ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Διπλωματική Εργασία

Στατιστικά Μοντέλα Επιβίωσης και Εφαρμογή σε Ασθενείς με Πολλαπλούν Μυέλωμα

Αιμιλία Πέππα

Επιβλέπουσα Καθηγήτρια: Καρώνη Χρυσηΐς

Τριμελής Επιτροπή

Χ. Καρώνη,	Κ. Παυλοπούλου,	Π. Στεφανέας,
Καθηγήτρια, Ε.Μ.Π	ΕΔΙΠ, Ε.Μ.Π	Αν. Καθηγητής, Ε.Μ.Π

Αθήνα, Ιούνιος 2023

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα εργασία αποτελεί διπλωματική εργασία στο πλαίσιο των σπουδών μου στην σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών με ειδίκευση στον τομέα της Στατιστικής και πραγματοποιήθηκε υπό την επίβλεψη της καθηγήτριας του τομέα Μαθηματικών, κυρία Χρυσηίδα Καρώνη, την οποία ευχαριστώ θερμά για την επίβλεψη και την καθοδήγηση που μου έδωσε καθ'όλη την διάρκεια της εκπόνησης της διπλωματικής εργασίας και για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον αντικείμενο.

Ευχαριστώ, επίσης, τους φίλους και συμφοιτητές που ήταν πάντα δίπλα μου ενθαρρύνοντας και στηρίζοντας με στις δυσκολίες όπου αντιμετώπισα αλλά και στις επιτυχίες που ήρθαν μέσα στα φοιτητικά μου χρόνια. Ιδιαίτερα ευχαριστώ τους Αλεξία και Πάνο για την ενθάρρυνση και την στήριξη τους.

Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου, Προκόπη και Χαρά, για την στήριξη και την εμπιστοσύνη που μου έδωσαν όλα αυτά τα χρόνια, καθώς και τον αδερφό μου Γιώργο για την βοήθεια και την συμπαράσταση στην διάρκεια των φοιτητικών μου χρόνων.

Αιμιλία Πέππα

© (2021) Εθνικό Μετσόβιο Πολυτεχνείο. All rights Reserved. Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς το συγγραφέα. Οι απόψεις και τα συμπεράσματα που περιέχονται σ'αυτό το έγγραφο εκφράζουν το συγγραφέα και δεν πρέπει να ερμηνευτεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Η ανάλυση επιβίωσης είναι ένας κλάδος της στατιστικής που ασχολείται με τη μελέτη του χρόνου πριν συμβεί ένα γεγονός και βασίζεται σε πολλές κατανομές με και χωρίς παραμέτρους και βρίσκει πολυάριθμες εφαρμογές σε δεδομένα διάρκειας ζωής που επηρεάζονται από διάφορες μεταβλητές. Χρησιμοποιείται κυρίως για βιοϊατρικά δεδομένα μέσω της βιοστατιστικής, αλλά αυτό δεν επηρεάζει τη χρησιμότητά του σε άλλα επιστημονικά πεδία όπως η αναλογιστική και μηχανολογικές εφαρμογές.

Στόχος αυτής της εργασίας είναι η ανάλυση δεδομένων από ασθενείς με πολλαπλούν μυέλωμα χρησιμοποιώντας τεχνικές ανάλυσης επιβίωσης. Στα επόμενα κεφάλαια αναλύονται οι βασικοί ορισμοί της ανάλυσης επιβίωσης και μέθοδοι που χρησιμοποιούνται για την εξαγωγή αξιόπιστων συμπερασμάτων. Γίνεται βιβλιογραφική ανάλυση των παραμετρικών και μη-παραμετρικών μοντέλων και μεθόδων που χρησιμοποιούνται, καθώς και οι αιτίες της επιρροής διάφορων συμμεταβλητών στην επιβίωση των ασθενών του δείγματος. Επιπλέον, αναφέρονται εκτενώς οι έλεγχοι με τους οποίους εξετάζουμε την προσαρμογή των μοντέλων και την επιρροή των συμμεταβλητών.

Ακόμη, γίνεται αναφορά της χρήσης παραμετρικών μοντέλων επιβίωσης στην αναλογιστική επιστήμη και πώς αυτά συνδράμουν στην δημιουργία και ορθή εφαρμογή των ασφαλιστικών συμβολαίων ζωής.

Τέλος, πραγματοποιείται μια ανάλυση των δεδομένων, με την βοήθεια των στατιστικών πακέτων R και MINITAB, με τις μεθόδους που αναφέρθηκαν με σκοπό την εύρεση των παραμέτρων που επηρεάζουν την επιβίωση των ασθενών με πολλαπλούν μυέλωμα και παρουσιάζονται τα συμπεράσματα από την μελέτη αυτή.

Λέξεις Κλειδιά: Ανάλυση Επιβίωσης, Εκτιμήτρια Kaplan-Meier, Εκτιμήτρια Nelson-Aalen, Μοντέλο Αναλογικής Διακινδύνευσης του Cox, Μοντέλο Επιταχυνόμενης Διακοπής Weibull, Υπόλοιπα Cox-Snell, Υπόλοιπα Schoenfeld, Πολλαπλουν Μυέλωμα.

Abstract

Survival analysis is a branch of statistics that deals with the study of time before an event occurs. It is based on various distributions with and without parameters and finds numerous applications in life duration data influenced by various variables. It is primarily used for biomedical data through bio-statistics, but its usefulness extends to other scientific fields such as actuarial and engineering applications.

The objective of this study is the analysis of data from patients with multiple myeloma using survival analysis techniques. The following chapters analyze the basic definitions of survival analysis and the methods used to draw reliable conclusions. A literature review is conducted on parametric and non-parametric models and methods used, as well as the factors influencing patient survival in the sample. Additionally, extensive discussions are provided on the tests used to examine model fitting and the influence of variables.

Furthermore, the use of parametric survival models in actuarial science is mentioned, highlighting how they contribute to the creation and proper implementation of life insurance contracts.

Finally, a data analysis is performed using the statistical packages R and MINITAB, applying the methods mentioned to identify the parameters that affect the survival of patients with multiple myeloma, and the conclusions of this study are presented.

Keywords: Survival Analysis, Kaplan-Meier Estimator, Nelson-Aalen Estimator, Cox Proportional Hazard Model, Weibull Accelerated Failure Time Model, Cox-Snell Residuals, Schoenfeld Residuals, Multiple Myeloma.

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ ΚΑΙ ΓΡΑΦΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΠΕΡΙΕΧΟΜΕΝΑ

1. Ανάλυση Επιβίωσης

1.1 Εισαγωγή

Σε έναν κόσμο ο οποίος περικλείεται συνεχώς από πληροφορίες και δεδομένα, η χρήση της Στατιστικής Επιστήμης σε συνδυασμό με την Θεωρία Πιθανοτήτων είναι αναπόσπαστο κομμάτι της καθημερινότητας σχεδόν όλων των επαγγελματικών τομέων.

Σε τομείς όπως η ψυχολογία, τα οικονομικά, η βιολογία, η ιατρική, η επιδημιολογία κ.α., η στατιστική επιστήμη έχει προσφέρει πληθώρα εφαρμογών.

Αυτό το είδος ανάλυσης δεδομένων ονομάζεται *Ανάλυση Αξιοπιστίας (Reliability Analysis),* αν αυτή εφαρμόζεται στον κλάδο των θετικών επιστημών, ενώ αν εφαρμοστεί σε βιοϊατρικά δεδομένα ονομάζεται *Ανάλυση Επιβίωσης (Survival Analysis).*

Ο όρος Ανάλυση Επιβίωσης προήλθε, αρχικά, από την ανάγκη να μελετήσουμε τον χρόνο μεταξύ της θεραπείας μέχρι ότου επέλθει ο θάνατος, για αυτόν τον λόγο ονομάστηκε αντίστοιχα. Χρησιμοποιείται, επίσης για να περιγράψει την ανάλυση των δεδομένων με τη μορφή χρόνων από μια καλά καθορισμένη χρονική προέλευση μέχρι την εμφάνιση συγκεκριμένων συμβάντων. Όσον αφορά την ιατρική επιστήμη, σε αυτούς τους χρόνους συνήθως αντιστοιχούν η στρατολόγηση ενός ατόμου σε μια πειραματική μελέτη, όπως για παράδειγμα μια κλινική δοκιμή για την σύγκριση δύο φαρμάκων ή θεραπειών γενικότερα. Στην περίπτωση που το τελικό σημείο ενός ασθενή είναι ο θάνατος τότε έχουμε κυριολεκτικά χρόνους επιβίωσης.

Επιπλέον, οι τεχνικές αυτές μπορούν να φανούν χρήσιμες και σε άλλες περιοχές και περιστατικά στα οποία δεν υπάρχει υποχρεωτικά ως τελικό σημείο ο θάνατος του ατόμου. Τέτοια μπορεί να είναι η επανεμφάνιση συμπτωμάτων ή η δράση ενός φαρμάκου. Για αυτόν τον λόγο οι παρατηρήσεις που μελετάμε αναφέρονται ως δεδομένα χρόνου μέχρι το συμβάν.

Ωστόσο, αυτό το οποίο κάνει τα δεδομένα επιβίωσης ή αξιοπιστίας να χρήζουν διαφορετικής μεθοδολογίας και που δεν μας επιτρέπει την χρήση των κλασσικών στατιστικών τεχνικών είναι το γεγονός ότι στους χρόνους των παρατηρήσεων ενός δείγματος είναι πιθανόν να υπάρξουν αποκομμένες παρατηρήσεις, καθώς η παρακολούθηση των ατόμων ίσως

ξεκινά αφού έχει επέλθει κάποιο χρονικό διάστημα. Αργότερα θα αναλύσουμε περαιτέρω τις περιπτώσεις αυτές.

Στην παρούσα εργασία θα ασχοληθούμε με την στατιστική ανάλυση στους χώρους της ιατρικής και της αναλογιστικής,οι οποίοι αν και φαίνεται να μην σχετίζονται μεταξύ τους, έχουν πολλά κοινά, καθώς και στα δύο αυτά πεδία χρειάζεται σε κάποιες περιπτώσεις να μελετήσουμε την διάρκεια ζωής μέχρις ότου να συμβεί ένα συγκεκριμένο γεγονός (π.χ. θάνατος, λήξη ασφαλιστικού συμβολαίου, βλάβη σε μηχάνημα κ.ο.κ). Θα αναφερθούμε στην ανάλυση επιβίωσης αναλύοντας παραμετρικά και μη-παραμετρικά μοντέλα που χρησιμοποιούνται, καθώς και μεθοδολογίες για τις περιπτώσεις αυτές. Τέλος, θα πραγματοποιήσουμε μια ανάλυση σε δεδομένα ασθενών οι οποίοι πάσχουν από πολλαπλό μυέλωμα χρησιμοποιώντας τις μεθόδους που αναφέραμε και με την χρήση της επιστημονικής γλώσσας προγραμματισμού R.

1.2 Αποκομμένες Παρατηρήσεις

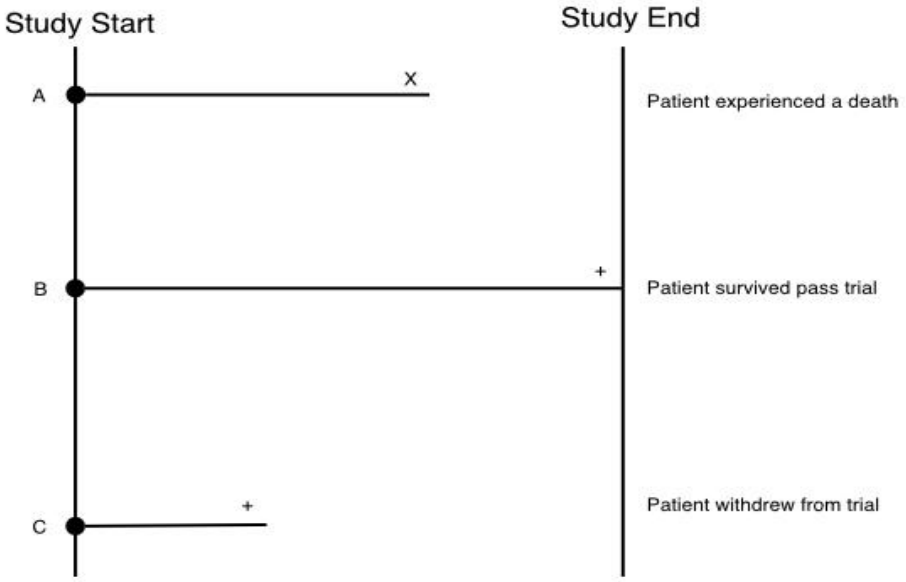
Τα δεδομένα χρόνου επιβίωσης παρουσιάζουν ένα σημαντικό χαρακτηριστικό το οποίο τα κάνει να διαφέρουν από άλλα στατιστικά δεδομένα και να χρήζουν διαφορετικών πρακτικών από τις κλασσικές στατιστικές μεθόδους. Αυτό το χαρακτηριστικό ονομάζεται αποκοπή παρατηρήσεων και συμβαίνει όταν ο χρόνος επιβίωσης ενός ατόμου δεν παρατηρείται στο τελικό σημείο της έρευνας.

Αυτό μπορεί να συμβαίνει επειδή κάποια άτομα είναι ακόμα ζωντανά ενώ κάποια άλλα όχι, ή ακόμη επειδή το άτομο δεν παρακολουθείται εκείνη την στιγμή και δεν είναι γνωστή η κατάσταση του. Για παράδειγμα κατά την διάρκεια μιας κλινικής δοκιμής κάποιο άτομο μετακόμισε σε κάποια άλλη πόλη, επομένως, αποχώρησε από την μελέτη, και το μόνο δεδομένο που έχουμε για την κατάστασή του είναι η τελευταία φορά που έγινε κάποια εξέταση ή έλεγχος.

1.2.1 Δεξιά Αποκοπή Παρατηρήσεων

Σε κάθε περίπτωση ένα άτομο το οποίο εισήλθε στην μελέτη την χρονική στιγμή t_0 θα πεθάνει την χρονική στιγμή $t_0 + t$ όπου το t είναι άγνωστο επειδή είναι ακόμα ζωντανός ή δεν παρακολουθείται πια. Τότε, λέμε ότι η τελευταία φορά που είχαμε κάποια πληροφορία για την κατάσταση του ήταν η χρονική στιγμή $t_0 + c$ και θα ονομάσουμε τον χρόνο c ως αποκομμένος χρόνος επιβίωσης. Αυτός ο τύπος αποκοπής

ονομάζεται δεξιά αποκοπή και ο πραγματικός χρόνος επιβίωσης είναι μεγαλύτερος από τον αποκομμένο, αλλά παρόλα αυτά παραμένει άγνωστος. *(Collett, 2014)*



Σχήμα 1.1: *Απεικόνιση δεξιά αποκομμένων δεδομένων τριών ασθενών.*

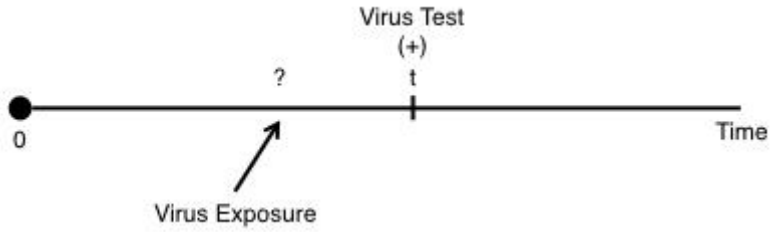
Στο *Σχήμα 1.1* βλέπουμε ένα παράδειγμα με τρεις ασθενείς και την πορεία τους μέχρι το τέλος της μελέτης:

- Ο Α ασθενής απεβίωσε πριν το πέρας της μελέτης,
- Ο Β ασθενής επιβιώνει μετά το πέρας της μελέτης,
- ενώ ο C ασθενής έχει αποχωρήσει από την μελέτη πριν το τέλος της.

1.2.2 Αριστερή Αποκοπή Παρατηρήσεων

Υπάρχουν περιπτώσεις, όμως, όπου τα άτομα τα οποία μελετάμε εισέρχονται στην έρευνα την χρονική στιγμή $t = c$ και δεν μπορούμε γνωρίζουμε τι συνέβη σε αυτές προτού ξεκινήσει η παρακολούθηση τους. Για παράδειγμα, σε μια μελέτη για την υποτροπή κάποιου όγκου μετά από επέμβαση προ τριών μηνών, για τους ασθενείς για τους οποίους υπάρχει όντως υποτροπή, ο χρόνος που συνέβη αυτή είναι άγνωστος. Ένα παράδειγμα από την καθημερινότητα μας είναι η διεξαγωγή ενός τεστ για να εξετάσουμε αν ένα άτομο έχει προσβληθεί από έναν ιό. Στο *Σχήμα 1.2* παρατηρούμε την χρονική στιγμή κατά την οποία γίνεται η διάγνωση του θετικού τεστ. Πριν διεξαχθεί το τεστ, δεν μπορούμε να γνωρίζουμε την ακριβή χρονική στιγμή όπου εκτέθηκε ο ασθενής στον ιό.

Τέτοιες παρατηρήσεις ονομάζονται <u>αριστερά αποκομμένες</u> και έχουμε δεδομένα μόνο για $t > c$.

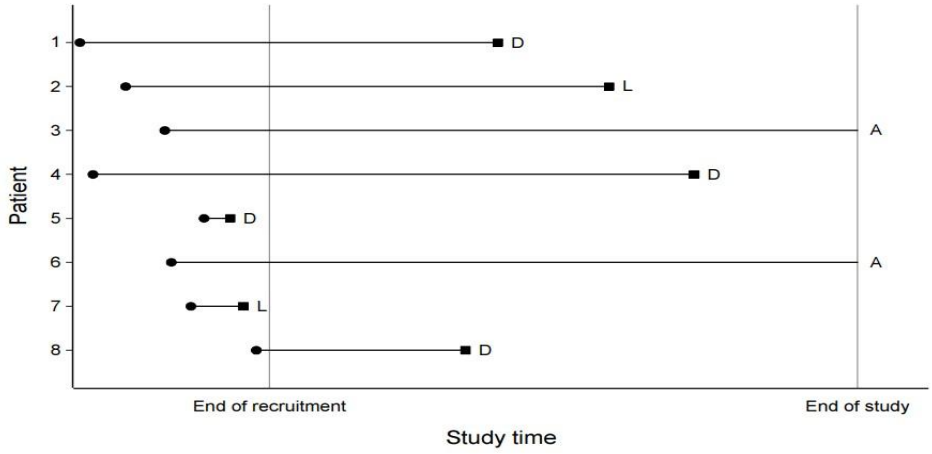


Σχήμα 1.2: *Απεικόνιση αριστερά αποκομμένης παρατήρησης.*

1.2.3 <u>Αποκομμένα Διαστήματα Επιβίωσης</u>

Εδώ θα πρέπει να σημειώσουμε ότι υπάρχει το ενδεχόμενο να έχουμε συνδυασμό αυτών των ειδών αποκοπής. Δηλαδή, σε μία έρευνα που μπορεί να πραγματοποιούμε να υπάρχουν δεξιά και αριστερά αποκομμένες παρατηρήσεις. Για παράδειγμα, στα ασφαλιστικά συμβόλαια ζωής, ο πελάτης θα προσέλθει την χρονική στιγμή t_A όπου συμβολίζει την αρχή συμβολαίου, και η παρακολούθηση θα ισχύσει μέχρις ότου να λήξει το ασφαλιστικό συμβόλαιο, δηλαδή, την t_B.

Αφού, λοιπόν, δεν γνωρίζουμε τι συνέβη πριν την χρονική στιγμή t_A (αρχή συμβολαίου) και μετά την χρονική στιγμή t_B (λήξη συμβολαίου), θα έχουμε αριστερά αποκομμένες παρατηρήσεις για $t < t_A$ και δεξιά αποκομμένες για $t > t_B$ όπως απεικονίζονται στο *Σχήμα 1.3*.



Σχήμα 1.3: *Απεικόνιση παρατηρήσεων σε αποκομμένα διαστήματα.*

1.3 Βασικές Συναρτήσεις

Οι δύο πιο σημαντικές συναρτήσεις που χρησιμοποιούνται για την ανάλυση επιβίωσης, οι οποίες περιγράφουν την κατανομή του χρόνου είναι η συνάρτηση επιβίωσης (ή συνάρτηση αξιοπιστίας) και η συνάρτηση διακινδύνευσης.

1.3.1 Συνάρτηση Επιβίωσης

Η συνάρτηση επιβίωσης αναφέρεται στην πιθανότητα επιβίωσης ενός οργανισμού ή ενός συστήματος σε σχέση με τον χρόνο και χρησιμοποιείται για να αναλύσει την ποσοτική σχέση μεταξύ επιβίωσης και χρόνου ή να αξιολογήσει την επίδραση άλλων παραγόντων στην επιβίωση του ατόμου ή του συστήματος.

Θεωρούμε τ.μ $T > 0$ η οποία εκφράζει την "διάρκεια ζωής" ενός πληθυσμού και είναι συνεχής τ.μ με συνάρτηση πυκνότητας πιθανότητας $f(t), t \geq 0$ και κατά συνέπεια με συνάρτηση κατανομής:

$$F(t) = P[T \leq t] = \int_0^t f(u)du$$

Ισχύει ότι η *F(t)* είναι αύξουσα με $\lim_{t\to 0} F(t) = 0$ και $\lim_{t\to\infty} F(t) = 1$.

Τότε,η συνάρτηση επιβίωσης όπου ορίζεται ως η πιθανότητα επιβίωσης ενός ατόμου μετά την χρονική στιγμή *t*, δηλαδή, η πιθανότητα η διάρκεια ζωής να είναι μεγαλύτερη του *t*, δίνεται από την σχέση:

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

όπου ισχύει ότι η συνάρτηση πυκνότητας πιθανότητας της τ.μ. T προκύπτει:

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

1.3.2 <u>Συνάρτηση Διακινδύνευσης</u>

Στην ανάλυση επιβίωσης, η συνάρτηση Διακινδύνευσης αναφέρεται στην πιθανότητα να συμβεί ένα συμβάν, όπως θάνατος ή αποχώρηση από την παρακολούθηση σε ένα δεδομένο χρονικό διάστημα, στο χρόνο που έχει ήδη παρέλθει.

Αν η συνάρτηση διακινδύνευσης είναι σταθερή προκύπτει ότι η πιθανότητα να συμβεί ένα συμβάν είναι ίδια για κάθε χρονικό διάστημα, ανεξαρτήτως του πότε έχουν συμβεί τα προηγούμενα συμβάντα. Ωστόσο, στην πραγματικότητα είναι πολύ δύσκολο να παραμένει σταθερή και μπορεί να μεταβάλλεται με το χρόνο.

Ορίζουμε ως συνάρτηση διακινδύνευσης της τ.μ Τ την :

$$h(t) = \lim_{\delta t \to 0} \frac{\frac{(S(t) - S(t+\delta t)}{S(t)}}{\delta t} = \frac{f(t)}{S(t)} \qquad (1.2)$$

και εκφράζει τον στιγμιαίο ρυθμό διακοπής ενός αντικειμένου στο χρονικό διάστημα *(t, t+δt)*, αν θεωρήσουμε ότι η επιβίωση την στιγμή t είναι δεδομένη.

Από την συνάρτηση διακινδύνευσης προκύπτει η σωρευτική συνάρτηση διακινδύνευσης η οποία ορίζεται ως:

$$H(t) = \int_0^t h(u)du \qquad (1.3)$$

Δεδομένου ότι η συνάρτηση επιβίωσης είναι η πιθανότητα να επιβιώσει κάποιος για ένα δεδομένο διάστημα, η συνάρτηση διακινδύνευσης είναι η αντίστροφη της συνάρτησης επιβίωσης και μπορεί να υπολογιστεί συνδυάζοντας τις εξισώσεις *(1.1), (1.2), (1.3)*:

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

$$= \int_0^t \frac{-S'(u)}{S(u)} du$$

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

Καταλήγουμε στην σχέση:

$$H(t) = - lnS(t)$$

η οποία είναι ισοδύναμη με την σχέση:

$$S(t) = e^{-H(t)}$$

(Καρώνη, 2009)

1. Μη-Παραμετρικά Μοντέλα Επιβίωσης

Για να πραγματοποιηθεί η ανάλυση και η επιλογή ενός μοντέλου για ένα σύνολο δεδομένων διάρκειας ζωής θα χρειαστεί αρχικά να δημιουργήσουμε γραφικές παραστάσεις. Αυτές οι γραφικές παραστάσεις θα μας βοηθήσουν να αξιολογήσουμε την συμπεριφορά των συναρτήσεων επιβίωσης και διακινδύνευσης.

Σε αυτό το κεφάλαιο θα αναλύσουμε τις μεθόδους οι οποίες χρησιμοποιούνται για την εκτίμηση των συναρτήσεων επιβίωσης και διακινδύνευσης. Αυτές οι μέθοδοι ονομάζονται Μη-Παραμετρικές καθώς δεν απαιτούν συγκεκριμένες υποθέσεις για την κατανομή των χρόνων επιβίωσης.

2.1 Εκτιμώμενη Συνάρτηση Επιβίωσης

Δεδομένου ότι τα δεδομένα χρόνων επιβίωσης τα οποία θέλουμε να μελετήσουμε είναι μη-αποκομμένα η συνάρτηση επιβίωσης δίνεται:

$$\hat{S}(t) = \frac{j}{n}$$

όπου j=”Αριθμός ατόμων με χρόνους επιβίωσης $\geq t$ “

και n=”Συνολικός Αριθμός ατόμων στο δείγμα”.

2.2 Εκτιμήτρια Kaplan-Meier

Όπως αναφέραμε σε προηγούμενο κεφάλαιο, είναι πολύ συχνό το φαινόμενο των αποκομμένων παρατηρήσεων σε ένα δείγμα. Σε αυτές τις περιπτώσεις χρησιμοποιούμε την μέθοδο που αναπτύχθηκε από τους Kaplan και Meier (*1958*) και ονομάζεται εκτιμήτρια Kaplan-Meier της συνάρτησης επιβίωσης.

Η εκτιμήτρια Kaplan-Meier χρησιμοποιείται για να εκτιμήσει την πιθανότητα επιβίωσης σε μια δεδομένη χρονική περίοδο μιας ομάδας ατόμων που έχουν αντιμετωπίσει ένα συγκεκριμένο γεγονός, όπως για

παράδειγμα ο θάνατος. Η μέθοδος αυτή επιτρέπει την σύγκριση της πιθανότητας επιβίωσης μεταξύ διαφορετικών ομάδων ατόμων σε σχέση με μια συγκεκριμένη μεταβλητή, όπως για παράδειγμα το φύλο, την ηλικία κ.ο.κ. Για παράδειγμα, αν θέλουμε να ελέγξουμε πως η ηλικία μπορεί να επηρεάσει την πιθανότητα επιβίωσης μπορούμε να διαχωρίσουμε τους ασθενείς σε διαφορετικές ηλικιακές ομάδες και να χρησιμοποιήσουμε την εκτιμήτρια Kaplan-Meier για να συγκρίνουμε τις πιθανότητες επιβίωσης μεταξύ των διαφορετικών ηλικιακών ομάδων.

Για την εκτίμηση της Kaplan-Meier κατασκευάζουμε μια σειρά χρονικών διαστημάτων στα οποία υπάρχει ένας χρόνος θανάτου ο οποίος συμβαίνει στην αρχή του διαστήματος.

Δηλαδή, έστω ότι έχουμε ένα τυχαίο δείγμα n ατόμων όπου κάποιες από αυτές τις παρατηρήσεις καταστρέφονται ή πεθαίνουν, δεδομένου ότι αναφερόμαστε σε ιατρικά δεδομένα, τις χρονικές στιγμές $t_1 < t_2 < \cdots < t_k$ με $k \leq n$.

Θεωρούμε ότι την χρονική στιγμή t_j πεθαίνουν d_j άτομα και ο αριθμός των ατόμων που είναι ζωντανά λίγο πριν την χρονική στιγμή t_j, συμπεριλαμβανομένων εκείνων οι οποίοι πρόκειται να πεθάνουν εκείνη την χρονική στιγμή ή αργότερα, συμβολίζεται με n_j, με $j = 1,2,\cdots,k$

Για τον υπολογισμό της συνάρτησης επιβίωσης εργαζόμαστε ως εξής.

