ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΑΝΑΛΥΣΗ ΤΟΥ ΧΡΟΝΟΥ ΕΠΙΒΙΩΣΗΣ ΑΣΘΕΝΩΝ ΜΕ ΚΑΡΚΙΝΟ ΤΟΥ ΜΑΣΤΟΥ ΜΕΤΑ ΑΠΟ ΧΕΙΡΟΥΡΓΙΚΗ ΕΠΕΜΒΑΣΗ

Ερνίντα Μ. Μάρκο

Διπλωματική Εργασία
που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Μάρτιος 2011

UNIVERSITY OF PIRAEUS

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN APPLIED STATISTICS

SURVIVAL ANALYSIS FOR PATIENTS WITH BREAST CANCER AFTER SURGERY

By
Ernida M. Marko

MSc Dissertation
Submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfillment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
March 2011

Στην οικογένειά μου

Ευχαριστίες

Ευχαριστώ τον επιπλέποντα καθηγητή μου κ. Κωνσταντίνο Πολίτη για τις πολύτιμες συμβουλές και την καθοδήγήση του. Επίσης, ευχαριστώ τον καθηγητή κ. Αριστείδη Περπέρογλου για την ευγενική παραχώρηση των δεδομένων που πραγματεύτηκα, τα οποία προέρχονται από το αρχείο του νοσοκομείου ΙΑΣΩ.

Περίληψη

Η ανάλυση επιβίωσης αναπτύχθηκε ως μια ανάγκη να αναλύσουμε και να ερμηνεύσουμε δεδομένα που αφορούν χρόνους ζωής, κάποια από τα οποία είναι λογοκριμένα και για τα οποία δεν είναι γνωστές οι ακριβείς τιμές όλων των παραμέτρων.

Σκοπός της παρούσας εργασίας είναι η ανάλυση δεδομένων που χρησιμοποιούνται στην ιατρική στατιστική, με στόχο την μελέτη του χρόνου επιβίωσης σε γυναίκες με καρκίνο του μαστού. Η ανάλυση του χρόνου επιβίωσης των ασθενών θα βασιστεί κυρίως στο μοντέλο αναλογικού κινδύνου του Cox. Όμως προκειμένου να έχουμε μια πλήρη εικόνα της ανάλυσης επιβίωσης γενικά, αρχικά ορίζονται οι βασικές έννοιες που την θεμελιώνουν, γίνεται αναφορά στον Kaplan – Meier εκτιμητή και στα βασικά τεστ ελέγχου συναρτήσεων επιβίωσης.

Στην συνέχεια εξετάζουμε, τόσο θεωρητικά όσο και πρακτικά, το μοντέλο αναλογικού κινδύνου του Cox. Κατά τη διαδικασία αυτή, περιγράφονται και αναλύονται τεχνικές επιλογής μεταβλητών ώστε να εντοπίσουμε το καλύτερο μοντέλο για την ερμηνεία των δεδομένων μας. Έπειτα, ελέγχουμε αν το μοντέλο αυτό είναι επαρκές, ικανοποίει δηλαδή την υπόθεση αναλογικού κινδύνου. Η υπόθεση ελέγχεται μέσω γραφικών μεθόδων, Schoenfeld και Scaled Schoenfeld υπόλοιπα και μέσω στατιστικών ελέγχων.

Τέλος, στοχεύοντας σε ένα καλύτερο μοντέλο τόσο από άποψη προσαρμογής όσο και αναλογικότητας ορίζουμε και εφαρμόζουμε τα μοντέλα ευπάθειας (frailty models).

Η ανάλυση των δεδομένων γίνεται με εφαρμογή του προγράμματος εντολών στην **R**.

Abstract

Survival analysis was developed as a means to analyze and interpret data for lifetimes, when some of these lifetimes are censored and not all the exact values of the parameters for the underlying distribution are known. .

The purpose of this study is to analyze the data used in medical statistics, in order to study the survival time in women with breast cancer. The analysis of survival time of patients will rely primarily on Cox's proportional hazards model. But to get a complete picture of overall survival analysis for these data, first we define the basic concepts on which survival analysis is based on, while reference is made to Kaplan - Meier estimator and to the basic survival functions test.

Then we consider both theoretically and practically Cox's proportional hazards model. Variable selection techniques are described and analyzed in order to identify the best model for the interpretation of our data. Subsequently we check if the model is adequate, if it satisfies the proportional hazards assumption. This assumption is controlled by graphical methods, Schoenfeld and Scaled Schoenfeld residuals and through other statistical tests.

Finally, aiming at a better model both in terms of adaptation and proportionality we define and implement frailty models.

The data analysis was conducted with the implementation of the programme **R.**

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Γραφημάτων

Δομή της διπλωματικής εργασίας

Στην παρούσα διπλωματική εργασία εξετάζουμε το χρόνο επιβίωσης ασθενών με καρκίνο του μαστού ύστερα από επέμβαση αφαίρεσεις του όγκου. Οι μεταβλητές που εξετάζουμε είναι οι: grade (δηλώνει την διαφοροποίηση του όγκου, είναι μια κατηγορική μεταβλητή που παίρνει τέσσερεις τιμές την 0 αν δεν είναι γωστή η διαφοροποίηση, την 1 για μεγάλη διαφοροποίηση, την 2 για μεσαία διαφοροποίηση και την 3 για χαμηλή διαφοροποίηση. , την size (μέγεθος του όγκου σε mm), την positive (δηλώνει τον αριθμό των διηθημένων λεμφαδένων, διακριτή) και την other (δίτιμη, παίρνει την τιμή 1 αν ο θάνατος έχει προελθεί από άλλη αιτία εκτός καρκίνου).

Οι παραπάνω μεταβλητές αποτελούν τους βασικούς προγνωστικούς παράγοντες που συνήθως μπαίνουν σε ένα μοντέλο του cox.

Η διπλωματική εργασία αποτελείται από οκτώ κεφάλαια και παράρτημα όπου συγκεντρώνονται πίνακες αποτελεσμάτων και εντολές που εφαρμόζουμε στην R.

Συγκεκριμένα, στο **1ο κεφάλαιο** δίνεται μια εισαγωγή στο καρκίνο, όπου περιγράφονται εν συντομία τα βασικά χαρακτηριστικά του, τα είδη, τα στάδια του, τους παράγοντες που τον επηρεάζουν καθώς επίσης και οι συνήθες τεχνικές που εφαρμόζονται για την θεραπεία του.

Το **2ο κεφάλαιο** περιλμβάνει τη θεωρία της ανάλυση επιβίωσης, ορισμός συναρτήσεων επιβίωσης – κινδύνου, περιγράφονται οι σχέσεις με τις οποίες συνδέονται τα δύο μεγέθη. Ακόμη στο κεφάλαιο αυτό γίνεται αναφορά στα λογοκριμένα δεδομένα, τα οποία αποτελέσαν τον κυρίαρχο λόγο ανάπτυξης της ανάλυσης επιβίωσης καθώς σε διαφορετική περίπτωση θα εφραμόσαμε τις κλασσικές μεθόδους ανάλυσης δεδομένων επιβίωσης, περιγράφονται τα τρία βασικά είδη λογοκρισίας.

Στο **3ο κεφάλαιο** ορίζονται οι έννοιες των πινάκων επιβίωσης και προσδιορίζεται η χρήση τους στην ανάλυση επιβίωσης. Γίνεται εκτενής αναφορά στον Kaplan – Meier εκτιμητή, ορίζονται οι σχέσεις που χρησιμοποιούμε για να εκτιμήσουμε μη – παραμετρικά την συνάρτηση επιβίωσης δεξιά λογοκριμένων δεδομένων, αναφέρονται επίσης οι διάφορες σχέσεις εκτίμησης της διακύμανσης του εκτιμητή καθώς και το διάστημα εμπιστοσύνης της συνάρτησης επιβίωσης. Τέλος,

περιγράφονται τα πιο βασικά τεστ που χρησιμοποιούμε για να συγκρίνουμε τις συναρτήσεις επιβίωσης υπό ομάδων ενός δείγματος.

Το **4ο κεφάλαιο** περιγράφει και ορίζει θεωρητικά το μοντέλο αναλογικού κινδύνου του Cox. Μας δίνεται η δυνατότητα να εξετάσουμε αν ο χρόνος επιβίωσης ασθενών επηρεάζεται από μία ή περισσότερες επεξηγηματικές μεταβλητές. Ακόμη περιλαμβάνει την θεωρία των πιο συχνά χρησιμοποιούμενων τεχνικών επιλογής μεταβλητών. Στο τέλος του κεφαλαίου περιγράφεται εν συντομία το στρωματοποιημένο μοντέλο του καθώς και κάποιες γενικεύσεις του μοντέλου αναλογικού κινδύνου.

Στο **5ο κεφάλαιο** προχωράμε στην αξιολόγηση του μοντέλου αναλογικού κινδύνου τόσο γραφικά όσο και μέσω της ανάλυσης υπολοίπων. Συγκεκριμένα για την αξιολόγηση της υπόθεσης αναλογικού κινδύνου γραφικά εφαρμόζουμε το log(-logS(t)) διαγράμματα, παρόλο που η ισχύς τους αμφισβητείται, και τα διαγράμματα των υπολοίπων (π.χ. Schoenfeld residuals plot, scaled schoenfeled residualas plot) συναρτήσει κάποιας συναρτησιακής μορφής του χρόνου επιβίωσης. Αναφέρονται συνοπτικά τα είδη υπολοίπων που ορίζονται για το μοντέλο του Cox (Cox snell residuals, martingale, deviance, scaled schoenfeld κτλ.) και προσδορίζεται ο λόγος ύπαρξης τους. Μέσω των scaled schoenfeld υπολοίπων μπορούμε να αξιολογήσουμε στατιστικά την υπόθεση αναλογικού κινδύνου.

Το **6ο και το 7ο κεφάλαιο** αφορά την εφαρμογή των δεδομένων χρησιμοποιώντας το πρόγραμμα R, βασιζόμενη στην θεωρία που αναφέραμε στα 4 τελευταία κεφάλαια. Συγκεκριμένα το **6ο κεφάλαιο** περιλαμβάνει το κομάτι της θεωρίας που αναφέρεται πάνω στην ανάλυση επιβίωσης. Ενώ στο **7ο κεφάλαιο** παρουσιάζονται τα πιο σημαντικά αποτελέσματα από την εφαρμογή του μοντέλου αναλογικού κινδύνου του Cox.

Το **8ο κεφάλαιο** είναι ο επίλογος της διπλωματικής εργασίας, όπου συγκεντώνονται τα συμπεράσματα από το 6ο και 7ο κεφάλαιο.

Τέλος, έχουμε το παράρτημα όπου δίνονται οι εντολές που χρησιμοποιούμε στην R καθώς και γραφήματα και πίνακες που είναι λογότερο σημαντικά για την ανάλυσή μας και δεν προσθέτουν κάτι επιπλέον.

ΚΕΦΑΛΑΙΟ 1^{ο}

ΕΙΣΑΓΩΓΗ – ΓΝΩΡΙΖΟΝΤΑΣ ΤΟΝ ΚΑΡΚΙΝΟ

1.1 Γενικά για τον καρκίνο

Η προέλευση της λέξης Καρκίνος, αποδίδεται στον Έλληνα ιατρό Ιπποκράτη, που έμεινε στην ιστορία ως «πατέρας της ιατρικής». Ο Ιπποκράτης χρησιμοποίησε τους όρους «καρκίνος» και «καρκίνωμα» για να περιγράψει διάφορους όγκους που εμφάνιζαν εσωτερικά ή εξωτερικά έλκη και διογκώσεις. Αναφορές για την συγκεκριμένη νόσο έχουν βρεθεί πολύ νωρίτερα σε παπύρους στην αρχαία Αίγυπτο όπου αναφέρονται 8 περιπτώσεις όγκων ή ελκών στο στήθος οι οποίοι αντιμετωπίζονταν με καυτηριασμό, το λεγόμενο «τρυπάνι της φωτιάς», ακόμα στον πάπυρο αναφέρονταν ότι η νόσος δεν είχε θεραπεία.

Ο καρκίνος (όγκος) ορίζεται ως η ανώμαλη ανάπτυξη κύτταρων με αποτέλεσμα τη δημιουργία όγκων σε διάφορα σημεία του σώματος**.** Με τον όρο «καρκίνος» περιγράφεται μία ομάδα νοσημάτων, που η αιτία τους βρίσκεται σε κυτταρικό επίπεδο. Για να μπορέσουμε να τον κατανοήσουμε καλύτερα ακολουθεί μία μικρή αναφορά σχετικά με το τι συμβαίνει όταν φυσιολογικά κύτταρα μετατρέπονται σε καρκινικά.

Ο ανθρώπινος οργανισμός αποτελείται από κύτταρα. Φυσιολογικά, τα κύτταρα αναπτύσσονται και διαιρούνται, ώστε να προκύψουν θυγατρικά κύτταρα και να διατηρηθεί η υγεία του οργανισμού. Μερικές φορές, η διαδικασία αυτή εκτρέπεται από το φυσιολογικό, οπότε προκύπτουν νέα κύτταρα (χωρίς να τα χρειάζεται ο οργανισμός) και - παράλληλα - δεν πεθαίνουν τα παλιά κύτταρα. Τα πλεονάζοντα κύτταρα σχηματίζουν μάζες, που καλούνται όγκοι. Σε μερικές περιπτώσεις αυτά τα παθολογικά κύτταρα κάνουν μετάσταση, δηλαδή εξαπλώνονται και σε άλλα μέρη του σώματος δημιουργώντας δευτερεύοντες όγκους (μεταστατικούς όγκους) παρόμοιους με αυτούς του αρχικού καρκίνου.

