ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

ΤΟΜΕΑΣ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΕΠΙΧΕΙΡΗΣΙΑΚΗΣ ΕΡΕΥΝΑΣ

Μεταπτυχιακό Δίπλωμα Ειδίκευσης

στη Στατιστική και Επιχειρησιακή Έρευνα

ΣΥΓΚΡΙΣΗ ΜΟΝΤΕΛΩΝ ΓΙΑ ΤΗΝ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΕΠΙΒΙΩΣΗΣ ΜΕ ΔΙΑΚΕΚΟΜΜΕΝΕΣ ΠΑΡΑΤΗΡΗΣΕΙΣ ΠΙΘΑΝΟΝ ΠΛΗΡΟΦΟΡΙΑΚΕΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΦΩΤΕΙΝΗ-ΘΕΑΝΩ ΒΑΡΕΛΤΖΗ

Επιβλέπων: Λέκτορας Φώτιος Σιάννης

Αθήνα 2012

Η παρούσα Διπλωματική Εργασία

Εκπονήθηκε στα πλαίσια των σπουδών

για την απόκτηση του

Μεταπτυχιακού Διπλώματος Ειδίκευσης

στ……

……..

που απονέμει το

Τμήμα Μαθηματικών

του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών

Εγκρίθηκε την ……..από Εξεταστική Επιτροπή

αποτελούμενη από τους :

Ονοματεπώνυμο	Βαθμίδα	Υπογραφή
……..(επιβλέπων Καθηγητής)	……	……..
……	……..	…….....
……	……	……...

Στους γονείς μου,
Ηλία και Μαρία.

Φωτεινή-Θεανώ Βαρελτζή

Σύγκριση μοντέλων για την ανάλυση δεδομένων επιβίωσης με διακεκομμένες παρατηρήσεις πιθανόν πληροφοριακές

Αθήνα 2012

Πρόλογος

Στην ανάλυση επιβίωσης υπάρχει πλήθος στατιστικών μοντέλων για την ανάλυση δεδομένων στα οποία συνυπάρχουν πληροφοριακές διακεκομμένες παρατηρήσεις. Στόχος αυτής της εργασίας είναι να καταγράψει και να περιγράψει συγκεκριμένα μοντέλα προκειμένου να αντιμετώπιστούν δεδομένα επιβίωσης με αυτό το ιδιαίτερο χαρακτηριστικό. Η πληροφορία που μπορεί να κρύβεται στις διακεκομμένες παρατηρήσεις είναι μια παράμετρος που μπορεί να έχει τεράστιο αντίκτυπο στην προτεινόμενη ανάλυση των δεδομένων. Σημαντική παράμετρος επίσης μπορεί να είναι και το πλήθος των διακεκομμένων παρατηρήσεων. Για το λόγο αυτό στην παρούσα εργασία διερευνάται η συμπεριφορά των μοντέλων αυτών σε προσομοιωμένα δεδομένα διαφόρων μεγεθών και με διαφορετικό ποσοστό διακεκομμένων παρατηρήσεων κάθε φορά, ώστε να διαπιστώθεί εάν κάποια από αυτά αναλύουν τα δεδομένα ακριβέστερα από κάποια άλλα με βάση κριτήρια που ορίζονται εκ των προτέρων.

Η παρούσα εργασία αποτελείται από τέσσερα κεφάλαια. Το πρώτο από αυτά αποτελεί μια εισαγωγή σε βασικές έννοιες, χαρακτηριστικά και μεθόδους ανάλυσης δεδομένων επιβίωσης. Στο δεύτερο κεφάλαιο γίνεται μια περιγραφή σε διάφορους τύπους ελλιπών δεδομένων που συναντώνται

στη βιβλιογραφία. Στο τρίτο κεφάλαιο παρουσιάζονται λεπτομερώς δύο μέθοδοι ανάλυσης δεδομένων με πληροφοριακές διακεκομμένες παρατηρήσεις. Τέλος, στο τέταρτο κεφάλαιο καταγράφονται πρακτικές εφαρμογές και συμπεράσματα που προκύπτουν από τις προτεινόμενες μεθόδους στα προσομοιωμένα δεδομένα.

Σε αυτό το σημείο αισθάνομαι βαθιά την ανάγκη να ευχαριστήσω θερμά και ουσιαστικά τους καθηγητές μου, κ. Απόστολο Μπουρνέτα, κ.Αντώνη Οικονόμου και κ. Ευτυχία Βαγγελάτου που μου έδωσαν την ευκαιρία να ξεκινήσω τον κύκλο των μεταπτυχιακών μου σπουδών καθώς και την κ. Λουκία Μελιγκοτσίδου και τον κ. Φώτη Σιάννη που με ενέπνευσαν σε αυτή τη διαδρομή. Ευχαριστώ διπλά τον επιβλέποντα αυτής της διπλωματικής εργασίας κ. Φώτη Σιάννη που δίχως τη δική του υπομονή, κατανόηση και επιμονή τούτη η εργασία δεν θα είχε ολοκληρωθεί. Η συμβολή του υπήρξε καθοριστική για το αξιόλογο του αποτελέσματος. Δεν μπορώ να μην ευχαριστήσω επίσης τους συμφοιτητές μου για το γνήσιο ενδιαφέρον τους και τη συνεχή υποστήριξη καθόλη τη διάρκεια των μεταπτυχιακών μου σπουδών. Τέλος, ιδιαίτερες ευχαριστίες οφείλω στο φίλο και συνάδελφο Απόστολο Δάμιαλη τόσο για την συνεχή παρότρυνσή του όσο και για τις επιστημονικές υποδείξεις του στην τελική διαμορφωση του κειμένου.

Περιεχόμενα

Κεφάλαιο 1

Ανάλυση επιβίωσης

Η ανάλυση δεδομένων επιβίωσης αναφέρεται στη μελέτη δεδομένων που αφορούν στο χρόνο που μεσολαβεί μεταξύ μιας καθορισμένης χρονικής αφετηρίας και την εκδήλωση κάποιου συμβάντος που μας ενδιαφέρει. Αρχικά, η ανάλυση επιβίωσης εφαρμόστηκε σε ιατρικές έρευνες και χρησιμοποιήθηκε ως εργαλείο για να περιγραφούν κυριολεκτικά οι χρόνοι επιβίωσης. Ωστόσο, το υπό μελέτη γεγονός δεν είναι πάντα ο θάνατος, αλλά μπορεί για παράδειγμα να είναι η ανακούφιση ενός ασθενή από επώδυνα συμπτώματα ή η ημέρα που ένας ασθενής θα πάρει εξιτήριο από το νοσοκομείο που νοσηλευόταν. Όσο για την καθορισμένη αφετηρία της μελέτης, μπορεί για παράδειγμα να είναι η στιγμή που σε κάποιον διαγνώστηκε μια ασθένεια ή η στιγμή που κάποιος ασθενής ξεκίνησε μια συγκεκριμένη θεραπεία.

Οι μέθοδοι που χρησιμοποιούνται στην ανάλυση επιβίωσης εφαρμόζονται σε δεδομένα που άπτονται διαφόρων επιστημονικών πεδίων και κλάδων. Έτσι, μέσω της ανάλυσης επιβίωσης μπορούμε να μελετήσουμε το χρόνο που χρειάζεται κάποιος για να φέρει σε πέρας μια συγκεκριμένη εργασία,

το χρόνο αντοχής ενός μηχανήματος σε δεδομένες συνθήκες χρήσης κτλ.

Η μεθοδολογία που χρησιμοποιούμε στην ανάλυση δεδομένων επιβίωσης διαφέρει από τις καθιερωμένες μεθόδους που χρησιμοποιούμε στη στατιστική ανάλυση δεδομένων κι αυτό διότι η φύση των δεδομένων επιβίωσης είναι διαφορετική. Τα παρακάτω ιδιαίτερα χαρακτηριστικά θα μπορούσαμε να πούμε, ότι συνυπάρχουν σε δεδομένα επιβίωσης και λίγο ή πολύ περιπλέκουν τη στατιστική τους ανάλυση. Συγκεκριμένα, το χρονικό διάστημα παρακολούθησης σε μια μελέτη διαφέρει από ασθενή σε ασθενή (άλλος παρακολουθείται για έναν μήνα και άλλος για έναν χρόνο), το συμβάν που μας ενδιαφέρει και βάσει του οποίου θέλουμε να βγάλουμε συμπεράσματα εκδηλώνεται απροσδόκητα στο παρατηρούμενο δείγμα (μπορεί να συμβεί στα τρία ή στα πέντε χρόνια) και επιπλέον οι διακεκομμένοι χρόνοι επιβίωσης, που είναι σχεδόν σίγουρο ότι θα εμπεριέχονται στο παρατηρούμενο δείγμα, παρατηρήσεις δηλαδή με άγνωστη ή εν μέρει καταγεγραμμένη διάρκεια ζωής, αποτελούν μερικά από τα χαρακτηριστικά των δεδομένων επιβίωσης που καθιστούν την ανάλυση τους ιδιαίτερη.

Σε αυτό το κεφάλαιο αναφερόμαστε σε βασικές έννοιες που σχετίζονται με δεδομένα επιβίωσης, στις βασικές συναρτήσεις που περιγράφουν την κατανομή του χρόνου επιβίωσης και στις μεθόδους εκτίμησης αυτών των συναρτήσεων που έχουν αναπτυχθεί προκειμένου να καταλήξουμε σε όσο το δυνατόν ορθά συμπεράσματα σχετικά με το γεγονός που μας ενδιαφέρει. Οι μέθοδοι αυτές χωρίζονται σε τρεις κατηγορίες, τις μη παραμετρικές, τις ημιπαραμετρικές και τις παραμετρικές ανάλογα από το αν υποθέτουμε ή όχι κάποια κατανομή για τους χρόνους επιβίωσης που έχουμε παρατηρήσει στο δείγμα μας.

Σε αυτό το σημείο είναι αναγκαίο να τονίσουμε ότι όλες αυτές οι μέθοδοι που περιγράφονται σε αυτό το κεφάλαιο εφαρμόζονται όταν στην ανάλυση μας οι διακεκομμένες παρατηρήσεις θεωρούνται μη πληροφοριακές, δηλαδή όταν κρίνουμε ότι η διακοπή της καταγραφής του χρόνου επιβίωσης δεν σχετίζεται με το υπό μελέτη συμβάν. Στα επόμενα κεφάλαια θα αναφερθούμε στις μεθόδους ανάλυσης δεδομένων επιβίωσης όταν οι διακεκομμένες παρατηρήσεις κρύβουν σημαντική πληροφορία για το συμβάν που μας ενδιαφέρει, κάτι που στην πραγματικότητα συμβαίνει πολύ συχνά.

1.1 Βασικές έννοιες

Ας θεωρήσουμε ότι έχουμε ένα υποθετικό δείγμα που περιλαμβάνει τους χρόνους παρακολούθησης ασθενών που πάσχουν από κάποιο συγκεκριμένο ιό, από τη στιγμή ανίχνευσης του ιού έως τη στιγμή εκδήλωσης πυρετού μεγαλύτερου από 39 βαθμούς Κελσίου. Το συμβάν αυτό, η στιγμή εκδήλωσης του συγκεκριμένου πυρετού, ονομάζεται *αποτυχία*. Η χρονική αφετηρία που ο κάθε ασθενής εισέρχεται στη μελέτη πρέπει να είναι ακριβώς καθορισμένη. Στο παράδειγμά μας, αυτή η αφετηρία είναι η στιγμή που ο ιός ανιχνεύεται στο αίμα του κάθε ασθενή και φυσικά διαφέρει από ασθενή σε ασθενή. Μετά τη λήξη της μελέτης καταγράφεται *ο χρόνος του κάθε ασθενή*, δηλαδή ο χρόνος που ο κάθε ασθενής παρέμεινε στη μελέτη, θεωρώντας ότι όλοι οι ασθενείς του δείγματος έχουν κοινή αφετηρία, τη χρονική στιγμή 0. Ο χρόνος που μεσολαβεί από τη χρονική αφετηρία μέχρι την αποτυχία ονομάζεται *χρόνος επιβίωσης ή χρόνος μέχρι την αποτυχία*. Εάν η παρατήρηση είνια διακεκομμένη ο αντίστοιχος χρόνος ονομάζεται χρόνος διακοπής. Ας θεωρήσουμε επίσης ότι οι ασθενείς συγκαταλέγονταν στη μελέτη από την

1η Ιανουαρίου του 2011 έως τις 28 Φεβρουαρίου του 2011 και ότι ανεξάρτητα από το εάν κάποιος από αυτούς εκδήλωσε πυρετό μεγαλύτερο από 39 βαθμούς ή όχι, στις 15 Μαρτίου του 2011 ήταν η προκαθορισμένη λήξη της μελέτης. Ονομάζουμε *χρόνο της μελέτης* τον ημερολογιακό χρόνο που διήρκεσε η μελέτη (από την 1η Ιανουαρίου έως την 15η Μαρτίου).

1.1.1 Βασικές συναρτήσεις

Για να περιγράψουμε μαθηματικά τα δεδομένα επιβίωσης χρησιμοποιούμε δύο θεμελιώδεις συναρτήσεις, τη *συνάρτηση επιβίωσης* $S(t)$ και τη *συνάρτηση κινδύνου* $h(t)$. Ο πραγματικός χρόνος επιβίωσης t ενός ασθενή, μπορεί να περιγραφεί με τη βοήθεια μιας μεταβλητής T, η οποία παίρνει μη αρνητικές τιμές. Έστω T η συνεχής τυχαία μεταβλητή που εκφράζει τη διάρκεια ζωής του υπό μελέτη ασθενή, δηλαδή το χρονικό διάστημα από την αφετηρία έως ότου προκύψει το γεγονός που μας ενδιαφέρει. Οι διαφορετικές τιμές που παίρνει η τυχαία μεταβλητή T συνιστούν μία κατανομή πιθανότητας με αντίστοιχη συνάρτηση πυκνότητας πιθανότητας $f(t)$. Τότε η αθροιστική συνάρτηση κατανομής $F(t)$ της τυχαίας μεταβλητής T είναι

$$F(t) = P(T \leq t) = \int_0^t f(u)\,\mathrm{d}u \tag{1.1}$$

και εκφράζει την πιθανότητα ο πραγματικός χρόνος επιβίωσης να είναι μικρότερος ή ίσος από κάποια τιμή t.

Η συνάρτηση επιβίωσης $S(t)$ ορίζεται ως η πιθανότητα ο πραγματικός χρόνος επιβίωσης να είναι μεγαλύτερος από κάποια τιμή t, δηλαδή

$$S(t) = P(T > t) = \int_t^\infty f(u)\,\mathrm{d}u \tag{1.2}$$

και εκφράζει την πιθανότητα ένας ασθενής να επιζήσει από τη χρονική αφετηρία έως κάποια στιγμή μετά τη στιγμή t.

Η σχέση που συνδέει τις δύο παραπάνω συναρτήσεις είναι

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t), \tag{1.3}$$

επομένως η συνάρτηση επιβίωσης είναι μια φθίνουσα συνάρτηση, που σημαίνει ότι η πιθανότητα ένας ασθενής να επιβιώσει πέραν της χρονικής στιγμής t κατά κανόνα μειώνεται καθώς ο χρόνος περνάει. Το πόσο απότομα θα φθίνει η συνάρτηση επιβίωσης εξαρτάται κάθε φορά από την κατανομή που ακολουθούν οι χρόνοι επιβίωσης στο εξεταζόμενο δείγμα. Συνήθως, η τιμή της συνάρτησης επιβίωσης είναι 1 τη στιγμή που ένας ασθενής εισέρχεται στη μελέτη και 0 όταν ο χρόνος τείνει στο άπειρο. Για την τυχαία μεταβλητή T η σχέση που συνδέει τη συνάρτηση πυκνότητας πιθανότητας με τη συνάρτηση επιβίωσης είναι

$$f(t) = \frac{\mathrm{d}F(t)}{\mathrm{d}t} = \frac{\mathrm{d}(1 - S(t))}{\mathrm{d}t} = -\frac{\mathrm{d}S(t)}{\mathrm{d}t}. \tag{1.4}$$

Η συνάρτηση κινδύνου $h(t)$ ορίζεται ως η πιθανότητα σε έναν ασθενή να εκδηλωθεί το υπό μελέτη γεγονός, την αμέσως επόμενη χρονική στιγμή, δηλαδή

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \tag{1.5}$$

και εκφράζει τον κίνδυνο του συμβάντος σε κάποια χρονική στιγμή t. Σύμφωνα με βασικές σχέσεις από τη θεωρία πιθανοτήτων ισχύει ότι

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{P(T \geq t)\Delta t} = \frac{f(t)}{S(t)} \tag{1.6}$$

και επομένως

$$h(t) = -\frac{1}{S(t)} \frac{\mathrm{d}S(t)}{\mathrm{d}t} = -\frac{\mathrm{d}\log S(t)}{\mathrm{d}t}. \tag{1.7}$$

Μία επιπρόσθετη συνάρτηση που χρησιμοποιούμε ευρύτατα στην ανάλυση δεδομένων επιβίωσης και προκύπτει άμεσα από τη συνάρτηση κινδύνου είναι η *αθροιστική συνάρτηση κινδύνου* $H(t)$

$$H(t) = \int_0^t h(u)\,\mathrm{d}u. \tag{1.8}$$

Αποδεικνύουμε χρησιμοποιώντας την σχέση (1.6) ότι

$$S(t) = \exp(-H(t)) \tag{1.9}$$

ή

$$H(t) = -\log S(t). \tag{1.10}$$

Χρησιμοποιώντας όλα τα παραπάνω καταλήγουμε στην εξής ισότητα που περιγράφει τη σχέση μεταξύ των πιο βασικών συναρτήσεων στην ανάλυση δεδομένων επιβίωσης

$$h(t) = \frac{f(t)}{S(t)} = \frac{-\,\mathrm{d}\log S(t)}{\mathrm{d}t} = \frac{\mathrm{d}H(t)}{\mathrm{d}t}. \tag{1.11}$$

Επομένως, αν γνωρίζουμε μία εκ των τριών συναρτήσεων $h(t)$, $S(t)$, $H(t)$ μπορούμε να υπολογίσουμε και τις υπόλοιπες.

1.2 Μη παραμετρικές μέθοδοι

Για την τετριμμένη περίπτωση όπου έχουμε ένα μονό δείγμα με όλους τους χρόνους επιβίωσης παρατηρημένους, δηλαδή χωρίς καμία διακεκομμένη παρατήρηση, τότε η συνάρτηση επιβίωσης $S(t)$ μπορεί να εκτιμηθεί από την *εμπειρική συνάρτηση επιβίωσης*

$$\hat{S}(t) = \frac{\text{πλήθος ασθενών με χρόνο επιβίωσης} \geq t}{\text{πλήθος ασθενών σε όλο το δείγμα}}. \tag{1.12}$$

Μπορούμε να παρατηρήσουμε ότι η συνάρτηση $\hat{S}(t)$ παίρνει την τιμή 1 στο χρονικό διάστημα από την αφετηρία της μελέτης έως και προτού το πρώτο παρατηρημένο συμβάν(πρώτη αποτυχία) και την τιμή 0 στο χρονικό διάστημα από το τελευταίο συμβάν(τελευταία αποτυχία) έως και το τέλος της μελέτης. Επίσης, η τιμή της εμπειρικής συνάρτησης επιβίωσης είναι σταθερή κατά το χρονικό διάστημα μεταξύ δύο διαδοχικών αποτυχιών οπότε στο αντίστοιχο γράφημα αναμένουμε μια κλιμακωτή απεικόνιση. Για τις εκτιμήσεις των συναρτήσεων ισχύει

$$\hat{S}(t) = 1 - \hat{F}(t), \tag{1.13}$$

όπου $\hat{F}(t)$ είναι η *εμπειρική αθροιστική συνάρτηση* κατανομής. Για την περιπτωση στην οποία στο δείγμα μας υπάρχουν και διακεκομμένες παρατηρήσεις η παραπάνω εκτιμήτρια δεν μπορεί να εφαρμοστεί. Τότε συχνά κάνουμε χρήση δύο άλλων εκτιμητών της συνάρτησης επιβίωσης, του *πίνακα ζωής ή αναλογιστικού εκτιμητή* και της πολύ δημοφιλούς και χρήσιμης εκτιμήτριας *Kaplan–Meier*.

1.2.1 Ο εκτιμητής πίνακας ζωής

Για τον υπολογισμό του εκτιμητή πίνακα ζωής, χωρίζουμε την περίοδο που πραγματοποιήθηκε η μελέτη σε k διαδοχικά χρονικά διαστήματα, συνήθως ισομήκη. Έστω ότι στο j-οστό από αυτά τα διαστήματα, δηλαδή από τη στιγμή t_j μέχρι τη στιγμή t_{j+1} παρατηρήσαμε d_j αποτυχίες και c_j διακεκομμένες παρατηρήσεις αντίστοιχα. Επιπλέον έστω ότι οι ασθενείς που επιβιώνουν μέχρι και ακριβώς πριν τη στιγμή t_j και άρα είναι σε κίνδυνο για αποτυχία στην αρχή του j-οστού διαστήματος είναι n_j. Τότε κάνοντας την υπόθεση ότι οι διακεκομμένες παρατηρήσεις εκδηλώνονται ομοιόμορφα σε

κάθε χρονικό διάστημα (αναλογιστικός ισχυρισμός), προκύπτει ότι ο μέσος αριθμός των ασθενών που βρίσκονται σε κίνδυνο στο j-οστό διάστημα είναι

$$n'_j = n_j - \frac{c_j}{2} \tag{1.14}$$

και η πιθανότητα να συμβεί μια αποτυχία στο j-οστό διάστημα είναι d_j/n'_j οπότε και η αντίστοιχη πιθανότητα επιβίωσης στο ίδιο διάστημα είναι $1-d_j/n'_j$ ή $(n'_j - d_j)/n'_j$. Επομένως εάν θέλουμε να υπολογίσουμε την πιθανότητα ενός ασθενή να επιβιώσει μετά την αρχή του m-οστού διαστήματος, θα λάβουμε υπόψη ότι έχει επιβιώσει σε όλο το προηγούμενο χρονικό διάστημα, οπότε ο εκτιμητής πινακας ζωής της συνάρτησης επιβίωσης είναι

$$\hat{S}(t) = \prod_{j=1}^{m} \frac{(n'_j - d_j)}{n'_j}, \tag{1.15}$$

με $t_m \leq t < t_{m+1}$ και $m = 1, \ldots, k$.

Αξίζει να παρατηρήσουμε ότι και αυτή η εκτίμηση της συνάρτησης επιβίωσης, $\hat{S}(t)$, είναι μια σταθερή ανά διαστήματα συνάρτηση. Ακόμα, αξιοσημείωτο είναι το γεγονός ότι η παραπάνω μέθοδος εύρεσης εκτιμητή είναι ιδιαίτερα ευαίσθητη στο πως θα γίνει η επιλογή των χρονικών διαστημάτων. Για το λόγο αυτό αυτός ο εκτιμητής εφαρμόζεται με επιτυχία όταν δεν γνωρίζουμε τους ακριβείς χρόνους των αποτυχιών αλλά τον αριθμό των αποτυχιών και των διακεκομμένων παρατηρήσεων που παρατηρήθηκαν σε μια σειρά από διαδοχικά χρονικά διαστήματα. Παρόλα αυτά μπορούμε να χρησιμοποιήσουμε τον παραπάνω εκτιμητή και όταν γνωρίζουμε με ακρίβεια τις χρονικές στιγμές των αποτυχιών, ωστόσο η ομαδοποίηση των χρόνων επιβίωσης έχει ως αποτέλεσμα την απώλεια μέρους της συνολικής πληροφορίας.

1.2.2 Η εκτιμήτρια Kaplan–Meier

Με παρόμοιο τρόπο θα υπολογίσουμε και την εκτιμήτρια Kaplan–Meier της συνάρτησης επιβίωσης, η οποία είναι πιο αποτελεσματική συγκριτικά με την προηγούμενη στην περίπτωση που γνωρίζουμε τους ακριβείς χρόνους των αποτυχιών. Η διαφοροποίηση των δύο εκτιμητών έγκειται στο ότι η επιλογή των διαστημάτων για την εκτιμήτρια Kaplan–Meier, γίνεται σύμφωνα με τους παρατηρούμενους χρόνους επιβίωσης. Έτσι κάθε διάστημα κατασκευάζεται με τέτοιον τρόπο ώστε να περιλαμβάνει μία αποτυχία την οποία παίρνουμε στην αρχή του κάθε διαστήματος. Οπότε σε κάθε διάστημα εκδηλώνεται μία αποτυχία στην αρχή του, εκτός εάν στο παρατηρούμενο δείγμα πραγματοποιήθηκαν ταυτόχρονα πολλές αποτυχίες και φυσικά εκτός από το πρώτο διάστημα που έχει ως αρχή την αφετηρία της μελέτης και ως πέρας τη στιγμή ακριβώς πριν την πρώτη αποτυχία. Ως εκ τούτου αυτό το διάστημα δεν περιλαμβάνει αποτυχία.

Για τον υπολογισμό της εκτιμήτριας Kaplan–Meier χωρίζουμε την περίοδο που πραγματοποιήθηκε η μελέτη σύμφωνα με τον τρόπο που ορίστηκε παραπάνω, σε k διαδοχικά χρονικά διαστήματα, όσα και τα γεγονότα που εκδηλώθηκαν. Έτσι στο j-οστό από αυτά τα διαστήματα, όπως και σεκάθε άλλο διάστημα, δηλαδή από τη στιγμή t_j μέχρι τη στιγμή t_{j+1} παρατηρήσαμε 1 αποτυχία. Η συνεισφορά των διακεκομμένων παρατηρήσεων σε αυτή τη μέθοδο έγκειται στο γεγονός ότι μέσω αυτών γνωρίζουμε τον ακριβή αριθμό των ασθενών που συνεχίζουν να συμετέχουν στη μελέτη ακριβώς πριν τη χρονική έναρξη του κάθε διαστήματος. Έστω επομένως ότι οι ασθενείς που συνεχίζουν να συμμετέχουν στη μελέτη(όσοι επιβιώνουν, εξαιρώντας όσους διέκοψαν) μέχρι και ακριβώς πριν τη στιγμή t_j είναι n_j. Τότε κά-

νοντας την υπόθεση ότι οι αποτυχίες συμβαίνουν ανεξάρτητα η μία από την άλλη, προκύπτει ότι η πιθανότητα για αποτυχία στο j-οστό διάστημα είναι d_j/n_j, οπότε και η αντίστοιχη πιθανότητα επιβίωσης στο ίδιο διάστημα είναι $(1-d_j)/n_j$ ή $(n_j-d_j)/n_j$.

Επομένως εάν θέλουμε να υπολογίσουμε την πιθανότητα ενός ασθενή να επιβιώσει μετά την αρχή του m-οστού διαστήματος, θα λάβουμε προφανώς υπόψη ότι έχει επιβιώσει σε όλο το προηγούμενο χρονικό διάστημα, οπότε η εκτιμήτρια Kaplan–Meier της συνάρτησης επιβίωσης είναι

$$\hat{S}(t) = \prod_{j=1}^{m} \frac{(n_j - d_j)}{n_j}, \tag{1.16}$$

με $t_m \leq t < t_{m+1}$ και $m = 1, \ldots, k$.

Μία εκτίμηση της διασποράς της εκτιμήτριας Kaplan–Meier δίνεται από τον τύπο του Greenwood

$$\hat{V}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}. \tag{1.17}$$

Για μεγάλο δείγμα μπορούμε να θεωρήσουμε ότι η εκτιμήτρια Kaplan–Meier ακολουθεί προσεγγιστικά την κανονική κατανομή με μέση τιμή $S(t)$ και διασπορά που εκτιμάται από τον τύπο του Greenwood. Ακολούθως προκύπτει ότι ένα $100(1-a)\%$ διάστημα εμπιστοσύνης για τη συνάρτηση επιβίωσης $S(t)$ είναι

$$\text{από} \quad \hat{S}(t) - z_{(1-a/2)}\sqrt{\hat{V}(\hat{S}(t))} \quad \text{έως} \quad \hat{S}(t) + z_{(1-a/2)}\sqrt{\hat{V}(\hat{S}(t))}.$$

Η αθροιστική συνάρτηση κινδύνου $H(t)$ μπορεί να εκτιμηθεί άμεσα μέσω της εκτιμήτριας Kaplan–Meier

$$\hat{H}(t) = -\log \hat{S}(t), \tag{1.18}$$

εντούτοις συχνά χρησιμοποιούμε την εκτιμήτρια που προτάθηκε από τους Nelson και Aalen

$$\hat{H}(t) = \sum_{j=1}^{m} \frac{d_j}{n_j}, \tag{1.19}$$

η οποία είναι επίσης μια σταθερή ανά διαστήματα συνάρτηση. Για να κάνουμε πιο κατανοητή τη διαδικασία υπολογισμού της εκτιμήτριας Kaplan–Meier παραθέτουμε ένα παράδειγμα με λίγα δεδομένα.

Παράδειγμα 1.1

Θεωρούμε τα δεδομένα από 18 γυναίκες, των οποίων μελετήθηκε ο χρόνος σε εβδομάδες, από τη στιγμή που ξεκίνησαν να χρησιμοποιούν μία ενδομήτρια συσκευή με σκοπό την αντισύλληψη έως ότου να πάψουν να την χρησιμοποιούν λόγω προβλημάτων αιμορραγίας κατά την διάρκεια της εμμηνόρροιας. Η μελέτη διήρκεσε δύο χρόνια.

