Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Εφαρμοσμενων Μαθηματικων και Φυσικων Επιστημων
Τομεας Μαθηματικων

Παλινδρόμηση Κορυφογραμμής, Τεχνική Lasso και Δέντρα Επιβίωσης σε μοντέλο αναλογικής διακινδύνευσης του Cox

Διπλωματικη Εργασια

του

ΤΟΥΦΕΞΗ ΓΕΩΡΓΙΟΥ ΧΡΗΣΤΟΥ

Επιβλέπουσα: Χ. Καρώνη-Ρίτσαρντσον
Καθηγήτρια, ΕΜΠ

Αθήνα, Ιούλιος 2021

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Εφαρμοσμενων Μαθηματικων και Φυσικων Επιστημων
Τομεας Μαθηματικων

Παλινδρόμηση Κορυφογραμμής, Τεχνική Lasso και Δέντρα Επιβίωσης σε μοντέλο αναλογικής διακινδύνευσης του Cox

Διπλωματικη Εργασια

του

ΤΟΥΦΕΞΗ ΓΕΩΡΓΙΟΥ ΧΡΗΣΤΟΥ

Επιβλέπουσα: Χ. Καρώνη-Ρίτσαρντσον
Καθηγήτρια, ΕΜΠ

Η Επιβλέπουσα	**Μέλος 1**	**Μέλος 2**
Χ. Καρώνη-Ρίτσαρντσον	Β. Παπανικολάου	Κ. Παυλοπούλου
Καθηγήτρια, ΕΜΠ	Καθηγητής, ΕΜΠ	Ε.ΔΙ.Π., ΕΜΠ

Αθήνα, Ιούλιος 2021

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Εφαρμοσμενων Μαθηματικων και Φυσικων Επιστημων
Τομεας Μαθηματικων

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας πτυχιακής εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην πτυχιακή μου εργασία και κατά συνέπεια αποτυχία απόκτησης του τίτλου σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η πτυχιακή εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

(Υπογραφή)

.....
Τουφεξής Γεώργιος
Χρήστος

14 Ιουλίου 2021

Περίληψη

Η ανάλυση επιβίωσης εξετάζει και μοντελοποιεί τον χρόνο που παίρνει ένα γεγονός να γίνει. Το μοντέλο αναλογικής διακινδύνευσης του Cox είναι ένα από τα πιο σημαντικά και διαδεδομένα εργαλεία για την μελέτη της εξάρτησης του χρόνου επιβίωσης με τις συμμεταβλητές του μοντέλου. Στη σύγχρονη εποχή έχουν αναπτυχθεί εναλλακτικές τεχνικές που μας βοηθούν να αντιμετωπίσουμε το πρόβλημα της πολυσυγγραμικότητας και να μειωθεί η διασπορά των συντελεστών καθώς η ανάγκη για καλύτερη πρόβλεψη παρά το μεγάλο αριθμό συμμεταβλητών είναι μεγάλη. Στα πλαίσια της παρούσας εργασίας θα γίνει σύντομη παρουσίαση αυτών των τεχνικών, θα υλοποιηθούν σε περιβάλλον R και θα γίνει σύγκριση των διάφορων τεχνικών με τη χρήση πραγματικών δεδομένων. Πιο συγκεκριμένα, θα χρησιμοποιηθούν οι τεχνικές Ridge και Lasso καθώς και τεχνικές που βασίζονται στα δέντρα επιβίωσης (Survival Trees).

Στόχος της διπλωματικής εργασίας είναι να δούμε αν όλες οι διαφορετικές τεχνικές θα μας οδηγήσουν στο ίδιο τελικό μοντέλο παλινδρόμησης και να εξάγουμε πιο ολοκληρωμένα και σωστά συμπεράσματα για τα δεδομένα μας μέσω των διαφορετικών τεχνικών.

Λέξεις Κλειδιά

Ανάλυση Επιβίωσης, Μοντέλο αναλογικής διακινδύνευσης του Cox, Δέντρα παλινδρόμησης, Μέθοδοι συρρίκνωσης, Τεχνική Ridge, Τεχνική Lasso, Δέντρα αποφάσεων

Abstract

Survival analysis examines and models the time it takes until an event happens. The Cox proportional-hazards regression model is one of the most important and popular tools for studying the dependency of the survival time on the covariates of the model. Modern methodology includes several alternative techniques that have been developed that help us tackle the problem of multicollinearity and reduce the variance of the regression coefficients as the need for better prediction despite the number of covariates is high. In this diploma thesis, these techniques are presented, followed by comparison of the results of their application to real data using the R environment. Ridge Lasso and techniques that are based on Survival Trees will be used in particular.

This diploma thesis aims to see if all the different techniques will lead us to the same final regression model and to draw more complete and correct conclusions for our data using the different techniques.

Keywords

Survival analysis, Cox proportional-hazards regression model, Regression Trees , Shrinkage Methods , Ridge technique, Lasso Technique, Decision Trees

στους γονείς μου

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω την καθηγήτρια Καρώνη Χρυσηίς για την επίβλεψη αυτής της διπλωματικής εργασίας και για την εξαιρετική συνεργασία που είχαμε. Θα ήθελα επίσης να ευχαριστήσω την οικογένεια, τους φίλους μου και την κοπέλα μου για την καθοδήγηση και την ηθική συμπαράσταση που μου προσέφεραν καθόλη την διάρκεια των σπουδών μου.

Αθήνα, Ιούλιος 2021

Τουφεξής Γεώργιος Χρήστος

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Πρόλογος

Η παρούσα διπλωματική εργασία με τίτλο «**Παλινδρόμηση κορυφογραμμής, Τεχνική Lasso και Δέντρα Επιβίωσης σε μοντέλο αναλογικής διακινδύνευσης του Cox**» εκπονήθηκε κατά την θερινή περίοδο του ακαδημαϊκού έτους 2020-2021 σε συνεργασία με τον τομέα των μαθηματικών της σχολής Ε.Μ.Φ.Ε υπό την επίβλεψη της καθηγήτριας του Ε.Μ.Π Χ. Καρώνη-Ρίτσαρντσον.

Κεφάλαιο 1

Εισαγωγή

Η ανάλυση επιβίωσης είναι ένας κλάδος της στατιστικής και αναφέρεται στην ανάλυση δεδομένων που αφορούν τον χρόνο μέχρι να συμβεί κάποιο γεγονός. Αυτό το γεγονός σχετίζεται συνήθως με ένα ανεπιθύμητο ενδεχόμενο.

Αρχικά, εξού και το όνομα, αναφερόταν για ασθενείς και για χρόνους θανάτου μετά την έναρξη κάποιας θεραπείας. Έπειτα όμως άρχισε να χρησιμοποιείται και σε άλλους κλάδους όπως η μηχανική ή η γεωργία και εξέταζε γεγονότα όπως η βλάβη μιας μηχανής, η θραύση κάποιου αντικειμένου, ο χρόνος για να αναπτυχθεί ένας καρπός σε ένα δέντρο κλπ. Για αυτήν την ανάλυση έχει αναπτυχθεί μια στατιστική θεωρία με το όνομα ανάλυση αξιοπιστίας όταν πρόκειται για θετικές επιστήμες ή ανάλυση επιβίωσης, όταν πρόκειται για βιοϊατρική.

Συνήθως η μεταβλητή που μας ενδιαφέρει είναι ο χρόνος, αλλά καλύπτονται και περιπτώσεις τυχαίων μεταβλητών που δεν αφορούν μόνο τον χρόνο όπως για παράδειγμα το φορτίο που ασκείται σε κάποιο υλικό ή η παραμόρφωση ενός υλικού . Παρ'όλα αυτά χρησιμοποιείται ο όρος **διάρκεια ζωής** γιατί συνήθως μας ενδιαφέρει ο χρόνος μέχρις ότου προκύψει οποιοδήποτε γεγονός.

Γενικότερα η ανάλυση επιβίωσης πραγματεύεται τυχαίες μεταβλητές που παίρνουν θετικές τιμές όπως:

- Χρόνος θανάτου ενός ασθενούς.
- Χρόνος αποβολής ενός μοσχεύματος.
- Χρόνος παραμονής σε νοσοκομείο.
- Χρόνος αστοχίας κάποιου οργάνου.
- Χρόνος από διάγνωση του HIV μέχρι την ανάπτυξη του AIDS.
- Χρόνος μεταξύ δύο ανακοπών καρδιάς.
- Χρόνος μέχρι να υπάρξει βλάβη μιας μηχανής.
- Χρόνος μετά από θεραπεία κατά του καρκίνου μέχρι τον θάνατο.

Οι χρόνοι επιβίωσης δεν αφορούν μόνο ανθρώπους αλλά μπορούν να αφορούν ζώα, φυτά καθώς και μηχανικά εξαρτήματα.

Καταλαβαίνουμε από τα παραπάνω πόσο σημαντική είναι η ανάλυση επιβίωσης καθώς και το πόσο σημαντικό να έχουμε τα κατάλληλα εργαλεία και τεχνικές για να είναι η ανάλυση των δεδομένων μας ακριβής ώστε να μπορούμε να κάνουμε σωστές προβλέψεις.

1.1 Αντικείμενο της διπλωματικής

Το βασικό ζήτημα που προκύπτει είναι αν υπάρχουν θεωρίες και τεχνικές που θα μας βοηθήσουν την στατιστική μας ανάλυση να είναι σωστή. Ένα από τα πιο σημαντικά και διαδεδομένα εργαλεία στην ανάλυση επιβίωσης είναι το μοντέλο αναλογικής διακινδύνευσης του Cox, που είναι ένα ημι-παραμετρικό μοντέλο παλινδρόμησης. Στην παρούσα διπλωματική εργασία, αφού πρώτα παρουσιάσουμε ένα θεωρητικό υπόβαθρο, θα χρησιμοποιήσουμε διάφορες τεχνικές σε αληθινά δεδομένα μέσω του προγραμματιστικού περιβάλλοντος R για να κάνουμε μια εμπεριστατωμένη ανάλυση των δεδομένων. Πιο συγκεκριμένα:

1. Θα κάνουμε την βασική μη-παραμετρική ανάλυση των δεδομένων μας (Kaplan-Meier, Log-rank test).
2. Θα προσαρμόσουμε τα δεδομένα μας στο μοντέλο αναλογικής διακινδύνευσης του Cox.
3. Θα χρησιμοποιήσουμε τις λεγόμενες μεθόδους συρρίκνωσης Ridge, Lasso.
4. Θα φτιάξουμε ένα Survival Tree για το μοντέλο παλινδρόμησης μας.

Σκοπός της εργασίας είναι μέσω όλων των παραπάνω να κάνουμε μια στατιστική ανάλυση και να δούμε αν όλες οι τεχνικές μας βοηθάνε για ένα σωστό αποτέλεσμα και θα χρησιμοποιήσουμε τεχνικές που χρησιμοποιούνται σχεδόν σε όλα τα μοντέλα παλινδρόμησης στο μοντέλο της αναλογικής διακινδύνευσης του Cox.

1.2 Οργάνωση του τόμου

Η εργασία αυτή είναι οργανωμένη σε έξι κεφάλαια. Στο **Κεφάλαιο 2** δίνεται το θεωρητικό υπόβαθρο των βασικών θεωριών και τεχνικών που σχετίζονται με τη διπλωματική αυτή. Αρχικά περιγράφονται η εκτιμήτρια Kaplan-Meier καθώς και ο έλεγχος Log-rank. Έπειτα γίνεται μια αναφορά στο μοντέλο αναλογικής διακινδύνευσης του Cox καθώς και στους ελέγχους υποθέσεων της αναλογικής διακινδύνευσης (PH Assumption) και στα υπόλοιπα στο μοντέλο του Cox. Στη συνέχεια, αναλύονται δύο μέθοδοι συρρίκνωσης οι Ridge, Lasso και τέλος αναφέρονται τα βασικά των δέντρων αποφάσεων όπως τα δέντρα παλινδρόμησης και τα δέντρα ταξινόμησης. Στο **Κεφάλαιο 3** γίνεται η περιγραφή του θέματος και στη συνέχεια δίνεται ο στόχος της συγκεκριμένης εργασίας. Στο **Κεφάλαιο 4** παρουσιάζονται και περιγράφονται τα δεδομένα του πειράματος μας. Στο **Κεφάλαιο** 5 γίνεται η βασική επεξεργασία των δεδομένων του πειράματος μας στο προγραμματιστικό περιβάλλον της R. Τέλος, στο **Κεφάλαιο 6** εξάγονται τα συμπεράσματα καθώς και μελλοντικές επεκτάσεις της παρούσας διπλωματικής εργασίας.

Μέρος Ι

Θεωρητικό Μέρος

Κεφάλαιο 2

Θεωρητικό υπόβαθρο

Στο κεφάλαιο αυτό παρουσιάζονται όλα τα θεωρητικά εργαλεία που έχουν σχέση με την εργασία αυτή, δηλαδή η εκτιμήτρια Kaplan-Meier, ο έλεγχος Log-Rank βασικά πράγματα στο μοντέλο αναλογικής διακινδύνευσης του Cox , μέθοδοι συρρίκνωσης Ridge, Lasso και τα βασικά των δέντρων αποφάσεων.

2.1 Μη-παραμετρική Ανάλυση Δεδομένων Διάρκειας Ζωής

2.1.1 Η εκτιμήτρια Kaplan-Meier

Πολύ σημαντικό στην ανάλυση δεδομένων διάρκειας ζωής είναι να δούμε ποιο είναι το κατάλληλο θεωρητικό μοντέλο που προσαρμόζεται στα δεδομένα μας. Συνήθως το πρώτο βήμα είναι η κατασκευή γραφικών παραστάσεων που μας δείχνουν πώς συμπεριφέρεται η συνάρτηση επιβίωσης και η συνάρτηση διακινδύνευσης.

Αρχικά ας δούμε πως θα μπορούσαμε να εκτιμήσουμε την συνάρτηση επιβίωσης. Έστω ότι έχουμε ένα δείγμα χρόνων επιβίωσης όπου όλες οι παρατηρήσεις είναι μη-αποκομμένες. Η συνάρτηση επιβίωσης S(t) είναι η πιθανότητα να επέζησε ένα άτομο για μια τιμή του χρόνου μεγαλύτερη ή ίση του t. Αυτή η συνάρτηση μπορεί να εκτιμηθεί ως εξής:

$$\hat{S}(t) = \frac{\text{Αριθμός ατόμων με χρόνο επιβίωσης} \geq t}{\text{Αριθμός ατόμων στα δεδομένα μας}} \tag{2.1}$$

ή αλλιώς $\hat{S}(t) = 1 - \hat{F}(t)$, όπου $\hat{F}(t)$ είναι συνάρτηση κατανομής (εμπειρική). Παρατηρούμε ότι πριν τον πρώτο θάνατο η $\hat{S}(t)$ ισούται με την μονάδα και μετά τον τελευταίο θάνατο ισούται με 0. Επίσης μεταξύ δύο θανάτων η συνάρτηση είναι σταθερή άρα καταλήγουμε ότι η $\hat{S}(t)$ είναι μια κλιμακωτή συνάρτηση.

Χρησιμοποιώντας την 2.1 παρατηρούμε ότι προκύπτει ένα πρόβλημα. Αυτή η μέθοδος δεν μπορεί να χρησιμοποιηθεί για να εκτιμήσει αποκομμένες παρατηρήσεις και επειδή το φαινόμενο της αποκοπής των δεδομένων (συνήθως από δεξιά) είναι πολύ συχνό, χρησιμοποιείται μια εκτιμήτρια που είναι πλέον από τις πιο σημαντικές που μπορεί να αντιμετωπίσει το πρόβλημα της αποκοπής. Αυτή η εκτιμήτρια είναι η **Kaplan-Meier** (Kaplan and Meier, 1958).

Έστω τυχαίο δείγμα n ατόμων, μερικοί εκ των οποίων πεθαίνουν κατά τις διακεκριμένες χρονικές στιγμές $t_{(1)} < t_{(2)} < ... < t_{(k)}$, $k \leq n$.

Έστω επίσης ότι κατά την $t_{(j)}$, πεθαίνουν d_j άτομα ενώ πριν από την $t_{(j)}$ ζούσαν n_j άτομα. Συνήθως $d_j = 1, \forall j$.

Είναι σημαντικό να καταλάβουμε ότι ο αριθμός n_j περιλαμβάνει όλα τα άτομα που ζούσαν εκείνη τη στιγμή, ανεξαρτήτως αν μετέπειτα θα πεθάνουν ή θα συνεχίζουν να ζουν μετά το τέλος του πειράματος. Δεν περιλαμβάνει τα νεκρά άτομα, ούτε τα άτομα με αποκομμένες τιμές πριν την στιγμή $t_{(j)}$.

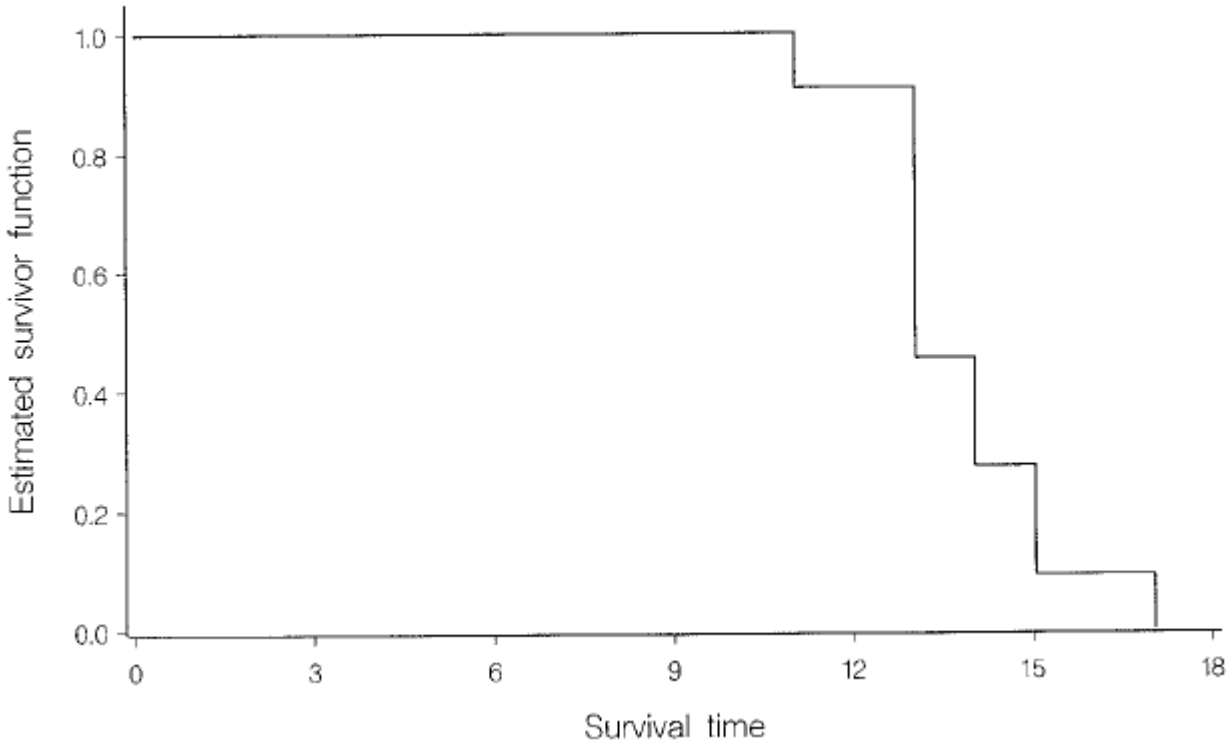
Ένα παράδειγμα για τα παραπάνω (Collett, 2003):

Παράδειγμα 2.1. *Μια επιπλοκή στους ασθενείς με οστεοσάρκωμα είναι ότι ο καρκίνος συχνά κάνει μετάσταση στους πνεύμονες. Αυτή η μετάσταση συνήθως είναι επικίνδυνη για την ζωή του ασθενή. Σε μια μελέτη από τους Burdette, Gehan (1970) έδωσε τους εξής χρόνους επιβίωσης σε μήνες για 11 άντρες ασθενείς.*

11 13 13 13 13 13 14 14 15 15 17

Χρησιμοποιώντας την 2.1 παίρνουμε τις εξής εκτιμώμενες τιμές για την συνάρτηση επιβίωσης:1.000, 0.909, 0.455, 0.273, 0.091. Μια γραφική παράσταση για την εκτιμώμενη συνάρτηση επιβίωσης δίνεται στο σχήμα 2.1

Σχήμα 2.1: *Εκτιμώμενη συνάρτηση επιβίωσης για τα δεδομένα του παραδείγματος 2.1*



Υπολογίζουμε την συνάρτηση επιβίωσης

$$S(t_{(j)}) = P(T > t_{(j)}) \tag{2.2}$$

και από τύπο πιθανότητας $P(A \cap B) = P(A)P(B|A)$ η 2.2 $\Rightarrow$

$$S(t_{(j)}) = P(T > t_{(1)})P(T > t_{(2)}|T > t_{(1)})...P(T > t_{(j)}|T > t_{(j-1)}) \tag{2.3}$$

Χρησιμοποιώντας έπειτα μια απλή εκτιμήτρια για τη P.

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

ομοίως

$$\hat{P}(T > t_{(2)}|T > t_{(1)}) = 1 - p_2 = 1 - \frac{d_2}{n_2} = \frac{n_2 - d_2}{n_2}$$

όπου p_i είναι η σχετική συχνότητα στο $(t_{(i-1)}, t_{(i)}]$ και τελικά η εκτιμήτρια Kaplan-Meier

$$\hat{S}(t) = \frac{n_1 - d_1}{n_1} \times \frac{n_2 - d_2}{n_2} \times ... \times \frac{n_i - d_i}{n_i}, i : t_{(i)} \leq t < t_{(i+1)}$$

$$= \begin{cases} \prod_{j:t_{(j)} \leq t} \frac{n_j - d_j}{n_j} & t \geq t_{(1)} \\ 1 & t < t_{(1)} \end{cases} \tag{2.4}$$

Η εξίσωση 2.4 είναι η εκτιμήτρια Kaplan-Meier , παρατηρούμε ότι αν δεν υπάρχουν αποκομμένες παρατηρήσεις από τον τύπο 2.4 ξαναγυρνάμε στον εμπειρικό τύπο 2.1. Παρατηρούμε επίσης ότι η εκτιμήτρια είναι και αυτή κλιμακωτή συνάρτηση. Επειδή η Kaplan-Meier αποτελεί εκτίμηση προφανώς θα χρειαστεί και τυπικό σφάλμα για να κατασκευαστούν διαστήματα εμπιστοσύνης. Ο τύπος του Greenwood (Καρώνη, 2009) δίνεται ως εξής:

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

Ένα ακόμη παράδειγμα (Collett, 2003):

Παράδειγμα 2.2. *Ο παγκόσμιος οργανισμός υγείας (WHO, 1987) έδωσε για έρευνα δεδομένα από κλινικές για την χρήση ενός αντισυλληπτικού μηχανισμού IUD γνωστό ως Multiload 250. Τα δεδομένα αφορούν 18 γυναίκες, όλες εκ των οποίων είχαν ηλικία 18-35. Στο παρακάτω πίνακα δίνεται ο χρόνος σε εβδομάδες από την πρώτη χρήση του αντισυλληπτικού μηχανισμού μέχρις ότου προκύψει κάποιο πρόβλημα αιμορραγίας:*

10	13*	18*	*19*	23*	*30*	*36*	38*	54*
56*	*59*	*75*	*93*	*97*	104*	*107*	107*	107*

