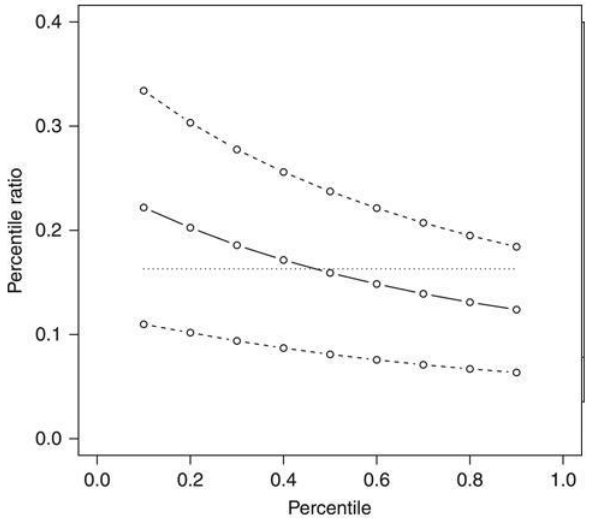
Τζουμέρκας Γιώργος

Μέθοδοι μετα-ανάλυσης και λόγος των ποσοστημορίων στη Βιοστατιστική

Διπλωματική Εργασία στο ΜΠΣ
της Στατιστικής και Επιχειρησιακής Έρευνας



Πανεπιστήμιο Αθηνών
Τμήμα Μαθηματικών
Αθήνα 2014

Η παρούσα Διπλωματική Εργασία εκπονήθηκε
στα πλαίσια των σπουδών για την απόκτηση του
Μεταπτυχιακού Διπλώματος Ειδίκευσης στην
Στατιστική και Επιχειρησιακή Έρευνα
που απονέμει το Τμήμα Μαθηματικών του
Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών

Εγκρίθηκε στις ..... από Εξεταστική Επιτροπή αποτελούμενη από τους:

Ονοματεπώνυμο	**Βαθμίδα**	**Υπογραφή**
Λουκία Μελιγκοτσίδου	Επίκουρη Καθηγήτρια	.....
Απόστολος Μπουρνέτας	Καθηγητής	.....
Φώτης Σιάννης (επιβλέπων)	Επίκουρος Καθηγητής	.....

Εισαγωγή

Η μετα-ανάλυση δεδομένων από πολλαπλές στατιστικές έρευνες πάνω στο ίδιο ερώτημα έχει τύχει μεγάλης χρήσης στη βιοστατιστική τα τελευταία χρόνια. Το πιο διαδεδομένο εργαλείο μοντελοποίησης της κάθε έρευνας ξεχωριστά είναι τα αναλογικά μοντέλα, όπου θεωρούμε πως οι καμπύλες επιβίωσης των ομάδων προς έλεγχο είναι ανάλογες η μια προς την άλλη μέσα στο χρόνο. Σε μια μετα-ανάλυση κάτω από αυτή την υπόθεση για κάθε έρευνα, χρησιμοποιείται ο λόγος των συναρτήσεων κινδύνου των δύο ομάδων από κάθε έρευνα ώστε να παράγουμε τα εκ νέου συμπεράσματα. Παρόλα αυτά σε μια μετα-ανάλυση όπου συμπεριλαμβάνουμε ένα πλήθος ερευνών, σε κάποιες από αυτές ενδέχεται η υπόθεση της αναλογικότητας να μην ευσταθεί και ο λόγος του κινδύνου μεταξύ των δύο ομάδων να εξαρτάται από τον χρόνο διάρκειας της έρευνας. Ένα νέο εναλλακτικό μέτρο για την μετα-ανάλυση μπορεί να χρησιμοποιηθεί όταν η υπόθεση των αναλογικών κινδύνων δεν είναι η κατάλληλη, αυτό είναι ο λόγος των ποσοστημορίων.

Τα ποσοστημόρια μιας συνάρτησης κατανομής είναι συσχετισμένα το ένα με το άλλο, άρα και ο λόγος των ποσοστημορίων δύο ομάδων από ποσοστημόριο σε ποσοστημόριο είναι συσχετισμένος και αυτός με την σειρά του. Θα προσπαθήσουμε να εντάξουμε αυτήν τη συσχέτιση σε μια μετα-ανάλυση ώστε αυτό να μπορέσει να αυξήσει την ακρίβεια της εκτίμησης και μάλιστα κάτω από μια μοντελοποίηση στην μορφή των γενικευμένων γραμμικών μοντέλων και πιο συγκεκριμένα με χρήση των τακτικών που χρησιμοποιούνται στην ανάλυση διαχρονικών δεδομένων, όπου έχουμε συσχετισμένα δεδομένα.

Στην παρούσα διπλωματική εργασία θα υπενθυμίσουμε στα τρία πρώτα κεφάλαια την κύρια θεωρία πίσω από τους τομείς της Μετα-ανάλυσης [4], της Ανάλυσης Διαχρονικών Δεδομένων [10], αφού πρώτα γίνει μια μικρή υπενθύμιση των γενικευμένων γραμμικών μοντέλων [8], και της Ανάλυσης Επιβίωσης [12], για περαιτέρω εμβάνθυνση προτείνονται οι αντίστοιχες παραθέσεις στην βιβλιογραφία, ενώ για πιο εισαγωγικές έννοιες πάνω στις πιθανότητες και τη στατιστική προτείνονται τα βιβλία [1],[2],[3]. Στο τέταρτο κεφάλαιο θα μιλήσουμε εκτενέστερα για το λόγο των ποσοστημορίων, τους τρόπους με τους οποίους μπορεί να χρησιμοποιηθεί σε μια μετα-ανάλυση και θα

εισάγουμε το πως μπορεί να μοντελοποιηθεί μια μετα-άναλυση με λόγο των ποσοστημορίων με χρήση τακτικών που χρησιμοποιούνται στην ανάλυση των διαχρονικών δεδομένων.

Περιεχόμενα

Κεφάλαιο 1

Μετα-ανάλυση

Μετα-ανάλυση είναι ο κλάδος της στατιστικής ο οποίος ασχολείται με τη συλλογή και την επεξεργασία ήδη υπάρχουσων στατιστικών μελετών για κάποιο θέμα, με σκοπό την δημιούργια μιας εκ νέου ενιαίας στατιστικής μελέτης η οποία λόγω του μεγαλύτερου δείγματος που θα έχει και μέσω των επιπλέον παραμέτρων που γίνεται να προσαρμοστούν σε αυτή, θα δώσει πιο ακριβή και ενδεχομένως περισσότερα συμπεράσματα.

1.1 Μέγεθος της επίδρασης (Effect size)

Το μέγεθος της επίδρασης (effect size) είναι ουσιαστικά η μονάδα μέτρησης στην μετα-ανάλυση. Είναι σαν παράδειγμα αυτό που μας δείχνει το αντίκτυπο μιας θεραπείας ή γενικότερα το μέγεθος μιας σχέσης μεταξύ δύο μεταβλητών. Το υπολογίζουμε για κάθε μια από τις ξεχωριστές μελέτες μας και ύστερα δουλεύουμε με όλα τα μεγέθη της επίδρασης για να εξετάσουμε την συνέπεια μεταξύ όλων των μελετών και για να καταλήξουμε σε ένα συγκεντρωτικό μέγεθος της επίδρασης, της μεταβλητής που ενδιαφερόμαστε το summary effect.

Η επιλογή του τι είδος effect size θα χρησιμοποιήσουμε γίνεται με βάση το πόσο τηρούνται τα εξής κριτήρια:

1. να μπορούν να είναι συγκρίσιμα από την μια μελέτη στην άλλη, δηλαδή κοινώς να μετράνε το ίδιο μέγεθος,

2. να μπορούν να υπολογιστούν από την υπάρχουσα δημοσιευμένη πληροφορία,

3. να έχουν καλά τεχνικά χαρακτηριστικά (π.χ. γνωστή κατανομή),

4. να έχουν λογικό νόημα.

Σε κάθε μετα-ανάλυση εμφανίζονται γύρω στις 2-3 επιλογές ανάλογα με το τι πληροφορία μας παρέχεται από τις αρχικές μελέτες. Συνήθως αν στις αρχικές μελέτες έχουμε μέσους (means) και τυπικές αποκλίσεις (standar deviations) μεταξύ δύο ομάδων τότε μπορούμε να χρησιμοποιήσουμε είτε την καθαρή (raw) διαφορά μεταξύ των δύο μέσων των δύο ομάδων, είτε την τυποποιημένη (standardized) διαφορά μεταξύ των μέσων, είτε των λόγο των μέσων των δυο ομάδων (response ratio). Όταν έχουμε δεδομένα 0 - 1 (γεγονότος - μη γεγονότος) χρησιμοποιούμε είτε τον λόγο του ρίσκου (risc ratio), είτε τον λόγο των προγνωστικών (odds ratio), είτε την διαφορά του ρίσκου (risc difference). Στην περίπτωση που οι αρχικές μελέτες μας δίνουν την συσχέτιση δύο μεταβλητών, τότε η συσχέτιση μπορεί να χρησιμοποιηθεί σαν effect size. Για τον συμβολισμό χρησιμοποιούμε τα γράμματα θ για το effect size και το y για τον δειγματικό εκτιμητή του.

1.1.1 Καθαρή διαφορά των μέσων (Raw mean difference)

Σε περιπτώσεις όπου τα αποτελέσματα των δοθέντων μελετών είναι σε ένα λογικά εξηγήσιμο και όμοιας μονάδας μέτρησης μέγεθος η μετα-ανάλυση μπορεί να γίνει απευθείας από την καθαρή διαφορά των μέσων.

Ας θεωρήσουμε την περίπτωση όπου έχουμε δύο είδη θεραπείας την συνήθη και την καινούρια και θέλουμε να ελέγξουμε την διαφορά των μέσων των δύο αυτών ομάδων. Αν συμβολίσουμε με μ_1, μ_2 τους πραγματικούς μέσους του πληθυσμού των δύο ομάδων αυτό που ζητάμε είναι το:

$$\Delta = \mu_1 - \mu_2 \tag{1.1}$$

Στην περίπτωση που έχουμε δύο ανεξάρτητες ομάδες.
Έστω ότι οι δειγματικοί μέσοι της πρώτης ομάδας μεγέθους n_1 και της δεύτερης μεγέθους n_2 είναι αντίστοιχα οι $\overline{X}_1, \overline{X}_2$, με αντίστοιχη δειγματική διασπορά S_1^2, S_2^2,

τότε έχουμε την εκτίμηση:

$$\hat{\Delta} = D = \overline{X}_1 - \overline{X}_2 \tag{1.2}$$

όπου η εκτίμηση για την διασπορά της δίνεται από την σχέση

$$V_D = \hat{V}(\overline{X}_1 - \overline{X}_2) = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \tag{1.3}$$

ενώ στην συνηθισμένη περίπτωση όπου θεωρούμε ότι οι δύο ομάδες έχουν ίδια διασπορά έχουμε:

$$V_D = \frac{n_1 + n_2}{n_1 n_2} S_{pooled} \tag{1.4}$$

όπου

$$S_{pooled} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \tag{1.5}$$

και προφανώς και για τις δύο περιπτώσεις ισχύει για το τυπικό σφάλμα

$$SE_D = \sqrt{V_D}. \tag{1.6}$$

Στην περίπτωση όπου οι δύο ομάδες έχουν εξάρτηση μεταξύ τους. Στην περίπτωση αυτή όπου μεταξύ των ομάδων του δείγματος υπάρχει κάποια μορφή σχέσης (π.χ. συγγενείς, ασθενείς στο ίδιο στάδιο μιας ασθένειας), τα αποτελέσματα επηρεάζονται από την συσχέτιση μεταξύ των μελών των ομάδων και το κάθε μέλος της ομάδας τοποθετείται σε διαφορετική υποομάδα ώστε να μειωθεί το σφάλμα και να αυξηθεί η στατιστική ισχύς. Οπότε αν υπολογίσουμε τον μέσο για κάθε ζευγάρι μπορούμε να έχουμε τον δειγματικό μέσο X_{diff} και την αντίστοιχη τυπική απόκλιση S_{diff} τους. Τότε για n υποομάδες έχουμε:

$$\hat{\Delta} = D = \overline{X}_{diff}, \tag{1.7}$$

$$V_D = \frac{S_{diff}^2}{n}. \tag{1.8}$$

Εναλλακτικά, αν έχουμε τους μέσους για κάθε ομάδα $(\overline{X}_1, \overline{X}_2)$ τότε παίρνουμε:

$$\hat{\Delta} = D = \overline{X}_1 - \overline{X}_2, \tag{1.9}$$

$$V_D = \frac{S_{diff}}{n}, \tag{1.10}$$

με το S_{diff} να δίνεται από τον τύπο

$$S_{diff} = (S_1^2 + S_2^2 - 2r \cdot S_1 S_2)^{\frac{1}{2}}, \tag{1.11}$$

όπου S_1, S_2 είναι οι τυπικές αποκλίσεις των διαφορετικών ομάδων και r είναι ο συντελεστής συσχέτισης των δύο ομάδων. Παρατηρούμε ότι όσο το $r \to 1$ το τυπικό σφάλμα μειώνεται.
Αν υποθέσουμε ότι $S_1 = S_2$,τότε:

$$S_{diff} = \sqrt{2S_{pooled}^2(1-r)}, \tag{1.12}$$

και προφανώς για όλα τα παραπάνω ισχύει

$$SE_D = \sqrt{V_D}. \tag{1.13}$$

1.1.2 Τυποποιημένη διαφορά των μέσων (Standardized mean difference)

Σε περιπτώσεις όπου τα αποτελέσματα των δοθέντων μελετών είναι σε διαφορετική μονάδα μέτρησης σε κάθε μελέτη, για να μπορέσουμε να έχουμε μια κοινή μονάδα μέτρησης τις τυποποιούμε διαρώντας σε κάθε μελέτη την διαφορά των μέσων των ομάδων με την αντίστοιχη τους τυπική απόκλιση. Κάνοντας τη συνηθισμένη υπόθεση ότι οι δύο πληθυσμοί έχουν την ίδια τυπική απόκλιση ($\sigma_1 = \sigma_2 = \sigma$) έχουμε ότι η πραγματική τυποποιημένη διαφορά μέσων δίνεται από την σχέση:

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}. \tag{1.14}$$

Στην περίπτωση που έχουμε δύο ανεξάρτητες ομαδες.
Έχουμε:

$$\hat{\delta} = d = \frac{\overline{X_1} - \overline{X_2}}{S_{within}}, \tag{1.15}$$

όπου

$$S_{within} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}. \tag{1.16}$$

Παρατηρούμε ότι εκτιμούμε και το S_1 και το S_2 γιατί οι δειγματικές εκτιμήσεις τους είναι σχεδόν απίθανο να είναι οι ίδιες. Ακόμα μπορεί να δειχτεί ότι μια πολύ καλή προσέγγιση της διασποράς του d δίνεται από την σχέση:

$$V_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}. \tag{1.17}$$

Σε αυτή την εξίσωση ο πρώτος όρος στα δεξιά της ισότητας αντικατοπτρίζει πρώτον στον αριθμητή την αβεβαιότητα της εκτίμησης του μέσου από την σχέση (1.15) και δεύτερο την αβεβαιότητα της εκτίμησης του S_{within}.
Έχει παρατηρηθεί ότι το d τείνει να υπερεκτιμήσει το δ και για να διορθωθεί το σφάλμα εκτίμηση ($bias$) χρησιμοποιούμε την διόρθωση $J.Hedges$ [5]:

$$J = 1 - \frac{3}{4df - 1}, \tag{1.18}$$

όπου df οι βαθμοί ελευθερίας που χρησιμοποιούνται για τον υπολογισμό του S_{within}, στην περίπτωση που αναφερόμαστε έχουμε $df = n_1 + n_2 - 2$ και έτσι:

$$\hat{\delta} = g = J \cdot d, \tag{1.19}$$

και μέσω της μεθόδου Δέλτα [1] (σελ. 302) έχουμε ότι η διορθωμένη διασπορά ισούται με

$$V_g = J^2 \cdot V_d. \tag{1.20}$$

Στην περίπτωση όπου οι ομάδες έχουν εξάρτηση μεταξύ τους.
Δουλεύοντας όπως στην αντίστοιχη περίπτωση προηγουμένως έχουμε:

$$d = \frac{\overline{Y}_{diff}}{S_{within}} = \frac{\overline{Y_1} - \overline{Y_2}}{S_{within}}, \tag{1.21}$$

όπου

$$S_{within} = \frac{S_{diff}}{\sqrt{2(1-r)}}, \tag{1.22}$$

και

$$V_d = \left(\frac{1}{n} + \frac{d^2}{2n}\right) \cdot 2 \cdot (1-r), \tag{1.23}$$

σε αυτήν την περίπτωση οι βαθμοί ελευθερίας είναι $df = n - 1$.

1.1.3 Λόγος των μέσων (Response ratio)

Σε μελέτες που μας δίνουν σαν αποτέλεσμα φυσικές και λογικά αντιλήψιμες κλίμακες (π.χ. μήκος, μάζα, όγκο) οι οποίες σχεδόν ποτέ δεν παίρνουν την τιμή 0, συνηθίζεται να χρησιμοποιούμε το λόγο των μέσων. Είναι σημαντικό να κατανοήσουμε ότι ο λόγος των μέσων έχει νόημα όταν τα αποτελέσματα που μας δίνονται είναι σε πραγματική κλίμακα. Παραδείγματος χάρην όταν έχουμε σαν αποτελέσματα βαθμολογίες διαγωνισμάτων, μεγέθη μέτρησης συμπεριφοράς κλπ δεν μπορούμε να τον χρησιμοποιήσουμε. Οι πράξεις γίνονται στο λογαριθμικό επίπεδο και έπειτα τις επαναφέρουμε ξανά στην φυσική τους κλίμακα ώστε να βγάλουμε τα συμπεράσματά μας. Έτσι έχουμε:

$$R = \frac{\overline{X}_1}{\overline{X}_2}, \tag{1.24}$$

$$lnR = ln(\overline{X}_1) - ln(\overline{X}_2), \tag{1.25}$$

$$V_{lnR} = S_{pooled} \cdot \left(\frac{1}{n_1(\overline{X}_1)^2} + \frac{1}{n_2(\overline{X}_2)^2}\right), \tag{1.26}$$

όπου το S_{pooled} υπολογίζεται όπως στην σχέση (1.5).

Η επαναφορά στην φυσική κλίμακα γίνεται ως εξής:

$$R = exp(lnR),$$

και αν τα άκρα του διαστήματος εμπιστοσύνης για τον λογάριθμο του R τα συμβολίσουμε με LL_{lnR}, UL_{lnR}, έχουμε για τα αντίστοιχα άκρα του R

$$LL_R = exp(LL_{lnR}),\ UL_R = exp(UL_{lnR}).$$

Μελέτες με διακριτά δίτιμα αποτελέσματα.
Σε μελέτες που βασίζονται σε δεδομένα 0 - 1, γεγονότος - μη γεγονότος για δυο ομάδες, όπου τα αποτελέσματα δίνονται σαν ο αριθμός των γεγονότων για την καθέμια από τις δύο ομάδες μπορούμε να χρησιμοποιήσουμε για effect size κάποιο από τις τρεις κατηγορίες που θα περιγραφούν παρακάτω. Θα συμβολίσουμε με A το πλήθος των ατόμων της πρώτης ομάδας που τους συνέβει το γεγονός που εξετάζουμε και B το πλήθος των ατόμων της πρώτης ομάδας που δεν τους συνέβει το γεγονός σε σύνολο n_1 ατόμων, ενώ για τη δεύτερη ομάδα αντίστοιχα θα συμβολίσουμε τις όμοιες καταστάσεις με C, D, n_2.

	$Events$	Non_Events	N
$Group\_1$	A	B	n_1
$Group\_2$	C	D	n_2

1.1.4 Λόγος του ρίσκου (Risc ratio)

Ο λόγος του ρίσκου είναι ο λόγος του ποσοστού από την πρώτη ομάδα που συνέβει το γεγονός προς το ποσοστό της δεύτερης ομάδας που συνέβει το γεγονός και εδώ οι πράξεις γίνονται στο λογαριθμικό επίπεδο και έπειτα επαναφέρουμε τα αποτελέσματα στην μορφή που τα θέλουμε με την αντίστροφη συνάρτηση. Έχουμε:

$$RiskRatio = \frac{A/n_1}{C/n_2}, \tag{1.27}$$

$$LogRiskRatio = ln(RiskRatio), \tag{1.28}$$

$$V_{LogRiskRatio} = \frac{1}{A} - \frac{1}{n_1} + \frac{1}{C} - \frac{1}{n_2}. \tag{1.29}$$

1.1.5 Λόγος των προγνωστικών (Odds ratio)

Ο λόγος των προγνωστικών είναι ο λόγος δύο προγνωστικών (*odds*). Αν και το αποτέλεσμα που μας δίνει δεν είναι το ίδιο νοητικά κατανοητό με τον λόγο του ρίσκου, έχει στατιστικές ιδιότητες που το κάνουν καλύτερη επιλογή. Έχουμε:

$$OddsRatio = \frac{A \cdot D}{B \cdot C}, \tag{1.30}$$

$$LogOddsRatio = ln(OddsRatio), \tag{1.31}$$

$$V_{LogOddsRatio} = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}, \tag{1.32}$$

και εδώ τα αποτελέσματα τα φέρνουμε στην μορφή που θέλουμε με την αντίστροφη της λογαριθμικής συνάρτησης.

1.1.6 Διαφορά του ρίσκου (Risc Difference)

Η διαφορά του ρίσκου είναι η διαφορά δύο λόγων του ρίσκου, εδώ δεν χρειάζεται η μετατροπή στην λογαριθμική κλίμακα .Έχουμε:

$$RiskDiff = \frac{A}{n_1} - \frac{C}{n_2}, \tag{1.33}$$

$$V_{RiskDiff} = \frac{A \cdot B}{n_1^3} + \frac{C \cdot D}{n_2^3}. \tag{1.34}$$

1.1.7 Μέγεθος της επίδρασης με βάση την συσχέτιση

Σε μελέτες όπου μας δίνεται ο συντελεστής συσχέτισης μεταξύ δύο συνεχών μεταβλητών μπορούμε να τον χρησιμοποιήσουμε σαν effect size. Η συσχέτιση είναι μια έννοια κατανοητή και μπορούμε να βγάλουμε συμπεράσματα μέσα από αυτήν. Έχουμε:

$$V_r = \frac{(1-r^2)^2}{n-1}, \tag{1.35}$$

όπου n είναι το μέγεθος του δείγματος.
Συνήθως χρησιμοποιείται ο μετασχηματισμός Fischer's z-scale [6]:

$$z = 0.5 \cdot ln\left(\frac{1+r}{1-r}\right), \tag{1.36}$$

όπου έτσι προσεγγιστικά έχουμε ότι:

$$V_z = \frac{1}{n-3}. \tag{1.37}$$

Για να φέρουμε πίσω στις αρχικές τιμές χρησιμοποιούμε την σχέση:

$$r = \frac{e^{2z}-1}{e^{2z}+1}. \tag{1.38}$$

1.1.8 Μετρατροπή από ένα effect size σε άλλο

Στην περίπτωση όπου σε διαφορετικές μελέτες της μετα-ανάλυσης μας, παρέχονται διαφορετικού είδους effect size πρέπει να τα μετατρέψουμε όλα ώστε όλες οι μελέτες να έχουν το ίδιο effect size. Αυτό που πρέπει να αναφερθεί είναι ότι αυτές οι μετατροπές γίνονται κάτω από υποθέσεις για την φύση των επιδράσεων, οπότε κι ας θεωρήσουμε πως αυτές οι υποθέσεις δεν είναι ακριβώς σωστές και ενδέχεται να επηρεάσουν την πρόβλεψή μας, συνήθως αυτό είναι προτιμότερο από την εναλλακτική του να αφαιρέσουμε όσες μελέτες δεν μας παρέχουν το μέγεθος της επίδρασης που χρησιμοποιούμε, αφού αυτό θα οδηγήσει στο χάσιμο πληροφορίας και το συστηματικό χάσιμο την πληροφορίας θα μας οδηγήσει σε σφάλματα εκτίμησης. Παρακάτω

αναφέρονται οι μαθηματικές σχέσεις με τις οποίες μπορούν να γίνουν αυτές οι μετατροπές, αναλόγα την κάθε περίπτωση.

Μετατροπή από LogOdds Ratio σε standardized mean difference
Μπορούμε να μετατρέψουμε ένα μέγεθος της επίδρασης από τον λόγο των προγνωστικών σε τυποποιημένη διαφορά των μέσων, μέσω της σχέσης η οποία αρχικά προτάθηκε από τους Hasselblad και Hedges [7]

$$d = LogOddsRatio \cdot \frac{\sqrt{3}}{\pi}$$

Για την διασπορά θα έχουμε μέσω της μεθόδου Δέλτα

$$V_d = V_{LogOddsRatio} \cdot \frac{3}{\pi^2}.$$

Μετατροπή από standardized mean difference σε LogOdds Ratio
Για την αντίστροφη μετατροπή από τυποποιημένη διαφορά των μέσων σε λόγο των προγνωστικών, χρησιμοποιούμε τις αντίστοιχες σχέσεις

$$LogOddsRatio = d \cdot \frac{\pi}{\sqrt{3}}$$

$$V_{logOddsRatio} = V_d \cdot \frac{\pi^2}{3}$$

Μετατροπή από συντελεστή συσχέτισης σε standardized mean difference
Για την μετατροπή από την συσχέτιση r στην τυποποιημένη διαφορά των μέσων d χρησιμοποιούμε την σχέση

$$d = \frac{2r}{\sqrt{1 - r^2}}$$

και εδώ με μέθοδο Δέλτα η αντίστοιχη διασπορά μας δίνεται από την σχέση

$$V_d = \frac{4V_r}{(1-r^2)^3}.$$

Χρησιμοποιώντας αυτήν την μετατροπή υποθέτουμε ότι τα συνεχή δεδομένα που χρησιμοποιήθηκαν για να εκτιμηθεί το r ακολουθούν διδιάστατη κανονική κατανομή και οι δύο ομάδες δημιουργούνται διχοτομώντας την μια από τις δύο ομάδες.

Μετατροπή από standardized mean difference σε συντελεστή συσχέτισης

Για την αντίστροφη μετατροπή της προηγούμενης έχουμε ότι

$$r = \frac{d}{\sqrt{d^2 + a}}$$

όπου το a είναι μια σταθερά διόρθωσης:

$$a = \frac{(n_1 + n_2)^2}{n_1 n_2}$$

για $n_1 \neq n_2$. Η σταθερά διόρθωσης εξαρτάται από το λόγο του n_1 προς το n_2, παρά από τις απόλυτες τιμές αυτών. Για αυτό το λόγο αν n_1, n_2 είναι άγνωστα θέτουμε $n_1 = n_2$ και τότε $a = 4$. Για την διασπορά έχουμε:

$$V_r = \frac{a^2 \cdot V_d}{(d^2 + a)^3}.$$

1.2 Ακρίβεια (Precision)

Το μέγεθος της επίδρασης σε κάθε μία από τις μελέτες της μετα-ανάλυσης *φέρνει μαζί* της και ένα διάστημα εμπιστοσύνης, το οποίο αντικατοπτρίζει την ακρίβεια με την οποία έχει εκτιμηθεί στην αντίστοιχη μελέτη του. Όσο μικρότερο διάστημα εμπιστοσύνης έχει ένα effect size, τόσο μεγαλύτερη θεωρείται η ακρίβειά του. Το διάστημα εμπιστοσύνης εξαρτάται από το τυπικό σφάλμα και σε ένα κλασσικό παράδειγμα όπου

θεωρούμε πως το effect size (έστω Y) κατανέμεται κανονικά δίνεται από τη σχέση $[\overline{Y} - z_{a/2}SE_Y, \overline{Y} + z_{a/2}SE_Y]$, όπου a είναι το επίπεδο στατιστικής σημαντικότητας και $z_{a/2}$ η συνάρτηση κατανομής της τυποποιημένης κανονικής. Κύριοι παράγοντες που επηρεάζουν την ακρίβεια της κάθε μελέτης είναι το μέγεθος δείγματος της μελέτης (όσο μεγαλύτερο δείγμα τόσο καλύτερη ακρίβεια) και ο τρόπος ο οποίος έχει κατασκευαστεί η μελέτη ώστε να μπορέσει να μικρύνει την διασπορά (υποθέσεις ανεξαρτησίας, εμφωλευμένα δεδομένα κλπ).

1.3 Μοντελοποίηση

Οι περισσότερες μετα-αναλύσεις βασίζονται σε ένα από τα δύο στατιστικά μοντέλα το fixed-effect model (μοντέλο με μοναδικό μέγεθος της επίδρασης) ή το random-effect model (μοντέλο με τυχαία μεγέθη της επίδρασης). Στο πρώτο υποθέτουμε ότι υπάρχει ένα και πραγματικό (true) effect size σε όλες τις μελέτες της μετα-ανάλυσης και ότι οι όποιες διαφορές στα παρατηρούμενα μεγέθη της επίδρασης ανά μελέτη οφείλονται σε δειγματικά λάθη (sampling error). Από την άλλη στο random-effect model επιτρέπουμε το πραγματικό effect size να διαφέρει από μελέτη σε μελέτη. Για παράδειγμα το effect size μπορεί να είναι μεγαλύτερο (ή μικρότερο) σε άτομα μιας από τις μελέτες με ηλικιωμένα άτομα από μια με μικρότερης ηλικίας άτομα, ή κάποιας που τα άτομα που συμμετέχουν έχουν υψηλότερη μόρφωση από μιας άλλης μελέτης κλπ. Είναι δηλαδή λογικό να υπάρχουν διαφορές αφού το δείγμα από μελέτη σε μελέτη ενδέχεται να είναι διαφορετικό.

1.3.1 Μοντέλο με μοναδικό μέγεθος της επίδρασης (Fixed-effect model)

Κάτω από το Fixed-effect model υποθέτουμε ότι όλες οι μελέτες της μετα-ανάλυσης έχουν ένα κοινό πραγματικό effect size. Δηλαδή όλοι οι παράγοντες που μπορούν να επηρεάσουν το effect size είναι ίδιοι σε κάθε μελέτη. Συμβολίζουμε το πραγματικό effect size με το γράμμα θ.

Αφού όλες οι μελέτες έχουν ένα πραγματικό effect size αυτό έπεται ότι οι όποιες διαφορές στο παρατηρούμενο ανά μελέτη effect size οφείλονται στο τυχαίο σφάλμα που κληρονομείται από κάθε μελέτη. Αν συμβολίσουμε με Y_i το παρατηρούμενο στην

i-οστη μελέτη μέγεθος της επίδρασης, τότε έχουμε την σχέση

$$Y_i = \theta + \epsilon_i \tag{1.39}$$

δηλαδή έχουμε ότι είναι ίσο με το πραγματικό effect size συν το τυχαίο σφάλμα της i-οστής μελέτης. Για το ϵ_i αφού είναι τυχαίο μπορούμε να υποθέσουμε κατανομή.
Για να μπορέσουμε να πετύχουμε μεγαλύτερη ακρίβεια στην εκτιμήση μας, δηλαδή στο summary effect, μικραίνοντας την διασπορά υπολογίζουμε το μέσο των effect size με βάρος (σταθμισμένος μέσος), όπου το βάρος που δίνεται στην κάθε μελέτη είναι αντιστρόφως ανάλογο της διασποράς της και έτσι το βάρος μιας μελέτης δίνεται από τον τύπο:

$$W_i = \frac{1}{V_{Y_i}} \tag{1.40}$$

και οπότε για τον σταθμισμένο μέσο έχουμε

$$M = \frac{\sum_{i=1}^{k} W_i Y_i}{\sum_{i=1}^{k} W_i}, \tag{1.41}$$

όπου k είναι ο αριμός των μελετών της μετα-ανάλυσης μας.
Για την διασπορά του σταθμισμένου μέσου έχουμε

$$V_M = \frac{1}{\sum_{i=1}^{k} W_i}, \tag{1.42}$$

και προφανώς

$$SE_M = \sqrt{V_M}. \tag{1.43}$$

Έτσι για τα διαστήματα εμπιστοσύνης έχουμε

$$LL_M = M - 1.96 \cdot SE_M$$

$$UL_M = M + 1.96 \cdot SE_M.$$

Ενώ μπορούμε να κάνουμε και τον στατιστικό έλεγχο $\theta = 0$, με ελεγχοσυνάρτηση

$$Z = \frac{M}{SE_M}.$$

1.3.2 Μοντέλο με τυχαία μεγεθη της επίδρασης (Random-effect model)

Στην απόφαση μας να ενώσουμε ένα σύνολο από μελέτες σε μια μετα-ανάλυση έχουμε θεωρήσει πως αυτές που διαλέξαμε μοιράζονται αρκετά κοινά γνωρίσματα ώστε να έχει λογική το εγχείρημά μας, αλλά δεν υπάρχει λόγος απαραίτητα να υποθέσουμε ότι αυτά τα κοινά γνωρίσματα είναι *ακριβώς* τα ίδια σε όλες τις μελέτες, δηλαδή ότι όλες οι μελέτες μοιράζονται το ίδιο ακριβώς πραγματικό effect size.

Σαν παράδειγμα ας θεωρήσουμε πως έχουμε μελέτες με τα αποτελέσματα μιας διδακτικής μεθόδου. Στις διάφορες μελέτες ενδέχεται να υπάρχουν διαφορές που να οφείλονται σε παράγοντες όπως οι ηλικίες των παιδιών, το μέγεθος των τάξεων κλπ τα οποία θα διαφέρουν από μελέτη σε μελέτη. Για να μπορέσουμε να περιλάβουμε στην μετα-ανάλυση μας αυτή την διασπορά ανάμεσα στις μελέτες χρησιμοποιούμε τα random-effect model. Στα random-effect model συνήθως υποθέτουμε πως τα πραγματικά effect size κατανέμονται κανονικά.
Έτσι το παρατηρούμενο effect size Y_i για την ι-οστη μελέτη δίνεται από τον συνολικό μέσο μ όλων των effect size, συν το τυχαίο σφάλμα της μελέτης ϵ_i, συν την πραγματική διασπορά ανάμεσα στις μελέτες ζ_i. Δηλαδή

$$Y_i = \mu + \epsilon_i + \zeta_i, \tag{1.44}$$

όπου το ζ_i εξαρτάται από την διασπορά της κατανομής των πραγματικών effect size μεταξύ των μελετών, συμβολίζεται με τ^2 και ονομάζεται διασπορά μεταξύ των ερευνών (between study variance), ενώ η ϵ_i εξαρτάται από την V_Y την διασπορά της κατανομής των σφαλμάτων στην κάθε μελέτη ξεχωριστά ή απλά εσωτερική διασπορά (within study variance).

Όπως και προηγουμένως η V_Y καθορίζει τα βάρη που θα δωθούν στις μελέτες, ενώ η διασπορά μεταξύ των ερευνών εκτιμάται με διάφορους τρόπους με πιο συνηθισμένο την εκτίμηση με την μέθοδο των στιγμών. Παρακάτω δίνονται οι μαθηματικές σχέσεις που χρησιμοποιούμε σε αυτή την μέθοδο, τις οποίες θα εξηγήσουμε πιο αναλυτικά στο επόμενο κεφάλαιο που θα μιλήσουμε για την Ετερογένεια:

$$T^2 = \frac{Q - df}{C}, \tag{1.45}$$

στην περίπτωση όπου το $Q - df$ είναι μικρότερο του 0 τότε $T^2 = 0$,
όπου

$$Q = \sum_{i=1}^{k} W_i Y_i^2 - \frac{(\sum_{i=1}^{k} W_i Y_i)^2}{\sum_{i=1}^{k} W_i}, \tag{1.46}$$

$$df = k - 1, \tag{1.47}$$

$$C = \sum_{i=1}^{k} W_i - \frac{\sum_{i=1}^{k} W_i^2}{\sum_{i=1}^{k} W_i}. \tag{1.48}$$

Έτσι όπως στο προηγούμενο μοντέλο έχουμε:

$$W_i^* = \frac{1}{V_{Y_i}^*} \tag{1.49}$$

όπου

$$V_{Y_i}^* = V_{Y_i} + T^2 \tag{1.50}$$

και αντίστοιχα για τις σχέσεις αυτού του μοντέλου στις οποίες για να τις ξεχωρίσουμε από τις προηγούμενες του μοντέλου με μοναδικό μέγεθος επίδρασης, τους τοποθετούμε σαν εκθέτη έναν αστερίσκο, καθότι είναι παρόμοιες αλλά στις διασπορές των μελετών προστίθεται και η ετερογένεια. Έχουμε

$$M^* = \frac{\sum_{i=1}^{k} W_i^* Y_i}{\sum_{i=1}^{k} W_i^*}, \tag{1.51}$$

$$V_M^* = \frac{1}{\sum_{i=1}^{k} W_i^*}, \tag{1.52}$$

$$SE_{M^*} = \sqrt{V_{M^*}}, \tag{1.53}$$

$$LL_{M^*} = M^* - 1.96 \cdot SE_{M^*},$$

$$UL_{M^*} = M^* + 1.96 \cdot SE_{M^*}$$

και για τον έλεγχο $\mu = 0$ έχουμε την ελεγχοσυνάρτηση

$$Z^* = \frac{M^*}{SE_{M^*}}.$$

1.4 Ετερογένεια (Heterogeneity)

Στη μετα-ανάλυση είναι σημαντικό να κατανοήσουμε τη μεταβλητότητα (*dispersion*) από μελέτη σε μελέτη και να την λάβουμε υπ΄ όψην στα αποτελέσματα. Αν τα effect size είναι συνεπής τότε συνήθως επικεντρώνουμε το ενδιαφέρον μας στο summary effect και αναφέρουμε ότι είναι *robust* (ευσταθές). Αν διαφέρουν λίγο μεταξύ τους πάλι στρεφόμαστε στο summary effect, αλλά αναφέρουμε τις όποιες μικροδιαφορές. Αν πάλι διαφέρουν σημαντικά τότε επικεντρώνουμε στη διασπορά. Το πρόβλημα που θέλουμε να αναδείξουμε είναι ότι η παρατηρούμενη διασπορά των εκτιμήσεων μας είναι μερικώς εσφαλμένη και περιέχει την πραγματική διασπορά και το τυχαίο σφάλμα. Θα δούμε πως απομονώνεται η πραγματική διασπορά και πως χρησιμοποιείται ώστε με κατάλληλα μέτρα να μπορούμε να περιγράψουμε την μεταβλητότητα.