10	13*	18*	19	23*	30	36	38*	54*
56*	59	75	93	97	104*	107	107*	107*

Οι ηλικίες των γυναικών του δείγματος είναι μεταξύ 18 και 35 ετών και όλες είχαν ήδη την εμπειρία δύο προηγούμενων κυήσεων. Οι παρατηρήσεις που σημειώνονται με * είναι διακεκομμένες. Η εκτιμήτρια Kaplan–Meier και οι υπολογισμοί που οδηγούν σε αυτή, σύμφωνα με την παραπάνω διαδικασία παρουσιάζονται στον παρακάτω πίνακα.

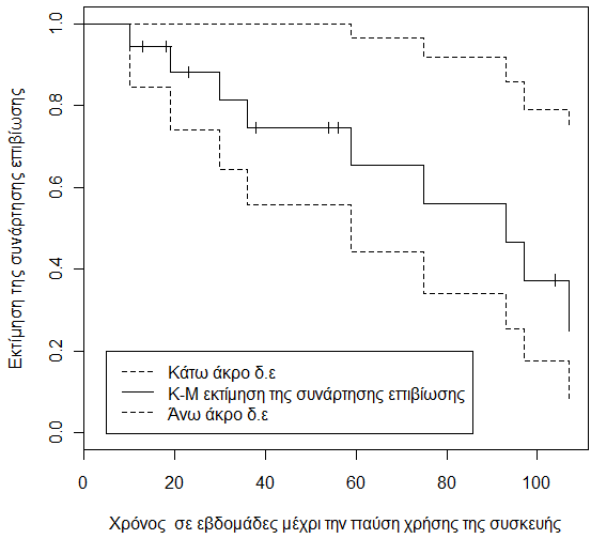
Διάστημα	n_j	d_j	$(n_j - d_j)/n_j$	$\hat{S}(t)$
0−	18	0	1.0000	1.0000
10−	18	1	0.9444	0.9444
19−	15	1	0.9333	0.8815
30−	13	1	1.9231	0.8137
36−	12	1	0.9167	0.7459
59−	8	1	0.8750	0.6526
75−	7	1	0.8571	0.5594
93−	6	1	0.8333	0.4662
97−	5	1	0.8000	0.3729
107−	3	1	0.6667	0.2486

Αν θεωρήσουμε ότι μας ενδιαφέρει η πιθανότητα μία ασθενής να χρησιμοποίησε την εν λόγω συσκευή για τουλάχιστον 30 εβδομάδες, ή ισοδύναμα η πιθανότητα μία ασθενής να μην εμφάνισε προβλήματα αιμορραγίας μέχρι και την 30η εβδομάδα χρήσης της συσκευής, τότε αυτή είναι ίση με $\hat{S}(t) = 1 \times 0.9444 \times 0.8815 \times 0.8137 \simeq 0.6774$.

Με χρήση του στατιστικού πακέτου R προκύπτει στο γράφημα 1.1, η γραφική παράσταση της εκτιμήτριας Kaplan–Meier για τη συνάρτηση του χρόνου επιβίωσης, όπως αυτή ορίστηκε για τις παραπάνω ασθενείς.

1.3 Ένα μοντέλο αναλογικών κινδύνων

Οι μη παραμετρικές μέθοδοι που περιγράψαμε έως τώρα για την εκτίμηση του χρόνου ζωής είναι πολύ χρήσιμες όταν έχουμε να αναλύσουμε ένα δείγμα στο οποίο περιλαμβάνονται χρόνοι επιβίωσης από δύο ή περισ-



Σχήμα 1.1: Η εκτιμήτρια Kaplan–Meier της συνάρτησης επιβίωσης για τα δεδομένα των ασθενών του παραδείγματος 1.1

σότερες ομάδες ασθενών σύμφωνα με κάποιο συγκεκριμένο κριτήριο (για παράδειγμα σύμφωνα με τη θεραπεία) και θέλουμε να συγκρίνουμε την επιβίωση μεταξύ των ομάδων. Ωστόσο, στις περισσότερες ιατρικές έρευνες, καταχωρούνται επιπρόσθετες πληροφορίες για τους ασθενείς. Σε μια κλινική μελέτη, για παράδειγμα, στην οποία το αντικείμενο των ερευνητών είναι η αναζήτηση της πιο αποτελεσματικής θεραπείας μεταξύ των δύο επικρατέστερων οι επιπρόσθετες πληροφορίες μπορεί να είναι η ηλικία, το φύλο, κάποιες μετρήσεις σχετικές με την φυσιολογία του ασθενή ή το περιβάλλον του αλλά και παράγοντες σχετικοί με τον τρόπο ζωής του. Το ερώτημα που εύλογα προκύπτει είναι ποιες από αυτές τις πληροφορίες, οι οποίες καταγράφονται με τη μορφή μεταβλητών επιδρούν στην επιβίωση των ασθενών. Ο Cox πρότεινε ένα μοντέλο, γνωστό ως *μοντέλο αναλογικών κινδύνων*, με το οποίο εκτιμάμε τον κίνδυνο της εκδήλωσης του συμβάντος που μας ενδιαφέρει σε οποιαδήποτε στιγμή μετά το ξεκίνημα μιας μελέτης.

Πριν αναφερθούμε αναλυτικά στο μοντέλο του Cox, καλό θα ήταν να εξηγήσουμε τι εννοούμε με τον όρο *αναλογικοί κίνδυνοι*. Ας θεωρήσουμε την περίπτωση που έχουμε ένα δείγμα ασθενών, οι οποίοι λαμβάνουν είτε την καθιερωμένη θεραπεία είτε μια νέα εναλλακτική της με αντίστοιχες συναρτήσεις κινδύνου $h_s(t)$ και $h_n(t)$. Τότε εάν υποθέσουμε ότι οι κίνδυνοι είναι αναλογικοί μεταξύ τους, το μοντέλο αναλογικών κινδύνων στην πιο απλή του μορφή είναι

$$h_n(t) = \psi h_s(t), \tag{1.20}$$

με την τιμή του ψ να είναι σταθερή. Μια συνέπεια της υπόθεσης των αναλογικών κινδύνων είναι ότι οι αντίστοιχες συναρτήσεις που απεικονίζουν τις εκτιμήσεις των πραγματικών χρόνων επιβίωσης των ασθενών που λαμβά-

νουν τη μία και την άλλη θεραπεία δεν τέμνονται. Αυτή είναι μια αναγκαία αλλά όχι και ικανή συνθήκη για να υποθέσουμε το μοντέλο αναλογικών κινδύνων. Ακόμα και στην περίπτωση που οι συναρτήσεις με τις εκτιμημένες τιμές είναι παράλληλες δεν συνεπάγεται κατάνάγκη αναλογικότητα στους κινδύνους. Εάν οι γραφικές παραστάσεις των συναρτήσεων που απεικονίζουν τις εκτιμήσεις των πραγματικών χρόνων επιβίωσης των δύο γκρουπ ασθενών τέμνονται, τότε ιδιαίτερη προσοχή πρέπει να δοθεί στην ερμηνεία τους.

Η τιμή του ψ ισούται με το λόγο του κινδύνου για αποτυχία ενός ασθενή που ακολουθεί τη νέα θεραπεία σε σχέση με τον κίνδυνο ενός ασθενή που ακολουθεί την καθιερωμένη, γι αυτό και το ψ ονομάζεται συχνά *σχετικός κίνδυνος ή αναλογία κινδύνου*. Το ψ δηλαδή, φανερώνει πόσο περισσότερο ή λιγότερο αποτελεσματική είναι η νέα θεραπεία σε σχέση με την καθιερωμένη. Έτσι, εάν $\psi < 1$ η νέα θεραπεία είναι περισσότερο ευεργετική από την καθιερωμένη ενώ, το αντίστροφο ισχύει $\psi > 1$. Επίσης, η τιμή του σχετικού κινδύνου ψ δεν μπορεί να πάρει αρνητικές τιμές γι΄ αυτό θέτουμε $\psi = exp(\beta)$. Άρα $\beta = ln(\psi)$, οπότε το β, δηλαδή ο λογάριθμος του σχετικού κινδύνου παίρνει αρνητικές τιμές όταν $\psi < 1$ και θετικές τιμές όταν $\psi < 1$. Στην περίπτωση που $\psi < 1$ η νέα θεραπεία είναι πιο αποτελεσματική από την καθιερωμένη ενώ στην περίπτωση που $\psi > 1$ η καθιερωμένη θεραπεία είναι η πιο αποτελεσματική.

1.3.1 Το μοντέλο του Cox με χρονικά ανεξάρτητες μεταβλητές

Στην γενική περίπτωση, υποθέτουμε ότι ο κίνδυνος για αποτυχία κάποιου ασθενή σε μια δεδομένη στιγμή εξαρτάται από ένα σύνολο μεταβλητών, $X_1, X_2, \ldots, X_p$, οι οποίες ονομάζονται ερμηνευτικές ή επεξηγηματικές. Οι τιμές των μεταβλητών αυτών συνιστούν για κάθε ασθενή ένα διάνυσμα $\boldsymbol{x}$, έτσι ώστε $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)'$. Υποθέτουμε ότι οι τιμές των μεταβλητών αυτών παραμένουν αμετάβλητες στο χρόνο και δρουν στην συνάρτηση κινδύνου με βάση τη σχέση

$$h_i(t) = \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}) h_0(t), \tag{1.21}$$

ή ισοδύναμα

$$h_i(t) = \exp(\sum_{j=1}^{p} \beta_j x_{ji}) h_0(t), \tag{1.22}$$

όπου $h_0(t)$ είναι μια *βασική συνάρτηση κινδύνου* και $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ είναι το διάνυσμα των συντελεστών των ερμηνευτικών μεταβλητών που εκφράζει την ποσοτική επίδραση τους στη συνάρτηση κινδύνου.

Επομένως ο κίνδυνος ενός ασθενή εξαρτάται από δύο μέρη, τη συνάρτηση $h_0(t)$ η οποία εξαρτάται αποκλειστικά από το χρόνο και τη συνάρτηση $\eta_i = \sum_{j=1}^{p} (\beta_j x_{ji})$ η οποία εξαρτάται από τις τιμές των ερμηνευτικών μεταβλητών και είναι γνωστή ως *γραμμικό μέρος ή προγνωστικός δείκτης*. Η βασική συνάρτηση κινδύνου είναι αυθαίρετη και ίδια για όλους τους ασθενείς που συμμετέχουν στη μελέτη ενώ, από τη σχέση (1.21) προκύπτει ότι για έναν ασθενή με $\boldsymbol{x} = (0, 0, \ldots, 0)'$,

$$h_i(t) = h_0(t). \tag{1.23}$$

Άρα η βασική συνάρτηση κινδύνου $h_0(t)$ μπορεί να οριστεί και ως η συνάρτηση κινδύνου ενός ασθενή με τιμή όλων των ερμηνευτικών μεταβλητών ίση με 0. Αυτό είναι το μοντέλο αναλογικών κινδύνων του Cox το οποίο εφαρμόζεται ευρύτατα στην ανάλυση δεδομένων επιβίωσης.

1.3.2 Η εκτίμηση των παραμέτρων στο μοντέλο του Cox

Οι συντελεστές β_j, $j = 1, \ldots, p$ εκτιμώνται με την μέθοδο μερικής πιθανοφάνειας και μάλιστα το μέγιστο επιτυγχάνεται χρησιμοποιώντας αριθμητικές μεθόδους. Πιο αναλυτικά, θεωρούμε ότι έχουμε δεδομένα επιβίωσης n ασθενών και μελετάμε αν και κατά πόσο η επιβίωση αυτών των ασθενών εξαρτάται από ένα σύνολο p μεταβλητών. Υποθέτουμε επίσης ότι σε $r < n$ από αυτούς τους ασθενείς εκδηλώνεται το συμβάν κατά τις διακεκριμένες και διατεταγμένες χρονικές στιγμές $t_{(1)}, t_{(2)}, \ldots, t_{(r)}$ ώστε την στιγμή $t_{(j)}$ να παρατηρείται η j-οστή εκδήλωση του συμβάντος. Συμβολίζουμε με $R(t_{(j)})$ το σύνολο των ασθενών που βρίσκονται σε κίνδυνο αμέσως πριν τη χρονική στιγμή $t_{(j)}$ ενώ, υποθέτουμε ότι τη χρονική στιγμή $t_{(j)}$ εκδηλώνεται το συμβάν μόνο σε έναν ασθενή, $d_j = 1$, δηλαδή δεν παρατηρούνται ίσοι χρόνοι επιβίωσης. Η πιθανότητα να εκδηλωθεί το συμβάν σε κάποιον ασθενή τη j-οστή από τις διατεταγμένες στιγμές εκδήλωσής του, δεδομένου ότι σε κάποιον ασθενή από το σύνολο $R(t_{(j)})$ πράγματι εκδηλώνεται το συμβάν είναι

$$P = \frac{\exp(\boldsymbol{\beta'} \boldsymbol{x}_{(j)})}{\sum_{l \epsilon R(t_{(j)})} \exp(\boldsymbol{\beta'} \boldsymbol{x}_l)}, \tag{1.24}$$

όπου με $\boldsymbol{x}_{(j)}$ συμβολίζουμε το διάνυσμα των ερμηνευτικών μεταβλητών για τον ασθενή στον οποίο εκδηλώνεται το συμβάν τη στιγμή $t_{(j)}$.

Αν πάρουμε το γινόμενο των παραπάνω πιθανοτήτων προκύπτει η συνάρτηση πιθανοφάνειας για το μοντέλο αναλογικών κινδύνων του Cox

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta' x}_{(j)})}{\sum_{l \epsilon R(t_{(j)})} \exp(\boldsymbol{\beta' x}_l)}, \tag{1.25}$$

η οποία μπορεί να γραφτεί και ως εξής

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp(\boldsymbol{\beta' x}_{(i)})}{\sum_{l \epsilon R(t_i)} \exp(\boldsymbol{\beta' x}_l)} \right\}^{\delta_i}, \tag{1.26}$$

εάν υποθέσουμε ότι στα δεδομένα εμπεριέχονται n παρατηρούμενοι χρόνοι επιβίωσης, $t_1, t_2, \ldots, t_n$ και ότι δ_i, $i = 1, \ldots, n$, είναι ένας δείκτης εκδήλωσης του συμβάντος που παίρνει την τιμή 1, εάν ο i-οστός χρόνος εκδήλωσης του γεγονότος πράγματι παρατηρήθηκε και 0, διαφορετικά (διακεκομμένη παρατήρηση).

Η συνάρτηση αυτή, καλείται *μερική πιθανοφάνεια* διότι σε αυτήν δεν γίνεται άμεση χρήση των πραγματικών χρόνων επιβίωσησ(διακεκομμένων ή μη) που παρατηρήθηκαν στο δείγμα. Η μερική πιθανοφάνεια δεν αποτελεί μια συνηθισμένη συνάρτηση πιθανοφάνειας, ωστόσο ο Cox απέδειξε ότι μπορεί να χρησιμοποιηθεί ως τέτοια για την εκτίμηση του β κι έτσι ο εκτιμητής $\hat{\beta}$ είναι αμερόληπτος, συνεπής και ασυμπτωτικά κανονικός. Για να εκτιμήσουμε επομένως το διάνυσμα των συντελεστών των ερμηνευτικών μεταβλητών, μεγιστοποιούμε το λογάριθμο της μερικής πιθανοφάνειας ως προς την κάθε συνιστώσα του. Το σύστημα των p εξισώσεων που προκύπτει λύνεται με αριθμητικές μεθόδους.

Για τον υπολογισμό της μερικής πιθανοφάνειας, υποθέσαμε ότι το υπό μελέτη γεγονός εκδηλώνεται μόνο σε έναν ασθενή την κάθε χρονική στιγμή που αυτό συμβαίνει. Ωστόσο, αυτό δεν είναι πάντα ρεαλιστικό, αν

σκεφτούμε για παράδειγμα την περίπτωση που μελετάται ο χρόνος εκδήλωσης μιας ασθένειας σε μήνες, τότε είναι πολύ πιθανό σε κάποιο μήνα να εμφανίσουν την ασθένεια περισσότεροι από ένας ασθενείς. Στην περίπτωση αυτή, που παρατηρούνται ίσοι χρόνοι επιβίωσης, χρησιμοποιούνται κάποιες προσεγγίσεις της μερικής πιθανοφάνειας.

Στα περισσότερα στατιστικά πακέτα η εφαρμογή του μοντέλου αναλογικών κινδύνων είναι εφικτή και έτσι υπολογίζουμε τις εκτιμήσεις των συντελεστών των παραμέτρων καθώς και τις τυπικές αποκλίσεις τους. Έλεγχοι της σημαντικότητας των μεταβλητών μπορούν να γίνουν με χρήση του λόγου των πιθανοφανειών ή της ελεγχοσυνάρτησης Wald και η επιλογή των στατιστικά σημαντικότερων από αυτές με τις διαδικασίες της κατά βήματα επιλογής που χρησιμοποιούνται ευρύτατα στην ανάλυση παλινδρόμησης. Υπενθυμίζουμε ότι ενώ το μοντέλο του Cox βασίζεται στον ισχυρισμό των αναλογικών κινδύνων, εντούτοις καμία ιδιαίτερη κατανομή πιθανότητας δεν προσδιορίστηκε για τη συνάρτηση επιβίωσης, γι αυτό και το μοντέλο αυτό είναι γνωστό ως *ημιπαραμετρικό*.

1.3.3 Το μοντέλο του Cox με χρονικά εξαρτημένες μεταβλητές

Στο μοντέλο του Cox, που παρουσιάσαμε στην προηγούμενη παράγραφο, οι μεταβλητές που επιδρούν στην επιβίωση, θεωρήθηκαν ανεξάρτητες από το χρόνο. Μάλιστα η τιμή τους καταγράφεται συνήθως κατά την έναρξη της μελέτης και έκτοτε θεωρείται ότι παραμένει σταθερή. Υπάρχουν ωστόσο περιπτώσεις, που η τιμή μιας μεταβλητής μεταβάλλεται κατά τη διάρκεια της μελέτης (για παράδειγμα το μέγεθος ενός κακοήθη όγκου) και

τότε η καταγραφή των τιμών της σε τακτά χρονικά διαστήματα μπορεί να οδηγήσει σε ένα πιο ικανοποιητικό μοντέλο. Στην περίπτωση που οι ερμηνευτικές μεταβλητές δεν είναι σταθερές αλλά εξαρτώνται από το χρόνο, το μοντέλο του Cox τροποποιείται κατάλληλα. Όπως και στα προηγούμενα, θεωρούμε ότι έχουμε n παρατηρήσεις και p μεταβλητές που υποθέτουμε πως επιδρούν στην επιβίωση.

Η σχέση (1.22) για το μοντέλο αναλογικών κινδύνων του Cox γενικεύεται άμεσα στη περίπτωσης ύπαρξης χρονικά εξαρημένων μεταβλητών ως εξής

$$h_i(t) = \exp(\sum_{j=1}^{p} \beta_j x_{ji}(t))h_0(t). \tag{1.27}$$

Σε αυτήν την περίπτωση η βασική συνάρτηση κινδύνου $h_0(t)$ φανερώνει τον κίνδυνο ενός ασθενή να εκδηλώσει το γεγονός, δεδομένου ότι στο ξεκίνημα της μελέτης όλες οι τιμές των ερμηνευτικών μεταβλητών που του αντιστοιχούν ήταν 0 και παρέμειναν μηδενικές καθόλη τη διάρκεια της έρευνας. Η υπόθεση των αναλογικών κινδύνων παύει πλέον να ισχύει, αφού όπως εύκολα μπορούμε να παρατηρήσουμε η ποσότητα $h_i(t)/h_0(t)$, που ονομάζεται *σχετικός κίνδυνος*, δεν είναι πλέον σταθερή.

Αν αναλογιστούμε το λόγο των συναρτήσεων κινδύνου δύο ασθενών τη στιγμή t, έστω του ασθενή r και του ασθενή s, τότε

$$\frac{h_r(t)}{h_s(t)} = \exp(\Big(\beta_1(x_{1r}(t) - x_{1s}(t)) + \cdots + \beta_p(x_{pr}(t) - x_{ps}(t))\Big), \tag{1.28}$$

οπότε ο συντελεστής β_j μπορεί να ερμηνευτεί ως ο λογάριθμος του σχετικού κινδύνου για δύο ασθενείς, για τους οποίους η j-οστή ερμηνευτική μεταβλητή για κάθε χρονική στιγμή t μεταβάλλεται κατά μία μονάδα, όταν όλες οι άλλες την ίδια στιγμή παραμένουν σταθερές. Στην πράξη η εκτίμηση του διανύσματος $\boldsymbol{\beta}$ αποτελεί μια αρκετά πολύπλοκη διαδικασία.

1.3.4 Έλεγχοι καταλληλότητας του μοντέλου αναλογικών κινδύνων

Η προσαρμογή του μοντέλου αναλογικών κινδύνων του Cox στο υπό μελέτη σύνολο δεδομένων, δε θα είχε νόημα αν η υπόθεση των αναλογικών κινδύνων δεν ισχύει σε αυτό. Επομένως, για να αποφανθούμε για τη καταλληλότητα του μοντέλου, πρέπει να εφαρμόσουμε ελέγχους για την υπόθεση των αναλογικών κινδύνων. Σε μια προκαταρκτική ανάλυση δεδομένων επιβίωσης και στην πιο απλή περίπτωση που έχουμε να συγκρίνουμε την αποτελεσματικότητα δύο θεραπειών, μπορούμε να σχεδιάσουμε τις καμπύλες που αντιστοιχούν στις εκτιμημένες συναρτήσεις των χρόνων επιβίωσης των δύο ομάδων και απλά να παρατηρήσουμε εάν με το πέρασμα του χρόνου αυτές αποκλίνουν ή όχι. Εάν αποκλίνουν ισχύει η αναλογικότητα ενώ, εάν αυτές τέμνονται ή είναι παράλληλες δεν ισχύει η αναλογικότητα. Στην περίπτωση αυτή θα επιλέξουμε ένα παραμετρικό μοντέλο για την περιγραφή των δεδομένων.

Ένας άλλος γραφικός τρόπος ελέγχου της αναλογικότητας προκύπτει από τη μορφή που λαμβάνει η συνάρτηση επιβίωσης στο μοντέλο αναλογικών κινδύνου. Άπό το μοντέλο αναλογικών κινδύνων του Cox

$$h_i(t) = \exp(\boldsymbol{\beta'} \boldsymbol{x}_i) h_0(t),$$

προκύπτει άμεσα ότι

$$\int_0^t h_i(u)\,\mathrm{d}u = \exp(\boldsymbol{\beta'} \boldsymbol{x}_i) \int_0^t h_0(u)\,\mathrm{d}u, \tag{1.29}$$

ή ισοδύναμα

$$\log H_i(t) = \boldsymbol{\beta'} \boldsymbol{x}_i + \log H_0(t). \tag{1.30}$$

Καθώς η ποσότητα $\boldsymbol{\beta}'\boldsymbol{x}_i$ είναι σταθερή στο χρόνο, όταν οι ερμηνευτικές μεταβλητές δεν είναι χρονικά εξαρτημένες, η καμπύλη $\log H_i(t)$ για οποιοδήποτε $\boldsymbol{x}_i$, είναι παράλληλη με την καμπύλη $\log H_0(t)$. Αυτό σημαίνει πως αν ισχύει η υπόθεση των αναλογικών κινδύνων, οι καμπύλες που απεικονίζουν τους λογάριθμους των αθροιστικών συναρτήσεων κινδύνου, για ασθενείς με διαφορετικές τιμές στις ανεξάρτητες μεταβλητές, σε σχέση με το χρόνο θα πρέπει να είναι παράλληλες. Επομένως, ένας απλός τρόπος γραφικού ελέγχου για την ισχύ των αναλογικών κινδύνων είναι ο προσδιορισμός των εκτιμήσεων Kaplan–Meier για επιλεγμένα x_i και ο έλεγχος της παραλληλίας που απαιτείται. Μειονέκτημα αυτής της μεθόδου, είναι ότι κατά τον υπολογισμό της εκτιμήτριας Kaplan–Meier δε λαμβάνουμε υπόψη τις επεξηγηματικές μεταβλητές και επιπλέον ότι απαιτείται μεγάλος αριθμός δεδομένων με την ίδια τιμή επεξηγηματικών μεταβλητών για να είναι έγκυρες οι εκτιμήσεις που προκύπτουν. Γι αυτό ο γραφικός έλεγχος εφαρμόζεται στην περίπτωση που υπάρχουν λίγες ερμηνευτικές μεταβλητές και οι τιμές τους μπορούν να ομαδοποιηθούν κατάλληλα.

Ένας πιο τυπικός έλεγχος καταλληλότητας του μοντέλου, προκύπτει από την εξέταση των υπολοίπων, τα οποία μας δείχνουν κατά πόσο τα δεδομένα βρίσκονται σε συμφωνία με τις προϋποθέσεις και τις προβλέψεις του προσαρμοσμένου μοντέλου. Κατάλληλα για τον έλεγχο της υπόθεσης των αναλογικών κινδύνων είναι τα υπόλοιπα Schoenfeld τα οποία έχουν νόημα για κάθε μη διακεκομμένη παρατήρηση

$$r_{Pji} = \delta_i\{x_{ji} - \hat{a}_{ji}\}, \tag{1.31}$$

όπου x_{ji} είναι η τιμή της j-οστής ερμηνευτικής μεταβλητής, $j = 1, \ldots, p$,

για τον i-οστό ασθενή,

$$\hat{a}_{ji} = \frac{\sum_{l \epsilon R(t_i)} x_{jl} \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l)}{\sum_{l \epsilon R(t_i)} \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l)},$$

και $R(t_i)$ είναι το σύνολο των ασθενών σε κίνδυνο τη στιγμή t_i.

Εναλλακτικά, χρησιμοποιούνται τα κλιμακοποιημένα υπόλοιπα Schoenfeld, που προτάθηκαν από τους Grampsch και Therneau (1994),

$$\boldsymbol{r}^*_{Pi} = rvar(\hat{\boldsymbol{\beta}}) \boldsymbol{r}_{Pi}, \tag{1.32}$$

όπου r είναι το πλήθος των μη διακεκομμένων παρατηρήσεων, $var(\hat{\boldsymbol{\beta}})$ είναι ο πίνακας διασποράς του δινύσματος των συντελεστών των ερμηνευτικών μεταβλητών και $\boldsymbol{r}_{Pi} = (r_{P1i}, r_{P2i}, \ldots, r_{Ppi})'$ είναι το διάνυσμα των υπολοίπων που υπολογίσαμε παραπάνω. Στη μορφή αυτή τα υπόλοιπα Schoenfeld είναι πολύ χρήσιμα για τον έλεγχο της υπόθεσης των αναλογικών κινδύνου. Οι Grambsch και Therneau έδειξαν ότι η αναμενόμενη τιμή του i-οστού κλιμακοποιημένου υπολοίπου Schoenfeld, για την j-οστή ερμηνευτική μεταβλητή X_j, είναι $E(r^*_{Pji}) \approx \beta_j(t_i) - \hat{\beta}_j$, όπου ο συντελεστής $\beta_j(t)$ λαμβάνεται ως ένας χρονικά εξαρτημένος συντελεστής της ερμηνευτικής μεταβλητής X_j και $\beta_j(t_i)$ είναι η τιμή του συντελεστή την i-οστή στιγμή εκδήλωσης του γεγονότος, t_i.

Το μοντέλο αναλογικών κινδύνων του Cox απαιτεί $\beta_j(t) = \beta_j$, άρα εάν ισχύει ο ισχυρισμός των αναλογικών κινδύνων για την ερμηνευτική μεταβλητή j, τότε η γραφική παράσταση των $r^*_{Pji} + \beta_j$ επί των t_i θα πρέπει είναι μια οριζόντια γραμμή. Αξίζει να παρατηρήσουμε ότι τα υπόλοιπα Schoenfed δε προσδιορίζονται από τις τιμές της εξαρτημένης μεταβλητής, δηλαδή του χρόνου επιβίωσης t, αλλά από τις ερμηνευτικές μεταβλητές. Επιπλέον, δεν ορίζονται για όλο το δείγμα, αλλά μόνο για τις μη-διακεκομμένες

παρατηρήσεις και γι αυτές αποτελούν διανύσματα. Κάθε μη διακεκομμένη παρατήρηση έχει τόσα υπόλοιπα, όσες είναι και οι ερμηνευτικές μεταβλητές.

Τέλος, στην απλή περίπτωση που έχουμε να συγκρίνουμε την αποτελεσματικότητα δύο θεραπειών, δηλαδή όταν έχουμε το είδος της θεραπείας ως μοναδική ερμηνευτική μεταβλητή στο μοντέλο, μπορούμε να ελέγξουμε την υπόθεση της αναλογικότητας προσθέτοντας μια χρονικά εξαρτημένη μεταβλητή στο μοντέλο. Η συνάρτηση κινδύνου για τον i-οστό ασθενή κάτω από το μοντέλο αναλογικών κινδύνων είναι

$$h_i(t) = \exp(\beta_1 x_{1i}) h_0(t),$$

όπου x_{1i} είναι η τιμή της ερμηνευτικής μεταβλητής X_1 και ισούται με 1 όταν ο ασθενής ακολουθεί τη νέα θεραπεία και με 0 όταν ακολουθεί την καθιερωμένη. Ορίζουμε μία νέα ερμηνευτική μεταβλητή, τη X_2, έτσι ώστε $X_2 = X_1 t$. Εάν προσθέσουμε τη μεταβλητή X_2 στο υπάρχον μοντέλο προκύπτει η εξής συνάρτηση κινδύνου

$$h_i(t) = \exp(\beta_1 x_{1i} + \beta_2 x_{2i} t) h_0(t). \tag{1.33}$$

Τότε ο σχετικός κίνδυνος, δηλαδή ο λόγος του κινδύνου που ελλοχεύει για τον ασθενή που ακολουθεί τη νέα θεραπεία σε σχέση με τον αντίστοιχο κίνδυνο για τον ασθενή που ακολουθεί την καθιερωμένη είναι $\exp(\beta_1 + \beta_2 t)$, αφού η μεταβλητή X_2 παίρνει την τιμή t, για έναν ασθενή που ακολουθεί τη νέα θεραπεία. Αυτή η μορφή σχετικού κινδύνου εξαρτάται από το χρόνο t άρα το μοντέλο αυτό δεν είναι μοντέλο αναλογικών κινδύνων.

