Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Εφαρμοσμένων Μαθηματικών & Φυσικών Επιστημών
Τομέας Μαθηματικών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μαρίνα Ντούσκα

Ανάλυση Επιβίωσης: το Μοντέλο του Cox, Ποινικοποιημένες Μέθοδοι και Δένδρα Επιβίωσης

Τριμελής Επιτροπή:
Χρυσηίς Καρώνη, Καθηγήτρια (Επιβλέπουσα)
Σχολή Εφαρμοσμένων Μαθηματικών & Φυσικών Επιστημών

Βασίλειος Παπανικολάου, Καθηγητής
Σχολή Εφαρμοσμένων Μαθηματικών & Φυσικών Επιστημών

Καλλιόπη Παυλοπούλου, Ε.ΔΙ.Π
Σχολή Εφαρμοσμένων Μαθηματικών & Φυσικών Επιστημών

Αθήνα, Ιούνιος 2022

Περίληψη

Στόχος της παρούσας διπλωματικής εργασίας είναι να αναλύσει τεχνικές που χρησιμοποιούνται στην ανάλυση δεδομένων διάρκειας ζωής ή αλλιώς στην ανάλυση επιβίωσης.

Στο πρώτο κεφάλαιο αναλύεται η ιδιαιτερότητα των δεδομένων διάρκειας ζωής, η αποκοπή δεδομένων και επιπλέον αναφέρονται οι βασικοί ορισμοί που χρειάζονται στην μη-παραμετρική ανάλυση επιβίωσης. Στο δεύτερο κεφάλαιο ορίζεται το ημιπαραμετρικό μοντέλο αναλογικής διακινδύνευσης του Cox. Το μοντέλο και αναπτύσσεται, παρουσιάζοντας πως εκτιμώνται οι συντελεστές του και επίσης πως μπορούν να πραγματοποιηθούν οι έλεγχοι υποθέσεων του μοντέλου. Στη συνέχεια, αναφέρονται οι προεκτάσεις που μπορούν να γίνουν στο μοντέλο του Cox, οι έλεγχοι για την ισχύ της υπόθεσης αναλογικής διακινδύνευσης και τα βασικά στοιχεία που χρησιμοποιούνται στην εκτίμηση της προβλεπτικής ικανότητας του μοντέλου μέσω της καμπύλης ROC. Στο τρίτο κεφάλαιο παρουσιάζονται οι ποινικοποιημένες μέθοδοι παλινδρόμησης, Ridge και Lasso, για την εκτίμηση των συντελεστών παλινδρόμησης του μοντέλου και η εφαρμογή τους στο μοντέλο του Cox. Στο τέταρτο κεφάλαιο, περιγράφεται πως μπορούν να δημιουργηθούν δένδρα παλινδρόμησης με σκοπό την πρόβλεψη της πορείας της υγείας ασθενών και τον εντοπισμό των σημαντικών παραγόντων που επηρεάζουν την εξέλιξη των ασθενειών τους. Τέλος, στο πέμπτο κεφάλαιο της εργασίας, η μεθοδολογία που προηγήθηκε εφαρμόζεται στην ανάλυση δεδομένων ενός συνόλου από ασθενείς που πάσχουν από καρκίνο του μαστού.

Λέξεις Κλειδιά

ανάλυση επιβίωσης, μοντέλο του Cox, παλινδρόμηση Ridge, δένδρα επιβίωσης, καρκίνος του μαστού

Abstract

The purpose of this study is to analyse techniques that are used in survival analysis.

In the first chapter the particular feature of survival data, censored observations and also basic definitions of non-parametric survival analysis are presented. The second chapter defines Cox's semi-parametric, propotional harazards model. The model is developed, showing how its coefficients are estimated and also how hypothesis testing can be carried out. Next, reference is made to extensions that can be made to the Cox model, tests for the proportional hazards assumption and the basic tools for estimating the predictive ability of the model through the ROC curve. In the third chapter the penalty methods, Ridge and Lasso, for the estimation of regression coefficients and their application to the Cox model are presented. The fourth chapter describes how regression trees can be created in order to predict the course of patients' health and to identify the important factors that influence the progression of their disease. Finally, in the fifth chapter of the study, the preceding methodology is applied to the analysis of data on a group of patients suffering from breast cancer.

Keywords

survival analysis; Cox model; Ridge regression; Lasso; survival trees; breast cancer

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά την καθηγήτρια του Εθνικού Μετσόβιου Πολυτεχνείου κ. Χρυσηίδα Καρώνη, για την πολύτιμη καθοδήγηση και την καθοριστική βοήθεια της σε όλη τη διάρκεια εκπόνησης της παρούσας διπλωματικής εργασίας.

Επιπλέον θα ήθελα να ευχαριστήσω τους φίλους μου, για την ενθάρρυνση και την πίστη τους σε μένα και τους συναδέλφους, που έγιναν πολύτιμοι φίλοι καθώς μεγαλώσαμε μαζί, ακαδημαϊκά και διαπροσωπικά.

Φυσικά το μεγαλύτερο ευχαριστώ το οφείλω στους γονείς μου, για την αμέριστη υποστήριξη και την απεριόριστη κι άνευ όρων αγάπη τους όλα αυτά τα χρόνια και στα αδέρφια μου που βρισκόντουσαν πάντα δίπλα μου να με βοηθούν και να μου δείχνουν το δρόμο.

Μαρίνα Ντούσκα

Περιεχόμενα

Κατάλογος Σχημάτων

Κεφάλαιο 1

Ανάλυση Επιβίωσης

1.1 Εισαγωγή

Μια περιοχή της Στατιστικής η οποία είναι άξια προσοχής είναι η ανάλυση δεδομένων διάρκειας ζωής. Ο όρος ανάλυση διάρκειας ζωής αναφέρεται στην μελέτη του χρόνου μέχρι να συμβεί ένα επιθυμητό ή ανεπιθύμητο γεγονός. Τα δεδομένα διάρκειας ζωής προκύπτουν σε διάφορους τομείς της επιστήμης. Ιδιαίτερο ενδιαφέρον παρουσιάζουν οι εφαρμογές στην βιοϊατρική επιστήμη, με το όνομα Ανάλυση Επιβίωσης (Survival Analysis), όπου μελετάται κυρίως η ανάλυση επιβίωσης ασθενών, εξετάζοντας ποιοι παράγοντες και σε ποιο βαθμό επηρεάζουν την πορεία της υγείας τους, συγκρίνοντας τις διάφορες θεραπείες τις οποίες μπορεί να ακολουθούν αλλά και συγκρίνοντας σύνολα με διαφορετικά χαρακτηριστικά ως προς τη διάρκεια ζωής. Επιπλέον πολλές εφαρμογές συναντώνται σε τεχνολογικές επιστήμες όπου μελετώνται η λειτουργία των μηχανημάτων, η αξιοπιστία συστημάτων και η αντοχή των υλικών, ο τομέας αυτός ονομάζεται Ανάλυση Αξιοπιστίας (Reliability Analysis).

Η παρούσα διπλωματική εργασία θα περιοριστεί στην Ανάλυση Επιβίωσης. Όπως αναφέρθηκε και παραπάνω σκοπός είναι να μελετηθεί ο χρόνος μέχρι να συμβεί ένα συμβάν ενδιαφέροντος, η σύγκριση του χρόνου αυτού μεταξύ δύο διαφορετικών ομάδων δεδομένων και γενικότερα η επιρροή των παραγόντων (συμμεταβλητών) που εμπεριέχονται στο μοντέλο σε αυτόν τον χρόνο. Συμβάν ενδιαφέροντος μπορεί να είναι για παράδειγμα είτε ο θάνατος κάποιου ασθενή, είτε η ανακούφιση από τον πόνο λόγω κάποιας θεραπείας, είτε ο χρόνος υποτροπής της ασθένειας κ.ά.

Βασικός στόχος για την ανάλυση της διάρκειας ζωής, στην ανάλυση επιβίωσης, είναι η εκτίμηση της *συνάρτησης επιβίωσης* $S(t)$. Έστω λοιπόν ότι T ορίζεται η τυχαία μεταβλητή που εκφράζει τη διάρκεια ζωής μιας μονάδας και μπορεί να λάβει μόνο θετικές τιμές. Η συνάρτηση επιβίωσης $S(t)$ εκφράζει την πιθανότητα η διάρκεια ζωής T, ενός τυχαία επιλεγμένου μέλους από έναν πληθυσμό μονάδων, να υπερβαίνει το χρόνο t. Ορίζουμε λοιπόν την $S(t)$:

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(u)du,$$

όπου $F(t) = P(T < t), t > 0$, είναι η συνάρτηση κατανομής της τυχαίας μεταβλητής T.

Η συνάρτηση πυκνότητας πιθανότητας της τ.μ T μπορεί να βρεθεί ως εξής:

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

Επιπλέον, βασικό εργαλείο στην ανάλυση επιβίωσης αποτελεί και η *συνάρτηση διακινδύνευσης*, $h(t)$ και εκφράζει την τάση προς διακοπή που έχει μια μονάδα μέσα στο χρονικό διάστημα $(t, t + \delta t]$, υπό την προϋπόθεση της επιβίωσης της μέχρι τη χρονική στιγμή t ή απλούστερα εκφράζει τον στιγμαίο ρυθμό διακοπής. Ορίζεται λοιπόν ως εξής:

$$h(t) = \lim_{\delta t \to\ 0} \frac{P[t < T \leq (t + \delta t)]}{P[T > t]} = \lim_{\delta t \to\ 0} \frac{(S(t) - S(t + \delta t))}{\delta t} = \frac{f(t)}{S(t)}$$

Σημειώνεται πως η ποσότητα $h(t)\delta t$ εκφράζει την πιθανότητα της επικείμενης διακοπής μιας μονάδας, δεδομένου ότι επιβίωσε μέχρι τη χρονική στιγμή t.

Μια ακόμη συνάρτηση η οποία είναι εξαιρετικά χρήσιμη στην ανάλυση επιβίωσης είναι και η *σωρευτική συνάρτηση διακινδύνευσης* (σ.σ.δ) η οποία είναι ιδιαίτερα βοηθητική στην κατάλληλη επιλογή στατιστικού μοντέλου στην ανάλυση ενός συνόλου δεδομένων και ορίζεται ως :

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

Όπως φαίνεται από τους παραπάνω ορισμούς οι συναρτήσεις $h(t), f(t), S(t), F(t), H(t)$ είναι ισοδύναμες καθώς γνωρίζοντας μια από αυτές μπορεί να προσδιοριστούν και οι υπόλοιπες.

1.2 Η ιδιαιτερότητα των δεδομένων διάρκειας ζωής

1.2.1 Η κατανομή τους

Ένα χαρακτηριστικό το οποίο ξεχωρίζει τα δεδομένα διάρκειας ζωής από τα υπόλοιπα είναι πως δεν είναι συμμετρικά κατανεμημένα (Collett, 2015). Στις περισσότερες περιπτώσεις η κατασκευή ενός ιστογράμματος (histogram) από δεδομένα διάρκειας ζωής, θα εμφάνιζε μια θετική λοξότητα (skewness), έτσι θα εμφανιζόταν μια μεγαλύτερη

'ουρά' στη δεξιά μεριά του ιστογράμματος, όπου θα εμπεριέχονταν και οι περισσότερες παρατηρήσεις. Κάτι τέτοιο, θα έκανε την υπόθεση της κανονικής κατανομής των δεδομένων να μοιάζει απαγορευτική. Μια λύση θα ήταν ο μετασχηματισμός των δεδομένων, χρησιμοποιώντας για παράδειγμα λογαρίθμους. Όμως μια καλύτερη προσέγγιση θα ήταν η υπόθεση μιας διαφορετικής κατανομής για τα δεδομένα που υπάρχουν.

1.2.2 Αποκομμένα Δεδομένα (Censored Data)

Μια επιπλέον ιδιαιτερότητα που εμφανίζουν τα δεδομένα διάρκειας ζωής, συγκριτικά με τα υπόλοιπα δεδομένα που συναντώνται σε στατιστικές αναλύσεις, είναι η αποκοπή δεδομένων (censoring). Κατά τη διάρκεια ενός πειράματος στο οποίο καταγράφεται η διάρκεια ζωής των ασθενών που συμμετέχουν σε αυτό, είναι πολύ σύνηθες κάποιοι από αυτούς να συνεχίσουν να ζουν μετά τον τερματισμό του πειράματος. Ως αποτέλεσμα η τιμή μιας παρατήρησης ή μιας μεταβλητής είναι μερικώς γνωστή. Είναι ιδιαίτερα σημαντικό όμως στην ανάλυση των δεδομένων, η πληροφορία αυτή να ληφθεί υπόψιν.

Η αποκοπή ως επί το πλείστον συμβαίνει όταν ένα άτομο αποσύρεται από την μελέτη πριν αυτή ολοκληρωθεί ή η μελέτη διακόπτεται σε ένα συγκεκριμένο χρονικό διάστημα και έτσι ο χρόνος επιβίωσης δεν είναι γνωστός, γνωρίζουμε όμως ότι έχει ξεπεράσει τη διάρκεια του πειράματος. Η περίπτωση αυτή ονομάζεται *από δεξιά αποκοπή*. Πιο σπάνια συναντάται *η από αριστερά αποκοπή* κατά την οποία ο χρόνος επιβίωσης είναι μικρότερος από κάποιο χρονικό διάστημα και είναι αρκετά πιο σπάνια από την δεξιά. Επίσης υπάρχει κι *η εντός διαστημάτων αποκοπή* κατά την οποία το γεγονός ενδιαφέροντος έχει συμβεί εντός κάποιου χρονικού διαστήματος. Η τελευταία προκύπτει είτε λόγω ανακριβών μετρήσεων είτε λόγω περιοδικής επίβλεψης του πειράματος (Καρώνη, 2009).

Η αποκοπή δεδομένων οφείλει να είναι τυχαία, δηλαδή να είναι ανεξάρτητη από την διάρκεια ζωής του ασθενή μετά το πέρας του πειράματος και ονομάζεται 'μη-πληροφοριακή' (non-informative censoring). Εν αντιθέσει, αν ένας ασθενής αποσύρεται από το πείραμα σε περίπτωση επιδείνωσης της κατάστασης του, η αποκοπή δεν θεωρείται τυχαία και ονομάζεται 'πληροφοριακή αποκοπή' (informative censoring). Είναι πολύ σημαντικό να διασφαλίζεται η non-informative αποκοπή, ώστε η ανάλυση που χρησιμοποιείται να είναι έγκυρη.

Παρακάτω αναφέρονται οι δύο βασικοί μηχανισμοί αποκοπής παρατηρήσεων, ανεξάρτητοι της διάρκειας ζωής της μονάδας: Τύπος Ι και Τύπος ΙΙ.

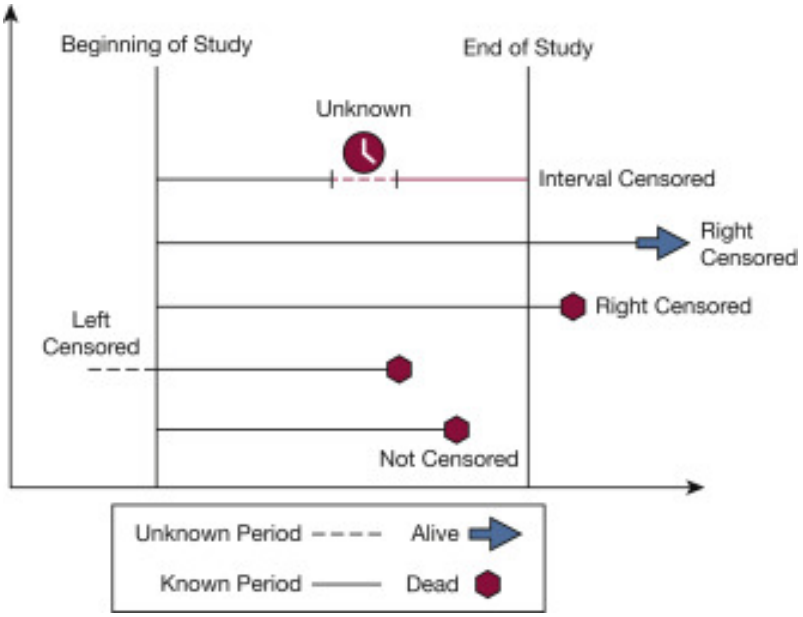
Αποκοπή Τύπου Ι

Έχοντας προκαθορίσει ένα χρονικό διάστημα c παρακολουθούμε τους ασθενείς που συμμετέχουν στο πείραμα.Γνωρίζουμε την ακριβή διάρκεια ζωής του ασθενούς αν $T_i < c$, διαφορετικά αν $T_i > c$ τότε γνωρίζουμε πως η διάρκεια ζωής υπερέβη το c.

Σε γενικότερη περίπτωση θα μπορούσε κάθε ασθενής να έχει δικό της χρόνο παρακολούθησης c_i. Ο αριθμός των ασθενών για τους οποίους καταγράφεται το γεγονός ενδιαφέροντος είναι τυχαίος και ο προκαθορισμένος χρόνος παρακολούθησης ονομάζεται χρόνος αποκοπής.

Αποκοπή Τύπου II

Έχοντας προκαθορίσει τον αριθμό κ των ασθενών που μας ενδιαφέρει, η διάρκεια παρακολούθησης του πειράματος είναι τυχαία και τελειώνει όταν καταγραφεί το γεγονός ενδιαφέροντος και για τις κ μονάδες.



Σχήμα 1.1: Απεικόνιση αποκομμένων παρατηρήσεων

1.3 Κολοβά δεδομένα (Truncated Data)

Πρόκειται για δεδομένα τα οποία προκύπτουν όταν το γεγονός ενδιαφέροντος για μια μονάδα, συμβαίνει εκτός ενός ορισμένου χρονικού διαστήματος στο οποίο εξελίσσεται η μελέτη. Βασική διαφορά κολοβών και αποκομμένων παρατηρήσεων είναι πως για τις τελευταίες υπάρχουν έστω και ελλιπείς πληροφορίες οι οποίες μπορούν να χρησιμοποιηθούν ενώ για τα κολοβά υπάρχουν παρατηρήσεις για τις οποίες δεν υπάρχει καμία πληροφορία.

1.4 Μη-παραμετρική άναλυση διάρκειας ζωής

Οι μέθοδοι που χρησιμοποιούνται παρακάτω χαρακτηρίζονται ως μη-παραμετρικοί καθώς δεν γίνεται γίνεται κάποια υπόθεση για την κατανομή που ακολουθεί η συνάρτηση επιβίωσης. Η συνάρτηση επιβίωσης καθώς και η γραφική της παράσταση χρησιμοποιούνται ιδιαίτερα όταν γίνεται η προσαρμογή του μοντέλου έτσι ώστε να δοθούν κάποιες ενδείξεις για τη συμπεριφορά των δεδομένων. Αναφέρονται χαρακτηριστικά τα εργαλεία της μη-παραμετρικής ανάλυσης που θα χρησιμοποιηθούν στην παρούσα διπλωματική για την ανάπτυξη του στατιστικού μοντέλου ανάλυσης επιβίωσης που θα χρησιμοποιηθεί, το μοντέλο του Cox.

1.4.1 Εκτίμηση της συνάρτησης επιβίωσης

Ένας απλοϊκός τρόπος για την εκτίμησης της συνάρτησης επιβίωσης περιγράφεται παρακάτω (Collett, 2015). Έστω ότι υπάρχει ένα σύνολο δεδομένων χωρίς αποκομμένα δεδομένα. Η εμπειρική εκτιμήτρια της συνάρτησης επιβίωσης βασισμένη στον ορισμό ότι πρόκειται για την πιθανότητα κάποιος να επιβιώσει για χρόνο μεγαλύτερο του t, δίνεται παρακάτω:

$$\hat{S}(t) = \frac{\text{Αριθμός ατόμων με χρόνο επιβίωσης} \geq t}{\text{Αριθμός ατομών στο σύνολο δεδομένων}}$$

Εύκολα μπορεί κάποιος να παρατηρήσει πως η παραπάνω εκτιμήτρια είναι ίση με τη μονάδα για τιμές του t πριν τον πρώτο θάνατο και ίση με μηδέν μετά τον τελευταίο θάνατο. Επιπλέον είναι μια σταθερή συνάρτηση μεταξύ δύο γειτονικών θανάτων και σε μια γραφική παράσταση της συνάρτησης αυτής με τον χρόνο θα είχε κλιμακωτή μορφή.

Σε περιπτώσεις όμως που υπάρχουν αποκομμένα δεδομένα η παραπάνω προσέγγιση δεν θα ήταν χρήσιμη. Παρακάτω εξετάζονται μεθοδολογίες για την εκτίμηση σε σύνολα με αποκομμένα δεδομένα.

1.4.2 Η εκτιμήτρια Kaplan-Meier

Εξ' αιτίας της αποκοπής δεδομένων που συναντάται στα δεδομένα διάρκειας ζωής, για την εκτίμηση της συνάρτησης επιβίωσης, χρησιμοποιείται ευρέως η εκτιμήτρια Kaplan-Meier (Kaplan & Meier, 1958).

Έχοντας ένα τυχαίο δείγμα n μονάδων, σε κάποιες από τις οποίες συμβαίνει το γεγονός ενδιαφέροντος σε διακεκριμένες χρονικές στιγμές $t_{(1)} < t_{(2)} < ... < t_{(k)}$, $k \leq n$
Ορίζονται ως:

- n_j : αριθμός μονάδων που λειτουργούν και άρα βρίσκονται σε ρίσκο τη χρονική στιγμή t_j. Δεν αφορά μονάδες που έχουν καταστραφεί, ούτε αποκομμένες τιμές πριν την στιγμή t_j.

- d_j : αριθμός μονάδων που τους συμβαίνει το γεγονός ενδιαφέροντος στο t_j, στις περισσότερες εφαρμογές $d_j = 1$.

Και η εκτιμήτρια Kaplan - Meier διαμορφώνεται ως:

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

Η παραπάνω είναι μια δειγματική εκτίμηση έτσι είναι καλό να υπάρχει και μια εκτίμηση της ακρίβειας της. Μπορεί να οριστεί το τυπικό σφάλμα για τη δειγματική εκτιμήτρια $\hat{S}$ γνωστό και ως τύπος του Greenwood:

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

1.4.3 Η εκτιμήτρια Nelson-Aalen

Για την εκτίμηση της σωρευτικής συνάρτησης διακινδύνευσης H(t) χρησιμοποιείται η εκτιμήτρια Nelson-Aalen ((Nelson, 1972), (Aalen, 1978).
Ορίζεται ως:

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

Η εκτιμήτρια διασποράς για την Nelson-Aalen ορίζεται ως:

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

Είναι δυνατόν να δειχθεί πως η παραπάνω εκτιμήτρια οδηγεί και σε μία ακόμη εκτίμηση της συνάρτησης επιβίωσης η οποία ορίζεται ως εξής:

$$\tilde{S}(t) = \prod_{j=1}^{k} exp(-\frac{d_j}{n_j})$$

Σε μικρότερα δείγματα πολλές φορές η εκτιμήτρια Nelson-Aalen της συνάρτησης επιβίωσης αποδίδει καλύτερα από την Kaplan - Meier, όμως γενικότερα προτιμάται η τελευταία.

1.4.4 Log-Rank έλεγχος υποθέσεων

Είναι πολύ σύνηθες σε διάφορες αναλύσεις να εμφανίζεται η ανάγκη να εντοπιστούν οι διαφορές ανάμεσα σε δύο ομάδες εκ των δεδομένων.Ένας πολύ διαδεδομένος τρόπος για να γίνει αυτή η μελέτη είναι να εκτελεσθεί ένας Log-Rank έλεγχος υποθέσεων. Πρόκειται για έναν έλεγχο υποθέσεων όπου ελέγχουμε :

H_0= Δεν υπάρχει διαφοροποίηση στη συνάρτηση επιβίωσης μεταξύ των μονάδων των δύο ομάδων.

H_1= Υπάρχει διαφοροποίηση στη συνάρτηση επιβίωσης μεταξύ των μονάδων των δύο ομάδων.

Έστω και πάλι $t_{(1)} < t_{(2)} < ... < t_{(k)}$ διακεκριμένες στιγμές διακοπής, κατά την διάρκεια των οποίων διακόπτεται η λειτουργία μονάδων που προέρχονται από δύο διαφορετικές ομάδες. Θεωρείται τότε ότι μέχρι τη χρονική στιγμή t_j, υπάρχουν n_{ij} μονάδες σε κίνδυνο και d_{ij} από αυτές σταματούν να λειτουργούν τη στιγμή t_j.
Ορίζεται τότε η στατιστική συνάρτηση ελέγχου:

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

Αποδεικνύεται πως κάτω από την μηδενική υπόθεση $H_0 : S_1(t) = S_2(t)$, το στατιστικό ελέγχου $\frac{u}{\sqrt{v}}$ ακολουθεί ασυμπτωτικά την τυποποιημένη κανονική κατανομή $N(0, 1)$. Άρα η log-rank $\frac{u^2}{v}$ ασυμπτωτικά ακολουθεί την χ_1^2.

1.4.5 Wilcoxon έλεγχος υποθέσεων

Το Wilcoxon test, που είναι γνωστό και ως test Breslow κάνει τον ίδιο έλεγχο υποθέσων με τον Log-Rank. Το στατιστικό ελέγχου βασίζεται στην ποσότητα:

$$U_W = \sum_{j=1}^{k} n_j (d_{1j} - e_{1j}), \text{ όπου } e_{1j} = \frac{n_{1j} d_j}{n_j}$$

Με την διασπορά της παραπάνω να ορίζεται ως :

$$V_W = \sum_{j=1}^{k} n_j^2 v_{1j}, \text{ όπου } v_{1j} \text{ όπως ορίστηκε στον έλεγχο Log-Rank}$$

Ο όρος e_{1j}, είναι σταθμισμένος από n_j, το σύνολο δηλαδή των μονάδων που βρίσκονται σε κίνδυνο την στιγμή t_j. Με τον τρόπο αυτό δίνεται λιγότερο βάρος στις διαφορές

μεταξύ των d_{1j} και e_{1j} τις στιγμές όπου ο συνολικός αριθμός των μονάδων που έχουν τους μεγαλύτερους χρόνους επιβίωσης, είναι μικρός.

Το στατιστικό ελέγχου του Wilcoxon ελέγχου υποθέσεων ορίζεται ως:

$$W_W = \frac{U_W^2}{V_W}$$

Το οποίο ακολουθεί και αυτό την χ_1^2 κατανομή.

Σε περιπτώσεις όπου η εναλλακτική της μηδενικής υπόθεσης, ότι δεν υπάρχει διαφορά μεταξύ των δύο ομάδων δεδομένων, είναι πως ο κίνδυνος για μια μονάδα στη μία ομάδα είναι ανάλογος με τον κίνδυνο για μια άλλη μονάδα της άλλης ομάδας, προτιμάται ο Log-Rank έλεγχος υποθέσεων. Η παραπάνω υπόθεση, γνωστή και ως υπόθεση αναλογικής διακινδύνευσης θα μελετηθεί παρακάτω. Σε ελέγχους όπου η εναλλακτική υπόθεση δεν έχει σχέση με την παραπάνω ο έλεγχος Wilcoxon είναι προτιμότερος.

1.5 Το μοντέλο αναλογικής διακινδύνευσης

Στα μοντέλα παλινδρόμησης οι συμμεταβλητές, οι παράγοντες δηλαδή που επηρεάζουν την μεταβλητή ενδιαφέροντος Y, εισάγονται στο μοντέλο βάσει της επίδρασης που έχουν στην παράμετρο μ της κατανομής της Y. Έτσι η ιδέα αυτή μπορεί να επεκταθεί και σε μοντέλα διάρκειας ζωής.

Σε μια εφαρμογή βιοϊατρικών δεδομένων η επιβίωση και γενικότερα η πορεία της υγείας ενός ασθενή επηρεάζεται από πολλούς διαφορετικούς παράγοντες. Αυτός είναι κι ένας λόγος που χειριζόμαστε τελείως διαφορετικά τα δεδομένα αυτά. Σκοπός είναι να εκτιμηθεί ο κίνδυνος θανάτου ή ο κίνδυνος υποτροπής και για τον λόγο αυτό μοντελοποιείται η συνάρτηση διακινδύνευσης, για να βρεθούν οι παράγοντες οι οποίοι την επηρεάζουν. Στα μοντέλα αναλογικής διακινδύνευσης τα οποία θα μελετηθούν στην παρούσα εργασία έχει γίνει η παραδοχή ότι οι συναρτήσεις διακινδύνευσης δύο διαφορετικών μονάδων από δύο διαφορετικά σύνολα δεδομένων, είναι ανάλογες την ίδια χρονική στιγμή.

Η πιο γνώστη μορφή του μοντέλου αυτού, στην οποία θα επικεντρωθεί και η ανάλυση αυτής της εργασίας, είναι το μοντέλο αναλογικής διακινδύνευσης του Cox, όπως παρουσιάστηκε από τον ίδιο το 1972, στο οποίο η συνάρτηση διακινδύνευσης παραμένει ακαθόριστη. Για τον λόγο αυτό αυτό το μοντέλο χαρακτηρίζεται ως ημι-παραμετρικό. Συγκεκριμένα οι συμμετεβλητές $\mathbf{x}$ δρουν στην συνάρτηση διακινδύνευσης μέσω της σχέσης :

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

Κεφάλαιο 2

Το μοντέλο του Cox

Η ενασχόληση με δεδομένα διάρκειας ζωής στις βιοϊατρικές επιστήμες παρουσιάζει σημαντικές διαφορές από τις τεχνολογικές εφαρμογές δεδομένων. Όπως μπορεί κανείς εύκολα να αναλογιστεί κάθε πληθυσμός που μελετάται σε ένα βιοϊατρικό πείραμα είναι διαφορετικός. Κάθε ασθενής είναι μοναδικός και έτσι δεν μπορεί να αναπτυχθεί a priori κάποια θεωρία για να περιγράψει την εξέλιξη της υγείας του. Επιπρόσθετα η συλλογή δεδομένων από ασθενείς, σπάνια γίνεται υπό αυστηρές συνθήκες διεξαγωγής του πειράματος. Επομένως δεν μπορούν να ελεγχθούν όλοι οι εξωτερικοί παράγοντες που μπορεί να επηρεάζουν το αποτέλεσμα της μελέτης και οι σημαντικοί παράγοντες που επηρεάζουν την πορεία της υγείας κάποιου ασθενή δεν είναι πάντοτε γνωστοί. Όλα τα παραπάνω λοιπόν, δημιουργούν την ανάγκη υιοθέτησης ενός μοντέλου, το οποίο δεν θα είναι παραμετρικό, δηλαδή ένα μοντέλο για το οποίο δεν θα χρειαστεί να γίνει κάποια εκτίμηση για τις παραμέτρους της κατανομής που ακολουθεί η μεταβλητή που περιγράφει την διάρκεια ζωής.

2.1 Ορισμός Μοντέλου

Στην δυσκολία υιοθέτησης ενός βασικού μοντέλου για την περιγραφή βιοϊατρικών δεδομένων έρχεται να δώσει λύση το ημι-παραμετρικό μοντέλο αναλογικής διακινδύνευσης του Cox, το οποίο αποτελεί την πιο διαδεδομένη επιλογή μοντέλου για την ανάλυση επιβίωσης. Χρησιμοποιώντας το μοντέλο αυτό, μπορεί να μελετηθεί η επιβίωση ασθενών σε σχέση με τις διάφορες επεξηγηματικές μεταβλητές του μοντέλου παρά το γεγονός ότι υπάρχουν αποκομμένες παρατηρήσεις.

Το μοντέλο αυτό είναι ιδιαίτερα χρήσιμο καθώς μπορεί να μοντελοποιήσει την συνάρτηση διακινδύνευσης $h(t)$ και βοηθά στη μελέτη της σχέσης μεταξύ της συνάρτησης διαδικινδύνευσης και πολλών επεξηγηματικών μεταβλητών καθώς και στην εύρεση των διαφορών που προκύπτουν μεταξύ διαφορετικών χαρακτηριστικών που έχουν ομάδες των δεδομένων. Επιπλέον είναι ένα σημαντικό εργαλείο για την εκτίμηση γενικότερα

των γεγονότων ενδιαφέροντος που βρίσκονται υπό μελέτη.

