ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Χρήση μεθόδων με ποινή σε μοντέλα παλινδρόμησης αναλογικής διακινδύνευσης

του φοιτητή
Ρουσή Δημητρίου

Επιβλέπουσα: Καρώνη Χρυσηίς,
Καθηγήτρια Ε.Μ.Π

Τριμελής επιτροπή:

Χ. Καρώνη	**Φ. Βόντα**	**Β. Παπανικολάου**
Καθηγήτρια ΕΜΠ	**Αναπλ. Καθηγήτρια ΕΜΠ**	**Καθηγητής ΕΜΠ**

Αθήνα, 2018

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω την καθηγήτρια του Ε.Μ.Π, κα Χρυσηίδα Καρώνη για την ανάθεση, επίβλεψη και διαρκή καθοδήγησή της σε όλα τα στάδια εκπόνησης της παρούσας εργασίας.
Επιπλέον, θα ήθελα να ευχαριστήσω την οικογένεια μου για την αμέριστη συμπαράσταση που μου έδειξε καθ' όλη τη διάρκεια συγγραφής αλλά και γενικότερα των σπουδών μου.
Τέλος, ξεχωριστή αναφορά θα ήθελα να κάνω στον πατέρα μου, ο οποίος ήταν πάντα δίπλα μου στις δύσκολες στιγμές και με ενέπνεε μέσα από τις δυσκολίες να συνεχίζω να πιστεύω στα όνειρα μου.

Περίληψη

Η παρούσα διπλωματική εργασία ασχολείται με τη στατιστική ανάλυση δεδομένων διάρκειας ζωής κάνοντας χρήση μιας ειδικής κατηγορίας μοντέλων στα οποία έχει επιβληθεί κάποιος συγκεκριμένος περιορισμός για τις εκτιμήσεις των παραμέτρων του -μια «ποινή» όπως συχνά αναφέρεται στη βιβλιογραφία.

Πιο αναλυτικά, το πρώτο κεφάλαιο αναφέρεται στις βασικές έννοιες του κλάδου της Ανάλυσης Αξιοπιστίας και Επιβίωσης (δεδομένα διάρκειας ζωής, συνάρτηση επιβίωσης, συνάρτηση διακινδύνευσης, σωρευτική συνάρτηση διακινδύνευσης κλπ). Επιπλέον, παραθέτονται στοιχεία από την μη-παραμετρική ανάλυση δεδομένων διάρκειας ζωής (εκτιμήτρια Kaplan-Meier, εκτιμήτρια Nelson-Aalen, μη παραμετρικός έλεγχος Log-rank, γραφικοί έλεγχοι).

Το δεύτερο κεφάλαιο χωρίζεται σε δύο σκέλη. Στο πρώτο σκέλος, γίνεται εκτενής παρουσίαση του μοντέλου αναλογικής διακινδύνευσης (προσαρμογή μοντέλου, γραφικός έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης, ορισμός υπολοίπων, κριτήρια επιλογής μεταβλητών, μέτρα καλής προσαρμογής του μοντέλου κλπ). Στο δεύτερο σκέλος εισάγεται μια ειδική κατηγορία μοντέλου αναλογικής διακινδύνευσης, το οποίο θα προσαρμοστεί στο σύνολο δεδομένων μας, το οποίο είναι το μοντέλο του Cox.

Στο τρίτο κεφάλαιο παρουσιάζεται η έννοια της μεθόδου ποινής, για την αντιμετώπιση των προβλημάτων της πολυσυγγραμμικότητας (multicollinearity) και της υπερπροσαρμογής (overfitting) μοντέλου. Γίνεται αναφορά στα είδη αυτών των μεθόδων, που έχουν αναπτυχθεί πολύ τα τελευταία χρόνια, και αναλύονται οι μέθοδοι Ridge, Lasso και (naïve) Elastic Net, που θα χρησιμοποιηθούν μετέπειτα και στο πρόβλημά μας. Το κεφάλαιο ολοκληρώνεται με την παρουσίαση της μεθόδου cross-validation για την επιλογή του βέλτιστου συντελεστή λ που χρησιμοποιείται σε όλες τις τεχνικές με ποινή.

Στο τέταρτο, και τελευταίο κεφάλαιο, μελετάται ένα σύνολο δεδομένων διάρκειας ζωής από ασθενείς που πάσχουν από οξεία μυελοπλαστική λευχαιμία (acute myeloblastic leukaemia) στο οποίο εφαρμόζονται όλες οι προαναφερθείσες μεθόδους. Πιο συγκεκριμένα, γίνεται μια πρώτη μη-παραμετρική ανάλυση των δεδομένων, προσαρμόζεται το κλασικό μοντέλο του Cox, πραγματοποιούνται όλοι οι κατάλληλοι έλεγχοι υποθέσεων και τέλος εφαρμόζονται όλες οι τεχνικές ποινής που αναλύθηκαν ελέγχοντας αν και κατά πόσο αυτές επηρεάζουν τα προηγούμενα αποτελέσματα που εξήχθησαν. Για την ανάλυση των δεδομένων χρησιμοποιήθηκαν τα στατιστικά πακέτα της R και του Minitab.

Abstract

The current thesis deals with the statistical analysis of lifetime data making use of a special category of models in which has been imposed some concrete restriction for the estimates of its parameters - a “penalty”, as it is often cited in the bibliography.

More specifically, the first chapter refers to the basic principles of Reliability and Survival Analysis (lifetime data, survival function, hazard function, cumulative hazard function etc). Moreover, techniques from the not-parametric lifetime data analysis are represented (Kaplan-Meier estimator, Nelson-Aalen estimator, non parametric Log-rank test, graphical tests).

The second chapter is separated in two parts. In the first part, the proportional hazards model is analyzed thoroughly (model adjustment, graphical test for the proportional hazards hypothesis, residuals definition, variables selection criteria, metres of goodness of fit etc). The second part is referred to a special case of proportional hazards model that will be adapted in our dataset, which is the Cox model.

The third chapter deals with penalized methods which are used for the confrontation of multicollinearity and overfitting problems in regression models. It combines the definition and the analysis of Ridge, Lasso and (naïve) Elastic Net penalized methods that have been developed over the past few years, and will be used in our dataset. The chapter is completed with the presentation of cross-validation method which provides us with the optimal choice of penalty factor λ.

The fourth, and last chapter, studies a lifetime dataset of patients that suffer from acute myeloblastic leukaemia. More specifically, the chapter starts with a non-parametric data analysis. Secondly, the classic Cox model is adjusted and, in order to confirm the model hypotheses, many analytical and graphical tests are included. Finally, all the penalized techniques are applied in order to check their influence over the previous results that were exported. The statistical packages R and Minitab were used for the data analysis.

Περιεχόμενα

1. Ανάλυση Επιβίωσης

1.1 Εισαγωγικές έννοιες

1.1.1 Δεδομένα Διάρκειας ζωής

Η ανάλυση δεδομένων διάρκειας ζωής ασχολείται με τη μελέτη του χρόνου (ως μια συνεχής τυχαία μεταβλητή T>0) έως ότου προκύψει ένα γεγονός. Στα πιο πολλά προβλήματα αυτό το υπό μελέτη γεγονός σχετίζεται με ένα ανεπιθύμητο ενδεχόμενο όπως θάνατος ή επιπλοκές υγείας ασθενή, καταστροφή ή βλάβη μηχανήματος και άλλα. Για την μελέτη, ανάλυση και εξαγωγή συμπερασμάτων σε τέτοιας μορφής σύνολα δεδομένων έχει αναπτυχθεί ολόκληρος κλάδος στη στατιστική που ονομάζεται Ανάλυση Αξιοπιστίας και Επιβίωσης. Ό όρος Αξιοπιστία (Reliability) χρησιμοποιείται για δεδομένα που αφορούν συνήθως τη βιομηχανία (πχ διάρκεια ζωής μπαταρίας) ενώ ο όρος Επιβίωση (Survivability) έχει να κάνει με βιοϊατρικές εφαρμογές.

Όμως, κάτι επίσης εξαιρετικά χρήσιμο είναι ότι η τ.μ T>0 δεν είναι ανάγκη να αναπαριστά μονάχα χρόνο και αυτό είναι που κάνει την Ανάλυση Αξιοπιστίας και Επιβίωσης τόσο ευέλικτη σε πάρα πολλές εφαρμογές σε όλο το φάσμα των επιστημών. Στην αρχική δήλωση της T θα μπορούσαμε αντί για «χρόνο» να γενικεύσουμε τον ορισμό σε «διάρκεια λειτουργίας» (operating time). Έτσι, η T μπορεί να μην εκφράζεται σε μονάδες χρόνου αλλά σε οποιαδήποτε άλλη μονάδα ταιριάζει στην περιγραφή του προβλήματός μας. Επομένως, η τμ μπορεί να είναι σε μονάδες φορτίου μέχρι την θραύση ενός υλικού, πλήθος χιλιομέτρων μέχρι την πρώτη βλάβη ενός αυτοκινήτου, κυβικά μέτρα νερού έως την υπερχείλιση ενός φράγματος κλπ.

Μια τελευταία παρατήρηση που θα πρέπει να κάνουμε είναι ότι για την περιγραφή δεδομένων διάρκειας ζωής είναι φυσικό να μην είναι ικανοποιητική μια συμμετρική κατανομή όπως η κανονική ή student κατανομή. Χρειαζόμαστε μια κατανομή που να δέχεται αυστηρά θετικές τιμές και επίσης περιμένουμε για μεγάλες τιμές της τ.μ T η πυκνότητα πιθανότητας να είναι ασύμμετρα μικρότερη της πυκνότητας για μικρές τιμές της T. Άρα, χρησιμοποιούμε κατανομές που είναι δεξιά λοξές, χαρακτηριστικά παραδείγματα των οποίων είναι η εκθετική, η Λογαριθμο-Κανονική και η Weibull κατανομή.

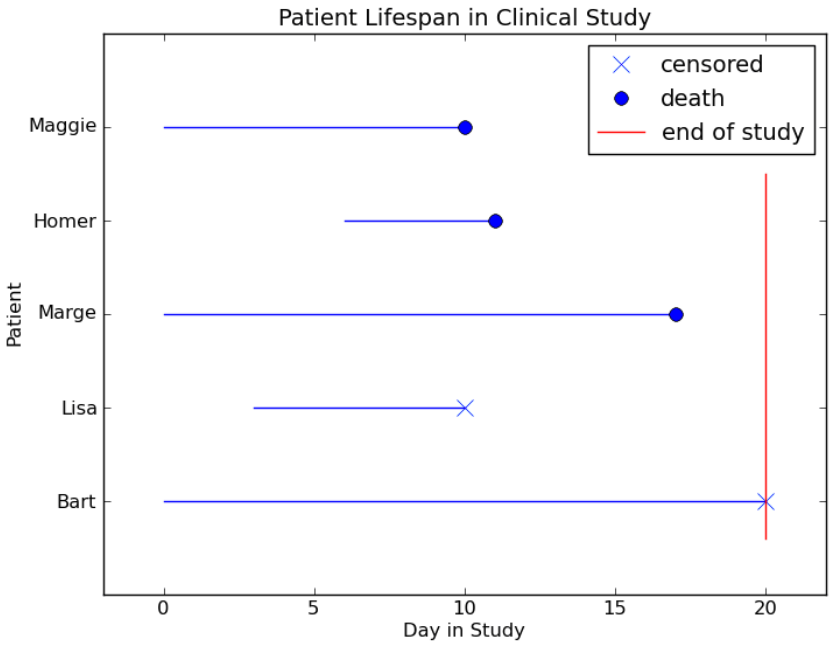
1.12 Αποκομμένα δεδομένα

Ένα βασικό χαρακτηριστικό των δεδομένων διάρκειας ζωής το οποίο δεν παρατηρείται σε άλλους κλάδους στη στατιστική είναι η λεγόμενη αποκοπή δεδομένων. Πολλές φορές, κατά την εκτέλεση ενός πειράματος στο οποίο καταγράφεται ο χρόνος λειτουργίας μέχρι να συμβεί ένα γεγονός, πολλές μονάδες συνεχίζουν να λειτουργούν και μετά τη λήξη του πειράματος. Αν και δεν είμαστε σε θέση να ξέρουμε πότε ακριβώς συνέβη το γεγονός στις εν λόγω μονάδες μετά το τέλος της παρακολούθησης ξέρουμε ότι μέχρι κάποια χρονική στιγμή ήταν ακόμα λειτουργικές. Αυτές η πληροφορίες δεν είναι καθόλου ασήμαντες, για αυτό και δεν τις εξαιρούμε από την μελέτη μας.

Είναι φυσικό σε βιοϊατρικές μελέτες όπου καταγράφεται ο χρόνος μέχρι να συμβεί ένα γεγονός (θάνατος, επιπλοκή, απόρριψη μοσχεύματος κλπ) να μην περιμένουμε απεριόριστο χρονικό διάστημα έως ότου συμβεί το γεγονός σε όλους τους ασθενείς. Αυτό συμβαίνει συνήθως διότι, χρειαζόμαστε άμεσα αποτελέσματα για την απόδοση μιας συγκεκριμένης θεραπείας ώστε να βγάλουμε ιατρικά συμπεράσματα. Επίσης, δεδομένου ότι τα υποκείμενα της μελέτης μας είναι ασθενείς και εμείς καταγράφουμε το χρόνο μέχρι να συμβεί ένα γεγονός είναι πιθανό κάποιος ασθενής να σταματήσει για κάποιο τυχαίο λόγο τη θεραπεία κάποια τυχαία χρονική στιγμή. Και σε αυτή τη περίπτωση δεν γνωρίζουμε πότε και αν θα συμβεί το γεγονός στον συγκεκριμένο ασθενή αλλά ξέρουμε ότι μέχρι τη χρονική στιγμή που σταμάτησε να ακολουθεί

τη θεραπεία δεν του είχε συμβεί. Τέτοιου είδους δεδομένα ονομάζονται αποκομμένα δεδομένα (censored data).

Ο πιο συνηθισμένος τύπος αποκοπής είναι αυτός των δεξιά αποκομμένων παρατηρήσεων, δηλαδή έχουμε εκείνη την περίπτωση όπου κάποιες μονάδες παραμένουν σε λειτουργία πέραν της χρονικής στιγμής που σταματάμε το πείραμα (υπάρχουν, επιπλέον, οι περιπτώσεις αριστερά αποκοπής και αποκοπής σε διάστημα αλλά δε θα ασχοληθούμε στην μελέτη μας με αυτές). Αυτό που πρέπει να επισημάνουμε είναι ότι η αποκοπή θα πρέπει να είναι τυχαία και να μην σχετίζεται με την μετέπειτα διάρκεια ζωής της μονάδας. Αυτό ονομάζεται και μη-πληροφοριακή αποκοπή. Σε αντίθετη περίπτωση, αν η αποκοπή δεν είναι τυχαία τότε απλώς θα μιλάμε για την απόσυρση μιας μονάδας όταν καταλάβουμε ότι αυτή αρχίζει να υπολειτουργεί.



Σχήμα 1.1: Περίπτωση αποκομμένων δεδομένων. Εδώ φαίνεται ότι στους πρώτους 3 ασθενείς έχει συμβεί το γεγονός, ο 4ος σταμάτησε πρόωρα την ιατρική παρακολούθηση και ο 5ος μέχρι το τέλος της παρακολούθησης δεν του είχε συμβεί το γεγονός

Παρακάτω παρουσιάζονται οι δύο πιο βασικοί μηχανισμοί αποκοπής δεδομένων, ανεξάρτητοι της διάρκειας ζωής των μονάδων:

Αποκοπή τύπου Ι: Η παρακολούθηση των μονάδων γίνεται σε προκαθορισμένο χρόνο c. Γνωρίζουμε την ακριβή διάρκεια ζωής των μονάδων αν $T_i < c$, αλλιώς γνωρίζουμε ότι κάποιες μονάδες έχουν υπερβεί το c. Αυτό μπορεί να γενικευτεί ώστε κάθε μονάδα που συμμετέχει στο πείραμα να έχει και το δικό της χρόνο c_i.

Αποκοπή τύπου ΙΙ: Σε αυτή την περίπτωση ο αριθμός των γεγονότων κ είναι προκαθορισμένος και ο χρόνος παρακολούθησης c είναι τυχαίος.

Παρατήρηση: Όλες οι περιπτώσεις αποκοπής δεδομένων δεν επηρεάζουν το είδος της κατανομής των δεδομένων παρά μόνο την εκτίμηση των παραμέτρων της κατανομής.

1.1.3 Η συνάρτηση επιβίωσης S(t)

Ο βασικός σκοπός της ανάλυσης αξιοπιστίας και επιβίωσης είναι να υπολογίσουμε την πιθανότητα η τυχαία μεταβλητή T, που εκφράζει τη διάρκεια ζωής μια μονάδας, να ξεπεράσει κάποιο συγκεκριμένο χρόνο t. Για αυτό το λόγο ορίζουμε την συνάρτηση:

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

Η συνάρτηση S ονομάζεται συνάρτηση αξιοπιστίας ή συνάρτηση επιβίωσης (survival function) γιατί εκφράζει ακριβώς αυτό: τη πιθανότητα μια μονάδα να επιβιώσει πέραν του χρόνου t. Αν τώρα F(t) και f(t) είναι αντίστοιχα οι συναρτήσεις κατανομής πιθανότητας και πυκνότητας πιθανότητας της τ.μ T>0 τότε η σχέση (1.1) μπορεί να γραφεί και ως:

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

1.1.4 Η συνάρτηση διακινδύνευσης h(t)

Η συνάρτηση διακινδύνευσης (hazard function) h(t) εκφράζει τον ρυθμό επικείμενης διακοπής δοθείσης της επιβίωσης της μονάδας μέχρι τη χρονική στιγμή t.
Ξέρουμε ότι η πιθανότητα: $P[t < T \le t + \delta t] \approx f(t) \cdot \delta t$
Άρα λοιπόν από τύπο δεσμευμένης πιθανότητας και τη σχέση (1.1) προκύπτει ότι:

$$P[t < T \le t + \delta t \mid T > t] = \frac{P[t < T \le t + \delta t]}{P(T > t)} = \frac{S(t) - S(t + \delta t)}{S(t)} \cong$$

$$\cong \frac{-dS(t) \cdot \delta t}{S(t)} \overset{(1.2)}{=} \frac{f(t) \cdot \delta t}{S(t)}$$

Οπότε ορίζουμε την συνάρτηση διακινδύνευσης:

$$h(t) = \lim_{\delta t \to 0} \frac{(S(t) - S(t + \delta t)) / \mathrm{S(t)}}{\delta t} = \frac{f(t)}{S(t)} \quad (1.4)$$

Η συμπεριφορά της h(t) συνήθως ανήκει σε μια από τις παρακάτω κατηγορίες:

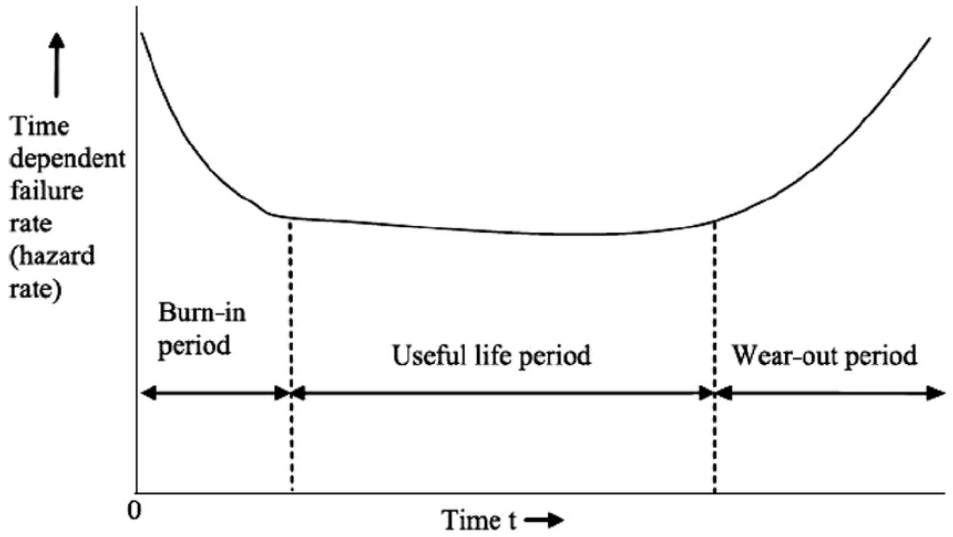
Α) h(t)=λ= σταθερή. Σε αυτή τη περίπτωση το μοντέλο έχει την ιδιότητα ο στιγμιαίος ρυθμός διακοπής λειτουργίας μιας μονάδας να είναι σταθερός στο χρόνο, δηλαδή είναι ανεξάρτητος της ηλικίας της μονάδας. Με άλλα λόγια η μονάδα δείχνει να μην γερνά με το πέρασμα του χρόνου. Δεν είναι ρεαλιστικό αλλά μπορεί να συμβαίνει σε κάποιο συγκεκριμένο χρονικό διάστημα $[\mathrm{t}_1, \mathrm{t}_2]$. Χαρακτηριστικό παράδειγμα ενός τέτοιου μοντέλου είναι η εκθετική κατανομή.

Β) h(t) αύξουσα συνάρτηση. Εδώ, ο στιγμιαίος ρυθμός διακοπής αυξάνεται καθώς αυξάνεται η ηλικία της μονάδας, λογική συνέπεια της γήρανσης.

Γ) h(t) φθίνουσα συνάρτηση. Σε αυτή τη περίπτωση ο ρυθμός διακοπής μειώνεται καθώς αυξάνεται η ηλικία της μονάδας. Η συμπεριφορά αυτή μπορεί να ερμηνευθεί στις βιοϊατρικές εφαρμογές ως βελτίωση της υγείας τους ασθενούς προϊόντος της αποδοτικής θεραπείας. Σε δεδομένα από τη βιομηχανία, μια τέτοια συμπεριφορά, μπορεί να προκύψει μόνο στο πρώτο

στάδιο της παραγωγής κατά το οποίο αποσύρονται άμεσα οι ελαττωματικές μονάδες και παραμένουν μόνο οι ποιοτικά καλύτερες άρα ο ρυθμός επικείμενης διακοπής μειώνεται.

Δ) h(t) συνδυασμός όλων των προηγούμενων περιπτώσεων. Ένας πολύ γνωστός συνδυασμός των παραπάνω περιπτώσεων είναι η λεγόμενη συνάρτηση διακινδύνευσης «μπανιέρας» (bathtub hazard function). Εδώ, έχουμε αρχικά μια χρονική περίοδο [0,t] όπου η συνάρτηση διακινδύνευσης είναι φθίνουσα, ακολουθούμενη από μια περίοδο σταθεροποίησης της h(t) και καταλήγει σε ένα στάδιο αύξουσας διακινδύνευσης. Λόγω αυτής της ιδιότυπης συμπεριφοράς, όπου γραφικά έχει το σχήμα μπανιέρας, αποκόμισε και το όνομά της.



Σχήμα 1.2: Συνάρτηση διακινδύνευσης bathtub

1.1.5 Η συνάρτηση σωρευτικής διακινδύνευσης H(t)

Η σωρευτική συνάρτηση διακινδύνευσης H(t) ορίζεται ως:

$$H(\mathrm{t}) = \int_0^t h(u)du \quad (1.5)$$

όπου h(t) είναι η συνάρτηση διακινδύνευσης. Βάσει του ορισμού (1.4) της συνάρτησης διακινδύνευσης, η σχέση (1.5) γράφεται και ως:

$$H(\mathrm{t}) = \int_0^t \frac{f(u)}{S(\mathrm{u})}du = \int_0^t \frac{-dS(u)}{S(u)}du = [-\ln S(u)]_0^t = -\ln S(t) \Leftrightarrow$$
$$S(t) = \exp\{-H(t)\} \quad (1.6)$$

Παρατήρηση: Για να καταλήξουμε στη σχέση (1.6) έχουμε σιωπηλά δεχτεί ότι το ln(S(0))=0. Αυτό προκύπτει από μια πολύ λογική παραδοχή για την συνάρτηση επιβίωσης. Σε ένα ρεαλιστικό μοντέλο που περιγράφει δεδομένα διάρκειας ζωής, θα πρέπει για t=0 που ξεκινάει ο χρόνος λειτουργίας των μονάδων η πιθανότητα κάποια μονάδα να έχει επιζήσει ως τη χρονική στιγμή 0 να είναι 100%. Άρα, με άλλα λόγια, θα πρέπει το S(0)=P[T>0]=1. Επομένως, ln(S(0))=ln(1)=0.

1.2 Μη-παραμετρική ανάλυση δεδομένων διάρκειας ζωής

1.2.1 Η εκτιμήτρια Kaplan-Meier της S(t)

Όπως είδαμε στο κεφάλαιο 1.1.3 η συνάρτηση επιβίωσης S(t) ορίζεται ως: S(t)=1-F(t), όπου F(t) είναι η συνάρτηση κατανομής των δεδομένων μας. Όταν όμως δεν γνωρίζουμε εκ των προτέρων την κατανομή και δεν μπορούμε να διακρίνουμε ότι υποβόσκει κάποιο συγκεκριμένο μοντέλο θα πρέπει να μπορούμε να υπολογίζουμε με έναν μη-παραμετρικό τρόπο την κομβικής σημασίας συνάρτηση επιβίωσης S(t). Εφόσον, μάλιστα, τα δεξιά δεδομένα αποκοπής τα συναντάμε πολύ συχνά στην μελέτη μας θα πρέπει αυτή η μη-παραμετρική εκτιμήτρια να μπορεί να συμπεριλάβει το γεγονός της ύπαρξης τέτοιων δεδομένων. Για αυτό το λόγο χρησιμοποιούμε την μη-παραμετρική εκτιμήτρια Kaplan-Meier.

Έστω λοιπόν n-το πλήθος μονάδες ενός δείγματος, εκ των οποίων στις k<n συνέβη το υπό μελέτη γεγονός. και καταγράφουμε τους διακεκριμένους χρόνους λειτουργίας $t_{(1)} < t_{(2)} < ... < t_{(k)}$. Ορίζουμε ως d_j το πλήθος των μονάδων στις οποίες συνέβη το γεγονός τη χρονική στιγμή $t_{(j)}, j = 1,...,k$ και ως n_j τον αριθμό των μονάδων που ήταν σε κίνδυνο αμέσως πριν τη χρονική στιγμή $t_{(j)}$.

Ο αριθμός n_j αφορά όλες εκείνες τις μονάδες με χρόνο ζωής μεγαλύτερο ή ίσο της j-οστής χρονικής στιγμής ($t \geq t_{(j)}$), ανεξαρτήτως της τελικής κατάληξής τους. Δηλαδή συμπεριλαμβάνονται όλες οι μονάδες στις οποίες πρόκειται να συμβεί το γεγονός και σε όλες εκείνες που δε θα συμβεί έως και το τέλος του πειράματος (αποκομμένες παρατηρήσεις). Αντίθετα, δεν λογαριάζονται στο n_j όλες οι μονάδες στις οποίες συνέβη το γεγονός και επίσης οι μονάδες με αποκομμένες τιμές πριν τη στιγμή $t_{(j)}$.

Η τιμή της εκτιμήτριας Kaplan-Meier τη χρονική στιγμή $t_{(j)}$ θα είναι:

$$S_{KM}(t_{(j)}) = P[T > t_{(j)}] =$$
$$= P[\{T > t_{(1)}\} \bigcap \{T > t_{(2)}\} \bigcap ... \bigcap \{T > t_{(j)}\}] =$$
$$= P[T > t_{(1)}] \cdot P[T > t_{(2)} \mid T > t_{(1)}] \cdot P[T > t_{(3)} \mid \{T > t_{(1)}\} \bigcap \{T > t_{(2)}\}] \cdot ....$$
$$P[T > t_{(j)} \mid \{T > t_{(1)}\} \bigcap ... \bigcap \{T > t_{(j-1)}\}] \quad (1.7)$$

Ο τύπος (1.7) προέκυψε από την απλή εφαρμογή του τύπου των πιθανοτήτων:
$P(A_1 \bigcap A_2) = P(A_1) \cdot P(A_2 \mid A_1)$ επεκτείνοντας τον για k-το πλήθος ενδεχόμενα όπου το κάθε $A_i = \{T > t_{(i)}\}$. Επιπλέον, είναι προφανές ότι εφόσον οι χρονικές στιγμές $t_{(1)} < t_{(2)} < ... < t_{(k)}$ τα ενδεχόμενα $\{T > t_{(p)}\}$ και $\{T > t_{(q)}\}$ είναι ξένα μεταξύ τους αν $p \neq q$ και μάλιστα το ένα περιέχεται στο άλλο. Δηλαδή $\{T > t_{(p)}\} \subset \{T > t_{(q)}\}$ αν p<q.
Άρα η σχέση (1.7) απλοποιείται αρκετά:

$$S_{KM}(t_{(j)}) = P[T > t_{(1)}] \cdot P[T > t_{(2)} \mid T > t_{(1)}] \cdot ... \cdot P[T > t_{(j)} \mid T > t_{(j-1)}] \quad (1.8)$$

Αν ορίσουμε ως p_1 τη σχετική συχνότητα των μονάδων στις οποίες έχει συμβεί το γεγονός από την αρχή του πειράματος ως τη χρονική στιγμή $t_{(1)}$,τότε με βάση τους ορισμούς των d_j και n_j θα ισχύει ότι: $p_1 = d_1 / n_1$.

Μια εκτίμηση της πιθανότητας $P[T > t_{(1)}]$ είναι η $P[T > t_{(1)}] = 1 - p_1 = \frac{n_1 - d_1}{n_1}$.Με την ίδια λογική μπορούμε να βρούμε οποιαδήποτε $P[T > t_{(i)} \mid T > t_{(i-1)}] = \frac{n_i - d_i}{n_i}$ (1.9). Με βάση τον τύπο (1.9) αντικαθιστώντας στη σχέση (1.8) παίρνουμε την τελική έκφραση της εκτιμήτριας Kaplan-Meier (Kaplan & Meier, 1958):

$$S_{KM}(t_{(j)}) = \frac{n_1 - d_1}{n_1} \cdot .... \cdot \frac{n_j - d_j}{n_j} \Rightarrow S_{KM}(t) = \begin{cases} \prod_{j:t_{(j)} \le t} \frac{n_j - d_j}{n_j}, & t \ge t_{(1)} \\ 1, & t < t_{(1)} \end{cases} \quad (1.10)$$

Παρατήρηση: Η Kaplan-Meier εκτιμήτρια ως δειγματοσυνάρτηση είναι και αυτή τυχαία μεταβλητή και συχνά χρησιμοποιείται για να εντοπίσουμε αν υπάρχει κάποιο υποβόσκον μοντέλο στα δεδομένα του προβλήματός μας.

Πρόταση: Συχνά για την κατασκευή 95% διαστημάτων εμπιστοσύνης για τις τιμές της Kaplan-Meier εκτιμήτριας θεωρούμε ότι προσεγγιστικά: $S_{KM}(t) \sim N(S(t), V(S(t)))$ (1.11)

όπου $V(S(t)) = (S(t))^2 \cdot (\sum_{t_{(j)} < t} \frac{d_j}{n_j(n_j - d_j)})$ και το τυπικό σφάλμα $se(S(t)) = \sqrt{V(S(t))}$

Άρα ένα 95% δ.ε για την τιμή S(t) είναι το ακόλουθο:

$$S(t) - 1.96 \cdot se(S(t)) < S(t) < S(t) + 1.96 \cdot se(S(t)) \quad (1.12)$$

1.2.2 Η εκτιμήτρια Nelson-Aalen της H(t)

Έχοντας εκτιμήσει την συνάρτηση επιβίωσης S(t) από την εκτιμήτρια Kaplan-Meier μπορούμε να πάρουμε μια επίσης μη-παραμετρική εκτίμηση της σωρευτικής συνάρτησης διακινδύνευσης H(t). Αυτό μπορεί να γίνει μέσω της σχέσης (1.6) αλλά προτιμότερη είναι η εκτιμήτρια Nelson-Aalen (Nelson, 1972 Aalen, 1978)

Ξεκινώντας από τη σχέση $S(t) = \exp\{-H(t)\} \Leftrightarrow H(t) = -\ln(S(t))$ και εισάγοντας τον τύπο της εκτιμήτριας Kaplan-Meier από τη σχέση (1.10) παίρνουμε:

$$H(t) = -\ln(S_{KM}(t)) \Rightarrow H(t) = -\sum_{j:t_{(j)} \le t} \ln(1 - \frac{d_j}{n_j}) \quad (*)$$

Αν τώρα αναπτύξουμε σε σειρά Taylor την συνάρτηση f(x)=ln(1-x) γύρω από το μηδέν και αγνοώντας τους υψηλόβαθμους όρους καταλήγουμε ότι: $\ln(1-x) \cong -x$

Οπότε αν θεωρήσουμε ότι κάθε χρονική στιγμή ο όρος d_j είναι πολύ μικρότερος του n_j (κάτι που είναι αρκετά λογικό εφόσον οι μονάδες που καταστρέφονται κάθε στιγμή είναι πολύ

λιγότερες από όσες βρισκόντουσαν σε κίνδυνο ακριβώς πριν αυτή τη δεδομένη στιγμή). Άρα με βάση αυτή την παρατήρηση η σχέση (*) τελικά γίνεται:

$$H_{NA}(t) = \sum_{j:t_{(j)} \le t} \frac{d_j}{n_j} \quad (1.13)$$

1.2.3 Ο μη-παραμετρικός έλεγχος Log-rank

Πολλές φορές στα προβλήματα χρειαζόμαστε να συγκρίνουμε την επιβίωση δύο διαφορετικών ομάδων δεδομένων. Κάτι τέτοιο είναι πολύ σημαντικό διότι με αυτόν τον τρόπο μπορούμε να αποφανθούμε κατά πόσο ένας υπό μελέτη παράγοντας επιδρά στη διάρκεια ζωής των ατόμων. Έτσι λοιπόν, χωρίζουμε σε δύο ομάδες Α και Β τα δεδομένα μας και πραγματοποιούμε τον έλεγχο $\{H_0 : S_A(t) = S_B(t)\ vs\ H_1 : S_A(t) \neq S_B(t)\}$. Όταν δεν είναι γνωστή η μορφή των συναρτήσεων S_A και S_B τότε χρησιμοποιούμε τον μη παραμετρικό έλεγχο Log-rank (Καρώνη, 2009).

Έστω $t_{(1)} < t_{(2)} < ... < t_{(k)}$ οι διακεκριμένες χρονικές στιγμές κατά τις οποίες συμβαίνει το γεγονός και στις δύο ομάδες ατόμων. Αμέσως πριν τη χρονική στιγμή $t_{(j)}$, θεωρούμε ότι στην ομάδα i (=1,2) υπάρχουν n_{ij} μονάδες σε κίνδυνο εκ των οποίων στις d_{ij} συμβαίνει το γεγονός. Ορίζουμε ως:

$$n_j = n_{1j} + n_{2j} \text{ και } d_j = d_{1j} + d_{2j}$$

Για κάθε χρονική στιγμή $t_{(j)}$,j=1,…k κατασκευάζουμε έναν 2X2 πίνακα συνάφειας της μορφής:

Γεγονός \ Ομάδα	A	B	Άθροισμα
Ναι	D_{1j}	d_{2j}	d_j
Όχι	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
Άθροισμα	N_{1j}	n_{2j}	n_j

Πίνακας 1.1: 2X2 πίνακας συνάφειας

Δημιουργώντας έναν τέτοιο πίνακα συνάφειας, ουσιαστικά, θεωρούμε υπό την Ho ότι το γεγονός, αν θα συμβεί ή όχι, είναι ανεξάρτητο της ομάδας που ανήκουν τα άτομα. Για να ελέγξουμε την ισχύ αυτής της υπόθεσης υπολογίζουμε τις αναμενόμενες συχνότητες και τις συγκρίνουμε με τις παρατηρηθείσες συχνότητες κάθε κελιού και εκτελούμε τον λεγόμενο X^2 -independence test. Οι αναμενόμενες συχνότητες ορίζονται ως:

$$αναμενόμενες\ συχνότητες = \frac{(άθροισμα\ γραμμής) \times (άθροισμα\ στήλης)}{μέγεθος\ δείγματος}$$

όπου το άθροισμα είναι ως προς όλα τα κελιά.
Αν θέλουμε δηλαδή στον πίνακα 1.1 να υπολογίσουμε την αναμενόμενη συχνότητα του κελιού Ομάδα Β, Όχι θα είχαμε: $E(d_{2j}) = d_{2j} = d_j \cdot n_{2j} / n_j$

Άρα η απόκλιση από την παρατηρούμενη d_{2j} είναι: $u_j = d_{2j} - d_{2j} = d_{2j} - d_j \cdot n_{2j} / n_j$
Θεωρώντας ότι οι παρατηρηθείσες συχνότητες ακολουθούν την υπεργεωμετρική κατανομή μπορούμε να υπολογίσουμε την διασπορά:

$$V(d_{2j}) = \frac{n_{2j} d_j (n_j - n_{2j})(n_j - d_j)}{n_j^2 (n_j - 1)} = \frac{n_{2j} d_j \, \mathrm{n}_{1j} (n_j - d_j)}{n_j^2 (n_j - 1)} = v_j$$

Η ελεγχοσυνάρτηση του ελέγχου Log-rank ορίζεται λοιπόν ως εξής:

$$z = \frac{u}{\sqrt{v}} \quad (1.13)$$

όπου ορίζουμε ως: $u = \sum_{j=1}^{k} u_j$, $v = \sum_{j=1}^{k} v_j$
Αν το k (πλήθος μη αποκομμένων παρατηρήσεων) είναι μεγάλο τότε από Κ.Ο.Θ θα ισχύει ότι:
z~N(0,1) ή ισοδύναμα ότι: $z^2 = \frac{u^2}{v} \sim X^2_{(1)}$

Σχόλιο: Ο έλεγχος Log-rank είναι προφανώς ένας μη-παραμετρικός έλεγχος καθώς εξετάζει την υπόθεση $\{H_0 : S_A(t) = S_B(t)\}$ χωρίς να προϋποθέτει οποιαδήποτε γνώση της συναρτησιακής μορφής των $S_A(t)$ και $S_B(t)$. Τα μόνα στοιχεία που χρησιμοποιούνται είναι τα παρατηρούμενα n_{ij} και d_{ij}.

1.2.4 Γραφικοί έλεγχοι

Όπως αναφέραμε και στην παράγραφο 1.2.1, η εκτιμήτρια Kaplan-Meier είναι μια μη παραμετρική εκτιμήτρια της συνάρτησης επιβίωσης S(t). Μπορεί να χρησιμοποιηθεί σε κάθε πιθανό σύνολο δεδομένων δίνοντας μας αξιόπιστα αποτελέσματα εφόσον δεν προϋποθέτει καμία εκ των προτέρων γνώση της πιθανής παραμετρικής μορφής των δεδομένων μας. Όμως, επειδή πολλές φορές η χρήση του σωστού παραμετρικού μοντέλου μπορεί να είναι πιο αποτελεσματική για την εκτίμηση των χαρακτηριστικών της διάρκειας ζωής μπορούμε να κάνουμε χρήση της εκτιμήτριας Kaplan-Meier ώστε να διενεργήσουμε διάφορους γραφικούς ελέγχους ώστε να προσδιορίσουμε κάποια γνωστή κατανομή που υποβόσκει στα δεδομένα μας.

Ας υποθέσουμε λοιπόν ότι η τ.μ T>0 που περιγράφει τη διάρκεια ζωής ακολουθεί την εκθετική κατανομή αλλά εμείς δεν το γνωρίζουμε αυτό. Υπολογίσουμε αρχικά την εκτιμήτρια Kaplan-Meier. Ξέρουμε ότι η εκθετική κατανομή με παράμετρο λ, έχει συνάρτηση επιβίωσης της μορφής: $S(t) = \exp(-\lambda t), \lambda > 0$. Άρα ισοδύναμα ισχύει ότι: $-\ln(S(t)) = \lambda t \Rightarrow y = \beta_0 + \beta_1 x$
όπου: $y = -\ln(S(t))$, $x = t$, $\beta_0 = 0$ και $\beta_1 = \lambda$. Οπότε, αν υποβόσκει το μοντέλο της εκθετικής κατανομής η γραφική παράσταση των $-\ln(S_{KM}(t_{(j)}))$ έναντι των $t_{(j)}$ θα είναι μια ευθεία γραμμή ($t_{(j)}$ συμβολίζουμε τους χρόνους διακοπής). Όσο μεγαλύτερη η σύμπτωση των σημείων πάνω στην ευθεία τόσο καλύτερη η προσαρμογή των δεδομένων στην εκθετική κατανομή.

Παρόμοιους γραφικούς ελέγχους μπορούμε προφανώς να κάνουμε και για άλλα γνωστά παραμετρικά μοντέλα χρησιμοποιώντας και πάλι την Kaplan-Meier εκτιμήτρια των δεδομένων

μας. Στον πίνακα 1.2 φαίνονται οι γραφικοί έλεγχοι που μπορούν να γίνουν ανάλογα με το είδος της υποψήφιας κατανομής:

Κατανομή	**Γραφική παράσταση**
Εκθετική	$-\ln(S(t))$ vs t
Weibull	$\ln(-\ln(S(t)))$ vs lnt
Λογαριθμο-κανονική	$\Phi^{-1}(1-S(t))$ vs lnt
Κανονική	$\Phi^{-1}(1-S(t))$ vs t
Λογιστική	$\ln((1-S(t))/S(t))$ vs t
Λογαριθμο-λογιστική	$\ln((1-S(t))/S(t))$ vs lnt

Πίνακας 1.2

όπου Φ(z) είναι η συνάρτηση κατανομής της τυποποιημένης κανονικής κατανομής, δηλαδή: $\Phi(z) = P[Z \le z]$ με Z~N(0,1).

Παρόμοιες γραφικές παραστάσεις μπορούν να αναπτυχθούν κάνοντας χρήση και της μη παραμετρικής εκτιμήτριας Nelson-Aalen της σωρευτικής συνάρτησης διακινδύνευσης H(t). Παραδείγματος χάρη, αν υποθέσουμε πάλι ότι υποβόσκει το μοντέλο της εκθετικής κατανομής, η οποία έχει $\mathrm{H(t)} = \lambda t$, μπορούμε να κάνουμε τη γράφημα της $H_{NA}(\mathrm{t}_{(j)})$ έναντι του $\mathrm{t}_{(j)}$ αναμένουμε μια ευθεία.

Αφού επιβεβαιώσουμε την ύπαρξη ενός παραμετρικού μοντέλου μπορούμε να εκτιμήσουμε τις παραμέτρους της κατανομής από τα χαρακτηριστικά της ευθείας αν και συνήθως προτιμούμε να κάνουμε χρήση της κλασικής μεθόδου μέγιστης πιθανοφάνειας η οποία μπορεί να μετασχηματιστεί ώστε να λαμβάνει υπόψη και αποκομμένες παρατηρήσεις οι οποίες είναι πολύ συχνές σε δεδομένα διάρκειας ζωής.

2. Παραμετρικά μοντέλα αναλογικής διακινδύνευσης και το ημι-παραμετρικό μοντέλο του Cox

2.1 Μοντέλα Αναλογικής Διακινδύνευσης για δεδομένα διάρκειας ζωής

2.1.1 Ορισμός του μοντέλου

Όπως ξέρουμε από την γραμμική παλινδρόμηση, όταν θέλουμε να περιγράψουμε την επίδραση των $x_i, i = 1,..,\mathrm{k}$ επεξηγηματικών μεταβλητών στην μεταβλητή απόκρισης Y προσαρμόζουμε το γενικό γραμμικό μοντέλο:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \varepsilon_i \quad (2.1)$$

όπου θεωρούμε ότι τα υπόλοιπα ε_i είναι ανεξάρτητες και ισόνομες τ.μ που ακολουθούν την κανονική κατανομή, δηλαδή: $\varepsilon_i \sim N(0,\sigma^2)$. Εφόσον, οι συμμεταβλητές του μοντέλου θεωρούνται μη-στοχαστικές, τότε για η μεταβλητή απόκρισης $\mathrm{Y} \sim N(\mu,\sigma^2)$, όπου έχουμε συμβολίσει ως $\mu = \mu(\underset{\rightarrow}{x}) = \underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}$

Αν θέλαμε να προσαρμόσουμε ένα μοντέλο της μορφής (2.1) για την περιγραφή του χρόνου ζωής T που εξαρτάται από τις συμμεταβλητές x_i ναι μεν θα ήταν λογικό αλλά παρόλα αυτά καθόλου λειτουργικό στην πράξη. Όπως είδαμε για τη μεταβλητή απόκρισης $\mathrm{Y} \sim N(\mu,\sigma^2)$ όπου $\mu \in \mathbb{R}$, άρα το μ μπορεί εν γένει να λάβει και αρνητικές τιμές κάτι το οποίο δε θα είχε νόημα καθώς η τ.μ T>0 εκφράζει διάρκεια ζωής. Άρα χρειαζόμαστε μια διαφορετική προσέγγιση του προβλήματος.

