ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΔΠΜΣ: Μαθηματική Προτυποποίηση σε Σύγχρονες Τεχνολογίες και τη Χρηματοοικονομική

Κατεύθυνση: Μαθηματικά Επιστήμης Δεδομένων

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Παλινδρόμηση, Τεχνικές Συρρίκνωσης και Δέντρα στην Ανάλυση Επιβίωσης

Νοταρά Σοφία

Επιβλέπουσα: Καρώνη Χρυσηίς,
Καθηγήτρια Ε.Μ.Π

Τριμελής επιτροπή:

Χ. Καρώνη, Καθηγήτρια Ε.Μ.Π	Β. Παπανικολάου, Καθηγητής Ε.Μ.Π	Κ. Παυλοπούλου, Ε.ΔΙ.Π. Ε.Μ.Π

ΑΘΗΝΑ, ΙΟΥΝΙΟΣ 2022

Πρόλογος

Η παρούσα μεταπτυχιακή εργασία εκπονήθηκε κατά το ακαδημαϊκό έτος 2021-2022 στον τομέα Μαθηματικών της σχολής Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών του Εθνικού Μετσόβιου Πολυτεχνείου στα πλαίσια του Διεπιστημονικού Προγράμματος Μεταπτυχιακών Σπουδών «Μαθηματική Προτυποποίηση σε Σύγχρονες Τεχνολογίες και τη Χρηματοοικονομική».

Πρωτίστως, θα ήθελα να ευχαριστήσω θερμά την καθηγήτρια του Ε.Μ.Π. κυρία Χρυσηίδα Καρώνη στην οποία οφείλω την οικοδόμηση και ολοκλήρωση της συγκεκριμένης διπλωματικής εργασίας. Με την καθοδήγηση της διευκόλυνε το έργο μου, με παρότρυνε να εμβαθύνω στην εργασία μου και μου διόρθωσε με επιμέλεια τα λάθη μου.

Παράλληλα θα ήθελα να ευχαριστήσω την οικογένεια μου τόσο για την υλική όσο και για την πνευματική βοήθεια που μου προσέφερε όχι μόνο κατά τη διάρκεια της εκπόνησης της μεταπτυχιακής εργασίας αλλά κυρίως κατά τη διάρκεια τη φοίτησης μου στο Εθνικό Μετσόβιο Πολυτεχνείο.

Αθήνα, Ιούνιος 2022

Σοφία Μ. Νοταρά

Περίληψη

Σκοπός της παρούσας εργασίας είναι η στατιστική ανάλυση δεδομένων διάρκειας ζωής με χρήση διάφορων μεθόδων και μοντέλων. Η ανάλυση επιβίωσης και αξιοπιστίας είναι ο κλάδος της Στατιστικής με τον οποίο ασχοληθήκαμε και έχει εφαρμογή σε βιοϊατρικές αλλά και τεχνολογικές επιστήμες.

Πιο αναλυτικά, στο πρώτο κεφάλαιο δίνονται οι βασικοί ορισμοί της ανάλυσης επιβίωσης καθώς και οι ορισμοί που αφορούν αποκομμένα δεδομένα. Γίνονται αναφορές στη συνάρτηση επιβίωσης, τη συνάρτηση διακινδύνευσης, τη σωρευτική συνάρτηση διακινδύνευσης κ.τ.λ. Ακόμα περιγράφονται κάποιες μη-παραμετρικές τεχνικές που είναι πολύ χρήσιμες στην Ανάλυση Επιβίωσης όπως η εκτιμήτρια Kaplan-Meier, η εκτιμήτρια Nelson-Aalen, ο έλεγχος log-rank και ο έλεγχος του Wilcoxon.

Στο δεύτερο κεφάλαιο παρουσιάζονται αναλυτικά τα Παραμετρικά Μοντέλα Αναλογικής Διακινδύνευσης και γίνεται εκτενής ανάλυση του ημι-παραμετρικού μοντέλου του Cox. Το μοντέλο του Cox μοντελοποιεί τη συνάρτηση διακινδύνευσης σε σχέση με άλλες μεταβλητές. Η εκτίμηση των συντελεστών παλινδρόμησης επιτυγχάνεται μέσω της συνάρτησης μερικής πιθανοφάνειας. Επιπλέον περιγράφονται μέθοδοι που εξετάζουν αν ισχύει η υπόθεση αναλογικότητας των κινδύνων όπως και με τα υπόλοιπα, που χρησιμοποιούνται για διάφορους ελέγχους που αφορούν την καταλληλόλητα του μοντέλου.

Το τρίτο κεφάλαιο πραγματεύεται τις μεθόδους ποινής όταν στα δεδομένα εμφανίζεται το φαινόμενο της πολυσυγγραμμικότητας. Οι μέθοδοι ποινής που αναλύονται είναι οι: Ridge, Lasso και Elastic Net. Τέλος, παρουσιάζεται η μέθοδος Cross Validation για την εύρεση του βέλτιστου λ που χρησιμοποιείται στις παραπάνω μεθόδους.

Στο τέταρτο κεφάλαιο παρουσιάζονται τα δέντρα αποφάσεων για δεδομένα διάρκειας ζωής. Τα δέντρα αποφάσεων αποτελούν μία από τις πιο διαδομένες μεθόδους ταξινόμησης διότι είναι εύκολα στην κατανόηση.

Στο πέμπτο και τελευταίο κεφάλαιο της εργασίας γίνεται εφαρμογή των παραπάνω μεθόδων σε ένα δείγμα ασθενών που πάσχουν από πολλαπλό μυέλωμα. Σκοπός της ανάλυσης είναι η ανάδειξη των μεταβλητών που επηρεάζουν περισσότερο την πορεία της υγείας των ασθενών. Οι προς εξέταση ανεξάρτητες μεταβλητές είναι η ηλικία, το φύλο, τα επίπεδα αζώτου ουρίας αίματος, η ποσότητα του ασβεστίου, η τιμή της αιμοσφαιρίνης, το ποσοστό των καρκινικών κυττάρων και η ύπαρξη μίας συγκεκριμένης πρωτεΐνης στις εξετάσεις του ασθενή. Για την ανάλυση των δεδομένων χρησιμοποιήθηκε το στατιστικό πακέτο της R.

Abstract

The purpose of this work is the statistical analysis of lifetime data using various methods and models. Survival and reliability analysis is the field of Statistics that we are dealing with, which has application in the biomedical and technological sciences.

In more detail, the first chapter gives the basic definitions of survival analysis as well as the definition of censored data. Reference is made to the survival function, the hazard function, the cumulative hazard function, etc. Some non-parametric techniques that are very useful in Survival Analysis such as the Kaplan-Meier estimator, the Nelson-Aalen estimator, the log-rank test and the Wilcoxon test are also described.

The second chapter presents in detail parametric Proportional Hazards risk models and provides an extensive analysis of Cox's semi-parametric model. In the Cox model, which models the hazard function relative to other variables, the estimation of the regression coefficients is achieved through the partial likelihood. In addition, we describe methods that examine whether the proportional hazards assumption is valid, and also the residuals that are used for various checks on the suitability of the model.

The third chapter concerns penalised estimation methods when the phenomenon of multicollinearity appears in the data. The penalised methods analyzed are Ridge, Lasso and Elastic Net. Finally, the Cross Validation method is presented to find the optimal λ which is used in the above methods.

The fourth chapter presents decision trees for lifetime data. Decision trees are one of the most common classification methods because they are easy to understand.

In the fifth and last chapter of the work, the above methods are applied to a sample of patients suffering from multiple myeloma. The purpose of the analysis is to highlight the variables that most affect the course of the patient's health. The independent variables to be examined are age, sex, blood urea nitrogen, the amount of calcium, hemoglobin levels, cancer cell percentage, and the detection of a specific protein in the examination of the patient. The R statistical package was used for data analysis.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1. Εισαγωγή στην Ανάλυση Επιβίωσης και Αξιοπιστίας

1.1 Γενικά

Η ανάλυση επιβίωσης και αξιοπιστίας αφορά την μελέτη δεδομένων διάρκειας ζωής δηλαδή ασχολείται με τη μελέτη του χρόνου έως ότου προκύψει κάποιο γεγονός. Το γεγονός αυτό συνήθως είναι δυσάρεστο, όπως για παράδειγμα θάνατος ή υποτροπή ενός ασθενή, μηχανική βλάβη, χρεωκοπία μιας εταιρείας κ.τ.λ. Μπορεί, όμως, το γεγονός να μην είναι τόσο δυσάρεστο όπως η αποθεραπεία ενός ασθενή. Ο όρος Αξιοπιστία (Reliability) χρησιμοποιείται για δεδομένα που αφορούν συνήθως τη βιομηχανία (όπως για παράδειγμα την εμπλοκή ενός μηχανήματος) ή τη γεωργία (όπως την ανάλυση του χρόνου μέχρι να καρποφορήσει ένα δέντρο) ,ενώ ο όρος Επιβίωση (Survival) έχει να κάνει με βιοϊατρικές εφαρμογές.

Πατέρας της επιστήμης της Ανάλυσης Επιβίωσης μπορεί να θεωρηθεί ο Bernoulli (1700-1782), ο οποίος ιστορικά πρώτος χρησιμοποιεί δεδομένα θνησιμότητας από τον ιό της ευλογιάς για να αποδείξει την αποτελεσματικότητα των εμβολίων, ενώ στην συνέχεια και ο Graunt τον 17^{ο} αιώνα με την χρήση διαγραμμάτων ζωής προσπαθεί να κατανοήσει την διάρκεια της ανθρώπινης ζωής.

1.2 Αποκομμένα Δεδομένα

Ο χρόνος επιβίωσης είναι ιδιαίτερος διότι είναι περιορισμένος στο να είναι πάντα θετικός, και γιατί τα δεδομένα περιέχουν αποκομμένες (censored) παρατηρήσεις. Τα αποκομμένα δεδομένα είναι αυτά για τα οποία δεν είναι γνωστός ο χρόνος που συμβαίνει το γεγονός. Το μόνο που μπορούμε να πούμε είναι ότι ο χρόνος επιβίωσής τους είναι μεγαλύτερος από την τιμή που έχει καταγραφεί. Πολλές φορές, κατά την εκτέλεση ενός πειράματος στο οποίο καταγράφεται ο χρόνος λειτουργίας μέχρι να συμβεί ένα γεγονός, πολλές μονάδες συνεχίζουν να λειτουργούν και μετά τη λήξη του πειράματος. Αν και δεν είμαστε σε θέση να ξέρουμε πότε ακριβώς συνέβη το γεγονός στις εν λόγω μονάδες, μετά το τέλος της παρακολούθησης, ξέρουμε ότι μέχρι κάποια χρονική στιγμή ήταν ακόμα λειτουργικές. Αυτές οι πληροφορίες δεν είναι καθόλου ασήμαντες, για αυτό και δεν τις εξαιρούμε από την μελέτη μας. Τα δεδομένα που δεν είναι αποκομμένα ονομάζονται μη-αποκομμένα ή πλήρη δεδομένα. (Καρώνη, 2009)

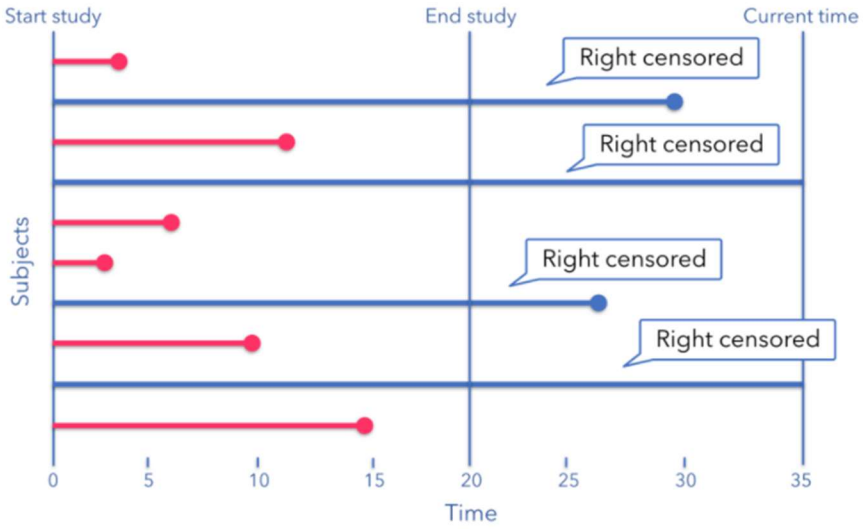
Η έννοια των αποκομμένων δεδομένων εισήχθη για πρώτη φορά το 1949 από τον Hald. Υπάρχουν δύο είδη αποκομμένων δεδομένων, τα αποκομμένα δεδομένα τύπου I και τα αποκομμένα δεδομένα τύπου II.

Η αποκοπή δεδομένων τύπου I αφορά την παρακολούθηση των μονάδων σε ένα συγκεκριμένο χρονικό διάστημα. Ο ερευνητής στην περίπτωση αυτή έχει προκαθορίσει ένα συγκεκριμένο ποσοστό επιτυχίας, και μόλις το πετύχει τερματίζει το πείραμα. Έτσι, οι χρόνοι διάρκειας ζωής των υπό εξέταση μονάδων είναι δεδομένοι, ενώ ο αριθμός των μονάδων που δεν επιβιώνει είναι τυχαίος.

Αντίθετα, όταν η παρακολούθηση του πειράματος διακόπτεται όταν καταστραφεί ένα συγκεκριμένο πλήθος μονάδων χωρίς όμως να γνωρίζουμε τη διάρκεια εκτέλεσης του πειράματος μπορούμε να πούμε ότι η αποκοπή δεδομένων είναι τύπου II.

Μία ακόμη περίπτωση στην οποία θα μπορούσαμε να πούμε ότι τα δεδομένα είναι αποκομμένα, είναι όταν η χρονική διάρκεια της έρευνας είναι προκαθορισμένη αλλά κάποιες μονάδες αποχωρούν από το πείραμα για άλλους λόγους. Ένα σχετικό παράδειγμα θα μπορούσε να ήταν η αποχώρηση ενός ασθενή από την έρευνα για προσωπικούς λόγους πριν αυτή τελειώσει.

Ένα παράδειγμα αποκομμένων δεδομένων δίνεται στο Διάγραμμα 1.1. Το πείραμα αφορούσε ασθενείς οι οποίοι έπασχαν από κάποια ασθένεια. Ο χρόνος επιβίωσης τους είναι το αντικείμενο της συγκεκριμένης μελέτης. Παρατηρούμε ότι ο 2^{ος}, ο 4^{ος}, ο 7^{ος} και ο 9^{ος} ασθενής απεβίωσαν αφότου η έρευνα είχε πραγματοποιηθεί. Αυτές οι μονάδες καλούνται δεξιά αποκομμένες παρατηρήσεις. Μάλιστα ο 4^{ος} και ο 9^{ος} ασθενείς όπως φαίνεται από την εικόνα δεν έχουν ακόμα φύγει από τη ζωή.



Διάγραμμα 1.1. Δεξιά αποκομμένα δεδομένα

Τέλος υπάρχουν και αριστερά αποκομμένες παρατηρήσεις, οι οποίες όμως είναι πιο σπάνιες και δεν θα ασχοληθούμε περεταίρω στη συγκεκριμένη εργασία. Το μόνο που γνωρίζουμε σε αυτό τον τύπο αποκοπής δεδομένων είναι ότι ο χρόνος επιβίωσης είναι μικρότερος από ένα χρονικό διάστημα. Ο ακριβής χρόνος επιβίωσης δεν είναι γνωστός.

1.3 Βασικές συναρτήσεις

1.3.1 Η συνάρτηση επιβίωσης ή αξιοπιστίας S(t)

Η συνάρτηση η οποία εκφράζει την πιθανότητα η διάρκεια ζωής T κάποιας τυχαίας μονάδας από το δείγμα, για παράδειγμα ασθενών, να ξεπερνάει το χρόνο t καλείται συνάρτηση επιβίωσης ή αξιοπιστίας (survival function) και συμβολίζεται με S(t). Ο τύπος της συνάρτησης αυτής δίνεται ως εξής:

$$S(t) = P[T > t] \tag{1.1}$$

Αν θεωρήσουμε ως F(t) τη συνάρτηση κατανομής πιθανότητας και f(t) τη συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής T>0 τότε η σχέση (1.1) μπορεί να γραφεί και ως:

$$S(t) = 1 - F(t) = \int_t^{\infty} f(u)du \tag{1.2}$$

Η εκτίμηση της συνάρτησης επιβίωσης αποτελεί τον σημαντικότερο σκοπό στην ανάλυση επιβίωσης και αξιοπιστίας.

1.3.2 Η συνάρτηση διακινδύνευσης h(t)

Η συνάρτηση Διακινδύνευσης (hazard function) εκφράζει τον κίνδυνο διακοπής μίας μονάδας στο χρονικό διάστημα (t, t + δt] με δεδομένο ότι έχει επιβιώσει η μονάδα μέχρι τη χρονική στιγμή t. Ο τύπος της συνάρτησης Διακινδύνευσης υπολογίζεται ως εξής:

$$h(t) = \lim_{\delta t \to 0} \left(\frac{P[t < T \leq t + \delta t \mid T > t]}{\delta t} \right) \tag{1.3}$$

Υπολογίζοντας την δεσμευμένη πιθανότητα έχουμε:

$$P[t < T \leq t + \delta t \mid T > t] = \frac{P[t < T \leq t + \delta t]}{P[\, T > t]} \simeq \frac{f(t)\delta t}{S(t)} \tag{1.4}$$

Τελικά η συνάρτηση Διακινδύνευσης δίνεται (συνδυάζοντας τις σχέσεις 1.3 και 1.4) από τον τύπο:

$$h(t) = \frac{f(t)}{S(t)} \tag{1.5}$$

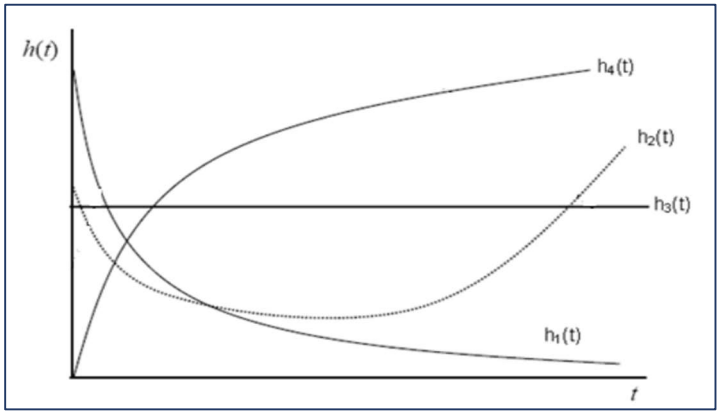
Η συνάρτηση h(t) μπορεί να παρατηρηθεί σε μία από τις παρακάτω μορφές οι οποίες συνοψίζονται και στο Διάγραμμα 1.2.

1) Η συνάρτηση h(t) να είναι φθίνουσα συνάρτηση διακινδύνευσης. Στην περίπτωση αυτή η στιγμιαία πιθανότητα να συμβεί το γεγονός μειώνεται με την πάροδο του χρόνου. Δεν είναι συνηθισμένη η μορφή αυτή της συνάρτησης διακινδύνευσης παρά μόνο σε λίγες περιπτώσεις. Για παράδειγμα θα μπορούσε να συμβεί σε ένα σύνολο ασθενών εάν η φαρμακευτική αγωγή που λάμβαναν βοηθούσε στην βελτίωση της υγείας τους. Στο Διάγραμμα 1.2 η συγκεκριμένη συνάρτηση διακινδύνευσης συμβολίζεται με $h_1(t)$.

2) Η συνάρτηση h(t) να είναι η διακινδύνευση μπανιέρας (bathtub hazard function). Η συνάρτηση αυτή έλαβε το όνομα της από τη γραφική της παράσταση. Στο πρώτο χρονικό διάστημα έχει μία φθίνουσα τάση και στη συνέχεια η γραφική σταθεροποιείται και καταλήγει σε ένα στάδιο αύξουσας διακινδύνευσης. Η περίπτωση αυτή είναι το πιο ρεαλιστικό μοντέλο διακινδύνευσης διότι είναι σαν να περιγράφει την ανθρώπινη ζωή. Στο Διάγραμμα 1.2 η συγκεκριμένη συνάρτηση διακινδύνευσης συμβολίζεται με $h_2(t)$.

3) Η συνάρτηση h(t) να είναι σταθερή, δηλαδή να μην εμφανίζει ούτε ανοδική ούτε πτωτική τάση. Στην περίπτωση αυτή η διακινδύνευση παραμένει σταθερή και έτσι η μονάδα φαίνεται να μην γερνά στο πέρασμα του χρόνου κάτι που αντικειμενικά δεν μπορεί να συμβεί. Παρόλα αυτά το συναντάμε για μικρά χρονικά διαστήματα σε μοντέλα όπως η εκθετική κατανομή. Η παραπάνω γραφική είναι η $h_3(t)$ στο Διάγραμμα 1.2.

4) Τέλος έχουμε την $h_4(t)$ η οποία είναι μία αύξουσα συνάρτηση διακινδύνευσης. Στην περίπτωση αυτή η διακινδύνευση αυξάνεται με την πάροδο του χρόνου. (Collett, 2003)



Διάγραμμα 1.2. Οι μορφές της συνάρτησης διακινδύνευση

1.3.3 Σωρευτική συνάρτηση διακινδύνευσης

Η σωρευτική συνάρτηση διακινδύνευσης δίνεται από τον τύπο:

$$H(t) = \int_0^t h(u)du \tag{1.6}$$

όπου h(u) είναι η συνάρτηση διακινδύνευσης. Κάνοντας αντικατάσταση τον ορισμό της συνάρτησης διακινδύνευσης από τη σχέση 1.5 στη σχέση 1.6 λαμβάνουμε τα ακόλουθα:

$$\begin{aligned} H(t) &= \int_0^t \frac{f(u)}{S(u)}du \\ &= \int_0^t \frac{-S'(u)}{S(u)}du \\ &= [-\ln S(u)]_0^t \\ &= -ln\, S(t) \end{aligned}$$

$$\Rightarrow H(t) = -ln\, S(t) \tag{1.7}$$

Τέλος, η συνάρτηση επιβίωσης μπορεί να δοθεί από την συνάρτηση σωρευτικής διακινδύνευσης από τον τύπο:

$$S(t) = exp\{-H(t)\} \tag{1.8}$$

1.4 Μη παραμετρική ανάλυση δεδομένων διάρκειας ζωής

Στην ανάλυση των δεδομένων διάρκειας ζωής ένας από μας στόχους που καλούμαστε να βγάλουμε εις πέρας είναι η εύρεση του μοντέλου που προσαρμόζεται καταλληλότερα σε αυτά. Για να επιλέξουμε το κατάλληλο μοντέλο μελετάμε τη συμπεριφορά των συναρτήσεων επιβίωσης και διακινδύνευσης. Σε αυτό μας βοηθούν οι εκτιμήτριες των συναρτήσεων αυτών, τις οποίες θα δούμε στη συνέχεια της παρούσας εργασίας. Οι μέθοδοι που θα χρησιμοποιήσουμε για να εκτιμήσουμε τις συναρτήσεις επιβίωσης και διακινδύνευσης ονομάζονται μη-παραμετρικές καθώς δεν χρειάζεται να γνωρίζουμε την κατανομή που ακολουθούν τα δεδομένα μας.

1.4.1 Η εκτιμήτρια Kaplan – Meier

Μία μη-παραμετρική μέθοδος που χρησιμοποιείται κατά κόρων όταν έχουμε δεξιά αποκομμένες παρατηρήσεις είναι αυτή των Kaplan – Meier. Η μέθοδος αυτή είναι ικανή να συμπεριλάβει τέτοιου είδους δεδομένα στην μελέτη και για αυτό το λόγο χρησιμοποιείται ευρέως.

Ας υποθέσουμε ότι έχουμε ένα τυχαίο δείγμα μεγέθους n, με μερικές από τις μονάδες του να καταστρέφονται κατά τις διακεκριμένες χρονικές στιγμές $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$ $με$ $k \leq n$. Ορίζουμε ως d_j το πλήθος των μονάδων στις οποίες συνέβη το γεγονός τη χρονική στιγμή $t_{(j)}$, $j = 1, .., k$ και ως $n_{(j)}$ τον αριθμό των μονάδων που ήταν σε κίνδυνο αμέσως πριν τη χρονική στιγμή $t_{(j)}$.

Η ποσότητα $n_{(j)}$ σχετίζεται με τις μονάδες που έχουν χρόνο ζωής μεγαλύτερο ή ίσο της j-οστής χρονικής στιγμής, ανεξαρτήτως της τελικής κατάληξής τους. Δηλαδή συμπεριλαμβάνονται όλες οι μονάδες στις οποίες πρόκειται να συμβεί το γεγονός και σε όλες εκείνες που δε θα συμβεί έως και το τέλος του πειράματος (αποκομμένες παρατηρήσεις). Εντούτοις, δεν υπολογίζονται στο $n_{(j)}$ όλες οι μονάδες στις οποίες συνέβη το γεγονός πριν τη στιγμή $t_{(j)}$ ούτε επίσης τις ήδη αποκομμένες παρατητήσεις.

Η εκτιμήτρια Kaplan – Meier τη χρονική στιγμή $t_{(j)}$ υπολογίζεται με την ακόλουθη διαδικασία.

$$S(t_{(j)}) = P(T > t_{(j)})$$
$$= P[\{T > t_{(1)}\} \cap \{T > t_{(2)}\} \cap \ldots \cap \{T > t_{(j)}\}]$$
$$= P[T > t_{(1)}]\, P[T > t_{(2)} | T > t_{(1)}] \ldots P[T > t_{(j)} | \{T > t_{(1)}\} \cap \ldots \cap \{T > t_{(j-1)}\}]$$

Τελικά έχουμε:

$$S(t_{(j)}) = P(T > t_{(1)}) P(T > t_{(2)} | T > t_{(1)}) \ldots P(T > t_{(j)} | T > t_{(j-1)}) \quad (1.9)$$

Μία εκτιμήτρια της $P(T > t_{(1)})$ δίνεται ως εξής:

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

όπου p_1 η σχετική συχνότητα των κατεστραμμένων μονάδων στο διάστημα (0, $t_{(1)}$].

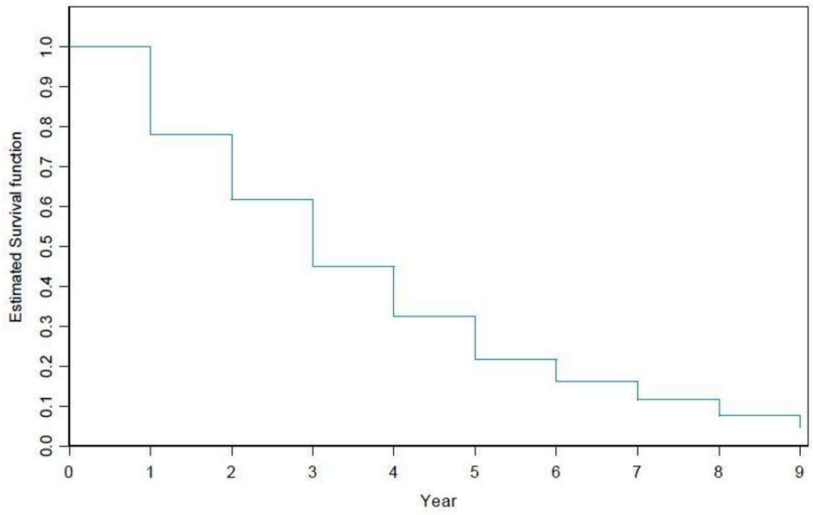
Επαναλαμβάνοντας την ίδια διαδικασία προκύπτει ότι

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

Τελικά προκύπτει η εκτιμήτρια Kaplan – Meier

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

Το διάγραμμα της εκτιμήτριας Kaplan-Meier συναρτήσει του χρόνου δεν είναι καμπύλη αλλά μία βαθμωτή συνάρτηση, όπου η $\hat{S}(t)$ θεωρείται σταθερή μεταξύ δύο γειτονικών χρονικών στιγμών και μειώνεται συνεχώς όπως φαίνεται στο Διάγραμμα 1.3.



Διάγραμμα 1.3: Εκτίμηση Kaplan-Meier

Έτσι για παράδειγμα, όταν η εκτίμηση της συνάρτησης επιβίωσης είναι ίση με 0,2 ο μέσος χρόνος επιβίωσης είναι περίπου 6 χρόνια, ενώ όταν $\hat{S}(t) = 0{,}7$ έχουμε χρονικό διάστημα επιβίωσης ίσο με 2 χρόνια. Από τη γραφική γίνεται φανερό ότι άτομα με μεγάλη εκτίμηση της συνάρτησης επιβίωσης έχουν μεγαλύτερο προσδόκιμο ζωής.

Τέλος, το τυπικό της σφάλμα υπολογίζεται από το τύπο του Greenwood για την εκτιμήτρια της διασποράς της $\hat{S}(t)$ που ορίζεται ως:

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

1.4.2 Η εκτιμήτρια της σωρευτικής συνάρτησης διακινδύνευσης Nelson-Aalen

Αφού έχουμε εκτιμήσει την συνάρτηση επιβίωσης S(t) από την εκτιμήτρια Kaplan-Meier μπορούμε να πάρουμε μια μη-παραμετρική εκτίμηση της σωρευτικής συνάρτησης διακινδύνευσης H(t). Η εκτίμηση αυτή μπορεί να γίνει από τον τύπο

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

Καταλήγουμε στο ότι η εκτιμήτρια της H(t) είναι η εκτιμήτρια Nelson-Aalen (Nelson, 1972, Aalen, 1978).

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

Οπότε και αυτή η συνάρτηση είναι κλιμακωτή.

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

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

Οπότε και το τυπικό σφάλμα έχει ως εξής:

$$se\left(\hat{S}(t)\right) = \left\{\hat{V}\left(\hat{S}(t)\right)\right\}^{1/2}$$

1.4.3 Έλεγχος log-rank

Ο έλεγχος Log-rank είναι ένας μη-παραμετρικός έλεγχος που χρησιμοποιείται για να συγκρίνουμε δύο ή και παραπάνω ομάδες στις οποίες έχουμε χωρίσει τα δεδομένα μας. Καλείται μη-παραμετρικός διότι δεν γνωρίζουμε τις συναρτήσεις επιβίωσης των ομάδων των δεδομένων μας, αλλά ελέγχουμε την ισότητα των συναρτήσεων επιβίωσης που έχουμε.

Έστω ότι έχουμε $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$ διακεκριμένες χρονικές στιγμές και έχουμε χωρίσει τα δεδομένα μας σε δύο ομάδες (ομάδα 1 και ομάδα 2). Σε αυτές τις χρονικές στιγμές παύουν να λειτουργούν οι παρατηρήσεις που προέρχονται και από τις δύο ομάδες. Θεωρούμε ότι αμέσως πριν τη χρονική στιγμή t_j για την ομάδα 1 υπάρχουν n_{1j} παρατηρήσεις σε κίνδυνο από τις οποίες παύουν να λειτουργούν d_{1j} μονάδες ακριβώς τη στιγμή $t_{(j)}$. Αντίθετα, για την ομάδα 2 αμέσως πριν τη χρονική στιγμή $t_{(j)}$ υπάρχουν n_{2j} παρατηρήσεις σε κίνδυνο από τις οποίες παύουν να λειτουργούν d_{2j} παρατηρήσεις ακριβώς τη στιγμή $t_{(j)}$. Επομένως, τη χρονική στιγμή $t_{(j)}$ παύουν να λειτουργούν συνολικά $d_j = d_{1j} + d_{2j}$ παρατηρήσεις από τις $n_j = n_{1j} + n_{2j}$ που ήταν σε κίνδυνο. Αυτό φαίνεται πιο αναλυτικά στον παρακάτω 2 x 2 πίνακα συνάφειας, ο οποίος είναι ο Πίνακας 1.1.

		Ομάδα		
		A	***B***	***Άθροισμα***
Διακοπή Λειτουργίας	***Ναι***	d_{1j}	d_{2j}	d_j
	Όχι	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
	Άθροισμα	n_{1j}	n_{2j}	n_j

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

Ο έλεγχος Log-rank που πραγματοποιούμε έχει σαν μηδενική υπόθεση την $H_0: S_A = S_B$ με εναλλακτική την $H_1: S_A \neq S_B$, όπου S_A, S_B οι συναρτήσεις επιβίωσης των ομάδων Α και Β αντίστοιχα. Δηλαδή ελέγχουμε αν υπάρχουν διαφοροποιήσεις μεταξύ των δύο ομάδων όσον αφορά την επιβίωση των μονάδων που ανήκουν στις ομάδες αυτές. Αποδεικνύεται ότι η ελεγχοσυνάρτηση Log-rank δίνεται από τον τύπο:

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

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

1.4.4 Έλεγχος Wilcoxon

Μια επέκταση της log-rank ελεγχοσυνάρτησης είναι η ελεγχοσυνάρτηση Wilcoxon. Αρχικά η ελεγχοσυνάρτηση log-rank γενικεύεται ως $\frac{(\sum w_j u_j)^2}{\sum {w_j}^2 v_j}$, με w_j οι συντελεστές στάθμισης. Στον έλεγχο log-rank το w_j λαμβάνει την τιμή 1, ενώ αν χρησιμοποιήσουμε τον έλεγχο του Wilcoxon θα έχουμε $w_j = n_j$. Οπότε στον έλεγχο του Wilcoxon ο συντελεστής στάθμισης είναι ίσος με τον αριθμό των μονάδων που βρίσκονται σε κίνδυνο πριν τη χρονική στιγμή $t_{(j)}$.

Όταν ισχύει η υπόθεση της αναλογικής διακινδύνευσης στις δύο ομάδες που θέλουμε να ελέγξουμε αν έχουν διαφοροποιήσεις στις συναρτήσεις επιβίωσής τους πιο κατάλληλος είναι ο έλεγχος log-rank. Αντιθέτως, όταν θέλουμε να εντοπίσουμε τις διαφοροποιήσεις στις συναρτήσεις επιβίωσης που εμφανίζονται νωρίς καλύτερο είναι να χρησιμοποιούμε τον έλεγχο του Wilcoxon καθώς αυτός δίνει μεγαλύτερη βάση στους τερματισμούς που προκύπτουν νωρίς στο πείραμα σε σύγκριση με αυτούς που θα συμβούν αργότερα στο πείραμα.

ΚΕΦΑΛΑΙΟ 2. Παραμετρικά Μοντέλα Αναλογικής Διακινδύνευσης - Το ημιπαραμετρικό μοντέλο του Cox

2.1 Μοντέλο Αναλογικής Διακινδύνευσης (PH)

Από τη γραμμική παλινδρόμηση έχει γίνει γνωστό ότι σε δείγμα μεγέθους n όταν θέλουμε να περιγράψουμε την εξάρτηση μιας μεταβλητής απόκρισης Y από κάποιες ανεξάρτητες μεταβλητές $x_i,\ i = 1,2,\dots,n$ προσαρμόζουμε στα δεδομένα ενός δείγματος το γενικό γραμμικό μοντέλο το οποίο παρουσιάζεται παρακάτω

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon\nu\acute{o}\varsigma + \beta_k x_{ik} + \varepsilon_i$$

Από την παραπάνω σχέση γνωρίζουμε ότι η μεταβλητή απόκρισης ακολουθεί την κανονική κατανομή. Στα δεδομένα διάρκειας ζωής η μεταβλητή απόκρισης είναι ο χρόνος ο οποίος λαμβάνει μόνο θετικές τιμές οπότε δεν είναι εφικτό να προσαρμόσουμε το γενικό γραμμικό μοντέλο.

Τα μοντέλα που επιλύουν την παραπάνω δυσκολία είναι τα μοντέλα αναλογικής διακινδύνευσης. Τα μοντέλα αυτά περιγράφουν την εξάρτηση του χρόνου επιβίωσης από τις επεξηγηματικές μεταβλητές και δίνονται από τον γενικό τύπο:

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

όπου $h_0(t)$ μια κοινή συνάρτηση διακινδύνευσης για τις παρατηρήσεις και $g(\vec{x})$ είναι μια συνάρτηση των ανεξάρτητων μεταβλητών x_i που εκφράζει τη μεταβλητότητα κάθε μονάδας. Αξίζει να σημειώσουμε ότι οι τιμές των ανεξάρτητων μεταβλητών δεν επηρεάζονται ούτε μεταβάλλονται κατά την πάροδο του χρόνου, αντιθέτως οι τιμές παραμένουν σταθερές και ίσες με τις τιμές που είχαν τη στιγμή που πραγματοποιήθηκε η δειγματοληψία.

Η συνάρτηση $g(\vec{x})$ τις περισσότερες φορές είναι ίση με $\exp(\vec{\beta}^T \cdot \vec{x})$ όπου β οι συντελεστές των μεταβλητών , οι οποίοι εκφράζουν ποσοτικά την επίδραση της καθεμιάς των συμμεταβλητών. Τελικά λαμβάνουμε το παρακάτω μοντέλο.

$$h(t;\vec{x}) = h_0(t) \cdot \exp(\vec{\beta}^T \cdot \vec{x})$$

Υπάρχουν δύο είδη μοντέλων Αναλογικής Διακινδύνευσης που βασίζονται στη συνάρτηση $h_0(t)$:

1) Παραμετρικά μοντέλα, όπου η συνάρτηση διακινδύνευσης $h_0(t)$ καθορίζεται από κάποιο γνωστό παραμετρικό μοντέλο όπως το Weibull ή το Log-Logistic.

2) Ημι-παραμετρικά μοντέλα, όπου η συνάρτηση $h_0(t)$ παραμένει ακαθόριστη, το πιο βασικό παράδειγμα ημι-παραμετρικού μοντέλου είναι αυτό του Cox με το οποίο θα ασχοληθούμε στη συνέχεια.

