ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΕ ΤΙΤΛΟ:

ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΜΟΝΤΕΛΩΝ ΕΥΠΑΘΕΙΑΣ ΜΕΣΩ ΓΕΝΙΚΕΥΜΕΝΗΣ ΣΥΝΑΡΤΗΣΗΣ ΠΙΘΑΝΟΦΑΝΕΙΑΣ

Επιβλέπουσα: Φιλία Βόντα, Καθηγήτρια Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2023.

Η παρούσα σελίδα αφέθηκε σκοπίμως κενή

Η παρούσα διπλωματική εκπονήθηκε
στα πλαίσια των σπουδών για την απόκτηση
του μεταπτυχιακού διπλώματος ειδίκευσης στην
Μαθηματική Προτυποποίηση Σύγχρονων Συστημάτων
με εξειδίκευση στα
Μαθηματικά της Επιστήμης των Δεδομένων.

ΠΕΡΙΛΗΨΗ

Η ανάλυση δεδομένων χρόνου ζωής (ή χρόνου μέχρι κάποιο γεγονός ενδιαφέροντος) έχει αναπτυχθεί ραγδαία τα τελευταία χρόνια λόγω της πληθώρας εφαρμογών που βρίσκει στην επιδημιολογία, στην βιολογία, στην δημογραφία, στις οικονομικές επιστήμες, στην μηχανική και στην ιατρική. Για αυτόν ακριβώς τον σκοπό έχει αναπτυχθεί εκτενώς και ο κλάδος της στατιστικής, γνωστός και ως «Ανάλυση επιβίωσης».

Στην παρούσα εργασία θα ασχοληθούμε με τον κλάδο της ανάλυσης επιβίωσης, και πιο συγκεκριμένα, με τα μοντέλα ευπάθειας. Τα μοντέλα ευπάθειας «Frailty Models» αποτελούν γενίκευση του μοντέλου του Cox. Όταν θεωρούμε ότι υπάρχει μια συμμεταβλητή που είτε είναι άγνωστη εντελώς στον ερευνητή είτε γνωρίζουμε ότι υπάρχει αλλά δεν έχουμε δεδομένα για αυτή, την θεωρούμε σαν μία τυχαία μεταβλητή που μπορεί να ενταχθεί στο μοντέλο του Cox η οποία ακολουθεί δυνητικά διάφορες κατανομές. Έτσι γεννιέται μια ολόκληρη κλάση μοντέλων ευπάθειας.

Η έννοια της ευπάθειας, παρέχει έναν βολικό τρόπο να περιγραφεί παρατηρούμενη ετερογένεια η οποία δεν είναι γνωστό που οφείλεται, καθώς και να περιγραφούν συσχετίσεις στα μοντέλα επιβίωσης. Η ευπάθεια είναι ένας άγνωστος παράγοντας ο οποίος τροποποιεί την συνάρτηση κινδύνου ενός ατόμου η μιας κλάσης σχετιζόμενων ατόμων από αυτόν που ορίζεται στο μοντέλο του Cox. Πιο συγκεκριμένα, ένα μοντέλο ευπάθειας, είναι ένα μοντέλο άγνωστων επιδράσεων για δεδομένα «χρόνου μέχρι κάποιο γεγονός», όπου μια άγνωστη επίδραση επιδρά πολλαπλασιαστικά στον αρχικό κίνδυνο.

Πιο συγκεκριμένα, στην παρούσα εργασία, επιχειρήσαμε να προτυποποιήσουμε ένα μοντέλο ευπάθειας σε ένα σετ δεδομένων ασθενών με λέμφωμα Non-Hodgkins. Για την πραγματοποίηση της ανάλυσης, γράφτηκε κώδικας σε R με δυναμικό τρόπο ώστε ο εκάστοτε φοιτητής/ ερευνητής να μπορεί να εισάγει τα δικά του δεδομένα και να πραγματοποιήσει αντίστοιχες αναλύσεις προχωρώντας ένα βήμα παραπέρα το πολλά υποσχόμενο αυτό πεδίο. Το πρόγραμμα της R που γράφθηκε μπορεί να αποτελέσει ένα πρώτο σκαλοπάτι για τον οποιοδήποτε που θέλει να ασχοληθεί με αυτό το αντικείμενο.

Η ανάλυση βασίστηκε σε γενικευμένη συνάρτηση πιθανοφάνειας που ορίζεται έτσι ώστε να δίνει ένα γενικό τύπο πιθανοφάνειας που ισχύει για όλη την κλάση των μοντέλων

ευπάθειας και μόνη αλλαγή που προϋποθέτει είναι η αλλαγή του μετασχηματισμού Laplace της κατανομής της μεταβλητής της ευπάθειας. Τα αποτελέσματα του αλγορίθμου που υιοθετήσαμε για την εκτίμηση των άγνωστων παραμέτρων μέσω της γενικευμένης συνάρτησης πιθανοφάνειας, βρήκαμε ότι είναι ταυτόσημα με τα αποτελέσματα έτοιμης βιβλιοθήκης της R, της "parfm" που εφαρμόζει παρόμοια τακτική. Η μέθοδός μας αποτελεί μια εναλλακτική μέθοδο του αλγορίθμου E-M για παραμετρικά μοντέλα ευπάθειας.

Στο πρώτο κεφάλαιο κάνουμε μία εισαγωγή για το θέμα της εργασίας, στο δεύτερο κάνουμε μια βιβλιογραφική ανασκόπηση. Στο τρίτο περιγράφουμε την μεθοδολογία και το θεωρητικό μαθηματικό υπόβαθρο που απαιτείται για την εργασία. Στο τέταρτο κεφάλαιο περιγράφουμε το σετ δεδομένων που χρησιμοποιήσαμε και τα αποτελέσματα της στατιστικής ανάλυσης και στο πέμπτο κεφάλαιο γίνεται μια σύγκριση των αποτελεσμάτων που προέκυψαν από τον προτεινόμενο αλγορίθμο και των αποτελεσμάτων από τη σχετική έτοιμη βιβλιοθήκη της R.

Λέξεις κλειδιά:

"Μοντέλα ευπάθειας", "Frailty models", "Gamma", "Inverse Gaussian", "Γενικευμένη συνάρτηση πιθανοφάνειας"

ABSTRACT

The analysis of lifetime (time-to-event or duration) data has developed rapidly in recent years due to the multitude of its applications in epidemiology, biology, demography, economics, engineering, and medicine. For this very purpose, the area of statistics, also known as *"Survival Analysis"*, has been extensively developed.

In this thesis we will deal with the area of survival analysis, and more specifically, with frailty models. Frailty Models are a natural generalization of the Cox model. When we suspect that there is a covariate that is either completely unknown or we know it exists but have no measurable data for it, we can treat it as a random variable that is introduced in the Cox model that could potentially follow various distributions. Thus, a class of frailty models is born.

The concept of frailty provides a convenient way to introduce some heterogeneity observed in the data that is unexplained as well as correlations between survival models. The frailty term is an unknown random factor that modifies the hazard function of an individual or a class of related individuals as compared to the hazard function of the Cox model. More specifically, a frailty model is a model of unknown effects for "time-to-event" data, where an unknown effect multiplicatively affects the initial hazard.

More specifically, in this work, we attempted to fit a frailty model on a data set related to patients with Non-Hodgkins lymphoma. To perform the analysis, a code was written in R in a dynamic way so that each student/researcher can enter his/her own data and perform corresponding analyses advancing this promising field. The R program written can be a first step for anyone who wants to deal with this topic.

The analysis was based on a generalized likelihood function defined to give a general likelihood formula that applies to the whole class of frailty models with the sole change that is required to be the change of the Laplace transform of the distribution of the frailty variable. The results of the algorithm we proposed to estimate the unknown parameters through the generalized likelihood function, were found to be identical to the results of a ready R library, “parfm” that uses a similar tactic. Our method is an alternative to the E-M algorithm for parametric frailty models.

In the first chapter we make an introduction about the topic of the thesis, in the second we make a bibliographic review. In the third chapter we describe the methodology

and the theoretical mathematical background required for our approach. In the fourth chapter we describe the data set that was used and the results of the statistical analysis, and in the fifth chapter a comparison is made between the results obtained from the proposed algorithm and the results obtained by the relevant ready library of R.

Keywords:

"Frailty models", "Gamma", "Inverse Gaussian", "Generalized Likelihood Function"

Ευχαριστίες

Η εκπόνηση της παρούσας εργασίας πραγματοποιήθηκε υπό την επίβλεψη της Καθηγήτριας του Ε.Μ.Π. Φιλίας Βόντα την οποία θα ήθελα να ευχαριστήσω θερμά για τις εκτενείς συζητήσεις μέσα από τις οποίες με καθοδήγησε, την κατανόηση που έδειξε και γενικότερα για την άρτια συνεργασία.

Επιπλέον, θα ήθελα να ευχαριστήσω την ομότιμη Καθηγήτρια κ. Χρυσηίδα Καρώνη και τον ομότιμο Καθηγητή κ. Βασίλειο Παπανικολάου για την τιμή που μου έκαναν να συμμετάσχουν στην τριμελή εξεταστική επιτροπή.

Γενικότερα θα ήθελα να ευχαριστήσω όλους όσους συναναστράφηκα κατά την πορεία του μεταπτυχιακού μου, από την γραμματεία, τους συμφοιτητές μου και τους καθηγητές καθώς όλοι, ανεξαιρέτως, συνεισέφεραν στην απόκτηση πληθώρας βιωμάτων μου που με καθόρισαν στις επαγγελματικές και όχι μόνο, επιλογές μου.

Ήταν ένας ενδιαφέρον κύκλος, ένας κύκλος που αποτελεί κρίκο μιας αλυσίδας η οποία συνεχίζεται...

Κουταλώνης Ιωάννης

Αθήνα, 2023

Περιεχόμενα

Contents

ΚΕΦΑΛΑΙΟ 1 - ΕΙΣΑΓΩΓΗ

1.1. Περί Ανάλυσης Επιβίωσης

Το παρόν κεφάλαιο αποτελεί μία εισαγωγή στις θεμελιώδεις έννοιες, πάνω στις οποίες αναπτύχθηκε η στατιστική μεθοδολογία της Ανάλυσης Επιβίωσης. Η δημιουργία ξεχωριστής στατιστικής θεωρίας για την ανάλυση των δεδομένων επιβίωσης αποτελεί αξιοσημείωτο γεγονός. Η Ανάλυση Επιβίωσης διαφέρει σημαντικά από τις τεχνικές που χρησιμοποιούνται σε άλλους κλάδους της Στατιστικής.

Πώς ένας παράγοντας κινδύνου επηρεάζει τον χρόνο έκβασης ενός γεγονότος?

Η απάντηση αυτής της ερώτησης αποτελεί την βάση της ανάλυσης επιβίωσης και αξιοπιστίας. Η ανάλυση επιβίωσης αποτελεί ξεχωριστή στατιστική περιοχή καθώς αναλύει δεδομένα που αφορούν τον χρόνο μέχρι να πραγματοποιηθεί ένα συγκεκριμένο συμβάν όπως η μηχανική αστοχία ή στην περίπτωση της ανάλυσης επιβίωσης, ο θάνατος. Γενικά, σαν “συμβάν” μπορεί να θεωρηθεί οτιδήποτε μεταβάλει την κατάσταση του υπό μελέτη αντικειμένου/ φαινομένου.

Πιο συγκεκριμένα, η ανάλυση επιβίωσης προσπαθεί να απαντήσει ερωτήματα όπως “ποιο είναι το ποσοστό του πληθυσμού που θα επιβιώσουν μετά από ένα ορισμένο χρονικό διάστημα?” ή “είναι δυνατό πολλαπλές αιτίες αποτυχίας η θανάτου να ληφθούν υπόψιν?” ή “πως ιδιαίτερες συνθήκες επηρεάζουν τις πιθανότητες επιβίωσης ή αποτυχίας?”.

Ως εκ τούτου, σημαντική έννοια για την μελέτη αυτή, αποτελεί ο επονομαζόμενος «χρόνος επιβίωσης» (Survival Time). Η έννοια του χρόνου επιβίωσης αναφέρεται σε μία τυχαία μεταβλητή που μετράει τον χρόνο που μεσολαβεί από την έναρξη της παρακολούθησης ενός ατόμου ή ενός δείγματος μέχρι την έκβαση του γεγονότος. Ο χρόνος επιβίωσης, καταλαβαίνουμε και διαισθητικά ότι οφείλει να είναι πάντα θετικός. Επίσης πρέπει να σημειώσουμε ότι τα δεδομένα χρόνου επιβίωσης έχουν ένα πολύ ιδιαίτερο χαρακτηριστικό. Τα δεδομένα μπορεί να είναι λογοκριμένα (Censored).

Για αυτόν τον λόγο, τα δεδομένα διάρκειας ζωής δεν υπόκεινται σε ανάλυση με τις συνηθισμένες στατιστικές μεθόδους, αφενός γιατί μπορεί να είναι λογοκριμένα (censored) ή/και αποκομμένα (truncated) και αφ’ ετέρου γιατί δεν έχουν συμμετρικές κατανομές. Ένα τυπικό ιστόγραμμα τέτοιων δεδομένων θα έδειχνε μια μακρύτερη ουρά στο δεξί του μέρος

όπου θα περιέχονται και οι περισσότερες παρατηρήσεις. Έτσι, δεν θα ήταν λογικό να θεωρήσουμε ότι τα δεδομένα ακολουθούν κανονική κατανομή.

Το κύριο ενδιαφέρον μας για τον χρόνο επιβίωσης είναι η εύρεση της κατανομής που ακολουθεί, η σύγκριση του χρόνου επιβίωσης μιας ομάδας σε σχέση με μία άλλη όπως και η μοντελοποίηση της σχέσης του χρόνου επιβίωσης σε σχέση με άλλες μεταβλητές.

Τα κλασσικά μοντέλα επιβίωσης ασχολούνται με την κλασσική περίπτωση ανεξάρτητων και ομοιόμορφα κατανεμημένων δεδομένων. Αυτό βασίζεται στην παραδοχή ότι ο υπό μελέτη πληθυσμός είναι ομοιογενής μέχρι έναν συγκεκριμένο αριθμό συμμεταβλητών. Ωστόσο, στην πραγματικότητα, εύκολα κάποιος μπορεί να παρατηρήσει ότι ο υπό μελέτη πληθυσμός χαρακτηρίζεται από μεγάλη ετερογένεια σε σχέση με κάποιο ή κάποια χαρακτηριστικά του πληθυσμού τα οποία ή είναι άγνωστα ή δεν είναι εύκολα μετρήσιμα όπως άλλες επεξηγηματικές μεταβλητές.

Αυτή η ετερογένεια συχνά αναφέρεται και ως μεταβλητότητα. Αυτό που θα μας απασχολήσει λοιπόν στα μοντέλα ευπάθειας είναι η μη παρατηρήσιμη ή άμεσα μετρήσιμη μεταβλητότητα στα δεδομένα επιβίωσης. Η ετερογένεια αυτή μπορεί να είναι δύσκολο να εκτιμηθεί αλλά είναι αρκετά σημαντική για να αγνοηθεί. Τις τελευταίες δεκαετίες, έχουν γίνει πολλές μελέτες πάνω σε μοντέλα ευπάθειας. Η βασική ιδέα των μοντέλων αυτών είναι ότι τα άτομα έχουν διαφορετικές ευπάθειες και εκείνο με την μεγαλύτερη ευπάθεια καταλήγει (πεθαίνει) νωρίτερα από εκείνο με την μικρότερη ευπάθεια.

1.2. Μοντέλα Ευπάθειας και Σκοπός της Παρούσας Εργασίας

Στην συγκεκριμένη εργασία θα αναλύσουμε με την μέθοδο των μοντέλων ευπάθειας ένα σετ δεδομένων ασθενών που πάσχουν από λέμφωμα Non-Hodgkins. Τα δεδομένα προέρχονται από το διεθνές project προγνωστικών παραγόντων για το λέμφωμα Non-Hodgkins (Ship et al, 1993).

Το συγκεκριμένο σετ δεδομένων έχει αναλυθεί από διάφορους επιστήμονες προκειμένου να προσπαθήσουν να εκτιμήσουν ποιος παράγοντας ή συνδυασμός παραγόντων θα μπορούσε να προβλέψει έστω και μερικώς την έκβαση του κάθε ασθενούς. Κατηγοριοποιώντας λοιπόν τους ασθενείς με ανάλογους βαθμούς “ρίσκου”, οι γιατροί θα μπορέσουν να προτείνουν την κατάλληλη θεραπεία για τον κάθε ασθενή.

Ο σκοπός μας εδώ είναι να ορίσουμε μία γενικευμένη συνάρτηση πιθανοφάνειας η οποία χειρίζεται όλη την κλάση των μοντέλων ευπάθειας με μικρές αλλαγές ως προς τον μετασχηματισμό Laplace της κατανομής της ευπάθειας που θα επιλέξουμε να χρησιμοποιήσουμε. Ο μετασχηματισμός Laplace της κατανομής της ευπάθειας εμπλέκεται άμεσα στον τύπο της γενικευμένης συνάρτησης πιθανοφάνειας.

Για τον υπολογισμό των παραμέτρων ενδιαφέροντος που υπεισέρχονται στα μοντέλα ευπάθειας έχει αναπτυχθεί κώδικας στη γλώσσα R, ο οποίος μπορεί να χρησιμοποιηθεί πέρα από το σετ πραγματικών δεδομένων που θεωρήσαμε στην παρούσα εργασία. Σκοπός μας επίσης είναι η σύγκριση των αποτελεσμάτων μας με εδραιωμένες μεθόδους εκτίμησης που αφορούν τα μοντέλα ευπάθειας μέσω της R.

1.3. Οδηγός Κεφαλαίων

Στο Κεφάλαιο 2 γίνεται μία βιβλιογραφική ανασκόπηση της περιοχής των μοντέλων που χρησιμοποιούνται στην ανάλυση επιβίωσης, ξεκινώντας από το μοντέλο του Cox (1972) και προχωρώντας σε διάφορες γενικεύσεις του όπως τα μοντέλα ευπάθειας, τα οποία θα μας απασχολήσουν στην παρούσα εργασία.

Στο Κεφάλαιο 3 θεμελιώνεται αναλυτικά το θεωρητικό υπόβαθρο πάνω στο οποίο βασίζεται η θεωρία που χρησιμοποιούμε στην παρούσα εργασία, ξεκινώντας από το πιο βασικό μοντέλο στην ανάλυση επιβίωσης, το μοντέλο του Cox (Cox 1972) και συνεχίζοντας με γενικεύσεις του μοντέλου του Cox και πιο συγκεκριμένα στα μοντέλα ευπάθειας τα οποία αποτελούν και το κύριο αντικείμενο αυτής της εργασίας. Στο ίδιο κεφάλαιο περιγράφεται αναλυτικά η μεθοδολογία που χρησιμοποιούμε για την εκτίμηση των παραμέτρων που εμπλέκονται στη γενικευμένη συνάρτηση πιθανοφάνειας για μια ολόκληρη κλάση μοντέλων ευπάθειας. Για τον σκοπό αυτό, επικεντρωνόμαστε σε δύο συγκεκριμένες κατανομές ευπάθειας, την Gamma και Inverse Gaussian γιατί αυτές έχουν χρησιμοποιηθεί στο κεφάλαιο 4 στην ανάλυση πραγματικών δεδομένων. Επίσης γίνεται αναφορά στον αλγόριθμο που δημιουργήθηκε στα πλαίσια της εργασίας για να αναλυθούν τα δεδομένα μέσω παραμετρικών μοντέλων ευπάθειας Gamma και Inverse Gaussian.

Στο Κεφάλαιο 4 περιγράφονται τα δεδομένα που χρησιμοποιήθηκαν σχετικά με το λέμφωμα Non-Hodgkins, τα αποτελέσματα του αλγορίθμου για τα παραμετρικά μοντέλα ευπάθειας Gamma και Inverse Gaussian και του μοντέλου του Cox όπως επίσης και τα αντίστοιχα αποτελέσματα από την βιβλιοθήκη “parfm” της R, που εφαρμόζει επίσης παραμετρικά μοντέλα ευπάθειας Gamma και Inverse Gaussian.

Στο Κεφάλαιο 5 γίνεται μία σύγκριση των αποτελεσμάτων του προτεινόμενου αλγορίθμου για τα παραμετρικά μοντέλα ευπάθειας Cox, Gamma και Inverse Gaussian, όπως επίσης και σύγκριση των αποτελεσμάτων του προτεινόμενου αλγορίθμου με τα αντίστοιχα αποτελέσματα της έτοιμης βιβλιοθήκης της R, “parfm” που εφαρμόζει παρόμοιες μεθόδους.

ΚΕΦΑΛΑΙΟ 2 – ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

Βασικός σκοπός πολλών ιατρικών μελετών είναι η προτυποποίηση ενός μοντέλου επιβίωσης. Το μοντέλο του Cox (1972) είναι πολύ διάσημο για τέτοιες αναλύσεις αλλά είναι περιορισμένες οι δυνατότητές του καθώς αρκετά προβλήματα εξαρτώνται από πλήθος παραγόντων, σε πολλούς από τους οποίους οι επιστήμονες δεν έχουν πρόσβαση. Την λύση σε αυτό το πρόβλημα έρχεται να δώσει η έννοια της ευπάθειας.

Αυτή η έννοια πάει πίσω στο βιβλίο των Greenwood and Yule (1920), στην εργασία των Vaupel et al. (1979) ενώ το πρώτο μόνο-μεταβλητό (univariate) μοντέλο ευπάθειας προτάθηκε από τον Beard (1959). Η έννοια της ευπάθειας αυτή καθ' αυτή, πρωτοεισήχθη από τους Vaupel et al. (1979) για την μελέτη ανεξάρτητων παρατηρήσεων (univariate case) και οι πρώτες εφαρμογές σε τέτοια προβλήματα ανάλυσης επιβίωσης πραγματοποιήθηκαν από τον Clayton (1978).

Οι κλασσικές μέθοδοι στην ανάλυση επιβίωσης υιοθετούν την άποψη ότι οι υπο μελέτη πληθυσμοί είναι ομοιογενείς, δηλαδή ότι όλα τα άτομα έχουν το ίδιο ρίσκο θνησιμότητας. Μερικές φορές όμως είναι αδύνατο να συμπεριλάβουμε όλους τους πιθανούς παράγοντες επικινδυνότητας, αφενός διότι μπορεί να μην είναι μετρήσιμοι (με τα ήδη υπάρχοντα μέσα) όπως π.χ. η γενετική προδιάθεση, και αφετέρου διότι μπορεί να μην έχουμε πληροφορίες για κάποιο παράγοντα λόγω του υπερβολικού κόστους που χρειάζεται να καταβληθεί για την απόκτησή του. Αυτό είναι πολύ σύνηθες στις ιατρικές και βιολογικές επιστήμες.

Για να αντιμετωπισθεί λοιπόν το πρόβλημα της μη μετρήσιμης ετερογένειας στον πληθυσμό που προκύπτει από μη παρατηρούμενες συμμεταβλητές, πρώτος ο Beard (1959) και έπειτα οι Vaupel et al. (1979) και Lancaster et al. (1979), πρότειναν ένα μοντέλο τυχαίων επιδράσεων για δεδομένα διάρκειας ζωής, ανεξάρτητα ο ένας από τον άλλον. Ο Beard (1959) χρησιμοποίησε τον όρο "Παράγοντας μακροζωίας" (longevity factor) αντί για "ευπάθεια" (Frailty). Ο σκοπός της εισαγωγής της τυχαίας επίδρασης ήταν να βελτιωθεί η προτυποποίηση των μοντέλων θνησιμότητας σε πληθυσμούς. Οι Vaupel et al. (1979) πρωτοεισήγαγαν την έννοια της ευπάθειας στην περιοχή της βιοστατιστικής και την εφάρμοσαν σε δεδομένα θνησιμότητας πληθυσμού.

Ο Lancaster (1979) ασχολήθηκε με το φαινόμενο των χρόνων ανεργίας και εισήγαγε τα μοντέλα ευπάθειας στην βιβλιογραφία της οικονομετρίας, το οποίο είναι γνωστό σαν

“mixed proportional hazards model” (MPH). Οσο η θεωρία αυτή ωρίμαζε στο μυαλό των επιστημόνων, υπήρξαν άτομα που δημιούργησαν τις δικές τους παραλλαγές στα μοντέλα ευπάθειας για να καλύψουν ο κάθε ένας τις δικές του ανάγκες.

Για παράδειγμα, προτάθηκε το αθροιστικό μοντέλο ευπάθειας όπου η ευπάθεια δρα αθροιστικά στην βασική συνάρτηση διακινδύνευσης. Περισσότερες πληροφορίες για αυτό το μοντέλο μπορούν να βρεθούν στα έργα των Rocha (1996), Silva και Amaral-Turkman (2004), και Tomazella et al. (2006).

Μοντέλα ευπάθειας αναλογικών κινδύνων περιγράφονται αναλυτικά στο έργο των Lam et al. (2002), Lam and Lee (2004). Στο έργο των Murphy et al. (1997), περιγράφεται αναλυτικά η σύνδεση μεταξύ των μοντέλων αναλογικών κινδύνων και των μοντέλων ευπάθειας. Τα μοντέλα ευπάθειας επιταχυνόμενης αποτυχίας (Accelerated Failure Time-AFT), αντιμετωπίζονται από τους Anderson and Louis (1995), οι οποίοι χρησιμοποιούν μοντέλο επιβίωσης δύο μεταβλητών με παραμετρικές και μη παραμετρικές κατανομές ευπάθειας. Οι Keiding et al. (1997) δίνουν έμφαση στην επίδραση της ετερογένειας που δημιουργείται από τις συμμεταβλητές που έχουν παραλειφθεί και παρατήρησαν ότι το μοντέλο AFT είναι πιο σταθερό από το μοντέλο αναλογικών κινδύνων υπό την παρουσία ετερογένειας. Οι Klein et al. (1999) θεωρούν ένα κανονικό μοντέλο παλινδρόμησης βασισμένο σε μία log-normal κατανομή ευπάθειας. Ο Pan (2001) προτείνει ένα μοντέλο σχετιζόμενων χρόνων αποτυχίας μοντελοποιώντας τον παράγοντα σφάλματος του AFM με μία προσέγγιση μοντέλων ευπάθειας. Λόγω των προβλημάτων αστάθειας που προκύπτουν από τον αλγόριθμο του Pan, Οι Zhang and Peng (2007) πρότειναν μία διαδικασία ημιπαραμετρικής εκτίμησης. Αυτήν την μεθοδολογία, οι Xu and Zhang (2010) την προχώρησαν ένα βήμα παραπέρα και ανέπτυξαν μία πιο σταθερή διαδικασία εκτίμησης στην ημιπαραμετρική ευπάθεια Gamma κατανομής του AFT Μοντέλου. Οι Lambert et al. (2004) μελέτησαν ένα παραμετρικό μοντέλο AFT με αθροιστικό όρο ευπάθειας. Άλλες ενδιαφέρουσες δημοσιεύσεις σε αυτήν την κατεύθυνση είναι τα έργα των Schnier et al. (2004), Chang (2004) και Komarek et al. (2007).

Ξεφεύγοντας από τις καθιερωμένες εφαρμογές, στην μελέτη του ανθρώπινου εγκεφάλου, οι Jonker et al. (2008) μελέτησαν τις ημικρανίες και πιο συγκεκριμένα την ηλικία πρωτοεμφάνισης ημικρανιών σε συγγενείς. Η ανάλυση έδειξε ότι υπάρχει στατιστικά σημαντική ένδειξη ότι η ηλικία πρωτοεμφάνισης φαίνεται να είναι γενετικά προκαθορισμένη.

Στον τομέα των εξαρτήσεων, στο έργο τους οι Nosyc et al. (2009) μελέτησαν άτομα εθισμένα σε οπιοειδή και πιο συγκεκριμένα τον χρόνο ελεγχόμενης λήψης μεθαδόνης μέχρι να υποχωρήσουν τα συμπτώματα στέρησης. Χρησιμοποιώντας μοντέλα ευπάθειας κατέληξαν σε μερικές σημαντικές συμμεταβλητές που καθορίζουν τον απαιτούμενο χρόνο απεξάρτησης όπως η ηλικία, το κοινωνικό status της γειτονιάς στην οποία διέμεναν και η διαθεσιμότητα των γιατρών της περιοχής. Σε άλλη εργασία, οι Li et al., (2011) μελέτησαν τα μακροπρόθεσμα αποτελέσματα της διακοπής τσιγάρου προκειμένου να εκτιμηθεί το ρίσκο υποτροπής στην επιβλαβή αυτή συνήθεια. Αναφέρουν χαρακτηριστικά ότι το μοντέλο τους δίνει εξαιρετικές εκτιμήσεις.

Στην κοινωνιολογία τώρα, αρκετές μελέτες έχουν γίνει στην θνησιμότητα σε παιδιά κάτω των 5 ετών χρησιμοποιώντας τόσο μοντέλα ευπάθειας όσο και μοντέλα του Cox. Να σημειωθεί ότι η θνησιμότητα κάτω από 5 έτη (Under 5 Mortality -U5M) είναι από τους πιο σημαντικούς δείκτες για την υγειονομική κατάσταση μίας κοινότητας. Ο λόγος της μελέτης της είναι να διερευνηθούν οι συμμεταβλητές που την καθορίζουν. Έτσι, εάν αυτές καθοριστούν, τότε, με τον έλεγχο των συμμεταβλητών θα μπορέσουν να βελτιωθούν οι πιθανότητες επιβίωσης του παιδιού. Στο έργο τους, οι Ayele et al. (2017) εξήγαγαν αρκετά ενδιαφέροντα αποτελέσματα καθώς με την χρήση των μοντέλων κατέληξαν στο ότι εάν η μητέρα δεν μείνει ξανά έγκυος πριν το παιδί της γίνει 5 ετών, αυτό μειώνει αρκετά την πιθανότητα θνησιμότητας. Με παρόμοιο τρόπο, οι Khan & Awan (2017) έκαναν την ίδια μελέτη στο Μπαγκλαντές. Βασισμένοι σε αυτή την μελέτη οι Yalew et al. (2022) χρησιμοποιώντας πάλι μοντέλα Cox αλλά και Gamma frailty μοντέλα, εντόπισαν και άλλες συμμεταβλητές που ήταν στατιστικά σημαντικοί δείκτες για την θνησιμότητα των παιδιών κάτω από 5 έτη στην Αιθιοπία.

Στην ογκολογία, στο έργο των Ghadimi et al. (2013) μελετήθηκαν οι παράγοντες που μπορεί να επηρεάζουν την επιβίωση ασθενών με καρκίνο του οισοφάγου με την χρήση μοντέλων ευπάθειας και πιο συγκεκριμένα με ευπάθεια κατανεμημένη με Inverse Gaussian κατανομή. Ο καρκίνος του οισοφάγου είναι από τους πιο συνήθεις τύπους θανατηφόρων καρκίνων στις αναπτυσσόμενες χώρες. Βρήκαν λοιπόν ότι το οικογενειακό ιστορικό και το φύλο ήταν από τους πιο βασικούς προγνωστικούς παράγοντες για την επιβίωση του ασθενούς. Σε συνέχεια της ογκολογίας, στο έργο τους οι Yazdani et al. (2019), εφάρμοσαν μοντέλα ευπάθειας για να μελετήσουν ασθενείς με καρκίνο του μαστού και πιο συγκεκριμένα, ποιες συμμεταβλητές καθόριζαν σε μεγαλύτερο βαθμό την επιβίωση αυτών.

Αντίστοιχη μελέτη έγινε και από τους Calsavara et al. (2019) σε ασθενείς με μελάνωμα. Αντίστοιχη δουλειά έχει γίνει και από την Gurmu (2018), η οποία χρησιμοποίησε διάφορα μοντέλα ευπάθειας για να εκτιμήσει τον χρόνο επιβίωσης γυναικών που πάσχουν από καρκίνο του τραχήλου της μήτρας. Έπειτα, οι Esayas Lelisho et al. (2022) έκαναν αντίστοιχη μελέτη σε ασθενείς με καρκίνο του στομάχου και οι Kim et al. (2016) μελέτησαν ασθενείς με καρκίνο της ουροδόχου κύστης.