Όπως αναφέρθηκε και παραπάνω πρόκειται για ένα μοντέλο παλινδρόμησης στο οποίο οι συμμεταβλητές $\mathbf{x}$ επιδρούν στη συνάρτηση διακινδύνευσης μέσω της :

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

$h_0(t)$: μία βασική συνάρτηση δικινδύνευσης, συγκεκριμένα είναι η συνάρτηση διακινδύνευσης, όταν οι τιμές όλων των συμμεταβλητών είναι ίσες με 0 ($\mathbf{x} = 0$)
$\mathbf{x}$: το διάνυσμα των συμμεταβλητών που επηρεάζουν το χρόνο ζωής
$\boldsymbol{\beta}$: ένα διάνυσμα p συντελεστών του μοντέλου, οι οποίοι εκφράζουν ποσοτικά την επίδραση κάθε συμμεταβλητής $\mathbf{x}$ και επηρεάζουν το γεγονός ενδιαφέροντος που εξετάζεται.

Συνεπώς συνάρτηση διακινδύνευσης $h(t;\mathbf{x})$, εξαρτάται από την $h_0(t)$, που είναι μια συνάρτηση του χρόνου και παραμένει ακαθόριστη και από μια δεύτερη ποσότητα που εξαρτάται από τις συμμεταβλητές μέσω του διανύσματος των παραμέτρων $\boldsymbol{\beta}$.

Βάσει της παραπάνω σχέσης η σωρευτική συνάρτηση διακινδύνευσης H(t) και η συνάρτηση επιβίωσης S(t) διαμορφώνονται κατ΄αντιστοιχία ως εξής:

$$H(t;\mathbf{x}) = H_0(t)e^{\boldsymbol{\beta}'\mathbf{x}}$$

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

Το μοντέλο χαρακτηρίζεται ως ημι-παραμετρικό καθώς καθ΄ όλη τη διάρκεια της ανάλυσης οι βασικές συναρτήσεις h_0 και S_0 δεν καθορίζονται και ως αναλογικής διακινδύνευσης καθώς παρατηρείται η βασική ιδιότητα του μοντέλου ότι ο λόγος $\frac{h(x_1;t)}{h(x_2;t)}$ παραμένει σταθερός και ανεξάρτητος του χρόνου t, όπου $x_1 = (x_{11}, ..., x_{1p})$ και $x_2 = (x_{21}, ..., x_{2p})$ τα διανύσματα των συμμεταβλητών, με p παραμέτρους για δύο ασθενείς). Επίσης σημειώνεται πως μόνο οι επιδράσεις των συμμεταβλητών $\mathbf{x}$ αναλύονται.

Σημειώνεται πως ο όρος $e^{(\beta x_i)}$ είναι πάντα θετικός (Collett, 2015). Έτσι θα μπορούσε να γραφεί ως $exp(\eta_i)$, όπου το η_i είναι ένας γραμμικός συνδυασμός των p επεξηγηματικών συμμεταβλητών $\mathbf{x}_i$. Έτσι:

$$\eta_i = \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}$$

ώστε $\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$. Ο όρος η_i ονομάζεται γραμμική συνιστώσα του μοντέλου και χαρακτηρίζεται επίσης ως 'risk score' ή προγνωστικός δείκτης (prognostic index) για την i-οστη μονάδα. Το μοντέλο τότε θα γίνει:

$$h_i(t) = exp(\beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi})h_0(t)$$

Η παραπάνω μορφή του μοντέλου μπορεί να γραφεί και ως:

$$\log(\frac{h_i(t)}{h_0(t)}) = \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}$$

και έτσι το μοντέλο αναλογικής διακινδύνευσης μπορεί να θεωρηθεί και ως ένα γραμμικό μοντέλο του λογαρίθμου του λόγου διακινδύνευσης.

Αξίζει ιδιαίτερης προσοχής το γεγονός πως στη γραμμική συνιστώσα του μοντέλου δεν υπάρχει σταθερός όρος. Επιπλέον δεν έχει γίνει καμία υπόθεση για τους συντελεστές του μοντέλου β_i. Παρακάτω θα δειχθεί πως μπορούν να εκτιμηθούν χωρίς να γίνουν υποθέσεις για αυτούς.

2.2 Ανάπτυξη Μοντέλου

2.2.1 Εκτίμηση Παραμέτρων

Για την ανάπτυξη του μοντέλου γίνεται η υπόθεση ότι διεξάγεται ένα πείραμα με n ασθενείς και πεθαίνουν k εξ' αυτών τις διακεκριμένες χρονικές στιγμές $t_{(1)} < t_{(2)} < ... < t_{(k)}$.

Τη χρονική στιγμή $t_{(j)}$ πεθαίνει ένας ασθενής, με συμμεταβλητές $\mathbf{x}_j$, οι συμμεταβλητές αυτές μπορεί να αναπαριστούν διάφορα χαρακτηριστικά. Την χρονική στιγμή αμέσως πριν από την $t_{(j)}$, το σύνολο των ασθενών που βρίσκονται σε κίνδυνο συμβολίζεται ως $R_{(j)}$ (risk set), δηλαδή το σύνολο των ασθενών υπό παρακολούθηση και έστω $\mathbf{x}_j = (x_{j1}, ..., x_{jp})$ το διάνυσμα των συμμεταβλήτων που αντιστοιχούν στον j ασθενή με χρόνο ζωής $t_{(j)}$. Η πιθανότητα να πεθάνει ένας συγκεκριμένος ασθενής j, δεδομένου ότι πεθαίνει ένας οποιαδήποτε ασθενής του συνόλου $R_{(j)}$, ορίζεται :

$$\frac{h(t; \mathbf{x}_j)}{\sum_{i \in R_{(j)}} h(t_j; \mathbf{x}_i)} = \frac{e^{\boldsymbol{\beta}' \mathbf{x_j}}}{\sum_{i \in R_{(j)}} e^{\boldsymbol{\beta}' \mathbf{x}_i}}$$

Η πιθανοφάνεια του συνόλου δεδομένων θα είναι:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{k} \frac{e^{\boldsymbol{\beta}' \mathbf{x}_j}}{\sum_{i \in R_{(j)}} e^{\boldsymbol{\beta}' \mathbf{x}_i}}$$

όπου $\mathbf{x}_j$ το διάνυσμα των συμμεταβλητών για τον ασθενή που πεθαίνει την διατεταγμένη χρονική στιγμή t_j. Οι παρατηρήσεις εκείνες που έχουν αποκοπεί δεν λαμβάνονται υπ'όψιν στον αριθμητή της πιθανοφάνειας, συμπεριλαμβάνονται όμως στο άθροισμα επάνω στο risk sets στις χρονικές στιγμές θανάτων που συμβαίνουν πριν μια αποκοπή. Η πιθανοφάνεια εξαρτάται από τις διατεταγμένες χρονικές στιγμές των θανάτων.

Με την μεγιστοποίηση του λογαρίθμου της παραπάνω συνάρτησης πιθανοφάνειας προκύπτει η εκτιμήτρια $\hat{\boldsymbol{\beta}}$ του διανύσματος των συντελεστών των παραμέτρων $\boldsymbol{\beta}$. Σημειώνεται ότι στην παραπάνω διαδικασία δεν εμφανίζεται η συνάρτηση $h_0(t)$ για αυτό και

ονομάζεται, μερική πιθανοφάνεια (Cox, 1975). Επιπλέον στον όρο $\boldsymbol{\beta}'\mathbf{x}$ δεν περιέχεται σταθερός όρος καθώς αυτός θα συμπεριλαμβάνεται στη βασική $h_0(t)$.

Κατά τη διαδικασία μεγιστοποίησης της λογαριθμοποιημένης μερικής πιθανοφάνειας, προκύπτουν οι εξισώσεις οι οποίες μπορούν να λυθούν με αριθμητικές μεθόδους, όπως παραδείγματος χάρη η μέθοδος Newton-Raphson, έτσι ώστε να προσδιοριστούν οι εκτιμήσεις των συντελεστών των συμμεταβλητών.

2.2.2 Ισόπαλοι χρόνοι διακοπής

Η παραπάνω μέθοδος εκτίμησης των παραμέτρων του μοντέλου μπορεί να ακολουθηθεί σε περιπτώσεις όπου δεν υπάρχουν δύο ή και παραπάνω θάνατοι(διακοπή) την ίδια χρονική στιγμή. Σε αντίθετη περίπτωση, όταν δηλαδή οι χρόνοι των θανάτων των ασθενών συμπίπτουν, δεν μπορεί να εφαρμοστεί η μέθοδος μερικής πιθανοφάνειας και ακολουθούνται διαφορετικές προσεγγίσεις για την εκτίμηση.

- Συνεχής μέτρηση χρόνου
  Έστω d_j οι ασθενείς που καταλήγουν την χρονική στιγμή t_j και είναι $d_j > 1$. Τότε είναι αναμενόμενο πως σε μια κλίμακα μεγαλύτερης ακρίβειας οι χρόνοι θανάτου καθενός θα ήταν διαφορετικοί. Συνεπώς οι θάνατοι θα έχουν γίνει με μια συγκεκριμένη σειρά, χωρίς να είναι γνωστό όμως ποία από τις δυνατές σειρές $d_j!$ θανάτων έχει προκύψει. Το να συμπεριληφθούν όλες στην μερική πιθανοφάνεια θα ήταν ιδιαίτερα πολύπλοκη ως λύση. Έτσι προτιμάται είτε η απλή προσέγγιση του Breslow η οποία διαμορφώνεται ως (Breslow, 1974) :

  $$\{\frac{e^{\boldsymbol{\beta}'\mathbf{z}_j}}{\sum_{i\in R_j} e^{\boldsymbol{\beta}'\mathbf{x}_i}}\}^{d_j}$$

  Με $\mathbf{z}_j = \sum_{k=1}^{d_j} \mathbf{x}_k$ και το x_k το διάνυσμα των συμμεταβλητών του ασθενούς k, με χρόνο θανάτου την στιγμή t_j. Η παραπάνω προσέγγιση θεωρείται καλή όταν ο λόγος $\frac{d_j}{n_j}$ είναι μικρός. Υπάρχει και η προσέγγιση του Effron η οποία είναι πιο δύσκολα υπολογίσιμη από αυτή του Breslow (Effron, 1977).

- Διακριτή μέτρηση χρόνου
  Όταν η παραδοχή ότι η ποσότητα $\frac{d_j}{n_j}$ είναι μικρή, χρησιμοποιείται η προσέγγιση του Cox στην οποία γίνεται η παραδοχή ότι η μέτρηση έγινε σε διακριτή κλίμακα (Cox, 1972). Υπολογίζονται οι πιθανές ομάδες σε ρίσκο R_j σε κάθε ισόπαλο χρόνο διακοπής. Η πιθανοφάνεια δίνεται από τον τύπο:

$$\prod_{j=1}^{k} \frac{e^{\boldsymbol{\beta}'\mathbf{z}_j}}{\sum_{q\in Q_j} e^{\boldsymbol{\beta}'\mathbf{z}_q}}$$

όπου Q_j το σύνολο των υποσυνόλων του R_j μεγέθους d_j.

2.2.3 Έλεγχοι υποθέσεων στο μοντέλο του Cox

Είναι σημαντικό στην ανάλυση του μοντέλου που έχει επιλεχθεί να μπορούν να ελεγχθούν βασικές υποθέσεις όπως αν ο συντελεστής β_i της συμμεταβλητής x_i είναι ίσος με το μηδέν, δηλαδή αν υπάρχει ανεξαρτησία μεταξύ της διάρκειας ζωής του ασθενούς και της συμμεταβλητής x_i.

1. Έλεγχος λόγου πιθανοφανειών Likelihood Ratio Test:
   Ο έλεγχος πραγματοποιείται ως εξής: γίνεται προσαρμογή του μοντέλου τόσο με την μεταβλητή x_i όσο και χωρίς, έτσι ώστε στο μοντέλο το οποίο δεν την περιέχει να ισχύει ο περιορισμός $\beta_i = 0$. Γίνεται υπολογισμός των μεγιστοποιημένων τιμών του λογαρίθμου της πιθανοφάνειας $\hat{\ell_1}$ και $\hat{\ell_0}$ με και χωρίς τη συμμεταβλητή x_i κατ΄ αντιστοιχία. Υπολογίζεται έτσι, η τιμή $-2(\hat{\ell_0} - \hat{\ell_1})$, η οποία συγκρίνεται με την χ_1^2 κατανομή. Σύμφωνα με την τιμή του παραπάνω ελέγχου, κρίνεται η αποδοχή ή η απόρριψη της μηδενικής υπόθεσης.

2. Έλεγχος Wald:
   Ένας επίσης συνηθισμένος έλεγχος για την ανεξαρτησία ή όχι της συμμεταβλητής x_i με την διάρκεια ζωής ενός ασθενούς είναι ο έλεγχος Wald. Στην συγκεκριμένη περίπτωση απαιτείται η προσαρμογή ενός μόνο μοντέλου. Η μηδενική υπόθεση όπως πριν είναι $H_0 : \beta_j = 0$. Η ελεγχοσυνάρτηση Wald ορίζεται ως εξής:

   $$\frac{\hat{\beta_j}}{se(\hat{\beta_j})}$$

   και η τιμή του συγκρίνεται με την τιμή της $N(0,1)$ κατανομής ή ισοδύναμα η παραπάνω τιμή του στατιστικού ελέγχου υψωμένη στο τετράγωνο με την τιμή της χ_1^2 κατανομής .

3. Έλεγχος Score:
   Ελέγχουμε και πάλι την μηδενική υπόθεση $H_0 : \beta = \beta_0$, για την ειδική περίπτωση όπου έχουμε μια συμμεταβλητή. Η ελεγχοσυνάρτηση σε αυτή την περίπτωση ορίζεται ως:

   $$S(\beta_0) = \frac{U(\beta_0)^2}{I(\beta_0)}$$

   όπου $U(\beta_0) = \frac{\vartheta \ln L(\beta_0)}{\vartheta \beta_0}$ και $I(\beta_0) = -\frac{\vartheta^2 \ln L(\beta_0)}{\vartheta^2 \beta_0}$.

Η στατιστική συνάρτηση ακολουθεί προσεγγιστικά την χ_1^2 κατανομή.

Οι παραπάνω έλεγχοι θεωρούνται ισοδύναμοι. Ο έλεγχος λόγου πιθανοφανειών είναι υπολογιστικά πιο δαπανηρός όμως προτιμάται ειδικότερα σε περιπτώσεις μικρών δειγμάτων. Ο έλεγχος Wald μπορεί να χρησιμοποιηθεί ως μια πρώτη ένδειξη σημαντικότητας των μεταβλητών σ΄ ένα μοντέλο, όταν υπάρχουν πολλές υποψήφιες.

2.2.4 Κριτήρια Επιλογής Μοντέλου

Κατά την προσαρμογή ενός μοντέλου επιβίωσης, σημαντικό κομμάτι αποτελεί η επιλογή των μεταβλητών εκείνων που χρειάζονται πραγματικά στο μοντέλο. Η συνάρτηση επιβίωσης μπορεί να εξαρτάται από διάφορους παράγοντες, οι οποίοι όμως μπορεί να κάνουν το μοντέλο ιδιαίτερα δύσχρηστο. Είναι σημαντικό λοιπόν να βρεθούν οι παράγοντες εκείνοι που μπορούν να περιγράψουν επαρκώς τα δεδομένα, χωρίς να κάνουν το μοντέλο περίπλοκο και δημιουργώντας ένα μοντέλο το οποίο είναι εύκολο να ερμηνευτεί. Όπως και στην ανάλυση παλινδρόμησης έτσι κι εδώ σκοπός είναι να δημιουργηθεί ένα μοντέλο όσο το δυνατόν πιο φειδωλό. Έχουν δημιουργηθεί πολλές και διαφορετικές μέθοδοι για την επιλογή των μεταβλητών σε ένα μοντέλο. Γενικότερα για την επιλογή του βέλτιστου μοντέλου γίνεται χρήση των γνωστών μεθόδων από την ανάλυση παλινδρόμησης και παρακάτω θα αναφερθούν κάποιες από αυτές.

Δύο κριτήρια τα οποία μπορούν να χρησιμοποιηθούν για την επιλογή του καλύτερου μοντέλου είναι το κριτήριο BIC (Bayesian Information Criterion) και το κριτήριο AIC (Akaike Information Criterion), στο οποίο και θα γίνει ιδιαίτερη αναφορά (Schwarz, 1978), (Akaike, 1998).

Σημειώνεται πως το κριτήριο BIC ορίζεται ως:

$$BIC = -2\hat{\ell} + k \ln n$$

Το κριτήριο AIC επιλέγει ένα μοντέλο με το μικρότερο δυνατό αριθμό συμμεταβλητών και ορίζεται ως εξής:

$$AIC = -2\hat{\ell} + 2k$$

όπου $\hat{\ell}$, η μεγιστοποιημένη λογαριθμοποιημένη πιθανοφάνεια του μοντέλου
και k, ο αριθμός των συμμεταβλητών του μοντέλου (Καρώνη & Οικονόμου, 2017).

Το AIC επιλέγει το μοντέλο εκείνο που με την μικρότερη τιμή του κριτηρίου. Γενικά η προσθήκη μεταβλητών στο μοντέλο οδηγεί σε καλύτερη προσαρμογή του μοντέλου είτε αυτές είναι σημαντικές είτε όχι. Υπάρχει έτσι μείωση της τιμής $-2\hat{\ell}$ όμως αυξάνεται με αυτόν τον τρόπο ο όρος $2k$, που θεωρείται μια 'ποινή' για την εισαγωγή πολλών μεταβλητών σε ένα μοντέλο. Έτσι μονό η εισαγωγή πρόσθετων μεταβλητών στο μοντέλο μειώνει την τιμή του κριτηρίου, μόνο αν αυτές οδηγούν σε καλύτερη προσαρμογή του μοντέλου, τόσο ώστε να ξεπερνά τον όρο της 'ποινής'.

Η επιλογή των σημαντικών μεταβλητών μπορεί να γίνει και από τις μεθόδους: προς τα εμπρός επιλογής (Forward Selection), της διαδοχικής αφαίρεσης (Backward Elimination) καθώς και της διαδικασίας κατά βήματα (Stepwise Selection) (Collett, 2015). Στην προς τα εμπρός επιλογή, οι μεταβλητές προσθέτονται διαδοχικά στο μοντέλο. Η μεταβλητή που προστίθεται κάθε φορά είναι εκείνη που κάνει μεγαλύτερη την μείωση της ποσότητας $-2\hat{\ell}$, έχοντας ξεκινήσει από ένα μοντέλο χωρίς καμία μεταβλητή. Η διαδικασία σταματά όταν η επομένη υποψήφια μεταβλητή για εισαγωγή στο μοντέλο, δεν μειώνει σημαντικά την τιμή $-2\hat{\ell}$. Στην διαδικασία της της διαδοχικής αφαίρεσης προσαρμόζεται το μοντέλο με τις περισσότερες δυνατές συμμεταβλητές. Στη συνέχεια αφαιρούνται διαδοχικά οι μεταβλητές που αυξάνουν λιγότερο την τιμή $-2\hat{\ell}$ με την αφαίρεση τους. Στη διαδικασία κατά βήματα συμβαίνει ένας συνδυασμός των παραπάνω. Κάθε φορά που προστίθεται μια καινούργια μεταβλητή στο μοντέλο, ελέγχεται αν η προηγούμενη που είχε συμπεριληφθεί, μπορεί τώρα να παραληφθεί.

2.3 Επεκτάσεις του μοντέλου του Cox

Υπάρχουν περιπτώσεις στις οποίες χρειάζεται στο μοντέλο του Cox να γίνουν κάποιες κατάλληλες τροποποιήσεις για την διαχείριση δεδομένων που παρουσιάζουν κάποια ιδιαίτερα χαρακτηριστικά (Καρώνη, 2009).

2.3.1 Στρωματοποιημένη ανάλυση

Είναι πολύ σύνηθες όταν προσαρμόζεται το μοντέλο του Cox να χρησιμοποιείται η δυνατότητα που δίνει να γίνει στρωματοποιημένη ανάλυση και έτσι να προκύψει το στρωματοποιήμενο μοντέλο Cox (stratified Cox model). Η ανάλυση αυτή προκύπτει όταν υπάρχουν ενδείξεις πως οι συναρτήσεις διακινδύνευσης μεταξύ δύο ή και περισσότερων κατηγοριών των δεδομένων δεν βρίσκονται σε αναλογία μεταξύ τους. Παραδείγματος χάρη, πολύ συχνά συναντάται αυτή η προσέγγιση όταν σε μια μελέτη υπάρχει η κατηγορική μεταβλητή 'φύλο', τότε λοιπόν τα δεδομένα χωρίζονται σε δύο στρώματα, γυναίκες και άνδρες. Οι συναρτήσεις διακινδύνευσης στην περίπτωση αυτή ορίζονται ως:

$$h(t;\mathbf{x}) = \begin{cases} e^{\boldsymbol{\beta}'\mathbf{x}} h_{0_1}(t), & \text{για τους άνδρες} \\ e^{\boldsymbol{\beta}'\mathbf{x}} h_{0_2}(t), & \text{για τις γυναίκες} \end{cases}$$

Όπως αναφέρθηκε και παραπάνω, η αναλογικότητα της διακινδύνευσης ισχύει για όλες τις μεταβλητές έκτος από αυτή του φύλου, για την οποία έχει γίνει η στρωματοποίηση, αφού οι βασικές συναρτήσεις διακινδύνευσης $h_{0_1}(t)$,$h_{0_2}(t)$, για τους άνδρες και τις γυναίκες αντίστοιχα δεν βρίσκονται σε αναλογία.

Σημειώνεται πως η στρωματοποίηση δεν γίνεται μόνο σε κατηγορικές μεταβλητές αλλά και σε ποσοτικές, χωρίζοντας την μεταβλητή ενδιαφέροντος σε ομάδες. Για παράδειγμα αν σε ένα μοντέλο περιέχεται ως μεταβλητή η ηλικία των ασθενών που συμμετέχουν στη μελέτη, μπορεί να κριθεί αναγκαίο οι ασθενείς να χωριστούν σε ηλικιακές ομάδες, όπως ασθενείς άνω των 50 ετών και κάτω των 50. Έτσι γενικότερα η συνάρτηση διακινδύνευσης ορίζεται για ένα άτομο που ανήκει στο m στρώμα ως:

$$h_m(t;\mathbf{x}) = h_{0_m}(t)e^{\boldsymbol{\beta}'\mathbf{x}}, m = 1, ...s$$

m: το στρώμα του παράγοντα
s: το πλήθος των επιπέδων της μεταβλητής
$h_{0m}(t)$: η βασική συνάρτηση κινδύνου στο m στρώμα

Τα άτομα που ανήκουν στο ίδιο στρώμα έχουν συνεπώς τις ίδιες βασικές συναρτήσεις διακινδύνευσης και επίσης οι συναρτήσεις διακινδύνευσης τους είναι ανάλογες σε αντίθεση με άτομα που βρίσκονται σε διαφορετικά στρώματα.

Για την εκτίμηση των παραμέτρων, χρησιμοποιείται και πάλι η λογαριθμοποιημένη συνάρτηση μερικής πιθανοφάνειας, υπολογισμένη για κάθε στρώμα m. Συνολικά για όλα τα στρώματα προκύπτει η συνάρτηση:

$$l(\boldsymbol{\beta}) = \sum_{m=1}^{s} l_m(\boldsymbol{\beta}) = \sum_{m=1}^{s}\sum_{j=1}^{k_m} \boldsymbol{\beta}'\mathbf{x}_{mj} - \sum_{m=1}^{s}\sum_{j=1}^{k_m} ln[\sum_{i\in R_{mj}} e^{\boldsymbol{\beta}'\mathbf{x_{mi}}}]$$

Η εκτίμηση των παραμέτρων γίνεται τελικά ακριβώς όπως είχε αναφερθεί στην παράγραφο 2.2.1.

2.3.2 Συμμεταβλητές εξαρτώμενες από το χρόνο

Μια άλλη περίπτωση στην οποία το μοντέλο του Cox χρειάζεται κάποια τροποποιήση για να μπορέσει να εφαρμοστεί, είναι όταν οι τιμές μιας συμμεταβλητής μεταβάλλονται με τον χρόνο, όπως για παράδειγμα η ηλικία ενός ασθενή ή η δοσολογία της φαρμακευτικής αγωγής που λαμβάνει ή και μια επαναλαμβανόμενη μέτρηση σε βάθος χρόνου. Οι μεταβλητές αυτές διακρίνονται σε δύο κατηγορίες :

- εξωτερικές (external) : οι μεταβλητές αυτές καλούνται εξωτερικές καθώς ελέγχονται από τον πειραματιστή και δεν επηρεάζονται από την πορεία του ασθενή στο πείραμα ή αλλάζουν με έναν προβλέψιμο τρόπο και είναι κοινές για όλους τους ασθενείς που συμμετέχουν στο πείραμα
- εσωτερικές (internal): οι εσωτερικές μεταβλητές επηρεάζονται από την πορεία της υγείας του ασθενή και διαφοροποιούνται για κάθε ασθενή που συμμετέχει στο πείραμα, όπως για παράδειγμα οι μετρήσεις της πίεσης ενός ασθενή ή η μόλυνση ενός ασθενή.

Για να συμπεριληφθούν τέτοιου είδους δεδομένα στην μερική συνάρτηση λογαριθμοποιήμενης πιθανοφάνειας θα πρέπει ο όρος $\mathbf{x}$ να αντικατασταθεί από το $\mathbf{x}(t_{(j)})$. Έτσι λοιπόν, η συνάρτηση διαμορφώνεται ως εξής:

$$l(\boldsymbol{\beta}) = \sum_{j=1}^{k} \boldsymbol{\beta}'\mathbf{x}_j - \sum_{j=1}^{k} ln[\sum_{i \in R_j} e^{\boldsymbol{\beta}'\mathbf{x}_i(t_{(j)})}]$$

Είναι εμφανές ότι για να μπορεί να υπολογιστεί η παραπάνω ποσότητα, θα πρέπει να είναι γνωστή σε κάθε χρονική στιγμή θανάτου (ή διακοπής) $t_{(j)}$ η τιμή κάθε συμμεταβλητής. Αυτό όμως πρακτικά είναι ακατόρθωτο καθώς είναι πολύ δύσκολη η συνεχής παρακολούθηση των τιμών των συμμεταβλητών. Ως επί το πλείστον στα πειράματα οι μετρήσεις γίνονται ανά διαστήματα και οι τιμές $\mathbf{x}(t_{(j)})$ στους ενδιάμεσους θανάτους (ή διακοπές), δεν είναι διαθέσιμες. Για να λυθεί το παραπάνω πρόβλημα μπορεί να χρησιμοποιηθεί η τελευταία μέτρηση πριν το θάνατο ή να γίνει παρεμβολή μεταξύ δυο μετρήσεων.

2.4 Έλεγχοι της υπόθεσης αναλογικής διακινδύνευσης

Πολύ σημαντικό κομμάτι στη διαδικασία προσαρμογής ενός μοντέλου είναι να επιβεβαιωθεί η υπόθεση, πως η επιλογή του μοντέλου ήταν αυτή που πραγματικά ταιριάζει στα δεδομένα (Καρώνη, 2009). Όταν προσαρμόζεται το μοντέλο του Cox έχει γίνει η αποδοχή, ότι η υπόθεση της αναλογικής διακινδύνευσης ισχύει. Δηλαδή, ότι ο λόγος :

$$\frac{h_l(t)}{h_j(t)} = \frac{h_0(t)e^{\boldsymbol{\beta}'\mathbf{x}_l}}{h_0(t)e^{\boldsymbol{\beta}'\mathbf{x}_j}} = e^{\boldsymbol{\beta}'(\mathbf{x_l}-\mathbf{x_j})}$$

δύο συναρτήσεων διακινδύνευσης για δύο ασθενείς i και j, είναι σταθερός και ανεξάρτητος του χρόνου.

Ο έλεγχος για την σωστή υιοθέτηση του μοντέλου αναλογικής διακινδύνευσης είναι απαραίτητος για να μπορούν τα αποτελέσματα της μελέτης να θεωρούνται αξιόπιστα. Για τον παραπάνω έλεγχο έχουν αναπτυχθεί δύο προσεγγίσεις οι οποίες και θα παρουσιαστούν παρακάτω. Σε περίπτωση που η υπόθεση της αναλογικής διακινδύνευσης δεν ισχύει θα πρέπει, είτε να γίνει κάποιος μετασχηματισμός στα δεδομένα προκειμένου να ισχύει, είτε να επιλεχθεί κάποιο άλλο πιο ταιριαστό στα δεδομένα μοντέλο.

1. Έλεγχος μέσω συμμεταβλητών εξαρτωμένων από τον χρόνο.
   Για κάθε μεταβλητή x_i ακολουθούνται τα παρακάτω:

   - δημιουργείται μια καινούργια μεταβλητή $z = x_i t$ ή $z = x_i lnt$ ή γενικότερα κάποια άλλη συνάρτηση του χρόνου, έτσι ώστε υπάρξει αλληλεπίδραση της x_i με το χρόνο. Προτιμούνται συναρτήσεις απλής μορφής.

- Προσαρμόζεται το μοντέλο του Cox, περιλαμβάνοντας και τη μετασχηματισμένη μεταβλήτη z μαζί με τις υπόλοιπες συμμεταβλητές.
- Πραγματοποιείται ένας έλεγχος υποθέσεων, με μηδενική υπόθεση $H_0 : \beta_z = 0$, όπου β_z ο συντελεστής της z.

Αν γίνει αποδεκτή η μηδενική υπόθεση $H_0 : \beta_z = 0$, σημαίνει πως η επίδραση της συμμεταβλητής x_i εκφράζεται μέσα από τον όρο $\beta_i x_i$ της συνάρτησης διακινδύνευσης $h(t)$, που είναι ανεξάρτητος από τον χρόνο και συνεπώς ισχύει η υπόθεση της αναλογικότητας. Μη αποδοχή της μηδενικής υπόθεσης, θα σήμαινε πως δεν ισχύει η αναλογικότητα και συνεπώς οι συναρτήσεις διακινδύνευσης μεταξύ δύο ατόμων δεν είναι ανάλογες.

2. Γραφικός Έλεγχος
Μια γνωστή προσέγγιση για τον έλεγχο της υπόθεσης της αναλογικής διακινδύνευσης είναι να γίνει μια γραφική παράσταση της μη-παραμετρικής συνάρτησης $ln\{-ln\hat{S}\}$ σε συνάρτηση με το χρόνο t, όπου $\hat{S}$ οι Kaplan-Meier εκτιμήτριες των αντίστοιχων ομάδων των δύο ασθενών που εξετάζονται. Είναι σύνηθες ο έλεγχος αυτός να προηγείται της προσαρμογής του μοντέλου καθώς μπορεί να δώσει πολύ ισχυρές ενδείξεις για το αν ισχύει η υπόθεση της αναλογικής διακινδύνευσης ή όχι. Όμως η εκτιμήτρια Kaplan-Meier λαμβάνει υπόψη της μόνο τις τιμές της συμμεταβλητής που καθορίζει τις ομάδες και καμία άλλη από αυτές που συμμετέχουν στο μοντέλο. Έτσι πρέπει να γίνουν κάποιες τροποποιήσεις στον παραπάνω έλεγχο για την βελτιστοποίηση του.