Ειδικότερα, εάν $\beta_2 < 0$ ο σχετικός κίνδυνος μειώνεται με το χρόνο που σημαίνει ότι ο κίνδυνος εκδήλωσης του γεγονότος για ασθενείς που ακολουθούν τη νέα θεραπεία μειώνεται με το πέρασμα του χρόνου σε σχέση

με τον κίνδυνο της άλλης ομάδας ασθενών. Το αντίστροφο συμβαίνει όταν το $\beta_2 > 0$. Εάν $\beta_1 < 0$, τότε μπορούμε να πούμε ότι τα ευεργετικά αποτελέσματα της νέας θεραπείας γίνονται πιο εμφανή με το πέρασμα του χρόνου. Ο έλεγχος της αναλογικότητας είναι ισοδύναμος με τον έλεγχο της υπόθεσης $\beta_2 = 0$, η οποία εάν τελικά ισχύει, σημαίνει ότι ο σχετικός κίνδυνος είναι $\exp(\beta_1)$, είναι δηλαδή σταθερός.

1.4 Μη παραμετρικοί έλεγχοι

Όπως έχουμε ήδη αναφέρει οι ερμηνευτικές μεταβλητές που επιδρούν στο χρόνο επιβίωσης μπορεί να είναι ποσοτικές, όπως το αποτέλεσμα μιας συγκεκριμένης εργαστηριακής εξέτασης ή η ηλικία ενός ασθενή, αλλά και ποιοτικές ή κατηγορικές όπως για παράδειγμα το φύλο του ασθενή ή το είδος της θεραπείας που ακολουθεί. Οι διαφορετικές τιμές των κατηγορικών μεταβλητών συνιστούν συνήθως ομάδες ασθενών.

1.4.1 Ο μη παραμετρικός έλεγχος log-rank

Στην ειδική περίπτωση, που υπάρχουν δύο (ή περισσότερες) ομάδες και δεν υπάρχουν άλλες ερμηνευτικές μεταβλητές, μπορεί να εφαρμοστεί ο έλεγχος log-rank προκειμένου να εξεταστεί εάν διαφοροποιείται η συνάρτηση επιβίωσης μεταξύ των δύο(ή περισσότερων) ομάδων.

Πιο συγκεκριμένα, έστω $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$ οι διακεκριμένες χρονικές στιγμές κατά τις οποίες συμβαίνει το υπό μελέτη γεγονός σε ασθενείς που προέρχονται από δύο ομάδες. Θεωρούμε ότι αμέσως πριν τη χρονική στιγμή $t_{(j)}$, $j = 1, \ldots, r$, υπάρχουν n_{ij} άτομα σε κίνδυνο, $i = 1, 2$ και d_{ij}

είναι σε πλήθος οι ασθενείς στους οποίους εκδηλώνεται το συμβάν τη στιγμή $t_{(j)}$. Τότε, τη δεδομένη στιγμή $t_{(j)}$ έχουμε d_j συμβάντα και n_j ασθενείς σε κίνδυνο, δηλαδή $n_j = n_{1j} + n_{2j}$ και $d_j = d_{1j} + d_{2j}$.

Έστω ότι η μηδενική υπόθεση που θέλουμε να εξετάσουμε είναι ότι δεν υπάρχει διαφοροποίηση στους χρόνους επιβίωσης των ασθενών μεταξύ των δύο ομάδων, δηλαδή ισοδύναμα θέλουμε να εξετάσουμε την ανεξαρτησία της επιβίωσης των ασθενών από την ομάδα στην οποία ανήκουν. Κάτω από την υπόθεση αυτής της ανέξαρτησίας, μπορούμε να θεωρήσουμε ότι το d_{1j} είναι μια τυχαία μεταβλητή που ακολουθει υπεργεωμετρική κατανομή, με ελαχιστη τιμή το 0 και μέγιστη το $\min\{d_j, n_{1j}\}$. Σύμφωνα με την κατανομή αυτή η πιθανότητα η τυχαία μεταβλητή που σχετίζεται με τον αριθμό των συμβάντων στους ασθενεις της πρώτης ομάδας να πάρει την τιμή d_{1j} είναι

$$\frac{\binom{d_j}{d_{1j}}\binom{n_j - d_j}{n_{1j} - d_{1j}}}{\binom{n_j}{n_{1j}}}.$$

Η μέση τιμή αυτής της τυχαίας μεταβλητής συμφωνα με τις ιδιότητες της υπεργεωμετρικής κατανομής είναι

$$e_{1j} = \frac{n_{ij} d_j}{n_j}. \tag{1.34}$$

Έτσι e_{1j} είναι ο αναμενόμενος αριθμός των ασθενών που προέρχονται από την 1η ομάδα, στους οποιούς εκδηλώνεται το συμβάν τη στιγμή t_j. Η απόκλιση αυτής της αναμενόμενης τιμής από την παρατηρούμενη είναι $d_{1j} - e_{1j}$, οπότε αθροίζοντας όλες αυτές τις αποκλίσεις για όλα τις στιγμές που συνέβη το γεγονός στους ασθενεις και των δύο ομάδων έχουμε ένα συνολικό μέτρο της διαφοράς των παρατηρούμενων τιμών d_{1j} από τις αναμενόμενες τιμές τους. Η ελεγχοσυνάρτηση που προκύπτει είναι

$$U_L = \sum_{j=1}^{r} (d_{1j} - e_{1j}), \tag{1.35}$$

με μέση τιμή 0, αφού $E(d_{1j}) = e_{1j}$, και διασπορά το άθροισμα των διασπορών των d_{1j}, αφού οι στιγμές εκδήλωσης του συμβάντος είναι ανεξάρτητες η μία από την άλλη.

Επειδή η τυχαία μεταβλήτή d_{1j} ακολουθεί υπεργεωμετρική κατανομή, η διασπορά της είναι

$$v_{1j} = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}, \tag{1.36}$$

οπότε και η διασπορά της ελεγχοσυνάρτησης U_L είναι

$$var(U_L) = \sum_{j=1}^{r} v_{1j}. \tag{1.37}$$

Μπορεί να αποδειχθεί ότι η ελεγχοσυνάρτηση U_L ακολουθεί την κανονική κατανομή όταν ο αριθμός των συμβάντων δεν είναι πολυ μικρός. Τότε προκύπτει επίσης ότι η ελεγχοσυνάρτηση, $U_L/\sqrt{var(U_L)}$, ακολουθεί κανονική κατανομή με μέση τιμή 0 και διασπορά 1, δηλαδή

$$\frac{U_L}{\sqrt{var(U_L)}} \sim N(0,1).$$

Η τελική μορφή του ελέγχου log-rank είναι το τετράγωνο αυτής της ελεγχοσυνάρτησης και ακολουθεί χ^2 κατανομή με 1 βαθμό ελευθερίας

$$W_L = \frac{U_L^2}{var(U_L)} \sim \chi_1^2.$$

Η μηδενική υπόθεση, ότι η επιβίωση ενός ασθενή δεν εξαρτάται από την ομάδα προέλευσης, απορρίπτεται αν η P-τιμή της κατανομής χ^2 με ένα βαθμό ελευθερίας που αντιστοιχεί στη τιμή της παραπάνω ελεγχοσυνάρτησης, είναι μικρότερη από τη τιμή a που ορίζεται ως επίπεδο σημαντικότητας. Ο λόγος που ο έλεγχος log-rank ανήκει στους μη παραμετρικούς ελέγχους, είναι ότι οι συναρτήσεις επιβίωσης, των οποίων η ισότητα ελέγχεται υπό την μηδενική υπόθεση, δε χρειάζεται να υπολογιστούν και για αυτό χρησιμοποιείται ευρύτατα.

1.4.2 Ο μη παραμετρικός έλεγχος Wilcoxon

Ο έλεγχος Wilcoxon αποτελεί μία σταθμισμένη παραλλαγή του ελέγχου log-rank μέσω του οποίου εξετάζεται η ίδια μηδενική υπόθεση, ότι δηλαδή δεν υπάρχει διαφοροποίηση μεταξύ των χρόνων επιβίωσης για ασθενείς που προέρχονται από δύο διαφορετικές ομάδες. Συγκεκριμένα, η ελεγχοσυνάρτηση λαμβάνει τη μορφή

$$U_W = \sum_{j=1}^{r} n_j(d_{1j} - e_{1j}), \tag{1.38}$$

όπου τα d_{1j} και e_{1j} ορίζονται όπως και στον έλεγχο log-rank.

Στον έλεγχο Wilcoxon, η κάθε διαφορά $d_{1j}-e_{1j}$ πολλαπλασιάζεται με το συνολικό αριθμό των ατόμων που βρίσκονται σε κίνδυνο, n_j, με αποτέλεσμα να δίνεται μεγαλύτερο βάρος στο αρχικό κομμάτι του δείγματος, όπου το πλήθος των υπολειπόμενων παρατηρήσεων είναι μεγαλύτερο απο ό,τι στους τελευταίους χρόνους εκδήλωσης του συμβάντος (υπενθυμίζουμε ότι είναι διατεταγμένοι), όπου απομένουν μόνο λίγα άτομα χωρίς να έχει διακοπεί η επιβίωση τους. Έτσι, στη περίπτωση που απαιτείται να δοθεί βαρύτητα στις πιθανές διαφορές μεταξύ των δύο συναρτήσεων επιβίωσης σε αρχικό στάδιο, ο έλεγχος Wilcoxon αποδίδει καλύτερα από τον έλεγχο log-rank.

Η διασπορά της ελεγχοσυνάρτησης U_W είναι

$$var(U_W) = \sum_{j=1}^{r} n_j^2 v_{1j}, \tag{1.39}$$

όπου $v_{1j} = n_{1j}n_{2j}d_j(n_j - d_j)/n_j^2(n_j - 1)$, όπως δείξαμε προηγούμενα. Η τελική μορφή του ελέγχου Wilcoxon είναι

$$W_W = \frac{U_W^2}{var(U_W)},$$

η οποία ακολουθεί χ^2 κατανομή με 1 βαθμό ελευθερίας όταν η μηδενική υπόθεση αληθεύει. Όπως και στον έλεγχο log-rank, η μηδενική υπόθεση, ότι η επιβίωση ενός ασθενή δεν εξαρτάται από την ομάδα προέλευσης, απορρίπτεται αν η P-τιμή της χ^2 με ένα βαθμό ελευθερίας που αντιστοιχεί στη τιμή της παραπάνω ελεγχοσυνάρτησης, είναι μικρότερη από τη τιμή a που ορίζεται ως επίπεδο σημαντικότητας.

1.5 Η κατανομή Weibull

Τόσο στο μοντέλο του Cox όσο και στις μη παραμετρικές τεχνικές εκτίμησης, που χρησιμοποιήσαμε στην αρχή αυτού του κεφαλαίου, δεν υποθέσαμε κάποια συγκεκριμένη κατανομή για τους χρόνους επιβίωσης. Το γεγονός αυτό μας επιτρέπει να εφαρμόζουμε ευρέως αυτές τις μεθόδους στην ανάλυση δεδομένων επιβίωσης. Ωστόσο, όταν η μορφή των δεδομένων επιτρέπει την υιοθέτηση κάποιας συγκεκριμένης κατανομής για τους χρόνους επιβίωσης, τότε τα συμπεράσματα που προκύπτουν είναι πιο ακριβή σε σχέση με εκείνα που θα προέκυπταν χωρίς την υιοθέτησή της. Τα μοντέλα στα οποία έχουμε υποθέσει μία συγκεκριμένη κατανομή για τους χρόνους επιβίωσης ονομάζονται παραμετρικά μοντέλα και με αυτά θα ασχοληθούμε στια επόμενες παραγράφους.

Μια κατανομή πιθανότητας που διαδραματίζει κυρίαρχο ρόλο στην ανάλυση δεδομένων είναι η κατανομή Weibull, γνωστή με το όνομα του Σουηδού Wallodi Weibull, ο οποίος τη χρησιμοποίησε στα μέσα της δεκαετίας του εξήντα για να περιγράψει την αντοχή των υλικών. Τα μοντέλα αναλογικών κινδύνων που βασίζονται στην κατανομή Weibull αποτελούν την πιο θεμελιώδη παραμετρική εκδοχή των μοντέλων αναλογικών κινδύνων στην

ανάλυση δεδομένων επιβίωσης γι αυτό και θα τα αναλύσουμε λεπτομερώς.

Στην πράξη μια λειτουργική μορφή της συνάρτησης κινδύνου είναι

$$h(t) = \lambda\gamma t^{\gamma-1}, \tag{1.40}$$

με τις παραμέτρους λ και γ να παίρνουν μόνο θετικές τιμές και το χρόνο t φυσικά να παίρνει μη αρνητικές τιμές. Το σχήμα της συνάρτησης αυτής, η οποία είναι μονότονη, εξαρτάται κυρίως από την τιμή της παραμέτρου γ γι αυτό και η γ ονομάζεται παράμετρος σχήματος ενώ, η παράμετρος λ ονομάζεται παράμετρος κλίμακας. Η γενική μορφή της συνάρτησης κινδύνου για τις διάφορες τιμές της παραμέτρου γ φαίνεται στο σχήμα 1.2. Η συνάρτηση κινδύνου είναι αύξουσα για $\gamma > 1$, φθίνουσα για $0 < \gamma < 1$ και σταθερή και ίση με λ για $\gamma = 1$. Ειδικά για την περίπτωση που η παράμετρος σχήματος είναι ίση με 1, η κατανομή Weibull συμπίπτει με την εκθετική κατανομή με συνάρτηση πυκνότητας πιθανότητας

$$f(t) = \lambda e^{-\lambda t} \tag{1.41}$$

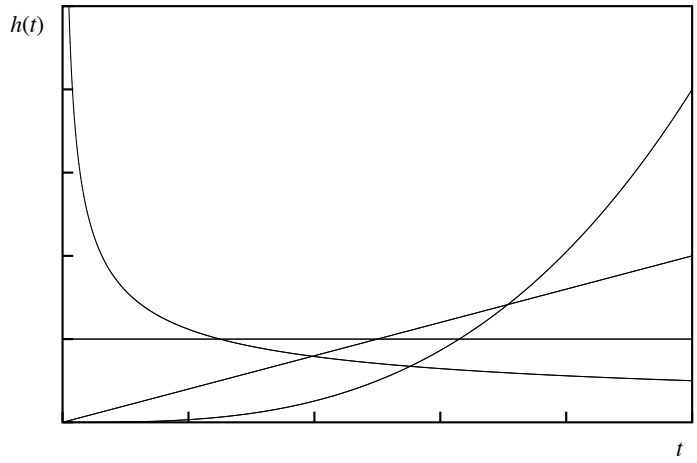
και μέση τιμή λ^{-1}. Όπως προαναφέραμε σε αυτήν την περίπτωση η συνάρτηση κινδύνου είναι σταθερή και ίση με λ, ένα πολύ βολικό και συχνά ρεαλιστικό σενάριο.

Για μια συνάρτηση κινδύνου της μορφής (1.40), η συνάρτηση επιβίωσης που προκύπτει είναι

$$S(t) = \exp\left\{ -\int_0^t \lambda\gamma u^{\gamma-1} \right\} \mathrm{d}u = \exp(-\lambda t^{\gamma}). \tag{1.42}$$

Η αντίστοιχη συνάρτηση πυκνότητας πιθανότητας μιας τυχαίας μεταβλητής T που περιγράφεται από τις δύο παραπάνω συναρτήσεις είναι

$$f(t) = \lambda\gamma t^{\gamma-1} exp(-\lambda t^{\gamma}), \tag{1.43}$$



Σχήμα 1.2: Η μορφή της συνάρτησης κινδύνου $h(t) = \lambda\gamma t^{\gamma-1}$, όταν οι χρόνοι επιβίωσης ακολουθούν την $W(\lambda, \gamma)$ για διάφορες τιμές της παραμέτρου σχήματος γ.

η οποία είναι η πυκνότητα μιας τυχαίας μεταβλητής που ακολουθεί την κατανομή Weibull με παράμετρο σχήματος γ, παράμετρο κλίμακας λ και συμβολίζεται $W(\lambda, \gamma)$. Η κατανομή αυτή είναι ασύμμετρη διότι η δεξιά ουρά της είναι πιο μακριά από την αριστερή και άρα παρουσιάζει θετικό συντελεστή λοξότητας που σημαίνει ότι οι περισσότερες τιμές της βρίσκονται δεξιά της επικρατούσας τιμής.

Η μέση τιμή μιας τυχαίας μεταβλητής T με κατανομή $W(\lambda, \gamma)$ αποδεικνύεται ότι είναι

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

όπου $\Gamma(x)$ είναι η συνάρτηση Γάμμα που ορίζεται από το ολοκλήρωμα

$$\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}\,\mathrm{d}u.$$

Ως αριθμητικό περιγραφικό μέτρο για την κατανομή Weibull, επειδή αυτή έχει θετικό συντελεστή λοξότητας, αντί για τη μέση τιμή που επηρεάζεται από ακραίες τιμές, συχνά χρησιμοποιούμε τη διάμεσο. Η τιμή της διαμέσου για τους χρόνους επιβίωσης είναι μια τιμή του χρόνου t, $t(50)$, για την οποία ισχύει ότι $S\{t(50)\} = 0.5$, έτσι ώστε

$$t(50) = \left\{\frac{1}{\lambda}\log 2\right\}^{1/\gamma}. \tag{1.44}$$

Η διάμεσος των χρόνων επιβίωσης, η τιμή $t(50)$, είναι αυτή η χρονική στιγμή μετά από την οποία περίπου το 50% του πληθυσμού στο δείγμα αναμένεται να επιβιώσει. Έπομένως ο μισός πλήθυσμός αναμένεται να εκδηλώσει το υπό μελέτη συμβάν, μετά την στιγμή $t(50)$. Γενικά το p ποσοστιαίο σημείο της κατανομής Weibull είναι

$$t(p) = \left\{\frac{1}{\lambda}\log\left(\frac{100}{100-p}\right)\right\}^{1/\gamma}. \tag{1.45}$$

1.5.1 Το παραμετρικό μοντέλο

Ας υποθέσουμε ότι σε ένα δείγμα παρατηρήθηκαν οι χρόνοι επιβίωσης n ασθενών και ότι σε r από αυτούς εκδηλώθηκε το συμβάν στις αντίστοιχες χρονικές στιγμές $t_1, \ldots, t_r$. Για τους εναπομείναντες $n-r$ ασθενείς, δεν καταγράφηκε ο χρόνος ζωής διότι από κάποιο χρονικό σημείο t^*, ξεχωριστό για κάθε ασθενή, και μετά χάθηκαν τα ίχνη τους (διακεκομμένες παρατηρήσεις).

Οι r χρόνοι επιβίωσης που καταγράφηκαν συνεισφέρουν στην συνάρτηση πιθανοφάνειας σύμφωνα με το γινόμενο

$$\prod_{j=1}^{r} f(t_j)$$

ενώ, οι χρόνοι επιβίωσης που δεν καταγράφηκαν συνεισφέρουν σύμφωνα με το γινόμενο

$$\prod_{l=1}^{n-r} S(t_l^*)$$

λαμβάνοντας υπόψη ότι μέχρι τη χρονική στιγμή t_l^*, ο εν λόγω ασθενής δεν εκδήλωσε το συμβάν άρα ο χρόνος επιβίωσής του είναι τουλάχιστον t^*. Οπότε η συνάρτηση πιθανοφάνειας για όλες τις παρατηρήσεις είναι το γινόμενο

$$\prod_{j=1}^{r} f(t_j) \prod_{l=1}^{n-r} S(t_l^*).$$

Μια εναλλακτική μορφή της συνάρτησης πιθανοφάνειας προκύπτει εάν θεωρήσουμε ότι το δείγμα μας αποτελείται από n ζευγάρια παρατηρήσεων, (t_i, δ_i), για $i = 1, \ldots, n$. Σύμφωνα με αυτόν τον συμβολισμό, το δ_i, είναι μια μεταβλητή δείκτης που παίρνει την τιμή 0, όταν ο αντίστοιχος χρόνος επιβίωσης του i-οστού ασθενή, t_i, είναι διακεκομμένος και την τιμή 1, όταν η αντίστοιχη παρατήρηση είναι γνωστή. Τότε η συνάρτηση πιθανοφάνειας γράφεται

$$\prod_{i=1}^{n} \{f(t_i)\}^{\delta_i} \{S(t_i)\}^{1-\delta_i}, \tag{1.46}$$

ή

$$\prod_{i=1}^{n} \left\{ \frac{f(t_i)}{S(t_i)} \right\}^{\delta_i} S(t_i),$$

ή σύμφωνα με τις ισχύουσες σχέσεις μεταξύ των βασικών συναρτήσεων στην ανάλυση επιβίωσης

$$\prod_{i=1}^{n} \{h(t_i)^{\delta_i} S(t_i)\}.$$

Οι άγνωστες παράμετροι που υπεισέρχονται στη συνάρτηση πιθανοφάνειας και που συνάδουν με την κατανομή που υποθέτουμε για τους πα-

ρατηρούμενους χρόνους επιβίωσης στο δείγμα, εκτιμώνται με τη μέθοδο μέγιστης πιθανοφάνειας.

Πιο αναλυτικά, αν η κατανομή που υποθέτουμε για τους χρόνους στο δείγμα που θεωρήσαμε παραπάνω είναι η Weibull με παραμέτρους λ και γ, τότε η συνάρτηση πυκνότητας πιθανότητας, η συνάρτηση επιβίωσης και η συνάρτηση κινδύνου είναι αντίστοιχα

$$f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda\gamma t^{\gamma}), S(t) = \exp(-\lambda t^{\gamma}), h(t) = \lambda\gamma t^{\gamma-1},$$

οπότε με αντικατάσταση αυτών των σχέσεων στη συνάρτηση πιθανοφάνειας (1.46) προκύπτει

$$\prod_{i=1}^{n}\left\{\lambda\gamma t_i^{\gamma-1}\exp(-\lambda t_i^{\gamma})\right\}^{\delta_i}\left\{\exp(-\lambda t_i^{\gamma})\right\}^{1-\delta_i},$$

ή ισοδύναμα

$$\prod_{i=1}^{n}\left\{\lambda\gamma t_i^{\gamma-1}\right\}^{\delta_i}\exp(-\lambda t_i^{\gamma}),$$

που είναι μια συνάρτηση των λ και γ, των άγνωστων παραμέτρων στην κατανομή Weibull, οπότε τη συμβολίζουμε $L(\lambda,\gamma)$.

Οι εκτιμητές των λ και γ που μεγιστοποιούν τη συνάρτηση πιθανοφάνειας βρίσκονται παραγωγίζοντας τον λογάριθμο της συνάρτησης πιθανοφάνειας ως προς λ και γ αντίστοιχα και θέτοντας τις αντίστοιχες παραγώγους ίσες με το μηδέν. Στην πράξη, για κατανομές που δίνουν κλειστους τύπους οι ζητούμενοι εκτιμητές $\hat{\lambda}$ και $\hat{\gamma}$ βρίσκονται με αριθμητικές μεθόδους, με πιο δημοφιλή τον αλγόριθμο Newton-Raphson. Στα περισσότερα στατιστικά πακέτα η εφαρμογή της κατανομής Weibull στους παρατηρούμενους χρόνους επιβίωσης είναι εφικτή και έτσι υπολογίζουμε τις εκτιμήσεις των άγνωστων παραμέτρων, τα εκτιμημένα ποσοστιαία σημεία που μας ενδιαφέ-

ρουν σύμφωνα με τον γενικό τύπο

$$\hat{t}(p) = \left\{ \frac{1}{\hat{\lambda}} \log\Big(\frac{100}{100-p}\Big) \right\}^{1/\hat{\gamma}},$$

καθώς και τις τυπικές αποκλίσεις τους.

Επιπλέον, μπορούμε να υπολογίσουμε τα αντίστοιχα $100(1-a)\%$ διαστήματα εμπιστοσύνης για τα ποσοστιαία σημεία, τα οποία είναι προτιμότερο να τα βρούμε μέσω των αντίστοιχων διαστημάτων για τις συναρτήσεις $\log t(p)$. Η τυπική απόκλιση για το λογάριθμο του ποσοστιαίου σημείου είναι

$$se\big\{\log t(p)\big\} = \frac{1}{\hat{t}(p)} se\big\{\hat{t}(p)\big\},$$

και τα αντίστοιχα άκρα του $100(1-a)\%$ διαστήματος εμπιστοσύνης για το $\log t(p)$ είναι

$$\log \hat{t}(p) \pm z_{a/2} se \log \hat{t}(p).$$

Υπολογιίζουμε τα αντίστοιχα $100(1-a)\%$ διαστήματα εμπιστοσύνης για τα ποσοστιαία σημεία $t(p)$ υψώνοντας στην κατάλληλη δύναμη. Για παράδειγμα, τα άκρα του $100(1-a)\%$ διαστήματος εμπιστοσύνης για τη διάμεσο των χρόνων επιβίωσης, $t(50)$, είναι $\hat{t}(50)\exp[\pm z_{a/2} se\big\{\log \hat{t}(50)\big\}]$.

1.5.2 Το παραμετρικό μοντέλο αναλογικών κινδύνων κάτω από την κατανομή Weibull

Όπως έχουμε ήδη διαπιστώσει από την παρούσα εργασία, ένα βολικό μοντέλο για να συγκρίνουμε δύο γκρουπ ασθενών είναι το μοντέλο αναλογικών κινδύνων. Σε αυτήν την παράγραφο, θα θεωρήσουμε ότι στο δείγμα μας υπάρχει μία ερμηνευτική μεταβλητή, η X, βάσει της οποίας γίνεται ο διαχωρισμός των ασθενών σε δύο γκρουπ, το γκρουπ I και το γκρουπ II.

Έστω ότι η τιμή της είναι 0, εάν ο ασθενής προέρχεται από το γκρουπ I και 1, εάν ο ασθενής προέρχεται από το γκρουπ II.

Σύμφωνα με το μοντέλο αναλογικών κινδύνων, ο κίνδυνος εκδήλωσης του συμβάντος στον i-οστό ασθενή τη στιγμή t είναι

$$h_i(t) = e^{\beta x_i} h_0(t), \tag{1.47}$$

όπου x_i είναι η τιμή της ερμηνευτικής μεταβλητής X για τον i-οστό ασθενή. Συνεπώς, η συνάρτηση κινδύνου για έναν ασθενή του γκρουπ I τη στιγμή t είναι $h_0(t)$ και για έναν ασθενή από το γκρουπ II είναι $\psi h_0(t)$, όπου $\psi = exp(\beta)$. Τότε η ποσότητα β είναι ο λογάριθμος του λόγου του κινδύνου που διατρέχει ένας ασθενής από το γκρουπ II ως προς τον κίνδυνο που διατρέχει ένας ασθενής από το γκρουπ I.

Ισχυριζόμενοι επιπλέον ότι οι παρατηρούμενοι χρόνοι επιβίωσης των ασθενών που προέρχονται από το γκρουπ I ακολουθούν την κατανομή Weibull, προκύπτει ότι η συνάρτηση κινδύνου, $h_0(t)$, αυτών των ασθενών είναι

$$h_0(t) = \lambda\gamma t^{\gamma-1} \tag{1.48}$$

και η συνάρτηση κινδύνου $\psi h_0(t)$ των ασθενών του γκρουπ II είναι

$$\psi h_0(t) = \psi\lambda\gamma t^{\gamma-1}. \tag{1.49}$$

Η σχέση (1.49) συμπίπτει με τη συνάρτηση κινδύνου μιας τυχαίας μεταβλητής T που ακολουθεί κατανομή Weibull με παράμετρο σχήματος γ και παράμετρο κλίμακας $\psi\lambda$. Συμπεραίνουμε ότι αν οι χρόνοι επιβίωσης των ασθενών του ενός γκρουπ ακολουθούν κατανομή Weibull με παράμετρο σχήματος γ και ο κίνδυνος εκδήλωσης του συμβάντος τη στιγμή t που διατρέχουν οι ασθενείς που προέρχονται από αυτό το γκρουπ είναι ανάλογος με

εκείνον του άλλου, τότε η κατανομή που ακολουθούν οι χρόνοι επιβίωσης του δεύτερου γκρουπ είναι και πάλι Weibull με την ίδια παράμετρο σχήματος γ. Αυτή η ιδιότητα της κατανομής Weibull είναι γνωστή ως ιδιότητα αναλογικών κινδύνων και εξαιτίας αυτής, η κατανομή αυτή διαδραματίζει τόσο σημαντικό ρόλο στην ανάλυση δεδομένων επιβίωσης.

Στην περίπτωση που στο παρατηρούμενο δείγμα n ασθενών εμπεριέχονται p ερμηνευτικές μεταβλητές $X_1, \ldots, X_p$ με αντίστοιχες τιμές $x_{1i}, \ldots, x_{pi}$ για κάθε ασθενή, το μοντέλο αναλογικών κινδύνων γενικεύεται ως εξής

$$h_i(t) = \exp\left(\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}\right) h_0(t), \tag{1.50}$$

για $i = 1, \ldots, n$. Παρόλο που η μορφή αυτού του μοντέλου είναι παρόμοια με το μοντέλο αναλογικών κινδύνων του Cox υπάρχει μεταξύ τους μια σημαντική διαφοροποίηση αναφορικά με την βασική συνάρτηση κινδύνου $h_0(t)$. Στο παραμετρικό μοντέλο, έχουμε υποθέσει την κατανομή Weibull για τους χρόνους επιβίωσης, γεγονός που επιφέρει μία συγκεκριμένη μορφή της βασικής συνάρτησης κινδύνου ενώ, στο μη παραμετρικό μοντέλο του Cox η μορφή της βασικής συνάρτησης κινδύνου είναι απροσδιόριστη.