Ένας πολύ διαδεδομένος τύπος μοντέλου που συνδέει τις επεξηγηματικές μεταβλητές με τον χρόνο διακοπής είναι τα λεγόμενα μοντέλα αναλογικής διακινδύνευσης (proportional hazard models, PH). Αυτά τα μοντέλα περιγράφονται από τη σχέση:

$$h(t;\underset{\rightarrow}{x}) = h_0(t) \cdot g(\underset{\rightarrow}{x}) \quad (2.2)$$

όπου $h_0(t)$ είναι μια κοινή συνάρτηση διακινδύνευσης για όλες τις μονάδες και g(.) είναι μια συνάρτηση των συμμεταβλητών που εκφράζει τη μεταβλητότητα κάθε μονάδας. Συνήθως επιλέγουμε ως $g(\underset{\rightarrow}{x}) = \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}) > 0$, αν και αυτή η επιλογή δεν είναι δεσμευτική. Το $\underset{\rightarrow}{\beta}$ είναι ένα διάνυσμα p=k+1 συντελεστών που εκφράζουν την επίδραση κάθε συμμεταβλητής $x_i, i = 0,...,k$. Άρα λοιπόν καταλήγουμε στο μοντέλο:

$$h(t;\underset{\rightarrow}{x}) = h_0(t) \cdot \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}) \quad (2.3)$$

Παρατηρούμε ότι η σχέση (2.3) ισοδύναμα μπορεί να γραφεί και ως:

$$\ln(h(t;\underset{\rightarrow}{x})) = \ln(h_0(t)) + \underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x} \quad (2.4)$$

το οποίο είναι ένα μοντέλο παλινδρόμησης για το lnh.

Ανάλογα με τη μορφή της ho(t) τα μοντέλα PH χωρίζονται σε:

Α) Παραμετρικά μοντέλα: Εδώ η βασική συνάρτηση διακινδύνευσης ho(t) καθορίζεται από κάποιο γνωστό παραμετρικό μοντέλο (πχ Weibull, Log-Normal)

Β) Ημι-παραμετρικά μοντέλα: Εδώ η μορφή της ho(t) δεν ακολουθεί κάποια συγκεκριμένη κατανομή ή μας είναι άγνωστη και παραμένει ακαθόριστη. Το πιο βασικό παράδειγμα ημι-παραμετρικού μοντέλου είναι αυτό του Cox με το οποίο θα ασχοληθούμε στη συνέχεια.

Χρησιμοποιώντας τη σχέση (2.3) και με βάση τον ορισμό (1.5) της σωρευτικής συνάρτησης διακινδύνευσης παίρνουμε ότι:

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

Ακριβώς, ανάλογα από τη σχέση (1.6) υπολογίζουμε την συνάρτηση επιβίωσης:

$$\mathrm{S(t;}\underset{\rightarrow}{x}) = \exp(-H(t;\underset{\rightarrow}{x})) = \exp(-H_0(t)\cdot\exp(\underset{\rightarrow}{\beta}^{\mathrm{T}}\cdot\underset{\rightarrow}{x})) = (\exp(-\mathrm{H}_0(t))^{\exp(\underset{\rightarrow}{\beta}^{\mathrm{T}}\cdot\underset{\rightarrow}{x})} = (S_0(t))^{\exp(\underset{\rightarrow}{\beta}^{\mathrm{T}}\cdot\underset{\rightarrow}{x})}$$

Ένα ιδιαίτερο χαρακτηριστικό των μοντέλων αναλογικής διακινδύνευσης είναι ότι ο λόγος:

$$\frac{h(t;\underset{\rightarrow}{x_1})}{h(t;\underset{\rightarrow}{x_2})} = \frac{\exp(\underset{\rightarrow}{\beta}^{\mathrm{T}}\cdot\underset{\rightarrow}{x_1})}{\exp(\underset{\rightarrow}{\beta}^{\mathrm{T}}\cdot\underset{\rightarrow}{x_2})} = \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}}\cdot(\underset{\rightarrow}{x_1}-\underset{\rightarrow}{x_2})) \quad (2.5)$$

για δύο διαφορετικά διανύσματα $\underset{\rightarrow}{x_1}$ και $\underset{\rightarrow}{x_2}$ είναι μη-στοχαστικός και ανεξάρτητος του χρόνου. Άρα οι συναρτήσεις διακινδύνευσης είναι σε αναλογία μεταξύ τους. Αυτή είναι και η βασική υπόθεση ενός PH μοντέλου και πρέπει να ελέγχεται πάντα (Caroni, 2017).

2.1.2 Εκτίμηση παραμέτρων

Η εκτίμηση των παραμέτρων ενός PH μοντέλου γίνεται με τη μέθοδο μέγιστης πιθανοφάνειας. Βέβαια, επειδή θέλουμε να καλύψουμε και την περίπτωση της ύπαρξης αποκομμένων παρατηρήσεων στα δεδομένα μας δεν χρησιμοποιούμε την κλασσική συνάρτηση πιθανοφάνειας $\mathrm{L} = \prod_{i=1}^{n} f(t_i)$, όπου f είναι η σ.π.π της τ.μ T και n το πλήθος των παρατηρήσεων, αλλά μια τροποποίηση αυτής:

$$L = \prod_{i=1}^{n} [f(t_i)]^{\delta_i}\cdot[S(t_i)]^{1-\delta_i} \quad (2.6)$$

όπου ορίζουμε ως $\delta_i = \begin{Bmatrix} 1, \textit{συνέβη το γεγονός} \\ 0, \textit{διαφορετικά} \end{Bmatrix}$ και S(t) η συνάρτηση επιβίωσης της T.

Λογαριθμίζοντας την σχέση (2.6) θα πάρουμε:

$$l = \ln L = \sum_{i=1}^{n} \{\delta_i \ln(\mathrm{f}(\mathrm{t}_i)) + (1-\delta_i)\ln(\mathrm{S}(\mathrm{t}_i))\} \quad (2.7)$$

Όμως επειδή έχουμε την περίπτωση PH μοντέλου ξέρουμε ότι: $S(t_i) = S(t_i; \underset{\rightarrow}{x_i}) = S_0(t)^{\exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x_i})}$ και επίσης $f(t_i) = f(t; \underset{\rightarrow}{x_i})$ άρα η σχέση (2.7) γράφεται ισοδύναμα και ως:

$$l = \sum_{i=1}^{n} \{\delta_i \ln(\mathrm{f}(\mathrm{t}; \underset{\rightarrow}{x_i})) + (1-\delta_i)\exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x_i})\ln(\mathrm{S}_0(t))\} \quad (2.8)$$

Τέλος μένει να λύσουμε το σύστημα εξισώσεων: $\frac{\partial l}{\partial \beta_j} = 0,\ j = 0,...,\mathrm{k}$ το οποίο συνήθως λύνεται με χρήση αριθμητικών μεθόδων.

2.1.3 Έλεγχοι υποθέσεων

Έχοντας προσαρμόσει το μοντέλο και εκτιμήσει τις παραμέτρους του, σημαντικό ρόλο για τη συνέχεια είναι να μπορούμε να εκτελούμε ελέγχους υποθέσεων που αφορούν τη σημαντικότητα των παραμέτρων που συμπεριλαμβάνονται στο εκτιμημένο μοντέλο καθώς επίσης και τον έλεγχο της υπόθεσης της αναλογικής διακινδύνευσης. Για τη σημαντικότητα των παραμέτρων χρησιμοποιούμε δύο πολύ γνωστούς ελέγχους: τον έλεγχο Wald (Wald test) και τον έλεγχο του λόγου των πιθανοφανειών (Likelihood ratio test) ενώ η υπόθεση αναλογικής διακινδύνευσης ελέγχεται μέσω μιας γραφικής μεθόδου.

2.1.3.1 Έλεγχος Wald

Έχοντας προσαρμόσει το μοντέλο αναλογικής διακινδύνευσης θα πρέπει να είμαστε σε θέση να ελέγξουμε την σημαντικότητα κάθε μεταβλητής ξεχωριστά που μετέχει στο μοντέλο. Ισοδύναμα, πρέπει να ελέγχουμε τις υποθέσεις: $\{H_o : \beta_i = 0 \ vs \ H_1 : \beta_i \neq 0\}, \mathrm{i} = 1,...,\mathrm{k}$. Αυτό ακριβώς κάνει ο έλεγχος Wald.

Αφού πρώτα υπολογίσουμε τις εκτιμήτριες $\beta_i, \mathrm{i} = 1,...,\mathrm{k}$ και τα τυπικά τους σφάλματα $se(\beta_i)$ χρησιμοποιούμε την ελεγχοσυνάρτηση:

$$w = \frac{\beta_i - \beta_i|_{H_0}}{se(\beta_i)} = \frac{\beta_i}{se(\beta_i)} \quad (2.9)$$

Υπό την μηδενική υπόθεση w~N(0,1) ή ισοδύναμα $w^2 \sim X^2_{(1)}$. Οπότε ορίζοντας ένα επίπεδο σημαντικότητας α (συνήθως 5%) μπορούμε να ελέγξουμε την σημαντικότητα των μεταβλητών του μοντέλου. Αν απορριφθεί η Ho τότε η i-οστή συμμεταβλητή είναι στατιστικά σημαντική (Collett, 2003).

2.1.3.2 Ο έλεγχος του λόγου των πιθανοφανειών

Ένας επίσης πολύ διαδεδομένος τρόπος εκτέλεσης ελέγχων υποθέσεων είναι ο έλεγχος του λόγου των πιαθανοφανειών (likelihood ratio test). Αν παραδείγματος χάρη θέλουμε να ελέγξουμε υποθέσεις της μορφής $H_o : \beta_i = 0$ ώστε να διαπιστώσουμε την σημαντικότητα της i-οστής συμμεταβλητής του μοντέλου μας θα πρέπει να ακολουθήσουμε την εξής διαδικασία:

Αρχικά προσαρμόζουμε το μοντέλο χωρίς την x_i ($\beta_i = 0$), και έπειτα υπολογίζουμε τη μεγιστοποιημένη τιμή του λογαρίθμου της πιθανοφάνειας, έστω $\hat{l}_0$.Κάνουμε το ίδιο και για το μοντέλο που περιέχει την x_i ($\beta_i \neq 0$) και υπολογίζουμε πάλι τη μεγιστοποιημένη τιμή του λογαρίθμου της πιθανοφάνειας, έστω $\hat{l}_1$. Ο έλεγχος του λόγου των πιθανοφανειών χρησιμοποιεί την ελεγχοσυνάρτηση:

$$z = -2(\hat{l}_0 - \hat{l}_1) \sim X^2_{(1)} \quad (2.10)$$

Αν η p-value είναι αρκετά μικρή (συνήθως μικρότερη του 0.05) έχουμε ισχυρές ενδείξεις ώστε να απορρίψουμε την μηδενική υπόθεση, και άρα η μεταβλητή x_i είναι στατιστικά σημαντική. Αν αντίθετα, η p-value προκύψει μεγάλη τότε θεωρούμε ότι η μεταβλητή x_i είναι στατιστικά μη σημαντική.

Ένα μεγάλο πλεονέκτημα του ελέγχου του λόγου των πιθανοφανειών είναι ότι είναι αρκετά ευπροσάρμοστος ώστε να χρησιμοποιηθεί σε πολλούς ελέγχους όχι μόνο της μορφής $\{H_o : \beta_i = 0 \; vs \; H_1 : \beta_i \neq 0\}$. Ένα χαρακτηριστικό παράδειγμα είναι η χρήση του για την επιλογή του καλύτερου παραμετρικού μοντέλου που προσαρμόζεται στα δεδομένα μας αν το ένα είναι εμφωλευμένο στο άλλο, δηλαδή σε περιπτώσεις που έχουμε να κάνουμε τον έλεγχο $\{H_o : \mu o \nu τέλο \, \mathrm{M}_0 \; vs \; H_1 : \mu o \nu τέλο \, \mathrm{M}_1\}$ όπου $\mathrm{M}_0 \subset \mathrm{M}_1$. Μια τέτοια περίπτωση είναι αν θέλουμε να ελέγξουμε αν τα δεδομένα μας προσαρμόζονται καλύτερα στην Weibull ή στην εκθετική κατανομή, μιας και η εκθετική κατανομή είναι ειδική περίπτωση της Weibull όταν η παράμετρος σχήματος η=1. Άρα μπορούμε να χρησιμοποιήσουμε τον έλεγχο του λόγου των πιθανοφανειών για την υπόθεση $\{H_0 : \eta = 1 \; vs \; H_1 : \eta \neq 1\}$.

2.1.3.3 Έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης

Όπως έχουμε ήδη αναφέρει, ένας πρωταρχικής σημασίας έλεγχος που πρέπει να γίνεται όταν προσαρμόζουμε ένα PH μοντέλο είναι ο έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης όπως αυτή περιγράφεται από τη σχέση (2.5).

Από την βασική ιδιότητα του μοντέλου PH: $h(t;\underset{\sim}{x}) = h_0(t) \cdot \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\sim}{x})$ και τη σχέση (1.6) που συνδέει τις συναρτήσει S(t) και H(t) καταλήγουμε:

$$S(t;\underset{\sim}{x}) = \exp(-H_0(t) \cdot \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\sim}{x})) \Rightarrow \ln(-\ln(S(t;\underset{\sim}{x})) - \ln(H_0(t)) = \underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\sim}{x}$$

Η τελευταία σχέση μας υποδεικνύει ότι η καμπύλη $\ln(-\ln(S(t;\underset{\sim}{x}))$ είναι παράλληλη με την $\ln(H_0(t))$ ως προς τον χρόνο t. Οπότε για να ελέγξουμε την υπόθεση της αναλογικής διακινδύνευσης για οποιαδήποτε συμμεταβλητή $\underset{\sim}{x}_j$, αρκεί να υπολογίσουμε την δική της εκτιμήτρια Kaplan-Meier $S_{KM\,j}(t)$ και να κάνουμε την γραφική παράσταση

$\ln(-\ln(S_{KM\,j}(t)))$ vs t (*). Άρα θα πρέπει να κάνουμε τόσα γραφήματα της μορφής (*) όσο και το πλήθος των συμμεταβλητών στο μοντέλο. Αν ισχύει η υπόθεση αναλογικής διακινδύνευσης για τη j-συμμεταβλητή του μοντέλου τότε οι ευθείες που σχηματίζονται στο γράφημα θα πρέπει να είναι παράλληλες.

Σημειώνεται ότι αυτός ο γραφικός έλεγχος της αναλογικότητας της διακινδύνευσης δεν προϋποθέτει καμία εκ των προτέρων γνώση του παραμετρικής μορφής της $h_0(t)$ άρα μπορεί να χρησιμοποιηθεί το ίδιο αποτελεσματικά σε ημι-παραμετρικά μοντέλα (Caroni, 2017).

2.1.4 Κριτήρια επιλογής μεταβλητών και μέτρα καλής προσαρμογής

Στην παράγραφο 2.1.3 παρουσιάστηκαν συνοπτικά οι έλεγχοι Wald και του λόγου των πιθανοφανειών που μας δίνουν μια πρώτη εκτίμηση για το ποιες είναι οι στατιστικά σημαντικές μεταβλητές του μοντέλου που προσαρμόσαμε. Όμως, θα πρέπει να έχουμε κατά νου ότι, η πρόσθεση μιας ή περισσοτέρων μεταβλητών στο μοντέλο, ακόμα κι αν είναι ελάχιστα στατιστικά σημαντικές, βελτιώνει το βαθμό επεξήγησης της εξαρτημένης μεταβλητής. Άρα, είναι πιθανό να καταλήξουμε σε ένα μοντέλο με πάρα πολλές συμμεταβλητές, γεγονός που αυξάνει την πολυπλοκότητα του προβλήματος. Αυτό που θα θέλαμε ιδανικά, είναι ένα μοντέλο με τον ελάχιστο δυνατό αριθμό συμμεταβλητών που να περιγράφουν πλήρως την εξαρτημένη μας μεταβλητή. Παρακάτω παρουσιάζονται 2 κριτήρια επιλογής του βέλτιστου μοντέλου και ένας βηματικός αλγόριθμος για την επιλογή του καλύτερου συνόλου μεταβλητών που πρέπει να περιέχονται στο μοντέλο.

2.1.4.1Τα κριτήρια AIC και BIC

Το κριτήριο AIC (Akaike's information criterion) αποτελεί ένα κριτήριο επιλογής του βέλτιστου μοντέλου με όσο το δυνατόν μικρότερο αριθμό παραμέτρων. Ορίζεται από τη σχέση (Akaike,1974):

$$AIC = 2d - 2l \quad (2.11)$$

όπου d είναι το πλήθος των παραμέτρων του μοντέλου και l είναι η μεγιστοποιημένη τιμή της λογαριθμοποιημένης συνάρτησης πιθανοφάνειας για το εκτιμημένο μοντέλο.
Συγκρίνοντας όλα τα υποψήφια μοντέλα, επιλέγουμε ως καλύτερο εκείνο με την μικρότερη τιμή του κριτηρίου AIC.

Γνωρίζουμε ότι η προσαρμογή του μοντέλου αυξάνεται όσο εισάγουμε επιπλέον μεταβλητές σε αυτό. Όμως, το κριτήριο AIC δεν προσμετρά μόνο την προσαρμογή, αλλά περιλαμβάνει και ένα είδος ποινής η οποία είναι μια αύξουσα συνάρτηση του αριθμού των παραμέτρων του εκάστοτε μοντέλου. Άρα εισάγοντας μια νέα μεταβλητή στο μοντέλο, ναι μεν βελτιώνεται η προσαρμογή άρα αυξάνεται η πιθανοφάνεια l=lnL ,άρα μειώνεται ο δεύτερος όρος του AIC αλλά από την άλλη ο παράγοντας ποινής d αυξάνεται, άρα αυξάνεται ταυτόχρονα και ο πρώτος όρος του AIC. Επομένως, η εισαγωγή μιας επιπλέον μεταβλητής μπορεί να οδηγήσει σε ένα καλύτερο μοντέλο (μικρότερο AIC) μόνο αν αυξάνει την προσαρμογή του τόσο ώστε να ξεπεράσει την ποινή του όρου 2d.

Το κριτήριο BIC (Bayesian information criterion), είναι ένα ακόμα κριτήριο επιλογής του βέλτιστου μοντέλου μεταξύ μοντέλων με διαφορετικό αριθμό παραμέτρων, όπως και το AIC. Αν και η αφετηρία του είναι διαφορετική από εκείνη του AIC, η μορφή και η λειτουργία του είναι παρόμοια με εκείνη του AIC. Η βασική τους διαφορά είναι ότι η εισαγωγή επιπλέον μεταβλητών στο μοντέλο αποθαρρύνεται σε μεγαλύτερο βαθμό από το AIC, δηλαδή η

αντίστοιχη ποινή βαραίνει περισσότερο το κομμάτι της αύξησης της προσαρμογής. Στη γενική περίπτωση το κριτήριο BIC ορίζεται ως (Buckland et al., 1997):

$$BIC = d \cdot \ln(n) - 2l \quad (2.12)$$

όπου πάλι d είναι το πλήθος παραμέτρων του υπό εξέταση μοντέλου, l είναι η μεγιστοποιημένη τιμή της λογαριθμοποιημένης συνάρτησης πιθανοφάνειας και n είναι το μέγεθος του δείγματος. Όπως και στο κριτήριο AIC, επιλέγουμε ως βέλτιστο μοντέλο εκείνο με τη μικρότερη τιμή του κριτηρίου BIC.

2.1.4.2 Η μέθοδος της διαδοχικής αφαίρεσης (backward elimination)

Ο αλγόριθμος της διαδοχικής αφαίρεσης, εκτελεσμένος με βάση το κριτήριο AIC, περιγράφεται από τα ακόλουθα βήματα:

Α) Αρχικά προσαρμόζουμε το μοντέλο που περιέχει όλες τις διαθέσιμες μεταβλητές.

Β) Αφαιρούμε τη λιγότερο στατιστικά σημαντική μεταβλητή, δηλαδή εκείνη που συμβάλει λιγότερο στο μοντέλο χρησιμοποιώντας το κριτήριο AIC. Καθορίζουμε, δηλαδή, ποια είναι η μεταβλητή η οποία αν απαλειφθεί από το μοντέλο θα μας δώσει τη μικρότερη αύξηση του κριτηρίου AIC αν επαναπροσαρμόσουμε το μοντέλο χωρίς αυτή.

Γ) Επαναπροσαρμόζουμε το μοντέλο που περιέχει όλες τις μεταβλητές εκτός από αυτή που απαλείψαμε από το προηγούμενο βήμα, και εκτελούμε ξανά τον ίδιο έλεγχο εντοπισμού της λιγότερο σημαντικής μεταβλητής χρησιμοποιώντας το κριτήριο AIC.

Δ) Επαναλαμβάνουμε τα δύο προηγούμενα βήματα μέχρι να καταλήξουμε σε ένα μοντέλο όπου η αφαίρεση οποιασδήποτε μεταβλητής αυξάνει την τιμή του AIC, οπότε και σταματάμε.

2.1.4.3 Η μέθοδος της διαδοχικής πρόσθεσης (forward selection)

Ο αλγόριθμος της διαδοχικής πρόσθεσης έχει παρόμοια λογική με την μέθοδο διαδοχικής αφαίρεσης ξεκινώντας απλά από διαφορετική αφετηρία. Ο αλγόριθμος της μεθόδου περιγράφεται από τα ακόλουθα βήματα:

Α) Αρχικά προσαρμόζουμε το σταθερό μοντέλο, δηλαδή αυτό που δεν περιέχει καμία μεταβλητή.

Β) Εισάγουμε τη στατιστικά σημαντικότερη μεταβλητή, δηλαδή εκείνη που συμβάλει περισσότερο στην επεξήγηση της μεταβλητής απόκρισης χρησιμοποιώντας το κριτήριο AIC. Καθορίζουμε, δηλαδή, ποια είναι η μεταβλητή η οποία αν προστεθεί στο μοντέλο θα δώσει τη μεγαλύτερη μείωση του κριτηρίου AIC αν επαναπροσαρμόσουμε το μοντέλο που την περιλαμβάνει.

Γ) Επαναπροσαρμόζουμε το μοντέλο το οποίο περιέχει μόνο την μεταβλητή που επιλέξαμε στο προηγούμενο βήμα και επαναπροσδιορίζουμε με τον ίδιο τρόπο ποια θα είναι η επόμενη μεταβλητή η οποία θα εισαχθεί στο μοντέλο.

Δ) Επαναλαμβάνουμε τα δύο προηγούμενα βήματα έως ότου η πρόσθεση οποιασδήποτε από τις εναπομένουσες μεταβλητές να αυξήσει την τιμή του AIC, οπότε και σταματάμε.

2.2 Το ημι-παραμετρικό μοντέλο του Cox

Συχνά στην ανάλυση επιβίωσης θέλουμε να προσαρμόσουμε μεγάλα σύνολα δεδομένων διάρκειας ζωής που αφορούν τον άνθρωπο. Σε αντίθεση με τα δεδομένα διάρκειας ζωής από τεχνολογικές εφαρμογές όπου, μπορούμε να καταλήξουμε σε ένα παραμετρικό μοντέλο με βάση προηγούμενης εμπειρίας που έχουμε με παρόμοιο υλικό το γεγονός ότι κάθε άτομο είναι διαφορετικό και φέρει τα δικά του μοναδικά χαρακτηριστικά κάνει το έργο της προσαρμογής ενός και μόνο γνωστού παραμετρικού μοντέλου ακατόρθωτο. Έτσι, θέλοντας να προσαρμόσουμε ένα μοντέλο αναλογικής διακινδύνευσης δεν είμαστε σε θέση να γνωρίζουμε την παραμετρική μορφή της κοινής συνάρτησης κινδύνου $h_0(t)$ και οδηγούμαστε στα λεγόμενα ημι-παραμετρικά μοντέλα PH. Το πιο γνωστό μοντέλο αυτής της μορφής είναι το μοντέλο αναλογικής διακινδύνευσης του Cox (the Cox proportional hazards model) που πρωτοεισήχθη από τον Άγγλο στατιστικολόγο David Cox (Cox, (1972)). Το μοντέλο του Cox χρησιμοποιείται ευρέως για τον προσδιορισμό διαφορών στην επιβίωση παρατηρούμενων μονάδων όταν υποβάλλονται σε διάφορες θεραπείες καθώς και σε πολλά άλλα προβλήματα βιοϊατρικής (Hosmer, Lemeshow & May, (2008)).

2.2.1 Ορισμός του μοντέλου και εκτίμηση παραμέτρων

Το μοντέλο του Cox αποτελεί ειδική περίπτωση μοντέλων αναλογικής διακινδύνευσης. Επομένως, θεωρούμε ότι οι συμμεταβλητές x_i δρουν στη συνάρτηση διακινδύνευσης μέσω της σχέσης:

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

όπου πάλι η $h_0(t)$ είναι η κοινή συνάρτηση διακινδύνευσης για όλες τις μονάδες του πληθυσμού και $\underset{\rightarrow}{\beta}$ είναι ένα διάνυσμα παραμέτρων όπου κάθε συνιστώσα του β_i αναπαριστά την ποσοτική επίδραση της αντίστοιχης συμμεταβλητής x_i στην διάρκεια ζωής. Ένας ισοδύναμος ορισμός της $h_0(t)$ είναι ότι αποτελεί τη συνάρτηση διακινδύνευσης ενός ατόμου όταν όλες οι συμμεραβλητές x_i λάβουν την τιμή μηδέν.

Αντίστοιχα με τα κοινά μοντέλα PH μπορούμε να βρούμε τη σωρευτική συνάρτηση διακινδύνευσης:

$$\mathrm{H}(t;\underset{\rightarrow}{x}) = H_0(t)\cdot \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}}\cdot \underset{\rightarrow}{x})$$

και επίσης την συνάρτηση επιβίωσης:

$$\mathrm{S}(t;\underset{\rightarrow}{x}) = [S_0(t)]^{\exp(\underset{\rightarrow}{\beta}^{\mathrm{T}}\cdot \underset{\rightarrow}{x})}$$

Το κύριο χαρακτηριστικό του μοντέλου του Cox, όπως αναφέραμε και στον πρόλογο, είναι ότι οι συγκεκριμένες παραμετρικές μορφές των συναρτήσεων $h_0(t), S_0(t)$ δεν καθορίζονται και μόνο η επίδραση του διανύσματος των συμμεταβλητών $\underset{\rightarrow}{x}$ αναλύεται. Παρόλα αυτά πρέπει να τονίσουμε ότι η υπόθεση της αναλογικότητας εξακολουθεί να ισχύει όπως και στα κλασσικά PH μοντέλα, καθώς αυτή δεν εξαρτάται από τον χρόνο.

Έστω λοιπόν ότι έχουμε n το πλήθος παρατηρήσεις εκ των οποίων στις k<n έχει συμβεί το υπό μελέτη γεγονός. Έστω επίσης, οι διακεκριμένοι χρόνοι διακοπής $t_{(1)} < t_{(2)} < ... < t_{(k)}$. Σε αυτό το σημείο θα θεωρήσουμε για απλότητα ότι κάθε τυχαία χρονική στιγμή $t_{(j)}$ συμβαίνει το γεγονός σε ακριβώς μια μονάδα του πληθυσμού, δηλαδή θα έχουμε:
$d_j = 1, \forall j = 1, ..., k$ αν και αυτό δεν είναι δεσμευτικό και μπορούμε να γενικεύσουμε το πρόβλημα μας όπως θα δούμε και στην παράγραφο 2.2.2

Ορίζουμε ως R_j το σύνολο των μονάδων που βρίσκονταν σε κίνδυνο ακριβώς πριν τη χρονική στιγμή $t_{(j)}$. Άρα λοιπόν, η πιθανότητα να συμβεί το γεγονός (το συμβολίζουμε ως A_j) σε μια δεδομένη μονάδα j, δοθέντος ότι συμβαίνει το γεγονός σε μια οποιαδήποτε μονάδα του συνόλου R_j θα είναι:

$$p_j = P(A_j \mid \bigcup_{i \in R_j} A_i) = \frac{P(A_j)}{P(\bigcup_{i \in R_j} A_i)} = \frac{P(A_j)}{\sum_{i \in R_j} P(A_i)} \quad (*)$$

η τελευταία ισότητα της σχέσης (*) προέκυψε από το γεγονός ότι: $A_k \bigcap A_\lambda = \varnothing, \forall k \neq \lambda$, οπότε $P(\bigcup_{i \in R_j} A_i) = \sum_{i \in R_j} P(A_i)$ διότι όλα τα ενδεχόμενα A_i είναι ξένα μεταξύ τους. Αυτό συμβαίνει επειδή από την αρχική μας υπόθεση θεωρήσαμε ότι κάθε χρονική στιγμή συμβαίνει το γεγονός ακριβώς σε μία μονάδα. Επιπλέον, γνωρίζουμε ότι η πιθανότητα να συμβεί το γεγονός σε μια μονάδα με διάνυσμα συμμεταβλητών $\underset{\rightarrow}{x}$ στο χρονικό διάστημα (t,t+δt), δεδομένου των μονάδων σε ρίσκο τη χρονική στιγμή t, είναι $h(t; \underset{\rightarrow}{x}) \cdot \delta t$. Άρα, η σχέση (*)γράφεται ισοδύναμα και ως:

$$p_j = \frac{\exp(\underset{\rightarrow}{\beta}^T \cdot \underset{\rightarrow}{x}_j)}{\sum_{i \in R_j} \exp(\underset{\rightarrow}{\beta}^T \cdot \underset{\rightarrow}{x}_i)} \quad (2.14)$$

Με αυτόν τον τρόπο είναι ουσιαστικά σαν να έχουμε ορίσει τις πιθανότητες p_j μιας διακριτής τυχαίας μεταβλητής για j=1,..,k όπου k είναι το πλήθος των μη αποκομμένων παρατηρήσεων. Ορίζουμε λοιπόν την συνάρτηση πιθανοφάνειας για την εκτίμηση των παραμέτρων του μοντέλου:

$$L(\underset{\rightarrow}{\beta}) = \prod_{j=1}^{k} \left(\frac{\exp(\underset{\rightarrow}{\beta}^T \cdot \underset{\rightarrow}{x}_j)}{\sum_{i \in R_j} \exp(\underset{\rightarrow}{\beta}^T \cdot \underset{\rightarrow}{x}_i)} \right) \quad (2.15)$$

Η παραπάνω συνάρτηση πιθανοφάνειας ονομάστηκε συνάρτηση μερικής πιθανοφάνειας (partial likelihood) (Cox, 1975) καθώς η κοινή συνάρτηση διακινδύνευσης $h_0(t)$ παραμένει άγνωστη αλλά παρόλα αυτά δεν επηρεάζει την ανάλυσή μας καθώς δεν συμμετέχει στην έκφραση της πιθανοφάνειας. Όπως και στην κλασσική μέθοδο μέγιστης πιθανοφάνειας θα εκτιμήσουμε το διάνυσμα παραμέτρων $\underset{\rightarrow}{\beta}$ k-συντελεστών από το $\underset{\rightarrow}{\beta}$ που μεγιστοποιεί την λογαριθμοποιημένη συνάρτηση πιθανοφάνειας ln(L). Θα πρέπει να τονίσουμε σε αυτό το σημείο ότι ο όρος $\underset{\rightarrow}{\beta}^T \cdot \underset{\rightarrow}{x}$ δεν περιλαμβάνει τον σταθερό όρο β_0 καθώς αυτός θεωρούμε ότι

έχει απορροφηθεί στην κοινή συνάρτηση διακινδύνευσης $h_0(t)$. Λογαριθμίζοντας την σχέση (2.15) καταλήγουμε λοιπόν:

$$l(\underset{\rightarrow}{\beta}) = \ln(L(\underset{\rightarrow}{\beta})) = \sum_{j=1}^{k}\{\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_j - \ln[\sum_{i\in R_j}\exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_i)]\} \quad (2.16)$$

Για να υπολογίσουμε το διάνυσμα των εκτιμητριών $\underset{\rightarrow}{\beta}^{T} = [\beta_1,...,\beta_k]$ θα πρέπει να βρούμε τις k-το πλήθος μερικές παραγώγους πρώτης τάξης της $l(\underset{\rightarrow}{\beta})$ ως προς β_j ,j=1,...,k και να λύσουμε το σύστημα εξισώσεων:

$$\frac{\partial l}{\partial \beta_r} = 0, \mathrm{r} = 1,...,k \Rightarrow \sum_{j=1}^{k} x_{jr} - \sum_{j=1}^{k}[(\sum_{i\in R_j} x_{ir} \cdot \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_i)) / (\sum_{i\in R_j}\exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_i))] = 0 \quad (2.17)$$

όπου έχουμε ορίσει ως x_{jr} την r-οστή συνιστώσα του διανύσματος συμμεταβλητών $\underset{\rightarrow}{x}_j$.Το σύστημα εξισώσεων (2.17) επιλύεται μόνο με αριθμητικές μεθόδους (Caroni, 2017).

Τέλος, πολλές φορές θέλουμε να υπολογίσουμε τα τυπικά σφάλματα $se(\beta_j), j = 1,..,k$ των εκτιμητριών. Ξέρουμε, γενικά ότι αν θ η εκτιμήτρια μέγιστης πιθανοφάνειας της παραμέτρου θ τότε ισχύει: $\theta \sim \mathrm{N}(\theta, V(\theta))$ και το τυπικό σφάλμα $se(\theta) = \sqrt{V(\theta)}$. Η εκτιμήτρια της διασποράς δίνεται από τον τύπο:

$$V(\theta) = (\frac{-\partial^2 l}{\partial \theta^2})^{-1}|_{\theta=\theta}$$

Άρα για τον υπολογισμό του τυπικού σφάλματος της εκτιμήτριας β_r χρειαζόμαστε την μερική παράγωγο δεύτερης τάξης:

$$-\frac{\partial^2 l}{\partial \beta_r^2} = \left[\frac{\sum_{i\in R_j} x_{ir}^2 \cdot \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_i)}{\sum_{i\in R_j}\exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_i)} - \left(\frac{\sum_{i\in R_j} x_{ir} \cdot \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_i)}{\sum_{i\in R_j}\exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_i)}\right)^2\right] \quad (2.18)$$

2.2.2 Ισόπαλοι χρόνοι διακοπής

Στην παράγραφο 2.2.1 πριν τον υπολογισμό της μερικής συνάρτησης πιθανοφάνειας θεωρήσαμε ως σύμβαση ότι σε κάθε διακεκριμένη χρονική στιγμή $t_{(j)}$,j=1,..,k το υπό μελέτη γεγονός συμβαίνει σε ακριβώς μια μονάδα του πληθυσμού μας, δηλαδή ότι: $d_j = 1, \forall j$. Συχνά, σε ιατρικές εφαρμογές όπου καταγράφεται ο χρόνος ζωής των ατόμων που πάσχουν από μια ασθένεια, παρότι ο χρόνος Τ είναι μια συνεχής τ.μ για λόγους ευκολίας γίνονται κατάλληλες στρογγυλοποιήσεις και έτσι ουσιαστικά καταλήγουμε σε διακριτούς χρόνους. Παραδείγματος χάριν, αν ένας ερευνητής μετράει το χρόνο ζωής ατόμων σε μήνες από την αρχή της ιατρικής τους παρακολούθησης, τότε αν κάποιο άτομο αποβιώσει σε 12 μήνες ακριβώς και κάποιο άλλο σε 12 μήνες και 5 μέρες, ενώ η διάρκεια ζωής τους είναι διαφορετική στην στρογγυλοποίηση

που θα γίνει θα καταγραφούν οι χρόνοι $t_1 = t_2 = 12$. Επομένως, καταλαβαίνουμε ότι το σενάριο να προκύψει $d_j > 1$ σε κάποια χρονική στιγμή $t_{(j)}$, j=1,…,k είναι άκρως ρεαλιστικό. Αυτοί οι χρόνοι $t_{(j)}$ ονομάζονται ισόπαλοι χρόνοι διακοπής.

Κατά συνέπεια, θα πρέπει να αλλάξει και η μορφή της συνάρτησης πιθανοφάνειας όπως αυτή ορίστηκε στην παράγραφο 2.2.1. Το βασικό πρόβλημα που αντιμετωπίζουμε είναι ότι τη χρονική στιγμή $t_{(j)}$ το πλήθος των μονάδων $d_j > 1$ στις οποίες συνέβη το γεγονός θεωρητικά θα προέκυπταν σε διαφορετικούς χρόνους αν δεν είχαμε χρησιμοποιήσει στρογγυλοποίηση και θέλαμε να κρατήσουμε περισσότερα σημαντικά ψηφία στις μετρήσεις μας. Αυτό όμως σημαίνει ότι θα μπορούσαν να είχαν προκύψει με οποιαδήποτε σειρά μεταξύ τους και άρα θα υπήρχαν $d_j !$ πιθανές σειρές εμφάνισης. Η συνάρτηση πιθανοφάνειας θα έπρεπε κανονικά να τις περιλαμβάνει όλες κάτι όμως που θα αύξανε την πολυπλοκότητα της και ταυτόχρονα τον υπολογισμό των εκτιμητριών β_j. Μια πολύ συνηθισμένη προσέγγιση, είναι αυτή του Breslow (Breslow, 1974) κατά την οποία η συνάρτηση πιθανοφάνειας λαμβάνει τη μορφή:

$$L_{Breslow} = \prod_{j=1}^{k} \frac{\exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{z}_j)}{\left(\sum_{i \in R_j} \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_i) \right)^{d_j}} \quad (2.19)$$

όπου έχουμε ορίσει ως $\underset{\rightarrow}{z}_j = \sum_{s=1}^{d_j} \underset{\rightarrow}{x}_s$, με το $\underset{\rightarrow}{x}_s$ να είναι το διάνυσμα συμμεταβλητών της μονάδας s, στην οποία συμβαίνει το γεγονός τη χρονική στιγμή $t_{(j)}$ και s=1,.., d_j. Η προσέγγιση αυτή θεωρείται ακριβής όσο η ποσότητα d_j / n_j παραμένει σχετικά μικρή.

Αν όμως η ποσότητα d_j / n_j δεν είναι μικρή τότε, χρησιμοποιούμε την προσέγγιση του Cox (1972) κατά την οποία κάνουμε την παραδοχή ότι τα δεδομένα μας παρατηρήθηκαν σε διακριτή αντί σε συνεχή κλίμακα. Σύμφωνα με την προσέγγιση αυτή, δεδομένου ότι την χρονική στιγμή $t_{(j)}$ συμβαίνουν d_j διακοπές λειτουργίας, η πιθανότητα να προκύψει ένα οποιοδήποτε σύνολο u αποτελούμενο από d_j μονάδες, είναι $P(u) \propto \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{z}_u)$. Όπως και στην προσέγγιση του Breslow $\underset{\rightarrow}{z}_u = \sum_{s=1}^{d_j} \underset{\rightarrow}{x}_s$, όπου $\underset{\rightarrow}{x}_s$ είναι το διάνυσμα συμμεταβλητών των μονάδων του συνόλου u. Τότε, η υπό συνθήκη πιθανότητα του παρατηρούμενου συνόλου μονάδων u^* με διακοπή δίνεται από τον τύπο:

$$P(u^* \mid d_j) = \frac{\exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{z}_j)}{\sum_{u \in R_j} \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{z}_u)} \quad (2.20)$$

Άρα η προσέγγιση της συνάρτησης πιθανοφάνειας Cox είναι:

$$L_{Cox} = \prod_{j=1}^{k} P(u^* \mid d_j) \quad (2.21)$$

Υπενθυμίζουμε ότι στη σχέση (2.20) ο παρανομαστής αποτελείται από το άθροισμα όλων των δυνατών $\begin{pmatrix} n_j \\ d_j \end{pmatrix}$ όρων (Collett, 2003).

2.2.3 Τα υπόλοιπα Schoenfeld

Τα υπόλοιπα που μελετάμε στο ημι-παραμετρικό μοντέλο του Cox, και θα τα χρησιμοποιήσουμε και στην παράγραφο 2.2.4 για τον έλεγχο της αναλογικότητας της διακινδύνευσης, είναι τα λεγόμενα υπόλοιπα Schoenfeld ή αλλιώς μερικά υπόλοιπα (partial residuals).

Όταν ορίσαμε τη συνάρτηση μερικής πιθανοφάνειας για την εκτίμηση των παραμέτρων στο μοντέλο του Cox είδαμε από την σχέση (2.14) ότι οι ποσότητες:

$$p_j = \frac{\exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_j)}{\sum_{i \in R_j} \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_i)}$$

εκφράζουν την πιθανότητα να συμβεί το υπό μελέτη γεγονός στην j-οστή μονάδα του πληθυσμού τη χρονική στιγμή $t_{(j)}, j = 1,...,k$, δεδομένου του συνόλου R_j των μονάδων που βρίσκονται σε κίνδυνο ακριβώς πριν τη χρονική στιγμή $t_{(j)}$. Εφόσον, όμως, δεν γνωρίζουμε σε ποια από όλες τις μονάδες του R_j πρόκειται να συμβεί το γεγονός, οι τιμές του διανύσματος συμμεταβλητών $\underset{\rightarrow}{x}$ της εν λόγω μονάδας αποτελούν μια διακριτή τυχαία μεταβλητή και οι πιθανότητες p_j είναι οι αντίστοιχες πιθανότητες εμφάνισης. Άρα μπορούμε να ορίσουμε την μέση τιμή της τ.μ ως εξής:

$$E(\underset{\rightarrow}{x} \mid R_j) = \sum_{s \in R_j} \underset{\rightarrow}{x}_s \cdot p_s = \frac{\sum_{s \in R_j} \underset{\rightarrow}{x}_s \cdot \exp(\underset{\rightarrow}{\beta}^{T} \cdot \underset{\rightarrow}{x}_s)}{\sum_{i \in R_j} \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_i)} \quad (2.22)$$

Σε αυτό το σημείο, είμαστε σε θέση να ορίσουμε ως υπόλοιπο, κατά τη γνωστή μεθοδολογία, την απόκλιση της παρατηρούμενης τιμής από την αναμενομένη. Άρα λοιπόν θα έχουμε:

$$\underset{\rightarrow}{r}_j = \underset{\rightarrow}{x}_j - E(\underset{\rightarrow}{x} \mid R_j)$$

Αντικαθιστώντας, εν τέλει το διάνυσμα παραμέτρων $\underset{\rightarrow}{\beta}$ του μοντέλου με το εκτιμημένο $\underset{\rightarrow}{\beta}$ προκύπτουν τα υπόλοιπα Schoenfeld (Schoenfeld, 1982):

$$\hat{\underset{\rightarrow}{r}}_j = \underset{\rightarrow}{x}_j - E(\underset{\rightarrow}{x} \mid R_j) \quad (2.23)$$

Θα πρέπει να επισημάνουμε σε αυτό το σημείο ότι σε αντίθεση με τα κλασσικά μοντέλα παλινδρόμησης όπου ορίζαμε τα υπόλοιπα $e_i = y_i - y_i$, στα υπόλοιπα Schoenfeld δεν χρησιμοποιούνται οι τιμές της εξαρτημένης μεταβλητής (εδώ του χρόνου t) αλλά τα αντίστοιχα διανύσματα συμμεταβλητών $\underset{\rightarrow}{x}_j$. Επιπλέον, θα πρέπει να ξέρουμε ότι τα υπόλοιπα Schoenfeld

υπολογίζονται στις μη αποκομμένες παρατηρήσεις αν και λόγω του ορισμού των πιθανοτήτων p_j προφανώς λαμβάνονται και αυτές υπόψη. Τέλος, είναι σημαντικό να παρατηρήσουμε ότι στον ορισμό (2.23) τα υπόλοιπα Schoenfeld έχουν διανυσματική μορφή καθώς για κάθε μη αποκομμένη παρατήρηση υπολογίζονται τόσα υπόλοιπα όσο και το πλήθος των συμμεταβλητών του μοντέλου που προσαρμόσαμε.