Εφαρμόζοντας τον βασικό τύπο της θεωρίας Πιθανοτήτων,

$$P(A \cap B) = P(A)P(B|A)$$

η συνάρτηση επιβίωσης γράφεται:

$$S(t_j) = P(T > t_j) = P(T > t_1)P(T > t_2|T > t_1)\cdots P(T > t_j|T > t_{j-1})$$

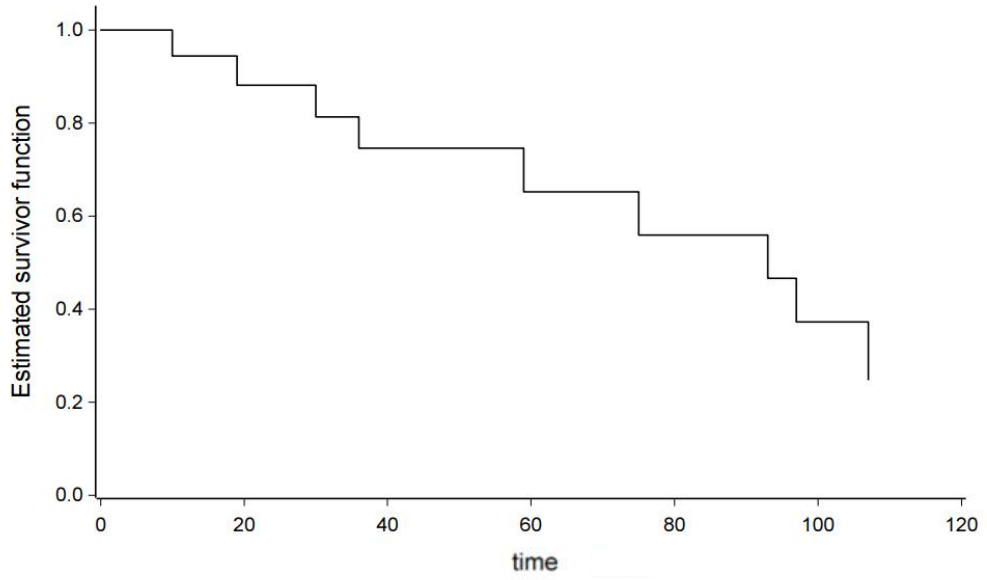
Τελικά, η εκτιμήτρια Kaplan-Meier της συνάρτησης επιβίωσης είναι:

$$\hat{S}(t) = \begin{cases} \prod_{j:t_j \leq t} \frac{n_j - d_j}{n_j}, & t \geq t_1 \\ 1, & t < t_1 \end{cases}$$

Το τυπικό σφάλμα της εκτίμησης αυτής δίνεται μέσω του τύπου του Greenwood *(1926)*:

$$se(\hat{S}(t)) = \hat{S}(t)\left\{\sum_{t_j \le t} \frac{d_j}{n_j(n_j - d_j)}\right\}^{1/2}$$

Η εκτιμήτρια Kaplan-Meier ακολουθεί προσεγγιστικά την κανονική κατανομή $N(\mu, \sigma^2)$ με $\mu = S(t)$ *και* $\sigma^2 = \left[se(\hat{S}(t))\right]^2$ και αποτελεί μια βαθμωτή συνάρτηση με γραφική παράσταση:



Σχήμα 2.1: *Γράφημα για την εκτίμηση της Kaplan-Meier.*

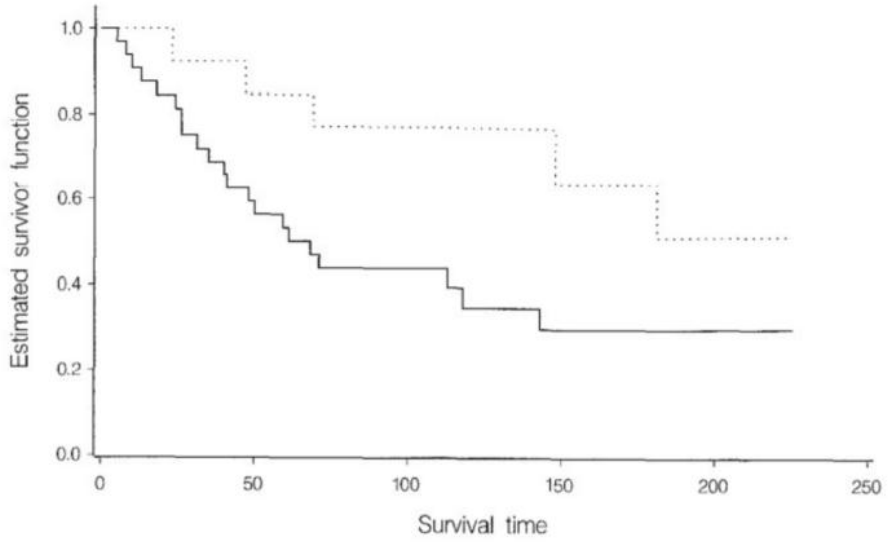
- ***<u>Σύγκριση Δύο Δειγμάτων</u>***

Πολλές φορές σε κάποια έρευνα χρειάζεται να χρησιμοποιήσουμε παραπάνω από ένα δείγμα και να συγκρίνουμε τα αποτελέσματα αυτών. Σε αυτή την περίπτωση, λοιπόν, ο απλούστερος τρόπος είναι μέσω της γραφικής απεικόνισης των εκτιμητριών.

<u>Παράδειγμα 1</u>:

Θα δανειστούμε το παράδειγμα του Collett *(2014)* το οποίο αφορά την πρόγνωση για γυναίκες με καρκίνο του μαστού. Τα δεδομένα των χρόνων επιβίωσης όπου έχουν συλλεχθεί είναι ομαδοποιημένα ανάλογα με το αν

τα τμήματα ενός όγκου είναι βαμμένα θετικά ή αρνητικά για το πεπτίδιο της HPA (Helix Pomatia Agglutinin) με τεχνικές ανοσοϊστοχημείας.



Σχήμα 2.2: *Εκτίμηση της Kaplan-Meier για γυναίκες με καρκίνο του μαστού.*

Στο *Σχήμα 2.2* παρουσιάζονται οι εκτιμήσεις των Kaplan-Meier για τις γυναίκες στις οποίες οι όγκοι είναι βαμμένοι θετικά, και συμβολίζονται με συνεχή γραμμή " ―― " και για εκείνες στις οποίες οι όγκοι είναι βαμμένοι αρνητικά, και συμβολίζονται με διακεκομμένη γραμμή " ….. ".

Παρατηρούμε ότι η εκτιμώμενη συνάρτηση επιβίωσης για τις γυναίκες για τις οποίες οι όγκοι βάφονται αρνητικά για το πεπτίδιο HPA είναι πάντα μεγαλύτερη από εκείνη για τις γυναίκες για τις οποίες οι όγκοι βάφονται θετικά.

Συμπερασματικά, ανά πάσα χρονική στιγμή *t*, η εκτιμώμενη πιθανότητα επιβίωσης, πέρα από το *t*, είναι μεγαλύτερη για τις γυναίκες με αρνητική χρώση ενώ εκείνες με θετική χρώση φαίνεται να έχουν χειρότερη πρόγνωση.

2.3 Εκτιμήτρια Nelson-Aalen

Από την συνάρτηση επιβίωσης μπορεί να εκτιμηθεί η σωρευτική συνάρτηση διακινδύνευσης, η οποία βασίζεται στους μεμονωμένους χρόνους συμβάντων, και μπορεί να πραγματοποιηθεί μέσω της εκτίμησης της Nelson-Aalen (*Nelson, 1972, Aalen, 1978*) και ορίζεται ως εξής.

$$\widehat{H}(t) = \begin{cases} \sum_{j:t_{(j)}} \frac{d_j}{n_j}, \text{όταν } t \geq t_{(1)} \\ 0\,, \qquad \text{όταν } t \leq t_{(1)} \end{cases}$$

Η εκτιμήτρια Nelson-Aalen χρησιμοποιείται συνήθως σε δεδομένα επιβίωσης για τον υπολογισμό του πραγματικού ρυθμού αποτυχίας στο χρόνο. Ο πραγματικός ρυθμός αποτυχίας δείχνει την πιθανότητα εμφάνισης μιας αποτυχίας (π.χ. θάνατος) σε μια μονάδα του χρόνου.

Γνωρίζουμε ότι η αθροιστική συνάρτηση κινδύνου και η συνάρτηση επιβίωσης συνδέονται μέσω της σχέσης:

$$H(t) = -\,lnS(t)$$

Χρησιμοποιώντας την εκτιμήτρια Kaplan-Meier έχουμε ότι για $t \geq t_1$, και θεωρώντας ότι το $\frac{d_j}{n_j}$ είναι πολύ μικρό έχουμε ότι:

$$\widehat{H}(t) = -\,ln\widehat{S}(t)$$

$$= -\sum_{j:t_j \leq t} ln\frac{(n_j - d_j)}{n_j}$$

$$= -\sum_{j:t_j \leq t} (1 - \frac{d_j}{n_j})$$

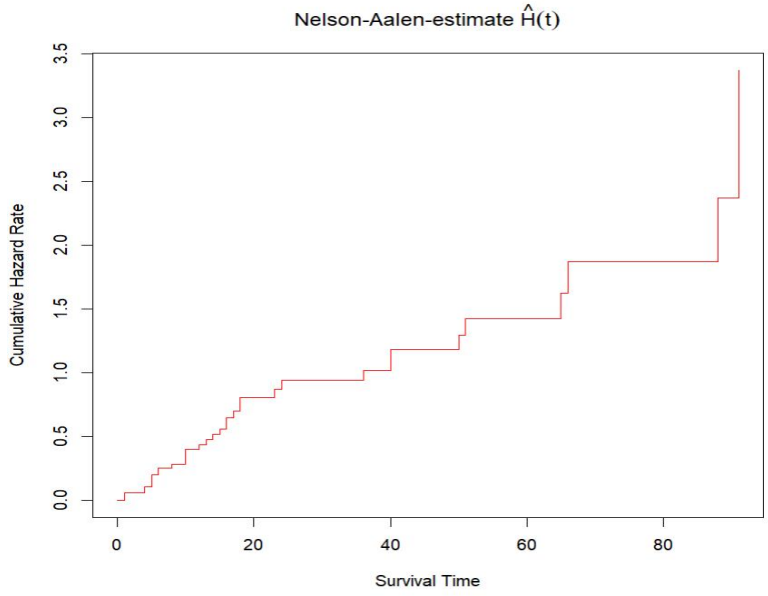
$$\approx \sum_{j:t_j \leq t} \left(\frac{d_j}{n_j}\right)$$

Ακόμη, η εκτιμήτρια της διασποράς της εκτιμήτριας Nelson-Aalen:

$$\widehat{V}(\widehat{H}) = \sum_{j:t_j \leq t} \left(\frac{d_j}{{n_j}^2}\right)$$

Η εκτιμήτρια Nelson-Aalen είναι, επίσης, βαθμιδωτή ή κλιμακωτή συνάρτηση. Στο *Σχήμα 2.3* παρουσιάζεται η γραφική παράσταση της εκτίμησης της Nelson Aalen.

Αξίζει να σημειωθεί ότι η εκτίμηση της Kaplan-Meier της συνάρτησης επιβίωσης μπορεί να θεωρηθεί ως μια προσέγγιση της εκτιμήτριας Nelson-Aalen. *(Collett,2014)*



Σχήμα 2.3: *Γραφική απεικόνιση της εκτίμησης της Nelson-Aalen της σωρευτικής συνάρτησης διακινδύνευσης.*

2.4 Μη-Παραμετρικοί Έλεγχοι

Για να συγκρίνουμε δύο ομάδες δεδομένων επιβίωσης, υπάρχει ένας αριθμός μεθόδων οι οποίοι μπορούν να χρησιμοποιηθούν για να ποσοτικοποιηθεί η έκταση των διαφορών μεταξύ των ομάδων. Θα αναφέρουμε δύο Μη-Παραμετρικές μεθόδους για την σύγκριση δύο ομάδων οι οποίες είναι ο έλεγχος Log-Rank και ο έλεγχος Wilcoxon.

2.4.1 Έλεγχος Log-Rank

Ο έλεγχος Log-Rank είναι ένας στατιστικός έλεγχος που χρησιμοποιείται για να ελέγξει εάν η κατανομή του χρόνου επιβίωσης μεταξύ δύο ή περισσότερων ομάδων είναι σημαντικά διαφορετική. Αυτός ο έλεγχος

είναι ευρέως χρησιμοποιούμενος στην ιατρική έρευνα, για παράδειγμα για να συγκρίνει τον χρόνο επιβίωσης μεταξύ δύο θεραπειών ή μεταξύ δύο ομάδων ασθενών.

Η βασική ιδέα του ελέγχου Log-Rank είναι να συγκρίνει την παρατηρούμενη κατανομή επιβίωσης με την αναμενόμενη κατανομή επιβίωσης, η οποία θα ήταν η ίδια αν οι δύο ομάδες είχαν την ίδια κατανομή επιβίωσης. Αυτό επιτυγχάνεται με τον υπολογισμό του log-rank στατιστικού τεστ, το οποίο βασίζεται στη σύγκριση της παρατηρούμενης κατανομής επιβίωσης με μια προσαρμοσμένη εκδοχή της κατανομής Kaplan-Meier για την περίπτωση που οι δύο ομάδες έχουν την ίδια κατανομή επιβίωσης.

Θεωρούμε τις χρονικές στιγμές θανάτου $t_1 < t_2 < \cdots < t_k$ στις δύο ομάδες δεδομένων με k οι διαφορετικοί χρόνοι θανάτου.

Την χρονική στιγμή t_j πεθαίνουν d_{1j} και d_{2j} άτομα από την Ομάδα I και Ομάδα II αντίστοιχα για j=1,2,…,k.

Υποθέτουμε , τώρα, ότι υπάρχουν n_{1j} και n_{2j} άτομα σε κίνδυνο αντίστοιχα για τις δύο ομάδες. Επομένως, την χρονική στιγμή t_j υπάρχουν $d_j = d_{1j} + d_{2j}$ θάνατοι από τα $n_j = n_{1j} + n_{2j}$ άτομα τα οποία κινδυνεύουν να πεθάνουν.

Λαμβάνοντας υπόψιν τα παραπάνω δεδομένα, δημιουργούμε τον Πίνακα Συνάφειας του *Πίνακα 2.1*.

Ομάδες	Αριθμός Θανάτων την χρονική στιγμή t_j	Αριθμός Επιζώντων την χρονική στιγμή t_j	Αριθμός ατόμων σε κίνδυνο πριν την χρονική στιγμή t_j
I	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
II	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Σύνολο	$d_j = d_{1j} + d_{2j}$	$n_j - d_j$	n_j

Πίνακας 2.1: *Πίνακας Συνάφειας για τον έλεγχο Log-Rank.*

Στην ανάλυση ενός πίνακα συνάφειας, χρησιμοποιώντας τον έλεγχο χ^2 θα υπολογίσουμε τις αναμενόμενες συχνότητες υπό την υπόθεση ότι οι πιθανότητες να διακοπούν οι λειτουργίες των μονάδων είναι ίδιες και για τις δύο ομάδες και κατά συνέπεια και οι συναρτήσεις επιβίωσης.

Μια τέτοια αναμενόμενη συχνότητα του πρώτου κελιού του *Πίνακα 2.1*, δηλ. τα άτομα από την *Ομάδα Ι* στα οποία συνέβη ο θάνατος, είναι η εξής:

$$E(d_{1j}) = n_{1j}d_j/n_j = \widehat{d}_{1j}$$

Με απόκλιση από την παρατηρούμενη d_{1j}:

$$u_j = d_{1j} - (n_{1j}d_j/n_j) \qquad (2.1)$$

Η διασπορά της παρατηρούμενης d_{1j} ορίζεται ως:

$$v_j = V(d_{1j}) = n_{1j}n_{2j}d_j(n_j - d_j)/n_j^2(n_j - 1) \qquad (2.2)$$

Τότε, ένας έλεγχος της υπόθεσης της ανεξαρτησίας πραγματοποιείται διαιρώντας το τετράγωνο της απόκλισης με την διασπορά της παρατηρούμενης d_{1j} που ορίσαμε στις σχέσεις *(2.1)* και *(2.2)*:

$$\frac{{u_j}^2}{v_j} = \frac{\{d_{1j} - (n_{1j}d_j/n_j)\}^2}{n_{1j}n_{2j}d_j(n_j - d_j)/n_j^2(n_j - 1)}$$

Η τελική μορφή της ελεγχοσυνάρτησης του ελέγχου Log-Rank προκύπτει αθροίζοντας ως προς όλες τις ανεξάρτητες χρονικές στιγμές t_j. Δηλαδή, u^2/v , όπου:

$$u = \sum_{\mathrm{j}} u_j = \sum_j \{d_{1j} - (n_{1j}d_j/n_j)\}$$

και $$\upsilon = \sum_{\mathrm{j}} v_j = \sum_j \{n_{1j}n_{2j}d_j(n_j - d_j)/n_j^2(n_j - 1)\}$$

Η ελεγχοσυνάρτηση u^2/v ακολουθεί την κατανομή χ_1^2 ασυμπτωτικά. (*Καρώνη, 2009*)

2.4.2 Έλεγχος Wilcoxon

Ο έλεγχος Wilcoxon, γνωστός και ως έλεγχος Breslow, είναι ένας μη παραμετρικός στατιστικός έλεγχος υπόθεσης που χρησιμοποιείται συνήθως στην ανάλυση επιβίωσης για τη σύγκριση των κατανομών επιβίωσης μεταξύ δύο ή περισσότερων ομάδων. Η βασική ιδέα του ελέγχου Wilcoxon είναι να υπολογίσει τη διαφορά των μέσων τιμών δύο ανεξάρτητων ομάδων και να ελέγξει εάν αυτή η διαφορά είναι σημαντική. Εάν η διαφορά είναι σημαντική, τότε οι δύο ομάδες έχουν στατιστικά σημαντικά διαφορετικές κατανομές επιβίωσης.

Ο έλεγχος Wilcoxon δίνεται από την σχέση:

$$W_w = \frac{U_w^2}{V_w}$$

όπου το στατιστικό U_w^2 ορίζεται ως:

$$U_w = \sum_{j=1}^{r} n_j(d_{1j} - e_{1j})$$

όπου το d_{1j} συμβολίζει τον αριθμό των θανάτων την δεδομένη στιγμή $t_{(j)}$ και το e_{1j} ορίζεται ως:

$$e_{1j} = n_{1j}d_j/n_j$$

Ο έλεγχος Wilcoxon είναι λιγότερο ευαίσθητος σε σχέση με τον έλεγχο Log-Rank στις αποκλίσεις του d_{1j} από το e_{1j} στα άκρα της κατανομής των χρόνων επιβίωσης.

Η διακύμανση του στατιστικού Wilcoxon όπου ακολουθεί την κατανομή χ^2 με ένα βαθμό ελευθερίας όταν η μηδενική υπόθεση είναι αποδεκτή, δίνεται από την σχέση:

$$W_w = U_w^2/V_w$$

Συμπερασματικά, ο έλεγχος Wilcoxon διεξάγεται με τον ίδιο τρόπο όπως και ο έλεγχος Log-Rank. Ωστόσο, στις περιπτώσεις όπου η εναλλακτική υπόθεση ότι ο κίνδυνος να υπάρξει ένα συμβάν (π.χ. θάνατος) σε

οποιαδήποτε χρονική στιγμή για ένα άτομο στην πρώτη ομάδα, είναι ανάλογος με τον κίνδυνο να υπάρξει το συμβάν για ένα άτομο στην δεύτερη ομάδα την ίδια χρονική στιγμή, είναι αποδεκτή, τότε ο καταλληλότερος έλεγχος είναι ο Log-Rank *(Collett, 2014)*. Αυτές οι περιπτώσεις σχετίζονται με τα μοντέλα αναλογικών κινδύνων όπου θα αναλύσουμε αργότερα σε αυτή την εργασία.

3. Παραμετρικά Μοντέλα Επιβίωσης

3.1 Βασικά Παραμετρικά Μοντέλα

Όπως είδαμε και στο προηγούμενο κεφάλαιο, με την χρήση Μη-Παραμετρικών μοντέλων και μεθόδων δεν χρειάζεται να υποθέσουμε μια συγκεκριμένη μορφή κατανομής πιθανοτήτων για τους χρόνους επιβίωσης. Ωστόσο, η χρήση του σωστού παραμετρικού μοντέλου μας οδηγεί σε ακριβέστερα συμπεράσματα.

Η ορθή επιλογή του παραμετρικού μοντέλου που ανταποκρίνεται στα δεδομένα μας γίνεται σε πρώτη φάση από την κατασκευή γραφικών παραστάσεων.

Με τον καθορισμό του μοντέλου κατανομής των χρόνων επιβίωσης και όντας γνωστή η συνάρτηση πυκνότητας πιθανότητας, μπορούμε να προσδιορίσουμε τις συναρτήσεις επιβίωσης και κινδύνου.

Έστω *f(x)* η σ.π.π της εκάστοτε κατανομής των χρόνων επιβίωσης.Τότε,η συνάρτηση επιβίωσης μπορεί να προσδιοριστεί ως εξής:

$$S(t) = 1 - \int_0^t f(u)du$$

Και η συνάρτηση διακινδύνευσης δίνεται από τον τύπο:

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

3.1.1 Εκθετικό Μοντέλο

Το απλούστερο μοντέλο διάρκειας ζωής για την συνάρτηση κινδύνου είναι να θεωρήσουμε ότι είναι σταθερή με την διάρκεια του χρόνου, δηλαδή, ότι ο κίνδυνος θανάτου για οποιαδήποτε χρονική στιγμή μετά την χρονική παρέλευση της έρευνας είναι ίδιος.

Η συνάρτηση πυκνότητας πιθανότητας της Εκθετικής Κατανομής ως γνωστόν είναι:

$$f(t) = \lambda e^{-\lambda t} \ , \ t > 0 \ και \ \lambda > 0$$

Επομένως, η συνάρτηση επιβίωσης γίνεται:

$$S(t) = e^{-\int_0^t h(u)du} = e^{-\int_0^t \lambda \, du} = e^{-\lambda t}$$

Και η συνάρτηση διακινδύνευσης προκύπτει ως εξής:

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

Το εκθετικό μοντέλο δεν συνηθίζεται να χρησιμοποιείται λόγω της ανεξαρτησίας της συνάρτησης κινδύνου από την ηλικία, δηλ. λόγω της σταθερής συνάρτησης κινδύνου.

3.1.2 <u>Μοντέλο Weibull</u>

Από τα πιο κοινά μοντέλα χρόνων επιβίωσης είναι το μοντέλο της κατανομής Weibull. Η κατανομή Weibull αποτελεί μια γενικότερη μορφή της Εκθετικής κατανομής όπως θα δούμε , καθώς , η σ.π.π ορίζεται ως:

$$f(t) = \eta \alpha^{-\eta} \, t^{\eta - 1} \, e^{-\left(\frac{t}{\alpha}\right)^{\eta}} \ , \ t > 0 \ , \ \alpha > 0 \ , \ \eta > 0$$

Η συνάρτηση επιβίωσης γίνεται:

$$S(t) = e^{-\left(\frac{t}{\alpha}\right)^{\eta}}$$

Και η συνάρτηση κινδύνου παίρνει την μορφή:

$$h(t) = \eta \alpha^{-\eta} t^{\eta}$$

Όπου ανάλογα με την τιμή του η η συνάρτηση κινδύνου θα είναι:

$$h(t) = \begin{cases} αύξουσα, \text{ όταν } \eta > 1 \\ σταθερή, \text{ όταν } \eta = 1 \\ φθίνουσα, \text{όταν } \eta < 1 \end{cases}$$

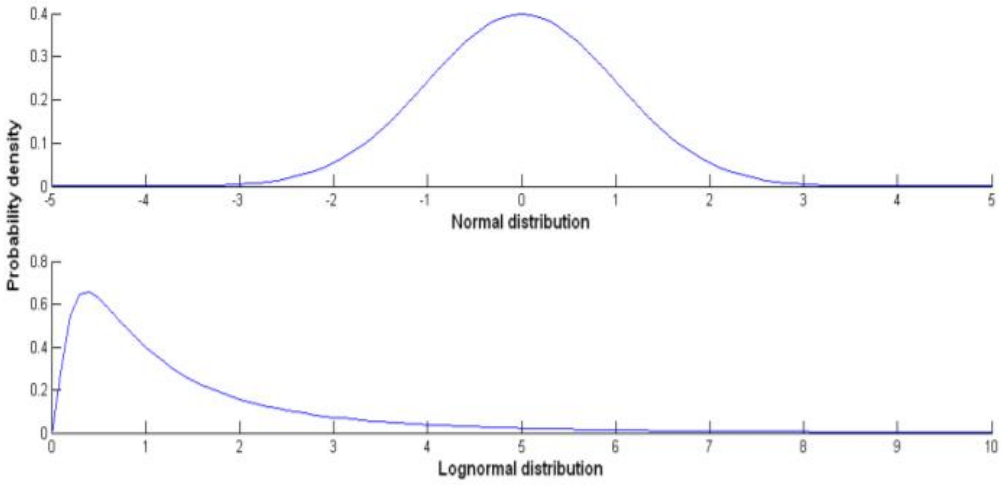
Για *η=1* η συνάρτηση κινδύνου παίρνει σταθερή τιμή και οι χρόνοι επιβίωσης ακολουθούν την Εκθετική κατανομή.

Η κατανομή Weibull έχει μεγάλη ευελιξία ως μοντέλο επιβίωσης και για αυτό χρησιμοποιείται αρκετά συχνά. *(Καρώνη, 2009)*

3.1.3 Λογαριθμο-Κανονικό Μοντέλο

Στην ανάλυση επιβίωσης και στα δεδομένα διάρκειας ζωής είναι αρκετά σπάνιο να χρησιμοποιήσουμε την Κανονική κατανομή ως μοντέλο επιβίωσης. Αυτό συμβαίνει διότι, ως γνωστόν, η Κανονική Κατανομή είναι συμμετρική ενώ στα δεδομένα επιβίωσης είναι λογικό η κατανομή των χρόνων επιβίωσης να παρουσιάζει λοξότητα, καθώς αναφερόμαστε σε διάρκεια επιβίωσης.

Για τον λόγο αυτό η χρήση της Λογαριθμο-Κανονικής κατανομής φαίνεται να είναι καλύτερη λύση, καθώς παρουσιάζει δεξιά λοξότητα, όπως μπορούμε να δούμε και στο *Σχήμα 3.1*.



Σχήμα 3.1: *Διαγράμματα των συναρτήσεων πιθανότητας της Κανονικής και της Λογαριθμο-Κανονικής κατανομής.*

Έστω *T* τ.μ η οποία ακολουθεί την Λογαριθμο-Κανονική κατανομή.
Δηλ. *T~Log-Normal*(μ, σ^2)

Τότε αν ορίσουμε την τ.μ *Y=lnT* η θα ακολουθεί την Κανονική κατανομή.
Δηλ. *Y~N*(μ, σ^2) με $\mu > 0$ η μέση τιμή και σ^2 η διασπορά, όπου θα έχουμε $\sigma > 0$

Θα έχουμε, τότε, την σ.π.π της Λογαριθμο-Κανονικής κατανομής η οποία ορίζεται ως:

$$f(t) = \frac{1}{\sigma_t\sqrt{2\pi}} exp \frac{-(lnt-\mu)^2}{2\sigma^2} \text{ , με } t > 0 \qquad (3.1)$$

Όπου, *exp(μ)* είναι η παράμετρος κλίμακας και $\sigma > 0$ η παράμετρος σχήματος.

Η σωρευτική συνάρτηση κατανομής δίνεται από την σχέση:

$$F(t) = \Phi\left[\frac{lnt-\mu}{\sigma}\right], t \in (0, \infty)$$

Όπου Φ είναι η συνάρτηση της αθροιστικής τυποποιημένης κανονικής κατανομής.

Τότε, η συνάρτηση επιβίωσης η οποία προκύπτει από την σχέσεις *(1.1),* που εισάγαμε στο Κεφάλαιο 1 και *(3.1)* είναι η εξής:

$$S(t) = 1 - \Phi\left[\frac{lnt-\mu}{\sigma}\right]$$

και τέλος η συνάρτηση διακινδύνευσης για την οποία προκύπτει η σχέση:

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

$$\Leftrightarrow \quad h(t) = \frac{\frac{1}{\sigma_t\sqrt{2\pi}} exp \frac{-(lnt-\mu)^2}{2\sigma^2}}{1-\Phi\left[\frac{lnt-\mu}{\sigma}\right]}$$

(Kurniasari et al., 2019)

Ωστόσο, η Λογαριθμο-Κανονική κατανομή παρουσιάζει και κάποια μειονεκτήματα, όπως ότι η συνάρτηση διακινδύνευσης $h(t)$ ενώ αρχικά και μέχρι κάποια χρονική στιγμή αυξάνεται, έπειτα τείνει να μηδενιστεί καθώς ο χρόνος $t \longrightarrow \infty$. Παρά το γεγονός αυτό, η Λογαριθμο-Κανονική κατανομή έχει καλή προσαρμογή σε περιπτώσεις που έχουμε σχετικά χαμηλές τιμές του t. *(Καρώνη, 2009)*

3.1.4 <u>Λογαριθμο-Λογιστικό Μοντέλο</u>

Στην ανάλυση δεδομένων επιβίωσης το Λογαριθμο-Λογιστικό μοντέλο είναι αρκετά χρήσιμο στις περιπτώσεις όπου ο βαθμός της θνησιμότητας φτάνει σε πολύ υψηλό επίπεδο, αλλά μετά από κάποιο χρονικό διάστημα φτάνει σε πολύ χαμηλά επίπεδα.

Η Λογαριθμο-Λογιστική κατανομή μοιάζει αρκετά, όσον αφορά το σχήμα της κατανομής, με την Λογαριθμο-Κανονική, ωστόσο, είναι πολύ πιο κατάλληλη να εφαρμοστεί σε δεδομένα επιβίωσης.

Ομοίως με την Λογαριθμο-Κανονική κατανομή, έστω T τ.μ η οποία ακολουθεί την Λογαριθμο-Λογιστική κατανομή.

Τότε η $Y=lnT$ η θα ακολουθεί την Λογιστική κατανομή.

Η σ.π.π της Λογαριθμο-Λογιστικής κατανομής δίνεται από τον τύπο:

$$f(t) = \frac{\alpha\gamma t^{\gamma-1}}{(1+\alpha t^{\gamma})^2}, \gamma\iota\alpha\ t > 0\ \kappa\alpha\iota\ \alpha > 0, \gamma > 0 \qquad (3.2)$$

Η συνάρτηση επιβίωσης προκύπτει από τις σχέσεις *(1.1)* και *(3.2):*

$$S(t) = (1+\alpha t^{\gamma})^{-1}$$

και η συνάρτηση διακινδύνευσης η οποία είναι:

$$h(t) = \frac{f(t)}{S(t)}$$
$$\Leftrightarrow \quad h(t) = \frac{\alpha\gamma t^{\gamma-1}}{(1+\alpha t^{\gamma})}$$

3.1.5 <u>Μοντέλο Gompertz</u>

Για να μοντελοποιήσουμε την ανθρώπινη ζωή, δηλ. να επιλέξουμε το κατάλληλο μοντέλο το οποίο θα ανταποκρίνεται στα χαρακτηριστικά της διάρκειας της ζωής ενός ανθρώπου, σίγουρα θέλουμε η συνάρτηση επιβίωσης να μην είναι αύξουσα ή σταθερή.

Η κατανομή η οποία είναι καταλληλότερη για την μοντελοποίηση της διάρκειας ζωής ενός ανθρώπου, κυρίως μεγαλύτερων ηλικιών, είναι η Gompertz.

Η σ.π.π για την κατανομή Gompertz δίνεται από τον τύπο:

$$f(t) = exp\left[(\lambda + \gamma t) - \frac{1}{\gamma}(e^{\lambda+\gamma t} - e^{\lambda})\right] \qquad (3.3)$$

Η συνάρτηση αξιοπιστίας που προκύπτει από τις σχέσεις *(1.1)* και *(3.3)* είναι:

$$S(t) = exp\left[-\frac{e^{\lambda}}{\gamma}(e^{\gamma t} - 1)\right], \mu\varepsilon\ t > 0, \gamma \neq 0$$

Και η συνάρτηση διακινδύνευσης που ορίζεται ως:

$$h(t) = exp(\lambda + \gamma t)$$

η οποία είναι φθίνουσα όταν $\gamma < 0$, αύξουσα όταν $\gamma > 0$ και σταθερή όταν $\gamma = 0$, όπου και στην τελευταία περίπτωση η κατανομή τελικά είναι η εκθετική. *(Καρώνη, 2009)*

3.2 <u>Παραμετρικά Μοντέλα Επιβίωσης στην Αναλογιστική Επιστήμη</u>

Η Αναλογιστική είναι ένας κλάδος της ασφαλιστικής ο οποίος ειδικεύεται στην χρήση μαθηματικών και στατιστικών μεθόδων, εκτός από χρηματοοικονομικά εργαλεία και γνώσεις, με σκοπό την εκτίμηση των κινδύνων στον τομέα της ασφάλισης. Το πόσο θα διατηρήσει ο κάθε ασφαλιζόμενος το συμβόλαιό του, πώς θα τροποποιήσει, αν θα το διακόψει ή αν το ανανεώσει εξαρτάται από πολλά κριτήρια, όπως την οικονομική της κατάσταση, το επαγγελματικό του προφίλ, τις ατομικές του ανάγκες, τα φυσικά του χαρακτηριστικά, την συγκεκριμένη διάρκεια ζωής του κ.α. Αυτό που ενδιαφέρει την κάθε ασφαλιστική εταιρεία, είναι η δημιουργία ενός αξιόπιστου μοντέλου που να λαμβάνει υπόψη όλους τους αναγκαίους παράγοντες και να προβλέπει τη διάρκεια ασφαλιστικού συμβολαίου, ανάλογα με το άτομο στο που αναφέρεται.