Η βασική συνάρτηση κινδύνου $h_0(t)$ στο παραμετρικό μοντέλο υπολογίζεται εύκολα θεωρώντας ότι για κάποιον ασθενή όλες οι τιμές των ερμηνευτικών μεταβλητών είναι ίσες με το 0. Έτσι, εάν υποθέσουμε την κατανομή Weibull, η βασική συνάρτηση κινδύνου για τον i-οστό ασθενή είναι

$$h_0(t) = \lambda\gamma t^{\gamma-1},$$

από την οποία προκύπτει άμεσα η βασική αθροιστική συνάρτηση κινδύνου

$$H_0(t) = \int_0^t \lambda\gamma u^{\gamma-1}\,\mathrm{d}u = \lambda t^{\gamma}. \tag{1.51}$$

Εφαρμόζοντας το μοντέλο αναλογικών κινδύνων του Cox η συνάρτηση κινδύνου για τον i-οστό ασθενή είναι

$$h_i(t) = \exp(\boldsymbol{\beta'} \boldsymbol{x_i}) \lambda \gamma t^{\gamma - 1} \tag{1.52}$$

και η αντίστοιχη αθροιστική συνάρτηση κινδύνου είναι

$$H_i(t) = \exp(\boldsymbol{\beta'} \boldsymbol{x_i}) \lambda t^{\gamma}. \tag{1.53}$$

Επομένως, όταν οι χρόνοι επιβίωσης ακολουθούν την κατανομή Weibull, η συνάρτηση επιβίωσης κάτω από το μοντέλο αναλογικών κινδύνων του Cox παίρνει τη μορφή

$$S_i(t) = \exp\big\{ - H_i(t)\big\} = \exp\big\{ - \exp(\boldsymbol{\beta'} \boldsymbol{x_i}) \lambda t^{\gamma}\big\}, \tag{1.54}$$

ή ισοδύναμα την πιο γενική μορφή

$$S_i(t) = \exp\big\{ - \exp(\boldsymbol{\beta'} \boldsymbol{x_i}) H_0(t)\big\}. \tag{1.55}$$

Και από τις παραπάνω σχέσεις αποδεικνύεται η ιδιότητα αναλογικών κινδύνων της κατανομής Weibull, αφού ο χρόνος επιβίωσης του ασθενή i στη μελέτη ακολουθεί αυτήν την κατανομή με παράμετρο κλίμακας $\lambda \exp(\boldsymbol{\beta'} \boldsymbol{x})$ και παράμετρο σχήματος γ. Επομένως, μπορούμε να πούμε ότι οι ερμηνευτικές μεταβλητές επιδρούν στο μοντέλο που περιγράφεται από τη σχέση (1.50) με τέτοιο τρόπο ώστε ενώ, μεταβάλλεται η παράμετρος κλίμακας της κατανομής Weibull των χρόνων επιβίωσης, η παράμετρος σχήματος της κατανομής παραμένει σταθερή.

Η διαδικασία εκτίμησης των άγνωστων παραμέτρων είναι ακριβώς ίδια με αυτήν που δείξαμε στην περίπτωση που δεν υπήρχαν ερμηνευτικές μεταβλητές. Η μόνη διαφορά που προκύπτει με τις ερμηνευτικές μεταβλητές

έγκειται στο ότι τότε πρέπει να μεγιστοποιήσουμε ταυτόχρονα τη συνάρτηση πιθανοφάνειας και ως προς το διάνυσμα των συντελεστών $\boldsymbol{\beta}$ και ως προς την παράμετρο γ.

1.5.3 Εκτίμηση των παραμέτρων του μοντέλου για τη σύγκριση δύο γκρουπ ασθενών

Το μοντέλο αναλογικών κινδύνων που περιγράφεται από τη σχέση (1.47) μπορεί να εκτιμηθεί χρησιμοποιώντας τη μέθοδο της μέγιστης πιθανοφάνειας. Θα περιγράψουμε αναλυτικά αυτή τη διαδικασία, αφού αυτή θα χρησιμοποιήσουμε για την ανάλυση των δεδομένων στο τελευταίο κεφάλαιο της παρούσας εργασίας. Υποθέτουμε ότι οι χρόνοι επιβίωσης και των δύο γκρουπ ασθενών ακολουθούν εκθετική κατανομή. Υποθέτουμε επιπλέον ότι οι ασθενείς του γκρουπ 1 είναι n_1 και οι αντίστοιχες παρατηρήσεις τους συμβολίζονται ως (t_{i1}, δ_{i1}), $i = 1, 2, \ldots, n_1$, όπου η μεταβλητή δ_{i1} παίρνει την τιμή 0, εάν ο αντίστοιχος χρόνος επιβίωσης του i-οστού ασθενή είναι διακεκομμένος και την τιμή 1 διαφορετικά. Όμοια οι παρατηρήσεις των n_2 ασθενών του γκρουπ 2 συμβολίζονται ως $(t_{i'2}, \delta_{i2})$, $i' = 1, 2, \ldots, n_2$. Έστω ότι για τους ασθενείς του γκρουπ 1 η συνάρτηση κινδύνου παίρνει την τιμή λ και η συνάρτηση πυκνότητας πιθανότητας καθώς και η συνάρτηση επιβίωσης είναι

$$f(t_{i1}) = \lambda e^{-\lambda t_{i1}}, S(t_{i1}) = e^{-\lambda t_{i1}}. \tag{1.56}$$

Για τους ασθενείς του γκρουπ 2 η συνάρτηση κινδύνου παίρνει την τιμή $\psi\lambda$ και η συνάρτηση πυκνότητας πιθανότητας καθώς και η συνάρτηση επιβίωσης είναι

$$f(t_{i'2}) = \psi\lambda e^{-\psi\lambda t_{i'2}}, S(t_{i'2}) = e^{-\psi\lambda t_{i'2}}. \tag{1.57}$$

Σύμφωνα με τη σχέση (1.46), η συνάρτηση πιθανοφάνειας για τις n_1+n_2 παρατηρήσεις, $L(\psi, \lambda)$, γράφεται

$$\prod_{i=1}^{n_1} \left\{\lambda e^{-\lambda t_{i1}}\right\}^{\delta_{i1}} \left\{e^{-\lambda t_{i1}}\right\}^{1-\delta_{i1}} \prod_{i'=1}^{n_2} \left\{\psi\lambda e^{-\psi\lambda t_{i'2}}\right\}^{\delta_{i'2}} \left\{e^{-\psi\lambda t_{i'2}}\right\}^{1-\delta_{i'2}},$$

ή πιο απλά

$$\prod_{i=1}^{n_1} \lambda^{\delta_{i1}} e^{-\lambda t_{i1}} \prod_{i'=1}^{n_2} (\psi\lambda)^{\delta_{i'2}} e^{-\psi\lambda t_{i'2}}.$$

Εάν ο αριθμός των συμβάντων που παρατηρήθηκαν στα δύο γκρουπ είναι r_1 και r_2 αντίστοιχα, τότε $r_1 = \sum_i \delta_{i1}$ και $r_2 = \sum_{i'} \delta_{i'2}$ και ο λογάριθμος της συνάρτησης πιθανοφάνειας παίρνει τη μορφή

$$\log L(\psi, \lambda) = r_1 \log\lambda - \lambda \sum_{i=1}^{n_1} t_{i1} + r_2 \log(\psi\lambda) - \psi\lambda \sum_{i'=1}^{n_2} t_{i'2}.$$

Με T_1 και T_2 συμβολίζουμε το συνολικό γνωστό χρόνο επιβίωσης για τους ασθενείς του γκρουπ 1 και του γκρουπ 2 αντίστοιχα. Τότε, T_1 και T_2 είναι το σύνολο των χρόνων, διακεκομμένων και επιβίωσης, για το κάθε γκρουπ ασθενών, έτσι ώστε ο λογάριθμος της συνάρτησης πιθανοφάνειας γίνεται

$$\log L(\psi, \lambda) = (r_1 + r_1) \log\lambda + r_2 \log\psi - \lambda(T_1 + \psi T_2).$$

Προκειμένου να πετύχουμε τις εκτιμημένες τιμές $\hat{\psi}$, $\hat{\lambda}$ για τις οποίες η παραπάνω συνάρτηση γίνεται μέγιστη, την παραγωγίζουμε ως προς ψ και λ και θέτουμε τις παραγώγους ίσες με μηδέν. Οι εξισώσεις που παίρνουμε είναι

$$\frac{r_2}{\hat{\psi}} - \hat{\lambda} T_2 = 0, \tag{1.58}$$

$$\frac{r_1 + r_2}{\hat{\lambda}} - (T_1 + \hat{\psi} T_2) = 0, \tag{1.59}$$

από τις οποίες προκύπτουν οι εκτιμήσεις

$$\hat{\lambda} = \frac{r_2}{\hat{\psi} T_2},$$

$$\hat{\psi} = \frac{r_2 T_1}{r_1 T_2} \tag{1.60}$$

και τελικά

$$\hat{\lambda} = \frac{r_1}{T_1}.$$

Οι δύο αυτοί εκτιμητές μπορούν να εξηγηθούν διαισθητικά. Η εκτιμημένη τιμή για την παράμετρο λ είναι το αντίστροφο του μέσου γνωστού χρόνου επιβίωσης για τους ασθενείς του γκρουπ 1, ενώ ο εκτιμημένος σχετικός κίνδυνος, $\hat{\psi}$, είναι ο λόγος του μέσου γνωστού χρόνου επιβίωσης για τους ασθενείς του γκρουπ 1 προς το μέσο γνωστό χρόνο επιβίωσης για τους ασθενείς του γκρουπ 2.

Ο ασυμπτωτικός πίνακας διασπορών και συνδιασπορών των εκτιμητών είναι ο αντίστροφος του πληροφοριακού πίνακα του οποίου τα στοιχεία του βρίσκονται από τις δεύτερες παραγώγους του λογάριθμου της συνάρτησης πιθανοφάνειας. Ισχύει ότι

$$\frac{\mathrm{d}\log L(\psi,\lambda)}{\mathrm{d}\psi^2} = -\frac{r_2}{\psi^2}, \frac{\mathrm{d}\log L(\psi,\lambda)}{\mathrm{d}\lambda^2} = -\frac{r_1+r_2}{\lambda^2}, \frac{\mathrm{d}\log L(\psi,\lambda)}{\mathrm{d}\lambda\,\mathrm{d}\psi} = -T_2.$$

Ο πληροφοριακός πίνακας προκύπτει από τις αρνητικές αναμενόμενες τιμές αυτών των μερικών παραγώγων

$$\boldsymbol{I}(\psi,\lambda) = \begin{pmatrix} r_2/\psi^2 & T_2 \\ T_2 & r_1 + r_2/\lambda^2 \end{pmatrix}$$

και ο αντίστροφος αυτού του πίνακα είναι

$$\frac{1}{(r_1+r_2)r_2 - T_2^2\psi^2\lambda^2} \begin{pmatrix} (r_1+r_2)\psi^2 & -T_2\psi^2\lambda^2 \\ -T_2\psi^2\lambda^2 & r_2\lambda^2 \end{pmatrix}.$$

Οι τυπικές αποκλίσεις των ψ και λ βρίσκονται εάν αντικαταστήσουμε στον παραπάνω πίνακα, τις πραμέτρους λ και ψ από τις εκτιμήσεις τους

και στη συνέχεια πάρουμε τις τετραγωνικές ρίζες αυτών. Έτσι, η τυπική απόκλιση του $\hat{\psi}$ είναι

$$se(\hat{\psi}) = \sqrt{\frac{(r_1 + r_2)\hat{\psi}^2}{(r_1 + r_2)r_2 - T_2^2\hat{\psi}^2\hat{\lambda}^2}}.$$

Αντικαθιστώντας σε αυτή τη σχέση τους εκτιμητές $\hat{\psi}$ και $\hat{\lambda}$ από τις εκφράσεις τους, προκύπτει ότι

$$se(\hat{\psi}) = \hat{\psi}\sqrt{\frac{r_1 + r_2}{r_1 r_2}}. \tag{1.61}$$

Όμοια η τυπική απόκλιση του $\hat{\lambda}$ είναι

$$se(\hat{\lambda}) = \hat{\lambda}\sqrt{r_1}.$$

Οι τυπικές αποκλίσεις αυτών των εκτιμητών δε μπορούν να χρησιμοποιηθούν άμεσα στην κατασκευή διαστημάτων εμπιστοσύνης των παραμέτρων ψ και λ, αφού και οι δύο παράμετροι πρέπει να παίρνουν θετικές τιμές και οι εκτιμήσεις τους τείνουν να έχουν ασύμμετρες κατανομές. Κατ΄ επέκταση ο ισχυρισμός της κανονικότητας, που συνήθως χρησιμοποιούμε για την κατασκευή διαστημάτων εμπιστοσύνης, δεν μπορεί να προϋποτεθεί. Η κατανομή για το λογάριθμο των εκτιμητών του ψ ή του λ είναι πολύ πιθανό να είναι συμμετρική, οπότε χρησιμοποιώντας την τυπική απόκλιση του λογάριθμου της εκτιμημένης παραμέτρου, υπολογίζουμε το διάστημα εμπιστοσύνης για το λογάριθμο της παραμέτρου. Στη συνέχεια, υψώνοντας στην κατάλληλη δύναμη παίρνουμε το διάστημα εμπιστοσύνης για την καθεαυτή παράμετρο. Η τυπική απόκλιση του λογαρίθμου της εκτιμημένης παραμέτρου βρίκεται χρησιμοποιώντας το γενικό αποτέλεσμα

$$var\{g(\hat{\lambda})\} \approx \left\{\frac{\mathrm{d}g(\hat{\lambda})}{\mathrm{d}\hat{\lambda}}\right\}^2 var(\hat{\lambda}), \tag{1.62}$$

δηλαδή

$$var(\log\hat{\psi}) \approx \hat{\psi}^{-2} var(\hat{\psi}).$$

Η σχέση (1.62) είναι άμεση απόρροια της *προσεγγιστικής σειράς Taylor*

$$var\{g(X)\} \approx \left\{\frac{\mathrm{d}g(X)}{\mathrm{d}X}\right\}^2 var(X),$$

που ισχύει για τη διασπορά μιας συνάρτησης μιας τυχαίας μεταβλητής X. Επομένως, η τυπική απόκλιση του $\log\hat{\psi}$ είναι

$$se(\log\hat{\psi}) = \hat{\psi}^{-1} se(\hat{\psi}) = \sqrt{\frac{r_1 + r_2}{r_1 r_2}}. \tag{1.63}$$

Τα όρια ενός $100(1-a)\%$ διαστήματος εμπιστοσύνης για το λογάριθμο του σχετικού κινδύνου είναι $\log\hat{\psi} \pm z_{a/2} se(\log\hat{\psi})$, από τα οποία εύκολα βρίσκονται και τα όρια του διαστήματος εμπιστοσύνης γαι το σχετικό κίνδυνο $\hat{\psi}$. Εάν απαιτηθεί, με παρόμοιο τρόπο βρίσκουμε το διάστημα εμπιστοσύνης για την παράμετρο λ.

1.5.4 Διερεύνηση της καταλληλότητας του παραμετρικού μοντέλου

Όπως και για το μοντέλο των αναλογικών κινδύνων του Cox έτσι και για το οποιοδήποτε παραμετρικό μοντέλο, καλό θα ήταν προτού ξεκινήσουμε την ανάλυση των δεδομένων να εξετάσουμε κατά πόσο η κατανομή που υποθέτουμε για τους χρόνους επιβίωσης είναι αληθοφανής. Ένας απλός έλεγχος που μπορούμε να κάνουμε είναι να εκτιμήσουμε τη συνάρτηση κινδύνου με μία από τις γνωστές μη παραμετρικές μεθόδους εκτίμησης που αναφέραμε στην αρχή του κεφαλαίου και από τη γραφική της απεικόνιση να εξάγουμε τα ανάλογα συμπεράσματα. Για παράδειγμα, εάν η συνάρτηση

κινδύνου προσομοιάζει με μια σταθερή συνάρτηση, τότε η υιοθέτηση της εκθετικής κατανομής για τους χρόνους επιβίωσης είναι μάλλον η πιο κατάλληλη επιλογή. Αντίθετα, εάν η γραφική της απεικόνιση είναι μονότονα αύξουσα ή φθίνουσα τότε η υιοθέτηση της κατανομής Weibull είναι μάλλον η πιο κατάλληλη επιλογή.

Ένας περισσότερο πληροφοριακός έλεγχος που μπορούμε να κάνουμε είναι να συγκρίνουμε τη συνάρτηση επιβίωσης των χρόνων επιβίωσης στο δείγμα με τη συνάρτηση επιβίωσης στο επιλεγμένο μοντέλο. Έστω ότι προβλέπεται ότι οι χρόνοι επιβίωσης στο δείγμα ακολουθούν την κατανομή Weibull(λ, γ). Τότε η συνάρτηση επιβίωσης όπως έχουμε ήδη δείξει είναι

$$S(t) = \exp\big\{ -\lambda t^{\gamma} \big\}$$

Παίρνοντας τον λογάριθμο αυτής, πολλαπλασιάζοντας με -1 και ξαναπαίρνοντας τον λογάριθμο προκύπτει

$$\log\big\{ -\log S(t) \big\} = \log\lambda + \gamma \log t \tag{1.64}$$

Στη σχέση (1.64) αντικαθιστούμε την συνάρτηση επιβίωσης με την Kaplan–Meier εκτίμησή της. Εάν ο ισχυρισμός μας περί της κατανομής Weibull είναι ορθός, τότε η εκτιμήσεις των τιμών της συνάρτησης επιβίωσης θα είναι πολύ κοντά στην παρατηρούμενες τιμές της και επομένως η γραφική παράσταση της $\log\big\{ -\log \hat{S}(t) \big\}$ επί της $\log t$ θα είναι μια κατά προσέγγιση ευθεία γραμμή. Επειδή, από τον ορισμό της αθροιστικής συνάρτησης κινδύνου ισχύει $H(t) = -\log S(t)$, το γράφημα που προκύπτει ονομάζεται γράφημα του λογάριθμου των αθροιστικών κινδύνων.

Στην περίπτωση που η απεικόνιση του γραφήματος είναι μια ευθεία γραμμή, ο σταθερός όρος της θεωρείται ως μια εκτίμηση για τον λογάριθμο της

παραμέτρου κλίμακας, λ, και η κλίση της ως μια εκτίμηση για την παράμετρο σχήματος, γ. Αξίζει να σημειώσουμε ότι αν η κλίση της ευθείας είναι κοντά στην μονάδα, τότε η εκθετική κατανομή είναι κατάλληλη για την περιγραφή των χρόνων επιβίωσης στο δείγμα. Επομένως, όταν οι χρόνοι επιβίωσης ακολουθούν την κατανομή Weibull(λ, γ) τότε το γράφημα του λογάριθμου των αθροιστικών κινδύνων είναι μια ευθεία με κλίση γ και σταθερό όρο $\log \lambda$.

Ας υποθέσουμε ένα σύνολο δεδομένων επιβίωσης με δύο γκρουπ ασθενών και έστω ότι χρόνοι επιβίωσης των ασθενών που προέρχονται από το πρώτο γκρουπ ακολουθούν την κατανομή Weibull(λ, γ) με το γράφημα του λογάριθμου των αθροιστικών κινδύνων να είναι μια ευθεία με κλίση γ και σταθερό όρο $\log \lambda$. Τότε, όπως έχουμε δείξει κάτω από το μοντέλο αναλογικών κινδύνων, οι χρόνοι επιβίωσης στο δεύτερο γκρουπ ακολουθούν την κατανομή Weibull$(\psi\lambda, \gamma)$ και επιπλέον το γράφημα του λογάριθμου των αθροιστικών κινδύνων είναι μια ευθεία με κλίση γ και σταθερό όρο $\log \psi\lambda$ ή $\log \psi + \log \lambda$. Εάν από το γράφημα με τις εκτιμήσεις του λογάριθμου των αθροιστικών συναρτήσεων κινδύνου επί του λογάριθμου των χρόνων επιβίωσης και για τα δύο γκρουπ προκύψουν παράλληλες ευθείες, τότε κατάλληλο για την ανάλυση είναι το παραμετρικό μοντέλο αναλογικών κινδύνων με χρόνους επιβίωσης που ακολουθούν την κατανομή Weibull.

Τέλος, τα υπόλοιπα Cox-Snell, θεωρούνται πολύ χρήσιμα ως προς τον έλεγχο της καταλληλότητας των παραμετρικών μοντέλων. Έστω για παράδειγμα ότι έχουμε θεωρήσει το μοντέλο αναλογικών κινδύνων με χρόνους επιβίωσης που ακολουθούν την κατανομή Weibull. Αν έχουμε n παρατηρήσεις, τότε το υπόλοιπο Cox-Snell για την i-οστή, μη διακεκομμένη

παρατήρηση, είναι

$$r_{Ci} = \hat{H}_i(t_i) = -\log \hat{S}_i(t_i), \tag{1.65}$$

οπότε για το συγκεκριμένο μοντέλο που θεωρήσαμε το i-οστό υπόλοιπο Cox-Snell είναι

$$r_{Ci} = \exp\left(\hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_p x_{pi}\right)\hat{\lambda} t^{\hat{\gamma}}. \tag{1.66}$$

Αν το μοντέλο που υποθέσαμε είναι κατάλληλο, τα υπόλοιπα Cox-Snell θα ακολουθούν την εκθετική κατανομή με παράμετρο 1.

Το μοντέλο αναλογικών κινδύνων όπως ήδη έχουμε αναφέρει είναι αρκετά ελκυστικό για την ανάλυση δεδομένων επιβίωσης εξαιτίας του γεγονότος ότι δεν είναι απαραίτητο να να υιοθετήσουμε μια συγκεκριμένη κατανομή πιθανότητας για τους παρατηρούμενους χρόνους επιβίωσης. Παρόλα αυτά, όταν υποθέτουμε κατανομή Weibull για τους χρόνους επιβίωσης η παραμετρική μορφή του μοντέλου αναλογικών κινδύνων αποτελεί μια κατάλληλη βάση προκειμένου να μοντελοποιήσουμε τα δεδομένα.

Κεφάλαιο 2

Ελλιπή δεδομένα

Πολύ συχνά στην ανάλυση δεδομένων επιβίωσης δεν είναι δυνατή η καταγραφή των χρόνων επιβίωσης για όλους τους ασθενείς. Δεν είναι σπάνιες οι φορές που φτάνει ο προκαθορισμένος χρόνος λήξης της μελέτης και σε κάποιους από τους ασθενείς που μελετάμε δεν έχει εκδηλωθεί ακόμα το συμβάν που μας ενδιαφέρει. Επίσης, πολλές φορές κάποιοι ασθενείς, είτε ηθελημένα είτε συμπτωματικά, παύουν να συμμετέχουν στη μελέτη και έτσι το μόνο που γνωρίζουμε γι αυτούς είναι μέχρι ποια χρονική στιγμή δεν έχουν εκδηλώσει το γεγονός. Τα παραπάνω φαινόμενα διακόπτουν κατά μία έννοια τη διαδικασία παρακολούθησης των ασθενών κατά τη διάρκεια μιας μελέτης, γι΄ αυτό και οι παρατηρήσεις που προκύπτουν με αυτόν τον τρόπο ονομάζονται διακεκομμένες. Ένα άλλο είδος παρατηρήσεων που συναντάμε στα δεδομένα επιβίωσης είναι οι περικομμένες παρατηρήσεις, οι οποίες δεν είναι ούτε καν εν μέρει διαθέσιμες στον ερευνητή. Συχνά στη στατιστική ανάλυση δεδομένων οι ερευνητές έρχονται αντιμέτωποι με σετ δεδομένων τα οποία εμπεριέχουν και παρατηρήσεις που λείπουν από το δείγμα, γνωστές

ως ελλιπή δεδομένα.

2.1 Διακεκομμένες παρατηρήσεις

Στην ανάλυση δεδομένων επιβίωσης, οι διακεκομμένες παρατηρήσεις γνωστές και ως διακεκομμένοι χρόνοι επιβίωσης συναντώνται όταν το μόνο που είναι γνωστό γι αυτούς είναι, είτε η αρχή τους είτε το τέλος τους, μέσα σε κάποιο συγκεκριμένο χρονικό διάστημα με αποτέλεσμα ο ερευνητής να μην μπορεί να καταγράψει το ζητούμενο χρόνο επιβίωσης (από τη συγκεκριμένη χρονική αφετηρία έως την εκδήλωση του συμβάντος). Το ερώτημα που προκύπτει είναι πώς αντιμετωπίζονται αυτές οι παρατηρήσεις από τον ερευνητή, δηλαδή πως μπορούν να ενσωματωθούν στη μελέτη και φυσικά αν πρέπει να ενσωματωθούν. Προτού απαντήσουμε σε αυτά τα ερωτήματα καλό θα ήταν να περιγράψουμε με λεπτομέρεια τα είδη των διακεκομμένων παρατηρήσεων που υπάρχουν στη βιβλιογραφία.

2.1.1 Δεξιά διακεκομμένες παρατηρήσεις

Οι δεξιά διακεκομμένες παρατηρήσεις ενός δείγματος εμφανίζονται όταν σε κάποιους ασθενείς δεν έχει εκδηλωθεί το γεγονός που μας ενδιαφέρει μέχρι και τη στιγμή του τελευταίου γνωστού χρόνου παρατήρησής τους. Αυτό σημαίνει ότι ο πραγματικός χρόνος επιβίωσης αυτών των ασθενών (ο χρόνος δηλαδή από την προκαθορισμένη χρονική αφετηρία της μελέτης μέχρι την εκδήλωση του συμβάντος) υπερβαίνει το χρόνο παρατήρησης τους, γι αυτό και ονομάζονται δεξιά διακεκομμένες. Αυτές οι παρατηρήσεις εμφανίζονται πολύ συχνά στη μελέτες δεδομένων επιβίωσης.

Ας αναλογιστούμε για παράδειγμα πόσο σύνηθες είναι να χάνονται τα ίχνη ορισμένων ασθενών κατά τη διάρκεια μιας μελέτης επειδή για κάποιο λόγο οι ασθενείς αυτοί σταμάτησαν να επικοινωνούν με τον ερευνητή. Οι λόγοι που συντελούν στην κατάσταση αυτή φυσικά ποικίλλουν από ασθενή σε ασθενή. Έτσι, κάποιος μπορεί απλά να μετακόμισε, κάποιος να είχε υπερβολικό φόρτο εργασίας και να το αμέλησε, κάποιος να αποφάσισε ότι δεν αξίζει τον κόπο η ταλαιπωρία της παρακολούθησης κτλ. Ας αναλογιστούμε επίσης την περίπτωση που φτάνει η στιγμή της ανάλυσης των δεδομένων, δηλαδή το τέλος της παρακολούθησης των ασθενών και σε πολλούς από αυτούς δεν έχει εκδηλωθεί ακόμα το υπό μελέτη γεγονός. Σε όλες τις παραπάνω εκδοχές το μόνο που γνωρίζουμε είναι ότι ο χρόνος επιβίωσης του ασθενή είναι μεγαλύτερος από το χρονικό διάστημα με άκρα την αφετηρία της μελέτης και τη στιγμή εμφάνισης της διακεκομμένης παρατήρησης. Με άλλα λόγια, το υπό μελέτη γεγονός πρόκειται να εκδηλωθεί αφού ο ασθενής θα έχει πάψει να παρακολουθείται από τον ερευνητή. Αυτό το είδος των διακεκομμένων παρατηρήσεων, είναι το πιο συνηθισμένο στην ανάλυση δεδομένων επιβίωσης γι αυτό και χρήζει περεταίρω διερεύνησης.

2.1.2 Αριστερά διακεκομμένες παρατηρήσεις

Οι αριστερά διακεκομμένες παρατηρήσεις ενός δείγματος εμφανίζονται όταν σε κάποιους ασθενείς έχει ήδη εκδηλωθεί το γεγονός που μας ενδιαφέρει προτού οι ασθενείς αυτοί εισαχθούν στη μελέτη, δηλαδή πριν την έναρξη της περιόδου παρακολούθησης. Αυτό σημαίνει ότι ο πραγματικός χρόνος επιβίωσης αυτών των ασθενών (ο χρόνος δηλαδή από την πρώτη εκδήλωση του συμβάντος μέχρι την επόμενη εκδήλωση του συμβάντος, ε-

άν αυτή παρατηρηθεί) υπολείπεται του χρόνου παρατήρησής τους, γι αυτό και ονομάζονται αριστερά διακεκομμένες. Αυτές οι παρατηρήσεις εμφανίζονται λιγότερο συχνά στη μελέτες δεδομένων επιβίωσης από ό,τι οι δεξιά διακεκομμένες παρατηρήσεις.