Για την εκτιμήση της $\hat{S}$ θα χρειαστεί να εκτιμηθούν τόσο οι συντελεστές $\boldsymbol{\beta}$ αλλά και η ακαθόριστη $S_0(t)$. Για την τελευταία χρησιμοποιείται η προσέγγιση Breslow (Breslow, 1974) που λαμβάνει υπόψη όλες τις συμμεταβλητές στο στρωματοποιημένο μοντέλο του Cox :

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

όπου

$$\hat{H}_0(t) = \sum_{t_{(j)} \leq t} \frac{d_j}{\sum_{i \in R_j} e^{\hat{\boldsymbol{\beta}}' \mathbf{x}_i}}$$

Κρίνεται απαραίτητο λοιπόν, προτού γίνει ο γραφικός έλεγχος της αναλογικής διακινδύνευσης του μοντέλου να αντικατασταθεί η εκτιμήτρια Kaplan-Meier με την κατά στρώματα εκτιμήτρια της συνάρτησης επιβίωσης.
Στη συνέχεια ακουλουθείται η εξής διαδικασία :

- Προσαρμόζεται το στρωματοποιημένο μοντέλο του Cox, όπου τα στρώματα καθορίζονται από τις τιμές της συμμεταβλητής, για την οποία ελέγχουμε

την υπόθεση αναλογικής διακινδύνευσης.

- Σε κάθε στρώμα γίνεται η εκτίμηση της βασικής συνάρτησης επιβίωσης $S_0(t)$. Για κάθε στρώμα k η βασική συνάρτηση επιβίωσης θα συμβολίζεται ως : $\hat{S}_{0k}(t)$
- Έστω $\bar{\mathbf{x}}_\mathbf{k}$ το διάνυσμα των μέσων τιμών των συμμεταβλητών στο k στρώμα και έτσι $\hat{S}_k(t) = \hat{S}_{0k}(t)e^{\hat{\boldsymbol{\beta}}'\bar{\boldsymbol{x}}_\boldsymbol{k}}$
- Προχωρώντας στην δημιουργία της γραφικής παράστασης για να ευσταθεί η υπόθεση της αναλογικής διακινδύνευσης για την συμμεταβλητή που μελετάται θα πρέπει οι καμπύλες των $ln\{-ln\hat{S}_k\}$, $k = 1, ..., s$ σε συνάρτηση με το χρόνο t να είναι παράλληλες.

2.5 Έλεγχοι καταλληλότητας του μοντέλου μέσω υπολοίπων

Η εξέταση των υπολοίπων αποτελεί μια βασική στρατηγική για τον έλεγχο της καταλληλότητας ενός μοντέλου (Καρώνη, 2009). Η χρήση τους επεκτείνεται και στον έλεγχο της υπόθεσης της αναλογικής διακινδύνευσης αλλά και γενικότερα μπορούν να δείξουν κατά πόσο τα δεδομένα συμφωνούν με τις προβλέψεις του μοντέλου αλλά και με τις προϋποθέσεις αυτού. Αποτελούν επίσης έναν τρόπο εντοπισμού άτυπων τιμών (outliers). Στην περίπτωση της γραμμικής παλινδρόμησης τα υπόλοιπα ορίζονται ως εξής:

$$\begin{aligned}\hat{\epsilon}_i &= y_i - \hat{y}_i \\ &= y_i - \boldsymbol{\beta}'\mathbf{x_i}\end{aligned}$$

Στα μοντέλα διάρκειας ζωής δεν προτιμούνται αυτά τα υπόλοιπα. Τα υπόλοιπα που χρησιμοποιούνται στο μοντέλο του Cox κατά βάση είναι τα υπόλοιπα Cox-Snell (ή γενικευμένα υπόλοιπα) και υπόλοιπα Schoenfeld (ή μερικά υπόλοιπα -partial residuals, τα υπόλοιπα Martingale και τα υπόλοιπα απόκλισης deviance, τα οποία και θα αναλυθούν παρακάτω.

2.5.1 Υπόλοιπα Cox-Snell

Οι Cox και Snell το 1968 πρότειναν μια διαφορετική προσέγγιση (Cox & Snell, 1968). Θεωρώντας μια τ.μ T_j με κατανομή που εξαρτάται από τις συμμεταβλητές $\mathbf{x_j}$ και παραμέτρους $\boldsymbol{\theta}$. Ορίζονται συναρτήσεις της μορφής:

$$w_j(T_j; \mathbf{x_j}, \boldsymbol{\theta})$$

Σε περίπτωση που είναι ανεξάρτητες, ισόνομες και οι κατανομές τους δεν εξαρτώνται από άγνωστους παραμέτρους, τότε ως υπόλοιπα μπορούν να χρησιμοποιηθούν οι ποσότητες :

$$\hat{\epsilon} = w_i(\mathbf{T_i}; \mathbf{x_i}, \hat{\theta})$$

Για το μοντέλο του Cox οι τιμές :

$$\begin{aligned}\hat{\epsilon_j} &= ln(\hat{S}(t_{(j)}; \mathbf{x_j}) \\ &= \hat{H}(t_{(j)}; \mathbf{x_j}) \\ &= \hat{H}_0(t_{(j)}) e^{\hat{\boldsymbol{\beta}'}\mathbf{x_j}}\end{aligned}$$

ορίζονται ως υπόλοιπα, όπου $\hat{S(.)}$ και $\hat{H(.)}$ οι εκτιμήτριες της συνάρτησης επιβίωσης και της συνάρτησης σωρευτικής διακινδύνευσης αντίστοιχα, οι οποίες υπολογίζονται με τον τρόπο που έχει αναφερθεί παραπάνω. Οι τιμές των υπολοίπων εξετάζονται γραφικά και αν υπάρχουν αρκετές ενδείξεις πως είναι της Εκθετικής κατανομής με παράμετρο 1, τότε μπορεί να γίνει η παραδοχή πως το μοντέλο που προσαρμόστηκε είναι κατάλληλο για τα δεδομένα.

Τα υπόλοιπα αυτά μπορούν να εφαρμοστούν σε μοντέλα οποιασδήποτε μορφής με ή και χωρίς συμμεταβλητές. Η μη-παραμετρική εκτίμηση όμως της $\hat{H_0(t)}$ δημιουργεί δυσκολίες και αβεβαιότητα και έτσι τα υπόλοιπα Cox-Snell φαίνεται να είναι λιγότερο χρήσιμα στο μοντέλο του Cox. Σημειώνεται όμως πως έχουν ιδιαίτερη εφαρμογή σε παραμετρικά μοντέλα επιβίωσης.

Είναι σημαντικό να αναφερθεί πως υπάρχουν και τροποποιημένα υπόλοιπα Cox-Snell με τα οποία μπορεί να ληφθούν υπόψη και τα λογοκριμένα δεδομένα.

2.5.2 Υπόλοιπα Schoenfeld

Το 1982, ο Schoenfeld έδωσε μια άλλη προσέγγιση για την εύρεση των υπολοίπων στο μοντέλο του Cox, προκειμένου να μην χρειάζεται η εκτίμηση της βασικής σωρευτικής συνάρτησης διακινδύνευσης $\hat{H}_0(t)$ εφόσον δεν γίνεται εκτίμηση ούτε για την συνάρτηση διακινδύνευσης $\hat{h}_0(t)$ (Schoenfeld, 1982). Παρακάτω αναλύεται η διαδικασία με την οποία ορίζονται.

Η πιθανότητα να σταματήσει η λειτουργία της μονάδας j, γνωρίζοντας πως διακόπτεται η λειτουργία μιας μονάδας την χρονική στιγμή $t_{(j)}$ από ένα σύνολο μονάδων $R_{(j)}$ σε ρίσκο αμέσως πριν αυτή τη στιγμή είναι :

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

Όμως επειδή το ποια ακριβώς μονάδα θα διακοπεί τη στιγμή $t_{(j)}$ από το σύνολο $R_{(j)}$ είναι άγνωστο, η τιμή των συμμεταβλητών θα είναι τ.μ με αναμενόμενη τιμή

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

Έτσι μπορούν να ορίστουν τα υπόλοιπα Schoenfeld:

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

Όπως είναι εμφανές για να οριστούν τα υπόλοιπα χρειάζονται τις συμμεταβλητές $\boldsymbol{x}$. Για κάθε μονάδα υπολογίζονται τόσα υπόλοιπα όσο και οι συμμεταβλήτες που υπάρχουν στο μοντέλο. Επίσης αποτελούν ενδείξεις για ακραίες τιμές, όταν παίρνουν πολύ μεγάλες τιμές.

Είναι αρκετά σύνηθες να χρησιμοποιούνται και τα κλιμακοποιημένα (scaled) υπόλοιπα Schoenfeld που δίνονται από τον τύπο (Grambsch & Therneau, 1994):

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

όπου k το πλήθος των μη-αποκομμένων παρατηρήσεων και $\hat{V}(\hat{\beta})$ ο εκτιμώμενος πίνακας διασποράς των $\hat{\boldsymbol{\beta}}$.

Χρησιμοποιώντας τα κλιμακοποιημένα υπόλοιπα μπορεί να γίνει και έλεγχος για την υπόθεση αναλογικής διακινδύνευσης (Grambsch & Therneau, 1994). Γίνεται αρχικά η υπόθεση ότι μια συμμεταβλητή δεν είναι σταθερή, εξαρτάται δηλαδή από το χρόνο. Τότε ορίζεται :

$$E(r_{ij})^* \simeq \beta_i(t_{(j)}) - \hat{\beta}_i$$

όπου $\beta_i(t_{(j)})$ ο συντελεστής της συμμεταβλητής i τη χρονική στιγμή $t_{(j)}$. Για να γίνει αποδεκτή η υπόθεση της αναλογικής διακινδύνευσης ότι δηλαδή, $\beta_i(t) = \beta_i \ \forall t$, θα πρέπει η γραφική παράσταση της ποσότητας $r_{ij}^* + \beta_i$ με τον χρόνο t να εμφανίζει μια οριζόντια γραμμή.

2.5.3 Υπόλοιπα Martingale

Σε περίπτωση που χρειάζεται να προσδιοριστεί η συναρτησιακή μορφή κάποιας μεταβλητής που πρέπει να εισαχθεί στο μοντέλο ή να προσδιοριστούν ακραίες τιμές αλλά και για να ελεγχθεί η υπόθεση της αναλογικής διακινδύνευσης γίνεται χρήση των υπολοίπων martingale. Ορίζονται ως η ποσότητα:

$$\hat{r}_M = \delta_i - \hat{r}_{CS}$$

όπου CS=Cox-Snell και

$$\delta_i = \begin{cases} 0, & \text{αν είναι αποκομμένη παρατήρηση} \\ 1, & \text{αν δεν είναι αποκομμένη παρατήρηση} \end{cases}$$

Παίρνουν τιμές από το $-\infty$ εως το 1 (για τις αποκομμένες παρατηρήσεις έχουν αρνητικές τιμές) και για μεγάλα δείγματα είναι ασυσχέτιστα, με αναμενόμενη μέση τιμή το 0. Έτσι μοιάζουν αρκετά με τα υπόλοιπα που γνωρίζουμε από την γραμμική παλινδρόμηση (Collett, 2015).

Μια ακόμη ερμηνεία των υπολοίπων αυτών είναι πως εκφράζουν τη διαφορά μεταξύ του παρατηρούμενου αριθμού των θανάτων για την i παρατήρηση στο διάστημα $(0, t_i)$ και τον αναμενόμενο αριθμό θανάτων για το προσαρμοσμένο μοντέλο. Σημειώνεται επίσης πως τα υπόλοιπα αυτά δεν είναι συμμετρικά κατανεμημένα γύρω από το 0 ακόμα και αν το μοντέλο που έχει προσαρμοστεί είναι το σωστό και αυτή η κυρτότητα κάνει τα γραφήματα των υπολοίπων δύσκολα να ερμηνευτούν (Collett, 2015).

2.5.4 Υπόλοιπα απόκλισης (deviance)

Το 1990 ο Therneau και άλλοι πρότειναν τα υπόλοιπα απόκλισης, τα οποία βασίζονται στα υπόλοιπα martingale, αλλά όμως είναι αρκετά πιο συμμετρικά από τα martingale. Τα υπόλοιπα απόκλισης χρησιμοποιούνται και αυτά για την εύρεση ακραίων τιμών (outliers) όμως δεν αποδείχτηκαν τόσο χρήσιμα όσο είχαν φανεί στην αρχή. Γενικότερα έχουν πολλές ομοιότητες με τα υπόλοιπα της γραμμικής παλινδρόμησης. Ορίζονται ως:

$$\hat{r}_{D_i} = sgn(\hat{r}_{M_i})\sqrt{-2(\hat{r}_{M_i} + \delta_i log(\delta_i - \hat{r}_{M_i})}$$

όπου $sgn(\hat{r}_{M_i}) = 1$ για $\hat{r}_{M_i} > 0$ και $sgn(\hat{r}_{M_i}) = -1$ για $\hat{r}_{M_i} < 0$. Ο όρος αυτός διασφαλίζει ότι τα υπόλοιπα απόκλισης και τα martingale υπόλοιπα είναι ομόσημα.

2.6 Καμπύλη ROC

Ένα πολύ σημαντικό εργαλείο για την προβλεπτική ικανότητα ενός μοντέλου είναι η καμπύλη ROC (Receiver Operating Characteristic). Χρησιμοποιώντας τις καμπύλες ROC μπορούμε να έχουμε πρόσβαση σε έναν 'δείκτη' ακρίβειας για την ικανότητα πρόβλεψης που έχει το μοντέλο μας. Η καμπύλη ROC καθιερώθηκε περίπου το 1941 και είχε την πρώτη εφαρμογή του σε στρατιωτικά ραντάρ, στην οποία οφείλει και το όνομα του.

Για την εφαρμογή της καμπύλης στο μοντέλο του Cox ορίζεται ο χρόνος επιβίωσης T_i για τον ασθενή i και γίνεται η υπόθεση ότι παρατηρούμε μόνο την ελάχιστη τιμή T_i και C_i, όπου C_i ένας ανεξάρτητος χρόνος αποκοπής. Έστω λοιπόν πως ορίζεται ο χρόνος παρακολούθησης $X_i = min(T_i, C_i)$ και έστω η δείκτρια $\Delta_i = \mathbb{1}(T_i \leq C_i)$ για το αν η παρατήρηση είναι αποκομμένη ή όχι. Επιπλέον για τον χρόνο επιβίωσης T_i μπορούμε να χρησιμοποιήσουμε μια αναπαράσταση μέσω της διαδικασία καταμέτρησης $N_i^* = \mathbb{1}(T_i \leq t)$ ή της αντίστοιχης αύξησης $dN_i^* = N_i^*(t) - N_i^*(t-)$. Η ανάλυση μας θα εστιάσει στην διαδικασία καταμέτρησης $N_i^*(t)$ που αφορά μόνο τον χρόνο επιβίωσης T_i, αντί της ευρέως χρησιμοποιούμενης $N_i^* = \mathbb{1}(X_i \leq t, \Delta = 1)$, που εξαρτάται από τον χρόνο αποκοπής (Fleming & Harrington, 2011). Τέλος, θα χρειαστεί να ορίσουμε τον δείκτη ρίσκου $R_i(t) = \mathbb{1}(X_i \geq t)$ καθώς επίσης να κάνουμε την υπόθεση πως για κάθε παρατήρηση i έχουμε ένα σύνολο από χρονικά αμετάβλητες συμμεταβλητές $Z_i = (Z_{i1}, Z_{i2}, ..., Z_{ip})$ (Heagerty & Zheng, 2005).

Παρακάτω θα επικεντρωθούμε στο μοντέλο του Cox για να δημιουργήσουμε ένα score για το μοντέλο καθώς και για να αξιολογηθεί η προβλεπτική του ικανότητα, όμως υπάρχουν μέθοδοι αξιολόγησης της ακρίβειας ενός προγνωστικού score που μπορούν να επεκταθούν σε οποιονδήποτε τύπο παλινδρόμησης. Εν προκειμένω διάφορες μέθοδοι συντελεστών, όπως η σταθμισμένη εκτίμηση μερικής πιθανοφάνειας, δίνουν μια προσέγγιση για την εκτίμηση ακριβών αποτελεσμάτων (Hastie & Tibshirani, 1993), (Cai & Sun, 2003).

Θα προχωρήσουμε λοιπόν, δίνοντας κάποιες βασικές αρχές για την εκτίμηση της μερικής πιθανοφάνειας. Υπό την προϋπόθεση της αναλογικής διακινδύνευσης, έχουμε:

$$\lambda(t|\boldsymbol{Z}_i) = \lambda_0(t) exp(Z_i^T \boldsymbol{\beta})$$

όπου

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

Έτσι οι συναρτήσεις score της μερικής πιθανοφάνειας μπορούν να γραφούν ως:

$$\mathbf{0} = \sum_i \Delta_i [\boldsymbol{Z}_i - \sum_i \pi_k(\boldsymbol{\beta}, X_i) \boldsymbol{Z}_i)]$$

		Πραγματική Κατάσταση		
		$Y=1$	$Y=0$	
Πρόβλεψη	$Y=1$	a	b	$a+b$
	$Y=0$	c	d	$c+d$
		$a+c$	$b+d$	n

όπου $\pi_k(\boldsymbol{\beta}, t) = R_k(t) \cdot exp(\boldsymbol{Z}_i^T \boldsymbol{\beta})/W(t)$ και $W(t) = \sum_j R_j(t) exp(\boldsymbol{Z}_j^T \boldsymbol{\beta})$. Λύνοντας τις παραπάνω εξισώσεις προκύπτουν οι εκτίμησης για τους συντελεστές $\hat{\boldsymbol{\beta}}$ της μέγιστης μερικής πιθανοφάνειας.

Για τα δυαδικά αποτελέσματα Y_i που έχουν προκύψει, η ακρίβεια της πρόβλεψης ορίζεται από δύο πολύ βασικές έννοιες την ευαισθησία (sensitivity) $P(\hat{p_i} > c|Y_i = 1)$, και την ειδικότητα (specificity) $P(\hat{p_i} \leq c|Y_i = 0)$,, όπου $\hat{p}$ είναι μια πρόβλεψη και c είναι ένα όριο για να ταξινομηθούν οι προβλέψεις ως θετικές αν $(\hat{p_i} > c)$ ή αρνητικές αν $(\hat{p_i} \leq c)$.

Ορίζουμε ως:

- Ευαισθησία (sensitivity): Το ποσοστό της ορθής πρόβλεψης της κατάστασης $Y = 1$, ή αλλιώς 'το ποσοστό των αληθώς θετικών αποτελεσμάτων (true positive rate)'. $TPR = \frac{a}{a+c}$

- 1−Ειδικότητα: Το ποσοστό των ψευδώς θετικών αποτελεσμάτων $(Y = 1)$ ενώ στην πραγματικότητα ισχύει $(Y = 0)$: $FPR = \frac{b}{b+d}$

- Ειδικότητα (specificity): ο ποσοστό της ορθής πρόβλεψης της κατάστασης $Y = 0$, ή αλλιώς 'το ποσοστό των αληθώς αρνητικών αποτελεσμάτων (true negative rate)'. $TNR = \frac{d}{b+d}$

- Θετική προβλεπόμενη τιμή: Η πιθανότητα εμφάνισης θετικού περιστατικού μεταξύ όλων των θετικών προβλέψεων. $PPV = \frac{a}{a+b}$

- Αρνητική προβλεπόμενη τιμή: Η πιθανότητα εμφάνισης αρνητικού περιστατικού μεταξύ όλων των αρνητικών προβλέψεων. $NPV = \frac{d}{c+d}$

- Ακρίβεια (Accuracy): Το ποσοστό των πραγματικών αποτελεσμάτων (τόσο αληθινά θετικά όσο και αληθινά αρνητικά) μεταξύ του συνόλου που εξετάστηκαν. $ACC = \frac{a+d}{n}$

- Θετικός λόγος πιθανοφανειών $(LR+)$: Δείχνει πόσες φορές πιο συχνά εμφανίζεται το θετικό αποτέλεσμα στην πραγματική κατάσταση $(Y = 1)$ σε σχέση με την πραγματική κατάσταση $(Y = 0)$. $LR+ = \frac{TPR}{FPR}$

- Αρνητικός λόγος πιθανοφανειών ($LR-$): Δείχνει πόσες φορές πιο συχνά εμφανίζεται το αρνητικό αποτέλεσμα στην πραγματική κατάσταση ($Y = 0$) σε σχέση με την πραγματική κατάσταση ($Y = 1$). $LR- = \frac{FPR}{TPR}$.

- DOR (diagnostic odds ratio): Ένα μέτρο αποτελεσματικότητας ενός διαγνωστικού τεστ. Τιμές μεγαλύτερες της μονάδας ή και υψηλότερες δείχνουν πως το τεστ είναι πολύ αποτελεσματικό.

Υπολογίζοντας τις τιμές της ευαισθησίας και της ειδικότητας για κάθε όριο c, στο εύρος $[0, 1]$, μπορεί να σχηματιστεί η χαρακτηριστική καμπύλη ROC, με την οποία απεικονίζεται η προβλεπτική ικανότητα του μοντέλου καθώς το όριο μεταβάλλεται. Η καμπύλη είναι μια γραφική παράσταση της ευαισθησίας σε συνάρτηση με την ειδικότητα ή εναλλακτικά η ευαισθησία σε συνάρτηση με την 1−ειδικότητα. Και στις δύο περιπτώσεις η κλίμακα του οριζόντιου άξονα θα είναι από το 0 έως το 1.

Βέλτιστη πρόβλεψη επιτυγχάνεται όταν παρατηρούμε τιμές του ορίου c με υψηλή ευαισθησία και ταυτόχρονα υψηλή ειδικότητα. Στην περίπτωση αυτή η καμπύλη ROC θα πλησίαζε την πάνω αριστερή γωνία στο σχήμα που έχει προκύψει. Ένα μέτρο για το πόσο πλησιάζει η καμπύλη την γωνία αυτή είναι το εμβαδόν κάτω από την καμπύλη (area under the curve, AUC), με μέγιστη τιμή το 1. Επιπλέον αν η καμπύλη πλησιάζει τη διαγώνιο που εμφανίζεται στο σχήμα που έχει δημιουργηθεί (πάνω στην οποία ισχύει $AUC = 0.5$), τότε τα ποσοστά των αληθώς θετικών και ψευδώς θετικών αποτελεσμάτων θα ήταν ίσα, γεγονός που θα σήμαινε ότι η πρόβλεψη είναι ανεξάρτητη από την πραγματική τιμή της Y (Καρώνη & Οικονόμου, 2017).

Κεφάλαιο 3

Ποινικοποιημένες Μέθοδοι

Στο κεφάλαιο αυτό θα αναλυθούν οι τρόποι με τους οποίους μπορεί να γίνει η εκτίμηση των συντελεστών ενός μοντέλου παλινδρόμησης. Όπως είναι γνωστό στο πολλαπλό γραμμικό μοντέλο παλινδρόμησης οι συντελεστές εκτιμώνται με τη μέθοδο των ελαχίστων τετραγώνων. Πιο συγκεκριμένα το πρόβλημα της εκτίμησης τίθεται ως εξής (Καρώνη & Οικονόμου, 2017) :

Έστω ένα μοντέλο με n παρατηρήσεις και με p επεξηγηματικές μεταβλητές. Το μοντέλο πολλαπλής παλινδρόμησης για την παραπάνω περίπτωση ορίζεται ως εξής:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i, \epsilon_i \sim N(0, \sigma^2).$$

όπου

- $x_{ij}, i = 1, ..n$ και $j = 1, ..., p$, οι τιμές για την i-οστη παρατήρηση των επεξηγηματικών μεταβλητών x_j
- $y_i, i = 1, ..., n$, οι τιμές της μεταβλητής απόκρισης για την i-οστη παρατήρηση
- $(\beta_0, \beta_1, \ldots, \beta_p)$ οι άγνωστες παράμετροι προς εκτίμηση και
- $\epsilon_i, i = 1, ..., n$ τα τυχαία σφάλματα που θεωρείται ότι ακολουθούν τις προϋποθέσεις του απλού γραμμικού μοντέλου.

Η εκτίμηση των συντελεστών $\boldsymbol{\beta}_i$γίνεται με την μέθοδο των ελαχίστων τετραγώνων (ε.τ), κατά την οποία γίνεται ελαχιστοποίηση του αθροίσματος τετραγώνων των υπολοίπων, δηλαδή της ποσότητας :

$$SSE = \sum_{i}^{n} (y_i - \boldsymbol{\beta}\mathbf{x_i})^2$$

όπου $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ και $\mathbf{x_i} = (1, x_{i1}, \ldots, x_{ip}), i = 1, 2, \ldots, n$

Οι παραπάνω εκτιμητές είναι οι συνηθέστεροι που χρησιμοποιούνται σε μια στατιστική ανάλυση, όταν δεν εμφανίζονται άλλα προβλήματα που να εμποδίζουν την χρήση τους, όπως για παράδειγμα η πολυσυγγραμικότητα, δηλαδή η ισχυρή γραμμική συσχέτιση μεταξύ των συμμεταβλητών. Όμως δεν είναι μόνο αυτή η παράμετρος που πρέπει να ληφθεί υπόψη για την επιλογή αυτής της μεθόδου για την εκτίμηση των συντελεστών. Οι εκτιμητές ελαχίστων τετραγώνων παρουσιάζουν αδυναμίες στην πρόβλεψη. Οι συντελεστές αυτοί έχουν μικρή μεροληψία (bias) όμως με αρκετά μεγάλη διασπορά γεγονός που τους κάνει λιγότερο ακριβείς. Ένας ακόμα λόγος που κάποιος μπορεί να μην επιλέξει αυτούς τους συντελεστές για την ανάλυση του, είναι η ερμηνεία των συντελεστών. Είναι αρκετά σύνηθες σε μια στατιστική ανάλυση να υπάρχουν πολλές συμμεταβλητές για ένα μοντέλο και από αυτές να πρέπει να επιλεχθούν αυτές οι οποίες χρειάζονται πραγματικά στο μοντέλο, να γίνει δηλαδή μια επιλογή μεταβλητών.

Για τους παραπάνω λόγους πολύ συχνά επιλέγονται κάποιες μέθοδοι **συρρίκνωσης** (shrinking) με τις οποίες κάποιοι από τους συντελεστές συρρικνώνονται προς το μηδέν ή και τίθενται ίσοι με το μηδέν.

Ως επί το πλείστον για την παραπάνω διαδικασία χρησιμοποιούνται **η μέθοδος επιλογής υποσυνόλου** (Subset Selection) καθώς και οι μέθοδοι **Ridge** και **Lasso**. Η μέθοδος επιλογής υποσυνόλου είναι πολλές φορές πολύ δαπανηρή υπολογιστικά, ειδικά σε περιπτώσεις όπου υπάρχουν πολλές υποψήφιες συμμεταβλητές. Επιπλέον είναι μια μέθοδος η οποία είναι εξαιρετικά μεταβλητή, πολύ μικρές αλλαγές στα δεδομένα μπορεί να φέρουν εντελώς διαφορετικά αποτελέσματα με αποτέλεσμα να μην υπάρχει καμία ακρίβεια στην προβλεπτική ικανότητα του μοντέλου. Η μέθοδος Ridge ή αλλιώς παλινδρόμηση κορυφογραμμής συρρικνώνει τους συντελεστές του μοντέλου χωρίς όμως να μπορεί να θέσει κανέναν ίσο με το μηδέν γεγονός που μπορεί να κάνει το μοντέλο να μην είναι εύκολα ερμηνεύσιμο. Τέλος η μέθοδος Lasso (Least Absolute Shrinkage and Selection Operator) φαίνεται να παρουσιάζει πλεονέκτημα συγκριτικά με τις άλλες δύο μεθόδους αφού μπορεί εκτός από το να συρρικνώνει προς το μηδέν κάποιους από τους συντελεστές να θέτει και κάποιους ίσους με το μηδέν, κάνοντας έτσι και μια επιλογή μεταβλητών και δημιουργώντας όσο το δυνατόν πιο απλά μοντέλα. Παρακάτω θα αναλυθούν οι δύο τελευταίες τεχνικές.

3.1 Παλινδρόμηση Ridge

Η παλινδρόμηση κορυφογραμμής ή Παλινδρόμηση Ridge όπως αναφέρθηκε χρησιμοποιείται σε περιπτώσεις που η εκτίμηση των συντελεστών με τη μέθοδο ελαχίστων τετραγώνων αποτυγχάνει. Μέσω αυτής της μεθόδου οι προς εκτίμηση εκτιμητές συρρικνώνονται προς το μηδέν μειώνοντας σημαντικά τη διασπορά και κατ΄ επέκταση βελτιώνοντας την προβλεπτική ικανότητα του μοντέλου. Το μοντέλο προτάθηκε από τους Hoerl και Kennard το 1970 στο paper τους, "Ridge Regression: Biased Estimation of Nonorthogonal Problems" (Hoerl & Kennard, 1970) και "Ridge Regression: Applications to Nonorthogonal Problems" (Hoerl & Kennard, 1970).

3.1.1 Το μοντέλο

Η τεχνική αυτή λειτουργεί με παρόμοιο τρόπο με αυτή των ελαχίστων τετραγώνων. Σε αυτή την περίπτωση όμως ελαχιστοποιείται διαφορετική ποσότητα (James & Witten & Hastie & Tibshirani, 2021). Ενώ στην περίπτωση των ελαχίστων τετραγώνων σκοπός της μεθόδου για την εκτίμηση των συντελεστών $\boldsymbol{\beta}$ είναι η ελαχιστοποίηση της ποσότητας του αθροίσματος τετραγώνων των υπολοίπων, RSS (Residual Sum of Squares), η οποία ορίζεται ως:

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

Για την παλινδρόμηση Ridge η ποσότητα που χρειάζεται να ελαχιστοποιηθεί για να εκτιμηθούν οι συντελεστές του μοντέλου $\hat{\boldsymbol{\beta}}^R$, δεν διαφέρει πολύ.
Συγκεκριμένα είναι η :

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2 \tag{3.1}$$

ή ισοδύναμα η ποσότητα:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2, \text{υπό τον περιορισμό}, \sum_{j=1}^{p}\beta_j^2 \leq s \tag{3.2}$$

όπου $\lambda \geq 0$ και ονομάζεται παράμετρος συντονισμού (tuning parameter) και $s \geq 0$. Οι παράμετροι αυτοί ελέγχουν το βαθμό στον οποίο θα συρρικνωθούν οι εκτιμήσεις των συντελεστών. Στόχος και στη μέθοδο αυτή είναι η καλή προσαρμογή των δεδομένων, μειώνοντας δραστικά το RSS. Επιπλέον ο όρος $\lambda\sum_{j=1}^{p}\beta_j^2$ καλείται ποινή συρρίκνωσης (shrinkage penalty), παίρνει χαμηλές τιμές όταν οι συντελεστές $\beta_1, \beta_2, ..., \beta_p$ είναι κοντά στο μηδέν και έτσι μπορεί να συρρικνώνει τους συντελεστές β_j κοντά στο μηδέν.

Ο συντελεστής συντονισμού λ έχει την ικανότητα να ελέγχει την επιρροή των δύο όρων που γίνεται προσπάθεια να ελαχιστοποιηθούν στους συντελεστές παλινδρόμησης. Όταν το $\lambda = 0$ δεν υπάρχει καμία ποινή και έτσι οι συντελεστές που προκύπτουν είναι στην πραγματικότητα οι συντελεστές ελαχίστων τετραγώνων. Καθώς όμως το $\lambda \to \infty$ τότε η επιρροή της ποινής θα μεγαλώνει και οι συντελεστές $\hat{\boldsymbol{\beta}}^{ridge}$ θα προσεγγίζουν το μηδέν. Για κάθε τιμή του λ θα προκύπτουν διαφορετικοί εκτιμητές των συντελεστών $\hat{\boldsymbol{\beta}}_{\lambda}^{ridge}$ και η σωστή επιλογή του λ είναι αυτή που θα οδηγήσει και σε καλύτερο μοντέλο.

