ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ & ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Ιεραρχικά Μοντέλα και Εφαρμογές

Διπλωματικη εργασια

της

Μπάτσιου Μαρίας

Επιβλέπων: Φουσκακης Δημητρης
Επ. Καθηγητης ΕΜΠ

Αθηνα, Ιουλιος 2013

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ & ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Ιεραρχικά Μοντέλα και Εφαρμογές

Διπλωματικη εργασια

της

Μπάτσιου Μαρίας

Εγκρίθηκε απο την τριμελή εξεταστική επιτροπή:

Φουσκακης Δημητρης .....
Επ. Καθηγητης ΕΜΠ

Βοντα Φιλια .....
Επ. Καθηγητρια ΕΜΠ

Κολετσος Ιωαννης .....
Επ. Καθηγητης ΕΜΠ

Αθηνα, Ιουλιος 2013

.....
Μπατσιου Μαρια
Διπλωματουχος Μαθηματικος Εφαρμογων ΣΕΜΦΕ-ΕΜΠ

Ευχαριστίες

Ολοκληρώνοντας την διπλωματική μου εργασία θα ήθελα να ευχαριστήσω όλους αυτούς τους ανθρώπους που με ενέπνευσαν να ασχοληθώ με τη στατιστική και με καθοδήγησαν όλα αυτά τα χρόνια ως μαθήτρια και μετέπειτα ως φοιτήτρια.

Αρχικά θα ήθελα να ευχαριστήσω τον καθηγητή μου κύριο Δημητρη Φουσκάκη για την καθοδήγηση του κατα τη διάρκεια της εκπόνησης αυτής της εργασίας.

Ευχαριστώ επίσης τους συμφοιτητές και φίλους μου, Νάντια Επισκόπου, Πυρπυρή Αλέξανδρο και Σπύρου Δημήτρη για τη τη βοήθεια που μου παρείχαν όποτε χρειάστηκε.

Τέλος ένα μεγάλο ευχαριστώ ανήκει στην οικογένεια μου για την ηθική και υλική υποστήριξη τους τα χρόνια των σπουδών μου και γενικότερα στη ζωή μου.

Περίληψη

Τα Ιεραρχικά ή Πολυεπίπεδα Μοντέλα θεωρούνται γενίκευση των γενικών και γενικευμένων γραμμικών μοντέλων, αποτελώντας βελτιωμένη έκδοση της κλασσικής παλινδρόμησης όσον αφορά την προβλεπτική ακρίβεια. Είναι ιδιαιτέρως διαδεδομένα στις κοινωνικές, ιατρικές και βιολογικές επιστήμες όπου τα δεδομένα έχουν ιεραρχική δομή, ενώ μπορούν να χρησιμοποιηθούν για πολλούς σκοπούς όπως η πρόβλεψη και η στατιστική συμπερασματολογία. Στο εισαγωγικό κεφάλαιο, παρουσιάζεται μέσω παραδειγμάτων η μορφή της δομής των δεδομένων και στη συνέχεια τα βασικά είδη των πολυεπίπεδων μοντέλων που αντιμετωπίζουν την πολυπλοκότητα της δομής αυτής. Στα κεφάλαια που ακολουθούν, αναλύονται τα δεδομένα που προέρχονται από την Έρευνα Σκοτών Αποφοίτων (Scottish School Leavers Survey) και την Δημογραφική Μελέτη για την Υγεία στο Μπανγκλαντές (Bangladesh Demographic and Health Survey) εφαρμόζοντας τις ιδέες της πολυεπίπεδης μοντελοποίησης.

Abstract

Multilevel (hierarchical) modeling is a generalization of linear and generalized linear modeling, outperforming classical regression in predictive accuracy. Considered as highly popular in the social, medical and biological sciences, where hierarchical structures are the norm, multilevel modeling can be used for a variety of purposes, including prediction or causal inference. In the introductory chapter, the data structrure is introduced though examples followed by the basic fundamental models that handle its complexity. In the later chapters, we analyze data from the Scottish School Leavers Survey and the Bangladesh Demographic and Health Survey, applying the ideas of multilevel modeling.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

1.1 Τι είναι τα πολυεπίπεδα μοντέλα παλινδρόμησης

Ορισμός 1. *Τα πολυεπίπεδα (ιεραρχικά) μοντέλα αποτελούν μια γενίκευση των γενικών και γενικευμένων γραμμικών μοντέλων,οι παράμετροι των οποίων (υπερπαράμετροι) αποτελούν και εκείνοι με τη σειρά τους μοντέλα παλινδρόμησης (με τις παραμέτρους τους να εκτιμώνται επίσης από τα δεδομένα).*

Τα πολυεπίπεδα μοντέλα καλούνται επίσης **ιεραρχικά**, για δύο διαφορετικούς λόγους: πρώτον, λόγω της δομής των δεδομένων (π.χ. μαθητές εμφωλευμένοι σε σχολεία), και δεύτερον, από το ίδιο το μοντέλο, που έχει τη δική του ιεραρχία, με τις παραμέτρους του κατώτερου επιπέδου (μαθητές) να ελέγχονται από τις υπερπαραμέτρους του ανώτερου επιπέδου του μοντέλου (σχολεία).

Για παράδειγμα, θεωρούμε δεδομένα από μαθητές από διάφορα σχολεία προβλέποντας σε κάθε σχολείο τους βαθμούς των μαθητών σε ένα τυποποιημένο διαγώνισμα, δεδομένων των βαθμών τους σε ένα pre-test αλλά και άλλων πληροφοριών. Θα προσαρμόσουμε για κάθε σχολείο, ένα ξεχωριστό μοντέλο παλινδρόμησης με τις παραμέτρους των οποίων να προσαρμόζονται βάσει των χαρακτηριστικών του κάθε σχολείου (λ.χ. αν το σχολείο είναι δημόσιο ή ιδιωτικό κ.λ.π.). Το γραμμικό μοντέλο για τους μαθητές (student-level) και εκείνο

για τα σχολεία (school-level) αποτελούν τα 2 επίπεδα του πολυεπίπεδου μοντέλου. Στο παράδειγμα αυτό, ένα πολυεπίπεδο μοντέλο μπορεί να εκφραστεί με 3 ισοδύναμους (τουλάχιστον) τρόπους σαν ένα μοντέλο student-level παλινδρόμησης:

- Μοντέλο όπου οι συντελεστές διαφέρουν ανά σχολείο (έτσι αντί για μοντέλο $y = \beta_0 + \beta_1 x + error$ έχουμε $y = \beta_{0j} + \beta_{1j} x + error$ όπου j ο δείκτης για τα σχολεία.

- Μοντέλο με περισσότερα από ένα στοιχεία διακύμανσης (student-level και school-level διακύμανση).

- Μοντέλο με πολυάριθμες επεξηγηματικές μεταβλητές, συμπεριλαμβανομένης μίας δείκτριας μεταβλητής για κάθε σχολείο.

Γενικότερα, θεωρούμε ένα πολυεπίπεδο μοντέλο ως ένα γενικό ή γενικευμένο γραμμικό μοντέλο παλινδρόμησης, οι παράμετροι του οποίου (οι συντελεστές του μοντέλου παλινδρόμησης) αποτελούν οι ίδιοι ένα μοντέλο. Το δεύτερο επίπεδο του μοντέλου έχει δικές του παραμέτρους, γνωστές ως υπερπαράμετροι, που εκτιμώνται επίσης από τα δεδομένα. Τα δύο μέρη κλειδιά για το πολυεπίπεδο μοντέλο είναι οι συντελεστές που διαφέρουν ανά επίπεδο (varying coefficients) και το μοντέλο για τους συντελεστές αυτούς (το οποίο με τη σειρά του μπορεί να περιέχει επεξηγηματικές μεταβλητές στο επίπεδο 2). Η κλασσική παλινδρόμηση μπορεί εξίσου να στεγάσει τέτοιου είδους συντελεστές, χρησιμοποιώντας δείκτριες μεταβλητές. Το χαρακτηριστικό που διακρίνει τα πολυεπίπεδα μοντέλα από εκείνα της κλασσικής παλινδρόμησης είναι η μεταβλητότητα μεταξύ των ομάδων.

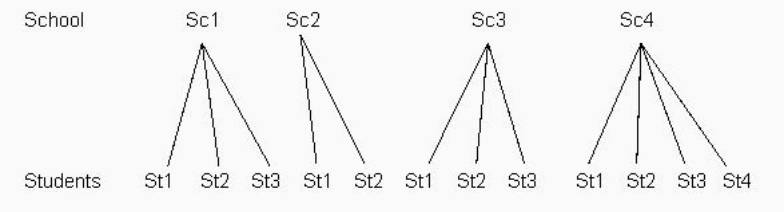
1.2 Πολυεπίπεδες δομές και ταξινομήσεις

Τα πολυεπίπεδα μοντέλα σχεδιάστηκαν για να εξερευνήσουμε και να αναλύσουμε δεδομένα που προέρχονται από πληθυσμούς με περίπλοκη δομή. Σε οποιαδήποτε περίπλοκη δομή, μπορούμε να αναγνωρίσουμε ατομικές μονάδες (atomic units). Αυτές είναι οι μονάδες που βρίσκονται στο χαμηλότερο επίπεδο του συστήματος. Συχνά, οι μονάδες αυτές είναι άτομα (π.χ. μαθητές). Τα άτομα αυτά ομαδοποιούνται σε υψηλότερα επίπεδα ή μονάδες (π.χ σχολεία). Κατά σύμβαση θα λέμε, ότι οι μαθητές βρίσκονται στο επίπεδο 1 και τα σχολεία στο επίπεδο 2 της δομής που κατασκευάσαμε. Ακολουθούν διαγράμματα μονάδων ή ταξινομήσεων. Αξίζει να σημειωθεί ότι το επίπεδο (level) υπονοεί

εμφωλευμένη ιεραρχική σχέση μεταξύ των μονάδων (οι μονάδες στα χαμηλότερα επίπεδα είναι εμφωλευμένες μόνο σε μία υψηλότερου επιπέδου μονάδα) Επιλέξαμε παραδείγματα όπου τα πολυεπίπεδα μοντέλα είναι ιδιαιτέρως χρήσιμα και συχνά απαραίτητα.

1.2.1 Διεπίπεδες ιεραρχικές δομές

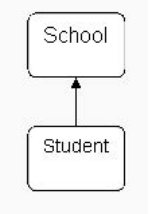
Οι ιεραρχικές δομές προκύπτουν όταν οι μονάδες χαμηλότερων επιπέδων εμφωλεύονται σε μια και μόνο μια υψηλότερου επίπεδου μονάδα. Το διάγραμμα-



Σχήμα 1.1: *Διάγραμμα-μονάδων διεπίπεδης εμφωλευμένης δομής: μαθητές σε σχολεία*

μονάδων (unit diagram, Σχ. 1.1) έχει ως στόχο να δείξει την υποκείμενη δομή ενός προβλήματος όσον αφορά την έρευνα των επιμέρους μονάδων. Οι κόμβοι στο διάγραμμα είναι συγκεκριμένες μονάδες του πληθυσμού. Στην περίπτωση αυτή, οι μονάδες είναι οι μαθητές και τα σχολεία που σχηματίζουν δύο επίπεδα (ή ταξινομήσεις). Οι χαμηλότερες μονάδες αποτελούν την κατάταξη των μαθητών (St1, St2 κλπ.) και οι υψηλότερες μονάδες αποτελούν την κατάταξη των σχολείων (Sc1, ..., Sc4). Αυτό το διάγραμμα μονάδων είναι απλά ένα σχήμα για να εκφράσουμε τη βασική διάρθρωση των μαθητών που είναι εμφωλευμένοι μέσα σε σχολεία. Σε ένα πραγματικό σύνολο δεδομένων θα έχουμε προφανώς πολλά περισσότερα από τέσσερα σχολεία και 12 μαθητές. Η ιεραρχική δομή υποδηλώνει ότι ένας μαθητής φοιτά σε μόνο ένα σχολείο. Μια τέτοια δομή μπορεί να προκύψει όταν μας ενδιαφέρει η σχολική επίδοση των μαθητών κάθε σχολείου οπότε και κάνουμε επαναλαμβανόμενες μετρήσεις αξιολογώντάς την. Η δομή αυτή είναι πιθανό να οδηγήσει είτε σε συσχετισμένα είτε μη ανεξάρτητα δεδομένα, με την έννοια ότι οι μαθητές που φοιτούν στο ίδιο σχολείο παρουσιάζουν συχνά παρόμοιες τάσεις όσον αφορά την επίδοση στις εξετάσεις. Η παραπάνω διεπίπεδη εμφωλευμένη δομή (Σχ. 1.1) μπορεί επίσης να παρουσιαστεί από ένα διάγραμμα ταξινόμησης (βλ. Σχήμα 1.2). Τα διαγράμματα ταξινόμησης έχουν ένα κόμβο ανά επίπεδο, ενώ κάθε κόμβος ενώνεται με

ένα μόνο βέλος υποδεικνύοντας μια εμφωλευμένη (αυστηρά ιεραρχική) σχέση μεταξύ των ταξινομήσεων. Τα διαγράμματα ταξινόμησης είναι πιο αφηρημένα



Σχήμα 1.2: *Διάγραμμα ταξινόμησης διεπίπεδης εμφωλευμένης δομής: μαθητές σε σχολεία*

από τα διαγράμματα μονάδων και είναι ιδιαίτερα χρήσιμα όταν ο πληθυσμός που μελετάται έχει μια περίπλοκη δομή με πολλές ταξινομήσεις.

Ο Πίνακας 1.1 δείχνει ένα πλαίσιο δεδομένων της δομής του Σχήματος 1.1 συμπεριλαμβανομένων μιας εξαρτημένης μεταβλητής y (βαθμολογία εξετάσεων τρέχοντος έτους), μιας επεξηγηματικής μεταβλητής στο επίπεδο-σχολείο (τύπος σχολείου), και δύο επεξηγηματικών μεταβλητών στο επίπεδο-μαθητής (φύλο και βαθμολογία εξετάσεων δύο χρόνια νωρίτερα). Παρατηρούμε ότι η μεταβλητή y αντιστοιχεί στο επίπεδο 1 (μαθητές), με το σχολείο 1 να έχει τρεις μαθητές και το σχολείο 4 τέσσερις μαθητές, δηλαδή τα δεδομένα δεν είναι ισορροπημένα. Στα πολυεπίπεδα μοντέλα δεν απαιτείται να υπάρχει ο ίδιος αριθμός χαμηλότερων μονάδων σε κάθε υψηλότερο επίπεδο μονάδας. Σε αυτό το παράδειγμα (και κατά κοινή σύμβαση), ο δείκτης i χρησιμοποιείται για να καταδείξει το χαμηλότερο επίπεδο μονάδων (μαθητής), ενώ ο δείκτης j τα σχολεία.

Με ένα τέτοιου είδους πλαίσιο δεδομένων θα μπορούσαμε να κάνουμε μια σειρά ερωτήσεων με τη χρήση ενός διεπίπεδου μοντέλου στο οποίο η σημερινή επίδοση ενός μαθητή έχει σχέση με την προηγούμενη και υπάρχουν διαθέσιμα δεδομένα για το φύλο του μαθητή και το δημόσιο/ιδιωτικό χαρακτήρα του σχολείου. Αυτές περιλαμβάνουν: Ποιος σημειώνει μεγαλύτερη πρόοδο; Τα αγόρια ή τα κορίτσια; Ποιοι μαθητές σημειώνουν μεγαλύτερη πρόοδο; Εκείνοι που φοιτούν σε ιδιωτικά ή σε δημόσια σχολεία; κ.ο.κ...

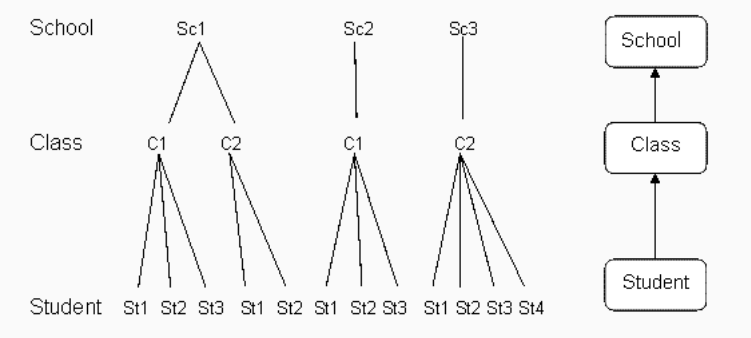
Άλλα κοινά παραδείγματα των διεπίπεδων δομών αποτελούν οι άνθρωποι μέσα στα νοικοκυριά, οι ασθενείς μέσα στα νοσοκομεία, και οι άνθρωποι μέσα σε γειτονιές. Όλα αυτά τα παραδείγματα, όπως και το παράδειγμα των μαθητών σε σχολεία, προκύπτουν όταν ο πραγματικός κόσμος έχει μια πολυστρωματική δομή, δηλαδή τα επίπεδα υπάρχουν στον πληθυσμό.

Επίπεδα		Απόκριση	Επεξηγηματικές Μεταβλητές		
Μαθητής i	Σχολείο j	Βαθμολογία εξέτασης ij	Προηγούμενη βαθμολογία μαθητή ij	Φύλο μαθητή ij	Τύπος σχολείου j
1	1	75	56	M	State
2	1	71	45	M	State
3	1	91	72	F	State
1	2	68	49	F	Private
2	2	37	36	M	Private
3	2	67	56	M	Private
1	3	82	76	F	State
2	3	85	50	F	State
1	4	54	39	M	Private
2	4	91	71	M	Private
3	4	43	41	M	Private
4	4	66	55	F	Private

Πίνακας 1.1: *Πλαίσιο Δεδομένων των Σχημάτων 1.1 και 1.2: Διεπίπεδη μελέτη που εξετάζει τις επιπτώσεις του σχολείου στην επίδοση των μαθητών.*

1.2.2 Τριεπίπεδες ιεραρχικές δομές

Στην ενότητα αυτή θα εξετάσουμε αυστηρά ιεραρχικές δομές τριών επιπέδων μέσω ενός διαγράμματος μονάδων ή ταξινόμησης και πλαισίων δεδομένων. Θα εξετάσουμε μια ανάλυση των μαθητών οι οποίοι είναι εμφωλευμένοι μέσα



Σχήμα 1.3: *Διάγραμμα-μονάδων & διάγραμμα ταξινόμησης τριεπίπεδης δομής: μαθητές σε τάξεις σε σχολεία*

σε τάξεις και αυτές με τη σειρά τους σε σχολεία. Η δομή έχει μια αυστηρή ιεραρχία, αν κάθε μαθητής ανήκει σε μία και μόνο μία τάξη και κάθε ομάδα στην τάξη βρίσκεται σε ένα και μόνο ένα σχολείο. Η μονάδα και τα διαγράμματα κατάταξης δίδονται στο Σχήμα 1.3. Το πλαίσιο των δεδομένων του Πίνακα 1.2 δείχνει τη δομή που θα μπορούσε να χρησιμοποιηθεί για την εκτίμηση της επίδρασης του τρόπου διδασκαλίας του καθηγητή (επίσημη ή ανεπίσημη) σχετικά με την πρόοδο και να αξιολογεί κατά πόσον η πρόοδος αυτή είναι διαφορετική στα δημόσια σε σχέση με τα ιδιωτικά σχολεία. Στη συγκεκριμένη μελέτη κάθε τάξη διδάσκεται από έναν δάσκαλο (όπως συμβαίνει συχνά σε ένα δημοτικό σχολείο) και έτσι είναι αδύνατο να διαχωριστούν οι διαφορές μεταξύ των καθηγητών και οι διαφορές μεταξύ των τάξεων αντίστοιχα καθώς αυτές περιπλέκονται.

Σε μελέτη τριών επιπέδων μπορεί να υπάρξουν ανισορροπίες σε κάθε ένα από τα υψηλότερα επίπεδα, έτσι μπορεί να υπάρξει ένας διαφορετικός αριθμός μαθητών σε κάθε τάξη και ένας διαφορετικός αριθμός τάξεων σε κάθε σχολείο. Είναι επίσης πιθανόν να υπάρχει εξάρτηση σε κάθε ένα από τα υψηλότερα επίπεδα, έτσι ώστε οι μαθητές στην ίδια τάξη είναι πιθανό να παρουσιάζουν περισσότερες ομοιότητες από μαθητές που επιλέγονται τυχαία από διαφορετικές τάξεις, ενώ τάξεις μέσα σε ένα σχολείο είναι επίσης πιθανό να παρουσιάζουν περισσότερες ομοιότητες μεταξύ τους σε σχέση με τάξεις από διαφορετικά σχολεία.

Επίπεδα			Απόκριση	Επεξηγηματικές Μεταβλητές			
Μαθητής i	Τάξη j	Σχολείο k	Βαθμολογία εξέτασης ijk	Προηγούμενη βαθμολογία μαθητή ijk	Φύλο μαθητή ij	Τρόπος διδα-σκαλίας jk	Τύπος σχο-λείου k
1	1	1	75	56	M	Formal	State
2	1	1	71	45	M	Formal	State
3	1	1	91	72	F	Formal	State
1	2	1	68	49	F	Informal	Private
2	2	1	37	36	M	Informal	Private
3	1	2	67	56	M	Formal	Private
2	1	2	82	76	F	Formal	State
3	1	2	85	50	F	Formal	State
1	1	3	54	39	M	Informal	Private
2	1	3	91	71	M	Informal	Private
3	1	3	43	41	M	Informal	Private
4	1	3	66	55	F	Informal	Private

Πίνακας 1.2: *Πλαίσιο Δεδομένων του Σχημάτων 1.3: Τριεπίπεδο μοντέλο μαθητών σε τάξεις, σε σχολεία με μεταβλητή απόκρισης και επιπλέον επεξηγηματικές μεταβλητές.*

ΚΕΦΑΛΑΙΟ 2

Τα Είδη και ο Σκοπός των Πολυεπίπεδων Μοντέλων

2.1 Εισαγωγή

Τα πολυεπίπεδα μοντέλα αντιπροσωπεύουν διαφορετικά επίπεδα συνάθροισης που μπορεί να υπάρχουν στα δεδομένα. Μερικές φορές οι ερευνητές βρίσκονται αντιμέτωποι με δεδομένα που συλλέγονται σε διαφορετικά επίπεδα, έτσι ώστε να παρέχονται τα χαρακτηριστικά των επιμέρους περιπτώσεων αλλά και των αντίστοιχων ομάδων. Επιπρόσθετα, οι ομάδες αυτές μπορούν επίσης να έχουν υψηλότερες ομάδες που συνδέονται με τα χαρακτηριστικά των δεδομένων. Αυτή η ιεραρχική δομή συναντάται συχνά σε δεδομένα κοινωνικών επιστημών ενώ συχνά αγνοείται από τους ερευνητές. Δυστυχώς, η αμέλεια ιεραρχιών στα δεδομένα μπορεί να έχει καταστροφικές συνέπειες για την μετέπειτα στατιστική συμπερασματολογία. Η συχνότητα των εμφωλευμένων δομών στις επιστήμες ανάλυσης δεδομένων είναι εντυπωσιακή. Ειδικότερα, στις κοινωνικές, ιατρικές και βιολογικές επιστήμες οι πολυεπίπεδες ή ιεραρχικές δομές αποτελούν κανόνα. Στις Ηνωμένες Πολιτείες και αλλού, οι επιμέρους ψηφοφόροι είναι εμφωλευμένοι σε εκλογικές περιφέρειες οι οποίες, με τη σειρά τους, είναι εμφωλευμένες σε περιοχές, οι οποίες είναι εμφωλευμένες σε πολιτείες, που είναι εμφωλευμένες τελικά στο κράτος. Στην υγειονομική περίθαλψη, οι

ασθενείς βρίσκονται εμφωλευμένοι σε πτέρυγες, οι οποίες στη συνέχεια είναι εμφωλευμένες σε κλινικές ή νοσοκομεία, κ.ο.κ. Όταν τα άτομα σχηματίζουν ομάδες ή συστάδες (clusters), αναμένεται ότι δύο τυχαία επιλεγμένα άτομα από την ίδια ομάδα τείνουν να παρουσιάσουν περισσότερες ομοιότητες από δύο άτομα που ανήκουν σε διαφορετικές ομάδες. Στο κλασσικό παράδειγμα, μαθητών σε τάξεις εμφωλευμένες σε σχολεία, οι μαθητές που ανήκουν στην ίδια τάξη, εξαρτώνται από τα χαρακτηριστικά των εκπαιδευτικών και την ικανότητα των συμμαθητών τους, πράγμα που επιδρά στο μορφωτικό επίπεδο ενός μαθητή. Για το λόγο αυτό περιμένουμε τη βαθμολογία των μαθητών που ανήκουν στην ίδια τάξη να είναι παρόμοια σε σχέση με εκείνη που λαμβάνουν μαθητές από διαφορετικές τάξεις. Ομοίως, οι μετρήσεις που λαμβάνονται στο ίδιο άτομο σε διαφορετικές περιπτώσεις, π.χ. φυσικές ιδιότητες ή κοινωνικές συμπεριφορές, τείνουν να έχουν υψηλότερο βαθμό συσχέτισης από δύο μετρήσεις σε διαφορετικά άτομα. Για το λόγο αυτό κρίνεται επιτακτική η ανάγκη πολυεπίπεδων (ή ιεραρχικών) μοντέλων, για την ανάλυση δεδομένων με ιεραρχική δομή. Τα πολυεπίπεδα μοντέλα αποτελούν μια ισχυρή και ευέλικτη επέκταση των συμβατικών πλαισίων παλινδρόμησης. Επεκτείνουν τα γραμμικά και γενικευμένα γραμμικά μοντέλα, ενσωματώνοντας επίπεδα απευθείας στο μοντέλο δικαιολογώντας έτσι τη συνάθροιση που παρουσιάζεται στα δεδομένα. Η δομή αυτή κατατάσσει τις περιπτώσεις σε γνωστές ομάδες, οι οποίες μπορεί να έχουν τις δικές τους επεξηγηματικές μεταβλητές στο επίπεδο της ομάδας. Έτσι, μια ιεραρχία καθορίζεται κατά τέτοιο τρόπο ώστε σε ορισμένες επεξηγηματικές μεταβλητές έχει ανατεθεί η ερμηνεία σε ατομικό επίπεδο, ενώ σε κάποιες άλλες η επεξήγηση των διαφορών στο επίπεδο της ομάδας. Με αυτόν τον τρόπο, λαμβάνονται υπόψιν οι συσχετίσεις μεταξύ των υποκειμένων εντός της ομάδας διαχωρίζοντάς τες από εκείνες ανάμεσα στις ομάδες. Με τις εμφωλευμένες δομές λοιπόν, η πολυεπίπεδη προσέγγιση παρέχει ένα σύνολο κρίσιμων πλεονεκτημάτων έναντι της συμβατικής μοντελοποίησης, όπου δεν λαμβάνεται υπόψιν η ετερογένεια και η συσχέτιση των δεδομένων. Ο πυρήνας ενός πολυεπίπεδου μοντέλου αποτελείται από μια εξίσωση παλινδρόμησης που συσχετίζει την μεταβλητή απόκρισης με ένα σύνολο επεξηγηματικών μεταβλητών θυμίζοντας γενικό ή γενικευμένο γραμμικό μοντέλο. Η παρέκκλιση προέρχεται μέσω της επεξεργασίας ορισμένων συντελεστών των επεξηγηματικών μεταβλητών.

Βασικές δομές μοντέλων

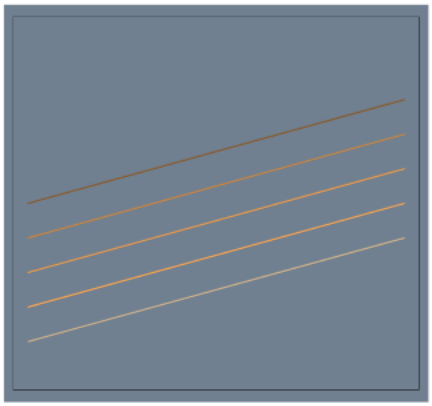
Η ανάπτυξη πολυεπίπεδων μοντέλων ξεκινά με τον προσδιορισμό ενός απλού γραμμικού μοντέλου για μεμονωμένα i:

$$y_i = \beta_0 + x_i\beta_1 + \varepsilon_i \ , \tag{2.1}$$

Ας υποθέσουμε πως υπάρχει ετερογένεια εφόσον κάθε περίπτωση i ανήκει σε μία από τις $j = 1, ..., J$ ομάδες με $J < n$. Εάν δε μας δίνεται πληροφορία για την επεξηγηματική μεταβλητή για τις j περιπτώσεις, η προσαρμογή του μοντέλου μπορεί να βελτιωθεί αντιστοιχώντας κάθε περίπτωση i στην αντίστοιχή της ομάδα. Χαλαρώνοντας τον ορισμό του σταθερού όρου, β_0 στην (2.1) σε j διακριτούς σταθερούς όρους, β_{0j} που στη συνέχεια ομαδοποιεί τις n περιπτώσεις δίνοντας τους έναν κοινό σταθερό όρο με άλλες περιπτώσεις αν βρίσκονται στην ίδια ομάδα. Για $i = 1, ..., n$:

$$y_i = \beta_{0j} + x_i\beta_1 + \varepsilon_i \tag{2.2}$$

όπου ο προστιθέμενος j δείκτης δείχνει ότι η i-οστή περίπτωση έχει j σταθερό όρο για την j ομάδα. Οι β_{0j} έχουν κοινή κανονική κατανομή με τον μέσο β_0 και τυπική απόκλιση σ_{u0}. Αφού ο συντελεστής β_1 δεν κατατάσσεται από τον όρο της ομάδας j, ξέρουμε ότι είναι ακόμα σταθερός για τις n περιπτώσεις και υπολογίζεται με την εκτιμήτρια τυπικού σημείου (standard point estimate). Αυτό αναπαρίσταται στο σχήμα παρακάτω και δείχνει ότι ενώ διαφορετικές



Σχήμα 2.1: Τυχαίες σταθερές

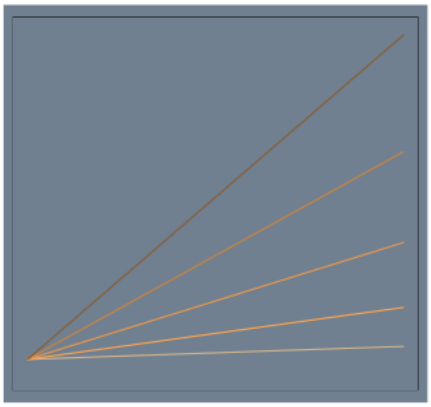
ομάδες ξεκινούν από διαφορετικούς σταθερούς όρους εξελίσσονται με τον ίδιο

ρυθμό (κλίση). Αυτό το μοντέλο είναι επαρκώς θεμελιώδες ώστε να έχει δική του ονομασία, *μοντέλο τυχαίων σταθερών (varying-intercept model ή random-intercepts model).*

Σε ένα διαφορετικό περιεχόμενο, μπορεί να θέλουμε να αντιστοιχίσουμε τις i περιπτώσεις σε j ομάδες αλλά το επιδράσεις δεν προκύπτει μέσω του σταθερού όρου όπου οι ομάδες ξεκινούν σε ένα μηδενικό επίπεδο της επεξηγηματικής μεταβλητής x. Τώρα έχουμε λόγο να πιστεύουμε ότι η ομαδοποίηση επηρεάζει τις κλίσεις αντί αυτού: καθώς το x αυξάνεται το y μεταβάλλεται με διαφορετικό τρόπο λόγω της ομάδας. Έτσι μπορούμε να χαλαρώσουμε τον ορισμό της μοναδικής κλίσης, β_1 στην (2.1) για να μπορέσουμε να αντιστοιχίσουμε τις i περιπτώσεις με τις j ομάδες σύμφωνα με :

$$y_i = \beta_0 + x_i\beta_{1j} + \varepsilon_i \tag{2.3}$$

όπου ο προστιθέμενος j δείκτης δείχνει ότι η i-οστή περίπτωση έχει κλίση j για την j-οστή ομάδα. Ο σταθερός όρος δε μεταβάλλεται μεταξύ των περιπτώσεων στα δεδομένα και οι κλίσεις έχουν κοινή κανονική κατανομή. Αυτό αναπαρίσταται στο παρακάτω σχήμα και δείχνει απόκλιση από το ίδιο σημείο εκκίνησης για τις ομάδες καθώς το x αυξάνει. Το μοντέλο αυτό είναι επίσης



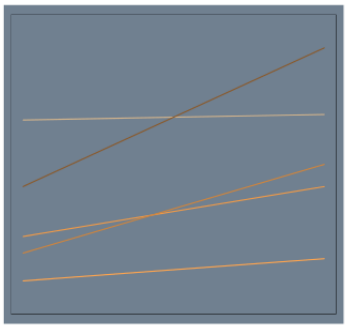
Σχήμα 2.2: Τυχαίες κλίσεις

θεμελιώδες αρκετά ώστε να έχει δική του ονομασία, *μοντέλο τυχαίων κλίσεων (varying-slope model ή random-slope model).* Αν υποθέσουμε πως η ετερογένεια στο δείγμα κατά μήκος των i περιπτώσεων είναι αρκετά πολύπλοκη, χρειάζεται να προσαρμοστεί και ως προς τις τυχαίες σταθερές και ως προς τις τυχαίες

κλίσεις. Αυτός είναι ένας απλός συνδυασμός των δύο προηγούμενων μοντέλων και έχει τη μορφή:

$$y_i = \beta_{0j} + x_i\beta_{1j} + \varepsilon_{ij}$$

όπου η συμμετοχή στην ομάδα j για την περίπτωση i έχει δύο επιδράσεις, από τη μία μεριά ότι είναι σταθερή και από την άλλη ότι διαφέρει από άλλες με αύξηση του x. Τα διανύσματα $(\beta_{0j},\ \beta_{1j})$ έχουν κοινή πολυμεταβλητή κανονική κατανομή. Ένα σύνθετο, πιθανώς υπερβολικό, αποτέλεσμα του μοντέλου δίνεται παρακάτω. Χωρίς να προκαλείται έκπληξη ονομάζεται *μοντέλο*



Σχήμα 2.3: Τυχαίες σταθερές-Τυχαίες κλίσεις

τυχαίων σταθερών και τυχαίων κλίσεων (varying- intercepts, varying-slope model ή random-intercepts, random-slope model)[1].

2.2 Το Φάσμα του Μοντέλου

Σύμφωνα με την έμφαση που δίνουν οι Gelman και Hill (2007, τα πολυεπίπεδα μοντέλα μπορούν να θεωρηθούν ότι υπάρχουν ανάμεσα από δύο άκρα που είναι διαθέσιμα όταν οι ομαδοποιήσεις είναι γνωστές: fully-pooled και fully-unpooled. Το fully-pooled μοντέλο αντιμετωπίζει τις μεταβλητές στο επίπεδο των ομάδων ως μεμονωμένες μεταβλητές, που σημαίνει ότι αγνοούμε τις διακρίσεις στο επίπεδο των ομάδων και αυτές οι επιδράσεις αντιμετωπίζονται σαν να είναι προσδιοριζόμενες ανά περίπτωση. Για ένα μοντέλο με μία επεξηγηματική μεταβλητή που μετριέται στο επίπεδο 1 (x_1) και μία που μετριέται

[1]Προς χάριν ευκολίας στα κεφάλαια που ακολουθούν θα ονομάζεται μοντέλο τυχαίων σταθερών(random- slope model).

στο επίπεδο 2 (x_2) ο προσδιορισμός είναι:

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i \tag{2.4}$$

Στον τελευταίο προσδιορισμό δεν έχουμε ορίσει τυχαίες επιδράσεις u_{0j}. Ισχυριζόμαστε ότι οι διακρίσεις μεταξύ των ομάδων δεν έχουν σημασία και όλες οι περιπτώσεις πρέπει να αντιμετωπίζονται με ομοιογένεια, αγνοώντας την (πιθανώς σημαντική) μεταβολή ανάμεσα στις κατηγορίες.[2] Στην άλλη άκρη του φάσματος είναι ένα σύνολο μοντέλων στα οποία αντιμετωπίζουμε κάθε ομάδα ως ξεχωριστό σύνολο δεδομένων και τα μοντελοποιούμε τελείως ξεχωριστά:

$$y_{ij} = \beta_{0j} + x_{ij}\beta_{1j} + \varepsilon_{ij} \tag{2.5}$$

για $j = 1, ..., J$. Ας σημειώσουμε ότι η επεξηγηματική μεταβλητή στο επίπεδο των ομάδων x_2 δεν εισάγεται σε αυτή την εξίσωση γιατί ο $x_{2i}\beta_2$ είναι σταθερός μέσα σε μια ομάδα και ως εκ τούτου εντάσσεται στο σταθερό όρο. Εδώ δεν υπάρχει επίπεδο 2 στην ιεραρχία και τα β_S θεωρούνται ότι είναι σταθερές παράμετροι. Η fully-unpooled προσέγγιση είναι η αντίθετη διάκριση από την fully-pulled προσέγγιση και ισχυρίζεται ότι οι ομάδες είναι τόσο διαφορετικές που δεν έχει νόημα να τις συσχετίσουμε στο ίδιο μοντέλο. Συγκεκριμένα, οι τιμές των κλίσεων και της σταθεράς από μια ομάδα δεν έχουν σχέση με αυτές στις άλλες ομάδες. Τέτοια ξεχωριστά μοντέλα παλινδρόμησης ξεκάθαρα υπερεκτιμούν την μεταβολή μεταξύ των ομάδων, και τις παρουσιάζουν περισσότερο διαφορετικές από όσο θα έπρεπε να είναι.

Μεταξύ αυτών των δύο πολικών διακρίσεων των ομάδων βρίσκεται το πολυεπίπεδο μοντέλο. Η λέξη "μεταξύ" εδώ σημαίνει ότι οι ομάδες αναγνωρίζονται ως διαφορετικές αλλά επειδή υπάρχει ένα μοναδικό μοντέλο συσχετίζονται με κοινά αποτελέσματα σε επίπεδο ατόμου και με υποθέσεις κατανομής στις τυχαίες επιδράσεις. Για αυτό το λόγο το μοντέλο που προκύπτει συμβιβάζεται μεταξύ πλήρους διάκρισης των ομάδων και πλήρους αγνόησής τους. Αυτό μπορεί να θεωρηθεί ως partial-pooling ή semi-pooling με την έννοια ότι οι ομάδες συλλέγονται μαζί σε ένα μοναδικό μοντέλο, αλλά οι διακρίσεις τους διατηρούνται.

Για να αναπαραστήσουμε το παραπάνω ας θεωρήσουμε ένα μοντέλο τυχαίων σταθερών χωρίς επεξηγηματικές μεταβλητές:

$$y_{ij} = \beta_{0j} + \varepsilon_{ij} \tag{2.6}$$

[2] Στα επόμενα κεφάλαια εξετάζουμε μοντέλα στα οποία ορίζονται επιδράσεις.