Είναι λοιπόν φανερό ότι για να καταλήξουμε σε ένα τελικό μοντέλο που να δίνει αξιόπιστα αποτελέσματα, πρέπει να μελετηθεί ένα μεγάλο φάσμα παραμέτρων και για το λόγο αυτό, τα μοντέλα επιβίωσης είναι αναπόσπαστο κομμάτι της.

Ένας σημαντικός παράγοντας για τους αναλογιστές είναι η δυνατότητα να χρησιμοποιούν αριστερά αποκομμένα δεδομένα, καθώς είναι πιο σύνηθες οι κάτοχοι ασφαλιστήριων συμβολαίων να ξεκίνησαν την ασφάλιση τους τουλάχιστον μετά την ενηλικίωση ή και ακόμα αργότερα. Τα περισσότερα μοντέλα που έχουν στην διάθεσή τους οι αναλογιστές δεν καλύπτουν αυτή την δυνατότητα και για αυτό τείνουν να χρησιμοποιούν μοντέλα επιβίωσης με την βοήθεια της μέγιστης πιθανοφάνειας, καθιστώντας απλή και γρήγορη την εργασία αυτή.

Ο Richards *(2012)* ανέλυσε στο άρθρο του *"A handbook of parametric survival models for actuarial use"* πώς οι παραδοσιακές αναλογιστικές τεχνικές για την ανάλυση θνησιμότητας αντικαθίστανται από στατιστικά μοντέλα, και συγκεκριμένα τα μοντέλα επιβίωσης που μοντελοποιούν τη θνησιμότητα σε ατομικό επίπεδο. Ισχυρίζεται ότι, στην ουσία, τα μοντέλα που χρησιμοποιούνται για την ανάλυση των ασφαλιστικών δεδομένων μοιάζουν πολύ με τα μοντέλα επιβίωσης τα οποία χρησιμοποιούνται στην ιατρική έρευνα, αλλά με κάποιες διαφορές. Το κυριότερο μεταξύ αυτών είναι η ύπαρξη μεγάλου όγκου δεδομένων.

Παράλληλα, στην αναλογιστική επιστήμη, δεδομένου ότι η διαχείριση των κεφαλαίων εξαρτάται από το χρόνο, δίνεται μεγάλη σημασία στη μορφή της συνάρτησης κινδύνου. Επεσήμανε, επίσης, την ανάγκη

προεπεξεργασίας των στοιχείων των ασφαλιστικών εταιρειών, καθώς είναι πολύ πιθανό να υπάρξουν διπλότυπα δεδομένα και αυτό μπορεί να επηρεάσει την έρευνα.

Όπως αναφέρθηκε και παραπάνω, το κύριο θέμα είναι η θνησιμότητα και ο τρόπος με τον οποίο οι αναλογιστικές στατιστικές μελετούν τη θνησιμότητα μέσω ποικίλων τεχνικών και μεθόδων που διαφέρουν από τις στατιστικές μεθόδους σε άλλους κλάδους. Στον *Πίνακα 3.2*, αναφέρονται 6 "νόμοι θνησιμότητας", όπως αναλύθηκαν στο παράρτημα.

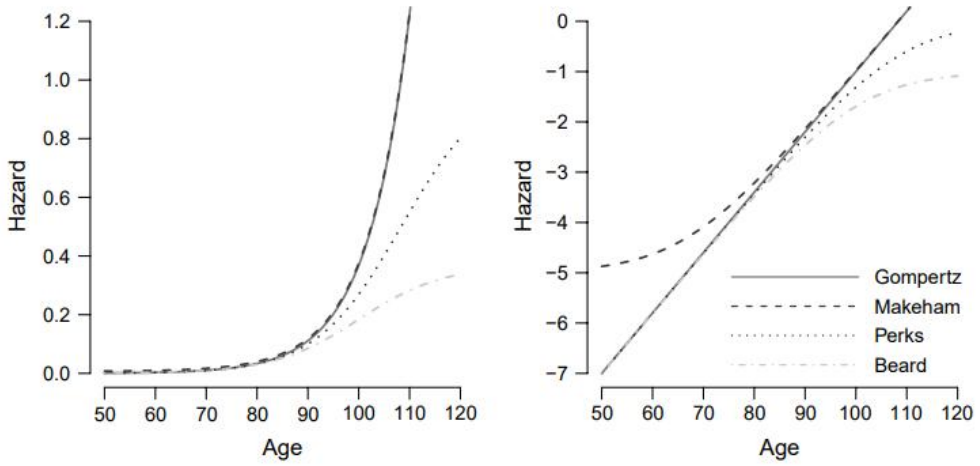
Mortality law	μ_x	$H_x(t)$
Gompertz (1825)	$e^{\alpha+\beta x}$	$\frac{(e^{\beta t}-1)}{\beta}e^{\alpha+\beta x}$
Makeham (1859)	$e^{\varepsilon}+e^{\alpha+\beta x}$	$te^{\varepsilon}+\frac{(e^{\beta t}-1)}{\beta}e^{\alpha+\beta x}$
Perks (1932)	$\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$	$\frac{1}{\beta}\log\left(\frac{1+e^{\alpha+\beta(x+t)}}{1+e^{\alpha+\beta x}}\right)$
Beard (1959)	$\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\rho+\beta x}}$	$\frac{e^{-\rho}}{\beta}\log\left(\frac{1+e^{\alpha+\rho+\beta(x+t)}}{1+e^{\alpha+\rho+\beta x}}\right)$
Makeham–Perks (1932)	$\frac{e^{\varepsilon}+e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$	$te^{\varepsilon}+\frac{(1-e^{\varepsilon})}{\beta}\log\left(\frac{1+e^{\alpha+\beta(x+t)}}{1+e^{\alpha+\beta x}}\right)$
Makeham–Beard (1932)	$\frac{e^{\varepsilon}+e^{\alpha+\beta x}}{1+e^{\alpha+\rho+\beta x}}$	$te^{\varepsilon}+\frac{(e^{-\rho}-e^{\varepsilon})}{\beta}\log\left(\frac{1+e^{\alpha+\rho+\beta(x+t)}}{1+e^{\alpha+\rho+\beta x}}\right)$

Πίνακας 3.2: *Οι "Νόμοι Θνησιμότητας" και οι αντίστοιχες συναρτήσεις διακινδύνευσης.*

Οι "Νόμοι Θνησιμότητας" που είδαμε σχετίζονται με Παραμετρικά Μοντέλα Επιβίωσης. Παρατηρώντας τα γραφήματα στο *Σχήμα 3.2* ο απλούστερος φαίνεται να είναι ο "νόμος" του Gompertz (*1825*), όπου η συνάρτηση κινδύνου για τη θνησιμότητα αυξάνεται εκθετικά σε άτομα ηλικίας άνω των 60 ετών. Στη συνέχεια ο νόμος του Makeham (*1859*), όπου προκύπτει προσθέτοντας έναν σταθερό όρο ανεξάρτητο από την ηλικία στη συνάρτηση του "νόμου" Gompertz. Ο "νόμος" του Perks (*1932*), όπου πρότεινε ένα μοντέλο παρόμοιο με το Gompertz, αλλά με πιο ομαλή κλίση εκεί όπου το Gompertz αναπτύσσεται εκθετικά και τέλος ο Beard (*1959*) ο οποίος πρόσθεσε μια παράμετρο ρ στο μοντέλο, η οποία πρόσθεσε μια μεγαλύτερη αλλαγή στο ρυθμό αλλαγής.

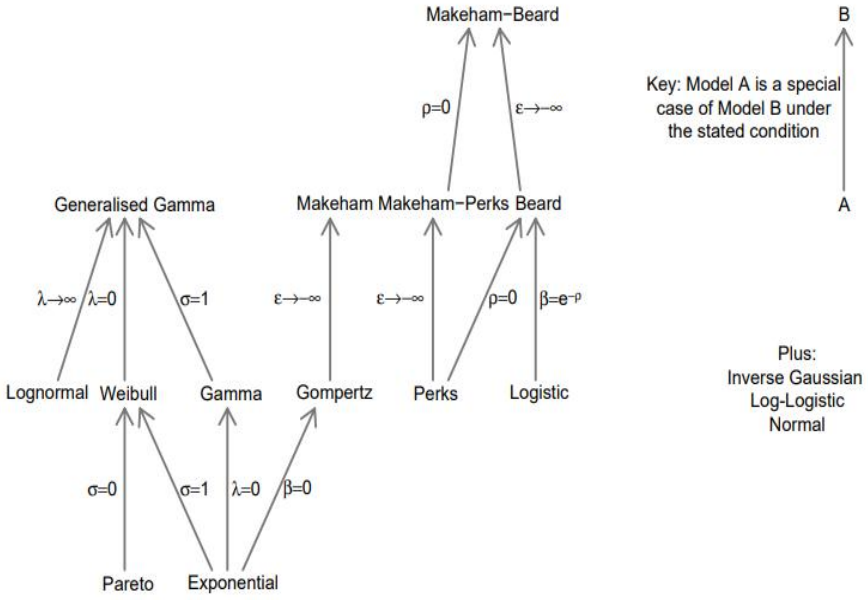
Όταν αυτοί οι "νόμοι" εφαρμόζονται σε πραγματικά δεδομένα θνησιμότητας, συνήθως δεν εφαρμόζονται πολύ καλά εκτός του ηλιακού εύρους των 60 με 90 ετών. Από την ηλικία των 90 ετών και πάνω παρατηρείται μια επιβράδυνση στον ρυθμό αύξησης, και γενικά σχετίζεται με την μείωση θνησιμότητας στα τέλη της ζωής (Gavrilov & Gavrilova (*2001*)) το οποίο έρχεται σε αντίθεση με τους "νόμους"

Gompertz και Makeham. Παρόμοια, και για τους συνταξιούχους κάτω των 60 ετών όπου το ποσοστό θνησιμότητας δεν μειώνεται εκθετικά με την μείωση της ηλικίας και ακυρώνοντας με αυτόν τον τρόπο τους "νόμους" Gompertz, Perks και Beard. Συνεπώς, όταν παραδείγματος χάριν, η έρευνα αφορά άτομα με ηλικιακό εύρος μεταξύ 50-110 χρειαζόμαστε έναν συνδυασμό ιδιοτήτων και για αυτόν τον λόγο οι "νόμοι" που προσαρμόζουν καλύτερα τέτοιου είδους δεδομένα είναι οι Makeham-Perks και Makeham-Beard.



Σχήμα 3.2: *Συναρτήσεις Διακινδύνευσης για τους "νόμους θνησιμότητας" των Gompertz, Makeham, Perks and Beard με παραμέτρους α = -13, β = 0.12, ρ = 1 και ε = -5 σε κανονική (αριστερά) και λογαριθμική κλίμακα (δεξιά).*

Στο *Σχήμα 3.3* μπορούμε να δούμε τις σχέσεις που έχουν μεταξύ τους οι "νόμοι" θνησιμότητας με τα μοντέλα επιβίωσης, ανάλογα με κάποιες αλλαγές στις παραμέτρους τους. Συνδυάζοντας το Σχήμα 3.3 με τους *Πίνακες 3.2* και *3.3* και ξεκινώντας από τα μοντέλα που βρίσκονται στο άνω μέρος του διαγράμματος, βλέπουμε ότι αυτά είναι ειδικές περιπτώσεις των μοντέλων που βρίσκονται στο κατώτερο μέρος. Συγκεκριμένα, το Weibull μοντέλο είναι ειδική περίπτωση του Εκθετικού όταν η παράμετρος $\sigma = 1$ και της Pareto όταν το $\sigma = 0$. Αντίστοιχα, το μοντέλο της Γενικευμένης Κατανομής Γάμμα είναι ειδική περίπτωση του μοντέλου Weibull όταν η παράμετρος $\lambda = 1$. Επίσης, όσον αφορά τον "νόμο" θνησιμότητας Makeham, ο οποίος είναι ειδική περίπτωση του μοντέλου Gompertz όταν η παράμετρος $\varepsilon = -\infty$ και η Gompertz με τη σειρά της είναι ειδική περίπτωση του εκθετικού μοντέλου για $\beta = 0$. Αντίστοιχη λογική ισχύει και στα υπόλοιπα μοντέλα του διαγράμματος.



Σχήμα 3.3: *Σχέσεις μεταξύ των μοντέλων επιβίωσης και των "νόμων" θνησιμότητας.*

Distribution	μ_x	$H_x(t)$
Exponential	e^{α}	te^{α}
Extreme value	See Gompertz hazard in Table 1	
Pareto	$\frac{e^{\alpha}}{x}$	$e^{\alpha}\log\left(\frac{x+t}{x}\right)$
Weibull	$e^{\alpha}x^{\sigma-1}$	$\begin{cases} e^{\alpha}\log\left(\frac{x+t}{x}\right), & \sigma=0; \\ \frac{e^{\alpha}}{\sigma}[(x+t)^{\sigma}-x^{\sigma}], & \text{otherwise.} \end{cases}$
Logistic	$\frac{1}{e^{\alpha}\left(1+\exp\left(-\frac{x+\alpha}{e^{\sigma}}\right)\right)}$	$\log\left(\frac{1+\exp\left(\frac{x+t+\alpha}{e^{\sigma}}\right)}{1+\exp\left(\frac{x+\alpha}{e^{\sigma}}\right)}\right)$
Log-Logistic	$\frac{e^{\alpha+\sigma}x^{e^{\sigma}-1}}{1+e^{\alpha}x^{e^{\sigma}}}$	$\log\left(\frac{1+e^{\alpha}(x+t)^{e^{\sigma}}}{1+e^{\alpha}x^{e^{\sigma}}}\right)$
Normal	$\frac{\frac{1}{e^{\sigma}\sqrt{2\pi}}\exp\left(-\frac{(x+\alpha)^2}{2e^{2\sigma}}\right)}{1-\Phi\left(\frac{x+\alpha}{e^{\sigma}}\right)}$	$\log\left(\frac{1-\Phi\left(\frac{x+\alpha}{e^{\sigma}}\right)}{1-\Phi\left(\frac{x+t+\alpha}{e^{\sigma}}\right)}\right)$
Lognormal	$\frac{\frac{1}{xe^{\sigma}\sqrt{2\pi}}\exp\left(-\frac{(\log x+\alpha)^2}{2e^{2\sigma}}\right)}{1-\Phi\left(\frac{\log x+\alpha}{e^{\sigma}}\right)}$	$\log\left(\frac{1-\Phi\left(\frac{\log x+\alpha}{e^{\sigma}}\right)}{1-\Phi\left(\frac{\log(x+t)+\alpha}{e^{\sigma}}\right)}\right)$
Inverse Gaussian	$\frac{\left(\frac{e^{\sigma}}{2\pi x^3}\right)^{\frac{1}{2}}\exp\left(-\frac{e^{\sigma}(x-e^{-\alpha})^2}{2e^{-2\alpha}x}\right)}{\mathrm{IGS}(x)}$	$\log\left(\frac{\mathrm{IGS}(x)}{\mathrm{IGS}(x+t)}\right)$
Gamma	$\frac{e^{\alpha}x^{e^{\lambda}-1}\exp(-xe^{\alpha e^{-\lambda}})}{\Gamma(e^{\lambda})-\gamma(e^{\lambda},xe^{\alpha e^{-\lambda}})}$	$\log\left(\frac{\Gamma(e^{\lambda})-\gamma(e^{\lambda},xe^{\alpha e^{-\lambda}})}{\Gamma(e^{\lambda})-\gamma(e^{\lambda},(x+t)e^{\alpha e^{-\lambda}})}\right)$
Generalised Gamma	$\frac{e^{\alpha}x^{e^{\lambda}\sigma-1}\exp\left(-x^{\sigma}\left(\frac{e^{\alpha}}{\sigma}\right)^{e^{-\lambda}}\right)}{\Gamma(e^{\lambda})-\gamma\left(e^{\lambda},x^{\sigma}\left(\frac{e^{\alpha}}{\sigma}\right)^{e^{-\lambda}}\right)}$	$\log\left(\frac{\Gamma(e^{\lambda})-\gamma\left(e^{\lambda},x^{\sigma}\left(\frac{e^{\alpha}}{\sigma}\right)^{e^{-\lambda}}\right)}{\Gamma(e^{\lambda})-\gamma\left(e^{\lambda},(x+t)^{\sigma}\left(\frac{e^{\alpha}}{\sigma}\right)^{e^{-\lambda}}\right)}\right)$

Πίνακας 3.3: *Μοντέλα επιβίωσης και οι αντίστοιχες συναρτήσεις διακινδύνευσης και σωρευτικές συναρτήσεις διακινδύνευσης.*

4. Μοντέλα Παλινδρόμησης για Δεδομένα Διάρκειας Ζωής

4.1 Συμμεταβλητές

Στις περισσότερες ιατρικές έρευνες για τους χρόνους διάρκειας ζωής ασθενών, σημαντικό ρόλο για την εξαγωγή συμπερασμάτων παίζουν και άλλοι παράγοντες όπως:

- Δημογραφικές μεταβλητές (π.χ. Φύλο,Ηλικία κτλ)
- Συνήθειες (π.χ. ιστορικό καπνίσματος,άσκηση-γυμναστική κτλ)
- Μεταβλητές που σχετίζονται με δείκτες της υγείας (π.χ. επίπεδο χοληστερόλης, ιστορικό καρκίνου κτλ)

Αυτοί οι παράγοντες οι οποίοι θα συμπεριληφθούν στην μελέτη ονομάζονται συμμεταβλητές (covariates). Σε αυτό το κεφάλαιο θα εξετάσουμε την επιρροή των συμμεταβλητών στην εξέλιξη του χρόνου επιβίωσης.

Συμπεριλαμβάνοντας αυτές τις επιπλέον παραμέτρους, το σύνολο των δεδομένων που θα προκύψει θα γίνει αρκετά πολύπλοκο έτσι ώστε να χρησιμοποιήσουμε τις Μη-Παραμετρικές μεθόδους που αναφέραμε σε προηγούμενο κεφάλαιο.

Τα μοντέλα που προκύπτουν με την προσθήκη παραπάνω από μίας μεταβλητής είναι κάπως διαφορετικά από εκείνα που χρησιμοποιούνται στην ανάλυση παλινδρόμησης, ωστόσο, πολλές από τις διαδικασίες είναι κοινές και για τα μοντέλα επιβίωσης.

Το βασικό μοντέλο για τα δεδομένα επιβίωσης το οποίο θα εξετάσουμε, είναι το μοντέλο αναλογικής διακινδύνευσης (PH). Το μοντέλο αναλογικής διακινδύνευσης είναι ένα συμπλήρωμα στο σύνολο των εργαλείων στην ανάλυση επιβίωσης και παρέχει ορισμένα ιδιαίτερα πλεονεκτήματα.

4.2 Μοντέλα Αναλογικής Διακινδύνευσης

Γενικότερα, από την προσθήκη των συμμεταβλητών στο μοντέλο παρατηρείται μια επιβράδυνση ή επιτάχυνση της κλίμακας του χρόνου.

Σε αυτό το υποκεφάλαιο θα ασχοληθούμε με τα μοντέλα Αναλογικής Διακινδύνευσης, τα οποία χωρίζονται σε δύο κατηγορίες μοντέλων και βασίζονται στην συνάρτηση διακινδύνευσης $h_0(t)$

i. **Παραμετρικά Μοντέλα** στα οποία η συνάρτηση διακινδύνευσης $h_0(t)$ ορίζεται μέσω μιας γνωστής κατανομής. (π.χ. Weibull)

ii. **Ημι-παραμετρικά Μοντέλα** στα οποία η συνάρτηση διακινδύνευσης $h_0(t)$ δεν είναι γνωστή (π.χ. μοντέλο αναλογικής διακινδύνευσης του Cox που θα αναλύσουμε σε επόμενο κεφάλαιο)

Ορισμός Μοντέλου Αναλογικής Διακινδύνευσης

Το μοντέλο αναλογικής διακινδύνευσης έχει την μορφή:

$$h(t|\lambda) = \lambda \cdot h_0(t)$$

Όπου η $h_0(t)$ είναι μια βασική συνάρτηση διακινδύνευσης και η τυχαία θετική ποσότητα $\lambda > 0$ ονομάζεται ευπάθεια (frailty) της κάθε μονάδας.

Το μοντέλο αναλογικής διακινδύνευσης έχει την ιδιότητα ο λόγος των συναρτήσεων διακινδύνευσης δύο μονάδων να παραμένει σταθερός και ανεξάρτητος του χρόνου.

$$\frac{h(t|\lambda_1)}{h(t|\lambda_2)} = \frac{\lambda_1}{\lambda_2}$$

Από την συνάρτηση διακινδύνευσης προκύπτει η συνάρτηση επιβίωσης:

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

όπου $H_0(t)$ είναι η σωρευτική βασική συνάρτηση διακινδύνευσης.

4.2.1 <u>Γραφικός Έλεγχος Καταλληλότητας στον Μοντέλο Αναλογικής Διακινδύνευσης</u>

Προκειμένου να προσαρμόσουμε ένα μοντέλο αναλογικής διακινδύνευσης θα πρέπει να ευσταθεί η υπόθεση της αναλογικότητας. Για να ελέγξουμε αυτή την υπόθεση αρκεί να ακολουθήσουμε τα παρακάτω βήματα.

Ως γνωστόν, οι συναρτήσεις διακινδύνευσης και επιβίωσης για το μοντέλο αναλογικής διακινδύνευσης είναι οι εξής:

$$h(t;x) = h_0(t)g(x)$$

και $$S(t;x) = exp\{-H_0(t)e^{\beta' x}\} \qquad (4.1)$$

όπου $H_0(t)$ είναι η σωρευτική συνάρτηση διακινδύνευσης και $g(x) = e^{\beta' x}$.

Αν λογαριθμήσουμε δύο φορές την σχέση *(4.1)*, τότε θα προκύψει η σχέση:

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

Τότε η καμπύλη $ln\{-lnS(t;x)\}$ θα είναι παράλληλη ως προς τον χρόνο t με την $lnH_0(t)$, για οποιοδήποτε x και σε αυτό συνεπάγεται ότι και οι καμπύλες $ln\{-lnS(t;x)\}$ θα είναι παράλληλες μεταξύ τους ως προς τον χρόνο t για οποιαδήποτε *x*.

Αν, επομένως, οι γραφικές παραστάσεις αυτών των καμπυλών φαίνονται να είναι παράλληλες μεταξύ τους, τότε, η υπόθεση της αναλογικής διακινδύνευσης επιβεβαιώνεται.

Για να πραγματοποιήσουμε αυτόν τον έλεγχο είναι απαραίτητο να εκτιμήσουμε τις $\hat{S}(t;x)$, συνήθως με την χρήση της Kaplan-Meier εκτιμήτριας για τις τιμές των *x*.

Ένας εναλλακτικός τρόπος αν υποψιαζόμαστε ότι κατανομή των δεδομένων διάρκειας ζωής είναι κάποια συγκεκριμένη, είναι η δημιουργία γραφικών παραστάσεων συναρτήσει του χρόνου t. Οι

γραφικές παραστάσεις αυτές θα παρουσιάζουν ευθείες γραμμές και επομένως, η προσαρμογή του μοντέλου θα ευσταθεί.

Ο *Πίνακας 4.1* μπορεί να βοηθήσει στην δημιουργία αυτών των διαγραμμάτων:

Κατανομή	**Γραφική Παράσταση**
Εκθετική	$-lnS(t;x) \quad με \quad t$
Weibull	$ln\{-lnS(t;x)\} \quad με \quad lnt$
Gumbel	$ln\{-lnS(t;x)\} \quad με \quad t$
Λογαριθμο-Κανονική	$\Phi^{-1}(1-S(t;x)) \quad με \quad lnt$
Λογαριθμο-Λογιστική	$ln[(1-S(t;x))/S(t;x)] \quad με \quad lnt$
όπου $\Phi(z)=P[Z \leq z]$ όταν $Z \sim N(0,1)$	

Πίνακας 4.1: *Συνιστώσες των γραφικών παραστάσεων της συνάρτησης επιβίωσης συναρτήσει του χρόνου για τις διάφορες κατανομές.*

Στις περιπτώσεις που μια μεταβλητή έχει πολλά επίπεδα ή έχουμε πολλές μεταβλητές, τότε ο γραφικός έλεγχος της υπόθεσης της αναλογικότητας δεν είναι εφικτός.

4.3 <u>Μοντέλο Επιταχυνόμενης Διακοπής</u>

Το Μοντέλο Επιταχυνόμενης Διακοπής (Accelerated Failure Time Model) ή πιο συχνά γνωστό και ως AFT, χρησιμοποιείται σε ένα σύνολο πειραμάτων στα οποία η αλλαγή της τιμής μιας συμμεταβλητής είναι δυνατόν να επισπεύσει τον θάνατο ή διακοπή λειτουργίας ενός ατόμου/μονάδας. Όταν θέλουμε να συνδέσουμε μία εξαρτημένη μεταβλητή Y με τις συμμεταβλητές X_i, το πιο σύνηθες μοντέλο που χρησιμοποιούμε είναι η γραμμική παλινδρόμηση.

Έτσι θεωρούμε το μοντέλο :

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k + \epsilon = B'x + \epsilon$$

Και τα υπόλοιπα ϵ είναι ανεξάρτητες και ισόνομες τ.μ. οι οποίες ακολουθούν κανονική κατανομή $\rightarrow \epsilon \sim N(0,\sigma^2)$

Παρατηρούμε ότι η εξαρτημένη μεταβλητή Y δεν είναι πάντα θετική (δηλ. $Y > 0$), και αυτό αποτελεί πρόβλημα καθώς εκφράζει χρόνους

διάρκειας ζωής. Ένας τρόπος να διορθώσουμε αυτό το πρόβλημα είναι ο μετασχηματισμός μέσω της μεταβλητής *lnT*.

$$lnT = \mu_x + \sigma\epsilon = \mu_0 + \beta' x + \sigma\epsilon$$

Με μ_0 η παράμετρος θέσης, σ η παράμετρος κλίμακας και ϵ μια τ.μ.

Για να εισάγουμε τις συμμεταβλητές που θέλουμε στο μοντέλο, χρησιμοποιούμε την σχέση $\mu_x = \mu(x)$, όπου στις περισσότερες περιπτώσεις είναι $\mu(\mathrm{x}) = exp(\beta'\mathrm{x})$.

Η συνάρτηση επιβίωσης, τότε, θα γίνει ως εξής:

$$\begin{aligned} S(t;x) &= P(T_x > t) \\ &= P(lnT_x > lnt) \\ &= P(\mu_0 + \beta' x + \sigma\epsilon > lnt) \\ &= P(\varepsilon > \frac{lnt - \mu - \beta' x}{\sigma}) \\ &= S_\varepsilon\left\{\frac{lnt-\mu-\beta' x}{\sigma}\right\} \end{aligned}$$

όπου S_ε είναι συνάρτηση επιβίωσης της τ.μ. *ε*.

Η γενική μορφή της συνάρτησης επιβίωσης γράφεται ως εξής:

$$S(t;x) = S_0(tg(x))$$

Όπου $g(x) > 0$ συνάρτηση των συμμεταβλητών.

Θα αναφέρουμε κάποιες ειδικές περιπτώσεις των μοντέλων επιταχυνόμενης διακοπής:

- <u>Κατανομές Gumbel και Weibull</u>

Έστω τ.μ. *T~Weibull(α,η),* τότε, αν θεωρήσουμε ότι *X=lnT* , η *X θ*α ακολουθεί την κατανομή *Gumbel*. Επομένως, θα έχουμε:

$$S(x) = exp\left[-e^{(x-\mu)/\sigma}\right], \mu\varepsilon\ \mu = ln\alpha\ \kappa\alpha\iota\ \sigma = \eta^{-1}$$

Τότε αποδεικνύεται εύκολα ότι αν τα υπόλοιπα ϵ ακολουθούν την τυποποιήμενη κατανομή *Gumbel* με *μ=0* και *σ=1*. Δηλ. $\epsilon \sim Gumbel(0,1)$

Και αν προσαρμόσουμε στο μοντέλο επιταχυνόμενης διακοπής

$$lnT_x = \mu_0 + \beta' x + \sigma\epsilon$$

Θα έχουμε ότι η T_x θα ακολουθεί την κατανομή *Weibull* για κάθε *x*.

Καταλήγουμε, λοιπόν, ότι η κατανομή Weibull είναι η μόνη που έχει την ιδιότητα να μπορεί να εκφραστεί ως μοντέλο αναλογικής διακινδύνευσης αλλά και ως μοντέλο επιταχυνόμενης διακοπής. *(Καρώνη, 2009)*

- <u>Λογιστική και Λογαριθμο-Λογιστική κατανομή</u>

Παρόμοια, έχουμε ότι για $\alpha = e^{\kappa}$, η συνάρτηση διακινδύνευσης της Λογαριθμο-Λογιστικής κατανομής είναι:

$$h_0(t) = \frac{\gamma t^{\gamma-1} e^{\kappa}}{1+e^{k} t^{\gamma}} \qquad (4.2)$$

Γενικά, για τα μοντέλα επιταχυνόμενης διακοπής έχουμε ότι:

$$\begin{aligned} h(t) &= \frac{f(t;x)}{S(t;x)} \\ &= \frac{-dln(S(t;x))}{dt} \\ &= \frac{-dln(S_0(te^{-\beta' x}))}{dt} \\ &= e^{-\beta' x} h_0(te^{-\beta' x}) \end{aligned}$$

Και με αντικατάσταση της $h_0(t)$ από την σχέση *(4.2)* καταλήγουμε στην σχέση:

$$h(t) = \frac{\gamma t^{\gamma-1} e^{\kappa-\gamma\beta' x}}{1 + (e^{\kappa-\gamma\beta' x}) t^{\gamma}}$$

Η οποία είναι η συνάρτηση διακινδύνευσης της Λογαριθμο-Λογιστικής κατανομής αλλά με $\alpha = e^{\kappa-\gamma\beta' x}$.

Τότε αποδεικνύεται ότι αν τα υπόλοιπα ϵ ακολουθούν την τυποποιημένη *Λογιστική κατανομή*, δηλ. $\epsilon \sim Logistic(0,1)$ με:

$$S_\epsilon(\epsilon) = (1 + e^\epsilon)^{-1}$$

Θα έχουμε ότι η T_x θα ακολουθεί την κατανομή *Λογαριθμο-Λογιστική* κατανομή για κάθε x.