Πίνακας 2.1: *Χρόνος σε εβομάδες χρήσης του IUD*

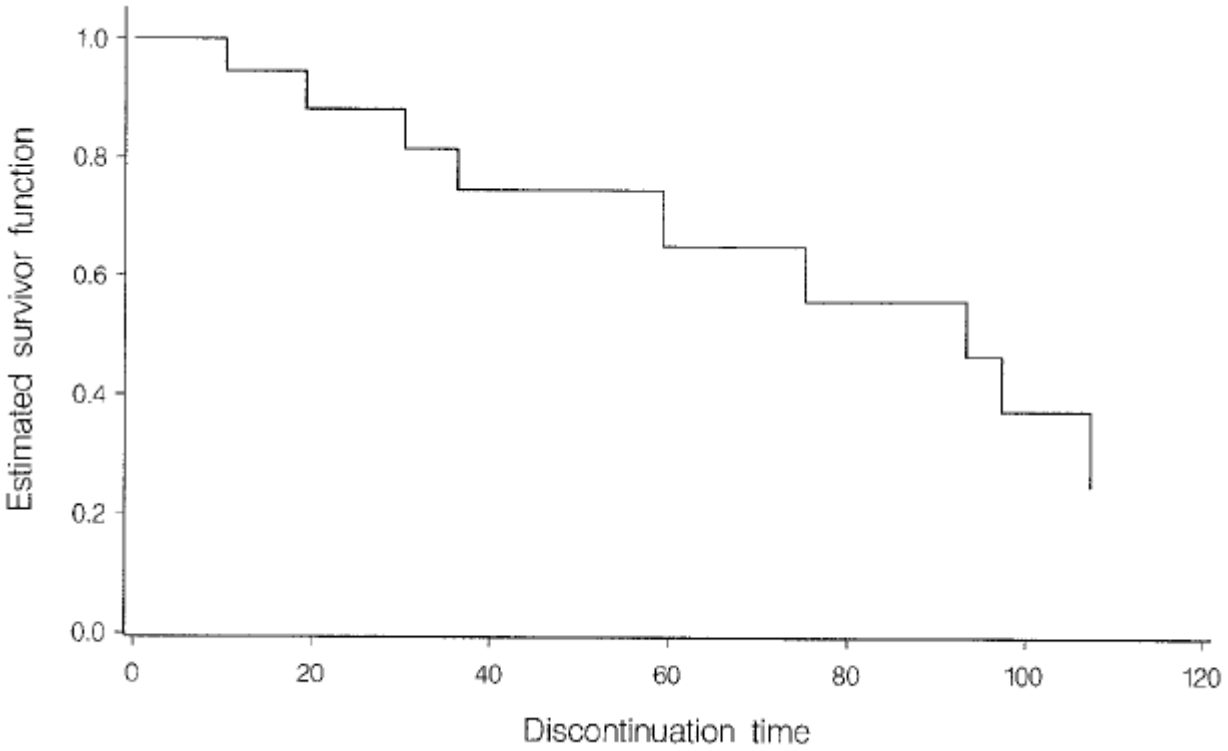
Στον παρακάτω πίνακα γίνεται χρήση του τύπου 2.4 για να υπολογίσουμε την εκτιμήτρια Kaplan-Meier:

Χρονικό Διάστημα	n_j	d_j	$(n_j - d_j)/n_j$	$\hat{S}(t)$
0-	*18*	*0*	*1.0000*	*1.0000*
10-	*18*	*1*	*0.9444*	*0.9444*
19-	*15*	*1*	*0.9333*	*0.8815*
30-	*13*	*1*	*0.9231*	*0.8137*
36-	*12*	*1*	*0.9167*	*0.7459*
59-	*8*	*1*	*0.8750*	*0.6526*
75-	*7*	*1*	*0.8751*	*0.5594*
93-	*6*	*1*	*0.8333*	*0.4662*
97-	*5*	*1*	*0.8000*	*0.3729*
107	*3*	*1*	*0.6667*	*0.2486*

Πίνακας 2.2: *Kaplan-Meier εκτίμηση του παραδείγματος 2.2*

Και καταλήγουμε στην γραφική παράσταση:

Σχήμα 2.2: *Γραφική παράσταση της Kaplan-Meier του παραδείγματος 2.2*



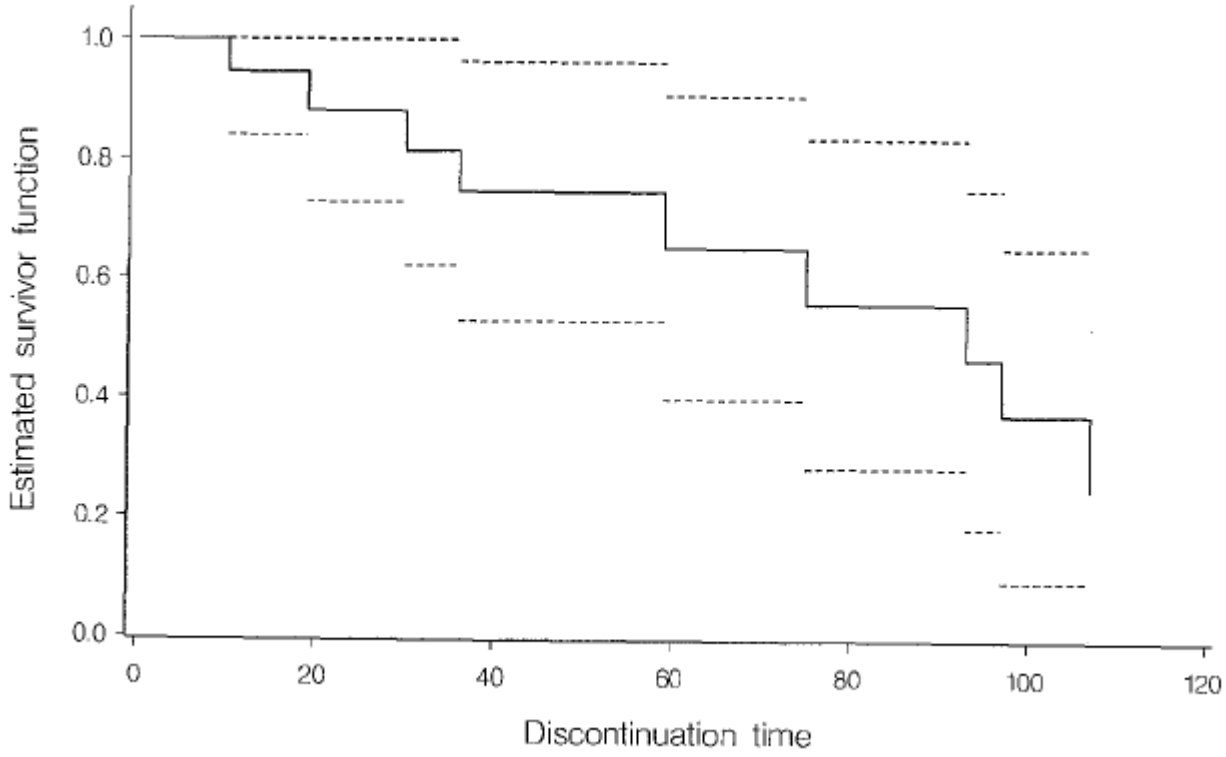
Χρησιμοποιώντας τώρα τον τύπο του Greenwood καταλήγουμε στον πίνακα :

Χρονικό Διάστημα	$\hat{S}(t)$	$se(\hat{S}(t))$	*95%Δ.Ε*
0-	*1.0000*	*0.0000*	
10-	*0.9444*	*0.0540*	*(0.839, 1.000)*
19-	*0.8815*	*0.0790*	*(0.727, 1.000)*
30-	*0.8137*	*0.09878*	*(0.622, 1.000)*
36-	*0.7459*	*0.1107*	*(0.529, 0.963)*
59-	*0.6526*	*0.1303*	*(0.397, 0.908)*
75-	*0.5594*	*0.1412*	*(0.283, 0.836)*
93-	*0.4662*	*0.1452*	*(0.182, 0.751)*
97-	*0.3729*	*0.1430*	*(0.093, 0.653)*
107	*0.2486*	*0.1392*	*(0.000, 0.522)*

Πίνακας 2.3: *Τυπικό σφάλμα του $\hat{S}(t)$ και διαστήματα εμπιστοσύνης για το S(t) του παραδείγματος 2.2*

Και η γραφική παράσταση με τα διαστήματα εμπιστοσύνης:

Σχήμα 2.3: *Γραφική παράσταση της Kaplan-Meier του παραδείγματος 2.2 με 95% Δ.Ε*



2.1.2 Έλεγχος Log-Rank

Ένας τρόπος να συγκρίνουμε τον χρόνο επιβίωσης σε δύο διαφορετικές ομάδες, και ίσως ο πιο απλός θα ήταν να κάνουμε την γραφική παράσταση των δύο εκτιμήσεων των συναρτήσεων επιβίωσης τους στον ίδιο άξονα. Ωστόσο υπάρχουν και άλλοι τρόποι που μας βοηθούν

στο να εξάγουμε τα σωστά συμπεράσματα.

Ένας τέτοιος, ο οποίος είναι πολύ χρήσιμος, είναι ο **log-rank** έλεγχος που είναι κατά βάση ένας μη-παραμετρικός έλεγχος υποθέσεων που χρησιμοποιούμε για τα δεδομένα μας. Είχε και άλλα ονόματα όπως Mantel and Haenszel (1959) ή αλλιώς Mantel-Cox ή Peto-Mantel-Haenszel αλλά κυρίως επικράτησε το όνομα **log-rank**.

Αρχικά ας δούμε πώς μπορούμε να τον κατασκευάσουμε τον παραπάνω έλεγχο. Έστω $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$ διακεκριμένες χρονικές στιγμές κατά τις οποίες κάποιοι άνθρωποι πεθαίνουν. Έστω επίσης ότι υπάρχουν δύο διαφορετικές ομάδες που προέρχονται αυτοί οι άνθρωποι. Θεωρούμε ότι στην ομάδα $i = 1$ ή $i = 2$ στην χρονική στιγμή $t_{(j)}$ υπάρχουν n_{ij} άνθρωποι σε κίνδυνο εκ των οποίων d_{ij} πεθαίνουν την χρονική στιγμή $t_{(j)}$. Ορίζουμε έπειτα ως

$$n_j = n_{1j} + n_{2j}$$

και

$$d_j = d_{1j} + d_{2j}.$$

Κατασκευάζουμε εν συνέχεια έναν πίνακα συνάφειας για να περιγράψουμε όλες τις πιθανές εκβάσεις:

	Ομάδα Α	Ομάδα Β	$\sum$
Πέθανε	d_{1j}	d_{2j}	d_j
Επέζησε	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
$\sum$	n_{1j}	n_{2j}	n_j

Κάνοντας τον γνωστό χ^2-έλεγχο για τον πίνακα συνάφειας υπολογίζουμε τις συχνότητες υπό την υπόθεση ανεξαρτησίας του θανάτου από την ομάδα στην οποία βρίσκεται. Οι συναρτήσεις επιβίωσης είναι ίδιες. Αυτός είναι και ο H_0 έλεγχος της υπόθεσης. Για την ομάδα Α η συχνότητα είναι:

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

και η απόκλιση :

$$u_j = d_{1j} - \frac{n_{1j}d_j}{n_j}$$

Για την διασπορά της d_{1j} :

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

Με την διαίρεση τώρα του τετράγωνο της ποσότητας u_j με την διασπορά της d_{1j} προκύπτει έλεγχος υπόθεσης της ανεξαρτησίας.

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

Αθροίζοντας, εν τέλει ως πρός όλες τις χρονικές στιγμές έχουμε:

$$u = \sum_j u_j = \sum_j \{d_{1j} - (n_{1j}d_j/n_j)\} \tag{2.5}$$

και

$$v = \sum_j v_j = \sum_j n_{1j}n_{2j}d_j(n_j - d_j)/n_j^2(n_j - 1) \tag{2.6}$$

Η $u/\sqrt{v} \sim N(0,1)$ και κατά συνέπεια η log-rank ελεγχοσυνάρτηση $u^2/v \sim \chi_1^2$ τα οποία αποδεικνύονται.

Ο παραπάνω έλεγχος είναι πάρα πολύ σημαντικός γιατί η μηδενική υπόθεση H_0 είναι η $S_1(t) = S_2(t)$ δίχως όμως να χρειάζεται να έχουν προσδιοριστεί αυτές οι δύο συναρτήσεις επιβίωσης. Παρατηρούμε τέλος, ότι ο έλεγχος log-rank μπορεί να επεκταθεί και σε άλλες ελεγχοσυναρτήσεις, αν παρατηρήσει κανείς ότι η log-rank ελεγχοσυνάρτηση είναι η:

$$\left(\sum w_j u_j\right)^2 / \sum w_j^2 v_j$$

με $w_j = 1$ τότε μπορεί να αντιληφθεί ότι με διαφορετική επιλογή του w_j αλλάζει και η ελεγχοσυνάρτηση η πιο γνωστή εκτός της log-rank είναι η **Wilcoxon** η οποία προκύπτει με την επιλογή $w_j = n_j$ (Καρώνη, 2009).

Παράδειγμα 2.3. *Το παράδειγμα αυτό αφορά του καρκίνο του μαστού στις γυναίκες που έχουν χωριστεί σε δύο ομάδες σύμφωνα με το έαν ένα μέρος του όγκου είχε θεωρηθεί θετικό η αρνητικό μέσω μιας ειδικής χρώσης (HPA stain).*

Χρόνος θανάτου	d_{1j}	n_{1j}	d_{2j}	n_{2j}	d_j	n_j	u_j	v_j
5	0	13	1	32	1	45	0.2889	0.2054
8	0	13	1	31	1	44	0.2955	0.2082
10	0	13	1	30	1	43	0.3023	0.2109
13	0	13	1	29	1	42	0.3095	0.2137
18	0	13	1	28	1	41	0.3171	0.2165
23	1	13	0	27	1	40	0.3250	0.2194
24	0	12	1	27	1	39	0.3077	0.2130
26	0	12	2	26	2	38	0.6316	0.4205
31	0	12	1	24	1	36	0.3333	0.2222
35	0	12	1	23	1	35	0.3429	0.2253
40	0	12	1	22	1	34	0.3529	0.2284
41	0	12	1	21	1	33	0.3636	0.2314
47	1	12	0	20	1	32	0.3750	0.2344
48	0	11	1	20	1	31	0.3548	0.2289
50	0	11	1	19	1	30	0.3667	0.2322
59	0	11	1	18	1	29	0.3793	0.2354
61	0	11	1	17	1	28	0.3929	0.2385
68	0	11	1	16	1	27	0.4074	0.2414
69	1	11	0	15	1	26	0.4231	0.2441
71	0	9	1	15	1	24	0.3750	0.2344
113	0	6	1	10	1	16	0.3750	0.2344
118	0	6	1	8	1	14	0.4286	0.2449
143	0	6	1	7	1	13	0.4615	0.2485
148	1	6	0	6	1	12	0.5000	0.2500
181	1	5	0	4	1	9	0.5556	0.2469
Σύνολο	5						9.5652	5.9289

Πίνακας 2.4: *Log-rank έλεγχος του παραδείγματος 2.3*

Άρα η τιμή της ελεγχοσυνάρτησης log-rank είναι 3.515 το οποίο έχει p-value: $P = 0.061$ που είναι αρκετά μικρό για να απορρίψουμε την μηδενική υπόθεση ότι οι συναρτήσεις επιβίωσης είναι ίδιες. Άρα έχει σημασία σε ποια ομάδα βρισκόμαστε.

2.2 Το Μοντέλο Αναλογικής Διακινδύνευσης του Cox

2.2.1 Ορισμός

Πολύ σημαντικό στην ανάλυση επιβίωσης είναι να δούμε σε ποιο μοντέλο προσαρμόζονται καλύτερα τα δεδομένα μας. Τα παραμετρικά μοντέλα χρησιμοποιούνται κυρίως σε ανάλυση τεχνολογικών δεδομένων, αντιθέτως σε ότι αφορά τον άνθρωπο επειδή οι συμμεταβλητές δεν είναι πάντα γνωστές και κάθε πληθυσμός είναι διαφορετικός από τον άλλον τα παραμετρικά μοντέλα, συνήθως, δεν ανταποκρίνονται στις ανάγκες. Ένα ευρέως γνωστό και χρησιμο-

ποιημένο μοντέλο κυρίως στην βιοιατρική αλλά και σε άλλες επιστήμες είναι το μοντέλο αναλογικής διακινδύνευσης του Cox.

Ο Sir David Cox παρατήρησε ότι εάν η υπόθεση αναλογικής διακινδύνευσης ισχύει τότε μπορούμε να εκτιμήσουμε την επίδραση των παραμέτρων χωρίς κάποια επιβάρυνση στην συνάρτηση διακινδύνευσης. Το μοντέλο παλινδρόμησης του Cox (1972) είναι ένα ημί-παραμετρικό μοντέλο παλινδρόμησης αναλογικής διακινδύνεσης. Αυτό σημαίνει ότι οι συμμεταβλητές επηρεάζουν την συνάρτηση διακινδύνευσης ως εξής:

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

όπου $h_0(t)$ είναι μια βασική συνάρτηση διακινδύνευσης και β είναι ένα διάνυσμα συντελεστών που εκφράζουν την επίδραση των συμμεταβλητών. Γνωρίζοντας την:

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

προκύπτει

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

και από την σχέση $S(t) = exp\{-H(t)\}$

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

Το σημαντικότερο πλεονέκτημα στο μοντέλο του Cox είναι ότι δεν καθορίζονται η βασική συνάρτηση διακινδύνευσης και η βασική συνάρτηση επιβίωσης. Αναλύεται η επίδραση όμως των συμμεταβλητών, εξού και ο προσδιορισμός ημι-παραμετρικό μοντέλο, το οποίο σε πολλές περιπτώσεις είναι πολύ σημαντικό.

2.2.2 Έλεγχοι της υπόθεσης αναλογικής διακινδύνευσης στο μοντέλο του Cox

Το μοντέλο αναλογικής διακινδύνευσης όπως είδαμε και πριν θεωρεί ότι ο λόγος μεταξύ των συναρτήσεων διακινδύνευσης δύο μονάδων είναι ανεξάρτητος του χρόνου δηλαδή για δύο μονάδες έστω i και j:

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

Υπάρχουν κάποιοι τρόποι να ελέγξουμε αυτήν την ιδιότητα. Ένας εκ των οποίων είναι ο εξής:

1. Φτιάχνουμε μια καινούρια μεταβλητή που εξαρτάται από τον χρόνο $z = x_i t$

2. Προσαρμόζουμε το μοντέλο του Cox ξανά συμπεριλαμβάνοντας αυτήν την φορά την νέα συμμεταβλητή z

3. Ελέγχουμε την μηδενική υπόθεση $H_0 : \beta_z = 0$

Τα παραπάνω τα κάνουμε για κάθε συμμεταβλητή στο μοντέλο μας. Η αποδοχή της H_0 μας δείχνει ότι η επίδραση της συμμεταβλητής x_i μέσω του ανεξάρτητου από τον χρόνο όρο $\beta_i x_i$ εκφράζεται επαρκώς.

Ένας ακόμη τρόπος για τον ίδιο έλεγχο προκύπτει από τη γραφική παράσταση της συνάρτησης $ln\{-ln\hat{S}\}$ έναντι του τ. $\hat{S}$ είναι η εκτιμήτρια Kaplan-Meier, ωστόσο στην περίπτωση του μοντέλου του Cox επειδή η Kaplan-Meier δεν συνυπολογίζει τις τιμές των άλλων συμμεταβλητών χρειαζόμαστε μια καλύτερη εκτίμηση καθώς αν χρησιμοποιούσαμε την Kaplan-Meier θα χρειαζόμασταν και την ακαθόριστη συνάρτηση $S_0(t)$. Για να το καταφέρουμε αυτό χρησιμοποιείται η μη-παραμετρική εκτιμήτρια του Breslow (1974) :

$$\hat{S}_0(t) = e^{-\hat{H}_0(t)}$$

με

$$\hat{H}_0(t) = \sum_{t_{(j)} \le t} \left(\frac{d_j}{\sum_{i \in R_j} e^{\hat{\beta}' x_i}} \right)$$

Άρα ένας γραφικός έλεγχος γίνεται τώρα με την κατά στρώματα εκτιμήτρια της συνάρτησης επιβίωσης.

2.2.3 Υπόλοιπα στο μοντέλο του Cox

Ένας βασικός τρόπος για να ελέγξουμε την καταλληλότητα ενός στατιστικού μοντέλου είναι να εξετάσουμε τα υπόλοιπα μετά την προσαρμογή του μοντέλου. Τα υπόλοιπα θα μας δείξουν κατά πόσο οι προϋποθέσεις αλλά και οι προβλέψεις του μοντέλου συμφωνούν με τα δεδομένα μας. Δεν αποτελεί, προφανώς, εξαίρεση και το ημι-παραμετρικό μοντέλο του Cox. Τα γενικευμένα υπόλοιπα του Cox & Snell (1968) είναι για την περίπτωση του Cox μοντέλου:

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

Τα συγκεκριμένα υπόλοιπα παρά την ευρεία τους χρήση κυρίως στα παραμετρικά μοντέλα, δεν βοηθάνε πολύ στο μοντέλο του Cox καθώς υπάρχει η δυσκολία λόγω της εκτίμησης (μη-παραμετρικής) της $\hat{H}_0(t)$, για αυτό ένα άλλο είδος υπολοίπων τα υπόλοιπα Schoenfeld ή αλλιώς «μερικά υπόλοιπα» αναπτύχθηκαν από τον Schoenfeld το 1982.

Παρακάτω θα δούμε πως κατασκευάζουμε τα υπόλοιπα Schoenfeld. Έστω:

$$p_j = \frac{e^{\beta' x_j}}{\sum_{i \in R_j} e^{\beta' x_i}}$$

είναι η πιθανότητα που το άτομο j πεθαίνει δεδομένου ότι πεθαίνει κάποιος την στιγμή $t_{(j)}$ όταν υπάρχει R_j άτομα σε κίνδυνο πριν από εκείνη τη στιγμή. Δεν ξέρουμε ποιο άτομο

θα πεθάνει από το σύνολο μας, άρα η τιμή των συμμεταβλητών x είναι μια τυχαία μεταβλητή με:

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

Τώρα με τον κλασσικό τρόπο της απόκλισης της παρατήρησης από την αναμενόμενη τιμή παίρνουμε τα υπόλοιπα:

$$r_j = x_j - E(x|R_j)$$

Τέλος, αν αντικαταστήσουμε τα β με τα $\hat{\beta}$ παίρνουμε τα υπόλοιπα Schoenfeld

$$\hat{r}_j = x_j - \hat{E}(x|R_j) \tag{2.7}$$

Κάποια ακόμα υπόλοιπα που θα μας απασχολήσουν στην παρούσα διπλωματική είναι τα martingale υπόλοιπα. Δίνονται από τον τύπο :

$$r_{M_i} = \delta_i - r_{C_i} \tag{2.8}$$

όπου r_{C_i} είναι τα υπόλοιπα Cox-snell που δίνονται από τον τύπο

$$r_{C_i} = exp(\hat{\beta}' x_i)\hat{H}_0(t_i) \tag{2.9}$$

και δ_i μια δείκτρια συνάρτηση που παίρνει την τιμή 0 όταν η παρατήρηση είναι αποκομμένη και 1 όταν δεν είναι αποκομμένη. Τα υπόλοιπα martingale παίρνουν τιμές από το -∞ εώς την μονάδα. Αποδεικνύεται επίσης ότι αθροίζονται στο 0. Τα υπόλοιπα Martingale μπορούν να χρησιμοποιηθούν για να εκτιμήσουμε την αληθινή μορφή της συνάρτησης μιας συμμεταβλητής (Thernau et al., 1990).

Τα υπόλοιπα Martingale είναι πολύ χρήσιμα και χρησιμεύουν και πέρα από τους συνήθεις σκοπούς όπως στο να αναγνωρίσουμε μια ακραία τιμή (outlier). Ωστόσο το μεγαλύτερο μειονέκτημα προκύπτει από την ασυμμετρία που υπάρχει καθώς όπως αναφέραμε πριν το άνω φράγμα είναι η μονάδα αλλά δεν έχει κάτω φράγμα.

2.3 Μέθοδοι Συρρίκνωσης

2.3.1 Παλινδρόμηση Κορυφογραμμής

Μια τεχνική που χρησιμοποιείται ευρέως σε πολλά μοντέλα παλινδρόμησης και θα χρησιμοποιήσουμε και στην παρούσα διπλωματική εργασία είναι η παλινδρόμηση κορυφογραμμής (τεχνική Ridge). Συνήθως σε ένα σύνολο δεδομένων παρουσιάζεται υψηλή συσχέτιση μεταξύ των επεξηγηματικών μεταβλητών και έτσι εμφανίζεται πολυσυγγραμικότητα. Η ύπαρξη πολυσυγγραμμικότητας οδηγεί σε υψηλά τυπικά σφάλματα για τις εκτιμήτριες ελαχίστων τετραγώνων με αποτέλεσμα να μην είναι εύκολο να βρεθούν ποιες μεταβλητές θα είναι στατιστικά σημαντικές. Το να γνωρίζουμε ποιες μεταβλητές είναι στατιστικά σημαντικές είναι ένα «όπλο» που μας βοηθάει πολύ στην στατιστική μας ανάλυση.