1.4.1 Απομονώνοντας τη διασπορά στα true effect

Για να κατανοήσουμε το πρόβλημα ας υποθέσουμε αρχικά πως όλες οι μελέτες της μετα-ανάλυσής μας έχουν το ίδιο true effect size, δηλαδή η πραγματική ετερογένεια είναι μηδεν. Κάτω από αυτή την υπόθεση δεν περιμένουμε τα παρατηρούμενα effect size να είναι τα ίδια, αλλά να πέφτουν σε ένα εύρος τιμών γύρω από το summary effect και αυτό λόγω του σφάλματος εσωτερικά στην κάθε μελέτη (within study error). Υποθέτοντας τώρα ότι το true effect size όντως διαφέρει από μελέτη σε μελέτη, τότε τα παρατηρούμενα effect size διαφέρουν για δύο λόγους ο ένας είναι ο προαναφερθής και ο άλλος είναι η ετερογένεια στα effect size. Οπότε άμα θέλουμε να απομονώσουμε την ετερογένεια πρέπει να ξεχωρίσουμε την παρατηρούμενη διασπορά σε αυτά τα δύο κομμάτια. Ο μηχανισμός για να το κάνουμε αυτό είναι ο εξής :

1. Υπολογίζουμε το συνολικό μέγεθος της παρατηρούμενης διασποράς από μελέτη σε μελέτη.

2. Εκτιμούμε πόσο τα παρατηρούμενα effect size θα διέφεραν άμα το true effect size ήταν ίδιο σε όλες τις μελέτες.

3. Η διασπορά που πλεονάζει (αν υπάρχει) θεωρούμε πως αντανακλά τις διαφορές στα effect size. (Αυτό είναι η ετερογένεια).

Υπολογίζοντας το Q

Το πρώτο βήμα για να ξεχωρίσουμε τις διασπορές είναι να υπολογίσουμε το Q, που ορίζεται ως:

$$Q = \sum_{i=1}^{k} W_i (Y_i - M)^2, \tag{1.54}$$

όπου W_i είναι το βάρος και Y_i το effect size της ι-οστης μελέτης, M το summary effect και k ο αριθμός των μελετών της μετα-ανάλυσής μας. Εναλλακτικά μπορεί να γραφτεί ως :

$$Q = \sum_{i=1}^{k} \left(\frac{Y_i - M}{SE_i} \right)^2, \tag{1.55}$$

ενώ για ευκολία στις πράξεις ο τύπος γράφεται

$$Q = \sum_{i=1}^{k} W_i Y_i^2 - \frac{(\sum_{i=1}^{k} W_i Y_i)^2}{\sum_{i=1}^{k} W_i}.$$

Η αναμενομενη τιμή του Q βασισμένη στο σφάλμα εσωτερικά μιας μελέτης

Στο επόμενο βήμα θα καθορίσουμε την αναμενόμενη τιμή με βάση την υπόθεση ότι το true effect size είναι ίδιο σε όλες τις μελέτες και η διασπορά οφείλεται μόνο από το σφάλμα μέσα στις μελέτες. Επειδή το Q είναι τυποποιημένο (*standardized*) η αναμενόμενη τιμή δεν εξαρτάται από την τιμή του effect size, αλλά απ΄τούς βαθμούς ελευθερίας, δηλαδή

$$df = k - 1.$$

Η πλεονάζουσα διασπορά

Προφανώς η διασπόρα που πλεονάζει βρίσκεται από την σχέση

$$Q - df. \tag{1.56}$$

1.4.2 Έλεγχος,εκτίμηση και μέγεθος ετερογένειας

Ελέγχοντας την ύπαρξη ομοιογένειας

Μπορεί να αποδειχτεί ότι

$$Q \sim {X^2}_{k-1} \tag{1.57}$$

και έτσι γίνεται δυνατό να ελέγξουμε την ύπαρξη ομοιογένειας, δηλαδή την

H_0: **όλες οι έρευνες μοιράζονται το ίδιο μέγεθος της επίδρασης**.

Κατά τα γνωστά απορρίπτουμε την H_0 αν το p value είναι μικρότερο του επίπεδου στατιστικής σημαντικότητας. Σε περίπτωση που δεν απορρίπτεται η H_0 αυτό δεν σημαίνει ότι απαραίτητα δεν υπάρχει ετερογένεια αφού το αποτέλεσμα ενδέχεται να οφείλεται στην έλλειψη στατιστικής δύναμης και γενικότερα είναι προτιμότερο αν πιστεύουμε πως υπάρχει διαφορά μεταξύ των μελετών στο effect size να συνεχίσουμε με αυτή μας την πεποίθηση.

Εκτιμώντας το τ^2

Το Q δεν είναι ένα αντιλήψιμο νοητικά μέγεθος, είναι ένα άθροισμα το οποίο εξαρτάται από το πλήθος των μελετών και επιπλέον είναι τυποποιημένο. Προηγουμένως μιλήσαμε για το τ^2 τη διασπορά που υπάρχει μεταξύ των μελετών, αφού δεν μπορούμε να γνωρίζουμε τα true effect size δεν γίνεται να βρούμε ακριβώς τη διασπορά τους, αλλά την εκτιμάμε από τα παρατηρούμενα effect size με την χρήση της πλεονάζουσας διασποράς, φέρνοντας την στο μετρικό σύστημα που ήταν πριν τυποποιηθεί, διαρώντας με το C, δηλαδή

$$T^2 = \frac{Q - df}{C},$$

όπου

$$C = \sum_{i=1}^{k} W_i - \frac{\sum_{i=1}^{k} W_i^2}{\sum_{i=1}^{k} W_i}.$$

Εδώ να αναφέρουμε πως υπάρχουν και άλλοι προτεινόμενοι τρόποι υπολογισμού του T^2, οι οποίοι βασίζονται κυρίως στην μέθοδο μεγιστοποίησης της πιθανοφάνειας, στους οποίους όμως δεν θα αναφερθούμε.

Η τυπική απόκλιση του T^2 δίνεται από τον τύπο:

$$T = \sqrt{T^2}. \tag{1.58}$$

Το I^2

Το T^2 όπως είπαμε είναι στο ίδιο μέτρο με το effect size, σε κάποιες περιπτώσεις όμως θα θέλαμε να δούμε την ετερογένεια σαν το ποσοστό που η παρατηρούμενη διασπορά αντανακλά τις πραγματικές διαφορές των effect size, δηλαδή το

$$I^2 = \frac{Q - df}{Q} \cdot 100\%, \tag{1.59}$$

αντικατοπτρίζει το μέγεθος του ανοίγματος των διαστήματως εμπιστοσύνης που δεν είναι εξαρτημένη από την πραγματική θέση ή το μέγεθος των true effect size. Είναι ένα μέγεθος στο οποίο παρατηρούμε την ασυνέπεια μεταξύ των αποτελεσμάτων των μελετών και όχι μέγεθος πραγματικής διασποράς. Αν τείνει στο μηδέν τότε σχεδόν όλη η διασπορά είναι λανθασμένη, άρα δεν υπάρχει κάτι να εξηγήσουμε, ενώ για μεγάλο μέγεθος πρέπει να υποθέσουμε τους λόγους και πιθανές τεχνικές επεξήγησης αυτού.

1.4.3 Διασπορά και διαστήματα εμπιστοσύνης των μέτρων της ετερογένειας

Υποθέτοντας ότι τα μεγέθη της επίδρασης κατανέμονται κανονικά το τυπικό σφάλμα του T^2 γίνεται να εκτιμηθεί μέσω των παρακάτω σχέσεων:

$$A = df + 2(sw1 - \frac{sw2}{sw1})T^2 + (sw2 - 2 \cdot \frac{sw3}{sw1} + \frac{(sw2)^2}{(sw1)^2})T^4, \tag{1.60}$$

όπου

$$sw1 = \sum_{i=1}^{k} W_i,\ sw2 = \sum_{i=1}^{k} W_i^2, sw3 = \sum_{i=1}^{k} W_i^3, \tag{1.61}$$

και έχουμε ότι

$$V_{T^2} = 2 \cdot \frac{A}{C^2}, \tag{1.62}$$

$$SE_{T^2} = \sqrt{V_{T^2}}. \tag{1.63}$$

Λόγω του ότι η κατανομή του T^2 δεν προσεγγίζεται καλά από την κανονική κατανομή, το διάστημα εμπιστοσύνης δεν θα είναι πολύ ακριβές αν χρησιμοποιήσουμε την προσέγγιση $T^2 \pm 1.96SE_{T^2}$, εκτός κι αν έχουμε μεγάλο αριθμό ερευνών. Υπάρχουν αρκετοί τρόποι ευρέσεως του διαστήματος εμπιστοσύνης, ένας από αυτούς είναι ο ακόλουθος:
Αν $Q > df + 1$ υπολόγισε το

$$B = 1/2 \cdot \frac{ln(Q) - ln(df)}{\sqrt{2Q} - \sqrt{2df - 1}}, \tag{1.64}$$

αλλιώς

$$B = \sqrt{\frac{1}{2(df-1)(1-(3(df-1)^2)^{-1})}}. \tag{1.65}$$

Υπολόγισε τα

$$L = exp\{\frac{1}{2}ln(\frac{Q}{df}) - 1.96B\}, \tag{1.66}$$

και

$$U = exp\{\frac{1}{2}ln(\frac{Q}{df}) + 1.96B\}. \tag{1.67}$$

Τελικά το διάστημα εμπιστοσύνης για το τ^2 μας δίνεται από τις σχέσεις:

$$LL_{T^2} = \frac{df(L^2-1)}{C} \tag{1.68}$$

$$UL_{T^2} = \frac{df(U^2-1)}{C} \tag{1.69}$$

Σε περίπτωση που κάποια τιμή υπολογιστεί μικρότερη του 0 τότε την θέτουμε 0. Αν το κάτω φράγμα είναι μεγαλύτερο του 0 τότε το T^2 θα πρέπει να είναι στατιστικά σημαντικό. Παρόλα αυτά αφού σχετίζεται με το Q και επειδή η κατανομή που ακολουθεί είναι γνωστή από την σχέση (1.57), είναι προτιμότερο να χρησιμοποιούμε αυτή την συνθήκη για να ελέγξουμε τη στατιστική σημαντικότητα. Σημειώνεται πως

το διάστημα εμπιστοσύνης του τ σχηματίζεται από την τετραγωνική ρίζα των δυο τελευταίων σχέσεων.

Υπάρχουν αρκετοί τρόποι υπολογισμού του διαστήματος εμπιστοσύνης του I^2, τα οποία επειδή δεν εκτιμάν κάποια συγκεκριμένη ποσότητα λέγονται διαστήματα αβεβαιότητας. Ένας από αυτούς είναι ο ακόλουθος:
Αν $Q > df + 1$ υπολόγισε το B μέσω της σχέσης (1.64), αλλιώς μέσω της σχέσης (1.65), έπειτα υπολόγισε τα L, U όπως πριν και τελικά:

$$LL_{I^2} = \frac{(L^2 - 1)}{L^2} \cdot 100\%, \tag{1.70}$$

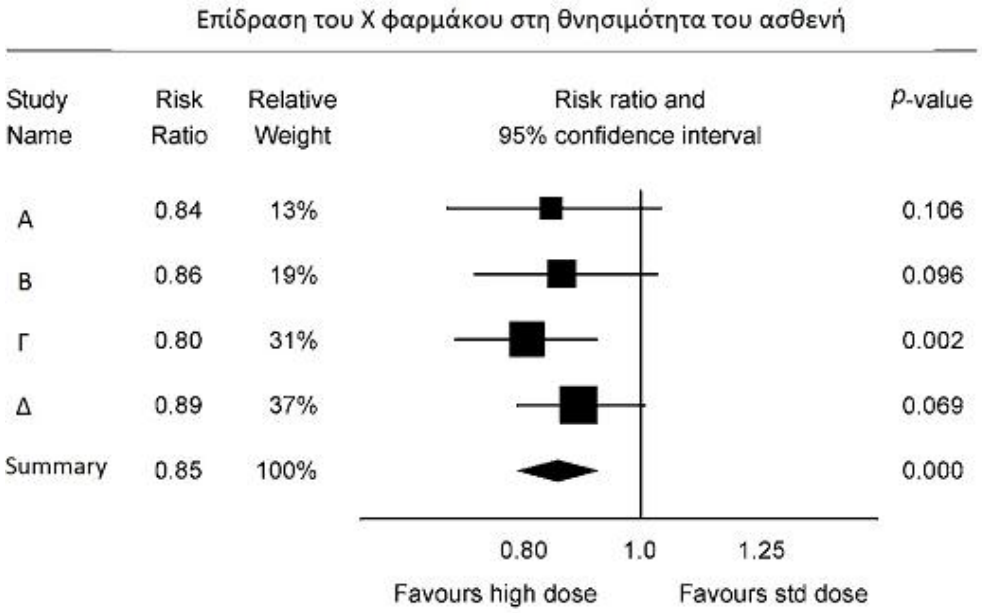
$$UL_{I^2} = \frac{(U^2 - 1)}{U^2} \cdot 100\%. \tag{1.71}$$

Κι εδώ σε περίπτωση που κάποια τιμή υπολογιστεί μικρότερη του 0 τότε την θέτουμε 0, όπως και για τον έλεγχο για το αν το $I^2 = 0$ προτιμούμε τον έλεγχο Q.

1.5 Forest plot

Ένας γραφικός τρόπος για να περιγραφτούν τα δεδομένα, αλλά και τα αποτελέσματα μιας μετα-ανάλυσης είναι το γράφημα δάσος της μετα-ανάλυσης (Forest plot). Τα κύρια χαρακτηριστικά του είναι:

1. Το μέγεθος των τετραγώνων είναι ανάλογο του βάρους που έχει η συγκεκριμένη μελέτη στη μετα-ανάλυση (όσο μεγαλύτερη -τόσο μεγαλύτερο μέγεθος),
2. το κέντρο του τετραγώνου αντιστοιχεί στο effect size της μελέτης που ανήκει ,
3. η οριζόντια γραμμή που τέμνει το τετράγωνο είναι το διάστημα εμποστοσύνης του,
4. το κέντρο του διαμαντιού αντιστοιχεί στο summary effect,
5. το μήκος του διαμαντιού είναι το διάστημα εμπιστοσύνης του summary effect,
6. το p-value είναι ο έλεγχος για την μηδενική υπόθεση.



Σχήμα 1.1: Παράδειγμα γραφήματος τύπου Δάσος μετα-ανάλυσης

1.6 Παράδειγμα Μετα-ανάλυσης

Έστω ότι μας δίνονται τα παρακάτω αποτελέσματα από έξι μελέτες, οι οποίες μας παρέχουν τους μέσους μεταξύ δύο ομάδων οι οποίες είναι ανεξάρτητες η μια από την άλλη. Θα χρησιμοποιήσουμε σαν μέγεθος της επίδρασης τον σταθμισμένο μέσο αυτών με βάση την διόρθωση Hedges.

$NoStudy$	$Mean(New)$	SD	n	$Mean(Standar)$	SD	n
$Study\,1$	94	22	60	92	20	60
$Study\,2$	98	21	65	92	22	65
$Study\,3$	98	28	40	88	26	40
$Study\,4$	94	19	200	82	17	200
$Study\,5$	98	21	50	88	22	45
$Study\,6$	96	21	85	92	22	85

Με τη χρήση των σχέσεων που είδαμε στην σχετική ενότητα, έχουμε τα εξής αποτελέσματα:

	Study 1	*Study* 2	*Study* 3	*Study* 4	*Study* 5	*Study* 6
S_{within}	21.02380	21.50581	27.01851	18.02776	21.47892	21.50581
d	0.09513	0.27899	0.37012	0.66564	0.46557	0.18599
V_d	0.033371	0.031069	0.050856	0.01055	0.043363	0.023631
J	0.99363	0.99412	0.990353	0.998114	0.9919137	0.9955291
df	118	128	78	398	93	168
g	0.094524	0.277356	0.366546	0.664385	0.461808	0.185165
V_g	0.032947	0.030704	0.0498797	0.0105140	0.0426646	0.02342

Θεωρώντας ότι το κατάλληλο μοντέλο είναι της μορφής Fixed Effect, δηλαδή ότι σε όλες τις μελέτες υπάρχει ένα μοναδικό μέγεθος της επίδρασης ($Y = g$, $V_Y = V_g$) έχουμε μέσω των σχέσεων που είδαμε στην σχετική ενότητα, τα παρακάτω αποτελέσματα:

$$W_1 = 30.35151, W_2 = 32.56811, W_3 = 20.04822,$$

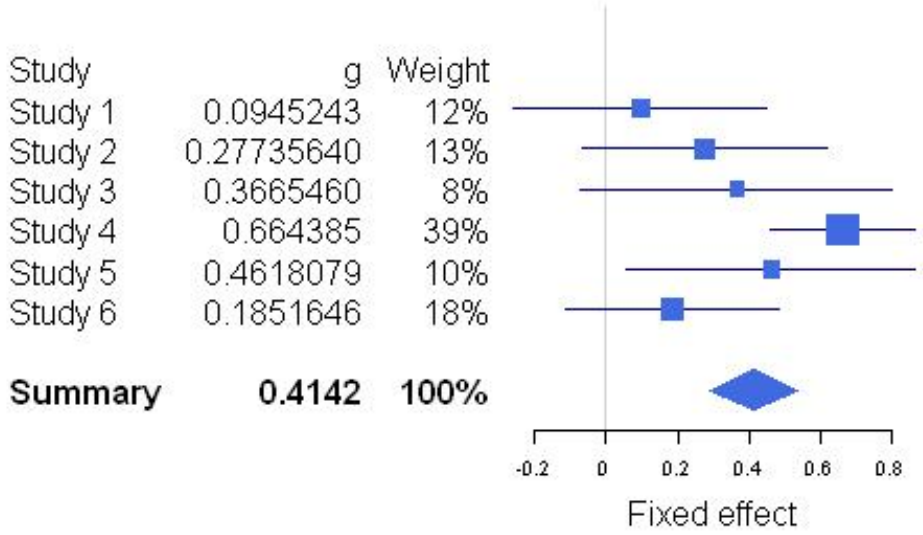
$$W_4 = 95.11053, W_5 = 23.43864, W_6 = 42.69795$$

$$M = 0.4142697$$

$$V_M = 0.004094753$$

$$SE_M = 0.06399026$$

με διάστημα εμπιστοσύνης: $[0.2888488, 0.5396906]$.



Αν είχαμε θεωρήσει ότι το κατάλληλο μοντέλο είναι της μορφής Random Effect, εκτιμάμε αρχικά την διασπορά μεταξύ των διαφορετικών ερευνών, δηλαδή το τ και απ΄τις σχέσεις που είδαμε για αυτό το μοντέλο έχουμε:

$$Q = 12.00325,\ C = 187.6978,\ df = 5$$

$$T^2 = 0.03731131$$

κι από εδώ έχουμε τα βάρη, τον σταθμισμένο μέσο και τα μέτρα θέσης αυτού και μέσω των σχετικών σχέσεων έχουμε:

$$M^* = 0.3582294$$

$$V_M^* = 0.01107621$$

$$SE_M^* = 0.1052$$

με διάστημα εμπιστοσύνης: $[0.1519521, 0.5645068]$.

Για το I^2 έχουμε

$$I^2 = 58.3\%.$$

Για τα διαστήματα εμπιστοσύνης των τ^2, I^2 έχουμε μέσω των σχέσεων που είδαμε:

$$sw1 = 244.21,\ sw2 = 13802.33, sw3 = 1021654,$$

$$A = 31.02016,\ V_{T^2} = 0.001760984,\ SE_{T^2} = 0.042.$$

Και αφού $Q > 5 + 1 = 6$

$$B = 0.2305011,\ L = 0.989,\ U = 2.434$$

κι έτσι

$$LL_{T^2} = -0.0007309904,\ UL_{T^2} = 0.1312144$$

άρα

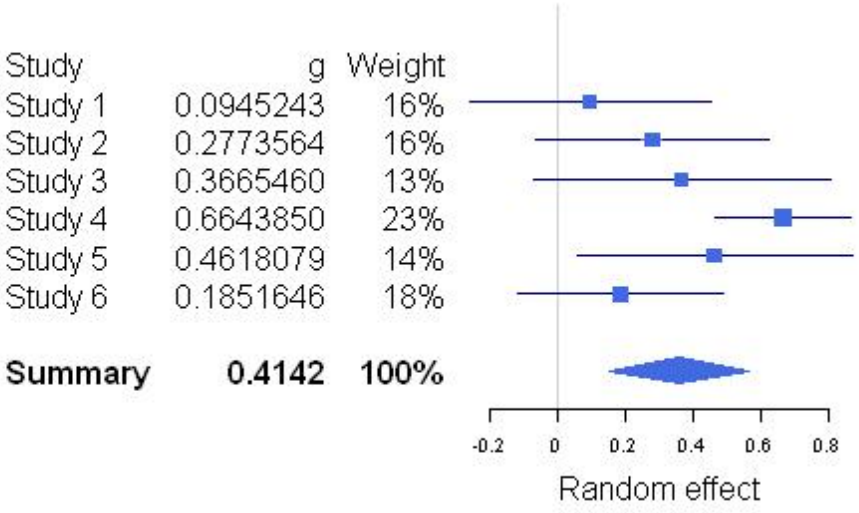
$$\tau \in [0, 0.1312144]$$

Ανάλογα

$$LL_{I^2} = -2.82\%,\ UL_{I^2} = 83.1$$

άρα

$$I^2 \in [0, 83.1\%].$$



Εδώ να αναφέρουμε ότι το ποιό είδος μοντελοποίησης θα χρησιμοποιηθεί εξαρτάται από την φύση των δεδομένων και από την διαίσθηση του επιστήμονα στο αν υπάρχει ετερογένεια ή όχι. Για ενδεικτικούς σκοπούς θα μπορούσε να χρησιμοποιηθεί ο έλεγχος της σχέσης (1.57). Στο παράδειγμά μας η H_0 απορρίπτεται στους 5 βαθμούς ελευθερίας με p value =0.03474325.

Κεφάλαιο 2

Ανάλυση διαχρονικών δεδομένων

(Longitudinal data analysis)

Η ανάλυση διαχρονικών δεδομένων είναι μια στατιστική ανάλυση στην οποία τα δεδομένα είναι συσχετισμένα μεταξύ τους ως προς το ότι προέρχονται από επαναλαμβανόμενες ανά χρονικά διαστήματα μετρήσεις διάφορων χαρακτηριστικών (μεταβλητών) πάνω στο ίδιο άτομο. Σε σύγκριση με τις μελέτες που χρησιμοποιούν διαφορετικά άτομα με τα ίδια συμπτώματα, εδώ οι ίδιοι ασθενείς παρακολουθούνται στην διάρκεια μιας χρονικής περιόδου και έτσι αποφεύγονται σφάλματα που μπορεί να οφείλονται σε άλλους παράγοντες όπως π.χ. διαφορές στις γενεές των ανθρώπων, για αυτό το λόγο τα παρατηρούμενα αποτελέσματα είναι πιο αξιόπιστα. Κύριος σκοπός της ανάλυσης διαχρονικών δεδομένων είναι να ελέγξουμε αν υπάρχει αλλαγή κάποιας κατάστασης και να ερευνήσουμε τα αίτια αυτών των αλλαγών.

2.1 Συνήθης γραμμικά και γενικευμένα γραμμικά μοντέλα

Πριν μιλήσουμε για την ανάλυση διαχρονικών δεδομένων, στην παρούσα ενότητα, θα θυμήσουμε ένα κομμάτι της θεωρίας για τα συνηθισμένα γραμμικά και για τα γενικευμένα γραμμικά μοντέλα, όπως και την φιλοσοφία πίσω από αυτά, ώστε να μπορέσουμε να τα δούμε σε αντιστοιχία με αυτά που χρησιμοποιούμε στην ανάλυση διαχρονικών

δεδομένων.

Στην στατιστική μοντελοποίηση σκοπός μας είναι να δημιουργήσουμε το κατάλληλο περιγραφικό μοντέλο που εξηγεί την μεταβλητότητα μιας απαντητικής μεταβλητής Y. Συνήθως για να ερμηνεύσουμε την Y χρησιμοποιούμε ανεξάρτητες/ ερμηνευτικές μεταβλητές $X = (x_1, ..., x_n)$, δηλαδή έχουμε την Y που ονομάζεται εξαρτημένη μεταβλητή, συναρτήσει των περιγραφικών μεταβλητών. Γενικά

$$Y = f(x_1, ..., x_n) + \epsilon. \tag{2.1}$$

Το ϵ είναι μια τυχαία συνιστώσα που ονομάζεται το σφάλμα, δηλαδή το κομμάτι της μεταβλητότητας των X που δεν ερμηνεύται μέσω της συνάρτησης $f(.)$ και είναι μια τυχαία μεταβλητή για την οποία πρέπει να υποθέσουμε κατανομή.

2.1.1 Το γενικό γραμμικό μοντέλο

Το γενικό γραμμικό μοντέλο, είναι ίσως το πιο συνηθισμένο μοντέλο που συναντάει κανείς και αυτό ορίζεται ως εξής

$$y_i = \beta_0 + \beta_1 x_{1_i} + \beta_2 x_{2_i} + ... + \beta_k x_{k_i} + \epsilon_i \tag{2.2}$$

με τις παρακάτω υποθέσεις

$$E(\epsilon_i) = 0,\ V(\epsilon_i) = 0,\ \forall i,\ COV(\epsilon_i, \epsilon_j) = 0,\ \forall i \neq j \tag{2.3}$$

για $i, j = 1, ..., n$, όπου n το πλήθος του δείγματος και k αριθμός των περιγραφικών μεταβλητών,
και με πιο ισχυρή υπόθεση

$$\epsilon_i \sim N(0, \sigma^2). \tag{2.4}$$

Σε μορφή πινάκων τα παραπάνω γράφονται:

$$y = X\beta + \epsilon \tag{2.5}$$

δηλαδή

$$\begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & . & . & . & x_{k1} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & x_{1n} & . & . & . & x_{kn} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ . \\ . \\ . \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ . \\ . \\ . \\ \epsilon_n k \end{pmatrix}$$

(2.6)

$$E(\epsilon) = 0,\ V(\epsilon) = E(\epsilon \cdot \epsilon') = \sigma^2 I_n, \epsilon \sim N_n(0, V(\epsilon))$$

Ελαχιστοποιώντας το άθροισμα των τετραγωνικών καταλοίπων (μέθοδος ελάχιστων τετραγώνων) $\sum_{i=1}^{n} \hat{\epsilon}_i^2 = \hat{\epsilon}'\hat{\epsilon}$, όπου $\hat{\epsilon} = Y - X\hat{\beta}$, έχουμε την εκτιμήτρια ελάχιστων τετραγώνων για τους συντελεστές των περιγραφικών μεταβλητών:

$$\hat{\beta} = (X'X)^{-1}X'y, \tag{2.7}$$

με την προϋπόθεση ότι ο πίνακας $(X'X)$ αντιστρέφεται, δηλαδή οι στήλες του είναι γραμμικώς ανεξάρτητες, διαφορετικά παρουσιάζεται το πρόβλημα της πολυσυγραμμικότητας, όπου δύο ή και παραπάνω μεταβλητές έχουν γραμμική σχέση η μια με την άλλη, μαθηματικά αυτό σημαίνει πως ο συντελεστής συσχέτισης των δύο μεταβλητών είναι 1. Γενικά η πολυσυγραμμικότητα είναι ένα πρόβλημα το οποίο παρουσιάζεται όταν δύο μεταβλητές έχουν άμεση σχέση η μια με την άλλη και αυτό μπορεί να οδηγήσει σε λάθος συμπεράσματα αφού η μια μεταβλητή θα απορροφάει σημαντικότητα από την άλλη και θέλει ιδιαίτερη προσοχή στην αντιμετώπισή της.
Η παραπάνω εκτιμήτρια είναι γραμμικός συνδυασμός της Y, είναι αμερόληπτη, δηλαδή $E(\hat{\beta}) = \beta$ και αποτελεσματική δηλαδή έχει την ελάχιστη διακύμανση από όλες τις γραμμικές αμερόληπτες εκτιμήτριες του β. Είναι και η εκτιμήτρια μέγιστης πιθανοφάνειας (ε.μ.π.). Η διασπορά του $\hat{\beta}$ δίνεται από την σχέση

$$V(\hat{\beta}) = (X'X)^{-1}\sigma^2 \tag{2.8}$$

και όπως είπαμε σαν γραμμικός συνδυασμός κανονικής ακολουθεί κανονική

$$\hat{\beta} \sim N_{k+1}(\beta, (X'X)^{-1}\sigma^2). \tag{2.9}$$

Η αμερόληπτη εκτιμήτρια του σ^2 δίνεται από την σχέση

$$\hat{\sigma}^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{n-k-1}, \tag{2.10}$$

η οποία είναι αποτελεσματική, ενώ η εκτιμήτρια μέγιστης πιθανοφάνειας του σ^2 είναι η

$$\hat{\sigma}^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{n} \tag{2.11}$$

η οποία δεν είναι αμερόληπτη, ούτε αποτελεσματική. Οι ε.μ.π. των β, σ^2 είναι συνεπείς, ασυμπτωτικά αποτελεσματικές και ασυμπτωτικά κανονικές.
Ορίζοντας ως

$$S(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1} \tag{2.12}$$

που τυπικά είναι η εκτιμήτρια της διασποράς της εκτιμήτριας των β, όταν η διασπορά των καταλοίπων είναι άγνωστη, αποδεικνύεται ότι ισχύει η σχέση

$$\frac{\hat{\beta}_i - \beta_i}{S(\hat{\beta}_i)} \sim t(n-k-1) \tag{2.13}$$

όπου $S(\hat{\beta}_i)$ είναι το $(i+1)$-οστό διαγώνιο στοιχείο του πίνακα $S(\hat{\beta}_i)$ και από εδώ μπορούμε να σχηματίσουμε τα σχετικά διαστήματα εμπιστοσύνης

$$\hat{\beta}_i \pm t_{a/2}(n-k-1)S(\hat{\beta}_i) \tag{2.14}$$

και τους σχετικούς ελέγχους υποθέσεων για $H_0 : \beta_i = \beta^*$,

$$\frac{\hat{\beta}_i - \beta^*}{S(\hat{\beta}_i)} \sim t_{a/2}(n-k-1), \tag{2.15}$$

όπου προφανώς για $H_0 : \beta_i = 0 \; vs \; H_1 : \beta_i \neq 0$ έχουμε τον έλεγχο στατιστικής σημαντικότητας της i μεταβλητής.
Στην περίπτωση που θέλουμε να πραγματοποιήσουμε έναν από κοινού έλεγχο στα-

τιστικών υποθέσων, πχ

$$H_0 : \begin{cases} \beta_1 = 0 \\ \beta_2 - \beta_3 = 0 \end{cases}$$

ή σε μορφή πινάκων

$$H_0 : R \cdot \beta = r \Leftrightarrow \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \beta = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

χρησιμοποιούμε για ελεγχοσυνάρτηση την σχέση

$$F = \frac{(R\hat{\beta} - r)'[RS(\hat{\beta})R']^{-1}(R\hat{\beta} - r)}{j} \tag{2.16}$$

η οποία κάτω από την H_0 ακολουθεί την κατανομή $F(j, n-k-1)$, που j είναι ο αριθμός των γραμμών του πίνακα r.

2.1.2 Γενικευμένα γραμμικά μοντέλα

Στο συνηθισμένο γραμμικό μοντέλο η εξαρτημένη μεταβλητή Y ακολουθούσε την κανονική κατανομή, τι συμβαίνει όμως στην περίπτωση όπου πχ η εξαρτημένη μεταβλητή είναι μια δίτιμη μεταβλητή γεγονότος - μη γεγονότος ή γενικότερα αν τα δεδομένα που μας δίνοται ακολουθούν κάποια άλλη κατανομή.

Αρχικά ας δούμε πότε μια τυχαία μεταβλητή ανήκει στην εκθετική οικογένεια κατανομών (E.O.K.), ένα σύνολο κατανομών στο οποίο περιέχεται και η κανονική αλλά και μερικές από τις πιο γνωστές κατανομές όπως η Benoulli, η Γάμμα, η Poisson κλπ.

Έστω λοιπόν ότι έχουμε την τυχαία μεταβλητή Y με κατανομή $f(y; \theta, \phi)$, όπου το θ ονομάζεται κανονική παράμετρος και το ϕ παράμετρος όχλησης, και 1) η κατανομη της μπορεί να γραφτεί στην μορφή

$$f(y; \theta, \phi) = exp\{\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\}, \tag{2.17}$$

για κάποιες πραγματικές συναρτήσεις $a(.), b(.), c(.)$ και 2) το στήριγμα της δεν εξαρτάται από τις δύο αυτές παραμέτρους, τότε η Y ανήκει στην εκθετική οικογένεια κατανομών.

Ισχύει επίσης ότι:

$$E(Y) = b'(\theta),\ V(Y) = a(\phi)b''(\theta). \tag{2.18}$$

Στα γενικευμένα γραμμικά μοντέλα η περιγραφική μας μεταβλητή θα ανήκει σε αυτήν την οικογένεια κατανομών, λόγω αυτού είναι προφανές ότι το συνηθισμένο γραμμικό μοντέλο είναι μια υποπερίπτωση των γενικευμένων. Στα γενικευμένα γραμμικά μοντέλα η τυχαία συνιστώσα μας δίνει πληροφορία για την κατανομή της Y και η $f(.)$ για την σχέση της Y με τις περιγραφικές. Για να δούμε την αντιστοιχία του συνηθισμένου γραμμικού μοντέλου με το γενικευμένο, στο συνηθισμένο είχαμε την σχέση $E(y_i) = x_i'\beta$, δηλαδή $E(y_i) = \mu_i = \eta_i$, ενώ στο γενικευμένο υπάρχει μια συνάρτηση $g(.)$ η οποία ονομάζεται συνάρτηση σύνδεσμος και η οποία είναι συναρτήση της μέσης τιμής, δηλαδή

$$\eta_i = g(\mu_i) \tag{2.19}$$

όπου η $g(.)$ πρέπει να είναι 1-1 και παραγωγίσιμη και το πεδίο ορισμού της εξαρτάται από την υπόθεση κατανομής.
Άρα αντιστρέφοντας έχουμε

$$E(y_i) = g^{-1}(x_i'\beta) \tag{2.20}$$

Η εκτίμηση μέγιστης πιθανοφάνειας λύνεται με την μέθοδο Newton Raphson στην σχέση

$$\frac{d(lnf(y;\beta,\phi))}{d\beta} = 0_k \tag{2.21}$$

Γενικευμένος έλεγχος λόγου πιθανοφανειών

Για να συγκρίνουμε δύο μοντέλα τα οποία είναι ΄φωλιασμένα΄ (nested), δηλαδή πρακτικά αναφερόμαστε σε μοντέλα κάτω από τα ίδια δεδομένα στα οποία το ένα περιέχει ένα υποσύνολο περιγραφικών μεταβλητών του άλλου, κάτω από την ίδια συνάρτηση σύνδεσμο της ίδιας κατανομής EOK, χρησιμοποιούμε τον γενικευμένο λόγο πιθανοφανειών. Δηλαδή θεωρώντας τον έλεγχο χωρίς βλάβη της γενικότητας $H_0 : \beta_{q+1} = \beta_{q+2} = ... = \beta_p = 0$ δηλαδή του μοντέλου με $n_i = \sum_{j=1}^{q} x_{ij}\beta_j$, έναντι του H_1 : του μοντέλου με όλες τις περιγραφικές μεταβλητές, δηλαδή με $n_i = \sum_{j=1}^{p} x_{ij}\beta_j$, έχουμε

την στατιστική συνάρτηση του λόγου των πιθανοφανειών

$$L_{01} = 2ln\left(\frac{max_{\theta\in\theta(1)}f_y(y;\theta)}{max_{\theta\in\theta(0)}f_y(y;\theta)}\right) = 2ln(f_y(y;\hat{\theta}(1)) - 2ln(f_y(y;\hat{\theta}(0)), \qquad (2.22)$$

όπου $\theta(0)$ είναι το σύνολο τιμών των παραμέτρων θ κάτω από την H_0 που είναι υποσύνολο του $\theta(1)$ του συνόλου τιμών κάτω από την εναλλακτική υπόθεση και $\hat{\theta}(0)$, $\hat{\theta}(1)$ τα αντίστοιχα διανύσματα των εκτιμήσεων των παραμέτρων κάτω από τις δύο υποθέσεις, εκτιμώμενες με ε.μ.π. αφού ισχύουν οι σχέσεις $b'(\hat{\theta}_i) = \hat{\mu}_i, g(\hat{\mu}_i) = \hat{\eta}_i$. Αποδεικνύεται ότι

$$L_{01} \sim X^2_{p-q} \qquad (2.23)$$

και μέσω αυτής της σχέσης έχουμε μια ελεγχοσυνάρτηση η οποία μπορεί να χρησιμοποιηθεί και για τους ελέγχους στατιστική σημαντικότητας ολόκληρου του μοντέλου, αλλά και των επι μέρος μεταβλητών του. Ας σημειωθεί ότι οι παραπάνω σχέσεις ισχύουν για παράγοντα όχλησης γνωστό, όταν αυτός είναι άγνωστος χρησιμοποιούνται οι ανάλογες διαδικασίες εκτίμησης.