Με τη βοήθεια της παραπάνω σχέσης και του ορισμού της σωρευτικής συνάρτησης διακινδύνευσης υπολογίζουμε τα εξής:

$$H(t;\vec{x}) = \int_0^t h(u;\vec{x})du = \int_0^t h_0(t) \cdot exp\,(\vec{\beta}^T \cdot \vec{x})\,du = H_0(t) \cdot exp\,(\vec{\beta}^T \cdot \vec{x})$$

Αντίστοιχα λαμβάνουμε την συνάρτηση επιβίωσης:

$$S(t;\vec{x}) = exp\big(-H(t;\vec{x})\big) = \exp\big(-H_0(t) \cdot ex\,p\big(\vec{\beta}^T \cdot \vec{x}\big)\big) = (S_0(t))^{ex\,p\left(\vec{\beta}^T \cdot \vec{x}\right)}$$

Θεωρούμε το λόγο των συναρτήσεων διακινδύνευσης δύο ατόμων (Hazard Ratio – HR) ως εξής:

$$HR_{(t)} = \frac{h(t;\vec{x}_1)}{h(t;\vec{x}_2)} = \frac{h_0(t) \cdot exp\,(\vec{\beta}^T \cdot\ \vec{x}_1)}{h_0(t) \cdot exp\,(\vec{\beta}^T \cdot\ \vec{x}_2)} = \frac{exp\,(\vec{\beta}^T \cdot\ \vec{x}_1)}{exp\,(\vec{\beta}^T \cdot\ \vec{x}_2)} = exp\,(\vec{\beta}^T \cdot (\vec{x}_1 - \vec{x}_2))$$

Παρατηρούμε ότι ο λόγος αυτός είναι σταθερός, επομένως και ανεξάρτητος του χρόνου. Όταν συμβαίνει αυτό έχουμε την ιδιότητα ενός αναλογικής διακινδύνευσης. Άρα οι συναρτήσεις διακινδύνευσης είναι σε αναλογία μεταξύ τους. Αυτή είναι και η βασική υπόθεση ενός μοντέλου αναλογικής διακινδύνευσης και πρέπει να ελέγχεται πάντα. (Caroni, 2017)

2.1.1 Γραφικός έλεγχος καταλληλόλητας του μοντέλου

Για να ελέγξουμε αν ισχύει η υπόθεση της αναλογικής διακινδύνευσης μπορούμε να κάνουμε ένα γραφικό έλεγχο. Αν εκμεταλλευτούμε τη σχέση από την παραπάνω ενότητα για τη συνάρτηση επιβίωσης λαμβάνουμε τα παρακάτω:

$$S(t;\vec{x}) = \exp\big(-H_0(t) \cdot ex\,p\big(\vec{\beta}^T \cdot \vec{x}\big)\big)$$

$$\Rightarrow ln\,\{-ln\,S(t;\vec{x})\} - ln\,H_0(t) = \vec{\beta}^T \cdot \vec{x}$$

Από την παραπάνω σχέση βγάζουμε το συμπέρασμα ότι οι καμπύλες $ln\,\{-ln\,S(t;\vec{x})\}$ για διάφορες τιμές του x είναι παράλληλες μεταξύ τους ως προς τον χρόνο t. Για το γραφικό έλεγχο χωρίζουμε τα δεδομένα μας σε ομάδες σύμφωνα με τις κοινές τιμές της μεταβλητής x και με αυτά τα διαφορετικά x θα κάνουμε τις καμπύλες $ln\,\{-ln\,S(t;\vec{x})\}$. Για τη συνάρτηση επιβίωσης S(t; x) θα χρησιμοποιήσουμε την εκτίμηση Kaplan-Meier.

Εναλλακτικός γραφικός έλεγχος είναι να κατασκευάσουμε τη γραφική παράσταση των καμπυλών $ln\{-ln\,S(t;\vec{x})\}$ συναρτήσει του $\ln(t)$. Και σε αυτή την περίπτωση θα έχουμε ευθείες οι οποίες θα είναι παράλληλες μεταξύ τους και ως προς τον άξονα x. Αν καταλήξουμε σε κάτι τέτοιο τότε εξάγουμε το συμπέρασμα ότι ισχύει η υπόθεση της αναλογικής διακινδύνευσης. Οι καμπύλες αυτές διαφέρουν μόνο ως προς τις οριζόντιες μετατοπίσεις. (Καρώνη 2009)

2.1.2 Εκτίμηση παραμέτρων

Η εκτίμηση παραμέτρων για το μοντέλο της αναλογικής διακινδύνευσης γίνεται με τη μέθοδο μέγιστης πιθανοφάνειας. Έστω ότι έχουμε δεξιά αποκομμένες παρατηρήσεις τότε η συνάρτηση Πιθανοφάνειας έχει ως εξής:

$$L = \prod_{i=1}^{n} \{f(t_i)\}^{\delta_i} \{S(t_i)\}^{1-\delta_i}$$

όπου δ_i ένας δίτιμος δείκτης που λαμβάνει τιμή ανάλογα με το αν η τιμή i είναι αποκομμένη ή όχι. Πιο συγκεκριμένα έχουμε:

$$\delta_i = \begin{cases} 0, & \delta\varepsilon\nu\ \sigma\upsilon\nu\acute{\varepsilon}\beta\eta\ \sigma\tau o\ \gamma\varepsilon\gamma o\nu\acute{o}\varsigma, \delta\eta\lambda\ \acute{\varepsilon}\chi o\upsilon\mu\varepsilon\ \alpha\pi o\kappa o\mu\mu\acute{\varepsilon}\nu\eta\ \tau\iota\mu\acute{\eta} \\ 1, & \sigma\upsilon\nu\acute{\varepsilon}\beta\eta\ \tau o\ \gamma\varepsilon\gamma o\nu\acute{o}\varsigma, \delta\eta\lambda\ \delta\varepsilon\nu\ \acute{\varepsilon}\chi o\upsilon\mu\varepsilon\ \alpha\pi o\kappa o\mu\mu\acute{\varepsilon}\nu\eta\ \tau\iota\mu\acute{\eta} \end{cases}$$

Λογαριθμίζοντας τη συνάρτηση Πιθανοφάνειας έχουμε:

$$l = \sum_{i=1}^{n} \{\delta_i lnf(t_i) + (1-\delta_i)\ lnS(t_i)\}$$

Δεδομένου ότι για το μοντέλο της αναλογικής διακινδύνευσης έχουμε

$$S(t_i) = (S_0(t))^{exp\left(\vec{\beta}^T \cdot \vec{x}_i\right)}$$

η λογαριθμοποιημένη πιθανοφάνεια λαμβάνεται ως εξής

$$l = \sum_{i=1}^{n} \left\{\delta_i lnf(t;\vec{x}_i) + (1-\delta_i) exp\left(\vec{\beta}^T \cdot \vec{x}_i\right)\ ln(S_0(t))\right\}$$

Τελικά παραγωγίζουμε την παραπάνω σχέση ως προς β_j εξισώνουμε με το μηδέν και λύνουμε το σύστημα με τη χρήση κατάλληλων αριθμητικών μεθόδων. (Xu et al. 2009)

2.2 Το ημιπαραμετρικό μοντέλο του Cox

Όταν η μελέτη μας αφορά ένα σύνολο ανθρώπων και το πως αυτοί αντιδρούν σε κάποια θεραπεία η βασική συνάρτηση διακινδύνευσης $\mathrm{h_0(t)}$ παραμένει ακαθόριστη. Αυτό συμβαίνει διότι ο κάθε οργανισμός έχει τα δικά του μοναδικά χαρακτηριστικά οπότε δεν είναι τόσο εύκολο να καταλήξουμε σε ένα παραμετρικό μοντέλο. Για το λόγο αυτό εισάγουμε το ημι-παραμετρικό μοντέλο του Cox το οποίο είναι ένα μοντέλο Αναλογικής Διακινδύνευσης που εισήχθη από τον Άγγλο στατιστικολόγο David Cox το 1972. Όπως και στα μοντέλα αναλογικής διακινδύνευσης η συνάρτηση διακινδύνευσης h(t,x) ορίζεται ως:

$$h(t;\vec{x}) = h_0(t) \cdot \exp\,(\vec{\beta}^T \vec{x})$$

με αντίστοιχο τρόπο παίρνουμε και τη σωρευτική συνάρτηση διακινδύνευσης

$$H(t;\vec{x}) = \int_0^t h_0(u) \cdot \exp\,(\vec{\beta}^T \vec{x})\, du = H_0(t)\exp\,(\vec{\beta}^T \vec{x})$$

Οπότε και η συνάρτηση επιβίωσης έχει ως εξής

$$S(t;\vec{x}) = \exp\big(-H(t;\vec{x})\big) = \exp\big(H_0(t)\exp\,(\vec{\beta}^T \vec{x})\big) = (S_0(t))^{\exp\,(\vec{\beta}^T \vec{x})}$$

Τελικά οι παραμετρικές μορφές των συναρτήσεων δεν καθαρίζονται, για αυτό το λόγο το μοντέλο του Cox καλείται ημι-παραμετρικό. Η μόνη ανάλυση που μπορούμε να κάνουμε είναι να ελέγξουμε την επίδραση των ανεξάρτητων μεταβλητών $\vec{x}$. (Hosmer, Lemeshow & May, 2008)

2.2.1 Εκτίμηση παραμέτρων του μοντέλου

Στην παράγραφο αυτή θα εκτιμήσουμε παραμέτρους του μοντέλου με τη μέθοδο της μέγιστης πιθανοφάνειας. Έστω ότι κατά όπως διακεκριμένες χρονικές στιγμές

$$t_1 < t_2 < \text{όπως} < t_k$$

διακόπτεται η λειτουργία k μονάδων. Δηλαδή τη χρονική στιγμή t_j διακόπτεται η λειτουργία μίας μόνο μονάδας με συμμεταβλητές $\vec{x}$. Ακόμα θεωρούμε R_j το σύνολο των παρατηρήσεων που βρίσκονται σε κίνδυνο τη χρονική στιγμή t_j. Η πιθανότητα να διακοπεί η λειτουργία μίας συγκεκριμένης μονάδας j , δοθέντος ότι σταματά να λειτουργεί μία οποιαδήποτε μονάδα από το σύνολο R_j , δίνεται από το τύπο:

$$\frac{h\big(t(j);\vec{x}_j\big)}{\sum_{i\epsilon R_j} h(t(j);\vec{x}_i)} = \frac{\exp\,(\vec{\beta}^T \vec{x}_j)}{\sum_{i\epsilon R_j} \exp\,(\vec{\beta}^T \vec{x}_i)}$$

Η συνάρτηση μερικής πιθανοφάνειας είναι η παρακάτω

$$L(\beta) = \prod_{j=1}^{k} \left\{ \frac{\exp\left(\vec{\beta}^T \vec{x}_j\right)}{\sum_{i \epsilon R_j} \exp\left(\vec{\beta}^T \vec{x}_i\right)} \right\}$$

Από την οποία παίρνουμε την εκτιμήτρια μεγίστης πιθανοφάνειας $\hat{\beta}$ του β. Ο λογάριθμος της πιθανοφάνειας είναι

$$l(\beta) = \sum_{j=1}^{k} \vec{\beta}^T \vec{x}_j - \sum_{j=1}^{k} ln \left\{ \sum_{i \epsilon R_j} \exp\left(\vec{\beta}^T \vec{x}_i\right) \right\}$$

Παραγωγίζοντας την τελευταία ως όπως β_r και εξισώνοντας με το μηδέν έχουμε

$$\sum_{j=1}^{k} \vec{x}_{jr} - \sum_{j=1}^{k} \left\{ \frac{\sum_{i \epsilon R_j} \vec{x}_{ir} \exp\left(\vec{\beta}^T \vec{x}_i\right)}{\sum_{i \epsilon R_j} \exp\left(\vec{\beta}^T \vec{x}_i\right)} \right\} = 0$$

Η τελευταία επιλύεται με κατάλληλες μεθόδους όπως του Newton Raphson.
(Cox 1972)

2.2.2 Έλεγχος υπόθεσης αναλογικής διακινδύνευσης

Σημαντικό βήμα της μελέτης είναι ο έλεγχος της καταλληλόλητας του μοντέλου του Cox. Για το λόγο αυτό πρέπει να ελέγχεται αν ισχύει η υπόθεση της αναλογικής διακινδύνευσης, όπως είχαμε δει και στα μοντέλα αναλογικής διακινδύνευσης. Όπως στην παράγραφο 2.1 έτσι και τώρα θέλουμε ο παρακάτω λόγος να είναι ανεξάρτητος του χρόνου

$$HR_{(t)} = \frac{h(t; \vec{x}_1)}{h(t; \vec{x}_2)} = \frac{h_0(t) \cdot exp\left(\vec{\beta}^T \cdot \vec{x}_1\right)}{h_0(t) \cdot exp\left(\vec{\beta}^T \cdot \vec{x}_2\right)} = \frac{exp\left(\vec{\beta}^T \cdot \vec{x}_1\right)}{exp\left(\vec{\beta}^T \cdot \vec{x}_2\right)} = exp\left(\vec{\beta}^T \cdot (\vec{x}_1 - \vec{x}_2)\right)$$

όπου 1, 2 δύο διαφορετικές μονάδες.

Ο έλεγχος αυτός πραγματοποιείται μέσω του γραφικού ελέγχου της υπόθεσης της αναλογικής διακινδύνευσης που είδαμε στην Παράγραφο 2.1.1. Εναλλακτικός τρόπος είναι μέσω των υπολοίπων Schoenfeld που θα δούμε στην παρακάτω παράγραφο.

2.2.3 Υπόλοιπα Schoenfeld

Τα μερικά υπόλοιπα (partial residuals) ή αλλιώς υπόλοιπα Schoenfeld εισήχθησαν από τον Schoenfeld το 1982 και είναι τα υπόλοιπα που χρησιμοποιούμε κατά κύριο λόγο στο ημι-παραμετρικό μοντέλο του Cox.

Το πλεονέκτημα των υπολοίπων Schoenfeld έναντι των άλλων υπολοίπων είναι ότι για τον υπολογισμό της δεν χρειάζεται η εκτίμηση της αθροιστικής αναφορικής συνάρτησης κινδύνου. Στα άλλα υπόλοιπα, υπολογίζεται ένα μόνο υπόλοιπο για κάθε άτομο. Τα υπόλοιπα Schoenfeld, υπολογίζουν ένα ξεχωριστό υπόλοιπο για κάθε άτομο για κάθε μεταβλητή. Δηλαδή αν έχουμε k μεταβλητές, τότε για κάθε άτομο υπολογίζονται k υπόλοιπα Schoenfeld. Με το γράφημα των υπολοίπων Schoenfeld συναρτήσει του χρόνου μπορούμε να ελέγξουμε την υπόθεση της αναλογικότητας των κινδύνων. Αν το γράφημα έχει μια τυχαία μορφή των υπολοίπων έναντι του χρόνου τότε ικανοποιείται η υπόθεση. Αντίθετα, σημαίνει ότι δεν ικανοποιείται η υπόθεση.

2.2.4 Καμπύλες ROC και AUC

Η προβλεπτική ικανότητα του μοντέλου του Cox μπορεί να εξεταστεί με τη βοήθεια της καμπύλης ROC και της τιμής AUC. Αν ένα μοντέλο έχει καλή προβλεπτική ικανότητα τότε μπορεί να προβλέψει σωστά (με μεγάλη πιθανότητα επιτυχίας) την τιμή της μεταβλητής απόκρισης σε νέα δεδομένα που εισάγουμε στο δείγμα.

Αν $\hat{p} = \hat{p}\ (Y = 1)$ η εκτιμημένη πιθανότητα επιτυχίας (δηλαδή όταν το γεγονός συμβαίνει – επέρχεται ο θάνατος του ατόμου) για κάθε μονάδα και p_0 ένα κατώφλι, διακρίνουμε τις εξής δύο περιπτώσεις:

- Αν $\hat{p} > p_0$, προβλέπεται Y=1 για την μονάδα αυτή
- Αν $\hat{p} < p_0$, προβλέπεται Y=0 για την μονάδα αυτή

Έτσι μπορούμε να κατασκευάσουμε τον πίνακα συνάφειας 2 x 2 ο οποίος παρατίθεται στον Πίνακα 1.2.

		Πραγματική κατάσταση	
		Y=1	Y=0
Πρόβλεψη	Y=1	a	b
	Y=0	c	d

Πίνακας 1.2: Πίνακας 2 x 2 συνάφειας

Από τον Πίνακα 1.2 ορίζονται οι ακόλουθες ποσότητες:

- Ευαισθησία (sensitivity) = $\frac{a}{a+c}$

Η ευαισθησία εκφράζει το πόσο συχνά προβλέπουμε σωστά ότι είναι Y=1 (true positive rate)

- Ειδικότητα (specificity) = $\frac{d}{b+d}$

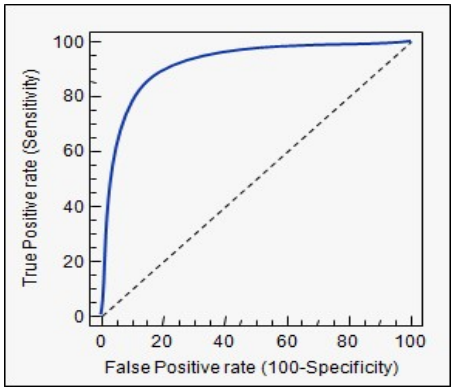
Η ευαισθησία εκφράζει το πόσο συχνά προβλέπουμε σωστά ότι είναι Y=1 (true positive rate). Ενώ η ειδικότητα εκφράζει το πόσο συχνά προβλέπουμε σωστά ότι είναι Y=0 (true negative rate) . ρ

Με τη βοήθεια της ειδικότητας μπορούμε εύκολα να ορίσουμε μία νέα ποσότητα η οποία καλείται 1-ειδικότητα (false positive rate) και εκφράζει το πόσο συχνά προβλέπουμε λάθος ότι είναι Y=1.

$$1\text{-ειδικότητα} = \frac{b}{b+d}$$

Όταν η ευαισθησία λαμβάνει μεγάλες τιμές η 1-ειδικότητα θα λαμβάνει μικρές τιμές και τότε το μοντέλο μας θα έχει υψηλή προβλεπτική ικανότητα.

Ένας τρόπος για να δούμε άμεσα την προβλεπτική ικανότητα του μοντέλου είναι να κατασκευάσουμε την καμπύλη ROC (Receiver Operating Characteristic) η οποία παρουσιάζεται στο Διάγραμμα 1.4. Σε αυτή τη γραφική παράσταση αξίζει να σημειωθεί ότι στον άξονα των x έχουμε το 1-ειδικότητα ενώ στον άξονα των y έχουμε την ευαισθησία. Η καμπύλη ROC είναι η μπλε γραμμή ενώ η διακεκομμένη ευθεία είναι αυτή για την οποία ισχύει ότι: ποσοστό αληθώς θετικών αποτελεσμάτων = ποσοστό ψευδώς θετικών αποτελεσμάτων. Το εμβαδόν που δημιουργείται κάτω από την καμπύλη ROC ονομάζεται AUC (Area Under the Curve) και όσο πιο κοντά είναι η τιμή του στο ένα τόσο καλύτερη είναι και η προβλεπτική ικανότητα του μοντέλου. Ακόμα αξίζει να σημειωθεί ότι το εμβαδό που δημιουργεί η γκρι γραμμή θα είναι πάντα ίσο με 0,5. Επομένως η τιμή του AUC θα κυμαίνεται από την τιμή 0,5 ως την τιμή 1. (Heagerty & Zheng, 2005)



Διάγραμμα 1.4: Καμπύλη ROC

2.3 Έλεγχοι υποθέσεων

Σημαντικό ρόλο κατά την προσαρμογή του μοντέλου είναι να μπορούμε να εκτελούμε ελέγχους που αφορούν τη σημαντικότητα των παραμέτρων που συμπεριλαμβάνονται στο εκτιμημένο μοντέλο. Για τη σημαντικότητα των μεταβλητών χρησιμοποιούμε τον έλεγχο Wald (Wald test), τον έλεγχο του λόγου των πιθανοφανειών (Likelihood ratio test), και τη μέθοδο της διαδοχικής αφαίρεσης (backward elimination).

2.3.1 Ο έλεγχος του Wald

Για να ελέγξουμε τη σημαντικότητα των συντελεστών του μοντέλου, οι οποίες αποτελούν σημειακές εκτιμήσεις, θα χρησιμοποιήσουμε τον έλεγχο του Wald. Με τον έλεγχο του Wald μπορούμε να εξετάσουμε τις εξής υποθέσεις:

H_0: β_i=0, δηλαδή η μεταβλητή x_i δε συμβάλει στο μοντέλο

H_1: $\beta_i \neq 0$, δηλαδή η μεταβλητή x_i συμβάλει στο μοντέλο

Ο έλεγχος γίνεται σε (1-α)% επίπεδο σημαντικότητας και η στατιστική συνάρτηση υπό την H_0 είναι η ακόλουθη:

$$\frac{\hat{\beta}_i}{se(\hat{\beta}_i)}$$

Η οποία καλείται Wald, ακολουθεί ασυμπτωτικά την κανονική κατανομή, ενώ το τετράγωνο της ακολουθεί την χ^2 κατανομή με 1 βαθμό ελευθερίας. Αν η p-value του ελέγχου Wald κριθεί ότι έχει λάβει μικρή τιμή τότε απορρίπτεται η μηδενική υπόθεση και έτσι ο συντελεστής β_i δεν μηδενίζεται οπότε η μεταβλητή x_i θεωρείται στατιστικά σημαντική. Σε αντίθετη περίπτωση, δηλαδή όταν η τιμή της p θεωρηθεί μεγάλη δεν απορρίπτουμε τη μηδενική υπόθεση. (Collett, 2003)

2.3.2 Ο έλεγχος του λόγου των πιθανοφανειών

Ένας ακόμα έλεγχος που προτιμάται σε σχέση με τον έλεγχο του Wald είναι ο έλεγχος του λόγου των πιθανοφανειών. Ο έλεγχος αυτός απαιτεί περισσότερους υπολογισμούς από τον έλεγχο Wald. Με το έλεγχο του λόγου των πιθανοφανειών μπορούμε να εξετάσουμε τις ίδιες υποθέσεις με τον έλεγχο του Wald:

H_0: β_i=0, δηλαδή η μεταβλητή x_i δε συμβάλει στο μοντέλο

H_1: $\beta_i \neq 0$, δηλαδή η μεταβλητή x_i συμβάλει στο μοντέλο

Ο έλεγχος γίνεται σε (1-α)% επίπεδο σημαντικότητας και χρησιμοποιούμε την ελεγχοσυνάρτηση

$$z = -2(\hat{l}_0 - \hat{l}_1) \sim X_d^2$$

Όπου $\hat{l}_0$ η μεγιστοποιημένη τιμή του λογαρίθμου της Πιθανοφάνειας στο μοντέλο που δεν περιέχει καμία μεταβλητή και $\hat{l}_1$ η μεγιστοποιημένη τιμή του λογαρίθμου της Πιθανοφάνειας στο μοντέλο που περιέχει μία μεταβλητή. Αν η τιμή της p είναι αρκετά μικρή απορρίπτουμε τη μηδενική υπόθεση και άρα η μεταβλητή με συντελεστή β_i είναι στατιστικά σημαντική.

2.4 Κριτήρια επιλογής βέλτιστου μοντέλου

2.4.1 Κριτήριο AIC

Ένας από τους δείκτες καλής προσαρμογής είναι το κριτήριο AIC (Akaike's Information Criterion) και στα γενικευμένα γραμμικά μοντέλα δίνεται από τον ακόλουθο τύπο:

$$\text{AIC} = -2\text{l}(\hat{\beta}) + 2\text{p}$$

όπου, $\text{l}(\hat{\beta})$ η μεγιστοποιημένη τιμή της συνάρτησης λογαριθμοπιθανοφάνειας και p ο αριθμός των επεξηγηματικών μεταβλητών που περιέχει το μοντέλο προς εκτίμηση (Akaike, 1974).

Το AIC αποτελεί ένα κριτήριο επιλογής του βέλτιστου μοντέλου με όσο το δυνατό μικρότερο αριθμό παραμέτρων. Από τον τύπο του AIC παρατηρούμε ότι η εισαγωγή παραμέτρων στο μοντέλο, είτε αυτές είναι στατιστικά σημαντικές είτε όχι, προκαλεί αύξηση του p και του $\text{l}(\hat{\beta})$ επομένως ο πρώτος όρος στον τύπο του AIC μειώνεται και ο δεύτερος αυξάνεται. Τελικά η εισαγωγή επιπλέον παραμέτρων στο μοντέλο οι οποίες οδηγούν σε βελτίωση της προσαρμογής του μοντέλου οδηγεί σε μείωση της τιμής του κριτηρίου AIC. Συνεπώς, επιλέγουμε το μοντέλο εκείνο με τη μικρότερη τιμή του κριτηρίου AIC. (Καρώνη & Οικονόμου, 2017)

2.4.2 Κριτήριο BIC

Ένα ακόμα κριτήριο που χρησιμοποιούμε για την επιλογή του βέλτιστου μοντέλου ανάμεσα σε μοντέλα με διαφορετικό αριθμό παραμέτρων είναι το BIC (Bayesian Information Criterion) και προτάθηκε από τον Schwarz (1978). Το κριτήριο BIC δίνεται από τον ακόλουθο τύπο στα γενικευμένα γραμμικά μοντέλα:

$$\text{BIC} = -2\text{l}(\hat{\beta}) + \text{pln}(\text{n})$$

όπου, $\text{l}(\hat{\beta})$ η μεγιστοποιημένη τιμή της συνάρτησης λογαριθμοπιθανοφάνειας, p ο αριθμός των επεξηγηματικών μεταβλητών που περιέχει το μοντέλο και n ο αριθμός των παρατηρήσεων. (Volinsky & Raftery, 1995)

Η προσθήκη παραμέτρων επηρεάζει την τιμή του κριτηρίου BIC με παρόμοιο τρόπο που επηρεάζει και την τιμή του κριτηρίου AIC. Πιο συγκεκριμένα, η εισαγωγή παραμέτρων στο μοντέλο, είτε αυτές είναι στατιστικά σημαντικές είτε όχι, προκαλεί αύξηση του p και του $l(\hat{\beta})$ επομένως ο πρώτος όρος στον τύπο του BIC μειώνεται και ο δεύτερος αυξάνεται. Τελικά η εισαγωγή επιπλέον παραμέτρων στο μοντέλο οι οποίες οδηγούν σε βελτίωση της προσαρμογής του μοντέλου οδηγεί σε μείωση της τιμής του κριτηρίου BIC. Συνεπώς, επιλέγουμε το μοντέλο εκείνο με τη μικρότερη τιμή του κριτηρίου BIC. Η μόνη διαφορά μεταξύ του AIC και του BIC είναι ότι στην περίπτωση του BIC η εισαγωγή επιπρόσθετων παραμέτρων αποθαρρύνεται σε μεγαλύτερο βαθμό από το AIC.

2.4.3 Μέθοδος της διαδοχικής αφαίρεσης (backward elimination)

Μία μέθοδος για να καταλήξουμε στο καταλληλότερο μοντέλο για την περιγραφή των δεδομένων είναι η διαδικασία της διαδοχικής αφαίρεσης (backward elimination). Σύμφωνα με τη διαδικασία αυτή προσαρμόζεται ένα μοντέλο το οποίο περιέχει όλες τις μεταβλητές και σε κάθε βήμα αφαιρείται η μεταβλητή που είναι περισσότερο στατιστικά μη σημαντική. Το μοντέλο στο οποίο καταλήγει η διαδικασία της διαδοχικής αφαίρεσης είναι το τελικό μοντέλο από το οποίο δεν μπορούμε να αφαιρέσουμε άλλες μεταβλητές καθώς όσες έχουν μείνει είναι στατιστικά σημαντικές.

ΚΕΦΑΛΑΙΟ 3. Μέθοδοι συρρίκνωσης

3.1 Το πρόβλημα της πολυσυγγραμμικότητας

Ορισμένες φορές κατά την ανάλυση των δεδομένων προκύπτει το πρόβλημα στις πολυσυγγραμμικότητας. Όταν δύο ή και περισσότερες ανεξάρτητες μεταβλητές του δείγματος παρουσιάζουν εξάρτηση η μία από την άλλη τότε λέμε ότι εμφανίζεται το φαινόμενο στις πολυσυγγραμμικότητας (multicollinearity). Στην περίπτωση αυτή είναι πιθανό να μπορούν να υπολογιστούν οι τιμές της μίας μεταβλητής από τις τιμές άλλων ανεξάρτητων μεταβλητών. Ακραία περίπτωση πολυσυγγραμμικότητας εμφανίζεται όταν μία επεξηγηματική μεταβλητή είναι γραμμικός συνδυασμός μερικών ή όλων των επεξηγηματικών μεταβλητών. Συχνός τρόπος για να ελέγξουμε αν εμφανίζεται το φαινόμενο της πολυσυγγραμμικότητας είναι ο υπολογισμός του συντελεστή συσχέτισης μεταξύ δύο ανεξάρτητων μεταβλητών. Όσο πιο κοντά στη μονάδα βρίσκεται ο συντελεστής συσχέτισης τόσο μεγαλύτερη είναι η και εξάρτηση της μίας μεταβλητής από την άλλη.

Το πρόβλημα έγκειται στο γεγονός ότι το μοντέλο ενδέχεται να μην είναι αξιόπιστο, ενώ η προβλεπτική του ικανότητα δεν επηρεάζεται ιδιαιτέρως. Επομένως, αν προσθέσουμε ή αφαιρέσουμε μία μεταβλητή υπάρχει μεγάλη πιθανότητα να μεταβληθούν σε μεγάλο βαθμό οι συντελεστές των ανεξάρτητων μεταβλητών του μοντέλου. Όταν προκύπτει το πρόβλημα αυτό είναι δύσκολο να εντοπίσουμε τις στατιστικά σημαντικές μεταβλητές του μοντέλου και για το λόγο αυτό εφαρμόζουμε κάποιες μεθόδους συρρίκνωσης.

Οι μέθοδοι συρρίκνωσης (shrinkage methods) περιορίζουν την εκτίμηση των συντελεστών των επεξηγηματικών μεταβλητών και σε κάποιες περιπτώσεις μηδενίζουν τους συντελεστές. Στις παρακάτω παραγράφους θα ασχοληθούμε με την παλινδρόμηση κορυφογραμμής (Ridge), την παλινδρόμηση Lasso και τη μέθοδο Elastic Net.

3.2 Η παλινδρόμηση κορυφογραμμής (Ridge)

Η παλινδρόμηση κορυφογραμμής Ridge μοιάζει με τη μέθοδο ελαχίστων τετραγώνων και πιο συγκεκριμένα αποτελεί μία βελτιωμένη έκδοσή της. Ο λόγος που καλούμαστε να χρησιμοποιήσουμε τη μέθοδο Ridge είναι ότι η μέθοδος ελαχίστων τετραγώνων επηρεάζεται από το φαινόμενο της πολυσυγγραμμικότητας και για το λόγο αυτό εκτιμάει τους συντελεστές β_j με πολύ μεγάλα τυπικά σφάλματα. Στο σημείο αυτό υπενθυμίζουμε ότι η μέθοδος ελαχίστων τετραγώνων εκτιμάει στους συντελεστές των ανεξάρτητων μεταβλητών ελαχιστοποιώντας το άθροισμα των τετραγώνων των υπολοίπων (RSS: Residual Sum of Squares), δηλαδή τη συνάρτηση:

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

Από την άλλη η μέθοδος Ridge για να εκτιμήσει τους συντελεστές των ανεξάρτητων μεταβλητών ελαχιστοποιεί την ακόλουθη συνάρτηση:

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

$$L_{Ridge} = RSS + \lambda \sum_{j=1}^{p}\beta_j^2$$

όπου n το μέγεθος του δείγματος και λ μία μη αρνητική παράμετρος συρρίκνωσης (tuning parameter) η οποία καθορίζει το μέγεθος της συρρίκνωσης. Όσο μεγαλύτερη είναι η τιμή της παραμέτρου λ τόσο μεγαλύτερη η συρρίκνωση. Η ποινή που εφαρμόζει η μέθοδος Ridge δίνεται από τον όρο $\lambda \sum_{j=1}^{p}\beta_j^2$.

Ο όρος $\sum_{j=1}^{p}\beta_j^2$ καλείται l_2 – penalty γιατί έχει να κάνει με την l_2- νόρμα. Υπενθυμίζουμε ότι η l_2- νόρμα δίνεται από τον ακόλουθο τύπο.

$$\|\beta\|_2 = \sqrt{\sum_{j=1}^{p}\beta_j^2}$$

Παρατηρούμε ότι όσο αυξάνεται η παράμετρος λ, τόσο η l_2- νόρμα των εκτιμήσεων των συντελεστών θα μειώνεται.

Η μέθοδος ποινής Ridge είναι ικανή να υπολογίσει τις εκτιμήσεις των συντελεστών β_j για κάθε μία από τις τιμές που λαμβάνει η τιμή λ. Η μέθοδος δεν εφαρμόζει ποινή στο σταθερό όρο β_0. Όταν η παράμετρος λ λαμβάνει πολύ μεγάλες τιμές δηλαδή τείνει στο άπειρο, οι εκτιμήσεις της μεθόδου Ridge των συντελεστών τείνουν στο 0, καθώς ο όρος της ποινής θα είναι πολύ μεγάλος. Αντιθέτως, όταν η τιμή της παραμέτρου λ είναι ίση με το μηδέν ο όρος της ποινής μηδενίζεται, επομένως η Ridge υπολογίζει τις ίδιες εκτιμήσεις με τη μέθοδο των ελαχίστων τετραγώνων. Αν λύσουμε την εξίσωση βρίσκουμε τις εκτιμήσεις των β_j , j=1,…,p της Παλινδρόμησης Κορυφογραμμής:

$$\hat{\beta}^R = (X'X + \lambda I)^{-1}X'y$$

Όπου I ο μοναδιαίος πίνακας με διαστάσεις pxp. Το πρόβλημα αυτό είναι εύκολο να επιλυθεί διότι ο πίνακας $(X'X + \lambda I)$ είναι πάντα αντιστρέψιμος.

Ενναλακτική συνάρτηση που η μέθοδος Ridge ελαχιστοποιεί είναι η

$$L_R = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2, \text{δεδομένου ότι } \sum_{j=1}^{p}\beta_j^2 \leq \lambda$$

3.3 Η μέθοδος Lasso

Η μέθοδος Lasso ή L_1 αποτελεί άλλη μία μέθοδο συρρίκνωσης (Least Absolute Shrinkage and Selection Operator) η οποία εισήχθη από τον Tibshirani το 1996. Η μέθοδος Lasso βρίσκει εφαρμογή σε πολλά μοντέλα μεταξύ των οποίων και το μοντέλο αναλογικής διακινδύνευσης του Cox.

Αν συγκρίνουμε τη μέθοδο αυτή με τη μέθοδο Ridge θα παρατηρήσουμε αρκετές ομοιότητες αλλά και μία σημαντική διαφορά. Ενώ η μέθοδος Ridge συρρικνώνει τις όλους του συντελεστές β_j χωρίς να μηδενίζει κανέναν (εκτός και αν το λ τείνει στο άπειρο), η μέθοδος Lasso συρρικνώνει κάποιους από τις συντελεστές του μοντέλου και όλους τους υπόλοιπους τους μηδενίζει. Το πλεονέκτημα της μεθόδου Lasso έναντι της Ridge είναι ότι μπορεί να χρησιμοποιηθεί και ως μέθοδος επιλογής του βέλτιστου μοντέλου καθώς τους συντελεστές των στατιστικά μη σημαντικών μεταβλητών τους μηδενίζει.

Η μέθοδος Lasso εκτιμάει τις συντελεστές β_j ελαχιστοποιώντας την ποσότητα

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

$$L_{Lasso} = RSS + \lambda \sum_{j=1}^{p}|\beta_j|$$

όπου n το μέγεθος του δείγματος και λ μία μη αρνητική παράμετρος συρρίκνωσης (tuning parameter) η οποία καθορίζει το μέγεθος της συρρίκνωσης, όπως και στη μέθοδο Ridge. Έτσι όσο μεγαλύτερη είναι η τιμή τις παραμέτρου λ, τόσο μεγαλύτερη θα είναι η συρρίκνωση. Η ποσότητα $\lambda \sum_{j=1}^{p}|\beta_j|$ είναι ο όρος ποινής τις μεθόδου Lasso και είναι το μόνο σημείο στο οποίο διαφέρει η μέθοδος Lasso με τη μέθοδο Ridge.

Στην περίπτωση αυτή έχουμε l_1 – penalty γιατί ο όρος $\sum_{j=1}^{p}|\beta_j|$ έχει να κάνει με την $l_1 -$ νόρμα. Παρακάτω δίνεται και ο τύπος τις $l_1 -$ νόρμας.

$$\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$$

Ενναλακτική συνάρτηση που η μέθοδος Lasso ελαχιστοποιεί είναι η

$$L_L = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2, \text{δεδομένου ότι } \sum_{j=1}^{p}|\beta_j| \le \lambda$$

Τα αποτελέσματα που λαμβάνουμε για τις εκτιμήσεις των συντελεστών είναι τα ίδια χρησιμοποιώντας οποιαδήποτε από τις δύο συναρτήσεις.

Συνοψίζοντας μπορούμε να πούμε ότι η μέθοδος Lasso είναι προτιμότερη σε περιπτώσεις όπου δεν υπάρχει μεγάλη συσχέτιση μεταξύ των μεταβλητών. Ενώ η Ridge σε περιπτώσεις όπου έχουμε μεγάλη συσχέτιση.

3.4 Η μέθοδος Elastic Net

Επειδή οι παραπάνω δύο μέθοδοι παρουσιάζουν μειονεκτήματα ανάλογα με τα χαρακτηριστικά του συνόλου δεδομένων που διαχειριζόμαστε χρήσιμο είναι να αναφέρουμε μία μέθοδο που συνδυάζει τις δύο προαναφερθείσες. Η μέθοδος αυτή καλείται Elastic Net και εισήχθη από τους Zou και Hastie το 2005. Η Elastic Net προσπαθεί να συνδυάσει τα πλεονεκτήματα των μεθόδων Ridge και Lasso. Στόχος λοιπόν είναι να γίνει η καλύτερη δυνατή επιλογή μεταβλητών όπως στη Lasso και ταυτόχρονα το μοντέλο να έχει την καλύτερη προβλεπτική ικανότητα όπως γίνεται με τη μέθοδο Ridge. (Zou & Hastie 2005)

Η μέθοδος Elastic Net εκτιμάει τους συντελεστές β_j ελαχιστοποιώντας την ποσότητα

$$L_{el.net} = \frac{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2}{2n} + \lambda\left\{\frac{1-\alpha}{2}\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}|\beta_j|\right\}$$

όπου n το μέγεθος του δείγματος και λ μία μη αρνητική παράμετρος συρρίκνωσης η οποία καθορίζει το μέγεθος της συρρίκνωσης όπως ακριβώς και στις δύο προηγούμενες μεθόδους. Το α είναι μία παράμετρος ανάμιξης (mixing parameter) που παίρνει τιμές μεταξύ του 0 και του 1. Όταν το α λαμβάνει την τιμή 0 οι λύσεις που προκύπτουν ταυτίζονται με εκείνες της μεθόδου Ridge, ενώ όταν το α παίρνει την τιμή 1 οι λύσεις είναι ίδιες με τις λύσεις της μεθόδου Lasso. Έχει αποδειχτεί ότι, για τιμές του α ανάμεσα στις τιμές 0.95 και 1, η μέθοδος συμπεριφέρεται με πανομοιότυπο τρόπο με τη μέθοδο Lasso, απαλείφοντας απλώς εκφυλιστικές συμπεριφορές στο μοντέλο λόγω πολύ υψηλής συσχέτισης των μεταβλητών (Noah et al., 2011).

