ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Δ.Π.Μ.Σ ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ

Αποκομμένα Δεδομένα σε Διάστημα-Ανάλυση και Εφαρμογές

Διπλωματική Εργασία

της

Αθανασίας-Μαρίας Ε. Σταθάκη

Επιβλέπουσα: Χρυσηίς Καρώνη
Καθηγήτρια Ε.Μ.Π.

Αθήνα, 2015

Η παρούσα Διπλωματική Εργασία εκπονήθηκε
στα πλαίσια των σπουδών μου για την απόκτηση του
Μεταπτυχιακού Διπλώματος Ειδίκευσης στις
Εφαρμοσμένες Μαθηματικές Επιστήμες.

Τριμελής Επιτροπή

Ονοματεπώνυμο	**Βαθμίδα**
Καρώνη Χρυσηίς	Καθηγήτρια (επιβλέπουσα)
Κουκουβίνος Χρήστος	Καθηγητής
Βόντα Φιλία	Αναπληρώτρια καθηγήτρια

Πίνακας περιεχομένων

Περίληψη

Η παρούσα εργασία επικεντρώνεται στην παρουσίαση μιας ειδικής περίπτωσης δεδομένων διάρκειας ζωής, που βρίσκονται σε ένα συγκεκριμένο διάστημα το οποίο πάντα είναι υποσύνολο της συνολικής διάρκειας ζωής της εκάστοτε παρατηρούμενης μονάδας. Οι προσεγγίσεις μέσω απόδοσης τιμών (imputation) χρησιμοποιούνται για να ανάγουν το πρόβλημα της ανάλυσης των χρόνων αποτυχίας των αποκομμένων δεδομένων σε διάστημα σε εκείνο της ανάλυσης χρόνων αποτυχίας με δεξιά αποκομμένα δεδομένα και αυτό γιατί ένα βασικό πλεονέκτημα των παραμετρικών προσεγγίσεων είναι ότι η εφαρμογή τους είναι απλή και γενικά ισχύει η συνήθης θεωρία μεγίστης πιθανοφάνειας. Έτσι, μπορούμε να αποφύγουμε τη χρήση της αποκοπής σε διάστημα (interval censoring) και να χρησιμοποιήσουμε τις υπάρχουσες συμπερασματικές διαδικασίες και του στατιστικού λογισμικού που έχει αναπτυχθεί για τα δεξιά αποκομμένα δεδομένα. Για την ανάλυση των χρόνων αποτυχίας βασιζόμαστε στην εκτίμηση της συνάρτησης επιβίωσης. Όταν έχουμε ένα παραμετρικό μοντέλο το πρόβλημα εκτίμησης είναι σχετικά εύκολο και χρησιμοποιείται συνήθως η εκτιμήτρια μεγίστης πιθανοφάνειας. Εδώ επειδή τα αποκομμένα δεδομένα σε διάστημα έχουν αναχθεί σε αυτά των δεξιά αποκομμένων δεδομένων η αντίστοιχη μη-παραμετρική εκτιμήτρια μεγίστης πιθανοφάνειας (NPMLE) της συνάρτησης επιβίωσης δίνεται από την εκτιμήτρια Kaplan-Meier. Γίνεται εκτενής αναφορά στις περιπτώσεις I και II των αποκομμένων δεδομένων σε διάστημα. Παρουσιάζονται σταθμισμένοι έλεγχοι log-rank (weighted log-rank tests) καθώς και η σταθμισμένη Kaplan-Meier για τη σύγκριση μεταξύ συναρτήσεων επιβίωσης και επιπλέον γίνεται αναφορά στην ανάλυση παλινδρόμησης για τις δύο αυτές περιπτώσεις των αποκομμένων δεδομένων. Ακολουθούν παραδείγματα και στατιστική ανάλυση δεδομένων.

Abstract

The present thesis focuses on a particular case of failure time data, namely, interval-censored data. The failure time is known to fall within an interval which is always a subset of the total life of each unit observed. Imputation approaches are used to reduce the problem of analyzing interval-censored failure time data to that of analyzing right-censored failure time data. This is because a main advantage of parametric approaches is that their implementation is straightforward in principle and in fact standard maximum likelihood theory generally applies. Thus, one can avoid dealing with interval censoring and use existing inference procedures and statistical software developed for right-censored data. The analysis of failure time data is based on the estimation of the survival function. When we have a parametric model, the estimation problem is relatively easy and the maximum likelihood approach is commonly used. Because the interval-censored data have been reduced to those of the right-censored data, the nonparametric maximum likelihood estimator (NPMLE) of a survival function is given by the Kaplan-Meier estimator. In addition cases I and II of interval-censored data and their regression analysis are discussed as well as the use of weighted log-rank tests and the weighted Kaplan-Meier for comparing survival functions in such cases. Examples of data analysis are presented.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά την επιβλέπουσα καθηγήτρια κυρία Χ.Καρώνη για την πολύτιμη καθοδήγηση και στήριξή της σε όλη τη διάρκεια εκπόνησης της παρούσας εργασίας.

Επιπλέον, θα ήθελα να ευχαριστήσω την οικογένειά μου και τους φίλους μου για τη διαρκή συμπαράσταση και υποστήριξή τους καθ᾽ όλη τη διάρκεια των σπουδών μου.

Κεφάλαιο 1

Εισαγωγή

Προτού αναφερθούμε στην έννοια και τη χρησιμότητα της στατιστικής ανάλυσης, θεωρώ σκόπιμο να δώσουμε την ερμηνεία για κάποιες εισαγωγικές έννοιες που θα μας βοηθήσουν στην κατανόηση του αντικειμένου που πραγματεύεται η παρούσα εργασία.

Όταν αναφερόμαστε σε ένα στατιστικό μοντέλο εννοούμε συνοπτικά τη μαθηματική διατύπωση υπό μορφή εξισώσεων της σχέσης μεταξύ μεταβλητών. Περιγράφει με μαθηματικό τρόπο πως σχετίζονται μία ή περισσότερες μεταβλητές (εξαρτημένες) με κάποιες άλλες μεταβλητές (ανεξάρτητες). Το μοντέλο αναφέρεται ως στατιστικό αφού οι μεταβλητές συνδέονται μεταξύ τους στοχαστικά και όχι ντετερμινιστικά. Παραδείγματος χάριν, το βάρος ενός ανθρώπου εξαρτάται από την ηλικία του. Αν γνωρίζουμε την ηλικία ενός ατόμου και διαθέτουμε ένα στατιστικό μοντέλο που να συνδέει τα δύο χαρακτηριστικά, τότε μπορούμε να βρούμε την πιθανότητα να έχει κάποιο συγκεκριμένο βάρος.

Όταν αναφερόμαστε σε στατιστικές αναλύσεις, εννοούμε την περαιτέρω μελέτη και επεξήγηση των δεδομένων τα οποία έχουμε συλλέξει. Ο λόγος για τον οποίο διεξάγουμε στατιστικές αναλύσεις είναι γιατί η απλή καταγραφή των δεδομένων δε μπορεί να μας εξασφαλίσει με σαφήνεια και λεπτομέρεια τη δυναμική που έχουν τα δεδομένα μας.

Η ανάλυσή τους όμως μας οδηγεί σε ασφαλή συμπεράσματα και καλύτερη κατανόηση των συνθηκών (της συλλογής των δεδομένων). Σημαντικό ρόλο παίζουν και οι συμμεταβλητές του μοντέλου, με το οποίο έχουμε αποφασίσει να πραγματοποιήσουμε την ανάλυσή μας. Ποιες από τις συμμεταβλητές πρέπει να συμπεριληφθούν στο μοντέλο; Γιατί είναι στατιστικά σημαντικές και επηρεάζουν την εξαρτημένη μεταβλητή; Ποιες μπορούν να παραληφθούν γιατί συνεισφέρουν από ελάχιστα έως και καθόλου; Επιθυμητό λοιπόν είναι ο αναλυτής να γνωρίζει τους λόγους για τους οποίους γίνεται η στατιστική ανάλυση.

Στην εργασία αυτή θα ασχοληθούμε αποκλειστικά με μία ειδική περίπτωση των δεδομένων διάρκειας ζωής, τη διάρκεια ζωής σε ένα συγκεκριμένο διάστημα το οποίο πάντα είναι υποσύνολο της συνολικής διάρκειας ζωής της εκάστοτε παρατηρούμενης μονάδας.

Οι καταστάσεις όπου η παρατηρούμενη απόκριση (σ.σ. αποτέλεσμα παρατήρησης) για κάθε μονάδα που υπόκειται σε μία μελέτη, είναι δύο. Είτε ένας συγκεκριμένος χρόνος επιβίωσης (ή αλλιώς χρόνος που παρατηρείται το συμβάν) είτε ένας αποκομμένος χρόνος (censored). Αξίζει να σημειωθεί ότι και τα δύο είναι αρκετά συνήθη στην πράξη.

Παρόλα αυτά, μπορούν να λάβουν χώρα και διάφορες άλλες καταστάσεις, κάποιες εκ των οποίων τις συναντάμε στις διαχρονικές μελέτες. Στις διαχρονικές μελέτες, λοιπόν, οι ασθενείς μπορούν να παρακολουθηθούν μόνο περιοδικά. Το γεγονός καταγράφεται μόνο σε κάποιο συγκεκριμένο χρονικό διάστημα. Το χρόνο αυτού του συμβάντος, θα τον αναφέρουμε ως το χρόνο αποτυχίας (failure time data) ή το χρόνο επιβίωσης (survival time).

1.1 Δεδομένα χρόνων αποτυχίας

Με τον όρο χρόνος αποτυχίας (failure time data) εννοούμε δεδομένα που αφορούν μη αρνητικές τυχαίες μεταβλητές που αντιπροσωπεύουν χρόνους συγκεκριμένων γεγονότων. Παραδείγματα γεγονότων, τα οποία συχνά αναφέρονται ως γεγονότα αποτυχίας ή επιβίωσης (failure or survival event) συμπεριλαμβανομένου και του θανάτου, είναι η αρχή μιας ασθένειας, ένα συγκεκριμένο ορόσημο, ή η αποτυχία-βλάβη ενός εξαρτήματος μιας μηχανής.

Η πραγματοποίηση του συμβάντος συνήθως αναφέρεται ως αποτυχία (failure). Τα δεδομένα του χρόνου αποτυχίας (failure time data) εμφανίζονται σε μεγάλο βαθμό σε διάφορες ιατρικές μελέτες, αλλά υπάρχουν εξίσου πολλές έρευνες οι οποίες παράγουν ανάλογα δεδομένα. Αυτές με τη σειρά τους, περιλαμβάνουν βιολογικές, δημογραφικές, οικονομικές ή κοινωνιολογικές μελέτες, καθώς και πειράματα αξιοπιστίας.

Η ανάλυση των δεδομένων του χρόνου αποτυχίας συνήθως μας οδηγεί στην αντιμετώπιση ενός εκ των τριών παρακάτω προβλημάτων. Αυτά είναι:

α) οι εκτιμήτριες των συναρτήσεων επιβίωσης,

β) η σύγκριση των θεραπειών ή των συναρτήσεων επιβίωσης και

γ) η εκτίμηση της επίδρασης των συμμεταβλητών ή η εξάρτηση από το χρόνο αποτυχίας των συμμεταβλητών.

Στη συνέχεια, αναφέρουμε μεθόδους οι οποίες μπορούν να χρησιμοποιηθούν για να αντιμετωπίσουν τέτοιου είδους προβλήματα αποκομμένων δεδομένων σε διάστημα (interval-censored data). Η συνάρτηση επιβίωσης, η οποία δίνεται παρακάτω, ορίζει την πιθανότητα ο χρόνος αποτυχίας να έπεται ή να προηγείται μιας συγκεκριμένης χρονικής στιγμής (η διαφορά τους ορίζεται με τα σύμβολα > ή <) και παίζει μεγάλο ρόλο στην ανάλυση του χρόνου αποτυχίας.

Για διάφορους λόγους, απαιτούνται συγκεκριμένες μέθοδοι για τη διαχείριση των δεδομένων χρόνου αποτυχίας. Ένας λόγος, ο οποίος αποτελεί και εξέχον χαρακτηριστικό ως προς το διαχωρισμό της ανάλυσης των δεδομένων του χρόνου αποτυχίας από άλλα στατιστικά πεδία, είναι η ύπαρξη των αποκομμένων δεδομένων, όπως τα δεξιά αποκομμένα (right censoring), στα οποία και θα αναφερθούμε παρακάτω.

Οι μηχανισμοί αποκομμένων δεδομένων μπορούν να είναι αρκετά περίπλοκοι και επιπλέον να χρήζουν συγκεκριμένων μεθόδων χρησιμοποίησης. Οι προσιτές μέθοδοι για άλλου είδους δεδομένων είναι συνήθως απλές αλλά όχι κατάλληλες για αποκομμένα δεδομένα (censored data). Truncation (κολόβωση) είναι ένα ακόμα χαρακτηριστικό μερικών δεδομένων χρόνου αποτυχίας το οποίο απαιτεί ειδικό τρόπο διαχείρισης. Θα επικεντρωθούμε στα αποκομμένα δεδομένα (και θα συζητήσουμε μόνο μερικά είδη κολοβών δεδομένων). Πριν προχωρήσουμε στην ανάλυση των όσων αναφέρθηκαν παραπάνω θα περιγράψουμε με συντομία δύο παραδείγματα δεδομένων με χρόνο αποτυχίας και τα χαρακτηριστικά τους.

Παράδειγμα 1.1

Πίνακας 1: Χρόνος μέχρι την υποτροπή σε βδομάδες για ασθενείς με οξεία λευχαιμία

Ομάδες	Χρόνοι επιβίωσης σε βδομάδες
6-MP	6 6 6 6* 7 9* 10 10* 11* 13 16 17* 19* 20* 22 23 25* 32* 32* 34* 35*
Placebo	1 1 2 2 3 4 4 5 5 8 8 8 8 11 11 12 12 15 17 22 23

1.1.1 Χρόνος μέχρι την υποτροπή ασθενών με οξεία λευχαιμία

Στον Πίνακα 1, αναπαραγόμενος από τους Freireich et al. (1963) και Gehan (1965), παρουσιάζεται ένα τυπικό σύνολο δεδομένων χρόνου αποτυχίας προερχόμενο από μία κλινική δοκιμή οξείας λευχαιμίας κάποιων ασθενών. Η μελέτη έγινε κατά τη διάρκεια ενός χρόνου και οι ασθενείς εντάσσονταν σε αυτή σε διαφορετικούς χρόνους. Η σύγκριση των δύο θεραπειών είναι ο πρωταρχικός σκοπός αυτής της ανάλυσης, δηλαδή εάν οι ασθενείς που τους χορηγείτο η θεραπεία 6-MP είχαν σημαντικά μεγαλύτερους χρόνους μέχρι την υποτροπή από ότι είχαν οι ασθενείς που ακολουθούσαν τη θεραπεία με placebo.

Για τις παρατηρηθείσες πληροφορίες που δίνονται στον Πίνακα 1, οι αριθμοί με αστερίσκο αντιπροσωπεύουν αποκομμένους χρόνους (censored times) ή αποκομμένους χρόνους που αφορούν στην υποτροπή της ασθένειας (censored remission times). Η διαφορά τους έγκειται στο γεγονός ότι μία τέτοια παρατήρηση είναι η ποσότητα του χρόνου από τη στιγμή που ο ασθενής εισέρχεται στη μελέτη έως το τέλος αυτής. Αυτοί οι χρόνοι που αφορούν στην υποτροπή της ασθένειας (remission times) ήταν αποκομμένοι γιατί οι ασθενείς ήταν ακόμα στο στάδιο της απαλλαγής στο τέλος της δοκιμής και έτσι για τους ελαττωμένους χρόνους το μόνο που γνωρίζαμε ήταν ότι ήταν μεγαλύτεροι από ότι οι αποκομμένοι χρόνοι. Για τους άλλους ασθενείς, οι ελαττωμένοι χρόνοι είχαν παρατηρηθεί ακριβώς. Αυτή η κατάσταση παρατηρείται, συχνά, σε μελέτες χρόνου διακοπής (failure time studies) και τα αποτελέσματα των δεδομένων συνήθως αναφέρονται σε δεξιά αποκομμένα δεδομένα χρόνου διακοπής.

Να παρατηρήσουμε επιπλέον ότι για τη συγκεκριμένη σύγκριση των δύο αγωγών/θεραπειών ένα απλό t-test δε μπορεί να έχει εφαρμογή, γιατί δε μπορεί να διαχειριστεί τους αποκομμένους χρόνους που αφορούν στην ελάττωση της ασθένειας (censored remission times) και σίγουρα το να αφαιρεθούν από την ανάλυση αυτοί οι χρόνοι δεν είναι το επιθυμητό.

Παράδειγμα 1.2

Πίνακας 2: Ηλικίες σε χρόνια έως την πρώτη χρήση ναρκωτικού (μαριχουάνα)

Ηλικία	**Αριθμός ακριβών παρατηρήσεων (exact obs.)**	**Αριθμός αριστερά αποκομμένων παρατηρήσεων**	**Αριθμός δεξιά αποκομμένων παρατηρήσεων**
10	4	0	0
11	12	0	0
12	19	2	0
13	24	15	1
14	20	24	2
15	13	18	3
16	3	14	2
17	1	6	3
18	0	0	1
>18	4	0	0

1.1.2 Χρόνοι μέχρι την πρώτη χρήση ενός ναρκωτικού (μαριχουάνα)

Οι Turnbull και Weiss (1978) αναφέρθηκαν σε μία ομάδα δεδομένων από χρόνους αποτυχίας (failure time data) από μία μελέτη που έγινε για τη χρήση ναρκωτικών από μαθητές γυμνασίου και τα οποία μπορούμε να δούμε στον Πίνακα 2. Στη μελέτη συμμετείχαν 191 αγόρια από την Καλιφόρνια και ρωτήθηκαν 'Πότε ήταν η πρώτη φορά που χρησιμοποίησαν το συγκεκριμένο ναρκωτικό;' Όπως ήταν αναμενόμενο, κάποια αγόρια θυμόντουσαν την ακριβή ηλικία που είχαν όταν έκαναν για πρώτη φορά χρήση ενώ κάποια άλλα δεν μπορούσαν να θυμηθούν. Επιπλέον, υπήρχαν και αγόρια που δεν είχαν κάνει χρήση αυτού του ναρκωτικού ποτέ. Έχοντας, λοιπόν, αυτές τις απαντήσεις είχαμε και τους παρακάτω τρεις τύπους παρατηρήσεων. Για την πρώτη περίπτωση, η ηλικία ήταν ακριβής. Για τη δεύτερη και την τρίτη περίπτωση γνωρίζαμε μόνο πως η ηλικία ήταν μικρότερη ή μεγαλύτερη από την τρέχουσα ηλικία των αγοριών, και αυτού του είδους τα δεδομένα συνήθως αναφέρονται ως αριστερά αποκομμένες ή δεξιά αποκομμένες παρατηρήσεις, αντίστοιχα.

Για τα συγκεκριμένα δεδομένα, ένα ερώτημα που προκύπτει και είναι ενδιαφέρον είναι η εκτίμηση της πιθανότητας να έχει γίνει χρήση του ναρκωτικού σε μία συγκεκριμένη ηλικία από μαθητές γυμνασίου. Είναι προφανές ότι μία απλή εμπειρική εκτίμηση δεν θα ήταν κατάλληλη, εκτός εάν κάποιος παρέλειπε κάποιες από τις αριστερά και τις δεξιά αποκομμένες παρατηρήσεις. Μεταξύ άλλων, να αναφέρουμε ότι οι Klein και Moeschberger (2003), Turnbull και Weiss (1978) ανέλυσαν τέτοια ομαδοποιημένα δεδομένα.

1.1.3 Αποκοπή και Κολόβωση

Όπως αναφέραμε και προηγουμένως, η αποκοπή είναι ένα από τα μοναδικά χαρακτηριστικά των δεδομένων του χρόνου αποτυχίας. Με την αποκοπή, εννοούμε ότι μία παρατήρηση ως προς ένα χρόνο επιβίωσης που μας ενδιαφέρει είναι ατελής, υπό την έννοια ότι ο χρόνος επιβίωσης γνωρίζουμε ότι έχει συμβεί σε ένα συγκεκριμένο εύρος αντί να έχει παρατηρηθεί ακριβώς. Να σημειώσουμε ότι τα αποκομμένα δεδομένα διαφέρουν από τα μη καταγεγραμμένα δεδομένα (άνευ κλινικής σημασίας) (missing data) γιατί μπορούν να παρέχουν μερική πληροφορία, σε αντίθεση με τις παρατηρήσεις που λείπουν οι οποίες δεν παρέχουν καμία πληροφορία σχετικά με τη μεταβλητή ενδιαφέροντος. Διαφορετικά είδη αποκοπής εμφανίζονται στην πράξη, αλλά αυτό στο οποίο επικεντρωνόμαστε είναι η δεξιά αποκοπή (right censoring).

Με τον όρο δεξιά αποκοπή εννοούμε ότι ο χρόνος αποτυχίας που μας ενδιαφέρει έχει παρατηρηθεί είτε ακριβώς είτε είναι μεγαλύτερος από ένα χρόνο αποκοπής (censoring time). Μία τυπική περίπτωση δεξιά αποκομμένων παρατηρήσεων είναι αυτή κατά την οποία η μελέτη επιβίωσης πρέπει να τελειώσει εξαιτίας, π.χ. περιορισμών του χρόνου ή περιορισμένων πόρων. Σε αυτή την περίπτωση, για άτομα των οποίων τα γεγονότα επιβίωσης (survival events) δεν έχουν ακόμα συμβεί ως το τέλος της μελέτης, οι χρόνοι επιβίωσης όπως είναι αναμενόμενο δεν έχουν παρατηρηθεί ακριβώς αλλά γνωρίζουμε ότι είναι μεγαλύτεροι από το τέλος χρόνου της μελέτης, δηλαδή είναι δεξιά αποκομμένοι. Για τα άτομα που έχουν ήδη αποτύχει, έως το τέλος της μελέτης, οι χρόνοι διακοπής είναι ακριβείς. Φυσικά, ο χρόνος που ολοκληρώνεται η μελέτη μπορεί να διαφέρει για διαφορετικά άτομα μερικά από τα οποία μπορεί να αποσυρθούν από τη μελέτη πριν από το τέλος της για διάφορους λόγους. Σε ένα γενικότερο πλαίσιο, που είναι πιο κατάλληλο για περισσότερες εφαρμογές, για κάθε άτομο, υπάρχει μία αποκομμένη μεταβλητή που αντιπροσωπεύει το δεξιά αποκομμένο χρόνο. Εάν η μεταβλητή επιβίωσης είναι μικρότερη από την αποκομμένη μεταβλητή, η παρατήρηση είναι ακριβής αλλιώς είναι δεξιά αποκομμένη. Αυτό συνήθως αναφέρεται ως μοντέλο τυχαίας αποκοπής (random censored model).

Είναι προφανές ότι για να καταλάβει κανείς τον τρόπο που η δεξιά αποκοπή συμβαίνει θα πρέπει να αναλύσει τα δεξιά αποκομμένα δεδομένα του χρόνου διακοπής (failure time data) σωστά. Για να απλοποιήσουμε την ανάλυση, συνήθως θεωρούμε ένα ανεξάρτητο δεξιά αποκομμένο μηχανισμό (right censoring mechanism). Με αυτό τον τρόπο εννοούμε ότι ο ρυθμός διακοπής ή διακινδύνευσης (failure rate or hazard) είναι το ίδιο για τα άτομα που είναι ακόμα στη μελέτη και για τα άτομα που έχουν αποκοπεί. Πιο συγκεκριμένα, υπό τον ανεξάρτητο δεξιά αποκομμένο μηχανισμό, έχουμε ότι:

$$\lim_{\Delta t \to 0^+} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t} = \lim_{\Delta t \to 0^+} \frac{P(t \le T < t + \Delta t \mid T \ge t, Y(t) = 1)}{\Delta t}$$

(Kalbfleisch και Prentice, 2002), όπου το T δηλώνει τη μεταβλητή επιβίωσης που μας ενδιαφέρει και με Y(t)=1 εννοούμε ότι το γεγονός δεν έχει συμβεί ούτε έχει αποκοπεί πριν από τη χρονική στιγμή t. Υπό το μοντέλο τυχαίας αποκοπής (random censorship model) η παραπάνω συνθήκη μπορεί να γραφεί και ως εξής:

$$\lim_{\Delta t \to 0^+} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t} = \lim_{\Delta t \to 0^+} \frac{P(t \le T < t + \Delta t \mid T \ge t, A \ge 1)}{\Delta t}$$

όπου Α δηλώνει τη μεταβλητή αποκοπής.

Υπάρχουν διαφορετικοί τύποι αποκοπής από δεξιά, όπως επίσης και άλλοι τύποι αποκοπής παρατηρήσεων. Για παράδειγμα, ο μηχανισμός της αποκοπής που σταματάει τη μελέτη σε ένα σταθερό σημείο για όλες τις μονάδες συνήθως αναφέρεται ως Τύπος 1 (Type I) αποκοπής. Ο Τύπος αποκοπής 2 (Type II) εννοεί ότι η μελέτη σταματάει εάν ένας προκαθορισμένος αριθμός ατόμων που συμμετείχαν στη μελέτη έχει αποτύχει. Για την αποκοπή σε διάστημα θα μιλήσουμε παρακάτω.

Η κολόβωση (truncation) αναφέρεται σε περιπτώσεις όπου ένα άτομο συμπεριλαμβάνεται στη μελέτη μόνο εάν ο αντίστοιχος χρόνος διακοπής ικανοποιεί συγκεκριμένες συνθήκες. Ένα σύνηθες και απλό παράδειγμα το οποίο αποδίδει τον κολοβό χρόνο διακοπής των δεδομένων είναι μία μελέτη κοορτής (cohort study) στην οποία τα άτομα περιλαμβάνονται στη μελέτη μόνο εάν αντιμετωπίζουν κάποια αρχική εκδήλωση της ασθένειας πριν από το γεγονός της επιβίωσης (survival event). Σε αυτή την περίπτωση για όλα τα άτομα της μελέτης, οι χρόνοι διακοπής είναι μεγαλύτεροι από τους χρόνους τέλεσης του συμβάντος της αρχικής εκδήλωσης. Αυτός ο τύπος κολόβωσης (truncation) συνήθως αναφέρεται ως αριστερή κολόβωση. Η ανεξάρτητη κολόβωση μπορεί να ορισθεί όμοια με την ανεξάρτητα δεξιά αποκοπή (μη-πληροφοριακή) και συνήθως χρησιμοποιείται για την ανάλυση των κολοβών χρόνων διακοπής των δεδομένων (truncated failure time data). (Kalbfleisch και Prentice, 2002; Lawless, 2003).

1.2 Χρόνοι αποτυχίας αποκομμένοι σε διάστημα

Όπως αναφέραμε και προηγουμένως, δεδομένα που αναφέρονται σε χρόνους αποτυχίας συμβαίνουν με πολλούς τρόπους και σε αρκετά πεδία, και υπάρχει ένας αριθμός λόγων γιατί ειδικές μέθοδοι χρειάζονται για την ανάλυσή τους. Θα επικεντρωθούμε στην αποκοπή σε διάστημα, που είναι πιο απαιτητική από τη δεξιά αποκοπή και για τέτοιου είδους δεδομένα οι μέθοδοι που έχουν αναπτυχθεί για τη δεξιά αποκοπή δεν μπορούν να εφαρμοσθούν γενικώς.

Με την αποκοπή σε διάστημα εννοούμε, ότι οι μονάδες της μελέτης ή οι διαδικασίες των χρόνων αποτυχίας (failure time processes) που μας ενδιαφέρουν δεν είναι υπό συνεχή παρακολούθηση. Ως συνέπεια, ο χρόνος αποτυχίας ή επιβίωσης δεν είναι πάντα με ακρίβεια παρατηρηθείς ή δεξιά αποκομμένος (right-censored). Για μία παρατήρηση που βρίσκεται σε διάστημα αποκομμένων δεδομένων, το μόνο που μπορούμε να γνωρίζουμε είναι ένα "παράθυρο" το οποίο είναι ένα διάστημα

μέσα στο οποίο το γεγονός της επιβίωσης (survival event) έχει πραγματοποιηθεί. Ο ακριβής ή δεξιά αποκομμένος χρόνος αποτυχίας μπορεί να θεωρηθεί ως μία ειδική περίπτωση των χρόνων αποτυχίας σε διάστημα αποκομμένων δεδομένων, όπου στις συγκεκριμένες περιπτώσεις το διάστημα περιορίζεται σε ένα και μοναδικό σημείο ή σε διάστημα ανοιχτό από δεξιά. Γενικότερα, θα μπορούσαμε να ορίσουμε μία παρατήρηση αποκομμένη σε ένα διάστημα σαν μία ένωση πολλαπλών μη επικαλυπτόμενων "παραθύρων" ή διαστημάτων (Turnbull, 1976).

Ένα τυπικό παράδειγμα από αποκομμένα δεδομένα σε διάστημα εμφανίζεται στις ιατρικές ή στις μελέτες υγείας που συνεπάγονται περιοδική παρακολούθηση πολλών κλινικών δοκιμών και διαχρονικών μελετών που εμπίπτουν σε αυτή την κατηγορία. Σε τέτοιες περιπτώσεις αποκομμένα δεδομένα σε διάστημα μπορούν να δημιουργηθούν με πολλούς τρόπους. Παραδείγματος χάριν, ένα άτομο μπορεί να χάσει μία ή περισσότερες προγραμματισμένες επισκέψεις σε κλινική για να παρατηρηθούν πιθανές αλλαγές στην πορεία της υγείας του. Ένα ακόμα παράδειγμα είναι η επίσκεψη στην κλινική από τα άτομα σε χρόνους που είναι βολικοί για τους ίδιους από ότι είναι οι προκαθορισμένες επισκέψεις τους. Στις δύο παραπάνω περιπτώσεις τα δεδομένα της αλλαγής της κατάστασης της υγείας των ατόμων είναι αποκομμένα δεδομένα σε διάστημα. Ακόμα όμως και αν όλα τα άτομα που υπόκεινται σε μία μελέτη ακολουθούν ακριβώς τους προκαθορισμένους χρόνους παρακολούθησής τους δε μπορούμε να παρατηρήσουμε τον ακριβή χρόνο αλλαγής της κατάστασης του ατόμου με δεδομένο ότι είναι μία συνεχής μεταβλητή.

1.3 Είδη διαστημάτων αποκοπής και οι σχηματισμοί τους

Έστω T μία μη αρνητική τυχαία μεταβλητή που αντιπροσωπεύει το χρόνο αποτυχίας ενός ατόμου σε μία μελέτη χρόνου αποτυχίας, δηλαδή τέλεσης του επιθυμητού συμβάντος (failure time study). Μία παρατήρηση στο T είναι αποκομμένη σε διάστημα εάν αντί να παρατηρήσουμε το T ακριβώς, παρατηρήσουμε ένα διάστημα (L,U] τέτοιο ώστε $T \in (L,U]$. (1.3.0.1)

Ο χρόνος T_i (i=1,..., n) λοιπόν της τέλεσης του γεγονότος είναι γνωστό (όποτε και αν συμβεί) ότι βρίσκεται μεταξύ επισκέψεων πχ. μεταξύ της επίσκεψης στο χρόνο και L_i της επίσκεψης στο χρόνο U_i. Η τέλεση του γεγονότος έχει συμβεί στο διάστημα $(L_i,U_i]$ με $L_i < T_i \le U_i$. Επιπλέον, εάν το γεγονός

συμβεί ακριβώς τη στιγμή της επίσκεψης, το οποίο έχει πολύ μικρή πιθανότητα τέλεσης αλλά μπορεί να συμβεί, τότε έχουμε ακριβή χρόνο επιβίωσης. Σε αυτή την περίπτωση υποθέτουμε ότι $T_i = L_i = U_i$.

Απ΄την άλλη πλευρά, για τα άτομα στα οποία το γεγονός τέλεσης δεν έχει συμβεί μέχρι και την τελευταία επίσκεψη αλλά μπορεί να συμβεί οποιαδήποτε στιγμή από εκεί και μετά, ο χρόνος αποτυχίας θα είναι δεξιά αποκομμένος (right censored). Γι΄ αυτό το λόγο υποθέτουμε ότι ο χρόνος T_i μπορεί να συμβεί μέσα στο διάστημα (L_i, ∞) με L_i να ισούται με την περίοδο του χρόνου από την αρχή της μελέτης έως την τελευταία επίσκεψη και με το αντίστοιχο $U_i = \infty$.

Ομοίως, για τα άτομα στα οποία το γεγονός τέλεσης έχει συμβεί πριν από την πρώτη επίσκεψη έτσι ώστε να έχουμε αριστερά αποκομμένα δεδομένα υποθέτουμε ότι ο χρόνος T_i βρίσκεται στο διάστημα $(0, L_i]$ με $L_i = 0$ αντιπροσωπεύοντας την αρχή της μελέτης με U_i να είναι ο χρόνος από την αρχή της μελέτης μέχρι την πρώτη επίσκεψη.

Να σημειώσουμε σε αυτό το σημείο ότι από όσα έχουμε πει μέχρι τώρα οι ακριβείς χρόνοι επιβίωσης όπως επίσης και τα δεξιά καθώς και τα αριστερά αποκομμένα δεδομένα είναι όλα ειδικές περιπτώσεις των δεδομένων επιβίωσης σε διάστημα με $L_i = U_i$ για συγκεκριμένους χρόνους, με $U_i = \infty$ για δεξιά αποκομμένες και $L_i = 0$ για αριστερά αποκομμένες παρατηρήσεις. Μπορούμε λοιπόν να πούμε ότι τα δεδομένα επιβίωσης σε διάστημα γενικεύονται σε οποιαδήποτε περίπτωση με συνδυασμό των χρόνων επιβίωσης (ακριβείς ή σε διάστημα) και σε δεξιά και σε αριστερά αποκομμένα που μπορούν να συμβούν σε μελέτες επιβίωσης.

1.3.1 Περίπτωση Ι αποκομμένων δεδομένων χρόνου αποτυχίας σε διάστημα

Ο όρος Περίπτωση Ι χρησιμοποιείται ευρέως όταν αναφερόμαστε σε αποκομμένα δεδομένα χρόνου αποτυχίας σε διάστημα στο οποίο όλα τα παρατηρηθέντα διαστήματα εμπεριέχουν είτε μηδενικό χρόνο είτε άπειρο (Groeneboom και Wellner, 1992; Huang, 1996). Με άλλα λόγια η παρατήρηση σε κάθε ατομικό χρόνο αποτυχίας είναι είτε αριστερά είτε δεξιά αποκομμένος τον οποίο συμβολίζουμε με L=0 είτε U= ∞. Την Περίπτωση Ι αποκομμένων δεδομένων σε διάστημα έχουμε όταν κάθε μονάδα της μελέτης παρατηρείται μόνο μία φορά και η μόνη παρατηρούμενη πληροφορία για το γεγονός επιβίωσης που μας ενδιαφέρει είναι εάν το γεγονός έχει συμβεί όχι νωρίτερα από τον παρατηρούμενο χρόνο. Αντί

των διαστημάτων στην (1.3.0.1) ένας πιο βολικός τρόπος παρουσίασης της Περίπτωσης Ι είναι $\{C,\ \delta=I(T \le V)\}$, όπου C δηλώνει τον παρατηρηθέντα χρόνο και Ι είναι η δείκτρια συνάρτηση.

Να σημειώσουμε ότι η Περίπτωση Ι διαφέρει από τη δεξιά αποκοπή ή την αριστερή, οι οποίες συνήθως περιλαμβάνουν μερικούς χρόνους αποτυχίας που έχουν παρατηρηθεί ακριβώς. Επιπλέον, τα αποκομμένα δεδομένα της Περίπτωσης Ι σε διάστημα συχνά αναφέρονται και ως δεδομένα τρέχουσας κατάστασης (current status data), ένας όρος προερχόμενος από δημογραφικές μελέτες.

1.3.2 Περίπτωση ΙΙ αποκομμένων δεδομένων χρόνου αποτυχίας σε διάστημα

Αποκομμένα δεδομένα σε διάστημα που περιλαμβάνουν τουλάχιστον ένα διάστημα (L,U] και με τα δύο άκρα να ανήκουν στο $(0,\infty)$ συνήθως αναφέρεται ως γενική ή Περίπτωση ΙΙ περίπτωση (Groeneboom και Wellner, 1992; Huang και Wellner, 1997; Sun, 1998, 2005). Με άλλα λόγια, η Περίπτωση ΙΙ αναφέρεται σε αποκομμένα δεδομένα διαστήματος που περιλαμβάνουν μερικά πεπερασμένα διαστήματα μακριά από το μηδέν. Ένας διαφορετικός τρόπος να παρουσιάσουμε τις παρατηρήσεις της περίπτωσης αυτής είναι :

$$\{R,\ V,\ \delta_1 = I(T \le R),\ \delta_2 = I(R < T \le V),\ \delta_3 = 1-\delta_1-\delta_2\} \qquad (1.3.2.1)$$

υποθέτοντας ότι κάθε μονάδα έχει παρατηρηθεί δύο φορές, όπου R και V είναι δύο τυχαίες μεταβλητές που ικανοποιούν την ανίσωση $R \le V$ με πιθανότητα 1. Αυτός ο σχηματισμός είναι βολικός και συχνά χρησιμοποιείται, για παράδειγμα, στη θεωρητική διερεύνηση μίας συμπερασματικής διαδικασίας.

Να σημειώσουμε ότι αποδεχόμενοι την ισότητα R=V=C η Περίπτωση Ι μπορεί να περιγραφεί από την (1.3.2.1). Ο Yu (2000) γενίκευσε αυτό το σχηματισμό για να συμπεριλάβει ακριβείς παρατηρήσεις. Μία ακόμα γενίκευση του σχηματισμού (1.3.2.1) γίνεται αν υποθέσουμε ότι υπάρχει μία ομάδα από παρατηρήσεις χρονικών σημείων, έστω $R_1 \le R_2 \le ... \le R_k$ για κάθε μονάδα μελέτης, όπου k είναι ένας τυχαίος ακέραιος. Η παρατηρηθείσα πληροφορία τότε έχει τη μορφή :

$$\{K,\ R_j,\ \delta_1 = I(R_j < T \le R_j),\ j = 1,...,\ k\} \qquad (1.3.2.2)$$

όπου $R_0 = 0$. Αυτός ο σχηματισμός ή ο τύπος των δεδομένων χρόνου αποτυχίας συνήθως αναφέρεται ως περίπτωση Κ ή μεικτή περίπτωση αποκομμένων δεδομένων σε διάστημα (Schick και Yu, 2000;

Wellner, 1995). Είναι προφανές ότι ο παραπάνω σχηματισμός περιλαμβάνει την απεικόνιση της (1.3.2.1) ως μία ειδική περίπτωση και παρέχει μία φυσική παρουσίαση των αποκομμένων δεδομένων χρόνου αποτυχίας σε διάστημα προερχόμενα από διαχρονικές μελέτες με περιοδική παρακολούθηση.

