ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Μεταπτυχιακή Εργασία

«ΜΟΝΤΕΛΑ ΕΥΠΑΘΕΙΑΣ ΜΕ ΕΦΑΡΜΟΓΗ ΣΕ ΔΕΔΟΜΕΝΑ ΣΧΕΤΙΚΑ ΜΕ ΤΟΝ ΚΑΡΚΙΝΟ»

ΠΑΝΑΓΙΩΤΟΠΟΥΛΟΥ ΜΑΡΙΑ – ΕΛΕΥΘΕΡΙΑ

Τριμελής Επιτροπή: Βόντα Φιλία, Αναπληρώτρια Καθηγήτρια (Επιβλέπουσα)

Κουκουβίνος Χρήστος, Καθηγητής

Καρώνη Χρυσηίς, Καθηγήτρια

Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών
«Εφαρμοσμένες Μαθηματικές Επιστήμες
με Ειδίκευση στη Στατιστική»

Αθήνα, Μάρτιος 2017

Περιεχόμενα

ΠΕΡΙΛΗΨΗ

Η Ανάλυση Επιβίωσης αποτελεί έναν κλάδο της Στατιστικής που ασχολείται αποκλειστικά με τη μελέτη και ανάλυση δεδομένων διάρκειας ζωής. Για την ανάλυση τέτοιου είδους δεδομένων έχουν αναπτυχθεί και εφαρμοστεί πολλές στατιστικές μέθοδοι οι οποίες συνεχώς εξελίσσονται τα τελευταία χρόνια.

Στην παρούσα εργασία γίνεται αρχικά αναφορά στις βασικές έννοιες της Ανάλυσης Επιβίωσης, καθώς και σε κάποιες βασικές κατανομές, που χρησιμοποιούνται για να περιγράψουν δεδομένα χρόνων επιβίωσης. Αναφέρονται οι βασικές ιδιότητες των εν λόγω κατανομών και απεικονίζονται γραφικά οι συναρτήσεις επιβίωσης και κινδύνου για κάθε μία από αυτές. Εν συνεχεία, αναφέρεται η θεωρητική προσέγγιση των μη παραμετρικών μεθόδων εκτίμησης των συναρτήσεων επιβίωσης και κινδύνου, Kaplan-Meier και Nelson-Aalen, αντίστοιχα. Επίσης, αναλύεται η μεθοδολογία του μοντέλου αναλόγων κινδύνων του Cox, καθώς και κάποιες βασικές τεχνικές για τον έλεγχο της προσαρμογής του μοντέλου. Επίσης, παρουσιάζεται μία εναλλακτική επιλογή μοντέλου για την ανάλυση δεδομένων επιβίωσης, το μοντέλο επιταχυνόμενου χρόνου αποτυχίας. Επιπλέον, παρουσιάζονται αναλυτικά τα μοντέλα ευπάθειας, τα οποία αποτελούν μία γενίκευση του μοντέλου αναλόγων κινδύνων του Cox, το οποίο είναι το πιο ευρέως διαδεδομένο μοντέλο παλινδρόμησης στην ανάλυση επιβίωσης, εισάγοντας έναν παράγοντα των τυχαίων επιδράσεων (*random effects*) ως μία τυχαία μεταβλητή για τον χειρισμό της αλληλεπίδρασης μεταξύ των επεξηγηματικών μεταβλητών, που δεν έχει ληφθεί υπόψη στο μοντέλο, αλλά και της πιθανής μη παρατηρούμενης ανομοιογένειας που χαρακτηρίζει τον πληθυσμό που εξετάζουμε. Αναλύονται παραμετρικά, αλλά και ημιπαραμετρικά μοντέλα ευπάθειας, καθώς και γενικεύσεις μοντέλων ευπάθειας για την περίπτωση των ομαδοποιημένων ή επαναλαμβανόμενων παρατηρήσεων.

Η ανάλυση των μοντέλων ευπάθειας, αλλά και η ανάλυση των διαφορετικών προσεγγίσεων, παραμετρικά και μη παραμετρικά παρουσιάζονται χρησιμοποιώντας ένα σετ δεδομένων, τα δεδομένα για λέμφωμα τύπου Non-Hodgkin με τη βοήθεια του στατιστικού πακέτου R.

ABSTRACT

Survival analysis deals with the analysis and interpretation of lifetime data and covers a wide variety of applications. For the analysis of such data, several statistical methods have been developed and applied in recent years that are constantly evolving.

In this dissertation, the basic ideas of the survival analysis and some useful distributions used to describe survival data are presented. In addition, the basic properties of these distributions and plots of the survival and hazard functions for each one of them are provided. Furthermore, the theoretical approach of non-parametric estimation methods of survival and cumulative hazard functions are presented using Kaplan-Meier and Nelson-Aalen estimators, respectively. Also, the methodology of the Cox proportional hazards model is presented, as well as some basic techniques for controlling the goodness-of-fit of the model. In addition, an alternative model is presented for analyzing survival data, the accelerated failure time model. Frailty models are presented in detail, which are a generalization of the Cox proportional hazards model which is broadly used in syrvival analysis to measure the effects of covariates. Frailty is included in the model as a random variable for handling the effect of the explanatory variables not taken into account in the model, but also in order to explain the possible heterogeneity that characterizes the population under examination. Parametric and semiparametric frailty models are examined as well as generalizations of frailty models for the case of grouped or repeated observations.

The methods and the different approaches we presented, parametric, nonparametric and semiparametric are then illustrated by a real data set related to the Non-Hodgkin type lymphoma using the statistical package R.

ΚΕΦΑΛΑΙΟ 1

1.1 Εισαγωγή στην Ανάλυση Επιβίωσης

Το παρόν κεφάλαιο αποτελεί μία εισαγωγή στις βασικές έννοιες, πάνω στις οποίες αναπτύχθηκε η Ανάλυση Επιβίωσης. Είναι αξιοσημείωτο ότι έχει δημιουργηθεί μία ξεχωριστή στατιστική θεωρία για την ανάλυση των δεδομένων επιβίωσης, η οποία διαφέρει σημαντικά από τις τεχνικές, που χρησιμοποιούνται στους υπόλοιπους κλάδους της Στατιστικής.

Τα κλασσικά εργαλεία στατιστικής συμπερασματολογίας είναι δύσκολο να εφαρμοσθούν λόγω της διαφορετικής μορφής των δεδομένων. Συνεπώς, η Ανάλυση Επιβίωσης περιλαμβάνει ένα σύνολο στατιστικών μεθόδων κατάλληλων για την ανάλυση και μελέτη δεδομένων διάρκειας ζωής (*survival data*). Κύριο χαρακτηριστικό αυτών των δεδομένων είναι ότι αντικατοπτρίζουν τη χρονική διάρκεια έως ότου συμβεί ένα γεγονός, δηλαδή μία εκ των υπό μελέτη μεταβλητή είναι ο χρόνος. Φυσικά, ο όρος επιβίωση έχει ευρεία έννοια, μπορεί να αναφέρεται στον χρόνο επιβίωσης ενός ασθενή από μια θανατηφόρα νόσο, αλλά μπορεί να προσδιορίζει και την αξιοπιστία μιας μηχανής ή το χρόνο μέχρι την ίαση μίας ασθένειας. Η ανάλυση επιβίωσης εφαρμόζεται σε διάφορους επιστημονικούς κλάδους, όπως η Οικονομία, η Επιδημιολογία, η Φαρμακευτική, η Δημοσιογραφία, η Βιομηχανία, αλλά και η Μηχανική.

1.2 Η Ανάλυση Επιβίωσης: Εισαγωγικές Έννοιες

Τα δεδομένα σε προβλήματα Ανάλυσης Επιβίωσης μπορεί να προέρχονται από συνεχείς, διακριτές ή και μεικτές τυχαίες μεταβλητές. Ωστόσο, ένα αρκετά συνηθισμένο πρόβλημα, που παρατηρείται σε προβλήματα ανάλυσης επιβίωσης είναι η έλλειψη πληροφορίας που μπορεί να διαθέτουμε για κάποιες παρατηρήσεις μας, οι οποίες όπως θα δούμε και παρακάτω μπορεί να είναι λογοκριμένες (*censored*) ή περικομμένες (*truncated*). Το γεγονός αυτό καθιστά δύσκολη την εκτίμηση των αγνώστων παραμέτρων των στατιστικών μοντέλων, που χρησιμοποιούμε με βάση τις συνηθισμένες τεχνικές της στατιστικής συμπερασματολογίας. Για παράδειγμα, μην έχοντας επαρκώς παρατηρήσει τα δεδομένα μας δεν μπορούμε να χρησιμοποιήσουμε τη μέθοδο Μέγιστης Πιθανοφάνειας για την εκτίμηση των παραμέτρων. Ένας τρόπος αντιμετώπισης του εν λόγω προβλήματος, όπως θα δούμε και παρακάτω είναι η εισαγωγή επιπλέον λανθανουσών μεταβλητών (*latent variables*), οι οποίες αντικατοπτρίζουν τα μη παρατηρούμενα δεδομένα και η χρήση του αλγορίθμου Ε.Μ. (*Expectation Maximization*).

Ο χρόνος επιβίωσης μπορεί να οριστεί ως το χρονικό διάστημα μέχρι την πραγματοποίηση ενός γεγονότος. Ως γεγονός μπορεί να θεωρηθεί η εμφάνιση ή η ίαση μιας ασθένειας, η εξέλιξη μίας θεραπείας, ο θάνατος του ασθενούς, η παύση λειτουργίας μίας μηχανής κ.α. Ο χρόνος επιβίωσης μπορεί να είναι είτε διακριτός, είτε συνεχής. Ωστόσο, στις περισσότερες περιπτώσεις είναι συνεχής και με αυτήν την παραδοχή θα συνεχίσουμε σε όλη την παρούσα διπλωματική. Στην Ανάλυση Επιβίωσης όλο το ενδιαφέρον συσσωρεύεται στην εκτίμηση της πιθανότητας ένα άτομο να επιβιώσει τουλάχιστον για χρόνο t και αυτό εκφράζεται μαθηματικά από τη συνάρτηση επιβίωσης. Αυτές οι έννοιες θα αποσαφηνιστούν και θα σχολιαστούν αναλυτικότερα παρακάτω.

1.3 Λογοκριμένα ή Περικομμένα Δεδομένα Επιβίωσης

Τα δεδομένα που εξετάζονται στην ανάλυση επιβίωσης ανήκουν στις εξής βασικές κατηγορίες:

1. Μη λογοκριμένα (*Observed-Uncensored*) δεδομένα.
2. Περικομμένα (*truncated*) ή λογοκριμένα (*censored*) δεδομένα.

- **Παρατηρούμενες – Μη λογοκριμένες παρατηρήσεις, Περικομμένες και Λογοκριμένες παρατηρήσεις**

Ξεκινώντας από τον ορισμό της μη περικομμένης παρατήρησης θα είναι πιο εύκολο να ξεκαθαρίσουμε πλήρως πότε μία παρατήρηση είναι λογοκριμένη ή περικομμένη. Στην ανάλυση επιβίωσης το πείραμα έχει ένα καλά ορισμένο χρονικό διάστημα με αρχή και τέλος δύο δεδομένες χρονικές στιγμές. Οι παρατηρήσεις λοιπόν οι οποίες αναφέρονται στην πραγματοποίηση του υπό μελέτη γεγονότος στο συγκεκριμένο χρονικό διάστημα του πειράματος είναι τα τυπικά δεδομένα επιβίωσης, τα οποία είναι οι ακριβείς χρόνοι επιβίωσης (*uncensored data*) (Hosmer, et al., 2008).

Στις περισσότερες περιπτώσεις όμως το δείγμα μας αποτελείται και από ελλιπείς παρατηρήσεις, οι οποίες διακρίνονται σε δύο κατηγορίες, τις λογοκριμένες (*censored*) και τις περικομμένες (*truncated*). Στην πρώτη περίπτωση δεν έχει παρατηρηθεί η ολοκλήρωση του υπό μελέτη συμβάντος πριν το πέρας του πειράματος ή δεν γνωρίζουμε την ακριβή χρονική στιγμή που παρουσιάστηκε το συμβάν. Αυτές οι μη ολοκληρωμένες παρατηρήσεις εμπίπτουν στην κατηγορία των λογοκριμένων δεδομένων (*censored data*). Γενικά, μία παρατήρηση καλείται λογοκριμένη όταν η έλλειψη πληροφορίας οφείλεται σε διάφορους τυχαίους παράγοντες που αφορούν την ιδιαιτερότητα του κάθε πειράματος. Δηλαδή, για παράδειγμα σε μία μελέτη του χρόνου επιβίωσης σε άτομα που πάσχουν από καρκίνο ενδέχεται να μην παρατηρηθεί ο χρόνος θανάτου για ορισμένους ασθενείς, καθώς αυτοί οι ασθενείς ήταν εν ζωή μετά το πέρας της έρευνας (Hosmer, et al., 2008).

Επιπλέον, μία παρατήρηση καλείται περικομμένη (*truncated*), όταν η έλλειψη πληροφορίας είναι συνυφασμένη με το σχεδιασμό της μελέτης, η οποία επηρεάζει τη διαδικασία επιλογής του δείγματος. Σε αυτές τις περιπτώσεις, η τιμή δεν μπορεί καν να ανιχνευθεί από τον ερευνητή λόγω κάποιων δυσκολιών που παρουσιάζει το πείραμα. Για παράδειγμα, σε μία έρευνα αστρονομίας - αστρομετρίας για την ανίχνευση γαλαξιών στον ουράνιο θόλο παρατηρείται ότι κάποιοι από αυτούς δεν είναι ανιχνεύσιμοι από το τηλεσκόπιο αν η φαινομενική φωτεινότητά τους είναι κάτω από μία ορισμένη τιμή. Συνεπώς, ο ερευνητής δεν έχει τη δυνατότητα να παρατηρήσει αυτές τις τιμές λόγω της ιδιαιτερότητας, που παρουσιάζει η συγκεκριμένη μελέτη (Klein & Moeschberger, 2003).

Οι περικομμένες / λογοκριμένες παρατηρήσεις μπορούν να διαχωριστούν σε τρείς κατηγορίες, οι οποίες είναι οι εξής:

- Περικομμένη/ λογοκριμένη παρατήρηση από αριστερά (*left truncated/ censored*).
- Περικομμένη/ λογοκριμένη παρατήρηση από δεξιά (*right truncated/ censored*).
- Περικομμένη/ λογοκριμένη παρατήρηση σε διάστημα (*interval truncated/ censored*).

- **Περικομμένη από αριστερά, από δεξιά ή σε διάστημα**

- Περικομμένη παρατήρηση από αριστερά (*left truncated*): Ένα συμβάν/γεγονός/αντικείμενο παρατηρήθηκε μόνο αν η υπό μέτρηση ποσότητα είναι μεγαλύτερη από μία συγκεκριμένη τιμή.
- Περικομμένη παρατήρηση από δεξιά (*right truncated*): Ένα συμβάν/γεγονός/αντικείμενο παρατηρήθηκε μόνο αν η υπό μέτρηση ποσότητα είναι μικρότερη από μία συγκεκριμένη τιμή.
- Περικομμένη παρατήρηση σε διάστημα (*interval truncated*): Ένα συμβάν/γεγονός/αντικείμενο παρατηρήθηκε μόνο αν η υπό μέτρηση ποσότητα λαμβάνει τιμή σε ένα συγκεκριμένο διάστημα τιμών.

- **Λογοκριμένη από δεξιά, από αριστερά ή σε διάστημα**

Η πιο συνηθισμένη παρατηρούμενη μορφή σε δεδομένα επιβίωσης είναι τα λογοκριμένα δεδομένα από δεξιά (*right censored data*), τα οποία λαμβάνουν το όνομα τους από την φύση τους, αφού είναι κομμένα από το δεξί μέρος της χρονικής περιόδου. Για παράδειγμα, σε μία μελέτη για τον χρόνο επιβίωσης ατόμων που πάσχουν από κάποια μορφή καρκίνου, παρατηρούνται ασθενείς που έχουν επιβιώσει μέχρι το τέλος του πειράματος. Άρα για αυτούς γνωρίζουμε ότι έχουν επιβιώσει μέχρι μία συγκεκριμένη χρονική στιγμή αλλά δεν έχουμε την παρατηρούμενη πληροφορία που

αναζητάμε, δηλαδή τη χρονική διάρκεια που επήλθε ο θάνατος μετά τη διάγνωση της νόσου. Αυτές οι παρατηρούμενες τιμές αποτελούν δεξιά λογοκριμένες παρατηρήσεις.

Μία άλλη μορφή λογοκριμένης παρατήρησης είναι η λογοκριμένη παρατήρηση από αριστερά (*left censored observation*), έστω C_1. Σε αυτήν την περίπτωση δεν είναι γνωστός ο ακριβής χρόνος έναρξης του υπό μελέτη γεγονότος, αλλά υπάρχει η πληροφορία ότι πραγματοποιήθηκε σίγουρα πριν την χρονική στιγμή C_1. Για παράδειγμα, υπάρχουν περιπτώσεις όπου ένας καρκινοπαθής εισήχθη στην μελέτη, αλλά μετά τη χρονική στιγμή που τον κατέβαλλε η ασθένεια. Συνεπώς, δεν είναι ακριβής ο χρόνος επιβίωσης, καθώς δεν γνωρίζουμε την ακριβή χρονική στιγμή που ξεκίνησε η ασθένεια. Αυτό είναι πολύ σύνηθες σενάριο, ιδιαίτερα στον κλάδο της Ιατρικής μελέτης.

Τέλος, υπάρχουν περιπτώσεις, όπου μπορεί να μην υπάρχει η δυνατότητα να συλλεχθούν οι παρατηρήσεις σε ένα χρονικό διάστημα κατά τη διάρκεια του πειράματος και αυτό φυσικά εξαρτάται αποκλειστικά από το είδος της κάθε έρευνας. Για παράδειγμα σε μία έρευνα για τον καθορισμό του διαστήματος μεταξύ του πρώτου σταδίου καρκίνου και του δεύτερου υπάρχει περίπτωση ο υπεύθυνος της έρευνας να αποφασίσει ότι θέλει να επικοινωνεί με τους ασθενείς κάθε τρείς μήνες για δύο χρόνια. Συνεπώς, οι παρατηρήσεις που συλλέγονται είναι διακριτές παρατηρήσεις και το μόνο που γνωρίζουμε είναι ότι το υπό μελέτη συμβάν έχει ή δεν έχει πραγματοποιηθεί σε ένα χρονικό διάστημα. Σε αυτήν την περίπτωση μιλάμε για λογοκριμένα δεδομένα σε διάστημα (Klein & Moeschberger, 2003).

ΚΕΦΑΛΑΙΟ 2

2.1 Βασικές έννοιες στην Ανάλυση Επιβίωσης

2.1.1 Αθροιστική Συνάρτηση Κατανομής, Συνάρτηση Επιβίωσης και Συνάρτηση Κινδύνου

Ο χρόνος επιβίωσης αποτελεί μία μη αρνητική τυχαία μεταβλητή, έστω T. Αυτή η μεταβλητή συνήθως είναι συνεχής, αλλά λόγω της δυσκολίας πολλές φορές να μετρηθεί σε όλο το χρονικό ορίζοντα του πειράματος μπορεί να εμφανισθεί και σε διακριτή μορφή. Θα μπορούσαμε να προβούμε σε απλά περιγραφικά εργαλεία (π.χ. μέση τιμή, διασπορά, διάμεσος) ή ακόμα και γραφήματα (π.χ. ιστογράμματα), ώστε να έχουμε μία εικόνα των δεδομένων μας και να μπορέσουμε να αντλήσουμε κάποια συμπεράσματα για την κατανομή αυτών. Ωστόσο, στην περίπτωση περικομμένων δεδομένων ο υπολογισμός αυτών των ποσοτήτων δεν μπορεί να πραγματοποιηθεί. Αυτό δημιουργεί προβλήματα στην αρχική επιλογή του κατάλληλου στατιστικού μοντέλου που θα χρησιμοποιηθεί.

Ας υποθέσουμε, χωρίς βλάβη της γενικότητας, ότι το υπό μελέτη χαρακτηριστικό T, από το οποίο έχουμε συλλέξει δεδομένα είναι μία συνεχής τυχαία μεταβλητή, η οποία ακολουθεί μία γνωστή συνάρτηση πυκνότητας πιθανότητας $f_T(\cdot)$. Τότε, όπως είναι γνωστό η αθροιστική συνάρτηση κατανομής της τυχαίας μεταβλητής T δίνεται από τη σχέση:

$$F_T(t) = P(T \leq t) = \int_0^t f_T(s)ds\,, \qquad t \geq 0 \tag{2.1}$$

και εκφράζει την πιθανότητα για ένα τυχαία επιλεγμένο άτομο από τον πληθυσμό να έχει πραγματοποιηθεί το συμβάν (δηλαδή να μην έχει επιβιώσει) μέχρι τη χρονική στιγμή t.

Ωστόσο, στην πλειονότητα των εφαρμογών της Ανάλυσης Επιβίωσης μας ενδιαφέρει περισσότερο να εξετάσουμε και να περιγράψουμε την πιθανότητα επιβίωσης των ατόμων του πειράματος, ώστε να βγάλουμε κάποια πιο χρήσιμα συμπεράσματα. Για το λόγο αυτό, χρησιμοποιείται συνήθως η συνάρτηση επιβίωσης:

$$S_T(t) = P(T > t) = 1 - F_T(t) = \int_t^\infty f_T(s)ds\,, \qquad t \geq 0 \tag{2.2}$$

και εκφράζει την πιθανότητα ένα τυχαία επιλεγμένο άτομο από τον πληθυσμό να έχει επιβιώσει μέχρι τη χρονική στιγμή t.

Από τη σχέση (2.2) έχουμε ότι:

$$f_T(t) = -\frac{dS_T(t)}{dt}.$$

Η συνάρτηση επιβίωσης είναι μία από τις πιο βασικές συναρτήσεις που μελετώνται στην Ανάλυση Επιβίωσης. Ωστόσο, σε ένα πείραμα μας ενδιαφέρει να ποσοτικοποιήσουμε κατά μία έννοια και τον κίνδυνο να μην επιβιώσει μία μονάδα του συστήματος που μελετάμε την αμέσως επόμενη στιγμή δεδομένου ότι έχει επιζήσει μέχρι εκείνη τη στιγμή. Μαθηματικά αν θέλουμε να εκφράσουμε την παραπάνω έννοια θα λέγαμε ότι θέλουμε να υπολογίσουμε ποια είναι η πιθανότητα να επιζήσει μία μονάδα του συστήματος το χρονικό διάστημα $\Delta t = [t, dt]$ δεδομένου ότι έχει επιζήσει μέχρι τη χρονική στιγμή t.

Την παραπάνω πληροφορία μας την δίνει η συνάρτηση κινδύνου, η οποία ορίζεται ως εξής:

$$h_T(t) = lim_{dt\to 0}\frac{P(t < T < t + dt | T \geq t)}{dt} = lim_{dt\to 0}\frac{P(t < T < t + dt)}{dtP(T \geq t)} = \frac{f_T(t)}{S_T(t)}, \qquad t \geq 0. \tag{2.3}$$

Στην πραγματικότητα η συνάρτηση κινδύνου δεν έχει ως αποτέλεσμα μία πιθανότητα, αλλά ένα ρυθμό και πιο συγκεκριμένα τον ρυθμό θνησιμότητας (*mortality rate*), δηλαδή τον ρυθμό αποτυχίας την αμέσως επόμενη χρονική στιγμή από χρόνο t, δεδομένου ότι η μονάδα του πληθυσμού έχει επιβιώσει μέχρι τη δεδομένη χρονική στιγμή t.

Εν συνεχεία θα δώσουμε κάποιες σχέσεις, που σθνδέουν όλες τις παραπάνω συναρτήσεις.

Γνωρίζουμε ότι η αθροιστική συνάρτηση κατανομής συνδέεται με τη συνάρτηση πυκνότητας πιθανότητας με την παρακάτω σχέση:

$$F_T(t) = \int_0^t f_T(s)ds \Leftrightarrow f_T(t) = \frac{dF_T(t)}{dt} = -\frac{dS_T(t)}{dt}, t \geq 0.$$

Συνεπώς,

$$h_T(t) = \frac{-\frac{dS_T(t)}{dt}}{S_T(t)} \Leftrightarrow$$

$$-h_T(t) = \frac{d[\ln S_T(t)]}{dt} \Leftrightarrow$$

$$-\int_0^t h_T(s)ds = \ln[S_T(t)] - \ln[S_T(0)] \Leftrightarrow$$

$$S_T(t) = \exp\left[-\int_0^t h_T(s)ds\right], t \geq 0. \quad (2.4)$$

Αν συμβολίσουμε την αθροιστική συνάρτηση κινδύνου με $H_T(t) = \int_0^t h_T(s)ds\,, t \geq 0$, τότε η παραπάνω σχέση μπορεί να γραφεί ως εξής:

$$S_T(t) = e^{-H_T(t)}, t \geq 0. \quad (2.5)$$

Δύο σημαντικές σχέσεις που προκύπτουν από τα παραπάνω είναι οι εξής:

$$h_T(t) = \frac{-dln\big(S_T(t)\big)}{dt}, t \geq 0.$$

$$H_T(t) = -\ln\big(S_T(t)\big), t \geq 0. \quad (2.6)$$

2.1.2 Πιθανοφάνεια και Λογοκριμένες/Περικομμένες Παρατηρήσεις

Αυτό που διαφοροποιεί την Ανάλυση Επιβίωσης από άλλους κλάδους Στατιστικής είναι σε πολλές περιπτώσεις η ύπαρξη λογοκριμένων, αλλά και περικομμένων δεδομένων. Στην πραγματικότητα τα λογοκριμένα/περικομμένα δεδομένα αποτελούν ατελείς παρατηρούμενες τιμές. Ωστόσο, μας δίνουν πληροφορία, έστω και ημιτελή. Υπάρχει μία ιδιαιτερότητα λοιπόν στην κατασκευή της συνάρτησης πιθανοφάνειας, όταν υπάρχουν τέτοιου είδους δεδομένα. Θα κάνουμε μία βασική υπόθεση στις τεχνικές που θα αναφέρουμε παρακάτω, ότι οι χρόνοι επιβίωσης, αλλά και οι

περικομμένοι/λογοκριμένοι χρόνοι επιβίωσης είναι ανεξάρτητοι. Σε περιπτώσεις, όπου αυτή η υπόθεση δεν ικανοποιείται θα πρέπει να ληφθούν υπόψη άλλες εξειδικευμένες τεχνικές. Είναι πολύ σημαντικό επίσης να συμπεριλάβουμε όλη την πληροφορία που έχουμε από μία περικομμένη/λογοκριμένη παρατήρηση. Στην περίπτωση των πραγματικών χρόνων επιβίωσης γνωρίζουμε την πιθανότητα ο χρόνος επιβίωσης ενός τυχαία επιλεγμένου ατόμου να πραγματοποιηθεί τη δεδομένη χρονική στιγμή, χρησιμοποιώντας φυσικά τη συνάρτηση πυκνότητας πιθανότητας. Για λογοκριμένα δεδομένα από δεξιά γνωρίζουμε ότι το υπό μελέτη γεγονός δεν έχει πραγματοποιηθεί μέχρι την παρατηρούμενη χρονική στιγμή (*censoring time*) και μπορούμε να αντλήσουμε πληροφορία με τη βοήθεια της συνάρτησης επιβίωσης. Στην περίπτωση των από αρστερά λογοκριμένων δεδομένων γνωρίζουμε ότι το υπό μελέτη συμβάν έχει πραγματοποιηθεί πριν από την παρατηρούμενη λογοκριμένη από αριστερά τιμή και σε αυτήν την περίπτωση μπορούμε να αντλήσουμε πληροφορία με τη βοήθεια της αθροιστικής συνάρτησης πιθανοφάνειας για αυτήν την χρονική στιγμή. Τέλος, στην περίπτωση λογοκριμένων δεδομένων ανά διάστημα, γνωρίζουμε ότι το συμβάν πραγματοποιήθηκε σε ένα καθορισμένο διάστημα και μπορούμε να υπολογίσουμε την πιθανότητα ο χρόνος επιβίωσης να εμπίπτει σε αυτό το διάστημα. Για περικομμένα δεδομένα, μπορούν να υπολογιστούν οι αντίστοιχες πιθανότητες χρησιμοποιώντας φυσικά και δεσμευμένες πιθανότητες (Klein & Moeschberger, 2003). Πιο κάτω παραθέτουμε τη συνεισφορά στην συνάρτηση πιθανοφάνειας παρατηρήσεων που εμπίπτουν σε διάφορες κατηγορίες:

Ακριβής Χρόνος Επιβίωσης (*exact lifetimes*) $\rightarrow f\ (x)$.

Λογοκριμένη Παρατήρηση από Δεξιά (*right censored observations*) $\rightarrow S\ (C_r\ )$, όπου C_r λογοκριμένες παρατηρήσεις από δεξιά.

Λογοκριμένη Παρατήρηση από Αριστερά (*left censored observations*) $\rightarrow\ 1 - S\ (C_l)$ όπου C_l λογοκριμένες παρατηρήσεις από αριστερά.

Λογοκριμένη Παρατήρηση σε Διάστημα (*interval censored observations*) $\rightarrow [S\ (L) - S\ (R)]$, όπου L λογοκριμένες παρατηρήσεις από αριστερά και R λογοκριμένες παρατηρήσεις από δεξιά.

Περικομμένη Παρατήρηση από Αριστερά (*left truncated observations*) $\rightarrow f\ (x)\ \ /S\ (Y_L)$, όπου Y_L περικομμένες παρατηρήσεις από αριστερά.

Περικομμένη Παρατήρηση από Δεξιά (*right truncated observations*) $\rightarrow f(x)/\ [1 - S\ (Y_R)]$, όπου Y_R περικομμένες παρατηρήσεις από δεξιά.

Περικομμένη Παρατήρηση σε Διάστημα (*interval truncated observations*) $\rightarrow f\ (x)/[S\ (Y_L) - S\ (Y_R\ )]$, όπου Y_L περικομμένες παρατηρήσεις από αριστερά και Y_R περικομμένες παρατηρήσεις από δεξιά, αντίστοιχα.

Αναφέραμε τα βασικότερα είδη λογοκριμένων δεδομένων και την πληροφορία που μπορούμε να αντλήσουμε από αυτά. Ωστόσο, θα επικεντρωθούμε περισσότερο στην κατασκευή της συνάρτησης πιθανοφάνειας σε λογοκριμένες παρατηρούμενες τιμές από δεξιά (*right censored data*) στο παρόν κεφάλαιο, αφού είναι αυτά που εμφανίζονται στις περισσότερες εφαρμογές δεδομένων ανάλυσης επιβίωσης.

Έστω $T_1, T_2, \dots, T_n$ οι χρόνοι που παρατηρούμε και στους οποίους συμβαίνει κάποιο γεγονός (*event times*) και $X_1, X_2, \dots, X_n$ οι χρόνοι επιβίωσης των ατόμων. Έστω, $C_1, C_2, \dots, C_n$ οι χρόνοι λογοκρισίας των ατόμων για τυχαίο δείγμα μεγέθους n. Για μη λογοκριμένα δεδομένα θα ισχύει $X_i \leq C_i, i = 1, 2, \dots, n$, και άρα $T_i = X_i$, ειδάλλως ισχύει $C_i < X_i$ και αρα $T_i = C_i$. Με άλλα λόγια, τα παρατηρούμενα δεδομένα μπορούμε να θεωρήσουμε ότι εκφράζονται από μία ακολουθία από ζεύγη τιμών όπου κάθε ένα από αυτά τα ζεύγη περιλαμβάνουν τη μικρότερη τιμή μεταξύ του χρόνου επιβίωσης και του χρόνου λογοκρισίας για κάθε μονάδα του δείγματος σε συνδυασμό με ένα δείκτη που υποδηλώνει αν η συγκεκριμένη παρατήρηση είναι λογοκριμένη ή όχι. Έστω, f η συνάρτηση πυκνότητας πιθανότητας και S αντίστοιχα, η συνάρτηση επιβίωσης των χρόνων επιβίωσης X.

Το παρατηρούμενο δείγμα γράφεται ως εξής:

$$(T_i, \delta_i), i = 1,2, \dots, n, \text{όπου}$$

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

$$\delta_i = I(X_i \le C_i) = \begin{cases} 1, & \alpha\nu \ X_i \le C_i \ \ (\chi\rho\acute{o}\nu o\varsigma \ \epsilon\pi\iota\beta\acute{\iota}\omega\sigma\eta\varsigma) \\ 0, & \alpha\nu \ X_i > C_i \ \ (\lambda o\gamma o\kappa\rho\iota\mu\acute{\epsilon}\nu\eta \ \pi\alpha\rho\alpha\tau\acute{\eta}\rho\eta\sigma\eta \ \alpha\pi\acute{o} \ \delta\epsilon\xi\iota\acute{\alpha}). \end{cases} \tag{2.7}$$

Για την περίπτωση, όπου τα δεδομένα επιβίωσης περιλαμβάνουν λογοκριμένες παρατηρήσεις από δεξιά, η συνάρτηση πιθανοφάνειας λαμβάνει την εξής μορφή:

$$L = \prod_{i=1}^{n} [f(t_i)(1 - G(t_i))]^{\delta_i} [g(t_i)S(t_i)]^{1-\delta_i} = \left\{ \prod_{i=1}^{n} (1 - G(t_i))^{\delta_i} g(t_i)^{1-\delta_i} \right\} \left\{ \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \right\}, \tag{2.8}$$

όπου θεωρούμε ότι X *και* C είναι ανεξάρτητες τυχαίες μεταβλητές, f και S είναι η συνάρτηση πυκνότητας πιθανότητας και η συνάρτηση επιβίωσης για τον χρόνο επιβίωσης X, ενώ g και G είναι η συνάρτηση πυκνότητας πιθανότητας και η αθροιστική συνάρτηση κατανομής για τον λογοκριμένο χρόνο C.

Η κατανομή των λογοκριμένων χρόνων μπορεί να μην συμπεριληφθεί στην πιθανοφάνεια, καθώς δεν εξαρτάται από τις παραμέτρους που εξετάζουμε. Έτσι, θα έχουμε τη σχέση (Klein & Moeschberger, 2003) (Janssen & Duchateau, 2008)

$$L \propto \prod_{i=1}^{n} [f(t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i}. \tag{2.9}$$

Χρησιμοποιώντας τις σχέσεις (2.3) και (2.6) σχέση (2.9) γράφεται και ως

$$L \propto \prod_{i=1}^{n} [h(t_i)]^{\delta_i} S(t_i) \ \text{ή}$$

$$L \propto \prod_{i=1}^{n} [h(t_i)]^{\delta_i} \exp(-H(t_i)), \tag{2.10}$$

όπου $H(t_i) = \int_0^{t_i} h(s)ds$ η αθροιστική συνάρτηση κινδύνου.

2.1.3 Στατιστικά Παραμετρικά Μοντέλα

Όπως έχουμε ήδη αναφέρει, το ενδιαφέρον μας επικεντρώνεται στην μελέτη του χρόνου επιβίωσης, το οποίο φυσικά συνεπάγεται στην εύρεση των προαναφερθέντων συναρτήσεων, της συνάρτησης επιβίωσης $S_T(t)$ και της συνάρτησης κινδύνου $h_T(t)$. Είναι δύσκολο να προσεγγίσουμε ακριβώς τη συμπεριφορά του υπό μελέτη συστήματος, γιατί είναι πολλοί παράγοντες που το επηρεάζουν, τους οποίους ίσως δεν γνωρίζουμε και ακόμη και στις περιπτώσεις που τους γνωρίζουμε ενδεχομένως να μην μπορούμε να τους ποσοτικοποιήσουμε. Ωστόσο, υπάρχουν κατάλληλα μοντέλα, τα οποία μπορούν να προσεγγίσουν αρκετά καλά την αξιοπιστία ενός συστήματος για το χρόνο επιβίωσης και πληρούν αρκετές ιδιότητες αυτού. Αυτά εκφράζονται με κατάλληλες κατανομές, οι οποίες χρησιμοποιούνται ευρέως για τη μελέτη δεδομένων επιβίωσης. Κάποιες από αυτές είναι η Εκθετική κατανομή (*Exponential*), η Weibull, η Λογαριθμολογιστική (*Log-Logistic*), η Gompertz, η Λογαριθμοκανονική (*Log-normal*), η Γάμμα (*Gamma*) και η Pareto.

Παρακάτω θα αναφερθούμε σε αυτές τις κατανομές καθώς και στις ιδιότητές τους.

Εκθετική κατανομή: $T \sim Exp(\lambda)$

Η Εκθετική κατανομή αποτελεί το πιο απλό μοντέλο για τη μελέτη δεδομένων επιβίωσης και λόγω των πολύ σημαντικών μαθηματικών ιδιοτήτων της χρησιμοποιείται σε πολλές πρακτικές εφαρμογές. Οι συναρτήσεις επιβίωσης, καθώς και κάποια χαρακτηριστικά της εν λόγω κατανομής δίνονται παρακάτω:

Έστω ο χρόνος επιβίωσης T ακολουθεί την Εκθετική κατανομή $T \sim Exp(\lambda)$, τότε

- Η συνάρτηση πυκνότητας πιθανότητας είναι η $f_T(t) = \lambda e^{-\lambda t}, \quad t \geq 0, \lambda > 0.$
- Η αθροιστική συνάρτηση κατανομής του χρόνου επιβίωσης είναι η $F_T(t) = 1 - e^{-\lambda t}, \quad t \geq 0, \lambda > 0.$
- Η συνάρτηση επιβίωσης είναι η $S_T(t) = e^{-\lambda t}, \quad t \geq 0, \lambda > 0.$
- Η συνάρτηση κινδύνου είναι η $h_T(t) = \lambda, \quad t \geq 0, \lambda > 0.$
- Η μέση τιμή είναι $E(T) = \frac{1}{\lambda}, \ t \geq 0, \lambda > 0.$
- Η διασπορά είναι $V(T) = \frac{1}{\lambda^2}, \ t \geq 0, \lambda > 0.$

Μία πολύ βασική παρατήρηση είναι ότι η συνάρτηση κινδύνου όπως βλέπουμε παραπάνω είναι σταθερή, δηλαδή $h_T(t) = \lambda, \ t \geq 0$. Συνεπώς, ο κίνδυνος να πραγματοποιηθεί το γεγονός την αμέσως επόμενη στιγμή είναι σταθερός και δεν εξαρτάται από την δεδομένη χρονική στιγμή. Αυτή η παρατήρηση είναι απόρροια μίας πολύ σημαντικής ιδιότητας της Εκθετικής κατανομής, η οποία καλείται «έλλειψη μνήμης» (Wienke, 2011). Άρα, αν $T \sim Exp(\lambda)$, τότε η δεσμευμένη πιθανότητα ο χρόνος επιβίωσης T να υπερβεί την χρονική στιγμή $t + t_0$ δεδομένου ότι έχει ήδη υπερβεί τον χρόνο t δεν εξαρτάται από το τι έχει συμβεί μέχρι εκείνη τη χρονική στιγμή. Μαθηματικά αυτό μπορεί να εκφραστεί ως εξής:

$$P(T > t + t_0 | T > t) = P(T > t_0).$$

Η παραπάνω σχέση αποδεικνύεται εύκολα ως εξής:

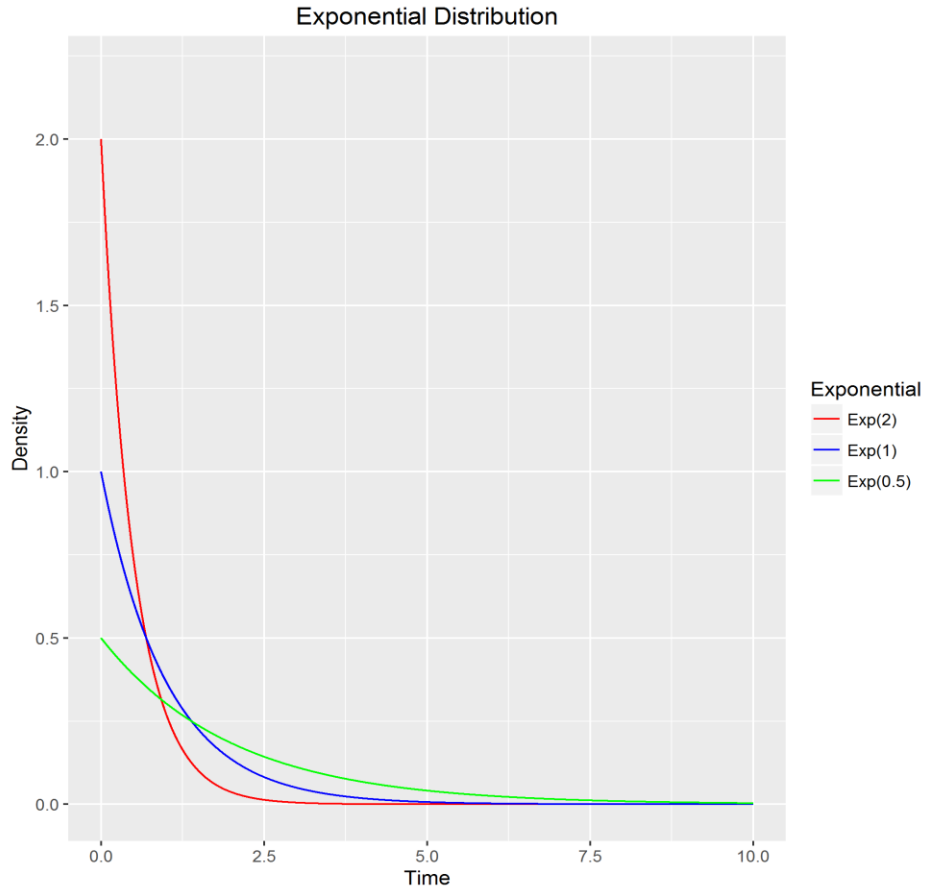
$$P(T > t + t_0 | T > t) = \frac{P(T > t + t_0, T > t)}{P(T > t)} = \frac{P(T > t + t_0)}{P(T > t)} = \frac{S(t + t_0)}{S(t)} = \frac{e^{-\lambda(t+t_0)}}{e^{-\lambda t}} = e^{-\lambda t_0} = P(T > t_0).$$

Η πιθανότητα να συμβεί το γεγονός πέρα από μία δεδομένη χρονική στιγμή είναι ανεξάρτητη από την δεδομένη ηλικία της υπό μελέτης μονάδας του δείγματος, το οποίο καθιστά αρκετά δύσκολη την επιλογή αυτού του μοντέλου. Αν το σκεφτούμε καλύτερα αυτή η ιδιότητα είναι αρκετά μη ρεαλιστική και κάνει την Εκθετική κατανομή ακατάλληλη για την μελέτη δεδομένων που αφορούν ανθρώπινες ζωές, εκτός από τις περιπτώσεις που λαμβάνουμε δεδομένα επιβίωσης για ένα πολύ μικρό χρονικό διάστημα. Ωστόσο, είναι μία ρεαλιστική κατανομή για την μελέτη του χρόνου επιβίωσης μηχανών και μηχανικών εξαρτημάτων.

Χρησιμοποιώντας τον παρακάτω κώδικα δημιουργούμε το Διάγραμμα 2.1.

```
#Exponential Distribution
x <- seq(0, 10, 0.001)
y1 = dexp(x, rate = 1)
df1 <- data.table(x, y1)
y2 <- dexp(x, rate = 0.5)
df2 <- data.table(x, y2)
y3 <- dexp(x, rate = 2)
```

```
df3 <- data.table(x, y3)
df5 <- merge(df1, df2, by = "x")
df <- merge(df3, df5, by = "x")
df_last <- melt.data.table(df, id = 1, value.factor = TRUE)
p12 <- ggplot(df_last, aes(x , value, colour = variable)) + geom_line() + ylim(0, 2.2)
p12 <- p12 + scale_color_manual(name = "Exponential", values = c("red", "blue", "green"),
labels = c('Exp(2)', 'Exp(1)', 'Exp(0.5)')) + ggtitle("Exponential Distribution") +
xlab("Time") + ylab("Density")
ggsave("Exponential_density.png")
```



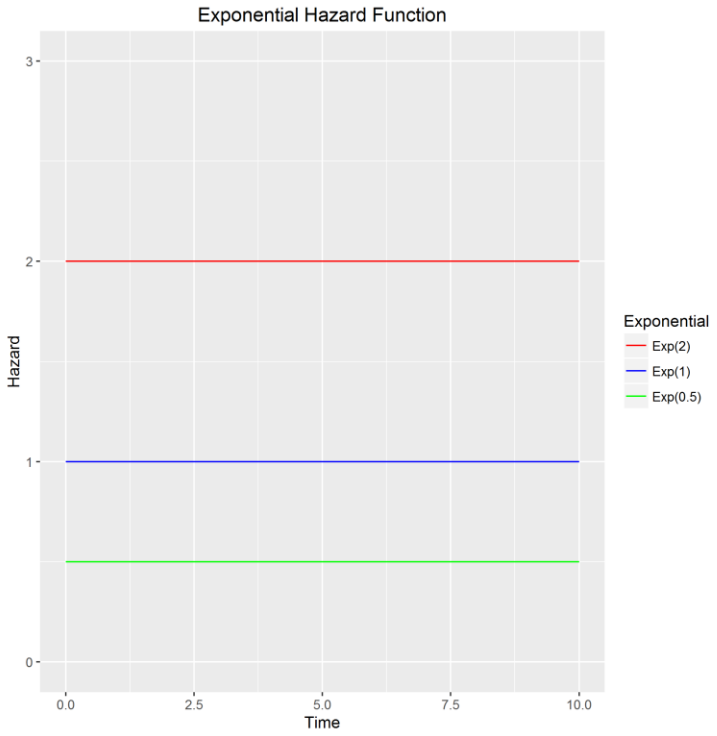
Διάγραμμα 2.1: *Οι συναρτήσεις πυκνότητας πιθανότητας της Εκθετικής κατανομής για διαφορετικές τιμές της παραμέτρου* λ.

Στο Διάγραμμα 2.1 απεικονίζονται οι συναρτήσεις πυκνότητας πιθανότητας της Εκθετικής κατανομής για τιμές την παραμέτρου $\lambda = 0.5, 1, 2$.

Χρησιμοποιώντας τον παρακάτω κώδικα δημιουργούμε το Διάγραμμα 2.2.

```
#hazard function
h.exp <- function(x, rate) {rate}
x <- seq(0, 10, 0.001)
y1 = h.exp(x, rate = 1)
df1 <- data.table(x, y1)
y2 <- h.exp(x, rate = 0.5)
df2 <- data.table(x, y2)
y3 <- h.exp(x, rate = 2)
```

```
df3 <- data.table(x, y3)
df5 <- merge(df1, df2, by = "x")
df <- merge(df3, df5, by = "x")
df_last <- melt.data.table(df, id = 1, value.factor = TRUE)
p13 <- ggplot(df_last, aes(x , value, colour = variable)) + geom_line() + ylim(0, 3)
p13 <- p13 + scale_color_manual(name = "Exponential", values = c("red", "blue", "green"),
labels = c('Exp(2)', 'Exp(1)', 'Exp(0.5)'))+ ggtitle("Exponential Hazard Function") +
xlab("Time") + ylab("Hazard")
ggsave("Exponential_hazard.png")
```



Διάγραμμα 2.2 : *Η συναρτήσεις κινδύνου που προκύπτουν αν θεωρήσουμε ότι ο χρόνος επιβίωσης ακολουθεί Εκθετική κατανομή για διαφορετικές τιμές της παραμέτρου λ.*

Στο Διάγραμμα 2.2 απεικονίζονται οι συναρτήσεις κινδύνου για την Εκθετική κατανομή για τιμές την παραμέτρου $\lambda = 0.5, 1, 2$. Όπως μπορούμε να δούμε όλες αποτελούν σταθερές συναρτήσεις, καθώς αυτό προκύπτει από την εν λόγω κατανομή.

Κατανομή Weibull: $T \sim Weib(\lambda, v)$

Η κατανομή Weibull έλαβε το όνομά της από τον Σουηδό μηχανικό Wallodi Weibull (1939). Αποτελεί μία γενίκευση της Εκθετικής κατανομής και περιλαμβάνει δύο παραμέτρους v και λ, εκ των οποίων η πρώτη v είναι η παράμετρος σχήματος (*shape parameter*), η οποία καθορίζει το σχήμα της εν λόγω κατανομής και η δεύτερη λ, η παράμετρος κλίμακας (*scale parameter*), η οποία καθορίζει κατά κάποιο τρόπο την μεταβλητότητα της κατανομής. Είναι φυσικά μία κατανομή, η οποία χρησιμοποιείται αρκετά σε πρακτικές εφαρμογές, αφού σε αντίθεση με την Εκθετική κατανομή, η συνάρτηση κινδύνου για τον χρόνο επιβίωσης είναι σταθερή.

