ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΔΙΑΤΑΚΤΙΚΑ PROBIT ΚΑΙ LOGIT ΜΟΝΤΕΛΑ

Θεοφάνης Θουκυδίδης Τσισμεντζόγλου

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

Σεπτέμβριος 2017

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN

APPLIED STATISTICS

ORDERED PROBIT AND LOGIT MODELS

Theofanis Thoukydidis Tsismentzoglou

MSc Dissertation

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

Applied Statistics

Piraeus, Greece

September 2017

Ευχαριστίες

Θα ήθελα να ευχαριστήσω όλους όσοι βοήθησαν στην πορεία μου μέχρι τώρα να μαθαίνω να ζω καλύτερα. Στους φίλους για τη χαρά που μου προσφέρουν, στους καθηγητές μου για τη γνώση και την παρότρυνση και στην οικογένεια μου για τη στήριξη και την αγάπη της.

Περίληψη

Σε πολλές κλινικές μελέτες τα αποτελέσματα ενός φαρμάκου δεν είναι ίαση ή όχι αλλά και διάφορα άλλα ενδιάμεσα στάδια τα οποία καταγράφονται με μια διατακτική κλίμακα. Τέτοια δεδομένα μελετώνται με την βοήθεια των ordered logit and probit γενικευμένων γραμμικών μοντέλων. Τα μοντέλα αυτά είναι γενικεύσεις των γνωστών logit και probit μοντέλων για τα οποία ομως η εξαρτημένη μεταβλητή είναι οποιαδήποτε μεταβλητή με διάταξη. Στην εργασία αυτή δίδεται το θεωρητικό υπόβαθρο των μοντέλων αυτών και έμφαση στην εφαρμογή τους σε πραγματικά δεδομένα.
Keywords: διατεταγμενα, logit, probit, γενικευμένα γραμμικά μοντέλα

Abstract

In many clinical trials the outcomes of a prescribed drug are not limited to curing or not the underlying condition. Instead they include other different intermediate stages that are documented by an ordinal scale. To model and study data such as these, one can use ordered logit and probit generalized linear models. These models are generalizations of known logit and probit models, but for which the dependent variable is any variable that is ordered. This thesis presents a general theoretical background given of these models and also there is given attention in real data application.

Keywords: ordered, logit, probit, glm, generalized linear models

Περιεχόμενα

Κατάλογος σχημάτων

R output

Κεφάλαιο 1

Εισαγωγή

Στο κεφάλαιο αυτό θα εισάγουμε την έννοια της διατακτικής ανάλυσης δεδομένων, θα αναφέρουμε τη χρησιμότητα της και θα κάνουμε μια σύντομη βιβλιογραφική ανασκόπηση επιστημονικών εργασιών που τη θεμελίωσαν. Ακόμα, θα αναφέρουμε μερικές διαφορές στη μοντελοποίηση διατακτικών και μη διατακτικών μοντέλων παλινδρόμησης που διαμορφώνουν την προτιμησή μας για αυτά και τέλος θα παραθέσουμε επιλεγμένες περιπτώσεις από τη βιβλιογραφία που η χρήση των διατακτικών μοντέλων είχε σημαντικό όφελος για την στατιστική ανάλυση των ερευνητών.

1.1 Κλίμακες μέτρησης μεταβλητών

Στη μελέτη της εξάρτησης μεταξύ μιας μεταβλητής απόκρισης και μιας ή και παραπάνω ανεξάρτητων μεταβλητών, η επιλογή ενός μοντέλου για την περιγραφή της σχέσης αυτής καθορίζεται σε μεγάλο βαθμό από το είδος της μεταβλητής απόκρισης. Αυτή μπορεί να είναι ένα από τα παρακάτω (Greenland, 1985) :

- Ονομαστική (ή αλλιώς ποιοτική) και η οποία δεν περιλαμβάνει διάταξη της μεταβλητής απόκρισης (π.χ αιτία θανάτου, ύπαρξη ασθένειας κ.α)

- Διατακτική, η οποία περιλαμβάνει μόνο τη διάταξη της μεταβλητής απόκρισης χωρίς όμως να προσδιορίζεται η απόσταση μεταξύ των τιμών που παίρνει η μεταβλητή. (π.χ κατηγορία κλινικής διάγνωσης: ακίνδυνη διάγνωση, ήπιος κίνδυνος, σοβαρός κίνδυνος, θανάσιμος κίνδυνος)

- Μεταβλητή διαστήματος (*interval scale*) (συμπεριλαμβανομένων και συνεχών αποκρίσεων), όπου περιλαμβάνεται τόσο η διάταξη των τιμών της μεταβλητής απόκρισης

όσο και η απόσταση μεταξύ αυτών των τιμών. (π.χ κλίμακα μέτρησης συστολικής πίεσης στις μονάδες mm Hg)

Στις φυσικές επιστήμες μεγάλο ποσοστό των δεδομένων έχει ποσοτικό χαρακτήρα, μετρημένα ωστόσο πολλές φορές σε αυθαίρετη κλίμακα. Στις κοινωνικές και σε ένα βαθμό και στις βιολογικές επιστήμες πιο συχνά συναντώνται δεδομένα ποιοτικού χαρακτήρα. Αυτά συνήθως παίρνουν τιμές σε περιορισμένο αριθμό είτε διατακτικών, είτε ονομαστικών κατηγοριών (Stevens, 1946).

Πολλές φορές, ο διαχωρισμός μεταξύ ονομαστικών και διατακτικών δεν είναι ξεκάθαρος. Για παράδειγμα, η κλίμακα των κατηγοριών στην αντίληψη για την ποιότητα φαγητού (εξαιρετική, καλή,..., κακή, φρικτή) είναι ξεκάθαρα διατακτική. Η κλίμακα των προτιμήσεων για τα ραδιοτηλεοπτικά προγράμματα μπορεί να αντιμετωπιστεί αρχικά ως ονομαστική, όμως αυτό μπορεί να μην ισχύει για όλες τις αναλύσεις όπως επίσης για τις πολιτικές προτιμήσεις και την αντίληψη περί ποιότητας. Η κλίμακα που περιγράφει το χρώμα των ματιών και μαλλιών ως φωτεινό-σκούρο μπορεί να διαταχθεί στην κλίμακα του γκρι, ωστόσο η συνάφεια της διάταξης μπορεί να εξαρτάται από το πλαίσιο στο οποίο εξετάζεται και εάν δεν υπάρχει ξεκάθαρη σύνδεση με το ηλεκτρομαγνητικό φάσμα ή κλίμακα του γκρι τα χρώματα θεωρείται ότι περιγράφονται από μια ονομαστική κλίμακα.

Στην εργασία αυτή μας ενδιαφέρει η μελέτη των διατακτικών μεταβλητών. Οι διατακτικές μετρήσεις περιγράφουν τη διάταξη των δεδομένων, αλλά όχι όμως και σχέσεις μεταξύ των κατηγοριών ή όπως αναφέραμε και παραπάνω την απόσταση μεταξύ των κατηγοριών. Οι διατακτικές κατηγορικές μεταβλητές επίσης, είναι πολύ χρήσιμες σε πολλά επιστημονικά πεδία όπου η ακριβής μέτρηση των δεδομένων δεν είναι πάντα εφικτή. Για παράδειγμα, συχνά τα δεδομένα σε ιατρικές και κοινωνικές μελέτες είναι ποσοτικά, βασιζόμενα όμως σε αυθαίρετες κλίμακες μέτρησης. Υπάρχει άφθονο υλικό στη βιβλιογραφία που ασχολείται με τη κατάλληλη επιλογή και εφαρμογή διατακτικών μοντέλων ανάλογα το είδος των εξεταζόμενων δεδομένων καθώς και υλικό που συγκρίνει τα διατακτικά μοντέλα με τα μη διατακτικά και το οποίο θα αναφερθεί στη συνέχεια.

1.2 Διατακτική ανάλυση

Στη συνήθη λογιστική παλινδρόμηση η μεταβλητή απόκρισης συνήθως έχει δυο κατηγορίες, την αποτυχία ή την επιτυχία. Όμως αυτές οι περιπτώσεις δεν είναι αποκλειστικές και μπορούμε να συναντήσουμε περισσότερες από δυο κατηγορίες. Σε αυτή τη περίπτωση θα μιλάμε για πολυτομικά (*polytomous*) δεδομένα. Ειδικότερα για τις πολυτομικές αποκρίσεις διακρίνουμε την περίπτωση όπου αυτές έχουν διατεταγμένες κατηγορίες..

Ωστόσο παρόλο που προσφέρουν μεγαλύτερη στατιστική σημαντικότητα στα αποτελέσματα αναλύσεων όπως θα δούμε και στην ενότητα, συχνά δεν χρησιμοποιούνται στην πράξη 1.4 και προτιμούνται. Έχει αναφερθεί (Svensson, 2001) πως οι κυριότεροι λόγοι που δεν χρησιμοποιούνται στις επιστήμες υγείας αυτού του είδους τα μοντέλα, είναι η έλλειψη γνώσεων, η τήρηση των παραδοσιακών μεθόδων μέσα σε ερευνητικές ομάδες και η ζήτηση πιο καθιερωμένων μοντέλων προκειμένου να γίνουν συγκρίσεις με προηγούμενες μελέτες .

1.2.1 Διατακτικές κατηγορίες μεταβλητών

Ο Anderson (1984) περιγράφει δυο τρόπους με τους οποίους δημιουργούνται οι διατακτικές μεταβλητές. Πρώτα, οι **ομαδοποιημένες συνεχείς** (*grouped continuous*) είναι απλώς μια κατηγοριοποίηση μιας συνεχούς μεταβλητής και είναι πολύ εύκολο να παρατηρηθεί. Για παράδειγμα, ο McCullagh (1980) αναφέρει τις κλάσεις εισοδήματος σε δολάρια (0-2000,2001-3000,...).

Η δεύτερη κατηγορία, η **αξιολογημένη** (*assessed*) είναι μια μορφη αξιολόγησης και προκύπτει όταν ένας εξωτερικός εκτιμητής αφού αξιοποιήσει τις διαθέσιμες πληροφορίες, συμπεραίνει το βαθμό σημαντικότητας της κάθε κατηγορίας της διατακτικής μεταβλητής. Οι Anderson and Philips (1981) όταν αναφέρονται στο βαθμό ανακούφισης από τον πόνο έπειτα από την θεραπεία (χειρότερα, ίδιο, μικρή βελτίωση, μέτρια βελτίωση, σημαντική βελτίωση, πλήρης ανακούφιση) στην πραγματικότητα αναφέρονται σε μια τέτοιου είδους μεταβλητή, αφού ο γιατρός που κάνει την αξιολόγηση του ασθενή, θα χρησιμοποιήσει το σύνολο των πληροφοριών που είναι διαθέσιμο για να εκτιμήσει την παρατηρούμενη κατηγορία πόνου.

Καταλαβαίνουμε πως οι εκτιμημένες μεταβλητές είναι πιθανό να περιλαμβάνουν σφάλμα που οφείλεται σε τρίτους παράγοντες (τον ερευνητή στην προκειμένη) και στις περισσότερες περιπτώσεις να επιδράσει η υποκειμενικότητα του ερευνητή στη δημιουργία της μεταβλητής.

1.3 Βιβλιογραφική και ιστορική ανασκόπηση

Τα μοντέλα παλινδρόμησης για διατεταγμένες μεταβλητές άρχισαν να χρησιμοποιούνται ευρέως στις βιοεπιστήμες. Οι Aitchison and Silvey (1957) πρότειναν το διατεταγμένο probit μοντέλο παλινδρόμησης για διατακτικές εξαρτημένες μεταβλητές και πιο συγκεκριμένα για την ανάλυση πειραμάτων στα οποία οι αποκρίσεις των συμμετεχόντων σε διάφορες δόσεις της δραστικής ουσίας, είναι χωρισμένες σε διατεταγμένες κλάσεις, ενώ οι Aitchison and Bennett (1970) για κατηγορικές μεταβλητές. Ο Snell (1964) ήταν αυτός που πρότεινε για λόγους μαθηματικής απλοποίησης τη χρήση της λογιστικής κατανομής αντί της κανονικής στις

προσεγγιστικές μεθόδους. Αυτοί που έθεσαν τα θεμέλια όμως και παρουσίασαν ενδελεχώς τα διατεταγμένα μοντέλα απόκρισης ήταν οι McKelvey and Zavoina (1975), γενικεύοντας τα μοντέλα των Aitchison και Silvey για περισσότερες από μια ανεξάρτητες μεταβλητές. Στην συγκεκριμένη εργασία υπέθεσαν την υποκείμενη ύπαρξη μιας συνεχούς λανθάνουσας μεταβλητής και η οποία συνδεόταν με ένα μονό πίνακα των ερμηνευτικών μεταβλητών και με έναν όρο σφάλματος. Στη συνέχεια, για να λάβουν την παρατηρηθείσα απόκριση της κατηγορικής εξαρτημένης μεταβλητής χώριζαν σε πεπερασμένα διακριτά διαστήματα την συνεχή μεταβλητή.

Ο Walker and Duncan (1967) και ο McCullagh (1980) ανέπτυξαν το αθροιστικό (*cumulative*) και αναλογικό μοντέλο (proportional) , στο οποίο μοντελοποίησαν τις αθροιστικές πιθανότητες της διατεταγμένης απόκρισης, ως ένα μονότονα αύξων μετασχηματισμό μιας γραμμικής προβλεπτικής μεταβλητής μετα πάνω στο διάστημα χρησιμοποιώντας τις probit και logit συναρτήσεις σύνδεσης. Αυτό το μοντέλο δίνει την ίδια συνάρτηση πιθανότητας με το μοντέλο των McKelvey και Zavoina που αναφέραμε παραπάνω και είναι συνεπώς ισοδύναμο. (Boes and Winkelmann, 2006)

Έχουν εισηγηθεί επίσης παραμετρικές γενικεύσεις πάνω στο θέμα, όπου κάνουν χρήση εναλλακτικών συναρτήσεων σύνδεσης, όπως γενικευμένες συναρτήσεις πρόβλεψης που περιλαμβάνουν παραμέτρους διασποράς Cox (1995), η log-log και η complementary log-log (McCullagh, 1980). Ακόμα, έχουν εισηγηθεί από τους (Olsson, 1979) και Ronning and Kukuk (1996) μοντέλα όπου τόσο οι εξαρτημένες όσο και οι ανεξάρτητες είναι διατεταγμένες και αποτελούν εφαρμογή των log-linear μοντέλων.
Από την άλλη χρησιμοποιούνται ημι-παραμετρικές και απαραμετρικές προσεγγίσεις για να παρακαμφθούν οι προϋποθέσεις κατανομής του συνηθισμένου μοντέλου. Αυτές αναφέρονται στις εργασίες των Agresti and Liu (1999), Barnhart and Sampson (1994) , Agresti et al. (1995), αλλά και των Winship, C. & Mare (1984), Bellemare et al. (2002), Stewart (2004)

1.4 Διαφορές στη μοντελοποίηση διατακτικών και μη διατακτικών μοντέλων παλινδρόμησης

Συνήθως οι διατακτικές μεταβλητές είτε μετατρέπονται σε συνεχείς είτε αντιμετωπίζονται ως ποιοτικές και υπολογίζονται απλώς οι αναλογίες σε κάθε επίπεδο της απόκρισης. Στη συνέχεια ελέγχονται με τεστ συσχέτισης χ^2 οι διαφορές στις αναλογίες (Heath 1985). Όμως χάνεται σημαντική πληροφορία έτσι καθώς τα αποτελέσματα εξαρτώνται από το μέγεθος της κάθε κατηγορίας , δεν παράγεται κάποιο μέτρο συσχέτισης και δεν λαμβάνεται υπόψιν η διάταξη της μεταβλητής (Lee, 1992).

Ακόμα, χρησιμοποιείται δίτιμη λογιστική παλινδρόμηση και συμπιέζεται η διατακτική κλίμακα σε δυο κατηγορίες. Σε αυτή τη περίπτωση η απόφαση για το που θα διχοτομηθεί η μεταβλητή είναι αυθαίρετη και αγνοείται η σχέση των λόγων σχετικών πιθανοτήτων (*odds ratio*) μεταξύ ενός επιπέδου διχοτόμησης και ενός άλλου επιπέδου.

Για να αποφευχθεί η αυθαίρετη διχοτόμηση, μπορούν απλώς να υπολογιστούν και να συγκριθούν οι λόγοι σχετικών πιθανοτήτων σε διάφορα επίπεδα διχοτόμησης. Σε αυτή την περίπτωση όμως οι διαφορές στις εκτιμήσεις μπορεί να οφείλονται σε τυχαίο σφάλμα.

Μια άλλη εναλλακτική είναι να χρησιμοποιηθούν τα πολυωνυμικά λογιστικά μοντέλα που χρησιμοποιούνται συνήθως σε πολυτομικές μεταβλητές απόκρισης, χωρίς ωστόσο να περιλαμβάνουν πληροφορίες για τη διάταξη της μεταβλητής. Ο Gurland et al. (1960) έδειξε πως στην βιολογική αξιολόγηση εάν η μεταβλητή απόκρισης περιλαμβάνει πάνω από δυο κατηγορίες, δηλαδή είναι πολύτομη, τότε είναι πιο αποτελεσματικό να χρησιμοποιηθούν όλες οι κατηγορίες παρά να ομαδοποιηθούν ορισμένες κατηγορίες για να διχοτομηθεί η μεταβλητή απόκρισης.

Ακόμα, χρησιμοποιούνται χ^2 τεστ για τάση, t - test, ανάλυση διακύμανσης και ανάλυση συνδιακύμανσης. Αυτά μετατρέπουν σε συνεχείς κλίμακες τις διατεταγμένες κατηγορίες για να ποσοτικοποιηθούν οι αποστάσεις μεταξύ των κατηγοριών. Ωστόσο η ποσοτικοποίηση των αποστάσεων είναι αυθαίρετη και σε αυτή την περίπτωση. Μάλιστα, οι τιμές που επιλέχθηκαν για να ποοσοτικοποιήσουν τις διατακτικές κατηγορίες μπορεί να έχουν σημαντική επίδραση στη συμπερασματολογία του μοντέλου (Everitt, 1992) και να οδηγήσουν σε παραπλανητικά συμπεράσματα (Hastie et al., 1989).

Πολλές φορές για την ανάλυση διατακτικών δεδομένων απόκρισης χρησιμοποιούνται τα σκορ μέσου όρου, πρακτική που είναι σπάνια δικαιολογημένη καθώς δεν είναι σίγουρη η ίση απόσταση μεταξύ των κατηγοριών. Στην έρευνά τους οι Christian (1984) μελετούν τους τραυματισμούς που προκύπτουν έπειτα από αυτοκινητιστικά ατυχήματα και τους κατανέμουν σε κατηγορίες ανάλογα την έκταση του τραυματισμού ως "καθόλου", "ελαφρός", "μέτριος", "σοβαρός" και "θανάσιμος", δίνοντάς τους τα σκορ 1,2,3,4,5 αντίστοιχα. Η σύγκριση του μέσου όρου της σοβαρότητας των σκορ, για παράδειγμα μεταξύ αυτών που χρησιμοποιούν ζώνη ασφαλείας και αυτών που δεν χρησιμοποιούν μπορεί να μην είναι έγκυρη λόγω αυθαίρετης θέσπισης των σκόρ. Συνεπώς, η διαφορά των σκορ 1 και 2 δε θα είναι η ίδια μεταξύ των σκορ 4 και 5. Δηλαδή, το μέσο σκορ 3 το οποίο προκύπτει από το σκορ 2 (ελαφρός) και 4 (σοβαρός) στους χρήστες ζώνης δε μπορεί να αντιστοιχηθεί με το μέσο σκορ υπολογισμένο από τα σκορ 1 (καθόλου) και 5 (θανάσιμος) σε μη χρήστες ζώνης.