Αν σε αυτό το σημείο ορίσουμε ως $V(\hat{\underset{\rightarrow}{r}}_j)$ τον τετραγωνικό πίνακα p-διάστασης (όπου p το πλήθος των συμμεταβλητών του μοντέλου) των εκτιμημένων διασπορών των $\hat{\underset{\rightarrow}{r}}_j$ τότε μπορούμε να μετασχηματίσουμε τα κλασσικά υπόλοιπα Schoenfeld ως εξής:

$$\tilde{\underset{\rightarrow}{r}}_j = (V(\hat{\underset{\rightarrow}{r}}_j))^{-1} \cdot \hat{\underset{\rightarrow}{r}}_j \quad (2.24)$$

Η σχέση (2.24) εκφράζει μια σταθμισμένη εκδοχή των κλασσικών υπολοίπων Schoenfeld που χρησιμοποιείται συχνά για τον εντοπισμό προβλημάτων στο μοντέλο του Cox. Όμως στον ορισμό (2.24) χρησιμοποιείται ο αντίστροφος πίνακας $(V(\hat{\underset{\rightarrow}{r}}_j))^{-1}$ αυτό μπορεί να κάνει τον υπολογισμό των $\tilde{\underset{\rightarrow}{r}}_j$ πολύ χρονοβόρο ειδικά σε μοντέλα με πολλές συμμεταβλητές. Για αυτό το λόγο χρησιμοποιείται η προσέγγιση: $(V(\hat{\underset{\rightarrow}{r}}_j))^{-1} \cong k \cdot V(\underset{\rightarrow}{\beta})$ (όπου k το πλήθος των μη αποκομμένων παρατηρήσεων) λόγω της οποίας καταλήγουμε στα λεγόμενα κλιμακοποιημένα (scaled) υπόλοιπα Schoenfeld (Grambsch & Therneau, 1994)

$$\underset{\rightarrow}{r}_j^* = k \cdot V(\underset{\rightarrow}{\beta}) \cdot \hat{\underset{\rightarrow}{r}}_j \quad (2.25)$$

2.2.4 Έλεγχος της υπόθεσης αναλογικής διακινδύνευσης στο μοντέλο του Cox

Όπως έχουμε αναφέρει και στα κλασσικά παραμετρικά μοντέλα αναλογικής διακινδύνευσης, πρωταρχικό ρόλο στην ανάλυση μας είναι ο έλεγχος της ιδιότητας της αναλογικής διακινδύνευσης που θα πρέπει να έχουν τα μοντέλα που προσαρμόζουμε. Στην παράγραφο 2.1.3.3 είδαμε έναν γραφικό τρόπο ελέγχου αυτής της υπόθεσης ο οποίος μπορεί να εφαρμοστεί και στο ημι-παραμετρικό μοντέλο του Cox, καθώς για την υλοποίηση του χρειάζονται μόνο οι μη-παραμετρικές εκτιμήτριες Kaplan-Meier για κάθε συμμεταβλητή του μοντέλου. Σε αυτήν τη παράγραφο θα εξετάσουμε δύο επιπλέον μεθοδολογίες ελέγχου της αναλογικότητας της διακινδύνευσης. Ο πρώτος είναι ένας γραφικός έλεγχος των κλιμακοποιημένων υπολοίπων Schoenfeld που ορίστηκαν στην προηγούμενη παράγραφο και ο δεύτερος είναι το λεγόμενο Global test που εξετάζει την εξάρτηση των παραμέτρων β_i του μοντέλου από τον χρόνο.

2.2.4.1 Γραφικός έλεγχος των υπολοίπων Schoenfeld

Η υπόθεση της αναλογικής διακινδύνευσης μπορεί να διατυπωθεί ισοδύναμα και ως:

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

Μπορεί να αποδειχθεί (Grambsch & Therneau, 1994) ότι ισχύει προσεγγιστικά ότι:

$$E(\mathrm{r}_{ij}^{*}) \cong \beta_i(t_{(j)}) - \beta_i \quad (2.26)$$

όπου $\beta_i(t_{(j)})$ είναι ο συντελεστής της i-οστής συμμεταβλητής του μοντέλου τη χρονική στιγμή $t_{(j)}$. Επομένως, για να αποδεχτούμε την υπόθεση $\beta_i(t) = \beta_i, \forall t$ θα πρέπει η γραφική παράσταση:

$$(\mathrm{r}_{ij}^{*} + \beta_i) \text{ ως προς } t_{(j)} \quad (2.27)$$

να δείχνει μια οριζόντια γραμμή (δηλαδή ανεξαρτησία ως προς τον χρόνο).

Επισημαίνουμε ότι θα χρειαστούμε τόσα γραφήματα της μορφής (2.27) όσο και το πλήθος των συμμεταβλητών που μετέχουν στο μοντέλο μας.

2.2.4.2 Ο έλεγχος Global

Για να ερμηνεύσουμε τα οπτικά αποτελέσματα που πήραμε από τον γραφικό έλεγχο των υπολοίπων Schoenfeld χρησιμοποιούμε το λεγόμενο στατιστικό ελέγχου Global. Συγκεκριμένα, εκφράζουμε την εξάρτηση της $\beta_i(t)$ από τη διάρκεια t ή γενικότερα από μια συνάρτηση g(t) μέσω της παλινδρόμησης

$$\beta_i(t) = \beta_i + \theta_i(g(t) - \bar{g}),\ i = 1, \ldots, p$$

όπου $\bar{g}$ συμβολίζουμε τον μέσο όρο των $g_j = g(t_{(j)})$ και έστω p το πλήθος των συμμεταβλητών.

Στη συνέχεια μια ανάλυση βασισμένη στην εκτίμηση της θ_i με την μέθοδο των γενικευμένων ελαχίστων τετραγώνων, οδηγεί σε έναν ισοδύναμο έλεγχο, που εξετάζει την αναλογικότητα της διακινδύνευσης στο μοντέλο του Cox ο οποίος είναι:

$$\{H_0 : \theta_i = 0, \forall i = 1, \ldots, p \ vs\ H_1 : \theta_j \neq 0, \text{ για κάποιο j}\} \quad (2.28)$$

Ο έλεγχος (2.28) χρησιμοποιεί την ελεγχοσυνάρτηση:

$$T = \frac{(\underrightarrow{g} - \underrightarrow{\bar{g}})^T S^{*} I(\underrightarrow{\beta}) \mathrm{S}^{*T} (\underrightarrow{g} - \underrightarrow{\bar{g}})}{k \sum_{j=1}^{k} (g_j - \bar{g})^2} \square\ \mathrm{X}^2_{(\mathrm{p})} \quad (2.29)$$

όπου έχουν οριστεί ως:

- k, ο αριθμός των μη αποκομμένων παρατηρήσεων
- p, το πλήθος των συμμεταβλητών
- $(\underrightarrow{g} - \underrightarrow{\bar{g}})$, το διάνυσμα k-διάστασης με j-οστό στοιχείο το $g_j = g(t_{(j)})$
- S^{*},ο kxp διαστάσεων πίνακας των κλιμακοποιημένων υπολοίπων

- $I^{-1}(\underrightarrow{\beta}) = V(\underrightarrow{\beta})$

Απορρίπτουμε την μηδενική υπόθεση, άρα και την υπόθεση της αναλογικότητας της διακινδύνευσης, για μεγάλες τιμές της ελεγχοσυνάρτησης T (Grambsch & Therneau, 1994)

Αν τώρα θέλουμε να κάνουμε τον παραπάνω έλεγχο για την i-οστή συμμεταβλητή του μοντέλου η ελεγχοσυνάρτηση (2.29) μετασχηματίζεται στην:

$$T_i = \frac{\left[\sum_{i=1}^{k}(g_j - \bar{g})r_{ij}^*\right]^2}{kI_{ii}\sum_{j=1}^{k}(g_j - \bar{g})^2} \sim X^2_{(1)} \quad (2.30)$$

όπου έχουμε ορίσει ως:

- r_{ij}^*, το (i,j)-οστό στοιχείο του πίνακα S^* που εισήγαμε παραπάνω
- I_{ii}, το i-οστό διαγώνιο στοιχείο του αντίστροφου πίνακα πληροφορίας $I^{-1}(\underrightarrow{\beta})$

2.2.5 Σημεία επιρροής στο μοντέλο του Cox

Ένα σημαντικό κριτήριο ελέγχου της καταλληλότητας ενός μοντέλου είναι εύρεση εκείνων των παρατηρήσεων που επηρεάζουν σε μεγάλο βαθμό τα αποτελέσματα. Δηλαδή, ας υποθέσουμε ότι αφαιρώντας μια παρατήρηση από το μοντέλο μας αυτό έχει σαν αποτέλεσμα να μεταβληθεί σημαντικά ο κίνδυνος να συμβεί το υπό μελέτη γεγονός. Αυτές οι παρατηρήσεις ονομάζονται σημεία επιρροής. Τέτοια σημεία ενδέχεται να συμπεριλαμβάνονται στο σύνολο δεδομένων μας αν έχει γίνει κακή προσαρμογή του μοντέλου και έχει εξαιρεθεί μια σημαντική μεταβλητή ή απλώς έχει γίνει λάθος κατά το στάδιο συλλογής των μετρήσεων.

Έστω λοιπόν ότι θέλουμε να ελέγξουμε κατά πόσο μια παρατήρηση επηρεάζει την εκτιμήτρια β_j της παραμέτρου β_j, j=1,..,p του μοντέλου του Cox. Σε αυτή την περίπτωση, πρέπει πρώτα να προσαρμόσουμε το μοντέλο του Cox που περιέχει το σύνολο των n παρατηρήσεων και εν συνεχεία να γίνει η αναπροσαρμογή έχοντας παραλείψει ακριβώς εκείνη την παρατήρηση που εικάζουμε ότι επηρεάζει σε μεγάλο βαθμό την j-οστή παράμετρο του μοντέλου. Η ίδια διαδικασία μπορεί να ακολουθηθεί για καθεμία από τις n παρατηρήσεις του συνόλου δεδομένων.

Αν ορίσουμε ως β_j την εκτιμήτρια της παραμέτρου β_j και $\beta_{j(i)}$ την εκτιμήτρια της β_j έχοντας εξαιρέσει την i-οστή παρατήρηση τότε μπορεί να οριστεί το εξής μέτρο επιρροής (Belsley et al., 1980):

$$DFBETAS_{ji} = \frac{\beta_j - \beta_{j(i)}}{\sqrt{S^2_{(i)}c_{jj}}},\ i = 1,...,n\ \ j = 1,...,\mathrm{p} \quad (2.31)$$

όπου $S^2_{(i)}$ είναι η εκτιμήτρια της διασποράς σ^2 που προκύπτει από την προσαρμογή του μοντέλου όταν έχει αφαιρεθεί η i-οστή παρατήρηση και c_{jj} το j-οστό διαγώνιο στοιχείο του πίνακα $(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}$, με X τον πίνακα σχεδιασμού.

Μεγάλη τιμή του $DFBETAS_{ji}$ και ιδιαίτερα αν $\left|DFBETAS_{ji}\right| > 2/\sqrt{n}$, υποδεικνύει ότι η i-οστή παρατήρηση έχει μεγάλη επιρροή στην εκτίμηση του j-οστού συντελεστή του μοντέλου (Καρώνη, 2017).

Ένα βασικό μειονέκτημα αυτής της μεθόδου είναι η ανάγκη αναπροσαρμογής του μοντέλου ώστε να μπορούμε να αποφανθούμε για την επιρροή κάποιας συγκεκριμένης παρατήρησης. Αυτό εν γένει ενδέχεται να είναι πολύ χρονοβόρο ιδιαίτερα αν μελετάμε ένα μεγάλο σύνολο δεδομένων και έχουμε στη διάθεση μας πάρα πολλές συμμεταβλητές που μετέχουν στο μοντέλο. Ένας τρόπος να παρακάμψουμε αυτή τη διαδικασία είναι αυτή που περιγράφεται από τους Cain & Lange (Cain & Lange, 1984). Αν ορίσουμε ως $\underset{\rightarrow}{\beta} - \underset{\rightarrow}{\beta}_{(j)}$ την επιρροή της j-οστής παρατήρησης στην εκτίμηση $\underset{\rightarrow}{\beta}$,τότε χρησιμοποιώντας τον πρώτο όρο της επέκτασης Taylor της συνάρτησης score η οποία εκφράζει πόσο ευαίσθητη είναι η συνάρτηση πιθανοφάνειας $L(\underset{\rightarrow}{\theta};\underset{\rightarrow}{x})$ ως προς την παράμετρο $\underset{\rightarrow}{\theta}$, καταλήγουμε στη σχέση:

$$\underset{\rightarrow}{\beta} - \underset{\rightarrow}{\beta}_{(j)} \cong I^{-1}\underset{\rightarrow}{d}_j \quad (2.32)$$

όπου I είναι ο πίνακας παρατηρούμενης πληροφορίας και το διάνυσμα $\underset{\rightarrow}{d}_j$ δίνεται από τον τύπο:

$$\underset{\rightarrow}{d}_j = \delta_j \underset{\rightarrow}{r}^*_j - \sum_{i\in D_j} \frac{\exp(\underset{\rightarrow}{x}_j \underset{\rightarrow}{\beta}) \cdot \{\underset{\rightarrow}{x}_j - E(\underset{\rightarrow}{x} \mid R_i)\}}{\sum_{k\in R_i} \exp(\underset{\rightarrow}{x}_k \underset{\rightarrow}{\beta})} \quad (2.33)$$

όπου έχουμε ορίσει ως D_j το σύνολο των μονάδων στις οποίες συνέβη το γεγονός πριν ή ακριβώς τη χρονική στιγμή $t_{(j)}$, και $\hat{\underset{\rightarrow}{r}}_j$ είναι τα υπόλοιπα Schoenfeld όπως αυτά ορίστηκαν στη σχέση (2.23).

Ο πρώτος όρος του τύπου (2.33) είναι τα υπόλοιπα Schoenfeld μόνο για τις μονάδες στις οποίες έχει συμβεί το γεγονός (καθώς πολλαπλασιάζονται με τη δείκτρια συνάρτηση δ_j). Ο δεύτερος όρος του τύπου αφορά όλες τις μονάδες και αναπαριστά τη συμβολή όλων των συνόλων κινδύνου R_i στα οποία περιέχεται η μονάδα j. Θα πρέπει να επισημάνουμε ότι, εν γένει αυτοί οι δύο όροι έχουνε αντίθετη επίδραση ως προς τον χρόνο. Ο δεύτερος όρος αυξάνεται για μεγαλύτερα $t_{(j)}$, διότι είναι το άθροισμα ενός αυξανόμενου αριθμού όρων. Οπότε, για μικρούς χρόνους διακοπής ο πρώτος όρος κυριαρχεί ενώ για μεγαλύτερους χρόνους διακοπής ο δεύτερος όρος αποκτά μεγαλύτερη βαρύτητα. Για εκείνες τις μονάδες που είναι αποκομμένες από νωρίς στη μελέτη, ο πρώτος όρος είναι μηδενικός και ο δεύτερος αρκετά μικρός. Έτσι μπορούμε να θεωρήσουμε ότι οι συγκεκριμένες μονάδες έχουν μικρή επιρροή στο μοντέλο μας.

2.2.6 Κριτήρια επιλογής μεταβλητών στο μοντέλο του Cox

2.2.6.1 Ο συντελεστής προσδιορισμού στο μοντέλου του Cox

Όπως και στα κλασσικά μοντέλα παλινδρόμησης, έτσι και στο μοντέλο του Cox, αφού γίνει η προσαρμογή θέλουμε να εξετάσουμε πόσο ικανοποιητικά μπορεί το μοντέλο μας να περιγράψει τα δεδομένα. Για αυτό το σκοπό, στα μοντέλα παλινδρόμησης χρησιμοποιείται ο συντελεστής προσδιορισμού:

$$R^2 = 1 - \frac{SSE}{SST}$$

ο οποίος λαμβάνει τιμές από 0 έως 1 και εκφράζει το ποσοστό της διασποράς της εξαρτημένης μεταβλητής y που εξηγείται από το μοντέλο. Όσο πιο κοντά στο 1 είναι η τιμή του R^2 τόσο καλύτερο το μοντέλο. Όμως, στην περίπτωση του μοντέλου του Cox δεν μπορεί να χρησιμοποιηθεί διότι επεξεργαζόμαστε δεδομένα διάρκειας ζωής που περιέχουν πολύ συχνά αποκομμένες παρατηρήσεις. Αντί του κλασσικού R^2 έχουν προταθεί άλλα στατιστικά ελέγχου καλής προσαρμογής του μοντέλου εκ των οποίων τα πιο γνωστά είναι του McFadden (McFadden, 1974) :

$$R^2_{McF} = 1 - \left(\frac{\ln L_1}{\ln L_0} \right) \quad (2.34)$$

και των Cox και Snell (Cox & Snell, 1989):

$$R^2_{CS} = 1 - \left(\frac{L_0}{L_1} \right)^{\left(\frac{2}{n}\right)} \quad (2.35)$$

όπου n είναι το συνολικό μέγεθος του δείγματος, L_0 είναι η μεγιστοποιημένη συνάρτηση πιθανοφάνειας του μοντέλου χωρίς συμμεταβλητές και L_1 η αντίστοιχη μεγιστοποιημένη συνάρτηση πιθανοφάνειας του υπό μελέτη μοντέλου. Όπως και σε όλες τις περιπτώσεις συντελεστών προσδιορισμού, καλύτερο μοντέλο θεωρείται αυτό με την μεγαλύτερη τιμή του συντελεστή.

Σε αυτό το σημείο θα πρέπει να παρατηρήσουμε τα εξής:

Α) Στον συντελεστή προσδιορισμού του McFadden το $\ln L_0$ αναπαριστά το συνολικό άθροισμα τετραγώνων SST και το $\ln L_1$ είναι το αντίστοιχο άθροισμα τετραγώνων των σφαλμάτων SSE του κλασικού συντελεστή προσδιορισμού. Ο λόγος των πιθανοφανειών καταδεικνύει το επίπεδο της βελτίωσης που προσφέρει το πλήρες μοντέλο με όλες τις συμμεταβλητές έναντι του σταθερού χωρίς καμία συμμεταβλητή. Μια μικρή τιμή λοιπόν του λόγου του λογαριθμοποιημένων πιθανοφανειών θα σημαίνει ότι το πλήρες μοντέλο με τις συμμεταβλητές είναι πολύ καλύτερο του σταθερού μοντέλου.

Β) Ο συντελεστής προσδιορισμού των Cox και Snell έχει παρόμοια πρακτική σημασία με τον συντελεστή του McFadden με μια βασική διαφορά. Ενώ, ο συντελεστής McFadden όπως και το κλασικό R^2 παίρνουν τιμές στο διάστημα [0,1] ο συντελεστής Cox-Snell έχει μέγιστη τιμή πάντα μικρότερη του 1. Αυτό διότι, αν υποθέσουμε ότι το πλήρες μοντέλο περιγράφει τέλεια τα δεδομένα μας, τότε θα έχει πιθανοφάνεια $L_1 = 1$, οπότε ο συντελεστής R^2_{CS} θα παίρνει την τιμή: $1 - \left(L_0\right)^{\frac{2}{n}}$ που είναι προφανώς μικρότερη της μονάδας (Hosmeret al., 2008).

2.2.6.2 Τα κριτήρια AIC και BIC στο μοντέλο του Cox

Στα τυπικά μοντέλα παλινδρόμησης ξέρουμε ότι μπορούν να οριστούν τα κριτήρια AIC και BIC ώστε να έχουμε ένα μέτρο σύγκρισης μεταξύ μοντέλων με διαφορετικό πλήθος

επεξηγηματικών μεταβλητών. Για το ημι-παραμετρικό μοντέλο του Cox οι τροποποιήσεις των κριτηρίων έχουν να κάνουν με τη χρήση της μερικής συνάρτησης πιθανοφάνειας και είναι οι ακόλουθες:

Κριτήριο AIC (Xu et al., 2009):

$$AIC = -2 \cdot \sum_{j=1}^{k} \{\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_j - \ln[\sum_{i \in R_j} \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_i)]\} + 2p \quad (2.36)$$

Κριτήριο BIC (Volinsky & Raftery, 2000):

$$BIC = -2 \cdot \sum_{j=1}^{k} \{\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_j - \ln[\sum_{i \in R_j} \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_i)]\} + p \cdot \ln(\mathrm{k}) \quad (2.37)$$

όπου το πρώτο σκέλος και στις δύο κριτήρια είναι η λογαριθμοποιημένη συνάρτηση μερικής πιθανοφάνειας όπως ορίστηκε στην σχέση (2.16), p είναι το πλήθος των συμμεταβλητών στο μοντέλο και k είναι ο αριθμός των μη αποκομμένων παρατηρήσεων.
Κατά τα γνωστά, συγκρίνοντας όλα τα υποψήφια μοντέλα, προτιμότερο είναι εκείνο με την μικρότερη τιμή των κριτηρίων.

2.2.7 Έλεγχος προβλεπτικής ικανότητας του μοντέλου: καμπύλες ROC

Έχοντας καταλήξει στο βέλτιστο μοντέλο με βάση τις τεχνικές και τα κριτήρια που ήδη αναφέραμε, μια επιπλέον πολύ βασική πτυχή του μοντέλου που θα πρέπει να εξετάσουμε είναι η ικανότητα του να προβλέπει νέα δεδομένα. Είναι πιθανό, ένα προσαρμοσμένο μοντέλο να έχει άψογη ικανότητα περιγραφής του υπάρχοντος συνόλου δεδομένων που έχουμε στα χέρια μας αλλά να μην μπορεί να μας δώσει αξιόπιστες προβλέψεις για δεδομένα εκτός αυτού. Η αξιολόγηση της προβλεπτικής ικανότητας του μοντέλου μπορεί να γίνει γραφικά μέσω των καμπύλων ROC (ROC curves).

Οι καμπύλες ROC (Receiver Operating Characteristic) αποτελούν χρήσιμη τεχνική για την οργάνωση, την επιλογή και απεικόνιση ταξινομητών με βάση τη γραφική τους παράσταση που χρησιμοποιούνται ευρέως στη διαγνωστική ιατρική καθώς συμβάλουν στην εξασφάλιση της επιθυμητής ακρίβειας στις προβλέψεις.

Έστω δίτιμη τ.μ Y, η οποία δέχεται τις τιμές Y=1 («επιτυχία») και Y=0 («αποτυχία») και έστω επίσης $p = P(Y = 1)$ η εκτιμημένη πιθανότητα επιτυχίας. Τότε αν ορίσουμε μια σταθερά p_0 μπορούμε να διακρίνουμε τις περιπτώσεις:

- Αν $p > \mathrm{p}_0$ προβλέπεται Y=1
- Αν $p \le \mathrm{p}_0$ προβλέπεται Y=0

Οι συχνότερα χρησιμοποιούμενες και αναφερόμενες συνιστώσες της διαγνωστικής ποιότητας μιας δοκιμασίας, που καθορίζουν τη διακριτική της ικανότητα είναι το ποσοστό των αληθώς θετικών αποτελεσμάτων, το ποσοστό δηλαδή των θετικών ενδείξεων στον πληθυσμό των ασθενών (true positive rate, TPR) ή αλλιώς της ευαισθησίας (sensitivity) της δοκιμασίας και δίνεται από τον τύπο:

$$SE = TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \quad (2.38)$$

και το ποσοστό των αληθώς αρνητικών αποτελεσμάτων, δηλαδή το ποσοστό των αρνητικών ενδείξεων στον πληθυσμό των ασθενών (true negative rate, TNR) ή αλλιώς της ειδικότητας (specificity) της δοκιμασίας και υπολογίζεται ως εξής:

$$SPC = TNR = \frac{TN}{N} = \frac{TN}{FP + TN} \quad (2.39)$$

Τα ποσοστά αυτά, καθώς και τα συμπληρωματικά τους, δηλαδή το ποσοστό ψευδώς αρνητικών (false negative rate, FNR) και ψευδώς θετικών αποτελεσμάτων (false positive rate, FPR) ονομάζονται λειτουργικά χαρακτηριστικά της διαγνωστικής δοκιμασίας.

Με βάση τα παραπάνω ορίζεται λοιπόν ο 2X2 πίνακας συνάφειας:

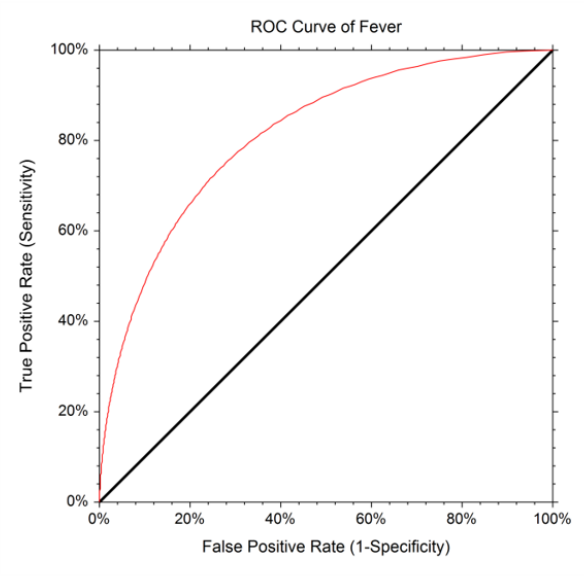
		Πραγματική κατάσταση	
		Θετικό	**Αρνητικό**
Πρόβλεψη	**Θετικό**	**TP**	**FP**
	Αρνητικό	**FN**	**TN**

Πίνακας 2.1: 2X2 πίνακας συνάφειας

Αν, λοιπόν, υπολογιστούν οι τιμές της ευαισθησίας και της ειδικότητας για κάθε p_0 στο εύρος [0,1], μπορεί να σχηματιστεί η χαρακτηριστική καμπύλη ROC η οποία απεικονίζει την προβλεπτική ικανότητα του μοντέλου καθώς το όριο p_0 μεταβάλλεται.

Στον άξονα x, ενός τέτοιου γραφήματος, έχουμε την ποσότητα (1-Specificity) και στον άξονα y έχουμε την Sensitivity. Σχεδιάζουμε την καμπύλη για τις διάφορες τιμές του p_0 και επιπλέον την ευθεία για την οποία ισχύει TPR=FPR και χωρίζει τη γραφική παράσταση σε δύο ισεμβαδικά χωρία.

Ένα μοντέλο θα έχει μεγάλη προβλεπτική ικανότητα αν για πολύ μικρές τιμές της ποσότητας (1-Specificity) το Sensitivity να πλησιάζει την τιμή 1. Ένας δείκτης που μετρά αυτή την επιθυμητή ιδιότητα ενός μοντέλου είναι το εμβαδόν κάτω από την καμπύλη (area under the curve, AUC). Όσο το AUC προσεγγίζει το 1 τόσο καλύτερη προβλεπτική ικανότητα έχει το μοντέλο μας. Θα πρέπει να τονίσουμε ότι η καμπύλη δεν μπορεί να βρεθεί ποτέ κάτω από την ευθεία TPR=FPR που διχοτομεί το χωρίο, άρα το AUC λαμβάνει τιμές από 0.5 έως και 1 (εφόσον το εμβαδόν κάτω από την ευθεία είναι πάντα 0.5).



Γράφημα 2.1: Παράδειγμα καμπύλης ROC

Όσα παρουσιάστηκαν παραπάνω αφορούν περιπτώσεις μοντέλων όπου η επεξηγηματική μεταβλητή Y είναι δίτιμη με τιμές 0 και 1. Για δεδομένα διάρκειας ζωής όμως θα πρέπει να γίνει μια κατάλληλη γενίκευση των ορισμών της ευαισθησίας (sensitivity) και της ειδικότητας (specificity). Αντί, λοιπόν, της απλής δίτιμου αποτελέσματος, ένας χρόνος επιβίωσης μπορεί να θεωρηθεί ως ένα χρονικά μεταβαλλόμενο δίτιμο αποτέλεσμα. Έτσι, λοιπόν μπορούμε να καταλήξουμε σε χρονο-εξαρτώμενες καμπύλες ROC από τις οποίες μπορούμε να εξάγουμε ανάλογα αποτελέσματα για την προβλεπτική ικανότητα του μοντέλου σε σχέση με την απλή περίπτωση δίτιμης μεταβλητής (Heagerty & Zheng, 2005).

2.2.8 Επεκτάσεις του μοντέλου του Cox

Μερικές φορές, αφού προσαρμόσουμε το ημι-παραμετρικό μοντέλο του Cox στα δεδομένα μας αντιλαμβανόμαστε ότι η πρωταρχικής σημασίας υπόθεση αναλογικής διακινδύνευσης δεν ισχύει καθολικά στο μοντέλο μας ή τουλάχιστον για κάποια συμμεταβλητή. Για να αντιμετωπίσουμε αυτό το πρόβλημα μπορούμε να γενικεύσουμε το κλασικό μοντέλο του Cox. Οι δύο πιο συνηθισμένοι τρόποι επίλυσης του προβλήματος είναι: α) η στρωματοποιημένη εκδοχή του μοντέλου του Cox (stratified Cox model) και β) η εισαγωγή συμμεταβλητών εξαρτημένων από το χρόνο.

2.2.8.1 Στρωματοποιημένο μοντέλο του Cox

Αυτή η επέκταση του μοντέλου του Cox χρησιμοποιείται όταν εικάζουμε ότι οι συναρτήσεις διακινδύνευσης μεταξύ δύο ή περισσοτέρων κατηγοριών μιας συμμεταβλητής δεν βρίσκονται σε αναλογία μεταξύ τους. Για αυτό το λόγο, χωρίζουμε τη κατηγορική μεταβλητή σε «στρώματα» (strata) ώστε κάθε στρώμα να έχει τη δική του βασική συνάρτηση διακινδύνευσης $h_0(t)$. Άρα παραδείγματος χάρη, αν έχουμε στην μελέτη μας την συμμεταβλητή που εκφράζει τα επίπεδα αιμοπεταλίων (χαμηλά, φυσιολογικά, υψηλά) και πως αυτή επηρεάζει τη διάρκεια ζωής ασθενών που πάσχουν από απλαστική αναιμία, τότε αν αντιληφθούμε ότι οι κατηγορίες δεν βρίσκονται σε αναλογία μεταξύ τους, η συνάρτηση διακινδύνευσης μέσω της στρωματοποιημένης ανάλυσης μπορεί να γραφεί ως:

$$h(t;\underset{\rightarrow}{x}) = \begin{cases} \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x})\, \mathrm{h}_{01}(t), \alpha\nu\ \upsilon\psi\eta\lambda\acute{o}\ \epsilon\pi\acute{\imath}\pi\epsilon\delta o \\ \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x})\, \mathrm{h}_{02}(t), \alpha\nu\ \varphi\upsilon\sigma\iota o\lambda o\gamma\iota\kappa\acute{o}\ \epsilon\pi\acute{\imath}\pi\epsilon\delta o \\ \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x})\, \mathrm{h}_{03}(t), \alpha\nu\ \chi\alpha\mu\eta\lambda\acute{o}\ \epsilon\pi\acute{\imath}\pi\epsilon\delta o \end{cases}$$

όπου $\mathrm{h}_{01}(t)$, $\mathrm{h}_{02}(t)$ και $\mathrm{h}_{03}(t)$ οι βασικές συναρτήσεις διακινδύνευσης των ασθενών με υψηλά, φυσιολογικά και χαμηλά επίπεδα αιμοπεταλίων αντίστοιχα. Επιπλέον, θεωρούμε ότι η ιδιότητα της αναλογικής διακινδύνευσης εξακολουθεί να ισχύει για τις υπόλοιπες συμμεταβλητές του μοντέλου μας. Είναι, επίσης, εμφανές ότι οι συντελεστές $\underset{\rightarrow}{\beta}$ των υπόλοιπων συμμεταβλητών είναι κοινοί και για τα 3 στρώματα της συμμεταβλητής που εκφράζει τα επίπεδα αιμοπεταλίων.

Για να εκτιμηθούν οι παράμετροι στο στρωματοποιημένο μοντέλο του Cox θα χρησιμοποιήσουμε παρόμοια τεχνική με την παράγραφο 2.2.1 με μικρές τροποποιήσεις. Πιο συγκεκριμένα, για κάθε στρώμα m=1,..,s, ο λογάριθμος της μερικής πιθανοφάνειας δίνεται από τον τύπο (Καρώνη, 2009):

$$l_m(\underset{\rightarrow}{\beta}) = \sum_{j=1}^{k_m} \underset{\rightarrow}{\beta}^{\mathrm{T}} \underset{\rightarrow}{x}_{mj} - \sum_{j=1}^{k_m} \ln\left\{ \sum_{i \in R_{mj}} \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \underset{\rightarrow}{x}_{mi}) \right\} \quad (2.40)$$

όπου έχει προστεθεί σε σχέση με την κλασική σχέση (2.16) της μερικής πιθανοφάνειας, ο δείκτης m στα σύμβολα $\underset{\rightarrow}{x}_j$, R_j και k ώστε να υποδηλώσουμε ότι μόνο τα δεδομένα του m-οστού στρώματος συμμετέχουν στην συνάρτηση. Συνολικά λοιπόν, για όλα τα στρώματα θα έχουμε:

$$l(\underset{\rightarrow}{\beta}) = \sum_{m=1}^{s} l_m(\underset{\rightarrow}{\beta}) = \sum_{m=1}^{s}\sum_{j=1}^{k_m} \underset{\rightarrow}{\beta}^{\mathrm{T}} \underset{\rightarrow}{x}_{mj} - \sum_{m=1}^{s}\sum_{j=1}^{k_m} \ln\left\{ \sum_{i \in R_{mj}} \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \underset{\rightarrow}{x}_{mi}) \right\} \quad (2.41)$$

Στη συνέχεια υπολογίζουμε τις εκτιμήτριες β_j των παραμέτρων του μοντέλου με παρόμοιες τεχνικές όπως και πριν.

Αυτή η τεχνική στρωματοποίησης, μπορεί να χρησιμοποιηθεί ως ένας εναλλακτικός τρόπος ελέγχου της υπόθεσης της αναλογικής διακινδύνευσης.

2.2.8.2 Συμμεταβλητές εξαρτημένες από το χρόνο

Σε μια μελέτη ανάλυσης επιβίωσης, είναι πιθανό να κληθούμε να διαχειριστούμε μεταβλητές οι οποίες έχουν εξάρτηση από το χρόνο. Τέτοιες μεταβλητές μπορεί να είναι η ηλικία ενός ατόμου, η δόση ενός φαρμάκου ή η θερμοκρασία υπό την οποία μελετώνται τα εξαρτήματα σε ένα πείραμα. Αυτού του είδους οι μεταβλητές καλούνται «εξωτερικές» (external) καθώς ενδέχεται να αλλάξουν με το πέρασμα του χρόνου αλλά αυτό θα γίνει είτε με απόλυτα αναμενόμενο τρόπο (ηλικία ατόμου) είτε από γνωστό εξωτερικό παράγοντα (δόση φαρμάκου). Μια άλλη κατηγορία μεταβλητών, είναι οι λεγόμενες «εσωτερικές» (internal) μεταβλητές οι οποίες μεταβάλλονται ανάλογα με την κατάσταση της ίδιας της μονάδας μέσα στο χρόνο. Τέτοιες μεταβλητές σε βιοϊατρικές μελέτες, μπορεί να είναι κλινικές μετρήσεις (πχ αριθμός λευκών αιμοσφαιρίων) ή οτιδήποτε σχετίζεται με την κατάσταση της υγείας του ασθενή (πχ μέγεθος κακοήθους όγκου).

Αυτές οι μεταβλητές προφανώς δεν μπορούν να πληρούν την ιδιότητα της αναλογικότητας της διακινδύνευσης καθώς είναι χρονοεξαρτώμενες, αλλά παρόλα μπορούν να συμπεριληφθούν στο μοντέλο του Cox ως επέκταση της κλασικής του μορφής. Για αυτό το σκοπό, η συνάρτηση μερικής πιθανοφάνειας (2.16) τροποποιείται ως εξής:

$$l(\underset{\rightarrow}{\beta}) = \sum_{j=1}^{k} \underset{\rightarrow}{\beta}^{\mathrm{T}} \underset{\rightarrow}{x}_j(t_{(j)}) - \sum_{j=1}^{k} \ln\left\{ \sum_{i \in R_j} \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \underset{\rightarrow}{x}_i(t_{(j)}) \right\} \quad (2.42)$$

όπου έχουμε αντικαταστήσει από τον κλασικό τύπο το $\underset{\rightarrow}{x}_j$ με το $\underset{\rightarrow}{x}_j(t_{(j)})$. Βασική προϋπόθεση λοιπόν είναι για κάθε χρόνο διακοπής $t_{(j)}$ να γνωρίζουμε τις τιμές που λαμβάνουν όλες οι συμμεταβλητές του μοντέλου. Αυτό εν γένει δεν είναι εύκολο, διότι συνεπάγεται διαρκή παρακολούθηση των μονάδων. Σε πολλές ιατρικές μελέτες, γίνεται παρακολούθηση των ατόμων ανά τακτά χρονικά διαστήματα (εβδομαδιαίως, μηνιαίως) και όσες τιμές $\underset{\rightarrow}{x}_j(t_{(j)})$ δεν είναι διαθέσιμες, είτε χρησιμοποιούνται οι τελευταίες μετρήσεις πριν συμβεί το γεγονός είτε χρησιμοποιείται η μέθοδος παρεμβολής μεταξύ δύο μετρήσεων.

3. Οι μέθοδοι ποινής

3.1 Τα φαινόμενα υπερπροσαρμογής και πολυσυγγραμμικότητας

Πολύ συχνά βρισκόμαστε αντιμέτωποι με το πρόβλημα προσαρμογής ενός μοντέλου παλινδρόμησης που περιέχει πάρα πολλές επεξηγηματικές μεταβλητές. Το προσαρμοσμένο μοντέλο μπορεί να ταιριάζει πολύ καλά στα δεδομένα μας αλλά να μην μπορεί να προσαρμοστεί σε νέα δεδομένα ή να έχει πολύ μικρή προβλεπτική ικανότητα από ένα μοντέλο που προσαρμόζεται «χειρότερα» στα δεδομένα μας. Αυτό είναι το λεγόμενο πρόβλημα υπερπροσαρμογής (overfitting). Η υπερπροσαρμογή έχει ως αποτέλεσμα μοντέλα που είναι πιο περίπλοκα από όσο χρειάζεται.

Για την αντιμετώπιση αυτού του προβλήματος υπάρχουν δυο λύσεις. Η πρώτη λύση είναι να μειώσουμε τον αριθμό των χαρακτηριστικών, είτε επιλέγοντας χειρωνακτικά τα χαρακτηριστικά που θα χρησιμοποιήσουμε είτε χρησιμοποιώντας κάποιον αλγόριθμο επιλογής. Η δεύτερη λύση είναι να κάνουμε κανονικοποίηση (regularization) του μοντέλου. Δηλαδή, κρατάμε όλα τα χαρακτηριστικά, αλλά μειώνουμε την αντίστοιχη παράμετρο β_i, δηλαδή τη βαρύτητα που έχει το αντίστοιχο χαρακτηριστικό κατά δημιουργία του μοντέλου. Η κανονικοποίηση δίνει καλά αποτελέσματα, όταν καθένα από τα πολλά χαρακτηριστικά συνεισφέρει από λίγο.

Ένα άλλο γνωστό πρόβλημα στην πολλαπλή παλινδρόμηση είναι το πρόβλημα της πολυσυγγραμμικότητας (multicollinearity). Πολυσυγγραμμικότητα υπάρχει όταν δύο ή περισσότερες ανεξάρτητες μεταβλητές είναι ισχυρά συσχετισμένες μεταξύ τους, και οι τιμές της μιας μπορούν να υπολογιστούν από την άλλη. Η πολυσυγγραμμικότητα δεν έχει επιπτώσεις στην προβλεπτική ικανότητα του μοντέλου, έχει όμως επιπτώσεις στους συντελεστές των ανεξάρτητων μεταβλητών καθώς τα αποτελέσματα που εξάγουμε μπορεί να μην είναι αξιόπιστα. Οι τιμές των συντελεστών μπορεί να αλλάξουν πολύ, αν προστεθεί ή αφαιρεθεί μια μεταβλητή ή αν συμβούν μικρές μεταβολές στα δεδομένα. Επιπλέον, η παρουσία αυτού του φαινομένου οδηγεί σε αυξημένα τυπικά σφάλματα $\mathrm{se}(\beta_i)$ των συντελεστών του μοντέλου και κατά συνέπεια δυσκολεύει την εκτίμηση της επίδρασης κάθε επεξηγηματικής μεταβλητής. Δηλαδή, γίνεται πιο δύσκολος ο καθορισμός των στατιστικά σημαντικών μεταβλητών.

Μια πολύ λογική προσέγγιση για την αντιμετώπιση των παραπάνω προβλημάτων θα ήτανε η επανασυλλογή περισσότερων δεδομένων ώστε να γίνει εκ νέου η μοντελοποίηση με τη χρήση, ίσως, διαφορετικών εκτιμητικών μεθόδων για τον καθορισμό των παραμέτρων του μοντέλου. Κάτι τέτοιο όμως μπορεί να είναι χρονοβόρο και κοστοβόρο.

3.2 Ορισμός της ποινής

Έχοντας αναφέρει όλα τα παραπάνω μπορούμε σε αυτό το σημείο να εισάγουμε την έννοια της «ποινής» για τους συντελεστές ενός μοντέλου παλινδρόμησης. Μια ποινή είναι ουσιαστικά ένας επιπλέον περιορισμός από τον οποίο πρέπει να διέπονται οι παράμετροι β_i του μοντέλου μας.

Σε σχέση με τα κλασικά μοντέλα παλινδρόμησης μια από τις πρώτες τεχνικές που αναπτύχθηκαν ήταν αυτή της παλινδρόμησης κορυφογραμμής (ridge regression) (Hoerl & Kennard, 1970) η οποία για την αντιμετώπιση του προβλήματος της πολυσυγγραμμικότητας επέβαλε μια L2 τύπου συνθήκη στην εκτίμηση των συντελεστών του μοντέλου. Όταν αναφερόμαστε σε L2 (και αντίστοιχα L1) τύπο συνθήκης αυτό έχει να κάνει με την χρήση της

διανυσματικής νόρμας που χρησιμοποιείται στον ορισμό της συνθήκης. Θυμίζουμε ότι αν έχουμε ένα διάνυσμα $\underset{\sim}{x} = (x_1, ..., x_n)^T$ τότε η L2 νόρμα του x ορίζεται ως εξής: $\|x\|_2 = (|x_1|^2 + ... + |x_n|^2)^{1/2}$ ενώ η L1 νόρμα του ίδιου διανύσματος θα ήτανε: $\|x\|_1 = (|x_1| + ... + |x_n|)$.

Μια παρόμοια τεχνική της παλινδρόμησης κορυφογραμμής είναι η τεχνική Lasso (Least Absolute Shrinkage and Selection Operator) (Tibshirani, 1996). Η τεχνική Lasso λειτουργεί με παρόμοιο τρόπο, με τη διαφορά ότι επιβάλλεται στην εκτίμηση των παραμέτρων του μοντέλου μια L1 τύπου ποινή, δηλαδή χρησιμοποιείται η L1 διανυσματική νόρμα. Επιπλέον, ο Tibshirani χρησιμοποίησε μια παραλλαγή αυτής της μεθόδου σε μοντέλα δεδομένων διάρκειας ζωής και πιο συγκεκριμένα στο μοντέλο αναλογικής διακινδύνευσης του Cox (Tibshirani, 1997) με την οποία θα ασχοληθούμε εκτενέστερα στις επόμενες παραγράφους.

Παρότι και οι δύο παραπάνω μέθοδοι είναι μέθοδοι «συρρίκνωσης» των εκτιμητριών των παραμέτρων του μοντέλου παλινδρόμησης (shrinkage methods) για την αντιμετώπιση του προβλήματος της υπερπαραμετροποίησης ή της πολυσυγγραμμικότητας, τα αποτελέσματα των L1 και L2 ποινών είναι αρκετά διαφορετικά στην πράξη. Εφαρμόζοντας L2 ποινή καταλήγουμε σε μικρές αλλά μη μηδενικές εκτιμήτριες παραμέτρων, ενώ αντίθετα η χρήση μιας L1 ποινής μας επιστρέφει πολλές παραμέτρους που έχουνε συρρικνωθεί ακριβώς στο μηδέν και κάποιες λιγότερες που έχουν υποστεί μικρή συρρίκνωση.