που επίσης λέγεται μοντέλο μέσου (mean model) αφού το β_{0j} αντιπροσωπεύει τον μέσο της j-οστής ομάδας. Αν υποθέσουμε ότι $\beta_{0j} = \beta_0$ είναι σταθερό μεταξύ όλων των περιπτώσεων, τότε γίνεται το fully-pooled μοντέλο. Αντίστροφα, αν δημιουργήσουμε J ξεχωριστά μοντέλα καθένα με το δικό του β_{0j} τα οποία δεν προέρχονται από κανονική κατανομή, τότε έχουμε την fully-unpooled προσέγγιση. Εκτιμώντας τη (2.6), όπως διατυπώθηκε, δίνει μέσους της ομάδας που είναι ένας σταθμισμένος μέσος όρος των n_j περιπτώσεων στην ομάδα j και το συνολικό μέσο από όλες τις περιπτώσεις. Ορίζουμε πρώτα:

- $\bar{y}_j$ fully-unpooled μέσος για την ομάδα j
- $\bar{y}$ fully-pooled μέσος
- σ_0^2 διακύμανση μέσα στην ομάδα
- σ_1^2 διακύμανση μεταξύ των μέσων εκτιμητριών $\bar{y}_j$
- n_j μέγεθος της ομάδας j.

Μετά μια προσέγγιση της εκτίμησης του πολυεπίπεδου μοντέλου για τον μέσο της ομάδας δίνεται από:

$$\hat{\beta}_{0j} = \frac{\frac{n_j}{\sigma_0^2}\bar{y}_j + \frac{1}{\sigma_1^2}\bar{y}}{\frac{n_j}{\sigma_0^2} + \frac{1}{\sigma_1^2}} \tag{2.7}$$

Αυτή είναι μια πολύ αποκαλυπτική έκφραση. Ο posterior μέσος για μια ομάδα είναι ο σταθμισμένος μέσος όρος της κατανομής από ένα ολικό δείγμα και της κατανομής από αυτή την ομάδα, όπου η στάθμιση εξαρτάται από τις σχετικές διακυμάνσεις και το μέγεθος της ομάδας. Καθώς το μέγεθος της αυθαίρετης j ομάδας μειώνεται, ο $\bar{y}_j$ γίνεται λιγότερο σημαντικός και η εκτιμήτρια των ομάδων ισχυροποιείται από το ολικό δείγμα. Μια ομάδα μηδενικού μεγέθους, ίσως μια υποτιθέμενη περίπτωση, βασίζεται τελείως στο μέγεθος του πλήρους δείγματος, αφού η (2.7) μειώνει το $\beta_{0j} = \bar{y}$. Από την άλλη μεριά, καθώς η ομάδα j μεγαλώνει, η εκτιμήτρια της κυριαρχεί της συνεισφοράς από τον fully-pooled μέσο, και είναι επίσης μια μεγάλη επιρροή σε αυτόν τον fully-pooled μέσο. Αυτό ονομάζεται η συρρίκνωση των επιδράσεων του μέσου προς τον κοινό μέσο. Επιπλέον, καθώς $\sigma_1^2 \to 0$ τότε, $\hat{\beta}_{0j} \to \bar{y}$ και καθώς $\sigma_1^2 \to \infty$ τότε $\hat{\beta}_{0j} \to \bar{y}_j$. Με αυτόν τον τρόπο η επίδραση των ομάδων που είναι στο κέντρο ενός πολυεπίπεδου μοντέλου είναι η ισορροπία ανάμεσα στο μέγεθος της ομάδας που σχετίζεται με το ολικό δείγμα και τις τυπικές αποκλίσεις στα επίπεδα ατόμων και ομάδων.

ΚΕΦΑΛΑΙΟ 3

Εισαγωγή στα Πολυεπίπεδα Μοντέλα

3.1 Συγκρίνοντας ομάδες με τη βοήθεια ιεραρχικών μοντέλων

3.1.1 Πολυεπίπεδο μοντέλο για τις επιδράσεις των ομάδων

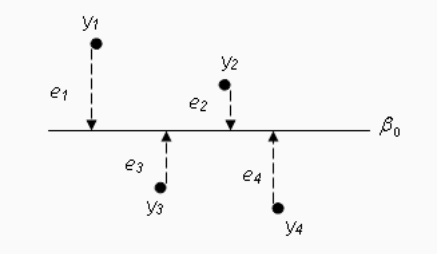
Πριν από την παρουσίαση των πολυεπίπεδων μοντέλων, ας θεωρήσουμε το απλούστερο δυνατό μοντέλο παλινδρόμησης: ένα μοντέλο για τη μέση τιμή της εξαρτημένης μεταβλητής y χωρίς όμως επεξηγηματικές μεταβλητές. Ένα τέτοιο "μηδενικό" (null) μοντέλο μπορεί να γραφεί ως εξής:

$$y_i = \beta_0 + \varepsilon_i \tag{3.1}$$

όπου y_i η τιμή της y για την i περίπτωση ($i = 1, \ldots, n$), β_0 η μέση τιμή του y στον πληθυσμό και ε_i τα υπόλοιπα για την i περίπτωση, τα οποία ακολουθούν κανονική κατανομή με μηδενική μέση τιμή και διασπορά σ^2, δηλαδή

$$\varepsilon_i \sim N(0, \sigma^2).$$

Η διακύμανση συνοψίζει τη μεταβλητότητα γύρω από τη μέση τιμή. Το ακόλουθο σχήμα απεικονίζει τα υπόλοιπα για 4 παρατηρήσεις (n=4).

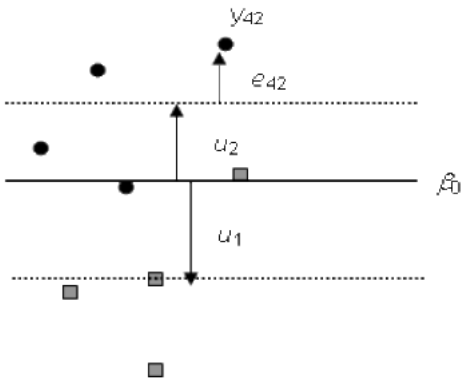


Σχήμα 3.1

Ας προχωρήσουμε τώρα στην απλούστερη μορφή ενός πολυεπίπεδου μοντέλου, το οποίο επιτρέπει διαφοροποιήσεις μεταξύ των ομάδων όσον αφορά τη μέση τιμή της y. Τα δεδομένα μας τώρα έχουν μια διεπίπεδη δομή, με τα άτομα στο επίπεδο 1, εμφωλευμένα σε ομάδες στο επίπεδο 2. Υποδεικνύουμε την ομάδα στην οποία ανήκει το άτομο i, προσθέτοντας ένα δεύτερο δείκτη j έτσι ώστε η y_{ij} να αποτελεί την τιμή της y για το i άτομο στην j ομάδα. Υποθέσουμε ότι υπάρχουν συνολικά j ομάδες με n_j άτομα στην j ομάδα.

Σ' ένα διεπίπεδο μοντέλο χωρίζουμε το υπόλοιπο σε δύο μέρη που αντιστοιχούν στα δύο επίπεδα των δεδομένων μας. Έστω u_j τα υπόλοιπα στο επίπεδο της ομάδας (group level residuals) που είναι γνωστά ως **τυχαίες επιδράσεις των ομάδων (group random effects)** και τα ατομικά υπόλοιπα ε_{ij}. Έτσι η διεπίπεδη προέκταση της σχέσης (3.1) έχει ως εξής:

$$y_{ij} = \beta_0 + u_j + \varepsilon_{ij}. \tag{3.2}$$



Σχήμα 3.2

Το σχήμα (3.2) απεικονίζει τις y-τιμές για 8 άτομα σε 2 ομάδες με τα

άτομα στη 2^{η} ομάδα να υποδηλώνονται με μαύρους κύκλους ενώ εκείνα της $1^{\eta\varsigma}$ ομάδας με γκρίζα τετράγωνα. Η συμπαγής οριζόντια γραμμή αντιπροσωπεύει την ολική μέση τιμή ενώ οι διακεκομμένες τις μέσες τιμές των ομάδων 1 και 2. Το ε_{42} αποτελεί το σχετικό σφάλμα του $4^{ου}$ ατόμου στη 2^{η} ομάδα. Τα υπόλοιπα και στα δύο επίπεδα, ακολουθούν κανονική κατανομή με μηδενική μέση τιμή

$$u_j \sim N(0, \sigma_u^2) \text{ και } \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2).$$

Διαμερίζοντας τη διακύμανση

Ο συντελεστής διαμέρισης διακύμανσης (Variance Partition Coefficient) υπολογίζει το ποσοστό της συνολικής διακύμανσης που οφείλεται στις διαφορές μεταξύ των ομάδων:

$$VPC = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2}.$$

Το εύρος του συντελεστή αυτού, κυμαίνεται μεταξύ 0 (όταν δεν υφίστανται διαφορές μεταξύ των ομάδων, δηλαδή $\sigma_u^2 = 0$) και 1 (όταν δεν εμφανίζονται διαφορές μεταξύ των ατόμων μέσα στις ομάδες, δηλαδή $y_{ij} = y_j$ ή σ_ε^2). Στην περίπτωση που η τιμή του ισούται με 0.2 για παράδειγμα, θα λέγαμε πως το 20% της διακύμανσης εμφανίζεται μεταξύ των ομάδων και το 80% μεταξύ των ατόμων μέσα στις ομάδες. Η συσχέτιση μεταξύ τυχαία επιλεγμένων ζευγών από άτομα της ίδιας ομάδας είναι 0.2.

΄Ελεγχοι για τις επιδράσεις των ομάδων

Μπορούμε να ελέγξουμε την μηδενική υπόθεση ($H_0 : \sigma_u^2 = 0$), δηλαδή ότι δεν υπάρχουν διαφορές μεταξύ των ομάδων, συγκρίνοντας τα μοντέλα (3.1) και (3.2) με τη βοήθεια του ελέγχου λόγου πιθανοφάνειας (Likelihood Ratio Test)

$$LR = -2\log L_1 - (-2\log L_2)$$

όπου $\log L_1$ και $\log L_2$ οι τιμές των λογαριθμοποιημένων πιθανοφανειών των μοντέλων (3.1) και (3.2) αντίστοιχα. Οι τιμές των πιθανοφανειών μπορούν να υπολογιστούν με τη βοήθεια του στατιστικού πακέτου R και μπορούν να χρησιμοποιηθούν για τη σύγκριση οποιουδήποτε ζεύγους από εμφωλευμένα μοντέλα.

Ο στατιστικός έλεγχος LR μπορεί να συγκριθεί με την $\mathcal{X}_n^2$ κατανομή με αριθμό βαθμών ελευθερίας ισοδύναμο με εκείνο των επιπλέον παραμέτρων σε ένα πιο σύνθετο μοντέλο. Στην περίπτωση αυτή, το (3.2) μοντέλο έχει μια

επιπλέον παράμετρο, την διακύμανση μεταξύ των ομάδων σ_u^2, επομένως θα έχουμε ένα βαθμό ελευθερίας.

Η απόρριψη της μηδενικής υπόθεσης ($H_1 : \sigma_u^2 \neq 0$) υπαινίσσεται πως υπάρχουν διαφορές μεταξύ των ομάδων και για το λόγο αυτό ένα πολυεπίπεδο μοντέλο είναι προτιμότερο από ένα απλό γραμμικό ή ένα γενικό γραμμικό μοντέλο. Εντούτοις, αξίζει να σημειωθεί πως στην περίπτωση αποδοχής της μηδενικής υπόθεσης, αν και η προσαρμογή των δεδομένων μας σε ένα γενικό γραμμικό μοντέλο είναι δικαιολογημένη, πρέπει να δοθεί η απαιτούμενη προσοχή στις διαφορές μεταξύ των ομάδων που μπορεί να εμφανιστούν με την προσθήκη νέων επεξηγηματικών μεταβλητών.

3.1.2 Εκτίμηση τυχαίων επιδράσεων των ομάδων

Σε ένα πολυεπίπεδο μοντέλο οι επιδράσεις των ομάδων u_j (υπόλοιπα επιπέδου 2) είναι τυχαίες μεταβλητές που θεωρείται ότι ακολουθούν κανονική κατανομή. Επομένως, η κατανομή τους συνοψίζεται από δύο παραμέτρους, το μέσο (που είναι σταθερός στο 0) και τη διακύμανση σ_u^2. Αυτή η διακύμανση εκτιμάται μαζί με τις άλλες παραμέτρους του μοντέλου (3.2): τη διακύμανση σ_e^2 της υποομάδας και τον ολικό μέσο (ο μέσος που αφορά όλα τα δεδομένα) β_0. Όμως, για να συγκρίνουμε ομάδες μεταξύ τους πρέπει να έχουμε μια εκτίμηση του u_j για κάθε ομάδα. Αυτές οι εκτιμήσεις προκύπτουν αφού προσαρμόσουμε το μοντέλο και βασίζονται στις εκτιμήσεις των παραμέτρων του μοντέλου (β_0, σ_u^2, σ_ε^2) και του δεδομένου y_{ij}

Σε ένα μονοεπίπεδο μοντέλο, έχουμε ένα μόνο σύνολο υπολοίπων. Για μία ατομική μονάδα δείγματος, το υπόλοιπο υπολογίζεται ως η διαφορά μεταξύ της αξίας y_{ij} της παρατήρησης και της αξίας που προβλέπεται από το προσαρμοσμένο μοντέλο, $\hat{y}_{ij}$. Σε ένα πολυεπίπεδο μοντέλο, το συνολικό υπόλοιπο (total residual) είναι $u_j + \varepsilon_{ij}$ που υπολογίζεται ως $r_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \hat{\beta}_0$. Πρέπει αυτό να το χωρίσουμε σε διαφορετικές εκτιμήσεις των u_j και ε_{ij}. Σημείο εκκίνησης για μία εκτίμηση του u_j θα ήταν να πάρουμε το μέσο του $y_{ij} - \hat{\beta}_0$ για την ομάδα j. Αυτό μερικές φορές ονομάζεται το πρώτο υπόλοιπο του μέσου (mean raw residual):

$$\bar{r}_j = \bar{y}_j - \hat{\beta}_0$$

όπου το $\bar{y}_j$ συμβολίζει το δειγματικό μέσο των y_{ij} στην ομάδα j. Για να έχουμε μια εκτίμηση του υπολοίπου για την ομάδα j, πολλαπλασιάζουμε το πρώτο υπόλοιπο (raw residual) με ένα παράγοντα k που ονομάζεται παράγοντας συρ-

ρίκνωσης (shrinkage factor):

$$\hat{u}_j = k\bar{r}_j, \quad \text{όπου } k = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_\varepsilon^2/n_j}.$$

Τα εκτιμώμενα υπόλοιπα $\hat{u}_j$ ονομάζονται συρρικνωμένα υπόλοιπα (shrunken residuals) ή μερικές φορές εμπειρικές Bayes εκτιμητές ή posterior (μεταγενέστεροι) εκτιμητές. Ο παράγοντας συρρίκνωσης k είναι πάντα μικρότερος ή ίσος με το 1 έτσι ώστε το $\hat{u}_j$ να είναι μικρότερο από ή ίσο με τα πρώτα υπόλοιπα του μέσου $\bar{r}_j$. Για μεγάλο n_j ο παράγοντας συρρίκνωσης θα είναι κοντά στο 1 και ως εκ τούτου το $\hat{u}_j$ θα είναι κοντά στο $\bar{r}_j$. Θα υπάρχει επίσης μικρή συρρίκνωση (k κοντά στο 1) όταν $\hat{\sigma}_\varepsilon^2$ είναι μικρό σε σχέση με το $\hat{\sigma}_u^2$. Ο παράγοντας συρρίκνωσης θα είναι αξιοσημείωτα μικρότερος από 1 όταν το n_j είναι μικρό ή το $\hat{\sigma}_\varepsilon^2$ είναι μεγάλο σε σχέση με το $\hat{\sigma}_u^2$ (μεγάλη μεταβλητότητα της υποομάδας). Σε κάθε περίπτωση έχουμε σχετικά λίγη πληροφορία για την ομάδα: λόγω συρρίκνωσης το πρώτο υπόλοιπο (raw residual) τείνει στο μηδέν με αποτέλεσμα ο μέσος της ομάδας $\hat{\beta}_0 + \hat{u}_j$ να τείνει στον ολικό μέσο $\hat{\beta}_0$. Αυτά τα υπόλοιπα συρρίκνωσης (shrinkage residuals) ονομάζονται επίσης σταθμισμένες εκτιμήσεις ακριβείας γιατί έχει ληφθεί υπόψη η αξιοπιστία τους στον υπολογισμό τους. Αναξιόπιστες εκτιμήσεις με μικρό n_j, για παράδειγμα, θα συρρικνωθούν προς τον ολικό μέσο. Αξιόπιστες εκτιμήσεις με μεγάλο n_j θα παραμείνουν κοντά στην αξία του πρώτου μέσου τους (raw mean) [1]. Όπως με κάθε εκτίμηση που βασίζεται σε δεδομένα δείγματος, η παρουσίαση των εκτιμώμενων υπολοίπων επιπέδου 2 $\hat{u}_j$ πρέπει να συνοδεύεται με στάνταρ σφάλματα ή διαστήματα εμπιστοσύνης που δείχνουν την αβεβαιότητα της εκτίμησης εξαιτίας της μεταβλητότητας του δείγματος. Τα βαθμονομημένα υπόλοιπα με διαστήματα εμπιστοσύνης μπορούν να αναπαρασταθούν γραφικά με το ονομαζόμενο caterpillar (κάμπια, λόγω σχήματος) γράφημα.

3.1.3 Παράδειγμα

Τα δεδομένα που θα αναλύσουμε προέρχονται από την Έρευνα Σκοτών Αποφοίτων (Scottish School Leavers Survey), μια αντιπροσωπευτική, σε εθνικό επίπεδο, μελέτη για νέους. Χρησιμοποιούμε στοιχεία από 7 μελέτες κοορτής νέων που συλλέγονται κατά την πρώτη πραγματοποίηση της μελέτης η οποία διεξήχθη κατά το τελευταίο έτος της υποχρεωτικής εκπαίδευσης (ηλικίες 16-17), όπου τα περισσότερα μέλη του δείγματος είχαν δώσει Τελικές εξετάσεις

[1] Αξίζει να σημειώσουμε πως έχοντας εκτιμήσει τα υπόλοιπα επιπέδου 2, οι εκτιμήσεις των υπολοίπων επιπέδου 1 λαμβάνονται από την αφαίρεση: $\hat{\varepsilon}_{ij} = r_{ij} - \hat{u}_j = y_{ij} - \hat{\beta}_0 - \hat{u}_j$

(Standard grades). Πρόκειται για εξετάσεις βάσει μαθήματος (συνήθως οκτώ τον αριθμό). Κάθε μάθημα βαθμολογείται με κλίμακα από 1 (υψηλότερη) έως 7 (χαμηλότερη). Η εξαρτημένη μεταβλητή είναι η συνολική βαθμολόγια που λαμβάνει ο υποψήφιος, καταχωρώντας το 7 στο '1', το 6 στο '2' κ.ο.κ. Έτσι μια υψηλή τιμή για τη μεταβλητή θα υποδηλώνει υψηλή βαθμολογία. Μέσω ενός

Όνομα μεταβλητής	Περιγραφή
caseid	Ανώνυμο αναγνωριστικό μαθητή
schoolid	Ανώνυμο αναγνωριστικό σχολείου
score	Η βαθμολογία υπολογίζεται από τις επιδόσεις στις τελικές εξετάσεις (Standard grades) που δόθηκαν από τους μαθητές στην ηλικία των 16. Η βαθμολογία κυμαίνεται μεταξύ 0 και 75, με τη μέγιστη βαθμολογία να καταδεικνύει την υψηλότερη επίδοση
cohort90	Το δείγμα περιέχει τις ακόλουθες κοορτές: 1984, 1986, 1988, 1990, 1996 and 1998. Η μεταβλητή **cohort90** υπολογίζεται αφαιρώντας την τιμή 1990 από κάθε έτος. Για αυτό οι τιμές κυμαίνονται από -6 (αντιστοιχεί στο 1984) έως 8 (1998), με το 1990 αντιστοιχεί στο μηδέν
female	Φύλο μαθητή (1=θύλη, 0=άρεν)
sclass	Κοινωνική τάξη ορίζεται η ανώτερη τάξη του πατέρα ή της μητέρας (1=ανώτερη, 2=μεσαία, 3=εργατική, 4=μη κατηγοριοποιημένη)
schtype	Είδος σχολείου ανάλογα με την χρηματοδότησης (1=ιδιωτικά, 0=δημόσια)
schurban	Κατηγοριοποίηση των σχολείων με βάση την τοποθεσία (1=αστική, 0=αγροτική)
schdenom	Κατηγοριοποίηση των σχολείων με βάση αν είναι θρησκευτικό ή όχι (1=Ρωμαιοκαθολικό, 0=μη-Ρωμαιοκαθολικό)

γενικού μοντέλου παλινδρόμησης μας δίνεται η δυνατότητα να εξετάσουμε τη σχέση μεταξύ της επίδοσης του μαθητή (**score**) και έτους (**cohort90**). Έτσι μπορούμε να απαντήσουμε στο ερώτημα για το αν η βαθμολογία άλλαξε με την πάροδο του χρόνου και για το αν η σχέση είναι γραμμική. Εδώ, επιτρέποντας την εξάρτηση της βαθμολογίας, μεταξύ των σχολείων θα εξετάσουμε την έκταση της διασποράς της βαθμολογίας ανάμεσα στα σχολεία. Επιπροσθέτως, θα λάβουμε υπόψιν μας και τις επιδράσεις μερικών μεταβλητών στην επίδοση στο

επίπεδο του σχολείου. Το σετ δεδομένων περιέχει τις μεταβλητές στο επίπεδο του μαθητή καθώς και 3 μεταβλητές επιπλέον στο επίπεδο του σχολείου. Θα ξεκινήσουμε με το πιο απλό πολυεπίπεδο μοντέλο εισάγοντας τις επιδράσεις για τα σχολεία στην επίδοση του μαθητή χωρίς την παρουσία επεξηγηματικών μεταβλητών. Αυτό το "μηδενικό" μοντέλο μπορεί να γραφεί:

$$score_{ij} = \beta_0 + u_{0j} + \varepsilon_{ij},$$

όπου $score_{ij}$ είναι η βαθμολογία του μαθητή i στο σχολείο j, β_0 η ολική μέση τιμή για τα σχολεία, u_{0j} η επίδραση του σχολείου j στη βαθμολογία και ε_{ij} το υπόλοιπο στο επίπεδο του μαθητή. Οι επιδράσεις για τα σχολεία u_{0j}, που αναφέρονται και ως υπόλοιπα για τα σχολεία (ή υπόλοιπα στο επίπεδο 2), θεωρούμε πως ακολουθούν κανονική κατανομή με μηδενική μέση τιμή και διακύμανση σ^2_{u0}. Η βασική εντολή που θα χρησιμοποιήσουμε στην R για να προσαρμόσουμε πολυεπίπεδα μοντέλα είναι μέρος του στατιστικού πακέτου $lme4$ που ανέπτυξαν οι Douglas Bates και Martin Maechel για την προσαρμογή γενικών γραμμικών μοντέλων μεικτών επιδράσεων. Με την προϋπόθεση της εγκατάστασης του πακέτου αυτού στον υπολογιστή μας, το καλούμε με την εντολή $library(lme4)$, έτσι ώστε αργότερα να είμαστε σε θέση να χρησιμοποιήσουμε την εντολή $lmer()$ και να προσαρμόσουμε το μοντέλο μας.

```
> library(lme4)
```

Η σύνταξη της συνάρτησης $lmer()$ στην R είναι παρόμοια με εκείνη της $lm()$ για την προσαρμογή ενός γενικού γραμμικού μοντέλου. Επιλέγουμε να αποθηκεύσουμε το "μηδενικό" μοντέλο ως ένα νέο αντικείμενο με την ονομασία **nullmodel**

```
> nullmodel<- lmer(score ~ (1 | schoolid), data = mydata, REML = FALSE)
```

Το τυχαίο κομμάτι του επιπέδου 2 του μοντέλου καθορίζεται από τη λίστα του τυχαίου κομματιού των επεξηγηματικών μεταβλητών και βρίσκεται σε παρενθέσεις. Ο σταθερός όρος καθορίζεται ρητά με τον αριθμό 1 που ακολουθείται από μία κάθετη γραμμή | και το αναγνωριστικό για το επίπεδο 2 ($schoolid$). Η επιλογή $data$ ορίζει το πλαίσιο δεδομένων που χρησιμοποιούμε, ενώ η επιλογή $REML = FALSE$ χρησιμοποιείται για να δηλώσουμε την εκτίμηση με τη μέθοδο της μέγιστης πιθανοφάνειας. Με τη βοήθεια της εντολής summary έχουμε

```
> summary(nullmodel)
```

Linear mixed model fit by maximum likelihood

$Formula:$ $score \sim (1|schoolid)$

$Data:$ $mydata$

AIC	BIC	$logLik$	$deviance$	$REMLdev$
286545	286570	−143270	286539	286539

Random effects:

$Groups$	$Name$	$Variance$	$Std.Dev.$
$schoolid$	$(Intercept)$	61.024	7.8118
$Residual$		258.357	16.0735

Number of obs: 33988, groups: schoolid, 508

Fixed effects:

	$Estimate$	$Std.\ Error$	$t\ value$
$(Intercept)$	30.6006	0.3693	82.85

Τα παραγόμενα αποτελέσματα αποτελούνται από τρία μέρη. Το πρώτο μέρος κάτω από τα $Formula$ και $data$ εκθέτει μια σειρά από μέτρα καταλληλότητας, δηλαδή αριθμητικές ποσότητες που χρησιμοποιούνται για την αξιολόγηση ενός μοντέλου (AIC, BIC, $Loglik$, ...). Το δεύτερο μέρος, κάτω από τα $Random\ effects$ συνοψίζει την διασπορά και τυπική απόκλιση για κάθε τυχαία επίδραση ξεχωριστά (συμπεριλαμβανομένων και των υπολοίπων στο επίπεδο 1). Στο τέλος του δεύτερου τμήματος δίνεται ο συνολικός αριθμός των παρατηρήσεων (δηλαδή 33988 μαθητές) καθώς και το πλήθος των ομάδων (δηλαδή 508 σχολεία). Το τρίτο και τελικό τμήμα, τα $FixedEffects$, περιέχει την εκτιμήτρια της παραμέτρου ($Estimate$), το τυπικό σφάλμα ($Std.Error$) και την t-τιμή ($t-value$) για κάθε παράμετρο μοντέλου.

Αναλυτικότερα τώρα, η συνολική μέση μαθητική επίδοση (μεταξύ των σχολείων) είναι 30.60. Η μέση τιμή για το σχολείο j εκτιμάται ως $30.60 + \hat{u}_{0j}$, όπου $\hat{u}_{0j}$ το υπόλοιπο για το σχολείο. Ένα σχολείο με $\hat{u}_{0j} > 0$ έχει μέση τιμή υψηλότερη του μέσου όρου, ενώ ένα σχολείο με $\hat{u}_{0j} < 0$ έχει μέση τιμή χαμηλότερη του μέσου όρου.

Η διακύμανση μεταξύ των σχολείων $schoolid$ $(Intercept)$ στην επίδοση των μαθητών είναι $\hat{\sigma}_{u0} = 61.02$ ενώ η διασπορά μεταξύ των μαθητών (μέσα στα σχολεία στο επίπεδο 1) $Residual$ είναι $\hat{\sigma}^2_\varepsilon = 258.36$. Έτσι, η συνολική διακύμανση είναι $61.02 + 258.36 = 319.08$. Επιπροσθέτως, ο VPC ισούται με $61.02/319.38 = 0.19$ που σημαίνει πως το 19% της διασποράς στη μαθητική επίδοση, μπορεί να αποδοθεί στις διαφορές μεταξύ των σχολείων. Αξίζει ωστόσο να σημειωθεί, ότι δεν έχουμε λάβει υπόψιν μας την αντιληπτική ικανότητα του εκάστοτε μαθητή (που μετράται μέσω εισαγωγικών εξετάσεων στη δευτεροβάθμια εκπαίδευση).

Προηγούμενες μελέτες έδειξαν πως η διακύμανση μεταξύ των σχολείων όσον αφορά την πρόοδο, δηλαδή έχοντας λάβει υπόψιν την αντιληπτική ικανότητα του μαθητή, είναι κοντά στο 10% .Για να ελέγξουμε τη σημαντικότητα των σχολικών επιδράσεων μπορούμε να χρησιμοποιήσουμε ελέγχους λόγου πιθανοφάνειας. Θα συγκρίνουμε το μηδενικό πολυεπίπεδο μοντέλο που ορίσαμε παραπάνω (null model) με το αντίστοιχο μηδενικό απλό γραμμικό. Για την προσαρμογή του τελευταίου, απομακρύνουμε τις τυχαίες επιδράσεις για τα σχολεία, u_{0j}, οπότε έχουμε:

$$score_{ij} = \beta_0 + \varepsilon_{ij}$$

Με τη βοήθεια της συνάρτησης $lm()$ προσαρμόζουμε το μοντέλο και το αποθηκεύουμε σε ένα νέο αντικείμενο με το όνομα fit.

```
> fit <- lm(score ~ 1, data = mydata)
```

Εν συνεχεία, με τη βοήθεια της $loglik$ εντολής λαμβάνουμε τις λογαριθμοποιημένες τιμές των πιθανοφανειών για κάθε μοντέλο :

```
> logLik(nullmodel)
'log Lik.' -143269.5 (df=3)
> logLik(fit)
'log Lik.' -145144.4 (df=2)
```

$LR = 2\log L_1 - (-2\log L_2) = 2[-143269.5 - (-145144.4)] = 3750$ με ένα βαθμό ελευθερίας.

Λαμβάνοντας υπόψιν ότι για την $\mathcal{X}^2_n$ με $n = 1$ βαθμό ελευθερίας και συντελεστή εμπιστοσύνης $\gamma = 0.95$ το στατιστικό ελέγχου για τη μηδενική υπόθεση είναι 3.84. Προφανώς λοιπόν, θα απορρίψουμε τη μηδενική υπόθεση για το ότι δεν υπάρχουν διαφορές μεταξύ των σχολείων και συμπεραίνουμε πως η ύπαρξη σχολικών επιδράσεων στη μαθητική επίδοση είναι εμφανής με αποτέλεσμα την επαναφορά στο πολυεπίπεδο με σχολικές επιδράσεις.

Εξετάζοντας τις σχολικές επιδράσεις

Για να εκτιμήσουμε τα υπόλοιπα στο επίπεδο του σχολείου $\hat{u}_{0j}$ με τα τυπικά τους σφάλματα, χρησιμοποιούμε την $ranef$ εντολή με την επιλογή $postVar$. Με τον τρόπο αυτό δημιουργούμε ένα αντικείμενο τυχαίων επιδράσεων που περιέχει ένα πίνακα διακύμανσης-συνδιακύμανσης στο $postVar$ χαρακτηριστικό.

```
> u0 <- ranef(nullmodel, postVar = TRUE)
> u0se <- sqrt(attr(u0[[1]], "postVar")[1, , ])
```

Τα υπόλοιπα των 508 σχολείων είναι αποθηκευμένα στη u_0 λίστα, που είναι για την ακρίβεια μια λίστα από λίστες. Το πρώτο στοιχείο της λίστας $u0[[1]]$ είναι η λίστα που αντιστοιχεί στο πρώτο σετ των τυχαίων επιδράσεων. Με τη βοήθεια της εντολής str θα πάρουμε μια περιγραφή της $u0[[1]]$

```
> str(u0[[1]]) List of 1 $ schoolid:'data.frame': 508 obs. of 1 variable:
..$ (Intercept): num [1:508] -11.84 3.21 3.4 -7.42 3.43 ...
..- attr(*, "postVar")= num [1, 1, 1:508] 5.71 1.7 2.24 4.29 2.66 ...
```

Η πρώτη γραμμή ($\$(Intercept)$) δημιουργεί λίστα από σχολικές επιδράσεις. Στην περίπτωση μας, έχουμε μόνο ένα σετ τυχαίων επιδράσεων για αυτό και το $u0[1]$ είναι μια λίστα με ένα αντικείμενο , το $u0[[1]]$. Το $u0[[1]]$ αποτελεί με τη σειρά του ένα πλαίσιο δεδομένων που περιέχει τα υπόλοιπα στο επίπεδο του σχολείου και τις διακυμάνσεις των υπολοίπων αυτών (posterior variances). Η τρίτη γραμμή αντιστοιχεί στο $postVar$ χαρακτηριστικό και δημιουργεί λίστα για τις προαναφερθείσες διακυμάνσεις. Έτσι λοιπόν, τα υπόλοιπα στο επίπεδο του σχολείου μαζί με τα τυπικά τους σφάλματα έχουν υπολογιστεί και αποθηκευτεί για κάθε σχολείο ξεχωριστά. Μπορούμε συνεπώς να προχωρήσουμε στη κατασκευή γραφήματος βασιζόμενοι στα δεδομένα αυτά. Αρχικά, δημιουργούμε ένα πλαίσιο δεδομένων που περιέχει ένα αναγνωριστικό, ένα υπόλοιπο και ένα τυπικό σφάλμα για κάθε σχολείο.

```
> schoolid <- as.numeric(rownames(u0[[1]]))
> u0tab <- cbind(schoolid, u0[[1]], u0se)
> colnames(u0tab) <- c("schoolid", "u0", "u0se")
```

Έπειτα ταξινομούμε το πλαίσιο δεδομένων μας με αύξουσα σειρά βάσει των u_0 τιμών.

```
> u0tab <- u0tab[order(u0tab$u0), ]
```

και δημιουργούμε μια νέα στήλη που περιέχει τις κατατάξεις (ranks).

```
> u0tab <- cbind(u0tab, c(1:dim(u0tab)[1]))
> colnames(u0tab)[4] <- "u0rank"
```

Τέλος κατατάσσουμε το πλαίσιο δεδομένων με βάση το αναγνωριστικό για το σχολείο

```
> u0tab <- u0tab[order(u0tab$schoolid), ]
```

Η εντολή $u0tab[1:10]$ στην R μας επιτρέπει να δούμε τα υπόλοιπα των 10 πρώτων σχολείων, τα τυπικά τους σφάλματα και την κατάταξη τους.

```
> u0tab[1:10, ]
```

	schoolid	*u0*	*u0se*	*u0rank*
1	1	−11.844059	2.390526	37
2	2	3.207216	1.303099	337
3	3	3.396920	1.497759	344
4	4	−7.416852	2.071609	73
5	5	3.427138	1.630506	345
6	6	12.437109	1.403491	487
7	7	−1.652372	1.460226	199
8	8	20.984041	2.021872	508
9	9	−8.693975	6.438403	59
10	10	1.737830	1.904961	291

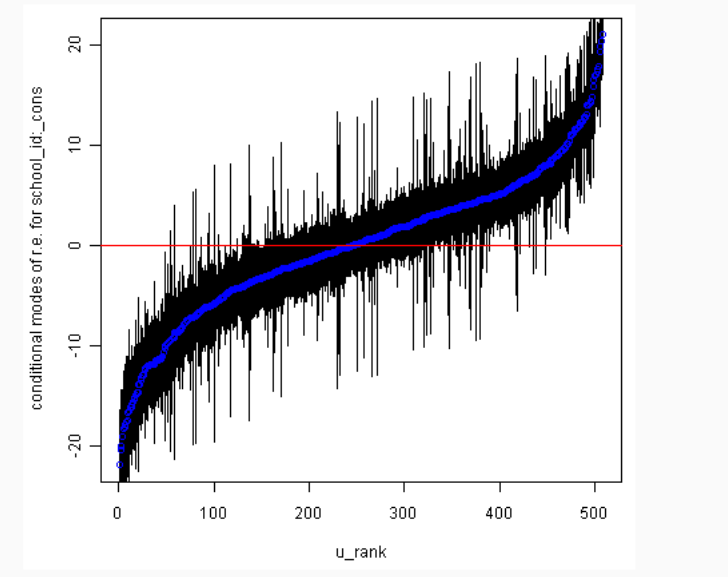
Παραδείγματος χάριν, το σχολείο 1 έχει υπόλοιπο -11.84 και κατατάσσεται 37^o από το τέλος. Για το σχολείο αυτό η μέση βαθμολογία ($score$) είναι $30.60 - 11.84 = 18.76$ σε αντίθεση με το σχολείο 8 (τον 1^o σε κατάταξη, υψηλότερα από όλα) έχει μέση βαθμολογία $30.60 + 20.98 = 51.58$ Με τη βοήθεια των εντολών $plot$ και $Segments$ δημιουργούμε ένα γράφημα κάμπιας (caterpillar plot) ώστε να παρουσιάσουμε τις σχολικές επιδράσεις με σειρά κατάταξης με 95% διάστημα εμπιστοσύνης. Με τη βοήθεια της εντολής

```
> plot(u0tab$u0rank, u0tab$u0, type = "n", xlab = "u_ rank", ylab = "conditional
modes of r.e. for school_ id:_ cons")
```

δημιουργούμε το γράφημα χωρίς τα δεδομένα αρχικά, ενώ με την εντολή

> segments(u0tab$u0rank, u0tab$u0 - 1.96*u0tab$u0se, u0tab$u0rank, u0tab$u0 + 1.96*u0tab$u0se)

προσθέτουμε 95% διάστημα εμπιστοσύνης. Τέλος, προσθέτουμε τα δεδομένα



Σχήμα 3.3: *Γράφημα κάμπια (caterpillar plot) των υπολοίπων για τα σχολεία (school residuals) και 95% δ.ε. για την επίδοση του μαθητή*

μας και μια οριζόντια γραμμή που αντιστοιχεί στο $y = 0$, το μέσο σχολείο

> points(u0tab$u0rank, u0tab$u0, col = "blue")
> abline(h = 0, col = "red")

Ας παρατηρήσουμε ότι τα διαστήματα εμπιστοσύνης γύρω από τις εκτιμήσεις των σφαλμάτων διαφοροποιούνται, άρα μικρότερα σχολεία έχουν ευρύτερα διαστήματα εμπιστοσύνης σε σχέση με τα μεγαλύτερα σχολεία.