- Λογαριθμο-Κανονική και Κανονική κατανομή

Παρόμοια με την Λογαριθμο-Λογιστική κατανομή, αν θεωρήσουμε τ.μ. διάρκειας ζωής *T* η οποία ακολουθεί την Λογαριθμο-Κανονική κατανομή, και *X=lnT* , όπου η *X* θα ακολουθεί την Κανονική κατανομή, μπορούμε να αποδείξουμε ότι αν τα υπόλοιπα ϵ ακολουθούν την τυποποιήμενη κανονική κατανομή με *μ=0* και *σ=1*. Δηλ. $\epsilon \sim N(0,1)$

Και αν προσαρμόσουμε στο μοντέλο επιταχυνόμενης διακοπής:

$$lnT_x = \mu_0 + \beta' x + \sigma\epsilon$$

Τότε, θα έχουμε ότι η T_x θα ακολουθεί την Λογαριθμο-Κανονική κατανομή για κάθε *x*.

4.3.1 Γραφικός Έλεγχος Καταλληλότητας στο Μοντέλο Επιταχυνόμενης Διακοπής

Για να ελέγξουμε την ορθότητα του μοντέλου επιταχυνόμενης διακοπής μπορούμε να χρησιμοποιήσουμε έναν γραφικό έλεγχο, όπως αντίστοιχα κάναμε και στο μοντέλο αναλογικής διακινδύνευσης.

Όπως αναφέραμε και προηγουμένως, έχουμε την γενική μορφή της συνάρτησης επιβίωσης:

$$S(t; x) = S_0(tg(x))$$

Όπου συνήθως η συνάρτηση $g(x) = e^{-\beta' x}$ και S_0 είναι μια βασική συνάρτηση επιβίωσης.

Τότε έχουμε ότι:

$$\begin{aligned} S(t;x) &= P(T_0 > tg(x)) \\ &= P(lnT_0 > lntg(x)) \\ &= P(lnT_0 > lnt + lng(x)) \\ &= S^*(y + lng(x)) \end{aligned}$$

αν $y = lnt$ και S^* είναι η συνάρτηση επιβίωσης της τ.μ $Y = lnT_0$

Προκύπτει, λοιπόν, ότι στο γράφημα της $S(t;x)$ ως προς lnt για ένα συγκεκριμένο x, θα έχει οριζόντια μετατόπιση της S^* ως προς το lnt.

Καταλήγουμε ότι για να ισχύει η υπόθεση της επιταχυνόμενης διακοπής θα πρέπει όλες οι καμπύλες $S(t;x)$ για τις διάφορες τιμές του x, να διαφέρουν μεταξύ τους μόνο ως προς τις οριζόντιες μετατοπίσεις.

Για να πραγματοποιήσουμε αυτόν τον έλεγχο, συνήθως χρησιμοποιούμε την εκτιμήτρια Kaplan-Meier για να εκτιμήσουμε τις τιμές της $\hat{S}(t)$ για κάθε ομάδα.

4.4 <u>Μοντέλο Αναλογικής Διακινδύνευσης του Cox</u>

Η επιλογή του κατάλληλου μοντέλου για την ανάλυση επιβίωσης ενός ανθρώπου διαφέρει από τις τεχνολογικές εφαρμογές, στις οποίες η επιλογή γίνεται ευκολότερα και συνήθως υιοθετείται κάποιο παραμετρικό μοντέλο με βάση προηγούμενη εμπειρία από παρόμοια δεδομένα.

Στην περίπτωση που αφορά τον άνθρωπο, έχουμε διαφορετικές περιπτώσεις πληθυσμών και κάθε μονάδα φέρει τα δικά του μοναδικά χαρακτηριστικά, επομένως απαιτείται διαφορετική αντιμετώπιση και καθιστά δυσκολότερη την προσαρμογή ενός γνωστού παραμετρικού μοντέλου.

Ως εκ τούτου, το πιο ευρέως διαδεδομένο μοντέλο αναλογικής διακινδύνευσης που χρησιμοποιείται είναι το μοντέλο του Cox *(1972)* και ανήκει στην κατηγορία των ημι-παραμετρικών μοντέλων καθώς η βασική συνάρτηση διακινδύνευσης $h_0(t)$ παραμένει ακαθόριστη.

Μέσω ενός διανύσματος x εισάγονται στο μοντέλο όλες οι συμμεταβλητές, ποσοτικές και ποιοτικές, που συμβάλουν στην αξιόπιστη πρόβλεψη του μοντέλου.

Η συνάρτηση διακινδύνευσης ορίζεται, όπως και στα μοντέλα αναλογικής διακινδύνευσης που είδαμε σε προηγούμενο κεφάλαιο, ως:

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

όπου $h_0(t)$ είναι μια βασική συνάρτηση διακινδύνευσης και β ένα διάνυσμα n συντελεστών οι οποίοι εκφράζουν ποσοτικά την επίδραση κάθε συμμεταβλητής.

Η σωρευτική συνάρτηση διακινδύνευσης θα είναι:

$$H(t;x) = \int_0^t h_0(u)e^{\beta' x}du = H_0(t)e^{\beta' x} \qquad (4.3)$$

με $H_0(t)$ είναι μια βασική συνάρτηση διακινδύνευσης.

Αντικαθιστώντας στην σχέση *(1.1)* που εισάγαμε στο Κεφάλαιο 1, την σχέση *(4.3)* θα έχουμε:

$$S(t;x) = e^{-H(t;x)} = e^{-H_0(t)e^{-\beta' x}} = \{S_0(t)\}^{\,e^{-\beta' x}}$$

όπου $S_0(t)$ είναι μια βασική συνάρτηση επιβίωσης.

Το κυριότερο χαρακτηριστικό του ημι-παραμετρικού μοντέλου αναλογικής διακινδύνευσης του Cox είναι ότι οι συγκεκριμένες παραμετρικές των συναρτήσεων $h_0(\mathrm{t})$ και $\mathrm{S_0(t)}$ δεν καθορίζονται και μόνο η επίδραση των συμμεταβλητών x μπορεί να αναλυθεί. *(Καρώνη, 2009)*

4.4.1 <u>Εκτίμηση Παραμέτρων στο μοντέλο του Cox</u>

Η εκτίμηση των παραμέτρων του μοντέλου αναλογικής διακινδύνευσης του Cox επιτυγχάνεται με την μέθοδο της μερικής πιθανοφάνειας.

Έστω ότι σταματούν να λειτουργούν k μονάδες τις διακεκριμένες χρονικές στιγμές $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$. Επίσης, ισχύει ότι $d_j = 1$ για κάθε j μονάδες που σταματούν να λειτουργούν την συγκεκριμένη

χρονική στιγμή $t_{(j)}$. Την ίδια χρονική στιγμή $t_{(j)}$ που σταματά να λειτουργεί μια μονάδα με συμμεταβλητές x_j , η πιθανότητα να σταματήσει να λειτουργεί μια συγκεκριμένη μονάδα j, δεδομένου ότι σταματάει να λειτουργεί μια μονάδα η οποία βρίσκεται σε κίνδυνο την συγκεκριμένη χρονική στιγμή, αφού η $h(t)dt$ εκφράζει τη στιγμιαία πιθανότητα διακοπής, είναι:

$$\mathrm{P} = \frac{h(t(j); x_j)}{\sum_{i \in R_j} h(t(j); x_j)} = \frac{e^{\beta' x_j}}{\sum_{i \in R_j} e^{\beta' x_j}}$$

με R_j το σύνολο των μονάδων οι οποίες βρίσκονται σε κίνδυνο αμέσως μετά την χρονική στιγμή $t_{(j)}$.

Τότε η συνάρτηση πιθανοφάνειας θα είναι:

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

και η Λογαριθμοποιημένη Πιθανοφάνεια θα είναι:

$$\ell(\beta) = \sum_{j=1}^{k} \beta' x_j - \sum_{j=1}^{k} ln \left\{ \sum_{i \in R_j} e^{\beta' x_i} \right\} \qquad (4.4)$$

Παραγωγίζοντας την σχέση *(4.4)* προκύπτουν οι πρώτες μερικές παράγωγοι:

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

Λύνοντας το σύστημα εξισώσεων $\frac{\partial \ell}{\partial \beta_r} = 0$, $r = 1,2,\cdots,p$ ως προς β θα προκύψουν οι εκτιμήσεις της μεγίστης πιθανοφάνειας $\widehat{\beta_r}$.

Σε πολλές περιπτώσεις χρειάζεται να εκτιμήσουμε τα τυπικά σφάλματα των εκτιμήσεων της μέγιστης πιθανοφάνειας,δηλ. την τυπική απόκλιση:

$$\text{se}(\widehat{\beta_r})=\sqrt{\widehat{Var}(\widehat{\beta_r})}$$

Αυτές προκύπτουν από τον αντίστροφο του πίνακα παρατηρούμενης πληροφορίας με (r,s) στοιχείο $-\frac{\partial^2 \ell}{\partial \beta_r \partial \beta_s}\Big|_{\hat{\beta}}$ όπου :

$$-\frac{\partial^2 \ell}{\partial \beta_r \partial \beta_s}\Bigg|_{\widehat{\beta}} = \sum_{j=1}^{k} \sum_{i \epsilon R_j} x_{ir} \left[x_{ir} - \frac{\sum_{l \epsilon R_j} x_{ls} e^{\beta' x_i}}{\sum_{l \epsilon R_j} e^{\beta' x_i}} \right] \frac{e^{\beta' x_i}}{\sum_{l \epsilon R_j} e^{\beta' x_i}}$$

Στην παραπάνω διαδικασία δεν παρατηρείται η συνάρτηση $h_0(t)$ και αυτό εξηγεί τον όρο "μερική πιθανοφάνεια" για την παραπάνω ανάλυση. *(Καρώνη, 2009)*

4.4.2 <u>Ισόπαλοι Χρόνοι Διακοπής</u>

Στην περίπτωση που οι χρόνοι διακοπής συμπίπτουν, δηλαδή, την χρονική στιγμή t_j σταματούν να λειτουργούν παραπάνω από μία μονάδα , τότε η συνάρτηση πιθανοφάνειας θα αλλάξει μορφή. Στην προηγούμενη περίπτωση είχαμε $d_j = 1$, ενώ τώρα θα έχουμε $d_j > 1$ την χρονική στιγμή t_j. Το κυριότερο πρόβλημα που καλούμαστε να αντιμετωπίσουμε είναι το γεγονός ότι το πλήθος των μονάδων οι οποίες σταμάτησαν να λειτουργούν πιθανόν να προέκυψαν σε διαφορετικούς χρόνους, αλλά λόγω της στρογγυλοποίησης στους χρόνους (λιγότερα σημαντικά ψηφία), να προκύπτει εν τέλει η ίδια χρονική στιγμή t_j. Σε αυτή, λοιπόν, την περίπτωση δεν μπορούμε να είμαστε σίγουροι για την ακριβή σειρά που σταμάτησαν να λειτουργούν. Επομένως, υπάρχουν $d_j!$ πιθανές σειρές εμφάνισης.

Για να αποφευχθεί το γεγονός να προκύψει μία πολύπλοκη συνάρτηση μερικής πιθανοφάνειας χρησιμοποιείται η απλή προσέγγιση του Breslow *(1974)*:

$$L_{Breslow}(\beta) = \prod_{j=1}^{k} \left\{ \frac{e^{\beta' z_j}}{\left[\sum_{i \in R_j} e^{\beta' x_i} \right]^{d_j}} \right\}$$

όπου ουσιαστικά ο όρος $\frac{e^{\beta' x_i}}{\sum_{l \epsilon R_j} e^{\beta' x_i}}$ ο οποίος ισχύει για d_j=1 αντικαθίσταται από τον όρο $\frac{e^{\beta' z_j}}{\left[\sum_{i \epsilon R_j} e^{\beta' x_i}\right]^{d_j}}$ για $d_j > 1$ με $z_j = \sum_{k=1}^{d_j} x_k$ και x_k είναι το διάνυσμα των συμμεταβλητών της μονάδας *k* που σταματά να λειτουργεί την χρονική στιγμή t_j *και k = 1,2,…*d_j

Η ακρίβεια της προσέγγισης του Breslow εξαρτάται από τον λόγο d_j/n_j . Όταν η ποσότητα d_j/n_j είναι αρκετά μικρή, τότε η προσέγγιση είναι ακριβής.

Στην αντίθετη περίπτωση όπου το d_j/n_j δεν είναι μικρό, χρησιμοποιούμε μια προσέγγιση του Cox στην οποία θεωρούμε ότι τα δεδομένα παρατηρήθηκαν σε διακριτή κλίμακα αντί για συνεχή.

Δεδομένου ότι την χρονική στιγμή t_j σταμάτησε η λειτουργία σε d_j μονάδες, η πιθανότητα να προκύψει ένα οποιοδήποτε σύνολο u που αποτελείται από d_j μονάδες θα είναι:

$$P(u) \propto e^{\beta' z_u}$$

Όπου z_u είναι το άθροισμα των συμμεταβλητών x των μονάδων του συνόλου u. Η υπό συνθήκη πιθανότητα του παρατηρούμενου συνόλου μονάδων u^* με διακοπή θα είναι:

$$P(u^* | d_j) = e^{\beta' z_j} \Big/ \sum_{u \epsilon R_j} e^{\beta' z_u}$$

Πρέπει να σημειωθεί ότι ο παρονομαστής αποτελείται από το άθροισμα όλων των δυνατών $\binom{n_j}{d_j}$ όρων, όπου n_j είναι ο αριθμός των ατόμων σε κίνδυνο αμέσως μετά την χρονική στιγμή t_j. *(Καρώνη, 2009)*

4.4.3 <u>Στρωματοποιημένη Ανάλυση στο μοντέλο του Cox</u>

Το βασικό μοντέλο του Cox επεκτείνεται και σε άλλες εφαρμογές με πιο σημαντική εκείνη της στρωματοποιημένης ανάλυσης (*Stratified Cox Model),* κυρίως *σ*τις περιπτώσεις όπου οι συναρτήσεις διακινδύνευσης

δεν βρίσκονται σε αναλογία μεταξύ τους. Δηλαδή, η υπόθεση της αναλογικής διακινδύνευσης δεν ισχύει για το σύνολο των δεδομένων, αλλά μπορεί να ισχύει για κάποια υποσύνολα αυτών.

Έστω ότι μια κατηγορική μεταβλητή είναι αν ο ασθενής είναι καπνιστής ή όχι. Μελετάμε, τότε, την μεταβλητή σε δύο στρώματα.

Η συνάρτηση διακινδύνευσης διαμορφώνεται ως εξής:

$$h(t;x) = \begin{cases} e^{\beta' h_{01}(t)} \text{ , } \alpha\nu \text{ } \epsilon\acute{\iota}\nu\alpha\iota \text{ } \kappa\alpha\pi\nu\iota\sigma\tau\acute{\eta}\varsigma \\ e^{\beta' h_{02}(t)} \text{ , } \alpha\nu \text{ } \epsilon\acute{\iota}\nu\alpha\iota \text{ } \mu\eta \text{ } \kappa\alpha\pi\nu\iota\sigma\tau\acute{\eta}\varsigma \end{cases}$$

όπου $h_{01}(t)$ $\kappa\alpha\iota$ $h_{02}(t)$ είναι οι βασικές συναρτήσεις διακινδύνευσης των καπνιστών και των μη-καπνιστών, αντίστοιχα. Θεωρούμε, ακόμη, ότι οι $h_{01}(t)$ και $h_{02}(t)$ δεν βρίσκονται σε αναλογία μεταξύ τους και η ιδιότητα της αναλογικής διακινδύνευσης συνεχίζει να ισχύει για τις υπόλοιπες συμμεταβλητές του μοντέλου. Επιπλέον, βλέπουμε ότι οι συντελεστές β των συμμεταβλητών είναι ίδιοι και στα δύο στρώματα.

Για να εκτιμήσουμε τις παραμέτρους του μοντέλου θα χρησιμοποιήσουμε την μέθοδο της μέγιστης πιθανοφάνειας και για κάθε στρώμα m η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας θα είναι :

$$\ell_m(\beta) = \sum_{j=1}^{k_m} \beta' x_{mj} - \sum_{j=1}^{k_m} ln \left\{ \sum_{i \epsilon R_{mj}} e^{\beta' x_{mi}} \right\}$$

Η διαφορά με την προηγούμενη εκτίμηση είναι η προσθήκη του δείκτη m, έτσι ώστε να διαχωριστούν οι περιπτώσεις των διαφορετικών δεδομένων του στρώματος m.

Για το σύνολο των στρωμάτων με $m = 1, \ldots, n$ ισχύει:

$$\ell(\beta) = \sum_{\mathrm{j}=1}^{k_m} \ell_m(\beta) = \sum_{m=1}^{n} \sum_{j=1}^{k_m} \beta' x_{mj} - \sum_{m=1}^{n} \sum_{j=1}^{k_m} ln \left\{ \sum_{i \epsilon R_{mj}} e^{\beta' x_{mi}} \right\}$$

Με αυτή την μέθοδο μπορούμε να ελέγξουμε αν ισχύει η υπόθεση της αναλογικής διακινδύνευσης.

4.4.4 Έλεγχοι Υπόθεσης Αναλογικής Διακινδύνευσης

Για να χρησιμοποιήσουμε το μοντέλο αναλογικής διακινδύνευσης του Cox θα πρέπει αρχικά να σιγουρευτούμε ότι είναι το κατάλληλο, δηλαδή, να ικανοποιείται η ιδιότητα της αναλογικής διακινδύνευσης η οποία υποστηρίζει ότι πρέπει ο λόγος,

$$\frac{h(t;x_i)}{h(t;x_j)} = \frac{e^{\beta' x_i} h_0(t)}{e^{\beta' x_j} h_0(t)} = e^{\beta'(x_i - x_j)}$$

να είναι ανεξάρτητος του χρόνου *t* μεταξύ δύο μονάδων *i* και *j*.

Υπάρχουν δύο τρόποι για να ελέγξουμε αν ισχύει η ιδιότητα της αναλογικής διακινδύνευσης στο μοντέλο του Cox.

i. Ο πρώτος τρόπος είναι να ελέγξουμε ξεχωριστά κάθε συμμεταβλητή x_i και να ορίσουμε μία καινούρια μεταβλητή $\boldsymbol{z = x_i t}$,η κάποια άλλη συνάρτηση του χρόνου.

Στην συνέχεια, προσαρμόζουμε το μοντέλο του Cox συμπεριλαμβάνοντας την μεταβλητή z που ορίσαμε στο προηγούμενο βήμα.

Τέλος, ελέγχουμε την υπόθεση $H_0 : \{\beta_z = 0\}$ όπου αν η μηδενική υπόθεση γίνει δεκτή, τότε συμπεραίνουμε ότι η ιδιότητα αναλογικής διακινδύνευσης ισχύει.

ii. Ένας δεύτερος τρόπος για τον έλεγχο της ιδιότητας αναλογικής διακινδύνευσης είναι μέσω της γραφικής παράστασης της μη-παραμετρικής συνάρτησης $ln\{-ln\,\hat{S}\}$ έναντι του χρόνου t, όπου $\hat{S}$ οι εκτιμήσεις Kaplan-Meier. Για να ισχύει η υπόθεση της αναλογικής διακινδύνευσης πρέπει οι καμπύλες που θα σχηματίσουμε να είναι παράλληλες μεταξύ τους.

Η εκτιμήτρια Kaplan-Meier δεν λαμβάνει υπόψιν τις άλλες συμμεταβλητές του μοντέλου, παρά μόνο αυτή η οποία καθορίζει τις ομάδες. Για αυτόν τον λόγο εργαζόμαστε ως εξής:

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

Δεδομένου ότι, στο ημι-παραμετρικό μοντέλο του Cox δεν γνωρίζουμε τη βασική συνάρτηση επιβίωσης $S_0(t;x)$, θα την εκτιμήσουμε μέσω της μη-παραμετρικής εκτιμήτριας του Breslow λαμβάνοντας υπόψιν όλες τις συμμεταβλητές του στρωματοποιημένου μοντέλου του Cox.

$$\hat{S}_0(t) = e^{-\widehat{H_0}(t)} \ ,$$

με $$\widehat{H_0}(t) = \sum_{t_j \le t} \left(\frac{d_j}{\sum_{i \epsilon R_j} e^{\beta' x_i}} \right)$$

Για κάθε στρώμα *m* εκτιμάμε τη βασική συνάρτηση επιβίωσης $S_0(t)$ όπου:

$$\hat{S}_m(t) = \hat{S}_{0m}(t) e^{\widehat{\beta}' \bar{x}_m}$$

Με $\bar{x}_m$ είναι το διάνυσμα μέσων τιμών των συμμεταβλητών στο στρώμα m και ακολουθούμε την ίδια διαδικασία αλλά αυτή την φορά σχεδιάζουμε τις καμπύλες των $ln\{ - ln\, \hat{S}_m(t) \}$ συναρτήσει του λογαριθμοποιημένου χρόνου lnt, για *m=1,2,… ,p*. Αν οι καμπύλες είναι παράλληλες μεταξύ τους, τότε ισχύει η υπόθεση της αναλογικής διακινδύνευσης. *(Καρώνη , 2009)*

4.4.5 <u>Έλεγχος μέσω Υπολοιπων</u>

Αφού προσαρμόσουμε το μοντέλο στα δεδομένα επιβίωσης χρόνου, είναι απαραίτητο να ελέγξουμε την καταλληλότητα του. Ένας αξιόπιστος τρόπος για να το επαληθεύσουμε είναι ο έλεγχος μέσω των υπολοίπων του μοντέλου. Στην ουσία ο έλεγχος που πραγματοποιούμε σχετίζεται με τα τυχαία σφάλματα που προκύπτουν από την προσαρμογή του μοντέλου παλινδρόμησης και τον έλεγχο κάποιων βασικών υποθέσεων που πρέπει να πληρούνται. Στην γενικότερη περίπτωση των μοντέλων γραμμικής παλινδρόμησης, τα υπόλοιπα ορίζονται ως εξής:

$$\begin{aligned} \hat{\varepsilon}_i &= y_i - \hat{y}_i \\ &= y_i - \widehat{\beta}' x_i \end{aligned}$$

όπου y_i είναι η παρατηρούμενη τιμή και $\hat{y}_i$ η προσαρμοσμένη τιμή.

Ένας τρόπος με τον οποίο μπορούμε να εξετάσουμε τα υπόλοιπα είναι συνήθως μέσω γραφικών παραστάσεων, ελέγχοντας την κατανομή τους και την ύπαρξη έκτοπων τιμών (outliers). Τα είδη υπολοίπων διαφέρουν σε κάθε περίπτωση, και σε αυτή την ενότητα θα ασχοληθούμε με τα διαφορετικά είδη υπολοίπων που χρησιμοποιούνται για το μοντέλο αναλογικής διακινδύνευσης του Cox.

4.4.5.1 Υπόλοιπα Cox-Snell

Οι Cox και Snell *(1968*) πρότειναν τα υπόλοιπα Cox-Snell, ως ένα συγκεκριμένο παράδειγμα του γενικού ορισμού των υπολοίπων και είναι εκείνα που χρησιμοποιούνται περισσότερο στην ανάλυση δεδομένων επιβίωσης, ιδιαίτερα για στην χρήση του μοντέλου αναλογικής διακινδύνευσης του Cox, τα οποία ορίζονται ως εξής:

$$-ln\hat{S}(t_{(j)}; x_j) = \hat{H}(t_{(j)}; x_j) = \hat{H}_0(t_{(j)})e^{\hat{\beta}' x_j} = \hat{\varepsilon} \qquad (4.5)$$

όπου $\hat{S}$ οι εκτιμήσεις της συνάρτησης επιβίωσης,
$\hat{H}$ οι εκτιμήσεις της συνάρτησης διακινδύνευσης,
$\hat{H}_0$ μια μη-παραμετρική εκτιμήτρια.

Ωστόσο, τα υπόλοιπα Cox-Snell είναι ιδιαίτερα χρήσιμα για τα παραμετρικά μοντέλα, σε αντίθεση με τα μη-παραμετρικά, στα οποία η βασική σωρευτική συνάρτηση διακινδύνευσης $\hat{H}_0$ παραμένει άγνωστη.

4.4.5.2 Υπόλοιπα Schoenfeld

Τα υπόλοιπα που προκύπτουν από το προσαρμοσμένο μοντέλο είναι πιο εύκολο να ερμηνευθούν όταν εκφράζουν με κάποιον τρόπο την διαφορά ανάμεσα στις παρατηρούμενες τιμές και τις προβλεπόμενες τιμές που προκύπτουν από το προσαρμοσμένο μοντέλο. Ένα είδος υπολοίπων όπου καθιστώνται ιδιαίτερα χρήσιμα για τον έλεγχο του μοντέλου αναλογικής διακινδύνευσης του Cox είναι τα υπόλοιπα Schoenfeld *(1982)*. *(Καρώνη, 2004)*

Όπως είδαμε και σε προηγούμενη ενότητα, γνωρίζουμε ότι η πιθανότητα να σταματήσει να λειτουργεί μια μονάδα *j*, δεδομένου ότι τη χρονική

στιγμή $t_{(j)}$ σταματάει να λειτουργεί μια μονάδα και πριν από αυτή τη χρονική στιγμή βρίσκονται σε κίνδυνο R_j μονάδες, είναι:

$$P = p_i = \frac{e^{x_{(j)}'\beta}}{\sum_{i\epsilon R_j} e^{x_{(j)}'\beta}}$$

Παρόλα αυτά, δεν είναι γνωστό ποια μονάδα από το σύνολο R_j θα είναι αυτή που θα σταματήσει η λειτουργία της την χρονική στιγμή $t_{(j)}$. Τότε, θα έχουμε ότι η τιμή των συμμεταβλητών της συγκεκριμένης μονάδας θα είναι τ.μ. και θα έχει αναμενόμενη τιμή:

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

Θα πρέπει να σημειωθεί, ότι τα υπόλοιπα Schoenfeld έχουν ένα ιδιαίτερο χαρακτηριστικό σε σχέση με τις άλλες κατηγορίες υπολοίπων, καθώς προσδιορίζονται από τις συμμεταβλητές x και όχι από τις τιμές της εξαρτημένης μεταβλητής $y_i - \hat{y}_i$, όπως στα υπόλοιπα μοντέλα παλινδρόμησης *(Καρώνη, 2004)*. Επιπλέον, τα συγκεκριμένα υπόλοιπα αποτελούν διανύσματα, επομένως, κάθε μη – αποκομμένη παρατήρηση έχει τον ίδιο αριθμό υπολοίπων όσα είναι και οι συμμεταβλητές. Για τον λόγο αυτό, τα υπόλοιπα Schoenfeld ορίζονται ως:

$$\hat{r}_j = x_i\text{-}\widehat{E}(x|R_j)$$

όπου $\widehat{E}(x|R_j)$ είναι η εκτιμώμενη τιμή των συμμεταβλητών όταν αντικαταστήσουμε τις παραμέτρους β με τις εκτιμώμενες παραμέτρους $\widehat{\beta}$, στην αναμενόμενη τιμή των συμμεταβλητών $E(x|R_j)$.

Μια άλλη εκδοχή των υπολοίπων Schoenfeld που προτιμούνται για την εξέταση της υπόθεσης της αναλογικότητας των κινδύνων παρουσιάζονται στη συνέχεια και ονομάζονται κλιμακοποιημένα υπόλοιπα (Scaled).

Θεωρούμε το διάνυσμα των υπολοίπων $\hat{r}_j$ την χρονική στιγμή $t_{(j)}$ με k το πλήθος των μη-αποκομμένων παρατηρήσεων και $\widehat{V}(\widehat{\beta})$ ο πίνακας διασποράς του $\widehat{\beta}$. Τότε, θα έχουμε ότι τα κλιμακοποιημένα υπόλοιπα θα είναι:

$$r_j^* = k\widehat{V}(\widehat{\beta})\hat{r}_j$$

4.4.5.3 <u>Υπόλοιπα Martingale</u>

Σε πολλές περιπτώσεις, οι υποθέσεις ενός μοντέλου παραβιάζονται χωρίς να γνωρίζουμε ακριβώς τα αίτια που συμβαίνει αυτό. Χρειαζόμαστε, λοιπόν, κάποια ένδειξη ώστε να εντοπίσουμε τα σημεία επιρροής του μοντέλου. Μία κατηγορία υπολοίπων που χρησιμοποιούνται για αυτό τον σκοπό κατά κόρον στην ανάλυση του μοντέλου αναλογικής διακινδύνευσης του Cox, είναι τα υπόλοιπα Martingale που προτάθηκαν από τους Barlow και Prentice *(1988)*. Προκύπτουν από την σχέση:

$$\widehat{M}_i(t) = N_i(t) - \int_0^t Y_i(s) e^{\widehat{\beta}' X_i(s)} d\widehat{H}_0(s) \quad για \quad i = 1, \ldots, n$$

όπου και $\widehat{H}_0$ είναι η σωρευτική συνάρτηση διακινδύνευσης και $N_i(t)$ είναι ο αριθμός των παρατηρήσεων για τα οποία έγινε το συμβάν κατά την διάρκεια του *t*.

Μια προσέγγιση για την εξέταση των υπολοίπων Martingale είναι να σημειωθεί η διαφορά μεταξύ του παρατηρούμενου αριθμού θανάτων για το i-οστό άτομο στην έρευνα, μέχρι την χρονική στιγμή t_i, και τον αντίστοιχο εκτιμώμενο αριθμό που προέβλεψε το μοντέλο.

Τα υπόλοιπα Martingale στηρίζονται στην εκτίμηση της Nelson-Aalen της σωρευτικής συνάρτησης διακινδύνευσης. Ορίζουμε ως υπόλοιπα Martingale την σχέση:

$$\widehat{M}_i = \delta_i - \hat{\varepsilon} \qquad (4.6)$$

όπου το δ_i εκφράζει την τελική κατάσταση.

Όπως είδαμε και στην *(4.5)* τα υπόλοιπα Cox-Snell ορίζονται ως:

$$\hat{\varepsilon} = \widehat{H}(t_{(j)}; x_j) = \widehat{H}_0(t_{(j)}) e^{\widehat{\beta}' x_j}$$

Επομένως, η σχέση *(4.6)* γίνεται:

$$\widehat{M}_i = \delta_i - \widehat{H}_0(t_{(j)}) e^{\widehat{\beta}' x_j} \qquad (4.7)$$

Το εύρος των τιμών που μπορεί να πάρει το δ_i είναι μεταξύ του $-\infty$ και του 1 για τα υπόλοιπα που εφαρμόζονται σε δεδομένα με αποκομμένες παρατηρήσεις. Όταν το $\delta_i = 0$ τότε η σχέση *(4.7)* θα πάρει αρνητική τιμή. Αυτό συμβολίζει ότι το άτομο θα ζήσει παραπάνω από την πρόβλεψη του μοντέλου. Αντίθετα, όταν το $\delta_i = 1$, σημαίνει ότι το άτομο απεβίωσε νωρίτερα από την πρόβλεψη του μοντέλου.