Και οι τρείς απεικονίσεις, από την (1.3.0.1) έως την (1.3.2.2) δημιουργούν την ίδια συνάρτηση πιθανοφάνειας. Επισημαίνουμε ότι, παρόλο που και οι δύο απεικονίσεις (1.3.2.1) και (1.3.2.2) φαίνονται λογικές δεν είναι σύνηθες να έχουμε αποκομμένα δεδομένα σε διάστημα που έχουν συλλεχθεί ή δοθεί σε τέτοιους σχηματισμούς στην πράξη. Ωστόσο, είναι πολύ πιο εύκολο και πιο φυσικό να κάνουμε υποθέσεις ως προς την ανεξαρτησία του T για αυτές από ότι ως προς την απεικόνιση (1.3.0.1), η οποία συχνά χρήζει προσδιορισμό των ασυμπτωτικών ιδιοτήτων των συμπερασματικών διαδικασιών. Για δεδομένα υπό την απεικόνιση (1.3.2.1) ή (1.3.2.2) ο καθένας μπορεί εύκολα να συμπεριλάβει τα ανταποκρινόμενα δεδομένα με την απεικόνιση (1.3.0.1). Από τη άλλη πλευρά, είναι φαινομενικώς αδύνατο να μετασχηματίσουμε την απεικόνιση (1.3.0.1) στην (1.3.2.2) χωρίς επιπλέον πληροφορία για τη παρατηρούμενη διαδικασία και δεν είναι άμεσα προς μετασχηματισμό οι παρατηρήσεις που προέρχονται από την απεικόνιση (1.3.0.1) σε αυτές της απεικόνισης (1.3.2.1).

1.3.3 Χρόνοι αποτυχίας διπλά αποκομμένων δεδομένων

Θεωρούμε μία μελέτη επιβίωσης που περιλαμβάνει δύο σχετιζόμενα γεγονότα. Έστω X και S οι χρόνοι της τέλεσης των δύο γεγονότων με $X \le S$. Ορίζουμε $T=S-X$ και υποθέτουμε ότι T είναι ο χρόνος επιβίωσης που μας ενδιαφέρει. Με τον όρο χρόνοι αποτυχίας διπλά αποκομμένων δεδομένων εννοούμε ότι οι παρατηρήσεις και στο X και στο S είναι αποκομμένες σε διάστημα (De Gruttola και Lagakos, 1989; Sun, 2004). Ειδικότερα, υποθέτουμε ότι αντί να παρατηρήσουμε τα X και S ακριβώς παρατηρούμε δύο διαστήματα (L,R] και (U,V] τέτοια ώστε $X \in (L, R]$, $S \in (U, V]$, $L \le R$ και $U \le V$ με πιθανότητα 1. Με άλλα λόγια, οι παρατηρήσεις στο T είναι διπλά αποκομμένες.

Η ειδική περίπτωση για τα διπλά αποκομμένα δεδομένα (doubly censored data) στην οποία το S είναι μόνο δεξιά αποκομμένο συμβαίνει συχνά, και σε αυτή την περίπτωση έχουμε είτε U=V ή $V=\infty$.

Ένας άλλος σχηματισμός για αυτή τη συγκεκριμένη περίπτωση που μπορεί να είναι πιο ρεαλιστικός είναι να υποθέσουμε ότι υπάρχει μία αποκομμένη μεταβλητή C, η οποία συνήθως θεωρούμε ότι είναι

ανεξάρτητη του S. Η παρατήρηση στο S τότε αποτελείται από το S* = {S,C} και δ= I(S*= S) όπου I είναι η δείκτρια συνάρτηση όπως προηγουμένως.

Μπορούμε να δούμε χρόνους αποτυχίας διπλά αποκομμένων δεδομένων σε μελέτες των οποίων η ασθένεια είναι σε εξέλιξη, όπου τα δύο γεγονότα θα μπορούσαν να αντιπροσωπεύουν τη μόλυνση και τη μετέπειτα εμφάνιση μίας συγκεκριμένης ασθένειας. Σε αυτές τις περιπτώσεις, λοιπόν διπλά αποκομμένες παρατηρήσεις συμβαίνουν κυρίως εξαιτίας της φύσης της ασθένειας ή της δομής του σχεδιασμού της μελέτης. Έστω ότι X και S αντιπροσωπεύουν τη μόλυνση από τον ιό HIV και τους χρόνους διάγνωσης του AIDS αντίστοιχα, και T είναι ο λανθάνων χρόνος του AIDS.

Χρόνοι αποτυχίας διπλά αποκομμένων δεδομένων περιλαμβάνουν ως ειδικές περιπτώσεις δεξιά αποκομμένους και αποκομμένους χρόνους αποτυχίας σε διάστημα. Για παράδειγμα, μειώνουν τα αποκομμένα δεδομένα σε διάστημα εάν ο χρόνος της τέλεσης του πρώτου γεγονότος X, μπορεί να παρατηρηθεί ακριβώς (L=R). Επιπλέον, εάν η παρατήρηση του χρόνου της τέλεσης του μεταγενέστερου γεγονότος S, είναι ακριβής ή δεξιά αποκομμένη, τότε έχουμε μία δεξιά αποκομμένη παρατήρηση στο T. Να σημειώσουμε ότι για διπλά αποκομμένα δεδομένα, εάν το X έχει παρατηρηθεί ακριβώς, μπορούμε να εξάγουμε συμπεράσματα για το T θεωρώντας ότι το X= 0 κάτι που συμβαίνει στην ανάλυση των δεδομένων που αφορούν σε χρόνους αποτυχίας.

1.4 Έννοιες και μερικά μοντέλα παλινδρόμησης

Έστω T μία μη αρνητική τυχαία μεταβλητή που αντιπροσωπεύει τους χρόνους αποτυχίας μιας μονάδας η οποία είναι η μεταβλητή επιβίωσης που μας ενδιαφέρει. Για την εξαγωγή συμπερασμάτων ως προς την T, η συνάρτηση επιβίωσης (survival function) και η συνάρτηση διακινδύνευσης (hazard function) είναι ιδιαιτέρως χρήσιμες για μοντελοποίηση. Η συνάρτηση επιβίωσης της T ορίζεται ως η πιθανότητα η T να υπερβαίνει μία τιμή t. Έστω, λοιπόν, $S(t)$ η συνάρτηση επιβίωσης της T. Τότε έχουμε $S(t) = P(T > t),\ 0 < t < \infty$.

Η συνάρτηση διακινδύνευσης ορίζεται διαφορετικά για συνεχείς και διακριτές μεταβλητές επιβίωσης, παρακάτω θα δώσουμε τον ορισμό της συνεχούς. Οι συναρτήσεις πυκνότητας πιθανότητας και κατανομής συχνά χρησιμοποιούνται και στην ανάλυση επιβίωσης αν και όχι τόσο συχνά όσο οι συναρτήσεις επιβίωσης και διακινδύνευσης.

Επιπλέον, για να επανεξετάσουμε αυτές τις συναρτήσεις μαζί με τις σχέσεις που τις συνδέουν, σε αυτή την ενότητα θα περιγράψουμε ημι-παραμετρικά μοντέλα παλινδρόμησης που συνήθως χρησιμοποιούμε στην ανάλυση επιβίωσης. Αυτό περιλαμβάνει ένα μοντέλο αναλογικής διακινδύνευσης όπως το μοντέλο του Cox, το μοντέλο αναλογικής σχετικής πιθανότητας (proportional odds model), το μοντέλο προσθετικής διακινδύνευσης (additive hazard model) και άλλα.

Εμείς θα αναφερθούμε στο μοντέλο αναλογικής διακινδύνευσης, αφού πρώτα παρουσιάσουμε τα βασικά ως προς τις συνεχείς μεταβλητές επιβίωσης.

1.4.1 Συνεχείς μεταβλητές επιβίωσης

Υποθέτουμε ότι T είναι μία συνεχής μεταβλητή και επιπλέον ότι η συνάρτηση πυκνότητας πιθανότητας f(t) υπάρχει. Εξ ορισμού, είναι εύκολο να δει κανείς ότι η συνάρτηση πυκνότητας πιθανότητας και η συνάρτηση επιβίωσης ικανοποιούν τη σχέση:

$$f(t) = -\frac{dS(t)}{dt} \quad \text{με} \quad S(t) = \int_t^{\infty} f(s)ds$$

Η συνάρτηση διακινδύνευσης της T τη χρονική στιγμή t ορίζεται ως :

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{\mathrm{P}(t \le T < t + \Delta t \mid T \ge t)}{\Delta t}$$

αντιπροσωπεύει τη στιγμιαία πιθανότητα όπου μία μονάδα αποτυγχάνει τη χρονική στιγμή t δοθέντος ότι η συγκεκριμένη μονάδα δεν έχει αποτύχει πριν από αυτή τη στιγμή. Οι συναρτήσεις επιβίωσης, πυκνότητας πιθανότητας και διακινδύνευσης έχουν μονοσήμαντη σχέση μεταξύ τους (one-to-one relationship). Ειδικότερα, δοθέντος της συνάρτησης πυκνότητας πιθανότητας ή επιβίωσης, έχουμε ότι:

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d \log S(t)}{dt}$$

από την άλλη πλευρά, μπορεί να αποδειχθεί ότι $S(t) = (-\int_0^t f(s)ds) = \exp(\Lambda(t))$ και $f(t) = \lambda(t)\exp(-\Lambda(t))$ όπου $\Lambda(t) = \int_0^t f(s)ds$ η σωρευτική συνάρτηση διακινδύνευσης της T.

1.4.2 Το μοντέλο αναλογικής διακινδύνευσης

Έστω Z ένα διάνυσμα από συμμεταβλητές που περιλαμβάνει, παραδείγματος χάριν, ένα δείκτη θεραπείας, ηλικίας και φύλου. Όπως επισημάναμε νωρίτερα, μία ανάλυση παλινδρόμησης μας παρέχει μία εκτίμηση των επιδράσεων των συμμεταβλητών στους χρόνους διακοπής, που είναι ένα από τα σημαντικά θέματα της ανάλυσης επιβίωσης. Για αυτό το λόγο, ένα μοντέλο παλινδρόμησης είναι συνήθως αναγκαίο για να καθορίσουμε πως οι συμμεταβλητές επηρεάζουν τους χρόνους αποτυχίας που μας ενδιαφέρουν. Ένα μοντέλο αναλογικής διακινδύνευσης (PH) υποθέτει ότι η συνάρτηση διακινδύνευσης της T έχει τον ακόλουθο τύπο :

$$\lambda(t;\mathbf{Z}) = \lambda_0(t)\exp(\mathbf{Z}'\boldsymbol{\beta}) \qquad (1.4.2.1)$$

δοθέντων των συμμεταβλητών **Z**.

Το **β** είναι ένα διάνυσμα παραμέτρων παλινδρόμησης και το $\lambda_0(t)$ καθορίζεται από το εκάστοτε παραμετρικό μοντέλο. Αυτό το μοντέλο ορίζει ειδικώς ότι οι συμμεταβλητές δρουν πολλαπλασιαστικά στη συνάρτηση διακινδύνευσης και προσαρμόζεται με τη μέθοδο μεγίστης πιθανοφάνειας.

Η εξίσωση (1.4.2.1) λέει ότι η αναλογία των συναρτήσεων διακινδύνευσης για δύο μονάδες με διαφορετικές συμμεταβλητές είναι συνεχής. Πιο συγκεκριμένα, για την περίπτωση του διπλού δείγματος όπου Z= 0 ή 1, έχουμε :

$$\frac{\lambda(t;Z=1)}{\lambda(t;Z=0)} = \exp(\beta)$$

Κάτω από το μοντέλο αναλογικής διακινδύνευσης (PH), η υπό συνθήκη πυκνότητα και οι εξισώσεις επιβίωσης της T δοθέντος του **Z** έχει τους ακόλουθους τύπους :

$$f(t;\mathbf{Z}) = \lambda_0(t)\exp(\mathbf{Z}'\boldsymbol{\beta})\exp[-\Lambda_0(t)\exp(\mathbf{Z}'\boldsymbol{\beta})]$$

και

$$S(t;\mathbf{Z}) = \exp[-\Lambda_0(t)\exp(\mathbf{Z}'\boldsymbol{\beta})] = [S_0(t)]^{\exp(\mathbf{Z}'\boldsymbol{\beta})}$$

όπου

$$\Lambda_0(t) = \int_0^t \lambda_0(s)ds \text{ και } S_0(t) = \exp[\int_0^t \lambda_0(s)ds]$$

είναι η αρχική ή βασική σωρευτική συνάρτηση διακινδύνευσης και η αρχική συνάρτηση επιβίωσης. Η υπό συνθήκη σωρευτική συνάρτηση διακινδύνευσης της T δοθέντος του Z έχει τον τύπο:

$$\Lambda(t;\mathbf{Z}) = \Lambda_0(t)\exp(\mathbf{Z}'\boldsymbol{\beta})$$

Αν στην εξίσωση (1.4.2.1) το $\lambda_0(t)$ είναι μία ακαθόριστη βασική συνάρτηση διακινδύνευσης θα έχουμε το μοντέλο του Cox (Cox, 1972) η ειδική περίπτωση του (1.4.2.1) και ίσως είναι το πιο σύνηθες μοντέλο παλινδρόμησης που χρησιμοποιούμε στην ανάλυση με δεδομένα χρόνους αποτυχίας. Ένας σημαντικός λόγος είναι η απλή και επαρκής συμπερασματική διαδικασία, η προσέγγιση της μερικής πιθανοφάνειας σε σχέση με την παράμετρο παλινδρόμησης **β** που είναι διαθέσιμη για δεδομένα δεξιά αποκομμένων χρόνων αποτυχίας.

Η προσέγγιση μερικής πιθανοφάνειας είχε προταθεί από τον Cox (1972, 1975) και έχει μελετηθεί από αρκετούς συγγραφείς. Είναι εν μέρει απλό γιατί η συνάρτηση μερικής πιθανοφάνειας που χρησιμοποιείται για τέτοια συμπεράσματα είναι μόνο μία συνάρτηση του **β** και επιπλέον, δε χρειάζεται να χρησιμοποιήσουμε την αρχική συνάρτηση διακινδύνευσης $\lambda_0(t)$.

Η προσέγγιση είναι επαρκής γιατί η εκτιμήτρια του **β** είναι ασυμπτωτικά ισοδύναμη με την εκτιμήτρια του **β** που δίνεται από την πλήρη συνάρτηση πιθανοφάνειας (full likelihood function). Επιπρόσθετα του Cox (1972, 1975), οι Cox και Oakes (1984) και Kalbfleisch και Prentice (2002) δίνουν άλλες αναφορές, παραπομπές για το μοντέλο (1.4.2.1) που συζητήσαμε νωρίτερα και τη χρησιμότητά του στην ανάλυση παλινδρόμησης των χρόνων αποτυχίας των δεξιά αποκομμένων δεδομένων.

Υπάρχουν πολλές γενικεύσεις του αναλογικού μοντέλου διακινδύνευσης. Μία από αυτές επιτρέπει στο **Z** να εξαρτάται από το χρόνο το οποίο θα μπορούσε να αντιστοιχεί στην περίπτωση, παραδείγματος χάριν όπου το **Z** θα εμπεριείχε το επίπεδο της μόλυνσης του αέρα ή την ποσότητα του χρόνου που ένα άτομο εξασκείται. Μία ακόμα γενίκευση επιτρέπει στην αρχική συνάρτηση διακινδύνευσης να είναι διαφορετική για άτομα από διαφορετικές υποομάδες ή υποπληθυσμούς. Για να είμαστε πιο σαφείς, ας

υποθέσουμε ότι ο πληθυσμός έχει διαιρεθεί σε κ-στρώματα (strata) και η συνάρτηση διακινδύνευσης της Τ για ένα άτομο από το j-οστό στρώμα έχει τον ακόλουθο τύπο :

$$\lambda(t;\mathbf{Z}) = \lambda_{0j}(t)\exp(\mathbf{Z}'\boldsymbol{\beta})$$

δοθέντων των συμμεταβλητών **Z**, j= 1, … , k.

Αυτό σημαίνει ότι, μπορεί η συνάρτηση διακινδύνευσης να έχει διαφορετικά σχήματα για άτομα από διαφορετικά στρώματα.

Κεφάλαιο 2

Συμπερασματολογία για παραμετρικά μοντέλα και προσεγγίσεις απόδοσης τιμών

2.1 Εισαγωγή

Ένα βασικό πλεονέκτημα των παραμετρικών προσεγγίσεων είναι ότι η εφαρμογή τους είναι καταρχήν απλή και γενικά ισχύει η συνήθης θεωρία μεγίστης πιθανοφάνειας. Οι προσεγγίσεις μέσω απόδοσης τιμών (imputation) χρησιμοποιούνται για να ανάγουν το πρόβλημα της ανάλυσης των χρόνων αποτυχίας των αποκομμένων δεδομένων σε διάστημα ως προς εκείνο της ανάλυσης χρόνων αποτυχίας με δεξιά αποκομμένα δεδομένα. Έτσι, μπορούμε να αποφύγουμε τη χρησιμοποίηση της αποκοπής σε διάστημα (interval censoring) και να χρησιμοποιήσουμε τις υπάρχουσες συμπερασματικές διαδικασίες και του στατιστικού λογισμικού που έχει αναπτυχθεί για τα δεξιά αποκομμένα δεδομένα.

2.2 Παραμετρικά μοντέλα χρόνων αποτυχίας

Σε αυτή την ενότητα θα περιγράψουμε συνήθη παραμετρικά μοντέλα που χρησιμοποιούμε για την T, μία τυχαία μη αρνητική μεταβλητή που αντιπροσωπεύει το χρόνο αποτυχίας μιας μονάδας. Αυτά περιέχουν το Εκθετικό, το Weibull, το Λογαριθμο-κανονικό (log-normal) και το Λογαριθμο-λογιστικό (log-logistic) μοντέλο. Μερικά ακόμη παραμετρικά μοντέλα μπορούμε να βρούμε στους Kalbfleisch και Prentice (2002) και Lawless (2003).

2.2.1 Το Εκθετικό μοντέλο

Το Εκθετικό μοντέλο μίας παραμέτρου υποθέτει ότι η συνάρτηση διακινδύνευσης της T είναι σταθερή με T>0. Αυτό είναι $\lambda(t) = \lambda > 0$.

Είναι το απλό μοντέλο χρόνων αποτυχίας και υποθέτει ότι ο στιγμιαίος ρυθμός αποτυχίας είναι ανεξάρτητος του χρόνου t. Υπό αυτό το μοντέλο, οι συναρτήσεις επιβίωσης και πυκνότητας πιθανότητας της T είναι αντίστοιχα :

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

Επιπλέον, μπορεί εύκολα να αποδειχθεί ότι η υπό συνθήκη πιθανότητα αποτυχίας μέσα σε χρονικό διάστημα συγκεκριμένου μήκους είναι το ίδιο ανεξαρτήτως του πόσο μεγάλη είναι η μονάδα της μελέτης. Αυτή η ιδιότητα συνήθως αναφέρεται ως "ιδιότητα της μη απομνημόνευσης" του Εκθετικού μοντέλου (memoryless property).

Υποθέτουμε ότι υπάρχει ένα διάνυσμα από συμμεταβλητές **Z** και μας ενδιαφέρει η επίδραση του **Z** στην T. Ένας τρόπος να ορίσουμε το Εκθετικό μοντέλο παλινδρόμησης είναι να υποθέσουμε ότι η υπό συνθήκη συνάρτηση διακινδύνευσης της T δοθέντος του Z έχει τον τύπο :

$$\lambda(t; \mathbf{Z}) = \lambda \exp(\mathbf{Z}'\boldsymbol{\beta})$$

η οποία ακολουθεί το μοντέλο διακινδύνευσης (PH) (1.4.2.1). Εδώ το **β** δηλώνει το διάνυσμα των παραμέτρων παλινδρόμησης.

Η υπό συνθήκη συνάρτηση πυκνότητας πιθανότητας της T τότε έχει τον εξής τύπο :

$$f(t; \mathbf{Z}) = \lambda \exp(\mathbf{Z}'\boldsymbol{\beta}) \exp[-\lambda t \exp(\mathbf{Z}'\boldsymbol{\beta})]$$

δοθέντος του **Z.**

Έστω $Y = \log(T)$ ο λογάριθμος του χρόνου επιβίωσης (log survival time). Τότε το παραπάνω μοντέλο μπορεί εξίσου να ορισθεί και από τη σχέση :

$$Y = a - Z'\boldsymbol{\beta} + W \qquad (2.2.1.1)$$

Όπου $a = -\log(\lambda)$ και με το W να ακολουθεί την κατανομή Gumbel με συνάρτηση πυκνότητας πιθανότητας που δίνεται από τον παρακάτω τύπο :

$$\exp(w - e^{w}),\ \text{με } -\infty < w < \infty.$$

Σε σχέση με την T, το (2.2.1.1) μοντέλο είναι ένα Λογαριθμο-γραμμικό μοντέλο, και για το Y είναι ένα γραμμικό μοντέλο με τυχαία μεταβλητή σφάλματος W η οποία είναι της κατανομής Gumbel.

2.2.2 Το μοντέλο Weibull

Το απλό Εκθετικό μοντέλο που περιγράψαμε παραπάνω εξαρτάται μόνο από μία παράμετρο και μπορεί να είναι πολύ περιοριστικό μερικές φορές. Μία σημαντική γενίκευση αυτού είναι το μοντέλο Weibull δύο παραμέτρων με συνάρτηση διακινδύνευσης :

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

με λ, γ > 0. Είναι εύκολο να δει κανείς ότι η συνάρτηση διακινδύνευσης είναι μονότονα φθίνουσα για γ < 1, αύξουσα για γ > 1 και ανάγεται στην εκθετική συνάρτηση διακινδύνευσης εάν γ = 1. Υπό το μοντέλο Weibull οι συναρτήσεις επιβίωσης και πυκνότητας πιθανότητας της T έχουν τους τύπους :

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

και

$$f(t) = \lambda\gamma(\lambda t)^{\gamma-1}\exp[-(\lambda t)^{\gamma}]$$

αντίστοιχα.

Για την ανάλυση παλινδρόμησης, όπως και στην περίπτωση του εκθετικού μοντέλου, η συνάρτηση διακινδύνευσης μπορεί να γενικευθεί σε

$$\lambda(t;\mathbf{Z}) = \lambda\gamma(\lambda t)^{\gamma-1}\exp(\mathbf{Z}'\boldsymbol{\beta})$$

Η ανταποκρινόμενη υπό συνθήκες συνάρτηση πυκνότητας πιθανότητας της T δοθέντος του $\mathbf{Z}$ είναι τότε

$$f(t;\mathbf{Z}) = \lambda\gamma(\lambda t)^{\gamma-1}\exp(\mathbf{Z}'\boldsymbol{\beta})\exp[-(\lambda t)^{\gamma}\exp(\mathbf{Z}'\boldsymbol{\beta})]$$

Όπως και στην περίπτωση του Εκθετικού μοντέλου, με $Y = \log(T)$, το μοντέλο παλινδρόμησης μπορεί να γραφτεί ως

$$Y = a - Z^{'}\boldsymbol{\beta}^{*} + \sigma W \qquad (2.2.2.1)$$

Όπου $a = \log(\lambda)$, $\sigma = \gamma^{-1}$, $\beta^{*} = -\sigma\boldsymbol{\beta}$ και με το W να ακολουθεί την κατανομή Gumbel.

Να παρατηρήσουμε ότι όπως στο μοντέλο αναλογικής διακινδύνευσης (PH), και στο Εκθετικό καθώς και στο Weibull μοντέλο παλινδρόμησης οι συμμεταβλητές έχουν πολλαπλασιαστικές επιδράσεις στη συνάρτηση διακινδύνευσης.

Από την άλλη πλευρά, όπως το μοντέλο επιταχυνόμενης διακοπής (με αντίστοιχο τύπο $\log(\mathrm{T}) = \mathbf{Z}^{'}\boldsymbol{\beta} + W$), και τα δύο μοντέλα είναι λογαριθμο-γραμμικά και υπό αυτά οι συμμεταβλητές επηρεάζουν προσθετικά το λογαριθμο-χρόνο επιβίωσης (log survival time) του Y. Το μοντέλο Weibull είναι η μόνη οικογένεια μοντέλων που ικανοποιεί αυτούς τους όρους (Kalbfleisch και Prentice, 2002).

2.2.3 Το Λογαριθμο-κανονικό μοντέλο

Το Λογαριθμο-κανονικό μοντέλο υποθέτει ότι ο λογαριθμο-χρόνος επιβίωσης Y=log(T) έχει τον τύπο $Y = a + \sigma W$ με το W να είναι μία τυχαία μεταβλητή της Κανονικής κατανομής. Η συνάρτηση πυκνότητας πιθανότητας της T είναι τότε :

$$f(t) = (2\pi)^{-1/2}\gamma t^{-1}\exp[\frac{-\gamma^{2}(\log \lambda t)^{2}}{2}]$$

Όπου $\lambda = \exp(-a)$ και $\sigma = \gamma^{-1}$ όπως και προηγουμένως. Οι συναρτήσεις επιβίωσης και διακινδύνευσης της T περιέχουν τη σταθερή κανονική συνάρτηση κατανομής Φ(w) με

$$S(t) = 1 - \Phi(\gamma \log \lambda t)$$

και με τις δύο να μην έχουν κλειστού τύπου μορφή. Η συνάρτηση διακινδύνευσης αυξάνεται από το μηδέν για t=0 σε ένα μέγιστο και μετά πέφτει στο μηδέν όσο το t μεγαλώνει. Στην περίπτωση όπου υπάρχουν συμμεταβλητές **Z**, είναι προφανές ότι ακολουθώντας τα μοντέλα (2.2.1.1) και (2.2.2.1) μπορεί κανείς να ορίσει το Λογαριθμο-κανονικό μοντέλο παλινδρόμησης ως εξής :

$$Y = a - \mathbf{Z}^{'}\boldsymbol{\beta} + \sigma W \qquad (2.2.3.1)$$

το σύνηθες γραμμικό μοντέλο παλινδρόμησης. Αυτό το μοντέλο είναι ιδιαίτερα εύκολο να εφαρμοσθεί εάν δεν υπάρχει καθόλου αποκοπή. Αλλά η ύπαρξη της αποκοπής δυσχεραίνει τον υπολογισμό και την εξαγωγή συμπερασμάτων.

2.2.4 Το Λογαριθμο-λογιστικό μοντέλο

Το Λογαριθμο-λογιστικό μοντέλο ορίζεται με τον ίδιο τρόπο που ορίσαμε και το Λογαριθμο-κανονικό μοντέλο εκτός από το ότι το W έχει συνάρτηση πυκνότητας πιθανότητας

$$f(w) = \frac{e^w}{(1+e^w)^2}$$

Η συνάρτηση πυκνότητας πιθανότητας είναι συμμετρική με μέσο το μηδέν και διασπορά $\pi^2/3$. Υπό το Λογαριθμο-λογιστικό μοντέλο η σ.π.π. της Τ είναι:

$$f(t) = \lambda\gamma(\lambda t)^{\gamma-1}[1+(\lambda t)^{\gamma}]^{-2}$$

όπου ορίζουμε ξανά $\lambda = \exp(-a)$ και $\sigma = \gamma^{-1}$.

Σε σύγκριση με το Λογαριθμο-κανονικό μοντέλο, το Λογαριθμο-λογιστικό αν και χρησιμοποιείται λιγότερο συχνά στην ανάλυση χρόνων αποτυχίας έχει το πλεονέκτημα ότι και η συνάρτηση επιβίωσης και η συνάρτηση διακινδύνευσης έχουν κλειστού τύπου μορφές. Επιπλέον είναι πιο βολικό από ότι το Λογαριθμο-κανονικό μοντέλο στο χειρισμό των αποκομμένων δεδομένων. Η συνάρτηση επιβίωσης και διακινδύνευσης αντίστοιχα είναι :

$$S(t) = \frac{1}{1+(\lambda t)^{\gamma}} \quad \text{και} \quad \lambda(t) = \frac{\lambda\gamma(\lambda t)^{\gamma-1}}{1+(\lambda t)^{\gamma}}$$

Εάν γ < 1, η λ(t) είναι μονότονα φθίνουσα από το ∞ και εάν γ = 1 είναι μονότονα φθίνουσα από το λ. Για γ > 1 όπως και η λογαριθμο-κανονική συνάρτηση διακινδύνευσης λ(t) αυξάνεται από το μηδέν ως το μέγιστο και μετά πέφτει στο μηδέν.

2.3 Πιθανοφάνεια βασισμένη σε συμπεράσματα για παραμετρικά μοντέλα

Ας θεωρήσουμε μία μελέτη επιβίωσης που αποτελείται από n ανεξάρτητες μονάδες. Έστω T_i να δηλώνει το χρόνο επιβίωσης του μονάδας i που μας ενδιαφέρει για i= 1, . .. , n και υποθέτουμε ότι το T_i ακολουθεί ένα παραμετρικό μοντέλο με συνάρτηση επιβίωσης S(t,θ), όπου $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)'$ δηλώνει άγνωστες παραμέτρους. Επίσης υποθέτουμε ότι μόνο αποκομμένα δεδομένα σε διάστημα είναι διαθέσιμα και τότε έχουμε τον τύπο $\{(L_i, R_i], \mathbf{Z}_i;\ i = 1, ..., n\}$, όπου $(L_i, R_i]$ δηλώνει το διάστημα στο οποίο το T_i έχει παρατηρηθεί ότι ανήκει και Z_i είναι το διάνυσμα συμμεταβλητών που σχετίζεται με την i-οστή μονάδα i=1,..., n. Τότε η συνάρτηση πιθανοφάνειας είναι ανάλογη με :

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\theta}) = \prod_{i=1}^{n} [S(L_i, \boldsymbol{\theta}) - S(R_i, \boldsymbol{\theta})]$$

θεωρώντας ότι $L_i < R_i$ για όλα τα i=1,..., n.

2.3.1 Συμπερασματολογία με γενικά παραμετρικά μοντέλα

Η κλασική προσέγγιση της συμπερασματολογίας για το **θ** είναι να εκτιμηθεί από την εκτιμήτρια μεγίστης πιθανοφάνειας, που ορίζεται ως η τιμή του **θ** που μεγιστοποιεί το L(**θ**), και η χρήση των στατιστικών συναρτήσεων του λόγου πιθανοφανειών και του ελέγχου score που προέρχονται από το L(**θ**). Δοθέντος του ανεξάρτητου μηχανισμού αποκοπής που θεωρούμε εδώ, γενικά ισχύει η κλασική ασυμπτωτική πιθανοθεωρία. Ειδικότερα ασυμπτωτικές προσεγγίσεις των κατανομών των εκτιμητριών μεγίστης πιθανοφάνειας και των ελεγχοσυναρτήσεων του λόγου πιθανοφανειών και score είναι διαθέσιμες και παρέχουν άμεσα συμπεράσματα. Έστω ότι το $\hat{\boldsymbol{\theta}}$ δηλώνει την εκτιμήτρια μεγίστης πιθανοφάνειας του **θ.** Ορίζουμε:

$$U(\boldsymbol{\theta}) = \sum_{i=1}^{n} U_i(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} L_i(\boldsymbol{\theta}) \text{ και } I(\boldsymbol{\theta}) = \sum_{i=1}^{n} I_i(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} L_i(\boldsymbol{\theta})$$

Τότε κάτω από συγκεκριμένες συνθήκες κανονικότητας η $\hat{\boldsymbol{\theta}}$ είναι συνεπής. Επιπλέον, όταν το n είναι μεγάλο είναι η μοναδική λύση $\hat{\boldsymbol{\theta}}$ της εξίσωσης U(**θ**)=0 και η κατανομή του μπορεί να

προσεγγισθεί από την πολυμεταβλητή Κανονική κατανομή με μέσο $\boldsymbol{\theta}$ και συνδιακύμανση τον πίνακα $I^{-1}(\boldsymbol{\theta})$. Με άλλα λόγια, έχουμε $\hat{\boldsymbol{\theta}} \sim \mathbf{N}(\boldsymbol{\theta}, I^{-1}(\boldsymbol{\theta}))$ που μπορεί να χρησιμοποιηθεί για τον έλεγχο υποθέσεων και εκτιμήσεων που απορρέουν από διάστημα για το $\boldsymbol{\theta}$. Για να καθορίσουμε το $\hat{\boldsymbol{\theta}}$ μπορούμε να χρησιμοποιήσουμε διαδικασία εξεύρεσης ρίζας (root-finding procedure) ή τον αλγόριθμο του Newton-Raphson. Υποθέτουμε ότι $I(\boldsymbol{\theta}_0)$ είναι μη μοναδικό. Στον αλγόριθμο Newton-Raphson μία αρχική τιμή, ας πούμε $\boldsymbol{\theta}_0$, του $\boldsymbol{\theta}$ έχει ενημερωθεί (updated) από:

$\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 - I^{-1}(\boldsymbol{\theta}_0)U(\boldsymbol{\theta}_0)$ επαναληπτικά έως ότου επιτευχθεί η σύγκλιση.

Για να ελέγξουμε την υπόθεση $H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, με $\boldsymbol{\theta}_0$ γνωστό, είναι βολικό να κάνουμε χρήση του ελέγχου score (score test)

$$U^{'}(\boldsymbol{\theta}_0)I^{-1}(\boldsymbol{\theta}_0)U(\boldsymbol{\theta}_0)$$

το οποίο έχει μία ασυμπτωτική χ^2 κατανομή με p βαθμούς ελευθερίας. Υποθέτουμε ότι $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^{'}, \boldsymbol{\theta}_2^{'})^{'}$ όπου $\boldsymbol{\theta}_1$ και $\boldsymbol{\theta}_2$ είναι συνιστώσες του $\boldsymbol{\theta}$ με διαστάσεις k και p-k αντίστοιχα.

Τότε στην πράξη, μία πιο συνήθη υπόθεση είναι $H_2 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$, με $\boldsymbol{\theta}_{10}$ γνωστό. Σε αυτή τη περίπτωση, ο μερισμός (partition) U($\boldsymbol{\theta}$) γίνεται όπως και για το $\boldsymbol{\theta}$, π.χ.

$$U^{'}(\boldsymbol{\theta}) = [U_1^{'}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), U_2^{'}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)]$$

με U_1 και U_2 να έχουν διαστάσεις k και p-k αντιπροσωπεύοντας τα $\boldsymbol{\theta}_1$ και $\boldsymbol{\theta}_2$ αντίστοιχα.

2.4 Συμπερασματολογία βασισμένη στην απόδοση ή αντικατάσταση τιμών (imputation)

Η απόδοση ή πολλαπλή απόδοση (imputation) είναι μία γενική προσέγγιση διαχείρισης μη καταγεγραμμένων δεδομένων σε προβλήματα (Rubin, 1987) και συχνά χρησιμοποιείται, παραδείγματος χάριν, σε μελέτες δειγματοληψίας. Τα μη καταγεγραμμένα δεδομένα (missing data) συνήθως αναφέρονται σε παρατηρηθέντα δεδομένα τα οποία δε μας παρέχουν καμία πληροφορία καταγεγραμμένη ως προς τις μεταβλητές απόκρισης που μας ενδιαφέρουν. Οι αποκομμένοι ή σε

διάστημα αποκομμένοι χρόνοι αποτυχίας διαφέρουν από τους χρόνους αποτυχίας που αντιστοιχούν σε μη καταγεγραμμένα δεδομένα γιατί οι πρώτοι παρέχουν έστω και ελλιπείς (incomplete) πληροφορίες για τους χρόνους αποτυχίας που μας ενδιαφέρουν. Με άλλα λόγια, τα αποκομμένα δεδομένα σε διάστημα είναι στην πραγματικότητα ατελή δεδομένα, όχι ακριβώς μη καταγεγραμμένα δεδομένα. Παρόλα αυτά, μπορεί κανείς να διαχειριστεί τους μη παρατηρούμενους πραγματικούς χρόνους διακοπής ως μη καταγεγραμμένες τιμές και να τους αντικαταστήσει με τιμές τεκμαρτές ή τιμές συμβατής απόδοσης (imputed values) βασιζόμενος στις παρατηρηθείσες τιμές.

Έστω ότι το T_i αντιπροσωπεύει τους χρόνους επιβίωσης που μας ενδιαφέρουν από n ανεξάρτητα άτομα, και ας υποθέσουμε ότι έχουν παρατηρηθεί μόνο αποκομμένα σε διάστημα δεδομένα και τα συμβολίζουμε ως εξής :

$$\{(L_i, R_i],\ \mathbf{Z}_i;\ i = 1, ...,\ n\}$$

Έστω πρόβλημα που αφορά σε ένα δείγμα. Θεωρούμε ότι $\mathbf{Z}_i = 0$ και ότι η $S(t;\boldsymbol{\theta})$ δηλώνει τη συνάρτηση επιβίωσης των T_i με άγνωστες παραμέτρους $\boldsymbol{\theta}$. Στις ακόλουθες περιπτώσεις εστιάζουμε κυρίως σε αυτές όπου η διάσταση του $\boldsymbol{\theta}$ είναι άπειρη. Αυτές συμπεριλαμβάνουν περιπτώσεις όπου το $\boldsymbol{\theta}$ αντιπροσωπεύει όλη τη συνάρτηση επιβίωσης ή αποτελείται από πεπερασμένης διάστασης διάνυσμα των παραμέτρων που μας ενδιαφέρουν. Το μη παραμετρικό πρόβλημα, που μόλις αναφέραμε, το οποίο αφορούσε σε ένα δείγμα αντιστοιχεί στην πρώτη περίπτωση και ένα παράδειγμα της τελευταίας περίπτωσης δίνεται με το $\boldsymbol{\theta}$ να γίνεται β και $\lambda_0(t)$ για την ανάλυση παλινδρόμησης υπό το μοντέλο (1.4.2.1).

