ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Διπλωματική Εργασία

Ανάλυση Επιβίωσης και Παλινδρόμηση Κατωφλιού

Ελένη Γιουρμετάκη

Επιβλέπουσα: **Καρώνη Χρυσηΐς**

Καθηγήτρια Ε.Μ.Π

Τριμελής επιτροπή:

Χ. Καρώνη, Χ. Κουκουβίνος, Φ. Βόντα

Καθηγήτρια Ε.Μ.Π, Καθηγητής Ε.Μ.Π, Αναπλ. καθηγήτρια Ε.Μ.Π

Αθήνα, Σεπτέμβριος 2017

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά την καθηγήτρια του Ε.Μ.Π, κα Χρυσηίδα Καρώνη, επιβλέπουσα της παρούσας εργασίας, για την βοήθεια της και για τη δυνατότητα που μου έδωσε να ασχοληθώ με ένα αντικείμενο που με ενδιαφέρει πολύ.

Ακόμη, θα ήθελα να ευχαριστήσω τους γονείς μου, την αδερφή μου και τους φίλους μου για την συμπαράσταση και τη στήριξή τους όλους αυτούς τους μήνες της προετοιμασίας.

Ελένη Γιουρμετάκη

Αθήνα, Σεπτέμβριος 2017

Περίληψη

Η ανάλυση επιβίωσης είναι κλάδος της Στατιστικής που ασχολείται με τη μελέτη των δεδομένων διάρκειας ζωής, δηλαδή τη μελέτη του χρόνου μέχρις ότου συμβεί κάποιο γεγονός. Το γεγονός που μελετάμε μπορεί να είναι κάτι ανεπιθύμητο όπως η υποτροπή ενός ασθενή αλλά και πιο σπάνια κάτι επιθυμητό όπως για παράδειγμα η απαλλαγή από τον πόνο. Η ανάλυση επιβίωσης έχει πολλές εφαρμογές σε βιοϊατρικές αλλά και τεχνολογικές επιστήμες.

Σκοπός της παρούσας εργασίας είναι η παρουσίαση των βασικών στοιχείων της ανάλυσης επιβίωσης και η πρακτική τους εφαρμογή σε συγκεκριμένα δεδομένα με τη βοήθεια της R και του MINITAB. Αναλυτικότερα, το Κεφάλαιο 1 αναφέρεται στις βασικές έννοιες της ανάλυσης επιβίωσης, δηλαδή στις συναρτήσεις επιβίωσης και διακινδύνευσης και τις εκτιμήτριές τους, στα παραμετρικά μοντέλα παλινδρόμησης και στο μοντέλο του Cox. Το Κεφάλαιο 2 αναφέρεται στο μοντέλο πρώτης μετάβασης και την παλινδρόμηση κατωφλιού (threshold regression) που είναι μια καινούργια μέθοδος και χρησιμοποιείται όλο και πιο συχνά ως εναλλακτικό μοντέλο του μοντέλου αναλογικής διακινδύνευσης του Cox, όταν η υπόθεση της αναλογικής διακινδύνευσης δεν ισχύει. Επίσης, στο Κεφάλαιο 2 γίνεται η σύγκριση των δύο μοντέλων και παρουσιάζονται μερικά παραδείγματα του μοντέλου του threshold που έχει χρησιμοποιηθεί σε πρόσφατες έρευνες. Τέλος, στο Κεφάλαιο 3 παρουσιάζεται η στατιστική ανάλυση των δεδομένων 137 ασθενών που πάσχουν από λευχαιμία και έχουν υποβληθεί σε μεταμόσχευση μυελού των οστών και εξετάζεται η επιβίωσή τους μετά τη μεταμόσχευση.

Abstract

Survival analysis is a branch of Statistics that studies lifetime data, in other words it studies the time until an event occurs. The event under study can be something undesirable, such as the relapse of a patient, or more rarely something desirable, for instance the relief of pain. Survival analysis has a lot of applications in biomedical and technological sciences.

The purpose of this thesis is the presentation of the special features of survival analysis and their application to specific data with the help of R and MINITAB. More specifically, Chapter 1 refers to the basic concepts of survival analysis, including the survival and hazard functions and their estimators, the parametric regression model and the Cox regression model. Chapter 2 deals with the first-hitting-time model and threshold regression. Threshold regression is a new method which is widely used as an alternative to Cox regression when the proportional hazards assumption does not hold. Furthermore, a comparison between Cox and threshold regression is presented in Chapter 2 along with examples from recent studies on threshold regression. Finally, in Chapter 3 we study the data of 137 leukemia patients who have undergone bone marrow transplantation and we examine their survival after the transplantation using the methods discussed in the previous chapters.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1

Βασικές Έννοιες της Ανάλυσης Επιβίωσης

Η ανάλυση επιβίωσης ή ανάλυση αξιοπιστίας είναι συγκεκριμένη στατιστική θεωρία που ασχολείται με δεδομένα διάρκειας ζωής. Τα δεδομένα διάρκειας ζωής είναι αυτά στα οποία μελετάμε το χρόνο μέχρις ότου συμβεί κάποιο γεγονός. Το γεγονός αυτό συνήθως είναι δυσάρεστο, δηλαδή θάνατος ή υποτροπή ενός ασθενή, μηχανική βλάβη, χρεωκοπία μιας εταιρείας κ.ά. Μπορεί, όμως, το γεγονός να μην είναι τόσο δυσάρεστο όπως για παράδειγμα η αποθεραπεία ενός ασθενή.

Η ανάλυση αυτή έχει πολλές εφαρμογές σε διάφορους επιστημονικούς χώρους, όπως στην ιατρική και τη μηχανική. Ο όρος ανάλυση επιβίωσης χρησιμοποιείται κυρίως σε ιατρικές εφαρμογές ενώ ο όρος ανάλυση αξιοπιστίας χρησιμοποιείται σε εφαρμογές των θετικών επιστημών. Η μελέτη του χρόνου μέχρι το θάνατο του ασθενή μετά από μία συγκεκριμένη θεραπεία και η σύγκριση των αποτελεσμάτων μεταξύ δύο ή και παραπάνω μεθόδων θεραπείας σε ασθενείς που πάσχουν από την ίδια ασθένεια είναι δύο χαρακτηριστικά παραδείγματα ανάλυσης επιβίωσης. Αντίθετα, ο χρόνος μέχρι τη βλάβη μιας μηχανής είναι ένα παράδειγμα ανάλυσης αξιοπιστίας (Καρώνη, 2009).

1.1 Αποκομμένα δεδομένα

Για τα δεδομένα διάρκειας ζωής δεν ακολουθούμε τις κλασικές μεθόδους που χρησιμοποιούμε γενικά στην ανάλυση δεδομένων. Αυτό συμβαίνει γιατί τα δεδομένα διάρκειας ζωής δεν είναι συμμετρικά κατανεμημένα. Πιο συγκεκριμένα, συνήθως η κατανομή που ακολουθούν αυτά τα δεδομένα έχει θετική ασυμμετρία, δηλαδή οι περισσότερες παρατηρήσεις βρίσκονται δεξιά της κορυφής του ιστογράμματος. Επομένως, δεν μπορούμε να υποθέσουμε ότι τα δεδομένα μας ακολουθούν την κανονική κατανομή που είναι συμμετρική. Για να λύσουμε, λοιπόν, αυτό το πρόβλημα προσαρμόζουμε ένα εναλλακτικό μοντέλο που είναι κατάλληλο για τα δεδομένα διάρκειας ζωής.

Ένα ακόμη χαρακτηριστικό των δεδομένων διάρκειας ζωής που τα διαφοροποιεί από τα υπόλοιπα είναι η αποκοπή. Όπως αναφέραμε παραπάνω, στην ανάλυση επιβίωσης ή αξιοπιστίας μελετάμε το χρόνο μέχρις ότου συμβεί ένα γεγονός. Πολλές φορές, όταν τερματίζουμε το πείραμά μας, υπάρχει περίπτωση σε κάποιες μονάδες να μην έχει συμβεί ακόμη το γεγονός που μελετάμε. Για παράδειγμα, έστω ότι στο πείραμα μελετάμε το χρόνο που λειτουργούν οι μηχανές ενός εργοστασίου μέχρι να υποστούν βλάβη. Την χρονική στιγμή που σταματάμε το πείραμα, έστω στιγμή c, υπάρχει περίπτωση μερικές μηχανές να μην έχουν υποστεί βλάβη ακόμη. Σ' αυτή την περίπτωση δε γνωρίζουμε την ακριβή χρονική στιγμή που οι μηχανές αυτές σταματούν να λειτουργούν, όμως γνωρίζουμε ότι τη στιγμή c λειτουργούσαν ακόμη. Τότε έχουμε δεξιά αποκομμένες παρατηρήσεις (right censored observations). Η περίπτωση της δεξιάς αποκοπής δεδομένων παρουσιάζεται αρκετά συχνά σε εφαρμογές και αυτήν θα χρησιμοποιήσουμε και εμείς στη συνέχεια.

Αντίθετα, μία πιο σπάνια περίπτωση είναι οι αριστερά αποκομμένες παρατηρήσεις (left censored observations). Σ' αυτή τη μορφή αποκοπής τερματίζουμε το πείραμά μας μία χρονική στιγμή c, χωρίς όμως να έχουμε υπό συνεχή επίβλεψη το πείραμα. Δηλαδή, στο παράδειγμα με τις μηχανές, τη στιγμή c διαπιστώνουμε ότι κάποιες μηχανές έχουν υποστεί βλάβη χωρίς, όμως να γνωρίζουμε την ακριβή χρονική στιγμή που σταμάτησαν να λειτουργούν.

Τέλος, υπάρχει και η περίπτωση της αποκοπής σε διάστημα (interval censoring). Πάλι στο παράδειγμα με τις μηχανές, η αποκοπή σε διάστημα είναι όταν γνωρίζουμε ότι κάποιες μηχανές υπέστησαν βλάβη στο χρονικό διάστημα (c_1, c_2) αλλά δεν μπορούμε να ξέρουμε την ακριβή χρονική στιγμή που έγινε η βλάβη.

1.2 Βασικές συναρτήσεις

1.2.1 Συνάρτηση επιβίωσης ή αξιοπιστίας

Έστω μία συνεχής τυχαία μεταβλητή (τ.μ) T>0 η οποία εκφράζει τη διάρκεια ζωής για τη μονάδα που μελετάμε. Έστω, επίσης, ότι η τ.μ T ακολουθεί μία κατανομή με συνάρτηση πυκνότητας πιθανότητας (σ.π.π) f(t), $t \geq 0$ και συνάρτηση κατανομής

(σ.κ) $F(t)$, η οποία ως γνωστόν ορίζεται ως $F(t) = P[T \leq t] = \int_0^t f(u)du$. Η συνάρτηση επιβίωσης ή αξιοπιστίας ορίζεται ως:

$$S(t) = 1 - F(t) = P[T > t] = \int_t^\infty f(u)du \qquad (1.1)$$

και αναπαριστά την πιθανότητα η διάρκεια ζωής να είναι μεγαλύτερη του χρόνου t. Επειδή ισχύει η σχέση $f(t) = \frac{dF(t)}{dt}$ έχουμε ότι:

$$f(t) = -\frac{dS(t)}{dt} \qquad (1.2)$$

1.2.2 Συνάρτηση διακινδύνευσης

Η συνάρτηση διακινδύνευσης συμβολίζεται ως h(t) και εκφράζει τον κίνδυνο που διατρέχει η μονάδα που μελετάμε να της συμβεί το γεγονός το χρονικό διάστημα (t, t+δt), δεδομένου ότι το γεγονός δεν έχει συμβεί μέχρι τη χρονική στιγμή t. Η συνάρτηση διακινδύνευσης είναι:

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t+\delta t \mid T \geq t}{\delta t} \right\} \qquad (1.3)$$

Από τις πιθανότητες γνωρίζουμε ότι $P(A|B) = \frac{P(AB)}{P(B)}$, άρα ο αριθμητής της σχέσης (1.3) γίνεται: $\frac{P(t \leq T < t+\delta t)}{P(T \geq t)} = \frac{F(t+\delta t) - F(t)}{S(t)}$, όπου F(t) η σ.κ της τ.μ T και S(t) η συνάρτηση επιβίωσης. Συνεπώς, η σχέση (1.3) γίνεται:

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{F(t+\delta t) - F(t)}{\delta t} \right\} \frac{1}{S(t)}$$

Τώρα, επειδή $\lim_{\delta t \to 0} \left\{ \frac{F(t+\delta t) - F(t)}{\delta t} \right\} = \frac{dF(t)}{dt} = f(t)$ έχουμε τη σχέση:

$$h(t) = \frac{f(t)}{S(t)} \qquad (1.4)$$

Η σχέση (1.4) μας δείχνει τον τρόπο που συνδέονται η συνάρτηση επιβίωσης με τη συνάρτηση διακινδύνευσης.

1.2.3 Σωρευτική συνάρτηση διακινδύνευσης

Η σωρευτική συνάρτηση διακινδύνευσης συμβολίζεται ως H(t) και ορίζεται ως:

$$H(t) = \int_0^t h(u)du \qquad (1.5)$$

Η σωρευτική συνάρτηση διακινδύνευσης μας βοηθά να επιλέξουμε το κατάλληλο στατιστικό μοντέλο κατά την ανάλυση των δεδομένων. Από τις σχέσεις (1.2), (1.4) και (1.5) έχουμε ότι:

$$H(t) = \int_0^t \frac{f(u)}{S(u)}du$$

$$= \int_0^t \frac{-\frac{dS(u)}{du}}{S(u)}du$$

$$= [-\ln S(u)]_0^t$$

$$\Rightarrow H(t) = -\ln S(t) \qquad (1.6)$$

Συνεπώς, η συνάρτηση επιβίωσης μπορεί να προκύψει από τη συνάρτηση σωρευτικής διακινδύνευσης από τη σχέση

$$S(t) = \exp(-H(t))$$

1.3 Μη-παραμετρική ανάλυση

Ένα σημαντικό βήμα στην ανάλυση των δεδομένων διάρκειας ζωής είναι η εύρεση του μοντέλου που προσαρμόζεται κατάλληλα σε αυτά. Για να επιλέξουμε το κατάλληλο μοντέλο μελετάμε τη συμπεριφορά των συναρτήσεων επιβίωσης και διακινδύνευσης. Σε αυτό βοηθούν οι εκτιμήτριες των συναρτήσεων που θα δούμε στις επόμενες παραγράφους. Οι μέθοδοι που θα χρησιμοποιήσουμε για να εκτιμήσουμε τις συναρτήσεις επιβίωσης και διακινδύνευσης ονομάζονται μη-παραμετρικές καθώς δεν χρειάζεται να γνωρίζουμε την κατανομή που ακολουθούν τα δεδομένα μας.

1.3.1 Εκτίμηση της συνάρτησης επιβίωσης

Έστω ότι έχουμε ένα διατεταγμένο τυχαίο δείγμα με μη-αποκομμένες παρατηρήσεις, $t_{(1)} \leq t_{(2)} \leq \ldots \leq t_{(n)}$, για την τυχαία μεταβλητή T. Από τον τύπο (1.1) έχουμε $S(t) = P(T > t) = 1 - P(T \leq t)$ και άρα για το σημείο $t_{(j)}$ λαμβάνουμε την εκτιμήτρια για την συνάρτηση επιβίωσης $\widehat{S}\left(t_{(j)}\right) = 1 - p_j$, όπου p_j η σχετική συχνότητα του γεγονότος ($T \leq t_{(j)}$) στο δείγμα. Υπάρχουν αρκετές εκτιμήτριες για την p_j, γενικά οι εκτιμήτριες αυτές είναι της μορφής $p_j = \frac{j-a}{n-2a+1}$, $j = 1, \ldots, n$ με διάφορες τιμές του α. Μία εκτιμήτρια που χρησιμοποιείται συχνά λέγεται τύπος του Hazen και είναι η $p_j = \frac{j-0.5}{n}$ (α=0.5). Η πιο απλή εκτιμήτρια είναι η $p_j = \frac{j}{n}$ που δεν είναι της μορφής $p_j = \frac{j-a}{n-2a+1}$.

Η εκτιμήτρια της συνάρτησης επιβίωσης, $\widehat{S}(t)$, θεωρείται σταθερή μεταξύ δύο γειτονικών χρονικών στιγμών, επομένως το διάγραμμα της $\widehat{S}(t)$ έναντι του χρόνου μας δείχνει ότι η $\widehat{S}(t)$ είναι κλιμακωτή ή βαθμωτή συνάρτηση.

1.3.2 Η Kaplan-Meier εκτιμήτρια

Η μέθοδος που περιγράψαμε στην προηγούμενη παράγραφο για την εκτίμηση της συνάρτησης επιβίωσης δεν χρησιμοποιείται για αποκομμένες παρατηρήσεις. Παρόλα αυτά, στην πραγματικότητα οι αποκομμένες παρατηρήσεις εμφανίζονται αρκετά συχνά, και πιο συγκεκριμένα οι δεξιά αποκομμένες, γι' αυτό και χρησιμοποιούμε την εκτιμήτρια Kaplan-Meier σε αυτές τις περιπτώσεις (Kaplan & Meier, 1958).

Έστω ότι έχουμε τυχαίο δείγμα n μονάδων με αποκομμένες παρατηρήσεις που συμβαίνουν σε μερικές από τις διακεκριμένες χρονικές στιγμές $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$, $k \leq n$. Θεωρούμε ότι ακριβώς τη χρονική στιγμή $t_{(j)}$ καταστρέφονται d_j μονάδες, ενώ λίγο πριν από αυτή τη χρονική στιγμή λειτουργούσαν n_j μονάδες. Συνήθως, $d_j = 1$.

Ο αριθμός n_j περιλαμβάνει και τις μονάδες που πρόκειται να καταστραφούν τη χρονική στιγμή $t_{(j)}$, ενώ δεν περιλαμβάνει αυτές που έχουν ήδη καταστραφεί και τις ήδη αποκομμένες. Η συνάρτηση επιβίωσης είναι

$$S(t_{(j)}) = P(T > t_{(j)}) = P(T > t_{(1)})P(T > t_{(2)} | T > t_{(1)}) \dots P(T > t_{(j)} | T > t_{(j-1)})$$

Μία εκτιμήτρια της $P(T > t_{(1)})$ είναι

$$\hat{P}(T > t_{(1)}) = 1 - p_1 = 1 - \frac{d_1}{n_1} = \frac{n_1 - d_1}{n_1}$$

όπου p_1 η σχετική συχνότητα των κατεστραμμένων μονάδων στο διάστημα $(0, t_{(1)}]$. Δουλεύοντας με τον ίδιο τρόπο, για τις άλλες εκτιμήτριες έχουμε

$$\hat{P}(T > t_{(2)} | T > t_{(1)}) = \frac{n_2 - d_2}{n_2} \quad \text{κ.ο.κ.}$$

Άρα τελικά η εκτιμήτρια Kaplan-Meier δίνεται από τον τύπο

$$\hat{S}(t) = \begin{cases} \prod_{j:\, t_j \le t} \frac{n_j - d_j}{n_j}, & \text{όταν } t \ge t_{(1)} \\ 1, & \text{όταν } t < t_{(1)} \end{cases}$$

Το διάγραμμα της εκτιμήτριας Kaplan-Meier συναρτήσει του χρόνου μας δίνει και πάλι μία κλιμακωτή συνάρτηση, όπου η $\widehat{S}(t)$ θεωρείται σταθερή μεταξύ δύο γειτονικών χρονικών στιγμών και μειώνεται συνεχώς.

Η εκτιμήτρια της διασποράς της εκτιμήτριας Kaplan-Meier δίνεται από τον τύπο του Greenwood:

$$\widehat{V}\left(\hat{S}(t)\right) = \hat{S}(t)^2 \sum_{t_{(j)} \le t} \frac{d_j}{n_j(n_j - d_j)}$$

και άρα το τυπικό σφάλμα είναι: $se\left(\hat{S}(t)\right) = \left\{\widehat{V}\left(\hat{S}(t)\right)\right\}^{1/2}$.

1.3.3 Η Nelson-Aalen εκτιμήτρια της σωρευτικής συνάρτησης διακινδύνευσης

Μία εκτίμηση για τη σωρευτική συνάρτηση διακινδύνευσης είναι με τη βοήθεια της $\widehat{S}(t)$ από τον τύπο (1.6). Παρόλα αυτά, συνήθως προτιμούμε την εκτιμήτρια

Nelson-Aalen (Aalen, 1978) (Nelson, 1972) που προκύπτει από τη σχέση (1.6) ως εξής:

$$\hat{H}(t) = -\ln \hat{S}(t) = -\sum_{j:\, t_{(j)} \le t} \ln(1 - \frac{d_j}{n_j}) \cong \sum_{j:\, t_{(j)} \le t} \frac{d_j}{n_j}$$

όπου n_j και d_j όπως τα ορίσαμε στην παράγραφο 1.3.2. Συνεπώς, η Nelson-Aalen εκτιμήτρια δίνεται από το τύπο

$$\hat{H}(t) = \begin{cases} \sum_{j:\, t_{(j)} \le t} \frac{d_j}{n_j} & ,\text{όταν } t \ge t_{(1)} \\ 0 & ,\text{ όταν } t < t_{(1)} \end{cases}$$

και είναι κι αυτή μία κλιμακωτή συνάρτηση.

Η εκτιμήτρια της διασποράς της εκτιμήτριας Nelson-Aalen δίνεται από τον τύπο:

$$\hat{V}(\hat{H}) = \sum_{j:\, t_{(j)} \le t} \frac{d_j}{{n_j}^2}$$

και άρα το τυπικό σφάλμα είναι $se(\hat{H}) = \{\hat{V}(\hat{H})\}^{1/2}$.

1.3.4 Έλεγχος Log-rank

Ο έλεγχος Log-rank είναι μη-παραμετρικός και τον χρησιμοποιούμε για να συγκρίνουμε δύο ή και παραπάνω ομάδες στις οποίες έχουμε χωρίσει τα δεδομένα μας. Λέγεται μη-παραμετρικός γιατί δεν γνωρίζουμε μαθηματικά τις συναρτήσεις επιβίωσης των ομάδων των δεδομένων μας. Έστω ότι έχουμε $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$ διακεκριμένες χρονικές στιγμές και έχουμε χωρίσει τα δεδομένα σε δύο ομάδες (ομάδα 1 και 2). Σε αυτές τις χρονικές στιγμές παύουν να λειτουργούν μονάδες που προέρχονται και από τις δύο ομάδες. Θεωρούμε ότι αμέσως πριν τη χρονική στιγμή $t_{(j)}$ για την ομάδα 1 υπάρχουν n_{1j} μονάδες σε κίνδυνο από τις οποίες παύουν να λειτουργούν d_{1j} μονάδες ακριβώς τη στιγμή $t_{(j)}$. Αντίθετα, για την ομάδα 2 αμέσως πριν τη χρονική στιγμή $t_{(j)}$ υπάρχουν n_{2j} μονάδες σε κίνδυνο από τις οποίες παύουν να λειτουργούν d_{2j} μονάδες ακριβώς τη στιγμή $t_{(j)}$. Επομένως, τη χρονική στιγμή $t_{(j)}$ παύουν να λειτουργούν συνολικά $d_j = d_{1j} + d_{2j}$ μονάδες από τις $n_j = n_{1j} + n_{2j}$

που ήταν σε κίνδυνο. Η κατάσταση που περιγράψαμε φαίνεται αναλυτικά στον Πίνακα 1.1.

Πίνακας 1.1: Πίνακας συνάφειας για τη χρονική στιγμή $t_{(j)}$

	Διακοπή λειτουργίας		
Ομάδα	ΝΑΙ	ΟΧΙ	Σε κίνδυνο πριν την στιγμή $t_{(j)}$
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Σύνολο	d_j	$n_j - d_j$	n_j

Ο έλεγχος Log-rank που πραγματοποιούμε είναι : H_0: S_1=S_2 vs H_1: $S_1 \neq S_2$, όπου S_1, S_2 οι συναρτήσεις επιβίωσης των ομάδων 1, 2 αντίστοιχα. Δηλαδή ελέγχουμε αν υπάρχουν διαφοροποιήσεις μεταξύ των δύο ομάδων όσον αφορά την επιβίωση των μονάδων που ανήκουν στις ομάδες αυτές. Αποδεικνύεται ότι η ελεγχοσυνάρτηση Log-rank δίνεται από τον τύπο

$$\frac{u^2}{v} = \frac{\left[\sum_j \left\{ d_{1j} - \left(\frac{n_{1j} d_j}{n_j}\right)\right\}\right]^2}{\sum_j \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{{n_j}^2 (n_j - 1)}}$$

και ακολουθεί την ${X_1}^2$ κατανομή ασυμπτωτικά, υπό την μηδενική υπόθεση H_0. Αν η p-τιμή του ελέγχου είναι αρκετά μικρή (<0.05) απορρίπτουμε την μηδενική υπόθεση και άρα θεωρούμε ότι υπάρχουν διαφοροποιήσεις μεταξύ των δύο ομάδων. Σε αντίθετη περίπτωση, δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση ότι οι συναρτήσεις επιβίωσης των δύο ομάδων είναι ίσες.

Αξίζει να αναφέρουμε ότι ο έλεγχος Log-rank μπορεί να επεκταθεί και σε άλλες μορφές. Πιο συγκεκριμένα, η ελεγχοσυνάρτηση Log-rank γενικεύεται ως $\frac{(\sum w_j u_j)^2}{\sum {w_j}^2 v_j}$, όπου w_j συντελεστές στάθμισης. Για τον έλεγχο Log-rank έχουμε $w_j = 1$, ενώ αν επιλέξουμε $w_j = n_j$ έχουμε την ελεγχοσυνάρτηση Wilcoxon. Δηλαδή, για τον έλεγχο

Wilcoxon ο συντελεστής στάθμισης είναι ο αριθμός των μονάδων που βρίσκονται σε κίνδυνο πριν από τη στιγμή $t_{(j)}$.

Όσον αφορά τη σύγκριση μεταξύ των ελέγχων Log-rank και Wilcoxon, ο έλεγχος Log-rank είναι καταλληλότερος όταν ισχύει η υπόθεση της αναλογικής διακινδύνευσης στις δύο ομάδες που θέλουμε να ελέγξουμε αν έχουν διαφοροποιήσεις στις συναρτήσεις επιβίωσής τους. Αντίθετα, ο έλεγχος Wilcoxon δίνει μεγαλύτερη σημασία στις διακοπές που προκύπτουν νωρίς στο πείραμα σε σύγκριση με αυτές που θα συμβούν αργότερα γι' αυτό και είναι ισχυρότερος του Log-rank όταν θέλουμε να εντοπίσουμε τις διαφοροποιήσεις στις συναρτήσεις επιβίωσης που εμφανίζονται νωρίς (Collett, 2003).

1.4 Παραμετρικά μοντέλα παλινδρόμησης

Η γραμμική παλινδρόμηση είναι το βασικό μοντέλο της στατιστικής στο οποίο η τιμή της εξαρτημένης μεταβλητής, Y, συνδέεται γραμμικά με τις συμμεταβλητές x_i, όπως φαίνεται στο τύπο

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon = \boldsymbol{\beta}'\boldsymbol{x} + \varepsilon \qquad (1.7)$$

όπου ε τα σφάλματα διαφορετικών παρατηρήσεων τα οποία είναι ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν την κατανομή $N(0,\sigma^2)$ με παράμετρο θέσης 0 και κλίμακας 1. Οι συμμεταβλητές x_i θεωρούνται μη-στοχαστικές και άρα $Y \sim N(\mu, \sigma^2)$, όπου $\mu = \mu(\boldsymbol{x}) = \boldsymbol{\beta}'\boldsymbol{x}$ η μέση τιμή και σ^2 η διασπορά.

Τώρα για τα μοντέλα διάρκειας ζωής μας ενδιαφέρουν τα μοντέλα παλινδρόμησης για δεδομένα διάρκειας ζωής. Πιο συγκεκριμένα θα ασχοληθούμε με δύο βασικά μοντέλα παλινδρόμησης, το μοντέλο επιταχυνόμενης διακοπής (accelerated failure model) και το μοντέλο αναλογικής διακινδύνευσης (proportional hazards model).

1. Το **μοντέλο επιταχυνόμενης διακοπής** προκύπτει ως εξής:

Στο κλασικό μοντέλο παλινδρόμησης (1.7) παρατηρούμε ότι η εξαρτημένη μεταβλητή y δεν μπορεί να είναι πάντα θετική όπως απαιτούν τα δεδομένα διάρκειας

ζωής. Γι' αυτό το λόγο χρησιμοποιούμε την συνάρτηση $\ln T$ σαν εξαρτημένη μεταβλητή και έτσι έχουμε το γραμμικό μοντέλο

$$\ln T_{\boldsymbol{x}} = \mu_{\boldsymbol{x}} + \sigma\varepsilon = \mu_0 + \boldsymbol{\beta}'\boldsymbol{x} + \sigma\varepsilon$$

όπου μ_0 και σ είναι οι παράμετροι θέσης και κλίμακας αντίστοιχα και ε είναι μία τυχαία μεταβλητή.

Η συνάρτηση επιβίωσης γράφεται:

$$S(t;\boldsymbol{x}) = P(T_{\boldsymbol{x}} > t) = P(\ln T_{\boldsymbol{x}} > \ln t) = P(\mu_0 + \boldsymbol{\beta}'\boldsymbol{x} + \sigma\varepsilon > \ln t)$$
$$= P(\ln T_0 + \boldsymbol{\beta}'\boldsymbol{x} > \ln t) = P\big(T_0 > texp(-\boldsymbol{\beta}'\boldsymbol{x})\big) = S_0(texp(-\boldsymbol{\beta}'\boldsymbol{x}))$$

Συνεπώς, έχουμε

$$S(t;\boldsymbol{x}) = S_0(tg(\boldsymbol{x})) \qquad (1.8)$$

όπου $g(\boldsymbol{x})$ μια θετική συνάρτηση των συμμεταβλητών και S_0 μια βασική συνάρτηση επιβίωσης. Η σχέση (1.8) περιγράφει το μοντέλο της επιταχυνόμενης διακοπής και έχει εφαρμογές σε προβλήματα στα οποία η αλλαγή της τιμής μιας συμμεταβλητής έχει σαν αποτέλεσμα την επίσπευση της διακοπής της λειτουργίας των πειραματικών μονάδων.