2.2 Βασικές έννοιες ανάλυσης διαχρονικών δεδομένων

Αφού θυμήσαμε τι γενικά ισχύει στα γραμμικά και γενικευμένα γραμμικά μοντέλα, ας προχωρήσουμε τώρα στο κύριο κομμάτι του κεφαλαίου την ανάλυση διαχρονικών δεδομένων.

2.2.1 Ειδικά χαρακτηριστικά - Στόχοι

Τα διαχρονικά δεδομένα είναι εμφωλευμένα (clustered). Αυτό συμβαίνει από το γεγονός ότι προέρχονται από το ίδιο άτομο σε διαφορετικές χρονικές στιγμές, π.χ. αν κάποιο άτομο ασθενεί σήμερα από μια χρονοβόρα πάθηση, έχει άμεση σχέση με το αν θα ασθενεί σε 6 μήνες, υποθέτοντας πως η εξέταση γίνεται κάθε έξι μήνες. Έτσι έχουμε συσχετισμένες παρατηρήσεις και η λογική λέει ότι είναι θετικά συσχετισμένες. Συνήθως ο λόγος που υπάρχει αυτή η συσχέτιση θεωρείται ασήμαντος, αλλά παρόλα

αυτά η συσχέτιση πρέπει να ληφθεί υπ΄όψην στην ανάλυση και έτσι η συνήθης υπόθεση περί ανεξαρτησίας μεταξύ των τιμών μιας μεταβλητής στις χρονικές στιγμές ενός ασθενή δεν μπορεί να χρησιμοποιηθεί εδώ. Προφανώς μεταξύ διαφορετικών ασθενών υποθέτουμε ανεξαρτησία.

Οι στόχοι μιας ανάλυσης διαχρονικών δεδομένων είναι να ελέγξουμε την ετερογένεια στις τιμές των μεταβλητών στον κάθε ασθενή, να ελέγξουμε τις αλλαγές μέσα στο χρόνο - πράγμα αδύνατο στις έρευνες που χρησιμοποιούνται τιμές από μια χρονική στιγμή, να ελέγξουμε τα αίτια για αυτές τις αλλαγές και να κάνουμε εκτιμήσεις για το πως η κατάσταση κάποιου ατόμου αλλάζει μέσα στο χρόνο.

2.2.2 Ισορροπημένη και μη ισορροπημένη έρευνα

Σε μια ανάλυση διαχρονικών δεδομένων οι μονάδες (π.χ. ασθενείς) για τις οποίες ελέγχουμε τις μεταβλητές τους αποκαλούνται είτε υποκείμενα, είτε άτομα. Οι μετρήσεις γίνονται σε διαφορετικούς χρόνους ή περιόδους. Αν όλα τα άτομα έχουν ακριβώς τον ίδιο αριθμό παρατηρήσεων τότε η έρευνα λέγεται ισορροπημένη (balanced), στην διαφορετική περίπτωση που είναι και η συνήθης ονομάζεται μη ισορροπημένη έρευνα (unbalanced). Τα δεδομένα συλλέγονται είτε αναδρομικά, είτε την ίδια στιγμή της μελέτης, προφανώς η δεύτερη περίπτωση είναι και καλύτερη, ενώ οι απουσιάζοντες παρατηρήσεις (missing data) σε μια μελέτη είναι κι αυτό ένα συνηθισμένο φαινόμενο. Ένα παράδειγμα ισορροπημένης έρευνας είναι μια έρευνα που γίνεται πάνω σε πειραματόζωα τα οποία πάντα θα είναι παρών στις διάφορες εξετάσεις ανά τα χρονικά διαστήματα που έχουν οριστεί, ενώ παράδειγμα μιας μη ισοροπημένης έρευνας είναι οι άνθρωποι ασθενείς που τους έχει ζητηθεί ανά εξάμηνο να πηγαίνουν στην κλινική και να εξετάζονται και σε πολλές περιπτώσεις αυτοί αμελούν να πάνε, στην περίπτωση αυτή τα στοιχεία που ήταν να δωθούν θεωρούνται missing data.

2.2.3 Συμβολισμοί

- N άτομα
- από τα οποία παίρνουμε n διαχρονικές μετρήσεις
- στους χρόνους t_i

- Y_{ij} είναι η παρατήρηση που μας ενδιαφέρει για τον έλεγχο (εξαρτημένη μεταβλητή) του i-οστου ατόμου στην j-οστη μέτρηση
- $x_{ij} = (x_{1ij}, x_{2ij}, ..., x_{pij})$ είναι οι περιγραφικές μεταβλητές του i-οστου ατόμου στην j-οστη μέτρηση
- δείκτρια απουσίας $\delta_{ij} = 1$ αν τα Y_{ij} και x_{ij} δεν απουσιάζουν, αλλιώς $\delta_{ij} = 0$

Subject	Time Point	Missing Indicator	Response	Covariates
1	1	δ_{11}	y_{11}	$x_{111} \dots x_{11p}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots \ddots \vdots$
	j	δ_{1j}	y_{1j}	$x_{1j1} \dots x_{1jp}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots \ddots \vdots$
	t_1	δ_{1t_1}	y_{1t_1}	$x_{1t_11} \dots x_{1t_1p}$
.....				
i	1	δ_{i1}	y_{i1}	$x_{i11} \dots x_{i1p}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots \ddots \vdots$
	j	δ_{ij}	y_{ij}	$x_{ij1} \dots x_{ijp}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots \ddots \vdots$
	t_i	δ_{it_i}	y_{it_i}	$x_{it_i1} \dots x_{it_ip}$
.....				
n	1	δ_{n1}	y_{n1}	$x_{n11} \dots x_{n1p}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots \ddots \vdots$
	j	δ_{nj}	y_{nj}	$x_{nj1} \dots x_{njp}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots \ddots \vdots$
	t_n	δ_{nt_1}	y_{nt_1}	$x_{nt_n1} \dots x_{nt_np}$

Σχήμα 2.1: Γενική μορφή δεδομένων διαχρονικής ανάλυσης

Το ενδιαφέρον μας βρίσκεται κυρίως στην αλλαγή της μέσης τιμής $\mu_j = E(Y_{ij})$ της εξαρτημένης μεταβλητής σε σχέση πάντα και με τις περιγραφικές μεταβλητές. Σε περίπτωση που θεωρούμε πως η μέση τιμή διαφέρει και μεταξύ των ασθενών τότε $\mu_{ij} = E(Y_{ij})$.

2.2.4 Εξάρτηση και συσχέτιση

Ας υποθέσουμε μια ανάλυση διαχρονικών δεδομένων με τα χαρακτηριστικά που προσδιορίσαμε προηγουμένως, τότε για τα γνωστά μέτρα θέσης έχουμε τα εξής:

$$\mu_j = E(Y_{ij}), \tag{2.24}$$

$$\sigma_j^2 = V(Y_{ij}) = E[(Y_{ij} - E(Y_{ij}))^2] = E[(Y_{ij} - \mu_j)^2], \tag{2.25}$$

$$\sigma_{jk} = Cov(Y_{ij}, Y_{ik}) = E[(Y_{ij} - \mu_j)(Y_{ik} - \mu_k)], \tag{2.26}$$

$$\rho_{jk} = \frac{\sigma_{jk}}{\sigma_j \sigma_k}, \tag{2.27}$$

έτσι ο πίνακας συνδιακύμανης του i-οστου ατόμου είναι της μορφής:

$$Cov \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ . \\ . \\ . \\ Y_{in} \end{pmatrix} = \begin{pmatrix} \sigma_1 & \sigma_{12} & . & . & . & \sigma_{1n} \\ \sigma_{21} & \sigma_2 & . & . & . & \sigma_{2n} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \sigma_{n1} & \sigma_{n2} & . & . & . & \sigma_n \end{pmatrix}. \tag{2.28}$$

2.3 Μοντελοποιώντας τη διασπορά

Αν και όπως αναφέραμε ο λόγος για τον οποίο υπάρχει συσχέτιση μεταξύ των τιμών της εξαρτημένης μεταβλητής ενός ατόμου δεν είναι σημαντικής σημασίας το να υποθέσουμε ύπαρξη διασποράς μεταξύ αυτών αυξάνει την ακρίβεια της εκτίμησης. Επιπλέον στην περίπτωση που έχουμε απουσιάζοντα δεδομένα (missing data) η *σωστή* περιγραφή της δομής της διασποράς είναι αναγκαία για την εκτίμηση των παραμέτρων της παλινδρόμησης. Κάποιες απ΄τις δομές για τον πίνακα συνδιακύμανσης που μπορούμε να επιλέξουμε περιγράφονται παρακάτω.

2.3.1 Χωρίς Δομή (Unstructed)

$$Cov\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ . \\ . \\ . \\ Y_{in} \end{pmatrix} = \begin{pmatrix} \sigma_1 & \sigma_{12} & . & . & . & \sigma_{1n} \\ \sigma_{21} & \sigma_2 & . & . & . & \sigma_{2n} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \sigma_{n1} & \sigma_{n2} & . & . & . & \sigma_n \end{pmatrix}. \tag{2.29}$$

Η συγκεκριμένη δομή είναι λογική όταν ο αριθμός των χρόνων είναι σχετικά μικρός και όλες οι μετρήσεις είναι στους ίδιους ακριβώς χρόνους. Έχει μεγάλο μειονέκτημα ότι πρέπει να εκτιμήσουμε $n(n+1)/2$ παραμέτρους και υπάρχει πρόβλημα όταν οι χρόνοι δεν είναι ίδιοι στις παρατηρήσεις.

2.3.2 Συμμετρικά Σύνθετη (Compound Symmetry)

$$Cov\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ . \\ . \\ . \\ Y_{in} \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & \rho & . & . & . & \rho \\ \rho & 1 & . & . & . & \rho \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \rho & \rho & . & . & . & 1 \end{pmatrix}, \tag{2.30}$$

όπου $\sigma^2 = Var(Y_{ij})$ και $\rho = \rho_{jk}$. Το πλεονέκτημα σε αυτή την δομή είναι ότι έχουμε να εκτιμήσουμε μόνο δύο παραμέτρους. Η συγκεκριμένη δομή όμως ουσιαστικά σημαίνει πως υποθέτουμε ότι οποιοδήποτε ζευγάρι τιμών έχει την ίδια συσχέτιση, χωρίς να έχει σημασία ο χρόνος που έχει περάσει μεταξύ των δύο, ενώ γενικά περιμένουμε ότι η συσχέτιση όσο μεγαλώνει η χρονική απόσταση δύο τιμών θα φθείρεται. Επιπλέον η σταθερή διασπορά δεν είναι τόσο ευσταθής, συνήθως η διασπορά αυξάνει με τον χρόνο.

2.3.3 Toepliz

$$Cov\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ . \\ . \\ . \\ . \\ Y_{in} \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & . & . & \rho_{n-1} \\ \rho_1 & 1 & . & . & . & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & . & . & \rho_{n-3} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & . & . & 1 \end{pmatrix}, \tag{2.31}$$

Υποθέτει ότι οποιοδήποτε ζευγάρι με ίσο χρονικό διάστημα έχει την ίδια συσχέτιση. Η διασπορά είναι και εδώ σταθερή και $\rho_k = \rho_{j,j+k}$. Είναι κατάλληλλη όταν οι μετρήσεις είναι σε ίδιες χρονικές περιόδους και χρειάζεται να εκτιμήσουμε n παραμέτρους.

2.3.4 Αυτο-οπισθοδρομική (Autoregressive)

$$Cov\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ . \\ . \\ . \\ . \\ Y_{in} \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & . & . & \rho^{n-1} \\ \rho & 1 & . & . & . & \rho^{n-2} \\ \rho^2 & \rho & 1 & . & . & \rho^{n-3} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & . & . & 1 \end{pmatrix} \tag{2.32}$$

Μια ειδική περίπτωση της Toepliz. Το πλεονέκτημά της είναι ότι οι παράμετροι προς εκτίμηση μειώνονται στους 2.

2.3.5 Συγκεντρωτική (Banded)

$$Cov\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ . \\ . \\ . \\ . \\ Y_{in} \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & \rho_1 & 0 & . & . & 0 \\ \rho_1 & 1 & \rho_1 & . & . & 0 \\ 0 & \rho_1 & 1 & . & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & 1 \end{pmatrix} \tag{2.33}$$

Η συγκεκριμένη δομή υποθέτει πως από ένα σημείο και μετά δεν υπάρχει συσχέτιση. Δηλαδή $\rho_{j,j+k} = 0$ από κάποιο k και μετά. Το συγκεκριμένο παράδειγμα είναι μια Toepliz στην οποία για $k = 2$ και μετά η συσχέτιση θεωρείται μηδενική. Και εδώ έχουμε το μειονέκτημα ότι κάνουμε μια πολύ ισχυρή υπόθεση.

2.3.6 Εκθετική

Όταν οι μετρήσεις δεν είναι ίσα χρονικώς κατανεμημένες μπορούμε να γενικεύσουμε την αυτο-οπισθοδρομική δομη υποθέτοντας πως

$$\rho_{jk} = \rho^{|t_{ij}-t_{ik}|} = exp\{log(\rho)|t_{ij} - t_{ik}|\} \tag{2.34}$$

για $\rho > 0$. Έτσι θεωρούμε πως ο χρόνος φθίνει εκθετικά με τη διαφορά του χρόνου μεταξύ των δυο τιμών που συσχετίζονται.

2.4 Μοντελοποίηση

Υπάρχουν δύο είδη μοντέλων που μπορούν να περιγράψουν τα διαχρονικά δεδομένα το ένα είναι αυτό που τα αποτελέσματά του αναφέρονται στο κάθε άτομο ξεχωριστά (subject-specific models, μοντέλα SA), ενώ στο δεύτερο αναφέρονται στον μέσο του πληθυσμού (marginal or population average models, μοντέλα PA).

2.4.1 Γενικευμένα Γραμμικά Μοντέλα Μικτών Επιδράσεων

(Generalized Linear Mixed Effects Models)

Η μοντελοποίηση με βάση τα GLMM χρησιμοποιείται όταν θέλουμε να βγάλουμε συμπεράσματα για το κάθε άτομο ξεχωριστά, δηλαδή είναι ένα subject-specific μοντέλο (SS). Υποθέτουμε πως υπάρχουν τυχαίες επιδράσεις (random effects) οι οποίες ακολουθούν μια πολυδιάστατη κανονική κατανομή, δηλαδή υποθέτουμε ότι στο κάθε άτομο ξεχωριστά επιδρά κάποιος άγνωστος παράγοντας τον οποίο δεν γνωρίζουμε και αυτό επηρεάζει το μοντέλο.
Η κατανομή που ακολουθούν τα Y_{ij} θεωρούμενη πως υπάρχουν random effects ανήκει στην εκθετική οικογένεια κατανομών και η διασπορά δίνεται από την σχέση

$$Var(Y_{ij}) = \phi u(E[Y_{ij}|b_i]), \tag{2.35}$$

όπου b_i είναι τα random effects τα οποία τα υποθέτουμε ανεξάρτητα της εξαρτημένης μεταβλητής.
Η μορφή του μοντέλου είναι

$$\eta_{ij} = X'_{ij}\beta + Z'_{ij}b_i, \tag{2.36}$$

και για κάποια δεδομένη ανάλογα την περίπτωση link function $g(.)$, έχουμε

$$g(E[Y_ij|b_i]) = \eta_{ij} = X'_{ij}\beta + Z'_{ij}b_i. \tag{2.37}$$

Ο πίνακας Z_{ij} καθορίζεται από το που θα υποθέσουμε πως θα υπάρχει τυχαία επίδραση, αυτή μπορεί να εμφανίζεται είτε στο σταθερό όρο, είτε και στις επεξηγηματικές μεταβλητές, τα αίτια για να υποθέσουμε τυχαίες επιδράσεις καθορίζονται είτε διαισθητικά, είτε από υπάρχουσα γνώση, είτε σχηματικά, είτε και με την χρήση στατιστικών ελέγχων. Τα random effects θεωρητικά μπορεί να ακολουθούν οποιαδήποτε πολυδιάστατη κατανομή, στην πράξη όμως ακολουθούν κανονική με μέση τιμή 0 και πίνακα συνδιασποράς G.

Μέσω της κατανομής των τυχαίων επιδράσεων,αλλά και της εξαρτημένης μεταβλητής καθορίζεται πλήρως η απο κοινού

$$f(Y_i, b_i) = f(Y_i|b_i)f(b_i), \tag{2.38}$$

όπου

$$f(Y_i|b_i) = f(Y_{i1}|b_i)f(Y_{i2}|b_i)...f(Y_{in_i}|b_i), \tag{2.39}$$

αφού έχουμε υποθέσει ανεξαρτησία.
Έτσι η πιθανοφάνεια παίρνει την μορφή

$$L(\beta, G, \phi) = \prod_{i=1}^{N} \int f(Y_i|b_i)f(b_i)db_i \tag{2.40}$$

και με αριθμητικές τεχνικές μπορούμε να εκτιμήσουμε τα β, G, ϕ.
Τέλος, δοθέντων των εκτιμήτριων μεγίστης πιθανοφάνειας των β, G, ϕ που βρήκαμε μπορούμε να εκτιμήσουμε και το b_i σαν

$$\hat{b}_i = E(b_i|Y_i; \hat{\beta}, \hat{G}, \hat{\phi}) \tag{2.41}$$

και εδώ χρησιμοποιούνται αντίστοιχες αριθμητικές τεχνικές.

2.4.2 Γενικευμένες εξισώσεις εκτίμησης (Generalized Estimating Equations)

Η μοντελοποίηση με βάση τις GEE χρησιμοποιείται όταν θέλουμε να περιγράψουμε συμπεράσματα για τον μέσο του πληθυσμού. Πρακτικά είναι μια προέκταση του Γενικευμένου Γραμμικού μοντέλου στα διαχρονικά δεδομένα. Δεν κάνουμε υπόθεση για την κατανομή της εξαρτημένης μεταβλητής, αλλά καθορίζουμε το μοντέλο που ο μέσος θα ακολουθεί. Η εξαρτημένη μεταβλητή μπορεί να είναι είτε διακριτή, είτε συνεχής και επιπλέον μπορεί να επεξεργαστεί και unbalanced δεδομένα (δεδομένα

μη ισορροπημένης έρευνας). Συμβολίζοντας παρόμοια με προηγουμένως έχουμε

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ . \\ . \\ . \\ . \\ Y_{in_i} \end{pmatrix} \tag{2.42}$$

ο πίνακας της εξαρτημένης μεταβλητής του i-οστού ατόμου, όπου n_i ο αριθμός των παρατηρήσεων που έχει το συγκεκριμένο άτομο, με $n_i \leq n$, και

$$X_{ij} = \begin{pmatrix} X_{ij1} \\ X_{ij2} \\ . \\ . \\ . \\ . \\ X_{ijp} \end{pmatrix} \tag{2.43}$$

ο πίνακας των περιγραφικών μεταβλητών του συγκεκριμένου ατόμου, όπου p είναι το πλήθος τους, και X_{ijk} είναι ένα διάνυσμα με τις τιμές της $k-$οστής περιγραφικής μεταβλητής του ατόμου αυτού για όλους τους χρόνους.
Τα είδη των περιγραφικών μεταβλητών χωρίζονται σε αυτά που μπορούν να αλλάζουν στον χρόνο (π.χ. κατάσταση ασθενή, ηλικία) και σε αυτά που δεν αλλάζουν (π.χ. φύλο, είδος θεραπείας).

Τα μοντέλα που αναφέρονται στο μέσο του πληθυσμού έχουν τρία κομμάτια που πρέπει να καθοριστούν:
1. Ο μέσος είναι της μορφής

$$g(\mu_{ij}) = \eta_{ij} = X'_{ij}\beta, \tag{2.44}$$

όπου ο $\mu_{ij} = E[Y_{ij}|X_{ij}]$ εξαρτάται από το γραμμικό η_{ij} μέσω από την συνάρτηση σύνδεσμο (link function) $g(.)$.

Παρακάτω δίνεται ο πίνακας με τα πιο συνηθισμένα link function

$Distribution$	$u(\mu)$	$link$
$Normal$	1	$\mu = \eta$
$Bernoulli$	$\mu(1-\mu)$	$log(\frac{\mu}{1-\mu}) = \eta$
$Poisson$	μ	$log(\mu) = \eta$

2. Η διασπορά έχει την μορφή

$$Var(Y_{ij}) = \phi u(\mu_{ij}), \tag{2.45}$$

όπου το $u(\mu_{ij})$ είναι μια γνωστή συνάρτηση του μέσου και το ϕ είναι ο $scale$ παράμετρος (παράμετρος όχλησης) που μπορεί να είναι γνωστός ή να θέλει εκτίμηση, διαφέρει ανάλογα την περίπτωση και μπορεί να εξαρτάται απ΄τον χρόνο.

3. Η σχέση εσωτερικά μεταξύ των διαχρονικών τιμών της μεταβλητής είναι μια συνάρτηση διαφορετικών παραμέτρων ας πούμε a, που μπορεί και αυτή να εξαρτάται από τους μέσους, η μοντελοποίηση της γίνεται με τους τρόπους που αναφέραμε στην παράγραφο 2.3.

Το γεγονός ότι δεν κάνουμε υπόθεση για την κατανομή του Y_{ij} οδηγεί στην μέθοδο Generalized Estimating Equations (GEE). Ο μέσος και η εσωτερική σχέση των μεταβλητών μοντελοποιείται ξεχωριστά. Έτσι η συνολική σχέση των τιμών της μεταβλητής μεταφράζεται σε έναν ενιαίο πίνακα συνδιασποράς με τον οποίο εργαζόμαστε

$$V_i = \sqrt{A_i} Corr(Y_i) \sqrt{A_i}, \tag{2.46}$$

όπου ο A_i είναι ο διαγώνιος πίνακας με τιμές τις $Var(Y_{ij}) = \phi u(\mu_{ij})$.
Ο εκτιμητής των συντελεστών των παραμέτρων β υπολογίζεται ελαχιστοποιώντας ως προς β το

$$\sum_{i=1}^{N} (y_i - \mu_i(\beta))' V_i^{-1} (y_i - \mu_i(\beta)) \tag{2.47}$$

όπου μ_i είναι ο πίνακας με τιμές $\mu_{ij} = g^{-1}(X'_{ij}\beta)$ και ο V_i θεωρείται γνωστός. Αυτό οδηγεί στις εξισώσεις

$$\sum_{i=1}^{N} D'_i V_i^{-1}(y_i - \mu_i) = 0, \tag{2.48}$$

όπου V_i είναι ο πίνακας συνδιασποράς που αναφέραμε στην σχέση (2.46) και $D_i = \frac{d\mu_i}{d\beta}$. Ο αλγόριθμος για την εύρεση των εκτιμητριών είναι ο εξής

Βήμα 1 Με βάση τις εκτιμήσεις των a, ϕ, εκτίμησε το V_i και έπειτα εκτίμησε το β με χρήση της σχέσης (2.48)

Βήμα 2 Με βάση την εκτίμηση του β εκτίμησε τα a, ϕ με χρήση της σχέσης για τα τυποποιημένα κατάλοιπα

$$e_{ij} = \frac{Y_{ij} - \hat{\mu}_{ij}}{\sqrt{u(\hat{\mu}_{ij})}}, \tag{2.49}$$

Επανέλαβε εώς ότου επιτευχθεί σύγκλιση.

Ιδιότητες

1. Η εκτιμήτρια $\hat{\beta}$ είναι συνεπής εκτιμήτρια του β. Αυτό ισχύει ανεξαρτήτως της επιλογής του V_i.

2. Για μεγάλο δείγμα το $\hat{\beta}$ ακολουθεί πολυδιάστατη κατανομή με μέσο β και

$$Cov(\hat{\beta}) = B^{-1} M B^{-1}, \tag{2.50}$$

όπου

$$B = \sum_{i=1}^{N} D'_i V_i^{-1} D_i, \tag{2.51}$$

$$M = \sum_{i=1}^{N} D'_i V_i^{-1} Cov(Y_i) V_i^{-1} D_i. \tag{2.52}$$

Οι παραπάνω πίνακες μπορούν να εκτιμηθούν αντικαθιστώντας τα a, β, ϕ από τις

εκτιμητριές τους και αντικαθιστώντας το $Cov(Y_i) = \Sigma_i$ με το $(Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)$.

3. Έτσι έχουμε τον εκτιμητή σάντουιτς

$$\hat{Cov}(\hat{\beta}) = (\sum_{i=1}^{N} D_i' V_i^{-1} D_i)^{-1} (\sum_{i=1}^{N} D_i' V_i^{-1} (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)' V_i^{-1} D_i) (\sum_{i=1}^{N} D_i' V_i^{-1} D_i)^{-1}. \quad (2.53)$$

4. Αν έχουμε διαλέξει το κατάλληλο μοντέλο $V_i = \Sigma_i$ και $\hat{Cov}(\hat{\beta}) = B^{-1}$.

Πλεονεκτήματα

Ο εκτιμητής $\hat{\beta}$ του GEE

1. Είναι όσο ακριβής όσο και ο εκτιμητής μέγιστης πιθανοφάνειας.
2. Είναι συνεπής ακόμα και αν έχουμε κάνει λάθος στην επιλογή δομής του πίνακα των συνδιασπορών.
3. Μπορούμε ακόμα και σε αυτήν την περίπτωση να εκτιμήσουμε καλά* τυπικά σφάλματα μέσω του εκτιμητή σάντουιτς.

*Όταν ο αριθμός των ατόμων δεν είναι πολύ μεγάλος σε σύγκριση με αυτό των διαχρονικών μετρήσεων τότε δεν πρέπει να βασιζόμαστε στον εκτιμητή σάντουιτς. Σε αυτήν την περίπτωση είναι καλύτερο να προτιμήσουμε την συνδιασπορά

$$Cov(\hat{\beta}) = B^{-1},$$

η οποία μας παρέχει καλές εκτιμήσεις όταν έχουμε επιλέξει σωστό πίνακα συνδιακύμανσης.

2.5 Παράδειγμα Ανάλυσης διαχρονικών δεδομένων

Όπως και στο προηγούμενο κεφάλαιο, έτσι κι εδώ θα παραθέσουμε ένα παράδειγμα μιας έρευνας η οποία θα χρησιμοποιεί longitudinal δεδομένα. Θα χρησιμοποιήσουμε

ένα υποσύνολο από πραγματικά δεδομένα της έρευνας που διεξήχθει το 1982 από τον Alfrent Sommer και τους συνεργάτες του στο Άσεχ της Ινδονησίας [11]. Το συνολικό δείγμα της έρευνας προήλθε από 3000 άτομα προσχολικής ηλικίας, τα οποία εξετάστηκαν ανά 3 μήνες για πάνω από 6 εξετάσεις. Κύριος σκοπός της έρευνας ήταν τα αποτελέσματα από την έλλειψη βιταμίνης Α στα παιδιά αυτής της ηλικίας. Τα στοιχεία που συλλέχθησαν ήταν η ύπαρξη λοίμωξης του αναπνευστικού, η ύπαρξη ξηροφθαλμίας, το ύψος και το βάρος των παιδιών. Το δείγμα που θα μελετήσουμε προέρχεται από 250 παιδιά, σε 6 διαδοχικές εξετάσεις ανά εποχή (δηλαδή κάθε εξέταση ήταν την επόμενη εποχή), σε σύνολο 1200 παρατηρήσεων.
Οι μεταβλητές με τις οποίες θα δουλέψουμε είναι οι εξής:

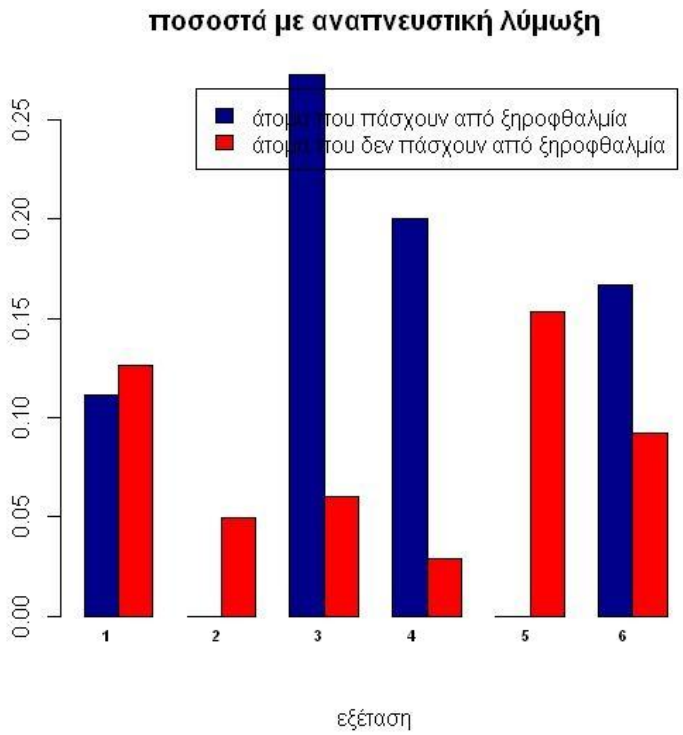
1. ID: Ο κωδικός αριθμός του ατόμου
2. time: Ο αριθμός της εξέτασης (1-6)
3. resp: Ύπαρξη αναπνευστικής λοίμωξης
4. agem: Ηλικία σε μήνες (-32 εώς 50, κεντραρισμένο στο 36)
5. xerop: Ύπαρξη ξηροφθαλμίας
6. sex: Φύλο (αγόρι=0)
7. height: Ύψος για την ηλικία, σαν ποσοστό της κλίμακας του National Center Health Statistics , (κεντραρισμένο στο 90%)
8. stunted: Δείκτρια του αν το άτομο είναι μη ανεπτυγμένο κανονικά, (προέρχεται από το αν το άτομο είναι μικρότερο από 85% στην προαναφερθήσα κλίμακα ύψος ανά ηλικία του NCHS)
9. season: Η εποχή που έγινε η εξέταση (Χειμώνας=1, Άνοιξη=2, Καλοκαίρι=3, Φθινόπωρο=4)

Η εξαρτημένη μεταβλητή θα είναι η δίτιμη resp της ύπαρξης αναπνευστικής λοίμωξης, ενώ οι περιγραφικές θα είναι αυτές της ηλικίας, ύπαρξης ξηροφθαλμίας, του φύλου, του ύψους για την ηλικία, της δείκτριας ανάπτυξης και της εποχής.

	ID	resp	agem	xeropht	sex	height	stunted	time	season
1	121013	0	31	0	0	-3	0	1	2
2	121013	0	34	0	0	-3	0	2	3
3	121013	0	37	0	0	-2	0	3	4
4	121013	0	40	0	0	-2	0	4	1
5	121013	1	43	0	0	-2	0	5	2
6	121013	0	46	0	0	-3	0	6	3
7	121113	0	-9	0	1	2	0	1	2
8	121113	0	-6	0	1	0	0	2	3
9	121113	0	-3	0	1	-1	0	3	4
10	121113	0	0	0	1	-2	0	4	1
11	121113	1	3	0	1	-3	0	5	2
12	121113	0	6	0	1	-3	0	6	3
13	121114	0	-26	0	0	8	0	1	2
14	121114	0	-23	0	0	5	0	2	3
15	121114	0	-20	0	0	3	0	3	4
16	121114	1	-17	0	0	0	0	4	1
17	121114	1	-14	0	0	0	0	5	2
18	121114	0	-11	0	0	0	0	6	3
19	121140	0	-19	0	1	5	0	1	2
20	121140	0	-16	1	1	4	0	2	3
21	121215	0	0	0	1	-10	1	1	2
22	121215	0	3	0	1	-8	1	2	3
23	121215	0	6	0	1	-7	1	3	4
24	121215	0	9	0	1	-9	1	4	1

Σχήμα 2.2: Ενδεικτικά δίνονται οι πρώτες 24 παρατηρήσεις του παραδείγματος

Θα ξεκινήσουμε την ανάλυση μας βγάζοντας συμπεράσματα από την γραφική μελέτη των δεδομένων, φτιάχνοντας κυρίως τα γραφήματα των ποσοστών της ύπαρξης αναπνευστικής λοίμωξης για τις διάφορες τιμές της κάθε μεταβλητής ανά εξέταση. Ενδεικτικά δίνουμε το παρακάτω γράφημα, όπου έχουμε τα ποσοστά των ατόμων με ξηροφθαλμία και χωρίς ξηροφθαλμία που έχουν αναπνευστική λοίμωξη και από εδώ μπορούμε να βγάλουμε κάποια πρώτα συμπεράσματα (τα υπόλοιπα γραφήματα υπάρχουν στο παράρτημα Σχήματα A.1 - A.4). Από το συγκεκριμένο γράφημα δεν μπορούμε να βγάλουμε κάποιο εμφανές συμπέρασμα για την σχέση ξηροφθαλμίας - αναπνευστικής λοίμωξης, ωστόσο πρέπει να παρατηρήσουμε ότι στα δεδομένα μας η ξηροφθαλμία παρουσιάζεται μόλις 55 φορές στα 1200 και αυτό μπορεί να επηρεάζει την στατιστική δύναμη, όσον αφορά τη συγκεκριμένη μεταβλητή.



Σχήμα 2.3: Ποσοστά των ατόμων με ξηροφθαλμία και χωρίς ξηροφθαλμία που έχουν αναπνευστική λοίμωξη

Από τα αντίστοιχα και ανάλογα γραφηματα για τις λοιπές μεταβλητές παρατηρούμε

- από τα γραφήματα σχετικά με την ηλικία ότι στις ηλικίες 0 -4 υπάρχει μια τάση τα ποσοστά να είναι μεγαλύτερα σε σχέση με τις ηλικίες 4-7,
- από το γράφημα σχετικά με το φύλο του ατόμου ότι υπάρχει μια τάση στα αγόρια τα ποσοστά στην εμφάνιση λοίμωξης να είναι μεγαλύτερα,
- από το γράφημα σχετικά με την εποχή (δεν βάλαμε το συνολικό ποσοστό ανά εποχή γιατί η Άνοιξη και το Καλοκαίρι εμφανίζονται δυο φορές, ενώ οι άλλες δύο μια φορά), παρατηρούμε ότι φαίνεται να υπάρχει κάποια σχέση της εποχής με την ύπαρξη της λοίμωξης του αναπνευστικού (πχ την Άνοιξη τα ποσοστά φαίνονται σταθερά μεγαλύτερα απ΄τις άλλες εποχές),
- από τα γραφήματα σχετικά με το ύψος ανά ηλικία και την ύπαρξη αναπνευστικής λοίμωξης γενικά δεν μπορούμε να διακρίνουμε κάτι,

- από τα γραφήματα σχετικά με το αν το άτομο δεν έχει αναπτυχθεί σωστά, γενικά δεν μπορούμε να διακρίνουμε κάποια τάση.

Όπως αναφέραμε προηγουμένως στο κεφάλαιο, υπάρχουν δύο τρόποι μοντελοποίησης αυτός που αναφέρεται στο κάθε άτομο ξεχωριστά και αυτός που αναφέρεται στον μέσο του πληθυσμού. Θα μοντελοποιήσουμε αρχικά με τον πρώτο τρόπο και πιο συγκεκριμένα με την χρήση των Γενικευμένων Γραμμικών Μοντέλων Μικτών Επιδράσεων GLMM.

Το μοντέλο μας θα έχει σαν εξαρτημένη μεταβλητή (Y) την μεταβλητή ύπαρξης αναπνευστικής λοίμωξης resp και σαν επεξηγηματικές αυτές της ηλικίας, της ύπαρξης ξηροφθαλμίας, του φύλου, του αν το άτομο είναι ανεπτυγμένο σωστά, του ύψους και της εποχής (agem,xeropht,sex,stunted,height,season). Προφανώς η συνδετική συνάρτηση link function θα είναι από κατανομή Bernoulli, αφού τα Y είναι 0 - 1.