3.5 Cross - Validation (cvl)

Στόχος μιας στατιστικής ανάλυσης δεν είναι μόνο να κατασκευάσουμε ένα μοντέλο που να περιγράφει καλύτερα τα δεδομένα αλλά να έχει και ικανοποιητική προβλεπτική ικανότητα. Η προβλεπτική ικανότητα του μοντέλου είναι απαραίτητη ώστε να μπορεί να προβλέψει τη μεταβλητή απόκρισης σε νέα δεδομένα που τυχόν εισάγουμε στο δείγμα μετά την προσαρμογή του μοντέλου. Η προβλεπτική ικανότητα σε μοντέλα παλινδρόμησης μπορεί να ελεγχθεί με τη βοήθεια του κριτηρίου AIC, ή με άλλα μέτρα όπως Allen's PRESS και Mallows Cp. Τα δύο τελευταία δεν μπορούν να χρησιμοποιηθούν στο μοντέλο της αναλογικής διακινδύνευσης. Ακόμα το κριτήριο AIC δεν είναι εύκολο να ερμηνευθεί. Για το λόγο αυτό το 1993 οι Verweij και Van Houwelingen πρότειναν την cross-validated λογαριθμημένη πιθανοφάνεια ως μέτρο προβλεπτικής ικανότητας του μοντέλου αναλογικής διακινδύνευσης. Για το λόγο αυτό χρησιμοποιούμε τη μέθοδο cross validation.

Στην ανάλυση επιβίωσης όταν εφαρμόζουμε τη μέθοδο κορυφογραμμής Ridge και τη Lasso στόχος μας είναι να επιλέξουμε τη βέλτιστη τιμή της παραμέτρου συρρίκνωσης λ ώστε να έχουμε το μικρότερο δυνατό μέσο τετραγωνικό σφάλμα (MSE). Αυτή τη

τιμή του λ θα τη βρούμε με τη μέθοδο cross validation. Θα επιλέξουμε ένα φάσμα τιμών του λ και θα υπολογίσουμε το σφάλμα της cross validation για κάθε τιμή του λ. Έπειτα, θα επιλέξουμε τη τιμή εκείνη που ελαχιστοποιεί το σφάλμα cross validation. Τέλος, θα ξανά προσαρμόσουμε τα δύο μοντέλα μας χρησιμοποιώντας τη τιμή του λ που επιλέξαμε.

Η μέθοδος cvl χωρίζει τα δεδομένα με τυχαίο τρόπο σε δύο ομάδες, ώστε και οι δύο ομάδες να περιέχουν αρκετές παρατηρήσεις. Η πρώτη ομάδα καλείται training set και χρησιμοποιείται για την προσαρμογή του μοντέλου ενώ το δεύτερο σετ δεδομένων ονομάζεται validation set και γίνεται χρήση του για τον έλεγχο της προβλεπτικής ικανότητας του μοντέλου.

Πιο αποτελεσματική και απλή έκδοση της cross – validation είναι η leave one out cross validation η οποία κρατάει μόνο μία παρατήρηση για το validation set ενώ όλες οι υπόλοιπες παρατηρήσεις θα ανήκουν στο training set. Αν το δείγμα έχει μέγεθος n τότε το training set θα αποτελείται από n-1 παρατηρήσεις ενώ το validation set θα περιλαμβάνει μόνο μία. Προσαρμόζουμε το μοντέλο με τις n-1 παρατηρήσεις και ελέγχουμε την προβλεπτική του ικανότητα με τη 1 παρατήρηση. Επαναλαμβάνουμε την ίδια διαδικασία επιλέγοντας κάθε φορά διαφορετική παρατήρηση για το validation set μέχρι να τις χρησιμοποιήσουμε όλες.

Η μέθοδος υπολογίζει την εκτίμηση για το μέσο τετραγωνικό σφάλμα (MSE). Από τη στιγμή που την έχουμε επαναλάβει τόσες φορές όσες και το μέγεθος του δείγματος η τελική εκτίμηση του MSE θα είναι ίση με τη μέση τιμή των εκτιμήσεων των MSE που προέκυψαν από κάθε μία επανάληψη. Έτσι έχουμε:

$$CV_n = \frac{1}{n}\sum_{i=1}^{n} MSE_i = (y_i - \hat{y}_i)^2$$

Η μέθοδος αυτή είναι προτιμότερη από την cross validation διότι έχει μικρότερη μεροληψία συγκριτικά με την cross validation. Αυτό συμβαίνει γιατί το training set της μεθόδου cvl περιέχει σχεδόν τις μισές παρατηρήσεις από το πλήθος n παρατηρήσεων που έχουμε και προσαρμόζει το μοντέλο με βάση αυτές. Ακόμα η μέθοδος cross validation κάνει τυχαία το διαχωρισμό οπότε κάθε φορά δίνει διαφορετικά αποτελέσματα. Αντιθέτως η leave one out cross validation καταλήγει κάθε φορά στα ίδια αποτελέσματα.

ΚΕΦΑΛΑΙΟ 4. Δέντρα Επιβίωσης

4.1 Εισαγωγή στα δέντρα

Τα δέντρα αποφάσεων αποτελούν μία μέθοδο κατηγοριοποίησης (Classification) και συνδέονται άρρηκτα με τη μηχανική μάθηση (Machine Learning). Τα δέντρα αποφάσεων μπορούν να χρησιμοποιηθούν στην ταξινόμηση, δηλαδή στην πρόβλεψη που αφορά σε ποια τάξη ανήκουν τα δεδομένα. Τα δέντρα αυτά ονομάζονται δέντρα ταξινόμησης (Classification trees) και βρίσκουν εφαρμογή όταν η εξαρτημένη μεταβλητή που μελετούμε είναι διακριτή. Από την άλλη όταν η εξαρτημένη μεταβλητή είναι συνεχής κατασκευάζουμε δέντρα για την πρόβλεψη κάποιας συγκεκριμένης τιμής της εξαρτημένης μεταβλητής από τις ανεξάρτητες. Τα τελευταία δέντρα καλούνται δέντρα παλινδρόμησης (Regression trees). Τα δέντρα εισήχθησαν αρχικά από στους Morgan και Sonquist (1963), αλλά έγιναν πραγματικά δημοφιλή από στους Breiman et al (1984).

Τα δέντρα απόφασης ανήκουν στην κατηγορία της επαγωγικής μάθησης. Επαγωγική μάθηση είναι η αυτόματη μάθηση που χρησιμοποιείται για την εξαγωγή γενικών κανόνων από μία βάση δεδομένων, οι οποίοι θα μπορούν αργότερα να εφαρμοστούν προκειμένου να εξαχθούν συμπεράσματα για νέα δεδομένα. Η μεθοδολογία των δέντρων απόφασης είναι μία γενική μη παραμετρική τεχνική, που βασίζεται στη φιλοσοφία των αλγορίθμων διαφοροποίησης (recursive partitioning algorithms) και είναι ικανή να παράγει ταξινομητές προκειμένου να εκτιμήσει νέες, άγνωστες καταστάσεις, ή να αποκαλύψει στους μηχανισμούς που χαρακτηρίζουν ένα πρόβλημα.

Τα δέντρα απόφασης είναι κατασκευασμένα με δομή από πάνω στους τα κάτω. Ο πρώτος κόμβος του δέντρου ονομάζεται ρίζα και ακολουθούν οι ενδιάμεσοι κόμβοι. Ένα δέντρο αποφάσεων μπορεί να περιέχει μηδενικούς ενδιάμεσους κόμβους ή και περισσότερους (internal nodes) και έναν ή περισσότερους τερματικούς κόμβους/ φύλλα (leaf). Κάθε ενδιάμεσος κόμβος, που ονομάζεται αλλιώς και κόμβος ελέγχου, περιέχει έναν έλεγχο ο οποίος ελέγχει την τιμή της έκφρασης των χαρακτηριστικών και δημιουργεί δύο ή περισσότερους κόμβους-απογόνους (child nodes). Ο αριστερός απόγονος προκύπτει όταν επαληθεύεται ο έλεγχος, ενώ ο άλλος απόγονος δηλαδή ο δεξιός προκύπτει όταν ο έλεγχος δεν επαληθεύεται. Τέλος, ένας τερματικός κόμβος αποτελείται από μία τιμή τάξης. Οι τερματικοί κόμβοι είναι αυτοί που οδηγούν στην ταξινόμηση της εξεταζόμενης περίπτωσης σε μία από τις προκαθορισμένες κλάσεις (τάξεις).

Τα δέντρα επιβίωσης (Survival Trees) είναι μία μορφή δέντρων ταξινόμησης και παλινδρόμησης που προσαρμόζονται για να χειρίζονται τα αποκομμένα δεδομένα. Η βασική διαίσθηση πίσω από τα μοντέλα των δένδρων είναι η κατανομή των δεδομένων με βάση ένα συγκεκριμένο κριτήριο διαίρεσης έτσι ώστε τα αντικείμενα που είναι παρόμοια μεταξύ στους με βάση το γεγονός που μας ενδιαφέρει, να τοποθετηθούν στον ίδιο κόμβο. Η αρχική απόπειρα χρήσης μιας δομής δέντρου για δεδομένα επιβίωσης έγινε από στους Ciampi et αl. (1981), Marubini, Morabito και Valsecchi (1983). Ωστόσο, στους Gordon & Olshen (1985) ανήκει η πρώτη αναφορά στην οποία συζητήθηκε η δημιουργία δένδρων επιβίωσης.

4.2 Διαφορά δέντρων απόφασης και δέντρων επιβίωσης

Η βασική διαφορά ανάμεσα σε ένα δέντρο επιβίωσης και ένα τυποποιημένου δέντρο αποφάσεων είναι η επιλογή του κριτηρίου διαίρεσης. Η μέθοδος του δέντρου αποφάσεων εκτελεί αναδρομικό διαχωρισμό στα δεδομένα, ορίζοντας ένα κατώφλι για κάθε χαρακτηριστικό, παρόλα αυτά δεν μπορεί να εξετάσει, ούτε τις αλληλεπιδράσεις μεταξύ των χαρακτηριστικών, ούτε τις αποκομμένες πληροφορίες στο μοντέλο (Safavian & Landgrebe, 1991).

4.3 Κατασκευή των δέντρων επιβίωσης

Στόχος είναι η κατασκευή ενός βέλτιστου δέντρου, δηλαδή ενός δέντρου που θα πετυχαίνει τη μεγαλύτερη ισορροπία μεταξύ της πολυπλοκότητας και της ακρίβειας. Με τον όρο πολυπλοκότητα αναφερόμαστε στο συνολικό αριθμό κόμβων του δέντρου ενώ η ακρίβεια έχει να κάνει με την ικανότητα ταξινόμησης. Συνεπώς, το βέλτιστο δέντρο αρκεί να πετυχαίνει την ιδανική ισορροπία ανάμεσα στο συνολικό αριθμό των κόμβων και στην ικανότητα ταξινόμησης.

Η κατασκευή ενός δέντρου επιβίωσης ξεκινά από τον αρχικό κόμβο δηλαδή τη ρίζα. Η μέθοδος προσπαθεί να βρει την καλύτερη μεταβλητή που θα ανατεθεί στον κόμβο αυτόν, σε συνδυασμό πάντα με το βέλτιστο κανόνα διαχωρισμού. Για να συμβεί αυτό, δοκιμάζονται όλες οι μεταβλητές και όλες οι πιθανές τιμές που αντιστοιχούν στη μεταβλητή αυτή. Η επιλογή του καλύτερου διαχωριστή γίνεται βάση του συνάρτησης-κριτηρίου που εφαρμόζεται και στους δύο κόμβους που προκύπτουν και θα την αναλύσουμε εκτενέστερα στην επόμενη παράγραφο.

Η διαδικασία που περιγράφηκε παραπάνω επαναλαμβάνεται αναδρομικά με σκοπό να κατασκευαστούν τα αντίστοιχα υπόδεντρα, μέχρι το σχηματισμό του τελικού δέντρου.

4.3.1 Κριτήρια διαίρεσης για δέντρα επιβίωσης

Στα δέντρα απόφασης οι πιο συνηθισμένες συναρτήσεις διαχωρισμού που χρησιμοποιούνται στην περίπτωση κατηγορικής μεταβλητής απόκρισης είναι το κριτήριο Gini και η εντροπία. Αντιθέτως, στην περίπτωση συνεχούς μεταβλητής απόκρισης χρησιμοποιείται συχνότερα ως κριτήριο διαίρεσης το άθροισμα των τετραγωνικών αποκλίσεων από τον μέσο όρο. Με αντίστοιχο τρόπο έχουμε κριτήρια διαίρεσης για τα δέντρα επιβίωσης τα οποία αναλύονται στη συνέχεια. Τονίζουμε ότι πέρα από ότι αναλυθεί στην πορεία η πιο διαδομένη συνάρτηση διαχωρισμού είναι αυτή της Kaplan – Meier.

Για τα δέντρα επιβίωσης έχουμε δύο κατηγορίες στις οποίες μπορούν να ομαδοποιηθούν τα κριτήρια διαίρεσης. Η πρώτη κατηγορία στοχεύει στην μεγιστοποίηση της ετερογένειας ανάμεσα στους κόμβους, ενώ η δεύτερη στην ελαχιστοποίηση της ομοιογένειας εντός του κόμβου.

Στην πρώτη κατηγορία ελαχιστοποιείται η συνάρτηση απώλειας χρησιμοποιώντας το κριτήριο ομοιογένειας εντός του κόμβου. Για το λόγο αυτό οι Gordon και Olsen (1985) μετρούσαν τις αποστάσεις ομοιογένειας και Hellinger μεταξύ των εκτιμώμενων λειτουργιών διανομής χρησιμοποιώντας τη μετρική Wasserstein. Το 1989 οι Davis και Anderson χρησιμοποίησαν μία εκθετική συνάρτηση loglikelihood για αναδρομικό διαχωρισμό με βάση το άθροισμα υπολοίπων από το μοντέλο Cox. Τέλος οι Leblanc και Crowley (1992) μέτρησαν την απόκλιση του κόμβου με βάση το πρώτο βήμα μιας διαδικασίας εκτίμησης πλήρους πιθανοφάνειας.

Στη δεύτερη κατηγορία κριτηρίων διαίρεσης, χρησιμοποιήθηκαν στατιστικές δοκιμής log-rank για μέτρα ετερογένειας μεταξύ κόμβων (Ciampi et al., 1986). Αργότερα, οι Ciampi et al. (1987) πρότειναν μια στατιστική αναλογία πιθανοτήτων για να μετρηθεί η ανομοιότητα μεταξύ δύο κόμβων. Με βάση την τάξη στατιστικών δύο δειγμάτων οι Tarone-Ware, Segal (Segal, 1988) εισήγαγαν μια διαδικασία μέτρησης της μεταξύ τους σχέσης. Η βελτιστοποίηση του επιπέδου της παράδοσης είναι η ικανότητά του να χειρίζεται τα αποκομμένα δεδομένα χρησιμοποιώντας τη δομή του δέντρου.

4.3.2 Επιλογή τελικού δέντρου

Μια σημαντική πτυχή κατά την κατασκευή ενός δέντρου είναι να αποφασίσουμε πότε θα σταματήσουμε το διαχωρισμό και ως εκ τούτου να επιλέξουμε ένα συγκεκριμένο δέντρο ως το τελικό μοντέλο. Εάν είναι πολύ μεγάλο το δέντρο, θα τείνει να ταιριάζουν υπερβολικά τα δεδομένα και έτσι θα αποτυγχάνει να γενικευτεί καλά στον πληθυσμό που μας ενδιαφέρει. Από την άλλη μεριά εάν το δέντρο είναι πολύ μικρό, μπορεί να χάνει σημαντικά χαρακτηριστικά της σχέσης μεταξύ των συμμεταβλητών και του αποτελέσματος. Υπάρχουν δύο βασικές προσεγγίσεις για την επιλογή ενός τελικού δέντρου. Η πρώτη είναι μια προς τα πίσω μέθοδος που δημιουργεί ένα μεγάλο δέντρο και στη συνέχεια επιλέγει ένα κατάλληλο υπόδεντρο κλαδεύοντας μερικά από τα κλαδιά του. Η δεύτερη είναι μια προς τα μπροστά μέθοδος που χρησιμοποιεί έναν ενσωματωμένο κανόνα διακοπής για να αποφασίσει πότε θα σταματήσει να χωρίζει έναν κόμβο περαιτέρω.

ΚΕΦΑΛΑΙΟ 5. Μελέτη Περίπτωσης

5.1 Περιγραφή δεδομένων

Τα δεδομένα με τα οποία ασχοληθήκαμε στην μελέτη μας λήφθηκαν από τους Krall, Uthoff και Harley το 1975 και αφορούν 48 άτομα με ηλικίες μεταξύ των 50 και 80 ετών. Η έρευνα πραγματοποιήθηκε από το ιατρικό κέντρο του Πανεπιστημίου της Δυτικής Βιρτζίνια στις Ηνωμένες Πολιτείες Αμερικής προκειμένου να διερευνηθεί η συσχέτιση που υπάρχει ανάμεσα σε ορισμένες μεταβλητές, που αφορούν χαρακτηριστικά των ασθενών και αποτελέσματα των εξετάσεων τους, και στον χρόνο επιβίωσης αυτών. Αντικείμενο της μελέτης μας αποτελεί ο χρόνος επιβίωσης ασθενών που πάσχουν από πολλαπλό μυέλωμα, από την ημέρα της διάγνωσής τους με την ασθένεια αυτή. Συμπερασματικά η μεταβλητή απόκρισης είναι ο χρόνος σε μήνες που μεσολαβεί από τη διάγνωση ως και το θάνατο των ασθενών.

5.1.1 Πολλαπλό μυέλωμα

Το Πολλαπλό Μυέλωμα (Multiple Myeloma), που περιγράφηκε για πρώτη φορά το 1848, είναι μια νεοπλασματική διαταραχή πλασματοκυττάρων που χαρακτηρίζεται από κλωνικό πολλαπλασιασμό των κακοήθων πλασματοκυττάρων στο μικροπεριβάλλον του μυελού των οστών, από μονοκλωνική πρωτεΐνη στο αίμα ή στα ούρα, και δυσλειτουργία των συναφών οργάνων. Τα πλασματοκύτταρα , είναι κύτταρα που υπό φυσιολογικές συνθήκες παράγουν τα αντισώματα, πρωτεΐνες δηλαδή που είναι ειδικευμένες για να προσκολλώνται σε διαφορετικό για κάθε ένα αντιγόνο, όπως σε αντιγόνα βακτηρίων, ιών και άλλων μικροοργανισμών με αποτέλεσμα να απενεργοποιούν αυτά τα αντιγόνα ή να στρέφουν άλλα κύτταρα του ανοσοποιητικού έναντι αυτών των εχθρικών αντιγόνων που βρίσκονται στην επιφάνεια των ιών ή των μικροβίων. Στην περίπτωση του πολλαπλού μυελώματος, τα πλασματοκύτταρα υφίσταται κακοήθη μετάλλαξη. Τα κακοήθη πλασματοκύτταρα παράγουν αντισώματα ή και απλά τμήματα αντισωμάτων, τις παραπρωτεΐνες. Παράγονται από έναν κλώνο πλασματοκυττάρων και έτσι λέγονται μονοκλωνικές πρωτεΐνες (Μ πρωτεΐνες), γιατί όλα τα πλασματοκύτταρα παράγουν την ίδια παραπρωτεΐνη. Αυτά τα αντισώματα έχουν όλα τις ίδιες φυσικοχημικές ιδιότητες.

Η παραγωγή μεγάλων ποσοτήτων παραπρωτεϊνών, και κυρίως ορισμένων τμημάτων που λέγονται ελαφρές αλυσίδες, μπορεί να προκαλέσει απόφραξη των νεφρικών σωληνάριων, με συνέπεια την βλάβη της λειτουργίας των νεφρών και ως αποτέλεσμα τη νεφρική ανεπάρκεια. Ακόμα, εμποδίζει τον σχηματισμό φυσιολογικών αντισωμάτων, καθιστώντας τον ασθενή πιο επιρρεπή σε λοιμώξεις. Στα συμπτώματα περιλαμβάνονται ακόμα πόνοι στα οστά, την μέση, κατάγματα, υπερασβεστιαιμία, αναιμία και νεφρική ανεπάρκεια. Η υπερασβεστιαιμία συνήθως προκαλεί διάφορες επιπλοκές, όπως βλάβες των νεφρών, κόπωση, σύγχυση, καρδιακές αρρυθμίες, ναυτία και έμετο. Αν και η νόσος παρατηρείται πιο συχνά μεταξύ συγγενών το πολλαπλό μυέλωμα δεν αποτελεί μια κληρονομική ασθένεια με την στενή έννοια του όρου.

Αντιπροσωπεύει περίπου το 1% των νεοπλασματικών ασθενειών και το 13% των αιματολογικών καρκίνων. Στις δυτικές χώρες, η ετήσια εμφάνιση της ασθένειας προσαρμοσμένη με κριτήριο την ηλικία είναι 5,6 περιπτώσεις ανά 100.000 άτομα. Η μέση ηλικία κατά τη διάγνωση είναι περίπου 70 χρόνια. Από τις 40 χώρες της Ευρώπης, η Ελλάδα κατέχει την 26η θέση στο ρυθμό εμφάνισης νέων περιστατικών (2.9 ανά 100.000 άτομα πληθυσμού).

5.1.2 Το δείγμα

Το παρόν δείγμα αποτελείται από 48 άτομα ηλικίας από 50 έως και 80 ετών που πάσχουν από πολλαπλό μελάνωμα. Κάποιοι από τους ασθενείς δεν είχαν πεθάνει μέχρι τη στιγμή της διεξαγωγής της έρευνας και για το λόγο αυτό θεωρήθηκαν ως δεξιά αποκομμένες παρατηρήσεις. Μέρος του δείγματος παρουσιάζεται στον Πίνακα 5.1. Η πηγή του δείγματός μας είναι Collett (2003).

patient	time	status	age	sex	bun	ca	hb	pcells	protein
1	13	1	66	1	25	10	14.6	18	1
2	52	0	66	1	13	11	12.0	100	0
3	6	1	53	2	15	13	11.4	33	1
4	40	1	69	1	10	10	10.2	30	1
5	10	1	65	1	20	10	13.2	66	0
6	7	0	57	2	12	8	9.9	45	0
7	66	1	52	1	21	10	12.8	11	1
8	10	0	60	1	41	9	14.0	70	1
9	10	1	70	1	37	12	7.5	47	0
10	14	1	70	1	40	11	10.6	27	0
11	16	1	68	1	39	10	11.2	41	0
12	4	1	50	2	172	9	10.1	46	1
13	65	1	59	1	28	9	6.6	66	0
14	5	1	60	1	13	10	9.7	25	0
15	11	0	66	2	25	9	8.8	23	0
16	10	1	51	2	12	9	9.6	80	0
17	15	0	55	1	14	9	13.0	8	0
18	5	1	67	2	26	8	10.4	49	0
19	76	0	60	1	12	12	14.0	9	0
20	56	0	66	1	18	11	12.5	90	0

Πίνακας 5.1: Μέρος του δείγματος

Η μεταβλητή Status περιγράφει την κατάσταση του ασθενή και είναι κωδικοποιημένη με 0 για τα άτομα των οποίων ο θάνατος δεν έχει ακόμα επέλθει (άρα οι παρατηρήσεις αυτές είναι αποκομμένες) και 1 για τους ασθενείς που έχουν φύγει από τη ζωή, οπότε αυτές οι παρατηρήσεις δεν είναι αποκομμένες.

Οι επεξηγηματικές μεταβλητές είναι επτά εκ των οποίων οι δύο αφορούν τα βασικά χαρακτηριστικά των ασθενών και είναι το φύλο «sex» και η ηλικία «age». Το φύλο, που είναι κατηγορική μεταβλητή, λαμβάνει την τιμή 1 για τους άντρες και την τιμή 0 για τις γυναίκες. Οι υπόλοιπες πέντε μεταβλητές έχουν να κάνουν με ευρήματα από ειδικές εξετάσεις των ασθενών τόσο ποσοτικές όσο και κατηγορικές.

Πιο συγκεκριμένα, μία από τις ποσοτικές επεξηγηματικές μεταβλητές αφορούν τα επίπεδα του αζώτου ουρίας αίματος «bun» (blood urea nitrogen). Το άζωτο ουρίας αποτελεί μέρος της ουρίας, μιας ουσίας που σχηματίζεται στο ήπαρ μέσω της ενζυμικής διαδικασίας καταβολισμού των πρωτεϊνών. Η ουρία φυσιολογικά διηθείται ελεύθερα μέσω των νεφρικών σπειραμάτων, με ένα μικρό ποσοστό να επαναρροφάται στα σωληνάρια και το υπόλοιπο να απεκκρίνεται στα ούρα. Η παθολογική κατάσταση που υπάρχουν αυξημένα επίπεδα αζώτου ουρίας στο αίμα ονομάζεται αζωθαιμία και υποδεικνύει υπολειτουργία των νεφρών. Υπερβολικά αυξημένα επίπεδα αζώτου ουρίας στην αιματική κυκλοφορία μπορεί να αποτελούν ένδειξη νεφρικού ή ηπατικού προβλήματος.

Η μέτρηση του ασβεστίου «ca» (serum calcium) αποτελεί άλλη μία ποσοτική μεταβλητή και αναφέρεται στην ποσότητα του ασβεστίου στον οργανισμό του ασθενούς. Η μέτρηση του ασβεστίου στο αίμα χρησιμοποιείται για τη διάγνωση και την παρακολούθηση ενός μεγάλου φάσματος διαταραχών, συμπεριλαμβανομένων νοσημάτων των οστών, των νεφρών, των παραθυρεοειδών αδένων και του γαστρεντερικού συστήματος.

Άλλη ποσοτική μεταβλητή είναι η τιμή της αιμοσφαιρίνης «hb» (haemoglobin) και είναι η χρωστική ουσία των ερυθροκυττάρων που μεταφέρει το οξυγόνο. Αποτελείται από αμινοξέα που σχηματίζουν μια ενιαία πρωτεΐνη που ονομάζεται σφαιρίνη και μία ένωση που ονομάζεται αίμη. Η αίμη περιέχει άτομα σιδήρου και την κόκκινη χρωστική πορφυρίνη. Κάθε ερυθρό αιμοσφαίριο περιέχει περίπου 300 εκατομμύρια μόρια αιμοσφαιρίνης.

Τέλος, το ποσοστό κυττάρων πλάσματος στο μυελό των οστών αποτελεί την τελευταία ποσοτική μεταβλητή «cells». Όσο μεγαλύτερο το ποσοστό των κυττάρων αυτών τόσο πιο δυσμενής και η κατάσταση για τον ασθενή καθώς υποδεικνύει μεγάλο αριθμό καρκινικών κυττάρων.

Όσον αναφορά τις κατηγορικές επεξηγηματικές μεταβλητές που σχετίζονται με αποτελέσματα εξετάσεων έχουμε μόνο μία και έχει να κάνει με την παρουσία της μονοκλωνικής πρωτεΐνης στον ορό της ουρίας «protein». Η μεταβλητή λαμβάνει την τιμή 1 όταν υπάρχει η πρωτεΐνη στον ορό της ουρίας και 0 όταν αυτή δεν υπάρχει.

Σκοπός της μελέτης είναι να εξετάσουμε κατά πόσο στατιστικά σημαντικές είναι οι ανεξάρτητες μεταβλητές «bun», «ca», «hb», «pcells» και «protein» στον χρόνο επιβίωσης των ασθενών που πάσχουν από πολλαπλό μυέλωμα. Ο χρόνος επιβίωσης των ασθενών ενδέχεται να επηρεάζεται τόσο από το φύλο των ασθενών, όσο και από την ηλικία τους. Ακόμα καθοριστικό ρόλο στην πορεία της υγείας τους ενδέχεται να διαδραματίζει η ύπαρξη ή όχι της μονοκλωνικής πρωτεΐνης στον ορό του αίματός τους. Για το λόγο αυτό θα εξετάσουμε αν το φύλο, η ηλικία και η ύπαρξη της πρωτεΐνης επηρεάζουν τη διάρκεια ζωής του ασθενούς. Με τον ίδιο τρόπο θα εργαστούμε και για τις υπόλοιπες ανεξάρτητες μεταβλητές του δείγματος.

Για να πετύχουμε το στόχο μας χρησιμοποιήσαμε μεθόδους ανάλυσης επιβίωσης με τη βοήθεια της R. Ξεκινήσαμε από τους μη παραμετρικούς ελέγχους για να βγάλουμε τα αρχικά συμπεράσματα και στη συνέχεια, προσαρμόσαμε το ημιπαραμετρκό μοντέλο του Cox στην R. Χρησιμοποιήσαμε ημιπαραμετρικό μοντέλο καθώς δεν γνωρίζουμε την κατανομή που ακολουθούν τα δεδομένα μας καθώς η ανταπόκριση των ασθενών σε κάθε ασθένεια αλλά και οι μεταβλητές που επηρεάζουν την πορεία της υγείας τους δεν είναι πλήρως καθορισμένες.

Αφού προσαρμοστεί το μοντέλο του Cox στα δεδομένα μας, εφαρμόζουμε γραφικούς και αναλυτικούς ελέγχους για να επαληθεύσουμε αν ισχύει η υπόθεση της αναλογικής διακινδύνευσης. Επιπλέον, εκτελούμε τους ελέγχους Wald και του λόγου των πιθανοφανειών ώστε να πάρουμε μια εκτίμηση των πιο στατιστικά σημαντικών μεταβλητών που συμμετέχουν στο μοντέλο και χρησιμοποιούμε τους αλγορίθμους επιλογής μεταβλητών με βήματα (backward elimination) ώστε να καταλήξουμε στο βέλτιστο δυνατό μοντέλο που να περιγράφει τα δεδομένα μας. Με τη βοήθεια των καμπύλων ROC θα ελέγξουμε την προβλεπτική ικανότητα του βέλτιστου μοντέλου στο οποίο καταλήξαμε.

Στη συνέχεια της μελέτης μας θα χρησιμοποιήσουμε τις διαφορετικές μεθόδους ποινών για να αντιμετωπίσουμε το φαινόμενο της πολυσυγγραμμικότητας. Πιο συγκεκριμένα θα ασχοληθούμε με τις μεθόδους: Ridge, Lasso και Elastic Net.

Τέλος θα εφαρμόσουμε τα δέντρα επιβίωσης ώστε λαμβάνοντας αποτελέσματα για τη στατιστική σημαντικότητα των ανεξάρτητων μεταβλητών να είμαστε πλέον σε θέση να συγκρίνουμε τα αποτελέσματα που προκύπτουν από τη χρήση διαφόρων μεθόδων.

5.2 Εκτίμηση Kaplan – Meier

Για να ξεκινήσουμε την ανάλυση του συνόλου δεδομένων επιβίωσης που μας δόθηκαν θα παρουσιάσουμε κάποια γραφικά αποτελέσματα για τους χρόνους επιβίωσης των ατόμων που βρίσκονται σε ένα γκρουπ με την βοήθεια των εκτιμήσεων της συνάρτησης επιβίωσης Kaplan - Meier. Τα αποτελέσματα αυτά προκύπτουν εύκολα από τις εκτιμήσεις των συναρτήσεων επιβίωσης. Στην συνέχεια θα συγκρίνουμε δύο διαφορετικές ομάδες ασθενών. Μπορούμε να πραγματοποιήσουμε μια σύγκριση της επιβίωσης για κάθε ομάδα χρησιμοποιώντας τις εκτιμήσεις των συναρτήσεων επιβίωσης με την βοήθεια του ελέγχου log-rank και του ελέγχου Wilcoxon. Σε πρώτη φάση θα κάνουμε μία εκτίμηση της Kaplan-Meier για όλο το δείγμα και έπειτα για κάθε ανεξάρτητη μεταβλητή θα χωρίσουμε τις παρατηρήσεις σε δύο ομάδες και θα ελέγξουμε αν υπάρχει διαφοροποίηση ή όχι στην επιβίωση ανάμεσα στις δύο ομάδες. Αν υπάρχει διαφορά ανάμεσα στις δύο ομάδες θα λέμε ότι η συγκεκριμένη μεταβλητή είναι στατιστικά σημαντική και άρα επηρεάζει το χρόνο επιβίωσης των ασθενών που πάσχουν από πολλαπλό μυέλωμα. Σε αντίθετη περίπτωση η μεταβλητή θα είναι στατιστικά μη σημαντική.

Ξεκινώντας με την εκτίμηση Kaplan – Meier της συνάρτησης επιβίωσης λαμβάνουμε τα αποτελέσματα του Πίνακα 5.2.

```
kaplan_meier_fit1 <- survfit (Surv(ti me, status) ~ 1, data=data)
summary (kaplan_meier_fit1)
```

Στον Πίνακα 5.2 η πρώτη στήλη time έχει να κάνει με τη χρονική στιγμή στην οποία αναφερόμαστε, η δεύτερη στήλη n.risk αφορά το πλήθος των ατόμων που κινδυνεύουν να τους συμβεί το γεγονός δηλαδή να φύγουν από τη ζωή. Η στήλη n.event δείχνει το πλήθος των ασθενών στους οποίους συνέβη το γεγονός τη χρονική στιγμή time. Οι τιμές της εκτιμήτριας Kaplan – Meier φαίνονται στην 4η στήλη η οποία και καλείται survival. Τέλος δίνονται τα τυπικά σφάλματα των εκτιμήσεων στη στήλη std.err και στις δύο τελευταίες στήλες τα 95% διαστήματα εμπιστοσύνης για τις εκτιμήσεις.

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	48	3	0.9375	0.0349	0.8715	1.000
4	44	2	0.8949	0.0445	0.8118	0.986
5	42	4	0.8097	0.0571	0.7051	0.930
6	38	2	0.7670	0.0616	0.6554	0.898
8	35	1	0.7451	0.0636	0.6304	0.881
10	34	4	0.6575	0.0696	0.5343	0.809
12	28	1	0.6340	0.0710	0.5091	0.789
13	26	1	0.6096	0.0723	0.4832	0.769
14	25	1	0.5852	0.0734	0.4577	0.748
15	24	1	0.5608	0.0743	0.4326	0.727
16	22	2	0.5098	0.0758	0.3810	0.682
17	20	1	0.4844	0.0762	0.3559	0.659
18	19	2	0.4334	0.0762	0.3071	0.612
23	15	1	0.4045	0.0764	0.2793	0.586
24	14	1	0.3756	0.0762	0.2524	0.559
36	13	1	0.3467	0.0756	0.2261	0.532
40	12	2	0.2889	0.0732	0.1758	0.475
50	9	1	0.2568	0.0718	0.1485	0.444
51	8	1	0.2247	0.0696	0.1224	0.412
65	5	1	0.1798	0.0687	0.0850	0.380
66	4	1	0.1348	0.0646	0.0527	0.345
88	2	1	0.0674	0.0576	0.0126	0.359
91	1	1	0.0000	NaN	NA	NA

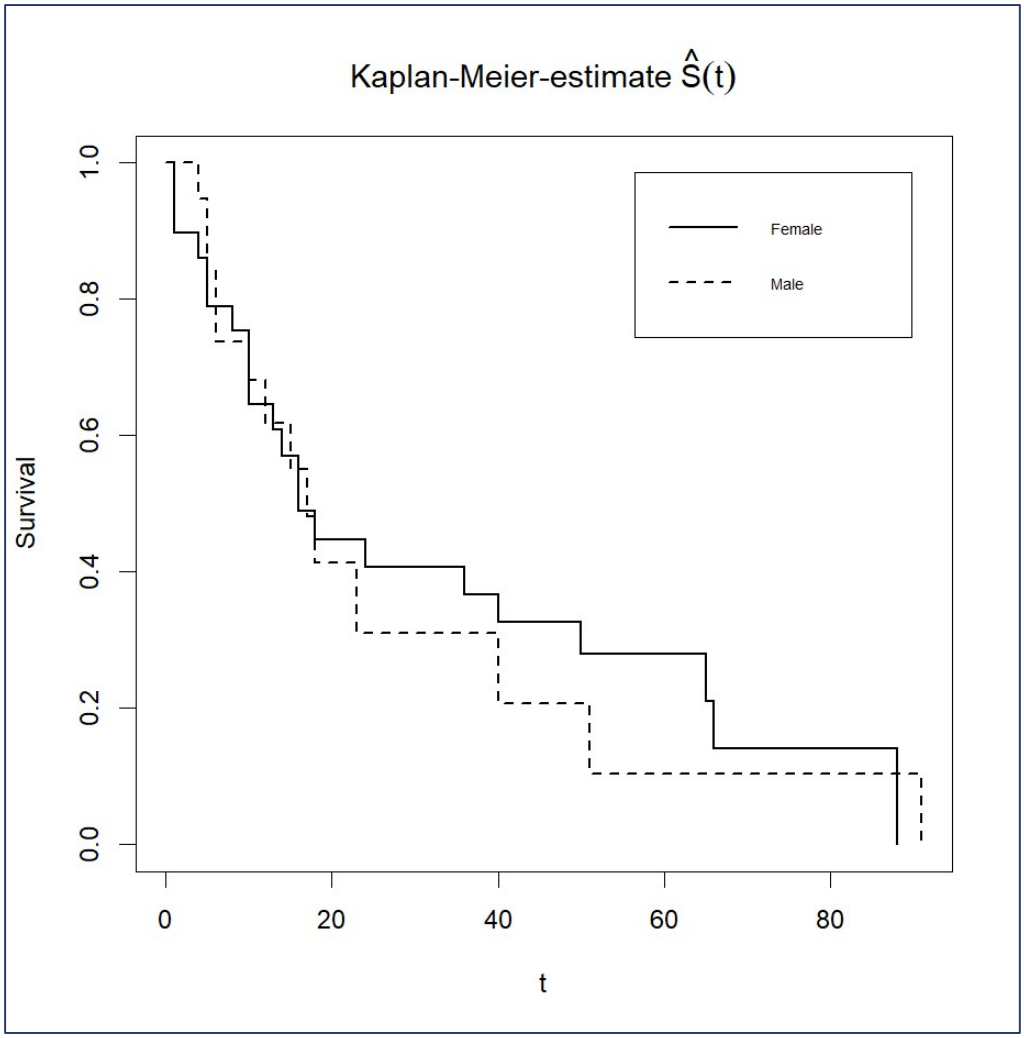
Πίνακας 5.2 : Εκτιμήσεις Kaplan-Meier της συνάρτησης επιβίωσης

Από τον Πίνακα 5.2 βλέπουμε για παράδειγμα ότι το 10ο μήνα 34 άτομα βρίσκονταν σε κίνδυνο να τους συμβεί το γεγονός, δηλαδή να πεθάνουν, ενώ 4 άτομα έφυγαν από τη ζωή. Ακόμα το μήνα εκείνο η εκτίμηση της συνάρτησης επιβίωσης είναι ίση με 0.6575.