Είναι σημαντικό να σημειωθεί πως η συρρίκνωση όπως αυτή γίνεται στη σχέση (3.1) αφορά τους συντελεστές $\beta_1, \beta_2, ..., \beta_p$ και όχι τον σταθερό όρο β_0. Το παραπάνω συμβαίνει καθώς σκοπός της τεχνικής είναι η συρρίκνωση της εκτιμώμενης συσχέτισης που κάθε συμμεταβλητή έχει με την μεταβλητή απόκρισης y. Ο σταθερός όρος β_0 όμως δεν κρίνεται αναγκαίο να συρρικνωθεί καθώς αποτελεί ένα μέτρο της μέσης τιμής της μεταβλητής απόκρισης όταν όλες οι επεξηγηματικές μεταβλητές (ή συμμεταβλητές) είναι μηδέν. Με την υπόθεση ότι οι τιμές του πίνακα σχεδιασμού $\mathbf{X}$ έχουν κεντραριστεί ώστε να έχουν μέση τιμή μηδέν, τότε ο εκτιμητής του σταθερού όρου β_0 θα είναι της μορφής: $\beta_0 = \bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$

Σημειώνεται ότι $\mathbf{X}$ ο πίνακας σχεδιασμού του μοντέλου, ορίζεται ως εξής :

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}$$

3.1.2 Εκτίμηση Συντελεστών

Ελαχιστοποιώντας την (3.1) οι συντελεστές των συμμεταβλητών δίνονται από την σχέση :

$$\hat{\boldsymbol{\beta}}_{\lambda}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \tag{3.3}$$

όπου $\mathbf{X}$ ο πίνακας σχεδιασμού του μοντέλου, $\mathbf{y}$ ένα διάνυσμα με τις τιμές της μεταβλητής απόκρισης και $\mathbf{I}$ ο μοναδιαίος $n \times n$.

Οι συντελεστές αυτοί έχουν κατά κανόνα μικρότερα σφάλματα πρόβλεψης από τις εκτιμήτριες ελαχίστων τετραγώνων. Πρόκειται για μεροληπτικές εκτιμήτριες με : πίνακα συνδιασποράς :

$$Var(\hat{\boldsymbol{\beta}}_{\lambda}^{ridge}) = \sigma^2(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}$$

και μέσο τετραγωνικό σφάλμα :

$$\begin{aligned} MSE &= Var(\hat{\boldsymbol{\beta}}_{\lambda}^{ridge}) + bias(\hat{\boldsymbol{\beta}}_{\lambda}^{ridge}) \\ &= \sigma^2 Tr(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1} + \lambda^2\boldsymbol{\beta}^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-2}\boldsymbol{\beta} \end{aligned} \tag{3.4}$$

Η σωστή επιλογή του λ είναι αυτή που θα οδηγήσει σε μεγαλύτερη μείωση της διασποράς συγκριτικά με την αύξηση της μεροληψίας που θα συμβεί όπως αναφέρθηκε και προηγουμένως. Όσο το λ μεγαλώνει και η συρρίκνωση των εκτιμητών συμβάλει σε σημαντική μείωση της διασποράς των προβλέψεων 'θυσιάζοντας' μια μικρή αύξηση της μεροληψίας. Οι Hoerl και Kennard έδειξαν πως αυτό συμβαίνει όταν ικανοποιείται ο περιορισμός :

$$MSE(\hat{\boldsymbol{\beta}}_{\lambda}^{ridge}) < Var(\hat{\boldsymbol{\beta}})$$

Ενδιαφερόν επίσης έχει το γεγόνος πως οι εκτιμητές ελαχίστων τετραγώνων είναι ισοδύναμης κλίμακας, δηλαδή ο πολλαπλασιασμός της συμμεταβλητής X_j με μια σταθερά c οδηγεί σε μια αλλαγή κλίμακας των ε.ε.δ κατά έναν παράγοντα $\frac{1}{c}$. Ανεξαρτήτως λοιπόν, με τον τρόπο που αλλάζει η j-οστή συμμεταβλητή, ο παράγοντας $X_j\hat{\beta}_j$ θα παραμείνει ίδιος. Κάτι τέτοιο όμως δεν ισχύει για τους εκτιμητές της παλινδρόμησης Ridge, για τους οποίους, όταν μια συμμεταβλητή πολλαπλασιάζεται με κάποια σταθερά μπορεί να αλλάξουν σημαντικά. Οι παράγοντες $X_j\hat{\beta}_{j,\lambda}^{ridge}$ επηρεάζονται από την κλίμακα της j-οστής συμμεταβλητής ή ακόμα και από την κλίμακα των υπολοίπων συμμεταβλητών. Είναι προτιμότερο λοιπόν, πριν κάποιος προβεί σε μία εκτίμηση των συντελεστών των συμμεταβλητών με την μέθοδο Ridge να κάνει μια κανονικοποίηση στις συμμεταβλητές, έτσι ώστε να βρίσκονται όλες στην ίδια κλίμακα και να έχουν όλες τυπική απόκλιση ίση με τη μονάδα. Με τον τρόπο αυτό η προσαρμογή του μοντέλου δεν θα εξαρτάται από την κλίμακα των συμμεταβλητών. Η κανονικοποίηση μπορεί να χρησιμοποιώντας την σχέση :

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x_j})^2}}$$

όπου ο παρανομαστής είναι η εκτίμηση της τυπικής απόκλισης για την j-οστή συμμεταβλητή.

3.1.3 Επιλογή της Παλινδρόμησης Ridge

Η παλινδρόμηση Ridge προσφέρει κάποια πλεονεκτήματα έναντι των εκτιμητριών ελαχίστων τετραγώνων όπως αναφέρθηκαν στην προηγούμενη ανάλυση και συγκεντρώνονται στα σημεία που θα αναφερθούν παρακάτω.

Όταν το λ είναι μηδέν οι εκτιμήτριες της τεχνικής Ridge συμπίπτουν με αυτές των ε.ε.τ. και έτσι παρατηρείται μεγάλη διασπορά και καθόλου όμως μεροληψία (James & Witten & Hastie & Tibshirani, 2021). Καθώς το λ μεγαλώνει, η προσαρμογή του μοντέλου Ridge μειώνεται και οι εκτιμήτριες των συντελεστών συρρικνώνονται οδηγώντας έτσι σε μείωση της διασποράς στις προβλέψεις αλλά και σε αύξηση της μεροληψίας.

Γενικότερα όταν η σχέση μεταξύ των συμμεταβλητών και της μεταβλητής απόκρισης είναι σχετικά γραμμική, οι εκτιμήτριες ελαχίστων τετραγώνων έχουν πολύ μεγάλη διασπορά, γεγονός που σημαίνει πως μια μικρή αλλαγή στα δεδομένα μπορεί να επιφέρει μεγάλες αλλαγές στις εκτιμήσεις των συντελεστών. Ειδικά όταν οι συμμεταβλητές p είναι όσες οι παρατηρήσεις n τότε οι ε.ε.τ. θα είναι εξαιρετικά μεταβλητές και σε περιπτώσεις όπου $p > n$, οι ε.ε.τ. δεν θα δίνουν καν μοναδική λύση. Σε περιπτώσεις όπως αυτές και γενικά όταν οι ε.ε.τ. έχουν υψηλή διασπορά, οι εκτιμητές της παλινδρόμησης Ridge φαίνεται να είναι μια πολύ καλύτερη εναλλακτική από αυτή των ελαχίστων τετραγώνων.

Ένα ακόμα πλεονέκτημα της τεχνικής αυτή είναι πως υπολογιστικά είναι πολύ πιο συμφέρουσα, καθώς για κάθε τιμή του λ προσαρμόζεται ένα μόνο μοντέλο. Μπορεί να αποδειχθεί ότι οι υπολογισμοί για την λύση τις (3.1) για όλες τις τιμές του λ είναι υπολογιστικά σχεδόν οι ίδιοι με την προσαρμογή ενός μοντέλου με την χρήση των ελαχίστων τετραγώνων.

3.2 Παλινδρόμηση Lasso

Παρόλο που η τεχνική Ridge παρουσιάζει πλεονεκτήματα σε σχέση με τις κλασσικές εκτιμήτριες ελαχίστων τετραγώνων, δεν επιλέγει ένα υποσύνολο συμμεταβλητών από τις p αλλά τις συμπεριλαμβάνει όλες στο μοντέλο, γεγονός που πολύ συχνά δεν είναι επιθυμητό. Η ποινή $\lambda \sum_{i=1}^{n} \beta_j^2$ ναι μεν συρρικνώνει τις εκτιμήτριες των συντελεστών προς το μηδέν, δεν τις θέτει όμως ίσες με το μηδέν. Αυτό μπορεί να δημιουργήσει πρόβλημα στην ερμηνεία του μοντέλου ειδικά αν ο αριθμός p είναι μεγάλος. Είναι προτιμότερο το μοντέλο να συμπεριλαμβάνει μόνο τις στατιστικά σημαντικές μεταβλητές έτσι ώστε να προκύψει ένα μοντέλο όσο το δυνατόν πιο φειδωλό.

Το 1995 ο Tibshirani, προτείνει μια καινούργια τεχνική η οποία καταφέρνει να κάνει και επιλογή καλύτερου μοντέλου, η μέθοδος ονομάζεται LASSO (Least Absolute Shrinkage and Selection Operator) (Tibshirani, 1996). Η μέθοδος αυτή συγκεντρώνει τα πλεονεκτήματα της παλινδρόμησης Ridge και της μεθόδου επιλογής υποσυνόλου καθώς μπορεί να συρρικνώνει τους συντελεστές του μοντέλου αλλά και να θέτει κάποιους ακριβώς ίσους με το μηδέν και έτσι κάνει και επιλογή μεταβλητών. Είναι πολύ χρήσιμη σε προβλήματα μεγάλων διαστάσεων και προσφέρει μοντέλα τα οποία είναι εύκολα ερμηνεύσιμα.

Οι μέθοδοι Lasso και Ridge αποτελούν ειδικές περιπτώσεις της παλινδρόμησης Bridge, που είχε προταθεί για πρώτη φορά από τους Frank και Friedman το 1993.

3.2.1 Το μοντέλο

Το μοντέλο και πάλι έχει ως εξής : Έστω, $x_{ij}, y_i, i = 1, 2, ..., n$ και $j = 1, ..., p$, το σύνολο των δεδομένων, όπου με $x_i = (x_{i1}, ..., x_{ip})$ συμβολίζονται οι επεξηγηματικές μεταβλητές (ή συμμεταβλητές) και y_i οι αντίστοιχες αποκρίσεις. Όπως και στην παλινδρόμηση Ridge οι μεταβλητές x_{ij} θεωρούνται κανονικοποιημένες ώστε να ισχύει: $\mu = \sum_{i=1}^{n} \frac{x_{ij}}{n} = 0$ και $\sigma^2 = \sum_{i=1}^{n} \frac{x_{ij}^2}{n}$.

Οι εκτιμήτριες lasso $\hat{\boldsymbol{\beta}}_\lambda^L$ προκύπτουν με την ελαχιστοποίηση της ποσότητας:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| = RSS + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3.5}$$

ή ισοδύναμα της:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2, \text{υπό τον περιορισμό}, \sum_{j=1}^{p} |\beta_j| \leq s \tag{3.6}$$

όπου και πάλι $\lambda \geq 0$ και ονομάζεται παράμετρος συντονισμού (tuning parameter) και $s \geq 0$. Ο παράγοντας β_0 μπορεί να παραληφθεί και αυτό γιατί για όλες τις τιμές

που μπορεί να πάρει το s, η λύση για το β_0 είναι : $\hat{\beta}_0 = \bar{y}$ και έτσι χωρίς βλάβη της γενικότητας γίνεται η υπόθεση ότι $\bar{y} = 0$. Αυτό έχει ως αποτέλεσμα να γίνει παράλειψη του β_0 από τις σχέσεις (3.5) και (3.6).

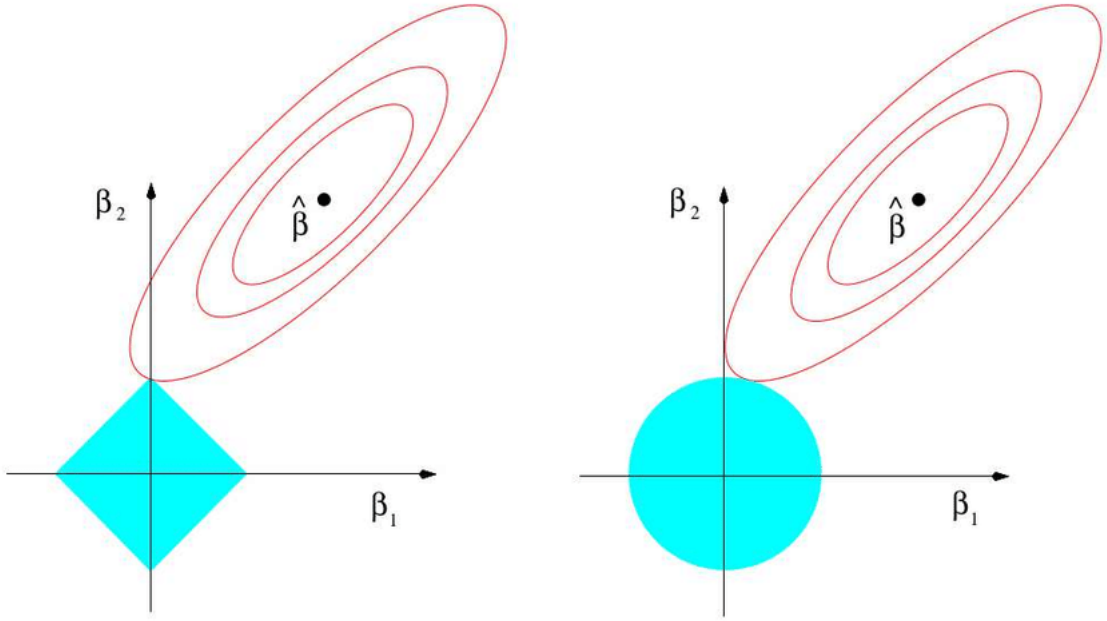
Και σε αυτή τη μέθοδο οι παράμετροι λ και s ελέγχουν τον βαθμό συρρίκνωσης των συντελεστών.

Όπως εύκολα μπορεί κανείς να παρατηρήσει οι σχέσεις (3.1) και (3.5) για την Ridge και την Lasso αντίστοιχα, μοιάζουν ιδιαίτερα. Διαφοροποιούνται στον παράγοντα της ποινής. Για την Lasso ο όρος $\lambda \sum_{j=1}^{p} |\beta_j|$ χρησιμοποιεί την ℓ_1 νόρμα, σε αντίθεση με την Ridge που ο όρος της ποινής της παραπέμπει στην ℓ_2 νόρμα. Η τεχνική της Lasso μπορεί να θέσει τις εκτιμήτριες των συντελεστών ακριβώς ίσες με το μηδέν λόγω αυτού του παράγοντα της ποινής, όταν η τιμή του λ είναι επαρκώς μεγάλη. Με τον τρόπο αυτό η Lasso καταφέρνει να κάνει επιλογή μεταβλητών, όπως και η μέθοδος επιλογής υποσυνόλου, δημιουργώντας έτσι μοντέλα εύκολα ερμηνεύσιμα. Γενικά, όταν το $\lambda = 0$ η Lasso δίνει τις εκτιμήτριες των ελαχίστων τετραγώνων (όπως και η Ridge), όταν το $\lambda = \infty$ δίνει το μηδενικό μοντέλο (όπως και η Ridge), ενώ στις ενδιάμεσες τιμές παράγει ένα μοντέλο με κάποιες από τις συμμεταβλητές σε αντίθεση με την Ridge που θα παράξει ένα μοντέλο με όλες τις συμμεταβλητές όπου το λ θα ελέγχει απλώς το μέγεθος συρρίκνωσης των συντελεστών.

3.2.2 Η γεωμετρία της Lasso

Ο λόγος που η Lasso θέτει συντελεστές ίσους με το μηδέν ενώ η Ridge όχι, όπως αναφέρθηκε και προηγουμένως έγκειται στους όρους ποινής της κάθεμιας. Αυτό φαίνεται καλύτερα στις σχέσεις (3.2) και (3.6) και στους περιορισμούς $\sum_{j=1}^{p} \beta_j^2 \leq s$ και $\sum_{j=1}^{p} |\beta_j| \leq s$ αντίστοιχα. Υπάρχουν τιμές του λ για τις οποίες θα προκύψουν οι ίδιες εκτιμήτριες Ridge και Lasso (James & Witten & Hastie & Tibshirani, 2021).

Έστω ένα μοντέλο με δύο επεξηγηματικές μεταβλητές, δηλαδή $p = 2$. Τότε για την (3.2) οι εκτιμήτριες των συντελεστών Ridge με το μικρότερο RSS βρίσκονται εντός του κύκλου που ορίζεται από τη σχέση : $\beta_1^2 + \beta_2^2 \leq t$ ένω οι εκτιμήτριες των συντελεστών Lasso εμπεριέχονται μέσα στο σχήμα διαμαντίου που ορίζεται από τη σχέση $|\beta_1| + |\beta_2| \leq t$.



Σχήμα 3.1: Εικόνα εκτιμήσεων για: (α) Lasso και (β) Ridge

Οι δύο περιορισμοί ουσιαστικά ψάχνουν την μικρότερη τιμή του RSS, υπό την προϋπόθεση ότι υπάρχει ένας περιοσμός t για το πόσο μεγάλος μπορεί να γίνει ο παράγοντας $\sum_{j=1}^{p} \beta_j^2$ και $\sum_{j=1}^{p} |\beta_j|$ αντίστοιχα. Όταν το t είναι πολύ μεγάλο ο περιορισμός δεν είναι πολύ αυστηρός, οι συντελεστές μπορεί να γίνουν αρκετά μεγάλοι και για την ακρίβεια αν το t είναι αρκετά μεγάλο τότε απλώς έχουμε τις εκτιμήτριες ελαχίστων τετραγώνων. Όταν το t είναι μικρό θα πρέπει $\sum_{j=1}^{p} \beta_j^2$ και $\sum_{j=1}^{p} |\beta_j|$ να είναι αρκετά μικρά για να μην ξεπεράσουν τον περιορισμό.

Όπως φαίνεται στο σχήμα 3.1 το μπλε διαμάντι και ο μπλε κύκλος αποτελούν τους περιορισμούς για Lasso και Ridge αντίστοιχα, ενώ το $\hat{\boldsymbol{\beta}}$ είναι η λύση των ελαχίστων τετραγώνων. Αν το t είναι αρκετά μεγάλο (αντίστοιχα μεγάλο με το $\lambda = 0$), τότε οι περιορισμένες περιοχές θα περιέχουν το $\hat{\boldsymbol{\beta}}$ και οι μέθοδοι Lasso και Ridge θα δίνουν τις ίδιες εκτιμήτριες με τα ελάχιστα τετράγωνα.

Οι ελλείψεις που έχουν ως κέντρο το $\hat{\boldsymbol{\beta}}$ απεικονίζουν ισοϋψείς καμπύλες, δηλαδή τα σημεία κάθε έλλειψης έχουν το ίδιο RSS. Όσο οι ελλείψεις απομακρύνονται από το $\hat{\boldsymbol{\beta}}$ το RSS μεγαλώνει. Οι εξισώσεις (3.2) και (3.6) υποδεικνύουν ότι οι εκτιμήτριες Lasso και Ridge προκύπτουν από το πρώτο σημείο όπου οι ελλείψεις συναντούν την περιοχή περιορισμού. Οι περιοχή περιορισμού για την Ridge είναι κυκλική και έτσι το παραπάνω δεν θα συμβεί πάνω σε έναν άξονα και αυτός είναι ο λόγος που δεν μπορεί κάποιος Ridge συντελεστής να είναι ακριβώς μηδέν. Αντιθέτως η περιοχή περιορισμού για την Lasso έχει γωνίες στους άξονες και για το λόγο αυτό συχνά η τομή των ελλείψεων με την περιοχή περιορισμού γίνεται πάνω στους άξονες. Παρομοίως συμπεριφέρεται η

μέθοδος και για μεγαλύτερες διαστάσεις.

3.3 Σύγκριση Ridge και Lasso

Η Lasso έχει ένα ξεκάθαρο πλεονέκτημα έναντι της Ridge και αυτό είναι η επιλογή μεταβλητών που μπορεί να κάνει, γεγονός που την κάνει να παράγει μοντέλα τα οποία είναι εύκολα να ερμηνευτούν. Όμως για να καταλήξει κανείς σε μία από τις δύο τεχνικές θα πρέπει να αναλογιστεί και ποία από τις δύο μπορεί να έχει καλύτερη προβλεπτική ικανότητα.

Για την σύγκριση των δύο τεχνικών θα πρέπει κάποιος να συμβουλευτεί και το μέσο τετραγωνικό σφάλμα (MSE). Γενικά η Lasso έχει παρόμοια αποτελέσματα με τα αυτά της Ridge, δηλαδή όσο το λ μεγαλώνει η διασπορά μειώνεται 'με αντάλλαγμα' μια μικρή αύξηση στην μεροληψία. Όμως κανείς δεν μπορεί να υποθέσει με σιγουριά πως η μία μέθοδος είναι καλύτερη από την άλλη. Θα μπορούσε να γίνει η υπόθεση, ότι σε προβλήματα που λίγες συμμεταβλητές είναι διάφορες του μηδενός και οι περισσότερες είναι είτε μηδέν είτε πολύ κοντά σε αυτό, η Lasso θα ήταν καλύτερη επιλογή. Αντιστοίχως η Ridge θα απέδιδε καλύτερα σε περιπτώσεις με πολλές μεταβλητές του ίδιου περίπου μεγέθους. Είναι γνωστό όμως πως στην πλειοψηφία των προβλημάτων ο αριθμός των συμμεταβλητών που έχουν σχέση με την μεταβλητή απόκρισης δεν είναι ποτέ γνωστός a priori. Μπορούν να χρησιμοποιηθούν μέθοδοι όπως η Cross-Validation για την διερεύνηση ποιας μεθόδου πρέπει να ακολουθηθεί.

Σημειώνεται επίσης πως και οι δύο τεχνικές είναι υπολογιστικά αποδοτικές και 'κοστίζουν' όσο η εφαρμογή ενός μοντέλου ελαχίστων τετραγώνων.

3.4 Μια ειδική περίπτωση για Ridge και Lasso

Έστω μια περίπτωση παλινδρόμησης στην οποία ο αριθμός των παρατηρήσεων n είναι όσος και ο αριθμός των συμμεταβλητών p και ο πίνακας σχεδιασμού $\mathbf{X}$ είναι διαγώνιος με άσσους στην διαγώνιο και μηδενικά όλα τα υπόλοιπα στοιχεία του. Επιπλέον γίνεται η υπόθεση ότι η παλινδρόμηση γίνεται χωρίς σταθερό όρο β_0. Έτσι για την εύρεση των συντελεστών $\beta_1, ..., \beta_p$ ελαχιστοποιούνται οι ποσότητες :

- Για τη μέθοδο ελαχίστων τετραγώνων

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 \tag{3.7}$$

Όπου η λύση είναι:

$$\hat{\beta}_j = y_j$$

- Για τη μέθοδο Ridge

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \tag{3.8}$$

Όπου οι εκτιμήτριες είναι:

$$\hat{\beta_j^R} = \frac{y_j}{(1+\lambda)}$$

- Για τη μέθοδο Lasso

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \tag{3.9}$$

Όπου οι εκτιμήτριες είναι:

$$\hat{\beta}_j = \begin{cases} y_j - \frac{\lambda}{2}, & \text{αν } y_j > \frac{\lambda}{2} \\ y_j + \frac{\lambda}{2}, & \text{αν } y_j < -\frac{\lambda}{2} \\ 0, & \text{αν } |y_j| \leq \frac{\lambda}{2} \end{cases}$$

Παρατηρείται εύκολα ότι στην προκειμένη περίπτωση, η συρρίκνωση που κάνουν οι δύο τεχνικές είναι εντελώς διαφορετικές. Για την Ridge οι εκτιμήτριες των συντελεστών συρρικνώνονται με την ίδια ακριβώς αναλογία ενώ για την Lasso οι εκτιμήτριες συρρικνώνονται κατά μια ποσότητα ίση με $\frac{\lambda}{2}$ ενώ αυτές οι εκτιμήτριες που η απόλυτη τιμή τους είναι μικρότερη από $\frac{\lambda}{2}$ τίθενται ίσες με το μηδέν. Σε γενικότερες περιπτώσεις από αυτές του πίνακα $\mathbf{X}$ τα παραπάνω αλλάζουν όμως η κεντρική ιδέα παραμένει. Η Ridge συρρικνώνει κάθε διάσταση με τη ίδια αναλογία ενώ η Lasso συρρικνώνει περισσότερο ή λιγότερο όλες τις εκτιμήτριες προς το μηδέν με μία παρόμοια ποσότητα, ενώ οι πολύ 'μικρές' εκτιμήτριες τίθενται μηδέν.

3.5 Cross-Validation

Η μέθοδος Cross-Validation, είναι η πιο διαδομένη μέθοδος αξιολόγησης της προβλεπτικής ικανότητας ενός μοντέλου (van Houwelingen & Sauerbrei, 2013).

Η γενική ιδέα της μεθόδου στηρίζεται στο διαχωρισμό των δεδομένων σε δύο σύνολα, ένα με το οποίο προσαρμόζεται το μοντέλο και ονομάζεται training set και ένα με το οποίο αξιολογείται η προβλεπτική ικανότητα του μοντέλου και ονομάζεται test set. Ο διαχωρισμός των δεδομένων μπορεί να γίνει με τους τρεις παρακάτω τρόπους (James & Witten & Hastie & Tibshirani, 2021):

- Χωρίζοντας τα δεδομένα τυχαία σε δύο σύνολα περίπου ίδιου μεγέθους (Validation Set Approach).

- Χωρίζοντας τα δεδομένα σε δύο σύνολα, αφήνοντας όμως στο test set μια μόνο παρατήρηση (Leave one out Cross-Validation) και τις υπόλοιπες $n - 1$ παρατηρήσεις χρησιμοποιούνται για την προσαρμογή του μοντέλου. Η διαδικασία επαναλαμβάνεται συνολικά n φορές και κάθε φορά υπολογίζεται το μέσο τετραγωνικό σφάλμα (MSE) κάθε μοντέλου. Ο μέσος όρος των MSE αποτελεί την εκτιμήτρια CV του MSE του training set.

- Χωρίζοντας τα δεδομένα σε k 'φακέλους' (folds) περίπου ίδιου μεγέθους. Ο πρώτος φάκελος χρησιμοποιείται ως test set και οι υπόλοιποι $k - 1$ χρησιμοποιούνται για την προσαρμογή του μοντέλου και υπολογίζεται το μέσο τετραγωνικό σφάλμα (MSE) του μοντέλου. Η διαδικασία επαναλαμβάνεται συνολικά k φορές, διαλέγοντας κάθε φορά ένα διαφορετικό φάκελο για test set και υπολογίζοντας το MSE για κάθε μοντέλο. Ο μέσος όρος των MSE αποτελεί την εκτιμήτρια CV του MSE του training set.

3.5.1 Η επιλογή του λ

Όπως αναφέρθηκε παραπάνω για τις μεθόδους συρρίκνωσης Ridge και Lasso η επιλογή του λ καθορίζει σε μεγάλο βαθμό την αποτελεσματικότητα της εκάστοτε μεθόδου. Για την καλύτερη τιμή του λ χρησιμοποιείται η μέθοδος Cross-Validation.

Η διαδικασία επιλογής του λ έχει ως εξής: από ένα πλέγμα τιμών για το λ υπολογίζεται το CV σφάλμα για κάθε τιμή με τον τρόπο που περιγράφηκε παραπάνω (James & Witten & Hastie & Tibshirani, 2021). Στη συνέχεια επιλέγεται το λ εκείνο που δίνει την μικρότερη τιμή σφάλματος. Τέλος το μοντέλο προσαρμόζεται ξανά με όλες τις παρατηρήσεις αυτή τη φορά και με την επιλεγμένη τιμή λ.

3.6 Μέθοδοι συρρίκνωσης στο μοντέλο του Cox

Με την ίδια λογική θα γίνει η προσπάθεια να προσαρμοστούν οι τεχνικές που αναφέρθηκαν παραπάνω στο μοντέλο του Cox (James & Witten & Hastie & Tibshirani, 2021). Η ποσότητα που θα γίνει προσπάθεια να ελαχιστοποιηθεί αυτή τη φορά είναι μια ποινικοποιημένη μορφή της αρνητικής λογαριθμοποιημένης πιθανοφάνειας η οποία διαμορφώνεται ως εξής:

$$-log \prod_{i:\delta_i=1} \frac{e^{\sum_{j=1}^{p} x_{ij}\beta_j}}{\sum_{i':y_{i'}\geq y_i} e^{\sum x_{i'j}\beta_j}} + \lambda P(\beta)$$

Το $P(\beta)$ διαμορφώνεται ως $P(\beta) = \sum_{j=1}^{p} \beta_j^2$, όταν πρόκειται για ποινή Ridge ενώ ως $P(\beta) = \sum_{j=1}^{p} |\beta_j|$, όταν πρόκειται για ποινή Lasso.

Συνήθως το λ είναι μια μη-αρνητική παράμετρος συντονισμού και κάνουμε την ελαχιστοποίηση για ένα εύρος τιμών του λ. Όταν το $\lambda = 0$ η παραπάνω σχέση ισοδυναμεί με τη συνηθισμένη στο μοντέλο του Cox. Όταν όμως το λ είναι θετικό τότε οι εκτιμήτριες θα είναι συρρικνωμένες. Για μεγάλες τιμές του λ, η μέθοδος Ridge θα δώσει εκτιμήτριες με τιμές πολύ κοντά στο μηδέν ενώ η Lasso θα δώσει συρρικνωμένες εκτιμήτριες με κάποιες από αυτές να είναι μηδέν.

Έχει παρατηρηθεί πως η μέθοδος Lasso αποδίδει πολύ καλύτερα ως μοντέλο από άλλες μεθόδους όπως η subset selection ή η stepwise selection (Tibshirani, 1997). Σημειώνεται επίσης πως είναι αρκετά χρήσιμη η κανονικοποίηση των συμμεταβλητών, έτσι ώστε η ποινή να είναι ίδια για όλες. Επιπλέον αν είναι γνωστό εξ΄ αρχής πως μια συμμεταβλητή είναι απαραίτητη στο μοντέλο, όπως για παράδειγμα ένας παράγοντας που επηρεάζει μια θεραπεία και ως εκ τούτου δεν είναι επιθυμητή η συρρίκνωση της, μπορεί να παραληφθεί από τον όρο της ποινής.

3.6.1 The elastic net

Αξίζει να αναφερθεί πως το 2005 , οι Zou και Hastie, παρουσίασαν τη μέθοδο elastic net (Zou & Hastie, 2005). Πρόκειται για έναν συνδυασμό των δύο ποινών που αναφέρθηκαν. Ο όρος της ποινής $\lambda P(\beta)$ τώρα ορίζεται ως $P(\beta) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$.

Η μέθοδος εκτελεί επιλογή μεταβλητών και εκτίμηση των συντελεστών των επεξηγηματικών μεταβλητών όπως και η lasso όμως επιπρόσθετα με τον δεύτερο όρο ποινής, δίνει έμφαση σε παραπάνω μεταβλητές. Με τον τρόπο αυτό η elastic net κατα κύριο λόγο επιλέγει παραπάνω μεταβλητές από την lasso.

Κεφάλαιο 4

Δένδρα Παλινδρόμησης

4.1 Εισαγωγή

Στην στατιστική αλλά και σε εφαρμογές της σε άλλες επιστήμες είναι πολύ σύνηθες από ένα σύνολο δεδομένων να χρειαστεί να γίνει πρόβλεψη για κάποια ποσότητα. Στις περιπτώσεις αυτές ως επί το πλείστον γίνεται χρήση κάποιου μοντέλου παλινδρόμησης. Πολλές φορές όμως δεν είναι δυνατή ή δεν είναι αρκετά ικανοποιητική αυτή η προσέγγιση.