Αρχικά θα ελέγξουμε την κάθε μεταβλητή ξεχωριστά για να δούμε ποιες είναι στατιστικά σημαντικές στο μοντέλο, όταν είναι μόνο αυτές σαν παράμετροι. Στην συνέχεια και αφού έχουμε υπ΄όψην ποιες είναι στατιστικά σημαντικές (και αφού στο συγκεκριμένο παράδειγμα ο αριθμός των περιγραφικών μεταβλητών είναι μικρός, οπότε δεν είναι ανάγκη να απορρίψουμε ακόμα μεταβλητές) θα τοποθετήσουμε όλες μαζί στο μοντέλο. Έπειτα θα καθορίσουμε με ελέγχους στατιστικής σημαντικότητας σε ποιες μεταβλητές και αν στο σταθερό όρο θα τοποθετήσουμε τις τυχαίες επιδράσεις. Στο συγκεκριμένο παράδειγμα δοκιμάσαμε να τοποθετήσουμε στην ηλικία (η οποία αλλάζει με τον χρόνο) και στον σταθερό όρο και τελικά αποφανθήκαμε πως θα έχουμε μόνο στον σταθερό όρο τυχαία επίδραση. Τώρα για την επιλογή του τελικού μοντέλου θα χρησιμοποιήσουμε την διαδικασία της αντίστροφης απόρριψης (backward elimination), όπου απορρίπτουμε κάθε φορά την πιο στατιστικά ασήμαντη μεταβλητή (αν υπάρχει) και επαναλαμβάνουμε μέχρι να φτάσουμε σε ένα μοντέλο που όλες οι μεταβλητές θα είναι στατιστικά σημαντικές. Δουλεύοντας λοιπόν με αντίστροφη απόρριψη θα εξαιρέσουμε αρχικά σαν μην στατιστικά σημαντική τη μεταβλητή του αν το άτομο είναι σωστά ανεπτυγμένο με το μεγαλύτερο p _value =0.7. Οπότε καταλήγουμε στο μοντέλο με τις 5 επεξηγηματικές μεταβλητές, από το οποίο κάνοντας πάλι τις σχετικές μαθηματικές μεθόδους εκτίμησης παρατηρούμε ότι το μεγαλύτερο p _value =0.23 το έχει η μεταβλητή ύπαρξης ξηροφθαλμίας την οποία και θα εξαιρέσουμε σαν μη στατιστικά σημαντική. Στη συνέχεια αφού πρώτα παρατηρήσουμε

πως σε επίπεδο στατιστικής σημαντικότητας a=5% η μεταβλητή του φύλου θεωρείται στατιστικά μη σημαντική, αλλά σε a=10% όχι, με p _value 0.08641 και αφού ελέγξουμε και την στατιστική σημαντικότητα των ψευδομεταβλητών της εποχής (στους 3 βαθμούς ελευθερίας) καταλήγουμε στο μοντέλο:

$$\eta_{ij} = b_i - 3.400146 - 0.032124 \cdot agem_{ij} - 0.474769 \cdot sex_{ij} - 0.058191 \cdot height_{ij}$$

$$+1.43196 \cdot seas2_{ij} + 0.703544 \cdot seas3_{ij} + 0.696069 \cdot seas4_{ij}$$

όπου η τυχαία επίδραση ακολουθεί

$$b_i \sim N(0, 0.87859).$$

Τα συμπεράσματα μας με βάση αυτό το μοντέλο είναι αρχικά για τις μεταβλητές που απορρίφθηκαν:

- (stunted): το αν είναι ένα άτομο μη ανεπτυγμένο σωστά ή όχι δεν φαίνεται να επηρεάζει την πιθανότητα εμφάνισης λοίμωξης του αναπνευστικού
- (xeropth): το αν ένα άτομο πάσχει από ξηροφθαλμία δε φαίνεται να επηρεάζει την πιθανότητα εμφάνιση της αναπνευστικής λοίμωξης. Όμως πρέπει να αναφερθεί το γεγονός ότι στο δείγμα τα κρούσματα ξηροφθαλμίας ήταν πολύ λίγα, δηλαδή το δείγμα για την ξηροφθαλμία ήταν πολύ μικρό για ασφαλή συμπεράσματα.

Για τις μεταβλητές που συμπεριλήφθηκαν στο μοντέλο συμπεράνουμε ότι:

- (agem) (αρνητικό πρόσημο στον συντελεστή της), η ηλικία έχει αρνητική σχέση με την ύπαρξη αναπνευστικής λοίμωξης, δηλαδή όσο αυξάνεται η ηλικία μειώνεται και η πιθανότητα για αναπνευστική λοίμωξη. Υπολογίζεται μέσω της αντίστροφης συνάρτησης της logODDS ($p = \frac{e^{\eta}}{1+e^{\eta}}$) ότι είναι γύρω στο 3% μείωση της πιθανότητας εμφάνισης για κάθε επιπλέον μήνα, με σταθερές όλες τις άλλες μεταβλητές.
- (sex): τα αγόρια έχουν μεγαλύτερη πιθανότητα εμφάνισης λοίμωξης του αναπνευστικού (υπενθυμίζουμε ότι σε ε.σ.σ. 5 % το φύλο μπορεί να θεωρηθεί μη στατιστικά σημαντικό)

- (height): Το ύψος και η εμφάνιση λοίμωξης έχουν αρνητική σχέση, όσο πιο κοντά στο σωστό ύψος κατά τον NCHS βρίσκεται το παιδί, η πιθανότητα λοίμωξης στο αναπνευστικό είναι μικρότερη.
- (season): Στις εποχές η σειρά που η πιθανότητα είναι μεγαλύτερη για ύπαρξη λοίμωξης είναι η εξής:
  Ανοιξη,Καλοκαίρι,Φθινόπωρο,Χειμώνας.
  Σημειώνεται ότι το Φθινόπωρο με το Καλοκαίρι έχουν σχεδόν την ίδια πιθανότητα.

Συνεχίσαμε την ανάλυσή μας, χρησιμοποιώντας τώρα Γενικευμένες εξισώσεις εκτίμησης, δηλαδή GEE μοντέλο, που θα μας δώσει συμπεράσματα για τον μέσο του πληθυσμού.
Για την συνδιακύμαση επιλέξαμε το μοντέλο μας την μορφή compound symmetry, δηλαδή ίδιες διασπορές και $Corr(Y_{ij}, Y_{ik}) = p$ για τα στοιχεία ενός ατόμου και με παρόμοια μεθοδολογία όπως στο προηγούμενο μοντέλο κι εδώ απορρίφθησαν οι μεταβλητές της ύπαρξης ξηροφθαλμίας και της μη σωστής ανάπτυξης και φτάσαμε στο μοντέλο το οποίο οι εκτιμήσεις των συντελεστών δίνονται παρακάτω:

	$\hat{\beta}$	$naiveSE$
$intercept$	-3.05429782	0.38762559
$agem$	-0.02958081	0.00663348
sex	-0.43145205	0.23715274
$height$	-0.05417625	0.02109744
$seas2$	1.36692359	0.39873082
$seas3$	0.68892760	0.41811553
$seas4$	0.66816920	0.46566441

Παρατηρούμε ανάλογα συμπεράσματα με τα προηγούμενα αλλά εδώ αναφέρονται στον μέσο του πληθυσμού και όχι για το κάθε άτομο ξεχωριστά. Κι εδώ η ηλικία έχει αρνητική σχέση, το φύλο αυτή τη φορά είναι στατιστικά σημαντικό και σε ε.σ.σ 5% και μας δείχνει όπως και πριν ότι τα αγόρια κινδυνεύουν πιο πολύ, το ύψος έχει αρνητική σχέση και οι εποχές έχουν κι αυτές παρόμοια συμπεράσματα με το προηγούμενο μοντέλο.

Οι τρόποι ελέγχου καλής προσαρμογής των μοντέλων, δεν κρίνεται σκόπιμο να αναφερθούν στην παρούσα διπλωματική εργασία, αφού σκοπός της υποενότητας είναι να δώσει ένα παράδειγμα του πως εφαρμόζονται αυτά τα οποία αναφέρονται στο κεφάλαιο και όχι μια πλήρη ανάλυση του κεφαλαίου των αναλύσεων διαχρονικών δεδομένων. Τελειώνοντας να αναφέρουμε ότι οι μεθοδολογίες εύρεσης κατάλληλου μοντέλου, όπως και το ποιο μοντέλο θα χρησιμοποιήσουμε είναι ευθύνη του εκάστοτε ερευνητή, της a priori γνώσης του σε κάποια θέματα και της διαίσθησής του, δεν υπάρχουν πάντα σταθερές τακτικές και μόνο ένας τρόπος μοντελοποίησης.

Κεφάλαιο 3

Ανάλυση Επιβίωσης

Ανάλυση Επιβίωσης είναι ο κλάδος της στατιστικής ο οποίος ασχολείται με την περιγραφή και την ανάλυση των δεδομένων σε σχέση με τον χρόνο από μια καλώς ορισμένη αρχική χρονική στιγμή εώς την πραγματοποίηση κάποιου συγκεκριμένου γεγονότος ή το τέλος ενός ορισμένου χρόνου. Στη Βιοστατιστική η αρχική χρονική στιγμή συνηθώς αντιστοιχεί στην εισαγωγή ενός ατόμου σε κάποια ιατροφαρμακευτική έρευνα, όπως π.χ. μια κλινική μελέτη για τη σύγκριση δύο φαρμάκων. Η έρευνα μπορεί να ελέγχει το αν θα παρουσιαστούν κάποια συμπτώματα, κάποιο γεγονός το οποίο τίθεται σε ελέγχο και λοιπά σχετικά θέματα, ενώ αν το καταληκτικό γεγονός είναι ο θάνατος τότε και κυριολεκτικά αναφερόμαστε σε χρόνους επιβίωσης (survival time). Στην ανάλυση επιβίωσης δεν έχουμε πάντα σαν καταληκτικό γεγονός το θάνατο για αυτό οι παρατηρήσεις λέγονται δεδομένα μέχρι το γεγονός (time to event data).

Τα δεδομένα επιβίωσης γενικά δεν είναι συμμετρικά κατανεμημένα. Συνήθως ένα ιστόγραμμα από χρόνους επιβίωσης κάποιας ομάδας ατόμων με ίδια χαρακτηριστικά τείνει να έχει θετικό συντελεστής λοξότητας, δηλαδή είναι λοξή προς τα δεξιά και περιέχει το μεγαλύτερο κομμάτι των παρατηρήσεων. Οπότε θα ήταν παράλογο να υποθέσουμε κανονική κατανομή. Αυτός είναι και ένας λόγος που δεν χρησιμοποιούμε τα συνηθισμένα μοντέλα.

3.1 Βασικές έννοιες

3.1.1 Λογοκριμένα δεδομένα (Censored data)

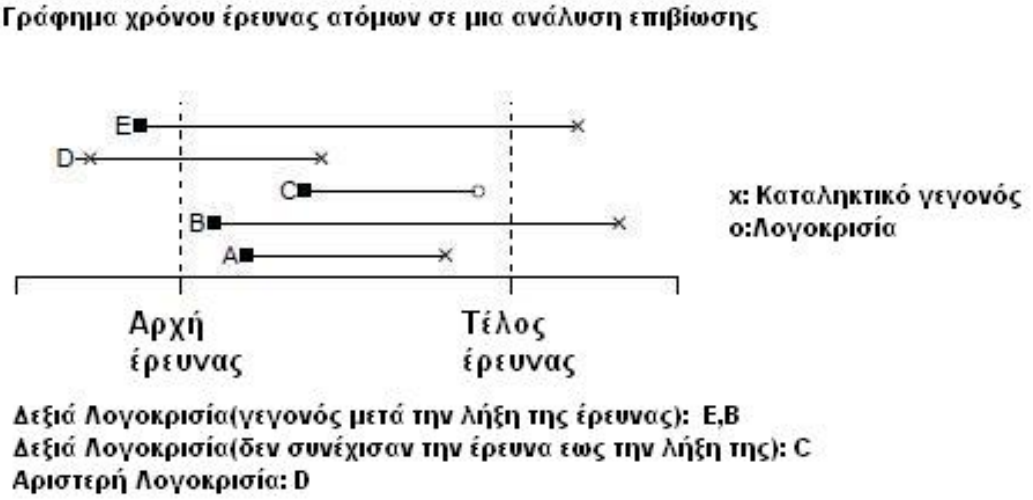
Ένα ακόμα χαρακτηριστικό που συναντάμε στα δεδομένα επιβίωσης και κάνουν τις συνηθισμένες στατιστικές μεθόδους μη εφαρμόσιμες είναι ότι τα δεδομένα είναι συνήθως censored (λογοκριμένα). Ένας χρόνος επιβίωσης ενός ατόμου λέγεται λογοκριμένος όταν το καταληκτικό γεγονός δεν παρατηρηθεί μέχρι την λήξη της έρευνας. Αυτό μπορεί να οφείλεται στο γεγονός ότι το άτομο είναι ακόμα ζωντανό όταν ολοκληρώνεται η έρευνα (ή στην γενική περίπτωση δεν του έχει συμβεί το καταληκτικό γεγονός της έρευνας), είτε γιατί το άτομο σταμάτησε να συμμετέχει στην έρευνα οπότε από την στιγμή που σταμάτησε και ύστερα η κατάληξή του είναι άγνωστη. Ακόμα ένας άλλος λόγος λογοκριμένων δεδομένων είναι όταν σαν παράδειγμα ο ασθενής πεθαίνει στην διάρκεια της έρευνας αλλά από αίτια που δεν έχουν να κάνουν με την έρευνα. Στην τελευταία περίπτωση πρέπει να βεβαιωθούμε ότι δεν υπάρχει όντως καμία σύνδεση με το αντικείμενο της έρευνα ή αν μέσα από τέτοιες περιπτώσεις παράγονται καινούριες αναλύσεις επιβίωσης για άλλα αίτια θανάτου.

Σε όλες αυτές τις περιπτώσεις το άτομο που μπαίνει στην έρευνα τη χρονική στιμή t_0 παθαίνει το καταληκτικό γεγονός στην χρονική στιγμή $t_0 + t$. Το t είναι άγνωστο, είτε γιατί το άτομο δεν έχει πάθει ακόμα το γεγονός, είτε γιατί το άτομο δεν συμμετέχει πλέον στην θεραπεία. Αν ο τελευταίος γνωστός χρόνος που γνωρίζουμε πως δεν έχει πάθει το γεγονός είναι η στιγμή $t_0 + c$, τότε το c λέγεται ένα λογοκριμένος χρόνος επιβίωσης. Αυτό το είδος λογοκριμένων δεδομένων όπου το άτομο συμμετέχει στην έρευνα και έχει τον τελευταίο του γνωστό χρόνο επιβίωσης και το καταληκτικό γεγονός θα συμβεί `δεξία` από αυτή την στιγμή, λέγεται δεξιά λογοκρισία.

Άλλη μορφή λογοκρισίας είναι η λεγόμενη αριστερή λογοκρισία, που συναντάται όταν το γεγονός που περιμένουμε συμβεί πριν το άτομο εισέλθει στην έρευνα. Σαν παράδειγμα μπορούμε να θεωρήσουμε ένα άτομο το οποίο έχει χειρουργηθεί για κάποια μορφή καρκίνου, για την οποία δεν γνωρίζουμε ποια στιγμή συνέβει η εμφάνιση αυτού και ελέγχουμε σαν γεγονός αν στο άτομο έχει επανέλθει η συγκεκριμένη μορφή καρκίνου. Η μορφή αυτού του είδους λογοκρισίας είναι πιο σπάνια από την δεξιά.

Τέλος υπάρχει και η λογοκρισία διαστήματος, αυτή συμβαίνει όταν ο έλεγχος σε

μια έρευνα γίνεται ανά χρονικά διαστήματα και κάτι συμβεί μέσα στο διάστημα αυτό αλλά δεν γνωρίζουμε την ακριβή στιγμή. Σαν παράδειγμα αναφέρουμε μια έρευνα όπου κάποια άτομα ελέγχονται για μια πάθηση κάθε έξι μήνες, όταν το άτομο βρεθεί με την πάθηση, δεν γνωρίζουμε ακριβώς πότε συνέβει μέσα στο διάστημα των έξι μηνών.



Σχήμα 3.1: Γραφική περιγραφή των ειδών λογοκρισίας

3.1.2 Συνάρτηση επιβίωσης - Συνάρτηση κινδύνου

Ας υποθέσουμε ότι το T συμβολίζει τον χρόνο επιβίωσης ενός ατόμου, με συνάρτηση κατανομής $F(T)$ και πυκνότητας $f(t)$, προφανώς αφού αναφερόμαστε σε χρόνο η τυχαία μεταβλητή δεν μπορεί να είναι αρνητική. Τότε ορίζεται η συνάρτηση επιβίωσης η οποία υποδηλώνει την πιθανότητα κάποιο άτομο να έχει επιβιώσει μετά την χρονική στιγμή t, δηλαδή έχουμε την σχέση

$$S(t) = P(T \geq t) = 1 - F(t). \tag{3.1}$$

Μια άλλη πολύ χρήσιμη σχέση που μας ενδιαφέρει στην ανάλυση επιβίωσης είναι η συνάρτηση κινδύνου (hazard function), η πιθανότητα να πεθάνει το άτομο (ή να

πάθει το καταληκτικό γεγονός της έρευνάς μας στην γενική περίπτωση) δεδομένου ότι έχει επιβιώσει μέχρι στιγμής. Αυτό μαθηματικά μεταφράζεται στη σχέση:

$$h(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt | T \geq t)}{dt} \tag{3.2}$$

από την οποία μπορούμε να καταλήξουμε στις παρακάτω σχέσεις:

$$h(t) = \frac{f(t)}{S(t)}, \tag{3.3}$$

$$h(t) = -\frac{dlog(S(t))}{dt}, \tag{3.4}$$

$$S(t) = exp\{-H(t)\}, \tag{3.5}$$

όπου

$$H(t) = \int_0^t h(u)du, \tag{3.6}$$

η οποία ονομάζεται και αθροιστικός κίνδυνος (cumulative hazard). Λόγω της (3.5) ισχύει:

$$H(t) = -log(S(t)). \tag{3.7}$$

Στην ανάλυση επιβίωσης η συνάρτηση επιβίωσης και κινδύνου εκτιμούνται από τους χρόνους επιβίωσης του δείγματος. Υπάρχουν μέθοδοι εκτίμησης με χρήση υπόθεσεων κατανομών για τον χρόνο T αλλά και χωρίς, για τους οποίους θα μιλήσουμε στην συνέχεια.

3.2 Μη παραμετρικοί μέθοδοι εκτίμησης

Ένα πρώτο βήμα στην ανάλυση επιβίωσης είναι η εύρεση τιμών και η δημιουργία γραφικών παραστάσεων των χρόνων επιβίωσης που θα μας παρέχουν κάποια αρχικά συμπεράσματα και θα μας δείξουν ενδεχομένως προς τα που θα κινηθούμε. Αυτές οι μεθόδοι λέγονται απαραμετρικές ή χωρίς ανάγκη κατανομής, αφού δεν κάνουμε υποθέσεις για αυτήν.

3.2.1 Εκτιμώντας την συνάρτηση επιβίωσης

Στην απλή περίπτωση όπου δεν έχουμε καθόλου λογοκρισία στα δεδομένα μας η εμπειρική συνάρτηση επιβίωσης δίνεται σαν ο λόγος των ατόμων που έχουν επιβιώσει μέχρι μια χρονική στιγμή t, προς το σύνολο των ατόμων της ανάλυσής μας. Μια άλλη μέθοδος εκτίμησης είναι η αναλογιστική εκτίμηση (actuarial ή life-table), στην οποία χωρίζουμε την χρονική περίοδο σε διαμερίσεις $j = 1, 2, ...m$, όχι απαραίτητα ίσων μεγεθών και συμβολίζουμε με d_j, c_j τους αριθμούς των θανάτων και των λογοκριμένων δεδομένων αντίστοιχα στη διαμέριση μεταξύ των χρονικών στιγμών t'_j, t'_{j+1} και n_j το πλήθος των ατόμων που είναι ζωντανοί στην αρχή της j διαμέρισης. Υποθέτουμε ότι η λογοκρισία κατανέμεται ομοιόμορφα στην j διαμέριση άρα ο μέσος αριθμός που βρίσκονται σε κίνδυνο σε αυτή την διαμέριση είναι

$$n'_j = n_j - 0.5c_j. \tag{3.8}$$

Οπότε η πιθανότητα επιβίωσης σε αυτό το διάστημα είναι $(n'_j - d_j)/n'_j$ και έτσι για να βρούμε την εκτίμηση θεωρώντας ότι είμαστε στην αρχή της k διαμέρισης,έχουμε

$$S^*(t) = \prod_{j=1}^{k} \frac{n'_j - d_j}{n'_j} \tag{3.9}$$

για $t \in [t'_k, t'_{k+1}),\ k = 1, 2, ..., m.$

Το τυπικό σφάλμα δίνεται από την σχέση

$$se(S^*(t)) \approx S^*(t)\sqrt{\sum_{j=1}^{k} \frac{d_j}{(n'_j - d_j)n'_j}}. \tag{3.10}$$

Η πιο συνηθισμένη μέθοδος εκτίμησης λογοκριμένων δεδομένων είναι αυτή με χρήση του εκτιμητή Kaplan-Meier. Για να παράγουμε τον εκτιμητή χωρίζουμε και εδώ τον χρόνο σε διαμερίσεις, αλλά η κάθε διαμέριση είναι έτσι φτιαγμένη ώστε ένας χρόνος που συνέβει το καταληκτικό γεγονός να περιέχεται στο διαστημά της και αυτός να βρίσκεται στην αρχή του διαστήματος. Συμβολίζουμε τους χρόνους που ξεκινάνε οι διαμερίσεις οι οποίοι είναι και οι χρόνοι που συμβαίνει κάποιος θάνατος

με $t(1) < t(2) < ... < t(r)$, στο διάστημα $[t_0, t(1))$ δεν συμβαίνει κανένας θάνατος. Με παρόμοια λογική με τον προηγούμενο εκτιμητή έχουμε τον τύπο

$$\hat{S}(t) = \prod_{j=1}^{k} \frac{n_j - d_j}{n_j}, \tag{3.11}$$

για $t \in [t(k), t(k+1))$, $k = 1, 2, ..., r$ και $\hat{S}(t) = 1$ για $t \in [t_0, t(1))$.
Επισημαίνουμε ότι την χρονική στιγμή $t(j)$ μπορεί να συμβούν πάνω από ένας θάνατοι (καταληκτικό γεγονός).

Το τυπικό σφάλμα δίνεται από την σχέση

$$se(\hat{S}(t)) \approx \hat{S}(t)\sqrt{\sum_{j=1}^{k} \frac{d_j}{(n_j - d_j)n_j}}. \tag{3.12}$$

Ένας εναλλακτικός εκτιμητής είναι ο Nelson Aalen ο οποίος δίνεται από τη σχέση

$$S^{\sim}(t) = \prod_{j=1}^{k} e^{-\frac{d_j}{n_j}}, \tag{3.13}$$

με τυπικό σφάλμα

$$se(S^{\sim}(t)) = S^{\sim}(t)\sqrt{\sum_{j=1}^{k} d_j/n_j^2}. \tag{3.14}$$

3.2.2 Εκτιμώντας την συνάρτηση κινδύνου

Ας υποθέσουμε πως έχουμε διαμερίσει τον χρόνο όπως στην αναλογιστική εκτίμηση της συνάρτησης επιβίωσης, μια λογική εκτίμηση του αναμενόμενου κινδύνου στην μονάδα του χρόνου σε μια διαμέριση είναι ο λόγος του παρατηρούμενου αριθμού των θανάτων, προς τον μεσό αριθμό αυτών που βρίσκονται σε κίνδυνο, άρα άμα συμβολίσουμε με τ_j το μήκος της $j-$οστής διαμέρισης και θεωρώντας ότι ο ρυθμός των θανάτων είναι σταθερός στην διαμέριση, έχουμε την εκτίμηση που ονομάζεται life-table εκτίμηση της συνάρτησης κινδύνου:

$$h^*(t) = \frac{d_j}{(n'_j - d_j/2)\tau_j}, \tag{3.15}$$

για $t \in [t'_k, t'_{k+1})$, $k = 1, 2, ..., m$.
Ένας άλλος τρόπος εκτίμησης είναι αν υποθέσουμε ότι η συνάρτηση κινδύνου είναι σταθερή μεταξύ των χρόνων που συμβαίνει το καταληκτικό γεγονός, ο κίνδυνος στην μονάδα του χρόνου μπορεί να βρεθεί διαρώντας τον αριθμό θανάτων προς των αριθμό των ατόμων σε ρίσκο. Έτσι έχουμε:

$$\hat{h}(t) = \frac{d_j}{n_j \tau_j}, \tag{3.16}$$

για $t \in [t(j), t(j+1))$ και $\tau_j = t(j+1) - t(j)$. Ο εκτιμητής αυτός λέγεται Kaplan-Meier εκτιμητής της συνάρτησης κινδύνου, αφού μέσω αυτού μπορούμε να φτάσουμε στην σχέση (3.11).

Η εκτίμηση της αθροιστικής συνάρτησης κινδύνου είναι πολύ σημαντική όταν θέλουμε να υποθέσουμε μοντέλα και κατανομές, παρακάτω δίνονται δύο τρόποι εκτίμησής της μέσω πρώτα της εκτίμησης Kaplan-Meier και ύστερα της εκτίμησης Nelson Aalen.

$$\hat{H}(t) = -\sum_{j=1}^{k} log\left(\frac{n_j - d_j}{n_j}\right) \tag{3.17}$$

για $t \in [t(k), t(k+1))$, $k = 1, 2, ..., r$.

$$H(t) = -\sum_{j=1}^{k} \frac{d_j}{n_j}, \tag{3.18}$$

για $k = 1, 2, ..., r$.

3.2.3 Εκτιμώντας τα ποσοστημόρια των χρόνων επιβίωσης

Αφού όπως αναφέραμε οι κατανομές των χρόνων επιβίωσης τείνουν να είναι δεξιά λοξές η διάμεσος φαίνεται να είναι ένα καλύτερο μέτρο για να περιγράψουμε τα δε-

δομένα μας σε σχέση με την μέση τιμή. Εκτιμώντας την συνάρτηση επιβίωσης είναι εύκολο να εκτιμήσουμε τη διάμεσο του χρόνου επιβίωσης. Είναι ο χρόνος όπου το μισό δείγμα των ατόμων εκτιμάται ότι θα έχει επιβιώσει και συμβολίζεται με $t(50)$. Βρίσκεται από την σχέση $S(t(50)) = 0.5$.

Επειδή η μη παραμετρική εκτίμηση της συνάρτησης επιβίωσης γίνεται βηματικά είναι σχεδόν αδύνατο να ισχύει η ισότητα και εκτιμάται από την σχέση:

$$\hat{t}(50) = min\{t(j)|\hat{S}(t(j)) < 0.5\} \tag{3.19}$$

όπου $t(j)$ ο $j-$οστος κατά σειρά χρόνος θανάτου, $j = 1, ..., r$.
Αντίστοιχα για τα άλλα ποσοστοιμόρια έχουμε $S(t(p)) = 1 - \frac{p}{100}$ και σαν εκτίμηση παίρνουμε τον μικρότερο εκτιμώμενο χρόνο για τον οποίο ισχύει $\hat{S}(t(p)) < 1 - \frac{p}{100}$.

3.3 Ημι-παραμετρικά μοντέλα με βάση την υπόθεση αναλογικότητας

Μοντελοποιώντας μια ανάλυση επιβίωσης μπορούμε παραδείγματος χάρην να ερευνήσουμε πως η επιβίωση μιας ομάδα από ασθενείς εξαρτάται από τις τιμές μιας ή και παραπάνω περιγραφικών μεταβλητών. Στην ανάλυση επιβίωσης το ενδιαφέρον μας επικεντρώνεται στο ρίσκο του κινδύνου (συνάρτηση κινδύνου) στο να πάθει το άτομο το καταληκτικό γεγονός κάθε χρονική στιγμή και έχουμε την απευθείας μοντελοποίηση της συνάρτησης κινδύνου, σε σύγκριση με της συνήθεις μεθόδους όπου μοντελοποιούμε την γραμμική σχέση κάποιων μεταβλητών. Παρόλα αυτά πολλές αρχές και διαδικασίες των γραμμικών μοντέλων ισχύουν και εδώ.

Υπάρχουν δύο κύριοι λόγοι για να μοντελοποιήσουμε τα δεδομένα από μια ανάλυση επιβίωσης, ο ένας είναι να δούμε ποιοι παράγοντες επηρεάζουν την επιβίωση ενός ατόμου μέσω των περιγραφικών μεταβλητών και ο άλλος να μπορούμε να εκτιμήσουμε την επιβίωση κάποιου ατόμου. Αυτό θα μας οδηγήσει να ενδιαφερόμαστε για ποσότητες όπως η διάμεσος που σαν παράδειγμα σε μια έρευνα για κάποια μορφή καρκίνου μας δηλώνει την εκτίμηση για το σε πόσους μήνες έχουν πεθάνει οι μισοί που έχουν αυτή τη μορφή καρκίνου.

Το μοντέλο που θα μιλήσουμε αρχικά είναι το μοντέλο αναλογικού κινδύνου (propotional hazards model), γνωστό και ως μοντέλο του Cox, το οποίο υποθέτει

την ύπαρξη αναλογικών κινδύνων, ενώ δεν υποθέτει κάποια κατανομή και για αυτό το λόγο λέγεται ημι-παραμετρικό. Για να περιγράψουμε τι σημαίνει η υπόθεση αναλογικού κινδύνου, σαν παράδειγμα ας υποθέσουμε μια μελέτη για κάποια πάθηση όπου δοκιμάζονται δύο είδη θεραπείας, η καινούργια (New) και η ήδη υπάρχουσα Standar, τότε με βάση την υπόθεση που αναφέραμε έχουμε για τις συναρτήσεις κινδύνων των ατόμων από την μια θεραπεία και από την άλλη, την σχέση

$$h_{New}(t) = \psi \cdot h_{Standar}(t), \tag{3.20}$$

όπου ψ είναι ένας σταθερός θετικός αριθμός και είναι ο λόγος των δύο συναρτήσεων κινδύνου. Ανάλογα του αν το ψ είναι μικρότερο ή μεγαλύτερο του 1 μπορούμε να βγάλουμε συμπέρασμα για το ποια θεραπεία είναι καλύτερη, αν π.χ. $\psi > 1$ τότε ο κίνδυνος θανάτου την χρονική στιγμή t είναι μεγαλύτερος για τα άτομα που βρίσκονται κάτω από την καινούρια θεραπεία, άρα η συνήθης θεραπεία είναι προτιμότερη. Αφού το ψ είναι θετικό μπορούμε να το γράψουμε στην μορφή $\beta = log\psi$. Ας θέσουμε τώρα την μεταβλητή X που παίρνει την τιμή 1 αν το άτομο βρίσκεται στην καινούρια θεραπεία, αλλιώς παίρνει την τιμή 0. Τότε η συνάρτηση επιβίωσης του i-οστού ατόμου θα έχει την μορφή

$$h_i(t) = e^{\beta x_i} h_0(t), \tag{3.21}$$

για $i = 1, ..., n$, όπου $h_0(t)$ η υπάρχουσα συνηθισμένη θεραπεία και n ο αριθμός των ασθενών στην έρευνα.

Γενικεύοντας το μοντέλο του Cox και θεωρώντας πως τα στοιχεία της έρευνας μας δεν αποτελούνται μόνο από χρόνους του καταληκτικού γεγονότος ή χρονους λογοκρισίας, αλλά και περιγραφικές μεταβλητές $x = (x_1, ..., x_p)'$, οι οποίες μπορεί να είναι συνεχείς, διακριτές και κατηγορικές (με δημιουργία ψευδομεταβλητών), έχουμε το γενικό αναλογικό μοντέλο:

$$h_i(t) = e^{\eta_i} h_0(t), \tag{3.22}$$

όπου

$$\eta_i = \beta_1 x_{1i} + ... + \beta_p x_{pi}. \tag{3.23}$$

Η $h_0(t)$ λέγεται baseline (αρχική ή μηδενική) συνάρτηση κινδύνου και είναι η συνάρτηση κινδύνου του ατόμου που έχει όλες τις τιμές των μεταβλητών 0, δηλαδή $x = (0, ..., 0)'$.

3.3.1 Εκτίμηση παραμέτρων

Η εκτίμηση των παραμέτρων $\beta_1, ..., \beta_p$ γίνεται με την χρήση μερικής πιθανοφάνειας (partial likelihood). Υποθέτουμε ότι μεταξύ δύο γεγονότων ο κίνδυνος είναι μηδενικός στο διάστημα αυτό κι έτσι αυτά τα διαστήματα δεν παρέχουν καμία πληροφορία. Οπότε θεωρούμε ότι το i-οστό άτομο με πίνακα μεταβλητών x_i μπορεί να έχει ένα γεγονός την χρονική στιγμή $t(j)$ δεδομένου ότι ένα άλλο άτομο εκείνη τη στιγμή παθαίνει το καταληκτικό γεγονός. Δηλαδή

P(το άτομο με μεταβλητές x_i να πεθάνει την στιγμή $t(j)$ δεδομένου ότι κάποιο άτομο πεθαίνει τη στιγμή $t(j)$)=P(το άτομο με μεταβλητές x_i να πεθάνει την στιγμή $t(j)$)/P(κάποιο άτομο πεθαίνει τη στιγμή $t(j)$) , αυτό μεταφράζεται στην σχέση:

$$\frac{h_i(t(j))}{\sum_{l \in R(t(j))} h_l(t(j))} = \frac{exp\{\beta' x_j\}}{\sum_{l \in R(t(j))} exp\{\beta' x_l\}}, \tag{3.24}$$

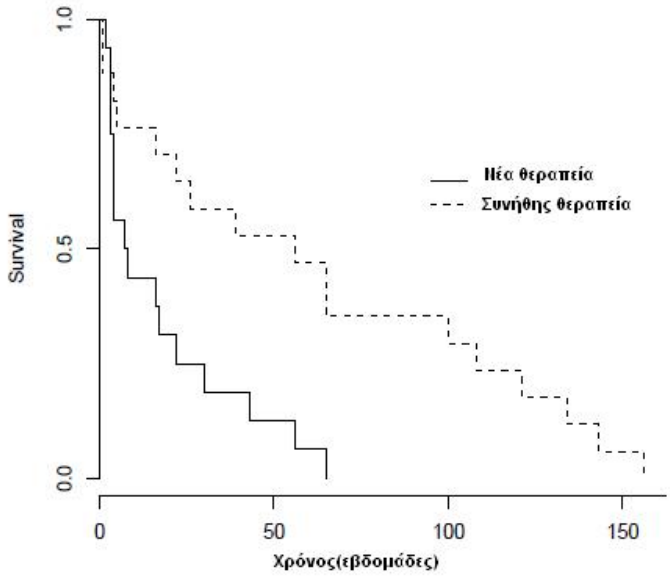
όπου $R(t)$ το σύνολο των ατόμων που βρίσκονται σε κίνδυνο και $\beta' = (\beta_1, ..., \beta_p)$. Μέσω αυτή της σχέσης μπορούμε να υπολογίσουμε την πιθανοφάνεια και να υπολογίσουμε τις εκτιμήσεις.

3.3.2 Αναλογικότητα (propotionality)

Η υπόθεση της αναλογικότητας δεν πρέπει να γίνεται χωρίς ενδείξεις ύπαρξης αυτής. Το πρώτο που έχουμε να κάνουμε όταν θέλουμε να υποθέσουμε αναλογικότητα είναι να κοιτάξουμε στα γραφήματα κατά πόσο για τις διαφορετικές τιμές μια μεταβλητής οι γραφικές παραστάσεις των συναρτήσεων επιβίωσης είναι ανάλογες μεταξύ τους μέσα στον χρόνο. Αν απ΄την άλλη τέμνονται ή είναι είναι απλά παράλληλες δεν μπορούμε να υποθέσουμε αναλογικότητα.

Σαν παράδειγμα ας θεωρήσουμε πως έχουμε τις δύο θεραπείες που μιλήσαμε προηγουμένως, ένας αρχικός τρόπος για να ελέγξουμε την αναλογικότητα είναι εκτιμώντας πχ με Kaplan-Meier την συνάρτηση επιβίωσης για τα άτομα από την καινούργια θεραπεία και με τον ίδιο τρόπο για τα άτομα από την συνήθης θεραπεία και σχεδιάζοντας

αυτές στους άξονες να δούμε αν υπάρχει τάση ή όχι για αναλογικότητα. Αυτό πρέπει να γίνει σε όλες τις μεταβλητές που θα συμπεριλάβουμε στην έρευνα. Σε περίπτωση μη κατηγορικών μεταβλητών μπορούμε να κατηγοριοποιήσουμε τις τιμές τους και να πράξουμε αναλόγως. Ένας ακόμα τρόπος ελέγχου γίνεται με την χρήση των καταλοίπων και χρήση τους σε σχετικά γραφημάτα, όπως υπάρχουν και τρόποι αντιμετώπισης άμα κάποια μεταβλητή δεν έχει αναλογικότητα, για τα οποία όμως δεν θα ασχοληθούμε στη συγκεκριμένη εργασία.