Στην ιατρική, οι Hanagal & Dabade (2014), (2015) εφάρμοσαν μοντέλα ευπάθειας για να μελετήσουν δεδομένα μόλυνσης νεφρών. Ο σκοπός ήταν να βρεθεί ποιοι ασθενείς ήταν πιο επιρρεπείς στο να νοσήσουν από μόλυνση νεφρών. Στην εργασία τους, οι Hanagal & Dabade (2014), η Gamma κατανεμημένη ευπάθεια περιέγραφε τα δεδομένα με τον βέλτιστο τρόπο. Σε αντίστοιχη μελέτη, οι Ferreira & Colugnati (2021) μελέτησαν την έκβαση χρόνιας νεφρικής ανεπάρκειας και οι Adeleke et al. (2019) μελέτησαν κρίσεις άσθματος σε παιδιά.

Το 2011, στο έργο τους, οι Usha et al. (2011) συζήτησαν την χρήση μοντέλων ευπάθειας σε εφαρμογές γενετικής και βιοϊατρικής καθώς αυτά τα μοντέλα έχουν τη δυνατότητα να εξηγήσουν την μη παρατηρούμενη ετερογένεια που παρατηρείται σε τέτοιου τύπου εφαρμογές. Πιο συγκεκριμένα, σε αυτήν την έρευνα υποστηρίζουν ότι θα μπορούσε αυτή η θεωρία να αποτελέσει θεμέλιο για την δημιουργία διαγνωστικών εργαλείων. Παραδείγματα τέτοιων αναλύσεων δίνονται και στην εργασία τους.

Στην επιδημιολογία, οι Wand & Ramjee (2015) μελέτησαν την μεταδοτικότητα του HIV και πιο συγκεκριμένα, τους παράγοντες που μπορεί να καθορίσουν ποιες ομάδες ατόμων είναι πιο επιρρεπείς στον ιό.

Στην οικονομία, τα μοντέλα ευπάθειας έχουν χρησιμοποιηθεί και στα δάνεια. Στο έργο τους οι Chen et al. (2016) προσπάθησαν να δημιουργήσουν ένα μοντέλο που να περιγράφει την αδυναμία αποπληρωμής στεγαστικών δανείων. Στο μοντέλο τους κατάφεραν και ενσωμάτωσαν επιτυχώς, 15 μεταβλητές.

Τέλος, αρκετή πρόσφατη θεωρητική δουλειά έχει γίνει σχετικά με την εφαρμογή μοντέλων ευπάθειας σε δεδομένα που παρουσιάζουν λογοκρισία από δεξιά στο έργο των Huber-Carol & Vonta (2004) όπως και στο έργο των Slud & Vonta (2004), και γενικότερα με μοντέλα επιβίωσης όπως μπορεί κάποιος να βρει στα έργα των Vonta & Karagrigoriou (2007), Vonta & Karagrigoriou (2010), Vonta & Karagrigoriou (2014) και στις παραπομπές εντός.

Πρέπει να παρατηρήσουμε λοιπόν ότι τα μοντέλα ευπάθειας είναι μία σχετικά νέα θεωρία η οποία έχει προσελκύσει αρκετούς επιστήμονες λόγω των δυνατοτήτων που προσφέρει στην ανάλυση των δεδομένων επιβίωσης. Στο ίδιο πλαίσιο στην παρούσα εργασία, θα προσπαθήσουμε να επεκτείνουμε το πεδίο γνώσεων και εφαρμογών στα μοντέλα ευπάθειας.

ΚΕΦΑΛΑΙΟ 3 – ΣΤΑΤΙΣΤΙΚΗ ΜΕΘΟΔΟΛΟΓΙΑ

3.1. Βασική Θεωρία

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

1. Μη λογοκριμένα δεδομένα (Observed or Uncensored data).
2. Λογοκριμένα δεδομένα (censored data) ή/και Αποκομμένα δεδομένα (truncated data).

3.1.1. Λογοκρισία Δεδομένων (Censoring)

Λογοκριμένες ονομάζονται οι παρατηρήσεις εκείνες για τις οποίες δεν έχει καταγραφεί η ακριβής χρονική στιγμή κατά την οποία επέρχεται το αναμενόμενο γεγονός. Η λογοκρισία δεδομένων είναι και η ειδοποιός διαφορά της ανάλυσης επιβίωσης από τα άλλα πεδία της στατιστικής. Λόγω της λογοκρισίας, ο ερευνητής έχει σαν πληροφορία μόνο το ότι ο υπό μελέτη χρόνος για ένα τερματικό γεγονός ή για ένα γεγονός ενδιαφέροντος δεν έλαβε χώρα πριν από μια χρονική στιγμή λόγω διαφόρων τυχαίων παραγόντων που αφορούν την ιδιαιτερότητα του πειράματος (λήξη του πειράματος ή εγκατάλειψη του υποκειμένου από το πείραμα) καθώς ο χρόνος του πειράματος είναι χρονικά περιορισμένος. Στην περίπτωση αυτή έχουμε όπως ονομάζεται λογοκρισία από δεξιά. Ως εκ τούτου, στην λογοκρισία δεδομένων, η τιμή μιας μεταβλητής ή παρατήρησης είναι μερικώς γνωστή.

Μία λογοκριμένη παρατήρηση περιλαμβάνει μόνο περιορισμένη πληροφορία για την άγνωστη μεταβλητή ενδιαφέροντος πράγμα το οποίο απαιτεί εξειδικευμένες μεθόδους. Πιο συγκεκριμένα, στην ανάλυση επιβίωσης και σχετικά με τα λογοκριμένα δεδομένα, έχουμε μερική πληροφορία για τον χρόνο επιβίωσης ενός ατόμου/αντικειμένου, αφού είναι γνωστό μόνο το κάτω φράγμα του χρόνου επιβίωσης. Εάν δεν γνωρίζουμε ακριβώς τη διάρκεια ζωής ενός ατόμου/αντικειμένου, διαθέτουμε την πληροφορία ότι έχει ξεπεράσει την χρονική διάρκεια κατά την οποία το άτομο ήταν υπό παρατήρηση.

Γενικά στα δεδομένα μπορεί να έχουμε λογοκριμένες παρατηρήσεις που μπορούν να διαφοροποιηθούν σε 3 κατηγορίες:

- **Λογοκρισία από δεξιά (Right Censoring)**

Την περίπτωση αυτή τη συζητήσαμε ήδη οπότε στη συνέχεια παρουσιάζονται οι τύποι της Λογοκρισίας από Δεξιά (οι τρόποι με τους οποίους μπορούν να δημιουργηθούν τέτοιου είδους δεδομένα).

1. **Λογοκρισία Τύπου I**: Ο χρόνος διάρκειας της μελέτης είναι προκαθορισμένος. Λογοκριμένα δεδομένα παράγονται όταν μέσα σε αυτό το ορισμένο χρονικό διάστημα δεν συμβαίνει το γεγονός σε κάποιους από τους υπό μελέτη ασθενείς.
2. **Λογοκρισία Τύπου II**: Η παρακολούθηση – μελέτη ολοκληρώνεται όταν ένας προκαθορισμένος αριθμός «γεγονότων» λάβουν χώρα. Λογοκριμένα δεδομένα παράγονται όταν μέσα σε αυτό το χρονικό διάστημα δεν συμβαίνει το γεγονός σε κάποιους από τους υπό μελέτη ασθενείς.
3. **Τυχαία Λογοκρισία (Random Censoring)**: Ο χρόνος λογοκρισίας είναι τυχαίος για κάθε ασθενή στη μελέτη.

4. **Λογοκρισία από Αριστερά (Left Censoring):** Σε αυτήν την περίπτωση, το συμβάν έχει λάβει χώρα πριν την έναρξη της μελέτης – παρακολούθησης, συνήθως χωρίς να γνωρίζουμε τον ακριβή χρόνο του "γεγονότος". Η συγκεκριμένη περίπτωση παρουσιάζεται λιγότερο συχνά κατά τις εφαρμογές μελετών. Με άλλα λόγια, ο χρόνος του 'γεγονότος' του ασθενή υπολείπεται του χρόνου παρακολούθησης – μελέτης.

5. **Λογοκρισία σε Διάστημα (Interval Censoring),** κατά την οποία το συμβάν λαμβάνει χώρα εντός ενός συγκεκριμένου χρονικού διαστήματος χωρίς όμως να γνωρίζουμε την ακριβή χρονική στιγμή. Τέτοιου είδους δεδομένα συνήθως οφείλονται σε περιπτώσεις κατά τις οποίες η μελέτη είναι περιοδική και δεν υπάρχει συνεχής παρακολούθηση των συμμετεχόντων. Με άλλα λόγια ο χρόνος επιβίωσης κυμαίνεται εντός ενός συγκεκριμένου διαστήματος.

Παράλληλα μπορεί τα δεδομένα να είναι και αποκομμένα (truncated), ή μόνο αποκομμένα, αλλά δεν θα ασχοληθούμε περαιτέρω με αυτού του είδους τα δεδομένα στην εργασία αυτή. Στα πλαίσια της εργασίας, θα ασχοληθούμε με την δεξιά λογοκρισία η οποία είναι η πιο συνήθης παρατηρούμενη μορφή λογοκρισίας και για την οποία και θα δώσουμε λεπτομερή στοιχεία παρακάτω.

Για ένα δείγμα από ασθενείς μεγέθους n, θεωρούμε T_i^*, $i = 1, \dots n$ θετικές ανεξάρτητες και ισόνομες τυχαίες μεταβλητές που αναπαριστούν τους χρόνος επιβίωσης και C_i ανεξάρτητες και ισόνομες τυχαίες μεταβλητές που αναπαριστούν τους χρόνους λογοκρισίας. Αυτό που παρατηρούμε στο πείραμα είναι ο χρόνος T_i ο οποίος είναι ο μικρότερος από τους χρόνους T_i^* και C_i. Δηλαδή

$$T_i = \min(T_i^*, C_i)$$

Επιπρόσθετα, μέσω της τυχαίας μεταβλητής δ_i έχουμε επίσης την πληροφορία για το εάν ο χρόνος T_i είναι χρόνος γεγονότος η χρόνος λογοκρισίας, πιο συγκεκριμένα:

$$\delta_i = \begin{cases} 1 \text{ εάν } T_i^* \leq C_i \text{ και άρα το } T_i \text{ θα αποτελεί τον μη λογοκριμένο χρόνο επιβίωσης.} \\ 0 \text{ εάν } T_i^* > C_i \text{ και άρα το } T_i \text{ θα αποτελεί τον λογοκριμένο χρόνο επιβίωσης.} \end{cases}$$

Ως εκ τούτου, τα δεδομένα αποτελούνται από ζεύγη δεδομένων $(T_1, \delta_1), (T_2, \delta_2), \dots, (T_n, \delta_n)$ όπου $T_i = \min(T_i^*, C_i)$ για κάθε ασθενή, "i".

Η λογοκρισία αποτελεί πρόβλημα σε περιπτώσεις όπως οι κλινικές μελέτες στις οποίες οι ασθενείς μπορεί να εισάγονται και να αποχωρούν από τις μελέτες σε διαφορετικούς χρόνους. Το ενδιαφέρον στις εκάστοτε μελέτες βρίσκεται στους χρόνους συμβάντων των ασθενών. Η λογοκρισία τους μπορεί να οφείλεται στους παρακάτω λόγους:

- Ο ασθενής δεν μπορεί να πάρει μέρος στην μελέτη μετά την εγγραφή του για κάποιον λόγο
- Η μελέτη διακόπτεται λόγω ανωτέρας βίας η λόγω πολύ ισχυρών παρενεργειών.
- Η μελέτη παύει λόγω περάτωσής της.
- Το συμβάν δεν μπορεί να καταγραφεί καθώς επισκιάζεται από άλλο συμβάν του ασθενή π.χ. θάνατος ή ατύχημα.

3.1.2. Συναρτήσεις που Περιγράφουν τους Χρόνους Επιβίωσης

Όπως προαναφέραμε, στην ανάλυση επιβίωσης ο χρόνος είναι η μεταβλητή απόκρισης, ο οποίος λόγω λογοκρισίας (ή/και αποκοπής) δεν είναι πάντα γνωστός όπως απαιτείται να είναι στις συμβατικές στατιστικές συναρτήσεις πιθανοφάνειας. Πιο

συγκεκριμένα, λόγω της λογοκρισίας, τα δεδομένα είναι μία μίξη συνεχών και διακριτών μετρήσεων και απαιτείται διαφορετικός χειρισμός τους στην συνάρτηση πιθανοφάνειας.

Έστω μία θετική τυχαία μεταβλητή T^* η οποία περιγράφει τον χρόνο από ένα καλώς ορισμένο χρονικό σημείο εκκίνησης μέχρι την χρονική στιγμή που λαμβάνει χώρα το υπό μελέτη γεγονός. Εάν το γεγονός είναι ο θάνατος, ονομάζουμε την μεταβλητή T^* "Χρόνο Ζωής". Αυτή η μεταβλητή είναι συνήθως είναι συνεχής, αλλά λόγω της δυσκολίας μέτρησης της σε όλο το χρονικό ορίζοντα του πειράματος, μπορεί να εμφανιστεί και υπό 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 είναι εφικτός. Αυτό το γεγονός λοιπόν δημιουργεί προβλήματα στην αρχική επιλογή του κατάλληλου στατιστικού μοντέλου που θα χρησιμοποιηθεί.

Η κατανομή των χρόνων επιβίωσης, κατά βάση χαρακτηρίζεται από 2 συναρτήσεις, τη συνάρτηση επιβίωσης και την συνάρτηση κινδύνου ή διακινδύνευσης. Ενίοτε, χρησιμοποιείται και η συνάρτηση πυκνότητας πιθανότητας. Ανάλογα με το πρόβλημα, χρησιμοποιείται και η καταλληλότερη συνάρτηση.

Ας υποθέσουμε, χωρίς βλάβη της γενικότητας, ότι ο χρόνος του γεγονότος T^* ακολουθεί συνεχή κατανομή και γενικότερα, όλες οι κατανομές του χρόνου μέχρι ένα γεγονός ορίζονται στο πεδίο $[0, \infty)$ εκτός αν ορίζεται κάτι διαφορετικό από το πρόβλημα. Η κατανομή μιας συνεχούς τυχαίας μεταβλητής ορίζεται μονοσήμαντα από την συνάρτηση πυκνότητας πιθανότητας f. Η συνάρτηση κατανομής της τυχαίας μεταβλητής T^* ορίζεται ως:

$$F(t) = P(T^* \leq t) = \int_0^t f(s)ds \quad (1)$$

Πρόκειται για την αθροιστική συνάρτηση κατανομής (Cumulative distribution function-C.D.F) του T^* όπου $P(A)$ συμβολίζει την πιθανότητα να συμβεί το γεγονός A.

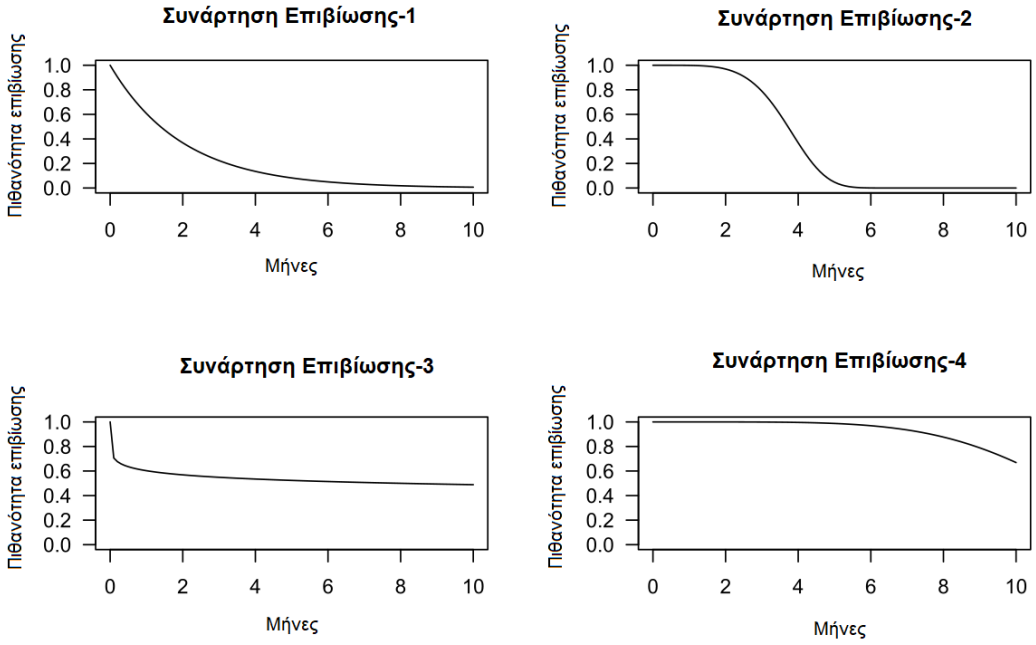
Στην ανάλυση επιβίωσης, ενδιαφερόμαστε κυρίως για την πιθανότητα ενός ατόμου να επιβιώσει πέρα από την χρονική στιγμή t. Η πιθανότητα αυτή ονομάζεται και συνάρτηση επιβίωσης η οποία δίνεται από τον τύπο:

$$S(t) = P(\delta\iota\acute{\alpha}\rho\kappa\epsilon\iota\alpha\ \zeta\omega\acute{\eta}\varsigma\ \mu\epsilon\gamma\alpha\lambda\acute{\upsilon}\tau\epsilon\rho\eta\ \tau\sigma\upsilon\ t) = P(T^* > t) = 1 - F(t) = \int_t^\infty f(s)ds$$

Άρα:

$$S(t) = \int_t^{\infty} f(s)ds \quad (2)$$

Πρόκειται για μία μη αρνητική και φθίνουσα συνάρτηση του χρόνου με $S(0) = 1$ και $S(\infty) = 0$. Η γραφική παράσταση της $S(t)$ σε σχέση με το t είναι γνωστή και ως καμπύλη επιβίωσης. Μερικές γραφικές παραστάσεις συναρτήσεων επιβίωσης μπορούμε να δούμε και στην Εικόνα 1 παρακάτω:



Εικόνα 1: Καμπύλες επιβίωσης της σχέσης (2) για διάφορες σ.π.π. f(s).

Η συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής T^* υπολογίζεται ως:

$$f(t) = \frac{d}{dt}F(t) = -\frac{d}{dt}S(t) \quad (3)$$

Της οποίας η καμπύλη ονομάζεται καμπύλη πυκνότητας πιθανότητας και η αναμενόμενη διάρκεια ζωής (μέσος χρόνος επιβίωσης) υπολογίζεται ως:

$$\mu = E(T^*) = \int_0^{\infty} tf(t)dt = -\int_0^{\infty} t\frac{d}{dt}S(t)dt = \int_0^{\infty} S(t)dt \quad (4)$$

3.1.3. Συνάρτηση Διακινδύνευσης

Η συνάρτηση κινδύνου ή διακινδύνευσης χαρακτηρίζει τον κίνδυνο θανάτου που μεταβάλλεται με τον χρόνο. Προσδιορίζει τον ρυθμό απρόσμενου θανάτου στο χρονικό διάστημα $(t, t+\epsilon)$, δοθέντος του ότι το άτομο επιβιώνει μέχρι την χρονική στιγμή t και ορίζεται ως:

$$h(t) = \lim_{\delta t \to 0} \frac{P(t < T^* \leq t + \delta t | T^* > t)}{\delta t} = \frac{f(t)}{1 - F(t)}$$

άρα

$$h(t) = \frac{f(t)}{1 - F(t)} \quad (5)$$

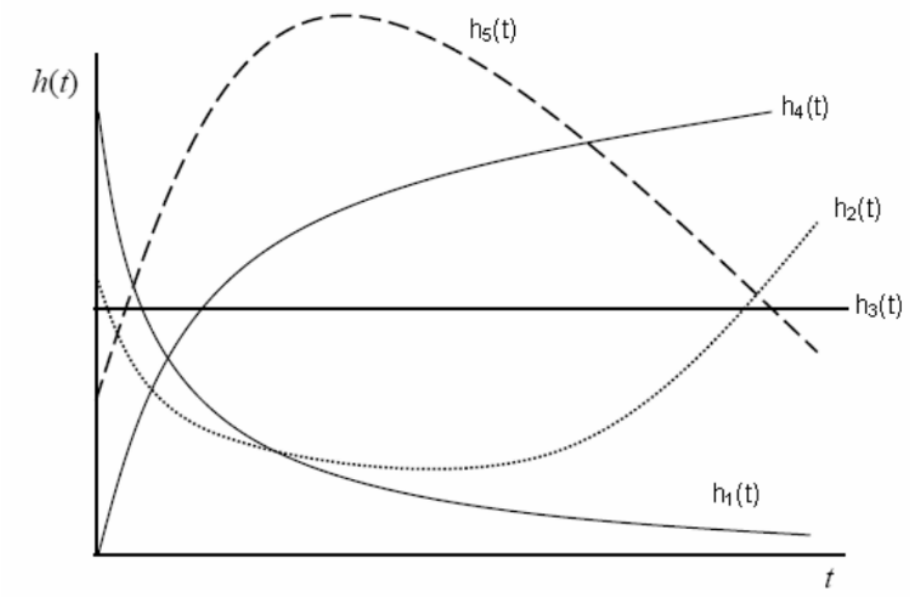
Για να εξηγήσουμε τα πιο πάνω, από τον ορισμό της δεσμευμένης πιθανότητας έχουμε ότι:

$$P[t < T^* < t + \delta t | T^* > t] = \frac{P[t < T^* < t + \delta t]}{P[T^* > t]} = \frac{S(t) - S(t + \delta t)}{S(t)}$$

Συνεπώς η συνάρτηση διακινδύνευσης αλλάζει σε:

$$h(t) = \lim_{\delta t \to 0} \frac{\frac{S(t) - S(t + \delta t)}{S(t)}}{\delta t} = \frac{f(t)}{S(t)} \quad (6)$$

Οπού η σχέση αυτή εκφράζει τον στιγμιαίο ρυθμό κινδύνου ή αποτυχίας.



Εικόνα 2: Καμπύλες συναρτήσεων διακινδύνευσης της σχέσης (5) για διάφορα f(t), S(t) .

Οι τρείς συναρτήσεις f, S και h που προαναφέρθηκαν, είναι μαθηματικά ισοδύναμες καθώς με την γνώση της μίας μπορούν να υπολογιστούν οι άλλες δύο.

Ήδη είδαμε ότι ισχύει

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

Λόγω του ότι η συνάρτηση πυκνότητας πιθανότητας οποιασδήποτε κατανομής ισούται με την παράγωγο της συνάρτησης κατανομής της, έχουμε ότι:

$$f(t) = \frac{d}{dt}F(t) = -\frac{d}{dt}S(t) \quad (7)$$

Συνεπώς:

$$h(t) = \frac{1}{S(t)}\left(-\frac{d}{dt}S(t)\right) \quad (8)$$

ή ισοδύναμα

$$h(t) = -\frac{d}{dt}\ln\big(S(t)\big) \quad (9)$$

Ολοκληρώνοντας την σχέση (9) στο $[0, t]$ με $S(0) = 1$, δηλαδή ορίζοντας την αθροιστική συνάρτηση κινδύνου ως

$$H(t) = \int_0^t h(x)dx$$

προκύπτει ότι

$$H(t) = -lnS(t) \text{ ή } \textit{ισοδύναμα} \quad S(t) = e^{-H(t)} \quad (10)$$

και

$$f(t) = h(t)e^{-H(t)} \quad (11)$$

3.1.4. Μοντέλο Αναλογικής Διακινδύνευσης του Cox

Τα μοντέλα που χρησιμοποιούνται στην ανάλυση επιβίωσης είναι χρήσιμα στην πιο απλή περίπτωση όπου οι χρόνοι επιβίωσης των ατόμων του δείγματος είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές. Στα περισσότερα προβλήματα όμως, ο υπό μελέτη πληθυσμός είναι μη ομογενής, και τα άτομα σχετίζονται και με άλλες μεταβλητές, που θα τις ονομάσουμε συμμεταβλητές (όπως φύλο, ηλικία, ιατρικό ιστορικό, οικογενειακή κατάσταση, σπουδές, δημογραφικά χαρακτηριστικά, κ.ο.κ). Ο ερευνητής λοιπόν στις

περισσότερες περιπτώσεις καλείται να εξετάσει κατά πόσο και πώς επηρεάζουν οι μεταβλητές αυτές τον χρόνο επιβίωσης.

Η σχέση μεταξύ του χρόνου επιβίωσης ενός ατόμου και άλλων συμμεταβλητών ορίζεται συνήθως μέσω ενός μοντέλου παλινδρόμησης. Όταν έχουμε λογοκριμένα δεδομένα επιβίωσης και συμμεταβλητές που εμπλέκονται, χρησιμοποιείται ευρέως το μοντέλο αναλογικής διακινδύνευσης του Cox. Το μοντέλο αυτό πρωτοεισήχθη από τον Cox (1972). Είναι ένα μοντέλο παλινδρόμησης όπου ο χρόνος του γεγονότος είναι η εξαρτώμενη μεταβλητή και επιτρέπει την εισαγωγή πληροφορίας για γνωστές (παρατηρούμενες) συμμεταβλητές για κάθε άτομο με έναν εύκολο τρόπο.

Για την χρήση του μοντέλου αυτού, ανάλογα με το τι υποθέσεις θέλει να κάνει ο ερευνητής για την κατανομή των χρόνων επιβίωσης, η συνάρτηση κινδύνου μπορεί να πάρει οποιαδήποτε μορφή, οδηγώντας σε παραμετρικά (parametric) ή ημιπαραμετρικά (semi-parametric) μοντέλα. Η συνάρτηση κινδύνου σε αυτό το μοντέλο μπορεί να πάρει οποιαδήποτε μορφή, μέχρι και τη μορφή κλιμακωτής συνάρτησης (step-function). Είναι σημαντικό να αναφέρουμε όμως ότι οι κίνδυνοι δύο ατόμων, με βάση τον ορισμό του μοντέλου, είναι ανάλογοι και ο λόγος των κινδύνων ανεξάρτητος του χρόνου. Η συνάρτηση πιθανοφάνειας που χρησιμοποιείται για την εκτίμηση των παραμέτρων αντικαθίσταται από τη μερική συνάρτηση πιθανοφάνειας (partial likelihood function) την οποία εισήγαγε ο Cox το 1975 (Cox, 1975). Το σημαντικό γεγονός είναι ότι η στατιστική συμπερασματολογία με βάση τη μερική συνάρτηση πιθανοφάνειας δεν υπολείπεται σε τίποτα από εκείνη που βασίζεται στην ολική συνάρτηση πιθανοφάνειας όσον αφορά τη συνέπεια και αποδοτικότητα των εκτιμητριών.

Για να εξηγήσουμε καλύτερα το μοντέλο, έστω $h(t|X)$ ο κίνδυνος ενός ατόμου σε μια χρονική στιγμή t οπου $X^T = (X_1, \dots, X_k)$ το διάνυσμα των συμμεταβλητών που πιστεύουμε ότι ασκούν επιρροή στον χρόνο ζωής αυτού του ατόμου, οι οποίες είναι γνωστές στον ερευνητή. Οι συμμεταβλητές μπορούν να αφορούν οποιοδήποτε χαρακτηριστικό όπως:

- Φυσικές ιδιότητες ατόμων (ηλικία / φύλο /…).
- Το αν ο υπό μελέτη ασθενής έχει υποβληθεί σε κάποια θεραπεία.
- Εξωγενείς παράγοντες.

Το μοντέλο αναλογικής διακινδύνευσης έχει την ακόλουθη γενική μορφή ως προς την συνάρτηση κινδύνου:

$$h(t|X) = h_0(t)g(X) \quad (12)$$

Οπού $h_0(t)$ είναι η βασική συνάρτηση διακινδύνευσης, αυτή που αντιστοιχεί στα άτομα με συμμεταβλητές $X = 0$. Επίσης, το μοντέλο του Cox υποθέτει ότι $g(x)$ είναι κάποια θετική συνάρτηση του x. Πιο συγκεκριμένα, η συνάρτηση $g(X)$ στο μοντέλο του Cox παίρνει την εξής μορφή

$$g(X) = e^{\beta^T X} \quad (13)$$

με $\beta' = (\beta_1, \beta_2, \dots, \beta_k)$ να αποτελεί το διάνυσμα των παραμέτρων της παλινδρόμησης. Σε αυτό το μοντέλο, οι συμμεταβλητές δρουν λοιπόν πολλαπλασιαστικά στη βασική συνάρτηση κινδύνου.

Με την προσθήκη επιπλέον παραγόντων ρίσκου, όπως αυτά καθορίζονται από τις προγνωστικές πληροφορίες των ατόμων, αυτό το μοντέλο επιτρέπει να έχουμε μια απλή και εύκολη ερμηνεία των αποτελεσμάτων. Υποθέτουμε σε αυτό το μοντέλο ότι όλη η διακύμανση του κινδύνου μπορεί να εξηγηθεί από ένα πεπερασμένο διάνυσμα των παρατηρούμενων συμμεταβλητών. Η βασική ιδέα πίσω από τον ορισμό του μοντέλου είναι αφενός ο διαχωρισμός της χρονικής επίδρασης μέσω μόνο της βασικής συνάρτησης διακινδύνευσης και αφετέρου ότι η επίδραση των συμμεταβλητών εμπλέκεται στο μοντέλο μέσω ενός εκθετικού όρου ο οποίος εξασφαλίζει τη θετικότητα της συνάρτησης κινδύνου για όποιες τιμές των συμμεταβλητών.

Όπως είπαμε πιο πριν και θα αποδείξουμε πιο κάτω, οι κίνδυνοι δύο ατόμων την χρονική στιγμή t συνδέονται με έναν συντελεστή αναλογίας που δεν εξαρτάται από το t. Η συνάρτηση διακινδύνευσης ενός ατόμου με διάνυσμα συμμεταβλητών X, ορίζεται από τις σχέσεις (12) και (13), δεδομένων των συμμεταβλητών, ως:

$$h(t|X) = h_0(t)e^{\beta^T X} \quad (14)$$

Η ανεξαρτησία της διακινδύνευσης (και άρα και της επιβίωσης) από μια δεδομένη συμμεταβλητή X_j, συνεπάγεται ότι β_j=0. Αυτός είναι και ο λόγος που γίνονται έλεγχοι υποθέσεων για αυτήν ακριβώς τη μηδενική υπόθεση για όλες τις συμμεταβλητές.

Τονίζουμε ξανά ότι η συνάρτηση διακινδύνευσης $h(t|X)$, εξαρτάται μόνο από τον χρόνο και τις επιμέρους συμμεταβλητές αλλά μέσω δυο διαφορετικών παραγόντων.

1. Ο πρώτος παράγοντας, $h_0(t)$, που αφήνεται αυθαίρετος, είναι μόνο συνάρτηση του χρόνου και θεωρείται ίδια για το πλήθος των ατόμων n.
2. Ο δεύτερος παράγοντας είναι μια ποσότητα που εξαρτάται από τις συμμεταβλητές μόνο μέσω του διανύσματος β^T.