Έτσι λοιπόν μια διαφορετική μέθοδος από αυτές που έχουν αναλυθεί μέχρι στιγμής προτείνεται για την πρόβλεψη ενός αποτελέσματος. Η μέθοδος αυτή χωρίζει τον χώρο των επεξηγηματικών μεταβλητών σε μικρότερες περιοχές (James & Witten & Hastie & Tibshirani, 2021). Ο τρόπος με τον οποίο μοντελοποιείται η μέθοδος βασίζεται σε μια ομάδα μεθόδων που χαρακτηρίζονται ως βασισμένες στη μοντελοποίηση με τη μορφή των δένδρων. Η διαδικασία με την οποία γίνεται ο διαχωρισμός του χώρου, μπορεί να οπτικοποιηθεί σαν ένα δένδρο, το οποίο ξεκινώντας από την 'ρίζα' του (κόμβο αρχής ή κορυφή), χωρίζεται σε δύο 'κλαδιά'('παιδιά' ή κόμβους), τα οποία με την σειρά τους χωρίζονται με τον ίδιο τρόπο μέχρι να τελειώσει η διαμέριση.

Η προσέγγιση που θα αναλυθεί λοιπόν, σε αυτό το κεφάλαιο, για την πρόβλεψη ενός αποτελέσματος σε ένα σύνολο δεδομένων, είναι τα δένδρα παλινδρόμησης (regression trees). Είναι επί τοις ουσίας, τα ευρέως διαδεδομένα δένδρα απόφασης (decision trees), σε περιπτώσεις όμως που η πρόβλεψη που απαιτείται να γίνει, είναι ποσοτική μεταβλητή.

Τα regression trees χρησιμοποιούνται συχνά, όταν δημιουργούνται προβλήματα στην παλινδρόμηση όπως, όταν στο σύνολο δεδομένων υπάρχουν υπάρχουν πολλές επεξηγηματικές μεταβλητές οι οποίες αλληλεπιδρούν με περίπλοκους και μη γραμμικούς τρόπους. Επιπλέον με τα regression trees παράγονται αποτελέσματα τα οποία είναι πολύ εύκολα να ερμηνευτούν, ακόμα και από ανθρώπους που δεν έχουν κανένα μαθηματικό υπόβαθρο και επιπλέον τα αποτελέσματα μπορούν να οπτικοποιηθούν σε κάθε βήμα της διαδικασίας.

Από το 1963, με την δουλειά τους οι Morgan και Sonquist, παρουσίασαν μια διαφορετική ομάδα μεθόδων παλινδρόμησης οι οποίες πλέον είναι γνωστές ως δένδρα (trees) ή recursive partitioning (Hothron & Hornik & Zeileis, 2006). Από τότε πάρα πολλοί επιστήμονες έχουν ασχοληθεί με την συγκεκριμένη μέθοδο προσπαθώντας να την βελτιώσουν με διαφορετικές προσεγγίσεις. Βασικός πυρήνας όλων των προσεγγίσεων είναι ένας αλγόριθμος δύο βημάτων, πρώτα 'χωρίζοντας' τις παρατηρήσεις με αναδρομικό τρόπο και στη συνέχεια προσαρμόζοντας ένα σταθερό μοντέλο στο αποτέλεσμα κάθε διαχωρισμού. Αξιοσημείωτες υλοποιήσεις του παραπάνω αλγορίθμου είναι αυτές ο CART των Breiman, Friedman, Olshen και Stone το 1984 και ο C4.5 του Quinlan το 1993. Αμφότεροι πραγματοποιούν εξαντλητική αναζήτηση στα πιθανά χωρίσματα μεγιστοποιώντας κάποιο μέτρο πληροφορίας για την μη-καθαρότητα (impurity) κόμβου, επιλέγοντας τη μεταβλητή που θα κάνει το βέλτιστο χώρισμα. Με τον τρόπο αυτό όμως δημιουργούνται προβλήματα overfitting καθώς και μεροληψίας για μεταβλητές που οδηγούν σε πολλά χωρίσματα.

Η μέθοδος που θα αναλυθεί και θα χρησιμοποιηθεί στην παρούσα εργασία θα κάνει χρήση αμερόληπτης recurvise partitioning που στηρίζεται σε 'υπό όρους έλεγους υποθέσεων' και μπορεί να εφαρμοστεί σε όλα τα είδη δεδομένων, όπως και τα αποκομμένα που αποτελούν αντικείμενο ενδιαφέροντος της εργασίας, έτσι ώστε να δημιουργήσει δένδρα επιβίωσης (Zeileis & Hothron & Hornik, 2008).

4.2 Ανάπτυξη μεθόδου για τη δημιουργία δένδρων επιβίωσης

Το πρόβλημα της πρόβλεψης αφορά και πάλι μοντέλα παλινδρόμησης για μια μεταβλητή απόκρισης $\boldsymbol{Y}$, δεδομένου ότι στο πρόβλημα είναι γνωστές m επεξηγηματικές μεταβλητές. Η μεταβλητή απόκρισης $\boldsymbol{Y}$ μπορεί να είναι και αυτή m διαστάσεων. Επιπλέον οι μεταβλητές $\boldsymbol{Y}$ και $\boldsymbol{X} = (X_1, ...X_m)$ μπορούν να έχουν μετρηθεί σε οποιαδήποτε κλίμακα. Γίνεται η υπόθεση πως η υπό συνθήκη κατανομή $D(\boldsymbol{Y}|\boldsymbol{X})$ της μεταβλητής απόκρισης $\boldsymbol{Y}$ δεδομένου των $\boldsymbol{X}$ συμμεταβλητών εξαρτάται από μία συνάρτηση f των συμμεταβλητών:

$$D(\boldsymbol{Y}|\boldsymbol{X}) = D(\boldsymbol{Y}|(X_1, ...X_m) = D(\boldsymbol{Y}|f(X_1, ...X_m)),$$

όπου θα δημιουργηθούν σχέσεις παλινδρόμησης που έχουν βασιστεί στη διαμέριση του χώρου των επεξηγηματικών μεταβλητών. Με τον τρόπο αυτό προκύπτουν r ξένα μεταξύ τους 'κελιά' διαμερίσεων $B_1, B_2, ..., B_r$, έτσι ώστε όλος ο χώρος $\mathcal{X}$ των συμμεταβλητών να είναι: $\mathcal{X} = \cup_{k=1}^{r} B_k$. Σ'ένα τυχαίο δείγμα n παρατηρήσεων οι οποίες είναι ανεξάρτητες και ισόνομες, όπου από αυτές ενδέχεται X_{ji} να λείπουν αποτελούν

ένα learning sample (δείγμα μάθησης):

$$\mathcal{L}_n = (\boldsymbol{Y}, X_{1i}, ..., X_{mi}); i = 1, ..., n$$

και βάσει αυτού προσαρμόζεται η σχέση παλινδρόμησης.

Στη συνέχεια θα δοθεί η γενική μορφή του αλγορίθμου του recursive binary partitioning κάνοντας χρήση μη αρνητικών ακεραίων βαρών $\boldsymbol{w} = (w_1, ...w_n)$. Κάθε κόμβος του δένδρου που θα σχηματιστεί από την διαμέριση του χώρου, αναπαριστάται από ένα διάνυσμα βαρών που έχει μη μηδενικά στοιχεία μόνο εαν οι αντίστοιχες παρατηρήσεις είναι στοιχεία του κόμβου ενώ σε διαφορετική περίπτωση είναι μηδέν. Ο αλγόριθμος έχει ως εξής:

1. Για τα βάρη $\boldsymbol{w}$ να γίνει ο καθολικός έλεγχος της μηδενικής υπόθεσης H_0 της ανεξαρτησίας μεταξύ οποιασδήποτε εκ των m συμμεταβλητών και της μεταβλητής απόκρισης. Σταμάτα αν η μηδενική υπόθεση δεν μπορεί να απορριφθεί. Αλλιώς επέλεξε την j^*-οστη συμμεταβλητή X_j* με τη μεγαλύτερη συσχέτιση με το $\boldsymbol{Y}$.

2. Διάλεξε ένα σύνολο $A* \subset \mathcal{X}_j*$ έτσι ώστε να χωριστεί η $\mathcal{X}_j*$ σε δυο ξένα σύνολο $A*$ και $\mathcal{X}_j * \backslash A*$. Τα βάρη $\boldsymbol{w}_{left}$ και $\boldsymbol{w}_{right}$ καθορίζουν τα δύο σύνολα με $\boldsymbol{w}_{left,i} = I(X_{j*i} \in A*)$ και $\boldsymbol{w}_{right,i} = I(X_{j*i} \notin A*)$, για όλα τα $i = 1, ..., n$. (όπου I μια δείκτρια συνάρτηση)

3. Αναδρομικά επανάλαβε τα βήματα 1 και 2 με τροποποιημένα τα $\boldsymbol{w}_{left}$ και $\boldsymbol{w}_{right}$ αντίστοιχα.

Ο τρόπος με τον οποίο γίνεται ο διαχωρισμός στα βήματα 1 και 2 του αλγορίθμου, είναι ο λόγος χάρη στον οποίο δημιουργούνται ευκόλως ερμηνεύσιμα δένδρα τα οποία δεν έχουν την τάση να χρησιμοποιούν τις συμμεταβλητές εκείνες που οδηγούν σε περισσότερες διαμερίσεις. Ο έλεγχος υποθέσεων διενεργείται σε ένα προκαθορισμένο επίπεδο σημαντικότητας a. Επιπλέον ο αλγόριθμος οδηγεί σε μια διαμέριση $B_1, B_2, ..., B_r$, όπου κάθε κελί $B \in B_1, B_2, ..., B_r$ είναι συσχετισμένο με ένα διάνυσμα από βάρη.

Σημειώνεται πως στο πρώτο βήμα χρειάζεται να αποφασιστεί αν η μεταβλητή απόκρισης μπορεί να εξηγηθεί από οποιαδήποτε από τις m συμμεταβλητές του προβλήματος (Hothron & Hornik & Zeileis, 2008). Σε κάθε κόμβο, που χαρακτηρίζεται από τα βάρη $\boldsymbol{w}$, για να γίνει ο καθολικός έλεγχος υποθέσεων για την ανεξαρτησία και των m συμμεταβλητών με την μεταβλητή απόκρισης, γίνονται m μερικοί έλεγχοι υποθέσεων για κάθε μια από τις m συμμεταβλητές. Συγκεκριμένα σε κάθε μερικό έλεγχο, ελέγχεται η μηδενική υπόθεση: $H_j = D(\boldsymbol{Y}|X_j) = D(\boldsymbol{Y})$ και στον καθολικό έλεγχο, η μηδενική υπόθεση: $H_0 = \cap_{j=1}^{m} H_0^j$. Η αναδρομή, όπως αναφέρθηκε και παραπάνω, η αναδρομή σταματάει, όταν δεν μπορεί να απορριφθεί η H_0 σε ένα επίπεδο σημαντικότητας a. Αν

μπορεί να απορριφθεί η H_0, τότε μετριέται η συσχέτιση μεταξύ της $\boldsymbol{Y}$ και καθεμίας από τις m συμμεταβλητές, κάνοντας χρήση των τιμών των στατιστικών συναρτήσεων από τους m ελέγχους υποθέσεων ή των p-values, που δείχνουν την απόκλιση από την μερική μηδενική υπόθεση H_0^j.

Ο τρόπος διεξαγωγής των ελέγχων υποθέσεων στο πρώτο βήμα, γίνεται βάσει των *ελέγχων μεταθέσεων*, που έχουν αναπτυχθεί από τους Strasser και Weber το 1999, όμως δεν θα αναπτυχθούν περαιτέρω καθώς ξεφεύγουν από τα πλαίσια της συγκεκριμένης εργασίας.

Η χρήση της μεθόδου των regression trees για την πρόβλεψη μιας ποσότητας έχει αρκετά πλεονεκτήματα. Όπως αναφέρθηκε και παραπάνω πρόκειται για μια μέθοδο που παράγει μοντέλα που είναι εύκολα να ερμηνευτούν και επίσης υπάρχει οπτική παρουσίαση της μεθόδου. Το γεγονός αυτό κάνει τη μέθοδο αυτή εύχρηστη ακόμα και για ανθρώπους που δεν έχουν κάποιο μαθηματικό υπόβαθρο. Επιπλέον η μέθοδος αυτή μπορεί χρησιμοποιηθεί για δεδομένα κάθε είδους χωρίς να γίνει κάποια υπόθεση για την κατανομή την οποία ακολουθούν και χωρίς να γίνει κάποια ειδική προετοιμασία για τα δεδομένα.

Κεφάλαιο 5

Εφαρμογή

5.1 Εισαγωγή

Στον παρόν κεφάλαιο θα γίνει μια στατιστική ανάλυση πάνω σε ένα σύνολο δεδομένων που αφορά γυναίκες που πάσχουν από καρκίνο του μαστού. Η ανάλυση που θα γίνει θα μελετά το χρόνο επιβίωσης των ασθενών μέχρι να υποτροπιάσουν και τους παράγοντες που επηρεάζουν την πορεία της υγείας τους.

5.1.1 Λίγα λόγια για την ασθένεια

Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας (Π.Ο.Υ) ο καρκίνος του μαστού εμφανίζεται στα κύτταρα των πόρων και των λοβιδίων στον αδενικό ιστό του μαστού (World Health Organization, 2022). Στη συνέχεια μπορεί να προχωρήσει σταδιακά σε 'διπλανά' όργανα ή ιστούς όπως δέρμα, μύες, λεμφαδένες και εν συνεχεία να εξαπλωθεί και σε άλλα όργανα όπως ήπαρ, πνεύμονες, οστά, εγκέφαλος κ.ά (Άλμα Ζωής, 2022).

Ο καρκίνος του μαστού κυρίως 'χτυπά' γυναίκες άνω των 40 ετών και ο κίνδυνος εμφάνισης του αυξάνεται από την παρουσία συγκεκριμένων παραγόντων όπως το κληρονομικό ιστορικό, το κάπνισμα, η έλλειψη άσκησης, η παχυσαρκία, η αυξημένη κατανάλωση αλκοόλ καθώς και η εκτεταμένη ορμονική θεραπεία (World Health Organization, 2022).

Τα στατιστικά στοιχεία που υπάρχουν για την ασθένεια αυτή, αναδεικνύουν την ανάγκη για πρόληψη της ασθένειας μέσω της συχνής εξέτασης. Αναφέρεται ενδεικτικά ότι στις ΗΠΑ το 2004 υπήρχαν περίπου 217.000 νέες περιπτώσεις καρκίνου του μαστού ενώ σήμερα οι ασθενείς είναι περισσότερες από 2.000.000. Στην Ελλάδα ανακύπτουν περίπου 7.772 νέες περιπτώσεις καρκίνου του μαστού το χρόνο. Παγκοσμίως 1 στις 7 γυναίκες θα παρουσιάσει καρκίνο του μαστού σε κάποια φάση της ζωής της (Άλμα Ζωής, 2022).

5.1.2 Το σύνολο δεδομένων της μελέτης

Στο σύνολο δεδομένων προέρχεται από την μελέτη "German Breast Cancer Study Group 2", στην οποία υπάρχουν παρατηρήσεις από 686 γυναίκες. Βάσει αυτών των παρατηρήσεων αναλύονται παράγοντες που σχετίζονται με τον καρκίνο του μαστού. Οι παράγοντες αυτοί είναι:

- horTh: Η ορμονοθεραπεία, μια κατηγορική μεταβλητή με δύο επίπεδα, 'ναι' αν η γυναίκα υπόκειται σε θεραπεία ή 'όχι' αν δεν υπόκειται.
- age: Η ηλικία κάθε ασθενούς, ποσοτική μεταβλητή.
- menostat: Η κατάσταση εμμηνόπαυσης, μια κατηγορική μεταβλητή με δύο επίπεδα "pre", πριν την εμμηνόπαυση και "post" μετά την εμμηνόπαυση.
- tsize: Το μέγεθος του όγκου, μια ποσοτική μεταβλητή μετρημένη σε χιλιοστά (millimetre).
- tgrade: Ο βαθμός του όγκου, μια κατηγορική μεταβλητή διάταξης (ordinal) με τρία επίπεδα, $I < II < III$
- pnodes: Ο αριθμός των θετικών όγκων, μια ποσοτική μεταβλητή.
- progrec: Υποδοχείς προγεστερόνης, πρόκειται για πρωτεΐνες στις οποίες δεσμεύεται η ορμόνη προγεστερόνη, είναι μια ποσοτική μεταβλητή μετρημένη σε fmol (γραμμομόρια).
- estrec: Υποδοχείς οιστρογόνων, πρόκειται για πρωτεΐνες στις οποίες δεσμεύονται τα οιστρογόνα, είναι μια ποσοτική μεταβλητή μετρημένη σε fmol (γραμμομόρια).
- time: Ο χρόνος μέχρι επιβίωσης μέχρι την υποτροπή, μια ποσοτική μεταβλητή μετρημένη σε μέρες.
- cens: Ένας δείκτης για το αν η παρατήρηση είναι αποκομμένη ή όχι, 0 αν η παρατήρηση είναι αποκομμένη, 1 αλλιώς.

5.2 Ανάλυση των δεδομένων

Αρχικά θα περάσουμε στην R τα δεδομένα μας. Θα γίνει χρήση της εντολής **read.table** για να εισαχθούν τα δεδομένα στην R από το αρχείο στο οποίο υπάρχουν. Παρακάτω φαίνονται οι πρώτες 20 γραμμές από τις 686 του πίνακα:

	id	age	tsize	pnodes	progrec	estrec	hormone	meno	tgrad	time	cens
1	1	70	21	3	48	66	no	yes	II	1814	1
2	2	56	12	7	61	77	yes	yes	II	2018	1
3	3	58	35	9	52	271	yes	yes	II	712	1
4	4	59	17	4	60	29	yes	yes	II	1807	1
5	5	73	35	1	26	65	no	yes	II	772	1
6	6	32	57	24	0	13	no	no	III	448	1
7	7	59	8	2	181	0	yes	yes	II	2172	0
8	8	65	16	1	192	25	no	yes	II	2161	0
9	9	80	39	30	0	59	no	yes	II	471	1
10	10	66	18	7	0	3	no	yes	II	2014	0
11	11	68	40	9	16	20	yes	yes	II	577	1
12	12	71	21	9	0	0	yes	yes	II	184	1
13	13	59	58	1	154	101	yes	yes	II	1840	0
14	14	50	27	1	16	12	no	yes	III	1842	0
15	15	70	22	3	113	139	yes	yes	II	1821	0
16	16	54	30	1	135	6	no	yes	II	1371	1
17	17	39	35	4	79	28	no	no	I	707	1
18	18	66	23	1	112	225	yes	yes	II	1743	0
19	19	69	25	1	131	196	yes	yes	I	1781	0
20	20	55	65	4	312	76	no	yes	I	865	1

Showing 1 to 20 of 686 entries, 11 total columns

Σχήμα 5.1: Οι 20 πρώτες γραμμές του πίνακα που περιέχει τα δεδομένα

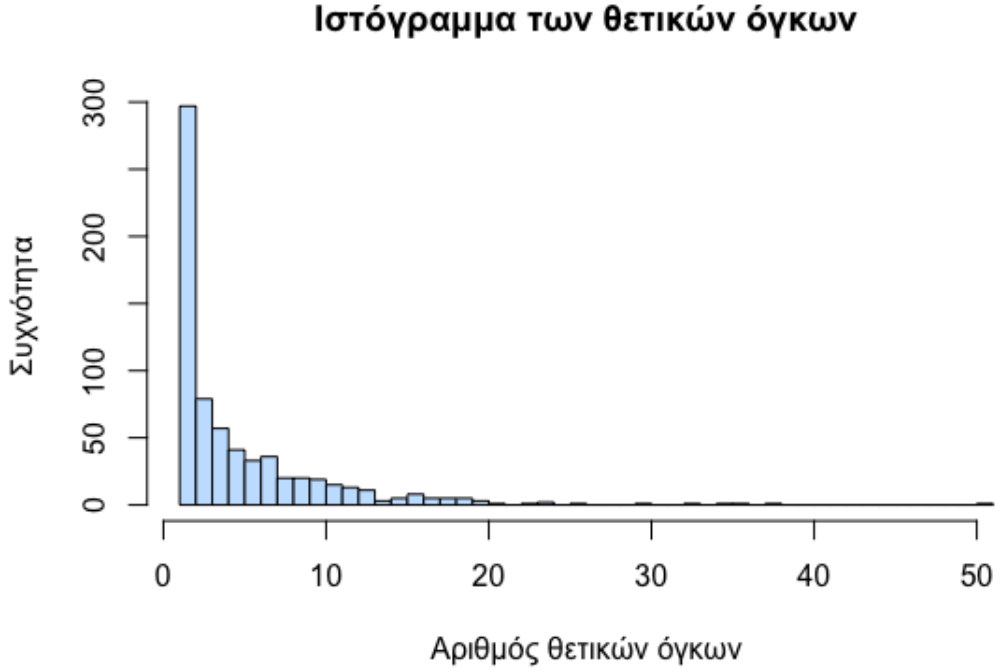
Σημειώνεται πως οι μεταβλητές οι οποίες είναι κατηγορικές καταχωρούνται ως τέτοιες, με την χρήση του ορίσματος **colClasses** στην εντολή **read.table** και στη συνέχεια τα επίπεδα κάθε κατηγορικής μεταβλητής έχουν μετατραπεί από '1' και '2' σε "no" και "yes" καθώς και από '1','2' και '3' σε "I","II" και "III".

5.2.1 Μια περιγραφική ανάλυση των δεδομένων

Χρησιμοποιώντας την εντολή **summary** για τον πίνακα των δεδομένων μας μπορούμε να πάρουμε σημαντικές πληροφορίες για τις ποσοτικές μεταβλητές που έχουμε στη διάθεση μας.

Καταρχάς η διάμεση ηλικία των γυναικών του δείγματος μας είναι τα 53 έτη. Επιπλέον το μέσο μέγεθος των όγκων που έχουν οι γυναίκες στο δείγμα, είναι 29.33 mm. Επιπλέον όπως απεικονίζεται και παρακάτω στο Ιστόγραμμα των αριθμών των

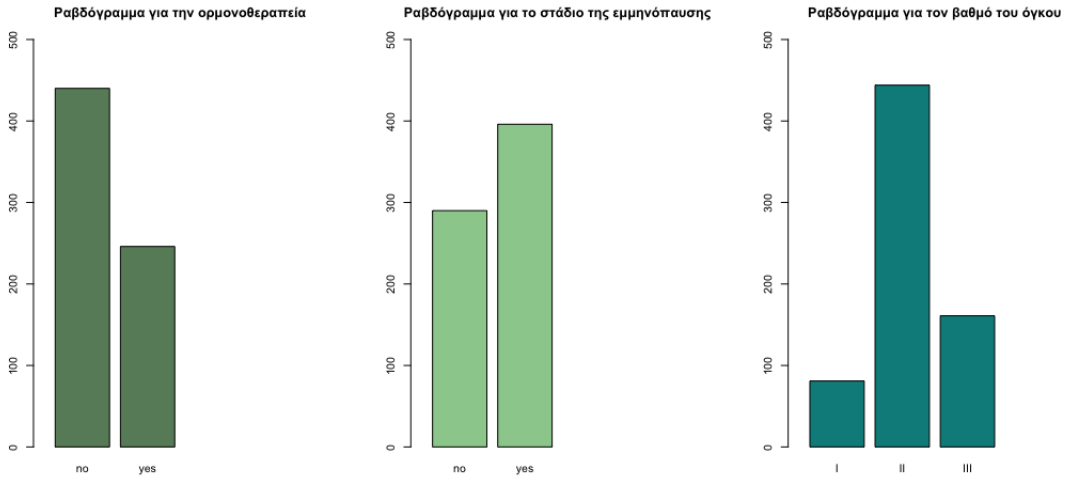
θετικών όγκων, δηλαδή στο Σχήμα 5.2, οι περισσότερες ασθενείς του δείγματος έχουν λιγότερους από 5 θετικούς όγκους. Ο μέσος όρος των υποδοχέων της προγεστερόνης είναι 110.0 και των οιστρογόνων είναι 96.25.



Σχήμα 5.2: Το ιστόγραμμα του βαθμού των θετικών όγκων στις ασθενείς του δείγματος

Παρατηρούμε επίσης πως 440 γυναίκες υποβάλλονται σε ορμοθεραπεία ενώ 226 όχι. Πριν την εμμηνόπαυση βρίσκονται 290 ασθενείς ενώ 396 έχουν περάσει το στάδιο της εμμηνόπαυσης. Επιπροσθέτως όγκο βαθμού "Ι" έχουν 81 γυναίκες, όγκο βαθμού "ΙΙ" έχουν 444 και όγκο βαθμού "ΙΙΙ" έχουν 161 ασθενείς. Τέλος, 387 παρατηρήσεις είναι αποκομμένες και 299 είναι μη-αποκομμένες, δηλαδή έχει υπάρξει υποτροπή στην ασθένεια τους.

Παρακάτω, στο Σχήμα 5.3, παρουσιάζονται τα ραβδογράμματα για τις μεταβλητές που αφορούν την ορμονοθεραπεία, την κατάσταση της εμμηνόπαυσης καθώς και τον βαθμό του όγκου κάθε ασθενούς.



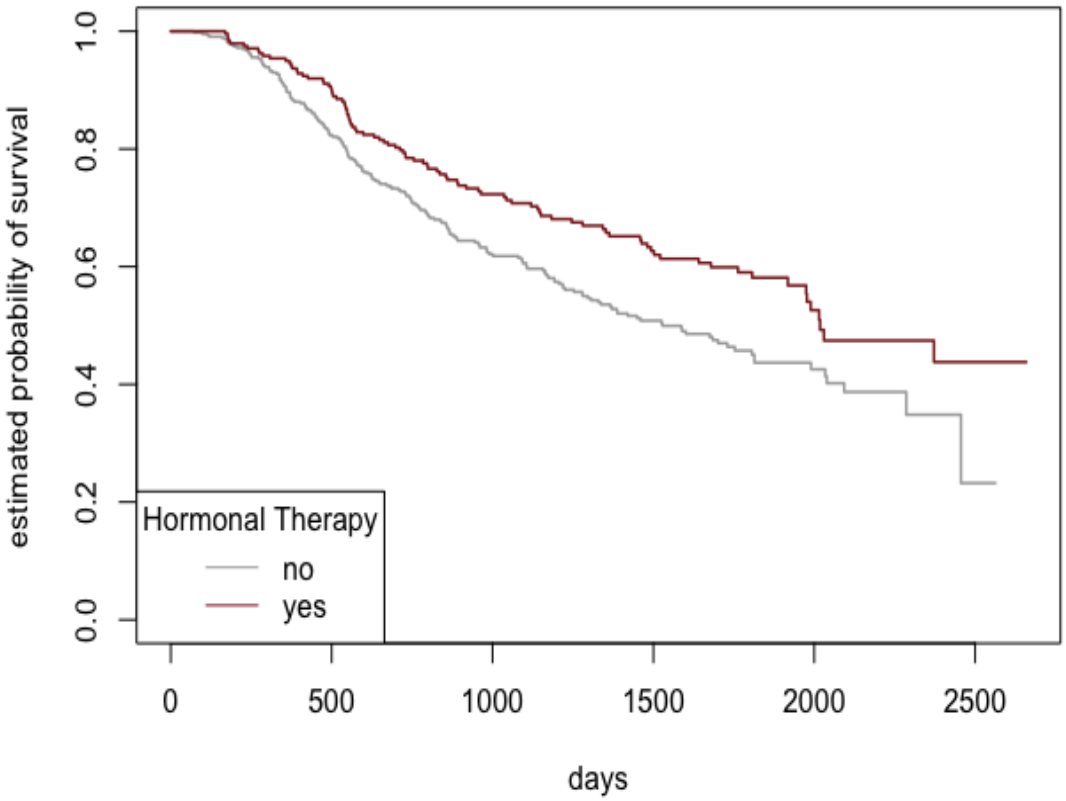
Σχήμα 5.3: Τα ραβδογράμματα για την ορμονοθεραπεία, την εμμηνόπαυση καθώς και το βαθμό του όγκου

5.2.2 Μη παραμετρική Ανάλυση

Οι εκτιμήσεις Kaplan-Meier

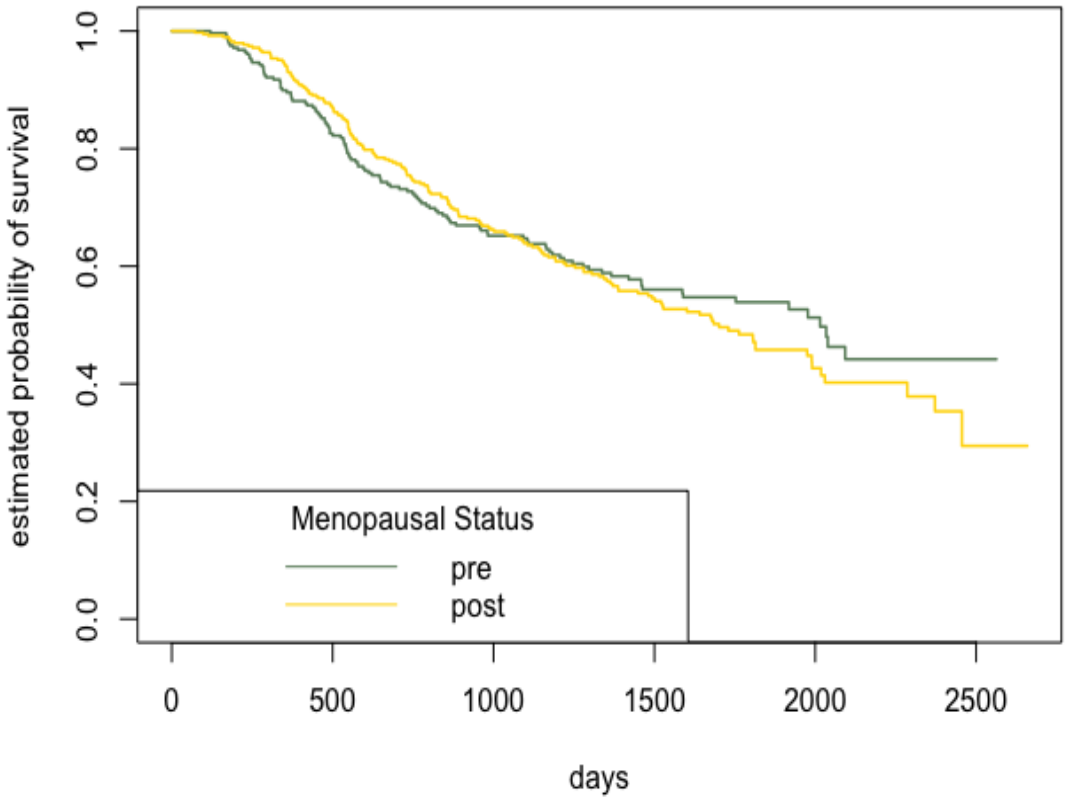
Στη συνέχεια είναι αρκετά σημαντικό για να πάρουμε μια πρώτη ένδειξη για την επιβίωση των γυναικών που συμμετέχουν στην μελέτη μέχρι την υποτροπή της ασθένειας τους, να αναπαραστήσουμε γραφικά τις εκτιμήσεις Kaplan-Meier του δείγματος μας.

Αρχικά παρουσιάζονται οι καμπύλες Kaplan-Meier για δύο ομάδες ασθενών αυτές που υποβάλλονται σε ορμονοθεραπεία και αυτές που δεν υποβάλλονται. Όπως φαίνεται και στο Σχήμα 5.4, οι ασθενείς οι οποίες έχουν υποβληθεί σε ορμονοθεραπεία παρουσιάζουν περισσότερες πιθανότητες να επιβιώσουν από αυτές που δεν έχουν υποβληθεί σε θεραπεία.