3.2 Μοντέλο τυχαίων σταθερών (random intercept model)

Ας επεκτείνουμε τώρα το μοντέλο επιδράσεων (group effects model) (3.2) προσθέτοντας μια επεξηγηματική μεταβλητή. Έστω x_{ij} μια συνεχής επεξηγη-

ματική μεταβλητή ορισμένη στο επίπεδο 1. Οι i και j δείκτες υποδηλώνουν πως οι τιμές της μεταβλητής x διαφέρουν ανά άτομο μέσα σε μια ομάδα. Το απλούστερο μοντέλο με μια επεξηγηματική μεταβλητή έχει ως εξής:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + \varepsilon_{ij} \tag{3.3}$$

Στο μοντέλο αυτό, η συσχέτιση μεταξύ y και x αντιπροσωπεύεται από μια ευθεία γραμμή με συντελεστές β_0 και β_1. Εντούτοις, ο σταθερός όρος β_{0j} για την j ομάδα ισούται με $\beta_0 + u_j$ με

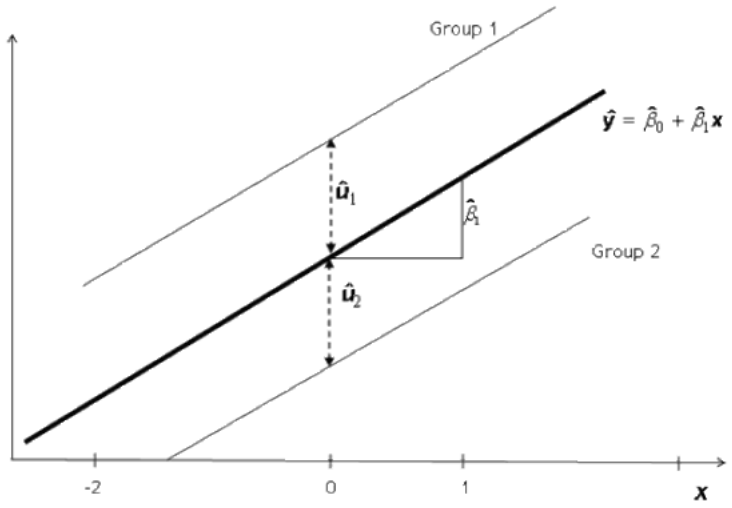
$$u_j \sim N(0, \sigma_u^2).$$

Θα μπορούσαμε να θεωρήσουμε πως ένα πολυεπίπεδο μοντέλο αποτελείται από δύο μέρη: το "σταθερό" κομμάτι $\beta_0 + \beta_1 x_{ij}$ που επεξηγεί τη σχέση της μέσης τιμής της μεταβλητής y με τις επεξηγηματικές μεταβλητές, με παραμέτρους β_0, β_1 και το "τυχαίο" κομμάτι $u_j + \varepsilon_{ij}$ που περιέχει τα υπόλοιπα για τα επίπεδα 1 και 2, με παραμέτρους σ_u^2 και σ_ε^2 αντίστοιχα. Το μοντέλο (3.3) καλείται μοντέλο τυχαίων σταθερών (random intercept model) διότι ο συντελεστής β_0 μπορεί να διαφέρει μεταξύ των ομάδων. Αυτό σημαίνει απλά πως ο σταθερός όρος των γραμμών παλινδρόμησης για τις ομάδες, επιτρέπεται να λαμβάνει διαφορετικές τιμές από την κατανομή. Επισημαίνουμε πως ο συντελεστής είναι προδιαγεγραμμένος για κάθε ομάδα οπότε το μοντέλο γράφεται ως

$$\begin{aligned} y_{ij} &= \beta_{0j} + \beta_1 x_{ij} + \varepsilon_{ij} \\ \beta_{0j} &= \beta_0 + u_j \end{aligned} \tag{3.4}$$

Ενώ λοιπόν ο β_{0j} συντελεστής διαφέρει ανά ομάδα, η κλίση β_1 παραμένει σταθερή για κάθε ομάδα. Έτσι, ένα γράφημα προσαρμογής του μοντέλου απεικονίζει μια σειρά από παράλληλες γραμμές.

$$\hat{y}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 x_{ij} + \hat{u}_j$$



Σχήμα 3.4

3.2.1 Προσθήκη επεξηγηματικών μεταβλητών στο επίπεδο του μαθητή

Ξεκινάμε με την προσθήκη της $cohort90$ στο ''μηδενικό'' μοντέλο της προηγούμενης παραγράφου, οπότε έχουμε

$$score_{ij} = \beta_0 + \beta_1 cohort90_{ij} + u_{oj} + \varepsilon_{ij}.$$

Με τη βοήθεια της εντολής

> fit <- lmer(score ~ cohort90 + (1 | schoolid), data = mydata, REML = FALSE)

προσαρμόζουμε το ανωτέρω μοντέλο στα δεδομένα μας. Καλώντας την ακόλουθη εντολή έχουμε

```
> summary(fit)
Linear mixed model fit by maximum likelihood
  Formula : score ~ cohort90 + (1|schoolid)
     Data :                    mydata
   AIC     BIC     logLik   deviance   REMLdev
 280922  280955  -140457     280914     280921
Random effects:
```

Groups	*Name*	*Variance*	*Std. Dev.*
schoolid	*(Intercept)*	45.988	6.7815
Residual		219.288	14.8084

Number of obs: 33988, groups: schoolid, 508

Fixed effects:

	Estimate	*Std. Error*	*t value*
(Intercept)	30.55913	0.32250	94.76
*cohort*90	1.21496	0.01553	78.24

Correlation of Fixed Effects:

	(Intr)
*cohort*90	−0.002

Στατιστική περιγραφή αποτελεσμάτων

Ξεκινώντας, μας δίνονται κάποια μέτρα καταλληλότητας του μοντέλου, δηλαδή αριθμητικές ποσότητες που χρησιμοποιούνται για την αξιολόγηση του μοντέλου, όπως τα κριτήρια AIC και BIC με τιμές 280922 και 280955 αντίστοιχα. Συγκρίνοντας τις τιμές αυτές με εκείνες για το μοντέλο του $score_{ij} = \beta_0 + u_{0j} + \varepsilon_{ij}$ της προηγούμενης παραγράφου, παρατηρούμε μια μικρή μείωση. Αυτό σημαίνει πως το παρόν μοντέλο είναι σαφώς καταλληλότερο του "μηδενικού". Εν συνεχεία, έχουμε $\hat{\sigma}^2_{u0} = 45.99$ (διακύμανση μεταξύ των σχολείων) με τυπική απόκλιση ίση με 6.78. Η διακύμανση μεταξύ των μαθητών (επίπεδο 1) είναι $\hat{\sigma}^2_{\varepsilon} = 219.29$ με τυπική απόκλιση ίση με 14.81. Ακολούθως, η συνολική μέση επίδοση είναι $\hat{\beta}_0 = 30.56$ με τυπική απόκλιση $se(\hat{\beta}_0) = 0.32$, ενώ έχουμε κλίση $\hat{\beta}_1 = 1.215$ και $se(\hat{\beta}_1) = 0.02$. Συνεπώς, το προσαρμοσμένο μοντέλο έχει ως εξής

$$\widehat{score}_{ij} = 30.559 + 1.215 cohort90_{ij}$$

Το γράφημα του παραπάνω προσαρμοσμένου μοντέλου θα απεικονίζει ένα σετ παραλλήλων γραμμών με σταθερή κλίση $\hat{\beta}_1 = 1.215$.Για την κατασκευή του γραφήματος αυτού πρέπει πρώτα να υπολογίσουμε την επίδοση $(\widehat{score})$ του κάθε μαθητή, βασιζόμενοι στην εκάστοτε κοορτή και το αντίστοιχο σχολείο. Δημιουργούμε λοιπόν μια νέα μεταβλητή $(predscore)$ μέσω της εντολής $fitted$, δηλαδή

```
> predscore <- fitted(fit)
```

Στη συνέχεια με την εντολή

```
> datapred <- unique(data.frame(cbind(predscore = predscore, cohort90 =
mydata$cohort90, schoolid = mydata$schoolid)))
```

δημιουργούμε τη *datapred* μεταβλητή ώστε να επιλέξουμε το μικρότερο ποσοστό δεδομένων που χρειάζονται για τη δημιουργία του γραφήματος. Θέλουμε το γράφημά μας, να απεικονίζει σχολεία για δύο ή παραπάνω κοορτές. Για να γίνει αυτό ταξινομούμε το πλαίσιο δεδομένων ως προς το αναγνωριστικό για το σχολείο (*schoolid*) και την *cohort*90 με τη βοήθεια της εντολής *order* και έπειτα δημιουργούμε μια νέα *multiplecohorts* μεταβλητή με αρχικές τιμές ίσες με το 0.

```
> datapred <- datapred[order(datapred$schoolid, datapred$cohort90), ]
> datapred$multiplecohorts <- rep(0, length(datapred$schoolid))
```

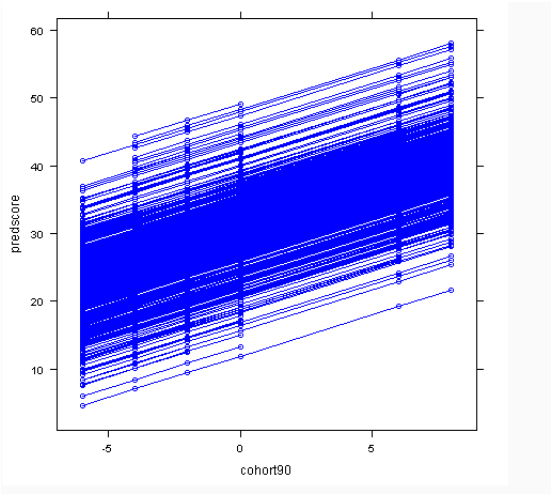
Με την εντολή

```
> datapred$multiplecohorts[datapred$schoolid %in%
unique(datapred$schoolid[duplicated(datapred$schoolid)])] <- 1
```

αντικαθιστούμε τη *multiplecohorts* με την τιμή 1. Στην εντολή αυτή χρησιμοποιήσαμε και τις εντολές *unique* και *duplicated*. Η *duplicated*(*datapred*$*schoolid*) επιστρέφει μια λίστα δεικτών της *schoolid* μεταβλητής. Εφαρμόζοντας την εντολή *unique* στις τιμές της *schoolid* για τις θέσεις που επιστρέφει η *dublicated*, παίρνουμε τη λίστα της μεταβλητής *scoolid* που εμφανίζεται σε περισσότερες από μία κοορτές. Η ακόλουθη εντολή δίνει το ζητούμενο γράφημα το οποίο περιέχει γραμμές σχολείων για περισσότερες από μία κοορτές

```
> xyplot(predscore ~ cohort90, data = datapred[datapred$multiplecohorts ==
1, ], groups = schoolid, type = c("p", "l"), col = "blue")
```

Επιστρέφοντας πάλι στα αποτελέσματα προσαρμογής του μοντέλου μας και συγκρίνοντάς το με το "μηδενικό" της παραγράφου 3.1, συνοψίζουμε πως με την προσθήκη της μεταβλητής *cohort*90 μειώθηκε η διασπορά και στο επίπεδο των σχολείων και σε εκείνο των μαθητών. Πιο συγκεκριμένα η διακύμανση μεταξύ των σχολείων μειώθηκε από 61.02 σε 45.99, ενώ η διακύμανση μεταξύ των μαθητών από 258.36 σε 219.29. Η μείωση της διασποράς "μέσα" στα σχο-



Σχήμα 3.5: Γραμμές πρόβλεψης σχολείων του μοντέλου τυχαίων σταθερών

λεία ήταν αναμενόμενη μιας και η $cohort90$ μεταβλητή αποτελεί μια μεταβλητή που ανήκει στο επίπεδο των μαθητών. Η σημαντική μείωση στη μεταξύ των σχολείων διασπορά προτείνει πως η κατανομή που ακολουθούν οι μαθητές ανά κοορτή διαφέρει από σχολείο σε σχολείο. Τέλος ο συντελεστής VPC μειώθηκε ελαφρά $45.99/(45.99 + 219.29) = 0.017$. Συνεπώς, λαμβάνοντας υπόψιν τις επιδράσεις των κοορτών το 17% της διασποράς στην επίδοση οφείλεται στις διαφορές μεταξύ των σχολείων.

3.3 Επιτρέποντας τη διαφοροποίηση των κλίσεων μεταξύ των ομάδων: Μοντέλα τυχαίων κλίσεων

3.3.1 Μοντέλο τυχαίων κλίσεων

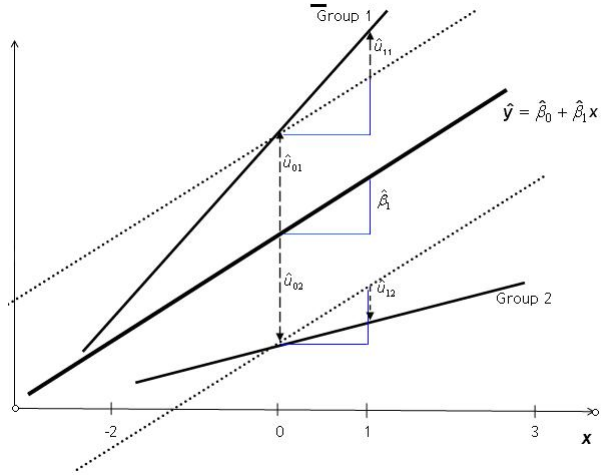
Σ' ένα μοντέλο τυχαίων σταθερών υποθέτουμε ότι η σχέση μεταξύ y και x είναι ίδια για κάθε ομάδα, δηλαδή, η παράμετρος β_1 του μοντέλου δεν διαφοροποιείται μεταξύ των ομάδων. Χαλαρώνουμε τώρα αυτό τον περιορισμό, επιτρέποντας στην β_1 να μεταβάλλεται τυχαία ανά ομάδες, καταλήγοντας σε ένα μοντέλο τυχαίων κλίσεων:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + u_{0j} + u_{1j}x_{ij} + \varepsilon_{ij} \tag{3.5}$$

το οποίο μπορεί επίσης να γραφεί ως εξής:

$$\begin{aligned} y_{ij} &= \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij} \\ \beta_{0j} &= \beta_0 + u_{0j} \\ \beta_{1j} &= \beta_1 + u_{1j} \end{aligned} \tag{3.6}$$

Συγκριτικά με το (3.2), το μοντέλο (3.5) περιέχει ένα επιπλέον όρο $u_{1j}x_{ij}$ ενώ συμπεριλάβαμε ένα επιπρόσθετο '0' δείκτη στον u_j. Οι τυχαίες επιδράσεις u_{0j} και u_{1j} ακολουθούν κανονική κατανομή με μηδενική μέση τιμή και διακύμανση $\sigma^2_{u_0}$ και $\sigma^2_{u_1}$ αντίστοιχα, και συνδιακύμανση $\sigma^2_{u_{01}}$. Ο όρος $u_{1j}x_{ij}$ ορίζεται ως η αλληλεπίδραση μεταξύ του ατόμου και της ομάδας. Στο ακόλουθο σχήμα έχουμε μεταβλητές κλίσεις ανά ομάδες σε αντίθεση με το Σχ. 3.4 όπου παρέμεναν σταθερές. Ας υποθέσουμε για παράδειγμα πως το Σχ. 3.6 απεικονίζει τις γραμμές πρόβλεψης της σχέσης μεταξύ της επίδοσης ενός 16-χρονου μαθητή (y) και την βαθμολογία εισαγωγής του σε σχολείο (x). Η διαφορά μεταξύ των 2 σχολείων διευρύνεται με την αύξηση του x. Έτσι η επιλογή του σχολείου είναι ιδιαιτέρως σημαντική ανάμεσα σε μαθητές με υψηλή βαθμολογία εισαγωγής.



Σχήμα 3.6: *Γραμμές πρόβλεψης ενός μοντέλου τυχαίων κλίσεων*

3.3.2 Παράδειγμα: μοντέλο τυχαίων κλίσεων για συνεχή επεξηγηματική μεταβλητή

Θα επεκτείνουμε τώρα το μοντέλο που προσαρμόσαμε στο τέλος της παραγράφου 3.2 επιτρέποντας την ταυτόχρονη τυχαία μεταβολή μεταξύ των σχο-

λείων, όχι μόνο του σταθερού όρου αλλά και της κλίσης. Το μοντέλο που θα προσαρμόσουμε είναι το ακόλουθο:

$$score_{ij} = \beta_0 + \beta_1 cohort90_{ij} + u_{0j} + u_{1j} cohort90_{ij} + \varepsilon_{ij}$$

Αξίζει να σημειωθεί ότι ο u_{1j} αποτελεί τον νέο όρο που προστέθηκε στο μοντέλο έτσι ώστε ο συντελεστής της $cohort90$ είναι ο

$$\beta_{ij} = \beta_1 + u_{1j}$$

και έτσι η διασπορά στο επίπεδο των σχολείων έχει αντικατασταθεί από ένα πίνακα με δύο παραμέτρους, σ^2_{u0} και σ_{u01}

$$\begin{pmatrix} u_{0j} \\ u_{ij} \end{pmatrix} \sim MVN(0, \Omega_u), \;\; 0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \;\; \Omega_u = \begin{pmatrix} \sigma^2_{u0} & \\ \sigma^2_{u01} & \sigma^2_{u1} \end{pmatrix}$$

Ας σημειώσουμε ότι το υπόλοιπο της κλίσης u_{1j} και η αντίστοιχη διακύμανση σ^2_{u1} και συνδιακύμανση σ_{u01} έχουν δείκτη '1' αφού η $cohort90$ είναι η πρώτη επεξηγηματική μεταβλητή στο μοντέλο. Με χρήση λοιπόν της εντολής

```
> fit <- lmer(score ~ cohort90 + (1 + cohort90 | schoolid), data = mydata, REML = FALSE)
```

προσαρμόζουμε το μοντέλο μας ενώ με την εντολή $summary(fit)$ έχουμε τα ακόλουθα αποτελέσματα:

```
> summary(fit)
 Linear mixed model fit by maximum likelihood
  Formula : score ~ cohort90 + (1 + cohort90|schoolid)
     Data :                    mydata
   AIC    BIC    logLik   deviance   REMLdev
 280698 280749 -140343     280686     280692
Random effects:
 Groups     Name        Variance  Std. Dev.  Corr
 schoolid   (Intercept)  42.85809  6.54661
            cohort90      0.16059  0.40074  -0.390
 Residual                215.73930 14.68807
Number of obs: 33988, groups: schoolid, 508
Fixed effects:
             Estimate  Std. Error  t value
(Intercept)  30.60970   0.31345    97.66
 cohort90     1.23391   0.02532    48.74
```

Correlation of Fixed Effects:
(*Intr*)
*cohort*90 −0.267

Στατιστική περιγραφή αποτελεσμάτων

Ξεκινώντας με τα μέτρα καταλληλότητας του μοντέλου έχουμε τα κριτήρια $AIC = 280698$ και $BIC = 280749$. Η τιμή της λογαριθμοποιημένης πιθανοφάνειας $logik$ ισούται με -140343 και θα χρησιμοποιηθεί παρακάτω στον έλεγχο λόγου πιθανοφάνειας. Η διακύμανση για το σταθερό όρο (intercept variance) ισούται με $\hat{\sigma}^2_{u0} = 42.86$ με τυπική απόκλιση 6.55. Η διακύμανση για την κλίση (slope variance) ισούται με $\hat{\sigma}^2_{u1} = 0.16$ με τυπική απόκλιση 0.40. Ο συντελεστής συσχέτισης είναι $\rho_{u01} = -0.39$. Εν συνεχεία, $\hat{\beta}_0$ ισούται με 30.61 ενώ το τυπικό σφάλμα με $se(\hat{\beta}_0)$ είναι ίσο με 0.31 τέλος παρατηρούμε ότι $\hat{\beta}_1 = 1.234$ και $se(\hat{\beta}_1) = 0.03$.

3.3.3 Ελέγχοντας τις τυχαίες κλίσεις (random slopes)

Μπορούμε να χρησιμοποιήσουμε και εδώ τον έλεγχο λόγου πιθανοφάνειας. Αυτή τη φορά όμως, για να ελέγξουμε αν οι επιδράσεις της μεταβλητής $cohort90$ διαφέρουν μεταξύ των σχολείων. Η μηδενική υπόθεση του ελέγχου αυτού θέλει τις $\sigma_{u01} = \sigma^2_{u1} = 0$. Με τη βοήθεια τώρα της εντολής $anova$, η R προχωράει στην διεξαγωγή ενός LR ελέγχου. Αρχικά όμως θα πρέπει να αποθηκεύσουμε σε ένα νέο αντικείμενο $fita$ το μοντέλο

$$score_{ij} = \beta_0 + \beta_1 cohort90_{ij} + u_{0j} + \varepsilon_{ij}$$

της προηγούμενης παραγράφου για να συγκρίνουμε. Χρησιμοποιούμε το Likelihood ratio test για να ελέγξουμε αν η επίδραση της $cohort90$ μεταβάλλεται μεταξύ των σχολείων.

$$H_0 : \sigma_{u01} = \sigma^2_{u1} = 0$$

Οπότε με τη βοήθεια της εντολής $anova$ έχω

```
> fita <- lmer(score ~ cohort90 + (1 | schoolid), data = mydata, REML = FALSE)
> anova(fit, fita)
Data: mydata
Models:
fita: score ~ cohort90 + (1 | schoolid)
```

fit: score cohort90 + (1 + cohort90 | schoolid)

	Df	AIC	BIC	$logLik$	$Chisq$	$Chi\,Df$	$Pr(>Chisq)$
$fita$	4	280922	280955	-140457			
fit	6	280698	280749	-140343	227.40	2	$< 2.2e-16 * **$

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Ο έλεγχος αυτός συγκρίνει το μοντέλο τυχαίων κλίσεων με εκείνο των τυχαίων σταθερών. Εφόσον η p-τιμή $< 2.2e^{-16} < 0.001$, συμπεραίνουμε πως η επίδραση της $cohort90$ στην μαθητική επίδοση διαφέρει μεταξύ των σχολείων.

3.3.4 Εκτίμηση τυχαίων επιδράσεων των κοορτών μεταξύ των σχολείων

Όπως αναφέραμε κατά τη στατιστική περιγραφή προηγουμένως, η εκτίμηση κοορτής για το σχολείο j είναι $1.234 + \widehat{u}_{ij}$ και διακύμανση μεταξύ των σχολείων $\widehat{\sigma}_{u1}$ ισούται με 0.160. Για το "μέσο" σχολείο λοιπόν προβλέπουμε μια αύξηση κατά 1.234 στην επίδοση για διαδοχικές κοορτές. Ένα 95% διάστημα εμπιστοσύνης για τις κλίσεις των σχολείων είναι $1.234 \pm 1.96\sqrt{0.160} = 0.45$ έως 2.018. Η $\widehat{\sigma}^2_{u0}$ διακύμανση του σταθερού όρου είναι 42.858 όταν η $cohort90 = 0$, δηλαδή για την κοορτή 1990.

3.3.5 Εξετάζοντας τα υπόλοιπα των σταθερών όρων και των κλίσεων των σχολείων

Η αρνητική τιμή της εκτιμήτριας της συνδιακύμανσης ($\widehat{\sigma}_{u01} = -1.024$) υποδηλώνει ότι τα σχολεία με χαμηλότερο σταθερό όρο (επίδοση κάτω του μέσου όρου το 1990) τείνουν να παρουσιάσουν μια πιο σημαντική αύξηση στην επίδοση (άνω του μέσου όρου). Με τη βοήθεια της εντολής $VarCorr$ στην R παράγουμε τον πίνακα συσχέτισης τυχαίων επιδράσεων.

```
> VarCorr(fit)
 $schoolid
             (Intercept)   cohort90
(Intercept)   42.858092 -1.0241779
 cohort90     -1.024178  0.1605916
attr(,"stddev")
(Intercept) cohort90
```

```
6.5466092 0.4007388
attr(,"correlation")
             (Intercept)   cohort90
(Intercept)    1.000000 -0.390389
cohort90      -0.390389  1.000000
attr(,"sc")
[1] 14.68807
```

οπότε η εκτιμήτρια συσχέτισης σταθερού όρου-κλίσης ισούται με

$$\hat{\rho}_{u01} = \frac{\hat{\sigma}_{u01}}{\sqrt{\hat{\sigma}^2_{u0}\hat{\sigma}^2_{u1}}} = -0.390.$$

Για να υπολογίσουμε τις εκτιμήτριες των σταθερών όρων και κλίσεων χρησιμοποιούμε την $raneff$ εντολή με την επιλογή $postVar$. Αποθηκεύουμε τα δεδομένα σε ένα νέο αντικείμενο με το όνομα $myrandomeff$ που θα περιέχει τιμές από τυχαίους σταθερούς όρους αλλά και κλίσεις μαζί με τις διακυμάνσεις τους.

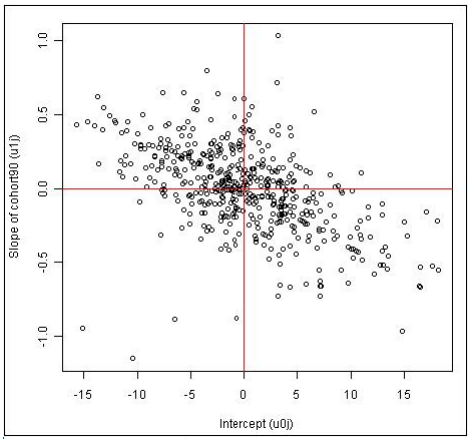
```
> myrandomeff <- ranef(fit, postVar = TRUE)
```

Με το ακόλουθο σύνολο εντολών που παραθέτουμε παρακάτω θα δημιουργήσουμε ένα γράφημα των κλίσεων των σχολείων $\hat{u}_{ij}$ συναρτήσει των σταθερών $\hat{u}_{0j}$

```
> plot(myrandomeff[[1]], xlab = "Intercept (u0j)", ylab = "Slope of cohort90
(u1j)")
> abline(h = 0, col = "red")
> abline(v = 0, col = "red")
```

Από το γράφημα αυτό μπορούμε να προσδιορίσουμε παραδείγματος χάριν, τα σχολεία εκείνα, οι μαθητές των οποίων σημείωσαν επίδοση χαμηλότερη του μέσου όρου το 1990, αλλά και άνω του μέσου όρου επίδοση σημειώνοντας βελτίωση με την πάροδο του χρόνου (όπως τα σχολεία στο δεύτερο τεταρτημόριο). Αντιθέτως, τα σχολεία που ανήκουν στο τέταρτο τεταρτημόριο αποτελούνται από μαθητές η επίδοση των οποίων ήταν χαμηλότερη του μετρίου το 1990, αλλά οι κάτω του μετρίου κλίσεις για τα σχολεία υποδηλώνουν πως οι μαθητές τους συνεχίζουν σε χαμηλά επίπεδα όσον αφορά τις επιδόσεις τους. Το παρακάτω σχήμα απεικονίζει την προσαρμογή του μοντέλου για σχολείο j

$$\widehat{score}_{ij} = (30.610 + \hat{u}_{0j}) + (1.234 + \hat{u}_{ij})cohort90_{ij}$$



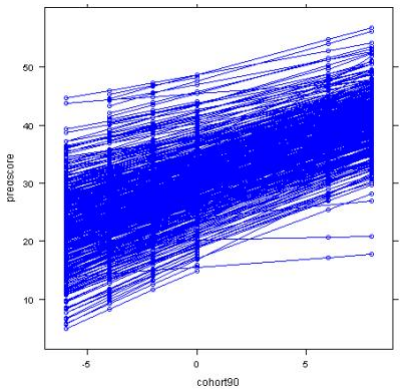
Σχήμα 3.7:

Πριν τη δημιουργία ενός γραφήματος για την προσαρμογή του μοντέλου πρέπει να υπολογίσουμε τη μεταβλητή $scôre$ κάθε μαθητή.

```
> predscore <- fitted(fit)
```

Τώρα όπως και στην προηγούμενη παράγραφο προσαρμόζουμε το μοντέλο μας και προκύπτει το ακόλουθο γράφημα.

```
> datapred <- cbind(predscore = predscore, cohort90 = mydata$cohort90, schoolid
= mydata$schoolid)
> datapred <- data.frame(unique(datapred))
> datapred <-datapred[order(datapred$schoolid,datapred$cohort90),]
> datapred$multiplecohorts <- rep(0, dim(datapred)[1])
> datapred$multiplecohorts[datapred$schoolid %in%
unique(datapred$schoolid[duplicated(datapred$schoolid)])] <- 1
> xyplot(predscore ~ cohort90, data = datapred[datapred$multiplecohorts ==
1,], groups = schoolid, type = c("p","l"), col = "blue")
```



Σχήμα 3.8

3.3.6 Παράδειγμα ενός μοντέλου τυχαίων κλίσεων για δίτιμη επεξηγηματική μεταβλητή

Προσθήκη τυχαίου συντελεστή για το φύλο

Αν προσαρμόσουμε τα δεδομένα μας με ένα γενικό γραμμικό μοντέλο, θα συμπεραίναμε ότι η μέση επίδοση των κοριτσιών στις εξετάσεις είναι καλύτερη από την αντίστοιχη των αγοριών. Για να εξετάσουμε αν η διαφορά του φύλου είναι ίδια μεταξύ των σχολείων, εισάγουμε μια σταθερή επίδραση για το φύλο.

$$score_{ij} = \beta_0 + \beta_1 cohort90_{ij} + \beta_2 female_{ij} + u_{0j} + u_{1j} cohort90_{ij} + \varepsilon_{ij}$$

Με τη βοήθεια τώρα της εντολής[2]

> (fit2a <- lmer(score ~ cohort90 + female + (1 + cohort90 | schoolid), data = mydata, REML = FALSE))

προσαρμόζουμε το ανωτέρω μοντέλο στα δεδομένα μας λαμβάνοντας τα παρακάτω αποτελέσματα

Linear mixed model fit by maximum likelihood
Formula: score ~ cohort90 + female + (1 + cohort90 | schoolid)
Data: mydata

AIC	*BIC*	*logLik*	*deviance*	*REMLdev*
280558	280617	−140272	280544	280552

[2]Οι παρενθέσεις στην εντολή υπάρχουν για την άμεση παρουσίαση των αποτελεσμάτων μετά την προσαρμογή, χωρίς την εντολή $summary()$

Random effects:

Groups	*Name*	*Variance*	*Std. Dev.*	*Corr*
schoolid	*(Intercept)*	42.57457	6.52492	
	cohort90	0.16127	0.40158	−0.393
Residual		214.83738	14.65733	

Number of obs: 33988, groups: schoolid, 508

Fixed effects:

	Estimate	*Std. Error*	*t value*
(Intercept)	29.58494	0.32405	91.30
cohort90	1.22734	0.02533	48.45
female	1.94453	0.16298	11.93

Correlation of Fixed Effects:

	(Intr)	*chrt90*
cohort90	−0.254	
female	−0.265	−0.022

Στατιστική περιγραφή αποτελεσμάτων

Ξεκινώντας κλασσικά με τα μέτρα καταλληλότητας του μοντέλου έχουμε $AIC = 280558$ και $BIC = 280617$. Η τιμή της λογαριθμοποιημένης πιθανοφάνειας ισούται με $logLik = -140272$. Η διακύμανση του σταθερού όρου ισούται με $\hat{\sigma}^2_{u0} = 42.57$ και τυπική απόκλιση ίση με 6.52. Η διακύμανση για την κλίση σ^2_{u1} ισούται με 0.16 και η τυπική απόκλιση 0.40. Η διακύμανση μεταξύ των μαθητών ισούται με $\sigma^2_\varepsilon = 214.83$ και έχει τυπική απόκλιση ίση με 14.66. Εν συνεχεία ο β_0 ισούται με 29.58 και η τυπική απόκλιση είναι ίση με $se(\hat{\beta}_0) = 0.32$. Τέλος, η $\hat{\beta}_1 = 1.22$ με $se(\hat{\beta}_1) = 0.02$ και $\hat{\beta}_2 = 1.94$ και $se(\hat{\beta}_2) = 0.16$.

Εισάγουμε τώρα και μία τυχαία επίδραση για το φύλο οπότε το προηγούμενο μοντέλο γίνεται

$$score_{ij} = \beta_0 + \beta_1 cohort90_{ij} + \beta_2 female_{ij} + u_{0j} + u_{ij} cohort90_{ij} + u_{2j} female_{ij} + \varepsilon_{ij}$$

Προσαρμόζοντάς το με την ακόλουθη εντολή προκύπτουν τα κάτωθεν αποτελέσματα.

```
> (fit2 <- lmer(score ~ cohort90 + female + (1 + cohort90 + female | schoolid),
data = mydata, REML = FALSE))
 Linear mixed model fit by maximum likelihood
Formula: score ~ cohort90 + female + (1 + cohort90 + female | schoolid)
```

Data: mydata

AIC	*BIC*	*logLik*	*deviance*	*REMLdev*
280559	280643	−140269	280539	280547

Random effects:

Groups	*Name*	*Variance*	*Std. Dev.*	*Corr*	
schoolid	*(Intercept)*	40.55760	6.3685		
	cohort90	0.16169	0.4021	−0.394	
	female	1.37140	1.1711	0.206	−0.113
Residual		214.51590	14.6464		

Number of obs: 33988, groups: schoolid, 508

Fixed effects:

	Estimate	*Std. Error*	*t value*
(Intercept)	29.58914	0.31766	93.15
cohort90	1.22777	0.02534	48.44
female	1.93142	0.17391	11.11

Correlation of Fixed Effects:

	(Intr)	*chrt90*
cohort90	−0.253	
female	−0.201	−0.046

Η επίδραση για το φύλο στο σχολείο j είναι $\widehat{\beta}_2 = 1.931 + \widehat{u}_{2j}$. Με την προσθήκη της τυχαίας επίδρασης για το φύλο προέκυψαν τρεις νέες τυχαίες παραμέτρους (σ_{u02}, σ_{u12}, σ^2_{u2}) με $\hat{\sigma}_{u02} = 40.55$, $\hat{\sigma}_{u12} = 0.16$ και $\hat{\sigma}^2_{u2} = 1.37$. Η συσχέτιση μεταξύ του σταθερού όρου και της $female$ μεταβλητής ισούται με $\rho_{02} = 0.206$, ενώ η αντίστοιχη συσχέτιση μεταξύ $cohort90$ και $female$ ισούται με $\rho_{u12} = -0.113$. Επιπρόσθετα, εκτελούμε ένα LR-έλεγχο (Likelihood Ratio test) για να συγκρίνουμε τα μοντέλα της παραγράφου αυτής. Έτσι, η μηδενική υπόθεση έχει ως εξής:

$$H_0 : \sigma_{u02} = \sigma_{u12} = \sigma^2_{u2} = 0$$

Οπότε καλώντας την εντολή $anova$ έχω

```
> anova(fit2, fit2a)
```

Data: mydata

Models:

fit2a: score ∼ cohort90 + female + (1 + cohort90 | schoolid)

fit2: score ∼ cohort90 + female + (1 + cohort90 + female | schoolid)

	Df	*AIC*	*BIC*	*logLik*	*Chisq*	*Chi Df*	*Pr(>Chisq)*
fit2a	7	280558	280617	−140272			
fit2	10	280559	280643	−140269	5.2362	3	0.1553

Παρατηρούμε ότι η p-τιμή $= 0.1553 < 0.05$ πράγμα που σημαίνει ότι δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση οπότε συμπεραίνουμε ότι η επίδραση του φύλου παραμένει η ίδια μεταξύ των σχολείων. Άρα, η επίδοση των κοριτσιών διαφορετικών σχολείων δεν διαφέρει από την επίδοση των αγοριών. Έτσι τελικά επιστρέφουμε σε ένα μοντέλο με σταθερό συντελεστή για τη $female$ μεταβλητή.

3.3.7 Προσθήκη τυχαίου συντελεστή για την κοινωνική τάξη

Σε περίπτωση προσαρμογής των δεδομένων με βοήθεια ενός γενικού γραμμικού μοντέλου θα βλέπαμε ότι η κοινωνική τάξη του γονέα επηρεάζει την επίδοση του μαθητή. Θα εξετάσουμε τώρα αν οι επιδόσεις της κοινωνικής τάξης είναι οι ίδιες μεταξύ των σχολείων. Πριν εισάγουμε τη μεταβλητή για την κοινωνική τάξη στο μοντέλο μας δημιουργούμε 3 ψευδομεταβλητές για τις κατηγορίες 1,2 και 4 αντίστοιχα (θεωρούμε την κατηγορία 3 ως κατηγορία αναφοράς).

```
> mydata$sclass1 <- mydata$sclass == 1
> mydata$sclass2 <- mydata$sclass == 2
> mydata$sclass4 <- mydata$sclass == 4
```

Ξεκινάμε κατά τα γνωστά προσαρμόζοντας το ακόλουθο μοντέλο[3] στην R

$$\begin{aligned} score_{ij} =& \beta_0 + \beta_1 cohort90_{ij} + \beta_2 female_{ij} + \beta_3 sclass1_{ij} \\ &+ \beta_4 sclass2_{ij} + \beta_5 sclass4_{ij} + u_{0j} + u_{1j} cohort90_{ij} + \varepsilon_{ij} \end{aligned}$$

με την εντολή

```
> (fit3a <- lmer(score ~ cohort90 + female + sclass1 + sclass2 + sclass4 + (1
+ cohort90 | schoolid), data = mydata, REML = FALSE))
```

λαμβάνουμε τα ακόλουθα αποτελέσματα

```
Linear mixed model fit by maximum likelihood
Formula: score ~ cohort90 + female + sclass1 + sclass2 + sclass4 + (1 + cohort90
| schoolid)
Data: mydata
   AIC    BIC    logLik  deviance  REMLdev
276712 276797 -138346    276692    276705
```

[3] Προσθήκη σταθερής επίδρασης στην $sclass$

Random effects:

Groups	*Name*	*Variance*	*Std. Dev.*	*Corr*
schoolid	*(Intercept)*	22.51334	4.74482	
	cohort90	0.15084	0.38839	−0.317
Residual		192.94571	13.89049	

Number of obs: 33988, groups: schoolid, 508

Fixed effects:

	Estimate	*Std. Error*	*t value*
(Intercept)	24.60993	0.27962	88.01
cohort90	1.18284	0.02432	48.64
female	1.96135	0.15428	12.71
sclass1TRUE	11.08565	0.20639	53.71
sclass2TRUE	5.87517	0.20405	28.79
sclass4TRUE	−3.73775	0.28453	−13.14

Correlation of Fixed Effects:

	(Intr)	*chrt90*	*female*	*s1TRUE*	*s2TRUE*
cohort90	−0.150				
female	−0.296	−0.023			
sclass1TRUE	−0.395	−0.054	0.008		
sclass2TRUE	−0.386	−0.020	0.009	0.539	
sclass4TRUE	−0.271	−0.036	0.013	0.358	0.357

Δεν θα προβούμε σε στατιστική περιγραφή των αποτελεσμάτων αυτών. Παρ' όλα αυτά είναι αναγκαία για την σύγκριση του μοντέλου αυτού με το ακόλουθο ώστε να εξετάσουμε την ύπαρξη ή μη, διαφοράς στην επίδοση λόγω κοινωνικής τάξης.