Τέλος, πρέπει να αναφέρουμε ότι τα τελευταία χρόνια έχει επίσης προταθεί η μέθοδος Elastic net (Zou & Hastie, 2005) η οποία ουσιαστικά συνδυάζει τις τεχνικές Ridge και Lasso εφαρμόζοντας έναν κυρτό συνδυασμό L1 και L2 περιορισμών για την εκτίμηση των παραμέτρων του μοντέλου. Με τη μέθοδο Elastic net καταλήγουμε σε ένα μικτό αποτέλεσμα, δηλαδή λιγότερες παράμετροι καταλήγουνε να είναι μηδενικές σε σχέση με μια απλή L1 περιορισμό αλλά η συρρίκνωση των μη μηδενικών παραμέτρων είναι μεγαλύτερη συγκριτικά με μια αμιγώς L2 συνθήκη.

3.3 Τύποι μεθόδων ποινών

Στα κλασικά μοντέλα πολλαπλής παλινδρόμησης για την εκτίμηση των παραμέτρων $\beta_j, j = 1, .., k$ χρησιμοποιούμε τη μέθοδο ελαχίστων τετραγώνων επιδιώκοντας να ελαχιστοποιήσουμε τα τυχαία σφάλματα ε_i. Θέλουμε δηλαδή να βρούμε το ελάχιστο της συνάρτησης:

$$S(\underset{\sim}{\varepsilon}) = \|\underset{\sim}{\varepsilon}\|_2^2 = \|\underset{\rightarrow}{y} - X\underset{\rightarrow}{\beta}\|_2^2 = (\underset{\rightarrow}{y} - X\underset{\rightarrow}{\beta})^{\mathrm{T}} \cdot (\underset{\rightarrow}{y} - X\underset{\rightarrow}{\beta}) = S(\underset{\rightarrow}{\beta}) \quad (3.1).$$

Από την σχέση (3.1) παίρνοντας την μερική παράγωγο ως προς $\underset{\rightarrow}{\beta}$ και εξισώνοντας με το μηδέν καταλήγουμε στη σχέση:

$$\mathrm{X}^{\mathrm{T}}\underset{\rightarrow}{y} = X^T X\underset{\rightarrow}{\beta} \quad (3.2)$$

Άρα αν ο πίνακας $X^T X$ είναι αντιστρέψιμος τότε προκύπτει η κλασική εκτιμήτρια ελαχίστων τετραγώνων:

$$\underset{\rightarrow}{\beta}^{OLS} = (X^T X)^{-1} \cdot \mathrm{X}^{\mathrm{T}}\underset{\rightarrow}{y} \quad (3.3)$$

Όμως πολλές φορές, λόγω υψηλής συσχέτισης μεταξύ των επεξηγηματικών μεταβλητών του μοντέλου είμαστε αναγκασμένοι να εφαρμόσουμε μια επιπλέον ποινή για την εκτίμηση των παραμέτρων $\beta_j, j=1,..,k$.Παρακάτω παρουσιάζονται οι 3 πιο γνωστές τεχνικές μέθοδοι συρρίκνωσης (shrinkage methods): Ridge, Lasso και (naïve) Elastic Net.

3.3.1 Η μέθοδος Ridge

Η πρώτη τεχνική συρρίκνωσης, όπως αναφέραμε και πρωτύτερα, είναι η μέθοδος Ridge. Όταν παρουσιάζεται το πρόβλημα της πολυσυγγραμμικότητας η κλασική μέθοδος ελαχίστων τετραγώνων (3.1) αδυνατεί να εκτιμήσει ικανοποιητικά την επίδραση των επεξηγηματικών μεταβλητών στην μεταβλητή απόκρισης. Αυτό συμβαίνει διότι η υψηλή συσχέτιση μεταξύ των μεταβλητών μας επιστρέφει εκτιμήτριες β_i με μεγάλα τυπικά σφάλματα $se(\beta_i)$. Όπως ξέρουμε από το θεώρημα Gauss-Markov, αν ισχύουν οι τυπικές υποθέσεις του πολλαπλού γραμμικού μοντέλου τότε η εκτιμήτρια $\underrightarrow{\beta}$ αποτελεί την καλύτερη γραμμική αμερόληπτη εκτιμήτρια του διανύσματος παραμέτρων $\underrightarrow{\beta}$ (Best Linear Unbiased Estimators, BLUE). Ο όρος αμερόληπτη σημαίνει ότι ισχύει: $\mathrm{E}(\underrightarrow{\beta}) = \underrightarrow{\beta}$ ενώ ο όρος καλύτερη έχει να κάνει με το ότι είναι η εκτιμήτρια ελάχιστης διασποράς δεδομένης της αμεροληψίας. Αυτή η δέσμευση της αμεροληψίας κάνει στην περίπτωσή μας τις εκτιμήσεις β_i να έχουνε μεγάλα τυπικά σφάλματα. Με τη μέθοδο Ridge θα καταλήξουμε σε εκτιμήτριες μη αμερόληπτες αλλά που θα έχουνε πολύ μικρότερη διασπορά από τους κλασικούς.

Στη μέθοδο Ridge θέλουμε να λύσουμε το πρόβλημα ελαχιστοποίησης:

$$S(\underrightarrow{\beta}) = \left\| \underrightarrow{y} - X\,\underrightarrow{\beta} \right\|_2^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \text{ δεδομένου: } \left\| \underrightarrow{\beta} \right\|_2^2 \le c^2 \Rightarrow \sum_{j=1}^{p} \left| \beta_j \right|^2 \le c^2 \quad (3.4)$$

όπου c είναι ένας αυθαίρετος πραγματικός αριθμός. Χρησιμοποιώντας σε αυτό το σημείο τη μέθοδο των πολλαπλασιαστών Lagrange, καταλήγουμε στο ισοδύναμο πρόβλημα ελαχιστοποίησης της συνάρτησης:

$$F(\underrightarrow{\beta}, \lambda) = \left(\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \cdot \beta_j)^2 + \lambda \cdot \sum_{j=1}^{p} \beta_j^2 \right) \quad (3.5)$$

όπου το $\lambda \ge 0$ είναι μια σταθερά, η οποία επιλέγεται κάθε φορά και ονομάζεται παράμετρος μεροληψίας (biasing parameter). Το λ ελέγχει το μέγεθος τη συρρίκνωσης: όσο μεγαλύτερη η τιμή του λ, τόσο μεγαλύτερη η συρρίκνωση.

Επιπλέον, πρέπει να παρατηρήσουμε ότι στη σχέση (3.5) ο σταθερός όρος του μοντέλου β_0 έχει μείνει εκτός του όρου ποινής. Μπορεί να αποδειχτεί ότι η λύση της (3.5) μπορεί να χωριστεί σε δύο μέρη, έπειτα από αναπαραμετροποίηση χρησιμοποιώντας αντί των x_{ij} τα κεντραρισμένα $(x_{ij} - \bar{x}_j)$. Έτσι εκτιμούμε το β_0 από το $\bar{y} = \sum_{i=1}^{n} y_i \Big/ n$. Οπότε υποθέτουμε ότι αυτό το κεντράρισμα έχει γίνει, και άρα ο πίνακας πληροφορίας X διαθέτει p (αντί για p+1) στήλες. Η σχέση (3.5) σε μορφή πινάκων γράφεται ως εξής:

$$RSS(\lambda) = (\underset{\rightarrow}{y} - X\underset{\rightarrow}{\beta})^{\mathrm{T}} \cdot (\underset{\rightarrow}{y} - X\underset{\rightarrow}{\beta}) + \lambda \underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{\beta} \quad (3.6)$$

Λύνοντας την (3.6) καταλήγουμε στην εκτιμήτρια της μεθόδου Ridge:

$$\underset{\rightarrow}{\beta}^{ridge} = (X^T X + \lambda \mathrm{I})^{-1} \mathrm{X}^{\mathrm{T}} \underset{\rightarrow}{y} \quad (3.7)$$

όπου I είναι ο pxp μοναδιαίος πίνακας. Προφανώς όταν το λ=0, η παραπάνω εκτιμήτρια ταυτίζεται με την κλασική εκτιμήτρια ελαχίστων τετραγώνων. Πρέπει να παρατηρήσουμε ότι, λόγω της επιλογής μιας τετραγωνικής ποινής $\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{\beta}$, η μέθοδος ridge καταλήγει σε μια λύση που είναι και πάλι γραμμική συνάρτηση του y. Σε σχέση με την κλασική εκτιμήτρια ελαχίστων τετραγώνων (3.3) στη λύση προστίθεται μια θετική σταθερά στα διαγώνια στοιχεία του πίνακα $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ πριν αυτός αντιστραφεί. Αυτό αυτομάτως κάνει το πρόβλημα επιλύσιμο διότι ακόμα και αν ο $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ δεν είναι μέγιστης τάξης, ο $(X^T X + \lambda \mathrm{I})$ θα είναι πάντα αντιστρέψιμος. Αυτό είναι και το βασικό πλεονέκτημα της μεθόδου ridge.

Μπορούμε να αποδείξουμε εύκολα ότι η εκτιμήτρια ridge δεν είναι αμερόληπτη, δηλαδή $\mathrm{E}(\underset{\rightarrow}{\beta}^{ridge}) \neq \underset{\rightarrow}{\beta}$ αλλά σχετίζεται γραμμικά με την εκτιμήτρια ελαχίστων τετραγώνων διότι:

$$\underset{\rightarrow}{\beta}^{ridge} = (X^T X + \lambda \mathrm{I})^{-1} \mathrm{X}^{\mathrm{T}} \underset{\rightarrow}{y} = (X^T X + \lambda \mathrm{I})^{-1} \cdot (X^T X) \cdot \underset{\rightarrow}{\beta}^{OLS} = W_\lambda \cdot \underset{\rightarrow}{\beta}^{OLS} \quad (3.8)$$

όπου $W_\lambda = (X^T X + \lambda \mathrm{I})^{-1} \cdot (X^T X)$

Επιπλέον, μπορεί να υπολογιστεί ο πίνακας συνδιασποράς της εκτιμήτριας ridge, ως εξής:

$$V(\underset{\rightarrow}{\beta}^{ridge}) = V(W_\lambda \underset{\rightarrow}{\beta}^{OLS}) = W_\lambda \cdot V(\underset{\rightarrow}{\beta}^{OLS}) \cdot W_\lambda^{\ T} = \sigma^2 W_\lambda \cdot (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1} \cdot W_\lambda^{\ T} \quad (3.9)$$

Στη σχέση (3.9) καταλήξαμε χρησιμοποιώντας την ιδιότητα της εκτιμήτριας ελαχίστων τετραγώνων: $\underset{\rightarrow}{\beta}^{OLS} \sim \mathrm{N}_p(\underset{\rightarrow}{\beta}, \sigma^2 (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1})$

Για να κάνουμε μια σύγκριση με την αμερόληπτη εκτιμήτρια ελαχίστων τετραγώνων θα θεωρήσουμε την ειδική περίπτωση όπου ο πίνακας είναι ορθοκανονικός, δηλαδή: $\mathrm{X}^{\mathrm{T}}\mathrm{X} = \mathrm{I}_p$

Άρα λοιπόν, ξέρουμε ότι $V(\underset{\rightarrow}{\beta}^{OLS}) = \sigma^2 (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1} = \sigma^2 \mathrm{I}_p$ ενώ η διασπορά της εκτιμήτριας ridge γίνεται:

$$V(\underset{\rightarrow}{\beta}^{ridge}) = \sigma^2 W_\lambda (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1} W_\lambda^{\mathrm{T}} = \sigma^2 (\mathrm{I}_p + \lambda \mathrm{I}_p)^{-1} \mathrm{I}_p [(\mathrm{I}_p + \lambda \mathrm{I}_p)^{-1}]^{\mathrm{T}} = \sigma^2 (1+\lambda)^{-2} \mathrm{I}_p$$

Άρα όπως αναμενόταν: $V(\underset{\rightarrow}{\beta}^{ridge}) \leq V(\underset{\rightarrow}{\beta}^{OLS})$

Γνωρίζουμε ότι, γενικά για μια οποιαδήποτε εκτιμήτρια T(X) μιας παραμέτρου θ (ισχύει ανάλογα και για διάνυσμα παραμέτρων) το μέσο τετραγωνικό σφάλμα MSE (mean square error) ορίζεται ως εξής:

$$MSE(T(\mathrm{X})) = \mathrm{E}[(\mathrm{T}(\mathrm{X}) - \theta)^2] = B^2(T(X)) + Var(T(X))$$

όπου B(T(X)) είναι το μέτρο της μεροληψίας της εκτιμήτριας: $\mathrm{B}(\mathrm{T}(\mathrm{X})) = \mathrm{E}[\mathrm{T}(\mathrm{X}) - \theta]$

Άρα, στην περίπτωσή μας η μεροληψία της εκτιμήτριας ridge θα είναι:

$$\mathrm{B}^2(\underset{\rightarrow}{\beta}^{ridge}) = \underset{\rightarrow}{\beta}^{\mathrm{T}}(\mathrm{W}_\lambda - \mathrm{I})^{\mathrm{T}}(\mathrm{W}_\lambda - \mathrm{I})\underset{\rightarrow}{\beta} \quad (3.10)$$

ενώ η συνολική διασπορά της εκτιμήτριας θα είναι το ίχνος του πίνακα (3.9), δηλαδή το άθροισμα των διαγώνιων στοιχείων του πίνακα συνδιασποράς, άρα:

$$Var(\underset{\rightarrow}{\beta}^{ridge}) = \sigma^2 tr\{\mathrm{W}_\lambda(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}W_\lambda^{\mathrm{T}}\} \quad (3.11)$$

Επομένως, το μέσο τετραγωνικό σφάλμα της εκτιμήτριας ridge συναρτήσει του λ θα είναι:

$$\mathrm{MSE}(\lambda) = \sigma^2 tr\{\mathrm{W}_\lambda(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}W_\lambda^{\mathrm{T}}\} + \underset{\rightarrow}{\beta}^{\mathrm{T}}(\mathrm{W}_\lambda - \mathrm{I})^{\mathrm{T}}(\mathrm{W}_\lambda - \mathrm{I})\underset{\rightarrow}{\beta} \quad (3.12)$$

Για μικρές τιμές του λ, ο πρώτος όρος της διασποράς κυριαρχεί στο MSE, ενώ για μικρές τιμές του λ, ο δεύτερος όρος της μεροληψίας κυριαρχεί στο MSE. Στόχος της μεθόδου ridge είναι η εύρεση εκείνης της τιμής της σταθεράς λ ώστε η μείωση του όρου της διασποράς να είναι μεγαλύτερη από την αύξηση του όρου της μεροληψίας. Κάτι τέτοιο επιτυγχάνεται για εκείνα τα λ όπου ισχύει:

$$\mathrm{MSE}(\underset{\rightarrow}{\beta}^{ridge}) < V(\underset{\rightarrow}{\beta}^{OLS}) \quad (3.13)$$

3.3.2 Η μέθοδος Lasso

Άλλη μια σημαντική μέθοδος συρρίκνωσης, εκτός της ridge, είναι η Lasso (Least Absolute Shrinkage and Selection Operator) η οποία πρωτοπαρουσιάστηκε από τον Tibshirani το 1996. Όπως και η μέθοδος ridge εφαρμόζεται στα κλασικά μοντέλα παλινδρόμησης αλλά έχει επεκταθεί η θεωρία της και για δεδομένα διάρκειας ζωής και στη συγκεκριμένα στο μοντέλο αναλογικής διακινδύνευσης του Cox.

Το χαρακτηριστικό της μεθόδου Lasso, που την κάνει ιδιαίτερα χρήσιμη και την ξεχωρίζει από την μέθοδο ridge, είναι ότι συρρικνώνει κάποιους από τους συντελεστές του μοντέλου και όλους τους υπόλοιπους τους μηδενίζει. Έτσι λοιπόν, μπορεί να χρησιμοποιηθεί και ως κριτήριο επιλογής μοντέλου καθώς ταυτόχρονα γίνεται συρρίκνωση του μοντέλου και μηδενισμός των στατιστικά μη σημαντικών μεταβλητών, εξ ου και η ονομασία της.

Ομοίως με τη μέθοδο ridge, θέλουμε με τη μέθοδο lasso να ελαχιστοποιήσουμε τη συνάρτηση:

$$S(\underset{\rightarrow}{\beta}) = \left\| \underset{\rightarrow}{y} - X\,\underset{\rightarrow}{\beta} \right\|_2^2 = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \text{ δεδομένου} \left\| \underset{\rightarrow}{\beta} \right\|_1 \le c \Rightarrow \sum_{j=1}^{p}\left|\beta_j\right| \le c \quad (3.14)$$

όπου c μια αυθαίρετη θετική σταθερά.
Η διαφορά με τη μέθοδο ridge είναι η L1 τύπου συνθήκη που επιβάλλουμε στο διάνυσμα παραμέτρων. Όπως και πριν, μπορούμε με αναπαραμετροποίηση να αφήσουμε εκτός τον σταθερό όρο του μοντέλου β_0 ο οποίος και πάλι εκτιμάται από το $\bar{y}$. Λόγω της μορφής του περιορισμού του προβλήματος (3.14), αν επιλέξουμε τη σταθερά c να είναι «αρκετά μικρή» αυτό θα αναγκάσει πολλά από τα εκτιμημένα β_j να είναι ακριβώς μηδέν. Αν το c επιλεγεί μικρότερο του $c_0 = \sum_{j=1}^{p}\left|\beta_j^{OLS}\right|$,όπου β_j^{OLS} είναι οι κλασικές εκτιμήτριες ελαχίστων

τετραγώνων, τότε οι εκτιμήτριες lasso είναι ακριβώς οι εκτιμήτριες ελαχίστων τετραγώνων. Αν όμως παραδείγματος χάριν, επιλεγεί ως $c = c_0/2$ τότε οι εκτιμήτριες ελαχίστων τετραγώνων συρρικνώνονται κατά 50% κατά μέσο όρο.

Για να κατανοήσουμε καλύτερα τα αποτελέσματα της μεθόδου lasso θα θεωρήσουμε ξανά την ειδική περίπτωση όπου ο nxp πίνακας σχεδιασμού X είναι ορθοκανονικός, δηλαδή ισχύει η σχέση: $\mathrm{X^T X = X X^T = I}$ (όπου I ο ταυτοτικός πίνακας). Τότε λοιπόν η λύση του προβλήματος (3.14) μπορεί να αποδειχτεί ότι είναι:

$$\beta_j^{lasso} = sign\left(\beta_j^{OLS}\right)\left(\left|\beta_j^{OLS}\right| - \gamma\right)^+ \quad (3.15)$$

όπου το γ καθορίζεται από τη συνθήκη: $\sum_{j=1}^{p}\left|\beta_j\right| \le c$

Στην ορθοκανονική περίπτωση σχεδιασμού που θεωρήσαμε, η μορφή της λύσης (3.15) μας υποδεικνύει ότι η μέθοδος επιλογής των k-το πλήθος καλύτερων μεταβλητών γίνεται διαλέγοντας απλώς τις k μεγαλύτερες κατά απόλυτη τιμή παραμέτρους που εκτιμήθηκαν από την κλασική μέθοδο ελαχίστων τετραγώνων και θέτοντας τις υπόλοιπες ίσες με μηδέν. Για κάποια επιλογή της σταθεράς λ, αυτό είναι ισοδύναμο με το να θέσουμε $\beta_j^{lasso} = \beta_j^{OLS}$ αν $\left|\beta_j^{OLS}\right| > \lambda$ και 0 διαφορετικά.

Έχουμε βρει στη σχέση (3.8) την εκτιμήτρια της μεθόδου ridge. Άρα κάνοντας σύγκριση με την εκτιμήτρια lasso στην ορθοκανονική περίπτωση όπου ισχύει $\mathrm{X^T X = X X^T = I}$ θα έχουμε

$$\underrightarrow{\beta}^{ridge} = (X^T X + \lambda \mathrm{I})^{-1} \cdot (X^T X) \cdot \underrightarrow{\beta}^{OLS} = (\mathrm{I} + \lambda \mathrm{I})^{-1} \cdot \mathrm{I} \cdot \underrightarrow{\beta}^{OLS} = (1+\lambda)^{-1} \cdot \underrightarrow{\beta}^{OLS} \quad (3.16)$$

Με βάση τη σχέση (3.15) για την εκτιμήτρια lasso και τη σχέση (3.16) μπορούμε να επιβεβαιώσουμε ότι η βασική διαφορά των δύο μεθόδων είναι ότι η ridge συρρικνώνει όλους τις παραμέτρους του μοντέλου κατά ένα συγκεκριμένο παράγοντα ενώ η lasso επιλέγει αυτές με τη μεγαλύτερη συνεισφορά στο μοντέλο και μηδενίζει τις υπόλοιπες.

Θέλοντας να προσαρμόσουμε μια L1 τύπου ποινή στο μοντέλο του Cox θα χρησιμοποιήσουμε τη μέθοδο Lasso όπως αναλύσαμε παραπάνω αλλά με κάποιες τροποποιήσεις. Η νέα προσέγγιση της μεθόδου θα έχει στόχο την μεγιστοποίηση του λογαρίθμου της μερικής συνάρτησης πιθανοφάνειας και να συρρικνώσει κάποιες από τις παραμέτρους του μοντέλου και να μηδενίσει όλες τις υπόλοιπες.

Στην περίπτωση του μοντέλου του Cox, η φύση του περιορισμού είναι η ίδια αλλά ο ορισμός του προβλήματος λίγο διαφορετικός. Οι lasso εκτιμήτριες των παραμέτρων του μοντέλου θεωρούμε ότι είναι η λύση του προβλήματος (Goeman, 2010):

$$\underrightarrow{\beta} = \arg\max l(\underrightarrow{\beta}) \text{ δεδομένου ότι } \left\|\underrightarrow{\beta}\right\|_1 = \sum_{j=1}^{p}\left|\beta_j\right| \le c \quad (3.17)$$

όπου l είναι ο λογάριθμος της μερικής συνάρτησης πιθανοφάνειας και p το πλήθος των συμμεταβλητών στο μοντέλο. Μπορούμε να ορίσουμε ένα ισοδύναμο πρόβλημα με το (3.17) για τον καθορισμό των lasso εκτιμητριών μέσω της ποινικοποιημένης συνάρτησης πιθανοφάνειας (penalized likelihood):

$$\underset{\rightarrow}{\beta} = \arg\max\{l(\underset{\rightarrow}{\beta}) - \lambda \left\| \underset{\rightarrow}{\beta} \right\|_1\} \quad (3.18)$$

για δοσμένη συνάρτηση πιθανοφάνειας οι ορισμοί (3.17) και (3.18) είναι ισοδύναμοι καθώς μπορούμε να καταλήξουμε στο πρόβλημα (3.18) κάνοντας χρήση της μεθόδου των πολλαπλασιαστών Lagrange για την βελτιστοποίηση του προβλήματος (3.17).
Έχοντας ορίσει λοιπόν τη συνάρτηση μερικής πιθανοφάνειας στο μοντέλο του Cox χωρίς ισόπαλους χρόνους διακοπής από τη σχέση (2.16), το πρόβλημα (3.18) παίρνει τελικά την εξής μορφή:

$$\underset{\rightarrow}{\beta} = \arg\max\{\sum_{j=1}^{k} (\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_j - \ln[\sum_{i \in R_j} \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_i)]) - \lambda \left\| \underset{\rightarrow}{\beta} \right\|_1\} \quad (3.19)$$

Για την επίλυση του προβλήματος χρησιμοποιείται η μέθοδος κλίσης (Gradient Ascent) έχοντας σταθερή τιμή του λ. Για την εύρεση της βέλτιστης τιμής του λ χρησιμοποιείται η μέθοδος cross validation που περιγράφεται στην παράγραφο 3.4

3.3.3 Η μέθοδος Elastic Net

Όπως αναφέραμε στις δύο προηγούμενες παραγράφους, μπορούμε να χρησιμοποιήσουμε είτε L2 τύπου ποινή (μέθοδος Ridge) είτε L1 τύπου ποινή (μέθοδος Lasso) για τη συρρίκνωση των παραμέτρων του μοντέλου. Όμως, έχει αποδειχθεί ότι, και οι δύο μέθοδοι παρουσιάζουν μειονεκτήματα ανάλογα με τα χαρακτηριστικά του συνόλου δεδομένων που διαχειριζόμαστε.

Ξέρουμε ότι η μέθοδος Ridge ναι μεν συρρικνώνει τις παραμέτρους του μοντέλου αλλά δεν μπορεί ουσιαστικά να κάνει επιλογή των σημαντικών μεταβλητών που θα μετέχουν στο μοντέλου καθώς τις διατηρεί όλες ανέπαφες μέσα σε αυτό. Από την άλλη μεριά, η μέθοδος Lasso φαίνεται ιδιαίτερα ελκυστική ακριβώς επειδή ταυτόχρονα γίνεται συρρίκνωση κάποιων παραμέτρων και όλες οι υπόλοιπες προκύπτουν να είναι μηδενικές ώστε να καταλήξουμε σε μια μορφή μοντέλου με πολύ λιγότερες μεταβλητές. Παρόλα αυτά, ένα βασικό μειονέκτημα της μεθόδου Lasso είναι ότι, όταν διαθέτουμε ένα σύνολο μεταβλητών με υψηλή συσχέτιση μεταξύ τους, τότε τείνει να επιλέξει μόνο μια μεταβλητή από το σύνολο χωρίς να ενδιαφέρεται ποια θα είναι αυτή. Επιπλέον, έχει παρατηρηθεί εμπειρικά ότι, για υψηλές συσχετίσεις μεταξύ των μεταβλητών η προβλεπτική ικανότητα του μοντέλου της μεθόδου Lasso είναι υποδεέστερη του μοντέλου της μεθόδου Ridge.

Για όλους τους παραπάνω λόγους, έχει προταθεί τα τελευταία χρόνια μια νέα μέθοδος ποινής, η λεγόμενη (naive) Elastic Net (Zou & Hastie 2005), η οποία είναι ουσιαστικά μια μίξη των μεθόδων Ridge και Lasso και προσπαθεί να συνδυάσει τα πλεονεκτήματα των δύο μεθόδων. Δηλαδή, θέλουμε να γίνεται σε κάθε περίπτωση συνόλου δεδομένων η καλύτερη δυνατή επιλογή μεταβλητών, όπως στην Lasso, και ταυτόχρονα να επιτυγχάνεται καλή προβλεπτική ικανότητα μοντέλου.

Για να ορίσουμε τις εκτιμήτριες naïve elastic net $\underset{\rightarrow}{\beta}^{net}$ θα θεωρήσουμε αρχικά χωρίς βλάβη της γενικότητας κάποιες παραδοχές που αναφέρθηκαν και σε προηγούμενες παραγράφους. Πιο συγκεκριμένα, θεωρούμε ότι ισχύουν οι συνθήκες:

$$\sum_{i=1}^{n} y_i = 0, \sum_{i=1}^{n} x_{ij} = 0, \sum_{i=1}^{n} x_{ij}^2 = 1, \ j = 1, \ldots, p$$

Για μη αρνητικές σταθερές λ_1, λ_2 ορίζουμε τις εκτιμήτριες naïve elastic net ως την λύση του προβλήματος ελαχιστοποίησης:

$$\underset{\rightarrow}{\beta}^{net} = \arg\min_{\underset{\rightarrow}{\beta}} \left\{ L(\lambda_1, \lambda_2, \underset{\rightarrow}{\beta}) \right\} \quad (3.20)$$

όπου έχουμε ορίσει:

$$L(\lambda_1, \lambda_2, \underset{\rightarrow}{\beta}) = \left\| \underset{\rightarrow}{y} - X \underset{\rightarrow}{\beta} \right\|_2^2 + \lambda_2 \left\| \underset{\rightarrow}{\beta} \right\|_2^2 + \lambda_1 \left\| \underset{\rightarrow}{\beta} \right\|_1$$

Αν ορίσουμε ως $\alpha = \lambda_1 / (\lambda_1 + \lambda_2)$ τότε μπορούμε να διαπιστώσουμε ότι το πρόβλημα είναι ισοδύναμο με το πρόβλημα βελτιστοποίησης:

$$\underset{\rightarrow}{\beta}^{net} = \arg\min \left\{ \left\| \underset{\rightarrow}{y} - X \underset{\rightarrow}{\beta} \right\|_2^2 \right\} \text{ δεδομένου ότι: } (1-a) \left\| \underset{\rightarrow}{\beta} \right\|_2^2 + a \left\| \underset{\rightarrow}{\beta} \right\|_1 \le c \quad (3.21)$$

όπου c μια αυθαίρετη σταθερά.

Η συνάρτηση $(1-a) \left\| \underset{\rightarrow}{\beta} \right\|_2^2 + a \left\| \underset{\rightarrow}{\beta} \right\|_1 \le c$ ονομάζεται elastic net ποινή, και είναι ένας κυρτός συνδυασμός των ποινών των μεθόδων ridge και lasso. Όταν το α=1, τότε η μέθοδος naive elastic net είναι ακριβώς η μέθοδος lasso.

Αν θεωρήσουμε ξανά την ειδική περίπτωση ορθοκανονικού σχεδιασμού τότε μπορούμε να δείξουμε ότι λύση του προβλήματος (3.21) είναι η εξής:

$$\beta_i^{net} = \frac{\left(\left| \beta_i^{OLS} \right| - \lambda_1 / 2 \right)^+}{1 + \lambda_2} sign\left(\beta_i^{OLS} \right) \quad (3.22)$$

Στην περίπτωση του μοντέλου του Cox, θέλουμε να βρούμε το $\underset{\rightarrow}{\beta}$ που μεγιστοποιεί την συνάρτηση μερικής πιθανοφάνειας:

$$L(\underset{\rightarrow}{\beta}) = \prod_{j=1}^{k} \left(\frac{\exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_j)}{\sum_{i \in R_j} \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_i)} \right)$$

δεδομένου του περιορισμού: $(1-a) \sum_{i=1}^{p} \beta_i^2 + a \sum_{i=1}^{p} |\beta_i| \le c$. Το πρόβλημα μεγιστοποίησης της μερικής συνάρτησης πιθανοφάνειας είναι ισοδύναμο με τη μεγιστοποίηση της λογαριθμοποιημένης μερικής συνάρτησης πιθανοφάνειας την οποία για ευκολία στις πράξεις έχει πολλαπλασιαστεί κατά ένα παράγοντα $2/n$ (n=πλήθος παρατηρήσεων). Οπότε, θέλουμε να βρούμε το $\underset{\rightarrow}{\beta}$ που μεγιστοποιεί τη συνάρτηση:

$$\frac{2}{n} l(\underset{\rightarrow}{\beta}) = \frac{2}{n}\left[\sum_{j=1}^{k} \{ \underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_j - \ln[\sum_{i \in R_j} \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_i)] \} \right]$$

Χρησιμοποιώντας τη μέθοδο πολλαπλασιαστών Lagrange, καταλήγουμε ότι οι εκτιμήτριες της μεθόδου naïve elastic net είναι οι λύσεις του προβλήματος:

$$\underset{\rightarrow}{\beta}^{net} = \arg\max_{\underset{\rightarrow}{\beta}} \left[\frac{2}{n} \left(\sum_{j=1}^{k} \{ \underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_j - \ln[\sum_{i \in R_j} \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_i)] \} - \lambda P_a(\underset{\rightarrow}{\beta}) \right) \right] \quad (3.23)$$

όπου ορίσαμε ως:

$$\lambda P_a(\underset{\rightarrow}{\beta}) = \lambda \left(a \sum_{i=1}^{p} |\beta_i| + \frac{1}{2}(1-a) \sum_{i=1}^{p} \beta_i^2 \right) \quad (3.24)$$

Για σταθερή τιμή του λ, όσο το α τείνει στο μηδέν τόσο οι λύσεις που προκύπτουν μοιάζουν με εκείνες τις μεθόδου ridge, ενώ αντίστοιχα όσο το α τείνει στη μονάδα τόσο οι λύσεις προσεγγίζουν τις αντίστοιχες λύσεις της μεθόδου lasso κρατώντας δηλαδή όλο και λιγότερες μεταβλητές και αυξάνοντας τη βαρύτητα των μη μηδενικών συντελεστών. Έχει αποδειχτεί ότι, για α=0.95 (ή και για τιμή ακόμα πλησιέστερη στη μονάδα), η μέθοδος συμπεριφέρεται πολύ παρόμοια με τη μέθοδο lasso, απαλείφοντας απλώς εκφυλιστικές συμπεριφορές στο μοντέλο λόγω πολύ υψηλής συσχέτισης των μεταβλητών (Noah et al., 2011).

Ως μέθοδος επίλυσης του προβλήματος (3.23) προτείνεται ο αριθμητικός αλγόριθμος της Newton-Raphson με κάποιες τροποποίησεις.

3.4 Η μέθοδος cross-validation (cvl) για την επιλογή του βέλτιστου λ

Όπως είδαμε και στις δύο παραπάνω μεθόδους, καθοριστικό ρόλο παίζει η παράμετρος μεροληψίας λ. Για την επιλογή λοιπόν της βέλτιστης τιμής του λ χρησιμοποιείται η μέθοδος cross validation (cvl).

Ας υποθέσουμε ότι έχουμε n το πλήθος παρατηρήσεις και ένα μοντέλο παλινδρόμησης που περιγράφει τα δεδομένα. Αν $l(\underset{\rightarrow}{\beta})$ είναι η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας και $\underset{\rightarrow}{\beta}$ το διάνυσμα παραμέτρων του μοντέλου τότε μπορούμε να ορίσουμε τη συνεισφορά της i-οστής παρατήρησης στην λογαριθμοποιημένη πιθανοφάνεια ως εξής (Verweij & van Houwelingen, 1993)

$$l_i(\underset{\rightarrow}{\beta}) = l(\underset{\rightarrow}{\beta}) - l_{(-i)}(\underset{\rightarrow}{\beta}) \quad (3.25)$$

όπου $l_{(-i)}(\underset{\rightarrow}{\beta})$ είναι η λογαριθμοποιημένη πιθανοφάνεια αν αφήσουμε εκτός την i-οστή παρατήρηση. Αντίστοιχα, θα συμβολίζουμε ως $\underset{\rightarrow(-i)}{\beta}$ εκείνη την επιλογή του $\underset{\rightarrow}{\beta}$ που μεγιστοποιεί την $l_{(-i)}(\underset{\rightarrow}{\beta})$. Αν οι παράμετροι του μοντέλου είναι ανεξάρτητες μεταξύ τους, όπως στην κλασική περίπτωση γραμμικής παλινδρόμησης, τότε η $l_i(\underset{\rightarrow}{\beta})$ αναπαριστά την συνεισφορά της i-οστής παραμέτρου και ισχύει:

$$\sum_{i=1}^{n} l_i(\underset{\rightarrow}{\beta}) = l(\underset{\rightarrow}{\beta}) \quad (3.26)$$

Σε αυτό το σημείο μπορούμε να ορίσουμε την cvl λογαριθμοποιημένη πιθανοφάνεια ως εξής:

$$\text{cvl} = \sum_{i=1}^{n} l_i(\underset{\rightarrow}{\beta}_{(-i)}) \quad (3.27)$$

Για ένα δεδομένο μοντέλο, η cvl προσμετράει πόσο καλά η κάθε παρατήρηση i=1,..,n μπορεί να προβλεθεί από το σύνολο όλων των υπόλοιπων παρατηρήσεων. Μπορούμε να επεκτείνουμε αυτή τη μέθοδο ακόμα και σε μοντέλο διάρκειας ζωής και πιο συγκεκριμένα στο μοντέλο αναλογικής διακινδύνευσης του Cox.

Ξέρουμε ότι στην περίπτωση του μοντέλου του Cox η μερική συνάρτηση πιθανοφάνειας χωρίς ισόπαλους χρόνους διακοπής δίνεται από τον τύπο:

$$L(\underset{\rightarrow}{\beta}) = \prod_{j=1}^{k} \left(\frac{\exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_j)}{\sum_{i \in R_j} \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \cdot \underset{\rightarrow}{x}_i)} \right)$$

Αφήνοντας εκτός την i-οστή μονάδα, τότε απαλείφεται ο i-οστός παράγοντας του μοντέλου και η μονάδα i εξαιρείται από όλα τα σύνολα κινδύνου πριν τη χρονική στιγμή $t_{(i)}$. Αν όλα τα $t_{(i)}$ ταξινομηθούν ώστε $t_{(j)} < t_{(i)}, \forall j < i$ και ορίσουμε ως $w_j = \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \underset{\rightarrow}{x}_j)$ τότε έχουμε:

$$L_{(-i)}(\underset{\rightarrow}{\beta}) = \prod_{j<i} \left(\frac{w_j}{\sum_{k \in R_j} w_k - w_i} \right) \cdot \prod_{j>i} \left(\frac{w_j}{\sum_{k \in R_j} w_k} \right) \quad (3.28)$$

Η συνεισφορά $L_i(\underset{\rightarrow}{\beta})$ της μονάδας i στην μερική πιθανοφάνεια ισούται με $L_i(\underset{\rightarrow}{\beta}) = L(\underset{\rightarrow}{\beta}) / L_{(-i)}(\underset{\rightarrow}{\beta})$ το οποίο μας οδηγεί στην έκφραση:

$$L_i(\underset{\rightarrow}{\beta}) = \prod_{j<i} (1 - p_{ij})\, p_{ii} \quad (3.29)$$

όπου ορίσαμε ως:

$$p_{ij} = \frac{w_i}{\sum_{k \in R_j} w_k} \quad (3.30)$$

Η σχέση (3.30) εκφράζει την πιθανότητα να διακοπεί η λειτουργία της μονάδας i τη χρονική στιγμή $t_{(j)}$, δεδομένων των συνόλων κινδύνου και των χρόνων επιβίωσης. Επομένως, η $L_i(\underset{\rightarrow}{\beta})$ εκφράζει την υπό συνθήκη πιθανότητα η i-οστή μονάδα να επιβιώσει ως τη χρονική στιγμή $t_{(i-1)}$ και, αν το $d_i = 1$, να διακόπτεται η λειτουργία σε χρόνο. Από τη σχέση (3.29) βρίσκουμε ότι η συνεισφορά $l_i(\underset{\rightarrow}{\beta})$ στην λογαριθμοποιημένη πιθανοφάνεια είναι:

$$l_i(\underset{\rightarrow}{\beta}) = \ln(L_i(\underset{\rightarrow}{\beta})) = \sum_{j<i}\left[\ln(1-p_{ij}) + \ln(p_{ii})\right] \quad (3.31)$$

Στο σταθερό μοντέλο χωρίς καμία συμμεταβλητή, οι πιθανότητες p_{ij} απλοποιούνται αρκετά και παίρνουν την εξής μορφή: $p_{ij} = (\mathrm{n} - \mathrm{j} + 1)^{-1}, \forall i$ και αν δεν υπάρχουν αποκομμένες παρατηρήσεις τότε ισχύει επιπλέον ότι: $l_i(0) = -\ln(n)$.

Για τον υπολογισμό της λογαριθμοποιημένης συνάρτησης πιθανοφάνειας cvl, χρειάζεται να υπολογίσουμε τις εκτιμήτριες $\underset{\rightarrow(-\mathrm{i})}{\beta}$. Ο καθορισμός αυτών των εκτιμητριών απαιτεί την προσαρμογή n το πλήθος μοντέλων του Cox με (n-1) το πλήθος παρατηρήσεων το καθένα.

4. Εφαρμογή

4.1 Παρουσίαση του προβλήματος και περιγραφή των μεταβλητών

Τα δεδομένα το προβλήματος που μελετάμε είναι από τον κλάδο της ιατρικής και πιο συγκεκριμένα είναι δεδομένα διάρκειας ζωής 51 ασθενών που πάσχουν από οξεία μυελοπλαστική λευχαιμία μέχρι να συμβεί το γεγονός= θάνατος ή το τέλος της ιατρικής παρακολούθησης (Lee, 1980). Εφαρμόζεται μια θεραπεία αλλά για τον κάθε ασθενή οι συνθήκες που επηρεάζουν την νοσηλεία είναι διαφορετικές καθώς αυτή επηρεάζεται από διάφορους παράγοντες όπως η ηλικία, η θερμοκρασία και μετρήσεις διαγνωστικών εξετάσεων. Συνολικά λοιπόν, έχουμε 7 μεταβλητές (age, smear, absinf, labindex, absblasts, temp, treat) που ενδέχεται να επηρεάζουν την διάρκεια ζωής του ασθενή (time) ενώ επιπλέον υπάρχει και ένας παράγοντας αποκοπής (status) που μας υποδηλώνει αν έχει συμβεί το γεγονός στο συγκεκριμένο άτομο. Η ερμηνεία των μεταβλητών του προβλήματος φαίνεται στον πίνακα 4.1

age	Η ηλικία, σε χρόνια, του ασθενή
smear	Ποσοστό επίστρωσης βλαστοκυττάρων (%)
absinf	Ποσοστό λευχαιμικών κυττάρων που εισήλθαν στο μυελό των οστών (%)
labindex	Ποσοστό κυττάρων που προήλθαν από μυελό των οστών (%)
absblasts	Ο απόλυτός αριθμός βλαστοκυττάρων($\times 10^3$)
temp	Η υψηλότερη θερμοκρασία σώματος ($\times 10^o F$)
treat	Ανταπόκριση σε θεραπεία (ναι=1, όχι=0)
status	Αν ο ασθενής έχει πεθάνει (ναι=1, όχι=0)
time	Χρόνος, σε μήνες, μέχρι να συμβεί το γεγονός= θάνατος ή τέλος ιατρικής παρακολούθησης

Πίνακας 4.1: Περιγραφή των μεταβλητών που μετέχουν στο πρόβλημα

Για την ανάλυση μας θα χρησιμοποιήσουμε τις μεθόδους της ανάλυσης επιβίωσης με χρήση των στατιστικών πακέτων της R και του Minitab. Αρχικά ασχολούμαστε με τις μη παραμετρικές τεχνικές τις οποίες αναλύσαμε στην παράγραφο 1.2 για μια πρώτη εξαγωγή συμπερασμάτων στο Minitab.

Στη συνέχεια, προσαρμόζουμε το ημιπαραμετρκό μοντέλο του Cox στην R. Είναι αρκετά λογικό να χρησιμοποιήσουμε ένα ημιπαραμετρικό μοντέλο εφόσον δεν έχουμε καμία εκ των προτέρων γνώση για το αν τα δεδομένα μας ακολουθούν μια συγκεκριμένη κατανομή ώστε να μπορεί να προσδιοριστεί η κοινή συνάρτηση κινδύνου $h_0(t)$ για όλα τα άτομα. Αυτό ήταν κάτι το οποίο περιμέναμε καθώς, όπως έχουμε ήδη αναφέρει, σε ό,τι αφορά τον άνθρωπο, ο κάθε πληθυσμός είναι διαφορετικός, δηλαδή το κάθε άτομο έχει τα δικά του μοναδικά χαρακτηριστικά που το ξεχωρίζουνε από τα υπόλοιπα άτομα του πληθυσμού και άρα είναι αδύνατο να προσαρμοστεί ένα παραμετρικό μοντέλο για όλους.

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

Στο τελευταίο σκέλος αυτού του κεφαλαίου, προσαρμόζουμε εξ αρχής το μοντέλο του Cox αυτή τη φορά όμως, χρησιμοποιώντας όλες τις τεχνικές συρρίκνωσης, όπως αυτές αναλύθηκαν στο κεφάλαιο 3, για την αντιμετώπιση προβλημάτων πολυσυγγραμμικότητας. Τέλος, κάνουμε σύγκριση των αποτελεσμάτων με τα αποτελέσματα του τυπικού μοντέλου Cox και παραθέτουμε γενικά συμπεράσματα.