Σχήμα 3.2: Γράφημα συναρτήσεων επιβίωσης δύο θεραπειών, με ύπαρξη ένδειξης για αναλογικότητα

3.4 Παραμετρικά μοντέλα με βάση την υπόθεση αναλογικότητας

Όταν χρησιμοποιούμε το μοντέλο του Cox σε μια ανάλυση επιβίωσης, δεν χρειάζεται να υποθέσουμε κάποια κατανομή για τους χρόνους επιβίωσης. Σαν αποτέλεσμα η συνάρτηση κινδύνου δεν περιορίζεται σε μια συναρτησιακή δομή και το μοντέλο

γίνεται εύκολο στην χρήση του και για αυτό είναι και ευρέως διαδεδομένο. Από την άλλη άμα η υπόθεση μιας συγκεκριμένης κατανομής είναι συμβατή με τα δεδομένα, τα συμπεράσματα ενός τέτοιου μοντέλου θα είναι πιο ακριβή. Συγκεκριμένα οι εκτιμήσεις όπως για την διάμεσο και για την αναλογική σχέση των συναρτήσεων κινδύνου θα έχουν μικρότερα τυπικά σφάλματα. Αυτά τα μοντέλα στα οποία υποθέτουμε κάποια κατανομή, αλλά και την ύπαρξη αναλογικότητας, όπως την ορίσαμε προηγουμένως ονομάζονται παραμετρικά μοντέλα με την υπόθεση αναλογικότητας του κινδύνου (parametric propotional hazard models).

Όταν έχουμε επιλέξει την κατανομή που θεωρούμε κατάλληλη για το μοντέλο, οι αντίστοιχες εξισώσεις επιβίωσης και κινδύνου δίνονται από τις σχέσεις που είχαμε αναφέρει στην ενότητα 3.1.2 και ο γενικός τύπος για το μοντέλο είναι όπως και στο Cox

$$h_i(t) = e^{\eta_i} h_0(t), \tag{3.25}$$

όπου

$$\eta_i = \beta_1 x_{1i} + ... + \beta_p x_{pi}. \tag{3.26}$$

με σημαντική διαφορά όμως ότι εδώ η αρχική συνάρτηση κινδύνου baseline function $h_0(t)$ ακολουθεί κάποια συγκεκριμένη κατανομή. Η εκτίμηση των παραμέτρων γίνεται με εκτίμηση της μεγίστης πιθανοφάνειας. Παρακάτω αναφέρουμε τις βασικές κατανομές που χρησιμοποιούνται στα παραμετρικά μοντέλα με υπόθεση αναλογικότητας του κινδύνου.

3.4.1 Εκθετική κατανομή

Η πιο απλή μορφή που μπορεί να πάρει η συνάρτηση κινδύνου είναι αυτή που υποθέτουμε πως μένει σταθερή στο χρόνο. Ο κίνδυνος οπότε του να πεθάνει κάποιος είναι ίδιος ανεξάρτητα του πόσο χρόνος έχει περάσει. Δηλαδή:

$$h(t) = \lambda. \tag{3.27}$$

Το λ είναι μια θετική σταθερά η οποία εκτιμάται με βάση τα δεδομένα. Με βάση τις σχέσεις που περιγράψαμε στην ενότητα 3.1 έχουμε ότι :

$$S(t) = e^{-\lambda t}, \tag{3.28}$$

και έτσι

$$f(t) = \lambda e^{-\lambda t}, \tag{3.29}$$

δηλαδή η τυχαία μεταβλητή του χρόνου T ακολουθεί την εκθετική κατανομή με μέση τιμή λ^{-1}.
Για τα ποσοστημόρια έχουμε, ότι αν το p ποσοστημόριο του χρόνου εκφράζει την τιμή $t(p)$ τέτοια ώστε $S(t(p)) = 1 - \frac{p}{100}$ τότε:

$$t(p) = \frac{1}{\lambda} log\left(\frac{100}{100 - p}\right). \tag{3.30}$$

3.4.2 Weibull κατανομή

Στην πράξη η υπόθεση σταθερής συνάρτησης κινδύνου, δηλαδή εκθετικών κατανεμημένων χρόνων επιβίωσης είναι πολύ χαλαρή. Μια πιο γενική μορφή για την συνάρτηση κινδύνου είναι η

$$h(t) = \lambda \gamma t^{\gamma - 1}, \tag{3.31}$$

μια συνάρτηση που εξαρτάται από δύο θετικούς παραμέτρους.
Η συνάρτηση τότε είτε είναι φθίνουσα είτε είναι αύξουσα, ανάλογα με την τιμή που πέρνει το γ, ενώ στην ειδική περίπτωση που ισούται με την μονάδα οι χρόνοι έχουν την εκθετική κατανομή. Το γ ονομάζεται παράγοντας σχήματος, αφού το σχήμα της συνάρτησης καθορίζεται από αυτό, ενώ το λ παράγοντας κλίμακας.
Για την συνάρτηση επιβίωσης έχουμε:

$$S(t) = e^{-\lambda t^{\gamma}}, \tag{3.32}$$

και για την συνάρτηση κατανομής του χρόνου

$$f(t) = \lambda \gamma t^{\gamma - 1} e^{-\lambda t^{\gamma}}, \tag{3.33}$$

η οποία ονομάζεται κατανομή Weibull με παράμετρους λ, γ και συμβολίζεται με $W(\lambda, \gamma)$.
Η μέση τιμή της κατανομή δίνεται από τον τύπο

$$E(T) = \lambda^{-1/\gamma} \Gamma(\gamma^{-1} + 1), \tag{3.34}$$

με $\Gamma(y) = \int_0^\infty u^{y-1}e^{-u}du$.
Για τα ποσοστημόρια ισχύει:

$$t(p) = \left(\frac{1}{\lambda}log\frac{100}{100-p}\right)^{1/\gamma}. \tag{3.35}$$

3.4.3 Log-logistic κατανομή

Ένας περιορισμός της συνάρτησης κινδύνου με χρόνους επιβίωσης από Weibull κατανομή, είναι ότι λόγω της μονοτονίας της δεν μπορεί να περιγράψει περιπτώσεις όπου η συνάρτηση κινδύνου αλλάζει διεύθυνση. Σαν παράδειγμα σε κάποιες μορφές καρκίνου στην περίοδο μετά από κάποια εγχείρηση αφαίρεσης όγκων υπάρχει μεγάλος κινδύνος να πεθάνει κάποιος ασθενής, ενώ όσο περνάει ο καιρός μειώνεται σταδιακά. Μια κατανομή που μπορεί να περιγράψει τέτοιες περιπτώσεις είναι η log-logistic, για την οποία ισχύουν οι σχέσεις:

$$h(t) = \frac{\lambda\tau(\lambda t)^{\tau-1}}{1+(\lambda t)^\tau}, \tag{3.36}$$

$$S(t) = \frac{1}{1+(\lambda t)^\tau}, \tag{3.37}$$

όπου λ, τ είναι θετικοί αριθμοί. Για $\tau \leq 1$ ο κίνδυνος μειώνεται, ενώ αν $\tau > 1$ ο κίνδυνος είναι μονοκόρυφος και φθίνει μονότονα. Αντίστοιχα ισχύουν οι σχέσεις:

$$f(t) = \frac{\lambda\tau(t\lambda)^{\tau-1}}{(1+(\lambda t)^\tau)^2} \tag{3.38}$$

με

$$E(T) = \frac{\pi}{\lambda\tau sin(\pi\tau^{-1})}. \tag{3.39}$$

Για τα ποσοστημόρια έχουμε

$$t(p) = \frac{1}{\lambda}\left(\frac{p}{100-p}\right)^{1/\tau}. \tag{3.40}$$

3.4.4 log-Normal κατανομή

Γενικά όπως αναφέραμε ο χρόνος επιβίωσης δεν ακολουθεί την κανονική κατανομή. Υπάρχουν όμως μετασχηματισμοί που μπορούμε να φέρουμε τα δεδομένα σε μορφή ώστε να ακολουθούν κανονική κατανομή. Ένας συνηθισμένος τρόπος να γίνει αυτό είναι λογαριθμίζοντας τα και στην περίπτωση που όντως φαίνεται πως ακολουθούν κανονική κανονική μπορούμε να χρησιμοποιήσουμε την λογαριθμο-κανονική κατανομή (log-Normal) που είναι η κατανομής μιας τυχαίας μεταβλητής όπου ο λογάριθμός της ακολουθεί την κανονική κατανομή. Έχουμε για αυτήν :

$$f(t) = \frac{1}{\sigma t\sqrt{2\pi}} exp\{-\frac{(logt-\mu)^2}{2\sigma^2}\}. \tag{3.41}$$

Οι συναρτήσεις κινδύνου και επιβίωσης δεν υπολογίζονται σε κλειστή μορφή, η σχέση που ισχύει είναι η:

$$S(t) = 1 - \Phi\left(\frac{logt-\mu}{\sigma}\right), \tag{3.42}$$

όπου $\Phi(.)$ η συνάρτηση κατανομής της κανονικής $N(0,1)$.

3.5 Accelerated failure time models (AFT)

Αν και τα propotional μοντέλα έχουν ευρεία χρήση στην ανάλυση επιβίωσης, είναι σχετικά λίγες οι κατανομές που μπορούμε να χρησιμοποιήσουμε για αυτού του είδους τα μοντέλα και συνήθως αναφέρονται σε μονότονες συναρτήσεις. Ένα είδος μοντελοποίησης που περιέχει μια μεγαλύτερη γκάμα από κατανομές είναι τα μοντέλα επιταχυνόμενου χρόνου αποτυχίας (accelerated failure time models). Ειδικά σε περιπτώσεις όπου η υπόθεση αναλογικότητας δεν φαίνεται σωστή αυτά τα μοντέλα είναι πολύ χρήσιμα.

3.5.1 Δομή μοντέλου

Τα AFT μοντέλα είναι μια γενική μορφή μοντέλου για την ανάλυση επιβίωσης στα οποία οι περιγραφικές μεταβλητές που συλλέγονται από τα άτομα υποτίθεται ότι επιδρούν πολλαπλασιαστικά στον χρόνο και έτσι επηρεάζουν το μέγεθος που το άτομο κινείται στον άξονα του χρόνου. Δηλαδή μπορούν να ερμηνευτούν σαν την ταχύτητα

της δράσης μιας ασθένειας, πράγμα το οποίο έχει και φυσική-λογική έννοια.

Για να δούμε την δομή του μοντέλου ας αναφέρουμε πάλι το παράδειγμα με τα δύο είδη θεραπείας την καινούργια New και την ήδη υπάρχουσα Standar. Κάτω από τα AFT μοντέλα ο χρόνος επιβίωσης ενός ατόμου στην νέα θεραπεία είναι πολλαπλάσιος του χρόνου επιβίωσης ενός ατόμου στην υπάρχουσα θεραπεία. Έτσι το αποτέλεσμα της καινούργιας θεραπείας θα είναι είτε να *αυξήθει* είτε να *μειώθει* στο πέρασμα του χρόνου. Κάτω από αυτήν την υπόθεση η πιθανότητα του ατόμου κάτω από την νέα θεραπεία να ζει μετά το χρόνο t είναι η πιθανότητα το άτομο στην υπάρχουσα θεραπεία να επιβιώσει μετά από χρόνο t/ϕ, όπου ϕ είναι άγνωστη θετική σταθερά. Δηλαδή ορίζοντας τις συναρτήσεις επιβίωσης για την νέα θεραπεία έχουμε την σχέση:

$$S_{New} = S_{Standar}(t/\phi), \tag{3.43}$$

για όλες τις τιμές του t. Μια ερμηνεία του μοντέλου αυτού είναι ότι η διάρκεια ζωής ενός ατόμου που ακολουθεί την νέα θεραπεία είναι ίση ϕ-φορές την διάρκεια ζωής που θα είχε αν είχε ακολουθήσει την συνηθισμένη θεραπεία. Έτσι το ϕ αντικατροπτίζει το μέγεθος της νέας θεραπείας στον αρχικό *baseline* χρόνο. Αν το καταληκτικό γεγονός είναι ο θάνατος κάποιου ασθενή, τότε όταν $\phi < 1$, αυτό σημαίνει μια επιτάχυνση του χρόνου προς τον θάνατο για τα άτομα της νέας θεραπείας σε σχέση με αυτά από την συνηθισμένη, η οποία την συγκεκριμένη περίπτωση θα θεωρηθεί καλύτερη. Στην περίπτωση που το καταληκτικό γεγονός είναι η αποθεραπεία του ασθενή τότε προφανώς η νέα θεραπεία είναι καλύτερη (για $\phi < 1$). Για αυτό και το ϕ^{-1} ονομάζεται παράγοντας επιτάχυνσης (acceleration factor).

Ο παράγοντας επιτάχυνσης μπορεί να εννοηθεί και όταν αναφερόμαστε για την διάμεσο στις δύο θεραπείες, ας τις πούμε $t_N(50), t_S(50)$. Για αυτές τις τιμές ισχύει $S_N(t_N(50)) = S_S(t_S(50)) = \frac{1}{2}$, και έτσι έχουμε ότι $t_N(50) = \phi \cdot t_S(50)$. Δηλαδή κάτω από τα AFT μοντέλα έχουμε ότι η διάμεσος της επιβίωσης ενός ασθενή στη νέα θεραπεία είναι ϕ-φορές την συνήθη. Η συγκεκριμένη ερμηνεία είναι πολύ χρήσιμη, αφού γίνεται κατανοητή και με την φυσική έννοια.
Με βάση τις μαθηματικές σχέσεις που έχουμε αναφέρει έχουμε:

$$f_N(t) = \phi^{-1} f_s(t/\phi), \tag{3.44}$$

$$h_N(t) = \phi^{-1} h_s(t/\phi). \tag{3.45}$$

Θέτοντας όπως κάναμε και στο μοντέλο του Cox στην αντίστοιχη περίπτωση, την μεταβλητή X που παίρνει την τιμή 1 αν το άτομο βρίσκεται στην καινούρια θεραπεία, αλλιώς την τιμή 0, η συνάρτηση επιβίωσης του i-ατόμου έχει την μορφή

$$h_i(t) = \phi^{-x_i} h_0(t/\phi^{x_i}), \tag{3.46}$$

και εδώ η $h_0(t)$ ονομάζεται αρχική συνάρτηση (baseline) και είναι η συνάρτηση κινδύνου των ατόμων από την συνήθη θεραπεία.
Όπως αναφέραμε το ϕ είναι θετικός αριθμός άρα μπορεί να γραφτεί $\phi = e^a$ κι έτσι έχουμε τον τύπο:

$$h_i(t) = e^{-ax_i} h_0(t/e^{ax_i}). \tag{3.47}$$

Γενικεύοντας τώρα όπως κάναμε αντίστοιχα στο μοντέλο του Cox και θεωρώντας τις περιγραφικές μεταβλητές $x = (x_1, ..., x_p)'$ που συνοδεύουν κάθε άτομο της έρευνας, έχουμε την συνάρτηση κινδύνου για το i-οστο άτομο:

$$h_i(t) = e^{-\eta_i} h_0(t/e^{\eta_i}), \tag{3.48}$$

όπου

$$\eta_i(t) = a_1 x_{1i} + ... + a_p x_{pi}, \tag{3.49}$$

για $i = 1, ..., n$. Και εδώ θεωρούμε την $h_0(t)$ baseline (αρχική ή μηδενική) συνάρτηση κινδύνου και είναι η συνάρτηση κινδύνου του ατόμου που έχει όλες τις τιμές των μεταβλητών ίσες με μηδέν.
Για την αντίστοιχη συνάρτηση επιβίωσης έχουμε

$$S_i(t) = S_0(t/e^{\eta_i}). \tag{3.50}$$

3.5.2 Log-linear (λογαριθμο-γραμμική) μορφή του μοντέλου

Ας υποθέσουμε ότι η τυχαία μεταβλητή που περιγράφει τον χρόνο ζωής του i-οστου ατόμου T_i έχει την μορφή

$$logT_i = \mu + a_1 x_{1i} + ... + a_p x_{pi} + \sigma\epsilon_i. \tag{3.51}$$

Σε αυτήν την μοντελοποίηση οι $a_1, ..., a_p$ είναι οι άγνωστοι συντελεστές των τιμών των περιγραφικών μεταβλητών $X_1, ...X_p$, ενώ τα μ, σ είναι δύο επιπλέον παράμετροι, γνωστοί και σαν intercept και scale αντίστοιχα. Το ϵ_i είναι μια τυχαία μεταβλητή και χρησιμοποιείται για να περιγραφτεί η απόκλιση των τιμών του λογαριθμικού χρόνου από το γραμμικό κομμάτι του μοντέλου, για το οποίο θεωρούμε πως ακολουθεί κάποια κατανομή. Σε αυτήν τη μορφή του μοντέλου οι τιμές των $a_1, ..., a_p$ μας δείχνουν την επίδραση που έχουν οι περιγραφικές μεταβλητές στον χρόνο επιβίωσης.

Ας δούμε τώρα τη σύνδεση αυτής της μορφής των AFT μοντέλων με αυτήν που περιγράψαμε αρχικά. Για την παραπάνω σχέση του $logT_i$ και θέτοντας

$$a'x_i = a_1 x_{1i} + ... + a_p x_{pi}$$

έχουμε:

$$S_i(t) = P(T_i \geq t) = P(exp\{\mu + a'x_i + \sigma\epsilon_i\} \geq t) = P(exp\{\mu + \sigma\epsilon_i\} \geq t/exp\{a'x_i\}),$$

οπότε για την αρχική συνάρτηση επιβίωσης έχουμε:

$$S_0(t) = P(exp\{\mu + \sigma\epsilon_i\} \geq t)$$

Άρα

$$S_i(t) = S_0(t/exp(a'x_i)) \tag{3.52}$$

ο οποίος είναι ο τύπος (3.50) της γενικής μοφης του μοντέλου για $\eta_i = a'x_i$.
Για την συνάρτηση κινδύνου έχουμε ότι

$$h_i(t) = e^{-a'x_i} h_0(t/e^{a'x_i}) \tag{3.53}$$

δηλαδή η σχέση (3.48) για $\eta_i = a'x_i$.

Η λογαριθμο-γραμμική μορφή του μοντέλου μπορεί να μας δώσει και μια γενική μορφή για την συνάρτηση επιβίωσης του i-οστου ατόμου

$$S_i(t) = P(\epsilon_i \geq \frac{logt - \mu - a'x_i}{\sigma}) = S_{\epsilon_i}(\frac{logt - \mu - a'x_i}{\sigma}). \tag{3.54}$$

Από αυτή τη σχέση οδηγούμαστε στα παρακάτω:

$$t_i(p) = exp\{\sigma\epsilon_i(p) + \mu + a'x_i\}t_0(p), \tag{3.55}$$

$$H_i(t) = H_{\epsilon_i}(\frac{logt - \mu - a'x_i}{\sigma}), \tag{3.56}$$

$$h_i(t) = \frac{1}{\sigma t}h_{\epsilon_i}(\frac{logt - \mu - a'x_i}{\sigma}). \tag{3.57}$$

Για τις πιο συνηθισμένες κατανομές που ακολουθεί ο χρόνος επιβίωσης έχουμε τον παρακάτω πίνακα:

T_i	$S_{\epsilon_i}(\epsilon)$	$h_{\epsilon_i}(\epsilon)$	$\epsilon_i(p)$
$Expontential$	$exp\{-e^{\epsilon}\}$	e^{ϵ}	$log(-log(100-p)/100))$
$Weibull$	$exp\{-e^{\epsilon}\}$	e^{ϵ}	$log(-log(100-p)/100)$
$Log-logistic$	$(1+e^{\epsilon})^{-1}$	$(1+e^{-\epsilon})^{-1}$	$log(p/(100-p))$
$Lognormal$	$1-\Phi(\epsilon)$	$\frac{exp\{-\epsilon^2/2\}}{(1-\Phi(\epsilon))\sqrt{2\pi}}$	$\Phi^{-1}(p/100)$

Η εκτίμηση των παραμέτρων γίνεται με την μέθοδο μέγιστης πιθανοφάνειας.

3.6 Παράδειγμα Ανάλυσης Επιβίωσης

Για να γίνουν πιο κατανοητές οι διαδικασίες για τις οποίες μιλήσαμε στα προηγούμενα μέρη του κεφαλαίου αυτού, θα δώσουμε ένα παράδειγμα μια έρευνας με πραγματικά δεδομένα. Μας δίνεται λοιπόν ένα τμήμα μελέτης που έλαβε χώρα στις ΗΠΑ από το Ακτινο-ογκολογικό Ινστιτούτο, με την πλήρη μελέτη να περιέχει ασθενείς με πλακώδες καρκίνωμα (*squamous carcinoma*) σε 15 σημεία στο στόμα και στον λαιμό, από 16 νοσοκομεία [13]. Τα δεδομένα που θα δουλέψουμε αναφέρονται σε μόνο 3

σημεία, από τα 6 μεγαλύτερα νοσοκομεία. Οι ασθενείς που συμμετείχαν στην μελέτη έλαβαν τυχαία μια από τις 2 θεραπείες, η πρώτη: μόνο ακτινοθεραπεία και η δεύτερη: ακτινοθεραπεία και φάρμακα χημειοθεραπείας.
Οι μεταβλητές οι οποίες θα χρησιμοποιήσουμε είναι οι εξής:

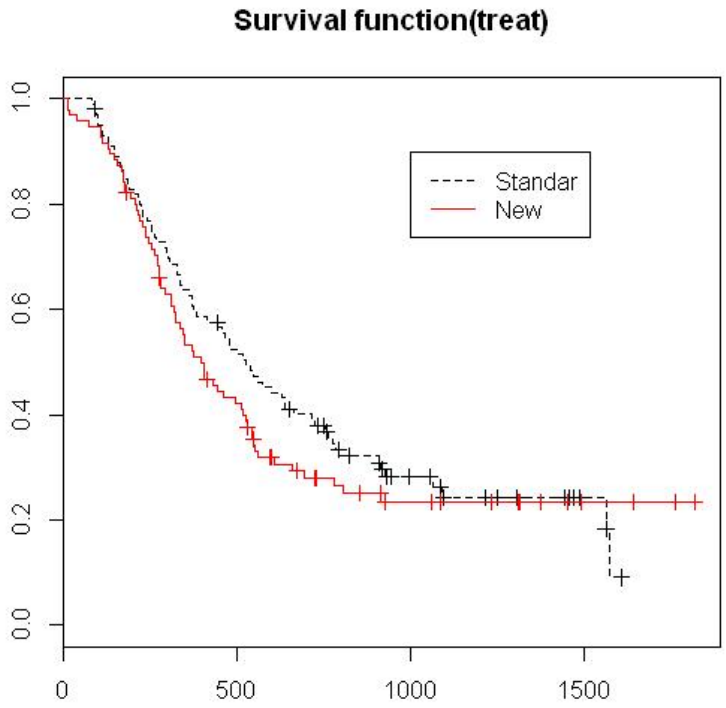
1. Case: Ο κωδικός του κάθε ασθενή (μέγεθος δείγματος 195)
2. Sex: Το φύλο του ασθενή : Άντρας=1, Γυναίκα=2 (76%, 24%)
3. Grade : Το μέγεθος που ο όγκος έχει καταλάβει τα κύτταρα:
   πολύ=1, μεσαία=2, λίγο=3 (25%, 56%, 18%, NA=1%)
4. age: Η ηλικία του ασθενή (20-90, η πλειοψηφία ανήκει στην κατηγορία 40-60)
5. cond: Η κατάσταση του ασθενή : καμία ανικανότητα=1, χρειάζεται να προσέχει=2, χρειάζεται βοήθεια από τρίτο=3 (74%, 22%, 3%, NA=1%)
6. site: Σημείο που υπάρχει ο όγκος: faucial arch=1, tonsillar fossa=2, pharyngeal tongue=4 (κοντινά ποσοστά εμφάνισης).
7. Tstage: Μέγεθος διόγκωσης του όγκου :2εκ. ή μικρότερος=1, 2-4εκ. με μια μικρή διείσδυση σε βάθος=2, 4εκ και μεγαλύτερος=3, εξαιρετικά επιθετικός όγκος=4 (5%, 13%, 48%, 34%)
8. Status: λογοκρισία=0, θάνατος=1 (27% *censored*)
9. time: Χρόνος γεγονότος, είτε ότι ο ασθενής πέθανε, είτε ότι ο ασθενής έφυγε από την έρευνα (11-1823)

Το πρώτο πράγμα που έχουμε να κάνουμε, όπως και σε μια συνηθισμένη στατιστική έρευνα είναι να βγάλουμε κάποια αρχικά συμπεράσματα μέσω των γραφικών παραστάσεων. Στην ανάλυση επιβίωσης και ειδικότερα όταν θέλουμε να δουλέψουμε με την υπόθεση ύπαρξης αναλογικότητας, το πρώτο που ελέγχουμε είναι η υπόνοια αναλογικότητας, αυτό γίνεται με βάση των γραφήματων της συνάρτησης επιβίωσης εκτιμώντας την με Kaplan-Meier για κάθε μεταβλητή ξεχωριστά, απ΄τα οποία μπορούμε επίσης να πάρουμε μια πρώτη άποψη για το κατά πόσο οι διάφορες τιμές μιας μεταβλητής επηρεάζουν τον χρόνο επιβίωσης. Χρησιμοποιώντας τις αντίστοιχες συναρτήσεις παραθέτουμε ενδεικτικά τα γραφήματα για την θεραπεία και για την κατάσταση του ασθενή (Σχήμα 3.4, 3.5), ενώ τα λοιπά υπάρχουν στο παράρτημα (Σχήμα A.5, A.6).

	case	sex	treat	grade	age	cond	site	tstage	status	time
1	1	2	1	1	51	1	2	3	1	631
2	2	1	2	1	65	1	4	2	1	270
3	3	1	1	2	64	2	1	3	1	327
4	4	1	1	1	73	1	1	4	1	243
5	5	1	2	2	64	1	1	4	1	916
6	6	1	2	1	61	1	2	3	0	1823
7	7	1	1	2	65	1	2	4	1	637
8	8	1	2	3	84	1	4	1	1	235
9	9	1	1	2	54	2	1	3	1	255
10	10	1	1	2	72	2	4	2	1	184
11	11	1	1	2	42	1	4	2	1	1064
12	12	1	1	2	61	1	1	4	1	414
13	13	1	2	1	71	1	2	3	1	216
14	14	1	2	2	83	3	4	3	1	324
15	15	1	1	3	43	1	2	4	1	480
16	16	1	2	2	52	1	4	4	1	245

Σχήμα 3.3: Ενδεικτικά δίνονται οι 16 πρώτες παρατηρήσεις

Από το γράφημα της συνάρτησης επιβίωσης ως προς το είδος θεραπείας παρατηρούμε πως γενικά οι δύο θεραπείες είναι σχετικά αναλογικές και πως δεν διαφέρουν σημαντικά, το οποίο ενδέχεται να σημαίνει ότι θα τις βρούμε στατιστικά μη σημαντικές στα μοντέλα μας. Μάλιστα παρατηρούμε ότι η συνήθης θεραπεία είναι γενικά καλύτερη από τη δοκιμάστική. Η τομή στις δύο ΄γραμμές΄ των συναρτήσεων επιβίωσης ενδέχεται να οφείλεται σε μεμονομένες ακραίες τιμές.
Όμοια συμπεράσματα παρατηρούμε και από το αντίστοιχο γράφημα για το φύλο, με τις γυναίκες να έχουν γενικά *καλύτερη* επιβίωση, αν και φαίνεται πως δεν διαφέρει σημαντικά.
Στο γράφημα του μέγεθους του όγκου παρατηρούμε πως γενικά οι ασθενείς που ο όγκος δεν έχει επεκταθεί πολύ στα κύτταρα έχουν καλύτερη επιβίωση σε σχέση με τις δύο άλλες κατηγορίες, που κινούνται σε παρόμοιες τιμές. Γενικά υπάρχει μια τάση για αναλογικότητα στα διαφορετικά επίπεδα της μεταβλητής και ενδέχεται λόγω της διαφοράς (γραφικά) που υπάρχει μεταξύ των ατόμων που το μέγεθος του όγκου είναι μικρό, σε σχέση με τις άλλες δύο περιπτώσεις να την βρούμε στατιστικά σημαντική κατά την μοντελοποίηση.
Στο γράφημα της κατάστασης του ασθενή, παρατηρούμε σημαντική διαφορά μεταξύ των ατόμων που δεν έχουν καμία ανικανότητα σε σχέση με τα άλλα,παρατηρούμε επίσης ότι τα άτομα που χρειάζονται βοήθεια από τρίτο έχουν πολύ μικρή *γραμμή*



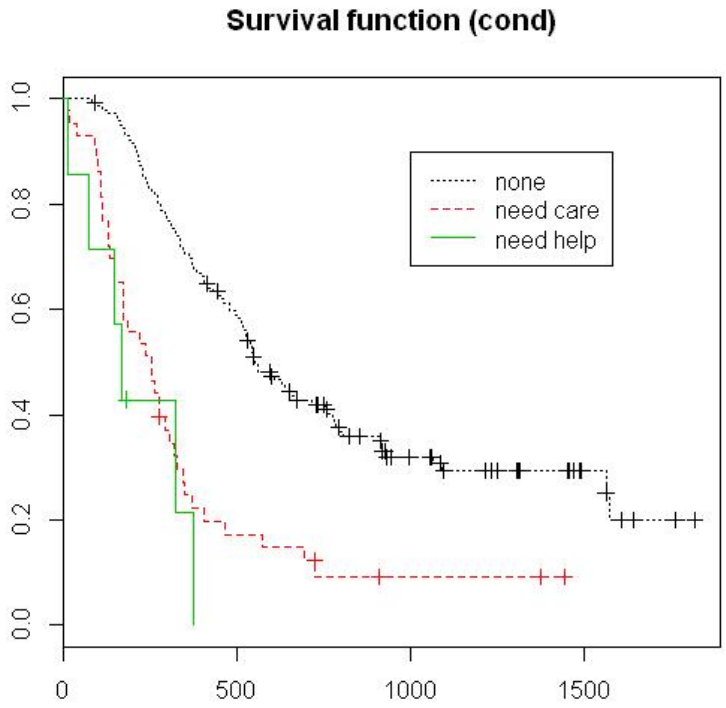
Σχήμα 3.4: Συναρτήσεις επιβίωσης με βάση την θεραπεία

επιβίωσης αν και αυτό μπορεί να οφείλεται στο μικρό δείγμα αυτών των ατόμων. Και εδώ παρατηρούμε μια αναλογικότητα στις τιμές της μεταβλητής.

Στο γράφημα που αναφέρεται στο σημείο που βρίσκεται ο όγκος, παρατηρούμε ότι οι τιμές για την κάθε τιμή της μεταβλητής γενικά δεν διατηρούν κάποια αναλογικότητα, τέμνονται σε πολλά σημεία και γενικά είναι κοντά η μια στην άλλη, αυτό μπορεί να σημαίνει ότι θα είναι μη στατιστικά σημαντικές.

Για το γράφημα του μεγέθους διόγκωσης του όγκου, αρχικά παρατηρούμε ότι η συνάρτηση επιβίωσης της περίπτωσης ΄2εκ. ή μικρότερος΄ η οποία χαλάει την αναλογικότητα ενδέχεται να επηρεάζεται από το γεγονός πως ασθενείς με αυτή την ιδιότητα είναι μόνο το 5% του δείγματος και έτσι παίρνοντας αυτό κατά νου παρατηρούμε μια αναλογικότητα όπως και διαφορά στην επιβίωση στα διάφορα επίπεδα, το οποίο μπορεί να σημαίνει πως στο μοντέλο μας η μεταβλητή θα είναι στατιστικά σημαντική.

Τέλος για την μεταβλητή της ηλικίας κατηγοριοποιήσαμε σε τιμές: μικρότεροι του



Σχήμα 3.5: Συναρτήσεις επιβίωσης με βάση την κατάσταση του ασθενή

40=1,(40-60)=2,(60-80)=3, μεγαλύτεροι του 80=4, γενικά υπάρχει μια αναλογικότητα, στις δύο τιμές 2,3 που κατέχουν και το μεγάλο μέρος του δείγματος παρατηρούμε ότι το 3(60-80) έχει καλύτερη επιβίωση πολύ οριακά. Τείνει να είναι στατιστικά ασήμαντη (γραφικά) και ενδέχεται να μην είναι λόγω κυρίως των τιμών για τα 1,4 που όμως δεν έχουν ικανοποιητικό δείγμα (η 1 έχει την χειρότερη επιβίωση).

Θα μοντελοποιήσουμε τα δεδομένα με χρήση του αναλογικού μοντέλου Cox. Αρχικά θα ελέγξουμε κατά πόσο οι διάφορες μεταβλητές είναι από μόνες τους στατιστικά σημαντικές με βάση αυτό το μοντέλο. Όσες δεν είναι θα της απορρίψουμε από το μοντέλο μας, έτσι έχουμε τον παρακάτω πίνακα:

Μεταβλητή	Στατ.Σημ.	p value
treat	ΟΧΙ	0,336
sex	ΟΧΙ	0,4008
grade	ΝΑΙ	0.06309
cond	ΝΑΙ	≈ 0
site	ΟΧΙ	0,5189
tstage	ΝΑΙ	0,01527
age	ΟΧΙ	0,5018

Το πιο σημαντικό που παρατηρούμε είναι ότι η θεραπεία είναι στατιστικά ασήμαντη, πράγμα για το οποίο είχαμε ήδη ενδείξεις και από το γράφημα, επίσης από την εκτίμηση του συντελεστή της ($e^{\beta} = 1.1761$) βλέπουμε ότι η συνάρτηση κινδύνου δηλαδή η πιθανότητα να πεθάνει κάποιος την χρονική στιγμή t, είναι μεγαλύτερη στη δοκιμαστική θεραπεία, αν και το γεγονός ότι είναι στατιστικά ασήμαντη θα μας κάνει να υποθέσουμε πως οι δύο θεραπείες δεν διαφέρουν ουσιαστικά. Και έτσι με βάση την φύση των δύο θεραπειών φαίνεται ότι η χημειοθεραπεία είναι περιττή (και ενδέχεται κιόλας να κάνει και μικρότερη την επιβίωση.
Το σημείο που υπάρχει ο όγκος δεν φαίνεται να επηρεάζει την συνάρτηση κινδύνου. Όμοια και το φύλο και η ηλικία του ασθενούς. Αξίζει να σημειωθεί πως τα παραπάνω αποτελέσματα τα είχαμε υποπτευθεί ήδη από την περιγραφική στατιστική.

Συνεχίζοντας την διαδικασία επιλογής μοντέλου έχουμε μείνει με τις μεταβλητές grade,cond,tstage τις οποίες θα τις τοποθετήσουμε στο ίδιο μοντέλο όπου θα χρησιμοποιήσουμε την διαδικασία της αντίστροφης απόρριψης (backward elimination), όπου απορρίπτουμε κάθε φορά την πιο στατιστικά ασήμαντη μεταβλητή (αν υπάρχει) και επαναλαμβάνουμε μέχρι να φτάσουμε σε ένα μοντέλο που όλες οι μεταβλητές θα είναι στατιστικά σημαντικές. Με βάση αυτή την διαδικασία καταλήγουμε στο μοντέλο με περιγραφικές μεταβλητές αυτές της κατάστασης τους ασθενή και του μεγέθους διόγκωσης του όγκου.

Σαν τελικό βήμα, δοκιμάζουμε μια -μια τις μεταβλητές που απορρίψαμε στο πρώτο βήμα, μήπως είναι τώρα στατιστικά σημαντικές στο μοντέλο που καταλήξαμε μετά τη διαδικασία αντίστροφης απόρριψης. Ύστερα και από αυτόν τον έλεγχο, παραμένουμε

στο μοντέλο με τις δύο περιγραφικές μεταβλητές που είπαμε προηγουμένως, για το οποίο έχουμε τις εκτιμήσεις των συντελεστών:

Μεταβλητή	συντελεστής	εκθετικός συντ.	p value
cond=2	1.0532	2.8668	≈ 0
cond=3	1.8636	6.4468	
tstage=2	-0.4041	0.6675	0.025
tstage=3	-0.3471	0.7067	
tstage=4	0.1867	1.2053	

Δηλαδή σύμφωνα με την σχέση (3.22) καταλήξαμε στο μοντέλο:

$$h_i(t) = e^{\{1.0532 \cdot cond2_i + 1.8636 \cdot cond32_i - 0.4041 \cdot tstage2_i - 0.3471 \cdot tstage3_i + 0.1867 \cdot tstage4_i\}} \cdot h_0(t)$$

Έτσι έχουμε τα εξής συμπεράσματα για τις μεταβλητές που επηρεάζουν τον χρόνο επιβίωσης του ασθενή. Για την μεταβλητή της κατάστασης του ασθενή παρατηρούμε ότι η πιθανότητα να πεθάνει κάποιος σε κάποιο χρόνο t είναι μεγαλύτερη στα άτομα που χρειάζονται βοήθεια από τρίτο 6.4 φορές σε σχέση με την baseline, μετά 2.8 φορές (σε σχέση με την baseline) τα άτομα που πρέπει να προσέχουν και προφανώς μικρότερο κίνδυνο έχουν τα άτομα που δεν έχουν καμία ανικανότητα.
Για την μεταβλητή του μέγεθος διόγκωσης του όγκου, έχουμε ότι τα άτομα με εξαιρετικά επιθετικό όγκο η πιθανότητα σε σχέση με την baseline αυξάνεται κατά περίπου 20% (1.20) ενώ στα άλλα δύο: 4εκ και μεγαλύτερος, 2-4εκ. με μια μικρή διείσδυση σε βάθος μειώνεται κατά περίπου 30%. Τα κύρια στοιχεία που θα κρατήσουμε είναι ότι ο επιθετικός όγκος αυξάνει κατά πολύ τον κίνδυνο σε σχέση με τα άλλα δύο, 4εκ και μεγαλύτερος, 2-4εκ. με μια μικρή διείσδυση, τα οποία δίνουν σχεδόν ίδιο κίνδυνο, ενώ το ότι στην δεύτερη θέση βρίσκεται το 2εκ. ή μικρότερος ναι μεν θα το λάβουμε υπ΄όψη αλλά λόγω της μικρής εμφάνισης του στο δείγμα (5%) και λόγω της μεγάλης διασποράς που αυξάνει συγχρόνως και τα διαστήματα εμπιστοσύνης θα κρατήσουμε μια μεγάλη επιφύλαξη κατά πόσο δεν επηρεάζεται από τα προαναφερθέντα.