Ο καρκίνος συνήθως δεν επηρεάζει μόνο ένα όργανο του σώματος και δεν έχει μία μορφή. Μπορεί να περιλαμβάνει οποιοδήποτε ιστό του σώματος και να έχει τελείως διαφορετική μορφή σε κάθε σημείο του σώματος. **Υπάρχουν πάνω από 200**

διαφορετικά είδη καρκίνου και το κάθε είδος έχει τον δικό του τρόπο θεραπευτικής αντιμετώπισης. Οι περισσότερες μορφές καρκίνου είναι όγκοι εκτός από ορισμένους τύπους καρκίνου όπως η λευχαιμία, των οποίων τα κύτταρα κυκλοφορούν μέσα στο αίμα και στα όργανα και τελικά αναπτύσσονται σε συγκεκριμένους ιστούς.

Οι καρκίνοι μπορούν τελικά να προκαλέσουν το θάνατο, εάν δεν θεραπευθούν, όμως η επιβίωση των ασθενών με καρκίνο τα τελευταία χρόνια χάρη στις επιστημονικές και τεχνολογικές ανακαλύψεις έχει βελτιωθεί σημαντικά. Παρόλα αυτά ο καρκίνος αποτελεί την δεύτερη αιτία θανάτου στις αναπτυγμένες χώρες μετά τα καρδιαγγειακά νοσήματα. (βλέπε Ιστοσελίδα 1)[1]

1.2 Καλοήθεις και κακοήθεις όγκοι

Οι όγκοι μπορεί να είναι **καλοήθεις** ή **κακοήθεις**.

Οι **καλοήθεις όγκοι** ή **καλοήθης νεοπλασία** , σχηματίζονται από πολλαπλασιαζόμενα κύτταρα εξαιτίας κάποιου ερεθισμού. Ωστόσο, όταν ο ερεθισμός αυτός πάψει να υφίσταται, η νεοπλασία διακόπτεται, ο όγκος δεν αναπτύσσεται περαιτέρω, δεν καταστρέφει περιβάλλοντες υγιείς ιστούς και δεν προκαλεί το σχηματισμό νέων όγκων σε άλλα σημεία του σώματος. Ακόμη, όταν ένας καλοήθης όγκος αφαιρείται από την περιοχή του οργανισμού στην οποία αναπτύχθηκε, **δεν παρατηρείται εκ νέου ανάπτυξή του**. Οι καλοήθεις όγκοι σπάνια προκαλούν σοβαρά προβλήματα σε έναν οργανισμό, δεν είναι επικίνδυνοι για τη ζωή του ατόμου και συνήθως αντιμετωπίζονται εύκολα.

Από την άλλη πλευρά **οι κακοήθεις όγκοι** γνωστοί και ως **νεοπλάσματα,** σχηματίζονται παρόμοια με τους καλοήθεις όγκους αλλά λειτουργούν τελείως διαφορετικά.

Τα πολλαπλασιαζόμενα κύτταρα φέρουν μια παθογένεια που προκαλεί τη συνεχή και ανεξέλεγκτη αύξησή τους. Η **κακοήθης νεοπλασία** - η διαδικασία του άναρχου

πολλαπλασιασμού των κυττάρων που καταλήγει σε σχηματισμό νέων ιστών – εξακολουθεί να δρα, ακόμη και όταν ο αρχικός όγκος αφαιρεθεί από το σώμα,

[1] Οι ιστοσελίδες που αναφέρονται στο παρόν κεφάλαιο βρίσκονται στο τέλος της διπλωματικής , στη βιβλιογραφία.

σχηματίζοντας στη θέση του ένα καινούργιο. Ταυτόχρονα, επιδρά καταστροφικά στους γειτνιάζοντες ιστούς, στους οποίους εισβάλει (διαδικασία που ονομάζεται *διήθηση*) και έχει τη δυνατότητα σχηματισμού νέων όγκων σε άλλα σημεία του σώματος. Η διαδικασία αυτή της μετακίνησης των καρκινικών κυττάρων μέσω του αίματος ή της λέμφου στο υπόλοιπο σώμα ονομάζεται **μετάσταση.** Οι κακοήθεις όγκοι σχηματίζονται από κύτταρα που είναι ουσιαστικά άχρηστα για τον οργανισμό, καθώς δεν αντικαθιστούν φθαρμένους ή κατεστραμμένους ιστούς. Δυστυχώς, όμως, εκτός από άχρηστα είναι και επιβλαβή καθώς είναι εν δυνάμει θανατηφόρα, ειδικά αν δεν διαγνωστούν έγκαιρα.

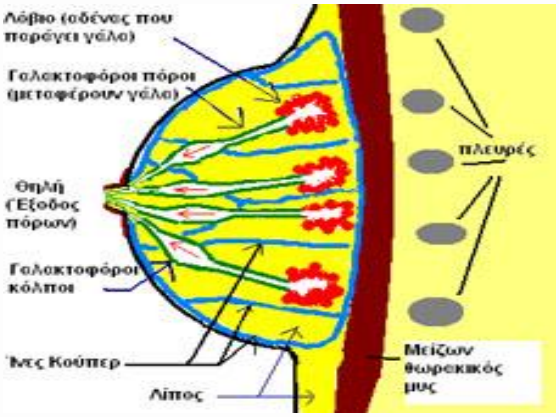
Να αναφέρουμε ότι οι περισσότεροι καρκίνοι παίρνουν το όνομά τους από τον τύπο του κυττάρου ή του οργάνου από το οποίο αρχίζουν. Αν κάνουν μετάσταση ο νέος όγκος φέρει το ίδιο όνομα με τον αρχικό. Ορισμένοι όγκοι παίρνουν το όνομά τους από τον επιστήμονα που τους ανακάλυψε (π.χ. Hodgkin, Brenner).

(Ιστοσελίδες 1 & 2)

1.3 Καρκίνος του μαστού

Ο μαστός είναι ένας αδένας σχεδιασμένος για να παράγει γάλα. Οι λοβοί στο μαστό παράγουν το γάλα, το οποίο στη συνέχεια μεταφέρεται μέσω των πόρων στη θηλή.

Ο καρκίνος του μαστού είναι ο κακοήθης όγκος που προκύπτει από ανεξέλεγκτο πολλαπλασιασμό των κυττάρων του μαστού. Εμφανίζεται κυρίως στους πόρους, τα σωληνάρια που μεταφέρουν το γάλα στη θηλή κατά τη διάρκεια του θηλασμού, αλλά και στους λοβούς. Ο καρκίνος του μαστού μπορεί να αναπτυχθεί σε οποιαδήποτε ηλικία αλλά ο κίνδυνος ανάπτυξής του αυξάνεται με την ηλικία της γυναίκας.



(Πηγή εικόνας: Γ.Π. Φύσσας –Ο μαστός και οι παθήσεις του–Απλή ανατομία και λειτουργία του μαστού)

Ο καρκίνος του μαστού είναι μιά μάστιγα για το δυτικό κόσμο. Είναι ο συχνότερος καρκίνος στις γυναίκες και προσβάλλει κάθε χρόνο πολλά εκατομμύρια σε όλο τον πλανήτη. Στην Ελλάδα εμφανίζονται περίπου 4500 νέες περιπτώσεις καρκίνου του μαστού το χρόνο, συγκεκριμένα εκτιμάται ότι κάθε μία εργάσιμη ώρα μία Ελληνίδα ανακαλύπτει ότι έχει καρκίνο του μαστού και κάθε έξι ώρες μία άλλη χάνει τη ζωή της εξαιτίας της νόσου. Η μαστογραφία βοηθά στην έγκαιρη διάγνωση και είναι ένας κρίσιμος παράγοντας σωτηρίας των ασθενών, καθώς μπορεί να εντοπίσει τον όγκο μόλις στο 1.1 εκατοστό όταν ο τυχαίος εντοπισμός του γίνεται συνήθως σε υπερτριπλάσιο μέγεθος 3.6 εκατοστά.. Σήμερα είναι γνωστό ότι αν ένας καρκίνος του μαστού βρεθεί σε μία προληπτική μαστογραφία πριν γίνει επιθετικός (διηθητικός) και μάλιστα πριν καν γίνει ψηλαφητός και αφαιρεθεί χειρουργικά (για το συγκεκριμένα στάδιο λέμε ότι ο καρκίνος είναι in situ) τότε γιατρεύεται οριστικά. Αυτό δεν συμβαίνει για τις εξαρχής πολύ επιθετικές μορφές καρκίνου, για τις οποίες ούτως ή άλλως δεν μπορεί η ιατρική να προσφέρει πολλά. Επομένως οι πιθανότητες ίασης συνδέονται άμεσα με το χρόνο που μεσολαβεί μέχρι την διάγνωση της ασθενείας, καθώς ο τρόπος που δρα ο καρκίνος είναι σιωπηλός και τις περισσότερες φορές τα πρώτα συμπτώματα εμφανίζονται σε αρκετά προχωρημένο στάδιο. Δυστυχώς όμως στην Ελλάδα το ποσοστό των κρουσμάτων του καρκίνου του μαστού που διαγιγνώσκεται σε πρώιμο στάδιο μόλις και μετά βίας προσεγγίζει το 5% σε αντίθεση με την Ευρώπη που είναι περίπου το 60%. Αυτό είναι ακόμα πιο στενάχωρο όταν το πενταετές ποσοστό επιβίωσης σε περιπτώσεις διάγνωσης σε πρώιμο στάδιο φθάνει ως και το 95%, στοιχείο που υποδηλώνει πως ο καρκίνος του μαστού μπορεί να αντιμετωπιστεί επιτυχώς για την πλειονότητα των γυναικών που φροντίζουν να τον εντοπίσουν έγκαιρα.

Όσον αφορά τα ποσοστά θανάτου από καρκίνο του μαστού αυτά χαρακτηρίζονται από πτωτική τάση από τις αρχές του 1990, με τις μεγαλύτερες μειώσεις να εντοπίζονται στις γυναίκες κάτω των 50. Οι ερευνητές αποδίδουν την πτώση αυτή στην έγκαιρη διάγνωση μέσω μαστογραφιών καθώς και στις βελτιώσεις που έχουν επέλθει στις σχετικές θεραπευτικές αγωγές.

(Ιστοσελίδες 3, 4, 5 & 6 , και Φύσσας 2006)

1.4 Παράγοντες που επηρεάζουν τον καρκίνο του μαστού

Πληθώρα μελετών έχουν προσδιορίσει ορισμένους παράγοντες οι οποίοι συντελούν στην αύξηση της πιθανότητας εμφάνισης κακοήθους νεοπλασίας στο μαστό.

Κάποιοι από τους παράγοντες αυτούς αναφέρονται παρακάτω:

- **Κληρονομικότητα,** εκτιμάται πως περίπου το 10% των κρουσμάτων καρκίνου του μαστού σχετίζεται με παράγοντες κληρονομικότητας. Η πλειονότητα των κρουσμάτων προκύπτει από βλάβες στο γενετικό υλικό των μαστικών κυττάρων, που προκαλούνται από διάφορους παράγοντες κατά τη διάρκεια της ζωής ενός ατόμου. Συγκεκριμένα, δύο γονίδια, γνωστά ως BRCA 1 και BRCA 2, έχουν προσδιοριστεί ως παράγοντες που συμβάλλουν στην εμφάνιση καρκίνου του μαστού όταν έχουν υποστεί γενετική αλλοίωση. Έτσι, γυναίκες με εξ αίματος συγγενείς που έχουν νοσήσει αντιμετωπίζουν **αυξημένο κίνδυνο εμφάνισης καρκίνου στο μαστό.**

  Το ποσοστό κινδύνου εμφανίζεται αντιστρόφως ανάλογο με την ηλικία εμφάνισης της νόσου: όσο νεότερο ήταν το μέλος της οικογένειας όταν νόσησε, τόσο μεγαλύτερο το ποσοστό κινδύνου για τις υπόλοιπες γυναίκες της οικογένειας. Το ίδιο σημαντικός φαίνεται πως είναι και ο αριθμός των κρουσμάτων στην οικογένεια: όσο περισσότερα τα κρούσματα, τόσο μεγαλύτερος είναι και ο κίνδυνος για τα υπόλοιπα μέλη.

- **Ηλικία,** υπάρχουν διαθέσιμα στοιχεία που αποδεικνύουν πως ο **καρκίνος του μαστού** μπορεί να προκύψει σε οποιαδήποτε ηλικία μετά την εφηβεία. Ωστόσο, τα ποσοστά εμφάνισής του αυξάνονται όσο ανεβαίνουμε τις ηλικιακές κλίμακες. Κατά συνέπεια γυναίκες σε προχωρημένη ηλικία έχουν μεγαλύτερο κίνδυνο από τις νεαρές, αλλά οι νεότερες που έχουν προσβληθεί από καρκίνο του μαστού έχουν ελαφρώς χειρότερη πρόγνωση.

- **Ηλικία εμμηναρχής – εμμηνόπαυσης,** η πρώιμη έναρξη της έμμηνης ρήσης (πριν από το $12^{ο}$ έτος της ηλικίας τους) και η καθυστερημένη εμμηνόπαυση (μετά τα 55) φαίνεται να συνδέονται με την αυξημένη πιθανότητα προσβολής από τη νόσο. Ακόμη, η λήψη οιστρογόνων μετά την εμμηνόπαυση έχει συσχετιστεί με αυξημένα ποσοστά εμφάνισης της νόσου, με τον κίνδυνο να είναι ανάλογος του διαστήματος λήψης των οιστρογόνων.