4.4.5.4 Υπόλοιπα Deviance

Αν και τα υπόλοιπα Martingale παρουσιάζουν πολλά πλεονεκτήματα, ένα πολύ σημαντικό μειονέκτημα σύμφωνα με τους Therneau et al. *(1990)* είναι η έλλειψη συμμετρίας γύρω από το μηδέν, δηλ. η ύπαρξη λοξότητας στο γράφημα. Αυτή η λοξότητα παραμορφώνει την εμφάνιση των γραφημάτων με αποτελέσματα να είναι αρκετά δύσκολο να ερμηνευθούν. Αυτό μπορεί να συμβεί ακόμα και όταν το μοντέλο προσαρμόζεται σωστά. Πρότειναν αντ'αυτών τα υπόλοιπα Deviance τα οποία κατανέμονται καλύτερα γύρω από το μηδέν και μπορούν να φανούν περισσότερο χρήσιμα για την ερμηνεία του αποτελέσματος.Τα υπόλοιπα Deviance ορίζονται από την σχέση:

$$D_i = sgn(\widehat{M}_i)\sqrt{\left[-2\{\widehat{M}_i + \delta_i ln(\delta_i - \widehat{M}_i)\}\right]}$$

και προκύπτει από την διαφορά της μέγιστης λογαριθμοποιημένης πιθανοφάνειας ℓ_i του τρέχοντος μοντέλου του i-οστού ατόμου και της μέγιστης λογαριθμοποιημένης πιθανοφάνειας $\hat{\ell}_i$ του μοντέλου με όλες τις συμμεταβλητές για το συγκεκριμένο άτομο. Δηλαδή,

$$D_i = -2\{ln\ell_i - ln\hat{\ell}_i\}$$

Γενικά, οι παρατηρήσεις που αντιστοιχούν σε σχετικά μεγάλα υπόλοιπα είναι τα σημεία επιρροής του μοντέλου και που αποκλίνουν από την προσαρμογή του. *(Fitrianto and Ting Jiin, 2013)*

Σημεία Επιρροής DF BETAS

Στην στατιστική αρκετά συχνά τυχαίνει η εκτίμηση των συντελεστών β να επηρεάζεται από κάποιες συγκεκριμένες παρατηρήσεις παραπάνω

απ'ότι οι υπόλοιπες. Αυτές τις παρατηρήσεις τις ονομάζουμε "σημεία επιρροής". Ένα είδος σημείων επιρροής στο μοντέλο παλινδρόμησης είναι και τα λεγόμενα DF BETAS τα οποία εισήγαγαν οι Belsley et al. *(1980)*.

Έστω ότι $\widehat{\beta}_{j(i)}$ είναι η εκτίμηση της παραμέτρου β_i, όταν παραλείψουμε από το μοντέλο την παρατήρηση i. Τότε, θα έχουμε ότι τα DF BETAS προκύπτουν από την σχέση:

$$DFBETAS_{ji} = \frac{\widehat{\beta}_j - \widehat{\beta}_{j(i)}}{\sqrt{S^2_{(i)} c_{jj}}}, με\ i = 1,2,\dots,n,\ j = 0,1,2,\dots,k$$

με n το μέγεθος του δείγματος, k ο αριθμός των επεξηγηματικών μεταβλητών, $S^2_{(i)}$ είναι η διασπορά των υπολοίπων όταν η παραλείπεται η παρατήρηση και c_{jj} είναι ο αριθμός των επαναλήψεων της x_j.

Όταν τα $DFBETAS_{ji}$ παρουσιάζουν μεγάλες τιμές, τότε, οι αντίστοιχες παρατηρήσεις είναι πιθανό να ασκούν επιρροή στο μοντέλο παλινδρόμησης *(Καρώνη, 2010)*.

Ειδικότερα, αν:

$$|DFBETAS_{ji}| > 2/\sqrt{n}$$

τότε, η παρατήρηση *i* φαίνεται να έχει μεγάλη επιρροή στην εκτίμηση του $\widehat{\beta}_j$.

4.5 <u>Κριτήρια Επιλογής Μεταβλητών</u>

Κατά την διάρκεια της διεξαγωγής μιας στατιστικής έρευνας, συνήθως πρέπει να λάβουμε υπόψιν αρκετές παραμέτρους και μεταβλητές με σκοπό να καταλήξουμε σε αξιόπιστα συμπεράσματα. Ωστόσο, όλες οι μεταβλητές δεν συντελούν το ίδιο στην προσαρμογή του μοντέλου μας, και ιδιαίτερα στην αρχή της έρευνας είναι δύσκολο να γνωρίζουμε εκ των προτέρων ποιες από τις μεταβλητές που έχουμε στην διάθεση μας είναι στατιστικά σημαντικές για το μοντέλο μας.

Ο βασικός μας στόχος είναι η εξεύρεση του μοντέλου με ένα σύνολο από τις διαθέσιμες μεταβλητές, οι οποίες θα είναι στατιστικά σημαντικές και θα συνδράμουν στην ακριβέστερη μελλοντική πρόβλεψη. Κατά κύριο λόγο, για τα μοντέλα που θα προκύψουν προτιμούμε να είναι όσο πιο απλά και ταυτόχρονα αξιόπιστα γίνεται, καθώς συνήθως όσο περίπλοκο και σύνθετο μπορεί να είναι ένα πρόβλημα το οποίο μας θέτεται, οι παράγοντες που συνδράμουν σε μεγαλύτερο βαθμό είναι λίγοι.

Σε αυτή την ενότητα θα εστιάσουμε σε κάποια κριτήρια επιλογής μεταβλητών που συντελούν στην κατάληξη ενός αξιόπιστου μοντέλου.

4.5.1 Κριτήριο AIC

Το κριτήριο AIC (*Akaike's Information Criterion)* αναπτύχθηκε από τον Akaike (1974) και χρησιμοποιείται ευρέως για την επιλογή του βέλτιστου μοντέλου. Στην περίπτωση όπου τα μοντέλα που θέλουμε να συγκρίνουμε έχουν τον ίδιο αριθμό παραμέτρων επιλέγει το μοντέλο με την καλύτερη εφαρμογή της μεθόδου ελαχίστων τετραγώνων. Ορίζεται από την σχέση:

$$AIC = 2d - 2lnL = 2d - 2\hat{\ell}$$

όπου d είναι το πλήθος των παραμέτρων του μοντέλου και L η μεγιστοποιημένη τιμή της συνάρτηση πιθανοφάνειας και $\hat{\ell} = l\mathrm{nL}$ η μεγιστοποίηση της λογαριθμοποιημένης συνάρτησης πιθανοφάνειας.

Μέσω της ελαχιστοποίησης του AIC επιλέγεται το καταλληλότερο μοντέλο, δηλ. το μοντέλο με την μικρότερη τιμή AIC είναι εκείνο που θα προτιμήσουμε.

4.5.2 Κριτήριο BIC

Το BIC (*Bayesian Information Criterion*) προτάθηκε από τον Schwarz *(1978)* και χρησιμοποιείται επίσης για την επιλογή του καλύτερου μοντέλου με παρόμοιο τρόπο με το AIC.

Ωστόσο, η βασική διαφορά τους είναι ότι η προσθήκη επιπλέον παραμέτρων δεν προτιμάται τόσο.

Η σχέση που μας δίνει την εκτίμηση του κριτηρίου BIC είναι η εξής :

$$BIC = dlnn - 2lnL$$

όπου το d πλήθος των παραμέτρων και n το πλήθος των αποκομμένων παρατηρήσεων.

Γενικεύοντας τα παραπάνω, μπορούμε να ορίσουμε την μορφή που εκφράζει τα κριτήρια AIC και BIC :

$$-2\hat{\ell} + q$$

Όπου ανάλογα με το κριτήριο που χρησιμοποιείται κάθε φορά ο όρος q διαφοροποιείται:

AIC : q=2p , BIC: q=plnn

5. Ανάλυση Καμπυλών ROC

5.1 Εισαγωγή

Κατά την διεξαγωγή μίας έρευνας, ένα από τα κυριότερα θέματα που προκύπτουν είναι η πραγματοποίηση προβλέψεων. Μια μέθοδος για τον χαρακτηρισμό της προγνωστικής ακρίβειας ενός μοντέλου παλινδρόμησης, ειδικότερα όταν υπάρχουν αποκομμένοι χρόνοι επιβίωσης, είναι η χρήση των καμπυλών ROC (Receiver Operating Characteristic).

Οι καμπύλες ROC είναι πολύ χρήσιμες καθώς μας βοηθούν στην πρόβλεψη της ικανότητας ενός μοντέλου ώστε να ανταποκρίνεται σε νέα δεδομένα και να πραγματοποιεί ορθές προβλέψεις με σκοπό την σωστή λήψη αποφάσεων.

Οι βασικές πληροφορίες που χρειάζονται για την χρήση των καμπυλών ROC είναι:

- Η κατάσταση στο τέλος της παρακολούθησης (δυαδική)
- Η διάρκεια της παρακολούθησης

5.2 Ορισμός Καμπυλών ROC

Έστω T_i ο χρόνος επιβίωσης μιας μονάδας i. Υποθέτουμε ότι παρατηρούμε τις ελάχιστες τιμές των T_i και C_i, όπου το C_i συμβολίζει έναν ανεξάρτητο αποκομμένο χρόνο επιβίωσης.

Στην συνέχεια, ορίζουμε τον χρόνο παρακολούθησης ως:

$$X_i = min(T_i\,, C_i)$$

και τον δείκτη αποκοπής ως:

$$\Delta_i = 1(T_i \leq C_i)$$

Ο χρόνος επιβίωσης μπορεί, επίσης, να αναπαρασταθεί μέσω της διαδικασίας καταμέτρησης:

$$N_i^*(t) = 1(T_i \leq \mathrm{t})$$

ή την αντίστοιχη προσαύξηση:

$$dN_i^*(t) = N_i^*(t) - N_i^*(t-)$$

Αξίζει να επισημάνουμε ότι εστιάζουμε κυρίως στην διαδικασία καταμέτρησης $N_i^*(t)$, η οποία ορίζεται μόνο από την άποψη του χρόνου επιβίωσης T_i, σε αντίθεση με τον πιο συνηθισμένο συμβολισμό $N_i(t) = 1(X_i \leq \mathrm{t}, \Delta_i = 1)$, όπου εξαρτάται από τον αποκομμένο χρόνο επιβίωσης.

Έστω, τώρα, ο δείκτης κινδύνου:

$$R_i(t) = 1(X_i \geq \mathrm{t})$$

Θεωρούμε ότι σε κάθε μονάδα i αντιστοιχεί ένα σύνολο από συμμεταβλητές $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ip})$ οι οποίες είναι χρονικά αμετάβλητες.

Θα εστιάσουμε σε μεθόδους του μοντέλου του Cox με σκοπό να δημιουργήσουμε ένα μοντέλο score, και επιπλέον, την προβλεπτική του ικανότητα.

Ωστόσο, οι εκτιμώμενες μέθοδοι που προτείνονται μπορούν να χρησιμοποιηθούν για την ακρίβεια του προγνωστικού score που προκύπτει από οποιαδήποτε παλινδρόμηση ή προγνωστική μέθοδο, και σε αυτή την περίπτωση διάφορες μέθοδοι συντελεστών *(Hastie and Tibshirani, 1993)*, όπως η σταθμισμένη μερική εκτίμηση πιθανοφάνειας *(Cai and Sun, 2003)*, μπορούν να μας δώσουν μια ικανοποιητική προσέγγιση για την εκτίμηση ακριβέστερων αποτελεσμάτων.

Επομένως, θα κάνουμε μια σύντομη εισαγωγή στις σχετικές πτυχές της εκτίμησης της μερικής πιθανοφάνειας. Θεωρούμε ότι υπό τις συνθήκες της υπόθεσης της αναλογικής διακινδύνευσης, έχουμε:

$$\lambda(t|Z_i) = \lambda_0(t) exp(Z_i^T \beta)$$

όπου, $\lambda(t|Z_i) = \lim_{\delta \to \infty} \delta^{-1} P\big[T_i \in [t, t+\delta) | Z_i, T_i \geq t\big]\big]$

Τότε, η μερική πιθανοφάνεια των εξισώσεων θα είναι:

$$0 = \sum_{\mathrm{i}} \Delta_i \left[Z_i - \left(\sum_{\kappa} \pi_\kappa(\beta, X_i) \mathrm{Z}_\kappa \right) \right]$$

όπου, $\pi_\kappa(\beta, t) = R_\kappa(t) exp(Z_i^T \beta)/W(t)$

και $W(t) = \sum_j R_j(t) exp(Z_i^T \beta)$

Επιλύοντας τις παραπάνω εξισώσεις προκύπτουν οι εκτιμήσεις της μέγιστης πιθανοφάνειας $\widehat{\beta}$. *(Heagerty & Zheng, 2005)*

Έστω *Y* μια τ.μ, η οποία δέχεται μόνο δύο τιμές *Y=0* (αν έχουμε "επιτυχία") και *Y=1* (αν έχουμε "αποτυχία").

Τότε, εκτιμώμενη πιθανότητα επιτυχίας ορίζεται ως :

$$\hat{p} = \widehat{P}(Y = 1) = \frac{e^{x'\widehat{\beta}}}{1+e^{x'\widehat{\beta}}}$$

Έστω, επίσης,μία θετική σταθερά p_0 για την οποία ισχύει ότι :

- $\hat{p} > p_0$, τότε προβλέπεται ότι θα είναι *Y = 1*
- $\hat{p} \leq p_0$, τότε προβλέπεται ότι θα είναι *Y = 0*

Διαμορφώνεται, λοιπόν, ο πίνακας συνάφειας *Πίνακας 5.1*.

Πρόβλεψη	***Παρατήρηση***			
		Y = 1	**Y = 0**	**Άθροισμα**
	Y = 1	a	b	a+b
	Y = 0	c	d	c+d
	Άθροισμα	a+c	b+d	n

Πίνακας 5.1: *Πίνακας Συνάφειας της Καμπύλης ROC.*

5.3 Πιθανοφάνειες των Καμπυλών ROC

Οι συχνότερα χρησιμοποιούμενες συνιστώσες και ορίζονται με βάση τον παραπάνω πίνακα είναι:

► ***Ευαισθησία (Sensitivity)***

Ορίζεται ως το ποσοστό των αληθώς θετικών αποτελεσμάτων (True Positive Rate), δηλ. το ποσοστό που είχε σωστή πρόβλεψη της κατάστασης *Y=1*:

$$SE = TPR = \frac{a}{a+ b}$$

► ***Ειδικότητα (Specificity)***

Ορίζεται ως το ποσοστό των αληθώς αρνητικών αποτελεσμάτων (True Negative Rate) δηλ. το ποσοστό που είχε σωστή πρόβλεψη της κατάστασης *Y=0*:

$$SPC = TNR = \frac{d}{c+d}$$

Επιπλέον,σημαντικό ρόλο έχουν τα συμπληρωματικά τους ,δηλαδή,το ποσοστό των ψευδώς αρνητικών αποτελεσμάτων (False Negative Rate) και το ποσοστό των ψευδώς θετικών αποτελεσμάτων (False Positive Rate) για τα οποία ισχύει:

- $FNR = 1 - SE = \frac{c}{a+c}$
- $FPR = 1 - SPC = \frac{b}{b+d}$

► ***Ακρίβεια (Accuracy)***

Η Ακρίβεια εκφράζει το ποσοστό των πραγματικών αποτελεσμάτων. Συμπεριλαμβάνει τα αληθώς θετικά και τα αληθώς αρνητικά αποτελέσματα στο σύνολο των παρατηρήσεων.

Ορίζεται ως:

$$ACC = \frac{a+d}{n}$$

► ***Θετικός Λόγος Πιθανοφανειών***

Για να προσεγγίσουμε τον ρυθμό με τον οποίο εμφανίζεται θετικό αποτέλεσμα στα άτομα τα οποία νοσούν, σε σχέση με τα άτομα που δεν νοσούν, χρησιμοποιούμε τον Θετικό Λόγο Πιθανοφανειών όπου ορίζεται ως:

$$LR_{+} = \frac{TPR}{FPR}$$

► ***Αρνητικός Λόγος Πιθανοφανειών***

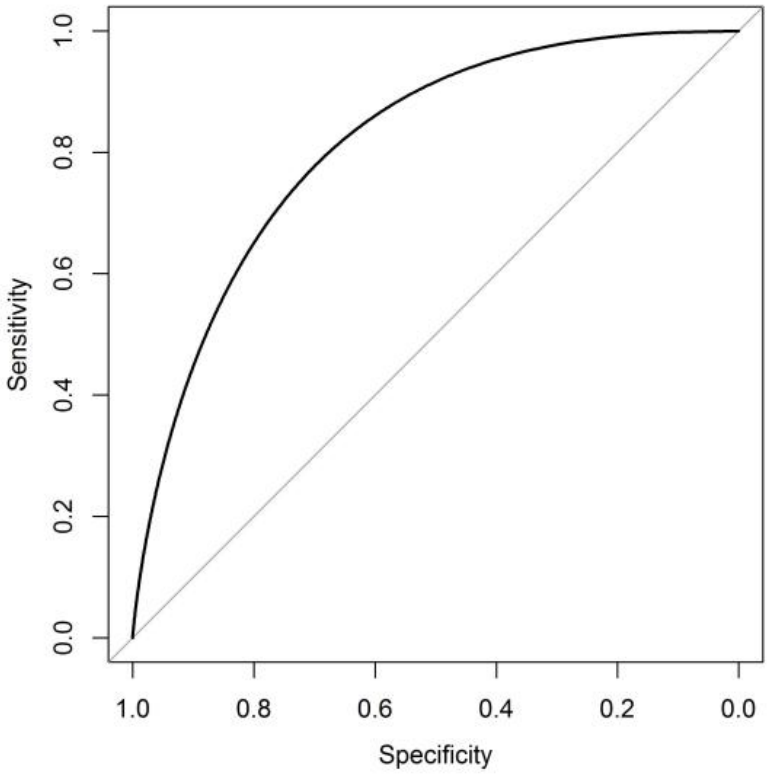
Αντίστοιχος του Θετικού Λόγου Πιθανοφανειών είναι ο Αρνητικός Λόγος Πιθανοφανειών, ο οποίος εκφράζει τον ρυθμό με τον οποίο εμφανίζεται ένα αρνητικό αποτέλεσμα στα άτομα τα οποία δεν νοσούν, σε σχέση με εκείνα που νοσούν. Ορίζεται ως:

$$LR_{-} = \frac{1}{LR_{+}} = \frac{FPR}{TPR}$$

Οι παραπάνω ποσότητες ονομάζονται *πιθανοφάνειες* (likelihoods) ή αλλιώς *λειτουργικά χαρακτηριστικά* (operating characteristics) της διαγνωστικής δοκιμασίας.

5.4 Προσαρμογή της καμπύλης ROC

Υπολογίζοντας τις πιθανοφάνειες για κάθε p_0 με $p_0 \epsilon [0,1]$ θα σχηματιστεί η καμπύλη ROC *(βλ. Σχήμα 5.1),* η οποία απεικονίζει την προβλεπτική ικανότητα του μοντέλου παλινδρόμησης. Η καμπύλη ROC είναι μια γραφική παράσταση συναρτήσει της ευαισθησίας και της ειδικότητας, δηλαδή, του ποσοστού των αληθώς θετικών τιμών, έναντι των ψευδώς θετικών τιμών.



Σχήμα 5.1: *Διάγραμμα της Καμπύλης ROC*

Το τεταρτημόριο στο οποίο σχηματίζεται η Καμπύλη ROC, χωρίζεται με μία διχοτόμο ευθεία που συμβολίζει τα καλά αποτελέσματα για τα σημεία πάνω από την ευθεία και τα κακά για τα σημεία κάτω από αυτήν.

Για να έχουμε μια επιτυχημένη πρόβλεψη, ιδανικά θέλουμε, η καμπύλη ROC να βρίσκεται κοντά στην πάνω αριστερή γωνία του διαγράμματος. Δηλαδή, να έχουμε υψηλή ευαισθησία και χαμηλή ειδικότητα και επομένως, καλύτερη προβλεπτική ικανότητα. Η περιοχή κάτω από την καμπύλη ROC εκφράζει πόσο κοντά είναι σε αυτή την γωνία με μέγιστη τιμή προφανώς την μονάδα, $\boldsymbol{AUC = 1}$, και ονομάζεται Καμπύλη AUC.

6. Εφαρμογή Μοντέλων Ανάλυσης Επιβίωσης σε Δεδομένα για Ασθενείς με Πολλαπλούν Μυέλωμα

6.1 Παρουσίαση δείγματος και μεταβλητών

Η θεωρία που αναλύθηκε παραπάνω θα χρησιμοποιηθεί για την ανάλυση πραγματικών δεδομένων όπου το δείγμα μας προέρχεται από μία μελέτη του Ιατρικού Κέντρου του Πανεπιστημίου Δυτικής Βιρτζίνια στις Η.Π.Α , προκειμένου να εξεταστεί η συσχέτιση μεταξύ συγκεκριμένων μεταβλητών με τον χρόνο επιβίωσης *(Collett, 2014)* και αποτελείται από 48 ασθενείς με ηλικία μεταξύ 50 και 80 ετών που πάσχουν από πολλαπλούν μυέλωμα.

Εμβαθύνοντας περισσότερο στην περίπτωση του πολλαπλού μυελώματος, το οποίο είναι μία κακοηθής νόσος που χαρακτηρίζεται από την συσσώρευση ανώμαλων πλασματοκυττάρων ενός τύπου λευκών αιμοσφαιρίων στον μυελό των οστών,όπου ο πολλαπλασιασμός τους μπορεί να προκαλέσει καταστροφή του οστικού ιστού καθώς και άλλες ασθένειες όπως αναιμία,αιμοραγίες, υποτροπιάζουσες λοιμώξεις και χωρίς την έγκαιρη αντιμετώπιση της οδηγεί στον θάνατο στις περισσότερες περιπτώσεις.

Τα δεδομένα παρουσιάζονται στον *Πίνακα 6.1* και παραθέτουμε ένα μικρό δείγμα 10 ασθενών.

Αριθμός Ασθενή	Χρόνος Επιβίωσης	Κατάσταση	Ηλικία	Φύλο	Επίπεδο Αζώτου Ουρίας	Ασβέστιο στο αίμα	Αιμοσφαιρίνη	Ποσοστό Πλασματοκυττάρων	Παρουσία Πρωτεΐνης Bence-Jones
1	13	1	66	1	25	10	14,6	18	1
2	52	0	66	1	13	11	12,0	100	0
3	6	1	53	2	15	13	11,4	33	1
4	40	1	69	1	10	10	10,2	30	1
5	10	1	65	1	20	10	13,2	66	0
6	7	0	57	2	12	8	9,9	45	0
7	66	1	52	1	21	10	12,8	11	1
8	10	0	60	1	41	9	14,0	70	1
9	10	1	70	1	37	12	7,5	47	0
10	14	1	70	1	40	11	10,6	27	0

Πίνακας 6.1 : *Δείγμα των δεδομένων της μελέτης.*

Η εξαρτημένη μεταβλητή που θα χρησιμοποιήσουμε είναι η θνητότητα και στον *Πίνακα 6.2* παρουσιάζεται η κωδικοποίηση του στο δείγμα.

Μεταβλητή	**Κωδικοποίηση**	
Θνητότητα (Status)	0	1
	Επιβίωση	Θάνατος

Πίνακας 6.2: *Κωδικοποίηση για την μεταβλητή "Θνητότητα".*

Στον *Πίνακα 6.3* παρουσιάζονται οι επεξηγηματικές μεταβλητές οι οποίες καταγράφηκαν κατά την στιγμή της διάγνωσης.

Age	Η ηλικία του ασθενούς (σε έτη)
Sex	Το φύλο του ασθενούς (Άνδρας = 1 , Γυναίκα = 0)
Bun	Επίπεδο Αζώτου Ουρίας στο Αίμα
Ca	Ασβέστιο στο αίμα
Hb	Αιμοσφαιρίνη
Pcells	Ποσοστό Πλασματοκυττάρων στον μυελό των οστών
Protein	Παρουσία Πρωτεΐνης Bence-Jones στα ούρα (Ναι = 0, Όχι = 1)

Πίνακας 6.3 : *Επεξηγηματικές μεταβλητές.*

Η μεταβλητή απόκρισης είναι ο χρόνος επιβίωσης t σε μήνες και εκφράζει τον χρόνο από την διάγνωση έως τον θάνατο.

Θα θεωρήσουμε ως υποθέσεις ελέγχου σε σχέση με τους ελέγχους που θα χρησιμοποιήσουμε, την μηδενική υπόθεση:

$H_0 = \{$Οι κατανομές ταυτίζονται μεταξύ τους.$\}$

με εναλλακτική υπόθεση την:

$H_1 = \{$Οι κατανομές διαφέρουν μεταξύ τους.$\}$

Χαρακτηριστικά Ασθενών

Το δείγμα μας αποτελείται από 48 ασθενείς εκ των οποίων το 60.4% είναι άνδρες και το 39.6% είναι γυναίκες. Η διάμεση ηλικία των ασθενών

στο σύνολο του δείγματος είναι 62.5 έτη και όσων αφορά την παρουσία της πρωτεΐνης Bence-Jones στα ούρα παρατηρείται ότι στις εξετάσεις του 69% των ασθενών ανιχνεύτηκε η πρωτεΐνη, ενώ στο υπόλοιπο 31% δεν ανιχνεύτηκε. Στον *Πίνακα 6.4* παρουσιάζονται αναλυτικά τα ποσοστά των χαρακτηριστικών των ασθενών.

Χαρακτηριστικά Ασθενών σε ποσοστό %	
Φύλο	
Άντρες	60.4
Γυναίκες	39.6
Παρουσία Πρωτεΐνης Bence-Jones στα ούρα	
Παρούσα	69
Απούσα	31

Πίνακας 6.4 : *Χαρακτηριστικά ασθενών σε ποσοστά.*

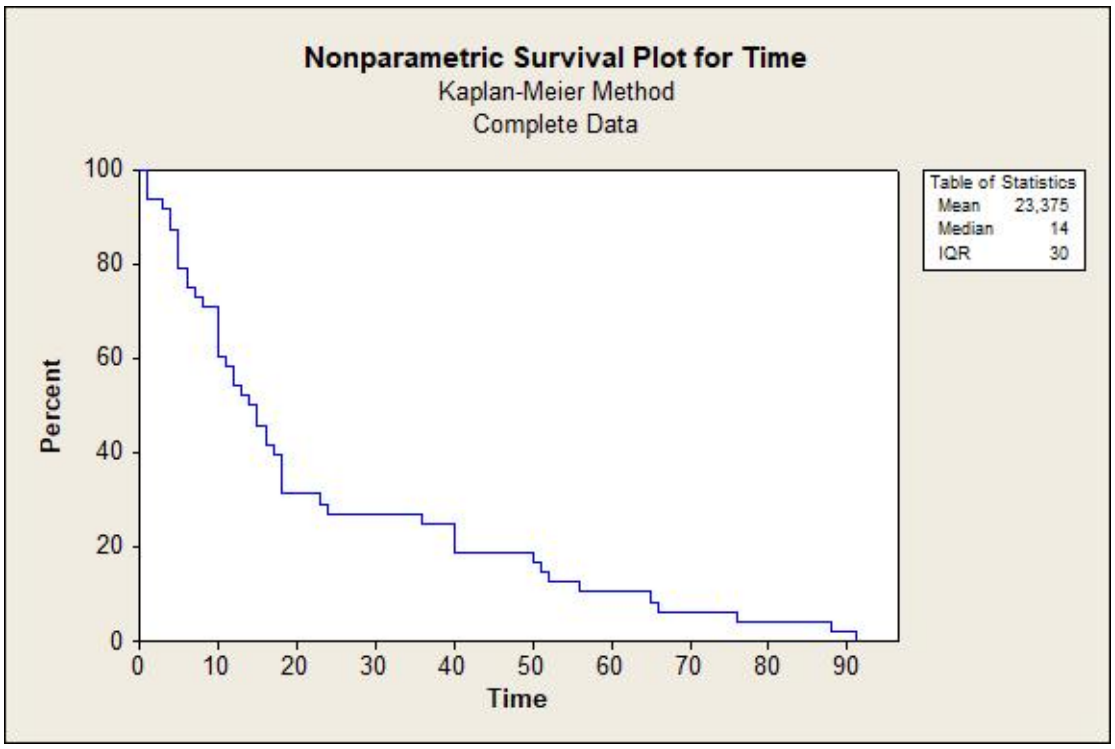
Μία πρώτη προσέγγιση των συνεχών μεταβλητών με τα βασικά μεγέθη παρουσιάζονται στον *Πίνακα 6.5*.

Μεταβλητές	**Ελάχιστη Τιμή**	**Διάμεσος**	**Μέση Τιμή**	**Μέγιστη Τιμή**
Ηλικία (Έτη)	50.00	62.50	62.90	77.00
Επίπεδα Αζώτου Ουρίας στο Αίμα	6.00	21.00	33.92	172.00
Ασβέστιο	8.00	10.00	9.938	15.00
Αιμοσφαιρίνη	4.90	10.20	10.25	14.60
Πλασματοκύτταρα στον μυελό των οστών (%)	3.00	33.00	42.94	100.00

Πίνακας 6.5 : *Χαρακτηριστικά ποσά για τις ποσοτικές μεταβλητές.*

6.2 *Εκτίμηση της Συνάρτησης Επιβίωσης με Kaplan-Meier*

Ένα από τα πρώτα βήματα που πρέπει να ακολουθήσουμε ώστε να προχωρήσουμε στην ανάλυση των παραπάνω δεδομένων είναι παρουσιάζοντας τις εκτιμήτριες Kaplan-Meier με την βοήθεια του ελέγχου Log-Rank για το σύνολο των δεδομένων. Για να εξάγουμε αποτελέσματα για τους χρόνους επιβίωσης των ασθενών ανάλογα με την ομάδα που θέλουμε να μελετήσουμε θα παρουσιάσουμε γραφικά τις εκτιμήσεις των συναρτήσεων επιβίωσης.