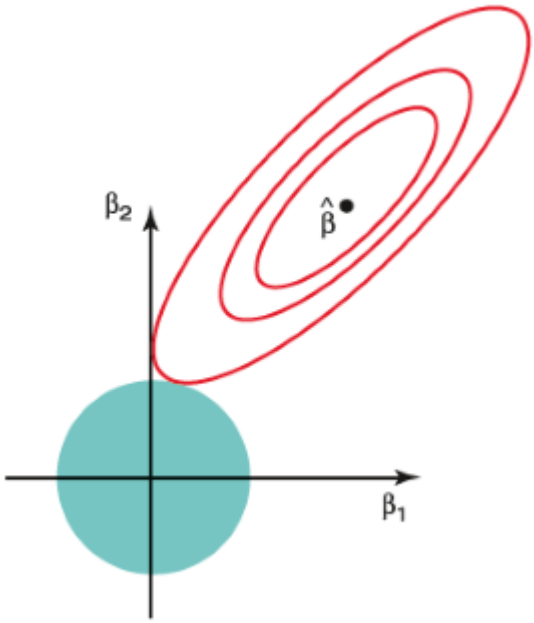
Ας δούμε τώρα τι ακριβώς είναι η τεχνική Ridge. Η τεχνική Ridge (Hoerl & Kennard, 1970), είναι μια μέδοθος συρρίκνωσης κάποιων συντελεστών παλινδρόμησης. Η τεχνική συρρίνκνωσης Ridge είναι αρκετά παρόμοια με την μέδοθο ελαχίστων τετραγώνων. Η ειδοποιός διαφορά τους είναι η ποσότητα που ελαχιστοποιείται. Οι εκτιμήτριες της παλινδρόμησης Ridge ($\hat{\beta}^R$) υπολογίζονται ελαχιστοποιώντας την ποσότητα:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2 \tag{2.10}$$

όπου λ≥ 0 είναι μια ρυθμιστική παράμετρος (Tuning Parameter). Όπως και στην μέθοδο ελαχίστων τετραγώνων θέλουμε να γίνει το RSS μικρό. Ο δεύτερος όρος της εξίσωσης γνωστός ως και Shrinkage penalty είναι μικρός όταν τα $\beta_1, ..., \beta_p$ τείνουν στο 0. Το λ λέγεται ρυθμιστική παράμετρος και παρατηρούμε ότι όταν το λ=0 θα πάρουμε την εκτίμηση των ελαχίστων τετραγώνων. Όταν το λ τείνει στο άπειρο οι εκτιμήσεις των συμμεταβλητών θα πάνε στο 0. Παρατηρούμε επίσης ότι ανάλογα με το λ που θα διαλέξουμε παίρνουμε διαφορετική τιμή. Άρα η επιλογή της παραμέτρου λ είναι πολύ σημαντική όπως θα δούμε και παρακάτω. Παρατηρούμε τέλος, από την εξίσωση ότι η παράμετρος δεν επιδρά στο β_0 καθώς δεν θέλουμε να συρρικνωθεί.

Στο σχήμα 2.4 παρουσιάζουμε γεωμετρικά την επίδραση που έχει ο περιορισμός $\sum_{k=1}^{p}\beta_k^2 \leq t$,όπου t εξαρτάται από λ, στην ελαχιστοποίηση της παράστασης (RSS):

Σχήμα 2.4: *Περιορισμός Ridge παλινδρόμησης σε 2 διαστάσεις*



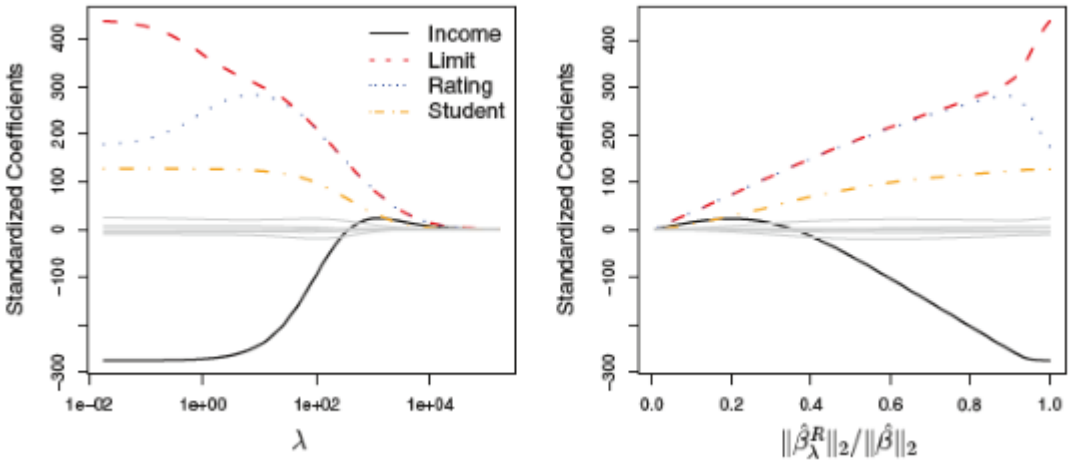
Ο περιορισμός απεικονίζεται από τον κύκλο ($\beta_1^2 + \beta_2^2 \leq t$), ενώ οι ελλείψεις του σχήματος αντιπροσωπεύουν τα σημεία στα οποία η RSS έχει την ίδια τιμή, η οποία προς το κέντρο των ελλείψεων παίρνει μικρότερη τιμή. Παρατηρούμε ότι η ελαχιστοποίηση επιτυγχάνεται για

κάποιο σημείο της κυκλικής περιοχής που ορίζει ο περιορισμός.

Όσο η τιμή της παραμέτρου λ αυξάνεται οι συντελεστές παλινδρόμησης συρρικνώνονται. Αυτό οδηγεί και στην μείωση της διασποράς των συντελεστών και για αυτό επιτυγχάνεται μεγαλύτερη ακρίβεια στο μοντέλο μας. Μια ποσότητα που θα εισάγουμε είναι η νόρμα. Η νόρμα l_2: $\|\beta\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$ μετράει την απόσταση του β από το 0. Όσο η ρυθμιστική παράμετρος λ αυξάνεται η νόρμα αυτή μειώνεται, επίσης μειώνεται και παίρνει και τιμές από το 1 εως το 0 η ποσότητα: $\left\|\hat{\beta}_{\lambda}^{R}\right\|_2 / \left\|\hat{\beta}\right\|_2$. Την τιμή 1 παίρνει όταν η ρυθμιστική παράμετρος λ=0 και την τιμή 0 όταν η ρυθμιστική παράμετρος λ τείνει στο άπειρο.

Τα παραπάνω φαίνονται στο σχήμα 2.5:

Σχήμα 2.5: *Γραφική παράσταση των ορθοκανονικοποιημένων μεταβλητών συναρτήσει του λ και της ποσότητας* $\left\|\hat{\beta}_{\lambda}^{R}\right\|_2 / \left\|\hat{\beta}\right\|_2$



Στο παραπάνω παράδειγμα (James et al., 2013) έχουν χρησιμοποιηθεί ορθοκανονικοποιημένες συμμεταβλητές για να είναι στην ίδια κλίμακα με τον τύπο :

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}} \tag{2.11}$$

Η τεχνική Ridge έχει ένα πλεονέκτημα σε σχέση με τα ελάχιστα τετράγωνα και έχει βάση στην «ανταλλαγή» μεροληψίας με διασπορά. Όσο αυξάνεται η ρυθμιστική παράμετρος λ μειώνεται η διασπορά και αυξάνεται η μεροληψία. Η παλινδρόμηση Ridge δουλεύει καλύτερα όταν οι εκτιμήσεις των ελαχίστων τετραγώνων έχουν πολύ μεγάλη διασπορά. Με την επιλογή κατάλληλου λ μπορεί να προκύψει μικρή αύξηση της μεροληψίας με μεγάλη μείωση της διασποράς πράγμα που μπορεί να αποδειχθεί ιδιαίτερα χρήσιμο.

2.3.2 Τεχνική Lasso

Μια παρόμοια τεχνική με την τεχνική Ridge είναι η τεχνική Lasso (Tibshirani, 1996). Η παλινδρόμηση κορυφογραμμής έχει ένα βασικό μειονέκτημα σε σχέση με άλλες τεχνικές που γενικά επιλέγουν μοντέλα που περιέχουν ένα **υποσύνολο** των μεταβλητών. Η παλιν-

δρόμηση Ridge θα έχει στο τέλος όλες τις μεταβλητές, μπορούν να μηδενιστούν μόνο όταν η ρυθμιστική παράμετρος λ τείνει στο άπειρο. Αυτό δεν είναι πρόβλημα για την πρόβλεψη αλλά είναι πρόβλημα για την ερμηνεία ειδικά όταν οι μεταβλητές του μοντέλου μας είναι πολλές.

Ένας τρόπος να αποφευχθεί αυτό το πρόβλημα είναι η τεχνική Lasso. Στην τεχνική Lasso έχουμε να ελαχιστοποιήσουμε μια διαφορετική ποσότητα από ότι αυτή της τεχνικής Ridge και υπερνικά το πρόβλημα που υπάρχει στην Ridge παλινδρόμηση, η ποσότητα αυτή είναι:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| = RSS+\lambda\sum_{j=1}^{p}|\beta_j| \tag{2.12}$$

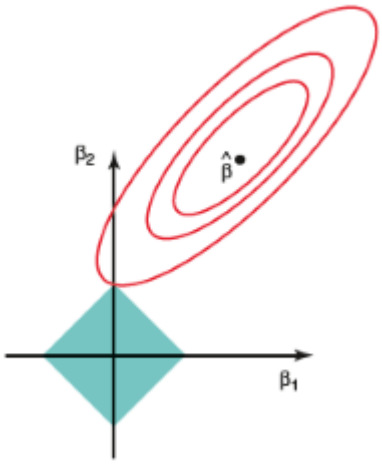
Βλέπουμε ότι η 2.10 με την 2.12 είναι παρόμοιες, η Lasso παρατηρούμε ότι χρησιμοποιεί μια διαφορετική νόρμα, την νόρμα l_1: $\|\beta\|_1 = \sqrt{\sum_{j=1}^{p}|\beta_j|}$

Όπως στην περίπτωση της Ridge παλινδρόμησης, η τεχνική Lasso συρρικνώνει τις εκτιμήσεις των συμμεταβλητών προς το 0. Ωστόσο, στην περίπτωση της Lasso, η ποινή αυτή την φορά μπορεί να κάνει κάποιες συμμεταβλητές **ακριβώς** 0. Άρα η ειδοποιός διαφορά είναι ότι η Lasso κάνει επιλογή μεταβλητών, άρα το μοντέλο που καταλήγουμε μπορεί να έχει λιγότερες μεταβλητές από το αρχικό μοντέλο πράγμα που κάνει την ερμηνεία του μοντέλου αρκετά πιο εύκολη.

Όπως και στην τεχνική Lasso καταλαβαίνουμε ότι η επιλογή της ρυθμιστικής παραμέτρου λ είναι βαρύνουσας σημασίας, καθώς επιλέγοντας κατάλληλο λ, θα φτάσουμε σε μοντέλο με λιγότερες μεταβλητές και καλύτερη ερμηνεία δίχως να έχουμε επηρεάσει πολύ την μεροληψία.

Στο σχήμα 2.6 παρουσιάζουμε γεωμετρικά την επίδραση που έχει ο περιορισμός $\sum_{k=1}^{p}|\beta_k| \leq t$,όπου t εξαρτάται από λ, στην ελαχιστοποίηση της παράστασης (RSS):

Σχήμα 2.6: *Περιορισμός Lasso παλινδρόμησης σε 2 διαστάσεις*

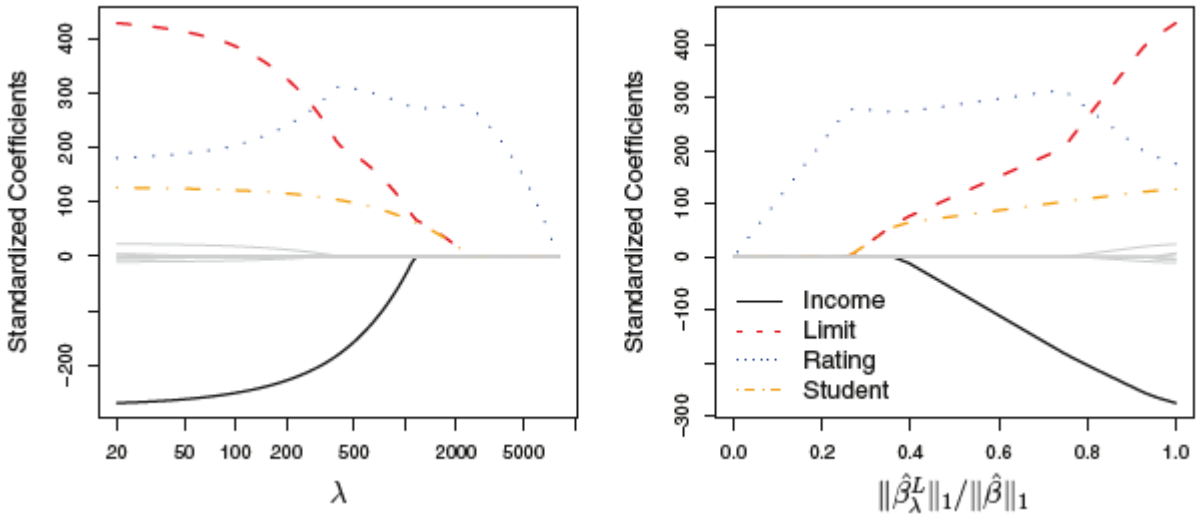


Ο περιορισμός απεικονίζεται από τον ρόμβο ($|\beta_1|+|\beta_2| \leq t$), ενώ οι ελλείψεις του σχήματος

αντιπροσωπεύουν τα σημεία στα οποία η RSS έχει την ίδια τιμή, η οποία προς το κέντρο των ελλείψεων παίρνει μικρότερη τιμή. Παρατηρούμε ότι η ελαχιστοποίηση επιτυγχάνεται σε κάποια κορυφή του ρόμβου που ορίζει ο περιορισμός.

Στο ίδιο παράδειγμα που χρησιμοποιήσαμε και στο προηγούμενο κεφάλαιο βλέπουμε και για την τεχνική Lasso:

Σχήμα 2.7: *Γραφική παράσταση των ορθοκανονικοποιημένων μεταβλητών συναρτήσει του λ και της ποσότητας* $\left\|\hat{\beta}_{\lambda}^{L}\right\|_1 / \left\|\hat{\beta}\right\|_1$



2.3.3 Σύγκριση τεχνικών Ridge, Lasso και επιλογή της παραμέτρου λ

Με βάση τα προηγούμενα κεφάλαια, παρατηρούμε ότι η τεχνική Lasso έχει ένα μεγάλο πλεονέκτημα σε σχέση με την παλινδρόμηση Ridge καθώς παράγει μοντέλα με μεταβλητές που είναι υποσύνολο των αρχικών μεταβλητών που έχει ως αποτέλεσμα την ευκολότερη ερμηνεία των μοντέλων. Παρ'όλα αυτά σημαντικό είναι να δούμε και ποια τεχνική έχει καλύτερη προβλεψιμότητα. Σε μοντέλα που όλες οι μεταβλητές παίζουν ρόλο (δηλαδή όλα τα β_j είναι διάφορα του μηδενός) η τεχνική Lasso, εμμέσως, μηδενίζει κάποιες μεταβλητές αυθαίρετα. Σε αυτές τις περιπτώσεις η τεχνική Ridge υπερτερεί της τεχνικής Lasso.

Σε περίπτωση που δεν παίζουν όλες οι μεταβλητές ρόλο αλλά μόνο κάποιες από το σύνολο, τότε η τεχνική Lasso υπερτερεί της τεχνικής Ridge. Αυτές οι δύο διαφορετικές περιπτώσεις μας δείχνουν ότι καμία από τις δύο τεχνικές δεν υπερτερεί από την άλλη. Γενικά, σε μοντέλα που ένας μικρός αριθμός μεταβλητών έχει σημαντικούς συντελεστές και οι υπόλοιπες έχουν συντελεστές πολύ μικρούς ή κοντά στο 0, υπερτερεί η Lasso τεχνική. Ενώ σε μοντέλα που σχεδόν όλες οι μεταβλητές έχουν σημαντικούς συντελεστές, υπερτερεί η τεχνική Ridge. Ωστόσο, σε πραγματικά δεδομένα δεν ξέρουμε ποτέ εκ των προτέρων ποιες μεταβλητές είναι σημαντικές άρα θα πρέπει να εξεταστούν και οι δύο τεχνικές.

Γενικότερα, η τεχνική Ridge συρρικνώνει κάθε συντελεστή κατά ίδια αναλογία που οδηγεί

σε καλύτερη πρόβλεψη στο μοντέλο ενώ η τεχνική Lasso συρρικνώνει όλους τους συντελεστές προς το 0 με ανάλογα ποσά και κάποιους μέχρι ακριβώς και 0 που οδηγεί σε μοντέλα με λιγότερες μεταβλητές και πιο εύκολη ερμηνεία.

Το να επιλέξεις την κατάλληλη ρυθμιστική παράμετρο λ είναι ένα δύσκολο πρόβλημα. Οι βέλτιστες ρυθμιστικές παράμετροι είναι «δύσκολο να ρυθμιστούν στην πράξη» (Lederer and Müller, 2015) και δεν είναι «πρακτικά εφικτές» (Fan and Tang, 2013). Πολλές τεχνικές έχουν εξεταστεί για παράδειγμα ο Tibshirani (2013) θεωρεί την cross validation τεχνική (που είναι αρκετά διαδεδομένη στην αναζήτηση των κατάλληλων ρυθμιστικών παραμέτρων) ένα εύκολο τρόπο για να εκτιμήσεις το λάθος της πρόβλεψης. Ωστόσο ο Chand (2012) θεωρεί ότι αυτή η τεχνική σχεδόν πάντα δεν κάνει σωστή επιλογή μεταβλητών. Η cross-validation τεχνική δουλεύει ως εξής, επιλέγουμε αρχικά, ένα πλέγμα από τιμές των λ, και υπολογίζουμε το σφάλμα του cross-validation για κάθε τιμή του λ. Επιλέγουμε έπειτα την τιμή της λ με το μικρότερο σφάλμα. Τέλος, ξαναφτιάχνουμε το μοντέλο χρησιμοποιώντας όλες τις διαθέσιμες παρατηρήσεις και την επιλεγμένη τιμή της ρυθμιστικής παραμέτρου. Οι Fan and Tang προτείνουν κάτι διαφορετικό:

1. Επιλέγουμε μια μέθοδο κανονικοποίησης όπως Ridge ή Lasso κλπ.
2. Χρησιμοποιούμε μια ακολουθία τιμών των ρυθμιστικών παραμέτρων για να φτιάξουμε μια σειρά από διαφορετικά μοντέλα.
3. Μελετάμε τα μοντέλα και επιλέγουμε το κατάλληλο με τα γνωστά κριτήρια όπως το AIC, BIC.

Ακούγεται απλή προσέγγιση αλλά το να επιλέξουμε μέθοδο και μετά να επιλέξουμε μοντέλο δεν δουλεύει σε περιπτώσεις που οι μεταβλητές p είναι πολλές και άρα αυξάνονται εκθετικά με το δείγμα. Σε αυτές τις περιπτώσεις οι Fang and Tang (2013) λένε ότι δεν υπάρχει τρόπος να πολεμήσουμε εύκολα αυτό το πρόβλημα. Παρατηρούμε δηλαδή ότι η επιλογή του λ είναι ένα δύσκολο πρόβλημα που απαιτεί αρκετά προσεχτική προσέγγιση και αντιμετώπιση.

2.4 Τα βασικά των δέντρων αποφάσεων

2.4.1 Δέντρα παλινδρόμησης

Πάρα πολύ χρήσιμη κατηγορία μεθόδων που χρησιμοποιείται πλέον ευρέως είναι οι Tree-based methods αυτές οι μέθοδοι είναι πολύ χρήσιμες όχι μόνο για την στατιστική ανάλυση αλλά και για την οπτικοποίηση των αποτελεσμάτων ώστε να μπορούν να τα καταλάβουν και άτομα που δεν έχουν σχέση με την στατιστική. Οι μέθοδοι αυτές είναι απλές και βοηθούν στην εύκολη ερμηνεία των αποτελεσμάτων. Χρησιμοποιούνται σε προβλήματα παλινδρόμησης αλλά και σε προβλήματα κατηγοριοποίησης.

Αρχικά, θα δούμε για τα προβλήματα παλινδρόμησης και το πως κατασκευάζουμε ένα δέντρο παλινδρόμησης. Η διαδικασία είναι η εξής:

1. Διαμερίζεται ο χώρος των επεξηγηματικών μεταβλητών $X_1, X_2, ...X_p$ σε Κ διακεκριμένους και μη επικαλυπτόμενους υποχώρους. $R_1, R_2, ..., R_k$
2. Για κάθε παρατήρηση που είναι στην ίδια περιοχή κάνουμε την ίδια πρόβλεψη, που είναι πρακτικά η μέση απόκριση σε εκείνη την περιοχή.

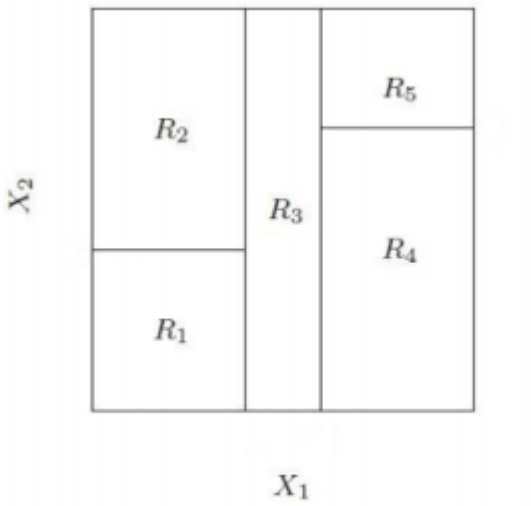
Εάν για παράδειγμα στην πρώτη περιοχή R_1 έχουμε μέση απόκριση 5, τότε για μια δοσμένη παρατήρηση $X = x$ με $x \in R_1$ προβλέπουμε ότι η τιμή θα είναι 5.

Όπως και στα προηγούμενα κεφάλαια, για την κατασκευή των Κ υποχώρων ελαχιστοποιούμε μια συνάρτηση κόστους. Αυτή η συνάρτηση κόστους είναι η RSS που δίνεται από τον τύπο:

$$RSS = \sum_{k=1}^{K} \sum_{i \in R_k} (y_i - \hat{y}_{R_k})^2 \tag{2.13}$$

όπου το $\hat{y}_{R_k}$ είναι η μέση απόκριση για τον υποχώρο R_k. Στο σχήμα 2.8 φαίνεται ένα παράδειγμα για δύο μεταβλητές που χωρίζονται σε 5 υποχώρους.

Σχήμα 2.8: *Διαμέριση χώρου για 2 μεταβλητές*



Επειδή είναι υπολογιστικά ανέφικτο να λάβουμε υπόψιν όλες τις πιθανές διαμερίσεις χρησιμοποιείται μια «άπληστη» όπως θα λέγαμε και στον προγραμματισμό μέθοδος, ξεκινάμε δηλαδή στην κορυφή (όπου όλες οι παρατηρήσεις ανήκουν σε ένα χωρίο) και έπειτα ο χώρος χωρίζεται σε δύο υποχώρους δημιουργώντας έτσι δύο νέα κλαδιά. Η μέθοδος θεωρείται άπληστη γιατί κοιτάμε μεμονωμένα σε κάθε βήμα ποια είναι η καλύτερη διαμέριση χωρίς να λαμβάνουμε υπόψιν ένα μελλοντικό βήμα.

Πιο συγκεκριμένα, η διαδικασία είναι η εξής: επιλέγουμε μια επεξηγηματική μεταβλητή X_j και ένα σημείο διαχωρισμού s ώστε να μειωθεί η RSS όσο περισσότερο γίνεται δημιουργώντας έτσι δύο υποχώρους $\{X|X_j < s\}$ και $\{X|X_j \geq s\}$. Δηλαδή εξετάζουμε όλες τις επεξηγηματικές μεταβλητές $X_1, X_2, ...X_p$ και όλες τις πιθανές τιμές του s για κάθε μεταβλητή και έπειτα επιλέγουμε την μεταβλητή και το s ώστε το δέντρο που προκύπτει να έχει την

χαμηλότερη RSS.

Ειδικότερα, για οποιοδήποτε j και s ορίζουμε τους χώρους:

$$R_1(j,s) = \{X|X_j < s\} \text{ και } R_2(j,s) = \{X|X_j \geq s\} \tag{2.14}$$

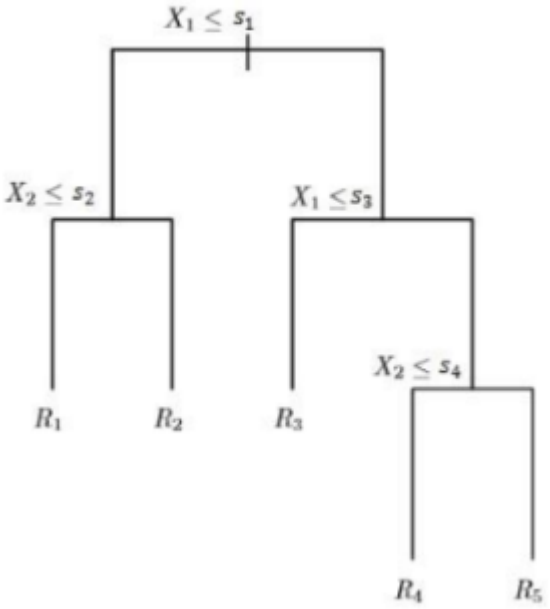
και ψάχνουμε τα j και s που ελαχιστοποιούν την συνάρτηση:

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 \tag{2.15}$$

οπου $\hat{y}_{R_1}$ και $\hat{y}_{R_2}$ είναι η μέση απόκριση στον χώρο $R_1(j,s)$ και $R_2(j,s)$ αντίστοιχα.