4.2 Εφαρμογή μη-παραμετρικών ελέγχων

Από την περιγραφή των μεταβλητών στον πίνακα 4.1 καταλαβαίνουμε ότι ίσως ο σημαντικότερος παράγοντας που ενδέχεται να επηρεάζει τη διάρκεια ζωής των ασθενών είναι ο παράγοντας treat που εκφράζει την ανταπόκριση ή όχι των ασθενών στη θεραπεία. Αρχικά λοιπόν, κατασκευάζουμε τη μη παραμετρική εκτιμήτρια Kaplan-Meier των συναρτήσεων επιβίωσης για τα δεδομένα μας κατηγοριοποιώντας τα σε ασθενείς που ανταποκρίθηκαν στη θεραπεία (treat=1) και σε αυτούς που δεν ανταποκρίθηκαν (treat=0) ώστε να συμπεράνουμε πως αυτό επιδρά στην διάρκεια ζωής. Σε αυτό το σύνολο δεδομένων ο παράγοντας αποκοπής είναι η μεταβλητή status που εκφράζει εάν στον συγκεκριμένο ασθενή συνέβη (status=1) ή όχι (status=0) το γεγονός.

Τα αποτελέσματα που παίρνουμε από το Minitab είναι τα εξής:

Distribution Analysis: time by treat

Variable: time
treat = 0

Censoring Information	Count
Uncensored value	27

Kaplan-Meier Estimates

Time	Number at Risk	Number Failed	Survival Probability	Standard Error	95,0% Normal CI Lower	Upper
0	27	1	0,962963	0,0363447	0,891729	1,00000
1	26	12	0,518519	0,0961590	0,330050	0,70699
2	14	3	0,407407	0,0945607	0,222072	0,59274
3	11	3	0,296296	0,0878772	0,124060	0,46853
4	8	3	0,185185	0,0747568	0,038665	0,33171
5	5	2	0,111111	0,0604812	0,000000	0,22965
7	3	1	0,074074	0,0504010	0,000000	0,17286
12	2	1	0,037037	0,0363447	0,000000	0,10827
13	1	1	0,000000	0,0000000	0,000000	0,00000

Distribution Analysis: time by treat

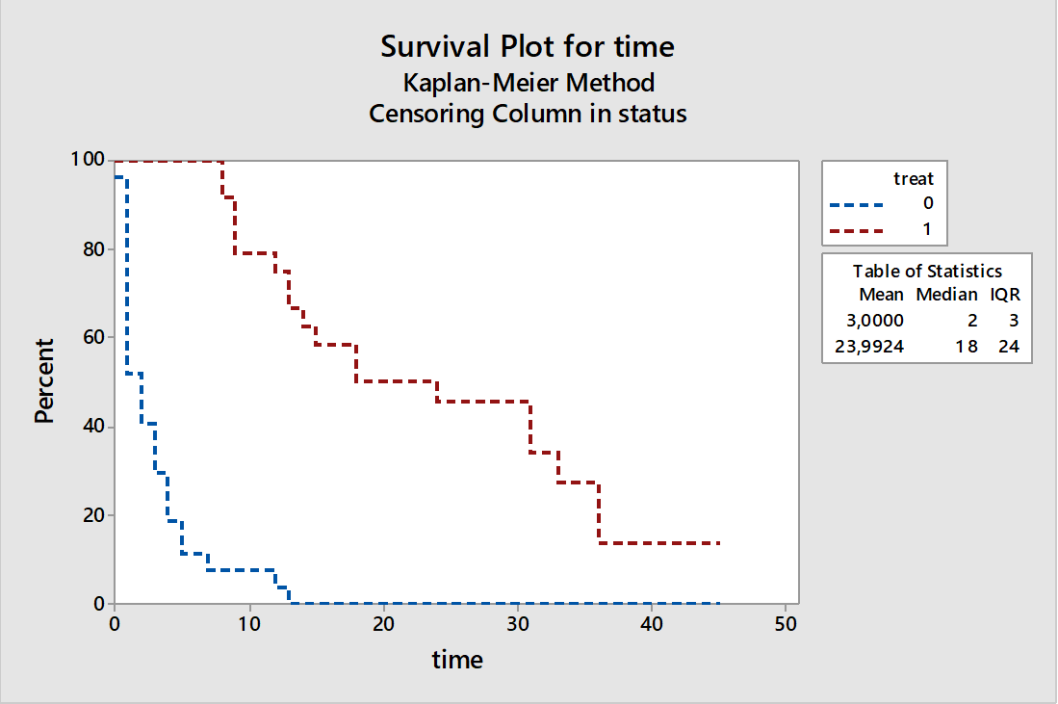
Variable: time
treat = 1

Censoring Information	Count
Uncensored value	18
Right censored value	6

Kaplan-Meier Estimates

Time	Number at Risk	Number Failed	Survival Probability	Standard Error	95,0% Normal CI Lower	Upper
8	24	2	0,916667	0,056417	0,806092	1,00000
9	22	3	0,791667	0,082898	0,629189	0,95414
12	19	1	0,750000	0,088388	0,576762	0,92324
13	18	2	0,666667	0,096225	0,478069	0,85526
14	16	1	0,625000	0,098821	0,431314	0,81869
15	15	1	0,583333	0,100635	0,386093	0,78057
18	14	2	0,500000	0,102062	0,299962	0,70004
24	11	1	0,454545	0,102407	0,253832	0,65526
31	8	2	0,340909	0,103641	0,137776	0,54404
33	5	1	0,272727	0,102925	0,070998	0,47446
36	4	2	0,136364	0,085423	0,000000	0,30379

Το γράφημα των Kaplan-Meier εκτιμητριών των δύο ομάδων ασθενών που μας επιστρέφει το Minitab είναι το ακόλουθο:



Γράφημα 4.1: Γράφημα των Kaplan-Meier εκτιμητριών για τις δύο ομάδες ασθενών

Σχόλια: 1) Όπως περιμέναμε ο παράγοντας ανταπόκριση σε θεραπεία δείχνει να επηρεάζει τη διάρκεια ζωής των ασθενών. Και οι δύο εκτιμήτριες σαν συναρτήσεις είναι φθίνουσες. Η πιθανότητα επιβίωσης των ασθενών και των δύο ομάδων μειώνεται με την πάροδο του χρόνου όμως η Kaplan-Meier εκτιμήτρια της ομάδας ασθενών που ανταποκρίθηκαν στη θεραπεία (treat=1) είναι σταθερά στο χρόνο πιο πάνω από την αντίστοιχη της ομάδας ασθενών που δεν ανταποκρίθηκαν (treat=0).

2) Είναι φανερό ότι για χρόνο T>10 μήνες η $S_{KM}(t)$ της ομάδας ασθενών που δεν ανταποκρίθηκαν στη θεραπεία μηδενίζεται, δηλαδή στους ασθενείς οι οποίοι δεν ανταποκρίνονται στη θεραπεία μετά τους 10 μήνες νοσηλείας είναι βέβαιο ότι θα συμβεί το γεγονός. Από την άλλη μεριά, παρατηρούμε ότι για τους ασθενείς που ανταποκρίθηκαν στη θεραπεία για χρόνο T>35 μήνες η εκτιμήτρια $S_{KM}(t)$ σταθεροποιείται. Αυτό μπορεί να είναι ένα πρώτο σημάδι ότι το διάστημα [0,35]μήνες είναι ο χρόνος που απαιτείται για να ξεκινήσει

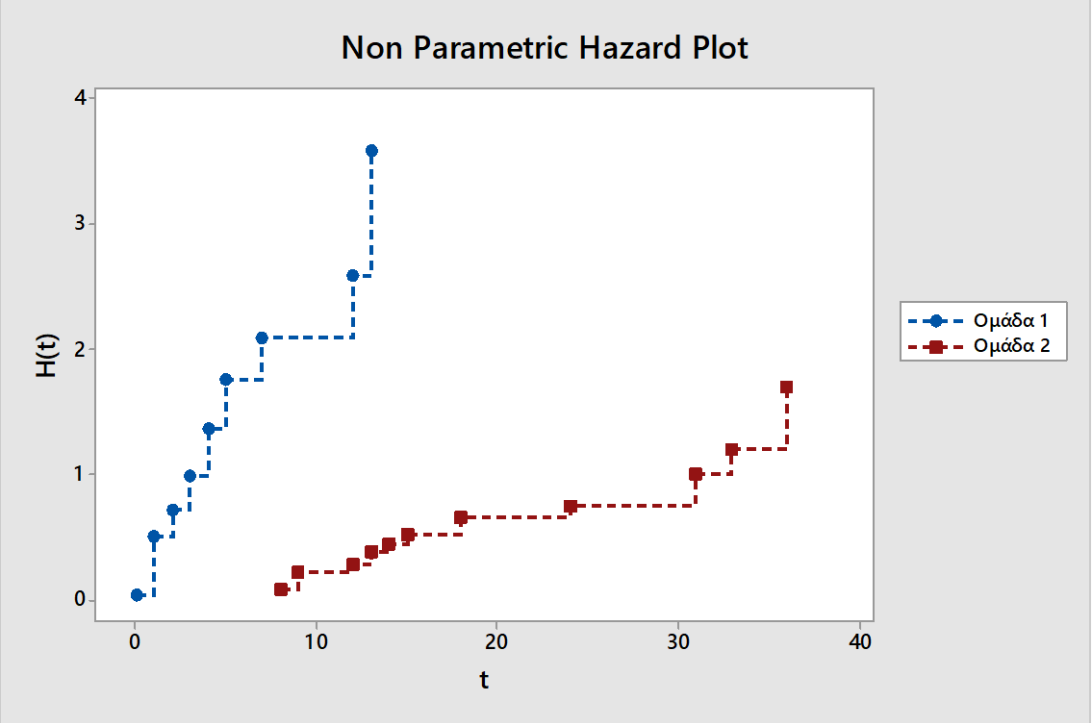
να αποδίδει αποτελέσματα η θεραπεία και μετά από αυτό η υγεία των ασθενών σταθεροποιείται.

3) Πρέπει να παρατηρήσουμε επίσης ότι για T<12 μήνες οι εκτιμήτριες και των δύο ομάδων έχουν παρόμοια συμπεριφορά. Και οι δύο παρουσιάζουν πολύ «απότομη» καθοδική κλίση. Όμως για T>12 μήνες ενώ η $S_{KM}(t)$ της ομάδας ασθενών που δεν ανταποκρίθηκαν στη θεραπεία μηδενίζεται οριστικά η εκτιμήτρια $S_{KM}(t)$ της ομάδας ασθενών που ανταποκρίθηκαν στη θεραπεία ναι μεν συνεχίζει να φθίνει αλλά πλέον με πολύ μικρότερη κλίση. Αυτό ίσως είναι ένα πρώτο σημάδι ότι η θεραπεία χρειάζεται περίπου ένα διάστημα 12 μηνών ώστε να σταματήσει τη ραγδαία επιδείνωση της υγείας των ασθενών.

Στη συνέχεια θα κατασκευάσουμε την μη παραμετρική εκτιμήτρια Nelson-Aalen της σωρευτικής συνάρτησης διακινδύνευσης για τις δύο ομάδες ασθενών. Για να υπολογίσουμε τις τιμές των Nelson-Aalen εκτιμητριών θα πρέπει να έχουμε τις τιμές των n_j και d_j για τις δύο ομάδες ασθενών. Στο εξής για λόγους συντομίας θα αναφερόμαστε στην ομάδα ασθενών που δεν ανταποκρίθηκε στη θεραπεία ως Ομάδα 1 και στην ομάδα ασθενών που ανταποκρίθηκαν στην θεραπεία ως Ομάδα 2. Άρα έχουμε τον κάτωθι συγκεντρωτικό πίνακα:

				Nelson-Aalen εκτιμήτρια
	$t_{(j)}$	n_j	d_j	$H_{NA}(t)$
Ομάδα 1	0	27	1	0,0370
	1	26	12	0,4986
	2	14	3	0,7129
	3	11	3	0,9856
	4	8	3	1,3606
	5	5	2	1,7606
	7	3	1	2,0940
	12	2	1	2,5939
	13	1	1	3,5939
Ομάδα 2	8	24	2	0,0833
	9	22	3	0,2197
	12	19	1	0,2723
	13	18	2	0,3834
	14	16	1	0,4459
	15	15	1	0,5126
	18	14	2	0,6555
	24	11	1	0,7464
	31	8	2	0,9964
	33	5	1	1,1964
	36	4	2	1,6964

Πίνακας 4.2: Πίνακας των τιμών της μη παραμετρικής εκτιμήτριας Nelson-Aalen της σωρευτικής συνάρτησης διακινδύνευσης H(t) για τις δύο ομάδες ασθενών



Γράφημα 4.2: Γράφημα των Nelson-Aalen εκτιμητριών για τις δύο ομάδες ασθενών

Σχόλιο: Όπως ήταν αναμενόμενο και οι δύο συναρτήσεις είναι σταθερά αυξητικές συναρτήσει του χρόνου, πράγμα που σημαίνει ότι ο κίνδυνος για την εμφάνιση του γεγονότος (θάνατος ασθενή) συνεχώς μεγαλώνει με το πέρασμα του χρόνου. Είναι εμφανές ότι κάθε χρονική στιγμή η σωρευτική συνάρτηση διακινδύνευσης της ομάδας 1 είναι μεγαλύτερη από την αντίστοιχη της ομάδας 2 κάτι που επιβεβαιώνει ξανά ότι η μεταβλητή treat δείχνει να επηρεάζει τη διάρκεια ζωής των ασθενών.

Στη συνέχεια εφαρμόζουμε τον μη παραμετρικό έλεγχο Log-rank για τα δεδομένα των δύο ομάδων ασθενών. Ξέρουμε ότι η δειγματοσυνάρτηση του ελέγχου Log-rank υπό την μηδενική υπόθεση $H_0 : S_1(t) = S_2(t)$ ακολουθεί την $X^2_{(1)}$ ασυμπτωτικά (όπου $S_1(t) = S_2(t)$ οι συναρτήσεις επιβίωσης των δύο ομάδων ασθενών). Άρα ελέγχοντας την ισχύ της μηδενικής υπόθεσης H_0 ελέγχουμε κατά πόσο ο παράγοντας treat επηρεάζει τη διάρκεια ζωής. Με αυτόν τον τρόπο θα έχουμε εκτός των δύο παραπάνω γραφικών ελέγχων που εκτελέσαμε και μια επιπλέον μαθηματική επιβεβαίωση για τη σημαντικότητα του παράγοντα treat..

Τα αποτελέσματα του μη παραμετρικού ελέγχου Log-rank παρουσιάζονται στον πίνακα 4.3

$t_{(j)}$	n_{1j}	n_{2j}	d_{1j}	d_{2j}	u_j	v_j
0	27	24	1	0	0,470588	0,249135
1	26	24	12	0	5,76	2,322808
2	14	24	3	0	1,894737	0,660328
3	11	24	3	0	2,057143	0,608499
4	8	24	3	0	2,25	0,52621
5	5	24	2	0	1,655172	0,275183
7	3	24	1	0	0,888889	0,098765
8	2	24	0	2	-0,15385	0,136331
9	2	22	0	3	-0,25	0,209239
12	2	19	1	1	0,809524	0,163719
13	1	18	1	2	0,842105	0,132964
14	0	16	0	1	0	0
15	0	15	0	1	0	0
18	0	14	0	2	0	0
24	0	11	0	1	0	0
31	0	8	0	2	0	0
33	0	5	0	1	0	0
36	0	4	0	2	0	0
Σύνολο					16,22431	5,383181

Πίνακας 4.3: Για την εκτέλεση του μη παραμετρικού ελέγχου Log-rank πρέπει να υπολογιστεί σε κάθε χρονική στιγμή $t_{(j)}$ το πλήθος των μονάδων n_{1j} και n_{2j} που βρίσκονταν σε κίνδυνο ακριβώς πριν αυτή τη χρονική στιγμή καθώς επίσης και το πλήθος των μονάδων d_{1j} και d_{2j} στις οποίες συνέβη το γεγονός και από τις δύο ομάδες.

Το στατιστικό ελέγχου Log-rank ορίζεται ως: $z = \frac{u^2}{v}$ όπου: $u = \sum_{j=1}^{k} u_j$ και $v = \sum_{j=1}^{k} v_j$

και ξέρουμε ότι για μεγάλο k ισχύει ότι: $z \sim X^2_{(1)}$. Άρα λοιπόν:

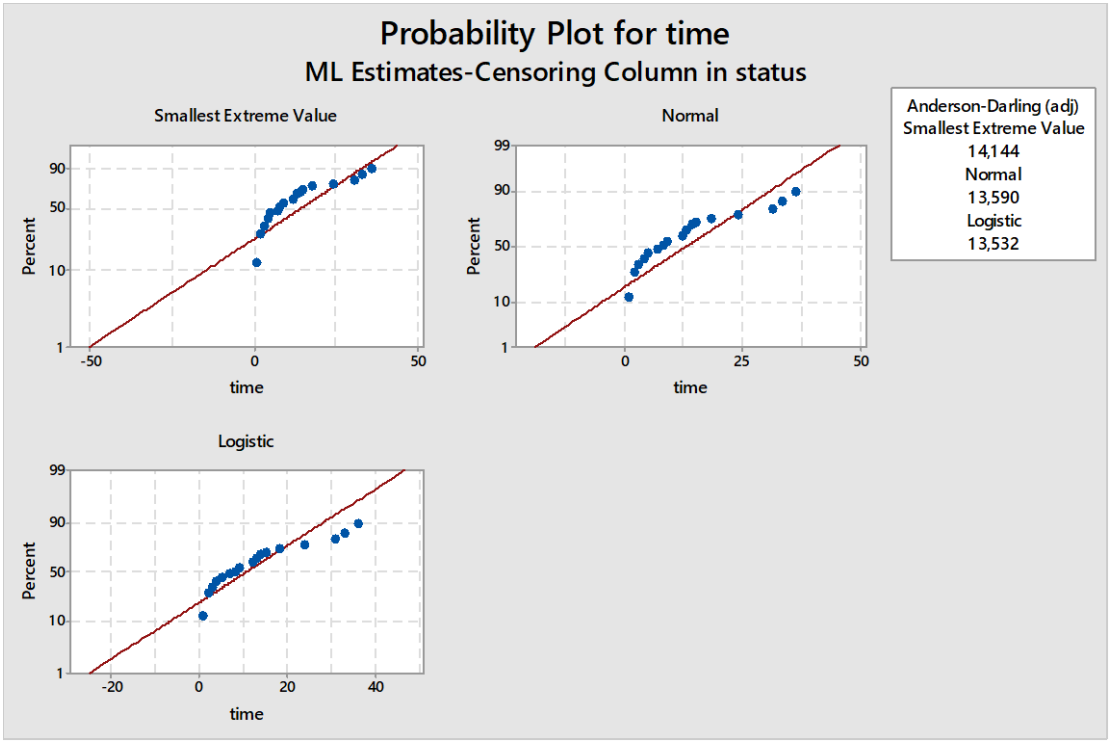
$$z = \frac{16.224^2}{5.383} = 48.898 \sim X^2_{(1)}$$ επομένως: p-value= $P[z > 48.898 \mid z \sim X^2_{(1)}] < 0.0001$

Η p-value είναι πάρα πολύ μικρή άρα απορρίπτουμε την Ho, άρα οι δύο ομάδες ασθενών διαφέρουν σημαντικά ως προς τη διάρκεια ζωής επομένως ο παράγοντας treat είναι στατιστικά σημαντικός.

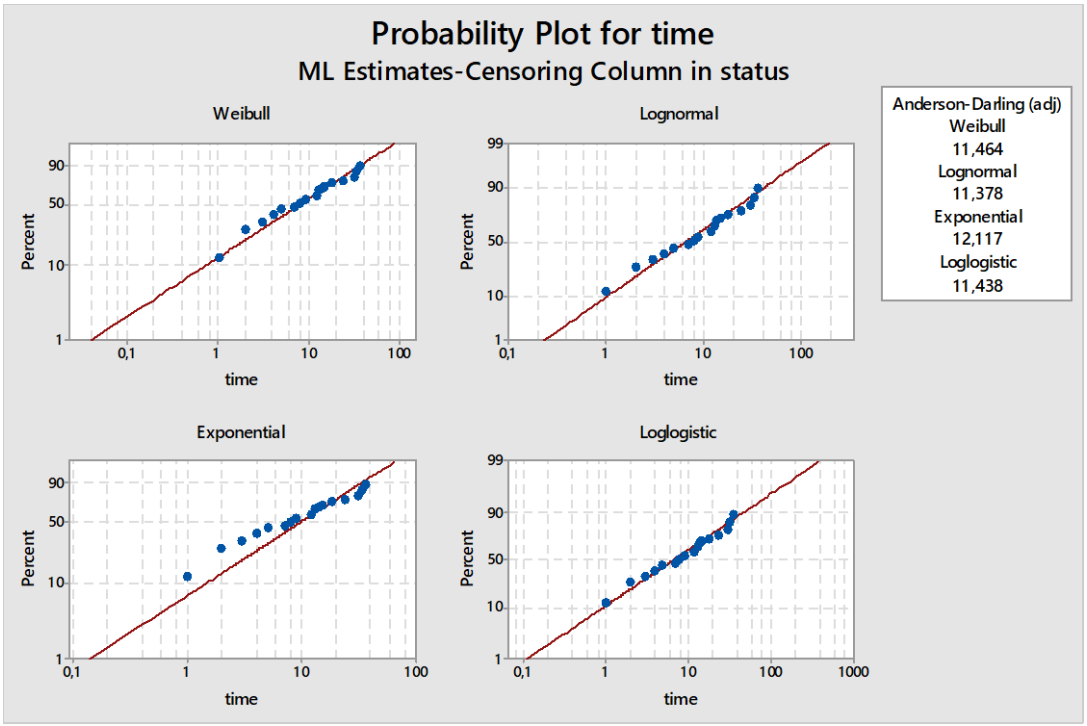
Τέλος, παρότι δεν προσδιορίζεται κάποιο παραμετρικό μοντέλο το οποίο ακολουθούν τα δεδομένα μας, μπορούμε να κάνουμε γραφικούς ελέγχους για την εύρεση κάποιας υποβόσκουσας κατανομής. Πιο συγκεκριμένα, θα χρησιμοποιήσουμε την εκτιμήτρια Kaplan-Meier για όλο το σύνολο ασθενών αυτή τη φορά, και θα διενεργηθούν με τη βοήθεια του Minitab γραφικοί έλεγχοι για να προσδιορίσουμε ποια κατανομή θα ήταν η καταλληλότερη για την περιγραφή των δεδομένων μας.

Το Minitab για να μας υποδείξει ποια κατανομή ταιριάζει καλύτερα εκτελεί τον έλεγχο Kolmogorov-Smirnov ο οποίος χρησιμοποιεί την ελεγχοσυνάρτηση Anderson-Darling. Όσο μικρότερη τιμή παίρνει το στατιστικό ελέγχου τόσο καλύτερα προσαρμόζεται το σύνολο δεδομένων στην εν λόγω κατανομή.

Σχόλιο: Για να μην επηρεαστούν ιδιαίτερα τα γραφικά αποτελέσματα από πολλαπλούς ισόπαλους χρόνους διακοπής, θεωρήσαμε ότι πρέπει να σχεδιαστεί σε κάθε τέτοια περίπτωση μόνο η διάμεσος αυτών των χρόνων.
Τα γραφικά αποτελέσματα αυτών των ελέγχων φαίνονται στα γραφήματα 4.3 και 4.4



Γράφημα 4.3: Έλεγχος προσαρμογής των δεδομένων στα παραμετρικά μοντέλα των κατανομών Smallest extreme value (Gumbel), Normal και Logistic



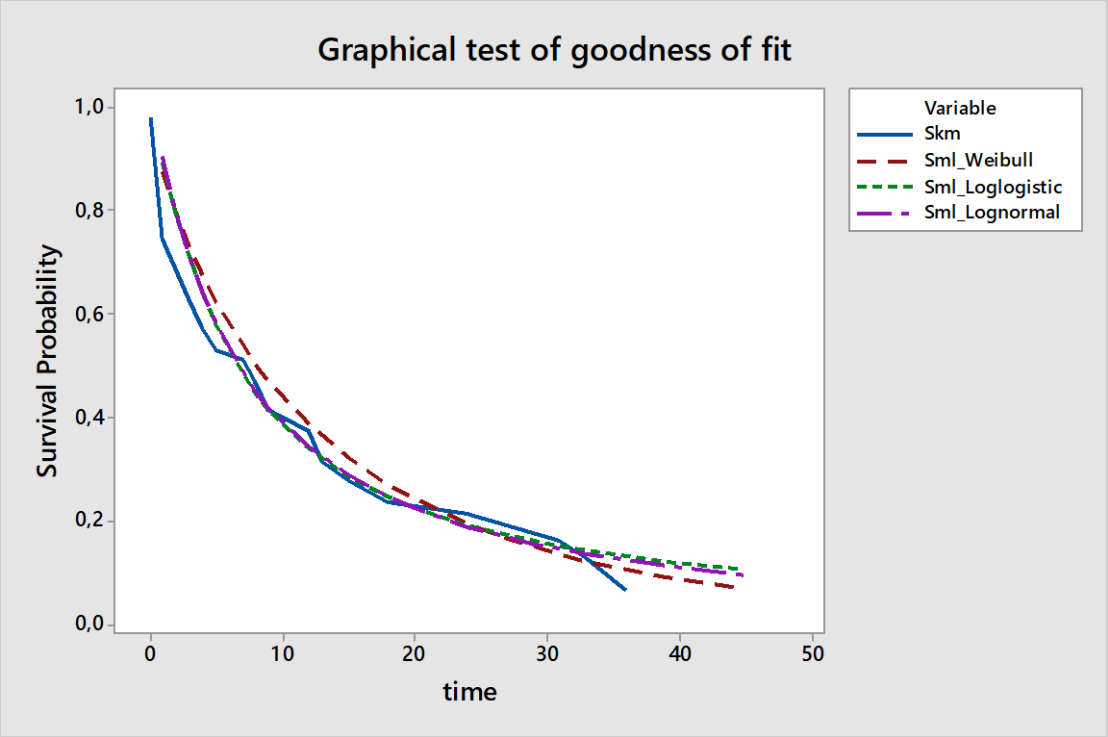
Γράφημα 4.4: Έλεγχος προσαρμογής των δεδομένων στα παραμετρικά μοντέλα των κατανομών Weibull, Lognormal, Exponential και Loglogistic

Όπως αναφέραμε και πρωτύτερα, θεωρούμε ως καταλληλότερη κατανομή για την περιγραφή των δεδομένων μας εκείνη η οποία λαμβάνει την μικρότερη τιμή της (προσαρμοσμένης) ελεγχοσυνάρτησης Anderson-Darling. Στον πίνακα 4.4 παρατίθενται οι τιμές της ελεγχοσυνάρτησης Anderson-Darling κατά φθίνουσα σειρά για κάθε κατανομή που εξετάζεται(από την χειρότερη στην καλύτερη).

Κατανομή	Anderson-Darling (adjusted)
Gumbel	14,144
Normal	13,59
Logistic	13,532
Exponential	12,117
Weibull	11,464
Loglogistic	11,438
Lognormal	11,378

Πίνακας 4.4: Πίνακας των τιμών της ελεγχοσυνάρτησης Anderson-Darling adjusted για τις διάφορες περιπτώσεις κατανομών

Όπως αναμενόταν, συμμετρικές κατανομές όπως η κανονική και η λογιστική κατανομή δεν μπορούν να προσαρμοστούν καλά σε δεδομένα διάρκειας ζωής ενώ το μοντέλο της εκθετικής κατανομής όπως αναφέραμε και στο κεφάλαιο 1 έχει σταθερή συνάρτηση επιβίωσης κάτι το οποίο στις εφαρμογές είναι μη ρεαλιστικό. Από τις κατανομές Weibull, Loglogistic, Lognormal καλύτερα φαίνεται να προσαρμόζεται η Lognormal στα δεδομένα μας αν και η διαφορά της με τις Weibull και Loglogistic όπως φαίνεται και από τον πίνακα 4.4 είναι πολύ μικρή.
Ένας επιπλέον γραφικός έλεγχος που μπορεί να γίνει είναι να υπολογίσουμε τις τιμές των εκτιμητριών μέγιστης πιθανοφάνειας για τις συναρτήσεις επιβίωσης των κατανομών Lognormal, Loglogistic και Weibull και να τις συγκρίνουμε με τις τιμές της μη παραμετρικής εκτιμήτριας Kaplan-Meier. Τα γραφικά αποτελέσματα φαίνονται στο γράφημα 4.5



Γράφημα 4.5: Γραφικός έλεγχος προσαρμογής των δεδομένων σε κάποια γνωστή κατανομή

Όσο πιο κοντά γραφικά στην μη παραμετρική εκτιμήτρια Kaplan-Meier βρίσκεται κάποια από τις παραμετρικές συναρτήσεις επιβίωσης τόσο καλύτερα προσαρμόζονται τα δεδομένα μας στην εν λόγω κατανομή. Άρα, παρατηρούμε ότι και τα τρία παραμετρικά μοντέλα έχουν ικανοποιητική προσαρμογή, αν και μπορούμε να πούμε ότι ίσως πιο καλά προσαρμόζεται το μοντέλο της Lognormal κατανομής (μωβ γραμμή), όπως άλλωστε βρήκαμε και από τον μη παραμετρικό έλεγχο Anderson-Darling.

4.3 Το μοντέλο του Cox χωρίς ποινή

Σε αυτό το σημείο θα προσαρμόσουμε το κλασικό μοντέλο του Cox στα δεδομένα για τη διάρκεια ζωής των ασθενών που πάσχουν από λευχαιμία. Για αυτό το σκοπό θα χρησιμοποιήσουμε το στατικό πακέτο της R και τις βιβλιοθήκες survival και splines.

4.3.1 Έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης

Πριν ξεκινήσουμε την προσαρμογή του μοντέλου θα πρέπει να ελέγξουμε την ισχύ της υπόθεσης της αναλογικής διακινδύνευσης του μοντέλου. Θα χρησιμοποιήσουμε την γραφική τεχνική που μελετήσαμε στην παράγραφο 2.1.3.3, δηλαδή δεδομένης της σχέσης:

$$S(t;\underset{\rightarrow}{x}) = \exp(-H_0(t)\cdot\exp(\underset{\rightarrow}{\beta}^{\mathrm{T}}\cdot\underset{\rightarrow}{x})) \Rightarrow \ln(-\ln(S(t;\underset{\rightarrow}{x})) - \ln(H_0(t)) = \underset{\rightarrow}{\beta}^{\mathrm{T}}\cdot\underset{\rightarrow}{x}$$

μπορούμε να κάνουμε το γράφημα της $\ln(-\ln(S_{KMi}(t)))$ (όπου S_{KMi} η εκτιμήτρια Kaplan-Meier της i-οστής συμμεταβλητής του μοντέλου) έναντι του χρόνου t. Αν για τις διάφορες κατηγορίες μιας μεταβλητής είναι το γραφικό αποτέλεσμα είναι παράλληλες ευθείες τότε ισχύει η υπόθεση της αναλογικότητας της διακινδύνευσης για την εν λόγω συμμεταβλητή. Για να γίνει το γράφημα πιο ομαλό μπορούμε να χρησιμοποιήσουμε μια συνάρτηση του χρόνου t στον οριζόντιο άξονα. Για αυτό το λόγο επιλέγουμε χρησιμοποιήσουμε τον λογάριθμο του χρόνου lnt. Τέτοιου είδους γραφήματα μπορούν να γίνουν αν έχουμε κατηγορικές συμμεταβλητές στο μοντέλο. Για τις ποσοτικές συμμεταβλητές θα πρέπει να γίνει κατάλληλη ομαδοποίηση.

Από τις συμμεταβλητές που διαθέτουμε μόνο η treat είναι κατηγορική και οι υπόλοιπες είναι ποσοτικές. Για τις υπόλοιπες 6 συμμεταβλητές θα δημιουργήσουμε 2 επίπεδα κατηγοριών για την καθεμία. Ένας τρόπος ομαδοποίησης θα ήταν να χρησιμοποιήσουμε τον δειγματικό μέσο κάθε συμμεταβλητής και να χωρίσουμε τα δεδομένα σε αυτά που είναι κάτω και πάνω από τον μέσο όρο. Όμως, επειδή ο δειγματικός μέσος επηρεάζεται από ακραίες παρατηρήσεις θα προτιμήσουμε την δειγματική διάμεσο. Άρα με τη βοήθεια του Minitab βρίσκουμε τη δειγματική διάμεσο για κάθε ποσοτική συμμεταβλητή:

Συμμεταβλητή	age	smear	absinf	labindex	absblasts	temp
Διάμεσος	50	69	61	9	2,6	990

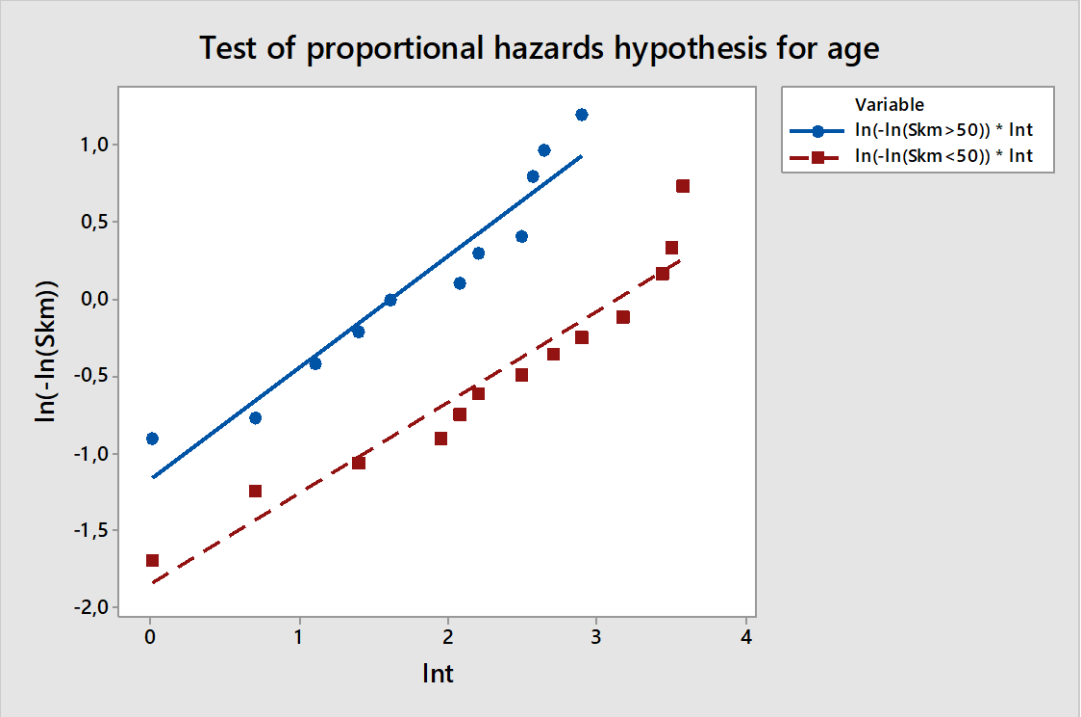
Άρα, μπορούμε πλέον να ορίσουμε τις εξής συμμεταβλητές:

$$age_1 = \begin{cases} 1, \alpha\nu\ age < 50 \\ 0, \alpha\nu\ \text{age} \geq 50 \end{cases}, smear_1 = \begin{cases} 1, \alpha\nu\ \text{smear} < 69 \\ 0, \alpha\nu\ smear \geq 69 \end{cases}, ab\sin f_1 = \begin{cases} 1, \alpha\nu\ ab\sin f > 61 \\ 0, \alpha\nu\ ab\text{sinf} \leq 61 \end{cases}$$

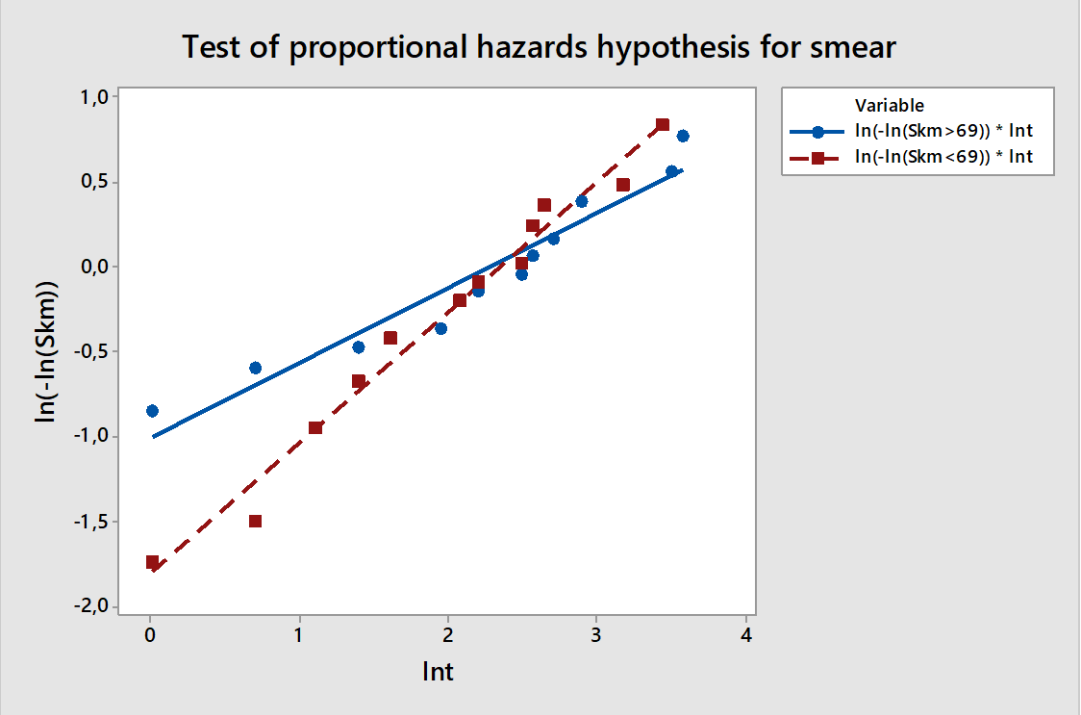
$$labindex_1 = \begin{cases} 1, \alpha\nu\ labindex < 9 \\ 0, \alpha\nu\ labindex \geq 9 \end{cases}, absblasts_1 = \begin{cases} 1, \alpha\nu\ absblasts > 2.6 \\ 0, \alpha\nu\ \text{absblasts} \leq 2.6 \end{cases}, temp_1 = \begin{cases} 1, \alpha\nu\ temp < 990 \\ 0, \alpha\nu\ temp \geq 990 \end{cases}$$

Με βάση αυτές τις νέες συμμεταβλητές μπορούμε να υπολογίσουμε την εκτιμήτρια S_{KM} καθεμίας από αυτές. Άρα, κάνουμε το γράφημα $\ln(-\ln(S_{KM\,i}(t)))$ έναντι του t και να ελέγξουμε την υπόθεση της αναλογικότητας του διακινδύνευσης.