Για τη συγκεκριμένη κατανομή έχουμε τα εξής (Wienke, 2011):

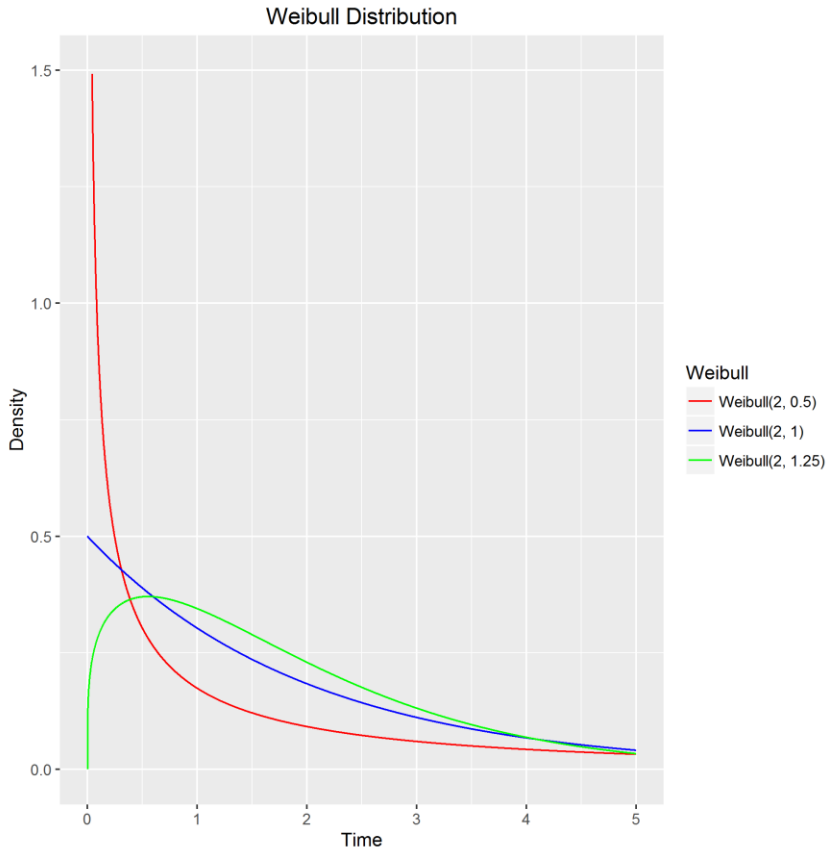
- η συνάρτηση πυκνότητας πιθανότητας είναι $f_T(t) = \lambda\nu t^{\nu-1}e^{-\lambda t^{\nu}}, \quad \lambda > 0, \nu > 0, t > 0.$
- η αθροιστική συνάρτηση κατανομής του χρόνου επιβίωσης είναι $F_T(t) = 1 - e^{-\lambda t^{\nu}}, \quad \lambda > 0, \nu > 0, t > 0.$
- η συνάρτηση επιβίωσης είναι η $S_T(t) = e^{-\lambda t^{\nu}}, \quad \lambda > 0, \nu > 0, t > 0.$
- η συνάρτηση κινδύνου είναι η $h_T(t) = \lambda\nu t^{\nu-1}, \quad \lambda > 0, \nu > 0, t > 0.$
- η μέση τιμή είναι $E(T) = \lambda^{-\frac{1}{\nu}}\Gamma\left(1 + \frac{1}{\nu}\right), \quad \lambda > 0, \nu > 0, t > 0.$
- η διασπορά είναι $V(T) = \lambda^{-\frac{2}{\nu}}\left(\Gamma\left(1 + \frac{2}{\nu}\right) - \Gamma\left(1 + \frac{1}{\nu}\right)^2\right), \ \lambda > 0, \nu > 0, t > 0,$ όπου η Γάμμα συνάρτηση είναι: $\Gamma(\kappa) = \int_0^{\infty} s^{k-1}e^{-s}ds.$

Η παράμετρος ν επιτρέπει στην συνάρτηση κινδύνου να λάβει μία πληθώρα από διαφορετικά σχήματα. Πιο συγκεκριμένα έχουμε ότι:

- Αν $\nu < 1$, τότε ο ρυθμός αποτυχίας μειώνεται με την πάροδο του χρόνου. Αυτό χρησιμοποιείται αν υπάρχουν ενδείξεις «βρεφικής θνησιμότητας» (*«infant mortality»*) στον υπό μελέτη πληθυσμό ή γενικώς στοιχεία που υποδηλώνουν ένα αυξανόμενο ρυθμό αποτυχίας εξαρχής, ο οποίος μειώνεται με το πέρασμα του χρόνου.
- Αν $\nu = 1$, τότε η συνάρτηση κινδύνου παραμένει σταθερή, μιας και η κατανομή Weibull συμπίπτει με την Εκθετική κατανομή με παράμετρο λ.
- Αν $\nu > 1$, τότε ο ρυθμός αποτυχίας αυξάνεται με την πάροδο του χρόνου. Αυτή η επιλογή παραμέτρου είναι χρήσιμη σε περιπτώσεις όπου παρατηρείται μια διαδικασία "γήρανσης" (*«aging process»*) στα δεδομένα, δηλαδή όσο περνάει ο χρόνος, οι μονάδες του πληθυσμού είναι πιο πιθανό να μην επιβιώνουν.

Χρησιμοποιώντας τον παρακάτω κώδικα δημιουργούμε το Διάγραμμα 2.3.

```
library(data.table)
library(ggplot2)
x <- seq(0, 5, 0.001)
#Weibull Distribution
#propability function
y1 = dweibull(x, scale = 2, shape = 1)
df1 <- data.table(x, y1)
y2 <- dweibull(x, scale = 2, shape = 1.25)
df2 <- data.table(x, y2)
y3 <- dweibull(x, scale = 2, shape = 0.5)
df3 <- data.table(x, y3)
df4 <- merge(df1, df2, by = "x")
df <- merge(df3, df4, by = "x")
df_last <- melt.data.table(df, id = 1, value.factor = TRUE)
p <- ggplot(df_last, aes(x, value, colour = variable)) + geom_line() + ylim(0, 1.5)
p <- p + scale_color_manual(name = "Weibull", values = c("red", "blue", "green"), labels
= c('Weibull(2, 0.5)', 'Weibull(2, 1)', 'Weibull(2, 1.25)'))
p <- p + ggtitle("Weibull Distribution") + xlab("Time") + ylab("Density")
ggsave("weibull_density.png")
```

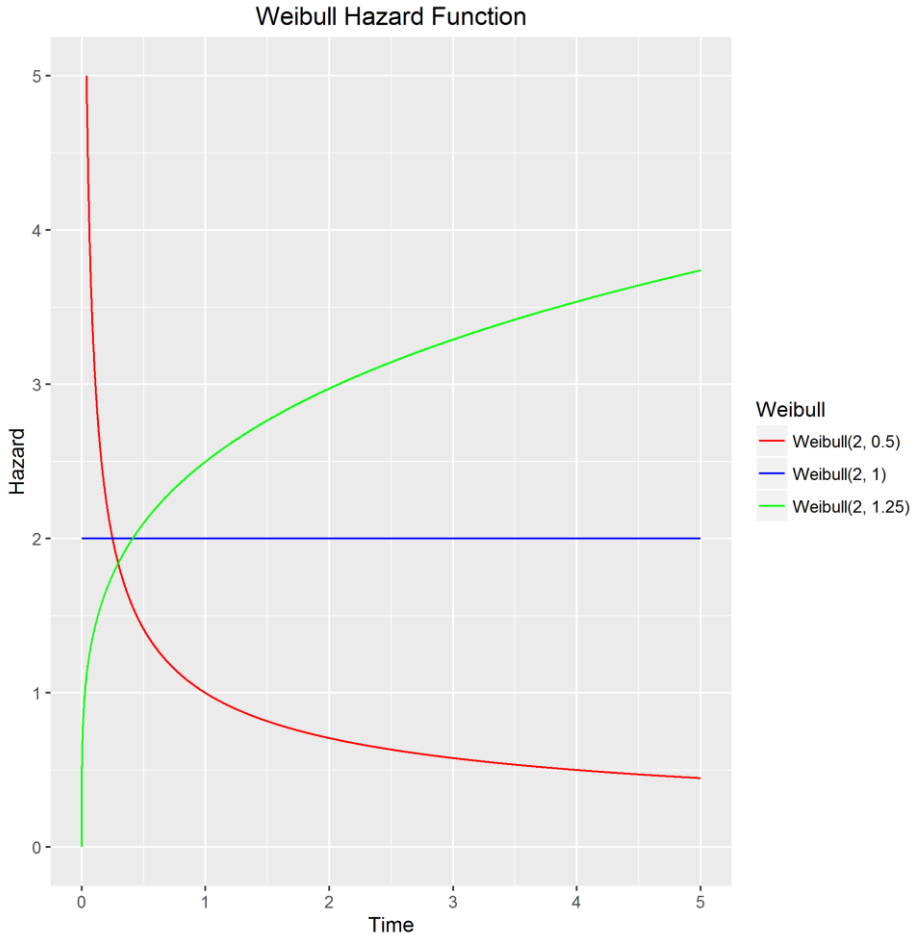


Διάγραμμα 2.3: *Οι συναρτήσεις πυκνότητας πιθανότητας της Weibull κατανομής για διαφορετική επιλογή παραμέτρων.*

Στο Διάγραμμα 2.3 απεικονίζονται οι συναρτήσεις πυκνότητας πιθανότητας των κατανομών $Weibull\ (\lambda = 2, \nu = 0.5), Weibull\ (\lambda = 2, \nu = 1)$ και $Weibull\ (\lambda = 2, \nu = 1.25)$.

Χρησιμοποιώντας τον παρακάτω κώδικα δημιουργούμε το Διάγραμμα 2.4.

```
#Hazard function
h <- function(t, scale, shape){ y <- scale * shape * (t ^ (shape - 1)); return(y)}
x <- seq(0, 5, 0.001)
#Weibull Distribution
#propability function
y1 = h(x, scale = 2, shape = 1)
df1 <- data.table(x, y1)
y2 <- h(x, scale = 2, shape = 1.25)
df2 <- data.table(x, y2)
y3 <- h(x, scale = 2, shape = 0.5)
df3 <- data.table(x, y3)
df4 <- merge(df1, df2, by = "x")
df <- merge(df3, df4, by = "x")
df_last <- melt.data.table(df, id = 1, value.factor = TRUE)
p1 <- ggplot(df_last, aes(x , value, colour = variable)) + geom_line() + ylim(0, 5)
p1 <- p1 + scale_color_manual(name = "Weibull", values = c("red", "blue", "green"),
labels = c('Weibull(2, 0.5)','Weibull(2, 1)','Weibull(2, 1.25)')) + ggtitle("Weibull
Hazard Function") + xlab("Time") + ylab("Hazard")
ggsave("weibull_hazard.png")
```



Διάγραμμα 2.4: *Οι συναρτήσεις κινδύνου για την επιλογή της Weibull κατανομής χρησιμοποιώντας διαφορετικές παραμέτρους.*

Στο Διάγραμμα 2.4 απεικονίζονται οι συναρτήσεις κινδύνου, όπως προκύπτουν για τις κατανομές $Weibull\ (\lambda = 2, \nu = 0.5)$, $Weibull\ (\lambda = 2, \nu = 1)$ και $Weibull\ (\lambda = 2, \nu = 1.25)$. Παρατηρούμε ότι για παράμετρο $\nu = 1$, η $Weibull$ ταυτίζεται με την Εκθετική με παράμετρο λ και έτσι η συνάρτηση κινδύνου είναι μία σταθερή συνάρτηση $h = \lambda = 2$. Επιπλέον, για $\nu = 0.5$ $(\nu < 1)$ παρατηρούμε ότι ο ρυθμός αποτυχίας μειώνεται με την πάροδο του χρόνου, ενώ για $\nu = 1.25$ $(\nu > 1)$ παρατηρούμε ότι ο ρυθμός αποτυχίας αυξάνεται με την πάροδο του χρόνου.

Λογαριθμολογιστική κατανομή: $T \sim logL(u, k)$

Η Λογαριθμολογιστική κατανομή αποτελεί μία εναλλακτική περίπτωση μοντέλου για την μελέτη δεδομένων επιβίωσης.

Για τη συγκεκριμένη κατανομή έχουμε (Wienke, 2011):

- η συνάρτηση πυκνότητας πιθανότητας είναι $f_T(t) = \frac{uk(ut)^{k-1}}{(1+(ut)^k)^2},\ u > 0,\ k > 0,\ t > 0.$
- η αθροιστική συνάρτηση κατανομής του χρόνου επιβίωσης είναι $F_T(t) = 1 - \frac{1}{1+(ut)^k},\ u > 0,\ k > 0, t > 0.$
- η συνάρτηση επιβίωσης είναι η $S_T(t) = \frac{1}{1+(ut)^k},\ u > 0,\ k > 0,\ t > 0.$

- η συνάρτηση κινδύνου είναι η $h_T(t) = \frac{uk(ut)^{k-1}}{1+(ut)^k}, \quad u > 0,\ k > 0,\ t > 0.$

- η μέση τιμή είναι $E(T) = \ln\left(\frac{1}{k}\right), \quad k > 0.$

- η διασπορά είναι $V(T) = \frac{1}{u^2},\ u > 0.$

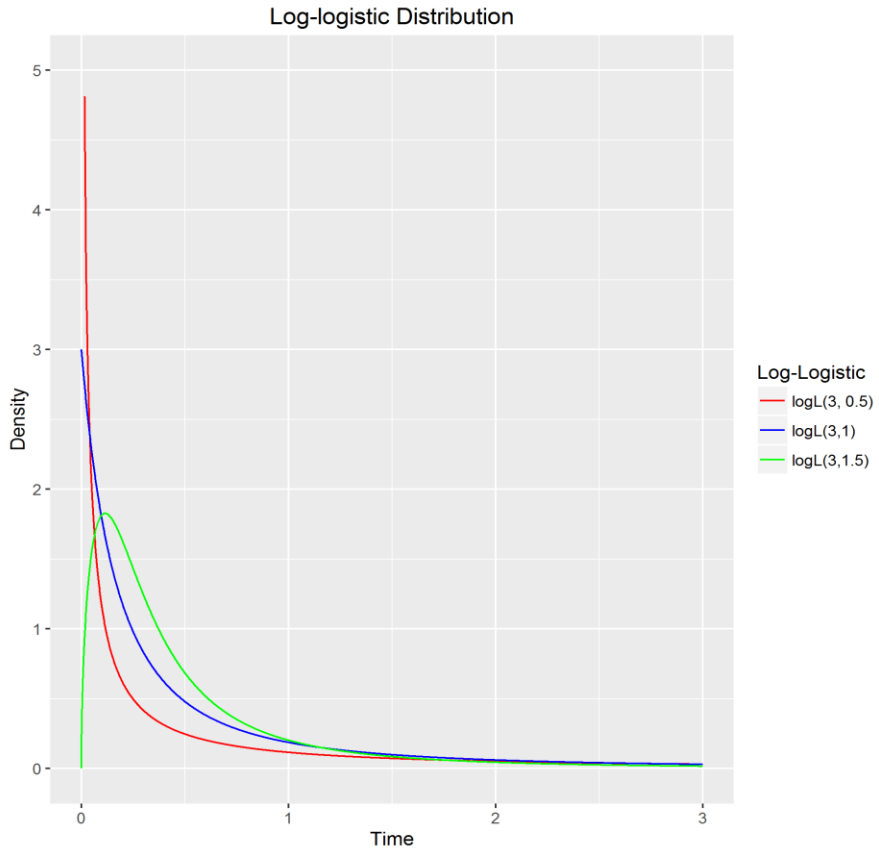
Είναι αξιοσημείωτο να αναφερθεί ότι για $k > 1$, η εν λόγω κατανομή παρουσιάζει την ίδια συμπεριφορά με την Λογαριθμοκανονική κατανομή (Log-normal), η οποία θα αναφερθεί παρακάτω.

Είναι μία κατανομή όπου ο ρυθμός αποτυχίας παρουσιάζει την εξής συμπεριφορά:

- Αν $k \leq 1$, τότε ο ρυθμός αποτυχίας μειώνεται με την πάροδο του χρόνου.
- Αν $k > 1$, τότε ο ρυθμός αποτυχίας έχει «κυρτή» συμπεριφορά, δηλαδή αρχικά αυξάνεται και έπειτα μειώνεται με την πάροδο του χρόνου.

Χρησιμοποιώντας τον παρακάτω κώδικα δημιουργούμε το Διάγραμμα 2.5.

```
###log-logistic
##density
#install.packages("actuar")
library(actuar)
x<-seq(0,3, 0.001)
d.log.log <- function(x, scale, shape)
  shape*scale* (scale*x)^(shape-1) * (1+(scale*x)^shape)^(-2)
h.log.log <- function(x,scale,shape)
  shape*scale* (scale*x)^(shape-1)  / ( 1 + (scale*x)^shape)
#propability function
y1=d.log.log(x, scale = 3, shape = 1)
df1<-data.table(x,y1)
y2<-d.log.log(x, scale = 3, shape = 1.5)
df2<-data.table(x,y2)
y3<-d.log.log(x, scale=3, shape = 0.5)
df3<-data.table(x,y3)
df4<-merge(df1,df2,by = "x")
df<-merge(df3,df4,by = "x")
df_last<-melt.data.table(df,id=1,value.factor=TRUE)
p2<-ggplot(df_last, aes(x , value, colour=variable))+ geom_line()+ ylim(0, 5)
p2<-p2+scale_color_manual(name="Log-Logistic",values=c("red","blue","green"),labels =
c('logL(3, 0.5)','logL(3,1)','logL(3,1.5)'))
p2<-p2+ggtitle("Log-logistic Distribution")+ xlab("Time")+ ylab("Density")
ggsave("log_logistic_density.png")
```

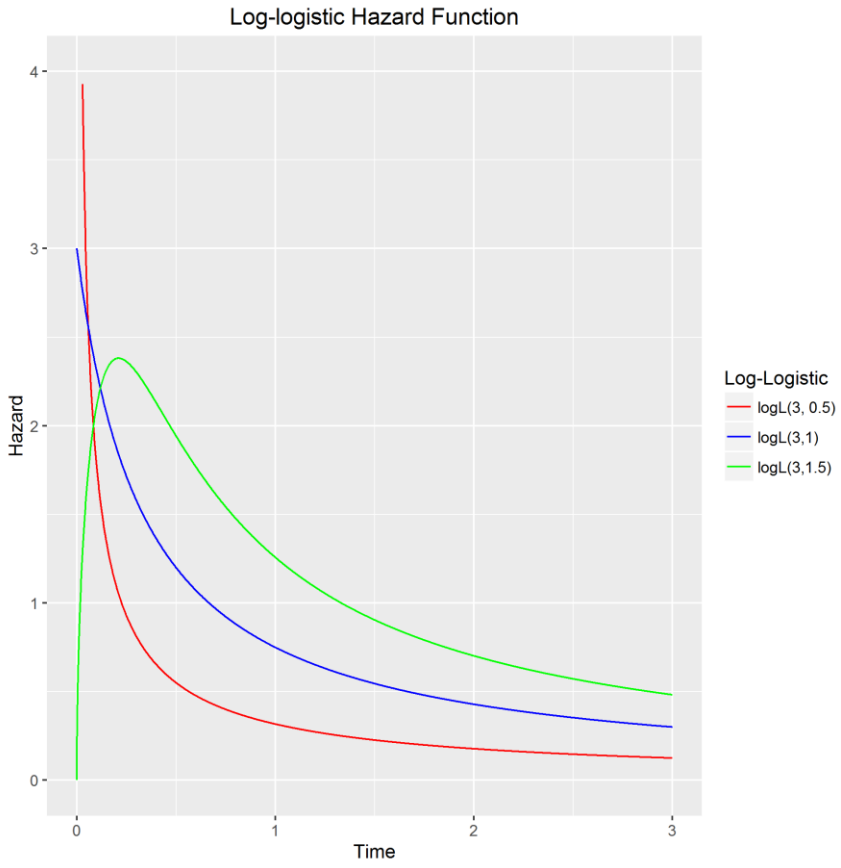


Διάγραμμα 2.5*: Οι συναρτήσεις πυκνότητας πιθανότητας της Log-Logistic κατανομής για διαφορετικές παραμέτρους.*

Στο Διάγραμμα 2.5 απεικονίζονται οι συναρτήσεις πυκνότητας πιθανότητας των κατανομών $Loglogistic\ (u = 3, k = 0.5)$, $Loglogistic\ (u = 3, k = 1)$ και $Loglogistic\ (u = 3, k = 1.5)$.

Χρησιμοποιώντας τον παρακάτω κώδικα δημιουργούμε το Διάγραμμα 2.6.

```
#hazard function
x<-seq(0,3,0.001)
y1=h.log.log(x, scale=3,shape=1)
df1<-data.table(x,y1)
y2<-h.log.log(x,scale=3, shape=1.5)
df2<-data.table(x,y2)
y3<-h.log.log(x,scale=3, shape=0.5)
df3<-data.table(x,y3)
df4<-merge(df1,df2,by="x")
df<-merge(df3,df4,by="x")
df_last<-melt.data.table(df,id=1,value.factor=TRUE)
p3<-ggplot(df_last, aes(x , value, colour=variable))+geom_line()+ylim(0, 4)
p3<-p3+scale_color_manual(name="Log-Logistic",values=c("red","blue","green"),labels
=c('logL(3, 0.5)','logL(3,1)','logL(3,1.5)'))
p3<-p3+ggtitle("Log-logistic Hazard Function")+xlab("Time")+ylab("Hazard")
ggsave("log_logistic_hazard.png")
```



Διάγραμμα 2.6: *Οι συναρτήσεις κινδύνου της Log-logistic κατανομής για επιλογή διαφορετικών παραμέτρων.*

Στο Διάγραμμα 2.6 απεικονίζονται οι συναρτήσεις κινδύνου, όπως προκύπτουν για τις κατανομές $Loglogistic\ (u = 3, k = 0.5)$, $Loglogistic\ (u = 3, k = 1)$ και $Loglogistic\ (u = 3, k = 1.5)$. Παρατηρούμε ότι για παράμετρο $k = 0.5$ και $k = 1$ $(k \leq 1)$, ο ρυθμός αποτυχίας μειώνεται με την πάροδο του χρόνου, ενώ για $k = 1.5$ $(k > 1)$ παρατηρούμε ότι ο ρυθμός αποτυχίας έχει μία κυρτή μορφή, όπου αρχικά αυξάνεται και έπειτα μειώνεται με την πάροδο του χρόνου.

Κατανομή Gompertz: $T \sim G(\lambda, \varphi)$

Η κατανομή Gompertz χρησιμοποιείται κυρίως για την μελέτη και περιγραφή δεδομένων που αφορούν τη διάρκεια ζωής ενηλίκων. Συνήθως βρίσκει εφαρμογή στην αναλογιστική, την δημογραφία, αλλά και στις επιστήμες της βιολογίας και της γεροντολογίας. Τα τελευταία χρόνια έχει παρατηρηθεί ότι στον κλάδο των υπολογιστών, αρκετοί επιστήμονες χρησιμοποιούν την κατανομή Gompertz για την μοντελοποίηση του ρυθμού αποτυχίας υπολογιστικών κωδίκων. Επιπλέον, έχουν γίνει προσπάθειες να περιγραφεί η συμπεριφορά του καταναλωτή με τη βοήθεια της κατανομής Gompertz στον χώρο του εμπορίου και της αγοράς.

Αν μία τυχαία μεταβλητή T ακολουθεί την κατανομή $Gompertz\ (\lambda, \varphi)$ με παραμέτρους λ, φ τότε για τις συναρτήσεις που μελετώνται στην ανάλυση επιβίωσης θα ισχύουν τα εξής (Wienke, 2011):

- η συνάρτηση πυκνότητας πιθανότητας είναι $f_T(t) = \lambda e^{\varphi t} e^{\frac{-\lambda}{\varphi}(e^{\varphi t} - 1)},\ \lambda > 0, \varphi > 0,\ t > 0.$
- η αθροιστική συνάρτηση κατανομής του χρόνου επιβίωσης είναι $F_T(t) = 1 - e^{\frac{-\lambda}{\varphi}(e^{\varphi t} - 1)},\ \lambda > 0, \varphi > 0,\ t > 0.$
- η συνάρτηση επιβίωσης είναι $S_T(t) = e^{\frac{-\lambda}{\varphi}(e^{\varphi t} - 1)},\ \lambda > 0, \varphi > 0, t > 0.$

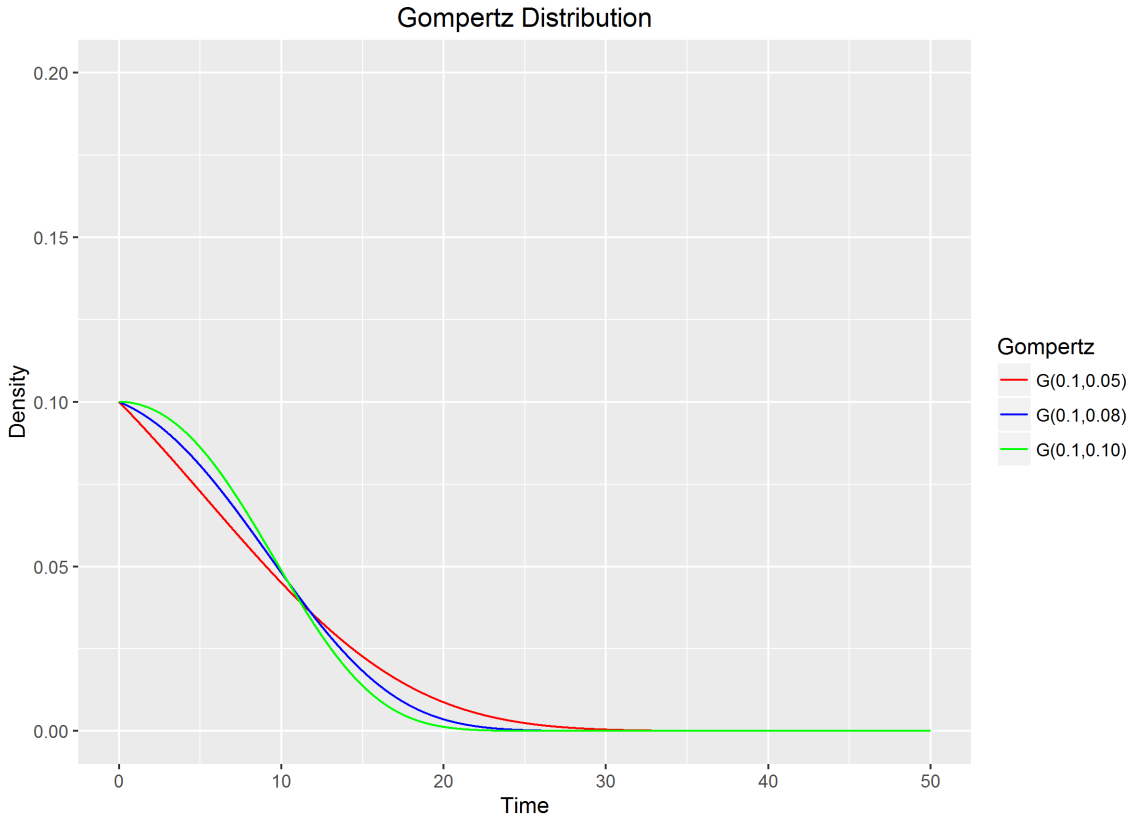
- η συνάρτηση κινδύνου είναι $h_T(t) = \lambda e^{\varphi t}, \quad \lambda > 0, \varphi > 0, \ t > 0.$
- η μέση τιμή είναι $E(T) = \frac{1}{\varphi} e^{\frac{\lambda}{\varphi}} E_1\left(\frac{\lambda}{\varphi}\right) \approx \frac{1}{\varphi} e^{\frac{\lambda}{\varphi}} \left(\frac{\lambda}{\varphi} - \ln\left(\frac{\lambda}{\varphi}\right) - \gamma\right), \lambda > 0, \varphi > 0, \ t > 0,$

όπου $E_s^j(z) = \frac{1}{\Gamma(j+1)} \int_1^\infty [\ln(x)]^j x^{-s} e^{-zx} dx$ είναι η γενικευμένη ολοκληρώσιμο - εκθετική συνάρτηση (*generalized integro - exponential function*), για την οποία ισχύει επίσης η σχέση $E_s^0(z) = E_s(z)$. Επιπλέον, όπου $\gamma \approx 0.57722$ είναι η σταθερά Euler-Mascheroni. Για μικρές τιμές των φ, λ η παραπάνω σχέση δίνει ακριβείς λύσεις για τη μέση τιμή.

Η συνάρτηση κινδύνου για αυτήν την κατανομή είναι μία αύξουσα συνάρτηση ως προς t που λαμβάνει ελάχιστη τιμή ίση με την τιμή της παραμέτρου λ. Για σταθερό λ, όσο μικρότερη είναι η τιμή της παραμέτρου φ, τόσο μικρότερος είναι ο ρυθμός αύξησης της συνάρτησης κινδύνου, αλλά και οι τιμές αυτής. Για μικρότερες τιμές της φ λαμβάνουμε μικρότερες τιμές για την συνάρτηση κινδύνου. Η Εκθετική κατανομή είναι μία ειδική περίπτωση της κατανομής Gompertz για παράμετρο $\varphi = 0$.

Χρησιμοποιώντας τον παρακάτω κώδικα δημιουργούμε το Διάγραμμα 2.7.

```
##Gompertz
x<-seq(0,50,0.001)
#density
d.gompertz<-function(x,l,f){

  l*exp(f*x)*exp((-l/f)*(exp(f*x)-1))

}
h.gompertz<-function(x,l,f){
  l*exp(f*x)
 }
#propability function
y1=d.gompertz(x, l=0.1, f=0.05)
df1<-data.table(x,y1)
y2<-d.gompertz(x,l=0.1, f=0.08)
df2<-data.table(x,y2)
y3<-d.gompertz(x,l=0.1, f=0.10)
df3<-data.table(x,y3)
df4<-merge(df1,df2,by="x")
df<-merge(df4,df3,by="x")
df_last<-melt.data.table(df,id=1,value.factor=TRUE)
p4<-ggplot(df_last, aes(x , value, colour=variable))+geom_line()+ylim(0, 0.2)+xlim(0,50)
p4<-p4+scale_color_manual(name="Gompertz",values=c("red","blue","green"),labels          =
c('G(0.1,0.05)','G(0.1,0.08)','G(0.1,0.10)'))
p4<-p4+ggtitle("Gompertz Distribution")+xlab("Time")+ylab("Density")
ggsave("Gompertz_density.png", plot=p4)
```

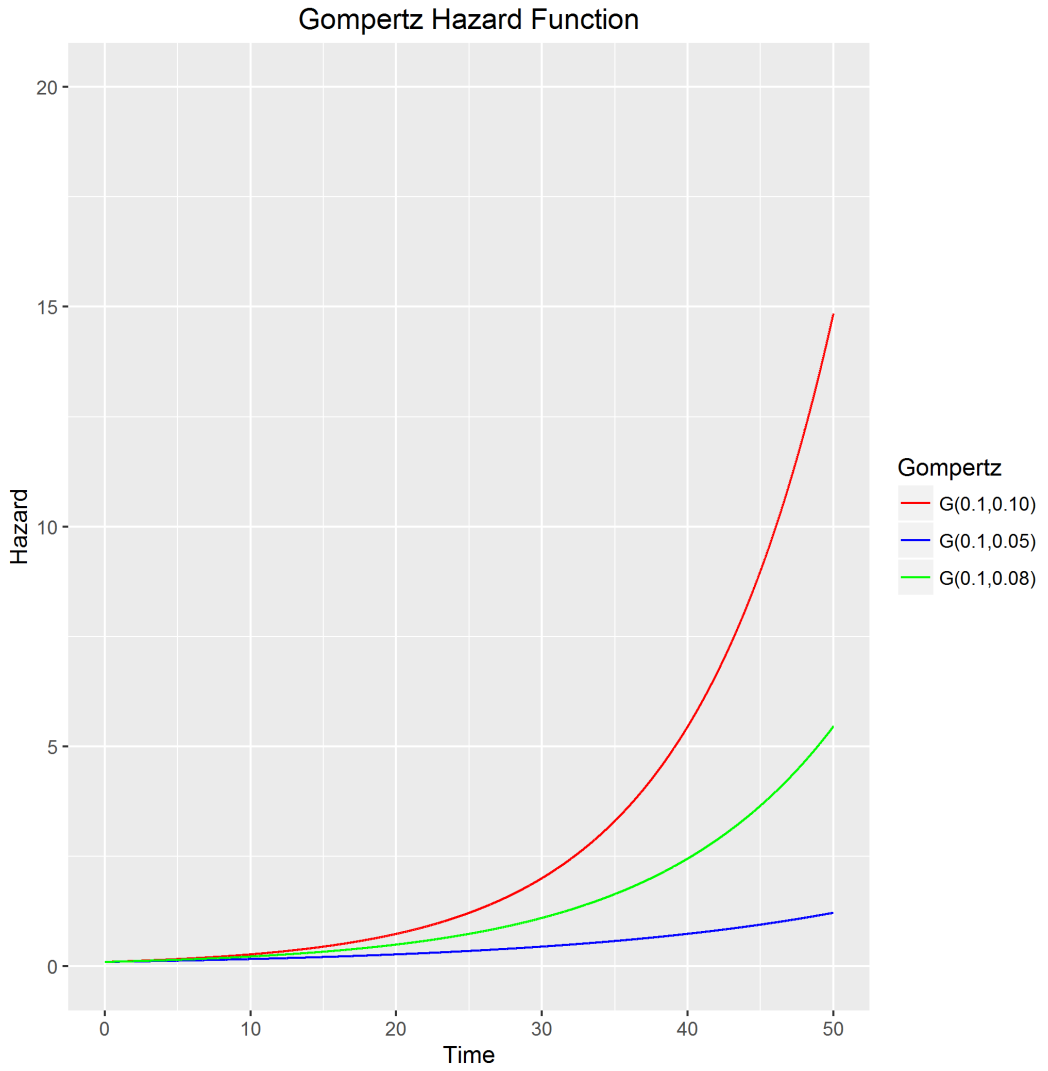


Διάγραμμα 2.7: *Οι συναρτήσεις πυκνότητας πιθανότητας για την κατανομή Gompertz με διαφορετικές τιμές στις παραμέτρους.*

Στο Διάγραμμα 2.7 απεικονίζονται οι συναρτήσεις πυκνότητας πιθανότητας των κατανομών $Gompertz\ (\lambda = 0.1,\ \varphi = 0.05),\ Gompertz\ (\lambda = 0.1,\ \varphi = 0.08)$ και $Gompertz\ (\lambda = 0.1,\ \varphi = 0.10)$.

Χρησιμοποιώντας τον παρακάτω κώδικα δημιουργούμε το Διάγραμμα 2.8.

```
#hazard function
y1=h.gompertz(x, l=0.1, f=0.05)
df1<-data.table(x,y1)
y2<-h.gompertz(x,l=0.1, f=0.08)
df2<-data.table(x,y2)
y3<-h.gompertz(x,l=0.1, f=0.10)
df3<-data.table(x,y3)
df4<-merge(df1,df2,by="x")
df<-merge(df3,df4,by="x")
df_last<-melt.data.table(df,id=1,value.factor=TRUE)
p5<-ggplot(df_last, aes(x , value, colour=variable))+geom_line()+ylim(0, 20)+xlim(0,50)
p5 <- p5 + scale_color_manual(name = "Gompertz", values = c("red", "blue", "green"),
labels = c('G(0.1, 0.10)', 'G(0.1, 0.05)', 'G(0.1, 0.08)'))
p5<-p5+ggtitle("Gompertz Distribution")+xlab("Time")+ylab("Hazard")
ggsave("Gompertz_hazard.png", plot=p5)
```



Διάγραμμα 2.8: *Οι συναρτήσεις κινδύνου της Gompertz κατανομής χρησιμοποιώντας διαφορετικές τιμές για τις παραμέτρους.*

Στο Διάγραμμα 2.8 απεικονίζονται οι συναρτήσεις κινδύνου, όπως προκύπτουν για τις κατανομές $Gompertz\ (\lambda = 0.1,\ \varphi = 0.05)$, $Gompertz\ (\lambda = 0.1,\ \varphi = 0.08)$ και $Gompertz\ (\lambda = 0.1,\ \varphi = 0.10)$. Παρατηρούμε ότι η συνάρτηση κινδύνου σε αυτήν την περίπτωση είναι μία αύξουσα συνάρτηση που λαμβάνει ελάχιστη τιμή ίση με την τιμή της παραμέτρου $\lambda = 0.1$. Επιπλέον, για σταθερό $\lambda = 0.1$, η συνάρτηση κινδύνου για $\varphi = 0.05$ έχει μικρότερο ρυθμό αύξησης συγκριτικά με $\varphi = 0.08$ και $\varphi = 0.10$. Γενικά, επιβεβαιώνεται ότι όσο μικρότερη είναι η τιμή της παραμέτρου φ, τόσο μικρότερος είναι ο ρυθμός αύξησης της συνάρτησης κινδύνου, αλλά και οι τιμές αυτής. Επίσης, παρατηρούμε ότι για μικρότερες τιμές της φ λαμβάνουμε μικρότερες τιμές για την συνάρτηση κινδύνου.

Λογαριθμοκανονική κατανομή

Όταν μία τυχαία μεταβλητή X ακολουθεί την κατανομή $N(m,\ s^2)$, τότε η τυχαία μεταβλητή $T = \exp(X)$ ακολουθεί τη Λογαριθμοκανονική κατανομή $logN\ (m,\ s^2)$ με παραμέτρους $m,\ s^2$.

Ισχύουν οι παρακάτω σχέσεις για την Λογαριθμοκανονική κατανομή (Wienke, 2011):

- η συνάρτηση πυκνότητας πιθανότητας είναι $f_T(t) = \frac{1}{\sqrt{2\pi}st} e^{\frac{-(logt-m)^2}{2s^2}}, m \in \mathcal{R}, s > 0, t > 0.$

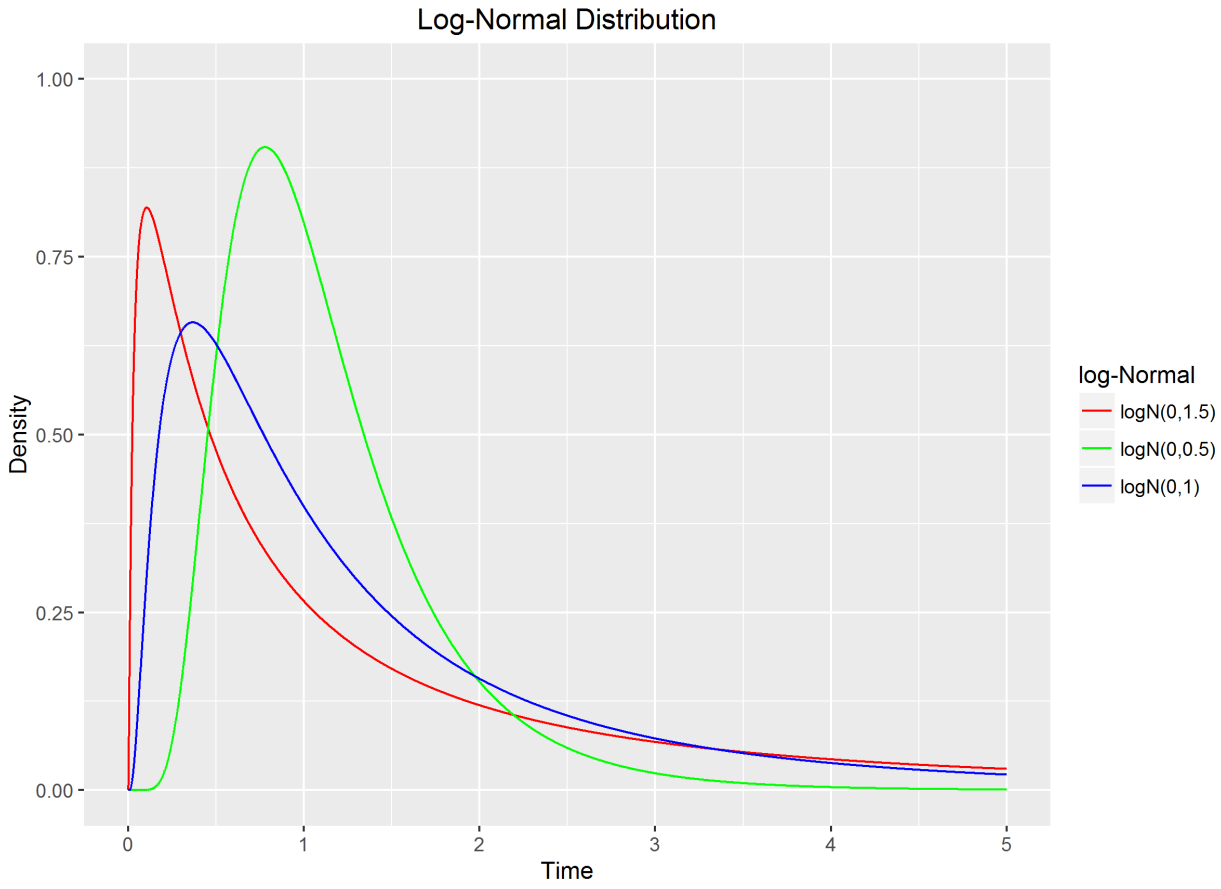
- η αθροιστική συνάρτηση κατανομής του χρόνου επιβίωσης είναι $F_T(t) = \Phi\left(\frac{logt-m}{s}\right), m \in \mathcal{R}, s > 0, t > 0.$
- η συνάρτηση επιβίωσης είναι $S_T(t) = 1 - \Phi\left(\frac{logt-m}{s}\right), m \in \mathcal{R}, s > 0, t > 0.$
- η συνάρτηση κινδύνου είναι $h_T(t) = \frac{\frac{1}{st}\Phi(\frac{logt-m}{s})}{1-\Phi(\frac{logt-m}{s})}, m \in \mathcal{R}, s > 0, t > 0.$
- η μέση τιμή είναι $E(T) = e^{m+\frac{s^2}{2}}, m \in \mathcal{R}, s > 0, t > 0.$
- η διασπορά είναι $V(T) = e^{m+\frac{s^2}{2}}\left(e^{s^2} - 1\right), m \in \mathcal{R}, s > 0, t > 0.$

Στα παραπάνω η συνάρτηση $\Phi(\cdot)$ αποτελεί την συνάρτηση κατανομής της τυποποιημένης κανονικής κατανομής.

Η συνάρτηση κινδύνου για αυτήν την κατανομή έχει ενδιαφέρουσα συμπεριφορά. Αρχικά, για $t = 0$ λαμβάνει τιμή 0, εν συνεχεία αυξάνεται έως ένα ολικό μέγιστο και μειώνεται τείνοντας προς το μηδέν, καθώς ο χρόνος τείνει στο άπειρο. Αυτό καθιστά την εν λόγω κατανομή ακατάλληλη για δεδομένα επιβίωσης ανθρώπων ιδιαίτερα για ανθρώπους μεγάλης ηλικίας. Ωστόσο, αν εξετάζουμε δεδομένα επιβίωσης νεότερων ανθρώπων ή βρεφών φαίνεται να είναι μία καλή επιλογή. Στον κλάδο της Φαρμακευτικής η εν λόγω κατανομή χρησιμοποιείται κυρίως για την μοντελοποίηση δεδομένων που εκφράζουν χρόνο αντίδρασης ενός ατόμου σε μια ασθένεια (*latent time*), για παράδειγμα τον χρόνο από την στιγμή που κατέβαλλε τον ασθενή η ασθένεια μέχρι την εμφάνιση των πρώτων συμπτωμάτων.

Χρησιμοποιώντας τον παρακάτω κώδικα δημιουργούμε το Διάγραμμα 2.9.

```
###Log-Normal
#propability function
x<-seq(0,5,0.001)
y1=dlnorm(x, 0,0.5)
df1<-data.table(x,y1)
y2<-dlnorm(x,0,1)
df2<-data.table(x,y2)
y3<-dlnorm(x,0,1.5)
df3<-data.table(x,y3)
df4<-merge(df1,df2,by = "x")
df<-merge(df3,df4,by = "x")
df_last<-melt.data.table(df, id = 1,value.factor = TRUE)
p6<-ggplot(df_last, aes(x , value, colour = variable))+ geom_line()+ ylim(0, 1)
p6<-p6+scale_color_manual(name="log-Normal",values=c("red","green","blue"),labels =
c('logN(0,1.5)','logN(0,0.5)','logN(0,1)'))
p6<-p6 + ggtitle("Log-Normal Distribution")+ xlab("Time")+ ylab("Density")
ggsave("logNormal_density.png", plot=p6)
```

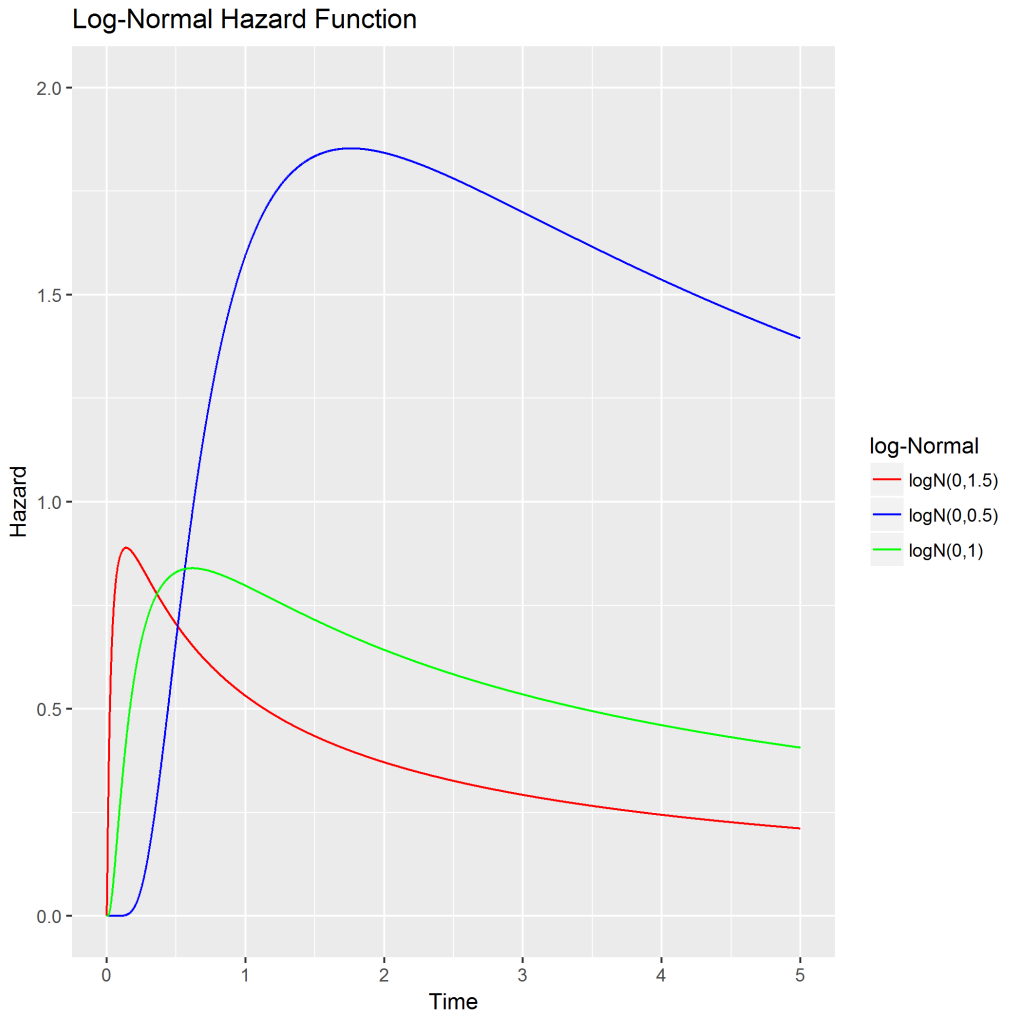


Διάγραμμα 2.9: *Οι συναρτήσεις πυκνότητας πιθανότητας για την Log-Normal κατανομή επιλέγοντας διαφορετικές παραμέτρους.*

Στο Διάγραμμα 2.9 απεικονίζονται οι συναρτήσεις πυκνότητας πιθανότητας των κατανομών $LogN\ (m = 0,\ s^2 = 0.25)$, $LogN\ (m = 0,\ s^2 = 1)$ και $LogN\ (m = 0,\ s^2 = 2.25)$. Στο περιθώριο αυτού και του επομένου γραφήματος δίνονται οι τυπικές αποκλίσεις των κατανομών.