Έπειτα συνεχίζουμε την διαδικασία, μόνο που πλέον αντί να κάνουμε διαμέριση από όλο τον χώρο που είχαμε αρχικά διαλέγουμε ένα από τους δύο νέους υποχώρους δημιουργώντας έτσι τρεις υποχώρους. Κοιτάμε τώρα ποιον από τους τρεις να διαλέξουμε για να ελαχιστοποιήσουμε την RSS κ.ο.κ. μέχρι να φτάσουμε ένα κριτήριο διακοπής που θα μας λήξει τον αλγόριθμο. Στο σχήμα 2.9 βλέπουμε ένα παράδειγμα με δύο μεταβλητές και 5 υποχώρους:

Σχήμα 2.9: *Δέντρο παλινδρόμησης για 2 μεταβλητές*



Η παραπάνω διαδικασία ίσως παράγει καλές προβλέψεις αλλά μπορεί να υπάρξει υπερπροσαρμογή δηλαδή, αν προσαρμόσουμε το μοντέλο σε μια άλλη ομάδα δεδομένων, να μην έχει καλή απόδοση καθώς το δέντρο που προέκυψε μπορεί να είναι περίπλοκο.

Συχνά θέλουμε μικρότερα δέντρα. Τα μικρότερα δέντρα έχουν χαμηλότερη διασπορά και καλύτερη ερμηνεία αποτελεσμάτων με κόστος μια μικρή μεροληψία. Η διαδικασία συνήθως

που επιλέγεται για να καταλήξουμε σε μικρότερα δέντρα είναι να κατασκευάζουμε μεγάλα δέντρα και έπειτα να τα «κλαδεύουμε». Υπάρχουν αρκετοί αλγόριθμοι για αυτό που δεν θα αναλυθούν στην παρούσα εργασία αλλά το «κλάδεμα» των δέντρων είναι μια πολύ σημαντική διαδικασία για μια στατιστική ανάλυση. Παρακάτω θα δώσουμε έναν από αυτούς του αλγόριθμους.

Τα βήματα του αλγόριθμου είναι τα εξής:

1. Κάνουμε την διαδικασία που αναφέραμε προηγουμένως για να καταλήξουμε στο μεγάλο δέντρο σταματώντας όταν σε κάποιο τερματικό φύλλο έχει λιγότερες από κάποιο ελάχιστο αριθμό παρατηρήσεων.

2. Κάνουμε το «κλάδεμα» του δέντρου μέχρι να αποκτήσουμε μια ακολουθία από τα καλύτερα υποδέντρα σαν μια συνάρτηση του a.

3. Επιλέγουμε το α κάνοντας K-φορές cross-validation δηλαδή χωρίζουμε τις παρατηρήσεις σε K υποχώρους και για κάθε $k = 1, 2..., K$:

   (α΄) Κάνω τα βήματα 1 και 2 σε όλα εκτός από τον κ-οστό υποχώρο.

   (β΄) Υπολογίζω το MSPE (Mean Squared Prediction error) στον κ-οστό υποχώρο σαν συνάρτηση του a. Παίρνουμε τον μέσο όρο των αποτελεσμάτων για κάθε τιμή του a, και διαλέγουμε το a που ελαχιστοποιεί το μέσο σφάλμα.

4. Επιστρέφουμε το υποδέντρο του βήματος 2 για την επιλεγμένη τιμή του a.

Έτσι λοιπόν παίρνουμε ένα δέντρο παλινδρόμησης που έχει «κλαδευτεί» σε μικρότερο δέντρο Για κάθε τιμή του a αντιστοιχεί ένα υποδέντρο $T \subset T_0$ έτσι ώστε να ελαχιστοποιείται η:

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + a|T| \tag{2.16}$$

Το $|T|$ είναι το πλήθος των τερματικών φύλλων του δέντρου και R_m ο υποχώρος που αντιστοιχεί στο m-οστό τερματικό φύλλο. Το a όπως και στα προηγούμενα κεφάλαια είναι μια ρυθμιστική παράμετρος που όταν ισούται με το 0 τότε $T = T_0$ ενώ όσο αυξάνεται «κλαδεύονται» τα δέντρα και έτσι παίρνουμε την ακολουθία των υποδέντρων σαν συνάρτηση του a όπως αναλύθηκε και στον παραπάνω αλγόριθμο.

2.4.2 Δέντρα ταξινόμησης

Παρόμοια με τα δέντρα παλινδρόμησης είναι τα δέντρα ταξινόμησης μόνο που τώρα αντί για μια ποσοτική μεταβλητή εξετάζουμε μια ποιοτική μεταβλητή. Όπως είδαμε στα δέντρα παλινδρόμησης, χρησιμοποιούσαμε την μέση απόκριση των παρατηρήσεων που βρισκόταν στην ίδια περιοχή. Εδώ κάτι τέτοιο δεν είναι εφικτό. Στα δέντρα ταξινόμησης, προβλέπουμε ότι κάθε παρατήρηση ανήκει στην *πιο συχνά εμφανιζόμενη τάξη*.

Η κατασκευή του δέντρου ταξινόμησης είναι παρόμοια αυτής του δέντρο παλινδρόμησης. Χρησιμοποιούμε πάλι δηλαδή την άπληστη μέθοδο που αναφέραμε στο προηγούμενο κεφάλαιο. Εδώ υπάρχει μια διαφορά όμως δεν μπορούμε να χρησιμοποιήσουμε σαν κριτήριο το RSS. Η εναλλακτική που χρησιμοποιείται είναι το ποσοστό σφάλματος ταξινόμησης που δίνεται από τον τύπο :

$$E = 1 - \max_k(\hat{p}_{mk}) \tag{2.17}$$

όπου $\hat{p}_{mk}$ είναι το ποσοστό των παρατηρήσεων στην m-οστή περιοχή και ανήκει στην k-οστή τάξη. Ωστόσο υπάρχουν κάποιες ποσότητες που έχουν αποδειχθεί καλύτερες από την παραπάνω ποσότητα για την μέδοθο των δέντρων. Η μια είναι η Gini index και η άλλη ονομάζεται εντροπία. Οι τύποι για τις δύο αυτές ποσότητες δίνονται παρακάτω :

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{2.18}$$

και

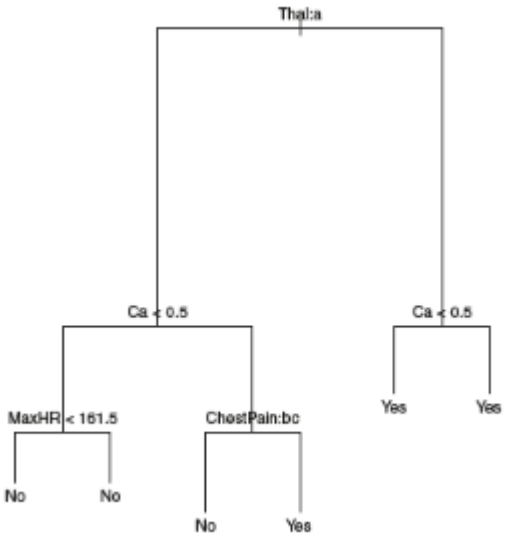
$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk} \tag{2.19}$$

Εφόσον $0 \leq \hat{p}_{mk} \leq 1 \rightarrow 0 \leq -\hat{p}_{mk} \log \hat{p}_{mk}$. Παρατηρούμε ότι η εντροπία μηδενίζεται όταν τα $\hat{p}_{mk}$ είναι κοντά στο 0 ή 1.

Όταν κατασκευάζουμε ένα δέντρο συνήθως χρησιμοποιούμε την 2.18 ή την 2.19. Όταν όμως «κλαδεύουμε» το δέντρο μπορούμε να χρησιμοποιήσουμε οποιαδήποτε από τις 3 αλλά η 2.17 είναι η καλύτερη αν μας νοιάζει η προβλεψιμότητα του τελικού δέντρου.

Στο σχήμα 2.10 βλέπουμε ένα δέντρο ταξινόμησης:

Σχήμα 2.10: *Δένδρο ταξινόμησης*



Στην παρούσα διπλωματική εργασία θα χρησιμοποιήσουμε ένα **δέντρο επιβίωσης** στο οποίο τα τερματικά φύλλα θα απεικονίζουν την συνάρτηση επιβίωσης και θα μας βοηθήσουν στην ερμηνεία των αποτελεσμάτων.

Υπάρχουν διάφορα κριτήρια για το πως φτιάχνουμε ένα τέτοιο δέντρο. Ο αναδρομικός διαχωρισμός για αποκομμένες μεταβλητές έχει τραβήξει μεγάλο ενδιαφέρον (π.χ. Segal, 1988, LeBlanc and Crowley, 1992). Δέντρα επιβίωσης που χρησιμοποιούνε το P-Value των log-rank ελέγχων χρησιμοποιούνται από τους Schumacher et al. (2012).

Το πακέτο του R που χρησιμοποιούμε στην παρούσα διπλωματική εργασία είναι το party package (Torsten Hothorn et al.) χρησιμοποιείται ο παρακάτω αναδρομικός αλγόριθμος:

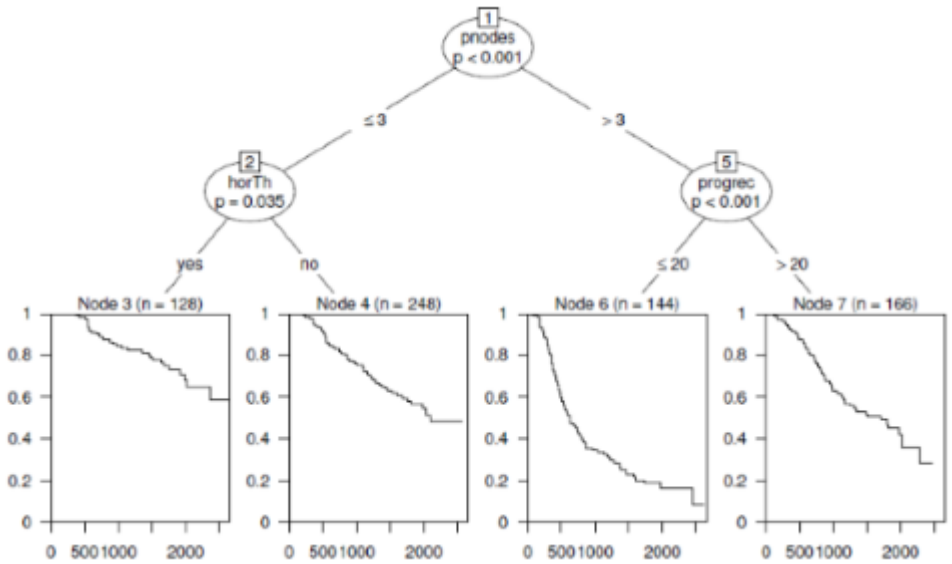
1. Έστω $w = (w_1, w_2, ..., w_n)$ μη αρνητικά βάρη. Για κάθε w έλεγξε την μηδενική υπόθεση ανεξαρτησίας μεταξύ των συμμεταβλητών m και της μεταβλητής απόκρισης. Σταμάτα εάν αυτή η υπόθεση δεν μπορεί να απορριφθεί. Αλλιώς διάλεξε την συμμεταβλητή X_{j^*} με την μεγαλύτερη συσχέτιση με την Y .

2. Επέλεξε ένα σύνολο $A^* \subset \chi_{j^*}$ ώστε να χωριστεί το χ_{j^*} σε δύο σύνολα A^* και $\chi_{j^*} \setminus A^*$. Τα βάρη w_{left}, w_{right} συνιστούν τις δυο υποομάδες με $w_{left,i} = w_i I(\chi_{j^* i \in A^*})$ και $w_{right,i} = w_i I(\chi_{j^* i \notin A^*})$ για κάθε $i = 1, 2..., n$. Το I είναι δείκτρια συνάρτηση.

3. Επανάλαβε τα βήματα 1 και 2 αναδρομικά με τα επεξεργασμένα w_{left}, w_{right} αντίστοιχα.

Ο διαχωρισμός της επιλογής των μεταβλητών και η διαδικασία διαχωρισμού στα 2 πρώτα βήματα του αλγορίθμου οδηγούν σε δέντρα που μπορούμε να ερμηνεύσουμε. Στο package που χρησιμοποιούμε για το δέντρο επιβίωσης το κριτήριο για τον διαχωρισμό των φύλλων είναι πάλι το log-rank. Στο σχήμα 2.11 βλέπουμε ένα παράδειγμα ενός survival-tree:

Σχήμα 2.11: *Παράδειγμα survival tree*



2.5 Καμπύλη ROC και AUC

2.5.1 Καμπύλη λειτουργικού χαρακτηριστικού δείκτη

Το 1971 ο Lusted εισήγαγε στην διαγνωστική ιατρική μία μέθοδο περιγραφής της ακρίβειας ενός ελέγχου, η οποία λαμβάνει υπόψιν όλα τα πιθανά σημεία απόφασης. Αυτός ο έλεγχος καλής προσαρμογής που χρησιμοποιείται συχνά κυρίως στη λογιστική παλινδρόμηση αλλά θα χρησιμοποιήσουμε και εμείς στην ανάλυση επιβίωσης είναι η **καμπύλη λειτουργικού χαρακτηριστικού δείκτη** (ROC-Receiver Operating Characteristic).

Η καμπύλη δημιουργείται με τη γραφική αναπαράσταση του πραγματικού θετικού ποσοστού (TPR-True Positive Rate) που είναι ο y-άξονας της γραφικής παράστασης, έναντι του ποσοστού ψευδών θετικών (FPR-False Positive Rate) που είναι ο x-άξονας σε διάφορα σημεία διαχωρισμού. Το πραγματικό θετικό ποσοστό ονομάζεται αλλιώς «ευαισθησία» (sensitivity), ενώ το ψευδώς θετικό ποσοστό είναι γνωστό ως «1-ειδικότητα», όπου ως «ειδικότητα» (specificity) ορίζεται το πραγματικά αρνητικό ποσοστό. Η σύνδεση των σημείων αυτών αποτελεί την εμπειρική καμπύλη **ROC**.

Η **ευαισθησία** ή αλλιώς **TPR** δίνεται από τον τύπο:

$$TPR = \frac{TP}{TP + FN} \tag{2.20}$$

όπου TP (True Possitive) είναι οι πραγματικά θετικές τιμές ενώ FN (False Negative) οι ψευδώς αρνητικές τιμές.

Η **1-ειδικότητα** ή αλλιώς **FPR** δίνεται από τον τύπο:

$$FPR = \frac{FP}{FP + TN} = 1 - TNR \tag{2.21}$$

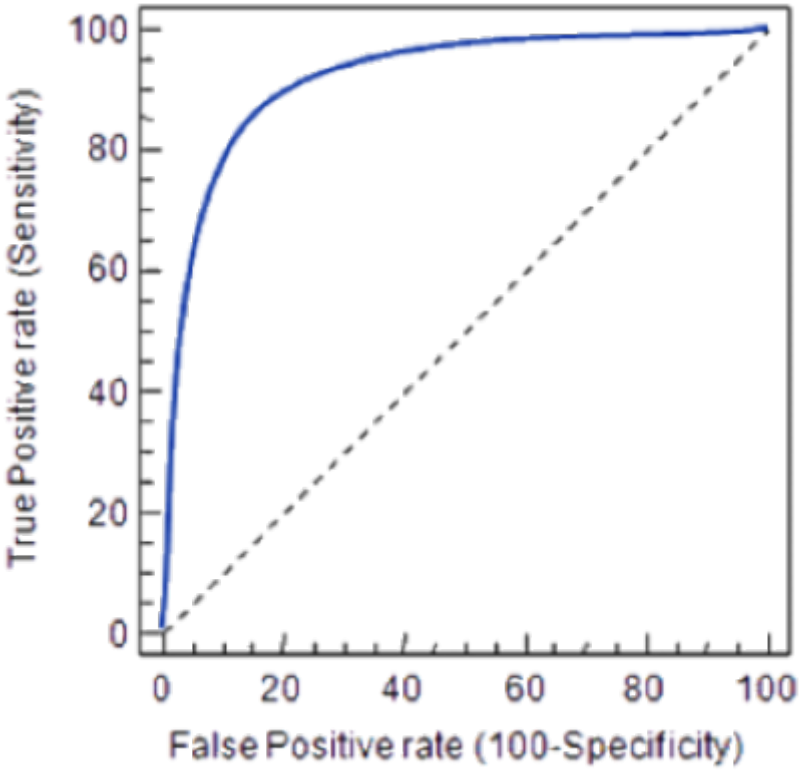
όπου FP (False Possitive) είναι οι ψευδώς θετικές τιμές ενώ TN (True Negative) οι πραγματικές αρνητικές τιμές και **TNR** ή αλλιώς **ειδικότητα** είναι:

$$TNR = \frac{TN}{TN + FP} \tag{2.22}$$

Άρα ο y-άξονας είναι η σχέση 2.20 και ο x-άξονας είναι η σχέση 2.21. Η καμπύλη ROC είναι σημαντική καθώς συμπεριλαμβάνει όλα τα ενδεχόμενα σημεία απόφασης και σε περιπτώσεις που συγκρίνουμε δύο ή περισσότερους ελέγχους μπορούμε άμεσα να συγκρίνουμε οπτικά τις δύο οι περισσότερες καμπύλες ROC.

Στο σχήμα 2.12 φαίνεται ένα παράδειγμα καμπύλης ROC.

Σχήμα 2.12: *Παράδειγμα καμπύλης ROC*



2.5.2 Η περιοχή κάτω από την καμπύλη ROC

Για να εξετάσουμε την ακρίβεια ενός ελέγχου υπάρχει ένα ακόμα μέτρο, το εμβαδόν της περιοχής κάτω από την καμπύλη ROC ή **AUC** που είναι συντομογραφία για το **Area under the ROC Curve** (Hanley and McNeil 1982). Λαμβάνει τιμές από 0 έως 1, και όσο μεγαλύτερη είναι η τιμή του εμβαδού κάτω από την καμπύλη, τόσο πιο ακριβής είναι ο διαγνωστικός μας έλεγχος. Η περιοχή κάτω από την καμπύλη ROC τυπικά υπολογίζεται προσθέτοντας διαδοχικές περιοχές τραπεζίων κάτω από την καμπύλη ROC. Αν η τιμή είναι κάτω από το 0.5 έχουμε ανακριβή έλεγχο άρα πρακτικά παίρνει τιμές από το [0.5–1.0]. Η περιοχή κάτω από την καμπύλη ROC μπορεί να χρησιμοποιηθεί σαν μια περίληψη της ικανότητας του μοντέλου μας.

Το εμβαδόν της περιοχής κάτω από την καμπύλη ROC αντιστοιχεί στη μέση τιμή της ευαισθησίας για όλες τις πιθανές τιμές ειδικότητας ή αντιστοιχεί στη μέση τιμή της ειδικότητας για όλες τις πιθανές τιμές ευαισθησίας.

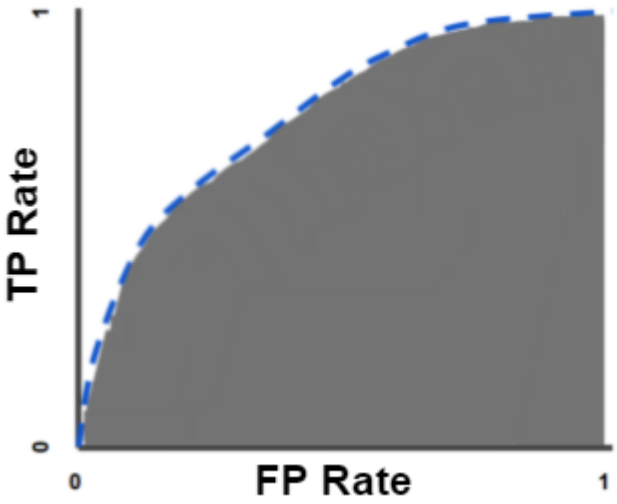
Σύμφωνα με την τιμή που παίρνει χωρίζονται στις εξής κατηγορίες (Hosmer and Lemeshow, 2013):

- 0.5= Καμία διάκριση
- 0.5-0.7= Φτωχή διάκριση
- 0.7-0.8= Αποδεκτή διάκριση
- 0.8-0.9= Τέλεια διάκριση
- >0.9 = Έξοχη διάκριση

Αυτές οι κατηγορίες δεν είναι απόλυτες αλλά χρησιμοποιούνται ως κατευθυντήριες γραμμές σε αρκετές στατιστικές αναλύσεις. Ανάλογα με το περιβάλλον της στατιστικής ανάλυσης ίσως να διαφέρουν λίγο.

Στο σχήμα 2.13 φαίνεται ένα παράδειγμα της περιοχής AUC.

Σχήμα 2.13: *Παράδειγμα περιοχής AUC*



Η σκιαγραφημένη περιοχή είναι η περιοχή κάτω από την καμπύλη ROC. Ανάλογα με την τιμή που παίρνει όπως αναφέραμε και πριν μπορούμε να εξάγουμε συμπεράσματα συνολικά για την ικανότητα του μοντέλου μας το οποίο είναι πάρα πολύ χρήσιμο.

Κεφάλαιο 3

Περιγραφή θέματος

Στο κεφάλαιο αυτό θα γίνει η περιγραφή του θέματος όσον αφορά το πρακτικό μέρος της εργασίας.

3.1 Σημασία και σκοπός της εργασίας

Εφόσον δόθηκε το θεωρητικό υπόβαθρο της παρούσας διπλωματικής εργασίας στα επόμενα κεφάλαια θα γίνει μια στατιστική ανάλυση χρησιμοποιώντας όλα τα προαναφερθέντα, δηλαδή αρκετές διαφορετικές τεχνικές στα ίδια δεδομένα.

Η ανάλυση επιβίωσης είναι πολύ σημαντική στις μέρες μας καθώς επεκτείνεται όλο και παραπάνω και έχει πλέον χρήση σε διάφορους κλάδους εκτός της βιοιατρικής. Σκοπός της εργασίας αυτής είναι με χρήση του προγραμματιστικού περιβάλλοντος R καθώς και με διάφορα στατιστικά πακέτα (libraries) να κάνουμε μια στατιστική ανάλυση. Η χρήση της R γίνεται καθώς είναι μια από τις πιο διαδεδομένες γλώσσες προγραμματισμού στην ανάλυση δεδομένων και είναι προσβάσιμη από όλους.

Συγκεκριμένα, αφού χρησιμοποιήσουμε κάποια δεδομένα θα διεξάγουμε μια στατιστική ανάλυση χρησιμοποιώντας όλες τις τεχνικές που αναλύθηκαν στα προηγούμενα κεφάλαια. Αυτές οι τεχνικές είναι διαδεδομένες και χρησιμοποιούνται ευρέως στην ανάλυση δεδομένων (data science). Σκοπός της εργασίας είναι να εξάγουμε συμπεράσματα αφού χρησιμοποιήσουμε όλες τις διαφορετικές τεχνικές και να δούμε πόσο σημαντική είναι να χρησιμοποιούμε ένα συνονθύλευμα από όλες τις τεχνικές.

Πιο συγκεκριμένα, θέλουμε να δούμε σε δεδομένα επιβίωσης που χρησιμοποιείται ευρέως το μοντέλο του Cox εάν οι τεχνικές Lasso, Ridge καθώς και τα δέντρα επιβίωσης (Survival Trees) θα μας βοηθήσουν στην εξαγωγή των συμπερασμάτων.

Εν κατακλείδι, θα καταλήξουμε σε ένα τελικό μοντέλο παλινδρόμησης για τα δεδομένα μας το οποίο θα είναι το καταλληλότερο και θα δούμε ποιες μεταβλητές τελικά ήταν χρήσιμες στο μοντέλο μας.

Μέρος II

Πρακτικό Μέρος

Κεφάλαιο 4

Δεδομένα πειράματος

Στο κεφάλαιο αυτό παρουσιάζεται το πείραμα που θα εκτελέσουμε καθώς και ποια είναι τα δεδομένα του πειράματος.

4.1 Περιγραφή δεδομένων του πειράματος

Στο πείραμα μας εξετάζουμε εάν υπήρξε υποτροπή ή θάνατος σε διάφορους τύπους λευχαιμίας που δέχθηκαν μόσχευμα μυελού των οστών. Παρακάτω θα αναφέρουμε και θα αναλύσουμε όλες τις μεταβλητές του μοντέλου μας. Οι μεταβλητές είναι οι εξής:

- **time**: χρόνος σε μέρες μέχρι την υποτροπή ή το θάνατο ενός ασθενούς μετά από μεταμόσχευση μυελού των οστών
- **indicator**: 1: έχει συμβεί το γεγονός, 0: αλλιώς
- **group**: τύπος λευχαιμίας (1=ALL, 2=AML low-risk, 3=AML high-risk)
- **recipient age**: ηλικία του ασθενή
- **donor age**: ηλικία του δότη
- **recipient sex**: φύλο του ασθενή (1=άντρας, 0=γυναίκα)
- **donor sex**: φύλο του δότη (1=άντρας, 0=γυναίκα)
- **recipient cmv**: κατάσταση του cmv του ασθενή (1=θετικό cmv, 0=αρνητικό cmv)
- **donor cmv**: κατάσταση του cmv του δότη (1=θετικό cmv, 0=αρνητικό cmv)
- **waiting time**: χρόνος αναμονής σε μέρες από τη διάγνωση μέχρι τη μεταμόσχευση
- **fab**: 1=fab βαθμού 4 ή 5, 0=αλλιώς
- **mtx**: 1=ναι, 0=όχι

Οι τύποι λευχαιμίας που μελετάμε είναι η οξεία λεμφοβλαστική λευχαιμία (ALL) και η οξεία μυελοκυτταρική λευχαιμία (AML low-risk και AML high-risk).