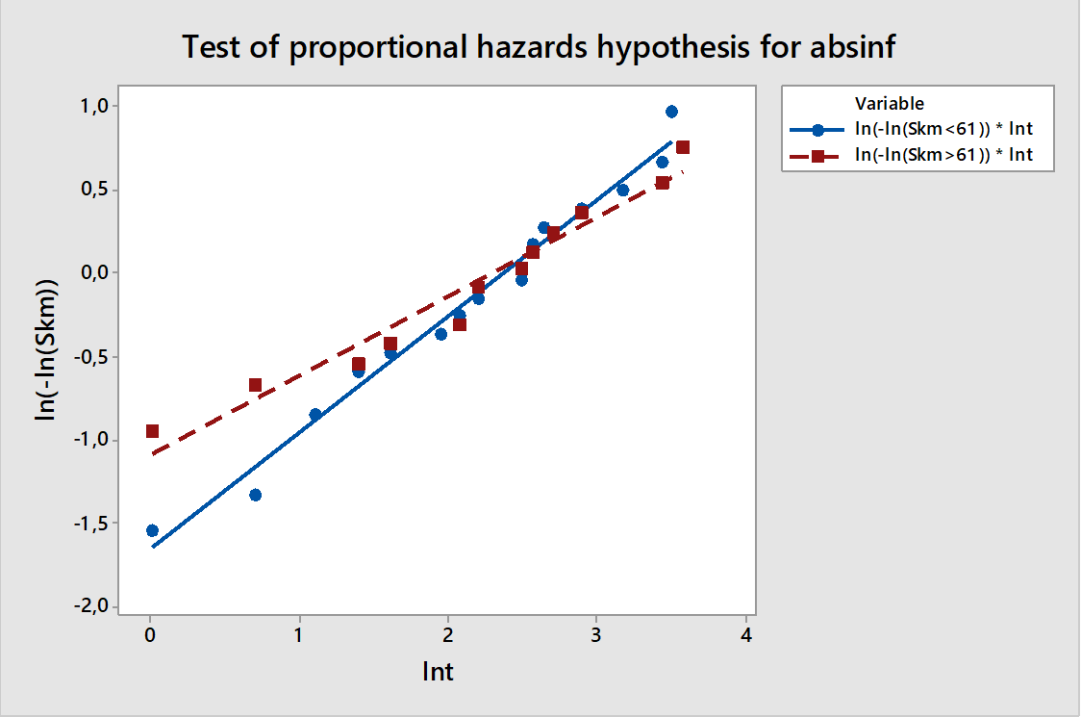
Για την κατασκευή των γραφικών παραστάσεων θα χρησιμοποιήσουμε τη βοήθεια του Minitab. Τα αποτελέσματα φαίνονται στα γραφήματα 4.6-4.12



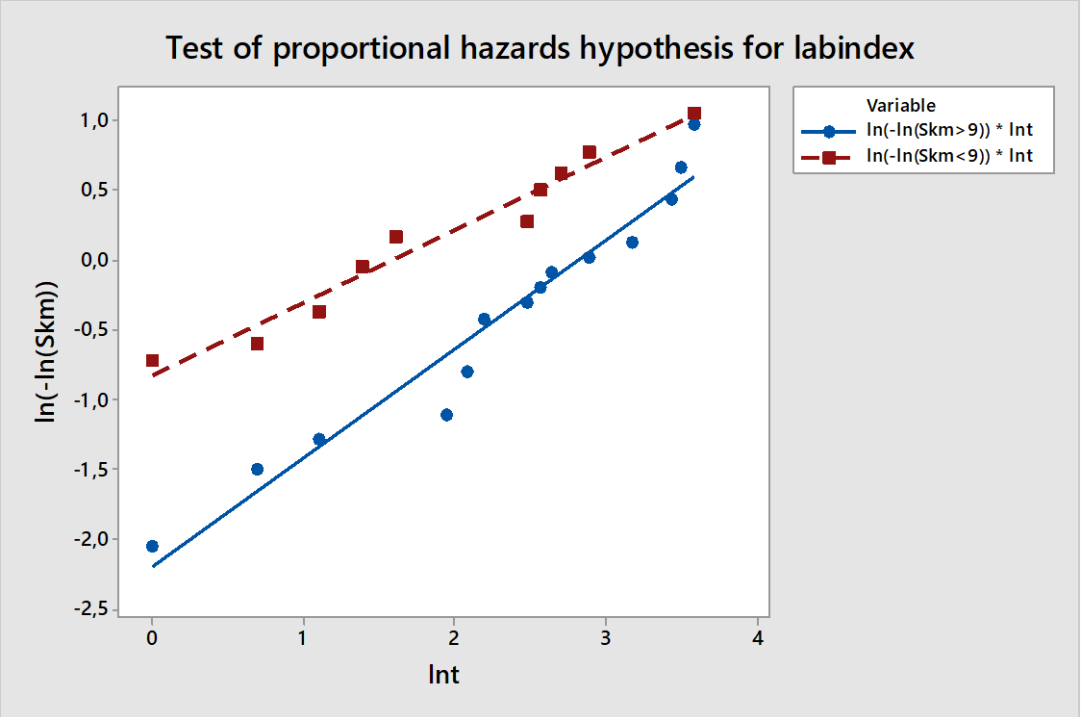
Γράφημα 4.6: Έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης για την μεταβλητή age



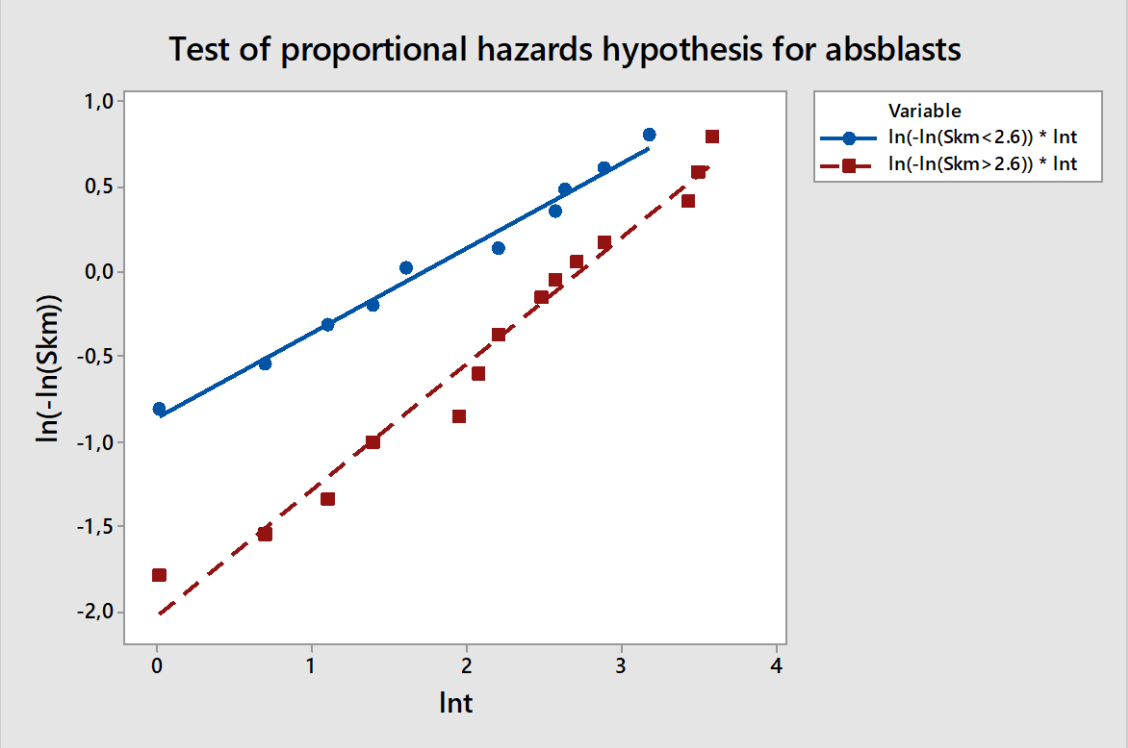
Γράφημα 4.7: Έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης για την μεταβλητή smear



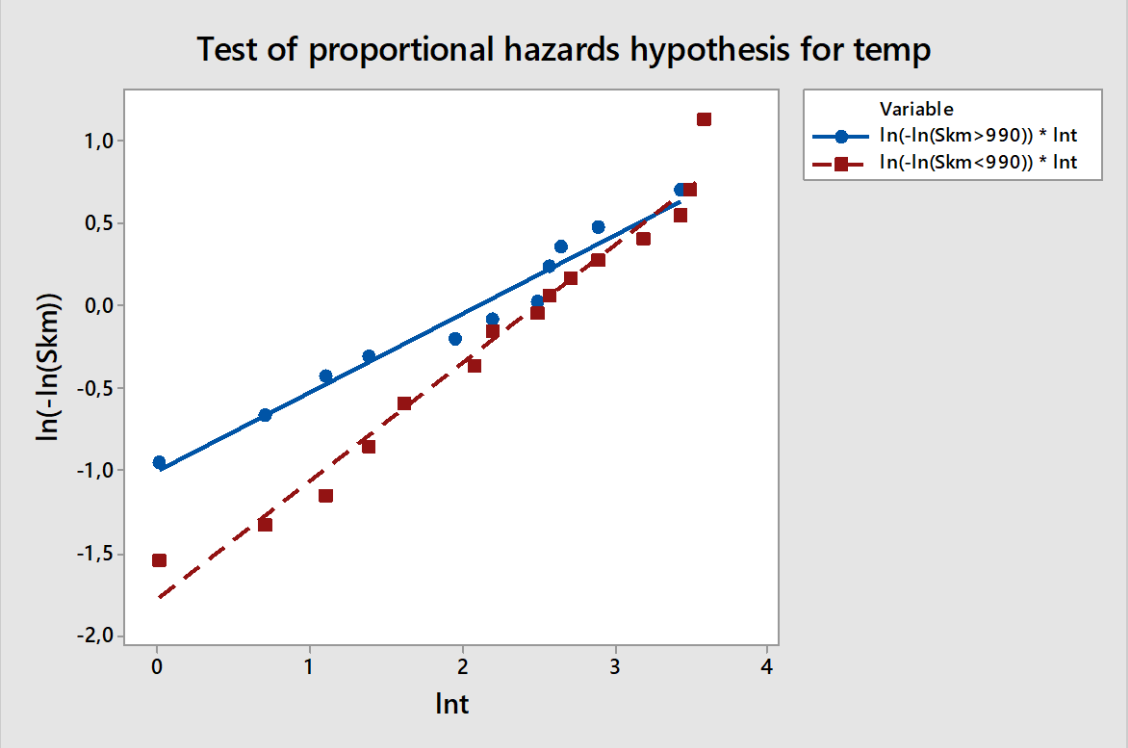
Γράφημα 4.8: Έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης για την μεταβλητή absinf



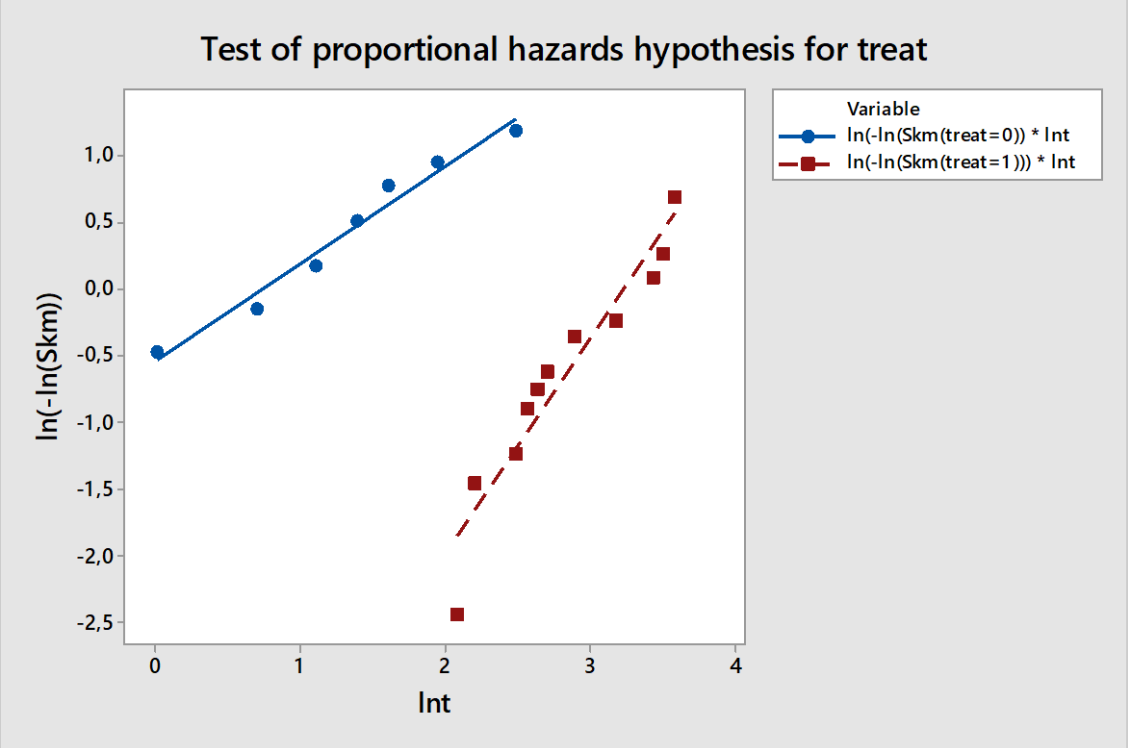
Γράφημα 4.9: Έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης για την μεταβλητή labindex



Γράφημα 4.10: Έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης για την μεταβλητή absblasts



Γράφημα 4.11: Έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης για την μεταβλητή temp



Γράφημα 4.12: Έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης για τη μεταβλητή treat

Από τα γραφήματα 4.6-4.12 διαπιστώνουμε ότι είναι εύλογη η υπόθεση της αναλογικής διακινδύνευσης για τις μεταβλητές treat, absblasts, labindex και age ενώ για τις μεταβλητές temp, absinf και smear φαίνεται να μην ισχύει καθώς οι ευθείες για τις δύο κατηγορίες ασθενών είναι τεμνόμενες. Ειδικά για το γράφημα 4.12 της μεταβλητής treat, η οποία εκφράζει την ανταπόκριση των ασθενών στη θεραπεία, μπορούμε να διαπιστώσουμε ότι εκπληρώνεται ικανοποιητικά η υπόθεση της αναλογικής διακινδύνευσης και επίσης οι ευθείες για τις δύο κατηγορίες ασθενών έχουν μεγάλη κατακόρυφη απόσταση κάτι που επιβεβαιώνει άλλη μια φορά το συμπέρασμά μας ότι ο παράγοντας treat επιδρά καθοριστικά στην διάρκεια ζωής.

Συμπερασματικά, με αυτές τις ενδείξεις μπορεί η υπόθεση της αναλογικότητας της διακινδύνευσης φαίνεται να μην εκπληρώνεται για όλες τις συμμεταβλητές του μοντέλου, παρόλα αυτά θα προσαρμόσουμε το ημιπαραμετρικό μοντέλο του Cox και θα την επανεξετάσουμε μέσω των κλιμακοποιημένων των υπολοίπων Schoenfeld.

4.3.2 Προσαρμογή του μοντέλου

Εισάγοντας τα δεδομένα μας στην R, μπορούμε εύκολα να προσαρμόσουμε το κλασικό μοντέλο του Cox χωρίς ποινή. Τα αποτελέσματα φαίνονται στον πίνακα 4.5

Μεταβλητή	Συντελεστής	Τυπικό σφάλμα	Στατιστικό ελέγχου Wald	p-value
age	0.023576	0.012218	1.930	0.0536
smear	0.029983	0.015533	1.930	0.0536
absinf	-0.004178	0.012164	-0.343	0.7313
labindex	0.072799	0.043968	1.656	0.0978
absblasts	-0.049489	0.024699	-2.004	0.0451
temp	0.008251	0.014446	0.571	0.5679
treat1	-3.365606	0.587477	-5.729	1.01e-08

Πίνακας 4.5: Αποτελέσματα από το μοντέλο του Cox που περιέχει και τις 7 συμμεταβλητές

Σχόλιο: Στον πίνακα αποτελεσμάτων υπάρχει και η μεταβλητή treat1. Επειδή η μεταβλητή treat είναι κατηγορική με 2 επίπεδα κατηγοριών, η R το αναγνωρίζει αυτόματα και παίρνει ως κατηγορία αναφοράς εκείνο το επίπεδο στο οποίο έχουμε αναθέσει την τιμή 0, άρα εν προκειμένω την κατηγορία treat=0 (ασθενείς που δεν ανταποκρίθηκαν στη θεραπεία). Άρα, όταν αναφερόμαστε στην μεταβλητή treat1 στο μοντέλο αυτό έχει να κάνει με την εκτίμηση της παραμέτρου treat για το επίπεδο treat=1 (ασθενείς που ανταποκρίθηκαν στη θεραπεία).

Με βάση τα αποτελέσματα του πίνακα 4.5, μπορούμε να πούμε ότι προσαρμόζοντας αρχικά το μοντέλο του Cox στα δεδομένα των ασθενών με λευχαιμία συμπεριλαμβάνοντας και τις 7 συμμεταβλητές (age, smear, absinf, labindex, absblasts, temp, treat) οι παράγοντες που προκύπτουν στατιστικά σημαντικοί σε επίπεδο σημαντικότητας 5% είναι οι: treat, και absblasts. Αυτό το συμπέρασμα μας βασίζεται στις p-values που μας δίνει η R από τον έλεγχο Wald κατά τον οποίο εξετάζεται η σημαντικότητα κάθε παράγοντα ελέγχοντας την υπόθεση $\{H_0 : \beta_j = 0 \; vs \; H_1 : \beta_j \neq 0\}$.

Το στατιστικό ελέγχου είναι το:

$$\mathrm{w} = \frac{\beta - \beta_{|\mathrm{H}_0}}{se(\beta)} = \frac{\beta}{se(\beta)} \sim N(0,1)$$

Παραδείγματος χάρη, για τον παράγοντα age το στατιστικό Wald είναι: $\mathrm{w} = \frac{0.0236}{0.0122} = 1.93$

Επομένως η p-value θα είναι: $p = P(\mathrm{Z} > |\,\mathrm{z}\,|) = 2 - 2 \cdot 0.9734 = 0.0532$ οπότε σε επίπεδο σημαντικότητας 5% δεχόμαστε οριακά την μηδενική υπόθεση και ο παράγοντας age δεν είναι στατιστικά σημαντικός. Βέβαια, είναι εμφανές ότι οι μεταβλητές age και smear απορρίφθηκαν οριακά ως στατιστικά μη σημαντικές άρα δεν είναι δεδομένο ότι δεν επιδρούν στην διάρκεια ζωής.

Σε ανάλογα συμπεράσματα για την σημαντικότητα των συμμεταβλητών θα είχαμε κατασκευάζοντας και τα 95% διαστήματα εμπιστοσύνης $[\beta_j \pm 1.96 se(\beta_j)]$ των παραμέτρων του μοντέλου. Τα αποτελέσματα φαίνονται στον πίνακα 4.6

Μεταβλητή	2.5%	97.5%
age	-0.0003701668	0.047523038

smear	-0.0004613779	0.060427484
absinf	-0.0280193658	0.019663638
labindex	-0.0133775151	0.158975110
absblasts	-0.0978979125	-0.001080328
temp	-0.0200620323	0.036563673
treat1	-4.5170394783	-2.214173284

Πίνακας 4.6: 95% διαστήματα εμπιστοσύνης για τις μεταβλητές του μοντέλου

Με βάση τα αποτελέσματα του πίνακα 4.6, μπορούμε να κρίνουμε ποιες μεταβλητές είναι σημαντικές και ποιες όχι. Σε όσες μεταβλητές το δ.ε τους περιέχει την τιμή 0 θεωρούνται μη σημαντικές και αντίστοιχα για όσες μεταβλητές το δ.ε δεν περιέχει το 0 κρίνονται ως στατιστικά σημαντικές. Άρα, πάλι μπορούμε να επιβεβαιώσουμε ότι επιδρούν σημαντικά στη διάρκεια ζωής των ασθενών μονάχα οι μεταβλητές treat1 και absblasts.

Επιπλέον η R μας επιστρέφει τη τιμή του ελέγχου του λόγου πιθανοφανειών για την σύγκριση του μοντέλου που περιέχει και τις 7 συμμεταβλητές και αυτού που δεν περιέχει καμία. Υπενθυμίζουμε ότι το στατιστικό ελέγχου σε αυτόν τον έλεγχο είναι το:

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

όπου $\hat{l}_0$ και $\hat{l}_1$ είναι οι μεγιστοποιημένες λογαριθμοποιημένες συναρτήσεις μερικής πιθανοφάνειας στο μοντέλο χωρίς καμία μεταβλητή και στο μοντέλο που τις περιέχει όλες αντίστοιχα. Οι βαθμοί ελευθερίας d της κατανομής είναι η διαφορά των παραμέτρων ανάμεσα στα δύο μοντέλα, άρα εν προκειμένω d=7.

Οπότε έχουμε ότι: $w = 58.66 \sim X^2_{(7)}$ και επομένως το p-value<0.0001.

Επομένως, απορρίπτουμε την μηδενική υπόθεση $H_0 : \beta_1 = ... = \beta_7 = 0$

4.3.3 Εύρεση βέλτιστου μοντέλου με χρήση βηματικών μεθόδων

Για να καταλήξουμε στο βέλτιστο μοντέλο θα εφαρμόσουμε τον αλγόριθμο της διαδοχικής απαλοιφής (backward elimination) ώστε να καταλήξουμε στο βέλτιστο μοντέλο. Ο όρος «βέλτιστο» έχει να κάνει με την επιλογή του μικρότερου συνόλου μεταβλητών που θα περιγράφουν καλύτερα τα δεδομένα μας.

Ο αλγόριθμος, πριν την εκτέλεση κάποιου βήματος, ξεκινάει με το μοντέλο που περιέχει όλες τις συμμεταβλητές όπως φαίνεται και στον πίνακα 4.7

Συμμεταβλητή που αφαιρείται	**Βαθμοί ελευθερίας**	**AIC**	**Μεταβολή της $-2\hat{l}$**	**p-value**
absinf	1	239.33	0.115	0.73398

temp	1	239.54	0.320	0.57133
καμία	-	241.22	-	-
labindex	1	241.89	2.674	0.10198
smear	1	242.83	3.616	0.05723
age	1	243.03	3.813	0.05085
absblasts	1	243.69	4.470	0.03449
treat1	1	277.62	38.400	5.762e-10

Πίνακας 4.7: Συνοπτικός πίνακας της κατάστασης πριν την εκτέλεση του πρώτου βήματος του αλγορίθμου

Στον πίνακα 4.7, υπολογίζονται οι τιμές που κριτηρίου AIC αν στο επόμενο βήμα αφαιρεθεί η εκάστοτε μεταβλητή, η μεταβολή του στατιστικού ελέγχου του λόγου των πιθανοφανειών και τέλος η p-value αυτού του ελέγχου που συγκρίνει το τρέχον μοντέλο με το μοντέλο που προκύπτει μετά την αφαίρεση της εν λόγω μεταβλητής.

Άρα λοιπόν, η τιμή -2 φορές του λογαρίθμου της πιθανοφάνειας, είναι χαμηλότερη κατά 0.115 στο μοντέλο χωρίς τη μεταβλητή absinf, σε σχέση με αυτό που περιέχει όλες τις συμμεταβλητές. Εξετάζεται δηλαδή η μηδενική υπόθεση Ho: $\beta_{absinf} = 0$ χρησιμοποιώντας την ελεγχοσυνάρτηση: $z = -2(\hat{l}_0 - \hat{l}_1) \sim \mathrm{X}^2_{(1)} \Rightarrow z = 0.115 \sim \mathrm{X}^2_{(1)}$

Άρα η p-value αυτού του ελέγχου θα είναι:

$$p - value = P[X^2 > z \mid X^2 \sim \mathrm{X}^2_{(1)}] = 1 - 0.265 = 0.735$$

Η p-value προκύπτει να είναι πολύ μεγάλη άρα σε επίπεδο σημαντικότητας 5% δεχόμαστε την μηδενική υπόθεση $\beta_{absinf} = 0$.

Σύμφωνα μάλιστα με τον πίνακα 4.7, η μεταβλητή absinf είναι και η λιγότερη σημαντική μεταβλητή διότι επιφέρει την μικρότερη μεταβολή της ελεγχοσυνάρτησης του λόγου των πιθανοφανειών. Άρα κατά το πρώτο βήμα του αλγορίθμου αφαιρείται η συμμεταβλητή absinf.

Μετά την αφαίρεση της absinf, η R αναπροσαρμόζει αυτόματα το μοντέλο του Cox με τις υπόλοιπες 6 μεταβλητές όπως φαίνεται στον πίνακα 4.8

Συμμεταβλητή που αφαιρείται	Βαθμοί ελευθερίας	AIC	Μεταβολή της $-2\hat{l}$	p-value
temp	1	237.62	0.290	0.59041
καμία	-	239.33	-	-
labindex	1	240.15	2.814	0.09344
age	1	241.04	3.711	0.05405
absblasts	1	241.71	4.378	0.03641
smear	1	243.67	6.338	0.01182
treat1	1	277.37	40.033	2.50e-07

Πίνακας 4.8: Συνοπτικός πίνακας της κατάστασης πριν την εκτέλεση του δεύτερου βήματος του αλγορίθμου

Από τα αποτελέσματα του πίνακα 4.8, και δουλεύοντας ανάλογα με πριν, η συμμεταβλητή που πρέπει να αφαιρεθεί από το μοντέλο καθώς είναι η λιγότερο στατιστικά σημαντική είναι η temp. Η p-value τιμή της για τον έλεγχο της υπόθεσης Ho: $\beta_{temp} = 0$ προκύπτει να είναι 0.59.

Άρα σε επίπεδο σημαντικότητας 5% δεχόμαστε την Ho και η συμμεταβλητή temp αφαιρείται από το μοντέλο.

Άρα η R αναπροσαρμόζει αυτόματα το μοντέλο με τις εναπομένουσες 5 συμμεταβλητές και τα αποτελέσματα που μας δίνει φαίνονται στον πίνακα 4.9.

Συμμεταβλητή που αφαιρείται	Βαθμοί ελευθερίας	AIC	Μεταβολή της $-2\hat{l}$	p-value
καμία	-	237.62	-	-
labindex	1	239.05	3.425	0.06423
age	1	239.58	3.958	0.04666
absblasts	1	240.15	4.527	0.03336
smear	1	241.81	6.185	0.01289
treat1	1	277.93	42.308	7,80e-11

Πίνακας 4.9: Συνοπτικός πίνακας αποτελεσμάτων μετά την εκτέλεση του δεύτερου βήματος του αλγορίθμου

Σε αυτό το σημείο ο αλγόριθμος σταματάει. Παρότι φαίνεται ότι η p-value για τον έλεγχο της υπόθεσης $\{H_0 : \beta_{labindex} = 0\}$ παίρνει οριακά μια τιμή πάνω από 0.05, (συγκεκριμένα 0.0642) η αντίστοιχη συμμεταβλητή labindex, η οποία φαίνεται να είναι η λιγότερο στατιστικά σημαντική, δεν απαλείφεται από το μοντέλο. Αυτό συμβαίνει διότι αν αφαιρεθεί, το νέο μοντέλο που θα προκύψει θα έχει τιμή του κριτηρίου AIC μεγαλύτερη από το ισχύον μοντέλο. Όπως ξέρουμε, το κριτήριο AIC μας δίνει ένα μέτρο σύγκρισης της καταλληλότητας μοντέλων με διαφορετικό πλήθος μεταβλητών. Μεταξύ υποψήφιων μοντέλων διαλέγουμε εκείνο με την μικρότερη τιμή του κριτηρίου AIC. Επομένως, ο αλγόριθμος σταματάει διότι οποιαδήποτε επιπλέον συμμεταβλητή κι αν αφαιρεθεί, δεν καταλήγουμε σε ένα «καλύτερο» μοντέλο.

Έχοντας καταλήξει στο βέλτιστο μοντέλο που θα περιέχει τις μεταβλητές labindex, age, absblasts, smear και treat1 μπορούμε να πάρουμε από την R ένα συνοπτικό πίνακα των αποτελεσμάτων, όπως φαίνεται στον πίνακα 4.10

Συμμεταβλητή	β	$se(\beta)$	Έλεγχος Wald	p-value
age	0.02365	0.01196	1.978	0.0479
smear	0.02554	0.01053	2.425	0.0153
labindex	0.08019	0.04242	1.890	0.0587
absblasts	-0.04205	0.02117	-1.986	0.0471
treat1	-3.44197	0.57888	-5.946	2.75e-09

Πίνακας 4.10: Συγκεντρωτικός πίνακας αποτελεσμάτων για το βέλτιστο μοντέλο

Επομένως το τελικό μοντέλο λαμβάνει την μορφή:

$$\mathrm{h}(t;\underline{x}) = h_0(t)\cdot\exp(0.02365age + 0.02554smear + 0.08019labindex - 0.04205absblasts - 3.44197treat_1)$$

4.3.4 Ερμηνεία των συντελεστών του βέλτιστου μοντέλου

Αφού έχει γίνει η προσαρμογή του μοντέλου του Cox και έχουμε καταλήξει και στη μορφή του βέλτιστου μοντέλου θα πρέπει να μπορούμε να αξιολογήσουμε τα τελικά αποτελέσματα ώστε να ερμηνεύσουμε σωστά την επίδραση κάθε συμμεταβλητής στη διάρκεια ζωής των ασθενών. Στον παρακάτω πίνακα 4.11 παρουσιάζονται συγκεντρωτικά κάποια σημαντικά αποτελέσματα από το βέλτιστο μοντέλο στο οποίο καταλήξαμε και θα μας βοηθήσει στην ερμηνεία των συντελεστών.

Συμμεταβλητή	β	$\exp(\beta)$	95% δ.ε του β		95% δ.ε του $\exp(\beta)$	
			2.5%	97.5%	2.5%	97.5%
Age	0.02365	1.0239	0.0002134603	0.0470957205	1.00021348	1.04822234
Smear	0.02554	1.0259	0.0049002691	0.0461832086	1.00491230	1.04726626
Labindex	0.08019	1.0835	-0.0029491814	0.1633237945	0.99705516	1.17741787
Absblasts	-0.04205	0.9588	-0.0835499322	-0.0005473484	0.91984516	0.99945280
Treat1	-3.44197	0.0320	-4.5765611853	-2.3073761427	0.01029022	0.09952204

Πίνακας 4.11: Συνοπτικός πίνακας αποτελεσμάτων για την επίδραση κάθε συμμεταβλητής του βέλτιστου μοντέλου στη διάρκεια ζωής

Για το μοντέλο που καταλήξαμε , ο συντελεστής της μεταβλητής age ισούται με 0.02365, άρα η συνάρτηση διακινδύνευσης επηρεάζεται κατά $h_0(t)\cdot\exp(0.02365)=1.0239\cdot h_0(t)$ αν η ηλικία του ασθενή αυξηθεί κατά έναν χρόνο και το ποσοστό επίστρωσης βλαστοκυττάρων, το ποσοστό κυττάρων που προήλθαν από μυελό των οστών, ο απόλυτός αριθμός βλαστοκυττάρων και η ανταπόκριση του ασθενή στη θεραπεία παραμείνουν σταθερά.

Ο συντελεστής της μεταβλητής smear ισούται με 0.02554 άρα η συνάρτηση διακινδύνευσης επηρεάζεται κατά $h_0(t)\cdot\exp(0.02554)=1.0259\cdot h_0(t)$ αν το ποσοστό επίστρωσης βλαστοκυττάρων αυξηθεί κατά 1% και η ηλικία του ασθενή, το ποσοστό κυττάρων που προήλθαν από μυελό των οστών, ο απόλυτος αριθμός βλαστοκυττάρων και η ανταπόκριση του ασθενή στη θεραπεία παραμείνουν σταθερά.

Ο συντελεστής της μεταβλητής labindex ισούται με 0.08019 άρα η συνάρτηση διακινδύνευσης επηρεάζεται κατά $h_0(t)\cdot\exp(0.08019)=1.0835\cdot h_0(t)$ αν το ποσοστό των κυττάρων που προήλθαν από μυελό των οστών αυξηθεί κατά 1% και η ηλικία του ασθενή, το ποσοστό επίστρωσης βλαστοκυττάρων, ο απόλυτος αριθμός βλαστοκυττάρων και η ανταπόκριση του ασθενή στη θεραπεία παραμείνουν σταθερά.

Ο συντελεστής της μεταβλητής absblasts ισούται με -0.04205 άρα η συνάρτηση διακινδύνευσης επηρεάζεται κατά $h_0(t)\cdot\exp(-0.04205)=0.9588\cdot h_0(t)$ αν ο απόλυτος αριθμός βλαστοκυττάρων αυξηθεί κατά 1000 και η ηλικία του ασθενή, το ποσοστό επίστρωσης βλαστοκυττάρων, το ποσοστό των κυττάρων που προήλθε από μυελό των οστών και η ανταπόκριση του ασθενή στη θεραπεία παραμείνουν σταθερά.

Τέλος ο συντελεστής της μεταβλητής treat1 ισούται με -3.44197 άρα η συνάρτηση διακινδύνευσης επηρεάζεται κατά $h_0(t)\cdot\exp(-3.44197)=0.032\cdot h_0(t)$ αν ο ασθενής περάσει από την κατηγορία treat=0 στην κατηγορία treat=1, δηλαδή αν αρχίσει να ανταποκρίνεται στη θεραπεία και η ηλικία του, το ποσοστό επίστρωσης βλαστοκυττάρων, το ποσοστό των κυττάρων που προήλθε από μυελό των οστών και ο απόλυτος αριθμός βλαστοκυττάρων παραμείνουν σταθερά.

Γνωρίζοντας ότι το μοντέλο του Cox εκφράζεται από τη σχέση:

$$h(t;\underset{\sim}{x}) = h_o(t)\exp(\underset{\rightarrow}{\beta}^{\mathrm{T}}\underset{\sim}{x})$$

συμπεραίνουμε ότι οι τιμές του $\exp(\underset{\rightarrow}{\beta}^{\mathrm{T}})$ φανερώνουν κατά πόσο πολλαπλασιάζεται η κοινή συνάρτηση κινδύνου $h_0(t)$. Πιο συγκεκριμένα, η τιμή $\exp(\beta_i)$ δείχνει κατά πόσο η i-οστή συμμεταβλητή του μοντέλου επιδρά στη διάρκεια ζωής, όταν όλες οι υπόλοιπες συμμεταβλητές παραμένουν σταθερές. Προφανώς $\exp(\beta_i) > 0, \forall \mathrm{i}$ όμως μπορούμε να κάνουμε την εξής κατηγοριοποίηση:

Α) Αν $\exp(\beta_i) > 1$ τότε η i-οστή συμμεταβλητή επιδρά αρνητικά στη διάρκεια ζωής καθώς η κοινή συνάρτηση κινδύνου για όλες τις μονάδες $h_o(t)$ πολλαπλασιάζεται με κάτι μεγαλύτερο της μονάδας άρα η συνάρτηση διακινδύνευσης μεγαλώνει.

Β) Αν $\exp(\beta_i) < 1$ τότε με την ίδια λογική η i-οστή μεταβλητή του μοντέλου επιδρά θετικά στη διάρκεια ζωής, διότι συρρικνώνεται η κοινή συνάρτηση διακινδύνευσης.

Γ) Αν $\exp(\beta_i) \approx 1$ τότε η i-οστή συμμεταβλητή ουσιαστικά δεν επηρεάζει τη διάρκεια ζωής, καθώς η κοινή συνάρτηση διακινδύνευσης παραμένει αμετάβλητη.

Κάνοντας λοιπόν αυτές τις παρατηρήσεις, μπορούμε να συμπεράνουμε ότι:

1) Οι μεταβλητές age, smear, labindex επιδρούν αρνητικά στη διάρκεια ζωής των ασθενών αλλά όχι και τόσο σημαντικά καθώς όλες είχανε μια τιμή $\exp(\beta_i)$ πολύ κοντά στο 1.
2) Η μεταβλητή absblasts επιδρά θετικά στη διάρκεια ζωής των ασθενών καθώς $\exp(\beta_{absblasts}) = 0.9588 < 1$, αλλά και αυτή είναι μια τιμή κοντά στη μονάδα άρα δεν την επηρεάζει καταλυτικά.
3) Η πλέον στατιστικά σημαντική μεταβλητή, όπως αναμενόταν άλλωστε, είναι η μεταβλητή treat. Παρατηρούμε λοιπόν ότι $\exp(\beta_{treat}) = 0.032 << 1$ άρα όταν η κατηγορική μεταβλητή treat=1, δηλαδή ο ασθενής ανταποκριθεί στη θεραπεία, αυτό επιδρά πολύ θετικά στην υγεία του (κάτι που επιβεβαιώσαμε και στους μη παραμετρικούς ελέγχους που προηγήθηκαν).

4.3.5 Γραφικός έλεγχος των υπολοίπων Schoenfeld

Όπως είδαμε στην παράγραφο 2.2.4.1 μπορούμε να ελέγξουμε την υπόθεση της αναλογικότητας της διακινδύνευσης μέσω των κλιμακοποιημένων υπολοίπων Schoenfeld. r_{ij}^{*} Αν $\beta_i(\mathrm{t})$ ο συντελεστής της συμμεταβλητής i κατά τη χρονική στιγμή t τότε η υπόθεση της αναλογικής διακινδύνευσης μπορεί να εκφραστεί ισοδύναμα και ως: $\beta_i(\mathrm{t}) = \beta_i, \forall t$. Ξέρουμε επίσης ότι: $E(r_{ij}^{*}) \approx \beta_i(\mathrm{t}_{(j)}) - \beta_i$, όπου $\beta_i(\mathrm{t}_{(j)})$ είναι ο συντελεστής της συμμεταβλητής i τη χρονική στιγμή $\mathrm{t}_{(j)}$. Επομένως, για να γίνει αποδεκτή η υπόθεση $\beta_i(\mathrm{t}) = \beta_i, \forall t$ θα πρέπει η γραφική παράσταση $(r_{ij}^{*} + \beta_i)$ ως προς $\mathrm{t}_{(j)}$ να δείχνει μια οριζόντια γραμμή.

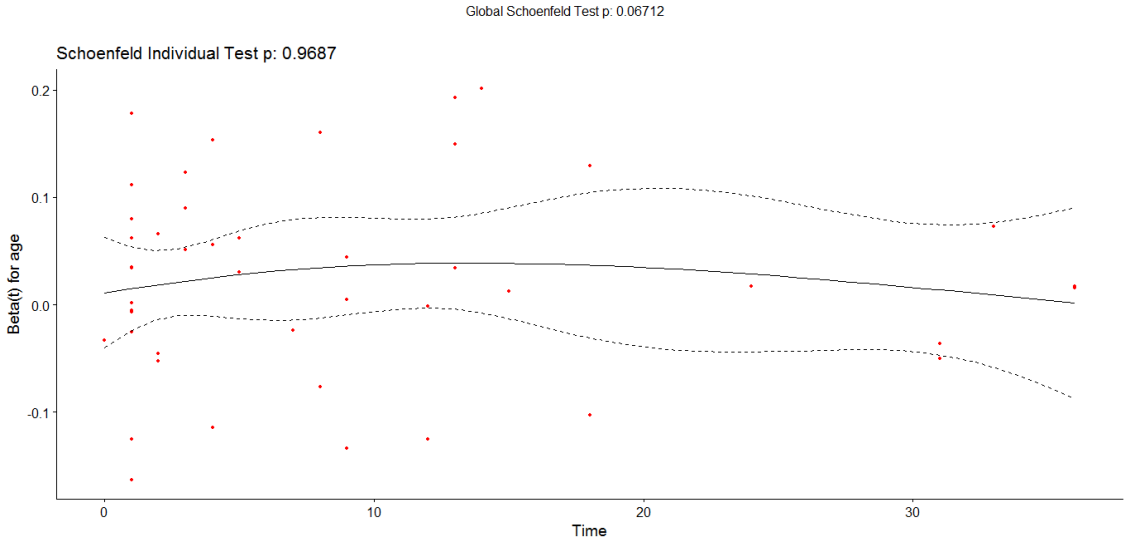
Μπορούμε λοιπόν στην R να κατασκευάσουμε έναν πίνακα ελέγχου της ισχύς υπόθεσης της αναλογικής διακινδύνευσης για κάθε συμμεταβλητή. Τα αποτελέσματα φαίνονται στον πίνακα 4.12

Συμμεταβλητή	rho	chisq	p-value
Age	-0.00523	0.00154	0.96869
Smear	-0.35674	7.30772	0.00687
Labindex	0.07367	0.31407	0.57520
Absblasts	0.26546	3.64327	0.05630
treat1	0.20882	1.84531	0.17433
GLOBAL	NA	10.30202	0.06712

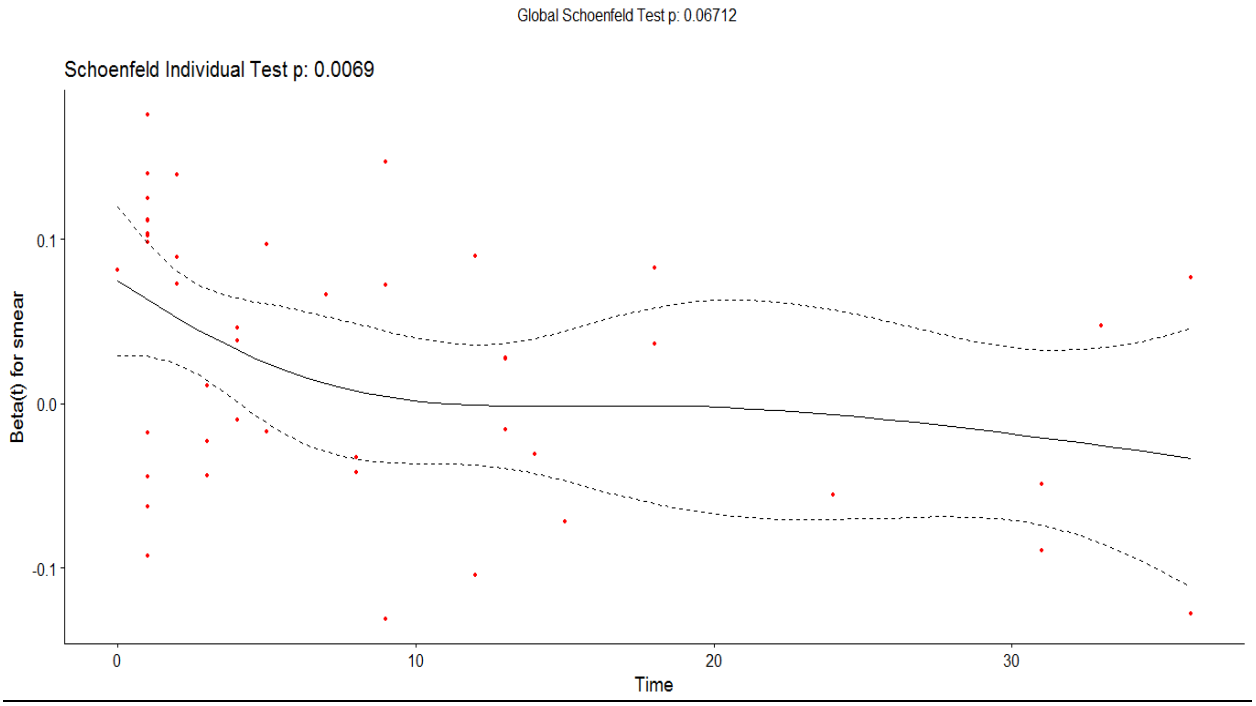
Πίνακας 4.12: Συνοπτικός πίνακας ελέγχου της υπόθεσης της αναλογικής διακινδύνευσης για το βέλτιστο μοντέλο με τον μετασχηματισμό "identity".

Στον πίνακα 4.12 η στήλη rho εκφράζει τον συντελεστή συσχέτισης ανάμεσα στα κλιμακοποιημένα υπόλοιπα Schoenfeld και τους μετασχηματισμένους χρόνους επιβίωσης για κάθε μεταβλητή χωριστά. Εμείς εδώ θεωρήσαμε ότι δεν παίρνουμε κάποια συνάρτηση του χρόνου αλλά τους ταυτοτικά τους χρόνους αποκοπής $t_{(j)}$ (transform='identity'). Επιπλέον, εκτελείται ένας X^2 έλεγχος και έχουμε σε ξεχωριστή στήλη τα p-values αυτών των ελέγχων. Τέλος, μας επιστρέφεται η τιμή του ελέγχου Global, όπως αυτός αναλύθηκε στην παράγραφο 2.2.4.2. Παρατηρούμε λοιπόν ότι η συμμεταβλητή smear είναι έντονα χρονικά εξαρτημένη ενώ επίσης οριακά μπορούμε να αποδεχτούμε την υπόθεση της χρονικής ανεξαρτησίας για τη συμμεταβλητή absblasts. Τέλος, ο έλεγχος Global μας επιστρέφει μια τιμή του στατιστικού ελέγχου $T = 10.3 \sim X_5^2$ και άρα υπολογίζεται η p-value=0.067 που είναι οριακά άνω του επιπέδου σημαντικότητας 5%.
Συμπερασματικά, μπορούμε να πούμε ότι εν γένει ισχύει η υπόθεση της αναλογικής διακινδύνευσης αλλά φαίνεται ότι τουλάχιστον μια συμμεταβλητή που μετέχει στο μοντέλο να είναι έντονα χρονοεξαρτημένη.