Πριν την προσαρμογή του μοντέλου μας είναι σημαντικό να πραγματοποιήσουμε γραφικό έλεγχο για να δούμε αν ισχύει η υπόθεση της επιταχυνόμενης διακοπής. Από τη σχέση (1.8) έχουμε

$$S(t;\boldsymbol{x}) = S_0\big(tg(\boldsymbol{x})\big) = P\big(T_0 > tg(\boldsymbol{x})\big) = P(\ln T_0 > \ln t + \ln g(\boldsymbol{x}))$$
$$S(t;\boldsymbol{x}) = S^*(y + \ln g(\boldsymbol{x}))$$

με $y = \ln t$ και S^* η συνάρτηση επιβίωσης της τυχαίας μεταβλητής $Y = \ln T_0$. Επομένως, για να ισχύει η υπόθεση της επιταχυνόμενης διακοπής πρέπει το γράφημα της $S(t;\boldsymbol{x})$ ως προς $\ln t$ για συγκεκριμένο διάνυσμα συμμεταβλητών $\boldsymbol{x}$ να είναι οριζόντια μετατόπιση της S^* ως προς $\ln t$. Δηλαδή πρέπει να βρούμε τις εκτιμήτριες Kaplan-Meier για κάθε ομάδα και να κάνουμε το γράφημα αυτών ως προς $\ln t$. Οι καμπύλες που θα προκύψουν πρέπει να είναι παράλληλες μεταξύ τους.

2. Το **μοντέλο αναλογικής διακινδύνευσης** είναι εκείνο για το οποίο ο λόγος των συναρτήσεων διακινδύνευσης για δύο διανύσματα συμμεταβλητών $\mathbf{x_1}$, $\mathbf{x_2}$ είναι σταθερός και ανεξάρτητος του χρόνου t. Δηλαδή, $\frac{h(t\,;\,\boldsymbol{x_1})}{h(t\,;\,\boldsymbol{x_2})} = \lambda$ = σταθερό. Το μοντέλο αναλογικής διακινδύνευσης ορίζεται ως:

$$h(t;\boldsymbol{x}) = g(\boldsymbol{x})h_0(t)$$

με $h_0(t)$ μια βασική συνάρτηση διακινδύνευσης και $g(\boldsymbol{x})$>0. Συνήθως $g(\boldsymbol{x}) = e^{\boldsymbol{\beta' x}}$.

Ανάλογα με το αν η βασική συνάρτηση διακινδύνευσης είναι γνωστή, δηλαδή γνωστής κατανομής, χωρίζουμε το μοντέλο αναλογικής διακινδύνευσης σε δύο κατηγορίες:

α. Παραμετρικό μοντέλο παλινδρόμησης, στην περίπτωση που η h_0 είναι συγκεκριμένης γνωστής κατανομής

β. Ημι-παραμετρικό μοντέλο παλινδρόμησης, στην περίπτωση που η h_0 είναι ακαθόριστη. Σε αυτή την κατηγορία ανήκει και μοντέλο του Cox.

1.4.1 Υπόλοιπα Cox-Snell

Ένας τρόπος να ελέγξουμε αν το μοντέλο μας είναι το κατάλληλο είναι να μελετήσουμε τα υπόλοιπα μετά την προσαρμογή του μοντέλου. Στην κλασική περίπτωση της γραμμικής παλινδρόμησης τα υπόλοιπα υπολογίζονται από τη σχέση

$$\widehat{\varepsilon_i} = y_i - \widehat{y_i} = y_i - \widehat{\boldsymbol{\beta}}'\boldsymbol{x_i}$$

και εκφράζουν τη διαφορά μεταξύ της παρατηρούμενης τιμής, y_i ,και της προσαρμοσμένης, $\widehat{y_i}$.

Τα υπόλοιπα που χρησιμοποιούνται για τα δεδομένα διάρκειας ζωής είναι τα Cox-Snell (Cox & Snell, 1968) και ορίζονται ως:

$$-\ln \hat{S}\,(t_i\,;\,\boldsymbol{x_i}) = \widehat{H}(t_i\,;\,\boldsymbol{x_i}) = \widehat{\varepsilon_i}$$

όπου $\hat{S}$, $\widehat{H}$ οι εκτιμήτριες των συναρτήσεων επιβίωσης και σωρευτικής διακινδύνευσης αντίστοιχα. Τα υπόλοιπα αυτά είναι πολύ χρήσιμα καθώς μπορούν να

εφαρμοστούν σε οποιασδήποτε μορφής μοντέλα με αλλά και χωρίς συμμεταβλητές. Αυτό που μας ενδιαφέρει για να είναι κατάλληλο το μοντέλο μας είναι να εξετάσουμε γραφικά αν οι τιμές των υπολοίπων ακολουθούν την εκθετική κατανομή με παράμετρο τη μονάδα. Αν ισχύει αυτό τότε το μοντέλο μας σωστά προσαρμόστηκε στα δεδομένα.

Σε περίπτωση που έχουμε από δεξιά αποκομμένες παρατηρήσεις τότε έχουμε τα διορθωμένα Cox-Snell υπόλοιπα που δίνονται από τον τύπο

$$1 - \ln \hat{S}\,(t_i\,;\,\boldsymbol{x_i}) = \widehat{\varepsilon_i}$$

1.4.2 Το μοντέλο του Cox

Το μοντέλο του Cox είναι ένα μοντέλο αναλογικής διακινδύνευσης που ανήκει στην κατηγορία των ημι-παραμετρικών μοντέλων. Ορίζεται ως εξής:

$$h(t;\boldsymbol{x}) = h_0(t)e^{\boldsymbol{\beta' x}} \qquad (1.9)$$

όπου h_0 είναι μία βασική συνάρτηση διακινδύνευσης, η οποία είναι ακαθόριστη γι' αυτό και το μοντέλο είναι ημι-παραμετρικό, το **β** είναι ένα διάνυσμα συντελεστών που αντιστοιχούν σε καθεμιά των συμμεταβλητών (Cox, 1972). Για τον προσδιορισμό της συνάρτησης επιβίωσης του μοντέλου του Cox ακολουθούμε τα παρακάτω βήματα:

Γνωρίζουμε ότι η σωρευτική συνάρτηση διακινδύνευσης συνδέεται με τη συνάρτηση διακινδύνευσης μέσω του τύπου $H(t) = \int_0^t h(u)du$, οπότε με τη βοήθεια της σχέσης (1.9) έχουμε ότι

$$H(t;\boldsymbol{x}) = \int_0^t h_0\,(t)e^{\boldsymbol{\beta' x}}du = H_0(t)e^{\boldsymbol{\beta' x}}$$

με H_0 μία βασική σωρευτική συνάρτηση διακινδύνευσης που αντιστοιχεί στην h_0. Η συνάρτηση επιβίωσης δίνεται από τον τύπο $S(t;\boldsymbol{x}) = \exp(-H(t;\boldsymbol{x}))$, επομένως

$$S(t;\boldsymbol{x}) = \exp\{-H_0(t)e^{\boldsymbol{\beta' x}}\}$$

$$= \{S_0(t)\}^{e^{\boldsymbol{\beta' x}}}$$

όπου S_0 μία βασική συνάρτηση επιβίωσης.

Στη συνέχεια μας ενδιαφέρει να εκτιμήσουμε τις παραμέτρους του μοντέλου του Cox με τη μέθοδο μεγίστης πιθανοφάνειας. Έστω ότι τις διακεκριμένες χρονικές στιγμές $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$ διακόπτεται η λειτουργία k μονάδων. Δηλαδή, τη χρονική στιγμή $t_{(j)}$ σταματά να λειτουργεί μία μονάδα με συμμεταβλητές $\boldsymbol{x}_j$. Με R_j συμβολίζουμε το σύνολο των μονάδων που βρίσκονται σε κίνδυνο αμέσως πριν τη χρονική στιγμή $t_{(j)}$. Γνωρίζουμε από τη θεωρία πιθανοτήτων ότι η πιθανότητα να διακοπεί η λειτουργία μίας μονάδας j, δοθέντος ότι σταματά να λειτουργεί μία οποιαδήποτε μονάδα από το σύνολο R_j, δίνεται από το τύπο:

$$\frac{h(t_{(j)}\,;\,x_{(j)})}{\sum_{i\in R_j} h(t_{(j)}\,;\,x_{(i)})} = \frac{e^{\boldsymbol{\beta}\prime \boldsymbol{x_j}}}{\sum_{i\in R_j} e^{\boldsymbol{\beta}\prime \boldsymbol{x_i}}}\,.$$

Επομένως, ο Cox (1972) έδειξε ότι η συνάρτηση πιθανοφάνειας για το σύνολο των δεδομένων υπολογίζεται από τον τύπο

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{k} \left\{ \frac{e^{\boldsymbol{\beta}\prime \boldsymbol{x_j}}}{\sum_{i\in R_j} e^{\boldsymbol{\beta}\prime \boldsymbol{x_i}}}\,. \right\}$$

Έπειτα, η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας είναι

$$l(\boldsymbol{\beta}) = \sum_{j=1}^{k} \boldsymbol{\beta}'\boldsymbol{x}_j - \sum_{j=1}^{k} \ln\left\{\sum_{i\in R_j} e^{\boldsymbol{\beta}\prime \boldsymbol{x_i}}\right\}.$$

Τέλος, βρίσκοντας τις πρώτες μερικές παραγώγους

$$\frac{\partial l}{\partial \beta_r} = \sum_{j=1}^{k} x_{jr} - \sum_{j=1}^{k} \left[\sum_{i\in R_j} x_{ir}\, e^{\boldsymbol{\beta}\prime \boldsymbol{x_i}} \Big/ \sum_{i\in R_j} e^{\boldsymbol{\beta}\prime \boldsymbol{x_i}} \right]$$

και λύνοντας τις εξισώσεις $\frac{\partial l}{\partial \beta_r} = 0,\ r = 1, \ldots, p$ με επαναληπτικές διαδικασίες ως προς $\boldsymbol{\beta}$ (συνήθως χρησιμοποιούμε τη μέθοδο Newton-Raphson), βρίσκουμε τις εκτιμήτριες μεγίστης πιθανοφάνειας $\widehat{\boldsymbol{\beta}}$ της παραμέτρου $\boldsymbol{\beta}$.

Το αντίστροφο του πίνακα παρατηρούμενης πληροφορίας με το (r,s) στοιχείο του να είναι

$$-\frac{\partial^2 l}{\partial \beta_r \partial \beta_s} = \sum_{j=1}^{k} \sum_{i\in R_j} x_{ir} \left[x_{is} - \frac{\sum_{l\in R_j} x_{ls} e^{\boldsymbol{\beta}\prime \boldsymbol{x_l}}}{\sum_{i\in R_j} e^{\boldsymbol{\beta}\prime \boldsymbol{x_l}}} \right] \frac{e^{\boldsymbol{\beta}\prime \boldsymbol{x_i}}}{\sum_{i\in R_j} e^{\boldsymbol{\beta}\prime \boldsymbol{x_l}}}$$

υπολογισμένο στο $\widehat{\boldsymbol{\beta}}$ μας δίνει τις εκτιμήσεις των διασπορών των εκτιμήσεων $\widehat{\boldsymbol{\beta}}$.

Το μοντέλο του Cox είναι ένας βασικός τρόπος ανάλυσης δεδομένων διάρκειας ζωής με συμμεταβλητές. Γι'αυτό το λόγο η ανακάλυψη του μοντέλου θεωρείται μία από τις μεγαλύτερες επιτυχίες στη στατιστική.

1.4.3 Διαγνωστικές μέθοδοι στο μοντέλο του Cox

Είναι σημαντικό να ελέγξουμε αν το μοντέλο αναλογικής διακινδύνευσης του Cox είναι το κατάλληλο ή όχι, δηλαδή αν ισχύει η υπόθεση της αναλογικής διακινδύνευσης. Υπάρχουν κάποιες διαγνωστικές μέθοδοι που μας βοηθούν σε αυτόν τον έλεγχο.

1. Γραφικός έλεγχος

Η συνάρτηση επιβίωσης δίνεται από τον τύπο

$$S(t;\boldsymbol{x}) = \exp(-H_0(t)e^{\boldsymbol{\beta}'\boldsymbol{x}})$$

όπου $H_0(t)$ η σωρευτική συνάρτηση διακινδύνευσης που αντιστοιχεί στη βασική συνάρτηση διακινδύνευσης $h_0(t)$. Οπότε, έχουμε

$$\ln\{-lnS(t;\boldsymbol{x})\} - lnH_0(t) = \boldsymbol{\beta}'\boldsymbol{x}$$

Αυτό μας δείχνει ότι η καμπύλη

$$\ln\{-lnS(t;\boldsymbol{x})\}$$

για οποιαδήποτε συμμεταβλητή **x** είναι παράλληλη της $lnH_0(t)$ ως προς το χρόνο. Συνεπώς, οι καμπύλες $\ln\{-lnS(t;\boldsymbol{x})\}$ θα είναι παράλληλες και μεταξύ τους ως προς το χρόνο για διαφορετικές τιμές του **x.**

Οπότε για να πραγματοποιήσουμε το γραφικό έλεγχο για την υπόθεση της αναλογικής διακινδύνευσης:

- υπολογίζουμε την εκτιμήτρια Kaplan-Meier $\hat{S}(t;\boldsymbol{x})$ για επιλεγμένες τιμές **x**
- πραγματοποιούμε τις γραφικές παραστάσεις των $\ln\{-ln\hat{S}(t;\boldsymbol{x})\}$ έναντι του χρόνου
- αν οι καμπύλες αυτές είναι παράλληλες μεταξύ τους ως προς το χρόνο, ισχύει η υπόθεση της αναλογικής διακινδύνευσης

Ένα μειονέκτημα του γραφικού ελέγχου είναι ότι οι εκτιμήσεις $\hat{S}(t; \boldsymbol{x})$ είναι έγκυρες μόνο όταν έχουμε ένα αρκετά μεγάλο δείγμα παρατηρήσεων για επιλεγμένες τιμές **x**. Γι'αυτό το λόγο ο γραφικός έλεγχος πραγματοποιείται όταν έχουμε λίγες συμμεταβλητές και χρειάζεται να ομαδοποιήσουμε τις τιμές των ποσοτικών συμμεταβλητών. Ο γραφικός έλεγχος γίνεται πάντα πριν την προσαρμογή του μοντέλου.

2. Υπόλοιπα Schoenfeld

Για τον έλεγχο της καταλληλότητας του μοντέλου, αφού γίνει η προσαρμογή του, μπορούμε να χρησιμοποιήσουμε τα υπόλοιπα. Τα υπόλοιπα μας δείχνουν τις διαφορές μεταξύ των παρατηρούμενων τιμών και των προβλεπόμενων τιμών από την προσαρμογή του μοντέλου.

Στην παράγραφο 1.4.1 είδαμε τα υπόλοιπα Cox-Snell που είναι πολύ χρήσιμα και δίνονται από το τύπο:

$$-\ln \hat{S}\left(t_{(j)};\, \boldsymbol{x_j}\right) = \hat{H}\left(t_{(j)};\, \boldsymbol{x_j}\right) = \widehat{H_0}(t_{(j)}) e^{\widehat{\boldsymbol{\beta}'} \boldsymbol{x_j}},$$

όπου $\widehat{H_0}$ μία μη-παραμετρική εκτιμήτρια. Παρόλο που τα υπόλοιπα Cox-Snell χρησιμοποιούνται συχνά και είναι αρκετά σημαντικά, στην περίπτωση του μοντέλου του Cox είναι λιγότερο χρήσιμα καθώς η βασική συνάρτηση διακινδύνευσης, h_0, δεν καθορίζεται και άρα ούτε και η $\widehat{H_0}$ μπορεί να καθοριστεί. Επομένως, για το μοντέλο του Cox χρησιμοποιούμε τα υπόλοιπα Schoenfeld.

Στην παράγραφο 1.4.2 είδαμε ότι η πιθανότητα να διακοπεί η λειτουργία μίας μονάδας j, δοθέντος ότι σταματά να λειτουργεί μία οποιαδήποτε μονάδα από το σύνολο R_j είναι $p_j = \frac{e^{\boldsymbol{\beta}' \boldsymbol{x_j}}}{\sum_{i \in R_j} e^{\boldsymbol{\beta}' \boldsymbol{x_i}}}$. Επειδή, όμως, δε ξέρουμε ποια είναι η μονάδα που θα σταματήσει να λειτουργεί τη χρονική στιγμή $t_{(j)}$ από το σύνολο R_j, θεωρούμε ότι η τιμή των συμμεταβλητών **x** αυτής της μονάδας είναι τυχαία μεταβλητή με αναμενόμενη τιμή

$$E(\boldsymbol{x}|R_j) = \sum_{k \in R_j} x_k\, p_k = \frac{\sum_{k \in R_j} x_k\ e^{\beta' x_k}}{\sum_{i \in R_j} e^{\beta' x_i}}.$$

Τα υπόλοιπα Schoenfeld ορίζονται ως

$$\hat{\boldsymbol{r}}_{\boldsymbol{j}} = \boldsymbol{x}_{\boldsymbol{j}} - \hat{E}(\boldsymbol{x}|R_j) \qquad (1.10)$$

όπου $\hat{E}(\boldsymbol{x}|R_j)$ είναι το $E(\boldsymbol{x}|R_j)$ με τη διαφορά ότι στη θέση των $\boldsymbol{\beta}$ βάζουμε τα $\widehat{\boldsymbol{\beta}}$ (Schoenfeld, 1982).

Όπως παρατηρούμε στη σχέση (1.10), τα υπόλοιπα Schoenfeld προσδιορίζονται με τη βοήθεια των συμμεταβλητών **x** και όχι από τις τιμές της εξαρτημένης μεταβλητής, όπως συνηθίζεται στα υπόλοιπα του μοντέλου της κλασικής παλινδρόμησης. Ακόμη, τα υπόλοιπα αυτά είναι διανύσματα και κάθε μη-αποκομμένη παρατήρηση έχει τόσα υπόλοιπα όσες είναι και οι συμμεταβλητές του μοντέλου. Τα υπόλοιπα Schoenfeld αναπαριστούν την απόκλιση μεταξύ της συμμεταβλητής της μονάδας που της συνέβη το γεγονός τη χρονική στιγμή t_j και του σταθμισμένου μέσου όρου όλων των συμμεταβλητών που ανήκουν στο σύνολο R_j. Οπότε, μία μεγάλη τιμή του υπολοίπου Schoenfeld μας δείχνει ότι η μονάδα εκείνη που της συνέβη το γεγονός τη χρονική στιγμή t_j ήταν μία ακραία παρατήρηση. Δηλαδή, τα υπόλοιπα Schoenfeld συνδέονται άμεσα με την έννοια των σημείων επιρροής στα οποία θα αναφερθούμε αργότερα.

Αρκετά συχνά χρησιμοποιούνται τα κλιμακοποιημένα υπόλοιπα Schoenfeld καθώς είναι πιο εύκολο να υπολογιστούν. Τα κλιμακοποιημένα Schoenfeld δίνονται από τον τύπο

$$\boldsymbol{r}_{\boldsymbol{j}}^{*} = k\hat{V}(\widehat{\boldsymbol{\beta}})\hat{\boldsymbol{r}}_{\boldsymbol{j}}$$

όπου k το πλήθος των μη-αποκομμένων παρατηρήσεων και $\hat{V}(\widehat{\boldsymbol{\beta}})$ ο εκτιμώμενος πίνακας διασποράς των $\widehat{\boldsymbol{\beta}}$ (Grambsch & Therneau, 1994).

Επιπλέον, με τα κλιμακοποιημένα υπόλοιπα Schoenfeld μπορούμε να ελέγξουμε αν ισχύει η υπόθεση της αναλογικής διακινδύνευσης στο μοντέλο του Cox. Αν υποθέσουμε ότι τουλάχιστον μία συνιστώσα του διανύσματος των συντελεστών των συμμεταβλητών, **β**(t), δεν είναι σταθερή, δηλαδή εξαρτάται από το χρόνο τότε το μοντέλο του Cox γίνεται

$$h(t;\boldsymbol{x}) = h_0(t)\exp(\boldsymbol{\beta}(\boldsymbol{t})'\boldsymbol{x})$$

Σε αυτή την περίπτωση δεν ισχύει η αναλογικότητα. Η υπόθεση της αναλογικής διακινδύνευσης ισχύει όταν **β**(t)=**β** και τότε το διάγραμμα των συντελεστών $\beta_i(t_{(j)})$ ως προς $t_{(j)}$ θα είναι μία οριζόντια γραμμή. Αποδεικνύεται ότι

$$E(r_{ij}^{*}) \cong \beta_i(t_{(j)}) - \widehat{\beta}_i,$$

όπου $\beta_i(t_{(j)})$ ο συντελεστής της συμμεταβλητής i τη χρονική στιγμή $t_{(j)}$ και r_{ij}^{*} τα κλιμακοποιημένα υπόλοιπα Schoenfeld (Grambsch & Therneau, 1994).

Μέσω της παλινδρόμησης

$$\beta_i(t) = \beta_i + \theta_i(g(t) - \bar{g}),\ i = 1, \dots, p$$

μπορούμε να μελετήσουμε την εξάρτηση του συντελεστή $\beta_i(t)$ από το χρόνο ή από μία συνάρτηση του χρόνου. Θεωρούμε $g(t)$ συνάρτηση του χρόνου και $\bar{g}$ είναι ο μέσος όρος των συναρτήσεων του χρόνου, $g_j = g(t_{(j)})$ για τις χρονικές στιγμές $t_{(j)}$. Επομένως, ένας έλεγχος που γίνεται για την αναλογικότητα στο μοντέλο του Cox είναι

H_0: $\theta_i = 0\ για\ κάθε\ i = 1, \dots, p$ vs H_1: διαφορετικά

με την ελεγχοσυνάρτηση να δίνεται από τον τύπο

$$T = \frac{(\boldsymbol{g}-\bar{\boldsymbol{g}})' S^{*} I(\hat{\beta}) S^{*\prime} (\boldsymbol{g}-\bar{\boldsymbol{g}})}{k \sum_{j=1}^{k} (g_j - \bar{g})^2} \qquad (1.11)$$

όπου κ ο αριθμός των γεγονότων, $\boldsymbol{g} - \bar{\boldsymbol{g}}$ το διάνυσμα k-διάστασης με j-οστό στοιχείο $g(t_{(j)}) - \bar{g}$, S^{*} ο kxp πίνακας των κλιμακοποιημένων Schoenfeld, p το πλήθος των συμμεταβλητών και $I^{-1}(\hat{\beta}) = \hat{V}(\hat{\beta})$. Η ελεγχοσυνάρτηση (1.11) ακολουθεί ασυμπτωτικά την κατανομή X_p^2 υπό την H_0. Όταν θέλουμε να ελέγξουμε την αναλογικότητα για μόνο την i-οστή συμμεταβλητή τότε η ελεγχοσυνάρτηση είναι

$$T_i = \frac{\left\{\sum_{j=1}^{k} (g_j - \bar{g}) r_{ij}^{*}\right\}^2}{k I^{ii} \sum_{j=1}^{k} (g_j - \bar{g})^2} \qquad (1.12)$$

και ακολουθεί ασυμπτωτικά την X_1^2 κατανομή, όπου r_{ij}^* το (i,j)-οστό στοιχείο του πίνακα S^* και I^{ii} το (i,j)-οστό διαγώνιο στοιχείο του πίνακα $I^{-1}(\hat{\beta})$ (Therneau & Grambsch, 2000).

3. Σημεία επιρροής

Για την καταλληλότητα του μοντέλου είναι σημαντικό να ελέγξουμε αν υπάρχουν κάποιες συγκεκριμένες παρατηρήσεις που επηρεάζουν σε μεγάλο βαθμό τα αποτελέσματα. Δηλαδή, αν υποθέσουμε ότι αφαιρούμε μία παρατήρηση από το μοντέλο και αυτό έχει σαν αποτέλεσμα να αυξηθεί ή να μειωθεί ο κίνδυνος να συμβεί το γεγονός που μελετάμε κατά μία σημαντική ποσότητα, τότε η παρατήρηση αυτή ονομάζεται σημείο επιρροής. Τέτοια σημεία ενδεχομένως να υπάρχουν εξαιτίας της παράλειψης κάποιας σημαντικής μεταβλητής ή σε κάποιο λάθος που έγινε κατά τη διάρκεια της μέτρησης (Collett, 2003).

Έστω ότι επιθυμούμε να ελέγξουμε αν κάποια συγκεκριμένη παρατήρηση έχει επίδραση στην εκτιμήτρια της παραμέτρου β_j, την $\widehat{\beta_j}$ με j=1,…,p, p το σύνολο των παραμέτρων, σε ένα μοντέλο παλινδρόμησης του Cox. Σε αυτή τη περίπτωση πρώτα προσαρμόζουμε το μοντέλο με όλες τις παρατηρήσεις, n, των δεδομένων μας και μετά προσαρμόζουμε το μοντέλο χωρίς την παρατήρηση εκείνη, έτσι στο δεύτερο μοντέλο θα έχουμε n-1 παρατηρήσεις. Από τα αποτελέσματα που θα προκύψουν μπορούμε να διαπιστώσουμε αν η παρατήρηση εκείνη είναι σημείο επιρροής. Την ίδια διαδικασία μπορούμε να επαναλάβουμε για όλες τις παρατηρήσεις. Αν $\widehat{\beta_j}$ είναι η εκτιμήτρια της παραμέτρου β_j και $\widehat{\beta_j}(i)$ είναι η εκτιμήτρια της β_j μετά από αφαίρεση της i-οστής παρατήρησης τότε ορίζουμε την ποσότητα

$$DFBETAS_{ji} = \frac{\widehat{\beta_j} - \widehat{\beta_j}(i)}{\sqrt{S(i)^2 c_{jj}}}$$

όπου $S(i)^2$ η εκτιμήτρια της διασποράς σ^2 που προκύπτει από την προσαρμογή του μοντέλου όταν έχει αφαιρεθεί η i-οστή παρατήρηση και c_{jj} το j-οστό διαγώνιο στοιχείο του συμμετρικού πίνακα $(X'X)^{-1}$, με X τον πίνακα σχεδιασμού. Αν

$$\left|DFBETAS_{ji}\right| > \frac{2}{\sqrt{n}}$$

τότε η i-οστή παρατήρηση έχει μεγάλη επιρροή στην εκτιμήτρια του j-οστού συντελεστή του μοντέλου (Καρώνη, 2017).

Όπως αναφέρθηκε προηγουμένως, στο μοντέλο του Cox για να βρούμε αν είναι μία παρατήρηση σημείο επιρροής χρειάζεται να κάνουμε προσαρμογή του μοντέλου δύο φορές, μία με όλες τις παρατηρήσεις και μία χωρίς την παρατήρηση αυτή. Υπάρχει, όμως, τρόπος να το αποφύγουμε αυτό χρησιμοποιώντας τον πρώτο όρο της επέκτασης της σειράς Taylor για την συνάρτηση score που μας δείχνει πόσο ευαίσθητη είναι η συνάρτηση πιθανοφάνειας L(**θ;X**) στην παράμετρο της **θ**. Έτσι έχουμε

$$\hat{\beta} - \widehat{\beta}\,(j) \cong I^{-1}\boldsymbol{d_j}$$

όπου

$$\boldsymbol{d_j} = \delta_j \hat{\boldsymbol{r}}_{\boldsymbol{j}} - \sum_{i \in D_j} \frac{e^{\beta' x_j}\{x_j - \widehat{\boldsymbol{E}}(\boldsymbol{x}|\boldsymbol{R_i})\}}{\sum_{k \in R_i} e^{\beta' x_k}} \qquad (1.13)$$

όπου D_j είναι το σύνολο εκείνων που τους συνέβη το γεγονός τη χρονική στιγμή t_j ή πριν τη στιγμή αυτή, $\boldsymbol{d_j}$ τα υπόλοιπα score, $I^{-1}\boldsymbol{d_j}$ είναι οι γραμμές του nxp πίνακα που έχει ως στοιχεία του τις τιμές $DFBETAS$ που αναφέρθηκαν προηγουμένως. Ο πρώτος όρος του τύπου (1.13) είναι τα υπόλοιπα Schoenfeld και υπάρχει μόνο για τις μονάδες j στις οποίες έχει συμβεί το γεγονός. Ο δεύτερος όρος του τύπου υπάρχει για όλα τις μονάδες και αναπαριστά τη συμβολή όλων των συνόλων R_i που βρίσκονται σε κίνδυνο και περιέχουν τη μονάδα j. Ο πρώτος όρος του τύπου (1.13) είναι πιο σημαντικός για τις μονάδες στις οποίες συμβαίνει το γεγονός νωρίς. Ο δεύτερος όρος είναι πιο σημαντικός για τις μονάδες στις οποίες αργεί να συμβεί το γεγονός. Για εκείνες τις μονάδες που είναι αποκομμένες από νωρίς στη μελέτη, ο πρώτος όρος του τύπου είναι μηδέν και ο δεύτερος αρκετά μικρός. Σε αυτή την περίπτωση οι συγκεκριμένες μονάδες δεν έχουν μεγάλη επιρροή στο μοντέλο μας (Caroni, 2004).