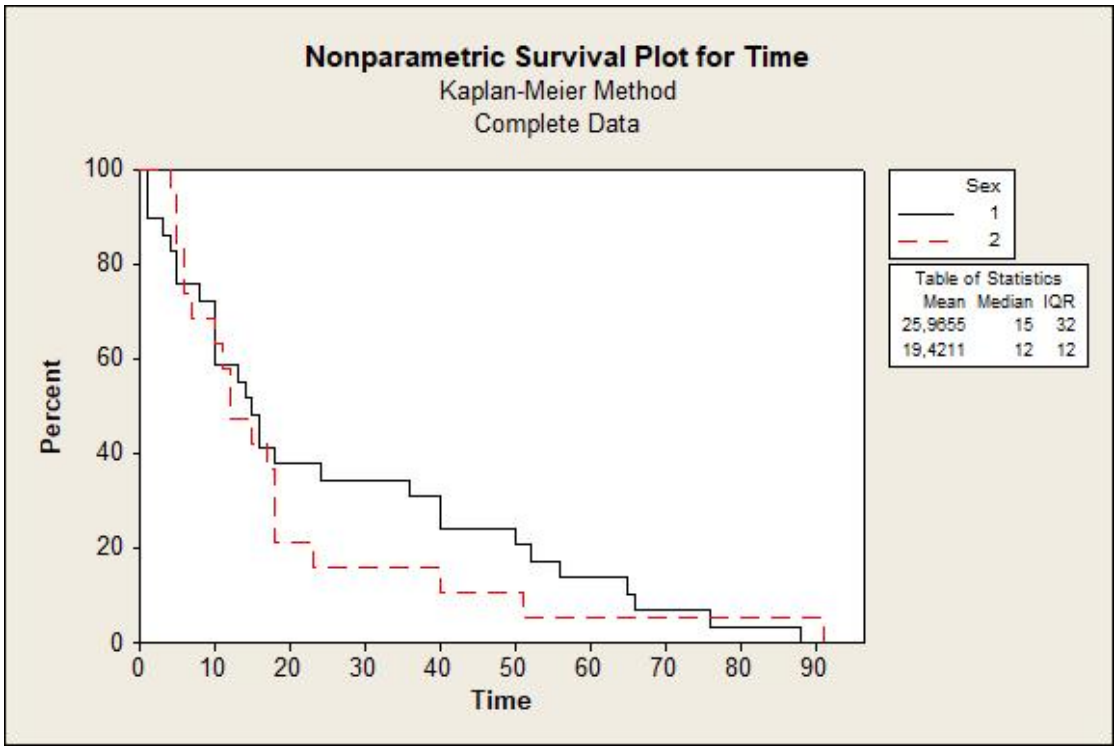
Σχήμα 6.1: *Εκτίμηση της Kaplan-Meier της συνάρτησης επιβίωσης S (t) για το σύνολο των παρατηρήσεων.*

Στο *Σχήμα 6.1* στο οποίο παρουσιάζεται η εκτίμηση της Kaplan-Meier της συνάρτησης επιβίωσης, παρατηρούμε ότι έχουμε αρκετούς θανάτους στο σύντομο χρονικό διάστημα των 18 μηνών με πιθανότητα επιβίωσης 0.43. Παρατηρείται μια σταθεροποίηση τον 24ο μήνα καθώς δεν έχουμε κάποιον θάνατο μέχρι τον 36ο μήνα, δηλ. για 12 συνεχόμενους μήνες. Επίσης, το ίδιο συμβαίνει και στον 51ο μήνα και στον 66ο μήνα, όπου

δεν παρατηρείται κάποιος θάνατος για 14 και 22 μήνες και με πιθανότητες επιβίωσης 0.22 και 0.13 αντίστοιχα.

Στην συνέχεια,θα συγκρίνουμε τις εκτιμήσεις Kaplan-Meier σε σχέση με το φύλο του ασθενή και κατά πόσο επιδρά στην συνάρτηση επιβίωσης. Από τους ελέγχους Log-Rank και Wilcoxon στον *Πίνακα 6.6* προκύπτει ότι δεν υπάρχει στατιστικά σημαντική διαφορά στην επιβίωση στους άντρες σε σχέση με τις γυναίκες με $p - value = 0.629$ για τον έλεγχο Log-Rank και αντίστοιχα $p - value = 0.720$ για τον έλεγχο Wilcoxon.

Τα αποτελέσματα των ελέγχων Log-Rank και Wilcoxon επιβεβαιώνονται και στο γράφημα στο *Σχήμα 6.2,* στο οποίο παρουσιάζονται οι εκτιμήσεις της Kaplan-Meier για τους άνδρες, με μαύρο χρώμα, και για τις γυναίκες, με κόκκινο χρώμα, καθώς παρατηρούμε ότι οι γραμμές και των δύο φύλων δεν απέχουν σημαντικά.

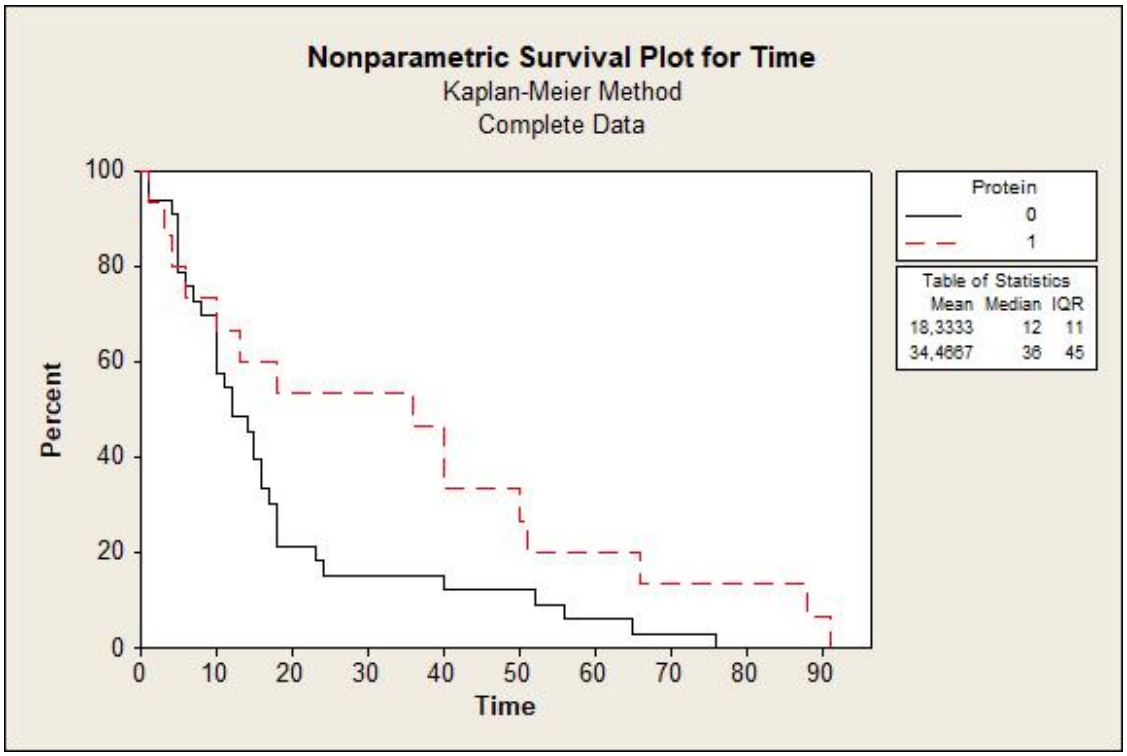


Σχήμα 6.2: *Εκτίμηση της Kaplan-Meier της συνάρτησης επιβίωσης S(t) ανάλογα το φύλο του ασθενούς. (Άντρες "—" , Γυναίκες "- -").*

Test Statistics			
Method	**Chi-Square**	**DF**	**P-Value**
Log-Rank	0.233945	1	0.629
Wilcoxon	0.128250	1	0.720

Πίνακας 6.6 : *Αποτελέσματα ελέγχων Log-Rank και Wilcoxon για την σύγκριση των πιθανοτήτων επιβίωσης ανδρών και γυναικών ασθενών.*

Τέλος, ελέγξαμε αν η παρουσία ή όχι της πρωτεΐνης Bence-Jones στα ούρα επηρεάζει την πιθανότητα επιβίωσης των ασθενών. Σύμφωνα με τα αποτελέσματα των ελέγχων Log-Rank και Wilcoxon που παρουσιάζονται στον *Πίνακα 6.7*, προκύπτει ότι δεν υπάρχει διαφοροποίηση μεταξύ της παρουσίας ή όχι της πρωτεΐνης Bence-Jones στα ούρα αφού έχουμε $p - value = 0.55$ για τον έλεγχο Log-Rank και $p - value = 0.187$ για τον έλεγχο Wilcoxon αντίστοιχα. Ωστόσο από το γράφημα του *Σχήματος 6.3* παρατηρούμε μια μικρή διαφορά στην επιβίωση εκείνων που παρουσιάζουν την πρωτεΐνη Bence-Jones, όπου και θα το ελέγξουμε και αργότερα για να επιβεβαιώσουμε τα αποτελέσματα.



Σχήμα 6.3 : *Εκτίμηση της Kaplan-Meier της συνάρτησης επιβίωσης ανάλογα με την ύπαρξη πρωτεΐνης Bence-Jones στα ούρα του ασθενούς. (Παρούσα "—", Απούσα "- -")*

Test Statistics			
Method	**Chi-Square**	**DF**	**P-Value**
Log-Rank	3.68547	1	0.055
Wilcoxon	1.74457	1	0.187

Πίνακας 6.7 : *Αποτελέσματα ελέγχων Log-Rank και Wilcoxon για την σύγκριση των πιθανοτήτων επιβίωσης ασθενών με παρουσία της πρωτεΐνης Bence-Jones.*

6.3 Εφαρμογή του Μοντέλου Αναλογικής Διακινδύνευσης του Cox

Στην συνέχεια θα προσαρμόσουμε το μοντέλο αναλογικής διακινδύνευσης του Cox και θα εφαρμόσουμε τις τεχνικές που αναφέραμε στην θεωρία για τον έλεγχο της καταλληλότητας του μοντέλου.

6.3.1 Εφαρμογή του Μοντέλου του Cox

Αφού έχουμε ορίσει όλες τις επεξηγηματικές και κατηγορικές μεταβλητές αλλά και την μεταβλητή απόκρισης και τον χρόνο επιβίωσης θα προχωρήσουμε στην προσαρμογή του μοντέλου.

Θα προσαρμόσουμε τώρα το μοντέλο του Cox χρησιμοποιώντας το στατιστικό πακέτο R, όπου η συνάρτηση επιβίωσης *S(t)* εξαρτάται από τον χρόνο επιβίωσης των ασθενών και την κατάσταση τους μετά το τέλος της μελέτης.

	Coef	**exp(Coef)**	**se(Coef)**	**z**	**p**
Ηλικία (Age)	-0.018056	0.982106	0.027833	-0.649	0.516521
Φύλο (Sex)	-0.249473	0.779211	0.403093	-0.619	0.535985
Επίπεδο Αζώτου Ουρίας στο Αίμα (Hb)	0.022661	1.022919	0.006110	3.709	0.000208***
Ασβέστιο (Ca)	0.013265	1.013353	0.132681	0.100	0.920363
Αιμοσφαιρίνη (Bun)	-0.133017	0.875450	0.068527	-1.941	0.052249*
Πλασματοκύτταρα (Pcells)	-0.001359	0.998642	0.006588	-0.206	0.836585
Πρωτείνη Bence-Jones (Protein)	-0.683269	0.504964	0.429395	-1.591	0.111556
Likelihood ratio test =17.53 **on 7 df, p** = 0.01428**n**= 48**, number of events**= 36 **AIC** = 211,15					

Πίνακας 6.8 : *Αποτελέσματα από την προσαρμογή του μοντέλου Cox.*

Από τα αποτελέσματα που παρουσιάζονται στον *Πίνακα 6.8* για το προσαρμοσμένο μοντέλο με όλες τις συμμεταβλητές, παρατηρούμε ότι από τις τιμές p των ελέγχων Wald, η πιο στατιστικά σημαντική μεταβλητή είναι η μεταβλητή που αφορά το επίπεδο αζώτου στο αίμα με p = 0.000208 και η λιγότερο στατιστικά σημαντική η μεταβλητή που αφορά το επίπεδο Αιμοσφαιρίνης με p-value = 0.052249.

Για να βελτιώσουμε το μοντέλο μας θα χρησιμοποιήσουμε την διαδικασία διαδοχικής αφαίρεσης κατά την οποία εισάγονται αρχικά όλες οι μεταβλητές του μοντέλου και αφαιρούνται μία-μία ξεκινώντας από την λιγότερη σημαντική και με βάση τον έλεγχο Wald, καταλήγοντας στο τελικό μοντέλο παίρνοντας τα αποτελέσματα στον *Πίνακα 6.9*.

	coef	**exp(coef)**	**se(coef)**	**z**	**p**
Επίπεδο Αζώτου Ουρίας στο Αίμα (Hb)	0.022042	1.022287	0.005926	3.719	0.0002 ***
Αιμοσφαιρίνη (Bun)	-0.109409	0.896364	0.061891	-1.768	0.0771
Πρωτεΐνη Bence-Jones (Protein)	-0.663097	0.515253	0.407855	-1.626	0.1040
Likelihood ratio test = 16.78 **on 3 df,** **p** = 8e-04					
Wald test = 19.25 **on 3 df,** **p** = 2e-04					
Score (logrank) test = 23.88 **on 3 df,** **p** = 3e-05					
AIC = 203.9					

Πίνακας 6.9: *Αποτελέσματα για το μοντέλο του Cox με την διαδικασία διαδοχικής αφαίρεσης.*

Με την διαδικασία της διαδοχικής αφαίρεσης καταλήγουμε στο μοντέλο με συμμεταβλητές το Επίπεδο Αζώτου, την Αιμοσφαιρίνη και την Πρωτεΐνη Bence-Jones. Από τα αποτελέσματα της προσαρμογής του μοντέλου, συμπεραίνουμε από τις τιμές των p-value ότι η μεταβλητή του επιπέδου Αζώτου είναι η πιο στατιστικά σημαντική για το μοντέλο, και ακολουθούν οι συμμεταβλητές Αιμοσφαιρίνη και πρωτεΐνη Bence-Jones, για την οποία επιβεβαιώνεται τελικά ότι επιδρά στην επιβίωση.

Επίσης, από την εφαρμογή του κριτηρίου AIC παρατηρούμε ότι το μοντέλο βελτιώθηκε αφού το *AIC = 203,9,* ενώ το μοντέλο με όλες τις συμμεταβλητές είχε το *AIC = 211,15*.

Τέλος, για να συγκρίνουμε τις συμμεταβλητές μεταξύ τους για τον βαθμό που επιδρούν στην επιβίωση των ασθενών, θα χρησιμοποιήσουμε τα αποτελέσματα του exp(coef) τα οποία εκφράζουν κατά πόσο μεταβάλλεται η συνάρτηση διακινδύνευσης. Δηλαδή, σε τι βαθμό επιδρά μια συμμεταβλητή στον χρόνο επιβίωσης αν θεωρήσουμε ότι οι υπόλοιπες μεταβλητές παραμένουν σταθερές.

Συγκεκριμένα για τα δεδομένα μας, παρατηρούμε ότι για κάθε αύξηση μίας μονάδας στον δείκτη του επιπέδου ουρίας αζώτου στο αίμα ενός ασθενούς, η συνάρτηση διακινδύνευσης πολλαπλασιάζεται κατά 1.023 φορές, δηλαδή, $h(t;x) = h_0(t) \cdot 1.023$. Αντίστοιχα, για κάθε αύξηση μιας μονάδας της αιμοσφαιρίνης, η συνάρτηση διακινδύνευσης πολλαπλασιάζεται κατά 0,896 φορές, δηλαδή, $h(t;x) = h_0(t) \cdot 0.896$. Και τέλος, η αύξηση κατά μία μονάδα της πρωτεΐνης Bence-Jones στα ούρα πολλαπλασιάζει την συνάρτηση διακινδύνευσης κατά 0.515 φορές, δηλαδή, $h(t;x) = h_0(t) \cdot 0.515$.

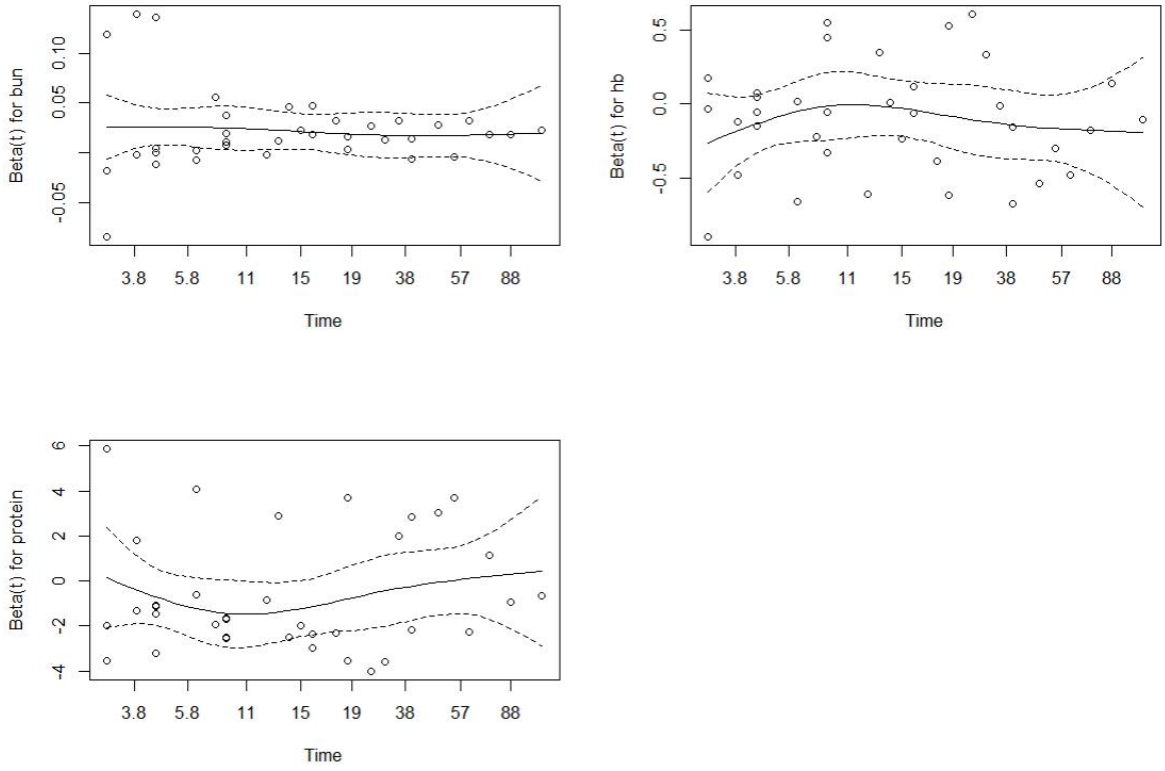
6.3.2 Έλεγχος για την υπόθεση αναλογικότητας

Θα εξετάσουμε αν το τελικό μοντέλο πληροί τις προϋποθέσεις για την υπόθεση της αναλογικότητας στο μοντέλο του Cox μέσω στατιστικών ελέγχων και γραφικών παραστάσεων με την βοήθεια του στατιστικού πακέτου R. Όπως βλέπουμε στα αποτελέσματα του ελέγχου X^2 στον *Πίνακα 6.10*, στην τρίτη στήλη για τις p-values, η υπόθεση είναι αποδεκτή για όλες τις συμμεταβλητές αλλά και για ολόκληρο το μοντέλο (GLOBAL).

	Chisq	df	p
Επίπεδο Αζώτου Ουρίας στο Αίμα (Hb)	0.7182	1	0.40
Αιμοσφαιρίνη (Bun)	0.0115	1	0.91
Πρωτείνη Bence-Jones (Protein)	0.2501	1	0.62
GLOBAL	1.3487	3	0.72

Πίνακας 6.10: *Αποτελέσματα για την υπόθεση της αναλογικότητας.*

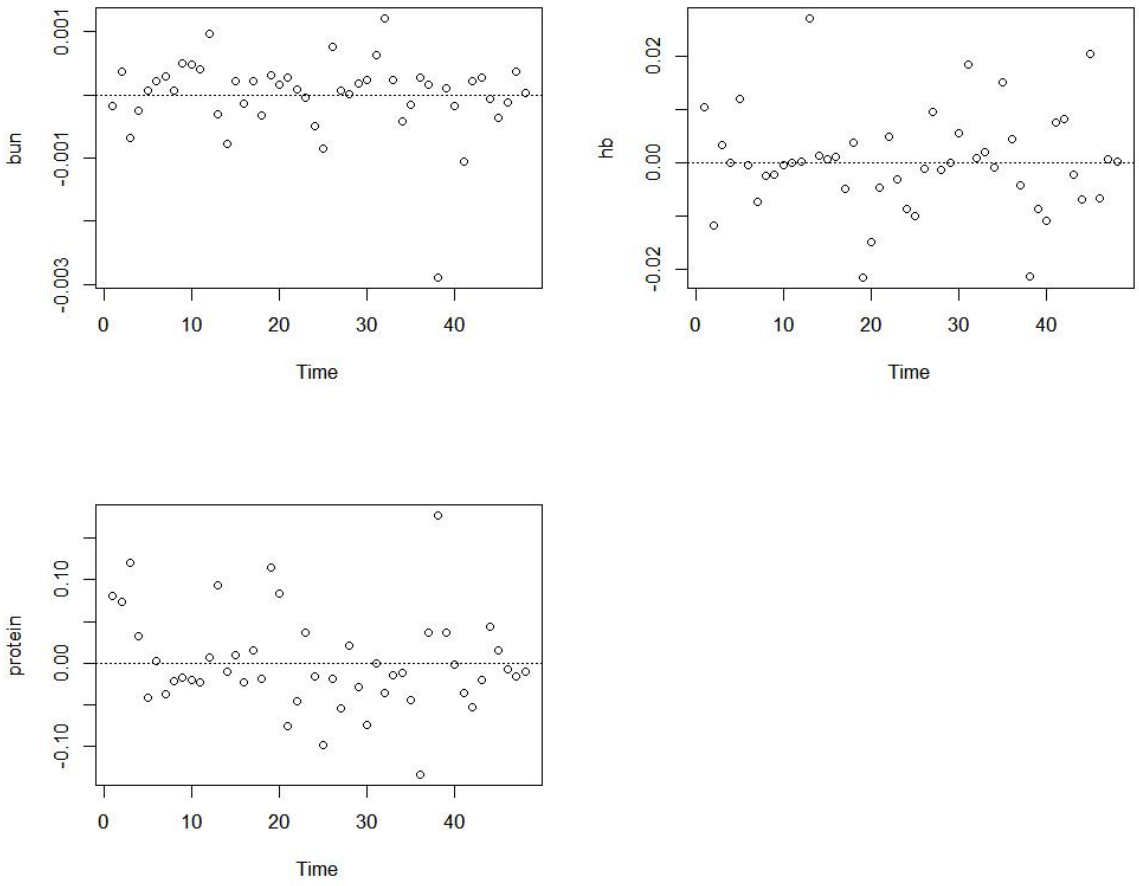
Στο *Σχήμα 6.4* παρουσιάζονται τα υπόλοιπα Schoenfeld σε συνάρτηση με τον χρόνο επιβίωσης για τις συμμεταβλητές του μοντέλου και αποτελούν μια γραφική επιβεβαίωση των ελέγχων της υπόθεσης αναλογικότητας της διακινδύνευσης. Για να επιβεβαιωθούν οι παραπάνω αριθμητικοί υπολογισμοί των ελέγχων θα πρέπει οι γραφικοί έλεγχοι να αποτελούνται από μια ευθεία γραμμή ή από μια εξομαλυμμένη καμπύλη, όπου βλέπουμε ότι όντως ισχύει και στο δείγμα μας. Οι διακεκομμένες γραμμές συμβολίζουν το ± 2 τυπικό σφάλμα της προσαρμογής.



Σχήμα 6.4 : *Γραφικός έλεγχος της υπόθεσης αναλογικότητας της διακινδύνευσης μέσω των υπολοίπων Schoenfeld.*

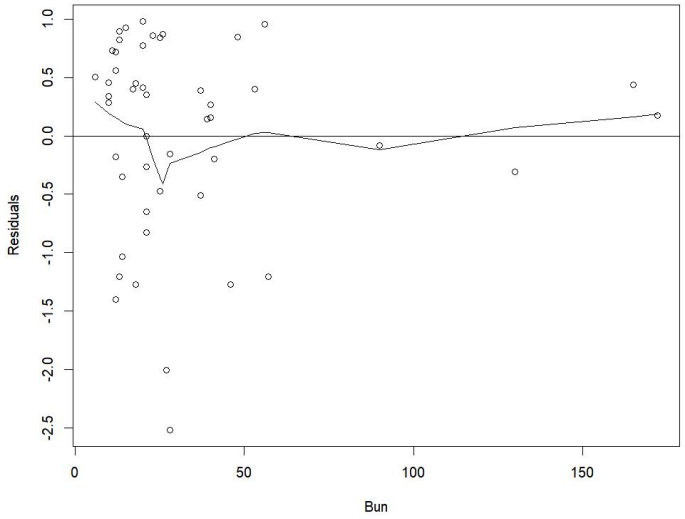
Στην συνέχεια, για να ελέγξουμε την επάρκεια του μοντέλου μας θα πρέπει να εξετάσουμε αν κάποιες από τις παρατηρήσεις μας αποτελούν σημεία επιρροής και επομένως θα έχουν μεγαλύτερη επίδραση στο μοντέλο. Για να το ελέγξουμε αυτό θα χρησιμοποιήσουμε τα υπόλοιπα DF-BETAS.

Οι παρατηρήσεις οι οποίες είναι μικρότερες κατ'απόλυτη τιμή από το σημείο αποκοπής $2/\sqrt{n}$ δεν αποτελούν σημεία επιρροής, όπου n είναι το πλήθος του δείγματος. Το σημείο επιρροής του δείγματός είναι το $2/\sqrt{n} = 0.29$. Από τα διαγράμματα των συμμεταβλητών του τελικού μοντέλου συναρτήσει του χρόνου στο *Σχήμα 6.5* συμπεραίνουμε ότι δεν παρατηρείται κάποια παρατήρηση που να ξεφεύγει από τα όρια του σημείου αποκοπής.

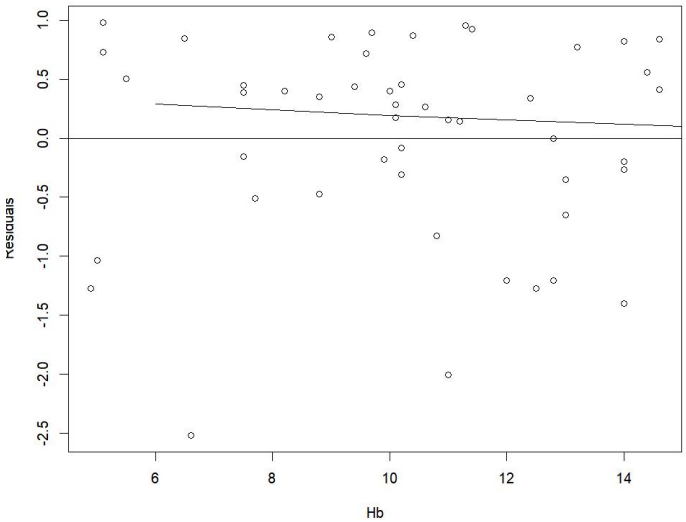


Σχήμα 6.5 : *Γραφικός έλεγχος των υπολοίπων DF-BETAS συναρτήσει του χρόνου.*

Για να αποκλείσουμε το πρόβλημα της μη-γραμμικότητας στο μοντέλο διακινδύνευσης του Cox, θα χρησιμοποιήσουμε τα υπόλοιπα Martingale. Οι συμμεταβλητή Πρωτεΐνη Bence-Jones είναι διτιμη, επομένως, δεν τίθεται θέμα μη-γραμμικότητας. Συνεπώς, θα χρησιμοποιήσουμε τις δύο συμμεταβλητές του μοντέλου, την Αιμοσφαιρίνη και το Επίπεδο Αζώτου Ουρίας στο Αίμα. Εφαρμόζουμε τα υπόλοιπα Martingale και προέκυψαν τα γραφήματα στα *Σχήματα 6.6* και *6.7* για τις συμμεταβλητές.

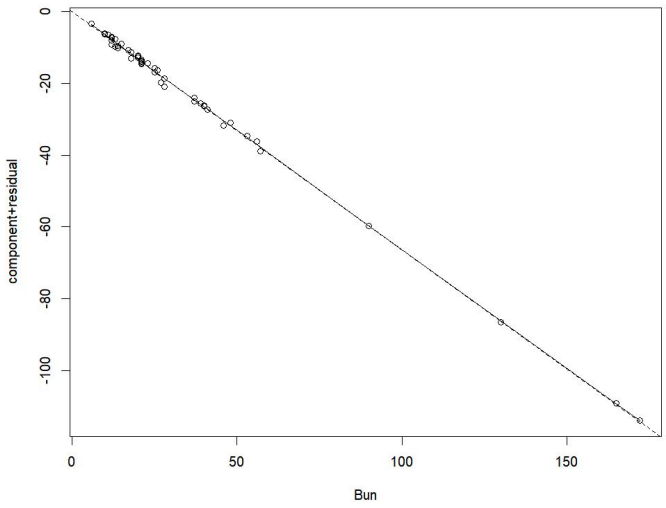


Σχήμα 6.6: *Υπόλοιπα Martingale για την μεταβλητή Επίπεδο του Αζώτου.*

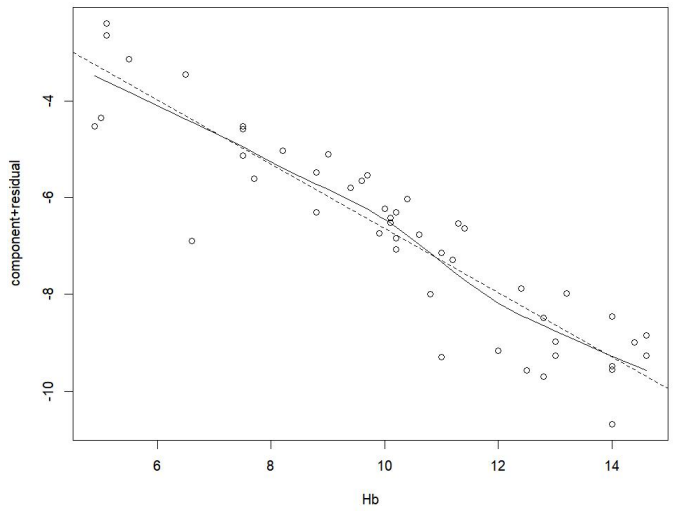


Σχήμα 6.7: *Υπόλοιπα Martingale για την μεταβλητή Αιμοσφαιρίνη.*

Χρησιμοποιώντας τα υπόλοιπα Martingale θα δημιουργήσουμε τα γραφήματα των Component υπολοίπων για να ελέγξουμε την γραμμικότητα.



Σχήμα 6.8: *Component-Residuals για την μεταβλητή Επίπεδο του Αζώτου.*



Σχήμα 6.9*: Component-Residuals για την μεταβλητή Αιμοσφαιρίνη.*

Παρατηρούμε στα *Σχήματα 6.8* και *6.9,* ότι και για τις δύο συμμεταβλητές η γραμμικότητα φαίνεται να ικανοποιείται, ιδιαίτερα για

την συμμεταβλητή Επίπεδο του Αζώτου Ουρίας στο Αίμα στην οποία φαίνεται η γραμμικότητα να είναι πολύ πιο ξεκάθαρη.