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

Ο Κυτταρομεγαλοϊός (CMV) είναι ένας κοινός ιός που ανήκει στην οικογένεια των ερπητοιών. Οι λήπτες μοσχεύματος είναι μια κατηγορία που διατρέχει υψηλό κίνδυνο για λοίμωξη από CMV επειδή πρέπει να παίρνουν φάρμακα που καταστέλλουν το ανοσοποιητικό σύστημά τους.

Χάριν συντομίας οι μεταβλητές έχουν μετονομαστεί σε :

1. time=t
2. indicator=c
3. recipient age=r.age
4. donor age=d.age
5. recipient sex=r.sex
6. donor sex=d.sex
7. recipient cmv=r.cmv
8. donor cmv=d.cmv
9. waiting time=w.time

Κεφάλαιο 5

Επεξεργασία δεδομένων

Στο κεφάλαιο αυτό γίνεται η βασική επεξεργασία των δεδομένων καθώς και η ανάλυση των δεδομένων. Η επεξεργασία και η ανάλυση των δεδομένων πραγματοποιήθηκε στο προγραμματιστικό περιβάλλον της R και πιο συγκεκριμένα χρησιμοποιήθηκε η RStudio.

5.1 Βασική ανάλυση δεδομένων

Στην ενότητα αυτή παρουσιάζεται η βασική ανάλυση δεδομένων. Πιο συγκεκριμένα, γίνονται οι Kaplan-Meier εκτιμήσεις, έλεγχοι αναλογικής διακινδύνευσης, ROC και AUC curves, εφαρμογή του μοντέλου αναλογικής διακινδύνευσης του Cox καθώς και τα υπόλοιπα Schoenfeld και Martingale.

Αρχικά, εγκαθιστούμε τα packages που χρειαζόμαστε και χρησιμοποιούμε την εντολή library() για να τα φορτώσουμε στο προγραμματιστικό περιβάλλον της R.

```
>library(survival)
>library(splines)
>library(parmsurvfit)
>library(party)
>library(risksetROC)
>library(glmnet)
>library(lattice)
```

Αυτά τα libraries είναι αναγκαία για την ανάλυση παρακάτω. Στην συνέχεια πρώτο βήμα είναι να περάσουμε τα δεδομένα μας στην R αυτό γίνεται με τις εξής εντολές:

```
> cc<-read.table("C:/bmt.txt",header=TRUE)
> attach(cc)
> cc
```

	id	group	t	c	r.age	d.age	r.sex	d.sex	r.cmv	d.cmv	w.time	fab	mtx
1	1	1	2081	0	26	33	1	0	1	1	98	0	0
2	2	1	1602	0	21	37	1	1	0	0	1720	0	0
3	3	1	1496	0	26	35	1	1	1	0	127	0	0
4	4	1	1462	0	17	21	0	1	0	0	168	0	0
5	5	1	1433	0	32	36	1	1	1	1	93	0	0
6	6	1	1377	0	22	31	1	1	1	1	2187	0	0

Έπειτα, χωρίζουμε τα δεδομένα μας στις 3 διαφορετικές ομάδες λευχαιμίας για να κάνουμε την βασική μας ανάλυση με τον παρακάτω κώδικα:

```
> ccdat1<-subset(cc,group==1)
> ccdat2<-subset(cc,group==2)
> ccdat3<-subset(cc,group==3)
```

και ενδεικτικά λίγα output μετά τον χωρισμό των ομάδων:

```
> ccdat1
   id group    t c r.age d.age r.sex d.sex r.cmv d.cmv w.time fab mtx
1   1     1 2081 0    26    33     1     0     1     1     98   0   0
2   2     1 1602 0    21    37     1     1     0     0   1720   0   0
3   3     1 1496 0    26    35     1     1     1     0    127   0   0
4   4     1 1462 0    17    21     0     1     0     0    168   0   0
5   5     1 1433 0    32    36     1     1     1     1     93   0   0
6   6     1 1377 0    22    31     1     1     1     1   2187   0   0
> ccdat2
   id group    t c r.age d.age r.sex d.sex r.cmv d.cmv w.time fab mtx
39 39     2 2569 0    19    13     1     1     1     0    270   1   0
40 40     2 2506 0    31    34     1     1     0     0     60   0   0
41 41     2 2409 0    35    31     1     1     1     1    120   0   0
42 42     2 2218 0    16    16     1     1     1     0     60   1   0
43 43     2 1857 0    29    35     0     0     1     0     90   0   0
44 44     2 1829 0    19    18     1     1     1     0    210   0   0
45 45     2 1562 0    26    30     1     1     1     1     90   0   0
46 46     2 1470 0    27    34     1     1     0     1    240   0   0
> ccdat3
     id group    t c r.age d.age r.sex d.sex r.cmv d.cmv w.time fab mtx
93   93     3 2640 0    18    23     1     1     0     0    750   0   0
94   94     3 2430 0    29    26     1     1     0     1     24   0   0
95   95     3 2252 0    35    31     1     0     0     0    120   0   0
96   96     3 2140 0    27    17     1     1     1     1    210   0   0
97   97     3 2133 0    36    39     0     1     0     0    240   0   0
98   98     3 1238 0    24    28     1     0     1     1    240   0   0
99   99     3 1631 0    27    21     1     0     1     0    690   1   1
100 100     3 2024 0    35    41     0     1     0     0    105   1   0
```

Βρίσκουμε τις εκτιμήτριες Kaplan-Meier για όλα τα δεδομένα καθώς και για κάθε group ξεχωριστά χρησιμοποιώντας την εντολή survfit() και με την χρήση της εντολής summary().

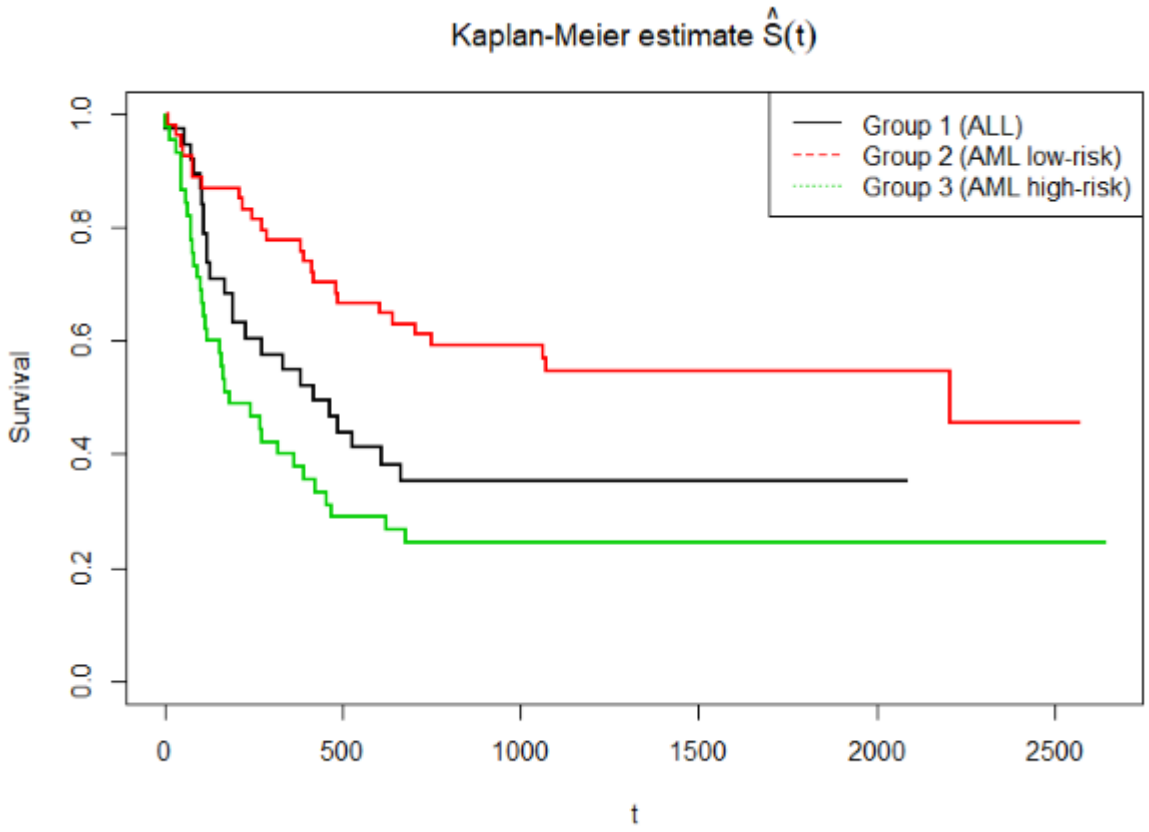
```
> outp1
Call: survfit(formula = Surv(t, c) ~ 1, data = cc)

      n  events  median 0.95LCL 0.95UCL
    137      83     481     381    1063
> outp1g
Call: survfit(formula = Surv(t, c) ~ group, data = cc)

         n events median 0.95LCL 0.95UCL
group=1 38     24    418     194      NA
group=2 54     25   2204     704      NA
group=3 45     34    183     115     456
> summary(outp1)
Call: survfit(formula = Surv(t, c) ~ 1, data = cc)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1    137       1    0.993 0.00727        0.979        1.000
    2    136       1    0.985 0.01025        0.966        1.000
   10    135       1    0.978 0.01250        0.954        1.000
   16    134       1    0.971 0.01438        0.943        0.999
   32    133       1    0.964 0.01602        0.933        0.995
   35    132       1    0.956 0.01748        0.923        0.991
   47    131       2    0.942 0.02003        0.903        0.982
> summary(outp1g)
Call: survfit(formula = Surv(t, c) ~ group, data = cc)

                group=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     38       1    0.974  0.0260        0.924        1.000
   55     37       1    0.947  0.0362        0.879        1.000
   74     36       1    0.921  0.0437        0.839        1.000
   86     35       1    0.895  0.0498        0.802        0.998
  104     34       1    0.868  0.0548        0.767        0.983

                group=2
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   10     54       1    0.981  0.0183        0.946        1.000
   35     53       1    0.963  0.0257        0.914        1.000
   48     52       1    0.944  0.0312        0.885        1.000
   53     51       1    0.926  0.0356        0.859        0.998

                group=3
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2     45       1    0.978  0.0220        0.936        1.000
   16     44       1    0.956  0.0307        0.897        1.000
   32     43       1    0.933  0.0372        0.863        1.000
   47     42       2    0.889  0.0468        0.802        0.986
```

Εφόσον έχουμε υπολογίσει τις εκτιμήτριες είμαστε έτοιμοι για τις γραφικές παραστάσεις των εκτιμήσεων που θα μας δώσουν μια καλύτερη εικόνα για τα πρώτα μας συμπεράσματα.

Αρχικά για κάθε ομάδα ξεχωριστά:

```
> plot(outp1g, main=expression(paste("Kaplan-Meier estimate",hat(S)(t))),xlab="t",ylab="Survival",lwd=2,col=1:3)
> legend("topright",c("Group 1", "Group 2", "Group 3"),lty=1:3,col=1:3)
```

και παίρνουμε:



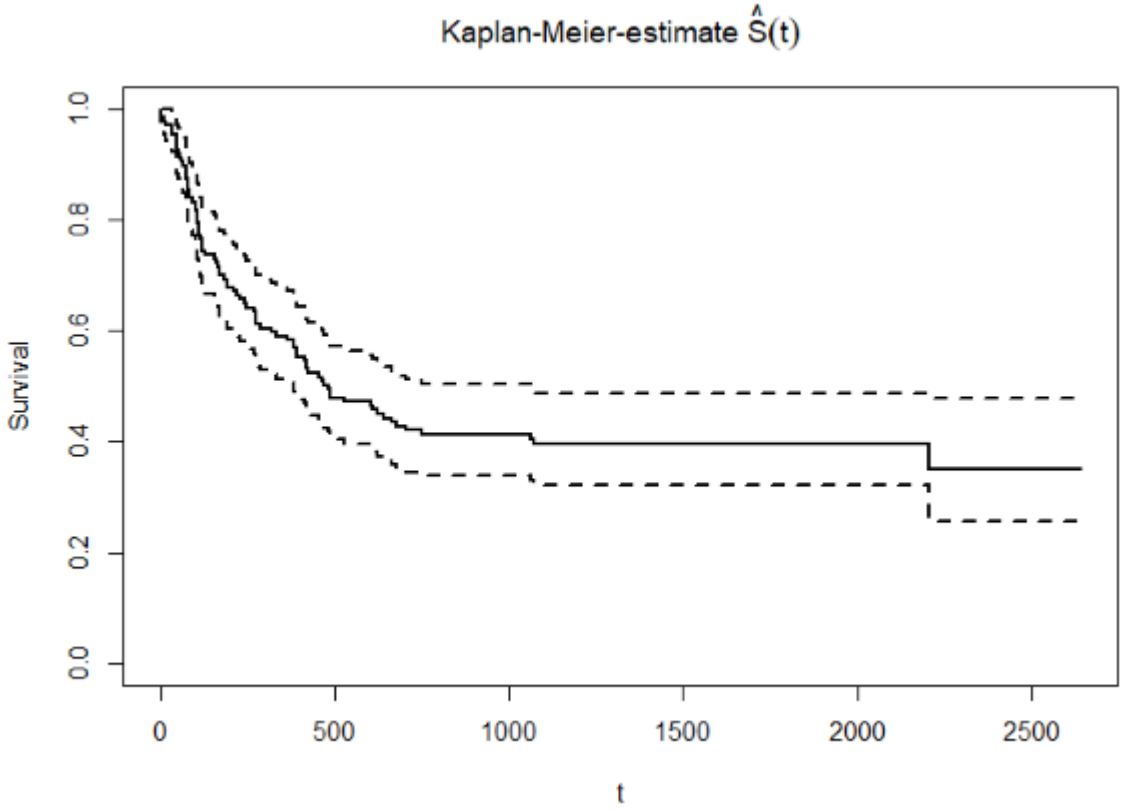
Σχήμα 5.1: *Γραφική παράσταση της εκτίμησης Kaplan-Meier των ομάδων*

Οι καμπύλες των εκτιμήσεων για αυτούς τους τρεις τύπους λευχαιμίας παρουσιάζονται στο Σχήμα 5.1. Παρατηρούμε ότι υπάρχουν έντονες διαφοροποιήσεις μεταξύ των τριών τύπων λευχαιμίας και μάλιστα φαίνεται πως η ομάδα 2 (AML, low risk) να «αντέχει» περισσότερο σε σχέση με τις άλλες δύο. Παρατηρούμε ότι την μικρότερη διάρκεια ζωής μετά το μόσχευμα την έχει η ομάδα 3 δηλαδή τα άτομα με οξεία μυελοκυτταρρική λευχαιμία (AML high-risk).

Έπειτα θα κάνουμε την γραφική παράσταση της εκτίμησης Kaplan-Meier για όλα τα δεδομένα χωρίς τον διαχωρισμό σε ομάδες

```
> plot(outp1, main=expression(paste("Kaplan-Meier estimate ",hat(S)(t))),xlab="t",ylab="Survival",lwd=2)
```

και παίρνουμε την γραφική παράσταση:



Σχήμα 5.2: *Γραφική παράσταση της εκτίμησης Kaplan-Meier όλων των δεδομένων*

Οι διακεκομμένες γραμμές δείχνουν ένα διάστημα εμπιστοσύνης ενώ η κανονική γραμμή είναι η κλιμακωτή συνάρτηση της εκτίμησης Kaplan-Meier.

Έπειτα εφόσον έχουμε χωρίσει τα δεδομένα σε ομάδες και αφού έχουμε προσδιορίσει για κάθε ομάδα τις εκτιμήσεις της Kaplan-Meier θα πραγματοποιήσουμε τις γραφικές παραστάσεις των $ln\{-ln\hat{S}_k(t)\}$ έναντι του t. Αυτό το αναφέραμε στην παράγραφο 2.2.2. Κάνοντας αυτό πραγματοποιείται ο πιο απλός έλεγχος ορθότητας της υπόθεσης αναλογικής διακινδύνευσης ώστε να έχει νόημα η προσαρμογή του μοντέλου στα δεδομένα.

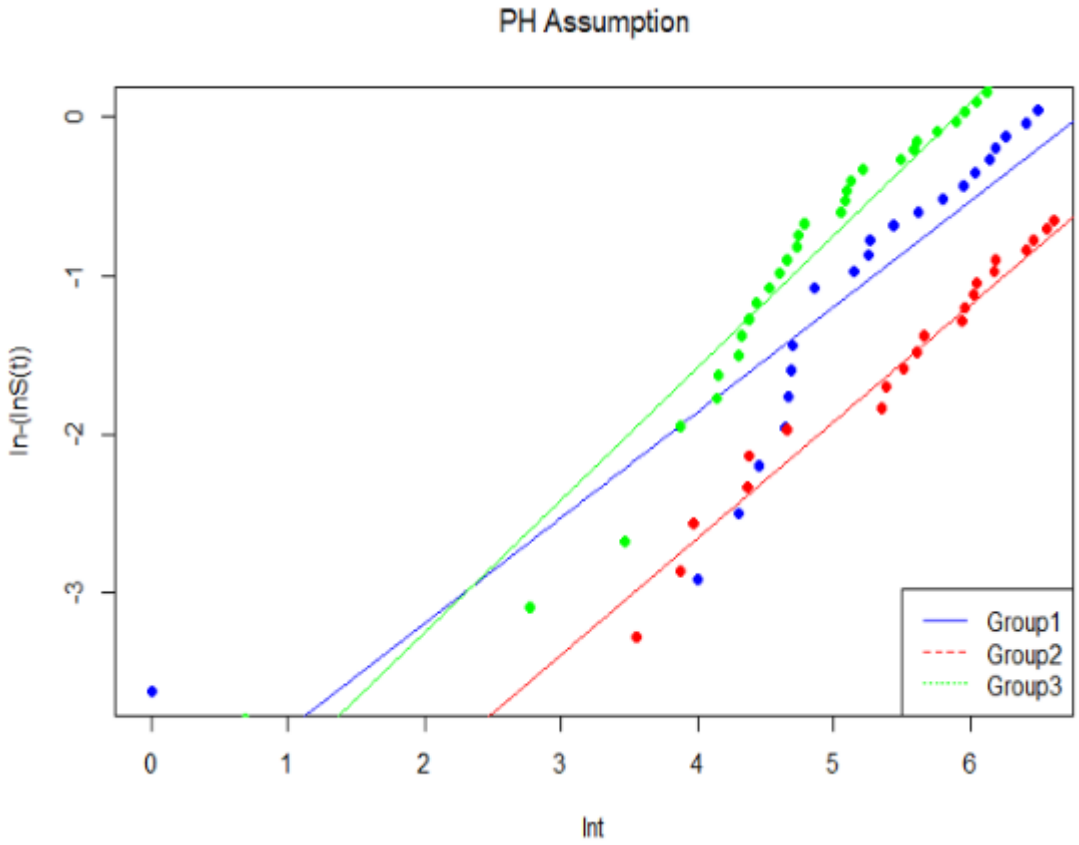
Το παραπάνω γίνεται με τον εξής κώδικα:

```
> group1<-Surv(t[group=="1"],c[group=="1"])
> group2<-Surv(t[group=="2"],c[group=="2"])
> group3<-Surv(t[group=="3"],c[group=="3"])
> outp1ph<-survfit(group1~1,type="kaplan-meier",data=cc)
> outp2ph<-survfit(group2~1,type="kaplan-meier",data=cc)
> outp3ph<-survfit(group3~1,type="kaplan-meier",data=cc)
> Utime1<-outp1ph$time[outp1ph$n.event==1]
> Utime2<-outp2ph$time[outp2ph$n.event==1]
> Utime3<-outp3ph$time[outp3ph$n.event==1]
> SKM1<-outp1ph$surv[outp1ph$n.event==1]
> SKM2<-outp2ph$surv[outp2ph$n.event==1]
> SKM3<-outp3ph$surv[outp3ph$n.event==1]
```

Και χρησιμοποιώντας τον κώδικα:

```
> plot(log(Utime1),log(-log(SKM1)),main=expression(paste("PH Assumption")),xlab="lnt",ylab="ln-(lnS(t))",col="blue",pch=19)
> abline(lm(log(-log(SKM1))~log(Utime1)),col="blue")
> points(log(Utime2),log(-log(SKM2)),col="red",pch=19)
> abline(lm(log(-log(SKM2))~log(Utime2)),col="red")
> points(log(Utime3),log(-log(SKM3)),col="green",pch=19)
> abline(lm(log(-log(SKM3))~log(Utime3)),col="green")
> legend("bottomright",c("Group1","Group2","Group3"),col=c("blue","red","green"),lty=1:3)
```

Πάιρνουμε την γραφική παράσταση:



Σχήμα 5.3: *Γραφικός έλεγχος του μοντέλου αναλογικής διακινδύνευσης για τις 3 ομάδες*

Στο σχήμα 5.3 θα θέλαμε να είναι οι 3 γραμμές παράλληλες βλέπουμε ότι υπάρχει ένα μικρό πρόβλημα ανάμεσα στην 3η ομάδα με την 1η. Παρατηρώντας και τον πίνακα των δεδομένων βλέπουμε ότι κάποια άτομα είναι εντελώς εκτός των προβλεπόμενων τιμών και ίσως για αυτό προκύπτει ένα μικρό πρόβλημα. Για αυτό δεν απορρίπτουμε ακόμα την αναλογική διακινδύνευση.

Λόγω αυτού του προβλήματος θα κάνουμε και άλλους ελέγχους για να ενισχύσουμε τα συμπεράσματα μας. Ένας τέτοιος έλεγχος που αναφέραμε και στο κεφάλαιο 2.1.2 είναι ο έλεγχος Log-Rank.

Κάνουμε στην συνέχεια έναν έλεγχο Log-Rank για τις διαφορετικές ομάδες με την εντολή survdiff():

```
> outp2<-survdiff(Surv(t,c) ~ group)
> outp2
Call:
survdiff(formula = Surv(t, c) ~ group)

         N Observed Expected (O-E)^2/E (O-E)^2/V
group=1 38       24     21.9     0.211     0.289
group=2 54       25     40.0     5.604    11.012
group=3 45       34     21.2     7.756    10.529

 Chisq= 13.8  on 2 degrees of freedom, p= 0.001
```

Από το p-value=0.001 του παραπάνω ελέγχου συμπεραίνουμε ότι οι ομάδες διαφέρουν σημαντικά στην επιβίωση καθώς η μηδενική υπόθεση είναι ότι δεν υπάρχει καμία διαφορά στις καμπύλες επιβίωσης των ομάδων. Με αυτό το p-value απορρίπτουμε την μηδενική υπόθεση.

Συμπεραίνουμε λοιπόν ότι η μεταβλητή της ομάδας θα είναι σημαντική στην στατιστική μας ανάλυση κάτι που ήταν αναμενόμενο.

Στην συνέχεια, θα προσαρμόσουμε τα δεδομένα μας σε ένα μοντέλο αναλογικής διακινδύνευσης του Cox με την βοήθεια του library(lattice) (Deepayan Sarkar et al.) και την εντολή coxph().