Εισάγουμε τώρα τυχαίους συντελεστές στις ψευδομεταβλητές της $sclass$ μεταβλητής. Το μοντέλο έχει ως εξής:

$$\begin{aligned} score_{ij} = {} & \beta_0 + \beta_1 cohort90_{ij} + \beta_2 female_{ij} + \beta_3 sclass1_{ij} + \beta_4 sclass2_{ij} + \beta_5 sclass4_{ij} \\ & + u_{0j} + u_{1j} cohort90_{ij} + u_{3j} sclass1_{ij} + u_{4j} sclass2_{ij} + u_{5j} sclass4_{ij} + \varepsilon_{ij} \end{aligned}$$

όπου

$$\begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{3j} \\ u_{4j} \\ u_{5j} \end{pmatrix} \sim \mathcal{N}(0, \Omega_u), \quad \Omega_u = \begin{pmatrix} \sigma^2_{u0} & & & & \\ \sigma_{u01} & \sigma^2_{u1} & & & \\ \sigma_{u03} & \sigma_{u13} & \sigma^2_{u3} & & \\ \sigma_{u04} & \sigma_{u14} & \sigma_{u34} & \sigma^2_{u4} & \\ \sigma_{u05} & \sigma_{u15} & \sigma_{u35} & \sigma_{u45} & \sigma^2_{u5} \end{pmatrix}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2_\varepsilon)$$

Προσαρμόζουμε τώρα το μοντέλο και έχουμε

```
> (fit3 <- lmer(score ~ cohort90 + female + sclass1 + sclass2 + sclass4 + (1
+ cohort90 + sclass1 +sclass2 + sclass4 | schoolid), data = mydata, REML =
FALSE))
 Linear mixed model fit by maximum likelihood
Formula: score ~ cohort90 + female + sclass1 + sclass2 + sclass4 + (1 + cohort90
+ sclass1 + sclass2 + sclass4 | schoolid)
Data: mydata
```

AIC	BIC	logLik	deviance	REMLdev
276657	276843	−138307	276613	276625

Random effects:

Groups	Name	Variance	Std. Dev.	Corr			
schoolid	(Intercept)	11.26710	3.35665				
	cohort90	0.15599	0.39495	−0.418			
	sclass1TRUE	7.13610	2.67135	0.537	−0.105		
	sclass2TRUE	3.32133	1.82245	0.827	−0.164	0.857	
	sclass4TRUE	7.18489	2.68046	0.898	−0.417	0.439	0.605
Residual		191.76775	13.84802				

Number of obs: 33988, groups: schoolid, 508

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	24.40121	0.23070	105.77
cohort90	1.18452	0.02448	48.38
female	1.96598	0.15408	12.76
sclass1TRUE	11.18175	0.24369	45.88
sclass2TRUE	6.07241	0.22040	27.55
sclass4TRUE	−3.18959	0.31406	−10.16

Correlation of Fixed Effects:

	(Intr)	chrt90	female	s1TRUE	s2TRUE
cohort90	−0.176				
female	−0.359	−0.023			
sclass1TRUE	−0.212	−0.074	0.007		
sclass2TRUE	−0.207	−0.056	0.009	0.585	
sclass4TRUE	−0.042	−0.158	0.013	0.346	0.377

Προχωράμε λοιπόν στη σύγκριση των μοντέλων με τον έλεγχο –LR.

```
> anova(fit3, fit3a)
 Data: mydata
Models:
```

fit3a: score ∼ cohort90 + female + sclass1 + sclass2 + sclass4 + (1 + fit3a: cohort90 | schoolid)
fit3: score ∼ cohort90 + female + sclass1 + sclass2 + sclass4 + (1 +fit3: cohort90 + sclass1 + sclass2 + sclass4 | schoolid)

	Df	AIC	BIC	$logLik$	$Chisq$	$Chi\ Df$	$Pr(>Chisq)$
$fit3a$	10	276712	276797	-138346			
$fit3$	22	276657	276843	-138307	79.122	12	$6.069e-12***$

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Εξαιτίας της πολύ μικρής p-τιμής $(6.069e-12)$ συμπεραίνουμε πως υπάρχουν διαφορές στην επίδοση μεταξύ των σχολείων εξαιτίας της κοινωνικής τάξης.

Σύμφωνα με τα αποτελέσματα προσαρμογής του προηγούμενου μοντέλου οι συντελεστές της $sclass1$, $sclass2$, $sclass3$ και $sclass4$ έχουν ένα σταθερό κομμάτι που εκφράζει τις αντιθέσεις με την κατηγορία αναφοράς (εργατική τάξη) και ένα τμήμα που καθορίζει το σχολείο. Παραδείγματος χάριν, λαμβάνοντας υπόψιν τις μεταβλητές κοορτής και φύλου, τα παιδιά ανώτερης τάξης $(sclass = 1)$ που φοιτούν στο j σχολείο αναμένεται να σημειώσουν βαθμολογία επίδοσης $11.2 + \hat{u}_{3j}$ μονάδες υψηλότερη από τα παιδιά της εργατικής τάξης του ίδιου σχολείου. Ομοίως, τα παιδιά της μεσαίας τάξης που φοιτούν στο j σχολείο σημειώνουν επίδοση $6.1 + \hat{u}_{3j}$ μονάδες υψηλότερη από εκείνα της εργατικής τάξης του ίδιου σχολείου. Αντιθέτως, παιδιά απροσδιόριστης κοινωνικής τάξης σχολείου j σημειώνουν επίδοση $-3.2 + \hat{u}_{3j}$ χαμηλότερη από εκείνα οι γονείς των οποίων ανήκουν στην εργατική τάξη και φοιτούν επίσης στο j σχολείο. Λόγο του αυξημένου πλήθους των παραμέτρων του τυχαίου μέρους του μοντέλου, ο ευκολότερος τρόπος ερμηνείας των τυχαίων συντελεστών της $sclass$ είναι ο υπολογισμός της διακύμανσης μεταξύ των σχολείων.

$$V(u_{0j} + u_{1j}cohort90_{ij} + u_{3j}sclass1_{ij} + u_{4j}sclass2_{ij} + u_{5j}sclass4_{ij}).$$

Θα ακολουθήσουμε αυτή τη διαδικασία για κάθε κατηγορία της $sclass$ μεταβλητής, κρατώντας σταθερή την $cohort90$. Για ευκολία, μηδενίζουμε την $cohort90$ ώστε η διακύμανση μεταξύ των σχολείων να αναφέρεται στο έτος 1990. Έτσι η προηγούμενη σχέση απλοποιείται και έχω:

$$V(u_{0j} + u_{3j}sclass1_{ij} + u_{4j}sclass2_{ij} + u_{5j}sclass4_{ij}).$$

Με την ακόλουθη εντολή έχω τις τιμές της διακύμανσης για τις τυχαίες επιδράσεις

```
> VarCorr(fit3)$schoolid
```

	(Intercept)	cohort90	sclass1TRUE	sclass2TRUE	sclass4TRUE
(Intercept)	11.2670991	−0.5536348	4.8130826	5.0589541	8.0768916
cohort90	−0.5536348	0.1559892	−0.1105532	−0.1179540	−0.4419864
sclass1TRUE	4.8130826	−0.1105532	7.1361021	4.1744010	3.1412509
sclass2TRUE	5.0589541	−0.1179540	4.1744010	3.3213350	2.9574009
sclass4TRUE	8.0768916	−0.4419864	3.1412509	2.9574009	7.1848898

```
attr(,"stddev")
```

(Intercept)	cohort90	sclass1TRUE	sclass2TRUE	sclass4TRUE
3.3566500	0.3949546	2.6713484	1.8224530	2.6804645

```
attr(,"correlation")
```

	(Intercept)	cohort90	sclass1TRUE	sclass2TRUE	sclass4TRUE
(Intercept)	1.0000000	−0.4176092	0.5367681	0.8269863	0.8976936
cohort90	−0.4176092	1.0000000	−0.1047837	−0.1638737	−0.4174954
sclass1TRUE	0.5367681	−0.1047837	1.0000000	0.8574470	0.4386944
sclass2TRUE	0.8269863	−0.1638737	0.8574470	1.0000000	0.6054019
sclass4TRUE	0.8976936	−0.4174954	0.4386944	0.6054019	1.0000000

Χρησιμοποιώντας την R για τον υπολογισμό διακύμανσης μεταξύ των σχολείων έχω

$$V(u_{0j} + u_{3j}) = \sigma^2_{u0} + 2\sigma_{u03} + \sigma^2_{u3}$$

```
> 11.267 + 2*4.813 + 7.136
[1] 28.029
```

που αποτελεί τη διακύμανση για την κατηγορία 1 (sclass1=0, sclass2=0 και sclass4=0). Ομοίως για τη δεύτερη, τρίτη και τέταρτη κατηγορία αντίστοιχα, η διακύμανση είναι:

$$V(u_{0j} + u_{4j}) = \sigma^2_{u0} + 2\sigma_{u04} + \sigma^2_{u4}$$

```
> 11.267 + 2*5.059 + 3.321
[1] 24.706
```

$$V(u_{0j}) = \sigma^2_{u0}$$

```
> 11.267
[1] 11.267
```

$$V(u_{0j} + u_{5j}) = \sigma^2_{u0} + 2\sigma_{u05} + \sigma^2_{u5}$$

```
> 11.267 + 2*8.077 + 7.184
[1] 34.605
```

Παρατηρούμε πως η διακύμανση είναι παρόμοια για τις δύο πρώτες κατηγορίες (ανώτερη και μεσαία τάξη), υψηλότερη για την κατηγορία 4 και χαμηλότερη για τα παιδιά της εργατικής τάξης (κατηγορία 3). Αυτό σημαίνει πως η επίδοση στις εξετάσεις για ένα σχολείο ενδιαφέρει περισσότερο τα παιδιά της τέταρτης κατηγορίας από εκείνα της τρίτης. Για παράδειγμα, η διαφορά στην επίδοση μεταξύ παιδιών απροσδιόριστης τάξης και εκείνων της εργατικής είναι $-3.190 + \hat{u}_{5j}$ με $\widehat{u}_{5j} = 7.182$. Το 95% διάστημα εμπιστοσύνης για τη διαφορά μεταξύ απροσδιόριστης και εργατικής τάξης είναι $-3.190 \pm 196\sqrt{7.184} = -8.443$ έως 2.063. Τέλος, αν κατατάξουμε τα σχολεία σύμφωνα με τις ταξικές τους διαφορές (διαφορά απροσδιόριστης και εργατικής τάξης), έτσι ώστε τα σχολεία με τη μεγαλύτερη διαφορά (υπέρ των παιδιών της εργατικής τάξης) να κατατάσσονται χαμηλότερα. Στο κατώτερο 2.5% των σχολείων τα παιδιά απροσδιόριστης τάξης αναμένεται να σημειώσουν μέση βαθμολογία επίδοσης 8.443 μονάδες χαμηλότερη από τα παιδιά της εργατικής τάξης. Αντίθετα, στο υψηλότερο 2.5% των σχολείων η διαφορά εκτιμάται να είναι μεγαλύτερη από 2.063 υπέρ των παιδιών απροσδιόριστης τάξης.

3.4 Προσθήκη επεξηγηματικών μεταβλητών στο επίπεδο 2

3.4.1 Συναφείς Επιδράσεις (Contextual effects)

Μέχρι στιγμής έχουμε θεωρήσει επεξηγηματικές μεταβλητές που ορίζονται στο χαμηλότερο σημείο της ιεραρχικής δομής. Τα πολυεπίπεδα μοντέλα που έχουμε θεωρήσει μέχρι στιγμής ελέγχουν για ομαδοποίηση, ενώ μας επιτρέπουν να ποσοτικοποιήσουμε το βαθμό της εξάρτησης και να διερευνήσουμε αν οι επιδράσεις των μεταβλητών επιπέδου 1 διαφέρουν μεταξύ αυτών των ομάδων. Ένα συγκεκριμένο πλεονέκτημα της πολυεπίπεδης μοντελοποίησης, παρόλα αυτά, είναι η δυνατότητα να εξερευνά τις επιδράσεις των μεταβλητών που ανήκουν στο επίπεδο των ομάδων. Οι μεταβλητές που ορίζονται σε επίπεδο 2 καλούνται συχνά *συναφείς* μεταβλητές και οι επιδράσεις τους στην y-τιμή ενός ατόμου ονομάζονται *συναφείς* επιδράσεις.

Παραδείγματα ερευνητικών ερωτημάτων που εμπλέκουν συναφείς επιδράσεις συμπεριλαμβάνουν:

- Πώς τα χαρακτηριστικά του δασκάλου (π.χ. ο αριθμός των χρόνων διδασκαλίας) επηρεάζουν την επίδοση των μαθητών; Επηρεάζονται οι επιδό-

σης των μαθητών από την ικανότητες των συμμαθητών τους; Για παράδειγμα, τα παιδιά με μεγάλες δυνατότητες αποδίδουν καλύτερα σε συνδιδασκαλία με άλλα παιδιά με αντίστοιχα μεγάλες δυνατότητες ή με μια ομάδα παιδιών με μικτές ικανότητες; Τα χαρακτηριστικά του δασκάλου και μαζί με εκείνα των συμμαθητών εξηγούν τη διακύμανση της επίδοσης μεταξύ των τάξεων;

- Με ποιους τρόπους το εισόδημα επηρεάζει τις ατομικές επιδράσεις στην υγεία; Τα άτομα με χαμηλό εισόδημα έχουν κακή υγεία γιατί τα πρότυπα του υλικού βιοτικού επιπέδου είναι χαμηλότερα από εκείνους με υψηλότερα εισοδήματα; Για παράδειγμα το χαμηλό εισόδημα, μπορεί να συνδεθεί με χαμηλά πρότυπα στέγασης ή κακή διατροφή. Η σχέση αποδίδει σε επίπεδο περιοχής; Σύμφωνα με τη σχετική εισοδηματική θεωρία, το εισόδημα κάποιου που σχετίζεται με αυτό του γείτονά του καθορίζει σημαντικά την υγεία που λειτουργεί υπεράνω οποιωνδήποτε επιδράσεων σε επίπεδο ατόμου. Με αυτό τον τρόπο, ένα άτομο με χαμηλό εισόδημα που ζει σε μια κοινωνία με μεγάλες εισοδηματικές ανισότητες προβλέπεται να έχει χειρότερη υγεία από κάποιον με χαμηλό εισόδημα που ζει σε μια κοινωνία με υψηλή εισοδηματική ισότητα. Για να δοκιμάσουμε τη σχετική εισοδηματική θεωρία, θα μπορούσαμε να συμπεριλάβουμε στο μοντέλο μας μια συνολική μέτρηση της εισοδηματικής ανισότητας. Για παράδειγμα, σε μια μελέτη των διαφορών στην υγεία για τα άτομα (επίπεδο 1) σε χώρες (επίπεδο 2) μπορούμε να συμπεριλάβουμε το τυπικό σφάλμα του εισοδήματος μεταξύ των χωρών.

- Ποιος είναι ο ρόλος του οικογενειακού ιστορικού στην παιδική υγεία; Ας υποθέσουμε ότι έχουμε μέτρα υγείας για όλα τα παιδιά (επίπεδο 1) σε μια οικογένεια (επίπεδο 2). Πρέπει πρώτα να υπολογίσουμε το ποσό της ολικής διακύμανσης στην υγεία που αποδίδεται στις διαφορές μεταξύ των οικογενειών (μερικός συντελεστής διακύμανσης ή ενδοοικογενειακή συσχέτιση). Στη συνέχεια μπορούμε να διερευνήσουμε αν οποιαδήποτε μεταβολή μέσα στην οικογένεια μπορεί να εξηγηθεί από τα οικογενειακά χαρακτηριστικά, που τα μοιράζονται όλα τα παιδιά στην οικογένεια, όπως η κοινωνικοοικονομική κατάσταση της οικογένειας, το μορφωτικό επίπεδο των γονέων και οι παράγοντες του τρόπου ζωής.

Μια επεξηγηματική μεταβλητή επιπέδου 2 μπορεί να συμπεριληφθεί σε ένα πολυεπίπεδο μοντέλο με τον ακριβώς ίδιο τρόπο για μια μεταβλητή επιπέδου 1. Αν έχουμε για παράδειγμα μια μεταβλητή επιπέδου 1 x_{1ij} και μία μεταβλητή

επιπέδου 2 x_{2j} το μοντέλο τυχαίων σταθερών (3.3) γίνεται[4]

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2j} + u_j + \varepsilon_{ij}. \tag{3.7}$$

Αν η συναφής μεταβλητή είναι η επιπέδου 2 μέση τιμή μιας επιπέδου 1 μεταβλητής, συμπεριλαμβάνεται επίσης στο μοντέλο ώστε η (3.7) γίνεται:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 \overline{x}_j + u_j + \varepsilon_{ij} \tag{3.8}$$

όπου $\overline{x}_j$ είναι ο μέση τιμή των x στην ομάδα j. Στην (3.8) ο β_1 είναι η επίδραση εντός των ομάδων για τη x και $\beta_1 + \beta_2$ είναι η επίδραση των x μεταξύ των ομάδων. Ο εντός της ομάδος συντελεστής μετρά τη σχέση μεταξύ των τιμών των x και y ενός ατόμου μέσα σε μια ομάδα. Ο μεταξύ των ομάδων συντελεστής μετρά τη σχέση μεταξύ των x και y στο επίπεδο των ομάδων, π.χ. του μέσου των x της ομάδας στον μέσο των y της ομάδας. Ο β_2 είναι η συναφής επίδραση των x, που είναι η επίδραση του μέσου των x της ομάδας σε ένα άτομο y το οποίο είναι πάνω από το αποτέλεσμα ενός ατόμου x στον y. Για να μπορούν καθέμιά από τις επιδράσεις εντός της ομάδας και μεταξύ των ομάδων να παρουσιαστούν από μία παράμετρο, η (3.8) μπορεί να επαναδιατυπωθεί ως

$$y_{ij} = \beta_0^* + \beta_1^*(x_{ij} - \overline{x}_j) + \beta_2^* \overline{x}_j + u_j + \varepsilon_{ij} \tag{3.9}$$

όπου $\beta_1^* = \beta_1$ η επίδραση εντός της ομάδας και $\beta_2^* = \beta_1 + \beta_2$ είναι το μεταξύ των ομάδων αποτέλεσμα. Τα μοντέλα (3.8) και (3.9) είναι ισοδύναμα, αλλά το (3.9) παράγει μια άμεση εκτίμηση (και τυπικό σφάλμα) το μεταξύ των ομάδων αποτέλεσμα των x.

3.4.2 Αλληλεπιδράσεις διασταυρούμενων επιπέδων (Cross-level interactions)

Όπως στην πολλαπλή παλινδρόμηση, έτσι κι εδώ, μπορούμε να επιτρέψουμε στην επίδραση μιας επεξηγηματικής μεταβλητής να εξαρτάται από την τιμή μιας άλλης επεξηγηματικής μεταβλητής. Τέτοιου είδους επιδράσεις ονομάζονται επιδράσεις αλληλεπιδράσεων (interaction effects). Οι αλληλεπιδράσεις μπορούν επίσης να συμπεριληφθούν σε ένα πολυεπίπεδο μοντέλο μεταξύ οποιουδήποτε ζεύγους μεταβλητών, ανεξάρτητα από το επιπέδο στο οποίο ορίζονται. Μια αλληλεπίδραση μεταξύ μιας μεταβλητής επιπέδου 1 και μιας μεταβλητής επιπέδου 2 λέγεται αλληλεπίδραση διασταυρούμενων επιπέδων.

[4] Αξίζει να σημειώσουμε πως μια μεταβλητή επιπέδου 2 δεν έχει δείκτη i, γιατί, από ορισμό, οι τιμές του δε μεταβάλλονται από άτομο σε άτομο μέσα στη μονάδα επιπέδου 2

Ένα παράδειγμα αλληλεπίδρασης διασταυρούμενων επιπέδων είναι το εξής: "Τα παιδιά υψηλών ικανοτήτων αποδίδουν καλύτερα στις εξετάσεις σε συνδιδασκαλία με παιδιά των ίδιων δυνατοτήτων με αυτά ή με παιδιά διαφορών επιπέδων δυνατοτήτων; Επιπλέον, μπορούμε να εξετάσουμε το βαθμό που μια αλληλεπίδραση μεταξύ των ατομικών δυνατοτήτων των μαθητών και των δυνατοτήτων της τάξης εξηγεί τις διαφορές μεταξύ σχολείων στη σχέση μεταξύ των αρχικών και μετέπειτα επιδόσεων των μαθητών. Για να διερευνήσουμε αυτές τις ερωτήσεις, θα συμπεριλάβουμε σε ένα μοντέλο τυχαίων κλίσεων για την αρχική επίδοση, ένα μέτρο της αρχικής επίδοσης του παιδιού (x_{ij}), τη μέση επίδοση στην τάξη ($\overline{x}_j$) και την αλληλεπίδρασή τους (x_{ij}*$\overline{x}_j$. Ένα μοντέλο τυχαίων κλίσεων με αλληλεπίδραση διασταυρούμενων επιπέδων είναι μία επέκταση του (3.7):

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2j} + \beta_3 x_{1ij}{}^{*} x_{2j} + u_{0j} + u_{1j} x_{1ij} + \varepsilon_{ij} \tag{3.10}$$

Στο μοντέλο (3.10) ο β_1 ερμηνεύεται ως η επίδραδη των x_1 σε ένα μέσο σχολείο (u_{1j}) όταν $x_2 = 0$. Για κάθε 1 μονάδα αύξησης της x_2 , επίδραση του μεταβάλλεται x_1 κατά μία ποσότητα β_3. Για παράδειγμα, αν $x_2 = 2$ η επίδραση (συντελεστής) του x_1 είναι $\beta_1 + 2\beta_3$.

3.4.3 Παράδειγμα

Στις προηγούμενες παραγράφους είδαμε πως εισάγουμε επεξηγηματικές στο επίπεδο 1 του μοντέλου μας, αλλά και πως ερμηνεύουμε μοντέλα τυχαίων σταθερών και τυχαίων κλίσεων. Εντούτοις, βασικό κίνητρο χρήσης πολυεπίπεδων μοντέλων αποτελεί η αξιολόγηση των επιδράσεων των επεξηγηματικών μεταβλητών του επιπέδου 2 στις μεταβλητές απόκρισης του επιπέδου 1 και το σημείο στο οποίο μπορούν να εξηγήσουν τη διακύμανση του επιπέδου 2. Στην εκπαίδευση για παράδειγμα, μας ενδιαφέρει η συναφή επίδραση (contextual effects) της επίδοσης των μαθητών στην μετέπειτα ακαδημαϊκή τους πορεία. Η πρόοδος ενός μαθητή μπορεί να επηρεαστεί από την απόδοση των συμμαθητών του και η επίδραση αυτή μπορεί να διαφέρει ανάλογα με την προσωπική επίδοση του μαθητή (αλληλεπιδράσεις μεταξύ των επιπέδων). Το σετ δεδομένων του παραδείγματος μας περιέχει 3 μεταβλητές στο επίπεδο του σχολείου που αποτελούν μελλοντικές επεξηγηματικές κατηγορικές μεταβλητές της επίδοσης ενός δεκαεξάχρονου μαθητή. Αυτές είναι: η *schtype* μεταβλητή, δηλαδή ο τύπος του σχολείου (δημόσιο ή ιδιωτικό), η *schurban* μεταβλητή, δηλαδή το είδος της σχολικής τοποθεσίας (αστική ή αγροτική) και η *schdenom* μεταβλητή, δηλαδή το είδος του σχολείου βάσει δόγματος (ρωμαιοκαθολικό ή μη).

Εφόσον κάθε μεταβλητή στο επίπεδο του σχολείου είναι δυαδική, πρέπει να εξετάσουμε το ποσοστό σε κάθε κατηγορία. Με τη βοήθεια λοιπόν των εντολών *table*, *prop.table* και *comsum* δημιουργούμε πίνακες συχνοτήτων, ποσοστών και αθροιστικών ποσοστών αντίστοιχα.

```
> mydata_un <- unique(mydata[,c(2, 7, 8, 9)])
> cbind(Freq = table(mydata_un$schtype), Perc = prop.table(
table(mydata_un$schtype)), Cum = cumsum(prop.table(table(mydata_un$schtype))))
   Freq      Perc       Cum
0   456 0.8976378 0.8976378
1    52 0.1023622 1.0000000
 > cbind(Freq = table(mydata_un$schurban), Perc = prop.table(
table(mydata_un$schurban)), Cum = cumsum(prop.table(table(mydata_un$schurban))))
   Freq      Perc       Cum
0   163 0.3208661 0.3208661
1   345 0.6791339 1.0000000
 > cbind(Freq = table(mydata_un$schdenom), Perc = prop.table(
table(mydata_un$schdenom)), Cum = cumsum(prop.table(table(mydata_un$schdenom))))
   Freq      Perc       Cum
0   425 0.8366142 0.8366142
1    83 0.1633858 1.0000000
```

Από τα παραπάνω αποτελέσματα προκύπτει το ποσοστό των σχολείων στην κατηγορία 1 κάθε μεταβλητής. Συνεπώς το 10% είναι ιδιωτικά, το 68% αυτών βρίσκονται σε αστική τοποθεσία, ενώ το 16% είναι καθολικά. Εισάγουμε τις μεταβλητές αυτές, μία προς μία, σε μια πιο απλοποιημένη μορφή από το μοντέλο που προσαρμόσαμε στην προηγούμενη παράγραφο. Αν και η επίδραση της κοινωνικής τάξης στην επίδοση του μαθητή διαφέρει μεταξύ των σχολείων, θα δουλέψουμε με ένα πιο απλό μοντέλο αφαιρώντας τους τυχαίους συντελεστές από τις ψευδομεταβλητές για την κοινωνική τάξη, το οποίο είναι:

$$score_{ij} = \beta_0 + \beta_1 cohort90_{ij} + \beta_2 female_{ij} + \beta_3 sclass1_{ij} + \beta_4 sclass2_{ij} + \beta_5 sclass4_{ij} + u_{0j} + u_{1j} cohort90_{ij} + \varepsilon_{ij}.$$

Ακολουθεί η προσαρμογή του και η παρουσίαση των αποτελεσμάτων με την εντολή

```
> (fit1a <- lmer(score ~ cohort90 + female + sclass1 + sclass2 + sclass4 + (1
+ cohort90 | schoolid), data = mydata, REML = FALSE))
Linear mixed model fit by maximum likelihood
```

Formula: score ∼ cohort90 + female + sclass1 + sclass2 + sclass4 + (1 + cohort90 | schoolid)
Data: mydata

AIC	BIC	$logLik$	$deviance$	$REMLdev$
276712	276797	−138346	276692	276705

Random effects:

$Groups$	$Name$	$Variance$	$Std.\ Dev.$	$Corr$
$schoolid$	$(Intercept)$	22.51334	4.74482	
	$cohort90$	0.15084	0.38839	−0.317
$Residual$		192.94571	13.89049	

Number of obs: 33988, groups: schoolid, 508
Fixed effects:

	$Estimate$	$Std.\ Error$	$t\ value$
$(Intercept)$	24.60993	0.27962	88.01
$cohort90$	1.18284	0.02432	48.64
$female$	1.96135	0.15428	12.71
$sclass1TRUE$	11.08565	0.20639	53.71
$sclass2TRUE$	5.87517	0.20405	28.79
$sclass4TRUE$	−3.73775	0.28453	−13.14

Correlation of Fixed Effects:

	$(Intr)$	$chrt90$	$female$	$s1TRUE$	$s2TRUE$
$cohort90$	−0.150				
$female$	−0.296	−0.023			
$sclass1TRUE$	−0.395	−0.054	0.008		
$sclass2TRUE$	−0.386	−0.020	0.009	0.539	
$sclass4TRUE$	−0.271	−0.036	0.013	0.358	0.357

Ως γνωστόν, ο πίνακας Fixed Effects περιέχει τις εκτιμήτριες των συντελεστών των μεταβλητών μας, του σταθερού όρου και τα τυπικά τους σφάλματα.

Συναφείς επιδράσεις

Όπως προαναφέραμε θα εισάγουμε μία-μία κάθε κατηγορική μεταβλητή στο μοντέλο μας. Ξεκινώντας λοιπόν με εκείνη για τον τύπο του σχολείου, δηλαδή αν είναι ιδιωτικό ή δημόσιο, την $schtype$.

$$score_{ij} = \beta_0 + \beta_1 cohort90_{ij} + \beta_2 female_{ij} + \beta_3 sclass1_{ij} + \beta_4 sclass2_{ij} + \beta_5 sclass4_{ij} + \beta_6 schtype_j + u_{0j} + u_{1j} cohort90_{ij} + \varepsilon_{ij}$$

Οπότε με χρήση της εντολής $update$, αναβαθμίζουμε και προσαρμόζουμε το προηγούμενο μοντέλο και προκύπτουν τα ακόλουθα

> (fit2 <- update(fit1a, . ~ . + schtype))
Linear mixed model fit by maximum likelihood
Formula: score ~ cohort90 + female + sclass1 + sclass2 + sclass4 + (1 + cohort90 | schoolid) + schtype
Data: mydata

AIC	BIC	logLik	deviance	REMLdev
276689	276782	−138333	276667	276678

Random effects:

Groups	Name	Variance	Std. Dev.	Corr
schoolid	(Intercept)	20.57103	4.53553	
	cohort90	0.14812	0.38486	−0.263
Residual		192.99407	13.89223	

Number of obs: 33988, groups: schoolid, 508
Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	24.27939	0.27872	87.11
cohort90	1.18404	0.02421	48.90
female	1.96377	0.15426	12.73
sclass1TRUE	11.03061	0.20695	53.30
sclass2TRUE	5.85641	0.20414	28.69
sclass4TRUE	−3.75024	0.28454	−13.18
schtype	4.24675	0.81687	5.20

Correlation of Fixed Effects:

	(Intr)	chrt90	female	s1TRUE	s2TRUE	s4TRUE
cohort90	−0.112					
female	−0.297	−0.023				
sclass1TRUE	−0.378	−0.055	0.008			
sclass2TRUE	−0.379	−0.021	0.009	0.540		
sclass4TRUE	−0.270	−0.036	0.013	0.357	0.357	
schtype	−0.214	0.007	0.001	−0.079	−0.034	−0.005

Από τα παραπάνω δεδομένα κρατάμε ότι $\hat{\beta}_6 = 4.25$, που σημαίνει πως ένα παιδί από ιδιωτικό σχολείο αναμένεται να σημειώσει βαθμολογία 4.25 πόντους μεγαλύτερη από ένα παιδί που πηγαίνει σε δημόσιο (από το ίδιο έτος, το ίδιο φύλο και το ίδιο κοινωνικό περιβάλλον). Το τυπικό σφάλμα $se(\hat{\beta}_6) = 0.82$ είναι περίπου πέντε φορές μικρότερο της $\hat{\beta}_6$ και για το λόγο αυτό, η επίδραση αυτή είναι στατιστικά σημαντική. Παρατηρούμε μια μικρή μείωση στη διακύμανση μεταξύ των σχολείων. Αφότου λάβαμε υπόψιν μας τον τύπο του σχολείου, η

διακύμανση για κοορτή 1990, δηλαδή η σ^2_{u0} μειώθηκε από το 2.5 (αποτελέσματα προσαρμογής $fit1a$) σε 20.6. Προσθέτουμε τώρα την $schurban$ μεταβλητή που προσδιορίζει την τοποθεσία του σχολείου (αστική ή αγροτική) και το μοντέλο θα γίνει:

$$score_{ij} = \beta_0 + \beta_1 cohort90_{ij} + \beta_2 female_{ij} + \beta_3 sclass1_{ij} + \beta_4 sclass2_{ij} + \beta_5 sclass4_{ij} + \beta_6 schtype_j + \beta_7 schurban_j + u_{0j} + u_{1j} cohort90_{ij} + \varepsilon_{ij}$$

και προσαρμόζοντας το παίρνω τα εξής αποτελέσματα:

```
> (fit3 <- update(fit2, . ~ . + schurban))
Linear mixed model fit by maximum likelihood
Formula: score ~ cohort90 + female + sclass1 + sclass2 + sclass4 + (1 + cohort90
| schoolid) + schtype + schurban
Data: mydata
  AIC    BIC    logLik   deviance   REMLdev
276682 276783 -138329     276658     276669
Random effects:
 Groups     Name        Variance  Std. Dev.  Corr
 schoolid   (Intercept) 19.95405  4.46700
            cohort90     0.14825  0.38503   -0.263
 Residual               193.01104 13.89284
Number of obs: 33988, groups: schoolid, 508
Fixed effects:
             Estimate    Std. Error  t value
(Intercept) 25.25993v0.42720  59.13
cohort90     1.18191    0.02423     48.78
female       1.96668    0.15425     12.75
sclass1TRUE 11.03299    0.20694     53.32
sclass2TRUE  5.84714    0.20418     28.64
sclass4TRUE -3.73999    0.28457    -13.14
schtype      4.39193    0.80922      5.43
schurban    -1.43718    0.47632     -3.02
Correlation of Fixed Effects:
```

	(Intr)	chrt90	female	s1TRUE	s2TRUE	s4TRUE	schtyp
cohort90	−0.099						
female	−0.190	−0.023					
sclass1TRUE	−0.252	−0.055	0.008				
sclass2TRUE	−0.264	−0.020	0.009	0.540			
sclass4TRUE	−0.165	−0.037	0.013	0.357	0.356		
schtype	−0.096	0.005	0.002	−0.080	−0.036	−0.004	
schurban	−0.763	0.035	−0.004	0.007	0.022	−0.015	−0.054

Από τα δεδομένα μας κρατάμε ότι $\hat{\beta}_7 = -1.44$. Αυτό σημαίνει ότι κατά μέσο όρο, ένας μαθητής που πηγαίνει σε σχολείο που βρίσκεται σε αστική τοποθεσία λαμβάνει βαθμολογία 1.44 πόντους χαμηλότερη από ένα μαθητή σχολείου αγροτικής τοποθεσίας. Η διακύμανση μεταξύ των σχολείων το 1990 μειώθηκε περαιτέρω, αλλά κατά ένα πολύ μικρό ποσοστό από $\hat{\sigma}^2_{u0} = 20.6$ σε 20.0. Τέλος, αναβαθμίζουμε το προηγούμενο μοντέλο εισάγοντας τη μεταβλητή *schdenom* ώστε να ελέγξουμε τις διαφορές στην επίδοση του μαθητή λόγω των θρησκευτικών δογμάτων του σχολείου.

$$\begin{aligned} score_{ij} = \beta_0 + \beta_1 cohort90_{ij} + \beta_2 female_{ij} + \beta_3 sclass1_{ij} + \beta_4 sclass2_{ij} + \beta_5 sclass4_{ij} \\ + \beta_6 schtype_j + \beta_7 schurban_j + \beta_8 schdenom_j + u_{0j} + u_{1j} cohort90_{ij} + \varepsilon_{ij}. \end{aligned}$$

Προσαρμόζουμε το ανωτέρω μοντέλο στα δεδομένο που προκύπτουν τα παρακάτω δεδομένα:

```
> (fit4 <- update(fit3, . ~ . + schdenom))
Linear mixed model fit by maximum likelihood
Formula: score ~ cohort90 + female + sclass1 + sclass2 + sclass4 + (1 + cohort90
| schoolid) + schtype + schurban + schdenom
Data: mydata
   AIC    BIC    logLik   deviance   REMLdev
276684 276793 −138329     276658      276668
Random effects:
 Groups      Name        Variance  Std. Dev.  Corr
 schoolid    (Intercept) 19.96576  4.46831
             cohort90     0.14818  0.38494   −0.267
 Residual                193.01279 13.89290
Number of obs: 33988, groups: schoolid, 508
Fixed effects:
```

	Estimate	*Std. Error*	*t value*
$(Intercept)$	25.24843	0.42934	58.81
$cohort90$	1.18203	0.02422	48.80
$female$	1.96666	0.15425	12.75
$sclass1TRUE$	11.03348	0.20696	53.31
$sclass2TRUE$	5.84725	0.20419	28.64
$sclass4TRUE$	−3.74055	0.28458	−13.14
$schtype$	4.39730	0.81079	5.42
$schurban$	−1.46221	0.48490	−3.02
$schdenom$	0.17054	0.60149	0.28

Correlation of Fixed Effects:

	$(Intr)$	$chrt90$	$female$	$s1TRUE$	$s2TRUE$	$s4TRUE$	$schtyp$	$schrbn$
$cohort90$	−0.100							
$female$	−0.189	−0.023						
$sclass1TRUE$	−0.252	−0.055	0.008					
$sclass2TRUE$	−0.263	−0.020	0.009	0.540				
$sclass4TRUE$	−0.164	−0.037	0.013	0.357	0.356			
$schtype$	−0.102	0.005	0.001	−0.079	−0.035	−0.005		
$schurban$	−0.726	0.034	−0.004	0.004	0.020	−0.013	−0.065	
$schdenom$	−0.101	0.001	−0.002	0.015	0.008	−0.007	0.066	−0.189

Ο λόγος $\frac{\hat{\beta}_8}{se(\hat{\beta}_8)} = 0.28$, δηλαδή δεν σημειώνονται διαφορές μεταξύ καθολικών και μη σχολείων. Για το λόγο αυτό απομακρύνουμε τη μεταβλητή από το μοντέλο μας.