Χρησιμοποιώντας τον παρακάτω κώδικα δημιουργούμε το Διάγραμμα 2.10.

```
#hazard function
#install.packages("eha")
library(eha)
y1 = hlnorm(x, 0, 0.5)
df1 <- data.table(x, y1)
y2 <- hlnorm(x,0,1)
df2 <- data.table(x, y2)
y3 <- hlnorm(x, 0, 1.5)
df3 <- data.table(x, y3)
df4 <- merge(df1,df2,by = "x")
df <- merge(df3,df4,by="x")
df_last <- melt.data.table(df, id = 1, value.factor = TRUE)
p7 <- ggplot(df_last, aes(x , value, colour=variable))+geom_line()+ylim(0, 2)
p7 <- p7 + scale_color_manual(name = "log-Normal", values =
c("red","blue","green"),labels = c('logN(0,1.25)','logN(0,0.5)','logN(0,1)'))
p7<-p7 + ggtitle("Log-Normal Hazard Function") + xlab("Time") + ylab("Hazard")
ggsave("logNormal_hazard.png", plot = p7)
```



Διάγραμμα 2.10: *Οι συναρτήσεις κινδύνου της Log-Normal κατανομής χρησιμοποιώντας διαφορετικές παραμέτρους.*

Στο Διάγραμμα 2.10 απεικονίζονται οι συναρτήσεις κινδύνου, όπως προκύπτουν για τις κατανομές $LogN\ (m = 0,\ s^2 = 0.25)$, $LogN\ (m = 0,\ s^2 = 1)$ και $LogN\ (m = 0,\ s^2 = 2.25)$. Η συνάρτηση κινδύνου και για τις τρείς επιλογές παραμέτρων έχει την ίδια συμπεριφορά. Αρχικά, για $t = 0$ λαμβάνει τιμή 0, εν συνεχεία αυξάνεται έως ένα ολικό μέγιστο και μειώνεται τείνοντας προς το μηδέν, καθώς ο χρόνος τείνει στο άπειρο. Για μεγαλύτερη διασπορά $s^2 = 2.25$, ο ρυθμός αύξησης είναι μεγαλύτερος, ενώ φαίνεται να φτάνει σε μικρότερο ολικό μέγιστο. Όσο μικρότερη είναι η διασπορά, θεωρώντας σταθερή τη μέση τιμή, τόσο μικρότερος είναι ο ρυθμός αύξησης του ρυθμού κινδύνου και τόσο μεγαλύτερη είναι η τιμή του ολικού μεγίστου, που λαμβάνει η συνάρτηση κινδύνου.

Κατανομή Γάμμα: $G(\alpha, \beta)$

Η κατανομή Γάμμα αποτελεί άλλη μία γενίκευση της Εκθετικής κατανομής. Λαμβάνει δύο παραμέτρους α (παράμετρος σχήματος) και β (παράμετρος κλίμακας) και οι συναρτήσεις που εξετάζονται στην ανάλυση επιβίωσης δίνονται από τις παρακάτω σχέσεις:

- η συνάρτηση πυκνότητας πιθανότητας είναι $f_T(t) = \frac{\beta^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\beta t}, t > 0, \alpha > 0, \beta > 0.$

- η αθροιστική συνάρτηση κατανομής του χρόνου αποτυχίας είναι

  $F_T(t) = \int_0^t \frac{\beta^\alpha}{\Gamma(\alpha)} s^{\alpha-1} e^{-\beta s}\, ds = \frac{1}{\Gamma(\alpha)} \int_0^{\beta t} s^{\alpha-1} e^{-s}\, ds, t > 0, \alpha > 0, \beta > 0.$

- η συνάρτηση επιβίωσης είναι η $S_T(t) = \int_t^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} s^{\alpha-1} e^{-\beta s}\, ds,\ t > 0, \alpha > 0,\ \beta > 0.$

- η συνάρτηση κινδύνου είναι η $h_T(t) = \frac{\beta^\alpha t^{\alpha-1} e^{-\beta t}}{\left(1 - \frac{1}{\Gamma(\alpha)} \int_0^{\beta t} s^{\alpha-1} e^{-s} ds\right) \cdot \Gamma(\alpha)}$, $t > 0, \alpha > 0,\ \beta > 0.$

- η μέση τιμή είναι $E(T) = \frac{a}{\beta}$, $\alpha > 0,\ \beta > 0.$

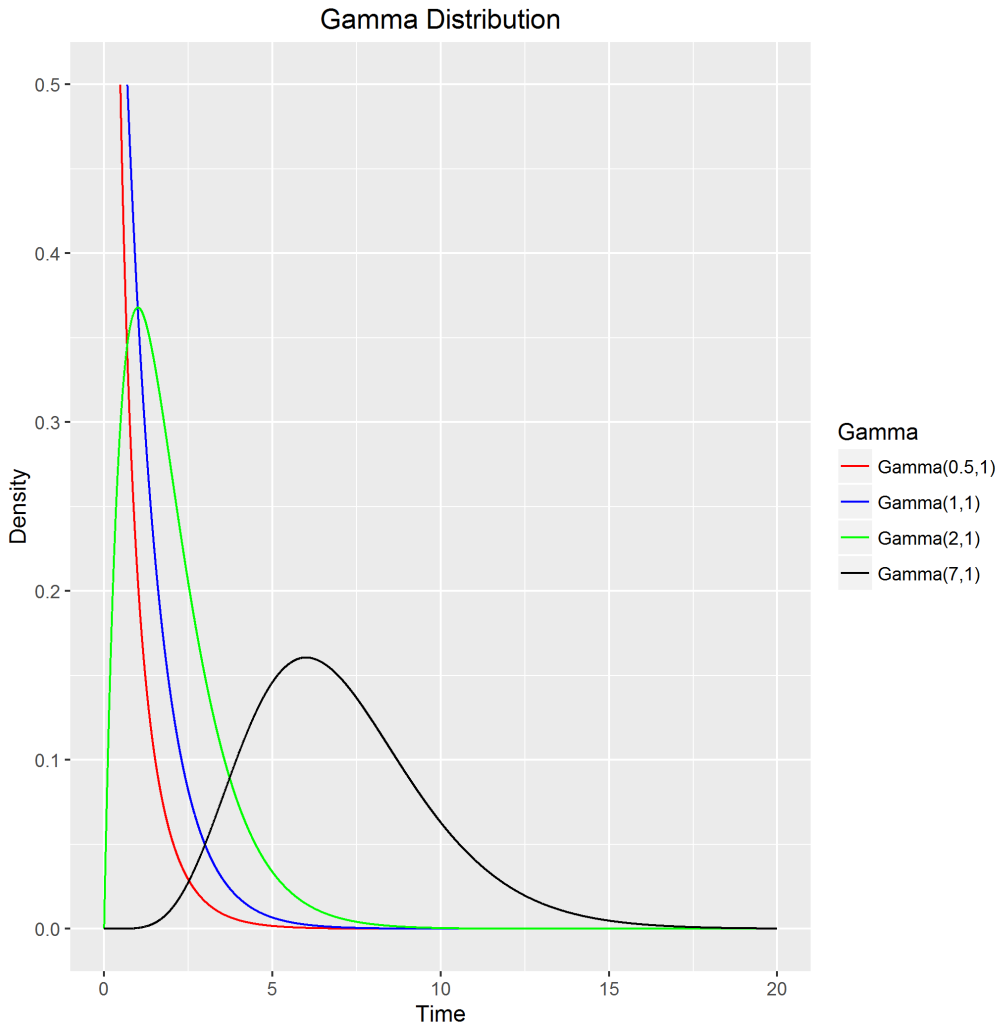
- η διασπορά είναι $V(T) = \frac{\alpha}{\beta^2}, \alpha > 0,\ \beta > 0.$

Παρατηρούμε ότι οι συναρτήσεις επιβίωσης και κινδύνου δεν υπολογίζονται σε κλειστή μορφή και αυτό αποτελεί μία επιπρόσθετη δυσκολία στην επιλογή της εν λόγω κατανομής για την ανάλυση δεδομένων επιβίωσης, καθώς υπάρχει δυσκολία στον υπολογισμό των εκτιμήσεων των παραμέτρων. Ωστόσο, με κατάλληλες αριθμητικές μεθόδους είναι εφικτός ο υπολογισμός αυτών.

Για $\alpha = 1$ η κατανομή Γάμμα συμπίπτει με την Εκθετική κατανομή με παράμετρο β.

Χρησιμοποιώντας τον παρακάτω κώδικα δημιουργούμε το Διάγραμμα 2.11.

```
#Gamma distribution
x <- seq(0,5,0.001)
h.gamma <- function(x , shape, scale)
{
  f.g <- dgamma(x, shape, scale)
  S.g <- 1 - pgamma(x, shape, scale)
  return(f.g/S.g)
}
###Gamma
#propability function
x<-seq(0,20,0.001)
y1=dgamma(x,shape=0.5,scale=1)
df1<-data.table(x,y1)
y2<-dgamma(x,shape=1, scale=1)
df2<-data.table(x,y2)
y3<-dgamma(x,shape=2, scale=1)
df3<-data.table(x,y3)
y4<-dgamma(x,shape=7, scale=1)
df4<-data.table(x,y4)
df5<-merge(df1,df2,by="x")
df6<-merge(df3,df4,by="x")
df<-merge(df5,df6,by="x")
df_last<-melt.data.table(df,id=1,value.factor=TRUE)
p8<-ggplot(df_last, aes(x , value, colour=variable))+geom_line()+ylim(0, 0.5)
p8<-p8+scale_color_manual(name="Gamma",values=c("red","blue","green","black"),labels     =
c('Gamma(0.5,1)','Gamma(1,1)','Gamma(2,1)','Gamma(7,1)'))
p8<-p8+ggtitle("Gamma Distribution")+xlab("Time")+ylab("Density")
ggsave("Gamma_density.png")
```



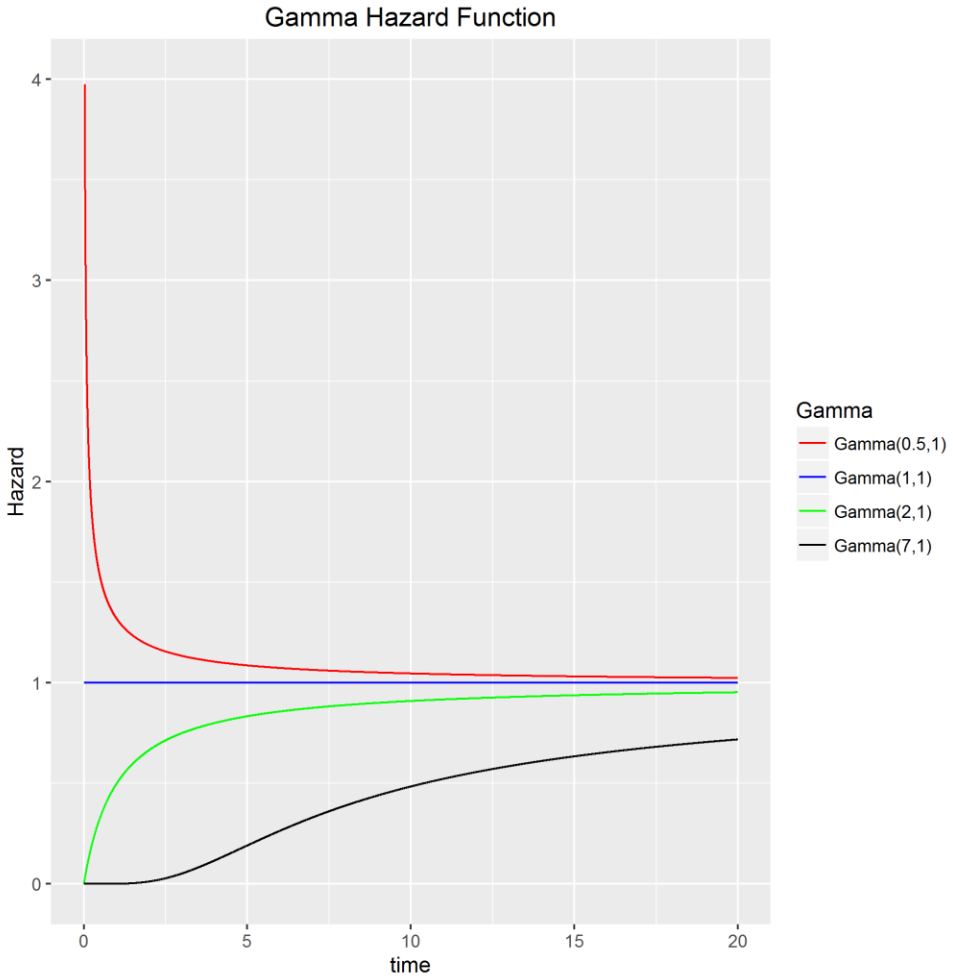
Διάγραμμα 2.11: *Οι συναρτήσεις πυκνότητας πιθανότητας με διαφορετική επιλογή παραμέτρων.*

Στο Διάγραμμα 2.11 απεικονίζονται οι συναρτήσεις πυκνότητας πιθανότητας των κατανομών $Gamma\ (a = 0.5,\ \beta = 1)$, $Gamma\ (a = 1,\ \beta = 1)$, $Gamma\ (a = 2,\ \beta = 1)$ και $Gamma\ (a = 7,\ \beta = 1)$.

Χρησιμοποιώντας τον παρακάτω κώδικα δημιουργούμε το Διάγραμμα 2.12.

```
#hazard function
x<-seq(0,20,0.001)
y1=h.gamma(x,shape=0.5,scale=1)
df1<-data.table(x,y1)
y2<-h.gamma(x,shape=1, scale=1)
df2<-data.table(x,y2)
y3<-h.gamma(x,shape=2, scale=1)
df3<-data.table(x,y3)
y4<-h.gamma(x,shape=7, scale=1)
df4<-data.table(x,y4)
df5<-merge(df1,df2,by="x")
df6<-merge(df3,df4,by="x")
df<-merge(df5,df6,by="x")
df_last<-melt.data.table(df,id=1,value.factor=TRUE)
p9<-ggplot(df_last, aes(x , value, colour=variable))+geom_line()+ylim(0, 4)
p9<-p9+scale_color_manual(name="Gamma",values=c("red","blue","green","black"),labels=
c('Gamma(0.5,1)','Gamma(1,1)','Gamma(2,1)','Gamma(7,1)'))
```

```
p9<-p9+ggtitle("Gamma Hazard Function")+xlab("Time")+ylab("Hazard")
ggsave("Gamma_hazard.png")
```



Διάγραμμα 12: *Οι συναρτήσεις κινδύνου της Gamma κατανομής για διαφορετική επιλογή παραμέτρων.*

Στο Διάγραμμα 2.12 απεικονίζονται οι συναρτήσεις κινδύνου για τις κατανομές $Gamma\ (a = 0.5,\ \beta = 1)$, $Gamma\ (a = 1,\ \beta = 1)$, $Gamma\ (a = 2,\ \beta = 1)$ και $Gamma\ (a = 7,\ \beta = 1)$.

Κατανομή Pareto

Η κατανομή Pareto περιλαμβάνει δύο παραμέτρους α και c και η συνάρτηση πυκνότητας πιθανότητας, η συνάρτηση κατανομής, η συνάρτηση επιβίωσης, η συνάρτηση κινδύνου, καθώς και η μέση τιμή και διασπορά αυτής δίνονται από τις παρακάτω σχέσεις.

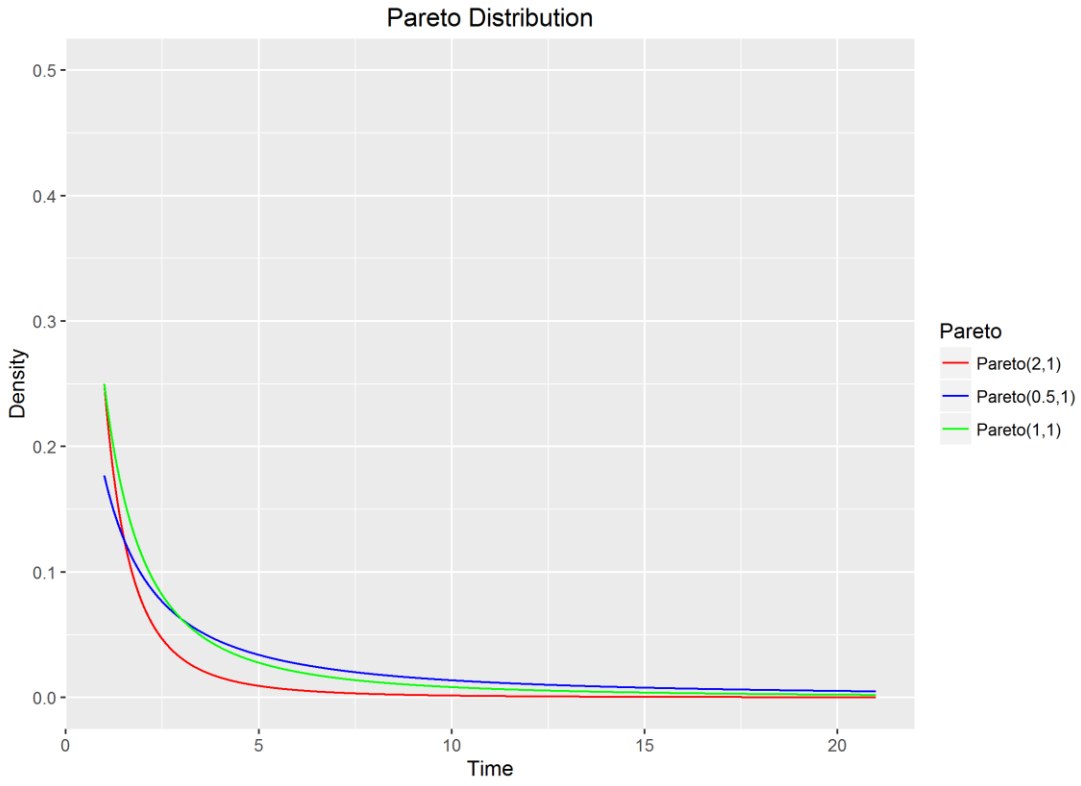
- η συνάρτηση πυκνότητας πιθανότητας είναι $f_T(t) = a\frac{c^a}{t^{a+1}}, \quad t \geq c, \alpha > 0, c > 0.$
- η αθροιστική συνάρτηση κατανομής του χρόνου επιβίωσης είναι $F_T(t) = 1 - \left(\frac{c}{t}\right)^a,\ t \geq c, \alpha > 0, c > 0.$
- η συνάρτηση επιβίωσης είναι $S_T(t) = \left(\frac{c}{t}\right)^a, t \geq c, \alpha > 0, c > 0.$
- η συνάρτηση κινδύνου είναι $h_T(t) = \frac{a\frac{c^a}{t^{a+1}}}{\left(\frac{c}{t}\right)^a} = a\frac{1}{t}, t \geq c, \alpha > 0, c > 0.$

- η μέση τιμή είναι $E(T) = \frac{ac}{a-1}$, $a > 1, c > 0$.
- η διασπορά είναι $V(T) = \frac{ac^2}{(a-1)^2(a-2)}$, $a > 2, c > 0$.

Η κατανομή Pareto επίσης μπορεί να θεωρηθεί ως μία μίξη Εκθετικής κατανομής με Γάμμα. Συχνά εφαρμόζεται για την περιγραφή της διανομής του πληθυσμού των πόλεων, τις διακυμάνσεις των τιμών των μετοχών και γενικότερα χρησιμοποιείται σε κοινωνικοοικονομικές μελέτες. Επιπροσθέτως, η κατανομή Pareto βρίσκει εφαρμογή στον ασφαλιστικό τομέα, αλλά και για την φροντίδα των απαιτήσεων ενός χαρτοφυλακίου (Wienke, 2011).

Χρησιμοποιώντας τον παρακάτω κώδικα δημιουργούμε το Διάγραμμα 2.13.

```
#Pareto distribution
#propability function
install.packages("actuar")
library(actuar)
x<-seq(0,20,0.001)
scale<-1
x<-x+scale
y1=dpareto(x,shape=0.5,scale=1)
df1<-data.table(x,y1)
y2<-dpareto(x,shape=1, scale=1)
df2<-data.table(x,y2)
y3<-dpareto(x,shape=2, scale=1)
df3<-data.table(x,y3)
df5<-merge(df1,df2,by="x")
df<-merge(df3,df5,by="x")
df_last<-melt.data.table(df,id=1,value.factor=TRUE)
p10<-ggplot(df_last, aes(x , value, colour=variable))+geom_line()+ylim(0, 0.5)
p10<-p10+scale_color_manual(name="Pareto",values=c("red","blue","green"),labels          =
c('Pareto(2,1)','Pareto(0.5,1)','Pareto(1,1)'))
p10<-p10+ggtitle("Pareto Distribution")+xlab("Time")+ylab("Density")
ggsave("Pareto_density.png", plot=p10)
```

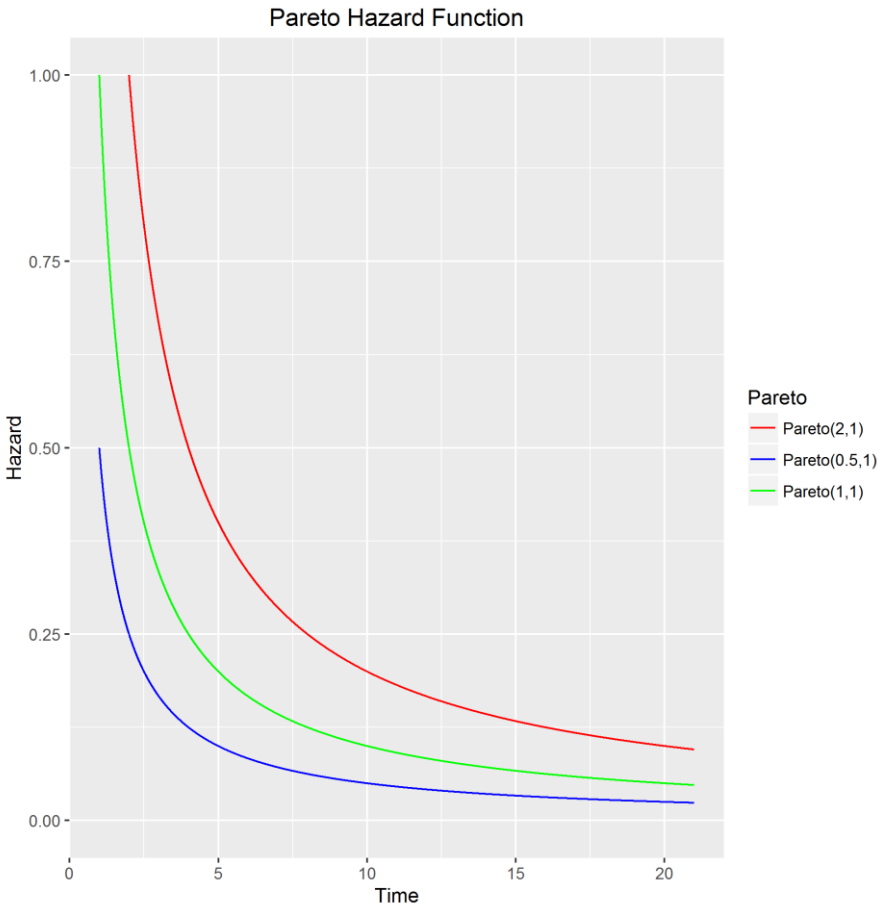


Διάγραμμα 2.13: *Οι συναρτήσεις πυκνότητας πιθανότητας της κατανομής Pareto με διαφορετικές παραμέτρους.*

Στο Διάγραμμα 2.13 απεικονίζονται οι συναρτήσεις πυκνότητας πιθανότητας για τις κατανομές $Pareto\ (shape = 2, scale = 1)$, $Pareto\ (shape = 0.5, scale = 1)$ και $Pareto\ (shape = 1, scale = 1)$.

Χρησιμοποιώντας τον παρακάτω κώδικα δημιουργούμε το Διάγραμμα 2.14.

```
#hazard function
h.pareto<-function(x, shape){
  shape*(1/x)
}
x<-seq(0,20,0.001)
scale<-1
x<-x+scale
y1=h.pareto(x,shape=0.5)
df1<-data.table(x,y1)
y2<-h.pareto(x,shape=1)
df2<-data.table(x,y2)
y3<-h.pareto(x,shape=2)
df3<-data.table(x,y3)
df5<-merge(df1,df2,by="x")
df<-merge(df3,df5,by="x")
df last<-melt.data.table(df,id=1,value.factor=TRUE)
p11<-ggplot(df_last, aes(x , value, colour=variable))+geom_line()+ylim(0, 1)
p11<-p11+scale_color_manual(name="Pareto",values=c("red","blue","green","black"),labels  =
c('Pareto(2,1)','Pareto(0.5,1)','Pareto(1,1)'))
p11<-p11+ggtitle("Pareto Hazard Function")+xlab("Time")+ylab("Hazard")
ggsave("Pareto_hazard.png", plot=p11)
```



Διάγραμμα 2.14: *Οι συναρτήσεις κινδύνου για την κατανομή Pareto με διαφορετική επιλογή παραμέτρων.*

Στο Διάγραμμα 2.14 απεικονίζονται οι συναρτήσεις κινδύνου για τις κατανομές $Pareto\,(\,shape = 2,\ scale = 1)$, $Pareto\,(\,shape = 0.5,\ scale = 1)$ και $Pareto\,(\,shape = 1,\ scale = 1)$.

2.2 Μη Παραμετρικές Μέθοδοι Εκτίμησης των Συναρτήσεων Επιβίωσης και Κινδύνου

2.2.1 Εισαγωγή

Στην προηγούμενη ενότητα είδαμε τις πιο συνηθισμένες επιλογές της κατανομής του χρόνου επιβίωσης, που αναφέρονται στη βιβλιογραφία. Με βάση αυτές τις κατανομές είδαμε σε κάθε περίπτωση πως μπορεί να υπολογιστεί η συνάρτηση επιβίωσης και η συνάρτηση κινδύνου. Εναλλακτικά, οι εν λόγω συναρτήσεις μπορούν να εκτιμηθούν μη παραμετρικά με τη βοήθεια του δείγματος, που διαθέτουμε χωρίς να υποθέσουμε μία κατανομή για τον χρόνο επιβίωσης. Σε αυτή την ενότητα λοιπόν μεταξύ άλλων, θα αναφερθούμε στον εκτιμητή Kaplan-Meier, που χρησιμοποιείται για την εκτίμηση της συνάρτησης επιβίωσης, αλλά και τον εκτιμητή Nelson-Aalen, που χρησιμοποιείται για την εκτίμηση της αθροιστικής συνάρτησης κινδύνου.

2.2.2 Η Μέθοδος Kaplan-Meier

Στην περίπτωση μη λογοκριμένων δεδομένων μπορεί να χρησιμοποιηθεί η εμπειρική κατανομή (*empirical estimator*) για την εκτίμηση της συνάρτησης επιβίωσης, η οποία δίνεται από την εξής σχέση:

$$\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} I\{t_i > t\},$$

όπου $I(\cdot)$ είναι η δείκτρια συνάρτηση, η οποία λαμβάνει την τιμή 1 για τις τιμές που πληρούν την σχέση $t_i > t$ και μηδέν οπουδήποτε αλλού. Συνεπώς, ο εκτιμητής είναι ουσιαστικά το ποσοστό αυτών που έχουν επιβιώσει μέχρι και την χρονική στιγμή t.

Ωστόσο, αναπτύχθηκε και ο εκτιμητής Kaplan-Meier (ή *product limit estimator*), ο οποίος είναι ουσιαστικά μία επέκταση του εμπειρικού εκτιμητή για δεδομένα που περιέχουν και λογοκριμένες τιμές. Η εκτιμήτρια Kaplan-Meier δίνει εκτιμήσεις των τιμών της συνάρτησης επιβίωσης στους μη λογοκριμένους χρόνους αλλά για τον υπολογισμό της λαμβάνονται υπόψην και οι ακριβείς χρόνοι επιβίωσης και οι λογοκριμένοι χρόνοι.

Αν θεωρήσουμε τους διατεταγμένους χρόνους επιβίωσης

$$t_{(1)} < t_{(2)} < t_{(3)} < \cdots < t_{(n)}$$

τότε η εκτίμηση Kaplan-Meier της συνάρτησης επιβίωσης για τους μη λογοκριμένους χρόνους δίνεται από τη σχέση:

$$\hat{S}(t) = \begin{cases} 1, & t_{(1)} > t \\ \prod_{i:t_{(i)}<t} \left(1 - \frac{d_i}{n_i}\right), & t_{(1)} \leq t \end{cases}$$

όπου d_i είναι ο αριθμός αυτών που απεβίωσαν τη χρονική στιγμή t_i και n_i είναι το σύνολο των επιζώντων ακριβώς πριν τη χρονική στιγμή $t_{(i)}$.

Παράδειγμα 2.1

Ας υποθέσουμε ότι έχουμε δεδομένα επιβίωσης καρκινοπαθών ασθενών για 6 μηνιαίες περιόδους. Οι χρόνοι επιβίωσης δίνονται στον Πίνακα 2.1:

6	**6**	**6**	**7**	**7***
9	10	10*	10*	10*
12	12*	12*	16	23
32				

Πίνακας 2.1: *Οι παρατηρήσεις του χρόνου επιβίωσης (μήνες). Με * δηλώνονται οι δεξιά λογοκριμένες παρατηρήσεις.*

Τότε για την εκτίμηση της συνάρτησης επιβίωσης με τη βοήθεια του εκτιμητή Kaplan-Meier δημιουργούμε τον Πίνακα 2.2.

t	d_i	t_i	c_i	n_i	$\widehat{S}(t)$
$0 \le t < 6$					1
$6 \le t < 7$	3	3	0	16	$1 - \frac{3}{16} = 0.8125$
$7 \le t < 9$	1	1	1	13	$0.8125 * \left(1 - \frac{1}{13}\right) = 0.75$
$9 \le t < 10$	1	1	0	11	$0.75 * \left(1 - \frac{1}{11}\right) = 0.682$
$10 \le t < 12$	1	1	3	10	$0.682 * \left(1 - \frac{1}{10}\right) = 0.614$
$12 \le t < 16$	1	1	2	6	$0.614 * \left(1 - \frac{1}{6}\right) = 0.512$
$16 \le t < 23$	1	1	0	3	$0.512 * \left(1 - \frac{1}{3}\right) = 0.384$
$23 \le t < 32$	1	1	0	2	$0.384 * \left(1 - \frac{1}{2}\right) = 0.256$
$32 \le t < 36$	1	1	0	1	$0.256 * \left(1 - \frac{1}{1}\right) = 0.000$

Πίνακας 2.2: *Υπολογισμός του Kaplan-Meier εκτιμητή για τα δεδομένα του Παραδείγματος 2.1.*

Η εκτιμήτρια Kaplan-Meier αποτελεί μία βηματική συνάρτηση και αυτό εκφράζεται και από τη γραφική της αναπαράσταση. Οι εκτιμήσεις απεικονίζονται στην καμπύλη επιβίωσης όπου οι μη λογοκριμένοι χρόνοι είναι χωρισμένοι σε διαστήματα και σε κάθε ένα από αυτά αντιστοιχεί μία τιμή για την συνάρτηση επιβίωσης.

Η μέθοδος Kaplan-Meier μας δίνει μία εκτίμηση για την συνάρτηση επιβίωσης. Ωστόσο, αυτό δεν είναι αρκετό, καθώς θα πρέπει να διερευνήσουμε την ακρίβεια της εκτιμήτριας. Ένα μέτρο ακρίβειας μίας εκτιμήτριας αντικατοπτρίζεται στο τυπικό της σφάλμα. Ουσιαστικά, μικρές τιμές του τυπικού σφάλματος υποδηλώνουν μικρές διακυμάνσεις της εκτιμήτριας από δείγμα σε δείγμα. Επιπλέον, με τη βοήθεια του τυπικού σφάλματος μπορούν να κατασκευαστούν και κατάλληλα διαστήματα εμπιστοσύνης για την ως προς εκτίμηση ποσότητα.

Η εκτίμηση Kaplan-Meier για τη συνάρτηση επιβίωσης για μία τυχαία τιμή t στο χρονικό διάστημα $[t_{(\kappa)}, t_{(\kappa+1)}]$ μπορεί να γραφεί στη μορφή (Collett, 2003):

$$\hat{S}(t) = \prod_{j=1}^{\kappa} \hat{p}_j,$$

όπου $\kappa = 1, \dots, n$, $p_j = (n_j - d_j)/n_j$ είναι η εκτίμηση της πιθανότητας ότι μία μονάδα του πληθυσμού επέζησε στο χρονικό διάστημα $\left[t_{(j)}, t_{(j+1)}\right]$, $j = 1, \dots, n$.

Λογαριθμίζοντας την παραπάνω σχέση λαμβάνουμε ότι:

$$\log(\hat{S}(t)) = \sum_{j=1}^{\kappa} \log(\hat{p}_j),$$

με διασπορά, που δίνεται από τη σχέση:

$$\mathrm{Var}\{\log(\hat{S}(t))\} = \sum_{j=1}^{\kappa} \mathrm{Var}\{\log(\widehat{p_j})\}. \tag{2.11}$$

Ο αριθμός αυτών που επιβίωσαν στο διάστημα $[t_{(j)}, t_{(j+1)}]$, $j = 1, \dots, n$ μπορούμε να θεωρήσουμε ότι ακολουθούν την Διωνυμική κατανομή με παραμέτρους n_j και p_j.

Άρα, η εκτιμήτρια της διασποράς θα είναι:

$$\mathrm{Var}\{n_j - d_j\} = n_j p_j (1 - p_j), \tag{2.12}$$

Όπως προκύπτει από την Διωνυμική κατανομή.

Συνεπώς, χρησιμοποιώντας τη σχέση (2.12) έχουμε ότι:

$$\mathrm{Var}\{p_j\} = \mathrm{Var}\left\{n_j - \frac{d_j}{n_j}\right\} = \frac{\mathrm{Var}\{n_j - d_j\}}{{n_j}^2} = p_j(1 - p_j)/n_j. \tag{2.13}$$

Σύμφωνα με την προσέγγιση κατά Taylor για την εύρεση της διασποράς μίας συνάρτησης της τυχαίας μεταβλητής X, έστω $g(\cdot)$ έχουμε ότι:

$$\mathrm{Var}\{g(X)\} \approx \left\{\frac{dg(X)}{dX}\right\}^2 \mathrm{Var(X)}. \tag{2.14}$$

Χρησιμοποιώντας τις σχέσεις (2.13) και (2.14) λαμβάνουμε ότι:

$$\mathrm{Var}\{\log(\widehat{p_j})\} \approx \frac{1 - \widehat{p_j}}{n_j \widehat{p_j}} = \frac{d_j}{n_j(n_j - d_j)}. \tag{2.15}$$

Η σχέση (2.11) χρησιμοποιώντας τη σχέση (2.15) δίνει:

$$\mathrm{Var}\{\log(\hat{S}(t))\} \approx \sum_{j=1}^{\kappa} \frac{d_j}{n_j(n_j - d_j)}. \tag{2.16}$$

και χρησιμοποιώντας τη σχέση (2.14) λαμβάνουμε ότι:

$$\mathrm{Var}\{\log(\hat{S}(t))\} \approx \frac{1}{[\hat{S}(t)]^2} Var\{\hat{S}(t)\}, \tag{2.17}$$

και τελικά,

$$\mathrm{Var}\{\hat{S}(t)\} \approx [\hat{S}(t)]^2 \sum_{j=1}^{\kappa} \frac{d_j}{n_j(n_j - d_j)}. \tag{2.18}$$

Συνεπώς, η διασπορά της εκτιμήτριας Kaplan-Meier για τη συνάρτηση επιβίωσης αποδεικνύεται ότι ικανοποιεί την παρακάτω σχέση:

$$Var\{\hat{S}(t)\} \approx [\hat{S}(t)]^2 \sum_{t_{(i)} < t} \frac{d_j}{n_j(n_j - d_j)}$$

και άρα το τυπικό σφάλμα της εκτιμήτριας Kaplan-Meier για τη συνάρτηση επιβίωσης δίνεται από την σχέση:

$$se\{\hat{S}(t)\} \approx \hat{S}(t)\left\{\sum_{t_{(i)}<t} \frac{d_j}{n_j(n_j - d_j)}\right\}^{1/2}.$$

Για κάθε χρονική στιγμή t μπορεί να κατασκευαστεί ένα διάστημα εμπιστοσύνης της συνάρτησης επιβίωσης είναι:

$\hat{S}(t) \pm z_{1-\frac{a}{2}} se\{\hat{S}(t)\}$, όπου $z_{1-\frac{a}{2}}$ το (1-α/2) ποσοστιαίο σημείο της τυποποιημένης Κανονικής κατανομής.

Παράδειγμα 2.2

Για το συγκεκριμένο παράδειγμα θα χρησιμοποιήσουμε δεδομένα επιβίωσης 228 ασθενών σε προχωρημένο στάδιο καρκίνου του πνεύμονα. Τα δεδομένα προέρχονται από το κέντρο NCCTG (Loprinzi, et al., 1994) και είναι διαθέσιμα στην R μέσω της εντολής *data(lung)*. Το δείγμα περιλαμβάνει 10 μεταβλητές, οι οποίες είναι οι εξής:

- inst: Ο κωδικός του ιδρύματος που νοσηλεύτηκε ο ασθενής (*institution code*).
- time: Ο χρόνος επιβίωσης σε μέρες (*survival time*).
- status: Μία δίτιμη μεταβλητή, η οποία υποδηλώνει πότε μία παρατήρηση είναι λογοκριμένη ή όχι (*censoring status*), για τιμή $1 =$ λογοκριμένες τιμές, $2 =$ χρόνος επιβίωσης.
- age: Ηλικία σε χρόνια.
- sex: Μία δίτιμη μεταβλητή που υποδηλώνει το φύλο του ασθενή και λαμβάνει τιμές, Άντρας $= 1$, Γυναίκα $= 2$.
- ph.ecog: Κλίμακα απόδοσης ECOG, $0 =$ καλό στάδιο, $5 =$ θάνατος.
- ph.karno: Κλίμακα απόδοσης Karnofsky, η οποία συμπληρώνεται από τον φυσικό επιστήμονα, $0 =$ άσχημο στάδιο, $\ldots, 100 =$ καλό στάδιο.
- pat.karno: Κλίμακα απόδοσης Karnofsky, η οποία συμπληρώνεται από τον ασθενή, $0 =$ άσχημο στάδιο,…, $100 =$ καλό στάδιο.
- meal.cal: Συνολικές θερμίδες που καταναλώθηκαν στα γεύματα.
- wt.loss: Το σύνολο απώλειας βάρους τους τελευταίους 6 μήνες.

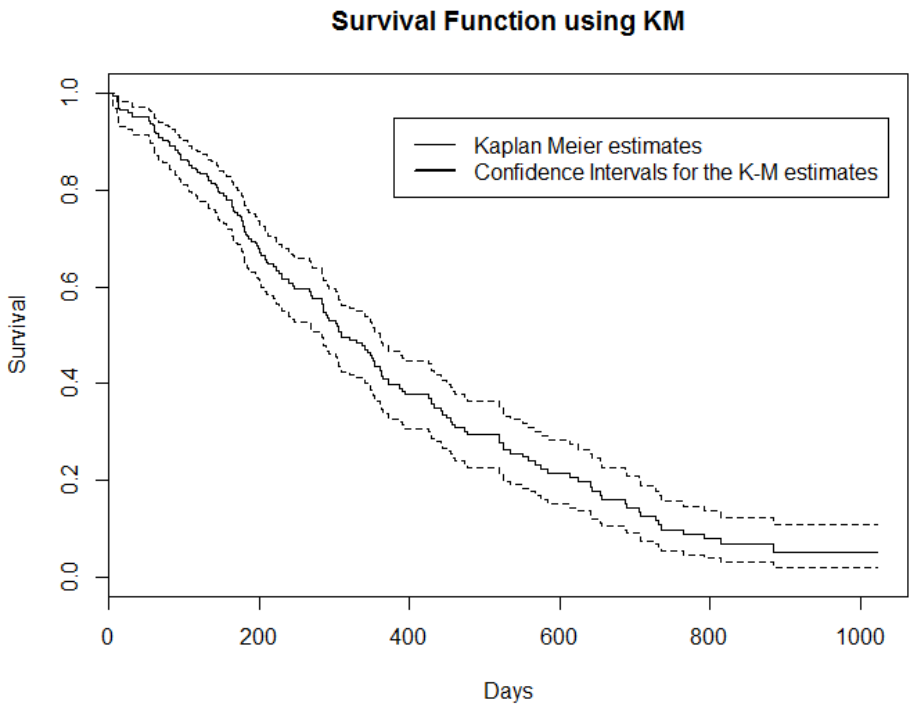
Τα παραπάνω δεδομένα θα χρησιμοποιηθούν για την κατασκευή της εκτιμήτριας Kaplan-Meier με τη βοήθεια του στατιστικού πακέτου R για όλο το δείγμα, αλλά και για άντρες και γυναίκες χωριστά. Επιπλέον, θα κατασκευαστεί και ένα διάστημα εμπιστοσύνης για τις υπό εκτίμηση ποσότητες και τέλος θα απεικονιστούν τα αποτελέσματα σε κατάλληλα γραφήματα.

```
#Kaplan_Meier Survival Analysis
#install the package
#install.packages("survival")
#load the package
library(survival)
data(lung)

#show first 6 rows of the dataset
head(lung)

## Add survival object
lung$SurvObj <- with(lung, Surv(time, status == 2))

## Check data
head(lung)
```

```
> head(lung)
  inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss SurvObj
1    3  306      2  74   1       1        90       100     1175      NA     306
```

```
2     3  455       2  68    1        0        90          90      1225       15     455
3     3 1010       1  56    1        0        90          90        NA       15    1010+
4     5  210       2  57    1        1        90          60      1150       11     210
5     1  883       2  60    1        0       100          90        NA        0     883
6    12 1022       1  74    1        1        50          80       513        0    1022+
```

```
#Kaplan Meier
## Kaplan-Meier estimator. The "log-log" confidence interval is preferred.
km.as.one <- survfit(SurvObj ~ 1, data = lung, conf.type = "log-log")
km.by.sex <- survfit(SurvObj ~ sex, data = lung, conf.type = "log-log")

## Show object
km.as.one

km.by.sex

## See survival estimates at given time
> km.as.one
Call: survfit(formula = SurvObj ~ 1, data = lung, conf.type = "log-log")

      n  events  median 0.95LCL 0.95UCL
    228     165     310     284     361
>
> km.by.sex
Call: survfit(formula = SurvObj ~ sex, data = lung, conf.type = "log-log")

        n events median 0.95LCL 0.95UCL
sex=1 138    112    270     210     306
sex=2  90     53    426     345     524

## Plotting without any specification
plot(km.as.one)
```



Διάγραμμα 2.15: *Γράφημα της εκτίμησης της συνάρτησης Επιβίωσης χρησιμοποιώντας την εκτιμήτρια Kaplan-Meier, καθώς και τα 95% διαστήματα εμπιστοσύνης.*

Στο Διάγραμμα 2.15 απεικονίζονται οι εκτιμήσεις της συνάρτησης επιβίωσης, χρησιμοποιώντας τον Kaplan-Meier εκτιμητή, καθώς και τα αντίστοιχα 95% διαστήματα εμπιστοσύνης για δεδομένα επιβίωσης 228 ασθενών σε προχωρημένο στάδιο καρκίνου του πνεύμονα του Παραδείγματος 2.2. Για την εν λόγω εκτίμηση είναι προφανές ότι λαμβάνονται υπόψη οι χρόνοι επιβίωσης, καθώς και οι λογοκριμένες παρατηρήσεις.

```
library(ggfortify)
library(ggplot2)
library(survival)

autoplot(km.by.sex,main="Survival function using KM for males and females",ylab="Survival
propability")
```

Διάγραμμα 2.16: *Γράφημα της εκτίμησης της συνάρτησης Επιβίωσης χρησιμοποιώντας την εκτιμήτρια Kaplan-Meier για τους άνδρες και τις γυναίκες χωριστά, καθώς και τα αντίστοιχα 95% διαστήματα εμπιστοσύνης.*

Στο Διάγραμμα 2.16 απεικονίζονται οι εκτιμήσεις της συνάρτησης επιβίωσης, χρησιμοποιώντας τον Kaplan-Meier εκτιμητή για άντρες ($sex = 1$) και γυναίκες ($sex = 2$). Παρατηρούμε ότι η επιβίωση των αντρών μειώνεται με μεγαλύτερο ρυθμό σε σχέση με των γυναικών. Η πιθανότητα να επιβιώσουν οι άντρες 500 ημέρες είναι 23%, ενώ οι γυναίκες είναι περίπου 38%.

2.2.3 Η Μέθοδος Nelson-Aalen

Η μέθοδος Nelson-Aalen αποτελεί μία μη παραμετρική μέθοδο για την εκτίμηση της αθροιστικής συνάρτησης κινδύνου $H_t(\cdot)$. Η εκτιμήτρια Nelson-Aalen προτάθηκε από τον Nelson (1972) ως ένας δείκτης αξιοπιστίας και επαναχρησιμοποιήθηκε από τον Aalen (1978) σε σύγχρονες τεχνικές απαρίθμησης (*modern counting process techniques*) (Wienke, 2011).

Κρατώντας τον συμβολισμό, που χρησιμοποιήσαμε παραπάνω, η εκτιμήτρια Nelson-Aalen για την αθροιστική συνάρτηση κινδύνου δίνεται από τη σχέση:

$$\widehat{H_T}(t) = \sum_{t_j \le t} \frac{d_j}{n_j},$$

όπου n_j είναι τα άτομα που βρίσκονται σε κίνδυνο ακριβώς πριν την χρονική στιγμή t_j. Συνεπώς, η εκτιμήτρια Nelson-Aalen αποτελεί μία βηματική μέθοδο, συνεχή από δεξιά, η οποία λαμβάνει εκτιμήσεις σημειακά για τους παρατηρούμενους πραγματικούς χρόνους επιβίωσης.

Η διασπορά αυτής της εκτιμήτριας δίνεται από τη σχέση:

$$Var\{\hat{H}(t)\} = \sum_{t_j \le t} \frac{d_j}{n_j^2}.$$

Συνεπώς, υποθέτοντας για μεγάλα δείγματα κανονική κατανομή μπορούμε να κατασκευάσουμε κατάλληλα διαστήματα εμπιστοσύνης για κάθε χρονική στιγμή t:

$\hat{H}(t) \pm z_{1-\frac{a}{2}} se\{\hat{H}(t)\}$, όπου $se\{\hat{H}(t)\}$ είναι η τετραγωνική ρίζα του $Var\{\hat{H}(t)\}$.

Για μικρά δείγματα μία καλή επιλογή διαστήματος εμπιστοσύνης είναι η χρήση του λογαριθμικού μετασχηματισμού.

$$\hat{H}(t) \exp\left[\pm z_{1-\frac{a}{2}} \frac{se\{\hat{H}(t)\}}{\hat{H}(t)}\right].$$

$z_{1-\frac{a}{2}}$ το (1-α/2) ποσοστιαίο σημείο της τυποποιημένης Κανονικής κατανομής.

Ο Breslow (1972) πρότεινε έναν διαφορετικό μη παραμετρικό εκτιμητή για την συνάρτηση επιβίωσης, ο οποίος δίνεται από τη σχέση:

$$\widehat{S_B}(t) = \exp\left(-\hat{H}(t)\right).$$

Αν για μία δεδομένη χρονική στιγμή ο εκτιμητής Nelson-Aalen είναι

$$\widehat{H_T}(t) = \sum_{t_j \le t} \frac{d_j}{n_j}$$

τότε η εκτιμήτρια Breslow μπορεί να γραφεί:

$$\widehat{S_B}(t) = \prod_{j:t_j \le t} \exp\left(-\hat{H}(t_j)\right).$$

2.3 Μοντέλο αναλόγων κινδύνων του Cox

Στις περισσότερες περιπτώσεις ο πληθυσμός που μελετάμε είναι ανομοιογενής και διαφοροποιείται ανάλογα με διάφορα χαρακτηριστικά, όπως ηλικία, φύλο, οικογενειακή κατάσταση, ιατρικό ιστορικό, σπουδές, δημογραφικά χαρακτηριστικά κ.ο.κ. Ο ερευνητής λοιπόν στις περισσότερες περιπτώσεις ενδιαφέρεται να εξετάσει κατά πόσο και πώς επηρεάζουν οι μεταβλητές αυτές τον χρόνο επιβίωσης. Στην ανάλυση επιβίωσης προσπαθούμε να κατασκευάσουμε μοντέλα για να εκτιμήσουμε την επίδραση μιας θεραπείας, μιας ασθένειας στη διάρκεια ζωής ενός ατόμου. Συνεπώς, είναι εύλογο ότι θα επιλέξουμε να μοντελοποιήσουμε την συνάρτηση κινδύνου, η οποία ουσιαστικά μας δίνει ένα μέτρο για τον κίνδυνο

θανάτου ενός ατόμου. Φυσικά, από την εκτίμηση της συνάρτησης κινδύνου μπορούν να προκύψουν και οι εκτιμήσεις της συνάρτησης επιβίωσης, καθώς και η συνάρτηση πυκνότητας πιθανότητας του χρόνου επιβίωσης.