Ας αναλογιστούμε ωστόσο, ένα παράδειγμα για να κατανοήσουμε καλύτερα τη φύση αυτών των παρατηρήσεων. Έστω λοιπόν, ότι το ενδιαφέρον μιας μελέτης επικεντρώνεται στο χρόνο που μεσολαβεί από μια χειρουργική επέμβαση για την αφαίρεση ενός κακοήθη όγκου που αρχικά είχε εντοπιστεί μέχρι τη στιγμή της ενδεχόμενης επανεμφάνισης του καρκινώματος. Ας υποθέσουμε ακόμη ότι η εξέταση για την εξακρίβωση της επανεμφάνισης προγραμματίζεται από τους θεράποντες ιατρούς να γίνει τρεις μήνες μετά τη χειρουργική επέμβαση. Είναι πιθανό σε κάποιους από τους ασθενείς στο δείγμα να εξακριβώθηκε η επανεμφάνιση του καρκινώματος. Για αυτούς τους ασθενείς ο χρόνος μέχρι την επανεμφάνιση, δηλαδή ο χρόνος επιβίωσης που μελετάμε, σαφώς υπολείπεται του χρονικού διαστήματος των τριών μηνών γι αυτό και λέμε ότι είναι αριστερά διακεκομμένος. Προφανώς σε ότι αφορά αυτό το συγκεκριμένο είδος των διακεκομμένων παρατηρήσεων, το γεγονός που μας ενδιαφέρει δεν είναι καταληκτικό για τον ασθενή οπότε και η διαδικασία παρακολούθησής του, παρότι έχει εκδηλωθεί το γεγονός, έχει ένα συγκεκριμένο κάθε φορά νόημα. Όπως προείπαμε οι αριστερά διακεκομμένες παρατηρήσεις δεν συναντώνται τόσο συχνά όσο οι δεξιά διακεκομμένες στην ανάλυση δεδομένων επιβίωσης, ωστόσο οι ερευνητές των κοινωνικών επιστημών έρχονται συχνά αντιμέτωποι μαζί τους.

2.1.3 Διακεκομμένες παρατηρήσεις εντός διαστήματος

Οι διακεκομμένες παρατηρήσεις εντός διαστήματος σε ένα δείγμα εμφανίζονται όταν σε κάποιους ασθενείς το γεγονός που μας ενδιαφέρει έχει εκδηλωθεί μεταξύ δύο χρονικών στιγμών, χωρίς όμως να γνωρίζουμε ακριβώς το πότε. Αυτό σημαίνει ότι δεν μπορούμε να υπολογίσουμε ακριβώς το χρόνο επιβίωσης αυτών των ασθενών (το χρόνο δηλαδή από την προκαθορισμένη χρονική αφετηρία της μελέτης μέχρι την εκδήλωση του συμβάντος), μπορούμε εντούτοις να προσδιορίσουμε το χρονικό πλαίσιο μέσα στο οποίο κυμαίνεται. Αυτές οι παρατηρήσεις εμφανίζονται συχνά σε μελέτες δεδομένων επιβίωσης στις οποίες οι ασθενείς παρατηρούνται σε χρονικά προκαθορισμένες επισκέψεις και το γεγονός που μας ενδιαφέρει έχει πιθανότατα συμβεί μεταξύ αυτών.

Ας αναλογιστούμε και πάλι το παράδειγμα που προαναφέραμε για να κατανοήσουμε καλύτερα τη φύση αυτών των παρατηρήσεων. Έστω λοιπόν, ότι το ενδιαφέρον μιας μελέτης επικεντρώνεται στο χρόνο που μεσολαβεί από μια χειρουργική επέμβαση για την αφαίρεση ενός κακοήθη όγκου που αρχικά είχε εντοπιστεί μέχρι τη στιγμή της ενδεχόμενης επανεμφάνισης του καρκινώματος. Εάν στην πρώτη επίσκεψη, δηλαδή τρεις μήνες μετά τη χειρουργική επέμβαση, σε κάποιους από τους ασθενείς δεν παρατηρήθηκε επανεμφάνιση του καρκινώματος ενώ, στη δεύτερη επίσκεψη, δηλαδή έξι μήνες μετά τη χειρουργική επέμβαση, σε κάποιους από αυτούς παρατηρήθηκε επανεμφάνιση, τότε ο χρόνος επιβίωσης αυτών των ασθενών (ο χρόνος δηλαδή από την προκαθορισμένη χρονική αφετηρία της μελέτης μέχρι την εκδήλωση του συμβάντος) είναι μεταξύ τριών και έξι μηνών. Γι αυτού τους

ασθενείς ο χρόνος μέχρι την επανεμφάνιση, δηλαδή ο χρόνος επιβίωσης που μελετάμε, κυμαίνεται μέσα σε ένα συγκεκριμένο χρονικό πλαίσιο, γι αυτό και λέμε ότι είναι διακεκομμένος εντός διαστήματος. Προφανώς σε ότι αφορά αυτό και αυτό το είδος των διακεκομμένων παρατηρήσεων, το γεγονός που μας ενδιαφέρει δεν είναι καταληκτικό για τον ασθενή αλλά συνήθως είναι η επανεμφάνιση κάποιου συμπτώματος μιας ασθένειας ή και της ίδιας της ασθένειας. Συναντάμε συχνά τις διακεκομμένες παρατηρήσεις εντός διαστήματος σε μελέτες διαχρονικών δεδομένων, δηλαδή σε δεδομένα που αφορούν επαναλαμβανόμενες μετρήσεις ανά τακτά χρονικά διαστήματα.

2.1.4 Περικομμένες παρατηρήσεις

Συχνά στην ανάλυση δεδομένων επιβίωσης εκτός από τις διακεκομμένες παρατηρήσεις συναντάμε και τις περικομμένες παρατηρήσεις και μάλιστα κάποιες φορές αυτές οι δύο ειδικές κατηγορίες παρατηρήσεων συγχέονται αφού και οι δεύτερες αποτελούν ελλιπή δεδομένα. Το φαινόμενο της περικοπής σε δεδομένα επιβίωσης συμβαίνει μόνο όταν οι χρόνοι επιβίωσης που παρατηρούνται εμπίπτουν σε ένα ορισμένο χρονικό διάστημα (Y_L, Y_R). Επομένως οι ασθενείς με χρόνο επιβίωσης που δεν εμπίπτει σε αυτό το διάστημα, με περικομμένους δηλαδή χρόνους επιβίωσης δεν καταγράφονται καθόλου στα δεδομένα. Το γεγονός αυτό διαφοροποιεί αυτές τις παρατηρήσεις από τις διακεκομμένες καθώς στις δεύτερες ο χρόνος επιβίωσης καταγράφεται μερικώς.

2.1.5 Αριστερά περικομμένες παρατηρήσεις

Όταν το Y_R είναι το άπειρο τότε εμφανίζονται οι αριστερά περικομμένες παρατηρήσεις. Στην περίπτωση αυτή παρατηρούμε μόνο εκείνους τους ασθενείς των οποίων ο χρόνος επιβίωσης υπερβαίνει το χρονική στιγμή Y_L, που ονομάζεται και χρόνος περικοπής. Δηλαδή καταγράφουμε το χρόνο επιβίωσης, T, μόνο εάν $Y_L < T$. Σε αυτό το είδος της περικομμένων παρατηρήσεων όλοι οι ασθενείς που εκδήλωσαν το υπό μελέτη γεγονός πριν από τον χρόνο της περικοπής απλά δεν μπορεί να καταγραφούν. Ο χρόνος της περικοπής ονομάζεται και καθυστερημένος χρόνος εισόδου αφού οι ασθενείς παρατηρούνται μόνο από αυτή τη στιγμή έως και τη στιγμή του συμβάντος ή τη στιγμή της διακοπής (διακεκομμένη παρατήρηση). Σε αντίθεση λοιπόν με τις αριστερά διακεκομμένες παρατηρήσεις που οι ασθενείς έχουν εκδηλώσει το γεγονός προτού εισαχθούν στη μελέτη αλλά παρατηρούνται στη συνέχεια, οι αριστερά περικομμένες παρατηρήσεις αντιστοιχούν σε ασθενείς που δεν συγκαταλέγονται καθόλου στη μελέτη.

Ας αναλογιστούμε το παράδειγμα που χρησιμοποιήσαμε παραπάνω (για τις αριστερά διακεκομμένες παρατηρήσεις) για να κατανοήσουμε καλύτερα τη φύση των αριστερά περικομμένων παρατηρήσεων. Έστω λοιπόν ότι το ενδιαφέρον μιας μελέτης επικεντρώνεται στο χρόνο που μεσολαβεί από μια χειρουργική επέμβαση για την αφαίρεση ενός κακοήθη όγκου που αρχικά είχε εντοπιστεί μέχρι τη στιγμή της ενδεχόμενης επανεμφάνισης του καρκινώματος. Κάποιοι από τους ασθενείς εξαιτίας της ταχύτατης εξέλιξης της νόσου ίσως να μην κατάφεραν τελικά να καταφύγουν στη χειρουργική επέμβαση οπότε ο ζητούμενος χρόνος δεν καταγράφηκε ποτέ γι αυτούς. Οι παρατηρήσεις αυτές θεωρούνται αριστερά περικομμένες και καλό θα ήταν

να ληφθούν υπόψη από τον αναλυτή.

2.1.6 Δεξιά περικομμένες παρατηρήσεις

Όταν το Y_L θεωρείται μηδέν τότε είναι πιθανό να εμφανιστούν οι δεξιά περικομμένες παρατηρήσεις. Στην περίπτωση αυτή παρατηρούμε μόνο εκείνους τους ασθενείς των οποίων ο χρόνος επιβίωσης υπολείπεται της χρονικής στιγμής Y_R, δηλαδή καταγράφουμε το χρόνο επιβίωσης μόνο όταν $T \leq Y_R$. Σε αυτό το είδος της περικομμένων παρατηρήσεων όλοι οι ασθενείς που εκδήλωσαν το υπό μελέτη γεγονός μετά από τη χρονική στιγμή Y_R δεν καταγράφονται.

Ας αναλογιστούμε ένα παράδειγμα για να κατανοήσουμε καλύτερα τη φύση των δεξιά περικομμένων παρατηρήσεων. Το ενδιαφέρον μιας πραγματικής μελέτης που έλαβε χώρα από το 1978 έως το 1986 επικεντρώνονταν στο χρόνο που μεσολάβησε από τη μόλυνση με τον ιό του AIDS μέσω μετάγγισης έως την έναρξη των κλινικών συμπτωμάτων της νόσου σε ένα δείγμα ασθενών. Οι ασθενείς μολύνθηκαν με τον ιό την 1η Ιουνίου 1986 και η καταληκτική ημερομηνία που ένας ασθενής μπορούσε να εγγραφεί στο δείγμα ήταν 30 Ιουνίου 1986, με συνέπεια μόνο όσοι ασθενείς είχαν ήδη εμφανίσει τα κλινικά συμπτώματα της νόσου πριν από αυτήν την ημερομηνία να εγγράφτηκαν στο παρατηρούμενο δείγμα. Όσοι ασθενείς μεταγγίστηκαν την 1η Ιουνίου 1986 και ανέπτυξαν τη νόσο μετά την 30η Ιουνίου 1986 δεν παρατηρήθηκαν και επομένως ο χρόνος επιβίωσης δεν καταγράφηκε γι αυτούς. Αυτό το είδος των παρατηρήσεων ονομάζονται δεξιά περικομμένες.

2.2 Μηχανισμοί που οδηγούν σε ελλιπή δεδομένα

Οι στατιστικοί αναλυτές έρχονται πολλές φορές αντιμέτωποι με δείγματα που περιέχουν ελλιπή δεδομένα, εν γένει. Οι λόγοι για τους οποίους οι συγκεκριμένες παρατηρήσεις δεν καταχωρήθηκαν συνολικά σε μια μελέτη είναι άλλες φορές γνωστοί και άλλες όχι. Η γνώση ή η απουσία γνώσης περί του μηχανισμού που οδηγεί σε κάποιες άγνωστες (ή εν μέρει γνωστές) τιμές δεδομένων είναι αυτό που καθορίζει πλήρως τη μέθοδο ανάλυσης που θα επιλέξει ο ερευνητής καθώς και την ερμηνεία των αποτελεσμάτων που θα προκύψουν.

Κάποιες φορές ο μηχανισμός που οδηγεί σε ελλιπή δεδομένα είναι απολύτως ελεγχόμενος από τον ερευνητή. Για παράδειγμα, σε δειγματοληπτικές έρευνες, ο μηχανισμός αυτός είναι απλά η διαδικασία επιλογής του δείγματος (κάποιες μονάδες του υπό μελέτη πληθυσμού επιλέγονται προς καταγραφή ενώ κάποιες άλλες όχι). Αν ο ερευνητής προχωρήσει στην ανάλυση με βάση ένα τυχαίο δείγμα του πληθυσμού (δεδομένου ότι αυτό μπορεί να υλοποιηθεί επιτυχώς), τότε θεωρούμε ότι ο μηχανισμός αυτός μπορεί απλά να αγνοηθεί. Σε άλλου είδους έρευνες ωστόσο, όπως και στην ανάλυση επιβίωσης, ο μηχανισμός που οδηγεί σε διακεκομμένες παρατηρήσεις δεν μπορεί να αγνοηθεί. Έτσι, η ανάλυση και η ερμηνεία των δεδομένων που ακολουθεί είναι απολύτως εξαρτημένη από τις παραδοχές που θα κάνουμε σχετικά με την προέλευση των ελλιπών παρατηρήσεων, γι αυτό και οι παραδοχές αυτές πρέπει να διατυπώνονται με σαφήνεια κάθε φορά.

2.2.1 Η προέλευση των ελλιπών δεδομένων

Υπάρχουν πολλοί λόγοι για τους οποίους δεδομένα μπορεί να λείπουν από μία μελέτη. Μπορεί απλά να υπήρξε δυσλειτουργεία στον ηλεκτρονικό εξοπλισμό, κάποιος από τους ερευνητές να αρρώστησε ή για οποιονδήποτε άλλο λόγο τα δεδομένα να μην καταχωρήθηκαν σωστά. Σε αυτήν την περίπτωση λέμε ότι τα δεδομένα είναι ένα εντελώς τυχαίο φαινόμενο(MCAR). Σε αυτήν την περίπτωση εννοούμε ότι η πιθανότητα να λείπει μία συγκεκριμένη παρατήρηση από το δείγμα δεν σχετίζεται με κανέναν τρόπο ούτε από την τιμή που θα έπαιρνε αυτή η παρατήρηση ούτε από καμία άλλη τιμή του δείγματος. Για παράδειγμα, σε μια έρευνα που μελετάται το εισόδημα των νοικοκυριών, τα δεδομένα δεν θα θεωρούνταν εντελώς τυχαίο φαινόμενο(MCAR) καθώς είναι πολύ πιθανό οι άνθρωποι με χαμηλά εισοδήματα να μην τα αναφέρουν σε σχέση με άλλους με υψηλότερα εισοδήματα. Όταν τα δεδομένα του δείγματος είναι εντελώς τυχαίο φαινόμενο, τότε το να λείπει μία παρατήρηση είναι εξίσου πιθανό με το να λείπει μία άλλη παρατήρηση. Επίσης, οι εκτιμήσεις που θα προκύψουν από την ανάλυση αυτών των δεδομένων θα είναι αμερόληπτες συγκριτικά με αυτές που θα προέκυπταν εάν δεν έλειπαν οι συγκεκριμένες τιμές.

Συχνά ένα σετ δεδομένων μπορεί να μην χαρακτηρίζεται ως εντελώς τυχαίο φαινόμενο(MCAR), ωστόσο να καταχωρείται απλά ως τυχαίο φαινόμενο(MAR). Σε αυτήν την περίπτωση εννοούμε ότι η πιθανότητα να λείπει μία συγκεκριμένη παρατήρηση από το δείγμα δεν σχετίζεται με την τιμή που θα έπαιρνε αυτή η παρατήρηση, αφού όμως ελέγξουμε πως μπορεί να σχετίζεται με τιμές άλλων μεταβλητών του δείγματος. Για παράδειγμα, άνθρωποι με συμπτώματα κατάθλιψης είναι πιθανό να μην έχουν την τάση

να δηλώνουν το εισόδημα τους με αποτέλεσμα το παρατηρούμενο εισόδημα να σχετίζεται με την κατάθλιψη. Οι ίδιοι άνθρωποι είναι πολύ πιθανό να έχουν όντως μικρότερο εισόδημα και έτσι να έχουμε υψηλά ποσοστά δεδομένων που λείπουν ανάμεσα σε ανθρώπους με συμπτώματα κατάθλιψης, με αποτέλεσμα το μέσο εισόδημα που θα προκύψεινα είναι μικρότερο από ό,τι θα ήταν εάν δεν υπήρχαν τα ελλιπή δεδομένα. Παρόλα αυτά, εάν η πθανότητα να δηλώθεί το εισόδημα από ανθρώπους που πάσχουν από συμπτωματα κατάθλιψης εξακολουθεί να μην σχετίζεται με το επίπεδο του εισοδήματος, τότε τα δεδομένα θα θεωρούνταν τυχαίο φαινόμενο(MAR), αλλά όχι εντελώς τυχαίο(MCAR).

Στην ανάλυση ενός δείγματος που είναι απλά τυχαίο φαινόμενο(MAR), μπορεί να παραπλανηθούμε καθώς η λέξη τυχαιότητα συνειρμικά μας φέρνει στο νου και αμεροληψία. Γενικά αυτό δεν ισχύει, αν και πολλές φορές βρίσκουμε τρόπους να παράγουμε στατιστικά σημαντικούς και σχετικά αμερόληπτους εκτιμητές. Όταν μια μεταβλητή που λείπει είναι τυχαίο φαινόμενο(MAR), ιδιαίτερη προσοχή πρέπει να δοθεί στην περαιτέρω ανάλυση.

Όταν τα δεδομένα που λείπουν δεν είναι ούτε τυχαίο φαινόμενο(MAR), ούτε εντελώς τυχαίο φαινόμενο(MCAR), τότε λέμε ότι δεν λείπουν τυχαία(MNAR). Για παράδειγμα, είναι πολύ πιθανό σε μία έρευνα που ξετάζουμε την ψυχική υγεία ασθενών, τα άτομα που έχουν διαγνωστεί με συμπτώματα κατάθλιψης να μην αναφέρουν το επίπεδο της κατάστασής τους. Σε μια τέτοια περίπτωση τα δεδομένα σίγουρα δεν λείπουν τυχαία. Ξεκάθαρα, το μέσο επίπεδο ψυχικής υγείας των ατόμων που συμμετέχουν στην έρευνα είναι μεροληπτικό σε σχέση με την εκτίμηση που θα παίρναμε εάν δεν υπήρχαν ελλιπείς παρατηρήσεις στο δείγμα. Το ίδιο συμβαίνει στις έρευνες

εισοδημάτων στις οποίες άνθρωποι με χαμηλά εισοδήματα είναι λιγότερο πιθανό να δηλώσουν το εισόδημα τους.

Όταν στο σετ δεδομένων που εξετάζουμε οι παρατηρήσεις που λείπουν δεν είναι τυχαίο φαινόμενο(MNAR), τότε η προσέγγιση που ακολουθούμε είναι να μοντελοποιήσουμε το μηχανισμό που οδηγεί σε ελλιπείς παρατηρήσεις. Με άλλα λόγια πρέπει να κατασκευάσουμε ένα μοντέλο στο οποίο να συνυπολογίζονται και οι παρατηρήσεις που λείπουν, μια διαδικασία αρκετά πολύπλοκη. Στην περίπτωση των διακεκομμένων παρατηρήσεων στην ανάλυση δεδομένων επιβίωσης, οι μηχανισμοί που οδηγούν σε αυτές είναι συνήθως μη ελεγχόμενοι από τον ερευνητή, εντούτοις κάποιες φορές είναι κατανοητοί. Για κάποιους ασθενείς στο δείγμα υπάρχει μερική πληροφόρηση ως προς τον χρόνο επιβίωσης τους, την οποία ο ερευνητής πρέπει να λάβει υπόψη ώστε να αποφύγει την εξαγωγή μεροληπτικών συμπερασμάτων.

2.2.2 Εκτίμηση δείγματος με ελλιπή δεδομένα βασισμένη στην πιθανοφάνεια

Η μέθοδος της εκτίμησης συναρτήσεων με τη μέθοδο της μέγιστης πιθανοφάνειας δεν διαφοροποιείται ουσιαστικά όταν εφαρμόζεται σε ένα δείγμα με ελλιπή δεδομένα. Ωστόσο τα συμπεράσματα είναι περισσότερο αμφισβητήσιμα αφού οι παρατηρήσεις δεν είναι εν γένει ανεξάρτητες και ταυτοτικά κατανεμημένες. Επίσης ορισμένα απλά συμπεράσματα που συνήθως προκύπτουν, όπως η κανονικότητα ενός μεγάλου δείγματος, δεν έχουν τόσο άμεση εφαρμογή. Ας μελετήσουμε θεωρητικά τι συμβαίνει με τη συνάρτηση πιθανοφάνειας σε ένα δείγμα με ελλιπή δεδομένα.

Έστω Y το σύνολο των δεδομένων συμπεριλαμβανομένων και των ελ-

λιπών. Γράφουμε $\boldsymbol{Y} = (\boldsymbol{Y_{obs}}, \boldsymbol{Y_{mis}})$, όπου με $\boldsymbol{Y_{obs}}$ συμβολίζουμε τις παρατηρούμενες τιμές, με $\boldsymbol{Y_{mis}}$ συμβολίζουμε τις ελλιπείς, με $f(\boldsymbol{Y}|\theta) \equiv f(\boldsymbol{Y_{obs}}, \boldsymbol{Y_{mis}}|\theta)$ συμβολίζουμε την συνάρτηση πυκνότητας πιθανότητας της από κοινού κατανομής των παρατηρούμενων και των ελλιπών τιμών και με θ την άγνωστη παράμετρο της κατανομής που έχουμε υποθέσει για το δείγμα. Εξ' ορισμού η περιθώρια συνάρτηση πυκνότητας πιθανότητας για τις παρατηρούμενες τιμές είναι

$$f(\boldsymbol{Y_{obs}}|\theta) = \int f(\boldsymbol{Y_{obs}}, \boldsymbol{Y_{mis}}|\theta)\, \mathrm{d}\boldsymbol{Y_{mis}}. \tag{2.1}$$

Ορίζουμε την συνάρτηση πιθανοφάνειας του θ η οποία βασίζεται στις παρατηρούμενες τιμές $\boldsymbol{Y_{obs}}$, αγνοώντας τον μηχανισμό εμφάνισης ελλιπών δεδομένων, να είναι μία οποιαδήποτε συνάρτηση ανάλογη της $f(\boldsymbol{Y_{obs}}|\theta)$,

$$L(\theta|\boldsymbol{Y_{obs}}) \propto f(\boldsymbol{Y_{obs}}|\theta), \tag{2.2}$$

οπότε τα συμπεράσματα που θα προκύψουν για την άγνωστη παράμετρο θ βασίζονται στην παραπάνω πιθανοφάνεια, δεδομένου ότι μπορούμε να αγνοήσουμε το μηχανισμό που οδηγεί σε ελλιπή δεδομένα.

Στη γενική περίπτωση εισάγουμε στη διαδικασία και μία νέα μεταβλητή R μέσω της οποίας δηλώνουμε εάν μια τιμή παρατηρήθηκε ή όχι. Για παράδειγμα εάν $Y = (Y_{ij})$ είναι ένας $(n \times k)$ πίνακας με n παρατηρήσεις για k μεταβλητές, ορίζουμε τη μεταβλητή $R = (R_{ij})$ ως εξής

$$R_{ij} = \begin{cases} 1, & \text{αν } y_{ij} \text{ παρατηρήθηκε,} \\ 0, & \text{αν } y_{ij} \text{ δεν παρατηρήθηκε.} \end{cases}$$

Η από κοινού κατανομή των τυχαίων μεταβλητών R και Y είναι το γινόμενο της συνάρτησης κατανομής του Y και της δεσμευμένης κατανομής του R

δεδομένου του Y, δηλαδή

$$f(Y, R|\theta, \psi) = f(Y|\theta) f(R|Y, \psi). \tag{2.3}$$

Με ψ συμβολίζουμε την άγνωστη παράμετρο που φανερώνει την κατανομή του μηχανισμού που οδηγεί σε ελλιπή δεδομένα. Όταν αυτή η κατανομή είναι γνώστη το ψ δεν είναι απαραίτητο.

Το πραγματικά παρατηρούμενο δείγμα εμπεριέχει τις τιμές των μεταβλητών (Y_{obs}, R). Εξ΄ ορισμού η συνάρτηση κατανομής των παρατηρούμενων τιμών Y_{obs} είναι

$$f(Y_{obs}, R|\theta, \psi) = \int f(Y_{obs}, \boldsymbol{Y_{mis}}|\theta) f(R|\boldsymbol{Y_{obs}}, \boldsymbol{Y_{mis}}, \boldsymbol{\psi}) \, \mathrm{d}\boldsymbol{Y_{mis}}. \tag{2.4}$$

Η συνάρτηση πιθανοφάνειας των θ και ψ είναι οποιαδήποτε συνάρτηση των θ και ψ ανάλογη στην παραπάνω σχέση, δηλαδή

$$L(\theta, \psi|\boldsymbol{Y_{obs}}) \propto f(\boldsymbol{Y_{obs}}, R|\theta, \psi). \tag{2.5}$$

Το ερώτημα που προκύπτει είναι πότε τα συμπεράσματα θα βασίζονται στη σχέση (2.5) και πότε στην απλούστερη (2.2) με την οποία αγνοείται ο μηχανισμός που οδηγεί στις ελλιπείς παρατηρήσεις. Αξίζει να παρατηρήσουμε ότι όταν αυτός ο μηχανισμός δεν εξαρτάται από τις τιμές $\boldsymbol{Y_{mis}}$ τότε

$$f(R|\boldsymbol{Y_{obs}}, \boldsymbol{Y_{mis}}, \psi) = f(R|\boldsymbol{Y_{obs}}, \psi). \tag{2.6}$$

Ακολούθως από τη σχέση (2.4) ισχύει

$$f(Y_{obs}, R, \theta, \psi) = f(R|\boldsymbol{Y_{obs}}, \psi) \times \int f(\boldsymbol{Y_{obs}}, \boldsymbol{Y_{mis}}, \theta) \, \mathrm{d}\boldsymbol{Y_{mis}} \tag{2.7}$$

$$= f(R|\boldsymbol{Y_{obs}}, \psi) f(\boldsymbol{Y_{obs}}|\theta). \tag{2.8}$$

Σε πολλές πρακτικές εφαρμογές οι παράμετροι θ και ψ είναι απολύτως διακεκριμένες με την έννοια ότι ο από κοινού παραμετρικός χώρος τους είναι το γινόμενο των επιμέρους παραμετρικών χώρων του θ και ψ αντίστοιχα. Εάν αυτό πράγματι συμβαίνει τότε τα συμπεράσματα για το θ που βασίζονται στην πιθανοφάνεια $L(\theta, \psi|Y_{obs}, R)$ θα είναι τα ίδια που θα προκύψουν για το θ από την πιθανοφάνεια $L(\theta, |Y_{obs})$ που σημαίνει ότι η σχέση (2.6) ικανοποιείται και ο μηχανισμός που οδηγεί σε ελλιπή δεδομένα αγνοείται.

Παράδειγμα 2.1 Ας υποθέσουμε ότι έχουμε ένα δείγμα n χρόνων επιβίωσης που ακολουθούν την εκθετική κατανομή με παράμετρο λ ή ισοδύναμα την κατανομή Weibull με παραμέτρους λ και $\gamma = 1$. Το διάνυσμα των παρατηρούμενων τιμών, $\boldsymbol{Y_{obs}}$, είναι $\boldsymbol{Y_{obs}} = (y_1, \ldots, y_m)'$ και το διάνυσμα των τιμών που δεν παρατηρήθηκαν, $\boldsymbol{Y_{mis}}$, είναι $\boldsymbol{Y_{mis}} = (y_{m+1}, \ldots, y_n)'$. Τότε η συνάρτηση πιθανοφάνειας του δείγματος είναι

$$f(\boldsymbol{Y}|\lambda) = \lambda^{-n} \exp\left(-\sum_{i=1}^{n} \frac{y_i}{\lambda}\right) \tag{2.9}$$

Η συνάρτηση πιθανοφάνειας αγνοώντας το μηχανισμό που οδηγεί στα ελλιπή δεδομένα είναι ανάλογη της κατανομής των παρατηρούμενων τιμών $\boldsymbol{Y_{obs}}$ δεδομένης της παραμέτρου λ, που δίνεται από τη σχέση

$$f(\boldsymbol{Y_{obs}}|\lambda) = \lambda^{-m} \exp\left(-\sum_{i=1}^{m} \frac{y_i}{\lambda}\right). \tag{2.10}$$

Με το διάνυσμα $\boldsymbol{R} = (R_1, \ldots, R_n)'$, όπου $R_i = 1, i = 1, \ldots, m$ και $R_i = 0, i = m+1, \ldots, n$ συμβολίζουμε τη μεταβλητή μέσω της οποίας δηλώνουμε αν μια τιμή παρατηρήθηκε ή όχι. Υποθέτουμε ότι ο κάθε ασθενής παρατηρείται με πιθανότητα ψ, δηλαδή ο μηχανισμός που οδηγεί στα ελλιπή δεδομένα δεν εξαρτάται από τις τιμές $\boldsymbol{Y_{mis}}$ που λείπουν. Τότε

$$f(\boldsymbol{R}|\boldsymbol{Y}, \psi) = \psi^m (1-\psi)^{n-m} \tag{2.11}$$

και

$$f(\boldsymbol{Y_{obs}}, \boldsymbol{R}|\lambda, \psi) = \lambda^{-m} \exp\,(-\sum_{i=1}^{m} \frac{y_i}{\lambda})\psi^m(1-\psi)^{n-m}. \tag{2.12}$$

Στην περίπτωση που οι παραμετρικοί χώροι των λ και ψ δεν έχουν κοινά σημεία, η συμπερασματολογία για την παράμετρο λ βασίζεται στην $f(\boldsymbol{Y_{obs}})$ αγνοώντας τον μηχανισμό που οδηγεί στις μη παρατηρούμενες τιμές. Πιο συγκεκριμένα, η εκτίμηση μέγιστης πιθανοφάνειας για το λ είναι απλά ο δειγματικός μέσος των παρατηρούμενων χρόνων επιβίωσης, δηλαδή $\hat{\lambda} = \sum_{i=1}^{m} y_i/m$.