Αρχικά όμως θα φτιάξουμε την κατηγορική μεταβλητή να έχεις ως κατηγορία αναφοράς την ομάδα 2 με την εντολή factor. Αυτά γίνονται με τον παρακάτω κώδικα:

```
> groupf<-factor(group,levels=c(2,1,3))
> mod1 <- coxph(Surv(t,c)~groupf+r.age+d.age+r.sex+d.sex+r.cmv+d.cmv+w.time+fab+mtx)
```

Έχοντας πλέον ορίσει το μοντέλο αναλογικής διακινδύνευσης του Cox και κάνοντας την χρήση ξανά της εντολής summary() παίρνουμε τα παρακάτω:

```
> summary(mod1)
Call:
coxph(formula = Surv(t, c) ~ groupf + r.age + d.age + r.sex +
    d.sex + r.cmv + d.cmv + w.time + fab + mtx)

  n= 137, number of events= 83

              coef  exp(coef)   se(coef)      z Pr(>|z|)
groupf1  1.0624620  2.8934861  0.3705119  2.868  0.00414 **
groupf3  0.8742014  2.3969603  0.2821845  3.098  0.00195 **
r.age    0.0139626  1.0140605  0.0205096  0.681  0.49601
d.age   -0.0021921  0.9978103  0.0186648 -0.117  0.90651
r.sex   -0.1093030  0.8964587  0.2412718 -0.453  0.65053
d.sex    0.0333095  1.0338704  0.2416459  0.138  0.89036
r.cmv   -0.0606449  0.9411574  0.2546450 -0.238  0.81176
d.cmv   -0.0480256  0.9531094  0.2472610 -0.194  0.84600
w.time  -0.0003417  0.9996584  0.0003927 -0.870  0.38424
fab      0.8018912  2.2297539  0.2822418  2.841  0.00450 **
mtx      0.2918278  1.3388724  0.2542193  1.148  0.25099
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        exp(coef) exp(-coef) lower .95 upper .95
groupf1    2.8935     0.3456    1.3997     5.981
groupf3    2.3970     0.4172    1.3787     4.167
r.age      1.0141     0.9861    0.9741     1.056
d.age      0.9978     1.0022    0.9620     1.035
r.sex      0.8965     1.1155    0.5587     1.438
d.sex      1.0339     0.9672    0.6438     1.660
r.cmv      0.9412     1.0625    0.5714     1.550
d.cmv      0.9531     1.0492    0.5870     1.547
w.time     0.9997     1.0003    0.9989     1.000
fab        2.2298     0.4485    1.2824     3.877
mtx        1.3389     0.7469    0.8135     2.204

Concordance= 0.677  (se = 0.032 )
Likelihood ratio test= 26.1  on 11 df,   p=0.006
Wald test            = 25.2  on 11 df,   p=0.009
Score (logrank) test = 26.65  on 11 df,   p=0.005
> AIC(mod1)
[1] 742.4931
```

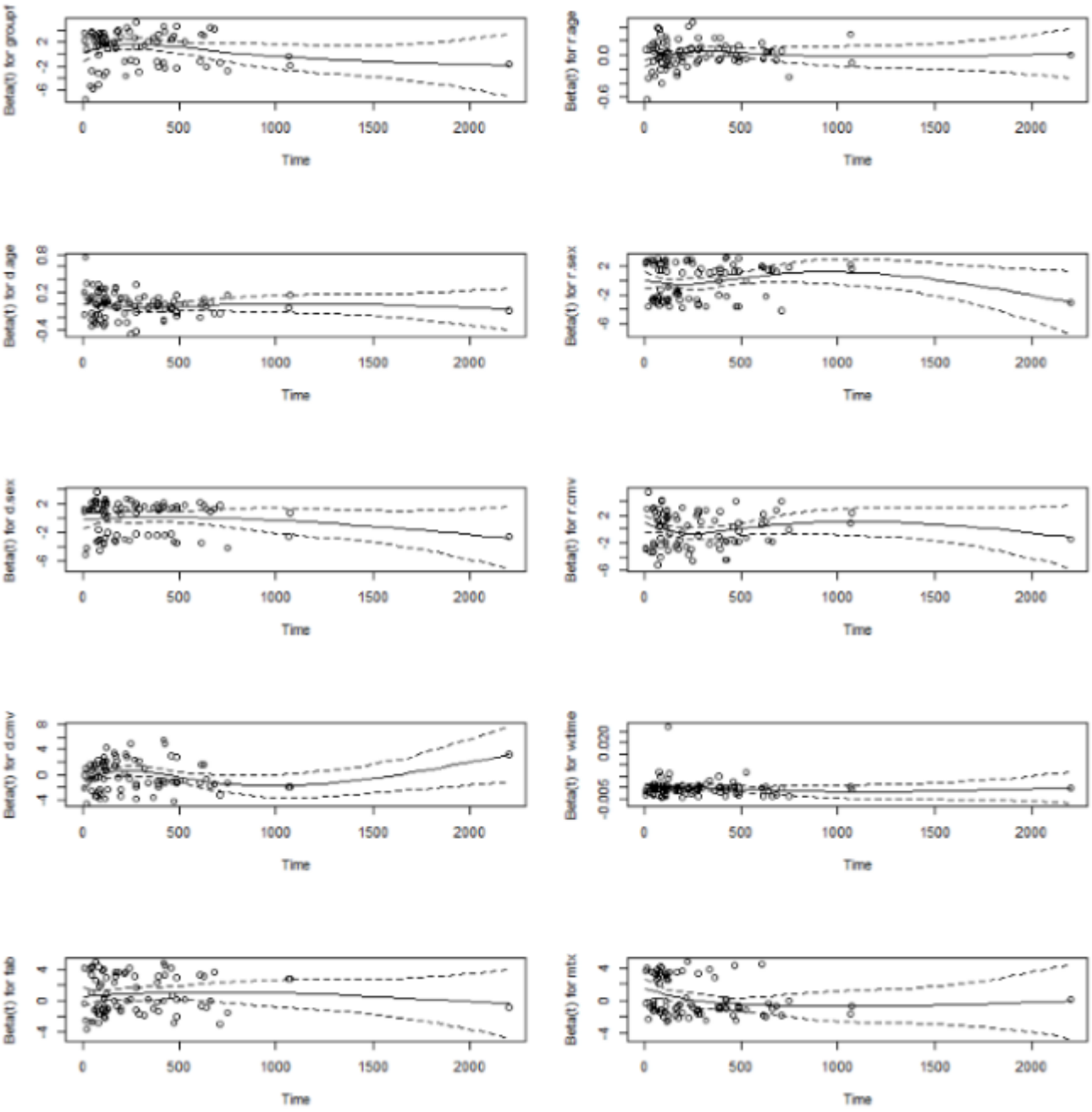
Παρατηρούμε από τα p-values ότι οι σημαντικές μεταβλητές του μοντέλου είναι η **groupf** η **fab**.

Για να ελέγξουμε την αναλογική διακινδύνευση στο μοντέλο του Cox θα εξετάσουμε τα Schoenfeld υπόλοιπα καθώς και τα martingale υπόλοιπα για να εξάγουμε περαιτέρω συμπεράσματα. Για τον έλεγχο υπόθεσης της αναλογικής διακινδύνευσης κάνουμε χρήση της cox.zph() και έπειτα κάνουμε το plot και παίρνουμε τα υπόλοιπα Schoenfeld.

```
> test.ph1<-cox.zph(mod1,transform='identity')
> test.ph1
         chisq df      p
groupf   3.864  2 0.1449
r.age    0.222  1 0.6373
d.age    2.537  1 0.1112
r.sex    0.034  1 0.8536
d.sex    0.556  1 0.4560
r.cmv    0.103  1 0.7486
d.cmv    0.257  1 0.6121
w.time   1.097  1 0.2950
fab      0.419  1 0.5173
mtx      8.373  1 0.0038
GLOBAL 19.531 11 0.0522
```

Παρατηρούμε από το global p-value ότι δεν είναι στατιστικά σημαντικό άρα μπορούμε να συμπεράνουμε ότι η υπόθεση αναλογικής διακινδύνευσης είναι σωστή.

Και η γραφική παράσταση των υπολοίπων Schoenfeld:



Σχήμα 5.4: *Υπόλοιπα Schoenfeld του μοντέλου μας*

Παρατηρούμε στο σχήμα 5.4 ότι δεν υπάρχει κάποιο μοτίβο με τον χρόνο άρα δεν έχουμε μεταβλητές εξαρτημένες από τον χρόνο και άρα η υπόθεση αναλογικής διακινδύνευσης ισχύει.

Για να βγάλουμε σίγουρα συμπεράσματα για τις μεταβλητές θα κάνουμε χρήση της διαδοχικής αφαίρεσης μεταβλητών και θα δούμε ποιο μοντέλο έχει το μικρότερο AIC. Αυτό μπορούμε να το κάνουμε στην R αρκετά εύκολα.

Κάνουμε το παραπάνω με τον εξής κώδικα:

```
> step(mod1,direction="backward",test="Chisq")
Start:  AIC=742.49
Surv(time, indicator) ~ groupf + r.age + d.age + r.sex + d.sex +
    r.cmv + d.cmv + w.time + fab + mtx

         Df    AIC     LRT Pr(>Chi)
- d.age   1 740.51  0.0138 0.906556
- d.sex   1 740.51  0.0190 0.890229
- d.cmv   1 740.53  0.0378 0.845824
- r.cmv   1 740.55  0.0567 0.811725
- r.sex   1 740.70  0.2045 0.651083
- r.age   1 740.95  0.4590 0.498110
- w.time  1 741.34  0.8435 0.358403
- mtx     1 741.78  1.2827 0.257406
<none>      742.49
- fab     1 748.73  8.2319 0.004116 **
- groupf  2 752.05 13.5597 0.001136 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=740.51
Surv(time, indicator) ~ groupf + r.age + r.sex + d.sex + r.cmv +
    d.cmv + w.time + fab + mtx

         Df    AIC     LRT  Pr(>Chi)
- d.sex   1 738.52  0.0139 0.9060142
- d.cmv   1 738.55  0.0453 0.8314118
- r.cmv   1 738.56  0.0498 0.8233938
- r.sex   1 738.71  0.2026 0.6525980
- r.age   1 739.28  0.7743 0.3788802
- w.time  1 739.35  0.8410 0.3591076
- mtx     1 739.81  1.3035 0.2535744
<none>      740.51
- fab     1 746.76  8.2483 0.0040791 **
- groupf  2 750.83 14.3207 0.0007768 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=738.52
Surv(time, indicator) ~ groupf + r.age + r.sex + r.cmv + d.cmv +
    w.time + fab + mtx

         Df    AIC     LRT Pr(>Chi)
- r.cmv   1 736.57  0.0467 0.828939
- d.cmv   1 736.57  0.0481 0.826410
- r.sex   1 736.72  0.1944 0.659283
- r.age   1 737.29  0.7653 0.381677
- w.time  1 737.44  0.9215 0.337072
- mtx     1 737.81  1.2940 0.255307
<none>      738.52
- fab     1 744.78  8.2545 0.004065 **
- groupf  2 748.86 14.3357 0.000771 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Step:  AIC=738.52
Surv(time, indicator) ~ groupf + r.age + r.sex + r.cmv + d.cmv +
    w.time + fab + mtx

         Df    AIC     LRT Pr(>Chi)
- r.cmv   1 736.57  0.0467 0.828939
- d.cmv   1 736.57  0.0481 0.826410
- r.sex   1 736.72  0.1944 0.659283
- r.age   1 737.29  0.7653 0.381677
- w.time  1 737.44  0.9215 0.337072
- mtx     1 737.81  1.2940 0.255307
<none>      738.52
- fab     1 744.78  8.2545 0.004065 **
- groupf  2 748.86 14.3357 0.000771 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=736.57
Surv(time, indicator) ~ groupf + r.age + r.sex + d.cmv + w.time +
    fab + mtx

         Df    AIC     LRT  Pr(>Chi)
- d.cmv   1 734.66  0.0878 0.7670079
- r.sex   1 734.81  0.2430 0.6220233
- r.age   1 735.29  0.7219 0.3955210
- w.time  1 735.51  0.9417 0.3318441
- mtx     1 735.82  1.2516 0.2632519
<none>      736.57
- fab     1 742.82  8.2507 0.0040735 **
- groupf  2 746.87 14.3066 0.0007823 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=734.66
Surv(time, indicator) ~ groupf + r.age + r.sex + w.time + fab +
    mtx

         Df    AIC     LRT  Pr(>Chi)
- r.sex   1 732.86  0.2074 0.6488214
- r.age   1 733.31  0.6554 0.4181812
- w.time  1 733.58  0.9216 0.3370610
- mtx     1 733.88  1.2239 0.2685911
<none>      734.66
- fab     1 741.05  8.3934 0.0037658 **
- groupf  2 744.94 14.2807 0.0007925 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Step:  AIC=732.86
Surv(time, indicator) ~ groupf + r.age + w.time + fab + mtx

         Df    AIC     LRT  Pr(>Chi)
- r.age   1 731.56  0.6944 0.4046566
- w.time  1 732.02  1.1550 0.2825080
- mtx     1 732.23  1.3637 0.2428957
<none>      732.86
- fab     1 739.76  8.8982 0.0028546 **
- groupf  2 743.14 14.2774 0.0007938 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=731.56
Surv(time, indicator) ~ groupf + w.time + fab + mtx

         Df    AIC     LRT  Pr(>Chi)
- w.time  1 730.89  1.3339 0.2481038
- mtx     1 731.51  1.9544 0.1621149
<none>      731.56
- fab     1 737.83  8.2737 0.0040224 **
- groupf  2 741.45 13.8897 0.0009636 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=730.89
Surv(time, indicator) ~ groupf + fab + mtx

         Df    AIC     LRT Pr(>Chi)
- mtx     1 730.85  1.9580 0.161729
<none>      730.89
- fab     1 737.20  8.3083 0.003947 **
- groupf  2 739.45 12.5559 0.001877 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=730.85
Surv(time, indicator) ~ groupf + fab

         Df    AIC     LRT  Pr(>Chi)
<none>      730.85
- fab     1 737.14  8.2902 0.0039859 **
- groupf  2 740.81 13.9572 0.0009316 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Call:
coxph(formula = Surv(time, indicator) ~ groupf + fab)

           coef exp(coef) se(coef)     z       p
groupf1 0.9045    2.4707   0.3203 2.824 0.00475
groupf3 0.8525    2.3456   0.2684 3.176 0.00149
fab     0.7695    2.1587   0.2703 2.847 0.00442

Likelihood ratio test=21.74  on 3 df, p=7.38e-05
n= 137, number of events= 83
```

Παρατηρούμε ότι το αρχικό μας μοντέλο έχει AIC 742.4931 και με την διαδοχική αφαίρεση καταλήγουμε σε μοντέλο με 2 συμμεταβλητές την groupf και την fab και το μοντέλο αυτό έχει AIC 730.85 που είναι αρκετά καλύτερο του αρχικού μας μοντέλου. Άρα τελικά η mtx (που είναι η τελευταία που αφαιρείται πριν το τελικό μοντέλο) καθώς και οι υπόλοιπες συμμεταβλητές δεν είναι σημαντικές στο μοντέλο μας.

Εισάγουμε έπειτα το τελικό μας μοντέλο στην R και κάνουμε κάποιους από του παραπάνω ελέγχους για το τελικό μας μοντέλο:

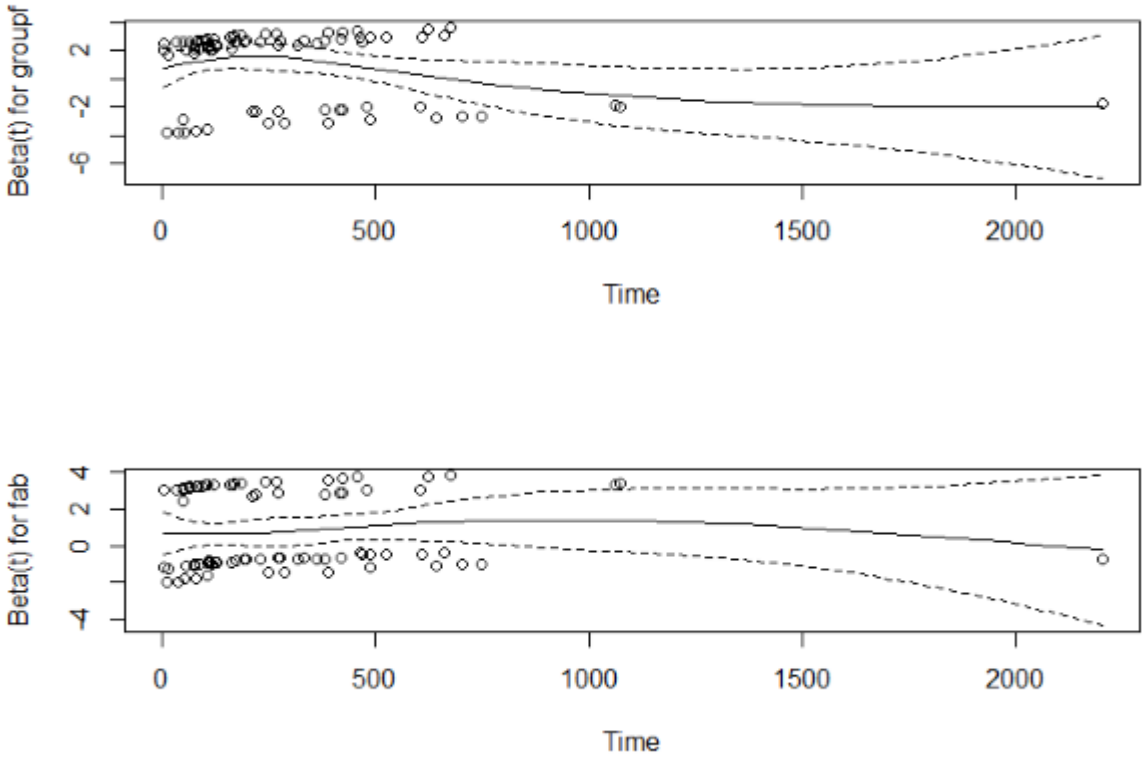
```
> modf<-coxph(Surv(t,c)~groupf+fab)
> AIC(modf)
[1] 730.8491
> modf
Call:
coxph(formula = Surv(t, c) ~ groupf + fab)

          coef exp(coef) se(coef)     z       p
groupf1 0.9045    2.4707   0.3203 2.824 0.00475
groupf3 0.8525    2.3456   0.2684 3.176 0.00149
fab     0.7695    2.1587   0.2703 2.847 0.00442

Likelihood ratio test=21.74  on 3 df, p=7.38e-05
n= 137, number of events= 83
> summary(modf)
Call:
coxph(formula = Surv(t, c) ~ groupf + fab)

  n= 137, number of events= 83

          coef exp(coef) se(coef)     z Pr(>|z|)
groupf1 0.9045    2.4707   0.3203 2.824  0.00475 **
groupf3 0.8525    2.3456   0.2684 3.176  0.00149 **
fab     0.7695    2.1587   0.2703 2.847  0.00442 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        exp(coef) exp(-coef) lower .95 upper .95
groupf1     2.471     0.4047     1.319     4.629
groupf3     2.346     0.4263     1.386     3.969
fab         2.159     0.4632     1.271     3.667

Concordance= 0.646  (se = 0.031 )
Likelihood ratio test= 21.74  on 3 df,   p=7e-05
Wald test            = 20.97  on 3 df,   p=1e-04
Score (logrank) test = 22.08  on 3 df,   p=6e-05
```

Για τον έλεγχο υπόθεσης της αναλογικής διακινδύνευσης κάνουμε χρήση της cox.zph() και έπειτα κάνουμε το plot και παίρνουμε τα υπόλοιπα Schoenfeld για το τελικό μας μοντέλο:

```
> test.phf<-cox.zph(modf,transform='identity')
> test.phf
       chisq df    p
groupf 3.834  2 0.15
fab    0.548  1 0.46
GLOBAL 3.861  3 0.28
```

Παρατηρούμε πάλι από το global p-value ότι δεν είναι στατιστικά σημαντικό άρα μπορούμε να συμπεράνουμε ότι η υπόθεση αναλογικής διακινδύνευσης είναι σωστή.

Παρακάτω έχουμε και τα υπόλοιπα Schoenfeld για το τελικό μας μοντέλο:



Σχήμα 5.5: *Υπόλοιπα Schoenfeld του τελικού μας μοντέλου*

Παρατηρούμε στο σχήμα 5.5 ότι δεν υπάρχει κάποιο μοτίβο με τον χρόνο άρα δεν έχουμε μεταβλητές εξαρτημένες από τον χρόνο και άρα η υπόθεση αναλογικής διακινδύνευσης ισχύει.

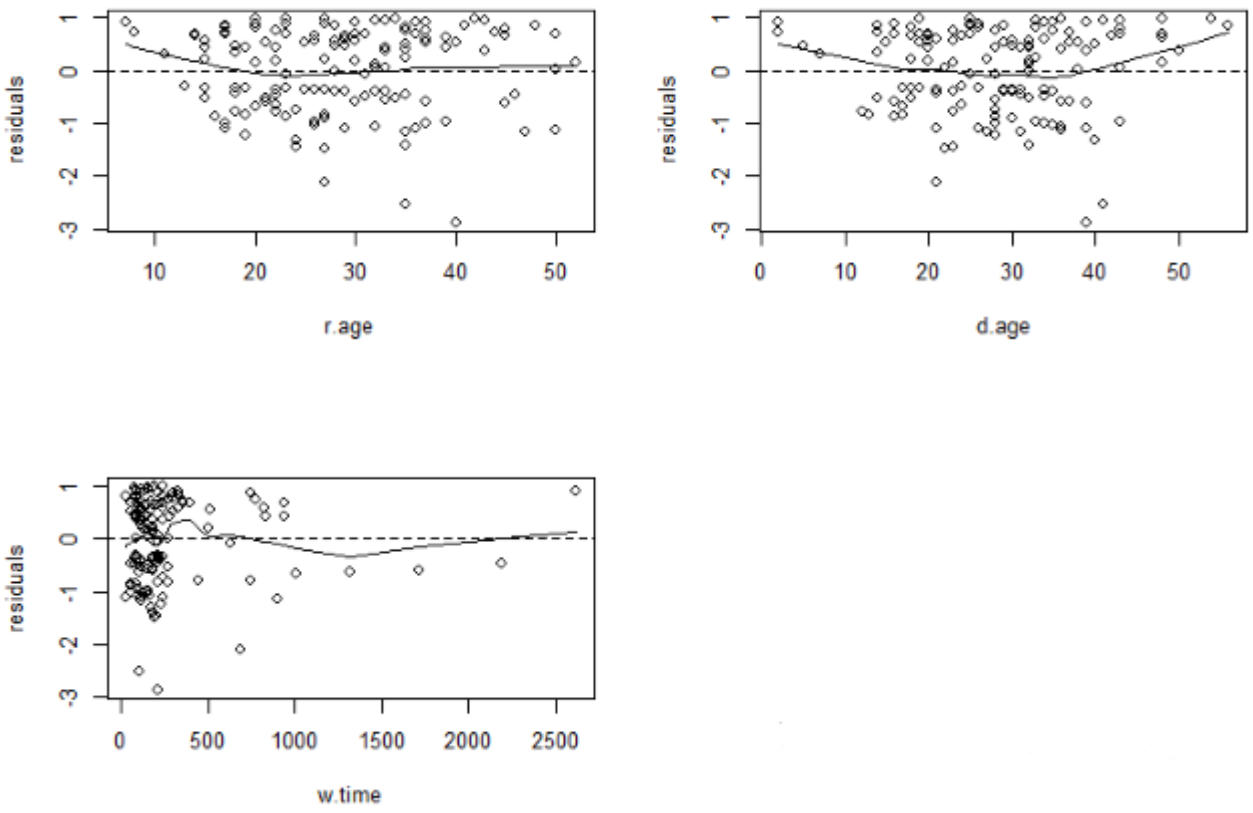
Με την χρήση της εντολής summary προηγουμένως μπορούμε να δώσουμε κάποιες ερμηνείες. Θυμίζουμε ότι το μοντέλο του Cox είναι : $h(t;x) = h_0(t)e^{\beta' x}$

Άρα οι τιμές του $e^{\hat{\beta}}$ μας δείχνουν κατά πόσο μια συμμεταβλητή επιδρά στην διάρκεια ζωής, όταν οι υπόλοιπες συμμεταβλητές είναι σταθερές π.χ για έναν ασθενή της ομάδας 1 δηλαδή με οξεία λεμφοβλαστική λευχαιμία (ALL) σε σχέση με έναν της ομάδας 2 (δηλαδή της ομάδας AML-low risk) η βασική συνάρτηση διακινδύνευσης πολλαπλασιάζεται κατά $h_0(t) * 2.5$, δηλαδή ο κίνδυνος μέχρι την υποτροπή ή το θάνατο αυξάνεται κατά 150%. Αντίστοιχα για έναν ασθενή της ομάδας 3 (AML high risk) σε σχέση με έναν της ομάδας 2 (AML- low risk) η βασική συνάρτηση διακινδύνευσης πολλαπλασιάζεται κατά 2.3, δηλαδή ο κίνδυνος αυξάνεται κατά 130%. Συνεπώς η ομάδα 2 δείχνει να αντέχει περισσότερο από τις άλλες δύο, το οποίο είναι και εμφανές από το Σχήμα 5.1 πιο πάνω. Για κάποιον που έχει βαθμό fab 4 ή 5 σε σχέση με κάποιον που δεν έχει η βασική συνάρτηση διακινδύνευσης πολλαπλασιάζεται κατά $h_0(t) * 2.159$ δηλαδή αυξάνεται κατά 120% περίπου .