Θεωρούμε τον "*Λόγο κινδύνου*" (Hazard Ratio - HR) ως τον λόγο των συναρτήσεων διακινδύνευσης δυο ατόμων. Στο μοντέλο αναλογικής διακινδύνευσης, όπως αναφέρεται και στο όνομα του μοντέλου, παρατηρείται η ιδιότητα της αναλογίας στις συναρτήσεις διακινδύνευσης δύο ατόμων, δηλαδή, αν $\frac{h(t|X_1)}{h(t|X_2)}$ ο λόγος των συναρτήσεων διακινδύνευσης δύο ατόμων και $X_1 = (x_{11}, x_{12}, \dots, x_{1k}), X_2 = (x_{21}, x_{22}, \dots, x_{2k})$ τα αντίστοιχα διανύσματα συμμεταβλητών τους, τότε προκύπτει ότι:

$$HR(t) = \frac{h(t|x_1)}{h(t|x_2)} = \frac{h_0(t)e^{\beta^T X_1}}{h_0(t)e^{\beta^T X_2}} = e^{\beta^T (X_1 - X_2)}$$

Αρά:

$$HR(t) = e^{\beta^T (X_1 - X_2)} \quad (15)$$

Αυτό σημαίνει ότι ο λόγος του κινδύνου για δύο ασθενείς είναι μια σταθερά ανεξάρτητη του χρόνου. Όπως μπορεί να αντιληφθεί κάποιος διαισθητικά, αυτό δεν μπορεί να ισχύει απόλυτα στην πραγματικότητα καθώς ο κίνδυνος ενός ατόμου εξαρτάται από πολλούς παράγοντες, πολλοί από τους οποίους αλλάζουν με τον χρόνο.

Στο μοντέλο του Cox ισχύει επίσης ότι η συνάρτηση επιβίωσης ενός ατόμου με διάνυσμα συμμεταβλητών X είναι η βασική συνάρτηση επιβίωσης υψωμένη σε κατάλληλη δύναμη. Θα το αποδείξουμε αυτό πιο κάτω στη σχέση (17).

Είναι γνωστό ότι η συνάρτηση διακινδύνευσης είναι μαθηματικά ισοδύναμη με την συνάρτηση επιβίωσης μέσω της σχέσης (9) όπως είδαμε πριν:

$$h(t) = -\frac{d}{dt} lnS(t)$$

Η αλλιώς μέσω της σχέσης (10):

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

όπου $H(t)$ η αθροιστική συνάρτηση διακινδύνευσης. Επομένως, ολοκληρώνοντας την σχέση (14) έχουμε:

$$H(t|X) = \int_0^t h_0(u)\, e^{\beta^T X} du = H_0(t)e^{\beta^T X}$$

Αρα:

$$H(t|X) = H_0(t)e^{\beta^T X} \quad (16)$$

και συνεπώς από την σχέση (10) μέσω της σχέσης (16) έχουμε:

$$S(t|x) = e^{-H(t|X)} = e^{-H_0(t)e^{\beta^T X}} = [S_0(t)]^{e^{\beta^T X}} \quad (17)$$

όπου $S_0(t)$ η βασική συνάρτηση επιβίωσης ίση με $e^{-H_0(t)}$ όπου $H_0(t)$ είναι η βασική αθροιστική συνάρτηση κινδύνου.

3.1.5. Η Εκτιμήτρια Kaplan Meier

Η μη-παραμετρική εκτιμήτρια Kaplan-Meier [Kaplan and Meier (1958)] της συνάρτησης επιβίωσης αποτέλεσε ένα γιγαντιαίο βήμα για την ανάλυση επιβίωσης γιατί λαμβάνει υπ' όψη της την λογοκρισία στην εκτίμηση της πιθανότητας επιβίωσης. Η εκτιμήτρια αυτή δεν βασίζεται σε κανένα μοντέλο αλλά μόνο στα δεδομένα και την εξέλιξη του φαινομένου που παρατηρούμε μέσα στο χρόνο. Είναι λογικό λοιπόν ότι η αποδοτικότητα όλων των μοντέλων στο να περιγράψουν τα δεδομένα και πιο συγκεκριμένα την πιθανότητα επιβίωσης των ασθενών, συγκρίνεται μέσω της εκτιμήτριας της πιθανότητας επιβίωσης Kaplan-Meier. Παρακάτω θα ορίσουμε την εκτιμήτρια Kaplan-Meier.

Για να εκτιμήσουμε την συνάρτηση επιβίωσης $S(t)$, θεωρούμε ένα δείγμα n ατόμων από τον πληθυσμό. Έστω διαδικασία $N(t)$ η οποία μετράει τον αριθμό των γεγονότων στο χρονικό διάστημα $[0, t]$, δηλαδή μετρά τον αριθμό των μη λογοκριμένων παρατηρήσεων. Επίσης έστω $Y(t)$ να είναι ο αριθμός των ατόμων που βρίσκονται σε κίνδυνο ακριβώς πριν τον χρόνο t (δηλαδη έχουν επιβιώσει μέχρι τον χρόνο t). Εστω $T_1 < T_2 < \cdots$ οι διακεκριμένοι χρόνοι σε αύξουσα σειρά που παρατηρήθηκαν γεγονότα, δηλαδή οι χρόνοι στους οποίους αυξήθηκε το $N(t)$.

Για να δώσουμε μία διαισθητική ερμηνεία της εκτιμήτριας Kaplan Meier του $S(t)$, χωρίζουμε το χρονικό διάστημα $[0, t]$ σε χρονικά διαστήματα $0 = t_0 < t_1 < \cdots < t_K = t$ και χρησιμοποιούμε τον πολλαπλασιαστικό νόμο πιθανοτήτων και εύκολα βλέπουμε ότι:

$$S(t) = \prod_{k=1}^{K} S(t_k | t_{k-1}) \qquad (17)$$

όπου

$$S(v|u) = \frac{S(v)}{S(u)} = \frac{P(T > v)}{P(T > u)} \; για \; v > u$$

ορίζεται ως η πιθανότητα το γεγονός να συμβεί αργότερα από τον χρόνο v δοθέντος του ότι δεν έχει πραγματοποιηθεί ακόμα στον χρόνο u. Θεωρούμε βέβαια ότι δεν υπάρχουν γεγονότα που να λαμβάνουν χώρα στον ίδιο χρόνο. Επίσης όλες οι λογοκρισίες που παρατηρούνται είναι από δεξιά.

Εάν δεν παρατηρείται κανένα γεγονός στο διάστημα $(t_k - 1, t_k]$ έχουμε ότι $S(t_k|t_{k-1}) = 1$. Αλλιώς εάν παρατηρείται γεγονός σε χρόνο $T_j \in (t_k - 1, t_k]$, τότε η εκτίμηση της $S(t_k|t_{k-1})$ είναι:

$$1 - \frac{1}{Y(t_k - 1)} = 1 - \frac{1}{Y(T_j)}$$

Και άρα από την (17) έχουμε ότι η εκτιμήτρια Kaplan-Meier της πιθανότητας επιβίωσης ορίζεται ως:

$$\hat{S}(t) = \prod_{T_j \leq t} \left(1 - \frac{1}{Y(T_j)}\right)$$

3.2. Μοντέλα Ευπάθειας

3.2.1. Η Ανάγκη για Γενίκευση του Μοντέλου του Cox

Οι μέθοδοι που χρησιμοποιούνται στην ανάλυση επιβίωσης είναι από τις βασικές ερευνητικές μεθόδους που χρησιμοποιούνται σε πολλά πεδία όπως η Ιατρική, η Βιολογία, η Δημογραφία και η Μηχανική. Η ονομασία που έχει δοθεί στην ανάλυση επιβίωσης προέρχεται από την εφαρμογή των μεθόδων που προαναφέραμε σε ιατρικές και δημογραφικές μελέτες θνησιμότητας. Ειδικά μετά από τα τέλη του 1970, η ανάλυση δεδομένων με την χρήση μεθόδων ανάλυσης επιβίωσης εξαπλώθηκε ραγδαία μετά την ανάπτυξη του μοντέλου αναλογικής διακινδύνευσης του Cox (1972) και μερικών επεκτάσεών του κατά την διάρκεια των τελευταίων τριών δεκαετιών. Στην παρούσα διπλωματική θα μελετήσουμε μία γενίκευση του μοντέλου του Cox που είναι τα μοντέλα ευπάθειας (*Frailty models*).

Τις τελευταίες δεκαετίες, έχουν γίνει πολλές δημοσιεύσεις πάνω στα μοντέλα ευπάθειας. Οι βασικές ιδέες σε αυτά τα μοντέλα είναι ότι:

- Τα υπό μελέτη άτομα χαρακτηρίζονται από διαφορετικές ευπάθειες.
- Τα περισσότερο ευπαθή άτομα τείνουν να πεθάνουν νωρίτερα από τα λιγότερο ευπαθή.

Τα κλασσικά μοντέλα επιβίωσης ασχολούνται με την κλασσική περίπτωση ανεξάρτητων και ομοιόμορφα κατανεμημένων δεδομένων. Αυτό βασίζεται στην παραδοχή ότι ο υπό μελέτη πληθυσμός είναι ομογενής μέχρι έναν συγκεκριμένο αριθμό συμμεταβλητών.

Ωστόσο, στην πραγματικότητα, εύκολα κάποιος παρατηρεί ότι ο υπό μελέτη πληθυσμός χαρακτηρίζεται από μεγάλη ετερογένεια σε χαρακτηριστικά (όπως για

παράδειγμα για ανθρώπους, ηλικία φύλο, εθνικότητα κτλ.…) ή και σε επιρροή επεξηγηματικών μεταβλητών.

Αυτή η ετερογένεια συχνά αναφέρεται και ως μεταβλητότητα. Αυτό που θα μας απασχολήσει λοιπόν στα μοντέλα ευπάθειας είναι η μη παρατηρούμενη μεταβλητότητα στην ανάλυση των δεδομένων αυτών. Η ετερογένεια αυτή μπορεί να είναι δύσκολο να εκτιμηθεί αλλά είναι παράλληλα σημαντική.

Γενικά, είναι δύσκολο και σχεδόν ακατόρθωτο να συμπεριλάβει κάποιος όλους τους παράγοντες κινδύνου, ίσως γιατί ο ερευνητής έχει λίγο έως καθόλου πληροφορία σε ατομικό επίπεδο (του κάθε ασθενή). Αυτό ισχύει κυρίως για μελέτες πληθυσμών όπου οι μόνες γνωστές μεταβλητές είναι το φύλο και η ηλικία. Επίσης μπορεί να μην γνωρίζουμε την σημαντικότητα ενός παράγοντα όπως επίσης και την ύπαρξή του! Σε άλλες περιπτώσεις, μπορεί να είναι και πολύ δύσκολο να μετρήσουμε τον παράγοντα κινδύνου λόγω του απαγορευτικού χρονικού και εργασιακού κόστους.

3.2.2. Η Γενίκευση

Ο σκοπός της εισαγωγής αγνώστων επιδράσεων στο μοντέλο του Cox ήταν να βελτιωθεί η προτυποποίηση μοντέλων θνησιμότητας πληθυσμών. Ο Vaupel et al. (1979) εισήγαγε την έννοια της ευπάθειας στην περιοχή της βιοστατιστικής και την εφάρμοσε σε δεδομένα πληθυσμιακής θνησιμότητας.

Θα προχωρήσουμε τώρα στον ορισμό του μοντέλου ευπάθειας. Αρχικά, Θα θεωρήσουμε ένα μοντέλο ευπάθειας χωρίς παρατηρούμενες συμμεταβλητές προκειμένου να εστιάσουμε στα βασικά σημεία της προτυποποίησης της ευπάθειας (*univariate frailty*). Το κλασσικό και πιο συχνά χρησιμοποιούμενο μοντέλο υποθέτει μία δομή αναλογικής διακινδύνευσης που εξαρτάται από την άγνωστη επίδραση (*Ευπάθεια*). Για να είμαστε πιο συγκεκριμένοι, η συνάρτηση διακινδύνευσης ενός ατόμου, εξαρτάται από μία μη-παρατηρούμενη, χρονοανεξάρτητη άγνωστη τυχαία μεταβλητή που θα την ονομάζουμε "Z". Υποθέτουμε ότι η τυχαία μεταβλητή Z επιδρά πολλαπλασιαστικά στην βασική συνάρτηση διακινδύνευσης $h_0(t)$, πιο συγκεκριμένα δεδομένης της τιμής της ευπάθειας Z, ισχύει ότι

$$h(t|Z) = Zh_0(t) \quad (18)$$

Εδώ, η ευπάθεια Z θεωρείται ότι είναι μία μή-αρνητική τυχαία μεταβλητή, που μεταβάλλεται στον πληθυσμό από άτομο σε άτομο. Ας σημειωθεί ότι η παράμετρος κλίμακας κοινή σε όλα τα άτομα του υπό μελέτη πληθυσμού μπορεί να απορροφηθεί στην βασική συνάρτηση διακινδύνευσης $h_0(t)$ προκειμένου οι κανονικοποιημένες κατανομές ευπάθειας να έχουν την ιδιότητα $EZ = 1$, εάν ορίζεται και υπάρχει η αναμενόμενη τιμή της ευπάθειας.

Η διασπορά της ευπάθειας $\sigma^2 = V(Z)$ (εάν υπάρχει) ερμηνεύεται σαν το μέγεθος της ετερογένειας του πληθυσμού σε σχέση με τον βασικό κίνδυνο. Όταν το σ^2 είναι μικρό, τότε οι τιμές του Z είναι πολύ κοντά στην μονάδα, ενώ οταν το σ^2 είναι μεγάλο, τοτε οι τιμες του Z έχουν μεγαλύτερη διασπορά, προσδίδοντας μεγαλύτερη ετερογένεια στους κινδύνους του κάθε ατόμου του υπό μελέτη πληθυσμού.

Είναι προφανές ότι ένα πολλαπλασιαστικό μοντέλο ευπάθειας αναπαριστά μια πολύ απλοποιημένη μορφή της ετερογένειας και του πως αυτή πιθανόν να δρα. Οι υποθέσεις ότι η ευπάθεια είναι χρονοανεξάρτητη και ότι δρα πολλαπλασιαστικά στον υποκείμενο βασικό κίνδυνο είναι αυθαίρετες αλλά λαμβάνονται σαν βάσεις για τις περισσότερες μελέτες σχετικές με την μη-παρατηρούμενη ετερογένεια στην ανάλυση επιβίωσης. Πιο συγκεκριμένα, με την χρήση των πολλαπλασιαστικών μοντέλων ευπάθειας, ο ερευνητής θα υπολογίσει σαν εκτιμήτρια της ευπάθειας μία τιμή που αντικατοπτρίζει την ολότητα των ατόμων και η οποία τιμή είναι και χρονικά σταθερή. Στην πραγματικότητα το κάθε άτομο της μελέτης έχει την δικιά του ευπάθεια, η οποία αλλάζει με τον χρόνο.

Όπως προαναφέραμε, μια κλάση μοντέλων γεννάται εάν εισάγουμε μια μη παρατηρούμενη μεταβλητή ευπάθειας "Z" μέσω της συνάρτησης διακινδύνευσης προκειμένου να εξηγήσουμε την ετερογένεια ενός πληθυσμού. Εισάγοντας λοιπόν και παρατηρούμενες συμμεταβλητές στο μοντέλο (18), κατ' αναλογία με το μοντέλο του Cox, η σχέση (16) γίνεται:

$$H(t|X,Z) = Ze^{\beta^T X} h_0(t), \qquad Z > 0 \qquad (19)$$

όπου $X = (X_1, X_2, \dots, X_\kappa)$ το διάνυσμα των συμμεταβλητών και $\beta = (\beta_1, \beta_2, \dots, \beta_\kappa)$ οι παράμετροι της παλινδρόμησης. Συνεπώς, τα μοντέλα ευπάθειας αποτελούν γενίκευση του γνωστού μοντέλου αναλογικής διακινδύνευσης. Το μοντέλο αναλογικής διακινδύνευσης προκύπτει εάν $Z = 1$ για όλα τα άτομα.

Συνεχίζοντας, αν θεωρήσουμε ότι $S(t|Z)$ δηλώνει την συνάρτηση επιβίωσης ενός ατόμου δεδομένης της ευπάθειάς του Z, η συνάρτηση επιβίωσης γίνεται από την σχέση (10) μέσω της (19):

$$S(t|Z) = e^{-\int_0^t h(S|Z)ds} = e^{-Z\int_0^t h_0(s)ds} = e^{-ZH_0(t)}$$

Άρα

$$S(t|Z) = e^{-ZH_0(t)} \qquad (20)$$

Σε αυτό το σημείο, η συνάρτηση $H_0(t) = \int_0^t h_0(s)ds$ θυμίζουμε ότι αποτελεί την αθροιστική βασική συνάρτηση διακινδύνευσης. Εάν υπάρχουν και παρατηρούμενες συμμεταβλητές έχουμε κατ΄ αναλογίαν ότι (μέσω της (19)):

$$S(t|Z,X) = e^{-\int_0^t h(S|Z,X)ds} = e^{-Ze^{\beta^T X}\int_0^t h_0(s)ds} = e^{-Ze^{\beta^T X}H_0(t)}$$

Θα δώσουμε τώρα ένα παράδειγμα για την ύπαρξη ετερογένειας σε ένα πληθυσμό. Για να αποδειχθεί η ύπαρξη μη παρατηρούμενης ετερογένειας σε έναν πληθυσμό, προτάθηκαν πειράματα που υπέβαλαν σε στρες διάφορους ζώντες οργανισμούς σε εργαστήρια (όπως σκουλήκια και έντομα). Η ιδέα του να χρησιμοποιούνται δεδομένα από πειράματα στρες για να τεστάρουν την υπόθεση της ετερογένειας, βασίζεται στην πεποίθηση ότι η ίδια έκθεση σε στρες παράγει διαφορετικά αποτελέσματα επιβίωσης σε ετερογενείς και ομογενείς πληθυσμούς. Με την έννοια της ετερογένειας, εννοούμε ότι τα στοιχεία του πληθυσμού που υποβάλλονται σε στρες, μπορεί να είναι μεν το ίδιο είδος, αλλά διαφέρουν ως προς ένα χαρακτηριστικό τους όπως π.χ. το φύλο ή κάποιον άλλον βιολογικό (η μη) παράγοντα.

Κατά την διάρκεια της έκθεσης, στον ετερογενή υπό μελέτη πληθυσμό τα πιο ευπαθή μέλη του θα πεθάνουν πρώτα. Στον ομοιογενή πληθυσμό από την άλλη, η θνησιμότητα είναι αποτέλεσμα τυχαιότητας. Ως εκ τούτου, μετά από την έκθεση σε στρες, ο ρυθμός θνησιμότητας της ομάδας που υπεβλήθη σε στρες θα πρέπει να είναι μικρότερος για τον ομογενή πληθυσμό.

Τα μοντέλα ευπάθειας εστιάζουν μόνο στην ιδέα της μη παρατηρούμενης ετερογένειας, η οποία βέβαια είναι μία υπεραπλούστευση. Για παράδειγμα, η έννοια της προσαρμοστικότητας που μπορεί να παρουσιάσει ο κάθε οργανισμός στο στρες, δεν υφίσταται. Το ίδιο ισχύει και με την έννοια της εξασθένισης, που είναι το αντίθετο. Ο λόγος

που δεν λαμβάνονται υπ' όψην αυτές οι δύο έννοιες στα μοντέλα ευπάθειας είναι διότι η ευπάθεια θεωρείται σταθερή με τον χρόνο, εξαιρώντας τις χρονοεξαρτημένες επιδράσεις της προσαρμογής και της εξασθένισης. Περισσότερες λεπτομέρειες υπάρχουν στα έργα των Khazaeli et al. (1995), Drapeau et al. (2000), Mueller et al. (2003), Wu et al. (2006), Rose et al. (2006), and Rockwood and Mitnitski (2007), Rockwood (2005).

Ο όρος ευπάθεια προέρχεται από τη γεροντολογία όπου χρησιμοποιείται για να δείξει την ευαισθησία στη θνησιμότητα και τη νοσηρότητα. Ωστόσο, οι απόψεις ως προς το τι καθορίζει την ιατρική και γεροντολογική έννοια της ευπάθειας διίστανται (βλέπετε τα έργα των Rockwood (2005), και Rockwood and Mitnitski (2007) και αναφορές σε αυτά).

Στους τομείς της βιοστατιστικής και της δημογραφίας, η ευπάθεια ερμηνεύεται ως μία τυχαία επίδραση. Αυτός ο στατιστικός ορισμός της αδυναμίας είναι διαφορετικός από αυτόν που χρησιμοποιείται στους τομείς της γεροντολογίας και της ιατρικής, και οι δύο έννοιες δεν πρέπει να συγχέονται. Η ευπάθεια στη βιοστατιστική και στη δημογραφία συνήθως θεωρείται ότι είναι σταθερή με την πάροδο του χρόνου για ένα άτομο (με ελάχιστες εξαιρέσεις) και ερμηνεύεται ως τυχαία επίδραση. Στο ιατρικό και γεροντολογικό πλαίσιο η ευπάθεια θεωρείται ότι αλλάζει και ότι συνήθως αυξάνεται με την πάροδο του χρόνου.

Μονομεταβλητά μοντέλα ευπάθειας χρησιμοποιούνται συχνά στη βιοστατιστική για να μοντελοποιήσουν την επίδραση της μη παρατηρούμενης ετερογένειας (μη παρατηρήσιμες συμμεταβλητές), ενώ στο ιατρικό πλαίσιο ο κύριος στόχος είναι η εύρεση υποκατάστατων μέτρων ευπάθειας για να εντοπιστούν τα πιο ευπαθή άτομα (συγκεκριμένα σκορ για παράδειγμα, σχετικά με την δραστηριότητα της καθημερινής ζωής). Αντίθετα, ο προσδιορισμός των ατόμων που είναι ευπαθή, φαίνεται να έχει μικρότερη σημασία για τα μοντέλα ευπάθειας στη βιοστατιστική και τη δημογραφία (Morley et al., (2002)).

Υπάρχουν διάφορες επιλογές για την κατανομή της ευπάθειας δηλαδή της τυχαίας μεταβλητής "Z". Μερικά βασικά αποτελέσματα μπορούν να βρεθούν στο έργο των Vaupel & Yashin (1985). Αξίζει να σημειωθεί ότι ένας παραμετρικός προσδιορισμός των κατανομών της ευπάθειας είναι καθαρά θέμα μαθηματικής ευκολίας. Ένας τέτοιος προσδιορισμός δεν είναι όμως και αναγκαίος. Για παράδειγμα στις οικονομικές επιστήμες, αυτό είναι ολόκληρο πεδίο ερευνών. Για παράδειγμα οι Heckman & Singer (1984) εκτιμούν την κατανομή της ευπάθειας μη παραμετρικά αλλά βάσει παραμετρικών υποθέσεων για τον βασικό κίνδυνο.

Ο Horowitz (1999) προτείνει στρατηγικές για να εκτιμηθεί τόσο η συνάρτηση κατανομής της ευπάθειας όσο και ο βασικός κίνδυνος μη παραμετρικά. Προκύπτει γενικά ότι

είναι πολύ δύσκολο να εκτιμήσουμε την συνάρτηση κατανομής της ευπάθειας σε εφαρμογές με πραγματικά δεδομένα, όπως μπορούμε να δούμε στις εργασίες των Heckman and Taber (1994), Kortram et al. (1995), and Horowitz (1999).

Η προσέγγιση που θα ακολουθήσουμε σε αυτή την εργασία θα είναι να απαλείψουμε την εξάρτηση από την συμμεταβλητή που είναι άγνωστη, την ευπάθεια "Z" δηλαδή. Για το σκοπό αυτό θα ολοκληρώσουμε την δεσμευμένη συνάρτηση επιβίωσης ως προς την ευπάθεια παίρνοντας ένα “μέσο όρο” ως προς τις τιμές της. Θα χρειαστεί να υποθέσουμε μία κατανομή για την ευπάθεια αλλά η μέθοδος αυτή εφαρμόζεται για οποιαδήποτε κατανομή. Η γενικευμένη συνάρτηση πιθανοφάνειας στην οποία θα βασιστούμε για την εκτίμηση των παραμέτρων (Likelihood function) δεν έχει επίσης άμεση εξάρτηση από την τυχαία μεταβλητή "Z". Η γενικευμένη συνάρτηση πιθανοφάνειας μπορεί να μεγιστοποιηθεί ως συνήθως αλλά λόγω της πολυπλοκότητάς της θα πρέπει να εφαρμοστούν αριθμητικές μέθοδοι μεγιστοποίησης ή μέθοδοι Monte Carlo. Λεπτομέρειες της μεθόδου θα δοθούν πιο κάτω.

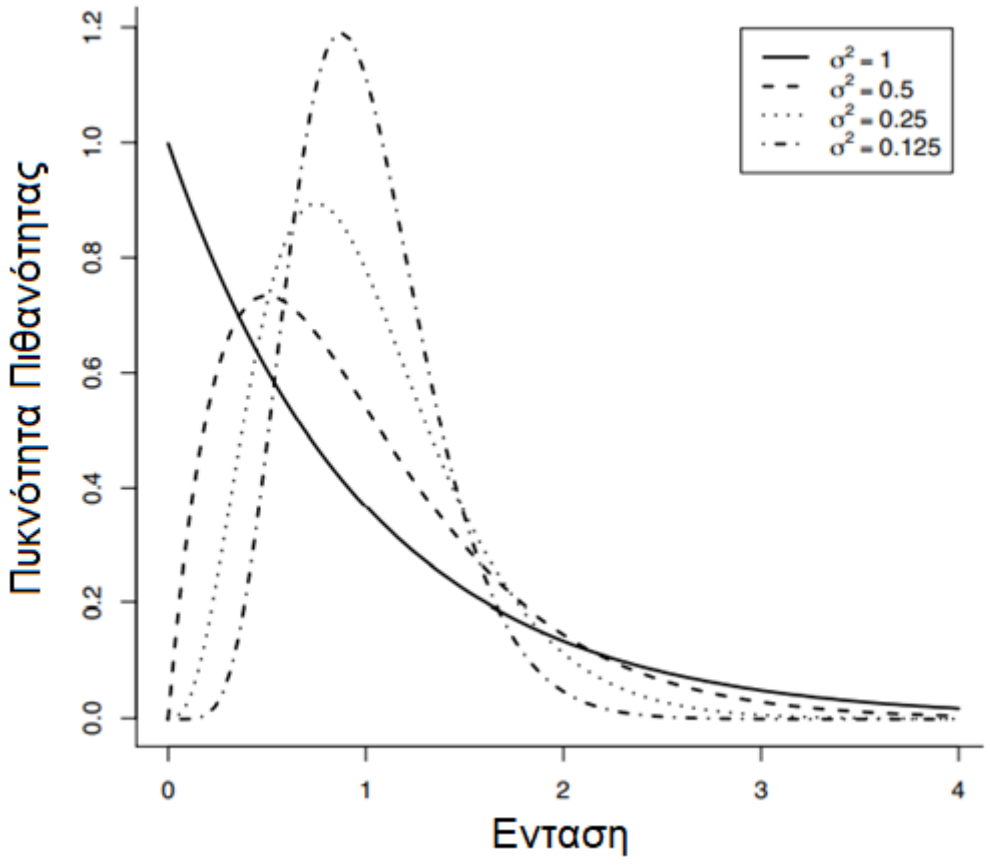
Οι πιο ευρέως χρησιμοποιούμενες κατανομές ευπάθειας στην ανάλυση επιβίωσης είναι οι Gamma, Inverse Gaussian και η Log-Normal. Μεγαλύτερες οικογένειες κατανομών δίνουν καλύτερες προτυποποιήσεις αλλά με το υπολογιστικό κόστος των περισσοτέρων παραμέτρων και πιο σύνθετων μοντέλων. Αυτό το πρόβλημα χρειάστηκε να αντιμετωπιστεί και σε αυτή την εργασία.

Υπάρχουν λίγες βιβλιογραφικές αναφορές εργασιών που συγκρίνουν μοντέλα με διαφορετικές κατανομές ευπάθειας. Είναι λοιπόν επιτακτική η ανάγκη τέτοιων μελετών σαν αντικείμενο έρευνας.

3.2.3. Το Gamma Μοντέλο Ευπάθειας

Η Gamma κατανομή, έχει χρησιμοποιηθεί εκτενώς σαν κατανομή της ευπάθειας όπως φαίνεται στα έργα των Greenwood and Yule (1920), Beard (1959), Vaupel et al. (1979), Congdon (1995), dos Santos et al. (1995), Hougaard (2000), Duchateau and Janssen (2008). Από υπολογιστική και αναλυτική σκοπιά, αυτή η κατανομή είναι πολύ χρήσιμη καθώς μπορούμε πολύ εύκολα να εξάγουμε εκφράσεις κλειστής μορφής για την μη δεσμευμένη επιβίωση και την συνάρτηση διακινδύνευσης. Αυτό συμβαίνει λόγω της απλότητας του μετασχηματισμού Laplace.

Αυτός είναι και ο λόγος που η κατανομή έχει χρησιμοποιηθεί στις περισσότερες εφαρμογές που έχουν δημοσιευθεί μέχρι και σήμερα. Η Gamma κατανομή $\Gamma(\kappa, \lambda)$ είναι μία ευέλικτη κατανομή που παίρνει πληθώρα σχημάτων καθώς το κ μεταβάλλεται. Όταν το $\kappa = 1$, είναι η γνωστή σε όλους εκθετική κατανομή και όταν το κ είναι μεγάλο, παίρνει το σχήμα της γνωστής σε όλους κανονικής κατανομής όπως φαίνεται στην Εικόνα 3 πιο κάτω.



Εικόνα 3: Συνάρτηση πυκνότητας πιθανότητας για όλες τις Gamma κατανομές με μέση τιμή 1 και τυπική απόκλιση 1, 0.5, 0.25 και 0.125. (Winke, 2010)

Παρ' όλα τα πλεονεκτήματα που έχει πρακτικά αυτή η συνάρτηση, δεν υπάρχει κάποιος βιολογικός λόγος που να καθιστά την συνάρτηση Gamma πιο προτιμητέα από τις υπόλοιπες κατανομές ευπάθειας. Σχεδόν όλα τα επιχειρήματα υπέρ της κατανομής Gamma βασίζονται σε μαθηματικά και υπολογιστικά στοιχεία. Στο έργο τους, οι Abbring and Van den Berg (2007) εκλογικεύουν την χρήση της Gamma κατανομής για ευπάθειες σε αναλύσεις δεδομένων επιβίωσης (για χρόνους μέχρι κάποιο γεγονός). Οι συγγραφείς δείχνουν ότι για μεγάλη κλάση μονομεταβλητών μοντέλων ευπάθειας, η κατανομή της ευπάθειας στους επιζήσαντες συγκλίνει στην Gamma κατανομή για χρόνους που τείνουν στο άπειρο.

Η ευπάθεια δεν μπορεί να είναι αρνητική και η Gamma κατανομή είναι μια από τις πιο πολυχρησιμοποιημένες κατανομές για να μοντελοποιηθούν μεταβλητές που είναι μη αρνητικές. Η πυκνότητα μιας Gamma-κατανεμημένης μεταβλητής με παράμετρο θέσης κ και παράμετρο ρυθμού λ δίνεται από την σχέση

$$f(z) = \frac{1}{\Gamma(\kappa)}\lambda^{\kappa} z^{\kappa-1} e^{-\lambda z}$$

Ως εκ τούτου, από τον μετασχηματισμό Laplace, ισχύει ότι:

$$L(u) = \frac{1}{\Gamma(\kappa)}\lambda^{\kappa}\int e^{-uz}\, z^{\kappa-1} e^{-\lambda z} dz$$

$$L(u) = \frac{\lambda^{\kappa}}{(\lambda+u)^{k}}\frac{1}{\Gamma(\kappa)}(\lambda+u)^{\kappa}\int z^{\kappa-1} e^{-(\lambda+u)z} dz$$

$$L(u) = (1+\frac{u}{\lambda})^{-\kappa}$$

Η συνέχεια στα τελευταία βήματα είναι συνέπεια της ολοκλήρωσης της πυκνότητας πιθανότητας της Gamma κατανομής με παραμέτρους κ και $\lambda + u$ η οποία δίνει αποτέλεσμα 1. Η πρώτη και δεύτερη παράγωγος του μετασχηματισμού Laplace Είναι:

$$L'(u) = -\frac{\kappa}{\lambda}(1+\frac{u}{\lambda})^{-k-1}$$

$$L''(u) = \frac{\kappa(\kappa+1)}{\lambda^2}(1+\frac{u}{\lambda})^{-k-2}$$

Για να είμαστε σίγουροι ότι το μοντέλο δεν θα έχει προβλήματα αναγνωρισιμότητας, έχει επιβληθεί ο περιορισμός $\kappa = \lambda$ στην Gamma κατανομή ο οποίος οδηγεί σε $EZ = 1$ όπως συζητήσαμε πιο πριν σε αυτό το κεφάλαιο. Ας σημειωθεί επίσης ότι η διασπορά της μεταβλητής της ευπάθειας είναι $\sigma^2 = \frac{1}{\lambda}$.