1.5 Επέκταση του μοντέλου του Cox

Μία επέκταση του κλασικού μοντέλου του Cox είναι η στρωματοποιημένη ανάλυση. Με άλλα λόγια, υπάρχουν περιπτώσεις που η υπόθεση της αναλογικής διακινδύνευσης δεν ισχύει στο σύνολο των δεδομένων μας αλλά σε διαφορετικά υποσύνολα των δεδομένων. Για παράδειγμα, αν υποθέσουμε ότι μελετάμε το κατά πόσο η ηλικία των ασθενών επηρεάζει τη θεραπεία στην οποία θα υποβληθούν τότε χωρίζουμε τη μεταβλητή ηλικία σε δύο στρώματα, τους νέους και τους ηλικιωμένους.

Σε αυτή τη περίπτωση θα έχουμε διαφορετικές συναρτήσεις διακινδύνευσης για τα δύο στρώματα, δηλαδή θα έχουμε

$$h(t;\boldsymbol{x}) = \begin{Bmatrix} e^{\boldsymbol{\beta}'\boldsymbol{x}}h_{01}(t), & για\ τους\ νέους \\ e^{\boldsymbol{\beta}'\boldsymbol{x}}h_{02}(t), & για\ τους\ ηλικιωμένους \end{Bmatrix}$$

όπου $h_{01}(t), h_{02}(t)$ οι βασικές συναρτήσεις διακινδύνευσης για τους νέους και τους ηλικιωμένους αντίστοιχα, οι οποίες όμως δεν βρίσκονται σε αναλογία μεταξύ τους. Παρόλα αυτά η υπόθεση της αναλογικής διακινδύνευσης ισχύει για τις υπόλοιπες συμμεταβλητές. Το διάνυσμα $\boldsymbol{\beta}$ των συντελεστών των άλλων συμμεταβλητών είναι κοινό και για τα δύο στρώματα.

Για την εκτίμηση των παραμέτρων του μοντέλου χρησιμοποιούμε τη μέθοδο της μεγίστης πιθανοφάνειας και έτσι για κάθε στρώμα m έχουμε ότι ο λογάριθμος της μερικής πιθανοφάνειας είναι

$$l_m(\boldsymbol{\beta}) = \sum_{\boldsymbol{j}=1}^{k_m} \boldsymbol{\beta}'\,\boldsymbol{x}_{mj} - \sum_{j=1}^{k_m} \ln\left\{ \sum_{i\in R_{mj}} e^{\beta' \boldsymbol{x}_{mi}} \right\}$$

δηλαδή είναι ο ίδιος τύπος που αναφερθήκαμε στο κλασικό μοντέλο του Cox με τη διαφορά ότι προσθέτουμε το δείκτη m για να τονίσουμε ότι αναφερόμαστε σε στρώματα. Τέλος, για όλα συνολικά τα στρώματα m=1,…,s έχουμε

$$l(\boldsymbol{\beta}) = \sum_{m=1}^{s} l_m(\boldsymbol{\beta})$$

και συνεχίζουμε με τη διαδικασία που περιγράψαμε στην Παράγραφο 1.4.2 λαμβάνοντας υπόψη τα στρώματα.

Για να ισχύει η υπόθεση της αναλογικής διακινδύνευσης στην περίπτωση της στρωματοποιημένης ανάλυσης πρέπει να ελέγξουμε αν οι καμπύλες των $\ln(-\ln \hat{S}_m(t)),\ m = 1, \dots, s$ έναντι του χρόνου t είναι παράλληλες. Αν οι καμπύλες δεν προκύπτουν παράλληλες αυτό σημαίνει ότι ήταν λάθος να χωρίσουμε τη μεταβλητή σε m στρώματα.

ΚΕΦΑΛΑΙΟ 2

Το Μοντέλο Παλινδρόμησης Threshold

Ένα στατιστικό μοντέλο είναι μία κατανομή πιθανότητας που κατασκευάζεται έτσι ώστε να εξάγουμε συμπεράσματα και να πάρουμε συγκεκριμένες αποφάσεις για τα δεδομένα μας (Davison, 2003). Είναι σημαντικό το γεγονός ότι ένα στατιστικό μοντέλο δεν είναι απαραίτητο να έχει άμεση σχέση με τη φυσική διαδικασία από την οποία παράχθηκαν τα δεδομένα. Για παράδειγμα έστω ότι έχουμε το κλασικό μοντέλο παλινδρόμησης $y = \boldsymbol{\beta}'\boldsymbol{x} + \varepsilon$. Τότε, δεν είναι απαραίτητο η μεταβλητή πρόβλεψης x να συνεισφέρει στην τιμή που θα πάρει η εξαρτημένη μεταβλητή y στην πραγματικότητα. Αυτό που απαιτείται είναι να υπάρχει στατιστική συσχέτιση που να καθιστά την μεταβλητή x ικανή να χρησιμοποιηθεί για την πρόβλεψη της y. Αυτού του είδους τα μοντέλα ο Cox τα ονομάζει εμπειρικά (Cox, 1990).

Παρόλα αυτά, στο χώρο της επιστήμης μοντέλο θεωρείται η αναπαράσταση μιας φυσικής διαδικασίας. Για παράδειγμα, γίνεται κάποια πειραματική διαδικασία στα ζώα και εξετάζεται η αντίδρασή τους σε αυτήν. Στη συνέχεια κατασκευάζεται ένα μαθηματικό μοντέλο με σκοπό να μελετηθεί η αντίστοιχη αντίδραση που θα έχει η ίδια διαδικασία στους ανθρώπους. Συνήθως, το μαθηματικό αυτό μοντέλο παράγεται με τη βοήθεια διαφορικών εξισώσεων. Το βασικό στοιχείο των μαθηματικών μοντέλων είναι να μιμηθούν σε κάποιο βαθμό τι ακριβώς γίνεται στο φυσικό σύστημα και να μπορούν οι παράμετροι του μοντέλου να είναι φυσικά ερμηνεύσιμες. Τέτοιου είδους μοντέλα μπορούν να χρησιμοποιηθούν στη στατιστική αν και είναι προτιμότερα σε εμπειρικά μοντέλα. Ο Cox έλεγε ότι «Τα πιο ελκυστικά μοντέλα είναι αυτά που συνδέονται άμεσα με το προς συζήτηση θέμα». Τα μοντέλα αυτά τα ονομάζει ουσιαστικά (substantive) (Cox, 1990).

Το ημι-παραμετρικό μοντέλο αναλογικής διακινδύνευσης του Cox είναι καθαρά περιγραφικό εμπειρικό μοντέλο και για αυτό το λόγο μπορεί να θεωρηθεί κατώτερο των άλλων μοντέλων. Ακόμη και ο Cox αναφέρει ότι τα μοντέλα επιταχυνόμενης διακοπής είναι πιο ελκυστικά εξαιτίας της άμεσης φυσικής τους ερμηνείας, ειδικότερα σε ότι αφορά τη μηχανική. Γι'αυτό το λόγο δημιουργήθηκε η ανάγκη

κατασκευής πιο ουσιαστικών μοντέλων. Μπορούμε να δούμε το γεγονός που μας ενδιαφέρει σαν το τελικό στάδιο μιας αναπτυσσόμενης διαδικασίας, η οποία είναι πολύ σημαντική ακόμα και όταν μας είναι άγνωστη. Αυτό που προτείνεται είναι η μελέτη της διάρκειας ζωής ως το αποτέλεσμα μιας υποβόσκουσας διαδικασίας, με το γεγονός που μας ενδιαφέρει να συμβαίνει όταν αυτή η διαδικασία φτάσει σε κάποιο όριο για πρώτη φορά. Δηλαδή η διάρκεια ζωής θα είναι ο χρόνος πρώτης διέλευσης της στοχαστικής διαδικασίας σε αυτό το όριο (Aalen & Gjessing, 2001). Η πρόταση αυτή των Aalen και Gjessing ήταν τόσο σημαντική που προκάλεσε το ενδιαφέρον πολλών να βρουν εναλλακτικά μοντέλα πέρα από το μοντέλο αναλογικής διακινδύνευσης (Caroni, 2017).

Συνεπώς, το γεγονός που μελετάμε συμβαίνει όταν μία στοχαστική διαδικασία συναντά μία οριακή τιμή (threshold). Ο χρόνος που απαιτείται για μία στοχαστική διαδικασία να φτάσει αυτή την οριακή τιμή για πρώτη φορά, ξεκινώντας από μία αρχική κατάσταση, ονομάζεται χρόνος πρώτης συνάντησης ή μετάβασης (first hitting time). Έτσι, λοιπόν, τα δεδομένα διάρκειας ζωής μπορούν να θεωρηθούν χρόνοι πρώτης μετάβασης, δηλαδή χρόνοι μέχρι μία στοχαστική διαδικασία να φτάσει κάποιο συγκεκριμένο όριο. Η στοχαστική διαδικασία μπορεί να είναι από διαδικασία Wiener μέχρι αλυσίδα Markov. Τα μοντέλα πρώτης μετάβασης έχουν αρκετές εφαρμογές σε πολλούς διαφορετικούς τομείς, όπως στην ιατρική, τη μηχανική, τις επιχειρήσεις, τα οικονομικά και την κοινωνιολογία, γι’ αυτό και το τελευταίο διάστημα χρησιμοποιούνται αρκετά συχνά. Τα μοντέλα αυτά μπορούν να περιγράψουν περιπτώσεις όπως τη διάρκεια παραμονής ασθενών σε νοσοκομείο, το χρόνο επιβίωσης ενός ασθενή μετά από μεταμόσχευση και το χρόνο που χρειάζεται μία μηχανή μέχρι να σταματήσει να λειτουργεί.

Για να γίνουν τα μοντέλα πρώτης μετάβασης χρήσιμα σε ακόμη περισσότερες εφαρμογές πρέπει να επεκταθούν έτσι ώστε να έχουν δομή μοντέλων παλινδρόμησης. Το μοντέλο παλινδρόμησης threshold αναφέρεται σε μοντέλα πρώτης μετάβασης με δομές παλινδρόμησης τα οποία μπορεί να περιλαμβάνουν δεδομένα με πολλές συμμεταβλητές. Οι παράμετροι της διαδικασίας, το σύνολο των οριακών τιμών και ο χρόνος μπορεί να εξαρτώνται από τις συμμεταβλητές. Η παλινδρόμηση threshold αναδύθηκε κυρίως στις αρχές του αιώνα που διανύουμε και από τότε έχει αυξηθεί ραγδαία η χρήση της.

Το μοντέλο πρώτης μετάβασης αποτελείται από δύο μέρη: τη στοχαστική διαδικασία και το σύνολο των ορίων. Η στοχαστική διαδικασία ορίζεται ως $\{X(t), t \in T, x \in X\}$ όπου T ο χρόνος και X ο χώρος των καταστάσεων, δηλαδή οι τιμές που μπορεί να πάρει η στοχαστική διαδικασία. Επίσης, θεωρούμε την αρχική τιμή $X(0) = x_0$. Το σύνολο των ορίων $B \subset X$. Σημαντικό επίσης χαρακτηριστικό είναι το αν η διαδικασία είναι παρατηρήσιμη ή όχι. Συνήθως η στοχαστική διαδικασία είναι λανθάνουσα, δηλαδή μη παρατηρήσιμη. Παίρνοντας την αρχική τιμή της διαδικασίας $X(0) = x_0$ να βρίσκεται εκτός του συνόλου B, έχουμε ότι ο πρώτος χρόνος μετάβασης με το B είναι η τυχαία μεταβλητή S και ορίζεται:

$$S = \inf\{t: X(t) \in B\} \qquad (2.1)$$

Επομένως, ο πρώτος χρόνος μετάβασης είναι ο χρόνος στον οποίο η στοχαστική διαδικασία συναντά για πρώτη φορά το σύνολο B. Στο σημείο που γίνεται η συνάντηση έχουμε $X(S) \in B$ και ονομάζεται σημείο κατωφλιού (threshold). Σε μερικές περιπτώσεις των μοντέλων πρώτης μετάβασης δεν είναι σίγουρο ότι η στοχαστική διαδικασία θα συναντήσει κάποια στιγμή το σύνολο B, γι' αυτό $P(S < \infty) < 1$. Έτσι, θα ορίσουμε $S = \infty$ να είναι η απουσία του χρόνου μετάβασης με πιθανότητα $P(S = \infty) = 1 - P(S < \infty)$.

Στο μοντέλο (2.1) θεωρούμε ότι το σύνολο B είναι ανεξάρτητο του χρόνου, παρόλα αυτά υπάρχουν και περιπτώσεις που έχουμε εξάρτηση από το χρόνο και τότε το σύνολο είναι το $B(t)$.

Όπως αναφέρθηκε παραπάνω η στοχαστική διαδικασία μπορεί να πάρει διάφορες μορφές. Θα αναφερθούμε μερικές από αυτές (Lee & Whitmore, 2006).

α. **Διαδικασία Bernoulli.** Σε αυτή την περίπτωση θεωρούμε ότι ο αριθμός των δοκιμών, S, που απαιτείται για να φτάσουμε την m-οστή επιτυχία σε μία διαδικασία Bernoulli $\{B_t: t=1,2,\ldots\}$ ακολουθεί την αρνητική διωνυμική κατανομή με παραμέτρους m και p, όπου p η πιθανότητα επιτυχίας σε κάθε δοκιμή. Έτσι, λοιπόν, θεωρούμε τη διαδικασία $\{X_t: t=0,1,2,\ldots\}$ με αρχική τιμή $X_0=x_0=m$ και $X_t=x_0-B_t$ t=1,2,.. , όπου B_t η διαδικασία Bernoulli. Ο χρόνος πρώτης μετάβασης είναι η πρώτη δοκιμή Bernoulli t=S για την οποία $X_t=0$.

β. **Διαδικασία Poisson.** Θεωρούμε ότι ο χρόνος S μέχρι να συμβεί το m-οστό γεγονός σε μία διαδικασία Poisson $\{N(t): t\geq 0\}$ με παράμετρο λ ακολουθεί την κατανομή Erlang με παραμέτρους m και λ. Έτσι, λοιπόν, θεωρούμε τη διαδικασία $\{X(t): t\geq 0\}$ με αρχική τιμή $X(0)=x_0=m$ και $X(t)=x_0-N(t)$, όπου $\{N(t): t\geq 0\}$ η διαδικασία Poisson. Ο χρόνος πρώτης μετάβασης είναι ο νωρίτερος χρόνος t=S για τον οποίο X(t)=0.

γ. **Διαδικασία Wiener.** Θεωρούμε τη διαδικασία Wiener $\{X(t): t\geq 0\}$ με μέσο μ, διασπορά σ^2 και αρχική τιμή $X(0)=x_0>0$. Ο χρόνος S που απαιτείται για να φτάσει η διαδικασία στο μηδέν για πρώτη φορά, δηλαδή να συμβεί το γεγονός, ακολουθεί την αντίστροφη Γκαουσιανή κατανομή αν $\mu<0$ έτσι ώστε η διαδικασία να τείνει στο μηδέν. Ένα παράδειγμα όπου μπορεί να χρησιμοποιηθεί το μοντέλο αυτό είναι η περιγραφή της διάρκειας μιας απεργίας. Έστω ότι {X(t)} είναι ένας τρόπος μέτρησης των διαφορών μεταξύ της διοίκησης και των εργατών σε χρόνο t μετά την έναρξη της απεργίας. Ως αρχική τιμή $X(0)=x_0>0$ θεωρούμε τη στιγμή της διαφωνίας των δύο ομάδων. Η απεργία λήγει όταν η διαδικασία φτάσει για πρώτη φορά στο μηδέν, δηλαδή όταν οι δύο ομάδες συμφωνήσουν (Lancaster, 1972).

δ. **Διαδικασία Γάμμα.** Θεωρούμε τη διαδικασία $\{X(t): t\geq 0\}$ με αρχική τιμή $X(0)=x_0>0$. Ορίζουμε $X(t)=x_0-Z(t)$, όπου $\{Z(t): t\geq 0\}$ είναι μία διαδικασία Γάμμα με παράμετρο κλίμακας β, παράμετρο σχήματος α και Z(0)=0. Ο πρώτος χρόνος μετάβασης της διαδικασίας {X(t)}, X=0, ακολουθεί την αντίστροφη Γάμμα κατανομή με συνάρτηση κατανομής $P(S>t)=P(Z(t)<x_0)$.

ε. **Διαδικασία Ornstein-Uhlenbeck (O.U).** Η διαδικασία O.U είναι μία παραλλαγή της Wiener που τείνει να παρασυρθεί προς ένα σταθερό επίπεδο ισορροπίας και άρα έχει την ιδιότητα της ομοιόστασης. Η κατανομή του χρόνου πρώτης μετάβασης για τη διαδικασία O.U θεωρείται ότι είναι Ricciardi-Sato (Ricciardi & Sato, 1988) (Aalen & Gjessing, 2004).

στ. **Αλυσίδα Markov.** Έστω έχουμε τη μαρκοβιανή αλυσίδα $\{X_t: t=0,1,2,\ldots\}$. Θεωρούμε *X* το χώρο των καταστάσεων, δηλαδή το σύνολο των τιμών που μπορεί να πάρει η αλυσίδα και *T* το χρονικό διάστημα στο οποίο γίνονται τα στάδια μετάβασης της αλυσίδας. Στην περίπτωση της μαρκοβιανής αλυσίδας ο χρόνος πρώτης

μετάβασης είναι ο ελάχιστος αριθμός βημάτων που απαιτούνται για να μετακινηθούμε από μία αρχική κατάσταση $X_0=x_0$ σε ένα σύνολο οριακών καταστάσεων *B*. Η κατανομή που ακολουθεί ο χρόνος πρώτης μετάβασης εξαρτάται από τον πίνακα μετάβασης της αλυσίδας.

ζ. **Ημι-μαρκοβιανή διαδικασία.** Η ημι-μαρκοβιανή διαδικασία {X(t): t≥0} επεκτείνει την αλυσίδα Markov καθώς περιλαμβάνει τον τυχαίο χρόνο στον οποίο η διαδικασία βρίσκεται στην εκάστοτε κατάσταση του χώρου *X*. Το μοντέλο της ημι-μαρκοβιανής αλυσίδας είναι αρκετά χρήσιμο παρόλο που η μαρκοβιανή ιδιότητα που ισχύει στην αλυσίδα Markov χάνεται. Ο χρόνος πρώτης μετάβασης είναι ο χρόνος εκείνος που η διαδικασία βρίσκεται στην αρχική ή τις μεταγενέστερες καταστάσεις πριν εισαχθεί στο σύνολο των οριακών καταστάσεων *B*.

Εμείς θα θεωρήσουμε ότι έχουμε τη διαδικασία Wiener. Έστω, λοιπόν, η διαδικασία Wiener $\{X(t), t \geq 0\}$ με μέσο μ, διασπορά σ^2 και αρχική τιμή $X(0) = x_0 > 0$. Θεωρούμε ότι η διαδικασία είναι λανθάνουσα. Ο χρόνος πρώτης μετάβασης ακολουθεί την αντίστροφη Γκαουσιανή κατανομή με παραμέτρους τις μ και σ^2 καθώς επίσης και την αρχική κατάσταση x_0. Η αρχική κατάσταση x_0 ορίζει το σημείο έναρξης της διαδικασίας. Σε περίπτωση που το γεγονός που μελετάμε είναι ο θάνατος από κάποια ασθένεια, τότε όσο πιο μακριά είναι η αρχική τιμή x_0 από το όριο (threshold) τόσο καλύτερη είναι η αρχική κατάσταση υγείας του ασθενούς. Η τάση ή κλίση μ περιγράφει το ρυθμό της κάθε μονάδας στο χρόνο με τον οποίο η διαδικασία φτάνει στο οριακό σημείο. Από τη θεωρία γνωρίζουμε ότι η συνάρτηση πυκνότητας πιθανότητας της αντίστροφης Γκαουσιανής όταν η διαδικασία ξεκινά από την αρχική τιμή x_0 και το όριο είναι στο επίπεδο του μηδενός είναι

$$f(t|\,\mu,\sigma^2,x_0) = \frac{x_0}{\sqrt{2\pi\sigma^2 t^3}} \exp\left(-\frac{(x_0+\mu t)^2}{2\sigma^2 t}\right) \mu\varepsilon\ -\infty < \mu < \infty, \sigma^2 > 0\ \kappa\alpha\iota\ x_0 > 0$$

(2.2)

Η συνάρτηση κατανομής που αντιστοιχεί στην παραπάνω συνάρτηση πιθανότητας είναι

$$F(t|\,\mu,\sigma^2,x_0) = \Phi\left(-\frac{(\mu t+x_0)}{\sqrt{\sigma^2 t}}\right) + \exp\left(\frac{-2x_0\mu}{\sigma^2}\right)\Phi\left(\frac{\mu t-x_0}{\sqrt{\sigma^2 t}}\right)$$

με Φ(.) τη συνάρτηση κατανομής της τυποποιημένης κανονικής κατανομής N(0,1) (Folks & Chhikara, 1978).

Αν το μ>0 τότε μπορεί να μη φτάσει η διαδικασία το πρώτο σημείο μετάβασης και έτσι η συνάρτηση πυκνότητας πιθανότητας θα είναι ακατάλληλη. Σ'αυτή την περίπτωση έχουμε $P(S = \infty) = 1 - \exp\left(-\frac{2x_0\mu}{\sigma^2}\right)$.

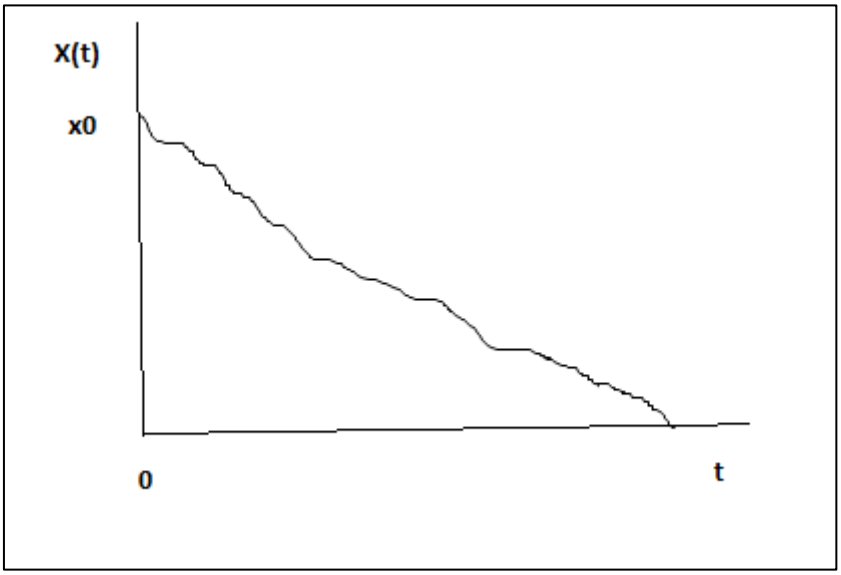
Θεωρούμε σ^2=1 και έτσι οι δύο άγνωστοι παράμετροι είναι η τάση μ (drift) και το x_0 (αρχική κατάσταση). Και οι δύο αυτές παράμετροι συνδέονται με k συμμεταβλητές παλινδρόμησης οι οποίες αναπαριστώνται από το διάνυσμα **z**=(1, z_1, …, z_k). Η πρώτη συνιστώσα του **z** είναι η μονάδα για να υπάρχει και ο σταθερός όρος στο μοντέλο παλινδρόμησης. Τώρα, η τάση μ συνδέεται με τις συμμεταβλητές μέσω της συνάρτησης σύνδεσης

$$\mu = \mathbf{z}\boldsymbol{\beta} = \beta_0 + \beta_1 z1 + \cdots + \beta_k z_k$$

ενώ η παράμετρος x_0 συνδέεται με τις συμμεταβλητές μέσω της λογαριθμικής συνάρτησης

$$\ln(x_0) = \boldsymbol{z\gamma} = \gamma_0 + \gamma_1 z1 + \cdots + \gamma_k z_k$$

με **β**=(β_0, β_1, …, β_k)' και **γ**=(γ_0, γ_1, …, γ_k)' τα διανύσματα των συντελεστών παλινδρόμησης (Lee & Whitmore, 2006).



Σχήμα 2.1: Η διαδικασία Wiener ως λανθάνουσα στοχαστική διαδικασία με αρχική κατάσταση x_0, τάση μ και σ^2=1

Στο Σχήμα 2.1 παρατηρούμε μία απεικόνιση της στοχαστικής διαδικασίας Wiener με αρχική κατάσταση x_0, τάση μ και $\sigma^2=1$. Το σημείο που το γράφημα τέμνει τον οριζόντιο άξονα είναι το οριακό σημείο (threshold).

2.1 Σύγκριση του μοντέλου αναλογικής διακινδύνευσης και του μοντέλου threshold

Το μοντέλο αναλογικής διακινδύνευσης του Cox θεωρείται από τα πιο σημαντικά μοντέλα για δεδομένα διάρκειας ζωής και μπορεί να χρησιμοποιηθεί σε πάρα πολλές εφαρμογές. Όπως είδαμε στην Παράγραφο 1.4.2 το μοντέλο του Cox προσδιορίζει τον τρόπο με τον οποίο το διάνυσμα των συμμεταβλητών **x** δρα πάνω στη βασική συνάρτηση διακινδύνευσης $h_0(t)$ μέσω της σχέσης

$$h(t;\boldsymbol{x}) = h_0(t)e^{\boldsymbol{\beta' x}}$$

Πρέπει να σημειωθεί ότι η συνάρτηση $h_0(t)$ οφείλει να είναι μονότονη, δηλαδή είτε να αυξάνεται (η πιο συχνή περίπτωση), είτε να μειώνεται, είτε να μένει σταθερή. Αν η $h_0(t)$ δεν είναι μονότονη τότε δεν μπορεί να ικανοποιηθεί η υπόθεση της αναλογικής διακινδύνευσης που είναι απαραίτητη για την προσαρμογή του μοντέλου. Συνεπώς, αυτός είναι ένας περιορισμός του μοντέλου του Cox που δεν τον έχει για παράδειγμα το μοντέλο της επιταχυνόμενης διακοπής. Έτσι, υπάρχουν και αρκετές εφαρμογές που το μοντέλο του Cox δεν μπορεί να χρησιμοποιηθεί.

Όπως αναφέρθηκε και στην αρχή του Κεφαλαίου, τα μοντέλα πρώτης μετάβασης είναι εναλλακτικά του μοντέλου του Cox. Εμείς θα θεωρήσουμε ως στοχαστική διαδικασία τη διαδικασία Wiener και έτσι ο χρόνος πρώτης μετάβασης θα ακολουθεί την αντίστροφη Γκαουσιανή κατανομή. Χρησιμοποιώντας μια διαφορετική παραμετροποίηση της αντίστροφης Γκαουσιανής κατανομής από αυτήν που αναπτύχθηκε στην αρχή του Κεφαλαίου, τύπος (2.2) έχουμε ότι η σ.π.π είναι η εξής:

$$f(t|\mu,\lambda) = \frac{\sqrt{\lambda}}{\sqrt{2\pi t^3}} \exp\left(-\frac{\lambda}{2\mu^2}\frac{(t-\mu)^2}{t}\right), \mu,\lambda > 0,\ t > 0$$

και η συνάρτηση επιβίωσης $S(t) = P(T > t)$ δίνεται από τον τύπο:

$$S(t) = \Phi\left[\sqrt{\lambda/t}\left(1-\frac{t}{\mu}\right)\right] - \exp\left(2\lambda/\mu\right)\Phi\left[-\sqrt{\lambda/t}\left(1+\frac{t}{\mu}\right)\right],$$

όπου Φ(.) η συνάρτηση κατανομής της τυποποιημένης κανονικής κατανομής. Η συνάρτηση διακινδύνευσης, $h(t) = \frac{f(t)}{S(t)}$, για την αντίστροφη Γκαουσιανή κατανομή μελετήθηκε από τους Chhikara and Folks (1978), οι οποίοι έδειξαν ότι πάντα πρώτα αυξάνεται και μετά μειώνεται (IDFR, increasingthen-decreasing-failure rate shape) για κάθε τιμή των παραμέτρων της κατανομής μ και λ. Αποδεικνύεται ότι

$$h(t) = \frac{3}{2t} + \frac{\lambda}{2t^2} - \frac{\lambda}{2\mu^2}$$

Παρατηρούμε ότι για μικρές τιμές του λόγου λ/μ το μέγιστο σημείο της συνάρτησης διακινδύνευσης συμβαίνει σχετικά νωρίς, στιγμή t^*, και έτσι το μεγαλύτερο ποσοστό των ατόμων επιβιώνουν μετά την χρονική στιγμή t^*. Συνεπώς, το γεγονός συμβαίνει νωρίς σε κάποιους αλλά μετά από αυτή το διάστημα ο κίνδυνος να συμβεί και στους υπόλοιπους μειώνεται. Αντίθετα, για μεγάλες τιμές του λόγου λ/μ έχει συμβεί το γεγονός σχεδόν σε όλα τα άτομα, δηλαδή πριν ακόμη η συνάρτηση διακινδύνευσης φτάσει στο μέγιστο σημείο. Άρα, έχουμε μία αύξηση του κινδύνου. Επομένως, παρόλο που η συνάρτηση διακινδύνευσης της αντίστροφης Γκαουσιανής κατανομής έχει πάντα μέγιστο, στην πράξη συμπεριφέρεται είτε σαν να μειώνεται είτε σαν να αυξάνεται. Γι' αυτό το λόγο και η αντίστροφη Γκαουσιανή κατανομή θεωρείται ευέλικτη ως προς το σχήμα της συνάρτησης διακινδύνευσης (Stogiannis et al., 2011).

