ενώ το δεύτερο με το ότι συνήθως δεν δίνεται σε κλειστή μορφή. Το πρώτο μπορεί να παρακαμφθεί αν αντικαταστήσουμε τις εν λόγω παραμέτρους με τις αντίστοιχες εκτιμήσεις τους. Στην περίπτωση αυτή, ο εκτιμητής $\hat{\mathbf{u}}$ που προκύπτει από την δεσμευμένη μέση τιμή $E(\mathbf{u}\,|\,\mathbf{y})$ συχνά αναφέρεται ως ο εμπειρικός Bayes εκτιμητής (*emprical* Bayes *predictor*, Carlin and Louis, 1996). Το δεύτερο μπορεί να αντιμετωπιστεί με αριθμητικό υπολογισμό της $E(\mathbf{u}\,|\,\mathbf{y})$ είτε με ακρίβεια (π.χ. χρησιμοποιώντας Monte-Carlo μεθόδους) είτε προσεγγιστικά.

Στην περίπτωση που υπάρχουν αρκετά δεδομένα διαθέσιμα για κάθε άτομο, μπορούμε να έχουμε ακριβείς προβλέψεις για τις τυχαίες επιδράσεις. Κατά συνέπεια, αντιλαμβανόμαστε ότι οι ελλιπείς παρατηρήσεις δεν βοηθάνε προς την κατεύθυνση αυτή. Έστω $\mathbf{Y}_i$ το σύνολο των μετρήσεων για το i-άτομο και u_i η αντίστοιχη επίδραση. Η σχετική αβεβαιότητα μετράται με την υπό συνθήκη διακύμανση $Var(u_i\,|\,\mathbf{Y}_i)$ ή το αντίστοιχο τυπικό σφάλμα. Αυτά τα τυπικά σφάλματα της πρόβλεψης μπορούν να υπολογιστούν ή να προσεγγιστούν χρησιμοποιώντας την δεσμευμένη κατανομή που έχουμε υποθέσει και αντικαθιστώντας με τις εκτιμήσεις των άγνωστων παραμέτρων, όπου χρειάζεται.

Η τακτική αυτή έχει δεχθεί αρκετή κριτική, αφού αντιμετωπίζοντας τις παραμέτρους ως γνωστές η αβεβαιότητα τείνει να υποεκτιμάται. Οι Booth and Hobert (1998) προτείνουν κάποια μέθοδο ώστε να προκύπτουν ακριβέστερα αποτελέσματα, αν και οι ρυθμίσεις που κάνουν είναι αρκετά πολύπλοκες. Συνιστάται να καταφεύγει κανείς στην μέθοδο αυτή μόνο

όταν είναι περιορισμένο το σύνολο των δεδομένων καθώς διαφορετικά οποιαδήποτε διόρθωση είναι ελάσσωνος πρακτικής σημασίας.

3.6 Παράδειγμα

Ας θεωρήσουμε το παράδειγμα των 9 κριτών που είδαμε στην παράγραφο 3.2. Οι συμμεταβλητές που έχουμε στη διάθεσή μας είναι η επαφή [X_1, ναι (1) ή όχι (2)], η θερμοκρασία [X_2, υψηλή (1) ή χαμηλή (2)] και το μπουκάλι [X_3, πρώτο (1) ή δεύτερο (2)]. Στη συνέχεια δίνεται ο πίνακας των δεδομένων. Στο μοντέλο θα εισάγουμε όρους τυχαίων επιδράσεων θέλοντας με τον τρόπο αυτό να καταγράφεται η κυμαινόμενη ευαισθησία των κριτών σχετικά με την πικρότητα του κρασιού. Θεωρούμε το μοντέλο

$$\log it[P(Y_{it} \le k \mid u_i)] = a_k + b_1 X_{i1} + b_2 X_{i2} + b_3 X_{i3} + u_i \ , \qquad (3.3)$$

όπου Y_{it} είναι η απόκριση του i-κριτή, $i = 1, 2, ..., 9$ στην t-δοκιμή ενώ $k = 1, 2, 3, 4$ και $u_i \sim N(0, s^2)$. Λόγω της μορφής του σταθερού όρου αναφέρεται και ως *random-intercept* μοντέλο.

	Χαμηλή θερμοκρασία				Υψηλή θερμοκρασία			
	Χωρίς Επαφή		Με Επαφή		Χωρίς Επαφή		Με Επαφή	
Κριτής	Μπ 1	Μπ 2	Μπ 1	Μπ 2	Μπ 1	Μπ 2	Μπ 1	Μπ 2
1	2	3	3	4	4	4	5	5
2	1	2	1	3	2	3	5	4
3	2	3	3	2	5	5	4	4
4	3	2	3	2	3	2	5	3
5	2	3	4	3	3	3	3	3
6	3	2	3	2	2	4	5	4
7	1	1	2	2	2	3	2	3
8	2	2	2	3	3	3	3	4
9	1	2	3	2	3	2	4	4

Randall (1989)

Η εκτίμηση των παραμέτρων του μοντέλου πραγματοποιείται με την μέθοδο της μέγιστης πιθανοφάνειας ενώ ο υπολογισμός του ολοκληρώματος με την μέθοδο των Gauss-Hermite με 10 *quadrature* σημεία. Η συνάρτηση πιθανοφάνειας είναι της μορφής

$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma}; \mathbf{Y}) = f(\mathbf{Y}; \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \int f(\mathbf{Y} \mid \mathbf{u}; \boldsymbol{\beta}) f(\mathbf{u}; \boldsymbol{\Sigma}) d\mathbf{u} \quad ,$$

όπου $\boldsymbol{\beta} = (a_1, a_2, a_3, a_4, b_1, b_2, b_3)'$, $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_9)'$, $\mathbf{u} = (u_1, u_2, ..., u_9)'$ και λόγω της ανεξαρτησίας των u_i ο πίνακας συνδιακύμανσης θα είναι διαγώνιος με το *i*-διαγώνιο στοιχείο να είναι η διακύμανση s^2. Τα αποτελέσματα που προκύπτουν δίνονται στους ακόλουθους πίνακες (Tutz and Hennevogl, 1996).

Εκτίμηση των παραμέτρων του μοντέλου (3.3)

	G-H(10)
α_1	4.139
α_2	0.969
α_3	1.777
α_4	3.649
β_1	1.546 (0.437)
β_2	0.925 (0.347)
β_3	0.123 (0.320)
$\hat{s}$	1.243 (0.479)

Στις παρενθέσεις δίνονται οι τιμές των τυπικών σφαλμάτων

Εκτίμηση των τυχαίων επιδράσεων του μοντέλου (3.3)

Κριτής	*Posterior means* (G-H(10))
1	-1.853 (0.505)
2	0.650 (0.563)
3	-0.952 (0.622)
4	0.101 (0.686)
5	-0.261 (0.648)
6	- 0.498 (0.560)
7	2.021 (0.494)
8	0.395 (0.567)
9	0.621 (0.484)

Από τον πρώτο πίνακα παίρνουμε τα εξής αποτελέσματα:

- Για δεδομένο *i*, δηλαδή κριτή, και οποιαδήποτε τιμή για το *k*, όταν υπάρχει επαφή $(X_1 = 1)$ ο εκτιμώμενος λόγος πιθανοτήτων $\frac{P(Y \le k)}{1 - P(Y \le k)}$ είναι ίσος με $e^{\hat{b}_1 - 2\hat{b}_1} = e^{-\hat{b}_1} = e^{-1.546} = \frac{1}{e^{1.546}} \approx \frac{1}{4.7} \approx 0.21$ φορές τον αντίστοιχο λόγο όταν δεν υπάρχει επαφή, όπου $X_1 = 2$. Διαφορετικά, ο δεύτερος λόγος είναι $\frac{1}{0.21} \approx 4.76$ φορές ο πρώτος.
- Με τις ίδιες προϋποθέσεις, όταν η θερμοκρασία είναι υψηλή $(X_2 = 1)$ ο εκτιμώμενος λόγος πιθανοτήτων είναι $e^{\hat{b}_2 - 2\hat{b}_2} = e^{-\hat{b}_2} = e^{-0.925} = \frac{1}{e^{0.925}} \approx \frac{1}{2.5} \approx 0.4$ φορές τον αντίστοιχο λόγο όταν η θερμοκρασία είναι χαμηλή $(X_2 = 2)$. Δηλαδή ο δεύτερος λόγος είναι $\frac{1}{0.4} = 2.5$ φορές ο πρώτος.
- Ομοίως, όταν η δοκιμή αφορά το πρώτο μπουκάλι $(X_3 = 1)$, ο εκτιμώμενος λόγος πιθανοτήτων είναι $e^{\hat{b}_3 - 2\hat{b}_3} = e^{-\hat{b}_3} = e^{-0.123} = \frac{1}{e^{0.123}} \approx \frac{1}{1.13} \approx 0.88$ φορές τον αντίστοιχο λόγο για το δεύτερο μπουκάλι $(X_3 = 2)$. Άρα ο δεύτερος λόγος είναι $\frac{1}{0.88} \approx 1.13$ φορές ο πρώτος.

Παρατηρούμε ότι και με τις τρεις ερμηνευτικές μεταβλητές συμβαίνει να αυξάνεται ο λόγος πιθανοτήτων, όταν αυτές παίρνουν τη μεγαλύτερη τιμή των σκορ που έχουμε αναθέσει στα επίπεδά τους. Επίσης, οι τιμές των a_1, a_2, a_3 και a_4 δεν παρουσιάζουν ενδιαφέρον και χρησιμοποιούνται μόνο για τον υπολογισμό εκτιμήσεων των αθροιστικών πιθανοτήτων.

3.7 Σχέση περιθώριων και μοντέλων τυχαίων επιδράσεων

Αν θεωρήσουμε το μοντέλο (3.1), τότε το διάνυσμα **β** των παραμέτρων σταθερής επίδρασης έχει υπό συνθήκη (*conditional*) ερμηνεία δοθέντων των τυχαίων επιδράσεων. Οι σταθερές

επιδράσεις διακρίνονται σε δύο περιπτώσεις. Αν υποθέσουμε ότι η ερμηνευτική μεταβλητή παίρνει διαφορετικές τιμές για καθεμία από τις παρατηρήσεις μιας ομάδας (π.χ. δόσεις *T* διαφορετικών φαρμάκων), τότε ο αντίστοιχος συντελεστής στο μοντέλο εκφράζει την μεταβολή της απόκρισης όταν η ερμηνευτική μεταβλητή αυξηθεί κατά μία μονάδα εντός της ομάδας (***within cluster***). Στην αντίθετη περίπτωση που η τιμή της ανεξάρτητης μεταβλητής είναι σταθερή (π.χ. το φύλο), τότε ο συντελεστής της στο μοντέλο αναφέρεται στην μεταβολή της απόκρισης κατά την αύξηση "μιάς μονάδας" μεταξύ των ομάδων (***between cluster***).

Στην πρώτη περίπτωση, οι τυχαίες επιδράσεις αλλά και οποιεσδήποτε άλλες ανεξάρτητες μεταβλητές υπάρχουν στο μοντέλο παραμένουν σταθερές για τις παρατηρήσεις εντός ομάδας ενώ στην δεύτερη περίπτωση, οι τυχαίες επιδράσεις παραμένουν σταθερές ανάμεσα στις ομάδες. Με αυτή την έννοια, τα μοντέλα τυχαίων επιδράσεων είναι υπό συνθήκη μοντέλα αφού εφαρμόζονται δοθείσης της τιμής των τυχαίων επιδράσεων. Το ίδιο δεν συμβαίνει, όμως, και με τα περιθώρια. Σε αυτά η μεταβολή των ερμηνευτικών μεταβλητών υπολογίζεται κατά μέσο όρο όλων των ομάδων (***population averaged***) και δεν έχει να κάνει με συγκρίσεις των ομάδων για δεδομένη τιμή των τυχαίων επιδράσεων.

Οι δύο τύποι μοντέλων συνδέονται κατά κάποιον τρόπο μεταξύ τους. Από το μοντέλο (3.1), αντιστρέφοντας την συνάρτηση σύνδεσης θα έχουμε

$$E(Y_{it} \mid \mathbf{u}_i) = g^{-1}(\mathbf{x}_{it}^{'}\boldsymbol{\beta} + \mathbf{z}_{it}^{'}\mathbf{u}_i)\,.$$

Παίρνοντας τον μέσο όρο των τυχαίων επιδράσεων προκύπτει ότι

$$E(Y_{it}) = E[E(Y_{it} \mid \mathbf{u}_i)] = \int g^{-1}(\mathbf{x}_{it}^{'}\boldsymbol{\beta} + \mathbf{z}_{it}^{'}\mathbf{u}_i)\, f(\mathbf{u}_i; \boldsymbol{\Sigma}) d\mathbf{u}_i \ ,$$

όπου $f(\mathbf{u}_i; \boldsymbol{\Sigma})$ είναι η συνάρτηση πιθανότητας της $N(\mathbf{0}, \boldsymbol{\Sigma})$ κατανομής των τυχαίων επιδράσεων. Αν ο σύνδεσμος ήταν ο ταυτοτικός, τότε θα είχαμε

$$E(Y_{it}) = E[E(Y_{it} \mid \mathbf{u}_i)] = \int (\mathbf{x}_{it}^{'}\boldsymbol{\beta} + \mathbf{z}_{it}^{'}\mathbf{u}_i)\, f(\mathbf{u}_i; \boldsymbol{\Sigma}) d\mathbf{u}_i = \mathbf{x}_{it}^{'}\boldsymbol{\beta}\,.$$

Το αντίστοιχο περιθώριο μοντέλο έχει την ίδια μορφή αλλά και την ίδια επίδραση $\boldsymbol{\beta}$ με το παραπάνω. Για οποιονδήποτε άλλο μη γραμμικό σύνδεσμο, όπως είναι ο *logit*, οι *population-averaged* επιδράσεις των περιθώριων μοντέλων συχνά είναι μικρότερες από αυτές των υπό συνθήκη (*conditional*) επιδράσεων. Παρόλα αυτά, όμως, από τα υπό συνθήκη μοντέλα μπορούμε να εξάγουμε πληροφορίες για τα περιθώρια μοντέλα.

Έχουμε αναφέρει ότι στα μοντέλα τυχαίων επιδράσεων η ερμηνεία των συντελεστών αφορά την επίδραση των συμμεταβλητών στην απόκριση για κάθε υποκείμενο ενώ στα περιθώρια η αντίστοιχη ερμηνεία αφορά την επίδραση των συμμεταβλητών στην μέση τιμή της απόκρισης για το σύνολο του πληθυσμού που μελετάται. Η επιλογή που καλείται να κάνει κάποιος ανάμεσα στις δύο αυτές προσεγγίσεις εξαρτάται από τη φύση του προβλήματος. Στις περισσότερες των εφαρμογών, πάντως, προτιμώνται τα περιθώρια μοντέλα. Για παράδειγμα, σε μια κλινική δοκιμή σχεδιασμένη για την σύγκριση ενός φαρμάκου με το ψευδοφάρμακο (*placebo*), οι ερευνητές ενδιαφέρονται περισσότερο για την μέση διαφορά ανάμεσα στις δύο αγωγές και επομένως, η προσαρμογή ενός περιθώριου μοντέλου είναι καταλληλότερη. Επιπλέον, το συγκεκριμένο μοντέλο συνιστάται όταν όλες οι συμμεταβλητές είναι ανεξάρτητες από το χρόνο (Zeger et al., 1988; Neuhaus et al., 1991; Graubard and Korn, 1994).

Κεφάλαιο 4ο
Διαθέσιμο Λογισμικό
(Software)

Οι ερευνητές δείχνουν πάντοτε ξεχωριστό ενδιαφέρον για την ανάλυση δεδομένων που προκύπτουν από ένα διαχρονικό σχεδιασμό. Παρ'ότι έχουν αναπτυχθεί πλήθος προσεγγίσεων βασισμένες στην πιθανοφάνεια όσον αφορά πολυμεταβλητές κανονικές αποκρίσεις, τα μοντέλα για διακριτού τύπου αποκρίσεις απαιτούν διαφορετική προσέγγιση. Έτσι, ιδιαίτερες προσπάθειες έχουν γίνει τα τελευταία χρόνια προκειμένου να αντιμετωπιστεί και πρακτικά, εκτός από θεωρητικά, η προσαρμογή τόσο των περιθώριων όσο και των μοντέλων τυχαίων επιδράσεων. Ωστόσο, η πολυπλοκότητα σε ορισμένες περιπτώσεις δεν μας αφήνει πολλά περιθώρια. Στην παράγραφο 4.1 αναφέρονται τα στατιστικά πακέτα και οι ρουτίνες για την προσαρμογή και εκτίμηση των περιθώριων μοντέλων και στην επόμενη 4.2 θίγουμε το ίδιο θέμα για τα μοντέλα τυχαίων επιδράσεων.

4.1 Περιθώρια μοντέλα

Ø *SPSS*

Με το στατιστικό πακέτο *SPSS* μπορούμε να αντιμετωπίσουμε μόνο επαναλαμβανόμενες μετρήσεις που αφορούν σε συνεχείς μεταβλητές μέσα από τη διαδικασία

Analyze ► General Linear Model ► Repeated Measures

Ανάλυση για διακριτές μεταβλητές δεν προβλέπεται από το σχετικό μενού.

Ø *S-Plus*

Αν θέλουμε να χρησιμοποιήσουμε το στατιστικό πακέτο *S-Plus*, οφείλουμε να εγκαταστήσουμε τη βιβλιοθήκη *GEE* στις ήδη υπάρχουσες. Μπορούμε να αναζητήσουμε το συγκεκριμένο φάκελο μέσα από το STATLIB και στη συνέχεια να τον μεταφέρουμε στον φάκελο *Library* μέσα από τη διαδικασία

Program Files ► *sp*2000 ► *Library* ,

όπου υπάρχουν και οι υπόλοιπες. Όταν πρόκειται να κάνουμε ανάλυση επαναλαμβανόμενων μετρήσεων και επιθυμούμε να χρησιμοποιήσουμε τη συγκεκριμένη βιβλιοθήκη, τότε επιλέγουμε

File ► Load Library

και την διαλέγουμε από τη λίστα. Η ρουτίνα με την οποία θα χειριστούμε το πρόβλημα είναι η *gee* και πληροφορίες για την σύνταξή της μπορούμε να πάρουμε αν στο παράθυρο των *Commands* πληκτρολογήσουμε *help*(*gee*), οπότε και θα εμφανιστεί το σχετικό παράθυρο. Σημειώνουμε ότι για την χρήση της εν λόγω εντολής, θα πρέπει η διατάξιμη μεταβλητή απόκρισης να μετατραπεί σε δίτιμη. Μια τέτοια προσέγγιση αλλά και εφαρμογές περιγράφονται από τον Clayton (MRC Biostatistics Unit, 1992).