Η συμβατή αντικατάσταση δεδομένων, σε τέτοιες περιπτώσεις, γίνεται με τη παραγωγή ενός ή πολλών ομάδων δεδομένων προερχόμενα από δεξιά αποκομμένους χρόνους αποτυχίας για τους T_i χρησιμοποιώντας τα παρατηρηθέντα δεδομένα. Στη συνέχεια, λοιπόν, μπορούμε να χρησιμοποιήσουμε τα νέα αυτά δεδομένα για να βγάλουμε συμπεράσματα για το $\boldsymbol{\theta}$. Είναι προφανές ότι αντί των δεξιά αποκομμένων δεδομένων, θα μπορούσαμε να δημιουργήσουμε ακριβή δεδομένα χρόνων αποτυχίας των T_i. Συνήθως κάτι τέτοιο δεν είναι αναγκαίο ούτε προτιμητέο. Ο κυριότερος λόγος είναι η ύπαρξη πολλών εγκεκριμένων μεθόδων δεξιά αποκομμένων δεδομένων για ποικίλα προβλήματα συμπερασματολογίας και υπάρχει η πιθανότητα εμφανών ελλείψεων στις προσεγγίσεις μέσω απόδοσης

τιμών (imputation approaches). Αν όμως επιλέξουμε να εισάγουμε το χρόνο επιβίωσης T_i από ένα πεπερασμένο διάστημα αποκομμένων δεδομένων, αλλά όχι δεξιά αποκομμένων παρατηρήσεων το πρόβλημα περιορίζεται στην ανάλυση των συμβατά αντικατεστημένων (δεξιά αποκομμένων) δεδομένων, τα οποία μπορούμε να χειριστούμε με μία από τις πολλές τεχνικές όπως για παράδειγμα τη μέθοδο μερικής πιθανοφάνειας. Στη συνέχεια θα συζητήσουμε δύο γενικές προσεγγίσεις της συμβατής απόδοσης τιμών. Στην παράγραφο 2.4.1 θα παρουσιαστεί η πρώτη προσέγγιση μέσω ενός σημείου (single point imputation approach) η οποία συχνά χρησιμοποιείται στην πράξη λόγω της απλότητάς της. Η δεύτερη προσέγγιση είναι με τη μέθοδο πολλαπλής προσέγγισης τιμών (Wei and Tanner, 1991) η εφαρμογή της οποίας αφορά στις τεχνικές αύξησης δεδομένων και αναλύεται στους Tanner και Wong (1987) και Tanner (1991) που θα συζητηθεί στην παράγραφο 2.4.2. Στη συνέχεια στις παραγράφους 2.4.3 και 2.4.4 θα αναφερθούμε στην εργασία των Wei και Tanner (1991) οι οποίοι με βάση το " Data augmentation algorithm" ανέπτυξαν δύο αλγορίθμους για την αντικατάσταση μη-καταγεγραμμένων δεδομένων:

α) Poor Man´s data augmentation (PMDA)

β) Asymptotic Normal data augmentation (ANDA)

2.4.1 Προσέγγιση συμβατής απόδοσης τιμών μέσω ενός σημείου

Για αποκομμένα δεδομένα χρόνων αποτυχίας σε διάστημα, η πιο απλή προσέγγιση απόδοσης τιμών ίσως είναι η υπόθεση ότι για τη γενική μονάδα i, ο υποβόσκων πραγματικός χρόνος αποτυχίας T_i είναι ίσος με μία τιμή μες στο παρατηρηθέν διάστημα $(L_i, R_i]$, $i = 1,..., n.$ Μία συνήθη επιλογή είναι να θεωρήσουμε ότι το T_i είναι το μεσαίο σημείο του διαστήματος για ένα πεπερασμένο διάστημα ή για μία αληθώς αποκομμένη παρατήρηση σε διάστημα. Για διαστήματα με $R_i = \infty$ ή με δεξιά αποκομμένες παρατηρήσεις, οι αρχικές παρατηρήσεις δεν τροποποιούνται. Στη συνέχεια έχουμε μια ομάδα από χρόνους αποτυχίας δεξιά αποκομμένων δεδομένων. Μία εναλλακτική επιλογή έναντι του μεσαίου σημείου είναι να θεωρήσουμε τον χρόνο αποτυχίας T_i να ταυτίζεται με το άνω άκρο του διαστήματος L_i, δηλαδή το τελευταίο αριστερό σημείο, ή με το R_i το τελευταίο δεξιό σημείο της προσέγγισης συμβατής απόδοσης τιμών. Είναι προφανές ότι οι τρεις παραπάνω μέθοδοι δε θα δώσουν πολύ διαφορετικά αποτελέσματα εάν όλα τα πεπερασμένα διαστήματα είναι περιορισμένα. Γενικότερα, η

επιλογή μίας εκ των τριών μεθόδων βασίζεται στο εάν το πραγματικό γεγονός που υπόκειται στη μελέτη μας είναι πιο πιθανό να συμβεί κοντά στο μεσαίο, στο αριστερό ή στο δεξιό σημείο του διαστήματος που παρατηρούμε. Φυσικά υπάρχουν περιπτώσεις όπου προγενέστερη πληροφορία δεν υπάρχει και σε αυτή την περίπτωση προτιμότερο είναι να επιλεχθεί τυχαία μία τιμή, παραδείγματος χάριν, από την Ομοιόμορφη κατανομή μέσα πάντα στο παρατηρηθέν διάστημα. Στις προαναφερθείσες μεθόδους αποδίδουμε τους χρόνους αποτυχίας T_i που είναι αποκομμένοι και βρίσκονται σε διάστημα. Αυτό όμως ισχύει μερικώς γιατί υπάρχουν αρκετές εγκεκριμένες προσεγγίσεις εξαγωγής συμπερασμάτων για δεξιά αποκομμένα δεδομένα που μπορούν να εφαρμοστούν στα συμβατά αντικατεστημένα δεξιά αποκομμένα δεδομένα. Ας υποθέσουμε ότι κάποιος θέλει να υπολογίσει μη παραμετρικά μία συνάρτηση επιβίωσης. Σε αυτή την περίπτωση μπορεί απλά να χρησιμοποιήσει την εκτιμήτρια Kaplan-Meier που δίνεται από τον τύπο

$$\hat{S} = \prod_{j:t_j \le t} (1 - \frac{d_j}{n_j})$$

(Kalbfleisch and Prentice, 2002).

Το t_j εδώ δηλώνει την ακριβή απόδοση της τιμής των χρόνων αποτυχίας, το d_j τον αριθμό των αποτυχιών και το n_j τον αριθμό των μονάδων που βρίσκονται σε κίνδυνο σε καθένα από τα t_j βασισμένα στη συμβατή αντικατάσταση των δεξιά αποκομμένων δεδομένων (imputed right-censored data). Είναι προφανές ότι η συνάρτηση $\hat{S}(t)$ είναι ασυνεχής στις ακριβείς αποδόσεις των τιμών των χρόνων αποτυχίας t_j.

Ας υποθέσουμε τώρα, ότι αντί του προβλήματος του ενός δείγματος κάποιος ενδιαφέρεται για την ανάλυση παλινδρόμησης υπό το μοντέλο αναλογικής διακινδύνευσης (PH) (1.4.2.1). Θεωρούμε πάλι τα t_j ορισμένα όπως προηγουμένως και $Z_{(j)}$ να δηλώνει το διάνυσμα της συμμεταβλητής του ατόμου του οποίου οι συμβατοί χρόνοι αποτυχίας είναι ίσοι με τους t_j δοθέντος ότι δεν υπάρχουν ακριβείς χρόνοι αποτυχίας από συμβατή αντικατάσταση. Επιπλέον, θεωρούμε $R_{(t_j)}$ να είναι η ομάδα κινδύνου των ατόμων το χρόνο t_j βασισμένη στη συμβατή αντικατάσταση των δεξιά αποκομμένων δεδομένων. Τότε

η παράμετρος παλινδρόμησης **β** στο μοντέλο (1.4.2.1) μπορεί να εκτιμηθεί με τη χρήση της εκτιμήτριας μερικής πιθανοφάνειας που ορίζεται ως την τιμή του **β** που μεγιστοποιεί τη μερική πιθανοφάνεια :

$$L_p(\boldsymbol{\beta}) = \prod_j \frac{\exp(\mathbf{Z}_j'\boldsymbol{\beta})}{\sum_{l \in R_{(t_j)}} \exp(\mathbf{Z}_{(l)}'\boldsymbol{\beta})}$$

(Cox, 1972). Για δεξιά αποκομμένα δεδομένα χρόνων αποτυχίας η εκτιμήτρια μερικής πιθανοφάνειας έχει εκτενώς μελετηθεί και γνωρίζουμε ότι είναι συνεχής και έχει ασυμπτωτική πολυμεταβλητή Κανονική κατανομή (Andersen, et al., 1982).

Εάν υπάρχουν ακριβείς χρόνοι αποτυχίας από συμβατή αντικατάσταση (imputed exact failure times), η μερική πιθανοφάνεια $L_p(\boldsymbol{\beta})$ χρειάζεται αναπροσαρμογή. Για αυτό το λόγο υπάρχουν αρκετοί τρόποι, ένας εκ των οποίων είναι να χρησιμοποιήσουμε την προσέγγιση :

$$\prod_j \frac{\exp(\mathbf{Z}_j'\boldsymbol{\beta})}{[\sum_{l \in R_{(t_j)}} \exp(\mathbf{Z}_l\boldsymbol{\beta})]^{d_j}}$$

(Breslow, 1974), όπου d_j είναι ο αριθμός των ατόμων των οποίων οι τεκμαρτοί από αντικατάσταση χρόνοι αποτυχίας είναι ίσοι με το t_j όπως αυτό ορίστηκε προηγουμένως.

Το μεγαλύτερο πλεονέκτημα της προσέγγισης απόδοσης τιμών μέσω ενός σημείου είναι η απλότητά της και στη συγκεκριμένη περίπτωση συμπεράσματα μπορούν να εξαχθούν χωρίς καμία δυσκολία με το ήδη υπάρχον λογισμικό. Εάν όλα τα διαστήματα είναι σχετικά μικρά η προσέγγιση μπορεί να μας παράσχει μία λογική και απλή εκτίμηση του αποτελέσματος βασιζόμενη στα παρατηρηθέντα δεδομένα.

Η μελέτη των δεδομένων επιβίωσης τα οποία δε βρίσκονται σε διάστημα είναι εξίσου σημαντική ως προς την εκτίμηση της συνάρτησης επιβίωσης $S(t)$ και στην αξιολόγηση της σημαντικότητας της δυναμικής των παραγόντων πρόγνωσης.

Ελάχιστα, όμως, στατιστικά λογισμικά επιτρέπουν τη χρήση τέτοιων δεδομένων και γι΄ αυτό το λόγο μία κοινή πρακτική εφαρμογή ανάλυσης δεδομένων είναι η υπόθεση ότι το γεγονός έχει συμβεί

είτε στο άνω είτε στο κάτω όριο του διαστήματος $(L_i, U_i]$ είτε ακόμα στο μέσον κάθε διαστήματος. Παρά την πρακτική αυτή εφαρμογή υπάρχουν συγγραφείς μεταξύ των οποίων οι Rucker και Messerer (1988), Odell (1992) και Dorey (1993) οι οποίοι υποστηρίζουν ότι το γεγονός της αποδοχής της υπόθεσης ότι οι χρόνοι επιβίωσης των διαστημάτων αντιστοιχούν στους ακριβείς χρόνους μπορεί να οδηγήσει σε μεροληπτικές εκτιμήσεις όπως επίσης ότι και τα αποτελέσματα και τα συμπεράσματα δεν μπορούν να είναι τελείως αξιόπιστα. Γιατί το μεγαλύτερο μειονέκτημα, της προσέγγισης συμβατής απόδοσης τιμών μέσω ενός σημείου (single imputation) είναι ότι συμπεριφέρεται στις μη καταγεγραμμένες ως τεκμαρτές, και ως εκ τούτου υποεκτιμάται η αληθινή μεταβλητότητα. Για τον προαναφερθέν λόγο, λοιπόν, μία λογική προέκταση είναι η επαναλαμβανόμενη προσέγγιση, που θα συζητήσουμε παρακάτω, έναντι της μίας και μοναδικής εφαρμογής της.

2.4.2 Πολλαπλή προσέγγιση απόδοσης ή αντικατάστασης τιμών

Σε αυτή την παράγραφο θα συζητήσουμε την εφαρμογή του αλγορίθμου για την αύξηση δεδομένων (data augmentation algorithm) όπως αναφέρεται από τους Tanner και Wong (1987) και Tanner (1991) για την ανάλυση αποκομμένων δεδομένων σε διάστημα. Αρχικά ο αλγόριθμος είχε σχεδιαστεί για να υπολογίζει την ύστερη κατανομή των παραμέτρων επαναλαμβάνοντας τα ακόλουθα δύο βήματα:

1) την απόδοση ή αντικατάσταση τιμών

2) τα ύστερα (posterior) βήματα

Πιο συγκεκριμένα, ακολουθώντας τον αλγόριθμο αύξησης δεδομένων μπορούμε να εκτιμήσουμε το **θ** ως ακολούθως:

βήμα 1. Δίνεται μία αρχική τιμή $\hat{\theta}^{(0)}$ και ορίζεται $\hat{S}^{(0)}(t) = S(t; \hat{\theta}^{(0)})$

βήμα 2. Στην i-οστή επανάληψη, για κάθε k και i, εάν $R_i = \infty$ π.χ. μία δεξιά αποκομμένη παρατήρηση έχει καταγραφεί για το T_i, ορίζουμε $T_i^{(k,l)} = L_i$ και $\delta_i^{(k,l)} = 0, \quad k = 1,..., M, \; i = 1,..., n.$ Διαφορετικά, ορίζουμε $T_i^{(k,l)}$ να είναι ένας τυχαίος αριθμός που παράγεται από τη $\hat{S}^{(l-1)}$ δεδομένου ότι $T_i^{(k,l)} \in (L_i, R_i]$ και $\delta_i^{(k,l)} = 1$. Αυτό δίνει M ομάδες από δεξιά αποκομμένα δεδομένα

$$\{T_i^{(k,l)},\ \delta_i^{(k,l)},\ Z_i;\ i=1,...,\ n\} \qquad (2.4.2.1)$$

k=1,..., N. Να σημειώσουμε ότι εδώ οι δεξιά αποκομμένοι δείκτες είναι πάντα οι ίδιοι.

βήμα 3. Για κάθε μία από τις M ομάδες των δεξιά αποκομμένων δεδομένων που δημιουργήθηκαν στο βήμα **1.** έχουμε και την εκτίμηση $\hat{\theta}^{(k,l)}$. Έστω $\hat{\Sigma}^{(k,l)}$ η εκτίμηση του πίνακα συνδιακύμανσης που περιέχει μία συνιστώσα του **θ** πεπερασμένης διάστασης.

βήμα 4. Καθορίζουμε την ανανεωμένη εκτιμήτρια του **θ** χρησιμοποιώντας τον τύπο

$$\hat{\theta}^{(l)} = \frac{1}{M}\sum_{k=1}^{M}\hat{\theta}^{(k,l)}$$

Ο αντίστοιχος πίνακας συνδιακύμανσης ή η συνάρτηση διακύμανσης μπορεί να εκτιμηθεί από

$$\hat{\Sigma}^{(l)} = \frac{1}{M}\sum_{k=1}^{M}\hat{\Sigma}^{(k,l)} + (1+\frac{1}{\mathrm{M}})\frac{\sum_{k=1}^{M}[\hat{\theta}_t^{(k,l)} - \hat{\theta}_t^{(l)}][\hat{\theta}_t^{(k,l)} - \hat{\theta}_t^{(l)}]}{M-1}$$

Το $\hat{\theta}_t^{(k,l)}$ δηλώνει τη συνιστώσα του **θ** που μας ενδιαφέρει ή την τιμή της συνάρτησης που εκπροσωπείται από το **θ** στο χρόνο t για ένα μη παραμετρικό πρόβλημα του ενός δείγματος.

βήμα 5. Επαναλαμβάνουμε τα βήματα **2.** έως **4.** μέχρι να επιτευχθεί η επιθυμητή σύγκλιση.

2.4.3 Poor Man´s Data Augmentation for Interval Data (PMDA)

Εδώ προτείνεται η χρήση της PMDA μεθόδου (Wei and Tanner, 1991) που αποδίδει χρόνους επιβίωσης από ένα πεπερασμένο διάστημα αποκομμένων δεδομένων. Αυτή η μέθοδος περιλαμβάνει τη δημιουργία πολλών ομάδων τεκμαρτών δεδομένων από έναν επαναληπτικό αλγόριθμο. Θα χρησιμοποιηθεί ο εκθέτης (i) και ο δείκτης (k) για να αντιπροσωπεύσουν την i-οστή επανάληψη και την k-οστή ομάδα συμβατά αντικατεστημένων δεδομένων αντίστοιχα. Οι παράμετροι θα πρέπει να εκτιμηθούν συμπεριλαμβανομένου του συντελεστή παλινδρόμησης **β** και της βασικής συνάρτησης επιβίωσης .

Ο αλγόριθμος που βασίζεται στην PMDA μέθοδο περιλαμβάνει τα ακόλουθα βήματα:

1. Υποθέτει ότι οι τρέχουσες εκτιμήσεις του συντελεστή παλινδρόμησης και της βασικής συνάρτησης επιβίωσης είναι:

$$\hat{\boldsymbol{\beta}}^{(i)} \text{ και } \hat{S}_0^{(i)}$$

2. Παράγει m ομάδες από πιθανές δεξιά αποκομμένες $\{T_1, \delta_1, Z\}, ..., \{T_m, \delta_m, Z\}$ παρατηρήσεις ως ακολούθως :

για κάθε παρατήρηση, (U_j, V_j, Z_j), $j = 1, ..., n$ και $\mathrm{k} = 1, ..., \mathrm{m}$.

α) Εάν $V_j < \infty$, δοκιμάζει X_j από την κατανομή $[\hat{\mathrm{S}}_0^{(i)}]^{\exp(Z_i\hat{\boldsymbol{\beta}}^{(i)})}$ δοθέντος ότι $\{U_j < X_j \le V_j\}$ και τότε $T_{(k),j} = X_j$ και $\delta_{(k),j} = 1$.

β) Εάν $V_j = \infty$ τότε $T_{(k),j} = U_j$ και $\delta_{(k),j} = 0$.

3. Χρησιμοποιεί κάθε $\{T_1, \delta_1, Z\}$ για να εφαρμόσει ένα μοντέλο του Cox που θα περιέχει μία εκτίμηση του $\hat{\boldsymbol{\beta}}_{(k)}^{(i)}$ και της συνδιακύμανσης του εκτιμητή $\hat{\Sigma}_{(k)}^{(i)}$.

4. Βασιζόμενος στην ομάδα $\{T_{(k)}, \delta_{(k)}, Z\}$ και στην εκτίμηση $\hat{\boldsymbol{\beta}}_{(k)}^{(i)}$, υπολογίζει την εκτιμήτρια Breslow της βασικής συνάρτησης επιβίωσης $\hat{\mathrm{S}}_{0,(k)}^{(i)}$ για $k = 1, ..., m$.

5. Έστω

$$\hat{\boldsymbol{\beta}}^{(i+1)} = \frac{1}{n}\left(\sum_{k=1}^{m} \hat{\boldsymbol{\beta}}_{(k)}^{(i)}\right) \text{ και } \hat{S}^{(i+1)} = \frac{1}{m}\left(\sum_{k=1}^{m} \hat{S}_{0,(k)}^{(i)}\right)$$

$$\hat{\Sigma}^{(i+1)} = \frac{1}{m}\left(\sum_{k=1}^{m} \hat{\Sigma}_{(k)}^{(i)}\right) + \left(1 + \frac{1}{m}\right)\frac{\sum_{k=1}^{m}[\hat{\boldsymbol{\beta}}_{(k)}^{(i)} - \hat{\boldsymbol{\beta}}^{(i+1)}]^2}{m-1}$$

6. Έστω $i \leftarrow i+1$ και επέστρεψε στο βήμα **1.** της μεθόδου PMDA έως ότου το $\hat{\boldsymbol{\beta}}^{(i)}$ να συγκλίνει. Τα $\hat{\boldsymbol{\beta}}^{(i)}$, $\hat{\Sigma}^{(i)}$ και $\hat{S}_0^{(i)}$ στη σύγκλιση είναι οι τελικές μας εκτιμήσεις.

Στο βήμα **1.** χρησιμοποιούμε την αρχική τιμή $\hat{\boldsymbol{\beta}}^{(0)} = 0$. Για την αρχική εκτίμηση της συνάρτησης επιβίωσης $\hat{S}_0^{(0)}$**,** πρέπει καταρχήν να δημιουργήσουμε m ομάδες από συμβατά αντικατεστημένα δεδομένα με έναν απλό τρόπο :

κρατάμε τις δεξιά αποκομμένες παρατηρήσεις, για κάθε αποκομμένη παρατήρηση σε διάστημα (U_j, V_j) ο ακριβής χρόνος αποτυχίας X_j έχει τυχαία επιλεγεί από την ομοιόμορφη κατανομή $U(U_j, V_j)$ και τότε $T_{(k),j} = X_j$ και $\delta_{(k),j} = 1$ για $k = 1, ..., m$.

Υποθέτουμε ότι η $\hat{S}_{0,(k)}^{(0)}$ είναι η εκτιμήτρια Breslow της βασικής συνάρτησης επιβίωσης από την k-οστή ομάδα των συμβατά αντικατεστημένων δεδομένων, τότε ισχύει η εξίσωση :

$$\hat{S}_0^{(0)} = \sum_{k=1}^{m} \frac{\hat{S}_{0,(k)}^{(0)}}{m}.$$

Στο βήμα **2.** δημιουργούμε m ομάδες από συμβατά δεδομένα (imputed data). Η γενική αίσθηση είναι ότι συχνά δε χρειάζεται οι ομάδες να είναι μεγάλες. Χρησιμοποιούμε ένα μέτριο αριθμό ομάδων, m=10.

Η συμβατή αντικατάσταση δεδομένων στο βήμα **2.** είναι απλή αφού η $\hat{S}_0$ είναι διακριτή. Υποθέτουμε τώρα ότι στο διάστημα $(U_j, V_j]$ η $[\hat{S}_0^{(i)}]^{\exp(Z_i\hat{\boldsymbol{\beta}}^{(i)})}$ έχει συνάρτηση πυκνότητας πιθανότητας $\{p_1, ..., p_{kj}\}$ στα σημεία $\{t_1, ..., t_{kj}\}$, τότε η X_j τυχαία επιλέγεται από τα $\{t_1, ..., t_{kj}\}$ με πιθανότητα ανάλογη της $\{p_1, ..., p_{kj}\}$.

Το βήμα **3.** μπορεί εύκολα να επιτευχθεί χρησιμοποιώντας μία από τις πολλές καθιερωμένες μεθόδους ή υπολογιστικά προγράμματα για δεξιά αποκομμένα δεδομένα.

Στο βήμα **5.** ως ένα άμεσο αποτέλεσμα έχουμε μία εκτίμηση της συνδιακύμανσης του εκτιμημένου συντελεστή παλινδρόμησης από το βήμα **1.** είναι το άθροισμα δύο όρων :

α) της εντός αντικατάστασης δεδομένων (within-imputation) και

β) της μεταξύ αντικατάστασης δεδομένων (between-imputation) των διακυμάνσεων.

Ο δεύτερος όρος μετρά την επιπρόσθετη μεταβλητότητα εξαιτίας των αποκομμένων δεδομένων σε πεπερασμένο διάστημα. Ένας παράγοντας μεγέθυνσης 1/m χρησιμοποιείται όταν εκτιμώντας την διακύμανση μεταξύ των συμβατά αντικατεστημένων δεδομένων (between-imputation variance) λαμβάνει υπόψη ένα πεπερασμένο αριθμό από συμβατές αντικαταστάσεις (Rubin, 1987; Schenker and Welsh, 1988; Tanner and Wong, 1987).

Συνήθως ως κριτήριο σύγκλισης χρησιμοποιείται το:

$$\left|\hat{\boldsymbol{\beta}}^{(i+1)} - \hat{\boldsymbol{\beta}}^{(i)}\right| < 0.01 \text{ ή } i > 50.$$

2.4.4 Asymptotic Normal Data Augmentation for Interval Censored Data (ANDA)

Είναι πλέον γνωστό ότι η PMDA μέθοδος μπορεί να υποεκτιμήσει την αληθινή μεταβλητότητα όταν η μη καταγραφή δεδομένων είναι σοβαρή. Σε αυτή την περίπτωση η μέθοδος ANDA είναι περισσότερο ακριβής (Wei and Tanner, 1991). Το σχέδιο της ANDA μπορεί να εφαρμοσθεί τροποποιώντας δύο βήματα του αλγορίθμου της PMDA μεθόδου ως ακολούθως :

1) Στο βήμα **5.** της PMDA μεθόδου η ύστερη κατανομή του συντελεστή παλινδρόμησης προσεγγίζεται από μια μίξη Κανονικών κατανομών:

$$\mathrm{g}^{(i+1)}(\boldsymbol{\beta}) = \frac{1}{m}\sum_{k=1}^{m} N(\hat{\boldsymbol{\beta}}_{(k)}^{(i)}, \hat{\Sigma}_{(k)}^{(i)})$$

2) Στο βήμα **2.** της PMDA μεθόδου, πρώτα παίρνουμε δείγμα m φορές από το $\mathrm{g}^{(i)}(\boldsymbol{\beta})$, k=1,..., m. Τότε για κάθε k=1, …, m και κάθε αποκομμένη παρατήρηση σε πεπερασμένο διάστημα (U_j, V_j, Z_j) το δείγμα X_j από την κατανομή $[\hat{S}_0^{(i)}]^{\exp(Z_i\hat{\boldsymbol{\beta}}_{(k)}^{(i)})}$, δοθέντος ότι $\{U_j < X_j \le V_j\}$ και διατηρώντας τις δεξιά αποκομμένες παρατηρήσεις. Χρησιμοποιούμε επίσης τις ίδιες αρχικές τιμές $\hat{\boldsymbol{\beta}}_{(k)}^{(0)}=0$, $\hat{\Sigma}_{(k)}^{(0)} = 0$ και $\hat{S}_{(0)}^{(0)}$.

Όλα τα υπόλοιπα βήματα παραμένουν ίδια.

2.4.5 Προσομοιώσεις και σχόλια

Με βάση προσομοιώσεων διερευνάται η επίδοση του τελικού δείγματος για τους δύο αυτούς αλγορίθμους. Η διαδικασία σύγκρισης που ακολουθείται σε αυτές τις περιπτώσεις υπολογίζει τη μη-παραμετρική εκτιμήτρια μεγίστης πιθανοφάνειας (NPMLE) η οποία βασίζεται σ'έναν αλγόριθμο επέκτασης της κυρτότητας (extended convex minorant algorithm) (Pan, 1999). Σαν βασική συνάρτηση κατανομής χρησιμοποιείται η Weibull (με παράμετρο σχήματος 2 και παράμετρο κλίμακας 1) καθώς και μία συμμεταβλητή: είτε δυαδική {0,1} (δηλαδή ίση με 0 ή 1 με μια αντίστοιχη πιθανότητα), είτε μια ομοιόμορφα συνεχής U(0,2). Να σημειώσουμε ότι οι επιδόσεις των δύο αλγορίθμων γενικώς είναι ισοδύναμες με ένα μικρό προβάδισμα στον αλγόριθμο ANDA ο οποίος και προτείνεται. Και οι δύο αλγόριθμοι φαίνεται να βελτιώνονται από την αύξηση του μεγέθους του δείγματος (Pan, 1999).

Συμπληρώνοντας υπάρχουν κάποιες επιφυλάξεις ως προς τη χρήση της εκτιμήτριας Breslow στο βήμα 4 της μεθόδου PMDA που πιθανώς να οδηγήσει στη συρρίκνωση της εκτιμώμενης βασικής συνάρτησης επιβίωσης καθώς οι επαναλήψεις θα συνεχίζονται. Για τη διόρθωση αυτής της τάσης προτάθηκε μία εξομάλυνση της εκτιμώμενης βασικής συνάρτησης επιβίωσης. Αντί της εκτιμήτριας Breslow μπορούμε απλά να χρησιμοποιήσουμε την εκτιμήτρια Link η οποία είναι μία γραμμική εξομάλυνση της εκτιμήτριας Breslow (Miller, 1981) και διορθώνει πιθανή εμφάνιση μεροληψίας.

2.4.6 Παράδειγμα

Θεωρούμε μία μελέτη δεδομένων που αφορά στον καρκίνο του μαστού (Finkelstein 1986; Finkelstein and Wolfe, 1985) και η οποία παρουσιάζεται στον πίνακα 3. Σε αυτή τη μελέτη υπήρξαν 94 ασθενείς νωρίς διαγνωσμένοι με καρκίνο του μαστού, στους οποίους δόθηκαν δύο ιατρικές θεραπείες αμέσως μετά την επέμβαση αφαίρεσης του όγκου. Η πρώτη αποτελείτο από τη ραδιοθεραπεία και η δεύτερη από βασική ραδιοθεραπεία σε συνδυασμό με ανοσοενισχυτική χημειοθεραπεία. Ένας εκ των στόχων της μελέτης ήταν να διερευνήσουν ποια από τις θεραπείες είχε καλύτερες μακροπρόθεσμες επιδράσεις. Τα δεδομένα είναι αποκομμένα δεδομένα σε διάστημα δεδομένου ότι οι ασθενείς μπορούσαν να επισκεφθούν την κλινική κάθε 4 έως 6 μήνες ή και σε διαστήματα μεγαλύτερα των μηνών αυτών εξαιτίας της απόστασης.

Πίνακας 3: Χρόνος επιδείνωσης των ασθενών με καρκίνο του μαστού που ακολουθούν δύο αγωγές.

Radiotherapy	(0,7]; (0,8]; (0,5]; (4,11]; (5,12]; (5,11]; (6,10]; (7,16]; (7,14]; (11,15]; (11,18]; ≥15; ≥17; (17,25]; (17,25]; ≥18; (19,35]; (18,26]; ≥22; ≥24; ≥24; (25,37]; (26,40]; (27;34]; ≥32; ≥33; ≥34; (36,44]; (36,48]; ≥36; ≥36; (37,44]; ≥37; ≥37, ≥37; ≥38, ≥40; ≥45; ≥46; ≥46; ≥46; ≥46; ≥46; ≥46; ≥46; ≥46
Radiotherapy + Chemotherapy	(0,22]; (0,5]; (4,9]; (4,8]; (5,8]; (8,12]; (8,21]; (10,35]; (10,17]; (11,13]; ≥11; (11,17]; ≥11; (11,20]; (12,20]; ≥13; (13,39]; ≥13; ≥13; (14,17]; (14,19]; (15,22]; (16,24]; (16,20]; (16,24]; (16,60]; (17,27]; (17,23]; (17,26]; (18,25]; (18,24]; (19,32]; ≥21; (22,32]; ≥23; (24,31]; (24,30]; (30,34]; (30,36]; ≥31; ≥32; (32,40]; ≥34; ≥34; ≥35; (35,39]; (44,48]; ≥48

Θα αναφερόμαστε στην ομάδα που ακολουθεί την θεραπεία της ραδιοθεραπείας ως την ομάδα αναφοράς με συμμεταβλητή $Z_i = 0$ και την άλλη ομάδα με $Z_i = 1$. Η NPMLE από το μοντέλο του Cox είναι $\tilde{\beta} = 0.795$ με εκτιμώμενο τυπικό σφάλμα 0.29 (Huang and Wellner, 1995). Η προτεινόμενη προσέγγιση μέσω πολλαπλής απόδοσης τιμών (multiple imputation approaches) χρησιμοποιώντας τους αλγορίθμους PMDA και ANDA αντίστοιχα αποδίδει τον εκτιμώμενο συντελεστή παλινδρόμησης (τυπικό σφάλμα) 0.90 (0.29) και 0.92 (0.29), τα οποία δεν είναι πολύ διαφορετικά από τα αποτελέσματα της NPMLE.

Τα δεδομένα αυτά αναλύονται περισσότερο όπως θα δούμε στην παράγραφο 3.4.4. Επιπλέον εφαρμόζοντας και τις δύο μεθόδους PMDA και ANDA με την εκτιμήτρια Link στα δεδομένα της μελέτης για τον καρκίνο του μαστού ο εκτιμώμενος συντελεστής παλινδρόμησης φαίνεται να είναι στην ουσία ίσος με τις αρχικές εκτιμήσεις.

Κεφάλαιο 3

Μη παραμετρική εκτίμηση μεγίστης πιθανοφάνειας

3.1 Εισαγωγή

Η εκτίμηση της συνάρτησης επιβίωσης είναι ίσως ο κλασικός στόχος που απαιτείται στην ανάλυση των δεδομένων χρόνων αποτυχίας. Μια εκτιμημένη συνάρτηση επιβίωσης μπορεί να χρησιμοποιηθεί για να αξιολογήσει την εγκυρότητα μιας υπόθεσης που αφορά σε ένα συγκεκριμένο παραμετρικό μοντέλο του οποίου η βασική συνάρτηση επιβίωσης μας ενδιαφέρει. Επίσης, μπορεί κάποιος να χρειαστεί να εκτιμήσει συναρτήσεις επιβίωσης για να εκτιμήσει στη συνέχεια συγκεκριμένες πιθανότητες επιβίωσης, για να συγκρίνει γραφικά αρκετές διαφορετικές θεραπείες ή για να προβλέψει πιθανότητες επιβίωσης για μελλοντικούς ασθενείς. Στην περίπτωση όπου ένα παραμετρικό μοντέλο μπορεί λογικώς να θεωρηθεί για τη βασική συνάρτηση επιβίωσης το πρόβλημα εκτίμησης είναι σχετικά εύκολο και για αυτό το πρόβλημα χρησιμοποιείται συνήθως η εκτιμήτρια μεγίστης πιθανοφάνειας που συζητήθηκε στην παράγραφο 2.3. Σε αυτό το κεφάλαιο θα εστιάσουμε την προσοχή μας στη μη παραμετρική εκτίμηση των συναρτήσεων επιβίωσης μαζί με την εκτίμηση των συναρτήσεων διακινδύνευσης.

Στην περίπτωση που έχουμε δεδομένα δεξιά αποκομμένων χρόνων αποτυχίας η μη παραμετρική εκτιμήτρια μεγίστης πιθανοφάνειας (NPMLE) της συνάρτησης επιβίωσης δίνεται από την εκτιμήτρια Kaplan-Meier (Kaplan and Meier, 1958; Kalbfleisch and Prentice, 2002) και η οποία έχει μελετηθεί εκτενώς. Επιπλέον η εκτιμήτρια του σημειακής διακύμανσης είναι διαθέσιμη και δίνεται από το γνωστό τύπο του Greenwood (Greenwood, 1926). Για δεδομένα αποκομμένα σε διάστημα, σε αντίθεση με την παραμετρική συμπερασματολογία, η μη παραμετρική συμπερασματολογία είναι πολύ περισσότερο σύνθετη από ότι τα δεξιά αποκομμένα δεδομένα βλέποντάς τα και από την πρακτική και από τη θεωρητική οπτική γωνία. Πιο συγκεκριμένα, η NPMLE εκτιμήτρια της συνάρτησης επιβίωσης δεν έχει κλειστού τύπου μορφή και μπορεί μόνο να ορισθεί χρησιμοποιώντας επαναληπτικούς αλγορίθμους.

Μερικές φορές το ενδιαφέρον μας στρέφεται και στην εκτίμηση της συνάρτησης διακινδύνευσης η οποία μπορεί να μας δώσει μεγαλύτερη επίγνωση ως προς τη μεταβλητή που μας ενδιαφέρει από ότι η συνάρτηση επιβίωσης. Για αυτό το λόγο, είναι σύνηθες και φυσικό να εφαρμόζουμε ασθενώς παραμετρικές προσεγγίσεις ή τεχνικές που εξομαλύνουν την εκτίμηση όπως η εκτίμηση του πυρήνα (kernel) και των spline. Και αυτό συμβαίνει γιατί οι παραμετρικές προσεγγίσεις συνήθως εμπλέκουν επιλογή μοντέλου ή εξέταση αυτού και οι μη παραμετρικές προσεγγίσεις συχνά δίνουν εκτιμήσεις που είναι μη ακριβείς. Σε αυτές τις περιπτώσεις συνηθίζονται τεχνικές εξομάλυνσης εκτιμήσεων.

Στην υποενότητα 3.2 θα ασχοληθούμε με την NPMLE εκτιμήτρια μιας συνάρτησης επιβίωσης βασιζόμενη στην Περίπτωση I (Case I) των αποκομμένων δεδομένων σε διάστημα (interval-censored) ή των δεδομένων της τρέχουσας κατάστασης (current status). Για αυτό το συγκεκριμένο είδος αποκομμένων δεδομένων σε διάστημα μία κλειστού τύπου μορφή είναι διαθέσιμη για την NPMLE.

Στη συνέχεια στην υποενότητα 3.3 θα θεωρήσουμε την NPMLE της συνάρτησης επιβίωσης βασιζόμενη σε εντός διαστημάτων αποκομμένα δεδομένα γενικού τύπου.

Και τέλος στην υποενότητα 3.4 θα αναφερθούμε σε τρεις αλγορίθμους καθορισμού της NPMLE για την Περίπτωση II (Case II) των αποκομμένων δεδομένων σε διάστημα.

3.2 NPMLE για δεδομένα τρέχουσας κατάστασης

Σε αυτή τη παράγραφο θεωρούμε την Περίπτωση I των αποκομμένων δεδομένων σε διάστημα παραδείγματος χάριν: δεδομένα τρέχουσας κατάστασης (current status), ή χρόνων αποτυχίας που προέρχονται από *n* ανεξάρτητα άτομα. Έστω ότι το T_i δηλώνει το χρόνο επιβίωσης που μας ενδιαφέρει με συνάρτηση επιβίωσης $S(t)$ και θεωρώντας ότι τα παρατηρηθέντα δεδομένα έχουν τη μορφή: $\{(C_i, \delta_i),\ i = 1, ..., n\}$, όπου C_i δηλώνει το χρόνο παρατήρησης του ατόμου i ανεξαρτήτως του T_i και $\delta_i = I(T_i \le C_i)$.

Τότε η συνάρτηση πιθανοφάνειας θα έχει τη μορφή

$$L_s(S(t)) = \prod_{i=1}^{n} [S(C_i)]^{1-\delta_i} [1 - S(C_i)]^{\delta_i}$$

Έστω $(s_j)_{j=0}^m$ ότι δηλώνει τα μοναδικώς διατεταγμένα στοιχεία του $\{0,\ C_i;\ i=1,...,\ n\}$. Ορίζουμε τη συνάρτηση $r_j=\sum_{i=1}^{n}\delta_i I(C_i=s_j)$ η οποία υπολογίζει τον αριθμό των ατόμων που έχουν παρατηρηθεί τη χρονική στιγμή s_j και βρέθηκαν να έχουν αποτύχει και τη συνάρτηση $n_j=\sum_{i=1}^{n}I(C_i=s_j)$ που υπολογίζει τον αριθμό των ατόμων που έχουν παρατηρηθεί τη χρονική στιγμή s_j, j=1,..., m. Τότε η συνάρτηση πιθανοφάνειας $L_s(S(t))$ μπορεί να ξαναγραφτεί ως:

$$L_s(S(t))=\prod_{i=1}^{m}[S(s_j)]^{n_j-r_j}[1-S(s_j)]^{r_j}=\prod_{j=1}^{m}[F(s_j)]^{r_j}[1-F(s_j)]^{n_j-r_j}$$

Είναι προφανές ότι η συνάρτηση πιθανοφάνειας L_s εξαρτάται από την S ή την F μόνο μέσω των τιμών που παίρνει το s_j. Αυτό σημαίνει ότι μπορούμε να εκτιμήσουμε την S ή την F μόνο σε αυτά τα s_j. Παρατηρώντας τον περιορισμό $F(s_1)\le...\le F(s_m)$ μπορούμε να δείξουμε ότι η μεγιστοποίηση της σε σχέση με το $F(s_j)_{j=1}^m$ είναι ίση με την ελαχιστοποίηση του

$$\sum_{j=1}^{m}n_j[\frac{r_j}{n_j}-F(s_j)]^2$$

που υπόκειται στον περιορισμό $F(s_1)\le...\le F(s_m)$ (Robertson et al., 1988). Το σύνολο των τιμών του $F(s_j)_{j=1}^m$ που ελαχιστοποιεί αυτό το άθροισμα συνήθως αναφέρεται ως ισοτονική παλινδρόμηση (isotonic regression) του $\{\frac{r_1}{n_1},...,\ \frac{r_m}{n_m}\}$ με βάρη $\{n_1,...,\ n_m\}$ (Barlow et al., 1972; Robertson et al., 1988). Χρησιμοποιώντας τη φόρμουλα μεγιστοποίησης-ελαχιστοποίησης για την ισοτονική παλινδρόμηση, η NPMLE της F τη χρονική στιγμή s_j έχει την τιμή:

$$\hat{F}(s_j)=\max_{u\le j}\min_{v\ge j}\frac{\sum_{j=u}^{v}d_j}{\sum_{j=u}^{v}n_j}$$

δοθέντος ότι η τιμή της NPMLE του S τη χρονική στιγμή s_j είναι $1-\hat{F}(s_j)$. Με άλλα λόγια, η NPMLE του S έχει κλειστού τύπου μορφή.

3.3 Χαρακτηριστικά της NPMLE για την Περίπτωση II των αποκομμένων δεδομένων σε διάστημα

Όπως έχουμε αναφέρει, τα αποκομμένα δεδομένα σε διάστημα εμφανίζονται σε μεγάλη κλίμακα σε κλινικές δοκιμές όπου, παραδείγματος χάριν, οι ασθενείς μπορούν να παρακολουθηθούν μόνο περιοδικά και το γεγονός που μας ενδιαφέρει γνωρίζουμε ότι έχει συμβεί εντός ενός χρονικού διαστήματος.