Μερικές απαραμετρικές μέθοδοι, όπως το τεστ Wilcoxon-Mann-Whitney και η ανάλυση διακύμανσης κατά Kruskal-Wallis που έχουν χρησιμοποιηθεί στην εργασία των (Siegel and Castellan, 1988) συμπεριλαμβάνουν την διατακτικότητα των δεδομένων, ωστόσο αυτές

απλώς συγκρίνουν το διάμεσο σκορ που σχηματίζει η απόκριση μεταξύ δυο ή περισσότερων ομάδων της μεταβλητής. Αυτού του είδους η προσέγγιση μπορεί να μην είναι τόσο αποτελεσματική καθώς δεν λαμβάνει υπόψιν της ολόκληρη την κατανομή συχνοτήτων των σκορ απόκρισης στις αντίστοιχες ομάδες έκθεσης. Για παράδειγμα, θα ήταν πιθανό τόσο οι χρήστες ζώνης ασφαλείας όσο και οι μη χρήστες να έχουν το ίδιο διάμεσο σκορ τραυματισμού ακόμα και εάν μια ομάδα (χρήστες ζώνης) εμφανίζουν μεγαλύτερες συχνότητες χαμηλών σκορ από ότι μια άλλη ομάδα (μη χρήστες ζώνης). Παρόλο που μερικές μη παραμετρικές μέθοδοι όπως αυτή της ανάλογης με την εμπειρική κατανομή ανάλυση (*ridit analysis*) [1] λαμβάνουν υπόψιν τους την κατανομή συχνοτήτων των σκορ απόκρισης (Fleiss et al., 2003, Siegel and Castellan, 1988) καμία δεν μεριμνά για στατιστική προσαρμογή των συγκεχυμένων μεταβλητών ή την αξιολόγηση της τροποποίησης των επιδράσεων, δυο επιτακτικές ανάγκες στην αιτιολογική συμπερασματολογία στην Ιατρική και την Βιολογία (Anderson and Philips, 1981) (Rothman1986).

Ένα ακόμα μειονέκτημα στην εφαρμογή μοντέλων γραμμικής παλινδρόμησης είναι η υπόθεση της ομοσκεδαστικότητας, καθώς η διακύμανση διατεταγμένων δεδομένων που περιγράφονται από μια πολυωνυμική κατανομή δεν είναι ομοιογενής. Η εφαρμογή τέτοιων μοντέλων μπορεί να έχει αμερόληπτες εκτιμήσεις των παραμέτρων, ωστόσο οι εκτιμήσεις των διακυμάνσεων θα είναι μεροληπτικές.

Η χρήση των διάφορων διατακτικών μοντέλων έχει πολλές φορές αμφιλεγόμενη προστιθέμενη αξία στην στατιστική ανάλυση. Πιο συγκεκριμένα, αναφέρεται πως η χρήση τους όταν η μεταβλητή απόκρισης έχει πάνω απο 4 ή 5 κατηγορίες δεν οφελεί σημαντικά. Σημαντική ωφέλεια όμως θα προκύψει όταν έχουμε πάνω από 2 κατηγορίες εάν αντιμετωπίσουμε την μεταβλητή απόκρισης ως διατακτική σε αντίθεση με τη πρακτική δίασπασης (*collapsing*) σε περαιτέρω δίτιμες μεταβλητές(Whitehead, 1993).

[1] Για παραπάνω από δυο ομάδες επιλέγεται μια ομάδα αναφοράς, ενώ για την άλλη ομάδα το μέσο σκορ ridit είναι εκτίμηση της πιθανότητας μιας τυχαία επιλεγμένης παρατήρησης της ομάδας να έχει τιμές της υποκείμενης συνεχούς κλίμακας (μεταβλητής) μεγαλύτερής ή ίσης από την τιμή μιας τυχαίας επιλεγμένης παρατήρησης της ομάδας αναφοράς (Bross 1958)

Δηλαδή εάν μεγαλύτερες τιμές της υποκείμενης κλίμακας υποδεικνύουν μια χειρότερη κατάσταση, το μέσο σκορ ridit είναι η εκτιμημένη πιθανότητα ένα τυχαίο άτομο από την ομάδα να είναι σε χειρότερη κατάσταση από ένα τυχαίο άτομο από την ομάδα αναφοράς.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

Τα διατακτικά λογιστικά μοντέλα παλινδρόμησης δουλεύουν με το ίδιο σκεπτικό όπως τα δίτιμα λογιστικά μοντέλα, αλλάζοντας απλώς τον τρόπο που καθορίζονται οι πιθανότητες. Αντί να χρησιμοποιείται η πιθανότητα κάποιου μεμονωμένου γεγονότος ή κατηγορίας της μεταβλητής απόκρισης, χρησιμοποιείται η πιθανότητα αυτού του γεγονότος μαζί με τη πιθανότητα όλων των γεγονότων πριν από αυτό.

2.1 Πολυωνυμικό λογιστικό μοντέλο

Επειδή τα διατακτικά μοντέλα παλινδρόμησης είναι μια ειδική περίπτωση πολυτομικής παλινδρόμησης, κρίνεται σκόπιμο να αναφερθούν οι βασικές έννοιες της πολυτομικής λογιστικής παλινδρόμησης. Εάν η απόκριση ενός συμμετέχοντα ή αντικειμένου σε μια έρευνα είναι μια επιλογή από ένα καθορισμένο σύνολο δυνατών επιλογών, λέμε τότε ότι η απόκριση είναι πολυτομική (*polytomous*) ή αλλιως πολυωνυμική (*polynomial*). Οι k δυνατές τιμές της μεταβλητής απόκρισης Y ονομάζονται ως κατηγορίες απόκρισης Συνήθως οι κατηγορίες αυτές ορίζονται ως ποιοτικές ή μη αριθμήσιμες. Παράδειγμα μιας ποιοτικής ταξινόμησης αποτελεί η κατηγοριοποίηση των ομάδων αίματος στις ομάδες O, AB, A, B. Ένα άλλο παράδειγμα αποτελεί η κλίμακα ILO $(0/0, 0/1, \ldots, 3/3)$ για την ταξινόμηση εικόνων ακτινογραφιών στήθους σχετικά με την σοβαρότητα της πνευμοκονίωσης, ασθένειας του πνεύμονα. Αυτές οι κατηγορίες λόγω του αυθαίρετου ορισμού τους δεν έχουν σαφή διαχωρισμό μεταξύ τους.

Εάν οι κατηγορίες έχουν διάταξη στη δομή τους, τότε δεν τίθεται λόγος ισότιμου χειρισμού των ακραίων κατηγοριών και των ενδιάμεσων. Όμως, εάν οι κατηγορίες είναι μια συλλογή ετικετών χωρίς κάποια δομή, τότε δεν χρειάζεται να γίνει εκ των προτέρων διαλογή κάποιου υποσυνόλου των κατηγοριών προς διαφορετική αντιμετώπιση. Λόγω της ποικιλίας στη μορφή των κατηγοριών απόκρισης που συναντούμε, υπάρχει και η ανάλογη ποι-

κιλία στην επιλογή των συναρτήσεων σύνδεσης. Οποιαδήποτε μορφή και εάν έχει η κλίμακα της μεταβλητής απόκρισης όμως, μπορούν να χρησιμοποιηθούν οι πιθανότητες απόκρισης $\pi_1, \pi_2, \ldots, \pi_k$.

Πρέπει να σημειωθεί πως το πολυωνυμικό μοντέλο για k κατηγορίες απόκρισης, ορίζεται από $k-1$ εξισώσεις και υπάρχει μόνο ένας συντελεστής β_{ij} για κάθε συνδυασμό των κατηγοριών και των συμεταβλητών. Έτσι, δεν είναι πιθανόν να συνοψίσουμε την επίδραση μιας συμμεταβλητής στην μεταβλητή απόκρισης Y με ένα μόνο μέτρο, όπως για παράδειγμα έναν λόγο σχετικών πιθανοτήτων. Αν και το πολυωνυμικό μοντέλο προσφέρει την δυνατότητα της ταυτόχρονης εξέτασης της επίδρασης μιας συμμεταβλητής σε όλες τις κατηγορίες της μεταβλητής απόκρισης, παρόλα αυτά παράγει μεγάλο αριθμό στατιστικής πληροφορίας κάνοντας δύσκολη την ανάλυση.

Χρησιμοποιείται ένα σύνολο I από ανεξάρτητες παρατηρήσεις από ερμηνευτικές μεταβλητές και αποκρίσεις. Η *n-οστή* παρατήρηση, για $i = 1, \ldots, I$ περιλαμβάνει παρατηρήσεις από ένα διάνυσμα διάστασης p ερμηνευτικών μεταβλητών $(\mathbf{X}_i = x_{i1}, \ldots, x_{ip})$ και μια μεταβλητή απόκρισης Y_i η οποία ανήκει σε μια από τις J διακριτές κατηγορίες $(j = 1, \ldots, J)$ της μεταβλητής απόκρισης. Δοθέντος του x_i, η πιθανότητα μιας απόκρισης $Y_i = j$ θα γράφεται ως $\pi_j(x_i)$. Τα διάφορα μοντέλα για τις πολύτομες αποκρίσεις διαφέρουν στον τρόπο που ορίζουν την σχέση ανάμεσα στις ερμηνευτικές μεταβλητές x_i και τις πιθανότητες $\pi_j(x_i)$.

Στην βιβλιογραφία συνήθως οι παρατηρήσεις θεωρούνται πως είναι ανεξάρτητες. Υπό αυτή την υπόθεση για μια γνωστή τιμή x του διανύσματος των ερμηνευτικών μεταβλητών, το διάνυσμα $(n_1(x), \ldots, n_J(x))$ των τιμών των κατηγορικών αποκρίσεων ακολουθεί την πολυωνυμική κατανομή με το διάνυσμα πιθανοτήτων $(\pi_1(x), \ldots, \pi_j(x))$. Ο συνολικός αριθμός παρατηρήσεων στην κατηγορία j συμβολίζεται ως $n_j(x)$ και λαμβάνει υπόψιν όλες τις παρατηρήσεις i για τις οποίες ισχύει $Y_i = j$ και για τις οποίες το διάνυσμα των ερμηνευτικών μεταβλητών x_i γράφεται ως x. Να σημειωθεί εδώ πως οι McCullagh and Nelder (1989) πρότειναν την εκτιμήτρια ημιμεγίστης πιθανοφάνειας (*quasi-maximum likelihood estimator*) για υπερδιεσπαρμένα πολύτομα δεδομένα ως προσέγγιση που δεν χρειάζεται να ικανοποιεί την υπόθεση ανεξαρτησίας.

2.1.1 Μοντέλα λογιστικής συνάρτησης

Να θυμίσουε ότι η logit συνάρτηση είναι ο λογάριθμος της σχετικής πιθανότητας για ένα γεγονός (*odds*), δηλαδή ο λογάριθμος της πιθανότητας να συμβεί ένα γεγονός και ορίζεται

ως:

$$\text{odds} = \text{logit}(\pi) = \log \frac{\pi}{1-\pi} = \alpha + \beta x \Leftrightarrow \tag{2.1}$$

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \tag{2.2}$$

Οι συντελεστές β δείχνουν πόσο αλλάζει το logit βασισμένο στις τιμές των επεξηγηματικών μεταβλητών.

2.1.2 Κατηγορική λογιστική παλινδρόμηση

Η κατηγορική λογιστική παλινδρόμηση χρησιμοποιείται όταν δεν υπάρχει φυσική διάταξη στις κατηγορίες της μεταβλητής απόκρισης. Έστω Y μια κατηγορική μεταβλητή απόκρισης με J κατηγορίες. Μια κατηγορία επιλέγεται αυθαίρετα ως *κατηγορία αναφοράς*. Εάν για παράδειγμα η κατηγορία αναφοράς είναι η πρώτη, έχουμε:

$$\text{logit}(\pi_j) = \log(\frac{\pi_j}{\pi_1}) = \alpha_j + \boldsymbol{\beta}_j^T \mathbf{X}_j,\ j = 2, \ldots, J \tag{2.3}$$

όπου με $\boldsymbol{\beta}_j^T$ συμβολίζουμε τον ανάστροφο του διανύσματος $\boldsymbol{\beta}_j$.

Τα πολυτομικά μοντέλα logit για ονομαστικές αποκρίσεις περιγράφουν ταυτόχρονα τις λογαριθμικές σχετικές συχνότητες (*log odds*) (ή ισοδύναμα τους συντελεστές β_j) για όλα τα $\binom{J}{2}$ ζευγάρια κατηγοριών. Αφού υπολογιστούν οι εκτιμήσεις των β_j τότε μπορούν να υπολογιστούν οι γραμμικές εκτιμήσεις $\mathbf{X}_j \boldsymbol{\beta}_j^T$. Από (2.3)

$$\hat{\pi}_j = \hat{\pi}_1\ \exp(\boldsymbol{\beta_j^T X_j}),\ j = 2, \ldots, J. \tag{2.4}$$

Αλλά $\hat{\pi}_1 + \hat{\pi}_2 + \cdots + \hat{\pi}_j = 1$, οπότε

$$\hat{\pi}_1 = \frac{1}{1 + \sum_{j=2}^{J} \exp(\boldsymbol{\beta_j^T X_j})} \tag{2.5}$$

$$\text{και } \pi_j = \frac{\exp(\boldsymbol{X}_j \boldsymbol{\beta}_j^T)}{1 + \sum_{j=2}^{J} \exp(\boldsymbol{\beta_j^T X_j})},\ j = 2, \ldots, J. \tag{2.6}$$

Σχόλιο: Οι παραπάνω σχέσεις δίνουν $J-1$ εξισώσεις γιατι η Y μεταβλητή έχει J κατηγορίες.

Οι εκτιμώμενες τιμές (προσδοκώμενες συχνότητες) για κάθε κατηγορία j της συμμεταβλητής μπορούν να υπολογιστούν πολλαπλασιάζοντας τις εκτιμημένες πιθανότητες $\hat{\pi}_j$ με την συνολική συχνότητα της συμμεταβλητής.

2.1.3 Logit κατηγορίας αναφοράς (*baseline-category logit*)

Τα logit κατηγορίας αναφοράς είναι μια μέθοδος συμπερασματολογίας στην περίπτωση της πολυτομικής λογιστικής παλινδρόμησης, και όπως θα δούμε στη συνέχεια στην ενότητα 2.4.1, σχετίζονται με τα logit μοντέλα γειτονικής κατηγορίας (adjacent-category logits).

Έστω Y μια κατηγορική μεταβλητή απόκρισης με J κατηγορίες. Τα πολυτομικά μοντέλα logit για ονομαστικές αποκρίσεις περιγράφουν ταυτόχρονα τις λογαριθμικές σχετικές συχνότητες (*log odds*) για όλα τα $\binom{J}{2}$ ζευγάρια κατηγοριών. Δοθέντος $J-1$ κατηγοριών, η εξέταση των υπόλοιπων ζευγαριών είναι περιττή.

Αν θέσουμε ως $\pi_j(\mathbf{X}) = \Pr(J = j \mid \mathbf{X})$ σε ενα σύνολο $\mathbf{X}$ επεξηγηματικών μεταβλητών και $\sum_j \pi_j(\mathbf{X}) = 1$. Για παρατηρήσεις σε αυτό το σύνολο, θεωρούμε τις συχνότητες στις J κατηγορίες της Y ως πολυτομικές με πιθανότητες $\{\pi_1(\mathbf{X}), \ldots, \pi_j(\mathbf{X})\}$.

Τα logit μοντέλα συγκρίνουν κάθε κατηγορία της απόκρισης με μια κατηγορία αναφοράς, συνήθως την πιο πολυπληθή ή την τελευταία (J κατηγορία). Το μοντέλο σε αυτή την περίπτωση είναι το

$$\log \frac{\pi_j(\mathbf{X})}{\pi_J(\mathbf{X})} = \alpha_j + \boldsymbol{\beta}_j'\mathbf{X},\ j = 1, \ldots, J-1 \tag{2.7}$$

και περιγράφει ταυτόχρονα τις επιδράσεις του $\mathbf{X}$ στα $J-1$ logit. Οι επιδράσεις διαφέρουν ανάλογα με την απόκριση που συγκρίνεται με την κατηγορία αναφοράς. Οι $J-1$ εξισώσεις που προκύπτουν, καθορίζουν τις παραμέτρους για τα logits με άλλα ζευγάρια κατηγοριών (Agresti, 2002) απόκρισης, αφού:

$$\log \frac{\pi_\alpha(\mathbf{X})}{\pi_\beta(\mathbf{X})} = \log \frac{\pi_\alpha(\mathbf{X})}{\pi_J(\mathbf{X})} - \log \frac{\pi_\beta(\mathbf{X})}{\pi_J(\mathbf{X})} \tag{2.8}$$

2.1.4 Εκτίμηση πιθανοτήτων απόκρισης

Για να εκφράσουμε πολυτομικά logit μοντέλα απευθείας σε όρους κατηγοριών απόκρισης $\{\pi_j(\mathbf{X})\}$ χρησιμοποιούμε τον ακόλουθο τύπο:

$$\pi_j(\mathbf{X}) = \frac{\exp(\alpha_j + \boldsymbol{\beta}_j^T\mathbf{X})}{1 + \sum_{h=1}^{J-1} \exp(\alpha_h + \boldsymbol{\beta}_h^T\mathbf{X})} \tag{2.9}$$

με $\alpha_J = 0$ και $\boldsymbol{\beta}_J = 0$. Αυτό προκύπτει από το την (2.7) εάν χρησιμοποιήσουμε το γεγονός πως η (2.7) ισχύει στην περίπτωση που $j = J$ θέτοντας $\alpha_J = 0$ και $\boldsymbol{\beta}_J = 0$. Ακόμα, παρατηρούμε πως ο παρανομαστής στην (2.9) είναι ο ίδιος για διάφορα j και πως οι αριθμητές για το κάθε j αθροίζουν στον παρανομαστή, οπότε ισχύει πως $\sum_j \pi_j(\mathbf{x}) = 1$. Τέλος, για $J = 2$ κατηγορίες, η 2.9 απλοποιείται στον τύπο που χρησιμοποιείται για την δίτιμη λογιστική παλινδρόμηση $\left(\pi(x) = \dfrac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}\right)$.