Σε περιβάλλον *S-Plus* λειτουργούν και εναλλακτικά προγράμματα, όπως είναι το Oswald. Η τελευταία έκδοσή του αλλά και *up-to-date* πληροφορίες διατίθενται από την ιστοσελίδα του http://www.maths.lancs.ac.uk/Software/Oswald/. Οι απαιτήσεις για την λειτουργία του αφορούν την έκδοση 3.0 του *S-Plus* ή μεγαλύτερη και έναν *ANSI C compiler*. Επιπλέον, λειτουργεί κάτω από *UNIX* και *DOS* συστήματα. Αποτελεί ένα σύνολο συναρτήσεων (*functions*) και κατηγοριών δεδομένων (*data types*) για την ανάλυση διαχρονικών δεδομένων εντός του στατιστικού περιβάλλοντος του *S-Plus*. Οι ήδη υπάρχουσες συναρτήσεις είναι σχεδιασμένες για την ανάλυση των νέων κατηγοριών και οι πρόσθετες συναρτήσεις που περιλαμβάνει το Oswald δεν είναι πολλές, ώστε αν ο χρήστης διαθέτει ήδη κάποια οικειότητα με το *S-Plus* να καταβάλλει την ελάχιστη προσπάθεια για την ανάλυση.

Στο Oswald, τα περιθώρια μοντέλα προσαρμόζονται με την βοήθεια της εντολής *gee.fit* και παρουσιάζει ομοιότητες στη σύνταξη της με την αντίστοιχη του *S-Plus*. Είναι, επίσης, διαθέσιμη από το STATLIB. Στην περίπτωση που η μεταβλητή απόκρισης είναι δίτιμη, συνηθίζεται η εφαρμογή της εντολής *alr.fit* (*Alternating Logistic Regression-ALR*). Η συγκεκριμένη ανάλυση εισήχθη από τους Carey, Zeger and Diggle (1993). Η εντολή αυτή

υποστηρίζει μόνο την ***exchangeable*** δομή συσχέτισης ενώ πέραν του καθορισμού του μοντέλου για την περιθώρια μέση τιμή δεν χρειάζεται κάτι επιπλέον. Περισσότερες πληροφορίες μπορεί να αναζητήσει κανείς στο εγχειρίδιο (***manual***) του προγράμματος (Smith, Robertson and Diggle, 1996). Εφαρμογές για τις δύο παραπάνω εντολές υπάρχουν στις ενότητες 5.2 και 5.3 αυτού.

Ø *R*

Πρόκειται για ένα λογισμικό με ιδιαίτερες δυνατότητες, το οποίο λειτουργεί μέσω εντολών σε περιβάλλον που θυμίζει αρκετά το παράθυρο των *Commands* του ***S-Plus***. Εν τούτοις υπάρχουν αρκετές διαφορές ανάμεσά τους, τις οποίες μπορεί κανείς να αναζητήσει στην ιστοσελίδα http://www.stats.ox.ac.uk/pub/R. Όσον αφορά διαχρονικά δεδομένα, υπάρχει το πακέτο εντολών ***GEEPACK***. Δεν περιλαμβάνεται μέσα στα ήδη υπάρχοντα και επομένως καλείται ο ενδιαφερόμενος να το εγκαταστήσει. Οι εντολές που μας ενδιαφέρουν είναι οι ***geese*** και ***ordgee***. Αφορούν επαναλαμβανόμενες δίτιμες και διατάξιμες αποκρίσεις αντίστοιχα. Η σύνταξη της πρώτης είναι εντελώς όμοια με αυτή της εντολής ***gee*** του ***S-Plus***. Οι διαφορές ανάμεσα στις εντολές ***geese*** και ***gee*** για την εκτίμηση του διανύσματος **β** είναι αρκετά μικρές ενώ και με τις δύο μεθόδους για τον πίνακα συνδιακύμανσης παίρνουμε την ανθεκτική εκτίμηση.

Για την δεύτερη, όμως, χρειάζεται να γίνουν ορισμένες ρυθμίσεις. Η μεταβλητή απόκρισης θα πρέπει να δηλωθεί ως διατάξιμη και επιπλέον τα δεδομένα πρέπει να είναι αρχικά ταξινομημένα κατά υποκείμενο και στη συνέχεια χρονικά (***by time within subject***). Ο δεύτερος όρος είναι ιδιαιτέρως αναγκαίος στην περίπτωση που υποθέτουμε την ***AR***-1 δομή και λιγότερο για τις υπόλοιπες. Επιπλέον, και οι δύο εντολές υποθέτουν ότι δεν υπάρχουν ελλιπείς τιμές και έτσι, αν για κάποιο υποκείμενο λείπει κάποια παρατήρηση, αυτό θα πρέπει να εξαιρεθεί. Περισσότερες πληροφορίες για τις δύο εντολές δίνοναι από το μενού της βοήθειας του πακέτου. Τέλος, οι ενδιαφερόμενοι μπορούν να ανατρέξουν στο εγχειρίδιο του ***S-Plus/R*** που συνοδεύει το βιβλίο ***Categorical Data Analysis*** (Agresti, 2002) και έχει επιμεληθεί η Laura Thompson, στο οποίο υπάρχουν παραδείγματα εφαρμογής των συγκεκριμένων (και όχι μόνο) εντολών για δεδομένα που περιγράφονται στο εν λόγω βιβλίο.

Ø *SAS*

Με βάση τα σημερινά δεδομένα, η γλώσσα προγραμματισμού *SAS* προσφέρει τις περισσότερες δυνατότητες όσον αφορά τις μεθόδους για επαναλαμβανόμενα κατηγορικά δεδομένα. Η διαδικασία *GENMOD* μπορεί να πραγματοποιήσει *GEE* αναλύσεις για περιθώρια μοντέλα παρέχοντας αρκετές επιλογές αναφορικά με την οικογένεια κατανομής των δεδομένων, την συνάρτηση σύνδεσης (*link function*) και τη δομή του πίνακα συσχέτισης. Παρ'όλα αυτά, δεν δίνεται η δυνατότητα να προσαρμόσουμε τα συγκεκριμένα μοντέλα με τη μέθοδο *ML*. Η έκδοση που μπορεί να χρησιμοποιηθεί για τις εκτιμήσεις είναι η *SAS/STAT Release* 6.12 (*SAS Institute* 1996) καθώς επίσης και οι μεταγενέστερες. Περισσότερες πληροφορίες μπορούν να αναζητηθούν στην σχετική ιστοσελίδα http://www.sas.com αλλά και από την εργασία των Horton and Lipsitz, (1999).

Ø *SUDAAN*

Πρόκειται για ένα διαφορετικό λογισμικό πρόγραμμα με εντολές που κινείται στην φιλοσοφία του *SAS* με περιορισμένες, όμως, δυνατότητες. Οι διαδικασίες *PROC LOGISTIC*, *PROC MULTILOG* και *PROC REGRESS* επιτρέπουν την προσαρμογή και εκτίμηση μοντέλων χρησιμοποιώντας τη μέθοδο των *GEE* για δίτιμες και συνεχείς αποκρίσεις. Οι επιλογές για την οικογένεια κατανομών περιορίζονται στην Gauss, Bernoulli/Διωνυμική και Poisson. Υποστηρίζει τους *cumulative logit* και *generalized logit* συνδέσμους. Όταν, όμως, πρόκειται για την διωνυμική κατανομή, περιορίζεται μόνο στον κανονικό. Επίσης, πέραν της ανεξάρτητης και της *exchangeable* δομής για τον πίνακα συσχέτισης δεν υπάρχουν εναλλακτικές επιλογές. Περαιτέρω πληροφορίες διατίθενται από τη σελίδα του Research Triangle Institute (http://www.rti.org/patents/sudaan/sudaan.html) και την εργασία των Horton and Lipsitz, (1999).

Ø *RMORD* (*Repeated Measures ORDinal Data*)

Πρόκειται για ένα εύχρηστο πρόγραμμα *FORTRAN* για την ανάλυση ομαδοποιημένων διατάξιμων αποκρίσεων με βάση τη μέθοδο των Stram, Wei and Ware, (1988). Αποτελεί μια επέκταση του *proportional-odds* μοντέλου (Mc Cullagh, 1980) στην περίπτωση που οι

ομάδες των αποκρίσεων συσχετίζονται. Χειρίζεται ελλιπή δεδομένα, αρκεί να είναι *MCAR* (*missing completely at random*), συμμεταβλητές που μεταβάλλονται ανά ομάδα ή ακόμα και χρονοεξαρτώμενες. Το πρόγραμμα διατίθενται από το STATLIB. Πληροφορίες για την χρήση του αλλά και παραδείγματα μπορούν οι ενδιαφερόμενοι να αναζητήσουν στην εργασία των Davis and Hall, (1996).

Ø *MAREG/WinMAREG*

Πρόκειται για ένα εναλλακτικό λογισμικό εργαλείο, που δημιούργησαν οι Kastner, Fieger and Heumann, (1997). Με την βοήθεια του μπορούμε να διαχειριστούμε δίτιμα, κατηγορικά ακόμα και συνεχή δεδομένα με περιορισμένες, όμως, επιλογές όσον αφορά την συνάρτηση σύνδεσης. Παρά το γεγονός ότι σχεδιάστηκε για την ανάλυση συσχετισμένων δεδομένων, εντούτοις προβλέπεται και η ανάλυση αντίστοιχων μη συσχετισμένων. Το εν λόγω λογισμικό χωρίζεται σε δύο μέρη: το *MAREG*, που χρησιμοποιείται για την εκτίμηση του μοντέλου που επιθυμούμε να αναλύσουμε και το *WinMAREG*, που χρησιμοποιείται για τον καθορισμό του.

Το *MAREG* είναι διαθέσιμο και εφαρμόζεται σε πλατφόρμες όπως το *DOS* και το *UNIX* (ή *Sun Solaris*, όπως αποκαλείται σήμερα). Αντιθέτως, το *WinMAREG* εφαρμόζεται στα *Windows*. Η τρέχουσα έκδοση υποστηρίζει μόνο *dBase* και *Paradox database αρχεία*. Τέτοιου είδους αρχεία μπορεί να δημιουργήσει σε προγράμματα όπως το *Excel* ή το *SPSS*. Ειδικότερα, τα αρχεία πρέπει να σώζονται ως *dBase IV* αρχεία ενώ η επέκταση *dbf* πρέπει να εμφανίζεται με καφαλαία έτσι, ώστε να διαβάζονται από το *WinMAREG*. Σημειώνουμε ότι παρουσιάζει ανεπαρκές επικοινωνιακό υπόβαθρο. Τα μηνύματα σφάλματος που εμφανίζονται δεν είναι ιδιαίτερα κατατοπιστικά, ώστε να εντοπίζει ο χρήστης το λάθος του ή που οφείλεται η αδυναμία του προγράμματος να διεκπεραιώσει την ανάλυση.

Διαθέσιμες συναρτήσεις σύνδεσης είναι η ταυτοτική για συνεχείς, ο *logit* σύνδεσμος για δίτιμες, ο αθροιστικός *logit* για διατάξιμες και ο πολυωνυμικός (*multinomial*) *logit* για γενικές περιπτώσεις πολύτιμων αποκρίσεων. Επίσης, για την *GEE* προσέγγιση εφαρμόζεται η μέθοδος *IEE* (*Independence Estimating Equations*) και η μέθοδος του Prentice, (1988). Τέλος, αν στα δεδομένα υπάρχουν ελλιπείς τιμές, το πρόγραμμα διενεργεί την ανάλυση μόνο για τις πλήρεις περιπτώσεις, εκτός από μια ειδική περίσταση που θα αναφέρουμε στο επόμενο κεφάλαιο. Το συγκεκριμένο πρόγραμμα μπορεί κανείς να το αποκτήσει μέσω της ηλεκτρονικής διεύθυνσης ftp.stat.uni-muenchen.de και ειδικότερα από τον κατάλογο

/*pub*/*sfb*386/*c*3/***mareg***. Στην ίδια σελίδα μπορεί να αναζητήσει πληροφορίες και για την εγκατάστασή του.

4.2 Μοντέλα τυχαίων επιδράσεων

Ø *SPSS*

Για τα συγκεκριμένα μοντέλα προβλέπεται αυτή τη φορά διαχείριση των δεδομένων με βάση το στατιστικό πακέτο *SPSS*. Ακολουθώντας την διαδικασία

Analyze ► *Mixed models* ► *Linear*

μπορούμε να δηλώσουμε τις μεταβλητές των οποίων οι τιμές δηλώνουν τις πολλαπλές παρατηρήσεις για κάθε υποκείμενο. Συνεχίζοντας έχουμε τη δυνατότητα να ορίσουμε την εξαρτημένη μεταβλητή, τις συμμεταβλητές αλλά και καθορίσουμε τους παράγοντες σταθερών και τυχαίων επιδράσεων. Μας παρέχεται η δυνατότητα για την εκτίμηση των παραμέτρων να χρησιμοποιήσουμε την μέθοδο μέγιστης πιθανοφάνειας (*ML*) ή την μέθοδο περιορισμένης μέγιστης πιθανοφάνειας (*REML*) αλλά και τον καθορισμό στοιχείων που αφορούν τη σύγκλιση των εν λόγω αλγορίθμων. Επίσης, το πακέτο υπολογίζει πλήθος στατιστικών, όπως πίνακες συνδιακύμανσης των παραμέτρων του μοντέλου, των τυχαίων επιδράσεων αλλά και των σφαλμάτων (***residuals***). Τέλος, μέσα από ένα σύνολο επιλογών μπορούμε να ορίσουμε και τη δομή του χρησιμοποιούμενου πίνακα συσχέτισης (***working correlation matrix***).

Ø *S-Plus*

Μέσω του *S-Plus* μπορούμε να κατασκευάσουμε ένα μικτό (***mixed***) μοντέλο είτε γραμμικό είτε μη γραμμικό, σε αντίθεση με το *SPSS* που προσαρμόζει μόνο γραμμικό. Η διαδικασία που ακολουθούμε είναι

Statistics ► *Mixed effects* ► *Linear/Nonlinear*

Όπως στο *SPSS*, έτσι και στο *S-Plus* υπάρχουν διαθέσιμες οι μέθοδοι *ML* και *REML* για την εκτίμηση των παραμέτρων. Επίσης, για τον πίνακα συσχέτισης υπάρχουν αρκετές

εναλλακτικές επιλογές, αν και λιγότερες απ' ότι προβλέπονται στο *SPSS*. Το συγκεκριμένο πρόγραμμα υποστηρίζει μόνο συνεχείς μεταβλητές απόκρισης.

- *SAS*

Τα τελευταία χρόνια αποτελεί την πιο δημοφιλή επιλογή για την προσαρμογή και εκτίμηση μοντέλων τυχαίων επιδράσεων, με τη βοήθεια της μακροεντολής *GLIMMIX*. Αποδίδει εκτιμητές που προκύπτουν προσεγγίζοντας την πιθανοφάνεια με χρήση των μεθόδων των Breslow and Clayton, (1993) και Wolfinger and O'Connell, (1993). Προσαρμόζει μοντέλα δίνοντας αρκετές επιλογές για την κατανομή της εξαρτημένης μεταβλητής και τη συνάρτηση σύνδεσης, υποθέτοντας πάντα κανονικότητα για τις τυχαίες επιδράσεις. Σε μεταγενέστερη έκδοση (*Version* 7) εισήχθη η μακροεντολή *PROC NLMIXED*. Για τον καθορισμό της πιθανοφάνειας χρησιμοποιεί μια προσαρμόσιμη (*adaptive*) έκδοση της μεθόδου Gauss-Hermite για το ολοκλήρωμα. Η συγκεκριμένη εντολή είναι σημαντικά καλύτερη της προηγούμενης, όταν οι συνιστώσες της διασποράς είναι μεγάλες ή τα δεδομένα απέχουν μακράν από την κανονική κατανομή. Παρουσιάζει, όμως, και αδυναμίες. Οι *quadrature* μέθοδοι είναι υπολογιστικά εφικτές για ολοκληρώματα μικρής διάστασης ενώ προς το παρόν η εντολή *NMIXED* δεν υιοθετεί ένθετα (*nested*) μοντέλα.

- *MIXOR*

Πρόκειται για ένα πρόγραμμα *FORTRAN* που οφείλεται στους Hedeker and Gibbons, (1996) και είναι διαθέσιμο για αθροιστικά (*cumulative*) logit μοντέλα. Για την εκτίμηση των παραμέτρων χρησιμοποιεί την μέγιστη περιθώρια πιθανοφάνεια (*marginal likelihood*) και τη μέθοδο αριθμητικής ολοκλήρωσης των Gauss-Hermite, αλλά τα τυπικά σφάλματα βασίζονται στην αναμενόμενη (*expected*) πληροφορία ενώ το *NLMIXED* του *SAS* χρησιμοποιεί την παρατηρούμενη (*observed*). Για την κατανομή των τυχαίων επιδράσεων εκτός από την κανονική μας παρέχεται η δυνατότητα να υποθέσουμε και την ομοιόμορφη. Επίσης, υπάρχουν και αρκετές επιλογές για την συνάρτηση σύνδεσης, όπως οι *log-log*, *probit*, *logistic* και *complementary log-log* σύνδεσμοι. Αναφέρουμε ότι για τις τυχαίες επιδράσεις χρησιμοποιεί Bayes εκτιμητές. Τέλος, προβλέπεται η προσαρμογή μοντέλων στα οποία οι

συμμεταβλητές μεταβάλλονται χρονικά. Περισσότερες πληροφορίες μπορούν να αναζητηθούν στην ηλεκτρονική διεύθυνση http://www.uic.edu/~hedeker/mixdos.html.

Κεφάλαιο 5ο
Εφαρμογή

5.1 Αριθμητική εφαρμογή με την χρήση των GEE

5.1.1 Περιγραφή των δεδομένων

Στην εργασία των Koch et al., (1989) περιγράφεται μια τυχαιοποιημένη κλινική δοκιμή για ένα νέο φάρμακο που αφορά την αντιμετώπιση αναπνευστικών διαταραχών. Στη έρευνα αυτή έλαβαν μέρος 111 ασθενείς, οι οποίοι με τυχαίο τρόπο χωρίστηκαν σε δύο ομάδες. Στην πρώτη ανήκουν εκείνοι που λαμβάνουν το υπό εξέταση φάρμακο και στη δεύτερη εκείνοι που λαμβάνουν εικονικό φάρμακο (***placebo***).

Κατά την διάρκεια της ***follow-up*** περιόδου, πραγματοποιήθηκαν τέσσερεις επισκέψεις των ασθενών. Σε κάθε μία από αυτές αξιολογήθηκε η κατάσταση της υγείας τους με βάση μια διατάξιμη κλίμακα 5 επιπέδων (0 = πολύ κακή, 1 = κακή, 2 = μέτρια, 3 = καλή, 4 = πολύ καλή). Στη συγκεκριμένη εφαρμογή θα περιορίσουμε την κλίμακα σε τρία επίπεδα και ειδικότερα με τιμές 1 = κακή, 2 = καλή και 3 = πολύ καλή. Το ερώτημα που πρέπει να απαντηθεί είναι αν στο πέρασμα του χρόνου οι ασθενείς που λαμβάνουν το νέο φάρμακο παρουσίαζουν σημαντική βελτίωση σε σχέση με τους υπόλοιπους.

Τα δεδομένα του προβλήματος θα αναλυθούν με την χρήση των ***GEE*** μέσα από το περιβάλλον των ***MAREG***, ***S-Plus*** και ***R***.