Για την εκτίμηση της συνάρτησης επιβίωσης από αποκομμένα δεδομένα σε διάστημα οι Peto (1973) και Turnbull (1976) πρότειναν αντίστοιχα έναν Newton-Raphson αλγόριθμο και έναν EM αλγόριθμο για να υπολογίσουν τη μη-παραμετρική εκτιμήτρια μεγίστης πιθανοφάνειας (NPMLE). Αργότερα, ο Finkelstein (1986) θεώρησε το μοντέλο αναλογικής διακινδύνευσης του Cox για τις ίδιες συνθήκες δεδομένων και πρότεινε έναν αλγόριθμο Newton-Raphson υπολογισμού της NPMLE του συντελεστή παλινδρόμησης. Μία αντίστοιχη αναφορά γίνεται από τους Huang και Wellner (1995) οι οποίοι θεσπίζουν την ασυμπτωτική κανονικότητα της Normal Probability Maximum Likelihood Estimator του συντελεστή παλινδρόμησης. Οι Kooperberg και Clarkson (1997) διαμόρφωσαν τη βασική συνάρτηση διακινδύνευσης κάνοντας χρήση των splines. Και οι Satten et al. (1996) και Goggins et al. (1998) πρότειναν την οριακή προσέγγιση προσάπτοντας τις τάξεις (βαθμούς) των αποκομμένων χρόνων επιβίωσης.

Ας θεωρήσουμε τώρα μία μελέτη με χρόνους αποτυχίας που αποτελείται από n ανεξάρτητα άτομα από έναν ομοιογενή πληθυσμό με συνάρτηση επιβίωσης S(t). Έστω T_i ο χρόνος επιβίωσης που μας ενδιαφέρει για το i άτομο με i=1,..., n. Επιπλέον τα αποκομμένα δεδομένα του διαστήματος ως προς τα T_i θεωρούμε ότι έχουν παρατηρηθεί και δίνονται από τη σχέση:

$$\mathbf{O} = \{(L_i, R_i];\ i = 1, ...,\ n\}$$

όπου το διάστημα $(L_i, R_i]$ έχει παρατηρηθεί ότι περιέχει τον T_i. Επιπλέον υποθέτουμε ότι ο στόχος μας είναι να παράγουμε την NPMLE από τη συνάρτηση επιβίωσης S(t).

Έστω $\{s_j\}_{j=0}^{m}$ τώρα να δηλώνει τα μοναδικώς διατεταγμένα στοιχεία του $\{0, L_i, R_i;\ i=1,...,\ n\}$ όπως ακριβώς και στην προηγούμενη ενότητα.

Ορίζουμε $a_{ij} = I(s_j \in (L_i, R_i))$ και $p_j = S(s_{j-1}) - S(s_j)$ με $i=1,...,\ n,\ j=1,...,\ m$.

Τότε η συνάρτηση πιθανοφάνειας είναι:

$$L_s(p) = \prod_{i=1}^{n}[S(L_i) - S(R_i)] = \prod_{i=1}^{n}\sum_{j=1}^{m} a_{ij}\, p_j \qquad (3.3.1)$$

όπου $\mathbf{p} = \{p_1,...,\ p_m\}'$.

Όσον αφορά τα δεδομένα τρέχουσας κατάστασης (current status data), είναι εύκολο να δείξουμε ότι η συνάρτηση πιθανοφάνειας L_s εξαρτάται από την S μόνο μέσω των τιμών της $S(s_j)_{j=1}^{m}$ και όχι από το πως το S μεταβάλλεται μεταξύ των s_j. Με άλλα λόγια, η NPMLE του S μπορεί να προσδιορισθεί μοναδικά μόνο από τις τιμές σε αυτές τις χρονικές στιγμές s_j και ο προσδιορισμός της είναι ίσος με τη μεγιστοποίηση της $L_s(\mathbf{p})$ σε σχέση με το **p** άτομο υπό τους περιορισμούς:

$$\sum_{j=1}^{m} p_j = 1 \text{ και } p_j \geq 0 \ (j=1,...,\ m)$$

(Gentleman and Geyer, 1994; Turnbull, 1976).

Στα ακόλουθα, ως συνήθως, θεωρείται ότι η NPMLE του S ή του F(t)=1-S(t) αντιστοιχεί σε μία διακριτή κατανομή με άλματα μόνο στα $\{s_j\}_{j=1}^{m}$ εκτός και αν ορίζεται διαφορετικά. Στη συνέχεια, προσδιορίζεται η NPMLE μεγιστοποιώντας τη συνάρτηση πιθανοφάνειας που δίνεται από τη σχέση (3.3.1) πάνω από όλες τις διακριτές συναρτήσεις επιβίωσης ή κατανομής που είναι σταθερές μεταξύ των σημείων $s_0 < s_1 < ... < s_m$. Σε μερικές περιπτώσεις, μπορεί να μας ενδιαφέρει να μεγιστοποιήσουμε τη συνάρτηση πιθανοφάνειας ως προς ένα διαφορετικό σύνολο των συναρτήσεων επιβίωσης ή κατανομής. Παραδείγματος χάριν, μερικές φορές μπορεί να είναι λογικό να υποθέσουμε ότι η F είναι εξομαλυμένη και έτσι να θέλουμε η NPMLE να είναι και αυτή "στρωτή".

Για τον προσδιορισμό της NPMLE του S ή του F χρειάζεται να καταλήξουμε στην επιλογή ενός αλγορίθμου. Επιπλέον μπορεί να χρειαστεί να διαχειριστούμε και κάποια προβλήματα που θα προκύψουν, όπως το εξής: εάν το $\hat{\mathbf{p}} = (\hat{p}_1, ..., \hat{p}_m)'$ είναι η NPMLE τότε κάποια στοιχεία του $\hat{\mathbf{p}}$ θα μπορούσε να είναι μηδέν και θα ήταν πολύ χρήσιμο και βοηθητικό εάν γνωρίζαμε αυτά τα μηδενικά στοιχεία πριν "τρέξουμε" ένα προκαθορισμένο πρόγραμμα. Για το συγκεκριμένο παράδειγμα, θα μπορούσαμε να βασιστούμε στο γεγονός ότι κάθε $\hat{p}_j$ μπορεί να είναι μη μηδενικό εκτός εάν $s_{j-1} = L_i$ για κάποιο i και $s_j = R_k$ για κάποιο k διαφορετικό από το i με i, k=1,..., n (Peto, 1973; Turnbull, 1976). Με αυτό τον τρόπο μπορούμε να εστιάσουμε στα $\hat{p}_j$ που ικανοποιούν τη συνθήκη, αλλά φυσικά, κάποια από αυτά θα μπορούσαν να ήταν μηδέν. Η χρήση αυτού του γεγονότος μπορεί να μειώσει σημαντικά τον αριθμό των χρονικών στιγμών m που πρέπει να ληφθούν υπόψη όσον αναφορά την υπολογιστική διαδικασία. Επιπρόσθετα, κάποιος θα μπορούσε να εφαρμόσει το πολλαπλασιαστικό κριτήριο του Lagrange (Lagrange multiplier criterion) που περιγράφουμε παρακάτω.

Μία εναλλακτική προσέγγιση για τον προσδιορισμό μη μηδενικών $\hat{p}_j$ είναι να βρούμε όλες τις χρονικές στιγμές s_j ή ένα σύνολο από ξένα μεταξύ τους διαστήματα που αποτελεί το πιθανό στήριγμα της NPMLE του S ή του F. Τα διαστήματα είναι αυτά των οποίων τα αριστερά και τα δεξιά άκρα δίνονται από μερικά από τα L_i και R_i αντίστοιχα και τα οποία δεν περιέχουν άλλα εκτός από τα άκρα. Αυτό συνήθως αναφέρεται και ως προσέγγιση του Turnbull και ο υπολογισμός της NPMLE αποτελείται από δύο βήματα:

1. Το πρώτο είναι η εξεύρεση των πιθανών άκρων ή των διαστημάτων και

2. το δεύτερο είναι η μεγιστοποίηση της $L_s(\mathbf{p})$.

Για να αποφασίσουμε εάν εκτιμήτρια $\hat{\mathbf{p}} = (\hat{p}_1, ..., \hat{p}_m)'$ του $\mathbf{p}$ είναι η NPMLE, ορίζουμε

$$d_j(\mathbf{p}) = \sum_{i=1}^{n} \frac{a_{ij}}{\sum_{l=1}^{m} a_{il} p_l}, \; j = 1, ..., \; m. \tag{3.3.2}$$

Το πολλαπλασιαστικό κριτήριο Lagrange προερχόμενο από τη θεωρία γραφημάτων λέει ότι το $\hat{\mathbf{p}}$ είναι η NPMLE εάν: $d_j(\hat{\mathbf{p}}) = n$ για όλα τα j=1, . . . , m (Gentleman and Geyer, 1994).

Επιπλέον, μπορεί να αποδειχθεί ότι η $\hat{\mathbf{p}}$ είναι NPMLE εάν και μόνο εάν $d_j(\hat{\mathbf{p}}) = n$ για όλα τα j (Böhning et al., 1996).

Για να εξηγήσουμε το πολλαπλασιαστικό κριτήριο Lagrange, θεωρούμε το σύνολο δεδομένων {(0,1],(1,3],(1,3],(0,2],(0,2],(2,3]} από τους Gentleman και Geyer (1994). Αποτελείται από 6 παρατηρηθέντα διαστήματα και για αυτά τα δεδομένα έχουμε ότι $\{s_j\}_{j=0}^{3} = \{0,1,2,3\}$

$$\mathbf{A} = (a_{ij}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{και} \quad L_s(\mathbf{p}) = \prod_{i=1}^{6} (a_{i1}p_1 + a_{i2}p_2 + a_{i3}p_3)$$

Για την εξεύρεση της NPMLE χρειάζεται να μεγιστοποιήσουμε τη συνάρτηση πιθανοφάνειας L_s ως προς $\mathbf{p} = (p_1, p_2, p_3)'$ δοθέντος ότι $\sum_{j=1}^{3} p_j = 1$ και $p_1,\ p_2,\ p_3 \geq 0$.

Υποθέτουμε ότι μας έχουν δοθεί 2 εκτιμήτριες :

$$\hat{p}_1 = (\frac{1}{2}, 0, \frac{1}{2})' \text{ και } \hat{p}_2 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})'$$

Για να ελέγξουμε ποια από τις 2 είναι η εκτιμήτρια NPMLE σημειώνουμε ότι ισχύουν τα κάτωθι

$$d_1 = \frac{1}{p_1} + \frac{2}{p_1 + p_2},\ d_2 = \frac{2}{p_2 + p_3} + \frac{2}{p_1 + p_2},\ d_3 = \frac{2}{p_2 + p_3} + \frac{1}{p_3}.$$

Είναι προφανές ότι δεν είναι η εκτιμήτρια NPMLE η $\hat{p}_1$ γιατί $d_2 = 8 > n = 6$ ενώ η δεύτερη εκτιμήτρια ικανοποιεί το πολλαπλασιαστικό κριτήριο Lagrange και είναι η NPMLE γιατί:

$$d_1 = d_2 = d_3 = 6 = n\,.$$

Η μοναδικότητα της εκτιμήτριας NPMLE είναι ένα ακόμη θέμα που συχνά πρέπει να ελεγθεί. Ως προς αυτό να σημειώσουμε ότι η $\log L_s(\mathbf{p})$ είναι μία κοίλη συνάρτηση αλλά μπορεί να μην είναι αυστηρά κοίλη. Ορίζουμε $\mathbf{A} = (a_{ij})$ τον nxm πίνακα, τότε η εκτιμήτρια NPMLE είναι μοναδική εάν ο βαθμός του A είναι ίσος με m, αυτό εγγυάται ότι η $p_j = 0$ συνάρτηση λογαριθμο-πιθανοφάνειας (log likelihood) είναι αυστηρά κοίλα και έτσι έχει μοναδικό μέγιστο. Στην πράξη ο βαθμός του A είναι συνήθως μικρότερος του m. Για αυτό το λόγο δηλώνουμε με A* τον υποπίνακα του A που αποτελείται από όλες τις στήλες του A έτσι ώστε είτε το αντίστοιχο $p_j > 0$ ή $d_j(\mathbf{p}) = n$ εάν το αντίστοιχο $p_j = 0$. Τότε μία επαρκής συνθήκη για τη μοναδικότητα της εκτιμήτριας NPMLE, που δίνεται από τους Gentleman και Geyer (1994), είναι ότι ο βαθμός του A* είναι ίσος με τον αριθμό των στηλών του.

3.4 Αλγόριθμοι για την Περίπτωση II των αποκομμένων δεδομένων σε διάστημα

Σε αυτή την ενότητα θα παρουσιάσουμε τρεις αλγορίθμους που μπορούμε να χρησιμοποιήσουμε για τον προσδιορισμό της εκτιμήτριας NPMLE των S, F, **p.** Ο πρώτος αλγόριθμος είναι ο αυτο-συνεπής (self-consistency) που αναπτύχθηκε από τον Turnbull (1976) και μπορεί να θεωρηθεί σαν μία εφαρμογή του EM (Expectation-Maximisation) αλγορίθμου (Dempster et al., 1977). Ο δεύτερος είναι ο αλγόριθμος ICM που τον εισήγαγαν πρώτοι οι Groeneboom και Wellner (1992) και στη συνέχεια τροποποιήθηκε από τον Jongbloed (1998). Μετατρέπει τη μεγιστοποίηση της συνάρτησης πιθανοφάνειας (3.3.1) σε μεγιστοποίηση της τετραγωνικής συνάρτησης χρησιμοποιώντας τη θεωρία της ισοτονικής παλινδρόμησης. Ο τρίτος αλγόριθμος είναι υβριδικός που έχει προταθεί από τους Wellner και Zhan (1997), ο οποίος αναφέρεται ως αλγόριθμος EM-ICM και συνδυάζει τον αυτο-συνεπή αλγόριθμο και τον ICM αλγόριθμο.

3.4.1 Ο αυτο-συνεπής αλγόριθμος του Turnbull

Μία αυτο-συνεπής εκτίμηση (self-consistent) συνήθως αναφέρεται ως μία εκτίμηση που μπορεί να χαρακτηρισθεί από την αντίστοιχη εξίσωση και το όριο των επαναλήψεων προκύπτει από αυτή την εξίσωση (Efron, 1967). Συνεπώς, η εκτίμηση μπορεί να προσδιοριστεί επαναληπτικά. Για να δημιουργηθεί η αυτο-συνεπής εξίσωση για αποκομμένα δεδομένα σε διάστημα μία άμεση και διαισθητική προσέγγιση βασισμένη σε εμπειρικές εκτιμήσεις είναι αναγκαία. Μια πιο γενική προσέγγιση συμπεριφέρεται στα αποκομμένα δεδομένα σε διάστημα σα να είναι ατελή (incomplete data) και στη συνέχεια εφαρμόζει τον ΕΜ αλγόριθμο. Στα ακόλουθα, χρησιμοποιούμε τη γενική προσέγγιση και στη συνέχεια ακολουθούν κάποια σχόλια ως προς την άμεση προσέγγιση.

Ορίζουμε

$$\mathbf{C_p} = \{\mathbf{p} \in [0,1]^m;\ \sum_{j=1}^{m} p_j = 1,\ p_j \geq 0\}$$

ένα υπό-διάστημα του R^m.

Στη συνέχεια μπορούμε να βρούμε την εκτιμήτρια NPMLE του **p** μεγιστοποιώντας τη συνάρτηση πιθανοφάνειας που δίνεται από τη σχέση (3.3.1) πάνω στην περιοχή του $\mathbf{C_p}$. Για να εφαρμόσουμε τον ΕΜ αλγόριθμο, υποθέτουμε ότι ο ακριβής χρόνος αποτυχίας των δεδομένων $\{T_i\}_{i=1}^n$ που μελετάμε είναι γνωστός. Η συνάρτηση λογαριθμο-πιθανοφάνειας για τα ολοκληρωμένα δεδομένα είναι:

$$l_s(\mathbf{p}; T_1, ...,\ T_n) = \log[\prod_{i=1}^{n} dF(T_i)] = \sum_{j=1}^{m} d_j^* \log p_j$$

όπου $\{T_i\}_{i=1}^n$ ο ακριβής χρόνος αποτυχίας

$$d_j^* = \sum_{i=1}^{n} I(T_i = s_j),\ j = 1, ...,\ m.$$

Έστω τώρα $\hat{\mathbf{p}}^c$ να δηλώνει την τρέχουσα εκτίμηση του **p** και $\hat{\mathbf{p}}^u$ την αναβαθμισμένη εκτίμηση του **p.**

Στο βήμα Ε του αλγορίθμου ΕΜ, χρειάζεται να υπολογισθεί η υπό όρους προσδοκία του $l_s(\mathbf{p};T_i^{'}s)$ δοθέντος $\hat{\mathbf{p}}^c$ και τα παρατηρηθέντα δεδομένα **O,** που έχουν τον τύπο:

$$E[l_s(\mathbf{p};T_i^{'}s)\,|\,\hat{\mathbf{p}}^c,\mathbf{O}]=\sum_{j=1}^{m}\log(p_j)E(d_j^*\,|\,\hat{\mathbf{p}}^c,\mathbf{O})=\sum_{j=1}^{m}d_j(\hat{\mathbf{p}}^c)p_j^c\log(p_j)$$

με $d_j(\mathbf{p})$ να έχει ορισθεί από τη σχέση (3.3.2).

Στο βήμα Μ του αλγορίθμου ΕΜ, χρειάζεται να μεγιστοποιηθεί η υπό όρους προσδοκία όπως έχει ήδη δοθεί πάνω στην περιοχή $\mathbf{C_p}$. Χρησιμοποιώντας την προσέγγιση Lagrange μεγιστοποιούμε την

$$\sum_{j=1}^{m}d_j(\hat{\mathbf{p}}^c)p_j^c\log(p_j)+\lambda(1-\sum_{j=1}^{m}p_j)$$

Διαφοροποιώντας την παραπάνω συνάρτηση σε σχέση με το p_j και παραγωγίζοντας στο 0 έχουμε:

$$p_j=\frac{d_j(\hat{\mathbf{p}}^c)p_j^c}{\lambda}$$

Τότε ακολουθεί από $\sum_{j=1}^{m}p_j=1$ ότι λ=n και

$$\hat{p}_j^u=\frac{d_j(\hat{\mathbf{p}}^c)p_j^c}{n}=\frac{1}{n}E[\sum_{i=1}^{n}I(T_i=s_j)\,|\,\hat{\mathbf{p}}^c,\mathbf{O}] \qquad (3.4.1.1)$$

η οποία εισάγει τον ακόλουθο αυτο-συνεπή αλγόριθμο για την εκτιμήτρια NPMLE.

Βήμα **1.** Επιλέγουμε μία αρχική εκτίμηση $\hat{\mathbf{p}}^0$ του **p.**

Βήμα **2.** Στην *l*-ιοστή επανάληψη, ορίζουμε την αναβαθμισμένη εκτίμηση που συμβολίζεται

$\hat{\mathbf{p}}^{(l)}=(\hat{p}_1^{(l)},.....\hat{p}_m^{(l)})$ του **p**

$$\hat{p}_j^{(l)} = \frac{d_j(\hat{\mathbf{p}}^{(l-1)})\, p_j^{(l-1)}}{n} = \frac{1}{n}\sum_{i=1}^{n} \frac{a_{ij}\,\hat{p}_j^{(l-1)}}{\sum_{k=1}^{m} a_{ik}\,\hat{p}_k^{(l-1)}} \quad \text{με } j = 1,\ldots,\ m.$$

Βήμα **3.** Επαναλαμβάνουμε το βήμα **2.** έως ότου η επιθυμητή σύγκλιση επιτευχθεί.

Η εκτίμηση $\hat{\mathbf{p}} = (\hat{p}_1,\ldots,\ \hat{p}_m)'$ που προκύπτει από αυτό τον αλγόριθμο είναι η λύση της αυτο-συνεπής εξίσωσης

$$\hat{p}_j = \frac{1}{n} E[\sum_{i=1}^{n} I(T_i = s_j) \mid \hat{\mathbf{p}}, \mathbf{O}] \qquad (3.4.1.2)$$

Επίσης, μπορεί να δειχθεί ότι η συνάρτηση πιθανοφάνειας αυξάνεται μετά από κάθε επανάληψη. Αν και η εκτίμηση $\hat{p}$ δεν είναι η εκτιμήτρια NPMLE της **p,** μπορεί να ελεγχθεί χρησιμοποιώντας το κριτήριο που αναφέραμε στην υποενότητα 3.3.

Μία άμεση προσέγγιση που επίσης δίνει ο αυτο-συνεπής αλγόριθμος και η εξίσωση (3.4.1.2) φαίνεται από το δεύτερο μέλος της εξίσωσης (3.4.1.1). Δοθέντος ότι η $\mathbf{p} = \hat{\mathbf{p}}^c$ ποσότητα αντιπροσωπεύει τις φορές όπου ο εκτιμώμενος αριθμός ατόμων των οποίων οι χρόνοι επιβίωσης είναι ίσοι με s_j, τους οποίους αποδίδει ο αλγόριθμος βασιζόμενος σε εμπειρικές εκτιμήσεις με $\hat{F}(t) = \sum_{s_j \le t} \hat{p}_j$ η εξίσωση (3.4.1.2) δίνει:

$$\hat{F}(t) = \frac{1}{n} E[\sum_{i=1}^{n} I(T_i \le t) \mid \hat{F}, \mathbf{O}] \qquad (3.4.1.3)$$

3.4.2 Ο επαναληπτικός αλγόριθμος Convex Minorant

Για να περιγράψουμε τον ICM αλγόριθμο (Iterative Convex Minorant), ορίζουμε

$$C_x = \{x_1,\ldots, x_{m-1})' \in R^{m-1}; 0 \le x_1 \le \ldots \le x_{m-1} \le 1\}$$

ένα υποδιάστημα του R^{m-1} και έστω $\beta_j = F(s_j),\ j = 1,\ldots,\ m.$ Με $\beta_0 = 0$, $\beta_m = 1$, $\beta = (\beta_1,\ldots,\ \beta_{m-1})'$ η συνάρτηση πιθανοφάνειας (3.3.1) μπορεί τότε να ξαναγραφτεί ως εξής:

$$L_s(\boldsymbol{\beta}) = \prod_{i=1}^{n} \sum_{j=1}^{m} a_{ij}(\beta_j - \beta_{j-1}) \qquad (3.4.2.1)$$

και η εκτιμήτρια NPMLE λαμβάνεται μεγιστοποιώντας την $L_s(\boldsymbol{\beta})$ πάνω στο C_x.

Ο ICM αλγόριθμος βασίζεται στα εξής δύο γεγονότα:

Πρώτα υποθέτουμε ότι το g είναι μία διαφοροποιημένη κοίλη απεικόνιση από το R^{m-1} στο R και C είναι ένας κοίλος κώνος. Υποθέτουμε επίσης ότι g(**x**) λαμβάνει το μέγιστο στην περιοχή C στο $\hat{\mathbf{x}}$. Έστω **W** ένας θετικά ορισμένος (m-1)x(m-1) πίνακας και **y** ένα σταθερό σημείο R^{m-1}. Ορίζουμε:

$$g^*(\mathbf{x} \mid \mathbf{y}, \mathbf{W}) = -\frac{1}{2}(\mathbf{x}-\mathbf{y})'\mathbf{W}(\mathbf{x}-\mathbf{y}) \text{ για } \mathbf{x} \in R^{m-1} \qquad (3.4.2.2)$$

και επιπλέον υποθέτουμε ότι το $\hat{\mathbf{x}}^* \in C$ μεγιστοποιεί την $g^*(\mathbf{x} \mid \mathbf{y}, \mathbf{W})$ στο C.

Το $\mathbf{y} = \hat{\mathbf{x}} + \mathbf{W}^{-1}\nabla g(\hat{\mathbf{x}})$ όπου $\nabla g(\mathbf{x})$ είναι το διάνυσμα των παραγώγων του g στο **x** και το $\hat{\mathbf{x}}^*$ μεγιστοποιεί την $g^*(\mathbf{x} \mid \mathbf{y}, \mathbf{W})$ πάνω στο C εάν και μόνο εάν $\hat{\mathbf{x}}^* = \hat{\mathbf{x}}$ (Groeneboom και Wellner, 1992).

Το δεύτερο γεγονός αφορά τη μεγιστοποίηση της τετραγωνικής συνάρτησης στην περιοχή του C_x. Όπως και πριν $C = C_{\mathbf{x}}$ και $\mathbf{W} = diag(w_j)$ ένας θετικά ορισμένος διαγώνιος πίνακας.

Ορίζουμε ακόμα $P_0 = (0,0)$ και $P_u = (\sum_{i=1}^{u} w_i, \sum_{i=1}^{u} w_i y_i),\ 1 \le u \le m-1$ σημεία του R^2 για κάποια σταθερά $\mathbf{y} = (y_1, ..., y_{m-1})' \in R^{m-1}$. Το σύνολο των σημείων $\{P_u;\ u = 0, ..., m-1\}$ συνήθως αναφέρεται ως αθροιστικό διάγραμμα γιατί οι συντεταγμένες του P_u είναι αθροίσματα των διανυσμάτων $(w_1, ..., w_{m-1})'$ και $(w_1 y_1, ..., w_{m-1} y_{m-1})'$.

Τότε το $\hat{\mathbf{x}}_i^*$ δίνεται από το αριστερό παράγωγο κυρτής π.χ. της συνάρτησης μεγίστης κυρτότητας κάτω από το αθροιστικό διάγραμμα $\{P_u;\ u = 0, ..., m-1\}$ αξιολογούμενο στο P_j.

Το πρώτο γεγονός προτείνει ότι εάν το $\hat{\mathbf{x}}$ είναι γνωστό, η μεγιστοποίηση της γενικής συνάρτησης g(**x**) είναι ίση με τη μεγιστοποίηση της τετραγωνικής συνάρτησης $g^*(\mathbf{x})$. Το δεύτερο γεγονός δίνει τη μεγιστοποίηση στο σημείο $\hat{\mathbf{x}}^*$ για μία συγκεκριμένη τετραγωνική συνάρτηση. Φυσικά το $\hat{\mathbf{x}}$ είναι άγνωστο και το πρώτο γεγονός δε δίνει μια άμεση διαδικασία μεγιστοποίησης, αλλά μπορεί να δώσει μια επαναληπτική μέθοδο. Τα δύο γεγονότα μαζί ενεργοποιούν τον ICM αλγόριθμο που θα περιγράψουμε.

Βήμα **1.** Επιλέγουμε μία αρχική εκτίμηση $\hat{\boldsymbol{\beta}}^0$ του **β.**

Βήμα **2.** Στην *l*-ιοστή επανάληψη, ορίζουμε την αναβαθμισμένη εκτίμηση που συμβολίζεται $\hat{\boldsymbol{\beta}}^{(l)} = (\hat{\beta}_1^{(l)}, ..., \hat{\beta}_{m-1}^{(l)})$ του $\boldsymbol{\beta}$ καθώς το $\hat{\mathbf{x}}^*$ μεγιστοποιεί την $g^*(\mathbf{x} \mid \mathbf{y}, \mathbf{W}(\hat{\boldsymbol{\beta}}^{(l-1)}))$ με $\mathbf{y} = \hat{\boldsymbol{\beta}}^{(l-1)} - W^{-1}(\hat{\boldsymbol{\beta}}^{(l-1)}) \nabla l_s(\hat{\boldsymbol{\beta}}^{(l-1)})$ και όπως πριν ο W είναι ένας θετικά ορισμένος διαγώνιος πίνακας και η $l_s(\boldsymbol{\beta})$ συνάρτηση λογαριθμο-πιθανοφάνειας από τη σχέση (3.4.2.1).

Βήμα **3.** Επιστέφουμε στο βήμα **2.** έως ότου η επιθυμητή σύγκλιση επιτευχθεί.

Ο Jongbloed (1998) δείχνει ότι ο αλγόριθμος ICM μπορεί να μην αυξάνει τη συνάρτηση λογαριθμο-πιθανοφάνειας μετά από κάθε επανάληψη και να μη συγκλίνει ολικά, γι΄αυτό το λόγο προτείνει την προσθήκη αναζήτησης μιας γραμμής μες στον αλγόριθμο βασισμένη στην επίτευξη της ολικής σύγκλισης. Έστω g, g*, C και $\hat{\mathbf{x}}$ όπως είχαν δοθεί στο πρώτο γεγονός. Για δεδομένο **x,** θετικά ορισμένο διαγώνιο πίνακα **W(x)** και **A(x)** να είναι διάνυσμα **z** στο οποίο η $g^*(\mathbf{z} \mid \mathbf{y}, \mathbf{W})$ επιτυγχάνει το μέγιστό της με $\mathbf{y} = \mathbf{x} - \mathbf{W}^{-1}(\mathbf{x}) \nabla l_s(\mathbf{x})$. Τότε για $\mathbf{x} \neq \hat{\mathbf{x}}$ και επαρκώς μικρό λ>0, $g(\mathbf{x} + \lambda(\mathrm{A}(\mathbf{x}) - \mathbf{x})) > g(\mathbf{x})$. Αυτό προτείνει ότι η αναζήτηση γραμμής, που δίνεται στο βήμα **2.1** παρακάτω, μπορεί να ενσωματωθεί στον ICM αλγόριθμο για να εγγυηθεί την αύξηση της συνάρτησης λογαριθμο-πιθανοφάνειας καθώς και την ολική σύγκλιση του αλγορίθμου.

Έστω 0<ε<0.5 ένας σταθερός αριθμός που ελέγχει τη διαδικασία αναζήτησης γραμμής. Το ακόλουθο βήμα προστίθεται ανάμεσα στα βήματα 2 και 3 του ICM αλγορίθμου όπως αυτός έχει ήδη περιγραφεί.

Βήμα **2.1** Εάν $l_s(\hat{\boldsymbol{\beta}}^{(l)}) > l_s(\hat{\boldsymbol{\beta}}^{(l-1)}) + (1-\varepsilon)[\nabla l_s(\hat{\boldsymbol{\beta}}^{(l-1)})]'(\hat{\boldsymbol{\beta}}^{(l)} - \hat{\boldsymbol{\beta}}^{(l-1)})$ τότε πήγαινε στο βήμα **3.**, διαφορετικά βρες ένα σημείο **z** τέτοιο ώστε $\mathbf{z} = \hat{\beta}^{(l-1)} + \lambda(\hat{\beta}^{(l)} - \hat{\beta}^{(l-1)})$ για $0 \le \lambda \le 1$ που ικανοποιεί

$$\varepsilon[\nabla l_s(\hat{\boldsymbol{\beta}}^{(l-1)})]'(\mathbf{z} - \hat{\boldsymbol{\beta}}^{(l-1)}) \le l_s(\mathbf{z}) - l_s(\hat{\boldsymbol{\beta}}^{(l-1)}) \le (1-\varepsilon)[\nabla l_s(\hat{\boldsymbol{\beta}}^{(l-1)})]'(\mathbf{z} - \hat{\boldsymbol{\beta}}^{(l-1)})$$

με $\hat{\boldsymbol{\beta}}$ να δηλώνει την εκτίμηση που δίνεται από τον ICM αλγόριθμο. Τότε η εκτιμήτρια NPMLE του F και **p** δίνεται από:

$\hat{F}(t) = \hat{\beta}_j$ εάν $s_j \le t < s_{j+1}$ για $j = 0,..., m-1$ και $\hat{p}_j = \hat{\beta}_j - \hat{\beta}_{j-1}$ για $j = 1,..., m$ αντίστοιχα.

Στον αλγόριθμο ICM μία λογική επιλογή για το **W(β)** είναι να πάρουμε:

$$w_j = w_j(\boldsymbol{\beta}) = -\frac{\partial^2}{\partial \beta_j^2} l_s(\boldsymbol{\beta})$$

υποθέτοντας ότι υπάρχει, με j=1,..., m-1. Ο Jongbloed (1998) μελέτησε αυτή την επιλογή και άλλες χρησιμοποιώντας προσομοιώσεις και πρότεινε τη συγκεκριμένη ως την καλύτερη σε σχέση με τις υπόλοιπες και πιο συγκεκριμένα φάνηκε ότι η αναζήτηση γραμμής ή το βήμα **2.1** χρησιμοποιήθηκαν λιγότερο υπό αυτή.

3.4.3 Ο EM Iterative Convex Minorant αλγόριθμος

Σε αυτή την ενότητα θα παρουσιάσουμε τον τρίτο αλγόριθμο, τον αλγόριθμο EM-ICM, που μπορεί να χρησιμοποιηθεί για τον καθορισμό της εκτιμήτριας NPMLE της F. Είναι ένας αλγόριθμος που συνδυάζει τον αυτο-συνεπή αλγόριθμο με τον ICM αλγόριθμο. Πιο συγκεκριμένα αποτελείται από τα ακόλουθα βήματα:

Βήμα **1.** Επιλέγουμε μία αρχική εκτίμηση του **β** τη $\hat{\boldsymbol{\beta}}^0$ και $\hat{\mathbf{p}}^0$ *του* **p**.

Βήμα **2.** Εφαρμόζουμε τα βήματα **2.** και **2.1** του ICM αλγορίθμου ως προς την τρέχουσα εκτίμηση για να λάβουμε μία αναβαθμισμένη εκτίμηση.

Βήμα **3.** Εφαρμόζουμε το βήμα **2.** του αυτο-συνεπή αλγορίθμου για την αναβαθμισμένη εκτίμηση που βρήκαμε στο αμέσως προηγούμενο βήμα για να λάβουμε μία εκ νέου αναβαθμισμένη εκτίμηση. Η αναβαθμισμένη εκτίμηση του βήματος **2.** του αυτό-συνεπή αλγορίθμου είναι $\hat{\mathbf{p}}^{(l-1)}$.

Βήμα **4.** Πάμε πίσω στο βήμα **2.** έως ότου η επιθυμητή σύγκλιση επιτευχθεί.

Χρησιμοποιώντας ένα γενικό θεώρημα για την ολική σύγκλιση της σύνθετης απεικόνισης, οι Wellner και Zhan (1997) δείχνουν ότι ο αλγόριθμος EM-ICM συγκλίνει στην εκτιμήτρια NPMLE εάν υπάρχει και είναι μοναδική και η συνάρτηση λογαριθμο-πιθανοφάνειας είναι συνεχώς διαφορίσιμη. Επίσης έδειξαν χρησιμοποιώντας προσομοιώσεις ότι εφαρμόζοντας τον EM-ICM αλγόριθμο μπορεί να παραλειφθεί η αναζήτηση γραμμής και πάλι όμως να επιτευχθεί η επιθυμητή σύγκλιση.

Για να εφαρμόσουμε τους αλγορίθμους που είδαμε χρειάζεται να διαλέξουμε ένα κριτήριο σύγκλισης. Ένα απλό κριτήριο σύγκλισης βασίζεται στην εγγύτητα των διαδοχικών εκτιμήσεων του F ή του **β** που μπορεί να μετρηθεί π.χ. ως εξής:

$$\sum_{j=1}^{m-1}\left|\hat{\beta}_j^{(l)}-\hat{\beta}_j^{(l-1)}\right|<\varepsilon \qquad (3.4.3.1)$$

με ε μία θετική σταθερά. Ένα ακόμα κριτήριο που χρησιμοποιείται συχνά στη μεγιστοποίηση της συνάρτησης πιθανοφάνειας είναι να ανάγουμε τη σύγκλιση στην αλλαγή της συνάρτησης λογαριθμο-πιθανοφάνειας. Σε αυτή την περίπτωση, οι επαναλήψεις σταματούν εάν:

$$\left|l_s(\hat{\boldsymbol{\beta}}^{(l)})-l_s(\hat{\boldsymbol{\beta}}^{(l-1)})\right|<\varepsilon$$

Τα παραπάνω κριτήρια δε μπορούν να μας πουν εάν η εκτίμηση που δίνεται από τη σύγκλιση είναι τοπικό ή ολικό μέγιστο. Εάν υπάρχουν τοπικά μέγιστα συνήθως προτιμάται το κριτήριο που βασίζεται στις βέλτιστες συνθήκες του Fenchel (Robertson et al., 1988). Υπό αυτό το κριτήριο, σταματάμε τις επαναλήψεις και δεχόμαστε $\hat{\boldsymbol{\beta}}^{(l)}=(\hat{\beta}_1^{(l)},...,\ \hat{\beta}_{m-1}^{(l)})'$ ως την εκτιμήτρια NPMLE της F εάν

$$\left|\sum_{j=1}^{m-1}\hat{\beta}_j^{(l)}\frac{\partial}{\partial\beta_j}l_s(\hat{\boldsymbol{\beta}}^{(l)})\right|<\varepsilon \quad \text{και} \quad \max\{\sum_{u=j}^{m-1}\frac{\partial}{\partial\beta_u}l_s(\hat{\boldsymbol{\beta}}^{(l)});\ j=1,...,\ m-1\}<\varepsilon\,.$$

3.4.4 Παράδειγμα

Σε αυτή την παράγραφο εφαρμόζονται οι τρεις αλγόριθμοι

α) ο αυτό-συνεπής αλγόριθμος του Turnbull (παράγραφος 3.4.1)

β) ο επαναληπτικός αλγόριθμος Convex Minorant (ICM) (παράγραφος 3.4.2)

γ) ο Expectation Maximisation Iterative Convex Minorant αλγόριθμος (EM-ICM) (παράγραφος 3.4.3)

που περιγράψαμε και θα δούμε την ανάλυση και τη σύγκρισή τους. Η μελέτη που χρησιμοποιείται για να εξηγηθούν οι παραπάνω αλγόριθμοι είναι μία αναδρομική έρευνα που παρουσιάστηκε από τους Klein και Moeschberger (1997) και σκοπός της διεξαγωγής της ήταν να συγκρίνουν τις αισθητικές/κοσμικές επιδράσεις που αντιστοιχούν στις δύο θεραπείες:

1) θεραπεία μόνο με ακτινοθεραπεία RT (radiation therapy)

2) θεραπεία ακτινοθεραπείας και χημειοθεραπείας RCT (radiation therapy plus adjuvant chemotherapy).

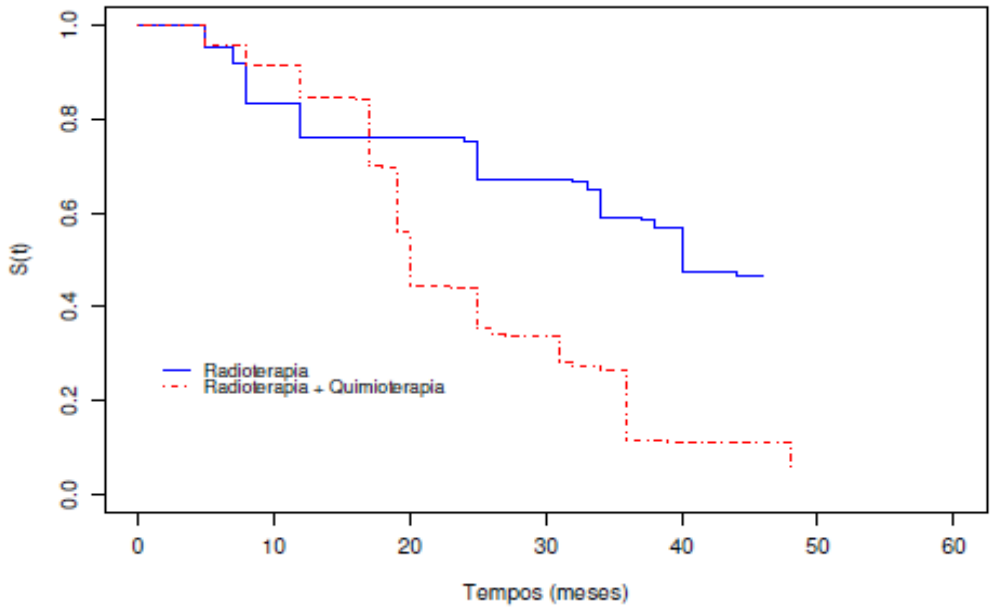
σε γυναίκες με πρόωρη εμφάνιση του καρκίνου του μαστού.