Η συνάρτηση πυκνότητας πιθανότητας της Gamma κατανεμημένης άγνωστης μεταβλητής $Z \sim \Gamma(\frac{1}{\sigma^2}, \frac{1}{\sigma^2})$ δίνεται απο την σχέση:

$$f(z) = \frac{1}{\Gamma\left(\frac{1}{\sigma^2}\right)}\left(\frac{1}{\sigma^2}\right)^{\frac{1}{\sigma^2}} z^{\frac{1}{\sigma^2}-1} exp\left(-\frac{z}{\sigma^2}\right) \quad (21)$$

και απεικονίζεται στην Εικόνα-3. Η μη δεσμευμένη συνάρτηση επιβίωσης μπορεί να προκύψει από τον μετασχηματισμό Laplace:

$$S(t) = L\big(H_0(t)\big) = \frac{1}{(1+\sigma^2 H_0(t))^{\frac{1}{\sigma^2}}}$$

Αυτό υποδεικνύει ότι η συνάρτηση πυκνότητας πιθανότητας είναι η:

$$f(t) = \frac{h_0(t)}{(1+\sigma^2 H_0(t))^{\frac{1}{\sigma^2}+1}}$$

και η συνάρτηση διακινδύνευσης:

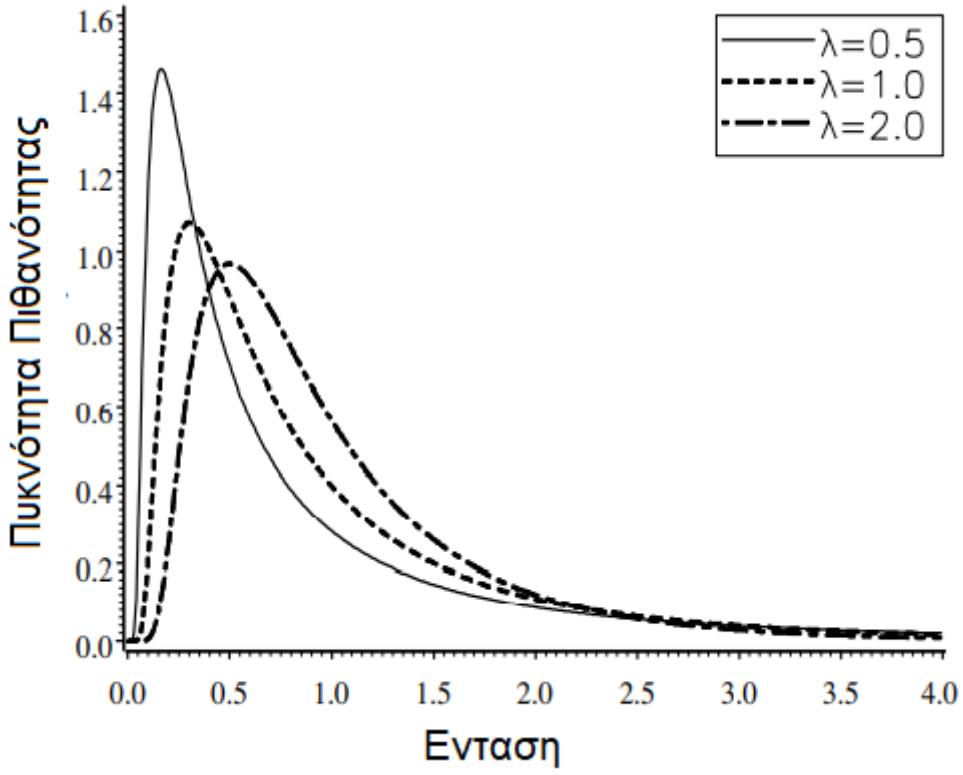
$$h(t) = \frac{h_0(t)}{1+\sigma^2 H_0(t)}$$

3.2.4. Το Inverse Gaussian Μοντέλο Ευπάθειας

Στην θεωρία πιθανοτήτων, η Inverse Gaussian Κατανομή (ή αλλιώς, κατανομή Wald), είναι μία οικογένεια συνεχών κατανομών στο $(0, \infty)$ με δύο παραμέτρους. Η συνάρτηση πυκνότητας πιθανότητας, από το έργο των Androulakis et al. (2012), δίνεται από την σχέση:

$$f_u(u) = \sqrt{\frac{b_2}{\pi u^3}} exp\left(\sqrt{4b_1 b_2}\right) exp\left(-b_1 u - b_2\left(\frac{1}{u}\right)\right) \qquad (22)$$

με $b_1 \geq 0, \quad b_2 \geq 0$



Εικόνα 4: Η συνάρτηση πυκνότητας πιθανότητας της inverse Gaussian κατανομής με μέση τιμή 1 και τυπική απόκλιση 0.5, 1 και 2 (Winke, 2010)

Για λόγους αναγνωρισιμότητας, στην σχέση (22) υποθέτουμε $b_1 = b_2 = b$ έτσι ώστε η μέση τιμή της ευπάθειας να είναι 1 και η διασπορά της να είναι $\sigma^2 = \frac{1}{2b}$ (Androulakis et al. (2012)).

Η Inverse Gaussian Κατανομή έχει αρκετές ιδιότητες ανάλογες με την Gaussian Κατανομή. Το όνομα της βέβαια είναι παραπλανητικό καθώς είναι αντίστροφη (=*inverse*)

μόνο υπό την έννοια ότι ενώ η Gaussian περιγράφει το επίπεδο μιας κίνησης Brown σε μία συγκεκριμένη χρονική στιγμή, η Inverse Gaussian περιγράφει την κατανομή του χρόνου που χρειάζεται μια κίνηση Brown με θετική τάση για να φτάσει ένα καθορισμένο θετικό επίπεδο.

3.3. Τα Μοντέλα Επιβίωσης Σαν Μοντέλα Γραμμικής Παλινδρόμησης

Σε αυτό το σημείο θα δούμε πώς μια ποικιλία από μοντέλα ευπάθειας μπορεί να γραφτεί υπό μορφή μοντέλου γραμμικής παλινδρόμησης. Η μέθοδος ανάλυσης των χρόνων επιβίωσης μίας ομάδας ατόμων βασίζεται στην συνάρτηση κινδύνου που αντιπροσωπεύει την πιθανότητα θανάτου ενός ατόμου της ομάδας τη χρονική στιγμή t δεδομένου ότι το άτομο έχει επιβιώσει μέχρι το t.

Όπως ήδη προαναφέραμε, ένα ευρέως χρησιμοποιούμενο μοντέλο για αυτόν τον σκοπό είναι το μοντέλο αναλογικής διακινδύνευσης του Cox (1972) για το οποίο η συνάρτηση κινδύνου δίνεται από την σχέση (14):

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

όπου $\beta \in R^k$ το διάνυσμα των παραμέτρων παλινδρόμησης που είναι και η παράμετρος ενδιαφέροντος, X είναι ένα *k*-διάστατο διάνυσμα συμμεταβλητών, το οποίο για το άτομο i ορίζεται ως $X_i = (x_{i1}, x_{i2}, \dots, x_{ik})$ και $h_0(t)$ είναι μία άγνωστη βασική συνάρτηση διακυνδύνευσης. Αντίστοιχα, η συνάρτηση επιβίωσης, όπως είδαμε, δίνεται από τη σχέση (17):

$$S(t|X) = e^{-e^{\beta^T X} H(t)}$$

Μία γενική κλάση μοντέλων ευπάθειας προκύπτει φυσικά από το μοντέλο του Cox όταν, προκειμένου να ερμηνεύσει κάποιος την ετερογένεια του πληθυσμού, εισάγει μια μη παρατηρούμενη μεταβλητή ευπάθειας, την οποία από εδώ και πέρα θα συμβολίζουμε ως η, μέσα στο μοντέλο μέσω της συνάρτησης διακινδύνευσης.

Έστω ότι το διάνυσμα συμμεταβλητών που υπεισέρχονται στη συνάρτηση κινδύνου του i ατόμου περιλαμβάνει τόσο όρους γνωστούς όσο και όρους άγνωστους, πιο συγκεκριμένα, $X_i = (x_{i,Γνωστά}, x_{i,Αγνωστα})$. Έτσι, η σχέση (14) με την εισαγωγή του όρου της ευπάθειας παίρνει την μορφή:

$$h(t|X_i) = h_0(t)e^{\beta^T x_{i\,Γνωστά}} e^{\beta^T x_{i\,Αγνωστά}} \quad (23)$$

Θέτοντας

$$\eta = e^{\beta^T x_{i\ Αγνωστά}}, \quad (24)$$

από τις σχέσεις (23) και (24), προκύπτει ότι:

$$h(t|\eta, X_i) = h_0(t)\eta e^{\beta^T x_{i\,\Gamma\nu\omega\sigma\tau\acute{\alpha}}} \equiv h_0(t)\eta e^{\beta^T x_i}$$

Άρα γενικά:

$$h(t|\eta, X) = h_0(t)\eta e^{\beta^T x_i} \quad (25)$$

Το η είναι τυχαία μεταβλητή η οποία θεωρούμε ότι διαφέρει από άτομο σε άτομο.

Για την συνάρτηση επιβίωσης, όπως είδαμε ισχύει από την σχέση (10) αντίστοιχα ότι

$$S(t|\eta, X) = e^{-\int_0^t h(s|\eta,X)\,ds} = e^{-\eta e^{\beta^T X}\int_0^t h_0(s)ds} \quad (26)$$

Άρα:

$$S(t|\eta, X) = e^{-\eta e^{\beta^T X}\,H_0(t)} \quad (27)$$

Με αυτόν το τρόπο ορίζεται ένα μοντέλο μη αναλογικής διακινδύνευσης δεδομένης της ευπάθειας.

Οι πιο πάνω συναρτήσεις κινδύνου και επιβίωσης ((25) και (27) αντίστοιχα) ορίζονται δεδομένων των συμμεταβλητών και της ευπάθειας. Αφού όμως η ευπάθεια δεν είναι μετρήσιμη μεταβλητή για κάθε άτομο, η προσέγγιση που θα ακολουθήσουμε σε αυτή την εργασία θα είναι να απαλοίψουμε την εξάρτηση από την ευπάθεια. Έχει προταθεί (Vonta (1996)) να ολοκληρώσουμε την δεσμευμένη συνάρτηση επιβίωσης ως προς η. Με αυτόν τον τρόπο, παίρνουμε κατά κάποιον τρόπο τον "Μέσο όρο" της συνάρτησης επιβίωσης ως προς η:

$$S(t|X) = \int_0^\infty e^{-\eta e^{\beta^T X} H_0(t)}\, dF_\eta(\eta) \equiv e^{-G(e^{\beta^T X} H_0(t))} \quad (28)$$

όπου $F_\eta(.)$ είναι η συνάρτηση κατανομής (*CDF*) της θετικής τυχαίας μεταβλητής η, και η συνάρτηση G ορίζεται ως

$$G(w) = -ln\left(\int_0^\infty e^{-yw} dF_\eta(y)\right) \quad (29)$$

Έτσι λοιπόν, μένουμε με μία συνάρτηση επιβίωσης εξαρτώμενη μόνο από τις παρατηρούμενες ποσότητες. Οι σχέσεις (28) και (29) ορίζουν μια κλάση μοντέλων ευπάθειας για διάφορες κατανομές ευπάθειας. Αξίζει να σημειωθεί ότι το ευρέως γνωστό μοντέλο Clayton-Cuzick (Clayton and Cuzick 1986) προκύπτει όταν το "η" θεωρείται ότι ακολουθεί την Γάμμα κατανομή με μέση τιμή 1 και τυπική απόκλιση "b".

Η συνάρτηση G σε αυτήν την περίπτωση ορίζεται ως $G(w) = \ln\left[(1 + bw)^{\frac{1}{b}}\right]$. Στην περίπτωση που $G(w) = w$ το μοντέλο που προκύπτει είναι εκείνο της αναλογικής

διακυνδύνευσης του Cox. Περισσότερα παραδείγματα μπορούν να βρεθούν στο έργο των Hougaard (1986) και στης παραπομπές εντός.

Η συνάρτηση κατανομής του η και ως εκ τούτου η συνάρτηση G μπορεί να θεωρηθεί γνωστή εκτός ίσως από κάποια παράμετρο που για την περίπτωση της Γάμμα κατανομής είναι το b. Η συνάρτηση G θεωρείται επίσης να είναι αυστηρώς αύξουσα και κοίλη με $G(0) = 0$ και $G(\infty) = \infty$.

3.4. Μοντέλα Ευπάθειας που Εξετάζονται στην Παρούσα Εργασία

Θέτοντας το απαραίτητο θεωρητικό υπόβαθρο, σε αυτό το υποκεφάλαιο θα δούμε με ποιες περιπτώσεις συναρτήσεων κατανομής της ευπάθειας θα ασχοληθούμε σε αυτήν την εργασία.

Αναφέραμε ότι σαν $F_\eta(.)$ στις σχέσεις (28) και (29) ορίζουμε την συνάρτηση κατανομής (*CDF*) της θετικής τυχαίας μεταβλητής η ενώ η συνάρτηση G ορίζεται στη σχέση (29). Ανάλογα με την συνάρτηση κατανομής $F_\eta(.)$ που θα χρησιμοποιήσουμε για την μελέτη μας, αντίστοιχα θα αλλάζει και η συνάρτηση $G(.)$, ορίζοντας μια ολόκληρη κλάση μοντέλων ευπάθειας.

3.4.1. Ευπάθεια που Ακολουθεί την Gamma Κατανομή

Εάν η ευπάθεια η ακολουθεί την $Gamma(\frac{1}{b},\frac{1}{b})$ κατανομή, τότε από τις σχέσεις (21) και (29) προκύπτει οτι

$$G(w)_{Gamma} = \ln\left[(1+bw)^{\frac{1}{b}}\right] \qquad (30)$$

Να υπενθυμίσουμε εδώ ότι στην Gamma κατανομή θεωρούμε ίσες τις παραμέτρους και ίσες με $\frac{1}{b}$ για λόγους αναγνωρισιμότητας, έτσι ώστε η μέση τιμή της ευπάθειας να είναι 1 και η διασπορά της ίση με b (Androulakis et al. (2012), (Vonta and Karagrigoriou (2007)).

3.4.2. Ευπάθεια που ακολουθεί την Inverse Gaussian Κατανομή

Εάν η ευπάθεια η ακολουθεί την $Inverse\ Gaussian(b_1,b_2)$ κατανομή, τότε για λόγους αναγνωρισιμότητας θεωρούμε $b_1 = b_1 = b$ έτσι ώστε η μέση τιμή της ευπάθειας να είναι 1 και η διασπορά της ίση με $\frac{1}{2b}$. (Androulakis et al. 2012).

Έτσι, από τις σχέσεις (22) και (29) προκύπτει οτι:

$$G(w)_{Inv\ Gaussian} = -\ln\left(e^{2b-\sqrt{4b(b+w)}}\right) = -2b+\sqrt{4b(b+w)}$$

Άρα:

$$G(w)_{Inv\ Gaussian} = -2b+\sqrt{4b(b+w)} \qquad (31)$$

3.4.3. Γενικευμένη Πιθανοφάνεια και Εκτίμηση Παραμέτρων

Έστω τυχαίο δείγμα n χρόνων επιβίωσης λογοκριμένων πιθανώς από δεξιά, με δεδομένα όπως αναφέρθηκε και στην υπό-ενότητα 3.1.1, να δίνονται ως $Data = (T_i, \delta_i, X_i)$, $i = 1,2, \dots, n$ με $T_i = \min(T_i^*, C_i)$.

$$\text{όπου } \begin{cases} T_i^* & o\ χρόνος\ επιβίωσης \\ C_i & o\ χρόνος\ λογοκρισίας\ (Censoring) \end{cases}$$

Επίσης

$$\delta_i = I(T_i^* \leq C_i) \begin{cases} 1 & εάν\ T_i^* \leq C_i\ \ (μή\ λογοκριμένη\ παρατήρηση) \\ 0 & εάν\ T_i^* > C_i\ \ (λογοκριμένη\ παρατήρηση\ απο\ δεξιά) \end{cases}$$

η δείκτρια της λογοκρισίας και X_i οι συμμεταβλητές του υποκειμένου i, διάστασης $dim = k$.

Στην περίπτωσή μας, όπου δηλαδή τα δεδομένα επιβίωσης περιλαμβάνουν λογοκριμένες παρατηρήσεις από δεξιά, η συνάρτηση πιθανοφάνειας τροποποιείται ως εξής:

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

$$= \left\{ \prod_{i=1}^{n} \left((1 - \mathrm{G}(t_i|X_i))^{\delta_i} \, g(t_i|X_i)^{1-\delta_i} \right) \right\} \left\{ \prod_{i=1}^{n} \left(f(t_i|X_i)^{\delta_i} \, S(t_i|X_i)^{1-\delta_i} \right) \right\}$$

όπου

- T, C είναι τυχαίες και ανεξάρτητες μεταβλητές.
- f είναι η συνάρτηση πυκνότητας πιθανότητας του χρόνου επιβίωσης T
- S είναι η συνάρτηση επιβίωσης του χρόνου επιβίωσης T
- G είναι η αθροιστική συνάρτηση κατανομής του χρόνου λογοκρισίας C
- g είναι η συνάρτηση πυκνότητας πιθανότητας του χρόνου λογοκρισίας C.

Ο πρώτος όρος στην παραπάνω συνάρτηση πιθανοφάνειας μπορεί να παραλειφτεί λόγω του ότι δεν εξαρτάται από τις παραμέτρους που μας ενδιαφέρουν και έτσι προκύπτει ότι η συνάρτηση πιθανοφάνειας απλοποιείται ως (Klein & Moeschberger, (2003)) (Janssen & Duchateau, (2008)):

$$L = \prod_{i=1}^{n} \left(f(t_i|X_i)^{\delta_i} \, S(t_i|X_i)^{1-\delta_i} \right)$$

Για την οποία και ισχύει από ιδιότητες δυνάμεων ότι:

$$L = \prod_{i=1}^{n} \left(f(t_i|X_i)^{\delta_i} \, S(t_i|X_i)^{1-\delta_i} \right) = \prod_{i=1}^{n} \left(\frac{f(t_i|X_i)^{\delta_i}}{S(t_i|X_i)^{\delta_i}} \, S(t_i|X_i) \right) \quad (32)$$

Όμως, $h(t_i|X_i) = \frac{f(t_i|X_i)}{S(t_i|X_i)}$ εξ ορισμού, άρα η σχέση (32) γίνεται:

$$L = \prod_{i=1}^{n} \left(h(t_i|X_i)^{\delta_i} \, S(t_i|X_i) \right) \quad (33)$$

Επίσης, $h(t_i|X_i) = -\frac{S'(t_i|X_i)}{S(t_i|X_i)}$, άρα:

$$L = \prod_{i=1}^{n} \left(\left(-\frac{S'(t_i|X)}{S(t_i|X)} \right)^{\delta_i} \, S(t_i|X) \right) \quad (34)$$

Συνεπώς, η σχέση (34) μέσω της σχέσης (28) γίνεται:

$$L = \prod_{i=1}^{n} \left(\frac{e^{-G\left(e^{\beta^T X_i} H(t_i)\right)} \; G'(w)|_{w=e^{\beta^T X_i} H(t_i)} e^{\beta^T X_i} \, h(t_i)}{e^{-G\left(e^{\beta^T X_H(t_i)}\right)}} \right)^{\delta_\iota} e^{-G\left(e^{\beta^T X_i} H(t_i)\right)}$$

$$= \prod_{i=1}^{n} \left(G'(w)|_{w=e^{\beta^T X_i} H(t_i)} e^{\beta^T X_i} \, h(t_i) \right)^{\delta_\iota} e^{-G\left(e^{\beta^T X_i} H(t_i)\right)}$$

Άρα:

$$L = \prod_{i=1}^{n} \left(G'(w)|_{w=e^{\beta^T X_i} H(t_i)} e^{\beta^T X_i} \, h(t_i) \right)^{\delta_\iota} e^{-G\left(e^{\beta^T X_i} H(t_i)\right)} \quad (35)$$

Για λόγους υπολογιστικής ευκολίας στην μέθοδο μεγίστης πιθανοφάνειας, λογαριθμίζουμε και έτσι βρίσκεται ότι η Log-Likelihood της σχέσης (35) είναι:

$$log(L) = \sum_{i=1}^{n} \delta_\iota \left\{ \ln\left(G'(w)|_{w=e^{\beta^T X_i} H(t_i)} \right) + \beta^T X_i + \ln(h(t_i)) \right\} - G\left(e^{\beta^T X_i} H(t_i) \right) \quad (36)$$

Στην εργασία αυτή, για λόγους υπολογιστικής ευκολίας, θα υποθέσουμε από εδώ και πέρα ένα παραμετρικό μοντέλο ευπάθειας και πιο συγκεκριμένα θα υποθέσουμε εκθετική κατανομή με παράμετρο λ για τη βασική αθροιστική συνάρτηση κινδύνου $H(t_i)$ ή ισοδύναμα για την βασική συνάρτηση κινδύνου, δηλαδή,

$$H(t_i) = \lambda t_i \quad (37)$$

και

$$h(t_i) = \lambda \quad (38)$$

Η (οχληρά) παράμετρος λ θεωρείται άγνωστη και θα εκτιμηθεί μαζί με την παράμετρο ενδιαφέροντος β μέσω της μεθόδου μεγίστης πιθανοφάνειας.

3.4.4. Ο Τύπος της Πιθανοφάνειας για την Gamma Κατανομή

Στην περίπτωση της Gamma κατανομής, όπως προαναφέραμε, ισχύει από την σχέση (30):

$$G(w)_{Gamma} = \ln\{(1+bw)^{\frac{1}{b}}\}$$

και

$$G'(w) = \frac{1}{1+bw}$$

και

$$G''(w) = -\frac{b}{(1+bw)^2}$$

Άρα στην περίπτωσή μας, από την σχέση (30) και για $w = e^{\beta^T X_i} H(t_i)$ έχουμε ότι η συνάρτηση G που εμπλέκεται στην συνάρτηση πιθανοφάνειας δίνεται από:

$$G\left(e^{\beta^T X_i} H(t_i)\right)_{Gamma} = \ln\{\left(1+be^{\beta^T X_i} H(t_i)\right)^{\frac{1}{b}}\} \quad (39)$$

Για την περίπτωση της βασικής συνάρτησης κινδύνου από την εκθετική κατανομή:

$$G\left(e^{\beta^T X_i}\lambda t_i\right)_{Gamma} = \left(\frac{1}{b}\right) ln\left(1+be^{\beta^T X_i}\lambda t_i\right) \quad (40)$$

Παραγωγίζοντας τη συνάρτηση G ως προς w και μετά για $w = e^{\beta^T X_i}\lambda t_i$ προκύπτει ότι:

$$G'(w)_{Gamma}|_{w=e^{\beta^T X_i}\lambda t_i} = \frac{1}{be^{\beta^T X_i}\lambda t_i + 1} \quad (41)$$

Άρα, κατά αυτόν τον τρόπο, η Log-Likelihood στην περίπτωση της Gamma Κατανομής, από τις σχέσεις $(36), (40), (41)$ προκύπτει ότι είναι η:

$$log(L) = \sum_{i=1}^{n} \delta_i \left\{ \ln\left(\frac{1}{be^{\beta^T X_i}\lambda t_i + 1}\right) + \beta^T X_i + \ln(\lambda) \right\} - \left(\frac{1}{b}\right) ln\left(1 + be^{\beta^T X_i}\lambda t_i\right) \quad (42)$$

3.4.5. Ο Τύπος της Πιθανοφάνειας για την Inverse Gaussian Κατανομή

Στην περίπτωση της Inverse Gaussian κατανομής, όπως προαναφέραμε, από την σχέση (31), ισχύει ότι:

$$G(w)_{Inv\ Gaussian} = -2b + \sqrt{4b(b+w)}$$

και

$$G'(w) = \frac{\sqrt{b}}{\sqrt{b+w}}$$

και

$$G''(w) = -\frac{\sqrt{b}}{2(b+w)^{3/2}}$$

Άρα σε αυτή την περίπτωση, από την σχέση (30) και για $w = e^{\beta^T X} H(t_i)$ έχουμε ότι η συνάρτηση G που εμπλέκεται στην συνάρτηση πιθανοφάνειας δίνεται από:

$$G\left(e^{\beta^T X_i} H(t_i)\right)_{Inv\ Gaussian} = -2b + \sqrt{4b\left(b + e^{\beta^T X_i} H(t_i)\right)} \quad (43)$$

Για την περίπτωση της βασικής συνάρτησης κινδύνου από την εκθετική κατανομή:

$$G\left(e^{\beta^T X} \lambda t_i\right)_{Inv\ Gaussian} = -2b + \sqrt{4b\left(b + e^{\beta^T X_i} \lambda t_i\right)} \quad (44)$$

Παραγωγίζοντας τη συνάρτηση G ως προς w και μετά για $w = e^{\beta^T X_i} \lambda t_i$ προκύπτει ότι:

$$G'(w)_{Inv\ Gaussian}\Big|_{w=e^{\beta^T X_i} \lambda t_i} = \frac{\sqrt{b}}{\sqrt{\left(b + e^{\beta^T X_i}\ \lambda\ t_i\right)}} \quad (45)$$

Άρα, κατά αυτόν τον τρόπο, η Log-Likelihood στην περίπτωση της Inverse Gaussian, από τις σχέσεις (36), (44) και (45) Κατανομής είναι η:

$$log(L) = \sum_{i=1}^{n} \delta_i \left\{ \ln\left(\frac{\sqrt{b}}{\sqrt{\left(b + e^{\beta^T X_i}\, \lambda\, t_i\right)}} \right) + \beta^T X_i + \ln(\lambda) \right\} + 2b$$

$$- \sqrt{4b\left(b + e^{\beta^T X_i} \lambda t_i\right)} \quad (46)$$

Στο συγκεκριμένο σημείο, θα συζητήσουμε την εύρεση των εκτιμητριών μέγιστης πιθανοφάνειας για τα $(\beta_1,\ \beta_2, \dots,\ \beta_k)$ που είναι οι παράμετροι παλινδρόμησης και για το "λ" που είναι οι παράμετροι ενδιαφέροντος και η οχληρά παράμετρος αντίστοιχα. Να τονίσουμε ξανά ότι η συνάρτηση $H(t_i)$ είναι γνωστής συναρτησιακής μορφής εκτός της παραμέτρου λ .

Είναι γνωστό το πρόβλημα σε αυτά τα μοντέλα, ότι αν και ο αριθμός των παραμέτρων προς εκτίμηση μπορεί να μην είναι πολύ μεγάλος, υπάρχουν αντιξοότητες ως προς την ταυτόχρονη μεγιστοποίηση ως προς όλες τις παραμέτρους και κυρίως ως προς την παράμετρο b που εμπλέκεται στην διασπορά της ευπάθειας και στην ουσία καθορίζει αν υπάρχει ετερογένεια στον πληθυσμό. Οι ερευνητές χρησιμοποιούν ευρέως τον ΕΜ αλγόριθμο για την επίτευξη μεγιστοποίησης θεωρώντας τις ευπάθειες των ατόμων ως μη γνωστά δεδομένα (missing data) ενώ εμείς εναλλακτικά σε αυτή την εργασία προσπαθήσαμε αρχικά να μεγιστοποιήσουμε την γενικευμένη συνάρτηση πιθανοφάνειας (που η γενική μορφή της ισχύει για όλες τις κατανομές ευπάθειας) ως προς όλες τις παραμέτρους ταυτόχρονα. Λόγω υπολογιστικών προβλημάτων της μεγιστοποίησης οδηγηθήκαμε στο να ακολουθήσουμε έναν αλγόριθμο ο οποίος περιγράφεται παρακάτω και κατά τον οποίο η συνάρτηση πιθανοφάνειας μεγιστοποιείται έμμεσα ως προς την παράμετρο b μέσω αναζήτησης του μεγίστου σε ένα πιθανό εκ των προτέρων για το b Grid τιμών.

3.5. Βελτίωση Ελαχιστοποίησης Μέσω Αλγορίθμου για την Εκτίμηση των Παραμέτρων

Για να παρακάμψουμε το πρόβλημα που προκύπτει κυρίως από την μεγιστοποίηση ως προς b, προτείνουμε τη χρήση της παρακάτω μεθόδου που πρότειναν οι Fan και Li (2002).

Παραγωγίζοντας την σχέση (36) (Log-likelihood) ως προς λ για την περίπτωση της εκθετικής κατανομής προκύπτει η ακόλουθη σχέση (έχοντας συμβολίσει την παράγωγο ως προς w με «'» , συμβολίζουμε τώρα την παράγωγο ως προς λ με « · »):

$$\frac{\partial(log(L))}{\partial\lambda} = \sum_{i=1}^{n} \delta_\iota \left\{ \frac{G'^{\cdot}(w)|_{w=e^{\beta^T X_i}\lambda t_i}}{G'(w)|_{w=e^{\beta^T X_i}\lambda t_i}} + \frac{1}{\lambda} \right\} - G^{\cdot}(w)|_{w=e^{\beta^T X_i}\lambda t_i} = 0 \quad (47)$$

Ορίζοντας την παράγωγο, $\frac{\partial(log(L))}{\partial\lambda} = 0$ και λύνοντας την σχέση ως προς λ παίρνουμε το μέγιστο ως προς λ, για δεδομένα όμως β *και* b. Για τα μοντέλα ευπάθειας Gamma και Inverse Gaussian, οι συναρτήσεις $G^{\cdot}(\dots), G'^{\cdot}(\dots)$ υπολογίζονται πιο κάτω.