2.2 Logit μοντέλα

2.2.1 Μοντελοποίηση αθροιστικών συχνοτήτων

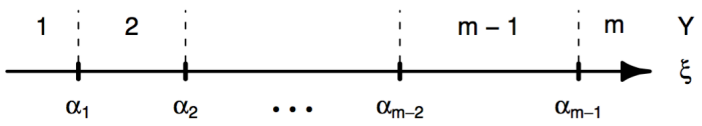
Τα διατακτικά αθροιστικά (*cumulative*) λογιστικά μοντέλα παλινδρόμησης δουλεύουν με το ίδιο σκεπτικό όπως τα δίτιμα λογιστικά μοντέλα, αλλάζοντας απλώς τον τρόπο που καθορίζονται οι πιθανότητες. Αντί να χρησιμοποιείται η πιθανότητα κάποιου μεμονωμένου γεγονότος ή κατηγορίας της μεταβλητής απόκρισης, χρησιμοποιείται η πιθανότητα αυτού του γεγονότος και η πιθανότητα όλων αυτών των γεγονότων πριν από αυτό.

Έστω ότι υπάρχει μια λανθάνουσα μεταβλητή ξ η οποία είναι γραμμική συνάρτηση των X και τυχαίων σφαλμάτων ε_i τα οποία ακολουθούν κάποια καθορισμένη συμμετρική κατανομή με μέσο όρο το 0, όπως η κανονική ή η λογιστική κατανομή:

$$\xi = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i \tag{2.10}$$

Αντί να χωρίσουμε σε δυο περιοχές το ξ και να δημιουργήσουμε μια δίτιμη απόκριση, η περιοχή του ξ χωρίζεται από $m-1$ άγνωστα όρια ή κατώφλια σε m γειτονικές περιοχές. Τα κατώφλια αυτά χωρίζουν επίσης τις γειτονικές κατηγορίες οι οποίες θα εκτιμηθούν με τους συντελεστές β στη (2.10). Σημειώνοντας τα κατώφλια (Σχήμα 2.1) με $\alpha_1 < \alpha_2 < \cdots < \alpha_{m-1}$ και με Y την τυχαία μεταβλητή απόκρισης που προκύπτει και η οποία περιγράφει την υποκείμενη τάση του φαινομένου έχουμε:

$$Y = \begin{cases} 1 & \text{αν } \xi \leq \alpha_1 \\ 2 & \text{αν } \alpha_1 < \xi \leq \alpha_2 \\ \vdots & \\ m-1 & \text{αν } \alpha_{m-2} < \xi \leq \alpha_{m-1} \\ m & \text{αν } \alpha_{m-1} < \xi \end{cases} \tag{2.11}$$



Σχήμα 2.1: διαμέριση του συνεχούς διαστήματος σε υποδιαστήματα

Χρησιμοποιώντας το μοντέλο για τη λανθάνουσα μεταβλητή σε συνάρτηση με τα κατώφλια παίρνουμε την αθροιστική συνάρτηση κατανομής της Y:

$$\begin{aligned}\Pr(Y_i \leq j) &= \Pr(\xi \leq a_j) \\ &= \Pr(\alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i \leq a_j) \\ &= \Pr(\varepsilon_i \leq \alpha_j - \alpha - \beta_1 X_{i1} - \cdots - \beta_k X_{ik})\end{aligned} \tag{2.12}$$

Εάν τώρα τα τυχαία σφάλματα ε_i είναι κατανεμημένα ανεξάρτητα και ακολουθούν την κανονική κατανομή οδηγούμαστε στο διατεταγμένο **probit** μοντέλο, ενώ εάν ακολουθούν την λογιστική κατανομή, τότε οδηγούμαστε στο διατεταγμένο **logit** μοντέλο (χρησιμοποιοείται η σχέση 2.1):

$$\begin{aligned}\operatorname{logit}\Big[\Pr(Y_i \leq j)\Big] &= \log\frac{\Pr(Y \leq i)}{1 - \Pr(Y \leq i)} = \log\frac{\Pr(Y_i \leq j)}{\Pr(Y_i > j)} \\ &= \alpha_j - \alpha - \beta_1 X_{i1} - \cdots - \beta_k X_{ik}\end{aligned} \tag{2.13}$$

2.2.2 Κατασκευή κατωφλίων

2.2.2.1 Συμμετρικά κατώφλια

Το βασικό μοντέλο αθροιστικής συνάρτησης σύνδεσης υποθέτει πως όλα τα κατώφλια είναι σταθερά για όλες τις τιμές του $\boldsymbol{\beta}^T\mathbf{x}$, είναι διατεταγμένα και πεπερασμένα αλλά χωρίς κάποια δομή. Για παράδειγμα, σε ερωτηματολόγια οι απαντήσεις στο ερώτημα "Πόσο συμφωνείτε με ..." τοποθετούνται σε κλίμακα μορφής "συμφωνώ απόλυτα" έως "διαφωνώ απόλυτα" με ενδιάμεσες κατηγορίες. Σε αυτό τον τύπο απαντήσεων η κλίμακα των απαντήσεων είναι χρήσιμο να είναι συμμετρική, για παράδειγμα οι ακραίες απαντήσεις έχουν την ίδια απόσταση από τις μεσαίες απαντήσεις.

Στην παρακάτω εικόνα, έχουμε έξι κατηγορίες απόκρισης και πέντε κατώφλια, όπου το κεντρικό κατώφλι θ_3 αντιστοιχεί στο c και τα a και b είναι μέτρα αποστάσεων που καθορίζουν την απόσταση έως τα υπόλοιπα κατώφλια. Αυτή η προσέγγιση είναι πιο φειδωλή, καθώς σε αυτή την περίπτωση χρησιμοποιούνται μόνο τρεις παράμετροι αντί για πέντε που θα χρειάζονταν κανονικά για να καθορίσυν τα κατώφλια.

θ_1	θ_2	θ_3	θ_4	θ_5
$-b+c$	$-a+c$	c	$a+c$	$b+c$

Σχήμα 2.2: Συμμετρικά κατώφλια με 6 κατηγορίες

2.2.2.2 Κατώφλια ίσης απόστασης

Ορισμένες περιπτώσεις διατακτικών μεταβλητών στηρίζονται σε μια διατακτική κλίμακα για την μεταβλητή κλίμακα απόκρισης. Σε αυτές τις περιπτώσεις θεωρούμε πως τα κατώφλια ισαπέχουν μεταξύ τους. Αυτού του είδους τα κατώφλια χρησιμοποιούν μόνο δυο παραμέτρους για να οριστούν

$$\theta_j = \alpha + b(j-1),\ j = 1, \ldots, J-1 \tag{2.14}$$

έτσι ώστε το $\theta_1 = \alpha$ είναι το πρώτο κατώφλι και *b* είναι η απόσταση μεταξύ γειτονικών κατωφλίων.

2.2.3 Αθροιστικά logit μοντέλα (*cumulative logit models*)

Το αθροιστικό μοντέλο logit όπως θα δούμε παρακάτω μοντελοποιεί τον λογάριθμο της σχετικής πιθανότητας να ανήκει μια παρατήρηση μέχρι κάποια κατηγορία *j*. Το μοντέλο που θα περιγράψουμε παρακάτω εκτός από το ότι απευθύνει το πρόβλημα της υποκείμενης διάταξης διατεταγμένων δεδομένων, αντιμετωπίζει και το πρόβλημα της στατιστικής προσαρμογής των συγκεχυμένων μεταβλητών ή την αξιολόγηση της τροποποίησης των επιδράσεων που αναφέραμε στην παράγραφο 1.4 .

Έχουμε μια μεταβλητή απόκρισης Y με $1, 2, \ldots, k$ κατηγορικές αποκρίσεις. Για κάθε υποκείμενο i που συμμετέχει στις μετρήσεις (ασθενής κ.α) θα ορίσουμε ως y_i ως την κατηγορία που προκύπτει για τη μεταβλητή απόκρισης, ενώ θα ορίσουμε ως $\mathbf{X}_i$ το διάνυσμα συμμεταβλητών p διάστασης με τις τιμές των ερμηνευτικών μεταβλητών.

Η Y ακολουθεί μια πολυωνυμική κατανομή με παράμετρο $\boldsymbol{\pi}$, όπου π_{ij} δηλώνει τη πιθανότητα η *i*-οστή παρατήρηση να βρίσκεται στην *j*-οστή κατηγορία.

Βασίζονται στις αθροιστικές πιθανότητες απο θεωρημα ολικης πιθανοτητας:

$$\gamma_{ij} = \Pr(Y_i \leq j) = p_{i1} + \cdots + p_{ij},\ j = 1, \ldots, k \tag{2.15}$$

Για λόγους απλότητας, αν και πρόκειται στην ουσία για μοντέλο δεσμευμένης πιθανότητας για κάθε δεδομένη κατηγορία της μεταβλητής απόκρισης στις επεξηγηματικές μεταβλητές, στη συνέχεια θα αναγράφουμε την πιθανότητα $\Pr(X_i < j \mid \mathbf{X_j})$ ως $\Pr(Y \leq j)$ εκτός και εάν χρειαστεί να αναφερθούμε σε ξεχωριστές μονάδες παρατήρησης.

Χρησιμοποιώντας την σχέση 2.2 σχηματίζονται οι αθροιστικές σχετικές πιθανότητες (*odds*):

$$\text{odds} = \Pr(Y \leq j) = \frac{\Pr(Y \leq j)}{1 - \Pr(Y \leq j)} = \frac{p_1 + \cdots + p_j}{p_{j+1} + \ldots p_k},\ j = 1, \ldots, k-1 \tag{2.16}$$

και ακολούθως τα αθροιστικά logit:

$$\text{logit}(\gamma_{ij}) = \text{logit}\left[\Pr(Y \leq j)\right] = \log \frac{\Pr(Y \leq j)}{1 - \Pr(Y \leq j)},\ j = 1, 2, \ldots, k-1. \tag{2.17}$$

Το παραπάνω *logit* (2.17) ισούται με το συνηθισμένο δίτιμο logit το οποίο εφαρμόζεται όταν διαχωρίζεται η μεταβλητή απόκρισης σε δυο αποτελέσματα $Y \leq j$ και $Y > j$. Κάθε αθροιστικό logit χρησιμοποιεί όλες τις k κατηγορίες απόκρισης αλλά ορίζονται για όλες εκτός της τελευταίας. [1]

Χρησιμοποιώντας την σχέση (2.16) η σχέση (2.17) έρχεται στην ακόλουθη μορφή και δημιουργείται ένα αθροιστικό μοντέλο με τη λογιστική συνάρτηση σύνδεσης. Ονομάστηκε μοντέλο σωρευτικών logits από τους Walker and Duncan (1967) και αργότερα στηρίχτηκε σε αυτά ο McCullagh (1980) για να ορίσει τα **proportional odds** που θα παρουσιαστούν στην επόμενη παράγραφο.

$$\text{logit}(\gamma_{ij}) = \text{logit}\left[\Pr(Y \leq j)\right] = \log \frac{\pi_1 + \cdots + p_j}{\pi_{j+1} + \cdots + \pi_k} = \alpha_j + \boldsymbol{\beta}_k^T \mathbf{X} \tag{2.18}$$

Το διάνυσμα $\boldsymbol{\beta}$ με τις παραμέτρους περιγράφει τις επιδράσεις των επεξηγηματικών μεταβλητών για κάθε k-οστή κατηγορία .

$$\text{odds} = \Pr(Y \leq y_j \mid \mathbf{X}) = \frac{exp(\alpha_j + \boldsymbol{\beta}_k^T \mathbf{X})}{1 + exp(\alpha_j + \boldsymbol{\beta}_k^T \mathbf{X})},\ j = 1, 2, \ldots, k \tag{2.19}$$

Παρατηρούμε πως το logit για κάθε αθροιστική πιθανότητα j, έχει το δικό του σταθερό όρο α_j. Τα α_j αυξάνονται μαζί με το j, αφού η πιθανότητα $\Pr(Y \leq j)$ αυξάνεται με κάθε j για κάθε δοσμένη τιμή του $\mathbf{X}$ και το logit είναι αύξουσα συνάρτηση αυτής της πιθανότητας. Για να ερμηνεύσουμε τα αποτελέσματα θα χρησιμοποιήσουμε την σχέση 2.2:

2.2.4 Αναλογικές σχετικές πιθανότητες (*Proportional Odds*)

Στο παραπάνω μοντέλο (2.19) περιέχονται ένα μεγάλο πλήθος παραμέτρων, δηλαδή ξεχωριστές παράμετροι για κάθε συνδυασμό των συμεταβλητών και των κατηγοριών. Υπάρχουν περιπτώσεις που θέλουμε να σχηματίσουμε πιο φειδωλά και οικονομικά μοντέλα. Ο McCullagh (1980) πρότεινε ένα μοντέλο για αυτή την περίπτωση που δίνεται παρακάτω, στο οποίο οι παράμετροι β της λογιστικής παλινδρόμησης δεν εξαρτώνται από την k -οστή κατηγορία παρά μόνο από την εκάστοτε συμμεταβλητή.

$$\text{odds} = \Pr(Y \leq y_j \mid \mathbf{X}) = \frac{exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{X})}{1 + exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{X})},\ j = 1, 2, \ldots, k \tag{2.20}$$

[1] αφού για $j = k$ ο παρανομαστής θα ήταν: $1 - \Pr(Y \leq k) = 1 - 1 = 0$

Οι *k* σχετικές πιθανότητες (*odds*) για κάθε μία από τις *j* κατηγορίες διαφέρουν μόνο σε σχέση με τους σταθερούς όρους α_j, είναι δηλαδή εξ΄ου και ο χαρακτηρισμός τους ως αναλογικών (*proportional*). Αυτή η σχετικά αυστηρή υπόθεση για την αναλογικότητα των σχετικών πιθανοτήτων ισχύει ειδικά για περιπτώσεις όπου η διατακτική μεταβλητή απόκρισης Y σχετίζεται με συνεχή λανθάνουσα μεταβλητή (McCullagh, 1980), για παράδειγμα εάν η Y είναι μια ομαδοποιημένη συνεχής μεταβλητή (π.χ ηλικιακές ομάδες, ομάδες εισοδήματος). Εάν ισχύει η υπόθεση της αναλογικότητας τότε οι επιδράσεις του μοντέλου δεν επηρεάζονται από την επιλογή του αριθμού των κατηγοριών ή των κατωφλίων, ισχυρισμός που θα αναλυθεί στο Κεφάλαιο 3.

Ο Anderson (1984) σημείωσε πως συχνά, για τις εκτιμημένες μεταβλητές (αναφέρονται στην ενότητα 1.2.1) οι αναλογικές σχετικές συχνότητες δεν είναι αρκετά ευέλικτες για να καλύψουν το εύρος των δυνατών προβλημάτων, η υπόθεση για την ύπαρξη της λανθάνουσας συνεχούς μεταβλητής δεν ικανοποιείται και η διάταξη που προτείνεται από τον ερευνητή δεν είναι σχετική με το εκάστοτε πρόβλημα. Έτσι πρότεινε μια γενική τάξη μοντέλων για διατακτικά μοντέλα, τα **στερεοτυπικά** (*stereotype*) διατακτικά μοντέλα τα οποία περιλαμβάνουν τα αναλογικά μοντέλα ως υποπερίπτωση και επιτρέπουν την εξέταση της διατακτικότητας της μεταβλητής απόκρισης.

Εάν και τα μοντέλα proportional odds πρέπει να εφαρμόζονται προσεκτικά καθώς η υπόθεση των αναλογικών σχετικών πιθανοτήτων δεν είναι αληθής για όλες τις διατακτικές κατηγορίες μεταβλητών. Ωστόσο είναι αρκετά δημοφιλής για δυο λόγους. Πρώτον, η επίδραση μιας συμμεταβλητής στην Y μπορεί να ποσοτικοποιηθεί από μόνο ένα συντελεστή παλινδρόμησης κάνοντας δυνατό τον υπολογισμό ενός κοινού λόγου σχετικών πιθανοτήτων. Έτσι, η παρουσίαση των αποτελεσμάτων γίνεται πιο περιεκτικά και απλά. Δεύτερον, είναι δυνατή η χρησιμοποίηση υπολογιστικών τεχνικών κλιμακωτής (*stewpwise*) επιλογής μεταβλητών.

Να σημειώσουμε τέλος πως ισχύει και πάλι η παρατήρηση που κάναμε πιο πάνω πως το logit για κάθε αθροιστική πιθανότητα j, έχει το δικό του σταθερό όρο α_j. Τα α_j αυξάνονται μαζί με το j, αφού η πιθανότητα $\Pr(Y \leq j)$ αυξάνεται με κάθε j για κάθε δοσμένη τιμή του $\mathbf{X}$ και το logit είναι αύξουσα συνάρτηση αυτής της πιθανότητας. Για να ερμηνεύσουμε τα αποτελέσματα θα χρησιμοποιήσουμε την σχέση (2.2):

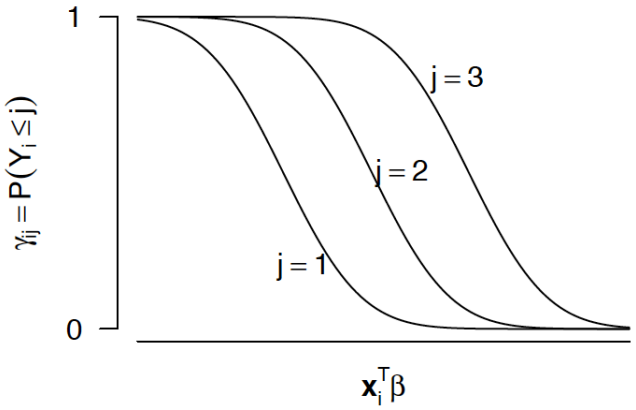
Χρησιμοποιώντας όσα αναφέρθηκαν για τα αθροιστικά logit μοντέλα στην προηγούμενη ενότητα (2.2.3) προσαρμόζουμε τη σχέση (2.18) για να ανταποκρίνεται στην ερμηνεία του McCullagh:

$$\text{logit}(\gamma_{ij}) = \alpha_j + \boldsymbol{\beta}^T\mathbf{X} \tag{2.21}$$

Για μικρές τιμές του $\boldsymbol{\beta}^T\mathbf{X}$ η απόκριση είναι πιθανότερο να πέσει μέσα στην πρώτη κα-

τηγορία και για μεγάλες τιμές του $\boldsymbol{\beta}^T\mathbf{X}$ είναι πιθανότερο να βρίσκεται στην τελευταία κατηγορία. Οι οριζόντιες μετατοπίσεις των καμπυλών περιγράφονται από τις τιμές του όρου α_j.

Να σημειωθεί εδώ πως ενώ η σχέση 2.18 όπως και τα διάφορα γραμμικά μοντέλα παλινδρόμησης να γράφονται με θετικό πρόσημο, είναι σύνηθες το παραπάνω μοντέλο να γράφεται με αρνητικό. Πρώτον, όσο πιο μεγάλη η τιμή του $\boldsymbol{\beta}^T\mathbf{X}$, τόσο μεγαλύτερη η πιθανότητα η κατηγορία απόκρισης να βρίσκεται στο υψηλότερο μέρος της κλίμακας απόκρισης. Έτσι, το $\boldsymbol{\beta}^T$ έχει την ίδια κατεύθυνση επίδρασης όπως και η παράμετρος παλινδρόμησης σε ένα κανονικό μοντέλο γραμμικής παλινδρόμησης ή ανάλυσης διασποράς (ANOVA). Ο δεύτερος λόγος σχετίζεται με την ερμηνεία της υποκείμενης μεταβλητής των μοντέλων αθροιστικής συνάρτησης σύνδεσης και την οποία θα αναλύσουμε στην 2.3.