Τα δεδομένα της αναδρομικής μελέτης για τις δύο θεραπείες παρουσιάζονται στον πίνακα 3 της παραγράφου 2.4.6 και αφορούν σε μια ομάδα με 46 ασθενείς που υποβλήθηκαν σε 46 ακτινοβολίες και μία ακόμα με 48 ακτινοβολίες με επιπρόσθετη χημειοθεραπεία. Οι ασθενείς παρακολουθούνταν αρχικά κάθε 4-6 μήνες, αλλά καθώς η περίοδος ανάρρωσης προχωρούσε το διάστημα μεταξύ των επισκέψεων επιμηκυνόταν. Το γεγονός ενδιαφέροντος (δηλαδή η τέλεση του συμβάντος) ήταν η πρώτη εμφάνιση της μέτριας ή σοβαρής συστολής του μαστού. Καθώς οι ασθενείς παρατηρούνταν μόνο σε τυχαίες χρονικές στιγμές, ο ακριβής χρόνος της συστολής του μαστού είναι μόνο γνωστό ότι βρίσκεται μέσα στο διάστημα μεταξύ των επισκέψεων.

Οι ασθενείς οι οποίοι δεν εμφάνισαν μέτρια ή σοβαρή συστολή του μαστού έως την τελευταία τους επίσκεψη κατηγοριοποιήθηκαν ως δεξιά αποκομμένα περιστατικά και ως εκ τούτου το άνω άκρο των διαστημάτων τους θεωρήθηκε ότι είναι το $U_i = \infty$ και το κάτω άκρο του διαστήματος ήταν ο χρόνος που μεσολάβησε από την πρώτη έως και την τελευταία επίσκεψη.

Χρησιμοποιώντας τον αλγόριθμο του Turnbull παίρνουμε τις εκτιμώμενες συναρτήσεις επιβίωσης για την ακτινοθεραπεία μόνο και αντίστοιχα για την ακτινοβολία με τη συμβολή της χημειοθεραπείας των ομάδων που συμμετείχαν όπως βλέπουμε στην πρώτη γραφική παράσταση (Διάγραμμα 1).

Διάγραμμα 1: Εκτιμώμενες καμπύλες επιβίωσης βασισμένες σε αποκομμένα δεδομένα διαστήματος.



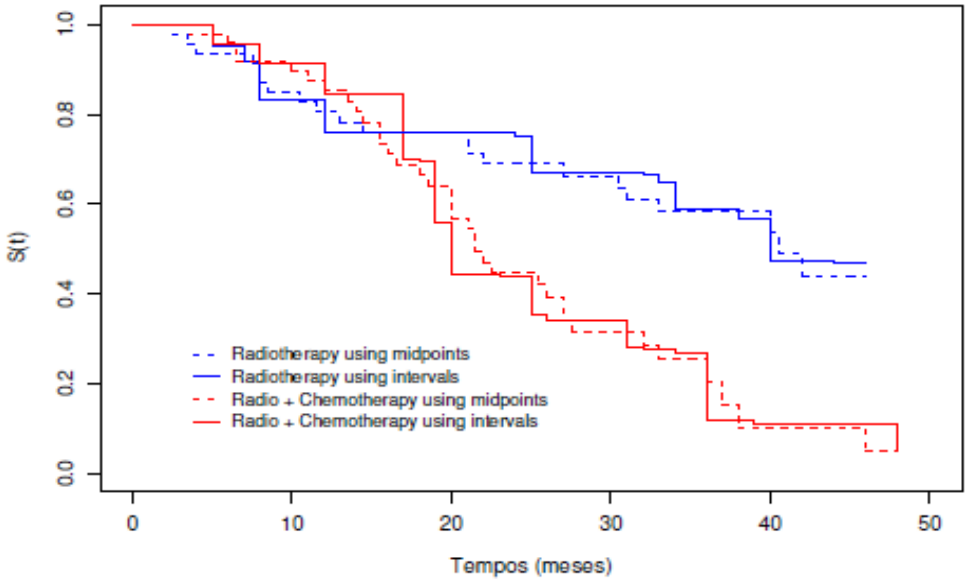
Όπως ήταν αναμενόμενο παρατηρούμε στο διάγραμμα ότι οι εκτιμήτριες NPMLE καθεμίας εκ των συναρτήσεων επιβίωσης που προκύπτουν από τους 3 αλγορίθμους ως προς τη σύγκλισή τους είναι σχεδόν πανομοιότυπες. Επιβεβαιώνεται ότι οι ασθενείς που ανήκαν στην ομάδα θεραπείας με RCT εμφάνισαν συστολή του καρκίνου του μαστού νωρίτερα από ότι οι ασθενείς που ανήκαν στην ομάδα θεραπείας με RT. Το κριτήριο σύγκλισης που χρησιμοποιήσαμε σε όλους τους αλγορίθμους είναι το κριτήριο (3.4.3.1) με $\varepsilon = 10^{-8}$.

Οι επαναλήψεις που απαιτήθηκαν για τον αυτο-συνεπή αλγόριθμο, τον ICM και τον EM-ICM είναι 416, 38 και 4 αντίστοιχα για την ομάδα θεραπείας με RT και 505, 7 και 3 για την RCT ομάδα. Βλέπουμε λοιπόν ότι οι αλγόριθμοι ICM και EM-ICT χρειάστηκαν λιγότερες επαναλήψεις από ότι ο αυτο-συνεπής αλγόριθμος. Όμως ως προς τον υπολογιστικό χρόνο (computing time for CPU) για την ομάδα RT αποδεικνύεται πιο γρήγορος ο αυτο-συνεπής αλγόριθμος από τον ICM γιατί απαιτεί πολύ λιγότερη προσπάθεια μέσα στην επανάληψη.

Χρησιμοποιώντας το μέσο κάθε διαστήματος, κοινή τακτική που χρησιμοποιούν οι αναλυτές εξαιτίας της έλλειψης καλής γνώσης των στατιστικών μεθοδολογιών και του ανάλογου διαθέσιμου λογισμικού, και μετά εφαρμόζοντας τη μέθοδο Kaplan-Meier λαμβάνουμε τις εκτιμήτριες καμπύλες επιβίωσης που παρουσιάζονται στο Διάγραμμα 2. Περιλαμβάνει και τις καμπύλες του πρώτου διαγράμματος.

Να σημειώσουμε ότι οι γραφικές παραστάσεις (Διαγράμματα 1 και 2) των εκτιμημένων συναρτήσεων επιβίωσης δεν παρουσιάζουν μεγάλες διαφορές μεταξύ των θεραπειών κατά τους 18 πρώτους μήνες. Όμως από τον 18ο μήνα και μετά παρατηρείται μία γρήγορη πτώση στην καμπύλη που αντιστοιχεί στους ασθενείς που εκτός από τη ακτινοθεραπεία έκαναν ταυτόχρονη χρήση και της χημειοθεραπείας, εν αντιθέσει με τους ασθενείς που ακολουθούν το πρόγραμμα μόνο της ακτινοθεραπείας. Να επισημάνουμε ότι σε ποσοστό μόλις 11,06 % των ασθενών που ανήκαν στην ομάδα που ακολουθούσαν και ακτινοθεραπεία και χημειοθεραπεία βρέθηκαν χωρίς καμία ένδειξη συστολής του μαστού σε χρόνο που αντιστοιχεί σε 40 μήνες, όπως αυτό φαίνεται από το Διάγραμμα 1. Σε αντίθεση με την ομάδα που ακολουθούσε μόνο τη μέθοδο της ακτινοθεραπείας με αντίστοιχο ποσοστό 47,37 % (4πλάσιο) στον ίδιο χρόνο. Βλέπουμε, λοιπόν, ότι οι ασθενείς που ακολούθησαν μόνο τη ακτινοθεραπεία είχαν για μεγαλύτερο χρονικό διάστημα επιδείνωση της ασθένειάς τους (Διάγραμμα 1)

Διάγραμμα 2: Εκτιμώμενες συναρτήσεις επιβίωσης που χρησιμοποιούν μέσους και διαστήματα.



Συγκρίνοντας τις καμπύλες, στο Διάγραμμα 2 όπου χρησιμοποιούν αμφότερες οι εκτιμήτριες τα μέσα και τα διαστήματα, σε πολλούς χρόνους παρατηρούμε να είναι ίδιες. Παρόλα αυτά έχουν μία τάση να είναι είτε κάτω είτε να έχουν υπερεκτιμηθεί από άλλους χρόνους. Εάν υποθέσουμε ότι το γεγονός συνέβη στο τέλος ή στην αρχή του κάθε διαστήματος αντί για το μέσο θα γίνει περισσότερο αντιληπτό το γεγονός της υπερεκτίμησης ή όχι που απεικονίζεται στο διάγραμμά μας. Ένας ακόμη παράγοντας που συμβάλει στην εμφάνιση τέτοιων διαφορών/αποκλίσεων είναι το εύρος κάθε διαστήματος. Επιτείνονται όσο το εύρος κάθε διαστήματος μεγαλώνει.

Από αυτά τα αποτελέσματα και σύμφωνα με κάποιους συγγραφείς όπως ο Lindsey (1998), καθώς και την ανάλυση των αποκομμένων δεδομένων σε διάστημα υποθέτουμε ότι το γεγονός συνέβη στο τέλος (ή στην αρχή ή στο μέσον) του κάθε διαστήματος και στη συνέχεια εφαρμόζουμε μεθόδους για συγκεκριμένους χρόνους γεγονότων τα οποία οδηγούν σε αναξιόπιστα συμπεράσματα.

Κεφάλαιο 4

Σύγκριση των συναρτήσεων επιβίωσης

4.1 Εισαγωγή

Η σύγκριση των θεραπειών είναι ένας από τους πρωταρχικούς στόχους των περισσότερων ιατρικών μελετών όπως είναι οι κλινικές δοκιμές. Σε αυτές τις περιπτώσεις η έλλειψη ύπαρξης σοβαρών στοιχείων που θα μπορούσαν να υποστηρίξουν ένα συγκεκριμένο παραμετρικό μοντέλο, μας οδηγεί στην εφαρμογή μη παραμετρικών μεθόδων. Για χρόνους αποτυχίας δεξιά αποκομμένων δεδομένων οι περισσότερες μη παραμετρικές μέθοδοι μπορούν να κατηγοριοποιηθούν σε δύο τύπους:

1) σταθμισμένοι έλεγχοι log-rank (weighted log-rank tests)

2) σταθμισμένη Kaplan-Meier

Η χρήση του ελέγχου log-rank είναι ίσως η πιο συνηθισμένη από τις μη παραμετρικές διαδικασίες. Περισσότερες πληροφορίες σχετικά με τους δύο στατιστικούς τύπους μπορούμε να βρούμε στους Fleming και Harrington (1991) και στους Kalbfleisch και Prentice (2002) μεταξύ άλλων. Αυτό το κεφάλαιο ασχολείται με αντίστοιχες μεθόδους που είναι κατάλληλες για χρόνους αποτυχίας αποκομμένων δεδομένων σε διάστημα. Στην παράγραφο 4.2 θα συζητήσουμε για μη παραμετρικές συγκρίσεις θεραπειών όταν τα τρέχοντα δεδομένα π.χ. Περίπτωση Ι: αποκομμένα δεδομένα σε διάστημα είναι διαθέσιμα και δύο τύποι μεθόδων μπορούν να εφαρμοσθούν. Η πρώτη προσέγγιση προϋποθέτει ότι οι παρατηρηθέντες χρόνοι κατά τη διάρκεια της θεραπείας των ομάδων ακολουθούν την ίδια κατανομή και η δεύτερη προσέγγιση επιτρέπει την ύπαρξη διαφορετικών κατανομών για αυτούς τους χρόνους.

4.2 Περίπτωση Ι: τρέχοντα δεδομένα

Θα χρησιμοποιήσουμε την τυχαία μεταβλητή T για να δηλώσουμε το χρόνο αποτυχίας και S την αντίστοιχη συνάρτηση επιβίωσης. Έστω C η μεταβλητή που αναφέρεται στον παρατηρηθέντα χρόνο που θεωρούμε ότι είναι ανεξάρτητος του T. Επίσης υποθέτουμε ότι τα δεδομένα έχουν παρατηρηθεί ανεξάρτητα από n αντικείμενα (μελέτης) και για κάθε αντικείμενο μπορούμε να δούμε μόνο το C και $\delta = I(T \le C)$.

Για τη σύγκριση των θεραπειών σκεφτόμαστε δύο ξεχωριστές περιπτώσεις γιατί οι μέθοδοι για αυτές είναι λίγο διαφορετικές. Θεωρούμε ότι το C ακολουθεί την ίδια κατανομή για όλα τα αντικείμενα της μελέτης και η άλλη επιτρέπει οι κατανομές του C να είναι διαφορετικές για τα αντικείμενα διαφορετικών θεραπειών.

Εδώ θα εστιάσουμε την προσοχή μας σε ένα πρόβλημα σύγκρισης 2 δειγμάτων και θα συζητήσουμε 2 μεθόδους. Η πρώτη είναι η διαδικασία Wilcoxon-type και η δεύτερη είναι μία γενίκευση της σταθμισμένης διαδικασίας Kaplan-Meier η ιδέα της οποίας αναπτύσσεται από τους Andersen και Ronn (1995). Με S_1, S_2 δηλώνουμε τις συναρτήσεις επιβίωσης που αντιστοιχούν στις 2 ομάδες θεραπείας και ο στόχος μας είναι να ελέγξουμε την υπόθεση $H_0 : S_1(t) = S_2(t)$ για όλα τα t ή S_1=S_2.

4.2.1 Η διαδικασία Wilcoxon-type

Υποθέτουμε ότι τα παρατηρηθέντα δεδομένα αποτελούνται από $\{(C_i, \delta_i, Z_i);\ i = 1, \ldots, n\}$ όπου Z_i=0 ή 1 να είναι ο δείκτης της ομάδας θεραπείας για το αντικείμενο της μελέτης i. Για να ελέγξουμε τη μηδενική υπόθεση, σημειώνουμε ότι κάτω από την H_0 και την υπόθεση ότι τα C_i ακολουθούν την ίδια κατανομή, τα δ_i είναι τυχαίες μεταβλητές όμοιων και ανεξάρτητων κατανομών (i.i.d). Αυτό προτείνει να χρησιμοποιήσουμε το ακόλουθο στατιστικό Wilcoxon:

$$\sum_i \sum_j (Z_i - Z_j)(\delta_i - \delta_j)$$

για να ελέγξουμε την H_0. Είναι προφανές ότι το παραπάνω στατιστικό είναι ίσο με

$$U_{cw} = \sum_{i=1}^{n} (Z_i - \bar{Z})\delta_i, \text{ όπου } \bar{Z} = n^{-1}\sum_{i=1}^{n} Z_i$$

Είναι εύκολο να δείξουμε ότι υπό την H_0 και για μεγαλύτερο *n,* η κατανομή του $n^{-1/2}U_{cw}$ μπορεί να προσεγγισθεί από μία κανονική κατανομή με μέσο το 0 και διασπορά

$$\hat{\sigma}_{cw}^2 = n^{-1}\sum_{i=1}^{n} (Z_i - \bar{Z})^2 \delta_i^2$$

Έτσι ένας μεγάλος δειγματικός έλεγχος της μηδενικής υπόθεσης χρησιμοποιώντας το στατιστικό $U_{cw} / \hat{\sigma}_{cw}$ μπορεί να εφαρμοσθεί με ποσοστιαία σημεία της τυποποιημένης Κανονικής κατανομής.

4.2.2 Γενίκευση της σταθμισμένης Kaplan-Meier διαδικασίας

Έστω *G* η συνάρτηση κατανομής του *C* και $\hat{G}_n$ η εμπειρική κατανομή του *C*. Υποθέτουμε ότι και το S_0 και το *G* είναι συνεχείς συναρτήσεις. Έστω $\hat{S}_1, \hat{S}_2$ να είναι οι εκτιμήτριες μεγίστης πιθανοφάνειας των S_1, S_2 αντίστοιχα βασισμένες στα αντικείμενα της μελέτης μέσα σε κάθε ομάδα θεραπείας ξεχωριστά. Έστω n_1, n_2 να δηλώνουν τον αριθμό των ατόμων της μελέτης που λαμβάνουν μία από τις 2 θεραπείες, αντίστοιχα, και θεωρούμε ότι $n_1 / n \to p(0 < p < 1),\ n \to \infty$ όπου $n_1 + n_2 = n$.

Για να ελέγξουμε την H_0, παρακινημένοι από τους Kolmogorov-Smirnov και τους σταθμισμένους ελέγχους Kaplan-Meier (Pepe and Fleming, 1989), μπορούμε να κατασκευάσουμε ένα απλό στατιστικό έλεγχο:

$$U_{cs} = \int_0^{\tau} [\hat{S}_2(t) - \hat{S}_1(t)] d\hat{G}_n(t) \qquad (4.2.1.2.1)$$

όπου τ είναι μία σταθερά τέτοια ώστε $S_0(\tau) > 0$ και συνήθως επιλέγεται ως ο μεγαλύτερος χρόνος παρατήρησης.

Η ασυμπτωτική κατανομή του U_{cs} ακολουθεί την πυκνότητα της $\hat{G}_n$ πιο συγκεκριμένα εάν το n είναι μεγάλο και η μηδενική υπόθεση αληθής η κατανομή του U_{cs} πολλαπλασιασμένη με $\sqrt{n}$ μπορεί να προσεγγισθεί από την κανονική κατανομή με μέσο 0 και διασπορά

$$\hat{\sigma}^2_{cw} = \frac{n^2}{n_1 n_2} \int_0^\tau \hat{S}_0(t)[1-\hat{S}_0(t)]d\hat{G}_n(t).$$

Ένας μεγάλος, λοιπόν, δειγματικός έλεγχος της H_0 μπορεί να εφαρμοσθεί χρησιμοποιώντας $U^*_{cs} = \sqrt{n}U_{cs} / \hat{\sigma}_{cs}$ με κανονικές τιμές.

Ακολουθώντας την παραπάνω ιδέα μπορούν να αναπτυχθούν παρόμοιοι στατιστικοί έλεγχοι. Για παράδειγμα, μπορούμε να χρησιμοποιήσουμε το στατιστικό των U_{cs} τετραγώνων των $\hat{S}_1, \hat{S}_2$. Μία ακόμα εναλλακτική είναι να λάβουμε υπόψη μας τη διαφορά ανάμεσα στους εμπειρικούς μέσους των δύο πληθυσμών. Οι μέθοδοι που βασίζονται σε αυτές τις στατιστικές, οι οποίες αναμένεται να εκτελεσθούν όπως όταν υπολογίζουμε το U_{cs} συζητούνται λεπτομερώς από τους Andersen και Ronn (1995) και Tang et al. (1995).

4.3 Περίπτωση II: Rank-based διαδικασίες σύγκρισης

Θεωρούμε τη γενική ή την Περίπτωση II των αποκομμένων δεδομένων χρόνων αποτυχίας σε διάστημα και υποθέτουμε ότι τα παρατηρηθέντα δεδομένα είναι

$$\{(L_i, R_i],\ \mathbf{Z}_i;\ i = 1,..., n\}$$

για n ανεξάρτητα άτομα καθένα από τα οποία δέχεται μία από τις p+1 διαφορετικές θεραπείες. Αφού δηλώσαμε το διάστημα μες στο οποίο το γεγονός της επιβίωσης που μας ενδιαφέρει έχει συμβεί στο άτομο i θεωρούμε και το p-διάστατο διάνυσμα των δεικτών της θεραπείας, το $\mathbf{Z}_i$. Στόχος μας είναι ο έλεγχος της υπόθεσης H_0: οι p+1 συναρτήσεις επιβίωσης που αντιστοιχούν στις θεραπείες να είναι πανομοιότυπες. Πολλές προσεγγίσεις μπορούν να αναπτύξουν διαδικασίες ελέγχων κάνοντας χρήση των βαθμολογιών (rankings) των μη-παρατηρηθέντων πραγματικών χρόνων αποτυχίας, T_i και είναι ίδιες με τις μεθόδους που έχουν αναπτυχθεί για αποκομμένους ή δεξιά αποκομμένους χρόνους

αποτυχίας. Σε αυτή την παράγραφο θα εστιάσουμε την προσοχή μας στην προσέγγιση που έδωσαν οι Zhao και Sun (2004) η οποία γενικεύει άμεσα τον έλεγχο log-rank που είναι ο πιο συνήθης έλεγχος για δεξιά αποκομμένα δεδομένα λόγω της απλότητας και της ευκολίας της ερμηνείας και της εκτέλεσής του.

4.3.1 Γενικευμένος έλεγχος log-rank

Δηλώνουμε με $\hat{S}_0$ την εκτιμήτρια μεγίστης πιθανοφάνειας της κοινής συνάρτησης επιβίωσης $S_0(t) = \Pr(T_i > t)$ κάτω από τη μηδενική υπόθεση και ορίζουμε τους $s_1 < ... < s_m$ διατεταγμένους-διακεκριμένους χρόνους του διαστήματος $(L_i, R_i]$, $i = 1, ..., n$ στο οποίο η $\hat{S}_0$ παρουσιάζει άλματα. Για ευκολία θεωρούμε ότι υπάρχει ένα χρονικό σημείο $s_{m+1} > s_m$ στο οποίο η $\hat{S}_0$ έχει την υπολειπόμενη πυκνότητα με $\hat{S}_0(t)$ να είναι ίση με το μηδέν για $t \geq s_{m+1}$.

Για κάθε ζεύγος (i,j) ορίζουμε:

$$a_{ij} = I(s_j \in (L_i, R_i])$$

με το δείκτη του γεγονότος να ανήκει στο διάστημά μας, όπως φαίνεται από την αμέσως προηγούμενη σχέση και με i=1,..., n, j=1,..., m+1. Για να αναπτύξουμε μία ελεγχοσυνάρτηση του τύπου log-rank για την H_0 θυμόμαστε ότι ο στατιστικός έλεγχος log-rank είναι το άθροισμα των παρατηρηθέντων μείον τον αναμενόμενο αριθμό των θανάτων-γεγονότων. Έτσι χρειάζεται να καθορίσουμε τον αριθμό των αποτυχιών και τον αριθμό των ατόμων που βρίσκονται σε κίνδυνο για κάθε καταγεγραμμένο χρόνο αποτυχίας.

Για αυτό το λόγο, ορίζουμε $\delta_i = 0$ εάν η παρατήρηση του χρόνου αποτυχίας T_i για το i-οστό άτομο της μελέτης είναι δεξιά αποκομμένη και 1 διαφορετικά, με i=1,..., n. Έχουμε δηλαδή $\delta_i = I(R_i \leq s_m)$. Επιπλέον ορίζουμε $\rho_{ij} = I(\delta_i = 0, L_i \geq s_j)$ που παίρνει την τιμή 1 εάν T_i είναι δεξιά αποκομμένος και το i-οστό άτομο βρίσκεται ακόμα σε κίνδυνο τη χρονική στιγμή s_j^-, i=1,..., n, j=1,..., m. Τότε εάν η μηδενική υπόθεση είναι αληθής και η $S_0(t)$ είναι γνωστή, δοθέντος της $\hat{S}_0(t)$ μπορεί να εκτιμήσει τους

συνολικά παρατηρηθέντες χρόνους αποτυχίας και τον αριθμό των ατόμων που βρίσκονται σε κίνδυνο τη χρονική στιγμή s_j^- σύμφωνα με τον ακόλουθο τύπο:

$$d_j = \sum_{i=1}^{n} \delta_i \frac{a_{ij}[\hat{S}_0(s_j-) - \hat{S}_0(s_j)]}{\sum_{u=1}^{m+1} a_{iu}[\hat{S}_0(s_u-) - \hat{S}_0(s_u)]}$$

και

$$n_j = \sum_{r=j}^{m+1} \sum_{i=1}^{n} \delta_i \frac{a_{ir}[\hat{S}_0(s_r-) - \hat{S}_0(s_r)]}{\sum_{u=1}^{m+1} a_{iu}[\hat{S}_0(s_u-) - \hat{S}_0(s_u)]} + \sum_{i=1}^{n} \rho_{ij}$$

αντίστοιχα, j=1,..., m.

Ομοίως εκτιμάει τους παρατηρηθέντες χρόνους αποτυχίας και τον αριθμό των ατόμων που βρίσκονται σε κίνδυνο τη χρονική στιγμή s_j με j=1,..., m για τις ομάδες θεραπείας l, με l=1,..., p+1 και είναι

$$d_{jl} = \sum_{i}^{l} \frac{a_{ij}[\hat{S}_0(s_j-) - \hat{S}_0(s_j)]}{\sum_{u=1}^{m+1} a_{iu}[\hat{S}_0(s_u-) - \hat{S}_0(s_u)]}$$

και

$$n_{jl} = \sum_{r=j}^{m+1} \sum_{i}^{l} \delta_i \frac{a_{ir}[\hat{S}_0(s_r-) - \hat{S}_0(s_r)]}{\sum_{u=1}^{m+1} a_{iu}[\hat{S}_0(s_u-) - \hat{S}_0(s_u)]} + \sum_{i}^{l} \rho_{ij}$$

αντίστοιχα όπου Σ_i^l δηλώνει το άθροισμα πάνω σε όλα τα άτομα του πληθυσμού. Στην περίπτωση των δεξιά αποκομμένων δεδομένων οι παραπάνω εκτιμήσεις ανάγονται στους παρατηρούμενους χρόνους αποτυχίας και τον αριθμό των ατόμων που βρίσκονται σε κίνδυνο, τα οποία έχουν χρησιμοποιηθεί για την κατασκευή του κλασικού στατιστικού ελέγχου log-rank. Για τον έλεγχο της ισότητας των p+1 συναρτήσεων επιβίωσης οι Zhao και Sun (2004) καταλήγουν στην ελεγχοσυνάρτηση $U_r^* = \mathbf{U}_r' \hat{\mathbf{V}}_r^{-1} \mathbf{U}_r \sim X_p^2$ υπό την H_0, με $\mathbf{U}_r = (U_{r,1},..., U_{r,p+1})'$ και

$$U_{r,1} = \sum_{j=1}^{m} (d_{j,l} - \frac{n_{jl} d_j}{n_j})$$

όπου $\hat{\mathbf{V}}_r^{-1}$ ο γενικευμένος αντίστροφος πίνακας διασποράς-συνδιασποράς των $\mathbf{U}_r$, στην r-οστή χρονική στιγμή.

Κεφάλαιο 5

Ανάλυση παλινδρόμησης για τρέχοντα δεδομένα

5.1 Εισαγωγή

Σε μερικές περιπτώσεις όπως τα πειράματα καρκινογένεσης σε κρυφούς όγκους τα τρέχοντα δεδομένα (current status) είναι η μόνη διαθέσιμη πληροφορία για τις μεταβλητές επιβίωσης που μας ενδιαφέρουν όπως ο χρόνος εμφάνισης του όγκου (Dinse και Lagakos, 1983), που όμως δε μπορούν να μετρηθούν άμεσα. Σε άλλες περιπτώσεις όπως αυτές που προέρχονται από μελέτες διασταύρωσης κάποιου σημαντικού γεγονότος τα τρέχοντα δεδομένα παρέχουν πιο κατανοητή και αξιόπιστη πληροφορία για το χρόνο του γεγονότος από ότι τα πλήρη δεδομένα. Ένα παράδειγμα των παραπάνω περιπτώσεων είναι οι επιδημιολογικές μελέτες όπου το γεγονός που μας ενδιαφέρει είναι η εμφάνιση μια χρόνιας πάθησης (Keiding, 1991; Keiding et al., 1996; Shiboski and Jewell, 1992). Ένα ακόμα παράδειγμα δίνεται από τις δημογραφικές μελέτες όπου το γεγονός μπορεί να είναι, π.χ. πρώτη εγκυμοσύνη ή πρώτος γάμος (Diamond and McDonald, 1991; Diamond et al., 1986).

Σε αυτό το κεφάλαιο θα μελετήσουμε την ανάλυση παλινδρόμησης της Περίπτωσης I (Case I) ή των χρόνων αποτυχίας των τρεχουσών δεδομένων υπό τα ημι-παραμετρικά μοντέλα και τα συμπεράσματα που προκύπτουν για τους συντελεστές παλινδρόμησης. Για αυτήν, όπως και σε πολλές άλλες περιπτώσεις, η πιο συχνά εφαρμοσμένη προσέγγιση είναι αυτή της ημι-παραμετρικής εκτιμήτριας μεγίστης πιθανοφάνειας. Η συγκεκριμένη προσέγγιση είναι άμεση, αλλά όχι εύκολη γιατί η πιθανοφάνεια είναι μία συνάρτηση πεπερασμένης διάστασης παραμέτρων παλινδρόμησης και μη πεπερασμένης διάστασης παραμέτρων θορύβου, της σωρευτικής συνάρτησης διακινδύνευσης ή της συνάρτησης επιβίωσης. Ως επακόλουθο, θα πρέπει να εκτιμηθούν ταυτόχρονα και οι παράμετροι παλινδρόμησης και οι παράγοντες θορύβου. Αυτό διαφέρει από την ανάλυση παλινδρόμησης των χρόνων αποτυχίας δεξιά αποκομμένων δεδομένων χρησιμοποιώντας το μοντέλο PH (1.4.2.1) όπου η προσέγγιση μερικής πιθανοφάνειας μπορεί να εφαρμοσθεί. Στην τελευταία προσέγγιση, για

συμπεράσματα ως προς τις παραμέτρους παλινδρόμησης, η μερική πιθανοφάνεια μπορεί να παραχθεί χωρίς να συμπεριλαμβάνονται σε αυτή οι παράμετροι θορύβου. Όμως για τα τρέχοντα δεδομένα η προσέγγιση μερικής πιθανοφάνειας δεν είναι διαθέσιμη και γι' αυτό το λόγο βασιζόμαστε στην πλήρη πιθανοφάνεια.

Για να αποφευχθεί η πιθανοφάνεια που εμπλέκει μη πεπερασμένης διάστασης παραμέτρους θορύβου θα γίνει χρήση της μεθόδου μεγίστης πιθανοφάνειας υποχώρων. Η βασική ιδέα της μεθόδου είναι η προσέγγιση των μη πεπερασμένων διάστασης παραμέτρων θορύβου από μία ακολουθία πεπερασμένης διάστασης παραμέτρων η οποία είναι η αρχική παράμετρος χώρου και η οποία προσεγγίζεται από μία αύξουσα ακολουθία πεπερασμένης διάστασης υποχώρων. Για το πρόβλημα που θα περιγραφεί παρακάτω υποθέτουμε ότι ένα ημι-παραμετρικό μοντέλο παλινδρόμησης έχει ορισθεί με παράμετρο παλινδρόμησης **β** και σωρευτική συνάρτηση διακινδύνευσης $\Lambda_0(t)$. Τότε η αρχική παράμετρος χώρου που συνδέεται με την $\Lambda_0(t)$ μπορεί να είναι μία συλλογή από μη φθίνουσες συναρτήσεις και οι υπόχωροι μπορεί να είναι π.χ. ομάδες από τμήματα συνεχών και μη φθινουσών γραμμικών συναρτήσεων. Για κάθε πεπερασμένο δείγμα η εκτίμηση του **β** και του $\Lambda_0(t)$ μπορεί να υπολογιστεί μεγιστοποιώντας τη συνάρτηση πιθανοφάνειας ως προς τις παραμέτρους του χώρου για το **β** και τον υπόχωρο. Με άλλα λόγια, χρειάζεται να δουλέψουμε με μία πεπερασμένης διάστασης παράμετρο χώρου με τη μέθοδο των υποχώρων. Ένα πλεονέκτημα αυτής της μεθόδου η εκτιμήτρια της σωρευτικής συνάρτησης $\Lambda_0(t)$ έχει γρηγορότερο ρυθμό σύγκλισης από ότι η εκτιμήτρια που προέρχεται από τη μεγιστοποίηση της πλήρης πιθανοφάνειας ως προς την αρχική παράμετρο χώρου (Huang and Rossini, 1997).

Σημείωση: Σε αυτό το κεφάλαιο θα θεωρούμε ότι υπάρχει μία μελέτη επιβίωσης που αποτελείται από *n* ανεξάρτητα άτομα. Για το *i*-οστό άτομο υποθέτουμε ότι υπάρχουν δύο μεταβλητές: μία θα είναι ο χρόνος επιβίωσης που μας ενδιαφέρει και θα συμβολίζεται με T_i και η άλλη C_i και θα δηλώνει το χρόνο παρατήρησης του ατόμου, με *i=1,..., n.* Επίσης για το *i*-οστό άτομο υποθέτουμε ότι υπάρχει ένα διάνυσμα συμμεταβλητών $\mathbf{Z}_i$. Η κατανομή των T_i καθορίζεται από την παράμετρο παλινδρόμησης **β** και τη σωρευτική συνάρτηση διακινδύνευσης $\Lambda_0(t)$ ή τη συνάρτηση επιβίωσης $S_0(t) = \exp(-\Lambda_0(t))$.

Επιπλέον, για συμπεράσματα ως προς το **β,** $\Lambda_0(t)$ ή $S_0(t)$ μόνο τα τρέχοντα δεδομένα είναι διαθέσιμα και δίνονται από τον τύπο:

$$\{(C_i,\ \delta_i = I(T_i \le C_i),\ \mathbf{Z}_i);\ i = 1,...,\ n\}$$

Με αυτό τον τρόπο, κάθε άτομο παρατηρείται μόνο μία φορά την C_i και σε αυτή τη χρονική στιγμή ξέρουμε μόνο εάν το γεγονός της επιβίωσης που μας ενδιαφέρει έχει συμβεί πριν ή τη στιγμή C_i. Γενικότερα, υποθέτουμε ότι τα $\mathbf{Z}_i$, T_i και C_i είναι ανεξάρτητα.

5.2 Ανάλυση με το μοντέλο αναλογικής διακινδύνευσης

Σε αυτή την ενότητα θα αναφερθούμε στην ανάλυση παλινδρόμησης των τρεχουσών δεδομένων χρησιμοποιώντας το μοντέλο PH (1.4.2.1) το πιο σύνηθες μοντέλο παλινδρόμησης για την ανάλυση των χρόνων αποτυχίας. Ως προς τη σωρευτική συνάρτηση διακινδύνευσης, το μοντέλο δηλώνεται

$$\Lambda(t; \mathbf{Z}_i) = \Lambda_0(t)\exp(\mathbf{Z}'\boldsymbol{\beta})$$

για δεδομένο $\mathbf{Z}_i$ και η συνάρτηση πιθανοφάνειας είναι ανάλογη του

$$L(\boldsymbol{\beta}, \Lambda_0) = \prod_{i=1}^{n} \exp[-(1-\delta_i)e^{\mathbf{Z}_i'\boldsymbol{\beta}}\Lambda_0(C_i)][1-\exp(-e^{\mathbf{Z}_i'\boldsymbol{\beta}}\Lambda_0(C_i))]^{\delta_i} \qquad (5.2.1)$$

Ως προς το **β** και την S_0 η παραπάνω συνάρτηση πιθανοφάνειας έχει τώρα τη μορφή

$$L(\boldsymbol{\beta}, S_0) = \prod_{i=1}^{n} [S_0(C_i)]^{(1-\delta_i)\exp^{\mathbf{Z}_i'\boldsymbol{\beta}}} \{1-[S_0(C_i)]^{\exp^{\mathbf{Z}_i'\boldsymbol{\beta}}}\}^{\delta_i} \qquad (5.2.2)$$

5.3 Εκτιμήτρια μεγίστης πιθανοφάνειας

Για την εκτίμηση του **β** και της S_0, η προσέγγιση μεγίστης πιθανοφάνειας μεγιστοποιεί τη συνάρτηση πιθανοφάνειας $L(\boldsymbol{\beta}, S_0)$ που δόθηκε από τον τύπο (5.2.2). Για αυτή και μία ομάδα τρεχουσών δεδομένων, όπως στην περίπτωση του ενός δείγματος που συζητήθηκε στις ενότητες 3.2 έως

3.4, μόνο οι τιμές της $S_0(t)$ στους παρατηρηθέντες χρόνους C_i επηρεάζουν τη συνάρτηση πιθανοφάνειας. Χωρίς βλάβη της γενικότητας, μπορούμε να εστιάσουμε στη μεγιστοποίηση της $L(\boldsymbol{\beta}, S_0)$.

Έστω $0 < s_1 < ... < s_m$ οι διατεταγμένοι διακριτοί χρόνοι του $\{C_i\}_{i=1}^n$ και Ω_s η συλλογή όλων των συναρτήσεων επιβίωσης $S_0(t)$ οι οποίες έχουν τη μορφή:

$$S_0(t) = \prod_{j:s_j \le t} e^{-\exp(a_j)} \tag{5.3.1}$$

όπου $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_m)'$ είναι άγνωστες παράμετροι. Για να καθοριστούν οι εκτιμητές μεγίστης πιθανοφάνειας του $\boldsymbol{\beta}$ και της S_0, χρειάζεται να μεγιστοποιήσουμε την $L(\boldsymbol{\beta}, S_0)$ και την S_0 στο Ω_s. Σε αυτή την περίπτωση, η λογαριθμική συνάρτηση πιθανοφάνειας μπορεί να γραφτεί ως

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \{\delta_i \log[1 - \prod_{j_i} e^{-\exp(a_j + \mathbf{Z}_i'\boldsymbol{\beta})}] - \sum_{j_i} (1 - \delta_i) e^{(a_j + \mathbf{Z}_i'\boldsymbol{\beta})}\}$$

με Π_{j_i} και Σ_{j_i} το γινόμενο και το άθροισμα ως προς $\{j;\ s_j \le C_i\}$, αντίστοιχα.