3.5.1. Βελτίωση Ελαχιστοποίησης Για την Gamma Κατανομή:

Όπως είδαμε, από την σχέση (40), για την Gamma κατανομή, η $G\left(e^{\beta^T X_i}\lambda t_i\right)_{Gamma}$ είναι:

$$G\left(e^{\beta^T X_i}\lambda t_i\right)_{Gamma} = \left(\frac{1}{b}\right) ln\left(1 + be^{\beta^T X_i}\lambda t_i\right)$$

Η πρώτη και δεύτερη μερική παράγωγος ως προς "λ" δίνονται από:

$$G^{\cdot}(w)_{Gamma}|_{w=e^{\beta^T X_i}\lambda t_i} = G'(w)|_{w=e^{\beta^T X_i}\lambda t_i} \cdot \frac{dw}{d\lambda} = \frac{e^{\beta^T X_i}\ t_i}{1 + be^{\beta^T X_i}\lambda t_i} \quad (48)$$

και από την:

$$G'^{\cdot}(w)_{Gamma}|_{w=e^{\beta^T X_i}\lambda t_i} = G''(w)|_{w=e^{\beta' X}\lambda t_i} \cdot \frac{dw}{d\lambda} = -\frac{be^{\beta^T X_i}\ t_i}{(1 + be^{\beta^T X_i}\lambda t_i)^2} (49)$$

Συνεπώς, στην περίπτωση της Gamma συνάρτησης, η νέα σχέση της οποίας αναζητούμε τη λύση ως προς λ προκύπτει από τις σχέσεις (47), (48) και (49) και είναι η:

$$\sum_{i=1}^{n} \delta_\iota \left\{ \frac{-\frac{b e^{\beta^T X_i}\, t_i}{\left(1 + b e^{\beta^T X_i} \lambda t_i\right)^2}}{\frac{1}{1 + b e^{\beta^T X_i} \lambda t_i}} + \frac{1}{\lambda} \right\} - \frac{e^{\beta^T X_i}\, t_i}{1 + b e^{\beta^T X_i} \lambda t_i} = 0 \;\; \rightarrow$$

$$\sum_{i=1}^{n} \delta_\iota \left\{ -\frac{b e^{\beta^T X_i}\, t_i}{1 + b e^{\beta^T X_i} \lambda t_i} + \frac{1}{\lambda} \right\} - \frac{e^{\beta^T X_i}\, t_i}{1 + b e^{\beta^T X_i} \lambda t_i} = 0 \qquad (50)$$

3.5.2. Βελτίωση Ελαχιστοποίησης Για την Inverse Gaussian Κατανομή:

Όπως είδαμε, από την σχέση (44), για την Inverse Gaussian κατανομή, η $G\left(e^{\beta^T X_i} \lambda t_i\right)_{Inv\ Gaussian}$ είναι:

$$G\left(e^{\beta^T X_i} \lambda t_i\right)_{Inv\ Gaussian} = -2b + \sqrt{4b\left(b + e^{\beta^T X_i} \lambda t_i\right)}$$

της οποίας η πρώτη και δεύτερη μερική παράγωγος ως προς “λ” είναι:

$$G^{\cdot}(w)_{Inv\ Gaussian}\big|_{w = e^{\beta^T X_i} \lambda t_i} = G'(w)\big|_{w = e^{\beta^T X_i} \lambda t_i} \cdot \frac{dw}{d\lambda} = \frac{\sqrt{b} e^{\beta^T X_i}\, t_i}{\sqrt{\left(b + e^{\beta^T X_i}\, \lambda\, t_i\right)}} \qquad (51)$$

και

$$G^{\prime\cdot}(w)_{Inverse\ Gaussian}\big|_{w = e^{\beta^T X_i} \lambda t_i} = G''(w)\big|_{w = e^{\beta^T X_i} \lambda t_i} \cdot \frac{dw}{d\lambda} = -\frac{\sqrt{b} e^{\beta^T X_i}\, t_i}{2\left(b + e^{\beta^T X_i}\, \lambda\, t_i\right)^{\frac{3}{2}}} \qquad (52)$$

Συνεπώς, στην περίπτωση της Inverse Gaussian συνάρτησης, η νέα μη γραμμική εξίσωση η οποία πρέπει να λυθεί ως προς λ , από τις σχέσεις (47), (51) και (52) δίνεται από την σχέση:

$$\sum_{i=1}^{n} \delta_\iota \left\{ \frac{-\frac{\sqrt{b} e^{\beta^T X_i}\ t_i}{2\left(b + t_i\, e^{\beta^T X_i} \lambda\right)^{\frac{3}{2}}}}{\frac{\sqrt{b}}{\sqrt{\left(b + e^{\beta^T X_i}\, \lambda\, t_i\right)}}} + \frac{1}{\lambda} \right\} - \frac{\sqrt{b} e^{\beta^T X_i}\ t_i}{\sqrt{\left(b + e^{\beta^T X_i}\, \lambda\, t_i\right)}} = 0$$

ή ισοδύναμα:

$$\sum_{i=1}^{n} \delta_\iota \left\{ -\frac{e^{\beta^T X_i}\ t_i}{2\left(b + t_i\, e^{\beta^T X_i} \lambda\right)} + \frac{1}{\lambda} \right\} - \frac{\sqrt{b}\, e^{\beta^T X_i}\ t_i}{\sqrt{\left(b + t_i\, e^{\beta^T X_i} \lambda\right)}} = 0 \quad (53)$$

Δίνουμε εδώ επίσης την παράγωγο της συνάρτησης πιθανοφάνειας που δίνεται στην (36) ως προς β_j, *j=1, ..., k:*

$$\frac{\partial(log(L))}{\partial \beta_j} = \sum_{i=1}^{n} \delta_\iota \left\{ \frac{G''(w)|_{w=e^{\beta^T X_i \lambda t_i}}}{G'(w)|_{w=e^{\beta^T X_i \lambda t_i}}} \frac{dw}{d\beta_j} + X_{ik} \right\} - G'(w)|_{w=e^{\beta^T X_i \lambda t_i}} \frac{dw}{d\beta_j} =$$

$$\sum_{i=1}^{n} \delta_\iota \left\{ \frac{G''(w)|_{w=e^{\beta^T X_i \lambda t_i}}}{G'(w)|_{w=e^{\beta^T X_i \lambda t_i}}} e^{\beta^T X_i}\, \lambda\, t_i X_{ij} + X_{ij} \right\} - G'(w)|_{w=e^{\beta^T X_i \lambda t_i}} e^{\beta^T X_i}\, \lambda\, t_i X_{ij}$$

και τελικά το μέγιστο ως προς β βρίσκεται από το σύστημα των k εξισώσεων:

$$\sum_{i=1}^{n} \delta_\iota \left\{ \frac{G''(w)|_{w=e^{\beta^T X_i \lambda t_i}}}{G'(w)|_{w=e^{\beta^T X_i \lambda t_i}}} e^{\beta^T X_i}\, \lambda\, t_i + 1 \right\} - G'(w)|_{w=e^{\beta^T X_i \lambda t_i}} e^{\beta^T X_i}\, \lambda\, t_i) X_{ij} = 0$$

Ή

$$\sum_{i=0}^{n} \left[\delta_\iota \left\{ \frac{G''(w)|_{w=e^{\beta^T X_i \lambda t_i}}}{G'(w)|_{w=e^{\beta^T X_i \lambda t_i}}} e^{\beta^T X_i}\, \lambda\, t_i + 1 \right\} - G'(w)|_{w=e^{\beta^T X_i \lambda t_i}} e^{\beta^T X_i}\, \lambda\, t_i \right] X_{ij} = 0 \quad (54)$$

για j=1,..,k.

Για την περίπτωση της Inverse Gaussian ευπάθειας το σύστημα (54) γίνεται:

$$\sum_{i=0}^{n}\left[\delta_\iota\left\{\frac{-\frac{\sqrt{b}}{2(b+e^{\beta^T X_i}\lambda\, t_i)^{3/2}}\, e^{\beta^T X_i}\lambda\, t_i+1}{\frac{\sqrt{b}}{\sqrt{(b+e^{\beta^T X_i}\lambda\, t_i)}}}\right\}-\frac{\sqrt{b}}{\sqrt{(b+e^{\beta^T X_i}\lambda\, t_i)}}e^{\beta^T X_i}\lambda\, t_i\right]X_{ij}=0$$

$$\Rightarrow$$

$$\sum_{i=0}^{n}\left[\delta_\iota\left\{-\frac{e^{\beta^T X_i}\lambda\, t_i}{2(b+e^{\beta^T X_i}\lambda\, t_i)}+1\right\}-\frac{\sqrt{b}e^{\beta^T X_i}\lambda\, t_i}{\sqrt{(b+e^{\beta^T X_i}\lambda\, t_i)}}\right]X_{ij}=0\ \Rightarrow$$

$$\sum_{i=0}^{n}\left[\delta_\iota\left\{\frac{e^{\beta^T X_i}\lambda\, t_i+2b}{2(b+e^{\beta^T X_i}\lambda\, t_i)}\right\}-\frac{\sqrt{b}e^{\beta^T X_i}\lambda\, t_i}{\sqrt{(b+e^{\beta^T X_i}\lambda\, t_i)}}\right]X_{ij}=0\ \Rightarrow\ (55)$$

Για j=1,..,k. Μαζί με την (53) που απλοποιούμε εδώ έχουμε ένα σύστημα με k+1 εξισώσεις:

$$\sum_{i=1}^{n}\delta_\iota\left\{\frac{e^{\beta^T X_i}\lambda\ t_i+2b}{2\lambda(b+\ e^{\beta^T X_i}\lambda t_i)}\right\}-\frac{\sqrt{b}\,e^{\beta^T X_i}\ t_i}{\sqrt{(b+\ e^{\beta^T X_i}\lambda t_i)}}=0\quad(56)$$

Για την περίπτωση της Gamma ευπάθειας η (54) γίνεται:

$$\sum_{i=0}^{n}\left[\delta_\iota\left\{\frac{-\frac{b}{(1+be^{\beta^T X_i}\lambda\, t_i)^2}\, e^{\beta^T X_i}\lambda\, t_i+1}{\frac{1}{1+be^{\beta^T X_i}\lambda\, t_i}}\right\}-\frac{1}{1+be^{\beta^T X_i}\lambda\, t_i}e^{\beta^T X_i}\lambda\, t_i\right]X_{ij}=0\quad\Rightarrow$$

$$\sum_{i=0}^{n}\left[\delta_\iota\left\{-\frac{be^{\beta^T X_i}\lambda\, t_i}{1+be^{\beta^T X_i}\lambda\, t_i}+1\right\}-\frac{e^{\beta^T X_i}\lambda\, t_i}{1+be^{\beta^T X_i}\lambda\, t_i}\right]X_{ij}=0\Rightarrow$$

$$\sum_{i=0}^{n}\left[\delta_{\iota}\left\{\frac{1}{1+be^{\beta^T X_i}\,\lambda\,t_i}\right\}-\frac{e^{\beta^T X_i}\,\lambda\,t_i}{1+be^{\beta^T X_i}\,\lambda\,t_i}\right]X_{ij}=0 \quad (57)$$

Για j=1, …, k. Μαζί με την (50) που απλοποιούμε λίγο εδώ, προκύπτει ένα σύστημα με k+1 εξισώσεις:

$$\sum_{i=1}^{n}\delta_{\iota}\left\{\frac{1}{\lambda(1+be^{\beta^T X_i}\lambda t_i)}\right\}-\frac{e^{\beta^T X_i}\;t_i}{1+be^{\beta^T X_i}\lambda t_i}=0 \qquad (58)$$

3.5.3. Ο Αλγόριθμος

Τα βήματα του αλγορίθμου που χρησιμοποιούμε για την εκτίμηση των παραμέτρων β, λ και b, με βάση τη γενικευμένη συνάρτηση πιθανοφάνειας (36) και μέσω της μεθόδου μεγίστης πιθανοφάνειας, για τον οποίο δημιουργήθηκε κώδικας στην R, αναλυτικά, είναι τα εξής:

Βήμα-1:

Χρησιμοποιώντας την σχέση (16), εφαρμόζουμε το μοντέλο αναλογικής Διακινδύνευσης του Cox στα δεδομένα, έχοντας σαν χρόνο τον "*χρόνο επιβίωσης*" συνοδευόμενο από την "*Λογοκρισία*" και σαν συμμεταβλητές τις υπόλοιπες στήλες των δεδομένων. Με αυτόν τον τρόπο υπολογίζουμε τους αρχικούς συντελεστές παλινδρόμησης β_0 και την αρχική παράμετρο λ_0.

Ο λόγος που το κάνουμε αυτό είναι διότι τα μοντέλα ευπάθειας έχουν σαν ειδική περίπτωση το μοντέλο αναλογικής διακινδύνευσης του Cox για $G(w)=w$. Είναι λοιπόν λογικό αυτές οι εκτιμήσεις να χρησιμοποιηθούν στον αλγόριθμο σαν αρχικές τιμές, καθώς και οι καινούριες τιμές θα κυμαίνονται γύρω από τις τιμές της περίπτωσης του μοντέλου του Cox.

Πρέπει να θυμίσουμε εδώ όμως ότι το μοντέλο του Cox είναι γενικά ένα ημιπαραμετρικό μοντέλο αλλά μπορεί να οριστεί και ως παραμετρικό αν υποθέσει ο ερευνητής συγκεκριμένη παραμετρική μορφή για τη συνάρτηση κινδύνου. Θα μπορούσαμε

λοιπόν εδώ να υποθέσουμε εκθετική μορφή για τη συνάρτηση κινδύνου του Cox για να πάρουμε αρχικές τιμές αφού στη συνέχεια θα εξετάσουμε παραμετρικά μοντέλα ευπάθειας με εκθετική συνάρτηση κατανομής της συνάρτησης κινδύνου.

Βήμα-2:

Θέτουμε ένα Grid Τιμών για την παράμετρο "b" η οποία σχετίζεται με τη διασπορά της ευπάθειας και η οποία θα αυξάνεται για κάθε βήμα (Step), "κ". Στην συγκεκριμένη περίπτωση $b_k = b_0 + \ Step \ * \ \kappa$ όπου για λόγους ακρίβειας θέσαμε $Step = 0{,}1$.

Βήμα-3:

Χρησιμοποιώντας τους συντελεστές παλινδρόμησης β_0, και δεδομένη τιμή της παραμέτρου b_k, αντικαθιστούμε τα προαναφερθέντα και λύνουμε την $\frac{\partial(log(L))}{\partial\lambda} = 0$ (σχέσεις (50) και (53) για το Gamma και Inverse Gaussian μοντέλο αντίστοιχα) ως προς λ. Την τιμή αυτή του λ τη συμβολίζουμε με λ_1. Να σημειώσουμε εδώ ότι η τιμή λ_0 δεν χρησιμοποιήθηκε άμεσα στον αλγόριθμο αλλά θα μπορούσε να χρησιμοποιηθεί και να οδηγηθεί ο αλγόριθμος κατ΄ευθείαν στο Βήμα-4 για να πάρουμε την επόμενη τιμή για το β , δηλαδή το β_1 για δεδομένο $\lambda = \lambda_0$.

Σε αυτό το σημείο αντικαθιστούμε το λ_1 σαν καινούρια τιμή του λ και την χρησιμοποιούμε σαν input για το επόμενο βήμα.

Βήμα-4:

Χρησιμοποιώντας λοιπόν την τιμή λ_1, και την δεδομένη τιμή για το b_k, αντικαθιστούμε τα προαναφερθέντα και λύνουμε την $\frac{\partial(log(L))}{\partial\beta} = 0$ (σχέσεις (56) και (57) για το Γάμμα και Inverse Gaussian μοντέλο αντίστοιχα) ως προς β. Η αρχική τιμή β_0 χρησιμοποιείται και σαν αρχική τιμή στην εντολή constroptim της R η οποία καλείται για να βρεθεί το ελάχιστο ως προς β της $-log(L)$. Την τιμή αυτή του β που βρίσκουμε στο Βήμα-4 τη συμβολίζουμε με β_1.

Σε αυτό το σημείο αντικαθιστούμε το καινούριο β_1. σαν καινούρια τιμή του β και την χρησιμοποιούμε σαν input για το επόμενο βήμα.

Βήμα-5:

Ο αλγόριθμος συνεχίζει με εναλλαγές μεταξύ του βήματος 3 και 4 για δεδομένο b_k, λαμβάνοντας μία ακολουθία από παραμέτρους (λ_2, β_2), $(\lambda_3, \beta_3), \dots$, $(\lambda_{\mathrm{k}}, \beta_{\mathrm{k}})$ Ο αλγόριθμος σταματάει όταν για δύο διαδοχικά βήματα ισχύει ότι:

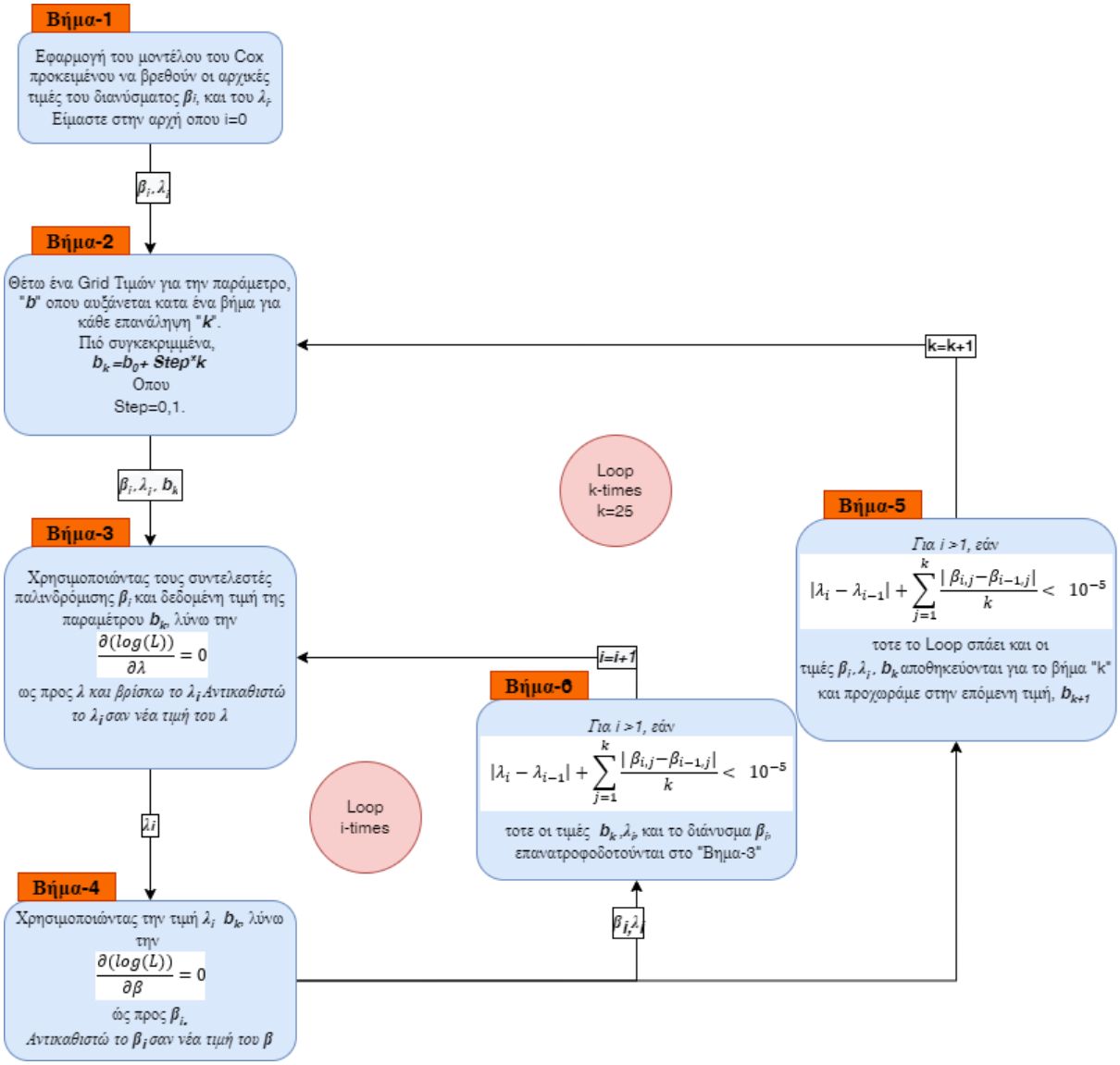
$$|\lambda_i - \lambda_{i-1}| + \sum_{j=1}^{k} \frac{|\beta_{i,j} - \beta_{i-1,j}|}{p} < 10^{-5} \qquad (59)$$

Βήμα-6:

Αποθηκεύουμε το διάνυσμα β_i που βρέθηκε από το Βήμα-4 και την τιμή λ_i που βρέθηκε από το Βήμα-3 τα οποία ικανοποιούν την (59) και αποτελούν τις εκτιμήτριες μεγίστης πιθανοφάνειας των β και λ για δεδομένο b_k από το Grid του Βήματος-2. Επαναλαμβάνουμε τα Βήματα-3 και 4 για την επόμενη τιμή b_{k+1} του Grid μέχρι να εξαντληθούν όλες οι τιμές του b στο grid. Σκοπός μας είναι να εντοπίσουμε το μέγιστο και ως προς την παράμετρο b. Αν το μέγιστο δεν εντοπιστεί στο grid που έχουμε θέσει, συνεχίζουμε θέτοντας περισσότερες τιμές στο grid μέχρι να εντοπιστεί το μέγιστο.

Έμπνευση για αυτό τον αλγόριθμο αποτελεί το έργο των Fan και Li (2002), από όπου και υιοθετήσαμε την ιδέα του Grid πιθανών τιμών για τις τιμές του b. Η μέθοδος αυτή ακολουθείται και στο έργο των Androulakis et al. (2012). Η ανάγκη για την υιοθέτηση του Grid είναι οι δυσκολίες στη μεγιστοποίηση ως προς την παράμετρο b που υπεισέρχεται στην συνάρτηση $G(x)$.

Το σύνολο των βημάτων μπορεί να φανεί και στην Εικόνα-4 παρακάτω:



Εικόνα 4: Τα βήματα του αλγορίθμου που περιγράψαμε παραπάνω.

Το τελικό output του αλγορίθμου είναι ένα σετ υπολογισμένων παραμέτρων, δηλαδή των εκτιμητριών μεγίστης πιθανοφάνειας $\hat{b}, \hat{\lambda}, \hat{\beta}$.

3.6. Ημι-Παραμετρικά Μοντέλα Ευπάθειας και ο Αλγόριθμος ΕΜ

Προκειμένου να εκτιμηθούν οι παράμετροι στο ημι-παραμετρικό μοντέλο επιβίωσης του Cox το οποίο δεν περιλαμβάνει μεταβλητή ευπάθειας, χρησιμοποιείται η μέθοδος μέγιστης πιθανοφάνειας με βάση τη μερική πιθανοφάνεια (Cox 1975). Ωστόσο, στην περίπτωση των ημι-παραμετρικών μοντέλων ευπάθειας, θα πρέπει να λάβουμε υπόψη τη συμμετοχή του παράγοντα ευπάθειας στο μοντέλο.

Η εκτίμηση των παραμέτρων σε αυτή την περίπτωση γίνεται με τη χρήση του ΕΜ αλγόριθμου (Expectation Maximization algorithm). Η βασική συνάρτηση κινδύνου στην περίπτωση των ημιπαραμετρικών μοντέλων, θεωρείται μία (άγνωστη) οχληρή παράμετρος (nuisance parameter) όταν δεν κάνουμε κάποια υπόθεση για την κατανομή της. Θεωρούμε λοιπόν, ότι δεν υπάρχει για αυτήν καμία παρατηρούμενη πληροφορία.

Ο ΕΜ αλγόριθμος αποτελείται από δύο βήματα:

1. Το Expectation step και
2. Το Maximization step.

Στο expectation step, υπολογίζονται οι αναμενόμενες τιμές των μη παρατηρούμενων μεταβλητών ευπάθειας, δεδομένων των παρατηρήσεων και των εκτιμημένων παραμέτρων. Στο Maximization step οι αναμενόμενες τιμές που έχουν υπολογιστεί λαμβάνονται υπόψη ως πραγματικές και χρησιμοποιούνται σαν input για τον υπολογισμό νέων εκτιμήσεων των παραμέτρων μεγιστοποιώντας την συνάρτηση πιθανοφάνειας. Παρακάτω θα παρουσιάσουμε τον ΕΜ αλγόριθμο για το Gamma μοντέλο ευπάθειας που είναι και το default σε βιβλιοθήκη της R “Frailtyem” (Janssen & Duchateau, 2008) (Klein & Moeschberger, 2003) (Wienke, 2011).

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

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

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

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

όπου $Z = (Z_1, Z_2, \dots, Z_n)$ είναι το τυχαίο δείγμα των ευπαθειών και $\beta = (\beta_1, \beta_2, \dots, \beta_k)$ το διάνυσμα των άγνωστων παραμέτρων παλινδρόμησης του μοντέλου ευπάθειας.

Αν τα $Z_i,\ i = 1,2, \dots, n$ είχαν παρατηρηθεί, τότε οι εκτιμήσεις των παραμέτρων β μπορούν εύκολα να προκύψουν αν αντικαταστήσουμε στη σχέση της L_1 τους όρους $Z_i exp(\beta^T X_i)$ από $\exp((\beta^T X_i) + log(Z_i))$ και ακολούθως εφαρμόσουμε τη μέθοδο μέγιστης μερικής 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 θα χρησιμοποιηθούν για την μεγιστοποίηση της συνάρτησης μερικής πιθανοφάνειας, ώστε τελικά να υπολογιστούν οι άγνωστες παράμετροι του μοντέλου.

Expectation Step

Η μη παρατηρούμενη ευπάθεια για κάθε ασθενή $Z_i,\ i = 1, 2, \dots, n$ εκτιμάται από την αναμενόμενη τιμή που δίνεται από τη σχέση

$$E_{(\kappa+1)}(Z_i) = \frac{\frac{1}{{\sigma_{(\kappa)}}^2} + \delta_i}{\frac{1}{{\sigma_{(\kappa)}}^2} + H_{0\kappa}(t_i; \theta)\exp\left({\beta^T}_{(\kappa)} X_i\right)}$$

Να σημειώσουμε εδώ ότι $H_{0\kappa}(\dots)$ είναι ένας μη παραμετρικός εκτιμητής της αθροιστικής βασικής συνάρτησης κινδύνου με βάση τις εκτιμήσεις μέχρι το κ – βήμα.

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:

$$L(\beta|Z) = \prod_{i=1}^{n} \left(\frac{e^{\beta^T X_i + \log(Z_i)}}{\sum_{j \in (t_i)} Z_j e^{\beta^T X_j}} \right)^{\delta_i}$$

Καθώς οι τυχαίες μεταβλητές Z_i και $\log(Z_i)$ μπορούν να αντικατασταθούν από τις αναμενόμενες τιμές τους, ο λογάριθμος της παραπάνω πιθανοφάνειας γράφεται ως:

$$Log\left(L(\beta, \sigma^2)\right) = \sum_{i=1}^{n} \delta_i \left[\beta^T X_i + E_{(\kappa)} \log(Z_i) - log\left(\sum_{j \in (t_i)} E_{(\kappa)}(Z_j) e^{\beta^T X_j} \right) \right]$$

όπου το $E_{(\kappa)}$ δηλώνει την αναμενόμενη τιμή στο $\kappa -$ βήμα του αλγορίθμου. Μεγιστοποιώντας την παραπάνω σχέση, λαμβάνουμε εκτιμήσεις για τις παραμέτρους με διάφορες αριθμητικές επαναληπτικές μεθόδους (Βλέπετε Moon, (1996) και αναφορές εντός). Τέλος, να κάνουμε την παρατήρηση ότι ο αλγόριθμος Ε-Μ μπορεί να χρησιμοποιηθεί και σε συνδυασμό με παραμετρικά μοντέλα ευπάθειας.

ΚΕΦΑΛΑΙΟ 4 – ΤΑ ΔΕΔΟΜΕΝΑ ΚΑΙ Η ΑΝΑΛΥΣΗ ΤΟΥΣ

4.1. Η Αρχική Μορφή των Δεδομένων

Τα δεδομένα που θα αναλυθούν σε αυτό το κεφάλαιο προέρχονται από το διεθνές project προγνωστικών παραγόντων για το λέμφωμα Non-Hodgkins (Shipp et al., 1993). Παρόλο που πολλοί ασθενείς με λέμφωμα Non-Hodgkins γιατρεύονται με συνδυασμό χημειοθεραπειών, οι περισσότεροι δεν θεραπεύονται και τελικά υποκύπτουν στην ασθένειά τους.

Το project αυτό είχε ως σκοπό την ανάπτυξη ενός μοντέλου για την πρόβλεψη της έκβασης της κατάστασης ασθενών που έπασχαν από επιθετικής μορφής λέμφωμα Non-Hodgkins βάση ορισμένων κλινικών χαρακτηριστικών τους πριν τους χορηγηθεί η εκάστοτε θεραπεία.

Οι συμμετέχοντες στην έρευνα ήταν ενήλικες που έπασχαν από επιθετικής μορφής λέμφωμα Non-Hodgkins και που νοσηλεύτηκαν σε κάποιο από 16 ινστιτούτα και ομίλους στις Ηνωμένες Πολιτείες της Αμερικής, την Ευρώπη και τον Καναδά. Επιπρόσθετα, οι ασθενείς αυτοί υπεβλήθησαν σε θεραπευτική αγωγή μεταξύ του 1982 και του 1987 με συνδυασμό χημικοθεραπευτικών προγραμμάτων που περιείχαν σαν δραστική ουσία την doxorubicin και στους οποίους εκτιμήθηκαν ορισμένα κλινικά χαρακτηριστικά που προέβλεπαν την άνευ υποτροπής επιβίωσή τους.

Η συνδυαστική χημειοθεραπεία έχει μετατρέψει το επιθετικής μορφής λέμφωμα Non-Hodgkins από μία συχνά θανατηφόρα ασθένεια σε μία που είναι συχνά ιάσιμη. Παρόλα αυτά, ακόμα και σήμερα, πολλοί ασθενείς καταλήγουν από την ασθένεια αυτή. Αυτό το γεγονός υποδεικνύει την ανάγκη για πιο ακριβείς μεθόδους αναγνώρισης ασθενών με διαφορετικές προγνώσεις στο πέρας του χρόνου.

Ο διαχωρισμός ασθενών σε ομάδες “υψηλού” και “χαμηλού” ρίσκου μπορεί να έχει σημαντικές θεραπευτικές επιπτώσεις. Ασθενείς σε ομάδα υψηλού ρίσκου που δεν αντιμετωπίζονται αποτελεσματικά από τις ήδη υπάρχουσες θεραπείες μπορεί να βοηθηθούν από άλλες πειραματικές προσεγγίσεις ενώ εκείνοι σε ομάδα χαμηλού ρίσκου, μπορεί να παρουσιάσουν τοξικές αντιδράσεις εάν αντιμετωπιστούν με πειραματικές θεραπείες και αντ’ αυτού να βοηθηθούν πιο πολύ με τις ήδη υπάρχουσες θεραπείες. Επίσης η αναγνώριση

διαφόρων ομάδων κινδύνου μπορεί να βοηθήσει στον σχεδιασμό και την κατανόηση διαφόρων θεραπευτικών δοκιμών.

Στην μελέτη αυτή, το στάδιο του καρκίνου υπολογίζεται σύμφωνα με την κατηγοριοποίηση Ann Arbor η οποία αρχικά αναπτύχθηκε για το λέμφωμα Non-Hodgkins. Αυτή η κατηγοριοποίηση δίνει έμφαση στην κατανομή των νοσούντων λεμφαδένων καθώς η νόσος του Hodgkins συνήθως εξαπλώνεται κατά μήκος γειτονικών ομάδων λεμφαδένων (Carbone et al., 1971).

Τα υπόλοιπα χαρακτηριστικά που επιστήμονες προσπάθησαν να συσχετίσουν με την πορεία της ασθένειας είναι:

- **Η ηλικία της διάγνωσης** (*Age*). Η ηλικία κωδικοποιήθηκε ως εξής:
  - Τιμή 0 για ασθενή $<= 60$ ετών
  - Τιμή 1 για ασθενή > 60 ετών

  Αυτή η διχοτόμηση είχε χρησιμοποιηθεί και σε προηγούμενες αναλύσεις για τον λόγο ότι οι ασθενείς παρουσιάζουν διαφορετική δυναμική και απόκριση στις θεραπείες μετά τα 60 τους έτη. Για αυτόν ακριβώς τον λόγο στο Project αυτό, επιχείρησαν οι ερευνητές στο παρελθόν να δημιουργήσουν 2 μοντέλα, ένα για ασθενείς κάτω των 60 ετών και ένα για ασθενείς 60 ετών και άνω.
- **Κατάσταση ικανότητας** (*Performance_Status*). Αντιπροσωπεύει την κατάσταση ικανότητας του ασθενούς και τα συνοδά προβλήματα υγείας. Σύμφωνα με την κλίμακα του Eastern Cooperative Oncology Group Scale Ισοδυναμεί με:
  - 0 εάν ο ασθενής δεν έχει συμπτώματα και είναι Περιπατητικός
  - 1 εάν ο ασθενής έχει συμπτώματα και είναι μερικώς η πλήρως μη περιπατητικός. Αυτό περιλαμβάνει τις περιπτώσεις οπού:
    1. Ο ασθενής έχει συμπτώματα αλλά είναι αυτοεξυπηρετούμενος
    2. Ο ασθενής είναι κατάκοιτος για τουλάχιστον την μισή ημέρα
    3. Ο ασθενής είναι κατάκοιτος για πάνω από την μισή ημέρα
    4. Ο ασθενής είναι χρόνια κατάκοιτος και χρειάζεται βοήθεια για να καλύψει τις καθημερινές του ανάγκες
- **Συγκέντρωση γαλακτικής αφυδρογονάσης ορού** (*LDH_Level*). Πιο συγκεκριμένα, το επίπεδο *LDH* εκφράστηκε σαν ο λόγος της μετρούμενης τιμής του ασθενή προς το άνω φράγμα του ορίου που μετρήθηκε στο εκάστοτε εργαστήριο και ο διαχωρισμός ορίστηκε ως:

- 0 εαν ο λόγος είναι $<\ 1$ φορά του κανονικού
- 1 εάν ο λόγος είναι $>= 1$ φορά του κανονικού

- **Ο αριθμός λεμφικών και έξω-λεμφικών σημείων ασθένειας** (*Nodes*). Ο διαχωρισμός έγινε ως εξής:
  - 0 για $<= 1$ σημείο
  - 1 για $>\ 1$ σημεία
- **Η διάκριση μεταξύ τοπικής η εκτεταμένης νόσου** (*Stage*). Η τοπική νόσος ισοδυναμεί σε στάδιο κατά Ann Arbor I, II και η εκτεταμένη νόσος ισοδυναμεί σε στάδιο κατά Ann Arbor III, IV. Πιο συγκεκριμένα, ο διαχωρισμός ήταν:
  - 0 για στάδιο καρκίνου I, II (In situ καρκίνος)
  - 1 για στάδιο καρκίνου III, IV (Μεταστατικός καρκίνος)

Πιο συγκεκριμένα, τα δεδομένα όλων των προαναφερθέντων παραγόντων, ήταν πλήρη για 1872 ασθενείς από τους οποίους πάρθηκε τυχαίο δείγμα 1385 ασθενών (74%). Τα δεδομένα αποτελούνται από 1385 ανώνυμους ασθενείς που δέχθηκαν ένα πρωτόκολλο συνδυαστικής χημειοθεραπείας που περιείχε Doxorubicin σαν μέρος της φάσης 2 και 3 μίας μελέτης μεταξύ του 1982 και 1987. Το ποσοστό λογοκρισίας ήταν 54.7%.