- **Τεκνοποίηση,** μελέτες δείχνουν πως οι άτεκνες γυναίκες, ή οι γυναίκες που δεν είχαν πλήρεις κυήσεις ή γέννησαν μετά τα 35 τους χρόνια έχουν μεγαλύτερο κίνδυνο εμφάνισης της νόσου σε σχέση με τις πολύτεκνες. Ακόμη προκύπτει πως ο τοκετός σε νεαρή ηλικία (μικρότερη των 25 ετών) είναι ένας σημαντικός προασπιστικός παράγοντας εναντίον του καρκίνου του μαστού για όλη την μετέπειτα ζωή της γυναίκας.
- **Θηλασμός,** σε αντίθεση με ότι πίστευαν στο παρελθόν για το θηλασμό και την προστασία του από το καρκίνο στοιχεία έχουν δείξει ότι ο θηλασμός δεν παίζει προστατευτικό ρόλο έναντι του καρκίνου και ούτε έχει σχέση αν η ίδια γυναίκα έχει θηλάσει από την μητέρα της.
- **Παχυσαρκία και αύξηση του σωματικού βάρους,** οι παχύσαρκες γυναίκες έχουν μεγαλύτερο κίνδυνο σε σχέση με αυτές κανονικού βάρους. Συγκεκριμένα, σημαντικό ρόλο παίζει η ηλικία στην οποία προσλαμβάνονται τα κιλά: όσο μεγαλύτερη ηλικία, τόσο μεγαλύτερος ο κίνδυνος. Η αύξηση του βάρους μετά την ηλικία των 20 ετών αυξάνει αναλογικά τον κίνδυνο εμφάνισης της νόσου και ο κίνδυνος αυτός τριπλασιάζεται εάν ο δείκτης μάζας σώματος βρίσκεται στα μέγιστα επίπεδά του μετά την ηλικία των 50 ετών**.**
- **Κάπνισμα,** σύμφωνα με στοιχεία πρόσφατων ερευνών η κατανάλωση ενός πακέτου τσιγάρων ημερησίως, από γυναίκες προ της εμμηνόπαυσης και για εννέα περίπου χρόνια, αυξάνει δραστικά τον κίνδυνο εμφάνισης του καρκίνου του μαστού. Η διακοπή του καπνίσματος έχει αποδειχτεί πως είναι σε θέση να μειώσει το σχετικό κίνδυνο.
- **Ψυχολογικοί παράγοντες,** και κοινωνικές συνθήκες ζωής μπορούν να επηρεάσουν τη σωματική υγεία. Διάφοροι ψυχολογικοί παράγοντες όπως άγχη, ψυχικές διαταραχές, μοναξιά, φαίνονται να συνδέονται με την εμφάνιση του καρκίνου του μαστού, αν και η ψυχική ένταση δεν μπορεί εύκολα να μετρηθεί και να ποσοτικοποιηθεί.
- **Διατροφή,** πολλές μελέτες έχουν δείξει το ρόλο που παίζει η διατροφή στην ανάπτυξη καρκίνου του μαστού. Το κίνδυνο εμφάνισης φαίνονται να επηρεάζουν ιδιαίτερα τροφές πλούσιες σε ζωικό λίπος και πρωτεΐνες.
- **Έκθεση του μαστού σε ακτινοβολία,** η ιονίζουσα ακτινοβολία έχει προσδιοριστεί ως παράγοντας αύξησης του κινδύνου εμφάνισης της νόσου,

ιδιαίτερα όταν εκτίθεται η γυναίκα σε μικρή ηλικία (βέβαια η δόση ακτινοβολίας για διαγνωστικούς σκοπούς δεν συντελεί στην αύξηση του κινδύνου και για αυτό δεν θα πρέπει να αποφεύγεται).

- **Προηγούμενος καρκίνος της μήτρας ή των ωοθηκών,** στοιχεία υποδηλώνουν πως γονιδιακές μεταλλάξεις που προκαλούν καρκίνο των ωοθηκών και της μήτρας είναι σε θέση να προκαλέσουν κακοήθη νεοπλασία και στο μαστό.

(Ιστοσελίδες 3 και 6)

1.5 Θεραπευτικές μέθοδοι

Η θεραπεία που θα επιλεγεί για την αντιμετώπιση μιας κακοήθους νεοπλασίας στο μαστό εξαρτάται τόσο από τον τύπο όσο και το βαθμό του καρκίνου, καθώς και από τη γενικότερη βιολογική και σωματικής κατάσταση της ασθενούς: παράγοντες όπως η ηλικία, το οικογενειακό ιστορικό και μια πιθανή εγκυμοσύνη διαδραματίζουν σημαντικό ρόλο στην επιλογή θεραπευτικής αγωγής. Υπάρχουν διαθέσιμες αρκετές θεραπευτικές επιλογές και πολλές φορές χρησιμοποιούνται ένα ή περισσότερα θεραπευτικά σχήματα ώστε να παράσχουν πιο ολοκληρωμένη θεραπεία στην ασθενή. Ακολουθούν ορισμένες από τις τυπικές επιλογές αντιμετώπισής του καρκίνου στο μαστό:

1. Ογκεκτομή απλή ή ακολουθούμενη από ακτινοθεραπεία

Η ογκεκτομή περιλαμβάνει την εκτομή - δηλαδή την αφαίρεση του όγκου και των περιβαλλόντων ιστών - και όχι ολόκληρου του μαστού. Με τον τρόπο αυτό διασώζεται το μεγαλύτερο μέρος του μαστικού αδένα. Ο συνδυασμός ογκεκτομής και ακτινοθεραπείας αποτελεί την πιο κοινή και συνήθη θεραπευτική αντιμετώπιση. Να αναφέρουμε πως ο περιβάλλων ιστός που αφαιρείται περιλαμβάνει στρώματα υγιών κυττάρων, εξασφαλίζοντας με τον τρόπο αυτό πως δεν παραμένουν καθόλου καρκινικά κύτταρα στο μαστό

Συχνά, την εγχείριση ακολουθεί ακτινοθεραπεία, η οποία έχει υπολογιστεί πως μειώνει τον κίνδυνο επανεμφάνισης του όγκου κατά 60%. Το δεύτερο αυτό στάδιο αντιμετώπισης της νόσου μπορεί να εφαρμοστεί είτε εξωτερικά είτε εσωτερικά:

i. Στην περίπτωση **εξωτερικής ακτινοθεραπείας** οι ακτίνες κατευθύνονται στο στήθος μέσω ενός γραμμικού επιταχυντή, έξω από το σώμα της ασθενούς. Η θεραπεία διαρκεί συνήθως μερικές εβδομάδες (το πολύ επτά) και περιλαμβάνει 5 περίπου εφαρμογές την εβδομάδα.

ii. Στην περίπτωση της **εσωτερικής ακτινοβολίας**, μία μικρή πηγή ακτινών εισάγεται στην παθούσα περιοχή, προσβάλλοντας μόνον αυτή και όχι ολόκληρο το μαστό. Με αυτό το τρόπο το επιθυμητό αποτέλεσμα επιτυγχάνεται σε πιο σύντομο χρονικό διάστημα (5 – 10 ημέρες) και με λιγότερες παρενέργειες.

2. Μαστεκτομή

Μαστεκτομή ονομάζεται η ολική χειρουργική αφαίρεση του στήθους. Αυτή η θεραπευτική προσέγγιση προτιμάται σε περιπτώσεις κατά τις οποίες τα καρκινικά κύτταρα ενδέχεται να έχουν εξαπλωθεί εκτενώς ή χαρακτηρίζονται από πολλαπλές εστίες, και σε περιπτώσεις βεβαρυμμένου οικογενειακού ιστορικού ή ύπαρξης γονιδίων που αποδεδειγμένα προκαλούν καρκινογένεση στο στήθος.

Η μαστεκτομή ως χειρουργική επέμβαση μπορεί να διαφέρει στην έκταση και την εφαρμογή της:

i. *Υποδόριος μαστεκτομή:* συνίσταται στην αφαίρεση ολόκληρου του παθογόνου μαστού, ακόμη και της θηλής, δίχως ωστόσο την απομάκρυνση του περιβάλλοντος δερματικού ιστού

ii. *Απλή μαστεκτομή* ονομάζεται η αφαίρεση του συνόλου του παθογόνου αδένα, δηλαδή του μαστού, καθώς και του περιβάλλοντος δέρματος.

iii. *Ολική ριζική μαστεκτομή* είναι η διαδικασία κατά την οποία αφαιρούνται το σύνολο του μαστού και του περιβάλλοντος δέρματος, οι γειτνιάζοντες θωρακικοί μύες όπισθεν του μαστού καθώς και οι λεμφαδένες της μασχαλιαίας περιοχής.

iv. *Συντηρητική μερική μαστεκτομή,* περιλαμβάνει την εκτομή του όγκου και των γειτονικών ιστών, αλλά όχι των λεμφαδένων της μασχάλης.

v. *Τροποποιημένη ριζική μαστεκτομή* είναι η ριζική μαστεκτομή που δεν περιλαμβάνει αφαίρεση των θωρακικών μυών όπισθεν του μαστού.

3. **Ορμονική θεραπεία**

Οι αγωγές αυτές στοχεύουν στη μείωση των παραγόμενων οιστρογόνων στο σώμα της ασθενούς προκειμένου τα καρκινικά κύτταρα να μην έχουν την απαραίτητη ορμόνη που χρειάζονται για να ξεκινήσει η ανάπτυξη και ο πολλαπλασιασμός τους, δηλαδή τα οιστρογόνα. Εφαρμόζεται σε περιπτώσεις ασθενών που έχουν κάνει εξέταση ορμονικών υποδοχέων και τα αποτελέσματα ήταν θετικά. Ορισμένες από τις αγωγές αυτές περιλαμβάνουν τη έκχυση ορμονικών υποκατάστατων που προσκολλώνται στους υποδοχείς ορμονών, καταλαμβάνοντας τη θέση των οιστρογόνων. Άλλες ουσίες στοχεύουν στη γενικότερη μείωση της παραγωγής οιστρογόνων στο σώμα της ασθενούς.

4 . **Χημειοθεραπεία**

Η χημειοθεραπεία περιλαμβάνει τη χρήση φαρμακευτικών ουσιών που αποσκοπούν στην καταστροφή των καρκινικών κυττάρων. Όταν έπεται της χειρουργικής επέμβασης ονομάζεται «συμπληρωματική θεραπεία». Όταν, ωστόσο, ο όγκος είναι πολύ μεγάλος και καρκινικά κύτταρα έχουν προσβάλει σοβαρά τους λεμφαδένες και έχουν διασπαρθεί και σε άλλα μέρη του σώματος, η χημειοθεραπεία προηγείται της επέμβασης και αποσκοπεί στη συρρίκνωση του όγκου και στην εξόντωση των διασπαρμένων καρκινικών κυττάρων. Οι φαρμακευτικές χημειοθεραπευτικές ουσίες διοχετεύονται στο σώμα με τους εξής τρόπους:

- Στοματικά, μέσω χαπιού ή πόσιμου υγρού.
- Μέσω ενέσεων.
- Ενδοφλέβια (η μέθοδος αυτή χρησιμοποιείται συνήθως στην πράξη)

Οι χημειοθεραπευτικές ουσίες διατρέχουν το κυκλοφορικό σύστημα και καταστρέφουν όποια καρκινικά κύτταρα έχουν διασπαρθεί. Δυστυχώς, όμως, συχνά προσβάλλουν και υγιή κύτταρα, με αποτέλεσμα ορισμένες δυσάρεστες παρενέργειες, οι περισσότερες από τις οποίες, ωστόσο, είναι αντιμετωπίσιμες.

Η χημειοθεραπεία συστήνεται όταν το διηθητικό καρκίνωμα είναι αρκετά μεγάλο (μεγαλύτερο του ενός εκατοστού) ή έχει επεκταθεί και στους λεμφαδένες. Διαρκεί,

συνήθως, 3 με έξι μήνες και η φύση της εξαρτάται από τη σοβαρότητα της κάθε περίπτωσης και από τη φυσιολογία της ασθενούς, την ηλικία της, τη γενικότερη κατάσταση της υγείας της κτλ.

Πρόσφατα έχουν δημιουργηθεί δύο νέες εξετάσεις οι οποίες είναι σε θέση να βοηθήσουν τόσο το γιατρό όσο και την ασθενή να προσδιορίσουν το αν και κατά πόσον απαιτείται χημειοθεραπεία για την αντιμετώπιση της νόσου: οι εξετάσεις αυτές είναι γνωστές ως Ογκότυπος DX and MammaPrint και είναι σε θέση να εντοπίσουν γονίδια που έχει αποδειχτεί πως συσχετίζονται με την επανεμφάνιση της νόσου.

1.5.1 Μετά τη Θεραπεία

Μετά την αγωγή, θα ακολουθήσει μια περίοδος επιφυλακής, κατά την οποία η υγεία του οργανισμού της πρώην ασθενούς και η κατάσταση του μαστού θα εξετάζονται περιοδικά και συστηματικά.

Κατά τα πρώτα 5 χρόνια μετά το πέρας της θεραπείας η ενδιαφερόμενη θα πρέπει να επισκέπτεται το γιατρό της ανά 4 ή 6 μήνες για ειδικές εξετάσεις. Μόλις ολοκληρωθεί η πρώτη πενταετία, οι επισκέψεις αυτές θα γίνουν ετήσιες. Πρώην ασθενής που έχει υποστεί ογκεκτομή ή επέμβαση διατήρησης του στήθους θα πρέπει να κάνει μαστογραφία 6 μήνες μετά το πέρας της ακτινοθεραπείας, και στη συνέχεια μία φορά κάθε χρόνο. Σε περίπτωση μαστεκτομής, η μαστογραφία πρέπει να γίνεται μία φορά το χρόνο.