Για τη συνάρτηση διακινδύνευσης αποδεικνύεται ότι το όριό της τείνει στο $\frac{\mu^2}{2}$ όπου μ η τάση. Οπότε έχουμε

α. το όριο, $\frac{\mu^2}{2}$, είναι ανεξάρτητο της αρχικής κατάστασης της διαδικασίας

β. για μεγάλο χρόνο t, ο λόγος $\frac{h(t;\mathbf{x_1})}{h(t;\mathbf{x_2})}$ είναι ανεξάρτητος του χρόνου με $\mathbf{x_1}$ και $\mathbf{x_2}$ τα διανύσματα συμμεταβλητών για δύο μονάδες

Από το α. συμπεραίνουμε ότι οι διαφορές στις αρχικές καταστάσεις της διαδικασίας μεταξύ των μονάδων χάνουν τη σημαντικότητά τους όσο περνά ο χρόνος. Αντίθετα, στο μοντέλο αναλογικής διακινδύνευσης η επιρροή των αρχικών καταστάσεων της διαδικασίας μεταξύ των μονάδων παραμένει η ίδια για πάντα. Αυτό είναι και ένα πλεονέκτημα των μοντέλων πρώτης μετάβασης έναντι των μοντέλων αναλογικής διακινδύνευσης, είναι πιο ρεαλιστικά.

Από το β. συμπεραίνουμε ότι για μεγάλο χρόνο t ισχύει η υπόθεση της αναλογικής διακινδύνευσης με τη σταθερά της αναλογίας να εξαρτάται μόνο από την παράμετρο μ (τάση) της διαδικασίας και όχι από τις αρχικές συνθήκες. Έχει αποδειχθεί ότι τα μοντέλα αναλογικής διακινδύνευσης είναι μία ειδική περίπτωση των μοντέλων πρώτης μετάβασης (Lee & Whitmore, 2010) (Stogiannis et al., 2011).

Το μοντέλο του Cox έχει πολλά πλεονεκτήματα και άρα όταν ικανοποιείται η υπόθεση της αναλογικής διακινδύνευσης πρέπει να χρησιμοποιείται. Παρόλα αυτά το μοντέλο παλινδρόμησης threshold έχει επίσης αρκετά οφέλη που οι ερευνητές πρέπει να μελετήσουν, όπως:

1. Αποδεικνύεται ότι μπορούν να κατασκευαστούν παραλλαγές του μοντέλου πρώτης μετάβασης ώστε να ικανοποιείται η υπόθεση της αναλογικής διακινδύνεσης. Αυτό σημαίνει ότι αν μπορούμε να τοποθετήσουμε το μοντέλο του Cox σε ένα πλαίσιο του μοντέλου πρώτης μετάβασης, τότε μπορούμε να δώσουμε μία ερμηνεία για τη βασική συνάρτηση διακινδύνευσης.

2. Όταν το μοντέλο πρώτης μετάβασης μπορεί να χρησιμοποιηθεί και οι συναρτήσεις επιβίωσης δεν ικανοποιούν την υπόθεση της αναλογικότητας τότε το μοντέλο του Cox θεωρείται ακατάλληλο. Παρόλα αυτά, το μοντέλο παλινδρόμησης threshold και το μοντέλο του Cox δεν διαφέρουν τόσο πολύ στο εύρος των δεδομένων ώστε να θεωρηθούν στατιστικά διαφορετικά.

3. Το μοντέλο παλινδρόμησης threshold είναι ένα μοντέλο που εξηγεί τα δεδομένα με τον ελάχιστο αριθμό μεταβλητών πρόβλεψης, κάτι που δεν το κάνει το μοντέλο του Cox. Το μοντέλο παλινδρόμησης threshold είναι παραμετρικό και έχει δύο σύνολα συντελεστών για τις συμμεταβλητές, τους συντελεστές που σχετίζονται με την παράμετρο $\ln(x_0)$ και αυτούς που σχετίζονται με την παράμετρο μ. Αντίθετα, το μοντέλο του Cox είναι ημι-παραμετρικό καθώς η βασική συνάρτηση διακινδύνευσης, h_0, είναι ακαθόριστη. Το μοντέλο του Cox περιλαμβάνει πολλές παραμέτρους για την εκτίμηση της h_0, όμως περισσότερη προσοχή δίνεται στην εκτίμηση του διανύσματος των συντελεστών παλινδρόμησης και όχι στην κατανόηση της απροσδιόριστης συνάρτησης h_0.

4. Με τη βοήθεια του μοντέλου παλινδρόμησης threshold οι ερευνητές μπορούν να απαντήσουν σε σημαντικά ερωτήματα όπως: Ποιες είναι οι συμμεταβλητές που επηρεάζουν την αρχική κατάσταση υγείας των ασθενών, $\ln(x_0)$, και ποιες επηρεάζουν την παράμετρο μ; Ποια είναι η στοχαστική διαδικασία και ποιο είναι το όριο (threshold) στο οποίο φτάνει; Αντίθετα, το μοντέλο του Cox δεν επιτρέπει στους ερευνητές να ψάξουν τόσο βαθιά.

5. Το μοντέλο του Cox συμπεριλαμβάνεται στα περισσότερα στατιστικά πακέτα και είναι εύκολο στη χρήση του. Μερικοί υποστηρίζουν πως το μοντέλο παλινδρόμησης threshold θα είναι δύσκολο να χρησιμοποιηθεί γιατί θα απαιτεί πολύ προγραμματισμό. Παρόλα αυτά στην R υπάρχει το πακέτο threg που πραγματοποιεί την προσαρμογή του μοντέλου threshold και είναι αρκετά εύχρηστο καθώς μέσω αυτού μπορούν να κατασκευαστούν οι γραφικές παραστάσεις των εκτιμήσεων των συναρτήσεων διακινδύνευσης, επιβίωσης και πυκνότητας πιθανότητας αλλά και να υπολογιστούν οι αναλογίες διακινδύνευσης. Στο παράρτημα Β, στο τέλος της εργασίας, φαίνεται το πακέτο threg και χρησιμοποιείται στο Κεφάλαιο 3, Παράγραφος 3.10 (Lee et al., 2010).

2.2 Παραδείγματα του μοντέλου παλινδρόμησης threshold

Ένα παράδειγμα της εφαρμογής του μοντέλου πρώτης μετάβασης είναι η μελέτη που έγινε με αφορμή την ανησυχία ότι η έκθεση των ανθρώπων στο πετρέλαιο κίνησης (diesel) μπορεί να αυξήσει τον κίνδυνο προσβολής από καρκίνο του πνεύμονα. Οι Lee et al (2004) αναφέρουν την κλινική μελέτη που πραγματοποιήθηκε σε άνδρες που δούλευαν σε βιομηχανία σιδηροδρόμων από 10 έως 20 χρόνια πριν το 1959. Κατέληξαν στο συμπέρασμα ότι οι άνθρωποι που ήταν στο προσωπικό για τη λειτουργία των τρένων (μηχανικοί, πυροσβέστες, εργολάβοι κ.τ.λ) το 1959 είχαν σημαντικά διαφορετική κατάσταση υγείας συγκριτικά με τους υπόλοιπους εργαζόμενους στη βιομηχανία των σιδηροδρόμων. Επίσης, οι εργάτες μεγαλύτερης ηλικίας είχαν χειρότερη κατάσταση υγείας από τους υπόλοιπους κάτι που ήταν αναμενόμενο. Αξιοσημείωτο είναι το γεγονός ότι η εκτιμήτρια της αρχικής κατάστασης υγείας και κατ' επέκταση και η εκτιμήτρια του μέσου χρόνου επιβίωσης για αυτούς που είχαν καρκίνο του πνεύμονα για το προσωπικό που δούλευε για τη

λειτουργία των τρένων ήταν 10% χαμηλότερη από την αντίστοιχη εκτιμήτρια των υπόλοιπων εργαζομένων. Τα ευρήματα της μελέτης αυτής έδωσαν το έναυσμα για περαιτέρω έρευνα πάνω στο συγκεκριμένο θέμα (Lee et al, 2004).

Μια επέκταση της προηγούμενης έρευνας έγινε αργότερα, το 2009, και είχε ως θέμα την προσαρμογή του μοντέλου παλινδρόμησης threshold για τη μελέτη του αυξημένου κινδύνου εμφάνισης καρκίνου του πνεύμονα αλλά και των ασθενειών του κυκλοφορικού συστήματος του προσωπικού που είχε εκτεθεί στο πετρέλαιο κίνησης. Στα δεδομένα αυτής της έρευνας υπήρχαν εργάτες που πέθαναν από καρκίνο του πνεύμονα, από ασθένειες του κυκλοφορικού συστήματος ή από άλλες αιτίες. Επίσης, υπήρχαν πληροφορίες για το αν οι εργάτες κάπνιζαν ή αν είχαν εκτεθεί σε αμίαντο. Κατέληξαν στο συμπέρασμα ότι η κατάσταση υγείας των μηχανικών των τρένων που πέθαναν από καρκίνο του πνεύμονα και από ασθένειες του κυκλοφορικού συστήματος επηρεάστηκε σημαντικά από το αν κάπνιζαν και από το αν είχαν εκτεθεί σε αμίαντο σε σχέση με τους υπόλοιπους εργαζόμενους (Lee et al, 2009).

Μία ακόμη εφαρμογή του μοντέλου παλινδρόμησης threshold έγινε σε έρευνα με σκοπό τον έλεγχο της σοβαρότητας της αρχικής βλάβης και του ρυθμού ανάρρωσης των εργαζομένων που επιστρέφουν στη δουλειά μετά από τραυματισμό των άκρων. Το μοντέλο threshold χρησιμοποιήθηκε με σκοπό να εκτιμηθεί ποιοι παράγοντες συνέλαβαν στην αρχική επιδείνωση του τραυματισμού των εργαζομένων αλλά και στο ρυθμό ανάρρωσής τους μετά τον τραυματισμό των άκρων. Το γεγονός που μελετήθηκε είναι το αν οι εργαζόμενοι επέστρεψαν στη δουλειά μετά τον τραυματισμό και αν επέστρεψαν πόσο ήταν το χρονικό διάστημα από τον τραυματισμό μέχρι την επιστροφή. Η έρευνα κατέληξε στο συμπέρασμα ότι οι νεότεροι σε ηλικία ασθενείς που είχαν τραυματιστεί στα άνω άκρα, είχαν δουλειά μερικής απασχόλησης, εκπαίδευση πάνω από 12 χρόνια και υψηλότερη αυτό-αποτελεσματικότητα, είχαν λιγότερες βλάβες μετά τον τραυματισμό τους συγκριτικά με τους υπόλοιπους. Ακόμη, οι νεότεροι εργαζόμενοι με τραυματισμούς στα άνω άκρα, δουλειά μερικής απασχόλησης και υψηλό επίπεδο μόρφωσης είχαν γρηγορότερη ανάρρωση και μεγαλύτερη πιθανότητα να επιστρέψουν στη δουλειά τους μετά τον τραυματισμό σε σύγκριση με τους υπόλοιπους (Hou et al, 2016).

ΚΕΦΑΛΑΙΟ 3

Στατιστική Ανάλυση

3.1 Πρόβλημα

Η μεταμόσχευση μυελού των οστών χρησιμοποιείται για τη θεραπεία της οξείας λευχαιμίας. Παρόλα αυτά, η ανάρρωση του ασθενή μετά τη μεταμόσχευση είναι μία πολύπλοκη διαδικασία επειδή εξαρτάται από πολλούς παράγοντες, όπως η ηλικία, το φύλο του ασθενή αλλά και του δότη κ.α. Σκοπός της παρούσας εργασίας είναι να εξετάσουμε κατά πόσο επηρεάζεται η επιβίωση 137 ασθενών με λευχαιμία μετά από μεταμόσχευση μυελού των οστών από συγκεκριμένους παράγοντες. Οι τύποι λευχαιμίας που μελετάμε είναι η οξεία λεμφοβλαστική λευχαιμία (ALL) και η οξεία μυελοκυτταρική λευχαιμία (AML low_risk και AML high_risk). Το γεγονός που μας ενδιαφέρει είναι η υποτροπή ή ο θάνατος του ασθενούς (Klein & Moeschberger, 2003). Οι μεταβλητές που χρησιμοποιούμε φαίνονται στον Πίνακα 3.1.

Πίνακας 3.1: Οι συμμεταβλητές του προβλήματος

ΜΕΤΑΒΛΗΤΕΣ
time: χρόνος σε μέρες μέχρι να συμβεί το γεγονός
indicator: 1: έχει συμβεί το γεγονός, 0: αλλιώς
group: τύπος λευχαιμίας (1=ALL, 2=AML low_risk, 3=AML high_risk)
recipient_age: ηλικία του ασθενή
donor_age: ηλικία του δότη
recipient_sex: φύλο του ασθενή (1=άντρας, 0=γυναίκα)
donor_sex: φύλο του δότη (1=άντρας, 0=γυναίκα)
recipient_cmv: κατάσταση του cmv του ασθενή (1=θετικό cmv, 0=αρνητικό cmv)
donor_cmv: κατάσταση του cmv του δότη (1=θετικό cmv, 0=αρνητικό cmv)
waiting_time: χρόνος αναμονής σε μέρες από τη διάγνωση μέχρι τη μεταμόσχευση
fab: 1=fab βαθμού 4 ή 5, 0=αλλιώς
mtx: 1=ναι, 0=όχι

Η μεταβλητή fab είναι μία ταξινόμηση των ασθενών με μυελοκυτταρική λευχαιμία (AML) που βασίζεται σε μορφολογικά κριτήρια. Οι ασθενείς με fab βαθμού 4 ή 5 (M4 ή M5) διατρέχουν μεγαλύτερο κίνδυνο υποτροπής ή θανάτου μετά τη μεταμόσχευση του μυελού των οστών. Η μεταβλητή mtx μας δείχνει το αν οι ασθενείς έλαβαν κάποια προφύλαξη για την ασθένεια μοσχεύματος έναντι ξενιστή (gvhd) μετά τη μεταμόσχευση ή όχι.

3.2 Έλεγχος καλής προσαρμογής του μοντέλου

Αρχικά, με τη βοήθεια του MINITAB υπολογίζουμε τις εκτιμήσεις Kaplan-Meier της συνάρτησης επιβίωσης.

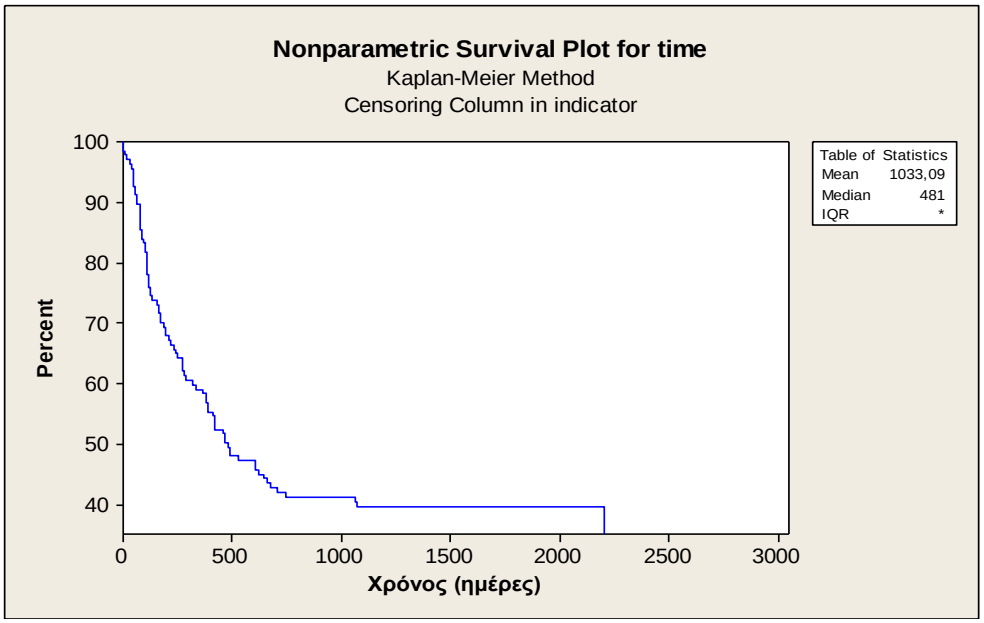
Αποτελέσματα 3.1: Εκτιμήσεις Kaplan-Meier της συνάρτησης επιβίωσης

Distribution Analysis: time

Kaplan-Meier Estimates

Time	Number at Risk	Number Failed	Survival Probability	Standard Error	95,0% Normal CI Lower	Upper
1	137	1	0,992701	0,0072726	0,978447	1,00000
2	136	1	0,985401	0,0102471	0,965318	1,00000
10	135	1	0,978102	0,0125035	0,953596	1,00000
16	134	1	0,970803	0,0143838	0,942611	0,99899
32	133	1	0,963504	0,0160211	0,932103	0,99490
35	132	1	0,956204	0,0174836	0,921937	0,99047
47	131	2	0,941606	0,0200336	0,902341	0,98087
48	129	2	0,927007	0,0222239	0,883449	0,97057
53	127	1	0,919708	0,0232167	0,874204	0,96521
55	126	1	0,912409	0,0241526	0,865070	0,95975
63	125	1	0,905109	0,0250381	0,856036	0,95418
64	124	1	0,897810	0,0258783	0,847090	0,94853
74	123	2	0,883212	0,0274392	0,829432	0,93699
76	121	1	0,875912	0,0281666	0,820707	0,93112
79	120	1	0,868613	0,0288622	0,812044	0,92518
80	119	2	0,854015	0,0301666	0,794889	0,91314
84	117	1	0,846715	0,0307792	0,786389	0,90704
86	116	1	0,839416	0,0313675	0,777937	0,90090
93	115	1	0,832117	0,0319327	0,769530	0,89470
100	114	1	0,824818	0,0324761	0,761166	0,88847
104	113	1	0,817518	0,0329988	0,752842	0,88219
105	112	2	0,802920	0,0339858	0,736309	0,86953
107	110	1	0,795620	0,0344518	0,728096	0,86314
109	109	1	0,788321	0,0349004	0,719918	0,85672
110	108	1	0,781022	0,0353323	0,711772	0,85027
113	107	1	0,773723	0,0357481	0,703658	0,84379
115	106	1	0,766423	0,0361484	0,695574	0,83727
120	105	1	0,759124	0,0365336	0,687519	0,83073
122	104	2	0,744526	0,0372609	0,671496	0,81756
129	102	1	0,737226	0,0376037	0,663524	0,81093
157	101	1	0,729927	0,0379332	0,655579	0,80427
162	100	1	0,722628	0,0382497	0,647660	0,79760
164	99	1	0,715328	0,0385536	0,639765	0,79089
168	98	1	0,708029	0,0388450	0,631894	0,78416
172	97	1	0,700730	0,0391243	0,624048	0,77741
183	96	1	0,693431	0,0393918	0,616224	0,77064

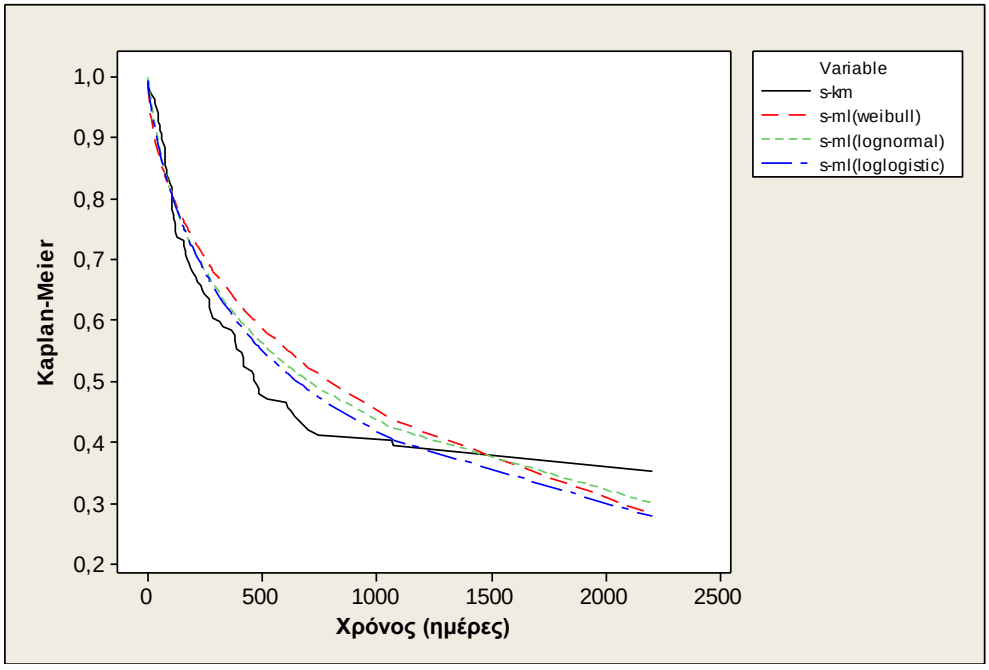
192	95	1	0,686131	0,0396476	0,608423	0,76384
194	94	1	0,678832	0,0398921	0,600645	0,75702
211	93	1	0,671533	0,0401254	0,592889	0,75018
219	92	1	0,664234	0,0403477	0,585154	0,74331
230	90	1	0,656853	0,0405688	0,577340	0,73637
242	89	1	0,649473	0,0407788	0,569548	0,72940
248	88	1	0,642092	0,0409778	0,561777	0,72241
268	87	1	0,634712	0,0411660	0,554028	0,71540
272	86	1	0,627332	0,0413437	0,546300	0,70836
273	85	1	0,619951	0,0415108	0,538592	0,70131
276	84	1	0,612571	0,0416675	0,530904	0,69424
288	83	1	0,605191	0,0418140	0,523237	0,68714
318	82	1	0,597810	0,0419504	0,515589	0,68003
332	81	1	0,590430	0,0420767	0,507961	0,67290
363	80	1	0,583049	0,0421930	0,500353	0,66575
381	79	1	0,575669	0,0422995	0,492764	0,65857
383	78	1	0,568289	0,0423962	0,485194	0,65138
390	77	2	0,553528	0,0425604	0,470111	0,63694
414	75	1	0,546148	0,0426280	0,462598	0,62970
418	74	1	0,538767	0,0426861	0,455104	0,62243
421	73	1	0,531387	0,0427346	0,447629	0,61515
422	72	1	0,524006	0,0427736	0,440172	0,60784
456	71	1	0,516626	0,0428032	0,432733	0,60052
466	70	1	0,509246	0,0428232	0,425314	0,59318
467	69	1	0,501865	0,0428339	0,417912	0,58582
481	68	1	0,494485	0,0428351	0,410530	0,57844
486	67	1	0,487105	0,0428268	0,403166	0,57104
487	66	1	0,479724	0,0428092	0,395820	0,56363
526	65	1	0,472344	0,0427820	0,388493	0,55620
606	63	1	0,464846	0,0427549	0,381048	0,54864
609	62	1	0,457349	0,0427176	0,373624	0,54107
625	61	1	0,449851	0,0426702	0,366219	0,53348
641	60	1	0,442354	0,0426126	0,358835	0,52587
662	59	1	0,434856	0,0425448	0,351470	0,51824
677	58	1	0,427359	0,0424668	0,344125	0,51059
704	57	1	0,419861	0,0423784	0,336801	0,50292



Σχήμα 3.1: Οι εκτιμήσεις Kaplan-Meier της συνάρτησης επιβίωσης

Στα Αποτελέσματα 3.1 βλέπουμε τις τιμές της εκτιμήτριας Kaplan-Meier στη στήλη "survival probability", τον αριθμό των ασθενών που κινδυνεύουν να τους συμβεί το γεγονός (θάνατος ή υποτροπή) στη στήλη "number at risk", τον αριθμό των ασθενών που τους συνέβη το γεγονός τη χρονική στιγμή "time" στη στήλη "number failed", τα τυπικά σφάλματα των εκτιμητριών στη στήλη "standard error" και τέλος τα 95% διαστήματα εμπιστοσύνης για τις εκτιμήτριες. Ακόμη, στο Σχήμα 3.1 παρατηρούμε τις εκτιμήσεις Kaplan-Meier συναρτήσει του χρόνου και προκύπτει μία κλιμακωτή συνάρτηση όπως υποδεικνύει και η θεωρία.

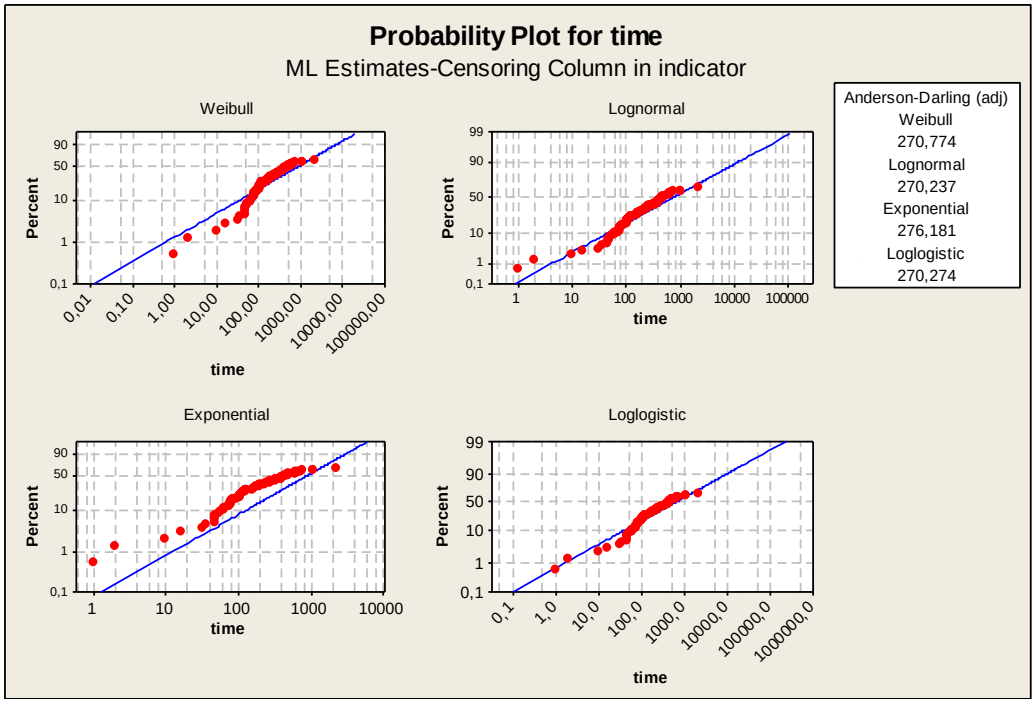
Στη συνέχεια βρίσκουμε τις τιμές της εκτιμήτριας μεγίστης πιθανοφάνειας της συνάρτησης επιβίωσης για τις κατανομές Weibull, Log-normal και Log-logistic και τις συγκρίνουμε με τις τιμές της εκτιμήτριας Kaplan-Meier με σκοπό να ελέγξουμε ποια κατανομή ακολουθούν τα δεδομένα μας.



Σχήμα 3.2: Οι εκτιμήσεις Kaplan-Meier σε σύγκριση με τις εκτιμήσεις μεγίστης πιθανοφάνειας των κατανομών Weibull, Log-normal και Log-logistic συναρτήσει του χρόνου

Από τα Σχήμα 3.2 παρατηρούμε ότι οι εκτιμήσεις μεγίστης πιθανοφάνειας της συνάρτησης επιβίωσης για την κατανομή Log-logistic είναι πιο κοντά στις τιμές της

εκτιμήτριας Kaplan-Meier συγκριτικά με τις τιμές της εκτιμήτριας μεγίστης πιθανοφάνειας για τις κατανομές Log-normal και Weibull.



Σχήμα 3.3: Γράφημα πιθανοτήτων της προσαρμοσμένης με τη μέθοδο μεγίστης πιθανοφάνειας συνάρτησης επιβίωσης των κατανομών Weibull, Log-normal, Εκθετικής και Log-logistic αντίστοιχα έναντι του χρόνου

Πίνακας 3.2: Οι τιμές της ελεγχοσυνάρτησης Anderson- Darling για τις τέσσερις κατανομές

Κατανομή	Διορθωμένη ελεγχοσυνάρτηση Anderson- Darling
Weibull	270,774
Log-normal	270,237
Εκθετική	276,181
Log-logistic	270,274

Ακόμη, στον Πίνακα 3.2 παρατηρούμε τις τιμές για την ελεγχοσυνάρτηση Anderson-Darling (AD). Από τη θεωρία γνωρίζουμε ότι το καταλληλότερο μοντέλο είναι αυτό με τη μικρότερη τιμή AD, στην περίπτωσή μας είναι αυτό υπό την

κατανομή Log-normal με οριακή διαφορά από το μοντέλο της Log-logistic κατανομής. Από τα διαγράμματα όμως βλέπουμε ότι το γράφημα πιθανοτήτων για την Log-logistic είναι καλύτερο σε σύγκριση με τις υπόλοιπες κατανομές. Συνεπώς, φαίνεται ότι τα δεδομένα μας είναι πιο κοντά στη Log-logistic κατανομή.

3.3 Εκτιμήσεις Kaplan-Meier για τους τρεις τύπους λευχαιμίας

Σε αυτή την ενότητα υπολογίζουμε ξεχωριστά τις εκτιμήτριες Kaplan-Meier της συνάρτησης επιβίωσης για τους τρεις τύπους λευχαιμίας (ALL, AML low_risk, AML high_risk) και ελέγχουμε τυχόν διαφορές μεταξύ αυτών των τύπων. Στα Αποτελέσματα 3.2, 3.3 και 3.4 παρατηρούμε τις εκτιμήτριες Kaplan-Meier, τα τυπικά σφάλματα και τα διαστήματα εμπιστοσύνης των συναρτήσεων επιβίωσης για τις τρεις διαφορετικές ομάδες λευχαιμίας αντίστοιχα. Στο Σχήμα 3.4 συγκρίνουμε τις εκτιμήτριες Kaplan-Meier των τριών τύπων λευχαιμίας και συμπεραίνουμε ότι η ομάδα 3 (AML high_risk) είναι αυτή υπό την οποία το γεγονός (υποτροπή ή θάνατος) συμβαίνει πιο γρήγορα στους ασθενείς, ενώ η ομάδα 2 (AML low_risk) είναι αυτή που έχει την καλύτερη πρόγνωση για τους ασθενείς.