Ορίζουμε D_j να είναι το σύνολο των δεικτών των ατόμων για τα οποία $s_j = C_i$ και $\delta_i = 1$ και το R_j σύνολο των δεικτών των ατόμων για τα οποία $s_j = C_i, j = 1, ..., m$. Έστω $a_j = \sum_{k=1}^{j} \exp(a_k),\ j = 1, ..., m$. Τότε η λογαριθμική συνάρτηση πιθανοφάνειας $l(\boldsymbol{\beta}, \boldsymbol{\alpha})$ μπορεί να ξαναγραφτεί ως

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{j=1}^{m} \{\sum_{i \in D_j} \log[\frac{1 - e^{-a_j \exp(\mathbf{Z}_i'\boldsymbol{\beta})}}{e^{-a_j \exp(\mathbf{Z}_i'\boldsymbol{\beta})}}] - a_j \sum_{i \in R_j} e^{\mathbf{Z}_i'\boldsymbol{\beta}}\} \tag{5.3.2}$$

Για να μεγιστοποιήσουμε την $l(\boldsymbol{\beta}, \boldsymbol{\alpha})$ μια λογική προσέγγιση είναι να εφαρμόσουμε τον αλγόριθμο Newton-Raphson και για αυτό θα χρειαστούμε την πρώτη και τη δεύτερη παράγωγο της $l(\boldsymbol{\beta}, \boldsymbol{\alpha})$. Έχουμε λοιπόν

$$\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} = \sum_{j=1}^{m} a_j \{ \sum_{i \in D_j} \mathbf{Z}_i e^{\mathbf{Z}_i' \boldsymbol{\beta}} [q(a_j, \mathbf{Z}_i) + 1] + \sum_{i \in R_j} \mathbf{Z}_i e^{\mathbf{Z}_i' \boldsymbol{\beta}}$$

$$\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial a_j} = e^{a_j} \sum_{k=j}^{m} \{ \sum_{i \in D_k} e^{\mathbf{Z}_i' \boldsymbol{\beta}} [q(a_k, \mathbf{Z}_i) + 1] + \sum_{i \in R_k} e^{\mathbf{Z}_i' \boldsymbol{\beta}}$$

$$\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \sum_{j=1}^{m} a_j \{ \sum_{i \in D_j} \mathbf{Z}_i \mathbf{Z}_i' e^{\mathbf{Z}_i' \boldsymbol{\beta}} [q(a_j, \mathbf{Z}_i) + 1 - a_j e^{\mathbf{Z}_i' \boldsymbol{\beta}} [q(a_j, \mathbf{Z}_i) + 1]] - \sum_{i \in R_j} \mathbf{Z}_i \mathbf{Z}_i' e^{\mathbf{Z}_i' \boldsymbol{\beta}}$$

$$\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial a_j \partial \boldsymbol{\beta}'} = e^{a_j} \sum_{k=j}^{m} \{ \sum_{i \in D_k} \mathbf{Z}_i e^{\mathbf{Z}_i' \boldsymbol{\beta}} [q(a_k, \mathbf{Z}_i) + 1 - a_j e^{\mathbf{Z}_i' \boldsymbol{\beta}} q(a_k, \mathbf{Z}_i)[q(a_k, \mathbf{Z}_i) + 1]] - \sum_{i \in R_k} \mathbf{Z}_i e^{\mathbf{Z}_i' \boldsymbol{\beta}}$$

$$\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial a_j^2} = \frac{\partial l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial a_j} - e^{2a_j} \sum_{k=j}^{m} \{ \sum_{i \in D_k} e^{2\mathbf{Z}_i' \boldsymbol{\beta}} [q(a_k, \mathbf{Z}_i) + 1] q(a_k, \mathbf{Z}_i) \}$$

και

$$\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial a_j \partial a_k} = -e^{a_j + a_k} \sum_{l=k}^{m} \{ \sum_{i \in D_l} e^{2\mathbf{Z}_i' \boldsymbol{\beta}} q(a_l, \mathbf{Z}_i)[q(a_l, \mathbf{Z}_i) + 1], \ k > j$$

όπου

$$q(a_j, \mathbf{Z}_i) = \frac{e^{-a_j \exp(\mathbf{Z}_i' \boldsymbol{\beta})}}{1 - e^{-a_j \exp(\mathbf{Z}_i' \boldsymbol{\beta})}} \ j = 1, ..., \ m, \ i = 1, ..., \ n.$$

Για να εφαρμόσουμε τον αλγόριθμο Newton-Raphson πρέπει να επιλέξουμε αρχικές εκτιμήσεις για το **β** και το **α** όπως επίσης ένα κριτήριο σύγκλισης και να υπολογίσουμε τον αντίστροφο ενός πίνακα (p+m)x(p+m). Ένα λογικό σύνολο αρχικών εκτιμήσεων είναι αυτό που συζητήθηκε στην παράγραφο 2.4.1 από την προσέγγιση συμβατής απόδοσης τιμών μέσω ενός σημείου και ένα κριτήριο σύγκλισης θα μπορούσε να είναι αυτό της παραγράφου 3.4.3. Για τον αντίστροφο πίνακα, μια απλοποίηση μπορεί να επιτευχθεί χρησιμοποιώντας το γεγονός ότι για ένα συμμετρικό 2x2 πίνακα

$$\mathbf{A} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \text{ με } \mathbf{A}^{-1} = \begin{pmatrix} A_{11|2}^{-1} & -A_{11|2}^{-1} A_{12} A_{22}^{-1} \\ -A_{22}^{-1} A_{21} A_{11|2}^{-1} A_{22}^{-1} + & A_{22}^{-1} A_{21} A_{11|2}^{-1} A_{12} A_{22}^{-1} \end{pmatrix}$$

όλοι οι αντίστροφοι υπάρχουν (Rao, 1973, pp. 33), με $A_{11|2} = A_{11} - A_{12}A_{22}^{-1}A_{21}$. Έστω $\hat{\boldsymbol{\beta}}_n$ και $\hat{\boldsymbol{\alpha}}_n$ οι εκτιμήσεις των **β** και **α** που ορίσθηκαν παραπάνω και $\hat{S}_n(t)$ η εκτίμηση της συνάρτησης επιβίωσης του τύπου (5.3.1) με **α** να αντικαθίσταται από το $\hat{\boldsymbol{\alpha}}$ και $\hat{\Lambda}_n(t) = -\log[\hat{S}_n(t)]$ μία εκτίμηση της σωρευτικής συνάρτησης διακινδύνευσης. Για να βγάλουμε συμπεράσματα για το **β** χρειάζεται να ξέρουμε την ασυμπτωτική κατανομή και μία εκτίμηση του πίνακα διασποράς-συνδιασποράς του $\hat{\boldsymbol{\beta}}_n$. Για την εκτίμηση της συνδιασποράς, μια γενική προσέγγιση είναι να χειριστούμε το $l(\boldsymbol{\beta}, \boldsymbol{\alpha})$ σαν μια παραμετρική συνάρτηση πιθανοφάνειας εξαρτώμενη από τα **β** και **α.** Τότε μπορούμε να χρησιμοποιήσουμε τον πίνακα πληροφορίας (Fisher). Στην περίπτωση που το $\mathbf{Z}_i$ παίρνει τις τιμές 0 ή 1 μπορεί να ακολουθηθεί η εξής εναλλακτική του Huang (1996). Έστω $g_0(c)$ και $g_1(c)$ οι συναρτήσεις πυκνότητας των C_i για τα άτομα με $\mathbf{Z}_i = 0$ ή 1 και $\hat{g}_0(c)$ και $\hat{g}_1(c)$ οι βασικοί εκτιμητές τους, αντίστοιχα. Για δεδομένη συμμεταβλητή **Z**, ορίζουμε:

$$\hat{R}(c; Z) = \frac{\exp[-e^{Z\hat{\beta}_n}\hat{\Lambda}_n(c)]}{1-\exp[-e^{Z\hat{\beta}_n}\hat{\Lambda}_n(c)]}\hat{\Lambda}_n^2(c)e^{2Z\hat{\beta}_n}$$

και

$$\hat{\mu}(c) = \frac{\hat{R}(c; \mathbf{Z}=1)\hat{g}_1(c)n_1}{\hat{R}(c; \mathbf{Z}=1)\hat{g}_1(c)n_1 + \hat{R}(c; \mathbf{Z}=0)\hat{g}_0(c)(n-n_1)}$$

όπου $n_1 = \sum_{i=1}^{n}\mathbf{Z}_i$, ο αριθμός των ατόμων με $\mathbf{Z}_i = 1$. Τότε μια σταθερή εκτίμηση της διασποράς του $\hat{\boldsymbol{\beta}}_n$ δίνεται από $(n\hat{\sigma}_n^2)^{-1}$ με

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\{\hat{R}(C_i; \mathbf{Z}_i)[\mathbf{Z}_i - \hat{\mu}(C_i)]^2\} \qquad (5.3.3)$$

Η εκτιμήτρια μεγίστης πιθανοφάνειας $\hat{\boldsymbol{\beta}}_n$ είναι ασυμπτωτικά επαρκής. Μια απλούστευση προκύπτει εάν $g_0(c) = g_1(c)$, το οποίο υπονοεί ότι τα C_i είναι ανεξάρτητα από τα $\mathbf{Z}_i$. Σε αυτή την περίπτωση μπορούμε να δούμε ότι το $\hat{\mu}(c)$ και έτσι και το $\hat{\sigma}_n$ δεν περιέχουν τις εκτιμήσεις των $g_0(c)$ και $g_1(c)$.

Στην πραγματικότητα, για διάφορα $\mathbf{Z}_i$, εάν C_i είναι ανεξάρτητο του $\mathbf{Z}_i$, ο πίνακας διασποράς-συνδιασποράς του $\hat{\boldsymbol{\beta}}_n$ μπορεί απλά να εκτιμηθεί από $(n\hat{\Sigma}_n)^{-1}$ με

$$\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}\{\hat{R}(C_i;\mathbf{Z}_i)[\mathbf{Z}_i - \frac{\sum_{j=1}^{n}\mathbf{Z}_j\hat{R}(C_i;\mathbf{Z}_j)}{\sum_{j=1}^{n}\hat{R}(C_i;\mathbf{Z}_j)}]^2\} \qquad (5.3.4)$$

(Haung, 1996).

Για να αποφευχθούν κάποια προβλήματα που μπορεί να προκληθούν στον αλγόριθμο Newton-Raphson όταν έχουμε μεγάλο αριθμό διαφορετικών χρονικών παρατηρήσεων μπορούμε να μεγιστοποιήσουμε τη συνάρτηση πιθανοφάνειας της $L(\boldsymbol{\beta},\Lambda_0)$ που δίνεται από τον τύπο (5.2.1). Τότε το αποτέλεσμα τη λογαριθμικής συνάρτησης πιθανοφάνειας είναι της μορφής:

$$l(\boldsymbol{\beta},\Lambda_0) = \sum_{i=1}^{n}\{\delta_i \log[1-\exp(-e^{\mathbf{Z}_i'\boldsymbol{\beta}}\Lambda_0(C_i))] - (1-\delta_i)\exp(\mathbf{Z}_i'\boldsymbol{\beta})\Lambda_0(C_i)\}$$

5.4 Παραδείγματα

Για να εξηγήσουμε καλύτερα την προσέγγιση μεγίστης πιθανοφάνειας θα δούμε δύο παραδείγματα, το πρώτο αναφέρεται σε δεδομένα για όγκο στον πνεύμονα και το δεύτερο αναφέρεται σε ένα σύνολο τρεχουσών δεδομένων που προέρχονται από μια μελέτη ασβεστοποίησης των ενδοφθάλμιων φακών υδρογέλης.

Οι Hoel και Walberg (1972) έδωσαν ένα σύνολο δεδομένων (παράγραφος 3.1) για 144 αρσενικά RFM ποντίκια ενός πειράματος ογκογονικότητας που περιείχε όγκους του πνεύμονα. Τα δεδομένα παρουσιάζονται στον Πίνακα 4 που αποτελείται από το χρόνο θανάτου κάθε ζώου καταγεγραμμένο σε μέρες και ένα δείκτη για την παρουσία του όγκου του πνεύμονα (1) ή την απουσία (0) αυτού κατά το χρόνο του θανάτου. Το πείραμα περιλαμβάνει δύο θεραπείες, σε φυσικό περιβάλλον (CE, 96 ποντίκια) και σε αντιμικροβιακό περιβάλλον (GE, 48 ποντίκια). Οι όγκοι του πνεύμονα σε αρσενικά ποντίκια είναι κυρίως μη θανατηφόροι, που σημαίνει ότι η εμφάνιση του όγκου δεν αλλάζει το ρυθμό θανάτου.

Το γεγονός "εμφάνιση του όγκου" συμβαίνει πριν και μετά τους χρόνους εξέτασης θανάτου.

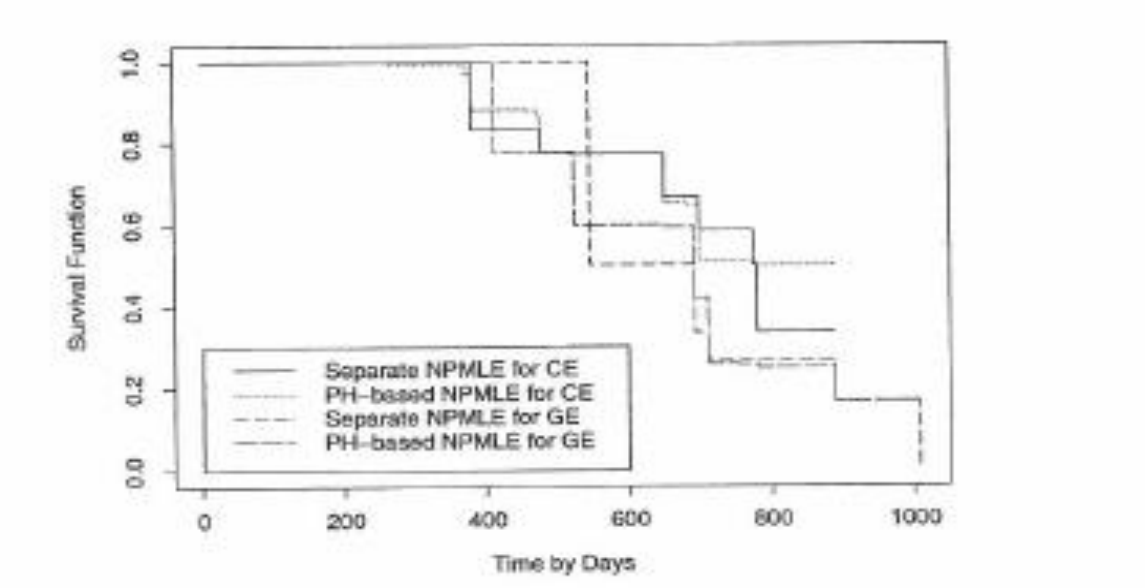
Πίνακας 4: Χρόνοι θανάτου σε μέρες για 144 αρσενικά ποντίκια διαγνωσμένα με όγκους του πνεύμονα.

Group	Tumor status	Death times
CE	**With tumor**	381, 477, 485, 515, 539, 563, 565, 582, 603, 616, 624, 650, 651, 656, 659, 672, 679, 698, 702, 709, 723, 731, 775, 779, 795, 811, 839
	No tumor	45, 198, 215, 217, 257, 262, 266, 371, 431, 447, 454, 459, 475, 479, 484, 500, 502, 503, 505, 508, 516, 531, 541, 553, 556, 570, 572, 575, 577, 585, 588, 594, 600, 601, 608, 614, 616, 632, 632, 638, 642, 642, 642, 644, 644, 647, 647, 653, 659, 660, 662, 663, 667, 667, 673, 673, 677, 689, 693, 718, 720, 721, 728, 760, 762, 773, 777, 815, 886
GE	**With tumor**	546, 609, 692, 692, 710, 752, 773, 781, 782, 789, 808, 810, 814, 842, 846, 851, 871, 873, 876, 888, 888, 890, 894, 896, 911, 913, 914, 914, 916, 921, 921, 926, 936, 945, 1008
	No tumor	412, 524, 647, 648, 695, 785, 814, 817, 851, 880, 913, 942,986

Ορίζεται $Z_i = 0$ για τα ποντίκια που ζουν στο φυσικό περιβάλλον (CE) και 1 για εκείνα στο αντιμικροβιακό περιβάλλον (GE), με T_i ορίζεται ο χρόνος εμφάνισης του όγκου για κάθε ζώο στη μελέτη και υποθέσαμε ότι μπορούν να περιγραφούν από το μοντέλο PH (1.4.2.1). Για την εκτίμηση της επίδρασης του περιβαλλοντικού παράγοντα στην αύξηση του όγκου, η προσέγγιση μεγίστης πιθανοφάνειας δίνει $\hat{\beta} = 0.6934$ με την εκτιμημένη τυπική απόκλιση να είναι ίση με 0.320 βασισμένη στην προσέγγιση της παρατηρηθείσας πληροφορίας (Fisher). Τα αποτελέσματα δείχνουν ότι τα ζώα στο αντιμικροβιακό περιβάλλον έχουν σημαντικά υψηλότερες επιπτώσεις στον όγκο του πνεύμονα από αυτά που βρίσκονται σε φυσικό περιβάλλον.

Στο Διάγραμμα 3 παρουσιάζεται η εκτιμήτρια μεγίστης πιθανοφάνειας των συναρτήσεων επιβίωσης των χρόνων των ζώων που έχουν καταγραφεί με όγκο στον πνεύμονα για τα δύο περιβαλλοντικά σύνολα. Συγκρίνονται με ξεχωριστές NPMLE των δύο συναρτήσεων επιβίωσης με βάση την παράγραφο 3.2 οι οποίες συμπεριλαμβάνονται και αυτές στο διάγραμμα.

Διάγραμμα 3: Εκτιμήτριες των συναρτήσεων επιβίωσης του χρόνου εμφάνισης του όγκου του πνεύμονα.



Μπορούμε να δούμε από το διάγραμμα ότι οι ξεχωριστές εκτιμήτριες από την παράγραφο 3.2 και οι εκτιμήτριες που δόθηκαν υπό το μοντέλο (1.4.2.1) φαίνονται κοντά η μία στην άλλη, δείχνοντας ότι το μοντέλο (1.4.2.1) μας παρέχει μια αποδεκτή προσέγγιση του προβλήματός μας. Επιπλέον, βλέπουμε ότι οι διαφορές μεταξύ των επιπτώσεων του όγκου του πνεύμονα εμφανίζονται κυρίως στα τελευταία στάδια του πειράματος.

Στο δεύτερο παράδειγμα βλέπουμε την ασβεστοποίηση των ενδοφθαλμιαίων φακών υδρογέλης (IOL) η οποία αναφέρεται ως μία σπάνια επιπλοκή της θεραπείας του καταρράκτη. Η μελέτη αποτελείται από 379 ασθενείς οι οποίοι υπεβλήθησαν σε IOL εμφύτευση και εξετάστηκαν από έναν έμπειρο οφθαλμολόγο ως προς την κατάσταση της ασβεστοποίησης. Για κάθε ασθενή τα δεδομένα δίνουν το χρόνο εξέτασης που κυμαίνεται από 0 έως 37 μήνες από την IOL εμφύτευση και το βαθμό της

σοβαρότητας της ασβεστοποίησης με δείκτες 0 και 1. Με $\delta_i = 0$ δηλώνεται η απουσία ή η ελάχιστου βαθμού ασβεστοποίηση που έχει ο ασθενής κατά το χρόνο της εξέτασης, ενώ με 1 δηλώνεται η ύπαρξη ήπιου ή σοβαρού βαθμού ασβεστοποίηση κατά το χρόνο της εξέτασης. Όλα αυτά τα δεδομένα καθώς και το φύλο των ασθενών παρουσιάζονται στον πίνακα 5 και σκοπός της μελέτης είναι να εκτιμήσει κατά πόσον το φύλο επηρεάζει την IOL ασβεστοποίηση και να αξιολογήσει εάν ο κίνδυνος της επιπλοκής μεταξύ των ανδρών και των γυναικών ασθενών είναι ίδια.

Πίνακας 5: Αριθμός ασθενών με και χωρίς IOL.

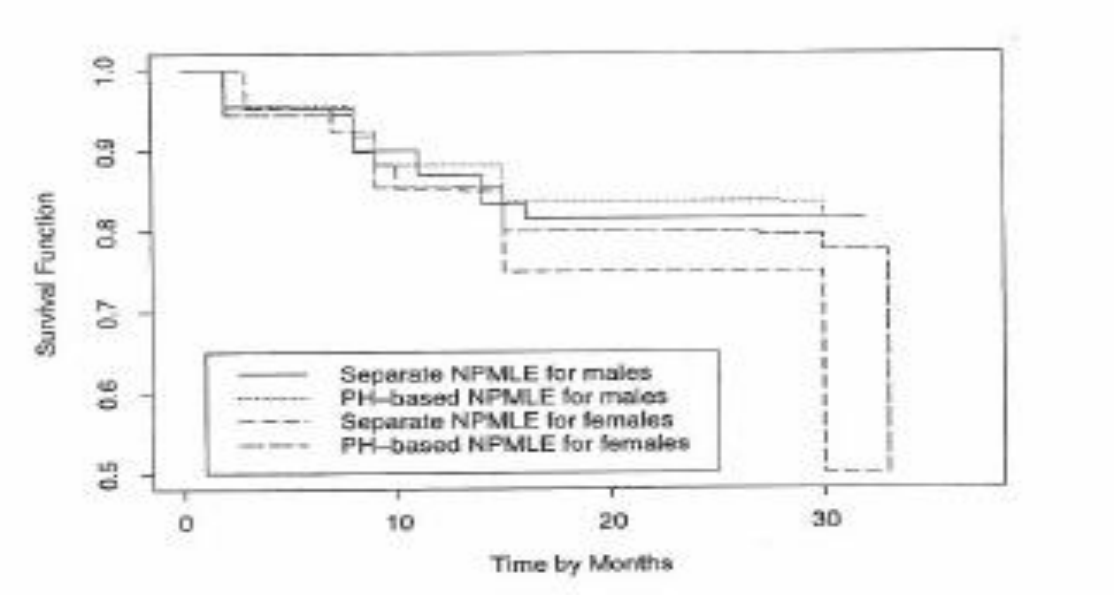
Male patients			**Female patients**		
Exam time	$\delta_i = 0$	$\delta_i = 1$	**Exam time**	$\delta_i = 0$	$\delta_i = 1$
1	3		2	9	
4	11		5	10	
7	9		8	12	1
10	6	1	11	16	5
13	7	1	14	7	
16	5	3	17	7	
19	4		20	2	
26	3		23	1	
32	1		26	2	
2	4	1	30	1	1
5	5	1	37	1	
8	7	1	3	6	1
11	15	1	6	10	
14	8	2	9	12	2
17	6	1	12	19	1

22	1		15	10	5
28	2		18	2	
3	5		21	1	
6	4		24	2	
9	6	1	27		1
12	5	1	31	1	
15	5		4	7	1
18	1	1	7	13	1
24	3		10	15	3
30	1		13	17	3
			16	14	3
			19	3	1
			22	1	
			25	1	
			29	2	
			33		1

Ορίζεται T_i ο χρόνος εμφάνισης της IOL ασβεστοποίησης και υποθέτουμε ότι ακολουθούν το PH μοντέλο (1.4.2.1). Τότε για τα T_i έχουμε διαθέσιμα μόνο τρέχοντα δεδομένα. Επιπλέον ορίζεται $Z_i = 0$ για τις γυναίκες ασθενείς και 1 για τους άντρες. Η εκτιμήτρια μεγίστης πιθανοφάνειας δίνει $\hat{\beta} = -0.2241$ και το εκτιμώμενο τυπικό σφάλμα είναι 0.295 βασισμένο στην προσέγγιση της παρατηρηθείσας πληροφορίας. Χρησιμοποιώντας την Κανονική κατανομή για $\beta = 0$ η p-τιμή είναι 0.448 που σημαίνει ότι δεν υπάρχει σημαντική διαφορά μεταξύ των γυναικών και των αντρών ασθενών.

Το διάγραμμα 4 περιέχει τις ξεχωριστές NPMLE των συναρτήσεων επιβίωσης του χρόνου της IOL ασβεστοποίησης των αντρών και των γυναικών ασθενών της παραγράφου 3.2 καθώς και την εκτιμήτρια μεγίστης πιθανοφάνειας για τις ίδιες συναρτήσεις επιβίωσης υπό το PH μοντέλο.

Διάγραμμα 4: Εκτιμήτριες των συναρτήσεων επιβίωσης του χρόνου IOL ασβεστοποίησης



Από το διάγραμμα επιβεβαιώνεται το παραπάνω συμπέρασμα για μη αισθητή διαφορά ανάμεσα στα δύο φύλα ως προς το χρόνο της IOL ασβεστοποίησης και επιπλέον το μοντέλο PH φαίνεται να έχει λογική προσέγγιση του προβλήματος.

Κεφάλαιο 6

Ανάλυση παλινδρόμησης της Περίπτωσης II για αποκομμένα δεδομένα σε διάστημα

6.1 Εισαγωγή

Αυτό το κεφάλαιο αναφέρεται στην ανάλυση παλινδρόμησης της Περίπτωσης II των αποκομμένων χρόνων αποτυχίας σε διάστημα. Συγκρίνοντάς τα με τα τρέχοντα δεδομένα είναι προφανές ότι στην Περίπτωση II τα αποκομμένα δεδομένα σε διάστημα παρέχουν περισσότερη πληροφορία για το βασικό χρόνο επιβίωσης που μας ενδιαφέρει. Έτσι διαισθητικά, η ανάλυση παλινδρόμησης της Περίπτωσης II φαίνεται να είναι απλούστερη από ότι για τα τρέχοντα δεδομένα. Από την άλλη πλευρά όμως, στην Περίπτωση II χρειάζεται ο χειρισμός δύο ή και παραπάνω μεταβλητών που αντιπροσωπεύουν τους παρατηρηθέντες χρόνους εν αντιθέσει με τη μία μεταβλητή στην περίπτωση των τρεχουσών δεδομένων. Σαν αποτέλεσμα, η ανάλυση παλινδρόμησης στην Περίπτωση II είναι πιο πολύπλοκη και δύσκολη απ΄ότι για τα τρέχοντα δεδομένα και ως προς τη θεωρία και ως προς την πράξη.

Για την ανάλυση της Περίπτωσης II, όπως και στο κεφάλαιο 5, θα γίνει πρώτα μια αναφορά στο PH μοντέλο (1.4.2.1) μαζί με την προσέγγιση μεγίστης πιθανοφάνειας. Αν και η διαδικασία εξαγωγής συμπερασμάτων είναι όμοια με αυτή που περιγράφηκε στην παράγραφο 5.2, η εκτέλεση και ο υπολογισμός στην Περίπτωση II είναι πιο σύνθετοι απ΄ότι για τα τρέχοντα δεδομένα. Επίσης η παραγωγή ασυμπτωτικών ιδιοτήτων είναι πολύ πιο δύσκολη.

6.2 Ανάλυση με το μοντέλο αναλογικής διακινδύνευσης

Μία μελέτη επιβίωσης αποτελείται από *n* ανεξάρτητες μονάδες που παράγουν το εξής διάστημα αποκομμένων δεδομένων:

$$\{(L_i, R_i],\ \mathbf{Z}_i;\ i = 1,...,\ n\} \tag{6.1}$$

που αναφέρεται στους χρόνους επιβίωσης που μας ενδιαφέρουν. Όπως και προηγουμένως $(L_i, R_i],\ i = 1,...,\ n$ δηλώνει το διάστημα μέσα στο οποίο το γεγονός της επιβίωσης για την i-οστή μονάδα έχει παρατηρηθεί ότι έχει συμβεί και το $\mathbf{Z}_i$ αντιπροσωπεύει το *p*-διάστατο διάνυσμα των συμμεταβλητών της μονάδας *i*, i=1,..., n. Επίσης, η $S(t;\mathbf{Z})$ δηλώνει τη συνάρτηση επιβίωσης για μία μονάδα με συμμεταβλητές **Z**. Τότε η συνάρτηση πιθανοφάνειας είναι ανάλογη με

$$L = \prod_{i=1}^{n} [S(L_i, \mathbf{Z}_i) - S(R_i, \mathbf{Z}_i)]$$

θεωρώντας ότι $L_i < R_i$ για όλα τα $i = 1,...,\ n$.

Σε αυτή την παράγραφο θεωρήθηκε ότι $S(t;\mathbf{Z})$ καθορίζεται από το PH μοντέλο (1.4.2.1). Ο λογάριθμος της συνάρτησης πιθανοφάνειας έχει τότε τη μορφή

$$l(\boldsymbol{\beta}, S_0) = \sum_{i=1}^{n} \log\{[S_0(L_i)]^{\exp(\mathbf{Z}_i'\boldsymbol{\beta})} - [S_0(R_i)]^{\exp(\mathbf{Z}_i'\boldsymbol{\beta})}\}$$

ως προς την παράμετρο παλινδρόμησης **β** και τη βασική συνάρτηση επιβίωσης $S_0(t)$. Τα συμπεράσματα για το **β** και την S_0 βασίζονται στην προσέγγιση μεγίστης πιθανοφάνειας που πρώτα μελετήθηκε από τον Finkelstein (1986) και στη συνέχεια παραθέτονται και κάποιες σχετικές ασυμπτωτικές ιδιότητες.

6.2.1 Εκτιμήτρια μεγίστης πιθανοφάνειας

Η πιθανοφάνεια εξαρτάται από την S_0 μόνο μέσω των τιμών της στα διαφορετικά χρονικά παρατηρηθέντα σημεία. Έτσι χρειάζεται η εκτίμηση των τιμών της S_0 σε αυτά τα χρονικά σημεία. Έστω $s_0 = 0 < s_1 < ... < s_{m+1} = \infty$ να δηλώνουν τα διατεταγμένα και διακεκριμένα χρονικά σημεία από όλα τα παρατηρηθέντα εντός του διαστήματος τελικά σημεία $\{L_i, R_i; i = 1, ..., n\}$ και $a_{ij} = I(s_j \in (L_i, R_i]),\ j = 1, ..., m,\ i = 1, ..., n.$ Έστω ότι η $S_0(s_j)$ μπορεί να ξαναγραφτεί ως εξής:

$$S_0(s_j) = \prod_{k=1}^{j} e^{-\exp(a_k)} = e^{-\sum_{k=1}^{j}\exp(a_k)}$$

Στη συνέχεια, ως προς τις παραμέτρους $\boldsymbol{\beta}$ και $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_m)'$, ο λογάριθμος της συνάρτησης πιθανοφάνειας $l(\boldsymbol{\beta}, S_0)$ ξαναγράφεται ως:

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n}\{\log \sum_{j=1}^{m+1} a_{ij}[e^{-a_{j-1}\exp(\mathbf{Z}_i'\boldsymbol{\beta})} - e^{-a_j \exp(\mathbf{Z}_i'\boldsymbol{\beta})}]\}$$

όπου $a_j = \sum_{k=0}^{j}\exp(a_k),\ a_0 = -\infty$ και $a_{m+1} = \infty.$

Για τη μεγιστοποίηση του λογαρίθμου της συνάρτησης πιθανοφάνειας μπορεί να θεωρηθεί ότι προέρχεται από ένα παραμετρικό μοντέλο και να χρησιμοποιηθεί ο αλγόριθμος Newton-Raphson, θα χρειαστούν όμως πρώτα οι **score** συναρτήσεις των **β** και **α** καθώς και ο πίνακας πληροφορίας Fisher. Οι συναρτήσεις είναι

$$U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{Z}_i g_i^{-1} \sum_{j=1}^{m+1} a_{ij}(f_{ij-1} - f_{ij})$$

και

$$U_{a_j}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial a_j} = \sum_{i=1}^{n} g_i^{-1} b_{ij} c_{ij}$$

Όπου

$$f_{ij} = S(s_j;\mathbf{Z}_i)\log S(s_j;\mathbf{Z}_i),\ f_{i0} = f_{im+1} = 0,\ b_{ij} = \exp(a_j + \mathbf{Z}_i'\boldsymbol{\beta}),\ c_{ij} = \sum_{l=i}^{m+1}(a_{il} - a_{il+1})S_{il}(s_j;\mathbf{Z}_i),\ a_{im+2} = 0$$

και $$g_i = \sum_{j=1}^{m+1} a_{ij}[S(s_{j-1};\mathbf{Z}_i) - S(s_j;\mathbf{Z}_i)],\ j = 1,...,\ m,\ i = 1,...,\ n.$$

Τότε οι εκτιμήτριες μεγίστης πιθανοφάνειας των $\boldsymbol{\beta}$ και $\boldsymbol{\alpha}$ προσδιορίζονται λύνοντας τις score συναρτήσεις $U_{\boldsymbol{\beta}}(\boldsymbol{\beta},\boldsymbol{\alpha}) = 0$, $U_{a_j}(\boldsymbol{\beta},\boldsymbol{\alpha}) = 0$, j=1,..., m.

Ο πίνακας πληροφορίας Fisher είναι:

$$I(\boldsymbol{\beta},\boldsymbol{\alpha}) = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}$$

όπου $I_{11} = -\dfrac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'},\ I_{12} = I_{21}' = -\dfrac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial a_j \partial\boldsymbol{\beta}},\ I_{22} = -\dfrac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial a_j \partial a_k}$

και μετά από πράξεις μπορεί να εφαρμοσθεί ο αλγόριθμος Newton-Rapshon.

6.2.2 Ασυμπτωτικές ιδιότητες και συγκρίσεις επιβίωσης

Έστω $\hat{\boldsymbol{\beta}}_n$ και $\hat{\boldsymbol{\alpha}}_n = (\hat{\alpha}_1,...,\ \hat{\alpha}_m)'$ οι εκτιμήτριες μεγίστης πιθανοφάνειας των $\boldsymbol{\beta}$ και $\boldsymbol{\alpha}$ όπως ορίσθηκαν και στην παράγραφο 6.2.1 για δοσμένο n. Η εκτιμήτρια της βασικής συνάρτησης επιβίωσης $S_0(t)$ είναι η $\hat{S}_n(t)$ που αντιστοιχεί στη δεξιά συνεχή συνάρτηση με άλματα μόνο στα s_j και

$$\hat{S}_n(s_j) = \prod_{k=1}^{j} e^{-\exp(\hat{a}_k)},\ j = 1,...,\ m.$$

Επιπλέον $\hat{\Lambda}_n(t) = -\log[\hat{S}_n(t)]$, μία εκτιμήτρια της βασικής σωρευτικής συνάρτησης διακινδύνευσης $\Lambda_0(t)$. Θεωρείται ότι η $S_0(t)$ είναι συνεχής και τα $\mathbf{Z}_i$ φραγμένα. Για την περιγραφή των συνθηκών που απαιτούνται για τις ασυμπτωτικές ιδιότητες των $\hat{\boldsymbol{\beta}}_n$ και $\hat{S}_n(t)$ υποθέτεται ότι τα παρατηρηθέντα

δεδομένα δίνονται με την (1.3.2.1) διατύπωση, δηλαδή τα δεδομένα δίνονται ως προς τα U και V. Τότε μπορεί να δειχθεί ότι $\hat{\boldsymbol{\beta}}_n$ και $\hat{S}_n(t)$ είναι συνεπείς εκτιμήτριες (Huang and Wellner, 1997).

Χρειάζονται περισσότερες συνθήκες για να στηρίξουν την ασυμπτωτική κανονικότητα του $\hat{\boldsymbol{\beta}}_n$ και αυτές περιλαμβάνουν:

α) η ένωση της στήριξης των U, V περιέχεται σε ένα φραγμένο διάστημα μακριά από το μηδέν και

β) S_0 έχει αυστηρά θετική και συνεχή παράγωγο στο διάστημα που ορίσθηκε στο (α). Επιπλέον, έχει θεωρηθεί ότι ο ακριβής χρόνος αποτυχίας δεν έχει παρατηρηθεί. Υπό κάποιες συνθήκες κανονικότητας και καθώς το $n \to \infty$ ισχύει ότι:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to N(0, \Sigma^{-1})$$

κατά κατανομή. Το Σ δηλώνει το κατώτερο φράγμα πληροφορίας για το **β** και έτσι η $\hat{\boldsymbol{\beta}}_n$ είναι ασυμπτωτικά αποτελεσματική.

Όπως έχει αναφερθεί και στο κεφάλαιο 4 η σύγκριση πολλών συναρτήσεων επιβίωσης συχνά έχει ενδιαφέρον στην πράξη. Υπό το μοντέλο που έχει θεωρηθεί εδώ, εάν το $\mathbf{Z}_i$ είναι το διάνυσμα δείκτης για την i-οστή μονάδα, η σύγκριση είναι τότε ισοδύναμη με τον έλεγχο **β**=0 που μπορεί να εκτελεσθεί εφαρμόζοντας τον έλεγχο score. Ένα βασικό πλεονέκτημα αυτού του ελέγχου είναι ότι περιλαμβάνει μόνο την εκτιμήτρια μεγίστης πιθανοφάνειας $\hat{\boldsymbol{\alpha}}_0$ της **α** στο **β**=0, αλλά όχι την εκτιμήτρια μεγίστης πιθανοφάνειας του **β.** Αυτό μπορεί να μειώσει σε μεγάλο βαθμό την υπολογιστική προσπάθεια σε σύγκριση με τον έλεγχο Wald που βασίζεται στο $\hat{\boldsymbol{\beta}}$. Ο στατιστικός έλεγχος score που ορίζεται ως $U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ με **β**=0 και $\hat{\boldsymbol{\alpha}}_0 = \boldsymbol{\alpha}$ έχει τη μορφή:

$$U_{PH} = \sum_{i=1}^{n} \mathbf{Z}_i \frac{\sum_{j=1}^{m+1} a_{ij}[\hat{S}_0(s_{j-1}) \log \hat{S}_0(s_{j-1}) - \hat{S}_0(s_j) \log \hat{S}_0(s_j)]}{\sum_{j=1}^{m+1} a_{ij}[\hat{S}_0(s_{j-1}) - \hat{S}_0(s_j)]} \qquad (6.2)$$

όπου $\hat{S}_0(t) = \hat{S}_n(t)$ στο **β**=0. Ο Finkelstein (1986) πρώτα αναφέρθηκε στον έλεγχο score και πρότεινε την προσέγγιση του πίνακα διασποράς-συνδιασποράς του U_{PH} από το $I_{11|2}$ και έτσι η κατανομή $U_{PH}^{'} I_{11|2}^{-1} U_{PH}$ προσεγγίζεται από την χ^2 κατανομή με p βαθμούς ελευθερίας. Όπου το στοιχείο $I_{11|2}$ είναι του αντίστροφου 2x2 πίνακα διαμέρισης $I^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \begin{pmatrix} I_{11|2}^{-1} & I_{12|2} \\ I_{21|1} & I_{22|1} \end{pmatrix}$ με

$I_{11|2} = I_{11} - I_{12} I_{22}^{-1} I_{21}$, $I_{12|2} = I_{21|1}^{'} = -I_{11|2}^{-1} I_{12} I_{22}^{-1}$ και $I_{22|1} = I_{22}^{-1} + I_{22}^{-1} I_{12} I_{11|2}^{-1} I_{12} I_{22}^{-1}$.

Κεφάλαιο 7

Στατιστική Ανάλυση

7.1 Ανάλυση των αρχικών δεδομένων

Στο κεφάλαιο αυτό θα δούμε πως εφαρμόζονται οι διάφορες κλασικές μέθοδοι αλλά και οι νέες που περιγράψαμε στο θεωρητικό μέρος για την ανάλυση ενός μοντέλου. Τα δεδομένα (Lee, 1980) αφορούν 51 ασθενείς που πάσχουν από οξεία μυελοβλαστική λευχαιμία και υποβάλλονται σε μία θεραπεία, στης οποίας το τέλος βλέπουμε αν έχουν ανταποκριθεί ή όχι. Η μεταβλητή Response (V7) δηλώνει την ανταπόκριση στη θεραπεία: 1 ανταποκρίνεται, 0 δεν ανταποκρίνεται, η μεταβλητή Time (V8) δηλώνει το χρόνο επιβίωσης του ασθενούς σε μήνες και η μεταβλητή Status (V9) την κατάσταση του ασθενούς στο πέρας της έρευνας: 1 δεν έχει επιβιώσει, 0 έχει επιβιώσει. Επιπλέον διαθέτουμε τις ακόλουθες έξι συμμεταβλητές :

1. την ηλικία του ασθενούς, Age (V1)
2. το ποσοστό επίστρωσης των βλαστοκυττάρων, Smear (V2)
3. το ποσοστό των κυττάρων στο μυελό των οστών, Absolute infiltrate (V3)
4. το ποσοστό των κυττάρων που προήλθαν από τον μυελό των οστών, Labelling index (V4)
5. τα απόλυτα βλαστοκύτταρα, Absolute blasts (V5)
6. τη θερμοκρασία του σώματος, Temperature X 10F (V6)

Σε κάθε παρένθεση αναφέρεται το όνομα της μεταβλητής όπως αυτή εμφανίζεται στον παρακάτω πίνακα που φαίνονται τα δεδομένα των 51 ασθενών. Στα δεδομένα αυτά προστέθηκαν και οι στήλες V81 και V82 που δείχνουν το χρόνο των ασθενών ως προς το 12μηνο της θεραπείας που διανύουν. Η Status12 αναφέρεται στην κατάσταση των ασθενών στους 12 μήνες και πρόκειται για δεδομένα που περιέχουν αποκομμένες παρατηρήσεις όπως διαπιστώνουμε από τη μεταβλητή Status.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V81	V82	status12
1	33	92	92	5	2.6	980	1	45	0	42	48	0
2	27	95	95	6	7.5	980	1	36	1	36	42	0
3	28	70	70	14	10.0	1010	1	39	0	36	42	0
4	33	42	38	12	2.5	984	1	36	1	36	42	0
5	..	..	..	..	..	..	..	..	..	..	..	..
..	..	..	..	..	..	..	..	..	..	..	..	..