Σχήμα 2.3: Απεικόνιση ενος αθροιστικού μοντέλου logit με 4 κατηγορίες απόκρισης

2.2.4.1 Λόγοι σχετικών πιθανοτήτων και αναλογικές σχετικές πιθανότητες

Ο λόγος σχετικών πιθανοτήτων για ένα γεγονός $Y \leq j$ στο $\mathbf{X}_1$ σε σχέση με το ίδιο γεγονός στο $\mathbf{X}_2$ είναι:

$$\text{odds} = \frac{\dfrac{\gamma_j(\mathbf{X}_1)}{1-\gamma_j(\mathbf{X}_1)}}{\dfrac{\gamma_j(\mathbf{X}_2)}{1-\gamma_j(\mathbf{X}_2)}} = \frac{\exp(\alpha_j - \boldsymbol{\beta}^T\mathbf{X}_1)}{\exp(\alpha_j - \boldsymbol{\beta}^T\mathbf{X}_2)} = \exp[\boldsymbol{\beta}^T(\mathbf{X}_2 - \mathbf{X}_1)] \tag{2.22}$$

ο οποίος είναι ανεξάρτητος του j. Έτσι ο αθροιστικός λόγος σχετικών πιθανοτήτων είναι αναλογικός με την απόσταση μεταξύ των $\mathbf{X}_1$ και $\mathbf{X}_2$. Εάν η $\mathbf{X}$ αντιπροσωπεύει μια μετα-

βλητή θεραπείας με δυο επίπεδα (π.χ placebo, θεραπεία) τότε ισχύει $\mathbf{X}_2 - \mathbf{X}_1 = 1$ και ο λόγος σχετικών πιθανοτήτων είναι $\exp(-\beta_{\text{θεραπεία}})$. Ομοίως ο λόγος σχετικών πιθανοτήτων του γεγονότος $Y \geq j$ είναι $\exp(\beta_{\text{θεραπεία}})$.

Τα διαστήματα εμπιστοσύνης (δ.ε) για τους λόγους σχετικών πιθανοτήτων δίνονται μετασχηματίζοντας τα όρια των διαστημάτων εμπιστοσύνης για το $\boldsymbol{\beta}^T$, το οποίο θα οδηγήσει σε μη συμμετρικά δ.ε για τα odds ratios. Συμμετρικά δ.ε κατασκευασμένα από το τυπικό σφάλμα των odds ratios δεν είναι κατάλληλα και θα πρέπει να αποφεύγονται .

2.2.5 Μερικώς αναλογικό μοντέλο σχετικών πιθανοτήτων (PPOM)

Είναι σπάνιο για όλες τις επεξηγηματικές μεταβλητές που περιλαμβάνονται σε ένα μοντέλο να εμφανίζουν την υπόθεση των όμοιων αναλογικών σχετικών συχνοτήτων. Η έννοια του μερικώς αναλογικού μοντέλου σχετικών πιθανοτήτων (*partial proportional odds*) εισήχθηκε από τους Peterson and Harrell (1990) για να επιτραπεί σε ορισμένες μόνο επεξηγηματικές μεταβλητές ενός μοντέλου να υπακούουν την υπόθεση του αναλογικού μοντέλου και για τις υπόλοιπες που δεν την υπακούουν ορίζει ειδικές παραμέτρους οι οποίες διαφέρουν για τις διάφορες κατηγορίες που συγκρίνονται. Ακολουθούν δυο τύποι μερικώς αναλογικών μοντέλων, το περιορισμένο και το μη περιορισμένο.

2.2.5.1 Μη περιορισμένο μερικώς αναλογικό μοντέλο σχετικών πιθανοτήτων

Όπως αναφέρθηκε παραπάνω, το μερικώς αναλογικό μοντέλο σχετικών πιθανοτήτων επιτρέπει μη αναλογικές σχετικές πιθανότητες για ένα υποσύνολο q των p ερμηνευτικών μεταβλητών $(q < p)$. Επιπλέον, επιτρέπει να εξεταστεί η υπόθεση των αναλογικών σχετικών συχνοτήτων για το υποσύνολο q.

Για μια διατακτική μεταβλητή Y με k κατηγορίες και $\mathbf{x}$ ένα διάνυσμα ερμηνευτικών μεταβλητών διάστασης p το προτεινόμενο μοντέλο για τις αθροιστικές πιθανότητες είναι το

$$\Pr(Y \leq y_j \mid \mathbf{X}) = \frac{exp(-\alpha_j - \boldsymbol{\beta}^T\mathbf{X} - \mathbf{t}^T\boldsymbol{\gamma_j})}{1 + exp(-\alpha_j - \boldsymbol{\beta}^T\mathbf{X} - \mathbf{t}^T\boldsymbol{\gamma_j})},\ j = 1, 2, \ldots, k \tag{2.23}$$

Σε αυτό το μοντέλο το $\mathbf{t}$ είναι ένα διάνυσμα διάστασης q ενός υποσυνόλου q συμμεταβλητών για τις οποίες η υπόθεση των αναλογικών σχετικών συχνοτήτων δεν υποτίθεται εκ των προτέρων ούτε εξετάζεται αν ισχύει. Το $\boldsymbol{\gamma}_j$ είναι ένα διάνυσμα διάστασης q συντελεστών παλινδρόμησης για τις μεταβλητές στο $\mathbf{t}$ και έτσι ο όρος $\mathbf{t}^T\boldsymbol{\gamma}_j$ είναι η προσάυξηση που σχετίζεται μόνο με το j-οστό αθροιστικό logit (ισχύει $1 \leq j \leq k$ και $\gamma_1 = 0$) και για αυτό το μοντέλο ονομάστηκε μη περιορισμένο.

Εάν $\gamma_j = 0$ για όλα τα j τότε το μοντέλο (2.23) μετατρέπεται στο αναλογικό μοντέλο (2.20). Ένας έλεγχος για την υπόθεση των αναλογικών σχετικών συχνοτήτων για τις q μεταβλητές στο $\mathbf{t}$ βασίζεται στην υπόθεση $H_0 : \gamma_j = 0,\ \forall j \in (2 \leq j \leq k)$. Καθώς ισχύει $\gamma_1 = 0$, το μοντέλο χρησιμοποιεί μόνο την ποσότητα $(\alpha + \boldsymbol{\beta}^T\mathbf{x})$ για να εκτιμήσει το λόγο σχετικών πιθανοτήτων που σχετίζεται με τη διχοτόμηση της Y σε $y_j = 1$ έναντι της $y_j > 1$. Παρόλα αυτά η εκτίμηση των λόγων σχετικών πιθανοτήτων που σχετίζονται με τις υπόλοιπες αθροιστικές πιθανότητες περιλαμβάνει την αύξηση του $(\alpha + \boldsymbol{\beta}^T\mathbf{x})$ κατά $\mathbf{t}\boldsymbol{\gamma}_j$.

2.2.5.2 Περιορισμένο μερικώς αναλογικό μοντέλο σχετικών πιθανοτήτων

Όταν η σχέση μεταξύ μιας επεξηγηματικής μεταβλητής και μιας κατηγορίας απόκρισης δεν είναι αναλογική, συνήθως περιμένουμε την ύπαρξη κάποιου είδους τάσης. Αυτός ο τύπος μοντέλου που πρότειναν οι Peterson and Harrell (1990) επιτρέπει την ύπαρξη μιας γραμμικής σχέσης μεταξύ του logit για μια επεξηγηματική μεταβλητή και της κατηγορίας απόκρισης (Ananth, 1997).

Σε αυτή την περίπτωση μπορούν να εισαχθούν στο μοντέλο περιορισμοί για να συμπεριλάβουν αυτή τη γραμμική σχέση.

$$\Pr(Y \leq y_j \mid \mathbf{X}) = \frac{exp(-\alpha_j - \boldsymbol{\beta}^T\mathbf{X} - \mathbf{t}^T\boldsymbol{\gamma}\boldsymbol{\Gamma}_j)}{1 + exp(-\alpha_j - \boldsymbol{\beta}^T\mathbf{X} - \mathbf{t}^T\boldsymbol{\gamma}\boldsymbol{\Gamma}_j)},\ j = 1, 2, \ldots, k \tag{2.24}$$

όπου τα $\boldsymbol{\Gamma}_j$ είναι προκαθορισμένοι αριθμοί, σταθεροί για κάθε logit ($\Gamma_1 = 0$). Η παράμετρος γ είναι και πάλι είναι ένα διάνυσμα διάστασης q αλλά σε αυτή την περίπτωση δεν εξαρτάται από την κατηγορία j.

2.2.6 Ταυτόχρονη χρήση διατακτικών logit από τα διατακτικά μοντέλα

Για κάθε τύπο διατακτικού logit το οποίο εφαρμόζεται σε μια μεταβλητή c κατηγοριών, δημιουργούνται $c-1$ logit, τα οποία ενσωματώνονται σε ένα μοντέλο. Στην επόμενη ενότητα δείχνουμε πως αυτή η προσέγγιση δίνει πιο φειδωλά και απλούστερα στην ερμηνεία μοντέλα σε σχέση με την εφαρμογή $c - 1$ ξεχωριστών μοντέλων, ένα για κάθε logit.

Είδαμε ως τώρα μοντέλα για αθροιστικά logit και στη συνέχεια θα παρουσιαστούν μοντέλα και για άλλα διατακτικά logit. Κάθε μοντέλο έχει το δικό του λόγο σχετικών πιθανοτήτων για την περιγραφή των επιδράσεων. Για παράδειγμα,όπως θα δούμε και παρακάτω, αφού το logit μοντέλο των γειτονικών κατηγοριών χρησιμοποιεί ζευγάρια γειτονικών κατηγοριών, θα περιγράφονται χρησιμοποιώντας τοπικούς λόγους σχετικών πιθανοτήτων Agresti (2010).

2.3 Μοντέλα probit

Μετά το logit, η πιο συνηθισμένη συνάρτηση σύνδεσης για αθροιστικής συνάρτησης σύνδεσης μοντέλα είναι το probit, που είναι η αντίστροφη της αθροιστικής συνάρτησης κατανομής της τυποποιημένης κανονικής. Θα δούμε πως προκύπτει όταν η λανθάνουσα μεταβλητή μοντελοποιείται από ένα γραμμικό μοντέλο παλινδρόμησης για το οποίο η συνάρτηση κατανομής της λανθάνουσας μεταβλητής είναι η κανονική, με σταθερή διακύμανση.

Συνεχίζοντας με τις έννοιες από την ενότητα 2.2.1, υποθέτουμε πως υπάρχει μια λανθάνουσα μεταβλητή S με αθροιστική συνάρτηση σύνδεσης F. Τότε, η διατακτική μεταβλητή απόκρισης Y_i μπορεί να παρατηρηθεί στην κατηγορία j εάν η S_i είναι μεταξύ των κατωφλίων $\theta_{j-1} < S_i < \theta_j$ και ισχύει

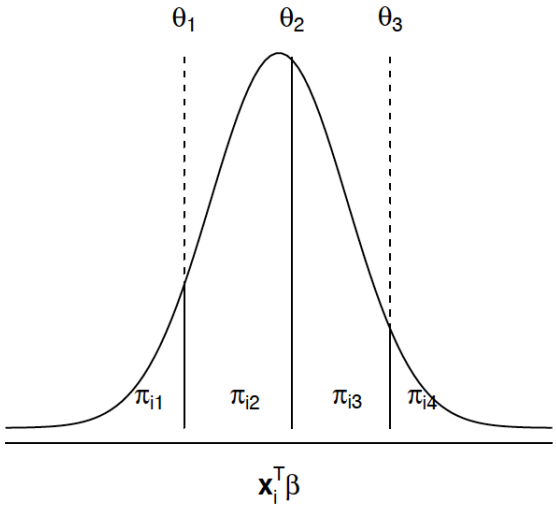
$$-\infty \equiv \theta_0 < \theta_1 < \cdots < \theta_{J-1} < \theta_J \equiv \infty \tag{2.25}$$

Έτσι η S βρίσκεται σε $J+1$ διαστήματα, το οποίο μπορούμε να διακρίνουμε και στην Εικόνα 2.4 για 4 κατηγορίες απόκρισης. Τα 3 κατώφλια χωρίζουν την περιοχή κάτω από την καμπύλη σε τέσσερις περιοχές, κάθε μια από τις οποίες αντιπροσωπεύει την πιθανότητα μια παρατήρηση να βρεθεί σε μια περιοχή. Τα όρια είναι σταθερά πάνω στην ευθεία του άξονα χ , ωστόσο η θέση της λανθάνουσας κατανομής και κατ επέκταση και των περιοχών κάτω από την καμπύλη αλλάζει μαζί με το $\mathbf{x}_i$.

Ένα γραμμικό μοντέλο για την λανθάνουσα μεταβλητή είναι το

$$S_i = \alpha + \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2) \tag{2.26}$$

όπου ε_i τυχαία σφάλματα και α είναι ο σταθερός όρος, δηλαδή η μέση τιμή του S_i όταν το $\mathbf{x}_i$ αντιστοιχεί σε ένα επίπεδο αναφοράς για κατηγορικούς παράγοντες και σε μηδέν για συνεχείς συμμεταβλητές. Μπορούμε να γράψουμε δηλαδή και $S_i \sim N(\alpha + \boldsymbol{\beta}^T \mathbf{x}_i, \sigma^2)$.



Σχήμα 2.4: Απεικόνιση ενος αθροιστικού μοντέλου σε σχέση με υποκείμενη μεταβλητή

Έτσι λοιπόν, η αθροιστική πιθανότητα μιας παρατήρησης να βρεθεί στην κατηγορία j ή και πιο κάτω είναι

$$\gamma_{ij} = \Pr(Y_i \leq j) = \Pr(S_i) \leq \theta_j = \Pr\left(Z_i \leq \frac{\theta_j - \alpha - \boldsymbol{\beta}^T \mathbf{x}_i}{\sigma}\right) \tag{2.27}$$

$$= \Phi\left(\frac{\theta_j - \alpha - \boldsymbol{\beta}^T \mathbf{x}_i}{\sigma}\right) \tag{2.28}$$

όπου $Z_i = \frac{S_i - \alpha - \boldsymbol{\beta}^T \mathbf{x}_i}{\sigma} \sim N(0,1)$ και Φ η συνάρτηση κατανομής της τυπικής κανονικής.

Καθώς η θέση (α) και η κλίμακα (σ) της λανθάνουσας μεταβλητής δεν μπορεί να προσδιοριστούν από τις παρατηρήσεις, θα ορίσουμε ένα μοντέλο που μπορούμε να παρατηρήσουμε ως εξής

$$\gamma_{ij} = \Phi(\theta_j - \boldsymbol{\beta}^T \mathbf{x}_i) \tag{2.29}$$

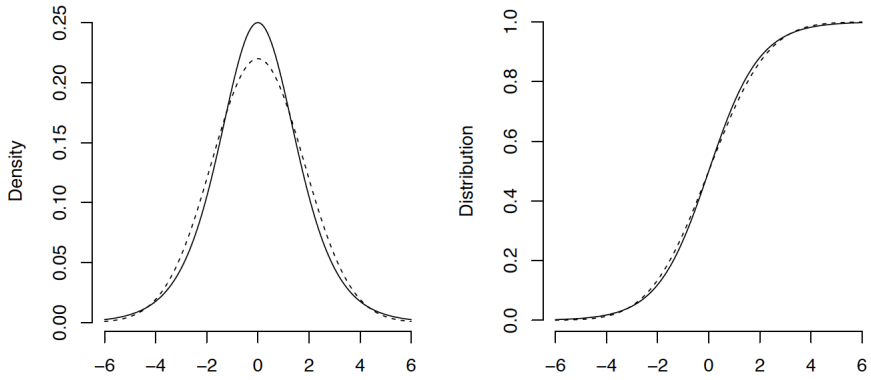
με εκτιμημένες παραμέτρους

$$\hat{\theta}_j = \frac{(\theta_j - \alpha)}{\sigma} \quad \text{και} \quad \hat{\boldsymbol{\beta}} = \frac{\boldsymbol{\beta}}{\sigma}. \tag{2.30}$$

Παρατηρούμε πως το μείον στη (2.29) διατηρείται και εδώ και δείχνει πως μια θετική εκτιμημένη παράμετρος β σημαίνει μια θετικής κατεύθυνσης μετατόπιση της κατανομής της λανθάνουσας μεταβλητής.

Το (2.29) είναι ένα αθροιστικό μοντέλο με συνάρτηση σύνδεσης την probit.

Η συνήθης μορφή της λογιστικής κατανομής έχει μέσο όρο 0 και διακύμανση $\pi^2/3$. Ακόμα, η κατανομή της είναι συμμετρική και μοιάζει με την κανονική, αλλά με πιο βαριές ουρές.



Σχήμα 2.5: Συνάρτ. πυκνότητας και κατανομή της λογιστικής με συνεχή γραμμή και κανονικής με διακεκομμένη, με μέσο όρο 0 και διακύμανση $\pi^2/3$.

2.4 Μοντέλα εναλλακτικής συνάρτησης σύνδεσης

Τα μοντέλα για διατακτικές κατηγορίες απόκρισης μπορούν και να μη χρησιμοποιούν τις αθροιστικές πιθανότητες. Σε αυτή την ενότητα θα δείξουμε τα εναλλακτικά μοντέλα logit που υπάρχουν.

2.4.1 Μοντέλα γειτονικών κατηγοριών (*adjacent categories models*)

Χρησιμοποιώντας αυτά που γράφτηκαν στην ενότητα 2.1.3 θα δούμε την κατηγορία των διατεταγμένων logit μοντέλων γειτονικών κατηγοριών (*adjacent-categories*). Θεωρούνται υποκατηγορία των μοντέλων βασικής κατηγορίας με μειωμένο αριθμό παραμέτρων μετά από αξιοποίηση της διάταξης για να βρεθεί μια κοινή επίδραση και δεν στηρίζονται στην ύπαρξη κάποιας λανθάνουσας, συνεχούς εξαρτημένης μεταβλητής. Επίσης, σε αυτή την κατηγορία μοντέλων, συγκρίνουμε κάθε απόκριση με την αμέσως μεγαλύτερη απόκριση.

Χρησιμοποιούν πιθανότητες μιας κατηγορίας αντί για αθροιστικές πιθανότητες και έτσι είναι πιο απλό να εξηγήσει κάποιος τις επιδράσεις σε όρους σχετικών πιθανοτήτων σχετιζόμενων με συγκεκριμένες κατηγορίες απόκρισης. Αυτά τα μοντέλα έλαβαν αναγνώριση τις δεκαετίες του 1980 και 1990, εν μέρει λόγω της σύνδεση με ορισμένα λογαριθμογραμ-

μικά μοντέλα. Για παράδειγμα ο Agresti (1992) μοντελοποίησε ζευγαρωμένες παρατηρήσεις προτίμησης επεκτείνοντας το μοντέλο των Bradley-Terry για διατακτικές αποκρίσεις. Άλλες εργασίες για τέτοια δεδομένα είναι των Böckenholt and Dillon (1997), Fahrmeir et al. (2001).