Αποτελέσματα 3.2: Εκτιμήσεις Kaplan-Meier για την ομάδα ALL

Kaplan-Meier Estimates

Time	Number at Risk	Number Failed	Survival Probability	Standard Error	95,0% Normal CI Lower	Upper
1	38	1	0,973684	0,0259672	0,922789	1,00000
55	37	1	0,947368	0,0362235	0,876372	1,00000
74	36	1	0,921053	0,0437441	0,835316	1,00000
86	35	1	0,894737	0,0497845	0,797161	0,99231
104	34	1	0,868421	0,0548361	0,760944	0,97590
107	33	1	0,842105	0,0591528	0,726168	0,95804
109	32	1	0,815789	0,0628861	0,692535	0,93904
110	31	1	0,789474	0,0661348	0,659852	0,91910
122	30	2	0,736842	0,0714338	0,596834	0,87685
129	28	1	0,710526	0,0735704	0,566331	0,85472
172	27	1	0,684211	0,0754053	0,536419	0,83200
192	26	1	0,657895	0,0769602	0,507055	0,80873
194	25	1	0,631579	0,0782518	0,478208	0,78495
230	23	1	0,604119	0,0795218	0,448259	0,75998
276	22	1	0,576659	0,0805088	0,418865	0,73445
332	21	1	0,549199	0,0812232	0,390005	0,70839
383	20	1	0,521739	0,0816721	0,361665	0,68181
418	19	1	0,494279	0,0818598	0,333837	0,65472
466	18	1	0,466819	0,0817882	0,306517	0,62712
487	17	1	0,439359	0,0814566	0,279707	0,59901
526	16	1	0,411899	0,0808617	0,253413	0,57039
609	14	1	0,382478	0,0802600	0,225171	0,53978
662	13	1	0,353057	0,0792956	0,197640	0,50847

Αποτελέσματα 3.3: Εκτιμήσεις Kaplan-Meier για την ομάδα AML low_risk

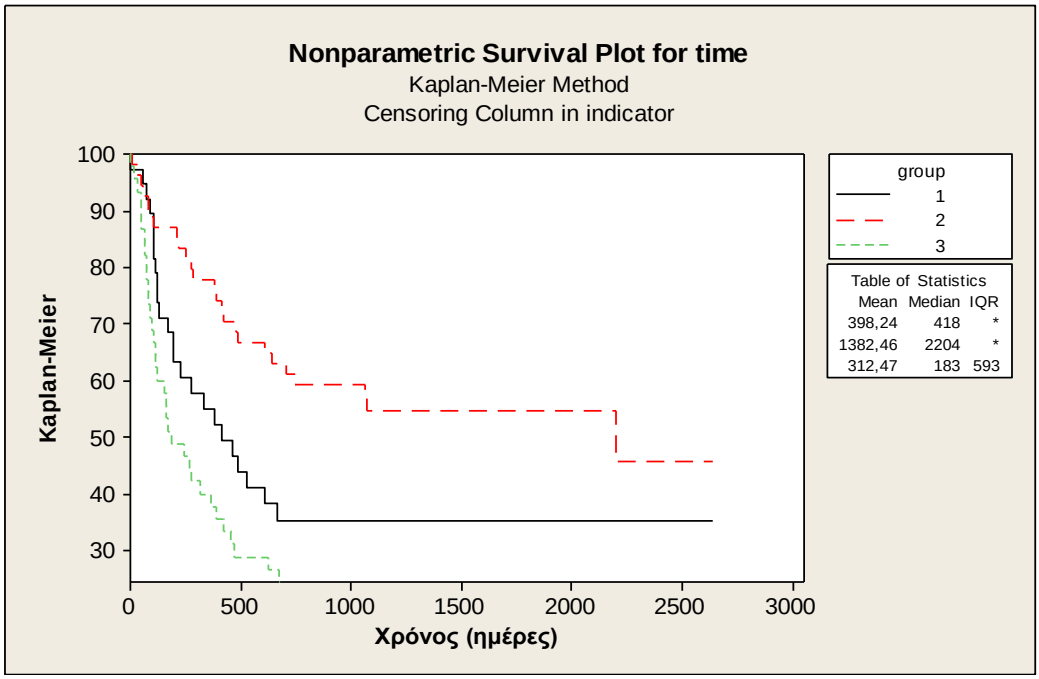
Kaplan-Meier Estimates

Time	Number at Risk	Number Failed	Survival Probability	Standard Error	95,0% Normal CI Lower	Upper
10	54	1	0,981481	0,018346	0,945523	1,00000
35	53	1	0,962963	0,025700	0,912593	1,00000
48	52	1	0,944444	0,031171	0,883350	1,00000
53	51	1	0,925926	0,035639	0,856075	0,99578
79	50	1	0,907407	0,039445	0,830097	0,98472
80	49	1	0,888889	0,042767	0,805068	0,97271
105	48	1	0,870370	0,045710	0,780781	0,95996
211	47	1	0,851852	0,048343	0,757101	0,94660
219	46	1	0,833333	0,050715	0,733934	0,93273
248	45	1	0,814815	0,052861	0,711209	0,91842
272	44	1	0,796296	0,054807	0,688876	0,90372
288	43	1	0,777778	0,056575	0,666893	0,88866
381	42	1	0,759259	0,058180	0,645229	0,87329
390	41	1	0,740741	0,059635	0,623858	0,85762
414	40	1	0,722222	0,060952	0,602759	0,84169
421	39	1	0,703704	0,062139	0,581914	0,82549
481	38	1	0,685185	0,063203	0,561310	0,80906
486	37	1	0,666667	0,064150	0,540935	0,79240
606	36	1	0,648148	0,064986	0,520778	0,77552
641	35	1	0,629630	0,065715	0,500831	0,75843
704	34	1	0,611111	0,066340	0,481087	0,74114
748	33	1	0,592593	0,066865	0,461541	0,72364
1063	26	1	0,569801	0,068067	0,436393	0,70321
1074	25	1	0,547009	0,069055	0,411664	0,68235
2204	6	1	0,455840	0,101182	0,257527	0,65415

Αποτελέσματα 3.4: Εκτιμήσεις Kaplan-Meier για την ομάδα AML high_risk

Kaplan-Meier Estimates

Time	Number at Risk	Number Failed	Survival Probability	Standard Error	95,0% Normal CI Lower	Upper
2	45	1	0,977778	0,0219739	0,934710	1,00000
16	44	1	0,955556	0,0307207	0,895344	1,00000
32	43	1	0,933333	0,0371849	0,860452	1,00000
47	42	2	0,888889	0,0468486	0,797067	0,98071
48	40	1	0,866667	0,0506745	0,767347	0,96599
63	39	1	0,844444	0,0540284	0,738551	0,95034
64	38	1	0,822222	0,0569937	0,710517	0,93393
74	37	1	0,800000	0,0596285	0,683130	0,91687
76	36	1	0,777778	0,0619748	0,656309	0,89925
80	35	1	0,755556	0,0640644	0,629992	0,88112
84	34	1	0,733333	0,0659218	0,604129	0,86254
93	33	1	0,711111	0,0675660	0,578684	0,84354
100	32	1	0,688889	0,0690122	0,553627	0,82415
105	31	1	0,666667	0,0702728	0,528934	0,80440
113	30	1	0,644444	0,0713576	0,504586	0,78430
115	29	1	0,622222	0,0722744	0,480567	0,76388
120	28	1	0,600000	0,0730297	0,456864	0,74314
157	27	1	0,577778	0,0736283	0,433469	0,72209
162	26	1	0,555556	0,0740741	0,410373	0,70074
164	25	1	0,533333	0,0743698	0,387571	0,67910
168	24	1	0,511111	0,0745172	0,365060	0,65716
183	23	1	0,488889	0,0745172	0,342838	0,63494
242	22	1	0,466667	0,0743698	0,320905	0,61243
268	21	1	0,444444	0,0740741	0,299262	0,58963
273	20	1	0,422222	0,0736283	0,277913	0,56653
318	19	1	0,400000	0,0730297	0,256864	0,54314
363	18	1	0,377778	0,0722744	0,236122	0,51943
390	17	1	0,355556	0,0713576	0,215697	0,49541
422	16	1	0,333333	0,0702728	0,195601	0,47107
456	15	1	0,311111	0,0690122	0,175850	0,44637
467	14	1	0,288889	0,0675660	0,156462	0,42132
625	13	1	0,266667	0,0659218	0,137462	0,39587
677	12	1	0,244444	0,0640644	0,118880	0,37001



Σχήμα 3.4: Σύγκριση των εκτιμητριών Kaplan-Meier των τριών τύπων λευχαιμίας, group1=ALL, group2=AML low_risk, group3=AML high_risk

3.4 Έλεγχος Log-rank

Στη συνέχεια, πραγματοποιούμε τον έλεγχο Log-rank: H_0: $S_1=S_2=S_3$ vs H_1: $S_1 \neq S_2 \neq S_3$, όπου S_1, S_2, S_3 οι συναρτήσεις επιβίωσης για τις τρεις ομάδες λευχαιμίας, με σκοπό να ελέγξουμε αν υπάρχουν διαφοροποιήσεις μεταξύ αυτών των συναρτήσεων. Από τα Αποτελέσματα 3.5 του MINITAB λαμβάνουμε ότι η ελεγχοσυνάρτηση Log-rank είναι $\frac{u^2}{v} = 13{,}8037$ και ακολουθεί την X^2 κατανομή με δύο βαθμούς ελευθερίας. Επιπλέον, η p-value=0,001 είναι πολύ μικρή και άρα απορρίπτουμε την μηδενική υπόθεση. Αυτό σημαίνει ότι υπάρχουν διαφοροποιήσεις στις συναρτήσεις επιβίωσης των τριών ομάδων λευχαιμίας, δηλαδή η επιβίωση των ασθενών μετά τη μεταμόσχευση εξαρτάται από τον τύπο της λευχαιμίας από την οποία πάσχουν. Κάτι που επιβεβαιώνεται και από το Σχήμα 3.4 που απεικονίζει τις γραφικές παραστάσεις των εκτιμητριών Kaplan-Meier για τις τρεις ομάδες συναρτήσει του χρόνου. Τέλος, παρατηρούμε ότι στα αποτελέσματα του MINITAB εμφανίζεται και ο έλεγχος Wilcoxon ο οποίος είναι μία επέκταση του Log-rank όπως αναφέραμε στο Κεφάλαιο 1, Παράγραφος 1.3.4. Η ελεγχοσυνάρτηση του Wilcoxon ισούται με 16,2407 και η p-

value<0,001 οπότε και οι δύο έλεγχοι καταλήγουν στο ίδιο συμπέρασμα, υπάρχουν στατιστικά σημαντικές διαφοροποιήσεις στις συναρτήσεις επιβίωσης των τριών ομάδων λευχαιμίας.

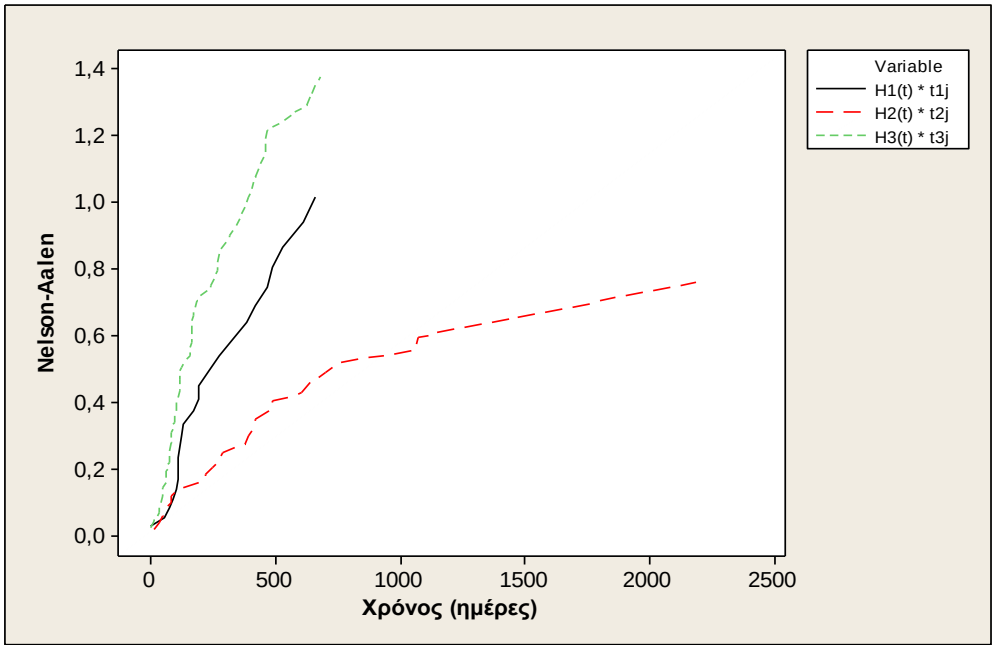
Αποτελέσματα 3.5: Αποτελέσματα του ελέγχου Log-rank

```
Test Statistics

Method     Chi-Square  DF  P-Value
Log-Rank      13,8037   2    0,001
Wilcoxon      16,2407   2    0,000
```

3.5 Εκτιμήσεις Nelson-Aalen

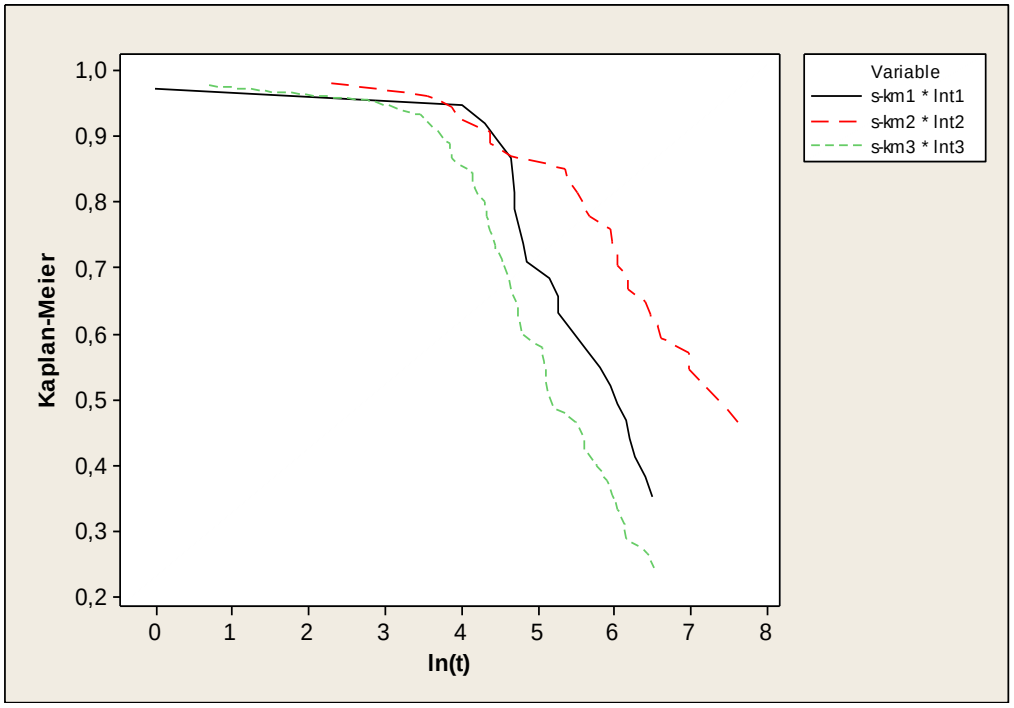
Όπως αναφέραμε και στη θεωρία η σωρευτική συνάρτηση διακινδύνευσης, H(t), εκτιμάται με την εκτιμήτρια Nelson-Aalen από τον τύπο $\hat{H}(t) = \sum_{j:\, t(j)\le t} \frac{d_j}{n_j}$. Στο Σχήμα 3.5 βλέπουμε τις εκτιμήτριες αυτές συναρτήσει του χρόνου και παρατηρούμε ότι μεγαλύτερο κίνδυνο διατρέχουν οι ασθενείς που πάσχουν από τη λευχαιμία της ομάδας 3 (AML high_risk). Στο ίδιο συμπέρασμα είχαμε καταλήξει και στην Παράγραφο 3.3 που συγκρίναμε τις συναρτήσεις επιβίωσης των τριών ομάδων.



Σχήμα 3.5: Οι εκτιμήσεις Nelson-Aalen συναρτήσει του χρόνου

3.6 Προσαρμογή του μοντέλου επιταχυνόμενης διακοπής

Πριν την προσαρμογή του μοντέλου επιταχυνόμενης διακοπής κάνουμε τον γραφικό έλεγχο για να ελέγξουμε αν είναι κατάλληλο αυτό το μοντέλο (όπως αναφέραμε στο Κεφάλαιο 1, Παράγραφος 1.4). Για την πραγματοποίηση αυτού του ελέγχου χρειαζόμαστε τις εκτιμήτριες Kaplan-Meier της συνάρτησης επιβίωσης για τους τρεις τύπους λευχαιμίας που υπολογίσαμε στην Παράγραφο 3.3. Στο Σχήμα 3.6 παρατηρούμε το γράφημα των εκτιμητριών αυτών ως το φυσικό λογάριθμο του χρόνου αντίστοιχα για τις τρεις ομάδες. Για να ισχύει η υπόθεση της επιταχυνόμενης διακοπής πρέπει οι καμπύλες να διαφέρουν μόνο ως προς τις οριζόντιες μετατοπίσεις.



Σχήμα 3.6: Γραφικός έλεγχος του μοντέλου επιταχυνόμενης διακοπής

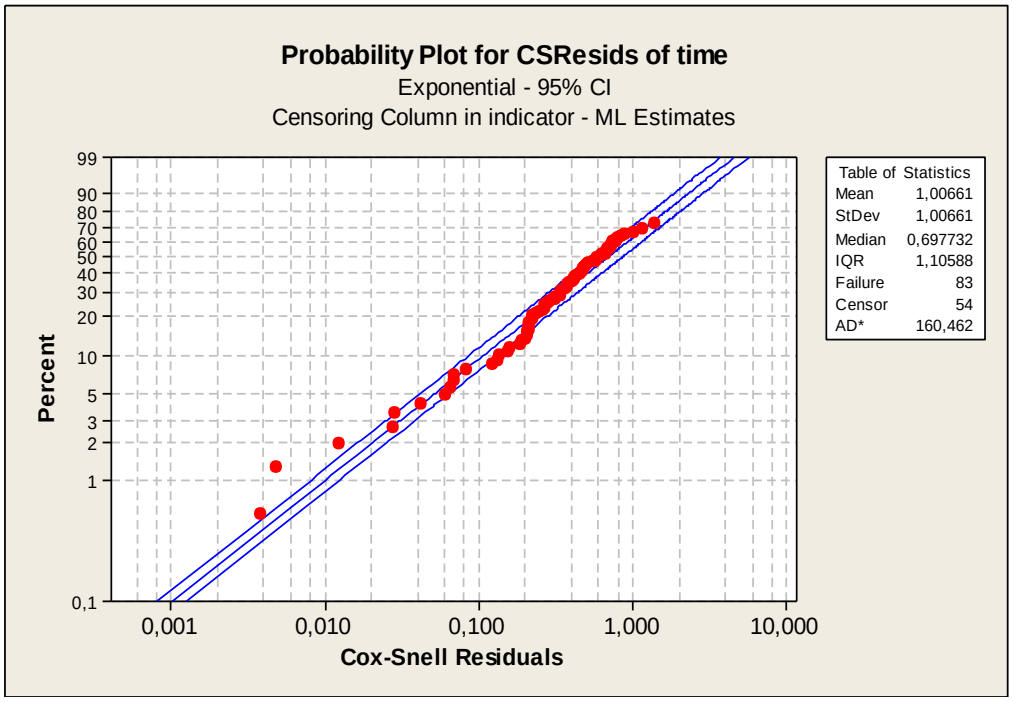
Η μεταβλητή groupf χωρίζεται σε groupf1 και groupf2 όπου

$$\text{groupf1}=\begin{cases}1, & \text{όταν } group = 1\ (ALL)\\ 0, & \text{αλλιώς}\end{cases}$$

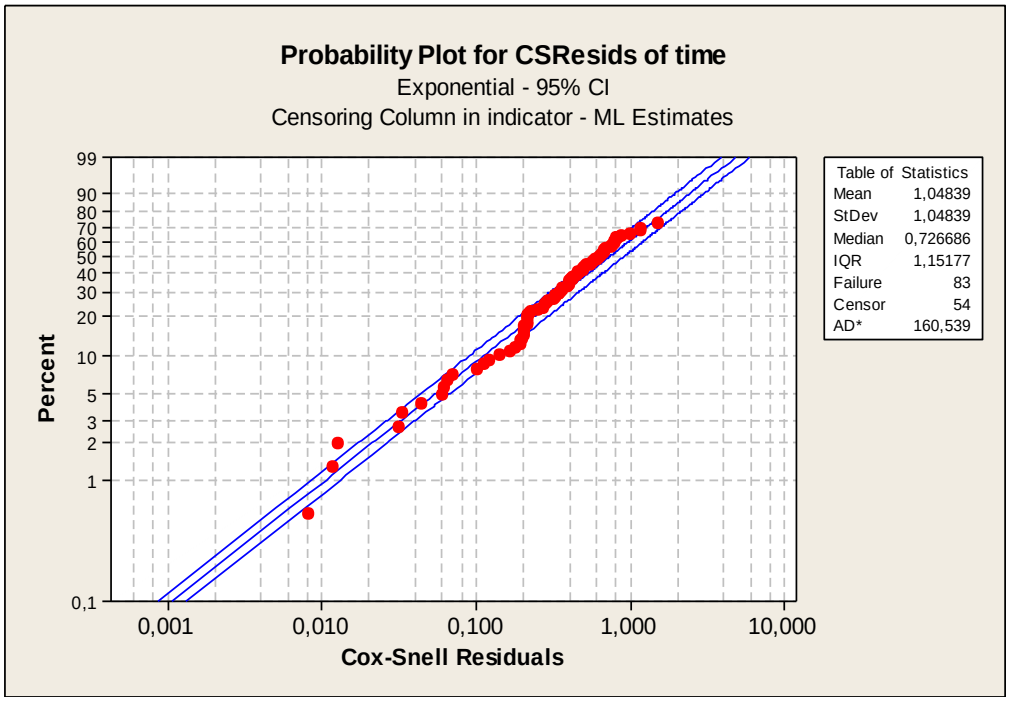
$$\text{groupf2}=\begin{cases}1, & \text{όταν } group = 2(AML\ low\_risk)\\ 0, & \text{αλλιώς}\end{cases},$$

δηλαδή η κατηγορία 3 (AML high_risk) της μεταβλητής group είναι η κατηγορία αναφοράς.

Προσαρμόζουμε το μοντέλο της επιταχυνόμενης διακοπής υπό την κατανομή Log-logistic αλλά και υπό την Log-normal. Στα Σχήματα 3.7 και 3.8 βλέπουμε τις γραφικές παραστάσεις των υπολοίπων Cox-Snell για την Log-normal και την Log-logistic αντίστοιχα. Παρατηρούμε ότι το διάγραμμα των υπολοίπων για την Log-logistic είναι καλύτερο. Συνεπώς, μας ενδιαφέρει η προσαρμογή του μοντέλου υπό την Log-logistic κατανομή.



Σχήμα 3.7: Γραφικός έλεγχος για τα υπόλοιπα Cox-Snell υπό την Log-normal κατανομή



Σχήμα 3.8: Γραφικός έλεγχος για τα υπόλοιπα Cox-Snell υπό την Log-logistic κατανομή

Μετά την προσαρμογή του μοντέλου υπό την Log-logistic κατανομή πρέπει να αποφασίσουμε ποιες συμμεταβλητές είναι στατιστικά σημαντικές κάνοντας χρήση κάποιων ελέγχων. Αρχικά, στον Πίνακα 3.3 βλέπουμε τους συντελεστές των συμμεταβλητών, τα τυπικά σφάλματα, το στατιστικό ελέγχου z και την p-τιμή του ελέγχου Wald. Για κάθε μία από τις συμμεταβλητές γίνεται ο έλεγχος Wald

$$H_0: \beta_i=0 \text{ vs } H_1: \beta_i\neq 0$$

με την ελεγχοσυνάρτηση να είναι $z = \frac{\beta_i}{se(\beta_i)}$ και ακολουθεί την τυποποιημένη κανονική κατανομή, N(0,1), υπό την H_0. Παρατηρούμε από τα αποτελέσματα του Πίνακα 3.3 ότι οι συμμεταβλητές fab, groupf και mtx είναι στατιστικά σημαντικές καθώς έχουν τις μικρότερες p-τιμές.

Πίνακας 3.3: Το μοντέλο της επιταχυνόμενης διακοπής

	συντελεστής	τυπικό σφάλμα	z	p-value
σταθερά	6.794	0.801	8.487	<0.001
recipient_age	-0.011	0.030	-0.359	0.720
donor_age	-0.011	0.028	-0.383	0.701
recipient_sex	0.267	0.358	0.746	0.455
donor_sex	0.084	0.364	0.231	0.817
recipient_cmv	0.282	0.386	0.731	0.465
donor_cmv	-0.068	0.370	-0.185	0.853
groupf1	-0.252	0.539	-0.468	0.640
groupf2	1.414	0.431	3.283	0.001
waiting_time	0.001	0.001	0.953	0.340
fab	-1.264	0.427	-2.956	0.003
mtx	-0.849	0.416	-2.040	0.041

Επιπλέον, μπορούμε να κάνουμε και έναν έλεγχο της πιθανοφάνειας για τη σύγκριση του μοντέλου με τις κατά Wald στατιστικά σημαντικές συμμεταβλητές εναντίον του μοντέλου που περιλαμβάνει όλες τις συμμεταβλητές:

H_0: το μοντέλο με τις συμμεταβλητές groupf, fab, mtx

H_1: το μοντέλο με όλες τις συμμεταβλητές

η ελεγχοσυνάρτηση είναι $-2(\widehat{l_0} - \widehat{l_1})$, όπου $\widehat{l_0}$ η μεγιστοποιημένη λογαριθμοποιημένη συνάρτηση πιθανοφάνειας υπό την H_0 και $\widehat{l_1}$ η μεγιστοποιημένη λογαριθμοποιημένη συνάρτηση πιθανοφάνειας υπό την H_1. Από τη προσαρμογή των δύο μοντέλων έχουμε τα αποτελέσματα του Πίνακα 3.4.

Πίνακας 3.4: Έλεγχος του λόγου των πιθανοφανειών

	$\hat{l}$
Μοντέλο με συμμεταβλητές groupf, fab, mtx	-637,664
Μοντέλο με όλες τις συμμεταβλητές	-636,018
$-2(\widehat{l_0} - \widehat{l_1})$	3,29

Η p-τιμή του ελέγχου αυτού είναι: $P(X_7^2 > 3.29) = 0.85$ Συνεπώς, δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση και άρα το μοντέλο με τις συμμεταβλητές groupf, fab και mtx είναι καλύτερο από εκείνο με όλες τις συμμεταβλητές.

Ακόμη, για την επιλογή των στατιστικά σημαντικών συμμεταβλητών μπορούμε να χρησιμοποιήσουμε τη διαδικασία της διαδοχικής αφαίρεσης (backward elimination) των μη σημαντικών μεταβλητών ξεκινώντας από το μοντέλο που περιλαμβάνει όλες τις συμμεταβλητές. Η διαδικασία της διαδοχικής αφαίρεσης μπορεί να γίνει με βάση την τιμή του $-2(\widehat{l_0} - \widehat{l_1}) \sim X_1^2$ όπου αφαιρούμε τη συμμεταβλητή εκείνη που έχει τη μεγαλύτερη p-value, δηλαδή εκείνη που δεν είναι στατιστικά σημαντική. Στον Πίνακα 3.5 βλέπουμε τις τιμές των p-value στο πρώτο βήμα της διαδοχικής αφαίρεσης. Παρατηρούμε ότι σε αυτό το πρώτο βήμα η συμμεταβλητή donor_cmv έχει την μεγαλύτερη p-τιμή και άρα αφαιρείται πρώτη από το μοντέλο. Συνεχίζουμε τη διαδικασία για τις υπόλοιπες 9 συμμεταβλητές και καταλήγουμε στο τελικό μοντέλο που φαίνεται στον Πίνακα 3.6 με τις μεταβλητές mtx, fab και groupf να προκύπτουν στατιστικά σημαντικές. Στο ίδιο αποτέλεσμα καταλήξαμε και προηγουμένως με τον έλεγχο Wald και τον έλεγχο της πιθανοφάνειας.

Πίνακας 3.5: Το πρώτο βήμα της διαδικασίας της διαδοχικής αφαίρεσης

Αφαίρεση συμμεταβλητής	Μεταβολή της $-2\hat{l}$	p-value
donor_cmv	0.034	0.853
donor_sex	0.053	0.817
recipient_age	0.129	0.720
donor_age	0.147	0.701
recipient_cmv	0.533	0.466
recipient_sex	0.553	0.457
waiting_time	0.965	0.326
mtx	4.085	0.043*
fab	8.690	0.003**
groupf	15.060	0.0005***

Πίνακας 3.6: Το τελευταίο βήμα της διαδικασίας της διαδοχικής αφαίρεσης

Αφαίρεση συμμεταβλητής	Μεταβολή της $-2\hat{l}$	p-value
mtx	5.118	0.024*
fab	7.696	0.006**
groupf	13.981	0.0009***

Οι αστερίσκοι που υπάρχουν δίπλα στις p-τιμές δείχνουν ότι αυτές οι p-τιμές είναι αρκετά μικρές και άρα οι αντίστοιχες συμμεταβλητές είναι στατιστικά σημαντικές.