5.2.1 Σύγκριση των εκτιμήσεων Kaplan-Meier μεταξύ ανδρών και γυναικών

Αρχικά ελέγχουμε γραφικά αν το φύλο επιδρά στη διάρκεια ζωής των ασθενών. Οι εντολές που χρησιμοποιήθηκαν ήταν οι ακόλουθες:

```
km_sex <- survfit(Surv(time, status) ~ sex, type ="kaplan-meier", data=data)

plot (km_sex, lty = 1:2, main = expression(paste("Kaplan-Meier-estimate ",
hat(S)(t))), xlab="t", ylab="Survival", lwd=1.5)
legend ("topright", inset=0.05,   c("Female", "Male"), lty = 1:2, lwd=1.5, cex=0.6)
```



Διάγραμμα 5.1: Σύγκριση των εκτιμήσεων Kaplan-Meier μεταξύ ανδρών και γυναικών

Στο Διάγραμμα 5.1 συγκρίνουμε γραφικά τις εκτιμήσεις Kaplan-Meier των δύο φύλων και παρατηρούμε ότι δεν υπάρχει μεγάλη διαφοροποίηση ανάμεσα στα δύο φύλα. Έτσι συμπεραίνουμε ότι το φύλο δεν επηρεάζει την διάρκεια επιβίωσης των ασθενών που πάσχουν από πολλαπλό μυέλωμα.

Στη συνέχεια, υπολογίζουμε ξεχωριστά τις εκτιμήσεις Kaplan-Meier της συνάρτησης επιβίωσης για τα δύο φύλα (γυναίκες και άντρες) και ελέγχουμε αν υπάρχουν διαφορές μεταξύ αυτών των δύο φύλων. Για να λάβουμε τα συμπεράσματα αυτά χρησιμοποιήσαμε την εντολή:

```
summary(km_sex)
```

Στα αποτελέσματα στους Πίνακες 5.3 και 5.4 παρατηρούμε τις εκτιμήσεις Kaplan-Meier, τα τυπικά σφάλματα και τα διαστήματα εμπιστοσύνης των συναρτήσεων επιβίωσης για τα δύο φύλα. Συγκεκριμένα στον Πίνακα 5.3 έχουμε λάβει αποτελέσματα σχετικά με το γυναικείο φύλλο, ενώ στον Πίνακα 5.4 παρουσιάζονται τα αποτελέσματα που αφορούν τους άντρες.

sex=1

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	29	3	0.897	0.0566	0.7923	1.000
4	25	1	0.861	0.0647	0.7428	0.997
5	24	2	0.789	0.0766	0.6522	0.954
8	22	1	0.753	0.0811	0.6098	0.930
10	21	3	0.646	0.0902	0.4908	0.849
13	17	1	0.608	0.0926	0.4507	0.819
14	16	1	0.570	0.0942	0.4118	0.788
16	14	2	0.488	0.0968	0.3311	0.720
18	12	1	0.448	0.0969	0.2928	0.684
24	11	1	0.407	0.0962	0.2559	0.647
36	10	1	0.366	0.0948	0.2204	0.608
40	9	1	0.325	0.0926	0.1864	0.568
50	7	1	0.279	0.0903	0.1479	0.526
65	4	1	0.209	0.0907	0.0894	0.490
66	3	1	0.139	0.0831	0.0434	0.448
88	1	1	0.000	NaN	NA	NA

Πίνακας 5.3: Εκτιμήσεις Kaplan-Meier της συνάρτησης επιβίωσης για τις γυναίκες

sex=2

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
4	19	1	0.947	0.0512	0.8521	1.000
5	18	2	0.842	0.0837	0.6931	1.000
6	16	2	0.737	0.1010	0.5632	0.964
10	13	1	0.680	0.1080	0.4983	0.928
12	11	1	0.618	0.1145	0.4301	0.889
15	9	1	0.550	0.1207	0.3574	0.845
17	8	1	0.481	0.1236	0.2906	0.796
18	7	1	0.412	0.1236	0.2291	0.742
23	4	1	0.309	0.1287	0.1368	0.699
40	3	1	0.206	0.1202	0.0657	0.646
51	2	1	0.103	0.0944	0.0171	0.621
91	1	1	0.000	NaN	NA	NA

Πίνακας 5.4: Εκτιμήσεις Kaplan-Meier της συνάρτησης επιβίωσης για τους άντρες

Το συμπέρασμα που βγάλαμε νωρίτερα επιβεβαιώνεται από στους Πίνακες 5.3 και 5.4 και άρα το φύλο δεν επηρεάζει το χρόνο ζωής των ασθενών.

Ένας τελευταίος τρόπος για να στηρίξουμε το συμπέρασμά μας είναι ο έλεγχος log-rank και ο έλεγχος του Wilcoxon. Οι εντολές που χρησιμοποιήσαμε είναι οι ακόλουθες.

```
Log_rank <- survdiff(Surv(time,status) ~ sex)
log_rank

wilcoxon <- survdiff(Surv(time,status) ~ sex,rho=1)
Wilcoxon
```

Και η p-value που λάβαμε από τον έλεγχο log-rank είναι ίση με p=0,8 δεχόμαστε τη μηδενική υπόθεση και άρα δεν υπάρχει διαφοροποίηση στους χρόνους επιβίωσης ανάμεσα στα δύο φύλλα. Σε κάτι αντίστοιχο καταλήγουμε και με τον έλεγχο του Wilcoxon αφού λάβαμε p-value=1. Οπότε έχουμε αρκετές ενδείξεις για να θεωρήσουμε τη μεταβλητή sex στατιστικά μη σημαντική.

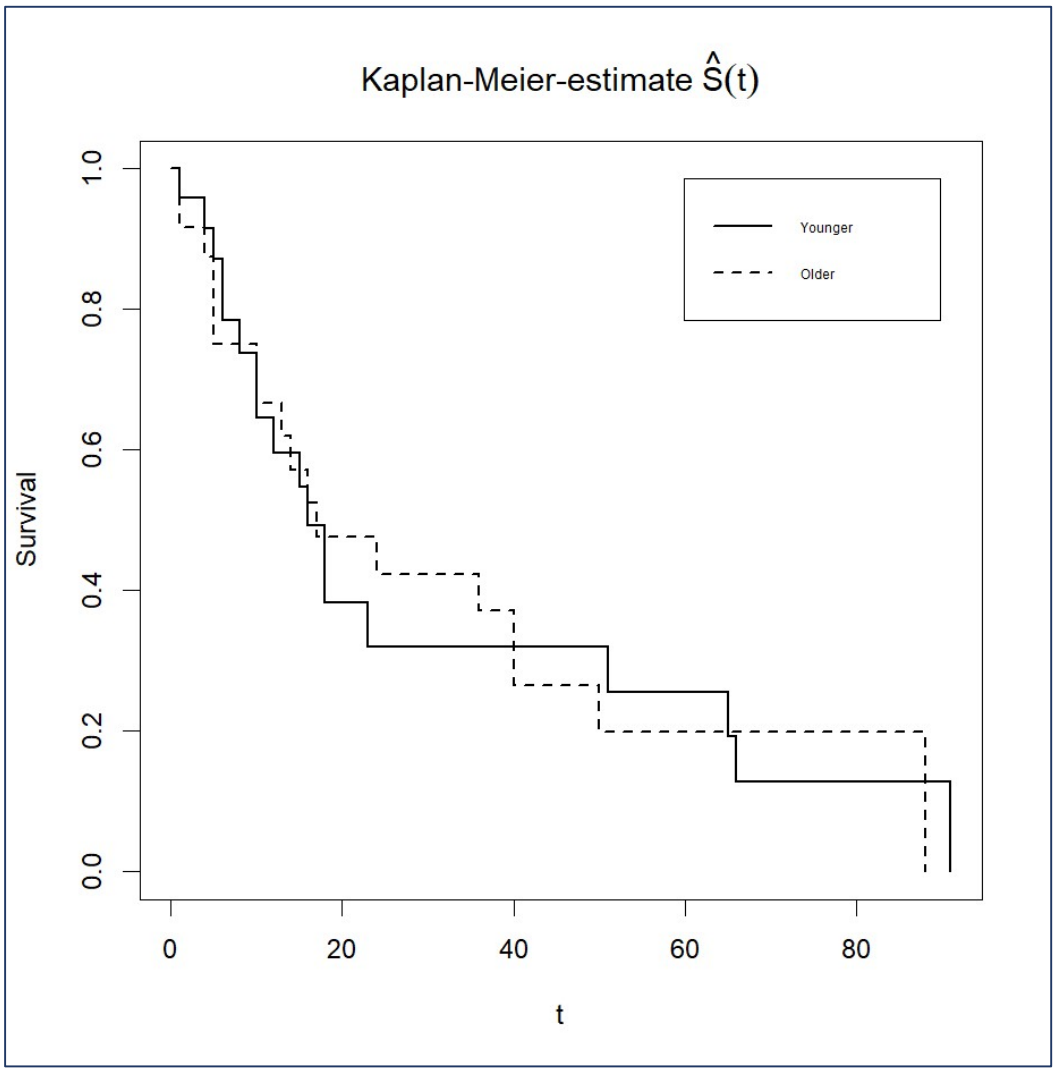
5.2.2 Σύγκριση των εκτιμήσεων Kaplan-Meier μεταξύ νεότερων και μεγαλύτερων ασθενών

Ελέγχουμε γραφικά αν η ηλικία επιδρά στη διάρκεια ζωής των ασθενών. Για να μπορέσουμε να κατασκευάσουμε τη γραφική αυτή παράσταση χωρίζουμε τους ασθενείς σε δύο κατηγορίες. Το μέτρο που χρησιμοποιήσαμε για το διαχωρισμό αυτό ήταν η μέση τιμή της ηλικίας των ασθενών καθώς δεν έχουμε ακραίες παρατηρήσεις για να χρειαστεί να χρησιμοποιήσουμε τη διάμεσο. Η δειγματικός μέσος της ηλικίας των ασθενών βρέθηκε με τη βοήθεια της εντολής sum στην R ίση με 62.89583. Έτσι ομαδοποιήσαμε τους ασθενείς σε δύο κατηγορίες. Η πρώτη κατηγορία περιλάμβανε τα άτομα που είχαν μικρότερη ηλικία από το δειγματικό μέσο ενώ η δεύτερη ομάδα αποτελούταν από τους ασθενείς που είχαν ηλικία μεγαλύτερη από τη μέση τιμή. Στην πρώτη ομάδα δόθηκε η κωδικοποίηση 0 όσον αναφορά την ηλικία και στην δεύτερη η τιμή 1. Έτσι έχουμε:

$$age = \begin{cases} 0, & \text{όταν } age \leq 63 \\ 1, & \text{όταν } age > 63 \end{cases}$$

Πλέον είμαστε έτοιμοι να λάβουμε τη γραφική παράσταση από την οποία θα εξάγουμε σημαντικά για τη μελέτη μας αποτελέσματα. Οι εντολές που χρησιμοποιήθηκαν ήταν οι ακόλουθες:

```
km_age <- survfit(Surv(time, status)~data2[,4], type="kaplan-meier",data=data2)

plot(km_age, lty = 1:2, main=expression(paste("Kaplan-Meier-estimate ",
hat(S)(t))), xlab="t", ylab="Survival", lwd=1.5)

legend("topright", inset=0.05,   c("Younger", "Older"), lty = 1:2, lwd=1.5, cex=0.6)
```



Διάγραμμα 5.2: Σύγκριση των εκτιμήσεων Kaplan-Meier μεταξύ νεότερων και μεγαλύτερων ασθενών

Στο Διάγραμμα 5.2 συγκρίνουμε γραφικά τις εκτιμήσεις Kaplan-Meier μεταξύ των δύο ηλικιακών ομάδων και παρατηρούμε ότι δεν υπάρχει μεγάλη διαφοροποίηση ανάμεσα στις δύο αυτές κατηγορίες. Αυτό γίνεται φανερό από το γεγονός ότι σε ορισμένα σημεία του διαγράμματος η γραφική της πρώτης ομάδας (δηλαδή των ατόμων που είναι μικρότερα από τα 63 έτη) βρίσκεται πάνω από τη γραφική της δεύτερης ομάδας (δηλαδή των ατόμων που είναι μεγαλύτερα από τα 63 έτη) την ίδια στιγμή που σε άλλα διαστήματα του Διαγράμματος συμβαίνει το αντίθετο. Οπότε δεν μπορούμε να καταλήξουμε σε κάποιο συμπέρασμα για το ποια από τις δύο ηλικιακές ομάδες έχει μεγαλύτερο προσδόκιμο ζωής. Έτσι συμπεραίνουμε ότι η ηλικία δεν επηρεάζει την διάρκεια επιβίωσης των ασθενών που πάσχουν από πολλαπλό μυέλωμα.

Στη συνέχεια, υπολογίζουμε ξεχωριστά τις εκτιμήσεις Kaplan-Meier της συνάρτησης επιβίωσης για τις δύο ηλικιακές ομάδες και ελέγχουμε αν υπάρχουν διαφορές μεταξύ αυτών των δύο κατηγοριών. Για να λάβουμε τα συμπεράσματα αυτά χρησιμοποιήσαμε την εντολή:

```
summary(km_age)
```

Στα αποτελέσματα στους Πίνακες 5.5 και 5.6 παρατηρούμε τις εκτιμήσεις Kaplan-Meier, τα τυπικά σφάλματα και τα διαστήματα εμπιστοσύνης των συναρτήσεων επιβίωσης για τις δύο ηλικιακές ομάδες. Συγκεκριμένα στον Πίνακα 5.6 έχουμε λάβει αποτελέσματα σχετικά με τα νεότερα άτομα, ενώ στον Πίνακα 5.7 παρουσιάζονται τα αποτελέσματα που αφορούν τους μεγαλύτερους ασθενείς.

data2[, 4]=0

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	24	1	0.958	0.0408	0.8816	1.000
4	22	1	0.915	0.0577	0.8084	1.000
5	21	1	0.871	0.0695	0.7452	1.000
6	20	2	0.784	0.0856	0.6331	0.971
8	17	1	0.738	0.0921	0.5778	0.943
10	16	2	0.646	0.1011	0.4751	0.878
12	13	1	0.596	0.1048	0.4223	0.841
15	12	1	0.546	0.1072	0.3719	0.803
16	10	1	0.492	0.1095	0.3178	0.761
18	9	2	0.382	0.1091	0.2187	0.669
23	6	1	0.319	0.1079	0.1641	0.619
51	5	1	0.255	0.1035	0.1151	0.565
65	4	1	0.191	0.0952	0.0721	0.508
66	3	1	0.127	0.0821	0.0361	0.450
91	1	1	0.000	NaN	NA	NA

Πίνακας 5.5: Εκτιμήσεις Kaplan-Meier της συνάρτησης επιβίωσης για τα νεότερα άτομα

data2[, 4]=1

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	24	2	0.917	0.0564	0.8125	1.000
4	22	1	0.875	0.0675	0.7522	1.000
5	21	3	0.750	0.0884	0.5953	0.945
10	18	2	0.667	0.0962	0.5024	0.885
13	14	1	0.619	0.1004	0.4504	0.851
14	13	1	0.571	0.1034	0.4008	0.815
16	12	1	0.524	0.1052	0.3534	0.776
17	11	1	0.476	0.1058	0.3080	0.736
24	9	1	0.423	0.1065	0.2585	0.693
36	8	1	0.370	0.1055	0.2119	0.647
40	7	2	0.265	0.0984	0.1276	0.548
50	4	1	0.198	0.0934	0.0789	0.499
88	1	1	0.000	NaN	NA	NA

Πίνακας 5.6: Εκτιμήσεις Kaplan-Meier της συνάρτησης επιβίωσης για τα μεγαλύτερα άτομα

Το συμπέρασμα που βγάλαμε νωρίτερα επιβεβαιώνεται από τους Πίνακες 5.5 και 5.6 και άρα η ηλικία δεν επηρεάζει το χρόνο ζωής των ασθενών.

Ένας τελευταίος τρόπος για να στηρίξουμε το συμπέρασμά μας είναι ο έλεγχος log-rank και ο έλεγχος του Wilcoxon. Οι εντολές που χρησιμοποιήσαμε είναι οι ακόλουθες.

```
log_rank2 <- survdiff(Surv(time,status) ~ data2[,4])
log_rank2

wilcoxon2 <- survdiff(Surv(time,status) ~ data2[,4],rho=1)
wilcoxon2
```

Και η p-value που λάβαμε από τον έλεγχο log-rank είναι ίση με p=0,9 οπότε δεχόμαστε τη μηδενική υπόθεση και άρα δεν υπάρχει διαφοροποίηση στους χρόνους επιβίωσης ανάμεσα στις δύο ηλικιακές ομάδες. Σε κάτι αντίστοιχο συμπέρασμα καταλήγουμε και με τον έλεγχο του Wilcoxon αφού λάβαμε p-value=1. Οπότε έχουμε αρκετές ενδείξεις για να θεωρήσουμε τη μεταβλητή age στατιστικά μη σημαντική.

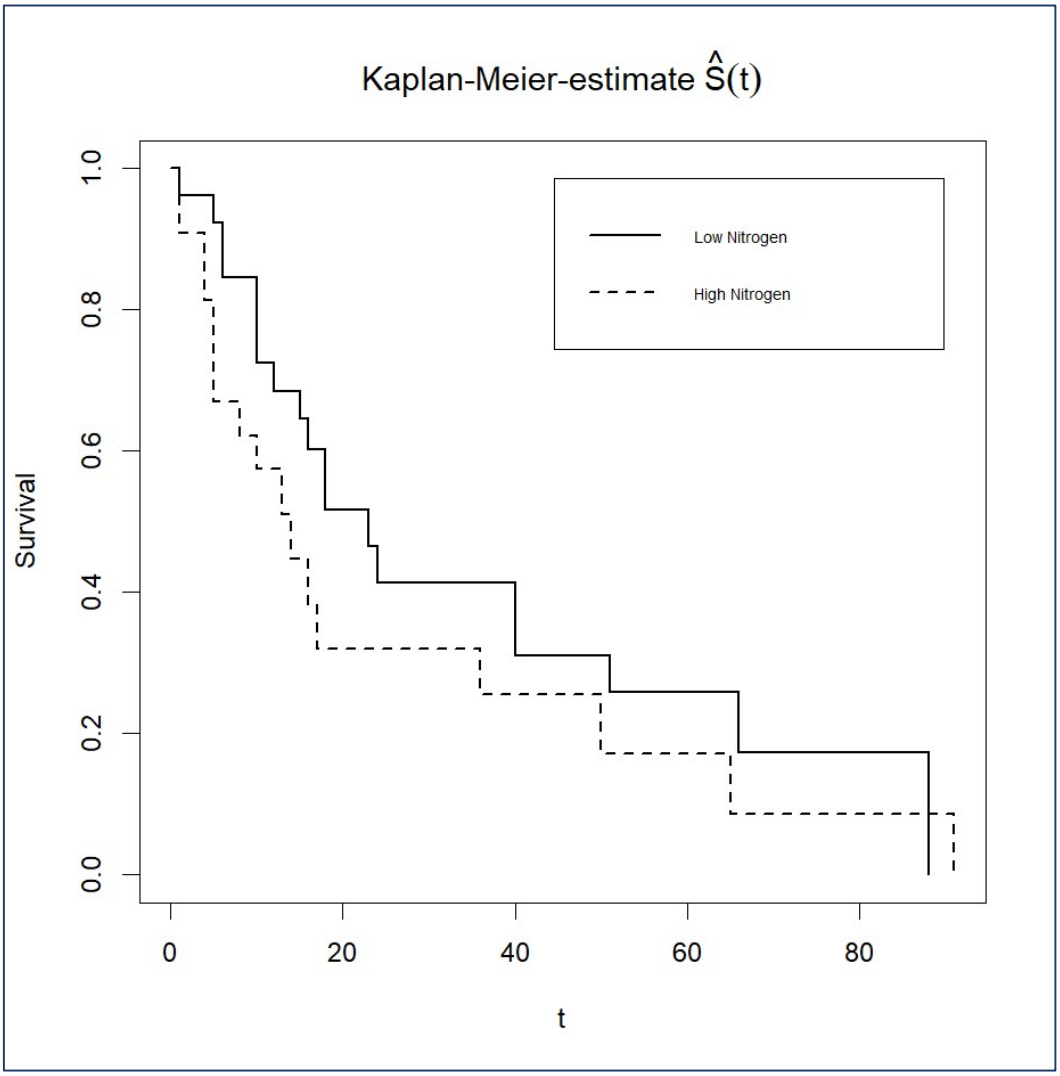
5.2.3 Σύγκριση των εκτιμήσεων Kaplan-Meier μεταξύ ατόμων που έχουν χαμηλή και υψηλή τιμή αζώτου ουρίας αίματος

Ελέγχουμε γραφικά αν η τιμή του αζώτου ουρίας αίματος επιδρά στη διάρκεια ζωής των ασθενών. Για να μπορέσουμε να κατασκευάσουμε τη γραφική αυτή παράσταση χωρίζουμε τους ασθενείς σε δύο κατηγορίες. Το μέτρο που χρησιμοποιήσαμε για το διαχωρισμό αυτό δεν ήταν η μέση τιμή διότι έχουμε ακραίες παρατηρήσεις. Για το λόγο αυτό θα χρησιμοποιήσουμε τη διάμεσο. Η διάμεσος των τιμών του αζώτου ουρίας των ασθενών βρέθηκε με τη βοήθεια στις εντολής median στην R ίσος με 21. Έτσι ομαδοποιήσαμε τους ασθενείς σε δύο κατηγορίες. Η πρώτη κατηγορία περιλάμβανε τα άτομα που είχαν τιμή αζώτου ουρίας μικρότερη ή ίση με διάμεσο, ενώ η δεύτερη ομάδα αποτελούταν από τους ασθενείς που είχαν τιμή αζώτου ουρίας μεγαλύτερη από τη διάμεσο. Στην πρώτη ομάδα δόθηκε η κωδικοποίηση 0 όσον αναφορά το άζωτο και στην δεύτερη η τιμή 1. Έτσι έχουμε:

$$bun = \begin{cases} 0, & \text{όταν } bun \leq 21 \\ 1, & \text{όταν } bun > 21 \end{cases}$$

Στη συνέχεια ελέγχουμε γραφικά αν η τιμή του αζώτου ουρίας αίματος επιδρά στη διάρκεια ζωής των ασθενών. Οι εντολές που χρησιμοποιήθηκαν ήταν οι ακόλουθες:

```
km_bun <- survfit(Surv(time, status)~data2[,6], type="kaplan-meier",data=data)

plot(km_bun, lty=1:2, main=expression(paste("Kaplan-Meier-estimate ", hat(S)(t))), xlab="t", ylab="Survival", lwd=1.5)

legend("topright", inset= 0.05, c("Low Nitrogen", "High Nitrogen"), lty = 1:2, lwd=1.5, cex=0.6)
```



Διάγραμμα 5.3: Σύγκριση των εκτιμήσεων Kaplan-Meier μεταξύ ατόμων που έχουν χαμηλή και υψηλή τιμή αζώτου ουρίας αίματος

Παρατηρούμε στο Διάγραμμα 5.3 ότι από τους πρώτους κιόλας μήνες οι Kaplan-Meier εκτιμήσεις στην ομάδα ασθενών που έχουν χαμηλές τιμές αζώτου ουρίας αίματος στις εξετάσεις τους (και μάλιστα μικρότερη από 21) είναι πιο πάνω από την αντίστοιχη στην ομάδα ασθενών που έχουν τιμή αζώτου μεγαλύτερη του 21. Για το λόγο αυτό και οι χρόνοι επιβίωσης στην πρώτη ομάδα είναι μεγαλύτεροι. Το συμπέρασμα αυτό ήταν λίγο πολύ αναμενόμενο καθώς από σχετική βιβλιογραφία ενημερωθήκαμε ότι οι υψηλές τιμές αζώου ουρίας αίματος επιδρούν αρνητικά στην κατάσταση στις υγείας του ασθενούς, επομένως μειώνουν και το χρόνο επιβίωσης.

Στη συνέχεια, υπολογίζουμε ξεχωριστά στις εκτιμήσεις Kaplan-Meier στις συνάρτησης επιβίωσης των ατόμων που έχουν χαμηλή και υψηλή τιμή αζώτου ουρίας αίματος, και ελέγχουμε αν υπάρχουν διαφορές μεταξύ αυτών των δύο αυτών ομάδων. Στα αποτελέσματα του Πίνακα 5.7 και 5.8 παρατηρούμε στις εκτιμήσεις Kaplan-Meier, τα τυπικά σφάλματα και τα διαστήματα εμπιστοσύνης των συναρτήσεων επιβίωσης για στις δύο κατηγορίες. Για να λάβουμε τα συμπεράσματα αυτά χρησιμοποιήσαμε την εντολή:

```
summary(km_bun)
```

```
               data2[, 6]=0
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     26       1    0.962  0.0377       0.8904        1.000
    5     25       1    0.923  0.0523       0.8261        1.000
    6     24       2    0.846  0.0708       0.7182        0.997
   10     21       3    0.725  0.0886       0.5708        0.922
   12     18       1    0.685  0.0924       0.5258        0.892
   15     17       1    0.645  0.0953       0.4825        0.861
   16     15       1    0.602  0.0982       0.4370        0.829
   18     14       2    0.516  0.1013       0.3510        0.758
   23     10       1    0.464  0.1034       0.2999        0.718
   24      9       1    0.413  0.1040       0.2517        0.676
   40      8       2    0.309  0.1004       0.1639        0.584
   51      6       1    0.258  0.0960       0.1243        0.535
   66      3       1    0.172  0.0950       0.0582        0.508
   88      1       1    0.000     NaN           NA           NA
```

Πίνακας 5.7: Εκτιμήσεις Kaplan-Meier της συνάρτησης επιβίωσης για τα άτομα που έχουν χαμηλή τιμή αζώτου ουρίας αίματος

data2[, 6]=1						
time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	22	2	0.9091	0.0613	0.7966	1.000
4	19	2	0.8134	0.0843	0.6639	0.997
5	17	3	0.6699	0.1023	0.4965	0.904
8	14	1	0.6220	0.1056	0.4459	0.868
10	13	1	0.5742	0.1078	0.3974	0.830
13	9	1	0.5104	0.1131	0.3305	0.788
14	8	1	0.4466	0.1156	0.2689	0.742
16	7	1	0.3828	0.1153	0.2121	0.691
17	6	1	0.3190	0.1124	0.1599	0.636
36	5	1	0.2552	0.1065	0.1126	0.578
50	3	1	0.1701	0.0993	0.0542	0.534
65	2	1	0.0851	0.0780	0.0141	0.513
91	1	1	0.0000	NaN	NA	NA

Πίνακας 5.8: Εκτιμήσεις Kaplan-Meier της συνάρτησης επιβίωσης για τα άτομα που έχουν υψηλή τιμή αζώτου ουρίας αίματος

Το συμπέρασμα που βγάλαμε νωρίτερα επιβεβαιώνεται και από στις Πίνακες 5.7 και 5.8 οπότε η τιμή του αζώτου ουρίας φαίνεται να επηρεάζει το χρόνο ζωής των ασθενών.

Ένας τελευταίος τρόπος για να στηρίξουμε το συμπέρασμά μας είναι ο έλεγχος log-rank και ο έλεγχος του Wilcoxon. Οι εντολές που χρησιμοποιήσαμε είναι οι ακόλουθες.

```
Log_rank3<-survdiff(Surv(time,status)~data2[,6])
log_rank3

wilcoxon3<-survdiff(Surv(time,status)~data2[,6],rho=1)
wilcoxon3
```

Η p-value που λάβαμε από τον έλεγχο log-rank είναι ίση με p=0,02 οπότε απορρίπτουμε τη μηδενική υπόθεση και άρα υπάρχει διαφοροποίηση ανάμεσα στις δύο ομάδες ασθενών. Σε κάτι αντίστοιχο καταλήγουμε και με τον έλεγχο του Wilcoxon αφού λάβαμε p-value=0,01. Οπότε έχουμε αρκετές ενδείξεις για να κρίνουμε τη μεταβλητή bun στατιστικά σημαντική. Επομένως, η μεταβλητή αυτή επηρεάζει τη διάρκεια επιβίωσης των ασθενών που πάσχουν από πολλαπλό μυέλωμα.

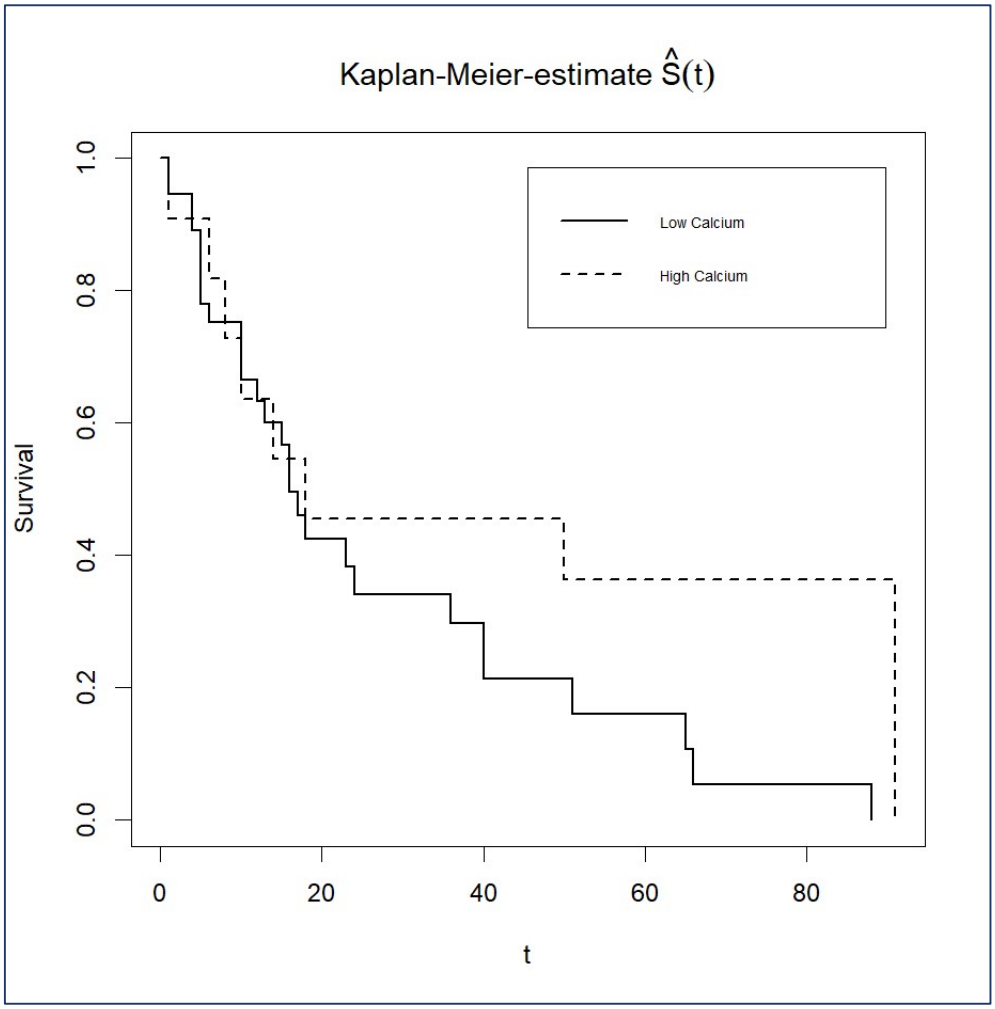
5.2.4 Σύγκριση των εκτιμήσεων Kaplan-Meier μεταξύ ατόμων με χαμηλή και υψηλή τιμή ασβεστίου

Ελέγχουμε γραφικά αν η τιμή του ασβεστίου επιδρά στη διάρκεια ζωής των ασθενών. Για να μπορέσουμε να κατασκευάσουμε τη γραφική αυτή παράσταση χωρίζουμε τους ασθενείς σε δύο κατηγορίες. Το μέτρο που χρησιμοποιήσαμε για το διαχωρισμό αυτό δεν ήταν η μέση τιμή διότι έχουμε ακραίες παρατηρήσεις. Για το λόγο αυτό θα χρησιμοποιήσουμε τη διάμεσο. Η διάμεσος των τιμών του ασβεστίου των ασθενών βρέθηκε με τη βοήθεια της εντολής median στην R ίση με 10. Έτσι ομαδοποιήσαμε τους ασθενείς σε δύο κατηγορίες. Η πρώτη κατηγορία περιλάμβανε τα άτομα που είχαν τιμή ασβεστίου μικρότερη από τη διάμεσο ενώ η δεύτερη ομάδα αποτελούταν από τους ασθενείς που είχαν τιμή ασβεστίου μεγαλύτερη από τη διάμεσο. Στην πρώτη ομάδα δόθηκε η κωδικοποίηση 0 όσον αναφορά την ηλικία και στην δεύτερη η τιμή 1. Έτσι έχουμε:

$$ca = \begin{cases} 0, & \text{όταν } ca \leq 10 \\ 1, & \text{όταν } ca > 10 \end{cases}$$

Πλέον είμαστε έτοιμοι να λάβουμε τη γραφική παράσταση από την οποία θα εξάγουμε σημαντικά για τη μελέτη μας αποτελέσματα. Οι εντολές που χρησιμοποιήθηκαν ήταν οι ακόλουθες:

```
km_ca <- survfit(Surv(time, status)~data2[,7], type="kaplan-meier",data=data)

plot(km_ca, lty = 1:2, main=expression(paste("Kaplan-Meier-estimate ", hat(S)(t))), xlab="t", ylab="Survival", lwd=1.5)

legend("topright", inset= 0.05, c("Low Calcium", "High Calcium"), lty = 1:2, lwd=1.5, cex=0.6)
```



Διάγραμμα 5.4: Σύγκριση των εκτιμήσεων Kaplan-Meier μεταξύ ατόμων με χαμηλή και υψηλή τιμή ασβεστίου

Στο Διάγραμμα 5.4 συγκρίνουμε γραφικά τις εκτιμήσεις Kaplan-Meier μεταξύ των δύο ομάδων και παρατηρούμε ότι δεν υπάρχει μεγάλη διαφοροποίηση ανάμεσα στις δύο αυτές κατηγορίες μέχρι και λίγο πριν τον 20ο μήνα. Αυτό γίνεται φανερό από το γεγονός ότι σε ορισμένα διαστήματα του διαγράμματος η γραφική της πρώτης ομάδας βρίσκεται πάνω από τη γραφική της δεύτερης ομάδας την ίδια στιγμή που σε άλλα διαστήματα του Διαγράμματος συμβαίνει το αντίθετο. Αντιθέτως, από τον 20ο μήνα και έπειτα είναι φανερό ότι τα άτομα στα οποία η τιμή του ασβεστίου είναι υψηλή, και συγκεκριμένα μεγαλύτερη από 10, έχουν πολύ καλύτερη πορεία στην υγεία τους σε σχέση με τα άτομα των οποίων το ασβέστιο είναι χαμηλό. Το αποτέλεσμα αυτό ήταν αναμενόμενο, καθώς η ύπαρξη ασβεστίου σε έναν οργανισμό βελτιώνει τη συνολική υγεία του ατόμου.

Στη συνέχεια, υπολογίζουμε ξεχωριστά τις εκτιμήτριες Kaplan-Meier της συνάρτησης επιβίωσης για τις δύο ομάδες και ελέγχουμε αν υπάρχουν διαφορές μεταξύ αυτών των δύο κατηγοριών. Για να λάβουμε τα συμπεράσματα αυτά χρησιμοποιήσαμε την εντολή:

```
summary(km_ca)
```

Στα αποτελέσματα στους Πίνακες 5.9 και 5.10 παρατηρούμε τις εκτιμήσεις Kaplan-Meier, τα τυπικά σφάλματα και τα διαστήματα εμπιστοσύνης των συναρτήσεων επιβίωσης για τις δύο ομάδες. Συγκεκριμένα στον Πίνακα 5.9 έχουμε λάβει αποτελέσματα σχετικά με τα άτομα που έχουν χαμηλή τιμή ασβεστίου στο αίμα τους, ενώ στον Πίνακα 5.10 παρουσιάζονται τα αποτελέσματα που αφορούν τους ασθενείς με υψηλές τιμές ασβεστίου.

data2[, 7]=0

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	37	2	0.9459	0.0372	0.87582	1.000
4	34	2	0.8903	0.0518	0.79439	0.998
5	32	4	0.7790	0.0690	0.65485	0.927
6	28	1	0.7512	0.0719	0.62265	0.906
10	26	3	0.6645	0.0791	0.52616	0.839
12	21	1	0.6329	0.0815	0.49176	0.814
13	19	1	0.5996	0.0837	0.45603	0.788
15	18	1	0.5663	0.0854	0.42131	0.761
16	16	2	0.4955	0.0882	0.34954	0.702
17	14	1	0.4601	0.0887	0.31528	0.671
18	13	1	0.4247	0.0887	0.28207	0.639
23	10	1	0.3822	0.0894	0.24167	0.605
24	9	1	0.3398	0.0890	0.20334	0.568
36	8	1	0.2973	0.0874	0.16707	0.529
40	7	2	0.2123	0.0805	0.10104	0.446
51	4	1	0.1593	0.0759	0.06261	0.405
65	3	1	0.1062	0.0666	0.03105	0.363
66	2	1	0.0531	0.0502	0.00832	0.339
88	1	1	0.0000	NaN	NA	NA

Πίνακας 5.9: Εκτιμήσεις Kaplan-Meier της συνάρτησης επιβίωσης για ασθενείς με χαμηλή τιμή ασβεστίου

data2[, 7]=1

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	11	1	0.909	0.0867	0.754	1.000
6	10	1	0.818	0.1163	0.619	1.000
8	9	1	0.727	0.1343	0.506	1.000
10	8	1	0.636	0.1450	0.407	0.995
14	7	1	0.545	0.1501	0.318	0.936
18	6	1	0.455	0.1501	0.238	0.868
50	5	1	0.364	0.1450	0.166	0.795
91	1	1	0.000	NaN	NA	NA

Πίνακας 5.10: Εκτιμήσεις Kaplan-Meier της συνάρτησης επιβίωσης για ασθενείς με υψηλή τιμή ασβεστίου

Το συμπέρασμα που βγάλαμε νωρίτερα επιβεβαιώνεται από τους Πίνακες 5.9 και 5.10 καθώς παρατηρούμε ότι από τον 20ο μήνα και μετά οι τιμές τις συνάρτησης επιβίωσης είναι μεγαλύτερες για τα άτομα που έχουν υψηλές τιμές ασβεστίου σε σχέση με εκείνα που δεν έχουν. Κάτι τέτοιο βέβαια δεν συμβαίνει και πριν τον 20ο μήνα οπότε ακόμα δεν μπορούμε να είμαστε βέβαια για το αν η μεταβλητή ca είναι στατιστικά σημαντική, δηλαδή δεν είμαστε σε θέση να κρίνουμε αν οι τιμές του ασβεστίου στο αίμα του ασθενούς επηρεάζουν το χρονικό διάστημα επιβίωσής του.