Για το σκοπό αυτό έχουν αναπτυχθεί διάφορα μοντέλα παλινδρόμησης κατάλληλα για δεδομένα επιβίωσης. Ένα από τα πιο διαδεδομένα μοντέλα στην ανάλυση επιβίωσης είναι το μοντέλο αναλόγων κινδύνων του Cox (1972 - *Cox proportional hazards regression model*).

Το μοντέλο του Cox λοιπόν αποτελεί μία στατιστική τεχνική, ώστε να προβλέψουμε την επίδραση της θεραπείας στην επιβίωση των ασθενών, δεδομένου των τιμών των επεξηγηματικών μεταβλητών που επηρεάζουν τον χρόνο επιβίωσης του ασθενούς και κατ' επέκταση τον κίνδυνο θανάτου του.

Το μοντέλο του Cox λαμβάνει την παρακάτω δομή:

Έστω, T η τυχαία μεταβλητή, που εκφράζει τον χρόνο επιβίωσης ασθενών και $\boldsymbol{X} = (X_1, X_2, \dots, X_p)$ ένα τυχαίο διάνυσμα p επεξηγηματικών μεταβλητών για τους εν λόγω ασθενείς.

Για να μπορέσουμε να εκτιμήσουμε την σχέση της μεταβλητής απόκρισης (T) με τις επεξηγηματικές μεταβλητές χρησιμοποιούμε ένα σύνολο παρατηρήσεων $(t_i, \delta_i, \boldsymbol{x}_i), i = 1, 2, \dots, n$, όπου t_i ο χρόνος επιβίωσης που παρατηρήθηκε, δ_i ένας δείκτης που προσδιορίζει αν η παρατήρηση είναι λογοκριμένη ή μη και $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ το διάνυσμα των παρατηρούμενων τιμών των p επεξηγηματικών μεταβλητών που αντιστοιχούν στον $i-$ασθενή.

$$\delta_i = \begin{cases} 1, & \alpha\nu\ \eta\ t_j\ \epsilon\acute{\iota}\nu\alpha\iota\ \pi\rho\alpha\gamma\mu\alpha\tau\iota\kappa\acute{\eta}\ \tau\iota\mu\acute{\eta} \\ 0, & \alpha\nu\ \eta\ t_j\ \epsilon\acute{\iota}\nu\alpha\iota\ \lambda o\gamma o\kappa\rho\iota\mu\acute{\epsilon}\nu\eta \end{cases}$$

Τότε το μοντέλο του Cox δίνεται στη γενική του μορφή από τη σχέση

$$h(t|\boldsymbol{X}) = h_0(t)g(\boldsymbol{X}),$$

όπου $h(t|\boldsymbol{X})$ είναι ο κίνδυνος θανάτου ενός ασθενούς την χρονική στιγμή t δεδομένου όλων των παραγόντων που επηρεάζουν τον χρόνο επιβίωσής του. Επιπλέον, η ποσότητα $h_0(t)$ είναι το επίπεδο αναφοράς της συνάρτησης κινδύνου για έναν ασθενή (*arbitrary baseline hazard rate*) και αντιστοιχεί στην πιθανότητα θανάτου ενός ασθενή τη χρονική στιγμή t όταν όλες οι επεξηγηματικές μεταβλητές λαμβάνουν την τιμή 0. Αποτελεί τη βασική συνάρτηση κινδύνου του μοντέλου παλινδρόμησης και χρησιμοποιείται ως μία βάση σύγκρισης για τον κίνδυνο θανάτου ενός ασθενή. Κατά μία έννοια αποτελεί μία γενίκευση του σταθερού όρου του μοντέλου (*intercept-constant term*) όπως αυτός ορίζεται στα παραμετρικά μοντέλα παλινδρόμησης. Η $g(\cdot)$ είναι μία γνωστή συνάρτηση, η οποία μπορεί να επιλεχθεί κατάλληλα για την κατασκευή του μοντέλου παλινδρόμησης, ώστε η συνάρτηση κινδύνου να λαμβάνει θετικές τιμές, αφού εκφράζει ρυθμό. Η πιο συνήθης επιλογή είναι η εκθετική συνάρτηση και έτσι η $g(\cdot)$ λαμβάνει τη μορφή $g(\boldsymbol{X}) = \exp(\beta_1 X_1 + \cdots + \beta_p X_p)$, όπου $\beta_1, \dots, \beta_p$ είναι οι συντελεστές του μοντέλου που αντιστοιχούν σε κάθε μία μεταβλητή από τις μεταβλητές $X_1, X_2, \dots, X_p$.

Η συνάρτηση κινδύνου λοιπόν εκφράζεται ως γινόμενο δύο όρων, της βάσης $h_0(t)$ η οποία εξαρτάται από τον χρόνο t και της συνάρτησης $g(\boldsymbol{X})$ η οποία δεν εξαρτάται από το χρόνο επιβίωσης, αλλά από το διάνυσμα των άγνωστων παραμέτρων του μοντέλου. Από τον ορισμό του μοντέλου αναλόγων κινδύνων του Cox παρατηρούμε ότι το επίπεδο αναφοράς της συνάρτησης κινδύνου $h_0(t)$ δεν εξαρτάται από τις τιμές των επεξηγηματικών μεταβλητών (Klein & Moeschberger, 2003).

Για αυτό το λόγο το παραπάνω μοντέλο είναι ημι-παραμετρικό, καθώς θεωρούμε μια θεωρητική κατανομή μόνο για την συνάρτηση g, ενώ ο βασικός ρυθμός κινδύνου, η οποία είναι μια άγνωστη συνάρτηση, εκτιμάται μη παραμετρικά.

- **Εκτίμηση Παραμέτρων του μοντέλου**

Ας συμβολίσουμε με $f(\cdot)$ τη συνάρτηση πυκνότητας πιθανότητας του χρόνου επιβίωσης και $S(\cdot)$ η αντίστοιχη συνάρτηση επιβίωσης. Τότε, όπως έχει αναφερθεί και σε προηγούμενη ενότητα, η συνάρτηση πιθανοφάνειας έχει την παρακάτω μορφή (ως προς μία σταθερά αναλογίας):

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \{[f(t_i, \boldsymbol{x}_i, \delta_i)]^{\delta_i} [S(t_i, \boldsymbol{x}_i, \delta_i)]^{1-\delta_i}\}.$$

Λογαριθμίζοντας έχουμε:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{\delta_i \ln[f(t_i, \boldsymbol{x}_i, \delta_i)] + (1-\delta_i) \ln[S(t_i, \boldsymbol{x}_i, \delta_i)]\}.$$

Επιπλέον, για το μοντέλο του Cox, η συνάρτηση Επιβίωσης μπορεί να γραφεί ως εξής:

$$S(t_i, x_i, \boldsymbol{\beta}) = [S_0(t_i)]^{\exp(\boldsymbol{\beta}' \boldsymbol{x}_i)}, \tag{2.10}$$

όπου $S_0(t_i) = \exp(-H_0(t_i))$ είναι το επίπεδο αναφοράς της συνάρτησης επιβίωσης.

Χρησιμοποιώντας την σχέση: $f(t, x, \beta) = h(t, x, \beta) \cdot S(t, x, \beta)$, τον ορισμό του μοντέλου του Cox και την έκφραση (2.10) λαμβάνουμε ότι:

$$\mathrm{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} [h_0(t_i) \exp(\boldsymbol{\beta}' \boldsymbol{x}_i)]^{\delta_i} \exp(-H_0(t_i) \exp(\boldsymbol{\beta}' \boldsymbol{x}_i)) \tag{2.11}$$

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{\delta_i \ln[h_0(t_i)] + \delta_i \boldsymbol{\beta}' \boldsymbol{x}_i + e^{\boldsymbol{\beta}' \boldsymbol{x}_i} \ln[S_0(t_i)]\}. \tag{2.12}$$

Η παραπάνω συνάρτηση εξαρτάται εκτός από την άγνωστη παράμετρο $\boldsymbol{\beta}$ και από την άγνωστη συνάρτηση $h_0(t_i)$, αλλά και από τη $S_0(t_i)$. Όπως έχει αναφερθεί στην βιβλιογραφία λοιπόν είναι δύσκολο να εφαρμόσουμε τη μέθοδο Μέγιστης Πιθανοφάνειας δεδομένου ότι έχουμε πολλές άγνωστες «παραμέτρους» να χειριστούμε και αυτή τελικά που μας ενδιαφέρει να εκτιμήσουμε είναι το διάνυσμα συντελεστών $\boldsymbol{\beta}$.

Για να μπορέσει να αντιμετωπίσει αυτήν την δυσκολία, ο Cox πρότεινε μία άλλη μορφή συνάρτησης πιθανοφάνειας για να περιγράψει τα δεδομένα, την οποία αποκάλεσε μερική συνάρτηση πιθανοφάνειας (*Partial likelihood function*). Η συνάρτηση αυτή εξαρτάται μόνο από την παράμετρο που μας ενδιαφέρει, δηλαδή την $\boldsymbol{\beta}$. Απέδειξε ότι οι εκτιμήτριες των παραμέτρων που προκύπτουν από την πλήρη συνάρτηση πιθανοφάνειας παρουσιάζουν ακριβώς τις ίδιες ιδιότητες με τις εκτιμήτριες που προκύπτουν από την μερική συνάρτηση πιθανοφάνειας (Hosmer, et al., 2008).

Η μερική συνάρτηση Πιθανοφάνειας δίνεται από τη σχέση:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ \frac{\exp(\boldsymbol{\beta}' \boldsymbol{x}_i)}{\sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}' \boldsymbol{x}_i)} \right\}^{\delta_i},$$

όπου $R(t_i)$ είναι το σύνολο των μονάδων του πληθυσμού που βρίσκονται σε κίνδυνο λίγο πριν τη χρονική στιγμή t_i (Collett, 2003).

Ο λογάριθμος της μερικής συνάρτησης πιθανοφάνειας εκφράζεται ως εξής:

$$l(\boldsymbol{\beta}) = \ln\big(L(\boldsymbol{\beta})\big) = \sum_{i=1}^{n} \delta_i \left\{ \boldsymbol{\beta}' \boldsymbol{x}_i - ln\left(\sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}' \boldsymbol{x}_i) \right) \right\}.$$

Οι εκτιμήσεις των συντελεστών προκύπτουν μεγιστοποιώντας την παραπάνω συνάρτηση χρησιμοποιώντας επαναληπτικές – αριθμητικές μεθόδους, όπως την επαναληπτική μέθοδο Newton – Raphson.

- **Wald test**

Στην περίπτωση που θέλουμε να ελέγξουμε τη μηδενική υπόθεση ότι ένας συντελεστής του μοντέλου είναι ίσος με $H_0: \beta_i = 0$, με εναλλακτική ότι $H_1: \beta_i \neq 0$ για κάθε συντελεστή χωριστά, χρησιμοποιούμε το Wald test. Το στατιστικό ελέγχου δίνεται υπό τη μηδενική υπόθεση από την παρακάτω σχέση:

$$T = \left\{ \frac{\left(\hat{\beta}_j - \beta_0\right)}{se(\hat{\beta}_j)} \right\}^2,$$

όπου $se(\hat{\beta}_j)$ είναι το τυπικό σφάλμα της εκτίμησης του μονοδιάστατου συντελεστή β_j. Το εν λόγω στατιστικό ακολουθεί την X^2 κατανομή με 1 βαθμό ελευθερίας. Το Wald test γενικεύεται στην περίπτωση που η παράμετρος $\boldsymbol{\beta}$ που μας ενδιαφέρει είναι διάνυσμα διαστάσεως p (Collett, 2003).

- **Λόγος αναλογικών κινδύνων (proportional hazard ratios)**

Για δύο διανύσματα τιμών των επεξηγηματικών μεταβλητών $\boldsymbol{X}$ και $\boldsymbol{X}^*$ λαμβάνουμε ότι:

$$\frac{h(t|\boldsymbol{X})}{h(t|\boldsymbol{X}^*)} = \frac{h_0(t)\exp(\sum_{i=1}^{p} \beta_i X_i)}{h_0(t)\exp(\sum_{i=1}^{p} \beta_i X_i^*)} = \exp\left\{ \sum_{i=1}^{p} \beta_i \left(X_i - X_i^*\right) \right\} = \gamma.$$

Η παραπάνω σχέση αποτελεί μία σταθερά αναλογίας, έστω γ και παρατηρείται ότι είναι μία σχέση ανεξάρτητη από τον χρόνο t. Συνεπώς, ο λόγος των ρυθμών κινδύνου δύο διαφορετικών ατόμων είναι σταθερός για γνωστές τιμές των επεξηγηματικών μεταβλητών και εκφράζει το σχετικό κίνδυνο (*relative risk*) θανάτου ενός ατόμου με παράγοντες κινδύνου $\boldsymbol{X}$ συγκριτικά με ένα άτομο με παράγοντες κινδύνου $\boldsymbol{X}^*$. Υπάρχει λοιπόν μία σχέση αναλογίας μεταξύ των δύο συναρτήσεων κινδύνου για τα δύο διανύσματα τιμών των επεξηγηματικών μεταβλητών και γι' αυτό τον λόγο το μοντέλο παλινδρόμησης του Cox καλείται και μοντέλο αναλόγων κινδύνων. (Klein & Moeschberger, 2003)

- **Λόγος στιγμιαίων κινδύνων (hazard ratios)**

Το μοντέλο εξετάζει την μορφή της συνάρτησης κινδύνου, αλλά και τις επιδράσεις των επεξηγηματικών μεταβλητών στην συνάρτηση κινδύνου και κατ' επέκταση στον χρόνο επιβίωσης. Με τη χρήση νεπέριων λογαρίθμων η συνάρτηση κινδύνου $\ln\{h(t|\boldsymbol{X})\}$ συνδέεται γραμμικά με τις επεξηγηματικές μεταβλητές $X_1, X_2, \dots, X_p$. Συνεπώς, οι συντελεστές του μοντέλου μπορούν να ερμηνευθούν ως εξής στην περίπτωση ποσοτικών επεξηγηματικών μεταβλητών:

$$\ln\{h(t|\boldsymbol{X})\} = \ln\big(h_0(t)g(\boldsymbol{X})\big) = \ln\big(h_0(t)\exp\big(\beta_1 X_1 + \cdots + \beta_p X_p\big)\big) = ln\big(h_0(t)\big) + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$\ln\left\{\frac{h(t|\boldsymbol{X})}{h_0(t)}\right\} = \beta_1 X_1 + \cdots + \beta_p X_p$$

Υποθέτουμε ότι για το μοντέλο του Cox, οι επεξηγηματικές μεταβλητές $X_1, X_2, \dots, X_p$ επιδρούν προσθετικά στην ποσότητα $\ln\{h(t|\boldsymbol{X})\}$, δηλαδή δεν υπάρχουν αλληλεπιδράσεις (*interactions*) μεταξύ των μεταβλητών και ότι η ποσότητα $\ln\{h(t|\boldsymbol{X})\}$ συνδέεται γραμμικά με τους συντελεστές του μοντέλου $\beta_1, \dots, \beta_p$. Επιπλέον, το μοντέλο του Cox δεν υποθέτει μία ιδιαίτερη κατανομή για τους χρόνους επιβίωσης, αλλά υποθέτει ότι οι επιδράσεις διαφορετικών μεταβλητών στην επιβίωση είναι σταθερές ως προς τον χρόνο.

- **Ερμηνεία των συντελεστών του μοντέλου**

Ας υποθέτουμε ότι το μοντέλο αναλόγων κινδύνων περιλαμβάνει μία συνεχή επεξηγηματική μεταβλητή, έστω X. Τότε, το μοντέλο δίνεται από τη σχέση:

$$h(t|X) = h_0(t)e^{\beta X}.$$

Ο λογάριθμος του λόγου των συναρτήσεων κινδύνου γράφεται στη μορφή ενός γραμμικού μοντέλου.

$$\ln\left\{\frac{h(t|X)}{h_0(t)}\right\} = \beta X.$$

Ας υποθέσουμε ότι θέλουμε να συγκρίνουμε τις συναρτήσεις κινδύνου για δύο μονάδες του πληθυσμού, όπου η μία λαμβάνει την τιμή x για την επεξηγηματική μεταβλητή X και η άλλη την τιμή $x + 1$. Τότε,

$$\frac{h(t|x+1)}{h(t|x)} = \frac{h_0(t)e^{\widehat{\beta}(x+1)}}{h_0(t)e^{\widehat{\beta}x}} = \frac{e^{\widehat{\beta}(x+1)}}{e^{\widehat{\beta}x}} = e^{\widehat{\beta}}.$$

Συνεπώς, η εκτίμηση της παραμέτρου β είναι η μεταβολή του λογαρίθμου του λόγου της συνάρτησης κινδύνου, όταν η επεξηγηματική μεταβλητή αυξηθεί κατά μία μονάδα.

Αν υποθέσουμε ότι το μοντέλο μας περιλαμβάνει μία επεξηγηματική μεταβλητή, η οποία είναι κατηγορική και ας θεωρήσουμε την πιο απλή περίπτωση, όπου έχουμε δύο κατηγορίες $X = 0$ ή $X = 1$.

Τότε η βασική συνάρτηση κινδύνου αναφέρεται στους ασθενείς, που λαμβάνουν την τιμή 1 για την επεξηγηματική μεταβλητή X. Επιπροσθέτως, το $\exp(\beta)$ είναι η μεταβολή στην τιμή της συνάρτησης κινδύνου ενός ασθενή με τιμή 1 για την επεξηγηματική μεταβλητή X σε σχέση με έναν ασθενή με τιμή 0 για την επεξηγηματική μεταβλητή X (Collett, 2003).

- **Deviance**

Σε αυτό το σημείο θα θέλαμε να δούμε αν το μοντέλο που επιλέξαμε προσαρμόζεται καλά στα δεδομένα μας. Με άλλα λόγια είναι σημαντικό να κάνουμε επιλογή των κατάλληλων μεταβλητών, ώστε να κατασκευάσουμε ένα μοντέλο το οποίο θα περιγράφει στον μεγαλύτερο δυνατό βαθμό τα δεδομένα μας.

Ένας τρόπος για να ελέγξουμε την προσαρμογή του μοντέλου είναι χρησιμοποιώντας την στατιστική συνάρτηση Deviance, η οποία στηρίζεται στον έλεγχο του λόγου των πιθανοφανειών δύο επιλεγμένων μοντέλων. Ουσιαστικά, ελέγχεται η μηδενική υπόθεση $H_0: \boldsymbol{\beta}_h = \boldsymbol{0}$ με εναλλακτική την $H_0: \boldsymbol{\beta}_h \neq \boldsymbol{0}$. Με άλλα λόγια γίνεται μία σύγκριση μεταξύ του μοντέλου που προσαρμόζουμε και του ιδανικού μοντέλου (*saturated model or full model or maximal model*).

Συγκρίνεται η μεγιστοποιημένη πιθανοφάνεια υπό το μοντέλο της επιλογής μας με την μεγιστοποιημένη πιθανοφάνεια για το καλύτερο μοντέλο. Το saturated μοντέλο είναι ουσιαστικά το καλύτερο μοντέλο που θα μπορούσαμε να πάρουμε από τα δεδομένα μας. Είναι ένα μοντέλο με αριθμό επεξηγηματικών μεταβλητών ίσο με το μέγεθος του δείγματος. Άρα, το saturated μοντέλο αποτελεί το μοντέλο με την καλύτερη προσαρμογή στα δεδομένα και συγκρίνοντας το με το μοντέλο της επιλογής μας λαμβάνουμε μία εικόνα του πόσο καλά περιγράφει το μοντέλο μας τα δεδομένα μας. Για το μοντέλο του Cox χρησιμοποιώντας τον λόγο των μερικών συναρτήσεων Πιθανοφάνειας λαμβάνουμε ότι:

$$Deviance = -2\ln\left\{\frac{L_{fitted}}{L_{saturated}}\right\}.$$

Αποδεικνύεται ότι η εν λόγω στατιστική συνάρτηση ακολουθεί αυμπτωτικά την κατανομή X^2με βαθμούς ελευθερίας ίσους με τη διαφορά των παραμέτρων των δύο μοντέλων, $d_{saturated} - d_{fitted}$, όπου $d_{saturated}$ ο αριθμός των παραμέτρων του full μοντέλου και d_{fitted} ο αριθμός παραμέτρων του υπό μελέτη μοντέλου. Χρησιμοποιώντας την ελεγχοσυνάρτηση Deviance μπορεί να γίνει και σύγκριση δύο εμφωλευμένων μοντέλων. Σε αυτήν την περίπτωση η ελεγχοσυνάρτηση $Deviance_1 - Deviance_2$ ακολουθεί την X^2 κατανομή με βαθμούς ελευθερίας ίσους με τον αριθμό της διαφοράς των παραμέτρων των δύο μοντέλων.

2.4 Κριτήρια επιλογής μοντέλου

Για την επιλογή του κατάλληλου μοντέλου, αλλά και για την σύγκριση διαφορετικών μοντέλων ως προς τη σπουδαιότητά τους χρησιμοποιούνται τα μέτρα καταλληλότητας. Πρόκειται για κάποιες αριθμητικές ποσότητες, που χρησιμοποιούνται για την αξιολόγηση ενός μοντέλου, αλλά κυρίως για την επιλογή του βέλτιστου μοντέλου μεταξύ άλλων. Θα παρουσιάσουμε κάποια από αυτά και συγκεκριμένα τα κριτήρια AIC και BIC.

2.4.1 Δείκτες Καλής Προσαρμογής AIC και BIC

- **Akaike's information criterion (*AIC*)**

Το AIC αποτελεί ένα κριτήριο επιλογής του βέλτιστου μοντέλου με το όσο το δυνατόν μικρότερο αριθμό παραμέτρων. Ορίζεται από τη σχέση

$$AIC = 2d - 2\log L,$$

όπου L η μεγιστοποιημένη τιμή της συνάρτησης πιθανοφάνειας για το εκτιμημένο μοντέλο και d ο αριθμός παραμέτρων του μοντέλου.
Συγκρίνοντας όλα τα υποψήφια μοντέλα με βάση το παραπάνω κριτήριο φαίνεται να είναι βέλτιστο εκείνο με το μικρότερο AIC. Η εισαγωγή περισσότερων παραμέτρων στο μοντέλο αυξάνει την προσαρμογή του, ανεξάρτητα αν είναι στατιστικά σημαντικές ή όχι, καθώς αυξάνει ο όρος $\log L$ με την επιπλέον πρόσθεση μεταβλητών. Ωστόσο, αυξάνεται και ο πρώτος όρος του AIC, το d, δηλαδή ο αριθμός των μεταβλητών και μειώνεται ο δεύτερος όρος του AIC. Τελικά, η εισαγωγή επιπλέον παραμέτρων στο μοντέλο μειώνει την τιμή του AIC στην περίπτωση που η προσαρμογή του μοντέλου βελτιώνεται. Η ποσότητα 2d καλείται ποινή (*penalty*).

- **Bayesian information criterion (*BIC*)**

Το κριτήριο BIC προτάθηκε από τον Schwarz (1978) και ορίζεται από τη σχέση

$$BIC = d\log n - 2\log L,$$

όπου L η μεγιστοποιημένη τιμή της συνάρτησης πιθανοφάνειας για το εκτιμημένο μοντέλο και d αριθμός παραμέτρων του μοντέλου και n ο αριθμός των παρατηρήσεων.
Αποτελεί ένα ακόμη κριτήριο επιλογής βέλτιστου μοντέλου ανάμεσα σε μοντέλα με διαφορετικό αριθμό παραμέτρων, όχι απαραίτητα εμφωλευμένων. Η λογική και η χρήση του είναι όμοια με το κριτήριο AIC. Ωστόσο, η διαφορά τους έγκειται στο γεγονός ότι στην περίπτωση του BIC η εισαγωγή επιπρόσθετων παραμέτρων αποθαρρύνεται σε μεγαλύτερο βαθμό από το AIC.

Σύμφωνα με την κλίμακα του Raftery μπορούμε να κρίνουμε κατά πόσο ένα μοντέλο έχει καλύτερη προσαρμογή στα δεδομένα σε σχέση με ένα άλλο λαμβάνοντας την απόλυτη τιμή της διαφοράς των τιμών, που λαμβάνουν τα κριτήρια BIC για δύο μοντέλα (Hardin & Hilbe, 2007). Η κλίμακα αυτή αναγράφεται στον παρακάτω πίνακα.

Διαφορά των τιμών BIC	**Ένδειξη**
$0-2$	Ασθενής (Weak)
$2-8$	Θετική (Positive)
$6-10$	Ισχυρή (Strong)
>10	Πολύ Ισχυρή (Very Strong)

Πίνακας 2.3: *Κλίμακα τιμών του κριτηρίου BIC.*

2.5 Μοντέλα επιταχυνόμενου χρόνου αποτυχίας

Όπως είπαμε και στην προηγούμενη ενότητα στο μοντέλο του Cox γίνεται η παραδοχή ότι η συνάρτηση αναφοράς της συνάρτησης κινδύνου $h_0(\cdot)$ είναι ανεξάρτητη των τιμών των επεξηγηματικών μεταβλητών. Σε αρκετές εφαρμογές της θεωρίας Αξιοπιστίας, π.χ. στο χώρο της βιομηχανίας η εν λόγω παραδοχή μπορεί να μην είναι ρεαλιστική. Για το λόγο αυτό, συνήθως προσαρμόζονται μοντέλα επιταχυνόμενου χρόνου επιτυχίας (*accelerated failure time models*), βλ. (Kalbfleisch & Prentice, 1980), (Wienke, 2011), τα οποία αποτελούν μία κλάση από λογαριθμογραμμικά μοντέλα του χρόνου επιβίωσης και στα οποία η συνάρτηση κινδύνου έχει την παρακάτω μορφή:

Έστω, το τυχαίο δείγμα μεγέθους n $\boldsymbol{T} = T_1, T_2, \dots, T_n$, όπου T_j ο χρόνος επιβίωσης για τον $j-$ασθενή και τυχαίο διάνυσμα επεξηγηματικών μεταβλητών $\boldsymbol{X} = (X_1, X_2, \dots, X_p)$ για κάθε ασθενή.

Ας θεωρήσουμε την τυχαία μεταβλητή $\boldsymbol{Y}$, η οποία είναι ένας μετασχηματισμός της $\boldsymbol{T}$, $\boldsymbol{Y} = \log(\boldsymbol{T})$. Τότε, θεωρούμε ότι η $\boldsymbol{Y}$ συνδέεται γραμμικά με τις επεξηγηματικές μεταβλητές υπό τη σχέση $\boldsymbol{Y} = \boldsymbol{\beta'X} + \boldsymbol{W}$, όπου $\mathbf{W}$ είναι η μεταβλητή που εκφράζει το σφάλμα του μοντέλου και ακολουθεί μια συνάρτηση πυκνότητας πιθανότητας, έστω f. Εφαρμόζοντας εκθετικό μετασχηματισμό η παραπάνω σχέση μπορεί να γραφεί στη μορφή $\boldsymbol{T} = \exp(\boldsymbol{\beta'X}) * \exp(\boldsymbol{W})$, όπου $\exp(\boldsymbol{W})$ έχει βασική συνάρτηση κινδύνου $h_0(w)$ ανεξάρτητη από τους συντελεστές του μοντέλου $\boldsymbol{\beta}$.

Συνεπώς, η συνάρτηση κινδύνου του χρόνου επιβίωσης μπορεί να γραφεί στην μορφή:

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

ΚΕΦΑΛΑΙΟ 3

3.1 Μοντέλα Ευπάθειας

Έχουν αναπτυχθεί πολλά μοντέλα παλινδρόμησης για δεδομένα διάρκειας ζωής, με σκοπό την εκτίμηση της συνάρτησης κινδύνου, και κατά συνέπεια της συνάρτησης επιβίωσης. Σε αυτά τα μοντέλα χρησιμοποιείται ως μεταβλητή απόκρισης ο χρόνος επιβίωσης και ως επεξηγηματικές μεταβλητές διάφοροι παράγοντες που θεωρούμε ότι επηρεάζουν το χρόνο επιβίωσης. Ωστόσο, σίγουρα υπάρχουν και άλλοι σημαντικοί παράγοντες που επιδρούν στον χρόνο επιβίωσης και είναι αδύνατον να συμπεριληφθούν στην ανάλυση, ίσως γιατί ο ερευνητής δεν έχει πληροφορία για αυτούς σε επίπεδο ατόμου. Ενδεχομένως να μην γνωρίζει ακόμα και την ύπαρξη αυτών των παραγόντων ή να χρειάζεται οικονομικό και προσωπικό κόστος για να συλλέξει αυτά τα δεδομένα και έτσι να μην μπορεί να έχει την πληροφορία τελικά. Συνεπώς, λαμβάνουμε δύο πηγές μεταβλητότητας στα δεδομένα μας, την υπολογίσιμη μεταβλητότητα από παρατηρούμενους παράγοντες και την μεταβλητότητα, η οποία δεν μπορεί να υπολογιστεί αφού υπάρχουν αστάθμητοι παράγοντες. Οι άγνωστοι αυτοί παράγοντες κινδύνου μπορούν να συμπεριληφθούν στο μοντέλο μέσω μίας λανθάνουσας μεταβλητής (*latent variable*), η οποία αποτελεί τον παράγοντα τυχαίων επιδράσεων και καλείται ευπάθεια (*frailty*) στην ανάλυση επιβίωσης και τα εν λόγω μοντέλα καλούνται ευπαθή μοντέλα (*frailty models*). Οι Vaupel et al. (1979) ήταν αυτοί που εισήγαγαν τον όρο ευπάθεια προκειμένου να χαρακτηρίσει αυτά τα μοντέλα τυχαίων επιδράσεων. Ο όρος αυτός χρησιμοποιήθηκε για να προσδιορίσει ότι διαφορετικές μονάδες του πληθυσμού μπορεί να βρίσκονται σε κίνδυνο όταν εκτίθενται σε διάφορους παράγοντες ακόμα και αν φαινομενικά εμφανίζουν ίδια χαρακτηριστικά, όπως ηλικία, φύλο, βάρος κ.λ.π. (Hanagal, 2011). Για παράδειγμα, σε μία έρευνα που εξετάζει δύο θεραπείες για την αντιμετώπιση του καρκίνου συλλέγουμε δεδομένα για κάποια χαρακτηριστικά των ασθενών, όπως φύλο, ηλικία κ.λ.π. και προσαρμόζουμε το μοντέλο αναλόγων κινδύνων του Cox. Συνεπώς, για τους ασθενείς ίδιας ηλικίας και φύλου, που λαμβάνουν την ίδια θεραπεία, η κατανομή του χρόνου επιβίωσης θα είναι ίδια. Ωστόσο, αυτή η παραδοχή ιδιαίτερα για μελέτες σε ζωντανούς οργανισμούς δεν είναι σωστή, καθώς σίγουρα θα υπάρχουν πολλοί παράγοντες που επηρεάζουν τον χρόνο επιβίωσης αυτών και τις περισσότερες φορές δεν μπορούν να συμπεριληφθούν στο μοντέλο.

Η γενίκευση του μοντέλου αναλόγων κινδύνων του Cox είναι το πιο ευρέως διαδεδομένο μοντέλο παλινδρόμησης ευπάθειας, εισάγοντας τον παράγοντα των τυχαίων επιδράσεων (*random effects*) ως μία τυχαία μεταβλητή για τον χειρισμό της αλληλεπίδρασης μεταξύ των μεταβλητών, που δεν έχει ληφθεί υπόψη στο μοντέλο, αλλά και της μη παρατηρούμενης ανομοιογένειας που χαρακτηρίζει τον πληθυσμό που εξετάζουμε. Στα μονοδιάστατα μοντέλα επιβίωσης (*univariate survival models*), ένα μοντέλο ευπάθειας μπορεί να χρησιμοποιηθεί για την εκτίμηση της ανομοιογένειας μεταξύ των διαφορετικών μονάδων του πληθυσμού. Στα πολυδιάστατα (*multivariate survival models*), χρησιμοποιούνται τα από κοινού μοντέλα ευπάθειας (*shared frailty models*). Το από κοινού μοντέλο ευπάθειας είναι ένα μοντέλο τυχαίων επιδράσεων, το οποίο υποθέτει μεταβλητότητα μεταξύ διαφορετικών ομάδων (*frailty*) και μεταβλητότητα μεταξύ των μονάδων του πληθυσμού, η οποία εξηγείται από την συνάρτηση κινδύνου (*hazard function*). Παρακάτω θα δούμε το εν λόγω μοντέλο πιο αναλυτικά.

Για την μοντελοποίηση του χρόνου επιβίωσης μπορούν να χρησιμοποιηθούν και παραμετρικά, αλλά και ημιπαραμετρικά μοντέλα παλινδρόμησης. Στην περίπτωση των μοντέλων ευπάθειας για την παραμετρική προσέγγιση υποθέτουμε μία γνωστή κατανομή για την βασική συνάρτηση κινδύνου (*baseline hazard function*) και εκτιμούμε τις παραμέτρους αυτής χρησιμοποιώντας τα δεδομένα, ενώ στην περίπτωση των ημιπαραμετρικών μοντέλων δεν γίνεται καμία υπόθεση για την κατανομή της βασικής συνάρτησης κινδύνου και η εκτίμηση αυτής γίνεται με άλλες τεχνικές, όπως ο αλγόριθμος ΕΜ.

3.2 Παραμετροποίηση για το μοντέλο ευπάθειας

Έστω T μία τυχαία μεταβλητή που δηλώνει χρόνο επιβίωσης και ακολουθεί μία συνεχή κατανομή. Μία μη αρνητική τυχαία μεταβλητή Z καλείται ευπάθεια (*frailty*) (Vaupel et al., 1979) αν η δεσμευμένη συνάρτηση κινδύνου δεδομένου του Z λαμβάνει την εξής μορφή:

$$h(t|\, Z) = Zh_0(t), \tag{3.1}$$

όπου $h_0(t)$ είναι η βασική συνάρτηση κινδύνου.

Τότε, η δεσμευμένη συνάρτηση επιβίωσης δεδομένου του Z δίνεται από τη σχέση:

$$S(t|\, Z) = e^{-ZH_0(t)} \tag{3.2}$$

όπου $H_0(t) = \int_0^t h_0(s)ds$ είναι η αθροιστική συνάρτηση βαθμού κινδύνου για το επίπεδο αναφοράς.

Η συνάρτηση επιβίωσης $S(t)$ προκύπτει ολοκληρώνοντας τη δεσμευμένη συνάρτηση επιβίωσης $S(t|\, Z)$ ως προς Z, υπολογίζοντας δηλαδή την αναμενόμενη τιμή

$$S(t) = \int_0^\infty \exp\big(-zH_0(t)\big)\, f_Z(z)dz = E_Z[S(t|\, Z)] = E_Z[e^{-ZH_0(t)}], \tag{3.3}$$

όπου $f_Z(z)$ η συνάρτηση πυκνότητας πιθανότητας της ευπάθειας Z.

Η σχέση (3.3) μπορεί να γραφεί χρησιμοποιώντας τον μετασχηματισμό Laplace ως εξής:

$$S(t) = \boldsymbol{L}[H_0(t)]. \tag{3.4}$$

Χρησιμοποιώντας λοιπόν τον μετασχηματισμό Laplace και τις παραγώγους αυτού μπορούμε να έχουμε τη συνάρτηση πυκνότητας πιθανότητας και τη συνάρτηση κινδύνου του χρόνου επιβίωσης, καθώς και τη μέση τιμή και διασπορά για την μεταβλητή ευπάθειας.

Η $\kappa -$ οστή παράγωγος του μετασχηματισμού Laplace της ευπάθειας δίνεται από την σχέση:

$$\boldsymbol{L}^{(\kappa)}(s) = (-1)^\kappa \boldsymbol{E}(Z^\kappa \exp(-Zs)),$$

όπου $s = H_0(t)$. Επιπλέον έχουμε:

$$f(t) = (-1)\frac{dS(t)}{dt} = (-1)\int_0^\infty \frac{d}{dt}\big\{\exp\big(-zH_0(t)\big)\, f_Z(z)\big\}dz =$$
$$(-1)\int_0^\infty -z\cdot h_0(t)\exp\big(-zH_0(t)\big)\, f_Z(z)dz = (-1)h_0(t)\int_0^\infty -z\cdot \exp\big(-zH_0(t)\big)\, f_Z(z)dz = -\, h_0(t)\boldsymbol{L}^{(1)}\big(H_0(t)\big).$$

Άρα, η συνάρτηση πυκνότητας πιθανότητας δίνεται από τη σχέση:

$$f(t) = -h_0(t)\boldsymbol{L}'\big(H_0(t)\big), \text{ όπου } \boldsymbol{L}' = \boldsymbol{L}^{(1)} \tag{3.5}$$

Η συνάρτηση κινδύνου δίνεται από τη γνωστή σχέση $h(t) = f(t)/S(t)$ όπως έχει αναφερθεί σε προηγούμενο κεφάλαιο. Συνεπώς, χρησιμοποιώντας τις σχέσεις (3.4) και (3.5) λαμβάνουμε ότι

$$h(t) = \frac{f(t)}{S(t)} = -\frac{h_0(t)\boldsymbol{L}'\big(H_0(t)\big)}{\boldsymbol{L}[H_0(t)]} = -h_0(t)\frac{\boldsymbol{L}'\big(H_0(t)\big)}{\boldsymbol{L}\big(H_0(t)\big)}. \quad (3.6)$$

Εύκολα προκύπτουν η μέση τιμή και η διασπορά της τυχαίας μεταβλητής ευπάθειας Z, χρησιμοποιώντας τη σχέση, που δίνει την παράγωγο του μετασχηματισμού Laplace είναι άμεσο ότι:

$$\boldsymbol{E}(Z) = -\boldsymbol{L}'(0) \quad (3.7)$$

Επιπλέον, $\boldsymbol{L}''(0) = (-1)^2\boldsymbol{E}(Z^2) = \boldsymbol{E}(Z^2)$, άρα $V(Z) = \boldsymbol{E}(Z^2) - (\boldsymbol{E}(Z))^2$

$$\boldsymbol{V}(Z) = \boldsymbol{L}''(0) - \big(\boldsymbol{L}'(0)\big)^2 \quad (3.8)$$

Σε όλα τα παραπάνω θεωρούμε ότι υπάρχουν και ορίζονται τα ολοκληρώματα και οι παράγωγοι των παραπάνω εκφράσεων (Janssen & Duchateau, 2008) (Hanagal, 2011).

Σημείωση για το Μετασχηματισμό Laplace:

Ο μετασχηματισμός Laplace $\boldsymbol{L}(s)$ μίας συνάρτησης g ορίζεται ως εξής:

$$\boldsymbol{L}(s) = \int_0^\infty g(u)e^{-us}\,du$$

στην περίπτωση που το ολοκλήρωμα υπάρχει. Στην περίπτωση που η συνάρτηση g είναι μία συνάρτηση πυκνότητας πιθανότητας τότε το εν λόγω ολοκλήρωμα για $s = 0$ υπάρχει πάντα και ισχύει ότι $L(0) = 1$. Αν συμβολίσουμε τη μέση τιμή και την διασπορά της κατανομής μ, σ^2, τότε ισχύουν τα εξής:

$$\boldsymbol{L}'(0) = -\mu$$

$$\boldsymbol{L}''(0) = \mu^2 + \sigma^2.$$

3.3 Μοντέλα Ευπάθειας στην Μονομεταβλητή περίπτωση

Το μοντέλο παλινδρόμησης ευπάθειας ή μοντέλο τυχαίων επιδράσεων για δεδομένα επιβίωσης ως επέκταση του μοντέλου αναλόγων κινδύνων του Cox ορίζεται ως εξής:

$$h(t|\boldsymbol{X}, Z) = Zh_0(t)g(\boldsymbol{X}) \quad (3.9)$$

όπου $h(t|\boldsymbol{X}, Z)$ είναι ο κίνδυνος θανάτου ενός ασθενούς την χρονική στιγμή t δεδομένων των παραγόντων που επηρεάζουν τον χρόνο επιβίωσής του. Επιπλέον, η ποσότητα $h_0(t)$ είναι το επίπεδο αναφοράς της συνάρτησης κινδύνου (βασική συνάρτηση κινδύνου) για έναν ασθενή (*arbitrary baseline hazard rate*) και αντιστοιχεί στην πιθανότητα θανάτου ενός ασθενή τη χρονική στιγμή t όταν όλες οι επεξηγηματικές μεταβλητές λαμβάνουν την τιμή 0. Αποτελεί τη βάση του μοντέλου παλινδρόμησης και χρησιμοποιείται ως μία βάση σύγκρισης για τον κίνδυνο θανάτου ενός ασθενή. Η ευπάθεια λειτουργεί πολλαπλασιαστικά στην συνάρτηση $h_0(t)$. Η τυχαία μεταβλητή Z είναι μη αρνητική και καλείται μεταβλητή ευπάθειας (*frailty variable*). Εκφράζει τη μη παρατηρούμενη πληροφορία, που επηρεάζει τη συνάρτηση κινδύνου και λαμβάνει διαφορετική τιμή για κάθε μονάδα του πληθυσμού.

Αν $0 < Z < 1$ τότε ο ασθενής είναι λιγότερο ευπαθής συγκριτικά με το επίπεδο αναφοράς, δηλαδή η πιθανότητα να επέλθει το υπό μελέτη γεγονός για αυτόν τον ασθενή μειώνεται, ενώ αν $Z > 1$ ο ασθενής είναι περισσότερο ευπαθής από το μέσο ασθενή – το επίπεδο αναφοράς.

Η $g(\cdot)$ ορίζεται όπως και στο μοντέλο του Cox λαμβάνει τη μορφή $g(\widetilde{\boldsymbol{X}}) = \exp(\beta_1 X_1 + \cdots + \beta_p X_p)$, όπου $\beta_1, \ldots, \beta_p$ είναι οι συντελεστές του μοντέλου που αντιστοιχούν σε κάθε μία επεξηγηματική μεταβλητή $X_1, X_2, \ldots, X_p$.

Σε τυχαίο δείγμα μεγέθους n και χρησιμοποιώντας τους συμβολισμούς του προηγούμενου κεφαλαίου, η συνάρτηση πιθανοφάνειας για το μοντέλο ευπάθειας (3.9) μπορεί να γραφεί ως εξής:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} [Z_i h_0(t_i) \exp(\boldsymbol{\beta}' \boldsymbol{X}_i)]^{\delta_i} \exp(-Z_i H_0(t_i) \exp(\boldsymbol{\beta}' \boldsymbol{X}_i)). \tag{3.10}$$

Η παραπάνω συνάρτηση πιθανοφάνειας (3.10) αποτελεί τη βάση για την εκτίμηση των παραμέτρων του μοντέλου.

Η ευπάθεια σαν τυχαία μεταβλητή μπορεί να ακολουθεί διάφορες κατανομές, οι οποίες χρησιμοποιούνται ευρέως σε πολλές εφαρμογές στην Ανάλυση Επιβίωσης, όπως η Γάμμα, η Positive stable, η Inverse Gaussian, η Log-Normal, η Compound Poisson, Weibull κλπ. Θα αναλύσουμε παρακάτω τις πιο συνήθεις σε πρακτικές εφαρμογές, οι οποίες είναι η Γάμμα, η Positive stable και η Inverse Gaussian. Χρησιμοποιώντας τον μετασχηματισμό Laplace για αυτές τις κατανομές είναι εύκολο να υπολογιστούν οι εκτιμήσεις των παραμέτρων των μοντέλων παλινδρόμησης ευπάθειας.

3.3.1 Γάμμα μοντέλο ευπάθειας

Η κατανομή Γάμμα αποτελεί μία γενίκευση της Εκθετικής κατανομής. Λαμβάνει δύο παραμέτρους α και β και οι συναρτήσεις που εξετάζονται στην ανάλυση επιβίωσης δίνονται από τις παρακάτω σχέσεις:

Η συνάρτηση πυκνότητας πιθανότητας είναι

$$f_Z(z) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{\alpha-1} e^{-\beta z}, z > 0, \alpha > 0, \beta > 0. \tag{3.11}$$

Χρησιμοποιώντας τον μετασχηματισμό Laplace λαμβάνουμε ότι:

$$\begin{aligned} \boldsymbol{L}(s) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty e^{-sz} z^{\alpha-1} e^{-\beta z} dz \\ &= \left(1 + \frac{s}{\beta}\right)^{-\alpha}. \end{aligned} \tag{3.12}$$

Η πρώτη και η δεύτερη παράγωγος του μετασχηματισμου Laplace είναι:

$$\boldsymbol{L}'(s) = -\frac{a}{\beta}\left(1 + \frac{s}{\beta}\right)^{-\alpha-1}$$
$$\boldsymbol{L}''(s) = \frac{a(\alpha+1)}{\beta^2}\left(1 + \frac{s}{\beta}\right)^{-\alpha-2}.$$

Για $s = 0$, λαμβάνουμε ότι:

$$\boldsymbol{E}(Z) = \frac{\alpha}{\beta}$$
$$\boldsymbol{V}(Z) = \frac{a(\alpha+1)}{\beta^2} - \frac{\alpha^2}{\beta^2} = \frac{a}{\beta^2}.$$

Για να είναι το μοντέλο καλώς ορισμένο (identifiable), τότε θα πρέπει να κάνουμε την εξής παραδοχή για την Γάμμα κατανομή, $\alpha = \beta$. Έτσι, $\boldsymbol{E}(Z) = 1$ και $\boldsymbol{V}(Z) = \frac{1}{\beta}$.

Χρησιμοποιώντας τις σχέσεις (3.4) και (3.6) και τις παραπάνω σχέσεις για την τυχαία μεταβλητή Z και αν θεωρήσουμε ότι η διασπορά είναι η άγνωστη παράμετρος για την κατανομή Γάμμα με $\frac{1}{\sigma^2} = \beta = \alpha$, λαμβάνουμε τις εξής σχέσεις για την συνάρτηση επιβίωσης, την συνάρτηση κινδύνου, καθώς και την συνάρτηση πυκνότητας πιθανότητας για τον χρόνο επιβίωσης.

$$S(t) = \boldsymbol{L}[H_0(t)] = \left(1 + \sigma^2 H_0(t)\right)^{-\frac{1}{\sigma^2}} \tag{3.13}$$

$$h(t) = -h_0(t)\frac{\boldsymbol{L}'\left(H_0(t)\right)}{\boldsymbol{L}\left(H_0(t)\right)} = \frac{h_0(t)}{1 + \sigma^2 H_0(t)}$$

$$f(t) = -h_0(t)\boldsymbol{L}'\left(H_0(t)\right) = \frac{h_0(t)}{\left(1 + \sigma^2 H_0(t)\right)^{\frac{1}{\sigma^2}+1}} \ . \tag{3.14}$$

Γενικεύοντας τα παραπάνω αποτελέσματα στην περίπτωση του μοντέλου (3.9) χρησιμοποιώντας ως συνάρτηση αναφοράς την αθροιστική συνάρτηση κινδύνου δεδομένου των συμμεταβλητών $H_0(t|\,\boldsymbol{X}) = H_0(t)\exp(\boldsymbol{\beta'X})$ λαμβάνουμε κάποια αποτελέσματα χρήσιμα για την εκτίμηση των παραμέτρων του μοντέλου, ιδιαίτερα στην περίπτωση ενός ημι-παραμετρικού Γάμμα μοντέλου παλινδρόμησης ευπάθειας (Wienke, 2011).