Α) Αρχικά φτιάχνουμε δύο κατηγορίες για κάθε συμμεταβλητή, κάνοντας χρήση της διαμέσου και εφαρμόζουμε τη συνάρτηση

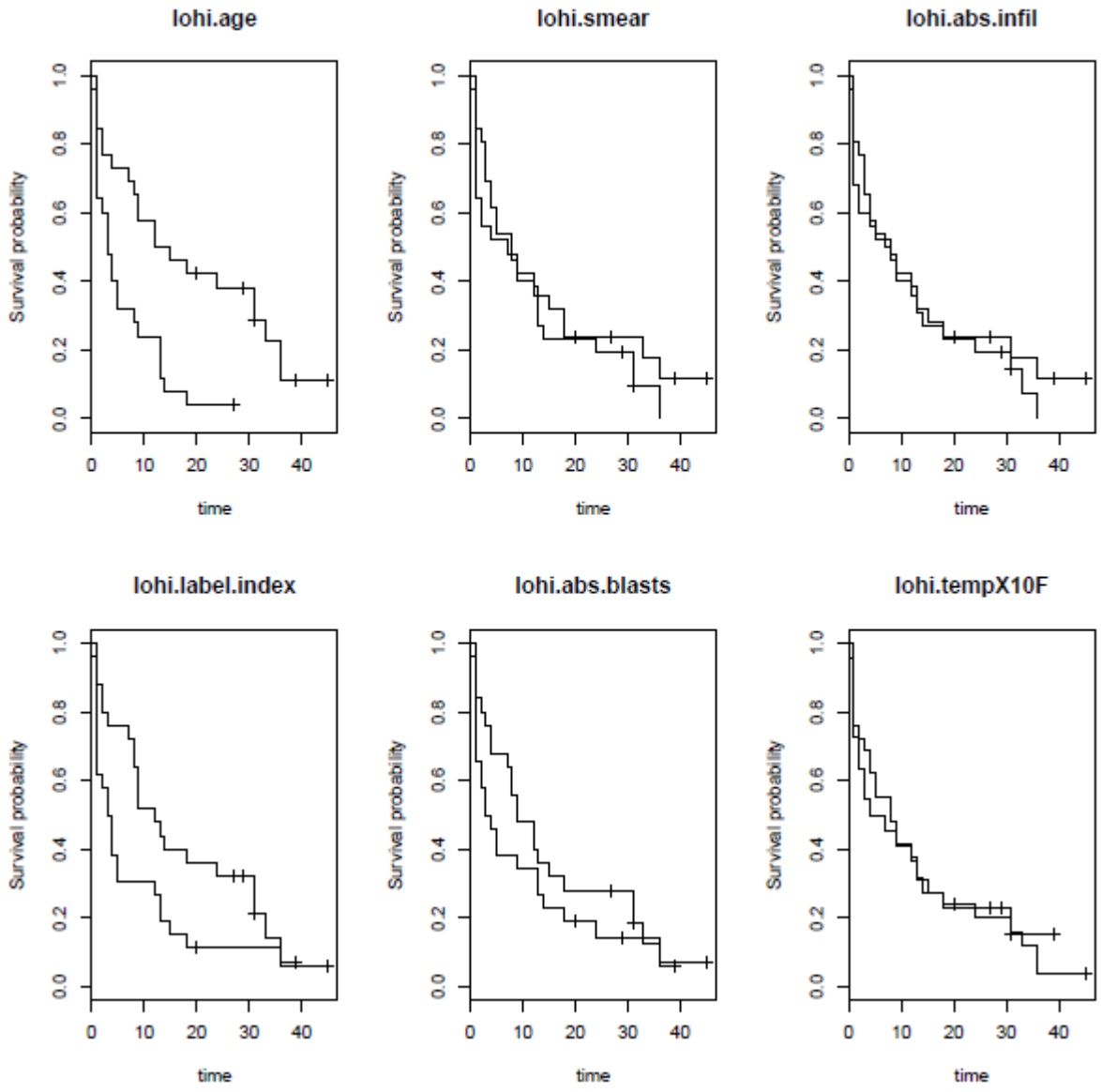
```
cut.at.median <- function(x) {cuts <- c(min(x), median(x), max(x))
cut(x, cuts, include.lowest = TRUE)}.
```

Στη συνέχεια κάνουμε για κάθε μεταβλητή τις αντίστοιχες Kaplan-Meier εκτιμήτριες με την εντολή

```
S1 <- Surv(stime, dstatus == 1,type="right")
```

καθώς και τους κλασικούς ελέγχους log-rank. Τέλος αναλύουμε το μοντέλο του Cox με όλες τις συμμεταβλητές που αφορούν τα αρχικά μας δεδομένα.

Κατασκευάζοντας και τις 6 Kaplan-Meier εκτιμήτριες των συμμεταβλητών παρατηρούμε ότι για τη συμμεταβλητή age οι πιθανότητες επιβίωσης διαφέρουν σημαντικά για τους διαφορετικούς χρόνους των 2 ομάδων. Βλέπουμε ότι στα πρώτα στάδια της μελέτης η πιθανότητα επιβίωσης των ατόμων που έχουν ηλικία κάτω των 50 ετών είναι μεγαλύτερη σε σύγκριση με την πιθανότητα επιβίωσης των ατόμων που ανήκουν στην ηλικιακή ομάδα άνω των 50 ετών. Όσο περνά ο χρόνος αυτή η κατάσταση εντείνεται. Επίσης και για τη συμμεταβλητή labelling index παρατηρούμε διαφορές στις συναρτήσεις επιβίωσης που, όμως, όσο περνά ο χρόνος και πλησιάζοντας προς το τέλος της μελέτης η διαφορά τείνει να εξαλειφθεί. Ως προς τις άλλες συμμεταβλητές οι διαφορές δεν είναι μεγάλες. Αντιλαμβανόμαστε, λοιπόν, ότι η ηλικία του ασθενούς και το ποσοστό των κυττάρων που προήλθαν από τον μυελό των οστών είναι σημαντικοί δείκτες για την πορεία της υγείας του ασθενούς.



Σχήμα 7.1.1: Διαγράμματα Kaplan-Meier εκτιμητριών ως προς τις συμμεταβλητές V1-V6 με χρήση της διαμέσου

Ακολουθώντας την παραπάνω διαδικασία η R τυπώνει τα Αποτελέσματα 7.1.1, στα οποία παρουσιάζεται η ανάλυση του μοντέλου μας.

Αποτελέσματα 7.1.1

```
MV1 <- coxph(S1 ~  age + smear + abs.infil + label.index +
               abs.blasts + tempX10F,
```

```
             ties="breslow")
print(summary(MV1))
```

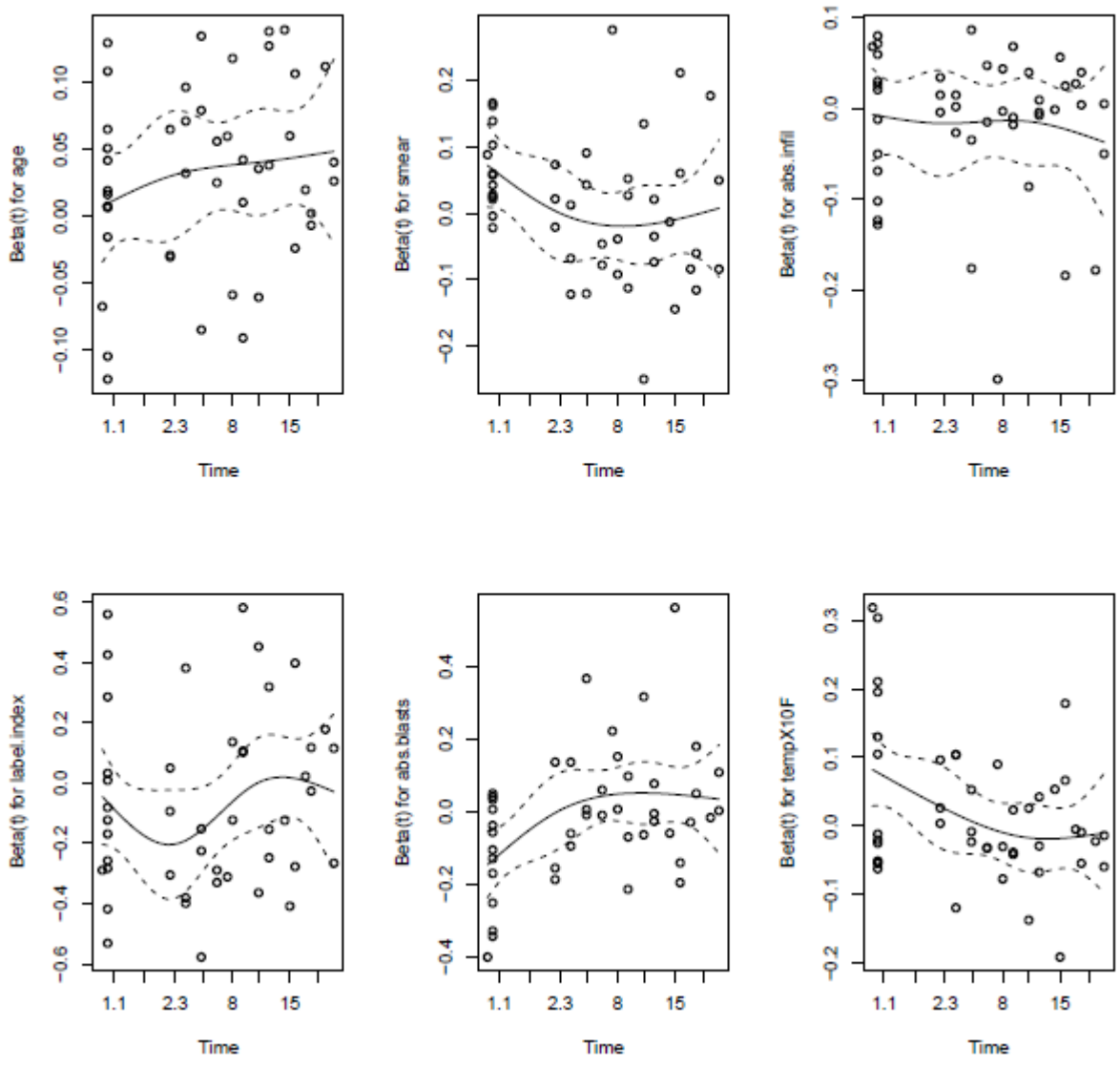
```
## Call:
## coxph(formula = S1 ~ age + smear + abs.infil + label.index +
##     abs.blasts + tempX10F, ties = "breslow")
##
##   n= 51, number of events= 45
##
##                  coef exp(coef) se(coef)      z Pr(>|z|)
## age          0.03093   1.03142  0.01031  2.999  0.00271 **
## smear        0.01238   1.01246  0.01538  0.805  0.42056
## abs.infil   -0.01608   0.98404  0.01249 -1.288  0.19778
## label.index -0.06460   0.93744  0.03881 -1.665  0.09601
## abs.blasts  -0.01306   0.98702  0.02236 -0.584  0.55901
## tempX10F     0.01907   1.01925  0.01300  1.467  0.14242
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##             exp(coef) exp(-coef) lower .95 upper .95
## age            1.0314     0.9695    1.0108     1.052
## smear          1.0125     0.9877    0.9824     1.043
## abs.infil      0.9840     1.0162    0.9602     1.008
## label.index    0.9374     1.0667    0.8688     1.012
## abs.blasts     0.9870     1.0131    0.9447     1.031
## tempX10F       1.0193     0.9811    0.9936     1.046
##
## Concordance= 0.722  (se = 0.057 )
## Rsquare= 0.295   (max possible= 0.997 )
## Likelihood ratio test= 17.83  on 6 df,   p=0.006685
## Wald test            = 16.94  on 6 df,   p=0.009512
## Score (logrank) test = 18.22  on 6 df,   p=0.005714
```

```
cox.zph(MV1)
```

```
##                 rho chisq       p
## age          0.2027  1.76 0.18437
## smear       -0.2710  3.29 0.06959
## abs.infil   -0.0834  0.28 0.59655
## label.index  0.1086  0.67 0.41308
## abs.blasts   0.3917  9.70 0.00185
## tempX10F    -0.3605  8.20 0.00418
## GLOBAL           NA 18.79 0.00452
```

Από τα αποτελέσματα της ανάλυσης φαίνεται οι συμμεταβλητές smear, abs.infiltrate, abs.blasts και temperature να μην είναι στατιστικά σημαντικές, ενώ η συμμεταβλητή age φαίνεται να είναι στατιστικά σημαντική με p-τιμή= 0.00271 και δεύτερη να έρχεται η συμμεταβλητή labelling index, κάτι που είδαμε και νωρίτερα από τις συναρτήσεις επιβίωσης. Επίσης η p-τιμή του ελέγχου log-rank είναι 0.0057

δηλαδή απορρίπτουμε την $H_0 : S_1(t) = S_2(t)$ που σημαίνει ότι οι δύο ομάδες διαφέρουν σημαντικά. Για να σιγουρευτούμε ότι το μοντέλο μας είναι κατάλληλο για την περιγραφή των δεδομένων μας πρέπει να ελέγξουμε ότι οι προϋποθέσεις της αναλογικότητας της διακινδύνευσης αληθεύουν. Τα αντίστοιχα διαγράμματα για τα υπόλοιπα Schoenfeld απεικονίζονται στο σχήμα 7.1.2.



Σχήμα 7.1.2: Διαγράμματα υπολοίπων Schoenfeld ως προς τις συμμεταβλητές V1-V6

Παρατηρούμε ότι μόνο στην περίπτωση της συμμεταβλητής age τα υπόλοιπα έχουν τυχαία ομοσκεδαστική μορφή γύρω από το μηδέν, παρόλ' αυτά η καμπύλη δεν είναι εντελώς ευθεία (οπότε θα

ίσχυε το μοντέλο αναλογικής διακινδύνευσης) αλλά παρουσιάζει μια μικρή αλλά αισθητή αυξητική τάση προς τα δεξιά. Φαίνεται λοιπόν ότι οι μεγαλύτερες ηλικίες έχουν υψηλότερες τιμές υπολοίπων γεγονός που σημαίνει ότι ίσως να υπάρχει διαφοροποίηση μεταξύ της επιβίωσης των ατόμων στα οποία η διάγνωση γίνεται σε μικρότερη ηλικία σε σχέση με αυτά στα οποία η ασθένεια διαγιγνώσκεται σε μεγαλύτερη ηλικία.

Β) Σε αυτή τη φάση θα χρησιμοποιήσουμε τα interval censored δεδομένα (Case II) και αφού αντικαταστήσουμε κάθε διάστημα με την κεντρική του τιμή χρησιμοποιώντας την εντολή

```
stime.c<-apply(mleuk[,10:11],c(1),mean)
S2<-Surv(stime.c,dstatus==1,type="right")
```

η R τυπώνει τα Αποτελέσματα 7.1.2 που αντιστοιχούν στη συμμεταβλητή age, ανάλογα είναι τα αποτελέσματα για τις υπόλοιπες συμμεταβλητές (Παράρτημα (0.2)) και στη συνέχεια έχουμε τα Αποτελέσματα 7.1.3 αντίστοιχα των Αποτελεσμάτων 7.1.1 έχοντας επαναλάβει τα παραπάνω βήματα, όπου ο έλεγχος log-rank εξετάζει αν οι δύο κατηγορίες ασθενών (όπως έχουν ορισθεί για κάθε συμμεταβλητή) διαφοροποιούνται μεταξύ τους ως προς την επιβίωση.

Αποτελέσματα 7.1.2

```
## Kaplan Meier tables for variable
## lohi.age
## Call: survfit(formula = S2 ~ mleuk[, x], data = mleuk)
## mleuk[, x]=[20,50]
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   1.5     26       6    0.769  0.0826       0.6232        0.949
##   4.5     20       1    0.731  0.0870       0.5787        0.923
##   9.0     19       4    0.577  0.0969       0.4151        0.802
##  15.0     15       3    0.462  0.0978       0.3047        0.699
##  21.0     12       1    0.423  0.0969       0.2701        0.663
##  27.0     10       1    0.381  0.0960       0.2323        0.624
##  33.0      8       3    0.238  0.0886       0.1147        0.494
##  39.0      4       2    0.119  0.0742       0.0351        0.404
## mleuk[, x]=(50,80]
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   1.5     25      10     0.60  0.0980      0.43566        0.826
```

```
##   4.5     15       7     0.32  0.0933      0.18071        0.567
##   9.0      8       2     0.24  0.0854      0.11947        0.482
##  15.0      6       4     0.08  0.0543      0.02117        0.302
##  21.0      2       1     0.04  0.0392      0.00586        0.273
```

Αποτελέσματα 7.1.3

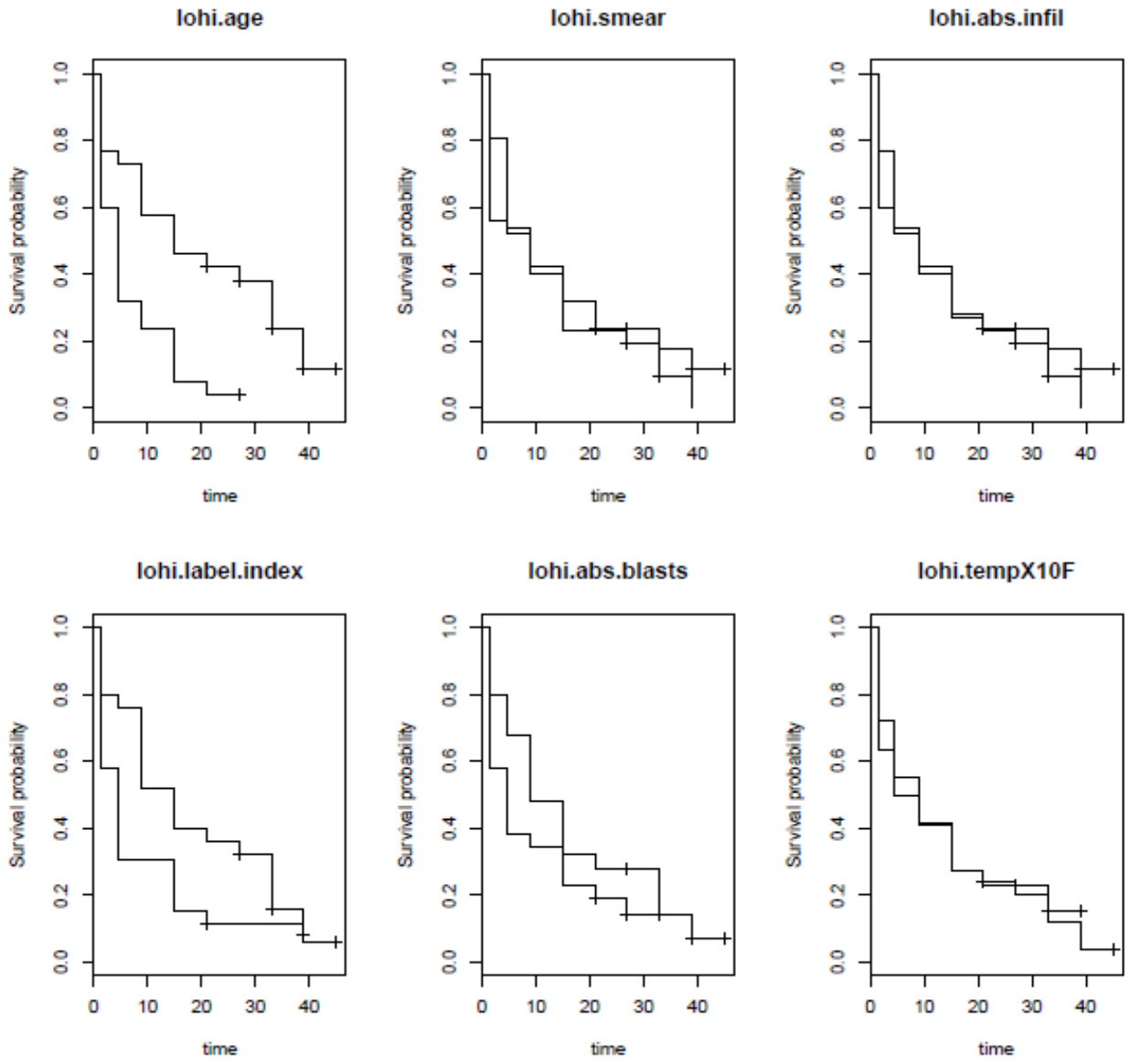
```
MV2 <- coxph(S2 ~  age + smear + abs.infil + label.index + abs.blasts + tempX10F,
             ties="breslow")
print(summary(MV2))
```

```
Call:
## coxph(formula = S2 ~ age + smear + abs.infil + label.index +
##     abs.blasts + tempX10F, ties = "breslow")
##
##   n= 51, number of events= 45
##
##                 coef exp(coef) se(coef)      z Pr(>|z|)
## age          0.02845   1.02886  0.01017  2.796  0.00517 **
## smear        0.01205   1.01213  0.01509  0.799  0.42450
## abs.infil   -0.01427   0.98583  0.01233 -1.157  0.24721
## label.index -0.04986   0.95137  0.03768 -1.323  0.18579
## abs.blasts  -0.01287   0.98721  0.02184 -0.589  0.55580
## tempX10F     0.01366   1.01376  0.01230  1.111  0.26668
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##             exp(coef) exp(-coef) lower .95 upper .95
## age            1.0289     0.9720    1.0085     1.050
## smear          1.0121     0.9880    0.9826     1.043
## abs.infil      0.9858     1.0144    0.9623     1.010
## label.index    0.9514     1.0511    0.8836     1.024
## abs.blasts     0.9872     1.0130    0.9458     1.030
## tempX10F       1.0138     0.9864    0.9896     1.038
##
## Rsquare= 0.239   (max possible= 0.997 )
## Likelihood ratio test= 13.94  on 6 df,   p=0.03032
## Wald test            = 13.18  on 6 df,   p=0.04021
## Score (logrank) test = 13.99  on 6 df,   p=0.0298
```

```
cox.zph(MV2)
```

```
##                 rho  chisq      p
## age          0.1381  0.773 0.3791
## smear       -0.1892  1.562 0.2113
## abs.infil   -0.0807  0.269 0.6040
## label.index  0.1263  0.839 0.3597
## abs.blasts   0.2700  4.239 0.0395
```

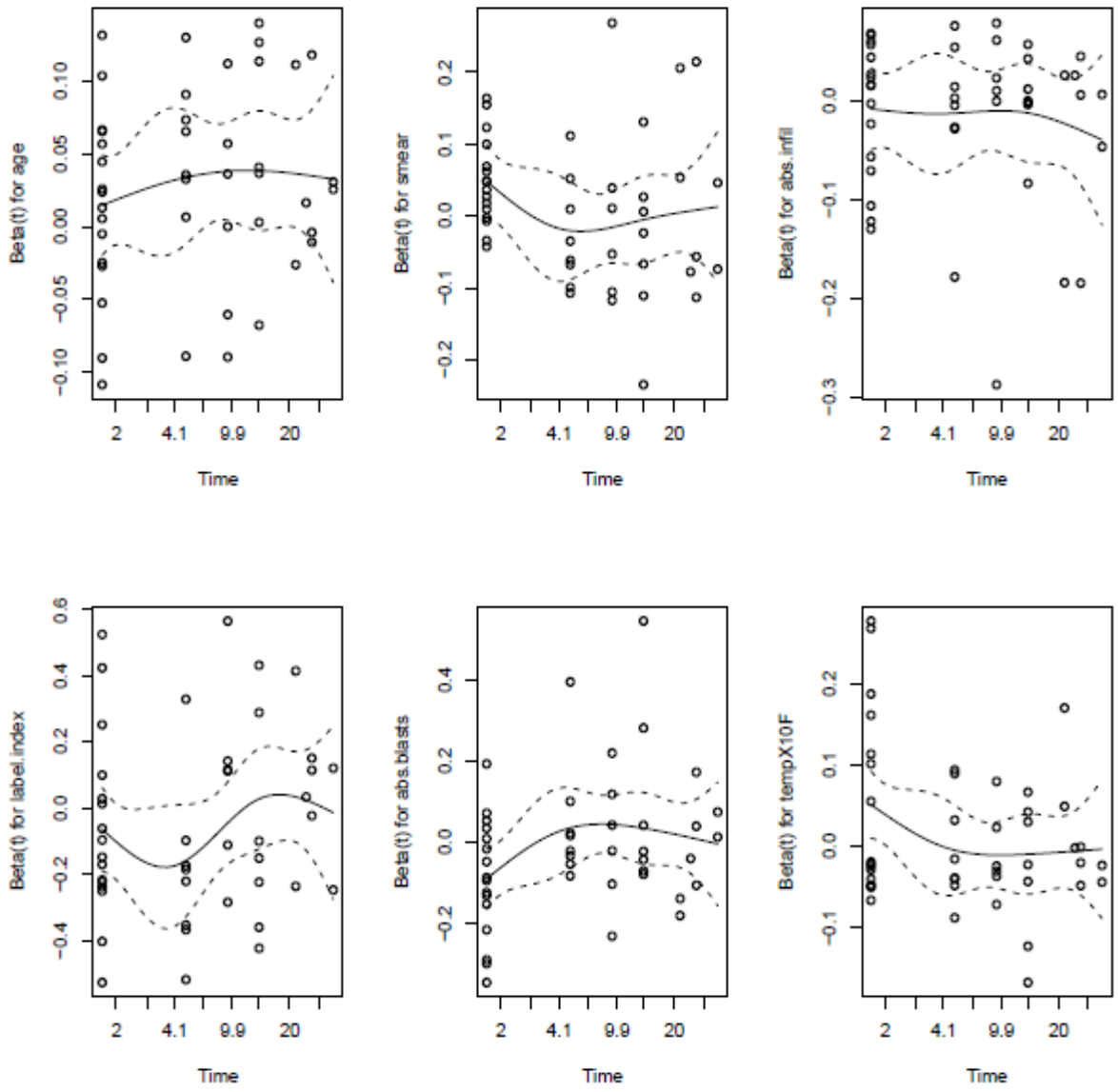
```
## tempX10F    -0.2658  3.953 0.0468
## GLOBAL           NA 11.482 0.0746
```



Σχήμα 7.1.3: Διαγράμματα Kaplan-Meier εκτιμητριών ως προς τις συμμεταβλητές V1-V6 με χρήση της κεντρικής τιμής του διαστήματος

Βλέπουμε ότι στις συναρτήσεις επιβίωσης δεν υπάρχουν αισθητές αλλαγές μετά την αντικατάσταση της κεντρικής τιμής του διαστήματος σε σχέση με αυτές του σχήματος 7.1.1. Η συμμεταβλητή age παραμένει στατιστικά σημαντική, όμως η p-τιμή των συμμεταβλητών abs.infil, label.index και tempx10F αυξήθηκε, γεγονός που δηλώνει μια ασθενέστερη συμβολή στην περιγραφή του μοντέλου. Οι

συμμεταβλητές smear και abs.blasts θα μπορούσαμε να πούμε ότι δε μεταβλήθηκαν καθόλου. Αντίστοιχα η p-τιμή του log-rank ελέγχου παραμένει χαμηλή (p=0.05) οδηγώντας μας ξανά στην απόρριψη της μηδενικής υπόθεσης. Το ίδιο ισχύει και για τα υπόλοιπα αν και υπάρχει μια μικρή τάση βελτίωσης ως προς την προσαρμογή στα δεδομένα (σχήμα 7.1.4).



Σχήμα 7.1.4: Διαγράμματα υπολοίπων Schoenfeld ως προς τις συμμεταβλητές V1-V6

Γ) Σε αυτό το σημείο θα κάνουμε χρήση της εντολής icfit που υπολογίζει τη μη-παραμετρική εκτιμήτρια μεγίστης πιθανοφάνειας (NPMLE) για την κατανομή των αποκομμένων δεδομένων σε διάστημα (παράγραφος 3.4.1) χρησιμοποιώντας την αυτο-συνεπή εκτιμήτρια (self-consistent estimator) έτσι ώστε η σχετική κατανομή επιβίωσης να γενικεύει την Kaplan-Meier εκτιμήτρια. Επιπλέον, θα εφαρμόσουμε και την εντολή ictest για τα νέα log-rank πάλι ως προς τα αποκομμένα δεδομένα σε διάστημα (Case II) για να δούμε αν οι δύο κατηγορίες ασθενών διαφοροποιούνται μεταξύ τους ως προς την επιβίωση (παράγραφος 4.3). Η συγκεκριμένη εντολή εκτελεί μερικούς διαφορετικούς ελέγχους για τα αποκομμένα δεδομένα σε διάστημα. Η προεπιλογή για την ictest είναι να εκτελεί μια δοκιμαστική μετάθεση, είτε ασυμπτωτική είτε ακριβής εξαρτώμενη από το μέγεθος των δεδομένων. Χρησιμοποιώντας την εντολή

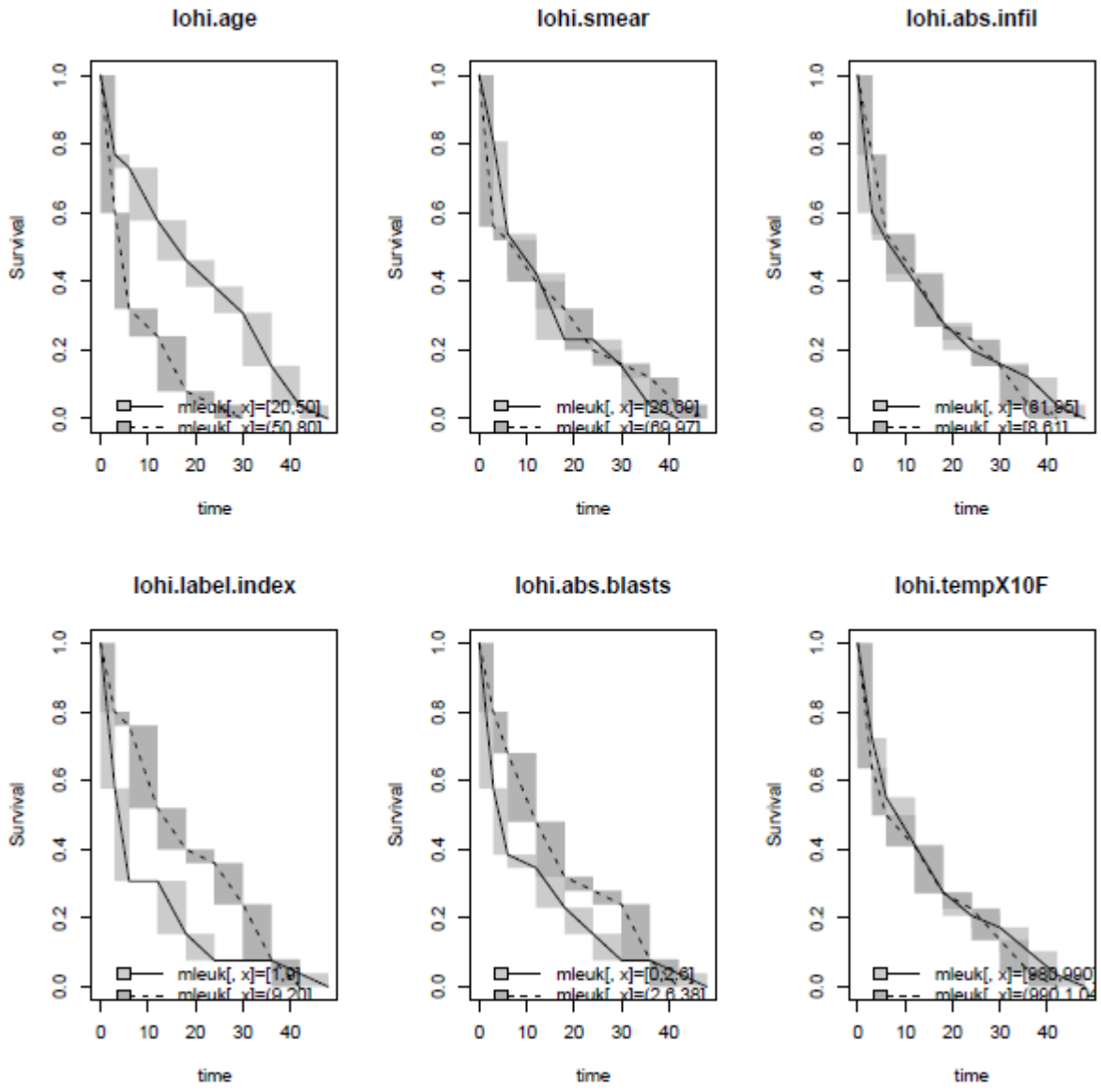
```
S3<-Surv(left,right,type="interval2")
```

για τη δημιουργία ενός νέου αντικειμένου επιβίωσης και τις εντολές icfit και ictest βλέπουμε στα Αποτελέσματα 7.1.4, ενδεικτικά για τη συμμεταβλητή age,

Αποτελέσματα 7.1.4

```
## Logrank trend test for variable
##
## lohi.age
## Asymptotic Logrank two-sample test (permutation form), Sun's
## scores
##
## data:  {left,right} by group
## Z = -3.1523, p-value = 0.00162
## alternative hypothesis: survival distributions not equal
##
##          n Score Statistic*
## [20,50] 26        -9.249215
## (50,80] 25         9.249215
## * like Obs-Exp, positive implies earlier failures than expected
```

ότι η p-τιμή =0.00162 οπότε απορρίπτουμε τη μηδενική υπόθεση όπως και προηγουμένως. Καταλήγουμε, λοιπόν, ότι η συμμεταβλητή age συμβάλλει στην καλή ερμηνεία του μοντέλου και σχετίζεται άμεσα με την επιβίωση του εκάστοτε ασθενή. Όλες οι υπόλοιπες συμμεταβλητές έχουν μεγάλες p-τιμές όπως φαίνεται στο παράρτημα (03) και δε δείχνουν να συμβάλουν στην καλή ερμηνεία του μοντέλου. Στη συνέχεια βλέπουμε τις συναρτήσεις επιβίωσης (σχήμα 7.1.5) που επιβεβαιώνουν τα Αποτελέσματα 7.1.4 ως προς τη σημαντικότητα της συμμεταβλητής age.



Σχήμα 7.1.5: Διαγράμματα συναρτήσεων επιβίωσης των συμμεταβλητών V1-V6

Παρατηρούμε από το διάγραμμα για τη συμμεταβλητή age ότι η ηλικιακή ομάδα των ασθενών που βρίσκεται στο διάστημα (50,80] ετών έχει σημαντικές αποκλίσεις σε σχέση με την ηλικιακή ομάδα των ασθενών που είναι στο διάστημα [20,50], δηλαδή η πιθανότητα επιβίωσης τους είναι σημαντικά μικρότερη σε σχέση με αυτήν της μικρότερης ηλικιακής ομάδας. Επιπλέον, να παρατηρήσουμε ότι το ποσοστό επίστρωσης των βλαστοκυττάρων smear φαίνεται να είναι ο λιγότερο σημαντικός παράγοντας

με p-τιμή =0.8398 και σε αυτή την περίπτωση δεχόμαστε τη μηδενική υπόθεση, δηλαδή ότι οι δύο ομάδες ασθενών δε διαφοροποιούνται μεταξύ τους ως προς την επιβίωση για τη συμμεταβλητή αυτή.

Δ) Τέλος, χρησιμοποιώντας το γενικευμένο γραμμικό μοντέλο για δυαδική εξαρτημένη μεταβλητή (complementary log-log) με όλες τις συμμεταβλητές όπως είχαν δοθεί αρχικά για τα current status (Case I) δεδομένα τη χρονική στιγμή του 12μηνού της θεραπείας που ακολουθούν βλέπουμε στο σχήμα 7.1.6 τα διαγράμματα των υπολοίπων έχοντας εφαρμόσει την εντολή

```
GLM1 <- glm(dstatus12 ~ age + smear + abs.infil + label.index +
abs.blasts + tempX10F,
family=binomial(link=cloglog), data=mleuk)
print(summary(GLM1))
```

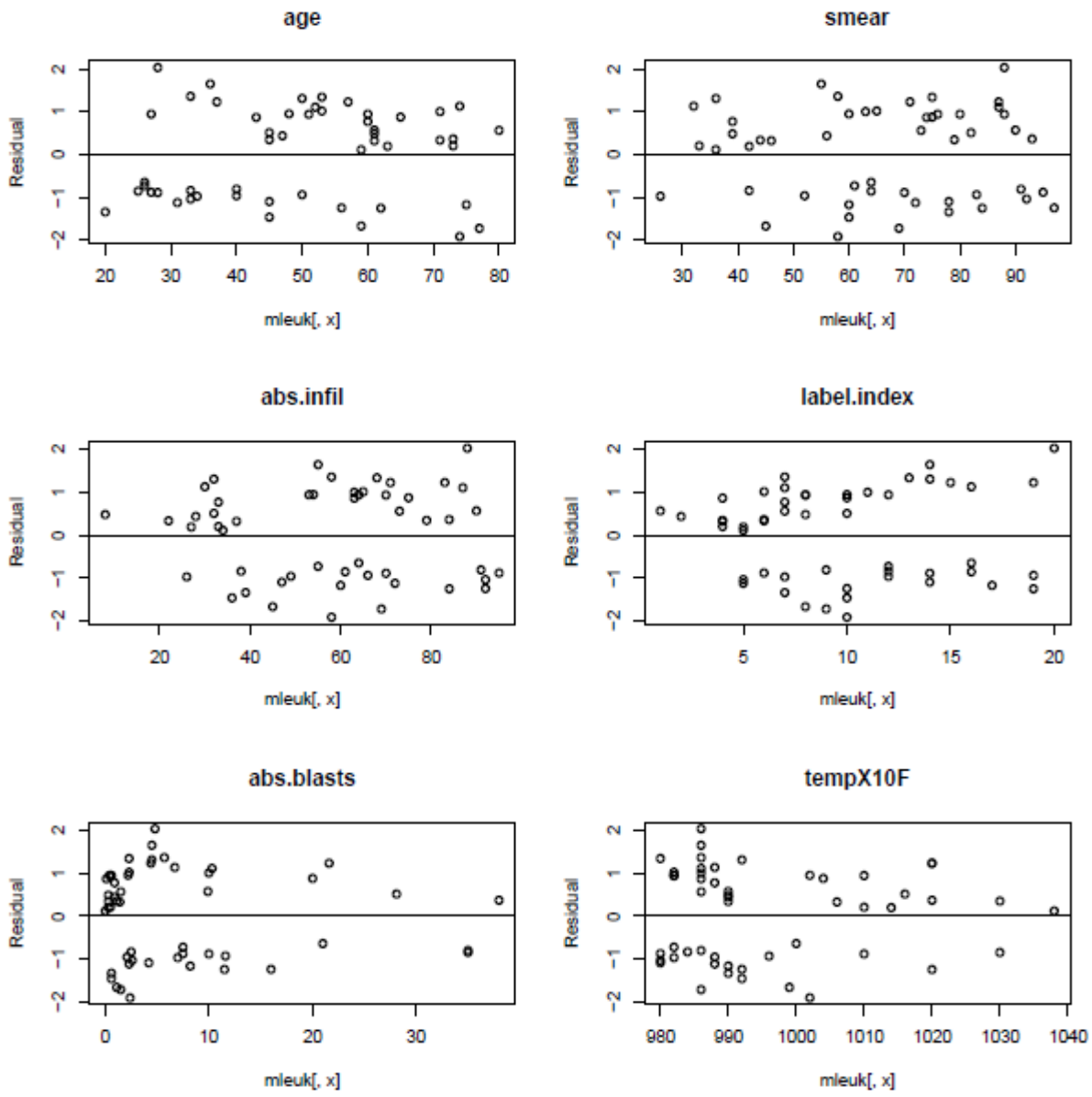
και τα αντίστοιχα Αποτελέσματα 7.1.5.