5.1.2 Εφαρμογή με το MAREG

Το συγκεκριμένο σύνολο δεδομένων διατίθεται στο φάκελο ***Examples*** του *MAREG*. Πρόκειται για ένα αρχείο ***dbf*** με το όνομα Miller, καθώς αυτός τα ανέλυσε το 1993. Ανοίγουμε το εν λόγω αρχείο μέσα από τη διαδικασία

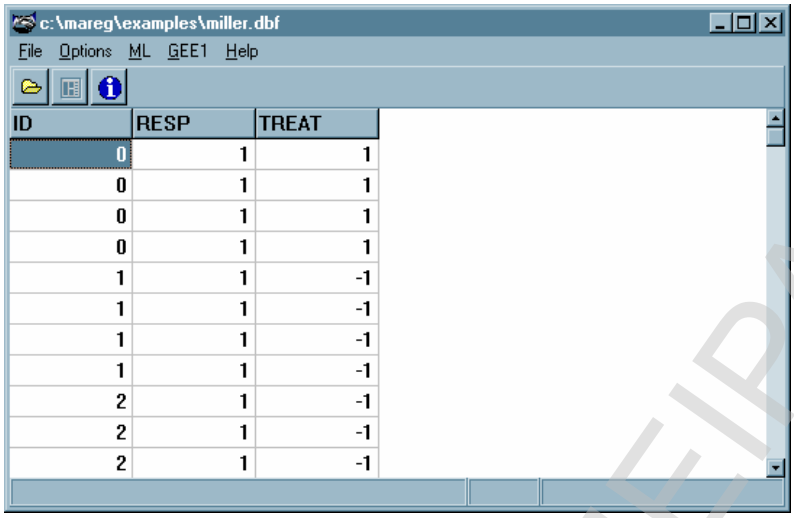
C ►*Program Files* ►*MAREG* ►*Examples* ► Miller

Ανοίγοντας το πρόγραμμα βρισκόμαστε μπροστά στην κύρια οθόνη (Εικόνα 1) και επιλέγοντας από το μενού

File ►*Open* ►*Examples*

το σχετικό αρχείο εμφανίζονται στην οθόνη (Εικόνα 2).

Εικόνα 1



ID	RESP	TREAT
0	1	1
0	1	1
0	1	1
0	1	1
1	1	-1
1	1	-1
1	1	-1
1	1	-1
2	1	-1
2	1	-1
2	1	-1

Εικόνα 2

Η μεταβλητή *ID*, που δίνεται στην πρώτη στήλη, εισάγεται ώστε να ξεχωρίζουν οι μετρήσεις για κάθε ασθενή και στην ουσία είναι αυτή που διαμορφώνει τις ομάδες (*clusters*). Η επόμενη, με την ονομασία *RESP* από το *RESPONSE*, είναι η μεταβλητή απόκρισης με την κωδικοποίηση 1,2 και 3, όπως περιγράφηκε στην ενότητα 5.1.1. Τέλος, η τελευταία μεταβλητή αναφέρεται στην θεραπεία που έλαβε ο κάθε ασθενής. Αν παίρνει το νέο φάρμακο δίνεται η τιμή 1 και διαφορετικά η τιμή -1. Για παράδειγμα, οι τέσσερεις πρώτες γραμμές αφορούν τον πρώτο ασθενή (*ID* = 0) για τον οποίο και στις τέσσερεις επισκέψεις η κατάσταση της υγείας του κρίνεται κακή (*RESP* = 1) ενώ ανήκει στην ομάδα εκείνων που λαμβάνουν την ενεργό θεραπεία (*TREAT* = 1).

Θα αντιμετωπίσουμε το πρόβλημα με τη βοήθεια της μεθόδου των *GEE*. Στη σειρά των εργαλείων αναφέρεται ως *GEE*1. Υπενθυμίζουμε ότι η συγκεκριμένη ονομασία έχει ως σκοπό να την διαχωρίσει από μεταγενέστερες επεκτάσεις. Επιλέγοντας τη μέθοδο αυτή, στη συνέχεια καλούμαστε με βάση τον τύπο της απόκρισης να διαλέξουμε την σύναρτηση σύνδεσης και κατά συνέπεια τον τύπο του μοντέλου που θα προσαρμόσουμε. Στην περίπτωση μας, όπου η απόκριση είναι κατηγορική και διατάξιμη, θα προτιμήσουμε τον αθροιστικό *logit*. Δηλαδή

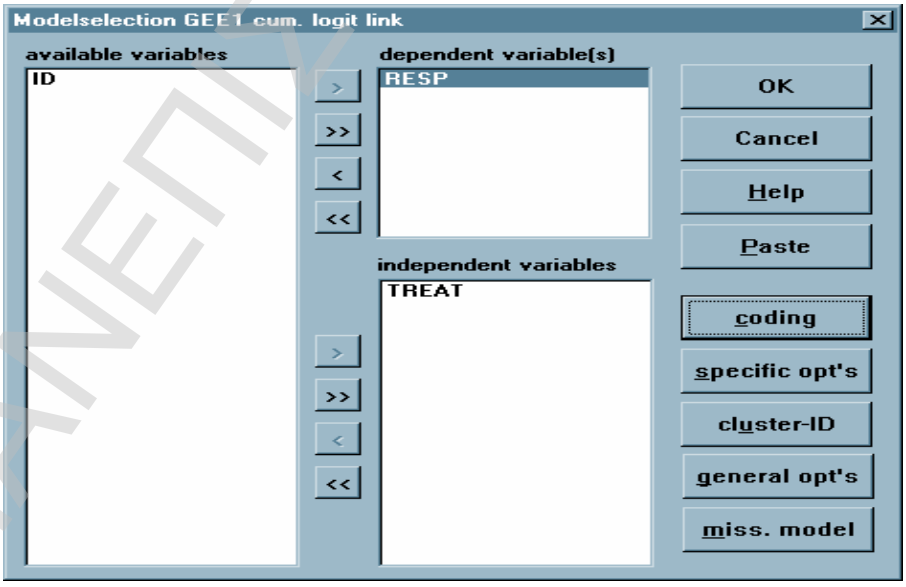
GEE1 ►*Categorical* ►*Cum. logit* .

Άρα, αν Y_{it} είναι η κατάσταση της υγείας του *i*-ασθενή με i = 1,2,…,111 κατά την *t*-επίσκεψη, όπου t = 1,2,3 και 4, και ως x συμβολίσουμε την θεραπεία, τότε το μοντέλο που θα προσαρμόσουμε θα είναι

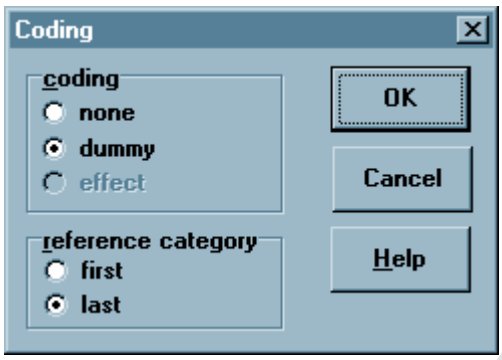
$$\log it[P(Y_{it} \le k \mid x_i)] = a_k + b_t x_i \,, \qquad (5.1)$$

όπου k = 1,2. Στο μοντέλο αυτό ο συντελεστής της θεραπείας φέρει τον δείκτη *t*, επιτρέποντας διαφορετικές επιδράσεις για κάθε χρονική στιγμή. Αντίθετα, η μεταβλητή x δεν φέρει τον ίδιο δείκτη, εφόσον ο κάθε ασθενής λαμβάνει σταθερά την ίδια θεραπεία και στις τέσσερεις μετρήσεις.

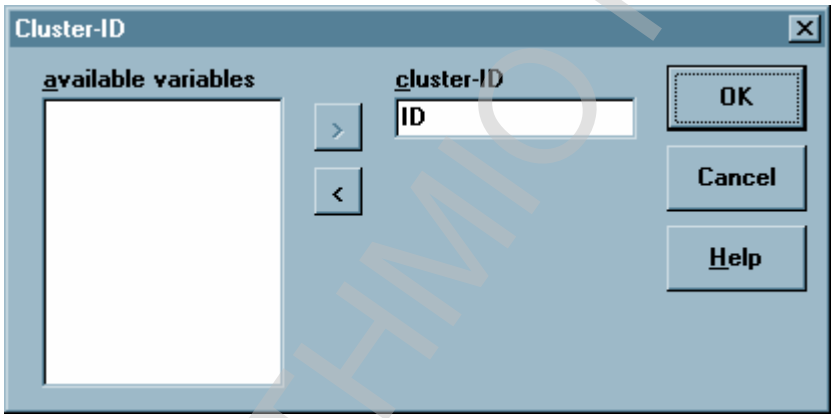
Στη συνέχεια εμφανίζεται το παράθυρο (Εικόνα 3) στο οποίο δηλώνουμε την μεταβλητή απόκρισης αλλά και την μοναδική ερμηνευτική που διαθέτουμε. Υποχρεούμαστε μολις ορίσουμε την πρώτη, να την δηλώσουμε ως δείκτρια (*dummy*). Αυτό γίνεται επιλέγοντας *Coding* από τα δεξιά (Εικόνα 4) (αφού έχουμε μαρκάρει την μεταβλητή) ή κάνοντας δεξί κλικ πάνω στην μεταβλητή *RESP*. Στο συγκεκριμένο σημείο μπορούμε να ορίσουμε το επίπεδο της μεταβλητής που θα αποτελεί αυτό της αναφοράς. Έχουμε τη δυνατότητα να επιλέξουμε το πρώτο ή το τελευταίο. Για την μεταβλητή *TREAT* δεν χρειάζεται να κάνουμε το ίδιο, εφόσον έχει μόνο δύο επίπεδα. Έπειτα, επιλέγουμε *Cluster-ID* και δηλώνουμε την *ID* μεταβλητή, ως αυτή που ορίζει τις ομάδες (Εικόνα 5).



Εικόνα 3

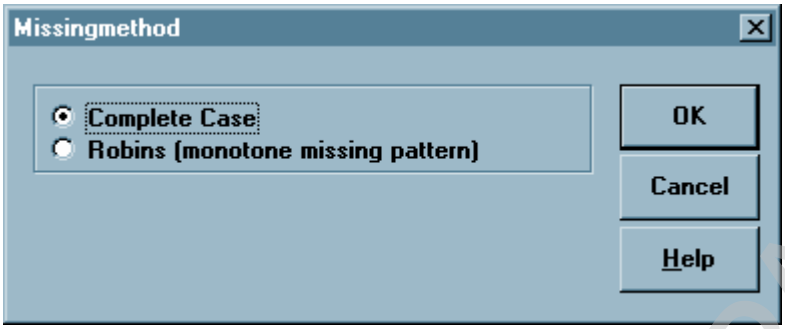


Εικόνα 4



Εικόνα 5

Όσον αφορά ελλιπή δεδομένα (*miss.model*, Εικόνα 6), το *MAREG* υποστηρίζει δύο τρόπους χειρισμού αυτών. Στην περίπτωση που λείπουν κάποιες μετρήσεις της μεταβλητής απόκρισης βάση ενός μονότονου (monotone) μηχανισμού, προσφέρεται η μέθοδος των Robins et al., (1995). Σημειώνουμε ότι οι συμμεταβλητές δεν επιτρέπεται να έχουν ελλιπείς τιμές. Αν, όμως, δεν υφίσταται τέτοιος μηχανισμός, το πρόγραμμα κάνει την ανάλυση με βάση τα άτομα που έχουν όλες τις μετρήσεις διαθέσιμες. Στη δική μας εφαρμογή, συγκεκριμένα, δεν υπάρχουν ελλιπή δεδομένα.

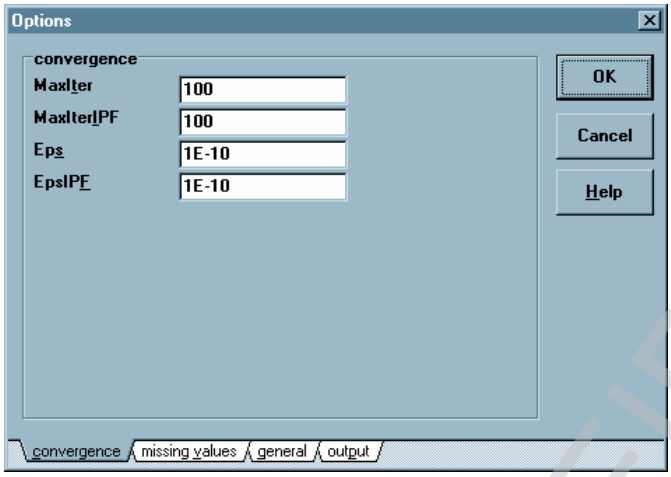


Εικόνα 6

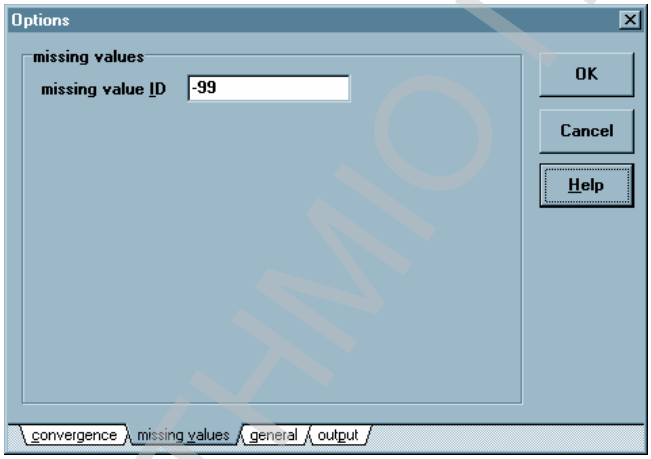
Επίσης, μέσω της επιλογής *general options* (*general opt's*, Εικόνα 7) μας δίνεται η δυνατότητα να κάνουμε αρκετές ρυθμίσεις. Στο κάτω μέρος του παραθύρου υπάρχουν τέσσερεις καρτέλες (*convergence, missing values, general* και *output*). Στην πρώτη μπορούμε να ορίσουμε τον μέγιστο αριθμό επαναλήψεων (*iterations*) για τον αλγόριθμο καθώς επίσης και την τιμή του έψιλον για την σύγκλισή του. Στην δική μας περίπτωση θα διατηρήσουμε τις εξ ορισμού τιμές.

Στη δεύτερη καρτέλα (Εικόνα 8) ορίζουμε την τιμή που θέλουμε να αντιλαμβάνεται το πρόγραμμα ως ελλιπές δεδομένο. Η τιμή που προτείνει το *MAREG* είναι -99. Μπορεί, όμως, να αλλάξει από το χρήστη. Οι τιμές που έχουν δηλωθεί ως ελλιπείς αντιμετωπίζονται από το πρόγραμμα βάσει της αντίστοιχης επιλογής που έχει γίνει προηγουμένως..

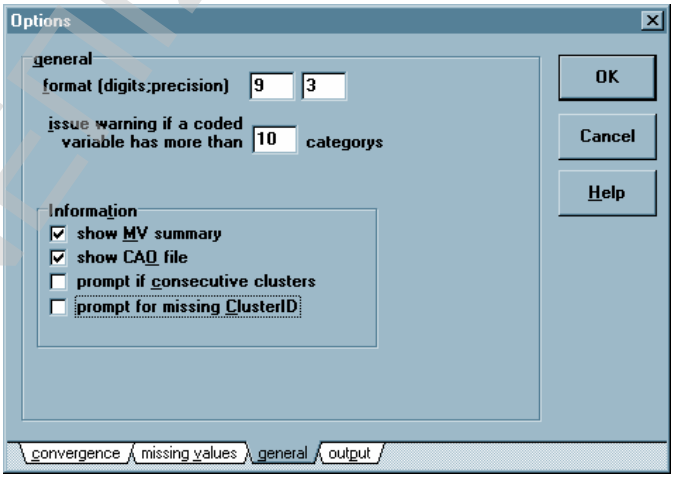
Στην τρίτη (Εικόνα 9) μπορούμε να κάνουμε επιλογές για την ακρίβεια των αριθμών και για άλλες πληροφορίες ή προειδοποιήσεις που θέλουμε να μας γνωστοποιούνται. Για περισσότερες λεπτομέρειες δεν έχουμε παρά να επιλέξουμε τη βοήθεια στα δεξιά. Στην τέταρτη, τέλος, επιλέγουμε αν θέλουμε το output που εμαφανίζεται μετά το πέρας της ανάλυσης να έχει πολλά στοιχεία ή όχι αλλά και αν θέλουμε ή όχι να παρακολουθούμε πως τρέχει ο αλγόριθμος στο παράθυρο του *DOS*.



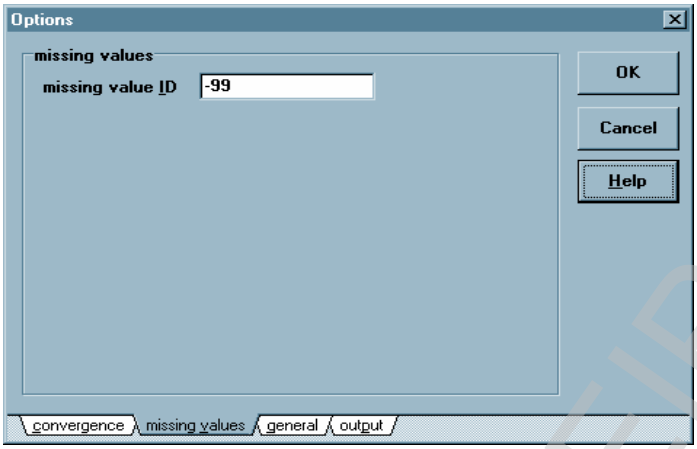
Εικόνα 7



Εικόνα 8



Εικόνα 9

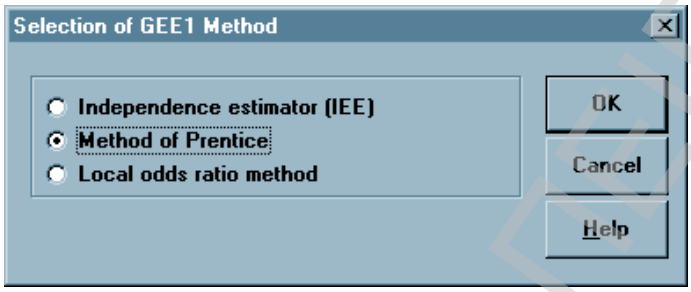


Εικόνα 10

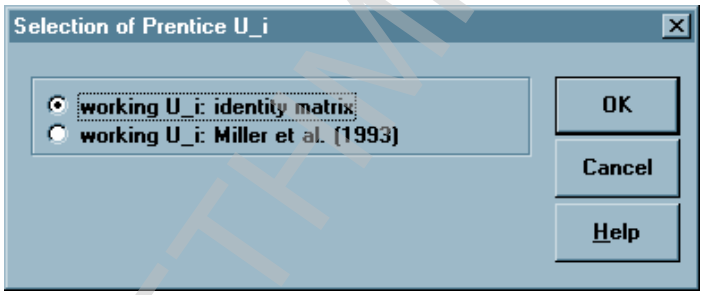
Το πιο σημαντικό βήμα, ωστόσο, έχει να κάνει με την επιλογή της δομής του πίνακα συσχέτισης αλλά και τη μέθοδο εκτίμησής του. Επιλέγοντας από τα δεξιά specific options (***specific opt's***) εμφανίζονται οι τρεις εναλλακτικές μέθοδοι (Εικόνα 11). Για την εφαρμογή μας θα συνεχίσουμε με την μέθοδο του Prentice. Ενδεικτικά, αναφέρουμε ότι είναι μια μέθοδος, η οποία για την εκτίμηση του πίνακα συσχέτισης χρησιμοποιεί εξισώσεις που έχουν δομή όμοια με αυτή των εξισώσεων (2.6) της παραγράφου 2.4.3. Κατ'αναλογία, λοιπόν, υπάρχει ένας πίνακας συνδιακύμανσης (στο πρόγραμμα συμβολίζεται με *U_i*) και του οποίου τα στοιχεία υπολογίζονται όπως περιγράφεται στην εργασία των Miller et al., (1993). Επιλέγοντας *OK*, εμφανίζεται το παράθυρο στο οποίο καλούμαστε να αποφασίσουμε αν θα χρησιμοποιήσουμε ως πίνακα *U_i* τον ταυτοτικό ή την ακριβή εκτίμησή του. Με βάση τους δημιουργούς του ***MAREG***, και οι δύο επιλογές είναι εξίσου λογικές. Επιλέγουμε, λοιπόν, την ταυτοτική δομή. Μετά το *OK*, εμφανίζεται το τελευταίο παράθυρο, όπου δηλώνουμε αν οι παράμετροι στο μοντέλο είναι σταθερές στο χρόνο (***fixed effects***) ή μεταβαλλόμενες (π.χ. ***varying intercepts*** ή/και ***varying covariate effects***). Συγκεκριμένα, ορίζουμε ***varying covariate effects*** και επιπλέον, επιλέγουμε τη δομή του πίνακα συσχέτισης μεταξύ των ***exchangeable***, ***stationary***, ***undefined*** και ***userdefined***. Για την εφαρμογή μας επιλέγουμε την πρώτη. Μετά το *OK*, γυρίζουμε στο αρχικό παράθυρο με τα δεδομένα.