Θα ελέγξουμε επίσης για τις ποσοτικές μεταβλητές του μοντέλου μας τα martingale υ-πόλοιπα. Αυτά γίνονται με τον εξής κώδικα:

```
> par(mfrow=c(2, 2))
> resmar<-residuals(mod1, type="martingale")
> X<- as.matrix(cc[ , c("r.age", "d.age","w.time")])
> for (j in 1:3) { # residual plots
+ plot(X[, j], resmar, xlab=c("r.age", "d.age","w.time")[j], ylab="residuals")
+ abline(h=0, lty=2)
+ lines(lowess(X[, j], resmar, iter=0))
+ }
```

και παίρνουμε την γραφική παράσταση των υπολοίπων martingale:



Σχήμα 5.6: *Υπόλοιπα Martingale των ποσοτικών μεταβλητών του μοντέλου μας*

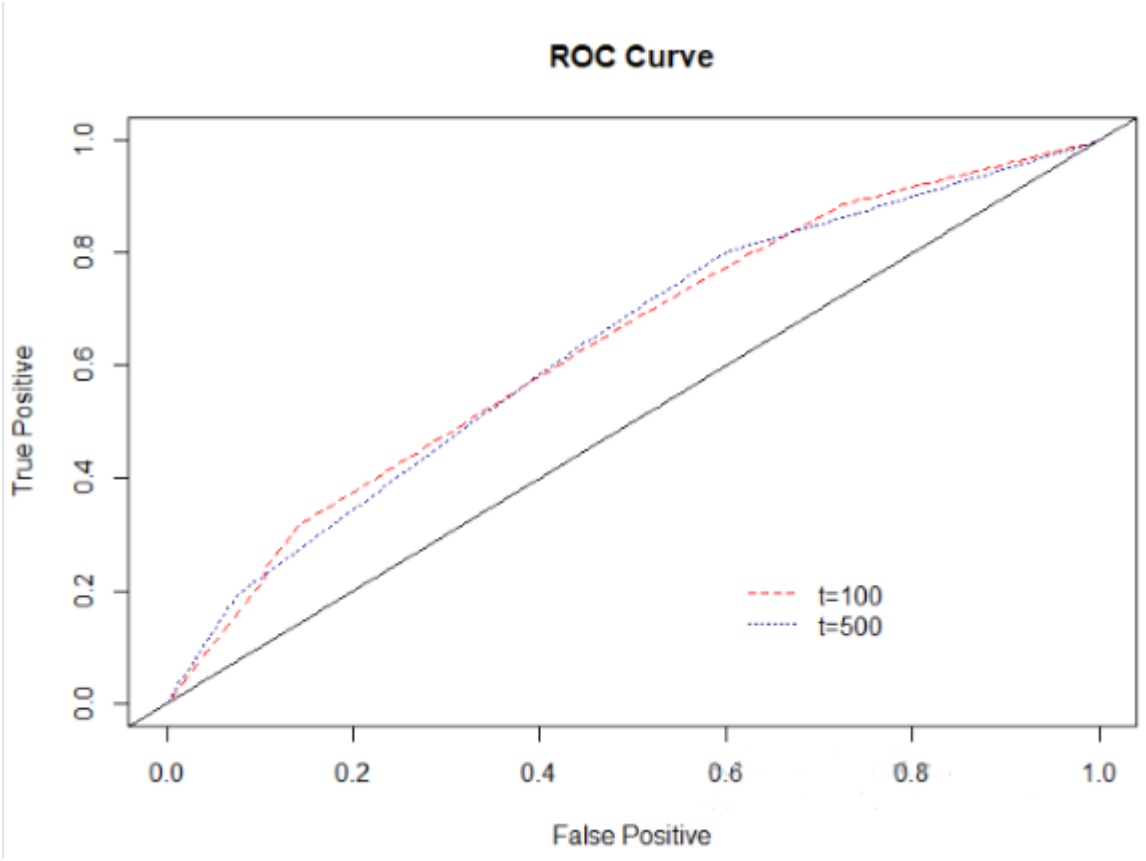
Ήταν αναμενόμενο καθώς οι μεταβλητές αυτές δεν είναι σημαντικές στο μοντέλο μας και δεν συμβάλλουν ιδιαίτερα, τα υπόλοιπα να μην μας αλλάζουν τα συμπεράσματα μας μέχρι τώρα.

Τέλος για την προβλεπτική ικανότητα του τελικού μας μοντέλου δίνουμε την καμπύλη ROC και το εμβαδόν της περιοχής κάτω από την καμπύλη ROC (AUC). Για να το κάνουμε αυτό θα φορτώσουμε το package risksetROC.

Αφού το φορτώσουμε φτιάχνουμε την καμπύλη ROC με τον παρακάτω κώδικα.

```
> eta<-mod$linear.predictor
> ROC100=risksetROC(Stime= t, status=c,marker=eta, predict.time=100, method="Cox",main="ROC Curve", lty=2, col="red", ylab="True Positiv
e", xlab="False Positive")
> ROC500=risksetROC(Stime= t, status=c,marker=eta, predict.time=500, method="Cox",plot=FALSE)
> lines(ROC500$FP,ROC500$TP, lty=3,col="darkblue")
> legend(.6,.25,lty=c(2,3),col=c("red","darkblue"), legend=c("t=100","t=500"), bty="n")
```

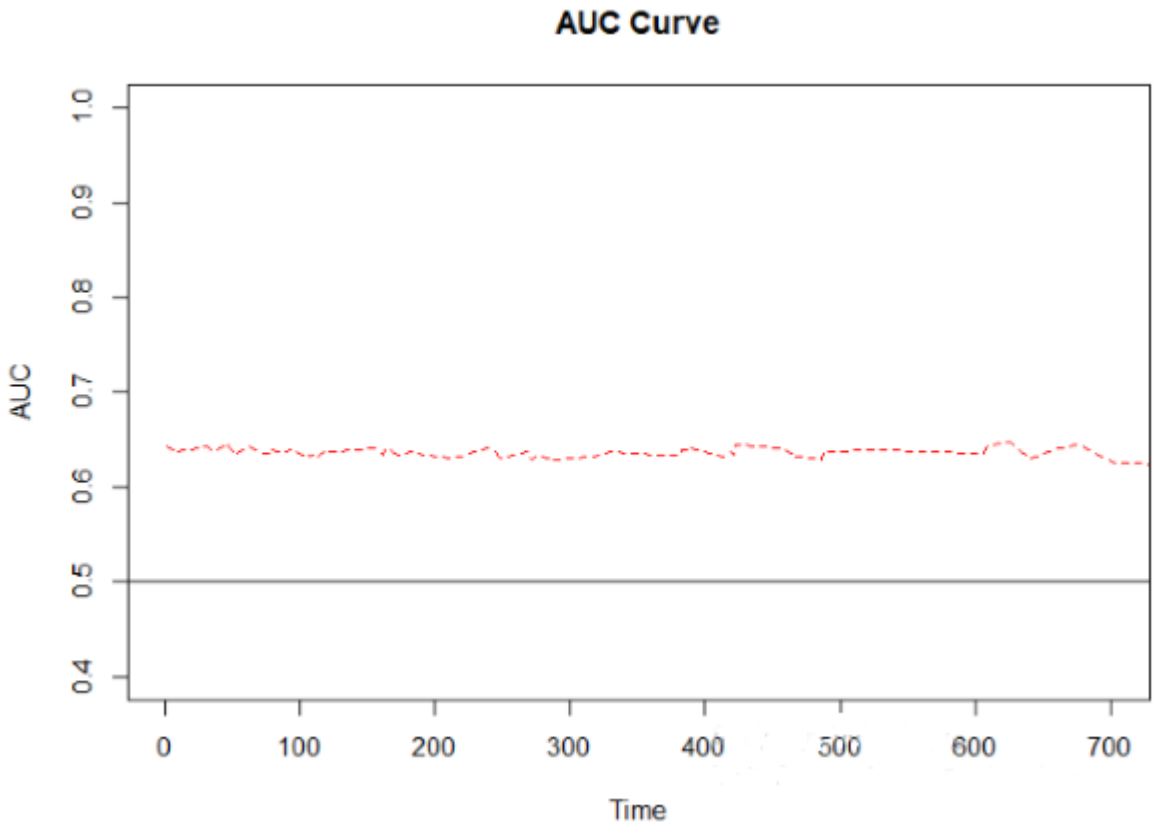
και παίρνουμε την γραφική παράσταση της καμπύλης ROC στο σχήμα 5.7:



Σχήμα 5.7: *Καμπύλη ROC του μοντέλου μας*

Παρατηρούμε από την καμπύλη ROC σε δυο διαφορετικές χρονικές στιγμές ότι δεν διαφέρει αρκετά, επίσης φαίνονται καλές οι καμπύλες αν και καλύτερα συμπεράσματα θα πάρουμε και από το AUC . Όσο πιο πάνω αριστερά τείνει μια καμπύλη ROC τόσο καλύτερη είναι αλλά ας δούμε και την τιμή του AUC για καλύτερα συμπεράσματα.

Υπολογίζουμε το AUC με τον παρακάτω κώδικα και βλέπουμε την τιμή του από την γραφική παράσταση.



Σχήμα 5.8: *Εμβαδόν της περιοχής κάτω από την καμπύλη ROC του μοντέλου μας*

Παρατηρούμε ότι η τιμή κυμαίνεται κοντά στο 0.7 που σύμφωνα με τα κριτήρια που αναφέραμε σε προηγούμενο κεφάλαιο είναι σχεδόν αποδεκτή διάκριση.

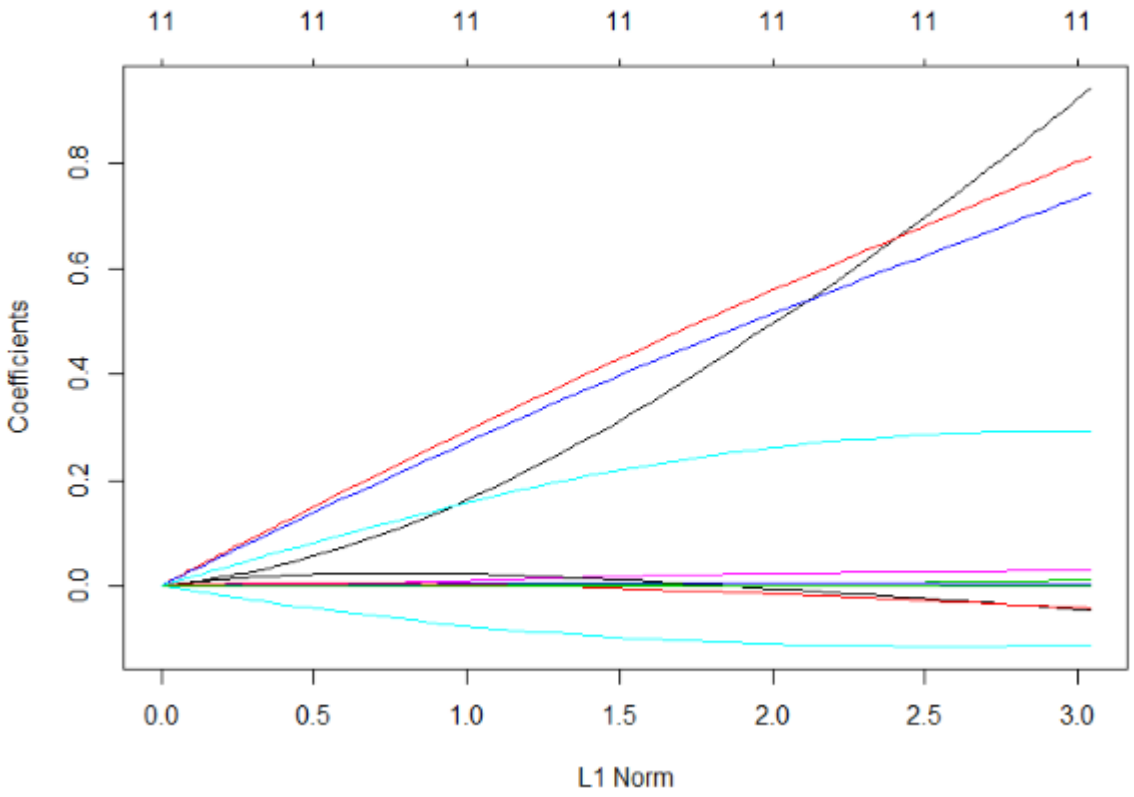
5.2 Ridge και Lasso στο μοντέλο του πειράματος

Σε αυτό το κεφάλαιο θα εφαρμόσουμε την τεχνική Ridge και την Lasso καθώς και θα επιλέξουμε σωστή ρυθμιστική παράμετρο για το μοντέλο μας.

Θα χρησιμοποιήσουμε το πακέτο glmnet και με τον εξής κώδικα θα πάρουμε τις γραφικές μας παραστάσεις.

```
> y<-Surv(t,c)
> x<-model.matrix(y~groupf+r.age+d.age+r.sex+d.sex+r.cmv+d.cmv+w.time+fab+mtx)
> fit3<-glmnet(x, y, family="cox")
> fit4<-glmnet(x,y,family="cox",alpha=0)
> plot(fit3)
> plot(fit4)
```

Αλλάζοντας την παράμετρο alpha=0 στον παραπάνω κώδικα κάνουμε την παλινδρόμηση κορυφογραμμής και παίρνουμε την γραφική παράσταση στο σχήμα 5.9:



Σχήμα 5.9: *Παλινδρόμηση κορυφογραμμής στο μοντέλο μας*

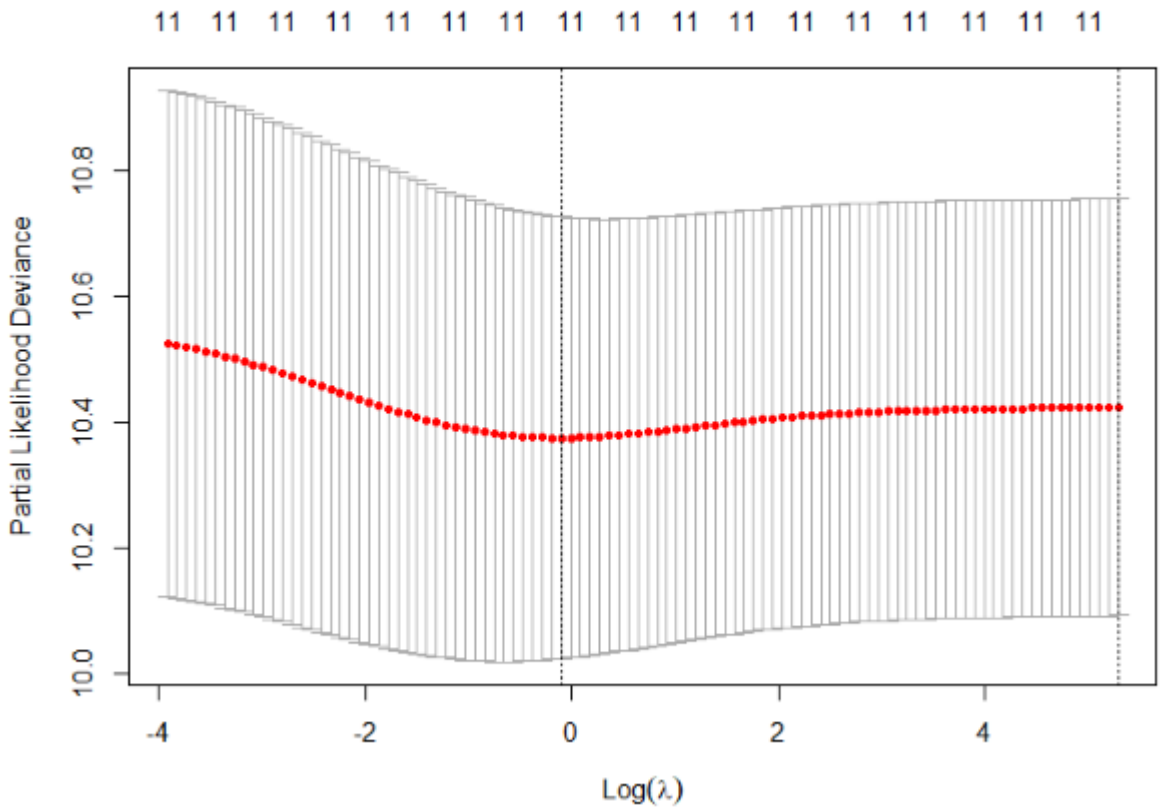
Παρατηρούμε από την γραφική παράσταση ότι οι συντελεστές των μεταβλητών δεν μηδενίζονται ποτέ εντελώς και επίσης παρατηρούμε ότι μόνο τρεις με τέσσερις μεταβλητές φαίνονται να επιδρούν σημαντικά στο μοντέλο μας.

Στην τεχνική Ridge οι μεταβλητές συρρικνώνονται αρκετά ώστε να μπορούμε να δούμε με αυτόν τον τρόπο ποιες είναι οι σημαντικές μεταβλητές στο μοντέλο μας.

Με τις παρακάτω εντολές επιλέγουμε την κατάλληλη ρυθμιστική παράμετρο για την παλινδρόμηση κορυφογραμμής και παίρνουμε τους συντελεστές τις παλινδρόμησης μας:

```
> cv_fit4 <- cv.glmnet(x, y,family="cox", alpha = 0)
> cv_fit4$lambda.min
[1] 0.9033998
```

Για την επιλογή του λ βλέπουμε και την γραφική παράσταση:



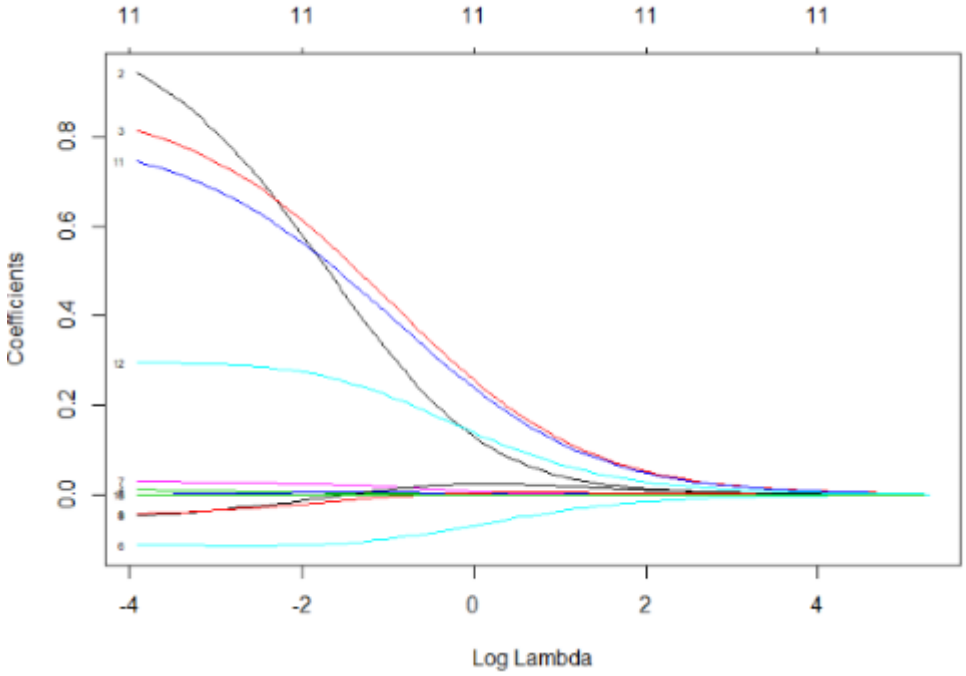
Σχήμα 5.10: *Επιλογή ρυθμιστικής παραμέτρου λ για το μοντέλο μας με την τεχνική Ridge*

Η γραφική παράσταση μας βοηθάει να δούμε ποια είναι η κατάλληλη ρυθμιστική παράμετρος. Η παραπάνω γραφική παράσταση γίνεται με την χρήση της εντολής plot του cvfit4.

Με την παρακάτω εντολή επιλέγοντας το λ που πήραμε από το παραπάνω κώδικα βλέπουμε ποιοι είναι οι συντελεστές τις παλινδρόμησης μας στην τεχνική Ridge.

```
> coef(fit4, s=0.9033998)
12 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept)  .
groupf1      0.144533230
groupf3      0.273626486
r.age        0.002212498
d.age        0.003258500
r.sex       -0.072878717
d.sex        0.007859781
r.cmv        0.023001120
d.cmv        0.002590811
w.time      -0.000035521
fab          0.254368561
mtx          0.147212869
```

Χρησιμοποιώντας στην εντολή plot() την παράμετρο xvar="lambda" παίρνουμε την γραφική παράσταση στο σχήμα 5.11. Παρατηρούμε ότι όλες οι μεταβλητές εκτός των **groupf1**, **groupf3**, **fab** **mtx** σχεδόν μηδενίζονται. Άρα καταλαβαίνουμε ότι οι μεταβλητές **groupf**, **fab** και λιγότερο η **mtx** είναι οι πιο σημαντικές για το μοντέλο μας.

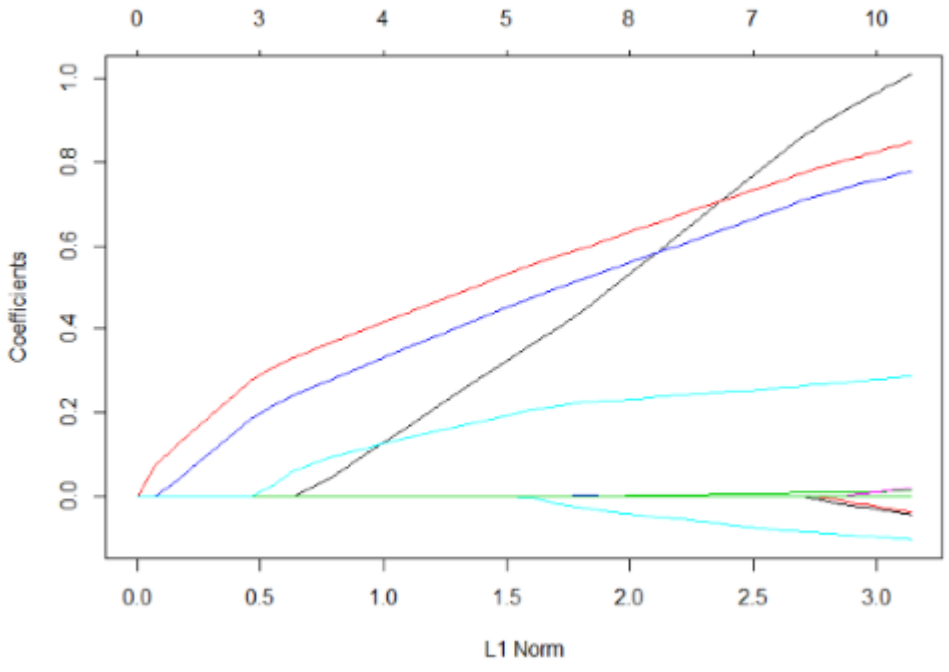


Σχήμα 5.11: *Γραφική παράσταση των συντελεστών παλινδρόμησης συναρτήσει του λογαρίθμου της ρυθμιστικής παραμέτρου λ στην παλινδρόμηση κορυφογραμμής*

Τέλος, με την εντολή print() βλέπουμε τις διαφορετικές τιμές του λ καθώς και το ποσοστό της απόκλισης :

```
> print(fit4)

Call:  glmnet(x = x, y = y, family = "cox", alpha = 0)

    Df %Dev  Lambda
1   11 0.00 199.200
2   11 0.02 181.500
3   11 0.02 165.400
4   11 0.03 150.700
5   11 0.03 137.300
6   11 0.03 125.100
7   11 0.03 114.000
8   11 0.04 103.900
9   11 0.04  94.640
10  11 0.04  86.230
11  11 0.05  78.570
12  11 0.05  71.590
13  11 0.06  65.230
14  11 0.06  59.440
```

Για την τεχνική Lasso, αρχικά με την εντολή plot του fit3 παίρνουμε την γραφική παράσταση:



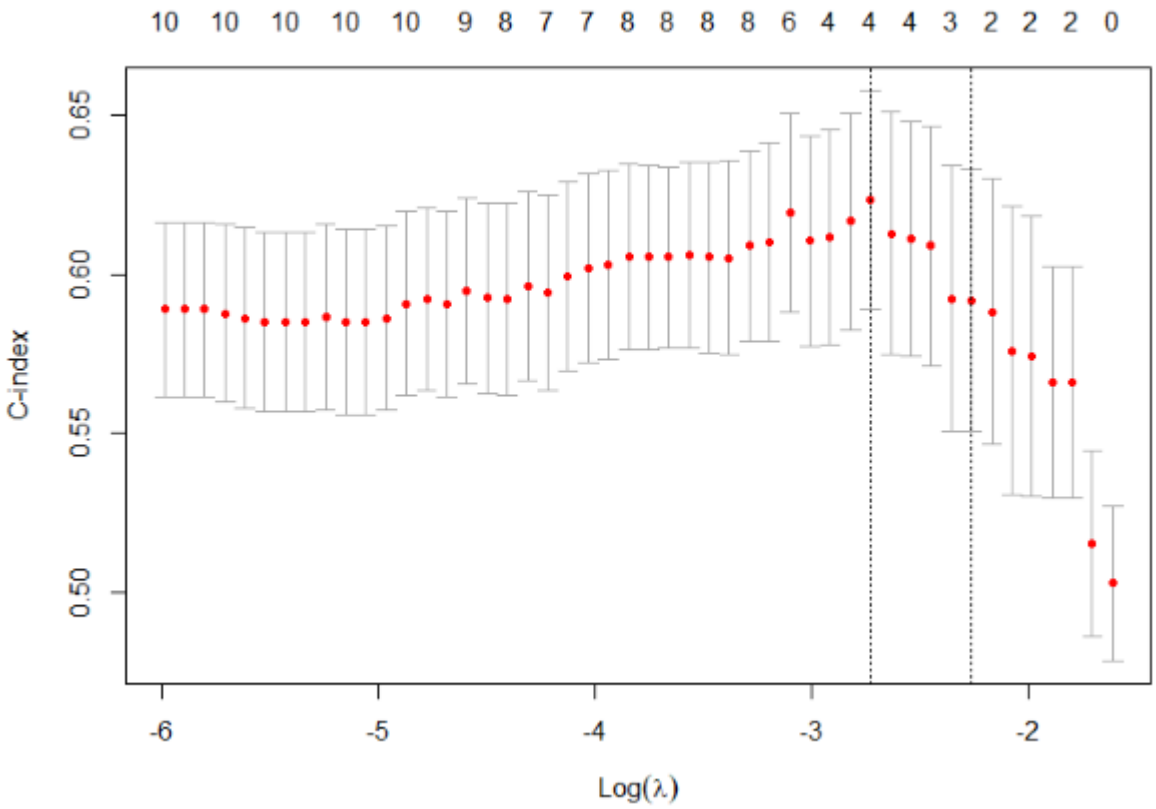
Σχήμα 5.12: *Τεχνική Lasso στο μουτέλο μας*

Στο σχήμα 5.12 κάθε καμπύλη ανταποκρίνεται σε μια μεταβλητή του μοντέλου μας. Δείχνει το μονοπάτι από τον συντελεστή των μεταβλητών έναντι της νόρμας από το διάνυσμα των συντελεστών όσο η ρυθμιστική παράμετρος λ διαφοροποιείται. Παρατηρούμε ότι αρκετοί συντελεστές τείνουν να πάνε στο 0. Και μόνο τρεις με τέσσερις μένουν μετά την τεχνική.