Μία λύση για να αποφανθούμε για το αν η τιμή του ασβεστίου επηρεάζει την έκβαση της υγείας του ασθενούς είναι ο έλεγχος log-rank και ο έλεγχος του Wilcoxon. Οι εντολές που χρησιμοποιήσαμε είναι οι ακόλουθες.

```
log_rank4 <- survdiff(Surv(time,status) ~ data2[,7])
log_rank4

wilcoxon4 <- survdiff(Surv(time,status) ~ data2[,7],rho=1)
wilcoxon4
```

Και η p-value που λάβαμε από τον έλεγχο log-rank είναι ίση με p=0,2 οπότε δεχόμαστε τη μηδενική υπόθεση και άρα δεν υπάρχει διαφοροποίηση στους χρόνους επιβίωσης ανάμεσα στις δύο ομάδες. Ακόμα με τον έλεγχο του Wilcoxon λαμβάνουμε p-value=0,5 τιμή η οποία μπορεί να υποδείξει αποδοχή της μηδενικής υπόθεσης. Τελικά μπορούμε να πούμε ότι η επεξηγηματική μεταβλητή ca δεν είναι στατιστικά σημαντική, δηλαδή η τιμή του ασβεστίου δεν επηρεάζει την έκβαση της υγείας των ασθενών.

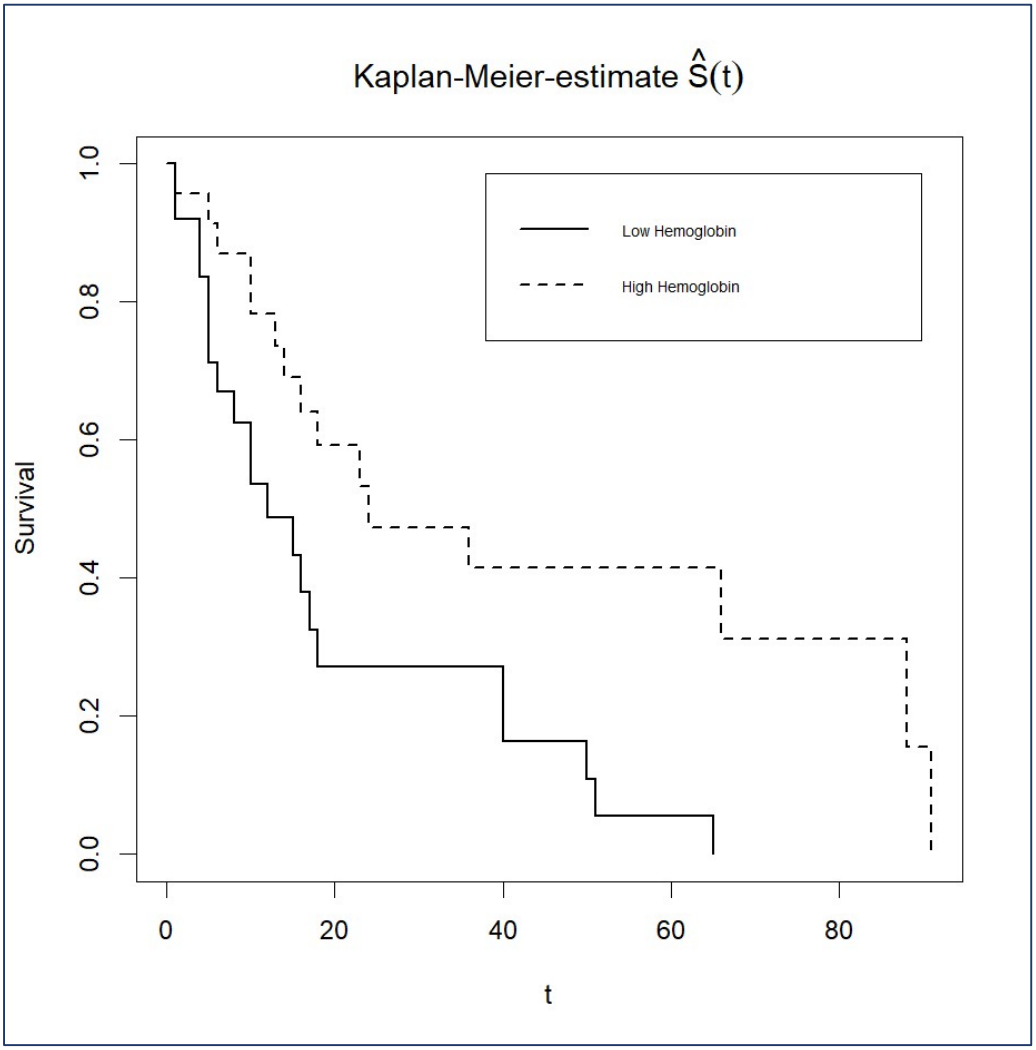
5.2.5 Σύγκριση των εκτιμήσεων Kaplan-Meier μεταξύ ασθενών με χαμηλή και υψηλή αιμοσφαιρίνη

Ελέγχουμε γραφικά αν η τιμή της αιμοσφαιρίνης επιδρά στη διάρκεια ζωής των ασθενών. Για να μπορέσουμε να κατασκευάσουμε τη γραφική αυτή παράσταση χωρίζουμε τους ασθενείς σε δύο κατηγορίες. Το μέτρο που χρησιμοποιήσαμε για το διαχωρισμό αυτό ήταν η μέση τιμή της αιμοσφαιρίνης των ασθενών καθώς δεν έχουμε ακραίες παρατηρήσεις για να χρειαστεί να χρησιμοποιήσουμε τη διάμεσο. Ο δειγματικός μέσος της αιμοσφαιρίνης των ασθενών βρέθηκε με τη βοήθεια της εντολής sum στην R ίσος με 10,2. Έτσι ομαδοποιήσαμε τους ασθενείς σε δύο κατηγορίες. Η πρώτη κατηγορία περιλάμβανε τα άτομα που είχαν τιμή αιμοσφαιρίνης μικρότερη από το δειγματικό μέσο ενώ η δεύτερη ομάδα αποτελούταν από τους ασθενείς που είχαν αιμοσφαιρίνη μεγαλύτερη από τη μέση τιμή. Στην πρώτη ομάδα δόθηκε η κωδικοποίηση 0 όσον αναφορά την ηλικία και στην δεύτερη η τιμή 1. Έτσι έχουμε:

$$hb = \begin{cases} 0, & \text{όταν } hb \leq 10{,}2 \\ 1, & \text{όταν } hb > 10{,}2 \end{cases}$$

Στη συνέχεια ελέγχουμε γραφικά αν η τιμή της αιμοσφαιρίνης επιδρά στη διάρκεια ζωής των ασθενών. Οι εντολές που χρησιμοποιήθηκαν ήταν οι ακόλουθες:

```
km_hb  <- survfit(Surv(time, status)~data2[,8], type="kaplan-meier",data=data)

plot(km_hb, lty = 1:2, main=expression(paste("Kaplan-Meier-estimate ", hat(S)(t))), xlab="t", ylab="Survival", lwd=1.5)

legend("topright", inset= 0.05, c("Low Hemoglobin", "High Hemoglobin"), lty = 1:2, lwd=1.5, cex=0.6)
```



Διάγραμμα 5.5: Σύγκριση των εκτιμήσεων Kaplan-Meier μεταξύ ατόμων που έχουν χαμηλή και υψηλή τιμή αιμοσφαιρίνης

Παρατηρούμε στο Διάγραμμα 5.5 ότι από τους πρώτους κιόλας μήνες η Kaplan-Meier εκτιμήτρια της ομάδας ασθενών που έχουν χαμηλές τιμές αιμοσφαιρίνης στις εξετάσεις τους (και μάλιστα μικρότερη από 10,2) είναι πιο κάτω από την αντίστοιχη στην ομάδας ασθενών που έχουν τιμή αιμοσφαιρίνης μεγαλύτερη του 10,2. Για το λόγο αυτό και οι χρόνοι επιβίωσής τους είναι μικρότεροι. Το συμπέρασμα αυτό ήταν λίγο πολύ αναμενόμενο καθώς γνωρίζουμε ότι οι υψηλές τιμές της αιμοσφαιρίνης επιδρούν θετικά στην κατάσταση της υγείας του ασθενούς, επομένως αυξάνουν και το χρόνο επιβίωσης. Αντιθέτως, χαμηλές τιμές αιμοσφαιρίνης επιφέρουν γρηγορότερα το θάνατο καθώς επιδρούν αρνητικά στην υγεία των ασθενών.

Στη συνέχεια, υπολογίζουμε ξεχωριστά τις εκτιμήσεις Kaplan-Meier της συνάρτησης επιβίωσης των ατόμων που έχουν χαμηλή και υψηλή τιμή αιμοσφαιρίνης, και ελέγχουμε αν υπάρχουν διαφορές μεταξύ αυτών των δύο ομάδων. Στα αποτελέσματα του Πίνακα 5.11 και 5.12 παρατηρούμε τις εκτιμήσεις Kaplan-Meier, τα τυπικά σφάλματα και τα διαστήματα εμπιστοσύνης των συναρτήσεων επιβίωσης για στις δύο κατηγορίες. Για να λάβουμε τα συμπεράσματα αυτά χρησιμοποιήσαμε την εντολή:

```
summary(km_hb)
```

data2[, 8]=0

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	25	2	0.9200	0.0543	0.81957	1.000
4	22	2	0.8364	0.0749	0.70170	0.997
5	20	3	0.7109	0.0923	0.55123	0.917
6	17	1	0.6691	0.0959	0.50529	0.886
8	15	1	0.6245	0.0993	0.45727	0.853
10	14	2	0.5353	0.1032	0.36679	0.781
12	11	1	0.4866	0.1047	0.31920	0.742
15	9	1	0.4325	0.1061	0.26744	0.700
16	8	1	0.3785	0.1057	0.21891	0.654
17	7	1	0.3244	0.1035	0.17356	0.606
18	6	1	0.2703	0.0994	0.13151	0.556
40	5	2	0.1622	0.0840	0.05875	0.448
50	3	1	0.1081	0.0713	0.02968	0.394
51	2	1	0.0541	0.0523	0.00812	0.360
65	1	1	0.0000	NaN	NA	NA

Πίνακας 5.11: Εκτιμήσεις Kaplan-Meier της συνάρτησης επιβίωσης για τα άτομα που έχουν χαμηλή τιμή αιμοσφαιρίνης

data2[, 8]=1						
time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	23	1	0.957	0.0425	0.8767	1.000
5	22	1	0.913	0.0588	0.8049	1.000
6	21	1	0.870	0.0702	0.7423	1.000
10	20	2	0.783	0.0860	0.6310	0.971
13	17	1	0.737	0.0925	0.5759	0.942
14	16	1	0.691	0.0975	0.5237	0.911
16	14	1	0.641	0.1022	0.4691	0.876
18	13	1	0.592	0.1056	0.4172	0.840
23	10	1	0.533	0.1104	0.3549	0.800
24	9	1	0.474	0.1129	0.2968	0.756
36	8	1	0.414	0.1132	0.2425	0.708
66	4	1	0.311	0.1235	0.1426	0.677
88	2	1	0.155	0.1260	0.0317	0.762
91	1	1	0.000	NaN	NA	NA

Πίνακας 5.12: Εκτιμήσεις Kaplan-Meier της συνάρτησης επιβίωσης για τα άτομα που έχουν υψηλή τιμή αιμοσφαιρίνης

Το συμπέρασμα που βγάλαμε νωρίτερα επιβεβαιώνεται και από τους Πίνακες 5.11 και 5.12 οπότε η τιμή της αιμοσφαιρίνης φαίνεται να επηρεάζει το χρόνο ζωής των ασθενών και μάλιστα θετικά.

Ένας τελευταίος τρόπος για να στηρίξουμε το συμπέρασμά στις είναι ο έλεγχος log-rank και ο έλεγχος του Wilcoxon. Οι εντολές που χρησιμοποιήσαμε είναι οι ακόλουθες.

```
Log_rank5 <- survdiff(Surv(time,status) ~ data2[,8])
log_rank5

wilcoxon5 <- survdiff(Surv(time,status) ~ data2[,8],rho=1)
wilcoxon5
```

Και η p-value που λάβαμε από τον έλεγχο log-rank είναι ίση με p=0,005 οπότε απορρίπτουμε τη μηδενική υπόθεση και άρα υπάρχει διαφοροποίηση ανάμεσα στις δύο ομάδες ασθενών. Σε κάτι αντίστοιχο καταλήγουμε και με τον έλεγχο του Wilcoxon αφού λάβαμε p-value=0,02. Οπότε έχουμε αρκετές ενδείξεις για να θεωρήσουμε τη μεταβλητή hb στατιστικά σημαντική. Επομένως, η μεταβλητή αυτή επηρεάζει τη διάρκεια επιβίωσης των ασθενών που πάσχουν από πολλαπλό μυέλωμα.

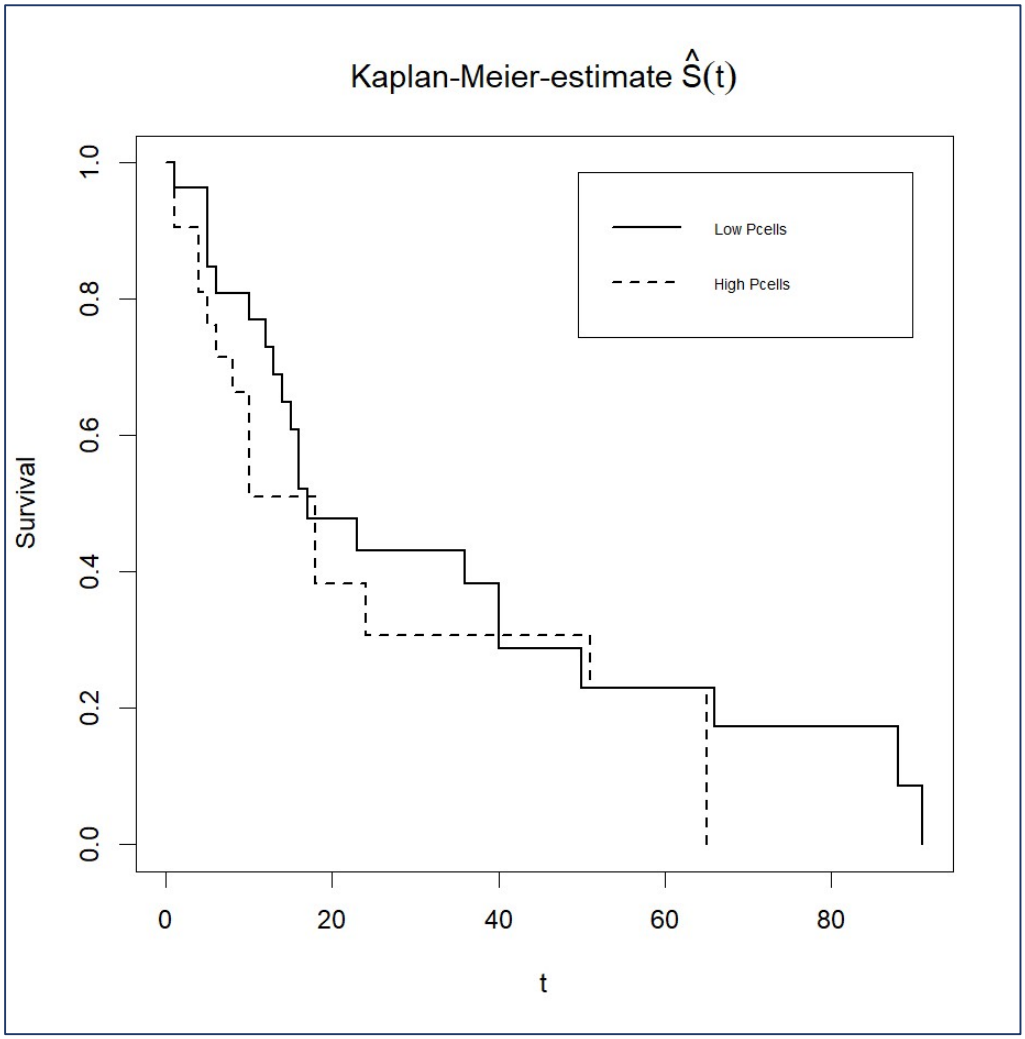
5.2.6 Σύγκριση των εκτιμήσεων Kaplan-Meier μεταξύ ασθενών με χαμηλό και υψηλό ποσοστό καρκινικών κυττάρων

Ελέγχουμε γραφικά αν το ποσοστό καρκινικών κυττάρων επιδρά στη διάρκεια ζωής των ασθενών. Για να μπορέσουμε να κατασκευάσουμε τη γραφική αυτή παράσταση χωρίζουμε τους ασθενείς σε δύο κατηγορίες. Το μέτρο που χρησιμοποιήσαμε για το διαχωρισμό αυτό ήταν η μέση τιμή της αιμοσφαιρίνης των ασθενών. Ο δειγματικός μέσος του ποσοστού καρκινικών κυττάρων των ασθενών βρέθηκε με τη βοήθεια της εντολής sum στην R ίσος με 43%. Έτσι ομαδοποιήσαμε τους ασθενείς σε δύο κατηγορίες. Η πρώτη κατηγορία περιλάμβανε τα άτομα που είχαν τιμή ποσοστού καρκινικών κυττάρων μικρότερη από το δειγματικό μέσο ενώ η δεύτερη ομάδα αποτελούταν από τους ασθενείς που είχαν ποσοστό καρκινικών κυττάρων μεγαλύτερο από τη μέση τιμή. Στην πρώτη ομάδα δόθηκε η κωδικοποίηση 0 όσον αναφορά το ποσοστό των καρκινικών κυττάρων και στην δεύτερη η τιμή 1. Έτσι έχουμε:

$$pcells = \begin{cases} 0, & \text{όταν } pcells \leq 43\% \\ 1, & \text{όταν } hb > 43\% \end{cases}$$

Στη συνέχεια ελέγχουμε γραφικά αν η τιμή της αιμοσφαιρίνης επιδρά στη διάρκεια ζωής των ασθενών. Οι εντολές που χρησιμοποιήθηκαν ήταν οι ακόλουθες:

```
km_pcells <- survfit(Surv(time, status)~data2[,9], type="kaplan-meier",data=data)

plot(km_pcells, lty = 1:2, main=expression(paste("Kaplan-Meier-estimate ", hat(S)(t))), xlab="t", ylab="Survival", lwd=1.5)

legend("topright", inset= 0.05, c("Low Pcells", "High Pcells"), lty = 1:2, lwd=1.5, cex=0.6)
```



Διάγραμμα 5.6: Σύγκριση των εκτιμήσεων Kaplan-Meier μεταξύ ατόμων με χαμηλό και υψηλό ποσοστό καρκινικών κυττάρων

Παρατηρούμε στο Διάγραμμα 5.6 ότι από τους πρώτους κιόλας μήνες η Kaplan-Meier εκτίμηση της ομάδας ασθενών που έχουν χαμηλό ποσοστό καρκινικών κυττάρων στις εξετάσεις τους είναι πιο ψηλά από την αντίστοιχη της ομάδας ασθενών που έχουν ποσοστό καρκινικών κυττάρων μεγαλύτερο του 43%. Για το λόγο αυτό και οι χρόνοι επιβίωσης τους είναι μεγαλύτεροι. Το συμπέρασμα αυτό ήταν λίγο πολύ αναμενόμενο καθώς γνωρίζουμε ότι τα υψηλά ποσοστά καρκινικών κυττάρων επιδρούν αρνητικά στην κατάσταση της υγείας του ασθενούς, επομένως ελαττώνουν και το χρόνο επιβίωσης. Αντιθέτως, χαμηλά ποσοστά καρκινικών κυττάρων αυξάνουν το χρόνο επιβίωσης των ασθενών. Βέβαια σε ορισμένα σημεία η γραφική παράσταση των ατόμων με μεγάλο ποσοστό καρκινικών κυττάρων βρίσκεται ψηλότερα από την γραφική των ατόμων με χαμηλό ποσοστό καρκινικών κυττάρων. Για το λόγο αυτό δεν είμαστε ακόμα βέβαιοι για το αν η μεταβλητή «pcells» είναι στατιστικά σημαντική.

Στη συνέχεια, υπολογίζουμε ξεχωριστά τις εκτιμήσεις Kaplan-Meier της συνάρτησης επιβίωσης των ατόμων που έχουν χαμηλό και υψηλό ποσοστό καρκινικών κυττάρων, και ελέγχουμε αν υπάρχουν διαφορές μεταξύ αυτών των δύο αυτών ομάδων. Στα αποτελέσματα του Πίνακα 5.13 και 5.14 παρατηρούμε τις εκτιμήτριες Kaplan-Meier, τα τυπικά σφάλματα και τα διαστήματα εμπιστοσύνης των συναρτήσεων επιβίωσης για τις δύο κατηγορίες. Για να λάβουμε τα συμπεράσματα αυτά χρησιμοποιήσαμε την εντολή:

```
summary(km_pcells)
```

data2[, 9]=0

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	27	1	0.963	0.0363	0.8943	1.000
5	25	3	0.847	0.0703	0.7203	0.997
6	22	1	0.809	0.0769	0.6713	0.975
10	21	1	0.770	0.0823	0.6248	0.950
12	19	1	0.730	0.0874	0.5771	0.923
13	18	1	0.689	0.0915	0.5314	0.894
14	17	1	0.649	0.0947	0.4874	0.864
15	16	1	0.608	0.0970	0.4449	0.831
16	14	2	0.521	0.1008	0.3569	0.761
17	12	1	0.478	0.1013	0.3154	0.724
23	10	1	0.430	0.1018	0.2704	0.684
36	9	1	0.382	0.1011	0.2277	0.642
40	8	2	0.287	0.0958	0.1490	0.552
50	5	1	0.229	0.0922	0.1043	0.504
66	4	1	0.172	0.0851	0.0652	0.454
88	2	1	0.086	0.0742	0.0158	0.467
91	1	1	0.000	NaN	NA	NA

Πίνακας 5.13: Εκτιμήσεις Kaplan-Meier της συνάρτησης επιβίωσης για τα άτομα που έχουν χαμηλό ποσοστό καρκινικών κυττάρων

data2[, 9]=1

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	21	2	0.905	0.0641	0.7875	1.000
4	19	2	0.810	0.0857	0.6579	0.996
5	17	1	0.762	0.0929	0.5999	0.968
6	16	1	0.714	0.0986	0.5450	0.936
8	14	1	0.663	0.1039	0.4879	0.902
10	13	3	0.510	0.1113	0.3327	0.783
18	8	2	0.383	0.1143	0.2130	0.687
24	5	1	0.306	0.1142	0.1473	0.636
51	4	1	0.230	0.1083	0.0911	0.579
65	1	1	0.000	NaN	NA	NA

Πίνακας 5.14: Εκτιμήσεις Kaplan-Meier της συνάρτησης επιβίωσης για τα άτομα που έχουν υψηλό ποσοστό καρκινικών κυττάρων

Το συμπέρασμα που βγάλαμε νωρίτερα επιβεβαιώνεται και από τους Πίνακες 5.13 και 5.14, οπότε το ποσοστό των καρκινικών κυττάρων φαίνεται να επηρεάζει το χρόνο ζωής των ασθενών και μάλιστα αρνητικά.

Ένας τελευταίος τρόπος για να στηρίξουμε το συμπέρασμά μας είναι ο έλεγχος log-rank και ο έλεγχος του Wilcoxon. Οι εντολές που χρησιμοποιήσαμε είναι οι ακόλουθες.

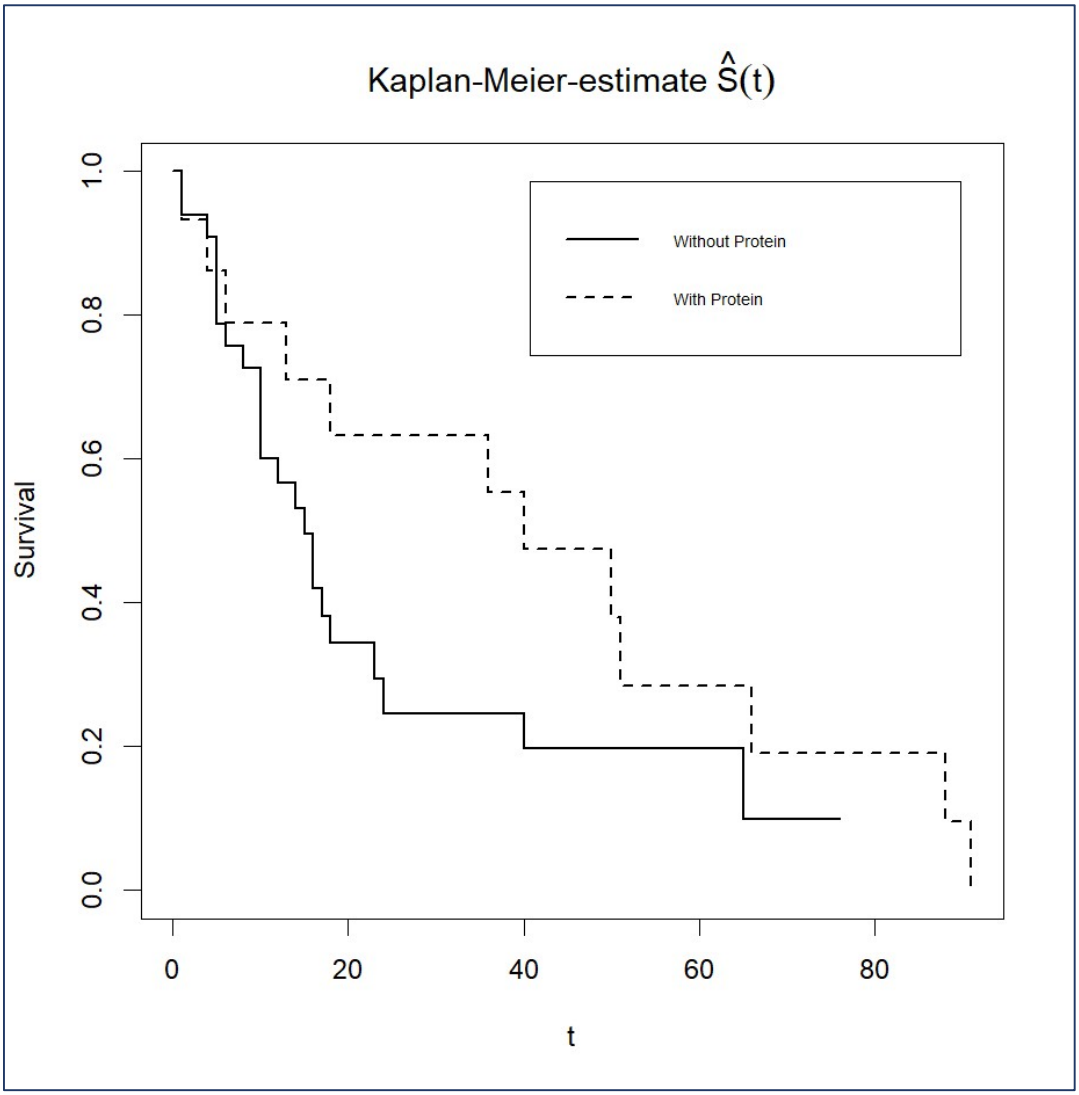
```
log_rank6 <- survdiff(Surv(time,status) ~ data2[,9])
log_rank6

wilcoxon6 <- survdiff(Surv(time,status) ~ data2[,9],rho=1)
wilcoxon6
```

Και η p-value που λάβαμε από τον έλεγχο log-rank είναι ίση με p=0,04 οπότε απορρίπτουμε τη μηδενική υπόθεση και άρα υπάρχει διαφοροποίηση ανάμεσα στις δύο ομάδες ασθενών. Σε κάτι αντίστοιχο καταλήγουμε και με τον έλεγχο του Wilcoxon αφού λάβαμε p-value=0,03. Οπότε έχουμε αρκετές ενδείξεις για να θεωρήσουμε τη μεταβλητή pcells στατιστικά σημαντική. Επομένως, η μεταβλητή αυτή επηρεάζει τη διάρκεια επιβίωσης των ασθενών που πάσχουν από πολλαπλό μυέλωμα.

5.2.7 Σύγκριση των εκτιμήσεων Kaplan-Meier μεταξύ ατόμων που έχουν την μονοκλωνική πρωτεΐνη και αυτών που δεν την έχουν

Αρχικά ελέγχουμε γραφικά αν η ύπαρξη της μονοκλωνικής πρωτεΐνης επιδρά στη διάρκεια ζωής των ασθενών. Οι εντολές που χρησιμοποιήθηκαν ήταν οι ακόλουθες:

```
km_protein <- survfit(Surv(time, status)~protein, type="kaplan-meier",data=data)

plot(km_protein, lty = 1:2, main=expression(paste("Kaplan-Meier-estimate", hat(S)(t))), xlab="t", ylab="Survival", lwd=1.5)

legend("topright", inset=c(0.05), c("Without Protein", "With Protein"), lty = 1:2, lwd=1.5, cex=0.6)
```



Διάγραμμα 5.7: Σύγκριση των εκτιμήσεων Kaplan-Meier μεταξύ ασθενών που έχουν τη μονοκλωνική πρωτεΐνη και εκείνων που δεν την έχουν

Παρατηρούμε στο Διάγραμμα 5.7 ότι τους πρώτους μήνες δεν παίζει ιδιαίτερο ρόλο η ύπαρξη της μονοκλωνικής πρωτεΐνης. Αντιθέτως μετά τον 10^{ο} μήνα η Kaplan-Meier εκτιμήτρια της ομάδας ασθενών που δεν έχουν τη μονοκλωνική πρωτεΐνη είναι πιο πάνω από την αντίστοιχη της ομάδας ασθενών που έχουν την πρωτεΐνη αυτή. Για το λόγο αυτό και οι χρόνοι επιβίωσής τους είναι μεγαλύτεροι. Το συμπέρασμα αυτό ήταν λίγο πολύ αναμενόμενο καθώς από σχετική βιβλιογραφία ενημερωθήκαμε ότι η ύπαρξη της πρωτεΐνης αυτής επιδρά αρνητικά στην κατάσταση της υγείας του ασθενούς, επομένως μειώνει και το χρόνο επιβίωσης.

Στη συνέχεια, υπολογίζουμε ξεχωριστά τις εκτιμήσεις Kaplan-Meier της συνάρτησης επιβίωσης για τα άτομα που δεν έχουν στον ορό του αίματός τους τη μονοκλωνική πρωτεΐνη και τα άτομα που την έχουν, και ελέγχουμε αν υπάρχουν διαφορές μεταξύ αυτών των δύο αυτών ομάδων. Στα αποτελέσματα του Πίνακα 5.15 και 5.16 παρατηρούμε τις εκτιμήτριες Kaplan-Meier, τα τυπικά σφάλματα και τα διαστήματα εμπιστοσύνης των συναρτήσεων επιβίωσης για τις δύο κατηγορίες. Για να λάβουμε τα συμπεράσματα αυτά χρησιμοποιήσαμε την εντολή:

```
summary(km_protein)
```

protein=0

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	33	2	0.939	0.0415	0.8614	1.000
4	31	1	0.909	0.0500	0.8161	1.000
5	30	4	0.788	0.0712	0.6600	0.940
6	26	1	0.758	0.0746	0.6246	0.919
8	24	1	0.726	0.0779	0.5883	0.896
10	23	4	0.600	0.0862	0.4525	0.795
12	18	1	0.566	0.0876	0.4183	0.767
14	16	1	0.531	0.0890	0.3823	0.738
15	15	1	0.496	0.0898	0.3474	0.707
16	13	2	0.419	0.0908	0.2744	0.641
17	11	1	0.381	0.0902	0.2398	0.606
18	10	1	0.343	0.0888	0.2066	0.570
23	7	1	0.294	0.0887	0.1629	0.531
24	6	1	0.245	0.0864	0.1228	0.489
40	5	1	0.196	0.0818	0.0865	0.444
65	2	1	0.098	0.0805	0.0196	0.490

Πίνακας 5.15: Εκτιμήσεις Kaplan-Meier της συνάρτησης επιβίωσης για τα άτομα που δεν έχουν την μονοκλωνική πρωτεΐνη στο αίμα στις

protein=1						
time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	15	1	0.9333	0.0644	0.8153	1.000
4	13	1	0.8615	0.0911	0.7003	1.000
6	12	1	0.7897	0.1081	0.6039	1.000
13	10	1	0.7108	0.1228	0.5066	0.997
18	9	1	0.6318	0.1321	0.4193	0.952
36	8	1	0.5528	0.1372	0.3399	0.899
40	7	1	0.4738	0.1385	0.2672	0.840
50	5	1	0.3791	0.1395	0.1843	0.780
51	4	1	0.2843	0.1330	0.1137	0.711
66	3	1	0.1895	0.1177	0.0561	0.640
88	2	1	0.0948	0.0892	0.0150	0.599
91	1	1	0.0000	NaN	NA	NA

Πίνακας 5.16: Εκτιμήσεις Kaplan-Meier της συνάρτησης επιβίωσης για τα άτομα που έχουν την μονοκλωνική πρωτεΐνη στο αίμα στις

Το συμπέρασμα που βγάλαμε νωρίτερα επιβεβαιώνεται και από στις Πίνακες 5.15 και 5.16 οπότε η ύπαρξη της μονοκλωνικής πρωτεΐνης φαίνεται να επηρεάζει το χρόνο ζωής των ασθενών.

Ένας τελευταίος τρόπος για να στηρίξουμε το συμπέρασμά μας είναι ο έλεγχος log-rank και ο έλεγχος του Wilcoxon. Οι εντολές που χρησιμοποιήσαμε είναι οι ακόλουθες.

```
Log_rank9 <- survdiff(Surv(time,status) ~ protein
Log_rank9

wilcoxon9 <- survdiff(Surv(time,status) ~ protein,rho=1)
wilcoxon9
```

Και η p-value που λάβαμε από τον έλεγχο log-rank είναι ίση με p=0,02 οπότε απορρίπτουμε τη μηδενική υπόθεση και άρα υπάρχει διαφοροποίηση ανάμεσα στις δύο ομάδες ασθενών. Σε κάτι αντίστοιχο καταλήγουμε και με τον έλεγχο του Wilcoxon αφού λάβαμε p-value=0,01. Οπότε έχουμε αρκετές ενδείξεις για να θεωρήσουμε τη μεταβλητή protein στατιστικά σημαντική. Επομένως, η μεταβλητή αυτή επηρεάζει τη διάρκεια επιβίωσης των ασθενών που πάσχουν από πολλαπλό μυέλωμα.

5.3 Προσαρμογή του μοντέλου του Cox

Αρχικά καλούμε τις απαραίτητες βιβλιοθήκες που θα μας χρειαστούν και εισάγουμε τα δεδομένα στην R με τη βοήθεια των ακόλουθων εντολών:

```
data<-read.table("C:/Users/wwwva/OneDrive/Υπολογιστής/ΕΡΓΑΣΙΑ/mult-myel.txt",
header=TRUE)
attach(data)
data
```

Για να προσαρμόσουμε το μοντέλο του Cox εργαστήκαμε χρησιμοποιώντας την παρακάτω εντολή:

```
model_cox_1<-coxph(Surv(time,status)~age+sex+bun+ca+hb+pcells+protein,
ties="breslow")
model_cox_1
summary(model_cox_1)
```

Και τα αποτελέσματα που λάβαμε φαίνονται στον Πίνακα 5.17.

	Συντελεστής	Τυπικό Σφάλμα	Στατιστικό ελέγχου Wald	p - value
age	-0.019358	0.027924	-0.693	0.488159
sex2	-0.250899	0.402286	-0.624	0.532836
bun	**0.020826**	**0.005929**	**3.513**	**0.000443**
ca	0.013125	0.132442	0.099	0.921061
hb	**-0.135241**	**0.068891**	**-1.963**	**0.049635**
pcells	-0.001594	0.006577	-0.242	0.808533
protein1	-0.640438	0.426687	-1.501	0.133367

Πίνακας 5.17: Αποτελέσματα από την προσαρμογή του μοντέλου του Cox με όλες τις επεξηγηματικές μεταβλητές

Αξίζει να αναφερθεί ότι στις κατηγορικές μεταβλητές η R επιλέγει σαν κατηγορία αναφοράς την πρώτη κατηγορία κάθε φορά. Έτσι όταν εμφανίζεται το όνομα της επεξηγηματικής μεταβλητής sex_1 γνωρίζουμε ότι η R αναφέρεται στην κατηγορία εκείνη στην οποία έχουμε αναθέσει την τιμή 0 άρα στις γυναίκες. Αντίθετα, η κατηγορία sex_2 έχει να κάνει με τους άντρες. Επίσης η κατηγορική μεταβλητή protein λαμβάνει τις τιμές 0 και 1. Όπως αναφέρθηκε και στην ανάλυση του δείγματος, η μεταβλητή $protein_1$ αφορά τους ασθενείς που δεν έχουν την πρωτεΐνη στο αίμα τους (και η μεταβλητή λαμβάνει την τιμή 0) ενώ η μεταβλητή $protein_2$ έχει να κάνει με τα άτομα που εμφανίζουν την συγκεκριμένη πρωτεΐνη στις εξετάσεις τους (και η μεταβλητή λαμβάνει την τιμή 1).