Εδώ θα αναφέρουμε ότι υπάρχουν πολλές άλλες διαδικασίες οι οποίες μπορούν να

ληφθούν υπ΄όψην σε μια ανάλυση επιβίωσης, ενδεικτικά αναφέρουμε τα απαραμετρικά κριτήρια που χρησιμοποιούνται για μια αρχική υπόνοια στατιστική σημαντικότητας, τα martingale κατάλοιπα για εύρεση τιμών που προεξέχουν (outliers) και τους διάφορους ελέγχους καλής προσαρμογής μέσω κυρίως των καταλοίπων (residuals), οι οποίες δεν θα αναφερθούν στην παρούσα εργασία αφού σκοπός είναι μια γενική ιδέα μια ανάλυσης επιβίωσης και όχι η εμβάνθυση σε αυτή. Για τον ίδιο λόγο δεν θα παραθέσουμε και την παραμετρική ανάλυση του παραδείγματος.

Κεφάλαιο 4

Μέθοδοι μετα-ανάλυσης στη Βιοστατιστική με χρήση του λόγου των ποσοστημορίων

Στο κεφάλαιο αυτό θα περιγράψουμε διάφορες μεθόδους όπου η μετα-ανάλυση, μπορεί να χρησιμοποιηθεί στην παραγωγή γενικών συμπερασμάτων από κάποιο σύνολο μελετών τα οποία ανήκουν στον τομέα της Βιοστατιστικής, όπως νέα θεραπεία εναντίον παλιάς θεραπείας, καινούριο φάρμακο εναντίον προϋπάρχοντος φαρμάκου και άλλες παρόμοιες στατιστικές έρευνες στις οποίες η ανάλυση επιβίωσης είναι μια από τις πιο βασικές στατιστικές μεθόδους. Το μέγεθος της επίδρασης στο οποίο θα επικεντρώσουμε το ενδιαφέρον μας θα είναι ο λόγος των ποσοστημορίων. Πριν όμως μιλήσουμε για τον λόγο των ποσοστημορίων στην μετα-ανάλυση ας αναφέρουμε το πιο διαδεδομένο μέγεθος επίδρασης που συναντάται στη βιβλιογραφία.

4.1 Λόγος του Κινδύνου (Hazard Ratio)

Το πιο συνηθισμένο και προτεινόμενο στατιστικό μέτρο ελέγχου σε μελέτες που αφορούν δυο διαφορετικές ομάδες είναι ο λόγος του κινδύνου (Hazard Ratio - HR), δηλαδή αν θεωρήσουμε ότι η μελέτη μας αναφέρεται στο παράδειγμα που μιλήσαμε και στο προηγούμενο κεφάλαιο της καινούριας θεραπείας New, εναντίον της συνηθισμένης Standar, το HR δηλώνει το κατά πόσο περισσότερο ή λίγοτερο κινδυνεύουν

τα άτομα της μιας θεραπείας έναντι της άλλης κάποια χρονική στιγμή να τους συμβεί το καταληκτικό γεγονός. Είναι κατάλληλο για τέτοιου είδους μελέτες αφού μπορεί να υπολογιστεί από δεδομένα μέχρι το γεγονός (time to event data), στα οποία υπάρχει λογοκρισία και περιγράφει πρακτικά το μέγεθος της διαφοράς μεταξύ των δυο Kaplan Meier καμπυλών των δυο θεραπειών. Δηλαδή για κάποια χρονική στιγμή t έχουμε

$$HR(t) = \frac{h_{New}(t)}{h_{Standar}(t)}. \tag{4.1}$$

Ανάλογα με την τιμή που μας δίνει το HR, βγάζουμε και τα σχετικά συμπεράσματά, με τιμές μικρότερες του 1 να σημαίνουν ότι ο κίνδυνος μικραίνει με την καινούρια θεραπεία, ενώ για τιμές μεγαλύτερες του 1 ο κίνδυνος μεγαλώνει, ενώ για τιμές ίσες με την μονάδα (ή πολύ κοντά) παρατηρούμε πως οι δυο θεραπείες έχουν την ίδια επίδραση τουλάχιστον στην συγκεκριμένη χρονική περίοδο. Προφανώς, εκτός από του αν είναι χειρότερη η μια θεραπεία από την αλλή μας παρέχεται και η πληροφοριά του κατά πόσο μεγαλύτερος ή μικρότερος σε σχέση με την άλλη θεραπεία είναι ο κίνδυνος. Η σχέση (4.1) μεταφράζεται στη μετα-ανάλυση, ανάλογα με το τι μας παρέχεται από κάθε διαθέσιμη μελέτη και πως έχει μοντελοποιηθεί αυτή.

Μη παραμετρική εκτίμηση του HR

Η εκτίμηση των Cox-Mantel για το HR μας δίνεται από την σχέση

$$HR = \frac{O_N/E_N}{O_S/E_S}, \tag{4.2}$$

όπου O_i είναι ο αριθμός των θανάτων που συνέβησαν στην $i-$οστή ομάδα και E_i είναι ο εκτιμώμενος αριθμός των θανάτων της $i-$οστής ομάδας, ο οποίος δίνεται από τον τύπο

$$E_i = \frac{n_i \cdot d}{N}, \tag{4.3}$$

όπου n_i είναι το σύνολο των ατόμων της $i-$οστής ομάδας που βρίσκονται σε κίνδυνο, d το σύνολο των ατόμων που πάθανε το καταληκτικό γεγονός και N ο συνολικός αριθμός των ατόμων σε κίνδυνο και από τις δύο ομάδες μαζί. Εδώ να παρατηρήσουμε ότι γενικά το HR το μετατρέπουμε στην λογαριθμική κλίμακα και δουλεύουμε με αυτήν, καθότι προσεγγίζεται από την κανονική κατανομή και στην συνέχεια το

επαναφέρουμε στο αρχικό μέτρο για να βγάλουμε τα συμπεράσματά μας. Για την διασπορά του $lnHR$ από την οποία θα παράγουμε και τα αντίστοιχα βάρη για κάθε έρευνα της μετα-ανάλυσης ισχύει η σχέση

$$V_{lnHR} = \frac{1}{E_N} + \frac{1}{E_S}, \tag{4.4}$$

κι έτσι για το διαστήματα εμπιστοσύνης έχουμε

$$lnHR \pm z_{a/2} SE_{lnHR}, \tag{4.5}$$

όπου

$$SE_{lnHR} = \sqrt{V_{lnHR}}. \tag{4.6}$$

Εναλλακτικά μπορούμε να χρησιμοποιήσουμε την εκτίμηση των Mantel-Haenszel για το HR. Η οποία δίνεται από τις σχέσεις

$$HR = exp\{\frac{O_N - E_N}{V}\}, \tag{4.7}$$

με

$$V = \frac{n_N \cdot n_S \cdot d \cdot s}{N^2(N-1)}, \tag{4.8}$$

όπου s είναι ο αριθμός των ατόμων που δεν τους συνέβει το καταληκτικό γεγονός. Σε αυτή την εκτίμηση η διασπορά δίνεται από τον τύπο

$$V_{lnHR} = \frac{1}{V}. \tag{4.9}$$

Εκτίμηση του HR με βάση τα αναλογικά μοντέλα

Το πιο συνηθισμένο βέβαια στις διάφορες έρευνες που μας παρέχονται για να παράγουμε μέσω αυτών μια μετα-ανάλυση είναι ότι θα έχουν χρησιμοποιηθεί παραμετρικές ή ημί-παραμετρικές μέθοδοι και μοντελοποιήσεις. Όπως αναφέραμε στο προηγούμενο κεφάλαιο η συνάρτηση κινδύνου για τις δυο ομάδες στα μοντέλα αλλά και στα

παραμετρικά αναλογικά μοντέλα δίνεται από την σχέση

$$h_N(t) = e^{\beta} h_S(t),$$

οπότε έχουμε

$$\frac{h_N(t)}{h_S(t)} = e^{\beta} = HR, \tag{4.10}$$

δηλαδή ο λόγος του κινδύνου των δύο ομάδων ισούται με τον συντελεστή $\psi = e^{\beta}$, οποίος παραμένει σταθερός όλες τις χρονικές στιγμές (κάτω από την συγκεκριμένη μοντελοποίηση). Μέσω της εκτίμησης του β από τις μεθόδους που περιγράφτηκαν στο προηγούμενο κεφάλαιο για τα ημι-παραμετρικά και για παραμετρικά μοντέλα (αναλόγως και της κατανομής της baseline) και με χρήση της μεθόδου Δέλτα, έχουμε για την διασπορά αυτού

$$V_{HR} = e^{2\hat{\beta}} \cdot var(\hat{\beta}). \tag{4.11}$$

Εκτίμηση του HR με βάση τα AFT μοντέλα
Στα μοντέλα επιταχυνόμενου χρόνου αποτυχίας όπως είδαμε ισχύει η σχέση

$$h_N(t) = \phi^{-1} h_s(t/\phi),$$

αυτό μας δείχνει ότι το HR δεν είναι σταθερό μέσα στο χρόνο, όπως ήταν στα αναλογικά μοντέλα, η εκτίμηση του HR και της διασποράς του κάποια χρονική στιγμή, καθορίζεται και από την κατανομή που ακολουθεί η baseline συνάρτηση κινδύνου (στο συγκεκριμένο παράδειγμα η συνάρτηση κινδύνου της υπάρχουσας θεραπείας).

4.2 Λόγος των ποσοστημορίων (Percentile Ratio)

Όπως αναφέραμε ο λόγος του κινδύνου, είναι το πιο συνηθισμένο μέτρο και επιπλέον έχει και μια κατανοητή φυσική ερμηνεία σε μια μετα-ανάλυση. Όμως σε μια μετα-ανάλυση η υπόθεση της αναλογικότητας γίνεται ακόμα πιο πολύπλοκη, αφού έχουμε

να κάνουμε με ένα σύνολο από μελέτες στις οποίες μπορεί αυτή η υπόθεση να μην ισχύει. Αν τελικά σε κάποιες από τις μελέτες δεν υπάρχει αναλογικότητα (πχ AFT μοντέλα) ο εκτιμώμενος λόγος του κινδύνου στηρίζεται στον χρόνο που διήρκησε η έρευνα και μπορεί να δημιουργήσει πρόβλημα στην μετα-ανάλυσή μας.

Ένα νέο μέτρο το οποίο μπορούμε να χρησιμοποιήσουμε στη μετα-ανάλυσή μας και να παρακάμψει αυτό το πρόβλημα είναι ο λόγος των ποσοστημορίων (percentile ratio - PR). Ο λόγος των ποσοστημορίων q_k στο $k-$οστό ποσοστημόριο των συνάρτησεων επιβίωσης των δύο ομάδων που έχουμε ορίσει (νέα θεραπεία-συνήθης θεραπεία) ορίζεται ως εξής

$$q_k = \frac{t_N(k)}{t_S(k)},\ k \in [0,1] \tag{4.12}$$

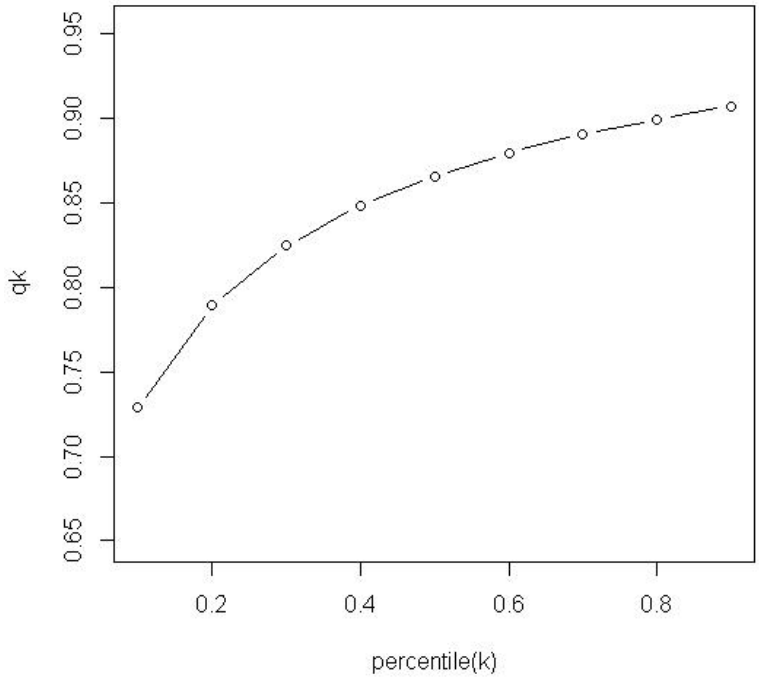
όπου

$$S_N(t_N(k)) = k = S_s(t_S(k)).$$

Για να γίνει πιο κατανοητή η φυσική ερμηνεία του λόγου των ποσοστημορίων, το $q_{0.5}$, δηλαδή το PR της διαμέσου των δύο ομάδων, αντικατοπτρίζει το πόσες φόρες η διάμεσος της μιας ομάδας είναι μεγαλύτερη ή μικρότερη από την διάμεσο της αλλής ομάδας, με τιμές μεγαλύτερης της μονάδας να σημαίνουν ότι ο χρόνος για να συμβεί το καταληκτικό γεγονός στον μισό πληθυσμό των ατόμων που έλαβαν την καινούρια θεραπεία ήταν μεγαλύτερος από τον αντίστοιχο χρόνο των ατόμων της συνήθους θεραπείας, ενώ έχουμε αντίθετα συμπεράσματα για τιμές μικρότερες του 1 και τέλος για τιμές πολύ κοντά ή ίσες με 1 παρατηρούμε παρόμοιες χρονικές στιγμές για τις δυο ομάδες. Επίσης γίνεται να χρησιμοποιηθεί οποιοδήποτε ποσοστημόριο, παραδείγματος χάρην αν έχουμε $q_{0.2} = 1.5$ αυτό μεταφράζεται ως εξής: Το 20 % του πληθυσμού των ατόμων που λαμβάναν την νέα θεραπεία έχει επιβιώσει κατά 50% περισσότερο χρόνο σε σύγκριση με το αντίστοιχο 20 % των ατόμων που λαμβάναν την συνήθη θεραπεία, δηλαδή για να γίνει ακόμα πιο κατανοητό άμα το 0.2 ποσοστημόριο της συνήθους θεραπείας είναι 1000 μέρες το αντίστοιχο της νέας θα είναι 1500 ημέρες.

Σε ένα γενικότερο μοτίβο, το q_k παίρνει τιμές σαν μια συνάρτηση του $k \in [0,1]$, αν και συνήθως επιλέγουμε να χρησιμοποιήσουμε τις τιμές στο διάστημα [0.1,0.9], αφού οι ακραίες τιμές δεν μας δίνουν σωστή πληροφορία καθότι η μια υποδηλώνει την αρχή μια έρευνας όπου όλοι είναι ζωντανοί, ενώ η άλλη την χρονική στιγμή όπου όλος ο πληθυσμός μια ομάδας θα έχει πάθει το καταληκτικό γεγονός που πρακτικά η

έρευνα θα έχει τελειώσει πριν συμβεί αυτό. Οπότε γενικά δουλεύουμε στο διάστημα $k \in (0, 1)$.



Σχήμα 4.1: Γράφημα του q_k από αναλογικό μοντέλο με log-logistic baseline

4.2.1 Εκτίμηση λόγου των ποσοστημορίων απαραμετρικά

Έχοντας ορίσει ως ποσοστημόριο της συνάρτησης επιβίωσης την χρονική στιγμή t_k για την οποία ισχύει η σχέση $S(t_k) = k \Rightarrow t_k = S^{-1}(k)$, έχουμε ότι το $k-$οστό ποσοστημόριο εκτιμάται από την σχέση

$$\hat{t}_K = min\{t : \hat{S}(t) < k\} \tag{4.13}$$

Επίσης γνωρίζουμε ότι λόγω της μεθόδου Δέλτα ισχύει για συνεχής συνάρτηση $g(.)$

$$var(g(X)) \approx \left(\frac{dg(X)}{dX}\right)^2 var(X)$$

κι έτσι

$$var(\hat{S}(\hat{t}_k)) = \left(\frac{d\hat{S}(\hat{t}_k)}{d\hat{t}_k}\right)^2 var(\hat{t}_k) = (\hat{f}(\hat{t}_k))^2 var(\hat{t}_k)$$

λύνοντας την παραπάνω σχέση ως προς $var(\hat{t}_k)$ έχουμε την διασπορά του $k-$οστού ποσοστημορίου

$$var(\hat{t}_k) = \left(\frac{1}{\hat{f}(\hat{t}_k)}\right)^2 var(\hat{S}(\hat{t}_k)) \tag{4.14}$$

και για το τυπικό σφάλμα έχουμε την σχέση

$$se(\hat{t}_k) = \frac{se(\hat{S}(\hat{t}_k))}{\hat{f}(\hat{t}_k)}. \tag{4.15}$$

Για το $se(\hat{S}(\hat{t}_k))$ χρησιμοποιούμε την σχέση (3.12), ενώ η εκτιμήτρια της συνάρτησης πυκνότητας του χρόνου επιβίωσης δίνεται από την σχέση

$$\hat{f}(\hat{t}_k) = \frac{\hat{S}(\hat{u}_k) - \hat{S}(\hat{l}_k)}{\hat{l}_k - \hat{u}_k} \tag{4.16}$$

όπου

$$\hat{u}_k = max\{t|\hat{S}(t) \geq k + \epsilon\} \tag{4.17}$$

$$\hat{l}_k = min\{t|\hat{S}(t) \leq k - \epsilon\} \tag{4.18}$$

για μικρές τιμές του ϵ (συνήθως $\epsilon = 0.05$).

Εκτίμηση του log PR

Όπως και στην περίπτωση του λόγου κινδύνου, έτσι και στον λόγο των ποσοστημορίων θα χρησιμοποιήσουμε την λογαριθμική κλίμακα αυτού. Για την εκτίμηση του $logq_k$ έχουμε

$$log\hat{q}_k = log\frac{\hat{t}_N(k)}{\hat{t}_S(k)} = log\hat{t}_N(k) - log\hat{t}_S(k) \tag{4.19}$$

και για την διασπορά λόγω του ότι

$$Var(log\hat{t}_N(k) - log\hat{t}_S(k)) = Var(log\hat{t}_N(k)) + Var(log\hat{t}_S(k))$$

έχουμε

$$Var(log\hat{q}_k) = Var(log\hat{t}_N(k)) + Var(log\hat{t}_S(k)) \tag{4.20}$$

όπου με χρήση της μεθόδου Δέλτα στην σχέση (4.14) μπορούμε να βρούμε την

$$Var(log\hat{t}(k)) \approx \frac{Var(\hat{t}(k))}{\hat{t}^2(k)} = \frac{Var(\hat{S}(\hat{t}(k))}{[\hat{f}(\hat{t}(k))\hat{t}(k)]^2} \tag{4.21}$$

Εκτίμηση του πίνακα συνδιακύμανσης με χρήση της μεθόδου bootstrap

Ένας εναλλακτικός τρόπος για να εκτιμήσουμε τη διασπορά αλλά και λοιπά μέτρα θέσης είναι η μέθοδος bootstrap. Για να περιγράψουμε την μέθοδο ας υποθέσουμε την γενική περίπτωση όπου έχουμε παρατηρήσει ανεξάρτητο και ισόνομο δείγμα $X_1, X_2, \ldots X_n$ από κάποια κατανομή με συνάρτηση πιθανότητας F. Τα X_i μπορεί να ανήκουν στους πραγματικούς, στο διδιάστατο χώρο ή ακόμα και σε πολύπλοκους χώρους διαστάσεων. Έστω ότι θέλουμε να εκτιμήσουμε κάποια μέτρα θέσης από την παράμετρο $\theta(F)$, (πχ την μέση τιμή, διάμεσο κλπ) και χρησιμοποιούμε κάποια σχέση $\hat{\theta} = \theta(\hat{F})$, όπου $\hat{F}$ η εμπειρική συνάρτηση πιθανότητας, όπου θέτουμε για κάθε τιμή του δείγματος αθροιστικά $1/n$. Δηλαδή αν θεωρήσουμε πως $Y_1, Y_2, \ldots Y_n$ είναι το διατεταγμένο δείγμα των X_i τότε η εμπειρική συνάρτηση πιθανότητας είναι η

$$\hat{F}(x) = \begin{cases} 0, & x < y_1 \\ k/n, & y_k \leq x < y_{k+1} \\ 1, & y_n \leq x \end{cases}$$

Έστω τώρα ότι θέλουμε να εκτιμήσουμε το μέτρο θέσης $\sigma(F)$, το οποίο ας υποθέσουμε για να γίνει και πιο κατανοητή η διαδικασία ότι είναι η τυπική απόκλιση. Η εκτίμηση που θα πάρουμε με την μέθοδο bootstrap είναι η $\hat{\sigma}_{bootstrap} = \sigma(\hat{F})$, δηλαδή η τυπική απόκλιση που θα πέρναμε μέσω της μεθόδου μέγιστης πιθανοφάνειας αν η

εμπειρική συνάρτηση πιθανότητας ήταν η πραγματική.
Για να υπολογίσουμε την $\hat{\sigma}_{bootstrap}$ συνήθως χρησιμοποιούνται μέθοδοι προσομοίωσης. Έχουμε τα εξής τέσσερα βήματα:

1. Παράγουμε ένα δείγμα $X_1^*, X_2^*, \ldots, X_n^*$ από την $\hat{F}$, όπου το κάθε X_i^* παίρνει την τιμή x_j με πιθανότητα $1/n$, δηλαδή πρακτικά δημιουργούμε ένα δείγμα από n X_i τα οποία επιλέχθηκαν με επανάθεση από το αρχικό μας δείγμα.
2. Αυτό μας δίνει μια bootstrap εμπειρική συνάρτηση πιθανότητας από την οποία μπορούμε να υπολογίσουμε την $\hat{\theta}^* = \theta(\hat{F}^*)$.
3. Τα βήματα 1) και 2) επαναλαμβάνονται N φορές (για μεγάλο N) και έτσι δημιουργούνται οι bootstrap τιμές $(\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_N^*)$.
4. Η τιμή του $\hat{\sigma}_{bootstrap}$ προσεγγίζεται από την σχέση που υπολογίζεται το μέτρο θέσης που επιθυμούμε να εκτιμήσουμε, π.χ. στην περίπτωση που το ζητούμενο μέτρο θέσης είναι η τυπική απόκλιση $\sigma(F)$, υπολογίζεται μέσω των bootstrap τιμών $\hat{\theta}^*$, από τον τύπο του δειγματικού μέσου

$$\hat{\sigma}_{bootstrap} = \frac{\sum_{j=1}^{N}(\hat{\theta}_j^*)^2 - (\sum_{j=1}^{N}\hat{\theta}_j^*)^2/N}{N-1}.$$

Με ανάλογο τρόπο μπορούν να υπολογιστούν και άλλα μέτρα θέσης. Στην περίπτωση που υπάρχουν δεξιά λογοκριμένα δεδομένα η διαδικασία είναι παρόμοια [14], αλλά στο βήμα 1) παράγουμε δείγμα από την διδιάστατη μεταβλητή (t_i, d_i) όπου t_i ο χρόνος καταληκτικού γεγονότος και d_i η δείκτρια, του αν το γεγονός είναι λογοκριμένο ή όχι.

4.2.2 Λόγος των ποσοστημορίων στα παραμετρικά μοντέλα

Στο προηγούμενο κεφάλαιο μιλήσαμε για τα παραμετρικά μοντέλα όπου η baseline συνάρτηση κινδύνου ακολουθεί κάποια γνωστή κατανομή, αυτά μπορεί να είναι αναλογικά, όπως και AFT. Λόγω του ότι οι κατανομές των χρόνων εκτιμούνται μπορούμε να βρούμε τις εκτιμήσεις των ποσοστημορίων για κάθε ομάδα, όπως και τον λόγο των ποσοστημορίων των δύο ομάδων. Θα αναφέρουμε παρακάτω τα αποτελέσματα για τις κύριες κατανομές που μιλήσαμε στο προηγούμενο κεφάλαιο, αφού πρώτα δείξουμε

την γενική σχέση που ισχύει για τις συναρτήσεις κινδύνου στα αναλογικά μοντέλα. Λόγω του ότι

$$h_N(t) = e^{\beta} h_S(t)$$

έχουμε

$$S_N(t) = exp\{-H_N(t)\} = exp\{-e^{\beta} \int_0^t h_S(t)\} = (S_S(t))^{e^{\beta}}. \tag{4.22}$$

Εκθετική κατανομή

Στα αναλογικά μοντέλα τα οποία ακολουθούν εκθετική κατανομή ο λόγος των ποσοστημορίων είναι σταθερός, αφού αν θεωρήσουμε ότι $S_S(t) = e^{-\lambda t}$, τότε λόγω της (4.22) έχουμε

$$S_N(t) = exp\{-e^{\beta}\lambda t\}$$

άρα

$$S_N(t_N) = S_S(t_S) = k \Rightarrow exp\{-e^{\beta}\lambda t_N\} = exp\{-\lambda t_S\}$$

κι έτσι βρίσκουμε ότι

$$q_k = \frac{t_N(k)}{t_S(k)} = e^{-\beta}. \tag{4.23}$$

Έτσι με τι διαδικασίες που περιγράφηκαν στην παράγραφο 3.4 εκτιμώντας της παραμέτρους του αναλογικού μοντέλου, έχουμε την εκτίμηση του λόγου των ποσοστημορίων

$$\hat{q}_k = e^{-\hat{\beta}}. \tag{4.24}$$

Για τη διασπορά της εκτίμησης του q_k, μέσω της μεθόδου δέλτα βρίσκουμε ότι ισχύει η σχέση

$$V(\hat{q}_k) = e^{-2\hat{\beta}} \cdot V(\hat{\beta}) = \hat{q}_k^2 \cdot V(\hat{\beta}). \tag{4.25}$$

Weibull κατανομή

Το ότι το PR παραμένει σταθερό μπορούμε να το δείξουμε και στα αναλογικά μοντέλα με Weibull κατανομή, αφού θεωρώντας ότι $S_S(t) = e^{-\lambda t^{\gamma}}$, έχουμε αντίστοιχα

$$S_N(t) = exp\{-e^{\beta}\lambda t^{\gamma}\}$$

$$S_N(t_N) = S_S(t_S) = k \Rightarrow exp\{-e^{\beta}\lambda t_N^{\gamma}\} = exp\{-\lambda t_S^{\gamma}\}$$

κι έτσι

$$q_k = \frac{t_N(k)}{t_S(k)} = e^{-\frac{\beta}{\gamma}}, \tag{4.26}$$

και άρα η εκτιμήτρια του λόγου ποσοστημορίων μέσω του αναλογικού μοντέλου είναι η

$$\hat{q}_k = e^{-\frac{\hat{\beta}}{\hat{\gamma}}}. \tag{4.27}$$

Η διασπορά της εκτίμησης του q_k βρίσκεται μέσω της πολυδιάστατης μεθόδου δέλτα, αφού έχουμε δύο εκτιμημένες παραμέτρους στην σχέση που μας δίνει το q_k, το β και το γ. Δηλαδή έχουμε

$$V(\hat{q}_k) = V(e^{-\frac{\hat{\beta}}{\hat{\gamma}}}) = \nabla h^T(\hat{\beta}, \hat{\gamma}) \cdot Cov(\hat{\beta}, \hat{\gamma}) \cdot \nabla h(\hat{\beta}, \hat{\gamma}) \tag{4.28}$$

όπου

$$\nabla h(\hat{\beta}, \hat{\gamma}) = \left(\frac{d(e^{-\frac{\hat{\beta}}{\hat{\gamma}}})}{d\hat{\beta}}, \frac{d(e^{-\frac{\hat{\beta}}{\hat{\gamma}}})}{d\hat{\gamma}} \right) = \left(-\frac{\hat{q}_k}{\hat{\gamma}}, \frac{\hat{\beta}\hat{q}_k}{\hat{\gamma}^2} \right) \tag{4.29}$$

log-logistic κατανομή

Τα αναλογικά μοντέλα με log-logistic κατανομή της baseline παρουσιάζουν ίσως το μεγαλύτερο ενδιαφέρον αφού ο λόγος των ποσοστημορίων δεν είναι σταθερός και είναι μια συνάρτηση του $k \in [0, 1]$. Θεωρώντας ότι $S_S(t) = \frac{1}{1+(\lambda t)^{\tau}}$ και $h_S(t) = \frac{\lambda\tau(\lambda t)^{\tau-1}}{1+(\lambda t)^{\tau}}$ και αφού παρατηρήσουμε πως

$$H_S(t) = \int_0^t \frac{\lambda\tau(\lambda u)^{\tau-1}}{1+(\lambda u)^{\tau}} du = log(1 + (\lambda t)^{\tau})$$

δουλεύοντας όπως στην σχέση (4.22) παίρνουμε

$$S_N(t) = exp\{-e^{\beta} log(1 + (\lambda t)^{\tau})\}.$$

Για να βρούμε τώρα το $k-$οστό ποσοστημόριο λύνουμε την εξίσωση ως προς t

$$S_N(t) = k \Rightarrow -e^{\beta} log(1 + (\lambda t)^{\tau}) = logk \Rightarrow log(1 + (\lambda t)^{\tau}) = logk^{-exp\{-\beta\}}$$

$$\Rightarrow 1+(\lambda t)^{\tau}=k^{-exp\{-\beta\}}\Rightarrow t=\left(\frac{k^{-exp\{-\beta\}}-1}{\lambda^{\tau}}\right)^{\frac{1}{\tau}}$$

με όμοιο τρόπο βρίσκουμε και το $k-$οστό ποσοστημόριο για την $S_S(t)$ κι έτσι έχουμε το PR

$$q_k=\left(\frac{k^{-exp\{-\beta\}}-1}{k^{-1}-1}\right)^{\frac{1}{\tau}}, \tag{4.30}$$

οπότε για την εκτιμήτρια ισχύει η σχέση

$$\hat{q}_k=\left(\frac{k^{-exp\{-\hat{\beta}\}}-1}{k^{-1}-1}\right)^{\frac{1}{\hat{\tau}}}, \tag{4.31}$$

και για την διασπορά, με πολυδιάστατη μέθοδο Δέλτα έχουμε

$$V(\hat{q}_k)=\nabla h^T(\hat{\beta},\hat{\tau})\cdot Cov(\hat{\beta},\hat{\tau})\cdot\nabla h(\hat{\beta},\hat{\tau}) \tag{4.32}$$

όπου

$$\nabla h(\hat{\beta},\hat{\tau})=\left(\frac{d(\hat{q_k})}{d\hat{\beta}},\frac{d(\hat{q_k})}{d\hat{\tau}}\right) \tag{4.33}$$

AFT μοντέλα

Για τα AFT μοντέλα ισχύει ότι γενικά το PR παραμένει σταθερό. Αυτό ισχύει αφού λόγω της (3.54), έχουμε

$$S_N=S_S\left(\frac{logt_N-\mu-\alpha}{\sigma}\right)=k=S_S\left(\frac{logt_S-\mu}{\sigma}\right)\Rightarrow logt_N-\alpha=logt_S$$

άρα τελικά

$$q_k=\frac{t_N}{t_s}=e^{\alpha}. \tag{4.34}$$

και έτσι μέσω των μεθόδων που περιγράψαμε στην ενότητα 3.5 έχουμε την εκτιμήτρια του

$$\hat{q}_k=e^{\hat{\alpha}} \tag{4.35}$$

ενώ για την διασπορά ισχύει η σχέση

$$V(\hat{q}_k) = \hat{q}_k^2 \cdot V(\hat{\alpha}). \tag{4.36}$$

Εδώ να σημειωθεί ότι ανάλογα με τις παραμετροποιήσεις που έχουμε σε ένα μοντέλο, όπως παραδείγματος χάρην είχαμε αναφέρει την λογαριθμο-γραμμική μορφή των AFT μοντέλων στην παράγραφο 3.5.2, αλλά και τα σχετικά που ισχύουν στα αναλογικά μοντέλα, οι διάφορες εκτιμήσεις του λόγου ποσοστημορίων και των διασπορών τους υπολογίζονται με όμοιο τρόπο και προφανώς τα αποτελέσματα είναι παρόμοια. Επίσης σημαντικό είναι το πως έχουμε ορίσει τα ποσοστημόρια αφού ανάλογα την βιβλιογραφία μπορεί να υπολογίζονται είτε μέσω της σχέσης $S(t) = k$, είτε μέσω της σχέση $S(t) = 1 - k$.

4.3 Μέθοδοι μετα-ανάλυσης για τον λόγο των ποσοστημορίων

Αφού περιγράψαμε στην προηγούμενη παράγραφο τους διάφορους τρόπους υπολογισμού της εκτίμησης του λόγου ποσοστημορίων και της διασποράς του, κυρίως όσον αφορά περιπτώσεις όπου μελετάμε την διαφορά μεταξύ δύο ομάδων, πχ περίπτωση νέας θεραπείας - συνηθισμένης θεραπείας θα δούμε παρακάτω πως μπορούμε να τα χρησιμοποιήσουμε όταν μας δίνονται σαν αποτελέσματα κάποιων μελετών με σκοπό να τα συμπεριλάβουμε σε μια μετα-ανάλυση από την οποία να παράγουμε νέα και πιο γενικά συμπεράσματα, για τα οποία τα πλεονεκτήματά τους περιγράφθηκαν στο πρώτο κεφάλαιο, όπου μιλήσαμε για την μετα-ανάλυση. Στην πράξη θα αναφερθούμε σε τεχνικές μετα-ανάλυσης όπου το μέγεθος της επίδρασης effect size, θα είναι ο λόγος των ποσοστημορίων $q_k,\ k = 0.1, \cdots, 0.9$ και επιθυμούμε να βγάλουμε ένα γενικό συμπέρασμα summary effect των q_k για το σύνολο κάποιων μελετών.

Στην πιο απλή μορφή μιας μετα-ανάλυσης, θα είχαμε ένα σύνολο ερευνών όπου το αποτέλεσμα που μας δίνεται για την κάθε μελέτη είναι ο λόγος των ποσοστημορίων και η αντίστοιχη διασπορά, όμως λόγω του ότι ο λόγος των ποσοστημορίων δεν είναι τόσο πολύ διαδεδομένος, αυτό θα ήταν πολύ σπάνιο. Μια περίπτωση από την οποία θα μπορούσαμε να παράγουμε δεδομένα για την μετα-ανάλυσή μας είναι αυτή όπου

μας δίνονται για κάθε έρευνα το μοντέλο που χρησιμοποιήθηκε και οι εκτιμήσεις των παραμέτρων του και μέσω αυτών παράγουμε το effect size q_k της κάθε μετα-ανάλυσης και τις όποιες εκτιμήσεις περιγραφικών μέτρων θέσης χρειαζόμαστε.

Μια άλλη περίπτωση η οποία είναι και η ιδανική γενικότερα για μια μετα-ανάλυση είναι όταν για κάθε έρευνα δεν μας δίνονται μόνο τα αποτελέσματα αυτής αλλά και όλα τα δεδομένα που χρησιμοποιήθηκαν. Στην βιοστατιστική αυτά λέγονται individual patient data, δηλαδή τα δεδομένα του κάθε ασθενή τα οποία θα αναφέρουμε για συντομογραφία σαν IPD. Μέσω αυτών μπορούμε να μοντελοποιήσουμε αρχικά είτε απαραμετρικά, είτε παραμετρικά ανάλογα πως κρίνουμε εμείς σωστό την κάθε έρευνα και να παράγουμε τα ζητούμενα αποτελέσματα, στην συγκεκριμένη περίπτωση το q_k, είτε όπως θα δούμε αργότερα να συνδυάσουμε όλα τα δεδομένα σε μια κοινή διαδικασία. Στα IPD δεδομένα την περίπτωση όπου πρώτα μοντελοποιούμε με κάποιο τρόπο και μετά παράγουμε μια μετα-ανάλυση την λέμε μετα-ανάλυση δύο φάσεων (two stage meta-analysis), ενώ όταν συνδυάζουμε όλα τα δεδομένα σε μια διαδικασία την λέμε μετα-ανάλυση μιας φάσης.