Αποτελέσματα 7.1.5

```
## Call:
## glm(formula = dstatus12 ~ age + smear + abs.infil + label.index +
## abs.blasts + tempX10F, family = binomial(link = cloglog),data = mleuk)
## Deviance Residuals:
## Min       1Q   Median   3Q     Max
##-1.9094 -0.9455 0.3394 0.9449 2.0332
##
## Coefficients:
##               Estimate  Std. Error  z value  Pr(>|z|)
## (Intercept)  -24.86621  16.21677    -1.533   0.1252
## age            0.02726   0.01351     2.017   0.0437 *
## smear          0.01901   0.02108     0.902   0.3671
## abs.infil     -0.01964   0.01691    -1.161   0.2454
## label.index   -0.09027   0.04829    -1.869   0.0616
## abs.blasts    -0.01472   0.02798    -0.526   0.5989
## tempX10F       0.02437   0.01622     1.503   0.1328
## ---
## Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
## Null deviance:     69.104 on 50 degrees of freedom
## Residual deviance: 55.978 on 44 degrees of freedom
## AIC: 69.978
## Number of Fisher Scoring iterations: 9
   coefficients(GLM1)
## (Intercept)  age        smear      abs.infil   label.index
## -24.86621480 0.02725761 0.01901438 -0.01963720 -0.09026665
## abs.blasts  tempX10F
## -0.01471719 0.02437493
```

Στο πάνω μέρος των αποτελεσμάτων παρουσιάζονται μερικά περιγραφικά στοιχεία για τα υπόλοιπα και πιο συγκεκριμένα η ελάχιστη και η μέγιστη τιμή, η διάμεσος καθώς και τα τεταρτημόρια. Στη συνέχεια παρουσιάζονται οι εκτιμήσεις των παραμέτρων των μοντέλων, τα τυπικά σφάλματά τους, οι τιμές των ελέγχων Wald για τους συντελεστές του μοντέλου και οι αντίστοιχες p-τιμές των ελέγχων. Όσον αφορά τις p-τιμές των ελέγχων βλέπουμε ότι μόνο η συμμεταβλητή age συμβάλλει σημαντικά (p=0.0437) σε αντίθεση με τις υπόλοιπες συμμεταβλητές, γεγονός που μας οδηγεί στο συμπέρασμα ότι η age σχετίζεται πράγματι με την επιβίωση των ασθενών. Επιπλέον βλέπουμε από τα Αποτελέσματα 7.1.5 ότι η τιμή της deviance είναι αρκετά μικρότερη σε σχέση με αυτή του μοντέλου που περιλαμβάνει μόνο το σταθερό όρο (null deviance) ένδειξη ότι το μοντέλο μας περιλαμβάνει και συμμεταβλητές στατιστικά σημαντικές. Προς αυτή την κατεύθυνση οδηγούμαστε και από την τιμή του κριτηρίου AIC το οποίο αποτελεί ένα κριτήριο επιλογής του βέλτιστου μοντέλου με όσο το δυνατόν μικρότερο αριθμό παραμέτρων και προτιμητέο μοντέλο με βάση αυτό το κριτήριο είναι εκείνο με το μικρότερο AIC. Ακόμα μικρότερη τιμή παίρνει το κριτήριο AIC όταν το μοντέλο μας συμπεριλαμβάνει μόνο το σταθερό όρο και τη συμμεταβλητή age (AIC: 47.267) από αυτή που έχει τώρα η deviance, η οποία περιλαμβάνει όλες τις συμμεταβλητές, ένα ακόμη γεγονός που επιβεβαιώνει τη σημαντικότητα αυτής της συμμεταβλητής.

Στο σχήμα 7.1.6 που φαίνεται παρακάτω η υπόθεση της ανεξαρτησίας των υπολοίπων δεν παραβιάζεται, αφού τα υπόλοιπα deviance δεν παρουσιάζουν κάποια ιδιαίτερη τάση στα γραφήματα. Από την άλλη, η υπόθεση της ομοσκεδαστικότητας δεν παραβιάζεται, ίσως όμως σε κάποιο βαθμό ως προς τη συμμεταβλητή abs.blasts.



Σχήμα 7.1.6: Διαγράμματα υπολοίπων deviance

.

7.2 Συμπεράσματα

Σχετικά με τα αποτελέσματα για το σύνολο των δεδομένων που μελετήθηκε, θα λέγαμε ότι στην περίπτωση που ενδιαφερόμαστε για το χρόνο επιβίωσης των ασθενών, τότε η ηλικία φαίνεται να παίζει το σημαντικότερο ρόλο, ενώ εμφανίζεται μια μικρή διαφοροποίηση μεταξύ των ατόμων νεώτερης και μεγαλύτερης ηλικίας. Και στην πρώτη φάση, κάνοντας χρήση της διαμέσου, και στη δεύτερη φάση αντικαθιστώντας κάθε διάστημα με την κεντρική του τιμή η συμμεταβλητή που συμβάλλει περισσότερο στο μοντέλο ήταν η ηλικία (age), δεύτερη ήταν το ποσοστό των κυττάρων που προήλθαν από το μυελό των οστών (labelling index) και στη συνέχεια οι συμμεταβλητές της θερμοκρασίας (tempX10F) και το ποσοστό των κυττάρων στο μυελό των οστών (absolute infiltrate) με την τελευταία να συμβάλλει λίγο περισσότερο στην περιγραφή του μοντέλου στην περίπτωση που τα αποκομμένα δεδομένα σε διάστημα αντικαθίστανται από την κεντρική του τιμή. Αυτό που κατηγορηματικά μπορούμε να πούμε είναι ότι οι συμμεταβλητές του ποσοστού επίστρωσης των βλαστοκυττάρων (smear) και των απόλυτων βλαστοκυττάρων (absolute blasts) σε καμία περίπτωση δεν επηρεάζουν τα αποτελέσματα. Όλες οι μέθοδοι καταλήγουν σε παρόμοια συμπεράσματα.

Αξίζει να σημειώσουμε ότι δεδομένης της μεταβλητής του χρόνου επιβίωσης των ασθενών, ο οποίος είναι διαφορετικός για κάθε άτομο, η ανάλυση γίνεται ως προς το χρόνο επιβίωσης. Είναι επίσης σημαντικό να δούμε ότι κάνοντας χρήση της εντολής icfit για τον υπολογισμό της μη-παραμετρικής εκτιμήτριας μεγίστης πιθανοφάνειας (NPMLE) για την κατανομή των αποκομμένων δεδομένων σε διάστημα (Case II) σε σύγκριση με τη δεύτερη φάση όπου έγινε χρήση της κεντρικής τιμής οι p-τιμές της ηλικίας (age) και του ποσοστού των κυττάρων που προήλθαν από το μυελό των οστών (labelling index) είναι αισθητά μικρότερες με την εφαρμογή της εντολής icfit και οι υπόλοιπες p-τιμές πολύ μεγαλύτερες σε σχέση με αυτές που προέρχονται από τη χρήση της κεντρικής τιμής. Έτσι διαπιστώθηκε ότι στην περίπτωση που κάποιες μεταβλητές σαφώς υπερισχύουν κάποιων άλλων με την έννοια ότι είναι ξεκάθαρα στατιστικά σημαντικές η εντολή icfit τα κατάφερε καλύτερα.

Όλες αυτές οι μέθοδοι έχουν το πλεονέκτημα να βρίσκουν εφαρμογή και σε γενικευμένα αλλά και σε μοντέλα δεδομένων επιβίωσης. Η εφαρμογή όλων των μεθόδων που περιγράψαμε έχει διευκολυνθεί πλέον σημαντικά από τη χρήση των αντίστοιχων στατιστικών πακέτων που έχουν ενσωματωθεί στα περισσότερα στατιστικά περιβάλλοντα. Η R σαφώς προσφέρει μια ολοκληρωμένη χρήση αυτών των μεθόδων.

Παράρτημα

✓ Εισαγωγή στην R

- Η R είναι μία *υπολογιστική γλώσσα και πακέτο* που προσφέρει δυνατότητες διαχείρισης και *στατιστικής ανάλυσης δεδομένων* καθώς και δυνατότητες κατασκευής γραφημάτων.
- Βασίζεται στην *γλώσσα προγραμματισμού* S (που χρησιμοποιεί και το στατιστικό πακέτο S plus) και πρόκειται για *λογισμικό ανοικτού κώδικα* (open source) που διατίθεται ελεύθερα.
- Μπορεί να χρησιμοποιηθεί είτε με κατευθείαν εντολές που υπάρχουν είτε με προγράμματα που ο χρήστης μπορεί να προγραμματίσει για επίλυση πιο πολύπλοκων στατιστικών προβλημάτων.
- Στις συγκεκριμένες σημειώσεις χρησιμοποιούμε την έκδοση 3.1.1

✓ Παράρτημα κώδικα

```
library(survival)

## Loading required package: splines

library(KMsurv)

# read data from *.csv file
mleuk <- read.csv("~/Projects/Sonia/Data/mleuk.csv", header=T)
setwd("~/Projects/Sonia/")

# Create binary variables
# This function splits its argument at the median value
cut.at.median <- function(x) {
  cuts <- c(min(x), median(x), max(x))
  cut(x, cuts, include.lowest = TRUE)
}

# The function cut.at.median is applied to the respective columns
# New values are appended to the data, by adding extra columns
mleuk <- cbind(mleuk,sapply(mleuk[,1:6], cut.at.median))

# Define names for the initial data
names.cont <- c("age", "smear", "abs.infil", "label.index", "abs.blasts",
                "tempX10F", "treat.resp", "stime",
                "dstatus", "left", "right", "dstatus12")

# Define names for the binary data as the original data with prefix "lohi"
names.bin <- paste0("lohi.",names.cont[1:6])
# Assign the names, by concatenating the two character vectors
```

```r
names(mleuk) <- c(names.cont,names.bin)

# Attach data, in order to be able to refer to variables by using their name
attach(mleuk)
```

0.1 Ανάλυση των αρχικών δεδομένων

```r
# Create survival object
S1 <- Surv(stime, dstatus == 1,type="right")

# Kaplan-Meier estimators

#png(filename="1.KM-initial_data.png",width=1400,height=1400,pointsize=32)
par(mfrow=c(2,3))

# create a for loop for repeating the procedure for all binary variables
for (x in 13:18) {
  var <- names(mleuk)[x]
  km.res <- survfit(S1 ~ mleuk[,x], data=mleuk)
  plot(km.res, main=var, xlab="time", ylab="Survival probability")
  summary(km.res)
  survdiff(S1 ~ lohi.age, data = mleuk)
}

#dev.off()


# COX-PH model for continuous independent variables
MV1 <- coxph(S1 ~  age + smear + abs.infil + label.index +
               abs.blasts + tempX10F,
             ties="breslow")
print(summary(MV1))
```

```
## Call:
## coxph(formula = S1 ~ age + smear + abs.infil + label.index +
##     abs.blasts + tempX10F, ties = "breslow")
##
##   n= 51, number of events= 45
##
##                 coef exp(coef) se(coef)      z Pr(>|z|)
## age          0.03093   1.03142  0.01031  2.999  0.00271 **
## smear        0.01238   1.01246  0.01538  0.805  0.42056
## abs.infil   -0.01608   0.98404  0.01249 -1.288  0.19778
## label.index -0.06460   0.93744  0.03881 -1.665  0.09601 .
## abs.blasts  -0.01306   0.98702  0.02236 -0.584  0.55901
## tempX10F     0.01907   1.01925  0.01300  1.467  0.14242
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
##              exp(coef) exp(-coef) lower .95 upper .95
## age             1.0314     0.9695    1.0108     1.052
## smear           1.0125     0.9877    0.9824     1.043
## abs.infil       0.9840     1.0162    0.9602     1.008
## label.index     0.9374     1.0667    0.8688     1.012
## abs.blasts      0.9870     1.0131    0.9447     1.031
## tempX10F        1.0193     0.9811    0.9936     1.046
##
## Concordance= 0.722  (se = 0.057 )
## Rsquare= 0.295   (max possible= 0.997 )
## Likelihood ratio test= 17.83  on 6 df,   p=0.006685
## Wald test            = 16.94  on 6 df,   p=0.009512
## Score (logrank) test = 18.22  on 6 df,   p=0.005714
```

```
cox.zph(MV1)
```

```
##                 rho chisq       p
## age          0.2027  1.76 0.18437
## smear       -0.2710  3.29 0.06959
## abs.infil   -0.0834  0.28 0.59655
## label.index  0.1086  0.67 0.41308
## abs.blasts   0.3917  9.70 0.00185
## tempX10F    -0.3605  8.20 0.00418
## GLOBAL           NA 18.79 0.00452
```

```
#png(filename = "COXPH-test.png",width=1400,height=1400,pointsize=32)
par(mfrow=c(2,3))
plot(cox.zph(MV1))

#dev.off()
```

0.2 Interval censored (Case II) δεδομένα

```
# Create new time variable (mean of interval)
stime.c <- apply(mleuk[,10:11],c(1),mean)
S2 <- Surv(stime.c, dstatus == 1,type="right")

# Kaplan-Meier estimators
#png(filename="2.KM-interval_data.png",width=1400,height=1400,pointsize=32)
par(mfrow=c(2,3))
for (x in 13:18) {
  var <- names(mleuk)[x]
  cat(paste("Kaplan Meier tables for variable \n\n", var ),"\n\n")
  km.res2 <- survfit(S2 ~ mleuk[,x], data=mleuk)
  plot(km.res2, main=var, xlab="time", ylab="Survival probability")
  print(summary(km.res2))
  km.logrank2 <- survdiff(S2 ~ mleuk[,x], data = mleuk)
  km.logrank2
}
```

```
## Kaplan Meier tables for variable
##
##  lohi.age

## Call: survfit(formula = S2 ~ mleuk[, x], data = mleuk)
##
##                 mleuk[, x]=[20,50]
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   1.5     26       6    0.769  0.0826       0.6232        0.949
##   4.5     20       1    0.731  0.0870       0.5787        0.923
##   9.0     19       4    0.577  0.0969       0.4151        0.802
##  15.0     15       3    0.462  0.0978       0.3047        0.699
##  21.0     12       1    0.423  0.0969       0.2701        0.663
##  27.0     10       1    0.381  0.0960       0.2323        0.624
##  33.0      8       3    0.238  0.0886       0.1147        0.494
##  39.0      4       2    0.119  0.0742       0.0351        0.404
##
##                 mleuk[, x]=(50,80]
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   1.5     25      10     0.60  0.0980      0.43566        0.826
##   4.5     15       7     0.32  0.0933      0.18071        0.567
##   9.0      8       2     0.24  0.0854      0.11947        0.482
##  15.0      6       4     0.08  0.0543      0.02117        0.302
##  21.0      2       1     0.04  0.0392      0.00586        0.273
##
## Kaplan Meier tables for variable
##
##  lohi.smear

## Call: survfit(formula = S2 ~ mleuk[, x], data = mleuk)
##
##                 mleuk[, x]=[26,69]
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   1.5     26       5   0.8077  0.0773       0.6696        0.974
##   4.5     21       7   0.5385  0.0978       0.3772        0.769
##   9.0     14       3   0.4231  0.0969       0.2701        0.663
##  15.0     11       5   0.2308  0.0826       0.1144        0.466
##  27.0      6       1   0.1923  0.0773       0.0875        0.423
##  33.0      4       2   0.0962  0.0617       0.0273        0.338
##  39.0      1       1   0.0000     NaN           NA           NA
##
##                 mleuk[, x]=(69,97]
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   1.5     25      11     0.56  0.0993       0.3956        0.793
##   4.5     14       1     0.52  0.0999       0.3568        0.758
##   9.0     13       3     0.40  0.0980       0.2475        0.646
##  15.0     10       2     0.32  0.0933       0.1807        0.567
##  21.0      8       2     0.24  0.0854       0.1195        0.482
##  33.0      4       1     0.18  0.0825       0.0733        0.442
##  39.0      3       1     0.12  0.0736       0.0360        0.400
```

```
##
## Kaplan Meier tables for variable
##
##  lohi.abs.infil

## Call: survfit(formula = S2 ~ mleuk[, x], data = mleuk)
##
##                 mleuk[, x]=(61,95]
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   1.5     25      10     0.60  0.0980       0.4357        0.826
##   4.5     15       2     0.52  0.0999       0.3568        0.758
##   9.0     13       3     0.40  0.0980       0.2475        0.646
##  15.0     10       3     0.28  0.0898       0.1493        0.525
##  21.0      7       1     0.24  0.0854       0.1195        0.482
##  33.0      4       1     0.18  0.0825       0.0733        0.442
##  39.0      3       1     0.12  0.0736       0.0360        0.400
##
##                 mleuk[, x]=[8,61]
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   1.5     26       6   0.7692  0.0826       0.6232        0.949
##   4.5     20       6   0.5385  0.0978       0.3772        0.769
##   9.0     14       3   0.4231  0.0969       0.2701        0.663
##  15.0     11       4   0.2692  0.0870       0.1429        0.507
##  21.0      7       1   0.2308  0.0826       0.1144        0.466
##  27.0      6       1   0.1923  0.0773       0.0875        0.423
##  33.0      4       2   0.0962  0.0617       0.0273        0.338
##  39.0      1       1   0.0000     NaN           NA           NA
##
## Kaplan Meier tables for variable
##
##  lohi.label.index

## Call: survfit(formula = S2 ~ mleuk[, x], data = mleuk)
##
##                 mleuk[, x]=[1,9]
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   1.5     26      11   0.5769  0.0969       0.4151        0.802
##   4.5     15       7   0.3077  0.0905       0.1729        0.548
##  15.0      8       4   0.1538  0.0708       0.0625        0.379
##  21.0      4       1   0.1154  0.0627       0.0398        0.334
##  39.0      2       1   0.0577  0.0514       0.0101        0.331
##
##                 mleuk[, x]=(9,20]
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   1.5     25       5     0.80  0.0800       0.6576        0.973
##   4.5     20       1     0.76  0.0854       0.6097        0.947
##   9.0     19       6     0.52  0.0999       0.3568        0.758
##  15.0     13       3     0.40  0.0980       0.2475        0.646
##  21.0     10       1     0.36  0.0960       0.2135        0.607
##  27.0      9       1     0.32  0.0933       0.1807        0.567
```

```
##  33.0      6       3     0.16  0.0803       0.0599        0.428
##  39.0      2       1     0.08  0.0694       0.0146        0.438
##
## Kaplan Meier tables for variable
##
##  lohi.abs.blasts

## Call: survfit(formula = S2 ~ mleuk[, x], data = mleuk)
##
##                 mleuk[, x]=[0,2.6]
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   1.5     26      11   0.5769  0.0969       0.4151        0.802
##   4.5     15       5   0.3846  0.0954       0.2365        0.625
##   9.0     10       1   0.3462  0.0933       0.2041        0.587
##  15.0      9       3   0.2308  0.0826       0.1144        0.466
##  21.0      6       1   0.1923  0.0773       0.0875        0.423
##  27.0      4       1   0.1442  0.0714       0.0547        0.380
##  39.0      2       1   0.0721  0.0622       0.0133        0.391
##
##                 mleuk[, x]=(2.6,38]
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   1.5     25       5     0.80  0.0800       0.6576        0.973
##   4.5     20       3     0.68  0.0933       0.5197        0.890
##   9.0     17       5     0.48  0.0999       0.3192        0.722
##  15.0     12       4     0.32  0.0933       0.1807        0.567
##  21.0      8       1     0.28  0.0898       0.1493        0.525
##  33.0      6       3     0.14  0.0727       0.0506        0.387
##  39.0      2       1     0.07  0.0614       0.0125        0.391
##
## Kaplan Meier tables for variable
##
##  lohi.tempX10F

## Call: survfit(formula = S2 ~ mleuk[, x], data = mleuk)
##
##                 mleuk[, x]=[980,990]
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   1.5     29       8   0.7241  0.0830      0.57844        0.907
##   4.5     21       5   0.5517  0.0923      0.39742        0.766
##   9.0     16       4   0.4138  0.0915      0.26832        0.638
##  15.0     12       4   0.2759  0.0830      0.15297        0.497
##  21.0      8       1   0.2414  0.0795      0.12661        0.460
##  27.0      6       1   0.2011  0.0757      0.09618        0.421
##  33.0      5       2   0.1207  0.0633      0.04318        0.337
##  39.0      3       2   0.0402  0.0390      0.00601        0.270
##
##                 mleuk[, x]=(990,1.04e+03]
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   1.5     22       8    0.636  0.1026       0.4640        0.873
##   4.5     14       3    0.500  0.1066       0.3292        0.759
```

```
##   9.0     11       2    0.409  0.1048       0.2476        0.676
##  15.0      9       3    0.273  0.0950       0.1378        0.540
##  21.0      6       1    0.227  0.0893       0.1052        0.491
##  33.0      3       1    0.152  0.0859       0.0499        0.460
```

```
#dev.off()

# COX-PH model for continuous independent variables
MV2 <- coxph(S2 ~  age + smear + abs.infil + label.index + abs.blasts + tempX10F,
             ties="breslow")
print(summary(MV2))
```

```
## Call:
## coxph(formula = S2 ~ age + smear + abs.infil + label.index +
##     abs.blasts + tempX10F, ties = "breslow")
##
##   n= 51, number of events= 45
##
##                 coef exp(coef) se(coef)      z Pr(>|z|)
## age          0.02845   1.02886  0.01017  2.796  0.00517 **
## smear        0.01205   1.01213  0.01509  0.799  0.42450
## abs.infil   -0.01427   0.98583  0.01233 -1.157  0.24721
## label.index -0.04986   0.95137  0.03768 -1.323  0.18579
## abs.blasts  -0.01287   0.98721  0.02184 -0.589  0.55580
## tempX10F     0.01366   1.01376  0.01230  1.111  0.26668
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##             exp(coef) exp(-coef) lower .95 upper .95
## age            1.0289     0.9720    1.0085     1.050
## smear          1.0121     0.9880    0.9826     1.043
## abs.infil      0.9858     1.0144    0.9623     1.010
## label.index    0.9514     1.0511    0.8836     1.024
## abs.blasts     0.9872     1.0130    0.9458     1.030
## tempX10F       1.0138     0.9864    0.9896     1.038
##
## Concordance= 0.734  (se = 0.067 )
## Rsquare= 0.239   (max possible= 0.997 )
## Likelihood ratio test= 13.94  on 6 df,   p=0.03032
## Wald test            = 13.18  on 6 df,   p=0.04021
## Score (logrank) test = 13.99  on 6 df,   p=0.0298
```

```
cox.zph(MV2)
```

```
##                 rho  chisq      p
## age          0.1381  0.773 0.3791
## smear       -0.1892  1.562 0.2113
## abs.infil   -0.0807  0.269 0.6040
## label.index  0.1263  0.839 0.3597
## abs.blasts   0.2700  4.239 0.0395
```

```
## tempX10F    -0.2658  3.953 0.0468
## GLOBAL           NA 11.482 0.0746
```

```
# png(filename = "2.COXPH-test.png",width=1400,height=1400,pointsize=32)
par(mfrow=c(2,3))
plot(cox.zph(MV2))
```

```
dev.off()
```

```
## null device
##           1
```

0.3 Interval censored (Case II) (Fay and Shaw)

```
library(interval)
```

```
## Loading required package: perm
## Loading required package: Icens
## Loading required package: MLEcens
```

```
# Create new survival object
S3 <- Surv(left,right,type="interval2")

#png(filename = "3.IntervalCensoring.png",width=1400,height=1400,pointsize=32)
par(mfrow=c(2,3))

for (x in 13:18) {
  var <- names(mleuk)[x]
  cat(paste("Logrank trend test for variable \n\n", var ),"\n\n")
  icfit(S3~mleuk[,x], data=mleuk)
  plot(icfit(S3~mleuk[,x], data=mleuk), main=var)
  print(ictest(left,right,mleuk[,x]))
}
```

```
## Logrank trend test for variable
##
##  lohi.age
```

```
##
##  Asymptotic Logrank two-sample test (permutation form), Sun's
##  scores
##
## data:  {left,right} by group
## Z = -3.1523, p-value = 0.00162
## alternative hypothesis: survival distributions not equal
##
##          n Score Statistic*
## [20,50] 26        -9.249215
## (50,80] 25         9.249215
```

```
## * like Obs-Exp, positive implies earlier failures than expected
## Logrank trend test for variable
##
##  lohi.smear

##
##  Asymptotic Logrank two-sample test (permutation form), Sun's
##  scores
##
## data:  {left,right} by group
## Z = -0.2022, p-value = 0.8398
## alternative hypothesis: survival distributions not equal
##
##          n Score Statistic*
## (69,97] 25       -0.5932816
## [26,69] 26        0.5932816
## * like Obs-Exp, positive implies earlier failures than expected
## Logrank trend test for variable
##
##  lohi.abs.infil

##
##  Asymptotic Logrank two-sample test (permutation form), Sun's
##  scores
##
## data:  {left,right} by group
## Z = -0.2071, p-value = 0.836
## alternative hypothesis: survival distributions not equal
##
##          n Score Statistic*
## (61,95] 25       -0.6075673
## [8,61]  26        0.6075673
## * like Obs-Exp, positive implies earlier failures than expected
## Logrank trend test for variable
##
##  lohi.label.index

##
##  Asymptotic Logrank two-sample test (permutation form), Sun's
##  scores
##
## data:  {left,right} by group
## Z = 1.7271, p-value = 0.08415
## alternative hypothesis: survival distributions not equal
##
##         n Score Statistic*
## [1,9]  26         5.067524
## (9,20] 25        -5.067524
## * like Obs-Exp, positive implies earlier failures than expected
## Logrank trend test for variable
```

```
##
##  lohi.abs.blasts

##
##  Asymptotic Logrank two-sample test (permutation form), Sun's
##  scores
##
## data:  {left,right} by group
## Z = 1.1301, p-value = 0.2585
## alternative hypothesis: survival distributions not equal
##
##           n Score Statistic*
## [0,2.6]  26          3.31572
## (2.6,38] 25         -3.31572
## * like Obs-Exp, positive implies earlier failures than expected
## Logrank trend test for variable
##
##  lohi.tempX10F

##
##  Asymptotic Logrank two-sample test (permutation form), Sun's
##  scores
##
## data:  {left,right} by group
## Z = -0.5347, p-value = 0.5929
## alternative hypothesis: survival distributions not equal
##
##                 n Score Statistic*
## [980,990]      29        -1.554244
## (990,1.04e+03] 22         1.554244
## * like Obs-Exp, positive implies earlier failures than expected
```

```
#dev.off()
```

0.4 Current status (Case I)

Status12 is 1 = dead at 12 months, 0 = alive at 12 months. Nobody was censored before 12 months. There are 21 alive at 12 months

```
GLM1 <- glm(dstatus12 ~ age + smear + abs.infil + label.index +
abs.blasts + tempX10F,
family=binomial(link=cloglog), data=mleuk)
print(summary(GLM1))
##
## Call:
## glm(formula = dstatus12 ~ age + smear + abs.infil + label.index +
## abs.blasts + tempX10F, family = binomial(link = cloglog),
## data = mleuk)
## Deviance Residuals:
## Min        1Q   Median   3Q   Max
## -1.9094 -0.9455 0.3394 0.9449 2.0332
##
```

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -24.86621 16.21677   -1.533  0.1252
## age           0.02726  0.01351    2.017  0.0437 *
## smear         0.01901  0.02108    0.902  0.3671
## abs.infil    -0.01964  0.01691   -1.161  0.2454
## label.index  -0.09027  0.04829   -1.869  0.0616 .
## abs.blasts   -0.01472  0.02798   -0.526  0.5989
## tempX10F      0.02437  0.01622    1.503  0.1328
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
## Null deviance:     69.104 on 50 degrees of freedom
## Residual deviance: 55.978 on 44 degrees of freedom
## AIC: 69.978
## Number of Fisher Scoring iterations: 9
coefficients(GLM1)
## (Intercept)     age       smear     abs.infil   label.index
## -24.86621480 0.02725761 0.01901438 -0.01963720 -0.09026665
## abs.blasts  tempX10F
## -0.01471719 0.02437493
f <- fitted.values(GLM1)
r <- residuals(GLM1)
#png(filename = "4.GLMresiduals.png", width=1400,height=1400,pointsize=32)
par(mfrow=c(3,2))
for (x in 1:6) {
plot(mleuk[,x],r, main=names(mleuk)[x],ylab = "Residual")
abline(0,0)
}
#dev.off()
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Abayomi K., Gelman A. and Levy M. (2008). Diagnostics for multivariate imputations. *Applied Statistics,* **57** (3), 273-291.

[2] Adamczyk A. and Palmer I. (2008). Religion and initiation into marijuana use: The dettering role of religious friends. *Journal of Drug Issues,* **38** (3), 717-741.

[3] Alati R., Mamum AA., Williams GM., O'Callaghan M., Najman JM. and Bor W. (2006). In utero alcohol exposure and prediction of alcohol disorders in early adulthood: A birth cohort study. *Archives of General Psychiatry,* **63** (9), 1009-1016.

[4] Albert PS. and Follman D. (2009). *Shared-Parameter Models*. In G. Fitzmaurice, M. Davidian, G. Verbeke, G. Molenberghs (eds.) *Longitudinal Data Analysis,* chapter 18, pp. 433-452. CRC Press, Boca Raton, FL.

[5] Amber G. and Omar RZ., Royston P., Kinsman R., Keogh BE. and Taylor KM. (2005). "Generic, simple risk stratification model for heart valve surgery. *Circulation,* **112** (2), 224-231.

[6] Andersen, P. K. and Gill, R. D. (1982). Cox regression model for counting processes: a large sample study. *The Annals of Statistics*, **10**, 1100-1120.

[7] Andersen, P. K. and Ronn, B. B. (1995). A nonparametric test for comparing two samples where all observations are either left- or right-censored. *Biometric,* **51**, 323-329.

[8] Arnold BC., Castillo E. and Sarabia JM. (1999). *Conditional Specification of Statistical Models.* Springer-Verlag, New York.

[9] Arnold BC. and Press SJ. (1989). Compatible conditional distributions. *Journal of the American Statistical Association,* **84**, 152-156.

[10] Barlow, R. E. Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972), *Statistical Inference under order restrictions.* New York: Wiley.

[11] Böhning, D., Schlattmann, P. and Dietz, E. (1996). Interval censored data: A note on the nonparametric maximum likelihood estimator of the distribution function. *Biometrika*, **83**, 462-466.

[12] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, **34**, 187-220.

[13] Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**, 269-276.

[14] Cox, D. R. and Oakes, D. (1984). *Analysis of survival data*, Chapman &Hall: London.

[15] De Gruttola, V. and Lagakos, S. W. (1989). Analysis of doubly-censored survival data, with application to AIDS. *Biometrics*, **45**, 1-12.

[16] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1-38.

[17] Diamond, I. D. and McDonald, J.W. (1991). The analysis of current status data. *Demographic Applications of Event History Analysis*, eds. Trussel, J., Hankinson, R. and Tilton, J. Oxford University Press: Oxford, U.K.

[18] Dinse, G. E. and Lagakos, S. W. (1983). Regression analysis of tumor prevalence data. *Applied Statistics*, **32**, 236-248.

[19] Dorey, F. J., Little, Roderick J. A., and Schenker, N. (1993). Multiple imputation for threshold-crossing data with interval censoring. *Statistics in Medicine*, **12**, 1589-1603.

[20] Efron, B. (1967). The two sample problem with censored data. In Proc. 5^{Th} *Berkeley Symp. On Math. Statist. Prob.*. Berkeley: University of California Press, 831-853.

[21] Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, **42**, 845-854.

[22] Finkelstein, D. M. and Wolfe, R. A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics*, **41**, 933-945.

[23] Fleming, T. R. and Harrington, D. P. (1991). *Counting process and survival analysis*, John Wiley: New York.

[24] Freireich, E. O. et al. (1963). The effect of 6-mercaptopmine on the duration of steroid induced remission in acute leukemia., *Blood*, **21**, 699-716.

[25] Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily single-censored samples. *Biometrika*, **52**, 203-223.

[26] Gentleman, R. and Geyer, C. J. (1994). Maximum likelihood for interval censored data: Consistency and computation. *Biometrika*, **81**, 618-623.

[27] Goggins, W. B., Finkelstein, D. M., Schoenfeld, D. A. and Zaslavsky, A. M. (1998). A Markov chain Monte Carlo EM algorithm for analyzing interval censored data under the Cox proportional hazards model. *Biometrics*, **54**, 1498-1507.

[28] Greenwood M. (1926). The natural duration of cancer. *Reports on Public Health and Medical Subjects*, **33**, 1-26.

[29] Groeneboom, P. and Wellner, J. A. (1992). *Information bounds and nonparametric maximum likelihood estimation*. DMV Seminar, Band 19, Birkhauser, New York.

[30] Hoel, D. G. and Walburg, H. E. (1972). Statistical analysis of survival experiments. *Journal of National Cancer Institute,* **49**, 361-372.

[31] Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *The Annals of Statistics*, **24**, 540-568.

[32] Huang, J. and Rossini, A. J. (1997). Sieve estimation for the proportional odds failure- time regression model with interval censoring. *Journal of the American Statistical Association,* **92**, 960-967.

[33] Huang, J. and Wellner, J. A. (1995). Asymptotic normality of the NPMLE of linear functionals for interval censored data, case I. *Statistica Neerlandica,* **49**, 153-163.

[34] Huang, J. and Wellner, J. A. (1997). Interval censored survival data: a review of recent progress. *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis,* eds. Lin, D. and Fleming, T. Springer-Verlag, New York, 123-169.

[35] Jongbloed, G. (1998). The iterative convex minorant algorithm for nonparametric estimation. *Journal of Computational and Graphical Statistics*, **7,** 310-321.

[36] Kalbfleisch, J. D. and Prentice, R. L. (2002). *The statistical analysis of failure time data.* Second edition, John Wiley: New York.

[37] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457-481.

[38] Keiding, N. (1991). Age-specific incidence and prevalence: A statistical perspective (with discussion). *Journal of the Royal Statistical Society, Series A,* **154**, 371-412.

[39] Klein, J. P. and Moeschberger, M. L. (2003). *Survival analysis*, Springer-Verlag: New York.

[40] Kooperberg, C. and Clarkson, D. B. (1997). Hazard regression with interval-censored data. *Biometrics,* **53**, 1485-1494.

[41] Lawless, J. F. (2003). *Statistical models and methods for lifetime data*. John Wiley: New York.

[42] Lee E.T. (1980). *Statistical Methods for Survival Data Analysis,* Life Learning Publications: Belmont, California.

[43] Lindsey, J. K. (1998). A study of interval censoring in parametric regression models. *Lifetime Data Analysis*, **4**, 329-354.

[44] Odell, P. M. , Anderson, K. M. , and D' Agostino, R. B. (1992). Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics,* **48**, 951-959.

[45] Pan, W. (1999). A comparison of some two-sample tests with interval-censored data. *Journal of Nonparametric Statistics,* **12**, 133-146.

[46] Pepe, M. S. and Fleming, T. R. (1989). Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. *Biometrics* **45**, 497-507.

[47] Peto, R. (1973). Experimental survival curves for interval-censored data. *Applied Statistics*, **22**, 86-91.

[48] Rao, C. R. (1973). *Linear statistical inference and its applications.* 2Nd ed. John Wiley & Sons, New York.

[49] Robertson, T., Wright, F. T. and Dykstra, R. (1988). *Order restricted statistical inference.* John Wiley: New York.

[50] Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley: New York.

[51] Rucker, G. and Messerer, D. (1988). Remission duration: An example of interval censored observations. *Statistics in Medicine,* **7,** 1139-1145.

[52] Satten, G. A. (1996). Rank-based inference in the proportional hazards model for interval-censored data. *Biometrika*, **83**, 355-370.

[53] Schick, A. and Yu, Q. (2000). Consistency of the GMLE with mixed case interval- censored data. *Scandinavian Journal of Statistics*, **27**, 45-55.

[54] Shiboski, S. C. and Jeweel, N. P. (1992). Statistical analysis of the time dependence of HIV infectivity based on partner study data. *Journal of the American Statistical Assosiation,* **51**, 1384-1399.

[55] Sun, J. (1998). Interval censoring. *Encyclopedia of Biostatistics*, John Wiley, First Edition, 2090-2095.

[56] Sun, J. (2004). Statistical analysis of doubly interval-censored failure time data. Advances in survival analysis, *Handbook of Statistics,* **23**, 105-122.

[57] Sun, J. (2005). Interval censoring. *Encyclopedia of Biostatistics*, John Wiley, Second Edition, 2603-2609.

[58] Sun, J. and Kalbfleisch, J. D. (1993). The analysis of current status data on point processes. *Journal of the American Statistical Associaton*, **88**, 1449-1454.

[59] Sun, J. and Kalbfleisch, J. D. (1996). Nonparametric tests of tumor prevalence data. *Biometrics*, **52**, 726-731.

[60] Tang, M. X., Tsai, W. Y., Mander, K. and Mayeux, R. (1995). Linear rank tests for doubly censored data. *Statistics in Medicine*, **14**, 2555-2563.

[61] Tanner, M. A. (1991). *Tools for statistical inference: observed data and data augmentation methods.* New York : Springer-Verlag.

[62] Tanner, M. A. and Wong, W. H. (1987). The application of imputation to an estimation problem in grouped lifetime analysis. *Technometrics*, **29**, 23-32.

[63] Turnbull, B. W. (1976). The empirical distribution with arbitrarily grouped censored data and truncated data. *Journal of the Royal Statistical Society, Series B,* **38**, 290-295.

[64] Turnbull, B. W. (1978). A likelihood ratio statistic for testing goodness of fit with randomly censored data. *Biometrics*, **34**, 367-375.

[65] Wei, G. C. G. and Tanner, M.A. (1991). Application of multiple imputation to the analysis of censored regression data. *Biometrics*, **47**, 1297-1309.

[66] Wellner, J. A. (1995). Interval censoring case 2: alternative hypotheses. *Analysis of Censored Data (Pune, 1994/1995),* eds. H. L. Koul and J. V. Deshoande, IMS Lecture Notes, Monogragh Series **27**, 271-219.

[67] Wellner, J. A. and Zhan, Y. (1997). A hydrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data. *Journal of the American Statistical Assosiation,* **92**, 945-959.

[68] Yu, Q., Li, L. and Wong, G. Y. C. (2000). On consistency of the self-consistent estimator of survival functions with interval-censored data. *Scandinavian Journal of Statistics,* **27**, 35-44.

[69] Zhao, Q. and Sun, J. (2004). Generalized log-rank test for mixed interval-censored failure time data. *Statistics in Medicine*, **23**, 1621-1629.

[70] Καρώνη, Χ. (2009) *Μοντέλα Αξιοπιστίας και Επιβίωσης*. Εκδόσεις Συμεών: Αθήνα.