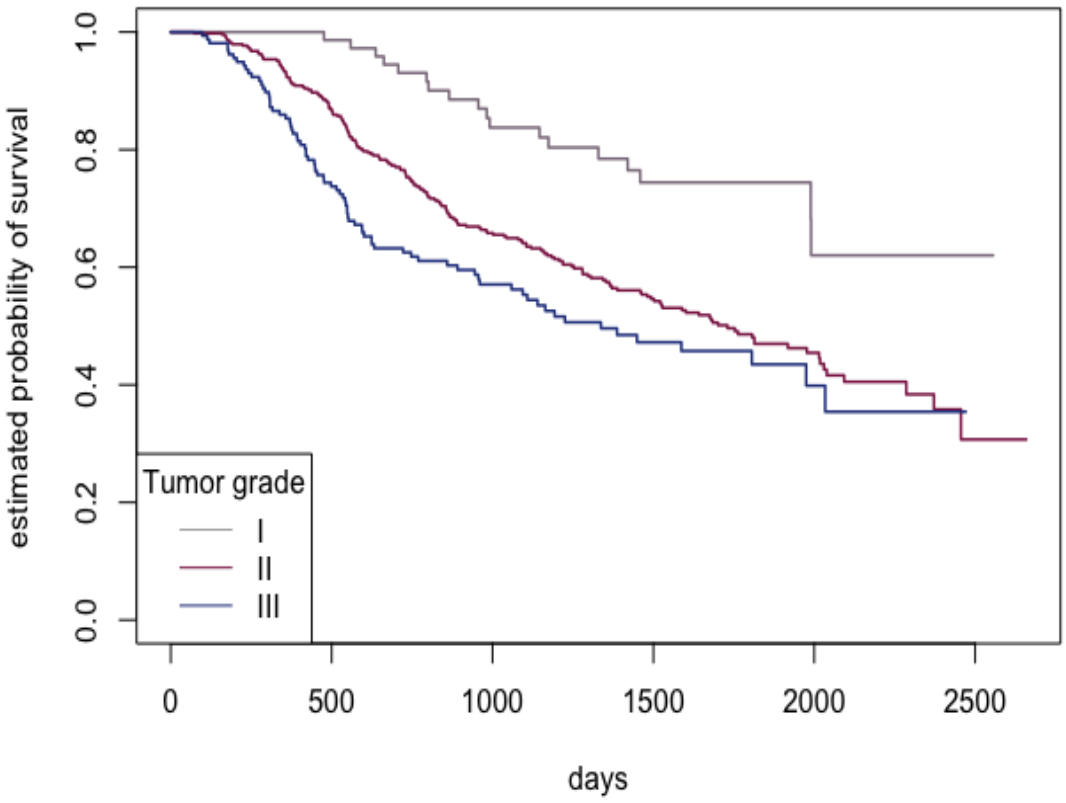
Σχήμα 5.4: Οι εκτιμήσεις Kaplan-Meier της επιβίωσης των ασθενών του δείγματος για τις περιπτώσεις που οι ασθενείς υποβάλλονται σε ορμονοθεραπεία ή όχι.

Στη συνέχεια παρατίθεται στο Σχήμα 5.5 η γραφική παράσταση των εκτιμητριών Kaplan-Meier, για τις ασθενείς εκείνες που βρίσκονται στο στάδιο πριν την εμμηνόπαυση και για εκείνες που έχουν περάσει την εμμηνόπαυση. Είναι εμφανές πως για τις ασθενείς που βρίσκονται στο στάδιο πριν την εμμηνόπαυση, οι πιθανότητες επιβίωσης είναι μεγαλύτερες από εκείνες που έχουν περάσει αυτό το στάδιο και το ρίσκο να υποτροπιάσουν είναι μεγαλύτερο. Είναι ιδιαίτερα σημαντικό όμως να συγκρατήσουμε πως φαίνεται και οι δύο ομάδες ασθενών να ακολουθούν μια περίπου ίδια πορεία, αφού όπως βλέπουμε οι δυο καμπύλες 'συμβαδίζουν ' στο μεγαλύτερο κομμάτι του διαγράμματος.



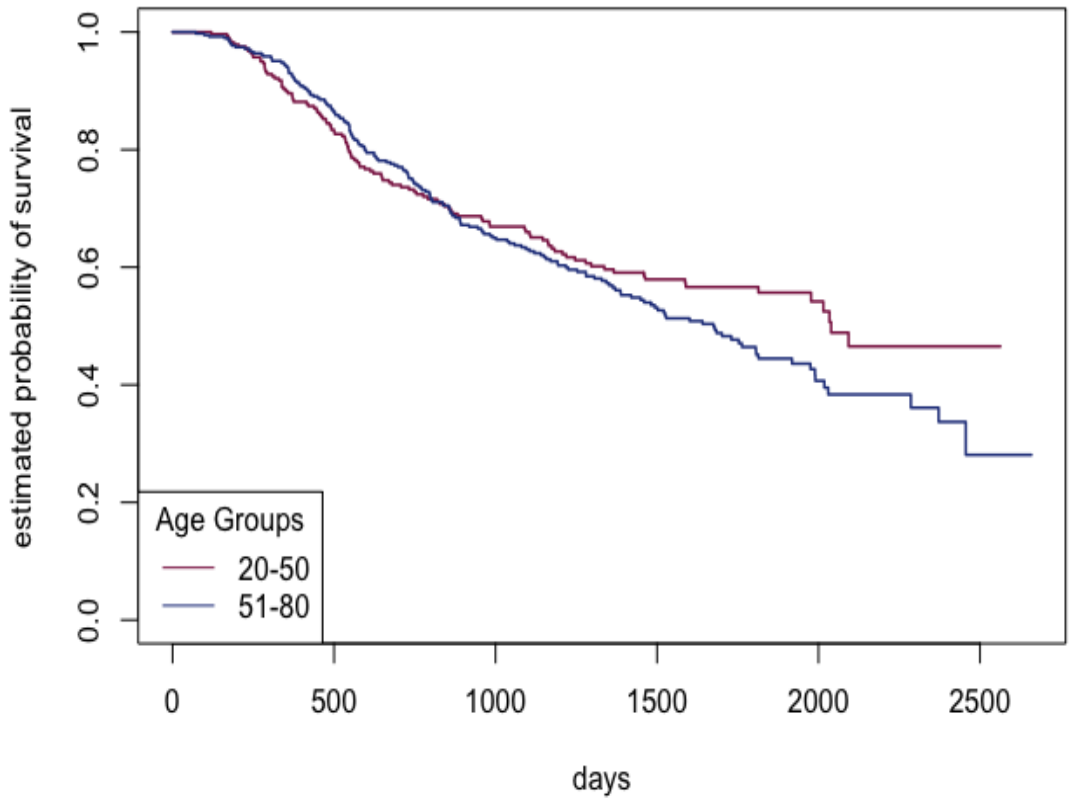
Σχήμα 5.5: Οι εκτιμήσεις Kaplan-Meier της επιβίωσης των ασθενών του δείγματος για την κατάσταση της εμμηνόπαυσης των ασθενών του δείγματος

Επιπλέον, παρακάτω στο Σχήμα 5.6 φαίνεται η γραφική παράσταση των εκτιμήσεων Kaplan-Meier, για την υποτροπή των ασθενών βάσει των τριών βαθμών όγκων που μπορεί να έχουν. Οι ασθενείς εκείνες με βαθμό όγκου 'I' παρουσιάζουν τις περισσότερες πιθανότητες επιβίωσης στη συνέχεια ακολουθούν αυτές με βαθμό 'II' και όπως παρατηρούμε οι ασθενείς με βαθμό όγκου 'III' φτάνουν στην υποτροπή πιο γρήγορα από τις άλλες δύο ομάδες ασθενών.



Σχήμα 5.6: Οι εκτιμήσεις Kaplan-Meier της επιβίωσης των ασθενών του δείγματος για τους 3 βαθμούς των θετικών όγκων

Τέλος, χωρίζοντας σε δυο κατηγορίες τις ασθενείς, βάσει των ηλικιών τους θα εκτιμήσουμε και πάλι τις εκτιμήσεις Kaplan-Meier για τις δύο ομάδες ασθενών για να δούμε πως επηρεάζει η ηλικία, το χρόνο ως την υποτροπή μιας ασθενούς. Χρησιμοποιώντας την εντολή cut της R χωρίζουμε την μεταβλητή age που αφορά την ηλικία των ασθενών σε δύο ομάδες, $20-50$ ετών και $51-80$ ετών. Όπως παρατηρούμε στο Σχήμα 5.7 παρακάτω, οι πιθανότητες επιβίωσης των γυναικών ηλικίας $20-50$ είναι μεγαλύτερες από αυτές της ηλικιακής ομάδας $51-80$ όπου αντιμετωπίζουν μεγαλύτερο ρίσκο και φτάνουν πιο γρήγορα στην υποτροπή.



Σχήμα 5.7: Οι εκτιμήσεις Kaplan-Meier της επιβίωσης των ασθενών του δείγματος, για τις ηλικιακές ομάδες ασθενών $20-50$ ετών και $51-80$ ετών

Έλεγχοι Υποθέσεων

Θα προχωρήσουμε και σε ένα Log-Rank έλεγχο υποθέσεων μεταξύ των διάφορων ομάδων που προκύπτουν μέσα από τα δεδομένα μας όπως κάναμε προηγουμένως με τις εκτιμήσεις Kaplan-Meier.

Είναι ιδιαιτέρως σημαντικό, να κατανοήσουμε ποιοι ακριβώς είναι οι παράγοντες εκείνοι που οδηγούν πιο γρήγορα σε υποτροπή της κατάστασης της υγείας των ασθενών. Συγκεκριμένα, στο σύνολο των δεδομένων που έχουμε στη διάθεση μας, θα προσπαθήσουμε να διαπιστώσουμε αν υπάρχουν διαφορές ανάμεσα στις ασθενείς που λαμβάνουν ορμονοθεραπεία και σε αυτές που δεν λαμβάνουν, αν η κατάσταση της εμμηνόπαυσης τους επηρεάζει την ασθένεια, καθώς και αν ο βαθμός των θετικών όγκων που έχουν, διαφοροποιεί την εξέλιξη της ασθένειας και τέλος πως διαμορφώνεται η πορεία μέχρι την υποτροπή για τις ασθενείς άνω των 50 ετών και για αυτές που είναι κάτω των 50.

- Αρχικά λοιπόν θα ελέγξουμε τον παράγοντα της ορμονοθεραπείας. Συγκεκριμένα θα ελέγξουμε αν υπάρχει η επιβίωση των ασθενών είναι διαφορετική αν λαμβάνουν θεραπεία με ορμόνες ή όχι. Ο Log-Rank έλεγχος υποθέσεων, ελέγχει την μηδενική υπόθεση, ότι δεν υπάρχει διαφορά στο συνάρτηση επιβίωσης μεταξύ των δύο ομάδων δεδομένων. Επιλέγουμε το επίπεδο σημαντικότητας του ελέγχου να είναι 5%.

```
> logrank.test <- survdiff(Surv(table$time,table$cens)~table$hormone)
> logrank.test
Call:
survdiff(formula = Surv(table$time, table$cens) ~ table$hormone)

                    N Observed Expected (O-E)^2/E (O-E)^2/V
table$hormone=no  440      205      180      3.37      8.56
table$hormone=yes 246       94      119      5.12      8.56

 Chisq= 8.6  on 1 degrees of freedom, p= 0.003
```

Σχήμα 5.8: Log-Rank έλεγχος υποθέσεων για τον παράγοντα της ορμονοθεραπείας.

Ο έλεγχος μας οδηγεί στο συμπέρασμα πως υπάρχουν αρκετές ενδείξεις να απορρίψουμε την μηδενική υπόθεση ότι δεν υπάρχει διαφορά στο χρόνο επιβίωσης μεταξύ των ασθενών που λαμβάνουν θεραπεία με ορμόνες και σε αυτές που δεν λαμβάνουν. Η πολύ μικρή τιμή p-value του ελέγχου ($p - value = 0.003$), μας οδηγεί στην απόρριψη της μηδενικής υπόθεσης, γεγονός που συμφωνεί και με το διάγραμμα των εκτιμήσεων των καμπυλών Kaplan-Meier που έχει προηγηθεί.

- Παρακάτω παρατίθεται ο Log-Rank έλεγχος υποθέσεων για τον παράγοντα της εμμηνόπαυσης. Και για τον παράγοντα αυτό ελέγχουμε και πάλι σε επίπεδο σημαντικότητας 5% αν οι γυναίκες που βρίσκονται πριν την εμμηνόπαυση διαφέρουν στο χρόνο επιβίωσης σε σχέση με αυτές που βρίσκονται μετά την εμμηνόπαυση. Όπως και παραπάνω η μηδενική υπόθεση του ελέγχου είναι πως δεν υπάρχει διαφορά μεταξύ των δύο αυτών ομάδων ασθενών.

```
> logrank.test <- survdiff(Surv(table$time,table$cens)~table$meno)
> logrank.test
Call:
survdiff(formula = Surv(table$time, table$cens) ~ table$meno)

                 N Observed Expected (O-E)^2/E (O-E)^2/V
table$meno=no  290      119      124     0.164      0.28
table$meno=yes 396      180      175     0.115      0.28

 Chisq= 0.3  on 1 degrees of freedom, p= 0.6
```

Σχήμα 5.9: Log-Rank έλεγχος υποθέσεων για τον παράγοντα της εμμηνόπαυσης

Η p-value τιμή του ελέγχου είναι αρκετά μεγάλη, συγκεκριμένα $p-value = 0.6$, δεν μας δίνει αρκετές ενδείξεις έτσι ώστε να μπορούμε να απορρίψουμε την μηδενική υπόθεση, ότι δηλαδή δεν υπάρχει διαφορά στο χρόνο επιβίωσης των ασθενών είτε αυτές βρίσκονται πριν, είτε αυτές βρίσκονται μετά την περίοδο της εμμηνόπαυσης. Το αποτέλεσμα αυτό, δηλαδή πως ο χρόνος επιβίωσης των ασθενών δεν επηρεάζεται από την κατάσταση της εμμηνόπαυσης μιας ασθενούς, συμφωνεί και με το διάγραμμα των εκτιμήσεων Kaplan-Meier, όπου όπως μπορεί κάποιος να παρατηρήσει, κατά το μεγαλύτερο κομμάτι του γραφήματος οι δύο καμπύλες σχεδόν ταυτίζονται.

- Στη συνέχεια θα προβούμε σε έναν ακόμα Log-Rank έλεγχο υποθέσεων για τον παράγοντα του βαθμού των θετικών όγκων που παρουσιάζει κάθε ασθενής. Θα ελέγξουμε δηλαδή, αν η επιβίωση των ασθενών διαφοροποιείται αναλόγως του βαθμού των θετικών όγκων που έχει η καθεμία ή αν ο παράγοντας αυτός δεν παίζει ρόλο στην εξέλιξη της ασθένειας.

```
> logrank.test <- survdiff(Surv(table$time,table$cens)~table$tgrad)
> logrank.test
Call:
survdiff(formula = Surv(table$time, table$cens) ~ table$tgrad)

                  N Observed Expected (O-E)^2/E (O-E)^2/V
table$tgrad=I    81       18     42.2   13.8469    16.159
table$tgrad=II  444      202    198.2    0.0725     0.215
table$tgrad=III 161       79     58.6    7.0788     8.848

 Chisq= 21.1  on 2 degrees of freedom, p= 3e-05
```

Σχήμα 5.10: Log-Rank έλεγχος υποθέσεων για τον παράγοντα του βαθμού των θετικών όγκων.

Παρατηρούμε πως για τον συγκεκριμένο έλεγχο η p-value τιμή του ελέγχου είναι εξαιρετικά μικρή, γεγονός που μας οδηγεί στο συμπέρασμα πως υπάρχουν αρκετές ενδείξεις, ώστε να απορρίψουμε την μηδενική υπόθεση ότι δεν υπάρχουν διαφορές μεταξύ των τριών βαθμών θετικών όγκων που μπορεί να έχει κάθε ασθενής. Και σε αυτή την περίπτωση ο έλεγχος αυτός συμφωνεί με το διάγραμμα των Kaplan-Meier εκτιμήσεων και συνεπώς καταλήγουμε πως η επιβίωση των ασθενών εξαρτάται από τον βαθμό των όγκων καθώς και τα δύο ευρήματα που έχουμε στη διάθεση μας συγκλίνουν σε αυτό το συμπέρασμα.

- Τέλος, θα προχωρήσουμε σε έναν Log-Rank για τις δύο ηλικιακές ομάδες που έχουμε δημιουργήσει, χωρίζοντας τα δεδομένα μας στις ασθενείς που είναι κάτω των 50 ετών και σε αυτές που είναι άνω των 50. Θα εξετάσουμε λοιπόν αν η ηλικία των ασθενών επηρεάζει την επιβίωση τους.

```
> logrank.test <- survdiff(Surv(table$time,table$cens)~agecat)
> logrank.test
Call:
survdiff(formula = Surv(table$time, table$cens) ~ agecat)

               N Observed Expected (O-E)^2/E (O-E)^2/V
agecat=20-50 289      114      125     0.988       1.7
agecat=51-80 397      185      174     0.711       1.7

 Chisq= 1.7  on 1 degrees of freedom, p= 0.2
```

Σχήμα 5.11: Log-Rank έλεγχος υποθέσεων για τις δύο ηλικιακές ομάδες

Τα αποτελέσματα του ελέγχου μας κάνουν να πιστέψουμε πως δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση, ότι οι δύο ηλικιακές ομάδες δεν διαφοροποιούνται. Η p-value τιμή του ελέγχου είναι αρκετά μεγάλη συγκεκριμένα 0.2 (πολύ μεγαλύτερη του επιπέδου σημαντικότητας 0.05) και όπως και σε όλες τις παραπάνω περιπτώσεις ελέγχου και εδώ τα αποτελέσματα συμφωνούν με το διάγραμμα των εκτιμήσεων Kaplan-Meier. Συνεπώς έχουμε αρκετές ενδείξεις για να καταλήξουμε στο ότι ο χρόνος επιβίωσης των δύο ηλικιακών ομάδων δεν διαφοροποιείται.

Παρά τις ενδείξεις για διαφοροποιήσεις μεταξύ ομάδων των δεδομένων μας, δεν θα προχωρήσουμε σε στρωματοποιημένη ανάλυση καθώς όπως μπορούμε να παρατηρήσουμε στις γραφικές παραστάσεις των εκτιμήσεων Kaplan-Meier δεν παρουσιάζεται μεγάλη απόκλιση στην επιβίωση των ασθενών.

5.2.3 Το μοντέλο του Cox

Θα συνεχίσουμε την ανάλυση μας, προσαρμόζοντας το μοντέλο αναλογικής διακινδύνευσης του Cox στα δεδομένα μας. Η παραπάνω προσαρμογή θα γίνει χρησιμοποιώντας την βιβλιοθήκη survival της R και θα μας βοηθήσει να καταλήξουμε σε κάποια συμπεράσματα για τις μεταβλητές εκείνες που επηρεάζουν την πορεία της υγείας των ασθενών του δείγματος μας. Θα γίνει η προσαρμογή του μοντέλου, θα προσπαθήσουμε να εντοπίσουμε τις συμμεταβλητές εκείνες που χρειάζονται να συμπεριληφθούν στο μοντέλο και εκείνες που δεν επηρεάζουν σε σημαντικό βαθμό την πορεία της υγείας των ασθενών. Τέλος, θα ελέγξουμε την υπόθεση αναλογικής διακινδύνευσης για το μοντέλο στο οποίο θα έχουμε καταλήξει από την ανάλυση μας.

Παρακάτω στο Σχήμα 5.12 δίνονται τα αποτελέσματα της μεθόδου “summary”, για το μοντέλο που προσαρμόστηκε. Το Σχήμα 5.12 μας δίνει πληροφορίες για τους συντελεστές των συμμεταβλητών, τα εκθετικά αυτών, την τυπική τους απόκλιση καθώς και τα στατιστικά ελέγχου των Wald ελέγχων υποθέσεων για κάθε μεταβλητή και την p-value τιμή των ελέγχων αυτών.

```
> summary(fit.all)
Call:
coxph(formula = Surv(table$time, table$cens) ~ table$hormone +
    table$age + table$tsize + table$pnodes + table$progrec +
    table$estrec + table$meno + table$tgrad)

  n= 686, number of events= 299

                      coef  exp(coef)   se(coef)      z Pr(>|z|)
table$hormoneyes -0.3462784  0.7073155  0.1290747 -2.683 0.007301 **
table$age        -0.0094592  0.9905854  0.0093006 -1.017 0.309126
table$tsize       0.0077961  1.0078266  0.0039390  1.979 0.047794 *
table$pnodes      0.0487886  1.0499984  0.0074471  6.551  5.7e-11 ***
table$progrec    -0.0022172  0.9977852  0.0005735 -3.866 0.000111 ***
table$estrec      0.0001973  1.0001973  0.0004504  0.438 0.661307
table$menoyes     0.2584448  1.2949147  0.1834765  1.409 0.158954
table$tgradII     0.6361117  1.8891211  0.2492025  2.553 0.010693 *
table$tgradIII    0.7796542  2.1807181  0.2684801  2.904 0.003685 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                 exp(coef) exp(-coef) lower .95 upper .95
table$hormoneyes    0.7073     1.4138    0.5492    0.9109
table$age           0.9906     1.0095    0.9727    1.0088
table$tsize         1.0078     0.9922    1.0001    1.0156
table$pnodes        1.0500     0.9524    1.0348    1.0654
table$progrec       0.9978     1.0022    0.9967    0.9989
table$estrec        1.0002     0.9998    0.9993    1.0011
table$menoyes       1.2949     0.7723    0.9038    1.8553
table$tgradII       1.8891     0.5293    1.1591    3.0788
table$tgradIII      2.1807     0.4586    1.2885    3.6909

Concordance= 0.692  (se = 0.015 )
Likelihood ratio test= 104.8  on 9 df,   p=<2e-16
Wald test            = 114.8  on 9 df,   p=<2e-16
Score (logrank) test = 120.7  on 9 df,   p=<2e-16
```

Σχήμα 5.12: Το μοντέλο αναλογικής διακινδύνευσης του Cox.

Από τα αποτελέσματα στου Σχήματος 5.12 καταλήγουμε στα ακόλουθα συμπεράσματα:

- Αρχικά όπως φαίνεται από τα αποτελέσματα των τριών ελέγχων υποθέσεων, Likelihood ratio test, Wald test καθώς και Score test, μπορούμε να υποστηρίξουμε πως υπάρχουν αρκετές ενδείξεις να απορρίψουμε την μηδενική υπόθεση, πως όλοι συντελεστές των συμμεταβλητών είναι μηδέν (δηλαδή ότι ισχύει $\beta_1 = \beta_2 = ... = \beta_9 = 0$). Και οι τρεις έλεγχοι συμφωνούν, αφού οι p-value τιμές των ελέγχων είναι πολύ μικρότερες από το επίπεδο σημαντικότητας που και σε αυτό τον έλεγχο θεωρούμε πως είναι 0.05. Αυτό το αποτέλεσμα, μας υποδεικνύει πως σίγουρα κάποιες από αυτές τις μεταβλητές χρειάζονται στο μοντέλο μας.

- Στη συνέχεια μπορούμε να πάρουμε μια αρχική ένδειξη για το ποια μεταβλητή χρειάζεται στο μοντέλο μας και ποια όχι. Για κάθε μια από τις συμμεταβλητές πραγματοποιείται ένας έλεγχος Wald και σημειώνεται η τίμη p-value κάθε ελέγχου. Όπως και παραπάνω σε επίπεδο σημαντικότητας 0.05 ελέγχουμε την μηδενική υπόθεση ότι $\beta_i = 0$, με εναλλακτική πως είναι $\beta_i \neq 0$ και συνεπώς αν μια συμμεταβλητή συμβάλλει και χρειάζεται στο μοντέλο μας ή όχι. Καταλήγουμε πως οι συμμεταβλητές έχουμε αρκετές ενδείξεις πως χρειάζεται να συμπεριληφθούν στο μοντέλο μας σύμφωνα με τον έλεγχο Wald, είναι η ορμονοθεραπεία (*hormoneyes*), ο αριθμός θετικών όγκων (*pnodes*), ο υποδοχέας προγεστερόνης (*progrec*), και ο βαθμός θετικών όγκων ($tgradII$ και $tgradIII$). Σημειώνεται πως και η μεταβλητή $tsize$ θα μπορούσε να θεωρηθεί σημαντική αφού η p-value τιμή της είναι 0.047, πολύ κοντά δηλαδή στο επίπεδο σημαντικότητας 0.05.

- Τέλος παρατηρούμε πως το ποσοστό συμφωνίας της πρόβλεψης και του μοντέλου είναι αρκετά ικανοποιητικό καθώς όπως φαίνεται από η τιμή *Concordance* είναι αρκετά υψηλή και συγκεκριμένα 0.692.

Στη συνέχεια αφού έχουμε κάποιες αρχικές ενδείξεις για το ποιες συμμεταβλητές είναι σημαντικές και χρειάζονται στο μοντέλο μας από τους ελέγχους Wald, θα προχωρήσουμε σε διάφορες δοκιμές για να δούμε πως μεταβάλλεται η τιμή του κριτηρίου AIC, όταν εναλλάσσονται οι συμμεταβλητές στο μοντέλο. Υπενθυμίζεται πως βέλτιστο θεωρείται το μοντέλο εκείνο που θα πετύχει μικρή τιμή του κριτηρίου.

Αρχικά στα μοντέλα που θα υπολογίσουμε την τιμή του κριτηρίου AIC θα συμπεριλάβουμε σίγουρα τις μεταβλητές *hormone*, *pnodes*, *progrec* και *tgrad* καθώς όπως είδαμε χρειάζονται στο μοντέλο αφού επηρεάζουν το χρόνο επιβίωσης των ασθενών. Οι παρακάτω έλεγχοι θα μας βοηθήσουν να προσδιορίσουμε το τελικό μας μοντέλο.

Συμμεταβλητές στο μοντέλο	Τιμή κριτηρίου AIC
hormone, *pnodes*, *progrec* και *tgrad*	3487.045
hormone, *pnodes*, *progrec*, *tgrad* και *meno*	3488.1
hormone, *pnodes*, *progrec*, *tgrad* και *estrec*	3488.929
hormone, *pnodes*, *progrec*, *tgrad* και *tsize*	3485.681
hormone, *pnodes*, *progrec*, *tgrad* και *age*	3489.026
καμία συμμεταβλητή	3576.209
hormone, *pnodes*, *progrec*, *tgrad*, *meno*,*estrec* και *tsize*	3489.464

Παρατηρούμε πως το μοντέλο που περιέχει τις συμμεταβλητές *hormone*, *pnodes*, *progrec*, *tgrad* καθώς και *tsize* είναι αυτό που μας δίνει τη χαμηλότερη τιμή του κριτηρίου AIC, γεγονός που ενισχύει την πεποίθηση μας πως αυτές είναι οι μεταβλητές που πρέπει οπωσδήποτε να συμπεριληφθούν στο μοντέλο μας.

Στη συνέχεια, όπως φαίνεται στο Σχήμα 5.13, 5.14 και 5.15, θα χρησιμοποιήσουμε την μέθοδο StepWise Selection, ωστέ να μπορέσουμε να καταλήξουμε στο βέλτιστο μοντέλο για την προσαρμογή των δεδομένων μας.

```
> coxstep<-step(fit.all,direction = "both")
Start:  AIC=3489.46
Surv(table$time, table$cens) ~ table$hormone + table$age + table$tsize +
    table$pnodes + table$progrec + table$estrec + table$meno +
    table$tgrad

                Df    AIC
- table$estrec   1 3487.7
- table$age      1 3488.5
- table$meno     1 3489.4
<none>             3489.5
- table$tsize    1 3491.2
- table$hormone  1 3494.9
- table$tgrad    2 3495.3
- table$progrec  1 3507.5
- table$pnodes   1 3519.5
```

Σχήμα 5.13: Stepwise διαδικασία

```
Step:  AIC=3487.65
Surv(table$time, table$cens) ~ table$hormone + table$age + table$tsize +
    table$pnodes + table$progrec + table$meno + table$tgrad

                Df    AIC
- table$age      1 3486.6
- table$meno     1 3487.7
<none>             3487.7
- table$tsize    1 3489.3
+ table$estrec   1 3489.5
- table$hormone  1 3493.0
- table$tgrad    2 3493.5
- table$progrec  1 3506.8
- table$pnodes   1 3517.6

Step:  AIC=3486.56
Surv(table$time, table$cens) ~ table$hormone + table$tsize +
    table$pnodes + table$progrec + table$meno + table$tgrad

                Df    AIC
- table$meno     1 3485.7
<none>             3486.6
+ table$age      1 3487.7
- table$tsize    1 3488.1
+ table$estrec   1 3488.5
- table$hormone  1 3492.3
- table$tgrad    2 3492.6
- table$progrec  1 3506.6
- table$pnodes   1 3516.4
```

Σχήμα 5.14: Stepwise διαδικασία

```
Step:  AIC=3485.68
Surv(table$time, table$cens) ~ table$hormone + table$tsize +
    table$pnodes + table$progrec + table$tgrad

                Df    AIC
<none>             3485.7
+ table$meno     1 3486.6
- table$tsize    1 3487.0
+ table$estrec   1 3487.4
+ table$age      1 3487.7
- table$hormone  1 3490.5
- table$tgrad    2 3491.8
- table$progrec  1 3506.1
- table$pnodes   1 3515.9
```

Σχήμα 5.15: Stepwise διαδικασία

Βάσει των αποτελεσμάτων και στη διαδικασία κατά βήματα (Stepwise selection), παρατηρούμε πως μικρότερη τιμή του κριτηρίου AIC εμφανίζεται στο μοντέλο που περιέχει τις συμμεταβλητές $hormone, tsize, pnodes, progrec,, tgrad$ καθώς και την μεταβλητή $tsize$, όπου όπως αναφέρθηκε και παραπάνω είχαμε στοιχεία που μας έδειχναν ότι πρόκειται για σημαντική μεταβλητή.

Όλα τα παραπάνω μας οδηγούν στο συμπέρασμα πως το τελικό μοντέλο που θα χρησιμοποιήσουμε συμπεριλαμβάνει τις μεταβλητές $hormone, pnodes, progrec,, tgrad$ και $tsize$. Στο Σχήμα 5.16 φαίνεται η προσαρμογή αυτού ακριβώς του μοντέλου :

```
> summary(fit.model)
Call:
coxph(formula = Surv(table$time, table$cens) ~ table$hormone +
    table$pnodes + table$progrec + table$tgrad + table$tsize)

  n= 686, number of events= 299

                       coef  exp(coef)   se(coef)      z Pr(>|z|)
table$hormoneyes -0.3235201  0.7235974  0.1258239 -2.571   0.0101 *
table$pnodes      0.0489976  1.0502178  0.0074529  6.574 4.89e-11 ***
table$progrec    -0.0022168  0.9977856  0.0005538 -4.003 6.26e-05 ***
table$tgradII     0.6440346  1.9041479  0.2490184  2.586   0.0097 **
table$tgradIII    0.7882290  2.1994977  0.2682585  2.938   0.0033 **
table$tsize       0.0073129  1.0073397  0.0038897  1.880   0.0601 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                 exp(coef) exp(-coef) lower .95 upper .95
table$hormoneyes    0.7236     1.3820    0.5655    0.9260
table$pnodes        1.0502     0.9522    1.0350    1.0657
table$progrec       0.9978     1.0022    0.9967    0.9989
table$tgradII       1.9041     0.5252    1.1688    3.1022
table$tgradIII      2.1995     0.4546    1.3001    3.7211
table$tsize         1.0073     0.9927    0.9997    1.0150

Concordance= 0.689  (se = 0.015 )
Likelihood ratio test= 102.5  on 6 df,   p=<2e-16
Wald test            = 112.2  on 6 df,   p=<2e-16
Score (logrank) test = 118.5  on 6 df,   p=<2e-16
```

Σχήμα 5.16: Το τελικό μοντέλο

Ερμηνεία Αποτελεσμάτων

Στον παραπάνω Σχήμα 5.16, μπορούμε να δούμε τις εκτιμήτριες των συντελεστών β_i του τελικού μοντέλου, καθώς και τα εκθετικά των συντελεστών αυτών που στην προκειμένη μας ενδιαφέρουν περισσότερο.