Επιλέγοντας πάλι το *OK*, το πρόγραμμα θα μας ζητήσει να αποθηκεύσουμε, όπου εμείς θέλουμε, κάποια αρχεία σχετικά με τα δεδομένα και τα αποτελέσματα. Όλα θα έχουν την ίδια ονομασία. Πιο συγκεκριμένα, δημιουργούνται τέσσερα αρχεία με επεκτάσεις ***cad***, ***cai***, ***cal*** και ***cao***. Το πρώτο περιέχει το σύνολο των δεδομένων χωρίς τις ονομασίες των μεταβλητών. Το

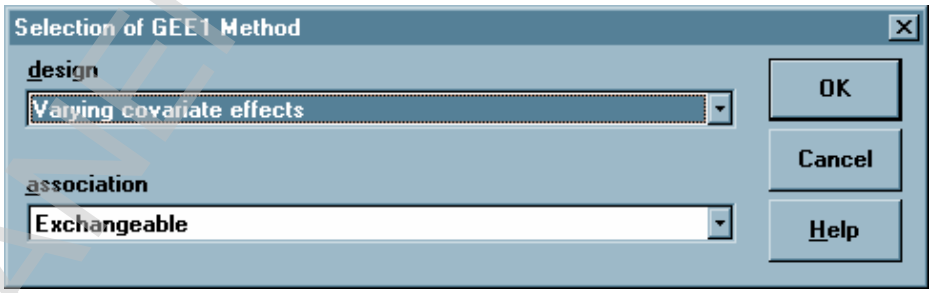
δεύτερο, που ανοίγει στο *Notepad*, περιέχει όλες τις επιλογές που κάναμε για το μοντέλο, τις παραμέτρους κτλ. Το τρίτο εμφανίζει τις επαναλήψεις του αλγόριθμου, όπως αυτές γίνονται στο *DOS* μέχρι την σύγκλιση και στο τελευταίο περιέχονται τα αποτελέσματα. Με τον τρόπο αυτό δεν χάνουμε καμιά πληροφορία .



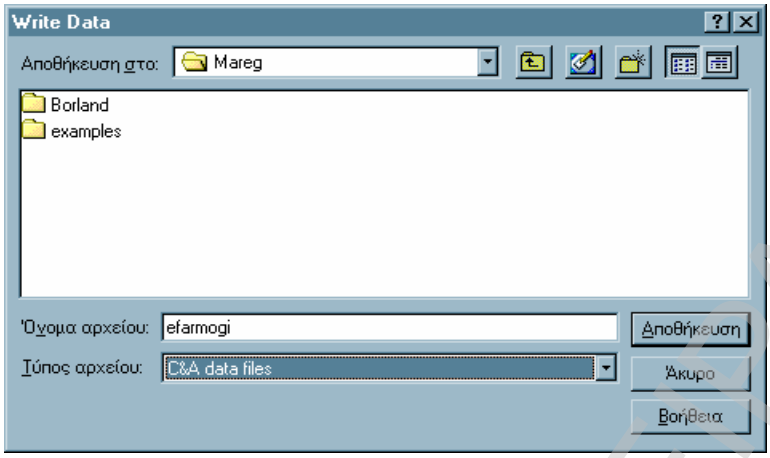
Εικόνα 11



Εικόνα 12



Εικόνα 13



Εικόνα 14

Στη συνέχεια και έπειτα από τη σύγκλιση του αλγορίθμου παίρνουμε τον Πίνακα 1 των αποτελεσμάτων.

MAREG version 0.2.0 (c) 01.06.1999 SFB386-C3. All rights reserved.
----- GEE Prentice estimator :
Inifile: h.cai
Out file: h.cao
Log file: h.cal
Data file: h.cad
Orig. sample size, #clusters: 444, 111
Act. sample size, #clusters: 444, 111
Estimation method: GEE
Special estimatior: Method of Prentice
Link: cumulative logit link
Variance function: Multinomial variance function
Design: Varying covariate effects
User given epsilon: 1e-010
User given maxiter: 100
Tolerance beta reached: 2.07364e-011
Iterations needed: 13
Estimated overdispersion: no overdispersion was estimated

Var.	beta	std.	r. std.	Wald	r. Wald	p	r. p
resp(1)	-1.612	0.184	0.198	76.869	66.351	0.000	0.000
resp(2)	0.676	0.163	0.164	17.266	17.097	0.000	0.000
treat_1	-0.223	0.176	0.172	1.593	1.668	0.207	0.197
treat_2	-0.755	0.184	0.185	16.909	16.631	0.000	0.000
treat_3	-0.541	0.177	0.199	9.327	7.383	0.002	0.007
treat_4	-0.389	0.176	0.198	4.858	3.857	0.028	0.050

Association model: Exchangeable correlation model, method of Prentice

Tolerance alpha reached: 9.03175e-011
Score equation for alpha uses: Identity matrix

alpha	std.	r. std.	Wald	r. Wald	p	r. p	corr.
0.898	0.094	0.203	90.912	19.627	0.000	0.000	0.421
-0.201	0.078	0.105	6.593	3.664	0.010	0.056	-0.100
-0.113	0.078	0.133	2.118	0.727	0.146	0.394	-0.057
0.660	0.086	0.121	58.579	29.745	0.000	0.000	0.319

Πίνακας 1: Αποτελέσματα της ανάλυσης του μοντέλου (5.1) με το MAREG

Το πρώτο τμήμα του πίνακα αναγράφει τις επιλογές που κάναμε για το μοντέλο και την ανάλυση. Στο δεύτερο τμήμα εμφανίζονται οι εκτιμήσεις των παραμέτρων του επιλεγμένου μοντέλου, οι τυπικές αποκλίσεις αυτών, τα στατιστικά τεστ και τα αντίστοιχα *p-values*.

Καθώς η μεταβλητή απόκρισης έχει τρία επίπεδα και σύμφωνα με το μοντέλο (5.1), υπάρχουν δύο σταθερές (*intercepts*) ενδεικνυόμενες με το όνομα της μεταβλητής απόκρισης. Δηλαδή *resp(1)* και *resp(2)* αντίστοιχα. Ακολούθως, δίνονται οι εκτιμήσεις για την επίδραση της θεραπείας κατά τις τέσσερεις χρονικές στιγμές. Από τον πίνακα παρατηρούμε ότι η παράμετρος b_1 (*treat_1*) είναι στατιστικά μη σημαντική. Το αποτέλεσμα αυτό είναι λογικό να συμβαίνει αφού σε μικρό χρονικό διάστημα δεν είναι πάντα εφικτό να καταγράφεται σημαντική διαφορά ανάμεσα στις δύο θεραπείες.

Η μέγιστη επίδραση της θεραπείας παρατηρείται στη δεύτερη επίσκεψη ενώ σταδιακά στη συνέχεια ελαττώνεται. Πιο συγκεκριμένα:

- Για $t=2$, θα έχουμε $\frac{P(Y_{i2} \le k)}{P(Y_{i2} > k)} = e^{\hat{a}_k + \hat{b}_2}$, όταν $x_i = 1$ (δηλαδή για τους ασθενείς που λαμβάνουν το ενεργό φάρμακο) και $\frac{P(Y_{i2} \le k)}{P(Y_{i2} > k)} = e^{\hat{a}_k - \hat{b}_2}$, όταν $x_i = -1$ (δηλαδή για τους ασθενείς που παίρνουν *placebo*). Ο λόγος των παραπάνω *odds* είναι ίσος με $e^{\hat{a}_k + \hat{b}_2 - \hat{a}_k + \hat{b}_2} = e^{2\hat{b}_2} = e^{2(-0.755)} = e^{-1.51} = 0.221$. Δηλαδή, το *odds* του ενδεχομένου η απόκριση Y_{i2} του *i*-υποκειμένου κατά την δεύτερη χρονική στιγμή να είναι σε κατηγορία μικρότερη ή και ίση με την k ($Y_{it} \le k$) έναντι του ενδεχομένου να είναι σε μεγαλύτερη ($Y_{it} > k$) για τους ασθενείς που λαμβάνουν *placebo* είναι ίσο με $\frac{1}{0.221} \approx 4.52$ φορές το αντίστοιχο *odds* για τους ασθενείς που λαμβάνουν το ενεργό φάρμακο.
- Ομοίως, για $t=3$ θα έχουμε ότι κατά την τρίτη επίσκεψη το ίδιο *odds* για τους ασθενείς που λαμβάνουν placebo είναι ίσο με $\frac{1}{e^{2\hat{b}_3}} = \frac{1}{e^{2(-0.541)}} = e^{1.082} = 2.951$ φορές το αντίστοιχο *odds* για τους υπόλοιπους ασθενείς.
- Τέλος, για $t=4$ προκύπτει ότι στην τελευταία επίσκεψη το ίδιο *odds* για τους ασθενείς που λαμβάνουν *placebo* είναι ίσο με $\frac{1}{e^{2\hat{b}_4}} = \frac{1}{e^{2(-0.389)}} = e^{0.778} = 2.177$ φορές το αντίστοιχο *odds* για τους ασθενείς υπό την ενεργή αγωγή. Σημειώνουμε ότι τα παραπάνω συμπεράσματα δεν εξαρτώνται από το επίπεδο k της απόκρισης (σοβαρότητα ασθένειας).

Το τελευταίο τμήμα του *output* αναφέρεται στη δομή της συσχέτισης (***association structure***). Το *MAREG* λειτουργεί σύμφωνα με τη μέθοδο του Prentice, (1988), όπου αν η απόκριση έχει *k* κατηγορίες, τότε για το κάθε υποκείμενο δημιουργούνται *k*-1 δείκτριες, όπως ακριβώς στην μέθοδο των Stram et al., (1988). Οπότε για τον *i*-ασθενή με $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})'$ και εφόσον $k = 3$, θα έχουμε

$$\underset{Y_{i1}}{\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix}}, \underset{Y_{i2}}{\begin{pmatrix} Y_{i1} \\ Y_{i3} \end{pmatrix}}, \underset{Y_{i3}}{\begin{pmatrix} Y_{i1} \\ Y_{i4} \end{pmatrix}}, \underset{Y_{i4}}{\begin{pmatrix} Y_{i2} \\ Y_{i3} \end{pmatrix}}.$$

Επομένως, ο πίνακας συσχέτισης για τις μετρήσεις του *i*-ασθενή θα είναι

$$\begin{pmatrix} \mathbf{1} & \text{corr}(\mathbf{Y}_{i1}, \mathbf{Y}_{i2})(\mathbf{a}) & \text{corr}(\mathbf{Y}_{i1}, \mathbf{Y}_{i3})(\mathbf{a}) & \text{corr}(\mathbf{Y}_{i1}, \mathbf{Y}_{i4})(\mathbf{a}) \\ \text{corr}(\mathbf{Y}_{i2}, \mathbf{Y}_{i1})(\mathbf{a}) & \mathbf{1} & \text{corr}(\mathbf{Y}_{i2}, \mathbf{Y}_{i3})(\mathbf{a}) & \text{corr}(\mathbf{Y}_{i2}, \mathbf{Y}_{i4})(\mathbf{a}) \\ \text{corr}(\mathbf{Y}_{i3}, \mathbf{Y}_{i1})(\mathbf{a}) & \text{corr}(\mathbf{Y}_{i3}, \mathbf{Y}_{i2})(\mathbf{a}) & \mathbf{1} & \text{corr}(\mathbf{Y}_{i3}, \mathbf{Y}_{i4})(\mathbf{a}) \\ \text{corr}(\mathbf{Y}_{i4}, \mathbf{Y}_{i1})(\mathbf{a}) & \text{corr}(\mathbf{Y}_{i4}, \mathbf{Y}_{i2})(\mathbf{a}) & \text{corr}(\mathbf{Y}_{i4}, \mathbf{Y}_{i3})(\mathbf{a}) & \mathbf{1} \end{pmatrix},$$

όπου, όμως, η κάθε συσχέτιση $\text{corr}(\mathbf{Y}_{ik}, \mathbf{Y}_{il})(\mathbf{a})$ με $\kappa,\lambda = 1,2,3,4$ και $k \neq l$ αντιστοιχεί στον 2×2 πίνακα των συσχετίσεων, σε συνάρτηση με το διάνυσμα **a**, ανάμεσα στις συνιστώσες των $\mathbf{Y}_{ik}, \mathbf{Y}_{il}$ και ο οποίος είναι κοινός για όλες αυτές, αφού έχουμε υποθέσει την ***exchangeable*** δομή. Για παράδειγμα, θα είναι

$$\text{corr}(\mathbf{Y}_{i1}, \mathbf{Y}_{i2})(\mathbf{a}) = \begin{pmatrix} \text{corr}(Y_{i11}, Y_{i21}) & \text{corr}(Y_{i11}, Y_{i22}) \\ \text{corr}(Y_{i12}, Y_{i21}) & \text{corr}(Y_{i12}, Y_{i22}) \end{pmatrix}(\mathbf{a})$$

και με βάση τα αποτελέσματα θα είναι

$$\text{corr}(\mathbf{Y}_{i1}, \mathbf{Y}_{i2})(\mathbf{a}) = \begin{pmatrix} 0.421 & -0.100 \\ -0.057 & 0.319 \end{pmatrix}.$$

Σημειώνουμε ότι τα τυπικά σφάλματα, τα σταστιστικά τεστ του Wald αλλά και τα ***p-values*** αντιστοιχούν στις παραμέτρους a_1, a_2, a_3 και a_4 (***association parameters***) που συνδέονται με τις συσχετίσεις.

5.1.3 Εφαρμογή με το R

Προτού συνεχίσουμε με την ανάλυση των δεδομένων με τη βοήθεια του λογισμικού προγράμματος *R*, υπενθυμίζουμε ότι χρειάζεται να υπάρχει το πακέτο *GEEPACK* μέσα στα ήδη διαθέσιμα. Ανοίγοντας το πρόγραμμα, κάλουμε το συγκεκριμένο πακέτο με την εντολή

library(geepack) .

Στην περίπτωση που έχουμε κάποιο αρχείο στο *Excel*, στο *SPSS* ή στο *S-Plus* και επιθυμούμε να το διαβάσουμε μέσα από το *R*, υπάρχει ο εξής εύκολος τρόπος. Μετατρέπουμε το συγκεκριμένο σε αρχείο txt, κάνοντας αντιγραφή και επικόλληση των δεδομένων στο Σημειωματάριο (*Notepad*). Στη συνέχεια το αποθηκεύουμε εντός του φακέλου *R*, που υπάρχει στο φάκελο *Program Files* του δίσκου. Για να το διαβάσουμε από το πρόγραμμά μας, γράφουμε

όνομα<-*read.table*(“*όνομα.txt*”,*header=T/F*) .

Αν το *txt* αρχείο στην πρώτη γράμμη έχει τα ονόματα των μεταβλητών, τότε γράφουμε *header=T* (*True*). Διαφορετικά, *header=F* (*False*).

Αναφορικά με τα δικά μας δεδομένα, συμβαίνει να υπάρχουν εξ ορισμού εντός του συγκεκριμένου πακέτου. Εμφανίζονται στην οθόνη μετά την εντολή

data(*respdis*)

και είναι της μορφής

	y_1	y_2	y_3	y_4	trt
1	1	1	1	1	1
2	1	1	1	1	0
3	1	1	1	1	0
M	M	M	M	M	M
107	3	3	3	3	0
108	3	3	3	3	0
109	3	3	3	3	0
110	3	3	3	3	0
111	3	3	3	3	0

Στην πρώτη γραμμή καταγράφονται οι τέσσερεις μετρήσεις για το πρώτο άτομο (y_1,y_2,y_3,y_4) και η δείκτρια για τη θεραπεία (*trt*). Ομοίως και για τα υπόλοιπα άτομα. Τα δεδομένα με τη συγκεκριμένη μορφή δεν είναι άμεσα επεξεργάσιμα. Πληκτρολογώντας ***help***("*ordgee*"), εμφανίζεται το παράθυρο της βοήθειας και στο τέλος υπάρχουν ορισμένες εντολές που αναφέρονται στο συγκεκριμένο παράδειγμα για την διαμόρφωση των δεδομένων αλλά και την διάταξη αυτών ως προς τη μεταβλητή που διαχωρίζει τις ομάδες (δηλαδή, η μεταβλητή *ID* που είδαμε κατά την εφαρμογή του *MAREG*) και το χρόνο. Σημειώνουμε ότι η θεραπεία δηλώνεται εδώ με τα επίπεδα 1 για το ενεργό φάρμακο και 0 για το *placebo* (αντί για -1 που είχαμε στο *MAREG*). Παίρνουμε τα εξής αποτελέσματα

	trt	time	resp	id
1.1	1	1	1	1
2.1	0	1	1	2
3.1	0	1	1	3
4.1	0	1	1	4
5.1	0	1	1	5
M	M	M	M	M
107.4	0	4	3	107
108.4	0	4	3	108
109.4	0	4	3	109
110.4	0	4	3	110
111.4	0	4	3	111

Στις πρώτες 111 γραμμές εμφανίζονται οι μετρήσεις που έγιναν για τους ασθενείς κατά την πρώτη επίσκεψη, στις επόμενες 111 κατά την δεύτερη επίσκεψη κοκ. Επίσης, οι ασθενείς καταγράφονται πάντοτε με την ίδια σειρά σύμφωνα με την *id* μεταβλητή. Στη συνέχεια προγραμματίζουμε το μοντέλο

$$\log \text{it}[P(Y_{it} \le k \mid x_i)] = a_k + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2 \quad , \qquad (5.2)$$

όπου X_1 και X_2 είναι η θεραπεία και ο χρόνος αντίστοιχα ενώ έχουμε συμπεριλάβει και τον όρο αλληλεπίδρασης, σύμφωνα με την εντολή *ordgee* ως εξής:

*model.fit <-ordgee(ordered(resp)~ trt * time, id=id, data=respdis1, mean.link=''logit'', corstr="exchangeable", int.const=TRUE, control=geese.control(maxit=*100*)) .*

Το προτελευταίο όρισμα είναι μια λογική μεταβλητή και παίρνει την τιμή *TRUE* αν θέλουμε να έχουμε σταθερές χρονικά τις παραμέτρους $a_k, k=1,2$ του μοντέλου (5.2). Διαφορετικά, παίρνει την τιμή *FALSE*. Το τελευταίο όρισμα έχει να κάνει με τη σύγκλιση του αλγόριθμου. Στην συγκεκριμένη περίπτωση έχουμε καθορίσει τον μέγιστο αριθμό επαναλήψεων σε 100. Περισσότερες πληροφορίες αλλά και ρυθμίσεις (*arguments*) στην παραπάνω εντολή μπορούμε να δούμε μέσα από την βοήθεια.