Προτού αναλύσουμε τα αποτελέσματα που λάβαμε θα κάνουμε μία μικρή υπενθύμιση του στατιστικού ελέγχου Wald. Κατά αρχάς ο έλεγχος αυτός χρησιμοποιείται για να βγάλουμε το συμπέρασμα ποιες από τις επεξηγηματικές μεταβλητές είναι στατιστικά σημαντικές και ποιες όχι. Η μηδενική υπόθεση του ελέγχου αναφέρει ότι ο συντελεστής της μεταβλητής είναι ίσος με 0 και άρα η μεταβλητή αυτή δεν είναι στατιστικά σημαντική. Από την άλλη η εναλλακτική υπόθεση υποστηρίζει ότι ο συντελεστής της μεταβλητής είναι διάφορος του μηδενός, οπότε η μεταβλητή αυτή είναι στατιστικά σημαντική.

H_0: β_i=0, δηλαδή η μεταβλητή x_i δε συμβάλει στο μοντέλο

H_1: $\beta_i \neq 0$, δηλαδή η μεταβλητή x_i συμβάλει στο μοντέλο

Ο έλεγχος γίνεται σε (1-α)% επίπεδο σημαντικότητας και η στατιστική συνάρτηση υπό την H_0 είναι η ακόλουθη:

$$\frac{\hat{\beta}_i}{se(\hat{\beta}_i)}$$

Η οποία καλείται Wald, ακολουθεί ασυμπτωτικά την κανονική κατανομή, ενώ το τετράγωνο της ακολουθεί την χ^2 κατανομή με 1 βαθμό ελευθερίας. Αν η p-value του ελέγχου Wald κριθεί ότι έχει λάβει μικρή τιμή τότε απορρίπτεται η μηδενική υπόθεση και έτσι ο συντελεστής β_i δεν μηδενίζεται οπότε η μεταβλητή x_i θεωρείται στατιστικά σημαντική. Σε αντίθετη περίπτωση, δηλαδή όταν η τιμή της p θεωρηθεί μεγάλη δεν απορρίπτουμε τη μηδενική υπόθεση.

Από τον παραπάνω πίνακα παρατηρούμε ότι η p-value του ελέγχου Wald είναι μικρότερη από 0,05 για τις μεταβλητές hb και bun. Επομένως για τις δύο αυτές μεταβλητές απορρίπτουμε τη μηδενική υπόθεση άρα αυτές οι μεταβλητές είναι στατιστικά σημαντικές και αρά επιδρούν στη διάρκεια ζωής των ασθενών. Αντίθετα, για τις μεταβλητές age, sex, ca, pcells και protein έχουμε μικρές τιμές για την p-value και άρα δεχόμαστε τη μηδενική υπόθεση. Οι μεταβλητές αυτές είναι στατιστικά μη σημαντικές.

Στη συνέχεια παραθέτουμε τον Πίνακα 5.18 στον οποίο παρουσιάζονται τα διαστήματα εμπιστοσύνης για κάθε μία μεταβλητή.

	2.5 %	97.5 %
age	**-0.074087881**	**0.0353718880**
sex2	**-1.039364690**	**0.5375675154**
bun	**0.009206017**	**0.0324457938**
ca	**-0.246456948**	**0.2727062174**
hb	**-0.270265317**	**-0.0002157272**
pcells	**-0.014484848**	**0.0112972859**
protein1	**-1.476729286**	**0.1958527923**

Πίνακας 5.18: Διαστήματα εμπιστοσύνης για το μοντέλο του Cox με όλες τις μεταβλητές

Παρατηρούμε ότι τα διαστήματα εμπιστοσύνης για τις μεταβλητές age, sex, ca, pcells και protein περιέχουν το μηδέν. Η ύπαρξη του μηδενός στα διαστήματα εμπιστοσύνης σε όλες τις παραπάνω μεταβλητές υποδεικνύει ότι αυτές είναι στατιστικά μη σημαντικές όπως άλλωστε είχαμε διαπιστώσει και νωρίτερα από τον Πίνακα 5.18. Σε αντιδιαστολή με τα παραπάνω τα διαστήματα εμπιστοσύνης των μεταβλητών bun και hb δεν περιλαμβάνουν το μηδέν, γεγονός που επαληθεύει ότι οι επεξηγηματικές αυτές είναι στατιστικά σημαντικές και άρα επηρεάζουν τον χρόνο επιβίωσης των ασθενών.

Στα αποτελέσματα μας δόθηκε η τιμή του ελέγχου του λόγου πιθανοφανειών για την σύγκριση του μοντέλου που περιέχει και τις 7 συμμεταβλητές και αυτού που δεν περιέχει καμία επεξηγηματική μεταβλητή. Ο έλεγχος αυτός έχει τις ακόλουθες υποθέσεις:

H_0: $\beta_1=\beta_2=\ldots=\beta_7=0$, όλοι οι συντελεστές είναι ίσοι με το μηδέν

$H_1: \beta_i \neq 0$, τουλάχιστον ένας από τους συντελεστές είναι διάφορος του μηδενός

Το στατιστικό του ελέγχου για το λόγο των πιθανοφανειών δίνεται ως εξής:

$$w = -2(\hat{l}_0 - \hat{l}_1) \sim X^2_d$$

όπου $\hat{l}_0$ η μεγιστοποιημένη λογαριθμοποιημένη συνάρτηση μερικής πιθανοφάνειας στο μοντέλο χωρίς καμία μεταβλητή και $\hat{l}_1$ η μεγιστοποιημένη λογαριθμοποιημένη συνάρτηση μερικής πιθανοφάνειας στο μοντέλο που περιλαμβάνει όλες τις μεταβλητές. Οι βαθμοί ελευθερίας d της κατανομής είναι η διαφορά των παραμέτρων ανάμεσα στα δύο μοντέλα, οπότε στην προκειμένη περίπτωση έχουμε d=7.

Η R για τον έλεγχο αυτό μας έδωσε p-value = 0.02, επομένως απορρίπτουμε την μηδενική υπόθεση και άρα το μοντέλο που περιλαμβάνει όλες τις συμμεταβλητές είναι προτιμότερο από το μοντέλο που δεν περιέχει καμία μεταβλητή. Συμπεραίνουμε λοιπόν ότι ενώ δεν είναι όλες οι μεταβλητές στατιστικά σημαντικές όπως προέκυψε από τον έλεγχο του Wald, η παρουσία όλων στο μοντέλο φαίνεται να βελτιώνει την προσαρμογή του.

5.3.1 Εύρεση βέλτιστου μοντέλου με τη διαδικασία διαδοχικής αφαίρεσης

Μία μέθοδος για να καταλήξουμε στο καταλληλότερο μοντέλο για την περιγραφή των δεδομένων είναι της διαδοχικής αφαίρεσης (backward elimination). Σύμφωνα με αυτήν προσαρμόζεται ένα μοντέλο το οποίο περιέχει όλες τις μεταβλητές και σε κάθε βήμα αφαιρείται η μεταβλητή που είναι περισσότερο στατιστικά μη σημαντική. Το μοντέλο στο οποίο καταλήγει η διαδικασία της διαδοχικής αφαίρεσης είναι το τελικό μοντέλο από το οποίο δεν μπορούμε να αφαιρέσουμε άλλες μεταβλητές καθώς όσες έχουν μείνει είναι στατιστικά σημαντικές.

Η εντολή με τη βοήθεια της οποίας πραγματοποιήσαμε τη διαδικασία της διαδοχικής αφαίρεσης είναι η ακόλουθη:

```
step(model_cox_1, direction = 'backward',test = 'Chisq')
```

	Df	AIC	Μεταβολή της $-2\hat{l}$	p-value
ca	**1**	**211.71**	**0.0098**	**0.921275**
pcells	**1**	**211.76**	**0.0591**	**0.807854**
sex	**1**	**212.09**	**0.3929**	**0.530797**
age	**1**	**212.18**	**0.4737**	**0.491303**
καμία	**-**	**213.70**	**-**	**-**
protein	**1**	**214.06**	**2.3608**	**0.124416**
hb	**1**	**215.54**	**3.8387**	**0.050081.**
bun	**1**	**221.19**	**9.4859**	**0.002071 ****

Πίνακας 5.19: Πίνακας αποτελεσμάτων του πρώτου βήματος του backward elimination για το μοντέλο του Cox

Από τον Πίνακα 5.19 παρατηρούμε ότι όταν αφαιρούμε τη μεταβλητή ca από το μοντέλο, το μοντέλο βελτιώνεται αφού μειώνεται η τιμή του κριτηρίου AIC. Ακόμα η p-value του ελέγχου των πιθανοφανειών είναι μεγάλη οπότε μπορούμε να αφαιρέσουμε τη μεταβλητή ca από το μοντέλο. Η τιμή για το AIC του βήματος αυτού είναι ίση με 213.7.

Έτσι περνάμε στο επόμενο βήμα της διαδικασίας που παρατίθεται στον Πίνακα 5.20.

	Df	AIC	Μεταβολή της $-2\hat{l}$	p-value
pcells	**1**	**209.77**	**0.0552**	**0.814190**
sex	**1**	**210.16**	**0.4448**	**0.504820**
age	**1**	**210.18**	**0.4660**	**0.494831**
καμία	**-**	**211.71**	**-**	**-**
protein	**1**	**212.06**	**2.3535**	**0.125000**
hb	**1**	**213.59**	**3.8837**	**0.048758 ***
bun	**1**	**219.26**	**9.5485**	**0.002001 ****

Πίνακας 5.20: Πίνακας αποτελεσμάτων του δεύτερου βήματος του backward elimination για το μοντέλο του Cox

Από τον Πίνακα 5.20 παρατηρούμε πως όταν αφαιρούμε τη μεταβλητή «pcells» από το μοντέλο, το μοντέλο βελτιώνεται αφού μειώνεται η τιμή του κριτηρίου AIC. Ακόμα η p-value του ελέγχου των λόγων πιθανοφανειών είναι μεγάλη οπότε μπορούμε να αφαιρέσουμε τη μεταβλητή «pcells» από το μοντέλο. Η τιμή για το AIC του βήματος αυτού είναι ίση με 211.71.

Ακολούθως, περνάμε στο επόμενο βήμα τους διαδικασίας στον Πίνακα 5.21.

	Df	AIC	Μεταβολή της $-2\hat{l}$	p-value
sex	**1**	**208.17**	**0.4004**	**0.526865**
age	**1**	**208.19**	**0.4234**	**0.515221**
καμία	**-**	**209.77**	**-**	**-**
protein	**1**	**210.11**	**2.3462**	**0.125590**
hb	**1**	**211.60**	**3.8367**	**0.050142 .**
bun	**1**	**217.28**	**9.5137**	**0.002039 ****

Πίνακας 5.21: Πίνακας αποτελεσμάτων του τρίτου βήματος του backward elimination για το μοντέλο του Cox

Από τον Πίνακα 5.21 παρατηρούμε ότι όταν αφαιρούμε τη μεταβλητή «sex» από το μοντέλο, το μοντέλο βελτιώνεται αφού μειώνεται η τιμή του κριτηρίου AIC. Ακόμα η p-value του ελέγχου των πιθανοφανειών είναι μεγάλη οπότε μπορούμε να αφαιρέσουμε τη μεταβλητή «sex» από το μοντέλο. Η τιμή για το AIC του βήματος αυτού είναι ίση με 209.77.

Στη συνέχεια, περνάμε στο επόμενο βήμα τους διαδικασίας το οποίο παρατίθεται στον Πίνακα 5.22.

	Df	AIC	Μεταβολή της $-2\hat{l}$	p-value
age	**1**	**206.50**	**0.3363**	**0.561991**
καμία	**-**	**208.17**	**-**	**-**
protein	**1**	**208.67**	**2.5028**	**0.113641**
hb	**1**	**209.60**	**3.4368**	**0.063761 .**
bun	**1**	**215.78**	**9.6162**	**0.001929 ****

Πίνακας 5.22: Πίνακας αποτελεσμάτων του τέταρτου βήματος του backward elimination για το μοντέλο του Cox

Από τον Πίνακα 5.22 παρατηρούμε ότι όταν αφαιρούμε τη μεταβλητή «age» από το μοντέλο, το μοντέλο βελτιώνεται αφού μειώνεται η τιμή του κριτηρίου AIC. Ακόμα η p-value του ελέγχου των πιθανοφανειών είναι μεγάλη οπότε μπορούμε να αφαιρέσουμε τη μεταβλητή «age» από το μοντέλο. Η τιμή για το AIC του βήματος αυτού είναι ίση με 208.17.

Ακολούθως, περνάμε στο πέμπτο βήμα τους διαδικασίας στο οποίο έχουμε αφαιρέσει τη μεταβλητή «age». Το τελευταίο βήμα παρατίθεται στον Πίνακα 5.23.

	Df	AIC	Μεταβολή της $-2\hat{l}$	p-value
καμία	**-**	**206.50**	**-**	**-**
protein	**1**	**206.94**	**2.4356**	**0.118609**
hb	**1**	**207.64**	**3.1382**	**0.076480 .**
bun	**1**	**213.83**	**9.3266**	**0.002259 ****

Πίνακας 5.23: Πίνακας αποτελεσμάτων του πέμπτου βήματος του backward elimination για το μοντέλο του Cox

Από τα αποτελέσματα του Πίνακα 5.23 συμπεραίνουμε ότι όλες οι μεταβλητές που απέμειναν στο μοντέλο είναι στατιστικά σημαντικές και δεν μπορούν να αφαιρεθούν από το μοντέλο. Παρατηρούμε ότι η συμμεταβλητή «protein» η οποία φαίνεται να είναι η λιγότερο στατιστικά σημαντική θα συμπεριληφθεί στο τελικό μοντέλο. Αυτό συμβαίνει διότι αν αφαιρεθεί, το νέο μοντέλο που θα προκύψει θα έχει τιμή του κριτηρίου AIC μεγαλύτερη από το τελευταίο μοντέλο. Όπως γνωρίζουμε, το κριτήριο AIC μας δίνει ένα μέτρο σύγκρισης της καταλληλόλητας μοντέλων με διαφορετικό πλήθος μεταβλητών. Μεταξύ υποψήφιων μοντέλων διαλέγουμε εκείνο με την μικρότερη τιμή του κριτηρίου AIC. Στο βήμα αυτό η τιμή του κριτηρίου AIC λαμβάνει τη μικρότερη τιμή του η οποία είναι ίση με 206.5. Επομένως, ο αλγόριθμος σταματάει διότι οποιαδήποτε επιπλέον συμμεταβλητή κι αν αφαιρεθεί, δεν καταλήγουμε σε ένα καλύτερο μοντέλο. Επομένως φτάσαμε στο τελικό μας μοντέλο το οποίο είναι και το καταλληλότερο.

Τελικά η R με την βοήθεια του κριτηρίου AIC κατέληξε στο βέλτιστο μοντέλο το οποίο περιέχει τις μεταβλητές: «protein», «hb» και «bun». Δηλαδή καταλήγουμε στο συμπέρασμα ότι οι μεταβλητές που επηρεάζουν το χρόνο επιβίωσης των ασθενών που πάσχουν από πολλαπλό μυέλωμα είναι η ύπαρξη ή όχι τους μονοκλωνικής πρωτεΐνης, η τιμή τους αιμοσφαιρίνης και τα επίπεδα του αζώτου ουρίας αίματος. Αντιθέτως οι παράγοντες: ηλικία, φύλο, ποσότητα του ασβεστίου στον οργανισμό του ασθενούς και το ποσοστό κυττάρων πλάσματος στο μυελό των οστών δεν είναι στατιστικά σημαντικοί και άρα δεν επιδρούν στη χρονική διάρκεια που θα ζήσουν οι ασθενείς μετά τη διάγνωσή τους. Ο πίνακας των συγκεντρωτικών αποτελεσμάτων δίνεται στον Πίνακα 5.24.

	β	**exp(β)**	**se(β)**	**z**	**p-value**
bun	0.020150	1.020354	0.005746	3.507	0.000454
hb	-0.110490	0.895395	0.061992	-1.782	0.074694
protein1	-0.616665	0.539742	0.406033	-1.519	0.128824

Πίνακας 5.24: Συγκεντρωτικός πίνακας αποτελεσμάτων για το βέλτιστο μοντέλο

Τελικά το βέλτιστο προσαρμοσμένο μοντέλου του Cox δίνεται ακολούθως.

$$h(t;x) = h_0(t) \cdot \exp\left(0.020150\text{bun} - 0.110490hb - 0.616665\, protein1\right)$$

5.3.2 Ερμηνεία συντελεστών

Αφού καταλήξαμε στο βέλτιστο μοντέλο του Cox θα το ξαναπροσαρμόσουμε στα δεδομένα με τη βοήθεια της εντολής

```
model_cox_2 <- coxph(Surv(time,status)~bun+hb+protein, ties="breslow")
```

Και στη συνέχεια θα αναλύσουμε το μοντέλο αυτό για να εξάγουμε συμπεράσματα για τους συντελεστές ώστε να εκτιμήσουμε την επίδραση κάθε μεταβλητής στη διάρκεια ζωής των ασθενών.

```
summary(model_cox_2)
```

Ο πίνακας που λαμβάνουμε θα μας βοηθήσει να εξάγουμε τα απαραίτητα αποτελέσματα. Ο πίνακας αυτός είναι εκείνος που λάβαμε στην προηγούμενη ενότητα και πιο συγκεκριμένα ο Πίνακας 5.25.

Για την ερμηνεία των συντελεστών του μοντέλου αρχικά θα εξηγήσουμε τον τρόπο με τον οποίο καταλαβαίνουμε ποιες συμμεταβλητές επιδρούν θετικά και ποιες αρνητικά στη διάρκεια ζωής των ασθενών που πάσχουν από πολλαπλό μυέλωμα. Αρχικά υπενθυμίζουμε ότι το μοντέλο του Cox δίνεται από τον τύπο:

$$h(t;\vec{x}) = h_0(t) \cdot \exp\left(\vec{\beta}^T \vec{x}\right)$$

από την οποία παρατηρούμε ότι η ποσότητα $\exp(\vec{\beta}^T)$ πολλαπλασιάζεται με τη συνάρτηση κινδύνου $h_0(t)$. Επομένως, μπορούμε να βγάλουμε συμπέρασμα για το αν η υπό εξέταση μεταβλητή συμβάλει θετικά ή αρνητικά στο χρονικό διάστημα επιβίωσης των ασθενών, εφόσον θεωρήσουμε ότι όλες οι άλλες μεταβλητές παραμένουν σταθερές.

Όπως είναι λογικό, όταν η τιμή της ποσότητας $\exp(\beta_i)$ είναι μεγαλύτερη της μονάδας τότε η συνάρτηση κινδύνου αυξάνεται και άρα η i-οστή μεταβλητή επιδρά αρνητικά στη διάρκεια ζωής των ασθενών.

Αντιθέτως, όταν η τιμή της ποσότητας $\exp(\beta_i)$ είναι μικρότερη της μονάδας τότε η συνάρτηση κινδύνου μειώνεται και άρα η i-οστή μεταβλητή επιδρά θετικά στη διάρκεια ζωής των ασθενών.

Τέλος, όταν η τιμή της ποσότητας $\exp(\beta_i)$ είναι ίση με τη μονάδα τότε η συνάρτηση κινδύνου διατηρείται σταθερή, οπότε η i-οστή μεταβλητή δεν συμβάλει στη διάρκεια ζωής των ασθενών.

Όσον αναφορά το μοντέλο αυτό παρατηρούμε ότι για τη μεταβλητή «bun» έχουμε $\exp(\beta_1)$ = 1.020354 > 1 , οπότε η ποσότητα του αζώτου στον ορό της ουρίας επιδρά αρνητικά στη διάρκεια ζωής των ασθενών αλλά όχι σε πολύ μεγάλο βαθμό καθώς η τιμή αυτή είναι αρκετά κοντά στην τιμή 1. Από την άλλη για τη μεταβλητή «hb» δίνεται $\exp(\beta_2)$ = 0.895395 < 1 , άρα η τιμή της αιμοσφαιρίνης επηρεάζει θετικά τη διάρκεια επιβίωσης των ασθενών. Αξίζει να σημειωθεί ότι και πάλι η επιρροή δεν είναι μεγάλη καθώς η τιμή 0.895 βρίσκεται αρκετά κοντά στην τιμή 1. Τέλος, τα άτομα που δεν εμφανίζουν στο αίμα τους τη μονοκλωνική πρωτεΐνη φαίνεται να έχουν καλύτερη έκβαση στην πορεία της υγείας τους καθώς η τιμή $\exp(\beta_3)$ = 0.539742 είναι μικρότερη του 1 και άρα η συνάρτηση διακυνδύνευσης μειώνεται, επομένως η κατηγορία 1 της μεταβλητής αυτής φαίνεται να επιδρά θετικά στη διάρκεια ζωής των ασθενών.

Ας δούμε και πιο αναλυτικά κατά πόσο επιδρά η κάθε μεταβλητή στη συνάρτηση διακινδύνευσης. Αρχικά, ο συντελεστής της μεταβλητής «bun» (ποσότητα αζώτου στο αίμα) είναι ίσος με 0.020150, οπότε exp(0.020150) = 1.020354. Συνεπώς, η συνάρτηση διακινδύνευσης αλλάζει κατά $\mathrm{h_0(t)}\exp(0.020150) = \mathrm{h_0(t)} \cdot 1.020354$ αν η ποσότητα του αζώτου στο αίμα αυξηθεί κατά μία μονάδα και η τιμή της αιμοσφαιρίνης και της ύπαρξης της μονοκλωνικής πρωτεΐνης παραμείνουν σταθερά.

Όσο αναφορά τη μεταβλητή «hb» (η τιμή της αιμοσφαιρίνης) ο συντελεστής της είναι ίσος με -0.110490, οπότε exp(-0.110490) = 0.895395. Συνεπώς, αν αυξηθεί κατά μία μονάδα η τιμή της αιμοσφαιρίνης και η τιμή της ποσότητας του αζώτου στο αίμα και της ύπαρξη της μονοκλωνικής πρωτεΐνης παραμείνουν σταθερά η συνάρτηση διακινδύνευσης αλλάζει κατά $\mathrm{h_0(t)} \cdot \exp(-0.110490) = \mathrm{h_0(t)} \cdot 0.895395.$

Τέλος για την μεταβλητή «protein1» ο συντελεστής της είναι ίσος με -0.616665, οπότε exp(-0.616665) = 0.539742. Συνεπώς, αν ο ασθενής περάσει από την τιμή της protein=0 στην τιμή της protein=1 και η τιμή της ποσότητας του αζώτου στο αίμα και της ύπαρξης της μονοκλωνικής πρωτεΐνης παραμείνουν σταθερά η συνάρτηση διακινδύνευσης αλλάζει κατά $\mathrm{h_0(t)} \cdot \exp(-0.616665) = \mathrm{h_0(t)} \cdot 0.539742.$

Στη συνέχεια θα κατασκευάσουμε τα διαστήματα εμπιστοσύνης για τους συντελεστές των μεταβλητών του βέλτιστου μοντέλου με τις εντολές.

```
confint(model_cox_2)

exp(confint(model_cox_2))
```

	2.5 % για το β	97.5 % για το β
bun	**0.008887796**	**0.03141196**
hb	**-0.231991121**	**0.01101111**
protein1	**-1.412475644**	**0.17914587**

Πίνακας 5.25: 95% Διαστήματα εμπιστοσύνης για τους συντελεστές β

	2.5 % για το exp(β)	97.5 % για το exp(β)
bun	**1.0089274**	**1.031911**
hb	**0.7929532**	**1.011072**
protein1	**0.2435396**	**1.196195**

Πίνακας 5.26: 95% Διαστήματα εμπιστοσύνης για exp(β)

5.3.3 Γραφικός έλεγχος των υπολοίπων Schoenfeld

Για να πραγματοποιήσουμε τον έλεγχο αναλογικής διακινδύνευσης χρησιμοποιούμε τα κλιμακοποιημένα υπόλοιπα Schoenfeld. Στόχος μας είναι να ελέγξουμε αν οι συντελεστές των ανεξάρτητων μεταβλητών εξαρτώνται από το χρόνο. Η ιδιότητα της αναλογικής διακινδύνευσης για το μοντέλο του Cox μπορεί να διατυπωθεί ισοδύναμα και ως:

$$\beta_i\,(t) = \beta_i\,, \forall t$$

Επομένως, για να αποδεχτούμε την παραπάνω υπόθεση $(\beta_i\,(t) = \beta_i\,, \forall t)$ θα πρέπει η γραφική παράσταση να είναι μία οριζόντια γραμμή, δηλαδή να υπάρχει ανεξαρτησία των τιμών της ως προς το χρόνο t.

Αρχικά κατασκευάζουμε έναν πίνακα ελέγχου της ισχύς υπόθεσης της αναλογικής διακινδύνευσης για κάθε επεξηγηματική μεταβλητή του τελικού βέλτιστου μοντέλου, χρησιμοποιώντας το μετασχηματισμό “identity”. Ο μετασχηματισμός “identity” αξιοποιεί της ταυτοτικούς χρόνους αποκοπής t_j. Η εντολή που μας βοήθησε να λάβουμε τα αποτελέσματα του Πίνακα 5.27 ήταν η ακόλουθη.

```
Vv <- cox.zph(model_cox_2,transform='identity', terms=FALSE)
Vv
```

	chisq	df	p
bun	0.3586	1	0.55
hb	0.0937	1	0.76
protein1	0.8669	1	0.35
GLOBAL	2.5254	3	0.47

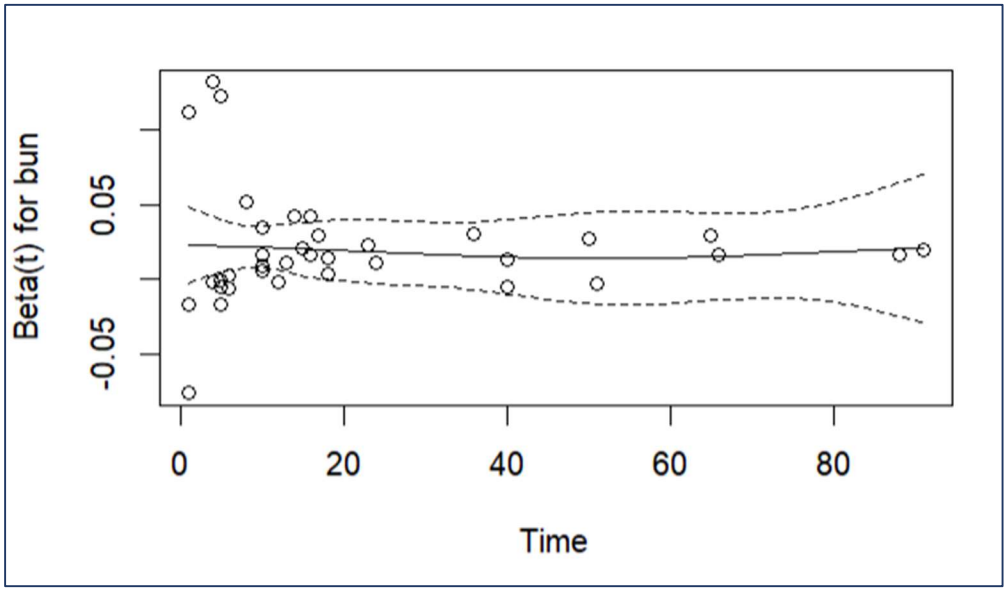
Πίνακας 5.27: Πίνακας ελέγχου της υπόθεσης της αναλογικής διακινδύνευσης για το βέλτιστο μοντέλο με τον μετασχηματισμό "identity"

Η πρώτη στήλη του Πίνακα 5.27 αναφέρεται στην τιμή του X^2 ελέγχου, ενώ η τρίτη στήλη δίνει τα p-values του ελέγχου αυτού. Παρατηρούμε ότι για τις τρεις συμμεταβλητές η p τιμή του ελέγχου είναι μεγάλη και άρα δεχόμαστε τη μηδενική υπόθεση. Οπότε ισχύει η υπόθεση της αναλογικής διακινδύνευσης και για τις τρεις μεταβλητές. Το αποτέλεσμα αυτό επιβεβαιώνεται και από τον έλεγχο Global καθώς και εκεί έχουμε p-value=0,47 οπότε οι συντελεστές των μεταβλητών του βέλτιστου μοντέλου δεν εξαρτώνται από τον χρόνο.

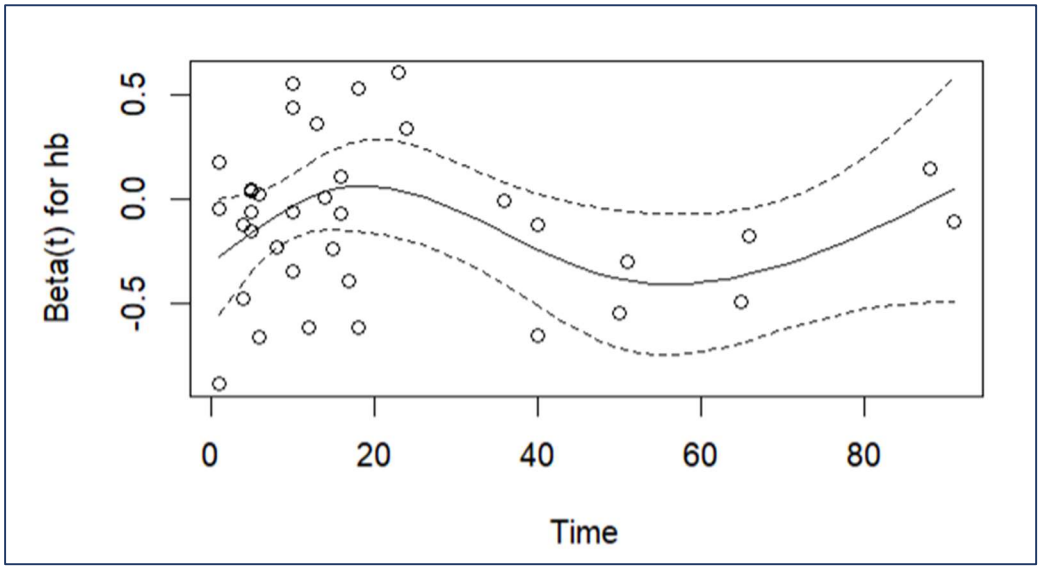
Σε κάτι αντίστοιχο πρέπει να καταλήξουμε και με τις γραφικές παραστάσεις $\left(r_{ij}^{*} + \beta_j\right)$ ως της t_j για κάθε συμμεταβλητή του μοντέλου που κατασκευάσαμε με τη βοήθεια της εντολής

```
plot(vv)
```

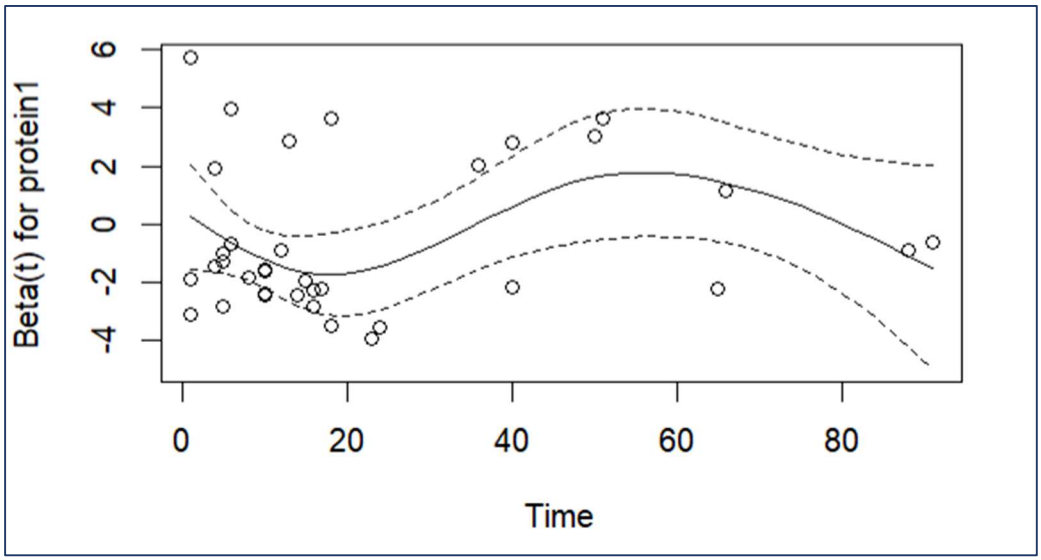
Και τα γραφήματα που λάβαμε παρατίθενται παρακάτω.



Διάγραμμα 5.8: Έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης για την μεταβλητή bun με τον μετασχηματισμό "identity"



Διάγραμμα 5.9: Έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης για την μεταβλητή hb με τον μετασχηματισμό "identity"



Διάγραμμα 5.10: : Έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης για την μεταβλητή protein με τον μετασχηματισμό "identity"

Στα Διαγράμματα 5.8, 5.9 και 5.10 η συμπαγής γραμμή είναι αποτέλεσμα της λεγόμενης εξομάλυνσης μέσω συναρτήσεων splines. Οι συναρτήσεις splines, αποτελούνται από δύο ή περισσότερες συνεχόμενες, συνήθως τρίτου βαθμού καμπύλες ή τόξα τα οποία ενώνονται μεταξύ της με ομαλό τρόπο. Οι διακεκομμένες γραμμές σχηματίζουν μια περιοχή $\pm$ 2 τυπικών αποκλίσεων από την εξομαλυμένη καμπύλη.

Από το Διάγραμμα 5.8 παρατηρούμε ότι η γραφική παράσταση είναι οριζόντια γραμμή οπότε για τη μεταβλητή «bun» ισχύει η υπόθεση της ανεξαρτησίας από το χρόνο. Όσον αναφορά της μεταβλητές «hb» και «protein» παρατηρούμε ότι στα Διαγράμματα 5.9 και 5.10 υπάρχει μία μικρή καμπυλότητα η οποία μας δημιουργεί αμφιβολίες για το αν ισχύει η υπόθεση της αναλογικής διακινδύνευσης για τις μεταβλητές αυτές.

Για το λόγο αυτό καταφεύγουμε στην επανάληψη της παραπάνω διαδικασίας με τη διαφορά ότι τώρα δεν θα χρησιμοποιήσουμε το μετασχηματισμό identity. Έτσι χρησιμοποιούμε τον προκαθορισμένο από τη R μετασχηματισμό των χρόνων διακοπής, που είναι ο μετασχηματισμός “km”, δηλαδή ο μετασχηματισμός της Kaplan-Meier. Για να κατασκευάσουμε τον πίνακα ελέγχου της ισχύς υπόθεσης της αναλογικής διακινδύνευσης για κάθε επεξηγηματική μεταβλητή του τελικού βέλτιστου μοντέλου, χρησιμοποιώντας το μετασχηματισμό “km” τρέξαμε την ακόλουθη εντολή.

```
Vv <- cox.zph(model_cox_2, terms=FALSE)
Vv
```

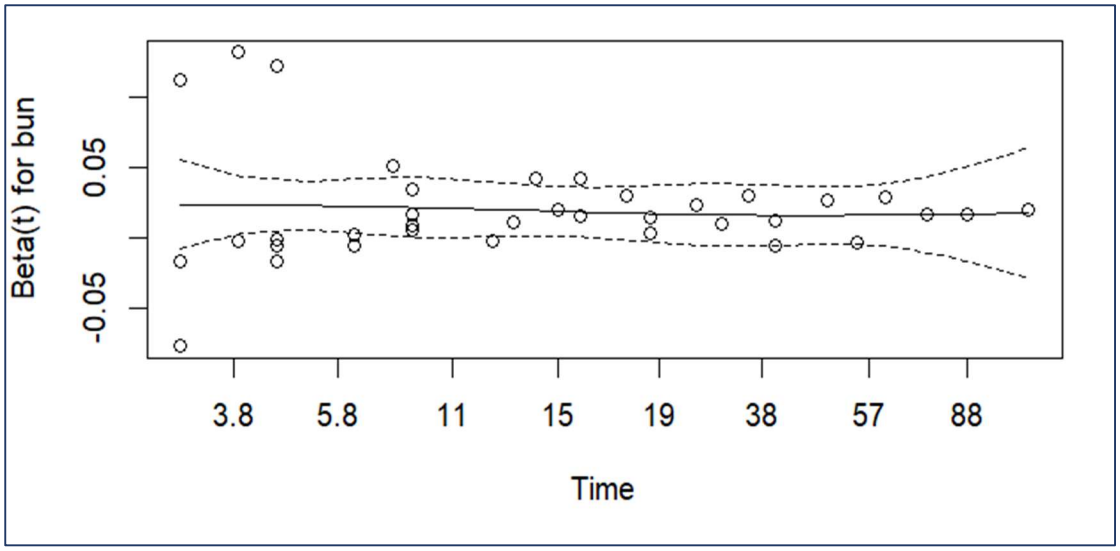
	chisq	df	p
bun	0.6187	1	0.43
hb	0.0158	1	0.90
protein1	0.1933	1	0.66
GLOBAL	1.0818	3	0.78

Πίνακας 5.28: Πίνακας ελέγχου της υπόθεσης αναλογικής διακινδύνευσης για το βέλτιστο μοντέλο με τον μετασχηματισμό “km”

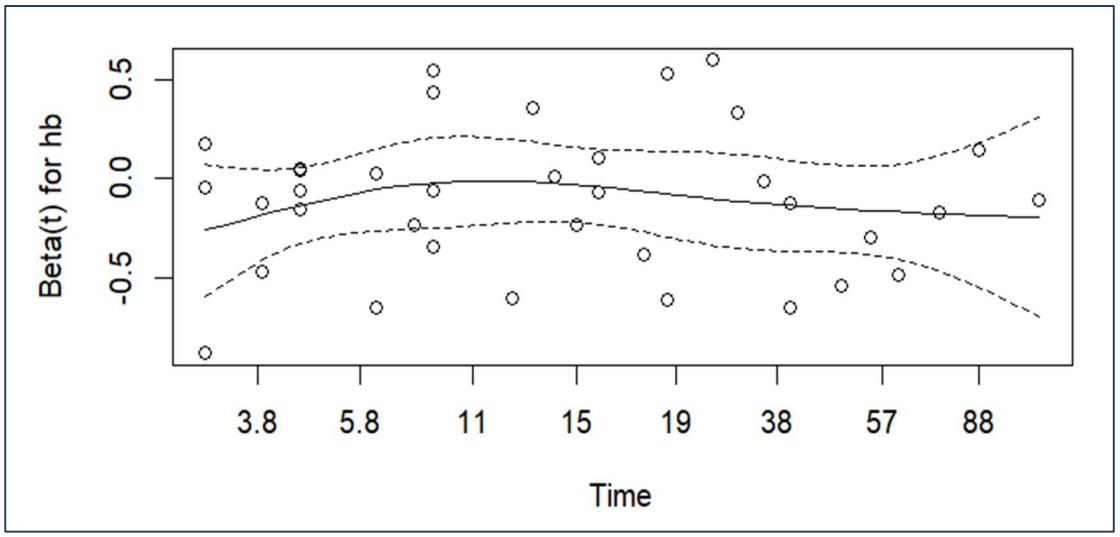
Από τον Πίνακα 5.28 παρατηρούμε ότι για τις τρεις συμμεταβλητές η p τιμή του ελέγχου είναι μεγάλη και άρα δεχόμαστε τη μηδενική υπόθεση. Άρα ισχύει η υπόθεση της αναλογικής διακινδύνευσης και για τις τρεις μεταβλητές. Το αποτέλεσμα αυτό επιβεβαιώνεται και από τον έλεγχο Global καθώς και εκεί έχουμε p-value=0,78 οπότε οι συντελεστές των μεταβλητών του βέλτιστου μοντέλου δεν εξαρτώνται από τον χρόνο.