Σύμφωνα με τον Goodman (1983), Simon (1974) το μοντέλο γράφεται ως :

$$\text{logit}\left[\frac{\Pr(Y=j|\mathbf{x})}{\Pr(Y=j+1|\mathbf{x})}\right] = \log\frac{\pi_j}{\pi_{j+1}},\ j=1,\ldots,J-1. \tag{2.31}$$

Αυτά τα logit είναι ισοδύναμα με τα logit μοντέλα γειτονικής κατηγορίας τα οποία χρησιμοποιούνται για ονομαστικές μεταβλητές απόκρισης. Οι συνδέσεις είναι οι ακόλουθες :

$$\log\frac{\pi_j}{\pi_J} = \log\frac{\pi_j}{\pi_{j+1}} + \log\frac{\pi_j}{\pi_{j+2}} + \cdots + \log\frac{\pi_{j-1}}{\pi_J} \tag{2.32}$$

και

$$\log\frac{\pi_j}{\pi_{j+1}} = \log\frac{\pi_j}{\pi_J} - \log\frac{\pi_{j+1}}{\pi_J},\ j=1,\ldots,J-1 \tag{2.33}$$

Κάθε σετ καθορίζει τα logit για όλα τα $\binom{J}{2}$ ζευγάρια κατηγοριών απόκρισης. Η σχέση των logit γειτονικής κατηγορίας (2.31) μπορεί να γραφτεί ως:

$$\log\frac{\pi_j}{\pi_{j+1}} = \alpha_j + \boldsymbol{\beta}^T\mathbf{X},\ j=1,\ldots,J-1 \tag{2.34}$$

με ένα κοινό συντελεστή $\boldsymbol{\beta}$. Αθροίζοντας $J-j$ όρους όπως στην σχέση (2.32) μπορούμε να πάρουμε το αντίστοιχο μοντέλο βασικής κατηγορίας.

$$\log\frac{\pi_j(\mathbf{x})}{\pi_J(\mathbf{x})} = \sum_{k=j}^{J-1}\alpha_k + \boldsymbol{\beta}^T(J-j)\mathbf{x},\ j=1,\ldots,J-1 \tag{2.35}$$

το οποίο μπορεί να γραφεί ως εξής

$$\log\frac{\pi_j(\mathbf{x})}{\pi_J(\mathbf{x})} = \alpha_j^* + \boldsymbol{\beta}^T\mathbf{u}_j,\ j=1,\ldots,J-1 \tag{2.36}$$

Το logit μοντέλο γειτονικών κατηγοριών αντιστοιχεί σε ένα logit βασικής κατηγορίας με προσαρμοσμένο πίνακα αλλά και με ξεχωριστή παράμετρο για κάθε μεταβλητή πρόβλεψης (Agresti, 2002). Έτσι, το μοντέλο (2.34) μπορεί να εφαρμοστεί και εφαρμόζοντας το αντίστοιχο μοντέλο βασικής κατηγορίας.

Η κατασκευή των μοντέλων γειτονικής κατηγορίας αναγνωρίζει την διάταξη των κατηγοριών της Y. Για να υπάρχει ωφέλεια από αυτό το οικονομικότερο μοντέλο χρειάζεται κατάλληλος καθορισμός του γραμμικού μοντέλου. Εάν για παράδειγμα μια επεξηγηματική

μεταβλητή έχει παρόμοια επίδραση σε κάθε logit, τότε μπορούμε να περιγράψουμε την επίδραση αυτή με μια παράμετρο αντί για $J-1$ παραμέτρους.

Εάν χρησιμοποιηθεί με αυτή την μορφή αναλογικών σχετικών συχνοτήτων, το μοντέλο (2.34) με logit γειτονικών κατηγοριών μπορεί να εφαρμοστεί καλά σε παρόμοιες καταστάσεις με το μοντέλο (2.21) με αθροιστικά logit.

Η επιλογή του κατάλληλου μοντέλου θα πρέπει να στηρίζεται λιγότερο στην καλή προσαρμογή του μοντέλου και περισσότερο στην επιλογή μοντελοποίησης επιδράσεων σε ξεχωριστές κατηγορίες της μεταβλητής απόκριση, όπως κάνουν τα μοντέλα γειτονικών κατηγοριών ή αντίθετα σε ομάδες κατηγοριών που κάνουν χρήση υποκείμενης μεταβλητής ή ολόκληρης κλίμακας, όπως τα αθροιστικά logit. Πάντως οι επιδράσεις στα αθροιστικά logit είναι μεγαλύτερες αφού χρησιμοποιούν όλο το εύρος της μεταβλητής απόκρισης. Ένα πλεονέκτημα των αθροιστικών logit είναι η σχετική σταθερότητα των εκτιμήσεων των επιδράσεων στην επιλογή και τον αριθμό των κατηγοριών απόκρισης, πράγμα που δε συμβαίνει με τα logit γειτονικών κατηγοριών.

Τα logit μοντέλα αθροιστικής πιθανότητας και γειτονικής κατηγορίας υποδηλώνουν στοχαστική διάταξη των κατανομών απόκρισης για διαφορετικές τιμές πρόβλεψης. Οι επιδράσεις σε ένα logit μοντέλο γειτονικής κατηγορίας αφορούν στην επίδραση που έχει αύξηση μιας μονάδας μιας ανεξάρτητης προβλεπτικής μεταβλητής στο λογάριθμο της σχετικής πιθανότητας της απόκρισης, στη μικρότερη αντί για τη μεγαλύτερη από οποιεσδήποτε δυο γειτονικές κατηγορίες, ενώ η επίδραση σε ένα logit μοντέλο αθροιστικής πιθανότητας όπως το περιγράφει ο McCullagh, δηλαδή το μοντέλο αναλογικών σχετικών πιθανοτήτων αναφέρεται σε ολόκληρη την κλίμακα αποκρίσεων. Όταν η μεταβλητή απόκρισης έχει δυο κατηγορίες μόνο, τότε τα μοντέλα αθροιστικής πιθανότητας και γειτονικής κατηγορίας απλοποιούνται στο κανονικό μοντέλο λογιστικής παλινδρόμησης.

2.4.2 Λόγος συνέχειας (*Continuation Ratio*)

Τα logit λόγου συνέχειας (*continuation-ratio*), είναι τα logit των δεσμευμένων πιθανοτήτων πως μια παρατήρηση θα βρίσκεται στην *j*-οστή κατηγορία, δεδομένου ότι βρίσκεται τουλάχιστον στην *j*-οστή κατηγορία για $j = 1, , J-1$ ($\Pr(Y = j \mid Y \geq j)$) και ορίζονται Fienberg (2007)

$$\log \frac{\pi_j}{\pi_{j+1} + \cdots + \pi_J}, \; j = 1, \ldots, J-1 \tag{2.37}$$

ή αλλιως και ως

$$\log \frac{\pi_{j+1}}{\pi_1 + \cdots + \pi_j},\ j = 1, \ldots, J-1. \tag{2.38}$$

Τα μοντέλα αυτής της παραγράφου είναι χρήσιμα όταν ένας ακολουθιακός μηχανισμός καθορίζει την μεταβλητή απόκρισης, όπως η επιβίωση μέσα από διάφορες ηλικιακές περιόδους Tutz (1990, 1991). Όταν το μοντέλο περιλαμβάνει ξεχωριστές επιδράσεις ${\boldsymbol{\beta_j}}^T$, η πολυωνυμική πιθανοφάνεια μετατρέπεται σε ένα γινόμενο διωνυμικών πιανοφανειών για τα ξεχωριστά logit για κάθε j. Τότε, η ξεχωριστή εφαρμογή μοντέλων για διαφορετικά continuation-ratio logit δίνει τα ίδια αποτελέσματα με την ταυτόχρονη εφαρμογή των μοντέλων. Οι McCullagh (1980), Thompson and Baker (1981) αντιμετώπισαν το μοντέλο αναλογικών σχετικών πιθανοτήτων ως μια ειδική περίπτωση του πολυμεταβλητού γενικευμένου γραμμικού μοντέλου

$$g(\boldsymbol{\mu_i}) = \boldsymbol{\beta}^T \mathbf{X}_i \tag{2.39}$$

όπου $\boldsymbol{\mu_i}$ είναι ο μέσος όρος ενός διανύσματος indicator αποκρίσεων $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{i,j-1})$ για την παρατήρηση i (1 είναι η τιμή της dummy μεταβλητής για την κατηγορία στην οποία βρίσκεται η παρατήρηση) και $\mathbf{g}$ είναι ένα διάνυσμα συναρτήσεων σύνδεσης. Τα logit μοντέλα γειτονικής κατηγορίας και continuation ratio μπορούν να συμπεριληφθούν σε αυτή την κατηγορία μοντέλων όπως ανέφεραν πιο λεπτομερώς οι Fahrmeir et al. (2001).

Κεφάλαιο 3

Ερμηνεία μοντέλων

Ο σχεδιασμός των γενικευμένων γραμμικών μοντέλων είναι τέτοιος ώστε η επίδραση κάθε x πάνω στην συνάρτηση σύνδεσης η είναι πάντα γραμμική. Έτσι η ερμηνεία των εκτιμημένων παραμέτρων ως γραμμικές επιδράσεις στην προβλέπουσα συνάρτηση η θα είναι κοινή σε όλα τα γενικευμένα γραμμικά μοντέλα. Για τα logit και probit μοντέλα συγκεκριμένα, η γραμμική επίδραση που δίνει μια εκτιμημένη παράμετρος στην η δείχνει την επίδραση του αντίστοιχου x στην logit $\left[\log\left(\frac{\mu}{1-\mu}\right)\right]$ και στην probit $[\Phi^{-1}(\mu)]$. Κάθε εκτίμηση δίνει την μερική επίδραση ενός συντελεστή θεωρώντας τις επιδράσεις των άλλων μεταβλητών x σταθερών.

Παρατηρώντας τις παραμέτρους π.χ (2.30) στα μοντέλα αθροιστικής συνάρτησης βλέπουμε πως είναι αναλογίες σήματος-θορύβου. Αυτό σημαίνει, πως προσθέτοντας μια συμμεταβλητή σε ένα μοντέλο αθροιστικής συνάρτησης σύνδεσης που μειώνει το θόρυβο των καταλοίπων θα έχει ως αποτέλεσμα την αύξηση της αναλογίας σήματος θορύβου. Αυτό έχει ως αποτέλεσμα συνήθως όταν προσθέτονται συμεταβλητές να αυξάνονται οι εκτιμήσεις των συντελεστών των υπόλοιπων συμεταβλητών. Στα γραμμικά μοντέλα από την άλλη σε ορθογονικούς σχεδιασμούς προσθέτοντας μια μεταβλητή δεν επηρεάζει τις τιμές των άλλων συντελεστών. Ο Bauer (2009) επεκτείνοντας το έργο των Winship, C. & Mare (1984) προτείνει ένα τρόπο να αλλάξει την κλίμακα των συντελεστών έτσι ώστε να είναι συγκρίσιμοι κατά τη διάρκεια της κατασκευής του μοντέλου.

Ανεξάρτητα από το πως τα κατώφλια διαχωρίζουν σε διακριτές περιοχές την κλίμακα της λανθάνουσας μεταβλητής, οι παράμετροι παλινδρόμησης β έχουν την ίδια ερμηνεία, ανεξαρτήτως του αριθμού των κατηγοριών. Επιπλέον, οι ερμηνείες δε θα αλλάξουν ούτε εάν συγχωνευτούν δυο ή παραπάνω κατηγορίες, παρόλο που οι εκτιμήσεις των παραμέτρων θα διαφέρουν. Έτσι, οι παράμετροι από διαφορετικές έρευνες με διαφορετικό αριθμό κατηγοριών απόκρισης μπορούν να συγκριθούν, όσο το επιτρέπει το επίπεδο του θορύβου. Ο Farewell (1982) συζητά για την περίπτωση που τα κατώφλια διαφέρουν ανάλογα το πως

διαφορετικά υποκείμενα της έρευνας αντιλαμβάνονται τα όρια των κατηγοριών. Αυτό το πρόβλημα πλέον μπορεί να αντιμετωπιστεί με την εισαγωγή τυχαίων επιδράσεων στο μοντέλο.

3.1 Ερμηνεία των εκτιμημένων παραμέτρων

Λόγω της ομοιότητας μεταξύ των διατεταγμένων probit και logit μοντέλων, θα δοθεί παράλληλα η ερμηνεία των αποτελεσμάτων αυτών των μοντέλων. Θα χρησιμοποιηθεί ως βάση για την ερμηνεία ένα probit μοντέλο και όπου είναι εφαρμόσιμο θα δοθεί η ερμηνεία της logit περίπτωσης.

3.1.1 Περιθωριακή (*marginal*) επίδραση στην συνάρτηση σύνδεσης η

Θα χρησιμοποιήσουμε logit μοντέλα επειδή είναι πιο εύκολη η ερμηνεία των σε αυτού του είδους την ερμηνεία. Εάν ισχύει η υπόθεση των αναλογικών σχετικών συχνοτήτων, η ερμηνεία των εκτιμημένων παραμέτρων είναι παρόμοια με εκείνη που ισχύει για δίτιμα logit μοντέλα και η μερική επίδραση του x είναι αμετάβλητη ασχέτως της επιλογής κατηγορίας απόκρισης, όπως συζητήσαμε και στην ενότητα 2.2.4, όπου τα εκτιμημένα β είναι ίδια ανεξάρτητα της κατηγορίας που επιλέγεται. Για δίτομες περιπτώσεις, η περιθωριακή επίδραση της x_k ερμηνεύεται ως η αναμενόμενη αλλαγή στις σχετικές πιθανότητες να ανήκει μια παρατήρηση στην κατηγορία 1 παρά στην κατηγορία 2, που είναι η πολλαπλασιαστική επίδραση της e^{β}, δοθέντος αλλαγής μιας μονάδας της x_k.

Προσαρμόζοντας την εξίσωση (2.20) για δυο κατηγορίες η πιθανότητα $\Pr(Y \leq y_1)$ ισούται με $\Pr(Y = y_1)$. Κάτι τέτοιο όμως δεν είναι δυνατόν να ισχύει για τρεις ή παραπάνω κατηγορίες. Για τρεις κατηγορίες, η εξίσωση (2.20) υποδεικνύει πως η επίδραση της x_k θα προκαλέσει αλλαγή στις σχετικές πιθανότητες να ανήκει η απόκριση στην κατηγορία 1 αντί στην 2 και 3, ή στην 1 ή στην 2 αντί στην 3 κατά e^{β}. Η αντίθεση θα είναι πάντα μεταξύ της πιθανότητας να ανήκει μια παρατήρηση στην πρώτη έως και τη *j*-οστή κατηγορία και της πιθανότητας να ανήκει στις υπόλοιπες κατηγορίες.

Κεφάλαιο 4

Αξιολόγηση μοντέλων

4.1 Εκτίμηση μέγιστης πιθανοφάνειας μοντέλων αθροιστικής συνάρτησης σύνδεσης

Η πρώιμη βιβλιογραφία πρότεινε για την εκτίμηση των μοντέλων στάθμιση ελάχιστων τετραγώνων (πχ Williams and Grizzle (1972)) αλλά πλέον τα μοντέλα αθροιστικής συνάρτησης σύνδεσης και όχι μόνο εκτιμούνται συνήθως με τη μέθοδο της μέγιστης πιθανοφάνειας. Οι (Walker and Duncan, 1967) και (McCullagh, 1980) χρησιμοποιούσαν αλγορίθμους σκορ του Fischer για να εκτιμήσουν την μέγιστη πιθανοφάνεια. Με αυτή την μέθοδο, μπορούν να υπολογιστούν τεστ σημαντικότητας και διαστήματα εμπιστοσύνης για τις παραμέτρους $\boldsymbol{\beta}$ του μοντέλου βασιζόμενα σε λόγους πιθανοφανειών, σκορ, ή στατιστικών συναρτήσεων του Wald. Η λογαριθμική συνάρτηση πιθανοφάνειας μπορεί να γραφτεί ως εξής

$$\ell(\boldsymbol{\theta}, \boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{n} w_i \log \pi_i \tag{4.1}$$

όπου i είναι ο δείκτης για βαθμωτές παρατηρήσεις (όχι πολυωνυμικές διανυσματικές παρατηρήσεις) , w_i είναι πιθανά βάρη για τις κατηγορίες και π_i είναι η πιθανότητα η *i*-οστή παρατήρηση να βρεθεί στην κατηγορία που βρέθηκε, δηλαδή είναι τα μη μηδενικά στοιχεία όπου ισχύει $\pi_{ij} I(Y_i = j)$ με $I(\cdot)$ να συμβολίζει τη δείκτρια συνάρτηση που παίρνει την τιμή 1 εάν η συνθήκη ισχύει και 0 εάν δεν ισχύει. Οι εκτιμητές μέγιστης πιθανοφάνειας (ΕΜΠ) των παραμέτρων , $\hat{\boldsymbol{\beta}}$ και $\hat{\boldsymbol{\theta}}$ είναι οι τιμές του $\boldsymbol{\beta}$ και $\boldsymbol{\theta}$ που μεγιστοποιούν τον λογάριθμο της πιθανοφάνειας στην (4.1).

Στις περιπτώσεις που τα δεδομένα δεν μπορούν να ομαδοποιηθούν σε έναν πίνακα και μια συνεχής μεταβλητή παίρνει μια ξεχωριστή τιμή για κάθε παρατήρηση, κάθε σειρά του επακόλουθου πίνακα θα περιέχει μόνο το 1 και 0 για τις υπόλοιπες παρατηρήσεις. Τα βάρη σε αυτή την περίπτωση εκτός και εάν ζητείται αλλιώς, θα παίρνουν την τιμή 1.

Σύμφωνα με τη γνωστή θεωρία πιθανοφάνειας ο πίνακας διακυμάνσεων-συνδιακυμάνσεων των παραμέτρων μπορεί να βρεθεί από τον αντίστροφο πίνακα της πληροφορίας Fischer. Αυτός ο πίνακας μπορέι να δοθεί από την Εσσιανή μήτρα του λογαρίθμου της συνάρτησης πιθανοφάνειας [1] υπολογισμένης στις εκτιμήτριες μέγιστης πιθανοφάνειας. Τα τυπικά σφάλματα μπορούν να βρεθούν ως η τετραγωνική ρίζα της διαγωνίου στον πίνακα διακυμάνσεων-συνδιακυμάνσεων.