Τέλος, πληκτρολογώντας

summary(model.fit)

παίρνουμε τα ακόλουθα αποτελέσματα

```
ordgee(formula = ordered(resp) ~ trt * time, id = id, data = respdis1, mean.link = "logit",
        corstr = "exchangeable", control = geese.control(maxit = 100), int.const = TRUE)

Mean Model:
Mean Link:                  logit
Variance to Mean Relation:  binomial

Coefficients:
             estimate        san.se       wald            p
Inter:1     1.28350838   0.27671753  21.5141440  3.512286e-06
Inter:2    -1.00153999   0.29083644  11.8587513  5.739101e-04
trt1        1.27811140   0.47267573   7.3115729  6.851193e-03
time       -0.05468851   0.08257273   0.4386509  5.077743e-01
trt1:time  -0.08596422   0.14070213   0.3732794  5.412221e-01

Scale is fixed.

Correlation Model:
Correlation Structure:     exchangeable
Correlation Link:          log

 Estimated Correlation Parameters:
          estimate      san.se      wald      p
alpha     2.512394   0.2864575  76.92279      0

Returned Error Value:  0
 Number of clusters:   111   Maximum cluster size: 4
```

Πίνακας 2: Αποτελέσματα ανάλυσης του μοντέλου (5.2) με το R

Από τα παραπάνω αποτελέσματα παρατηρούμε ότι η επίδραση της θεραπείας (παράμετρος *trt*1) είναι στατιστικά σημαντική. Η συγκεκριμένη μεταβλητή παρατηρείται (μετράται) μόνο κατά την πρώτη χρονική στιγμή (***baseline***) και διαφοροποιείται μεταξύ των ασθενών. Κατά συνέπεια, το αποτέλεσμα αυτό αποτελεί μια ένδειξη ότι οι θεραπείες αρχικά έχουν κάποια διαφορά.

Αναφορικά με την αλληλεπίδραση ανάμεσα στο χρόνο και τη θεραπεία, παρατηρούμε ότι για τον αντίστοιχο συντελεστή προκύπτει ένα ***p-value*** μεγαλύτερο από το αποδεκτό επίπεδο 5% και άρα ο συγκεκριμένος όρος κρίνεται μη στατιστικά σημαντικός. Στο ίδιο συμπέρασμα καταλήγουμε και για τον συντελεστή που αντιστοιχεί στο χρόνο. Πιο αναλυτικά, θα έχουμε:

- Για $t=1$ θα είναι $\log it[P(Y_{i1} \le k)] = \hat{a}_k + \hat{b}_1 + \hat{b}_2 + \hat{b}_3$ για την ομάδα του ενεργού φαρμάκου και $\log it[P(Y_{i1} \le k)] = \hat{a}_k + \hat{b}_2$ για την ομάδα του ***placebo***. Άρα, το εκτιμώμενο ***odds*** του ενδεχομένου η απόκριση υπό το φάρμακο να είναι μικρότερη οποιασδήποτε κατηγορίας ***k*** έναντι του να είναι μεγαλύτερη είναι ίσο με $\exp(\hat{b}_1 + \hat{b}_3) = \exp(1.28 - 0.086) = \exp(1.194) \approx 3.3$ φορές το αντίστοιχο εκτιμώμενο ***odds*** για την ομάδα του ***placebo***.
- Για $t=2$ θα έχουμε $\log it[P(Y_{i2} \le k)] = \hat{a}_k + \hat{b}_1 + 2\hat{b}_2 + 2\hat{b}_3$ και $\log it[P(Y_{i2} \le k\,|)] = \hat{a}_k + 2\hat{b}_2$ για το ενεργό φάρμακο και του ***placebo*** αντίστοιχα. Επομένως, ο λόγος των ***odds*** που αναφέραμε προηγουμένως θα είναι $\exp(\hat{b}_1 + 2\hat{b}_3) = \exp(1.28 - 2*0.086) = \exp(1.108) \approx 3.03$.
- Για $t=3$ θα είναι $\log it[P(Y_{i3} \le k)] = \hat{a}_k + \hat{b}_1 + 3\hat{b}_2 + 3\hat{b}_3$ και $\log it[P(Y_{i3} \le k\,|)] = \hat{a}_k + 3\hat{b}_2$ για τις ομάδες του φαρμάκου και του ***placebo*** αντίστοιχα. Δηλαδή, ο λόγος των ***odds*** θα είναι ίσος με $\exp(\hat{b}_1 + 3\hat{b}_3) = \exp(1.28 - 3*0.086) = \exp(1.022) \approx 2.78$.
- Στην τελευταία χρονική στιγμή $(t=4)$ θα είναι $\log it[P(Y_{i4} \le k)] = \hat{a}_k + \hat{b}_1 + 4\hat{b}_2 + 4\hat{b}_3$ και $\log it[P(Y_{i4} \le k\,|)] = \hat{a}_k + 4\hat{b}_2$ για τις ομάδες του φαρμάκου και του ***placeb***ο αντίστοιχα. Οπότε ο λόγος των ***odds*** θα ισούται με $\exp(\hat{b}_1 + 4\hat{b}_3) = \exp(1.28 - 4*0.086) = \exp(1.022) \approx 0.936$.

Τα αποτελέσματα περιλαμβάνουν και την εκτίμηση της παραμέτρου συσχέτισης α που εμφανίζεται στον αντίστοιχο πίνακα σύμφωνα με την δομή που του έχουμε ορίσει. Δεν πρόκειται για την ίδια τη συσχέτιση ανάμεσα στις παρατηρήσεις ενός ατόμου αλλά για κάποια παράμετρο που σχετίζεται με αυτή. Στην ενότητα 2.4.5 είδαμε ότι ο Prentice, (1988) εφάρμοσε το παρακάτω σύστημα εξισώσεων για την εκτίμηση των παραμέτων **a**.

$$\mathbf{U}(\mathbf{a}) = \sum_{i=1}^{N} \mathbf{E}_i^{'} \mathbf{W}_i^{-1} (\mathbf{Z}_i - \boldsymbol{\eta}_i(\mathbf{a})) = \mathbf{0}.$$

Οι Lipsitz et al., (1991) προτείνουν την αποφυγή περιορισμών για το διάστημα των παραμέτρων συσχέτισης χρησιμοποιώντας τον αντίστροφο μετασχηματισμό του Fisher (*inverse of Fisher's z transformation*). Ο εν λόγω σύνδεσμος εξασφαλίζει ότι οι συσχετίσεις θα είναι εντός του διαστήματος (-1,1).

5.1.4 Εφαρμογή με το S-Plus

Για την ανάπτυξη του παραδείγματος που παρουσιάσαμε στην προηγούμενη ενότητα θα χρησιμοποιήσουμε την ρουτίνα *gee* του πακέτου *GEE*. Η εν λόγω ρουτίνα προσφέρεται για δίτιμη μεταβλητή απόκρισης. Συνεπώς, προκειμένου να αναλυθεί η εφαρμογή μας με την *gee* πρέπει η απόκριση να μετασχηματιστεί σε δίτιμη. Για το λόγο αυτό, θα προσαρμόσουμε δύο μοντέλα λογιστικής παλινδρόμησης όπου θα καθορίζουμε κάθε φορά διαφορετική απόκριση.

Θεωρούμε τα δεδομένα όπως εμφανίζονται στο *MAREG*. Χωρίς βλάβη της γενικότητας στην μεταβλητή *TREAT* κωδικοποιούμε το *placebo* με την τιμή 0 (αντί για -1). Για το πρώτο μοντέλο ορίζουμε την μεταβλητή (*RESP*1)

$$Y_{i1t}^{*} = I(Y_{it} \le 1) = \begin{cases} 1, & an\ Y_{it} \le 1 \\ 0, & an\ Y_{it} > 1 \end{cases}.$$

Δηλαδή, ως "επιτυχία" θεωρούμε το ενδεχόμενο η απόκριση Y_{it} του i-υποκειμένου $(i = 1,2,...,111)$ κατά την t-χρονική στιγμή $(t = 1,2,3,4)$ να είναι ≤ 1 και ως "αποτυχία" το αντίθετο. Άρα θα έχουμε τα δεδομένα της μορφής

ID	RESP	RESP1	TREAT	OCCASION
0	1	1	1	1
0	1	1	1	2
0	1	1	1	3
0	1	1	1	4
1	1	1	0	1
1	1	1	0	2
1	1	1	0	3
1	1	1	0	4
M	M	M	M	M
109	3	0	0	1
109	3	0	0	2
109	3	0	0	3
109	3	0	0	4
110	3	0	0	1
110	3	0	0	2
110	3	0	0	3
110	3	0	0	4

Για την περιγραφή των δεδομένων θα εφαρμόσουμε ένα ανάλογο με το μοντέλο (2.1), το

$$\log it(P(Y^*_{i1t}=1)) = a_1 + b_1 TREAT_i + b_2 OCCASION + b_3 TREAT_i \times OCCASION\,, \qquad (5.3)$$

όπου με $TREAT_i$ συμβολίζουμε την θεραπεία που λαμβάνει ο *i*-ασθενής (με τιμές 1 για το ενεργό φάρμακο και 0 για το *placebo*) και με *OCCASION* την περίσταση κατά την οποία έγινε η μέτρηση με τιμές 1,2,3 και 4. Στο μοντέλο περιλαμβάνεται και όρος αλληλεπίδρασης.

Στο παράθυρο των *Commands* αλλάζουμε τα *contrasts* και στη συνέχεια καλούμε την βιβλιοθήκη *GEE* ενώ επισυνάπτουμε το αρχείο των δεδομένων

```
> options(contrasts=c("contr.treatment", "contr.poly"))
> library(gee)
> attach(efarmogi1) .
```

Σημειώνουμε ότι η μεταβλητή της θεραπείας πρέπει να δηλωθεί ως παράγοντας (*factor*). Για το λόγο αυτό πληκτρολογούμε

```
> as.factor(TREAT) .
```

Το επιθυμητό μοντέλο εφαρμόζεται μέσω της εντολής *gee*. Οπότε θα έχουμε

*fit.gee*1<-*gee*(*RESP*1~*TREAT***OCCASION, id=ID, family=binomia*l, +*corstr*="*exchangeabl*e",*data=efarmogi*1) .

Παρατηρούμε ότι η σύνταξή της παρουσιάζει ομοιότητες με αυτή της εντολής ***ordgee*** του ***R.*** Τα αποτελέσματα που παίρνουμε μετά την εντολή

> s*ummary*(*fit.gee*1)

παρουσιάζονται στον Πίνακα 3.1.

Το παρακάτω output περιλαμβάνει τις εκτιμήσεις των παραμέτρων, τα τυπικά σφάλματα αυτών και τις τιμές του στατιστικού τεστ *z* αν η δομή για τον πίνακα συσχέτισης που υποθέσαμε είναι η σωστή (***naive SE*** και ***naive z***). Επίσης, υπολογίζονται οι αντίστοιχες τιμές για τα τυπικά σφάλματα και το στατιστικό τεστ στην περίπτωση που η εν λόγω δομή είναι λανθασμένη (***robust SE*** και ***robus***t ***z***). Σύμφωνα με τις τιμές των τελευταίων τυπικών σφαλμάτων, η αλληλεπίδραση χρόνου και θεραπείας είναι μη στατιστικά σημαντική. Στο ίδιο συμπέρασμα καταλήγουμε και για την επίδραση της θεραπείας, καθώς η τιμή για το τυπικό σφάλμα είναι αρκετά μεγάλη. Αυτό σημαίνει ότι οι ομάδες της ενεργής αγωγής και του ***placebo*** δεν παρουσιάζουν σημαντικές διαφορές στην αρχή (***baseline***) της μελέτης. Αναφέρουμε ότι η παράμετρος κλίμακας (***Estimated Scale Parameter***) αντιστοιχεί στην παράμετρο f , την οποία συναντήσαμε στη θεωρία των *GEE*.

Coefficients:

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	-1.4519326	0.3665826	-3.9607239	0.3612614	-4.0190641
TREAT	-1.6202046	0.7323272	-2.2124053	0.7618256	-2.1267396
OCCASION	0.1485782	0.1037263	1.4324064	0.1101359	1.3490450
TREAT:OCCASION	0.1848570	0.1939309	0.9532104	0.2149849	0.8598602

Estimated Scale Parameter: 1.014745
Number of Iterations: 2

Working Correlation

	[,1]	[,2]	[,3]	[,4]
[1,]	1.0000000	0.4367172	0.4367172	0.4367172
[2,]	0.4367172	1.0000000	0.4367172	0.4367172
[3,]	0.4367172	0.4367172	1.0000000	0.4367172
[4,]	0.4367172	0.4367172	0.4367172	1.0000000

Πίνακας 3.1: Αποτελέσματα της ανάλυσης του μοντέλου (5.3) με το S-Plus

Για το δεύτερο μοντέλο εργαζόμαστε ανάλογα. Ως απόκριση ορίζουμε την μεταβλητή (*RESP*2)

$$Y_{i2t}^{*} = I(Y_{it} \le 2) = \begin{cases} 1, & an\ Y_{it} \le 2 \\ 0, & an\ Y_{it} > 2 \end{cases}.$$

Δηλαδή, ως "επιτυχία" θεωρούμε στην περίπτωση αυτή το ενδεχόμενο η απόκριση Y_{it} του *i*-υποκειμένου κατά την *t*-χρονική στιγμή να είναι ≤ 2 και ως "αποτυχία" το αντίθετο. Επομένως, το μοντέλο που θα εφαρμόσουμε θα είναι

$$\log it(P(Y_{i2t}^{*} = 1)) = a_2 + g_1 TREAT_i + g_2 OCCASION + g_3 TREAT_i \times OCCASION, \qquad (5.4)$$

και η εντολή που θα το εκτιμήσει θα είναι σε αναλογία με την προηγούμενη

*fit.gee*2<-*gee*(*RESP*2~*TREAT***OCCASION*, *id=ID*, *family=binomia*l, +*corstr*="*exchangeabl*e",*data=efarmogi*2).

Τα αποτελέσματα δίνονται στον Πίνακα 3.2.

Coefficients:

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	1.16839506	0.34601328	3.3767347	0.3438938	3.3975464
TREAT	-0.60581023	0.46622846	-1.2993849	0.4765568	-1.271223
OCCASION	-0.03700888	0.09617027	-0.3848267	0.0966784	-0.3828041
TREAT:OCCASION	-0.08382797	0.12973219	-0.6461617	0.1425794	-0.5879390

Estimated Scale Parameter: 1.008957
Number of Iterations: 2

Working Correlation

	[,1]	[,2]	[,3]	[,4]
[1,]	1.0000000	0.5051151	0.5051151	0.5051151
[2,]	0.5051151	1.0000000	0.5051151	0.5051151
[3,]	0.5051151	0.5051151	1.0000000	0.5051151
[4,]	0.5051151	0.5051151	0.5051151	1.0000000

Πίνακας 3.2: Αποτελέσματα της ανάλυσης του μοντέλου (5.4) με το S-Plus

Η ανάλυση αυτή δεν είναι η πλέον ενδεδειγμένη για διατάξιμες μεταβλητές απόκρισης και ο σκόπος της ήταν απλώς να δούμε την εφαρμογή της ρουτίνας ***gee***. Παρατηρούμε ότι μερικά από τα αποτελέσματα, όπως η μη σημαντικότητα της αλληλεπίδρασης και η επίδραση του χρόνου στο δεύτερο μοντέλο, συμφωνούν με αυτά που προέκυψαν από την ανάλυση μέσω του ***R*** και της εντολής ***ordgee***. Οι διαφορές ανάμεσα στις εκτιμήσεις για τις παραμέτρους της συσχέτισης και για τις τρεις αναλύσεις των δεδομένων των Koch et al., (1989) οφείλονται στο διαφορετικό θεωρητικό υπόβαθρο που έχουν.

Τα δεδομένα των Koch et al., (1989) αναλύθηκαν μέσω τριών διαφορετικών προγραμμάτων. Σε γενικές γραμμές, οι τρεις αναλύσεις μας έδωσαν ποιοτικά παρόμοια συμπεράσματα. Επιμέρους διαφορές μεταξύ αυτών οφείλονται στις επιλογές των μοντέλων που χρησιμοποιήθηκαν και στις διαφορετικές υποθέσεις που κάνει κάθε πρόγραμμα για τον " χρησιμοποιούμενο" πίνακα συσχετίσεων (" ***working***" ***correlation matrix***).

5.2 Ανάλυση με την χρήση μοντέλων τυχαίων επιδράσεων

5.2.1 Περιγραφή των δεδομένων

Πρόκειται για μια μελέτη που διεξήχθη στο NIMH Schizophrenia Collaborative Study και στην οποία έλαβαν μέρος 437 ασθενείς με ιστορικό σχιζοφρένιας. Με τυχαίο τρόπο καθένας από αυτούς έλαβε μία από τις τρεις φαρμακευτικές αγωγές: chlorpromazine, fluphenazine ή thioridazine ενώ υπάρχει και η ομάδα ***placebo***. Στην εν λόγω εφαρμογή θα θεωρήσουμε ότι υπάρχουν δύο μόνο ομάδες ασθενών. Η πρώτη θα περιλαμβάνει όσους έλαβαν το ***placebo*** και η άλλη όσους έλαβαν κάποια από τις ενεργές αγωγές.

Η μεταβλητή απόκρισης αντιστοιχεί στο Στοιχείο 79 (***Item*** 79) της Πολυδιάστατης Ψυχιατρικής κλίμακας Εσωτερικών ασθενών (Inpatient Multidimensional Psychiatric Scale, IMPS; Lorr and Klett, 1966) και είναι η σοβαρότητα της ασθένειας. Τα επίπεδά της διακρίνονται σε 1 = φυσιολογική, 2 = οριακή, 3 = ελαφρά, 4 = μέτρια, 5 = σημαντική, 6 = σοβαρή και 7 = πολύ σοβαρή κατάσταση. Πραγματοποιήθηκαν επτά μετρήσεις με αφετηρία την εβδομάδα 0 και πέρας την εβδομάδα 6. Στον παρακάτω πίνακα (Hedeker and Gibbons, 1997) δίνεται ο πειραματικός σχεδιασμός και τα μεγέθη του δείγματος ανά εβδομάδα

Ομάδα	Μέγεθος δείγματος ανά εβδομάδα						
	0	1	2	3	4	5	6
Placebo (n=108)	107	105	5	87	2	2	70
Ενεργό φάρμακο (n=329)	327	321	9	287	9	7	265
Ενεργό φάρμακο: chlorpromazine, fluphenazine ή thioridazine							

Οι περισσότερες μετρήσεις έχουν καταγραφεί κατά τις εβδομάδες 0,1,3 και 6 ενώ στις υπόλοιπες παρατηρείται σημαντικός αριθμός ελλιπών δεδομένων. Επιπλέον, ο αριθμός αυτών που δεν ολοκλήρωσαν τη μελέτη (***drop-out***), δηλαδή δεν έχει καταγραφεί η τελευταία μέτρηση για αυτούς, ανέρχεται στους 102 ασθενείς. Ειδικότερα, το ποσοστό αυτών που ολοκλήρωσαν τη μελέτη είναι 65% (70/108) και 81% (265/329) για την ομάδα ***placebo*** και αυτή του ενεργού φαρμάκου αντίστοιχα.