Ας υποθέσουμε τώρα ότι από κάποιο γνωστό χρόνο διακοπής, c, και μετά εμφανίζονται οι διακεκομμένοι χρόνοι επιβίωσης. Τότε

$$f(\boldsymbol{R}|\boldsymbol{Y}, \psi) = \prod_{i=1}^{n} f(R_i|y_i, \psi) \tag{2.13}$$

όπου

$$f(R_i|\boldsymbol{Y}, \psi) = \begin{cases} 1, & \text{αν } R_i = 1 \text{ και } y_i < c \text{ ή } R_i = 0 \text{ και } y_i < c \\ 0, & \text{διαφορετικά .} \end{cases}$$

Οπότε

$$f(\boldsymbol{Y_{obs}}, \boldsymbol{R}|\lambda) = \prod_{i=1}^{m} f(y_i, R_i|\lambda) \prod_{i=m+1}^{n} f(R_i|\lambda) \tag{2.14}$$

$$= \prod_{i=1}^{m} f(y_i|\lambda) f(R_i|\psi_i) \prod_{i=m+1}^{n} P(y_i > c|\lambda) \tag{2.15}$$

$$= \lambda^{-m} \exp\left\{ -\sum_{i=1}^{m} \frac{y_i}{\lambda} \right\} \exp\left\{ -\frac{(n-m)c}{\lambda} \right\} \tag{2.16}$$

Σε αυτήν την περίπτωση ο μηχανισμός που οδηγεί σε ελλιπή δεδομένα δεν μπορεί να αγνοηθεί και η σωστή συνάρτηση πιθανοφάνειας που μόλις υπολογίσαμε διαφέρει από αυτήν στη σχέση (2.10). Μάλιστα η εκτίμηση μέγιστης πιθανοφάνειας για το λ σε αυτήν την περίπτωση είναι

$\hat{\lambda} = \sum_{i=1}^{m} y_i + (n-m)c/m$, η οποία μπορεί να συγκριθεί με την εκτιμήτρια $\sum_{i=1}^{m} y_i/m$ που υπολογίσαμε προηγουμένα.

Κεφάλαιο 3

Παραμετρικά μοντέλα

Οι παραμετρικές μέθοδοι που χρησιμοποιούνται για την ανάλυση δεδομένων επιβίωσης ως επί των πλείστων, προυποθέτουν ότι ο μηχανισμός που οδηγεί σε δεξιά διακεκομμένες παρατηρήσεις αγνοείται από τον ερευνητή ως μη πληροφοριακός λόγω της τυχαιότητας του. Το γεγονός αυτό φαίνεται ξεκάθαρα από τη συνεισφορά των δεξιά διακεκομμένων παρατηρήσεων στη συνάρτηση πιθανοφάνειας, που δεν είναι άλλη από την πιθανότητα ο χρόνος επιβίωσης να ξεπερνά τον χρόνο διακοπής. Το γεγονός δηλαδή, ότι η διακεκομμένη παρατήρηση συνέβη τη στιγμή που συνέβει, δεν επιδρά με κάποιον τρόπο στην κατανομή του χρόνου επιβίωσης και επομένως δεν σχετίζεται με την συμπερασματολογία για την κατανομή αυτή. Η μεθοδολογία αυτή σε αρκετές πρακτικές εφαρμογές οδηγεί σε όχι και τόσο ακριβή συμπεράσματα ενώ, σε άλλες μπορεί να αποδειχθεί και παραπλανητική.

Σε μια κλινική δοκιμή για παράδειγμα, ένας ασθενής μπορεί να πάψει να συμμετέχει στην εν λόγω διαδικασία, είτε εξαιτίας της επιβάρυνσης της υγείας του, είτε εξαιτίας παρενεργειών από μια συγκεκριμένη φαρμακευτι-

κή αγωγή που ακολουθεί. Η διακοπή αυτή, τη στιγμή που συμβαίνει είναι πολύ πιθανό να είναι ενδεικτική ενός παρατηρούμενου χρόνου επιβίωσης ελλαττωμένου σε σχέση με αυτόν που θα καταγράφονταν σε περίπτωση μη διακοπής. Σε αντίθετη περίπτωση, που ένας ασθενής αισθάνθεί καλύτερα και επίσης πάψει να συμμετέχει στην εν λόγω διαδικασία είναι πολυ πιθανό, η διακοπή αυτή να επιφέρει το αντίστροφο αποτέλεσμα, δηλαδή ο παρατηρούμενος χρόνος επιβίωσης να είναι αυξημένος σε σχέση με αυτόν που θα καταγράφονταν σε περίπτωση μη διακοπής.

Στη βιβλιογραφία υπάρχουν πολλά παρόμοια παραδείγματα που καταδεικνύεται ότι ο ισχυρισμός περί της τυχαιότητας του μηχανισμού που οδηγεί σε διακεκομμένες παρατηρήσεις δεν ευσταθεί και έτσι χάνεται πολύτιμη πληροφορία ως προς την εκτίμηση της συνάρτησης επιβίωσης. Τα συμπεράσματα επομένως μιας ανάλυσης μπορεί να είναι σε μικρό ή μεγάλο βαθμό μεροληπτικά όταν αγνοείται η εξάρτηση που υπάρχει μετάξυ του παρατηρημένου χρόνου επιβίωσης και του διακεκομμένου χρόνου παρατήρησης.

3.1 Ένα παραμετρικό μοντέλο που επιτρέπει την εξάρτηση μεταξύ χρόνου επιβίωσης και χρόνου διακοπής

Με τη βοήθεια του μοντέλου που περιγράφεται σε αυτήν την παράγραφο, μελετάται η μεροληψία στα συμπεράσματα, που προκύπτει από την εξάρτηση μεταξύ του χρόνου επιβίωσης T και του χρόνου διακοπής C κατά τις καθιερωμένες μεθόδους ανάλυσης. Για κάθε ασθενή, θεωρούμε ότι υπάρχει ένας εν δυνάμει τυχαίος χρόνος διακοπής C και ένας εν δυνάμει τυχαίος

χρόνος επιβίωσης T. Ο μηχανισμός διακοπής δεν είναι πληροφοριακός όταν οι χρόνοι C και T είναι ανεξάρτητοι και οι αντίστοιχες παράμετροι των συναρτήσεων κατανομής είναι απολύτως διακεκριμένες. Παρατηρούμε το χρόνο $Y = min(T, C)$ και τη μεταβλητή δείκτη του μηχανισμού διακοπής I, που παίρνει την τιμή $I = 1$, όταν $T \leq C$ και την τιμή $I = 0$, όταν $T > C$. Δυστυχώς,το επίπεδο της εξάρτησης μεταξύ των μεταβλητών T και C δεν είναι δυνατόν να εκτιμηθεί, αφού η από κοινού κατανομή των μεταβλητών Y και I δεν επαρκεί για να προσδιοριστεί η από κοινού κατανομή των T και C.

Έτσι, με αφετηρία το γεγονός ότι το επίπεδο της εξάρτησης μεταξύ του χρόνου επιβίωσης και του χρόνου διακοπής δεν μπορεί να εκτιμηθεί, ώστε να προχωρήσουμε στην κατασκευή ενός συγκεκριμένου νέου μοντέλου για την περιγραφή των δεδομένων, καταφεύγουμε στην ανάλυση ευαισθησίας των καθιερωμένων παραμετρικών μοντέλων στην ανάλυση επιβίωσης. Το πρώτο βήμα προς αυτή την κατεύθυνση είναι να μοντελοποιήσουμε τη δεσμευμένη κατανομή του χρόνου διακοπής C, δεδομένης της τιμής του χρόνου επιβίωσης T (που πιθανόν να μην έχει παρατηρηθεί), ώστε να εισάγουμε τη σχέση μεταξύ του χρόνου επιβίωσης και του μηχανισμού διακοπής. Η ζητούμενη μοντελοποίηση επιτυγχάνεται επιτρέποντας στην παράμετρο της περιθώριας κατανομής του C να εξαρτάται από το T μέσω μιας συνάρτησης $B(t, \theta)$ που φανερώνει τη μεροληψία και μιας παραμέτρου δ που φανερώνει την ίδια την εξάρτηση. Ακολουθεί μια σχετικά απλή ανάλυση ευαισθησίας της παραμέτρου που μας ενδιαφέρει να εκτιμήσουμε, βασισμένη σε γραμμικές προσεγγίσεις για μικρές τιμές της παραμέτρου δ. Ουσιαστικά, η εκτίμηση μέγιστης πιθανοφάνειας της παραμέτρου που μας ενδιαφέρει, για μικρές

τιμές της παραμέτρου δ, προκύπτει να είναι σχεδόν ίση με την εκτίμηση που θα προέκυπτε από την καθιερωμένη ανάλυση αγνοώντας δηλαδή το μηχανισμό διακοπής, προσθέτοντας δ φορές έναν δείκτη ευαισθησίας που εξαρτάται κάθε φορά από το παρατηρούμενο μοτίβο των διακεκομμένων παρατηρήσεων.

Πριν περιγράψουμε αναλυτικά το παραμετρικό μοντέλο, ορίζουμε εκ νέου τις βασικές συναρτήσεις που περιγράφουν τα δεδομένα επιβίωσης προκειμένου να εντάξουμε και την παράμετρο που μας ενδιαφέρει να εκτιμήσουμε. Έστω $S_T(t,\theta)$ η συνάρτηση επιβίωσης που περιγράφει το χρόνο της τυχαίας μεταβλητής T και θ η παράμετρος που μας ενδιαφέρει να εκτιμήσουμε. Ισοδύναμα, η κατανομή της μεταβλητής T περιγράφεται από τις συναρτήσεις

$$f_T(t,\theta) = -\frac{\mathrm{d}S_T(t,\theta)}{\mathrm{d}t}, h_T(t,\theta) = -\frac{\mathrm{d}\log S_T(t,\theta)}{\mathrm{d}t}, H_T(t,\theta) = -\log S_T(t,\theta)$$

που είναι αντίστοιχα η συνάρτηση πυκνότητας πιθανότητας, η συνάρτηση κινδύνου και η αθροιστική συνάρτηση κινδύνου. Από τη συνάρτηση πυκνότητας πιθανότητας $f_T(t,\theta)$ προκύπτουν επίσης οι γνωστές συναρτήσεις

$$s_T(t,\theta) = \frac{\vartheta}{\vartheta\theta}\log f_T(t,\theta) \quad \text{και} \quad i_\theta = Var_T\big\{s_T(T,\theta)\big\}$$

που είναι αντίστοιχα η συνάρτηση score και η συνάρτηση πληροφορία. Όμοια έστω $S_C(c,\gamma)$ η συνάρτηση επιβίωσης που περιγράφει το χρόνο της τυχαίας μεταβλητής C με παράμετρο γ. Ακολούθως η κατανομή της μεταβλητής C, του χρόνου των διακεκομμένων παρατηρήσεων, περιγράφεται από τις συναρτήσεις

$$f_C(c,\gamma) = -\frac{\mathrm{d}S_C(c,\gamma)}{\mathrm{d}t}, h_C(c,\gamma) = -\frac{\mathrm{d}\log S_C(c,\gamma)}{\mathrm{d}t}, H_C(c,\gamma) = -\log S_C(c,\gamma)$$

που είναι αντίστοιχα η συνάρτηση πυκνότητας πιθανότητας, η συνάρτηση

κινδύνου και η αθροιστική συνάρτηση κινδύνου. Από τη συνάρτηση πυκνότητας πιθανότητας $f_C(c,\gamma)$ προκύπτουν επίσης οι γνωστές συναρτήσεις

$$s_C(c,\gamma) = \frac{\vartheta}{\vartheta\gamma}\log f_C(c,\gamma) \;\text{ και }\; i_\gamma = Var_C\{s_C(C,\gamma)\}$$

που είναι αντίστοιχα η συνάρτηση score και η συνάρτηση πληροφορία. Στην περαιτέρω ανάλυση θεωρούμε ότι οι παράμετροι γ και θ είναι εντελώς διακεκριμένες μεταξύ τους, με την έννοια ότι η μία δεν προσδίδει καμία πληροφόρηση για την άλλη, όπως επίσης και ότι η παράμετρος γ χρησιμοποιείται απλά ως βοηθητική για τους υπολογισμούς, αφού η παράμετρος που μας ενδιαφέρει είναι η παράμετρος της συνάρτησης του χρόνου επιβίωσης.

Η μέθοδος που χρησιμοποιούμε για την εξαγωγή των συμπερασμάτων βασίζεται στον ισχυρισμό ότι η δεσμευμένη κατανομή της μεταβλητής C δεδομένης της T, έχει ακριβώς την ίδια παραμετρική μορφή με την περιθώρια κατανομη $f_C(c,\gamma)$, με τη διαφορά ότι η παράμετρός της εξαρτάται από τη μεταβλητή T μέσω της συνάρτησης $B(t,\theta)$ και της προαναφερθείσας παραμέτρου δ. Αναλυτικά, η δεσμευμένη πυκνότητα έχει τη μορφή

$$P(C=c|T=t) = f_C\big(c, \gamma + \delta i_\gamma^{-\frac{1}{2}} B(t,\theta)\big). \tag{3.1}$$

Αξίζει να σημειωθεί ότι όταν η παράμετρος δ, η οποία θεωρείται ένα μέτρο μεγέθους της εξάρτησης των μεταβλητών C και T, είναι 0, τότε οι μεταβλητές T και C είναι ανεξάρτητες, με αποτέλεσμα ο μηχανισμός που οδηγεί σε διακεκομμένες παρατηρήσεις να αγνοείται στην περαταίρω ανάλυση. Χάριν απλότητας των συμβολισμών υποθέτουμε ότι οι παράμετροι θ και γ είναι μονοδιάστατες. Ωστόσο, στη γενική περίπτωση που αυτές είναι πολυδιάστατες τότε και η συνάρτηση $B(t,\theta)$ που φανερώνει τη μεροληψία είναι διάνυσμα στήλη με τόσες γραμμές όσες και οι διαστάσεις της παραμετρου γ.

Η συνάρτηση $f_C\big(c, \gamma + \delta i_\gamma^{-\frac{1}{2}} B(t,\theta)\big)$ είναι αμετάβλητη κάτω από αυθαίρετους γραμμικούς μετασχηματισμούς της συνάρτησης μεροληψίας, οπότε προτού προχωρήσουμε στην ανάλυση ευαισθησίας είναι αναγκαίο να βρούμε τους αντίστοιχους περιορισμούς θέσης και κλίμακας της για τη συνάρτηση $B(t,\theta)$. Αν επιτρέψουμε στην παράμετρο δ να παίρνει μικρές τιμές κοντά στο 0, εντός του διαστήματος $[-0.3, 0.3]$, οδηγούμαστε σε μια ανάλυση ευαισθησίας για την παράμετρο που μας ενδιαφέρει που βασίζεται σε απλές γραμμικές προσεγγίσεις. Έτσι με μια πρώτης τάξης προσέγγιση του αναπτύγματος *Taylor*, η από κοινού συνάρτηση πυκνότητας των μεταβλητών T και C είναι

$$f_{T,C}(t,c) = f_T(t,\theta) f_C(c, \gamma + \delta i_\gamma^{-\frac{1}{2}} B(t,\theta)) \tag{3.2}$$

$$\simeq f_T(t,\theta) f_C(c,\gamma)[1 + \delta i_\gamma^{-\frac{1}{2}} s_C(c,\gamma) B(t,\theta)], \tag{3.3}$$

αφού η σειρά *Taylor* για τη συνάρτηση $f_C(c, \gamma + \delta i_\gamma^{-\frac{1}{2}} B(t,\theta))$ γύρω από το σημείο (c,γ), για τους όρους πρώτης τάξης είναι

$$\begin{aligned} & f_C(c, \gamma + \delta i_\gamma^{-\frac{1}{2}} B(t,\theta)) \\ & \simeq f_C(c,\gamma) + (c-c)\frac{\vartheta}{\vartheta c} f_C(c,\gamma) + (\gamma + \delta i_\gamma^{-\frac{1}{2}} B(t,\theta) - \gamma)\frac{\vartheta}{\vartheta \gamma} f_C(c,\gamma) \\ & \simeq f_C(c,\gamma) + \delta i_\gamma^{-\frac{1}{2}} B(t,\theta)\frac{\vartheta}{\vartheta\gamma} f_C(c,\gamma) \end{aligned}$$

και

$$f_C(c,\gamma)\frac{\vartheta}{\vartheta\gamma}\log f_C(c,\gamma) = \frac{\vartheta}{\vartheta\gamma} f_C(c,\gamma)$$

ή

$$f_C(c,\gamma) s_C(c,\gamma) = \frac{\vartheta}{\vartheta\gamma} f_C(c,\gamma),$$

άρα

$$f_C(c, \gamma + \delta i_\gamma^{-\frac{1}{2}} B(t,\theta)) \simeq f_C(c,\gamma) + f_C(c,\gamma)\delta i_\gamma^{-\frac{1}{2}} s_C(c,\gamma) B(t,\theta).$$

Ολοκληρώνοντας τη σχέση (3.3) ως προς τη μεταβλητή T προκύπτει η περιθώρια πυκνότητα της μεταβλητής C

$$\int_0^\infty f_T(t,\theta) f_C(c,\gamma)[1+\delta i_\gamma^{-\frac{1}{2}} s_C(c,\gamma) B(t,\theta)]\,\mathrm{d}t \tag{3.4}$$

$$= f_C(c,\gamma) + \delta i_\gamma^{-\frac{1}{2}} s_C(c,\gamma) \int_0^\infty B(t,\theta) f_T(t,\theta)\,\mathrm{d}t. \tag{3.5}$$

Οπότε θεωρούμε ότι

$$E_T[B(T,\theta)] = \int_0^\infty B(t,\theta) f_T(t,\theta)\,\mathrm{d}t = 0, \tag{3.6}$$

ώστε η περιθώρια πυκνότητα της μεταβλητής C να είναι ίση με $f_C(c,\gamma)$ για τους όρους πρώτης τάξης της παραμέτρου δ. Ακόμα θέλουμε η συνάρτηση $B(t,\theta)$ να έχει πεπερασμένη διασπορά και έτσι χωρίς περιορισμό της γενικότητας υποθέτουμε ότι

$$Var_T[B(T,\theta)] = E_T[B^2(T,\theta)] = 1. \tag{3.7}$$

3.1.1 Ανάλυση ευαισθησίας

Υποθέτουμε ότι έχουμε ένα δείγμα από $i = 1, 2, \ldots, n$ ασθενείς, τους χρόνους $t_i = min(T, C)$ και τη μεταβλητή δείκτη του μηχανισμού διακοπής I, που ορίζεται να παίρνει την τιμή $I = 1$, όταν $T_i \leq C_i$ και την τιμή $I = 0$, όταν $T_i > C_i$. Τότε η συνάρτηση του λογάριθμου της πιθανοφάνειας του δείγματος είναι

$$L_\delta(\theta,\gamma) = \sum_{i=1}^{n} I_i log P(T = t_i \cap T < C) + (1 - I_i) log P(C = t_i \cap C < T). \tag{3.8}$$

Υπολογίζοντας κάθεμια από αυτές τις πιθανότητες, σύμφωνα με την (3.3) και επεκτείνοντας ως προς τους όρους πρώτης τάξης της παραμέτρου δ, βρίσκουμε

$$L_\delta(\theta,\gamma) \simeq$$
$$\simeq L_0(\theta,\gamma)+\delta i_\gamma^{-\frac{1}{2}} \sum_{i=1}^{n} \left\{ (1-I_i)\mu_i(t_i,\theta)s_C(t_i,\gamma) - I_i B(t_i,\theta)\frac{\vartheta H_C(t_i,\gamma)}{\vartheta\gamma} \right\}. \tag{3.9}$$

όπου

$$\mu_i(t_i,\theta) = \frac{\int_t^\infty B(u,\theta) f_T(u,\theta)\,\mathrm{d}u}{S(t,\theta)}$$

και

$$L_0(\theta,\gamma) = \sum_{i=1}^{n} \left\{ I_i log h_T(t_i,\theta) + (1-I_i) log h_C(t_i,\gamma) - H_T(t_i,\theta) - H_C(t_i,\gamma) \right\}. \tag{3.10}$$

Η σχέση (3.10) είναι ο λογάριθμος της συνάρτησης πιθανοφάνειας των παραμέτρων θ, γ στην περίπτωση που οι μεταβλητές T, C είναι ανεξάρτητες. Αξίζει να παρατηρήσουμε ότι είναι ένα άθροισμα από δύο μέρη. Το ένα από αυτά είναι η συνήθης μορφή του λογάριθμου της συνάρτησης πιθανοφάνειας για τις παρατηρούμενες τιμές τις μεταβλητής T, υποθέτοντας ότι αγνοούμε το μηχανισμό διακοπής και το άλλο ακριβώς το αντίστροφο, δηλαδή η συνήθης μορφή του λογάριθμου της συνάρτησης πιθανοφάνειας για τις παρατηρούμενες τιμές τις μεταβλητής C, υποθέτοντας ότι αγνοούμε το μηχανισμό εμφάνισης συμβάντων. Επιπλέον, αξίζει να σημειώσουμε ότι στη συνάρτηση πιθανοφάνειας (3.9) ο όρος που πολλαπλασιάζεται με την παράμετρο δ, εξαρτάται και από τις δύο παραμέτρους θ,γ, οπότε στις εφαρμογές αντικαθιστούμε με τις εκτιμήσεις αυτών.

Θεωρώντας σταθερή την παράμετρο δ, με $\hat{\theta}_\delta$ συμβολίζουμε την τιμή της παραμέτρου θ που μεγιστοποιεί τη συνάρτηση πιθανοφάνειας (3.9) και με $\hat{\theta}_0$ την τιμή της παραμέτρου θ που μεγιστοποιεί τη συνάρτηση πιθανοφάνειας (3.10). Στη συνέχεια παραγωγίζοντας την (3.9) ως προς θ προκύπτει

$$\hat{\theta}_\delta - \hat{\theta}_0 \simeq$$
$$\simeq \delta i_\gamma^{-\frac{1}{2}} \iota(\theta)^{-1} \sum_{i=1}^{n} \left\{ (1 - I_i) \frac{\vartheta \mu(t_i, \theta)}{\vartheta \theta} s_C(t_i, \gamma) - I_i \frac{\vartheta B(t_i, \theta)}{\vartheta \theta} \frac{\vartheta H_C(t_i, \gamma)}{\vartheta \gamma} \right\}, \tag{3.11}$$

όπου

$$\iota(\theta) = -\frac{\vartheta^2 L_0(\theta, \gamma)}{\vartheta \theta^2}$$

είναι η παρατηρούμενη πληροφορία για το θ, βασισμένη στο μέρος εκείνο της πιθανοφάνειας που αγνοείται. Οι ανάλογες προσεγγίσεις φυσικά μπορούν να τεκμηριωθούν και για τις εκτιμήσεις $\hat{\gamma}_\delta$ και $\hat{\gamma}_0$ της παραμέτρου γ.

3.1.2 Επιλογή της συνάρτησης μεροληψίας $B(t, \theta)$

Σύμφωνα με τον ορισμό του μοντέλου, η συνάρτηση μεροληψίας $B(t, \theta)$ επιλέγεται έτσι ώστε να αντανακλά τη μορφή της ενδεχόμενης εξάρτησης των μεταβλητών που μας ενδιαφέρουν. Για παράδειγμα, στην περίπτωση που η μεταβλητή C ακολουθεί μια συνεχή κατανομή με παράμετρο-ρυθμό γ, τότε η συνάρτηση μεροληψίας $B(t, \theta)$ επιλέγεται να είναι μια φθίνουσα ή αύξουσα συνάρτηση του t έτσι ώστε να είναι πληροφοριακή ως προς ποιες τιμές του t είναι πιο πιθανή η εμφάνιση των διακεκομμένων παρατηρήσεων. Γενικά, στην παρούσα εργασία θα χρησιμοποιήσουμε μια συγκεκριμένη μορφή για

τη συνάρτηση $B(t,\theta)$, σύμφωνα με Siannis et al. (2005), που βασίζεται στον παρακάτω ισχυρισμό.

Ας υποθέσουμε ότι για έναν συγκεκριμένο ασθενή οι χρόνοι επιβίωσης και διακοπής, T και C, είναι ανεξάρτητοι με αντίστοιχες συναρτήσεις πυκνότητας πιθανότητας

$$g_T(t,\theta+\epsilon_T\iota_\theta^{-\frac{1}{2}}) \quad \text{και} \quad g_C(c,\gamma+\epsilon_C\iota_\gamma^{-\frac{1}{2}})$$

όπου ϵ_T και ϵ_C είναι τυχαίες επιδράσεις με αντίστοιχες μέσες τιμές 0, διασπορές σ_T^2 και σ_C^2 και συνδιασπορά σ_{TC}.

Θεωρούμε ότι οι διασπορές σ_T^2, σ_C^2 και η συνδιασπορά σ_{TC} είναι πολύ μικρές και της ίδιας τάξης μεγέθους. Τότε η περιθώρια κατανομή της μεταβλητής T, για κάποιο t, είναι

$$f_T(t,\theta)=E\left\{g_T(t,\theta+\epsilon_T\iota_\theta^{-\frac{1}{2}})\right\}\simeq g_T(t,\theta)+\frac{\sigma_T^2}{2\iota_\theta}\frac{\vartheta^2 g_T(t,\theta)}{\vartheta\theta^2}$$

με αντίστοιχη σχέση να ισχύει για την περιθώρια κατανομή $f_C(c,\gamma)$της μεταβλητής C. Η από κοινού κατανομή των μεταβλητών T και C είναι

$$\begin{aligned} f_{T,C}(t,c) &= E\left\{g_T(t,\theta+\epsilon_T\iota_\theta^{-\frac{1}{2}})g_C(c,\gamma+\epsilon_C\iota_\gamma^{-\frac{1}{2}})\right\} \\ &\simeq f_T(t,\theta)f_C(c,\gamma)+\sigma_{TC}(\iota_\theta\iota_\gamma)^{-\frac{1}{2}}\frac{\vartheta g_T(t,\theta)}{\vartheta\theta}\frac{\vartheta g_C(c,\gamma)}{\vartheta\gamma} \\ &\simeq f_T(t,\theta)f_C(c,\gamma)[1+\sigma_{TC}(\iota_\theta\iota_\gamma)^{-\frac{1}{2}}s_T(t,\theta)s_C(c,\gamma)]. \end{aligned}$$

Αυτή η πιθανοφάνεια με τον ανάλογο κατάλληλο ορισμό για την παράμετρο δ, είναι ακριβώς ίδια με την πιθανοφάνεια που περιγράφεται από τη σχέση (3.3) όταν η συνάρτηση μεροληψίας $B(t,\theta)$ είναι ίση με την τυποποιημένη συνάρτηση score

$$B(t,\theta)=\iota_\theta^{-\frac{1}{2}}s_T(t,\theta). \tag{3.12}$$

3.1.3 Ανάλυση ευαισθησίας βασισμένη στο μοντέλο αναλογικών κινδύνων

Σε αυτήν την παράγραφο αλλά και στις πρακτικές εφαρμογές που θα αναλύσουμε στις επόμενες παραγράφους, εξετάζουμε την ειδική περίπτωση που οι κατανομές των μεταβλητών T και C ακολουθούν το πρότυπο των μοντέλων των αναλογικών κινδύνων, δηλαδή

$$h_T(t,\theta) = e^{\theta} h_T^*(t) \quad , \quad h_C(c,\gamma) = e^{\gamma} h_C^*(c), \tag{3.13}$$

με $h_T^*(t)$ και $h_T^*(t)$ να είναι γνωστές βασικές συναρτήσεις κινδύνων(για παράδειγμα, η συνάρτηση κινδύνου όταν η κατανομή του χρόνου επιβιωσης είναι η εκθετική). Στην περίπτωση αυτή, ο ρόλος των παραμέτρων θ και γ είναι να αυξάνουν ή να μειώνουν τον κίνδυνο εμφάνισης του συμβάντος και όχι να τροποποιούν τη μορφή του κινδύνου όσο περνάει ο χρόνος. Εάν ισχύει η (3.13), τότε

$$s_T(t,\theta) = 1 - H_T(t,\theta) \quad , \quad s_C(c,\gamma) = 1 - H_C(c,\gamma) \tag{3.14}$$

και

$$\iota_\theta = \iota_\gamma = 1. \tag{3.15}$$

Συνδυάζοντας τις σχέσεις (3.13),(3.14), (3.15) και (3.12) προκύπτει η εξής απλή μορφή για τη συνάρτηση $B(t,\theta)$

$$B(t,\theta) = 1 - H_T(t,\theta). \tag{3.16}$$

Η από κοινού κατανομή των μεταβλητών παίρνει έτσι τη νέα συμμετρική μορφή

$$f_{T,C}(t,c) \simeq f_T(t,\theta) f_C(c,\gamma)\big[1 + \delta[1 - H_C(c,\gamma)][1 - H_T(t,\theta)]\big]. \tag{3.17}$$

Ακόμα, σε αυτήν την περίπτωση ισχύει ότι $\mu(t,\theta) = -H_T(t,\theta)$, οπότε ο λογάριθμος της συνάρτησης πιθανοφάνειας παίρνει την εξής μορφή

$$\begin{aligned} &L_\delta(\theta,\gamma) \simeq \\ &\simeq L_0(\theta,\gamma)+\delta\iota_\gamma^{-\frac{1}{2}}\sum_{i=1}^{n}\Big\{H_T(t_i,\theta)H_C(t_i,\gamma)-I_iH_C(t_i,\gamma)-(1-I_i)H_T(t_i,\theta)\Big\}, \end{aligned} \tag{3.18}$$

συνεπώς μετά τις κατάλληλες παραγωγίσεις παίρνουμε την τελική εκτίμηση για την μεροληψία της παραμέτρου θ, σύμφωνα με τη υπόθεση της εξάρτησης των μεταβλητών T και C

$$\hat{\theta}_\delta - \hat{\theta}_0 \simeq \delta\iota(\theta)^{-1}\iota_\gamma^{-\frac{1}{2}}\sum_{i=1}^{n}\Big\{H_T(t_i,\theta)H_C(t_i,\gamma)-(1-I_i)H_T(t_i,\theta)\Big\}. \tag{3.19}$$

Η παραπάνω ανάλυση ευαισθησίας γενικεύεται στην περίπτωση που στο υπό μελέτη μοντέλο περιλαμβάνονται και ανεξάρτητες ερμηνευτικές μεταβλητές. Τότε, το μοντέλο αναλογικών κινδύνων της σχέσης (3.13) γράφεται

$$h_T(t,\theta,x) = e^{\theta^T x}h_T^*(t) \quad , \quad h_C(c,\gamma,x) = e^{\gamma^T x}h_C^*(c), \tag{3.20}$$

όπου $\boldsymbol{x}$ είναι το διάνυσμα των ανεξάρτητων ερμηνευτικών μεταβλητών και $\boldsymbol{\theta^T}$, $\boldsymbol{\gamma^T}$ είναι το ανάστροφο διάνυσμα της παραμέτρου θ και γ αντίστοιχα. Όπως και για την περίπτωση χωρίς ανεξάρτητες μεταβλητές, η μορφή της βασικής συναρτήσης κινδύνου είναι γνωστή και για τις δύο μεταβλητές T,C.