Γενικά, τα εκθετικά των συντελεστών υποδεικνύουν κατά πόσο πολλαπλασιάζεται η συνάρτηση διακινδύνευσης, δηλαδή κατά πόσο μια συμμεταβλητή επιδρά στη διάρκεια ζωής, όταν όλες οι άλλες συμμεταβλητές θεωρούνται σταθερές. Συνεπώς το μοντέλο μας συμπεραίνουμε ότι:

- Για τις ασθενείς που έχουν λάβει θεραπεία ορμονών, σε σχέση με αυτές που δεν έχουν λάβει, ο κίνδυνος να υποτροπιάσει η ασθενής μειώνεται κατά $e^{-0.3235201} = 0.7235974$. Δηλαδή, οδηγούμαστε σε μείωση του κινδύνου κατά περίπου 28% $(0.7235974 - 1 = -0.2764026)$.

- Αύξηση κατά μία μονάδα της μεταβλητής *pnodes* (υπό την προϋπόθεσή πως οι υπόλοιπες μεταβλητές παραμένουν σταθερές), δηλαδή αύξηση κατά μία μονάδα στον αριθμό των θετικών όγκων μιας ασθενούς, οδηγεί σε αύξηση του κινδύνου κατά $e^{0.0489976} = 1.0502178$, σε αύξηση δηλαδή της τάξης του 5%.

- Επίσης, αύξηση κατά μία μονάδα στην μεταβλητή του επιπέδου της προγεστερόνης, δεδομένου πως όλες οι υπόλοιπες μεταβλητές παραμένουν σταθερές, οδηγούν σε μείωση του κινδύνου κατά $e^{-0.0022168} = 0.9977856$, δηλαδή 0.22%.

- Κρατώντας σταθερές τις υπόλοιπες μεταβλητές, οι ασθενείς με βαθμό όγκου τύπου II αντιμετωπίζουν αυξημένο κίνδυνο υποτροπής σε σχέση με αυτές που έχουν όγκο τύπου I κατά $e^{0.6440346} = 1.9041479$, σε ποσοστό δηλαδή 90%.

- Υπό την προϋπόθεση πως οι υπόλοιπες μεταβλητές παραμένουν σταθερές, οι ασθενείς που έχουν όγκο τύπου III έχουν αυξημένο κίνδυνο υποτροπής κατά $e^{0.7882290} = 2.1994977$ συγκριτικά με αυτές που έχουν όγκο τύπου I, δηλαδή αντιμετωπίζουν 120% παράπανω κίνδυνο υποτροπής.

- Τέλος, αύξηση κατά μία μονάδα στο μέγεθος του όγκου μιας ασθενούς, δεδομένου πως οι υπόλοιπες μεταβλητές του μοντέλου παραμένουν ίδιες, αυξάνει τον κίνδυνο κατά $e^{0.0073129} = 1.0073397$, περίπου κατά 0.7%.

Έλεγχος υπόθεσης αναλογικής διακινδύνευσης

Καταρχάς θα προχωρήσουμε σε ένα Global test υποθέσεων για να ελέγξουμε την υπόθεση αναλογικής διακινδύνευσης του μοντέλου μας. Σε επίπεδο σημαντικότητας 0.05, θα ελέγξουμε αν μπορούμε να αποδεχθούμε την μηδενική υπόθεση H_0, ότι δηλαδή ίσχυει η αναλογική διακινδύνευση και οι συμμεταβλητές του μοντέλου μας δεν έχουν κάποια εξάρτηση από το χρόνο. Τα αποτελέσματα του ελέγχου φαίνονται στο Σχήμα 5.17

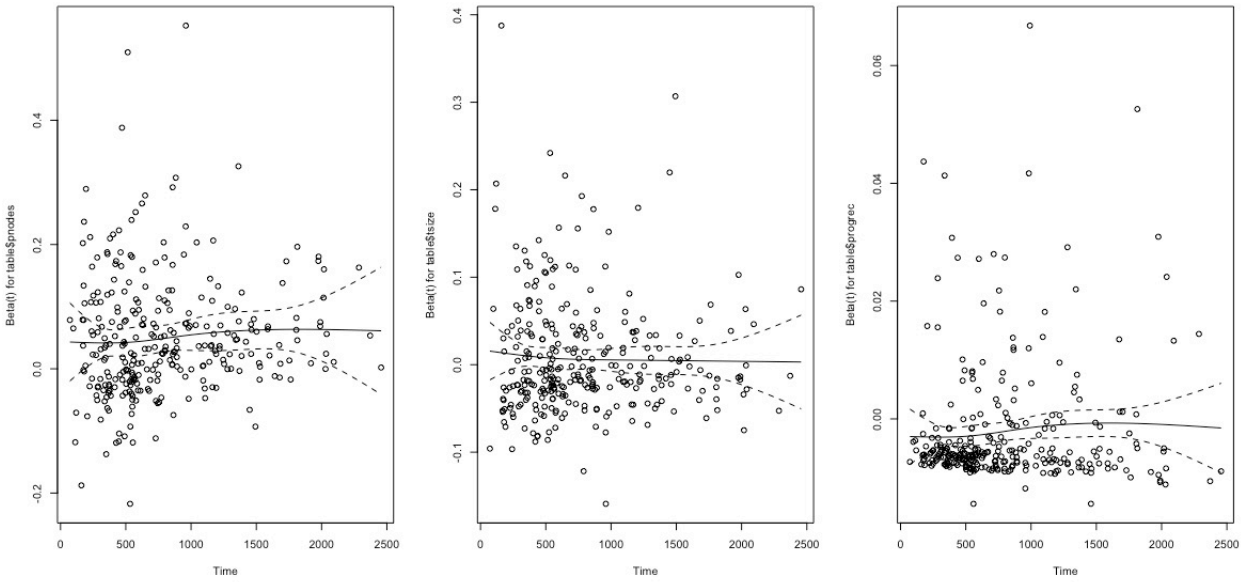
```
> res<-cox.zph(fit.model, transform="identity", global=TRUE)
> res
               chisq df     p
table$hormone  0.169  1 0.681
table$pnodes   1.282  1 0.257
table$progrec  3.194  1 0.074
table$tgrad    7.518  2 0.023
table$tsize    0.103  1 0.748
GLOBAL        11.623  6 0.071
```

Σχήμα 5.17: Αποτελέσματα του ελέγχου υποθέσων για την αναλογική διακινδύνευση

Ο έλεγχος συγκρίνεται την τιμή της στατιστικής συνάρτησης ελέγχου με την $\mathcal{X}_5^2$ κατανομή και μας δίνει τις τιμές p-value. Φαίνεται πως για κάθε μεταβλητή μπορούμε να αποδεχθούμε την μηδενική υπόθεση εκτός ίσως από τη συμμεταβλητή για τον τύπο του όγκου($tgrad$). Επιπλέον και ο ολικός έλεγχος για το μοντέλο μας υποδεικνύει πως έχουμε αρκετές ενδείξεις για να αποδεχθούμε την μηδενική υπόθεση, πως ισχύει η αναλογική διακινδύνευση.

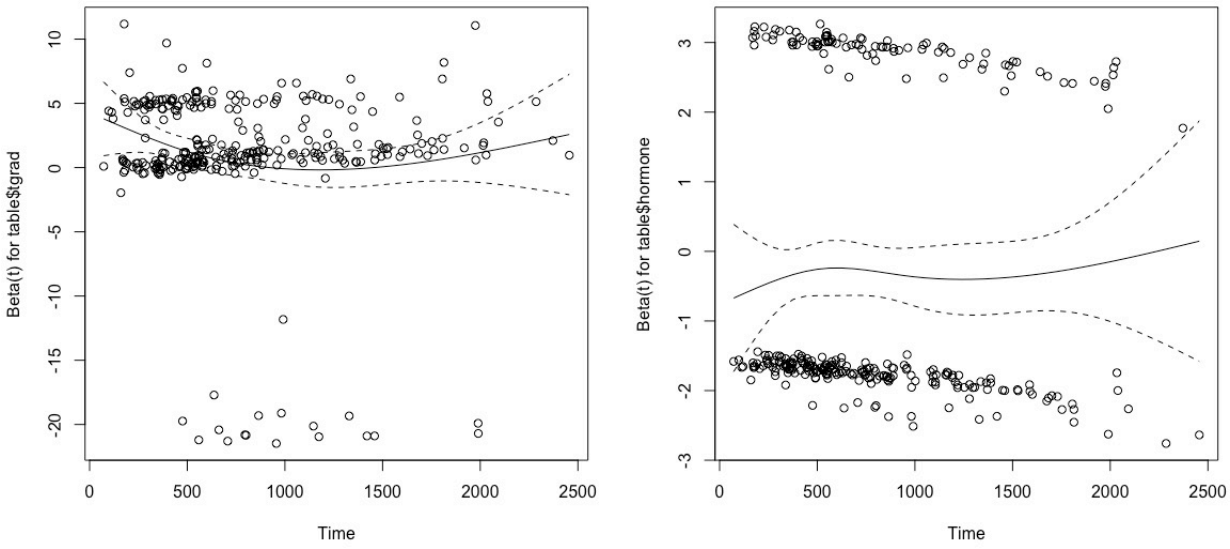
Στη συνέχεια της ανάλυσης μας θα χρησιμοποιήσουμε τη γραφική μορφή των υπολοίπων του μοντέλου, για να ελέγξουμε την υπόθεση αναλογικής διακινδύνευσης, της καλής προσαρμογής του μοντέλου στα δεδομένα μας αλλά και για να μπορέσουμε να εντοπίσουμε ακραίες τιμές και να εκτιμήσουμε τη συναρτησιακή μορφή κάποιας συμμεταβλητής.

- **Κλιμακοποιημένα Υπόλοιπα Schoenfeld**



Σχήμα 5.18: Υπόλοιπα Schoenfeld για τις συμμεταβλητές *pnodes*, *progrec* και *estrec*

Παρατηρούμε, στο Σχήμα 5.18, πως και για τις τρεις συμμεταβλητές οι γραμμές που έχουν δημιουργηθεί είναι σχεδόν οριζόντιες, γεγονός που οδηγεί στο συμπέρασμα πως και για τις τρεις μπορεί να υποστηριχθεί η υπόθεση της ανεξαρτησίας από το χρόνο και κατ΄ επέκταση της αναλογικής διακινδύνευσης.

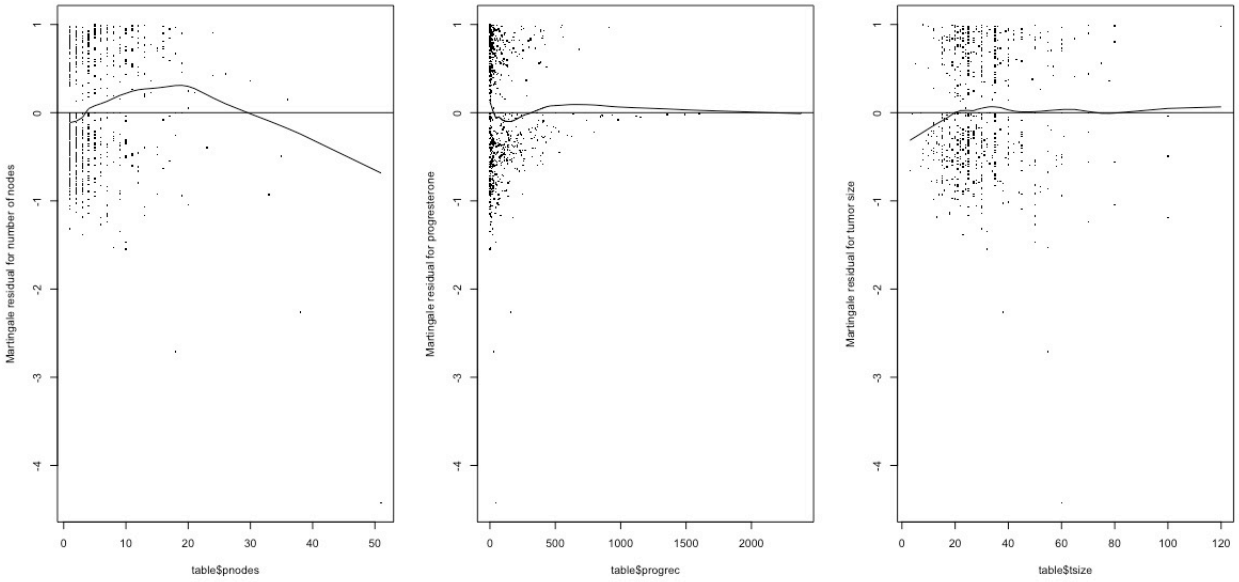


Σχήμα 5.19: Υπόλοιπα Schoenfeld για τις συμμεταβλητές *tgrad* και *hormone*

Για αυτές τις συμμεταβλητές και ιδιαίτερα για την συμμεταβλητή *tgrad* παρατηρούμε στο Σχήμα 5.19 πως η υπόθεση της αναλογικής διακινδύνευσης ίσως και να μην ισχύει καθώς παρατηρούμε μια ελαφριά καμπυλότητα και στις δύο γραφικές παραστάσεις.

- **Υπόλοιπα Martingale**

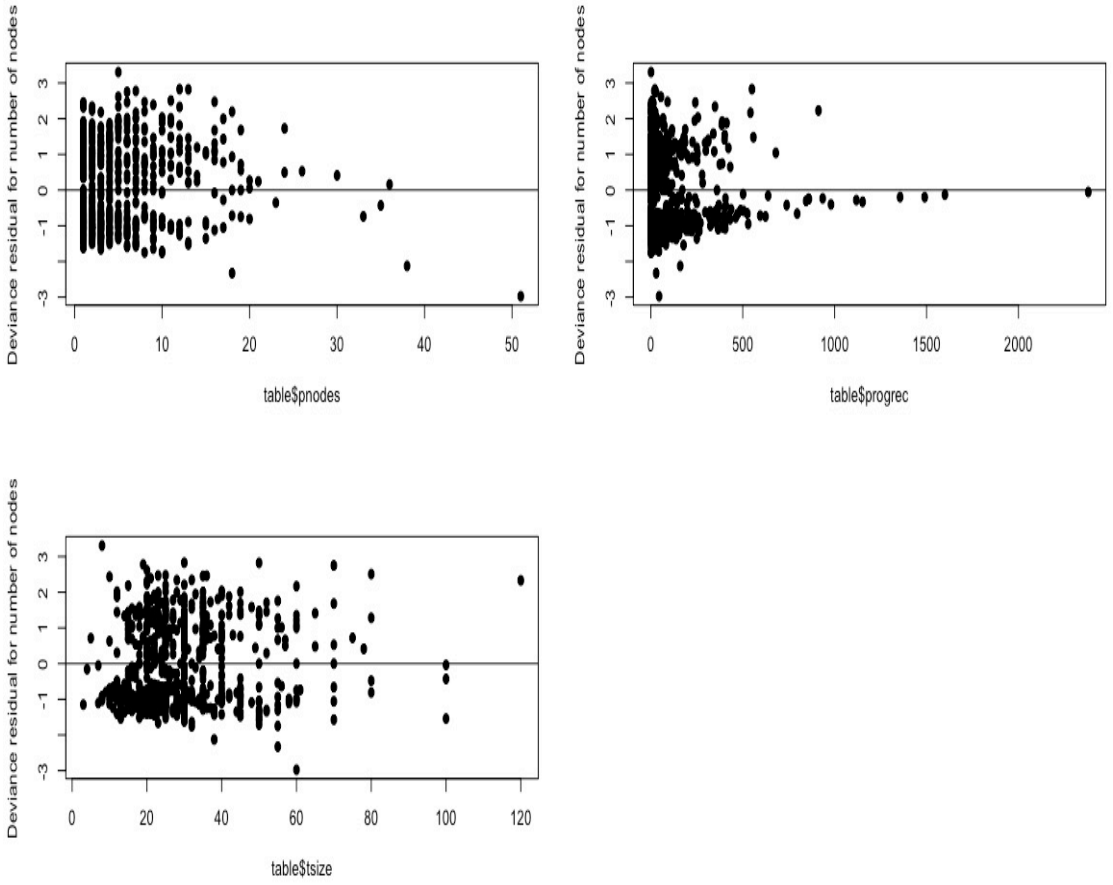
  Τα υπόλοιπα Martingale θα μας βοηθήσουν να κατανοήσουμε την συναρτησιακή μορφή των συμμεταβλητών του μοντέλου.



Σχήμα 5.20: Υπόλοιπα Martingale για τις συμμεταβλητές *pnodes*, *progrec* και *tsize*

Οι καμπύλες εξομάλυνσης, όπως φαίνονται στο Σχήμα 5.20, μας βοηθούν να καταλάβουμε πως για την συμμεταβλητή *pnodes* ίσως θα έπρεπε να σκεφτούμε μια τετραγωνική μορφή για την συναρτησιακή μορφή της συμμεταβλητής ενώ για τις άλλες δύο βλέπουμε μια σταθερή μορφή.

- **Υπόλοιπα Deviance**

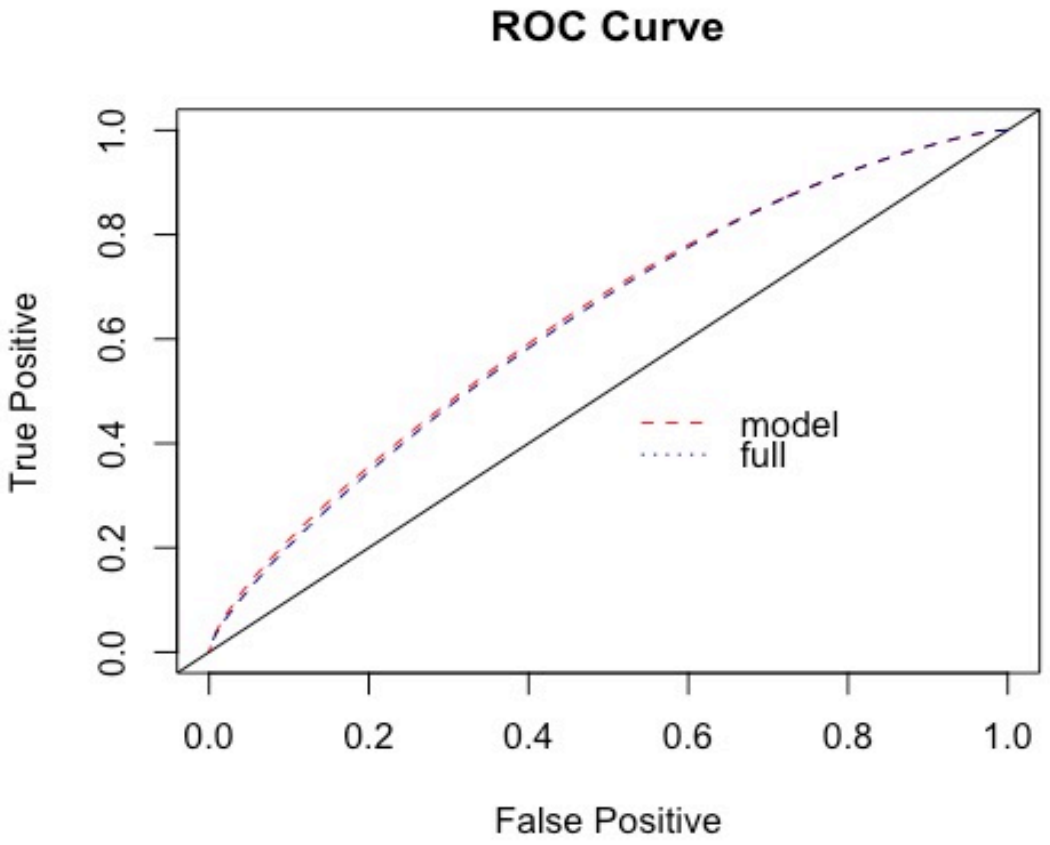


Σχήμα 5.21: Υπόλοιπα Deviance για τις συμμεταβλητές *pnodes*, *progrec* και *tsize*

Τα υπόλοιπα Deviance, μπορούν να βοηθήσουν στον εντοπισμό outliers τιμών. Όπως μπορούμε να δούμε στο Σχήμα 5.21, ιδιαίτερα στις συμμεταβλητές *pnodes* και *progrec* εντοπίζονται ακραίες τιμές.

Καμπύλη ROC

Στο Σχήμα 5.22, παραθέτεται η καμπύλη ROC για το μοντέλο στο οποίο έχουμε καταλήξει με τις συμμεταβλητές *pnodes*, *progrec*, *tsize*, *tgrad* και *hormone* και για το μοντέλο με όλες τις συμμεταβλητές.



Σχήμα 5.22: Καμπύλη ROC

Βλέπουμε πως ελαφρώς καλύτερη προβλεπτική ικανότητα έχει το μοντέλο με τις συμμεταβλητές *pnodes*, *progrec*, *tsize*, *tgrad* και *hormone*, γεγονός που συνηγορεί στο να το επιλέξουμε συγκριτικά με το μοντέλο με όλες τις συμμεταβλητές καθώς εκτός από μεγαλύτερη ακρίβεια είναι και πιο απλό και άρα πιο εύκολα ερμηνεύσιμο, πράγμα που αποζητούμε από ένα μοντέλο. Επιπλέον η προβλεπτική ικανότητα του μοντέλου θα μπορούσε να χαρακτηριστεί ικανοποιητική καθώς μπορεί να μην πλησιάζει αρκετά την επάνω αριστερή γωνία του διαγράμματος όμως αποκλίνει αρκετά από την διαγώνιο.

Επιπλέον για το μοντέλο που έχουμε επιλέξει η τιμή AUC (Area Under the Curve, του εμβαδού δηλαδή κάτω από την καμπύλη είναι 0.6389828, ενώ για το πλήρες μοντέλο

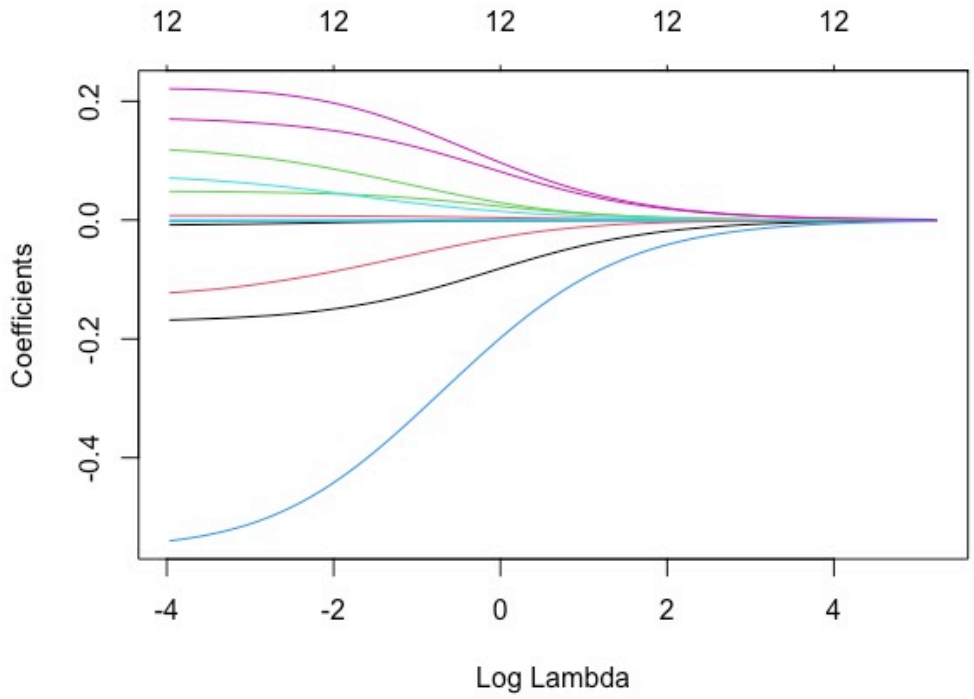
είναι 0.6414626. Παρατηρούμε πως και οι δύο τιμές είναι πολύ κοντά περίπου αφού και οι δύο μπορούν να στρογγυλοποιηθούν στο 0.64.

5.2.4 Ποινικοποιημένες Μέθοδοι Παλινδρόμησης

Σε αυτό το σημείο της ανάλυσης μας, θα χρησιμοποιήσουμε ποινικοποιημένες μεθόδους παλινδρόμησης για την προσαρμογή του μοντέλου μας. Με τον τρόπο ο οποίος έχει περιγραφεί στη θεωρία θα προσαρμόσουμε τα δεδομένα μας με την μέθοδο Ridge και Lasso. Για την υλοποίηση των μεθόδων αυτών στην R, θα γίνει χρήση του πακέτου glmnet.

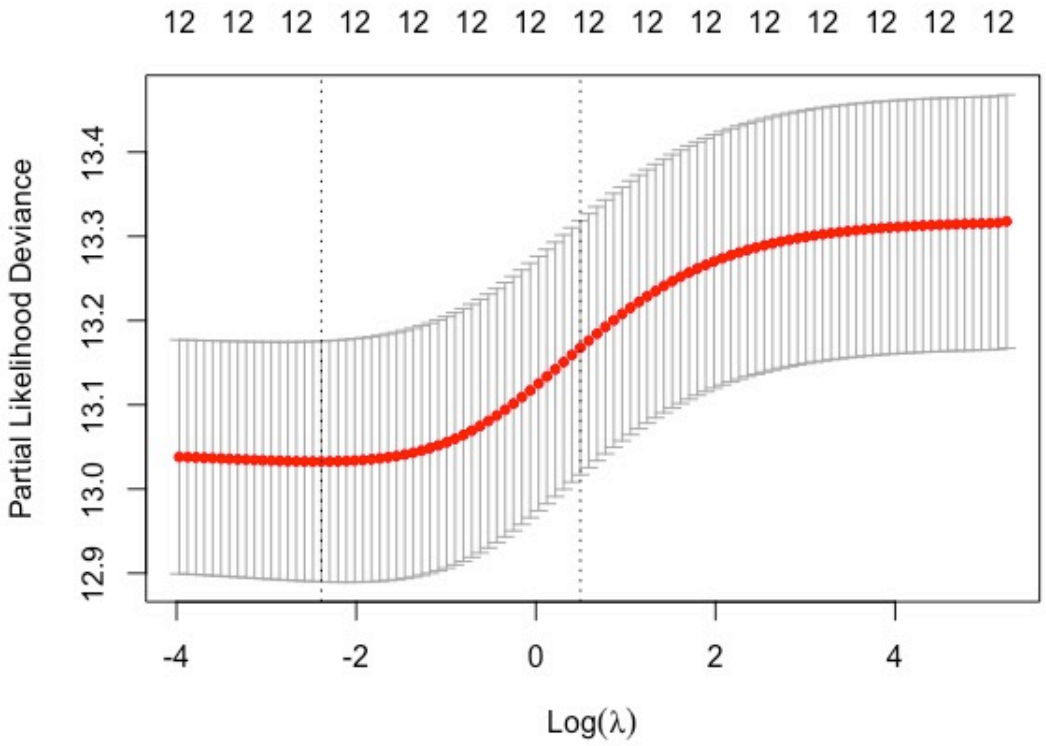
Μέθοδος Ridge

Στο Σχήμα 5.23 φαίνεται η αύξηση στην τιμή του λογαρίθμου της ποινής λ της μεθόδου και παρατηρούμε πως όσο μεγαλώνει η τιμή οι συντελεστές συρρικνώνονται προς το 0.



Σχήμα 5.23: Οι συντελεστές Ridge του μοντέλου συναρτήσει του λογαρίθμου της ποινής λ

Και πάλι με τη χρήση του πακέτου glmnet, θα εφαρμόσουμε Cross-Validation ώστε να εντοπίσουμε δύο τιμές του λ που εμφανίζουν ιδιαίτερο ενδιαφέρον. Συγκεκριμένα, την τιμή που δίνει το CV-σφάλμα με την μικρότερη τιμή και αυτή, που σύμφωνα με τον ευρέως χρησιμοποιούμενο κανόνα στην στατιστική, βασίζεται στην τυπική απόκλιση 1 από τη μικρότερη τιμή. Τα αποτελέσματα φαίνονται στο Σχήμα 5.24.



Σχήμα 5.24: Ο λογάριθμος της ποινής λ

Στη συνέχεια, στο Σχήμα 5.25, εμφανίζουμε τις δύο τιμές αυτές που μας ενδιαφέρουν για την τιμή της ποινής λ και για αυτές τις δύο θα εκτιμήσουμε του συντελεστές του μοντέλου.

```
> cv.fit2$lambda.min
[1] 0.09171503
> cv.fit2$lambda.1se
[1] 1.64046
```

Σχήμα 5.25: Οι δύο τιμές ενδιαφέροντος

Για την πρώτη περίπτωση και για την τιμή λ_{min} έχουμε τα παρακάτω αποτελέσματα, στο Σχήμα 5.26:

```
> lambda.min2
12 x 1 sparse Matrix of class "dgCMatrix"
                      1
age         -5.282788e-03
tsize        7.109285e-03
pnodes       4.558074e-02
progrec     -1.342055e-03
estrec      -6.456384e-05
hormoneno    1.569652e-01
hormoneyes  -1.562004e-01
menono      -9.658424e-02
menoyes      9.535233e-02
tgradI      -4.748445e-01
tgradII      5.277802e-02
tgradIII     2.068821e-01
```

Σχήμα 5.26: Οι συντελεστές του μοντέλου για την τιμή λ_{min}

Στην περίπτωση αυτή κανένας συντελεστής δεν έχει μηδενιστεί, πράγμα το όποιο περιμέναμε λόγω της 'φύσης' της μεθόδου. Μεγαλύτερη συρρίκνωση έχουν δεχθεί οι μεταβλητές $age, tsize, progrec$ και $estrec$. Δηλαδή για τη μέθοδο αυτή, σημαντικότερες μεταβλητές για το μοντέλο μας φαίνονται να είναι αυτές που αφορούν των αριθμό των θετικών όγκων ($pnodes$), την ορμονοθεραπεία ($hormone$), την κατάσταση της εμμηνόπαυσης ($meno$) και τον τύπο του όγκου ($tgrad$).

Για την τιμή λ_{1se} έχουμε τα παρακάτω αποτελέσματα στο Σχήμα 5.27:

```
> lambda.1se2
12 x 1 sparse Matrix of class "dgCMatrix"
                           1
age         -0.0006687555
tsize        0.0030260586
pnodes       0.0164101646
progrec     -0.0002602642
estrec      -0.0001179490
hormoneno    0.0607572074
hormoneyes  -0.0607454410
menono      -0.0189300308
menoyes      0.0189284810
tgradI      -0.1434205636
tgradII      0.0100812219
tgradIII     0.0703253394
```

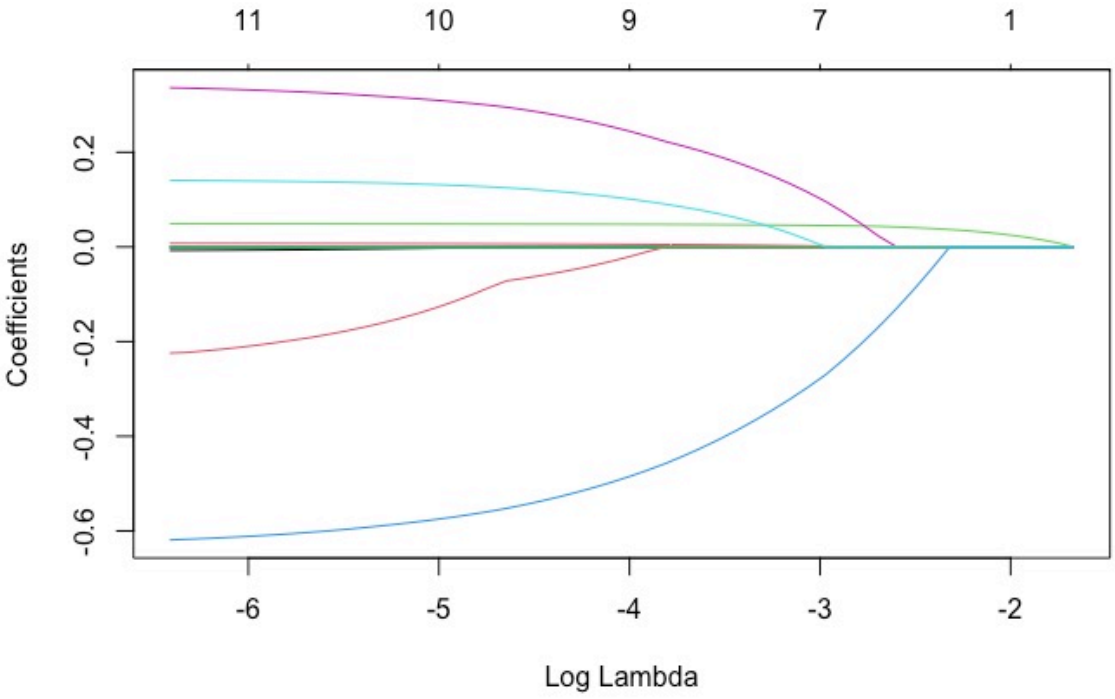
Σχήμα 5.27: Οι συντελεστές του μοντέλου για την τιμή λ_{1se}

Και στην περίπτωση αυτή παρατηρούμε πως οι σημαντικές μεταβλητές που δεν συρρικνώνονται σημαντικά και πρέπει να συμπεριληφθούν στο μοντέλο είναι ο αριθμός των θετικών όγκων ($pnodes$), η ορμονοθεραπεία ($hormone$), η κατάσταση της εμμηνόπαυσης ($meno$) και ο τύπος του όγκου ($tgrad$).