Αν οριστεί ως $\boldsymbol{\alpha} = [\boldsymbol{\theta}, \boldsymbol{\beta}]$ το σύνολο των παραμέτρων, τότε η Εσσιανή μήτρα δίνεται ως η δεύτερη παράγωγος του λογαρίθμου της συνάρτησης πιθανοφάνειας για τους εκτιμητές μέγιστης πιθανοφάνειας:

$$\boldsymbol{H} = \left. \frac{\partial^2 \ell(\boldsymbol{\alpha}; \mathbf{y})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} \right|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}} . \tag{4.2}$$

Τότε παρατηρηθέν πίνακας του πίνακα πληροφορίας Fischer είναι ο $\boldsymbol{I}(\hat{\boldsymbol{\alpha}}) = -\boldsymbol{H}$ και τα τυπικά σφάλματα δίνονται από

$$\mathrm{s.e}(\hat{\boldsymbol{\alpha}}) = \sqrt{\mathrm{diag}[-\boldsymbol{H}](\hat{\boldsymbol{\alpha}})^{-1}]}. \tag{4.3}$$

Ένας άλλος τρόπος για να βρεθεί ο πίνακας διακυμάνσεων-συνδιακυμάνσεων των παραμέτρων είναι να χρησιμοποιηθεί ο αναμενόμενος πίνακας της πληροφορίας Fischer. Η επιλογή για το εάν θα χρησιμοποιηθεί ο παρατηρηθέν η ο αναμενόμενος πίνακας της πληροφορίας Fischer συνήθως καθορίζεται από τον αλγόριθμο εφαρμογής του μοντέλου. Οι μέθοδοι επαναληπτικής στάθμισης ελαχίστων τετραγώνων συνήθως παράγουν τον αναμενόμενο πίνακα πληροφορίας Fischer ως υποπροϊόν του αλγορίθμου και οι αλγόριθμοι Newton-Raphson (όπως αυτός που χρησιμοποιείται στη συνάρτηση **clm** στο πακέτο **ordinal** στην **R**) παράγουν τον παρατηρηθέν πίνακα της πληροφορίας Fischer. Οι Efron and Hinkley (1978) εξέτασαν την επιλογή ανάμεσα στην χρήση του παρατηρηθέντος και του αναμενόμενου πίνακα πληροφορίας του Fischer και υποστήριξαν πως ο πίνακας με τις παρατηρημένες πληροφορίες περιέχουν πιο συναφείς πληροφορίες και θα πρέπει να προτιμούνται σε σχέση με τις εκτιμημένες.

Ο Pratt (1981) έδειξαν πως η λογαριθμική συνάρτηση πιθανοφάνειας των μοντέλων αθροιστικής συνάρτησης σύνδεσης που χρησιμοποιούν την logit, probit, log-log, clog-log είναι κοίλη. Αυτό σημαίνει πως έχουν ένα μοναδικό ολικό μέγιστο και δεν υπάρχει κίνδυνος σύγκλισης σε τοπικό μέγιστο. Ακόμα, κάθε βήμα ενός επαναληπτικού αλγορίθμου Newton-Raphson θα βρίσκει σίγουρα μεγαλύτερη πιθανοφάνεια, αν και το βήμα μπορεί να είναι πολύ μεγάλο για να προκαλέσει αύξηση στην πιθανοφάνεια. Σε αυτή την περίπτωση εάν μειώσουμε στο μισό το βήμα αυτό, είναι εγγυημένη ουσιαστικά η σύγκλιση.

[1] ισοδύναμα η Εσσιανή της αρνητικής λογαριθμικής πιθανοφάνειας

4.2 Απόκλιση και σύγκριση μοντέλων

4.2.1 Σύγκριση μοντέλων με ελέγχους λόγου πιθανοφανειών

Ένας τρόπος για να συγκριθούν διαφορετικά μοντέλα μεταξύ τους, είναι μέσω της στατιστικής συνάρτησης του λόγου πιθανοφανειών. Για δυο μοντέλα m_0 και m_1, όπου το m_0 εμπεριέχεται στο m_1, δηλαδή το m_0 είναι πιο απλό από το m_1, η στατισιτκή συνάρτηση του λόγου πιθανοφανειών για τη σύγκριση των m_0 και m_1 είναι η

$$LR = -2(\ell_0 - \ell_1) \tag{4.4}$$

όπου ℓ_0 και ℓ_1 έιναι ο λογάριθμος της πιθανοφάνειας της m_0 και m_1 αντίστοιχα. Η στατιστική αυτή συνάρτηση υπολογίζει την επιπλεόν περιπλοκότητα στο m_1 μοντέλο έναντι στο m_0. Επίσης ακολουθεί ασυμπτωτικά μια κατανομή χ^2 με βαθμούς ελευθερίας ίσους με τη διαφορά στον αριθμό των παραμέτρων των m_0 και m_1. Τα τεστ των λόγων πιθανοφανειών γενικά είναι πιο ακριβή από τα τεστ με τη στατιστική συνάρτηση του Wald.

4.2.2 Απόκλιση και πίνακες απόκλισης

Στα γραμμικά μοντέλα οι πίνακες διακύμανσης (ANOVA) και τα F-τεστ βασίζονται στην διάσπαση των τετραγωνικών αθροισμάτων. Η εξέταση των τετραγωνικών αθροισμάτων δεν έχει μεγάλη χρησιμότητα για κατηγορικές παρατηρήσεις οπότε για τα γενικευμένα γραμμικά μοντέλα και για πίνακες συνάφειας χρησιμοποιείται η απόκλιση για τη σύγκριση μοντέλων και τη δημιουργία πινάκων απόκλισης (ANODE-analysis of deviance). Οι (McCullagh and Nelder, 1989) αναφέρουν πως η απόκλιση είναι στενά συνδεδεμένη με το άθροισμα τετραγώνων για γραμμικά μοντέλα.

Στην προηγούμενη σχέση (4.4), το m_0 και m_1 αντιπροσωπεύουν το μειωμένο μοντέλο και το πλήρες (ή κορεσμένο) αντίστοιχα. Το πλήρες μοντέλο περιλαμβάνει μια παράμετρο για κάθε παρατήρηση και έτσι μπορεί και περιγράφει πλήρως τα δεδομένα μας χωρίς καμία αμφιβολία ενώ το μειωμένο μοντέλο περιέχει μια πιο συνοπτική περιγραφή των δεδομένων με σαφώς λιγότερες παραμέτρους.

Μια ακόμα υποπερίπτωση μοντέλου είναι το μηδενικό μοντέλο οποίο δεν περιγράφει καμία δομή στα δεδομένα πέρα από τον σταθερό όρο. Η απόκλιση που αντιστοιχεί σε αυτό το μοντέλο λέγεται μηδενική απόκλιση και είναι ανάλογη με το συνολικό άθροισμα τετραγώνων για γραμμικά μοντέλα και για αυτό μπορεί να συναντηθεί και ως συνολική απόκλιση. Συναντάται επίσης και η απόκλιση των καταλοίπων που θα μπορούσε να παρομοιαστεί με το άθροισμα τετραγώνων των καταλοίπων στην περίπτωση των γραμμικών μοντέλων και ορίζεται ως

$$D_{\text{καταλ}} = D_{\text{ολική}} - D_{\text{μειωμένη}} \tag{4.5}$$

Η διαφορά στην απόκλιση μεταξύ δυο εμφωλευμένων μοντέλων είναι ίδια με τη στατιστική συνάρτηση του λόγου πιθανοφανειών για την σύγκριση των μοντέλων. Έτσι και αυτή όπως και ο λόγος πιθανοφανειών, ακολουθεί ασυμπτωτικά μια κατανομή χ^2 με βαθμούς ελευθερίας ίσους με τη διαφορά στον αριθμό των παραμέτρων των δυο μοντέλων. Η απόκλιση στην (4.4) είναι η στατιστική συνάρτηση του λόγου πιθανοφανειών για την σύγκριση του πλήρους και του μειωμένου μοντέλου.

Η πιθανοφάνεια για ένα μειωμένο μοντέλο είναι δυνατό να βρεθεί από τις εκτιμήσεις των αθροιστικών μοντέλων σύνδεσης αλλά δεν είναι δυνατό να εκφραστεί το πλήρες μοντέλο ως ένα αθροιστικό μοντέλο σύνδεσης, έτσι ο λογάριθμος της πιθανοφάνειας του πλήρους μοντέλου θα πρέπει να βρεθεί με άλλο τρόπο. Για ένα πίνακα συνάφειας με *h* γραμμές και *j* στήλες, ο λογάριθμος της πιθανοφάνειας του πλήρους μοντέλου σε αντιστοιχία με την πιθανοφάνεια στην (4.1) δίνεται από

$$\ell_{\text{full}} = \sum_h \sum_j w_{hj} \log \hat{\pi}_{hj} \tag{4.6}$$

όπου: $\hat{\pi}_{hj} = \frac{w_{hj}}{w_{h.}}$, w_{hj} το πλήθος των παρατηρήσεων στο (*h,j*)-οστό κελί και $w_{h.}$ το άθροισμα των παρατηρήσεων στη γραμμή h.

4.2.3 Έλεγχος καλής προσαρμογής με την απόκλιση

Καθώς η απόκλιση ακολουθεί ασυμπτωτικά την κατανομή χ_2 όπως αναφέρθηκε και παραπάνω, μπορεί να χρησιμοποιηθεί για την αξιολόγηση καλής προσαρμογής ενός μειωμένου μοντέλου. Η ασυμπτωτική χρήση της κατανομής χ_2 θεωρείται καλή πρακτική εάν οι αναμενόμενες συχνότητες του μειωμένου μοντέλου δεν είναι πολύ μικρές (άτυπος κανόνας θεωρείται μεγαλύτερες από 5).

Ένα πρόβλημα με την απόκλιση για ένα μειωμένο μοντέλο είναι πως αυτή εξαρτάται από την επιλογή του πλήρους μοντέλου, καθώς συνήθως ο τρόπος που σχηματίζεται ένας πίνακας επηρεάζει και την συνολική απόκλιση του μοντέλου. Οι διαφορές στην απόκλιση για εμφωλευμένα μοντέλα είναι ανεξάρτητα της πιθανοφάνειας του πλήρους μοντέλου και έτσι οι διαφορές αυτές δεν επηρεάζονται από την επιλογή του πλήρους μοντέλου και μπορούν επομένως να χρησιμοποιηθούν για την αξιολόγηση των μοντέλων. Ο Collett (2002) προτείνει να συγχωνεύονται τα δεδομένα όσο το δυνατόν περισσότερο για την εκτίμηση των αποκλίσεων και της καλής προσαρμογής.

4.3 Μέτρα αξιολόγησης καλής προσαρμογής

Κάθε μοντέλο παλινδρόμησης στοχεύει στην περιγραφή σχέσεων μεταξύ μιας μεταβλητής απόκρισης και διάφορων συμμεταβλητών. Χρησιμοποιείται ένα συστηματικό μέρος και ένα μέρος σφάλματος. Το μέρος του σφάλματος είναι αυτό που συνιστά την απόκλιση των πραγματικών δεδομένων από το συστηματικό κομμάτι. Εάν η απόκλιση αυτή είναι αρκετά μεγάλη τότε το συγκεκριμένο μοντέλο δεν προσαρμόζεται ικανοποιητικά στα δεδομένα και δεν μπορεί να τα περιγράψει επαρκώς, πράγμα που οδηγεί σε αμφισβητήσιμα συμπεράσματα. Έτσι, η αξιολόγηση της καλής προσαρμογής έχει μεγάλη σημασία στην διαδικασία σχηματισμού μοντέλων και θα πρέπει να πραγματοποιείται πριν τον στατιστικό έλεγχο υποθέσεων. Σημαντικές μέθοδοι για την αξιολόγηση της προσαρμογής των μοντέλων παλινδρόμησης είναι τα κατάλοιπα και συγκρίσεις των παρατηρηθέντων τιμών απόκρισης με τις αντίστοιχες εκτιμημένες τιμές. Η αξιολόγηση της καλής προσαρμογής έχει δυο κύρια μέρη, την ολική και την επιμέρους αξιολόγηση καλής προσαρμογής. Ακόμα και όταν η ολική προσαρμογή του μοντέλου είναι επαρκής, υπάρχει η περίπτωση οι επιμέρους τιμές να μην έχουν καλή προσαρμογή.

Η επιλογή της κατάλληλης μεθόδου για την αξιολόγηση της καλής προσαρμογής εξαρτάται από το μοντέλο που χρησιμοποιείται. Οι Harrell (2001) παρουσιάζουν μια σύνοψη στην ανάπτυξη μοντέλων πολλαπλής παλινδρόμησης και αξιολόγησης των υποθέσεων των μοντέλων καθώς και αξιολόγησης καλής προσαρμογής. Στην περίπτωση του δίτιμου λογιστικού μοντέλου όλες οι μέθοδοι αξιολόγησης της καλής προσαρμογής συγκρίνουν τις παρατηρούμενες δεσμευμένες πιθανότητες με τις αντίστοιχες εκτιμημένες πιθανότητες. Εάν υπάρχουν μόνο κατηγορικές μεταβλητές και συνεπώς μικρός αριθμός συνδυασμών συμμεταβλητών, η ολική αξιολόγηση καλής προσαρμογής μπορεί να εξετασθεί από γνωστές μεθόδους όπως ο έλεγχος χ^2 του Pearson και το τεστ με τη στατιστική συνάρτηση του λόγου πιθανοφανειών. Όμως εάν ο αριθμός των συνδυασμών των συμμεταβλητών είναι μεγάλος και επακόλουθα ο αριθμός των αναπαραγόμενων μετρήσεων είναι μικρός, καθιστώντας τις μεθόδους αυτές άκυρες καθώς αυτές χρειάζονται μεγάλο αριθμό μετρήσεων . Να σημειωθεί πως αυτές οι μέθοδοι δεν μπορούν να χρησιμοποιηθούν για συνεχείς μετρήσεις και σε αυτή την περίπτωση μπορούν να χρησιμοποιηθούν τα τεστ των Hosmer και Lemeshow καθώς και του Brown τα οποία συγκεντρώνουν τις παρατηρήσεις ανάλογα με τις εκτιμημένες πιθανότητες.

Για τα πολυωνυμικά και διατακτικά λογιστικά μοντέλα παλινδρόμησης δεν υπάρχει διαθέσιμος κάποιος τρόπος στα διαθέσιμα λογισμικά, οπότε πριν εφαρμοστεί το *proportional odds model*, θα πρέπει να εξεταστούν τα δίτιμα λογιστικά μοντέλα παλινδρόμησης για κάθε διχοτομημένη απόκριση. Λόγω των αυστηρών προϋποθέσεων εφαρμογής το *proportional odds model* δεν είναι η σωστή μέθοδος για να ξεκινήσει μια έγκυρη ανάλυση (Greenland,

1994). Μόνο εάν τα ξεχωριστά δίτιμα μοντέλα είναι έγκυρα και τηρούν τις προϋποθέσεις εφαρμογής είναι δυνατόν να εξεταστεί η επάρκεια του *proportional odds model*. Η αναλογική υπόθεση στο *proportional odds model* μπορεί να εξεταστεί με ένα σκόρ τεστ Peterson and Harrell (1990). Ακόμα, στην αξιολόγηση της εγκυρότητας της αναλογικής υπόθεσης του *proportional odds model* μπορεί να χρησιμοποιηθεί η μοντελοποίηση των διχοτομημένων αποκρίσεων Brant (1990).

Εάν το αθροιστικό μοντέλο logit δεν έχει καλή εφαρμογή τότε μπορούν να χρησιμοποιηθούν ξεχωριστές επιδράσεις, αντικαθιστώντας στη σχέση (2.20) το β με το β_j. Αυτό δίνει μη παράλληλες γραμμές για τα $c-1$ logit. Ένα τέτοιο μοντέλο δεν μπορεί να ισχύει για μεγάλο εύρος τιμών των επεξηγηματικών μεταβλητών καθώς όταν τέμνονται οι γραμμές για τα logit είναι ένδειξη πως οι αθροιστικές πιθανότητες δεν είναι πλέον σε τάξη. Σε αυτή την περίπτωση χρησιμοποιούνται τα προαναφερθέντα τεστ των Peterson and Harrell (1990) και Brant (1990) όπου συγκρίνονται παρατηρηθείσες με εκτιμημένες μετρήσεις για μια διαμέριση των πιθανών τιμών απόκρισης.

Όταν η υπόθεση περί αναλογικότητας του μοντέλου αθροιστικών πιθανοτήτων κρίνεται ανεπαρκής μπορούν να δοκιμαστούν ως εναλλακτικές i) να χρησιμοποιηθούν διαφορετικές συναρτήσεις σύνδεσης, ii) προσθήκη επιπλέον όρων, όπως αλληλεπιδράσεων, iii) γενίκευση του μοντέλου προσθέτοντας παραμέτρους διακύμανσης Cox (1995), McCullagh (1980), iv) εφαρμόζοντας διαφορετικές επιδράσεις σε κάθε logit για κάποιες από τις ερμηνευτικές μεταβλητές(partial proportional odds Peterson and Harrell (1990)) v) χρησιμοποιώντας το κανονικό μοντέλο για ονομαστικές αποκρίσεις και το οποίο εφαρμόζει logit βασικής κατηγορίας αντιστοιχίζοντας κάθε κατηγορία με μια βάση .

Ακόμα όταν δεν ισχύει η υπόθεση των περί αναλογικότητας μπορεί να χρησιμοποιηθεί το continuation ratio logit (Cox, 1988) (ενότητα 2.4.2) ενώ εάν παραβιάζονται οι υποθέσεις τόσο των λόγων συνέχειας όσο και των αναλογικών σχετικών όρων μπορούν να ακολ

Υπάρχει ποικιλία άλλων ελέγχων αξιολόγησης καλής εφαρμογής στην διεθνή βιβλιογραφία, όπως ο υπολογισμός ποσοστών σωστής ταξινόμησης και γραφικές μέθοδοι. Για να αξιολογήσει την επιμέρους καλή εφαρμογή ο Pregibon (1980) γενίκευσε τις διαγνωστικές μεθόδους γραμμικής παλινδρόμησης για τα δίτιμα λογιστικά μοντέλα παλινδρόμησης και οι οποίες περιγράφονται στις εργασίες των (Hosmer and Lemeshow 1989, Hosmer et al. 1991 και Harrell et al. 1996)

Κεφάλαιο 5

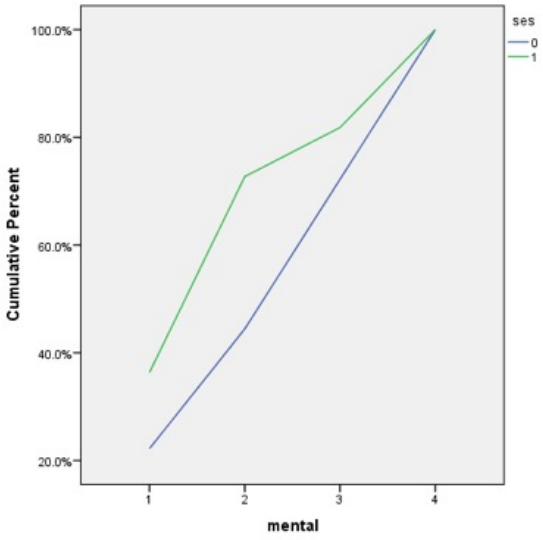
Εφαρμογή

Θα εξετάσουμε τα δεδομένα από το βιβλίο του Agresti (2002) προερχόμενα από μια έρευνα κατοίκων στην Alachua County της Φλόριντα και σχετίζει νοητική στέρηση με δυο ερμηνευτικές μεταβλητές. Η μεταβλητή **mental** που δηλώνει τη νοητική στέρηση είναι διατακτική με 4 κατηγορίες:
1=καλή κατάσταση
2=ήπια εμφάνιση συμπτωμάτων
3=μέτρια εμφάνιση συμπτωμάτων,
4=προβληματική κατάσταση.
Η μεταβλητή **life** είναι σύνθετη και μετράει τον αριθμό και σοβαρότητα από σημαντικών γεγονότων στη ζωή κάποιου ατόμου, όπως γέννηση παιδιού, νέα δουλειά κ.α που συνέβη στον εξεταζόμενο από την έρευνα μέσα στα τελευταία 3 χρόνια. Τέλος, η **ses** μετράει την κοινωνικοοικονομική κατάσταση του συμμετέχοντα και είναι δίτιμη (1=υψηλή, 0=χαμηλή).