Χάριν απλότητας υποθέτουμε ότι η συνάρτηση μεροληψίας εξαρτάται από το διάνυσμα των μεταβλητών $\boldsymbol{x}$ μόνο μέσω του γραμμικού προγνωστικού παράγοντα $\theta^T x$ και ότι η παράμετρος δ που φανερώνει την εξάρτηση των μεταβλητών T και C, δεν εξαρτάται από το $\boldsymbol{x}$, ωστόσο καμία από αυτές τις υποθέσεις δεν είναι περιοριστική. Επεκτείνοντας την προαναφερθείσα

ανάλυση ευαισθησίας και θεωρώντας την (3.16), η από κοινού κατανομή των μεταβλητών T, C γίνεται

$$f_{T,C}(t,c,x) \simeq f_T(t,\theta,x) f_C(c,\gamma,x)\big\{1+\delta[1-H_C(c,\gamma,x)][1-H_T(t,\theta,x)]\big\}, \tag{3.21}$$

και η συνάρτηση του λογάριθμου της πιθανοφάνειας (3.18) αντίστοιχα γενικεύεται ως εξής

$$\begin{aligned} &L_\delta(\theta,\gamma) \simeq L_0(\theta,\gamma) \\ &+\delta\sum_{i=1}^{n}\Big\{H_T(t_i,\theta,x_i)H_C(t_i,\gamma,x_i)-I_iH_C(t_i,\gamma,xi)-(1-I_i)H_T(t_i,\theta,x_i)\Big\}. \end{aligned} \tag{3.22}$$

Η τελική εκτίμηση για την μεροληψία της παραμέτρου θ και για αυτό το μοντέλο, σύμφωνα με τη υπόθεση της εξάρτησης των μεταβλητών T και C, είναι

$$\hat{\theta}_\delta-\hat{\theta}_0 \simeq \delta\iota(\theta)^{-1}\sum_{i=1}^{n}\Big\{x_i\big[H_T(t_i,\theta,x_i)H_C(t_i,\gamma,x_i)-(1-I_i)H_T(t_i,\theta,x_i)\big]\Big\}. \tag{3.23}$$

3.1.4 Δείκτης ευαισθησίας και διαστήματα εμπιστοσύνης

Όπως φαίνεται από τις εξισώσεις (3.11) και (3.19), η εκτίμηση μέγιστης πιθανοφάνειας $\hat{\theta}_\delta$ της παραμέτρου θ, κάτω από την προτεινόμενη ανάλυση ευαισθησίας είναι σχεδόν ίση με την εκτίμηση μέγιστης πιθανοφάνειας $\hat{\theta}_\delta$ της παραμέτρου θ, κάτω από την καθιερωμένη ανάλυση, προσθέτοντας δ φορές έναν παράγοντα που μπορεί να επίσης να εκτιμήθει από την καθιερωμένη

ανάλυση, χωρίς δηλαδή να λάβουμε υπόψη την εξάρτηση των μηχανισμών των χρόνων επιβίωσης και χρόνων διακοπής

$$\hat{\theta}_\delta - \hat{\theta}_0 = \delta U + O(\delta^2). \tag{3.24}$$

Με U συμβολίζεται ο διορθωτικός παράγοντας πρώτης τάξης για τη μεροληψία και ονομάζεται δείκτης ευαισθησίας. Σε πολλές εφαρμογές, η παράμετρος θ είναι απλά ένας βολικός τρόπος να παραμετρικοποιούμε ερμηνεύσιμες ποσότητες όπως η διάμεσος των χρόνων επιβίωσης. Στη γενική περίπτωση που ενδιαφερόμαστε να εκτιμήσουμε μια συνάρτηση του θ, έστω τη $J(\theta)$, τότε η ανάλυση ευαισθησίας που προκύπτει για τη συνάρτηση αυτή είναι

$$J(\hat{\theta}_\delta) \simeq J(\hat{\theta}_0) + \delta J'(\hat{\theta}_0)U. \tag{3.25}$$

Επιπλέον, το ασυμπτωτικό διάστημα εμπιστοσύνης για την παράμετρο θ, όταν η παράμετρος εξάρτησης δ είναι γραμμικής μορφής είναι

$$\text{από } \hat{\theta}_0 - \delta|U| - z_{(a)}\{\iota(\theta)\}^{-\frac{1}{2}} \quad \text{έως} \quad \hat{\theta}_0 + \delta|U| + z_{(a)}\{\iota(\theta)\}^{-\frac{1}{2}},$$

όπου $z_{(a)}$ είναι το κατάλληλο ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής. Παρατηρούμε ότι το παραπάνω διάστημα εμπιστοσύνης είναι το καθιερωμένο διάστημα εμπιστοσύνης διορθωμένο κατά τη μεροληψία που φαίνεται στην (3.24) και είναι έγκυρο όταν η παράμετρος δ παίρνει μικρές τιμές. Στις εφαρμογές, προσδίδουμε στην παράμετρο δ αληθοφανείς τιμές εντός ενός διαστήματος $(-\delta, \delta)$ που επιθυμούμε να συμπεριλάβουμε στην ανάλυση.

3.1.5 Ερμηνεία της παραμέτρου εξάρτησης

Η εν λόγω ανάλυση επιβίωσης είναι ένα χρήσιμο εργαλείο στη στατιστική συμπερασματολογία όταν με βάση κάποιο κριτήριο μπορούμε να απο-

φανθούμε εάν η εξάρτηση των μεταβλητών T και C που φανερώνεται από την παράμετρο δ είναι μικρή ή μεγάλη. Ένα σύνηθες μέτρο για τη συνάφεια των δύο μεταβλητών είναι ο συντελεστής συσχέτισης μεταξύ T και C, που ωστόσο μπορεί να μην είναι πάντα το κατάλληλο αφού πολλές φορές η εξάρτηση δεν έχει γραμμική μορφή και οι κατανομές έχουν μεγάλο συντελεστή λοξότητας. Αντί να χρησιμοποιήσουμε ως μέτρο το συντελεστή συσχέτισης των δύο μεταβλητών, μπορούμε να χρησιμοποιήσουμε τη συσχέτιση μεταξύ δύο συναρτήσεων των δύο μεταβλητών, έστω της $A(T, \theta)$ και της $D(C, \gamma)$. Όταν η παράμετρος δ έχει γραμμική μορφή, τότε

$$|Corr(A(T, \theta)), D(C, \gamma))| \leq |\delta|,$$

μέσω της οποίας προκύπτει ότι η παράμετρος δ μπορεί να ερμηνευτεί ως ένα άνω όριο για όλες τις συσχετίσεις μεταξύ του μηχανισμού γέννησης των χρόνων επιβίωσης και του μηχανισμού γέννησης των χρόνων διακοπής.

3.2 Η στάθμιση των δεδομένων επιβίωσης μέσω αντίστροφων πιθανοτήτων

Η μέθοδος της στάθμισης των δεδομένων μέσω αντίστροφων πιθανοτήτων είναι ένας τρόπος να παίρνουμε εκτιμήσεις των ποσοτήτων που μας ενδιαφέρουν σε δεδομένα επιβίωσης που περιέχουν διακεκομμένες παρατηρήσεις. Στην περίπτωση που ο μηχανισμός διακοπής είναι τυχαίος, η μέθοδος αυτή δίνει τα ίδια απότελέσματα με αυτά που παίρνουμε από τη μέθοδο Kaplan–Meier. Η ιδέα της μεθόδου αυτής βασίζεται στον ισχυρισμό, ότι η καμπύλη θνησιμότητας, που είναι το συμπλήρωμα της καμπύλης επι-

βίωσης, δίνεται από το άθροισμα των ποσοστών των αθροιστικών κινδύνων εμφάνισης του συμβάντος με το πέρασμα του χρόνου.

Ας υποθέσουμε ότι το υπό μελέτη γεγονός συνέβη τη στιγμή t και ότι ακριβώς στους μισούς από τους ασθενείς στο δείγμα παρατηρήθηκαν μη πληροφοριακοί διακεκομμένοι χρόνοι επιβίωσης μέχρι εκείνη τη στιγμή. Αυτό σημαίνει ότι μόνο οι μισοί από τους ασθενείς εκτίθενται στον κίνδυνο εμφάνισης του συμβάντος από τη στιγμή t και μετά. Επομένως, κατά μέσο όρο θα αναμέναμε ότι μετά από ίσο χρόνικό διάστημα με αυτό που μεσολάβησε μέχρι τη στιγμή t, θα παρατηρηθεί για δεύτερη φορά το γεγονός και κανένας ασθενής δεν θα έχει μείνει πλέον στη μελέτη. Κατά συνέπεια και κατά μέσο όρο, θα αναμέναμε ο συνολικός αριθμός των συμβάντων στο δείγμα να είναι 2. Γενικά, έαν P είναι η αναλογία των ασθενών στο δείγμα που βρίσκονται σε κίνδυνο ακριβώς μετα την πρώτη στιγμή εμφάνισης του συμβάντος, τότε ο αναμενόμενος αριθμός των συμβάντων στο δείγμα, E, είναι $E = 1/P$. Στην αντίθετη περίπτωση που P είναι η αναλογία των ασθενών στο δείγμα που δεν εμφάνισαν διακεκομμένο χρόνο επιβίωσης μεχρι τη στιγμή του πρώτου συμβάντος και PC είναι η αναλογία των ασθενών που εμφάνισαν διακεκομμένο χρόνο επιβίωσης ($P + PC = 1$), τότε $E = 1/(1 - PC)$.

Για παράδειγμα, εάν τα ποσοστά του μηχανισμού διακοπής είναι μηδενικά, τότε $PC = 0$ και $E = 1/(1 - 0) = 1$), που σημαίνει ο αναμενόμενος συνολικός αριθμός συμβάντων στο δείγμα παραμένει στο 1. Όταν υπάρχει ήπιο ποσοστό μηχανισμού διακοπής, έστω $PC = 0.1$, τότε το 1 παρατηρημένο γεγονός αναμένεται να αυξηθεί έτσι ώστε ο αναμενόμενος συνολικός αριθμός συμβάντων στο δείγμα να γίνει $E = 1/0.9 \simeq 1.1$. Η μικρή αυτή

αύξηση οφείλεται στο γεγονός ότι παραμένει μόνο ένα μικρό ποσοστό ασθενών που μελλοντικά ενδεχομένως μπορεί να αυξήσει τα παρατηρημένα συμβάντα. Σε περίπτωση που υπάρχει μεγάλο ποσοστό μηχανισμού διακοπής, έστω $PC = 0.9$, τότε τότε το 1 παρατηρημένο γεγονός αναμένεται να αυξηθεί και να γίνει $E = 1/0.1 = 10$ παρατηρημένα γεγονότα. Όλα τα πατραπάνω παραδείγματα προυποθέτουν ομοιογένεια στον κίνδυνο εμφάνισης των συμβάντων. Εάν ο μηχανισμός διακοπής δεν είναι τυχαίος και κατ' επέκταση πληροφοριακός, τότε η αναλογία PC ποικίλλει από ασθενή σε ασθενή και θα πρέπει να εκτιμηθεί ατομικά σύμφωνα με τους παράγοντες που επηρεάζουν τον κίνδυνο εμφάνισης του συμβάντος.

3.2.1 Η μέθοδος της στάθμισης των δεδομένων

Στις αναλύσεις που χρησιμοποιείται η μέθοδος της στάθμισης των δεδομένων μέσω αντίστροφων πιθανοτήτων, κάθε ασθενής i σταθμίζεται με ένα βάρος w_i, όπου w_i είναι το αντίστροφο της πιθανότητας ο εν λόγω ασθενής να δέχεται την έκθεσή του σε μια μεταβλητή X(π.χ σε μια θεραπεία), δεδομένης της ύπαρξης ενός διανύσματος ανέξαρτητων ερμηνευτικών μεταβλητών $\boldsymbol{Z}$ που ενδεχομένως τον επηρεάζουν. Ειδικότερα, $w_i = 1/f(X|Z)$, όπου $f(X|Z)$ είναι η δεσμευμένη συνάρτηση πυκνότητας της μεταβλητής X δεδομένων των παρατηρημένων τιμών των ανεξάρτητων μεταβλητών. Καθώς τα πραγματικά βάρη είναι τυπικά άγνωστα μπορούμε σε γενικές γραμμές να τα εκτιμήσουμε μη παραμετρικά από τις αναλογίες που προκύπτουν στο παρατηρούμενο δείγμα. Ωστόσο, όταν υπάρχουν ποικίλλες και συνεχείς ανεξάρτητες ερμηνευτικές μεταβλητές, αυτό δεν είναι δυνατό, οπότε για την εκτίμηση αυτών των βαρών μπορεί να χρησιμοποιηθεί ένα ημιπαραμετρικό

ή ένα παραμετρικό μοντέλο

Εν συντομία, εφαρμόζουμε ένα μοντέλο λογιστικής παλινδρόμησης με εξαρτημένη μεταβλητή, τη μεταβλητή έκθεσης του ασθενή X και ανεξάρτητες μεταβλητές το διάνυσμα των ερμηνευτικών μεταβλητών $\boldsymbol{Z}$. Έτσι παίρνουμε τις εκτιμήσεις των πιθανοτήτων $P(X = x|\boldsymbol{Z})$ από το μοντέλο που επιλεξαμε και στη συνέχεια υπόλογίζουμε τα βάρη $\hat{w}_i$ απλά αντιστρέφοντας αυτές τις πιθανότητες, $\hat{P}(X = x|\boldsymbol{Z})^{-1}$. Παρόλο που τα βάρη $\hat{w}_i$ είναι ασυμπτωτικά αμερόληπτα, στην πράξη παρουσιάζουν μεγάλη μεταβλητότητα. Προκειμένου να απαλλάχτουμε από αυτή τη μεταβλητότητα και να επιτύχουμε μια σχετική σταθερότητα των βαρών, υπολογίζουμε τα σταθμισμένα βάρη, αντικαθιστώντας τον αριθμητή του κλάσματος με την περιθώρια συνάρτηση πιθανότητας της μεταβλητής που δηλώνει την έκθεση στην παρατηρημένη κατάσταση, $s\hat{w}_i = f(X)/f(X|\boldsymbol{Z})$. Στη συνέχεια ένα ξεχωριστό μοντέλο λογιστικής παλινδρόμησης χρησιμοποιείται για την εκτίμηση του αριθμητή των σταθμισμένων βαρών sw_i.

Στην περίπτωση που τα δεδομένα επιβίωσης εμπεριέχουν διακεκομμένες παρατηρήσεις πιθανόν πληροφοριακές, ενσωματώνουμε το μηχανισμό διακοπής στην παραπάνω διαδικασία. Για να το επιτύχουμε αυτό, πολλαπλασιάζουμε το σταθμισμένο βάρος του κάθε ασθενή $s\hat{w}_i = f(X)/f(X|\boldsymbol{Z})$ με την αντίστροφη πιθανότητα των διακεκομμένων βαρών, που μεταβάλλεται με το χρόνο και είναι της μορφής $cw_i(t) = \prod_{j=1}^{t} P[C(t) = 0]/P[C(t) = 0|\boldsymbol{Z}]$ με $C(t)$ να είναι ένας δείκτης για τον μηχανισμό διακοπής. Δύο μοντέλα λογιστικής παλινδρόμησης χρησιμοποιούνται και πάλι για την εκτίμηση των όρων του κλάσματος.

Κεφάλαιο 4

Η ανάλυση των δεδομένων με τις προτεινόμενες μεθόδους

Σε αυτό το κεφάλαιο εφαρμόζουμε τις μεθόδους ανάλυσης δεδομένων επιβίωσης στις οποίες αναφερθήκαμε στα προηγούμενα κεφάλαια. Η πρώτη μέθοδος, που είναι και η καθιερωμένη είναι η υιοθέτηση ενός παραμετρικού μοντέλου για τους χρόνους επιβίωσης και η εκτίμηση των παραμέτρων που μας ενδιαφέρουν ακολουθώντας τη διαδικασία της μέγιστης πιθανοφάνειας. Η δεύτερη μέθοδος, εναλλακτική της πρώτης, είναι η υιοθέτηση ενός μοντέλου που επιτρέπει την εξάρτηση μεταξύ του χρόνου επιβίωσης και του χρόνου διακοπής με αποτέλεσμα να καταλήγουμε σε ένα εύρος για την εκτίμηση της παραμέτρου που μας ενδιαφέρει, ανάλογα με το μέγεθος της εξάρτησης. Η τρίτη μέθοδος και αυτή, εναλλακτική της πρώτης, είναι και πάλι η υιοθέτηση ενός παραμετρικού μοντέλου στο οποίο ενσωματώνεται ένα σταθμισμένο βάρος για τον κάθε ασθενή που εξαρτάται από το πόσο πιθανό είναι να παρατηρηθεί ή όχι ο αντίστοιχος χρόνος επιβίωσης

του. Στόχος μας είναι να συγκρίνουμε τις εκτιμήσεις των τριών μεθόδων σε προσομοιωμένα δεδομένα κατασκευασμένα με μια συγκεκριμένη μορφή εξάρτησης μεταξυ των μεταβλητών T και C και με διαφορετικά ποσοστά δεξιά διακεκομμένων παρατηρήσεων κάθε φορά.

4.1 Προσομοίωση δεδομένων με τέτοιο τρόπο ώστε να υπάρχει εξάρτηση μεταξύ του χρόνου επιβίωσης και του χρόνου διακοπής

Προτού εφαρμόσουμε τις δύο μεθόδους ανάλυσης δεδομένων επιβίωσης που αναφέραμε στο προηγούμενο κεφάλαιο προσομοιάζουμε το κατάλληλο σετ δεδομένων με συγκεκριμένα χαρακτηριστικά. Δημιουργούμε δύο ομάδες με 100 ασθενείς, έστω το γκρουπ A και το γκρουπ B και κατασκευάζουμε τους εν δυνάμει αληθινούς χρόνους επιβίωσής τους σε ημέρες. Οι χρόνοι επιβίωσης και για τα δύο γκρουπ ασθενών είναι εκθετικοί με παραμέτρους 0.01 και 0.02 αντίστοιχα. Αυτό σημαίνει ότι η μέση τιμή του χρόνου επιβίωσης για τους ασθενείς του γκρουπ A είναι περίπου 100 και για τους ασθενείς του γκρουπ B είναι περίπου 50 ημέρες. Εξ' αρχής δηλαδή, η συνάρτηση για το χρόνο επιβίωσης του γκρουπ B παίρνει μικρότερες τιμές από αυτή του γκρουπ A. Επίσης, το κάθε γκρουπ ασθενών ακολουθεί τη δική του θεραπεία. Κατασκευάσαμε τα δεδομένα με τέτοιο τρόπο ώστε οι ασθενείς του γκρουπ A να ακολουθούν τη θεραπεία με ένδειξη 0 και οι ασθενείς του γκρουπ B τη θεραπεία με ένδειξη 1.

Οι παρατηρούμενοι χρόνοι που περιλαμβάνονται στο σετ δεδομένων που θα αναλύσουμε είναι τέτοιοι ώστε να υπάρχει εξάρτηση μεταξύ του χρόνου επιβίωσης και του χρόνου διακοπής. Ο μηχανισμός εξάρτησης προκύπτει από τον παρακάτω συλλογισμό. Όταν η παρατήρηση που κατασκευάζουμε υποδηλώνει χρόνο επιβίωσης, τότε το διάνυσμα των παρατηρούμενων χρονικών τιμών είναι απλά το διάνυσμα που περιλαμβάνει τους τυχαίους εκθετικούς χρόνους των δύο γκρουπ ασθενών. Όταν όμως η παρατήρηση είναι δεξιά διακεκομμένη, δηλαδή υποδηλώνει χρόνο διακοπής, τότε το διάνυσμα των παρατηρούμενων χρονικών τιμών είναι το διάνυσμα που περιλαμβάνει τους αληθινούς (τυχαίους) χρόνους επιβίωσης για το γκρουπ A αλλά μικρότερους από τους αληθινούς (τυχαίους) χρόνους επιβίωσης για το γκρουπ B.

Με αυτόν τον τρόπο οι χρόνοι των διακεκομμένων παρατηρήσεων και το γεγονός που μελετάμε συσχετίζονται με ένα συγκεκριμένο μοτίβο. Αμέσως μετά τη στιγμή που εγκαταλείπουν τη μελέτη οι ασθενείς του γκρουπ A συμβαίνει το γεγονός. Δηλαδή, ο χρόνος της διακοπής για τους ασθενείς αυτού του γκρουπ είναι και ενδεικτικός του συμβάντος, το οποίο συμβαίνει αμέσως μετά. Για τους ασθενείς του γκρουπ B ο αληθινός χρόνος επιβίωσης είναι μεγαλύτερος από τον αντίστοιχο διακεκομμένο χρόνο παρατήρησής τους. Οπότε για να κατασκευάσουμε τους διακεκομμένους παρατηρούμενους χρόνους του γκρουπ B, αφαιρούμε ένα διάνυσμα με τυχαίους αριθμούς από την ομοιόμορφη κατανομή με άκρα το 0 και τις αληθινές τιμές των χρόνων επιβίωσης τους, από το διάνυσμα των αληθινών τιμών.

Επομένως, τα δεδομένα επιβίωσης που θα αναλύσουμε εκτός από τους παρατηρούμενους χρόνους περιλαμβάνουν ένα δίτιμο διάνυσμα, που παίρνει

την τιμή 1 όταν ο αντίστοιχος χρόνος είναι επιβίωσης και την τιμή 0 όταν είναι διακεκομμένος και ένα ακόμα δίτιμο διάνυσμα, που παίρνει την τιμή 1 και την τιμή 0, ανάλογα με τη θεραπεία που ακολουθει ο κάθε ασθενής. Τα προσομοιωμένα σετ δεδομένων προκύπτουν με χρήση του στατιστικού πακέτου R, μετά από 1000 επαναλήψεις και διαφορετικά ποσοστά διακεκομμένων παρατηρήσεων στο δείγμα. Ενδεικτικά ένα δείγμα από τα σετ δεδομένων που προκύπτουν, με ποσοστό διακεκομμένων παρατηρήσεων 50% παρουσιάζεται παρακάτω.

time	event	treat
41.4466545	0	0
326.1074073	1	0
7.8401580	1	0
130.1403305	1	0
39.4173229	0	0
11.61153094	1	0
1.2196944	1	0
207.0936305	1	0
49.6063717	1	0
125.6582380	0	0

4.2 Ανάλυση των προσομοιωμένων δεδομένων

Σύμφωνα με το τρόπο κατασκευής των δεδομένων, οι συναρτήσεις κατανομής των χρόνων επιβίωσης για το γκρουπ A και το γκρουπ B είναι

εκθετικές με παραμέτρους e^{b_0} και $e^{(b_0+b_1)}$ αντίστοιχα

$$f_{T_0}(t) = e^{b_0}e^{-e^{b_0}t}, f_{T_1}(t) = e^{(b_0+b_1)}e^{-e^{(b_0+b_1)}t}, \tag{4.1}$$

με δεδομένο ότι η παράμετρος που μας ενδιαφέρει να εκτιμήσουμε είναι το $b(x) = b_0 + b_1x$, με $b(x) = b_0$ για τους ασθενείς του γκρουπ A και $b(x) = b_0 + b_1$ για τους ασθενείς του γκρουπ B. Επίσης, σύμφωνα με τον τρόπο κατασκευής των δεδομένων οι αληθινές τιμές για αυτές τις παραμέτρους είναι $e^{b_0} = 0.01$ και $e^{(b_0+b_1)} = \log 0.02$ ή $b_1 = \log 2 = 0.6931472$, όταν $e^{b_0} = 0.01$. Ο σχετικός κίνδυνος ψ επομένως είναι 2 που σημαίνει ότι οι ασθενείς του γκρουπ B διατρέχουν 2 φορές μεγαλύτερο κίνδυνο να εκδηλώσουν το γεγονός σε σχέση με τον κίνδυνο που διατρέχουν οι ασθενείς του γκρουπ A. Η εκτίμηση για την παράμετρο b_1 που ουσιαστικά ερμηνεύει την επίδραση της θεραπείας στο χρόνο επιβίωσης και που θα προκύψει από την εκθετική παλινδρόμηση σε τρία διαφορετικά μοντέλα καθορίζει την καταλληλότητα του μοντέλου για την εξαγωγή των συμπερασμάτων.

Στην πρώτη μέθοδο ανάλυσης, υποθέτουμε ότι ο χρόνος διακοπής δεν εξαρτάται από το χρόνο επιβίωσης. Οι συναρτήσεις κινδύνου για τους ασθενείς του γκρουπ A και του γκρουπ B είναι αντίστοιχα $h_{T_0} = e^{b_0}$ και $h_{T_1} = e^{b_0+b_1}$. Με χρήση του στατιστικού πακέτου R εφαρμόζουμε μια εκθετική παλινδρόμηση με εξαρτημένη μεταβλητή το χρόνο time και ανεξάρτητη μεταβλητή, τη μεταβλητή treat για τα δύο γκρουπ ασθενών. Το εκθετικό μοντέλο που χρησιμοποιούμε είναι το

$$ER1 = survreg(Surv(data\$time, data\$event)\ data\$treat, \\ dist = "exponential"),$$

από το οποίο παίρνουμε τις εκτιμήσεις που θέλουμε για τις παραμέτρους b_0

και κυρίως για τη b_1. Συγκρίνοντας τις εκτιμήσεις αυτών των παραμέτρων με τις αληθινές τιμές τους, βρίσκουμε πόσο μεροληπτικά μπορεί να είναι τα αποτέλεσματα αυτής της μεθόδου για διαφορετικά ποσοστά διακεκομμένων παρατηρήσεων στο δείγμα.

Στην δεύτερη μέθοδο ανάλυσης, υποθέτουμε ότι ο χρόνος διακοπής εξαρτάται απο το χρόνο επιβίωσης, μέσω μιας παραμέτρου δ. Θεωρούμε επιπλέον τις συναρτήσεις κατανομής των διακεκομμένων χρόνων επιβίωσης για το γκρουπ A και το γκρουπ B, που είναι εκθετικές με παραμέτρους e^{γ_0} και $e^{(\gamma_0+\gamma_1)}$ αντίστοιχα

$$f_{C_0}(c) = e^{\gamma_0} e^{-e^{\gamma_0} c}, f_{C_1}(c) = e^{(\gamma_0+\gamma_1)} e^{-e^{(\gamma_0+\gamma_1)} c}, \tag{4.2}$$

αφού $c(x) = c_0 + c_1 x$, με $c(x) = c_0$ για τους ασθενείς του γκρουπ A και $c(x) = c_0 + c_1$ για τους ασθενείς του γκρουπ B. Με χρήση του στατιστικού πακέτου R εφαρμόζουμε μια δεύτερη εκθετική παλινδρόμηση με εξαρτημένη μεταβλητή το διακεκομμένο χρόνο time και ανεξάρτητη μεταβλητή, τη μεταβλητή treat για τα δύο γκρουπ ασθενών. Το εκθετικό μοντέλο που χρησιμοποιούμε είναι το

$$ER2 = survreg(Surv(data\$time, data\$1 - event)\ data\$treat, \\ dist = "exponential"),$$

από το οποίο παίρνουμε τις εκτιμήσεις που θέλουμε για τις παραμέτρους γ_0 και γ_1, τις οποίες μαζί με τις εκτιμήσεις $\hat{b}_0$ και $\hat{b}_1$ που ήδη έχουμε υπολογίσει από την πρώτη παλινδρόμηση, θα χρησιμοποιήσουμε για να εκτιμήσουμε την τελική τιμή της παραμέτρου b_1 που μας ενδιαφέρει. Η σχέση (3.23) για το γκρουπ A και για τα συγκεκριμένα προσομοιωμένα δεδομένα γίνεται

$$\hat{\theta}_{\delta 1} \simeq \hat{\theta}_{01} + \delta \frac{\sum_{i=1}^{200} \left\{ e^{\gamma_0} t_i^2 - (1 - I_i) t_i \right\}}{\sum_{i=1}^{200} t_i}, \tag{4.3}$$

και για το γκρουπ B

$$\hat{\theta}_{\delta 2} \simeq \hat{\theta}_{02} + \delta \frac{\sum_{i=1}^{200} \left\{ e^{(\gamma_0+\gamma_1)} t_i^2 - (1-I_i) t_i \right\}}{\sum_{i=1}^{200} t_i}, \tag{4.4}$$

όπου με $\hat{\theta}_{\delta 1}$ συμβολίζουμε την εκτίμηση $\hat{b}_0$ και με $\hat{\theta}_{\delta 2}$ την εκτίμηση $\hat{b}_1$ που έχουμε βρει από την πρώτη παλινδρόμηση. Η παράμετρος δ που φανερώνει την εξάρτηση μεταξύ των χρόνων επιβίωσης και των χρόνων διακοπής, όπως έχουμε ήδη αναφέρει παίρνει μικρές τιμές εντός του διαστήματος $[-0.3, 0.3]$. Για το κάθε γκρουπ ασθενών θα εκτιμήσουμε ένα εύρος τιμών που μπορούν να πάρουν οι εκτιμήσεις των παραμέτρων b_0 και b_1, όταν $\delta = -0.3$ και $\delta = 0.3$.