Αλληλεπιδράσεις διασταυρούμενων επιπέδων

Με τη μέχρι τώρα ανάλυση μας συμπεραίνουμε πως η επίδοση ενός δεκαεξάχρονου μαθητή συσχετίζεται με το έτος διεξαγωγής των εξετάσεων (κοορτή), το φύλο του μαθητή και την κοινωνική τάξη του γονέα. Στο επίπεδο του σχολείου υπάρχουν διαφορές στην επίδοση των μαθητών που προέρχονται είτε από ιδιωτικά ή δημόσια σχολεία, είτε από σχολεία αστικών ή αγροτικών περιοχών. Παρόλο αυτά, θεωρήσαμε μόνο τις βασικές επιδράσεις των μεταβλητών αυτών. Πρακτικά, η σχέση μεταξύ των y και x_1 μεταβλητών, μπορεί να εξαρτάται από την τιμή μιας άλλης x_2 μεταβλητής που ορίζεται στο ίδιο ή σε διαφορετικά επίπεδα. Όταν λοιπόν βρίσκονται σε διαφορετικά επίπεδα, η αλληλεπίδραση ονομάζεται αλληλεπίδραση διασταυρούμενων επιπέδων (cross-level interaction). Για να παρουσιάσουμε τις διασταυρούμενες επιδράσεις και έπειτα να προχωρήσουμε στην ερμηνεία, θα ελέγξουμε για αλληλεπίδραση μεταξύ της μεταβλητής για κοορτή που βρίσκεται στο επίπεδο 1 και αυτής για τον τύπο του σχολείου στο επίπεδο 2. Θα εξετάσουμε επίσης αν η αλληλεπίδραση

τύπου σχολείου-κοορτής μπορεί να εξηγήσει τις διαφορές μεταξύ των σχολείων στις τάσεις επίδοσης του μαθητή (δηλαδή αν μια τέτοιου είδους αλληλεπίδραση μειώνει τη διακύμανση του τυχαίου τμήματος της κλίσης της κοορτής). Αρχικά δημιουργούμε μια νέα μεταβλητή για την αλληλεπίδραση με την εντολή

```
> mydata$cohort90Xschtype <- mydata$cohort90*mydata$schtype
```

και έπειτα εισάγουμε την αλληλεπίδραση αυτή στο μοντέλο:

$$\begin{aligned} score_{ij} = {} & \beta_0 + \beta_1 cohort90_{ij} + \beta_2 female_{ij} + \beta_3 sclass1_{ij} + \beta_4 sclass2_{ij} + \beta_5 sclass4_{ij} \\ & + \beta_6 schtype_j + \beta_7 schurban_j + \beta_8 cohort90Xschtype_{ij} \\ & + u_{0j} + u_{1j} cohort90_{ij} + \varepsilon_{ij} \end{aligned}$$

και το προσαρμόζουμε με την εντολή

```
> (fit5 <- update(fit3, . ~ . + cohort90Xschtype))
Linear mixed model fit by maximum likelihood
Formula: score ~ cohort90 + female + sclass1 + sclass2 + sclass4 + (1 + cohort90
| schoolid) + schtype + schurban + cohort90Xschtype
Data: mydata
```

AIC	*BIC*	*logLik*	*deviance*	*REMLdev*
276651	276761	−138313	276625	276638

Random effects:

Groups	*Name*	*Variance*	*Std. Dev.*	*Corr*
schoolid	*(Intercept)*	20.41431	4.5182	
	cohort90	0.13801	0.3715	−0.233
Residual		192.85131	13.8871	

Number of obs: 33988, groups: schoolid, 508

Fixed effects:

	Estimate	*Std. Error*	*t value*
(Intercept)	25.18705	0.43201	58.30
cohort90	1.21354	0.02442	49.69
female	1.97026	0.15418	12.78
sclass1TRUE	11.01937	0.20687	53.27
sclass2TRUE	5.83072	0.20411	28.57
sclass4TRUE	−3.74284	0.28444	−13.16
schtype	5.29101	0.83071	6.37
schurban	−1.40365	0.48284	−2.91
cohort90Xschtype	−0.59937	0.10380	−5.77

Correlation of Fixed Effects:

	$(Intr)$	$chrt90$	$female$	$s1TRUE$	$s2TRUE$	$s4TRUE$	$schtyp$	$schrbn$
$cohort90$	-0.090							
$female$	-0.188	-0.022						
$sclass1TRUE$	-0.249	-0.056	0.008					
$sclass2TRUE$	-0.260	-0.023	0.009	0.540				
$sclass4TRUE$	-0.163	-0.037	0.013	0.357	0.356			
$schtype$	-0.100	0.044	0.002	-0.080	-0.037	-0.005		
$schurban$	-0.765	0.038	-0.004	0.007	0.021	-0.014	-0.050	
$chrt90Xscht$	0.027	-0.235	-0.005	0.009	0.013	0.002	-0.175	-0.014

Η εκτιμήτρια $\hat{\beta}_8 = -0.6$ είναι περίπου 6 φορές μεγαλύτερη από το τυπικό σφάλμα $se(\hat{\beta}_8) = 0.10$. Έτσι η επίδραση της μεταβλητής $cohort90$ διαφέρει για ιδιωτικά και δημόσια σχολεία. Ισοδύναμα, μπορούμε να πούμε ότι η διαφορά μεταξύ ιδιωτικών και δημοσίων μεταβάλλεται σε σχέση με τις μελέτης κοορτής. Η διακύμανση στο επίπεδο των σχολείων του συντελεστή της μεταβλητής $cohort90$ έχει μειωθεί ελαφρά από $0.148 - 0.138$. Για να αντιληφθούμε τη φύση της αλληλεπίδρασης, θεωρούμε το σταθερό μέρος του μοντέλου που περιέχει τις μεταβλητές $cohort90$ και $schtype$:

$$1.213 cohort90 + 5.219 schtype - 0.599 cohort90 \cdot schtype.$$

Για $schtype = 0$ (δημόσια σχολεία) η παραπάνω εξίσωση γίνεται

$$1.213 cohort90.$$

Έτσι στο "μέσο" σχολείο (π.χ. με $u_{1j} = 0$)[5] αναμένουμε μια αύξηση στην επίδοση της βαθμολογίας κατά 1.213 μονάδες με την πάροδο του χρόνου. Για $schtype = 1$ (ιδιωτικά σχολεία) η παραπάνω εξίσωση γίνεται

$$1.213 cohort90 + 5.219 - 0.599 cohort90 = 0.614 cohort90 + 5.219.$$

Έτσι για το "μέσο" ιδιωτικό σχολείο αναμένεται αύξηση της επίδοσης στη βαθμολογία 0.614 μονάδων. Η εκτιμήτρια του συντελεστή της $schtype$ (5.291) είναι η αναμενόμενη διαφορά στην επίδοση ανάμεσα σε ιδιωτικά και δημόσια το 1990 (δηλαδή όταν $cohort90 = 0$). Συμπερασματικά, η μέση επίδοση είναι υψηλότερη στα ιδιωτικά σχολεία από ότι στα δημόσια, παρ' όλα αυτά τα ιδιωτικά σχολεία σημειώνουν μικρότερη αύξηση στην επίδοση σε σχέση με την κοορτή.

[5] Η επίδραση της μεταβλητής $cohort90$ μεταβάλλεται τυχαία μεταξύ των σχολείων, έτσι θα σταθεροποιήσουμε το υπόλοιπο $u_{1j} = 0$ για να ελέγξουμε την αλληλεπίδραση σχολείου κοορτής

ΚΕΦΑΛΑΙΟ 4

Μοντέλα Παλινδρόμησης για δυαδικές μεταβλητές απόκρισης

4.1 Λογιστική Παλινδρόμηση

Το **μοντέλο της λογιστικής παλινδρόμησης** (logistic regression), αποτελεί ειδική περίπτωση των γενικευμένων γραμμικών μοντέλων, αλλά λόγω της σπουδαιότητάς του παρουσιάζεται ξεχωριστά. Αποτελεί τον πλέον καθιερωμένο τρόπο προσομοίωσης δυαδικών δεδομένων, όπου η εξαρτημένη μεταβλητή παίρνει δύο τιμές, οι οποίες αντιστοιχούν σε δύο ενδεχόμενα. Οι τιμές της μεταβλητής αποτελούν μια αυθαίρετη κωδικοποίηση των δύο ενδεχομένων, συνήθως 0 και 1. Επικεντρώνουμε την προσοχή μας σε ένα από τα δύο ενδεχόμενα, την επιτυχία $y = 1$ με πιθανότητα $p = P(\text{επιτυχία})$. Η y είναι τ.μ. της κατανομής Bernoulli, δηλαδή $y \sim B(p)$, με $E(y) = p$ και $V(y) = p(1-p)$. Επεκτείνοντας σε μια σειρά από n δοκιμές (δηλαδή, πραγματοποιήσεων των ενδεχομένων), ορίζουμε τ.μ. y=αριθμός επιτυχιών σε n δοκιμές. Υπό την υπόθεση ότι η πιθανότητα επιτυχίας p είναι ίδια σε κάθε δοκιμή και οι δοκιμές είναι ανεξάρτητες μεταξύ τους, τότε ισχύει η Διωνυμική (binomial) κατανομή

$$y \sim b(n, p)$$

με σ.π.

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y}, y = 0, 1, 2, \ldots, n$$

όπου η πιθανότητα επιτυχίας p είναι παράμετρος της κατανομής. Η Διωνυμική κατανομή αποτελεί τη βασική κατανομή για την περιγραφή και ανάλυση μιας μεταβλητής αυτής της φύσης. Η μέση τιμή της y είναι ίση με $E(y) = np$ και η διασπορά με $V(y) = np(1-p)$. Στην ειδική περίπτωση για $n = 1$, μιλάμε για **δεδομένα**, αλλιώς για **διωνυμικά δεδομένα**. Σε πολλές περιπτώσεις η τ.μ. y ενδέχεται να εξαρτάται από κάποιες επεξηγηματικές μεταβλητές. Η εξάρτηση της y από τις επεξηγηματικές μεταβλητές x (ανεξάρτητες μεταβλητές ή συμμεταβλητές) εισάγεται μέσω της πιθανότητας επιτυχίας p από τις x. Πιο συγκεκριμένα, κατασκευάζεται το αποκαλούμενο **μοντέλο της λογιστικής παλινδρόμησης**, το οποίο είναι ένα γενικευμένο γραμμικό μοντέλο και εκφράζεται μέσω της σχέσης

$$n_x = g(E(y_x)) = g(\mu_x) = \mathbf{x}'\beta$$

με την ακόλουθη δομή:

1. $y_x \sim b(n_x \mu_x)$ ($n_x > 1$, διωνυμικά δεδομένα) ή $y_x \sim B(\mu_x)$ ($n_x = 1$, δυαδικά δεδομένα)
2. $n_x = g(\mu_x) = \ln \frac{\mu_x}{n_x - \mu_x} = \ln \frac{p_x}{1-p_x} = logit(p_x) = \mathbf{x}'\beta$ συνάρτηση logit
3. ανεξαρτησία μεταξύ των παρατηρήσεων

όπου n_x ο αριθμός των επαναλήψεων της τιμής του διανύσματος x των επεξηγηματικών μεταβλητών. Αντιστρέφοντας τη συνάρτηση σύνδεσης προκύπτει

$$p_x = \frac{e^{n_x}}{1 + e^{n_x}}$$

από την οποία είναι φανερό ότι ισχύει ο απαραίτητος περιορισμός $0 < p_x < 1$. Για κάθε παρατήρηση i το μοντέλο γράφεται ως

$$\ln \frac{p_i}{1 - p_1} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}, i = 1, \ldots, n,$$

όπου

$$p_i = p_{x_i} = \frac{exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{1 + exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})} \tag{4.1}$$

η πιθανότητα επιτυχίας και συνεπώς

$$E(y_i) = n_i p_i = n_i \frac{e^{\mathbf{x}_i'\beta}}{1 + e^{\mathbf{x}_i'\beta}}$$

Η παραπάνω συνάρτηση $logit$ είναι η κανονική συνάρτηση σύνδεσης (link function) της Διωνυμικής κατανομής και αποτελεί τη συνηθέστερη επιλογή.

4.2 Εκτίμηση παραμέτρων και ερμηνεία

Η προσαρμογή του μοντέλου στα δεδομένα γίνεται με τη μέθοδο της μέγιστης πιθανοφάνειας όπως και με όλα τα γενικευμένα γραμμικά μοντέλα. Η συνάρτηση πιθανοφάνειας L ενός δείγματος τιμών $y_1, y_2, \ldots, y_n$ με μέσες τιμές $E(y_i) = \mu_i = n_i p_i$ και συμμεταβλητές $\mathbf{x}_i' = (x_{i0}, x_{i1}, \ldots, x_{ik})$, όπου n_i ο αριθμός των δοκιμών της στατιστικής μονάδας i, p_i η αντίστοιχη πιθανότητα επιτυχίας και $x_{i0} \equiv 1$ γράφεται ως

$$L(\beta) = \prod_i^n \binom{n_i}{y_i} p_i^{n_i} (1-p_i)^{n_i - y_i}.$$

Η πιθανοφάνεια εξαρτάται από τις άγνωστες πιθανότητες επιτυχίας p_i, οι οποίες με τη σειρά τους εξαρτώνται από τα β της σχέσης (4.1). Έτσι η συνάρτηση πιθανοφάνειας μπορεί να θεωρηθεί ως συνάρτηση των β με

$$\begin{aligned} \ell = \ln L(\beta) &= \sum_{i=1}^n \left\{ \ln \binom{n_i}{y_i} + y_i \ln p_i + (n_i - y_i) \ln (1-p_i) \right\} \\ &= \sum_{i=1}^n \left\{ \ln \binom{n_i}{y_i} + y_i \ln \frac{p_i}{1-p_i} + n_i \ln (1-p_i) \right\} \\ &= \sum_{i=1}^n \left\{ \ln \binom{n_i}{y_i} + y_i \mathbf{x}_i' \beta - n_i \ln (1 + e^{\mathbf{x}_i' \beta}) \right\} \end{aligned} \tag{4.2}$$

Παραγωγίζοντας ως προς τις συνιστώσες του βέχουμε

$$\begin{aligned} \frac{\partial \ln L(\beta)}{\partial \beta_j} &= \sum_{i=1}^n y_i x_{ij} - \sum_{i=1}^n n_i x_{ij} e^{\mathbf{x_i'}\beta} (1 + e^{\mathbf{x_i'}\beta})^{-1}, \qquad j = 0, 1, \ldots, k \\ &= \sum_{i=1}^n \left[y_i - n_i e^{\mathbf{x_i'}\beta} (1 + e^{\mathbf{x_i'}\beta})^{-1} \right] x_{ij} = \sum_{i=1}^n (y_i - n_i p_i) x_{ij}. \end{aligned}$$

Οι εκτιμήτριες της μέγιστης πιθανοφάνειας των β_j προκύπτουν με την ικανοποίηση των εξισώσεων

$$\sum_{i=1}^n (y_i - n_i \widehat{p}_i) x_{ij} = \sum_{i=1}^n (y_i - \widehat{\mu}_i) x_{ij} = 0, \quad j = 0, 1, \ldots, k.$$

4.2.1 Ερμηνεία των συντελεστών β

Ένα μεγάλο πλεονέκτημα της λογιστικής παλινδρόμησης έναντι των άλλων μοντέλων για διωνυμικά ή δυαδικά δεδομένα είναι η δυνατότητα ερμηνείας των τιμών των συντελεστών $\hat{\beta}$ Εφόσον εκτιμηθούν οι παράμετροι $\hat{\beta}$, η σχέση μεταξύ της προσαρμοσμένης πιθανότητας απόκρισης και των τιμών των $x_0, x_1, x_2, \ldots, x_k$ επεξηγηματικών μεταβλητών μπορεί να εκφραστεί ως

$$\hat{p} = \frac{e^{\mathbf{x}'\hat{\beta}}}{1 + e^{\mathbf{x}'\hat{\beta}}}$$

ή ισοδύναμα μέσω του **λόγου των συμπληρωματικών ή σχετικών πιθανοτήτων** (odds)

$$\frac{\hat{p}}{1-\hat{p}} = e^{\mathbf{x}'\hat{\beta}} = e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\cdots+\hat{\beta}_k x_k},$$

όπου $x_0 \equiv 1$. Από το odds προκύπτει ότι η ποσότητα $e^{\hat{\beta}_j}$ είναι ο παράγοντας επί τον οποίο πολλαπλασιάζεται η σχετική πιθανότητα πραγματοποίησης του γεγονότος «επιτυχία», όταν η αντίστοιχη ανεξάρτητη μεταβλητή x_j αυξηθεί κατά μία μονάδα, με δεδομένο πάντα ότι οι υπόλοιπες συμμεταβλητές παραμένουν σταθερές. Αν ο εκτιμημένος συντελεστής $\widehat{\beta}_j$ είναι θετικός, ο παράγοντας $e^{\widehat{\beta}_j}$ είναι μεγαλύτερος από τη μονάδα, γεγονός που σημαίνει πως το odds $\frac{\widehat{p}}{1-\widehat{p}}$ αυξάνεται με την αύξηση της x_j, αντίθετα αν το $\widehat{\beta}_j$ είναι αρνητικό, ο παράγοντας $e^{\widehat{\beta}_j}$ είναι μικρότερος της μονάδας και η σχετική πιθανότητα μειώνεται με την αύξηση της x_j.

Οι παράμετροι της παλινδρόμησης μπορούν να εκφραστούν και μέσα από το **λόγο του λόγου των συμπληρωματικών πιθανοτήτων**, δηλαδή μέσα από το **λόγο των odds** (odds ratio). Γενικώς ο λόγος των odds ενός ατόμου με τιμές συμμεταβλητών $\mathbf{x_1}$ σε σχέση με ένα άτομο με τιμές $\mathbf{x_2}$ των ίδιων συμμεταβλητών προκύπτει ως

$$\frac{\hat{p}_1}{1-\hat{p}_1} / \frac{\hat{p}_2}{1-\hat{p}_2} = \frac{odds(y=1 \mid x_1)}{odds(y=1 \mid x_2)} = \frac{exp(\mathbf{x_1}'\hat{\beta})}{exp(\mathbf{x_2}'\hat{\beta})} = exp(\mathbf{x_1} - \mathbf{x_2})'\beta.$$

Αν θεωρήσουμε ένα μοντέλο με δύο συμμεταβλητές x_1 και x_2, όπου η x_2 είναι μια ποσοτική ενώ η x_1 είναι μια δείκτρια μεταβλητή με $x_1 = 0, 1$, τότε

$$odds(y=1 \mid x_1=0, x_2) = exp(\hat{\beta}_0+\hat{\beta}_2 x_2) odds(y=1 \mid x_1=1, x_2) = exp(\hat{\beta}_0+\hat{\beta}_1+\hat{\beta}_2 x_2)$$

και ο λόγος των odds είναι $e^{\hat{\beta}_1}$, ανεξάρτητος της x_2. Αυτό το μοντέλο δείχνει ότι για κάθε τιμή της x_1 οι ευθείες $\ln(odds)$ ή $logit(\hat{p})$έχουν την ίδια κλίση, δηλαδή θα είναι παράλληλες.

4.3 Ελεγχοσυνάρτηση deviance

Μέσω της σχέσης $\ln\left(\frac{p_i}{1-p_i}\right) = \mathbf{x_i}'\beta$ το μοντέλο επιβάλλει μια δομή στα δεδομένα. Χωρίς καμία δομή έχουμε απλώς ανεξάρτητες τιμές από διαφορετικές Διωνυμικές κατανομές

$$y_i \sim b(n_i, p_i)$$

όπου $\mu_i = n_i p_i$ είναι χωρίς άλλο περιορισμό εκτός του ότι $\mu_i > 0$ και συνεπώς θα εκτιμάται από το $\tilde{\mu}_i = y_i$. Τότε η συμβολή της i-οστής παρατήρησης στη μεγιστοποιημένη τιμή του λογαρίθμου της πιθανοφάνειας υπό την υπόθεση H_S του κορεσμένου μοντέλου είναι

$$\tilde{\ell}_{iS} = \ln \binom{n_i}{y_i} + y_i \ln \tilde{p}_i + (n_i - y_i)\ln(1 - \tilde{p}_i).$$

Η αντίστοιχη συμβολή της i-οστής παρατήρησης στη μεγιστοποιημένη τιμή του λογαρίθμου της πιθανοφάνειας υπό την H_0 του υπό εξέταση μοντέλου είναι

$$\hat{\ell}_{i0} = \ln \binom{n_i}{y_i} + y_i \ln \hat{p}_i + (n_i - y_i)\ln(1 - \hat{p}_i).$$

Η διαφορά των δύο τιμών είναι

$$\begin{aligned}
\hat{\ell}_{i0} - \tilde{\ell}_{iS} &= y_i(\ln \hat{p}_i - \ln \tilde{p}_i) + (n_i - y_i)[\ln(1 - \hat{p}_i) - \ln(1 - \tilde{p}_i)] \\
&= y_i \ln\left(\frac{\hat{p}_i}{\tilde{p}_i}\right) + (n_i - y_i)\ln\left(\frac{1 - \hat{p}_i}{1 - \tilde{p}_i}\right) \\
&= y_i \ln\left(\frac{\hat{\mu}_i}{y_i}\right) + (n_i - y_i)\ln\left(\frac{n_i - \hat{\mu}_i}{n_i - y_i}\right),
\end{aligned}$$

όπου $\tilde{p}_i = \frac{y_i}{n_i}$ και $\hat{p}_i = \frac{\hat{\mu}_i}{n_i}$. Όπως και με το μοντέλο της παλινδρόμησης Poisson ορίζουμε την τυποποιημένη ελεγχοσυνάρτηση deviance ως

$$\begin{aligned}
D(\hat{\beta}) = D(\mathbf{y}; \hat{\mu}) = -2\left\{\hat{\ell}_0 - \hat{\ell}_S\right\} &= 2\sum_{i=1}^{n}\left\{y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (n_i - y_i)\ln\left(\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right)\right\} \\
&= 2\sum_{i=1}^{n} d_i(\hat{\beta}),
\end{aligned}$$

η οποία αποτελεί ένα μέτρο σύγκρισης μεταξύ των παρατηρήσεων y_i και των εκτιμηθέντων $\hat{\mu}_i$ και η οποία ταυτίζεται με τη συνάρτηση deviance, αφού για τη

Διωνυμική κατανομή ισχύει $a(\phi) = 1$. Υπενθυμίζουμε ότι η συνάρτηση deviance χρησιμοποιείται κυρίως για τη σύγκριση και ανάπτυξη μοντέλων.

Εδώ σημειώνουμε ότι **στην ειδική περίπτωση των δυαδικών δεδομένων**, δηλαδή όταν $n_i = 1, \forall i$, η **ελεγχοσυνάρτηση deviance δε μας παρέχει πληροφορίες για την καταλληλότητα ενός μοντέλου** και αυτό διότι εξαρτάται μόνο από τις προσαρμοσμένες ή εκτιμημένες τιμές $\hat{\mu}_i$ ως ακολούθως. Η συνάρτηση πιθανοφάνειας για n δυαδικές παρατηρήσεις είναι

$$L = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}$$

με $E(y_i) = p_i = \mu_i$ και επομένως η λογαριθμοποιημένη πιθανοφάνεια είναι

$$\ell = \ln L = \sum_{i=1}^{n} \{y_i \ln \mu_i + (1 - yi) \ln (1 - \mu_i)\}. \tag{4.3}$$

Εκτιμώντας το πλήρες ή κορεσμένο μοντέλο ισχύει $\tilde{\mu}_i = y_i$ και αφού οι όροι $y_i \ln y_i$ και $1 - y_i \ln (1 - y_i)$ είναι ίσοι με το μηδέν για τις δύο δυνατές τιμές του $y_i, 0$ και 1, θα ισχύει $\tilde{\ell}_S = 0$. Επομένως η ελεγχόμενη deviance για δυαδικές παρατηρήσεις ($n_i = 1$) δίνεται από τη σχέση

$$\begin{aligned} D(\hat{\beta}) &= -2 \sum_{i=1}^{n} \{y_i \ln \hat{\mu}_i + (1 - y_i) \ln (1 - \hat{\mu}_i)\} \\ &= -2 \sum_{i=1}^{n} \left\{ y_i \ln \left[\frac{\hat{\mu}_i}{(1 - \hat{\mu}_i)} \right] + \ln (1 - \hat{\mu}_i) \right\}. \end{aligned} \tag{4.4}$$

Για το υπό εξέταση μοντέλο, όπου $n_i = 1$, η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας των σχέσεων (4.2) και (4.3) γράφεται ως

$$\ell = \ln L(\beta) = \sum_{i=1}^{n} \left\{ y_i \mathbf{x'}_i \beta - \ln (1 + e^{\mathbf{x'}_i \beta}) \right\}$$

και παραγωγίζοντας ως προς τις παραμέτρους β_j έχουμε ότι

$$\frac{\partial \ln L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} (y_i - p_i) x_{ij} = \sum_{i=1}^{n} (y_i - \mu_i) x_{ij},$$

από την οποία συνεπάγεται ότι

$$\sum_{j=1}^{p} \beta_j \frac{\partial \ln L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} (y_i - \mu_i) \sum_{j=1}^{p} \beta_j x_{ij} = \sum_{i=1}^{n} (y_i - \mu_i) \ln \left[\frac{\mu_i}{(1 - \mu_i)} \right],$$

όπου $\mu_i = p_i = \frac{e^{\mathbf{x}_i'\beta}}{1+e^{\mathbf{x}_i'\beta}}$. Επειδή η $\hat{\beta}$ είναι η εκτιμήτρια μέγιστης πιθανοφάνειας της β, η παράγωγος στην αριστερή πλευρά της εξίσωσης μηδενίζεται στο $\hat{\beta}$. Επομένως οι προσαρμοσμένες πιθανότητες $\hat{\mu}_i = \hat{p}_i$ πρέπει να ικανοποιούν την εξίσωση

$$\sum_{i=1}^{n}(y_i - \hat{\mu}_i)logit(\hat{\mu}_i) = 0$$

και άρα

$$\sum_{i=1}^{n} y_i logit(\hat{\mu}_i) = \sum_{i=1}^{n} \hat{\mu}_i logit(\hat{\mu}_i).$$

Τέλος, αντικαθιστώντας την $\sum_{i=1}^{n} y_i logit(\hat{\mu}_i)$ στη σχέση (4.4) για την $D(\hat{\beta})$ λαμβάνουμε την τελική έκφραση για την ελεγχοσυνάρτηση deviance

$$D(\hat{\beta}) = -2\sum_{i=1}^{n}\left\{\hat{\mu}_i logit(\hat{\mu}_i) + \ln\left(1-\hat{\mu}_i\right)\right\},$$

η οποία εξαρτάται μόνο από τις προσαρμοσμένες τιμές $\hat{\mu}_i$ και όχι άμεσα από τις παρατηρήσεις y_i. Συνεπώς δεν μπορεί να κριθεί η καλή προσαρμογή του μοντέλου. Σε αυτήν την περίπτωση των δυαδικών παρατηρήσεων, όπου όλα τα $n_i = 1$, μπορεί να δειχθεί ότι η deviance δεν ακολουθεί την κατανομή $\mathcal{X}^2$ ούτε προσεγγιστικά. Ωστόσο **μεταβολές** της deviance εξακολουθούν να είναι της κατανομής $\mathcal{X}^2$.

4.4 Παράδειγμα

Τα δεδομένα που θα αναλύσουμε προέρχονται από την Δημογραφική μελέτη για την υγεία στο Μπανγκλαντές (Bungladesh Demographic and Health Survey), μια αντιπροσωπευτική, σε εθνικό επίπεδο έρευνα, για γυναίκες σε αναπαραγωγική ηλικία (16-49 ετών). Η εξαρτημένη μεταβλητή είναι ένας δυαδικός δείκτης για το κατά πόσο μια γυναίκα έλαβε προγεννητική φροντίδα από έναν ιατρικά εκπαιδευμένο πάροχο (γιατρό, νοσοκόμα, μαία) τουλάχιστον μια φορά πριν γεννήσει. Για να ελαχιστοποιήσουμε τα σφάλματα, επικεντρωθήκαμε μόνο στα παιδιά που γεννήθηκαν εντός 5 ετών από την έρευνα. Για το λόγο αυτό το δείγμα μας περιορίζεται σε γυναίκες που γέννησαν εντός μιας πενταετίας πριν την έρευνα. Σε περίπτωση που γέννησε περισσότερες από μια φορές κατά τη διάρκεια της προαναφερθείσας περιόδου

Όνομα μεταβλητής	Περιγραφή
comm	Αναγνωριστικό κοινότητας
womid	Αναγνωριστικό γυναίκας
antemed	Λήψη προγεννητικής φροντίδας τουλάχιστον μία φορά από ιατρικά εκπαιδευμένο πάροχο, π.χ. γιατρός, νοσοκόμα ή μαία (1 = ναι, 0 = όχι)
bord	Σειρα γέννησης του παιδιού (από 1 έως 13)
mage	Ηλικία μητέρας κατά τη γέννηση (σε χρόνια)
urban	Τύπος περιοχής κατοικίας (1 = αστική, 0 = αγροτική)
meduc	Επίπεδο μόρφωσης της μητέρας κατά την έρευνα (1 = κανένα, 2 = πρωτοβάθμια, 3 = δευτεροβάθμια και άνω)
islam	Θρήσκευμα μητέρας (1 = Ισλάμ, 0 = άλλο)
wealth	Οικογενειακό εισόδημα (από 1 = φτωχότεροι μέχρι 5 = πλουσιότεροι)

4.4.1 Εφαρμογή των Logit και Probit μοντέλων για την ανάλυση της πρόληψης προγεννητικής φροντίδας

Στην παράγραφο αυτή, θα εφαρμόσουμε τα *logit* και *probit* μοντέλα στα δεδομένα για την προγεννητική φροντίδα. Ξεκινάμε κατασκευάζοντας πίνακες συνάφειας, ώστε να αξιολογήσουμε τη σχέση της προγεννητικής φροντίδας (*antemed* μεταβλητή) ως προς τη μόρφωση της μητέρας (*meduc* μεταβλητή). Έτσι υπολογίζουμε το λόγο των συμπληρωματικών πιθανοτήτων (odds) και το λόγο των σχετικών συμπληρωματικών πιθανοτήτων (odds ratio). Έπειτα, θα προσαρμόσουμε τα δεδομένα μας με τη βοήθεια ενός *logit* μοντέλου (με τη μόρφωση της μητέρας (*meduc*) ως μοναδική επεξηγηματική μεταβλητή). Έτσι ακολουθώντας μια διαφορετική διαδικασία θα λάβουμε τελικά τις ίδιες τιμές όσον αφορά τους λόγους των odds. Εν συνεχεία, θα συγκρίνουμε τους συντελεστές των *probit* και *logit* μοντέλων, καταλήγοντας σε παρόμοια αποτελέσματα μετά την προσαρμογή τους.

Πιθανότητες, λόγοι πιθανοτήτων (odds) και λόγοι σχετικών πιθανοτήτων (odds ratios)

Με τη βοήθεια των παρακάτω εντολών κατασκευάζουμε πίνακες συνάφειας ή συχνοτήτων, οι γραμμές των οποίων αποτελούνται από τις κατηγορίες της μεταβλητής *antemed* και οι στήλες τους από τις κατηγορίες της *meduc*. Στο εσωτερικό τους παρατίθενται οι συχνότητες που αντιστοιχούν σε όλους τους

δυνατούς συνδιασμούς των κατηγοριών των δύο μεταβλητών.

```
> table(mydata$antemed, mydata$meduc)
      1    2    3
0  1272  856  485
1   594  793 1366
> table(mydata$antemed, mydata$meduc)/rbind(colSums(table(mydata$antemed,
mydata$meduc)), colSums(table(mydata$antemed, mydata$meduc)))
   1  2  30  0.6816720  0.5191025  0.26202051  0.3183280  0.4808975  0.7379795
> rowSums(table(mydata$antemed, mydata$meduc))/
sum(rowSums(table(mydata$antemed, mydata$meduc)))
        0         1
0.4869549 0.5130451
```

Σύμφωνα λοιπόν με τους παραπάνω πίνακες, η ολική πιθανότητα της λήψης προγεννητικής φροντίδας από ιατρικά εκπαιδευμένο πάροχο είναι 0.513, ενώ οι πιθανότητες λήψης προγεννητικής φροντίδας για μητέρες χωρίς σχολική εκπαίδευση, για μητέρες με πρωτοβάθμια ή δευτεροβάθμια εκπαίδευση και για μητέρες με τριτοβάθμια και άνω εκπαίδευση είναι 0.318, 0.481 και 0.738 αντίστοιχα.

Ο λόγος πιθανοτήτων (odds) της λήψης προγεννητικής φροντίδας για μητέρες χωρίς σχολική εκπαίδευση ισούται με το λόγο της τιμής για $antamed = 1$ προς την τιμή με $antamed = 0$ για $meduc = 1$.

```
> 594/1272
[1] 0.4669811
```

Με παρόμοια διαδικασία προκύπτουν οι λόγοι πιθανοτήτων της λήψης προγεννητικής φροντίδας για τις δύο εναπομείνασες κατηγορίες της μεταβλητής για τη μόρφωση της μητέρας. Οι τιμές των λόγων της λήψης προγεννητικής φροντίδας για μητέρες με πρωτοβάθμια ή δευτεροβάθμια εκπαίδευση και για μητέρες με τριτοβάθμια και άνω εκπαίδευση είναι 0.926 και 2.817 αντίστοιχα. Δεδομένου πως ένας λόγος ίσος με τη μονάδα υποδεικνύει (π.χ. πως η πιθανότητα λήψης προγεννητικής φροντίδας ισούται με την πιθανότητα μη λήψης προγεννητικής φροντίδας. Έτσι μπορούμε να δούμε που έχουν λάβει μόρφωση που αντιστοιχεί στο επίπεδο τουλάχιστον της δευτεροβάθμιας εκπαίδευσης είναι πιθανότερο να λάβουν προγεννητική φροντίδα από ότι να μη λάβουν.

Υπολογίζουμε τώρα τους λόγους των σχετικών πιθανοτήτων (odds ratios), συγκρίνοντας το odds της προγεννητικής φροντίδας για μία κατηγορία της με-

ταβλητής $meduc$ ως προς μια άλλη κατηγορία. Υπολογίζουμε λοιπόν το λόγο των odds της προγεννητικής φροντίδας για μητέρες με πρωτοβάθμια ή δευτεροβάθμια εκπαίδευση προς εκείνες χωρίς και ισούται με 1.984.

```
> (793/856)/(594/1272)
[1] 1.98381
```

Έτσι μια μητέρα με μόρφωση επιπέδου πρωτοβάθμιας ή δευτεροβάθμιας εκπαίδευση σχεδόν διπλασιάζει την πιθανότητα λήψης προγεννητικής φροντίδας. Με ανάλογες διαδικασίες προχωράμε στον υπολογισμό των υπολοίπων λόγων των odds με τα αποτελέσματα να συνοψίζονται στο στον ακόλουθο πίνακα.

Επίπεδο μόρφωσης	Πιθανότητα	Odds	Odds ratio
Κανένα	0.318	0.467	-
Πρωτοβάθμια	0.481	0.926	1.984
Δευτεροβάθμια και άνω	0.738	2.816	6.031

Πίνακας 4.1

4.4.2 Ερμηνεία ενός λογιστικού (logit) μοντέλου

Αρχικά, με τη βοήθεια της εντολής $glm()$ προσαρμόζουμε το $logit$ μοντέλο της σχέσης μεταξύ προγεννητικής φροντίδας και της μόρφωσης της μητέρας στην R. Η σύνταξη της εντολής δεν διαφέρει ιδιαίτερα αυτής για την $lm()$ που μας βοηθά στην προσαρμογή ενός απλού γραμμικού μοντέλου. Η βασική διαφορά είναι στην οικογένεια κατανομής που ορίζεται μέσω της $family$ επιλογής. Για ένα $logit$ μοντέλο, έχουμε τη διωνυμική οικογένεια κατανομής, δεδομένης της δυαδικής φύσης των δεδομένων. Μπορούμε να χρησιμοποιήσουμε διάφορες συναρτήσεις σύνδεσης, αλλά η προκαθορισμένη είναι η συνάρτηση $logit$. Με τη βοήθεια λοιπόν της εντολής

```
> fit <- glm(antemed ~ meduc2 + meduc3, data = mydata, family = binomial(logit))
```

προσαρμόζουμε το μοντέλο

$$\log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 meduc2_i + \beta_2 meduc3_i.$$

Καλώντας την $summary(fit)$ εντολή λαμβάνουμε τα εξής αποτελέσματα:

```
> summary(fit)
Call:
glm(formula = antemed ~ meduc2 + meduc3, family = binomial(logit), data =
mydata)
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.6367  -0.8755  0.7795  1.2100  1.5131
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.76147    0.04970 -15.323  < 2e-16 ***
meduc2       0.68502    0.06999   9.787  < 2e-16 ***
meduc3       1.79696    0.07255  24.768  < 2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 7435.2 on 5365 degrees of freedom
Residual deviance: 6747.6 on 5363 degrees of freedom
AIC: 6753.6
Number of Fisher Scoring iterations: 4
```

Το βασικό δέλεαρ των $logit$ μοντέλων είναι πως ύστερα από μια απλή προσαρμογή οι συντελεστές τους ερμηνεύονται ως λόγοι σχετικών πιθανοτήτων (odds ratios). Σημειώνουμε πως αναφερόμαστε σε λογαριθμοποιημένους λόγους σχετικών πιθανοτήτων (log-odds). Χρησιμοποιώντας την εντολή $exp(fit\$coefficients)$ απολογαριθμίζουμε τις εκτιμήτριες των συντελεστών του μοντέλου μετατρέποντας τες έτσι από log-odds σε λόγους odds.

```
> exp(fit$coefficients)
(Intercept)      meduc2      meduc3
  0.4669811   1.9838101   6.0312819
```

Οπότε μπορούμε να προχωρήσουμε στην ερμηνεία των συντελεστών. Ο $\hat{\beta}_1 = 0.68$ που τον μετατρέψαμε σε $\exp(\hat{\beta}_1) = 1.984$ μας εξηγεί πως στην περίπτωση μιας μητέρας με επίπεδο μόρφωσης πρωτοβάθμιας ή δευτεροβάθμιας εκπαίδευσης, σε σχέση με μία μητέρα χωρίς εκπαίδευση, αυξάνεται η πιθανότητα

λήψης προγεννητικής φροντίδας κατά 1.984. Όμοια ο $\hat{\beta}_2 = 1.800$ μετατρέπεται σε $\exp(\hat{\beta}_2) = 6.031$ που σημαίνει πως στην περίπτωση μιας μητέρας με μόρφωση επιπέδου τριτοβάθμιας ή ανώτερης εκπαίδευσης σε σχέση με εκείνη που δεν έχει λάβει τη στοιχειώδη μόρφωση επιπέδου πρωτοβάθμιας ή δευτεροβάθμιας εκπαίδευσης, αυξάνεται η πιθανότητα λήψης προγεννητικής φροντίδας κατά 6.031. Σημειώνουμε πως οι τιμές για τους λόγους των odds ταυτίζονται με εκείνες της προηγούμενης παραγράφου. Συμπερασματικά, όσο πιο μορφωμένη είναι μια μητέρα τόσο πιθανότερο είναι να λάβει προγεννητική φροντίδα.

4.4.3 Σύγκριση probit και logit συντελεστών

Θα προσαρμόσουμε το *probit* μοντέλο

$$F^{-1}(\pi_i) = \beta_0 + \beta_1 meduc2_i + \beta_2 meduc3_i$$

στα δεδομένα μας με τη βοήθεια της εντολής

```
> fit2 <- glm(antemed ~ meduc2 + meduc3,
data = mydata, family = binomial(probit))
```

Η σύνταξη της εντολής διαφέρει από εκείνη για το *logit* μοντέλο στο σημείο που ορίζουμε τη συνάρτηση σύνδεσης και μόνο. Στη θέση δηλαδή της *logit* θα βάλουμε την *probit*. Με χρήση της εντολής $summary(fit2)$ παίρνουμε τα εξής αποτελέσματα:

```
> summary(fit2)
 Call:
glm(formula = antemed ~ meduc2 + meduc3, family = binomial(probit), data
= mydata)
Deviance Residuals:
    Min      1Q  Median      3Q     Max
-1.6367 -0.8755  0.7795  1.2100  1.5131
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.47238    0.03022 -15.631  < 2e-16 ***
meduc2       0.42448    0.04321   9.825  < 2e-16 ***
meduc3       1.10951    0.04357  25.465  < 2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
```

Null deviance: 7435.2 on 5365 degrees of freedom
Residual deviance: 6747.6 on 5363 degrees of freedom
AIC: 6753.6
Number of Fisher Scoring iterations: 4

Οι εκτιμήτριες των μεταβλητών $meduc2$ και $meduc3$ έχουν την ίδια ερμηνεία με τις αντίστοιχες του μοντέλου $logit$. Παρ' όλα αυτά, οι $probit$ συντελεστές είναι κάπως μικρότεροι κατά απόλυτη τιμή σε σχέση με τους $logit$-συντελεστές κάθε μοντέλου, μαζί με τον λόγο τους συνοψίζονται στον ακόλουθο πίνακα:

Μεταβλητές	Logit	Probit	Probit:Logit
$\hat{\beta}_0$ $cons$	-0.761	-0.472	1.61
$\hat{\beta}_1$ $meduc2$	0.685	0.424	1.62
$\hat{\beta}_2$ $meduc3$	1.797	1.110	1.62

Πίνακας 4.2: *Εκτιμήτριες συντελεστών logit και probit για την πρόσληψη προγεννητικής φροντίδας*

4.4.4 Έλεγχοι στατιστικής σημαντικότητας και διαστήματα εμπιστοσύνης

Ας επανέλθουμε πάλι στο μοντέλο $logit$

$$\log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 meduc2_i + \beta_2 meduc3_i.$$

Αναφέρουμε ότι σε αυτή την παράγραφο όλα μπορούν να εφαρμοστούν εξίσου και στο $probit$ μοντέλο. Παραθέτουμε ξανά τα αποτελέσματα της προσαρμογής του μοντέλου της παραγράφου 4.4.2

> summary(fit)
Call: glm(formula = antemed ~ meduc2 + meduc3, family = binomial(logit), data = mydata)
Deviance Residuals:

Min	$1Q$	$Median$	$3Q$	Max
-1.6367	-0.8755	0.7795	1.2100	1.5131

Coefficients:

	$Estimate$	$Std.\ Error$	$z\ value$	$Pr(>\|z\|)$
$(Intercept)$	-0.76147	0.04970	-15.323	$< 2e-16 ***$
$meduc2$	0.68502	0.06999	9.787	$< 2e-16 ***$
$meduc3$	1.79696	0.07255	24.768	$< 2e-16 ***$

—

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 7435.2 on 5365 degrees of freedom
Residual deviance: 6747.6 on 5363 degrees of freedom
AIC: 6753.6
Number of Fisher Scoring iterations: 4

z-έλεγχοι

Για να ελέγξουμε τη μηδενική υπόθεση για το αν ένας συντελεστής ισούται με μηδέν χρησιμοποιούμε ένα z-έλεγχο. Σύμφωνα λοιπόν με τα παραπάνω η p-τιμή κάθε συντελεστή είναι σαφώς μικρότερη του 0.05 οπότε προφανώς είναι στατιστικά σημαντικοί σε 5% επίπεδο σημαντικότητας. Άρα απορρίπτω τη μηδενική υπόθεση οπότε οι $\hat{\beta}_0 \neq 0, \hat{\beta}_1 \neq 0$ και $\hat{\beta}_2 \neq 0$

Διαστήματα εμπιστοσύνης για τους λόγους των σχετικών πιθανοτήτων (odds ratios)

Κατασκευάζουμε 95% διάστημα εμπιστοσύνης για τις παραμέτρους του μοντέλου μέσω της εντολής $confint$.

```
> confint(fit)
Waiting for profiling to be done...
```

	2.5 %	97.5 %
$(Intercept)$	-0.8594640	-0.6646200
$meduc2$	0.5480822	0.8224776
$meduc3$	1.6554694	1.9398952

Για να πάρουμε διαστήματα εμπιστοσύνης για τους λόγους των odds απο-

λογαριθμίζουμε τα όρια του διαστήματος εμπιστοσύνης για κάθε παράμετρο.

```
> exp(fit$coefficients)
(Intercept)    meduc2    meduc3
  0.4669811  1.9838101  6.0312819
> exp(confint(fit))
Waiting for profiling to be done...
                2.5 %    97.5 %
(Intercept)  0.423389  0.514469
meduc2       1.729932  2.276132
meduc3       5.235537  6.958022
```

Έτσι, η μηδενική υπόθεση για $\beta = 0$ είναι ισοδύναμη με την αντίστοιχη μηδενική για $\exp(\beta) = 1$. Εφόσον κανένα από τα 95% διαστήματα εμπιστοσύνης για τους συντελεστές του μοντέλου δεν περιέχει την τιμή 1, απορρίπτουμε τη μηδενική υπόθεση σε κάθε περίπτωση. Ας σημειώσουμε ότι και τα δύο διαστήματα εμπιστοσύνης είναι μικρά, πράγμα που οφείλεται στα μικρά τυπικά σφάλματα.