Η συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής Z δεδομένου των επεξηγηματικών μεταβλητών του μοντέλου για αυτούς που έχουν επιβιώσει μέχρι τη χρονική στιγμή t $(T > t)$ μπορεί να γραφεί στη μορφή

$$\begin{aligned} f(z|\boldsymbol{X}, T > t) &= \frac{S(t|\boldsymbol{X}, z)f(z)}{S(t|\boldsymbol{X})} \\ &= \frac{\exp(-zH_0(t)\exp(\boldsymbol{\beta'X}))\, z^{\frac{1}{\sigma^2}-1}\exp\left(-\frac{z}{\sigma^2}\right)}{\Gamma\left(\frac{1}{\sigma^2}\right)\sigma^{\frac{1}{\sigma^2}}(1 + \sigma^2 H_0(t)\exp(\boldsymbol{\beta'X}))^{-\frac{1}{\sigma^2}}} \\ &= \frac{(\frac{1}{\sigma^2} + H_0(t)\exp(\boldsymbol{\beta'X}))^{\frac{1}{\sigma^2}}}{\Gamma\left(\frac{1}{\sigma^2}\right)} z^{\frac{1}{\sigma^2}-1}\exp(-z\left(\frac{1}{\sigma^2} + H_0(t)\exp(\boldsymbol{\beta'X})\right), \end{aligned}$$

η οποία αποτελεί Γάμμα κατανομή με παραμέτρους $\alpha = \frac{1}{\sigma^2}$ και $\beta = \frac{1}{\sigma^2} + H_0(t)\exp(\boldsymbol{\beta'X})$.

Επιπλέον, για $T = t$, δηλαδή για αυτούς που έχουν χρόνο επιβίωσης t, η συνάρτηση πυκνότητας πιθανότητας της Z δεδομένου των επεξηγηματικών μεταβλητών δίνεται από τη σχέση:

$$\begin{aligned} f(z|\boldsymbol{X}, T = t) &= \frac{f(t|\boldsymbol{X}, z)f(z)}{f(t|\boldsymbol{X})} \\ &= \frac{zh_0(t)\exp(-zH_0(t)\exp(\boldsymbol{\beta'X}))\, z^{\frac{1}{\sigma^2}-1}\exp\left(-\frac{z}{\sigma^2}\right)}{\Gamma\left(\frac{1}{\sigma^2}\right)\sigma^{\frac{1}{\sigma^2}}h_0(t)(1 + \sigma^2 H_0(t)\exp(\boldsymbol{\beta'X}))^{-\frac{1}{\sigma^2}-1}} \end{aligned}$$

$$= \frac{\left(\frac{1}{\sigma^2} + H_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{X})\right)^{\frac{1}{\sigma^2}+1}}{\Gamma\left(\frac{1}{\sigma^2}\right)\frac{1}{\sigma^2}} z^{\frac{1}{\sigma^2}}\exp(-z\left(\frac{1}{\sigma^2} + H_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{X})\right)$$

$$= \frac{(\frac{1}{\sigma^2} + H_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{X}))^{\frac{1}{\sigma^2}+1}}{\Gamma\left(\frac{1}{\sigma^2}+1\right)} z^{\frac{1}{\sigma^2}}\exp(-z\left(\frac{1}{\sigma^2} + H_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{X})\right),$$

η οποία αποτελεί Γάμμα κατανομή με παραμέτρους $\alpha = \frac{1}{\sigma^2} + 1$ και $\beta = \frac{1}{\sigma^2} + H_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{X})$.

Συνεπώς, η μέση τιμή και η διασπορά της τυχαίας μεταβλητής ευπάθειας και στις δύο περιπτώσεις, δηλαδή για χρόνο επιβίωσης μεγαλύτερο από την τιμή t και ίσο με t δίνονται από τις σχέσεις:

$$E(Z|\boldsymbol{X}, T > t) = \frac{a}{\beta} = \frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma^2} + H_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{X})} = \frac{1}{1 + \sigma^2 H_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{X})} \tag{3.15}$$

$$V(Z|\boldsymbol{X}, T > t) = \frac{a}{\beta^2} = \frac{\frac{1}{\sigma^2}}{\left(\frac{1}{\sigma^2} + H_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{X})\right)^2} = \frac{\sigma^2}{(1 + \sigma^2 H_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{X}))^2} \tag{3.16}$$

$$E(Z|\boldsymbol{X}, T = t) = \frac{a}{\beta} = \frac{\frac{1}{\sigma^2} + 1}{\frac{1}{\sigma^2} + H_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{X})} = \frac{1 + \sigma^2}{1 + \sigma^2 H_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{X})} \tag{3.17}$$

$$V(Z|\boldsymbol{X}, T = t) = \frac{a}{\beta^2} = \frac{\frac{1}{\sigma^2} + 1}{\left(\frac{1}{\sigma^2} + H_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{X})\right)^2} = \frac{\sigma^2(1 + \sigma^2)}{(1 + \sigma^2 H_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{X}))^2} \tag{3.18}$$

3.3.2 Γάμμα παραμετρικά μοντέλα ευπάθειας

Όπως έχουμε ήδη αναφέρει μπορούμε να χρησιμοποιήσουμε είτε παραμετρική προσέγγιση, είτε ημι- παραμετρική, ώστε να εκτιμήσουμε τις παραμέτρους του μοντέλου. Θα παρουσιάσουμε και τις δύο εκδοχές για το Γάμμα μοντέλο ευπάθειας, όπως και για τα υπόλοιπα μοντέλα ευπάθειας, που θα αναλύσουμε στο συγκεκριμένο κεφάλαιο.

Με βάση τη σχέση (3.10) η συνάρτηση πιθανοφάνειας για το μοντέλο ευπάθειας στη γενική του μορφή δίνεται από τη σχέση:

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}|Z_1, Z_2, \dots, Z_n) = \prod_{i=1}^{n}[z_i h_0(t_i; \boldsymbol{\theta})\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)]^{\delta_i}\exp(-z_i H_0(t_i; \boldsymbol{\theta})\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)), \tag{3.19}$$

όπου $\boldsymbol{\theta}$ είναι το άγνωστο διάνυσμα παραμέτρων που υπεισέρχεται στην βασική συνάρτηση κινδύνου (*baseline hazard function*).

Στην περίπτωση που η ευπάθεια ακολουθεί Γάμμα κατανομή και χρησιμοποιώντας τις σχέσεις (3.11) και (3.12) λαμβάνουμε την εξής μορφή για τη συνάρτηση Πιθανοφάνειας:

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) = \prod_{i=1}^{n} \left[\frac{h_0(t_i; \boldsymbol{\theta}) \exp(\boldsymbol{\beta}' \boldsymbol{x}_i)}{(1 + \sigma^2 H_0(t_i; \boldsymbol{\theta}) \exp(\boldsymbol{\beta}' \boldsymbol{x}_i))^{\frac{1}{\sigma^2}}} \right]^{\delta_i} (1 + \sigma^2 H_0(t_i; \boldsymbol{\theta}) \exp(\boldsymbol{\beta}' \boldsymbol{x}_i))^{-\frac{1}{\sigma^2}} \quad (3.20)$$

Χρησιμοποιώντας τις σχέσεις (3.15) και (3.17) η ευπάθεια για κάθε ασθενή $Z_i, i = 1, \dots, n$ εκτιμάται από την αναμενόμενη τιμή που τελικά δίνεται από τη σχέση

$$\hat{Z}_i = \frac{\frac{1}{\hat{\sigma}^2} + \delta_i}{\frac{1}{\hat{\sigma}^2} + H_0(t_i; \hat{\boldsymbol{\theta}}) \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{X}_i)}, \quad (3.21)$$

όπου $\hat{\boldsymbol{\theta}}$ είναι η εκτίμηση του διανύσματος παραμέτρων που υπεισέρχεται στην βασική συνάρτηση κινδύνου (*baseline hazard function*), $\hat{\boldsymbol{\beta}}$ είναι η εκτίμηση των παραμέτρων του μοντέλου, δ_i είναι ένας δείκτης που λαμβάνει την τιμή 1 αν έχουμε χρόνο επιβίωσης και τιμή 0 αν έχουμε λογοκριμένο χρόνο και $\hat{\sigma}^2$ είναι η εκτίμηση της διασποράς της μεταβλητής της ευπάθειας.

3.3.3 Γάμμα ημι-παραμετρικά μοντέλα ευπάθειας

Για την εκτίμηση των παραμέτρων στα ημι-παραμετρικά μοντέλα επιβίωσης που δεν περιλαμβάνουν μεταβλητή ευπάθειας χρησιμοποιείται η μέθοδος μέγιστης πιθανοφάνειας με βάση τη μερική πιθανοφάνεια. Ωστόσο, στην περίπτωση των ημι-παραμετρικών μοντέλων ευπάθειας πρέπει να λάβουμε υπόψη τη συμμετοχή του παράγοντα-ευπάθεια στο μοντέλο. Σε αυτήν την περίπτωση η εκτίμηση των παραμέτρων γίνεται χρησιμοποιώντας την μερική πιθανοφάνεια με χρήση του ΕΜ αλγορίθμου (*Expectation Maximization algorithm*). Στην περίπτωση των ημι-παραμετρικών μοντέλων η βασική συνάρτηση κινδύνου θεωρείται μία (άγνωστη) οχληρή μεταβλητή (*nuisance*) και δεν κάνουμε καμία υπόθεση για την κατανομή της. Θεωρούμε ότι δεν υπάρχει για αυτήν παρατηρούμενη πληροφορία.

Ο ΕΜ αλγόριθμος αποτελείται από δύο βήματα, Expectation step και Maximization step. Στο expectation step, υπολογίζονται οι αναμενόμενες τιμές για τις μη παρατηρούμενες μεταβλητές ευπάθειας δεδομένων των παρατηρήσεων και των εκτιμημένων παραμέτρων. Στο Maximization step οι αναμενόμενες τιμές που έχουν υπολογιστεί λαμβάνονται υπόψη ως πραγματική πληροφορία και υπολογίζονται νέες εκτιμήσεις για τις παραμέτρους μεγιστοποιώντας την συνάρτηση πιθανοφάνειας. Παρακάτω θα παρουσιάσουμε τον ΕΜ αλγόριθμο για το Γάμμα μοντέλο ευπάθειας (Janssen & Duchateau, 2008) (Klein & Moeschberger, 2003) (Wienke, 2011).

Θεωρούμε τις μεταβλητές ευπάθειας ως παρατηρούμενες τυχαίες μεταβλητές Z_i. Τότε, η από κοινού συνάρτηση πιθανοφάνειας για το τυχαίο δείγμα $(t_i,\ \delta_i,\ Z_i),\ i = 1,2, \dots, n$ μπορεί να γραφεί στην εξής μορφή:

$$L(\boldsymbol{\beta}, \sigma^2 | t_i, \delta_i, Z_i) = \prod_{i=1}^{n} f(t_i, \delta_i, Z_i; \boldsymbol{\beta}, \sigma^2)$$

$$= \prod_{i=1}^{n} f(t_i, \delta_i, \boldsymbol{\beta} |\ Z_i) \prod_{i=1}^{n} f(\ Z_i; \sigma^2), \quad (3.22)$$

$$= L_1(\boldsymbol{\beta} | \boldsymbol{Z}) L_2(\sigma^2 | \boldsymbol{Z}),$$

όπου $\boldsymbol{Z} = (Z_1, Z_2, \dots, Z_n)$ το τυχαίο δείγμα των ευπαθειών και $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)$ το διάνυσμα των αγνώστων παραμέτρων του μοντέλου ευπάθειας.

Από τη σχέση (3.10) προκύπτει ότι

$$L_1(\boldsymbol{\beta}|\boldsymbol{Z}) = \prod_{i=1}^{n} [Z_i h_0(t_i) \exp(\boldsymbol{\beta'} \boldsymbol{X}_i)]^{\delta_i} \exp(-Z_i H_0(t_i) \exp(\boldsymbol{\beta'} \boldsymbol{X}_i))$$

ενώ ο δεύτερος όρος της σχέσης (3.22) προκύπτει από τη συνάρτηση πυκνότητας πιθανότητας της μεταβλητής ευπάθειας

$$L_2(\sigma^2|\boldsymbol{Z}) = \prod_{i=1}^{n} f_Z(Z_i; \sigma^2).$$

Αν τα $Z_i, i = 1, 2, \dots, n$ είχαν παρατηρηθεί τότε οι εκτιμήσεις των παραμέτρων $\boldsymbol{\beta}$ θα μπορούσαν εύκολα να προκύψουν αν αντικαταστήσουμε στη σχέση L_1 τους όρους $Z_i exp(\boldsymbol{\beta'} \boldsymbol{X}_i)$ από $exp(\boldsymbol{\beta'} \boldsymbol{X}_i) + log(Z_i)$ και ακολούθως εφαρμόσουμε τη μέθοδο μέγιστης Πιθανοφάνειας στη Μερική συνάρτηση Πιθανοφάνειας, η οποία προκύπτει θεωρώντας τους όρους $\log(Z_i)$ ως fixed offset values. Υπολογίζουμε λοιπόν εκτιμήσεις για τις μεταβλητές ευπάθειας οι οποίες αντιστοιχούν στις αναμενόμενες τιμές των τυχαίων μεταβλητών Z_i και $\log(Z_i)$. Οι εκτιμήσεις αυτές χρησιμοποιούνται για την μεγιστοποίηση της συνάρτησης Πιθανοφάνειας, ώστε τελικά να υπολογίσουμε τις άγνωστες παραμέτρους του μοντέλου.

Maximization step:

Με βάση τη Μερική συνάρτηση Πιθανοφάνειας που παρουσιάστηκε στο μοντέλο του Cox μπορεί να γραφεί η αντίστοιχη συνάρτηση για το μοντέλο ευπάθειας. Τότε, λαμβάνουμε τη σχέση:

$$L(\boldsymbol{\beta}|\boldsymbol{Z}) = \prod_{i=1}^{n} \left(\frac{e^{\boldsymbol{\beta'} \boldsymbol{X}_i + log(Z_i)}}{\sum_{j \in R(t_i)} Z_j e^{\boldsymbol{\beta'} \boldsymbol{X}_j}} \right)^{\delta_i}.$$

Οι τυχαίες μεταβλητές Z_i και $logZ_i$ μπορούν να αντικατασταθούν από τις αναμενόμενες τιμές τους. Τότε ο λογάριθμος της παραπάνω πιθανοφάνειας γράφεται:

$$logL(\beta, \sigma^2) = \sum_{i=1}^{n} \delta_i \left[\boldsymbol{\beta'} \boldsymbol{X}_i + \boldsymbol{E}_{(\kappa)}\big(log(Z_i)\big) - \log\left(\sum_{j \in R(t_i)} \boldsymbol{E}_{(\kappa)}(Z_j) \mathrm{e}^{\boldsymbol{\beta}' \boldsymbol{X}_j} \right) \right].$$

όπου ο συμβολισμός $\boldsymbol{E}_{(\kappa)}$ δηλώνει την αναμενόμενη τιμή στο $\kappa-$βήμα του αλγορίθμου. Μεγιστοποιώντας την παραπάνω σχέση μπορούμε να λάβουμε εκτιμήσεις για τις παραμέτρους με διάφορες αριθμητικές επαναληπτικές μεθόδους.

Expectation step:

Όμοια με την παραμετρική προσέγγιση η μη παρατηρούμενη ευπάθεια για κάθε ασθενή $Z_i, i = 1, \dots, n$ εκτιμάται από την αναμενόμενη τιμή που τελικά δίνεται από τη σχέση

$$\boldsymbol{E}_{(\kappa+1)}(Z_i) = \frac{\frac{1}{\sigma_{(\kappa)}{}^2} + \delta_i}{\frac{1}{\sigma_{(\kappa)}{}^2} + H_{0\kappa}(t_i; \boldsymbol{\theta}) \exp\left(\boldsymbol{\beta}'_{(\kappa)} \boldsymbol{X}_i\right)}, \tag{3.23}$$

Να σημειώσουμε εδώ ότι $H_{0\kappa}(\cdot)$ είναι ένας μη παραμετρικός εκτιμητής της αθροιστικής βασικής συνάρτησης κινδύνου με βάση τις εκτιμήσεις μέχρι το $\kappa-$ βήμα. Για παράδειγμα χρησιμοποιώντας τον εκτιμητή Nelson-Aalen μία εκτίμηση στο $\kappa-$ βήμα θα μπορούσε να είναι η εξής:

$$H_{0\kappa}(t)=\sum_{i:t_i<t}\frac{\delta_i}{\sum_{j\in R(t_i)}\boldsymbol{E}_{(\kappa)}(Z_j)e^{\boldsymbol{\beta}'_{(\kappa)}\boldsymbol{X}_j}}.$$

3.3.4 Inverse Gaussian μοντέλο ευπάθειας

Η συνάρτηση πυκνότητας πιθανότητας της Inverse Gaussian κατανομής δίνεται από τη σχέση:

$$f(z)=\frac{\sqrt{\lambda}}{\sqrt{2\pi z^3}}\exp\left(-\frac{\lambda}{2\mu^2 z}(z-\mu)^2\right), z>0, \lambda>0, \mu>0. \tag{3.24}$$

$$f(z)=\frac{1}{\sqrt{2\pi\theta}}z^{-3/2}\exp\left(-\frac{1}{2\theta z}(z-1)^2\right), z>0, \theta>0.$$

Ο μετασχηματισμός Laplace για μία τυχαία μεταβλητή που ακολουθεί την Inverse Gaussian κατανομή αποδεικνύεται ότι δίνεται από τη σχέση:

$$L(s)=E(e^{-sZ})=\exp\left(\frac{\lambda}{\mu}\left(1-\sqrt{1+\frac{2\mu^2 s}{\lambda}}\right)\right). \tag{3.25}$$

Χρησιμοποιώντας την πρώτη και τη δεύτερη μερική παράγωγο της παραπάνω σχέσης και θέτοντας $s=0$ λαμβάνουμε τις εκφράσεις για τη μέση τιμή και τη διασπορά της τυχαίας μεταβλητής ευπάθειας Z,

$$E(Z)=-L'(0)=\mu \tag{3.26}$$

και

$$V(Z)=L''(0)-\left(L'(0)\right)^2=\frac{\mu^3}{\lambda}. \tag{3.27}$$

Υπό την υπόθεση $E(Z)=\mu=1$, έχουμε ότι $V(Z)=\frac{1}{\lambda}=\sigma^2$ και έτσι ο μετασχηματισμός Laplace μπορεί να γραφεί σε πιο απλή μορφή:

$$L(s)=\exp\left(\lambda\left(1-\sqrt{1+\frac{2s}{\lambda}}\right)\right)=\exp\left(\frac{1}{\sigma^2}\left(1-\sqrt{1+2s\sigma^2}\right)\right).$$

Συνεπώς, η συνάρτηση επιβίωσης και η συνάρτηση κινδύνου για το χρόνο επιβίωσης από τις σχέσεις (3.4) και (3.6) μπορούν να γραφούν ως εξής:

$$S(t)=\exp\left(\frac{1}{\sigma^2}\left(1-\sqrt{1+2H_0(t)\sigma^2}\right)\right) \tag{3.28}$$

και

$$h(t)=\frac{h_0(t)}{(1+2H_0(t)\sigma^2)^{1/2}} \tag{3.29}$$

Η συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής Z δεδομένων των επεξηγηματικών μεταβλητών του μοντέλου για αυτούς που έχουν επιβιώσει μέχρι τη χρονική στιγμή t $(T > t)$ μπορεί να γραφεί στη μορφή

$$f(z|\boldsymbol{X}, T > t) = \frac{S(t|\boldsymbol{X}, z) f(z)}{S(t|\boldsymbol{X})}$$

$$= \frac{1}{\sqrt{2\pi^2 z^3}} \exp\left(- \frac{\left(z - (1 + 2\sigma^2 H_0(t) \exp(\boldsymbol{\beta}'\boldsymbol{X}))^{-\frac{1}{2}}\right)^2}{\frac{2\sigma^2 z}{1 + 2\sigma^2 H_0(t) \exp(\boldsymbol{\beta}'\boldsymbol{X})}} \right).$$

Συνεπώς, η παραπάνω σχέση αποτελεί τη συνάρτηση πυκνότητας πιθανότητας μίας Inverse Gaussian κατανομής με μέση τιμή διασπορά που δίνεται από τις σχέσεις (Wienke, 2011):

$$E(Z|\boldsymbol{X}, T > t) = \frac{1}{\sqrt{1 + \sigma^2 H_0(t) \exp(\boldsymbol{\beta}'\boldsymbol{X})}} \tag{3.30}$$

$$V(Z|\boldsymbol{X}, T > t) = \frac{\sigma^2}{(1 + \sigma^2 H_0(t) \exp(\boldsymbol{\beta}'\boldsymbol{X}))^2} \tag{3.31}$$

3.3.5 Positive Stable μοντέλο ευπάθειας

Η συνάρτηση πυκνότητας πιθανότητας της Positive stable κατανομής δίνεται από τη σχέση:

$$f(z) = \frac{1}{\pi} \sum_{\kappa=1}^{\infty} (-1)^{k+1} \frac{\Gamma(\kappa\gamma + 1)}{\kappa!} z^{-\kappa\gamma - 1} \sin(\kappa\gamma\pi), \quad z \geq 0, \ \ 0 < \gamma \leq 1.$$

$$f(z) = -\frac{1}{\pi z} \sum_{\kappa=1}^{\infty} (-z^{\nu-1})^k \frac{\Gamma(\kappa(1-\nu) + 1)}{\kappa!} \sin((1-\nu)\kappa\pi), \quad z \geq 0, \ \ 0 < \nu < 1.$$

Η παραπάνω έκφραση αποτελεί μια δυναμοσειρά, η οποία συγκλίνει γρήγορα για μεγάλες τιμές του z και παρουσιάζει πιο αργή σύγκλιση για μικρές τιμές του z. Παρ' όλο που η μορφή της συνάρτησης καθιστά δύσκολους τους υπολογισμούς, ο μετασχηματισμός Laplace έχει πολύ απλή μορφή, για αυτό αποτελεί μία καλή επιλογή για μοντέλα ευπάθειας στην ανάλυση Επιβίωσης.

Ο μετασχηματισμός Laplace για μία τυχαία μεταβλητή που ακολουθεί την Positive Stable κατανομή αποδεικνύεται ότι δίνεται από τη σχέση:

$$L(s) = E\{e^{-sZ}\} = e^{-s^{\gamma}} \tag{3.32}$$

Η πρώτη παράγωγος του μετασχηματισμού Laplace τείνει στο άπειρο, άρα και η αναμενόμενη τιμή της μεταβλητής ευπάθειας απειρίζεται και έτσι η διασπορά της Z δεν ορίζεται. Αυτός είναι ο κύριος λόγος που η εν λόγω κατανομή χρησιμοποιήθηκε στα μοντέλα ευπάθειας, καθώς η πεπερασμένη μέση τιμή της κατανομής της ευπάθειας είναι μία μόνο απαίτηση (ανάμεσα σε άλλες) για την αναγνωρισιμότητα των παραμέτρων σε μονομεταβλητά μοντέλα ευπάθειας (*univariate frailty models*) (Wienke, 2011).

Η συνάρτηση πυκνότητας πιθανότητας, η συνάρτηση επιβίωσης και η συνάρτηση κινδύνου για το χρόνο επιβίωσης, χρησιμοποιώντας τον μετασχηματισμό Laplace που δόθηκε παραπάνω λαμβάνουν την εξής μορφή:

$$S(t) = e^{-H_0(t)^\gamma} \tag{3.33}$$

$$f(t) = \gamma h_0(t) H_0(t)^{\gamma-1} e^{-H_0(t)^\gamma} \tag{3.34}$$

$$h(t) = \gamma h_0(t) H_0(t)^{\gamma-1}. \tag{3.35}$$

3.4 Από κοινού μοντέλα ευπάθειας

Η μονοδιάστατη ανάλυση που αναφέραμε παραπάνω για τα μοντέλα ευπάθειας στηρίζεται στην παραδοχή ότι οι χρόνοι επιβίωσης διαφορετικών μονάδων του πληθυσμού είναι ανεξάρτητες τυχαίες μεταβλητές. Ωστόσο, υπάρχουν περιπτώσεις όπου έχουμε πληροφορία για επαναλαμβανόμενα γεγονότα ή γεγονότα που μπορούν να ομαδοποιηθούν. Για παράδειγμα μπορεί να εξετάζουμε άτομα τα οποία έχουν κάποια κοινά γενετικά χαρακτηριστικά και η ανάλυση εστιάζει σε αυτά, ή να γίνεται ανάλυση σε παντρεμένα ζευγάρια τα οποία έχουν ως κοινό χαρακτηριστικό το περιβάλλον στο οποίο ζούνε το οποίο δεν είναι κάτι μετρήσιμο. Ένα άλλο παράδειγμα θα μπορούσε να είναι η εξέταση μιας θεραπείας σε μία ομάδα ατόμων σε διαφορετικές χρονικές στιγμές. Για αυτά τα δεδομένα στην ανάλυση επιβίωσης χρησιμοποιείται πολυδιάστατη ανάλυση επιβίωσης και τα από κοινού μοντέλα επιβίωσης (*shared models*) για την καλύτερη προσαρμογή του μοντέλου.

Ας υποθέσουμε ότι έχουμε n ομάδες, κάθε ένα με $n_i, i = 1,2, \dots, n$ παρατηρήσεις. Έστω ότι τα άτομα στην ίδια ομάδα i μοιράζονται την ίδια μεταβλητή ευπάθειας $Z_i, i = 1, \dots, n$. Οι χρόνοι επιβίωσης μέσα σε κάθε ομάδα είναι ανεξάρτητοι δεδομένης της μεταβλητής ευπάθειας. Ένα από κοινού (shared) μοντέλο ευπάθειας αποτελεί ένα μοντέλο τυχαίων επιδράσεων (*random effect model*) το οποίο περιλαμβάνει μεταβλητότητα μεταξύ των groups (*frailty*) και μεταβλητότητα μεταξύ των διαφορετικών μονάδων του πληθυσμού, που εκφράζεται από την συνάρτηση κινδύνου (Wienke, 2011). Εστω ότι το διάνυσμα $\boldsymbol{X}_{ij}$ $i = 1, \dots, n, j = 1, \dots, n_i$ είναι το διάνυσμα των επεξηγηματκών μεταβλητών για το χρόνο T_{ij} της $j-$ παρατήρησης στην $i-$ ομάδα. Οι χρόνοι επιβίωσης στην $i-$ ομάδα θεωρούνται ανεξάρτητες τυχαίες μεταβλητές δεδομένης της αντίστοιχης μεταβλητής ευπάθειας Z_i που επηρεάζει τη συγκεκριμένη ομάδα. Η συνάρτηση κινδύνου για τους χρόνους επιβίωσης στην $i-$ ομάδα λαμβάνει την εξής μορφή για το από κοινού μοντέλο ευπάθειας

$$h\left(t\middle|\boldsymbol{X}_{ij}, Z_i\right) = Z_i h_0(t) e^{\boldsymbol{\beta}'\boldsymbol{X}_{ij}} \tag{3.36}$$

για $j = 1, \dots, n_i$, όπου $h_0(t)$ η βασική συνάρτηση κινδύνου και $\boldsymbol{\beta}$ το διάνυσμα των συντελεστών του μοντέλου προς εκτίμηση. Οι μεταβλητές ευπάθειας Z_i $(i = 1, 2, \dots, n)$ αποτελούν ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με συνάρτηση πυκνότητας πιθανότητας $f(z)$ (Wienke, 2011).

Σε αυτό το σημείο θα παρουσιάσουμε τη στατιστική συμπερασματολογία για τα από κοινού μοντέλο ευπάθειας, όπως αναφέρεται στο βιβλίο του Hanagal (2011). Για να αποδοθεί η συνάρτηση πιθανοφάνειας στη γενική της μορφή θεωρούμε ότι ο κοινός παράγοντας (*frailty variable*) δημιουργεί εξάρτηση μεταξύ των μονάδων σε κάθε ομάδα και δεσμεύοντας ως προς αυτόν, όλοι οι χρόνοι επιβίωσης των μονάδων που ανήκουν σε μία ομάδα είναι ανεξάρτητες μεταξύ τους τυχαίες μεταβλητές δεδομένου των $\boldsymbol{X}_{ij}$ και Z_i. Η από κοινού συνάρτηση επιβίωσης των χρόνων επιβίωσης $T_{i1}, T_{i2}, \dots, T_{in_i}$ για την $i-$ ομάδα μπορεί να γραφεί ως εξής:

$$P(T_{i1} > t_{i1}, \dots, T_{in_i} > t_{in_i} | \boldsymbol{X}_i, Z_i) = S(t_{i1}, \dots, t_{in_i} | \boldsymbol{X}_i, Z_i) = S(t_{i1} | \boldsymbol{X}_{i1}, Z_i) \cdots S(t_{in_i} | \boldsymbol{X}_{in_i}, Z_i)$$
$$= \exp\left\{-Z_i \sum_{j=1}^{n_i} H_0(t_{ij}) e^{\boldsymbol{\beta}' X_{ij}}\right\} \tag{3.37}$$

όπου $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \dots, \boldsymbol{X}_{in_i})$ είναι ο πίνακας των επεξηγηματικών μεταβλητών για τα άτομα της $i-$ομάδας. Ολοκληρώνοντας ως προς την μεταβλητή ευπάθειας την παραπάνω έκφραση λαμβάνουμε τη μη δεσμευμένη από κοινού συνάρτηση επιβίωσης των χρόνων επιβίωσης για την $i-$ομάδα.

$$S(t_{i1}, \dots, t_{in_i} | \boldsymbol{X}_i) = P(T_{i1} > t_{i1}, \dots, T_{in_i} > t_{in_i} | \boldsymbol{X}_i) = \int_0^\infty P(T_{i1} > t_{i1}, \dots, T_{in_i} > t_{in_i} | \boldsymbol{X}_i, Z_i) f(z) dz$$
$$= \mathbf{E}\{S(t_{i1}, \dots, t_{in_i} | \boldsymbol{X}_i, Z_i)\} = \mathbf{E}_Z\left(\exp\left\{-Z_i \sum_{j=1}^{n_i} H_0(t_{ij}) e^{\boldsymbol{\beta}' \boldsymbol{X}_{ij}}\right\}\right) \tag{3.38}$$
$$= \boldsymbol{L}\left(\sum_{j=1}^{n_i} H_0(t_{ij}) e^{\boldsymbol{\beta}' \boldsymbol{X}_{ij}}\right),$$

όπου $\boldsymbol{L}(\cdot)$ ο μετασχηματισμός Laplace της συνάρτησης πυκνότητας πιθανότητας $f(z)$ και $H_0(t) = \int_0^t h_0(s) ds$.

Η από κοινού συνάρτηση επιβίωσης για όλες τις ομάδες λόγω της ιδιότητας της ανεξαρτησίας μεταξύ των διαφορετικών ομάδων μπορεί να γραφεί ως εξής:

$$S(t_{i1}, \dots, t_{nn_i} | \boldsymbol{X}_{ij}) = \prod_{i=1}^{n} \boldsymbol{L}\left(\sum_{\boldsymbol{j=1}}^{n_i} H_0(t_{ij}) e^{\boldsymbol{\beta}' X_{ij}}\right). \tag{3.39}$$

Οι περιθώριες μονομεταβλητές δεσμευμένες συναρτήσεις επιβίωσης, χρησιμοποιώντας τον μετασχηματισμό Laplace μπορούν να γραφούν ως εξής:

$$S(t_{ij} | \boldsymbol{X}_{ij}) = \boldsymbol{E}[S(t_{ij} | \boldsymbol{X}_{ij}, Z_i)]$$
$$= \boldsymbol{E}\left[exp\left(-Z_i H_0(t_{ij}) e^{\boldsymbol{\beta}' X_{ij}}\right)\right]$$
$$= \boldsymbol{L}\left(H_0(t_{ij}) e^{\boldsymbol{\beta}' X_{ij}}\right). \tag{3.40}$$

Οι μη δεσμευμένη συνάρτηση επιβίωσης για την $i-$ομάδα δίνεται από τη σχέση:

$$S(t_{i1}, \dots, t_{nn_i} | \boldsymbol{X}_{ij}) = \boldsymbol{L}\left(\boldsymbol{L}^{-1}(S(t_{i1} | \boldsymbol{X}_{i1})) + \cdots + \boldsymbol{L}^{-1}\left(S(t_{in_i} | \boldsymbol{X}_{in_i})\right)\right),$$

όπου $\boldsymbol{L}^{-1}$ είναι η αντίστροφη συνάρτηση του μετασχηματισμού Laplace $\boldsymbol{L}$. Στην παραπάνω σχέση έχει χρησιμοποιηθεί η ιδιότητα της αντίστροφης συνάρτησης, που δίνει $H_0(t_{ij}) e^{\boldsymbol{\beta}' X_{ij}} = \boldsymbol{L}^{-1}\left(S(t_{ij} | \boldsymbol{X}_{ij})\right)$.

3.4.1 Εκτιμητική προσέγγιση για τα μοντέλα ευπάθειας

Έχουν αναπτυχθεί διάφορες μέθοδοι για την εκτίμηση των υπό μελέτη παραμέτρων στα μοντέλα ευπάθειας. Για τα παραμετρικά μοντέλα μπορεί να χρησιμοποιηθεί η μέθοδος μεγιστοποίησης της περιθώριας πιθανοφάνειας. Ωστόσο, η περιθώρια συνάρτηση πιθανοφάνειας είναι άμεσα υπολογίσιμη για ορισμένες επιλογές κατανομών, όπως την Γάμμα

κατανομή, ενώ οδηγεί σε πιο περίπλοκες μορφές για άλλες κατανομές και φυσικά σε ορισμένες περιπτώσεις δεν μπορεί να υπολογιστεί η εν λόγω συνάρτηση. Για το λόγο αυτό χρησιμοποιείται ο Ε-Μ αλγόριθμος για την εκτίμηση παραμέτρων αυτών των μοντέλων. Έχουν αναπτυχθεί και άλλες μέθοδοι εκτίμησης παραμέτρων, χρησιμοποιώντας τον Ε-Μ αλγόριθμο, αλλά και την μέθοδο ποινικοποιημένης μερικής πιθανοφάνειας (*penalized partial likelihood*), οι οποίες χρησιμοποιούνται σε ημιπαραμετρικά μοντέλα ευπάθειας (Janssen & Duchateau, 2008). Σε αυτήν την ενότητα θα αναφέρουμε τη θεωρία της μεθόδου μεγιστοποίησης της περιθώριας πιθανοφάνειας.

- Μέθοδος μεγιστοποίησης της περιθώριας πιθανοφάνειας (*marginal likelihood*)

Χρησιμοποιώντας τις σχέσεις (2.9) και (2.10) λαμβάνουμε ότι για την $j-$μονάδα του πληθυσμού, η συνάρτηση πιθανοφάνειας δίνεται από τη σχέση:

$$L \propto \prod_{j=1}^{n} \left(f(t_j)\right)^{\delta_j} \left(S(t_j)\right)^{1-\delta_j} \tag{3.41}$$

Με βάση την παραπάνω σχέση, η συνάρτηση πιθανοφάνειας για την $j-$μονάδα του πληθυσμού της $i-$ομάδας δίνεται από τη σχέση

$$L_\iota = \prod_{j=1}^{n_i} \left(f(t_{ij})\right)^{\delta_{ij}} \left(S(t_{ij})\right)^{1-\delta_{ij}} \tag{3.42}$$

Χρησιμοποιώντας τη σχέση (2.3), η δεσμευμένη συνάρτηση πιθανοφάνειας μπορεί να γραφεί στην εξής μορφή:

$$L_\iota = \prod_{j=1}^{n_i} \left(h(t_{ij})\right)^{\delta_{ij}} S(t_{ij}) \tag{3.43}$$

Λαμβάνοντας υπόψη τη σχέση (2.3) η γενική μορφή ενός από κοινού μοντέλου ευπάθειας, η οποία δίνεται από τη σχέση (3.36) μπορεί να γραφεί στην εξής μορφή:

$$\frac{f(t_{ij})}{S(t_{ij})} = z_i h_0(t_{ij}) \exp(\boldsymbol{\beta}' \boldsymbol{X}_{ij}) \tag{3.44}$$

Ολοκληρώνοντας και τα δύο μέρη της παραπάνω σχέσης λαμβάνουμε ότι:

$$\int_0^{t_{ij}} \frac{f(s)}{S(s)} ds = \int_0^{t_{ij}} z_i h_0(s) \exp(\boldsymbol{\beta}' \boldsymbol{X}_{ij})\, ds \tag{3.45}$$

Από σχέσεις (2.3) και (2.6) λαμβάνουμε ότι:

$$-\ln(S(t_{ij})) = H_0(t_{ij}) z_i \exp(\boldsymbol{\beta}' \boldsymbol{X}_{ij}) \tag{3.46}$$

Και τελικά

$$S(t_{ij}) = \exp\{-H_0(t_{ij}) z_i \exp(\boldsymbol{\beta}' \boldsymbol{X}_{ij})\} \tag{3.47}$$

Η δεσμευμένη πιθανοφάνεια για την $i-$ομάδα ασθενών δίνεται από τη σχέση:

$$L_i(\boldsymbol{\xi}, \boldsymbol{\beta} | z_i) = \prod_{j=1}^{n_i} (h_0(t_{ij}) z_i \exp(\boldsymbol{\beta}' \boldsymbol{X}_{ij}))^{\delta_{ij}} \exp(-H_0(t_{ij}) z_i \exp(\boldsymbol{\beta}' \boldsymbol{X}_{ij})) \tag{3.48}$$

όπου $H_0(t) = \int_0^t h(s)ds$ το επίπεδο αναφοράς της αθροιστικής συνάρτησης κινδύνου, $\boldsymbol{\xi}$ το διάνυσμα των παραμέτρων της κατανομής, που έχουμε θεωρήσει για το επίπεδο αναφοράς της συνάρτησης κινδύνου $h_0(t)$, $\boldsymbol{\beta}$ το διάνυσμα των συντελεστών του μοντέλου και z_i η μεταβλητή ευπάθειας για την $i-$ ομάδα ασθενών.

Ολοκληρώνοντας την παραπάνω σχέση ως προς την μεταβλητή ευπάθειας, λαμβάνουμε την περιθώρια συνάρτηση πιθανοφάνειας, η οποία μπορεί να γραφεί ως εξής:

$$L_{marginal,i}(\boldsymbol{\xi}, \theta, \boldsymbol{\beta}) = \int_0^{\infty} \prod_{j=1}^{n_i} (h_0(t_{ij}) s \exp(\boldsymbol{\beta}' \boldsymbol{x}_{ij}))^{\delta_{ij}} \exp(-H_0(t_{ij}) s \exp(\boldsymbol{\beta}' \boldsymbol{x}_{ij})) \times f_z(s;\theta) ds \quad (3.49)$$

όπου θ είναι η άγνωστη παράμετρος της κατανομής, που έχουμε υποθέσει για την μεταβλητή ευπάθειας. Συνήθως, για τις περισσότερες επιλογές κατανομών ισοδυναμεί με την διασπορά της κατανομής, δηλαδή τη διασπορά της μεταβλητής ευπάθειας.

Αν θεωρήσουμε ότι το ολοκλήρωμα υπάρχει και μπορούμε να το υπολογίσουμε σε κλειστή μορφή, τότε ας υποθέσουμε ότι λαμβάνουμε μία σχέση της μορφής:

$$L_{marginal,i}(\boldsymbol{\xi}, \theta, \boldsymbol{\beta}) = G(\boldsymbol{\xi}, \theta, \boldsymbol{\beta}) \quad (3.50)$$

Λογαριθμίζοντας την παραπάνω σχέση και αθροίζοντας για όλες τις $i = 1, \dots, n$ ομάδες ασθενών λαμβάνουμε την περιθώρια συνάρτηση πιθανοφάνειας για όλο τον πληθυσμό:

$$l_{marginal}(\boldsymbol{\xi}, \theta, \boldsymbol{\beta}) = \sum_{i}^{n} \log\{G(\boldsymbol{\xi}, \theta, \boldsymbol{\beta})\} \quad (3.51)$$

Με βάση τη μέθοδο μέγιστης πιθανοφάνειας μπορούμε να βρούμε τις εκτιμήσεις για τις άγνωστες παραμέτρους της κατανομής της $h_0(t)$, για την παράμετρο θ, που αντιστοιχεί στην κατανομή που έχουμε θεωρήσει για τον παράγοντα ευπάθειας και για το διάνυσμα $\boldsymbol{\beta}$ των άγνωστων συντελεστών του μοντέλου (Janssen & Duchateau, 2008).

Ο ασυμπτωτικός πίνακας διασποράς-συνδιασποράς μπορεί να υπολογιστεί από την λογαριθμοποιημένη πιθανοφάνεια. Ας συμβολίσουμε ως $\boldsymbol{H}(\boldsymbol{\xi}, \theta, \boldsymbol{\beta})$ τον Hessian matrix της μίξης των μερικών δεύτερων παραγώγων της περιθώριας λογαριθμοποιημένης συνάρτησης πιθανοφάνειας $l_{marginal}(\boldsymbol{\xi}, \theta, \boldsymbol{\beta})$.

Για λόγους απλότητας και καλύτερης έκφρασης των εξισώσεων, που ακολουθούν θα θεωρήσουμε $\boldsymbol{\zeta} = (\zeta_1, \zeta_2, \dots, \zeta_q)$ το διάνυσμα των παραμέτρων $(\boldsymbol{\xi}, \theta, \boldsymbol{\beta})$.

Το $H_{(i,j)} -$ στοιχείο του $(q \times q) -$ Hessian πίνακα δίνεται από την έκφραση:

$$H_{(i,j)} = \frac{\partial^2}{\partial \zeta_i \partial \zeta_j} l_{marginal}(\boldsymbol{\zeta}) \quad (3.52)$$

Ο πίνακας πληροφορίας κατά Fisher δίνεται από τη σχέση:

$$\boldsymbol{I}(\boldsymbol{\zeta}) = -\boldsymbol{E}(\boldsymbol{H}(\boldsymbol{\zeta})) \quad (3.53)$$

όπου $\boldsymbol{E}(\boldsymbol{H}(\boldsymbol{\zeta}))$ είναι η αναμενόμενη τιμή του Hessian matrix.

Ο παρατηρούμενος πίνακας πληροφορίας κατά Fisher δίνεται από τη σχέση:

$$\boldsymbol{I}(\boldsymbol{\zeta}) = -\boldsymbol{H}(\boldsymbol{\zeta}) \quad (3.54)$$

Ο ασυμπτωτικός πίνακας διασποράς – συνδιασποράς του διανύσματος των εκτιμώμενων παραμέτρων $\hat{\boldsymbol{\zeta}}$ είναι ο αντίστροφος του πίνακα πληροφορίας κατά Fisher και αντίστοιχα ο παρατηρούμενος πίνακας διασποράς – συνδιασποράς είναι ο αντίστροφος του παρατηρούμενου πίνακα πληροφορίας κατά Fisher.

Εκτίμηση παραμέτρων για το Γάμμα από κοινού μοντέλο ευπάθειας χρησιμοποιώντας την περιθώρια συνάρτηση πιθανοφάνειας

Η συνάρτηση πυκνότητας πιθανότητας για την Γάμμα κατανομή όπως έχει ήδη οριστεί στην ενότητα 2.1.3 με $\boldsymbol{E}(Z) = 1$ και $\boldsymbol{Var}(\mathrm{Z}) = \theta$ δίνεται από τη σχέση

$$f_Z(z) = \frac{z^{\frac{1}{\theta}-1}\exp(-z/\theta)}{\theta^{\frac{1}{\theta}}\Gamma(1/\theta)} \tag{3.55}$$

όπου $\Gamma(\cdot)$ η συνάρτηση Γάμμα. Μεγαλύτερες τιμές της παραμέτρου θ υποδηλώνουν μεγαλύτερο βαθμό μεταβλητότητας μεταξύ των διαφορετικών ομάδων ασθενών και μεγαλύτερο βαθμό ανομοιογένειας στις επιμέρους ομάδες ασθενών. Η πιο συνήθης επιλογή για την κατηγορία αναφοράς της συνάρτησης κινδύνου γα το εν λόγω μοντέλο είναι η Weibull κατανομή. Ωστόσο, υπάρχουν διάφορες επιλογές κατανομών, όπως η Εκθετική, η Λογαριμοκανονική, η Λογαριθμολογιστική κ.ο.κ.

Παρακάτω θα δείξουμε την μέθοδο εκτίμησης παραμέτρων του μοντέλου με την περιθώρια συνάρτηση πιθανοφάνειας χρησιμοποιώντας δύο επιλογές για την κατανομή της κατηγορίας αναφοράς της συνάρτησης κινδύνου, την Εκθετική και την Weibull κατανομή.

Με βάση τη σχέση (3.48), η περιθώρια συνάρτηση πιθανοφάνειας χρησιμοποιώντας Γάμμα κατανομή για τον παράγοντα ευπάθεια μπορεί να γραφεί στην μορφή:

$$L_{marginal,i}(\boldsymbol{\xi}, \theta, \boldsymbol{\beta}) = \int_0^{\infty} \prod_{j=1}^{n_i} (h_0(t_{ij}) s \exp(\boldsymbol{\beta}' \boldsymbol{x}_{ij}))^{\delta_{ij}} \exp(-H_0(t_{ij}) s \exp(\boldsymbol{\beta}' \boldsymbol{x}_{ij})) \times \frac{s^{\frac{1}{\theta}-1}\exp(-s/\theta)}{\theta^{\frac{1}{\theta}}\Gamma(1/\theta)} ds \tag{3.56}$$

όπου $\xi = \lambda$ για επιλογή της Εκθετικής κατανομής για την κατηγορία αναφοράς της συνάρτησης κινδύνου και $\boldsymbol{\xi} = (\lambda, \nu)$ για την επιλογή της Weibull κατανομής για την κατηγορία αναφοράς της συνάρτησης κινδύνου.

Αναδιατάσσοντας κατάλληλα τους όρους της σχέσης (3.56) λαμβάνουμε την εξής σχέση:

$$L_{marginal,i}(\boldsymbol{\xi}, \theta, \boldsymbol{\beta}) = \\ = \prod_{j=1}^{n_i} (h_0(t_{ij}) \exp(\boldsymbol{\beta}' \boldsymbol{x}_{ij}))^{\delta_{ij}} \int_0^{\infty} \frac{s^{\frac{1}{\theta}+d_i-1} \exp\left(-\frac{s}{\theta}\right) \exp(-\sum_{j=1}^{n_i} H_0(t_{ij}) s \exp(\boldsymbol{\beta}' \boldsymbol{x}_{ij}))}{\theta^{\frac{1}{\theta}}\Gamma(1/\theta)} ds \tag{3.57}$$

όπου $d_i = \sum_{j=1}^{n_i} \delta_{ij}$.

Έστω, $u = \frac{1}{\theta} + \sum_{j=1}^{n_i} H_0(t_{ij}) \exp(\boldsymbol{\beta}' \boldsymbol{x}_{ij})$ τότε η σχέση (3.57) μπορεί να γραφεί ως εξής:

$$L_{marginal,i}(\boldsymbol{\xi}, \theta, \boldsymbol{\beta}) = \frac{\Gamma(d_i + \frac{1}{\theta}) \prod_{j=1}^{n_i} (h_0(t_{ij}) \exp(\boldsymbol{\beta}' \boldsymbol{x}_{ij}))^{\delta_{ij}}}{u^{\frac{1}{\theta}+d_i}\, \theta^{\frac{1}{\theta}}\Gamma(1/\theta)} \int_0^{\infty} \frac{(us)^{\frac{1}{\theta}+d_i-1}\exp(-us)}{\Gamma(d_i + \frac{1}{\theta})} d(us) \tag{3.58}$$

Αφού το ολοκληρωμα ισούται με 1 σαν ολοκληρωμα συνάρτησης πυκνότητας πιθανότητας της Γάμμα κατανομής λαμβάνουμε ότι:

$$L_{marginal,i}(\boldsymbol{\xi},\theta,\boldsymbol{\beta}) = \frac{\Gamma(d_i+\frac{1}{\theta})\prod_{j=1}^{n_i}(h_0(t_{ij})\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij}))^{\delta_{ij}}}{\left(\frac{1}{\theta}+\left(\sum_{j=1}^{n_i}H_0(t_{ij})\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\right)\right)^{\frac{1}{\theta}+d_i}\theta^{\frac{1}{\theta}}\Gamma(1/\theta)} \quad (3.59)$$

Λογαριθμίζοντας την σχέση (3.59) και αθροίζοντας για τις n ομάδες των ασθενών έχουμε ότι:

$$l_{marginal}(\boldsymbol{\xi},\theta,\boldsymbol{\beta}) = \\ = \sum_{i=1}^{n}\left[d_i log\theta - log\Gamma\left(\frac{1}{\theta}\right) + log\Gamma\left(\frac{1}{\theta}+d_i\right) - \left(\frac{1}{\theta}+d_i\right)log\left(1+\theta\sum_{j=1}^{n_i}H_{ij,c}(t_{ij})\right) + \sum_{j=1}^{n_i}\delta_{ij}(\boldsymbol{\beta}'\boldsymbol{x}_{ij} + log h_0(t_{ij}))\right], \quad (3.60)$$

όπου $H_{ij,c}(t_{ij}) = H_0(t_{ij})\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})$.