Η τρίτη μέθοδος αφορά τη στάθμιση των δεδομένων μέσω αντίστροφων πιθανοτητών και είναι μια παραλλαγή της πρώτης μεθόδου. Οι πιθανότητες που χρειάζεται να υπολογίσουμε προκειμένου να εκτιμήσουμε τα σταθμισμένα βάρη, όπως τα περιγράψαμε στο προηγούμενο κεφάλαιο, είναι $P(E=1)/P(E=1|\boldsymbol{X}=\mathbf{0})$, $P(E=1)/P(E=1|\boldsymbol{X}=\mathbf{1})$, $P(E=0)/P(E=0|\boldsymbol{X}=\mathbf{0})$ και $P(E=0)/P(E=0|\boldsymbol{X}=\mathbf{1})$. Με E συμβολίζουμε τη μεταβλητή event που παίρνει την τιμή 1 όταν σημειώνεται εκδήλωση του γεγονότος και 0 διαφορετικά.

Οι πιθανότητες $P(E=1)$ και $P(E=0)$ υπολογίζονται εύκολα μέσω των τύπων που ισχύουν για τη λογιστική παλινδρόμηση, $e^{\hat{b}_0}/1+e^{\hat{b}_0}$ και $1/1+e^{\hat{b}_0}$, όπου $\hat{b}_0$ είναι η εκτίμηση που παίρνουμε από το γραμμικό μοντέλο

$$L1 = glm(event \backsim 1, family = binomial, data = data)$$

αγνοώντας την επίδραση της θεραπείας στην εκδήλωση ή όχι του γεγονότος. Οι πιθανότητες $P(E=1|\boldsymbol{X}=\mathbf{0})$, $P(E=1|\boldsymbol{X}=\mathbf{1})$, $P(E=0|\boldsymbol{X}=\mathbf{0})$ και $P(E=0|\boldsymbol{X}=\mathbf{1})$ υπολογίζονται επίσης μέσω των τύπων

που ισχύουν για τη λογιστική παλινδρόμηση, $e^{\hat{b}_0}/1+e^{\hat{b}_0}$, $e^{\hat{b}_0+\hat{b}_1}/1+e^{\hat{b}_0+\hat{b}_1}$, $1/1+e^{\hat{b}_0}$ και $1/1+e^{\hat{b}_0+\hat{b}_1}$ αντίστοιχα. Όμοια, $\hat{b}_0$ και $\hat{b}_1$ είναι οι εκτιμήσεις που παίρνουμε από το γραμμικό μοντέλο

$$L2 = glm(event \backsim treat, family = binomial, data = data)$$

θεωρώντας τη θεραπεία ως παράγοντα για την εκδήλωση ή όχι του γεγονότος.

Με χρήση του στατιστικού πακέτου R, τα σταθμισμένα βάρη υπολογίζονται εύκολα. Στη συνέχεια εφαρμόζουμε μια εκθετική παλινδρόμηση, έχοντας ενσωματώσει τα σταθμισμένα βάρη στην ανάλυση, μέσω του μοντέλου

$$ER3 = survreg(Surv(data\$time, data\$event) \backsim data\$treat, weights = w1, dist = "exponential")$$

απο όπου παιρνουμε τις τελικές εκτιμήσεις γαι την παράμετρο που εξ΄ αρχής μας ενδιαφέρει.

4.2.1 Οι εκτιμήσεις των παραμέτρων

Από την ανάλυση των δεδομένων που κάναμε με χρήση του στατιστικού πακέτου R, ενδιαφερόμαστε να συγκρίνουμε τις εκτιμήσεις που υπολογίσαμε με τις τρεις διαφορετικές μεθόδους για την παράμετρο b_1, η οποία ευθύνεται για τη επίδραση που έχει η θεραπεία στην συνάρτηση επιβίωσης. Στον πίνακα 4.1 φαίνονται αναλυτικά οι εκτιμήσεις αυτές. Το ποσοστό διακεκομμένων παρατηρήσεων είναι 50%, 40%, 30%, 20% και 10% αντίστοιχα.

Επίσης, στους πίνακες 4.2, 4.3, 4.4 και 4.5 φαίνονται τα 95% διαστήματα εμπιστοσύνης που προέκυψαν για την παράμετρο b_1 για κάθε μέθοδο και για

Πίνακας 4.1: Οι εκτιμήσεις για την παράμετρο b_1

Πραγματική τιμή	Ανάλυση 1η	Ανάλυση 2η	Ανάλυση 3η
0.6931472	1.007137	[0.5753107, 1.4389641]	0.9928988
0.6931472	0.9321655	[0.6425226, 1.2218085]	0.930498
0.6931472	0.8743951	[0.6523872, 1.0964031]	0.8679793
0.6931472	0.8100575	[0.6300238, 0.9900911]	0.8080238
0.6931472	0.7629473	[0.6855454, 0.8403492]	0.7633686

το αντίστοιχο ποσοστό διακεκομμένων παρατηρήσεων στο δείγμα. Επίσης στους ίδιους πίνακες φαίνονται οι πιθανότητες οι συγκεκριμένες εκτιμήσεις να βρίσκονται εντός του αντίστοιχου διαστήματος εμπιστοσύνης.

Πίνακας 4.2: Δ.Ε για b_1, με 50% διακοπή

Ανάλυση	95% Δ.Ε	Πιθανότητα
1η	[0.5995035, 1.4147710]	0.961
2η	[0.2671258, 0.5502135]	0.523
2η	[0.9318812, 2.279329]	0.999
3η	[0.5852649, 1.400533]	0.989

Μία πρώτη παρατήρηση που μπορούμε να κάνουμε και η οποία είναι μια φυσική συνέπεια είναι ότι όσο πιο μικρό είναι το ποσοστό των διακεκομμένων παρατηρήσεων στο δείγμα, τόσο πιο κόντα είναι οι εκτιμημένες τιμές της παραμέτρου b_1 στις αληθινές και για τις τρεις μεθόδους ανάλυσης. Για παράδειγμα όταν το ποσοστό των διακεκομμένων παρατηρήσεων στο δείγμα είναι 10% η απόκλιση που παρατηρείται στην παρατηρημένη τιμή από την

Πίνακας 4.3: Δ.Ε για b_1, με 40% διακοπή

Ανάλυση	95% Δ.Ε	Πιθανότητα
1η	$[0.5784251, 1.285906]$	0.957
2η	$[0.3807947, 0.6526398]$	0.547
2η	$[0.7760555, 1.919172]$	1
3η	$[0.5767577, 1.284238]$	0.991

Πίνακας 4.4: Δ.Ε για b_1, με 30% διακοπή

Ανάλυση	95% Δ.Ε	Πιθανότητα
1η	$[0.5542205, 1.194570]$	0.953
2η	$[0.4017769, 0.6007679]$	0.456
2η	$[0.7066641, 1.788372]$	0.999
3η	$[0.5478047, 1.188154]$	0.963

πραγματική είναι περίπου 0.05, για την πρώτη και την τρίτη μέθοδο ανάλυσης, ενώ όταν το ποσοστό των διακεκομμένων παρατηρήσεων στο δείγμα είναι 50% αυτή η απόκλιση είναι περίπου 0.3.

Σχετικά με τη δεύτερη μέθοδο παρατηρούμε ότι όσο πιο μικρό είναι το ποσοστό των διακεκομμένων παρατηρήσεων στο δείγμα τόσο μικρότερο είναι το έυρος των εκτιμήσεων που παίρνουμε. Επίσης, παρατηρούμε ότι οι εκτιμήσεις της καθιερωμένης μεθόδου είναι σχεδόν ίσες με αυτές που προκύπτουν από την ανάλυση με τα σταθμισμένα βάρη. Αυτό που μπορούμε να συμπεράνουμε είναι ότι μόνο για τα συγκεκρίμενα προσομοιωμένα δεδομένα, με τη δεδομένη μορφή εξάρτησης μεταξύ χρόνου επτβίωσης και χρόνου διακοπής, οι εκτιμήσεις των δύο μεθόδων σχεδόν ταυτίζονται, ως εκ τούτου

Πίνακας 4.5: Δ.Ε για b_1, με 20% διακοπή

Ανάλυση	95% Δ.Ε	Πιθανότητα
1η	$[0.4902084, 1.129907]$	0.955
2η	$[0.3869015, 0.6470606]$	0.587
2η	$[0.5935154, 1.612752]$	0.997
3η	$[0.4881748, 1.127873]$	0.969

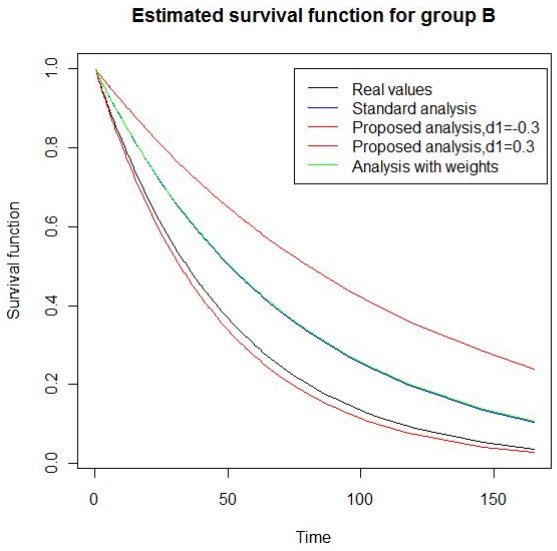
Πίνακας 4.6: Δ.Ε για b_1, με 10% διακοπή

Ανάλυση	95% Δ.Ε	Πιθανότητα
1η	$[0.4747443, 1.051150]$	0.946
2η	$[0.4429095, 0.707591]$	0.598
2η	$[0.506579, 1.394710]$	0.994
3η	$[0.4751655, 1.051572]$	0.96

μια τέτοια παρατήρηση δεν μπορεί να γενικευτεί.

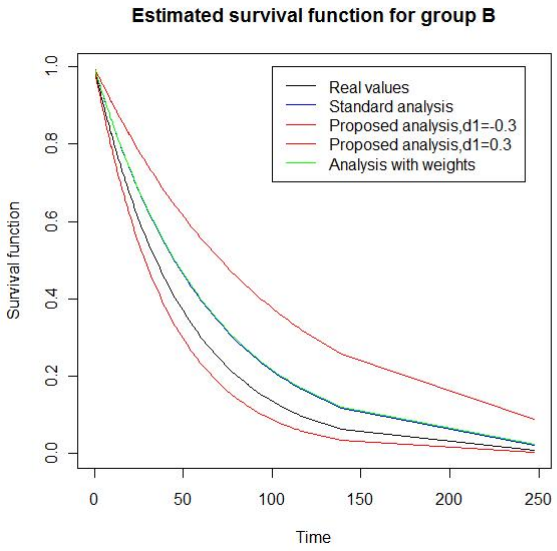
Επιπλέον, παρατηρούμε οι πραγματική τιμή της παραμέτρου b_1, είναι αρκετά πιο κοντά στο κάτω άκρο του εύρους των εκτιμήσεων που προκύπτουν από τη δεύτερη εναλλακτική μέθοδο ανάλυσης. Αυτό σημαίνει, ότι όταν η παράμετρος $\delta \in [-0.3, 0.3]$, που φανερώνει το μέγεθος της εξάρτησης παίρνει τιμές πιο κοντά στο -0.3, η εκτίμηση που παίρνουμε είναι πιο κοντά στη αληθινή από ό,τι αυτή που παίρνουμε όταν η παράμετρος δ παίρνει τιμές πιο κοντά στο 0.3. Ωστόσο, οι εκτιμήσεις που πήραμε όταν $\delta = -0.3$, έχουν όχι και τόσο μεγάλη πιθανότητα, (περίπου 50%), να βρίσκονται εντός του διαστήματος εμπιστοσύνης που κατασκευάζεται με βάση αυτή την εκτίμηση.

Στα παρακάτω διαγράμματα απεικονίζονται και η πραγματική συνάρτηση

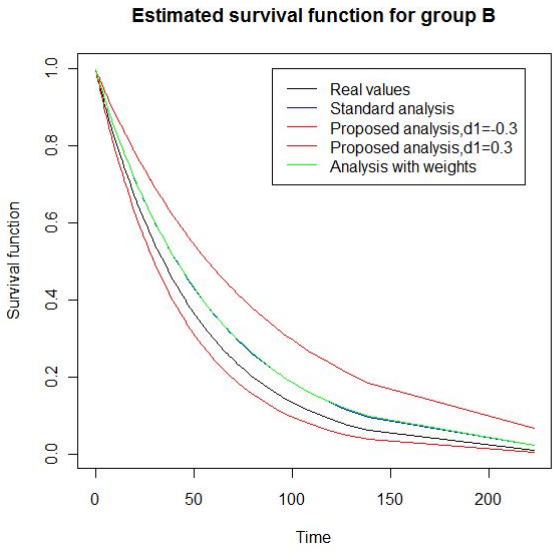


Σχήμα 4.1: Η εκτίμηση της συνάρτησης επιβίωσης για το γκρουπ B, όταν το ποσοστό των διακεκομμένων παρατηρήσεων στο δείγμα είναι 50%

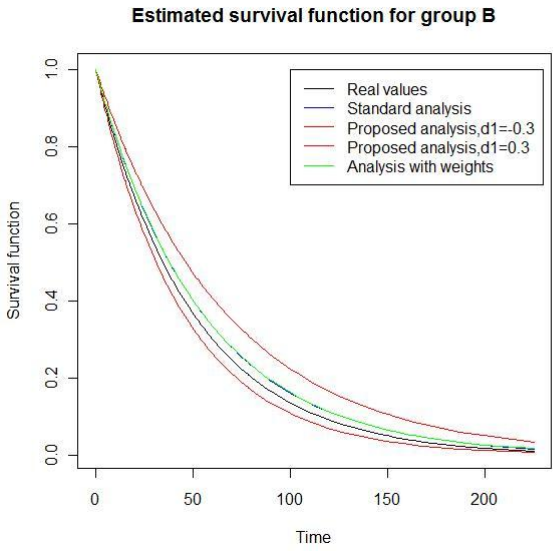
επιβίωσης του γκρουπ B και οι εκτιμήσεις για τη συνάρτηση αυτή, σύμφωνα με τα αποτελέσματα που προέκυψαν από την ανάλυση.



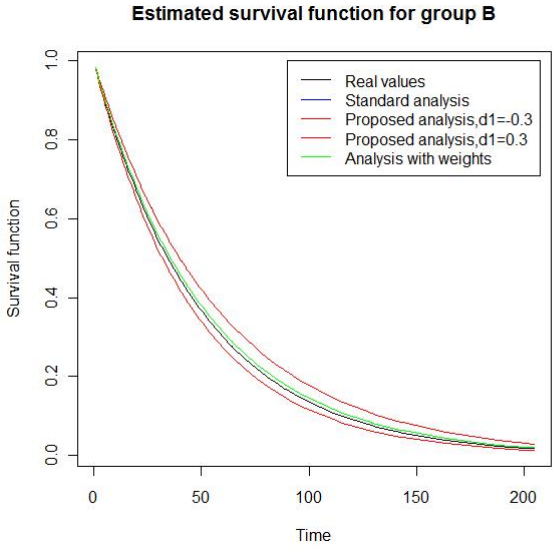
Σχήμα 4.2: Η εκτίμηση της συνάρτησης επιβίωσης για το γκρουπ B, όταν το ποσοστό των διακεκομμένων παρατηρήσεων στο δείγμα είναι 40%



Σχήμα 4.3: Η εκτίμηση της συνάρτησης επιβίωσης για το γκρουπ B, όταν το ποσοστό των διακεκομμένων παρατηρήσεων στο δείγμα είναι 30%



Σχήμα 4.4: Η εκτίμηση της συνάρτησης επιβίωσης για το γκρουπ B, όταν το ποσοστό των διακεκομμένων παρατηρήσεων στο δείγμα είναι 20%



Σχήμα 4.5: Η εκτίμηση της συνάρτησης επιβίωσης για το γκρουπ B, όταν το ποσοστό των διακεκομμένων παρατηρήσεων στο δείγμα είναι 10%

Βιβλιογραφία

[1] D. Collet. *Modelling Survival Data in Medical Research.* Chapman & Hall / CRC Press, second edition, 2003.

[2] J.P. Klein, M.L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data.* Springer, second edition, 2003.

[3] R.J.A. Little, D.B. Rubin. *Statistical Analysis with Missing Data.* John Wiley and sons, 1987.

[4] S.R. Cole, M.A. Hernan. Adjusted Survival Curves with inverse probability weighting. *Computer Methods and Programes in Biomedicine.* **75** (2004), pp. 45–49.

[5] F. Siannis, J. Copas, G. Lu. Sensitivity analysis for informative censoring in parametric survival models. *Biostatistics***6** (2005), pp. 77–91.

[6] W. van der Wal, R.B. Geskus. ipw: An R Package for inverse Probability Weighting. *Journal of Statistical Software* **43** (2011), pp. 1–23.

[7] D.O. Scharfstein, J.M. Robins. Estimation of the failure time distribution in the presence of informative censoring. *Biometrica* **89** (2002), pp. 617–634.

[8] G.L. Grunkemeier, M.J.C. Eijkemans, J.J.M. Takkenberg. Actual and Actuarial Probabilities of Competing Risks: Apples and Lemons. *Society of Thoracic Surgeons* **86** (2007), pp. 1586–1592.

Κώδικας R

```
library (survival)
library(ipw)


###DATA ΦΩΤΕΙΝΗ 06/12/11 #############
#######################################

a0=matrix(0,nrow=1000,ncol=1)
a1=matrix(0,nrow=1000,ncol=2)

r0=matrix(0,nrow=1000,ncol=1)
r1=matrix(0,nrow=1000,ncol=2)

v0=matrix(0,nrow=1000,ncol=1)
v1=matrix(0,nrow=1000,ncol=2)

cc=0.9

for(i in 1:1000){
n=100
T0=rexp(n,rate=0.01); sort(T0);
T1=rexp(n,rate=0.02); sort(T1);
mean(T0)
mean(T1)
#με πιθανότητα cc έχω παρατηρήσει το χρόνο επιβίωσης.
I=rbinom(2*n,size=1,prob=cc); I; mean(I)
fot=function(x){runif(1,0,x)}

#
dim(T0)=c(n,1)
dim(T1)=c(n,1)
U1=apply(T1,1,fot); U1; dim(U1)=c(n,1);
T=rbind(T0,T1); T;
U=rbind(T0,T1-U1)
W=I*T + (1-I)*U ; min(W)
W
########################################
########################################
data=data.frame(time=W,event=I,treat=c(rep(0,n),rep(1,n)),Censor=1-I)
data
table(data$event,data$treat)


#################################################################
#################################################################
ER1=survreg(Surv(data$time,data$event)~data$treat,dist="exponential")
summary(ER1)
```

```
#
a1[i,]=exp(-ER1$coef)
#a[i]=exp(-ER1$coef[1])

ER2=survreg(Surv(data$time,1-data$event)~data$treat,dist="exponential")
summary(ER2)
#
r1[i,]=exp(-ER2$coef)
#r[i]=exp(-ER2$coef[1])


sw1=ipwpoint(exposure=event,family='binomial',link='logit',numerator=~1,denomina
tor=~treat,data=data)

####stabilized weights####
w1=sw1$ipw.weights
table(w1)
ER3=survreg(Surv(data$time,data$event)~data$treat,weights=w1,dist="exponential")
summary(ER3)
#
v1[i,]=exp(-ER3$coef)
#v[i]=exp(-ER3$coef[1])
}

#####################################################################
#######################Standard estimates#####################
lhat1=mean(a1[,1]);
lhat2=lhat1*mean(a1[,2])
lhat1;lhat2;

thhat1=c(rep(log(lhat1),100));dim(thhat1)=c(100,1);thhat1[1,1]
thhat2=c(rep(log(lhat2),100));dim(thhat2)=c(100,1);thhat2[1,1]

#####################################
b12=thhat2[1,1]-thhat1[1,1];b12
#####################################
exp(b12)

datat0=data[data$treat==0,]$event;
s0=sum(datat0)
datat1=data[data$treat==1,]$event;
s1=sum(datat1)

psihat=exp(b12)

seb12=sqrt((s0+s1)/(s0*s1));seb12
b12-1.959964*seb12 ## kato akro ci##
b12+1.959964*seb12 ##ano akro ci##
```

```
cib12=c(b12-1.959964*seb12,b12+1.959964*seb12);dim(cib12)=c(1,2);cib12 ####ci
gia b12###

A=log(a1[,2])
AA=data.frame(est=A)
AAA=AA[AA<cib12[1,1]];AAA
AAAA=AA[AA>cib12[1,2]];AAAA
p=(1000-(length(AAA)+length(AAAA)))/1000;p

##########################
data1=data[data$treat==0,]$time;dim(data1)=c(100,1);d1=sort(data1)
S1=exp(-exp(thhat1)*d1);S1
plot(d1,S1,type="l",xlab="Time for group A",ylab="Survival function")
############################

data2=data[data$treat==1,]$time;dim(data2)=c(100,1);d2=sort(data2)
########################################
##########Estimates of Proposed Analysis#####
########################################

l2hat1=mean(r1[,1]);
l2hat2=l2hat1*mean(r1[,2])
l2hat1;l2hat2;

gamhat1=c(rep(log(l2hat1),100));dim(gamhat1)=c(100,1)
gamhat2=c(rep(log(l2hat2),100));dim(gamhat2)=c(100,1)
gamhat=c(gamhat1,gamhat2);dim(gamhat)=c(200,1);gamhat

e1=exp(gamhat1)
e2=exp(gamhat2)
e=exp(gamhat)

Datat0=data[data$treat==0,]$Censor;
S0=sum(Datat0)
Datat1=data[data$treat==1,]$Censor;
S1=sum(Datat1)

psi2=l2hat2/l2hat1      ###gamma1hat####
psi2

gamma12=log(psi2);gamma12
segamma12=sqrt((S0+S1)/(S0*S1));segamma12
kagamma=gamma12-1.959964*segamma12 ## kato akro ci gia gamma1##
pagamma=gamma12+1.959964*segamma12 ##ano akro ci gia gamma1##
cigam1=c(kagamma,pagamma);cigam1

datae1=data[data$event==1,]$time
datae0=data[data$event==0,]$time

s1=sum((data$time)^2)
```

```
s2=sum(datae0)
s3=sum(data$time)

si1=((e1*s1)-s2)/s3; si1 ###sensitivity index for g0###
si2=((e2*s1)-s2)/s3; si2 ###sensitivity index for g1###
si=((e*s1)-s2)/s3;si

nsik=(((e1*exp(kagamma))*s1)-s2)/s3; ## kato akro gia sensitivity index###
nsia=(((e1*exp(pagamma))*s1)-s2)/s3; ## ano akro gia sensitivity index###

cisi=c(nsik[1,1],nsia[1,1]);dim(cisi)=c(1,2);cisi###ci gia si###

thdeltahat10=thhat1-(0.3*si1);thdeltahat10[1,1] ###est of theta,g0,δ1###
thdeltahat11=thhat1+(0.3*si1);thdeltahat11[1,1] ###est of theta,g0,δ2###

thdeltahat20=thhat2-(0.3*si2);thdeltahat20[1,1] ###est of theta,g1,δ1###
thdeltahat21=thhat2+(0.3*si2);thdeltahat21[1,1] ###est of theta,g1,δ2###


ci1=c(cib12[1,1]+0.3*cisi[1,1],cib12[1,2]+0.3*cisi[1,2]);dim(ci1)=c(1,2)
ci2=c(cib12[1,1]-0.3*cisi[1,1],cib12[1,2]-0.3*cisi[1,2]);dim(ci2)=c(1,2)
ci1;ci2;

B1=log(a1[,2])+(0.3*si2[,1]);B1
BB1=data.frame(est=B1)
BBB1=BB1[BB1<ci1[1,1]];BBB1
BBBB1=BB1[BB1>ci1[1,2]];BBBB1
p=(1000-(length(BBB1)+length(BBBB1)))/1000;p

B2=log(a1[,2])-(0.3*si2[,1]);B2
BB2=data.frame(est=B2)
BBB2=BB2[BB2<ci2[1,1]];BBB2
BBBB2=BB2[BB2>ci2[1,2]];BBBB2
p=(1000-(length(BBB2)+length(BBBB2)))/1000;p

#####################################################
b120=thdeltahat20[1,1]-thdeltahat10[1,1];b120
b121=thdeltahat21[1,1]-thdeltahat11[1,1];b121
rangeb12=c(b120,b121);rangeb12
#####################################################
##############################IPW#####################
#####################################################
l3hat1=mean(v1[,1]);
l3hat2=l3hat1*mean(v1[,2])
l3hat1;l3hat2;

thhat31=c(rep(log(l3hat1),100));dim(thhat31)=c(100,1);thhat31[1,1]

thhat32=c(rep(log(l3hat2),100));dim(thhat32)=c(100,1);thhat32[1,1]
```

```
#####################################################
b13=thhat32[1,1]-thhat31[1,1];b13
#####################################################
seb13=sqrt((s0+s1)/(s0*s1));seb12
b13-1.959964*seb13 ## kato akro ci##
b13+1.959964*seb13 ##ano akro ci##

cib13=c(b13-1.959964*seb12,b13+1.959964*seb12);dim(cib13)=c(1,2);cib13 ####ci
gia b12###

C=log(v1[,2])
CC=data.frame(est=C)
CCC=CC[CC<cib13[1,1]];CCC
CCCC=CC[CC>cib13[1,2]];CCCC
p=(1000-(length(CCC)+length(CCCC)))/1000;p
#############################################################
S0=exp(-(0.02)*d2);S0
S1=exp(-exp(thhat2)*d2);S1
S20=exp(-exp(thdeltahat20)*d2);S20 ###est survival,g1,δ1###
S21=exp(-exp(thdeltahat21)*d2);S21 ###est survival,g1,δ2###
S3=exp(-exp(thhat32)*d2);S3

x1=d2;x2=d2;x3=d2;x4=d2;x5=d2
y1=S0;y2=S1;y3=S20;y4=S21;y5=S3
plot(c(x1,x2,x3,x4),c(y1,y2,y3,y4),main="Estimated survivor Function for group
B",type="n",xlab="Time",ylab="Survival function")
lines(x1,y1,lty=1,col="1")
lines(x2,y2,lty=1,col="blue")
lines(x3,y3,lty=1,col="red")
lines(x4,y4,lty=1,col="red")
lines(x5,y5,lty=1,col="green")
legend(100,0.9,c("Real values","Standard Analysis","Proposed Analysis,d1=-
0.3","Proposed Analysis,d1=0.3","Analysis with
weights"),lty=c(1,1,1,1),col=c("1","2","3","3"))
```

ΓΛΩΣΣΑΡΙ

1. Συνάρτηση Επιβίωσης	Survivor Function
2. Συνάρτηση Κινδύνου	Hazard Function
3. Αθροιστική Συνάρτηση Κινδύνου	Cumulative Hazard Function
4. Εμπειρική Συνάρτηση Επιβίωσης	Empirical Survivor Function
5. Εμπειρική Αθροιστική Συνάρτηση Κατανομής	Empirical Cumulative Distribution Function
6. Εκτιμητής Πίνακας Ζωής	Life - Table Estimate
7. Αναλογιστικός Εκτιμητής	Actuarial Estimate
8. Μοντέλο Αναλογικών Κινδύνων	Proportional Hazards Model
9. Σχετικός Κίνδυνος	Relative Hazard
10. Βασική Συνάρτηση Κινδύνου	Baseline Hazard Function
11. Γραμμικό Μέρος	Linear Component
12. Προγνωστικός Δείκτης	Prognostic Index
13. Παράμετρος Σχήματος	Shape Parameter
14. Παράμετρος Κλίμακας	Scale Parameter
15. Διακεκομμένες Παρατηρήσεις	Censored Observations
16. Ελλιπή Δεδομένα	Missing Data
17. Περικομμένες Παρατηρήσεις	Truncated Observations
18. Διακεκομμένοι Χρόνοι Επιβίωσης	Censored Survival Times
19. Δεξιά Διακεκομμένες Παρατηρήσεις	Right Censored Observations
20. Αριστερά Διακεκομμένες Παρατηρήσεις	Left Censored Observations
21. Εντός Διαστήματος Διακεκομμένες Παρατηρήσεις	Interval Censored Observations
22. Διαχρονικά Δεδομένα	Longitudinal Data

23. Αριστερά Περικομμένες Παρατηρήσεις	Left Truncated Observations
24. Δεξιά Περικομμένες Παρατηρήσεις	Right Truncated Observations
25. Ελλιπή Δεδομένα ως Τυχαίο Φαινόμενο	Missing at Random (MAR)
26. Ελλιπή Δεδομένα ως Εντελώς Τυχαίο Φαινόμενο	Missing Completely at Random (MCAR)
27. Ελλιπή Δεδομένα ως Μη Τυχαίο Φαινόμενο	Missing Not at Random (MNAR)