Έλεγχοι Wald

Οι z-έλεγχοι χρησιμοποιούνται για τον έλεγχο υποθέσεων ξεχωριστά για κάθε παράμετρο ($H_0 : \beta = 0, H_1 : \beta \neq 0$). Πολλές φορές όμως θέλουμε να ελέγξουμε αν :

1. δύο συντελεστές είναι ταυτόχρονα μηδέν
2. δύο οι περισσότεροι συντελεστές είναι ίσοι μεταξύ τους.

Ξεκινάμε λοιπόν με μηδενική υπόθεση $H_0 : \beta_1 = \beta_2 = 0$ για να ελέγξουμε αν οι παράμετροι των $meduc2$ και $meduc3$ μεταβλητών είναι ταυτόχρονα μηδέν.

Αρχικά εγκαθιστούμε το στατιστικό πακέτο $VGAM$ και το καλούμε με την εντολή $library(VGAM)$. Έπειτα καλούμε τη βιβλιοθήκη car $(library(car))$ και με την εντολή

```
> linear.hypothesis(fit, c("meduc2","meduc3"),c(0,0))
```

προχωράμε στον έλεγχο της μηδενικής υπόθεσης.

```
Linear hypothesis test
```

Hypothesis:
meduc2 = 0
meduc3 = 0
Model 1: antemed ~ meduc2 + meduc3
Model 2: restricted model

	$Res.Df$	Df	$Chisq$	$Pr(>Chisq)$
1	5363			
2	5365	-2	620.16	$< 2.2e-16 **$

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Παρατηρούμε πως η $textp- < 2.2e-16$, οπότε απορρίπτουμε την H_0 συμπεραίνοντας πως ένας από τους συντελεστές, ή και δυο, είναι μη μηδενικοί.

Τέλος, με την εντολή

```
> linear.hypothesis(fit, "meduc2 = meduc3")
```

θα έχουμε τη μηδενική υπόθεση $H_0 : \beta_1 = \beta_2, H_1 : \beta_1 \neq \beta_2$ και ελέγχουμε αν η πρόληψη προγεννητικής φροντίδας διαφέρει μεταξύ γυναικών με διαφορετικό επίπεδο μόρφωσης.

Linear hypothesis test
Hypothesis:
meduc2 - meduc3 = 0
Model 1: antemed ~ meduc2 + meduc3
Model 2: restricted model

	$Res.Df$	Df	$Chisq$	$Pr(>Chisq)$
1	5363			
2	5364	-1	236.72	$< 2.2e-16 **$

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Επειδή η p-τιμή $< 2.2e-16$ απορρίπτουμε την H_0 και δεχόμαστε πως μεταξύ των γυναικών που έχουν λάβει μόρφωση επιπέδου πρωτοβάθμιας ή μόρφωση επιπέδου δευτεροβάθμιας εκπαίδευσης και αυτών με τριτοβάθμια και άνω εκπαίδευσης υπάρχουν διαφορές όσον αφορά την πιθανότητα πρόληψης προγεννητικής φροντίδας από ιατρικά εκπαιδευμένο πάροχο.

4.4.5 Προσθήκη επιπλέον επεξηγηματικών μεταβλητών στα μοντέλα για την ανάλυση της προγεννητικής φροντίδας.

Στις προηγούμενες παραγράφους προσαρμόσαμε μοντέλα *logit* και *probit* στα δεδομένα μας, έχοντας μόνο μια επεξηγηματική μεταβλητή. Τώρα, θα επεκτείνουμε το *logit* μοντέλο των προηγούμενων παραγράφων με την προσθήκη επιπλέον επεξηγηματικών μεταβλητών. Αν και εστιάζουμε την προσοχή μας στα *logit* μοντέλα, ακολουθώντας ανάλογες διαδικασίες θα καταλήξουμε σε ίδια αποτελέσματα (ερμηνεία, έλεγχοι υποθέσεων) και για τα *probit* μοντέλα.

Επεκτείνοντας το logit μοντέλο

Θεωρούμε τις επεξηγηματικές μεταβλητές $meduc$, $mage$, $urban$ και $wealth$. Η κατηγορική μεταβλητή $wealth$ εισάγεται στο μοντέλο ως μια σειρά από ψευδομεταβλητές με την κατηγορία 1 ως κατηγορία αναφοράς. Δημιουργούμε λοιπόν με τις 4 ακόλουθες εντολές, τις 4 ακόλουθες εντολές, τις 4 ψευδομεταβλητές που θα εισάγουμε στο μοντέλο μας.

```
> mydata$wealth2 <- mydata$wealth == 2
> mydata$wealth3 <- mydata$wealth == 3
> mydata$wealth4 <- mydata$wealth == 4
> mydata$wealth5 <- mydata$wealth == 5
```

Οπότε το μοντέλο παίρνει την μορφή

$$log\frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 meduc2_i + \beta_2 meduc3_i + \beta_3 magec_i + \beta_4 magecsq_i + \beta_5 urban_i + \beta_6 wealth2_i + \beta_7 wealth3_i + \beta_8 wealth4_i + \beta_9 wealth5_i.$$

Με τη βοήθεια της εντολής

```
>fit <- glm(antemed ~ meduc2 + meduc3 + magec + magecsq + urban + wealth2
+ wealth3 + wealth4 + wealth5, data = mydata, family = binomial(logit))
```

προσαρμόζουμε το μοντέλο καλώντας την $summary(fit)$ εντολή λαμβάνουμε τα εξής αποτελέσματα:

```
> summary(fit)
Call:
```

glm(formula = antemed ∼ meduc2 + meduc3 + magec + magecsq + urban + wealth2 + wealth3 + wealth4 + wealth5, family = binomial(logit), data = mydata)
Deviance Residuals:

Min	$1Q$	$Median$	$3Q$	Max
-2.1623	-0.9475	0.4508	0.9233	2.2149

Coefficients:

	$Estimate$	$Std.\ Error$	$z\ value$	$Pr(>\|z\|)$
$(Intercept)$	-1.4362418	0.0815007	-17.622	$< 2e-16 * **$
$meduc2$	0.4511490	0.0771446	5.848	$4.97e-09 * **$
$meduc3$	1.1755466	0.0872954	13.466	$< 2e-16 * **$
$magec$	-0.0036639	0.0060086	-0.610	0.5420
$magecsq$	-0.0012959	0.0006384	-2.030	$0.0424*$
$urban$	0.7945408	0.0742342	10.703	$< 2e-16 * **$
$wealth2TRUE$	0.4756747	0.0970574	4.901	$9.54e-07 * **$
$wealth3TRUE$	0.6938627	0.0979619	7.083	$1.41e-12 * **$
$wealth4TRUE$	1.0479207	0.1015565	10.319	$< 2e-16 * **$
$wealth5TRUE$	1.6998585	0.1163810	14.606	$< 2e-16 * **$

—
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 7435.2 on 5365 degrees of freedom
Residual deviance: 6151.2 on 5356 degrees of freedom
AIC: 6171.2
Number of Fisher Scoring iterations: 4

Παρατηρούμε πως οι συντελεστές των μεταβλητών της ηλικίας δεν είναι πλέον στατιστικά σημαντικοί (η p-τιμή $magec$ είναι 0.54 και η p-τιμή της $magecqs$ 0.04 δηλαδή μεγαλύτερη από 0.05). Για το λόγο αυτό εφαρμόζουμε ένα έλεγχο-Wald με $H_0 : \beta_3 = \beta_4 = 0$. Όπως και στην παράγραφο 4.4.4, (έχοντας καλέσει την car βιβλιοθήκη προχωράμε με την εντολή

> linear.hypothesis(fit, c(”magec”, ”magecsq”),c(0,0))

στον έλεγχο της μηδενικής υπόθεσης παίρνοντας τα εξής αποτελέσματα:

Linear hypothesis test
Hypothesis:
magec = 0
magecsq = 0

Model 1: antemed ∼ meduc2 + meduc3 + magec + magecsq + urban + wealth2 + wealth3 + wealth4 + wealth5
Model 2: restricted model

	$Res.Df$	Df	$Chisq$	$Pr(>Chisq)$
1	5356			
2	5358	−2	7.4155	0.02453∗

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

οπότε απορρίπτουμε τη μηδενική υπόθεση διατηρώντας και τις δύο μεταβλητές στο μοντέλο.

Απλοποιούμε τώρα το μοντέλο μας, αντικαθιστώντας τις ψευδομεταβλητές $wealth2$ έως $wealth5$ με την αρχική $wealth$ 5-κατηγορική μεταβλητή και το μοντέλο γίνεται

$$log\frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 meduc2_i + \beta_2 meduc3_i + \beta_3 magec_i + \beta_4 magecsq_i + \beta_5 urban_i + \beta_6 wealth_i.$$

Με τις ακόλουθες εντολές προσαρμόζουμε το μοντέλο και λαμβάνουμε τα ακόλουθα αποτελέσματα τα οποία και θα ερμηνεύσουμε στην παράγραφο που ακολουθεί.

> fit2 <- glm(antemed ∼ meduc2 + meduc3 + magec + magecsq + urban + wealth, data = mydata, family = binomial(logit))
> summary(fit2)
Call:
glm(formula = antemed ∼ meduc2 + meduc3 + magec + magecsq + urban + wealth, family = binomial(logit), data = mydata)
Deviance Residuals:

Min	$1Q$	$Median$	$3Q$	Max
−2.1155	−0.9215	0.4751	0.9397	2.2134

Coefficients:

	Estimate	*Std. Error*	*z value*	$Pr(>\|z\|)$
(Intercept)	−1.8316062	0.0834646	−21.945	$< 2e-16 ***$
meduc2	0.4464140	0.0770187	5.796	$6.78e-09 ***$
meduc3	1.1832540	0.0871429	13.578	$< 2e-16 ***$
magec	−0.0029384	0.0059994	−0.490	0.624
magecsq	−0.0013050	0.0006386	−2.044	0.041*
urban	0.8390504	0.0721183	11.634	$< 2e-16 ***$
wealth	0.3865604	0.0253496	15.249	$< 2e-16 ***$

—

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 7435.2 on 5365 degrees of freedom
Residual deviance: 6160.7 on 5359 degrees of freedom
AIC: 6174.7
Number of Fisher Scoring iterations: 4

Ερμηνεία συντελεστών μοντέλου

Θα απολογαριμθήσουμε ξανά τους παραπάνω συντελεστές με την εντολή

```
> cbind(exp(fit2$coefficients), exp(confint(fit2)))
```

λαμβάνοντας τα παρακάτω αποτελέσματα:

Waiting for profiling to be done...

	2.5 %	97.5 %
(Intercept)	0.1601561	0.1358276	0.1884095
meduc2	1.5626983	1.3438444	1.8175434
meduc3	3.2649811	2.7534006	3.8747719
magec	0.9970659	0.9854143	1.0088671
magecsq	0.9986958	0.9974360	0.9999372
urban	2.3141684	2.0097402	2.6664684
wealth	1.4719094	1.4007693	1.5471375

Οι συντελεστές των $meduc2$ και $meduc3$ ερμηνεύονται με τον ίδιο τρόπο όπως και στην 4.1.1. Παρατηρούμε πως τα odds ratio της $meduc3$ μειώθηκε, λόγω της προσθήκης επιπλέον μεταβλητών στο μοντέλο.

Επειδή η μεταβλητή age παρουσιάζεται μέσω ενός γραμμικού και τετραγω-

νικού όρου, η ερμηνεία των συντελεστών ως λόγοι σχετικών πιθανοτήτων δεν έχει και τόσο νόημα. Για το λόγο αυτό θα ερμηνεύσουμε τις επιδράσεις της ηλικίας με την βοήθεια πιθανοτήτων πρόβλεψης. Πρώτα όμως πρέπει να ερμηνεύσουμε τους λόγους σχετικών πιθανοτήτων για τις $urban$ και $wealth$ μεταβλητές. Η πιθανότητα λήψης προγεννητικής φροντίδας είναι 2.3 φορές υψηλότερη για μια γυναίκα που κατοικεί σε αστική περιοχή σε σχέση με εκείνη σε μια αγροτική περιοχή (οι συντελεστές των μεταβλητών για τη μόρφωση, την ηλικία και το εισόδημα παραμένουν σταθεροί). Τέλος, $\hat{\beta}_6 = 1.5$ που σημαίνει η αύξηση του εισοδήματος προϋποθέτει την αύξηση της λήψης προγεννητικής φροντίδας κατά ένα παράγοντα 1.5.

Θα υπολογίσουμε την πιθανότητα πρόβλεψης λήψης προγεννητικής φροντίδας για ηλικίες μεταξύ 15 και 45 ετών κρατώντας τις υπόλοιπες μεταβλητές σταθερές. Έπειτα, θα κατασκευάσουμε το διάγραμμα των πιθανοτήτων πρόβλεψης ως προς την ηλικία της μητέρας κατά τον το τοκετό. Αρχικά αντιγράφουμε τα δεδομένα σε ένα πλαίσιο δεδομένων με ονομασία $mydata2$. Έπειτα αντικαθιστούμε τις μεταβλητές $meduc2$, $meduc3$, $urban$ και $wealth$ με τις μέσες τιμές τους. Οπότε έχουμε

```
> mydata2 <- mydata
> mean(mydata$meduc2)
[1]7 0.3073053
> mydata2$meduc2 <- mean(mydata$meduc2)
> mean(mydata$meduc3)
 [1] 0.3449497
> mydata2$meduc3 <- mean(mydata$meduc3)
> mean(mydata$urban)
 [1]0.3138278
> mydata2$urban <- mean(mydata$urban)
> mean(mydata$wealth)
 [1] 3.0082
> mydata2$wealth <- mean(mydata$wealth)
```

Θα υπολογίσουμε τις πιθανότητες πρόβλεψης με τη βοήθεια της εντολής $predict$ και έχουμε

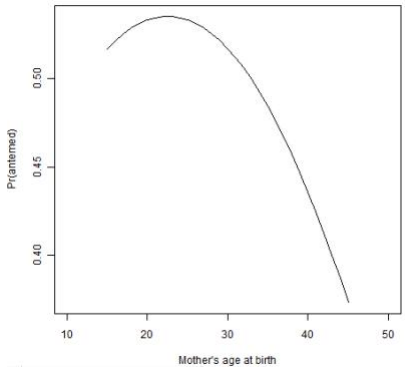
```
> mydata2$predprob <- predict(fit2, mydata2, type = "response")
```

Τελικά, παράγουμε το ζητούμενο γράφημα των πιθανοτήτων πρόβλεψης ως

προς την ηλικία της μητέρας.

```
> mydata2 <- mydata2[order(mydata2$mage), ]
> mydata2 <- mydata2[mydata2$mage %in% c(15:45), ]
> plot(mydata2$mage, mydata2$predprob, xlim = c(10, 50), xlab = "Mother's age at birth", ylab = "Pr(antemed)", type = "l")
```

Μπορούμε να δούμε πως οι γυναίκες σε ηλικία 25 περίπου ετών έχουν πιθανότητα άνω του 50% στο να λάβουν προγεννητική φροντίδα.



Σχήμα 4.1

ΚΕΦΑΛΑΙΟ 5

Πολυεπίπεδα Μοντέλα για δυαδικές μεταβλητές απόκρισης

5.1 Διεπίπεδα Random Intercept μοντέλα για δυαδικές μεταβλητές απόκρισης

Στο προηγούμενο κεφάλαιο είδαμε πως τα γενικά γραμμικά μοντέλα για συνεχείς μεταβλητές απόκρισης, μπορούν να γενικευτούν με τέτοιο τρόπο ώστε να διαχειριστούν δυαδικές μεταβλητές απόκρισης ορισμένες στο κατώτερο επίπεδο. Ωστόσο, υπάρχουν μοντέλα εμφωλευμένων δυαδικών δεδομένων, όπου οι επεξηγηματικές μεταβλητές ορίζονται στο επίπεδο 2. Ένας τρόπος ομαδοποίησης τέτοιων δεδομένων είναι η προσαρμογή ενός πολυεπίπεδου μοντέλου με τυχαίες επιδράσεις στο επίπεδο των ομάδων. Θα επιδιώξουμε την προσέγγιση αυτή εδώ, αλλά στόχος μας είναι να δείξουμε πώς τα πολυεπίπεδα μοντέλα μπορούν να εφαρμοστούν πιο γενικά σε διεπίπεδα δυαδικά δεδομένα με επεξηγηματικές μεταβλητές ορισμένες στα επίπεδο 1 και 2. Ένα παράδειγμα που θα μπορούσαμε να ερμηνεύσουμε μέσω των πολυεπίπεδων μοντέλων για δυαδικές μεταβλητές απόκρισης είναι:

- Ποίο είναι το εύρος της διακύμανσης μεταξύ των πολιτειών (ομάδες) στην προτίμηση ψήφου των Αμερικανών πολιτών (Ρεπουμπλικάνοι εναντίον Δη-

μοκρατικών);

- Μπορούν οι διαφορές μεταξύ των πολιτειών στην προτίμηση ψήφου να επεξηγηθούν μέσω των διαφορών στην εθνική ή θρησκευτική σύνθεση κάθε πολιτείας;

- Οι μεταβλητές στο επίπεδο 1, όπως το φύλο και η ηλικία έχουν διαφορετικές επιδράσεις σε διαφορετικές πολιτείες;

Ο πληθυσμός της παραπάνω μελέτης έχει διεπίπεδη δομή με τους ψηφοφόρους εμφωλευμένους στο επίπεδο 1 και τις πολιτείες στο δεύτερο επίπεδο. Στο κεφάλαιο αυτό, θα ενώσουμε τα πολυεπίπεδα μοντέλα για συνεχείς παρατηρήσεις και τα μοντέλα δυαδικών μεταβλητών απόκρισης. Θα δούμε επεκτάσεις των μοντέλων τυχαίων κλίσεων που εφαρμόζονται εξίσου σε δυαδικές μεταβλητές απόκρισης. Υπάρχουν εντούτοις, κάποια σημαντικά νέα ζητήματα που πρέπει να λάβουμε υπόψιν μας πριν προχωρήσουμε στην ερμηνεία των πολυεπίπεδων μοντέλων δυαδικών μεταβλητών απόκρισης.

5.1.1 Γενικευμένο Γραμμικό Μοντέλο Τυχαίων Σταθερών

Θεωρούμε μία δομή δύο επιπέδων όπου ένα σύνολο n ατόμων (επίπεδο 1) εμφωλεύεται σε J ομάδες (επίπεδο 2) με n_j άτομα στην ομάδα j. Στο κεφάλαιο αυτό, χρησιμοποιούμε τον όρο "ομάδα" σαν ένα γενικό όρο για κάθε μονάδα επιπέδου 2, π.χ. μια περιοχή ή ένα σχολείο. Συμβολίζουμε με y_{ij} τη μεταβλητή απόκρισης για το άτομο i στην ομάδα j, και με x_{ij} μια επεξηγηματική μεταβλητή στο επίπεδο 1. Ας θυμηθούμε την εξίσωση (3.3) της παραγράφου 3.2, το μοντέλο τυχαίων σταθερών για συνεχή μεταβλητή απόκρισης y

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + \varepsilon_{ij} \tag{5.1}$$

όπου u_j οι επιδράσεις των ομάδων (group effects) και ε_{ij} τα υπόλοιπα στο επίπεδο 1 που είναι ανεξάρτητα και ακολουθούν κανονική κατανομή με μηδενική μέση τιμή.

$$u_j \sim \mathcal{N}(0, \sigma_u^2) \quad \text{και} \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$$

Μπορούμε ακόμα να εκφράσουμε το μοντέλο με όρους μέσης τιμής ή αναμενόμενης τιμής y_{ij} για ένα άτομο στην ομάδα j και με τιμή x_{ij} στο x:

$$E(y_{ij}|x_{ij}, u_j) = \beta_0 + \beta_1 x_{ij} + u_j \tag{5.2}$$

Για μια δυαδική μεταβλητή απόκρισης y_{ij} , έχουμε $E(y_{ij}|x_{ij}, u_j) = \pi_{ij} = P(y_{ij} = 1)$ και ένα γενικευμένο γραμμικό μοντέλο τυχαίων σταθερών για την εξάρτηση της πιθανότητας απόκρισης π_{ij} στη x_{ij} γράφεται:

$$F^{-1}(\pi_{ij}) = \beta_0 + \beta_1 x_{ij} + u_j \tag{5.3}$$

όπου F^{-1} είναι η συνάρτηση σύνδεσης ($logit$ ή $probit$). Εδώ θα εστιάσουμε στη $logit$ συνάρτηση, αλλά ότι και να αναφέρουμε για τη $logit$ εφαρμόζεται εξίσου και στην $probit$.

5.1.2 Λογιστικό (logit) μοντέλο τυχαίων σταθερών

Σε ένα $logit$ μοντέλο $F^{-1}(\pi_{ij})$, είναι ο λογαριθμοποιημένος λόγος πιθανοτήτων (log-odds) για $y = 1$, έτσι η (5.3) γίνεται

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_1 x_{ij} + u_j \tag{5.4}$$

όπου $u_j \sim \mathcal{N}(0, \sigma_u^2)$.

Ερμηνεία του β_0 και β_1

Ο συντελεστής β_0 ερμηνεύεται ως το log-odds για $y = 1$ όταν $x = 0$ και $u = 0$ και αναφέρεται ως ολική σταθερά (overall intercept) της γραμμικής σχέσης μεταξύ του log-odds και του x. Εάν πάρουμε τον νεπέριο λογάριθμο του β_0, $\exp(\beta_0)$, εξασφαλίζουμε το λόγο πιθανοτήτων (odds) για $y = 1$ όταν $x = 0$ και $u = 0$.

Όπως και σε ένα $logit$ μοντέλο παλινδρόμησης (με ένα μόνο επίπεδο), το $\exp(\beta_1)$ μπορεί να ερμηνευθεί ως ένα odds ratio (για $y = 1$), συγκρίνοντας το λόγο πιθανοτήτων για δύο άτομα (στην ίδια ομάδα).

Ερμηνεία του u_j

Ενώ ο συντελεστής β_0 είναι η συνολική σταθερά (overall intercept) της γραμμικής σχέσης του log-odds και του x, η σταθερά (intercept) για μια δεδομένη ομάδα j είναι $\beta_0 + u_j$ η οποία θα είναι υψηλότερη ή χαμηλότερη της συνολικής σταθεράς και βασίζεται στο αν το u_j είναι μεγαλύτερο ή μικρότερο του μηδενός. Όπως και στην περίπτωση της συνεχούς μεταβλητής απόκρισης, καλούμε το συντελεστή u_j ως (τυχαία) επίδραση ομάδας, υπόλοιπο ομάδας (group residual) ή υπόλοιπο επίπεδου 2. Η διακύμανση των σταθερών όρων των ομάδων είναι

$V(u_j) = \sigma_u^2$, η οποία καλείται και διακύμανση των υπολοίπων μεταξύ των ομάδων ή απλά διακύμανση υπολοίπων στο επίπεδο 2.

Κατά την ανάλυση των ιεραρχικών δεδομένων, ενδιαφερόμαστε συχνά για το ποσό της διακύμανσης που αντιστοιχεί στα διαφορετικά επίπεδα της δομής δεδομένων και στο εύρος το οποίο η διακύμανση σε ένα δεδομένο επίπεδο μπορεί να εξηγηθεί από επεξηγηματικές μεταβλητές.

5.2 Παράδειγμα: Καθορισμός και Εκτίμηση ενός διεπίπεδου μοντέλου

Θα αναλύσουμε πάλι τα δεδομένα που χρησιμοποιήσαμε στο παράδειγμα του κεφαλαίου 4, απλώς εδώ θεωρούμε πολυεπίπεδα μοντέλα ώστε να ερευνήσουμε την μεταξύ των κοινοτήτων (between-community) διακύμανση για την προγεννητική φροντίδα. Τα δεδομένα έχουν μια διεπίπεδη ιεραρχική δομή με 5366 γυναίκες στο επίπεδο 1 εμφωλευμένες σε 361 κοινότητες στο επίπεδο 2. Σε αγροτικές περιοχές μια κοινότητα αντιστοιχεί σε ένα χωριό ενώ μια αστική κοινότητα αντιστοιχεί σε μία γειτονιά βάσει των ορισμών απογραφής. Θεωρούμε μια σειρά επεξηγηματικών μεταβλητών στο επίπεδο 1 όπως η ηλικία τη στιγμή του τοκετού και η μόρφωση. Οι μεταβλητές του επιπέδου 2 περιέχουν ένα δείκτη για τον αν η περιοχή κατοικίας μπορεί να χαρακτηριστεί ως αγροτική ή αστική.

Όνομα μεταβλητής	Περιγραφή
comm	Αναγνωριστικό κοινότητας
womid	Αναγνωριστικό γυναίκας
antemed	Λήψη προγεννητικής φροντίδας τουλάχιστον μία φορά από ιατρικά εκπαιδευμένο πάροχο, π.χ. γιατρός, νοσοκόμα ή μαία (1 = ναι, 0 = όχι)
bord	Σειρα γέννησης του παιδιού (από 1 έως 13)
mage	Ηλικία μητέρας κατά τη γέννηση (σε χρόνια)
urban	Τύπος περιοχής κατοικίας (1 = αστική, 0 = αγροτική)
meduc	Επίπεδο μόρφωσης της μητέρας κατά την έρευνα (1 = κανένα, 2 = πρωτοβάθμια, 3 = δευτεροβάθμια και άνω)
islam	Θρήσκευμα μητέρας (1 = Ισλάμ, 0 = άλλο)
wealth	Οικογενειακό εισόδημα (από 1 = φτωχότεροι μέχρι 5 = πλουσιότεροι)

5.2.1 Προσδιορισμός και ερμηνεία ενός διεπίπεδου μοντέλου

Ξεκινάμε προσαρμόζοντας ένα "μηδενικό" (null) διεπίπεδο μοντέλο που περιέχει ένα μόνο σταθερό όρο (intercept) β_0 τις επιδράσεις των κοινοτήτων (community-effects) u_{0j}

$$\log \frac{\pi_{ij}}{1-\pi_{ij}} = \beta_0 + u_{0j}$$

Ο συντελεστής β_0 παραμένει σταθερός ως προς όλες τις κοινότητες ενώ η τυχαία επίδραση u_{0j} είναι ορισμένη στην κοινότητα j. Η τυχαία επίδραση u_{0j} ακολουθεί κανονική κατανομή με διασπορά σ^2_{u0}. Η βασική εντολή στη R για την προσαρμογή πολυεπίπεδων μοντέλων σε δυαδικά δεδομένα είναι η $glmer$ της βιβλιοθήκης $lme4$. Αφού εγκαταστήσουμε στον υπολογιστή μας το πακέτο $lme4$ καλούμε τη βιβλιοθήκη μέσω της εντολής $library(lme4)$. Η σύνταξη της $glmer$ είναι όμοια με εκείνη για την $lmer$ που χρησιμοποιήσαμε σε προηγούμενο κεφάλαιο. Χρησιμοποιώντας λοιπόν την ακόλουθη εντολή, προσαρμόζουμε το μοντέλο στα δεδομένα μας.

```
> fit <- glmer(antemed ~ (1 | comm), family = binomial("logit"), data = mydata)
```

Στη συνέχεια με χρήση της summary εντολής λαμβάνουμε τα εξής αποτελέσματα

```
> summary(fit)
Generalized linear mixed model fit by the Laplace approximation
Formula: antemed ~ (1 | comm)
Data: mydata
  AIC   BIC  logLik  deviance
 6640  6653  -3318      6636
Random effects:
 Groups    Name        Variance  Std. Dev.
 comm.     (Intercept)  1.4644    1.2101
Number of obs: 5366, groups: comm, 361
Fixed effects:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  0.14811    0.07136     2.075    0.0379*
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5.2.2 Ερμηνεία διεπίπεδου "μηδενικού" (null) μοντέλου

Παρατηρούμε ότι η εκτιμήτρια του β_0 είναι $\hat{\beta}_0 = 0.148$. Μπορούμε λοιπόν να πούμε πως ο λογαριθμοποιημένος λόγος της σχετικής πιθανότητας (log-odds) της πρόσληψης προγεννητικής φροντίδας από ιατρικά εκπαιδευμένο πάροχο σε μία 'μέση' κοινότητα (για $u_{0j} = 0$) είναι 0.148. Η διακύμανση της u_{0j} τυχαίας επίδρασης είναι $\hat{\sigma}^2_{u0} = 1.464$. Προχωράμε στη διεξαγωγή ενός ελέγχου λόγου πιθανοφάνειας (likelihood ratio test) για να εξετάσουμε αν η διακύμανση $\hat{\sigma}^2_{u0} = 0$, δηλαδή

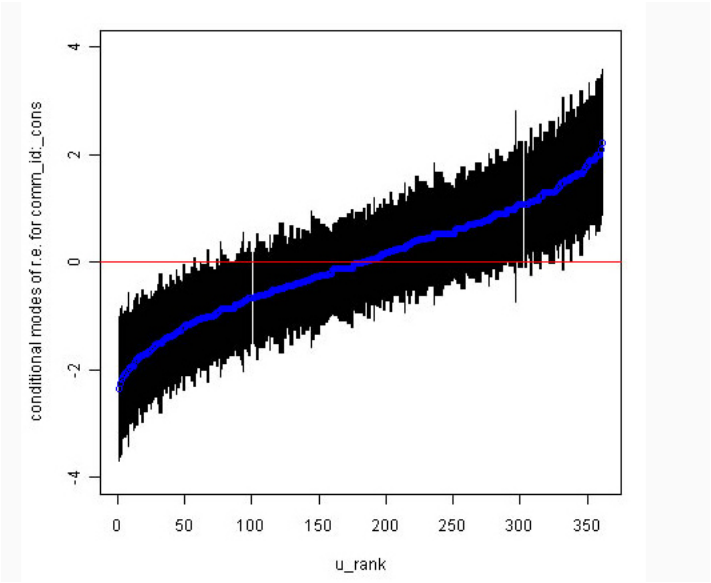
$$H_0 : \sigma^2_{u0} = 0$$
$$H_1 : \sigma^2_{u0} \neq 0$$

Έτσι συγκρίνουμε το διεπίπεδο μοντέλο μας με ένα γενικό γραμμικό χωρίς τυχαίες επιδράσεις στο επίπεδο 2.

```
> fita <- glm(antemed ~ 1, data = mydata, family = binomial("logit"))
> logLik(fita)-logLik(fit)
'log Lik.' -399.8392 (df=1)
```

Ο στατιστικός έλεγχος μας δίνει τιμή 799.8 $(= -2 \times (-399.83))$ με ένα βαθμό ελευθερίας. Για το λόγο αυτό απορρίπτουμε τη μηδενική υπόθεση και καταλήγουμε ότι $\sigma^2_{u0} \neq 0$. Όπως σε προηγούμενο κεφάλαιο, έτσι και εδώ, θα εξετάσουμε τις εκτιμήτριες των υπολοίπων των επιδράσεων των κοινοτήτων, $\widehat{u}_{0j}$, που μας έδωσε το "μηδενικό" μοντέλο. Θα δημιουργήσουμε λοιπόν, ένα caterpillar-γράφημα των επιδράσεων των κοινοτήτων σε 95% διάστημα εμπιστοσύνης. Για την κατασκευή του, θα χρησιμοποιήσουμε τις ίδιες ακριβώς εντολές με εκείνες για το διεπίπεδο μοντέλο τυχαίων σταθερών προηγούμενου κεφαλαίου.

```
> u0 <- ranef(fit, postVar = TRUE)
> u0se <- sqrt(attr(u0[[1]], "postVar")[1, , ])
> commid <- as.numeric(rownames(u0[[1]]))
> u0tab <- cbind("commid" = commid, "u0" = u0[[1]], "u0se" = u0se)
> colnames(u0tab)[2] <- "u0"
> u0tab <- u0tab[order(u0tab$u0), ]
> u0tab <- cbind(u0tab, c(1:dim(u0tab)[1]))
> u0tab <- u0tab[order(u0tab$commid), ]
> colnames(u0tab)[4] <- "u0rank"
> plot(u0tab$u0rank, u0tab$u0, type = "n", xlab = "u_ rank", ylab = "conditional
modes of r.e. for comm_ id:_ cons", ylim = c(-4, 4))
```



Σχήμα 5.1

```
> segments(u0tab$u0rank, u0tab$u0 - 1.96*u0tab$u0se, u0tab$u0rank, u0tab$u0
+ 1.96*u0tab$u0se)
> points(u0tab$u0rank, u0tab$u0, col = "blue")
> abline(h = 0, col = "red")
```

Το γράφημα δείχνει τις εκτιμήσεις των υπολοίπων για τις 361 κοινότητες του δείγματος. Για ένα σημαντικό αριθμό κοινοτήτων, το 95% διάστημα εμπιστοσύνης δεν επικαλύπτει την οριζόντια γραμμή στο 0, φανερώνοντας έτσι πως η πρόσληψη της προγεννητικής φροντίδας στις κοινότητες αυτές είναι μεγαλύτερη του μέσου όρου (πάνω από τη γραμμή 0) ή χαμηλότερη του μέσου όρου (κάτω από τη γραμμή 0).

Προσθήκη επεξηγηματικής μεταβλητής

Επεκτείνουμε το μοντέλο της προηγούμενης παραγράφου εισάγοντας μια επεξηγηματική μεταβλητή για την ηλικία της μητέρας ($magec$[1])

$$\log \frac{\pi_{ij}}{1-\pi_{ij}} = \beta_0 + \beta_1 magec_{ij} + u_{0j}.$$

[1]Αν και η συσχέτιση του log-odds της προγεννητικής φροντίδας με την ηλικία είναι καμπυλογραμμική (curvilinear) προσαρμόζουμε μια γραμμική επίδραση της ηλικίας

Προσαρμόζοντας το παραπάνω μοντέλο έχουμε τα ακόλουθα αποτελέσματα:

```
> (fit2 <- glmer(antemed ~ magec + (1 | comm), family = binomial("logit"),
data = mydata))
Generalized linear mixed model fit by the Laplace approximation
Formula: antemed ~ magec + (1 | comm)
Data: mydata
  AIC  BIC logLik deviance
 6603 6623  -3299     6597
Random effects:
 Groups Name        Variance Std.Dev.
 comm.  (Intercept) 1.4622   1.2092
Number of obs: 5366, groups: comm, 361

                    Estimate  Std. Error z value Pr(>|z|)
Fixed effects: (Intercept) 0.144680  0.071365  2.027  0.0426*
               magec       -0.032394 0.005163 -6.275 3.51e-10***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Correlation of Fixed Effects:
      (Intr)
magec 0.009
```

Συγκρίνοντας τα με τα προηγούμενα αποτελέσματα προσαρμογής του "μηδενικού" μοντέλου, παρατηρώ πως δεν σημειώνεται σημαντική μείωση στην $\hat{\sigma}^2_{u0}$ ($\hat{\sigma}^2_{u0} = 1.422$ από 1.4614). Οπότε η κατανομή της ηλικίας της μητέρας είναι όμοια μεταξύ των κοινοτήτων. Η προσαρμογή του μοντέλου έχει την εξής μορφή:

$$\log \frac{\hat{\pi}_{ij}}{1-\hat{\pi}_{ij}} = 0.144 - 0.032 magec_{ij}.$$

Ένα γράφημα των γραμμών πρόβλεψης των κοινοτήτων θα απεικονίζει ένα σετ παράλληλων γραμμών. Για την κατασκευή του θα υπολογίσουμε πρώτα την πιθανότητα πρόβλεψης λήψης προγεννητικής φροντίδας για κάθε μητέρα με την εντολή

```
predprob <- fitted(fit2)
```

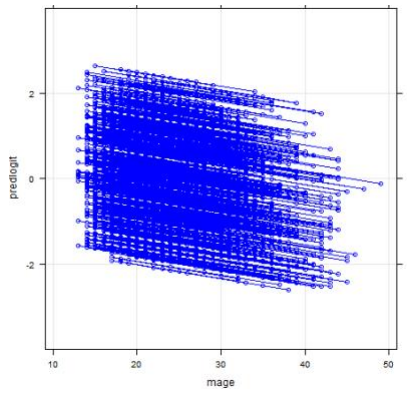
Η εντολή αυτή εξάγει τις προβλεπόμενες τιμές αποτελώντας εναλλακτική της εντολής *predict* που δεν είναι διαθέσιμη για μοντέλα που προσαρμόζονται με

εντολές όπως οι $lmer$ και $glmer$. Εν συνεχεία, μετασχηματίζουμε τις πιθανότητες πρόβλεψης για να προβλέψουμε το log-odds με τη βοήθεια της $logit()$ συνάρτησης της $VGAM$ βιβλιοθήκης. Δημιουργούμε λοιπόν ένα νέο πλαίσιο δεδομένων που περιέχει μόνο μια παρατήρηση για κάθε τιμή της $mage$ μεταβλητής μέσα σε κάθε κοινότητα.