Σε κάτι αντίστοιχο πρέπει να καταλήξουμε και με τις γραφικές παραστάσεις $(r_{ij}^{*} + \beta_j)$ ως προς t_j για κάθε συμμεταβλητή του μοντέλου που κατασκευάσαμε με τη βοήθεια της εντολής

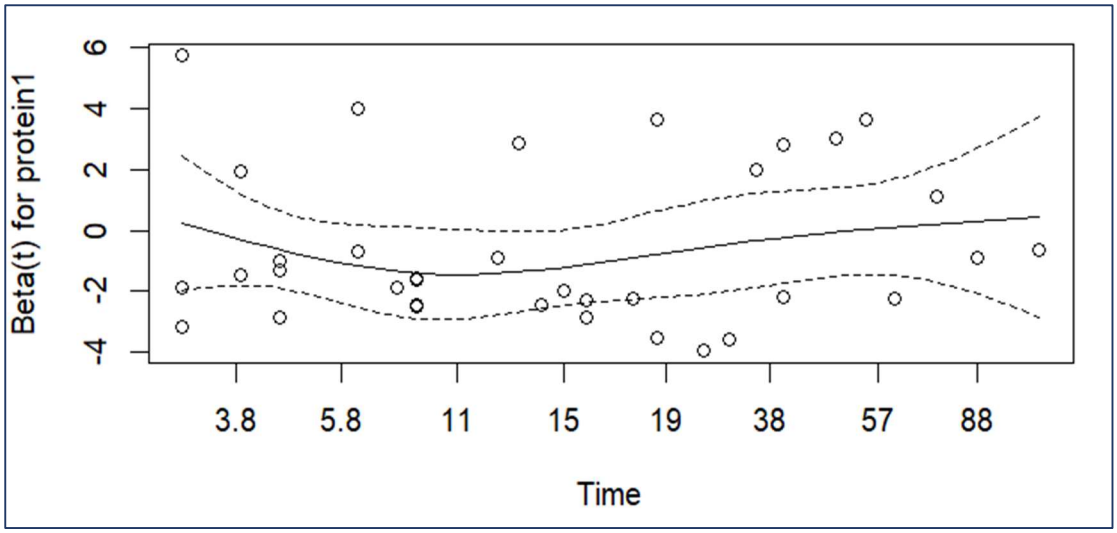
```
plot(vv)
```



Διάγραμμα 5.11: Έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης για την μεταβλητή bun με τον μετασχηματισμό "km"



Διάγραμμα 5.12: Έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης για την μεταβλητή hb με τον μετασχηματισμό "km"



Διάγραμμα 5.13: Έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης για την μεταβλητή protein με τον μετασχηματισμό "km"

Από τα Διαγράμματα 5.11, 5.12 και 5.13 είναι πλέον φανερό ότι και οι τρεις μεταβλητές του βέλτιστου μοντέλου είναι ανεξάρτητες από το χρόνο καθώς και τα τρία παραπάνω γραφήματα είναι σχεδόν μία οριζόντια γραμμή. Αξίζει να σημειωθεί ότι και σε αυτή την περίπτωση τα αποτελέσματα είναι ιδανικότερα για τη μεταβλητή «bun» καθώς η γραφική της συγκεκριμένης μεταβλητής είναι μία εντελώς οριζόντια γραμμή.

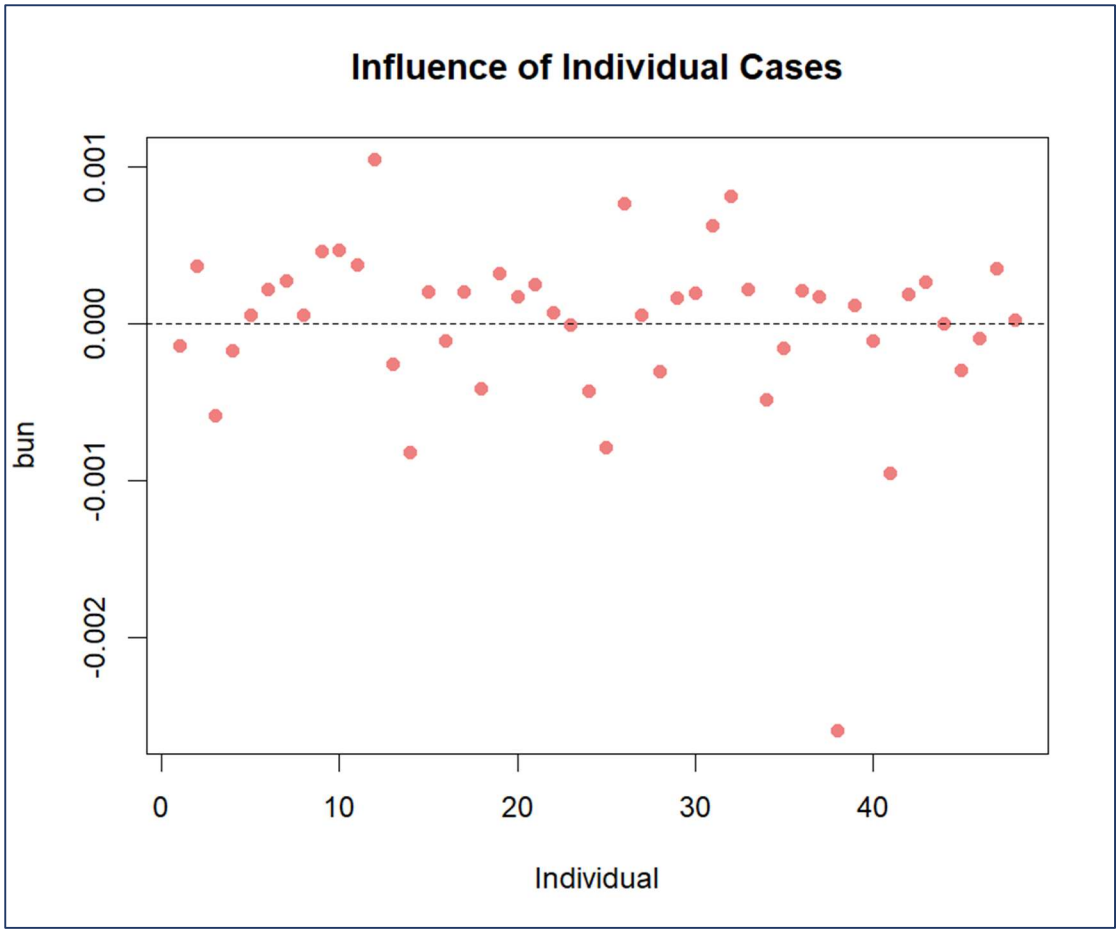
5.3.4 Σημεία επιρροής

Ως σημεία επιρροής χαρακτηρίζονται οι παρατηρήσεις εκείνες που όταν τις αφαιρέσουμε από το δείγμα μας τα αποτελέσματα που προκύπτουν είναι σημαντικά διαφοροποιημένα από τα αρχικά. Για να βρούμε τα σημεία επιρροής υπολογίζουμε τα dfbeta στο βέλτιστο μοντέλο στο οποίο έχουμε καταλήξει. Τις ποσότητες αυτές τις πήραμε με τη βοήθεια της εντολής:

```
dfbeta<-residuals(model_cox_2, type="dfbeta")
```

Στη συνέχεια κατασκευάσαμε τις γραφικές παραστάσεις για να συμπεράνουμε ποιες παρατηρήσεις επηρεάζουν την εκτίμηση των παραμέτρων του μοντέλου. Αρχικά για τη μεταβλητή bun έχουμε:

```
plot(dfbeta[,1], ylab="bun", xlab="Individual", col="lightcoral", pch=16,
main="Influence of Individual Cases")
 abline(h=0, lty=2)
```

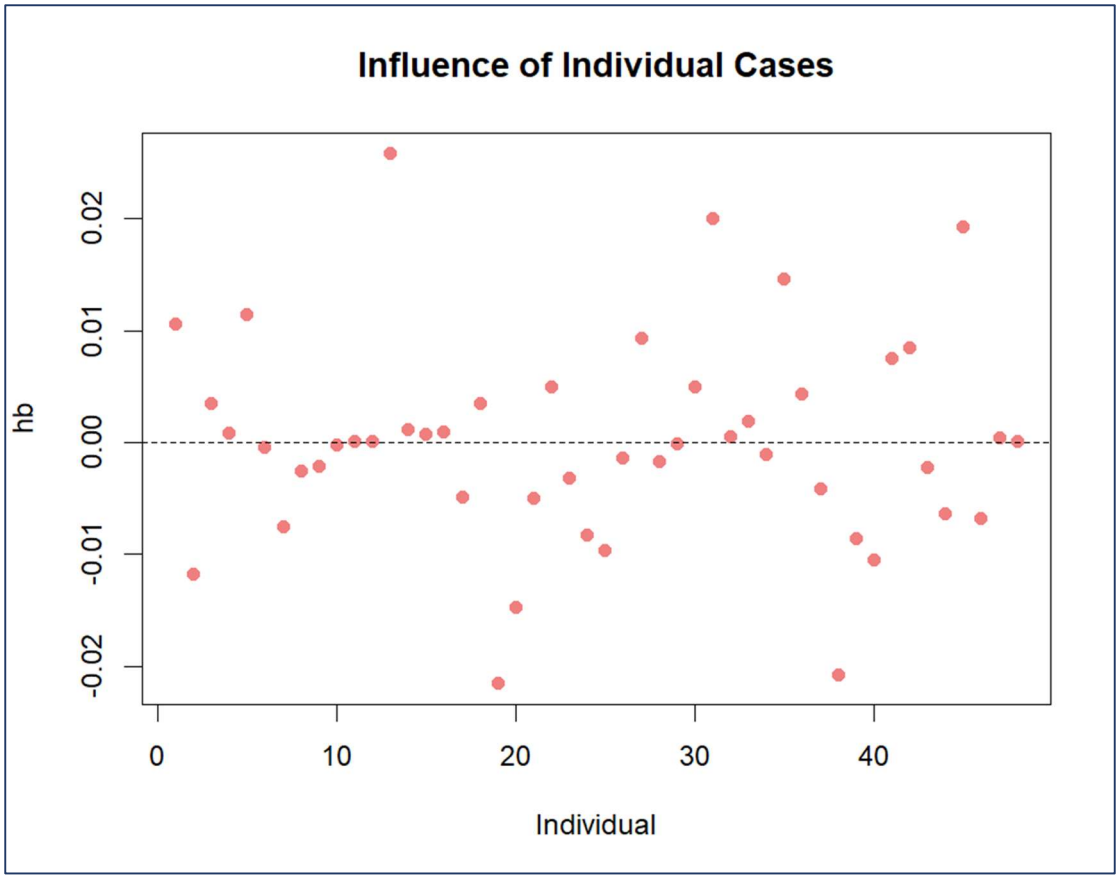


Διάγραμμα 5.14: Έλεγχος για τον εντοπισμό σημείων επιρροής για τη μεταβλητή bun του βέλτιστου μοντέλου

Από το Διάγραμμα 5.14 παρατηρούμε ότι ο 38ος ασθενής αποτελεί ένα σημείο επιρροής καθώς απέχει αρκετά από το μηδέν. Οριακά μπορούμε να πούμε ότι και ο 12ος και ο 41ος ασθενείς είναι σημεία επιρροής καθώς συγκριτικά με τις υπόλοιπες παρατηρήσεις απέχουν λίγο μεγαλύτερη απόσταση από το μηδέν.

Στη συνέχεια για τη μεταβλητή hb έχουμε:

```
plot(dfbeta[,2], ylab="hb", xlab="Individual", col="lightcoral", pch=16,
main="Influence of Individual Cases")
 abline(h=0, lty=2)
```

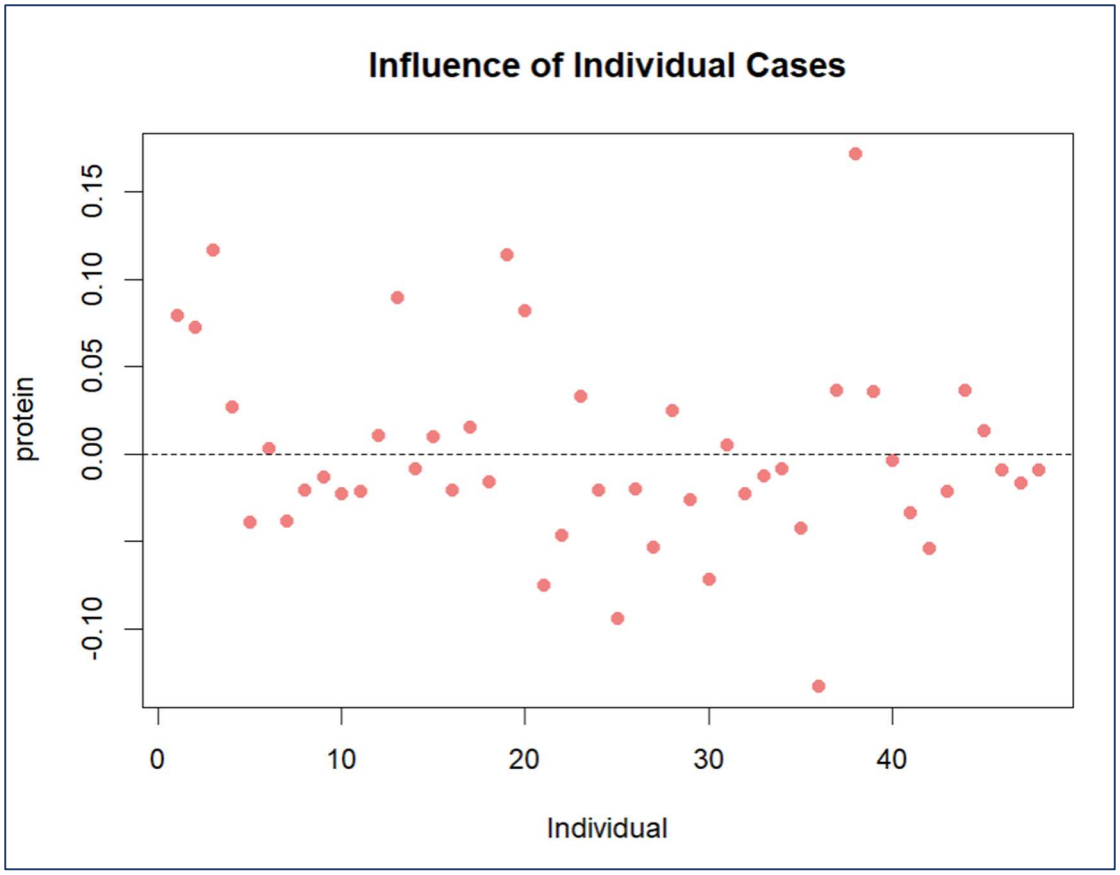


Διάγραμμα 5.15: Έλεγχος για τον εντοπισμό σημείων επιρροής για τη μεταβλητή hb του βέλτιστου μοντέλου

Στο Διάγραμμα 5.15 φαίνεται να είναι σημεία επιρροής ο 13ος, ο 19ος και ο 37ος ασθενής.

Τέλος για τη μεταβλητή protein:

```
plot(dfbeta[,3], ylab="pcells", xlab="Individual", col="lightcoral", pch=16,
main="Influence of Individual Cases")
 abline(h=0, lty=2)
```



Διάγραμμα 5.16: Έλεγχος για τον εντοπισμό σημείων επιρροής για τη μεταβλητή protein του βέλτιστου μοντέλου

Από το Διάγραμμα 5.16 ο 38ος και ο 36ος ασθενής φαίνεται να αποτελούν σημεία επιρροής καθώς απέχουν από το μηδέν σχετικά μεγάλη απόσταση.

Τελικά τα σημεία επιρροής δεν είναι πολλά σε σχέση με το μέγεθος του δείγματος. Χρησιμοποιώντας την ακόλουθη εντολή στην R μπορούμε να ελέγξουμε ποια από τα προαναφερόμενα σημεία αποτελούν όντως σημεία επιρροής.

```
which(dfbeta>0.29)
 a<-cbind(bun,time)
 max(a[bun==0,])
 mean(a[bun==0,])

 which(dfbeta>0.29)
 a<-cbind(hb,time)
 max(a[hb==0,])
 mean(a[hb==0,])

 which(dfbeta>0.29)
 a<-cbind(protein,time)
 max(a[protein==0,])
 mean(a[protein==0,])
```

Για τον έλεγχο αυτό έχουμε υπόψιν ότι ένα σημείο είναι σημείο επιρροής όταν η τιμή της ποσότητας dfbeta για την παρατήρηση αυτή είναι μεγαλύτερη από την ποσότητα $2/\sqrt{n}$, όπου n είναι το μέγεθος του δείγματος, δηλαδή όταν:

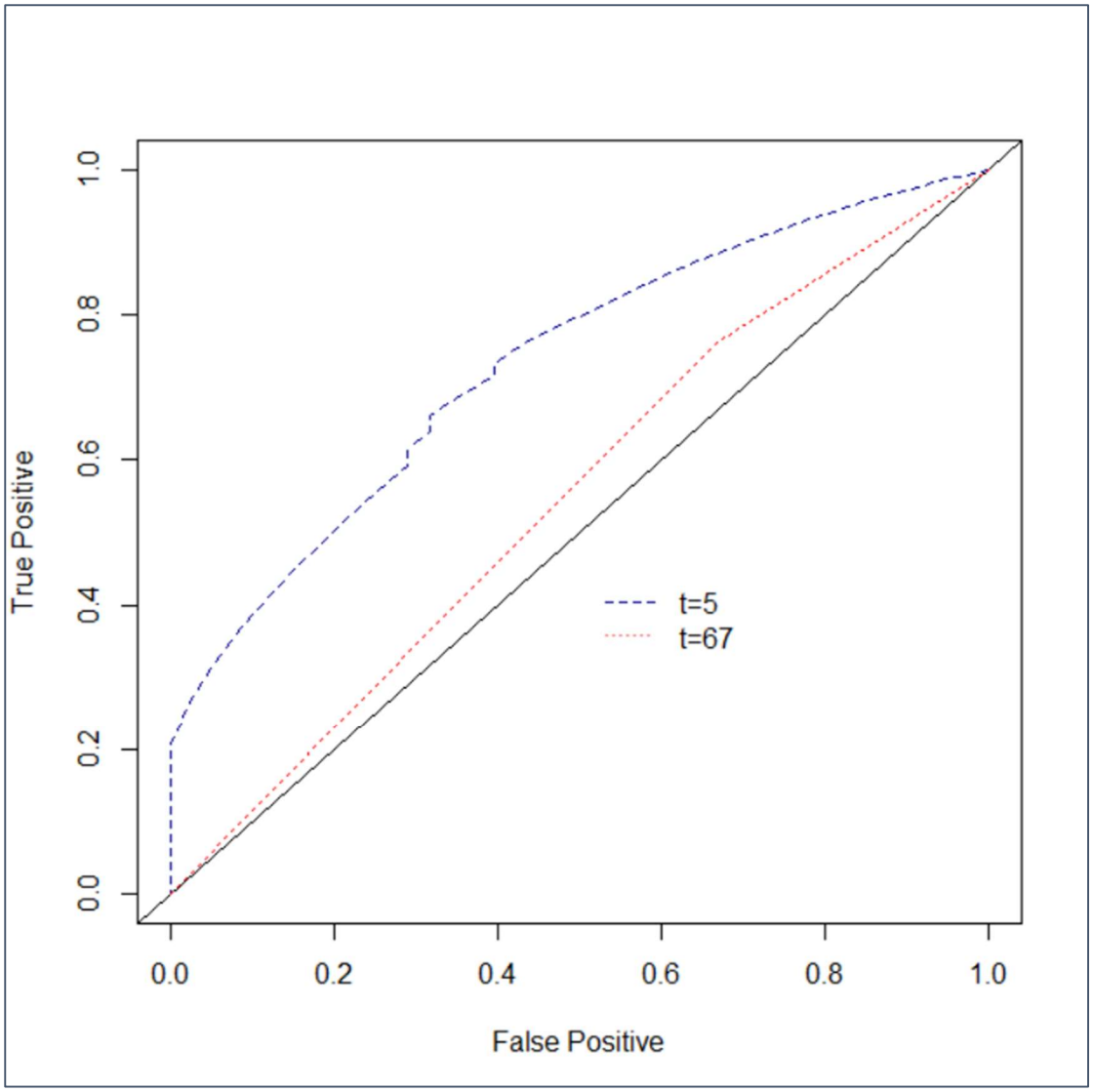
$$\text{dfbeta} > 2/\sqrt{48}$$

Από τα αποτελέσματα λάβαμε ότι κανένα σημείο δεν είναι σημείο επιρροής. Δηλαδή καμία παρατήρηση δεν επηρεάζει σε μεγάλο βαθμό τα αποτελέσματά μας. Οπότε οποιονδήποτε ασθενή και αν αφαιρέσουμε δεν θα διαφοροποιηθούν πολύ τα αποτελέσματα μας.

5.3.5 Προβλεπτική ικανότητα του μοντέλου

Για να ελέγξουμε την προβλεπτική ικανότητα του μοντέλου κατασκευάζουμε την καμπύλη ROC. Για το λόγο αυτό με τις παρακάτω εντολές κατασκευάσαμε δύο καμπύλες. Η μπλε καμπύλη αφορά τη χρονική στιγμή t = 5 μήνες, ενώ η κόκκινη καμπύλη έχει να κάνει με μία αρκετά μεγαλύτερη χρονική στιγμή που είναι ίση με t = 67 μήνες. Πήραμε δύο τιμές για το χρόνο μία μικρή χρονική στιγμή και μία πολύ μεγαλύτερη της πρώτης ώστε να ελέγξουμε πως μεταβάλλεται η προβλεπτική ικανότητα του μοντέλου από μικρούς χρόνους σε μεγάλους χρόνους.

```
install.packages("risksetROC")
 library(risksetROC)

 time<-c(data[,2])

  eta<-model_cox_2$linear.predictor
  ROC5=risksetROC(Stime=c(data[,2]),status=c(data[,3]),marker=eta,
predict.time=5, method="Cox",
              lty=2,col="darkblue",ylab="True Positive",xlab="False Positive")

ROC67=risksetROC(Stime=c(data[,2]),status=c(data[,3]),marker=eta,predict.time=6
7,method="Cox",
              lty=3,col="lightcoral",ylab="True Positive",xlab="False
Positive",plot=FALSE)
  lines(ROC67$FP,ROC67$TP, lty=3, col="red")
legend(.5,.45,lty=c(2,3),col=c("darkblue","lightcoral"),legend=c("t=5","t=67"),bty=
"n")
```

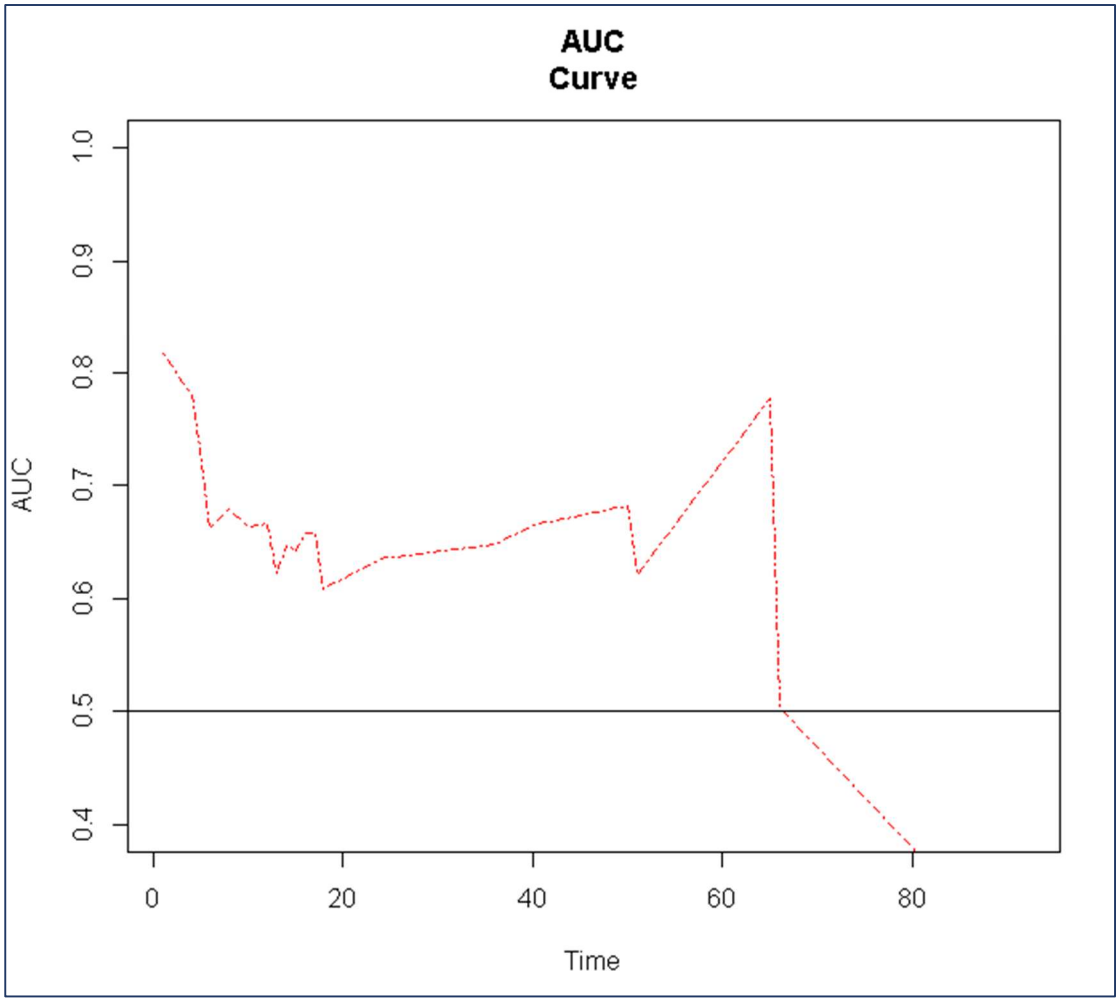


Διάγραμμα 5.17: Καμπύλη ROC του τελικού μοντέλου για t=5 και t=67 μήνες

Υπενθυμίζουμε ότι όσο μεγαλύτερο είναι το εμβαδόν κάτω από την καμπύλη που σχηματίζεται τόσο μεγαλύτερη είναι και η προβλεπτική ικανότητα του μοντέλου. Το εμβαδόν που δημιουργείται κάτω από την καμπύλη ROC ονομάζεται AUC (Area Under the Curve) και όσο πιο κοντά είναι η τιμή του στο ένα τόσο καλύτερη είναι και η προβλεπτική ικανότητα του μοντέλου. Ακόμα αξίζει να σημειωθεί ότι το εμβαδό που δημιουργεί η γκρι γραμμή θα είναι πάντα ίσο με 0,5. Επομένως η τιμή του AUC θα κυμαίνεται από την τιμή 0,5 ως την τιμή 1. Από το Διάγραμμα 5.17 παρατηρούμε ότι το μοντέλο έχει καλύτερη προβλεπτική ικανότητα για μικρούς χρόνους από ότι για μεγάλους χρόνους. Πιο συγκεκριμένα το εμβαδόν που σχηματίζεται κάτω από την μπλε γραμμή είναι ίσο με 0,68 ($\mathrm{AUC}_{\mathrm{t=5}} = 0{,}68$) ενώ το εμβαδόν κάτω από την κόκκινη είναι ίσο με 0,52 ($\mathrm{AUC}_{\mathrm{t=67}} = 0{,}52$). Ο λόγος που συμβαίνει αυτό είναι ότι για μεγάλους χρόνους δεν έχουμε στη διάθεση μας πολλά δεδομένα καθώς αρκετοί από τους ασθενείς έχουν ήδη φύγει από τη ζωή.

Στη συνέχεια κατασκευάσαμε ένα διάγραμμα του AUC σαν συνάρτηση του χρόνου. Η εντολή που χρησιμοποιήσαμε ήταν η ακόλουθη

```
risksetAUC(Stime=c(data[,2]),status=c(data[,3]),marker=eta,method="Cox",
tmax=91,main="AUCCurve",lty=10,col="red")
```



Διάγραμμα 5.18: Καμπύλη AUC συναρτήσει του χρόνου

Από το Διάγραμμα 5.18 παρατηρούμε ότι η τιμή του εμβαδού μειώνεται με την πάροδο του χρόνου κάτι που ήταν άλλωστε αναμενόμενο. Βέβαια η πτωτική τάση του εμβαδού δεν είναι καθολική καθώς σε ορισμένα διαστήματα παρατηρούμε ότι η γραφική είναι αύξουσα ενώ σε άλλα είναι φθίνουσα. Παρόλα αυτά η τιμή του AUC λαμβάνει την ελάχιστη τιμή της η οποία είναι ίση με 0.5 τη χρονική στιγμή t = 68 μήνες. Συμπεραίνουμε λοιπόν ότι μετά τον 68^{ο} μήνα η προβλεπτική ικανότητα του μοντέλου είναι σχεδόν μηδενική. Αυτό όπως αναφέραμε και παραπάνω ήταν αναμενόμενο καθώς όσο περνάει ο χρόνος «χάνουμε» δεδομένα αφού όλο και περισσότεροι ασθενείς έχουν φύγει από τη ζωή.

5.4 Εφαρμογή μεθόδων ποινών στο μοντέλο του Cox

Όταν οι μεταβλητές του μοντέλου φαίνεται να σχετίζονται μεταξύ τους παρουσιάζεται το πρόβλημα της πολυσυγγραμμικότητας. Η συσχέτιση των μεταβλητών έχει ως αποτέλεσμα η επιρροή της κάθε μεταβλητής, που σχετίζεται με κάποια άλλη, στο μοντέλο να εξασθενεί. Για το λόγο αυτό στην επόμενη παράγραφο θα ελέγξουμε ποιες μεταβλητές σχετίζονται μεταξύ τους σε μεγάλο βαθμό και ποιες σε μικρότερο βαθμό. Στη συνέχεια θα επιδιώξουμε να ξεπεράσουμε τυχόν προβλήματα πολυσυγγραμμικότητας εφαρμόζοντας τρεις μεθόδους ποινών, την Lasso, την Ridge και την Elastic Net.

5.4.1 Έλεγχος πολυσυγγραμμικότητας

Για να ελέγξουμε ποιες μεταβλητές σχετίζονται μεταξύ τους εκτελούμε τις παρακάτω εντολές.

```
install.packages("corrplot")
library(corrplot)
dataframe<-data.frame(cbind(age,sex,bun,ca,hb,pcells,protein))
correlation_table<-cor(dataframe)
```

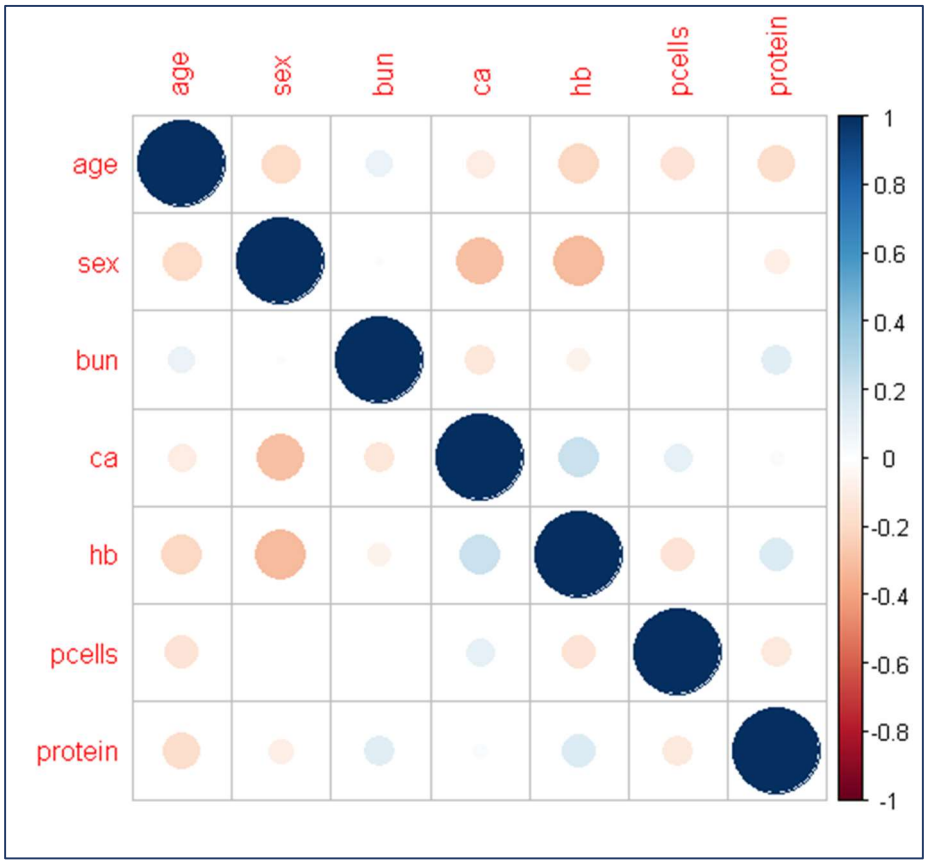
Και λαμβάνουμε τον Πίνακα 5.29 από τον οποίο παρατηρούμε ότι όλα τα διαγώνια στοιχεία στου πίνακα είναι ίσα με τη μονάδα αφού υπάρχει πλήρης εξάρτηση της κάθε μεταβλητής από τον εαυτό της. Γενικότερα όσο πιο κοντά στη μονάδα βρίσκεται η τιμή του κάθε κελιού τόσο μεγαλύτερη συσχέτιση υπάρχει και ανάμεσα στις συμμεταβλητές που βρίσκονται στις συγκεκριμένες γραμμές και στήλες. Έτσι έχουμε υψηλή θετική συσχέτιση ανάμεσα στις μεταβλητές «ca» και «hb» με τιμή ίση με 0.21, στις μεταβλητές «protein» και «hb» με τιμή 0.14 και στις μεταβλητές «bun» και «protein» με τιμή ίση με 0.12. Υψηλή αρνητική συσχέτιση υπάρχει ανάμεσα στις μεταβλητές «hb» και «sex» με τιμή -0.21, στις μεταβλητές «ca» και «sex» με τιμή ίση με -0.29 και τέλος στις μεταβλητές «hb» και «sex» με τιμή -0.32.

	age	sex	bun	ca	hb	pcells	protein
age	1.00000000	-0.185791817	0.089231799	-0.09979964	-0.20504961	-0.146032165	-0.17915196
sex	-0.18579182	1.000000000	-0.011290523	-0.29135236	-0.31773041	-0.004033536	-0.08616564
bun	0.08923180	-0.011290523	1.000000000	-0.12229262	-0.06775132	0.009469688	0.12049112
ca	-0.09979964	-0.291352363	-0.122292624	1.00000000	0.21383942	0.107936002	0.02936861
hb	-0.20504961	-0.317730406	-0.067751324	0.21383942	1.00000000	-0.145072268	0.14195106
pcells	-0.14603217	-0.004033536	0.009469688	0.10793600	-0.14507227	1.000000000	-0.11660300
protein	-0.17915196	-0.086165639	0.120491123	0.02936861	0.14195106	-0.116603002	1.00000000

Πίνακας 5.29 : Πίνακας ελέγχου της συσχέτισης των συμμεταβλητών του μοντέλου

Ελέγχουμε τα παραπάνω συμπεράσματα και γραφικά με την ακόλουθη εντολή

```
corrplot(correlation_table)
```



Διάγραμμα 5.19: Γραφικός έλεγχος του φαινομένου της πολυσυγγραμικότητας

Από το Διάγραμμα 19 επιβεβαιώνονται οι παρατηρήσεις μας για το ποιες μεταβλητές σχετίζονται περισσότερο η μία με την άλλη. Όσο μεγαλύτερος ο κύκλος στο Διάγραμμα 5.19 τόσο μεγαλύτερη και η συσχέτιση. Οι μπλε κύκλοι αφορούν τη θετική συσχέτιση, ενώ οι κόκκινοι κύκλοι την αρνητική συσχέτιση. Συνοψίζοντας οι μεταβλητές που φαίνεται να σχετίζονται περισσότερο με τις υπόλοιπες είναι οι: «hb» και «protein» οι οποίες ήταν και οι δύο από τις τρεις πιο στατιστικά σημαντικές συμμεταβλητές που προέκυψαν από το βέλτιστο μοντέλο του Cox.