Το ερώτημα που πρέπει να απαντηθεί είναι αν στο πέρασμα του χρόνου οι ασθενείς που έλαβαν κάποια από τις φαρμακευτικές αγωγές παρουσίασαν σημαντική βελτίωση σε σχέση με εκείνους που έλαβαν το *placebo*.

5.2.2 Εφαρμογή με το MIXOR

Για την περιγραφή των δεδομένων θα προσαρμόσουμε ένα μοντέλο με εισαγωγή τυχαίων επιδράσεων και σε αυτό θα μας βοηθήσει το πρόγραμμα *MIXOR*.

Θα εφαρμόσουμε το μοντέλο των αναλογικών ***odds*** με την μορφή που το είδαμε στην ενότητα 3.3.2. Δηλαδή αν Y_{it} είναι η μέρηση για τον *i*-ασθενή κατά την *t*-χρονική στιγμή θα έχουμε

$$\text{logit}[P(Y_{it} \le k)] = a + g_k + b_1 X_{i1} + b_2 X_{i2} + b_3 X_{i1} X_{i2} + u_i \ ,$$

όπου $i = 1, 2, ..., 437$, $k = 1, 2, ..., 6$ και $t = 0, 1, 2, ..., 6$. Στη συγκεκριμένη παράσταση με X_1 συμβολίζουμε τη θεραπεία με τιμές 1 για το ενεργό φάρμακο και 0 για το placebo. Η επόμενη, η X_2 έχει να κάνει με το χρόνο (δηλαδή με τις εβδομάδες) κατά τον οποίο πραγματοποιήθηκαν οι μετρήσεις. Το μοντέλο που προσαρμόζει το *MIXOR* υποκινείται από την ιδέα μιας ***latent*** μεταβλητής για την οποία έγινε λόγος στην ενότητα 3.3.2. Έτσι, εφόσον η απόκριση έχει 7 επίπεδα υποτίθεται ότι υπάρχει μια σειρά ***threshold*** τιμών $g_1, g_2, ..., g_6$ με $g_0 = -\infty$ και $g_7 = +\infty$. Η απόκριση Y_{it} θα ανήκει στην κατηγορία *k*, αν η ***latent*** μεταβλητή, έστω *y*, υπερβαίνει την τιμή g_{k-1} αλλά όχι την τιμή g_k. Για την ταυτοποίηση, το πρώτο ***threshold*** τίθεται μηδέν ($g_1 = 0$).

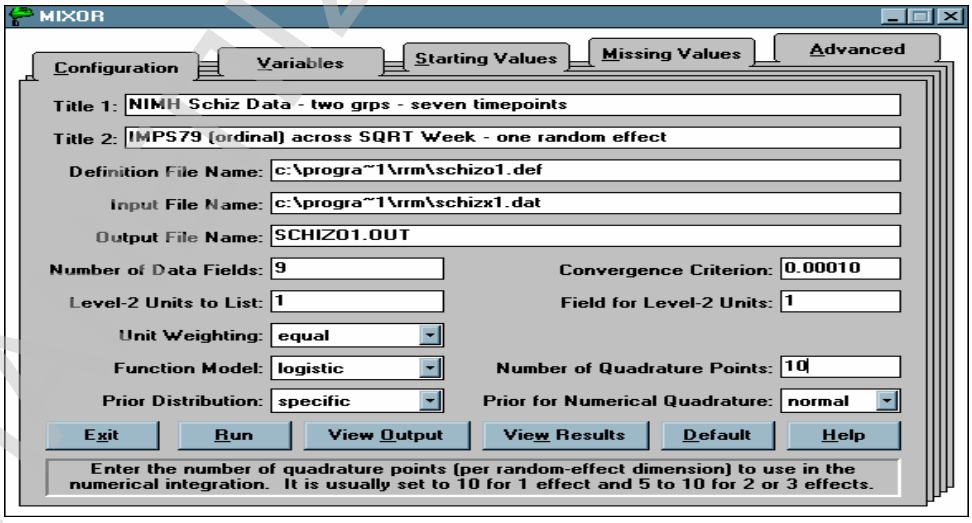
Στο σημείο αυτό προκύπτει το εξής πρόβλημα. Αν αναθέσουμε τα σκορ 0,1,…,6 στις εβδομάδες, παρατηρείται μη γραμμική σχέση ανάμεσα στο $\text{logit}[P(Y_{it} \le k)]$ και το χρόνο. Οι Hedeker and Gibbons (1997), που ασχολήθηκαν διεξοδικά με την ανάλυση των εν λόγω δεδομένων, πρότειναν ως μετρική του χρόνου τις τετραγωνικές ρίζες των αριθμών 0,1,…6 με αποτέλεσμα να προκύπτει, πλέον, μια λογική γραμμική σχέση. Στο μοντέλο έχουμε ακόμα εισάγει την αλληλεπίδραση θεραπείας και χρόνου ενώ η παράσταση ολοκληρώνεται με τις τυχαίες επιδράσεις. Αναφορικά με τις τελευταίες, υπάρχει ένας τέτοιος όρος για κάθε ασθενή

με σκοπό να δηλώνει την απόκλιση του από η γενική τάση που εμφανίζει η ομάδα στην οποία ανήκει (*placebo*/ενεργή αγωγή).

Είναι σημαντικό να έχουμε υπόψη ότι οι παράμεροι που εμφανίζονται στο μοντέλο είναι "*subject-specific*". Αυτό σημαίνει ότι η ερμηνεία τους γίνεται δοθείσης της τυχαίας επίδρασης του ασθενή. Για παράδειγμα, η παράμετρος b_1 της θεραπείας δηλώνει την αποτελεσματικότητα του ενεργού φαρμάκου έναντι στο placebo για τον υπο-πληθυσμό των ασθενών με το ίδιο επίπεδο της τυχαίας επίδρασης.

Τα δεδομένα του συγκεκριμένου προβλήματος βρίσκονται στο αρχείο *schizx.dat* στο φακέλο του προγράμματος. Στο αρχείο αυτό διακρίνει κανείς, μεταξύ άλλων, τον κωδικό (*ID*) για κάθε ασθενή, την μεταβλητή απόκρισης, την δείκτρια μεταβλητή για τη θεραπεία, την μεταβλητή για τον χρόνο με τιμές 0,1,…,6 και ακριβώς δίπλα τις τετραγωνικές ρίζες αυτών και την αλληλεπίδραση της θεραπείας με το χρόνο (με τη δεύτερη μετρική). Οι τιμές που λείπουν δηλώνονται με τον αριθμό -9. Παρά το γεγονός ότι η μεταβλητή απόκρισης είναι κατηγορική, σε ορισμένες περιπτώσεις έχουν καταγραφεί μετρήσεις με δεκαδικά ψηφία (π.χ. 5.3). Στην εφαρμογή μας θα χρησιμοποιήσουμε κλίμακα ακεραίων. Τέλος, στα δεδομένα περιλαμβάνεται και μία στήλη όπου οι κατηγορίες της απόκρισης έχουν μειωθεί σε 4 από 7, σύμφωνα με κάποια κωδικοποίηση. Για παράδειγμα, οι τρεις τελευταίες έχουν συμπτυχθεί σε μία και αντιστοιχούν στην τιμή 4.

Στη συνέχεια ακολουθεί η εκτίμηση του παραπάνω μοντέλου με τη χρήση του *MIXOR*. Ανοίγοντας το πρόγραμμα εμφανίζεται το παρακάτω παράθυρο



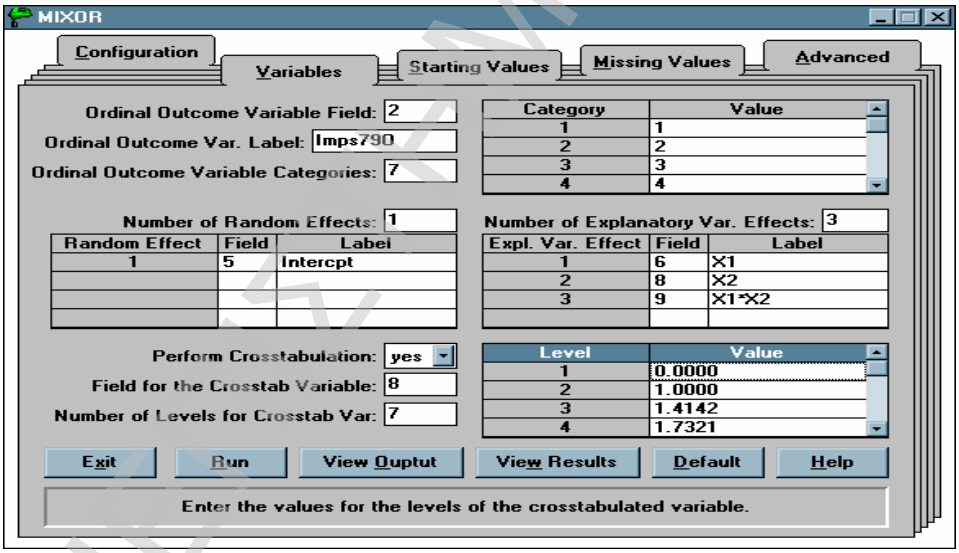
Εικόνα 1

Στις δύο πρώτες γραμμές καθορίζονται από το χρήστη αντίστοιχα τίτλος και υπότιτλος για την περιγραφή των δεδομένων και εμφανίζονται στη συνέχεια ως επικεφαλίδες στο *Output*. Στη θέση *Definition File Name* αναγράφεται το αρχείο που περιέχει τις ρυθμίσεις που κάνουμε κάθε φορά που τρέχουμε τα δεδομένα της επιλογής μας και στην ακόλουθη θέση αναγράφεται το αρχείο που θα διαβάσει το πρόγραμμα. Αν θέλουμε να διαβάσουμε κάποιο άλλο, κάνουμε διπλό αριστερό κλικ εντός των πεδίων. Σημειώνουμε ακόμα ότι όταν ο κέρσορας βρίσκεται εντός ενός πεδίου, στο τέλος της εικόνας παρέχεται βοήθεια για το πώς θα το συμπληρώσουμε.

Εν ολίγοις, στο συγκεκριμένο παράθυρο ορίζουμε:

- τον λογιστικό σύνδεσμο για την απόκριση (*Function Model*: *Logistic*)
- 10 *quadrature* σημεία για την αριθμητική ολοκλήρωση (*Number of Quadrature Points*)
- κανονική κατανομή για τις τυχαίες επιδράσεις (*Prior for Numerical Quadrature*: *normal*)

Στη συνέχεια, επιλέγοντας *Variables* στην αρχή του παραθύρου μεταφερόμαστε στην επόμενη καρτέλα.



Εικόνα 2

Ξεκινώντας από αριστερά κάνουμε τις ακόλουθες ρυθμίσεις:

- Στη θέση *Ordinal Outcome Variable Field* γράφουμε τον αριθμό 2, καθώς η μεταβλητή απόκρισης είναι στη δεύτερη στήλη του πίνακα των δεδομένων και στην επόμενη θέση

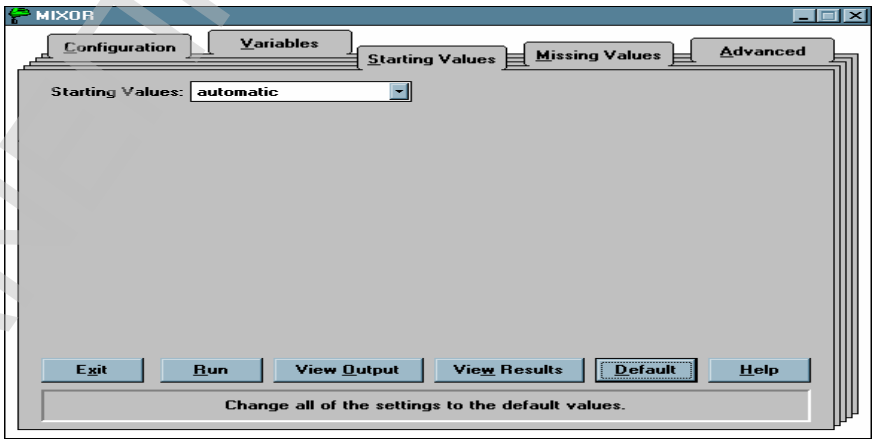
ορίζουμε κάποια ονομασία για αυτή (συγκεκριμένα *IMPS*790 από την κλίμακα που αναφέραμε στην προηγούμενη ενότητα).

- Στο επόμενο πεδίο, που αντιστοιχεί στον αριθμό των κατηγοριών της απόκρισης, συμπληρώνουμε τον αριθμό 7.
- Στη συνέχεια καθορίζουμε τον αριθμό των τυχαίων επιδράσεων. Στην περίπτωσή μας θα συμπληρώσουμε τον αριθμό 1. Επίσης, ορίζουμε τον αριθμό 5 στη θέση *Field*, διότι στον πίνακα των δεδομένων η 5η στήλη που αποτελείται από μονάδες αντιστοιχεί στους όρους των τυχαίων επιδράσεων.
- Αν θέλουμε να πάρουμε κάποιον πίνακα διασταύρωσης (crosstabulation) της εξαρτημένης μεταβλητής με μια ανεξάρτητη, επιλέγουμε *Perform Crosstabulation*: *yes* και στη συνέχεια καθορίζουμε σε ποια στήλη των δεδομένων βρίσκεται η ανεξάρτητη μεταβλητή και τον αριθμό των επιπέδων της. Για την εφαρμογή μας, έχει οριστεί η μεταβλητή του χρόνου με την μετρική των τετραγωνικών ριζών.

Στο δεξί μέρος του παραθύρου:

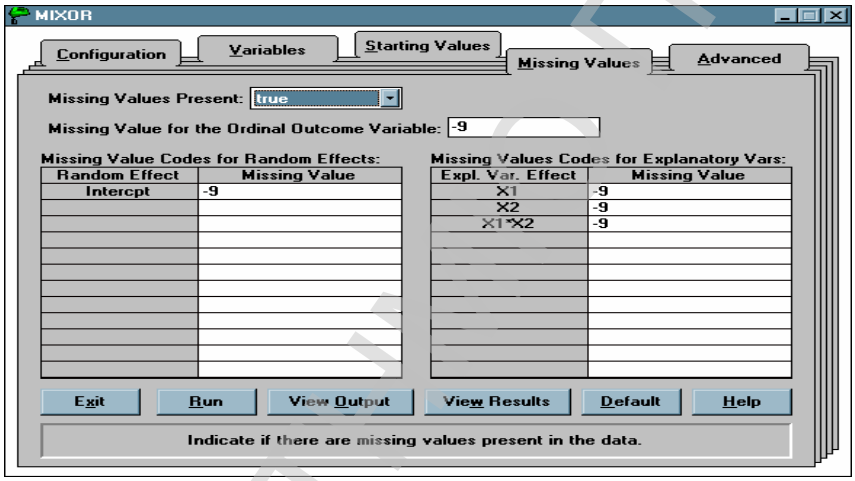
- Εμφανίζονται τα επίπεδα της απόκρισης
- Ο αριθμός των ερμηνευτικών μεταβλητών και οι ονομασίες αυτών, όπως θα εμφανίζονται στο *Output*. Στο μοντέλο που θα προσαρμόσουμε υπάρχουν τρεις και συμπληρώνουμε τους συμβολισμούς, όπως τους έχουμε αναφέρει στην παράσταση.
- Στο τελευταίο πεδίο καταγράφονται τα επίπεδα της ερμηνευτικής μεταβλητής που θα εμφανίζεται στον *Crosstabulation* πίνακα.

Στην καρτέλα *Starting Values* (Εικόνα 3) επιλέγουμε *automatic*

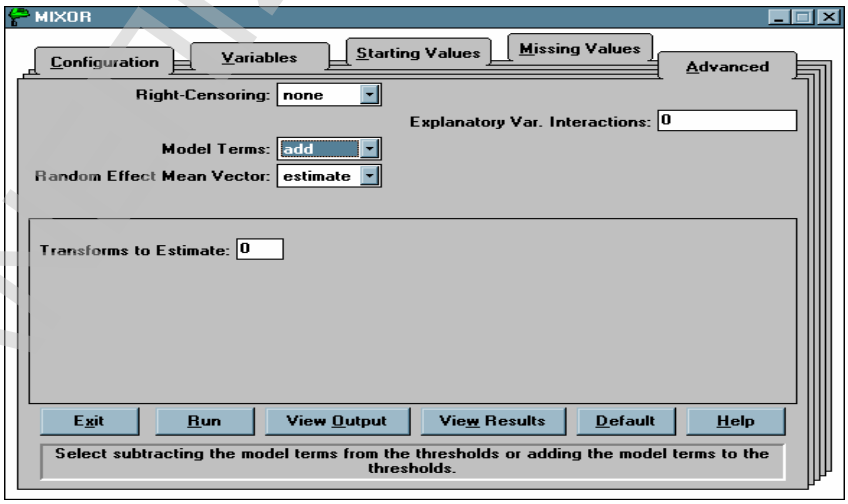


Εικόνα 3

και στην επόμενη (Εικόνα 4) ορίζουμε τον αριθμό -9 που εμφανίζεται στα δεδομένα και θα αναγνωρίζεται από το πρόγραμμα ως ελλιπή παρατήρηση, όπου και να συναντάται. Στην τελευταία καρτέλα ***Advanced*** (Εικόνα 5), μπορούμε να κάνουμε πιο προχωρημένες ρυθμίσεις, όπως δεξιά λογοκρισία ή αλληλεπίδραση ερμηνευτικών μεταβλητών με τις παραμέτρους a_k του μοντέλου. Η μόνη αλλαγή που θα κάνουμε στις ***default*** τιμές είναι στο πεδίο ***Model Terms***, όπου θα επιλέξουμε ***add***. Αυτό σημαίνει ότι οι όροι του μοντέλου θα προστίθενται στις παραμέτρους g_k, καθώς στη βιβλιογραφία συναντάται το μοντέλο των αναλογικών ***odds*** με τους όρους αυτούς να αφαιρούνται από αυτές.