4.3.1 Μετα-ανάλυση με χρήση της μεθόδου εύρεσης σταθμισμένου μέσου

Θα περιγράψουμε τη διαδικασία μια μετα-ανάλυσης έχοντας IPD δεδομένα, στην αντίθετη περίπτωση θα έπρεπε οι έρευνες που μας παρέχονταν να μας δίνουν τα απαραίτητα στοχεία ώστε να παράγουμε τον λόγο των ποσοστημοριών για κάθε έρευνα όπως και τη διασπορά τους, σε αυτήν την περίπτωση θα παραλείπαμε την φάση ένα και θα είχαμε απλά τη διαδικασία μετατροπής της απαντητικής πληροφορίας στην ζητούμενη ή αν μας δίνονταν κατευθείαν οι εκτιμήσεις των q_k θα περνούσαμε αμέσως στη δεύτερη φάση.

Ας υποθέσουμε πως έχουμε N έρευνες και για καθεμία από αυτές μας δίνονται τα δεδομένα τους. Ας θεωρήσουμε πως οι έρευνες μας είναι της μορφής σύγκρισης δύο ομάδων, νέα θεραπεία - συνήθης θεραπείας. Στην πρώτη φάση θα εκτιμήσουμε από κάθε έρευνα τα q_k για $k = 0.1, \cdots, 0.9$, αυτό μπορεί να γίνει είτε απαραμετρικά, είτε παραμετρικά και είναι ανάλογο της φύσης των δεδομένων. Θα δουλέψουμε στην λογαριθμική κλίμακα. Οπότε για την απαραμετρική περίπτωση εκτιμούμε τον λογάριθμο των λόγων των ποσοστημορίων για κάθε k, μέσω της σχέσης (4.19) και

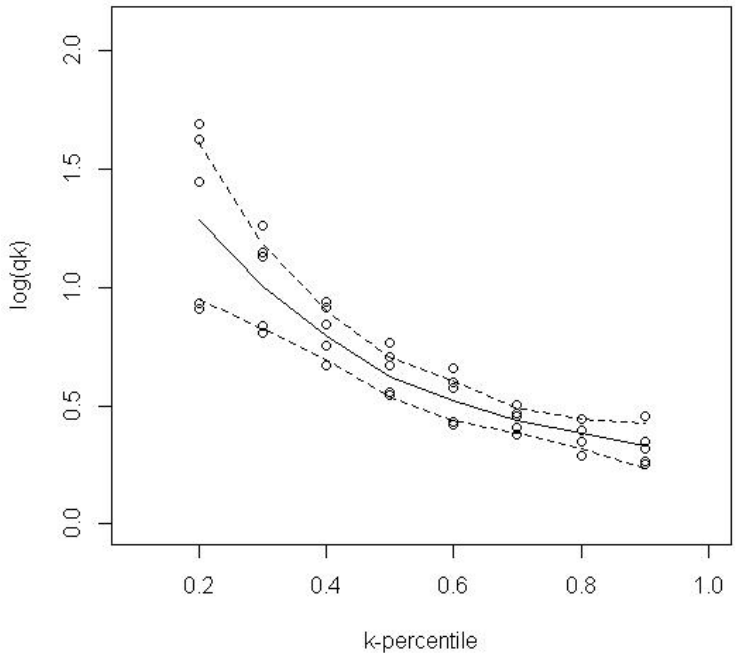
των λοιπών σχέσεων από την παράγραφο 4.2.1 και τις αντίστοιχες διασπορές είτε μέσω στων σχέσεων (4.20) - (4.21), είτε μέσω της διαδικασίας bootstrap την οποία περιγράψαμε στην ίδια αυτή παράγραφο. Συνήθως η μέθοδος bootstrap δίνει καλύτερα αποτελέσματα [16]. Αντίστοιχα τώρα στην περίπτωση που θέλουμε να εκτιμήσουμε τα ζητούμενα μέτρα επίδρασης παραμετρικά, ανάλογα και το παραμετρικό μοντέλο που θα χρησιμοποιήσουμε, ακολουθούμε τις διαδικασίες που περιγράφηκαν στην παράγραφο 4.2.2.

Έτσι έχουμε τις εκτιμήσεις των N διανυσμάτων q_k με τις αντίστοιχες διασπορές τους. Πρακτικά τα δεδομένα μας για την μετανάλυση θα είναι της μορφής:

$StudyNo$	$log\hat{q}_k$	$V(log\hat{q}_k)$
1	$(log\hat{q}_{1,0.1}, \cdots, log\hat{q}_{1,0.9})$	$(V(log\hat{q}_{1,0.1}), \cdots, V(log\hat{q}_{1,0.9}))$
.	.	.
.	.	.
.	.	.
N	$(log\hat{q}_{N,0.1}, \cdots, log\hat{q}_{N,0.9})$	$(V(log\hat{q}_{N,0.1}), \cdots, V(log\hat{q}_{N,0.9}))$

Οπότε μπορούμε να περάσουμε στη δεύτερη φάση της μετα-ανάλυσης με βάση τον σταθμισμένο λόγο των ποσοστημορίων. Όπως περιγράψαμε στην παράγραφο 1.3, υπάρχουν δύο είδη μοντέλων τα fixed effect και τα random effect, όπου στο πρώτο υποθέτουμε πως σε όλες τι δοθείσες έρευνες υπάρχει ένα κοινό μέτρο επίδρασης, ενώ στο δεύτερο ότι τα μέτρα επίδρασης διαφέρουν, πρακτικά αυτό σημαίνει ότι στο πρώτο υποθέτουμε πως υπάρχει ομοιογένεια μεταξύ των ερευνών, ενώ στο δεύτερο μοντέλο ετερογένεια, όπως αυτή ορίστηκε στην παράγραφο 1.4.
Ανάλογα με τα δεδομένα, τα συμπεράσματα που έχουμε από αυτά, την προϋπάρχουσα γνώση και τελικά την διαίσθησή μας επιλέγουμε πιο από τα δύο μοντέλα θα χρησιμοποιήσουμε. Παραδείγματος χάρην, άμα έχουμε κρίνει πως το κατάλληλο μοντέλο είναι αυτό με το μοναδικό μέγεθος της επίδρασης και θέλουμε να βγάλουμε συμπεράσματα για το $q_{0.5}$, τότε χρησιμοποιώντας τις σχέσεις της παραγράφου 1.3.1 για $Y_i = log\hat{q}_{i,0.5}$, $V_{Y_i} = V(log\hat{q}_{i,0.5})$, για $i = 1, \cdots, N$ και μετά αντιστρέφοντας την λογαριθμική σχέση έχουμε τα ζητούμενα αποτελέσματα της μετα-ανάλυσης. Επαναλαμβάνοντας εφ΄όσον είναι επιθυμητό την ίδια διαδικασία, μπορούμε να βγάλουμε συμπεράσματα για όσα $k = 0.1, \cdots, 0.9$ έχουμε την απαραίτητη πληροφορία. Αντίστοιχα μέσω των σχέσεων

της παραγράφου 1.3.2, δουλεύουμε στο μοντέλο με τυχαία μεγεθη της επίδρασης, όπου πρέπει να εκτιμηθεί και η ετερογένεια. Στο Σχήμα 4.2 έχουμε τα αποτελέσματα μιας τέτοια είδους μετα-ανάλυσης από 5 έρευνες στην οποία θεωρήσαμε σαν σωστή μοντελοποίηση τα random effect μοντέλα.



Σχήμα 4.2: Μετα-ανάλυση του λόγου των ποσοστημορίων βασισμένη σε 5 έρευνες. Τα σημεία είναι οι εκτίμησεις για κάθε ποσοστημόριο του q_k, η γραμμή είναι το summary effect, και οι διακεκομένες το διάστημα εμπιστοσύνης του.

Ας αναφέρουμε τώρα έναν εναλλακτικό τρόπο που μπορούμε να χρησιμοποιήσουμε στην πρώτη φάση της μετα-ανάλυσηςσε IPD δεδομένα, όταν έχουμε επιλέξει παραμετρικά μοντέλα, που εκτιμούμε τα effect size από κάθε έρευνα. Όπως είχαμε αναφέρει στο κεφάλαιο 3 η εκτίμηση των παραμέτρων γίνεται μέσω των εκτιμητών μέγιστης πιθανοφάνειας. Ας υποθέσουμε ότι η κατανομή του χρόνου καταληκτικού γεγονότος της $i-$οστής έρευνας ακολουθεί κατανομή με πυκνότητα $f_i(t;\beta,\omega)$, όπου β είναι ο

συντελεστής της δείκτριας για το αν ο χρόνος ανήκει στην νέα θεραπεία ή στην συνηθισμένη και ω ένα διάνυσμα με τις μεταβλητές της κατανομής. Τότε αφού β είναι μια συνάρτηση του q_k και του ω, η συνάρτηση πυκνότητας μπορεί να γραφτεί στην μορφή $f_i(t; q_k, \omega)$ (ή αν θέλουμε μπορούμε και του $logq_k$). Οπότε χρησιμοποιώντας την εκτίμηση μέγιστης πιθανοφάνειας μπορούμε να εκτιμήσουμε με χρήση των σχετικών αριθμητικών διαδικασιών την εκτιμήτρια του λόγου πιθανοφανειών για κάθε k, όπως και την διασπορά της. Η συνολική πιθανοφάνεια της $i-$οστής έρευνας δίνεται μέσω της σχέσης

$$\prod_{j=1}^{r} f_i(t_j; q_k, \omega) \cdot \prod_{l=1}^{n-r} S_i(t_l^*; q_k, \omega)$$

όπου t^* είναι οι λογοκριμένοι χρόνοι και συνεισφέρουν στην πιθανοφάνεια, αφού γνωρίζουμε ότι μέχρι αυτή την στιγμή τα άτομα στα οποία αναφέρονται δεν έχουν πάθει το καταληκτικό γεγονός, δηλαδή η πιθανότητα αυτή είναι ίση με $P(T \geq t^*) = S(t^*)$, όπου r το πλήθος των μη λογοκριμένων χρόνων και n το συνολικό πλήθος των χρόνων. Θεωρώντας την δείκτρια $\delta = 1$ για το αν ο χρόνος είναι μη λογοκριμένος, η παραπανω σχέση γίνεται

$$L(q_{i,k}) = \prod_{u=1}^{n} f_i^{\delta_u}(t_u; q_k, \omega) \cdot S_i^{1-\delta_u}(t_u; q_k, \omega). \tag{4.37}$$

4.3.2 Παραμετρική μετα-ανάλυση μιας φάσης, δεδομένων μέχρι το γεγονός

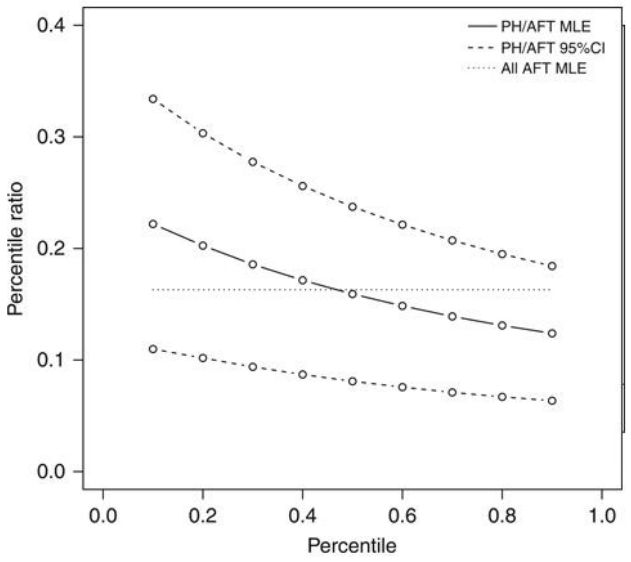
Η συγκεκριμένη μέθοδος μετα-ανάλυσης [17] η οποία αποτελεί και το εφαλτήριο της παρούσης εργασίας, μπορεί να εφαρμοστεί σε IPD δεδομένα και είναι μιας φάσης. Όπως περιγράψαμε και προηγουμένως η συνάρτηση της πυκνότητας της $i-$οστής έρευνας $f_i(t; \beta, \omega)$ που έχουμε μοντελοποιήσει κάτω από κάποιο παραμετρικό μοντέλο μπορεί να πάρει την μορφή $f_i(t; q_k, \omega)$, για $i = 1, \cdots, N$. Όπως έχουμε πει το q_k είναι μια ποσότητα με ξεκάθαρη έννοια η οποία δεν έχει να κάνει με την εκάστοτε υπόθεση κατανομής και έτσι σε μια παραμετρική μετα-ανάλυση είναι μια παράμετρος κοινή σαν έννοια σε όλες τις έρευνες. Οπότε για κάθε έρευνα έχουμε ότι η συνάρτηση της κατανομής του χρόνου καταληκτικού γεγονότος γράφεται στην μορφή $f_i(t; q_{ik}, \omega_i)$, όπου q_{ik} είναι ο λόγος των ποσοστημορίων της $i-$οστής έρευνας για κάποιο $k =$

$0.1, \cdots, 0.9$. Η πιο συνήθης υπόθεση στην μετα-ανάλυση είναι ότι η πραγματική τιμή του μεγέθους επίδρασης είναι ίδια σε όλες τις έρευνες, ενώ οι άλλοι παράμετροι διαφέρουν. Έτσι θεωρώντας $q_{ik} = q_k$ για κάθε i, έχουμε ότι η ολική πιθανοφάνεια όλων των ερευνών για συγκεκριμένο k παίρνει την μορφή

$$L(q_k) = \prod_{i=1}^{N} \prod_{u=1}^{n_i} f_i^{\delta_{iu}}(t_{iu}; q_k, \omega_i) \cdot S_i^{1-\delta_{iu}}(t_{iu}; q_k, \omega_i), \tag{4.38}$$

οπότε μπορούμε με χρήση της εκτίμησης μέγιστης πιθανοφάνειας να εκτιμήσουμε το q_k για τα επιθυμητά $k = 0.1, \cdots, 0.9$. Σε περιπτώσεις όπου όλες οι έρευνες ανήκουν σε μοντέλα με κατανομές που δίνουν σταθερό λόγο ποσοστημορίων αυτό είναι περιττό.

Γενικότερα το $\hat{q}_k$, που θα προέλθει από την μεγιστοποίηση της πάνω σχέσης για κάποιο k είναι το summary effect των $\hat{q}_{ik}$. Έτσι το $\hat{q}_k$ πρακτικά είναι μια συνεχής συνάρτηση του $k \in [0.1, 0.9]$.



Σχήμα 4.3: One-stage parametric meta-analysis βασισμένη σε έρευνες από αναλογικά μοντέλα με σταθερό και μή σταθερό PR.
Η συνεχής γραμμή είναι το summary effect και οι διακεκομένες το διάστημα εμπιστοσύνης του.

4.3.3 Πολυδιάστατη μετα-ανάλυση δύο φάσεων

Μια γενίκευση των μέθοδων μετα-ανάλυσης όπως περιγράφηκαν κυρίως στο πρώτο κεφάλαιο είναι η πολυδιάστατη μετα-ανάλυση (multivariate meta-analysis) [15]. Πρακτικά είναι αυτό που λέει το όνομα της, δηλαδή μια μετα-ανάλυση η οποία αναφέρεται σε πάνω από δύο μεγέθη που θέλουμε να βγάλουμε συμπεράσματα για αυτά, μεγέθη τα οποία ενδεχομένως να συσχετίζονται με κάποιο τρόπο. Η χρήση της δεν είναι ακόμα τόσο διαδεδομένη και όπως σε όλες τις στατιστικές αναλύσεις υπάρχουν πλεονεκτήματα και μειονεκτήμα, δεν θα εμβαθύνουμε σε αυτά και απλά παρακάτω θα κάνουμε μια σύντομη περιγραφή των τακτικών αυτής και το πως μπορεί να χρησιμοποιηθεί στην μετα-ανάλυση του λόγου των ποσοστημορίων, όπου έχουμε να εκτιμήσουμε το q_k για $k = 0.1, \cdots, 0.9$, και οι τιμές αυτές είναι συσχετισμένες.

Μοντελοποίηση

1) Μοντελοποιώντας εσωτερικά της έρευνας

Ας συμβολίσουμε το διάνυσμα των επιδράσεων (ή των εκτιμήσεων) για την $i-$οστή έρευνα ως Y_i. Αυτές ενδέχεται να είναι συσχετισμένες και έτσι υποθέτουμε ότι εσωτερικά σε κάθε έρευνα ισχύει

$$Y_i|\mu_i \sim N(\mu_i, S_i), \tag{4.39}$$

δηλαδή μια πολυδιάστατη κανονική κατανομή, όπου μ_i είναι η πραγματική επίδραση της $i-$οστής έρευνας και S_i ο πίνακας συνδιακύμανσης του Y_i. Οι πίνακες S_i πρακτικά είναι οι πίνακες συνδιακύμανσης εσωτερικά μιας έρευνας, μπορούν να εκτιμηθούν σε περιπτώσεις IPD δεδομένων για κάθε έρευνα ξεχωριστά και θεωρούνται γνωστοί.

2) Μοντελοποιώντας μεταξύ των ερευνών

Το πολυδιάστατο μοντέλο τυχαίων επιδράσεων το οποίο αναφέρουμε επιτρέπει τα μ_i να διαφέρουν από την μια έρευνα στην άλλη και υποθέτει ότι

$$\mu_i \sim N(\mu, \Sigma), \tag{4.40}$$

με το μ να είναι το (ολικό) διάνυσμα επίδρασης και το Σ να είναι ο πίνακας συνδιακύμανσης μεταξύ των ερευνών. Το μ πρακτικά μεταφράζεται σαν η μέση επίδραση από μια κανονική κατανομή των επιδράσεων των ερευνών. Ο Σ δεν θεωρείται πως έχει κάποια δομή, αλλά μπορούν να γίνουν υποθέσεις για αυτόν.

3) Οριακή μοντελοποίηση

Έτσι οριακά έχουμε το πολυδιάστατο μοντέλο μετα-ανάλυσης τυχαίων επιδράσεων

$$Y_i \sim N(\mu, S_i + \Sigma), \tag{4.41}$$

όπου τα Y_i θεωρούνται ανεξάρτητα αφού προέρχονται από διαφορετικές έρευνες. Στόχος μας είναι να εκτιμήσουμε το μ και το Σ. Από το $\hat{\Sigma}$ μπορούμε να βρούμε τις εκτιμήσεις των συσχετίσεων μεταξύ των ερευνών.

Η εκτίμηση των παραμέτρων γίνεται με διάφορους τρόπους οι οποίοι μπορούν να χωριστούν κυρίως σε δύο κατηγορίες, αυτούς που χρησιμοποιούν άμεσα τον εκτιμημένο πίνακας συνδιακύμανσης μεταξύ των ερευνών σαν να ήταν η πραγματική του τιμή και κάνουν συμπεράματα για το μέγεθος της επίδρασης, και αυτή θεωρείται η συνηθισμένη διαδικασία γιατί είναι πιο ευκολη στην χρήση και αυτές που δεν το κάνουν αυτό. Υποθέτοντας πως όλες οι έρευνες μας παρέχουν την επίδραση τους τότε το $\hat{\mu}$ δίνεται μέσω του $\hat{\Sigma}$ από την σχέση

$$\hat{\mu} = \left(\sum_{i=1}^{n} (S_i + \hat{\Sigma})^{-1} \right)^{-1} \left(\sum_{i=1}^{n} (S_i + \hat{\Sigma})^{-1} Y_i \right), \tag{4.42}$$

όπου n είναι το πλήθος των ερευνών.

Στην συνήθη διαδικασία, η οποία όμως πρέπει να έχει ένα επαρκές μεγάλο δείγμα ερευνών, οι εκτιμήσεις ασυμπτωτικά κατανέμονται στην κανονική με πίνακα συνδιακύμανσης

$$C = Var(\hat{\mu}) = \left(\sum_{i=1}^{n} (S_i + \hat{\Sigma})^{-1} \right)^{-1}. \tag{4.43}$$

Για την εκτίμηση του $\hat{\Sigma}$ έχουν προταθεί πολλές μεθόδοι όπως με μέθοδο μέγιστης πιθανοφάνειας, με μέθοδο περιορισμένης μέγιστης πιθανοφάνειας και μέθοδο στιγμών. Θα αναφέρουμε μόνο τον τύπο από τον οποίο παράγουμε τις εκτιμήσεις στην μέθοδο

περιορισμένης μέγιστης πιθανοφάνειας, οι λοιπές τεχνικές εκτίμησης του $\hat{\Sigma}$, αλλά και γενικότερα της εκτίμησης των παραμέτρων μπορούν να βρεθούν στην σχετική βιβλιογραφία [15].

$$\lambda_{REML} = -\frac{1}{2}\sum_{i=1}^{n} log|S_i+\Sigma| - \frac{1}{2}log|\sum_{i=1}^{n}(S_i+\Sigma)^{-1}| - \frac{1}{2}\sum_{i=1}^{n} {r_i}^T(S_i+\Sigma)^{-1}r_i, \quad (4.44)$$

όπου το r_i είναι τα κατάλοιπα.

Έχοντας τώρα μια ιδέα για τις τεχνικές πίσω από την πολυδιάστατη μετα-ανάλυση, θα περιγράψουμε την ανάλογη διαδικασία για την πολυδιάστατη μετα-ανάλυση του λόγου των ποσοστημορίων [16]. Αρχικά η πρώτη φάση είναι η εκτίμηση των λογαρίθμων των λόγων των ποσσοτημορίων των ερευνών μέσω των απαραμετρικών σχέσεων που περιγράφθηκαν στην παράγραφο 4.2.1, εκτιμώντας την συνδιακύμανση είτε με την εκτιμήτριά του, είτε με διαδικασία bootstrap. Ο τύπος της συνδιακύμανσης δίνεται εδώ ώστε να μπορούν να γίνουν οι ανάλογοι συσχετισμοί και να γίνει πιο κατανοητό πως χρησιμοποιούνται:

$$S_{k_1,k_2} = Cov\{log\hat{q}_{k_1}, log\hat{q}_{k_2}\} =$$

$$= Cov\{log\hat{t}_{Nk_1}, log\hat{t}_{Nk_2}\} + Cov\{log\hat{t}_{Sk_1}, log\hat{t}_{Sk_2}\} \quad (4.45)$$

όπου

$$Cov\{log\hat{t}_{k_1}, log\hat{t}_{k_2}\} \approx \frac{S(\hat{t}_{k_2})Var\{S(\hat{t}_{k_1})\}}{S(\hat{t}_{k_1})f(\hat{t}_{k_1})f(\hat{t}_{k_2})\hat{t}_{k_1}\hat{t}_{k_2}}. \quad (4.46)$$

Έτσι συνεχίζουμε στην δεύτερη φάση έχοντας με βάση τα προηγούμενα το διάνυσμα των $log\hat{q}_i$= $(log\hat{q}_{0.1,i}, \cdots, log\hat{q}_{0.9,i})$ από κάθε έρευνα μαζί με την αντίστοιχη εκτίμηση του πίνακα συνδιακύμανσης S_i. Για την πολυδιάστατη μετα-ανάλυση μας θεωρούμε ότι το $log\hat{q}_i$ ακολουθεί μια πολυδιάστατη κανονική κατανομή

$$log\hat{q}_i \sim N(logq_i, S_i), \quad (4.47)$$

όπου το $logq_i$ είναι το πραγματικό διάνυσμα των λόγων των ποσοστημορίων της i έρευνας. Το S_i είναι ο πίνακας συνδιακύμανσης εσωτερικά της έρευνας i. Το

πραγματικό διάνυσμα των $logq_i$ ακολουθεί την πολυδιάστατη κανονική

$$logq_i \sim N(logq, \Sigma), \tag{4.48}$$

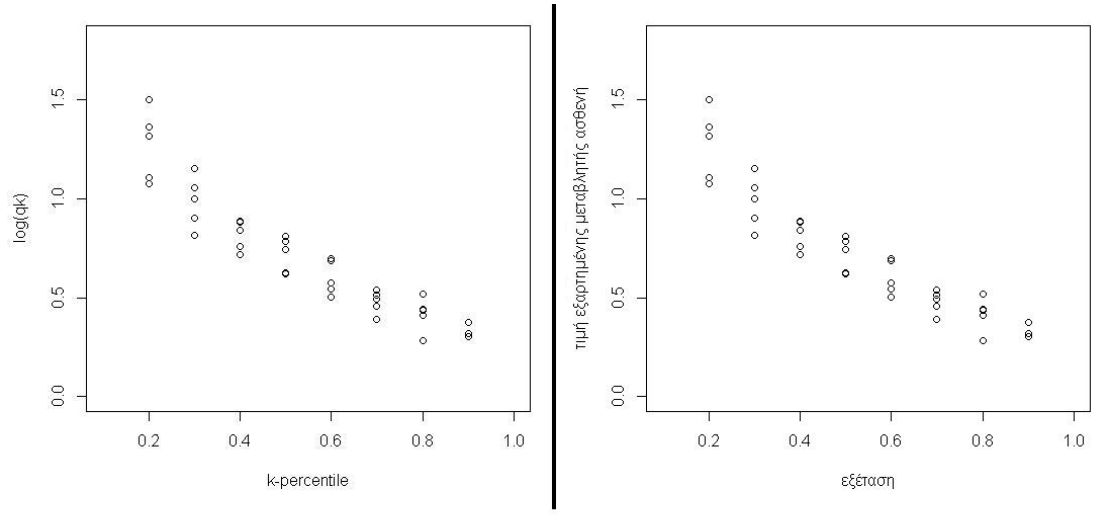
όπου Σ είναι ο πίνακας συνδιακύμανσης μεταξύ των ερευνών. Αφού ορίσαμε τις αντίστοιχες έννοιες που χρειάζονται για την πολυδιάστατη μετα-ανάλυση, όπως περιγράφθηκε, με χρήση των ανάλογων διαδικασιών μπορούμε να παράγουμε τα αποτελέσματα που επιθυμούμε.

4.4 Μετα-ανάλυση του λόγου των ποσοστημορίων με χρήση μοντέλων και τεχνικών ανάλυσης διαχρονικών δεδομένων

Στην παράγραφο αυτή η οποία αποτελεί και τον κύριο σκοπό της παρούσης εργασίας, θα περιγράψουμε μια μέθοδο μετα-ανάλυσης με μεγέθη επίδρασης τον λόγο των ποσοστημορίων, η οποία χρησιμοποιεί τις μεθόδους και τις τεχνικές της ανάλυσης διαχρονικών δεδομένων (longitudinal data) για τα οποία μιλήσαμε στο κεφάλαιο 2, για να δώσει εκτιμήσεις για το summary effect μέσω ενός μοντέλου γραμμικής σχέσης των μεταβλητών του, το οποίο μάλιστα θα μας δίνει τιμές όχι απλά για μια τιμή του k, αλλά για όσες είναι επιθυμητό.

Για να θυμήσουμε το σκεπτικό της ανάλυσης διαχρονικών δεδομένων και να μπορέσουμε να κάνουμε τον κατάλληλο παραλληλισμό, θα περιγράψουμε την πιο απλή μορφή της δομή της με ένα παράδειγμα. Έστω ότι σε ένα νοσοκομείο N ασθενείς κάθε 15 ημέρες εξετάζονται για να μετρηθεί η αύξηση ή μείωση της χοληστερίνης, αυτό πραγματοποιείται για ένα διάστημα 6 μηνών, δηλαδή συνολικά έχουμε 12 εξετάσεις σε κάθε ασθενή, οπότε και συλλέγουμε 12 τιμές από καθέναν από τους N ασθενείς. Ας θυμήσουμε πάλι, ότι ο λόγος των ποσοστημορίων είναι μια τιμή η οποία είναι συνάρτηση του $k \in [0.1, 0.9]$ και πρακτικά μας δίνει τιμές για κάποιες χρονικές στιγμές που υφίστανται σε όλες τις έρευνες. Άρα θεωρώντας τις $k = 0.1, \cdots, 0.9$ σαν τις χρονικές στιγμές που γίνονται οι εξετάσεις για να μας δωθεί η τιμή του q_k (ή του $logq_k$), έχουμε πρακτικά να κάνουμε με μια ανάλυση διαχρονικών δεδομένων, όπου μάλιστα οι μεταξύ τιμές των q_k σε κάθε έρευνα (ασθενή) από τις N είναι συ-

σχετισμένες η μια με την άλλη. Συμπεριλαμβάνοντας αυτές σε μια έρευνα μπορούμε να βρούμε ένα γραμμικό μοντέλο (όπως ορίστηκε στο κεφάλαιο 2 για τα διαχρονικά δεδομένα) με εξαρτημένη μεταβλητή το q_k και περιγραφική το k, από το μοντέλο αυτό θα μπορούμε να βγάλουμε συμπεράσματα για κάθε τιμή του k, για το summary effect δηλαδή το q_k.



Σχήμα 4.4: Παραλληλισμός των $logq_k$ ανά ποσοστημόριο με το παράδειγμα τιμών εξαρτημένης μεταβλητής κάποιας μονάδας μέτρησης ανά εξέταση.

4.4.1 Μοντελοποίηση με βάση τις GEE

Στην ενότητα αυτή θα περιγράψουμε πως θα δομηθεί το συγκεκριμένο μοντέλο και θα αναλύσουμε τις διάφορες διαδικασίες που θα χρησιμοποιήσουμε. Ας υποθέσουμε πως έχουμε N έρευνες οι οποίες αναφέρονται στην σύγκριση δύο ομάδων ως προς το χρόνο επιβίωσης ή γενικότερα κάποιου καταληκτικού γεγονότος και ας θεωρήσουμε πως είναι της μορφής νέα θεραπεία - συνήθης θεραπεία. Για καθεμία από τις N αυτές έρευνες μας έχουν δωθεί τα πλήρη δεδομένα τα οποία μας αναφέρουν τους χρόνους καταληκτικού γεγονότος του κάθε ασθενή, σε ποια ομάδα θεραπείας ανήκει, αν ο χρόνος είναι λογοκριμένος ή όχι και ενδεχομένως κάποιες άλλες μεταβλητές. Δηλαδή τα δεδομένα μας είναι IPD. Στην πρώτη φάση θα επεξεργαστούμε τα δεδομένα

για κάθε έρευνα ξεχωριστά χρησιμοποιώντας τις ανάλογες διαδικασίες εκτίμησης που μιλήσαμε στην ανάλυση επιβίωσης και στο παρών κεφάλαιο και με χρήση της σχέσης (4.19) θα παράγουμε τις εκτιμήσεις για τα $logq_k$ για $k = 0.1, \ldots, 0.9$. Στην συνέχεια με την μέθοδο bootstrap θα υπολογίσουμε τις εκτιμήσεις των πινάκων συνδιακύμανσης αυτών για κάθε έρευνα. Εδώ να αναφέρουμε ότι ανάλογα με την φύση των δεδομένων μπορούν οι εκτιμήσεις μας να είναι σε μικρότερο εύρος τιμών πχ στο $k = 0.2, \ldots, 0.9$. Έτσι έχουμε τα παρακάτω στοιχεία, τα οποία τα παρουσιάζουμε σε μορφή αντίστοιχης με την γενική μορφή των διαχρονικών δεδομένων (όπως στο Σχήμα 2.1 του κεφαλαίου 2)

$Study(ID)$	Y	k
1	$log\hat{q}_{1,0.1}$	0.1
1	$log\hat{q}_{1,0.2}$	0.2
.	.	.
.	.	.
1	$log\hat{q}_{1,0.9}$	0.9
2	$log\hat{q}_{2,0.1}$	0.1
.	.	.
.	.	.
2	$log\hat{q}_{2,0.9}$	0.9
.	.	.
.	.	.
N	$log\hat{q}_{N,0.1}$	0.1
.	.	.
.	.	.
N	$log\hat{q}_{N,0.9}$	0.9

όπως επίσης έχουμε εκτιμήσει και τους N πίνακες συνδιακύμανσης για κάθε έρευνα

$$Cov_i = \begin{pmatrix} V(log\hat{q}_{i0.1}) & Cov_{i,0.1,0.2} & . & . & . & Cov_{i,0.1,0.9} \\ Cov_{i,0.1,0.2} & V(log\hat{q}_{i0.2}) & . & . & . & Cov_{i,0.2,0.9} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ Cov_{i,0.1,0.9} & . & . & . & . & V(log\hat{q}_{i0.9}) \end{pmatrix}, \qquad (4.49)$$

όπου $Cov_{i,k_l,k_r} = \hat{Cov}(log\hat{q}_{i,k_l}, log\hat{q}_{i,k_r})$ οι εκτιμήτριες της συνδιασποράς μεταξύ των διαφορετικών τιμών των λόγων των ποσοστημορίων για τις διάφορες τιμές του k $(k_r, k_l = 0.1, \cdots 0.9, k_r \neq k_l)$, μέσω της μεθόδου bootstrap.

Αυτό που στην ουσία θέλουμε να κάνουμε στην συνέχεια είναι μέσω αυτών των δεδομένων να δημιουργήσουμε ένα γραμμικό μοντέλο, όπως το ορίσαμε στο κεφάλαιο 2 για τα διαχρονικά δεδομένα. Σε μια απλουστευμένη μορφή για να γίνουμε πιο κατανοητοί θέλουμε ένα μοντέλο της μορφής $logq_k = \alpha + \beta \cdot k$. Πρακτικά αυτό που θέλουμε να βρούμε είναι την εκτίμηση του πως αλλάζει ο μέσος λόγος των ποσοστημορίων ανά ποσοστημόριο, ώστε να έχουμε μια γραμμική σχέση η οποία για την συνεχή μεταβλητή $k \in [0.1, 0.9]$ θα μπορεί να μας δώσει το ζητούμενο. Για να το πετύχουμε αυτό θα χρησιμοποιήσουμε τις γενικευμένες εξισώσεις εκτίμησης GEE τις οποίες αναφέραμε στην παράγραφο 2.4.2, έτσι ώστε να εκτιμήσουμε το μέσο του $logq_k$ για κάθε k από το σύνολο (πληθυσμό) των N ερευνών. Όπως έχουμε ήδη αναφέρει στην μοντελοποίηση με βάση τις GEE δεν υποθέτουμε την κατανομή της εξαρτημένης μεταβλητής, αλλά καθορίζουμε την κατανομή του μέσου. Στην περίπτωση μας θα υποθέσουμε ότι ο μέσος της εξαρτημένης μεταβλητής δηλαδή του λογάριθμου του λόγου ποσοστημορίων ακολουθεί κανονική κατανομή και για την συνάρτηση σύνδεσμο θα ισχύει ότι $g(.) = \eta$ και ότι $u(\mu) = 1$, άρα ο μέσος θα είναι της μορφής

$$E[Y_{ij}|k] = E[logq_k] = \alpha + \beta \cdot k, \;\; k \in [0.1, 0.9]$$

Το κυριότερο πρόβλημα που έχουμε να αντιμετωπίσουμε είναι ότι το μικρό μέγεθος του δείγματος των ερευνών N, έρχεται σε ρίξη με τις GEE διαδικασίες, οι οποίες για την καλύτερη επιτυχία των εκτιμήσεών τους και ειδικότερα όσον αφορά την διασπορά προϋποθέτουν μεγάλο δείγμα ατόμων και μάλιστα πρέπει και να είναι μεγαλύτερο από των αριθμό των παρατηρήσεων (εξετάσεων ασθενή) για κάθε άτομο. Παρακάτω θα

δείξουμε πως επεμβαίνοντας στις GEE διαδικασίες, θα προσπαθήσουμε να αποφύγουμε αυτό το μειονέκτημα, αλλά και άλλα που εμφανίστηκαν κατά τις δοκιμές.