5.4.2 Εφαρμογή της μεθόδου Lasso

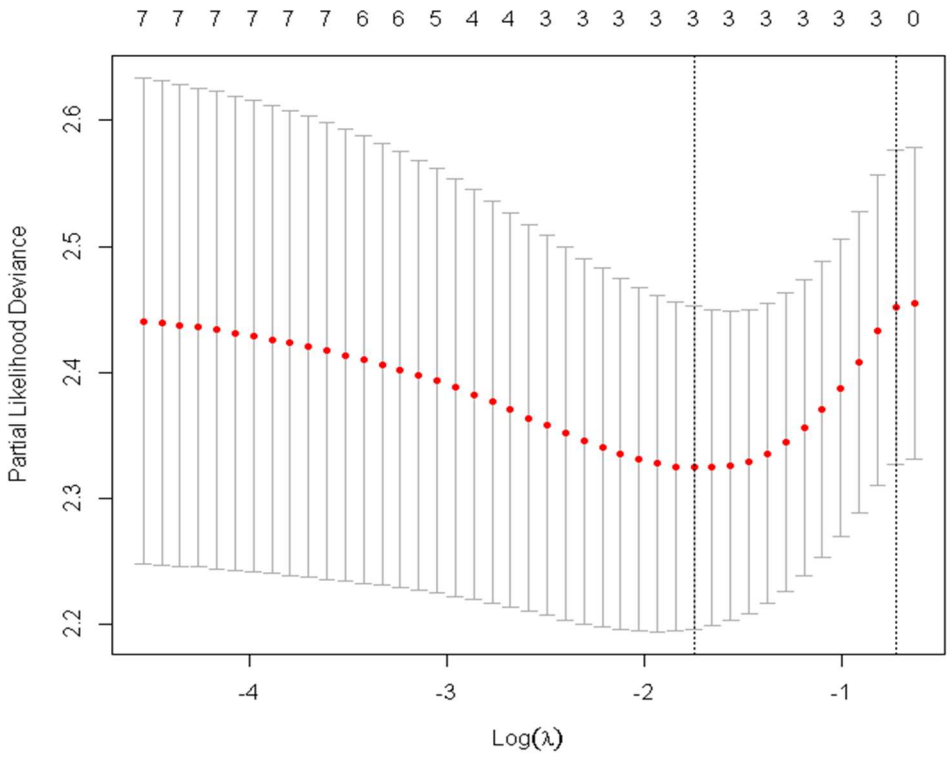
Για να αντιμετωπίσουμε το πρόβλημα της πολυσυγγραμμικότητας που παρουσιάστηκε στην προηγούμενη ενότητα θα εφαρμόσουμε τη μέθοδο ποινής Lasso. Με τη βοήθεια της μεθόδου αυτής θα ξανά προσαρμόσουμε το μοντέλο του Cox αφού πρώτα υπολογίσουμε τον βέλτιστο συντελεστή λ_1 με τη βοήθεια της μεθόδου cross validaton θεωρώντας α=1. Για την εύρεση της βέλτιστης τιμής του λ_1 όπως και για την κατασκευή του Διαγράμματος 5.20 χρειαστήκαμε τις ακόλουθες εντολές.

```
x<-cbind(age,sex,bun,ca,hb,pcells,protein)
y<-cbind(time, status)

install.packages('glmnet')
library(glmnet)

cv.lasso<-cv.glmnet(x,y,family="cox", alpha=1)
cv.lasso$lambda.min

plot(cv.lasso)
```

Η βέλτιστη τιμή του λ_1 προέκυψε ίση με 0.1744553 οπότε και log(λ)=log(0.1744553)= - 1.746087, ενώ το Διάγραμμα 5.20 που αναπαριστά την deviance της cross-validation συνάρτησης μερικής πιθανοφάνειας συναρτήσει της τιμής του log(λ) παρατίθεται παρακάτω.

Στο Διάγραμμα 5.20 παρατηρούμε δύο κάθετες γραμμές. Οι γραμμές αυτές απέχουν μεταξύ τους μία μονάδα και η αριστερή κάθετη γραμμή μας δίνει το λ για το οποίο η deviance της cvl παίρνει την ελάχιστη τιμή της, ενώ η δεξιά κάθετη γραμμή μας δίνει τη μεγαλύτερη τιμή της λ που έχει σφάλμα μέχρι μία τυπική απόκλιση από την ελάχιστη τιμή deviance της cvl.

Επιπρόσθετα, δίνεται το πλήθος των μη μηδενικών παραμέτρων του μοντέλου στο πάνω μέρος του Διαγράμματος. Παρατηρούμε ότι όσο αυξάνεται η ποσότητα log(λ) , κατ΄ επέκταση και το λ, τόσο μειώνεται το πλήθος των μη μηδενικών παραμέτρων. Έτσι όταν το λ λαμβάνει μικρές τιμές θα συμπεριλαμβάνουμε όλες τις συμμεταβλητές στο μοντέλο, ενώ για μεγαλύτερες τιμές του λ το πλήθος των συμμεταβλητών που περιλαμβάνουμε στο μοντέλο μειώνεται αισθητά.



Διάγραμμα 5.20: Deviance της cross-validation μερικής πιθανοφάνειας συναρτήσει του log(λ) (Lasso)

Κάνοντας χρήση των παρακάτω εντολών ξανά προσαρμόζουμε το μοντέλο του Cox εφαρμόζοντας της μέθοδο ποινής Lasso. Έτσι λαμβάνουμε τους συντελεστές του μοντέλου που φαίνονται στον πίνακα 5.31.

```
fit.lasso<-glmnet(x,y,family="cox", alpha=1)
coef(fit.lasso, s=cv.lasso$lambda.min)
```

Μεταβλητή	Συντελεστής
age	**0**
sex	**0**
bun	**0.01552760**
ca	**0**
hb	**-0.05866947**
pcells	**0**
protein	**-0.43927989**

Πίνακας 5.30: Παράμετροι του μοντέλου του Cox μετά την εφαρμογή της μεθόδου Lasso

Με την εφαρμογή της μεθόδου ποινής Lasso οι συντελεστές των μεταβλητών «age», «sex», «ca» και «pcells» έχουν μηδενιστεί. Αντιθέτως, οι συντελεστές των μεταβλητών «bun», «hb» και «protein» έχουν υποστεί συρρίκνωση. Οι μεταβλητές που δεν μηδενίστηκαν από τη μέθοδο Lasso είναι οι ίδιες με τις μεταβλητές που προέκυψαν στατιστικά σημαντικές όταν εφαρμόσαμε τη μέθοδο backward elimination στο μοντέλο του Cox.

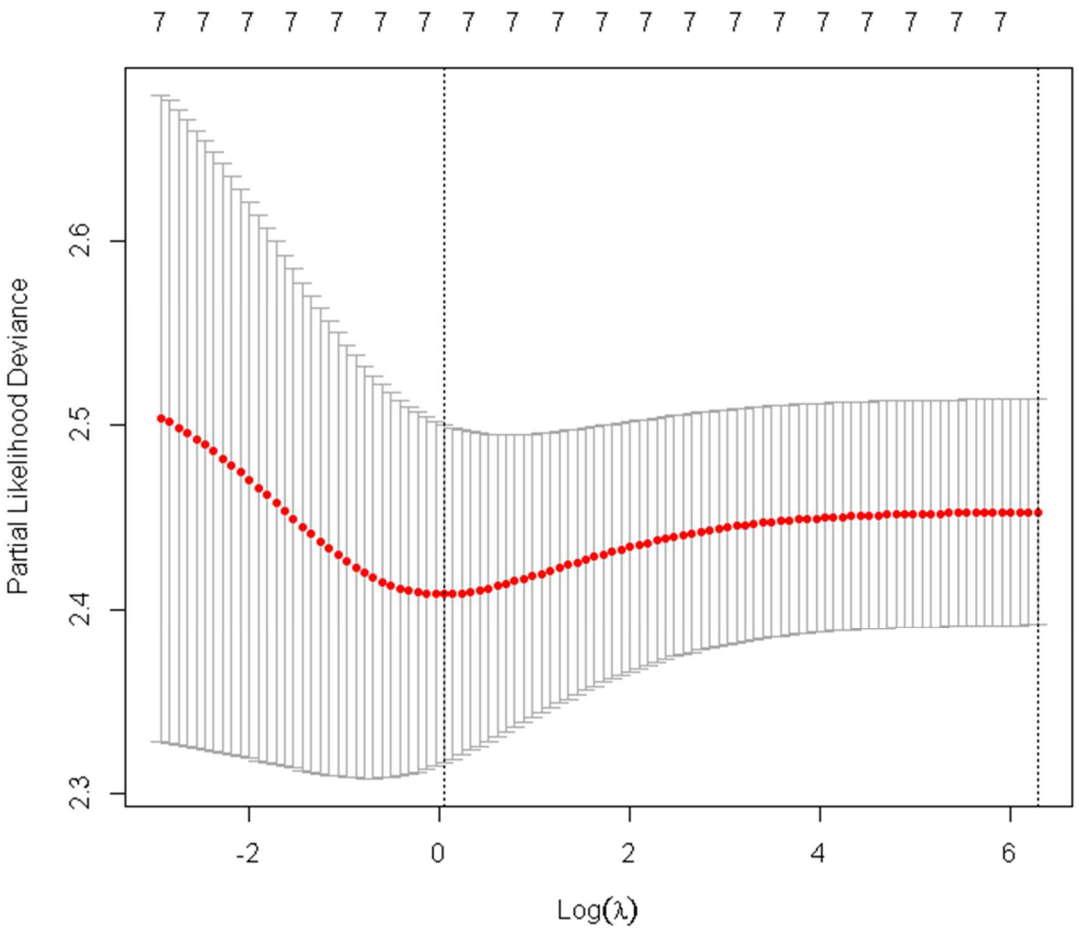
5.4.3 Εφαρμογή της μεθόδου Ridge

Η μέθοδος Ridge αποτελεί άλλη μία μέθοδος ποινής και δεν διαφέρει ιδιαιτέρα από τη μέθοδο Lasso. Η διαφορά της έγκειται στο γεγονός ότι τώρα θέτουμε α=0 και ψάχνουμε τη βέλτιστη τιμή του λ. Οι εντολές που χρειαστήκαμε ήταν οι ακόλουθες.

```
cv.ridge<-cv.glmnet(x,y,family="cox", alpha=0)
cv.ridge$lambda.min
```

Από τις οποίες λάβαμε ότι η βέλτιστη τιμή του λ είναι ίση με 1.045832 οπότε και log(λ)=log(1.045832)= 0.04481274. Για να κατασκευάσουμε τη γραφική παράσταση της deviance της cross-validation συνάρτησης μερικής πιθανοφάνειας συναρτήσει της τιμής του log(λ) τρέξαμε την ακόλουθη εντολή και λάβαμε το Διάγραμμα 21.

```
Plot(cv.ridge)
```



Διάγραμμα 5.21: Deviance της cross-validation μερικής πιθανοφάνειας συναρτήσει του log(λ) (Ridge)

Από το Διάγραμμα 5.21 παρατηρούμε ότι όσο και αν αυξηθεί η ποσότητα log(λ), οπότε κατ΄επέκταση και το λ, το πλήθος των παραμέτρων του μοντέλου παραμένει σταθερό και ίσο με 7. Δημιουργείται η υποψία ότι η μέθοδος Ridge κρατάει όλες τις συμμεταβλητές του μοντέλου. Για να επαληθεύσουμε τα αποτελέσματα που λάβαμε από το Διάγραμμα 5.21 θα προσαρμόσουμε εκ νέου το μοντέλο του Cox με τη μέθοδο Ridge και θα υπολογίσουμε τους συντελεστές για κάθε μεταβλητή.

```
Fit.ridge<-glmnet(x,y,family="cox", alpha=0)
coef(fit.ridge, s=cv.ridge$lambda.min)
```

Μεταβλητή	Συντελεστής
age	-0.0004258541
sex	-0.0173690315
bun	0.0842584514
ca	-0.0411337887
hb	-0.0601718447
pcells	0.0011333130
protein	-0.3584787240

Πίνακας 5.31: Παράμετροι του μοντέλου του Cox μετά την εφαρμογή της μεθόδου Ridge

Πράγματι καμία από τις μεταβλητές δεν μηδενίζεται, δηλαδή όλες οι συμμεταβλητές διατηρούνται στο μοντέλο. Αυτό μπορούμε να πούμε ότι είναι και ένα από τα μειονεκτήματα της μεθόδου Ridge καθώς δεν είναι ικανή να μηδενίσει τους συντελεστές των μεταβλητών που δε συμβάλλουν στο μοντέλο. Για το λόγο αυτό θα καταφύγουμε σε μία τελευταία μέθοδο ποινής τη μέθοδο Elastic Net.

5.4.4 Εφαρμογή της μεθόδου Elastic Net

Η μέθοδος ποινής Lasso έχει την τάση να μηδενίζει αρκετούς από τους συντελεστές των ανεξάρτητων μεταβλητών τη στιγμή που η μέθοδος Ridge αντιδρά με ακριβώς αντίθετο τρόπο. Η μέθοδος Ridge δεν μηδενίζει κανένα συντελεστή. Για να βρούμε μία μέση λύση εφαρμόζουμε τη μέθοδο ποινής Ridge η οποία αρχικά κατασκευάζει ένα διάνυσμα για το α το οποίο λαμβάνει τιμές από 0.01 έως 0.99. Στη συνέχεια για κάθε τιμή του διανύσματος α υπολογίζει τη βέλτιστη τιμή του λ εκτελώντας τη μέθοδο cross-validation. Και τελικά καταλήγει στα ολικά βέλτιστα α και λ. Για να λάβουμε τις βέλτιστες αυτές τιμές εκτελέσαμε τις εντολές.

```
# Δημιουργία ακολουθίας α
alphasOfInterest<-seq(0.01,0.99,by=0.01)

# Εκτέλεση της μεθόδου cross-validation για κάθε τιμή του α
cvs<-lapply(alphasOfInterest, function(curAlpha){cv.glmnet(x, y,
alpha=curAlpha,family="cox") })

#  Εύρεση του βέλτιστου λ για κάθε τιμή του α
optimumPerAlpha<-sapply(seq_along(alphasOfInterest), function(curi){
curcvs<-cvs[[curi]]
curAlpha<-alphasOfInterest[curi]
indOfMin<-match(curcvs$lambda.min, curcvs$lambda)
c(lam=curcvs$lambda.min, alph=curAlpha, cvup=curcvs$cvup[indOfMin])})

#  Εύρεση των βέλτιστων α και λ
posOfOptimum<-which.min(optimumPerAlpha["lam",])
overall.lambda.min<-optimumPerAlpha["lam",posOfOptimum]
overall.alpha.min<-optimumPerAlpha["alph",posOfOptimum]
```

Στη συνέχεια προσαρμόζουμε το μοντέλο του Cox με τη μέθοδο Elastic Net και υπολογίζουμε τους συντελεστές των ανεξάρτητων μεταβλητών.

```
fit.net<-glmnet(x,y,family="cox",alpha=overall.alpha.min)
coef(fit.net,s=overall.lambda.min)
```

Η βέλτιστη τιμή του α είναι ίση με 0.98 ενώ η βέλτιστη τιμή του λ είναι ίση με 0.1477917.

Μεταβλητή	Συντελεστής
age	**0**
sex	**0**
bun	**0.01643993**
ca	**0**
hb	**-0.06319458**
pcells	**0**
protein	**-0.48350664**

Πίνακας 5.32: Παράμετροι του μοντέλου του Cox μετά την εφαρμογή της μεθόδου Elastic Net

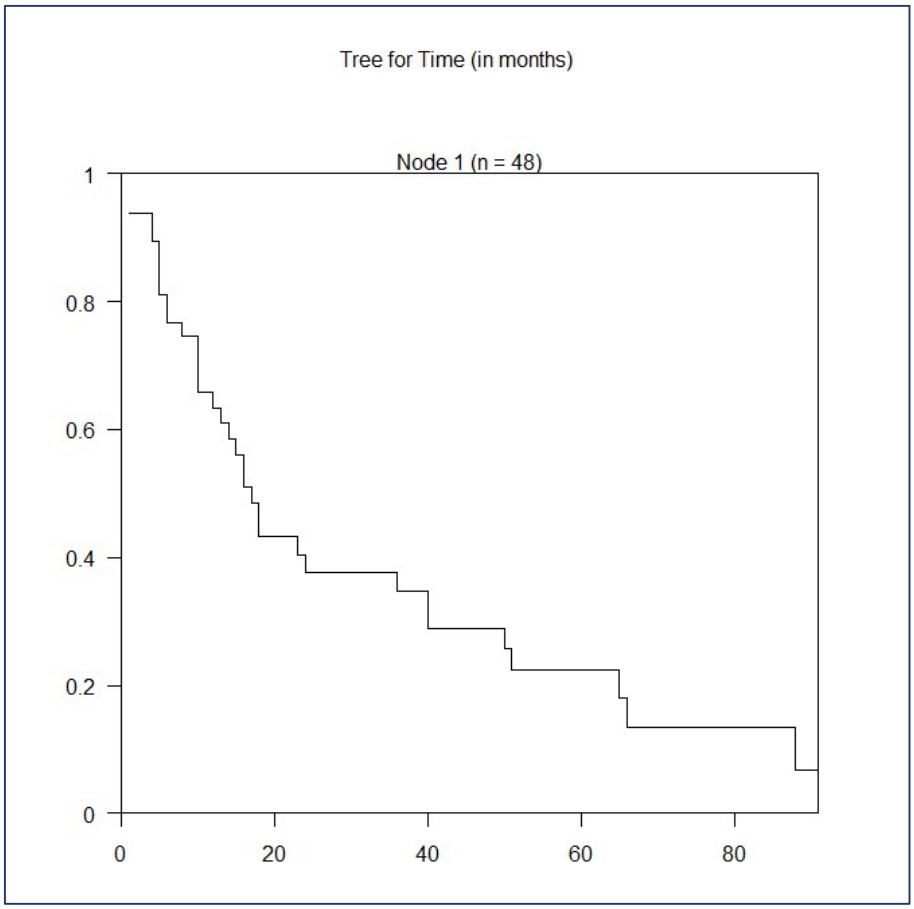
Από τον Πίνακα 5.32 γίνεται φανερό ότι όπως και με τη μέθοδο Lasso έτσι και με τη μέθοδο Elastic Net οι συντελεστές των μεταβλητών «age», «sex», «ca» και «pcells» μηδενίζονται. Αντιθέτως οι συντελεστές των μεταβλητών «bun», «hb» και «protein» υφίσταται συρρίκνωση.

5.5 Εφαρμογή των δέντρων επιβίωσης

Σκοπός της συγκεκριμένης παραγράφου είναι η κατασκευή του δέντρου επιβίωσης για τα δεδομένα μας. Από το δέντρο που θα λάβουμε θα μπορούμε να εξάγουμε συμπεράσματα για το ποιες είναι οι πιο στατιστικά σημαντικές μεταβλητές. Πιο συγκεκριμένα η μεταβλητή που θα χρησιμοποιείται στον πρώτο κόμβο θα είναι και η ανεξάρτητη μεταβλητή που επηρεάζει περισσότερο την έκβαση της υγείας των ασθενών σε σχέση με τις υπόλοιπες. Προκειμένου να λάβουμε το δέντρο αυτό εκτελέσαμε στην R τις παρακάτω εντολές.

```
Install.packages('party')
library(party)

surv.tree<-ctree(Surv(time, status)~ age+sex+bun+ca+hb+pcells+protein)
plot(surv.tree, main="Tree for Time (in months)")
```

Και από το Διάγραμμα 5.22 που παρατίθεται παρακάτω παρατηρούμε ότι λαμβάνουμε μόνο ένα τελικό φύλλο στο οποίο και συμπεριλαμβάνεται όλο το δείγμα.

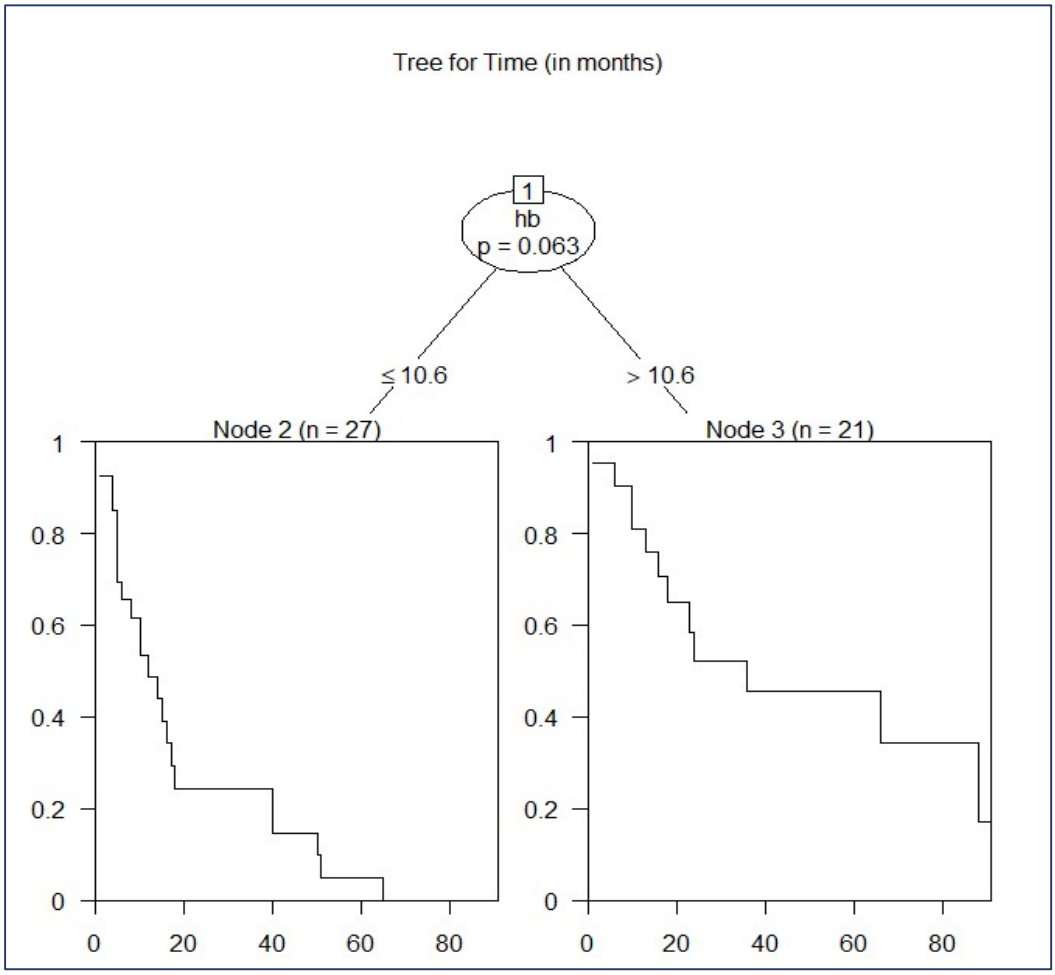


Διάγραμμα 5.22: Δέντρο επιβίωσης

Επειδή από το Διάγραμμα 5.22 δεν μπορούμε να λάβουμε τις πληροφορίες για το ποιες μεταβλητές συμβάλουν θα προχωρήσουμε σε μία περαιτέρω ανάλυση. Για το λόγο αυτό προσθέτουμε ένα alpha (1- 0.8=α=0.2) ώστε να επιτρέπει και στατιστικά μη σημαντικές διαφοροποιήσεις (έστω οριακές) εντός μιας μεταβλητής.

```
surv.tree<-ctree(Surv(time,status)~age+sex+bun+ca+hb+pcells+protein,
control=ctree_control(mincriterion = 0.8)))

surv.tree<-ctree(Surv(time,status)~bun+hb+protein,
control=ctree_control(mincriterion = 0.8)))
```

Όποια από τις δύο παραπάνω εντολές και εάν τρέξουμε λαμβάνουμε τα ίδια αποτελέσματα τα οποία είναι τα ακόλουθα.

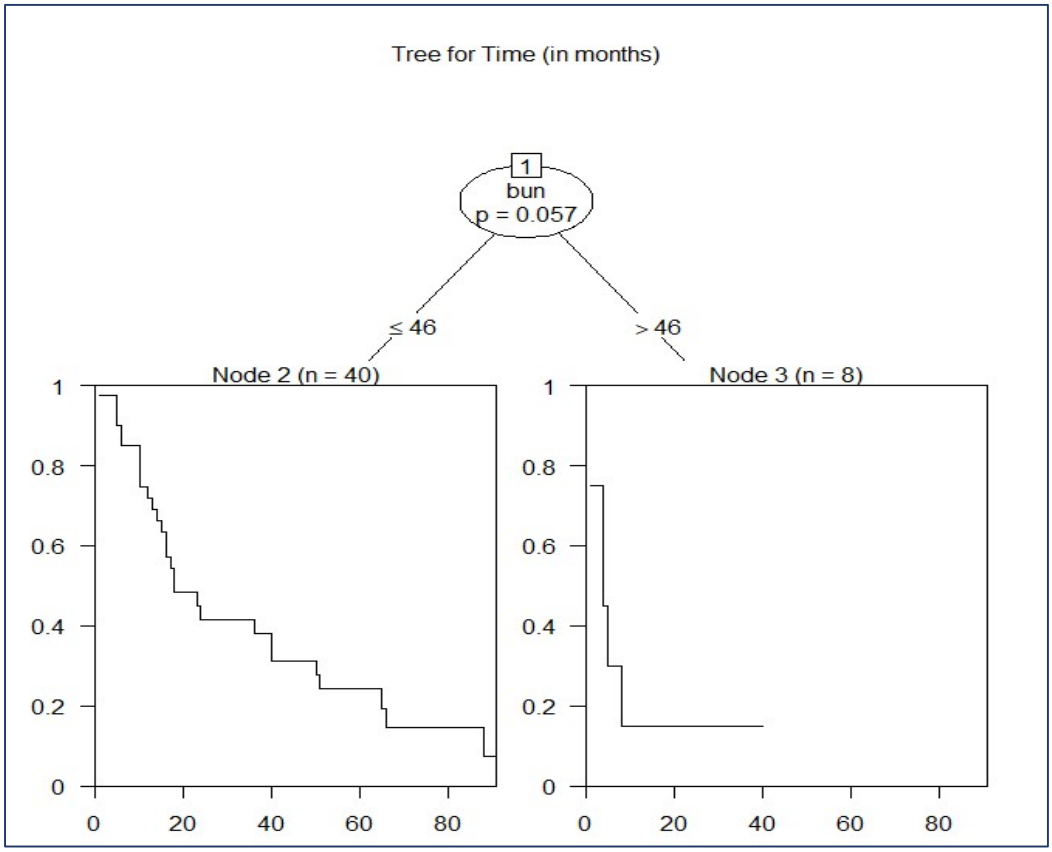


Διάγραμμα 5.23: Δέντρο επιβίωσης με α= 0.2

Από το Διάγραμμα 5.23 λαμβάνουμε το συμπέρασμα ότι στον κόμβο 2 κατατάσσονται τα άτομα με τιμή αιμοσφαιρίνης μικρότερη ή ίση με 10.6, ενώ στον κόμβο 3 τα άτομα που έχουν αιμοσφαιρίνη μεγαλύτερη από 10.6. Παρατηρούμε ότι στα άτομα που έχουν τιμή αιμοσφαιρίνης μικρότερη ή ίση με 10.6 το γεγονός συμβαίνει πιο γρήγορα από ότι σε αυτά που έχουν αιμοσφαιρίνη μεγαλύτερη από 10.6. Τελικά η μεταβλητή «hb» συμβάλει στο μοντέλο και άρα επηρεάζει την διάρκεια επιβίωσης των ασθενών που πάσχουν από πολλαπλό μυέλωμα. Έτσι όσο μεγαλύτερη είναι η τιμή της αιμοσφαιρίνης στο αίμα των ασθενών που πάσχουν από πολλαπλό μυέλωμα τόσο πιο πιθανό είναι ο χρόνος επιβίωσης τους να είναι μεγαλύτερος από αυτόν των ατόμων που έχουν χαμηλή αιμοσφαιρίνη.

Εκτελούμε μία τελευταία εντολή για την εξαγωγή συμπερασμάτων.

```
surv.tree<-ctree(Surv(time, status)~ bun , control=ctree_control(mincriterion = 0.8)))
```



Διάγραμμα 5.24: Δέντρο επιβίωσης με α = 0.2

Από το Διάγραμμα 5.24 παρατηρούμε ότι στον κόμβο 2 κατανέμονται τα άτομα που έχουν επίπεδα αζώτου ουρίας μικρότερα ή ίσα με 46 και στον κόμβο 3 εκείνα που έχουν άζωτο μεγαλύτερο από 46. Στην δεύτερη ομάδα παρατηρούμε ότι το γεγονός (δηλαδή ο θάνατος των ασθενών) επέρχεται πιο γρήγορα από ότι στη δεύτερη ομάδα. Συνεπώς υψηλές τιμές στο άζωτο ουρίας στον ορό του αίματος έχουν σαν αποτέλεσμα μικρό χρόνο επιβίωσης. Συμπερασματικά η μεταβλητή που αφορά τα επίπεδα αζώτου ουρίας (bun) συμβάλει στο μοντέλο και άρα επηρεάζει τη διάρκεια ζωής των ασθενών που πάσχουν από πολλαπλό μυέλωμα και μάλιστα αρνητικά.

5.6 Συμπεράσματα

Κατά τη διάρκεια της ανάλυσής μας εφαρμόσαμε τέσσερις διαφορετικές τεχνικές. Η πρώτη τεχνική ήταν η σύγκριση των εκτιμήσεων Kaplan-Meier ανάμεσα σε δύο διαφορετικές ομάδες ασθενών για κάθε ανεξάρτητη μεταβλητή ξεχωριστά, η δεύτερη αξιοποίησε το ημι-παραμετρικό μοντέλο του Cox, η τρίτη χρησιμοποίησε τις μεθόδους ποινών και πιο συγκεκριμένα τη μέθοδο κορυφογραμμής, τη μέθοδο Lasso και τη μέθοδο Elastic-Net. Τέλος η τέταρτη τεχνική αφορούσε τα δέντρα επιβίωσης.

Με τις εκτιμήσεις της συνάρτησης επιβίωσης Kaplan – Meier λάβαμε ως στατιστικά σημαντικές μεταβλητές τις ακόλουθες: το άζωτο ουρίας αίματος «bun», τα επίπεδα της αιμοσφαιρίνης «hb», το ποσοστό των καρκινικών κυττάρων «pcells» και τέλος την ύπαρξη της μονοκλωνικής πρωτεΐνης «protein». Επομένως σύμφωνα με τις εκτιμήσεις Kaplan – Meier οι παραπάνω τέσσερις ανεξάρτητες μεταβλητές επηρεάζουν τον χρόνο επιβίωσης των ασθενών που πάσχουν από πολλαπλό μυέλωμα.

Η δεύτερη τεχνική αφορούσε το ημι-παραμετρικό μοντέλο του Cox. Χρησιμοποιώντας τη μέθοδο της διαδοχικής αφαίρεσης καταλήξαμε στο συμπέρασμα ότι το βέλτιστο μοντέλο του Cox ήταν αυτό που περιείχε τις μεταβλητές «bun», «hb» και «protein». Τα αποτελέσματα της πρώτης και της δεύτερης τεχνικής συμπίπτουν με τη διαφορά ότι στο βέλτιστο μοντέλο του Cox η μεταβλητή «pcells» δεν θεωρείται στατιστικά σημαντική όπως προηγουμένως.

Στη συνέχεια για να άρουμε τη δυσκολία της πολυσυγγραμμικότητας εφαρμόσαμε μεθόδους που επιβάλλουν ποινή. Η πρώτη ήταν η μέθοδος Lasso η οποία επέδειξε ότι οι μεταβλητές: «bun», «hb» και «protein» συμβάλουν στο μοντέλο. Σε ακριβώς ίδια συμπεράσματα καταλήξαμε και όταν επικαλεστήκαμε τις μεθόδους Ridge και Elastic – Net.

Τέλος τα δέντρα επιβίωσης υπέδειξαν ως στατιστικά σημαντικές μεταβλητές την τιμή της αιμοσφαιρίνης «hb» και τα επίπεδα του αζώτου ουρίας αίματος «bun». Οι μεταβλητές αυτές ήταν ήδη υποψήφιες ως οι πιο στατιστικά σημαντικές και από την εφαρμογή των προηγούμενων μεθόδων.

Συμπερασματικά, ο χρόνος επιβίωσης των ασθενών που πάσχουν από πολλαπλό μυέλωμα επηρεάζεται σε μεγάλο βαθμό από την τιμή του αζώτου ουρίας αίματος και την τιμή της αιμοσφαιρίνης. Όσο ψηλότερη είναι η τιμή του αζώτου ουρίας τόσο μικρότερος και ο χρόνος επιβίωσης των ασθενών. Στον αντίποδα, άτομα με μεγάλη τιμή αιμοσφαιρίνης έχουν μεγαλύτερο χρόνο επιβίωσης. Άλλο χαρακτηριστικό του ασθενούς που συμβάλει στην πορεία της υγείας του είναι η ύπαρξη ή όχι της μονοκλωνικής πρωτεΐνης. Άτομα που έχουν την πρωτείνη έχουν πιο δυσμενή εξέλιξη της υγείας τους σε σχέση με αυτούς που δεν την έχουν. Τέλος το ποσοστό των καρκινικών κυττάρων που εμφανίζεται στις εξετάσεις του ασθενούς επηρεάζει (αρνητικά) σε μικρότερο βαθμό την πορεία της υγείας του και συνεπώς και το χρονικό διάστημα επιβίωσης του. Μεταβλητές όπως το φύλο και η ηλικία του ασθενούς φαίνεται να μην έχουν σημαντικό στατιστικό ρόλο και άρα δεν έχουν αντίκτυπο στη διάρκεια επιβίωσης των ασθενών. Τέλος ούτε η ποσότητα του ασβεστίου επιδρά στο χρόνο ζωής των ασθενών που πάσχουν από πολλαπλό μυέλωμα.

Βιβλιογραφία

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

2. Χ.Καρώνη. & Π. Οικονόμου. (2017), *Στατιστικά Μοντέλα Παλινδρόμησης* (**2η εκδ.**), Εκδόσεις Συμεών, Αθήνα

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

4. Akaike H. (1974), A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**, 716-723

5. Andersen P. K. & Gill R. D. (1982), Cox's regression model for counting processes: A large sample study, *Ann. Statist*., **10**, 1100-1120

6. Bickel P. J., Ritov Y. & Tsybakov A. (2008), Simultaneous analysis of Lasso and Dantzig selector, *Ann. Statist*., **36**, in the press

7. Breiman L., Friedman J., Olson R. & Stone C. (1984), Classification and Regression Trees, Wadsworth, Belmont, California.

8. Breslow N. E (1974), Covariance analysis of censored survival data, *Biometrics*, **30**, 89-99

9. Bou-Hamad I., Larocque D. & Ben-Ameur H. (2011), A review of survival trees *Statistics Surveys,* **5** , 44—71,

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

11. Ciampi A. , Bush R. S., Gospodarowicz M. & Till J. E. (1981), An approach to classifying prognostic factors related to survival experience for non-Hodgkin's lymphoma patients: Based on a series of 982 patients: 1967–1975. *Cancer*, 47: 621-627

12. Ciampi A., Chang CH., Hogg S. & McKinney S. (1987) *Recursive Partition: A Versatile Method for Exploratory-Data Analysis in Biostatistics*, In: MacNeill I.B., Umphrey G.J., Donner A., Jandhyala V.K. (eds) Biostatistics, The University of Western Ontario Series in Philosophy of Science (A Series of Books in Philosophy of Science, Methodology, Epistemology, Logic, History of Science and Related Fields), vol **38**. Springer, Dordrecht

13. Ciampi A., Thiffault J., Nakache J. P. & Asselain B. (1986), Stratification by stepwise regression, correspondence analysis and recursive partition: A comparison of three methods of analysis for survival data with covariates. *Computational Statistics & Data Analysis*, **4**, 185-204.

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

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

16. Cox D.R. (1975), Partial likelihood, *Biometrica*, **62**, 269-276

17. Davis R. B. & Anderson J. R. (1989), Exponential survival trees, *Statistics in Medicine*, **8**: 947-961.

18. Fu W. J. (1998), Penalized Regression: The Bridge Versus the LASSO, *Journal of Computational and Graphical Statistics*, **7**, pp. 397–416

19. Goeman J. (2010), L1 Penalized estimation in the Cox proportional hazards model, *Biometrical Journal*, **52**, 70-84

20. Gordon L. & Olshen R. A. (1985). Tree-structured survival analysis. *Cancer treatment reports 69*, **10**, 1065–1069

21. Harell F. E. Jr (2015), *Regression Modeling Strategies with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, Springer Series in Statistics, **2nd Edition**

22. Hoerl A. & Kennard R. (1970), Ridge regression: biased estimation for the non orthogonal problems*, Technometrics*, **12**, 55-67

23. Hosmer D.W., Lemeshow S. & May S. (2008), *Applied Survival Analysis: Regression Modeling of Time to Event Data*, (**2nd Edition**), Wiley, Hoboken, New Jersey

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

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

26. Lee E. (1980), *Statistical Methods for Survival Data Analysis*, Belmont, California: Lifetime Learning Publications.

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

28. Noah S., Friedman J., Hastie T. & Tibshirani R. (2011), Regularization Ppaths for Cox's proportional hazards model via coordinate descent, *Journal of Statistical Software*, Volume **39**, Issue 5

29. Raftery A., Madigan D. & Volinsky C. T. (1995), Accounting for model uncertainty in survival analysis improves predictive performance, *Bayesian Statistics*, **5**, 323–349.

30. Safavian S. R. & Landgrebe D. (1991), A survey of decision tree classifier methodology, *IEEE transactions on systems, man, and cybernetics 21*, **3**, 660–674

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

32. Schwartz G. (1978), Estimating the dimension of a model, *The Annals of Statistics*, **6**, pp. 461-464.

33. Segal M. (1988), Regression Trees for Censored Data, *Biometrics*, **44(1)**, 35-47.

34. Tibshirani R. (1996), Regression shrinkage and selection via the LASSO, *Journal of the Royal Statistical Society Series B*, **58**, 267-288

35. Tibshirani R. (1997), The LASSO method for variable selection in the Cox model, *Statistics in Medicine*, **16**, 385-395

36. Xu R., Vaida F., & Harrington D.P. (2009), Using profile likelihood for semiparametric model selection with application to proportional hazards mixed models, *Statistica Sinica*, **19**, 819-842

37. Zhang H. H. & Lu W. (2007), Adaptive lasso for Cox's proportional hazards model. *Biometrika*, **94**, 691-703.

38. Zou H. & Hastie T. (2005), Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society Series B*, **67**, 301-320