Ένας άλλος τρόπος για την επιλογή του καλύτερου μοντέλου είναι με βάση το κριτήριο AIC. Αν η αφαίρεση μιας συμμεταβλητής οδηγεί στην μείωση της τιμής του AIC, τότε αυτό σημαίνει ότι η συμμεταβλητή αυτή δεν είναι στατιστικά σημαντική καθώς το μοντέλο με το μικρότερο AIC είναι και το καλύτερο. Δουλεύοντας με το κριτήριο AIC λαμβάνουμε τα αποτελέσματα του Πίνακα 3.7. Παρατηρούμε ότι το μοντέλο με το μικρότερο AIC είναι αυτό με τις συμμεταβλητές mtx, fab και groupf, AIC=1287.3. Το συμπέρασμα αυτό συμβαδίζει με αυτό που καταλήξαμε χρησιμοποιώντας τους άλλους ελέγχους.

Πίνακας 3.7: Έλεγχος με το κριτήριο AIC

Συμμεταβλητές στο μοντέλο	**AIC**
όλες	1298
όλες εκτός της donor_cmv	1296.1
recipient_age, donor_age, recipient_cmv, recipient_sex, waiting_time, fab, groupf, mtx	1294.1
recipient_cmv, recipient_sex, waiting_time, donor_age, fab, groupf, mtx	1292.3
recipient_sex, donor_age, waiting_time, fab, groupf, mtx	1290.7
donor_age, waiting_time, fab, groupf, mtx	1289.4
waiting_time, fab, groupf, mtx	1288.2
fab, groupf, mtx	1287.3
fab	1312.8
groupf	1309.2
groupf, fab	1300.2
groupf, mtx	1308.9
fab, mtx	1311
mtx	1318.5

Καταλήξαμε, λοιπόν ότι το καλύτερο μοντέλο επιταχυνόμενης διακοπής είναι εκείνο με τις μεταβλητές fab, groupf, mtx, όποτε από τη προσαρμογή του μοντέλου αυτού έχουμε τις εκτιμήσεις των συντελεστών και τα τυπικά τους σφάλματα στον Πίνακα 3.8.

Πίνακας 3.8: Το τελικό μοντέλο της επιταχυνόμενης διακοπής

	Συντελεστής	Τυπικό σφάλμα
Σταθερά $\widehat{\beta_0}$	6.581	0.417
groupf1	0.020	0.503
groupf2	1.395	0.419
fab	-1.179	0.424
mtx	-0.892	0.390

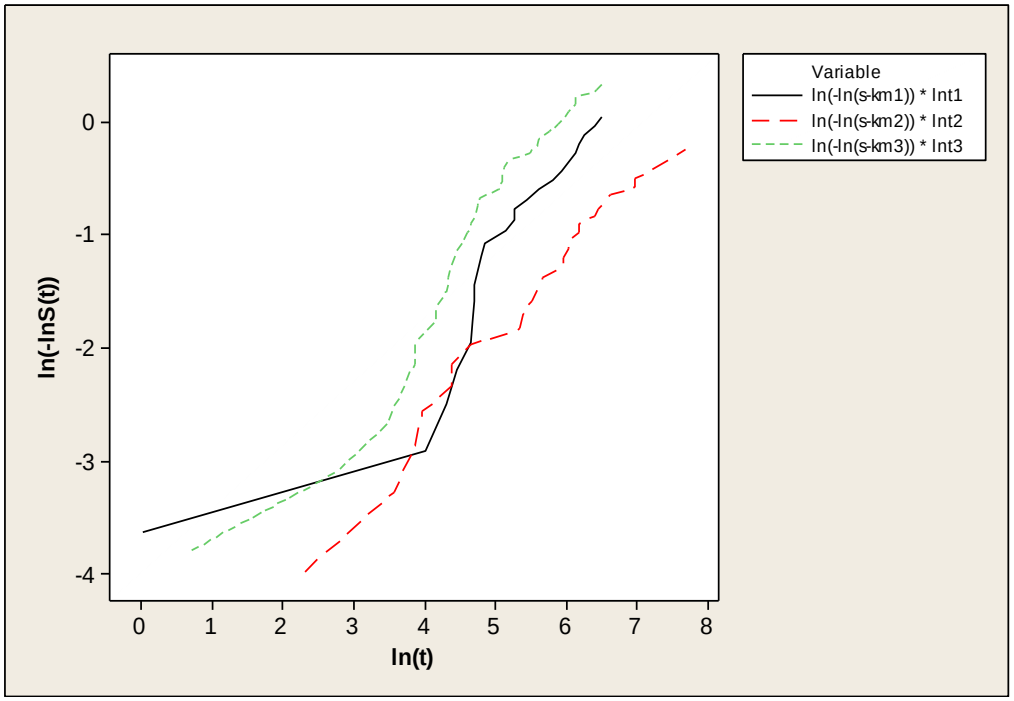
Όπου groupf1=$\begin{cases} 1, & \text{όταν } group = 1\ (ALL) \\ 0, & \text{αλλιώς} \end{cases}$ και

groupf2=$\begin{cases} 1, & \text{όταν } group = 2 (AML\ low\_risk) \\ 0, & \text{αλλιώς} \end{cases}$

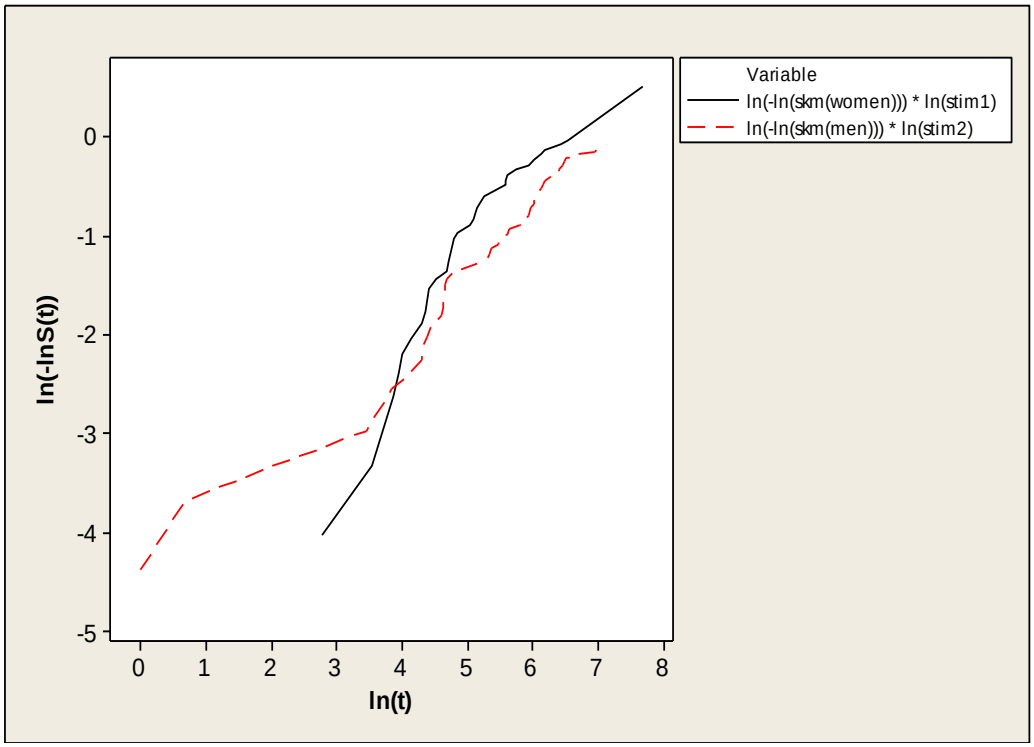
3.7 Προσαρμογή του μοντέλου αναλογικής διακινδύνευσης του Cox

Όπως και στην προσαρμογή του μοντέλου επιταχυνόμενης διακοπής, πραγματοποιούμε πρώτα τον γραφικό έλεγχο που φαίνεται στο Σχήμα 3.9 για να δούμε αν είναι κατάλληλο το μοντέλο. Ο γραφικός έλεγχος που κάνουμε για το μοντέλο της αναλογικής διακινδύνευσης είναι οι γραφικές παραστάσεις των $\ln\{-\ln \hat{S}\, kmi(t)\}$ με i=1,2,3 έναντι των $\ln t_i$, όπου $\hat{S}km1, \hat{S}km2, \hat{S}km3$ οι εκτιμήσεις Kaplan-Meier των συναρτήσεων επιβίωσης για τους τρεις τύπους λευχαιμίας που υπολογίσαμε στην Παράγραφο 3.3 και $\ln t_1$, $\ln t_2$, $\ln t_3$ οι λογάριθμοι των αντίστοιχων χρόνων για τους τρεις τύπους λευχαιμίας. Για να ισχύει η υπόθεση της αναλογικής διακινδύνευσης πρέπει οι καμπύλες να είναι παράλληλες, δηλαδή να διαφέρουν μόνο ως προς τις οριζόντιες μετατοπίσεις. Στο Σχήμα 3.9 παρατηρούμε ότι οι καμπύλες για τις κατηγορίες AML low_risk και AML high_risk είναι παράλληλες, ενώ η καμπύλη για την κατηγορία ALL τέμνει τις άλλες δύο.

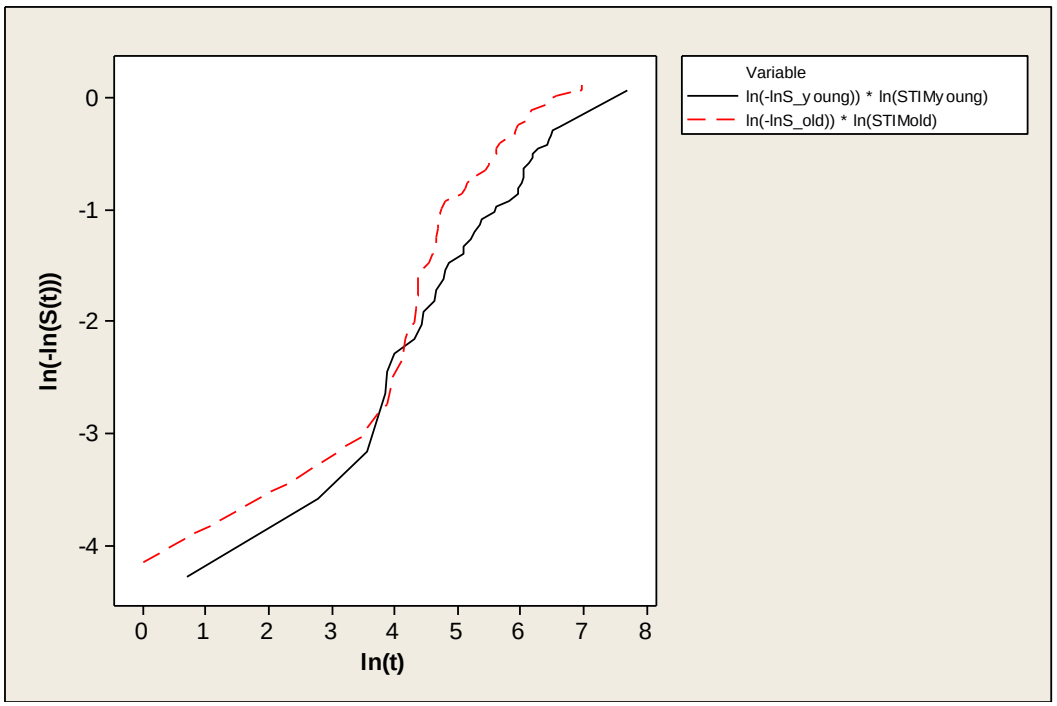
Πραγματοποιώντας τον ίδιο γραφικό έλεγχο αλλά για τη μεταβλητή recipient_sex, δηλαδή κάνουμε τις γραφικές παραστάσεις των $\ln\{-\ln \hat{S}\, kmj(t)\}$ έναντι των $\ln t_j$, όπου j=1,2 με j=1 για τις γυναίκες και j=2 για τους άνδρες, έχουμε το Σχήμα 3.10. Και πάλι βλέπουμε ότι οι καμπύλες τέμνονται. Τέλος, κάνουμε τον έλεγχο της αναλογικής διακινδύνευσης για τη συμμεταβλητή recipient_age. Φτιάχνουμε δύο κατηγορίες για αυτή τη συμμεταβλητή: young=7-28 χρονών και old=29-52 χρονών, δηλαδή κάνουμε τις γραφικές παραστάσεις των $\ln\{-\ln \hat{S}\, kmk(t)\}$ έναντι των $\ln t_k$, όπου k=young, old, και έχουμε το Σχήμα 3.11. Για ακόμη μία φορά παρατηρούμε ότι οι δύο καμπύλες τέμνονται.



Σχήμα 3.9: Γραφικός έλεγχος του μοντέλου αναλογικής διακινδύνευσης για τις τρεις ομάδες λευχαιμίας



Σχήμα 3.10: Γραφικός έλεγχος του μοντέλου αναλογικής διακινδύνευσης για το φύλο του ασθενούς



Σχήμα 3.11: Γραφικός έλεγχος του μοντέλου αναλογικής διακινδύνευσης για την ηλικία του ασθενούς

Στη συνέχεια προσαρμόζουμε το ημι-παραμετρικό μοντέλο του Cox με τη βοήθεια της R. Τα αποτελέσματα που λαμβάνουμε από την προσαρμογή του μοντέλου φαίνονται στον Πίνακα 3.9. Βλέπουμε τις εκτιμήσεις των συντελεστών των συμμεταβλητών, τα εκθετικά των συντελεστών, τα τυπικά σφάλματα, τις τιμές της ελεγχοσυνάρτησης Wald και τις p-τιμές των ελέγχων. Ο έλεγχος Wald γίνεται για κάθε μία από τις μεταβλητές ξεχωριστά με σκοπό να βρούμε ποιες συμμεταβλητές είναι στατιστικά σημαντικές. Για παράδειγμα, ο έλεγχος Wald που γίνεται για τη μεταβλητή donor_age είναι H_0: $\beta_{donor\_age}=0$ vs H_1: $\beta_{donor\_age} \neq 0$ με την ελεγχοσυνάρτηση να είναι $z = \frac{\beta_{donor\_age}}{se(\beta_{donor\_age})} = -0.117$ και να ακολουθεί την τυποποιημένη κανονική κατανομή, N(0,1), υπό την H_0. Επιπλέον, η p-τιμή του ελέγχου αυτού είναι p-value=0.910 και είναι αρκετά μεγάλη οπότε δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση ότι ο συντελεστής της μεταβλητής donor_age είναι μηδέν. Συνεπώς, η συμμεταβλητή donor_age δεν είναι στατιστικά σημαντική. Με την ίδια λογική, γίνονται οι έλεγχοι Wald για τις υπόλοιπες μεταβλητές και παρατηρούμε ότι οι groupf και fab είναι στατιστικά σημαντικές, ενώ οι υπόλοιπες δεν είναι.

Για την επιλογή του καλύτερου μοντέλου του Cox ακολουθούμε τη διαδικασία της διαδοχικής αφαίρεσης (backward elimination) με κριτήριο την τιμή $-2(\widehat{l_0} - \widehat{l_1})$. Το πρώτο βήμα της διαδικασίας παρουσιάζεται στον Πίνακα 3.10. Παρατηρούμε ότι η πρώτη μεταβλητή που αφαιρείται από το μοντέλο είναι η donor_age καθώς έχει τη μεγαλύτερη p-τιμή και άρα δεν είναι στατιστικά σημαντική. Η διαδικασία συνεχίζεται για τις υπόλοιπες μεταβλητές μέχρι το σημείο όπου δεν μπορεί να αφαιρεθεί κάποια άλλη γιατί όλες είναι σημαντικές, το τελευταίο βήμα φαίνεται στον Πίνακα 3.11. Ξεκινάμε με τη σύγκριση μεταξύ του μοντέλου που περιέχει τις 10 συμμεταβλητές και του μοντέλου που δεν περιλαμβάνει καμία. Δηλαδή πραγματοποιούμε τον έλεγχο H_0: $\beta_1=\beta_2=\ldots=\beta_{10}$ vs H_1: διαφορετικά. Η ελεγχοσυνάρτηση είναι $-2(\widehat{l_0} - \widehat{l_1}) = 26.1$ και ακολουθεί τη X^2 κατανομή με 11 βαθμούς ελευθερίας, αφού η μεταβλητή groupf χωρίζεται σε δύο ψευδομεταβλητές groupf1= $\begin{Bmatrix} 1, & \text{όταν } group = 1\ (ALL) \\ 0, & \text{αλλιώς} \end{Bmatrix}$ και groupf2= $\begin{Bmatrix} 1, & \text{όταν } group = 2(AML\ low\_risk) \\ 0, & \text{αλλιώς} \end{Bmatrix}$, με $\widehat{l_0}$ τη μεγιστοποιημένη συνάρτηση πιθανοφάνειας υπό την H_0 και $\widehat{l_1}$ τη μεγιστοποιημένη συνάρτηση πιθανοφάνειας υπό την H_1. Η p-τιμή είναι p=0.00627 και έτσι απορρίπτουμε τη

μηδενική υπόθεση. Συνεπώς, δεν μπορεί οι συντελεστές όλων των συμμεταβλητών να είναι μηδέν.

Πίνακας 3.9: Το μοντέλο αναλογικής διακινδύνευσης του Cox

Συμμεταβλητές	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	Wald	p- value
recipient_age	0.014	1.014	0.021	0.681	0.500
donor_age	-0.002	0.998	0.019	-0.117	0.910
recipient_sex	-0.109	0.896	0.241	-0.453	0.650
donor_sex	0.033	1.034	0.242	0.138	0.890
recipient_cmv	-0.061	0.941	0.255	-0.238	0.810
donor_cmv	-0.048	0.953	0.247	-0.194	0.850
groupf1	0.188	1.207	0.364	0.517	0.610
groupf2	-0.874	0.417	0.282	-3.098	0.002
waiting_time	-0.0003	1.000	0.0003	-0.870	0.380
fab	0.802	2.230	0.282	2.841	0.005
mtx	0.292	1.339	0.254	1.148	0.250

Πίνακας 3.10: Το πρώτο βήμα της διαδικασίας της διαδοχικής αφαίρεσης

Αφαίρεση συμμεταβλητής	Μεταβολή της $-2\hat{l}$	p-value
donor_age	0.014	0.907
donor_sex	0.019	0.890
donor_cmv	0.038	0.846
recipient_cmv	0.057	0.812
recipient_sex	0.205	0.651
recipient_age	0.459	0.498
waiting_time	0.844	0.358
mtx	1.283	0.257
fab	8.232	0.004**
groupf	13.560	0.001**

Πίνακας 3.11: Το τελικό βήμα της διαδικασίας της διαδοχικής αφαίρεσης

Αφαίρεση συμμεταβλητής	Μεταβολή της $-2\hat{l}$	p-value
fab	8.290	0.004 **
groupf	13.957	0.0009 ***

Από τον Πίνακα 3.11 καταλήγουμε ότι το καλύτερο μοντέλο αναλογικής διακινδύνευσης του Cox είναι εκείνο με τις συμμεταβλητές fab και groupf. Στην περίπτωση αυτή για τη συμμεταβλητή groupf γίνεται ο έλεγχος

$$H_0: \beta_{groupf1}=\beta_{groupf2}=0 \quad \text{vs} \quad H_1: \text{διαφορετικά}$$

δηλαδή εξετάζεται η αφαίρεση και των δύο ψευδομεταβλητών groupf1 και groupf2, οπότε η ελεγχοσυνάρτηση προκύπτει 13.957 και ακολουθεί την X^2 κατανομή με 2 βαθμούς ελευθερίας. Ακόμη, $P(X_2^2 > 13.957) = 0.0009$ οπότε απορρίπτουμε τη μηδενική υπόθεση και άρα η συμμεταβλητή groupf δεν αφαιρείται από το μοντέλο.

Εναλλακτικά, θα μπορούσαμε να χρησιμοποιήσουμε το κριτήριο AIC όπως κάναμε στην Παράγραφο 3.6 για την επιλογή του καλύτερου μοντέλου επιταχυνόμενης διακοπής. Συνεπώς, στον Πίνακα 3.12 βλέπουμε το τελικό μοντέλο αναλογικής διακινδύνευσης του Cox.

Πίνακας 3.12: Το τελικό μοντέλο αναλογικής διακινδύνευσης του Cox

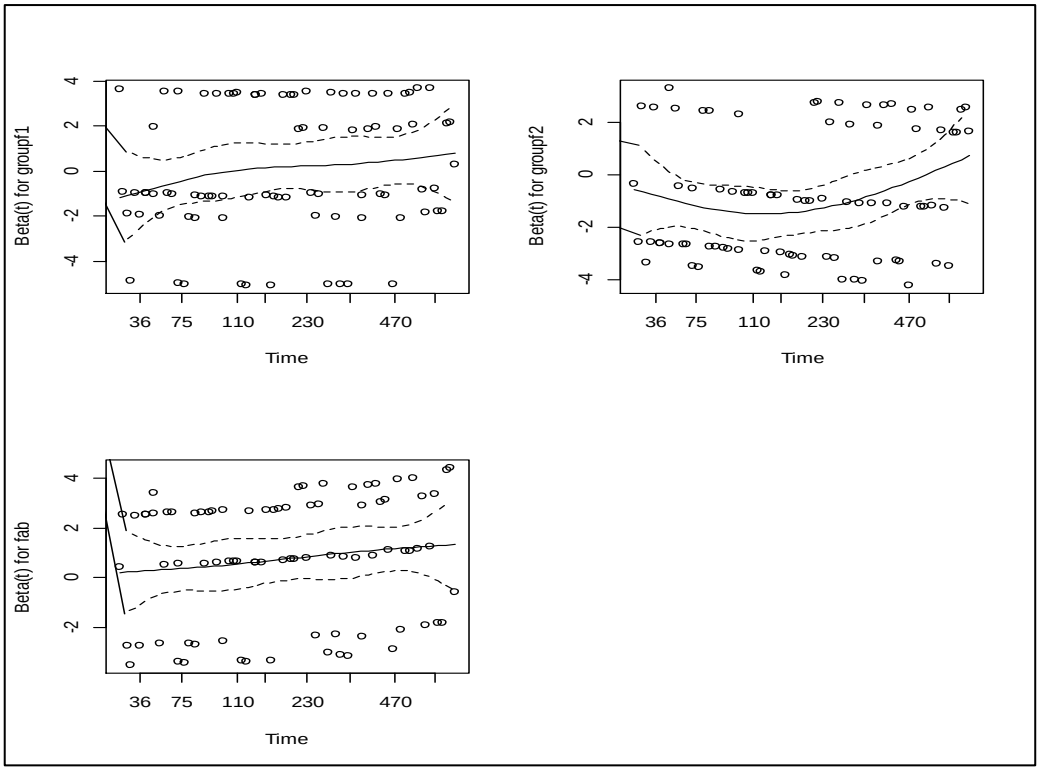
Συμμεταβλητές	$\hat{\beta}$	$\exp(\hat{\beta})$	$\text{se}(\hat{\beta})$	Wald	p- value	95% Δ.Ε του $\exp(\hat{\beta})$
groupf1	0.052	1.053	0.321	0.162	0.870	0.56-1.97
groupf2	-0.853	0.426	0.268	-3.176	0.002	0.25-0.72
fab	0.770	2.159	0.270	2.847	0.004	1.27-3.67

Τα εκθετικά των συντελεστών μας δείχνουν κατά πόσο πολλαπλασιάζεται η συνάρτηση διακινδύνευσης. Πιο συγκεκριμένα, τα εκθετικά των συντελεστών μας δείχνουν κατά πόσο μία συμμεταβλητή επιδρά στη διάρκεια ζωής όταν οι άλλες συμμεταβλητές του μοντέλου είναι σταθερές. Για παράδειγμα, κρατώντας τις άλλες συμμεταβλητές σταθερές, αν η μεταβλητή groupf1 αυξηθεί κατά μία μονάδα, δηλαδή

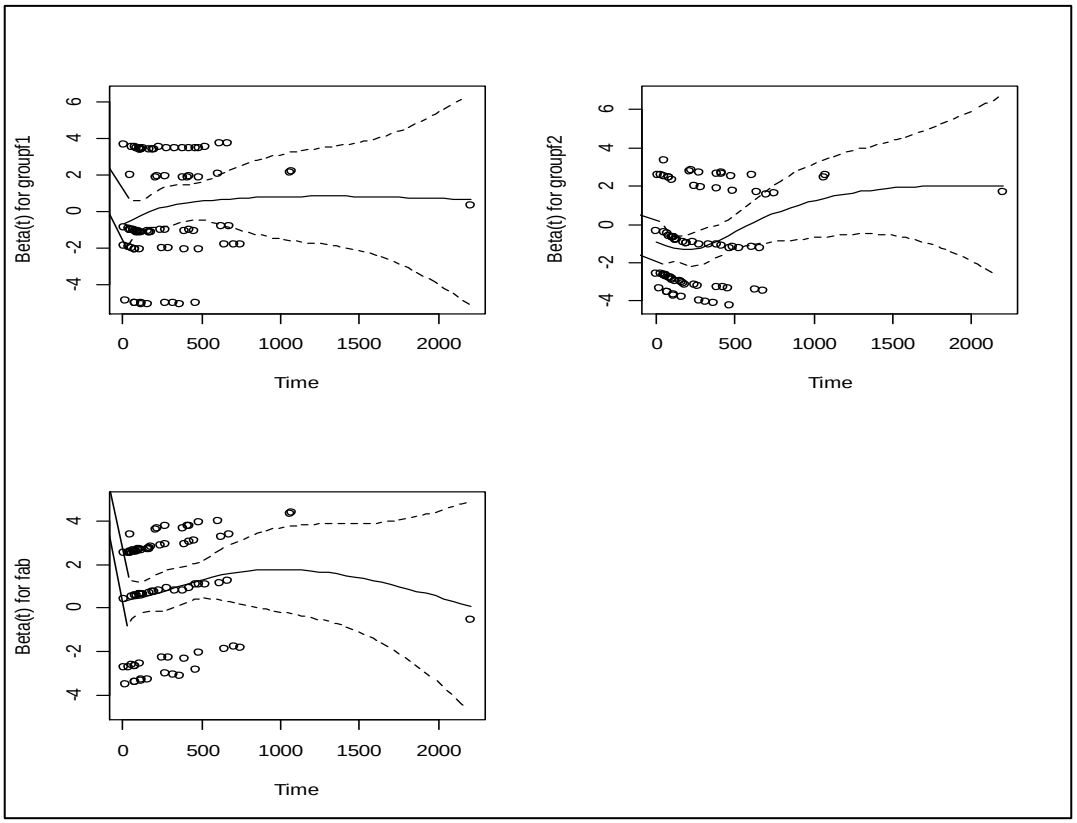
πηγαίνοντας από την κατηγορία αναφοράς (ομάδα 3: AML high_risk) στην ομάδα 1 (ALL) ο κίνδυνος να πεθάνει ή να υποτροπιάσει ο ασθενής (να συμβεί το γεγονός) αυξάνεται κατά $e^{\beta_{groupf1}} = 1.053$. Τα διαστήματα εμπιστοσύνης για τα e^{β}, που φαίνονται στην τελευταία στήλη του Πίνακα 3.10, υπολογίζονται από τον τύπο $\{\hat{\beta} \pm 1.96se(\hat{\beta})\}$.

Στο Κεφάλαιο 1, στην Παράγραφο 1.4.3 αναφέραμε ότι τα κλιμακοποιημένα υπόλοιπα Schoenfeld μπορούν να χρησιμοποιηθούν για τον έλεγχο της υπόθεσης της αναλογικότητας στο μοντέλο του Cox. Μας ενδιαφέρει να ελέγξουμε αν υπάρχει εξάρτηση των συντελεστών των συμμεταβλητών του μοντέλου από το χρόνο. Καταλήξαμε ότι το μοντέλο του Cox περιλαμβάνει δύο συμμεταβλητές, την groupf και την fab, με την groupf να χωρίζεται σε δύο ψευδομεταβλητές: groupf1 και groupf2.

Όπως είδαμε στην Παράγραφο 1.4.3 μελετάμε την εξάρτηση των υπολοίπων Schoenfeld από τον ίδιο το χρόνο ή από μία συνάρτηση του χρόνου g(t). Τα Σχήματα 3.12 και 3.13 απεικονίζουν τα κλιμακοποιημένα υπόλοιπα Schoenfeld για τις τρεις συμμεταβλητές του μοντέλου συναρτήσει μιας συνάρτησης του χρόνου, g(t), και συναρτήσει του ίδιου του χρόνου αντίστοιχα. Πιο συγκεκριμένα για την g(t) χρησιμοποιούμε την επιλογή "rank", βαθμός διάταξης, της R και για τον ίδιο το χρόνο την επιλογή "identity" της R (g(t)=t). Στα δύο σχήματα δεν παρατηρούμε κάτι που να δείχνει ότι έχουμε εξάρτηση από το χρόνο.



Σχήμα 3.12: Διάγραμμα των κλιμακοποιημένων υπολοίπων Schoenfeld των τριών μεταβλητών με κλίμακα το βαθμό διάταξης (συνάρτηση του χρόνου)



Σχήμα 3.13: Διάγραμμα των κλιμακοποιημένων υπολοίπων Schoenfeld των τριών μεταβλητών έναντι του χρόνου, g(t)=t

Το σημείο που φαίνεται να είναι απομακρυσμένο από τα υπόλοιπα στο Σχήμα 3.13, στα δεξιά, είναι η παρατήρηση 69 όπου ο ασθενής ανήκει στην ομάδα 2 (λευχαιμία τύπου AML low_risk).

Για να ενισχύσουμε τα αποτελέσματα μας θεωρούμε την παλινδρόμηση

$$\beta_i(t) = \beta_i + \theta_i(g(t) - \bar{g}),\ i = 1, \dots, p$$

και πραγματοποιούμε τον έλεγχο

H_0: $\theta_i = 0\ για\ κάθε\ i = 1, \dots, p$ vs H_1: διαφορετικά

Αν δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση τότε η αναλογικότητα του μοντέλου ισχύει. Με τη βοήθεια της R (παράρτημα Α), πραγματοποιούμε τον παραπάνω έλεγχο και έχουμε τον Πίνακα 3.13, όπου αντί του αρχικού χρόνου χρησιμοποιείται ως κλίμακα ο βαθμός διάταξης (rank), συνάρτηση του χρόνου, $g(t)$, και τον Πίνακα 3.14 όπου θεωρούμε τον αρχικό χρόνο.