Ομοίως με πριν επιλέγουμε την κατάλληλη ρυθμιστική παράμετρο για την τεχνική Lasso και παίρνουμε τους συντελεστές τις παλινδρόμησης μας:

```
> cvfit3 <- cv.glmnet(x, y, family = "cox", type.measure = "C")
> cvfit3$lambda.min
[1] 0.06523278
```

Το plot του cvfit μας βοηθάει με την επιλογή της παραμέτρου λ για να καταλήξουμε σε ποιες μεταβλητές θα μείνουν τελικά στο μοντέλο μας στην περίπτωση της Lasso τεχνικής.

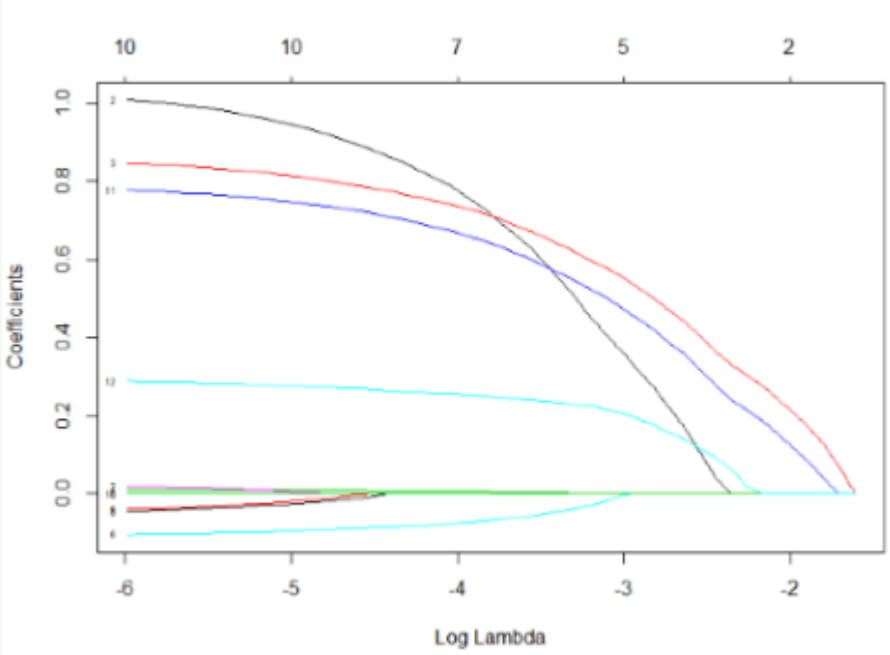


Σχήμα 5.13: *Επιλογή ρυθμιστικής παραμέτρου λ για το μοντέλο μας με την τεχνική Lasso*

Έπειτα, ομοιοτρόπως με πριν, θα δούμε ποιες μεταβλητές θα μείνουν στο μοντέλο μας καθώς κάποιες από αυτές θα μηδενιστούν. Αυτό γίνεται με τον εξής κώδικα:

```
> coef(fit3, s = 0.06523278)
12 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) .
groupf1     0.2214586
groupf3     0.4708893
r.age       .
d.age       .
r.sex       .
d.sex       .
r.cmv       .
d.cmv       .
w.time      .
fab         0.3890956
mtx         0.1585179
```

Χρησιμοποιώντας στην εντολή plot() την παράμετρο xvar="lambda" παίρνουμε την γραφική παράσταση στο σχήμα 5.14. Παρατηρούμε ότι οι σημαντικές μεταβλητές του μοντέλου μας είναι οι **groupf**, **fab** και λιγότερο η **mtx** όπως φάνηκε και στην προσαρμογή του μοντέλου του Cox με βήματα πιο πάνω.



Σχήμα 5.14: *Γραφική παράσταση των συντελεστών παλινδρόμησης συναρτήσει του λογαρίθμου της ρυθμιστικής παραμέτρου λ στην τεχνική Lasso*

Τέλος, με την εντολή print() βλέπουμε τις διαφορετικές τιμές του λ και το πώς αλλάζουν οι βαθμοί ελευθερίας του μοντέλου μας καθώς και το ποσοστό της απόκλισης:

```
> print(fit3)

Call:  glmnet(x = x, y = y, family = "cox")

   Df %Dev   Lambda
1   0 0.00 0.199200
2   1 0.24 0.181500
3   2 0.52 0.165400
4   2 0.77 0.150700
5   2 0.97 0.137300
6   2 1.13 0.125100
7   2 1.26 0.114000
8   3 1.41 0.103900
9   3 1.57 0.094640
10  4 1.80 0.086230
11  4 2.03 0.078570
12  4 2.23 0.071590
13  4 2.40 0.065230
14  4 2.54 0.059440
15  4 2.66 0.054160
16  5 2.76 0.049350
17  6 2.86 0.044960
18  7 2.96 0.040970
19  8 3.06 0.037330
20  8 3.14 0.034010
21  8 3.21 0.030990
22  8 3.27 0.028240
23  8 3.32 0.025730
24  8 3.36 0.023440
25  8 3.39 0.021360
26  8 3.42 0.019460
27  7 3.45 0.017730
28  7 3.46 0.016160
29  7 3.48 0.014720
30  7 3.49 0.013420
31  8 3.51 0.012220
32  8 3.52 0.011140
33  9 3.53 0.010150
34  9 3.54 0.009247
35  9 3.54 0.008425
36 10 3.55 0.007677
37 10 3.56 0.006995
38 10 3.56 0.006373
39 10 3.56 0.005807
40 10 3.57 0.005291
41 10 3.57 0.004821
42 10 3.57 0.004393
43 10 3.57 0.004003
44 10 3.58 0.003647
45 10 3.58 0.003323
46 10 3.58 0.003028
47 10 3.58 0.002759
48 10 3.58 0.002514
```

Η τεχνική Lasso τείνει να είναι χρήσιμη όταν υπάρχει μικρός αριθμός σημαντικών παραμέτρων ενώ η τεχνική Ridge όταν υπάρχει μεγάλος αριθμός παραμέτρων περίπου ίδιας τιμής.

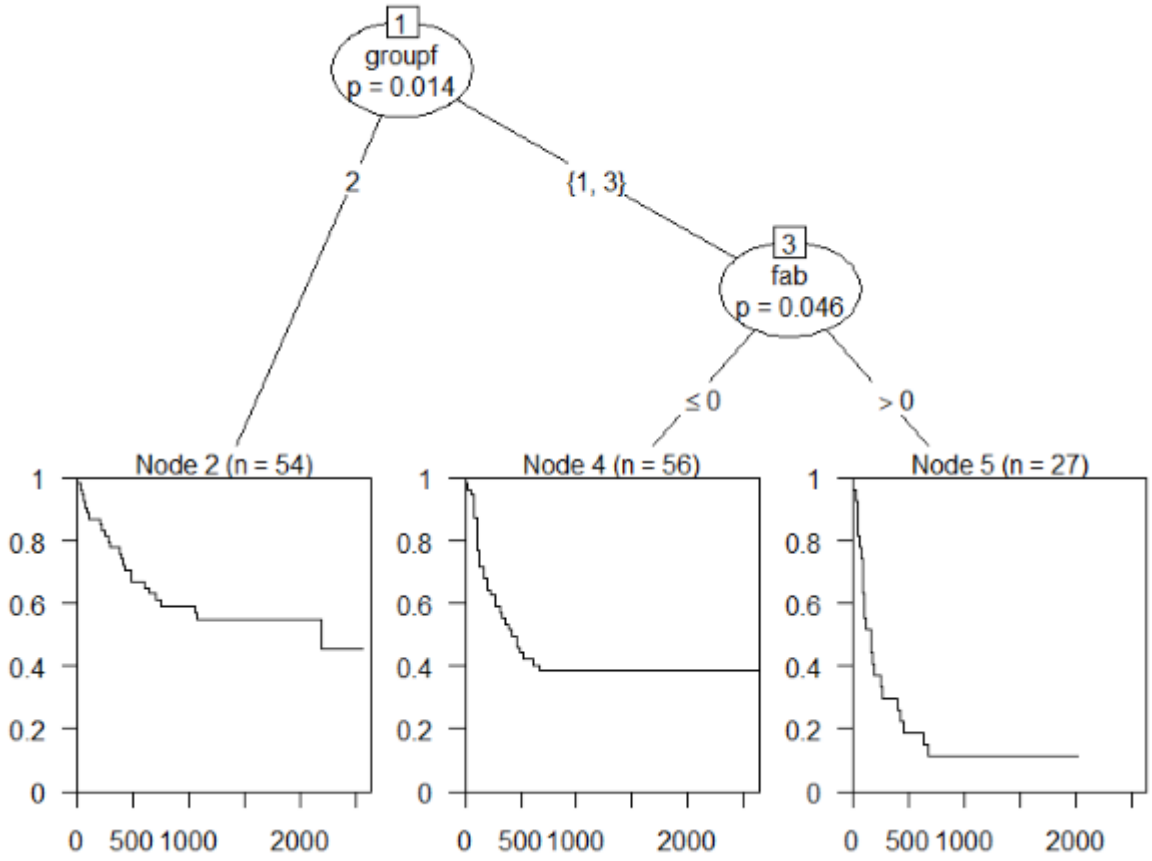
5.3 Survival Tree στο μοντέλο του πειράματος

Στο τελευταίο κομμάτι της ανάλυσης μας θα φτιάξουμε ένα δέντρο επιβίωσης για το μοντέλο μας. Θα χρησιμοποιήσουμε το πακέτο party και με τον εξής κώδικα θα καταλήξουμε σε ένα δέντρο επιβίωσης με τερματικά φύλλα την συνάρτηση επιβίωσης σε κάθε υποχώρο.

Αυτό γίνεται με τον εξής κώδικα:

```
> surv.tree<-ctree(Surv(t,c)~groupf + r.age + d.age + r.sex + d.sex + r.cmv
 + d.cmv + w.time + fab +mtx, data=cc)
> plot(surv.tree)
```

και παίρνουμε το παρακάτω δέντρο επιβίωσης:



Σχήμα 5.15: *Δέντρο επιβίωσης του μοντέλου μας*

Παρατηρούμε ότι και στο δέντρο επιβίωσης καταλήγουμε στο μοντέλο που καταλήξαμε και με την διαδοχική αφαίρεση. Και βλέπουμε την πιθανότητα επιβίωσης για τα διαφορετικά nodes. Βλέπουμε ότι στο Node 5 έχουμε την μικρότερη διάρκεια ζωής και μικρότερη πιθανότητα επιβίωσης. Αυτά είναι τα άτομα που ανήκουν στην ομάδα 1 (ALL) ή 3 (AML high-risk) και έχουν fab βαθμού 4 ή 5. Βλέπουμε ότι τα άτομα που δεν έχουν fab βαθμού 4 ή 5, στο Node 4 έχουν μεγαλύτερη διάρκεια ζωής. Τέλος στο Node 2 δηλαδή στα άτομα της ομάδας 2 (AML low-risk) έχουν την μεγαλύτερη διάρκεια ζωής και πιθανότητα επιβίωσης κάτι που διαπιστώθηκε και στα προηγούμενα.

Μέρος III

Επίλογος

Κεφάλαιο 6

Συμπεράσματα της διπλωματικής εργασίας

6.1 Συμπεράσματα

Στα προηγούμενα κεφάλαια, κάναμε χρήση πολλών τεχνικών. Αρχικά κάναμε την βασική ανάλυση των δεδομένων μας και προσαρμόσαμε το μοντέλο του Cox. Στη συνέχεια εφαρμόσαμε την παλινδρόμηση κορυφογραμμής, την τεχνική Lasso καθώς και κατασκευάσαμε ένα δέντρο επιβίωσης (Survival Tree).

Αρχικά παρατηρήσαμε στην παράγραφο 5.1 ότι με την διαδοχική αφαίρεση μεταβλητών στο μοντέλο του Cox καταλήξαμε στις συμμεταβλητές της ομάδας (groupf) και του fab καθώς το μοντέλο μόνο με αυτές έχει το καλύτερο AIC.

Στη παράγραφο 5.2 είδαμε για τις διάφορες τιμές τις ρυθμιστικής παραμέτρου λ πόσες συμμεταβλητές επιδρούν σημαντικά καθώς και παρατηρήσαμε στην περίπτωση της τεχνικής Lasso ότι μηδενίζονται όλοι οι συντελεστές για την καλύτερη τιμή του λ εκτός των συντελεστών του groupf, fab και mtx. Αυτό ενισχύεται και από τα σχήματα 5.11 και 5.14.

Στη παράγραφο 5.3 χρησιμοποιώντας τον αλγόριθμο από το πακέτο party της R βγαίνουμε στο ίδιο συμπέρασμα με την παράγραφο 5.1 καθώς το δέντρο επιβίωσης μας καταλήγει σε ένα μοντέλο που έχει δύο σημαντικές συμμεταβλητές την ομάδα (groupf) και την fab.

Συμπεραίνουμε όπως είχαμε πει και στην εισαγωγή ότι η χρήση πολλών τεχνικών για τα δεδομένα οδηγεί σε μια πιο εμπεριστατωμένη ανάλυση των δεδομένων μας. Η χρήση του δέντρου επιβίωσης (Survival Tree) μας δίνει και οπτικά μια καλύτερη εικόνα για το μοντέλο μας.

Το τελικό μοντέλο σε ένα μεγάλο ποσοστό χρησιμοποιώντας τις τεχνικές που αναφέραμε είναι το ίδιο. Δηλαδή το μοντέλο με τις συμμεταβλητές groupf και fab. Με τις τεχνικές Ridge και Lasso υπήρξε άλλη μια μεταβλητή που είναι σημαντική η mtx. Άρα οι τεχνικές που χρησιμοποιήσαμε μας βοηθήσαν να έχουμε μια πιο ολοκληρωμένη εικόνα και ίσως να εξετάσουμε και το πόσο επιδρά η μεταβλητή mtx τελικά στο μοντέλο μας.

6.2 Μελλοντικές Επεκτάσεις

Στις μέρες μας όλο και πιο αποδοτικοί αλγόριθμοι και τεχνικές έρχονται στην επιφάνεια και αναβαθμίζονται καθώς η ανάγκη για την σωστή και γρήγορη ανάλυση δεδομένων είναι μεγάλη. Η ανταγωνιστικότητα της σύγχρονης εποχής καθώς και η εύκολη πρόσβαση σε ειδική γνώση έχουν συνδράμει καταλυτικά.

Στόχος αυτής της εργασίας ήταν να χρησιμοποιήσουμε κάποιες σύγχρονες τεχνικές και να δούμε αν θα συνάδουν με τις παλιές και πιο χρησιμοποιημένες για πολλά χρόνια τεχνικές.

Η τεχνική Lasso, για παράδειγμα, φαίνεται σαν ένας άξιος «αντίπαλος» στην μέθοδο διαδοχικής αφαίρεσης μεταβλητών καθώς επιτυγχάνει σε πολλές περιπτώσεις τα ίδια αποτελέσματα και καταλήγει σε μοντέλα που μπορούν να ερμηνευθούν (Tibshirani, 1996). Στην πράξη η τεχνική Lasso όπως είδαμε και στην παρούσα διπλωματική εργασία πρέπει να χρησιμοποιείται ως συνδυασμός με άλλα εργαλεία κατασκευής μοντέλου. Επιπλέον οι τεχνικές Lasso και Ridge δίνουν λύσεις στο πρόβλημα της πολυσυγγραμικότητας.

Τα δέντρα επιβίωσης υπήρξαν και συνεχίζουν να είναι μια ενεργή «περιοχή» έρευνας. Έχουν προταθεί πολλές τεχνικές για την κατασκευή τους τα τελευταία 25 χρόνια (Bou-Hamad et al. , 2011). Όσον αφορά τα δέντρα επιβίωσης υπάρχει μια δυσκολία όταν υπάρχουν χρονικά μεταβαλλόμενες επιδράσεις. Σε αυτές τις περιπτώσεις χρειάζεται μελλοντική έρευνα καθώς δεν μπορούν να ερμηνευθούν οι «χαμένες» τιμές των χρονικά μεταβαλλόμενων μεταβλητών. Στην πράξη τα δέντρα επιβίωσης πρέπει να χρησιμοποιούνται ως συνδυασμός με μοντέλα που μπορούν να ερμηνευθούν (συνήθως παραμετρικά).

Εν κατακλείδι, καταλαβαίνουμε ότι οι περισσότερες τεχνικές δρουν ως συνδυασμός των ήδη υπαρχόντων τεχνικών και δεν μπορούν να υπάρξουν μεμονωμένα. Η αναζήτηση καλύτερων τεχνικών ή η αναβάθμιση των τεχνικών που υπάρχουν ήδη θα συνδράμουν στην καλύτερη ανάλυση των δεδομένων. Η ανάλυση επιβίωσης είναι πολύ σημαντική στις μέρες μας και η πληθώρα των ερευνών τα τελευταία χρόνια από τον επιστημονικό τομέα θα οδηγήσουν σε όλο καλύτερες τεχνικές και αλγόριθμους για να μπορούμε να αντιμετωπίσουμε κάθε πρόβλημα που μπορεί να προκύψει.

Βιβλιογραφία

[1] Χ. Καρώνη και Π. Οικονόμου. *Στατιστικά Μοντέλα Παλινδρόμησης*. Συμεών, Αθήνα, 2η έκδοση, 2017.

[2] Χ. Καρώνη. *Μοντέλα Αξιοπιστίας και Επιβίωσης*. Συμεών, Αθήνα, 2009.

[3] Α. Λιαπάτη. *Λείανση Επιφανειών ROC με Χρήση Πυρήνων*. Διπλωματική εργασία, Πανεπιστήμιο Αιγαίου, 2008.

[4] Γ.Α. Σταθόπουλος. *Στατιστικές Τεχνικές Παλινδρόμησης για την Ανάλυση Μεγάλων Δεδομένων*. Μεταπτυχιακή διπλωματική εργασία, Πανεπιστήμιο Πειραιώς, 2017.

[5] Δ. Φουσκάκης. *Ανάλυση Δεδομένων με Χρήση της R*. Τσότρας, Αθήνα, 2013.

[6] I. Bou-Hamad, D. Larocque and H. Ben-Ameur. *A review of survival trees*. *Statistics Surveys*, 5, 2011.

[7] S. Chand. *On tuning parameter selection of lasso-type methods - a monte carlo study*. *Proceedings of 2012 9th International Bhurban Conference on Applied Sciences Technology (IBCAST)*, σελίδες 120–129, 2012.

[8] D. Collet. *Modeling Survival Data in Medical Research*. Taylor and Francis, London, 2η έκδοση, 2003.

[9] D.R. Cox. *Regression Models and Life-Tables*. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[10] D.R. Cox and E.J. Snell. *A General Definition of Residuals*. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(2):248–275, 1968.

[11] Y. Fan and C.Y. Tang. *Tuning parameter selection in high dimensional penalized likelihood*. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):531–552, 2013.

[12] J. Friedman, T. Hastie, R. Tibshirani, B. Narasimhan, K. Tay, N. Simon and J. Qian (2021). *"glmnet ": Lasso and Elastic-Net Regularized Generalized Linear Models*. R package version 4.1-2. Available online at: https://glmnet.stanford.edu.

[13] J.A. Hanley and B. Mcneil. *The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve*. *Radiology*, 143:29–36, 1982.

[14] P.J. Heagerty (2015). *"risksetROC ": Riskset ROC curve estimation from censored survival data.* R package version 1.0.4. Available online at: https://cran.r-project.org/web/packages/risksetROC/risksetROC.pdf.

[15] A.E. Hoerl and R.W. Kennard. *Ridge Regression: Biased Estimation for Nonorthogonal Problems*. *Technometrics*, 12(1):55–67, 1970.

[16] T. Hothorn, K. Hornik, C. Strobl, A. Zeileis (2021). *"party ": A Laboratory for Recursive Partytioning*. R package version 1.3-7. Available online at: http://party.R-forge.R-project.org.

[17] A. Jacobson, V. Wilson and S. Pillegi (2018). *"parmsurvfit ": Parametric Models for Survival Data.* R package version 0.1.0. Available online at: https://github.com/apjacobson/parmsurvfit.

[18] G. James, D. Witten , T. Hastie and R. Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, New York, 8η έκδοση, 2013.

[19] E.L. Kaplan and P. Meier. *Nonparametric Estimation from Incomplete Observations*. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[20] M. LeBlanc and J. Crowley. *Relative Risk Trees for Censored Survival Data*. *Biometrics*, 48(2):411–425, 1992.

[21] J. Lederer and C. Müller. *Don't Fall for Tuning Parameters: Tuning-Free Variable Selection in High Dimensions With the TREX*. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), 2015.

[22] L.B. Lusted. *Signal Detectability and Medical Decision-Making*. *Science*, 171(3977):1217–1219, 1971.

[23] N. Mantel and W. Haenszel. *Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease*. *JNCI: Journal of the National Cancer Institute*, 22(4):719–748, 1959.

[24] S. Noah, F. Jerome, T. Hastie and R. Tibshirani. *Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent*. *Journal of Statistical Software*, 39, 2011.

[25] D. Sarkar, F. Andrews, K. Wright, N. Klepeis, J. Larsson and P. Murrell (2021). *"lattice ": Trellis Graphics for R.* R package version 0.20-44. Available online at: http://lattice.r-forge.r-project.org/.

[26] D. Schoenfeld. *Partial residuals for the proportional hazards regression model*. *Biometrika*, 69(1):239–241, 1982.

[27] M. Schumacher, N. Holländer, G. Schwarzer, H. Binder and W. Sauerbrei. *Prognostic Factor Studies*. In J. Crowley, A. Hoering, eds, Handbook of Statistics in Clinical Oncology, Chapman and Hall/CRC, 3rd edition, 2012.

[28] M.R. Segal, S.T. Weiss, F.E. Speizer and I.B. Tager. *Smoothing methods for epidemiologic analysis*. *Statistics in Medicine*, 7(5):601–611, 1988.

[29] T.M Therneau, T. Lumley, E. Atkinson and C. Crowson (2021). *"survival ": Survival analysis.* R package version 3.2-11. Available online at: https://github.com/therneau/survival.

[30] T.M. Therneau, P. M. Grambsch and T. R. Fleming. *Martingale-Based Residuals for Survival Models*. *Biometrika*, 77(1):147–160, 1990.

[31] R. Tibshirani. *Regression Shrinkage and Selection via the Lasso*. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

Παράρτημα

Συντομογραφίες - Αρκτικόλεξα - Ακρωνύμια

Ε.Μ.Φ.Ε	Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών
κ.λπ.	και λοιπά
κ.ο.κ	και ούτω καθεξής
Ε.Μ.Π	Εθνικό Μετσόβιο Πολυτεχνείο
AIC	Akaike information criterion
et al.	et alia - and others
RSS	Residual Sum of Squares
ROC	Receiver Operating Characteristic
AUC	Area Under the ROC Curve
MSPE	Mean Squared Prediction error
BIC	Bayesian Information Criterion