Σημειώνεται πως και οι δύο τιμές συμφωνούν στις σημαντικές μεταβλητές για το μοντέλο. Βλέπουμε επίσης πως υπάρχει συμφωνία με την ανάλυση που έχει γίνει με το μοντέλο του Cox πως οι συμμεταβλητές $pnodes$, $hormone$ και $tgrad$ είναι σημαντικές και χρειάζονται στο μοντέλο.

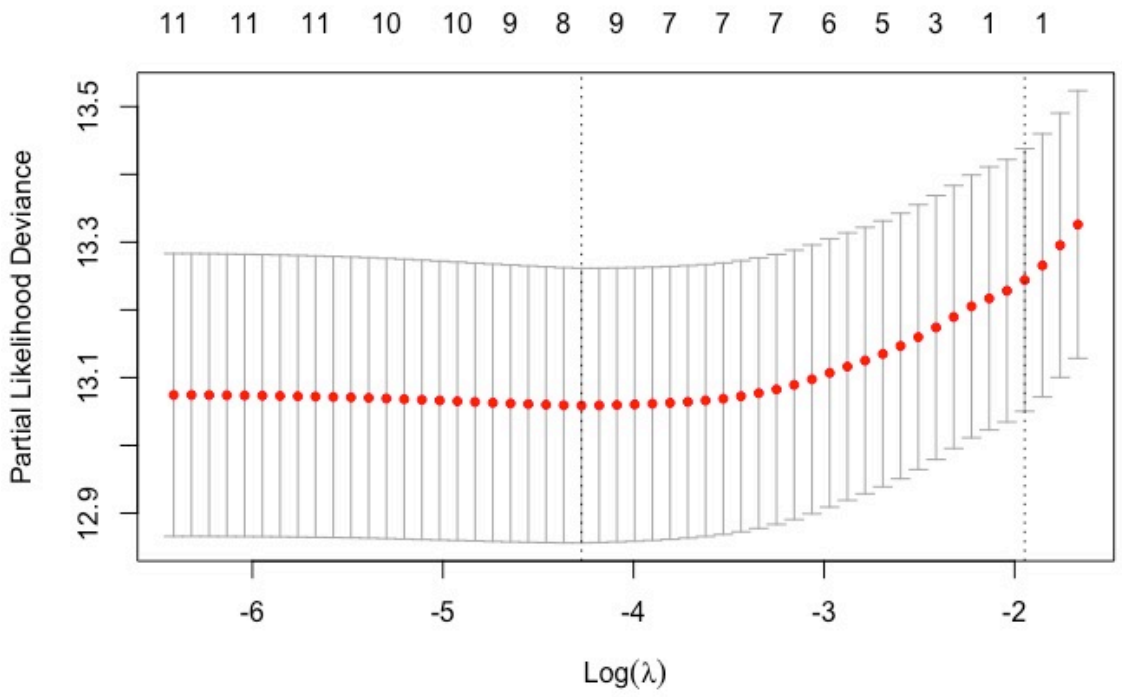
Μέθοδος Lasso

Στη συνέχεια αυτή τη φορά για τη μέθοδο Lasso παρατίθεται το γράφημα των εκτιμητριών Lasso των συντελεστών σε συνάρτηση με τον λογάριθμο της τιμής της ποινής λ, στο Σχήμα 5.28.



Σχήμα 5.28: Οι συντελεστές Lasso του μοντέλου συναρτήσει του λογαρίθμου της ποινής λ

Όσο ο λογάριθμος αυξάνεται οι συντελεστές συρρικνώνονται προς το μηδέν μέχρις ότου γίνουν όλοι μηδέν. Στο πάνω μέρος του γραφήματος φαίνονται οι μεταβλητές του μοντέλου μας, οι οποίες μειώνονται σταδιακά. Στόχος είναι να βρούμε του συντελεστές εκείνους που χρειάζονται πραγματικά στο μοντέλο και η μέθοδος να μηδενίσει τους υπόλοιπους. Και πάλι θα κάνουμε χρήση του πακέτου glmnet και θα εφαρμόσουμε Cross-Validation ώστε να εντοπίσουμε δύο τιμές του λ, την τιμή που δίνει την μικρότερη τιμή CV-σφάλματος και αυτή και αυτή που βασίζεται στην τυπική απόκλιση 1 από τη μικρότερη τιμή. Τα αποτελέσματα φαίνονται στο Σχήμα 5.29:



Σχήμα 5.29: Οι δύο τιμές ενδιαφέροντος

Παρακάτω, στο Σχήμα 5.30, βλέπουμε τις δύο τιμές που μας ενδιαφέρουν για την τιμή της ποινής λ και για αυτές τις δύο θα εκτιμήσουμε του συντελεστές Lasso του μοντέλου.

```
> cv.fit1$lambda.min
[1] 0.01393987
> cv.fit1$lambda.1se
[1] 0.1426789
```

Σχήμα 5.30: Οι δύο τιμές ενδιαφέροντος

Στο Σχήμα 5.31, όπου έχουν υπολογιστεί οι συντελεστές του μοντέλου βάσει της τιμής λ_{min}, βλέπουμε πως η μέθοδος πως αναμέναμε έχει θέσει κάποιους συντελεστές ίσους με το μηδέν. Συγκεκριμένα την μεταβλητή για την ηλικία (age) και την μεταβλητή για τα οιστρογόνα ($estrec$), γεγονός που μας υποδεικνύει πως δεν χρειάζονται στο μοντέλο. Επιπλέον μηδενίζονται οι συντελεστές της συμμεταβλητής που δείχνει αν μια ασθενής είναι στην φάση της εμμηνόπαυσης καθώς και αυτή που δείχνει αν μια ασθενής είχε όγκου τύπου II. Στο μοντέλο αυτό εμπεριέχονται όλες οι συμμεταβλητές που συμπεριλήφθηκαν στο μοντέλο του Cox μαζί με αυτή τη συμμεταβλητή που δείχνει ότι η ασθενής βρίσκεται πριν την εμμηνόπαυση.

```
> lambda.min
12 x 1 sparse Matrix of class "dgCMatrix"
                      1
age          .
tsize        5.977080e-03
pnodes       4.794611e-02
progrec     -1.765502e-03
estrec       .
hormoneno    2.698786e-01
hormoneyes  -2.925243e-14
menono      -4.628977e-02
menoyes      .
tgradI      -5.179319e-01
tgradII      .
tgradIII     1.134232e-01
```

Σχήμα 5.31: Οι συντελεστές του μοντέλου για την τιμή λ_{min}

Με την παρακάτω τιμή του λ, όπως φαίνεται στο Σχήμα 5.32, μηδενίζονται όλοι οι συντελεστές εκτός από την μεταβλητή για τον αριθμό των θετικών όγκων, γεγονός που δείχνει πως ίσως είναι η μεταβλητή που επηρεάζει παραπάνω απ΄ όλες τον χρόνο επιβίωσης ασθενών.

```
> lambda.1se
12 x 1 sparse Matrix of class "dgCMatrix"
                    1
age         .
tsize       .
pnodes      0.02210356
progrec     .
estrec      .
hormoneno   .
hormoneyes  .
menono      .
menoyes     .
tgradI      .
tgradII     .
tgradIII    .
```

Σχήμα 5.32: Οι συντελεστές του μοντέλου για την τιμή λ_{1se}

Και με την μέθοδο Lasso οι συντελεστές που φαίνονται να είναι σημαντικοί και να χρειάζονται οπωσδήποτε στο μοντέλο είναι οι συμμεταβλητές για τον αριθμό των θετικών όγκων ($pnodes$), για το αν οι ασθενείς έχουν υποβληθεί σε ορμονοθεραπεία ($hormone$) και για το βαθμό του όγκου ($tgrad$).

Συμπερασματικά παρατηρούμε πως και για τις δύο ποινικοποιημένες μεθόδους οι μεταβλητές που παρουσιάζουν το μεγαλύτερο ενδιαφέρον και θα έπρεπε οπωσδήποτε να συμπεριληφθούν στο μοντέλο είναι οι μεταβλητές για τον αριθμό των θετικών όγκων των ασθενών ($pnodes$), η μεταβλητή που δείχνει αν έχουν υποβληθεί σε ορμονοθεραπεία ($hormone$) και αυτή που δείχνει τον τύπο του όγκου ($tgrad$).

5.2.5 Δένδρο Επιβίωσης

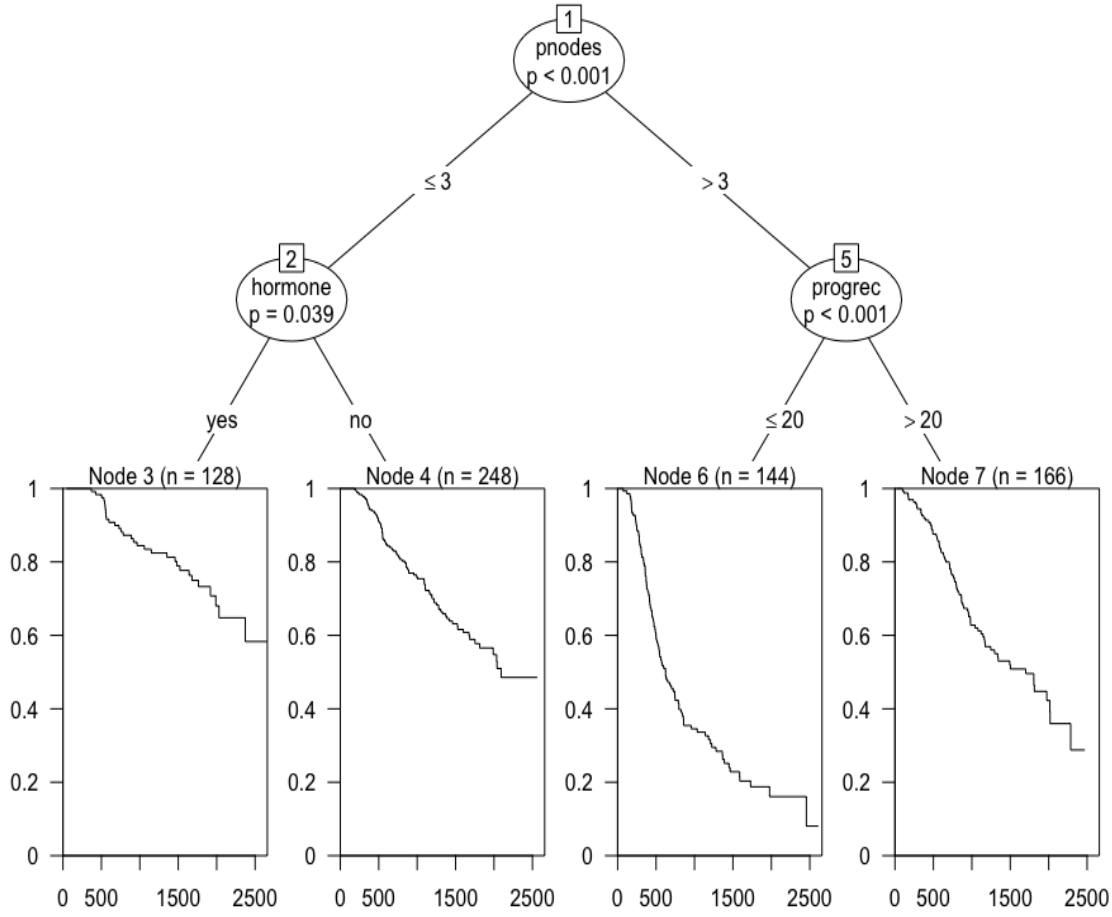
Όπως αναφέρθηκε σε προηγούμενο κεφάλαιο ένας τρόπος για να μπορέσουμε να εκτιμήσουμε τους σημαντικούς παράγοντες που επηρεάζουν την υγεία ενός ασθενή και να προβλέψουμε την πορεία της είναι να χρησιμοποιήσουμε αντί για κάποιο μοντέλο παλινδρόμησης, ένα δένδρο παλινδρόμησης και συγκεκριμένα ένα δένδρο επιβίωσης.

Στην προκειμένη περίπτωση βασιζόμενοι σε όσα έχουν αναλυθεί στην θεωρία, θα χρησιμοποιήσουμε το πακέτο 'party' της R για να δημιουργήσουμε το δένδρο παλινδρόμησης για τα δεδομένα του δείγματος μας.

Με τον τρόπο αυτό ευελπιστούμε να καταλήξουμε σε μια πρόβλεψη για την πορεία της υγείας των ασθενών του δείγματος μας και συγκεκριμένα στις ημέρες έως να υποτροπιάσουν. Θέλουμε να δημιουργήσουμε ένα απλό μοντέλο, με λίγα τερματικά φύλλα, έτσι ώστε να έχουμε στη διάθεση μας ένα εύκολα ερμηνεύσιμο μοντέλο για τους σημαντικούς παράγοντες που οδηγούν στην υποτροπή της ασθένειας.

Όπως περιγράφηκε παραπάνω, σε κάθε διαχωρισμό διεξάγεται ένας έλεγχος υποθέσεων για να γίνει ο διαμερισμός του χώρου των επεξηγηματικών μεταβλητών. Στην προκειμένη περίπτωση το πακέτο "party", χρησιμοποιεί κάποιους 'ελέγχους μεταθέσεων' σε συνδυασμό με κάποιους Log-Rank ελέγχους έτσι ώστε να καταλήξει στην πρώτη μεταβλητή που θα χρησιμοποιηθεί για το πρώτο διαχωρισμό και στη συνέχεια με τον τρόπο που έχει αναφερθεί θα προχωρήσει και στα υπόλοιπα 'χωρίσματα' τα οποία θέλουμε να είναι όσο το δυνατόν λιγότερα.

Παρακάτω λοιπόν, στο Σχήμα 5.33 παρατίθεται το δένδρο παλινδρόμησης που υλοποιήθηκε από το πακέτο της R "party". Η μελέτη που χρησιμοποιήθηκε αφορά 686 γυναίκες που πάσχουν από καρκίνο του μαστού και βάσει αυτών θα προσπαθήσουμε να προχωρήσουμε σε μια προγνωστική μοντελοποίηση του προβλήματος.



Σχήμα 5.33: Το δένδρο παλινδρόμησης για τα δεδομένα του δείγματος.

Βλέπουμε πως το δένδρο που δημιουργήθηκε έχει τέσσερα τερματικά φύλλα. Όπως παρατηρούμε οι παράγοντες που διαμερίζουν τον χώρο των επεξηγηματικών μεταβλητών είναι πρωταρχικά ο αριθμός των θετικών όγκων που έχει κάθε ασθενής και στη συνέχεια οι υποδοχείς προγεστερόνης καθώς επίσης σημαντικό ρόλο παίζει και η μεταβλητή που δείχνει αν οι ασθενείς έχουν υποβληθεί σε θεραπεία ορμονών ή όχι. Επιπλέον στο γράφημα του δένδρου που δημιουργήθηκε, αναπαριστώνται και οι γραφικές παραστάσεις των εκτιμήσεων Kaplan-Meier, για κάθε ομάδα ασθενών.

Καταλήγουμε λοιπόν στα παρακάτω συμπεράσματα:

- Οι ασθενείς που έχουν λιγότερους από τρεις θετικούς όγκους και έχουν υποβληθεί σε θεραπεία με ορμόνες έχουν τις καλύτερες πιθανότητες επιβίωσης από όλες τις υπόλοιπες.

- Εν αντιθέσει, οι ασθενείς που έχουν παραπάνω από τρεις θετικούς όγκους και τα επίπεδα των υποδοχέων προγεστερόνης τους είναι λιγότερο από 20 αντιμετωπίζουν το μεγαλύτερο ρίσκο και έχουν τις πιο δυσμενείς προβλέψεις επιβίωσης.

- Αρκετά καλές πιθανότητες επιβίωσης, έχουν επίσης οι ασθενείς που έχουν λιγότερους από τρεις θετικούς όγκους ακόμα και αν δεν έχουν υποβληθεί σε ορμοθεραπεία.

- Δύσκολη είναι επίσης η κατάσταση για τις ασθενείς με περισσότερους από τρεις θετικούς όγκους, που έχουν επίπεδα προγεστερόνης μεγαλύτερα από 20.

Το δένδρο παλινδρόμησης που δημιουργήθηκε συμφωνεί με το μοντέλο του Cox και με τις ποινικοποιημένες μεθόδους Ridge και Lasso πως οι μεταβλητές *pnodes* και *hormone* είναι σημαντικές και χρειάζονται σίγουρα στο μοντέλο μας καθώς επίσης υποδεικνύει πως σημαντικό ρόλο στην πρόβλεψη της πορείας της υγείας των ασθενών παίζει και η μεταβλητή *progrec*, αποτέλεσμα που συμφωνεί με το μοντέλο του Cox.

5.3 Συμπεράσματα

Εν κατακλείδι, για το σύνολο 686 ασθενών που πάσχουν από καρκίνο του μαστού καταλήγουμε στα παρακάτω:

- Για το μοντέλο του Cox, σημαντικότερες φαίνεται να είναι οι μεταβλητές *hormone*, *tsize*, *pnodes*, *progrec*, και *tgrad*. Συγκεκριμένα παρατηρούμε πως οι ασθενείς που έχουν υποβληθεί σε ορμονοθεραπεία και έχουν όγκο τύπου I αντιμετωπίζουν μειωμένο κίνδυνο υποτροπής συγκριτικά με τις υπόλοιπες ασθενείς.

- Χρησιμοποιώντας τις ποινικοποιημένες μεθόδους Ridge και Lasso, καταλήγουμε πως στο μοντέλο χρειάζονται οι μεταβλητές *pnodes*, *hormone* και *tgrad*. Γεγονός που συνηγορεί και με τα αποτελέσματα από το μοντέλο του Cox, πως ο αριθμός των θετικών όγκων, το αν μια ασθενής έχει υποβληθεί σε ορμονοθεραπεία ή όχι και ο τύπος του όγκου που έχει κάθε ασθενής επηρεάζει σημαντικά την εξέλιξη της ασθένειας και πρέπει να συμπεριληφθούν στο μοντέλο.

- Τέλος, δημιουργώντας ένα δένδρο παλινδρόμησης καταλήγουμε πως κυριότερη μεταβλητή για να διαχωρίσουμε την πορεία της υγείας των ασθενών είναι αυτή του αριθμού των θετικών όγκων, *pnodes*. Σημαντικό ρόλο παίζουν επίσης οι μεταβλητές για την ορμονοθεραπεία, *hormone*, καθώς και για τα επίπεδα προγεστερόνης, *progrec*. Σημειώνεται πως τις καλύτερες πιθανότητες επιβίωσης έχουν οι ασθενείς που έχουν υποβληθεί σε θεραπεία με ορμόνες και έχουν λιγότερους από τρεις θετικούς όγκους.

Βιβλιογραφία

Καρκίνος μαστού. (2022). *https* : *//almazois.gr/*. Πανελλήνιος Σύλλογος Γυναικών με Καρκίνο του Μαστού.

Καρώνη Χ. (2009). *Μοντέλα Αξιοπιστίας και Επιβίωσης.* Εκδόσεις ΣΥΜΕΩΝ.

Καρώνη Χ., Οικονόμου Π. (2017). *Στατιστικά Μοντέλα Παλινδρόμησης, Με χρήση Minitab και R.* Εκδόσεις ΣΥΜΕΩΝ.

Κοκολάκης Γ., Φουσκάκης Δ. (2009). *Στατιστική Θεωρία και Εφαρμογές.* Εκδόσεις Συμεών.

Φουσκάκης Δ. (2013). *Ανάλυση Δεδομένων με Χρήση της R.* Εκδόσεις Τσότρας.

References

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, *6*(4), 701 – 726.

Akaike, H. (1998). *Information Theory and An Extension of the Maximum Likelihood Principle*. Springer.

Benner, A., Zucknick, M., Hielscher, T., Ittrich, C., & Mansmann, U. (2010). High-dimensional Cox models: The choice of penalty as part of the model building process. *Biometrical Journal*, *52*, 50-69.

Breast cancer. (2021, 26 March). `https://www.who.int/news-room/fact-sheets/detail/breast-cancer`. World Health Organization.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, *30*(1), 89–99.

Cai, Z., & Sun, Y. (2003). Local linear estimation for time-dependent coefficients in Cox's regression models. *Scandinavian Journal of Statistics*, *30*(1), 93-111.

Collett, D. (2015). *Modelling Survival Data in Medical Research*. Chapman and Hall/CRC.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, *62*(2), 269–276.

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

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, *72*(359), 557–565.

Everitt, B. S., & Hothorn, T. (2009). *A Handbook of Statistical Analyses Using R, Second Edition*. Chapman & Hall/CRC.

Fleming, T. R., & Harrington, D. P. (2011). *Counting Processes and Survival Analysis*. John Wiley & Sons.

Grambsch, P. M., & Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, *81*(3), 515–526.

Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, *55*(4), 757–796.

Heagerty, P. J., & Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, *61*(1), 92-105.

Hoerl, A. E., & Kennard, R. W. (1970a). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, *12*(1), 69–82.

Hoerl, A. E., & Kennard, R. W. (1970b). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Hothorn, T., Hornik, K., & Zeileis, A. (2008, 01). Party: A laboratory for recursive part(y)itioning. *R package version 0.9-0*.

Hothorn, T., K.Hornik, & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651-674.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R*. Springer.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*(282), 457–481.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, *14*(4), 945–966.

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

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

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

van Houwelingen, J. H., & Sauerbrei, W. (2013, 01). Cross-validation, shrinkage and variable selection in linear regression revisited. *Open Journal of Statistics*, *03*, 79-102.

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492-514.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *67*(2), 301–320.

Παράρτημα Α΄

Κώδικας στην R

```
table<-read.table("breastcan.txt",header=TRUE, colClasses =
c("numeric","numeric","numeric","numeric","numeric","numeric","factor","factor","factor
str(table)
attach(table)

levels(table$hormone) <- c("no","yes")
levels(table$meno) <- c("no","yes")
levels(table$tgrad) <- c("I","II","III")

table(hormone)
table(cens)
f1<-table(table$hormone)
f2<-table(table$meno)
f3<-table(table$tgrad)

par(mfrow=c(1,3))
barplot(f1,xlim=c(0,5),ylim=c(0,500),col="darkseagreen4",main="Ραβδόγραμμα
για την ορμονοθεραπεία")
barplot(f2,xlim=c(0,5),ylim=c(0,500),col="darkseagreen3",main="Ραβδόγραμμα
για το στάδιο της εμμηνόπαυσης")
barplot(f3,xlim=c(0,5),ylim=c(0,500),col="darkcyan",main="Ραβδόγραμμα για
τον βαθμό του όγκου")

library(survival)
fit.surv <-survfit(Surv(table$time,table$cens)~1)
plot(fit.surv ,xlab="days",ylab="estimated probability of survival")
fit.surv

fit.meno <-survfit(Surv(table$time, table$cens)~table$meno)
plot(fit.meno,xlab = "days",ylab = "estimated probability of survival", col
= c("darkseagreen4","gold"),  lwd = 1.5)
legend("bottomleft", legend=c("pre","post"), col =
c("darkseagreen4","gold"), lty = 1, title="Menopausal Status")
fit.meno

fit.hormone <-survfit(Surv(table$time, table$cens)~table$hormone)
fit.hormone
plot(fit.hormone,xlab = "days",ylab = "estimated probability of survival",
col = c("darkgray","brown4"), lwd = 1.5)
legend("bottomleft", legend=c("no","yes"), col = c("darkgray","brown4"),
lty = 1, title="Hormonal Therapy")

fit.tgrad<-survfit(Surv(table$time, table$cens)~table$tgrad)
fit.tgrad
plot(fit.tgrad,xlab = "days",ylab = "estimated probability of survival",
col = c("thistle4","violetred4","royalblue4"), lwd = 1.5)
legend("bottomleft", legend=c("I","II","III"), col =
c("darkgray","violetred4","royalblue4"), lty = 1, title="Tumor grade")

agecat<-cut(table$age, breaks= 2)
levels(agecat) <- c("20-50","51-80")
fit.age<-survfit(Surv(table$time, table$cens)~agecat)
fit.age
```

```
plot(fit.age,xlab = "days",ylab = "estimated probability of survival", col
= c("violetred4","royalblue4"), lwd = 1.5)
legend("bottomleft", legend=c("20-50","51-80"), col =
c("violetred4","royalblue4"), lty = 1, title="Age Groups")

logrank.test <- survdiff(Surv(table$time,table$cens)~table$meno)
logrank.test
logrank.test <- survdiff(Surv(table$time,table$cens)~table$hormone)
logrank.test

logrank.test <- survdiff(Surv(table$time,table$cens)~table$tgrad)
logrank.test
logrank.test <- survdiff(Surv(table$time,table$cens)~agecat)
logrank.test

fit.cox <- coxph(Surv(table$time, table$cens) ~ table$hormone)
summary(fit.cox)

fit.all <- coxph(Surv(table$time,table$cens)~ table$hormone + table$age +
table$tsize + table$pnodes + table$progrec + table$estrec + table$meno +
table$tgrad)
fit.all
summary(fit.all)

fit.null<- coxph(Surv(table$time,table$cens)~ 1)
fit.null
summary(fit.null)

fit.m1<- coxph(Surv(table$time,table$cens)~ table$hormone +table$pnodes +
table$progrec + table$tgrad, data=table)
fit.m1
summary(fit.m1)

fit.m2<- coxph(Surv(table$time,table$cens)~ table$hormone +table$pnodes +
table$progrec + table$tgrad + table$meno)
fit.m2
summary(fit.m2)

fit.m3<- coxph(Surv(table$time,table$cens)~ table$hormone +table$pnodes +
table$progrec + table$tgrad + table$estrec)
fit.m3
summary(fit.m3)

fit.m4<- coxph(Surv(table$time,table$cens)~ table$hormone +table$pnodes +
table$progrec + table$tgrad + table$tsize)
fit.m4
summary(fit.m4)

fit.m5<- coxph(Surv(table$time,table$cens)~ table$hormone +table$pnodes +
table$progrec + table$tgrad + table$age)
fit.m5
summary(fit.m5)
```

```
coxstep<-step(fit.all,direction = "both")

fit.model<- coxph(Surv(table$time,table$cens)~ table$hormone +table$pnodes
+ table$progrec + table$tgrad + table$tsize)
fit.model
summary(fit.model)

res<-cox.zph(fit.model, transform="identity", global=TRUE)
res
par(mfrow=c(1,3))
plot1<-plot(res, var="table$pnodes")
plot2<-plot(res, var="table$tsize")
plot3<-plot(res, var="table$progrec")
par(mfrow=c(1,2))
plot4<-plot(res, var="table$tgrad")
plot5<-plot(res, var="table$hormone")

model.schoef1<-residuals(fit.model,type="schoenfeld")
plot(table$pnodes[table$cens==1],model.schoef1[,2],,ylab="Schoefeld
residual for number of nodes")
abline(h=0)

plot(table$progrec[table$cens==1],model.schoef1[,3],,ylab="Schoefeld
residual for progresterone")
abline(h=0)

plot(table$tsize[table$cens==1],model.schoef1[,6],,ylab="Schoefeld residual
for tumor size")
abline(h=0)

model.mart<-residuals(fit.model,type="martingale")

par(mfrow=c(2,2))
plot(table$pnodes,model.mart,pch=".",ylab="Martingale residual for number
of nodes")
abline(h=0)
lines(lowess(table$pnodes,model.mart))

plot(table$progrec,model.mart,pch=".",ylab="Martingale residual for
progresterone")
abline(h=0)
lines(lowess(table$progrec,model.mart))

plot(table$tsize,model.mart,pch=".",ylab="Martingale residual for tumor
size")
abline(h=0)
lines(lowess(table$tsize,model.mart))

model.dev<-residuals(fit.model,type="deviance")

par(mfrow=c(2,2))
plot(table$pnodes,model.dev,pch=19,ylab="Deviance residual for number of
nodes")
```

```
abline(h=0)
plot(table$progrec,model.dev,pch=19,ylab="Deviance residual for number of
nodes")
abline(h=0)
plot(table$tsize,model.dev,pch=19,ylab="Deviance residual for number of
nodes")
abline(h=0)


library(risksetROC)
eta1<-fit.model$linear.predictor
eta2<-fit.all$linear.predictor
ROC1=risksetROC(Stime= table$time, status=table$cens,marker=eta1,
predict.time=1000, method="Cox",
                main="ROC Curve", lty=2, col="red", ylab="True Positive",
xlab="False Positive")

ROC2=risksetROC(Stime= table$time, status=table$cens,
                marker=eta2, predict.time=1000,
method="Cox",col="blue",plot=FALSE)

lines(ROC1$FP,ROC2$TP, lty=2,col="darkblue")
legend(.5,.5,lty=c(2,3),col=c("red","darkblue"), legend=c("model","full"),
bty="n")

tmax=2600
AUC1<-risksetAUC(Stime= table$time, status=table$cens,
           marker=eta1, method="Cox", tmax=tmax,
           main="AUC Curve", lty=2, col="red")

data<-(subset(table, select = - c(1,10,
11),colClasses=c("numeric","numeric","numeric","numeric","numeric","factor","factor","f
levels(data$hormone) <- c("no","yes")
levels(data$meno) <- c("no","yes")
levels(data$tgrad) <- c("I","II","III")
str(data)
data1<-makeX(data)

library(glmnet)

set.seed(1)
fit1<- glmnet(data1, Surv(table$time,table$cens), family =
"cox",maxit=1000)
plot(fit1)
plot(fit1, xvar="lambda")
cv.fit1 <- cv.glmnet(data1, Surv(table$time, table$cens), family = "cox",
alpha = 1, maxit = 1000)
plot(cv.fit1)
cv.fit1$lambda.min
cv.fit1$lambda.1se
lambda.min<-coef(cv.fit1,s="lambda.min")
lambda.1se<-coef(cv.fit1,s="lambda.1se")
```

```
fit2<- glmnet(data1, Surv(table$time,table$cens), family =
"cox",maxit=1000, alpha = 0)
plot(fit2,xvar="lambda")
cv.fit2 <- cv.glmnet(data1, Surv(table$time, table$cens), family = "cox",
alpha = 0, maxit = 1000)
plot(cv.fit2)
cv.fit2$lambda.min
cv.fit2$lambda.1se
lambda.min2<-coef(cv.fit2,s="lambda.min")
lambda.1se2<-coef(cv.fit2,s="lambda.1se")

install.packages("party")
stree1 <-ctree(Surv(time, cens) ~ ., data = table)
plot(stree1)
```