Η συμπερίληψη μόνο των ασθενών που είχαν ολοκληρώσει την θεραπεία τους το 1987, οδήγησε σε ένα follow-up τουλάχιστον τριών ετών για όλους τους ασθενείς και κατά μέσο όρο 4.5 χρόνια follow-up για τους ασθενείς που επιβίωσαν. Με αυτόν τον τρόπο λοιπόν καθορίστηκαν τα στάδια του καρκίνου όπως επίσης και επανεξετάστηκαν τα παθολογικά τους στοιχεία σύμφωνα με τις οδηγίες των συμμετεχόντων στην μελέτη ιδρυμάτων.

Όλα τα δεδομένα κωδικοποιήθηκαν σαν $0,1$. Η δυαδική αυτή κωδικοποίηση είναι και η βάση του αρχικού μοντέλου που χρησιμοποιήθηκε στο project [Non-Hodgkin's Lymphoma Prognostic Factors Project (1993)]. Ένας κλινικός λόγος της δυαδικής κωδικοποίησης είναι ότι παρέχει μία απλή κατηγοριοποίηση του ρίσκου βασιζόμενη σε ένα πεπερασμένο σετ ομάδων ρίσκου.

Τα δεδομένα όταν κατέβηκαν από την βάση δεδομένων είχαν την μορφή που φαίνεται στην Εικόνα-5 παρακάτω, όπως αυτή πάρθηκε από το περιβάλλον της R, αφού εισήχθησαν τα δεδομένα στο πρόγραμμα:

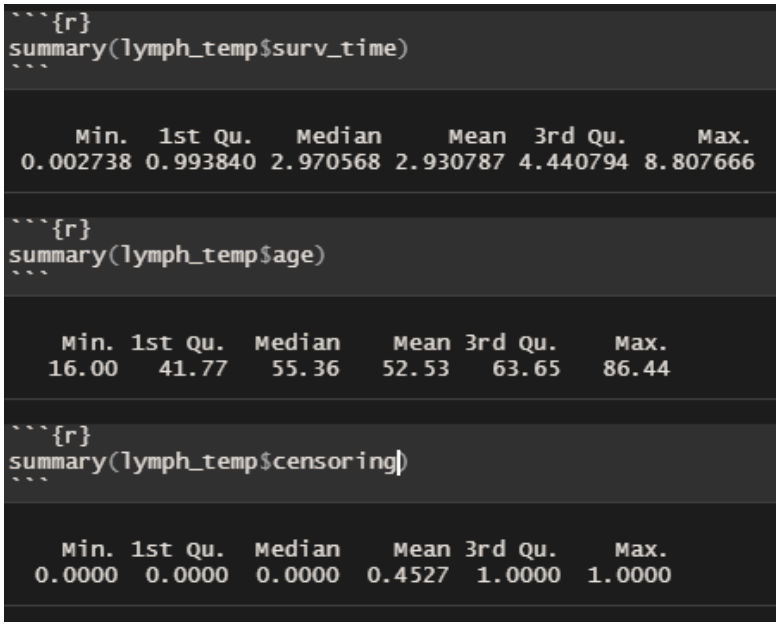
id	surv_time	censoring	age	mobility_status	ldh_level	nodes	tumor_stage
1	1.07323751	1	38.23135	2	2	2	4
2	4.15058179	1	54.85284	1	1	0	2
3	0.99110198	1	58.13826	2	1	2	4
4	4.55304586	0	38.11910	1	1	1	2
5	5.11978097	0	44.01916	2	1	0	2
6	0.77207392	1	27.82752	2	2	1	4
7	5.97125257	0	25.99042	2	2	2	4
8	0.81040383	1	57.57426	2	1	1	4
9	6.00684463	0	31.75359	1	1	1	2
10	1.83983573	1	43.31280	1	1	2	4
11	6.05612594	0	39.06913	1	1	0	3
12	2.82819986	1	48.47912	1	1	1	4
13	3.86858316	1	57.96030	1	1	0	3
14	5.16632444	0	47.05270	2	1	0	3
15	3.54004107	0	55.41958	1	2	0	3
16	4.91170431	0	57.37988	2	1	2	2
17	5.74948665	0	48.46543	2	1	0	2
18	0.78850103	1	59.49624	1	1	1	2
19	0.49007529	1	19.41410	1	1	1	2
20	0.34496920	1	53.31691	2	2	2	4

Εικόνα 5: Τα Δεδομένα στην ακατέργαστη μορφή τους όπως φαίνονται στο περιβάλλον της R. Οι στήλες εκφράζουν τα εξής: id: Είναι ο μοναδικός αριθμός του κάθε ασθενή, surv_time: Είναι ο χρόνος επιβίωσης του εκάστοτε ασθενή (σε χρόνια), censoring: Εκφράζει το εάν η μέτρηση είναι λογοκριμένη η όχι, mobility_status: Εκφράζει την κατάσταση απόδοσης, ldh_level: Εκφράζει την ποσότητα LDH στο αίμα, nodes: Είναι ο αριθμός των αδένων που έχουν νοσήσει, tumor_Stage: Είναι το στάδιο καρκίνου κατά Ann Arbor.

Να σημειωθεί ότι την στήλη “id” την εισήχθη από εμάς, αφότου εισήχθησαν τα δεδομένα στην R, καθώς θα χρειαστεί μετέπειτα στην ανάλυση. Πιο συγκεκριμένα, στην βιβλιοθήκη “parfm” της R, ζητάται μία παράμετρος “Cluster”. Στην περίπτωση που κάποιος θέλει να ορίσει ομάδες ασθενών που θα περιγράφονται από κοινή μεταβλητή ευπάθειας (Shared Frailty Models), το “id” είναι κοινό για τις ομάδες ασθενών αυτών. Στην περίπτωσή μας όμως, ο κάθε ασθενής περιγράφεται από την δική του μοναδική μεταβλητή ευπάθειας. Συνεπώς το “id” είναι ο αριθμός ασθενή, και άρα μοναδικός για κάθε ασθενή.

4.2. Περιγραφική Στατιστική των Δεδομένων

Με την χρήση κατάλληλων εντολών της R, μπορούμε να έχουμε μία πρώτη εικόνα ορισμένων βασικών περιγραφικών δεικτών για τις μεταβλητές surv_time (χρόνος επιβίωσης), age (ηλικία) και Censoring (λογοκρισία), όπως φαίνεται στην Εικόνα-6 παρακάτω:



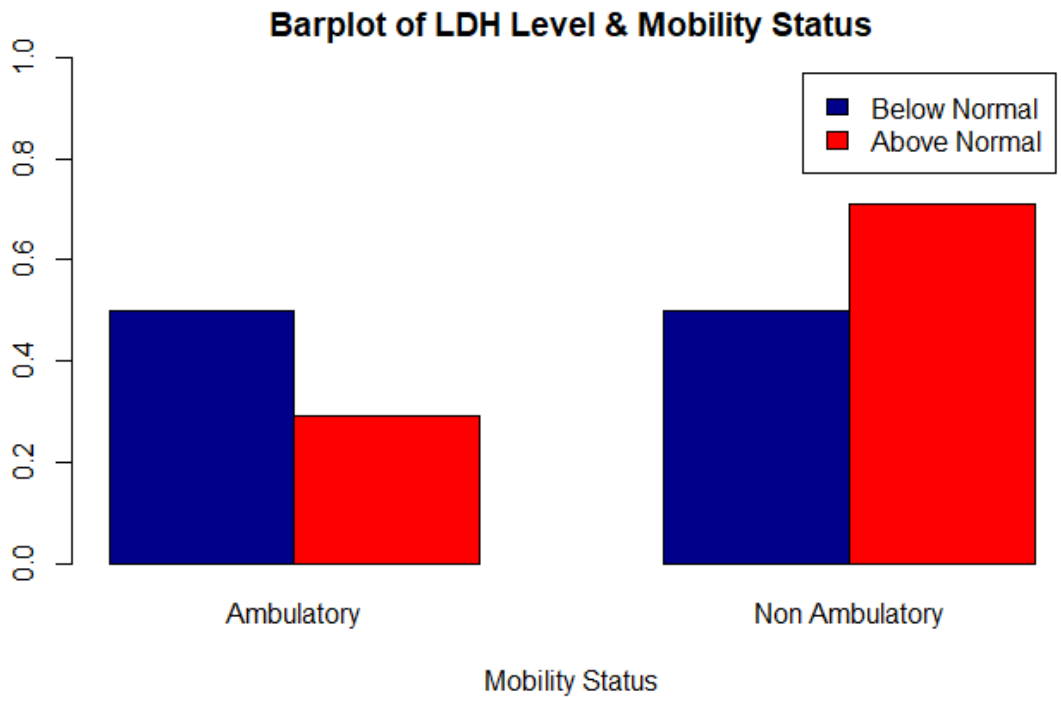
Εικόνα 6: Ορισμένα βασικά περιγραφικά χαρακτηριστικά για τις μεταβλητές surv_time (χρόνος επιβίωσης), age (ηλικία) και Censoring (λογοκρισία).

Παρατηρούμε ότι τα δεδομένα παρουσιάζουν υψηλό ποσοστό λογοκρισίας το οποίο θα βρούμε εύκολα καθώς παρατηρούμε ότι ο μέσος όρος του Censoring είναι $0{,}4527$. Συνεπώς:

$$Cencored\ Data = (1 - mean(Censoring)) = 1 - 0{,}4527 = 0{,}5473 = 54.7\%$$

Ίδιο ποσοστό με αυτό που προαναφέραμε στην παράγραφο 4.1. Επίσης, ο μεγαλύτερος χρόνος επιβίωσης όσων κατέληξαν είναι περίπου 8.8 έτη και η μέση ηλικία των ασθενών περίπου 52 έτη με ελάχιστο τα 16 έτη και μέγιστο τα 86 ετη.

Σε αυτό το σημείο, θα χρησιμοποιήσουμε τις εντολές της R προκειμένου να κατασκευάσουμε Barplots για μία πιο άμεση γραφική αναπαράσταση των κατηγορικών μεταβλητών προκειμένου να αντιληφθούμε καλύτερα το dataset.



Εικόνα 7: Barplot της κατάστασης ικανότητας του ασθενούς (Mobility Status) δοθέντων των επιπέδων της Γαλακτικής αφυδρογονάσης (LDH) στο αίμα του ασθενούς.

Το γράφημα της Εικόνας 7 αντιστοιχεί στα δεδομένα του πίνακα σχετικών συχνοτήτων της κατάστασης ικανότητας του ασθενούς δεσμεύοντας ως προς τα επίπεδα LDH, όπως αυτός φαίνεται στην Εικόνα 8 παρακάτω όπως αυτός πάρθηκε στην R:

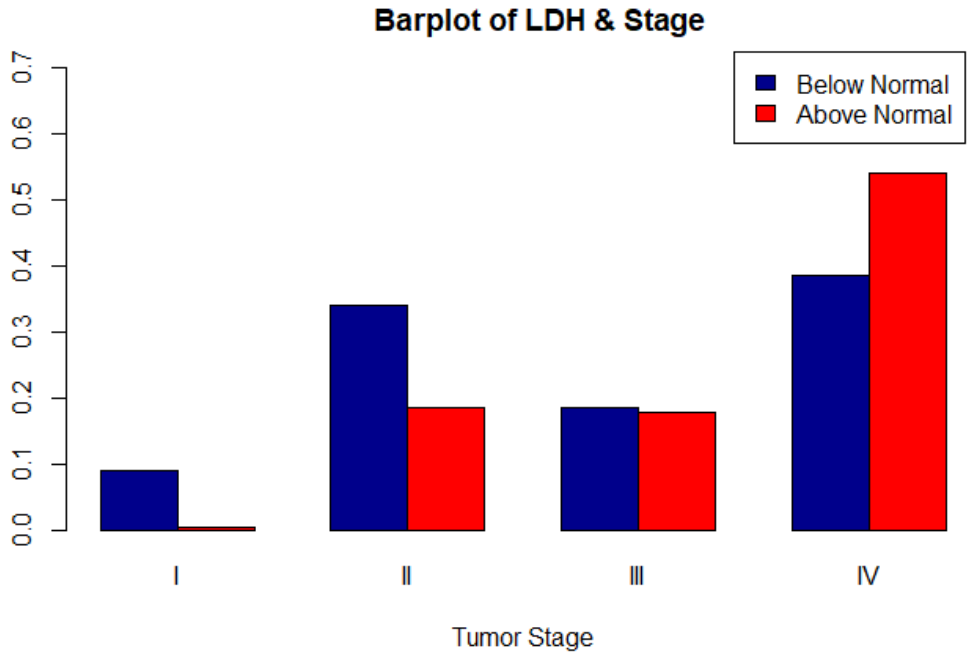
```
```{r}
table1<-table(lymph_temp$ldh_level,lymph_temp$mobility_status)
prop.table(table1,1)
```
```

```
            1         2
ldh_level   mobility_status
   1      0.5004757 0.4995243
   2      0.2904192 0.7095808
```

Εικόνα 8: Πίνακας σχετικών συχνοτήτων της κατάστασης ικανότητας του ασθενούς δεσμεύοντας ως προς τα επίπεδα LDH.

Από την Εικόνα 8, στην οποία περιγράφονται τα δεδομένα βάσει των οποίων δημιουργήθηκε το Barplot της Εικόνας 7, παρατηρούμε ότι από τους ασθενείς στο δείγμα μας που έχουν μικρότερη κυτταρική βλάβη το 50% είναι ικανοί για αυτοεξυπηρέτηση, ενώ το άλλο 50% δεν μπορούν να αυτοεξυπηρετηθούν.

Από την άλλη, από τους ασθενείς που παρουσιάζουν σοβαρότερη βλάβη των κυττάρων, το 29% μπορούν να αυτοεξυπηρετηθούν, ενώ το 70% δεν μπορούν. Σίγουρα συμπεραίνουμε ότι όσο μεγαλύτερα είναι τα επίπεδα *LDH* υπάρχει μεγαλύτερο πρόβλημα στην κατάσταση ικανότητας του ασθενούς.



Εικόνα 9: Barplot του σταδίου του non-Hodgkin λεμφώματος που βρίσκεται ο ασθενής (tumor_stage) δοθέντων των επιπέδων της γαλακτικής αφυδρογονάσης (LDH_level) στο αίμα του ασθενούς.

Το γράφημα της Εικόνας 9 αντιστοιχεί στα δεδομένα του πίνακα σχετικών συχνοτήτων του επιπέδου LDH του ασθενούς δεσμεύοντας ως προς το στάδιο καρκίνου, όπως αυτό φαίνεται στην Εικόνα 10 παρακάτω όπως αυτός πάρθηκε στην R:

```
```{r}
prop.table(table2,1)
```
```

		1	2	3	4
		tumor_stage			
1	ldh_	0.089438630	0.340627973	0.185537583	0.384395814
2	level	0.002994012	0.185628743	0.176646707	0.634730539

Εικόνα 10: Πίνακας σχετικών συχνοτήτων του επιπέδου LDH στο αίμα του ασθενούς δεσμεύοντας ως προς το στάδιο του non-Hodgkin λεμφώματος που βρίσκεται ο ασθενής (tumor_stage).

Στην Εικόνα 10, φαίνεται ο πίνακας σχετικών συχνοτήτων που περιγράφει τα δεδομένα του Barplot της Εικόνας 9. Παρατηρούμε ότι από τους ασθενείς στο δείγμα μας που έχουν μικρότερη κυτταρική βλάβη:

- Το 8% βρίσκονται στο Στάδιο-I
- Το 34% βρίσκονται στο Στάδιο-II
- Το 18% βρίσκονται στο Στάδιο-III
- Το 38% βρίσκονται στο Στάδιο-IV

Από την άλλη, από τους ασθενείς που παρουσιάζουν σοβαρότερη βλάβη των κυττάρων,

- Το 0.2% βρίσκονται στο Στάδιο-I
- το 18% βρίσκονται στο Στάδιο-II
- το 17% βρίσκονται στο Στάδιο-III
- το 63% βρίσκονται στο στάδιο-IV

Είναι αναμενόμενο ότι όσο μεγαλύτερα είναι τα επίπεδα *LDH* ο ασθενής τείνει να βρίσκεται σε πιο προχωρημένο στάδιο της νόσου. 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 (38%). Σίγουρα το ποσοστό αυτών που βρίσκονται στο αρχικό στάδιο της νόσου είναι πολύ μικρό για αυτούς που έχουν μικρότερη κυτταρική βλάβη (0.2%).

4.3. Η Επεξεργασμένη Μορφή των Δεδομένων

Αφού ελέγχθηκαν τα δεδομένα για την ορθότητα και πληρότητά τους, προχωρήσαμε στην απαραίτητη κωδικοποίηση για την καλύτερη ερμηνεία των αποτελεσμάτων. Πιο συγκεκριμένα, η κωδικοποίηση θα είναι αυτή που περιγράφεται στο κεφάλαιο 4.1.1. μετά την κωδικοποίηση, τα δεδομένα ήρθαν στην τελική τους μορφή όπως εκείνη φαίνεται στην Εικόνα 11 παρακάτω, όπως φαίνεται στο περιβάλλον της R:

surv_time	censoring	age_temp	tumor_stage_temp	nodes_temp	mobility_status_temp	ldh_level_temp
1.07323751	1	0	1	1	1	1
4.15058179	1	0	0	0	0	0
0.99110198	1	0	1	1	1	0
4.55304586	0	0	0	1	0	0
5.11978097	0	0	0	0	1	0
0.77207392	1	0	1	1	1	1
5.97125257	0	0	1	1	1	1
0.81040383	1	0	1	1	1	0
6.00684463	0	0	0	1	0	0
1.83983573	1	0	1	1	0	0
6.05612594	0	0	1	0	0	0
2.82819986	1	0	1	1	0	0
3.86858316	1	0	1	0	0	0
5.16632444	0	0	1	0	1	0
3.54004107	0	0	1	0	0	1
4.91170431	0	0	0	1	1	0
5.74948665	0	0	0	0	1	0

Εικόνα 11: Τα Δεδομένα στην καθαρή μορφή τους όπως φαίνονται στο περιβάλλον της R. Οι στήλες, στις οποίες έγιναν οι αλλαγές που εκφράζονται στο υποκεφάλαιο 3.5.1., εκφράζουν τα εξής: Surv_time: Είναι ο χρόνος επιβίωσης του εκάστοτε ασθενή, Censoring: Είναι το εάν η μέτρηση είναι λογοκριμένη η όχι, mobility_status_temp: Εκφράζει την κατάσταση απόδοσης, ldh_level_temp: Εκφράζει την ποσότητα LDH στο αίμα, nodes_temp: Είναι ο αριθμός των αδένων που έχουν νοσήσει, tumor_Stage_temp: Είναι το στάδιο καρκίνου κατά Ann Arbor

4.4. Η Ανάλυση των Δεδομένων – Αποτελέσματα Αλγορίθμων

4.4.1. Προσαρμογή Παραμετρικών Μοντέλων Ευπάθειας - Περίπτωση Μοντέλου του Cox.

Στη συνέχεια, τα δεδομένα αναλύθηκαν με την βοήθεια του αλγορίθμου που αναπτύχθηκε στην παράγραφο 3.5. Τα αποτελέσματα ήταν τα εξής:

Σύμφωνα με το **Βήμα-1** όπως αναφέρεται στην παράγραφο 3.5., η εφαρμογή του παραμετρικού μοντέλου του Cox, είχε τα ακόλουθα αποτελέσματα στο περιβάλλον της R όπως φαίνεται στην Εικόνα 12 παρακάτω:

```
```{r}
res.cox.exp <- survreg(Surv(surv_time, censoring) ~ mobility_status_temp + ldh_level_temp +
                  nodes_temp + tumor_stage_temp + age_temp,
                data =  lymph_temp, dist='exponential')
summary(res.cox.exp)
res.cox.exp$loglik

#Initial Beta Value
beta0<- -as.vector(res.cox.exp$coefficients[2:(d+1)])
beta0

#Initial Lambda Value
lambda0<- exp(-as.vector(res.cox.exp$coefficients)[1])
lambda0
```
```

```
Call:
survreg(formula = Surv(surv_time, censoring) ~ mobility_status_temp +
    ldh_level_temp + nodes_temp + tumor_stage_temp + age_temp,
    data = lymph_temp, dist = "exponential")
                    Value Std. Error     z       p
(Intercept)          3.2451     0.1130 28.72 < 2e-16
mobility_status_temp -0.6748     0.0881 -7.66 1.8e-14
ldh_level_temp       -0.6300     0.0872 -7.23 5.0e-13
nodes_temp           -0.3203     0.0925 -3.46 0.00054
tumor_stage_temp     -0.4968     0.0973 -5.10 3.3e-07
age_temp             -0.7523     0.0802 -9.38 < 2e-16

Scale fixed at 1

Exponential distribution
Loglik(model)= -1641.2   Loglik(intercept only)= -1798.1
        Chisq= 313.88 on 5 degrees of freedom, p= 1e-65
Number of Newton-Raphson Iterations: 5
n= 1385

[1] -1798.098 -1641.156
[1] 0.6748246 0.6299770 0.3202641 0.4967867 0.7523185
[1] 0.03896591
```

Εικόνα 12: Τα αποτελέσματα της εφαρμογής του παραμετρικού μοντέλου του Cox στα δεδομένα όπως εκείνα φαίνονται στο περιβάλλον της R.

Βλέπουμε λοιπόν από την Εικόνα 11, το διάνυσμα των παραμέτρων β είναι το:

$$\beta = (\beta_{mobility\_Status\_temp}, \beta_{ldh\_level\_temp}, \beta_{nodes\_temp}, \beta_{tumor\_stage\_temp}, \beta_{Age\_temp})$$

Παρατηρούμε ότι όλες οι μεταβλητές είναι ισχυρά στατιστικά σημαντικές και παραμένουν στο μοντέλο. Το διάνυσμα των εκτιμήσεων των παραμέτρων με βάση το παραμετρικό μοντέλο του Cox είναι:

$$\beta'_0 = (0.6748, 0.63000, 0.3203, 0.4968, 0.7523)$$

με τυπικό σφάλμα που κυμαίνεται απο 0.08 έως 0.0973.

Ενώ η τιμή της log-Likelihood που επιτυγχάνεται μέσω του παραμετρικού μοντέλου του Cox είναι: $-LogLik = -1641.156$. Στο παραμετρικό μοντέλο του Cox έχουμε υποθέσει εκθετική κατανομή για τους χρόνους επιβίωσης δηλαδή η συνάρτηση κινδύνου που έχουμε υποθέσει είναι σταθερή και ίση με λ. Η παράμετρος λ θεωρείται άγνωστη και εκτιμάται με βάση το μοντέλο του Cox ίση με $\lambda = 0.038$. Οι τιμές των β και λ που παίρνουμε στο Βήμα-1 μας δίνουν μία πρώτη ένδειξη για τις πραγματικές τιμές των παραμέτρων και ειδικά το διάνυσμα β_0 θα χρησιμοποιηθεί σαν αρχική τιμή στην μεγιστοποίηση της συνάρτησης πιθανοφάνειας ως προς β στη συνέχεια.

4.4.2. Προσαρμογή Παραμετρικών Μοντέλων Ευπάθειας - Περίπτωση Gamma Μοντέλου Ευπάθειας

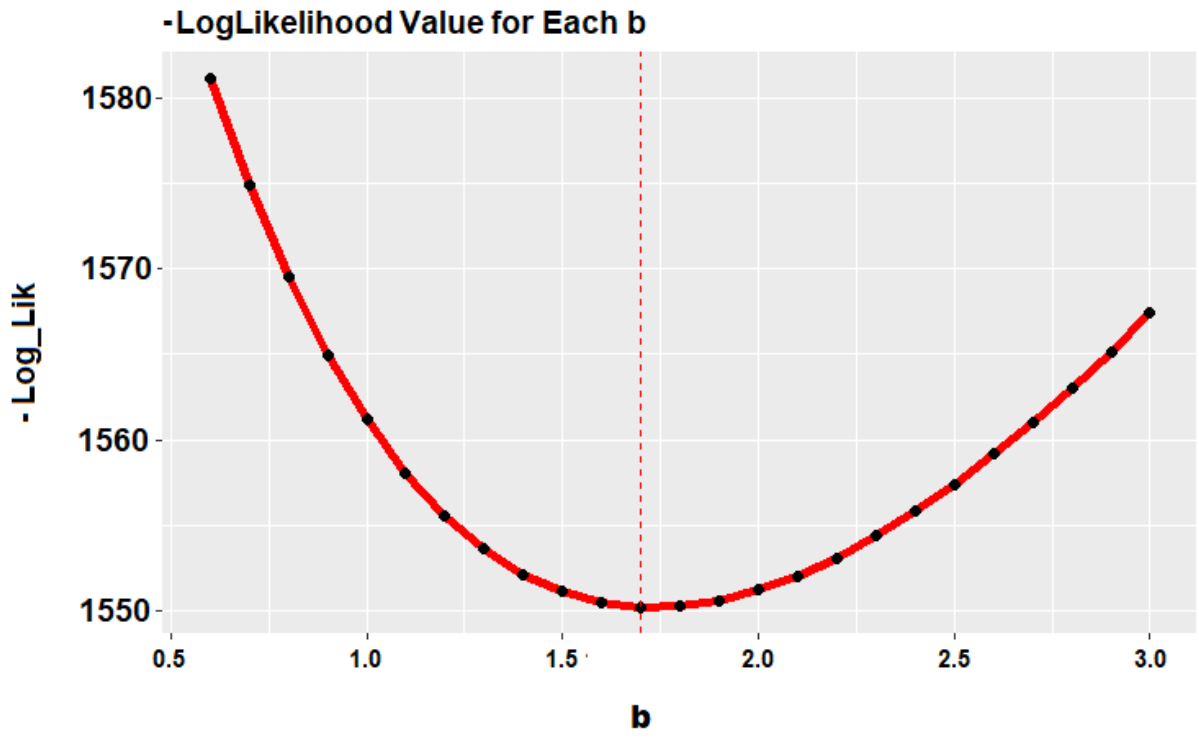
Έχοντας λοιπόν σαν Input το β_0, και προχωρώντας στα βήματα $2-6$ όπως φαίνονται στην παράγραφο $3.5.$, τα αποτελέσματα για το Gamma μοντέλο ευπάθειας με συνάρτηση κινδύνου που δίνεται από την εκθετική κατανομή για 25 τιμές της παραμέτρου b στο grid που επελέγη, φαίνονται στον Πίνακα-1 παρακάτω:

Number of grid value	b	λ	*β*					-Log-Likelihood
			Mobility_ Status	Ldh_ Level	Nodes	Tumor_ Stage	Age	
1	0.6	0.037	0.828	0.925	0.398	0.603	0.896	1581.167
2	0.7	0.037	0.844	0.958	0.406	0.614	0.910	1574.896
3	0.8	0.037	0.857	0.987	0.414	0.624	0.922	1569.530
4	0.9	0.038	0.870	1.011	0.421	0.632	0.932	1564.989
5	1.0	0.038	0.880	1.033	0.428	0.640	0.941	1561.194
6	1.1	0.038	0.890	1.052	0.433	0.647	0.950	1558.072
7	1.2	0.039	0.898	1.070	0.438	0.653	0.956	1555.558
8	1.3	0.039	0.906	1.085	0.443	0.657	0.963	1553.590
9	1.4	0.040	0.912	1.099	0.446	0.662	0.968	1552.113
10	1.5	0.040	0.917	1.111	0.449	0.665	0.973	1551.079
11	1.6	0.040	0.926	1.122	0.457	0.672	0.980	1550.443
12	1.7	0.041	0.931	1.133	0.460	0.675	0.985	1550.167
13	1.8	0.041	0.934	1.142	0.462	0.676	0.988	1550.215
14	1.9	0.042	0.938	1.151	0.464	0.678	0.991	1550.557
15	2.0	0.042	0.942	1.158	0.466	0.680	0.994	1551.164
16	2.1	0.043	0.945	1.166	0.469	0.682	0.997	1552.011
17	2.2	0.043	0.948	1.173	0.471	0.684	0.999	1553.077
18	2.3	0.044	0.951	1.179	0.472	0.685	1.002	1554.341
19	2.4	0.044	0.953	1.185	0.474	0.686	1.004	1555.784
20	2.5	0.045	0.956	1.190	0.476	0.688	1.007	1557.391
21	2.6	0.045	0.958	1.196	0.476	0.687	1.008	1559.146
22	2.7	0.046	0.959	1.201	0.478	0.687	1.009	1561.037

23	2.8	0.046	0.962	1.205	0.479	0.688	1.011	1563.051
24	2.9	0.047	0.963	1.210	0.481	0.689	1.013	1565.178
25	3.0	0.047	0.965	1.214	0.482	0.689	1.014	1567.407

Πίνακας 1: Το αποτέλεσμα του αλγορίθμου για 25 τιμές του Grid του b, όπως αυτό ορίζεται στο Βήμα-2 του κεφαλαίου 3.5. Παρατηρούμε ότι στη 12η τιμή του b στο grid παρατηρήθηκε ελάχιστο της -Log-Likelihood με τιμή 1550.167.