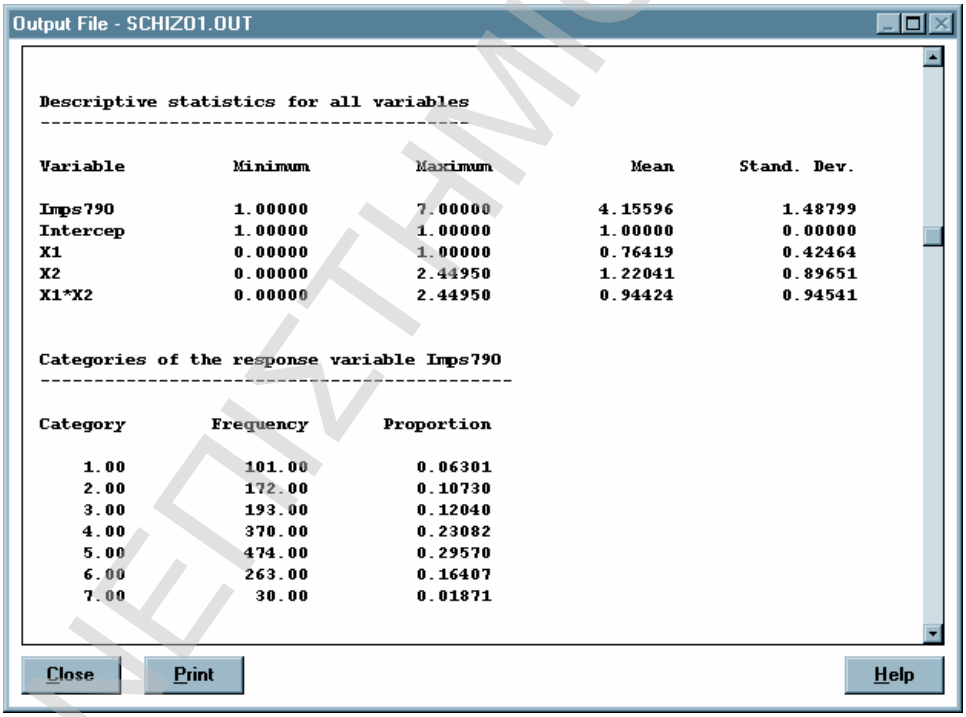


Εικόνα 4



Εικόνα 5

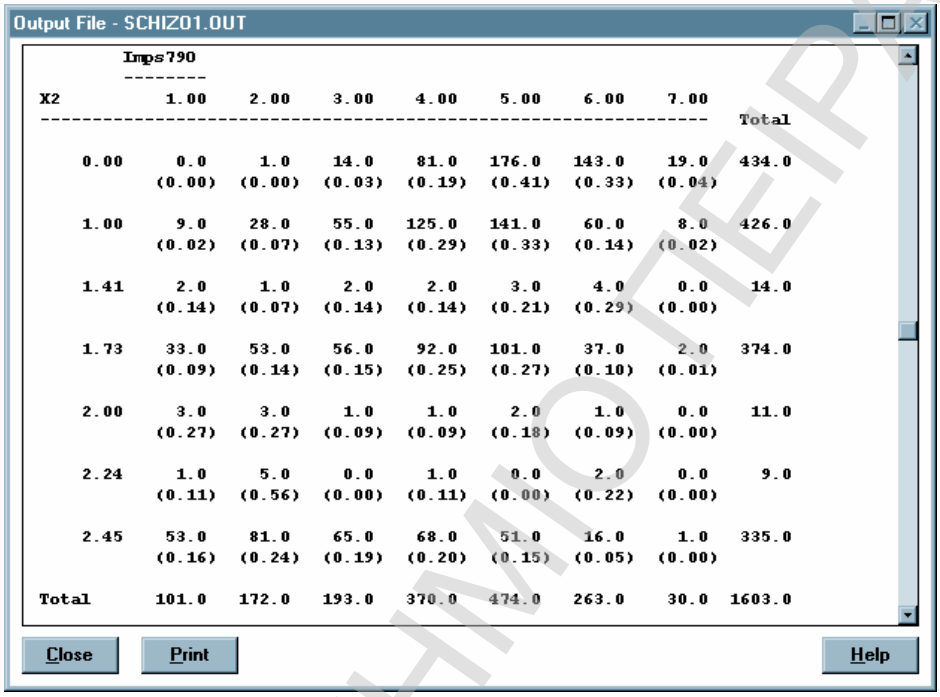
Επιλέγοντας ***Run*** παίρνουμε το output των αποτελεσμάτων, το οποίο θα παρουσιάσουμε και θα σχολιάσουμε τμηματικά. Στην Εικόνα 6 εμφανίζονται κάποια περιγραφικά στοιχεία (***Descriptive Statistics for all variables***) για την μεταβλητή απόκρισης, την θεραπεία την εβδομάδα αλλά και την αλληλεπίδραση των δύο τελεταίων. Πιο συγκεκριμένα δίνονται η ελάχιστη και μέγιστη τιμή αυτών, ο μέσος όρος και η τυπική απόκλισή τους. Στη συνέχεια παρουσιάζονται οι κατηγορίες της απόκρισης με τις συχνότητες που εμφανίζονται στα δεδομένα αλλά και το αντίστοιχο ποσοστό. Για παράδειγμα, καταγράφηκαν 101 παρατηρήσεις στην κατηγορία 1 της απόκρισης. Το σύνολο των παρατηρήσεων είναι 1603 (δεν υπολογίζονται οι ελλιπείς). Άρα το ποσοστό που αναλογεί σε αυτήν την κατηγορία είναι $\frac{101}{1603} \approx 0.063$. Παρατηρούμε ότι οι περισσότερες παρατηρήσεις συσσωρεύονται στις κατηγορίες 4 και 5 ενώ οι λιγότερες στην τελευταία.



Εικόνα 6

Στο επόμενο τμήμα του output (Εικόνα 7), εμφανίζεται ο πίνακας διασταύρωσης (***Crosstabulation***) της απόκρισης με τη μεταβλητή της εβδομάδας (X_2). Οι αριθμοί στις

παρανθέσεις αναφέρονται στα ποσοστά των συχνοτήτων επί του συνόλου των παρατηρήσεων για την κάθε εβδομάδα. Έτσι, από τους 434 ασθενείς που μετρήθηκαν στην αρχή της μελέτης (***baseline-Week*** 0), οι 176 από αυτούς, δηλαδή το 41%, ανήκουν στην κατηγορία 5 της απόκρισης.



Εικόνα 7

Στη συνέχεια (Εικόνα 8) εμφανίζονται οι αρχικές τιμές με τις οποίες αρχίζει ο αλγόριθμος για η εκτίμηση των παραμέτρων του μοντέλου και καθορίζονται από το πρόγραμμα (εφόσον έχουμε κάνει την αντίστοιχη επιλογή κατά τον καθορισμό του μοντέλου). Ακολούθως δίνονται ο αριθμός των επαναλήψεων, ο αριθμός των ***qudrature*** σημείων (του αλγόριθμου Gauss-Hermite, βλέπε ενότητα 3.3.3.1, σελίδα 43), ο λογάριθμος της πιθανοφάνειας (***Log Likelihood***) κατά τη σύγκλιση και το στατιστικό τεστ $-2\log l$ (***Deviance***). Επίσης δίνεται η τιμή ***ridge***. Αποτελεί μια ρύθμιση, η οποία γίνεται για τα διαγώνια στοιχεία του πίνακα πληροφορίας στην περίπτωση που το πρόγραμμα αντιμετωπίζει προβλήματα, όπως η αριθμητική ολοκλήρωση κατά τις επαναλήψεις. Η ρύθμιση αυτή συχνά βελτιώνει τις πιθανότητες σύγκλισης. Αρχικά ξεκινά με την τιμή μηδέν και κάθε φορά που παρουσιάζονται

δυσκολίες αυξάνεται κατά 0.1. Μετά τη σύγκλιση, ρυθμίζεται ξανά στην τιμή μηδέν προκειμένου να πάρουμε τα σωστά τυπικά σφάλματα για τις παραμέτρους του μοντέλου.

Στο επόμενο τμήμα του *output* (Εικόνα 9), δίνονται οι εκτιμήσεις των παραμέτρων του μοντέλου, τα τυπικά σφάλματα αυτών, το στατιστικό τεστ *z* αλλά και τα 2- *tailed p-values*. Έχουμε επισημάνει ότι στο μοντέλο των αναλογικών ***odds*** οι συντελεστές του μοντέλου δεν μεταβάλλονται κατά μήκος των κατηγοριών και για το λόγο αυτό δεν φέρουν τον δείκτη *k*. Κατά συνέπεια, η σχέση ανάμεσα στις ερμηνευτικές μεταβλητές και το αθροιστικό logit δεν εξαρτάται από την κατηγορία.

Ο συντελεστής b_1 της θεραπείας αντιπροσωπεύει την διαφορά στο ***logit*** κατά την εβδομάδα 0 (όπου $X_{i1} = 0$) ανάμεσα στους ασθενείς που παίρνουν κάποια από τις ενεργές αγωγές και αυτούς που παίρνουν ***placebo***. Ο αντίστοιχος συντελεστής b_2 για το χρόνο αφορά την εβδομαδιαία (σε μονάδες τετραγωνικών ριζών) μεταβολή στο ***logit*** για τους ασθενείς που λαμβάνουν το ***placebo***. Τέλος, ο συντελεστής b_3 τη αλληλεπίδρασης εκφράζει την διαφορά ανάμεσα στις δύο ομάδες θεραπείας στην εβδομαδιαία μεταβολή του ***logit***. Η τυχαία επίδραση u_i εκφράζει την απόκλιση του *i*-υποκειμένου από την τάση της ομάδας στην οποία ανήκει και εφόσον δεν εξαρτάται από τον χρόνο, η απόκλιση αυτή θεωρείται σταθερή κατά μήκος των εβδομάδων.

Από τα αποτελέσματα αποφαινόμαστε ότι οι ομάδες θεραπείας (ενεργό φάρμακο και ***placebo***) δεν διαφέρουν σημαντικά στην αρχή (***baseline***) της μελέτης καθώς η επίδραση της θεραπείας είναι μη σημαντική. Αν θεωρήσουμε το μοντέλο για την ομάδα του ***placebo***, παίρνει τη μορφή

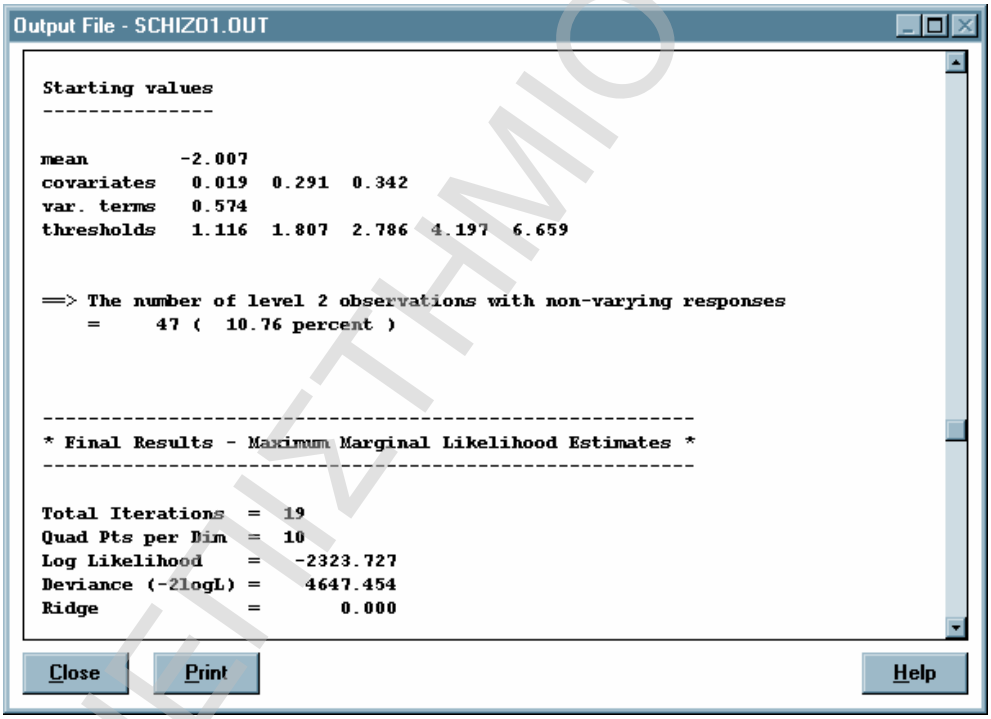
$$\text{logit}[P(Y_{it} \le k)] = a + g_k + b_2 X_{i2} + u_i .$$

Για ασθενείς στη συγκεκριμένη ομάδα και δεδομένου ότι ανήκουν στο ίδιο επίπεδο της τυχαίας επίδρασης, μπορούμε να συγκρίνουμε, για παράδειγμα, τις εβδομάδες 0 και 1. Οπότε ο εκτιμώμενος λόγος των ***odds*** $\dfrac{P(Y_{i1} \le k \mid X_{i2} = 1)}{1 - P(Y_{i1} \le k \mid X_{i2} = 1)}$ και $\dfrac{P(Y_{it} \le k \mid X_{i2} = 0)}{1 - P(Y_{it} \le k \mid X_{i2} = 0)}$ είναι ίσος με $\exp(\hat{b}_2) = \exp(0.739) \approx 2.09$ για κάθε κατηγορία *k*. Το θετικό πρόσημο της εκτίμησης του συντελεστή φανερώνει ότι η ομάδα του ***placebo*** βελτιώνεται με το πέρασμα του χρόνου.

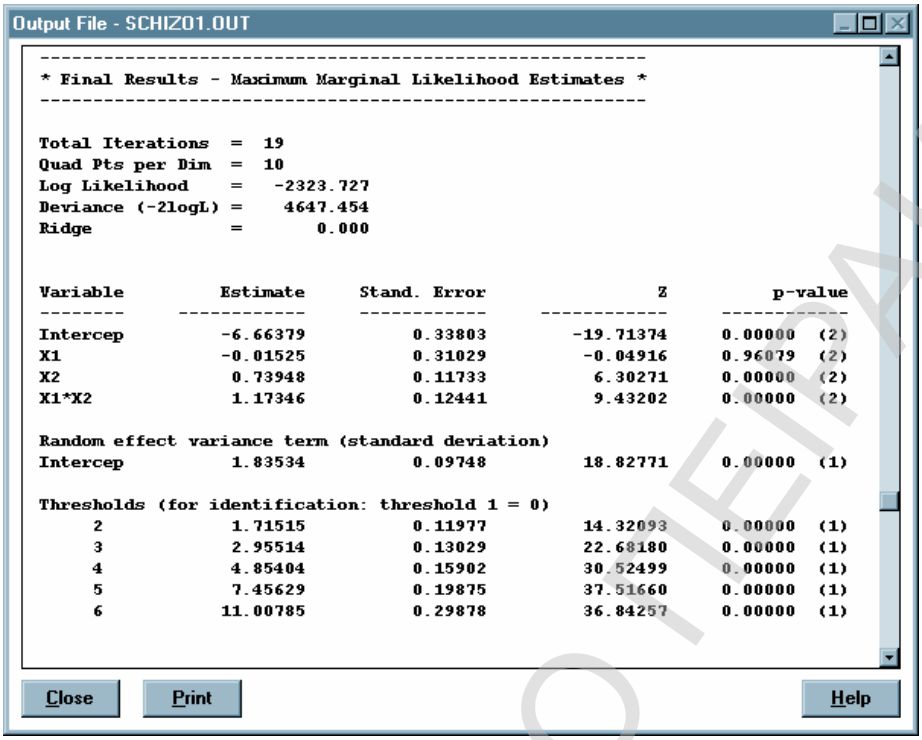
Αντίστοιχα, για τους ασθενείς που λαμβάνουν κάποια από τις ενεργές θεραπείες, το μοντέλο θα είναι της μορφής

$$\log it[P(Y_{it} \le k)] = a + g_k + b_1 + b_2 X_{i2} + b_3 X_{i2} + u_i \ .$$

Για την ομάδα του *placebo* κατά την *t*-εβδομάδα το εκτιμώμενο *odds* $\frac{P(Y_{it} \le k \mid X_{i2} = t)}{1 - P(Y_{it} \le k \mid X_{i2} = t)} = e^{\hat{a} + \hat{g}_k + \hat{b}_2 t}$ ενώ το ίδιο *odds* για την ομάδα των ενεργών φαρμάκων είναι ίσο με $e^{\hat{a} + \hat{g}_k + \hat{b}_1 + \hat{b}_2 + \hat{b}_3 t}$. Ο λόγος του δεύτερου *odds* προς το πρώτο είναι ίσος με $e^{\hat{b}_1 + \hat{b}_3 t}$. Η εκτίμηση για τον συντελεστή b_3 είναι θετική και κατά συνέπεια όσο ο χρόνος αυξάνεται τόσο η διαφορά ανάμεσα στα δύο *odds* θα μεγαλώνει. Καταλήγουμε, λοιπόν, ότι η κατάσταση των ασθενών που ανήκουν στην ομάδα των ενεργών αγωγών βελτιώνεται πιο γρήγορα σε σχέση με την άλλη.



Εικόνα 8

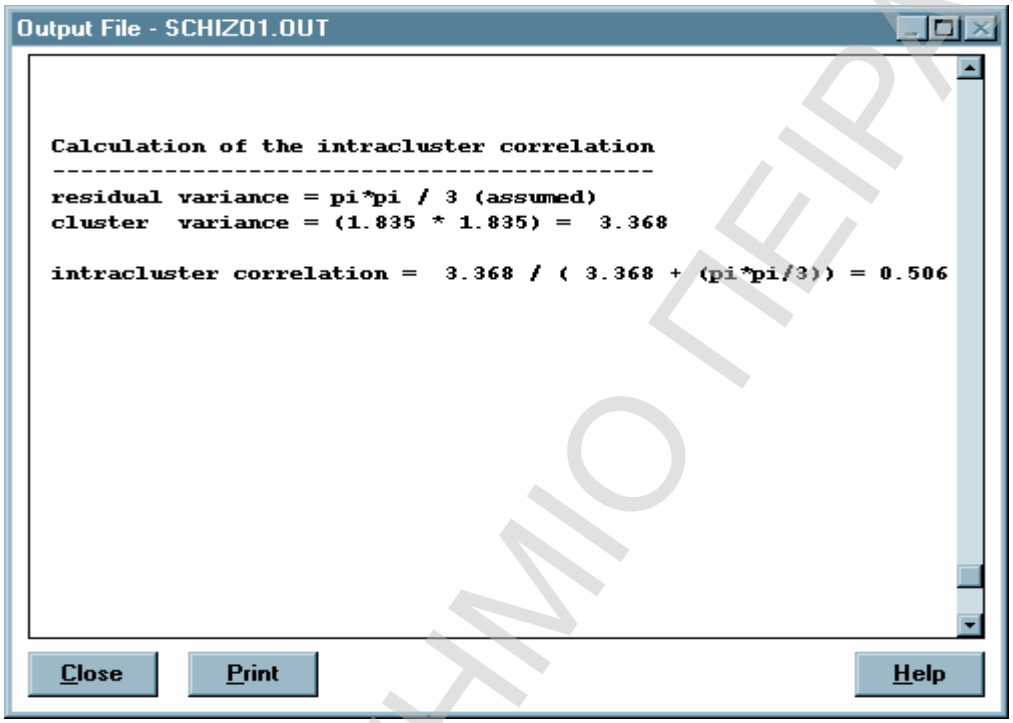


Εικόνα 9

Τέλος, το ***MIXOR*** δεν μας δίνει την εκτίμηση του πίνακα των διακυμάνσεων-συνδιακυμάνσεων για τις τυχαίες επιδράσεις αλλά τον παράγοντα Cholesky (δηλαδή την τετραγωνική ρίζα) αυτού. Στην περίπτωσή μας, όπου έχουμε μόνο έναν όρο τυχαίας επίδρασεις, ο παράγοντας αυτός αντιστοιχεί στην τετραγωνική ρίζα της διακύμανσης, δηλαδή την τυπική απόκλιση.