6.3.3 Στρωματοποιημένη Ανάλυση του Μοντέλου του Cox

Θα πραγματοποιήσουμε την στρωματοποιημένη ανάλυση στο τελικό μοντέλο του Cox (Stratified), έτσι ώστε να διαχωρίσουμε σε δύο στρώματα την συμμεταβλητή Πρωτεΐνη Bence-Jones για να ελέγξουμε αν οι συναρτήσεις διακινδύνευσης των δύο στρωμάτων διαφέρουν σημαντικά.

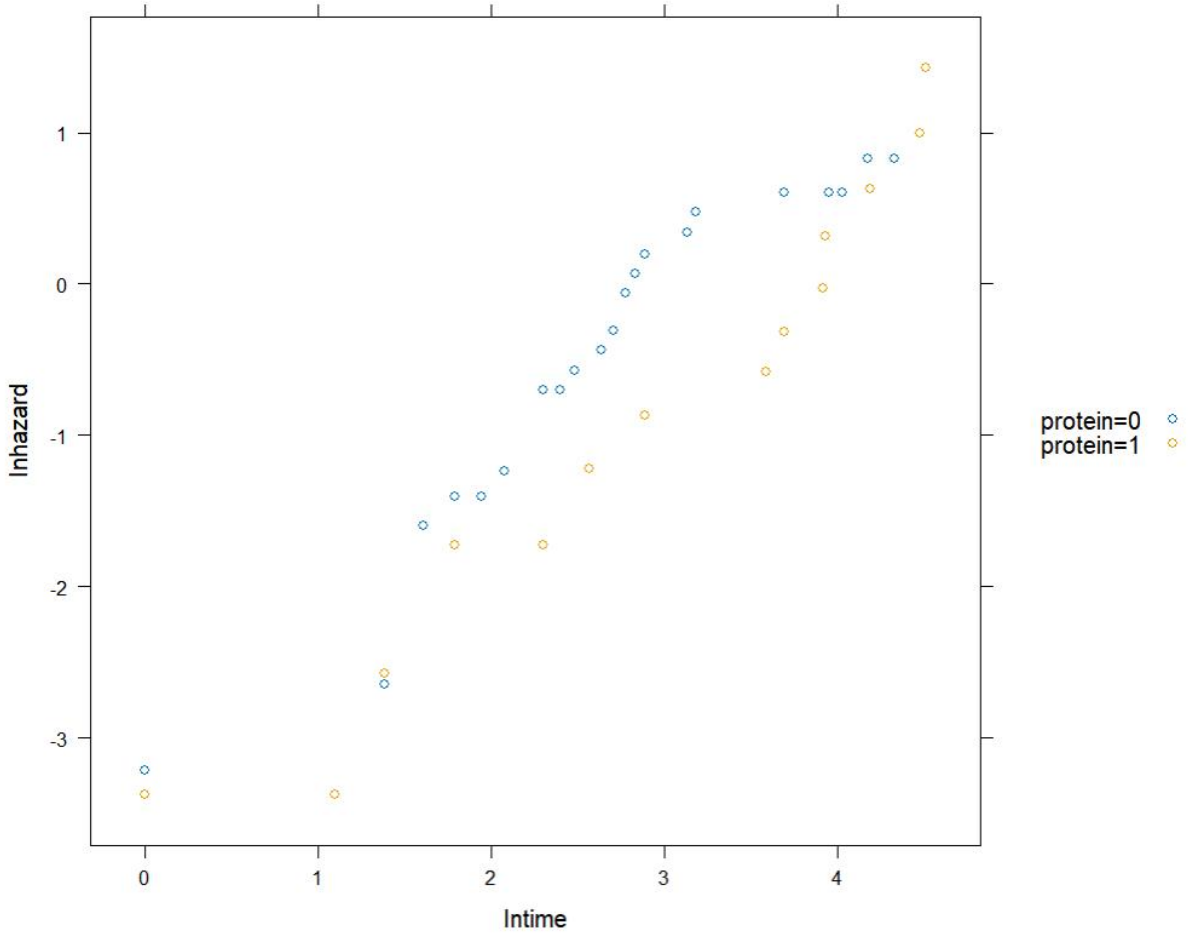
Θεωρούμε τα στρώματα m_1 και m_2:

- m_1 : {Ο ασθενής παρουσιάζει Πρωτεΐνη Bence-Jones στα ούρα.}
- m_2 : {Ο ασθενής δεν παρουσιάζει Πρωτεΐνη Bence-Jones στα ούρα.}

	Coef	**exp(Coef)**	**se(Coef)**	**z**	**p**
Επίπεδο Αζώτου Ουρίας στο Αίμα (Hb)	0.018816	1.018994	0.005802	3.243	0.00118
Αιμοσφαιρίνη (Bun)	-0.104944	0.900375	0.062088	-1.690	0.09098
Likelihood ratio test =12.21 **on 2 df,** **p** = 0.002233					
n = 48, **number of events** = 36					

Πίνακας 6.11: *Αποτελέσματα για το τελικό μοντέλο μετά την στρωματοποίηση.*

Στο *Σχήμα 6.10* παρατηρούμε ότι οι δύο κατηγορίες δεν φαίνονται να έχουν μεγάλες διαφορές καθώς τα σημεία και των δύο στρωμάτων είναι πολύ κοντά μεταξύ τους. Ωστόσο, αν έπρεπε να ξεχωρίσουμε κάποιο στρώμα, παρατηρούμε ότι το m_1, τα άτομα με παρουσία της πρωτεΐνης, φαίνεται να διατρέχει ελάχιστα παραπάνω κίνδυνο σε σχέση με το m_2, δηλ. τα άτομα που δεν παρουσιάζουν την πρωτεΐνη στα ούρα.

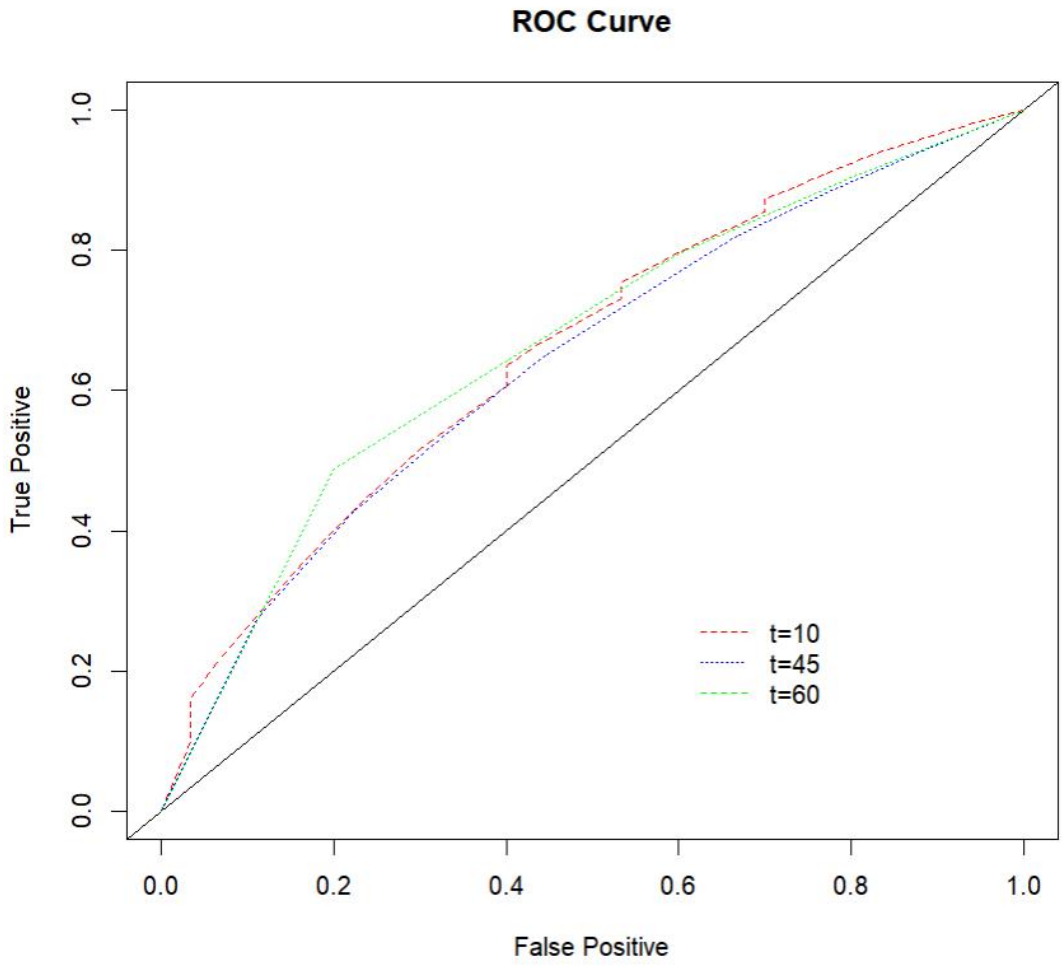


Σχήμα 6.10: *Στρωματοποίηση για την συμμεταβλητή Πρωτεΐνη Bence-Jones.*

6.4 Εφαρμογή της καμπύλης ROC

Στη συνέχεια θα εφαρμόσουμε τις καμπύλες ROC στα δεδομένα μας με την βοήθεια του στατιστικού πακέτου R στο τελικό μοντέλο του Cox το οποίο βρήκαμε με την μέθοδο διαδοχικής αφαίρεσης, έτσι ώστε να ελέγξουμε την αποτελεσματικότητα του.

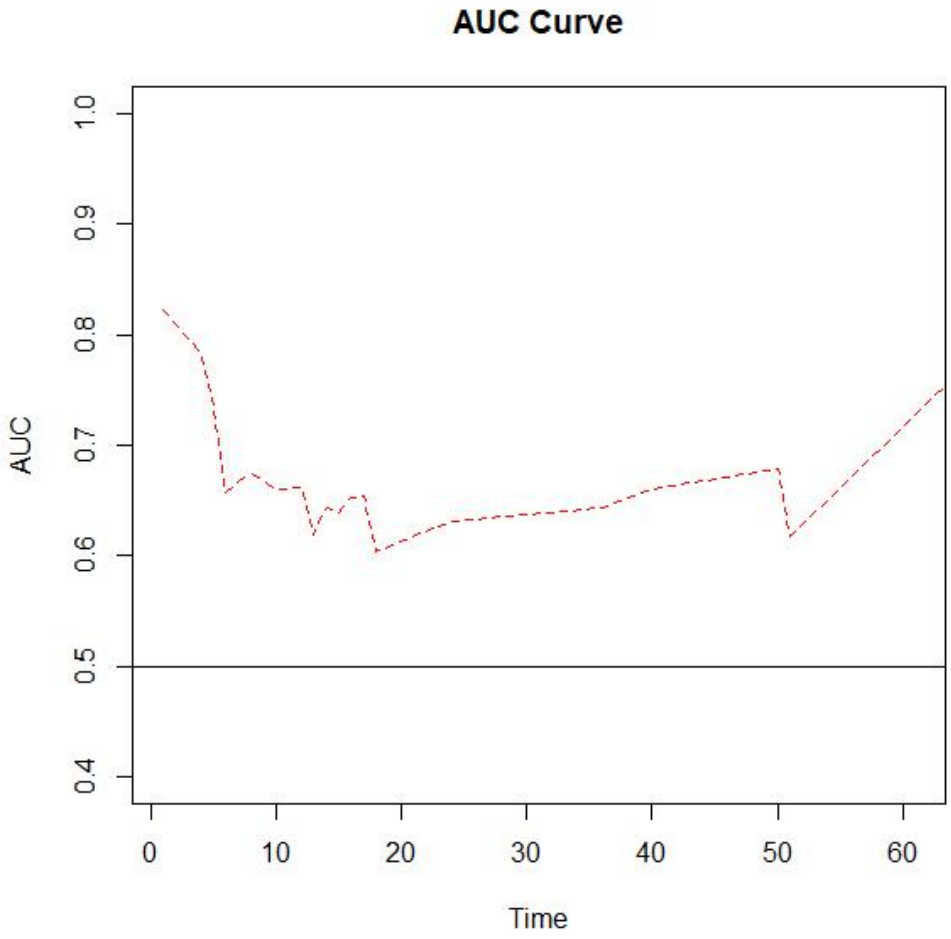
Για να γίνει αυτό θα επιλέξουμε τρεις διαφορετικές τιμές του χρόνου επιβίωσης των ασθενών οι οποίες θα απέχουν μεταξύ τους ώστε να ελέγξουμε την συμπεριφορά του μοντέλου μας για μικρούς και μεγάλους χρόνους. Οι τιμές που θα επιλέξουμε θα είναι $t_1 = 10,\ t_2 = 45$ και $t_3 = 60$ ημέρες.



Σχήμα 6.11 : *Καμπύλη ROC για το τελικό μοντέλο του Cox.*

Παρατηρούμε ότι οι καμπύλες που σχηματίζονται είναι αρκετά κοντά. Ωστόσο η καλύτερη καμπύλη είναι εκείνη η οποία πλησιάζει περισσότερο στην πάνω αριστερή γωνία του διαγράμματος, δηλαδή , η καμπύλη για $t_3 = 60$.

Στην συνέχεια, θα σχεδιάσουμε μια καμπύλη AUC έτσι ώστε να υπολογίσουμε το εμβαδόν που βρίσκεται κάτω από την καμπύλη ROC για μια μέγιστη τιμή του χρόνου. Για αυτή την εφαρμογή θα επιλέξουμε ως μέγιστη τιμή του t την $t_{max} = 60$.



Σχήμα 6.12 : *Καμπύλη AUC για το μοντέλο μετά την διαδικασία διαδοχικής αφαίρεσης.*

Στο διάγραμμα του *Σχήματος 6.12* παρατηρούμε ότι υπάρχουν μεγάλες διαφορές στις τιμές της καμπύλης AUC, καθώς υπάρχουν αρκετά μέγιστα και ελάχιστα σημεία. Αυτό σημαίνει ότι δεν έχουμε τον ίδιο αριθμό παρατηρήσεων σε όλες τις τιμές του χρόνου. Στα σημεία με τις περισσότερες παρατηρήσεις, δηλαδή, στις πολύ μικρές και στις πολύ μεγάλες τιμές του χρόνου επιβίωσης το μοντέλο μας έχει λίγο καλύτερη προβλεπτική ικανότητα καθώς όσες περισσότερες παρατηρήσεις έχουμε σε ένα χρονικό διάστημα τόσο θα αυξάνεται η τιμή του AUC. Οι τιμές των AUC που εκφράζουν το εμβαδόν κάτω από την καμπύλη για κάθε *t* είναι:

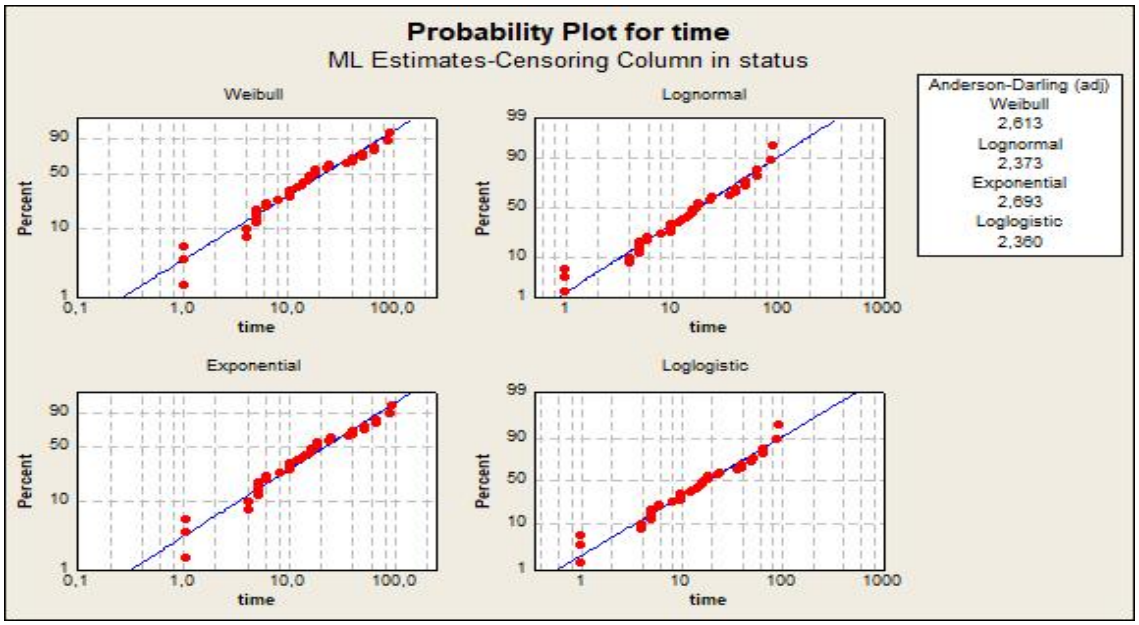
- $AUC_{t=10} = 0.66$
- $AUC_{t=45} = 0.64$
- $AUC_{t=60} = 0.67$

Επομένως, επιβεβαιώνεται και από τις τιμές AUC ότι το μοντέλο του Cox έχει καλύτερη προβλεπτική ικανότητα στις μεγάλες τιμές του χρόνου.

6.5 *Εφαρμογή του Μοντέλου Επιταχυνόμενης Διάρκειας Ζωής*

Σε αυτό το κεφάλαιο θα προσαρμόσουμε το Μοντέλο Επιταχυνόμενης Διάρκειας Ζωής στα δεδομένα μας με την χρήση της κατανομής Weibull, με σκοπό να ελέγξουμε εάν ένας οι περισσότεροι παράγοντες συμβάλλουν στην διάρκεια ζωής των ασθενών. Σε αντίθεση με το μοντέλο του Cox στο οποίο δεν είχαμε κάποια συγκεκριμένη κατανομή για τα δεδομένα μας, σε ένα παραμετρικό μοντέλο θα έχουμε πιο ακριβή αποτελέσματα εφόσον τα δεδομένα μας θα ακολουθούν μια συγκεκριμένη κατανομή.

Αρχικά, θα χρησιμοποιήσουμε ένα γράφημα πιθανότητας με σκοπό να ελέγξουμε την καταλληλότητα των κατανομών μέσω ενός ελέγχου καλής προσαρμογής Anderson-Darling. Μέσω του συγκεκριμένου ελέγχου μπορούμε να δούμε ποια κατανομή ταιριάζει καλύτερα στα δεδομένα μας συγκρίνοντας τα στατιστικά ελέγχου για κάθε κατανομή και επιλέγοντας εκείνο με την μικρότερη τιμή. Ουσιαστικά, η κατανομή με το μικρότερο στατιστικό είναι και εκείνη όπου τα σημεία βρίσκονται πιο κοντά στην προσαρμοσμένη ευθεία.



Σχήμα 6.13: *Διαγράμματα πιθανότητας για τις κατανομές Weibull, Λογαριθμο-Κανονική, Εκθετική και Λογαριθμο-Λογιστική.*

Σύμφωνα με τις γραφικές παραστάσεις των κατανομών στο *Σχήμα 6.13*, η κατανομή που προσαρμόζει καλύτερα τα δεδομένα μας φαίνεται να είναι η Log-Logistic και επιβεβαιώνεται μέσω του ελέγχου Anderson-Darling, αφού έχει στατιστικό ίσο με 2,360 και η αμέσως καλύτερη φαίνεται να είναι η Log-Normal κατανομή με στατιστικό ίσο με 2,373.

6.5.1 Εφαρμογή του Μοντέλου Weibull

Προσαρμόζοντας το μοντέλο μας για την κατανομή Weibull, συμπεριλαμβάνοντας στο μοντέλο όλες τις συμμεταβλητές, μας επιστρέφονται τα παρακάτω αποτελέσματα.

	Value	**Std. Error**	**z**	**p**
(Intercept)	2.28202	2.16337	1.05	0.291
Ηλικία (Age)	0.01218	0.02299	0.53	0.596
Φύλο (Sex)	-0.04220	0.32776	-0.13	0.898
Αιμοσφαιρίνη (Bun)	-0.01729	0.00405	-4.27	2e-05
Ασβέστιο (Ca)	-0.02497	0.11305	-0.22	0.825
Επίπεδο Αζώτου Ουρίας στο Αίμα (Hb)	0.08766	0.05725	1.53	0.126
Πλασματοκύτταρα (Pcells)	0.00101	0.00536	0.19	0.851
Πρωτεΐνη Bence-Jones (Protein)	0.62051	0.33350	1.86	0.063
Log(scale)	-0.17423	0.13059	-1.33	0.182
Scale = 0.84				
Weibull distribution				
Loglik(model) = -151.7		**Loglik(intercept only)** = -159.8		
Chisq = 16.26 **on 7 degrees of freedom , p** = 0.023				
Number of Newton-Raphson Iterations: 6				
N = 48				
AIC = 321.3086				

Πίνακας 6.12 : *Αποτελέσματα από την προσαρμογή του μοντέλου Weibull.*

Από τα αποτελέσματα της εφαρμογής του μοντέλου της κατανομής Weibull που παρουσιάζονται στον *Πίνακα 6.12*, και από τις τιμές των πιθανοτήτων p-value προκύπτει ότι η συμμεταβλητή η οποία είναι στατιστικά σημαντικότερη είναι η Αιμοσφαιρίνη (Bun) με p-value πολύ μικρό. Οι αμέσως επόμενες στατιστικά σημαντικές συμμεταβλητές

προκύπτουν ότι είναι η Πρωτεΐνη Bence-Jones με $\mathrm{p-value} = 0.063$ και το Επίπεδο Αζώτου Ουρίας στο Αίμα με $\mathrm{p-value} = 0.126$.

6.5.2 Εφαρμογή της μεθόδου διαδοχικής αφαίρεσης

Με σκοπό να επιτύχουμε το βέλτιστο μοντέλο θα εφαρμόσουμε την μέθοδο διαδοχικής αφαίρεσης (Backward Elimination) στο μοντέλο του Weibull με την βοήθεια του κριτηρίου AIC. Στον *Πίνακα 6.13* μπορούμε να δούμε τις διαφορετικές, και μικρότερες κάθε φορά, τιμές του AIC για κάθε καινούριο μοντέλο που δημιουργείται με την συμμεταβλητή που αφαιρείται.

Μοντέλο	AIC
Βήμα 1 (Αρχή Διαδικασίας)	321.31
Βήμα 2	319.33
Βήμα 3	317.36
Βήμα 4	315.41
Βήμα 5 (Τέλος διαδικασίας)	313.67

Πίνακας 6.13 : *Τιμές του ελέγχου AIC κατά την διαδικασία διαδοχικής αφαίρεσης.*

	Value	Std. Error	z	p
(Intercept)	2.86294	0.53475	5.35	8.6e-08
Αιμοσφαιρίνη (Bun)	-0.01723	0.00394	-4.37	1.2e-05
Επίπεδο Αζώτου Ουρίας στο Αίμα (Hb)	0.08073	0.05098	1.58	0.113
Πρωτεΐνη Bence Jones (Protein)	0.58341	0.32367	1.86	0.063
Log(scale)	-0.17407	0.12896	-1.35	0.177
Scale = 0.84				
Loglik(model) = -151.8 **Loglik(intercept only)** = -159.8				
Chisq = 15.9 **on** 3 **degrees of freedom,** **p** = 0.00119				
Number of Newton-Raphson Iterations: 5				
N = 48				

Πίνακας 6.14: *Αποτελέσματα για το τελικό μοντέλο Weibull*

Σύμφωνα με τα αποτελέσματα του *Πίνακα 6.14* για το τελευταίο βήμα της διαδικασίας, το βέλτιστο μοντέλο που προκύπτει για τα δεδομένα μας έχει τις συμμεταβλητές Αιμοσφαιρίνη (Bun), Ουρία στο Αίμα (Hb) και την παρουσία Πρωτεΐνης Bence Jones (Protein).

Το τελικό μοντέλο Weibull αποτελείται από τις συμμεταβλητές Αιμοσφαιρίνη, Επίπεδο Αζώτου Ουρίας στο Αίμα και την παρουσία Πρωτεΐνης Bence-Jones στα ούρα, όπως ακριβώς και στον μοντέλο αναλογικής διακινδύνευσης του Cox. Συνεπώς, καταλήγουμε ότι σε όλες τις περιπτώσεις αυτές οι συμμεταβλητές είναι οι πιο στατιστικά σημαντικές για την επιβίωση των ασθενών με πολλαπλούν μυέλωμα.

Συμπεράσματα

Σε αυτή την διπλωματική εργασία, πραγματοποιήθηκε αναλυτική παρουσίαση της θεωρίας των μεθόδων που χρησιμοποιούνται στην ανάλυση επιβίωσης και της εφαρμογής τους σε ιατρικά δεδομένα ασθενών που πάσχουν από Πολλαπλούν Μυέλωμα, για να διαπιστωθεί ποιες μεταβλητές συνέβαλαν σημαντικά στον χρόνο επιβίωσης των ασθενών. Πιο συγκεκριμένα, έγινε χρήση μη παραμετρικών μοντέλων όπως της εκτιμήτριας Kaplan-Meier και Nelson-Aalen, το ημιπαραμετρικό μοντέλο αναλογικής διακινδύνευσης του Cox, καθώς επίσης εφαρμόστηκαν οι μέθοδοι των καμπυλών ROC και τέλος, εφαρμόστηκε το παραμετρικό μοντέλο Weibull. Συνολικά, όλες οι μέθοδοι συμφωνούν σχετικά με το ποιες μεταβλητές επηρεάζουν τη διάρκεια ζωής, ωστόσο, υπάρχουν μικρές διαφορές ανάλογα την μέθοδο που χρησιμοποιήθηκε. Από την εφαρμογή της εκτιμήτριας Kaplan-Meier, μπορεί να φανεί ότι ο χρόνος επιβίωσης ενός ασθενούς δεν επηρεάζεται σημαντικά από κάποια μεταβλητή όπως το φύλο ή η Πρωτεΐνη Bence Jones, επειδή δεν υπάρχει μεγάλη διαφορά στην καμπύλη επιβίωσης. Στη συνέχεια εφαρμόσαμε το μοντέλο αναλογικής διακινδύνευσης του Cox και διαπιστώσαμε ότι ο χρόνος επιβίωσης επηρεάστηκε από τρεις μεταβλητές. Πρώτη μεταβλητή που επηρεάζει σημαντικά είναι το επίπεδο αζώτου ουρίας και ακολουθούν η αιμοσφαιρίνη και η παρουσία Πρωτεΐνης Bence-Jones στα ούρα. Βλέπουμε μια αύξηση 1.02 στη συνάρτηση διακινδύνευσης για τους ασθενείς με αυξημένες τιμές του αζώτου ουρίας σε σύγκριση με εκείνους που έχουν χαμηλότερα επίπεδα. Επιπλέον, κάθε αύξηση μιας μονάδας της τιμής της αιμοσφαιρίνης μειώνει τη συνάρτηση διακινδύνευσης κατά 0.89. Για τους ασθενείς που παρουσιάζουν την πρωτεΐνη Bence-Jones στα ούρα, μειώθηκε κατά 0.51 σε σύγκριση με εκείνους που δεν έχουν. Σε αυτό το μοντέλο, εφαρμόσαμε στρωματοποιημένη ανάλυση στη μεταβλητή που αφορά την παρουσία της Πρωτείνη Bence-Jones στα ούρα, και παρατηρήσαμε ότι δεν υπάρχει σημαντική διαφορά μεταξύ των κατηγοριών, και καταλήγει και το μοντέλο του Cox. Οι περισσότερα στατιστικά σημαντικές συμμεταβλητές φαίνεται να είναι το επίπεδο Αζώτου στο αίμα, η Αιμοσφαιρίνη και η παρουσία της πρωτεΐνης Bence-Jones στα ούρα, κάτι που προέκυψε από όλες τις μεθόδους που χρησιμοποιήσαμε. Επιπλέον, εφαρμόσαμε και την καμπύλη ROC στο τελικό μοντέλο, και ελέγξαμε την προγνωστική του ικανότητα, η οποία ήταν ικανοποιητική. Τέλος, προσαρμόσαμε το παραμετρικό μοντέλο Weibull, αρχικά πραγματοποιώντας κάποιους γραφικούς ελέγχους για να ελέγξουμε την

επάρκεια του και στη συνέχεια εφαρμόζοντας τη διαδικασία διαδοχικής αφαίρεσης μέσω της οποίας καταλήξαμε στο τελικό μοντέλο με συμμεταβλητές και πάλι ίδιες με το μοντέλο αναλογικής διακινδύνευσης του Cox.

ΒΙΒΛΙΟΓΡΑΦΙΑ

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

Akaike, H. (1974). *A new look at the statistical model identification. IEEE Trans. Automat. Control*, **19**, pp. 716-723.

Barlow, W.E. and Prentice, R. L. (1988). *Residuals for relative risk regression. Biometrika*, **75**, pp. 65 – 74.

Beard, R.E. (1959). *Note on some mathematical mortality models. In G. E. W. Wolstenholme & M. O'Connor (Eds.), The lifespan of animals,* Boston, MA: Little Brown, pp. 302-311

Belsley, D.A., Kuh, E. and Welsh, R.E, (2004). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* John Wiley, Hoboken, New Jersey.

Cai, Z. and Sun, Y. (2003). *Local linear estimation for time dependent coefficients in Cox's regression models*. *Scandinavian Journal of Statistics,* **30,** pp. 93–111.

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, pp.27-38

Collett, D. (2014). *Modelling Survival Data in Medical Research.(3rd edition)*. Chapman and Hall, Boca Raton.

Cox, D.R. (1972). *Regression models and life tables (with discussion). Journal of the Royal Statistical Society, Series B,* **24**, pp. 187-220

Cox, D.R. and Snell, J.E. (1968). *A general definition of residuals, Journal of the Royal Statistical Society, Series B,* **30** *(2)*, pp. 248 – 275

Fitrianto, A. and Ting Jiin, R.L. (2013). *Several Types of Residuals in Cox Regression-Model: An Empirical Study, Int. Journal of Math. Analysis*, **53**, pp. 2645 - 2654

Gavrilov, L.A. & Gavrilova, N.S. (2001). *The reliability theory of aging and longevity. Journal of Theoretical Biology*, **213**, pp. 527-545

Gompertz, B. (1825). *The nature of the function expressive of the law of human mortality*. *Philosophical Transactions of the Royal Society,* **115**, pp. 513-585.

Greenwood, M. (1926). *The natural duration of cancer. Reports on Public Health and Medical Subjects* **33**,. Her Majesty's Stationery Office, London. pp.1–26

Hastie, T. and Tibshirani, R. (1993). *Varying-coefficient models. Journal of the Royal Statistical Society*, Series B, **55**, pp. 757–796

Heagerty, P.J. and Zheng, Y. (2005). *Survival model predictive accuracy and ROC curves. Biometrics,* **61**, pp. 92-105

Kaplan, E.L. and Meier, P. (1958). *Nonparametric Estimation from incomplete observations*, *Journal of the J.Amer.Stat. Assoc.,* **53**, pp. 457-481.