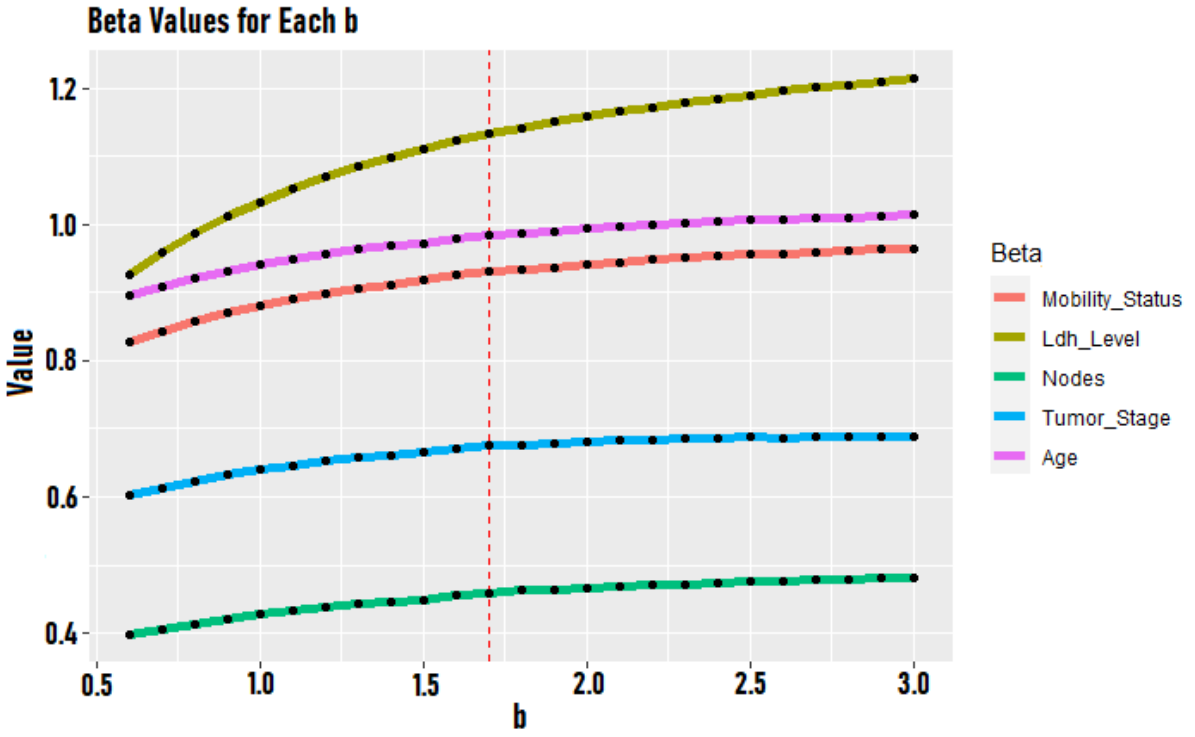
Πιο συγκεκριμένα, οι υπολογισμένες -Log-Likelihood φαίνονται στην Εικόνα 13 παρακάτω σαν συνάρτηση του b:



Εικόνα 13: Οι υπολογιζόμενες τιμές της -Log-Likelihood όπως αυτές υπολογίστηκαν από τον αλγόριθμο. Παρατηρείται και μία Κόκκινη κάθετη διακεκομμένη γραμμή για την τιμή του b στην οποία παρουσιάστηκε ελάχιστο στην τιμή της -Log-Likelihood.

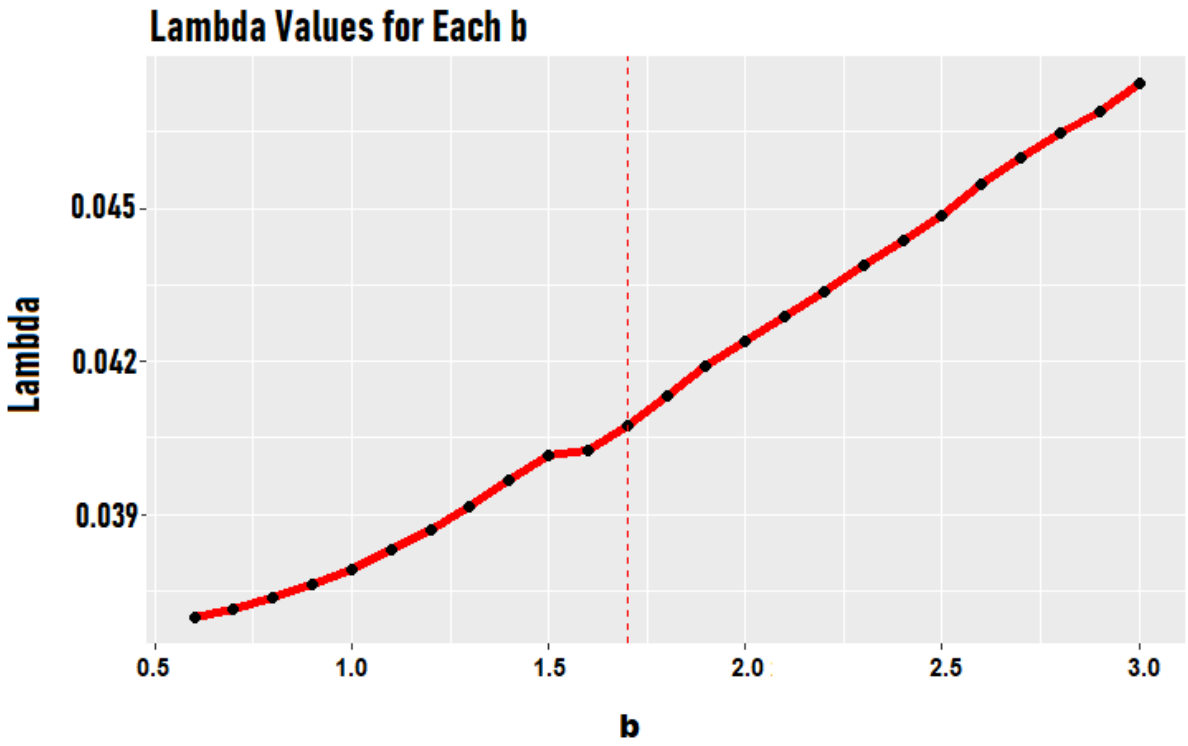
Να σημειωθεί σε αυτό το σημείο ότι η R ελαχιστοποιεί συναρτήσεις. Αρα έχουμε ζητήσει να ελαχιστοποιήσει την -Log-Likelihood και για αυτό και οι -Log-Likelihood που βλέπουμε στην Εικόνα 13 είναι θετικές και παρατηρούμε ελάχιστο αντί για μέγιστο. Επίσης να θυμίσουμε εδώ ότι η παράμετρος b η οποία είναι η διασπορά της τυχαίας μεταβλητής "ευπάθεια" είναι επίσης άγνωστη και πρέπει να εκτιμηθεί. Στο grid τιμών λοιπόν ξεκινάμε με μικρές τιμές για τη διασπορά (με μικρότερη το 0.6) και σταδιακά μεγαλώνουμε τη διασπορά (με μέγιστη τιμή το 3).

Μέσα σε αυτό το διάστημα διαπιστώθηκε ότι για $b = 1.7$ έχουμε μέγιστο για τη συνάρτηση πιθανοφάνειας το οποίο αντιστοιχεί σε "μέτρια" διασπορά και είναι παράλληλα και η εκτίμησή το "b" μέσω της μεθόδου μεγίστης πιθανοφάνειας. Στην συνέχεια, οι υπολογιζόμενες τιμές των συμμεταβλητών (διάνυσμα $\hat{\beta}$), Φαίνονται στην Εικόνα 14, παρακάτω για κάθε τιμή "b" του Grid:



Εικόνα 14: Οι υπολογιζόμενες εκτιμήτριες της παραμέτρου β όπως αυτές υπολογίστηκαν από τον αλγόριθμο για διάφορες τιμές της παραμέτρου b. Παρατηρείται και μία Κόκκινη κάθετη διακεκομμένη γραμμή στην τιμή του b για την οποία παρουσιάστηκε ελάχιστο στην τιμή της -Log-Likelihood.

Με τον ίδιο τρόπο, οι υπολογιζόμενες εκτιμήσεις της μεταβλητής λ, φαίνονται στην Εικόνα 15, παρακάτω:



Εικόνα 15: Οι υπολογιζόμενες τιμές της μεταβλητής λ όπως αυτές υπολογίστηκαν από τον αλγόριθμο για διάφορες τιμές της παραμέτρου b. Παρατηρείται και μία Κόκκινη κάθετη διακεκομμένη γραμμή στην τιμή του b για την οποία παρουσιάστηκε ελάχιστο στην τιμή της -Log-Likelihood.

Συμπερασματικά, βρίσκουμε ότι οι εκτιμήσεις με βάση το Γάμμα μοντέλο ευπάθειας και εκθετική κατανομή συνάρτησης διακινδύνευσης, όπως αυτές προκύπτουν από τον αλγόριθμό μας, είναι:

$$\hat{b} = 1.7, \hat{\lambda} = 0.04 \; και \; \hat{\beta} = (0.93, 1.13, 0.46, 0.67, 0.98)'$$

Με βάση τις εκτιμήσεις των παραμέτρων β οι οποίες είναι όλες θετικές, παρατηρούμε ότι ο κίνδυνος για όλες τις συμμεταβλητές αυξάνει από την ομάδα με τιμή 0 στην ομάδα 1. Πιο συγκεκριμένα:

- Για την μεταβλητή Mobility Status ο κίνδυνος αυξάνει κατά $e^{0.93} = 2.53$ φορές, με όλες τις άλλες συμμεταβλητές fixed, για αυτούς τους ασθενείς που έχουν κινητικά προβλήματα σε σχέση με τους ασθενείς που είναι περιπατητικοί.
- Για την μεταβλητή LDH Level ο κίνδυνος αυξάνει κατά $e^{1.13} = 3.10$ φορές, με όλες τις άλλες συμμεταβλητές fixed, για αυτούς τους ασθενείς που έχουν επίπεδο LDH στο αίμα τους πάνω μία φορά του κανονικού σε σχέση με τους ασθενείς που έχουν επίπεδο LDH στο αίμα τους κάτω από μία φορά του κανονικού

- Για την μεταβλητή Nodes ο κίνδυνος αυξάνει κατά $e^{0.46} = 1.58$ φορές, με όλες τις άλλες συμμεταβλητές fixed, για αυτούς τους ασθενείς που έχουν πάνω από μία εστία ασθενείας σε σχέση με τους ασθενείς που έχουν κάτω από μία εστία ασθενείας.
- Για την μεταβλητή Tumor_Stage ο κίνδυνος αυξάνει κατά $e^{0.67} = 1.95$ φορές, με όλες τις άλλες συμμεταβλητές fixed, για αυτούς τους ασθενείς που βρίσκονται στο στάδιο $III - IV$ σε σχέση με τους ασθενείς που βρίσκονται στο στάδιο $I - II$.
- Για την μεταβλητή Age ο κίνδυνος αυξάνει κατά $e^{0.98} = 2.66$ φορές, με όλες τις άλλες συμμεταβλητές fixed, για αυτούς τους ασθενείς που είναι άνω των 60 σε σχέση με τους ασθενείς που είναι κάτω των 60.

4.4.3. Προσαρμογή Παραμετρικών Μοντέλων Ευπάθειας - Περίπτωση Inverse Gaussian Μοντέλου Ευπάθειας

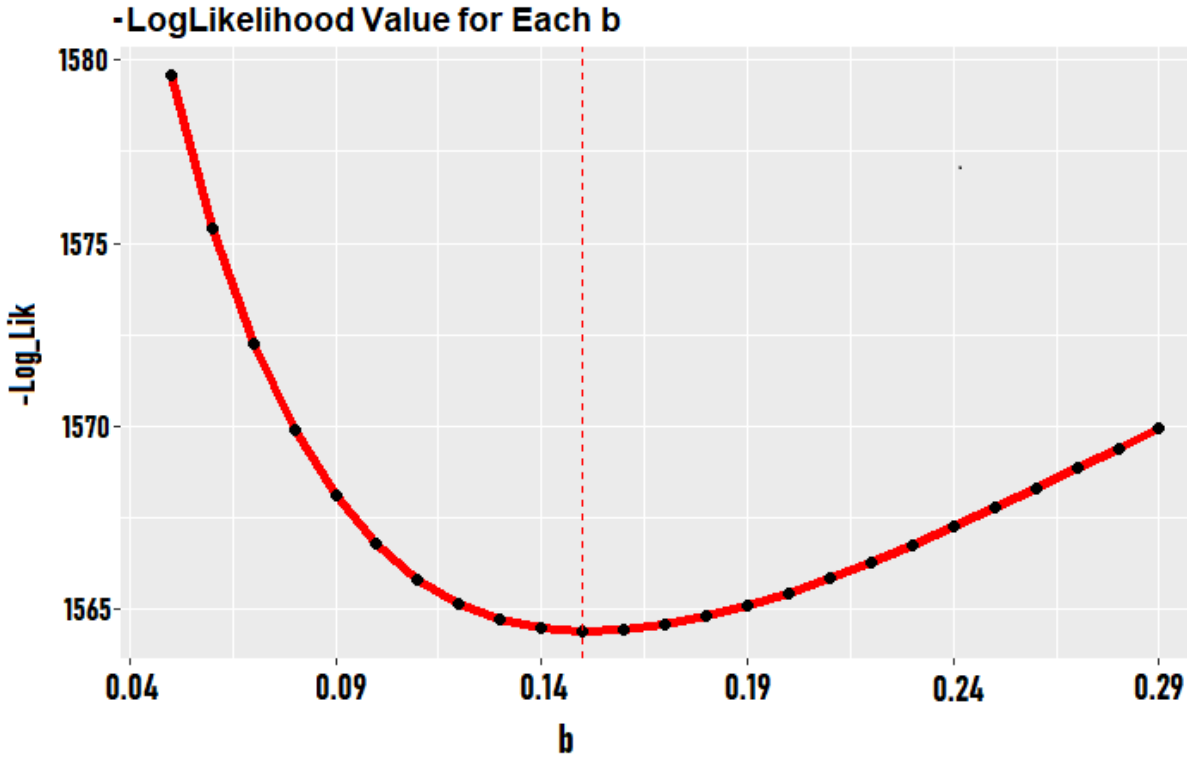
Έχοντας και πάλι σαν Input το β_0, και προχωρώντας στα βήματα $2-6$ όπως φαίνονται στην παράγραφο 3.5., τα αποτελέσματα για το Inverse Gaussian μοντέλο ευπάθειας με συνάρτηση κινδύνου που δίνεται από την εκθετική κατανομή για 25 Iteration Runs Φαίνονται στον Πίνακα-2 παρακάτω:

Number of grid value	b	λ	β					-Log-Likelihood
			Mobility_ Status	Ldh_ Level	Nodes	Tumor_ Stage	Age	
1	0.05	0.069	1.020	1.039	0.455	0.740	1.094	1579.580
2	0.06	0.064	1.001	1.030	0.452	0.723	1.076	1575.404
3	0.07	0.060	0.985	1.022	0.449	0.710	1.061	1572.257
4	0.08	0.057	0.970	1.013	0.444	0.698	1.046	1569.884
5	0.09	0.055	0.958	1.005	0.443	0.689	1.035	1568.105
6	0.10	0.053	0.947	0.997	0.439	0.681	1.023	1566.788
7	0.11	0.052	0.938	0.991	0.438	0.675	1.013	1565.835
8	0.12	0.050	0.929	0.984	0.436	0.669	1.005	1565.172
9	0.13	0.049	0.922	0.978	0.435	0.665	0.997	1564.742
10	0.14	0.048	0.916	0.972	0.433	0.660	0.990	1564.501
11	0.15	0.048	0.909	0.967	0.431	0.655	0.983	1564.413
12	0.16	0.047	0.904	0.961	0.429	0.652	0.978	1564.450
13	0.17	0.046	0.899	0.957	0.428	0.648	0.972	1564.589
14	0.18	0.046	0.894	0.952	0.426	0.645	0.967	1564.814
15	0.19	0.045	0.889	0.948	0.425	0.642	0.962	1565.108
16	0.20	0.045	0.885	0.943	0.423	0.639	0.957	1565.459
17	0.21	0.045	0.880	0.939	0.421	0.636	0.952	1565.857
18	0.22	0.044	0.876	0.935	0.419	0.633	0.948	1566.294
19	0.23	0.044	0.872	0.932	0.418	0.630	0.944	1566.763
20	0.24	0.044	0.869	0.928	0.416	0.628	0.940	1567.258
21	0.25	0.043	0.866	0.925	0.415	0.626	0.937	1567.774
22	0.26	0.043	0.863	0.921	0.414	0.624	0.934	1568.307

23	0.27	0.043	0.860	0.918	0.413	0.622	0.930	1568.853
24	0.28	0.043	0.858	0.915	0.412	0.620	0.928	1569.409
25	0.29	0.042	0.855	0.912	0.411	0.619	0.925	1569.972

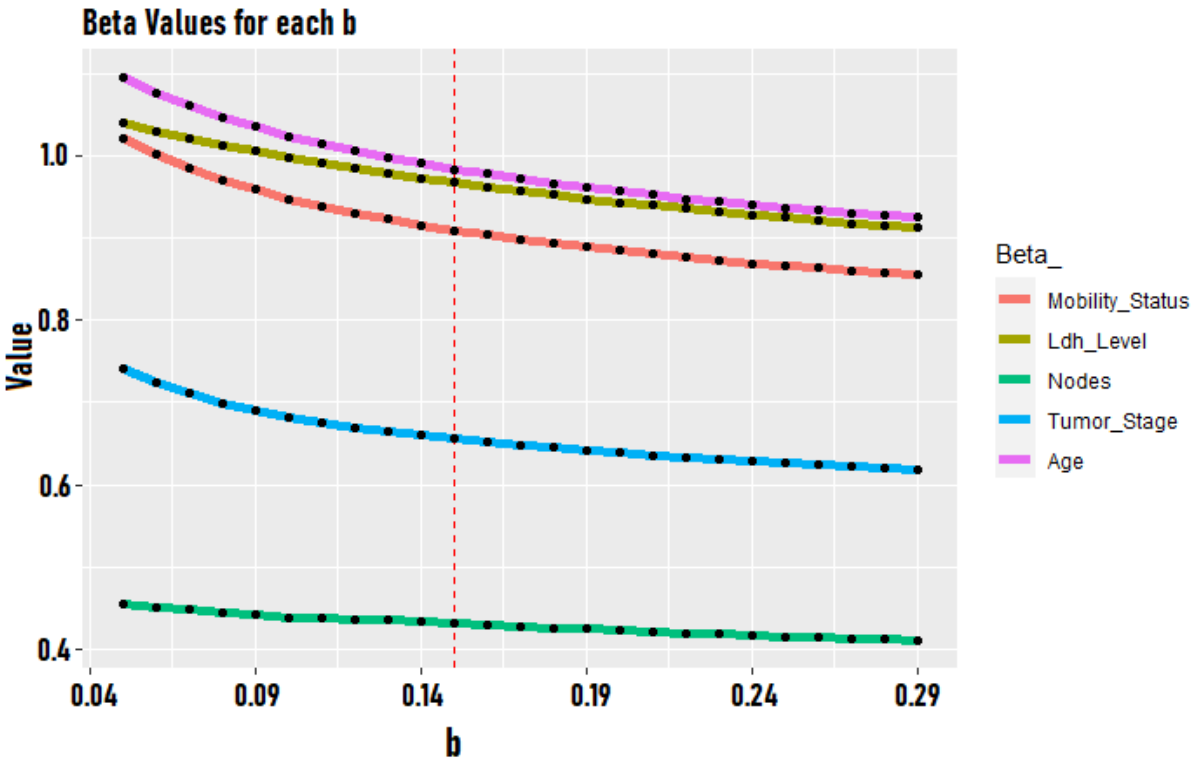
Πίνακας 2: Το αποτέλεσμα του αλγορίθμου για 25 τιμές του Grid (Iteration (k)) του b, όπως αυτό ορίζεται στο Βήμα-2 του κεφαλαίου 3.5. Παρατηρούμε ότι στο 11ο iteration run παρατηρήθηκε ελάχιστο της -Log-Likelihood με τιμή 1564.413.

Πιο συγκεκριμένα, οι υπολογισμένες -Log-Likelihood φαίνονται στην Εικόνα 16 παρακάτω:



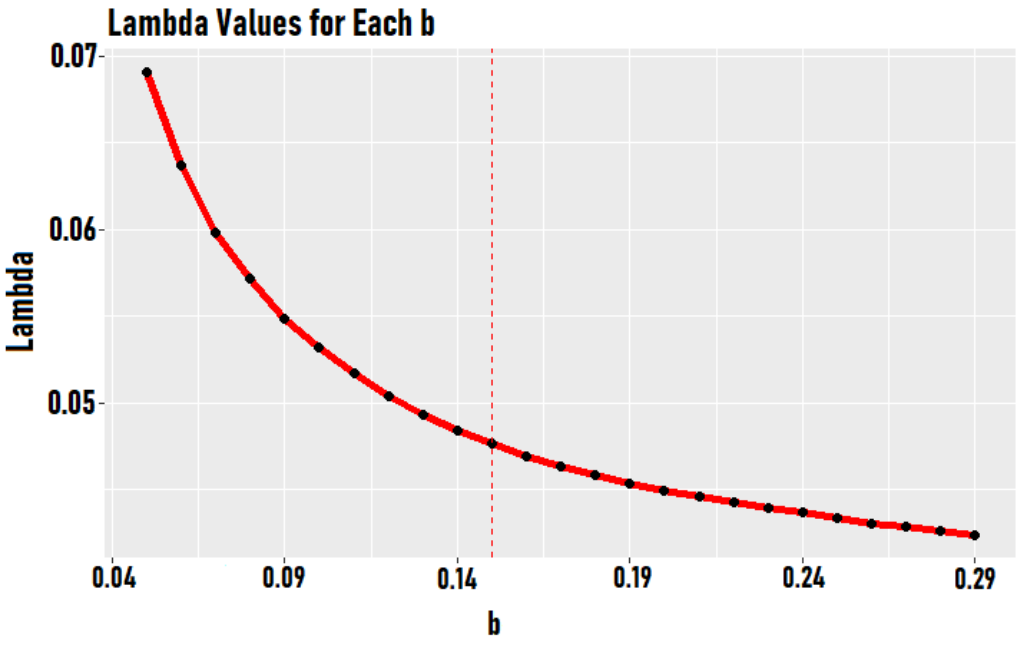
Εικόνα 16: Οι υπολογιζόμενες τιμές της -Log-Likelihood όπως αυτές υπολογίστηκαν από τον αλγόριθμο. Παρατηρείται και μία Κόκκινη κάθετη διακεκομμένη γραμμή για το Iteration στο οποίο παρουσιάστηκε ελάχιστο στην τιμή της -Log-Likelihood.

Να σημειωθεί σε αυτό το σημείο ότι η R μεγιστοποιεί συναρτήσεις. Για αυτό και οι Log-Likelihood που υπολογίζει είναι αρνητικές. Έτσι, εμείς αλλάζοντας το πρόσημο και θεωρώντας σαν optimum το μέγιστο, επιτυγχάνουμε το ίδιο αποτέλεσμα. Στην συνέχεια, οι υπολογιζόμενες τιμές των συμμεταβλητών (διάνυσμα beta), Φαίνονται στην Εικόνα 17, παρακάτω:



Εικόνα 17: Οι υπολογιζόμενες τιμές των συμμεταβλητών όπως αυτές υπολογίστηκαν από τον αλγόριθμο. Παρατηρείται και μία Κόκκινη κάθετη διακεκομμένη γραμμή στις τιμές για τις οποίες παρουσιάστηκε ελάχιστο στην τιμή της- Log-Likelihood.

Με τον ίδιο τρόπο, οι υπολογιζόμενες τιμές της μεταβλητής λ, φαίνονται στην Εικόνα 18, παρακάτω:



Εικόνα 18: Οι υπολογιζόμενες τιμές της μεταβλητής λ όπως αυτές υπολογίστηκαν από τον αλγόριθμο. Παρατηρείται και μία Κόκκινη κάθετη διακεκομμένη γραμμή στις τιμές για τις οποίες παρουσιάστηκε ελάχιστο στην τιμή της -Log-Likelihood.

Συμπερασματικά, βρίσκουμε ότι οι εκτιμήσεις με βάση το Inverse Gaussian μοντέλο ευπάθειας και εκθετική κατανομή συνάρτησης διακινδύνευσης, όπως αυτές προκύπτουν από τον αλγόριθμό μας, είναι:

$$\hat{b} = 0.15, \hat{\lambda} = 0.048 \ \kappa\alpha\iota \ \hat{\beta} = (0.909, 0.967, 0.431, 0.655, 0.983)'$$

Με βάση τις εκτιμήσεις των παραμέτρων β οι οποίες είναι όλες θετικές, παρατηρούμε ότι ο κίνδυνος για όλες τις συμμεταβλητές αυξάνει από την ομάδα με τιμή 0 στην ομάδα 1. Πιο συγκεκριμένα:

- Για την μεταβλητή Mobility Status ο κίνδυνος αυξάνει κατά $e^{0.91} = 2.48$ φορές, με όλες τις άλλες συμμεταβλητές fixed, για αυτούς τους ασθενείς που έχουν κινητικά προβλήματα σε σχέση με τους ασθενείς που είναι περιπατητικοί.
- Για την μεταβλητή LDH Level ο κίνδυνος αυξάνει κατά $e^{0.97} = 2.64$ φορές, με όλες τις άλλες συμμεταβλητές fixed, για αυτούς τους ασθενείς που επίπεδο LDH στο αίμα τους πάνω μία φορά του κανονικού σε σχέση με τους ασθενείς που έχουν επίπεδο LDH στο αίμα τους κάτω από μία φορά του κανονικού
- Για την μεταβλητή Nodes ο κίνδυνος αυξάνει κατά $e^{0.43} = 1.54$ φορές, με όλες τις άλλες συμμεταβλητές fixed, για αυτούς τους ασθενείς που έχουν πάνω από μία εστία ασθενείας σε σχέση με τους ασθενείς που έχουν κάτω από μία εστία ασθενείας.
- Για την μεταβλητή Tumor_Stage ο κίνδυνος αυξάνει κατά $e^{0.66} = 1.93$ φορές, με όλες τις άλλες συμμεταβλητές fixed, για αυτούς τους ασθενείς που βρίσκονται στο στάδιο $III - IV$ σε σχέση με τους ασθενείς που βρίσκονται στο στάδιο $I - II$.
- Για την μεταβλητή Age ο κίνδυνος αυξάνει κατά $e^{0.98} = 2.66$ φορές, με όλες τις άλλες συμμεταβλητές fixed, για αυτούς τους ασθενείς που είναι άνω των 60 σε σχέση με τους ασθενείς που είναι κάτω των 60.

4.5. Ανάλυση Δεδομένων – Αποτελέσματα Βιβλιοθήκης Parfm

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

Θα χρησιμοποιήσουμε σαν συνάρτηση κίνδυνου την εκθετική και για την μεταβλητή ευπάθειας, τις δύο κατανομές από τις πέντε, δηλαδή την Γάμμα και την Inverse Gaussian προκειμένου να συγκρίνουμε τα αποτελέσματά μας.

4.5.1. Προσαρμογή Παραμετρικών Μοντέλων Ευπάθειας - Περίπτωση Gamma Μοντέλου

Στην περίπτωση προσαρμογής του Gamma παραμετρικού Μοντέλου ευπάθειας με εκθετική συνάρτηση κινδύνου, τα αποτελέσματα, όπως αυτά παρουσιάζονται στο περιβάλλον της R, φαίνονται στην Εικόνα 19 παρακάτω:

```
```{r}
mod_gamma <- parfm(Surv(surv_time, censoring) ~ mobility_status_temp + ldh_level_temp +
                   nodes_temp + tumor_stage_temp + age_temp , cluster="id", data=lymph_temp, dist="exponential",
frailty="gamma")

summary(mod_gamma)

mod_gamma
```

    ESTIMATE   SE   p-val
 Min.   :0.04102   Min.   :0.006279   Min.   :0.0000000
 1st Qu.:0.56738   1st Qu.:0.126388   1st Qu.:0.0000000
 Median :0.93126   Median :0.135982   Median :0.0000000
 Mean   :0.85175   Mean   :0.122093   Mean   :0.0001441
 3rd Qu.:1.06058   3rd Qu.:0.140488   3rd Qu.:0.0000012
 Max.   :1.73401   Max.   :0.178635   Max.   :0.0007195
                                      NA's   :2

Frailty distribution: gamma
Baseline hazard distribution: Exponential
Loglikelihood: -1550.148

                    ESTIMATE SE   p-val
theta   1.734   0.179
lambda   0.041   0.006
mobility_status_temp 0.931   0.129 <.001 ***
ldh_level_temp   1.136   0.142 <.001 ***
nodes_temp   0.460   0.136 0.001 ***
tumor_stage_temp   0.675   0.139 <.001 ***
age_temp   0.985   0.123 <.001 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Εικόνα 19: Προσαρμογή Gamma παραμετρικού μοντέλου ευπάθειας με εκθετική συνάρτηση κινδύνου στο dataset, χρησιμοποιώντας την βιβλιοθήκη “parfm” της R.
```

Η εκτίμηση της παραμέτρου *"theta"*, η οποία ουσιαστικά εκφράζει μία εκτίμηση της διασποράς της μεταβλητής ευπάθειας υπό την παραδοχή της Gamma κατανομής, είναι στην περίπτωσή μας η μεταβλητή "b" όπως είδαμε στο κεφάλαιο $3.4.1$, και προκύπτει ίση με 1.734 και με τυπικό σφάλμα 0.179. Παρατηρούμε λοιπόν μία αξιοσημείωτη 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 α = 5% ($z - statistic = 9.68$).

Φαίνεται ότι όλες οι επεξηγηματικές μεταβλητές είναι στατιστικά σημαντικές για την προσαρμογή του Gamma παραμετρικού μοντέλου στα εν λόγω δεδομένα, με σημαντικότερη την ηλικία και το επίπεδο *LDH* στο αίμα. Επίσης παρατηρούμε ξανά ότι όλοι οι συντελεστές των επεξηγηματικών μεταβλητών έχουν θετικό πρόσημο, γεγονός το οποίο υποδηλώνει ότι αύξηση σε κάποια από τις μεταβλητές όπως στο *LDH*, στην ηλικία, στο mobility status, στο stage και στα nodes, επιδεινώνουν τον κίνδυνο άρα και την επιβίωση των ασθενών.

Να σημειωθεί ότι το αναγκαίο για την R, όρισμα *"id"* αναφέρεται στην περίπτωση που ο χρήστης θέλει να εφαρμόσει Shared Frailty models οπού τα δεδομένα σπάνε σε υποσύνολο δεδομένων που διέπονται από κοινή ευπάθεια. Στην περίπτωσή μας όμως, που δεν μας απασχολεί αυτό, βάζουμε σαν όρισμα το *"id"* που είναι μοναδικό για κάθε ασθενή, και έτσι ο κάθε ασθενής διέπεται από την δική του ευπάθεια.

Συμπερασματικά, βρίσκουμε ότι οι εκτιμήσεις με βάση το Gamma μοντέλο ευπάθειας και εκθετική κατανομή συνάρτησης διακινδύνευσης, όπως αυτά προκύπτουν από την βιβλιοθήκη "parfm" της R, είναι

$$\hat{b} = 1.73\ , \hat{\lambda} = 0.041\ \ \kappa\alpha\iota\ \ \hat{\beta} = (0.931, 1.136, 0.460, 0.675, 0.985)'$$

Με βάση τις εκτιμήσεις των παραμέτρων β οι οποίες είναι όλες θετικές, παρατηρούμε ότι ο κίνδυνος για όλες τις συμμεταβλητές αυξάνει από την ομάδα με τιμή 0 στην ομάδα 1. Πιο συγκεκριμένα:

- Για την μεταβλητή Mobility Status ο κίνδυνος αυξάνει κατά $e^{0.931} = 2.54$ φορές, με όλες τις άλλες συμμεταβλητές fixed, για αυτούς τους ασθενείς που έχουν κινητικά προβλήματα σε σχέση με τους ασθενείς που είναι περιπατητικοί.

- Για την μεταβλητή LDH Level ο κίνδυνος αυξάνει κατά $e^{1.136} = 3.11$ φορές, με όλες τις άλλες συμμεταβλητές fixed, για αυτούς τους ασθενείς που επίπεδο LDH στο αίμα τους πάνω μία φορά του κανονικού σε σχέση με τους ασθενείς που έχουν επίπεδο LDH στο αίμα τους κάτω από μία φορά του κανονικού
- Για την μεταβλητή Nodes ο κίνδυνος αυξάνει κατά $e^{0.460} = 1.58$ φορές, με όλες τις άλλες συμμεταβλητές fixed, για αυτούς τους ασθενείς που έχουν πάνω από μία εστία ασθενείας σε σχέση με τους ασθενείς που έχουν κάτω από μία εστία ασθενείας.
- Για την μεταβλητή Tumor_Stage ο κίνδυνος αυξάνει κατά $e^{0.675} = 1.96$ φορές, με όλες τις άλλες συμμεταβλητές fixed, για αυτούς τους ασθενείς που βρίσκονται στο στάδιο $III - IV$ σε σχέση με τους ασθενείς που βρίσκονται στο στάδιο $I - II$.
- Για την μεταβλητή Age ο κίνδυνος αυξάνει κατά $e^{0.985} = 2.68$ φορές, με όλες τις άλλες συμμεταβλητές fixed, για αυτούς τους ασθενείς που είναι άνω των 60 σε σχέση με τους ασθενείς που είναι κάτω των 60.