Στο τελευταίο τμήμα του *output* (Εικόνα 10), μας δίνονται πληροφορίες για την συσχέτιση μεταξύ των παρατηρήσεων για κάθε ασθενή. Αναφέρεται ως ***intraclass correlation***. Το συγκεκριμένο μέγεθος σχετίζεται με την διακύμανση των εν λόγω παρατηρήσεων (***cluster variance***) και δηλώνει το ποσοστό της διακύμανσης στο επίπεδο των ατόμων που παραμένει ανεξήγητο. Με άλλα λόγια απεικονίζει το μέγεθος της μεταξύ των ατόμων (***between-subjects***) διακύμανσης. Για το λογιστικό μοντέλο, όπου υποθέτουμε κανονικότητα των τυχαίων επιδράσεων, η ποσότητα αυτή ορίζεται ως $\frac{s_u^2}{s_u^2 + p^2/3}$, όπου s_u^2 είναι η διακύμανση της τυχαίας επίδρασης και υπολογίζεται ως το τετράγωνο της τυπικής απόκλισης που είδαμε στο προηγούμενο τμήμα. Ο δεύτερος όρος στον παρονομαστή του κλάσματος αντιπροσωπεύει

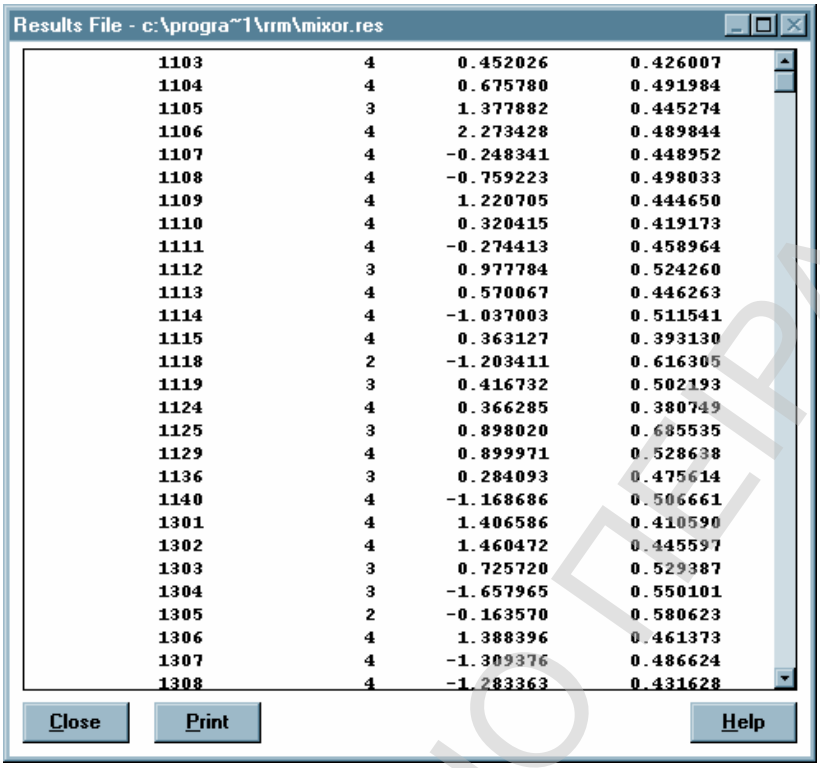
την διακύμανση της *latent* μεταβλητής, η οποία για το λογιστικό μοντέλο υποτίθεται ότι ακολουθεί τυπική λογιστική κατανομή (*standard logistic distribution*). Η εκτίμηση της συσχέτισης αυτής είναι 0.506.



Εικόνα 10

Τέλος, αν στο παράθυρο του *MIXOR* και έπειτα από τη σύγκλιση του αλγορίθμου επιλέξει κανείς *View Results*, εμφανίζεται το παράθυρο των εκτιμήσεων των επιδράσεων. Ταυτόχρονα, οι εκτιμήσεις αυτές αποθηκεύονται και στο αρχείο *MIXOR.RES* που βρίσκεται εντός του φακέλου μετά το πέρας του αλγορίθμου. Για μοντέλα με μία τυχαία επίδραση, το πρόγραμμα παράγει εμπειρικούς Bayes εκτιμητές για την τυχαία επίδραση κάθε υποκειμένου. Για το παράδειγμά μας το παράθυρο αυτό δίνεται στην Εικόνα 11.

Επίσης, μέσα στον ίδιο φάκελο δημιουργούνται δύο επιπλέον αρχεία, το *MIXOR.EST* και το *MIXOR.VAR*. Το πρώτο περιέχει τις εκτιμήσεις όλων των παραμέτρων και το δεύτερο περιέχει τον ασυμπτωτικό πίνακα διακυμάνσεων-συνδιακυμάνσεων αυτών.



Εικόνα 11: Παράθυρο των εκτιμήσεων των επιδράσεων του MIXOR

Οι στήλες με τη σειρά που τις βλέπουμε καταγράφουν:

1) τον κωδικό *ID* για κάθε ασθενή, 2) το πλήθος των μετρήσεων για τον κάθε ένα (χωρίς να προσμετρώνται οι ελλιπείς), 3) τους εμπειρικούς Bayes εκτιμητές (δηλαδή ο μέσος της αντίστοιχης *posterior* κατανομής) και 4) την ακρίβεια που συνδέεται με τους εν λόγω εκτιμητές (δηλαδή η τυπική απόκλιση της *posterior* κατανομής).

Βιβλιογραφία

Agresti, A. (1989). *A survey of models for repeated orderer categorical response data*. Statistics in Medicine, **8**: 1209-1224.

Agresti, A. (2002). *Categorical data analysis.* John Wiley and Sons, New York.

Agresti, A., Booth, J.G., Hobert, J.P and Caffo, B. (2000). *Random-effects modelling of categorical response data*. Sociological Methodology, **30**: 27-80.

Agresti, A. and Lang, J.B. (1993). *A proportional odds model with subject-specific effects for repeated ordered categorical responses*. Biometrika, **80**: 527-534.

Agresti, A. and Liu, I. (2005). *The analysis of ordered categorical data: an overview and a survey of recent developments.* Sociedad de Estadistica e Investigacion Operativa, **14**(1): 1-73.

Agresti, A. and Natarajan, R. (2001). *Modeling clustered ordered categorical data: A survey.* International Statistical Review, **69**: 345-371.

Albert, P.S. and McShane, L.M. (1995). *A generalized estimating equations approach for spatially correlated binary data: Applications to the analysis of neuroimaging data*. Biometrics, **56**: 627-638.

Anderson, J.A. and Philips, P.R. (1981). *Regression, discrimination and measurement models for ordered categorical variables*. Applied Statistics, **30**: 22-31.

Barnhart, H. X. and Williamson, J. (1998). *Goodness-of-fit tests for GEE modelling. Biometrics*, **54**(2): 720–729.

Bergsma, W. P. (1997). *Marginal Models for Categorical Data*. Tilburg Univ. Press.

Bock, R.D. (1975). *Multivariate statistical methods in behavioral research.* McGraw-Hill, New York.

Booth, J.G. and Hobert, J.P. (1998). *Standard errors of prediction in Generalized linear mixed models*. Journal of the American Statistical Association, **93**: 262-272.

Booth, J.G. and Hobert, J.P. (1999). *Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm.* Journal of the Royal Statistical Society, Series B, **61**: 265-285.

Breslow, N. and Clayton, D.G. (1993). *Approximate inference in generalized linear mixed models*. Journal of the American Statistical Association, **88**: 9-25.

Bryk, A.S. and Raudenbush, S.W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Sage Publications, Inc., Newbury Park, CA.

Cameron, A.C. and Trivedi, P.K. (1998). *Regression analysis of count data.* Cambridge University Press, New York.

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

Carlin, B.P. and Louis, T.A. (1996). *Bayes and empirical Bayes methods for data analysis*. Chapman and Hall, London.

Clayton, D. (1992). *Repeated ordinal measurements: a generalised estimating equation approach.* MRC Biostatistics Unit, Cambridge.

Cox, C. (1995). *Location-scale cumulative odds models for ordinal data: A generalized non-linear model approach* . Statistics in Medicine, **14**: 1191-1203.

Crouchley, R. (1995*) A random-effects model for ordered categorical data*. Journal of the American Statistical Association, **90**: 489-498.

Crowder, M.J. and Hand, D.J. (1990). *Analysis of repeated measures*. Chapman and Hall, London.

Davis, C. (2002) *Statistical methods for the analysis of repeated measurements.* Springer, New York.

Davis, C.S. and Hall, D.B. (1999). *A computer program for regression analysis of ordered categorical repeated measurements.* Computer Methods and Programs in Biomedicine, **51**: 153-169.

Dempster, A.P, Laird, N.M and Rubin, D.B. (1977) *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the American Statistical Association, **81**: 709-721.

Diggle, P.J., Liang, K.Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data.* Oxford University Press, Oxford.

Ekholm, A., Smith, P.W.F. and McDonald, J.W. (1995). *Marginal regression analysis of a multivariate binary response*. Biometrika, **82**: 847-854.

Ekholm, A., Jokinen J., McDonald, J.W., and Smith, P.W.F. (2003). *Joint regression and association modeling of longitudinal ordinal data.* Biometrics, **59**: 795-803.

Firth, D. (1993b). *Recent developments in quasi-likelihood methods* . Proc. ISI 49th Session, 341-358.

Fitzmaurice, G.M. (1995). *A caveat concerning independence estimating equations with multivariate binary data*. Biometrics, **51**: 309–317.

Follmann, D.A. and Lambert, D. (1989). *Generalizing logistic regression by non-parametric mixing*. Journal of the American Statistical Association, **84**: 295-300.

Girden, E.R. (1992). *ANOVA: Repeated Measures*. Sage Publications, Newbury Park, CA.

Goldstein, H. (1995). *Multilevel statistical models, 2nd edition*. Halstead Press, New York.

Graubard, B.I. and Korn, E.L. (1994). *Regression analysis with clustered data*. Statistics in Medicine, **13**: 509-522.

Haber, M. (1985). *Maximum likelihood methods for linear and log-linear models in categorical data* . Computational. Statistics in Data Analysis, **3**: 1-10.

Hagenaars, J.A. (1990). *Categorical Longitudinal Data:Log-linear, Panel, Trend and CohortAnalysis.* Sage Publications, Newbury Park, CA.

Hand, D.J. and Crowder, M.J. (1996). Practical longitudinal data analysis. Chapman and Hall, London.

Hardin, J.W. and Hilbe, J. M. (2003). *Generalized estimating equations*. Chapman and Hall, New York.

Harville, D.A. and Mee, R.W. (1984). *A mixed-model procedure for analyzing ordered categorical data*. Biometrics, **40**: 393-408.

Heagerty, P.J. and Zeger, S.L. (1996). *Marginal regression models for clustered ordinal measurements*. Journal of the American Statistical Association, **91**: 1024-1036.

Hedeker, D. and Gibbons, R.D. (1994). *A random-effects ordinal regression model for multilevel analysis*. Biometrics, **50**: 933–944.

Hedeker, D. and Gibbons, R.D. (1996). MIXOR: *A computer program for mixed-effects ordinal regression analysis.* Computer Methods and Programs in Biomedicine, **49**: 157-176.

Hedeker, D. and Gibbons, R.D. (1997). *Application of random-effects pattern-mixture models for missing data in longitudinal studies.* Psychological Methods, **2**(1): 64-78.

Hedeker, D. and Mermelstein, R.J. (2000). *Analysis of longitudinal substance use outcomes using ordinal random-effects regression models*. Addiction, **95**: 381–394.

Horton, N.J., Bebchuk, J.D., Jones, C.L., Lipsitz, S.R., Catalano, P.J., Zahner, G. E. P. and Fitzmaurice, G.M. (1999). *Goodness-of-fit for GEE: an example with mental health service utilization.* Statistics in Medicine, **8**: 213-222.

Horton, N.J. and Lipsitz S.R. (1999). *Review of software to fit generalized estimating equation regression models.* The American Statistician, **53** (2): 160-169.

Hosmer, D.W., Hosmer, T., le Cessie, S. & Lemeshow, S. (1997). *A comparison of goodness-of-fit tests for the logistic regression model*. Statistics in Medicine, **16**: 965-980.
Hosmer, D. W. and Lemeshow, S. (1980). *A goodness-of-fit test for the multiple logistic regression model*. Comm. Statist. Theory Methods, **10**: 1043-1069.

Hosmer, D.W. and Lemeshow, S. (2000). *Applied logistic regression, 2nd edition*. John Wiley and Sons, New York.

Huang, G.-H., Bandeen-Roche, K. and Rubin, G.S. (2002). *Building marginal models for multiple ordinal measurements*. Journal of the Royal Statistical Society, Series C: Applied Statistics, **51**: 37-57.

Jansen, J. (1990). *On the statistical analysis of ordinal data when extravariation is present*, Applied Statistics, **39**: 75–84.

Kastner, C., Fieger, A. and Heumann C. (1997). *MAREG and WinMAREG: A tool for marginal regression models.* Computational Statistics in Data Analysis, **24**: 237-241.

Kenward, M.G., Lesaffre, E. and Molenbergs, G. (1994). *An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random.* Biometrics, **50**: 945-953.

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

Koch, G.G., Carr, G.J., Amara, I.A., Stokes, M.E. and Uryniak, T.J. (1989). *Categorical data analysis.* Statistical Methodology in Pharmaceutical Sciences, D.A. Berry (ed), 391-475. New York: Marcel Dekker.

Kosorok, M. R. and Chao, W.H. (1996). *The analysis of longitudinal ordinal response data in continuous time.* Journal of the American Statistical Association, **91**: 807-817.

Lang, J.B. (1996a). *Maximum likelihood methods for a generalized class of log-linear models* Annals of Statistics, **24**: 726-752 .

Lang, J.B. and Agresti A. (1994). *Simultaneously modeling joint and marginal distributions of multivariate categorical responses.* Journal of the American Statistical Association, **89**: 625-632.

Lang, J.B. (2004). *Multinomial-Poisson homogeneous models for contingency tables.* Annals of Statistics, **32**: 340-383.

Liang, K.Y. and Zeger, S.L. (1986). *Longitudinal data analysis using generalized linear models.* Biometrica, **73**: 13-26.

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**(1): 153-160.

Little, R.J. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. John Wiley and Sons, New York.

Longford, N.T. (1993). *Random coe.cient models.* Oxford University Press, New York.

Lorr, M. and Klett, C.J. (1966). *Inpatient multidimensional psychiatric scale: Manual*. Consulting Psychologists Press, Palo Alto, CA.

Mark, S.D. and Gail, M.H. (1994). *A comparison of likelihood-based and marginal estimating equation methods for analysing repeated ordered categorical responses with missing data.* Statistics in Medicine, **13**: 479-493.

McCullagh, P. (1980). *Regression models for ordinal data.* Journal of the Royal Statistical Society, Series B, **42**: 109-142.

McCullagh, P. (1983). *Quasi-likelihood functions.* Annals of Statistics, **11**: 59-67.

McCullagh, P. and Nelder, J.A. (1989).*Generalized Linear Models*. Chapman and Hill, London.

McCulloch, C.E. (1997). *Maximum likelihood algorithms for generalized linear mixed models.* Journal of the American Statistical Association, **92**: 162-170.

Miller, J. (1977). *Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance*. The Annals of Statistics, **5**: 746-762.

Miller, M.E., Davis, C.S. and Landis, J.R. (1993). *The analysis of longitudinal polytomous data: Generalized estimating equations and connections with weighted least squares.* Biometrics, **49**: 1033-1044.

Neuhaus, J.M., Kalbfleisch, J.D. and Hauck, W.W. (1661).*A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data.* International Statistical Review, **59**: 25-35.

Neuhaus, J.M. (1992). *Statistical methods for longitudinal and clustered designs with binary responses.* Statistical Methods in Medical Research, **1**: 249-273.

Osius, G. and Rojek, D. (1992). *Normal goodness-of-fit tests for multinomial models with large degrees of freedom*. Journal of the American Statistical Association, **87**: 1145-1152.

Pan, W. (2002). *Goodness-of-fit tests for GEE with correlated binary data*. Scandinavian Journal of Statistics, **29**: 101-110.

Pendergast, J.F., Gange, S.J., Newton, M.A., Lindstrom, M.J., Palta, M. and Fisher, M.R.

(1996). *A survey of methods for analyzing clustered binary response data*. International Statistical Review, **64**: 89–118.

Pepe, M.S. and Anderson, G.L. (1994). *A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data.* Communications in Statistics-Simulation and Computation, **23**: 939-951.

Peterson, B. and Harrell, F.E. (1990). *Partial proportional odds models for ordinal response variables.* Applied Statistics, **39**: 205-217.

Pickles, A. (1990). *Longitudinal data and the analysis of change*. Oxford University Press, New York.

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

Randall, J.H. (1989). *The analysis of sensory data by generalized linear models*, Biometric Journal, **31**: 781-793.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1995). *Analysis of semi-parametric regression models for repeated outcomes in the presence of missing data*. Journal of the American Statistical Association, **90**: 106-121.

Royall, R.M. (1986) *Model robust confidence intervals using maximum likelihood estimators.* International Statistical Review, **54**: 221-226.

Searle, S.R., Casella, G. and McCulloch, C.E. (1992). *Variance Components*. John Wiley and Sons, New York.

Self, S.G. and Liang, K.Y. (1987). *Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions.* Journal of the American Statistical Association, **82**: 605-610.

Smith, D.M., Robertson,W.H. and Diggle, P.J. (1996). *Oswald: Object-Oriented Software for the Analysis of Longitudinal Data in S*, Technical ReportMA96/192, Department of Mathematics and Statistics, University of Lancaster, LA1 4YF, United Kingdom.

Stram, D.O., Wei, L.J. and Ware, J.H. (1988) *Analysis of repeated categorical outcomes with possibly missing observations and time-dependent covariates*. Journal of the American Statistical Association, **83**: 631-637.

Thompson L.(2005). *S-PLUS (and R) Manual to Accompany Agresti's Categorical Data Analysis* (2002), 2nd edition.

Toledano, A.Y. and Gatsonis, C. (1999). *Generalized estimating equations for ordinal categorical data: arbitrary patterns of missing responses and missingness in a key covariate*. Biometrics, **55**: 488-496.

Tsiatis, A. A. (1980). *A note on a goodness-of-®t test for the logistic regression model.* Biometrika, **67**: 250-251.

Tutz, G. and Hennewogl, W. (1996). *Random effects in ordinal regression models*. Computational Statistics and Data Analysis, **22**: 537-557.

Wedderburn, R.W. (1974). *Quasi-likelihood functions , generalized linear models and the Gauss-Newton method.* Biometrica, **61**: 439-447.

Wei, G. and Tanner, M. (1990). *A Monte-Carlo implementation of the EM algorithm and the Poor Man's data augmentation algorithms*. Journal of the American Statistical Association, **85**: 699-704.

Wolfinger, R. and O' Connell, M. (1993). *Generalized linear mixed models: A pseudo-likelihood approach*. Journal of Statistical Computational Simulation, **48**: 233-243.

Zeger, S.L., Liang, K.Y. and Albert, P.S. (1988). *Models for longitudinal data: a generalized estimating equation approach*. Biometrics, **44**: 1049-1060.