Φυσικά, η κάθε περίπτωση είναι διαφορετική. Ο θεράπων ιατρός είναι εκείνος που θα καθορίσει τη συχνότητα και το είδος των προληπτικών εξετάσεων, λαμβάνοντας υπόψη παράγοντες όπως η κληρονομικότητα, η ηλικία, η σοβαρότητα του αντιμετωπισθέντος όγκου κτλ. Αφού τα χαρακτηριστικά του καρκίνου μπορούν να επηρεάσουν τις θεραπευτικές επιλογές καθώς και την πιθανότητα επανεμφάνισής του.

(Ιστοσελίδες 6, 7 & 8)

1.6 Στάδια καρκίνου του μαστού

Ο καρκίνος του μαστού διακρίνεται σε πέντε στάδια, αρχίζοντας από το πιο ελαφρύ και προχωρώντας προς βαρύτερες καταστάσεις. Τα στάδια είναι γνωστά και αποδεκτά παγκοσμίως και βοηθούν:

- να καταλάβουν οι ασθενείς καλύτερα την πρόγνωσή τους και την πιθανή έκβαση της ασθένειας,
- να επιλεγούν οι κατάλληλες θεραπείες αντιμετώπισης του καρκίνου,
- να παραχθεί ένας κοινός τρόπος για να περιγραφεί η έκταση του καρκίνου του μαστού στους γιατρούς και τις νοσοκόμες σε όλο τον κόσμο, έτσι ώστε τα αποτελέσματα των εξετάσεων και της θεραπείας να μπορούν να συγκριθούν και κατανοηθούν.

Στάδιο 0 (in situ)

Αυτό το στάδιο χρησιμοποιείται για να περιγράψει έναν ακίνδυνο και ιάσιμο καρκίνο του μαστού (*πενταετής επιβίωση >95%)*. Οι καρκίνοι αυτοί είναι μικροί και ενδέχεται να αποκαλυφθούν τυχαία με την μαστογραφία πριν καν γίνουν ψηλαφητοί. Στο στάδιο αυτό δεν υπάρχει κανένα στοιχείο ότι έχουν διαλυθεί τα κύτταρα του καρκίνου ή ότι έχει εισβάλλει σε γειτονικούς υγιείς ιστούς.

Στάδιο I

Αυτό το στάδιο περιγράφει την παρουσία του καρκίνου στο στήθος (κύτταρα έχουν διαλυθεί από τον καρκίνο ή εισβολή καρκίνου σε γειτονικούς υγιείς ιστούς) κατά την οποία: το μέγεθος του όγκου να είναι μέχρι δύο εκατοστά, και κανένας λεμφαδένας δεν έχει προσβληθεί. Το στάδιο αυτό τις περισσότερες φορές είναι ιάσιμο χειρουργικά (*πενταετής επιβίωση περίπου 85%).*

Στάδιο II

Αυτό το στάδιο περιγράφει την εμφάνιση καρκίνου του μαστού κατά την οποία: το μέγεθος του όγκου είναι τουλάχιστον δύο εκατοστά, αλλά όχι πάνω από πέντε εκατοστά, επίσης στο στάδιο αυτό ο καρκίνος έχει επεκταθεί και στους λεμφαδένες κάτω από την μασχάλη στην ίδια πλευρά με τον καρκίνο του μαστού. Οι προσβληθέντες λεμφαδένες δεν έχουν κολλήσει ακόμη ο ένας στον άλλο ή στους περιβάλλοντες ιστούς, ένα σημάδι που δείχνει ότι ο καρκίνος δεν έχει προωθηθεί ακόμα στο στάδιο III (*πενταετής επιβίωση περίπου 66%*).

Στάδιο ΙΙΙ

Το χαρακτηριστικό του σταδίου ΙΙΙ είναι η ύπαρξη μεγάλου όγκου με προσβολή των σχετικών μασχαλιαίων λεμφαδένων που κολλάνε μεταξύ τους ή με γειτονικούς ιστούς. Το στάδιο ΙΙΙ διαιρείται σε υποκατηγορίες γνωστές ως ΙΙΙΑ και ΙΙΙΒ.

Στάδιο ΙΙΙΑ

Αυτό το στάδιο περιγράφει την ύπαρξη του καρκίνου του μαστού κατά την οποία το μέγεθος του όγκου είναι παραπάνω από πέντε εκατοστά, ή ο όγκος έχει επεκταθεί και στους λεμφαδένες, οι οποίοι συγκεντρώνονται ή κολλούν ο ένας πάνω στον άλλον ή στους περιβάλλοντες ιστούς. Το στάδιο αυτό είναι γενικά χειρουργήσιμο και θεραπεύεται επιθετικά (χειρουργικά, με ακτινοβολία και χημειοθεραπεία με ποικίλη σειρά).

Στάδιο ΙΙΙΒ

Αυτό το στάδιο περιγράφει την εισβολή του καρκίνου του μαστού κατά την οποία ένας όγκος (οποιουδήποτε μεγέθους) έχει επεκταθεί στο δέρμα των μαστών, στο θωρακικό τοίχωμα ή στους εσωτερικούς μαστικούς λεμφαδένες και περιλαμβάνει φλεγμονώδη καρκίνο του μαστού. Ο φλεγμονώδης καρκίνος του μαστού είναι ένα πολύ ασυνήθιστο αλλά πολύ σοβαρό, επιθετικό είδος καρκίνου του μαστού, ιδιαίτερο χαρακτηριστικό γνώρισμα του είναι η κοκκινίλα που περιβάλλει ένα μέρος του στήθους ή όλο το στήθος. Το στάδιο αυτό θεωρείται γενικά μη χειρουργήσιμο. Η χειρουργική περιορίζεται τις περισσότερες φορές στην αρχική διαγνωστική βιοψία και η θεραπεία περιλαμβάνει χημειοθεραπεία ή ακτινοβολία (*το πενταετές ποσοστό επιβίωσης είναι περίπου ίσων με 41 %*).

Στάδιο IV (Προχωρημένος καρκίνος του μαστού)

Αυτό το στάδιο περιλαμβάνει την εισβολή καρκίνου του μαστού κατά την οποία ένας όγκος έχει επεκταθεί πέρα από το στήθος, στην μασχάλη και στους εσωτερικούς μαστικούς λεμφαδένες. Ο όγκος μπορεί να έχει εξαπλωθεί και περισσότερο, όπως στη βάση του λαιμού, στους πνεύμονες, στο συκώτι, στα κόκαλα, ή στον εγκέφαλο.

Στο στάδιο αυτό η χειρουργική περιορίζεται στη βιοψία για να επιβεβαιωθεί ο κυτταρικός τύπος του όγκου και αν υπάρχουν ορμονικοί υποδοχείς. Προτεραιότητα στο συγκεκριμένο στάδιο έχει η ανάγκη να ελεγχθούν οι μεταστάσεις (*πενταετής επιβίωση γύρω στο 10%*).

(Ιστοσελίδα 9)

1.7 Ο καρκίνος του μαστού στους άντρες

Είναι πλέον αποδεκτό και γνωστό πως και οι άνδρες παρόλο που δεν έχουν μαστό μπορούν να αναπτύξουν καρκίνο του μαστού. Η συχνότητα του καρκίνου του μαστού στους άνδρες είναι πολύ πιο χαμηλή από ότι στις γυναίκες (σχετική συχνότητα εμφάνισης 1% δηλ. σε 100 γυναίκες με καρκίνο του μαστού αντιστοιχεί 1 άνδρας), συγκεκριμένα στις ΗΠΑ εκδηλώνονται κάθε χρόνο 1500 περιστατικά και εκτιμάται ότι από τη νόσο πεθαίνουν κάθε χρόνο 400 άντρες.

Δυστυχώς όμως ο καρκίνος του μαστού του άνδρα σε σχέση με αυτόν των γυναικών έχει συνήθως χειρότερη έκβαση για τους παρακάτω λόγους:

α) ο άνδρας δεν πάει στο γιατρό αμέσως μόλις διαπιστώσει κάποια ανωμαλία στο μαστό.

β) αργεί να διαγνωστεί διότι ασθενείς και γιατροί δεν τον σκέφτονται εύκολα

γ) λόγω καθυστερημένης διάγνωσης και λόγω του μικρού μεγέθους του μαστού, έχουμε πιο συχνή διήθηση του δέρματος ή του μυός που βρίσκεται πίσω από το μαστό, με αποτέλεσμα τη διευκόλυνση των μεταστάσεων.

1.7.1 Παράγοντες που συντελούν στην ανάπτυξη του καρκίνου

Η αιτιολογία της εμφάνισης καρκίνου του μαστού στον άνδρα δεν έχει ακόμα πλήρως κατανοηθεί, έχουν ωστόσο από μελέτες προσδιοριστεί παράγοντες που αυξάνουν τον κίνδυνο ανάπτυξης του καρκίνου στο μαστό.

Ως τέτοιοι παράγοντες αναφέρονται:

- Η **ηλικία,** συνήθως ο καρκίνος του μαστού στον άνδρα εμφανίζεται στις ηλικίες από 50 έως 70 ετών.

- Η **εθνικότητα,** μεγαλύτερη συχνότητα παρουσιάζουν οι άνδρες της μαύρης φυλής και ακόμα μεγαλύτερη οι άνδρες της Εβραϊκής φυλής.
- Το **οικογενειακό ιστορικό,** συγκεκριμένα επηρεάζει η ύπαρξη του γονιδίου BRCA 1 & 2, εκτιμάται ότι δωδεκαπλασιάζει τη συχνότητα ανάπτυξης καρκίνου του μαστού στον άνδρα. Τα γονίδια αυτά κληρονομούνται στους απογόνους και από το αρσενικό μέλος της οικογένειας. Επίσης έχουν πολύ υψηλό κίνδυνο άτομα που στο στενό οικογενειακό τους περιβάλλον έχουν άνδρες ή γυναίκες με καρκίνο του μαστού ή με καρκίνο ωοθηκών ή παχέως εντέρου.
- Η **υψηλή κοινωνικό-οικονομική κατάσταση**, αν και δεν έχει διευκρινιστεί ο ρόλος της, όμως ο καρκίνος του μαστού είναι συχνότερος σε αποφοίτους Πανεπιστημιακών Σχολών.
- Η **έκθεση σε ακτινοβολία,** επανειλημμένα κατά τη νεαρή τους ηλικία.

Ακόμη έχει παρατηρηθεί πως άνδρες που έχουν το **σύνδρομο Kleinefelter, βουβωνοκήλη**, **κρυψορχία ή μειωμένη στάθμη τεστοστερόνης**, η οποία επιτρέπει την αύξηση του μαζικού αδένα, εμφανίζουν 20 φορές μεγαλύτερη πιθανότητα να αναπτύξουν καρκίνο του μαστού. (Ιστοσελίδες 10 & 11)

Αντίστοιχα και άνδρες με αυξημένα επίπεδα οιστρογόνων στο αίμα τους που προκαλούν γυναικομαστία (αύξηση του μεγέθους του μαστού), είτε από κάποια συνοδό πάθηση ή από εξωγενή φαρμακευτική δράση έχουν υψηλές πιθανότητες να αναπτύξουν καρκίνο.

1.7.2 Συμπτώματα

Το πλέον συχνό σύμπτωμα είναι το ανώδυνο ψηλαφητό ογκίδιο στον ανδρικό μαστό που ανακαλύπτεται από τον ίδιο τον ασθενή και συνήθως εντοπίζεται κάτω από την περιοχή της θηλής και της θηλαίας άλω όπου υπάρχει και η μεγαλύτερη συγκέντρωση μαζικού ιστού. Υπάρχουν όμως και άλλα συμπτώματα, όπως η αλλαγή μεγέθους ή ακόμη και σχήματος του ενός μαστού, εξέλκωση (πληγή) στο δέρμα του μαστού, έκκριμα από τη θηλή του μαστού, εισολκή της θηλής, εξάνθημα ή τέλος ψηλαφητοί μασχαλιαίοι λεμφαδένες μπορεί να αποτελούν την κλινική εκδήλωση ενός καρκίνου του μαστού.

1.7.3 Θεραπευτική αντιμετώπιση

Επειδή ο καρκίνος του μαστού στους άντρες είναι πολύ σπάνιος, δεν υπάρχουν αρκετά στοιχεία για τη θεραπεία του.

Η εκλογή της θεραπευτικής αγωγής βασίζεται στην σταδιοποίηση της νόσου με τον ίδιο τρόπο που γίνεται και στη γυναίκα.

Η πιο συχνή θεραπευτική αντιμετώπιση στον άνδρα είναι η μαστεκτομή, που συνοδεύεται από λεμφαδενικό καθαρισμό της σύστοιχης μασχάλης, αφού ο μαζικός αδένας στον άνδρα είναι λίγος και δεν επιτρέπει την ασφαλή εξαίρεση μόνου του όγκου.

Στους άνδρες, οι καρκίνοι του μαστού είναι πολύ πιο συχνά ευαίσθητοι στις ορμόνες, οιστρογόνα ή προγεστερόνη, οι οποίες επηρεάζουν και την ανάπτυξή τους. Επειδή οι υποδοχείς για τις ορμόνες στους καρκίνους μαστού ανδρών είναι περισσότεροι, σε μεγάλο ποσοστό των περιστατικών, οι ορμονικές θεραπείες για τους άνδρες μπορεί να είναι ιδιαίτερα χρήσιμες.

Στις περιπτώσεις με μεταστάσεις, συστήνονται κατά κύριο λόγο οι ορμονικές θεραπείες. Η χημειοθεραπεία επιφυλάσσεται μόνο για τις περιπτώσεις που δεν έχουν ευαισθησία στις ορμόνες.