Πίνακας 3.13: Έλεγχος της αναλογικής διακινδύνευσης με κλίμακα τον βαθμό διάταξης

	rho	chisq	p-value
groupf1	0.168	2.19	0.139
groupf2	0.174	2.32	0.128
Fab	0.145	1.71	0.191
Global	NA	3.71	0.295

Πίνακας 3.14: Έλεγχος της αναλογικής διακινδύνευσης αν θεωρήσουμε τον ίδιο τον χρόνο, g(t)=t

	rho	chisq	p-value
groupf1	0.119	1.105	0.293
groupf2	0.254	4.909	0.027
Fab	0.104	0.882	0.348
Global	NA	5.409	0.144

Η ελεγχοσυνάρτηση που υπολογίζουμε και στους δύο πίνακες για κάθε μία από τις τρεις συμμεταβλητές μας δίνεται από τον τύπο (1.12) του Κεφαλαίου 1, Παράγραφος 1.4.3, με τη διαφορά ότι για τα αποτελέσματα του Πίνακα 3.14 έχουμε

$g(t) = t$. Στη στήλη «chisq» των Πινάκων 3.13 και 3.14 έχουμε την τιμή αυτής της ελεγχοσυνάρτησης και δίπλα την αντίστοιχη p-τιμή του ελέγχου. Στην τελευταία γραμμή των πινάκων έχουμε τον έλεγχο αυτό για ολόκληρο το μοντέλο και άρα χρησιμοποιείται ο τύπος (1.11) της Παραγράφου 1.4.3 για την τιμή της ελεγχοσυνάρτησης. Η στήλη «rho» μας δείχνει το συντελεστή συσχέτισης του Pearson μεταξύ των κλιμακοποιημένων Schoenfeld και της συνάρτησης $g(t_{(J)})$ για κάθε συμμεταβλητή i. Από τις p-τιμές του Πίνακα 3.13 παρατηρούμε ότι δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση του ελέγχου και άρα ισχύει η ιδιότητα της αναλογικής διακινδύνευσης. Αντίθετα, στον Πίνακα 3.14 παρατηρούμε ότι για τη συμμεταβλητή groupf2 έχουμε p-value<0.05 και άρα δεν ισχύει η υπόθεση της αναλογικότητας. Από τα Σχήματα 3.12 και 3.13 που φαίνονται τα υπόλοιπα Schoenfeld βλέπουμε ότι για τη συμμεταβλητή groupf2 υπάρχει μία τάση αύξησης. Βέβαια αυτό ίσως να οφείλεται στην απομακρυσμένη παρατήρηση στα δεξιά που αναφέραμε προηγουμένως (Fox, 2002).

Επειδή, λοιπόν για την groupf2 προκύπτει ότι δεν ισχύει η υπόθεση της αναλογικότητας θα ελέγξουμε μήπως ο συντελεστής της συμμεταβλητής εξαρτάται από το χρόνο. Έστω ότι $\beta(t) = a + bt$ ο συντελεστής μιας συμμεταβλητής, τότε έχουμε $\beta(t)x = ax + btx = ax + bz$ ο συντελεστής για την εξαρτημένη από το χρόνο μεταβλητή $z = \mathrm{t}x$ (Therneau et al., 2017).

Συνεπώς, για τη συμμεταβλητή groupf2 δημιουργούμε την εξαρτημένη από το χρόνο μεταβλητή, t*groupf2, με την εντολή tt (time-transform) και έχουμε τα αποτελέσματα του Πίνακα 3.15.

Πίνακας 3.15: Προσαρμογή του μοντέλου μετά τη δημιουργία της νέας μεταβλητής t*groupf2

	$\hat{\beta}$	exp($\hat{\beta}$)	se($\hat{\beta}$)	z	p-value
groupf1	0.034	1.035	0.322	0.106	0.920
groupf2	-1.462	0.232	0.429	-3.405	0.001
fab	0.740	2.096	0.271	2.734	0.006
tt(groupf2)	0.002	1.002	0.001	1.858	0.063

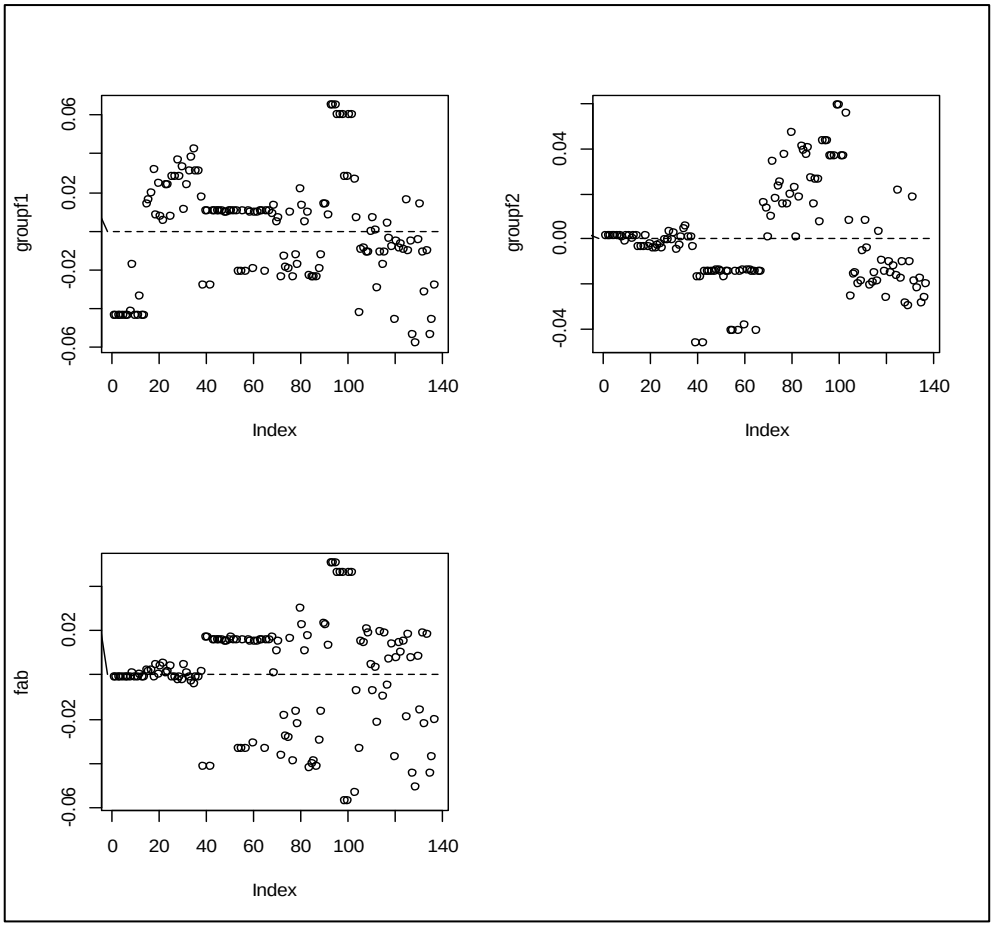
Από τον Πίνακα 3.15 παρατηρούμε ότι η καινούργια μεταβλητή δεν είναι στατιστικά σημαντική και άρα δεν έχουμε εξάρτηση του συντελεστή της groupf2 από το χρόνο. Ίσως υπάρχει οριακή εξάρτηση από το χρόνο αλλά όχι αξιοσημείωτη. Αυτό συμφωνεί και με τις αντίστοιχες γραφικές παραστάσεις των Σχημάτων 3.12 και 3.13.

3.8 Σημεία επιρροής

Όπως αναφέρθηκε στο Κεφάλαιο 1, Παράγραφος 1.4.3 τα σημεία επιρροής είναι αυτές οι παρατηρήσεις που αν αφαιρεθούν από το μοντέλο η προσαρμογή του μοντέλου θα μας δώσει διαφορετικά αποτελέσματα. Με τη βοήθεια των εντολών της R του Πίνακα 3.16 υπολογίζουμε τις ποσότητες DFBETAS, όπου mod το προσαρμοσμένο μοντέλο του Cox στο οποίο καταλήξαμε στην Παράγραφο 3.7, Πίνακας 3.12. Με το Σχήμα 3.14 ελέγχουμε αν υπάρχουν σημεία επιρροής για τις σημαντικές μεταβλητές του μοντέλου του Cox, δηλαδή για τις groupf και fab. Στο Σχήμα 3.14 παρατηρούμε ότι υπάρχουν ορισμένα σημεία που βρίσκονται σε απόσταση από τα σημεία στο κέντρο. Τα απομακρυσμένα αυτά σημεία είναι σχετικά λίγα συγκριτικά με το συνολικό πλήθος των παρατηρήσεων. Για τη μεταβλητή groupf1 τα ακραία σημεία είναι οι παρατηρήσεις 93, 94, 95, 96, 97, 98,101 και 102, για την groupf2 είναι οι παρατηρήσεις 99, 100 και 103 και τέλος για την fab είναι οι παρατηρήσεις 93, 94 και 95. Ωστόσο δε μας βάζουν σε ιδιαίτερες υποψίες καθώς οι τιμές των DFBETAS είναι χαμηλές.

Πίνακας 3.16: Εντολές για τον υπολογισμό τον DFBETAS

```
> dfbeta<-residuals(mod, type="dfbeta")
> par(mfrow=c(2,2))
> for (j in 1:3) {
+  plot(dfbeta[,j],ylab=names(coef(mod))[j])
+ abline(h=0, lty=2)
+ }
```



Σχήμα 3.14: Γραφήματα δείκτη των dfbetas για τις τρεις συμμεταβλητές

3.9 Στρωματοποιημένη ανάλυση

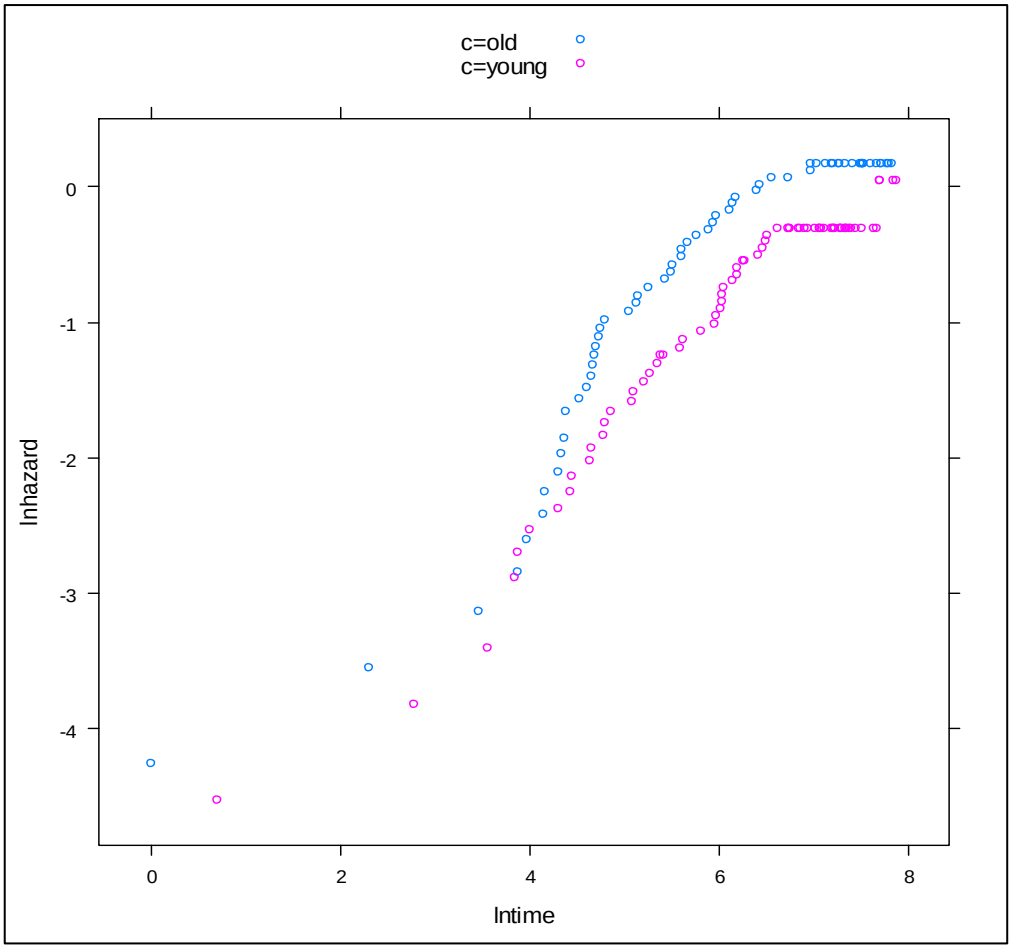
Σε αυτή τη παράγραφο προσαρμόζουμε ένα στρωματοποιημένο μοντέλο του Cox ως προς τη μεταβλητή recipient_age. Η διάμεσος για τη μεταβλητή recipient_age είναι τα 28 χρόνια, οπότε θεωρούμε τα εξής δύο στρώματα: τους νέους (7-28 χρονών) και τους μεγαλύτερους (29-52 χρονών). Κάθε στρώμα έχει τη δική του συνάρτηση διακινδύνευσης, ενώ οι συντελεστές των υπόλοιπων συμμεταβλητών παραμένουν σταθεροί μέσα στα στρώματα. Προσαρμόζοντας το στρωματοποιημένο μοντέλο του Cox παρατηρούμε στον Πίνακα 3.17 ότι όπως και στο κλασικό μοντέλο του Cox οι συμμεταβλητές groupf και fab είναι οι στατιστικά σημαντικές.

Πίνακας 3.17: Το στρωματοποιημένο μοντέλο του Cox ως προς τη συμμεταβλητή recipient_age

	$\hat{\beta}$	exp($\hat{\beta}$)	se($\hat{\beta}$)	z	p-value
donor_age	-0.007	0.993	0.016	-0.461	0.645
recipient_sex	-0.059	0.943	0.243	-0.244	0.807
donor_sex	0.006	1.006	0.238	0.024	0.981
recipient_cmv	-0.116	0.890	0.252	-0.462	0.644
donor_cmv	-0.016	0.984	0.250	-0.065	0.948
groupf1	0.262	1.300	0.359	0.732	0.464
groupf2	-0.900	0.407	0.283	-3.184	0.001**
waiting_time	-0.0003	0.999	0.0004	-0.779	0.436
fab	0.809	2.245	0.276	2.931	0.003**
mtx	0.287	1.333	0.253	1.137	0.256

Στον Πίνακα 3.17 βλέπουμε τα αποτελέσματα από την προσαρμογή του στρωματοποιημένου μοντέλου του Cox. Πιο συγκεκριμένα, έχουμε τους συντελεστές των συμμεταβλητών στην πρώτη στήλη στη συνέχεια τα εκθετικά των συντελεστών, τα τυπικά σφάλματα και τέλος την ελεγχοσυνάρτηση z και τις p-τιμές για τους ελέγχους Wald που γίνονται για να αποφασίσουμε ποιες συμμεταβλητές είναι σημαντικές. Οι έλεγχοι Wald είναι οι εξής: H_0: $\beta_i=0$ vs H_1: $\beta_i\neq0$, όπου i η συμμεταβλητή την οποία ελέγχουμε κάθε φορά. Η ελεγχοσυνάρτηση υπολογίζεται από τον τύπο $z = \frac{\beta_i}{se(\beta_i)}$ και ακολουθεί την τυποποιημένη κανονική κατανομή, N(0,1) υπό τη μηδενική υπόθεση.

Στη συνέχεια πραγματοποιούμε το Σχήμα 3.15 για να ελέγξουμε αν ισχύει η υπόθεση της αναλογικής διακινδύνευσης για τα δύο στρώματα που κατασκευάσαμε. Από το σχήμα αυτό βλέπουμε ότι οι δύο καμπύλες τέμνονται σε κάποια σημεία. Συνεπώς, θα ήταν λάθος να συμπεριλάβουμε την κατηγορική μεταβλητή της recipient_ age ως δείκτρια μεταβλητή στο μοντέλο του Cox.



Σχήμα 3.15: Εκτιμήσεις Breslow του ln(h(t)) έναντι του lnt για τις δύο κατηγορίες της ηλικίας των ασθενών

3.10 Παλινδρόμηση threshold

Επειδή η υπόθεση της αναλογικής διακινδύνευσης στο μοντέλο του Cox δεν ισχύει πάντα προσαρμόζουμε το μοντέλο παλινδρόμησης threshold θεωρώντας τη στοχαστική διαδικασία Wiener με αρχική τιμή x_0και τάση μ. Εξετάζουμε κατά πόσο οι συμμεταβλητές μας επηρεάζουν την αρχική κατάσταση της υγείας των ασθενών, x_0, και την παράμετρο μ. Υποθέτουμε ότι όλες οι συμμεταβλητές είναι μεταβλητές πρόβλεψης για τις παραμέτρους μ και x_0. Με τη βοήθεια της R και του πακέτου threg (παράρτημα Β) (Xiao, 2015) από την προσαρμογή του μοντέλου λαμβάνουμε τα αποτελέσματα του Πίνακα 3.18 για την τάση μ και τα αποτελέσματα του Πίνακα 3.19 για την παράμετρο x_0. Στους δύο Πίνακες παρουσιάζονται οι εκτιμημένες παράμετροι του μοντέλου καθώς και οι αντίστοιχοι έλεγχοι Wald.

Πίνακας 3.18: Το προσαρμοσμένο μοντέλο παλινδρόμησης threshold για την τάση μ

μ	$\hat{\beta}$	se($\hat{\beta}$)	z	p-value
Σταθερά	0.040	0.026	1.569	0.120
groupf1	-0.001	0.016	-0.820	0.410
groupf2	0.029	0.012	2.424	0.015
recipient_age	-0.004	0.001	-3.250	0.001
donor_age	0.003	9.58×10^{-4}	2.804	0.005
recipient_sex	0.003	0.001	0.243	0.810
donor_sex	0.004	0.011	0.381	0.700
recipient_cmv	0.010	0.012	0.915	0.360
donor_cmv	-5.35×10^{-4}	0.011	-0.048	0.960
fab	-0.028	0.012	-2.312	0.021
mtx	0.010	0.014	0.738	0.460
waiting_time	3.27×10^{-6}	2.87×10^{-5}	0.114	0.910

Παρατηρούμε ότι για την παράμετρο μ, που δείχνει πως εξελίσσεται η διαδικασία, στατιστικά σημαντικές είναι οι συμμεταβλητές recipient_age, donor_age και οριακά σημαντικές οι fab και groupf. Αυτό σημαίνει πως ο ρυθμός με τον οποίο αλλάζει η αρχική κατάσταση μέχρι να φτάσει στο οριακό σημείο επηρεάζεται σημαντικά από την ηλικία του ασθενούς, την ηλικία του δότη, την ομάδα λευχαιμίας από την οποία πάσχει ο ασθενής και τη συμμεταβλητή fab η οποία μας δείχνει το αν οι ασθενείς διατρέχουν μεγαλύτερο κίνδυνο θανάτου ή υποτροπής μετά τη μεταμόσχευση μυελού των οστών. Πιο συγκεκριμένα, η αύξηση της ηλικίας του ασθενή οδηγεί τη στοχαστική διαδικασία γρηγορότερα στο όριο (αρνητικός συντελεστής), η αύξηση της ηλικίας του δότη οδηγεί πιο αργά τη διαδικασία στο όριο (θετικός συντελεστής), η αύξηση της ψευδομεταβλητής groupf1 από το 0 (ομάδα AML high_risk) στο 1 (ομάδα ALL) οδηγεί τη διαδικασία γρηγορότερα στο όριο, ενώ η αύξηση της ψευδομεταβλητής groupf2 από το 0 (ομάδα AML high_risk) στο 1 (ομάδα AML low_risk) οδηγεί τη διαδικασία πιο αργά στο όριο. Τέλος, η αύξηση της συμμεταβλητής fab από το 0 στο 1 οδηγεί τη διαδικασία γρηγορότερα στο όριο.

Πίνακας 3.19: Το προσαρμοσμένο μοντέλο παλινδρόμησης threshold για την παράμετρο $ln\, x_0$

$\ln x_0$	$\hat{\beta}$	se($\hat{\beta}$)	z	p-value
Σταθερά	0.019	0.644	0.030	0.980
groupf1	-0.012	0.321	-0.038	0.970
groupf2	0.019	0.311	0.062	0.950
recipient_age	0.115	0.020	5.651	<0.001
donor_age	-0.067	0.016	-4.100	<0.001
recipient_sex	-8.57×10^{-4}	0.255	-0.003	1.0
donor_sex	0.004	0.282	0.014	0.990
recipient_cmv	0.021	0.271	0.076	0.940
donor_cmv	0.015	0.305	0.050	0.960
fab	0.015	0.313	0.049	0.960
mtx	-0.015	0.305	0.049	0.960
waiting_time	8.09×10^{-4}	1.97×10^{-4}	4.106	<0.001

Παρατηρούμε ότι για την αρχική κατάσταση της διαδικασίας στατιστικά σημαντικές είναι οι συμμεταβλητές recipient_age, donor_age και waiting_time καθώς είναι αυτές με τις μικρότερες p-τιμές. Αυτό σημαίνει ότι η ηλικία του ασθενούς, η ηλικία του δότη και ο χρόνος αναμονής μεταξύ της διάγνωσης και της μεταμόσχευσης επηρεάζουν σημαντικά το λογάριθμο της αρχικής κατάστασης της διαδικασίας. Αναλυτικότερα, η αύξηση της ηλικίας του ασθενούς σχετίζεται με καλή αρχική κατάσταση (θετικός συντελεστής), η αύξηση της ηλικίας του δότη σχετίζεται με φτωχότερη αρχική κατάσταση (αρνητικός συντελεστής) και τέλος η αύξηση του χρόνου αναμονής συνδέεται με καλή αρχική κατάσταση που είναι ένα λογικό συμπέρασμα.

Στην Παράγραφο 3.7 προσαρμόσαμε το μοντέλο του Cox και βρήκαμε ότι περιλαμβάνει τις μεταβλητές groupf και fab. Πιο κοντά στο μοντέλο του Cox, το οποίο μελετά τη συνάρτηση διακινδύνευσης με το πέρας του χρόνου, είναι το μοντέλο στο οποίο καταλήξαμε στον Πίνακα 3.18 μετά την προσαρμογή του μοντέλου παλινδρόμησης threshold για την παράμετρο μ. Παρόλα αυτά το μοντέλο του Πίνακα 3.18 έχει δύο συμμεταβλητές παραπάνω, την recipient_age και την donor_age.

Με την παλινδρόμηση threshold χωρίζουμε τις μεταβλητές μας σε δύο σύνολα. Αυτές που σχετίζονται με την κατάσταση της υγείας των ασθενών στην αρχή και αυτές που σχετίζονται με το ρυθμό της αλλαγής της κατάστασης της υγείας των ασθενών μετά τη μεταμόσχευση μυελού των οστών. Τα δύο αυτά σύνολα δεν είναι ξένα μεταξύ τους. Πιο συγκεκριμένα, έχουν κοινές τις συμμεταβλητές recipient_age και donor_age. Παρατηρούμε ότι τα πρόσημα των συντελεστών για αυτές τις δύο συμμεταβλητές είναι αντίθετα. Αναλυτικότερα, η συμμεταβλητή recipient_age έχει θετικό συντελεστή για την παράμετρο $\ln x_0$ και αρνητικό για την παράμετρο μ. Αυτό σημαίνει ότι η αύξηση της ηλικίας του ασθενή σχετίζεται με μία καλή αρχική κατάσταση υγείας, αλλά η στοχαστική διαδικασία φτάνει γρηγορότερα στο όριο (threshold). Για τη συμμεταβλητή donor_age έχουμε αρνητικό συντελεστή για την παράμετρο $\ln y_0$ και θετικό για την παράμετρο μ. Δηλαδή, η αύξηση της ηλικίας του δότη σχετίζεται με φτωχότερη αρχική κατάσταση υγείας, αλλά η στοχαστική διαδικασία φτάνει πιο αργά στο όριο. Επειδή, λοιπόν, έχουμε αντίθετα πρόσημα στους συντελεστές θα μπορούσαμε να ελέγξουμε τον αναμενόμενο χρόνο μέχρι να

συμβεί το γεγονός (θάνατος ή υποτροπή του ασθενούς) για κάθε συνδυασμό των τιμών των δύο συμμεταβλητών.

Στη συνέχεια προσαρμόζουμε το μοντέλο παλινδρόμησης threshold λαμβάνοντας υπόψη μόνο τις συμμεταβλητές που είναι στατιστικά σημαντικές για τις παραμέτρους μας μ και $\ln x_0$. Από την προσαρμογή αυτή λαμβάνουμε τα αποτελέσματα του Πίνακα 3.20, όπου παρατηρούμε τις εκτιμημένες παραμέτρους του μοντέλου και τους αντίστοιχους ελέγχους Wald.

Επίσης, κάνουμε έναν έλεγχο πιθανοφάνειας για το μοντέλο που περιλαμβάνει όλες τις συμμεταβλητές, M_1, και για εκείνο που έχει μόνο τις στατιστικά, κατά Wald, σημαντικές συμμεταβλητές, M_0. Δηλαδή ο έλεγχος που γίνεται είναι:

H_0: μοντέλο M_0

H_1: μοντέλο M_1

το στατιστικό ελέγχου δίνεται από τον τύπο $-2(\hat{l}_0 - \hat{l}_1)$, όπου $\hat{l}_0$ η μεγιστοποιημένη λογαριθμοποιημένη συνάρτηση πιθανοφάνειας υπό την H_0 και $\hat{l}_1$ η μεγιστοποιημένη λογαριθμοποιημένη συνάρτηση πιθανοφάνειας υπό την H_1. Οπότε από την προσαρμογή των δύο αυτών μοντέλων έχουμε τα αποτελέσματα του Πίνακα 3.21. Η p-τιμή του ελέγχου αυτού είναι $P(X_{14}^2 > 1.68) = 0.999$. Συνεπώς, δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση και άρα το μοντέλο M_0 με τις στατιστικά, κατά Wald, σημαντικές συμμεταβλητές είναι καλύτερο από το M_1 που περιέχει όλες τις συμμεταβλητές.

Πίνακας 3.20: Το προσαρμοσμένο μοντέλο παλινδρόμησης threshold μόνο για τις στατιστικά σημαντικές συμμεταβλητές

	$\hat{\beta}$	se($\hat{\beta}$)	z	p-value
lnx_0				
σταθερά	0.084	0.373	0.226	0.820
recipient_age	0.112	0.014	7.745	< 0.001
donor_age	-0.064	0.011	-6.076	< 0.001
waiting_time	0.001	0.0001	5.688	< 0.001
μ				
σταθερά	0.084	0.021	3.972	< 0.001
recipient_age	-0.004	0.001	-3.848	< 0.001
donor_age	0.002	0.001	2.341	0.019
fab	-0.030	0.011	-2.817	0.005
groupf1	-0.022	0.015	-1.506	0.130
groupf2	0.015	0.011	1.316	0.190

Πίνακας 3.21: Έλεγχος του λόγου των πιθανοφανειών

	$\hat{l}$
Μοντέλο M_0	-706.75
Μοντέλο M_1	-705.91
$-2(\hat{l}_0 - \hat{l}_1)$	1.68

Ακολουθώντας το άρθρο των Xiao et al. (2015) προσαρμόζουμε το μοντέλο του threshold με τις συμμεταβλητές recipient_age και fab ως μεταβλητές πρόβλεψης για την αρχική κατάσταση lnx_0 και τις συμμεταβλητές groupf και fab ως μεταβλητές πρόβλεψης για την τάση μ τότε έχουμε τα αποτελέσματα του Πίνακα 3.22.