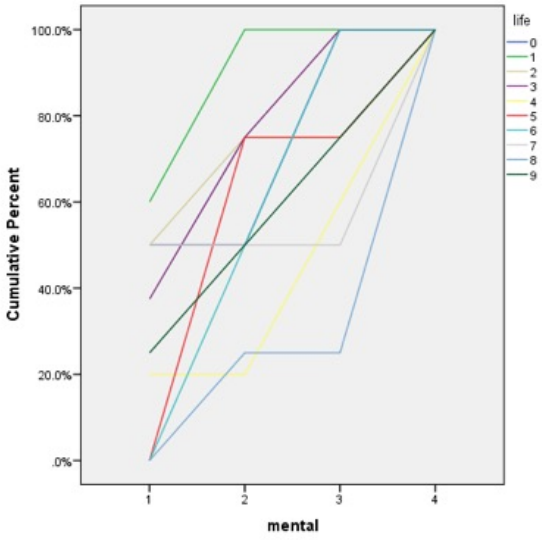
R code 5.1: Περιγραφή των δεδομένων

```
> summary(mentaldta)
mental           ses             life
Min.   :1.000   Min.   :0.00   Min.   :0.000
1st Qu.:1.000   1st Qu.:0.00   1st Qu.:2.000
Median :2.000   Median :1.00   Median :4.000
Mean   :2.325   Mean   :0.55   Mean   :4.275
3rd Qu.:3.000   3rd Qu.:1.00   3rd Qu.:6.250
Max.   :4.000   Max.   :1.00   Max.   :9.000
```

Αρχικά θα εξετάσουμε τα δεδομένα με ένα διάγραμμα αθροιστικών πιθανοτήτων της αξιολόγησης της νοητικής στέρησης με ξεχωριστές καμπύλες κάθε φορά για τα επίπεδα των μεταβλητών ses και life.



Σχήμα 5.1: Παρατηρούμενα ποσοστά για τη ses



Σχήμα 5.2: Παρατηρούμενα ποσοστά για τη life

Παρατηρούμε πως για τη ses αυτοί που έχουν χαμηλή κοινωνικοοικονομική κατάσταση βρίσκονται χαμηλότερα από αυτούς με υψηλή. Παρατηρούμε πως στην καλή κατάσταση νοητικής στέρησης βρίσκεται μεγαλύτερο ποσοστό ατόμων με καλή κοινωνικοοικονομική κατάσταση. Όσο προστίθενται επιπλέον ποσοστά (το αθροιστικό ποσοστό της ήπιας εμφάνισης είναι το άθροισμα της ήπιας εμφάνισης και της καλής κατάστασης) τα αθροιστικά ποσοστά για την ηπιότερα συμπτώματα νοητικής στέρησης παραμένουν μεγαλύτερα για αυτούς με καλή κατάσταση σε σχέση με αυτούς που έχουν κακή. Επειδή τα άτομα με καλή κοιν/κή κατάσταση έχουν ηπιότερη εμφάνιση συμπτωμάτων περιμένουμε να δούμε αρνητικούς συ-

ντελεστές για αυτή τη μεταβλητή. Ακόμα παρατηρούμε πως οι γραμμές στο διάγραμμα της μεταβλητής life τέμνονται και δεν μπορούμε να βγάλουμε εύκολο συμπερασμα.

5.0.1 Αναλογικό logit

Αρχικά θα εφαρμόσουμε το αθροιστικό μοντέλο logit αναλογικών σχετικών πιθανοτήτων.

R code 5.2: Αναλογικό logit

```
> prop.vgam<- vglm(mental ~ ses + life,
   family=cumulative(parallel = T), data=mentaldta)
> summary(prop.vgam)

Call:
vglm(formula = mental ~ ses + life, family = cumulative(parallel
   = T),
data = mentaldta)


Pearson residuals:
                 Min      1Q  Median     3Q   Max
logit(P[Y<=1]) -1.568 -0.7048 -0.2102 0.8070 2.713
logit(P[Y<=2]) -2.328 -0.4666  0.2657 0.6904 1.615
logit(P[Y<=3]) -3.688  0.1198  0.2039 0.4194 1.892

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept):1  -0.2819     0.6231  -0.452  0.65096
(Intercept):2   1.2128     0.6511   1.863  0.06251 .
(Intercept):3   2.2094     0.7171   3.081  0.00206 **
ses             1.1112     0.6143   1.809  0.07045 .
life           -0.3189     0.1194  -2.670  0.00759 **
---
Signif. codes:  0 ‘’*** 0.001 ‘’** 0.01 ‘’* 0.05 ‘’. 0.1 ‘’  1

Number of linear predictors:  3
```

```
Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2]),
   logit(P[Y<=3])

Residual deviance: 99.0979 on 115 degrees of freedom

Log-likelihood: -49.5489 on 115 degrees of freedom

Number of iterations: 5

No Hauck-Donner effect found in any of the estimates

Exponentiated coefficients:
ses      life
3.0380707 0.7269742
```

Οι εξισώσεις των logit θα είναι της μορφής

$$\begin{aligned}
\text{logit}\left[\Pr(Y \leq 1)\right] &= -0.2819 + 1.1112X_1 - 0.3189X_2 \\
\text{logit}\left[\Pr(Y \leq 1)\right] &= 1.2128 + 1.1112X_1 - 0.3189X_2 \\
\text{logit}\left[\Pr(Y \leq 3)\right] &= 2.2094 + 1.1112X_1 - 0.3189X_2
\end{aligned}$$

Οι εκτιμήσεις των συντελεστών α_i δίνονται από τα *Intercept* στη στήλη *Estimate* ($\alpha_1 = -0.2819, \alpha_2 = 1.2128, \alpha_3 = 2.2094$).

Οι εκτιμήσεις των παραμέτρων των επιδράσεων, δηλαδή τα $\hat{\beta}_1 = -0,319$ και $\hat{\beta}_2 = 1.11$ για τις μεταβλητές *ses* και *life* δείχνουν πως η αθροιστική πιθανότητα ξεκινώντας από το πρώτο στάδιο της κλίμακας όπου βρίσκονται τα άτομα με καλή κατάσταση μειώνεται όσο ο αριθμός των σημαντικών γεγονότων αυξάνεται και αυξάνεται στα ανώτερα επίπεδα της κοιν/κης κατάστασης. Για παράδειγμα, δοθέντος του επιπέδου της κοιν/κης κατάστασης, όσο αυξάνεται ο αριθμός των σημαντικών γεγονότων στη ζωή του ατόμου οι σχετικές πιθανότητες για πνευματική στέρηση σε οποιοδήποτε επίπεδο αυξάνονται κατά $e^{-0,32} = 0,72$.

Να σημειώσουμε πως η επίδραση των ερμηνευτικών μεταβλητών X είναι η μεταβολή των σχετικών πιθανοτήτων μια παρατήρηση να βρίσκεται σε μια συγκεκριμένη κατηγορία ή και στις αμέσως προηγούμενες, της εξαρτημένης μεταβλητής *mental*. Αυτή η επίδραση είναι αναλογική στις σχετικές συχνότητες (odds) για κάθε $Y \leq j$ για όλες τις κατηγορίες j και για αυτό έχουμε μόνο δυο εκτιμήσεις των συντελεστών των ses και life για όλα τα επίπεδα. Για παράδειγμα, δοθέντος του αριθμού των συμβάντων στη μεταβλητή *life*, στο υψηλό

επίπεδο κοιν/κής κατάστασης *ses* οι εκτιμημένες σχετικές πιθανότητες νοητικής στέρησης είναι $e^{1.111} = 3$ φορές οι τις σχετικές πιθανότητες στο χαμηλό επίπεδο νοητικής στέρησης, ανεξαρτήτως επιπέδου νοητικής στέρησης.

Παρατηρούμε ακόμα πως η μεταβλητή *ses* δεν κρίνεται απαραίτητη για την ερμηνεία του μοντέλου. Από το *p-value* των *intercept* κρίνουμε πως υπάρχει στατιστικά σημαντική διαφορά (*p-value* $\backsimeq 0$) σε επίπεδο εμπιστοσύνης 5% μόνο μεταξύ των κατηγοριών 4(προβληματική) και 1(καλή κατάσταση). Οι διαφορές των κατηγοριών 3 και 2 με την 1 δε φαίνεται να είναι στατιστικά σημαντικές.

Τέλος, το μοντέλο κρίνεται συνολικά επαρκές και με το κριτήριο της απόκλισης αφού η διαφορά της τιμής της απόκλισης (9,94) είναι μεγαλύτερη από την τιμή της στατιστικής συνάρτησης $\chi_{0,05,2} = 0,35$. Παρακάτω παρατίθεται το αποτέλεσμα για το μοντέλο μόνο με τους σταθερούς όρους με το οποίο θα συγκρίνουμε το μοντέλο με τις κύριες επιδράσεις.

R code 5.3: Αναλογικό logit μόνο με σταθερούς όρους

```
> prop.vgam0<- vglm(mental ~1, family=cumulative(parallel = T),
   data=mentaldta)
> summary(prop.vgam0)

Call:
vglm(formula = mental ~ 1, family = cumulative(parallel = T),
data = mentaldta)



Pearson residuals:
                Min      1Q  Median     3Q   Max
logit(P[Y<=1]) -1.024 -1.0236 -0.2563 1.4607 1.461
logit(P[Y<=2]) -1.749 -0.5799  0.3824 1.0727 1.073
logit(P[Y<=3]) -1.744  0.2315  0.2315 0.3671 1.216

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept):1  -0.8473     0.3450  -2.456  0.01406 *
(Intercept):2   0.4055     0.3227   1.256  0.20901
(Intercept):3   1.2368     0.3786   3.266  0.00109 **
---
Signif. codes:  0 ‘’*** 0.001 ‘’** 0.01 ‘’* 0.05 ‘’. 0.1 ‘’  1
```

```
Number of linear predictors:  3

Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2]),
   logit(P[Y<=3])

Residual deviance: 109.0421 on 117 degrees of freedom

Log-likelihood: -54.521 on 117 degrees of freedom

Number of iterations: 1

No Hauck-Donner effect found in any of the estimates
```

5.0.2 Αθροιστικό logit

Παρακάτω που εφαρμόζουμε το αθροιστικό logit χωρίς να ισχύει η υπόθεση της αναλογικότητας παρατηρούμε πως η απόκλιση των καταλοίπων μειώθηκε σε σχέση με το μοντέλο των αναλογικών σχετικών πιθανοτήτων (*proportional odds*) με τις κύριες επιδράσεις που εξετάσαμε παραπάνω. Πλέον υπάρχουν εκτιμήσεις των β για κάθε κατηγορία της *mental*. Οι εξισώσεις των logit θα είναι της μορφής

$$\begin{aligned}
\text{logit}\left[\Pr(Y \leq 1)\right] &= \alpha_1 + \beta_{11}X + \beta_{21}X \\
\text{logit}\left[\Pr(Y \leq 1)\right] &= \alpha_1 + \beta_{12}X + \beta_{22}X \\
\text{logit}\left[\Pr(Y \leq 1)\right] &= \alpha_1 + \beta_{11}X + \beta_{21}X \\
\text{logit}\left[\Pr(Y \leq 1)\right] &= \alpha_1 + \beta_{13}X + \beta_{23}X \\
&\vdots \\
\text{logit}\left[\Pr(Y \leq 3)\right] &= \alpha_3 + \beta_{11}X + \beta_{21}X
\end{aligned} \tag{5.1}$$

R code 5.4: Αθροιστικό logit

```
> cum.vgam<- vglm(mental ~ ses + life, family=cumulative,
   data=mentaldta)
> summary(cum.vgam)

Call:
```

```
vglm(formula = mental ~ ses + life, family = cumulative, data =
    mentaldta)

Pearson residuals:
                 Min      1Q  Median     3Q   Max
logit(P[Y<=1]) -1.509 -0.6707 -0.2126 0.8310 2.790
logit(P[Y<=2]) -2.583 -0.6027  0.2487 0.6277 1.793
logit(P[Y<=3]) -3.841  0.1128  0.1849 0.4095 2.063

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept):1  -0.1930     0.7387  -0.261   0.7938
(Intercept):2   0.8278     0.7036   1.176   0.2394
(Intercept):3   2.8049     0.9615      NA       NA
ses:1           0.9732     0.7720   1.261   0.2074
ses:2           1.4962     0.7460   2.006   0.0449 *
ses:3           0.7518     0.8358   0.899   0.3684
life:1         -0.3182     0.1597  -1.993   0.0463 *
life:2         -0.2739     0.1372  -1.997   0.0458 *
life:3         -0.3964     0.1592  -2.490   0.0128 *
---
Signif. codes:  0 ‘’*** 0.001 ‘’** 0.01 ‘’* 0.05 ‘’. 0.1 ‘’  1

Number of linear predictors:  3

Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2]),
    logit(P[Y<=3])

Residual deviance: 96.7486 on 111 degrees of freedom

Log-likelihood: -48.3743 on 111 degrees of freedom

Number of iterations: 14
```

Επίσης, και πάλι η μεταβλητή *ses* δεν κρίνεται επαρκής για την ερμηνεία του μοντέλου ενώ παρακάτω συγκρίνουμε το αναλογικό logit μοντέλο και το απλό αθροιστικό μοντέλο logit.

R code 5.5: Σύγκριση αναλογικού και αθροιστικού μοντέλου logit

```
> deviance(prop.vgam)-deviance(cum.vgam)
[1] 2.349308
> df.residual(prop.vgam)-df.residual(cum.vgam)
[1] 4
> qchisq(0.05,4)
[1] 0.710723
```

Παρατηρούμε πως μέχρι τώρα η απόκλιση των καταλοίπων (*Residual deviance*) δεν φαίνεται να διαφέρει από μοντέλο σε μοντέλο. Ωστόσο, με το κριτήριο των διαφορών στην απόκλιση, και συγκρίνοντας τις μεταξύ των δυο μοντελών (cumulative-proportional odds) παρατηρούμε πως η διαφορά τους είναι στατιστικά σημαντική και κρίνεται πως καλύτερη εφαρμογή έχει το μοντέλο *proportional odds*.

5.0.3 Μερικώς αναλογικά μοντέλα

Τα μοντέλα αυτά υιοθετούν την υπόθεση της αναλογικότητας μόνο για μερικές από τις ερμηνευτικές μεταβλητές, εδώ για την *life*. Οι εξισώσεις των logit θα είναι της μορφής

$$\begin{aligned}
\text{logit}\left[\Pr(Y \leq 1)\right] &= \alpha_1 + \beta_{11}X + \beta_2 X \\
\text{logit}\left[\Pr(Y \leq 1)\right] &= \alpha_1 + \beta_{12}X + \beta_2 X \\
\text{logit}\left[\Pr(Y \leq 1)\right] &= \alpha_1 + \beta_{11}X + \beta_2 X \\
\text{logit}\left[\Pr(Y \leq 1)\right] &= \alpha_1 + \beta_{13}X + \beta_2 X \\
&\vdots \\
\text{logit}\left[\Pr(Y \leq 3)\right] &= \alpha_3 + \beta_{11}X + \beta_2 X
\end{aligned} \tag{5.2}$$

R code 5.6: Μερικώς αναλογικά μοντέλα

```
> summary(par.vgam)
Call:
vglm(formula = mental ~ ses + life, family = cumulative(parallel
   = F ~
ses), data = mentaldta)


Pearson residuals:
```

```
                 Min      1Q  Median      3Q   Max
logit(P[Y<=1]) -1.488 -0.7076 -0.2117  0.8266 2.824
logit(P[Y<=2]) -2.832 -0.5098  0.2363  0.5662 1.952
logit(P[Y<=3]) -3.220  0.1256  0.2061  0.4120 1.609

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept):1  -0.1766     0.6951  -0.254  0.79943
(Intercept):2   1.0057     0.6633   1.516  0.12946
(Intercept):3   2.3956     0.7789   3.075  0.00210 **
ses:1           0.9824     0.7643   1.285  0.19868
ses:2           1.5415     0.7373   2.091  0.03656 *
ses:3           0.7362     0.8121   0.907  0.36464
life           -0.3241     0.1202  -2.697  0.00699 **
---
Signif. codes:  0 ‘’*** 0.001 ‘’** 0.01 ‘’* 0.05 ‘’. 0.1 ‘’  1

Number of linear predictors:  3

Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2]),
   logit(P[Y<=3])

Residual deviance: 97.3647 on 113 degrees of freedom

Log-likelihood: -48.6823 on 113 degrees of freedom
```

Η ερμηνεία των μοντέλων αυτών είναι παρόμοια με των προηγούμενων.Παρακάτω συγκρίνουμε το μοντέλο μερικών αναλογικών logit με το αναλογικό μοντέλο logit. (χρησιμοποιήθηκε μια συνάρτηση στην R για να υπολογίσει αυτόματα το p-value της χ^2 σε επίπεδι σημαντικότητας 5%)

R code 5.7: Σύγκριση μοντέλου μερικών αναλογικών logit με το αναλογικό logit

```
> pchisq(deviance(prop.vgam)-deviance(par.vgam),
+ df.residual(prop.vgam)-df.residual(par.vgam),lower.tail = F)
[1] 0.4203731
```

Βλέπουμε πως το μοντέλο των αναλογικών logit εφαρμόζεται καλύτερα στα δεδομένα μας αφού το *p-value* του ελέγχου πως διαφέρουν σημαντικά τα δυο μοντέλα. Να σημειώσουμε πως συγκρίναμε το μοντέλο των μερικών αναλογικών logit με το μοντέλο μόνο με τους σταθερούς όρους και βρήκαμε πως το μοντέλο των μερικών αναλογικών όρων δεν διαφέρει στατιστικώς σημαντικά (ε.σ = 5%) από το μοντέλο μόνο με τους σταθερούς όρους.

```
pchisq(deviance(par.vgam0)-deviance(par.vgam),
+  df.residual(par.vgam0)-     df.residual(par.vgam),lower.tail
   = F)
       [1] 0.01991874
```

5.0.4 Logit γειτονικής κατηγορίας

Να θυμίσουμε πως σε αυτή την περίπτωση, το κάθε ένα από $J-1$ logit που δημιουργούνται, δίνουν τον λογάριθμο της σχετικής πιθανότητας μια παρατήρηση να βρίσκεται είτε στην κατηγορία *j* είτε στην κατηγορία $j+1,\ j=1,\ldots,J-1$. Τα logit γειτονικής κατηγορίας διαφέρουν από μοντέλα αθροιστικών logit επειδή οι επιδράσεις των ερμηνευτικών μεταβλητών αναφέρονται σε συγκεκριμένες κατηγορίες απόκρισης και όχι σε αθροιστικές ομάδες και επειδή δε βασίζονται στην ύπαρξη μιας συνεχούς λανθάνουσας μεταβλητής.