4.5.1. Προσαρμογή Παραμετρικών Μοντέλων Ευπάθειας - Περίπτωση Inverse Gaussian Μοντέλου

Στην περίπτωση προσαρμογής του Inverse Gaussian παραμετρικού Μοντέλου ευπάθειας με εκθετική συνάρτηση κινδύνου, τα αποτελέσματα, όπως αυτά παρουσιάζονται στο περιβάλλον της R, φαίνονται στην Εικόνα 20 παρακάτω:

```
```{r}
mod_ig <- parfm(Surv(surv_time, censoring) ~ mobility_status_temp + ldh_level_temp +
                  nodes_temp + tumor_stage_temp + age_temp, cluster="id", data=lymph_temp, dist="exponential",
frailty="ingau")

summary(mod_ig)

mod_ig
```

    ESTIMATE   SE   p-val
 Min.   :0.04764   Min.   :0.007515   Min.   :0.0000000
 1st Qu.:0.54185   1st Qu.:0.122020   1st Qu.:0.0000000
 Median :0.90755   Median :0.130099   Median :0.0000000
 Mean   :1.04017   Mean   :0.182187   Mean   :0.0002339
 3rd Qu.:0.97372   3rd Qu.:0.134276   3rd Qu.:0.0000016
 Max.   :3.29483   Max.   :0.625100   Max.   :0.0011680
                                      NA's   :2

Frailty distribution: inverse Gaussian
Baseline hazard distribution: Exponential
Loglikelihood: -1564.411

                    ESTIMATE SE   p-val
theta   3.295   0.625
lambda   0.048   0.008
mobility_status_temp 0.908   0.126 <.001 ***
ldh_level_temp   0.966   0.130 <.001 ***
nodes_temp   0.430   0.132 0.001 **
tumor_stage_temp   0.654   0.136 <.001 ***
age_temp   0.982   0.118 <.001 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Εικόνα 20: Προσαρμογή Inverse Gaussian παραμετρικού μοντέλου ευπάθειας με εκθετική συνάρτηση κινδύνου στο dataset, χρησιμοποιώντας την βιβλιοθήκη "parfm" της R.

Η εκτίμηση της παραμέτρου *"theta"*, η οποία ουσιαστικά εκφράζει μία εκτίμηση της διασποράς της μεταβλητής ευπάθειας υπό την παραδοχή της Inverse Gaussian κατανομής, είναι στην περίπτωσή μας η ποσότητα $1/2b$ όπως είδαμε στο κεφάλαιο 3.4.2, και προκύπτει ίση με 3.295 και με τυπικό σφάλμα 0.625. Παρατηρούμε πάλι λοιπόν μία μεγάλη εκτίμηση για τη διασπορά της μεταβλητής ευπάθειας και επίσης η μηδενική υπόθεση ότι η διασπορά της ευπάθειας είναι 0 απορρίπτεται σε επίπεδο σημαντικότητας $\alpha = 5\%$, άρα είναι απαραίτητο να εισαγάγουμε τη μεταβλητή της ευπάθειας για να εξηγηθεί η ετερογένεια στον πληθυσμό.

Στην περίπτωση της Inverse Gaussian, όπως είδαμε στο κεφάλαιο 3.4.2, η διασπορά της ευπάθειας ισούται με $\frac{1}{2b} = 3.295$ το οποίο συνεπάγεται ότι η παράμετρος b είναι ίση με 0.15.
```

Φαίνεται ότι όλες οι επεξηγηματικές μεταβλητές είναι στατιστικά σημαντικές για την προσαρμογή του Inverse Gaussian παραμετρικού μοντέλου στα εν λόγω δεδομένα, με σημαντικότερη την ηλικία και δεύτερη την LDH. Επίσης παρατηρούμε ξανά ότι όλοι οι συντελεστές των επεξηγηματικών μεταβλητών έχουν θετικό πρόσημο, γεγονός το οποίο υποδηλώνει ότι αύξηση σε κάποια από τις μεταβλητές όπως στην ηλικία, στο LDH, στο mobility status, στο stage και στα nodes, επιδεινώνουν τον κίνδυνο άρα και την επιβίωση των ασθενών.

Συμπερασματικά, βρίσκουμε ότι οι εκτιμήσεις με βάση το Γάμμα μοντέλο ευπάθειας και εκθετική κατανομή συνάρτησης διακινδύνευσης, όπως αυτές προκύπτουν από την βιβλιοθήκη "parfm" της R, είναι:

$$\hat{b} = 1.5, \hat{\lambda} = 0.048 \; \kappa\alpha\iota \; \hat{\beta} = (0.908, 0.966, 0.430, 0.654, 0.982)'$$

Με βάση τις εκτιμήσεις των παραμέτρων β οι οποίες είναι όλες θετικές, παρατηρούμε ότι ο κίνδυνος για όλες τις συμμεταβλητές αυξάνει από την ομάδα με τιμή 0 στην ομάδα 1. Πιο συγκεκριμένα:

- Για την μεταβλητή Mobility Status ο κίνδυνος αυξάνει κατά $e^{0.908} = 2.48$ φορές, με όλες τις άλλες συμμεταβλητές fixed, για αυτούς τους ασθενείς που έχουν κινητικά προβλήματα σε σχέση με τους ασθενείς που είναι περιπατητικοί.
- Για την μεταβλητή LDH Level ο κίνδυνος αυξάνει κατά $e^{0.966} = 2.62$ φορές, με όλες τις άλλες συμμεταβλητές fixed, για αυτούς τους ασθενείς που έχουν επίπεδο LDH στο αίμα τους πάνω μία φορά του κανονικού σε σχέση με τους ασθενείς που έχουν επίπεδο LDH στο αίμα τους κάτω από μία φορά του κανονικού
- Για την μεταβλητή Nodes ο κίνδυνος αυξάνει κατά $e^{0.430} = 1.54$ φορές, με όλες τις άλλες συμμεταβλητές fixed, για αυτούς τους ασθενείς που έχουν πάνω από μία εστία ασθενείας σε σχέση με τους ασθενείς που έχουν κάτω από μία εστία ασθενείας.
- Για την μεταβλητή Tumor_Stage ο κίνδυνος αυξάνει κατά $e^{0.654} = 1.92$ φορές, με όλες τις άλλες συμμεταβλητές fixed, για αυτούς τους ασθενείς που βρίσκονται στο στάδιο $III - IV$ σε σχέση με τους ασθενείς που βρίσκονται στο στάδιο $I - II$.
- Για την μεταβλητή Age ο κίνδυνος αυξάνει κατά $e^{0.982} = 2.67$ φορές, με όλες τις άλλες συμμεταβλητές fixed, για αυτούς τους ασθενείς που είναι άνω των 60 σε σχέση με τους ασθενείς που είναι κάτω των 60.

Να σημειωθεί ότι το αναγκαίο για την R, όρισμα *“id”* αναφέρεται στην περίπτωση που ο χρήστης θέλει να εφαρμόσει Shared Frailty models οπού τα δεδομένα σπάνε σε υποσέτ δεδομένων που διέπονται από κοινή ευπάθεια. Στην περίπτωσή μας όμως, που δεν μας απασχολεί αυτό, βάζουμε σαν όρισμα το *“id”* που είναι μοναδικό για κάθε ασθενή, και έτσι ο κάθε ασθενής διέπεται από την δική του ευπάθεια.

ΚΕΦΑΛΑΙΟ 5 – ΣΥΓΚΡΙΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

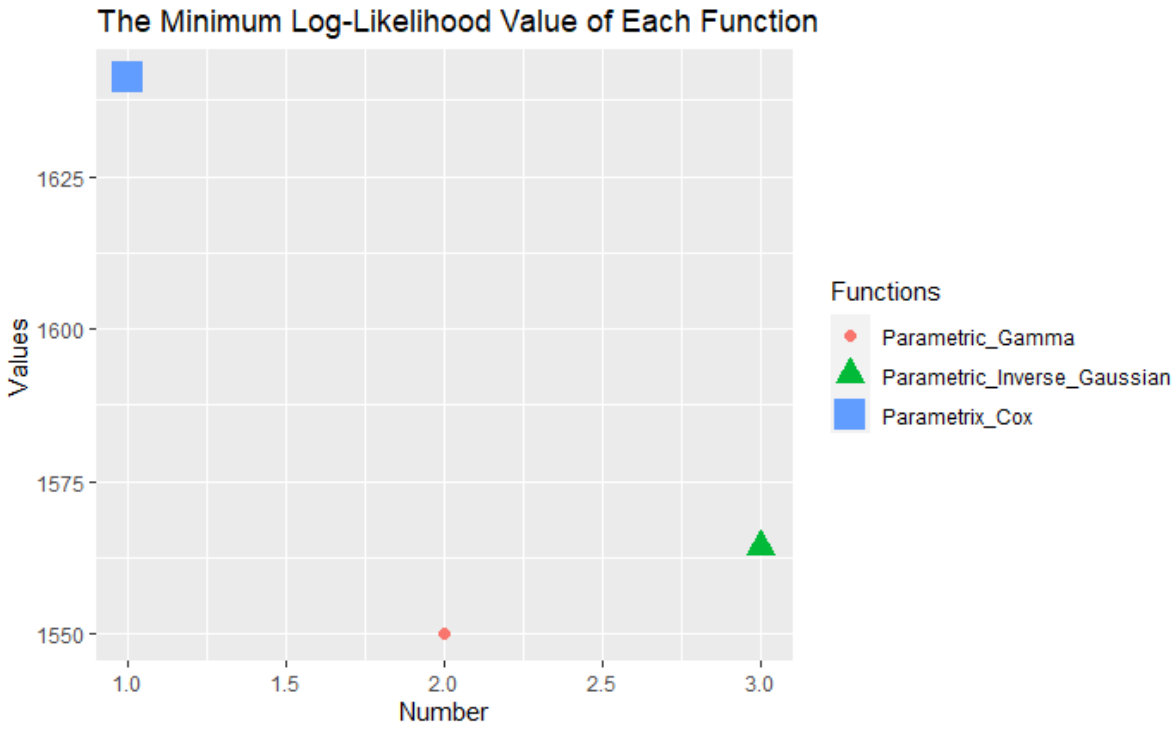
5.1. Σύγκριση Μοντέλων από Αλγορίθμους που Βασίζονται στην Προτεινόμενη Γενικευμένη Συνάρτηση Πιθανοφάνειας.

Τα αποτελέσματά των αλγορίθμων μας μπορούν να συνοψιστούν στον Πίνακα 3 παρακάτω:

Model	b	λ	β					-Log-Likelihood
			Mobility_ Status	Ldh_ Level	Nodes	Tumor_ Stage	Age	
Inverse Gaussian	0.15	0.048	0.909	0.967	0.431	0.655	0.983	1564.413
Gamma	1.7	0.041	0.931	1.133	0.460	0.675	0.985	1550.167
Cox	NA	0.038	0.674	0.63	0.320	0.496	0.752	1641.156

Πίνακας 3: Οι παράμετροι που υπολογίστηκαν στα ελάχιστα των μοντέλων Gamma, Inverse Gaussian όπως και οι παράμετροι που υπολογίστηκαν από το μοντέλο του Cox. Παρατηρούμε ότι την ελάχιστη -Log-Likelihood παρουσιάζει το Gamma Μοντέλο.

Παρατηρούμε ότι ελάχιστη τιμή της -Log-Likelihood παρουσιάζει το Gamma μοντέλο Ευπάθειας. Αυτό μπορεί εύκολα να παρατηρηθεί και στην Εικόνα 16 παρακάτω:



Εικόνα 16: σύγκριση των μοντέλων βάση της τιμής της - LogLikelihood. Παρατηρείται ότι η Gamma ευπάθεια είναι η πιο κατάλληλη να περιγράψει τα δεδομένα. Ακολουθεί η Inverse Gaussian και έπειτα το μοντέλο του Cox.

Συνεπώς συμπεραίνουμε ότι το Gamma Μοντέλο ευπάθειας περιγράφει πιο αποτελεσματικά τα δεδομένα αφού παρουσιάζει την ελάχιστη -Log-Likelihood από τα τρία παραμετρικά μοντέλα που συγκρίνουμε. Είναι ενδιαφέρον να παρατηρήσουμε εδώ ότι στο ίδιο συμπέρασμα είχαν και καταλήξει και οι Kosorok et al. (2004) στην ανάλυση αυτών των δεδομένων βάσει ημιπαραμετρικών μοντέλων ευπάθειας, όπως και οι Hanagal & Dabade (2014), ότι δηλαδή η Gamma κατανεμημένη ευπάθεια περιέγραφε τα δεδομένα με τον βέλτιστο τρόπο

5.2. Σύγκριση Μοντέλων από την Βιβλιοθήκη "parfm".

Τα αποτελέσματα του αλγορίθμου που κατασκευάσαμε για την Inverse Gaussian κατανομή με βάση την προτεινόμενη γενικευμένη συνάρτηση πιθανοφάνειας που δίνεται στη σχέση (36), σε σχέση με τα αποτελέσματα της βιβλιοθήκης *"parfm" για* Inverse Gaussian παραμετρικό μοντέλο ευπάθειας με εκθετική συνάρτηση κινδύνου φαίνονται στον Πίνακα 4 παρακάτω:

Inverse Gaussian	**b**	**λ**	***β***					**-Log-Likelihood**
			Mobility_ Status	**Ldh_ Level**	**Nodes**	**Tumor_ Stage**	**Age**	
Algorithm	0.15	0.048	0.909	0.967	0.431	0.655	0.983	1564.413
parfm Library	0.15	0.048	0.908	0.966	0.430	0.654	0.982	1564.411
Differences	0.00%	0.00%	0.11%	0.10%	0.23%	0.15%	0.10%	0.00%

Πίνακας 4: Τα αποτελέσματα του αλγορίθμου που κατασκευάσαμε για την Inverse Gaussian κατανομή σε σχέση με τα αποτελέσματα της βιβλιοθήκης "parfm" για Inverse Gaussian παραμετρικό μοντέλο ευπάθειας με εκθετική συνάρτηση κινδύνου

Τα αποτελέσματα του αλγορίθμου που κατασκευάσαμε για την Gamma κατανομή με βάση την προτεινόμενη γενικευμένη συνάρτηση πιθανοφάνειας που δίνεται στη σχέση (36), σε σχέση με τα αποτελέσματα της βιβλιοθήκης *"parfm"* για Gamma παραμετρικό μοντέλο ευπάθειας με εκθετική συνάρτηση κινδύνου φαίνονται στον Πίνακα 5 παρακάτω:

Gamma	b	λ	β					-Log-Likelihood
			Mobility_Status	Ldh_Level	Nodes	Tumor_Stage	Age	
Algorithm	1.7	0.041	0.931	1.133	0.46	0.675	0.985	1550.167
parfm Library	1.7	0.041	0.931	1.136	0.46	0.675	0.985	1550.148
Difference (%)	0.00%	0.00%	0.00%	0.26%	0.00%	0.00%	0.00%	0.00%

Πίνακας 5: Τα αποτελέσματα του αλγορίθμου που κατασκευάσαμε για την Inverse Gaussian κατανομή σε σχέση με τα αποτελέσματα της βιβλιοθήκης "parfm" για Inverse Gaussian παραμετρικό μοντέλο ευπάθειας με εκθετική συνάρτηση κινδύνου

Παρατηρούμε ότι τα αποτελέσματα είναι σχεδόν ίδια και στις δύο περιπτώσεις καθώς οι ποσοστιαίες διαφορές των υπολογιζόμενων τιμών του διανύσματος β, της παραμέτρου b και της $-Loglikelihood$, μεταξύ του προτεινόμενου αλγορίθμου και της βιβλιοθήκης, όταν υπάρχουν, είναι κάτω από 1%.

Αυτό το γεγονός πιστοποιεί ότι η μέθοδός μας προσέγγισε πολύ αποτελεσματικά την μέθοδο της έτοιμης βιβλιοθήκης της R.

Από πλευράς χρόνου, ο αλγόριθμός μας λειτουργεί λίγο πιο αργά από εκείνον της R αλλά πάντα υπάρχει περιθώριο βελτίωσης και μπορούμε να εισάγουμε οποιαδήποτε κατανομή ευπάθειας. Αυτό το σημείο παραμένει ανοικτό για περαιτέρω διερεύνηση στο μέλλον. Η βιβλιοθήκη "Parfm" έχει σαν δυνατότητες κατανομής της ευπάθειας αυτή τη στιγμή, μόνο τις συναρτήσεις "Gamma", "Inverse Gaussian", "Positive Stable" και "lognormal".

Λαμβάνοντας υπόψιν όμως την ραγδαία εξέλιξη του πεδίου, είναι χρήσιμο να υπάρξει ένα εργαλείο που να επιτρέπει την εισαγωγή οποιασδήποτε κατανομής ευπάθειας, όπως αυτό που δημιουργήσαμε στα πλαίσια της εργασίας αυτής.

Τελος

Βιβλιογραφία

Abbring, J. and van den Berg, G.J. (2007) The unobserved heterogeneity distribution in duration analysis. *Biometrika* 94, 87–99.

Adeleke, K.A. (2019) Parametric Frailty Models for Clustered Survival Data: Application to Recurrent Asthma Attack in Infants, Vol. 6, No. 3, PP:89-99. doi:10.18576/jsapl/060301

Anderson, J.E. and Louis, T.A. (1995) Survival analysis using a scale change random effects model. *Journal of the American Statistical Association* 90, 669–679.

Androulakis, E., Koukouvinos, C. and Vonta, F., (2012) Estimation and variable selection via frailty models with penalized likelihood, *Statistics in Medicine* 31, 2223–2239.

Ayele DG, Zewotir TT and Mwambi H. (2017). Survival analysis of under-five mortality using Cox and frailty models in Ethiopia. *J Health Popul Nutr*. Jun 2 2017;36(1):25. doi: 10.1186/s41043-017-0103-3. PMID: 28578680; PMCID: PMC5455089.

Beard, R.E. (1959). Note on some mathematical mortality models. In: The Lifespan of Animals. G.E.W. Wolstenholme, M.O'Conner (eds.), *Ciba Foundation Colloquium on Ageing*, Little, Brown, Boston, 302–311.

Bohdan Nosyk, Ying C. MacNab, Huiying Sun, Benedikt Fischer, David C. Marsh, Martin T. Schechter and Aslam H. Anis, (2009). Proportional Hazards Frailty Models for Recurrent Methadone Maintenance Treatment, *American Journal of Epidemiology*, Volume 170, Issue 6, 15 September 2009, Pages 783–792, https://doi.org/10.1093/aje/kwp186

Calsavara VF, Milani EA, Bertolli E and Tomazella V. (2020). Long-term frailty modeling using a non-proportional hazards model: Application with a melanoma dataset. *Statistical Methods in Medical Research*. 2020;29(8):2100-2118. doi:10.1177/0962280219883905

Carbone PP, Kaplan HS, Musshoff K, Smithers DW and Tubiana M. (1971). Report of the Committee on Hodgkin's Disease Staging Classification. *Cancer Res* 1971;31:1860-1861

Chang, S.-H. (2004). Estimating marginal effects in accelerated failure time models for serial sojurn times among repeated events. *Lifetime Data Analysis* 10, 175–190.

Chen, X., Ghysels, E., and Telfeyan, R. (2016). Frailty models for commercial mortgages. *Journal of Fixed Income*, 26(2), 16-31. 10.3905/jfi.2016.26.2.016

Clayton, D.G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 65, 141–151.

Clayton, DG. and Cuzick, J., (1986). The semi-parametric Pareto model for regression analysis of survival times. In: *Papers on semiparametric models* MS-R8614. Centrum voor Wiskunde en Informatica, Amsterdam, pp 19–31

Congdon, P. (1995). Modelling frailty in area mortality. *Statistics in Medicine* 14, 1859–1874.

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

Cox, D.R. (1975). Partial Likelihood. *Biometrika*, 62, 269-276.

Dos Santos, D., Davies, R and Francis, B. (1995). Nonparametric hazard versus nonparametric frailty distribution in modelling recurrence of breast cancer. *Journal of Statistical Planning and Inference* 47, 111–127.

Drapeau, M.D., Gassa, E.K., Simisona, M.D., Muellera, L.D. and Rosea, M.R. (2000). Testing the heterogeneity theory of late-life mortality plateaus by using cohorts of Drosophila melanogaster. *Experimental Gerontology* 35, 71–84.

Duchateau, L. and Janssen, P. (2008). The Frailty Model. *Springer*, New York.

Esayas Lelisho, M., Akessa, G. M., Kifle Demissie, D., Fikadu Yermosa, S., Andargie, S. A., Tareke, S. A., and Pandey, D. (2022). Application of Parametric Shared Frailty Models to Analyze Time-to-Death of Gastric Cancer Patients. *Journal of gastrointestinal cancer*, 10.1007/s12029-021-00775-y. Advance online publication. https://doi.org/10.1007/s12029-021-00775-y.

Fan J and Li R. (2002). Variable selection for Cox's proportional hazards model and frailty model. The *Annals of Statistics*; 30:74–99.

Ferreira R. and Colugnati F., (2021). Monitoring chronic kidney disease evolution using frailty models, *International Journal of Epidemiology*, 30:139-168, https://doi.org/10.1093/ije/dyab168.139

Ghadimi MR, Rasouli M, Mahmoodi M and Mohammad K. (2011). Prognostic factors for the survival of patients with esophageal cancer in Northern Iran. *J Res Med Sci*. 2011 Oct;16(10):1261-72. PMID: 22973319; PMCID: PMC3430015.

Greenwood, M. and Yule, G.U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society* 83, 255–279.

Gurmu, S.E. (2018). Assessing Survival Time of Women with Cervical Cancer Using Various Parametric Frailty Models: A Case Study at Tikur Anbessa Specialized Hospital, Addis Ababa, Ethiopia. *Ann. Data. Sci*. 5, 513–527. https://doi.org/10.1007/s40745-018-0150-7

Hanagal, D.D., and Dabade, A.D. (2014). Comparisons of frailty models for kidney infection data under Weibull baseline distribution. *Int. J. Math. Model*. Numer. Optimisation, 5, 342-373.

Hanagal D. D. and Dabade A. D., (2015). Comparison of Shared Frailty Models for Kidney Infection Data under Exponential Power Baseline Distribution, *Communications in Statistics - Theory and Methods*, Taylor & Francis Journals, vol. 44(23), pages 5091-5108, December.

Heckman, J.J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* 52, 271–320.

Heckman, J.J. and Taber, C.R. (1994). Econometric mixture models and more general models for unobservables in duration analysis. *Statistical Methods in Medical Research* 3, 279–302.

Horowitz, J.L. (1999). Semiparametric estimation of a proportional hazard model with unobserved heterogeneity. *Econometrica* 67, 1001–1028.

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

Hougaard, P., (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika* 73:387–396

Huber-Carol, C. and Vonta, F. (2004). Frailty models for arbitrarily censored and truncated Data, *Lifetime Data Anal*., 10, 369–388.

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

Keiding, N., Andersen, P. and Klein, J. (1997). The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in Medicine* 16, 215–224.

Khan, J.R. and Awan, N. (2017). A comprehensive analysis on child mortality and its determinants in Bangladesh using frailty models. *Arch Public Health* 75, 58. https://doi.org/10.1186/s13690-017-0224-6.

Khazaeli, A., Xiu, L. and Curtsinger, J.W. (1995). Stress experiments as a means of investigating age-specific mortality in Drosophila melanogaster. *Experimental Gerontology* 30, 177–184.

Kim, B., Ha, I. D., and Lee, D. (2016). Analysis of multi-center bladder cancer survival data using variable-selection method of multi-level frailty models. *Journal of the Korean Data and Information Science Society*. Korean Data and Information Science Society. https://doi.org/10.7465/jkdi.2016.27.2.499.

Klein, J. P. and Moeschberger, M. L., (2003). Survival Analysis Techniques for Censored and Truncated Data. United States of America: *Springer-Verlag* New York.

Klein, J.P., Pelz, C. and Zhang, M.-J. (1999). Random effects for censored data by a multivariate normal regression model. *Biometrics* 55, 497–506.

Komarek, A., Lesaffre, E. and Legrand, C. (2007). Baseline and treatment effect heterogeneity for survival times between centers using a random effects accelerated failure time model with flexible error distribution. *Statistics in Medicine* 26, 5457–5472.

Kortram, R. A., Lenstra, A. J., Ridder, G. and van Rooij, A. C. M. (1995). Constructive identification of the mixed proportional hazards model. *Statistica Neerlandica* 49, 269–281.

Kosorok, R. Michael, Bee Leng Lee and Jason P. (2004). Fine "Robust inference for univariate proportional hazards frailty regression models," *The Annals of Statistics*, Ann. Statist. 32(4), 1448-1491.

Lam, K. and Lee, Y. (2004). Merits of modelling multivariate survival data using random effects proportional odds model. *Biometrical Journal* 46, 331–42.

Lam, K.F., Lee, Y.W. and Leung, T.L. (2002). Modeling multivariate survival data by a semiparametric random effects proportional odds model. *Biometrics* 58, 316–323.

Lambert, P., Collett, D., Kimber, A. and Johnson, R. (2004). Parametric accelerated failure time models with random effects and an application to kidney transplant survival. *Statistics in Medicine* 23, 3177–3192.

Lancaster, T. (1979). Econometric methods for the duration of unemployment. *Econometrica* 47, 939–956.

Li Y, Wileyto EP, and Heitjan DF. (2011). Prediction of individual long-term outcomes in smoking cessation trials using frailty models. *Biometrics*. 2011 Dec;67(4):1321-9. doi: 10.1111/j.1541-0420.2011.01578.x. Epub 2011 Mar 14. PMID: 21401566.

Yalew M, Arefaynie M, Bitew G, Amsalu ET and Kefale B, et al. (2022). Time to under-five mortality and its predictors in rural Ethiopia: Cox-gamma shared frailty model. *Plos One* 17(4): e0266595. https://doi.org/10.1371/journal.pone.0266595

Morley, E., Perry, H. M. and Miller, D. K. (2002). Something about frailty. *Journal of Gerontology: Medical Sciences* 57A, M698–M704.

Moon, T. K. (1996). "The expectation-maximization algorithm," in *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47-60, doi: 10.1109/79.543975.

Mueller, L.D., Drapeau, M.D., Adams, C.S., Hammerle, C.W., Doyal, K.M., Jazayeri, A.J., Ly, T., Beguwala, S.A., Mamidi, A.R. and Rose, M.R. (2003). Statistical tests of demographic heterogeneity theories. *Experimental Gerontology* 38, 373–386.

Murphy, S.A., Rossini, A. and van der Vaart, A.W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association* 92, 968–976.

Pan, W. (2001). Using frailties in the accelerated failure time model. *Lifetime Data Analysis* 7, 55–64.

Rocha, C.S. (1996). Survival models for heterogeneity using the non-central chi-squared distribution with zero degrees of freedom. In: *Lifetime Data: Models in Reliability and Survival Analysis*. Jewell, N. et al. (eds.) Kluwer Academic Publishers, Dordrecht, pp. 275–279.

Rockwood, K. (2005) Frailty and its definition: a worthy challenge. Journal of the American Geriatric Society 53, 1069–1070.

Rockwood, K. and Mitnitski, A. (2007). Frailty in relation to the accumulation of deficits. *Journal of Gerontology* (A) 62, 722–727.

Schnier, C., Hielm, S. and Saloniemi, H.S. (2004). Comparison of the breeding performance of cows in cold and warm loose-housing systems in Finland. *Preventive Veterinary Medicine* 62, 135–151.

Shipp, M., Harrington, D. and Anderson, J., (1993). Development of a Predictive Model for Aggressive lymphoma. The International Non-Hodgkin's Lymphoma Prognostic Factors Project, *New England Journal of Medicine*, in press, pp. 987-994.

Silva, G.L. and Amaral-Turkman, M.A. (2004). Bayesian analysis of an additive survival model with frailty. *Communications in Statistics – Theory and Methods* 33, 2517–2533.

Slud, E. V. and Vonta, F. (2004), Consistency of the NPML Estimator in the Right-Censored Transformation Model, *Scand. J. Statist*., 31, 21-41

International Non-Hodgkin's Lymphoma Prognostic Factors Project. A predictive model for aggressive non-Hodgkin's lymphoma. (1993). *N Engl J Med*. 1993 Sep 30;329(14):987-94. doi: 10.1056/NEJM199309303291402. PMID: 8141877.

Jonker MA, Bhulai S, Boomsma DI, Ligthart RS, Posthuma D and Van der Vaart AW. Gamma frailty model for linkage analysis with application to interval-censored migraine data. (2009). *Biostatistics*. 2009 Jan;10(1):187-200. doi: 10.1093/biostatistics/kxn027. Epub 2008 Aug 19. PMID: 18714083.

Tomazella, V., Louzada-Neto, F. and da Silva, G.L. (2006). Bayesian modeling of recurrent events data with an additive gamma frailty distribution and a homogeneous Poisson process. Journal of Statistical Theory and Applications 5, 417–429.

Tomazella, V., Louzada-Neto, F., and da Silva, G.L. (2006). Bayesian modeling of recurrent events data with an additive gamma frailty distribution and a homogeneous Poisson process. *Journal of Statistical Theory and Applications* 5, 417–429.

Van Den Berg, G.J. (2001). Duration models: specification, identification, and multiple durations. In: *Handbook of Econometrics*. (Volume V) J.J. Heckman, E. Leamer (eds.) North-Holland, Amsterdam.

Vaupel, J., Manton, K. and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* 16, 439–54.

Vaupel, J.W. and Yashin, A.I. (1985). Heterogeneity's ruses: some surprising effects of selection on population dynamics. *The American Statistician* 39, 176–185.

Vonta, F., (1996). Efficient estimation in a non-proportional hazards model in survival analysis. *Scand J Stat* 23:49–61.

Vonta, F. and Karagrigoriou, A. (2007), Variable selection strategies in survival models with multiple imputations, Lifetime Data Analysis, 13, Issue 3, 295-315.

Vonta, F. and Karagrigoriou, A. (2010), Generalized Measures of Divergence in Survival Analysis and Reliability, *Journal of Applied Probability*, Vol. 47, No. 1 (pp. 216-234) published by Applied Probability Trust.

Vonta, I. and Karagrigoriou, A. (2014), Goodness-of-fit tests via ϕ-measures of divergence for censored data, *Journal of Statistical Computation and Simulation*, DOI:10.1080/00949655.2012.733396.

Wand H and Ramjee G. (2015). Biological impact of recurrent sexually transmitted infections on HIV seroconversion among women in South Africa: results from frailty models. *J Int AIDS Soc*. 2015 Apr 24;18(1):19866. doi: 10.7448/IAS.18.1.19866. PMID: 25912181; PMCID: PMC4410128.

Winke, A. (2010). Frailty models in survival analysis. *Chapman and Hall/CRC*, Florida.

Wu, D., Rea, S.L., Yashin, A.I. and Johnson, T.E. (2006). Visualizing hidden heterogeneity in isogenic populations of C. elegans*. Experimental Gerontology* 41, 261–270.

Xu, L., Zhang, J. (2010). An EM-like algorithm for the semiparametric accelerated failure time gamma frailty model. *Computational Statistics and Data Analysis* 54, 1467–1474.

Yazdani A, Yaseri M, Haghighat S, Kaviani A and Zeraati H. (2019). Investigation of Prognostic Factors of Survival in Breast Cancer Using a Frailty Model: A Multicenter Study. Breast Cancer: Basic and Clinical Research. 2019; 13. doi:10.1177/1178223419879112.

Zhang, J. and Peng, Y. (2007). An alternative estimation method for the accelerated failure time frailty model. *Computational Statistics and Data Analysis* 51, 4413–4423.