Συνοπτικά βλέπουμε ότι ο καρκίνος του μαστού στους άνδρες, αν και σπανιότερος, έχει πολλά κοινά σημεία με τον ίδιο καρκίνο στις γυναίκες

(Ιστοσελίδες 10 & 11)

ΚΕΦΑΛΑΙΟ 2^{ο}

ΑΝΑΛΥΣΗ ΕΠΙΒΙΩΣΗΣ

2.1 Εισαγωγή

Στο κεφάλαιο που ακολουθεί εισάγονται με απλό και κατανοητό τρόπο οι βασικές έννοιες τις οποίες συναντάμε στην ανάλυση επιβίωσης. Ορίζονται οι κύριες συναρτήσεις τις οποίες χρησιμοποιούμε για την ανάλυση δεδομένων που αναφέρονται σε χρόνους ζωής (και όχι μόνο) καθώς και οι διάφορες σχέσεις με τις οποίες οι συναρτήσεις αυτές συνδέονται.

Διαπιστώνουμε ότι η ανάλυση επιβίωσης αναπτύχθηκε ως μια ανάγκη να αναλύσουμε και να ερμηνεύσουμε δεδομένα που αφορούν χρόνους ζωής και για τα οποία δεν ήταν γνωστές οι ακριβείς τιμές όλων των παραμέτρων. Οπότε στην συνέχεια αναφερόμαστε στην έννοια της λογοκρισίας και στις διάφορες μορφές της. Προκύπτει πως λογοκριμένα δεδομένα απορρέουν από διάφορες καταστάσεις και πως τα δεδομένα αυτά μπορούν να γραφούν και με κάποια άλλη μορφή (π.χ. σε ζεύγη της μορφής (T_i,Δ_i)).

Τέλος, να σημειώσουμε ότι οι περισσότερες πληροφορίες σχετικά με όλα τα παραπάνω προέρχονται από τις σημειώσεις του Δ. Αντζουλάκο καθηγητή στο Πανεπιστήμιο Πειραιώς για το μάθημα ‘Ανάλυση Επιβίωσης’, την διπλωματική του απόφοιτου Α. Ξυραφάς, 2005, D. Collett, 2000 και το άρθρο των Kaplan –Meier, 1958.

2.2 Ορισμοί – Εισαγωγικές έννοιες

Με τον όρο ‘ανάλυση επιβίωσης ’ εννοούμε ένα σύνολο στατιστικών μεθόδων που χρησιμοποιούνται για την ανάλυση δεδομένων που αφορούν στο χρόνο που μεσολαβεί μέχρι κάποιο συγκεκριμένο συμβάν (ενδεχόμενο). Αρχικά η ανάλυση επιβίωσης αναφερόταν στο χρόνο μεταξύ της θεραπείας μέχρι και το θάνατο, τον οποίο ορίζουμε συνήθως ως σημείο τερματισμού μιας παρατηρήσεως, αυτός είναι και ο λόγος που πήρε το συγκριμένο όνομα.

Επίσης το ενδεχόμενο το οποίο ορίζουμε ως σημείο τερματισμού μιας παρατηρήσεως μπορεί να είναι και οτιδήποτε άλλο πέραν του θανάτου (π.χ. χρόνος υποτροπής μιας ασθένειας ή χρόνος ίασης για ένα συγκεκριμένο άτομο κτλ.).

Η ανάλυση επιβίωσης χρήζει ιδιαίτερης σημασίας καθώς τα δεδομένα με τα οποία ασχολείται όπως προαναφέραμε προκύπτουν ως χρόνοι επιβίωσης και δεν μπορούν να αναλυθούν με βάση τις κλασικές στατιστικές μεθόδους, καθώς δεν καλύπτουν κάποιες βασικές προϋποθέσεις. Όπως για παράδειγμα οι χρόνοι επιβίωσης είναι περιορισμένοι στο να είναι πάντα θετικοί και έτσι δεν ακολουθούν κάποια γνωστή συμμετρική κατανομή, π .χ την κανονική κατανομή. Θα μπορούσαμε βέβαια να μετατρέψουμε τα δεδομένα έτσι ώστε να έχουν μια πιο συμμετρική κατανομή, δηλ. θα μπορούσαμε να πάρουμε τους λογαρίθμους των χρόνων επιβίωσης. Όμως τότε ερχόμαστε αντιμέτωποι με ένα άλλο πιο ουσιαστικό και σημαντικό πρόβλημα τα δεδομένα χρόνων επιβίωσης συνήθως είναι λογοκριμένα. Τα λογοκριμένα δεδομένα είναι αυτά για τα οποία δεν είναι γνωστός ο χρόνος που συμβαίνει το γεγονός , το μόνο που μπορούμε να αναφέρουμε είναι ότι ο χρόνος επιβίωσης τους είναι μεγαλύτερος από την τιμή που έχει καταγραφεί.

Αρχικά θα ορίσουμε την τυχαία μεταβλητή που δηλώνει το χρόνο ζωής ενός ατόμου, τον οποίο συμβολίζουμε με Τ κεφαλαίο, όπου $T \geq 0$, αφού ο χρόνος δεν παίρνει αρνητικές τιμές.

Στην συνέχεια θα περιγράψουμε δύο πολύ σημαντικές ποσότητες που χρησιμοποιούνται προκειμένου να περιγράψουμε την κατανομή του χρόνου επιβίωσης ενός ατόμου, την συνάρτηση επιβίωσης και την συνάρτηση κινδύνου.

2.2.1 Συνάρτηση Επιβίωσης (survival function) S(t)

Ο πραγματικός χρόνος επιβίωσης ενός ατόμου, t, όπως αναφέραμε και παραπάνω μπορεί να θεωρηθεί ως η τιμή μιας μεταβλητής T, η οποία παίρνει μη αρνητικές τιμές. Οι διαφορετικές τιμές που μπορεί να πάρει η μεταβλητή T έχουν κάποια κατανομή πιθανότητας και έτσι μπορούμε να αποκαλέσουμε την T μια τυχαία μεταβλητή η οποία συνδέεται με το χρόνο επιβίωσης. Υποθέτουμε ότι η τυχαία μεταβλητή T έχει μια κατανομή πιθανότητας με αντίστοιχη συνάρτηση πυκνότητας την f(t). Η συνάρτηση κατανομής της T ορίζεται από την σχέση:

$$F(t)=P(T\leq t)\ ,\ t\geq 0 \qquad (2.1)$$

και αντιπροσωπεύει την πιθανότητα ο χρόνος επιβίωσης ενός ατόμου να είναι μικρότερος ή ίσον του t.

Έτσι η συνάρτηση επιβίωσης (survival function) S(t), που είναι η ουρά της κατανομής πιθανότητας, θα δηλώνει την πιθανότητα ο χρόνος ζωής ενός ατόμου να είναι μεγαλύτερος του χρόνου t, και θα δίνεται από την σχέση:

$$S(t)=P(T>t)\ ,\ t\geq 0 \qquad (2.2)$$

Οι παραπάνω σχέσεις, εφόσον έχουμε γνήσια ανισότητα, θα μπορούσαν να γραφτούν αντίστοιχα και ως εξής :

$$F(t)=P(T < t) \text{ και } S(t) =P(T \geq t) \text{ , για κάθε } t\geq 0$$

Επειδή όμως ο τρόπος γραφής που συναντάμε πιο πολύ είναι ο πρώτος επιλέγουμε να χρησιμοποιήσουμε αυτόν στην συγκεκριμένη εργασία.

Είναι εμφανές ότι η συνάρτηση επιβίωσης είναι φθίνουσα συνάρτηση του χρόνου με αρχή την τιμή 1 (αφού S(0)=1) και η οποία τείνει στο μηδέν στο τέλος του χρόνου ζωής του ατόμου, αφού το $\lim_{t\to\infty} S(t) = 0$.

Όταν ο χρόνος T είναι μια διακριτή τυχαία μεταβλητή, που παίρνει τις τιμές $t_1<t_2<\ldots$ τότε η συνάρτηση επιβίωσης S(t) της T ορίζεται από την σχέση

$$S(t) = \sum_{j:t_j>t} f(t_j),\ t \geq 0 \qquad (2.3)$$

όπου f(t) είναι η συνάρτηση πιθανότητας της T.

Ακόμα ισχύουν οι σχέσεις:

$$S(t) = 1,\ \text{για}\ 0 \leq t \leq t_1,$$

όπου t_1 ο χρόνος μέχρι την 1^{η} αποτυχία.

$$S(t_j) = f(t_{j+1}) + f(t_{j+2}) + ..., \; j = 1, 2, ...$$

$$f(t_j) = S(t_{j-1}) - S(t_j), \; j = 1, 2, ...$$

και
$$S(t_j) = P(T > t_j) = \prod_{j:t_j<t} \frac{S(t_{j+1})}{S(t_j)} \qquad (2.4)$$

Όταν ο χρόνος Τ είναι συνεχής τυχαία μεταβλητή: τότε η συνάρτηση επιβίωσης υπολογίζεται από την σχέση

$$S(t) = P(T > t) = \int_t^{\infty} f(u)du = 1 - F(t), \; t \geq 0$$

Και κατά συνέπεια η συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής Τ μπορεί να υπολογιστεί από την σχέση

$$f(t) = -\frac{d(S(t))}{dt} = -S'(t)$$

2.2.2 Συνάρτηση κινδύνου (hazard function ή hazard rate) h(t)

Μια άλλη σημαντική ποσότητα στην ανάλυση επιβίωσης είναι η συνάρτηση κινδύνου, η οποία χρησιμοποιείται ευρέως για να εκφράσει το ρίσκο ή τον κίνδυνο του θανάτου σε κάποια χρονική στιγμή t.

Η συνάρτηση αυτή είναι γνωστή και με άλλες ονομασίες οι οποίες συνδέονται κάθε φορά με το αντικείμενο που μελετάμε, έτσι στη θεωρία αξιοπιστίας είναι γνωστή ως δεσμευμένη βαθμίδα αποτυχίας (conditional failure rate), στις στοχαστικές διαδικασίες ως συνάρτηση εντάσεως (intensity function), στην δημογραφία και στη αναλογιστική επιστήμη ως ένταση θνησιμότητας (force of mortality) κ.ά.

Η συνάρτηση κινδύνου της τυχαίας μεταβλητής Τ, δηλώνει την στιγμιαία πιθανότητα που έχει ένα άτομο να αντιμετωπίσει το ενδεχόμενο (π.χ. θάνατος) στο χρόνο t, δοθέντος ότι αυτό επέζησε μέχρι τη χρονική στιγμή t. Για ένα πιο τυπικό ορισμό της συνάρτησης κινδύνου θα έπρεπε να θεωρήσουμε ότι η πιθανότητα με την οποία η τυχαία μεταβλητή, συνδέεται με το χρόνο επιβίωσης ενός ατόμου, Τ, παίρνει τιμές στο διάστημα t και t+Δt, δοθέντος ότι η Τ είναι μεγαλύτερη ή ίση του t, το οποίο γράφεται ως εξής $P(t \leq T < t+\Delta t | T \geq t)$. Προκειμένου να εκφράσουμε την

παραπάνω δεσμευμένη πιθανότητα σε πιθανότητα για κάθε μονάδα του πληθυσμού που μελετάμε θα πρέπει να την διαιρέσουμε με χρονικό διάστημα Δt, έτσι δίνουμε και κάποιο ρυθμό στην συνάρτηση, και στην συνέχεια να πάρουμε το όριο αυτού του πηλίκου για Δt→0.

Οπότε η h(t) θα υπολογίζεται από την σχέση

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T \le t + \Delta t \mid T \ge t)}{\Delta t},\ t \ge 0 \qquad (2.5)$$

Καθώς το Δt τείνει στο μηδέν και για κάθε t μη αρνητικό.

Από την εξίσωση (2.5) βλέπουμε ότι η h(t)* Δt είναι κατά προσέγγιση η πιθανότητα θανάτου ενός ατόμου στο διάστημα [t,t+Δt] γνωρίζοντας ότι το άτομο έχει επιβιώσει μέχρι την χρονική στιγμή t. Αυτός είναι και ο λόγος που η h(t) συνήθως αναφέρεται και ως ο κίνδυνος του θανάτου την χρονική στιγμή t.

$$h(t) \cong P(t \le T < t + \Delta t \mid T \ge t)$$

Από τον ορισμό της h(t) (2.5) προκύπτουν κάποιες χρήσιμες και ενδιαφέρουσες σχέσεις μεταξύ της συνάρτησης επιβίωσης και της συνάρτησης κινδύνου.

Για να καταλήξουμε στις σχέσεις αυτές θα χρησιμοποιήσουμε ορισμούς και θεωρήματα που είναι ευρέως γνωστά. Από την θεωρία πιθανοτήτων είναι γνωστό ότι η πιθανότητα πραγματοποίησης ενός ενδεχομένου Α δοθέντος ότι έχει πραγματοποιηθεί το ενδεχόμενο Β ισούται με P(A|B) = P(AB)/P(B), όπου P(AB) είναι η πιθανότητα να πραγματοποιηθούν ταυτόχρονα τα ενδεχόμενα Α και Β. Θα χρησιμοποιήσουμε το παραπάνω αποτέλεσμα προκειμένου να δείξουμε ότι η δεσμευμένη πιθανότητα στον ορισμό της συνάρτησης κινδύνου (2.5) ισούται με,

$$\frac{P(t \le T < t + \Delta t)}{P(T > t)}$$

Η οποία αλλιώς γράφεται ως εξής,

$$\frac{F(t + \Delta t) - F(t)}{S(t)},$$

όπου F(t) είναι η σ.κ της Τ

Τότε σύμφωνα με τα παραπάνω η συνάρτηση κινδύνου θα ισούται με

$$h(t) = \lim_{\Delta t \to 0} \left(\frac{F(t + \Delta t) - F(t)}{\Delta t} \right) \frac{1}{S(t)} \qquad (*)$$

Όμως από τον ορισμό της παραγώγου ξέρουμε ότι η ποσότητα

$$\lim_{\Delta t \to 0}\left(\frac{F(t+\Delta t)-F(t)}{\Delta t}\right) \qquad (**)$$

μας δίνει την παράγωγο της F(t) ως προς την t, η οποία είναι στην ουσία η σ.π.π της t δηλ. η f(t).