Οι εξισώσεις των logit θα είναι της μορφής

$$\log(\frac{\pi_1}{\pi_2}) = 0.19654 + 0.65760X_1 - 0.17595X_2$$
$$\log(\frac{\pi_2}{\pi_3}) = 0.96083 + 0.65760X_1 - 0.17595X_2$$
$$\log(\frac{\pi_3}{\pi_4}) = 0.41096 + 0.65760X_1 - 0.17595X_2$$

Βλέποντας τις εξισώσεις μπορούμε για παράδειγμα να καταλάβουμε πως δοθέντος του αριθμού των συμβάντων στη μεταβλητή *life*, στο υψηλό επίπεδο κοιν/κής κατάστασης *ses* οι εκτιμημένες σχετικές πιθανότητες νοητικής στέρησης είναι $e^{0.65760} = 1.930154$ φορές οι σχετικές πιθανότητες στο χαμηλό επίπεδο νοητικής στέρησης, για την κατηγορία *mental* 1 σε σχέση με την 2 .

R code 5.8: Logit γειτονικής κατηγορίας

```
> ad.vgam <-vglm(mental ~ ses + life,
   family=acat(reverse=T,parallel = T), data=mentaldta)
```

```
> summary(ad.vgam)

Call:
vglm(formula = mental ~ ses + life, family = acat(reverse = T,
parallel = T), data = mentaldta)



Pearson residuals:
                      Min       1Q  Median     3Q   Max
loge(P[Y=1]/P[Y=2]) -1.346 -0.87881 -0.1727 0.8659 2.589
loge(P[Y=2]/P[Y=3]) -2.168 -0.50716  0.2069 0.7692 1.600
loge(P[Y=3]/P[Y=4]) -4.299  0.03176  0.1571 0.3956 1.692

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept):1  0.19654    0.48505   0.405   0.6853
(Intercept):2  0.96083    0.56355   1.705   0.0882 .
(Intercept):3  0.41096    0.64672   0.635   0.5251
ses            0.65760    0.35087   1.874   0.0609 .
life          -0.17595    0.06959  -2.528   0.0115 *
---
Signif. codes:  0 ‘’*** 0.001 ‘’** 0.01 ‘’* 0.05 ‘’. 0.1 ‘’  1

Number of linear predictors:  3

Names of linear predictors:
loge(P[Y=1]/P[Y=2]), loge(P[Y=2]/P[Y=3]), loge(P[Y=3]/P[Y=4])

Residual deviance: 98.7022 on 115 degrees of freedom

Log-likelihood: -49.3511 on 115 degrees of freedom

Number of iterations: 4
```

Απορρίπτουμε σε επίπεδο στατιστικής σημαντικότητας 5% τη σημαντικότητα της μεταβλητής *ses*, ενώ παρακάτω συγκρίνουμε το αναλογικό logit μοντέλο και το μοντέλο logit γειτονικής κατηγορίας.

R code 5.9: Σύγκριση αναλογικού και γειτονικής κατηγορίας μοντέλου logit

```
 pchisq(deviance(ad.vgam)-deviance(prop.vgam),
+ df.residual(ad.vgam)- df.residual(prop.vgam),lower.tail = F)
        [1] 1
```

Με το κριτήριο των διαφορών στην απόκλιση, και συγκρίνοντας τις μεταξύ των δυο μοντελών (cumulative-proportional odds) παρατηρούμε πως η διαφορά τους είναι στατιστικά σημαντική σε ε.σ 5% κρίνουμε πως και κρίνεται πως καλύτερη εφαρμογή έχει το μοντέλο *γειτονικής κατηγορίας logit*.

5.0.5 Continuation ratio

Τα logit αυτά, είναι τα logit των δεσμευμένων πιθανοτήτων πως μια παρατήρηση θα βρίσκεται στην *j*-οστή κατηγορία, δεδομένου ότι βρίσκεται τουλάχιστον στην *j*-οστή κατηγορία για $j = 1, , J1$.

Οι εξισώσεις των logit θα είναι της μορφής

$$\log(\frac{\pi_1}{\pi_2+\pi_3+\pi_4}) = -0.19654 - 0.65760X_1 + 0.17595X_2$$
$$\log(\frac{\pi_2}{\pi_3+\pi_4}) = -0.96083 - 0.65760X_1 + 0.17595X_2$$
$$\log(\frac{\pi_3}{\pi_4}) = -0.41096 - 0.65760X_1 + 0.17595X_2$$

Βλέποντας τις εξισώσεις μπορούμε για παράδειγμα να καταλάβουμε πως δοθέντος του αριθμού των συμβάντων στη μεταβλητή *life*, στο υψηλό επίπεδο κοιν/κής κατάστασης *ses* οι εκτιμημένες σχετικές πιθανότητες νοητικής στέρησης είναι $e^{-0.65760} = 0.518$ φορές οι σχετικές πιθανότητες στο χαμηλό επίπεδο νοητικής στέρησης, για την κατηγορία *mental* 1 σε σχέση με τη 2 και τη 3.

```
> con.vgam <-vglm(mental ~ ses + life,
   family=acat(reverse=F,parallel = T), data=mentaldta)
> summary(con.vgam)

Call:
vglm(formula = mental ~ ses + life, family = acat(reverse = F,
parallel = T), data = mentaldta)
```

```
Pearson residuals:
                 Min      1Q  Median      3Q   Max
loge(P[Y=2]/P[Y=1]) -2.589 -0.8659  0.1727  0.87881 1.346
loge(P[Y=3]/P[Y=2]) -1.600 -0.7692 -0.2069  0.50716 2.168
loge(P[Y=4]/P[Y=3]) -1.692 -0.3956 -0.1571 -0.03176 4.299

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept):1 -0.19654    0.48505  -0.405   0.6853
(Intercept):2 -0.96083    0.56355  -1.705   0.0882 .
(Intercept):3 -0.41096    0.64672  -0.635   0.5251
ses           -0.65760    0.35087  -1.874   0.0609 .
life           0.17595    0.06959   2.528   0.0115 *
---
Signif. codes:  0 ‘’*** 0.001 ‘’** 0.01 ‘’* 0.05 ‘’. 0.1 ‘’  1

Number of linear predictors:  3

Names of linear predictors:
loge(P[Y=2]/P[Y=1]), loge(P[Y=3]/P[Y=2]), loge(P[Y=4]/P[Y=3])

Residual deviance: 98.7022 on 115 degrees of freedom

Log-likelihood: -49.3511 on 115 degrees of freedom
```

R code 5.10: Σύγκριση αναλογικού και continuation ratio logit

```
> pchisq(deviance(con.vgam)-deviance(prop.vgam),
+df.residual(con.vgam)-  df.residual(prop.vgam),lower.tail = F)
[1] 1
```

Με το κριτήριο των διαφορών στην απόκλιση, και συγκρίνοντας τις μεταξύ των δυο μοντελών (cumulative-proportional odds) παρατηρούμε πως η διαφορά τους είναι στατιστικά σημαντική σε ε.σ 5% κρίνουμε πως και κρίνεται πως καλύτερη εφαρμογή έχει το μοντέλο *continuation ratio logit*.

Παρακάτω μπορούμε να δούμε συνοπτικά τα μοντέλα που εφαρμόσαμε, την απόκλιση, τους βαθμούς ελευθερίας και εάν κρίναμε ότι έχουν καλή εφαρμογή.

Μοντέλο	**Deviance**	**df**	**Στατιστικά σημαντικό**
Αναλογικό logit	99.0979	115	ναι
Αθροιστικό logit	96.7486	117	όχι
Μερικώς αναλογικό logit	97.3647	113	όχι
Γειτονικής κατηγορίας logit	98.7022	115	ναι
Continuation ratio logit	98.7022	115	ναι

Βιβλιογραφία

Agresti, A. (1992), `Analysis of Ordinal Paired Comparison Data', *Applied Statistics* **41**(2), 287--297.
URL: *http://www.jstor.org/pss/2347562 http://www.jstor.org/stable/2347562?origin=crossref*

Agresti, A. (2002), *Categorical Data Analysis*, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., Hoboken, NJ, USA.
URL: *http://doi.wiley.com/10.1002/0471249688*

Agresti, A. (2010), Logistic Regression Models Using Cumulative Logits, *in* `Analysis of Ordinal Categorical Data', John Wiley & Sons, Inc., pp. 44--87.
URL: *http://doi.wiley.com/10.1002/9780470594001.ch3*

Agresti, A., Clogg, C. C. and Shihadeh, E. S. (1995), `Statistical Models for Ordinal Variables.'.

Agresti, A. and Liu, I. M. (1999), `Modeling a categorical variable allowing arbitrarily many category choices', *Biometrics* **55**(3), 936--943.
URL: *http://onlinelibrary.wiley.com/doi/10.1111/j.0006-341X.1999.00936.x/full*

Aitchison, J. and Bennett, J. A. (1970), `Polychotomous quantal response by maximum indicant', *Biometrika* **57**(2), 253--262.

Aitchison, J. and Silvey, S. D. (1957), `The Generalization of Probit Analysis to the Case of Multiple Responses', *Biometrika* **44**(1/2), 131--140.

Ananth, C. (1997), `Regression models for ordinal responses: a review of methods and applications', *International Journal of Epidemiology* **26**(6), 1323--1333.
URL: *https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/26.6.1323*

Anderson, J. A. (1984), `Regression and Ordered Categorical Variables', *Journal of the Royal Statistical Society. Series B (Methodological)* **46**(1), 1--30.
URL: *http://www.jstor.org/stable/2345457*

Anderson, J. A. and Philips, P. R. (1981), `Regression, Discrimination and Measurement Models for Ordered Categorical Variables', *Applied Statistics* **30**(1), 22.
URL: *http://www.jstor.org/stable/10.2307/2346654?origin=crossref*

Barnhart, H. X. and Sampson, A. R. (1994), `Overview of multinomial models for ordinal data', *Communications in Statistics - Theory and Methods* **23**(12), 3395--3416.
URL: *http://www.tandfonline.com/doi/abs/10.1080/03610929408831454*

Bauer, D. J. (2009), `A Note on comparing the estimates of models for cluster-correlated or longitudinal data with binary or ordinal outcomes', *Psychometrika* **74**(1), 97--105.
URL: *http://link.springer.com/10.1007/s11336-008-9080-1*

Bellemare, C., Melenberg, B. and van Soest, A. (2002), `Semi-parametric models for satisfaction with income', *Portuguese Economic Journal* **1**(2), 181--203.
URL: *https://www.ifs.org.uk/publications/2658*

Böckenholt, U. and Dillon, W. R. (1997), `Modeling within-subject dependencies in ordinal paired comparison data', *Psychometrika* **62**(3), 411--434.
URL: *http://link.springer.com/10.1007/BF02294559*

Boes, S. and Winkelmann, R. (2006), Ordered Response Models, *in* O. Hübler and J. Frohn, eds, `Modern Econometric Analysis: Surveys on Recent Developments', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 167--181.
URL: *https://doi.org/10.1007/3-540-32693-6_12*

Brant, R. (1990), `Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression', *Biometrics* **46**(4), 1171.
URL: *http://www.jstor.org/stable/2532457?origin=crossref*

Christian, M. S. (1984), `Morbidity and mortality of car occupants: comparative survey over 24 months.', *British medical journal (Clinical research ed.)* **289**(6457), 1525--6.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/6439299*

Collett, D. (2002), `Modelling binary data', p. 387.

Cox, C. (1988), `Multinomial Regression-Models Based on Continuation Ratios', *Statistics in Medicine* **7**(3), 435--441.

Cox, C. (1995), `Location—scale cumulative odds models for ordinal data: A generalized non□linear model approach', *Statistics in Medicine* **14**(11), 1191--1203.
URL: *http://doi.wiley.com/10.1002/sim.4780141105*

Efron, B. and Hinkley, D. V. (1978), `Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Fisher Information', *Biometrika* **65**(3), 457--482.
URL: *http://www.jstor.org/stable/2335893*

Everitt, B. S. (1992), The analysis of contingency tables, *in* `Monographs on statistics and applied probability', Vol. 45, Chapman & Hall, pp. 117--135.
URL: *http://onlinelibrary.wiley.com/doi/10.1002/bimj.4710350708/abstract*

Fahrmeir, L., Tutz, G. and Hennevogl, W. (2001), *Multivariate statistical modelling based on generalized linear models*, Springer New York.
URL: *http://www.library.wisc.edu/selectedtocs/bc025.pdf*

Farewell, V. T. (1982), `A note on regression analysis of ordinal data with variability of classification', *Biometrika* **69**(3), 533--538.
URL: *https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/69.3.533*

Fienberg, S. E. (2007), *The Analysis of Cross-Classified Categorical Data*, Springer New York, New York, NY.
URL: *http://link.springer.com/10.1007/978-0-387-72825-4*

Fleiss, J. L., Levin, B. and Paik, M. C. (2003), *Statistical Methods for Rates and Proportions*, Wiley Series in Probability and Statistics, 3 edn, John Wiley & Sons, Inc., Hoboken, NJ, USA.
URL: *http://doi.wiley.com/10.1002/0471445428*

Goodman, L. A. (1983), `The Analysis of Dependence in Cross-Classifications Having Ordered Categories, Using Log-Linear Models for Frequencies and Log-Linear Models for Odds', *Biometrics* **39**(1), 149--160.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/6871344 http://www.jstor.org/stable/2530815*

Greenland, S. (1985), `An Application of Logistic Models to the Analysis of Ordinal Responses', *Biometrical Journal* **27**(2), 189--197.

Greenland, S. (1994), `Alternative Models for Ordinal Logistic Regression', *Statistics in Medicine* **13**(16), 1665--1677.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/7973242*

Gurland, J., Lee, I. and Dahm, P. A. (1960), `Polychotomous Quantal Response in Biological Assay', *Biometrics* **16**(3), 382.
URL: *http://www.jstor.org/stable/2527689?origin=crossref*

Harrell, F. (2001), *Regression modeling strategies with applications to linear models, logistic regression models and survival analysis*, Springer New York.

Hastie, T. J., Botha, J. L. and Schnitzler, C. M. (1989), `Regression with an ordered categorical response', *Statistics in Medicine* **8**(7), 785--794.

Lee, J. (1992), `Cumulative logit modelling for ordinal response variables: applications to biomedical research', *Bioinformatics* **8**(6), 555--562.
URL: *http://dx.doi.org/10.1093/bioinformatics/8.6.555*

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

McCullagh, P. P. and Nelder, J. A. (1989), *Generalized Linear Models*, 2 edn, Chapman and Hall.

McKelvey, R. D. and Zavoina, W. (1975), `A statistical model for the analysis of ordinal level dependent variables', *Journal of mathematical sociology* **4**(1), 103--120.
URL: *http://www.tandfonline.com/doi/full/10.1080/0022250X.1975.9989847*

Olsson, U. (1979), `Maximum likelihood estimation of the polychoric correlation coefficient', *Psychometrika* **44**(4), 443--460.
URL: *http://link.springer.com/10.1007/BF02296207*

Peterson, B. and Harrell, F. E. (1990), `Partial Proportional Odds Models for Ordinal Response Variables', *Source Journal of the Royal Statistical Society. Series C (Applied Statistics) Appl. Statist* **39**(2), 205--217.
URL: *http://www.jstor.org/stable/2347760%5Cnhttp://www.jstor.org/page/*

Pratt, J. W. (1981), `Concavity of the log likelihood', *Journal of the American Statistical Association* **76**(373), 103--106.
URL: *http://www.jstor.org/stable/2287052?origin=crossref*

Pregibon, D. (1980), `Goodness of link tests for generalized linear models', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **29**(1), 15--23.
URL: *http://www.jstor.org/stable/10.2307/2346405?origin=crossref*

Ronning, G. and Kukuk, M. (1996), `Efficient estimation of ordered probit models', *Journal of the American Statistical Association* **91**(435), 1120--1129.
URL: *http://www.jstor.org/stable/2291731?origin=crossref*

Siegel, S. and Castellan, N. J. (1988), *Nonparametric statistics for the behavioral sciences*, McGraw-Hill.

Simon, G. (1974), `Alternative Analyses for the Singly Ordered Contingency Table', **69**(348), 971--976.
URL: *http://www.jstor.org/stable/2286174* .

Snell, E. J. (1964), `A Scaling Procedure for Ordered Categorical Data', *Biometrics* **20**(3), 592.
URL: *http://www.jstor.org/stable/2528498?origin=crossref*

Stevens, S. (1946), `On the Theory of Scales of Measurement', *Science* **103**(2684), 677--680.

Stewart, M. B. (2004), `A Comparison of Semiparametric Estimators for the Ordered Response Model', *Computational Statistics & Data Analysis* **49**(2), 555--573.
URL: *http://www.sciencedirect.com/science/article/pii/S0167947304001628*

Svensson, E. (2001), `Important Considerations for Optimal Communication Between Statisticians and Medical Researchers in Consulting, Teaching and Collaborative Research with a Focus on the Analysis of Ordered Categorical Data', *Training Researchers in the Use of Statistics* .

Thompson, R. and Baker, R. J. (1981), `Composite Link Functions in Generalized Linear Models', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **30**(2), 125--131.
URL: *http://www.jstor.org/stable/2346381*

Tutz, G. (1990), `Sequential item response models with an ordered response', *British Journal of Mathematical and Statistical Psychology* **43**(1), 39--55.
URL: *http://doi.wiley.com/10.1111/j.2044-8317.1990.tb00925.x*

Tutz, G. (1991), `Sequential models in categorical regression', *Computational Statistics & Data Analysis* **11**(3), 275--295.
URL: *https://econpapers.repec.org/RePEc:eee:csdana:v:11:y:1991:i:3:p:275-295*

Walker, S. H. and Duncan, D. B. (1967), `Estimation of the probability of an event as a function of several independent variables', *Biometrika* **54**(1-2), 167--179.
URL: *https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/54.1-2.167*

Whitehead, J. (1993), `Sample size calculations for ordered categorical data', *Statistics in Medicine* **12**(24), 2257--2271.

Williams, O. D. and Grizzle, J. E. (1972), `Analysis of Contingency Tables Having Ordered Response Categories', *Journal of the American Statistical Association* **67**(337), 55--63. **URL:** *http://www.tandfonline.com/doi/abs/10.1080/01621459.1972.10481205*

Winship, C. & Mare, R. (1984), `Regression Models with Ordinal Variables', *American Sociological Review* **49**(4), 512--525.