Πίνακας 3.22: Το μοντέλο του threshold αν θεωρήσουμε τις recipient_age και fab μεταβλητές πρόβλεψης για την παράμετρο $\ln x_0$ και τις groupf και fab για την τάση μ

	$\hat{\beta}$	se($\hat{\beta}$)	z	p-value
$\ln x_0$				
σταθερά	3.096	0.276	11.199	< 0.001
recipient_age	-0.032	0.008	-4.145	< 0.001
fab	-0.421	0.178	-2.366	0.018
μ				
σταθερά	0.028	0.010	2.759	0.006
groupf1	-0.013	0.013	-0.966	0.330
groupf2	0.014	0.011	1.310	0.190
fab	-0.024	0.011	-2.261	0.024

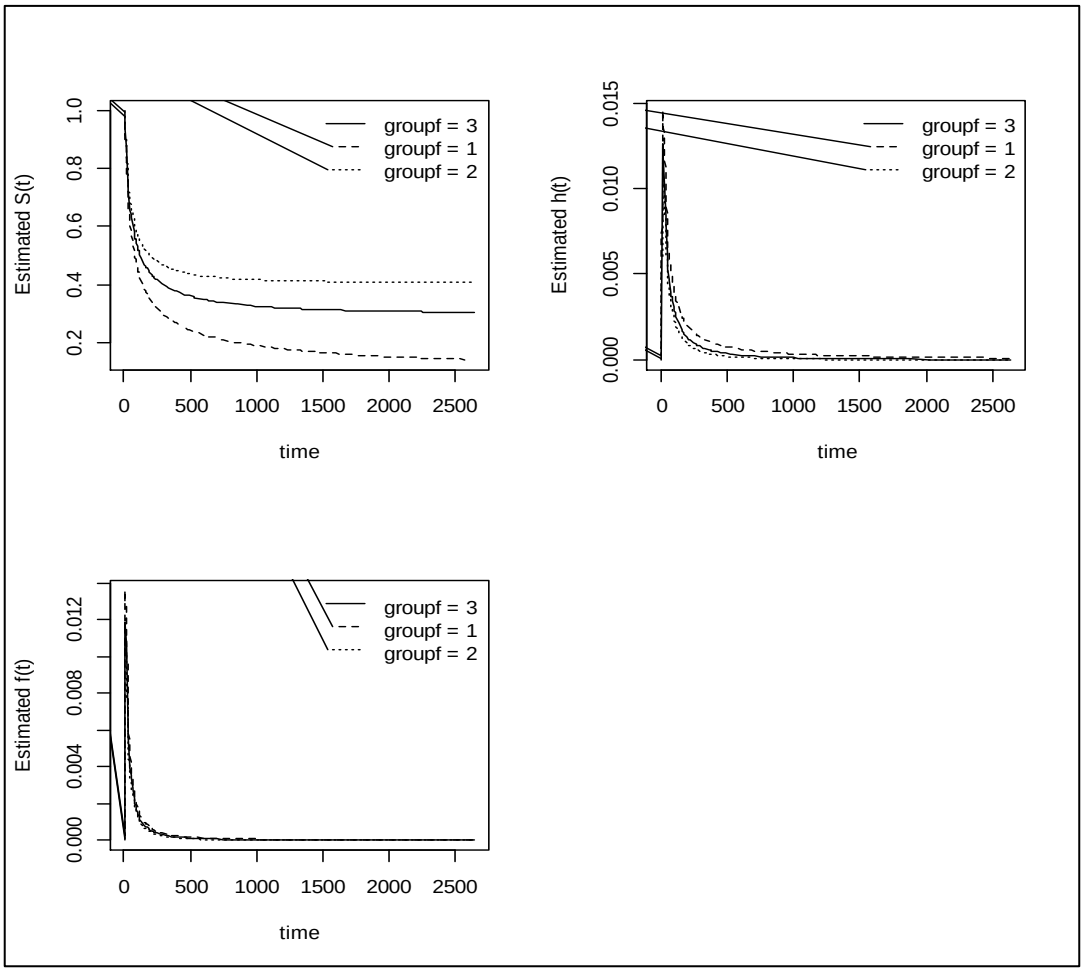
Από τον Πίνακα 3.22 παρατηρούμε ότι για την αρχική κατάσταση $\ln x_0$ στατιστικά σημαντικές είναι και οι δύο συμμεταβλητές πρόβλεψης, recipient_age και fab. Ενώ για την παράμετρο μ προκύπτει η συμμεταβλητή fab στατιστικά σημαντική. Αυτό σημαίνει ότι τα δύο μοντέλα παλινδρόμησης για τις δύο παραμέτρους έχουν κοινή συμμεταβλητή την fab. Ακόμη, παρατηρούμε ότι το πρόσημο του συντελεστή της fab και για τις δύο παραμέτρους είναι αρνητικό. Συνεπώς, συμπεραίνουμε ότι η αύξηση της τιμής της συμμεταβλητής fab από 0 σε 1, δηλαδή αναφερόμαστε σε ασθενείς που έχουν fab βαθμού 4 ή 5 και διατρέχουν μεγαλύτερο κίνδυνο να παρουσιάσουν υποτροπή ή να πεθάνουν, σχετίζεται με φτωχότερη αρχική κατάσταση υγείας και η στοχαστική διαδικασία φτάνει πιο γρήγορα στο όριο, κάτι που ήταν αναμενόμενο.

Στη συνέχεια, χρησιμοποιώντας τη συνάρτηση hr (hazard ratio) στην R (παράρτημα Β) για το μοντέλο παλινδρόμησης threshold λαμβάνουμε τις εκτιμήσεις του λόγου των συναρτήσεων διακινδύνευσης σε συγκεκριμένη χρονική στιγμή για συγκεκριμένες τιμές των συμμεταβλητών για τις δύο κατηγορίες της κατηγορικής μεταβλητής group. Η εντολή που χρησιμοποιούμε είναι:

hr.threg(fit1,var=groupf,timevalue=500,scenario=recipient_age(40)+
donor_age(45)+waiting_time(100)+fab(0))

όπου fit1 το μοντέλο παλινδρόμησης threshold που προσαρμόσαμε στον πίνακα 3.20. Θεωρούμε το εξής σενάριο: 40 ετών η ηλικία του ασθενή, 45 ετών η ηλικία του δότη, 100 ο χρόνος αναμονής σε μέρες από τη διάγνωση μέχρι τη μεταμόσχευση και 0 η τιμή του fab (όταν fab=1 είναι βαθμού 4 ή 5, fab=0 διαφορετικά). Υπολογίζουμε, λοιπόν, τις εκτιμήσεις των λόγων των συναρτήσεων διακινδύνευσης για τις δύο κατηγορίες τις συμμεταβλητής group και έχουμε για την groupf1 την τιμή 2.10909 και για την groupf2 την τιμή 0.5696079.

Στο Σχήμα 3.16 βλέπουμε τις γραφικές παραστάσεις των συναρτήσεων επιβίωσης, διακινδύνευσης και της συνάρτησης πυκνότητας πιθανότητας (σ.π.π) αντίστοιχα ως προς το χρόνο για τις 3 κατηγορίες της μεταβλητής group (group=1: ALL, group=2: AML low_risk, group=3: AML high_risk)



Σχήμα 3.16: Οι συναρτήσεις επιβίωσης, διακινδύνευσης και πυκνότητας πιθανότητας του μοντέλου παλινδρόμησης threshold συναρτήσει του χρόνου για τις κατηγορίες της μεταβλητής group

Από το Σχήμα 3.16 παρατηρούμε ότι μεγαλύτερο κίνδυνο διατρέχουν οι ασθενείς που πάσχουν από τη λευχαιμία της ομάδας 1 (ALL), ενώ καλύτερη πρόγνωση έχουμε για τους ασθενείς της ομάδας 2 (AML low_risk).

Τέλος, με τη βοήθεια της εντολής

predict.threg(fit1,timevalue=2000,scenario=recipient_age(40)+donor_age(45)+ waiting_time(100)+fab(0)+groupf1(0)+groupf2(1))

εκτιμούμε την αρχική κατάσταση υγείας του ασθενή, x_0, την τάση, μ, την τιμή της σ.π.π του μοντέλου, την τιμή της συνάρτησης επιβίωσης και τέλος την τιμή της συνάρτησης διακινδύνευσης για ένα συγκεκριμένο σενάριο μία συγκεκριμένη χρονική στιγμή. Για παράδειγμα, έχουμε το εξής σενάριο: 40 ετών η ηλικία του ασθενή, 45 ετών η ηλικία του δότη, 100 μέρες ο χρόνος αναμονής από τη στιγμή της διάγνωσης μέχρι τη μεταμόσχευση, 0 η τιμή της μεταβλητής fab και ο ασθενής ανήκει στον τύπο λευχαιμίας 2 (AML low_risk). Γι' αυτό το σενάριο λαμβάνουμε τις προβλέψεις του Πίνακα 3.23.

Πίνακας 3.23: Οι εκτιμήσεις της αρχικής κατάστασης x_0, της τάσης μ, της σ.π.π, της συνάρτησης επιβίωσης και της συνάρτησης διακινδύνευσης για συγκεκριμένο σενάριο και χρονική στιγμή

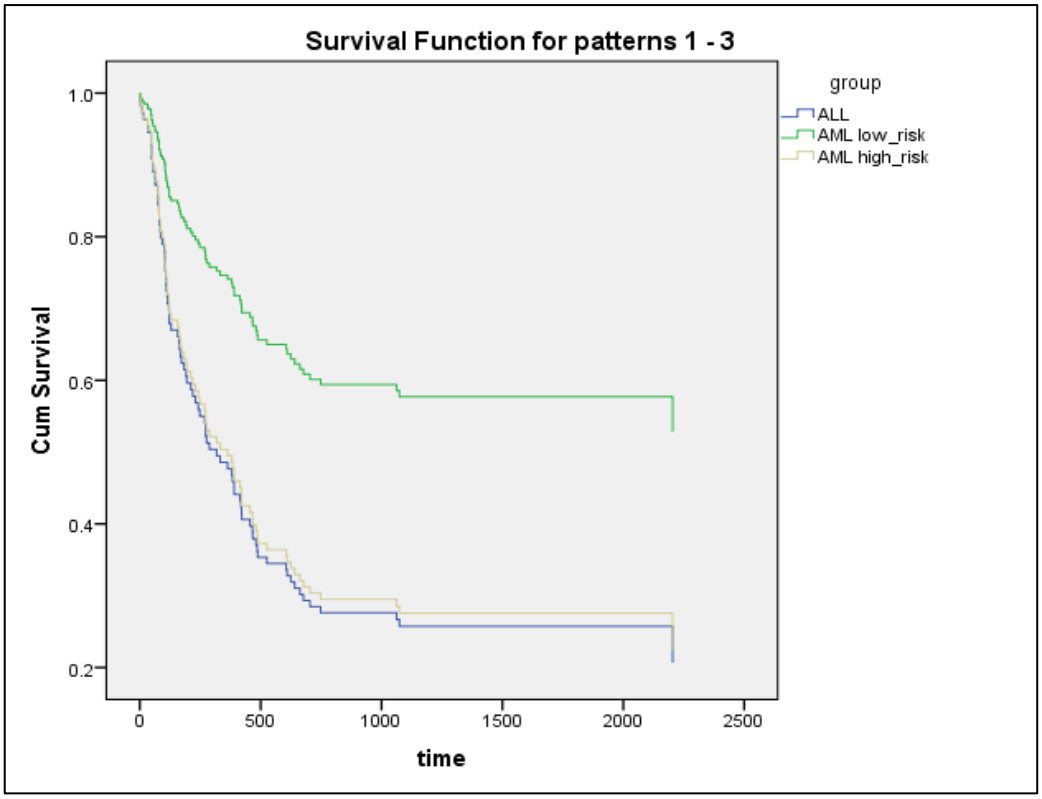
$x_0 = 5.68$
$\mu = 0.05$
$f(200 \mid \mu, x_0) = 2.30 \times 10^{-6}$
$S(200 \mid \mu, x_0) = 0.41$
$h(200 \mid \mu, x_0) = 5.62 \times 10^{-6}$

3.11 Συμπεράσματα και Σχόλια

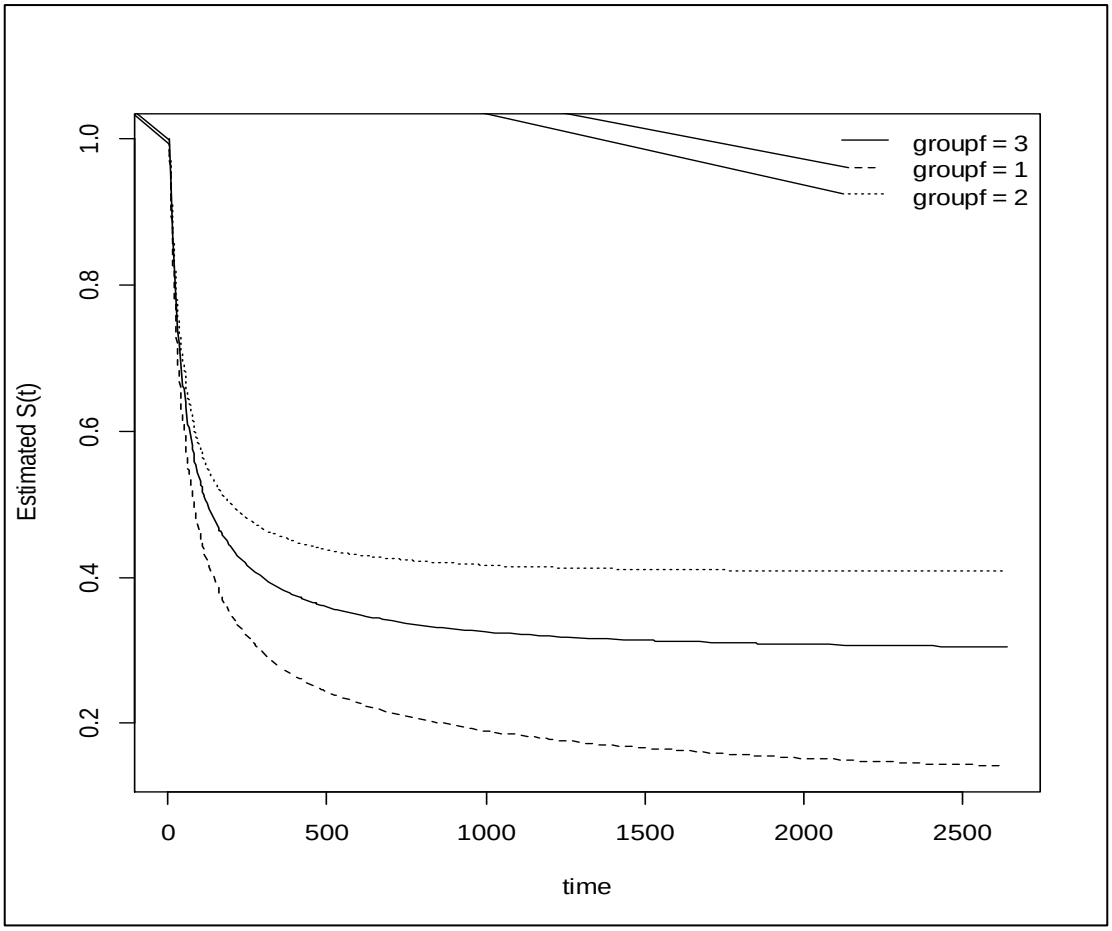
Συνοψίζοντας, από τις αναλύσεις που πραγματοποιήσαμε στις Παραγράφους 3.3, 3.4 και 3.5 συμπεραίνουμε ότι στους ασθενείς που πάσχουν από λευχαιμία τύπου AML high_risk το γεγονός (θάνατος ή υποτροπή) συμβαίνει πιο γρήγορα και άρα διατρέχουν μεγαλύτερο κίνδυνο συγκριτικά με τους ασθενείς που πάσχουν από τους άλλους τύπους λευχαιμίας. Στη συνέχεια, στα δεδομένα μας προσαρμόσαμε 3 διαφορετικά μοντέλα για να εξετάσουμε ποιοι παράγοντες επηρεάζουν την επιβίωση των ασθενών που πάσχουν από λευχαιμία μετά από μεταμόσχευση μυελού των οστών: το παραμετρικό μοντέλο της επιταχυνόμενης διακοπής, το ημι-παραμετρικό μοντέλο της αναλογικής διακινδύνευσης του Cox και το μοντέλο παλινδρόμησης threshold.

Από την προσαρμογή του μοντέλου της επιταχυνόμενης διακοπής καταλήξαμε στο συμπέρασμα ότι οι συμμεταβλητές fab, groupf και mtx είναι στατιστικά σημαντικές. Δηλαδή, βασικό ρόλο στην επιβίωση των ασθενών παίζουν το είδος της λευχαιμίας από την οποία πάσχουν οι ασθενείς (ALL, AML low_risk, AML high_risk), το αν έλαβαν κάποια προφύλαξη για την ασθένεια μοσχεύματος έναντι ξενιστή μετά τη μεταμόσχευση και ο βαθμός του κριτηρίου fab που δείχνει αν οι ασθενείς διατρέχουν μεγαλύτερο κίνδυνο υποτροπής ή θανάτου μετά τη μεταμόσχευση.

Από την προσαρμογή του μοντέλου του Cox συμπεραίνουμε ότι οι μεταβλητές fab και groupf είναι στατιστικά σημαντικές. Επειδή η υπόθεση της αναλογικής διακινδύνευσης δεν ισχύει πάντα, προσαρμόσαμε το μοντέλο παλινδρόμησης threshold ως εναλλακτικό του μοντέλου του Cox. Για το μοντέλο παλινδρόμησης threshold, όσον αφορά την αρχική κατάσταση σημαντικές προκύπτουν οι συμμεταβλητές recipient_age, donor_age και waiting_time και για την τάση μ οι συμμεταβλητές recipient_age, donor_age, fab και groupf.



Σχήμα 3.17: Οι συναρτήσεις επιβίωσης του μοντέλου του Cox συναρτήσει του χρόνου για τις κατηγορίες της μεταβλητής group



Σχήμα 3.18: Οι συναρτήσεις επιβίωσης του μοντέλου παλινδρόμησης threshold συναρτήσει του χρόνου για τις κατηγορίες της μεταβλητής group

Συγκρίνοντας τα Σχήματα 3.17 και 3.18 παρατηρούμε ότι και στις δύο περιπτώσεις η ομάδα 1 (ALL) είναι η ομάδα των ασθενών που τους συμβαίνει πιο γρήγορα το γεγονός. Ενώ η ομάδα 2 (AML low_risk) έχει την καλύτερη πρόγνωση για τους ασθενείς.

ΚΕΦΑΛΑΙΟ 4

Παράρτημα Α

Έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης στο μοντέλο του Cox στην R

Συνάρτηση: **cox.zph (fit, transform="km", global=TRUE)**

Περιγραφή: Η συνάρτηση αυτή ελέγχει αν ισχύει η υπόθεση της αναλογικής διακινδύνευσης στο μοντέλο του Cox.

fit: είναι το προσαρμοσμένο μοντέλο του Cox που δημιουργήθηκε με τη χρήση της συνάρτησης coxph.

transform: είναι ένας χαρακτήρας τύπου string που προσδιορίζει με ποιο τρόπο θα μετατραπούν οι χρόνοι επιβίωσης πριν γίνει ο έλεγχος. Πιθανές τιμές είναι οι "km" (Kaplan-Meier), "rank" (βαθμός διάταξης) και "identity", ή κάποια άλλη συνάρτηση.

global: πέρα από τον έλεγχο που γίνεται για κάθε μεταβλητή ξεχωριστά, γίνεται και ένας X^2 έλεγχος για όλες τις μεταβλητές συνολικά.

Το αντικείμενο που δημιουργείται από τη χρήση αυτής της συνάρτησης είναι ένας πίνακας με μία γραμμή για κάθε μεταβλητή και μία συμπληρωματική γραμμή σε περίπτωση που γίνει και το global test. Οι στήλες του πίνακα περιέχουν το συντελεστή συσχέτισης μεταξύ του χρόνου που έχει μετατραπεί και των κλιμακοποιημένων υπολοίπων Schoenfeld, ένα X^2 τεστ και μία p-τιμή για κάθε έλεγχο. Για το global test δεν υπάρχει κάποια συσχέτιση, οπότε μπαίνει η τιμή NA.

Παράρτημα Β

Πακέτο threg στην R (Xiao, 2015)

Περιγραφή: Σε αυτό το πακέτο γίνεται προσαρμογή του μοντέλου παλινδρόμησης threshold που σχετίζεται με το χρόνο πρώτης μετάβασης της στοχαστικής διαδικασίας Wiener με ένα όριο. Το μοντέλο αυτό έχει πολλές εφαρμογές στα δεδομένα διάρκειας ζωής.

Διαθέσιμες συναρτήσεις:

- **threg (formula, data)**

Η συνάρτηση αυτή προσαρμόζει το μοντέλο παλινδρόμησης threshold, θεωρώντας τη διαδικασία Wiener. Χρησιμοποιεί τη μέθοδο μεγίστης πιθανοφάνειας για τον υπολογισμό των συντελεστών παλινδρόμησης των συμμεταβλητών, των τυπικών σφαλμάτων των συντελεστών και τις p-τιμές.

formula: είναι ένα αντικείμενο formula που περιλαμβάνει την απόκριση και τις ανεξάρτητες μεταβλητές. Δηλαδή, η απόκριση είναι ένα αντικείμενο επιβίωσης που επιστρέφεται από τη συνάρτηση Surv. Μετά την απόκριση έχουμε το σύμβολο ~ δεξιά του οποίου έχουμε τις ανεξάρτητες μεταβλητές. Πιο συγκεκριμένα, χρησιμοποιούμε το σύμβολο | , οπότε μεταξύ των συμβόλων ~ και | βάζουμε τις μεταβλητές που θα χρησιμοποιήσουμε για την πρόβλεψη της παραμέτρου $\ln x_0$ στην παλινδρόμηση threshold, ενώ μετά το σύμβολο | βάζουμε τις μεταβλητές για την πρόβλεψη της τάση μ στην παλινδρόμηση threshold.

data: είναι το σύνολο των δεδομένων που θέλουμε να μελετήσουμε. Το σύνολο των δεδομένων μας πρέπει να έχει μία μεταβλητή που να δείχνει το χρόνο επιβίωσης και μία που να καθορίζει αν έχουμε αποκομμένες παρατηρήσεις ή όχι.

- **hr (object, var, timevalue, scenario)**

Η συνάρτηση αυτή υπολογίζει τις εκτιμήσεις των αναλογιών διακινδύνευσης σε συγκεκριμένο χρόνο και κάτω από συγκεκριμένες συνθήκες.

object: είναι ένα αντικείμενο threg.

var: αναφέρεται στην κατηγορική μεταβλητή που χρησιμοποιείται για τον υπολογισμό των αναλογιών διακινδύνευσης. Πρέπει να είναι μία κατηγορική

μεταβλητή που έχει χρησιμοποιηθεί στη συνάρτηση threg() και επιστρέφει το αντικείμενο threg.

timevalue: αναφέρεται στον χρόνο στον οποίο υπολογίζονται οι αναλογίες διακινδύνευσης. Επιτρέπεται να χρησιμοποιηθεί και διάνυσμα.

scenario: αναφέρεται σε συγκεκριμένες συνθήκες κάτω από τις οποίες θα υπολογιστούν οι αναλογίες διακινδύνευσης.

- **predict (object, timevalue, scenario,...)**

Η συνάρτηση αυτή χρησιμοποιείται με σκοπό να προβλέψει την τιμή της αρχικής κατάστασης της διαδικασίας x_0, την τιμή της παραμέτρου μ, τη συνάρτηση πυκνότητας πιθανότητας $f(t \mid \mu, x_0)$, τη συνάρτηση επιβίωσης $S(t \mid \mu, x_0)$ και τη συνάρτηση διακινδύνευσης $h(t \mid \mu, x_0)$ σε συγκεκριμένο χρόνο και κάτω από συγκεκριμένες συνθήκες. Σε αυτή τη συνάρτηση, όμως, πρέπει να βάλουμε τιμές σε όλες τις μεταβλητές του μοντέλου ακόμη και στις ψευδομεταβλητές της κατηγορικής μεταβλητής που έχουμε, ενώ στη συνάρτηση hr() δεν χρειάζεται.

object: είναι ένα αντικείμενο threg.

timevalue: αναφέρεται στον χρόνο στον οποίο υπολογίζονται οι εκτιμώμενες τιμές. Επιτρέπεται να χρησιμοποιηθεί και διάνυσμα. Αν παραλειφθεί η τιμή του timevalue τότε υπολογίζονται οι προβλεπόμενες τιμές για το χρόνο της μελέτης.

scenario: αναφέρεται στις τιμές όλων των μεταβλητών στη συγκεκριμένη χρονική στιγμή όπου θα υπολογιστούν οι προβλεπόμενες τιμές.

…: για μελλοντικές μεθόδους.

- **plot(x, var, scenario, graph, nolegend=0, nocolor=0,...)**

Η συνάρτηση αυτή απεικονίζει τις γραφικές παραστάσεις των εκτιμήσεων των συναρτήσεων επιβίωσης, διακινδύνευσης, πυκνότητας πιθανότητας στα διαφορετικά επίπεδα μιας κατηγορικής μεταβλητής η οποία έχει χρησιμοποιηθεί στο μοντέλο παλινδρόμησης threshold στη συνάρτηση threg().

x: ένα αντικείμενο threg.

var: αναφέρεται στην κατηγορική μεταβλητή για τα επίπεδα της οποίας θα κατασκευαστούν οι γραφικές παραστάσεις.

scenario: αναφέρεται στις συνθήκες κάτω από τις οποίες θα γίνουν οι γραφικές παραστάσεις.

graph: προσδιορίζει τον τύπο των γραφικών που θα κατασκευαστούν. Δηλαδή, η επιλογή “hz” είναι για τη συνάρτηση διακινδύνευσης, “sv” για τη συνάρτηση επιβίωσης και “ds” για τη συνάρτηση πυκνότητας πιθανότητας.

nolegend: αν θέσουμε nolegend=1 τότε δε θα μπει λεζάντα στις γραφικές παραστάσεις.

nocolor: αν θέσουμε nocolor=1 τότε όλες οι καμπύλες θα είναι σε μαύρο χρώμα.

…: για μελλοντικές μεθόδους

Το πακέτο threg περιλαμβάνει και δύο πακέτα δεδομένων:

bmt: για την επιβίωση 137 ασθενών οξείας λευχαιμίας που έκαναν μεταμόσχευση μυελού των οστών. Αυτά τα δεδομένα χρησιμοποιήθηκαν για την στατιστική ανάλυση που πραγματοποιήθηκε στο κεφάλαιο 3.

ikr: είναι δεδομένα 42 ασθενών που πάσχουν από λευχαιμία και ελέγχθηκαν για το αν παρουσίασαν υποτροπή και για το πόσο κράτησε η απαλλαγή τους από την ασθένεια.

Βιβλιογραφία

Aalen, O.O. & Gjessing, H.K. (2001). Understanding the shape of the hazard rate: A process point of view. *Statistical Science*, **16**, 1-22

Aalen, O.O. & Gjessing, H.K. (2004). Survival models based on the Ornstein–Uhlenbeck process. *Lifetime Data Analysis*, **10**, 407-423.

Aalen, O.O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics*, **6**, 701-726.

Caroni, C. (2004). Diagnostics for Cox's proportional hazards model. In M.S. Nikulin, N. Balakrishnan, M. Mesbah and N. Limnios (eds) *Parametric and Semiparametric Models with Applications to Reliability, Survival Analysis, and Quality of Life* in honour of Prof. Catherine Huber, Birkhauser, Boston, 27-38.

Caroni, C. (2017). *First Hitting Time Regression Models: Lifetime Data Analysis Based on Underlying Stochastic Processes.* London: Wiley-ISTE.

Collett, D. (2003). *Modelling Survival Data in Medical Research.* (2nd ed.) Boca Raton: Chapman & Hall/CRC.

Cox, D.R. & Snell, E.J. (1968). A general definition of residuals (with Discussion). *Journal of the Royal Statistical Society. Series B*, **30**, 248-275.

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

Cox, D.R. (1990). Role of models in statistical analysis. *Statistical Science*, **5**, 169-174.

Davison, A.C. (2003). *Statistical Models.* Cambridge: University Press.

Folks, J.L. & Chhikara, R.S. (1978). The inverse gaussian distribution and its statistical application--a review. *Journal of the Royal Statistical Society. Series B*, **40**, 263-289.

Fox, J. (2002). Cox proportional-hazards regression for survival data. *An R and S-PLUS Companion to Applied Regression*. (https://socserv.socsci.mcmaster.ca/jfox/Books/Companion/appendix/Appendix-Cox-Regression.pdf)

Grambsch, P.M. & Therneau, T.M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81**, 515-526.

Hou, W.H., Chuang, H.Y., & Lee, M. L. T. (2016). A threshold regression model to predict return to work after traumatic limb injury. *Injury*, **47**, 483-489.

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

Klein, J.P. & Moeschberger, M.L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data.* New York: Springer.

Lancaster, T. (1972). A stochastic model for the duration of a strike. *Journal of the Royal Statistical Society. Series A*, **135**, 257-271.

Lee, M.L.T. & Whitmore, G.A. (2006). Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. *Statistical Science*, **21**, 501-513.

Lee, M.L.T. & Whitmore, G.A. (2010). Proportional hazards and threshold regression: their theoretical and practical connections. *Lifetime Data Analysis*, **16**, 196-214.

Lee, M.L.T., Whitmore, G.A., Laden, F., Hart, J. E., & Garshick, E. (2004). Assessing lung cancer risk in railroad workers using a first hitting time regression model. *Environmetrics*, **15**, 501-512.

Lee, M.L.T., Whitmore, G. A., Laden, F., Hart, J. E., & Garshick, E. (2009). A case-control study relating railroad worker mortality to diesel exhaust exposure using a threshold regression model. *Journal of Statistical Planning and Inference*, **139**, 1633-1642.

Lee, M.L.T., Whitmore, G. A., & Rosner, B (2010). Benefits of threshold regression: a case-study comparison with Cox proportional hazards regression. In V.V. Rykov, N. Balakrishnan, M.S. Nikulin (eds). *Mathematical and Statistical Models and Methods in Reliability: Applications to Medicine, Finance, and Quality Control,* Birkhauser, New York, 359-370.

Nelson, W.B. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, **14**, 945-966.

Ricciardi, L.M. & Sato, S. (1988). First-passage-time density and moments of the Ornstein-Uhlenbeck process. *Journal of Applied Probability*, **25**, 43-57.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, **69**, 239-241.

Stogiannis, D., Caroni, C., Anagnostopoulos, C. E., & Toumpoulis, I. K. (2011). Comparing first hitting time and proportional hazards regression models. *Journal of Applied Statistics*, **38**, 1483-1492.

Therneau, T.M. & Grambsch, P.M. (2000). *Modeling Survival Data: Extending the Cox Model.* New York: Springer.

Therneau, T.M., Crowson, C., & Atkinson, E. (2017). Using time dependent covariates and time dependent coefficients in the Cox model. Survival V*ignettes*. (http://cran.es.r-project.org/web/packages/survival/vignettes/timedep.pdf)

Xiao, T. (2015). ***threg:*** *Threshold Regression*. R package version 1.0.3, http://CRAN.R-project.org/package=threg.

Xiao, T., Whitmore, G. A., He, X., & Lee, M. L. T. (2015). The R package threg to implement threshold regression models. *Journal of Statistical Software*, **66**, 1-16.

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

Καρώνη, Χ. (2017). *Στατιστικά Μοντέλα Παλινδρόμησης.* Αθήνα: Εκδόσεις Συμεών.