Άρα από (*), (**) έχουμε ότι η

$$h(t)=\frac{f(t)}{S(t)} \qquad (2.6)$$

Το οποίο συνεπάγεται $h(t)=-\frac{d}{dt}(\log S(t))$ (2.7)

Οι παραπάνω σχέσεις όταν ο **χρόνος Τ είναι διακριτή τυχαία μεταβλητή** ορίζονται ως εξής:

$$h(t)=P(T=t\,|\,T\geq t)=\frac{P(T=t)}{P(T\geq t)}=\frac{f(t)}{S(t)}$$

Και αφού $f(t_j)=S(t_{j-1})-S(t_j)$ και $S(0)=1$ έχουμε ότι

$$h(t_j)=1-\frac{S(t_{j-1})}{S(t_j)}$$

Ακόμα από την σχέση (2.4) έχουμε ότι $S(t_j)=\prod_{j:t_j\leq t}[1-h(t_j)]$

2.2.3 Κύριες μορφές της συνάρτησης κινδύνου

Η μελέτη της συνάρτησης κινδύνου παρουσιάζει ιδιαίτερο ενδιαφέρον στην συνεχή περίπτωση (αφού στην διακριτή περίπτωση h(t)=0, για κάθε $t\neq t_j$) καθώς στηριζόμαστε για την περιγραφή της στην μονοτονία της. Στην πράξη παρατηρούνται τέσσερεις κυρίες μορφές της συνάρτησης κινδύνου , οι οποίες είναι:

I. Αύξουσα συνάρτηση κινδύνου (increasing failure rate, IFR) , είναι η περίπτωση που συναντάμε πιο συχνά στην πράξη. Χρησιμοποιείται για να δηλώσει φυσική γήρανση ή υπολειπόμενο χρόνο ζωής σε μια συγκεκριμένη ομάδα ηλικιών.

II. Φθίνουσα συνάρτηση κινδύνου (decreasing failure rate, DFR), δεν συναντάται τόσο συχνά στην πράξη και εφαρμόζεται σε περιπτώσεις όπου υπάρχει μεγάλη πιθανότητα αποτυχίας στα πρώτα στάδια.

III. Συνάρτηση κινδύνου λεκανοειδής μορφής (bathtub-shaped). Είναι το πιο ρεαλιστικό μοντέλο από όλα για την περιγραφή του χρόνου ζωής ενός ανθρώπου από την γέννηση του και για μεγάλο χρονικό διάστημα. Στην περίπτωση αυτή η συνάρτηση κινδύνου είναι φθίνουσα αρχικά, σταθερή για κάποιο διάστημα και αύξουσα στο τελευταίο διάστημα.

IV. Συνάρτηση κινδύνου μορφής καμπούρας (upside-down bathtub hazard rate). Στην περίπτωση αυτή η συνάρτηση κινδύνου είναι αρχικά αύξουσα και στην συνέχεια φθίνουσα. Η μορφή αυτή συναντάται συνήθως όταν εξετάζουμε τον χρόνο αποτυχίας ύστερα από μια επιτυχημένη χειρουργική επέμβαση.

Επομένως η συνάρτηση κινδύνου έχει μια σαφή φυσική ερμηνεία και γενικά κάθε πληροφορία γύρω από την μορφή της συνάρτησης κινδύνου μπορεί να χρησιμοποιηθεί για την αναγνώριση και την κατάλληλη επιλογή ενός παραμετρικού μοντέλου. (Αντζουλάκος, 2009)

2.2.4 Αθροιστική συνάρτηση κινδύνου H(t)

Είναι μια ποσότητα που προέρχεται από την ολοκλήρωση της συνάρτηση κινδύνου και ορίζεται από την σχέση

$$H(t) = \int_0^t h(u)du,\ t \geq 0 \quad \acute{\eta} \qquad H(t) = -\ln[S(t)] \tag{2.8}$$

όταν ο χρόνος T είναι συνεχής τυχαία μεταβλητή.

Ακόμη από τις σχέσεις που δείξαμε παραπάνω προκύπτει ότι

$$S(t) = \exp[-H(t)] = \exp[-\int_0^t h(u)du] \tag{2.9}$$

Έχουμε ότι το $\lim_{t\to\infty} H(t) = \infty$, οπότε η αθροιστική συνάρτηση κινδύνου δεν είναι μια συνάρτηση κατανομής , δηλ. η ένταση θνησιμότητας είναι μια μη – ολοκληρώσιμη συνάρτηση στο $(0,\infty)$.

Τώρα, όταν ο χρόνος Τ είναι διακριτή τυχαία μεταβλητή τότε η αθροιστική συνάρτηση κινδύνου υπολογίζεται από το διακριτό ανάλογο της συνεχούς περίπτωσης (2.8) , επομένως θα ισούται με

$$H(t) = \sum_{j:t_j<t} h(t_j)$$

2.3 Λογοκριμένα δεδομένα (censored data)

(Για περισσότερα σχετικά με την θεωρεία που ακολουθεί βλέπε, M. Stevenson 2009)

Στην ανάλυση επιβίωσης συναντάμε συχνά λογοκριμένα ή ελλιπή δεδομένα, όπως προαναφέραμε αυτός ήταν και ένας από τους λόγους για τους οποίους τα δεδομένα αυτά δεν μπορούν να αναλυθούν με βάση τις κλασικές στατιστικές μεθόδους.

Έστω ότι μελετάμε το χρόνο ζωής μιας ομάδας ατόμων και το δείγμα που έχουμε στην διάθεση μας αποτελείται από $X_1,X_2,\ldots X_n$, χρόνους ζωής. Στην περίπτωση που όλες οι ακριβείς τιμές X_i στο δείγμα των n ατόμων είναι γνωστές, μπορούμε να βγάλουμε συμπεράσματα για την κατανομή του πληθυσμού χρησιμοποιώντας γνωστές παραμετρικές και μη παραμετρικές μεθόδους.

Όμως στην αντίθετη περίπτωση, όπου δεν είναι δυνατό να γνωρίζουμε τις ακριβείς τιμές των χρόνων ζωής όλου του δείγματος, όπως συνήθως συμβαίνει στην ανάλυση επιβίωσης, χρησιμοποιούμε άλλες μεθόδους για την ανάλυση των δεδομένων οι οποίες λαμβάνουν υπόψη την ύπαρξη λογοκριμένων παρατηρήσεων. Οι λογοκριμένες παρατηρήσεις σε μια μελέτη προκύπτουν όταν κάποια άτομα που συμμετέχουν σε αυτήν καταφέρουν να επιβιώνουν στο τέλος της μελέτης και έτσι για τα άτομα αυτά δεν γνωρίζουμε τον πλήρη χρόνο ζωής. Επίσης λογοκριμένες παρατηρήσεις έχουμε και όταν κάποια άτομα αποχωρούν από την μελέτη με την θέληση τους, άλλοι μπορεί να μεταναστεύουν ή να μετακομίσουν και έτσι δεν είμαστε σε θέση να τους παρακολουθούμε πλέον, ή και κάποιοι μπορεί να πεθάνουν κατά την διάρκεια της μελέτης από αίτια που δεν σχετίζονται με το ενδεχόμενο που μας ενδιαφέρει.

Μια λογοκριμένη παρατήρηση μας δίνει μόνο μερική πληροφόρηση για το χρόνο ζωής ενός ατόμου. Έτσι , ανάλογα με την γνώση μας για το χρόνο ζωής ενός ατόμου , αυτός μπορεί να είναι :

- είτε λογοκριμένος από δεξιά αν γνωρίζουμε ότι ο χρόνος ζωής του ατόμου είναι μεγαλύτερος από κάποιο χρόνο U
- είτε λογοκριμένος από τα αριστερά αν γνωρίζουμε ότι ο χρόνος ζωής του ατόμου είναι μικρότερος από κάποιο χρόνο U
- είτε λογοκριμένος σε διάστημα αν γνωρίζουμε ότι ο χρόνος ζωής του ατόμου βρίσκεται εντός ενός διαστήματος της μορφής (L , R) με L< R .

Τα δεδομένα που συναντάμε στην πράξη συνήθως αναφέρονται σε χρόνους λογοκριμένους από δεξιά, και εμείς για την ανάλυση μας θα θεωρούμε ότι έχουμε δεξιά λογοκρισία.

Ακόμη τα λογοκριμένα δεδομένα διακρίνονται και σε δεξιά λογοκρισία τύπου 1, δεξιά λογοκρισία τύπου 2 και τυχαία λογοκρισία.

Δεξιά λογοκρισία τύπου 1(type 1, right censoring)

Στον τύπο αυτό λογοκρισίας θέτουμε ένα συγκεκριμένο χρονικό διάστημα στον οποίον θα πρέπει να έχει ολοκληρωθεί η μελέτη, έστω u. Ο χρόνος u είναι σταθερός αριθμός και είναι γνωστός ως χρόνος λογοκρισίας. Επομένως αν έχουμε στην διάθεση μας το τυχαίο δείγμα $X_1,X_2,\ldots X_n$, (που παριστάνει τον πραγματικό χρόνο επιβίωσης των ατόμων στο δείγμα) για όσους αποτυγχάνουν πριν το χρόνο u (έστω αποτυγχάνουν κ άτομα) ο χρόνος επιβίωσης θα είναι γνωστός ενώ για τους υπόλοιπους (n-κ) θα ξέρουμε μόνο ότι οι χρόνοι ζωής τους είναι μεγαλύτεροι του χρόνου u.

Τότε τα λογοκριμένα δεδομένα μπορούν να γραφούν με την μορφή ζευγών (T_i,Δ_i) με i=1,2,…n, όπου

$$T_i = \min(X_i, u) \quad \text{και} \quad \Delta_i = \begin{cases} 1, & X_i \le u \\ 0, & X_i > u \end{cases}$$

δηλαδή η Δ_i είναι μια δίτιμη τυχαία μεταβλητή που δηλώνει αν ο χρόνος ζωής ενός ατόμου στο δείγμα είναι πλήρης (Δ_i=1) ή λογοκριμένος (Δ_i=0).

Μια πιο σύνθετη περίπτωση λογοκρισίας τύπου 1 εμφανίζεται όταν ο χρόνος λογοκρισίας δεν είναι ο ίδιος και για τα n άτομα στο δείγμα. Στην περίπτωση αυτή θεωρούμε ότι ο χρόνος λογοκρισίας για το άτομο i είναι u_i. Αυτό μπορεί να συμβεί όταν η χρονική αρχή παρακολούθησης του κάθε ατόμου δεν συμπίπτει με την χρονική αρχή έναρξης της μελέτης. Παρόμοια με παραπάνω τα λογοκριμένα δεδομένα μπορούν να γραφούν σε ζεύγη $(T_i\ \Delta_i)$ όπου τώρα

$$T_i = \min(X_i, u_i) \ \text{ και } \Delta_i = \begin{pmatrix} 1, & X_i \le u_i \\ 0, & X_i > u_i \end{pmatrix}$$

τώρα η $\Delta_i=1$ δηλώνει ότι γνωρίζουμε τον πλήρη χρόνο ζωής για το άτομο i.

- **<u>Δεξιά λογοκρισία τύπου 2 (type 2, right censoring)</u>**

Στο συγκεκριμένο είδος λογοκρισίας ο ερευνητής είναι αυτός που προκαθορίζει το πότε θα σταματήσει η έρευνα, ακριβέστερα ορίζει ότι θα σταματήσει την έρευνα όταν αποτύχουν τα πρώτα r άτομα (r<n) στο δείγμα. Τα δεδομένα μας θα αποτελούνται από τους διατεταγμένους πλήρεις χρόνους ζωής ($t_1 \le t_2 \le \ldots \le t_r$) των πρώτων r ατόμων που απέτυχαν, ενώ για τα υπόλοιπα n- r άτομα γνωρίζουμε μόνο ότι έχουν χρόνο ζωής μεγαλύτερο από το μέγιστο χρόνο ζωής των πρώτων r ατόμων που απέτυχαν. Τα δεδομένα που προκύπτουν μπορούν να γραφούν σε ζεύγη της μορφής $(T_{(i)},\ \Delta_i)$, με i=1,2,…n, όπου

$$T_{(i)} = \begin{pmatrix} X_{(i)}, & 1 \le i \le r \\ X_{(r)}, & r+1 \le i \le n \end{pmatrix} \text{ και } \Delta_i = \begin{pmatrix} 1, & 1 \le i \le r \\ 0, & r+1 \le i \le n \end{pmatrix}$$

Επίσης, συχνά στην πράξη χρησιμοποιείται ο συνδυασμός λογοκρισίας τύπου 1 και 2 (π.χ. στην βιοστατιστική) όπου η έρευνα σταματά είτε με την παρέλευση προκαθορισμένου χρόνου u είτε με την αποτυχία r ατόμων, όποιο συμβεί πρώτο.

Το πιο σημαντικό μειονέκτημα της λογοκρισίας τύπου 2 είναι ότι ο συνολικός χρόνος (T_r) που διαρκεί η έρευνα είναι άγνωστος.