Όπως αναφέραμε στο κεφάλαιο για τα διαχρονικά δεδομένα οι εκτιμήσεις των παραμέτρων στις GEE διαδικασίες δίνονται μέσα από την σχέση

$$U(\beta) = \sum_{i=1}^{N} \left(\frac{d\mu_i}{d\beta}\right)' [V_i(\hat{a})]^{-1}(y_i - \mu_i) = 0_2,$$

όπου στην συγκεκριμένη περίπτωση έχουμε ότι

$$\left(\frac{d\mu_i}{d\beta}\right) = \begin{pmatrix} 1 & X_{0.1} \\ 1 & X_{0.2} \\ . & . \\ . & . \\ . & . \\ 1 & X_{0.9} \end{pmatrix},$$

όπου X_k είναι το αντίστοιχο ποσοστημόριο, $X_k \in [0.1, 0.9]$, και

$$V_i(a) = \phi\sqrt{A_i}R_i(a)\sqrt{A_i},$$

ο πίνακας συνδιακύμανσης της $i-$οστής έρευνας. Πρακτικά δηλαδή για την εκτίμηση του πίνακα συνδιακύμανσης χρησιμοποιούμε μια δομή συσχέτισης $R_i(\hat{a})$ για όλες τις έρευνες η οποία πολλαπλασιάζεται με την παράμετρο ϕ η οποία είναι προς εκτίμηση και τον γνωστό μοναδιαίο πίνακα $\sqrt{A_i}$. Αντί να χρησιμοποιήσουμε όμως μια δομή συσχέτισης η οποία ενδέχεται να είναι λανθασμένη ή η εκτίμησή της ενδέχεται λόγω του μικρού δείγματος N να μην είναι αρκετά καλή, ενώ το ίδιο ισχύει και για το ϕ, θα χρησιμοποιήσουμε για κάθε έρευνα τον πίνακα συνδιασποράς της σχέσης (4.49) τον οποίο έχουμε εκτιμήσει μέσω bootstrap. Αυτό μας δίνει την σχέση

$$U^*(\beta) = \sum_{i=1}^{N} \left(\frac{d\mu_i}{d\beta}\right)' [Cov_i]^{-1}(y_i - \mu_i) = 0_2, \tag{4.50}$$

η οποία έχει δύο αγνώστους προς εκτίμηση.
Με παρόμοια λογική θα αντικαταστήσουμε σε όλες τις σχετικές σχέσεις την εκτίμη-

ση του πίνακα συνδιακύμανσης μέσω κατάλληλης δομής πίνακας συσχέτισης με τον αντίστοιχο πίνακα συνδιακύμανσης μέσω bootstrap, δηλαδή θα έχουμε τις σχέσεις

$$B^* = \sum_{i=1}^{N} \left(\frac{d\hat{\mu}_i}{d\beta}\right)' [Cov_i]^{-1} \left(\frac{d\hat{\mu}_i}{d\beta}\right), \tag{4.51}$$

$$M^* = \sum_{i=1}^{N} \left(\frac{d\hat{\mu}_i}{d\beta}\right)' [Cov_i]^{-1}(y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)'[Cov_i]^{-1} \left(\frac{d\hat{\mu}_i}{d\beta}\right), \tag{4.52}$$

από όπου μπορούμε να βρούμε τον αντίστοιχο εκτιμητή σάντουιτς, μέσω της

$$V_S^* = (B^*)^{-1} M^* (B^*)^{-1}. \tag{4.53}$$

Όπως αναφέραμε στην συγκεκριμένη ενότητα του κεφαλαίου 2 ο εκτιμητής σάντουϊτς δεν είναι τόσο κατάλληλος για μικρό δείγμα ατόμων, δηλαδή στην περίπτωσή μας για το δείγμα των N ερευνών. Ο Mancl και ο DeRouen [18], προτείνουν ένα διορθωμένο εκτιμητή σάντουιτς, όπου στην σχέση υπολογισμού του αντικαθιστούμε το $S_i S_i' = (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)'$ με το $(I - H_i)^{-1} S_i S_i' (I - H_i')^{-1}$, όπου I ο μοναδιαίος πίνακας και $H_i = \left(\frac{d\hat{\mu}_i}{d\beta}\right)(B)^{-1}\left(\frac{d\hat{\mu}_i}{d\beta}\right)' V_i^{-1}$. Το κίνητρο για αυτή τη διόρθωση είναι ότι το $S_i S_i'$ είναι μια μεροληπτική εκτίμηση του $Cov(Y_i)$ και ότι $E(S_i S_i') \approx (I - H_i) Cov(Y_i)(I - H_i')$. Έτσι καταλήγουμε στον εκτιμητή σάντουιτς με διόρθωση μεροληψίας.
Στην περίπτωση που εξετάζουμε έχουμε αρχικά ότι

$$H_i = \left(\frac{d\hat{\mu}_i}{d\beta}\right)(B^*)^{-1}\left(\frac{d\hat{\mu}_i}{d\beta}\right)' Cov_i^{-1} \tag{4.54}$$

και άρα ο εκτιμητής σάντουιτς με διόρθωση μεροληψίας δίνεται από την σχέση

$$V_{Sbc}^* = (B^*)^{-1} \sum_{i=1}^{N} \left[\left(\frac{d\hat{\mu}_i}{d\beta}\right)' [Cov_i]^{-1}(I - H_i)^{-1} S_i S_i' (I - H_i')^{-1} [Cov_i]^{-1} \left(\frac{d\hat{\mu}_i}{d\beta}\right)\right] (B^*)^{-1}. \tag{4.55}$$

Εδώ να αναφέρουμε ότι ενδέχεται το μοντέλο με την γραμμική σχέση $E[logq_k] = \alpha + \beta \cdot k$ να μην προσαρμόζει τόσο καλά για όλες τις τιμές του k όσο μια σχέση

με το ποσοστημορίο να είναι στην τετραγωνική του $E[logq_k] = \alpha + \beta \cdot k^2$ ή την κυβική του μορφή $E[logq_k] = \alpha + \beta \cdot k^3$ ή ακόμα και σε μια ευρύτερη πολυωνυμική μορφή, πχ $E[logq_k] = \alpha + \beta_1 \cdot k + \beta_2 k^2$, όπως ενδεχομένως για την αντιμετώπιση της μη σωστής προσαρμογής μπορούμε ενδεχομένως να εφαρμόσουμε άλλες τακτικές προσαρμογής ενός μοντέλου. Αυτό τίθεται προς έρευνα και μελέτη στα επόμενα βήματα που πρέπει να γίνουν πάνω στη δομή του μοντέλου, ώστε να καταλήξουμε στην κατάλληλη μορφή.

4.4.2 Προσομοίωση

Σε αυτή την ενότητα θα παρουσιάσουμε τα αποτελέσματα της προσομοίωσης που χρησιμοποιήσαμε ώστε να ελεγχθεί η απόδοση του μοντέλου που περιγράψαμε στην εκτίμηση του λόγου των ποσοστημορίων της διαμέσου $q_{0.5}$. Η δομή που ακολουθήσαμε για την προσομοίωση ήταν η εξής: γεννήσαμε δεδομένα από N (9-15) ανεξάρτητες έρευνες οι οποίες είναι βασισμένες στο μοντέλο του αναλογικού μοντέλου με baseline συνάρτηση να ακολουθεί την log-logistic κατανομή, η καθεμία από αυτές περιείχε 240 ασθενείς, 120 σε κάθε ομάδα. Το ποσοστό λογοκρισίας ήταν τυχαίο και ακολουθεί εκθετική κατανομή. Υποθέσαμε ότι ο σταθερός λόγος των ποσοστημορίων για την διάμεσο είναι 2, ενώ όλες οι άλλες παραμέτροι επιτρέψαμε να παίρνουν διάφορες τιμές. Οι τιμές που δώθηκαν στην πλειονότητα των περιπτώσεων στις παραμέτρους ήταν τέτοιες ώστε να αντικατοπτρίζουν λογικούς χρόνους επιβίωσης και επιπλέον λογικές διαφορές σε αυτούς μεταξύ των ερευνών. Παρακάτω θα αναφέρουμε μερικές από τις προσομοιώσεις που πραγματοποιήθηκαν για διάφορες τιμές των μεταβλητών, αλλά και για διάφορες τιμές λογοκρισίας.

Προσομοίωση 1

Παράμετροι :

Αριθμός ερευνών: 15

$r = (7, 7, 7, 7, 8, 6, 8, 8, 6, 6, 6, 7, 6, 6, 8)$

$\lambda^{-1} = (1130, 1335, 1310, 1260, 1236, 1268, 1265, 1353, 1255, 1119, 1330, 1191, 1389, 1358, 1186)$

censoring	*mean_est*	*coverage_prop*
0%	2.004146	0.9038
17%	2.029069	0.94
25%	2.012192	0.94

για την πιθανότητα σύγκλισης με βάση την εκτίμηση για την διασπορά μέσω των σχέσεων (4.51) και (4.53), δηλαδή με τον αντίστροφο του πίνακα B^* και με την εκτιμήτρια σάντουιτς χωρίς την διόρθωση έχουμε αντίστοιχα $(0\% : 0.8653846, 0.9038462)$, $(17\% : 0.77, 0.9)$ και $(25\% : 0.85, 0.92)$.

Προσομοίωση 2

Αριθμός ερευνών: 15

Παράμετροι :

$r = (1.441282, 0.8025609, 1.1171802, 1.9418977, 0.8925101, 1.9741414, 1.503814, 0.8950533, 1.6500516, 0.8623015, 1.4636805, 1.1998630, 1.9308729, 1.0117925, 1.9379657)$

$\lambda^{-1} = (1330, 1326, 1349, 1253, 1143, 1248, 1392, 1257, 1361, 1181, 1144, 1174, 1174, 1392, 1173)$

censoring	*mean_est*	*coverage_prop*
0%	2.021626	0.9375
17%	2.053227	0.90
22%	2.066189	0.9333333

αντίστοιχα όπως και πριν για τον αντίστροφο του πίνακα B^* και για την εκτιμήτρια σάντουιτς χωρίς την διόρθωση έχουμε $(0\% : 0.8541667, 0.875)$, $(17\% : 0.79, 0.86)$ και $(25\% : 0.8333333, 0.9)$.

Προσομοίωση 3

Αριθμός ερευνών: 13

Παράμετροι :

$r = (6, 8, 6, 8, 6, 7, 7, 7, 7, 8, 8, 7, 7)$

$\lambda^{-1} = (1369, 1164, 1267, 1109, 1397, 1306, 1203, 1368, 1238, 1208, 1162, 1179, 1249)$

censoring	*mean_est*	*coverage_prop*
0%	2.017833	0.94
17%	2.02707	0.96
25%	2.010423	0.98

σε αντιστοιχία με πριν έχουμε (0% : 0.88, 0.94), (17% : 0.78, 0.90) και (25%: 0.78, 0.92).

Προσομοίωση 4

Αριθμός ερευνών: 11

Παράμετροι :

$r = (6.545677, 7.962407, 6.952505, 7.541520, 7.3712217.698785,$
$6.688786, 7.215745, 6.444858, 6.071900, 7.179287)$

$\lambda^{-1} = (1313, 1258, 1193, 1207, 1164, 1363, 1149, 1336, 1289, 1118, 1268)$

censoring	*mean_est*	*coverage_prop*
0%	2.011396	0.93
16%	2.028643	0.9333333
24%	2.033123	0.94

κι εδώ έχουμε (0% : 0.78, 0.91), (16% : 0.8, 0.9333333) και (24% : 0.85, 0.91).

Προσομοίωση 5

Αριθμός ερευνών: 9

Παράμετροι :

$r = (7.359208, 5.294005, 7.277626, 5.933948, 6.812108, 6.126058, 5.495147, 6.907652, 5.504791)$

$\lambda^{-1} = (1379, 1299, 1253, 1365, 1141, 1168, 1240, 1389, 1165)$

censoring	*mean_est*	*coverage_prop*
0%	2.012607	0.92
16%	2.02758	0.93
24%	2.034371	0.91

κι εδώ έχουμε (0% : 0.81, 0.9), (16% : 0.77, 0.86) και (24% : 0.86, 0.82).

Από τα παραπάνω παρατηρούμε μέσω της μέσης τιμής των επαναλήψεων της σημειακής εκτίμησης του λόγου των ποσοστημορίων της διαμέσου, ότι είναι εμφανές ότι είναι πάρα πολύ κοντά στη ζητούμενη τιμή 2. Τα αποτελέσματα για την πιθανότητα σύγκλισης (coverage probability) υπολογίστηκαν σαν το ποσοστό των επαναλήψεων για τα οποία μέσα στο διάστημα εμπιστοσύνης 95% περιεχόταν η τιμή 2. Μπορούμε να παρατηρήσουμε ότι ο διορθωμένος εκτιμητής sandwich της συνδιασποράς των μεταβλητών, σε σύγκριση με τα αντίστοιχα coverage propability για την naive και την biased sandwich εκτίμηση, πετυχαίνει πολύ πιο *σωστές* τιμές, γύρω από το ζητούμενο ποσοστό του 95%. Γενικά ανάλογα και με τις τιμές που δώσαμε στις διάφορες παραμέτρους το coverage propability κυμαινόταν μεταξύ των τιμών 88% με 97%. Όπως αναφέρθηκε και προηγουμένως μεγάλη προσοχή πρέπει να δωθεί στο εύρος των ποσοστημορίων που θα χρησιμοποιηθεί καθότι ακραίες τιμές αυτών (τέλος της έρευνας) είτε μπορεί να μην ορίζονται, είτε λόγω της υπάρξεως της λογοκρισίας ενδέχεται να παρέχουν λάθος εκτίμηση. Στις συγκεκριμένες προσομοιώσεις χρησιμοποιήθηκαν οι τιμές που μας δίνονται από τα ποσοστημόρια για 0.2 εώς 0.9.

4.4.3 Μοντελοποίηση με χρήση των GLMM

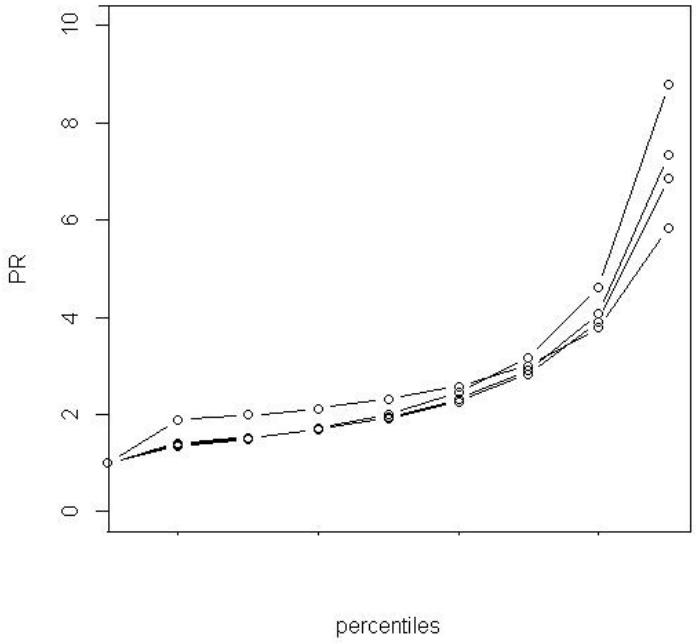
Σε αυτό το κεφάλαιο θα περιγράψουμε την μοντελοποίηση των δεδομένων μας, με χρήση της έτερης διαδικασίας της ανάλυσης διαχρονικών δεδομένων, αυτής των Γενικευμένων Γραμμικών Μοντέλων Μικτών Επιδράσεων.

Ας υποθέσουμε όπως πριν ότι έχουμε N έρευνες οι οποίες αναφέρονται στην σύγκριση δύο ομάδων με κάποιο καταληκτικό γεγονός, όλες οι έρευνες έχουν μια χρονική στιγμή εκκίνησης η οποία στην πράξη είναι η μέρα που ξεκινάει η εργαστηριακή έρευνα, εκείνη την χρονική στιγμή το PR είναι ίδιο για όλες τις N έρευνες, οπότε τοποθετώντας στους άξονες τις τιμές του PR για κάθε ποσοστημόριο και άμα υποθέσουμε πως εσωτερικά της κάθε έρευνας ισχύει μια μορφή γραμμικής σχέσης, παρατηρούμε πως από έρευνα σε έρευνα η διαφορά βρίσκεται στην κλίση της γραμμικής σχέσης. Αυτό με βάση τα GLMM σημαίνει ότι υπάρχει από έρευνα σε έρευνα μια τυχαία επίδραση $b_i \sim N(0, \sigma_i^2)$ η οποία επηρεάζει την κλίση της κάθε έρευνας ως προς τον λόγο των ποσοστημορίων, άρα το μοντέλο θα είναι της μορφής

$Y = \alpha + (b + \beta)X + \epsilon$, δηλαδή

$$q_{ik} = \alpha + (b_i + \beta)k + \epsilon_{ik},\ k \in [0.1, 0.9]$$

Τα δεδομένα θα εκτιμηθούν όπως στην προηγούμενη μοντελοποίηση απαραμετρικά και στην λογαριθμική κλίμακα, όπως και θα δομηθούν κατά τον ίδιο τρόπο ανάλογο με αυτόν που χρησιμοποιούμε στην ανάλυση διαχρονικών δεδομένων. Σε σύγκριση με το προηγούμενο μοντέλο που έγινε με βάση τις GEE διαδικασίες, εδώ δεν χρειάζεται η υπόθεση δομής πίνακα συνδιασποράς οπότε το βήμα όπου υπολογίζεται αυτό με χρήση bootstrap διαδικασιών δεν μας είναι χρήσιμο. Θα πρέπει να ελεγχθεί και αν κάποιο άλλο είδος γραμμικής σχέσης είναι πιο κατάλληλο (τετραγωνική, κυβική μορφή του k κλπ) ή αν πρέπει να γίνουν άλλες διαδικασίες ώστε να καταλήξουμε στο πιο σωστό μοντέλο.



Σχήμα 4.5: Εικονική περιγραφή μοντέλου. Θεωρητικά υποθέτουμε ότι όλες οι έρευνες ξεκινάνε από μια κοινή αρχή

4.4.4 Συζήτηση

Όπως αναφέραμε ο κύριος στόχος της παρούσης εργασίας ήταν αφού αρχικά υπενθυμίσουμε την κύρια θεωρία πίσω από τους σχετικούς τομείς της στατιστικής, να συζητήσουμε το κατά πόσο μπορούν οι τακτικές που χρησιμοποιούνται στην ανάλυση διαχρονικών δεδομένων να εισαχθούν σε μια μετα-ανάλυση η οποία θα χρησιμοποιεί για μέγεθος επίδρασης το λόγο των ποσοστημορίων, καθότι ο λόγος των ποσοστημορίων εσωτερικά μιας έρευνας είναι συσχετισμένος από ποσοστημόριο σε ποσοστημόριο και αυτό συμβαδίζει με το ζητούμενο της ανάλυσης διαχρονικών δεδομένων όπου οι τιμές της εξαρτημένης μεταβλητής είναι συσχετισμένες στο κάθε άτομο από εξέταση σε εξέταση. Με βάση αυτή τη μοντελοποίηση μιας μετα-ανάλυσης γίνεται προσπάθεια να παραχθεί μια *γραμμική σχέση* η οποία θα μας δίνει συμπεράσματα όχι μόνο για μια τιμή του λόγου των ποσοστημορίων αλλά για ένα σύνολο τιμών αυτού, το οποίο μάλιστα θα έχει αυξήσει την στατιστική του ακρίβεια αφού θα περιέχει και την συσχέτιση μεταξύ των τιμών των ποσοστημορίων σαν παραπάνω πληροφορία.

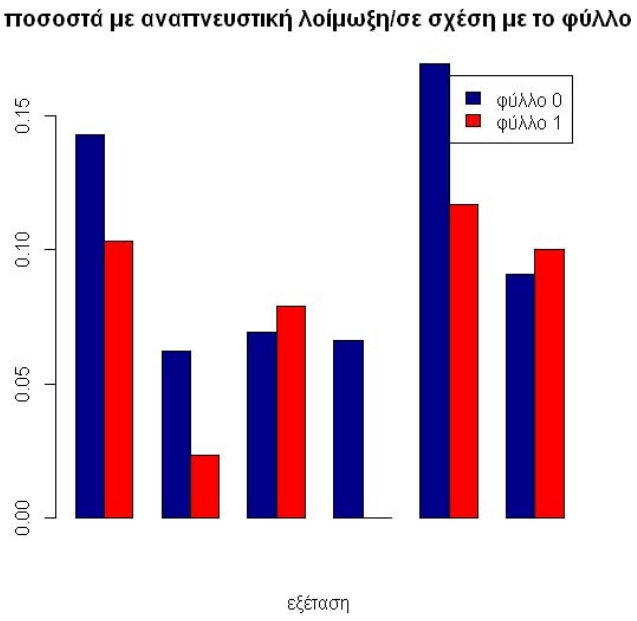
Όπως είδαμε στις παραπάνω προσομοιώσεις το μοντέλο αυτό φαίνεται να επιτυγχάνει το ζητούμενο αποτέλεσμα, κάτω από τους περιορισμούς που του θέσαμε. Παρατηρούμε ότι μπορεί να χρησιμοποιηθεί και σαν τρόπος μετα-ανάλυσης μιας μόνο τιμής του λόγου ποσοστημορίων στα GEE μοντέλα και ότι η σημειακή εκτίμηση είναι πολύ κοντά στην πραγματική τιμή, ενώ το διαστήμα εμπιστοσύνης που παράγει είναι σχετικά μικρό. Οι τιμές εκτός του λόγου των ποσοστημορίων για τη διάμεσο ($q_{0.5}$) αφού θα χρησιμοποιούν τα ίδια δεδομένα για το ίδιο εύρος των ποσοστημορίων, θα δίνουν τις ίδιες εκτιμήσεις για τις παραμέτρους του γραμμικού μοντέλου, παρόλα αυτά πρέπει να μελετηθεί η ευαισθησία σε αυτές και το κατά πόσο επηρεάζονται τελικά, ειδικότερα στις ακραίες τιμές εσωτερικά μιας έρευνας, δηλαδή για τα ποσοστημόρια προς το τέλος μιας μελέτης τα οποία στις δοκιμές με βάση αυτή τη μοντελοποίηση φαίνεται από ένα ποσοστημόριο και μετά ότι ενδεχομένως αποκλίνουν από την πραγματική τους τιμή. Αφότου φανεί κάτι τέτοιο πρέπει να προταθούν τρόποι αντιμετώπισης αυτού του φαινομένου.

Σε ένα επόμενο βήμα θα πρέπει να γενικεύσουμε τα αποτελέσματα και σε ένα μεγαλύτερο σύνολο κατανομών και σε συνδυασμούς αυτών. Ακόμα χρησιμοποιώντας τη γνώση μας από τις τακτικές της μετα-ανάλυσης μπορεί να ελεχθεί κατά πόσο και πως μπορούν να σταθμιστούν οι έρευνες ώστε να μην έχουν όλες την ίδια βαρύτητα.

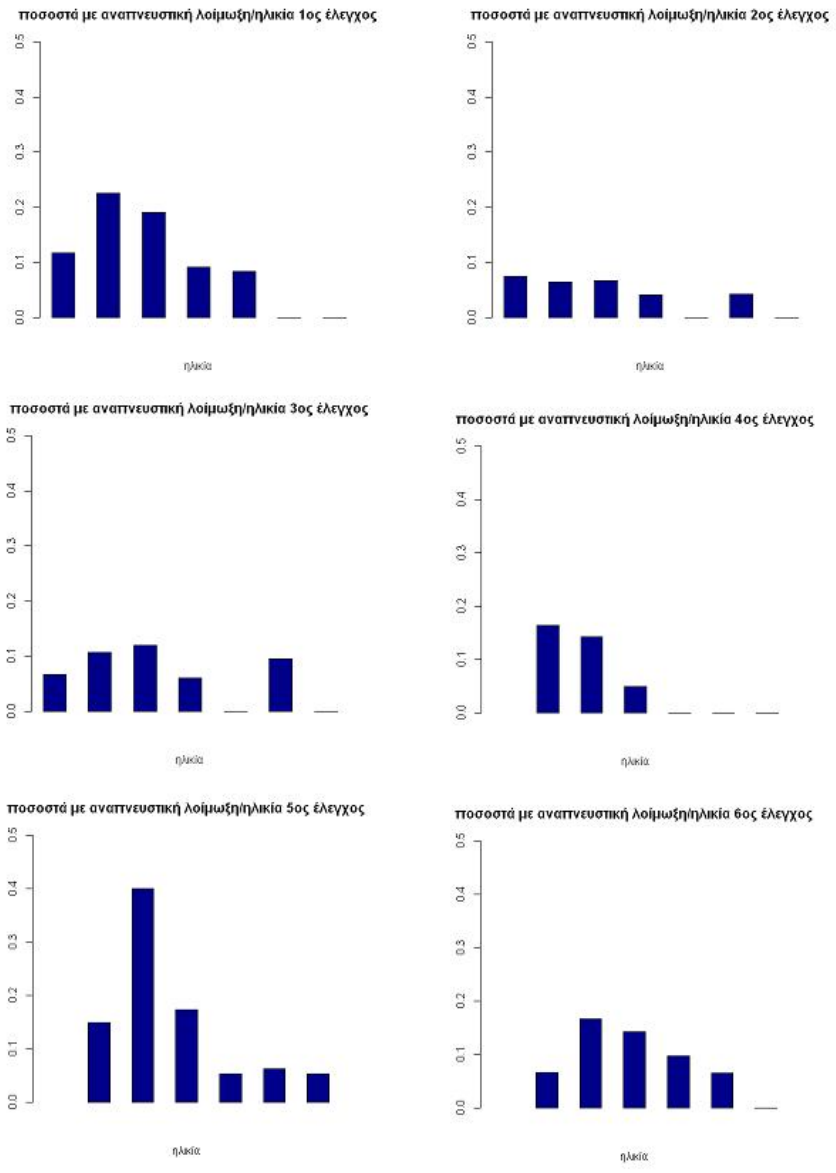
Η ετερογένεια μεταξύ των μελετών της μετα-ανάλυσης πρέπει να τεθεί και αυτή προς μελέτη και ενδεχομένως να βρεθούν τρόποι για την εκτίμηση αυτής.

Παράρτημα Α΄

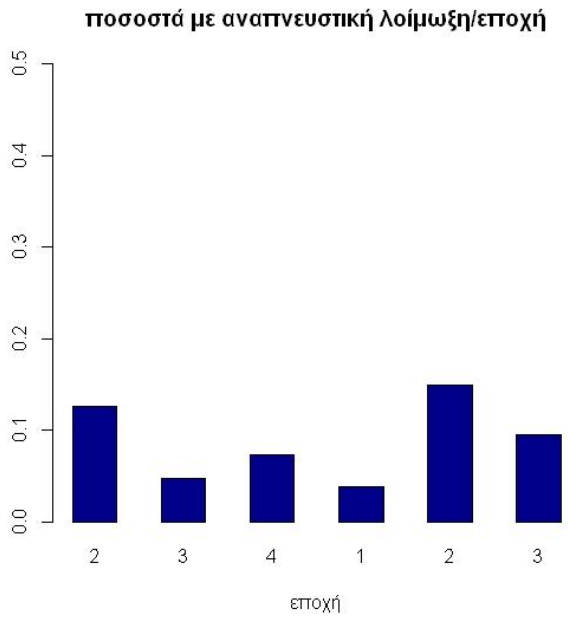
Λοιπά γραφήματα

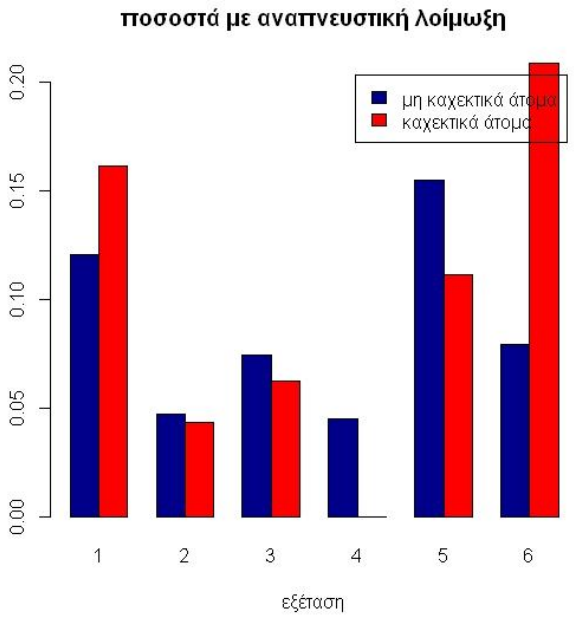


Σχήμα Α΄.1:

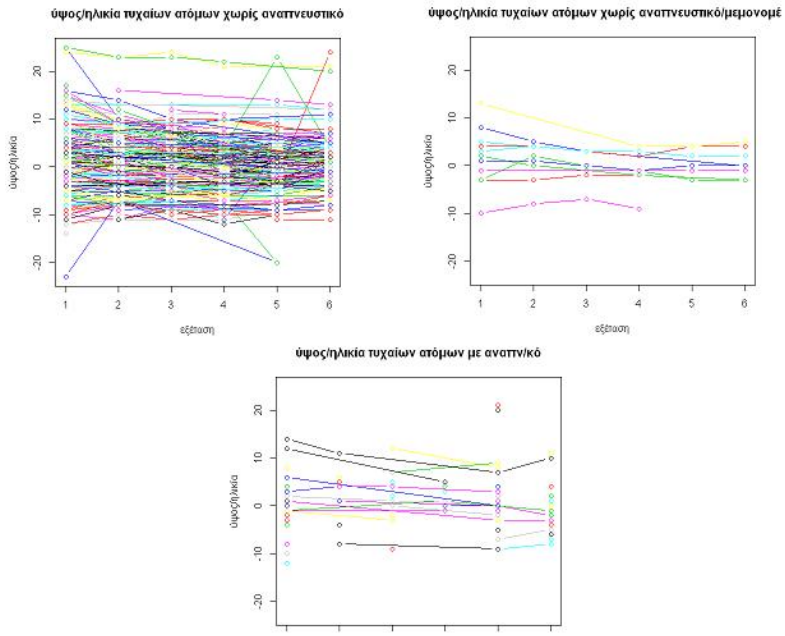

ποσοστά με αναπνευστική λοίμωξη/ηλικία 1ος έλεγχος
ποσοστά με αναπνευστική λοίμωξη/ηλικία 2ος έλεγχος
ποσοστά με αναπνευστική λοίμωξη/ηλικία 3ος έλεγχος
ποσοστά με αναπνευστική λοίμωξη/ηλικία 4ος έλεγχος
ποσοστά με αναπνευστική λοίμωξη/ηλικία 5ος έλεγχος
ποσοστά με αναπνευστική λοίμωξη/ηλικία 6ος έλεγχος
ηλικία
0.0
0.1
0.2
0.3
0.4
0.5

Σχήμα Α΄.2:

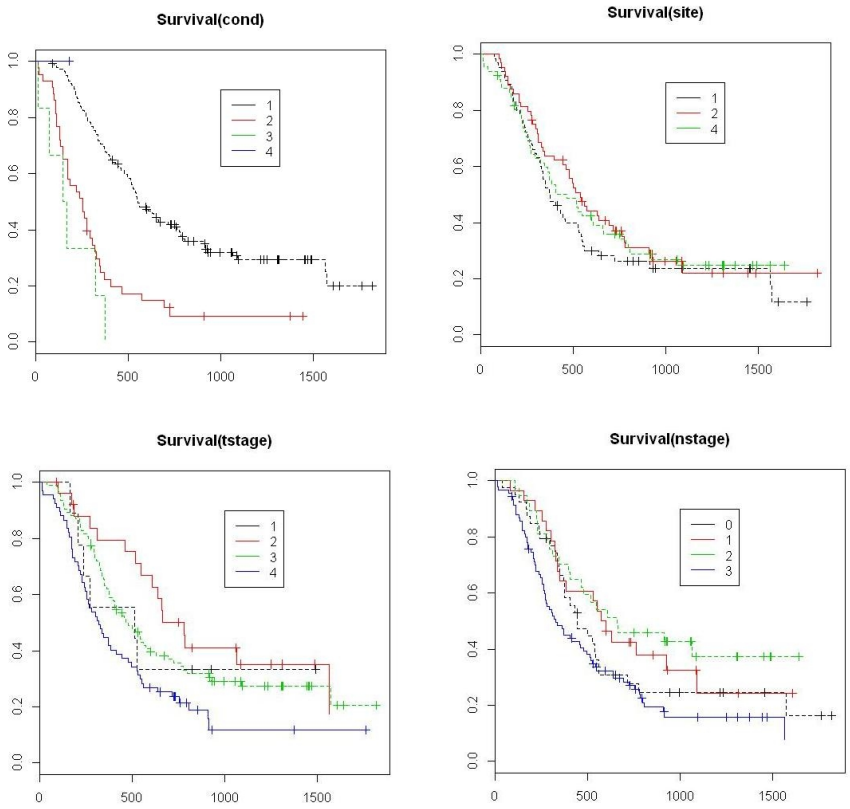

ποσοστά με αναπνευστική λοίμωξη/εποχή
0.0
0.1
0.2
0.3
0.4
0.5
2
3
4
1
2
3
εποχή


ποσοστά με αναπνευστική λοίμωξη
μη καχεκτικά άτομα
καχεκτικά άτομα
0.00
0.05
0.10
0.15
0.20
1
2
3
4
5
6
εξέταση

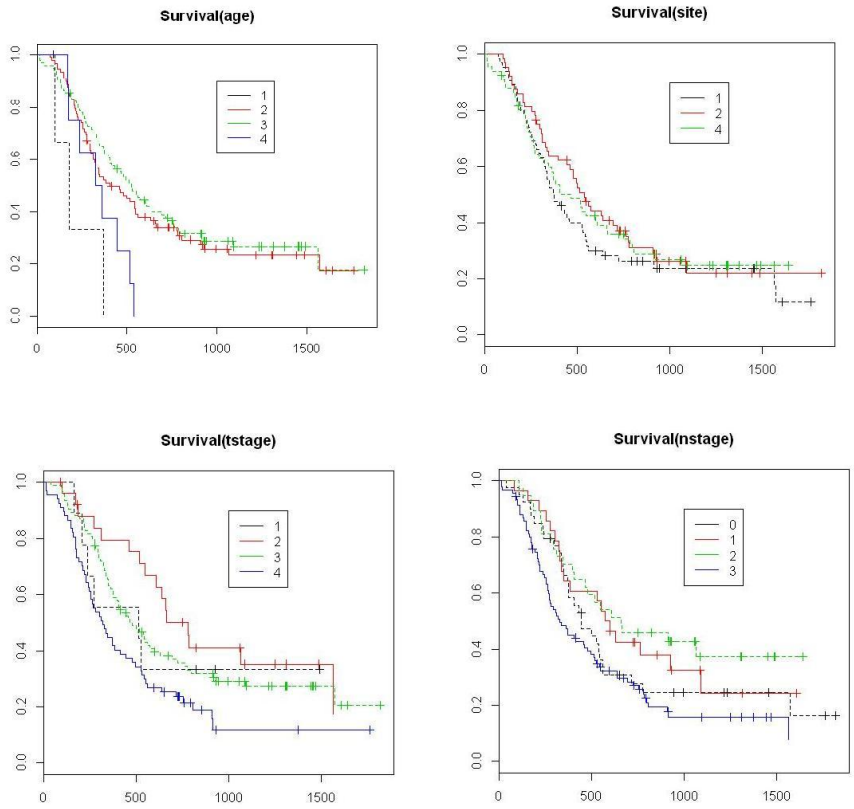
Σχήμα Α′.3:


ύψος/ηλικία τυχαίων ατόμων χωρίς αναπνευστικό
ύψος/ηλικία
εξέταση
ύψος/ηλικία τυχαίων ατόμων χωρίς αναπνευστικό/μεμονομέ
ύψος/ηλικία
εξέταση
ύψος/ηλικία τυχαίων ατόμων με αναπν/κό
ύψος/ηλικία

Σχήμα Α′.4:


Survival(cond)
1
2
3
4
Survival(site)
1
2
4
Survival(tstage)
1
2
3
4
Survival(nstage)
0
1
2
3
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500

Σχήμα Α΄.5:


Survival(age)
1
2
3
4
Survival(site)
1
2
4
Survival(tstage)
1
2
3
4
Survival(nstage)
0
1
2
3

Σχήμα Α΄.6:

Βιβλιογραφία

[1] Νικόλαος Δ. Παπαδάτος : Θεωρία Πιθανοτήτων, Αθήνα 2006

[2] Χαράλαμπος Δαμιανού, Μάρκος Κούτρας : Εισαγωγή στην Στατιστική I, II

[3] A.W. Van De Vaart : Asymptotic Statistics

[4] Michael Borenstein, Larry V. Hedges, Julian P. T. Higgins, Hannah R. Rothstein: Introduction to Meta-Analysis

[5] Larry V. Hedges (1981) : "Distribution theory for Glass's estimator of effect size and related estimators". Journal of Educational Statistics 6 (2): 107–128.

[6] Fisher, R.A. (1915). "Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population". Biometrika (Biometrika Trust) 10 (4): 507–521

[7] Meta-analysis of screening and diagnostic tests. Hasselblad, Vic; Hedges, Larry V. Psychological Bulletin, Vol 117(1), Jan 1995, 167-178

[8] S. K Sahu Statistical Methods II chapter 2,3 Σημειώσεις Μαθήματος Academic year 02-03

[9] Φώτης Σιάννης : Introduction to Longitudinal Data Analysis, Σημειώσεις Μαθήματος

[10] Charles S. Davis : Statistical Methods for the Analysis of Repeated Measurements

[11] Sommer A, Tarwotjo I, Hussaini G, Susanto D. Increased mortality in children with mild vitamin A deficiency. Lancet. 1983

[12] David Collet : Modelling Survival data in Medical Research

[13] A clinical Trial in the Trt. of Carcinoma of the Oropharynx, SOURCE: The Statistical Analysis of Failure Time Data, by JD Kalbfleisch RL Prentice, (1980)

[14] Bradley Efron : Censored Data and the Bootstrap, Journal of the American Statistical Association, Vol. 76, No. 374 (Jun., 1981), pp. 312- 319

[15] Jackson D, Riley R, White IR.: Multivariate meta-analysis: Potential and promise. Statistics in Medicine 2011; 30:2481–2498. DOI: 10.1002/sim.4172.

[16] Barrett JK, Farewell VT,,Siannis F,Tierney JF, Higgins J.P. T. : Two-stage meta-analysis of survival data from individual participants using percentile ratios, Stat Med. 2012 Dec 30;31(30):4296-308. doi: 10.1002/sim.5516. Epub 2012 Jul 24.

[17] Siannis F, Barrett JK, Farewell VT, Tierney JF. : One-stage parametric meta-analysis of time-to-event outcomes. Stat Med. 2010 Dec 20;29(29):3030-45. doi: 10.1002/sim.4086. Epub 2010 Oct 20.

[18] Lloyd A. Mancl, Timothy A. DeRouen : A Covariance Estimator for GEE with Improved Small-Sample Properties, Biometrics Volume 57, Issue 1, pages 126–134, March 2001