```
> datapred <- unique(data.frame(cbind(predlogit = predlogit,
comm = mydata$comm, mage = mydata$mage)))
```

Τέλος με την παρακάτω εντολή προκύπτει το γράφημα των πιθανοτήτων πρόβλεψης της προγεννητικής φροντίδας ως προς την ηλικία της μητέρας

```
> xyplot(predlogit ~ mage, data = datapred, groups = comm, type = c("p",
"l", "g"), col = "blue", xlim = c(9, 51), ylim = c(-4, 4))
```



Σχήμα 5.2

Σύμφωνα με το παραπάνω γράφημα για μια γυναίκα ηλικίας 22 ετών, το log-odds της λήψης προγεννητικής φροντίδας κυμαίνεται από -2.2 έως 2.5, αναλόγως την κοινότητα στην οποία ζει. Απόλογαριθμίζοντας τις τιμές έχουμε ότι η πιθανότητα λήψης προγεννητικής φροντίδας κυμαίνεται μεταξύ 0.10 και 0.92. Συνεπώς, υπάρχουν σημαντικές επιδράσεις μεταξύ των κοινοτήτων (strong community effects).

Όπως προείπαμε η επίδραση της $mage$ μεταβλητής παρουσιάζει μια καμπυλότητα. Για το λόγο αυτό θα μετασχηματίσουμε τη $mage$ σε μία τετραγωνική

συνάρτηση και έτσι έχουμε το ακόλουθο μοντέλο:

$$\log \frac{\pi_{ij}}{1-\pi_{ij}} = \beta_0 + \beta_1 magec_{ij} + \beta_2 magecsq_{ij} + u_{0j}$$

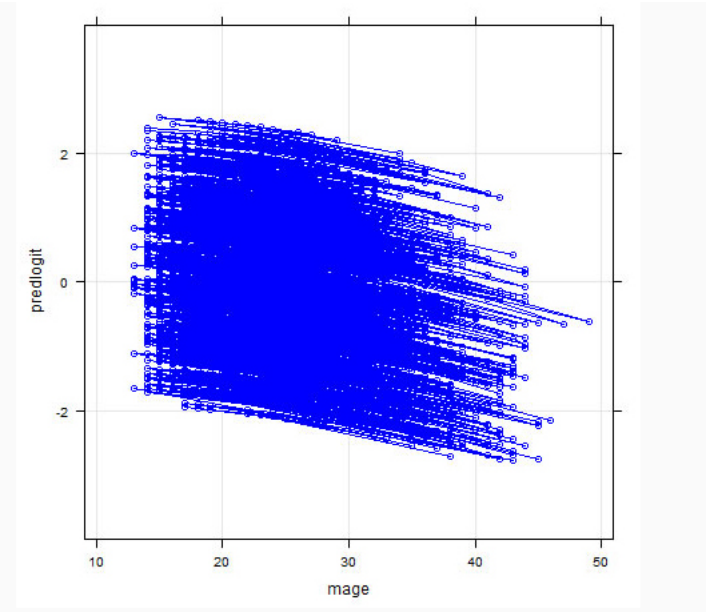
το οποίο και προσαρμόζουμε στην R ακολούθως:

```
> (fit3 <- glmer(antemed ~ magec + magecsq + (1 | comm),
family = binomial("logit"), data = mydata))
Generalized linear mixed model fit by the Laplace approximation
Formula: antemed ~ magec + magecsq + (1 | comm)
Data: mydata
  AIC   BIC  logLik deviance
 6603  6629 -3297     6595
Random effects:
 Groups Name        Variance Std. Dev.
 comm.  (Intercept) 1.4519   1.2049
Number of obs: 5366, groups: comm, 361
Fixed effects:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  0.1849150  0.0753437   2.454    0.0141*
magec       -0.0276123  0.0059222  -4.663    3.12e-06***
magecsq     -0.0010648  0.0006544  -1.6270.1037
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Correlation of Fixed Effects:
            (Intr) magec
magec        0.167
magecsq     -0.328 -0.481
```

Από τα παραπάνω δεδομένα εμμένουμε στο γεγονός ότι η $\hat{\beta}_2 = -0.001$ δεν είναι στατιστικά σημαντική σε επίπεδο σημαντικότητας α=5% (p-value $= 0.1 > 0.05$). Παρ' όλα αυτά δεν την απομακρύνουμε από το μοντέλο τουλάχιστον για τώρα.

"Τρέχουμε" πάλι τις εντολές που βοήθησαν στην κατασκευή του προηγούμενου γραφήματος, για να προκύψει ένα νέο γράφημα για την νέα *predlogit* μεταβλητή ως προς την *mage*.

```
> predprob <- lme4::fitted(fit3)
> predlogit <- logit(predprob)
> datapred <- unique(data.frame(cbind(predlogit = predlogit, comm = mydata$comm,
```



Σχήμα 5.3

mage = mydata$mage)))
> xyplot(predlogit ∼ mage, data = datapred, groups = comm, type = c(”p”, ”l”, ”g”), col = ”blue”, xlim = c(9, 51), ylim = c(-4, 4))

Παρατηρούμε πως οι γραμμές πρόβλεψης είναι ελαφρώς καμπυλόγραμμες λόγω της $magecsq$ (τετραγωνικός όρος $mage$ μεταβλητής), αλλά παραμένουν παράλληλες εξαιτίας της σχέσης της προγεννητικής φροντίδας με την ηλικία που θεωρείται η ίδια σε κάθε κοινότητα.

5.3 Διεπίπεδο μοντέλο τυχαίων κλίσεων (Two-level Random Slope Model)

Τα πολυεπίπεδα μοντέλα που έχουμε θεωρήσει έως τώρα, επιτρέπουν στην μεταβλητή απόκρισης να διαφέρει ανά ομάδα, συμπεριλαμβάνοντας ένα τυχαίο u_j όρο επιπέδου 2, για την επεξηγηματική μεταβλητή του μοντέλου. Παρ' όλα αυτά ο τυχαίος αυτός όρος επηρεάζει μόνο το σταθερό όρο του μοντέλου έτσι ώστε ο σταθερός όρος για την j ομάδα να είναι ο $\beta_0 + u_j$. Η επίδραση κάθε επεξηγηματικής μεταβλητής x παραμένει σταθερή για κάθε ομάδα. Θε-

ωρούμε τώρα μοντέλα τυχαίων κλίσεων που επιτρέπουν στην επίδραση μίας ή περισσοτέρων επεξηγηματικών μεταβλητών να διαφέρουν μεταξύ των ομάδων.

5.3.1 Λογιστικό (logit) μοντέλο τυχαίων κλίσεων

Στην παράγραφο 3.3.1 θεωρήσαμε μοντέλα τυχαίων κλίσεων για συνεχή μεταβλητή απόκρισης που συμπεριλάμβανε την προσάρτηση ενός τυχαίου όρου σε μία ή περισσότερες επεξηγηματικές μεταβλητές. Μπορούμε να κάνουμε το ίδιο και με ένα γενικευμένο γραμμικό μοντέλο για δυαδική μεταβλητή απόκρισης. Για παράδειγμα, το $logit$ μοντέλο τυχαίων σταθερών (5.4) μπορεί να επεκταθεί στο μοντέλο τυχαίων κλίσεων:

$$\log \frac{\pi_{ij}}{1-\pi_{ij}} = \beta_0 + \beta_1 x_{ij} + u_{0j} + u_{1j} x_{ij.} \tag{5.5}$$

Όπως και στην περίπτωση για συνεχή μεταβλητή απόκρισης έτσι και εδώ, προσθέσαμε ένα νέο $u_{1j}x_{ij}$ όρο και ένα δείκτη '0' στο υπόλοιπο του σταθερού όρου. Όπως και πριν, οι τυχαίες επιδράσεις u_{0j} και u_{1j} ακολουθούν κανονική κατανομή με μηδενική μέση τιμή και διασπορά σ^2_{u0} και σ^2_{u1} αντίστοιχα και συνδιακύμανση σ_{u01}. Επειδή οι u_{0j} και u_{1j} επιτρέπεται να συσχετίζονται (δηλαδή $\sigma_{u01} \neq 0$) ακολουθούν διμεταβλητή κανονική κατανομή που μπορεί συνοπτικά να εκφραστεί ως:

$$\mathbf{u} = \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim MVN(\mathbf{0}, \Omega_u)$$

όπου MVN "multivariate normal" δηλαδή διμεταβλητή κανονική κατανομή,

$$\mathbf{0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ και } \Omega_u = \begin{pmatrix} \sigma^2_{u0} & \sigma_{u01} \\ \sigma_{u01} & \sigma^2_{u1} \end{pmatrix}$$

ο πίνακας διακύμανσης τυχαίων επιδράσεων. Η κλίση της γραμμικής σχέσης μεταξύ της επεξηγηματικής μεταβλητής είναι του log-odds για $y = 1$ είναι $\beta_1 + u_{1j}$ για j ομάδα. Η συνδιακύμανση τυχαίων επιδράσεων σ_{u01}, είναι η διακύμανση εκείνη μεταξύ των σταθερών όρων και κλίσεων των ομάδων.

5.3.2 Επιτρέποντας στην επίδραση της wealth μεταβλητής να διαφέρει μεταξύ των κοινοτήτων

Στο μοντέλο αυτό, οι επιδράσεις της ηλικίας της μητέρας, της μόρφωσης και του οικογενειακού εισοδήματος παραμένουν ίδιες σε κάθε κοινότητα. Θα

προσαρμόσουμε τώρα μία τυχαία μεταβλητή $wealth$ επιτρέποντας έτσι στην επίδραση της να διαφέρει μεταξύ των κοινοτήτων.

$$\log \frac{\pi_{ij}}{1-\pi_{ij}} = \beta_0 + \beta_1 magec_{ij} + \beta_2 magecsq_{ij} + \beta_3 meduc2_{ij} + \beta_4 meduc3_{ij} + \beta_5 wealthc_{ij} + u_{0j} + u_{5j} wealthc_{ij}.$$

Σημειώνουμε την προσθήκη ενός νέου όρου, u_{5j} στο μοντέλο έτσι ώστε ο συντελεστής της $wealth$ μεταβλητής να γίνει $\beta_{5j} = \beta_5 + u_{5j}$ και η διακύμανση στο επίπεδο της κοινότητας να αντικατασταθεί από ένα πίνακα με δύο νέες παραμέτρους, σ^2_{u5} και σ_{u05}. Τα υπόλοιπα στο επίπεδο της κοινότητας ακολουθούν διμεταβλητή κανονική κατανομή με μηδενικό μέσο $\mathbf{0}$ και πίνακα διακύμανσης-συνδιακύμανσης Ω_u

$$\begin{pmatrix} u_{0j} \\ u_{5j} \end{pmatrix} \sim MVN(0, \Omega_{ij}), \quad 0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \Omega_u = \begin{pmatrix} \sigma^2_{u0} & \\ \sigma_{u05} & \sigma^2_{u5} \end{pmatrix}$$

Προσαρμόζοντας το μοντέλο μας με την ακόλουθη εντολή λαμβάνουμε τα παρακάτω αποτελέσματα.

```
> (fit <- glmer(antemed ~ magec + magecsq + meduc2 + meduc3 + wealthc
+ (1 + wealthc | comm), data = mydata, family = binomial("logit")))
Generalized linear mixed model fit by the Laplace approximation
Formula: antemed ~ magec + magecsq + meduc2 + meduc3 + wealthc + (1 +
wealthc | comm)
Data: mydata
```

AIC	*BIC*	*logLik*	*deviance*
5987	6046	−2984	5969

Random effects:

Groups	*Name*	*Variance*	*Std. Dev.*	*Corr*
comm	*(Intercept)*	0.842703	0.91799	
	wealthc	0.015906	0.12612	−0.946

Number of obs: 5366, groups: comm, 361

Fixed effects:

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	−0.4556524	0.0824081	−5.529	$3.22e-08$ ***
magec	−0.0001317	0.0065367	−0.020	0.984
magecsq	−0.0010578	0.0006819	−1.551	0.121
meduc2	0.5478625	0.0847046	6.468	$9.94e-11$ ***
meduc3	1.3102785	0.0968801	13.525	$< 2e-16$ ***
wealthc	0.4058495	0.0302306	13.425	$< 2e-16$ ***

—
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Correlation of Fixed Effects:

	$(Intr)$	$magec$	$magcsq$	$meduc2$	$meduc3$
$magec$	-0.022				
$magecsq$	-0.279	-0.489			
$meduc2$	-0.545	0.206	-0.047		
$meduc3$	-0.543	0.307	-0.069	0.553	
$wealthc$	0.008	-0.120	0.069	-0.159	-0.343

Αξίζει να σημειωθεί πως ο τρόπος ορισμού του τυχαίου όρου $((1+wealthc|comm))$ υποδηλώνει πως οι τυχαίες κλίσεις και τυχαίες σταθερές συνδιακυμαίνονται.

΄Ελεγχος τυχαίων κλίσεων

Χρησιμοποιούμε ένα LR-έλεγχο για να εξετάσουμε αν η επίδραση της $wealth$ μεταβλητής διαφέρει μεταξύ των ομάδων. Η μηδενική υπόθεση ελέγχει αν οι δύο νέες παράμετροι είναι ταυτόχρονα μηδενικές $\sigma^2_{u5} = \sigma_{u05} = 0$. Συγκρίνουμε λοιπόν το μοντέλο της προηγούμενης παραγράφου (χωρίς τυχαία κλίση) με αυτό το μοντέλο και συνεπώς $LR = 2[(-2984) - (-2990)] = 12$ με δύο βαθμούς ελευθερίας. Τελικά συμπεραίνουμε πως η επίδραση του εισοδήματος δεν διαφέρει μεταξύ των ομάδων.

5.3.3 Ερμηνεία μοντέλου τυχαίων κλίσεων

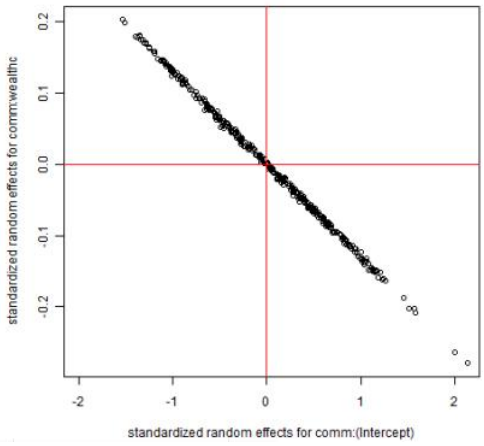
Από τα αποτελέσματα προσαρμογής του μοντέλου της παραγράφου 5.3.2 η εκτιμήτρια της επίδρασης της $wealth$ μεταβλητής είναι $0.405 + \hat{u}_{5j}$, ενώ η διακύμανση μεταξύ των ομάδων είναι 0.016. Τέλος, η διακύμανση του σταθερού όρου είναι $\hat{\sigma}^2_{u0} = 0.843$.

Εξετάζοντας τα υπόλοιπα του σταθερού όρου και των κλίσεων (intercept and slope residuals) των κοινοτήτων

Η αρνητική εκτιμήτρια διακύμανης του σταθερού όρου-κλίσης ($\hat{\sigma}_{u5} = -0.110$) υποδηλώνει ότι οι κοινότητες με πρόσληψη προγεννητικής φροντίδας κάτω του μέσου όρου (υπόλοιπο σταθερού όρου $\hat{u}_{0j} > 0$) τείνουν να έχουν επιδράσεις εισοδήματος άνω του μέσου όρου (υπόλοιπο κλίσης $\hat{u}_{5j} < 0$). Με άλλα λόγια,

στις κοινότητες με υψηλό ποσοστό πρόσληψης προγεννητικής φροντίδας, η βαθμίδα εισοδήματος είναι χαμηλότερη. Για την κατασκευή ενός γραφήματος των κλίσεων των κοινοτήτων της $wealth$ μεταβλητής ($\hat{u}_{5j}$ συναρτήσει $\hat{u}_{0j}$) θα χρησιμοποιήσουμε τις ακόλουθες εντολές:

```
> reffects <- ranef(fit, postVar = TRUE)
> plot(reffects[[1]], xlab = "standardized random effects for comm:(Intercept)",ylab
= "standardized random
effects for comm:wealthc", xlim = c(-2, 2.1))
> abline(v = 0, col = "red")
> abline(h = 0, col = "red")
```



Σχήμα 5.4

Εάν γνωρίζαμε την γεωγραφική τοποθεσία των κοινοτήτων θα ήταν πολύ ενδιαφέρον να ταυτοποιήσουμε τις κοινότητες με χαμηλή πρόσληψη και "απότομες" κλίσεις για το εισόδημα (δηλαδή κοινότητες στο 2ο τεταρτημόριο). Έτσι θα μπορούσαν να γίνουν προσπάθειες βελτίωσης των υπηρεσιών υγείας για τις μητέρες και να στοχοποιηθούν τέτοιες περιοχές.

Γραμμές πρόβλεψης των κοινοτήτων (Community prediction lines)

Το προσαρμοσμένο μοντέλο για την κοινότητα j, για μια γυναίκα μέσης ηλικίας ($magec = 0$,$magecsq = 0$) χωρίς μόρφωση (κατηγορία αναφοράς της

$meduc$: $meduc2 = 0$, $meduc3 = 0$) είναι:

$$log\frac{\widehat{\pi}_{ij}}{1-\widehat{\pi}_{ij}} = (-0.456 + \widehat{u}_{0j}) + (0.399 + \widehat{u}_{5j})wealthc_{ij}$$

Στην γραμμή προσαρμογής του μοντέλου για γυναίκες διαφορετικών ηλικιών ή μορφωτικού επιπέδου μόνο ο σταθερός όρος αλλάζει. Πιο συγκεκριμένα για μια γυναίκα με μόρφωση επιπέδου πρωτοβάθμιας εκπαίδευσης, έχουμε αύξηση του σταθερού όρου από -0.456 σε $-0.456+0.39 = 0.083$ Για να δημιουργήσουμε τώρα ένα γράφημα που θα απεικονίζει τις γραμμές πρόβλεψης των κοινοτήτων υπολογίζουμε αρχικά το $logit(\hat{\pi}_{ij})$ δηλαδή το log-odds πρόβλεψης, για κάθε γυναίκα βάσει της τιμής της μεταβλητής $wealthc$ και της κοινότητας της κατοικίας. Για την κατασκευή του γραφήματος ακολουθούμε στην R την παρακάτω διαδικασία. Αποθηκεύουμε αρχικά τα δεδομένα μας σε ένα νέο πλαίσιο δεδομένων με την ονομασία $mydatapred$ μέσω της εντολής

```
> mydatapred <- mydata
```

Εν συνεχεία, μηδενίζουμε τις μεταβλητές $magec$, $magevsq$, $meduc2$, και $meduc3$

```
> mydatapred$magec <- 0
> mydatapred$magecsq <- 0
> mydatapred$meduc2 <- 0
> mydatapred$meduc3 <- 0
```

Ακολούθως, υπολογίζουμε το log-odds πρόβλεψης για κάθε γυναίκα, βασιζόμενοι στο σταθερό τμήμα του μοντέλου

```
> X <- model.matrix(terms(fit), mydatapred)
> b <- fixef(fit)
> predlogit <- X %*% b
```

Συνδυάζουμε τα log-odds πρόβλεψης με τις τιμές των $wealthc$, $wealth$ και $comm$ μεταβλητών σε ένα νέο πλαίσιο δεδομένων $(datapred2)$

```
> mydatapred2 <- unique(data.frame(cbind(predlogit = predlogit, wealthc =
mydatapred$wealthc, comm = mydatapred$comm, wealth = mydatapred$wealth)))
> colnames(mydatapred2)[1] <- c("predlogit")
```

Ορίζουμε τον δείκτη για τις κοινότητες (ή τυχαίες επιδράσεις)

```
> re_id <- as.integer(rownames(reffects$comm))
```

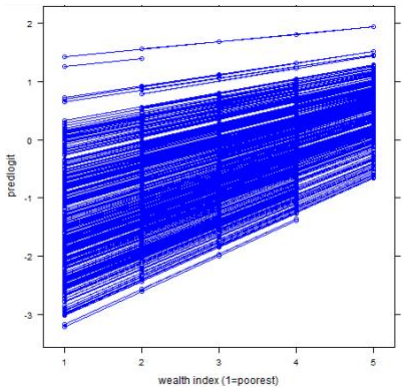
Προσθέτουμε το τυχαίο μέρος του μοντέλου πρόβλεψης αντικαθιστώντας τα log-odds με τα u_0 και u_1 της αντίστοιχης κοινότητας

```
> u0 <- data.frame(cbind(commid = re_id, u0 = reffects[[1]][, 1]))
> u1 <- data.frame(cbind(commid = re_id, u1 = reffects[[1]][, 2]))
> for (i in 1:dim(mydatapred2)[1]) { mydatapred2$predlogit[i] <-
mydatapred2$predlogit[i] + u0$u0[u0$commid == mydatapred2$comm[i]] +
u1$u1[u1$commid == mydatapred2$comm[i]] * mydatapred2$wealthc[i] }
```

Οπότε τελικά με την παρακάτω εντολή προκύπτει το ζητούμενο γράφημα γραμμών πρόβλεψης των κοινοτήτων

```
> xyplot(predlogit ~ wealth, data = mydatapred2, groups = comm, type =
c("p","l"), col = "blue", xlab = "wealth index (1 = poorest)")
```

Παρατηρούμε ότι κάποιες κάποιες γραμμές είναι πιο "κοντές" από κάποιες



Σχήμα 5.5

άλλες επειδή δεν περιέχουν όλες οι κοινότητες γυναίκες με υψηλό εισόδημα. Επίσης μπορούμε να δούμε ότι οι γραμμές των κοινοτήτων είναι πιο "απλωμένες" καθώς αυξάνεται το εισόδημα. Αυτό είναι αναμενόμενο λόγω της αρνητικής συσχέτισης μεταξύ των υπολοίπων κλίσης και του σταθερού όρου.

Διακύμανση μεταξύ των κοινοτήτων ως συνάρτηση εισοδήματος

Από το προηγούμενο γράφημα γραμμών πρόβλεψης για κάθε κοινότητα, μπορούμε να δούμε πως οι γραμμές είναι πιο απλωμένες για τα χαμηλότερα πεμπτημόρια του $wealth$ δείκτη από τα υψηλότερα. Με άλλα λόγια, η διακύμανση του λογαριθμοποιημένου λόγου πιθανότητας (log-odds) της πρόσληψης προγεννητικής φροντίδας μειώνεται με την αύξηση του εισοδήματος. Η προσθήκη τυχαίας κλίσης στην $wealth$ μεταβλητή, υποδηλώνει πως η διακύμανσης μεταξύ των κοινοτήτων είναι συνάρτηση της μεταβλητής $wealth$ και όχι σταθερή όπως στο μοντέλο τυχαίων σταθερών. Η συνάρτηση διακύμανσης στο επίπεδο κοινοτήτων παίρνει την ακόλουθη μορφή:

$$\begin{aligned} V(u_{0j} + u_{5j} wealth_{ij}) &= V(u_{0j}) + 2\text{cov}(u_{0j},\ u_{5j}) wealth_{ij} + V(u_{5j}) wealth_{ij}^2 \\ &= \sigma_{u0}^2 + 2\sigma_{u05} wealth_{ij} + \sigma_{u5}^2 wealth_{ij}^2 \end{aligned}$$

η οποία εκτιμάται (αντικαθιστώντας τις εκτιμήτριες των σ_{u0}^2, σ_{u05}, σ_{u5}^2)

$$0.843 - 0.219 wealth_{ij} + 0.016 wealth_{ij}^2$$

Έτσι μπορούμε να κατασκευάσουμε γράφημα της διακύμανσης μεταξύ των κοινοτήτων ως συνάρτηση του εισοδήματος ($wealth$ μεταβλητή) με τις ακόλουθες εντολές.

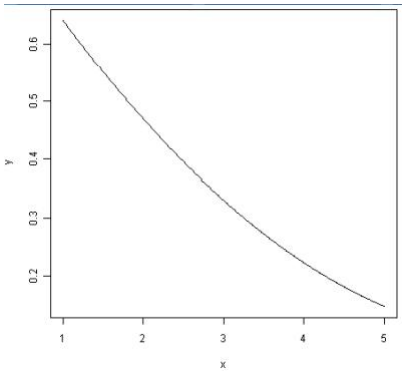
```
> x <- seq(1, 5, 0.01)
> y < −0.843 + (−0.219) ∗ x + 0.016 ∗ x^2
> plot(x, y, type = "l", xlim = c(1, 5))
```

Όπως ήταν αναμενόμενο, η διακύμανση μεταξύ των κοινοτήτων ακολουθεί το πρότυπο των γραμμών πρόβλεψης του προηγούμενου γραφήματος, δηλαδή μειώνεται ως συνάρτηση της μεταβλητής $wealth$.

Αν η διακύμανση είναι μία τετραγωνική συνάρτηση της $wealth$ μεταβλητής, το γράφημα απεικονίζει μια γραμμική μείωση. Αυτό συμβαίνει επειδή ο συντελεστής του γραμμικού όρου ($2\hat{\sigma}_{u05} = 0.219$) έχει μεγαλύτερη τιμή από τον τετραγωνικό ($\hat{\sigma}_{u5}^2 = 0.016$), σε βαθμό που η $\hat{\sigma}_{u5}^2$ να συνεισφέρει ελάχιστα στη διακύμανση μεταξύ των κοινοτήτων.

5.3.4 Προσαρμογή τυχαίων συντελεστών στη κατηγορική μεταβλητή wealth

Επειδή η μεταβλητή $wealth$ είναι κατηγορική, θα δημιουργήσουμε ψευδομεταβλητές για 4 από τις 5 κατηγορίες εισάγοντάς τες ως επεξηγηματικές



Σχήμα 5.6

μεταβλητές. Αντιμετωπίζουμε τη μεταβλητή $wealth$ ως συνεχή μεταβλητή, βασιζόμενοι στο γεγονός ότι οι συντελεστές των $wealth$ ψευδομεταβλητών εμφανίζει μια σχεδόν γραμμική αύξηση. Για την ανάλυση των τυχαίων κλίσεων, έχουμε θεωρήσει ότι η σχέση της $wealth$ μεταβλητής με τους log-odds της προγεννητικής φροντίδας είναι γραμμική σε όλες τις κοινότητες. Έχουμε όμως επιτρέψει στην κλίση αυτής της σχέσης να μεταβάλλεται μεταξύ των κοινοτήτων. Αυτό υποδηλώνει ότι η διακύμανση μεταξύ των κοινοτήτων είναι μια τετραγωνική συνάρτηση αν και το γράφημα της συνάρτησης διακύμανσης δείχνει πως η διακύμανση εξαρτάται γραμμικά από το εισόδημα. Στην παράγραφο αυτή, θα επανεξετάσουμε την υπόθεση ότι η σχέση εισοδήματος-προγεννητικής φροντίδας είναι γραμμική, με τη βοήθεια ψευδομεταβλητών. Έτσι, θα έχουμε μια πιο ευέλικτη συνάρτηση για τη διακύμανση μεταξύ των κοινοτήτων με το μειονέκτημα της προσθήκης πολλών παραμέτρων στο μοντέλο. Για το λόγο αυτό θα ερευνήσουμε αν μπορούμε να απλοποιήσουμε το μοντέλο έχοντας τυχαίους συντελεστές για τις $wealth$ ψευδομεταβλητές. Προσαρμόζουμε λοιπόν το ακόλουθο μοντέλο τυχαίων σταθερών:

$$\begin{aligned}\log\frac{\pi_{ij}}{1-\pi_{ij}} =& \beta_0 + \beta_1 magec_{ij} + \beta_2 magecsq_{ij} + \beta_3 meduc2_{ij} + \beta_4 meduc3_{ij} \\ &+ \beta_5 wealth2_{ij} + \beta_6 wealth3_{ij} + \beta_7 wealth4_{ij} + \beta_8 wealth5_{ij} + u_j\end{aligned}$$

στην R με τη βοήθεια της ακόλουθης εντολής, λαμβάνοντας τα εξής αποτελέσματα:

> (fit2 <- glmer(antemed ∼ magec + magecsq + meduc2 + meduc3 + wealth2 + wealth3 + wealth4 + wealth5 + (1 | comm), family = binomial("logit"), data = mydata))

Generalized linear mixed model fit by the Laplace approximation
Formula: antemed ∼ magec + magecsq + meduc2 + meduc3 + wealth2 + wealth3 + wealth4 + wealth5 + (1 | comm)
Data: mydata

AIC	*BIC*	*logLik*	*deviance*
5989	6055	−2985	5969

Random effects:

Groups	*Name*	*Variance*	*Std. Dev.*
comm.	(*Intercept*)	0.82392	0.9077

Number of obs: 5366, groups: comm, 361
Fixed effects:

	Estimate	*Std. Error*	*z value*	$Pr(>\|z\|)$
(*Intercept*)	−1.2513452	0.1016866	−12.306	$< 2e-16 ***$
magec	−0.0012448	0.0065373	−0.190	0.849
magecsq	−0.0010025	0.0006805	−1.473	0.141
meduc2	0.5529854	0.0843573	6.555	$5.55e-11 ***$
meduc3	1.3029666	0.0973852	13.380	$< 2e-16 ***$
wealth2	0.4683034	0.1058906	4.423	$9.76e-06 ***$
wealth3	0.6858771	0.1080670	6.347	$2.20e-10 ***$
wealth4	1.0569162	0.1141125	9.262	$< 2e-16 ***$
wealth5	1.7741832	0.1330637	13.333	$< 2e-16 ***$

—
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Correlation of Fixed Effects:

	(*Intr*)	*magec*	*magcsq*	*meduc2*	*meduc3*	*welth2*	*welth3*	*welth4*
magec	0.035							
magecsq	−0.251	−0.488						
meduc2	−0.317	0.205	−0.045					
meduc3	−0.228	0.312	−0.072	0.547				
wealth2	−0.507	−0.010	−0.013	−0.098	−0.113			
wealth3	−0.510	−0.030	0.021	−0.123	−0.194	0.544		
wealth4	−0.483	−0.067	0.038	−0.165	−0.281	0.525	0.572	
wealth5	−0.453	−0.119	0.060	−0.147	−0.325	0.462	0.507	0.550

Όπως και στην παράγραφο 4.4.5 η πιθανότητα λήψης προγεννητικής φροντίδας αυξάνεται με βάσει το εισόδημα. Η μεταβολή δεν είναι απολύτως γραμμική, εξαιτίας π.χ. της διαφοράς $\hat{\beta}_6 - \hat{\beta}_5 = 0.686 - 0.468 = 0.218$. Θα προχωρήσουμε τώρα στην εξέταση της σχέσης της πιθανότητας (probability) λήψης προγεννητικής φροντίδας και εισοδήματος βάσει ηλικίας και μορφωτικού επιπέδου.

Αρχικά, αποθηκεύουμε τα τρέχοντα αποτελέσματα, αντιγράφοντας τα δεδομένα μας σε ένα νέο πλαίσιο δεδομένων (*datapred*3).

```
> mydatapred3 <- mydata
```

Αντικαθιστούμε τις τιμές των εκτιμητριών των *magec*, *meduc*2, και *meduc*3 μεταβλητών με τις μέσες τιμές τους και θέτουμε την *magecsq* μεταβλητή ως το τετράγωνο της μέσης τιμής της *magec*:

```
> mydatapred3$magec <- 0
> mydatapred3$magecsq <- 0
> mean(mydatapred3$meduc2)
 [1] 0.3073053
 > mydatapred3$meduc2 <- mean(mydatapred3$meduc2)
> mean(mydatapred3$meduc3)
 [1] 0.3449497
> mydatapred3$meduc3 <- mean(mydatapred3$meduc3)
```

Υπολογίζουμε την πιθανότητα πρόβλεψης κάθε γυναίκας βασιζόμενοι μόνο στο σταθερό μέρος του μοντέλου.

```
> X <- model.matrix(terms(fit2), mydatapred3)
> b <- fixef(fit2)
> predlogit <- X %*% b
> predprob <- logit(predlogit, inverse = TRUE)
```

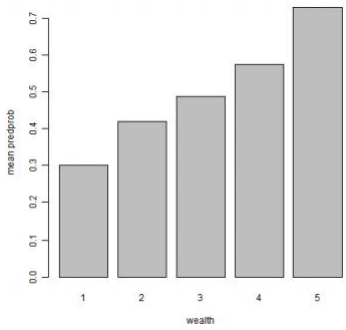
Έπειτα, συνδυάζουμε την πιθανότητα πρόβλεψης με την κατηγορική μεταβλητή *wealth* σε ένα νέο πλαίσιο δεδομένων.

```
> mydatapred3.2 <- unique(data.frame(cbind(predprob = predprob, wealth =
mydatapred3$wealth)))
> colnames(mydatapred3.2)[1] <- ”predprob”
```

Τέλος, με χρήση της *tapply* εντολής υπολογίζουμε τη μέση πιθανότητα κάθε κατηγορίας της μεταβλητής *wealth* και χρησιμοποιούμε την *barplot* εντολή για την κατασκευή ραβδογράμματος.

```
> mean.table <- tapply(mydatapred3.2$predprob, mydatapred3.2$wealth, mean)
```

> barplot(mean.table, ylab = ”mean predprob”, xlab = ”wealth”)



Σχήμα 5.7

Παρατηρούμε ότι η σχέση της πιθανότητας προγεννητικής φροντίδας και εισοδήματος είναι σχεδόν γραμμική. Σε ένα πιο γενικό μοντέλο τυχαίων συντελεστών[2], μπορούμε να επιτρέψουμε τη μεταβολή των συντελεστών και για τις 4 $wealth$ ψευδομεταβλητές μεταξύ των κοινοτήτων. Αυτό όμως είναι ένα αρκετά σύνθετο μοντέλο που οδηγεί σε ένα 5×5 πίνακα διακύμανσης στο επίπεδο των κοινοτήτων με 14 παραμέτρους περισσότερες από το μοντέλο τυχαίων σταθερών. Ξεκινάμε με ένα απλούστερο μοντέλο με τυχαίο συντελεστή μόνο για τη $wealth5$ μεταβλητή. Το μοντέλο αυτό υποδηλώνει ότι η διακύμανση μεταξύ των κοινοτήτων είναι η ίδια για τα πεμπτημόρια 1-4, αλλά διαφορετική στο 5. Προσαρμόζουμε το μοντέλο

$$\begin{aligned}\log \frac{\pi_{ij}}{1-\pi_{ij}} = \beta_0 &+ \beta_1 magec_{ij} + \beta_2 magecsq_{ij} + \beta_3 meduc2_{ij} + \beta_4 meduc3_{ij} \\ &+ \beta_5 wealth2_{ij} + \beta_6 wealth3_{ij} + \beta_7 wealth4_{ij} \\ &+ \beta_8 wealth5_{ij} + u_{0j} + u_{8j} wealth5_{ij}\end{aligned}$$

με τη βοήθεια της παρακάτω εντολής λαμβάνοντας τα εξής αποτελέσματα:

> (fit3 <- glmer(antemed ∼ magec + magecsq + meduc2 + meduc3 + wealth2 + wealth3 + wealth4 +
wealth5 + (1 + wealth5 | comm), family = binomial(”logit”), data = mydata))

[2]Σημειώστε την αλλαγή στην ορολογία από κλίση σε συντελεστή. Οι δύο όροι χρησιμοποιούνται εναλλακτικά, αλλά ο όρος κλίση είναι κατάλληλος μόνο σε γραμμική σχέση.

Generalized linear mixed model fit by the Laplace approximation
Formula: antemed m̃agec + magecsq + meduc2 + meduc3 + wealth2 + wealth3 +
wealth4 + wealth5 + (1 + wealth5 | comm)
Data: mydata

AIC	BIC	$logLik$	$deviance$
5985	6064	-2981	5961

Random effects:

$Groups$	$Name$	$Variance$	$Std.Dev.$	$Corr$
$comm$	$(Intercept)$	0.92694	0.96278	
	$wealth5$	0.40030	0.63270	-0.695

Number of obs: 5366, groups: comm, 361
Fixed effects:

	$Estimate$	$Std.Error$	$zvalue$	$Pr(>\|z\|)$
$(Intercept)$	-1.2627497	0.1039153	-12.152	$< 2e-16 * **$
$magec$	-0.0007792	0.0065613	-0.119	0.905
$magecsq$	-0.0010308	0.0006829	-1.509	0.131
$meduc2$	0.5526050	0.0848369	6.514	$7.33e-11 * **$
$meduc3$	1.3115543	0.0975750	13.442	$< 2e-16 * **$
$wealth2$	0.4681434	0.1066942	4.388	$1.15e-05 * **$
$wealth3$	0.6842719	0.1089179	6.282	$3.33e-10 * **$
$wealth4$	1.0596738	0.1153505	9.187	$< 2e-16 * **$
$wealth5$	1.8495106	0.1339629	13.806	$< 2e-16 * **$

—
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Correlation of Fixed Effects:

	$(Intr)$	$magec$	$magcsq$	$meduc2$	$meduc3$	$welth2$	$welth3$	$welth4$
$magec$	0.033							
$magecsq$	-0.245	-0.487						
$meduc2$	-0.311	0.206	-0.046					
$meduc3$	-0.224	0.312	-0.071	0.549				
$wealth2$	-0.500	-0.009	-0.014	-0.098	-0.113			
$wealth3$	-0.503	-0.029	0.020	-0.123	-0.193	0.544		
$wealth4$	-0.478	-0.066	0.037	-0.165	-0.279	0.523	0.572	
$wealth5$	-0.521	-0.117	0.058	-0.144	-0.314	0.457	0.497	0.528

Οι συντελεστές του μοντέλου β_5, β_6 και β_7 θεωρούνται σταθεροί μεταξύ των κοινοτήτων και ίσοι με $\hat{\beta}_5 = 0.468$, $\hat{\beta}_6 = 0.684$ και $\hat{\beta}_7 = 1.059$. Ο β_8 συντελεστής (δηλαδή η διαφορά μεταξύ των πεμπτημορίων 1 και 5) διαφέρει μεταξύ των κοινοτήτων και $\hat{\beta}_8 = 1.849 + \hat{u}_{8j}$ για την j κοινότητα. Μπορούμε επίσης να ελέγ-

ξουμε αν η διαφορά μεταξύ του πεμπτημορίου 5 και των υπολοίπων τεσσάρων πεμπτημορίων της $wealth$ μεταβλητής πράγματι μεταβάλλεται μεταξύ των κοινοτήτων. Αυτό θα γίνει, με τη βοήθεια ενός LR-ελέγχου με μηδενική υπόθεση $H_0 : \sigma^2_{u8} = \sigma_{u08} = 0$. Άρα με την εντολή $anova$ έχω:

```
> anova(fit2, fit3)
Data: mydata
Models:
fit2: antemed ~ magec + magecsq + meduc2 + meduc3 + wealth2 + wealth3 +
fit2:    wealth4 + wealth5 + (1 | comm)
fit3: antemed ~ magec + magecsq + meduc2 + meduc3 + wealth2 + wealth3 +
fit3:    wealth4 + wealth5 + (1 + wealth5 | comm)
     DfAIC   BIC   logLik     Chisq     ChiDf  Pr(>Chisq)
fit2    10  5989.2  6055.1   −2984.6
fit3    12  5985.0  6064.1  −2980.58.156    2      0.01694*
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Εφόσον p-value=0.01694, απορρίπτουμε την μηδενική υπόθεση και δεχόμαστε πως η τυχαία επίδραση της $wealth5$ μεταβλητής είναι στατιστικά σημαντική. Η διακύμανση μεταξύ των κοινοτήτων είναι

$$\begin{aligned} V(u_{0j} + u_{8j} wealth5_{ij}) &= \sigma^2_{u0} + 2\sigma_{u08} wealth5_{ij} + \sigma^2_{u8} wealth5^2_{ij} \\ &= 0.927 - 0.847 wealth5_{ij} + 0.400 wealth5^2_{ij} \end{aligned}$$

η οποία εξαιτίας της $wealth5$ μεταβλητής λαμβάνει τιμές 0 και 1, απλοποιείται σε:
0.927 στα πεμπτημόρια 1-4 ($wealth5 = 0$) και
0.480 στο πεμπτημόριο 5 ($wealth5 = 1$).