- **Τυχαία λογοκρισία (random censoring)**

Στην τυχαία λογοκρισία τόσο ο χρόνος έναρξης της έρευνας όσο και ο χρόνος τερματισμού της είναι προκαθορισμένα, εκείνο που είναι τυχαίο είναι οι χρονικές στιγμές στις οποίες τα άτομα εισέρχονται στην έρευνα κατά συνέπεια και ο χρόνος λογοκρισίας που αντιστοιχεί σε κάθε άτομο είναι τυχαίος. Έτσι για κάθε άτομα έχουμε μια τυχαία μεταβλητή X που δηλώνει το χρόνο ζωής του και μια άλλη τυχαία μεταβλητή, έστω U, που δηλώνει το χρόνο λογοκρισίας του ατόμου. Επομένως και πάλι τα δεδομένα που προκύπτουν μπορούν να γραφούν σε ζεύγη της μορφής (T_i, Δ_i) , με i=1,2,…n, και όπου

$$T_i = \min(X_i, U_i) \quad \text{και} \quad \Delta_i = \begin{pmatrix} 1, & X_i \le U_i \\ 0, & X_i > U_i \end{pmatrix}$$

Ακόμη θα πρέπει να αναφερθούμε και στα **περικομμένα δεδομένα (truncated data)**. Η περικοπή ορίζεται ως μια συνθήκη την οποία θα πρέπει να ικανοποιεί κάποιο άτομο προκειμένου να λάβει μέρος σε μια μελέτη. Τα άτομα που δεν ικανοποιούν την συγκεκριμένη συνθήκη δεν λαμβάνονται υπόψη στη ανάλυση. Η δυσκολία που παρουσιάζεται οφείλεται στο γεγονός ότι είναι αρκετά πιθανό τα άτομα να μην είναι υπό παρακολούθηση σε όλη τη διάρκεια επιβίωσής τους, ή μέχρι τη στιγμή που θα αντιμετωπίσουν το ενδεχόμενο που μελετάμε.

Για παράδειγμα κάποια άτομα (ασθενείς) που συμμετέχουν στην μελέτη μπορεί να μην αντιμετωπίζουν το ενδεχόμενο (π.χ. εμφάνιση συγκεκριμένης ασθένειας που μας ενδιαφέρει) μέχρι κάποια προκαθορισμένη χρονική στιγμή και έτσι δεν λαμβάνονται υπόψη στην μελέτη. Κάποιοι άλλοι δεν περιλαμβάνονται στην μελέτη καθώς δεν πληρούν πλέον τις αναγκαίες προϋποθέσεις, π.χ. αν μας ενδιαφέρουν άτομα ηλικίας ≥ 65, όσα άτομα δεν κατάφεραν να επιβιώσουν μέχρι την συγκεκριμένη ηλικία δεν θα τεθούν υπό παρακολούθηση.

Όταν εμφανίζονται οι παραπάνω περιπτώσεις σε μια μελέτη τότε δημιουργούνται και οι λεγόμενες περικομμένες παρατηρήσεις, οι οποίες δεν καθιστούν δυνατή την ανάλυση των δεδομένων με βάση τις κλασικές μεθόδους.

ΚΕΦΑΛΑΙΟ 3^{ο}

Εκτιμητής Kaplan-Meier (Μη – παραμετρική εκτίμηση της συνάρτησης επιβίωσης)

3.1 Εισαγωγή

Στο κεφάλαιο που ακολουθεί ορίζεται η έννοια των πινάκων επιβίωσης, γίνεται αναφορά στο σπουδαίο ρόλο τους στην ανάπτυξη της ανάλυσης επιβίωσης και παίρνουμε και μια πρώτη εκτίμηση για την συνάρτηση επιβίωσης ενός δείγματος. Ακολουθεί ο Kaplan – Meier εκτιμητής, ο οποίος χρησιμοποιείται για να εκτιμήσει μη – παραμετρικά την τιμή της συνάρτησης επιβίωσης που αναφέρεται σε δεξιά λογοκριμένους χρόνους ζωής. Στην συνέχεια εφόσον έχουμε κάποια εκτίμηση για την συνάρτηση επιβίωσης ενδιαφερόμαστε να δούμε αν υπάρχει κάποια σημαντική διαφορά στις εκτιμήσεις συναρτήσεων επιβίωσης που αναφέρονται σε υπό ομάδες του δείγματος που μελετάμε, χρησιμοποιώντας για αυτό το σκοπό το log-rank test καθώς και άλλα γνωστά τεστ.

Τέλος, να σημειώσουμε ότι οι περισσότερες έννοιες και στοιχεία που αναφέρονται στο κεφάλαιο αυτό προέρχονται από το βιβλίο του D, Collett, 2000 και τις σημειώσεις του Δ. Αντζουλάκος, 2009.

3.2 Ορισμοί- Υποθέσεις Πινάκων Επιβίωσης

Ο «πίνακας επιβίωσης» ή «πίνακας θνησιμότητας» που την ιδέα δημιουργίας του πρώτος συνέλαβε ο J. Graunt, επιτρέπει την ακριβή περιγραφή του τρόπου με τον οποίο εξαφανίζονται προοδευτικά τα μέλη μιας γενεάς εξαιτίας της θνησιμότητας. Ο πίνακας επιβίωσης είναι η ιστορία της ζωής μιας υποθετικής (συγχρονική ανάλυση) ή πραγματικής γενεάς (διαγενεακή ανάλυση) μειούμενης βαθμιαίως λόγω των θανάτων.

Οι υποθέσεις που υιοθετούνται κατά την κατάρτιση ενός πίνακα επιβίωσης είναι:

α) Η προς διερεύνηση πλασματική γενεά αποτελείται από ενός σταθερό αριθμό γεννήσεων που λαμβάνεται συνήθως ως δύναμη του δέκα (π.χ. 10^3, ή 10^4) και καλείται ¨ρίζα¨ του πίνακα επιβιώσεως

β) Ο πληθυσμός πεθαίνει σε κάθε ηλικία σύμφωνα με ένα προκαθορισμένο, σταθερό πρότυπο θνησιμότητας.

γ) Ο υπό παρατήρηση πληθυσμός είναι «κλειστός» στην επίδραση της μετανάστευσης και επομένως οι μεταβολές του αρχικού πληθυσμού οφείλονται μόνο στους θανάτους.

δ) Οι θάνατοι που συμβαίνουν κατά την διάρκεια κάθε ηλικίας ισοκατανέμονται (με εξαίρεση τα δύο πρώτα έτη ζωής)

ε) Ο συνολικός αριθμός θανάτων του πληθυσμού του πίνακα επιβίωσης είναι ίσος με το συνολικό αριθμό των γεννήσεων του πληθυσμού (δηλ. με την ρίζα του πίνακα επιβίωσης). (Μ. Παπαδάκης & Κ. Τσίμπος, 2004)

Οι πίνακες επιβίωσης διακρίνονται **αναλόγως του χρόνου** στον οποίο αναφέρονται σε:

α) *Πίνακες επιβίωσης περιόδου* (συγχρονική ανάλυση), οι οποίοι βασίζονται σε δεδομένα για τους κατά ηλικιακή ομάδα θανάτους μιας περιόδου (έτους, πενταετίας κ.τ.λ.) και στον κατά ηλικιακή ομάδα πληθυσμό στο μέσον της ίδιας περιόδου.

β) *Πίνακες επιβίωσης γενεάς* (διαγενεακή ανάλυση), οι οποίοι βασίζονται στους συντελεστές θνησιμότητας, οι οποίοι προκύπτουν από τη διαχρονική παρακολούθηση των μελών μιας γενεάς.

Οι πίνακες επιβίωσης επίσης διακρίνονται **αναλόγως του εύρους των ηλικιακών ομάδων** στο οποίο αναφέρονται: α) σε *πλήρεις,* όπου τα δεδομένα των θανάτων και του πληθυσμού υπάρχουν κατά μονοετείς ηλικιακές ομάδες και β) σε *συνεπτυγμένους* όπου τα δεδομένα των θανάτων και του πληθυσμού υπάρχουν κατά μεγαλύτερες του έτους ηλικιακές ομάδες, συνήθως πενταετείς.

Ακόμα να σημειώσουμε ότι οι πίνακες επιβίωσης συνήθως δημιουργούνται ξεχωριστά για το κάθε φύλο, λόγω των σημαντικά διαφορετικών κατά ηλικία επιπέδων θνησιμότητας στους άνδρες και τις γυναίκες.

Τέλος, οι πίνακες επιβίωσης (αναλυτικοί ή συνεπτυγμένοι) δύνανται να δημιουργηθούν και ανά αιτία θανάτου. Για την ταξινόμηση των θανάτων ανά αιτία χρησιμοποιείται το πρότυπο ταξινόμησης του Παγκόσμιου Οργανισμού Υγείας (Manual of the International Statistical Classification of Diseases, Injuries and Causes of Death), το οποίο ταξινομεί τους θανάτους σε μεγάλες ομάδες αιτιών.

3.2.1 Οι Πίνακες Επιβίωσης στην Ανάλυση Επιβίωσης

Οι πίνακες επιβίωσης (life tables) αποτελούν μία από τις παλαιότερες και τις πιο διαδεδομένες μεθόδους στην περιγραφή δεδομένων χρόνων επιβίωσης. Είναι μια παραλλαγή των συνηθισμένων πινάκων συχνοτήτων προσαρμοσμένα κατά τέτοιο τρόπο ώστε να εφαρμόζονται στην περίπτωση που υπάρχουν λογοκριμένα δεδομένα. Πρόκειται για ένα μηχανισμό, ο οποίος παίρνει ένα δείγμα ατόμων και ομαδοποιεί τους χρόνους ζωής τους καθώς και τους λογοκριμένους (ή περικεκομμένους) χρόνους σε διαστήματα.

Θεωρούμε ένα τυχαίο δείγμα μεγέθους n από ένα συγκεκριμένο πληθυσμό το οποίο περιέχει πλήρη και λογοκριμένα δεδομένα. Τα δεδομένα αυτά εμφανίζονται σε ένα πίνακα , ο οποίος αποτελείται από

k+1 διαστήματα της μορφής $I_j=[a_{j-1},a_j)$ για j=1,2,…,k+1, όπου $a_0=0$, $a_k=M$ (όπου M είναι μια προκαθορισμένη σταθερά, αν αναπαριστάναμε τα διαστήματα αυτά σε ένα άξονα χρόνου τότε το M θα αντιστοιχούσε στο άνω όριο χρόνου των παρατηρήσεων που έχουμε στο δείγμα) και $a_{k+1}=\infty$.

Κάθε παρατήρηση του τυχαίου δείγματός μας θα αντιστοιχεί είτε σε πλήρη χρόνο είτε σε λογοκριμένο. Έτσι για τα δεδομένα που ομαδοποιήσαμε είναι γνωστό μόνο σε ποιο διάστημα τα άτομα αποβίωσαν ή λογοκρίθηκαν. Συγκεκριμένα το τελευταίο διάστημα I_{k+1} θα αποτελείται αποκλειστικά από χρόνους επιβίωσης διότι τα άτομα που δεν αποβίωσαν μέχρι την χρονική στιγμή M, θα αποβιώσουν κάποια χρονική στιγμή στο διάστημα I_{k+1}.

Για κάθε διάστημα I_j , j=1,2,…,k+1 ορίζουμε τις ακόλουθες ποσότητες:

- r_j : αριθμός ατόμων που βρίσκονται σε κίνδυνο στην αρχή του διαστήματος I_j ,
- d_j : αριθμός θανάτων στο διάστημα I_j (δηλ. μας δίνει τους πλήρους χρόνους ζωής)

- c_j : αριθμός διαφυγών στο διάστημα I_j , δηλ. λογοκριμένοι χρόνοι επιβίωσης .

Από τα παραπάνω έχουμε τις σχέσεις:

$$r_1 = n \text{ , } r_j = r_{j-1} - d_{j-1} - c_{j-1}, \text{ για } j = 1,2,...k+1$$
$$\acute{\eta}\ r_{j+1} = r_j - d_j - c_j$$

Ακόμη, υποθέτοντας ότι η κατανομή των χρόνων επιβίωσης (T) των ατόμων του υπό μελέτη πληθυσμού έχει συνάρτηση επιβίωσης S(t) μπορούμε να ορίσουμε τις ποσότητες:

- P_j=P(άτομο να επιζεί πέραν του διαστήματος I_j)=
  P(T> a_j)=S(a_j)
- p_j=P(άτομο να επιζεί πέραν του διαστήματος I_j| επέζησε πέραν του διαστήματος I_{j-1})=P(T≥ a_j|T≥ a_{j-1})=P_j/P_{j-1}
- q_j=P(άτομο να πεθάνει στο διάστημα I_j| επέζησε πέραν του διαστήματος I_{j-1})=P(a_{j-1}≤T<a_j)/P(T≥a_{j-1})=1-P_j/P_{j-1}

(Αντζουλάκος, 2009)

Από τα παραπάνω έχουμε ότι P_0=1, $P_{κ+1}$=0, p_{k+1}=0 και q_{k+1}=1. Ακόμη ισχύει ότι

$$P_j = S(a_j) = p_1 p_2 ... p_j = (1-q_1)(1-q_2)...(1-q_j) = p_j P_{j-1} \text{ , για } j = 1,2,...k+1$$

αφού $P_j = \frac{P_1}{P_0} \cdot \frac{P_2}{P_1} \cdot \cdots \cdot \frac{P_j}{P_{j-1}}$, $j = 1,2,...k+1$.

Στην περίπτωση που όλοι οι χρόνοι επιβίωσης που μελετάμε είναι <u>πλήρεις</u> τότε προφανώς ισχύει ότι :

r_1=n, και r_j=r_{j-1}-d_{j-1}=n-d_1-d_2-…d_{j-1}, για j=1,2,…κ+1

και μια εκτιμήτρια της ποσότητας P_j θα ήταν η ποσότητα $\frac{r_{j+1}}{n}$, η οποία μας δίνει τον αριθμό των ατόμων που βρίσκονται εν ζωή την χρονική στιγμή a_j. . Στην περίπτωση όμως που υπάρχουν <u>λογοκριμένοι χρόνοι</u> ζωής τότε η παραπάνω ποσότητα θα υποεκτιμά την P_j , διότι η ποσότητα r_{j+1} μπορεί να μην συμπίπτει με τον πραγματικό αριθμό των ατόμων που βρίσκονται εν ζωή τη χρονική στιγμή a_j.