Kurniasari, D. Widyarini, R., Antonio, Y. (2019). *Characteritics of Hazard Rate Functions of Log-Normal Distributions, Journal of Physics: Conference Series,* **1338**, Article ID: 012036. https://doi.org/10.1088/1742-6596/1338/1/012036

Makeham, W. M. (1859). *On the law of mortality and the construction of annuity tables. Journal of the Institute of Actuaries,* **8**, pp. 301-310.

Nelson, W. (1972). *Theory and applications of hazard plotting for censored failure data. Technometrics,* **14** *(4),* pp. 945–965.

Perks, W. (1932). *On some experiments in the graduation of mortality statistics*. *Journal of the Institute of Actuaries,* **63***,* pp. 12-40.

Richards, S. J. (2012). *A handbook of parametric survival models for actuarial use. Scandinavian Actuarial Journal,* **4***,* pp. 233-257.

Schwarz, G. (1978). *Estimating the dimension of a model. Annals of Statistics,* **6**, pp. 416-464

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

Therneau, T.M., Grambsch, P.M., and Fleming, T.R. (1990). *Martingale-based residuals for survival models*, *Biometrika,* **77***,* pp. 147 – 160.

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

Καρώνη Χ. και Οικονόμου Π. (2010). *Στατιστικά Μοντέλα Παλινδρόμησης με χρήση Minitab και R*, Εκδόσεις Συμεών, Αθήνα.

Παράρτημα - Αποτελέσματα Minitab και R:

A. Αποτελέσματα Minitab για Kaplan-Meier

Nonparametric Survival Plot for time

Distribution Analysis: time

Variable: time

Censoring Information Count Uncensored value 36 Right censored value 12

Censoring value: status = 0

Nonparametric Estimates

Characteristics of Variable

Standard	95,0% Normal CI			
Mean(MTTF)	Error	Lower	Upper	
31,3455	4,79105	21,9552	40,7358	

Median = 17 IQR = 43 Q1 = 8 Q3 = 51

Kaplan-Meier Estimates

Number at Risk	Number Failed	Survival Probability	Standard Error	95,0% Normal CI Lower	95,0% Normal CI Upper
1	48	3	0,937500	0,0349386	0,869022	1,00000
4	44	2	0,894886	0,0444853	0,807697	0,98208
5	42	4	0,809659	0,0571220	0,697702	0,92162
6	38	2	0,767045	0,0615522	0,646405	0,88769
8	35	1	0,745130	0,0635755	0,620524	0,86974
10	34	4	0,657468	0,0695839	0,521086	0,79385
12	28	1	0,633987	0,0709500	0,494927	0,77305
13	26	1	0,609602	0,0722900	0,467917	0,75129
14	25	1	0,585218	0,0733958	0,441365	0,72907
15	24	1	0,560834	0,0742778	0,415252	0,70642
16	22	2	0,509849	0,0757709	0,361341	0,65836
17	20	1	0,484357	0,0761501	0,335105	0,63361
18	19	2	0,433372	0,0761919	0,284039	0,58271
23	15	1	0,404480	0,0763940	0,254751	0,55421
24	14	1	0,375589	0,0762049	0,226230	0,52495
36	13	1	0,346698	0,0756217	0,198482	0,49491
40	12	2	0,288915	0,0732289	0,145389	0,43244
50	9	1	0,256813	0,0717846	0,116118	0,39751
51	8	1	0,224711	0,0696203	0,088258	0,36116
65	5	1	0,179769	0,0686871	0,045145	0,31439
66	4	1	0,134827	0,0645654	0,008281	0,26137
88	2	1	0,067413	0,0575713	0,000000	0,18025
91	1	1	0,000000	0,0000000	0,000000	0,00000

Distribution Analysis: time by sex

Variable: timesex = 1

Censoring Information Count Uncensored value 22 Right censored value 7

Censoring value: status = 0

Nonparametric Estimates

Characteristics of Variable

Standard Error	95,0% Normal CI Lower	Mean(MTTF)	Upper
32,8322	6,16317	20,7526	44,9118

Median = 16 IQR = 55 Q1 = 10 Q3 = 65

Kaplan-Meier Estimates

Number at Risk	Number Failed	Survival Probability	Standard Error	95,0% Normal CI Lower	Upper
1	29	3	0,896552	0,0565523	0,785711	1,00000
4	25	1	0,860690	0,0646689	0,733941	0,98744
5	24	2	0,788966	0,0766285	0,638776	0,93915
8	22	1	0,753103	0,0811041	0,594142	0,91206
10	21	3	0,645517	0,0902209	0,468688	0,82235
13	17	1	0,607546	0,0925601	0,426131	0,78896
14	16	1	0,569574	0,0942425	0,384862	0,75429
16	14	2	0,488206	0,0967613	0,298558	0,67785
18	12	1	0,447522	0,0968739	0,257653	0,63739
24	11	1	0,406839	0,0962316	0,218228	0,59545
36	10	1	0,366155	0,0948192	0,180312	0,55200
40	9	1	0,325471	0,0926014	0,143975	0,50697
50	7	1	0,278975	0,0902942	0,102002	0,45595
65	4	1	0,209231	0,0907427	0,031379	0,38708
66	3	1	0,139488	0,0830810	0,000000	0,30232
88	1	1	0,000000	0,0000000	0,000000	0,00000

Distribution Analysis: time by sex

Variable: timesex = 2

Censoring Information Count Uncensored value 14 Right censored value 5

Censoring value: status = 0

Nonparametric Estimates

Characteristics of Variable

Standard Error	95,0% Normal CI Lower	Mean(MTTF)	Upper
27,2386	7,47733	12,5833	41,8939

Median = 17 IQR = 34 Q1 = 6 Q3 = 40

Kaplan-Meier Estimates

Number at Risk	Number Failed	Survival Probability	Standard Error	95,0% Normal CI Lower	Upper
4	19	1	0,947368	0,051228	0,846964	1,00000
5	18	2	0,842105	0,083655	0,678145	1,00000
6	16	2	0,736842	0,101023	0,538841	0,93484
10	13	1	0,680162	0,107988	0,468510	0,89181

12	11	1	0,618329	0,114513	0,393888	0,84277
15	9	1	0,549626	0,120651	0,313153	0,78610
17	8	1	0,480923	0,123593	0,238686	0,72316
18	7	1	0,412219	0,123565	0,170036	0,65440
23	4	1	0,309165	0,128661	0,056993	0,56134
40	3	1	0,206110	0,120156	0,000000	0,44161
51	2	1	0,103055	0,094443	0,000000	0,28816
91	1	1	0,000000	0,000000	0,000000	0,00000

Distribution Analysis: time by sex

Comparison of Survival Curves

Test Statistics

Method Chi-Square DF P-ValueLog-Rank 0,0358454 1 0,850Wilcoxon 0,0001360 1 0,991

Distribution Analysis: time by protein

Variable: timeprotein = 0

Censoring Information CountUncensored value 24Right censored value 9

Censoring value: status = 0

Nonparametric Estimates

Characteristics of Variable

Standard 95,0% Normal CIMean(MTTF) Error Lower Upper
23,4749 4,37638 14,8973 32,0524

Median = 15IQR = 16 Q1 = 8 Q3 = 24Kaplan-Meier Estimates

Number at Risk	Number Failed	Survival Probability	Standard Error	95,0% Normal CI Lower	Upper
1	33	2	0,939394	0,0415360	0,857985	1,00000
4	31	1	0,909091	0,0500438	0,811007	1,00000
5	30	4	0,787879	0,0711647	0,648398	0,92736
6	26	1	0,757576	0,0746009	0,611361	0,90379
8	24	1	0,726010	0,0778849	0,573359	0,87866
10	23	4	0,599747	0,0862092	0,430781	0,76871
12	18	1	0,566428	0,0876223	0,394692	0,73816
14	16	1	0,531026	0,0890107	0,356569	0,70548
15	15	1	0,495625	0,0898414	0,319539	0,67171
16	13	2	0,419375	0,0907677	0,241473	0,59728
17	11	1	0,381250	0,0901681	0,204524	0,55798
18	10	1	0,343125	0,0888464	0,168989	0,51726
23	7	1	0,294107	0,0886507	0,120355	0,46786
24	6	1	0,245089	0,0863706	0,075806	0,41437
40	5	1	0,196071	0,0818323	0,035683	0,35646
65	2	1	0,098036	0,0804961	0,000000	0,25581

Distribution Analysis: time by protein

Variable: timeprotein = 1

Censoring Information CountUncensored value 12Right censored value 3

Censoring value: status = 0

Nonparametric Estimates

Characteristics of Variable

Standard Error	95,0% Normal CI		Mean(MTTF)	Lower	Upper
42,0250	8,83509	24,7086	59,3415		

Median = 40IQR = 53 Q1 = 13 Q3 = 66

Kaplan-Meier Estimates

Time	Number at Risk	Number Failed	Survival Probability	Standard Error	95,0% Normal CI Lower	Upper
1	15	1	0,933333	0,064406	0,807100	1,00000
4	13	1	0,861538	0,091063	0,683058	1,00000
6	12	1	0,789744	0,108134	0,577805	1,00000
13	10	1	0,710769	0,122819	0,470048	0,95149
18	9	1	0,631795	0,132146	0,372793	0,89080
36	8	1	0,552821	0,137212	0,283890	0,82175
40	7	1	0,473846	0,138485	0,202420	0,74527
50	5	1	0,379077	0,139495	0,105671	0,65248
51	4	1	0,284308	0,132972	0,023687	0,54493
66	3	1	0,189538	0,117669	0,000000	0,42017
88	2	1	0,094769	0,089175	0,000000	0,26955
91	1	1	0,000000	0,000000	0,000000	0,00000

Distribution Analysis: time by protein

Comparison of Survival Curves

Test Statistics

Method	Chi-Square	DF	P-Value
Log-Rank	2,04590	1	0,153
Wilcoxon	1,90630	1	0,167

Β. Αποτελέσματα και Κώδικας σε R για το Μοντέλο Αναλογικής Διακινδύνευσης του Cox

```
> library(splines)
> library(survival)
> library(MASS)
> library(risksetROC)
> library(parmsurvfit)
> data<-read.csv("datasurvival.csv",header=TRUE)
> data<-as.data.frame(data)
> attach(data)
> summary(age)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
50.00   58.75   62.50   62.90   68.25   77.00
> summary(bun)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
6.00   13.75   21.00   33.92   39.25  172.00
> summary(ca)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
8.000   9.000  10.000   9.938  10.000  15.000
> summary(hb)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
4.90    8.65   10.20   10.25   12.57   14.60
```

```
> levels(sex)<-c("Male","Female")
> sex2<-table(sex)
> prop.table(sex2)
sex
1         2
0.6041667 0.3958333
> protein2<-table(protein)
> prop.table(protein2)
protein
0      1
0.6875 0.3125
> levels(status)<-c("Alive","Dead")
> status2<-table(status)
> prop.table(status2)
status
0    1
0.25 0.75
> summary(time)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.00    6.75   14.50   23.38   37.00   91.00
>
> #Προσαρμογή του Cox Model
> mod1<-coxph(Surv(time,status)~age+sex+bun+ca+hb+pcells+protein)
> mod1
Call:
  coxph(formula = Surv(time, status) ~ age + sex + bun + ca + hb +
          pcells + protein)

              coef exp(coef)  se(coef)      z        p
age      -0.018056  0.982106  0.027833 -0.649 0.516521
sex      -0.249473  0.779211  0.403093 -0.619 0.535985
bun       0.022661  1.022919  0.006110  3.709 0.000208
ca        0.013265  1.013353  0.132681  0.100 0.920363
hb       -0.133017  0.875450  0.068527 -1.941 0.052249
pcells   -0.001359  0.998642  0.006588 -0.206 0.836585
protein  -0.683269  0.504964  0.429395 -1.591 0.111556

Likelihood ratio test=17.53  on 7 df, p=0.01428
n= 48, number of events= 36
> #Διαδοχική αφαίρεση στο Cox Model
> mod2<-step(mod1, direction="backward")
Start:  AIC=211.15
Surv(time, status) ~ age + sex + bun + ca + hb + pcells + protein

          Df    AIC
- ca       1 209.16
- pcells   1 209.19
- sex      1 209.54
- age      1 209.56
<none>       211.15
- protein  1 211.80
- hb       1 212.90
- bun      1 219.69

Step:  AIC=209.16
Surv(time, status) ~ age + sex + bun + hb + pcells + protein

          Df    AIC
- pcells   1 207.20
- age      1 207.57
- sex      1 207.60
<none>       209.16
- protein  1 209.81
- hb       1 210.96
- bun      1 217.78
```

```
Step:  AIC=207.2
Surv(time, status) ~ age + sex + bun + hb + protein

          Df    AIC
- age      1 205.57
- sex      1 205.60
<none>       207.20
- protein  1 207.90
- hb       1 208.96
- bun      1 215.79

Step:  AIC=205.57
Surv(time, status) ~ sex + bun + hb + protein

          Df    AIC
- sex      1 203.90
<none>       205.57
- protein  1 206.31
- hb       1 206.99
- bun      1 213.88

Step:  AIC=203.9
Surv(time, status) ~ bun + hb + protein

Df    AIC
<none>       203.90
- protein  1 204.70
- hb       1 204.99
- bun      1 212.35
> mod3<-coxph(Surv(time,status)~bun+hb+protein)
> summary(mod3)
Call:
  coxph(formula = Surv(time, status) ~ bun + hb + protein)

n= 48, number of events= 36

              coef exp(coef)  se(coef)      z Pr(>|z|)
bun      0.022042  1.022287  0.005926  3.719   0.0002 ***
hb      -0.109409  0.896364  0.061891 -1.768   0.0771 .
protein -0.663097  0.515253  0.407855 -1.626   0.1040
---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        exp(coef) exp(-coef) lower .95 upper .95
bun        1.0223     0.9782    1.0105     1.034
hb         0.8964     1.1156    0.7940     1.012
protein    0.5153     1.9408    0.2317     1.146

Concordance= 0.696  (se = 0.051 )
Likelihood ratio test= 16.78  on 3 df,   p=8e-04
Wald test            = 19.25  on 3 df,   p=2e-04
Score (logrank) test = 23.88  on 3 df,   p=3e-05

> #Ελεγχος για την υπόθεση αναλογικότητας
> cox.zph(mod3)
          chisq df    p
bun      0.7182  1 0.40
hb       0.0115  1 0.91
protein 0.2501  1 0.62
GLOBAL   1.3487  3 0.72
> #Υπόλοιπα Schoenfeld
> par(mfrow=c(2,2))
> plot(cox.zph(mod3))
```

```
> cox.zph(mod1,transform="identity")
          chisq df    p
age      0.3623  1 0.55
sex      0.0136  1 0.91
bun      0.5061  1 0.48
ca       1.3454  1 0.25
hb       0.0520  1 0.82
pcells   1.4660  1 0.23
protein  1.2779  1 0.26
GLOBAL   8.2571  7 0.31
> cox.zph(mod2,transform = "identity")
         chisq df    p
bun      0.440  1 0.51
hb       0.103  1 0.75
protein 0.954  1 0.33
GLOBAL   2.947  3 0.40
> sresid<-resid(mod3,type="schoenfeld")
> sresid2<-resid(mod3,type="scaledsch")
> #df beta residuals
> par(mfrow=c(2,2))
> dfbetasCox<-residuals(mod3, type="dfbeta")
> for(j in 1:3){
  +   plot(dfbetasCox[,j],xlab="Time",ylab=names(coef(mod3))[j])
  +   abline(h=0,lty=3)
  + }
> n<-nrow(data)
> threshold<-2/sqrt(n)
> threshold
[1] 0.2886751
> #Martingale residuals
> abline(h=0,lty=2)
> par(mfrow=c(1,1))
> res<-residuals(mod3,type=("martingale"))
> X<-as.matrix(data[,c("bun")])
> Y<-as.matrix(data[,c("hb")])
> par(mfrow=c(1,1))
> for (j in 1:1){
  +   plot(X[,j],res,xlab="Bun"[j],ylab="Residuals")
  +   abline(h=0,lty=1)
  +   lines(lowess(X[,j],res,iter=0))
  + }
> for (j in 1:1){
  +   plot(Y[,j],res,xlab="Hb"[j],ylab="Residuals")
  +   abline(h=0,lty=1)
  +   lines(lowess(X[,j],res,iter=0))
  + }

> #Component residual
> b<-coef(mod3)[c(3)]
> for (i in 1:1){
  +   plot(X[,j],b[j]*X[,j]+res,xlab="Bun"[j],ylab="component+residual")
  +   abline(lm(b[j]*X[,j]+res ~X[,j]),lty=2)
  +   lines(lowess(X[,j],b[j]*X[,j]+res,iter=0))
  +
  + }
> for (i in 1:1){
  +   plot(Y[,j],b[j]*Y[,j]+res,xlab="Hb"[j],ylab="component+residual")
  +   abline(lm(b[j]*Y[,j]+res ~Y[,j]),lty=2)
  +   lines(lowess(Y[,j],b[j]*Y[,j]+res,iter=0))
  +
  + }
> # Στρωματοποιηση
> library(lattice)
> library(risksetROC)
> modStr<-coxph(Surv(time,status)~age+sex+bun+ca+hb+pcells+strata(protein),
                data=data,method="breslow")
```

```
> bh1<-basehaz(modStr,centered=TRUE)
> modStr
Call:
  coxph(formula = Surv(time, status) ~ age + sex + bun + ca + hb +
          pcells + strata(protein), data = data, method = "breslow")

              coef  exp(coef)   se(coef)      z       p
age     -0.0113021  0.9887615  0.0288829 -0.391 0.69557
sex     -0.3828102  0.6819423  0.4081326 -0.938 0.34827
bun      0.0195414  1.0197336  0.0060513  3.229 0.00124
ca       0.0083978  1.0084332  0.1333883  0.063 0.94980
hb      -0.1346668  0.8740071  0.0697803 -1.930 0.05362
pcells -0.0007812  0.9992191  0.0067613 -0.116 0.90802

Likelihood ratio test=13.27  on 6 df, p=0.03896
n= 48, number of events= 36
> Inhazard<-log(bh1[,1])
> Intime<-log(bh1[,2])
> xyplot(Inhazard~Intime,group=strata,auto.key=TRUE,data=bh1)
> modfinal<-coxph(Surv(time,status)~bun+hb+strata(protein),data=data,
                  method="breslow")
> bh2<-basehaz(modfinal,centered=TRUE)
> modfinal
Call:
  coxph(formula = Surv(time, status) ~ bun + hb + strata(protein),
        data = data, method = "breslow")

          coef exp(coef)  se(coef)      z       p
bun   0.018816  1.018994  0.005802  3.243 0.00118
hb   -0.104944  0.900375  0.062088 -1.690 0.09098

Likelihood ratio test=12.21  on 2 df, p=0.002233
n= 48, number of events= 36
> Inhazard<-log(bh2[,1])
> Intime<-log(bh2[,2])
> xyplot(Inhazard~Intime,group=strata,auto.key=TRUE,data=bh2)
> #Για την καμπύλη ROC:
> a<-mod3$linear.predictor
> ROC10=ROC10=risksetROC(Stime=time,status=status,marker=a,predict.time=10,
                         method="Cox",main="ROC Curve",lty=2,col="red",
                         ylab="True Positive",xlab="False Positive")
> ROC45=ROC45=risksetROC(Stime=time,status=status,marker=a,predict.time=45,
                         method="Cox",plot=FALSE)
> lines(ROC45$FP,ROC45$TP,lty=3,col="blue")
> ROC60=ROC60=risksetROC(Stime=time,status=status,marker=a,predict.time=60,
                         method="Cox",plot=FALSE)
> lines(ROC60$FP,ROC60$TP,lty=3,col="green")
> legend(.6,.30,lty=c(2,3),col=c("red","blue","green"),
         legend=c("t=10","t=45","t=60"),bty="n")
> ROC10
$marker
[1] -1.357865979 -1.335341973 -1.226575204 -1.184577689 -1.173636791
[6] -0.936914125 -0.917019015 -0.897384953 -0.893150534 -0.871108186
[11] -0.782457135 -0.762180864 -0.739656858 -0.712834661 -0.652290228
[16] -0.629284565 -0.610834426 -0.596782976 -0.585360421 -0.433050669
[21] -0.411751026 -0.345302877 -0.344660667 -0.315911840 -0.125842709
[26] -0.095428302 -0.037673317  0.008338009  0.096025744  0.135614973
[31]  0.170685401  0.269153482  0.369066534  0.851911013

$TP
[1] 1.00000000 0.98919364 0.97814111 0.96581863 0.95296761 0.93997521
[7] 0.92351268 0.90671935 0.88959304 0.87239405 0.85481175 0.83559958
[13] 0.81599389 0.79594159 0.77534416 0.75346114 0.73106886 0.70825960
[19] 0.68512758 0.66172981 0.63448263 0.60664886 0.57690276 0.54713755
[25] 0.51650420 0.47945828 0.44126832 0.40080777 0.35844207 0.31219368
[31] 0.26407762 0.21424418 0.15925401 0.09848576 0.00000000 0.00000000
```

```
$FP
[1] 1.00000000 0.96666667 0.93333333 0.90000000 0.86666667 0.83333333
[7] 0.80000000 0.76666667 0.73333333 0.70000000 0.70000000 0.66666667
[13] 0.63333333 0.60000000 0.56666667 0.53333333 0.53333333 0.50000000
[19] 0.46666667 0.43333333 0.40000000 0.40000000 0.36666667 0.33333333
[25] 0.30000000 0.26666667 0.23333333 0.20000000 0.16666667 0.13333333
[31] 0.10000000 0.06666667 0.03333333 0.03333333 0.00000000 0.00000000

$AUC
[1] 0.6596752
```

```
> ROC45
$marker
[1] -1.3578660 -1.2265752 -1.1736368 -0.8973850 -0.8931505 -0.6522902
[7] -0.5967830 -0.3159118  0.2691535

$TP
[1] 1.0000000 0.9462598 0.8849798 0.8203684 0.7351989 0.6496680
[7] 0.5408434 0.4258074 0.2734672 0.0000000 0.0000000

$FP
[1] 1.0000000 0.8888889 0.7777778 0.6666667 0.5555556 0.4444444
[7] 0.3333333 0.2222222 0.1111111 0.0000000 0.0000000

$AUC
[1] 0.6418436

> ROC60
$marker
[1] -1.3578660 -1.2265752 -0.8973850 -0.8931505  0.2691535

$TP
[1] 1.0000000 0.9038959 0.7943084 0.6419992 0.4890436 0.0000000 0.0000000

$FP
[1] 1.0 0.8 0.6 0.4 0.2 0.0 0.0

$AUC
[1] 0.6658494

> #Για το γράφημα AUC:
> risksetAUC(Stime=time,status=status,marker=a,method="Cox",tmax=60,
             main="AUC Curve",lty=2,col="red")
$utimes
[1]  1  4  5  6  8 10 12 13 14 15 16 17 18 23 24 36 40 50 51 65 66 88 91

$St
[1] 0.93750000 0.89488636 0.80965909 0.76704545 0.74512987 0.65746753
[7] 0.63398655 0.60960245 0.58521835 0.56083426 0.50984932 0.48435686
[13] 0.43337192 0.40448046 0.37558900 0.34669754 0.28891462 0.25681299
[19] 0.22471137 0.17976909 0.13482682 0.06741341 0.00000000

$AUC
[1] 0.8220968 0.7837784 0.7332062 0.6567409 0.6750128 0.6596752
[7] 0.6622700 0.6190686 0.6438591 0.6388768 0.6526107 0.6542534
[13] 0.6040499 0.6260597 0.6312026 0.6431438 0.6603232 0.6783694
[19] 0.6168648 0.7711813 0.5013424 0.3065641 0.0000000

$Cindex
[1] 0.6946294
```

C. Αποτελέσματα και Κώδικας σε R για εφαρμογή του Μοντέλου Weibull

```
> #Προσαρμογή του μοντέλο Weibull:
> modWeib<-survreg(Surv(time,status)~age+sex+bun+ca+hb+pcells+protein,
                   data=data,dist="weibull")
> modWeib
Call:
  survreg(formula = Surv(time, status) ~ age + sex + bun + ca +
           hb + pcells + protein, data = data, dist = "weibull")

Coefficients:
(Intercept)          age          sex          bun           ca
2.282021334  0.012183696 -0.042200093 -0.017286803 -0.024967818
hb       pcells      protein
0.087661382  0.001007144  0.620508851

Scale= 0.8401031

Loglik(model)= -151.7   Loglik(intercept only)= -159.8
Chisq= 16.26 on 7 degrees of freedom, p= 0.0229
n= 48

> summary(modWeib)

Call:
  survreg(formula = Surv(time, status) ~ age + sex + bun + ca +
           hb + pcells + protein, data = data, dist = "weibull")
              Value Std. Error     z     p
(Intercept)  2.28202    2.16337  1.05 0.291
age          0.01218    0.02299  0.53 0.596
sex         -0.04220    0.32776 -0.13 0.898
bun         -0.01729    0.00405 -4.27 2e-05
ca          -0.02497    0.11305 -0.22 0.825
hb           0.08766    0.05725  1.53 0.126
pcells       0.00101    0.00536  0.19 0.851
protein      0.62051    0.33350  1.86 0.063
Log(scale)  -0.17423    0.13059 -1.33 0.182

Scale= 0.84

Weibull distribution
Loglik(model)= -151.7   Loglik(intercept only)= -159.8
Chisq= 16.26 on 7 degrees of freedom, p= 0.023
Number of Newton-Raphson Iterations: 6
n= 48
Step:  AIC=319.33
Surv(time, status) ~ age + bun + ca + hb + pcells + protein

          Df    AIC    LRT Pr(>Chi)
- ca       1 317.36 0.0390 0.843439
- pcells   1 317.37 0.0485 0.825633
- age      1 317.63 0.3085 0.578590
<none>       319.33
- hb       1 320.10 2.7704 0.096022 .
- protein  1 320.58 3.2590 0.071032 .
- bun      1 327.25 9.9292 0.001627 **
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=317.36
Surv(time, status) ~ age + bun + hb + pcells + protein
```

```
          Df    AIC    LRT Pr(>Chi)
- pcells   1 315.41 0.0420 0.837591
- age      1 315.65 0.2868 0.592263
<none>       317.36
- hb       1 318.17 2.8059 0.093920 .
- protein  1 318.58 3.2200 0.072744 .
- bun      1 325.30 9.9316 0.001625 **
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=315.41
Surv(time, status) ~ age + bun + hb + protein

          Df    AIC    LRT Pr(>Chi)
- age      1 313.67 0.2630 0.608077
<none>       315.41
- hb       1 316.19 2.7877 0.094992 .
- protein  1 316.75 3.3424 0.067516 .
- bun      1 323.31 9.9010 0.001652 **
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Step:  AIC=313.67
Surv(time, status) ~ bun + hb + protein

          Df    AIC    LRT Pr(>Chi)
<none>       313.67
- hb       1 314.23 2.5584 0.109711
- protein  1 315.00 3.3272 0.068142 .
- bun      1 321.49 9.8216 0.001725 **
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Call:
  survreg(formula = Surv(time, status) ~ bun + hb + protein, data = data,
          dist = "weibull")

Coefficients:
(Intercept)         bun          hb     protein
2.86294303 -0.01722816  0.08073020  0.58341289

Scale= 0.8402385

Loglik(model)= -151.8   Loglik(intercept only)= -159.8
Chisq= 15.9 on 3 degrees of freedom, p= 0.00119
n= 48

> #Διαδικασία διαδοχικής αφαίρεσης:
> modWFinal<-step(modWeib, direction="backward", test="Chisq")
  Start:  AIC=321.31
  Surv(time, status) ~ age + sex + bun + ca + hb + pcells + protein

            Df    AIC    LRT Pr(>Chi)
  - sex      1 319.33 0.0165 0.897909
  - pcells   1 319.34 0.0354 0.850659
  - ca       1 319.36 0.0483 0.826022
  - age      1 319.59 0.2831 0.594666
  <none>       321.31
  - hb       1 321.77 2.4644 0.116455
  - protein  1 322.58 3.2745 0.070366 .
  - bun      1 329.25 9.9442 0.001614 **
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Step:  AIC=319.33
  Surv(time, status) ~ age + bun + ca + hb + pcells + protein
```

```
           Df    AIC    LRT Pr(>Chi)
- ca        1 317.36 0.0390 0.843439
- pcells    1 317.37 0.0485 0.825633
- age       1 317.63 0.3085 0.578590
<none>        319.33
- hb        1 320.10 2.7704 0.096022 .
- protein   1 320.58 3.2590 0.071032 .
- bun       1 327.25 9.9292 0.001627 **
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=317.36
Surv(time, status) ~ age + bun + hb + pcells + protein

           Df    AIC    LRT Pr(>Chi)
- pcells    1 315.41 0.0420 0.837591
- age       1 315.65 0.2868 0.592263
<none>        317.36
- hb        1 318.17 2.8059 0.093920 .
- protein   1 318.58 3.2200 0.072744 .
- bun       1 325.30 9.9316 0.001625 **
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=315.41
Surv(time, status) ~ age + bun + hb + protein

           Df    AIC    LRT Pr(>Chi)
- age       1 313.67 0.2630 0.608077
<none>        315.41
- hb        1 316.19 2.7877 0.094992 .
- protein   1 316.75 3.3424 0.067516 .
- bun       1 323.31 9.9010 0.001652 **
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Step:  AIC=313.67
Surv(time, status) ~ bun + hb + protein

           Df    AIC    LRT Pr(>Chi)
<none>        313.67
- hb        1 314.23 2.5584 0.109711
- protein   1 315.00 3.3272 0.068142 .
- bun       1 321.49 9.8216 0.001725 **
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> modWFinal
Call:
  survreg(formula = Surv(time, status) ~ bun + hb + protein, data = data,
          dist = "weibull")

Coefficients:
(Intercept)         bun          hb     protein
 2.86294303 -0.01722816  0.08073020  0.58341289

Scale= 0.8402385

Loglik(model)= -151.8   Loglik(intercept only)= -159.8
Chisq= 15.9 on 3 degrees of freedom, p= 0.00119
n= 48
> summary(modWFinal)

Call:
  survreg(formula = Surv(time, status) ~ bun + hb + protein, data = data,
          dist = "weibull")
              Value Std. Error     z       p
(Intercept)  2.86294    0.53475  5.35 8.6e-08
bun         -0.01723    0.00394 -4.37 1.2e-05
hb           0.08073    0.05098  1.58   0.113
protein      0.58341    0.31367  1.86   0.063
Log(scale)  -0.17407    0.12896 -1.35   0.177
```

```
Scale= 0.84

Weibull distribution
Loglik(model)= -151.8   Loglik(intercept only)= -159.8
Chisq= 15.9 on 3 degrees of freedom, p= 0.0012
Number of Newton-Raphson Iterations: 5
n= 48
```