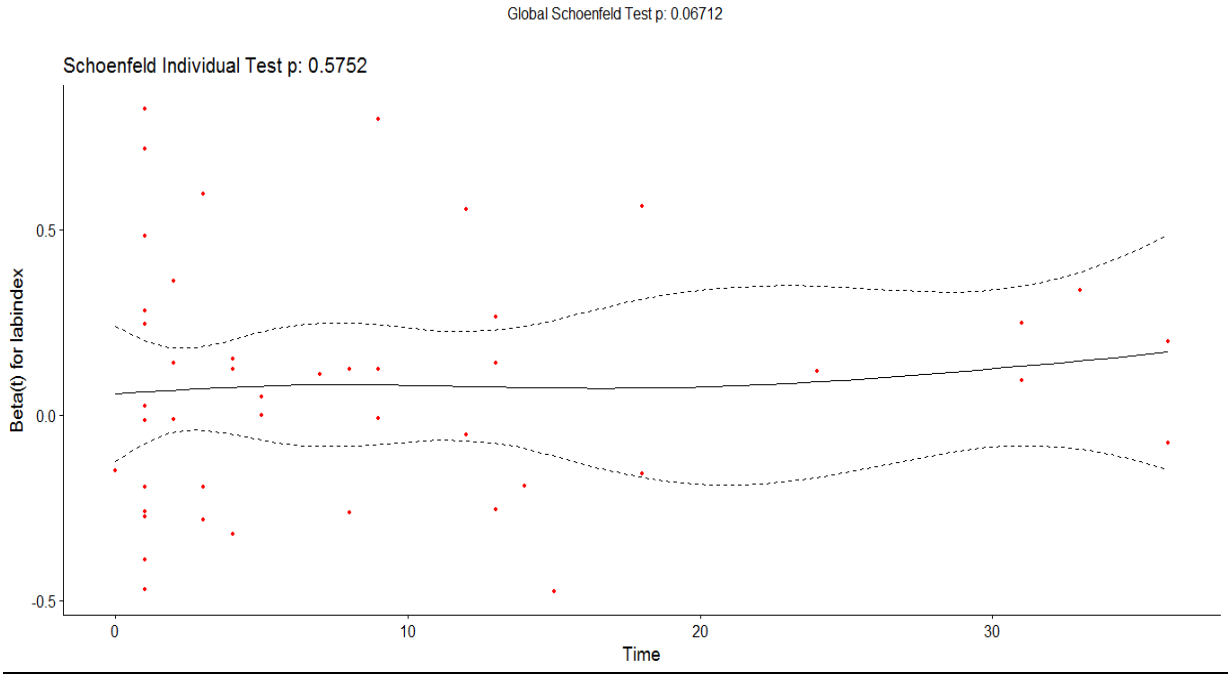
Τα γραφήματα $(r_{ij}^* + \beta_i)$ ως προς $t_{(j)}$ για κάθε συμμεταβλητή του μοντέλου που κατασκευάσαμε στην R είναι τα ακόλουθα:



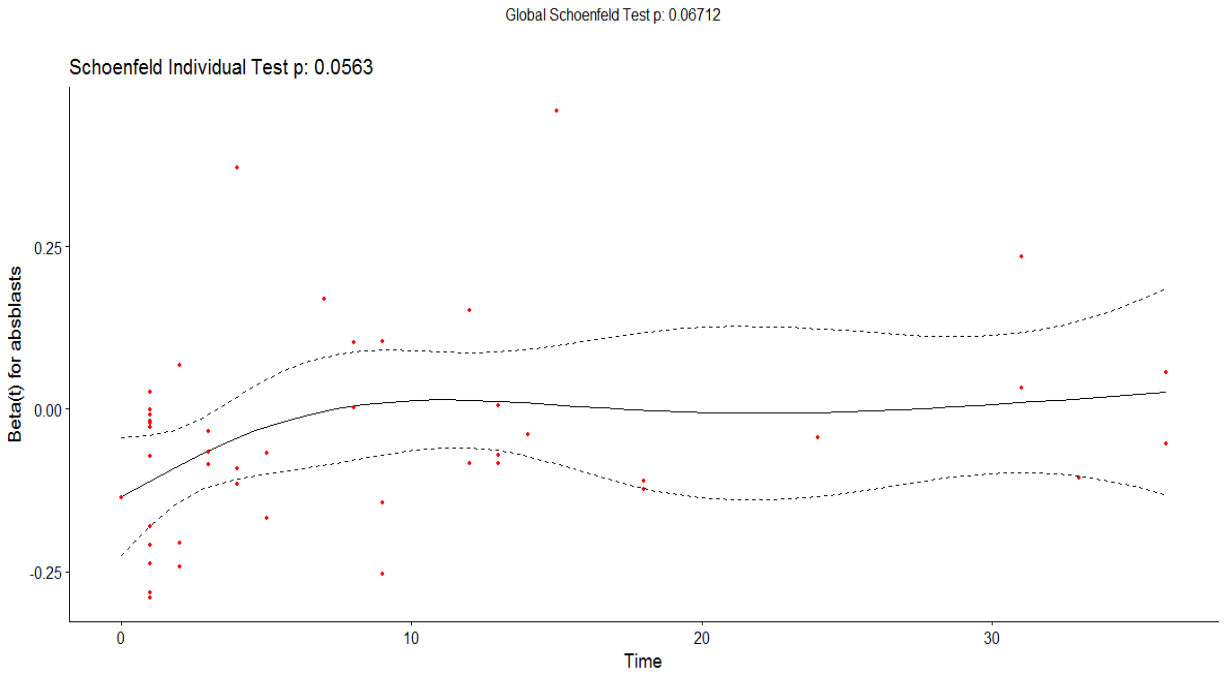
Γράφημα 4.13: Έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης για την μεταβλητή age



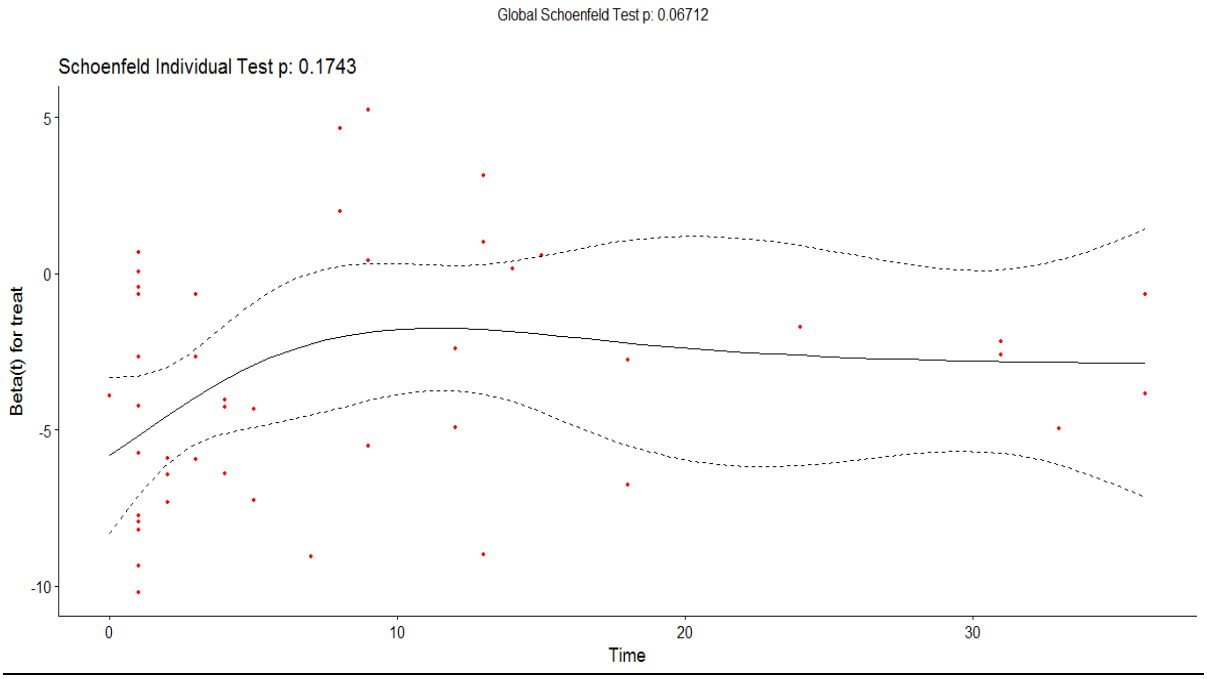
Γράφημα 4.14: Έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης για τη μεταβλητή smear



Γράφημα 4.15: Έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης για τη μεταβλητή labindex



Γράφημα 4.16: Έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης για τη μεταβλητή absblasts



Γράφημα 4.17: Έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης για τη μεταβλητή treat

Στα γραφήματα 4.13-4.17, η συμπαγής γραμμή στα γραφήματα είναι αποτέλεσμα της λεγόμενης εξομάλυνσης μέσω συναρτήσεων splines. Οι συναρτήσεις splines, αποτελούνται

από δύο ή περισσότερες συνεχόμενες, συνήθως τρίτου βαθμού (cubic) καμπύλες ή τόξα (arcs) τα οποία ενώνονται μεταξύ τους με ομαλό τρόπο. Οι διακεκομμένες γραμμές σχηματίζουν μια περιοχή $\pm 2-$ τυπικών αποκλίσεων από την εξομαλυμένη καμπύλη.

Σύμφωνα λοιπόν με τα γραφικά αποτελέσματα, μπορούμε να πούμε ότι φαίνεται να ισχύει η υπόθεση της ανεξαρτησίας από το χρόνο για τις μεταβλητές age και labindex καθώς η ευθεία που σχηματίζεται είναι σχεδόν οριζόντια. Επιπλέον, για τις μεταβλητές absblasts και treat παρατηρούμε ότι έχουμε παρόμοια συμπεριφορά αν και πρέπει να είμαστε πιο επιφυλακτικοί διότι για μικρές τιμές του χρόνου παρατηρείται μια καμπυλότητα στα γραφήματα αυτών των μεταβλητών. Τέλος, για τη μεταβλητή smear μπορούμε να πούμε ότι έχουμε ισχυρές ενδείξεις ενάντια στην υπόθεση της αναλογικής διακινδύνευσης καθώς παρατηρείται μεγάλη καμπυλότητα για χρόνους μικρότερους του t=10 μήνες.

Αυτά τα γραφικά αποτελέσματα συμφωνούν με τα αποτελέσματα του πίνακα 4.12 που σχολιάσαμε πρωτύτερα αλλά και των μη παραμετρικών γραφικών ελέγχων που διεξήγαμε στην παράγραφο 4.3.1.Δηλαδή, μπορούμε να δεχτούμε με ασφάλεια την υπόθεση της αναλογικότητας της διακινδύνευσης για τις συμμεταβλητές age και labindex, πιο οριακά για τις treat και abasblasts ενώ απορρίπτεται για την συμμεταβλητή smear. Βέβαια, σε αυτό το σημείο θα πρέπει να αναφέρουμε ότι, εφόσον το μέγεθος του δείγματος των ασθενών δεν είναι ιδιαίτερα μεγάλο (51 ασθενείς), ο γραφικός έλεγχος των κλιμακοποιημένων υπολοίπων Schoenfeld ενδέχεται να επηρεάζεται από ακραίες παρατηρήσεις.

Επειδή, όμως, ο έλεγχος Global έπαιρνε μια p-value=0.067 για την υπόθεση $\beta_i(\mathrm{t}) = \beta_i, \forall t$ που είναι οριακά άνω του 0.05 και τα γραφικά μας αποτελέσματα ήταν διφορούμενα θα διενεργήσουμε τον ίδιο έλεγχο αλλά αυτή τη φορά δεχόμενοι τον προκαθορισμένο από τη R μετασχηματισμό των χρόνων διακοπής, που είναι ο μετασχηματισμός “km” (Kaplan-Meier). Ο αναλυτικός πίνακας 4.13 των αποτελεσμάτων είναι ο ακόλουθος:

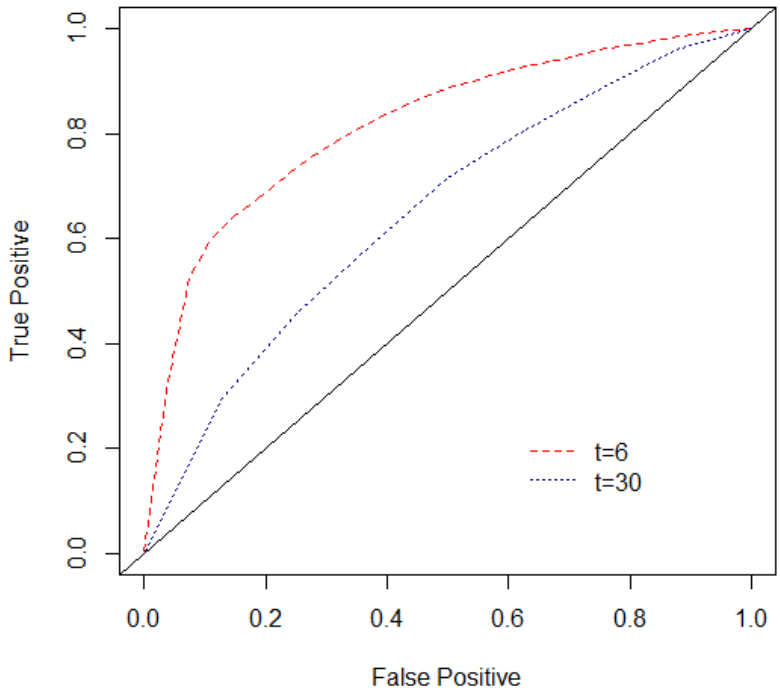
Συμμεταβλητή	rho	chisq	p-value
Age	0.0697	0.273	0.60102
Smear	-0.3910	8.780	0.00305
labindex	0.0438	0.111	0.73886
absblasts	0.3371	5.876	0.01534
treat1	0.2831	3.391	0.06554
GLOBAL	NA	13.654	0.01796

Πίνακας 4.13: Συνοπτικός πίνακας ελέγχου της υπόθεσης της αναλογικής διακινδύνευσης για το βέλτιστο μοντέλο με τον μετασχηματισμό “km”.

Τα αποτελέσματα του πίνακα 4.13 είναι πλέον ξεκάθαρα. Οι συμμεταβλητές age και labindex είναι χρονοανεξάρτητες καθώς λαμβάνουν μια πολύ μεγάλη p-value για τον έλεγχο της υπόθεσης $\beta_i(\mathrm{t}) = \beta_i, \forall t$ που είναι ισοδύναμη της υπόθεσης της αναλογικής διακινδύνευσης. Επίσης, μπορούμε να δεχτούμε οριακά και για τη συμμεταβλητή treat την ανεξαρτησία ως προς τον χρόνο. Τέλος, οι συμμεταβλητές absblasts και smear είναι φανερό ότι έχουν έντονη εξάρτηση ως προς τον χρόνο ενώ το τεστ Global μας δίνει p-value πολύ μικρή άρα απορρίπτουμε την υπόθεση της αναλογικότητας της διακινδύνευσης καθολικά για το μοντέλο του Cox.

4.3.6 Προβλεπτική ικανότητα μοντέλου

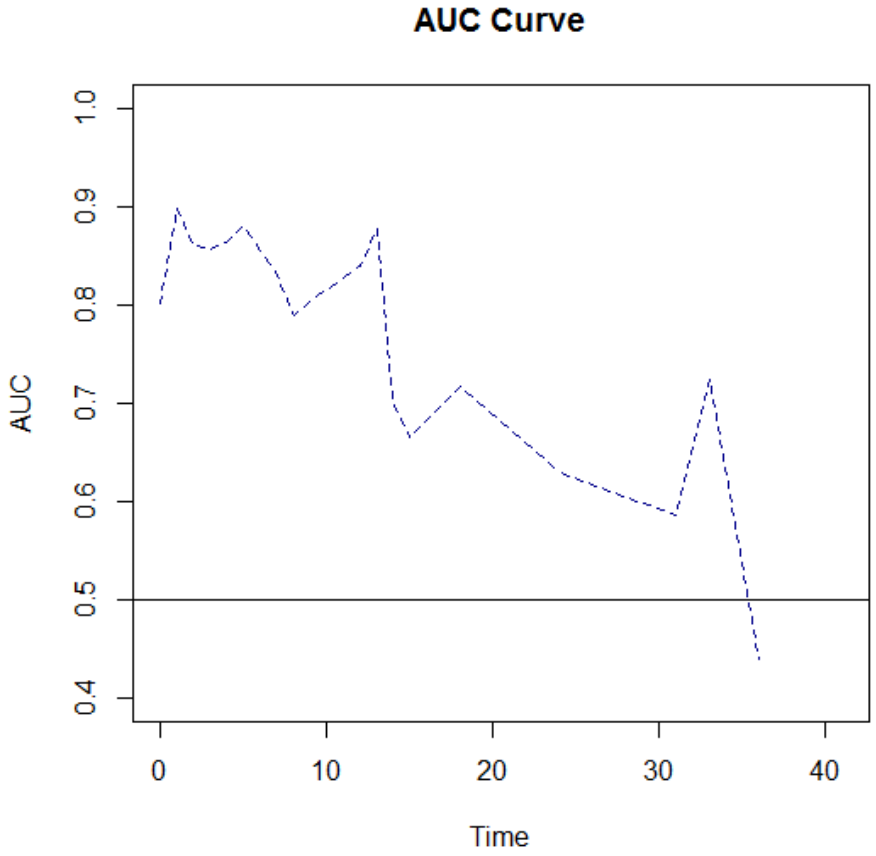
Όπως αναφέραμε και στην παράγραφο 2.1.5.1, ένας αποτελεσματικός τρόπος γραφικού ελέγχου της προβλεπτικής ικανότητας του μοντέλου είναι ο σχεδιασμός της καμπύλης ROC. Στο σημείο αυτό, εξετάζουμε πόσο καλά το βέλτιστο μοντέλο στο οποίο καταλήξαμε στην προηγούμενο παράγραφο μπορεί να προσαρμόζεται σε νέα δεδομένα. Για αυτό το σκοπό, σχεδιάζονται 2 καμπύλες ROC για 2 διαφορετικές χρονικές στιγμές $t_1 = 6,\ t_2 = 30$. Επιλέξαμε 2 χρονικές στιγμές που απέχουν πολύ μεταξύ τους ώστε να έχομε μια πρώτη εικόνα πως μεταβάλλεται η προβλεπτική ικανότητα του μοντέλου από μικρούς χρόνους σε μεγάλους. Με τη βοήθεια της R εγκαθιστούμε το πακέτο risksetROC και τα γραφικά αποτελέσματα φαίνονται παρακάτω.



Γράφημα 4.18: Καμπύλη ROC του τελικού μοντέλου για t=6 και t=30 μήνες

Παρατηρούμε ότι για t=6 μήνες η προβλεπτική ικανότητα του μοντέλου είναι ιδιαιτέρως ισχυρή καθώς το εμβαδόν κάτω από την καμπύλη προκύπτει να είναι AUC(6)=0.8164155 που είναι πολύ κοντά στη μέγιστη θεωρητική τιμή 1. Αντιθέτως, είναι εμφανές ότι για μεγάλες τιμές του χρόνου η προβλεπτική ικανότητα του μοντέλου μειώνεται αισθητά καθώς για t=30 το AUC(30)= 0.6498827. Αυτό ήταν κάτι το οποίο αναμέναμε καθώς το μέγεθος του δείγματος των ασθενών είναι σχετικά μικρό (51 ασθενείς) και δεν είχαμε στη διάθεση μας πολλές τιμές του χρόνου διάρκειας ζωής κοντά στην τιμή t=30. Είναι λογικό λοιπόν, να μην μπορούν να προσαρμοστούν ικανοποιητικά νέα δεδομένα για μεγάλες τιμές του χρόνου t.

Για να έχουμε μια πιο ξεκάθαρη εικόνα για την εξέλιξη της προβλεπτικής ικανότητας του μοντέλου κάνουμε το γράφημα του εμβαδού κάτω από την καμπύλη ROC συναρτήσει του χρόνου. Τα γραφικά αποτελέσματα φαίνονται παρακάτω:



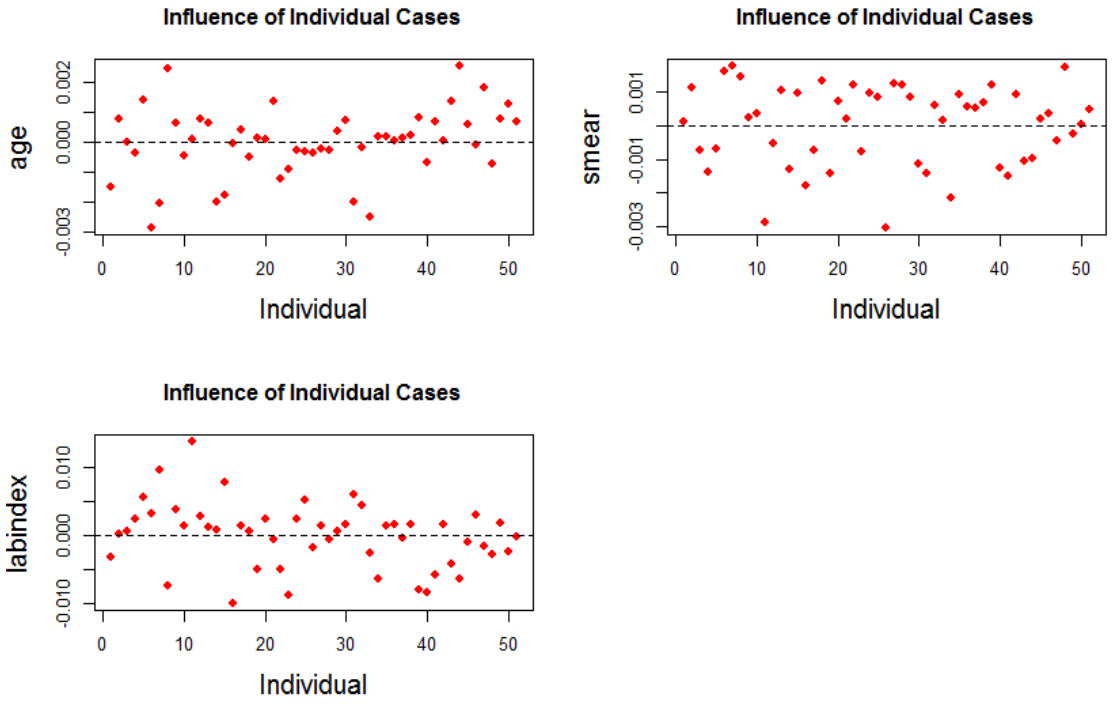
Γράφημα 4.19: Καμπύλη AUC σε σχέση με το χρόνο

Είναι εμφανές, από το γράφημα 4.19, ότι με το πέρασμα του χρόνου το AUC φθίνει, αν και δεν έχει εξ ολοκλήρου καθοδική πορεία και παρατηρούνται αρκετά τοπικά μέγιστα και ελάχιστα. Γενικά, όσες περισσότερες παρατηρήσεις έχουμε λοιπόν σε ένα διάστημα του χρόνου (τ_1, τ_2) τόσο αυξάνεται η τιμή του AUC σε αυτό το διάστημα και ισοδύναμα τόσο καλύτερη γίνεται η προβλεπτική ικανότητα του μοντέλου μας. Για αυτό το λόγο, η τιμή του AUC φτάνει στην ελάχιστη θεωρητική τιμή της 0.5 για t>35 μήνες καθώς μόλις 4 από τους 51 χρόνους ζωής που διαθέτουμε ξεπέρασαν τους 35 μήνες. Άρα, για πολύ μεγάλους χρόνους, άνω των 35 μηνών, το μοντέλο μας είναι παντελώς αναξιόπιστο ως προς την προβλεπτική του ικανότητα.

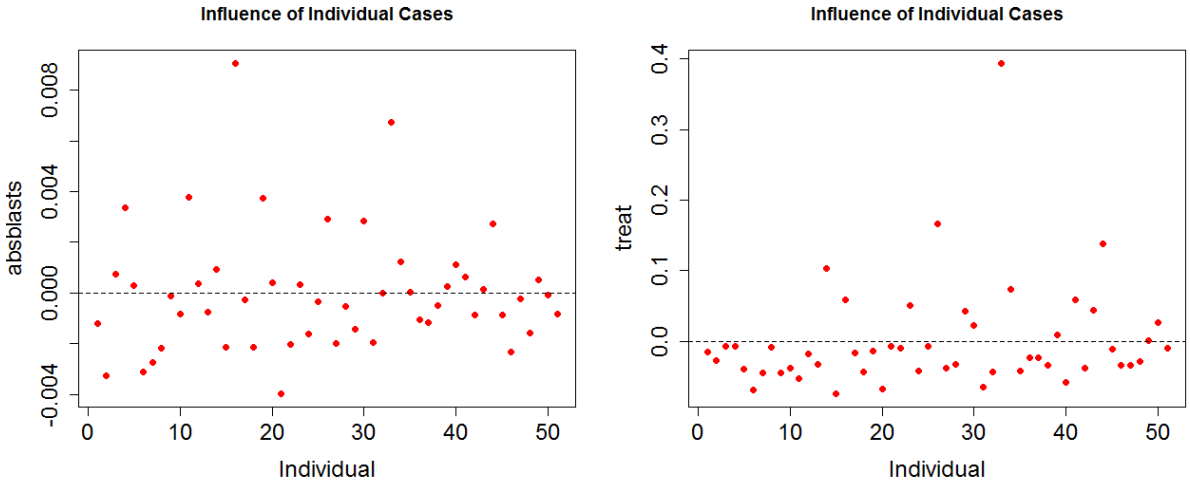
4.3.7 Σημεία επιρροής

Όπως αναλύσαμε και στην παράγραφο 2.2.5, σημεία επιρροής είναι εκείνες οι παρατηρήσεις του συνόλου δεδομένων μας, οι οποίες όταν αφαιρεθούν από το μοντέλο και γίνει αναπροσαρμογή του χωρίς αυτές τότε έχουμε σημαντική διαφοροποίηση των αποτελεσμάτων. Έχοντας καταλήξει στο βέλτιστο μοντέλο του Cox, που περιέχει τις πλέον σημαντικές μεταβλητές age,smear, absblasts, labindex και treat, με τη βοήθεια της R θα υπολογίσουμε τις ποσότητες DFBETAS και επιπλέον θα γίνει γραφική αναπαράσταση τους ώστε να έχουμε μια εποπτική εικόνα για το ποιες παρατηρήσεις ενδέχεται να επηρεάζουν σημαντικά την εκτίμηση των παραμέτρων του μοντέλου.

Τα αποτελέσματα που πήραμε από την R φαίνονται στα γραφήματα 4.20 και 4.21



Γράφημα 4.20: Έλεγχος για τον εντοπισμό σημείων επιρροής για τις σημαντικές συμμεταβλητές age, smear και labindex του βέλτιστου μοντέλου



Γράφημα 4.21: Έλεγχος για τον εντοπισμό σημείων επιρροής για τις σημαντικές συμμεταβλητές absblasts και treat του βέλτιστου μοντέλου

Από τα γραφήματα 4.20 και 4.21 παρατηρούμε ότι υπάρχουν κάποια σημεία που απέχουν αρκετά από το μηδέν για καθεμία από τις 5 συμμεταβλητές του βέλτιστου μοντέλου. Βέβαια, το πλήθος αυτών των σημείων δεν είναι μεγάλο συγκριτικά με το σύνολο των παρατηρήσεων και ούτε τα DFBETAS φαίνεται να λαμβάνουν πολύ μεγάλες τιμές εκτός από μια παρατήρηση στο γράφημα για τη συμμεταβλητή treat. Για να είμαστε σίγουροι όμως θα χρησιμοποιήσουμε το κριτήριο που είχαμε αναφέρει και στην παράγραφο 2.2.5, δηλαδή θα θεωρούμε ότι ένα σημείο είναι σημείο επιρροής αν: $\left|DFBETAS_{ji}\right| > 2/\sqrt{n} \Rightarrow \left|DFBETAS_{ji}\right| > 0.28$

Άρα μπορούμε να βρούμε στην R ότι η μόνη παρατήρηση που λαμβάνει τιμή DFBETAS μεγαλύτερη του 0.28 είναι η 33η παρατήρηση για τη συμμεταβλητή treat, όπου έχουμε $DFBETAS_{5,33}$ =0.393573004.

Είναι ακριβώς εκείνο το σημείο το οποίο και γραφικά παρατηρήσαμε ότι ξεφεύγει πολύ από το κεντρική γραμμή. Εν γένει ένα σημείο επιρροής μπορεί να εμφανιστεί ακόμα και λόγω σφαλμάτων στις μετρήσεις.

Όμως στην περίπτωσή μας δεν συμβαίνει κάτι τέτοιο. Η συυμεταβλητή για την οποία εμφανίζεται ένα σημείο επιρροής είναι η treat η οποία είναι κατηγορική μεταβλητή και εκφράζει την ανταπόκριση ή όχι στη θεραπεία. Λόγω λοιπόν της φύσης της μεταβλητής treat δεν μπορεί το σημείο επιρροής να οφείλεται σε λανθασμένη μέτρηση. Μπορούμε να διαπιστώσουμε με τη βοήθεια της R ότι ο 33ος ασθενής είναι ο ασθενής ο οποίος είχε την μεγαλύτερη διάρκεια ζωής από όσους δεν ανταποκρίθηκαν στη θεραπεία (treat=0). Πιο συγκεκριμένα, ο 33ος ασθενής είχε χρόνο ζωής 13 μήνες όταν ο μέσος όρος ζωής των ασθενών που δεν ανταποκρίθηκαν στη θεραπεία είναι 1,5 μήνας. Άρα, συμπεραίνουμε ότι υπάρχει μια παρατήρηση στο σύνολο δεδομένων μας που επηρεάζει την εκτίμηση της παραμέτρου για τη συμμεταβλητή treat. Αν εξαιρούσαμε τη συγκεκριμένη παρατήρηση και αναπροσαρμόζαμε το μοντέλο αναμένουμε η εκτίμηση της παραμέτρου για τη συμμεταβλητή treat να είναι στατιστικά σημαντικότερη και η τιμή της εκτιμήτριας να είναι μεγαλύτερη κατά απόλυτη τιμή καθώς πλέον θα επηρεάζεται περισσότερο η κοινή συνάρτηση διακινδύνευσης από το $\exp(\beta_{treat})$.

4.3.8 Προσθήκη συντελεστών αλληλεπίδρασης ως προς τον χρόνο

Όπως είδαμε στην παράγραφο 4.3.5, η υπόθεση της αναλογικής διακινδύνευσης δεν φαίνεται να ικανοποιείται ειδικά για τις συμμεταβλητές absblasts και smear. Όταν αντιλαμβανόμαστε ότι παραβιάζεται αυτή η βασική συνθήκη για το μοντέλο μας, δύο γνωστοί τρόποι επίλυσης του προβλήματος είναι: α) η στρωματοποίηση β) η προσθήκη συμμεταβλητών αλληλεπίδρασης ως προς τον χρόνο στο μοντέλο. Εμείς σε αυτή την παράγραφο θα ασχοληθούμε με τον δεύτερο τρόπο.

Επομένως, μπορούμε να αναπροσαρμόσουμε το ημιπαραμετρικό μοντέλο του Cox εισάγοντας αυτές τις 5 συμμεταβλητές (absblasts, smear, treat, age, labindex) αλλά επίσης και τις αλληλεπιδράσεις των συμμεταβλητών treat, absblasts και smear ως προς τον χρόνο, οι οποίες είδαμε ότι είναι οι πιο πιθανές να έχουνε κάποια χρονική εξάρτηση. Από τους πίνακες της R, μπορούμε να κρίνουμε την σημαντικότητα κάθε αλληλεπίδρασης και έτσι θα γίνει και ένας επιπλέον έλεγχος της υπόθεσης $\beta_i(\mathrm{t}) = \beta_i, \forall t$ αλλά ταυτόχρονα αποτελεί και λύση του προβλήματος (Allison, 1995).

Θα μπορούσαμε να πούμε ότι, από τις συμμεταβλητές που συμμετείχαν στο μοντέλο του Cox στο οποίο καταλήξαμε, αυτές που ήταν πιθανό να εξαρτώνται από τον χρόνο είναι οι smear, treat και absblasts. Αν θυμηθούμε την ερμηνεία κάθε μεταβλητής, οι μεταβλητές age και labindex είναι “εξωτερικές” (external), όπως αναλύσαμε και στην παράγραφο 2.2.6.2, δηλαδή είτε μεταβάλλονται με τον χρόνο αλλά με τρόπο αναμενόμενο (age=ηλικία ασθενή) είτε αλλάζουν από τον ίδιο τον πειραματιστή (labindex= ποσοστό κυττάρων που προήλθαν από μυελό των οστών). Αντίθετα, οι μεταβλητές treat=ανταπόκριση στη θεραπεία, absblasts=απόλυτος αριθμός βλαστοκυττάρων και smear=ποσοστό επίστρωσης βλαστοκυττάρων είναι “εσωτερικές” (internal) μεταβλητές, καθώς ενδέχεται να μεταβληθούν με την πάροδο του χρόνου αλλά αυτό είναι κάτι το οποίο εξαρτάται από τον ίδιο τον ασθενή και την πορεία της υγείας του.

Έχοντας αναφέρει όλα τα παραπάνω, μπορούμε να αναπροσαρμόσουμε το μοντέλο του Cox εισάγοντας και έναν παράγοντα εξάρτησης από τον χρόνο για τις συμμεταβλητές treat, smear και absblasts. Αν και λογικός θα ήταν λάθος, να ορίζαμε νέες μεταβλητές της μορφής treat*time, absblasts*time και smear*time και να τις εισάγαμε στο νέο μοντέλο ως παράγοντες αλληλεπίδρασης ως προς τον χρόνο. Το πρόβλημα με τις εν λόγω μεταβλητές είναι ότι, στην πραγματικότητα δεν θα δημιουργούσαμε μεταβλητές χρονοεξαρτώμενες αλλά στατικές ως προς τον χρόνο (time-static) για κάθε μονάδα. Μια γνήσια χρονικά εξαρτημένη συμμεταβλητή θα μπορούσε να δημιουργηθεί μόνο με την χρήση του ορίσματος tt (time transform) στην εντολή coxph η οποία είναι υπεύθυνη στην R για την προσαρμογή του μοντέλου του Cox (Therneau et al., 2017).

Τα αποτελέσματα, μετά την εισαγωγή αλληλεπίδρασης ως προς τον χρόνο για τις συμμεταβλητές treat, smear και absblasts, παρουσιάζονται συνοπτικά στον πίνακα 4.14.

Συμμεταβλητή	β	$se(\beta)$	Έλεγχος Wald	p-value
Age	0.0296044	0.0121885	2.429	0.015146
Smear	0.0461855	0.0140232	3.293	0.000989
Labindex	0.0636077	0.0451188	1.410	0.158604
Absblasts	-0.0755395	0.0265402	-2.846	0.004424
Treat	-6.0301690	1.6740703	-3.602	0.000316
tt(treat)	0.3380824	0.1847887	1.830	0.067315
tt(smear)	-0.0020449	0.0008871	-2.305	0.021152
tt(absblasts)	0.0030898	0.0018831	1.641	0.100841

Πίνακας 4.14: Συνοπτικός πίνακας αποτελεσμάτων μετά την προσαρμογή του μοντέλου του Cox προσθέτοντας και παράγοντες αλληλεπίδρασης ως προς τον χρόνο για τις συμμεταβλητές treat, smear και absblasts.

Παρατηρούμε από τα αποτελέσματα του πίνακα 4.14 ότι, η αλληλεπίδραση της συμμεταβλητής absblasts ως προς τον χρόνο φαίνεται να μην είναι στατιστικά σημαντική καθώς το p-value για τον έλεγχο Wald προκύπτει να είναι 0.10>0.05 ενώ η αντίστοιχη αλληλεπίδραση ως προς τον χρόνο της συμμεταβλητής smear προκύπτει στατιστικά σημαντική εφόσον το δικό της p-value για τον έλεγχο Wald είναι 0.02<0,05. Τέλος, από τον έλεγχο Wald για την αλληλεπίδραση ως προς τον χρόνο της συμμεταβλητής treat προκύπτει να είναι οριακά μη σημαντική στο μοντέλο μας. Μια ενδιαφέρουσα, επίσης, διαφοροποίηση σε σχέση με το μοντέλο χωρίς αλληλεπιδράσεις είναι ότι πλέον η συμμεταβλητή labindex παύει να είναι στατιστικά σημαντική. Αυτό πιθανότατα, προκύπτει λόγω της συσχέτισης που ενδέχεται να έχουν μεταξύ τους οι συμμεταβλητές του μοντέλου, αλλά αυτό είναι κάτι το οποίο θα εξετάσουμε όταν προσαρμόσουμε το μοντέλο του Cox εφαρμόζοντας τη μέθοδο Lasso.

Για να καταλήξουμε στο βέλτιστο μοντέλο που θα περιγράφει καλύτερα τα δεδομένα μας και θα περιέχει τις πλέον σημαντικές συμμεταβλητές που επηρεάζουν τη διάρκεια ζωής των ασθενών μπορούμε να εφαρμόσουμε ξανά τον αλγόριθμο της διαδοχικής αφαίρεσης. τα τελικά αποτελέσματα του αλγορίθμου φαίνονται στον πίνακα 4.15.

	Df	AIC		p-value

Συμμεταβλητή που αφαιρείται			Μεταβολή της $-2\hat{l}$	
Καμία	-	232.81	-	-
tt(absblasts)	1	233.82	3.015	0.0824923
tt(treat)	1	235.76	4.946	0.0261465
tt(smear)	1	237.57	6.763	0.0093069
Age	1	237.97	7.165	0.0074350
Absblasts	1	239.15	8.335	0.0038886
Smear	1	241.81	10.995	0.0009134
Treat	1	266.42	35.615	2,405e-09

Πίνακας 4.15: Πίνακας αποτελεσμάτων μετά την εκτέλεση του αλγορίθμου backward elimination

Οι συντελεστές του μοντέλου στο οποίο καταλήξαμε δίνονται από τον πίνακα 4.16

Συμμεταβλητή	β	$\exp(\beta)$	$se(\beta)$	Έλεγχος Wald	p-value
Age	0.0317334	1.0.322422	0.0119318	2.660	0.00782
Smear	0.0444685	1.0454721	0.0138558	3.209	0.00133
Absblasts	-0.0698323	0.9325502	0.0256981	-2.717	0.00658
Treat	-5.6292285	0.0035913	1.6382288	-3.436	0.00059
tt(treat)	0.3318660	1.3935661	0.1842910	1.801	0.07174
tt(smear)	-0.0021828	0.9978196	0.0008528	-2.560	0.01048
tt(absblasts)	0.0034158	1.0034216	0.0018205	1.876	0.06062

Πίνακας 4.16: Συγκεντρωτικός πίνακας αποτελεσμάτων τελικού μοντέλου

Οπότε, συμπερασματικά, δεχόμενοι ότι δεν μπορεί να ισχύει η υπόθεση της αναλογικής διακινδύνευσης καταλήξαμε σε μια επέκταση του κλασικού μοντέλου του Cox στο οποίο περιέχονται και όροι αλληλεπίδρασης ως προς τον χρόνο. Η τελική μορφή αυτού του μοντέλου δίνεται από τον τύπο:

$$\mathrm{h}(\mathrm{t};\underline{x}) = h_0(t)\cdot\exp(0.0317334age + 0.0444685smear - 0.0698323abssblasts - 5.6292285treat + 0.3318660tt(treat) - 0.0021828tt(smear) + 0.0034158tt(absblasts))$$

Θα πρέπει, σε αυτό το σημείο, ως τελευταία επισήμανση, να αναφέρουμε ότι το τελικό μοντέλο με τις αλληλεπιδράσεις των συμμεταβλητών ως προς τον χρόνο έχει τιμή του κριτηρίου AIC=232.8096 που είναι μικρότερη από την τιμή του ίδιου κριτηρίου για το αρχικό μοντέλο και με τις 7 συμμεταβλητές όπου είχαμε AIC=241.217 και επίσης από το μοντέλο με τις 5 συμμεταβλητές όπου AIC=237.622, άρα κρίνεται και ως το καταλληλότερο για την περιγραφή των δεδομένων μας.

4.4 Εφαρμογή μεθόδων ποινών στο μοντέλο του Cox

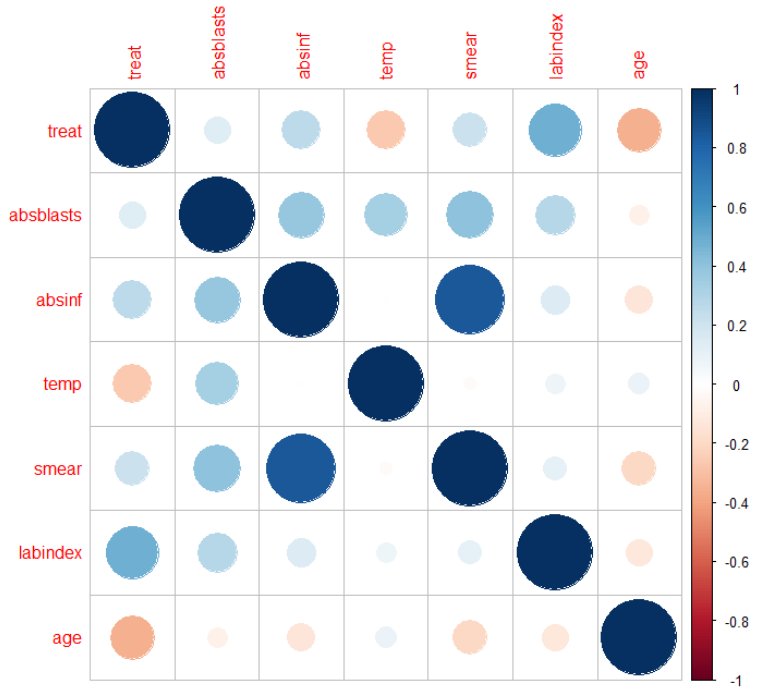
Αφού έχει προηγηθεί η ανάλυσή μας για την προσαρμογή του κλασικού μοντέλου του Cox μπορούμε να εφαρμόσουμε τις τεχνικές των μεθόδων ποινών. Εισάγοντας μια ποινή στην εκτίμηση των παραμέτρων, είδαμε ότι μπορούμε να αντιμετωπίσουμε τυχόν προβλήματα πολυσυγγραμμικότητας που ενδέχεται να διέπουν τις συμμεταβλητές του μοντέλου. Καταλήγοντας σε ένα νέο, διαφορετικό μοντέλο μπορούμε να κάνουμε σύγκριση με τα αποτελέσματα του κλασικού μοντέλο του Cox και να αναδείξουμε τις ιδιότητες των μεθόδων που μελετήσαμε στη θεωρία.

4.4.1 Έλεγχος του φαινομένου της πολυσυγγραμμικότητας

Αρχικά θα πρέπει να ελέγξουμε αν παρουσιάζεται το φαινόμενο της πολυσυγγραμμικότητας ανάμεσα στις συμμεταβλητές του μοντέλου, δηλαδή αν υπάρχει υψηλή συσχέτιση μεταξύ τους κάτι που θα μπορούσε να οδηγεί σε εξασθένηση της επιρροής τους όταν συνυπάρχουν στο προσαρμοσμένο μοντέλο. Με τη βοήθεια της R παρουσιάζονται αναλυτικά και γραφικά (με το πακέτο corrplot) τα αποτελέσματα για τη συσχέτιση των υποψήφιων συμμεταβλητών του μοντέλου:

Συμμεταβλητή	treat	absblasts	absinf	temp	smear	labindex	Age
treat	1	0.1321	0.2648	-0.2648	0.2153	0.4882	-0.3501
absbalsts	0.1321	1	0.3806	0.3313	0.4039	0.2879	-0.0705
absinf	0.2648	0.3806	1	-0.0067	0.8471	0.1444	-0.1370
temp	-0.2648	0.3313	-0.0067	1	-0.0282	0.0705	0.0846
smear	0.2153	0.4039	0.8471	-0.0282	1	0.1027	-0.2038
labindex	0.4882	0.2879	0.1444	0.0705	0.1027	1	-0.1243
age	-0.3501	-0.0705	-0.1370	0.0846	-0.2038	-0.1243	1

Πίνακας 4.17: Πίνακας ελέγχου της συσχέτισης των συμμεταβλητών του μοντέλου



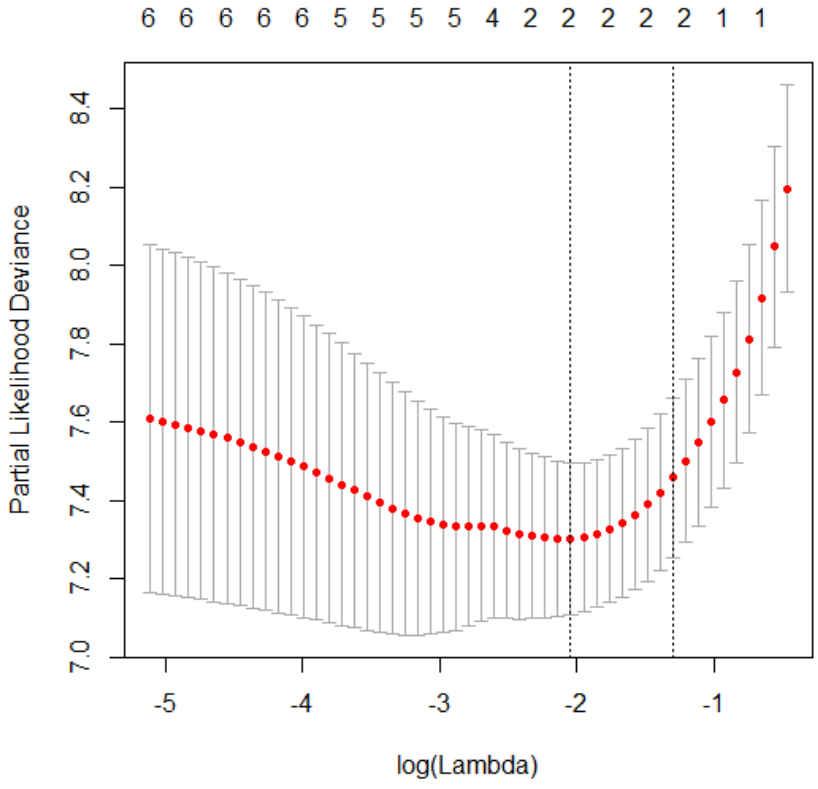
Γράφημα 4.17: Γραφικός έλεγχος του φαινομένου της πολυσυγγραμικότητας

Από τον πίνακα 4.17 και το βοηθητικό γράφημα 4.17 αντιλαμβανόμαστε ότι όλες οι υποψήφιες συμμεταβλητές του μοντέλου είναι από λίγο έως πολύ αλληλοσυσχετισμένες. Πιο συγκεκριμένα, υψηλή θετική συσχέτιση παρατηρούμε ανάμεσα στις συμμεταβλητές absinf και smear, labindex και treat, smear και absblasts ενώ υψηλή αρνητική συσχέτιση παρατηρείται μεταξύ των συμμεταβλητών age και treat και temp και treat. Συνολικά, θα μπορούσαμε να πούμε ότι, οι συμμεταβλητές treat και absblasts (οι οποίες θυμίζουμε ότι συμπεριλήφθηκαν στην τελική μορφή του κλασικού μοντέλου του Cox) είναι εκείνες που έχουν την πιο έντονη συσχέτιση με όλες τις υπόλοιπες.

4.4.2 Εφαρμογή της μεθόδου Lasso

Σε αυτό το σημείο θα προσαρμόσουμε από την αρχή το μοντέλο του Cox εφαρμόζοντας τη μέθοδο Lasso. Με αυτόν τον τρόπο θα αντιμετωπίσουμε το πρόβλημα πολυσυγγραμικότητας των συμμεταβλητών του μοντέλου. Η επιλογή του βέλτιστου συντελεστή λ_1, που θα χρησιμοποιηθεί στη μέθοδο Lasso, θα γίνει με τη μέθοδο cross-validation. Τα αποτελέσματα στα οποία θα καταλήξουμε θα συγκριθούν με τα αντίστοιχα του κλασικού μοντέλου του Cox που προσαρμόστηκε στην αρχή αυτού του κεφαλαίου.