Μεγιστοποιώντας την συνάρτηση (3.60) λαμβάνουμε τις εκτιμήσεις των παραμέτρων $(\boldsymbol{\xi},\theta,\boldsymbol{\beta})$.

Σε αυτό το σημείο θα θεωρήσουμε γνωστές κατανομές για την συνάρτηση αναφοράς της συνάρτησης κινδύνου. Η περιθώρια συνάρτηση πιθανοφάνειας λαμβάνει παραμετρική μορφή και σε αυτήν την περίπτωση μπορούν να χρησιμοποιηθούν οι κλασσικές μέθοδος μέγιστης πιθανοφάνειας για την εκτίμηση των υπό μελέτη παραμέτρων.

Για την επιλογή της Εκθετικής κατανομής, η συνάρτηση κινδύνου, καθώς και η αθροιστική συνάρτηση κινδύνου δίνονται από τις σχέσεις που έχουμε παραθέσει στην ενότητα 2.1.3

$$h(t) = \lambda \quad (3.61)$$

και

$$H(t) = \lambda t, \quad (3.62)$$

αντίστοιχα.

Συνεπώς, η περιθώρια συνάρτηση πιθανοφάνειας για το από κοινού Γάμμα μοντέλο ευπάθειας δίνεται από τη σχέση:

$$l_{marginal}(\lambda,\theta,\boldsymbol{\beta}) = \\ = \sum_{i=1}^{n}\left[d_i log\theta - log\Gamma\left(\frac{1}{\theta}\right) + log\Gamma\left(\frac{1}{\theta}+d_i\right) - \left(\frac{1}{\theta}+d_i\right)log\left(1+\theta\sum_{j=1}^{n_i}\lambda t_{ij}\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})\right) + \sum_{j=1}^{n_i}\delta_{ij}(\boldsymbol{\beta}'\boldsymbol{x}_{ij} + \log(\lambda))\right]. \quad (3.63)$$

Για την επιλογή της Weibull κατανομής, η συνάρτηση κινδύνου, καθώς και η αθροιστική συνάρτηση κινδύνου δίνονται από τις σχέσεις που έχουμε παραθέσει στην ενότητα 2.1.3

$$h(t) = \lambda\nu t^{\nu-1} \quad (3.64)$$

και

$$H(t) = \lambda t^{\nu}, \quad (3.65)$$

αντίστοιχα.

Συνεπώς, η περιθώρια συνάρτηση πιθανοφάνειας για το από κοινού Γάμμα μοντέλο ευπάθειας δίνεται από τη σχέση:

$$l_{marginal}(\lambda, \nu, \theta, \boldsymbol{\beta}) = \\ = \sum_{i=1}^{n} \left[d_i log\theta - log\Gamma\left(\frac{1}{\theta}\right) + log\Gamma\left(\frac{1}{\theta} + d_i\right) - \left(\frac{1}{\theta} + d_i\right) log\left(1 + \theta \sum_{j=1}^{n_i} \lambda {t_{ij}}^{\nu} \exp(\boldsymbol{\beta}' \boldsymbol{x}_{ij})\right) + \sum_{j=1}^{n_i} \delta_{ij} (\boldsymbol{\beta}' \boldsymbol{x}_{ij} + \log(\lambda \nu t_{ij}^{\nu-1})) \right]. \quad (3.66)$$

Οι πίνακες πληροφορίας κατά Fisher μπορούν να υπολογιστούν για κάθε μία από τις παραπάνω περιπτώσεις με βάση τη θεωρία, που έχουμε αναφέρει στην παρούσα ενότητα και συγκεκριμένα με βάση τις σχέσεις (3.52), (3.53) και (3.54).

ΚΕΦΑΛΑΙΟ 4

4.1 Τα δεδομένα για το λέμφωμα

Το λέμφωμα τύπου Non-Hodgkin (Non-Hodgkin Lymphoma) ή NHL ή λέμφωμα είναι ένας τύπος καρκίνου, ο οποίος εμφανίζεται στον ανθρώπινο οργανισμό προσβάλλοντας κάποια κύτταρα του οργανισμού, που αποτελούν μέρος του ανοσοποιητικού συστήματος, τα λεγόμενα λεμφοκύτταρα (Lymphocytes).

Προκειμένου να αναλυθούν οι παράγοντες κινδύνου που επιφέρουν αυτήν την ασθένεια και να αναπτυχθούν κατάλληλα στατιστικά μοντέλα για τον προσδιορισμό των παραγόντων που επηρεάζουν τους ασθενείς με λέμφωμα τύπου Non-Hodgkin (NHL) συλλέχθηκε ένα δείγμα 3273 ασθενών. Για την συλλογή και την ανάλυση του δείγματος συνεργάστηκαν 19 Ιδρύματα και ομάδες συνεργατών από τις Ηνωμένες Πολιτείες της Αμερικής, την Ευρώπη και τον Καναδά εφαρμόζοντας διαφορετικούς συνδυασμούς θεραπειών και χημειοθεραπειών από το 1982 μέχρι το 1985 (Shipp, et al., 1993).

Για την ανάλυση μας εμείς θα χρησιμοποιήσουμε ένα υπο-δείγμα από 1385 ασθενείς. Το δείγμα μας αποτελείται από 7 μεταβλητές, οι οποίες αναλυτικότερα είναι οι εξής:

1. ***Survival Time*** (*years*). Είναι ο χρόνος επιβίωσης (σε χρόνια) στην περίπτωση, που έχει παρατηρηθεί ή ο λογοκριμένος χρόνος αν αναφέρεται σε δεξιά λογοκριμένη παρατήρηση.
2. ***Status***. Αποτελεί τον δείκτη για να διακρίνουμε αν μία παρατήρηση είναι λογοκριμένη ή όχι και λαμβάνει δύο τιμές, $1 =$ dead (αναφέρεται στον παρατηρούμενο χρόνο επιβίωσης), $0 =$ alive (αναφέρεται στον παρατηρούμενο λογοκριμένο χρόνο).
3. ***Age*** (years). Η ηλικία του ασθενούς σε χρόνια.
4. ***LDH***. Είναι μία μεταβλητή, που προσδιορίζει τα επίπεδα της Γαλακτικής αφυδρογονάσης (*LDH*) στο αίμα. Η LDH είναι ένα ένζυμο, που βρίσκεται σχεδόν σε όλα τα κύτταρα του σώματος, αλλά μόνο μία μικρή ποσότητά της ανιχνεύεται στο αίμα. Η LDH απελευθερώνεται στην κυκλοφορία του αίματος και τα επίπεδα της αυξάνονται όταν τα κύτταρα υφίστανται βλάβη ή καταστρέφονται. Γενικά, χρησιμοποιείται σαν γενικός δείκτης κυτταρικής βλάβης. Συνεπώς, μικρότερα επίπεδα LDH δηλώνουν μικρότερη κυτταρική βλάβη και χαρακτηρίζονται στο δείγμα από τον δείκτη $1 = below\ normal$, ενώ μεγαλύτερα επίπεδα δηλώνουν σοβαρότερη βλάβη των κυττάρων και χαρακτηρίζονται στο δείγμα από τον δείκτη $2 = above\ normal$.
5. ***Performance Status***. Αντιπροσωπεύει την κατάσταση ικανότητας του ασθενούς και τα συνοδά προβλήματα υγείας. Ο δείκτης $1 =$ Ambulatory αναφέρεται σε ασθενή, όπου είναι συμπτωματικός, περιπατητικός, ικανός για αυτοεξυπηρέτηση, ενώ ο δείκτης $2 =$ Non-Ambulatory αναφέρεται στον ασθενή, όπου είναι συμπτωματικός, μη περιπατητικός και ενδεχομένως κλινήρης και ανίκανος να αυτοεξυπηρετηθεί.
6. ***Extra nodal sites***. Αντιπροσωπεύει το σύνολο των καταγεγραμμένων σημείων που διαγνώστηκε εξωλεμφαδενική λεμφοματώδης δραστηριότητα. Αυτά τα σημεία του ανθρώπινου οργανισμού μπορεί να είναι ο μυελός των οστών, το κεντρικό νευρικό σύστημα, τα πνευμόνια, το συκώτι κ.α. Στο δείγμα γίνεται ο εξής διαχωρισμός, κανένα σημείο λαμβάνει την τιμή 0, 1 σημείο λαμβάνει την τιμή 1 και 2 ή περισσότερα σημεία λαμβάνει την τιμή 2.
7. ***Stage***. Υπάρχουν τέσσερα στάδια μη-Hodgkin λεμφώματος, σύμφωνα με την Ταξινόμηση Ann Arbor, τα οποία περιγράφουν πού έχει εξαπλωθεί ο καρκίνος στο σώμα. Το στάδιο I αποτελεί το αρχικό στάδιο της νόσου, ενώ το στάδιο IV το πιο προχωρημένο. Τα στάδια εκφράζονται στο δείγμα με την εξής κωδικοποίηση 1=I, 2=II, 3=III, 4=IV. Τα διαφορετικά στάδια μπορούν να περιγραφούν συνοπτικά ως εκής:
   - **Στάδιο I**: Περιλαμβάνει μία μόνο περιοχή, συχνά ένα μεμονωμένο λεμφαδένα και την περιοχή που τον περιβάλλει. Κατά κανόνα, δεν υπάρχουν συμπτώματα.

- **Στάδιο II:** Περιλαμβάνει περισσότερες από μία περιοχές με προσβεβλημένους λεμφαδένες στη μια πλευρά του διαφράγματος ή μια περιοχή με προσβεβλημένο λεμφαδένα συν μια γειτονική περιοχή ή ένα όργανο.
- **Στάδιο III**: Περιλαμβάνει περιοχές με προσβεβλημένους λεμφαδένες και στις δύο πλευρές του διαφράγματος και ένα όργανο ή μια περιοχή κοντά στους λεμφαδένες, στον σπλήνα ή άλλο όργανο ή περιοχή.
- **Στάδιο IV**: Περιλαμβάνει ένα ή περισσότερα προσβεβλημένα όργανα και τον μυελό των οστών ή το δέρμα.

4.2 Μοντέλα ευπάθειας με χρήση του Στατιστικού πακέτου R

4.2.1 Περιγραφική Στατιστική

Αρχικά, δεδομένου ότι έχουμε τα δεδομένα σε μορφή CSV (*Comma delimited*) θα χρησιμοποιήσουμε την παρακάτω εντολή για να διαβάσουμε τα παρατηρούμενα δεδομένα στη μορφή ενός πλαίσιου δεδομένων με τη βοήθεια της R.

```
lymphoma_data <- read.csv( "lymphoma_data.csv", header = TRUE)
censored<-subset( lymphoma_data, Status == 0)
uncensored<-subset( lymphoma_data, Status == 1)
uncensored<-cbind( id =c( rep(c( 1:19 ), each = 33)), uncensored)*
censored<-cbind( id = c(rep( c( 1:18 ), each = 40), rep( 19,38 )), censored)*
lymphoma_data < -rbind( censored, uncensored)
```

Η ανάλυση των δεδομένων θα επικεντρωθεί κυρίως στο από κοινού μοντέλο ευπάθειας. Γι ‘αυτό το λόγο σε αυτό το σημείο προσθέτουμε την μεταβλητή id που αποτελεί ένα δείκτη απο το 1 έως το 19 και χαρακτηρίζει τα 19 clusters από τα οποία αποτελούνται τα δεδομένα. Έχουμε εισάγει τα δεδομένα σε ένα αντικείμενο, το οποίο καλούμε lymphoma_data.

Τα δεδομένα μας λοιπόν αν δούμε τις πρώτες 6 γραμμές και τις τελευταίες 6 γραμμές χρησιμοποιώντας τις συναρτήσεις head() και tail() της R θα είναι όπως φαίνεται παρακάτω:

```
> head(lymphoma_data)
   id Survival_Time Status Age LDH Performance_Status ENS Stage
1   1      5.834360      0  56   2                  1   0     2
3   1      5.971253      0  26   2                  2   2     4
7   1      4.553046      0  38   1                  1   1     2
8   1      5.119781      0  44   2                  1   0     2
10  1      6.006845      0  32   1                  1   1     2
12  1      6.056126      0  39   1                  1   0     3

> tail(lymphoma_data)
     id Survival_Time Status Age LDH Performance_Status ENS Stage
1375 19     1.4948665      1  63   1                  1   2     4
1376 19     1.7878166      1  74   2                  2   2     4
1380 19     1.2210815      1  67   2                  2   1     4
1381 19     1.8316222      1  61   2                  1   0     3
1383 19     1.2265572      1  61   2                  1   1     4
1385 19     0.1122519      1  66   2                  2   2     4
```

Χρησιμοποιώντας την έτοιμη συνάρτηση summary() μπορούμε να έχουμε μία πρώτη εικόνα ορισμένων βασικών περιγραφικών δεικτών για τις μεταβλητές Survival_Time και Age.

```
> summary(lymphoma_data$Survival_Time)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.002738 0.993800 2.971000 2.931000 4.441000 8.808000
```

```
> summary(lymphoma_data$Age)
    Min.   1st Qu.  Median       Mean     3rd Qu.      Max.
   16.00    42.00     55.00      52.53      64.00      86.00
```

Για τις κατηγορικές μεταβλητές, θα χρησιμοποιήσουμε τις εντολές table() και prop.table() της R με σκοπό να υπολογίσουμε τις συχνότητες, καθώς και τις σχετικές συχνότητες για κάθε κατηγορία. Χρησιμοποιώντας τις παραπάνω εντολές για τις μεταβλητές id, Status, LDH, Performance_Status, ENS και Stage λαμβάνουμε τα εξής αποτελέσματα:

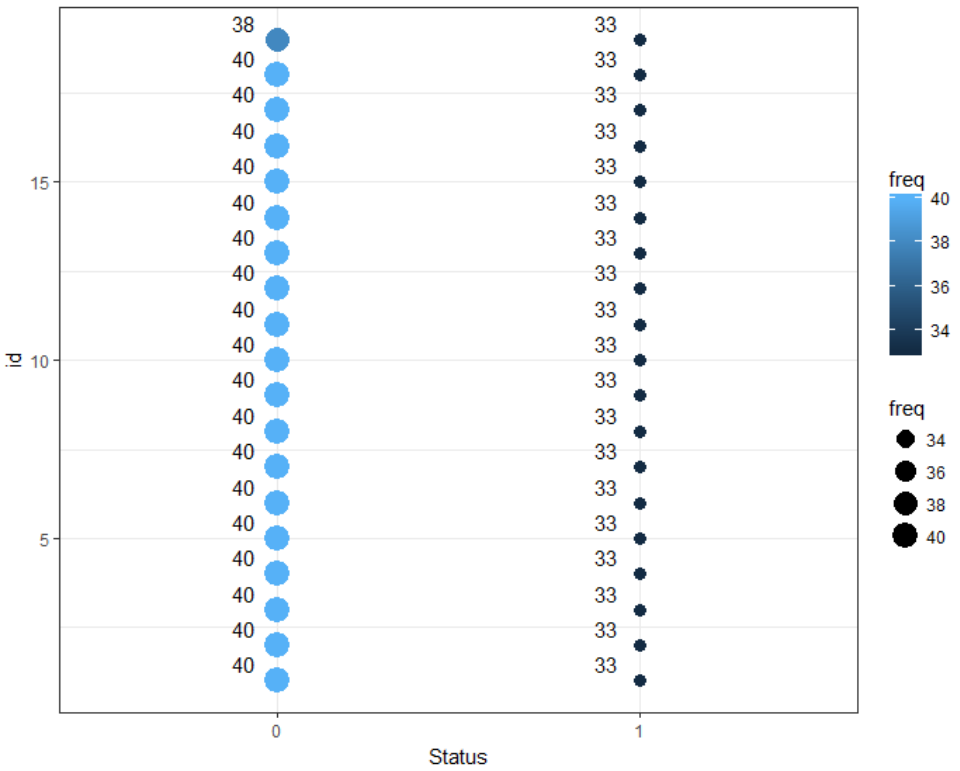
```
> table(id = lymphoma_data$id)
id
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19
73 73 73 73 73 73 73 73 73 73 73 73 73 73 73 73 73 73 71

> prop.table(table(id = lymphoma_data$id))
id
         1          2          3          4          5          6          7          8
0.05270758 0.05270758 0.05270758 0.05270758 0.05270758 0.05270758 0.05270758 0.05270758
         9
 0.05270758

        10         11         12         13         14         15         16         17
0.05270758 0.05270758 0.05270758 0.05270758 0.05270758 0.05270758 0.05270758 0.05270758

        18          19
 0.05270758   0.05126354

> table(lymphoma_data$Status)

     0   1
   758  627

> prop.table(table(lymphoma_data$Status))

       0          1
   0.5472924   0.4527076

> table(lymphoma_data$LDH)

    1    2
   623  762

> prop.table(table(lymphoma_data$LDH))

      1          2
  0.4498195    0.5501805

> table(lymphoma_data$Performance_Status)

     1    2
   1051  334

> prop.table(table(lymphoma_data$Performance_Status))
```

```
      1         2
  0.7588448 0.2411552

> table(lymphoma_data$ENS)

   0     1     2
  446   514   425

> prop.table(table(lymphoma_data$ENS))

       0         1         2
   0.3220217 0.3711191 0.3068592


> table(lymphoma_data$Stage)

    1    2    3    4
   95  420  254  616

> prop.table(table(lymphoma_data$Stage))

         1          2          3          4
    0.06859206 0.30324910 0.18339350 0.44476534
```

Τα δεδομένα παρουσιάζουν μεγάλο ποσοστό λογοκριμένων παρατηρήσεων ίσο με 54%.

Μπορούμε να κάνουμε και ένα γράφημα (*Bubble chart*), στο οποίο θα απεικονίζεται η συχνότητα των λογοκριμένων ή όχι παρατηρήσεων σε κάθε μία από τις 19 ομάδες ασθενών.

```
library(ggplot2)
data <- data.table(lymphoma_data)
frequencies <- with(data, table( id,Status))
data$Status <- as.factor(data$Status)
data1 <- count(data,c("id","Status"))
# data1 <- data[ , .(count = count(Status)), by = list( id, Status = as.factor(Status))]
ggplot(data1, aes(x = Status, y = id, label = freq)) + geom_point(aes( size = freq, color
= freq)) + geom_text( hjust = 2, vjust = -0.5, size = 4) + scale_size( range = c(3,6)) +
theme_bw()
```



Διάγραμμα 4.1: *Bubble chart, που απεικονίζει τις συχνότητες λογοκριμένων χρόνων (Status=0) και μη λογοκριμένων χρόνων (Status=1) για κάθε ομάδα ασθενών (id=1,…,19).*

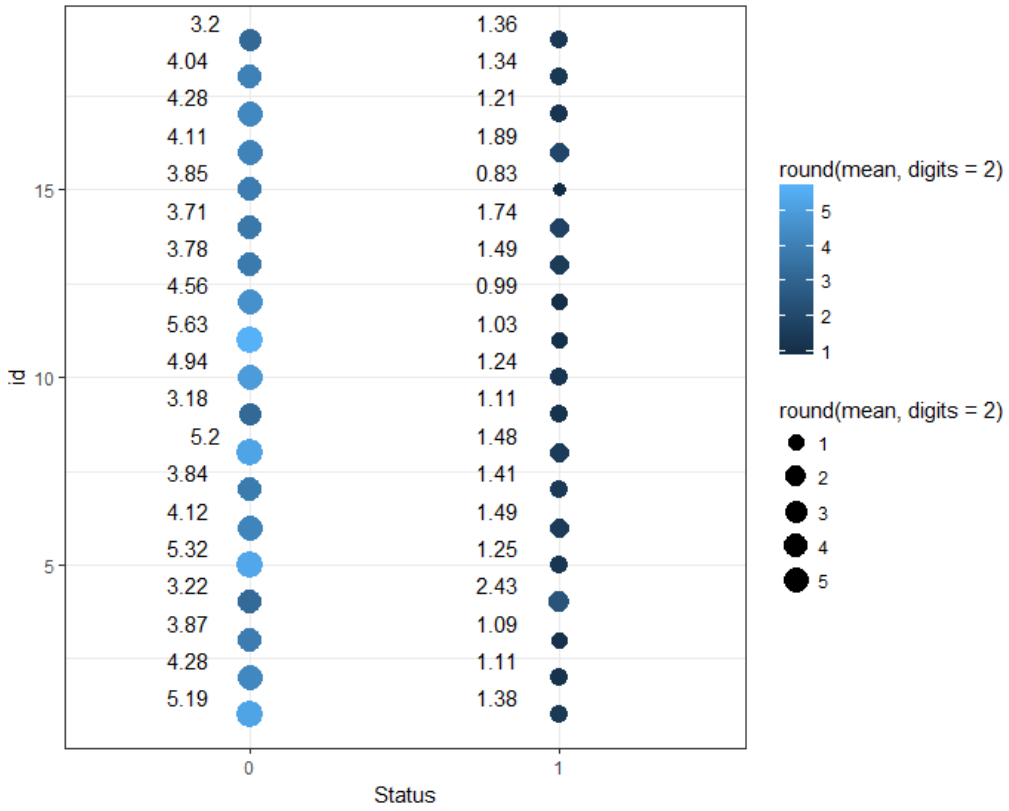
Στο Διάγραμμα 4.1 βλέπουμε τον αριθμό των παρατηρήσεων του χρόνου επιβίωσης και του λογοκριμένου χρόνου για κάθε ομάδα ασθενών από τις 19.

Επιπλέον, θα δούμε τη μέση τιμή και την τυπική απόκλιση του χρόνου επιβίωσης για κάθε ομάδα ασθενών.

```
> data[,list(mean = mean(Survival_Time),sd = sd(Survival_Time), by = id]
    id     mean       sd
 1:  1 3.467769 2.232846
 2:  2 2.844627 1.777503
 3:  3 2.614723 1.598663
 4:  4 2.859891 1.600279
 5:  5 3.479208 2.746459
 6:  6 2.930625 1.873605
 7:  7 2.741151 1.982219
 8:  8 3.516113 2.217320
 9:  9 2.241437 1.445000
10: 10 3.271619 2.031267
11: 11 3.550730 2.773682
12: 12 2.945440 2.101138
13: 13 2.741826 1.987332
14: 14 2.818036 1.735157
15: 15 2.484056 1.916106
16: 16 3.111361 2.149993
17: 17 2.890083 1.999321
18: 18 2.815748 1.956841
19: 19 2.344449 1.551488
```

Δημιουργούμε ένα Bubble chart, που απεικονίζει το μέσο παρατηρούμενο χρόνο για τους λογοκριμένους χρόνους και μη για κάθε ομάδα ασθενών.

```
##plot OF MEAN TIME PER GROUP
data<-data.table(lymphoma_data)
data[,mean:=mean(Survival_Time),by=list(id,Status)]
data1<-data[,list(id,Status,mean)]
data1<-unique(data1)
data1$Status<-as.factor(data1$Status)
ggplot(data1, aes(x=Status, y=id, label=round(mean,2)))+geom_point(aes(size=round(mean,di
gits=2), color=round(mean,digits=2)))+geom_text(hjust=2, vjust=-0.5, size=4)+scale_size(r
ange = c(3,6)) +
  theme_bw()
```

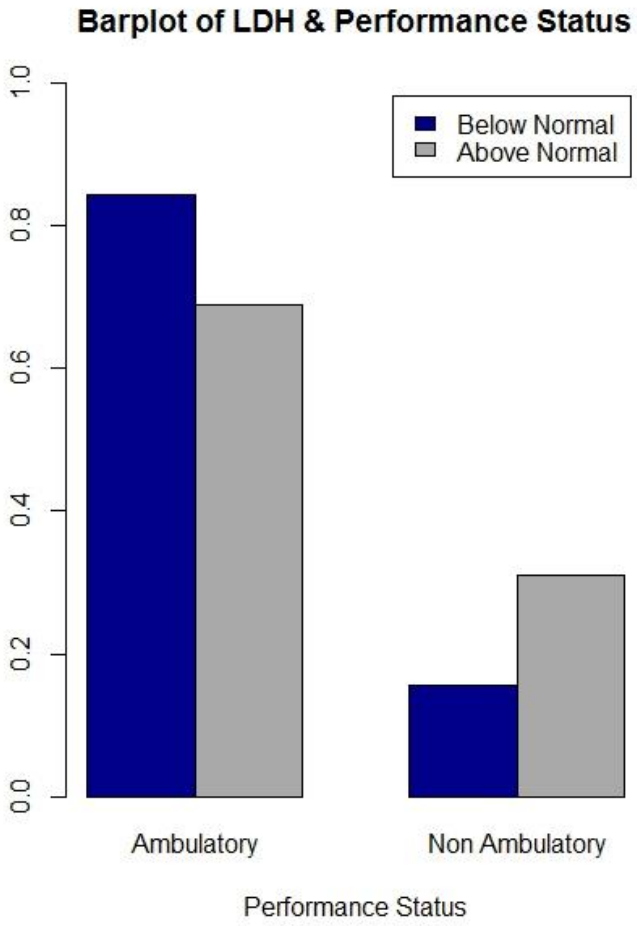


Διάγραμμα 4.2: *Bubble chart, που απεικονίζει τη μέση τιμή των παρατηρούμενων λογοκριμένων χρόνων (Status=0) και την μέση τιμή των χρόνων επιβίωσης (Status=1) για κάθε ομάδα ασθενών (id=1,…,19).*

Από το Διάγραμμα 4.2 συμπεραίνουμε ότι υπάρχει μία σχετική μεταβλητότητα του χρόνου επιβίωσης μεταξύ των ομάδων. Δεν ισχύει το ίδιο όσον αφορά τους λογοκριμένους χρόνους, καθώς παρατηρείται όλες οι ομάδες να έχουν περίπου κατά μέσο όρο ίδιο λογοκριμένο χρόνο ασθενών.

Σε αυτό το σημείο θα κατασκευάσουμε πίνακες συνάφειας διπλής εισόδου σχετικών συχνοτήτων για δύο κατηγορικές μεταβλητές, δηλαδή θα παρουσιάσουμε τις από κοινού πιθανότητες για κάθε κελί που εμφανίζεται στους συγκεκριμένους πίνακες. Επιπλέον, θα κατασκευάσουμε και τα αντίστοιχα ομαδοποιημένα ραβδογράμματα για μία πιο άμεση γραφική αναπαράσταση των κατηγορικών μεταβλητών. Για να επιτευχθούν τα παραπάνω στην R, θα χρησιμοποίησουμε κατάλληλα τις συναρτήσεις table() και barplot().

```
table1<-table(lymphoma_data$LDH,lymphoma_data$Performance_Status)

barplot(prop.table(table1,1), xlab="Performance Status", legend.text=c("Below Normal","Ab
ove Normal"),names.arg=c("Ambulatory","Non Ambulatory"), col=c("darkblue","darkgrey"), ma
in="Barplot of LDH & Performance Status",width=0.8, beside=TRUE, ylim=c(0,1))
```



Διάγραμμα 4.3: *Ομαδοποιημένο ραβδόγραμμα της κατάστασης ικανότητας του ασθενούς (Performance Status) δοθέντων των επιπέδων της Γαλακτικής αφυδρογονάσης (LDH) στο αίμα του ασθενούς.*

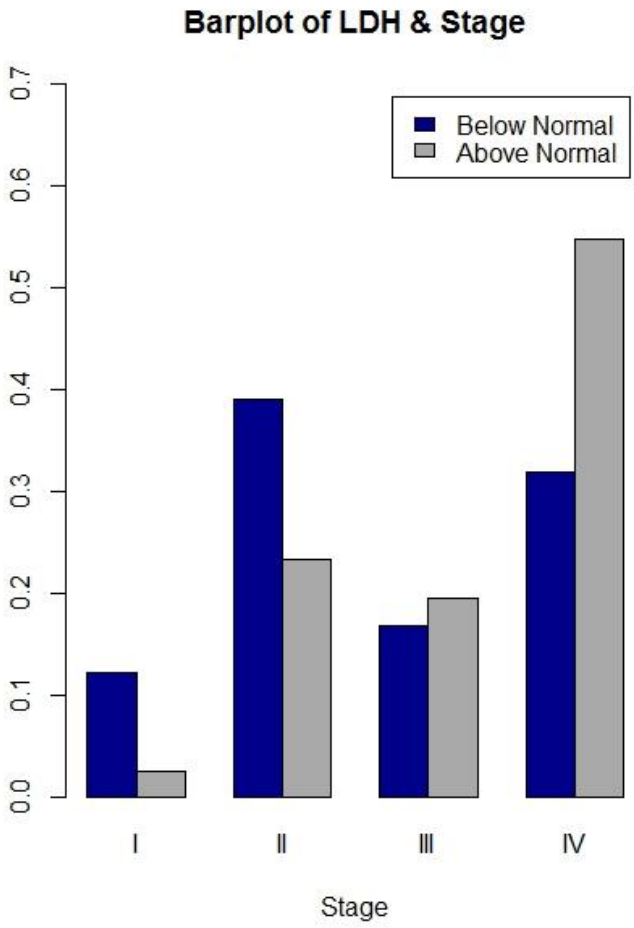
Το Διάγραμμα 4.3 αντιστοιχεί στα αποτελέσματα της εντολής prop.table(table1,1), δηλαδή στον πίνακα σχετικών συχνοτήτων της κατάστασης ικανότητας του ασθενούς δεσμεύοντας ως προς τα επίπεδα LDH.

```
> prop.table(table1,1)
   Performance_Status
LDH         1         2
  1 0.8443018 0.1556982
  2 0.6889764 0.3110236
```

Από τον παραπάνω πίνακα, ο οποίος περιγράφεται γραφικά με το Διάγραμμα 4.3, παρατηρούμε ότι από τους ασθενείς στο δείγμα μας που έχουν μικρότερη κυτταρική βλάβη το 84% είναι ικανοί για αυτοεξυπηρέτηση, ενώ μόλις 16% δεν μπορούν να αυτοεξυπηρετηθούν. Από τους ασθενείς που παρουσιάζουν σοβαρότερη βλάβη των κυττάρων, το 70%

μπορούν να αυτοεξυπηρετηθούν, ενώ το 30% είναι μη περιπατητικοί. Σίγουρα συμπεραίνουμε ότι όσο μεγαλύτερα είναι τα επίπεδα LDH υπάρχει μεγαλύτερο πρόβλημα στην κατάσταση ικανότητας του ασθενούς. Ωστόσο, παραμένει πολύ μεγάλο το ποσοστό αυτών που έχουν σχετικά καλή κατάσταση και μπορούν να αυτοεξυπηρετηθούν.

```
table2<-table(LDH, Stage)

barplot(prop.table(table2,1), xlab="Stage", legend.text=c("Below Normal","Above Normal"),
names=c("I", "II", "III", "IV"), col=c("darkblue","darkgrey"), main="Barplot of LDH & Sta
ge",width=0.30, beside=TRUE, ylim=c(0,0.7))
```



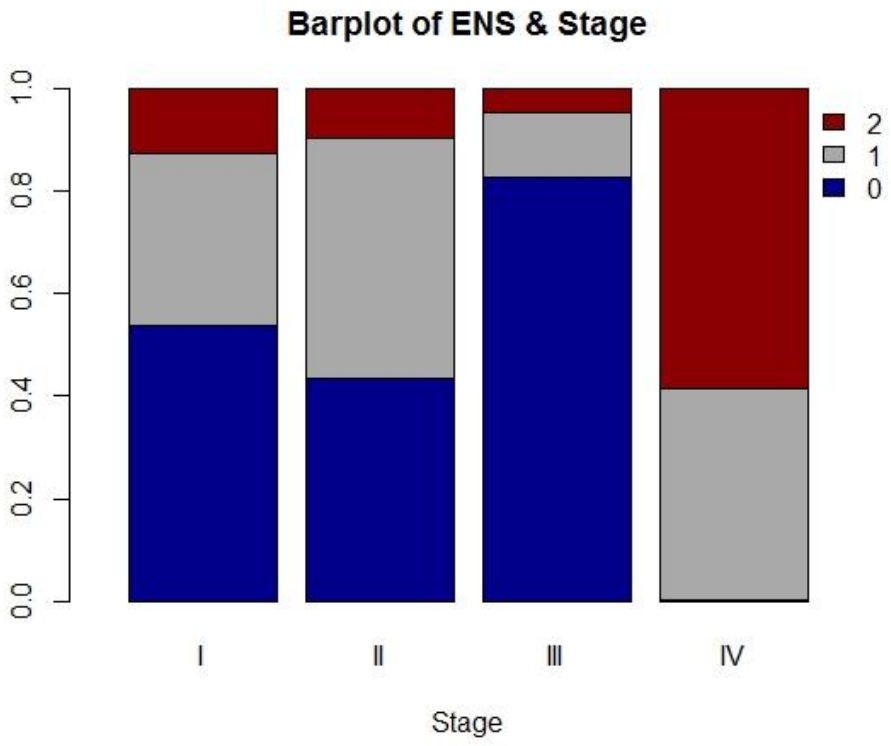
Διάγραμμα 4.4: *Ομαδοποιημένο ραβδόγραμμα του σταδίου μη-Hodgkin λεμφώματος που βρίσκεται ο ασθενής (Stage) δοθέντων των επιπέδων της Γαλακτικής αφυδρογονάσης (LDH) στο αίμα του ασθενούς.*

```
> prop.table(table2,1)
   Stage
LDH          1          2          3          4
  1 0.12199037 0.39004815 0.16853933 0.31942215
  2 0.02493438 0.23228346 0.19553806 0.54724409
```

Από τον παραπάνω πίνακα, ο οποίος περιγράφεται γραφικά με το Διάγραμμα 4.4, παρατηρούμε ότι από τους ασθενείς στο δείγμα μας που έχουν μικρότερη κυτταρική βλάβη το 12% βρίσκονται στο στάδιο I, δηλαδή στο αρχικό στάδιο της

νόσου, το 39% στο δεύτερο στάδιο, το 17% στο τρίτο και το 31% στο τέταρτο στάδιο. Από τους ασθενείς που παρουσιάζουν σοβαρότερη βλάβη των κυττάρων, το 2% βρίσκονται στο στάδιο 1, το 23% βρίσκονται στο στάδιο 2, το 20% στο στάδιο 3, ενώ το 55% στο τελευταίο και σοβαρότερο στάδιο, στο στάδιο 4. Είναι αναμενόμενο ότι όσο μεγαλύτερα είναι τα επίπεδα LDH υπάρχει μεγαλύτερο πρόβλημα στην κατάσταση ικανότητας του ασθενούς και ο ασθενής τείνει να βρίσκεται στο στάδιο 4. Ωστόσο, παραμένει σχετικά μεγάλο και το ποσοστό αυτών που δεν έχουν πολύ σοβαρό πρόβλημα κυτταρικής βλάβης σύμφωνα με τα επίπεδα LDH, αλλά παρ' όλα αυτά βρίσκονται στο τελευταίο στάδιο καρκίνου (32%). Σίγουρα το ποσοστό αυτών που βρίσκονται στο αρχικό στάδιο της νόσου είναι πολύ μικρό για αυτούς που έχουν μικρότερη κυτταρική βλάβη (2%).

```
table3<-table(ENS, Stage)

barplot(prop.table(table3,2), xlab="Stage", legend=levels(ENS),names=c("I", "II", "III",
"IV"), col=c("darkblue","darkgrey","darkred"), main="Barplot of ENS & Stage",width=0.25,x
lim=c(0,1.25), args.legend=list(
  x=1.35,
  y=max(colSums(prop.table(table3,2))),
  bty = "n"
))
```



Διάγραμμα 4.5: *Στοιβαγμένο ραβδόγραμμα των καταγεγραμμένων σημείων που διαγνώστηκε εξωλεμφαδενική λεμφοματώδης δραστηριότητα (Extra Nodal Nodes) δοθέντος του σταδίου μη-Hodgkin λεμφώματος που βρίσκεται ο ασθενής (Stage).*

```
> prop.table(table3,2)
   Stage
ENS           1           2           3           4
  0 0.536842105 0.435714286 0.826771654 0.003246753
  1 0.336842105 0.469047619 0.125984252 0.410714286
  2 0.126315789 0.095238095 0.047244094 0.586038961
```

Από το Διάγραμμα 4.5 παρατηρούμε ότι για τους ασθενείς που βρίσκονται στο στάδιο 3, το ποσοστό των ασθενών που δεν διαγνώστηκε εξωλεμφαδενική λεμφοματώδης δραστηριότητα είναι 83%. Για αυτούς που βρίσκονται στο πρώτο και στο δεύτερο στάδιο, το ποσοστό αυτών που έχουν παρουσιάσει εξωλεμφαδενική λεμφοματώδη δραστηριότητα σε 2 ή περισσότερα σημεία είναι 13% και 10%, αντίστοιχα. Τέλος, για αυτούς που βρίσκονται στο τελευταίο στάδιο, το 58.7% έχουν παρουσιάσει εξωλεμφαδενική λεμφοματώδη δραστηριότητα σε περισσότερα από 2 σημεία, το 41% παρουσιάζουν σε ένα σημείο, ενώ ένα πάρα πολύ μικρό ποσοστό 0.3% δεν παρουσιάζουν σε κανένα σημείο.

4.2.2 Προσαρμογή παραμετρικών μοντέλων AFT (μονομεταβλητή προσέγγιση)

Πριν προχωρήσουμε στην προσαρμογή και ανάλυση των κατάλληλων μοντέλων για τα δεδομένα μας, θα εφαρμόσουμε στις μεταβλητές μία άλλη κωδικοποίηση για καλύτερη ερμηνεία των αποτελεσμάτων. Συγκεκριμένα,

LDH: Θα κωδικοποιήσουμε τη μεταβλητή με τις τιμές 0 και 1, όπου με $0 = below\ normal$ θα χαρακτηρίζονται οι ασθενείς με μικρότερα επίπεδα LDH τα οποία δηλώνουν μικρότερη κυτταρική βλάβη, ενώ μεγαλύτερα επίπεδα τα οποία δηλώνουν σοβαρότερη βλάβη των κυττάρων θα χαρακτηρίζονται στο δείγμα από τον δείκτη $1 = above\ normal$.

Performance Status: Θα κωδικοποιήσουμε τη μεταβλητή με τις τιμές 0 και 1. Ο δείκτης $0 =$ Ambulatory αναφέρεται σε ασθενή, όπου είναι συμπτωματικός, περιπατητικός, ικανός για αυτοεξυπηρέτηση, ενώ ο δείκτης $1 =$ Non-Ambulatory αναφέρεται στον ασθενή, όπου είναι συμπτωματικός, μη περιπατητικός και ενδεχομένως κλινήρης και ανίκανος να αυτοεξυπηρετηθεί.

Extra nodal sites: Θα κωδικοποιήσουμε τη μεταβλητή με τις τιμές 0 και 1, όπου 0 αναφέρεται στους ασθενείς που έχουν καταγεγραμμένα σημεία διάγνωσης εξωλεμφαδενικής λεμφοματώδους δραστηριότητας με μικρότερο ή ίσο με 1 και αντίστοιχα με την τιμή 1, αν έχουν μεγαλύτερο ή ίσο αριθμό με 1.

Stage: Για τα στάδια I και II χρησιμοποιούμε την τιμή 0, ενώ για τα στάδια III και IV με την τιμή 1.

Age: κωδικοποιούμε κατάλληλα την επεξηγηματική μεταβλητή Age, ώστε να δημιουργήσουμε δύο ηλικιακές ομάδες, τους ασθενείς που είναι 60 ετών ή μικρότεροι από 60 ετών και αυτούς που είναι άνω των 60 ετών με δείκτες 0 και 1 αντίστοιχα.

Στην R θα χρησιμοποιήσουμε τις παρακάτω εντολές για να κάνουμε τις παραπάνω κωδικοποιήσεις.

```
#MAKE FACTORS
lymphoma_data$ENS<-as.factor(lymphoma_data$ENS)
lymphoma_data$Stage<-as.factor(lymphoma_data$Stage)
#RESCALE ENS AND STAGE
lymphoma_data$ENS1<-lymphoma_data$ENS
levels(lymphoma_data$ENS1)<-c("0","0","1")

lymphoma_data$Stage1<-lymphoma_data$Stage
levels(lymphoma_data$Stage1)<-c("0","0","1","1")

lymphoma_data_new<-lymphoma_data

lymphoma_data_new$Age1 <- ifelse(lymphoma_data_new$Age > 60, 1, 0)
lymphoma_data_new$Age1 <-as.factor(lymphoma_data_new$Age1)
```

Τα δεδομένα μας τώρα περιλαμβάνουν τα παρακάτω αντικείμενα με την εξής δομή:

```
> str(lymphoma_data_new)
'data.frame':  1385 obs. of  11 variables:
 $ id                : num  1 1 1 1 1 1 1 1 1 1 ...
 $ Survival_Time     : num  5.83 5.97 4.55 5.12 6.01 ...
 $ Status            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Age               : int  56 26 38 44 32 39 47 55 57 48 ...
 $ LDH               : Factor w/ 2 levels "1","1": 2 2 1 2 1 1 2 1 2 2 ...
 $ Performance_Status: Factor w/ 2 levels "1","1": 1 2 1 1 1 1 1 2 1 1 ...
 $ ENS               : Factor w/ 3 levels "0","1","2": 1 3 2 1 2 1 1 1 3 1 ...
 $ Stage             : Factor w/ 4 levels "1","2","3","4": 2 4 2 2 2 3 3 3 2 2 ...
 $ ENS1              : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 2 1 ...
 $ Stage1            : Factor w/ 2 levels "0","1": 1 2 1 1 1 2 2 2 1 1 ...
 $ Age1              : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

Να σημειώσουμε ότι θα δημιουργήσουμε ένα νέο object στην R, το οποίο καλούμε lymphoma_data_new και η ως άνω κωδικοποιήσεις θα ενσωματωθούν στις νέες μεταβλητές στο νέο πλαίσιο δεδομένων με ονόματα LDH, Performance_Status, ENS1, Stage1 και Age1. Συνεπώς, στο εξής θα χρησιμοποιούμε το πλαίσιο δεδομένων lymphoma_data_new, στο οποίο έχουμε εφαρμόσει νέα κωδικοποίηση στις επεξηγηματικές μεταβλητές.

Σε αυτό το σημείο τα μοντέλα που προσαρμόζουμε αποτελούν μοντέλα επιταχυνόμενου βαθμού κινδύνου και θεωρούμε διαφορετικές κατανομές για την μεταβλητή σφάλματος. Θα προσαρμόσουμε τέσσερα παραμετρικά μοντέλα στα δεδομένα μας χρησιμοποιώντας την Λογαριθμοκανονική, την Weibull, την Εκθετική και την Λογαριθμολογιστική κατανομή για την ευπαθή μεταβλητή. Για το σκοπό αυτό θα χρησιμοποιήσουμε τη συνάρτηση survreg() από το πακέτο survival, η οποία χρησιμοποιείται για την προσαρμογή AFT μοντέλων. Η παράμετρος formula χρησιμοποιείται για να ορίσουμε την μορφή του μοντέλου συμπεριλαμβάνοντας όλες τις κατάλληλες επεξηγηματικές μεταβλητές και η παράμετρος dist έχει διάφορες επιλογές κατανομών για την προσαρμογή του παραμετρικού μοντέλου (“Weibull”, “exponential”, “gaussian”, “logistic”, “lognormal” ή “loglogistic).

- ✓ Παραμετρικό μοντέλο χρησιμοποιώντας την Λογαριθμοκανονική κατανομή

```
> fit1 <- survreg(Surv(Survival_Time,Status) ~ Age1 + LDH + Performance_Status + ENS1 + S
tage1,data = lymphoma_data_new,
+                 dist='lognormal')
> summary(fit1)

Call:
survreg(formula = Surv(Survival_Time, Status) ~ Age1 + LDH +
    Performance_Status + ENS1 + Stage1, data = lymphoma_data_new,
    dist = "lognormal")
                     Value Std. Error     z         p
(Intercept)          3.245     0.1434 22.63 2.15e-113
Age11               -0.944     0.1223 -7.72  1.15e-14
LDH2                -0.932     0.1269 -7.35  2.00e-13
Performance_Status2 -1.042     0.1365 -7.64  2.24e-14
ENS11               -0.610     0.1341 -4.55  5.41e-06
Stage11             -0.545     0.1403 -3.89  1.02e-04
Log(scale)           0.645     0.0311 20.76  1.02e-95

Scale= 1.91
```

```
Log Normal distribution
Loglik(model)= -1552.9   Loglik(intercept only)= -1702.8
       Chisq= 299.76 on 5 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 4
n= 1385
```

- ✓ Παραμετρικό μοντέλο χρησιμοποιώντας την Weibull κατανομή

```
> fit2 <- survreg(Surv(Survival_Time,Status) ~ Age1 + LDH + Performance_Status + ENS1 + S
tage1,data = lymphoma_data_new,
+                 dist='weibull')
> summary(fit2)

Call:
survreg(formula = Surv(Survival_Time, Status) ~ Age1 + LDH +
    Performance_Status + ENS1 + Stage1, data = lymphoma_data_new,
    dist = "weibull")
                    Value Std. Error     z         p
(Intercept)          3.671     0.1544 23.78 5.19e-125
Age11               -0.927     0.1136 -8.17  3.19e-16
LDH2                -0.865     0.1236 -7.00  2.61e-12
Performance_Status2 -0.779     0.1222 -6.38  1.77e-10
ENS11               -0.550     0.1211 -4.54  5.56e-06
Stage11             -0.555     0.1398 -3.97  7.14e-05
Log(scale)           0.321     0.0347  9.26  2.13e-20

Scale= 1.38

Weibull distribution
Loglik(model)= -1590   Loglik(intercept only)= -1725.8
       Chisq= 271.51 on 5 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 5
n= 1385
```

- ✓ Παραμετρικό μοντέλο χρησιμοποιώντας την Λογαριθμολογιστική κατανομή

```
> fit3 <- survreg(Surv(Survival_Time,Status) ~ Age1 + LDH + Performance_Status + ENS1 + S
tage1,data = lymphoma_data_new,
+                 dist='loglogistic')
> summary(fit3)

Call:
survreg(formula = Surv(Survival_Time, Status) ~ Age1 + LDH +
    Performance_Status + ENS1 + Stage1, data = lymphoma_data_new,
    dist = "loglogistic")
                     Value Std. Error     z         p
(Intercept)          3.1735     0.1397 22.71 3.57e-114
Age11               -0.9493     0.1178 -8.06  7.87e-16
LDH2                -0.9122     0.1235 -7.39  1.49e-13
Performance_Status2 -1.0007     0.1330 -7.52  5.29e-14
ENS11               -0.5841     0.1289 -4.53  5.90e-06
Stage11             -0.5753     0.1370 -4.20  2.69e-05
Log(scale)           0.0751     0.0343  2.19  2.86e-02
```

```
Scale= 1.08

Log logistic distribution
Loglik(model)= -1557.7   Loglik(intercept only)= -1711
        Chisq= 306.61 on 5 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 4
n= 1385
```

- ✓ Παραμετρικό μοντέλο χρησιμοποιώντας την Εκθετική κατανομή

```
> fit13<-survreg(Surv(Survival_Time,Status) ~ Age1 + LDH + Performance_Status + ENS1 + St
age1,data = lymphoma_data_new,
+                dist='exponential')
> summary(fit13)

Call:
survreg(formula = Surv(Survival_Time, Status) ~ Age1 + LDH +
    Performance_Status + ENS1 + Stage1, data = lymphoma_data_new,
    dist = "exponential")
                     Value Std. Error     z         p
(Intercept)          3.098     0.0995 31.14 7.50e-213
Age11               -0.743     0.0809 -9.18  4.14e-20
LDH2                -0.678     0.0880 -7.70  1.34e-14
Performance_Status2 -0.607     0.0872 -6.96  3.32e-12
ENS11               -0.447     0.0876 -5.10  3.36e-07
Stage11             -0.421     0.1007 -4.18  2.89e-05