Θα επεκτείνουμε τώρα το μοντέλο, εισάγοντας ένα τυχαίο συντελεστή για την $wealth4$ μεταβλητή. Το μοντέλο αυτό επιτρέπει επιπρόσθετα, στη διαφορά μεταξύ τετάρτου πεμπτημορίου και των τριών πρώτων, να μεταβάλλεται μεταξύ των κοινοτήτων.

$$\begin{aligned} \log \frac{\pi_{ij}}{1-\pi_{ij}} = \beta_0 &+ \beta_1 magec_{ij} + \beta_2 magecsq_{ij} + \beta_3 meduc2_{ij} \\ &+ \beta_4 meduc3_{ij} + \beta_5 wealth2_{ij} + \beta_6 wealth3_{ij} + \beta_7 wealth4_{ij} \\ &+ \beta_8 wealth5_{ij} + u_{0j} + u_{7j} wealth4_{ij} + u_{8j} wealth5_{ij} \end{aligned}$$

Με τη βοήθεια της ακόλουθης εντολής

> (fit4 <- glmer(antemed ∼ magec + magecsq + meduc2 + meduc3 + wealth2 + wealth3
+ wealth4 + wealth5 + (1 + wealth4 + wealth5 | comm), family = binomial("logit"), data = mydata))

λαμβάνουμε τα παρακάτω αποτελέσματα:

Generalized linear mixed model fit by the Laplace approximation
Formula: antemed ∼ magec + magecsq + meduc2 + meduc3 + wealth2 + wealth3 +
wealth4 + wealth5 + (1 + wealth4 + wealth5 | comm)
Data: mydata

AIC	BIC	$logLik$	$deviance$
5983	$60827 - 2977$	5953	

Random effects:

$Groups$	$Name$	$Variance$	$Std.Dev.$	$Corr$	
$comm$	$(Intercept)$	1.0883	1.04323		
	$wealth4$	0.1548	0.39344	−0.833	
	$wealth5$	0.3993	0.63190	−0.764	0.278

Number of obs: 5366, groups: comm, 361
Fixed effects:

	$Estimate$	$Std.Error$	$zvalue$	$Pr(>\|z\|)$
$(Intercept)$	−1.2645335	0.1072542	−11.790	$< 2e-16 * **$
$magec$	−0.0006292	0.0065719	−0.096	0.924
$magecsq$	−0.0010570	0.0006847	−1.544	0.123
$meduc2$	0.5587145	0.0851097	6.565	$5.22e-11 * **$
$meduc3$	1.3152825	0.0974355	13.499	$< 2e-16 * **$
$wealth2$	0.4781463	0.1080120	4.427	$9.56e-06 * **$
$wealth3$	0.6909467	0.1104383	6.256	$3.94e-10 * **$
$wealth4$	1.0203383	0.1162116	8.780	$< 2e-16 * **$
$wealth5$	1.8483040	0.1355766	13.633	$< 2e-16 * **$

—
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Correlation of Fixed Effects:

	(Intr)	magec	magcsq	meduc2	meduc3	welth2	welth3	welth4
magec	0.032							
magecsq	−0.238	−0.486						
meduc2	−0.303	0.206	−0.045					
meduc3	−0.218	0.310	−0.069	0.551				
wealth2	−0.492	−0.008	−0.014	−0.097	−0.111			
wealth3	−0.498	−0.028	0.020	−0.120	−0.190	0.544		
wealth4	−0.552	−0.066	0.036	−0.167	−0.277	0.525	0.569	
wealth5	−0.537	−0.114	0.056	−0.141	−0.308	0.457	0.498	0.550

Οι τρεις νέες παράμετροι είναι: σ^2_{u7} (διακύμανση τυχαίας επίδρασης της $wealth4$ μεταβλητής), σ_{u07} (συνδιακύανση τυχαίας επίδρασης της $wealth4$ μεταβλητής και του σταθερού όρου) και σ_{u78} (συνδιακύμανση τυχαίων επιδράσεων των $wealth4$ και $wealth5$ μεταβλητών). Έπειτα, με τη βοήθεια ενός Wald-ελέγχου με $H_0 : \sigma^2_{u7} = \sigma_{u07} = \sigma_{u78} = 0$ θα εξετάσουμε αν η διαφορά μεταξύ του πεμπτημορίου 4 και των άλλων τριών μεταβάλλεται μεταξύ των κοινοτήτων. Με την εντολή $anova$ έχω:

```
> anova(fit3, fit4)
Data: mydata
Models:
fit3: antemed ~ magec + magecsq + meduc2 + meduc3 + wealth2 + wealth3 +
fit3:     wealth4 + wealth5 + (1 + wealth5 | comm)
fit4: antemed ~ magec + magecsq + meduc2 + meduc3 + wealth2 + wealth3 +
fit4:     wealth4 + wealth5 + (1 + wealth4 + wealth5 | comm)
     Df    AIC    BIClogLik  ChisqChiDfPr(>Chisq)
fit3 12 5985.0     6064.1          −2980.5
fit4 15 5983.2     6082.1          −2976.6          7.7833  3  0.05071.
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

Εφόσον p-value=0.05071 απορρίπτουμε τη μηδενική υπόθεση, συμπεραίνοντας ότι υπάρχει διαφορά μεταξύ των κοινοτήτων. Για να δούμε την εξάρτηση της

διακύμανσης βάσει εισοδήματος,

$$\begin{aligned} V(u_{0j} + u_{7j}wealth4_{ij} + u_{8j}wealth5_{ij}) &= \sigma^2_{u0} + 2\sigma_{u07}wealth4_{ij} + \sigma^2_{u7}wealth4^2_{ij} \\ &\quad + 2\sigma_{u08}wealth5_{ij} + 2\sigma_{u78}wealth4_{ij}wealth5_{ij} \\ &\quad + \sigma^2_{u8}wealth5^2_{ij} \\ &= 1.08 - 0.684wealth4_{ij} + 0.155wealth4^2_{ij} \\ &\quad - 1.006wealth5_{ij} + 0.138wealth4_{ij}wealth5_{ij} \\ &\quad + 0.399wealth5^2_{ij} \end{aligned}$$

όπου η $wealth5$ μπορεί να λάβει τιμές 0 και 1, η παραπάνω εξίσωση απλοποιείται σε
1.088 στα πεμπτημόρια 1-3 ($wealth4 = 0$, $wealth5 = 0$)
$1.088 - 0.684 + 0.155 = 0.559$ στο πεμπτημόριο 4 ($wealth4 = 1$, $wealth5 = 0$) και
$1.088 - 1.006 + 0.399 = 0.481$ στο πεμπτημόριο 5 ($wealth4 = 0$, $wealth5 = 1$).

Έτσι, η διακύμανση μεταξύ των κοινοτήτων για τα πεμπτημόρια 1-3 είναι 1.088, 0.559 για το τέταρτο και 0.481 για το πέμπτο. Το υψηλό ποσοστό διακύμανσης στα τρία χαμηλότερα πεμπτημόρια υποδηλώνει πως η κοινότητα κατοικίας έχει ισχυρότερη επίδραση στη πιθανότητα λήψης προγεννητικής φροντίδας από γυναίκες των οποίων το εισόδημα βρίσκεται στο χαμηλότερο 60% της κατανομής της $wealth$ μεταβλητής.

5.4 Προσθήκη Επεξηγηματικών Μεταβλητών Επιπέδου 2: Συναφείς Επιδράσεις

Μέχρι στιγμής έχουμε θεωρήσει τις επιδράσεις των επεξηγηματικών μεταβλητών επιπέδου 1 και των προσαρμοσμένων μοντέλων τυχαίων κλίσεων που επιτρέπουν στις επιδράσεις τους να διαφέρουν μεταξύ των κοινοτήτων. Ένα μεγάλο πλεονέκτημα της προσέγγισης της πολυεπίπεδης μοντελοποίησης είναι η δυνατότητα προσθήκης επεξηγηματικών μεταβλητών επιπέδου 1 και 2. Συγκεκριμένα, συχνά μας ενδιαφέρει αν οι μεταβλητές επιπέδου 2 μπορούν να εξηγήσουν τη διακύμανση επιπέδου 2.

Συναφείς Επιδράσεις

Θα ξεκινήσουμε εισάγοντας την $urban$ μεταβλητή στο μοντέλο που προσαρμόσαμε στο τέλος της παραγράφου 5.3.4.

$$\log \frac{\pi_{ij}}{1-\pi_{ij}} = \beta_0 + \beta_1 magec_{ij} + \beta_2 magecsq_{ij} + \beta_3 meduc2_{ij} + \beta_4 meduc3_{ij} + \beta_5 wealth2_{ij} + \beta_6 wealth3_{ij} + \beta_7 wealth4_{ij} + \beta_8 wealth5_{ij} + \beta_9 urban + u_{0j} + u_{7j} wealth4_{ij} + u_{8j} wealth5_{ij}$$

Η προσαρμογή του έχει ως εξής:

```
> (fit <- glmer(antemed ~ magec + magecsq + meduc2 + meduc3 + wealth2
+ wealth3 + wealth4 + wealth5 + urban + (1 + wealth4 + wealth5 | comm),
family = binomial("logit"), data = mydata))
 Generalized linear mixed model fit by the Laplace approximation
Formula: antemed ~ magec + magecsq + meduc2 + meduc3 + wealth2 + wealth3
+ wealth4 + wealth5 + urban + (1 + wealth4 + wealth5 | comm)
Data: mydata
  AIC  BIC logLik deviance
 5919 6024  -2943     5887
Random effects:
 Groups Name        Variance Std.Dev. Corr
 comm.  (Intercept) 0.94258  0.97086
        wealth4     0.17889  0.42295  -0.794
        wealth5     0.40963  0.64003  -0.893  0.435
Number of obs: 5366, groups: comm, 361
Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.5291684  0.1099727 -13.905  < 2e-16 ***
magec        0.0002318  0.0065795   0.035    0.972
magecsq     -0.0010812  0.0006858  -1.577    0.115
meduc2       0.5786081  0.0853165   6.782 1.19e-11 ***
meduc3       1.3708532  0.0978515  14.010  < 2e-16 ***
wealth2      0.4595245  0.1080121   4.254 2.10e-05 ***
wealth3      0.6527686  0.1103961   5.913 3.36e-09 ***
wealth4      0.9726351  0.1165610   8.344  < 2e-16 ***
wealth5      1.5299237  0.1366507  11.196  < 2e-16 ***
urban        0.9778041  0.1163354   8.405  < 2e-16 ***
—
```

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Correlation of Fixed Effects:

	(*Intr*)	*magec*	*magcsq*	*meduc2*	*meduc3*	*welth2*	*welth3*	*welth4*	*welth5*
magec	0.024								
magecsq	−0.233	−0.486							
meduc2	−0.312	0.207	−0.045						
meduc3	−0.245	0.311	−0.068	0.554					
welth2	−0.472	−0.008	−0.014	−0.097	−0.112				
welth3	−0.470	−0.029	0.021	−0.121	−0.192	0.543			
welth4	−0.514	−0.066	0.036	−0.167	−0.280	0.522	0.565		
welth5	−0.435	−0.115	0.056	−0.152	−0.328	0.456	0.498	0.552	
urban	−0.291	0.022	−0.002	0.053	0.111	−0.022	−0.044	−0.050	−0.278

Αξίζει να σημειώσουμε ότι αυτό το μοντέλο μπορεί να χρειαστεί λίγα λεπτά για να προσαρμοστεί. Συμπεραίνουμε, ότι οι γυναίκες που κατοικούν σε αστικές περιοχές είναι πιο πιθανό να χρησιμοποιήσουν υπηρεσίες προγεννητικής φροντίδας σε σχέση με τις γυναίκες αγροτικών περιοχών, από ένα ιατρικά εκπαιδευμένο πάροχο. Η διακύμανση του σταθερού όρου, που αντιπροσωπεύει τη διακύμανση μεταξύ των κοινοτήτων για οικογένειες που ανήκουν στο χαμηλότερο 60% (πεμπτημόριο 1-3) της κατανομής του εισοδήματος, έχει μειωθεί ελαφρώς από το 1.088 στο 0.942. Η διακύμανση μεταξύ των κοινοτήτων έχει μεταβληθεί στα 2 πρώτα πεμπτημόρια:

$$\begin{aligned} V(u_{0j} + u_{7j}wealth4_{ij} + u_{8j}wealth5_{ij}) &= \sigma^2_{u0} + 2\sigma_{u07}wealth4_{ij} + \sigma^2_{u07}wealth4^2_{ij} \\ &+ 2\sigma_{u08}wealth5_{ij} + 2\sigma_{u78}wealth4_{ij}wealth5_{ij} \\ &+ \sigma^2_{u08}wealth4^2_{ij} \\ &= 0.942 - 0.652wealth4_{ij} + 0.179wealth4^2_{ij} \\ &- 1.109wealth5_{ij} + 0.235wealth4_{ij}wealth5_{ij} \\ &+ 0.409wealth4^2_{ij} \end{aligned}$$

και επειδή οι $wealth4$ και $wealth5$ μπορούν να πάρουν τιμές μόνο 0 και 1, απλοποιείται σε:
0.942 στα πεμπτημόρια 1-3 ($wealth4 = 0$, $wealth5 = 0$)
0.469 στο πεμπτημόριο 4 ($wealth4 = 1$, $wealth5 = 0$)
0.942 στα πεμπτημόριο 5 ($wealth4 = 0$, $wealth5 = 1$)

Η προσθήκη της μεταβλητής $urban$ έχει εξηγήσει μερικώς τη διακύμανση μεταξύ των κοινοτήτων στην πρόληψη προγεννητικής φροντίδας για γυναίκες σε κάθε πεμπτημόριο, αλλά η μεγαλύτερη μείωση σημειώνεται στο υψηλότερο

πεμπτημόριο: η επίδραση των κοινοτήτων μεταξύ των πλουσιότερων γυναικών εξηγείται από τις αστικό-αγροτικές διαφορές στην πρόληψη. Για να διαλευκάνουμε περισσότερο αυτό το εύρημα, θα κοιτάξουμε τη σχέση μεταξύ $wealth$ και $urban$: Κατασκευάζουμε πίνακες συνάφειας για τις δύο μεταβλητές:

```
> table(mydata$wealth,mydata$urban)
     0   1
1  997 170
2  833 184
3  771 219
4  722 267
5  359 844
```

```
> table(mydata$wealth, mydata$urban)/ cbind(rowSums(table(mydata$wealth,
mydata$urban)), rowSums(table(mydata$wealth, mydata$urban)))
         0         1
1 0.8543273 0.1456727
2 0.8190757 0.1809243
3 0.7787879 0.2212121
4 0.7300303 0.2699697
5 0.2984206 0.7015794
```

Βρίσκουμε μια ισχυρή συσχέτιση μεταξύ εισοδήματος και τύπου περιοχής κατοικίας: μόνο το 30% των πιο πλουσιότερων γυναικών κατοικούν σε αγροτικές περιοχές, συγκριτικά με το 73%-85% των γυναικών στα χαμηλότερα 4 από τα 5 πεμπτημόρια του εισοδήματος.

Στη συνέχεια, ελέγχουμε αν υπάρχει συναφής επίδραση στο εισόδημα. Το να ζει μια γυναίκα σε μια καλύτερη κοινότητα (με υψηλή αναλογία οικογενειών στο υψηλότερο πεμπτημόριο του εισοδήματος) έχει επίδραση στην πιθανότητα να λάβει προγεννητική φροντίδα η οποία είναι πάνω από την επίδραση της οικονομικής της κατάστασης; Πρώτα χρειάζεται να συγκεντρώσουμε την ψευδομεταβλητή $wealth5$ στο επίπεδο κοινοτήτων αφού ταξινομήσουμε τα δεδομένα κατά μεταβλητή $comm$.

```
> mydata <- mydata[order(mydata$comm),]
> wealth5mean <- aggregate(mydata$wealth5,by=list(mydata$comm), mean)
> colnames(wealth5mean) <- c("commid","wealth5mean")
```

Τώρα προσθέτουμε αυτή τη νέα μεταβλητή στο πλαίσιο δεδομένων $mydata$:

> mydata$wealth5mean<-rep(0,dim(mydata)[1])
> for (i in 1:dim(mydata)[1]){
mydata$wealth5mean[i]<-wealth5mean$wealth5mean[wealth5mean$commid==
mydata$comm[i]]}

Η νέα μεταβλητή, $wealth5mean$, περιέχει την αναλογία γυναικών στην κοινότητα της οποίας οι οικογένειες ανήκουν στο υψηλότερο πεμπτημόριο εισοδήματος. Προσθέτωντας τώρα αυτή τη μεταβλητή στο μοντέλο

$$\begin{aligned}\log\frac{\pi_{ij}}{1-\pi_{ij}} = {} & \beta_0 + \beta_1 magec_{ij} + \beta_2 magecsq_{ij} + \beta_3 meduc2_{ij} + \beta_4 meduc3_{ij} \\ & + \beta_5 wealth2_{ij} + \beta_6 wealth3_{ij} + \beta_7 wealth4_{ij} + \beta_8 wealth5_{ij} \\ & + \beta_9 urban + \beta_{10} wealth5mean_j + u_{0j} + u_{7j} wealth4_{ij} + u_{8j} wealth5_{ij}\end{aligned}$$

θα προχωρήσουμε στην προσαρμογή του λαμβάνοντας τα ακόλουθα αποτελέσματα
> (fit2 <- glmer(antemed ∼ magec + magecsq + meduc2 + meduc3 + wealth2 + wealth3
+ wealth4 + wealth5 + urban + wealth5mean + (1 + wealth4 + wealth5 | comm), family = binomial("logit"), data = mydata))
Generalized linear mixed model fit by the Laplace approximation
Formula: antemed ∼ magec + magecsq + meduc2 + meduc3 + wealth2 + wealth3 +
wealth4 + wealth5 + urban + wealth5mean + (1 + wealth4 + wealth5 | comm)
Data: mydata

AIC	*BIC*	*logLik*	*deviance*
5890	6002	−2928	5856

Random effects:

Groups	*Name*	*Variance*	*Std.Dev.*	*Corr*	
comm	(*Intercept*)	0.85896	0.92680		
	wealth4	0.18550	0.43070	−0.791	
	wealth5	0.34162	0.58448	−0.921	0.490

Number of obs: 5366, groups: comm, 361
Fixed effects:

	Estimate	Std.Error	zvalue	Pr(>|z|)	
(Intercept)	−1.6300095	0.1100587	−14.810	< 2e − 16 ***	
magec	0.0002154	0.0065755	0.033	0.973872	
magecsq	−0.0011135	0.0006865	−1.622	0.104806	
meduc2	0.5868783	0.0853450	6.877	6.13e − 12 ***	
meduc3	1.3885848	0.0980236	14.166	< 2e − 16 ***	—
wealth2	0.4452634	0.1077504	4.132	3.59e − 05 ***	
wealth3	0.6233047	0.1101954	5.656	1.55e − 08 ***	
wealth4	0.9128371	0.1166641	7.824	5.10e − 15 ***	
wealth5	1.1930787	0.1457757	8.184	2.74e − 16 ***	
urban	0.4950151	0.1402982	3.528	0.000418 ***	
wealth5mean	1.4832497	0.2644775	5.608	2.04e − 08 ***	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

	(Intr)	magec	magcsq	meduc2	meduc3	welth2	welth3	welth4	welth5
magec	0.023								
magecsq	−0.232	−0.484							
meduc2	−0.318	0.206	−0.044						
meduc3	−0.255	0.313	−0.069	0.556					
wealth2	−0.465	−0.009	−0.014	−0.098	−0.114				
wealth3	−0.457	−0.030	0.022	−0.123	−0.196	0.542			
wealth4	−0.487	−0.067	0.037	−0.171	−0.288	0.521	0.564		
wealth5	−0.323	−0.111	0.057	−0.156	−0.333	0.434	0.482	0.545	
urban	−0.120	0.014	0.0030.025	0.056	−0.005	−0.008	0.008	0.021	
wealth5mean	−0.183	0.006	−0.008	0.032	0.063	−0.023	−0.048	−0.088	−0.388

	urban
magec	
magecsq	
meduc2	
meduc3	
wealth2	
wealth3	
wealth4	
wealth5	
urban	
wealth5mean	−0.583

Συμπεραίνουμε λοιπόν ότι υπάρχει πράγματι, μια θετική συναφής επίδραση στο εισόδημα η οποία είναι μεγαλύτερη της θετικής επίδρασης του οικογενειακού εισοδήματος. Ο συντελεστής της $wealth5mean$ (η εκτιμήτρια της οποίας είναι 1.483) είναι η διαφορά log-odds της προγεννητικής φροντίδας για μια γυναίκα σε μια κοινότητα όπου όλες οι οικογένειες ανήκουν στο υψηλότερο πεμπτημόριο του εισοδήματος και για μια γυναίκα που ανήκει σε μια κοινότητα

όπου καμία από τις οικογένειες δεν βρίσκεται στο υψηλότερο πεμπτημόριο.

5.4.1 Επιδράσεις διασταυρούμενων επιπέδων

Το τρέχον μοντέλο θεωρεί ότι η συναφής επίδραση του εισοδήματος παραμένει η ίδια για όλες τις γυναίκες, άνεξάρτητα από το εισόδημά τους. Θα τροποποιήσουμε αυτή την υπόθεση για να επιτρέψουμε στην επίδραση του εισοδήματος μεταξύ των κοινοτήτων, οσον αφορά την πιθανότητα μιας γυναίκας να χρησιμοποιήσει υπηρεσίες προγεννητικής φροντίδας, στηριζόμενη στα προσωπικά της έσοδα. Αυτό γίνεται, συμπεριλαμβάνοντας στο μοντέλο την αλληλεπίδραση του ατομικού εισοδήματος (μέσω των ψευδομεταβλητών $wealth2$ και $wealth5$) και του εισοδήματος μεταξύ των κοινοτήτων ($wealth5mean$), δηλαδή μια αλληλεπίδραση διασταυρούμενων επιπέδων. Παράγουμε πρώτα τους τέσσερις όρους της αλληλεπίδρασης:

```
> mydata$wealth2Xwealth5mean <- mydata$wealth2*mydata$wealth5mean
> mydata$wealth3Xwealth5mean <- mydata$wealth3*mydata$wealth5mean
> mydata$wealth4Xwealth5mean <- mydata$wealth4*mydata$wealth5mean
> mydata$wealth5Xwealth5mean <- mydata$wealth5*mydata$wealth5mean
```

Τώρα προσαρμόζουμε το μοντέλο:

$$\begin{aligned}\log\frac{\pi_{ij}}{1-\pi_{ij}} = {} & \beta_0 + \beta_1 magec_{ij} + \beta_2 magecsq_{ij} + \beta_3 meduc2_{ij} + \beta_4 meduc3_{ij} \\ & + \beta_5 wealth2_{ij} + \beta_6 wealth3_{ij} + \beta_7 wealth4_{ij} + \beta_8 wealth5_{ij} \\ & + \beta_9 urban_j + \beta_{10} wealth5mean_j \\ & + \beta_{11} wealth2Xwealth5mean_{ij} + \beta_{12} wealth3Xwealth5mean_{ij} \\ & + \beta_{13} wealth4Xwealth5mean_{ij} + \beta_{14} wealth5Xwealth5mean_{ij} \\ & + u_{0j} + u_{7j} wealth4_{ij} + u_{8j} wealth5_{ij}\end{aligned}$$

```
> (fit3 <- glmer(antemed ~ magec + magecsq + meduc2 + meduc3 + urban +
wealth2 +
wealth3 + wealth4 + wealth5 + wealth5mean + wealth2Xwealth5mean +
wealth3Xwealth5mean + wealth4Xwealth5mean + wealth5Xwealth5mean + (1
+ wealth4 +
wealth5 | comm), family = binomial("logit"), data = mydata))
Generalized linear mixed model fit by the Laplace approximation
Formula: antemed ~ magec + magecsq + meduc2 + meduc3 + urban + wealth2
```

+
wealth3 + wealth4 + wealth5 + wealth5mean + wealth2Xwealth5mean +
wealth3Xwealth5mean + wealth4Xwealth5mean + wealth5Xwealth5mean +
(1 + wealth4 + wealth5 | comm)
Data: mydata

AIC	BIC	$logLik$	$deviance$
5890	6028	-2924	5848

Random effects:

$Groups$	$Name$	$Variance$	$Std.Dev.$	$Corr$	
$comm$	$(Intercept)$	0.85694	0.92571		
	$wealth4$	0.18580	0.43104	-0.813	
	$wealth5$	0.36346	0.60287	-0.937	0.559

Number of obs: 5366, groups: comm, 361
Fixed effects:

	$Estimate$	$Std.Error$	$zvalue$	$Pr(>\|z\|)$
$(Intercept)$	-1.7742963	0.1253138	-14.159	$< 2e-16 * **$
$magec$	-0.0003509	0.0065829	-0.053	0.957484
$magecsq$	-0.0010751	0.0006863	-1.567	0.117204
$meduc2$	0.5844430	0.0854692	6.838	$8.03e-12 * **$
$meduc3$	1.3807666	0.0979890	14.091	$< 2e-16 * **$
$urban$	0.4754199	0.1392196	3.415	$0.000638 * **$
$wealth2$	0.5523079	0.1307463	4.224	$2.40e-05 * **$
$wealth3$	0.6677683	0.1368024	4.881	$1.05e-06 * **$
$wealth4$	1.0840033	0.1467476	7.387	$1.50e-13 * **$
$wealth5$	1.5088227	0.2039821	7.397	$1.39e-13 * **$
$wealth5mean$	2.9861124	0.6769591	4.411	$1.03e-05 * **$
$wealth2Xwealth5mean$	-1.2310274	0.7681490	-1.603	0.109025
$wealth3Xwealth5mean$	-0.6954785	0.7562838	-0.920	0.357782
$wealth4Xwealth5mean$	-1.6015881	0.7212115	-2.221	$0.026372*$
$wealth5Xwealth5mean$	-1.8008560	0.7173480	-2.510	$0.012058*$

—
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Correlation of Fixed Effects:

	(Intr)	magec	magcsq	meduc2	meduc3	urban	welth2	welth3	welth4
welth5	wlth5m	wlt2X5	wlt3X5	wlt4X5					
magec	0.042								
magecsq	−0.215	−0.485							
meduc2	−0.268	0.207	−0.044						
meduc3	−0.205	0.313	−0.068	0.557					
urban	−0.085	0.014	0.003	0.025	0.055				
wealth2	−0.514	−0.028	−0.002	−0.094	−0.113	−0.002			
wealth3	−0.503	−0.050	0.023	−0.124	−0.187	−0.007	0.530		
wealth4	−0.559	−0.079	0.040	−0.152	−0.260	−0.002	0.499	0.529	
wealth5	−0.467	−0.093	0.053	−0.111	−0.233	−0.030	0.364	0.385	0.430
wealth5mean	−0.495	−0.043	0.016	−0.012	−0.016	−0.256	0.354	0.359	0.391
0.360									
wlth2Xwlth5	0.322	0.040	−0.017	0.024	0.037	−0.002	−0.561	−0.308	−0.290
−0.218	−0.676								
wlth3Xwlth5	0.335	0.048	−0.010	0.042	0.053	0.005	−0.324	−0.582	−0.316
−0.238	−0.709	0.604							
wlth4Xwlth5	0.415	0.049	−0.019	0.029	0.055	0.018	−0.338	−0.350	−0.585
−0.306	−0.821	0.634	0.668						
wlth5Xwlth5	0.466	0.042	−0.021	0.018	0.029	0.046	−0.338	−0.346	−0.373
−0.608	−0.896	0.643	0.675	0.780					

Συγκρίνοντας τις εκτίμητριες των συντελεστών κάθε αλληλεπίδρασης με το τυπικό τους σφάλμα βρίσκουμε ότι μόνο οι μεταβλητές $wealth4Xwealth5mean$ και $wealth5Xwealth5mean$ είναι στατιστικά σημαντικές σε επίπεδο 5%. Για να ερμηνεύσουμε την επίδραση της αλληλεπίδρασης, θα υπολογίσουμε τις πιθανότητες πρόβλεψης λήψης προγεννητικής φροντίδας για τους διαφορετικούς συνδυασμούς των $wealth$ και $wealth5mean$, διατηρώντας τις $mage$, $meduc$ και $urban$ σταθερές. Θεωρούμε τις αναλογίες της κοινοτήτων στο υψηλότερο πεμπτημόριο του 0, 0.2 και 0.4.

Δημιουργούμε πρώτα ένα νέο πλαίσιο δεδομένων, $mydatapred$, και θέτουμε $mage$, $meduc$ και $urban$ μεταβλητές ίσες με τις μέσες τιμές τους, ενώ θέτουμε την $magecsq$ μεταβλητή ίση με τετράγωνο του δειγματικού μέσου της $magec$:

```
> mydatapred <- mydata
> mydatapred$magec<-0
> mydatapred$magecsq<-0
> mydatapred$meduc2<-mean(mydatapred$meduc2)
> mydatapred$meduc3<-mean(mydatapred$meduc3)
> mydatapred$urban<-mean(mydatapred$urban)
```

Στη συνέχεια επανακωδικοποιούμε τη wealth5mean για να πάρουμε μία από τις 3 αναλογίες που μας ενδιαφέρουν: 0, 0.2, ή 0.4. Ένας τρόπος για να το κάνουμε αυτό είναι να επανακωδικοποιήσουμε όλες τις τιμές του $wealth5mean$ στο εύρος από 0 μέχρι 0.1 για την πρώτη τιμή, όλες τις τιμές του $wealth5mean$ στο εύρος από 0.1 μέχρι 0.3 για τη δεύτερη και όλες τις τιμές του $wealth5mean$ στο εύρος από 0.3 μέχρι 1 στην τελευταία τιμή:

```
> mydatapred$wealth5mean[mydatapred$wealth5mean
>= 0 & mydatapred$wealth5mean <0.1]<-0
> mydatapred$wealth5mean[mydatapred$wealth5mean
>= 0.1 & mydatapred$wealth5mean <0.3]<-0.2
> mydatapred$wealth5mean[mydatapred$wealth5mean
>= 0.3 & mydatapred$wealth5mean <=1]<-0.4
```

Πρέπει επίσης να αλλάξουμε τις τιμές των τεσσάρων όρων των αλληλεπιδράσεων για να αντικατοπτρίσουμε τις αλλαγές που έχουμε ήδη κάνει στη μεταβλητή $wealth5mean$:

```
> mydatapred$wealth2Xwealth5mean<-mydatapred$wealth2*mydatapred$wealth5mean
> mydatapred$wealth3Xwealth5mean<-mydatapred$wealth3*mydatapred$wealth5mean
> mydatapred$wealth4Xwealth5mean<-mydatapred$wealth4*mydatapred$wealth5mean
> mydatapred$wealth5Xwealth5mean<-mydatapred$wealth5*mydatapred$wealth5mean
```

Τώρα που έχουμε θέσει όλες τις επεξηγηματικές μεταβλητές στις προδιαγεγραμμένες τους τιμές, μπορούμε να υπολογίσουμε τη μέση (δηλαδή συγκεκριμένη ανά ομάδα) πιθανότητα πρόβλεψης για κάθε γυναίκα.

```
> X <- model.matrix(terms(fit3), mydatapred)
> b <- fixef(fit3)
> predlogit <- X %*% b
> predprob <- logit(predlogit, inverse = TRUE)
```

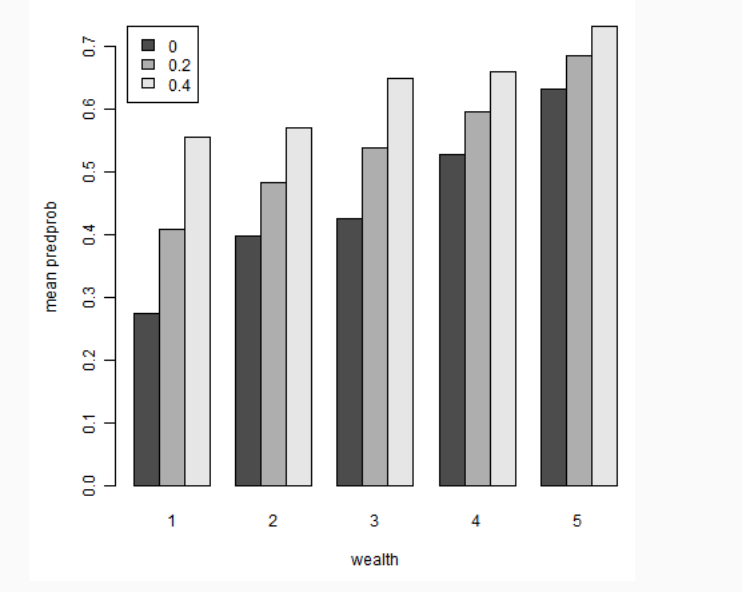
Έπειτα συνδυάζουμε τις πιθανότητες πρόβλεψης και τις τιμές του $wealth5mean$ στο $mydatapred2$ πλαίσιο δεδομένων, για να υπολογίσουμε τη μέση πιθανότητα για κάθε $wealth5mean$ τιμή χρησιμοποιώντας την εντολή $tapply$:

```
> mydatapred2 <- unique(data.frame(cbind(predprob = predprob, wealth5mean
= mydatapred$wealth5mean,wealth=mydatapred$wealth)))
```

```
> colnames(mydatapred2)[1] <- c("predprob")
> mean.table <- tapply(mydatapred2$predprob,list(mydatapred2$wealth5mean,
mydatapred2$wealth),mean)
```

Τώρα θα κάνουμε το γράφημα των πιθανοτήτων προβλέψης χρησιμοποιώντας την εντολή *barplot*. Αυτό δίνει συνολικά 15 κάθετες ράβδους:

```
> barplot(mean.table, ylab = "mean predprob", xlab = "wealth", beside = TRUE,
axes = TRUE, legend = TRUE,
args.legend = list(x = "topleft", inset = c(0.025,
0)))
```

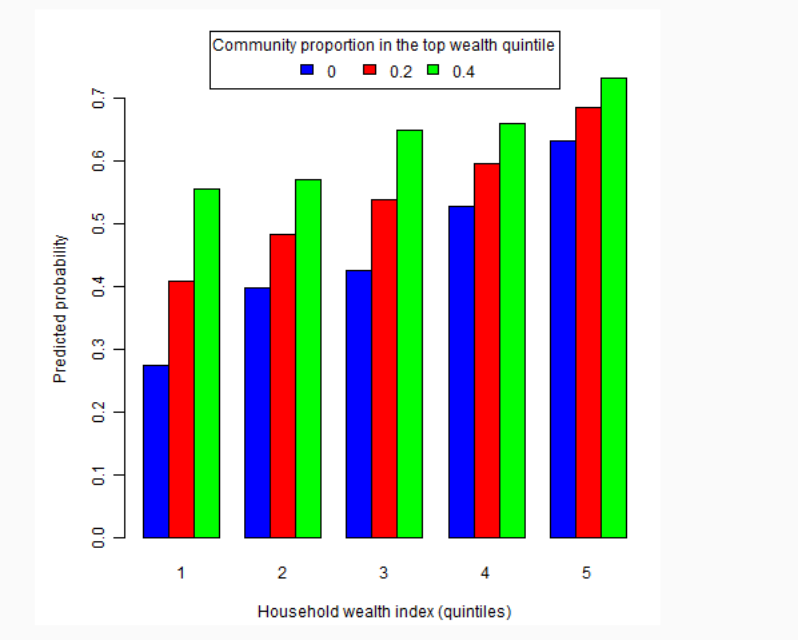


Σχήμα 5.8

Μια τελική και φυσικά πιο βελτιομένη εκδοχή του παραπάνω γραφήματος είναι η ακόλουθη:

```
> barplot(mean.table, ylab = "Predicted probability", xlab = "Household wealth
index (quintiles)", beside = TRUE, axes = TRUE, col = c("blue", "red", "green"),
legend = TRUE, args.legend = list(x = "top", inset = c(0, - 0.1), title =
"Community proportion in the top wealth quintile", horiz = TRUE))
```

Από το γράφημα λοιπόν μπορούμε να συμπεράνουμε ότι η συναφης επίδραση



Σχήμα 5.9

του εισοδήματος (δηλαδή η διαφορά των πιθανοτήτων πρόβλεψης για τις τιμές της μεταβλητής $wealth5mean$, 0, 0.2 και 0.4) είναι ασθενέστερη μεταξύ των γυναίκων στο υψηλότερο 40% της κατανομής του εισοδήματος ($wealth4$ και $wealth5$). Η ζωή σε μια κοινότητα στερήσεων (όπως υποδηλώνει η χαμηλή τιμή της μεταβλητής $wealth5mean$) σε αντίθεση με μια κοινότητα σε καλύτερη κατάσταση αποτελεί μεγαλύτερο εμπόδιο στη χρήση υπηρεσιών προγεννητικής φροντίδας από φτωχότερες γυναίκες. Εναλλακτικά, αλλά ισοδύναμα, μπορούμε να πούμε ότι η επίδραση του ατομικού εισοδήματος είναι ισχυρότερη σε φτωχότερες κοινότητες.

Βιβλιογραφία

[1] Marc A. Scott, Jeffrey S. Simanoff, Brian D. Marx *The SAGE Handbook of Multilevel Modelling*, SAGE Publication Ltd, 2013.

[2] Andrew Gelman, Jennifer Hill, E. Stein, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, 2007.

[3] Π. Οικονόμου, Χ. Καρώνη, *Στατιστικά Μοντέλα Παλινδρόμησης*, Αθήνα, 2010.

[4] Annette J. Dobson, *An Introduction to Generalized Linear Models*, Chapman & Hall/CRC, 2002.

[5] Andrew Gelman, *Multilevel (Hierarchical) Modeling: What it Can and Cannot Do*, June 1, 2005

[6] Sander Greenland, *Principles of multilevel modelling*, International Journal of Epidimiology 2000

[7] http://www.cmm.bris.ac.uk/lemma/