Για την υλοποίηση της μεθόδου Lasso, άλλα και των μεθόδων Ridge και (naïve) Elastic Net, χρησιμοποιήσαμε το πακέτο της R glmnet. Γενικά, το πακέτο glmnet χρησιμοποιείται για την εφαρμογή της μεθόδου Elastic Net, όμως επειδή έχουμε τη δυνατότητα να επιλέγουμε την τιμή της παραμέτρου α στην συνάρτηση ποινής $(1-a)\left\|\vec{\beta}\right\|_2^2 + a\left\|\vec{\beta}\right\|_1 \le c$, μπορούμε να εφαρμόσουμε τις μεθόδους Lasso και Ridge πολύ απλά θέτοντας εμείς εξ αρχής ως α=1 και α=0 αντίστοιχα.
Χρησιμοποιώντας την εντολή cv.glmnet μπορούμε να εφαρμόσουμε την γενικευμένη μέθοδο cross-validation η οποία μας επιστρέφει την βέλτιστη τιμή της παραμέτρου λ που θα χρησιμοποιήσουμε. Τα γραφικά αποτελέσματα της μεθόδου φαίνονται στο γράφημα 4.18



Γράφημα 4.18: Deviance της cross-validation μερικής πιθανοφάνειας συναρτήσει του log(λ)

Στο γράφημα 4.18, έχει σχεδιαστεί η καμπύλη της deviance της cvl συμπεριλαμβανομένου, για κάθε σημείο της ακολουθίας των λ που διαλέγει αυτόματα η R, ενός διαστήματος ± 1 τυπικής απόκλισης. Επιπλέον, επισημαίνονται μέσω 2 κάθετων γραμμών 2 συγκεκριμένες τιμές του λ. Η αριστερή κάθετη γραμμή μας δίνει το λ για το οποίο η deviance της cvl παίρνει την ελάχιστη τιμή της. Η δεξιά κάθετη γραμμή μας δίνει τη μέγιστη τιμή του λ σε απόσταση 1 τυπικής απόκλισης από την ελάχιστη τιμή της deviance. Αυτές τις τιμές μας τις επιστρέφει η R μέσω των εντολών cv.glmnet$lambda.min και cv.glmnet$lambda.1se. Τέλος, στο πάνω μέρος του γραφήματος εμφανίζεται το πλήθος των μη μηδενικών παραμέτρων καθώς το log(λ), άρα και το ίδιο το λ, αυξάνεται. Έτσι λοιπόν, για μικρές τιμές του λ στο μοντέλο συμπεριλαμβάνονται και οι 7 συμμεταβλητές ενώ όσο μεγαλώνει το λ τόσο λιγοστεύουν οι μη μηδενικές παράμετροι του μοντέλου.

Η τιμή του βέλτιστου λ που μας ενδιαφέρει προκύπτει να είναι λ=0.1296947 (log(λ)= -2.04)
Χρησιμοποιούμε την εντολή glmnet για την προσαρμογή του μοντέλου Cox εφαρμόζοντας τη μέθοδο Lasso και μέσω της εντολής coef μπορούμε να πάρουμε τις τιμές των παραμέτρων. Τα αποτελέσματα παρουσιάζονται στον πίνακα 4.18

Age	smear	absinf	absblasts	temp	labindex	Treat
0.01458	0	0	0	0	0	-1.8547

Πίνακας 4.18: Παράμετροι του μοντέλου του Cox μετά την εφαρμογή της μεθόδου Lasso

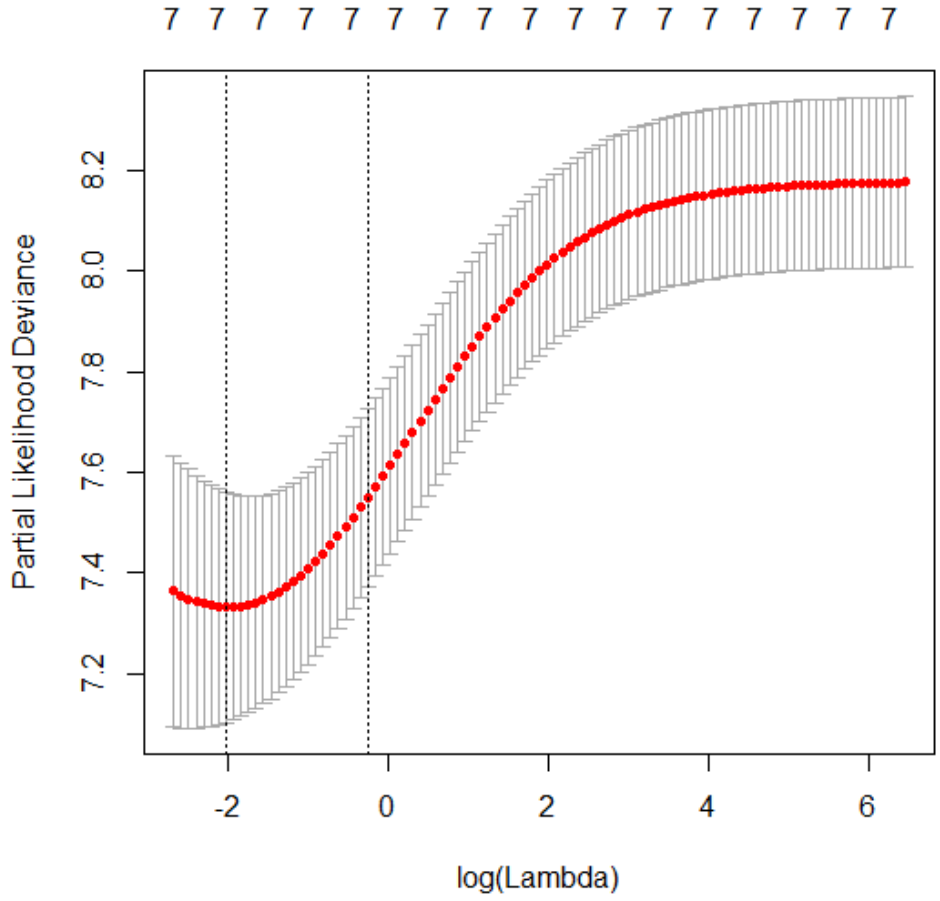
Όπως αναφέραμε και στην ανάλυση της μεθόδου Lasso στην παράγραφο 3.3.2, παρατηρούμε ότι με την εφαρμογή L1 τύπου ποινής οι περισσότεροι συντελεστές του μοντέλου έχουν μηδενιστεί ενώ οι υπόλοιποι έχουν υποστεί πολύ μικρή συρρίκνωση. Πιο συγκεκριμένα, οι παράμετροι των συμμεταβλητών smear, absinf, absblasts, temp και labindex έχουν γίνει ταυτοτικά μηδέν.

Σημείωση: Για την εφαρμογή της μεθόδου έπρεπε να εξαιρέσουμε από το σύνολο δεδομένων μας τα στοιχεία για τον 22ο κατά σειρά ασθενή για τον οποίο μας δίνεται ο χρόνος διάρκειας ζωής του σε μήνες να είναι T=0. Προφανώς, αυτό έχει γίνει για λόγους στρογγυλοποίησης από τον ίδιο τον μελετητή που έκανε τις μετρήσεις καθώς ο ασθενής απεβίωσε πριν την συμπλήρωση ενός μήνα παρακολούθησης. Παρόλα αυτά έπρεπε να αφαιρεθεί διότι το πακέτο glmnet είναι δομημένο έτσι ώστε να επεξεργάζεται αυστηρά θετικούς χρόνους διάρκειας ζωής.

Γενικά έχουμε αναφέρει ότι η μέθοδος Lasso θεωρείται αποτελεσματικότερη από την Ridge για την αντιμετώπιση του προβλήματος της πολυσυγγραμμικότητας. Όμως, στη συγκεκριμένη εφαρμογή είχαμε σχετικά λίγες συμμεταβλητές (συνολικά 7), οπότε λόγω υψηλής συσχέτισης είναι πιθανό η εφαρμογή της μεθόδου Lasso να καταλήγει σε μια μορφή μοντέλου το οποίο δεν είναι το πλέον ενδεδειγμένο για την περιγραφή των δεδομένων μας καθώς από αυτό έχουν εξαιρεθεί σημαντικές συμμεταβλητές. Για αυτό λοιπόν, κρίνεται σκόπιμο να χρησιμοποιήσουμε και τις μεθόδους Ridge και Elastic Net ώστε να κάνουμε σύγκριση των αποτελεσμάτων.

4.4.3 Εφαρμογή της μεθόδου Ridge

Για την εφαρμογή της μεθόδου Ridge στο μοντέλο του Cox, μπορούμε να χρησιμοποιήσουμε τις αντίστοιχες εντολές με αυτές που είδαμε για τη μέθοδο Lasso μέσω του πακέτου glmnet. Η βασική διαφορά είναι ότι τώρα θα θεωρήσουμε ότι η παράμετρος α=0 ώστε να εφαρμοστεί η μέθοδος Ridge. Άρα, αρχικά, εκτελούμε την εντολή cv.glmnet για να βρούμε τη βέλτιστη τιμή του λ.
Στο γράφημα 4.19 αναπαρίσταται η deviance της cross-validation συνάρτησης μερικής πιθανοφάνειας συναρτήσει της τιμής του log(λ).



Γράφημα 4.19: Υπολογισμός της deviance της συνάρτησης cvl συναρτήσει του log(λ)

Όπως και στη μέθοδο Lasso, για να βρούμε τη βέλτιστη τιμή του λ χρησιμοποιούμε την εντολή cv.glmnet$lambda.min από την οποία μας επιστρέφεται τελικά η τιμή λ= 0.1327466 (log(λ)= -2.02)

Τέλος, για την εφαρμογή της μεθόδου Ridge στο μοντέλο του Cox για τη δεδομένη τιμή του λ, χρησιμοποιούμε την εντολή glmnet. Χρησιμοποιώντας και την εντολή coef εμφανίζονται όλες οι παράμετροι του μοντέλου για να ελέγξουμε το μέγεθος της συρρίκνωσης που έχουν υποστεί συγκριτικά με το αρχικό μοντέλο του Cox που προσαρμόσαμε. Τα αποτελέσματα παρουσιάζονται στον πίνακα 4.19

Age	smear	absinf	absblasts	Temp	labindex	treat
0.02174	0.00915	-0.00294	-0.01443	0.00246	0.02471	-2.0045

Πίνακας 4.19: Παράμετροι του μοντέλου του Cox μετά την εφαρμογή της μεθόδου Ridge

Όπως αναμενόταν και από την ανάλυση της θεωρίας για τη μέθοδο ridge, καμία παράμετρος δεν έγινε ακριβώς μηδέν αν και παρατηρούνται τροποποιήσεις συγκριτικά με την αρχική

προσαρμογή του μοντέλου του Cox. Το βασικό μειονέκτημα της μεθόδου είναι ακριβώς ότι διατηρούνται όλες οι αρχικές παράμετροι του μοντέλου άρα δεν γίνεται καμία επιλογή του καλύτερου συνόλου συμμεταβλητών. Συγκριτικά με το πολύ "φτωχό" μοντέλο της μεθόδου Lasso και το "εκτεταμένο" μοντέλο της μεθόδου Ridge, τα αποτελέσματα μας μπορούν να βελτιωθούν εφαρμόζοντας την μέθοδο Elastic Net για μια τιμή της παραμέτρου $\alpha \in (0,1)$.

4.4.4 Εφαρμογή της μεθόδου Elastic Net

Όπως αναλύσαμε και στην παράγραφο 3.3.3 η μέθοδος Elastic Net αποτελεί μια μίξη των Lasso και Ridge η οποία προσπαθεί να κρατήσει μόνο τις καλές ιδιότητες των δύο αυτών μεθόδων. Για να προσαρμόσουμε με τη βοήθεια του πακέτου glmnet το μοντέλο του Cox εφαρμόζοντας την μέθοδο Elastic Net θα πρέπει να καθορίσουμε με κάποιον τρόπο την τιμή που θα παίρνει η παράμετρος α. Ξέρουμε ότι για τις Lasso και Ridge στην γενικευμένη συνάρτηση ποινής $(1-a)\left\|\underset{\rightarrow}{\beta}\right\|_2^2 + a\left\|\underset{\rightarrow}{\beta}\right\|_1 \le c$ πρέπει να θεωρήσουμε α=1 και α=0 αντίστοιχα. Για την Elastic Net η επιλογή της τιμής του α δεν είναι τετριμμένη. Γενικά, όσο το α τείνει στο 1 τόσο προσεγγίζουμε τα αποτελέσματα της μεθόδου Lasso και αντίστοιχα όσο το α τείνει στο μηδέν τόσο τα αποτελέσματα μας μοιάζουν με εκείνα της μεθόδου Ridge.

Μια λύση είναι να επιλέξουμε αυθαίρετα μια τιμή του $\alpha \in (0,1)$ (πχ α=0.5) και εφαρμόζοντας την μέθοδο cross-validation να πάρουμε τη βέλτιστη τιμή του λ που μας επιστρέφει η R. Αντί αυτής της απλουστευμένης αντιμετώπισης, μπορούμε να επιλέξουμε τις βέλτιστες τιμές των παραμέτρων α και λ που χρειαζόμαστε δουλεύοντας ως εξής:

- Αντί για μια αυθαίρετη τιμή της παραμέτρου α, χρησιμοποιούμε ένα διάνυσμα τιμών από 0.01 έως 0.99.
- Για κάθε τιμή του διανύσματος α βρίσκουμε τη βέλτιστη τιμή του λ εκτελώντας τη μέθοδο cross-validation.
- Εντοπίζουμε τα ολικά βέλτιστα α και λ τα οποία χρησιμοποιούμε για τη μέθοδο Elastic Net

Τελικά καταλήγουμε ότι οι βέλτιστες τιμές των α και λ είναι: α= 0.55, λ= 0.06410664

Προσαρμόζουμε το μοντέλο του Cox εφαρμόζοντας τη μέθοδο Elastic Net για τις παραπάνω τιμές των α και λ, οπότε καταλήγουμε στις τιμές των παραμέτρων του τελικού μοντέλου που φαίνονται στον πίνακα 4.20

age	smear	absinf	absblasts	temp	labindex	Treat
0.02103	0.00701	0	-0.010162	0	0.02953	-2.3565

Πίνακας 4.20: Παράμετροι του μοντέλου του Cox μετά την εφαρμογή της μεθόδου Elastic Net

Όπως είχαμε σχολιάσει και παραπάνω, η μέθοδος Elastic Net καταλήγει σε αποτελέσματα που είναι ανάμεσα στα αποτελέσματα της Ridge και της Lasso. Παρατηρούμε λοιπόν ότι, στο τελικό μοντέλο έχουμε τις 5 μη μηδενικές παραμέτρους age, smear, absinf, absblasts, labindex και treat ενώ οι άλλες δύο παράμετροι των συμμεταβλητών temp και absinf έχουν μηδενιστεί.

4.4 Γενικά συμπεράσματα

Στην εφαρμογή που μελετήσαμε στην παρούσα εργασία, αν και είχαμε σχετικά μικρό δείγμα (n=51 ασθενείς), αλλά και λίγες επεξηγηματικές μεταβλητές (7 στο σύνολο), παρόλα αυτά καταφέραμε και αναλύσαμε εις βάθος το κλασικό μοντέλο του Cox όπως επίσης και τα μοντέλα των μεθόδων ποινών Ridge, Lasso και (naïve) Elastic Net.

Καθ' όλη την πορεία της ανάλυσης, εφαρμόζοντας τις τεχνικές που αναπτύξαμε στη θεωρία, καταλήξαμε σε διάφορες μορφές μοντέλων από τις οποίες μπορούμε να πούμε ότι οι επικρατέστερες για την περιγραφή των δεδομένων μας είναι οι ακόλουθες:

Συμμεταβλητή	"Βέλτιστο" μοντέλο του Cox	Γενικευμένο μοντέλο του Cox	Μοντέλο μεθόδου Ridge	Μοντέλο μεθόδου Lasso	Μοντέλο μεθόδου Elastic Net
age	0.0237	0.0317	0.0217	0.0146	0.0210
smear	0.0255	0.0445	0.0092	0	0.0070
absblasts	-0.0421	-0.0698	-0.0144	0	-0.0102
absinf	0	0	-0.0029	0	0
labindex	0.0802	0	0.0247	0	0.0295
temp	0	0	0.0025	0	0
treat	-3.4420	-5.6292	-2.0045	-1.8547	-2.3565
tt(treat)	-	0.3319	-	-	-
tt(smear)	-	-0.0022	-	-	-
tt(absblasts)	-	0.0034	-	-	-

Πίνακας 4.21: Συγκεντρωτικός πίνακας αποτελεσμάτων όλων των μοντέλων

a) **Το "βέλτιστο" μοντέλο του Cox:** Προέκυψε μετά την προσαρμογή του κλασικού μοντέλου του Cox και την εκτέλεση της μεθόδου διαδοχικής αφαίρεσης (backward elimination) για την επιλογή των στατιστικά σημαντικότερων μεταβλητών. Περιέχει συνολικά 5 συμμεταβλητές. Παρουσιάζει αρκετά καλή προβλεπτική ικανότητα, ειδικά για χρόνους διάρκειας ζωής T<12 μηνών. Σημαντικό μειονέκτημά του η απουσία συντελεστών αλληλεπίδρασης ως προς τον χρόνο για εκείνες τις (σημαντικές) συμμεταβλητές για τις οποίες δεν ισχύει η υπόθεση της αναλογικής διακινδύνευσης

b) **Το γενικευμένο μοντέλο του Cox:** Περιέχει χρονικές αλληλεπιδράσεις κάποιων στατιστικά σημαντικών συμμεταβλητών οι οποίες μέσα από αναλυτικούς και γραφικούς τρόπους φαίνεται ότι δεν υπακούν στην υπόθεση της αναλογικότητας της διακινδύνευσης. Περιέχει συνολικά 8 συμμεταβλητές εκ των οποίων οι 3 είναι αλληλεπιδράσεις ως προς τον χρόνο. Θεωρείται πιο ρεαλιστικό σε σχέση με το μοντέλο a) αλλά λόγω αυτής της επιπλέον πολυπλοκότητάς του δεν μπορούμε να ξέρουμε αν είναι ιδιαίτερα εύχρηστο για μεγαλύτερα σύνολα δεδομένων.

c) **Το μοντέλο της μεθόδου Ridge:** Για την αντιμετώπιση του προβλήματος της πολυσυγγραμμικότητας των μεταβλητών του προβλήματός μας εφαρμόσαμε τις μεθόδους ποινών στο μοντέλο του Cox, μια εκ των οποίων είναι η μέθοδος Ridge. Η μέθοδος Ridge ναι μεν συρρικνώνει τις εκτιμήτριες των συντελεστών του μοντέλου αλλά δεν μηδενίζει καμία από αυτές οπότε αδυνατεί να επιλέξει το σύνολο των στατιστικά σημαντικότερων συμμεταβλητών για τη περιγραφή των δεδομένων μας. Για αυτό το λόγο το μοντέλο της μεθόδου Ridge περιέχει και τις 7 συμμεταβλητές.

d) **Το μοντέλο της μεθόδου Lasso:** Η μέθοδος Lasso, σε αντίθεση με την Ridge, καταλήγει σε μια μορφή μοντέλου όπου οι περισσότερες παράμετροι του μοντέλου

έχουν μηδενιστεί και όλοι οι υπόλοιποι έχουν υποστεί αμελητέα συρρίκνωση. Έτσι λοιπόν καταλήξαμε σε ένα μοντέλο με μόνο 2 συμμεταβλητές. Ένα πιθανό μειονέκτημα αυτού του μοντέλου είναι ότι λόγω του μικρού πλήθους των υποψήφιων συμμεταβλητών που είχαμε εξ αρχής που είχαμε στη διάθεσή μας να καταλήξαμε σε ένα "φτωχό" μοντέλο το οποίο δεν μπορεί να περιγράψει ικανοποιητικά τα δεδομένα.

e) **Το μοντέλο της μεθόδου (naïve) Elastic Net:** Η μέθοδος (naïve) Elastic Net προσπαθεί να συνδυάσει τα προτερήματα των Ridge και Lasso. Έτσι λοιπόν καταλήγουμε σε μια μορφή μοντέλου με 5 συμμεταβλητές οι παράμετροι των οποίων έχουν υποστεί μεγαλύτερη συρρίκνωση σε σχέση με τις αντίστοιχες της μεθόδου Ridge.

Παράρτημα

Α) Θεωρητικές αποδείξεις

1) Αλγόριθμος υπολογισμού των lasso εκτιμητριών

Για τον υπολογισμό των εκτιμητριών lasso θέλουμε να λύσουμε το εξής πρόβλημα:

$$\min \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \quad \text{δεδομένου ότι: } \left\|\underset{\rightarrow}{\beta}\right\|_1 \le c \Rightarrow \sum_{j=1}^{p}\left|\beta_j\right| \le c \quad (3.14)$$

Για κάθε τιμή της παραμέτρου c η εκτιμήτρια του σταθερού όρου β_0 είναι $\beta_0 = \bar{y}$. Άρα χωρίς βλάβη της γενικότητας μπορούμε να θεωρήσουμε ότι $\bar{y}$ =0 και επομένως να παραλείψουμε από το μοντέλο τον σταθερό όρο.

Αρχικά σταθεροποιούμε το $c \ge 0$. Έστω λοιπόν, $g(\underset{\rightarrow}{\beta}) = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$ και έστω $\underset{\rightarrow}{\delta}_i, i = 1,2,...,2^p$ οι p-πλειάδες της μορφής $(\pm 1, \pm 1, ..., \pm 1)$. Τότε η συνθήκη: $\sum_{j=1}^{p}\left|\beta_j\right| \le c$ μπορεί να γραφεί ισοδύναμα και ως: $\underset{\rightarrow}{\delta}_i^T \cdot \underset{\rightarrow}{\beta} \le c, \forall i$. Για δοσμένο $\underset{\rightarrow}{\beta}$, έστω τα σύνολα: $E = \{i : \underset{\rightarrow}{\delta}_i^T \cdot \underset{\rightarrow}{\beta} = c\}$ και $\mathrm{S} = \{i : \underset{\rightarrow}{\delta}_i^T \cdot \underset{\rightarrow}{\beta} < c\}$. Θεωρούμε ως: G_E τον πίνακα με γραμμές τα $\underset{\rightarrow}{\delta}_i$ για εκείνα τα $i \in E$. Ορίζουμε επίσης ως $\underset{\rightarrow}{1}$ το διάνυσμα μονάδων με μήκος όσο και το πλήθος των γραμμών του G_E.

Ο ακόλουθος αλγόριθμος (Tibshirani, 1996) ξεκινά με $E = \{i_0\}$ όπου $\underset{\rightarrow}{\delta}_{i_0} = sign(\underset{\rightarrow}{\beta}^{OLS})$, όπου συμβολίζουμε ως $\underset{\rightarrow}{\beta}^{OLS}$ τις τυπικές εκτιμήτριες ελαχίστων τετραγώνων. Ο αλγόριθμος λύνει το πρόβλημα ελαχίστων τετραγώνων δεδομένης της συνθήκης $\underset{\rightarrow}{\delta}_{i_0} \cdot \underset{\rightarrow}{\beta} \le c$ και μετά ελέγχει αν ισχύει: $\sum\left|\beta_j\right| \le c$. Αν ισχύει, ο αλγόριθμος τερματίζεται. Αν όχι, ο περιορισμός που παραβιάζεται προστίθεται στο σύνολο E και η διαδικασία συνεχίζεται έως ότου $\sum\left|\beta_j\right| \le c$. Τα βήματα του αλγορίθμου περιγράφονται συνοπτικά παρακάτω:

- Ξεκινάμε με το σύνολο $E = \{i_0\}$, όπου $\underset{\rightarrow}{\delta}_{i_0} = sign(\underset{\rightarrow}{\beta}^{OLS})$
- Βρίσκουμε το $\underset{\rightarrow}{\beta}$ που ελαχιστοποιεί τη συνάρτηση $g(\underset{\rightarrow}{\beta})$ δεδομένου: $\mathrm{G}_E \cdot \underset{\rightarrow}{\beta} \le c \cdot \underset{\rightarrow}{1}$
- Όσο ισχύει η συνθήκη: $\{\sum\left|\beta_j\right| > c\}$
  - ❖ Πρόσθεσε το i στο σύνολο E, όπου $\underset{\rightarrow}{\delta}_i = sign(\underset{\rightarrow}{\beta})$. Βρίσκουμε το $\underset{\rightarrow}{\beta}$ που ελαχιστοποιεί τη συνάρτηση $g(\underset{\rightarrow}{\beta})$ δεδομένου: $\mathrm{G}_E \cdot \underset{\rightarrow}{\beta} \le c \cdot \underset{\rightarrow}{1}$

Αυτή η διαδικασία θα πρέπει πάντα να συγκλίνει σε πεπερασμένο αριθμό βημάτων εφόσον σε κάθε βήμα του αλγορίθμου προστίθεται ένα στοιχείο στο σύνολο E και υπάρχουν συνολικά διαθέσιμα 2^p στοιχεία. Βέβαια, το γεγονός ότι ο αλγόριθμος τερματίζεται σε 2^p το πολύ

βήματα δεν είναι ιδιαίτερα εύχρηστο στην περίπτωση όπου ο αριθμός των παραμέτρων του μοντέλου p είναι πολύ μεγάλος. Όμως, στην πράξη έχει αποδειχτεί ότι ο αναμενόμενος αριθμός επαναλήψεων της μεθόδου είναι στο εύρος του διαστήματος $(0.5p, 0.75p)$, επομένως θεωρείται ικανοποιητικός για τις εφαρμογές.

2) Υπολογισμός των leave-one-out συντελεστών $\underset{\rightarrow}{\beta}_{(-i)}$

Υπάρχουν αρκετοί τρόποι για τον υπολογισμό των λεγόμενων leave-one-out συντελεστών $\underset{\rightarrow}{\beta}_{(-i)}$ παλινδρόμησης που χρησιμοποιούνται στην μέθοδο cross validation αλλά εμείς θα αναφέρουμε ενδεικτικά μόνο τον πρώτο που προτείνεται από τους Verweij και Van Houwelingen (Verweij & van Houwelingen 1993) . Αυτή η μέθοδος υπολογισμού βασίζεται στο γεγονός ότι το $\underset{\rightarrow}{\beta}_{(-i)}$ μεγιστοποιούν την $l_{(-i)}(\underset{\rightarrow}{\beta})$, άρα λοιπόν η πρώτη παράγωγος θα πρέπει να μηδενίζεται για $\underset{\rightarrow}{\beta} = \underset{\rightarrow}{\beta}_{(-i)}$. Οπότε η προσέγγιση Taylor πρώτης τάξης υπολογισμένη στο $\underset{\rightarrow}{\beta} = \underset{\rightarrow}{\beta}$ μας οδηγεί στη σχέση:

$$\underset{\rightarrow}{\beta}_{(-i)} = \underset{\rightarrow}{\beta} + \left(\frac{\partial^2 l}{\partial \underset{\rightarrow}{\beta}^2} - \frac{\partial^2 l_i}{\partial \underset{\rightarrow}{\beta}^2} \right)^{-1} \cdot \frac{\partial l_i}{\partial \underset{\rightarrow}{\beta}}$$

Στο μοντέλο του Cox η πρώτη παράγωγος όλων των $l_i(\underset{\rightarrow}{\beta})$, στη γενικότερη περίπτωση όπου έχουμε ισόπαλους χρόνους διακοπής, δίνεται από τη σχέση:

$$\frac{\partial l_i(\underset{\rightarrow}{\beta})}{\partial \underset{\rightarrow}{\beta}}(\underset{\rightarrow}{\beta}) = -\sum_{j<i} d_j \frac{p_{ij}}{1-p_{ij}} \left(\underset{\rightarrow}{x}_i - \bar{\underset{\rightarrow}{x}}_j \right) + d_i (\underset{\rightarrow}{x}_i - \bar{\underset{\rightarrow}{x}}_i)$$

όπου έχουμε συμβολίσει ως:

$$p_{ij} = \frac{w_i}{\sum_{k \in R_j} w_k} \ , \ w_j = \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \underset{\rightarrow}{x}_j) \ , \ \bar{\underset{\rightarrow}{x}}_j = \frac{\sum_{k \in R_j} w_k \underset{\rightarrow}{x}_k}{\sum_{k \in R_j} w_k} = \sum_{k \in R_j} p_{kj} \underset{\rightarrow}{x}_k$$

Το $\bar{\underset{\rightarrow}{x}}_j$ εκφράζει τον σταθμισμένο μέσο των συμμεταβλητών των μονάδων που ανήκουν στο σύνολο κινδύνου R_j. Τέλος, υπολογίζουμε και την δεύτερη παράγωγο του $l_i(\underset{\rightarrow}{\beta})$, η οποία προκύπτει να είναι:

$$\frac{\partial^2 l_i(\underset{\rightarrow}{\beta})}{\partial \underset{\rightarrow}{\beta}^2}(\underset{\rightarrow}{\beta}) = -\sum_{j<i} d_j \left[\frac{p_{ij}}{(1-p_{ij}^2)} (\underset{\rightarrow}{x}_i - \bar{\underset{\rightarrow}{x}}_j)(\underset{\rightarrow}{x}_i - \bar{\underset{\rightarrow}{x}}_j)^T - \frac{p_{ij}}{1-p_{ij}} \mathrm{var}(\underset{\rightarrow}{x}_j) \right] - d_i \, \mathrm{var}(\underset{\rightarrow}{x}_i)$$

όπου συμβολίσαμε ως:

$$\mathrm{var}(\underset{\rightarrow}{x}_j) = \sum_{k \in R_j} p_{kj} (\underset{\rightarrow}{x}_k - \bar{\underset{\rightarrow}{x}}_j)(\underset{\rightarrow}{x}_k - \bar{\underset{\rightarrow}{x}}_j)^T$$

είναι η σταθμισμένη διασπορά των συμμεταβλητών όλων των μονάδων που ανήκουν στο σύνολο R_j.

Επειδή ο υπολογισμός των δεύτερων παραγώγων μπορεί να είναι χρονοβόρος μπορούμε να επιλέξουμε την απλουστευμένη μορφή της εκτιμήτριας $\underset{\rightarrow}{\beta}_{(-i)}$ στην οποία έχει παραλειφθεί από το ανάπτυγμα Taylor η δευτεροβάθμια παράγωγος του $l_i(\underset{\rightarrow}{\beta})$. Άρα θα είχαμε:

$$\underset{\rightarrow}{\beta}_{(-i)} = \underset{\rightarrow}{\beta} + \left(\frac{\partial^2 l}{\partial \underset{\rightarrow}{\beta}^2} \right)^{-1} \cdot \frac{\partial l_i}{\partial \underset{\rightarrow}{\beta}}$$

3) Εύρεση σημείων επιρροής στο μοντέλο του Cox

Θεωρούμε ότι η επιρροή της j-οστής παρατήρησης στο $\underset{\rightarrow}{\beta}$ εκφράζεται ως: $\underset{\rightarrow}{\beta} - \underset{\rightarrow}{\beta}_{(j)}$. Έστω ότι χρησιμοποιούμε σταθμισμένη ανάλυση και η j-οστή παρατήρηση έχει βάρος w_j. Επιπλέον, δεχόμαστε ότι όλες οι υπόλοιπες παρατηρήσεις έχουν το ίδιο βάρος $w_i = 1, \forall i \neq j$. Θεωρώντας το $\underset{\rightarrow}{\beta}$ ως μια συνάρτηση του w_j, δηλαδή είναι $\underset{\rightarrow}{\beta} = \underset{\rightarrow}{\beta}(w_j)$, έχουμε ότι: $\underset{\rightarrow}{\beta}(1) = \underset{\rightarrow}{\beta}$ και $\underset{\rightarrow}{\beta}(0) = \underset{\rightarrow}{\beta}_{(j)}$. Η προσέγγιση του $\underset{\rightarrow}{\beta} - \underset{\rightarrow}{\beta}_{(j)}$ μέσω μιας πρώτης τάξης σειρά Taylor γύρω από το $w_j = 1$ θα είναι λοιπόν:

$$\underset{\rightarrow}{\beta} - \underset{\rightarrow}{\beta}_{(j)} \cong \frac{\partial \underset{\rightarrow}{\beta}}{\partial w_j}$$

Η παράγωγος $\partial \underset{\rightarrow}{\beta} / \partial w_j$ μπορεί να υπολογιστεί θεωρώντας το διάνυσμα score U ως μια συνάρτηση του $\underset{\rightarrow}{\beta}$ και του w_j. Η σχέση ανάμεσα στο $\underset{\rightarrow}{\beta}$ και το w_j δίνεται έμμεσα από την εξίσωση: $U(\underset{\rightarrow}{\beta}(w_j), w_j) = 0$. Παίρνοντας μερικές παραγώγους στη τελευταία σχέση θα έχουμε:

$$\frac{\partial U}{\partial \underset{\rightarrow}{\beta}} \frac{\partial \underset{\rightarrow}{\beta}}{\partial w_j} + \frac{\partial U}{\partial w_j} = 0 \Rightarrow \frac{\partial \underset{\rightarrow}{\beta}}{\partial w_j} = \left(-\frac{\partial U}{\partial \underset{\rightarrow}{\beta}} \right)^{-1} \frac{\partial U}{\partial w_j}$$

όπου ο όρος είναι ο πίνακας παρατηρούμενης πληροφορίας.

Για τον υπολογισμό του $\partial U / \partial w_j$ απαιτείται ο προσδιορισμός του τρόπου με τον οποίο τα βάρη εισάγονται στην συνάρτηση μερικής πιθανοφάνειας του Cox. Μια διόρθωση λοιπόν του τύπου της συνάρτησης μερικής πιθανοφάνειας είναι:

$$L(\underset{\rightarrow}{\beta}) = \prod_{i \in D} \left\{ \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \underset{\sim}{x}_i) \Big/ \sum_{k \in R_i} \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \underset{\sim}{x}_k) \right\}^{w_i}$$

όπου D συμβολίσαμε το σύνολο των μονάδων στις οποίες έχει συμβεί το γεγονός.

Η παράγωγος της λογαριθμοποιημένης μερικής πιθανοφάνειας ως προς $\underset{\rightarrow}{\beta}$ είναι το διάνυσμα score U

$$U = \sum_{i \in D} U_i = \sum_{i \in D} \left\{ w_i \underset{\rightarrow}{x}_i - w_i E(\underset{\rightarrow}{x} \mid R_i) \right\}$$

όπου:

$$E(\underset{\rightarrow}{x} \mid R_i) = \sum_{k \in R_i} w_k \underset{\rightarrow}{x}_k \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \underset{\rightarrow}{x}_k) \Big/ \sum_{k \in R_i} w_k \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \underset{\rightarrow}{x}_k)$$

Θέτοντας ως D_j το σύνολο των μονάδων στις οποίες συνέβη το γεγονός πριν ή ακριβώς τη χρονική στιγμή $t_{(j)}$, μπορούμε να υπολογίσουμε την παράγωγο του U ως προς w_j ως εξής:

$$\frac{\partial U}{\partial w_j} = \sum_{i \in D} \frac{\partial U_i}{\partial w_j} = \delta_j \left\{ \underset{\rightarrow}{x}_j - E(\underset{\rightarrow}{x} \mid R_j) \right\} - \sum_{i \in D_j} \frac{w_i \partial E(\underset{\rightarrow}{x} \mid R_i)}{\partial w_j} =$$

$$= \delta_j \hat{\underset{\rightarrow}{r}}_j - \sum_{i \in D_j} \frac{w_i \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \underset{\rightarrow}{x}_j)}{\sum_{k \in R_i} w_k \exp(\underset{\rightarrow}{\beta}^{\mathrm{T}} \underset{\rightarrow}{x}_k)} \left\{ \underset{\rightarrow}{x}_j - E(\underset{\rightarrow}{x} \mid R_i) \right\}$$

Β) Οι εντολές που χρησιμοποιήθηκαν

1) Για την προσαρμογή του μοντέλου του Cox χωρίς ποινή
   - model1<-coxph(Surv(time,status)~age+smear+absinf+labindex+absblasts+temp+treat)

2) Για την επιλογή βέλτιστου μοντέλου με τον αλγόριθμο backward elimination
   - model2<-step(model1,direction="backward", test="Chisq")

3) Για την ανασκόπηση των αποτελεσμάτων του βέλτιστου μοντέλου
   - summary(model2)

4) Για τα διαστήματα εμπιστοσύνης των β_j και $\exp(\beta_j)$
   - confint.deafault(model2)
   - exp(confint.deafault(model2))

5) Για τον έλεγχο της υπόθεσης της αναλογικής διακινδύνευσης του πίνακα 4.12
   - test.ph <- cox.zph(model2,transform='identity')

6) Για τα γραφήματα $(r_{ij}^{*} + \beta_i)$ ως προς $t_{(j)}$ για κάθε συμμεταβλητή
   - ggcoxzph(test.ph,var=1)
   - ggcoxzph(test.ph,var=2)
   - ggcoxzph(test.ph,var=3)
   - ggcoxzph(test.ph,var=4)
   - ggcoxzph(test.ph,var=5)

7) Για τον έλεγχο της υπόθεσης της αναλογικής διακινδύνευσης του πίνακα 4.13
   - test.ph <- cox.zph(model2)

8) Για την καμπύλη ROC του γραφήματος 4.18
   - status<-c
   - eta<-model2$linear.predictor
   - ROC6=risksetROC(Stime=time,status=c,marker=eta,predict.time=6,method="Cox",lty=2,col="red",ylab="True Positive",xlab="False Positive")
   - ROC30=risksetROC(Stime=time,status=c,marker=eta,predict.time=30,method="Cox",lty=3,col="darkblue",ylab="True Positive",xlab="False Positive",plot=FALSE)
   - lines(ROC30$FP,ROC30$TP, lty=3,col="darkblue")
   - legend(.6,.25,lty=c(2,3),col=c("red","darkblue"), legend=c("t=6","t=30"), bty="n")

9) Για το γράφημα 4.19 του AUC ως προς τον χρόνο
   - risksetAUC(Stime=time,status=c,marker=eta,method="Cox",tmax=40,main="AUC Curve",lty=2,col="darkblue")

10) Για την εύρεση των σημείων επιρροής του μοντέλου
   - dfbetas<-residuals(model2,type="dfbeta")
   - par(mfrow=c(2,2))
     for(j in 1:3){
   +plot(dfbetas[,j],ylab=names(coef(model2))[j],xlab="Individual",pch=16,col="red",main="Influence of Individual Cases",cex.lab=1.5)
   + abline(h=0, lty=2)
   + }

- par(mfrow=c(1,2))
  for(j in 4:5){
+
plot(dfbetas[,j],ylab=names(coef(model2))[j],xlab="Individual",pch=16,col="red",main="
Influence of Individual Cases",cex.lab=1.5)
+ abline(h=0, lty=2)
+ }

11) Για τη μελέτη του σημείου επιρροής
- which(dfbetas>0.28)
- a<-cbind(treat,time)
- max(a[treat==0,])
- mean(a[treat==0,])

12) Για την δημιουργία συντελεστών αλληλεπίδρασης ως προς τον χρόνο και τη αναπροσαρμογή του μοντέλου του Cox
- model3<-
  coxph(Surv(time,status)~age+smear+labindex+absblasts+treat+tt(treat)+tt(smear)+tt(
  absblasts),tt=function(x,t,...)x*t)

13) Για τα αποτελέσματα του πίνακα 4.15
- model4<-step(model3,direction="backward", test="Chisq")

14) Για τα αποτελέσματα του πίνακα 4.16
- summary(model4)

15) Για τον πίνακα 4.17 και το γράφημα 4.17
- mydataframe<-data.frame(cbind(treat,absblasts,absinf,temp,smear,labindex,age))
- correlation_table<-cor(mydataframe)
- corrplot(correlation_table)

16) Για την εξαίρεση όλων των μονάδων με μηδενικούς χρόνους διάρκειας ζωής
- x<-cbind(age,smear,absinf,labindex,absblasts,temp,treat)
- y<-cbind(time,status)
- which(y[,1]==0)
  [1] 22
- xnew<-rbind(x[1:21,],x[23:51,])
- ynew<-rbind(y[1:21,],y[23:51,])

17) Για τη μέθοδο cross-validation για την επιλογή του βέλτιστου λ της Lasso
- cv.lasso<-cv.glmnet(xnew,ynew,family=”cox”, alpha=1)
- cv.lasso$lambda.min
  [1] 0.1296947

18) Για το γράφημα 4.18
- plot(cv.lasso)

19) Για την εφαρμογή της μεθόδου Lasso και την εύρεση των παραμέτρων του μοντέλου
- fit.lasso<-glmnet(xnew,ynew,family=”cox”, alpha=1)
- coef(fit.lasso, s=cv.lasso$lambda.min)

20) Για την μέθοδο cross-validation για την επιλογή του βέλτιστου λ της Ridge
- cv.ridge<-cv.glmnet(xnew,ynew,family=”cox”, alpha=0)
- cv.ridge$lambda.min
  [1] 0.1327466

21) Για το γράφημα 4.19

- plot(cv.ridge)

22) Για την εφαρμογή της μεθόδου Ridge και την εύρεση των παραμέτρων του μοντέλου

- fit.ridge<-glmnet(xnew,ynew,family="cox", alpha=0)
- coef(fit.ridge, s=cv.ridge$lambda.min)

23) Δημιουργία ακολουθίας τιμών της παραμέτρου α

- alphasOfInterest<-seq(0.01,0.99,by=0.01)

24) Εκτέλεση της μεθόδου cross-validation για κάθε τιμή του α

- cvs<-lapply(alphasOfInterest, function(curAlpha){
  + cv.glmnet(xnew, ynew, alpha=curAlpha,family="cox")
  + })

25) Εύρεση του βέλτιστου λ για κάθε τιμή του α

- optimumPerAlpha<-sapply(seq_along(alphasOfInterest), function(curi){
  + curcvs<-cvs[[curi]]
  + curAlpha<-alphasOfInterest[curi]
  + indOfMin<-match(curcvs$lambda.min, curcvs$lambda)
  + c(lam=curcvs$lambda.min, alph=curAlpha, cvup=curcvs$cvup[indOfMin])
  + })

26) Εύρεση των βέλτιστων α και λ

- posOfOptimum<-which.min(optimumPerAlpha["lam",])
- overall.lambda.min<-optimumPerAlpha["lam",posOfOptimum]
- overall.alpha.min<-optimumPerAlpha["alph",posOfOptimum]

27) Για την εφαρμογή της μεθόδου Elastic Net για τα βέλτιστα α και λ και την εύρεση των παραμέτρων του μοντέλου

- fit.net<-glmnet(xnew,ynew,family="cox",alpha=overall.alpha.min)
- coef(fit.net,s=overall.lambda.min)

Βιβλιογραφία

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

Οικονόμου Π. και Καρώνη Χ. (2017), *Στατιστικά Μοντέλα Παλινδρόμησης με χρήση Minitab και R,* Εκδόσεις Συμεών, Αθήνα

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

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

Allison P. D. (1995), *Survival Analysis Using the SAS System: A Practical Guide*, Cary NC: SAS Institute

Anderson T. W., Darling D. A. (1952), Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes, *Annals of Mathematical Statistics*, **23**, 193–212

Belsley D.A., Kuh E.,Welsch R.E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, John Wiley, Hoboken, New Jersey

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

Buckland S.T., Burnham K.P. & Augustin N.H. (1997), Model selection: an integral part of inference, *Biometrics*, **53**, 603-618

Cain K.C. & Lange N.T. (1984), Approximate case influence for the proportional hazards regression model with censored data, *Biometrics*, **40**, 493-499

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

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

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

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

Cox D.R. & Snell E.J. (1989), *Analysis of Binary Data*, 2nd edition,Chapman & Hall, London

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

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

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

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

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

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

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

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

McFadden D. (1974), *Conditional logit analysis of qualitative choice behavior*. In: Zarembka P., editor, Frontiers in Economics, Academic Press, New York

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

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

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

Therneau T.M., Crowson C. & Atkinson E. (2017), Using time dependent covariates and time dependent coefficients in the Cox model. Survival Vignettes (http://cran.es.r-project.org/web/packages/survival/vignettes/timedep.pdf)

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

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

Verweij P. & van Houwelingen H. (1993), Cross-Validation in survival analysis, *Statistics in Medicine*, **12**, 2305-2314

Volinsky C.T. & Raftery A.E. (2000), Bayesian information criterion for censored survival models, *Biometrics*, **56**, 256-262

Xu R. et al. (2009), Using profile likelihood for semiparametric model selection with application to proportional hazards mixed models, *Statistica Sinica*, **19**, 819-842

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