Scale fixed at 1

Exponential distribution
Loglik(model)= -1639.3   Loglik(intercept only)= -1798.1
        Chisq= 317.61 on 5 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 5
n= 1385
```

Παρατηρούμε από τα παραπάνω αποτελέσματα ότι το μοντέλο με τη Λογαριθμοκανονική κατανομή για τη βασική συνάρτηση κινδύνου λαμβάνει τη μεγαλύτερη τιμή για τη συνάρτηση Πιθανοφάνειας σε λογαριθμική κλίμακα συγκριτικά με τα τρία άλλα μοντέλα (Lognormal – Likelihood = -1552, Weibull – Likelihood = -1590, Loglogistic – Likelihood = -1557, Exponential - Likelihood = -1639). Αναλύοντας, τα αποτελέσματα για το πρώτο μοντέλο, παρατηρούμε ότι το τεστ όλων των παραμέτρων ίσο με 0 απορρίπτεται με $p - τιμή \sim 0$ και τιμή για την X^2 ελεγχοσυνάρτηση ίση με 300 με 5 βαθμούς ελευθερίας. Σε όλα τα παραπάνω παραμετρικά μοντέλα, που προσαρμόστηκαν παρατηρούμε ότι όλοι οι επί μέρους έλεγχοι για κάθε μια μεταβλητή ξεχωριστά απορρίπτουν την υπόθεση ότι οι συντελεστές ειναι 0. Σε όλα τα μοντέλα, προκύπτει ότι όλες οι μεταβλητές είναι στατιστικά σημαντικές ως συνδυασμός του εν λόγω μοντέλου σε επίπεδο σημαντικότητας 5%. Ωστόσο, οι μεταβλητές Age, LDH και Performance status έχουν μεγαλύτερη βαρύτητα και προγνωστική αξία με την ηλικία να είναι η πιο σημαντική. Οι συντελεστές των μοντέλων είναι όλοι αρνητικοί κάτι το οποίο σημαίνει ότι όλες οι μεταβλητές επιδρούν αυξητικά στην συνάρτηση κινδύνου και κατά συνέπεια αρνητικά στην επιβίωση των ασθενών.

4.2.3 Προσαρμογή από κοινού παραμετρικών μοντέλων ευπάθειας AFT

Σε αυτό το σημείο λοιπόν θα εισάγουμε τον **παράγοντα ευπάθεια** στο εν λόγω μοντέλο (fit1). Όπως αναφέρθηκε και παραπάνω. Αρχικά, θα παρουσιάσουμε το Γάμμα μοντέλο ευπάθειας και εν συνεχεία το Gaussian μοντέλο ευπάθειας.

- ✓ Γάμμα παραμετρικό μοντέλο ευπάθειας χρησιμοποιώντας την Λογαριθμοκανονική κατανομή για τη βασική συνάρτηση κινδύνου

```
> fit4 <- survreg(Surv(Survival_Time,Status) ~ Age1 + LDH + Performance_Status + ENS1 + S
tage1 + frailty(id,dist="gamma"),data=lymphoma_data_new,dist='lognormal')
> summary(fit4)

Call:
survreg(formula = Surv(Survival_Time, Status) ~ Age1 + LDH +
    Performance_Status + ENS1 + Stage1 + frailty(id, dist = "gamma"),
    data = lymphoma_data_new, dist = "lognormal")
                    Value Std. Error      z        p
(Intercept)         4.862     0.3829  12.70 6.12e-37
Age11              -3.266     0.2690 -12.14 6.42e-34
LDH2               -0.883     0.1212  -7.29 3.20e-13
Performance_Status2 -0.925     0.1339  -6.91 4.99e-12
ENS11              -0.469     0.1296  -3.62 2.98e-04
Stage11            -0.718     0.1378  -5.21 1.89e-07
Log(scale)          0.578     0.0304  19.01 1.38e-80

Scale= 1.78

Log Normal distribution
Loglik(model)= -1479.9   Loglik(intercept only)= -1702.8
        Chisq= 445.88 on 21.6 degrees of freedom, p= 0
Number of Newton-Raphson Iterations:  6 16
n= 1385
```

Από τα αποτελέσματα του παραπάνω μοντέλου συμπεραίνουμε ότι ο έλεγχος υποθέσεων ότι όλοι οι συντελεστές του μοντέλου είναι ίσοι με 0 απορρίπτεται με $p-$*τιμή* περίπου ίση με 0 και X^2 ελεγχοσυνάρτηση ίση με 445.88. Οι σημαντικότεροι συντελεστές για το μοντέλο, όπως και πριν, είναι η ηλικία Age, η LDH και η Performance Status. Η σημαντικότητα της ηλικίας ενισχύεται μέσω του μοντέλου ευπάθειας, Το εν λόγω μοντέλο με την εισαγωγή της μεταβλητής της ευπάθειας είναι καλύτερο συγκριτικά με το μοντέλο χωρίς τον παράγοντα ευπάθειας (fit1). Η τιμή της συνάρτησης Πιθανοφάνειας σε λογαριθμική κλίμακα -1479.9 είναι μεγαλύτερη σε σχέση με την τιμή -1552 (για το μοντέλο fit1).

- ✓ Gaussian παραμετρικό μοντέλο ευπάθειας χρησιμοποιώντας την Λογαριθμοκανονική κατανομή για τη βασική συνάρτηση κινδύνου

```
> fit5 <- survreg(Surv(Survival_Time,Status) ~ Age1 + LDH + Performance_Status + ENS1 + S
tage1 + frailty(id, dist = "gaussian"),data=lymphoma_data_new ,dist='lognormal')
> summary(fit5)

Call:
```

```
survreg(formula = Surv(Survival_Time, Status) ~ Age1 + LDH +
    Performance_Status + ENS1 + Stage1 + frailty(id, dist = "gaussian"),
    data = lymphoma_data_new, dist = "lognormal")
                    Value Std. Error      z        p
(Intercept)          4.109     0.3606  11.39 4.45e-30
Age11               -3.329     0.2628 -12.67 8.60e-37
LDH2                -0.883     0.1214  -7.27 3.48e-13
Performance_Status2 -0.922     0.1342  -6.87 6.33e-12
ENS11               -0.466     0.1298  -3.59 3.25e-04
Stage11             -0.725     0.1382  -5.25 1.54e-07
Log(scale)           0.579     0.0304  19.03 8.87e-81

Scale= 1.78

Log Normal distribution
Loglik(model)= -1479.3   Loglik(intercept only)= -1702.8
       Chisq= 446.92 on 21.4 degrees of freedom, p= 0
Number of Newton-Raphson Iterations:  6 16
n= 1385
```

Από τα αποτελέσματα του παραπάνω μοντέλου συμπεραίνουμε ότι ο έλεγχος υποθέσεων ότι όλοι οι συντελεστές του μοντέλου είναι ίσοι με 0 απορρίπτεται με $p-$*τιμή* περίπου ίση με 0 και X^2 ελεγχοσυνάρτηση ίση με 446.92. Οι σημαντικότεροι συντελεστές για το μοντέλο είναι η ηλικία Age, η LDH και η Performance Status με την ηλικία να υπερέχει. Το εν λόγω μοντέλο με την εισαγωγή της μεταβλητής της ευπάθειας είναι καλύτερο συγκριτικά με το μοντέλο χωρίς τον παράγοντα ευπάθειας (fit1). Η τιμή της συνάρτησης Πιθανοφάνειας σε λογαριθμική κλίμακα -1479.3 είναι μεγαλύτερη σε σχέση με την τιμή -1552 (για το μοντέλο fit1).

4.2.4 Προσαρμογή από κοινού μοντέλων ευπάθειας

Χρησιμοποιώντας τα μοντέλα ευπάθειας, που θα προσαρμόσουμε εκτιμούμε τον κίνδυνο ένας ασθενής να παρουσιάσει λέμφωμα τύπου Non-Hodgkin (Non-Hodgkin Lymphoma) σε σχέση με τις επεξηγηματικές μεταβλητές LDH (επίπεδα της Γαλακτικής αφυδρογονάσης στο αίμα), Performance status (κατάσταση ικανότητας του ασθενούς), ENS (καταγεγραμμένα σημεία που διαγνώστηκε εξωλεμφαδενική λεμφοματώδης δραστηριότητα), Stage (στάδιο μη-Hodgkin λεμφώματος) και Age (ηλικία του ασθενούς).

Στη γενική του μορφή το μοντέλο γράφεται ως εξής:

$$h_{ij}\left(t|z_i,\bar{X}_j\right) = h_0(t)z_i\exp(\beta_1 Age + \beta_2 LDH + \beta_3 Performance\ Status + \beta_4 ENS + \beta_5 Stage) \qquad ()$$

όπου z_i είναι η ευπάθεια που αντιστοιχεί σε όλες τις μονάδες του πληθυσμού που ανήκουν στην $i-$ομάδα, $h_0(t)$ είναι η βασική συνάρτηση κινδύνου και $\beta_1,\beta_2,\beta_3,\beta_4,\beta_5$ οι συντελεστές του μοντέλου.

Να σημειώσουμε εδώ ότι αρχικά θα ασχοληθούμε με την παραμετρική περίπτωση των μοντέλων ευπάθειας, δηλαδή με μοντέλα όπου κάνουμε μία αρχική παραδοχή για την κατανομή του χρόνου T ή αντίστοιχα τη συναρτησιακή μορφή της $h_0(t)$. Ορίζουμε δηλαδή μία παραμετρική κατανομή για το χρόνο T και η βασική συνάρτηση κινδύνου είναι γνωστής συναρτησιακής μορφής αλλά εξαρτάται από κάποια άγνωστη παράμετρο. Παρακάτω θα αναφερθούμε και σε ημιπαραμετρικά μοντέλα, όπου δεν θεωρούμε μία γνωστή κατανομή για την βασική συνάρτηση κινδύνου, αλλά εκτιμάται η εν λόγω συνάρτηση μη παραμετρικά με διάφορες τεχνικές.

Σε αυτό το σημείο θα κάνουμε χρήση του πακέτου **parfm.** Το πακέτο για την βασική συνάρτηση κινδύνου υποστηρίζει τις εξής κατανομές: Εκθετική, Weibull, Gompertz, Λογαριθμοκανονική και Λογαριθμολογιστική, ενώ για την μεταβλητή ευπάθειας υποστηρίζει τις κατανομές: την Γάμμα, Positive stable και την Inverse Gaussian ή και καμία. Θα χρησιμοποίησουμε τις τρείς κατανομές για την μεταβλητή ευπάθειας, δηλαδή την Γάμμα, την Positive stable και την Inverse Gaussian.

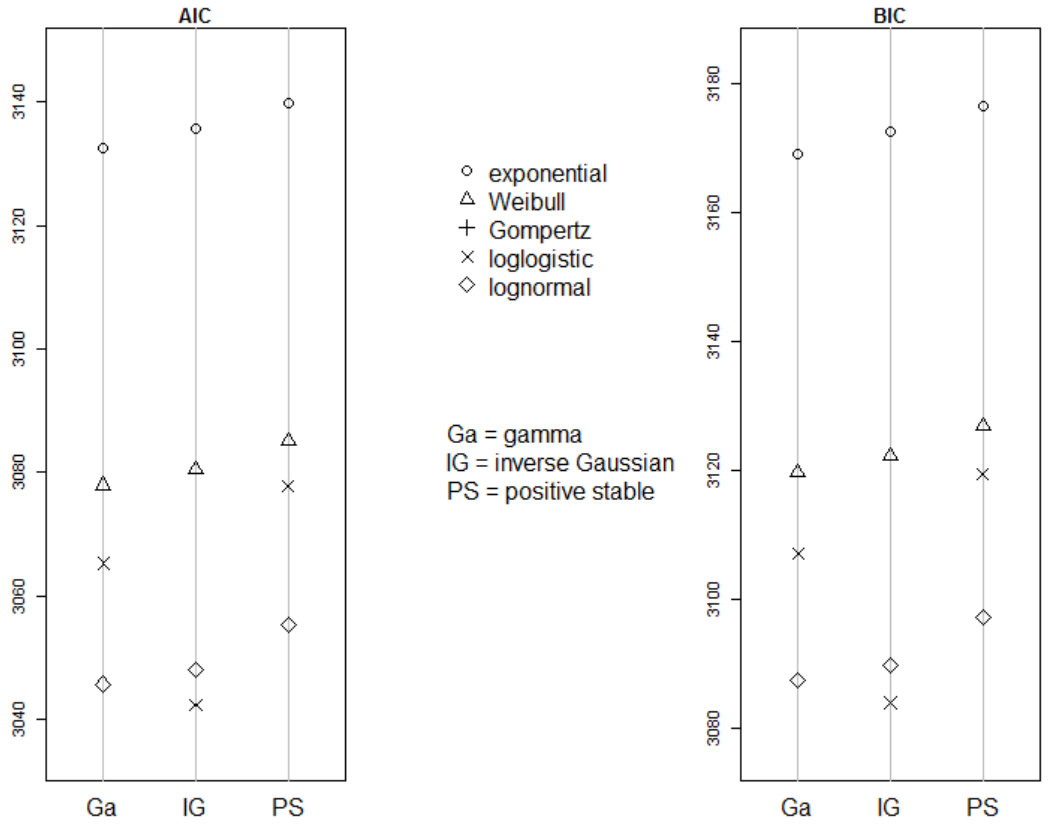
Αρχικά, θα κάνουμε σύγκριση μεταξύ όλων των διαφορετικών μοντέλων, που μπορούν να κατασκευαστούν με το parfm πακέτο με βάση τα κριτήρια AIC και BIC χρησιμοποιώντας στην R την συνάρτηση select.parfm().

```
#Comparison of possible models using parfm package
lymphoma.parfm <- select.parfm(Surv(Survival_Time,Status) ~ Age1 + LDH + Performance_Stat
us + ENS1 + Stage1,cluster = "id", data = lymphoma_data_new, dist = c("exponential", "wei
bull","gompertz", "loglogistic", "lognormal"),frailty = c("gamma", "ingau", "possta"), me
thod="BFGS", maxit=500)
lymphoma.parfm
### - Parametric frailty models - ###
Progress status:
  'ok' = converged
  'nc' = not converged
               Frailty
Baseline           gamma   invGau  posSta
exponential.........ok......ok......ok....
Weibull.............ok......ok......ok....
Gompertz............nc......nc......nc....
loglogistic.........ok......ok......ok....
lognormal...........ok......ok......ok....
> lymphoma.parfm
AIC:
              gamma    ingau   possta
exponential 3132.418 3135.808 3139.714
weibull     3077.884 3080.441 3085.018
gompertz        ----     ----     ----
loglogistic 3065.240 3042.189 3077.673
lognormal   3045.687 3048.090 3055.357

BIC:
              gamma    ingau   possta
exponential 3169.052 3172.442 3176.348
weibull     3119.752 3122.308 3126.886
gompertz        ----     ----     ----
loglogistic 3107.108 3084.057 3119.541
lognormal   3087.555 3089.958 3097.225
```

Παρατηρείται παραπάνω ότι για την επιλογή της Gompertz κατανομής για την ευπαθή μεταβλητή, το μοντέλο δεν συγκλίνει. Αυτό μπορεί να οφείλεται στην επιλογή της μεθόδου βελτιστοποίησης. Συνεπώς, με κατάλληλη επιλογή μεθόδου βελτίστοποίησης ή ακόμη και προσδιορισμό των αρχικών τιμών, με τις οποίες θα τρέξει ο αλγόριθμος της συνάρτησης parfm, το μοντέλο μπορεί να συγκλίνει. Βέβαια μπορεί να συγκλίνει και σε μη αποδεκτή τιμή, το οποίο σημαίνει ότι ο αλγόριθμος τείνει προς τοπικά μέγιστα και όχι στο ολικό μέγιστο της συνάρτησης της λογαριθμοποιημένης πιθανοφάνειας. Το ολικό μέγιστο θα μας δώσει τις αποδεκτές τιμές για την εκτίμηση των συντελεστών του μοντέλου.

Μπορούμε να κάνουμε και μία γραφική απεικόνιση των παραπάνω αποτελεσμάτων χρησιμοποιώντας την εντολή plot (lymphoma.parfm).

```
>plot(lymphoma.parfm)
```



Διάγραμμα 4.6: *Γραφική απεικόνιση των τιμών AIC και BIC, που προκύπτουν προσαρμόζοντας όλα τα μοντέλα χρησιμοποιώντας τις εξής επιλογές κατανομών για την βασική συνάρτηση κινδύνου και τον παράγοντα ευπάθεια: Βασική συνάρτηση κινδύνου – Εκθετική, Weibull,Gompertz, Λογαριθμολογιστική, Λογαριθμοκανονική. Παράγοντας ευπάθεια: Γάμμα, Inverse Gaussian, Positive Stable.*

Παρατηρούμε λοιπόν από το Διάγραμμα 4.6 ότι το καλύτερο μοντέλο για την περιγραφή της συνάρτησης κινδύνου χρησιμοποιώντας όλη τη διαθέσιμη πληροφορία των δεδομένων μας με βάση τα κριτήρια AIC και BIC είναι το Inverse Gaussian μοντέλο ευπάθειας με κατανομή Λογαριθμολογιστική για την βασική συνάρτηση κινδύνου. Επιπλέον, καλές επιλογές μοντέλου φαίνεται να είναι και το Γάμμα μοντέλο, το Positive stable μοντέλο ευπάθειας και το Inverse Gaussian μοντέλο με Λογαριθμοκανονική κατανομή για την βασική συνάρτηση κινδύνου.

Προσαρμόζουμε το «βέλτιστο» μοντέλο στην R χρησιμοποιώντας τις παρακάτω εντολές:

- ✓ Inverse-Gaussian παραμετρικό μοντέλο ευπάθειας χρησιμοποιώντας την Λογαριθμολογιστική κατανομή για τη βασική συνάρτηση κινδύνου

```
> fit6 <- parfm(Surv(Survival_Time,Status) ~ Age1 + LDH + Performance_Status + ENS1 + Sta
ge1 ,cluster="id", frailty = "ingau",data=lymphoma_data_new ,dist='loglogistic')

Execution time: 45.37 second(s)
> summary(fit6)
    ESTIMATE              SE                p-val
 Min.   : 0.01136   Min.   : 0.07661   Min.   :0.0000000
 1st Qu.: 0.49644   1st Qu.: 0.08988   1st Qu.:0.0000000
 Median : 0.59840   Median : 0.09845   Median :0.0000000
```

```
 Mean   : 7.37705   Mean   : 3.05446   Mean   :0.0003092
 3rd Qu.: 1.62816   3rd Qu.: 0.25207   3rd Qu.:0.0000000
 Max.   :52.75771   Max.   :23.45640   Max.   :0.0015460
                                       NA's   :3
> fit6

Frailty distribution: inverse Gaussian
Baseline hazard distribution: Loglogistic
Loglikelihood: -1513.095

                    ESTIMATE SE     p-val
theta               52.758   23.456
alpha                0.011    0.285
kappa                1.157    0.077
Age11                3.042    0.241 0.000 ***
LDH2                 0.607    0.090 0.000 ***
Performance_Status2  0.567    0.092 0.000 ***
ENS11                0.285    0.090 0.002 **
Stage11              0.590    0.105 0.000 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Kendall's Tau: 0.483
```

Η Inverse Gaussian κατανομή για την μεταβλητή ευπάθειας στην συνάρτηση parfm της R ορίζεται ως εξής:

$$f(z) = \frac{1}{\sqrt{2\pi \cdot theta}} \cdot z^{-\frac{3}{2}} \exp\left\{-\frac{(z-1)^2}{2 \cdot theta \cdot z}\right\}, \qquad z > 0, theta > 0.$$

$$E(Z) = \mu = 1 \; και \; V(Z) = \sigma^2 = \frac{1}{\lambda} = theta.$$

Η παραπάνω σχέση αποτελεί μία παραμετροποίηση της συνάρτησης πυκνότητας πιθανότητας της Inverse Gaussian κατανομής όπως δίνεται από τη σχέση (3.23)

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

θεωρώντας $E(Z) = \mu = 1 \; και \; V(Z) = \sigma^2 = \frac{1}{\lambda} = theta.$

Ερμηνεία των αποτελεσμάτων του παραπάνω μοντέλου, που προσαρμόστηκε είναι η παρακάτω:

Η εκτίμηση της παραμέτρου θ, η οποία ουσιαστικά εκφράζει μία εκτίμηση της διασποράς της μεταβλητής ευπάθειας υπό την παραδοχή της Inverse Gaussian κατανομής είναι 52.758 με τυπικό σφάλμα 23.456. Παρατηρούμε λοιπόν μία πολύ μεγάλη εκτίμηση για τη διασπορά της μεταβλητής ευπάθειας το οποίο σημαίνει ότι η ευπάθεια είναι απαραίτητη να υπεισέλθει στο μοντέλο για να εξηγήσει την ετερογένεια του πληθυσμού. Η μηδενική υπόθεση ότι η διασπορά της ευπάθειας είναι 0 απορρίπτεται σε επίπεδο σημαντικότητας $\alpha = 5\%$.

Σύμφωνα με την παραμετροποίηση της Λογαριθμολογιστικής κατανομής, που χρησιμοποιεί η συνάρτηση parfm της R, θεωρούμε ότι:

Μία τυχαία μεταβλητή T, που ακολουθεί την κατανομή $Loglog(\alpha, \kappa)$. Την συγκεκριμένη κατανομή έχουμε εισάγει το 1ο κεφάλαιο, αλλά με παραμέτρους κ και u, όπου $u^k = \nu = \exp(\alpha)$, $\nu = exp(alpha)$ και $alpha$ η εκτίμηση της R για την προσαρμογή του μοντέλου.

Η εκτίμηση των παραμέτρων $alpha$ και $kappa$, καθώς και τα αντίστοιχα τυπικά σφάλματα είναι $0.011\ (0.285)$ και $1.157(0.077)$, αντίστοιχα.

Φαίνεται ότι όλες οι επεξηγηματικές μεταβλητές είναι στατιστικά σημαντικές για την προσαρμογή του μοντέλου στα εν λόγω δεδομένα με σημαντικότερη την ηλικία. Επίσης παρατηρούμε ότι όλοι οι συντελεστές των επεξηγηματικών μεταβλητών είναι θετικοί το οποίο υποδηλώνει ότι αύξηση στην ηλικία, στο LDH, performance status, στο stage και στο ENS επιδεινώνουν τον κίνδυνο άρα και την επιβίωση των ασθενών.

Διαστήματα εμπιστοσύνης Wald για την εκθετική συνάρτηση όλων των συντελεστών του μοντέλου (δηλαδή για τα hazard ratios) δίνονται στην R με τις παρακάτω εντολές:

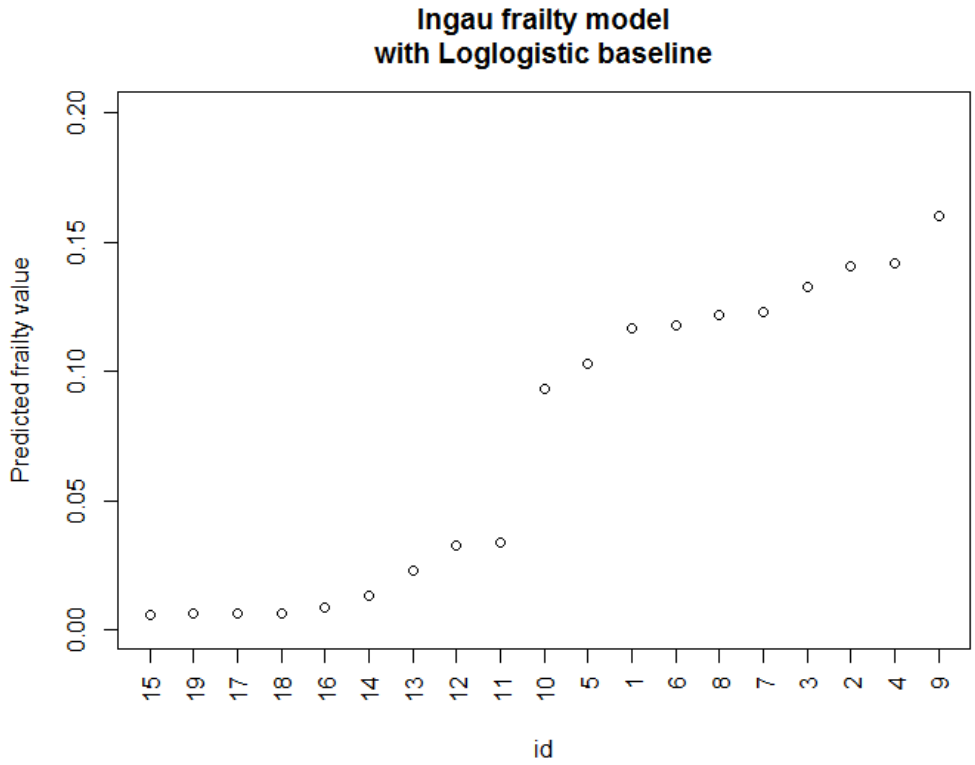
```
> ci.parfm(fit6, level=0.05)
                       low     up
Age11               13.057 33.591
LDH2                 1.539  2.187
Performance_Status2  1.472  2.111
ENS11                1.115  1.586
Stage11              1.469  2.215
```

Εκτιμήσεις της ευπάθειας μπορούν να δοθούν χρησιμοποιώντας την συνάρτηση predict() με όρισμα το εν λόγω παραμετρικό μοντέλο. Στην R δίνουμε την παρακάτω εντολή:

```
 > z6<-predict(fit6)
> z6
Ingau frailty model with Loglogistic baseline
 id frailty
 1  0.117
 2  0.141
 3  0.133
 4  0.142
 5  0.103
 6  0.118
 7  0.123
 8  0.122
 9  0.16
 10 0.093
 11 0.034
 12 0.033
 13 0.023
 14 0.013
 15 0.006
 16 0.009
 17 0.007
 18 0.007
 19 0.006
```

Μπορούμε να δούμε και γραφικά τις τιμές που λαμβάνει η μεταβλητή ευπάθειας, χρησιμοποιώντας τις παρακάτω εντολές στην R.

```
plot(z6, sort = "i",ylim=c(0.001,0.2))
```



Διάγραμμα 4.7: *Γραφική απεικόνιση των εκτιμώμενων τιμών του παράγοντα ευπάθεια για το Inverse Gaussian μοντέλο ευπάθειας με Λογαριθμολογιστική κατανομή για την βασική συνάρτηση, που προσαρμόστηκε στα Lymphoma data.*

Από το Διάγραμμα 4.7 βλέπουμε ότι υπάρχει μεγάλη ετερογένεια στον πληθυσμό αφού όλες οι τιμές της ευπάθειας στις διαφορετικές ομάδες διαφέρουν πολύ από τη μονάδα. Η διάμεσος και η μέση τιμή των τιμών της ευπάθειας λαμβάνουν τις τιμές 0.093 και 0.073, αντίστοιχα.

Σε αυτό το σημείο θα δώσουμε κάποιες εκτιμήσεις της δεσμευμένης συνάρτησης επιβίωσης, αλλά και της δεσμευμένης συνάρτησης κινδύνου $S(t_{ij}|Z_j, \boldsymbol{X}_{ij})$ και $h\left(t_{ij}\middle|Z_j, \boldsymbol{X}_{i_j}\right)$, όπως προκύπτει από το Inverse Gaussian μοντέλο ευπάθειας με Λογαριθμολογιστική κατανομή για την βασική συνάρτηση κινδύνου για διάφορες τιμές των συμμεταβλητών.

Η δεσμευμένη συνάρτηση επιβίωσης για μία j μονάδα του πληθυσμού στην i ομάδα για χρόνο επιβίωσης t_{ij} για δεδομένη τιμή της μεταβλητής ευπάθειας $Z_j = z_j$ δίνεται από τη σχέση:

$$S(t_{ij}|Z_j, \boldsymbol{X}_{ij}) = \exp(-z_j H_0(t_{ij})\eta_j), \text{όπου } \eta_j = \exp(\boldsymbol{\beta^T X_{ij}}).$$

Επιπλέον, η δεσμευμένη συνάρτηση κινδύνου μία j μονάδα του πληθυσμού στην i ομάδα για χρόνο επιβίωσης t_{ij} για δεδομένη τιμή της μεταβλητής ευπάθειας $Z_j = z_j$ δίνεται από τη σχέση:

$$h\left(t_{ij}\middle|Z_j, \boldsymbol{X}_{i_j}\right) = z_j h_0(t_{ij})\eta_j, \text{όπου } \eta_j = \exp(\boldsymbol{\beta}^T \boldsymbol{X}_{ij}).$$

Θα δημιουργήσουμε δύο συναρτήσεις στην R, οι οποίες καλούνται COND_HAZARD και COND_SURVIVAL. Οι δύο συναρτήσεις λαμβάνουν τις ίδιες παραμέτρους, οι οποίες είναι οι εξής:

t: Είναι ο χρόνος t_{ij} στις συναρτήσεις S και h.

frailty: Η τιμή για την μεταβλητή ευπάθειας, όπως προκύπτει από το μοντέλο για τη συγκεκριμένη ομάδα ασθενών, που αντιστοιχεί.

b1, b2, b3, b4, b5: Οι εκτιμήσεις των συντελεστών του μοντέλου.

Age, LDH, Performance_Status, ENS, Stage: Οι επεξηγηματικές μεταβλητές του μοντέλου, στις οποίες θα δώσουμε διάφορες τιμές για να εκτιμήσουμε την συνάρτηση επιβίωσης.

baseline_distr: Η κατανομή που θα χρησιμοποιηθεί για την βασική συνάρτηση κινδύνου. Οι διαθέσιμες επιλογές είναι baseline_distr = "exponential", baseline_distr = "Weibull", baseline_distr = "lognormal" και baseline_distr = "loglogistic".

mean, sd: Οι τιμές των παραμέτρων για την Λογαριθμοκανονική κατανομή, αν χρησιμοποιηθεί ως κατανομή για την βασική συνάρτηση κινδύνου.

l: Η τιμή της παραμέτρου της Εκθετικής κατανομής, αν χρησιμοποιηθεί ως κατανομή για την βασική συνάρτηση κινδύνου.

rho, lambda: Οι τιμές των παραμέτρων της Weibull κατανομής, αν χρησιμοποιηθεί ως κατανομή για την βασική συνάρτηση κινδύνου.

alpha, kappa: Οι τιμές των παραμέτρων της Λογαριθμολογιστικής κατανομής, αν χρησιμοποιηθεί ως κατανομή για την βασική συνάρτηση κινδύνου.

Η συνάρτηση στην R για την δεσμευμένη συνάρτηση κινδύνου είναι:

```
COND_HAZARD<-function(t, frailty, b1, b2, b3, b4, b5, Age, LDH, Performance_Status, ENS,
Stage, mean=NULL, sd=NULL, l=NULL, rho=NULL, lambda=NULL, alpha=NULL, kappa=NULL, baselin
e_distr)
  {

  hazard<-t
  if (baseline_distr=="exponential"){

    h<-l

  }else if (baseline_distr=="weibull"){

    h<-lambda*rho*t^(rho-1)

  }else if (baseline_distr=="lognormal"){

    h<-hlnorm(x=t, meanlog=mean, sdlog=sd)

  }else if (baseline_distr=="loglogistic"){
      v<-exp(alpha)
      h<-(v*kappa*(t^(kappa-1)))/(1+v*(t^kappa))
      }
```

```
  equation<-b1*Age+b2*LDH+Performance_Status*b3+ENS*b4+Stage*b5
  hazard<-frailty*h*exp(equation)

  #cat(paste0("Baseline Distrbution = ",baseline_distr),"\n",paste0("Conditional Hazard H
(t|X,Z) = ", hazard))
  return(hazard)
}
```

Η συνάρτηση στην R για την δεσμευμένη συνάρτηση επιβίωσης είναι:

```
COND_SURVIVAL<-function(t, frailty, b1, b2, b3, b4, b5, Age, LDH, Performance_Status, ENS
, Stage, mean=NULL, sd=NULL, l=NULL, rho=NULL, lambda=NULL,alpha=NULL, kappa=NULL, baseli
ne_distr)
{
  S<-t
  if (baseline_distr=="exponential"){

    H<-l*t

  }else if (baseline_distr=="weibull"){

    H<-lambda*(t^rho)

  }else if (baseline_distr=="lognormal"){

    H<-Hlnorm(x=t, meanlog=mean, sdlog=sd)
  }else if (baseline_distr=="loglogistic"){
    v<-exp(alpha)
    H<-(-log(1/(1+v*(t^kappa))))
  }

  equation<-b1*Age+b2*LDH+Performance_Status*b3+ENS*b4+Stage*b5
  S<-exp(-frailty*H*exp(equation))

  #cat(paste0("Baseline Distrbution = ",baseline_distr),"\n",paste0("Conditional Survival
S(t|X,Z) = ", S))
  return(S)
}
```

Μπορούμε να δώσουμε μία εκτίμηση να επιβιώσει κάποιος 10 χρόνια για το καλύτερο μοντέλο για $Age1 = 0$ χρόνων, με $LDH = 1$, $Performance\ Status = 0$, $ENS1 = 0$, $Stage1 = 0$. Να σημειώσουμε εδώ ότι για την ευπαθή μεταβλητή χρησιμοποιούμε την τιμή 0.040, που αντιστοιχεί στην εκτίμηση της διαμέσου της κατανομής $Inverse\ Gaussian\ (mean = 1, shape = 1/theta)$, $όπου\ theta = 1/\lambda =$ εκτίμηση της διασποράς της κατανομής όπως προκύπτει από το εν λόγω μοντέλο.

```
install.packages("statmod")
library(statmod)
theta<-52.758
p50<-qinvgauss(p=0.5, mean=1, disp=theta)

b1<-3.042
```

```
b2<-0.607
b3<-0.567
b4<-0.285
b5<-0.590
> estimate1<-COND_SURVIVAL(t=10,frailty=p50, b1, b2, b3, b4, b5, 0, 1, 0, 0, 0, baseline_
distr="loglogistic",alpha=0.011, kappa=1.157)
> estimate1
[1] 0.818122
```

Συνεπώς, η πιθανότητα ένας ασθενής να επιβιώσει 10 χρόνια, ο οποίος είναι μικρότερος των 60 χρόνων, με υψηλά επίπεδα LDH, δηλαδή σοβαρή κυτταρική βλάβη, με καλή φυσική κατάσταση, με ≤ 1 αριθμό σημείων προσβαλλόμενο με καρκίνο και σε $1^{ο}$ ή $2^{ο}$ στάδιο είναι ίση με 81%.

```
> estimate2<-COND_SURVIVAL(t=10,frailty=p50, b1, b2, b3, b4, b5, 1,1,1,1,1,baseline_distr
="loglogistic",alpha=0.011, kappa=1.157)
> estimate2
[1] 1.89104e-08
```

Συνεπώς, η πιθανότητα ένας ασθενής να επιβιώσει 10 χρόνια, ο οποίος είναι άνω των 60 χρόνων, με υψηλά επίπεδα LDH, δηλαδή μεγαλύτερη κυτταρική βλάβη, σε κακή φυσική κατάσταση, με περισσότερα από 1 σημεία προσβαλλόμενα με καρ κίνο και σε $3^{ο}$ ή $4^{ο}$ στάδιο είναι ίση με 0.
Επιπλέον, οι αντίστοιχες εκτιμήσεις για την δεσμευμένη συνάρτηση κινδύνου είναι:

```
> estimate3<-COND_HAZARD(t=10,frailty=p50, b1, b2, b3, b4, b5, 0, 1, 0, 0, 0, baseline_di
str="loglogistic",alpha=0.011, kappa=1.157)
> estimate3
[1] 0.007925283
```

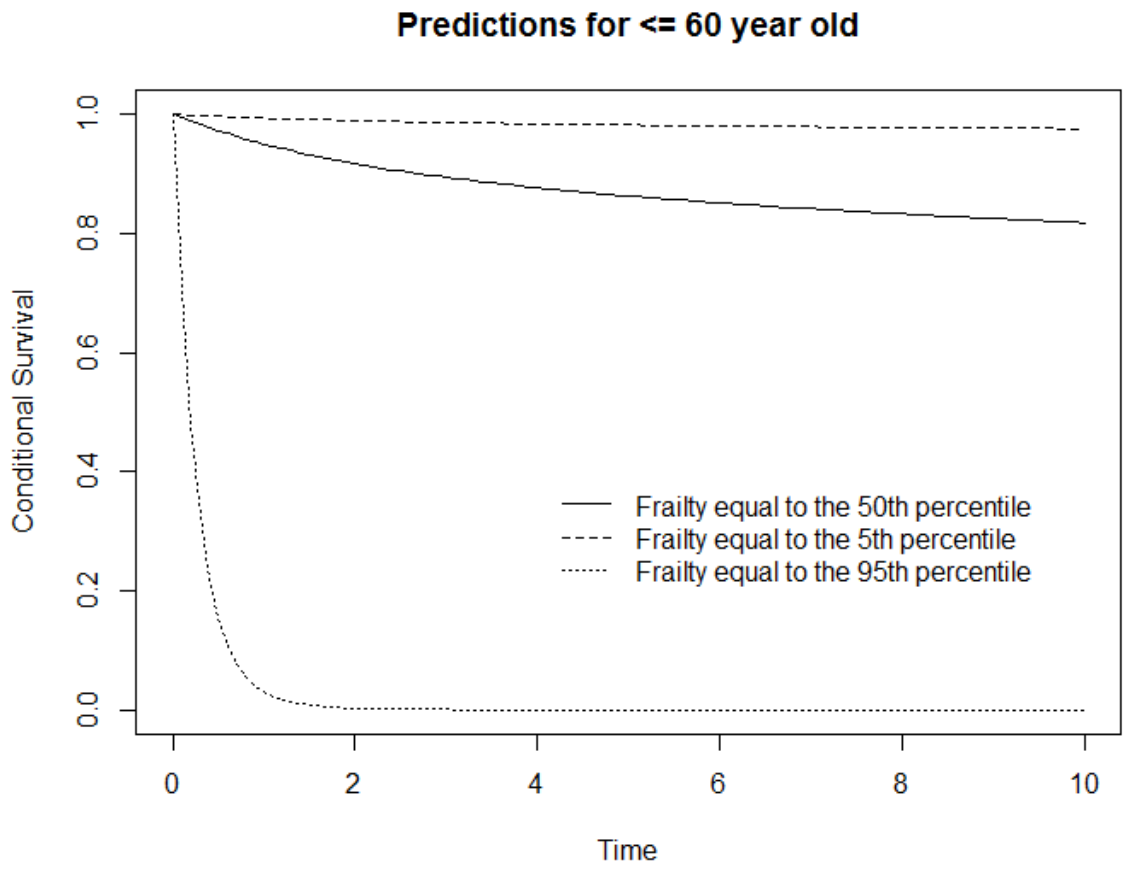
Συνεπώς, η πιθανότητα να πεθάνει ένας ασθενής στα 10 χρόνια δεδομένου ότι έχει επιβιώσει μέχρι εκείνη τη στιγμή, ο οποίος είναι μικρότερος των 60 χρόνων, με υψηλά επίπεδα LDH, δηλαδή σοβαρή κυτταρική βλάβη, με καλή φυσική κατάσταση, με ≤ 1 αριθμό σημείων προσβαλλόμενο με καρκίνο, σε $1^{ο}$ ή $2^{ο}$ στάδιο είναι ίση με 0.7%.

```
> estimate4<-COND_HAZARD(t=10,frailty=p50, b1, b2, b3, b4, b5, 1,1,1,1,1,baseline_distr="
loglogistic",alpha=0.011, kappa=1.157)
> estimate4
[1] 0.7020875
```

Η πιθανότητα να πεθάνει ένας ασθενής στα 10 χρόνια δεδομένου ότι έχει επιβιώσει μέχρι εκείνη τη στιγμή, ο οποίος είναι άνω των 60 χρόνων, με υψηλά επίπεδα LDH, δηλαδή μεγαλύτερη κυτταρική βλάβη, σε κακή φυσική κατάσταση, με περισσότερα από 1 σημεία προσβαλλόμενα με καρκίνο σε $3^{ο}$ ή $4^{ο}$ στάδιο είναι ίση με 70%.

Σε αυτό το σημείο θα κάνουμε το Διάγραμμα 4.8, το οποίο θα περιλαμβάνει την εκτίμηση της δεσμευμένης συνάρτησης επιβίωσης για διαφορετικές τιμές του χρόνου t θεωρώντας τιμές για την μεταβλητή ευπάθειας το 50%, 95% και 5% ποσοστιαία σημεία της κατανομής Inverse Gaussian($mean = 1, shape = 1/theta$), όπου $theta = 1/\lambda =$ εκτίμηση της διασποράς της κατανομής όπως προκύπτει από το εν λόγω μοντέλο. Οι εκτιμήσεις της συνάρτησης επιβίωσης που απεικονίζονται αφορούν τους ασθενείς, που είναι μικρότεροι των 60 χρόνων, με υψηλά επίπεδα LDH, δηλαδή σοβαρή κυτταρική βλάβη, με καλή φυσική κατάσταση, με ≤ 1 αριθμό σημείων προσβαλλόμενο με καρκίνο σε $1^{ο}$ ή $2^{ο}$ στάδιο.

```
t<-seq(0,10,0.1)

theta<-52.758
p50<-qinvgauss(p=0.5, mean=1, shape=1/theta)
p5<-qinvgauss(p=0.05, mean=1, shape=1/theta)
p95<-qinvgauss(p=0.95, mean=1, shape=1/theta)

estimate5<-COND_SURVIVAL(t,frailty=p50,  b1, b2, b3, b4, b5, 0, 1, 0, 0, 0, baseline_dist
r="loglogistic",alpha=0.011, kappa=1.157)
estimate6<-COND_SURVIVAL(t,frailty=p5, b1, b2, b3, b4, b5, 0, 1, 0, 0, 0, baseline_distr=
"loglogistic",alpha=0.011, kappa=1.157)
estimate7<-COND_SURVIVAL(t,frailty=p95,  b1, b2, b3, b4, b5, 0, 1, 0, 0, 0, baseline_dist
r="loglogistic",alpha=0.011, kappa=1.157)

plot(t, estimate7, type="l", lty=3, xlab="Time", ylab="Conditional Survival", main="Predi
ctions for <= 60 year old")
lines(t, estimate6, lty=2)
lines(t, estimate5)
legend( x=4, y=0.4,legend=c("Frailty equal to the 50th percentile", "Frailty equal to the
5th percentile", "Frailty equal to the 95th percentile"), lty=1:3, box.lty=0)
```

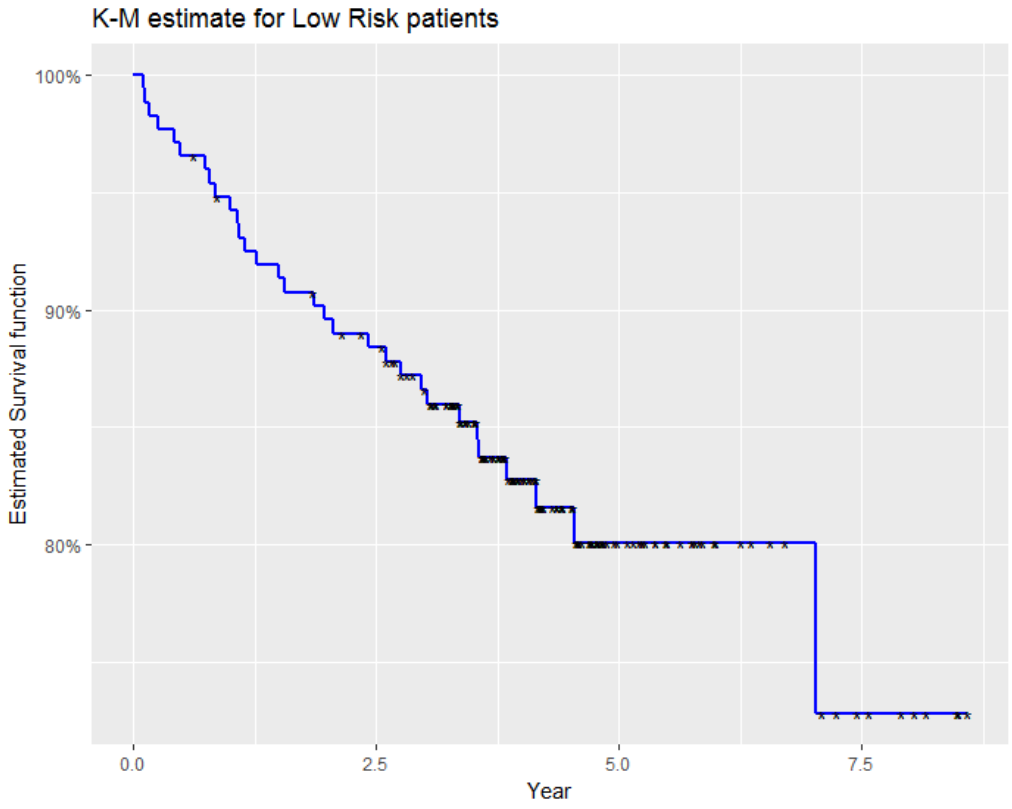


Διάγραμμα 4.8: *Συναρτήσεις επιβίωσης για το μοντέλο Inverse Gaussian με Λογαριθμολογιστική κατανομή για ασθενείς* $<= 60$ *χρονών, με* $LDH = 1$, $Performance\ Status = 0$, $ENS = 0$ *και* $Stage = 0$ *για διάφορες τιμές της μεταβλητής ευπάθειας.*

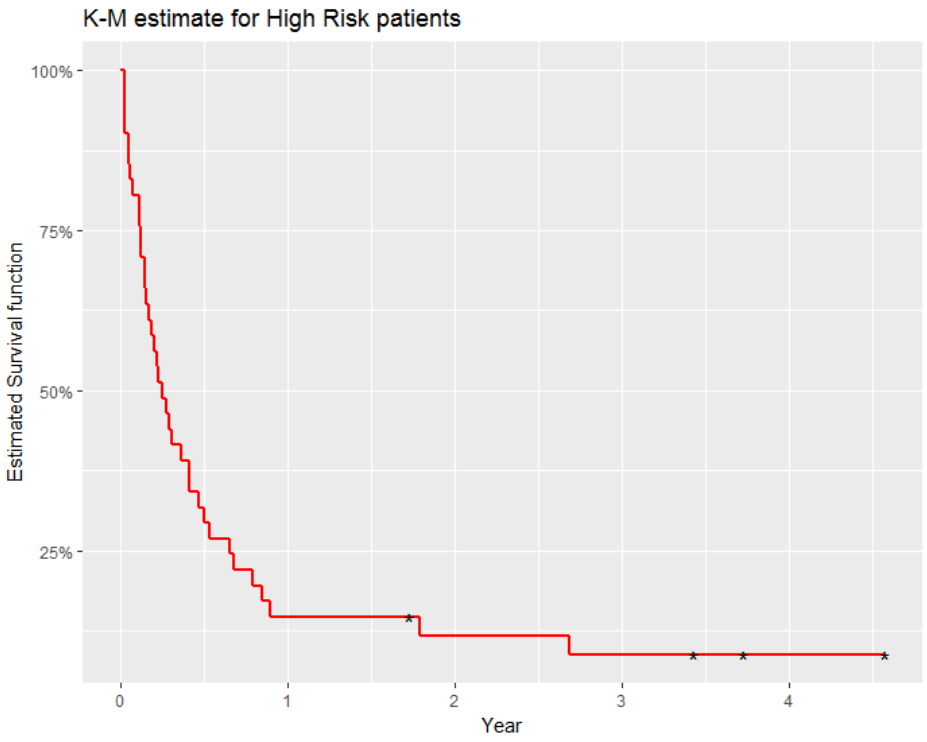
Να αναφέρουμε σε αυτό το σημείο ότι σε όλη την υπόλοιπη εργασία θεωρούμε ως Low Risk ασθενείς αυτούς που είναι <= 60 χρονών, με LDH = 0, Performance Status = 0, ENS = 0 και Stage = 0, ενώ High Risk ασθενείς αυτούς που > 60 χρονών, με LDH = 1, Performance Status = 1, ENS = 1 και Stage = 1.

Τα γραφήματα των εκτιμήσεων Kaplan Meier για Low Risk ασθενείς και High Risk ασθενείς δίνονται από τις παρακάτω εντολές στην R και παρουσιάζονται στα Διαγράμματα 4.9 και 4.10, αντίστοιχα.

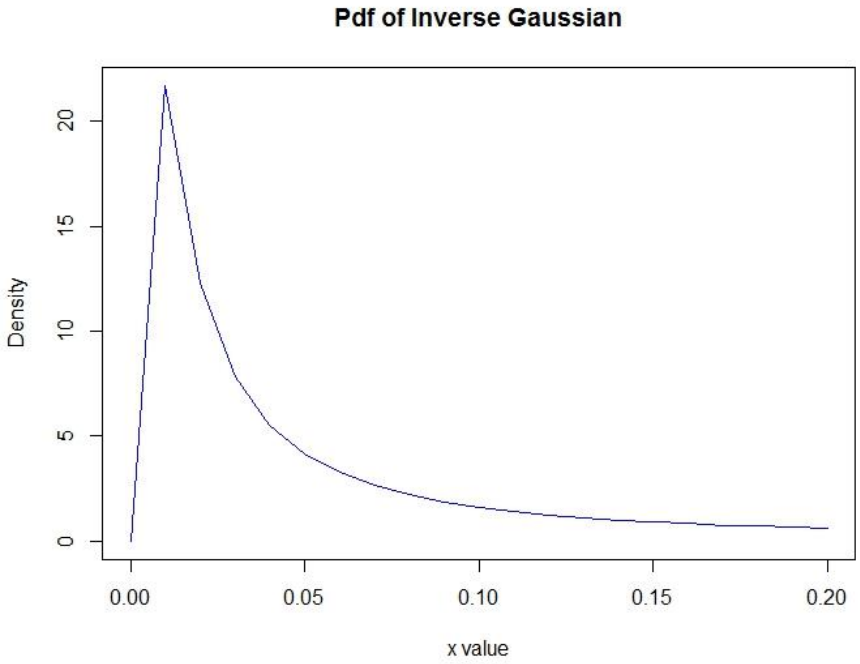
```
data0<-subset(lymphoma_data_new,Age1==0 & LDH==0 & Performance_Status==0 & ENS1==0 & Stag
e1==0 )
data1<-subset(lymphoma_data_new,Age1==1 & LDH==1 & Performance_Status==1 & ENS1==1 & Stag
e1==1 )

fit0 <- survfit(Surv(Survival_Time, Status) ~ 1 , data = data0)
fit1<-survfit(Surv(Survival_Time, Status) ~ 1 , data = data1)

autoplot(fit0, main="K-M estimate for Low Risk patients",conf.int = FALSE,surv.linetype=1
,surv.size=1,
         censor.shape = '*', censor.size = 4, facets = TRUE, ncol = 2, surv.colour="blue",
xlab="Year", ylab="Estimated Survival function")
autoplot(fit1, main="K-M estimate for High Risk patients",conf.int = FALSE,surv.linetype=
1,surv.size=1,
       censor.shape = '*', censor.size = 5, facets = TRUE, ncol = "red", surv.colour="red
", xlab="Year", ylab="Estimated Survival function")
```



Διάγραμμα 4.9: *Γραφική απεικόνιση των εκτιμήσεων του χρόνου επιβίωσης για Low Risk ασθενείς, χρησιμοποιώντας την εκτιμήτρια Kaplan-Meier.*



Διάγραμμα 4.10: *Γραφική απεικόνιση των εκτιμήσεων του χρόνου επιβίωσης για High Risk ασθενείς, χρησιμοποιώντας την εκτιμήτρια Kaplan-Meier.*

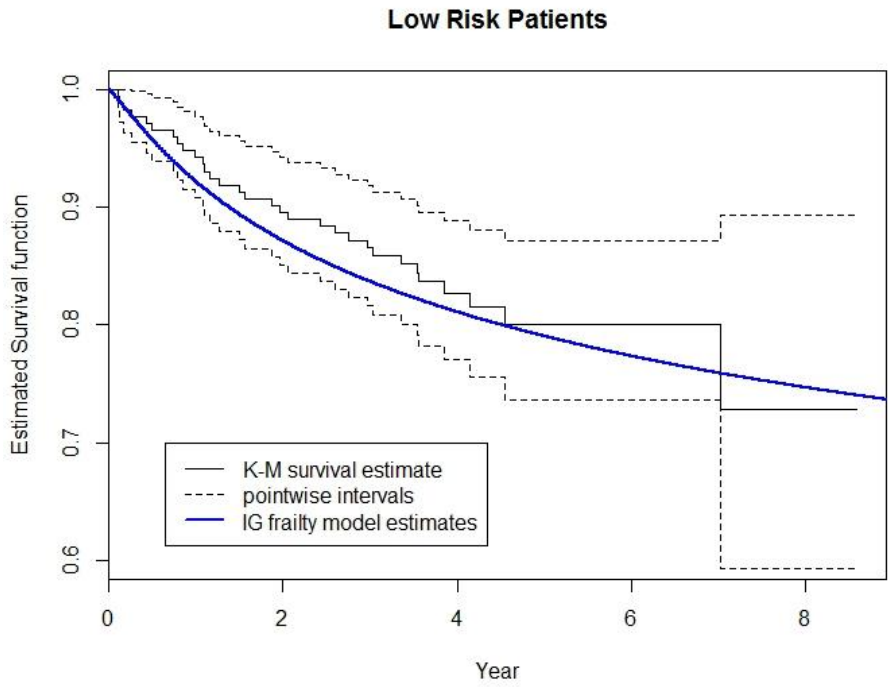


Διάγραμμα 4.11: *Γραφική απεικόνιση της συνάρτησης πυκνότητας πιθανότητας της ευπάθειας η οποία ακολουθεί Inverse Gaussian κατανομή με μέση τιμή 1 και διασπορά 52.758 (όπως προκύπτει από το Inverse-Gaussian παραμετρικό μοντέλο ευπάθειας χρησιμοποιώντας την Λογαριθμολογιστική κατανομή για τη βασική συνάρτηση κινδύνου που προσαρμόστηκε στα δεδομένα).*

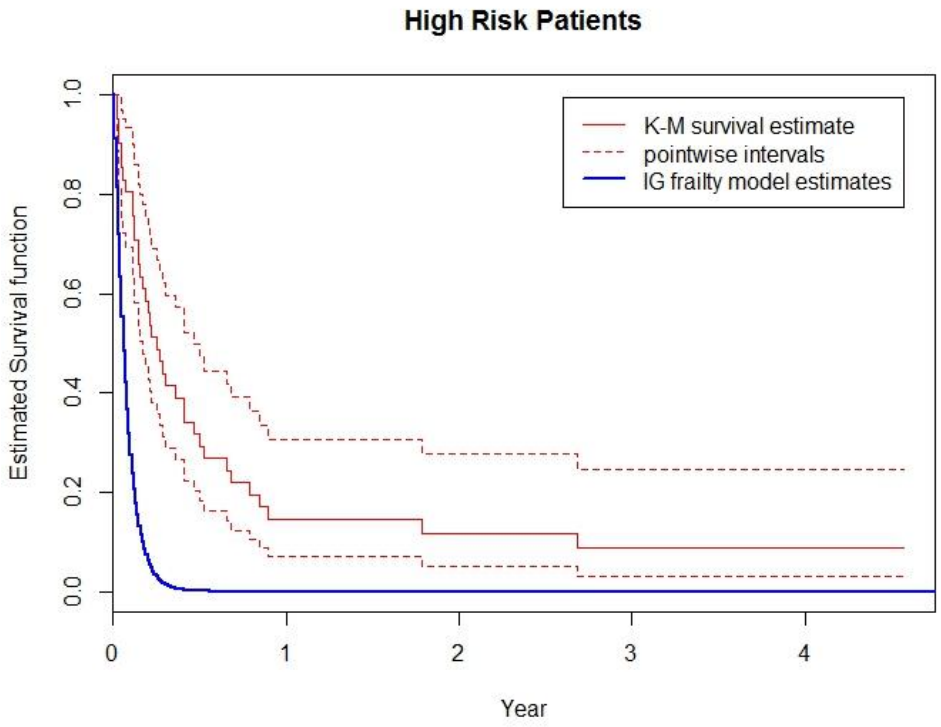
Στο Διάγραμμα 4.12 απεικονίζονται οι εκτιμήσεις του χρόνου επιβίωσης για Low Risk ασθενείς, χρησιμοποιώντας την εκτιμήτρια Kaplan-Meier, καθώς και διαστήματα εμπιστοσύνης αυτών, καθώς και απεικόνιση των εκτιμήσεων της συνάρτησης επιβίωσης, όπως προκύπτει από το μοντέλο Inverse Gaussian με Λογαριθμολογιστική κατανομή για την βασική συνάρτηση κινδύνου θεωρώντας ως τιμή για την μεταβλητή ευπάθειας το 70% ποσοστιαίο σημείο της Inverse Gaussian κατανομής (0.1167974). Παρατηρούμε ότι το Διάγραμμα 4.12 εμφανίζει την καμπύλη της εκτίμησης της δεσμευμένης συνάρτησης επιβίωσης ως προς το παράγοντα ευπάθεια κοντά στην εκτίμηση της συνάρτησης επιβίωσης, χρησιμοποιώντας τον εκτιμητή Kaplan-Meier. Στο συγκεκριμένο Διάγραμμα έχουμε χρησιμοποιήσει το 70% ποσοστιαίο σημείο της κατανομής της μεταβλητή ευπάθειας. Το σημείο αυτό επιλέχθηκε γιατί είναι κοντά στη διάμεσο των εκτιμήσεων της μεταβλητής ευπάθειας για τα διάφορα clusters.

Στο Διάγραμμα 4.13 απεικονίζονται οι εκτιμήσεις του χρόνου επιβίωσης για High Risk ασθενείς, χρησιμοποιώντας την εκτιμήτρια Kaplan-Meier, καθώς και τα αντίστοιχα 95% διαστήματα εμπιστοσύνης, καθώς και η απεικόνιση των εκτιμήσεων της συνάρτησης επιβίωσης, όπως προκύπτει από το μοντέλο Inverse Gaussian με Λογαριθμολογιστική κατανομή για την βασική συνάρτηση κινδύνου θεωρώντας ως τιμή για την μεταβλητή ευπάθειας το 70% ποσοστιαίο σημείο της Inverse Gaussian κατανομής.

```
p70<-qinvgauss(p=0.70, mean=1, shape=1/theta)
estimate8<-COND_SURVIVAL(t,frailty=p70,  b1, b2, b3, b4, b5, 0, 0, 0, 0, 0, baseline_dist
r="loglogistic",alpha=0.011, kappa=1.157)
plot(fit0, col="black", xlab="Year", ylab="Estimated Survival function", lwd="1", ylim=c(
0.6,1),main="Low Risk Patients")
lines(t, estimate8, col="black", lwd=2, type="s")
legend(0.65, 0.7, legend=c("K-M survival estimate","pointwise intervals","IG frailty mode
l estimates"), lty=c(1,2,1),lwd=c(1,1,2) ,col=c("black","black","blue"))
t<-seq(0, 6, 0.01)
estimate9<-COND_SURVIVAL(t,frailty=p70,  b1, b2, b3, b4, b5, 1, 1, 1, 1, 1, baseline_dist
r="loglogistic",alpha=0.011, kappa=1.157)
plot(fit1, col="red", xlab="Year", ylab="Estimated Survival function", lwd="1", ylim=c(0,
1),main="High Risk Patients")
lines(t, estimate9, col="blue", lwd=2, type="s")
legend(2.6, 0.9955, legend=c("K-M survival estimate","pointwise intervals","IG frailty mo
del estimates"), lty=c(1,2,1),lwd=c(1,1,2),col="red" ,col=c("red","red","blue"))
```



Διάγραμμα 4.12: *Γραφική απεικόνιση των εκτιμήσεων της συνάρτησης επιβίωσης για Low Risk ασθενείς, χρησιμοποιώντας την εκτιμήτρια Kaplan-Meier, καθώς και διαστήματα εμπιστοσύνης αυτών, καθώς και επεικόνιση των εκτιμήσεων της συνάρτησης επιβίωσης, όπως προκύπτει από το μοντέλο Inverse Gaussian με Λογαριθμολογιστική κατανομή για την βασική συνάρτηση κινδύνου και ως τιμή για την μεταβλητή ευπάθειας το 70% ποσοστιαίο σημείο της Inverse Gaussian κατανομής.*



Διάγραμμα 4.13: *Γραφική απεικόνιση των εκτιμήσεων του χρόνου επιβίωσης για High Risk ασθενείς, χρησιμοποιώντας την εκτιμήτρια Kaplan-Meier, καθώς και διαστήματα εμπιστοσύνης αυτών, καθώς και επεικόνιση των εκτιμήσεων της συνάρτησης επιβίωσης, όπως προκύπτει από το μοντέλο Inverse Gaussian με Λογαριθμολογιστική κατανομή για την βασική συνάρτηση κινδύνου και ως τιμή για την μεταβλητή ευπάθειας το 70% ποσοστιαίο σημείο της Inverse Gaussian κατανομής.*

Η εκτίμηση της πιθανότητας επιβίωσης με βάση το 70^{ο} ποσοστημόριο της μεταβλητής ευπάθειας είναι πολύ καλή για τους low-risk ασθενείς αλλά υποεκτιμάται σημαντικά για τους high-risk ασθενείς.

Εδώ λοιπόν θα χρησιμοποιήσουμε και έναν άλλο τρόπο για να εκτιμήσουμε την συνάρτηση επιβίωσης για low-risk και high-risk ασθενείς. Ο τρόπος αυτός βασίζεται στην ιδέα του να 'integrate out' την ευπάθεια από την αρχή.

Η συνάρτηση επιβίωσης, ως προς τις συμμεταβλητές μόνο, για την j μονάδα του δείγματος στην i ομάδα για χρόνο επιβίωσης t_{ij} για συγκεκριμένες τιμές των επεξηγηματικών μεταβλητών δίνεται από τη σχέση:

$$S(t_{ij}|\,\boldsymbol{X}_j) = \exp\left(-G\{\exp(\boldsymbol{\beta}^T\boldsymbol{X})\,H_0(t_{ij})\}\right),$$

όπου $G(u) = -\log\big(L(u)\big)$ *και* $L(u)$ ο Laplace μετασχηματισμός της κατανομής της μεταβλητής ευπάθειας (Vonta 1996), (Androulakis, et al., 2012).

Για την Positive Stable ο μετασχηματισμός Laplace δίνεται από τη σχέση:

$$L(u) = e^{-u^{\gamma}}.$$

Συνεπώς, η συνάρτηση επιβίωσης λαμβάνει την μορφή:

$$S(t|\bar{X}) = e^{-\left(e^{\beta^T X}H_0(t)\right)^{\gamma}}.$$

Για την $\Gamma\acute{\alpha}\mu\mu\alpha(\alpha,\beta)$ ο μετασχηματισμός Laplace δίνεται από τη σχέση:

$$L(u)=\frac{\beta^{a}}{(\beta+u)^{a}}.$$

Για Γάμμα κατανομή με $E(u)=1$ και $V(u)=\theta$, η συνάρτηση $G(u)=\frac{1}{\theta}\log[(1+\theta u)]$.

Συνεπώς, η συνάρτηση επιβίωσης λαμβάνει την μορφή:

$$S(t|\bar{X})=\exp\{\frac{1}{\theta}\log\left(1+\theta e^{\beta^{T}X}H_{0}(t)\right)\}.$$

Για Inverse Gaussian κατανομή με $E(u)=1$ και $V(u)=1/2b$, η συνάρτηση $G(u)=2(\sqrt{b(b+u)}-b)$

Συνεπώς, η συνάρτηση επιβίωσης λαμβάνει την μορφή:

$$S(t|\bar{X})=\exp\{2(\sqrt{b\left(b+e^{\beta^{T}X}H_{0}(t)\right)}-b)\}.$$

Συνεπώς, με βάση την γενική μορφή της δεσμευμένης συνάρτησης επιβίωσης, ως προς τις συμμεταβλητές μόνο, δημιουργήσαμε την παρακάτω συνάρτηση.

```
G<-function(u, distr, g=NULL, theta=NULL, b=NULL)
{
  res<-u

  if(distr=="Positive Stable"){

    res<-u^g

  }else if (distr=="gamma"){

    res<-1/theta*log(1+theta*u)

  }else if(distr=="Inverse Gaussian"){

    res<-2*(sqrt(b*(b+u))-b)
  }

  return(res)
}
```

```
UNCOND_SURVIVAL<-function(t, b1, b2, b3, b4, b5, Age, LDH, Performance_Status, ENS, Stage
, mean=NULL, sd=NULL, l=NULL, rho=NULL, lambda=NULL,baseline_distr, frailty_distr,g=NULL,
theta=NULL, b=NULL, alpha=NULL,kappa=NULL)
  {

  if (baseline_distr=="exponential"){
```

```
    H<-l*t

  }else if (baseline_distr=="weibull"){

    H<-lambda*(t^rho)

  }else if (baseline_distr=="lognormal"){

    H<-Hlnorm(x=t, meanlog=mean, sdlog=sd)

  }else if (baseline_distr=="loglogistic"){
    v<-exp(alpha)
    H<-(-log(1/(1+v*t^kappa)))
  }

  equation<-b1*Age+b2*LDH+Performance_Status*b3+ENS*b4+Stage*b5
  S<-exp(-G(exp(equation)*H,distr=frailty_distr, g=g, theta=theta, b=b ))

  #cat(paste0("Baseline Distrbution = ",baseline_distr),"\n",paste0("Conditional Survival
S(t|X) = ", S))
  return(S)
}
```

Μία συνάρτηση για το γράφημα της συνάρτησης επιβίωσης για Low risk και High Risk ασθενείς.

```
plot.frailty <- function(x, legend = c("High Risk Patients", "Low Risk Patients"),
                         xlab = "Year", ylab="Estimated Survival function", lwd = 2,
                         col = c("red", "black"), at = c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
),
                         ...)
{
  plot(x[, 1], x[, 2], type = "l", ylim = c(0, 1), xaxt = "n",
       xlab = xlab, ylab = ylab, col = col[2], lwd = 2, ...)
  axis(1, at = at)
  lines(x[, 1], x[, 3], col = col[1], lwd=2)
  legend("topright", legend = legend, lwd = lwd, col = col)
}
```

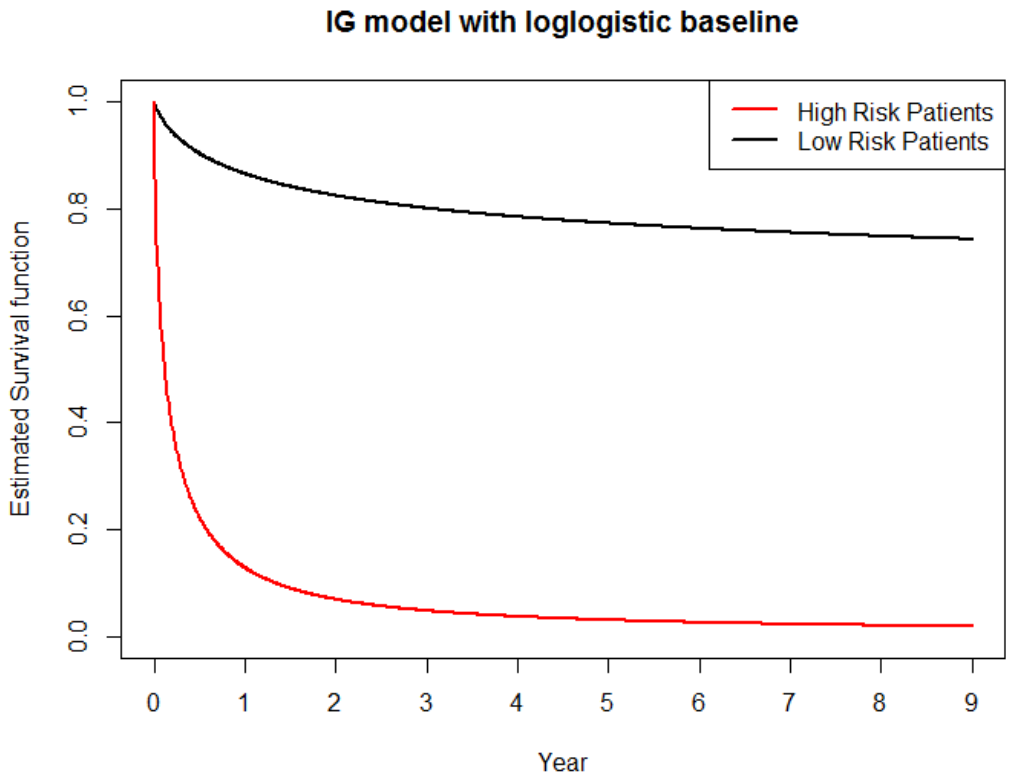
Υπολογίζουμε τις συναρτήσεις επιβίωσης για Low Risk και High Risk ασθενείς για διάφορες τιμές του t.

```
t <- seq(0, 10, 0.01)
b1<-3.042
b2<-0.607
b3<-0.567
b4<-0.285
b5<-0.590
theta<-52.758
b<-1/(2*theta)
s0.invg <-UNCOND_SURVIVAL(t, b1, b2, b3, b4, b5, Age=0, LDH=0, Performance_Status=0, ENS=
0, Stage=0,  baseline_distr="loglogistic", frailty_distr = "Inverse Gaussian",b=b, alpha=
0.011, kappa=1.157)
```

```
s1.invg <- UNCOND_SURVIVAL(t, b1, b2, b3, b4, b5,  Age=1, LDH=1, Performance_Status=1, EN
S=1, Stage=1, baseline_distr="loglogistic", frailty_distr = "Inverse Gaussian",b=b, alpha
=0.011, kappa=1.157)
```

Η γραφική απεικόνιση των δύο καμπυλών επιβίωσης δίνεται από τη συνάρτηση, που κατασκευάσαμε την οποία καλούμε plot.frailty και παρουσιάζεται στο Διάγραμμα 4.14.

```
plot.frailty(data.frame(t, s0.invg, s1.invg), main = "IG model with loglogistic baseline"
)
```

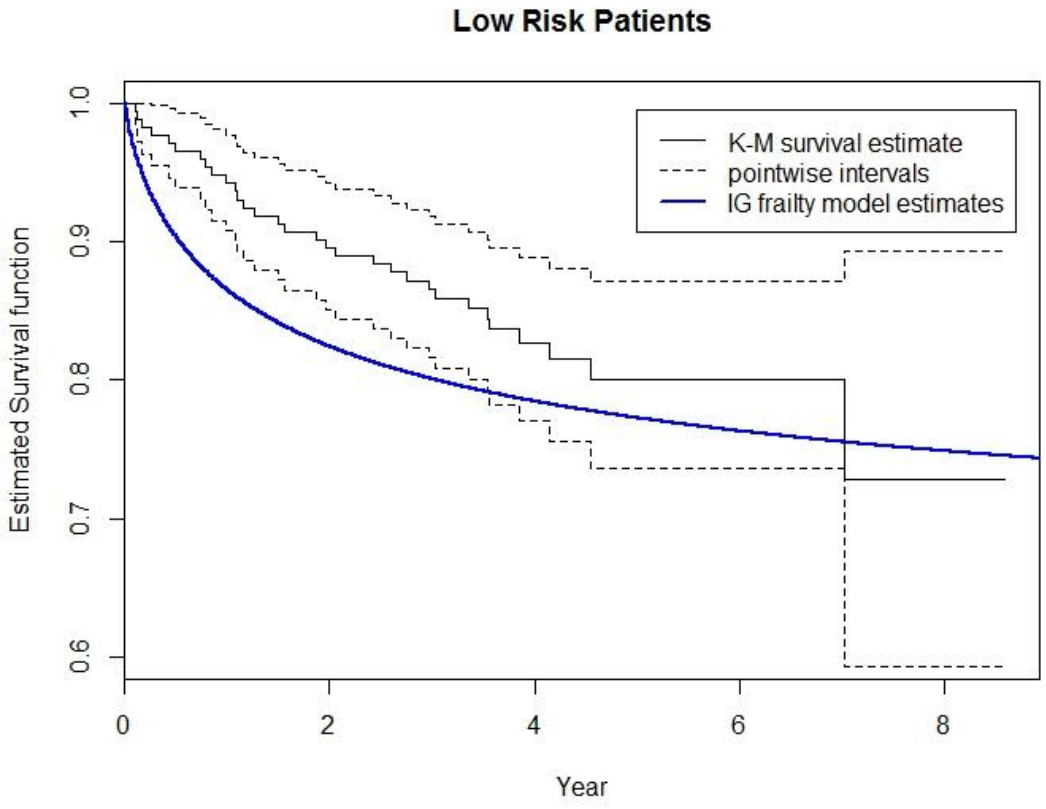


Διάγραμμα 4.14: *Γραφική απεικόνιση των εκτιμήσεων των συναρτήσεων Επιβίωσης για Low Risk και High Risk ασθενείς, που προκύπτουν προσαρμόζοντας το μοντέλο Inverse Gaussian με Λογαριθμολογιστική κατανομή για την βασική συνάρτηση κινδύνου.*

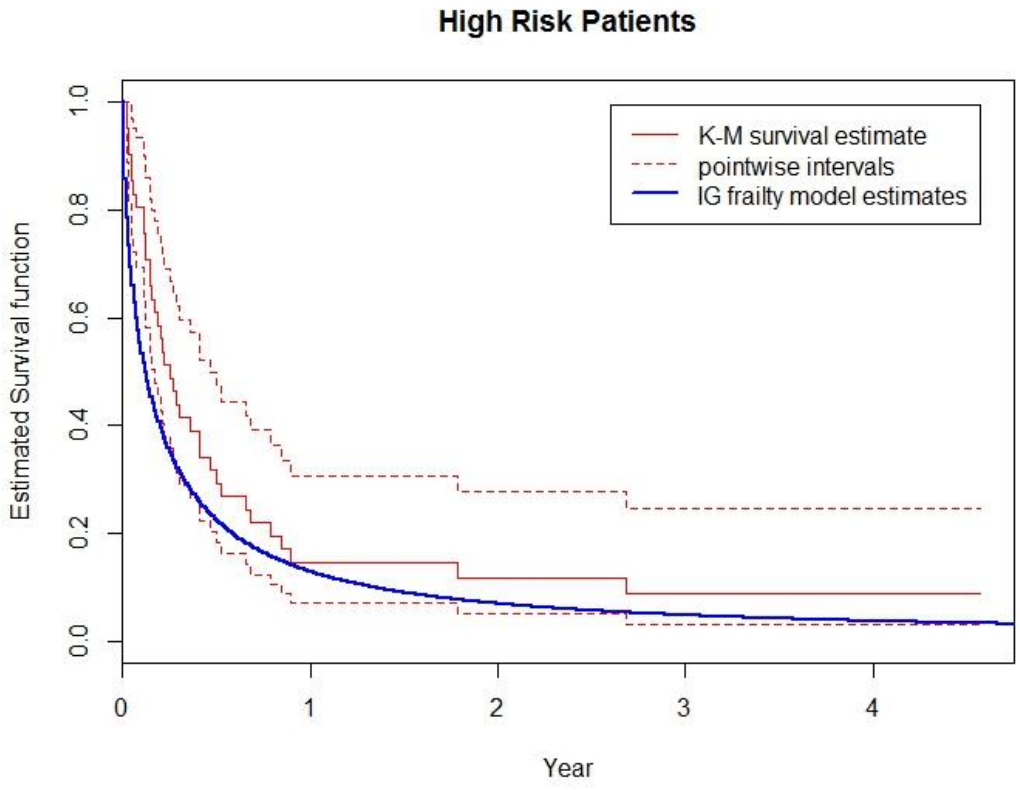
Συγκρίνοντας τις εκτιμήσεις για την συνάρτηση επιβίωσης και τις Kaplan-Meier εκτιμήσεις για Low Risk ασθενείς και High Risk ασθενείς λαμβάνουμε τα Διαγράμματα 4.15 και 4.16:

```
t <- seq(0, 9, 0.01)
plot(fit0, col="black", xlab="Year", ylab="Estimated Survival function", lwd="1", ylim=c(
0.6,1),main="Low Risk Patients")
lines(t, s0.invg, col="blue", lwd=2, type="s")
legend(5, 0.9955, legend=c("K-M survival estimate","pointwise intervals","IG frailty mode
l estimates"), lty=c(1,2,1),lwd=c(1,1,2), col=c("black","black","blue"))

plot(fit1, col="red", xlab="Year", ylab="Estimated Survival function", lwd="1", ylim=c(0,
1),main="High Risk Patients")
lines(t, s1.invg, col="blue", lwd=2, type="s")
```

```
legend(2.6, 0.9955, legend=c("K-M survival estimate","pointwise intervals","IG frailty mo
del estimates"), lty=c(1,2,1),lwd=c(1,1,2), col=c("red","red","blue"))
```



Διάγραμμα 4.15: *Γραφική απεικόνιση των εκτιμήσεων του χρόνου επιβίωσης για Low Risk ασθενείς, χρησιμοποιώντας την εκτιμήτρια Kaplan-Meier, καθώς και διαστήματα εμπιστοσύνης αυτών, καθώς και επεικόνιση των εκτιμήσεων της συνάρτησης επιβίωσης, όπως προκύπτει από το μοντέλο Inverse Gaussian με Λογαριθμολογιστική κατανομή για την βασική συνάρτηση κινδύνου.*



Διάγραμμα 4.16: *Γραφική απεικόνιση των εκτιμήσεων του χρόνου επιβίωσης για High Risk ασθενείς, χρησιμοποιώντας την εκτιμήτρια Kaplan-Meier, καθώς και διαστήματα εμπιστοσύνης αυτών, καθώς και επεικόνιση των εκτιμήσεων της συνάρτησης επιβίωσης, όπως προκύπτει από το μοντέλο Inverse Gaussian με Λογαριθμολογιστική κατανομή για την βασική συνάρτηση κινδύνου.*

- **Ημι παραμετρικα μοντέλα**

Για την προσαρμογή ενός ημι-παραμετρικού μοντέλου θα χρησιμοποιήσουμε την συνάρτηση frailtyPen του πακέτου frailtypack της R. Η εν λόγω συνάρτηση προσαρμόζει ένα Γάμμα από κοινού μοντέλο ευπάθειας εφαρμόζοντας την μέθοδο ποινικοποιημένης μερικής πιθανοφάνειας.

```
install.packages("frailtypack")
library(frailtypack)
fit_semipar<-frailtyPenal(Surv(Survival_Time,Status)~cluster(id)+Age1 + LDH + Performance
_Status + ENS1 + Stage1,data = lymphoma_data_new, n.knots=10,kappa = 1)

> fit_semipar
Call:
frailtyPenal(formula = Surv(Survival_Time, Status) ~ cluster(id) +
    Age1 + LDH + Performance_Status + ENS1 + Stage1, data = lymphoma_data_new,
    n.knots = 10, kappa = 1)


  Shared Gamma Frailty model parameter estimates
  using a Penalized Likelihood on the hazard function
```

```
                        coef exp(coef) SE coef (H) SE coef (HIH)        z          p
Age11                2.925025  18.63468   0.2515477     0.2515477 11.62811 0.0000e+00
LDH1                 0.606766   1.83449   0.0896835     0.0896835  6.76564 1.3272e-11
Performance_Status1  0.570098   1.76844   0.0920952     0.0920952  6.19031 6.0045e-10
ENS11                0.290029   1.33647   0.0900470     0.0900470  3.22087 1.2780e-03
Stage11              0.582692   1.79085   0.1050223     0.1050223  5.54827 2.8851e-08

    Frailty parameter, Theta: 0.957014 (SE (H): 0.310801 ) p = 0.0010378

      penalized marginal log-likelihood = -1500.18
      Convergence criteria:
      parameters = 4.17e-06 likelihood = 8.71e-05 gradient = 6.64e-09

      LCV = the approximate likelihood cross-validation criterion
            in the semi parametrical case     = 1.09612

      n= 1385
      n events= 627  n groups= 19
      number of iterations:  9

      Exact number of knots used:  10
      Value of the smoothing parameter:  1, DoF:  10.04
```

Για να συγκρίνουμε το Γάμμα ημιπαραμετρικό μοντέλο ευπάθειας και το Inverse-Gaussian παραμετρικό μοντέλο ευπάθειας με βασική συνάρτηση κινδύνου την Λογαριθμολογιστική κατανομή θα χρησιμοποιήσουμε το κριτήριο LCV (Likelihood Cross validation criterion).

Το κριτήριο LCV δίνεται από την σχέση:

$$LCV = \frac{1}{n}\left(tr\left({H_{pl}}^{-1}H_l\right) - l(\cdot)\right),$$

όπου H_{pl} είναι μείον ο Εσσιανός πίνακας της ποινικοποιημένης πιθανοφάνειας σε λογαριθμική κλίμακα, H_l είναι μείον ο Εσσιανός πίνακας της ολικής πιθανοφάνειας σε λογαριθμική κλίμακα και $l(\cdot)$ είναι η μεγιστοποιημένη τιμή της ολικής πιθανοφάνειας σε λογαριθμική κλίμακα.

Για το Γάμμα ημιπαραμετρικό μοντέλο ευπάθειας το εν λόγω κριτήριο λαμβάνει την τιμή:

$$LCV = 1.09612.$$

Για παραμετρικά μοντέλα το εν λόγω κριτήριο συνδέεται με το κριτήριο AIC χρησιμοποιώντας τη σχέση:

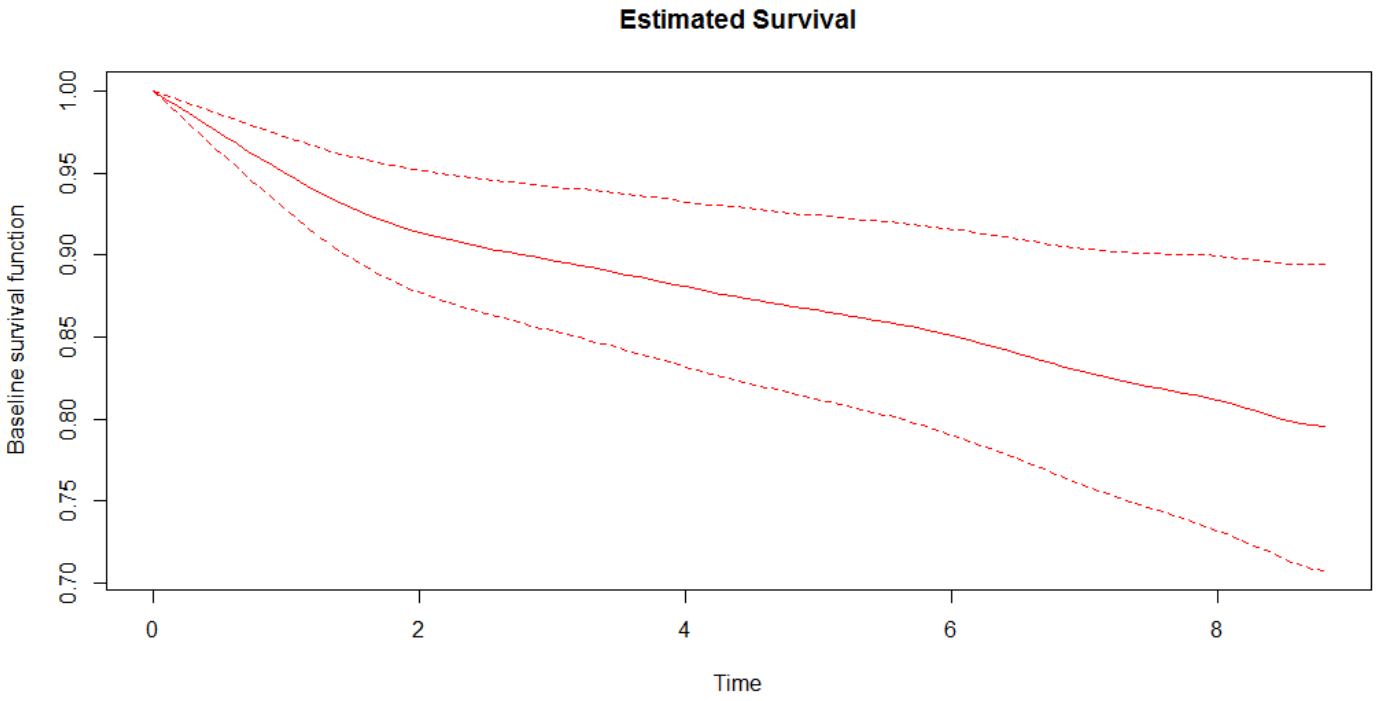
$$LCV \approx \frac{AIC}{2n}.$$

Για το Inverse-Gaussian παραμετρικό μοντέλο ευπάθειας με βασική συνάρτηση κινδύνου την Λογαριθμολογιστική κατανομή το εν λόγω κριτήριο λαμβάνει την τιμή $\text{LCV} = 1.09819$.

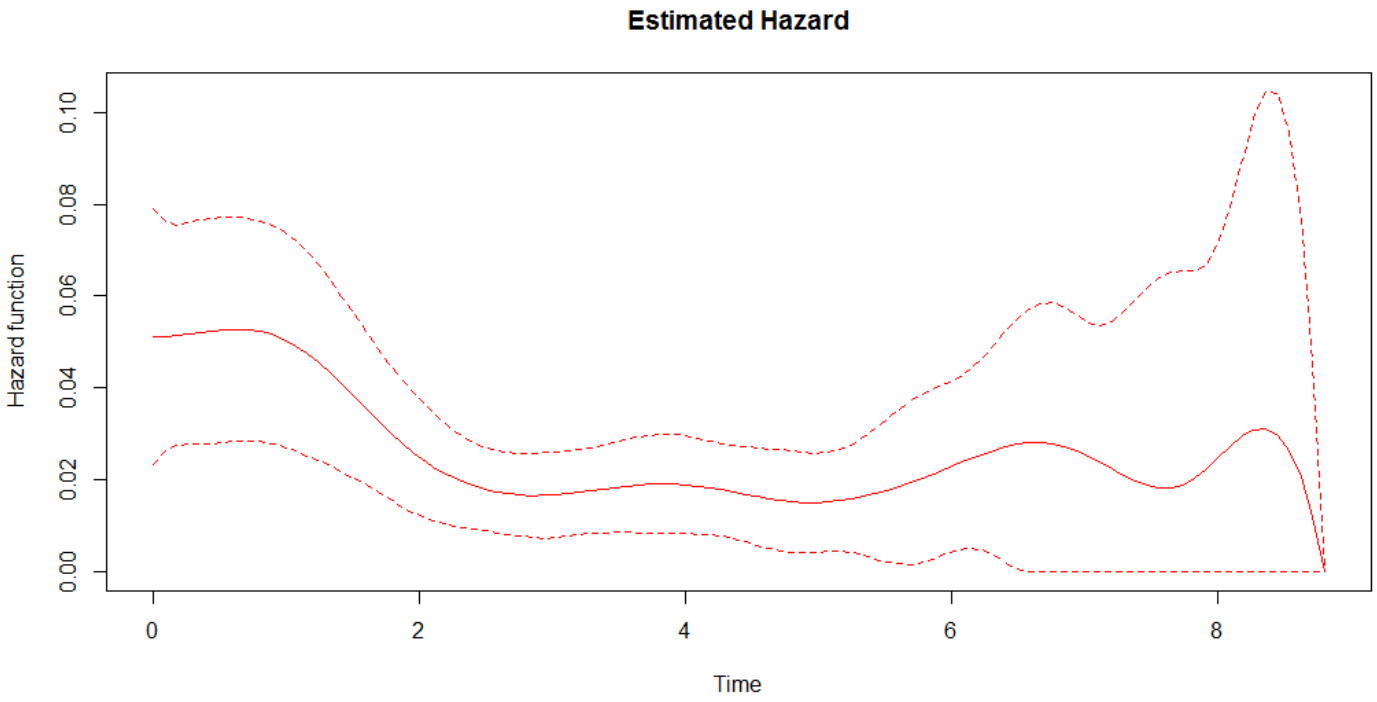
Παρατηρούμε ότι τα δύο μοντέλα έχουν παρόμοια προσαρμογή με βάση το κριτήριο LCV.

Χρησιμοποιώντας τις παρακάτω εντολές στην R, κατασκευάζουμε τα Διαγράμματα 4.17 και 4.18, που απεικονίζουν τις εκτιμώμενες βασικές συναρτήσεις επιβίωσης και κινδύνου για το ημιπαραμετρικό Γάμμα μοντέλο ευπάθειας, που προσαρμόστηκε παραπάνω.

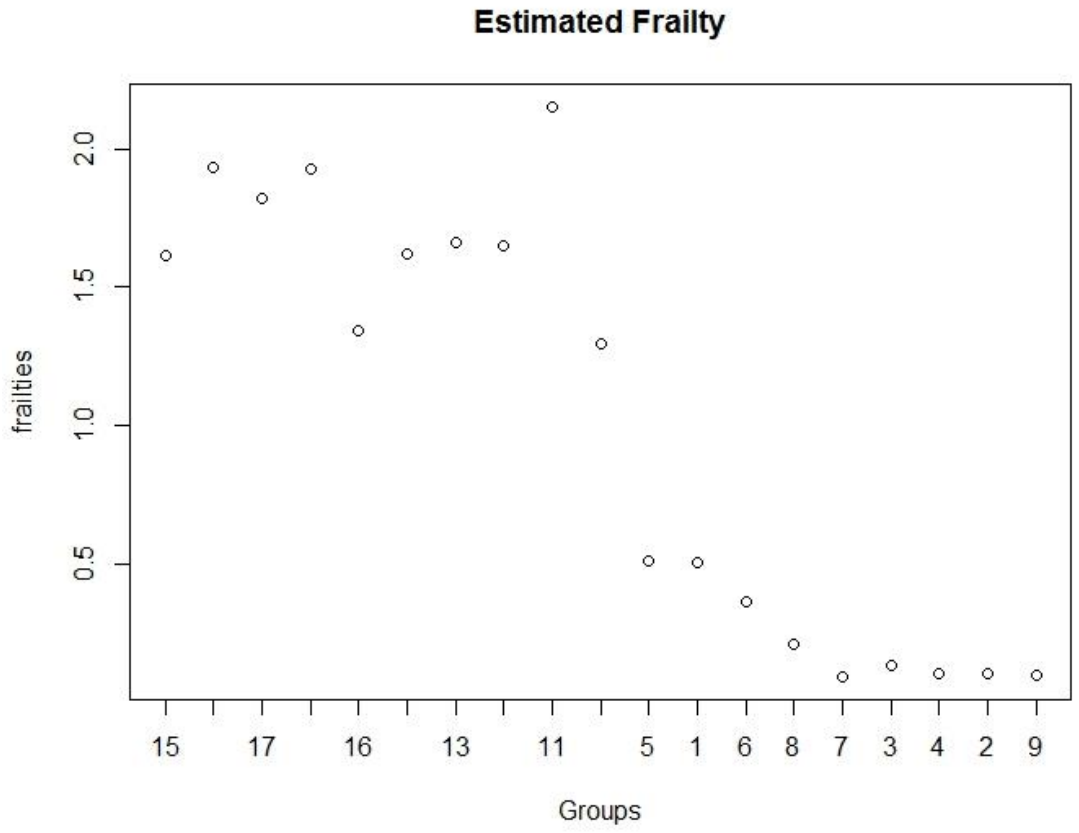
```
plot(fit_semipar,type.plot="Hazard", conf.bands=TRUE, main="Estimated Hazard")
plot(fit_semipar,type.plot="Survival", conf.bands=TRUE, main="Estimated Survival")
```



Διάγραμμα 4.17: *Γραφική απεικόνιση της εκτίμησης του χρόνου επιβίωσης, καθώς και διαστήματα εμπιστοσύνης αυτής όπως προκύπτουν από το ημιπαραμετρικό μοντέλο ευπάθειας Gamma.*



Διάγραμμα 4.18: *Γραφική απεικόνιση της εκτίμησης της συνάρτησης κινδύνου, καθώς και διαστήματα εμπιστοσύνης αυτής όπως προκύπτουν από το ημιπαραμετρικό μοντέλο ευπάθειας Gamma.*



Διάγραμμα 4.19: *Γραφική απεικόνιση των εκτιμώμενων τιμών του παράγοντα ευπάθεια για το ημιπαραμετρικό Γάμμα μοντέλο ευπάθειας,που προσαρμόστηκε στα Lymphoma data.*

Οι εκτιμήσεις της μεταβλητής ευπάθειας, όπως φαίνονται και στο Διάγραμμα 4.19 είναι πολύ διαφορετικές από το μοντέλο ευπάθειας Inverse Gaussian με Λογαριθμολογιστική κατανομή για τη βασική συνάρτηση κινδύνου. Επιπλέον, οι τιμές φαίνεται να διαφέρουν αρκετά ανά ομάδα, το οποίο υποδηλώνει ετερογένεια στον υπό μελέτη πληθυσμό μεταξύ των ομάδων.

Βιβλιογραφία

Androulakis, E., Koukouvinos, C. & Vonta, F., 2012. Estimation and variable selection via frailty models with penalized likelihood. *Statistics in Medicine*.

Anon., n.d. *http://stat.ethz.ch/R-manual/R-devel/library/stats/html/optim.html.* [Ηλεκτρονικό].

Breslow, N., 1974. Discussion of Professor Cox's paper. *Journal Royal Statistical Society,* B(34), pp. 216-17.

Collett, D., 2003. *Modelling Survival Data in Medical Research.* London: Chapman & Hall/CRC.

Cook, R. & Lawless, J., 2007. *The Statistical Analysis of Recurrent Events.* Berlin: Springer.

Deshpande, J. V. & Purohit, S. G., 2005. *Life Time Data: Statistical Models and Methods.* Singapore: World Scientific Publishing Co. Pte. Ltd..

Feller, W., 1971. *An Introduction to Probability Theory and its Applications.* New York: John Wiley and Sons.

Hanagal, D. D., 2011. *Modeling Survival Data Using Frailty Models.* United States of America: Chapman & Hall/CRC.

Hardin, J. & Hilbe, J., 2007. *Generalized Linear Models and Extensions.* Second Edition επιμ. s.l.:Stata Press.

Hosmer, D. W., Lemeshow, S. & May, S., 2008. *Applied Survival Analysis.* Hoboken, New Jersey: John Wiley & Sons.

Hougaard, P., 2000. *Analysis of Multivariate Survival Data.* New York: Springer.

Janssen, P. & Duchateau, L., 2008. *The Frailty Model.* New York, USA: Springer.

Kalbfleisch, J. D. & Prentice, R. L., 1980. *The Statistical Analysis of Failure Time Data.* Second επιμ. New York: John Wiley & Sons.

Kleinbaum, D. & Klein, M., 2005. *Survival Analysis.* 2nd Edition επιμ. United States of America: Springer.

Klein, J. P. & Moeschberger, M. L., 2003. *SURVIVAL ANALYSIS Techniques for Censored and Truncated Data.* United States of America: Springer-Verlag New York.

Lee, E. T. & Wenyu Wang, J., 2003. *Statistical Methods for Survival Data Analysis.* 3rd Edition επιμ. United States of America: Springer.

Loprinzi, C. και συν., 1994. Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. *Journal of Clinical Oncology*, pp. 12(3):601-7.

Marshall, A. & Olkin, I., 2007. *Life Distributions.* Berlin: Springer.

Martinussen, T. & Scheike, T., 2006. *Dynamic Regression Models for Survival.* Berlin: Springer.

Parzen, M. & Harrington, D., n.d. Residual Plots for Detecting Covariate Importance and Nonlinearity in Regression Models with Censored Data. *Harvard School of Public Health, Boston, MA 02115 and Dana Farber Cancer Institute & Emory University, Atlanta, GA 30322.*

Rondeau Virginie, Gonzalez Juan R., Mazroui Yassin, Mauguen Audrey, Krol Agnieszka, Diakite Amadou, Laurent Alexandre, Lopez Myriam;, 2016. *frailtypack: General Frailty Models Using a Semi-Parametrical Penalized Likelihood*

Estimation or a Parametrical Estimation. R-Package Version: 2.9.4. [Ηλεκτρονικό]
Available at: https://cran.r-project.org/web/packages/frailtypack/index.html

Rondeau, V. και συν., 2015. *Package 'frailtypack'.* [Ηλεκτρονικό]
Available at: https://cran.r-project.org/web/packages/frailtypack/frailtypack.pdf
[Πρόσβαση January 2016].

Rondeau, V., Mazroui, Y. & Gonzalez, J. R., 2012. frailtypack: An R Package for the Analysis of Correlated Survival Data with Frailty Models Using Penalized Likelihood Estimation or Parametrical Estimation. *Journal Of Statistical Software,* Volume 47(Issue 4).

Rotolo Federico, Munda Marco, 2015. *parfm: Parametric Frailty Models. R Package Version 2.5.10.*. [Ηλεκτρονικό]
Available at: https://cran.r-project.org/web/packages/parfm/index.html

Rotolo, F., Marco, M. & Legrand, C., 2012. parfm: Parametric Frailty Models in R. *Journal of Statistical Software, http://www.jstatsoft.org/,* November.51(11).

Schwarz, G., 1978. Estimating the dimension of a model. *Annals of Statistics,* Τόμος 6, pp. 461-464.

Selvin, S., 2008. *Survival Analysis for Epidemiological and Medical Research.* New York: Cambridge University Press.

Shipp, M., Harrington, D. & Anderson, J., 1993. Development of a Predictive Model for Aggressive lymphoma. *The International Non-Hodgkin's Lymphoma Prognostic Factors Project,* Τόμος New England Journal of Medicine, in press, pp. 987-994.

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

Therneau, T. M. & Lumley , T., 2016. *Package 'survival'.* [Ηλεκτρονικό]
Available at: https://cran.r-project.org/web/packages/survival/survival.pdf
[Πρόσβαση September 2016].

Vonta, F., 1996. Efficient estimation in a nonproportional hazards model in survival analysis, *Scand. J. Statist*., 23, 49-61.

Wienke, A., 2011. *Frailty Models in Survival Analysis.* United States of America: Chapman & Hall/CRC.