Προκειμένου να αντιμετωπίσουμε το συγκεκριμένο πρόβλημα θα επιδιώξουμε να εκτιμήσουμε πρώτα τις ποσότητες p_j και q_j και στην συνέχεια μέσω αυτών να οδηγηθούμε στην εκτιμήτρια της ποσότητας P_j.

Διακρίνουμε την ανάλυση μας σε δύο περιπτώσεις, στην πρώτη περίπτωση θεωρούμε ότι δεν υπάρχουν διαφυγές ($c_j=0$) και στην δεύτερη ότι υπάρχουν ($c_j>0$) και υπολογίσουμε για την κάθε περίπτωση τις ποσότητες που μας ενδιαφέρουν.

Στην $1^{η}$ περίπτωση μια λογική εκτιμήτρια της ποσότητας q_j δίνεται από την σχέση:

$$\hat{q}_j = \frac{d_j}{r_j}$$

Ενώ στην $2^{η}$ περίπτωση η παραπάνω ποσότητα δεν θα μπορούσε να χρησιμοποιηθεί, διότι θα υποεκτιμά την ποσότητα q_j , επειδή είναι πιθανό τα άτομα που έχουν διαφύγει στο διάστημα I_j να έχουν πεθάνει πριν το τέλος του διαστήματος I_j. Για να αντιμετωπίσουμε το συγκεκριμένο πρόβλημα εργαζόμαστε ως εξής, υποθέτουμε ότι οι λογοκριμένοι χρόνοι ζωής συμβαίνουν ομοιόμορφα κατά την διάρκεια του j διαστήματος και έτσι ο μέσος αριθμός ατόμων που είναι σε κίνδυνο κατά την διάρκεια του διαστήματος θα ισούται με

$$r_j' = r_j - c_j / 2$$

δηλ. θεωρούμε ότι οι μισές διαφυγές είναι σε ισχύ στο διάστημα I_j .

Η παραπάνω υπόθεση είναι γνωστή ως 'αναλογιστική υπόθεση'(actuarial assumprtion) (Collett, 2000)

Αντίστοιχα η εκτίμηση της ποσότητας q_j θα υπολογίζεται από την σχέση:

$$\hat{q}_j = \frac{d_j}{r_j'} = \frac{d_j}{r_j - \frac{c_j}{2}} \ , \ r_j>0.$$

Στην περίπτωση που $r_j=0$, τότε το $q_j = 1$ και το $p_j =1- q_j=1$, για κάθε τιμή του $r_j>0$ τότε η πιθανότητα επιβίωσης ενός ατόμου θα ισούται με:

$$\hat{p}_j = 1 - \frac{d_j}{r_j'} = \frac{r_j' - d_j}{r_j'} .$$

Αν θελήσουμε στην συνέχεια να υπολογίσουμε την πιθανότητα ενός άτομου να επιζήσει πέραν του χρόνου t_m με m=1,2,...k+1, (δηλ. να ζήσει κάποιο χρόνο μετά από την αρχή του m διαστήματος), αυτή θα ισούται με το γινόμενο των πιθανοτήτων ενός ατόμου να έχει επιζήσει πέραν της αρχής του m διαστήματος και σε όλα τα προηγούμενα m-1 διαστήματα.

Επομένως, η εκτίμηση της συνάρτησης επιβίωσης που παίρνουμε από τους πίνακες επιβίωσης θα ισούται με:

$$\hat{P}_j = \hat{p}_1 \hat{p}_2 \ldots \hat{p}_j = \prod_{j=1}^{m} \left(1 - \frac{d_j}{r'_j} \right)$$

Η εκτίμηση της συνάρτησης επιβίωσης με χρήση των πινάκων επιβίωσης είναι ευαίσθητη στην επιλογή των διαστημάτων I_j , τα διαστήματα αυτά δεν είναι απαραίτητο να είναι ίσα, συνήθως χρησιμοποιούμε 8 με 10 διαστήματα, εξαρτάται κάθε φορά από την φύση των δεδομένων.

Ακόμα από τους πίνακες επιβίωσης μας δίνεται η δυνατότητα να υπολογίσουμε και τα εκατοστημόρια (*percentiles*). Το p-οστό εκατοστημόριο μιας κατανομής του χρόνου T ορίζεται ως η τιμή t_p για την οποία ισχύει:

$$P(T \le t_p) = p \Rightarrow t_p = F^{-1}(p)$$

Για την τιμή p=0,5 παίρνουμε τη διάμεσο του δείγματός μας, η οποία δηλώνει την χρονική στιγμή πέρα από την οποία περιμένουμε να επιβιώσει το 50% των ατόμων του υπό μελέτη δείγματος.

3.3 Εκτιμητής Kaplan-Meier (product limit estimator ,PLE)

Ο εκτιμητής που χρησιμοποιείται για την εκτίμηση της συνάρτησης επιβίωσης πάνω σε δεδομένα που αφορούν χρόνους επιβίωσης, είναι γνωστός με την ονομασία Kaplan-Meier εκτιμητής (λόγω των δημιουργών του) ή ως εκτιμητής γινομένου ορίου (product limit estimator). Έτσι όπως και η ανάλυση επιβίωσης μπορεί να χρησιμοποιηθεί και σε άλλους τομείς εκτός της ιατρικής (όπως αναφέραμε και παραπάνω) έτσι και ο K-M εκτιμητής μπορεί να χρησιμοποιηθεί στην μηχανολογία (προκειμένου να εκτιμήσει το χρόνο έως την αποτυχία μιας μηχανής), στην οικονομία (προκειμένου να εκτιμήσει το χρονικό διάστημα που παραμένει κάποιος εκτός εργασίας ύστερα από απόλυση), στον αναλογισμό και την δημογραφία (όπου επίσης μας ενδιαφέρει η κατασκευή πινάκων θνησιμότητας ή η εκτίμηση π.χ. της έντασης θνησιμότητας για διάφορες ηλικίες) κτλ.

Ένα από τα πιο σημαντικά πλεονεκτήματα του συγκεκριμένου εκτιμητή είναι ότι μπορεί να αναλύσει λογοκριμένα δεδομένα και συγκεκριμένα δεξιά λογοκριμένα δεδομένα, δεδομένα τα οποία χάθηκαν στην διάρκεια της μελέτης πριν το τελικό αποτέλεσμα παρατηρηθεί. Αυτός ήταν και ένας από τους λόγους που οδήγησαν στην δημιουργία του διότι αν κάθε φορά τα δεδομένα που μελετούσαμε αναφερόντουσαν

σε πλήρεις χρόνους επιβίωσης τότε η συνάρτηση επιβίωσης θα συνέπιπτε με την εμπειρική συνάρτηση επιβίωσης (empirical survivor function, ESF) η οποία όπως είναι γνωστό ορίζεται από την σχέση:

$$S(t) = \frac{αριθμός\_ατόμων\_με\_χρόνο\_ζωής > t}{σύνολο\_ατόμων\_που\_συμμετέχουν\_στην\_μελέτη(n)}$$

Η ESF είναι μια σκαλωτή φθίνουσα συνάρτηση, η οποία μειώνεται κατά 1/n μετά από κάθε παρατηρούμενο χρόνο επιβίωσης , οπότε αν έχουμε κ χρόνους επιβίωσης ίσους ή μεγαλύτερους του t τότε η ESF θα μειωθεί κατά κ/ n στο χρόνο t.

Όμως στην περίπτωση που έχουμε λογοκριμένα δεδομένα η παραπάνω εκτίμηση της συνάρτησης επιβίωσης δεν μπορεί να χρησιμοποιηθεί καθώς δεν γνωρίζουμε των πραγματικό αριθμό των ατόμων που έχουν χρόνο ζωής > t. Αν αγνοούσαμε απλώς τις συγκεκριμένες παρατηρήσεις, αυτό θα είχε σαν συνέπεια την απώλεια χρήσιμης πληροφορίας. Παραδείγματος χάριν, μπορεί κάποιοι να χάθηκαν κατά το χρόνο t_2 ξέρουμε όμως ότι επιβίωσαν κατά την διάρκεια του t_1, κι θα έπρεπε αυτή η πληροφορία να ληφθεί υπόψη στην εκτίμηση της S(t). Οι περιπτώσεις που δεν ήταν τόσο ξεκάθαρες και θα δημιουργούσαν κάποια σύγχυση στην ανάλυση των δεδομένων ήταν αυτές όπου κάποια άτομα είχαν χαθεί και άλλοι είχαν πεθάνει ακριβώς στο χρόνο t. Για αυτό το λόγο οι Kaplan&Meier σε άρθρο τους το 1958 αποφάσισαν ότι οι θάνατοι που είχαν καταγραφεί στο χρόνο t να μεταχειρίζονται σαν να είχαν συμβεί λίγο πριν το χρόνο t, ενώ οι απώλειες που είχαν παρατηρηθεί στο χρόνο t να θεωρούνταν σαν να είχαν συμβεί λίγο μετά το χρόνο t. Με αυτόν τον τρόπο η ανάλυση των δεδομένων θα ήταν πιο ξεκάθαρη, πιο κατανοητή και πιο γρήγορη.

Ο εκτιμητής που πρότειναν για την ανάλυση τέτοιων δεδομένων (εκτιμητή Kaplan-Meier) έχει αναφερθεί και απεδείχθη ότι είναι ο μη παραμετρικός μέγιστης πιθανοφάνειας εκτιμητής της S(t), ορίζεται ως εξής:

Ορίζουμε ως S(t) την πιθανότητα ενός ατόμου από το σύνολο του πληθυσμού να έχει χρόνο επιβίωσης που να ξεπερνά το t.

Έστω ότι το μέγεθος του δείγματος είναι n, και έστω ότι οι παρατηρούμενοι χρόνοι μέχρι το θάνατο των n ατόμων είναι $t_1 \leq t_2 \leq \ldots \leq t_n$. Αντίστοιχα ορίζουμε και τις ποσότητες r_j και d_j , με την πρώτη να δηλώνει τον αριθμό των ατόμων που βρίσκονται σε κίνδυνο αμέσως πριν το χρόνο t_j και η δεύτερη τον αριθμό των θανάτων στο χρόνο

t_j. Είναι λογικό ότι τα διαστήματα ανάμεσα σε κάθε t_j δεν είναι ομοιόμορφα. Έτσι ο Kaplan-Meier εκτιμητής υπολογίζεται από την σχέση:

$$\hat{S}(t) = \prod_{t_j < t} \frac{r_j - d_j}{r_j}$$

Υπάρχει επίσης και ένας εναλλακτικός τρόπος ορισμού του K-M εκτιμητή, που ορίζεται ως εξής:

$$\hat{S}(t) = \prod_{tj \le t} \frac{r_j - d_j}{r_j} \qquad (3.1)$$

Η διαφορά στους δύο παραπάνω ορισμούς οφείλεται μόνο στους παρατηρούμενους χρόνους πραγματοποιήσεις του ενδεχομένου, ο δεύτερος ορισμός είναι δεξιά συνεχής ενώ ο πρώτος αριστερά.

Ο δεύτερος ορισμός είναι προτιμότερος διότι η εκτίμηση που παίρνουμε για την συνάρτηση επιβίωσης είναι πιο συμβατή με μια δεξιά συνεχή εκτίμηση της F(t), αφού όπως δείξαμε οι δύο συναρτήσεις συνδέονται μέσω της σχέσης

$$S(t) = P(T > t) = 1 - P(T \le t) = 1 - F(t)$$

Όσον αφορά την γραφική απεικόνιση του εκτιμητή όπως μπορούμε να δούμε και από την σχέση ορισμού του, αυτή θα είναι μια κλιμακωτή φθίνουσα καμπύλη η οποία θα ξεκινάει από την τιμή 1 την χρονική στιγμή t=0 και θα μειώνεται πολλαπλασιαστικά κατά $(r_j-d_j)/ r_j$ κάθε χρονική στιγμή t_j. Ο εκτιμητής θα αλλάζει τιμή μόνο στους χρόνους που παρατηρούνται θάνατοι (ενδεχόμενο) ενώ στο υπόλοιπο μέρος η τιμή του θεωρούμε ότι μένει σταθερή. Στο γράφημα οι απώλειες συνήθως δηλώνονται με κάθετα σημεία πάνω στην γραφική απεικόνιση , στην περίπτωση που δεν παρατηρείται ούτε λογοκρισία ούτε περικοπή τότε το γράφημα του εκτιμητή συμπίπτει με αυτό της εμπειρικής συνάρτησης επιβίωσης. Αν η τελευταία παρατήρηση είναι λογοκριμένος χρόνος τότε ο K-M εκτιμητής ορίζεται μέχρι αυτή την παρατήρηση.

3.3.1 Εκτίμηση της διακύμανσης του K-M εκτιμητή

Ύστερα από τον ορισμό του K-M εκτιμητή για να μπορεί να χρησιμοποιηθεί ο εκτιμητής αυτός για συμπερασματολογία, θα έπρεπε με κάποιο τρόπο να προσδιοριστεί και η εκτίμηση της διακύμανσης Var(S(t)) του K-M εκτιμητή της S(t).

Είδαμε παραπάνω ότι ο K-M εκτιμητής της συνάρτησης επιβίωσης ισούται με:

$$\hat{S}(t) = \prod_{t_j \le t} \frac{r_j - d_j}{r_j},$$

οπότε η

$$V(\hat{S}(t)) = V\left(\prod_{t_j \le t} \frac{r_j - d_j}{r_j} \right) = V\left(\prod_{t_j \le t} \hat{p}_j \right)$$

όμως ο υπολογισμός της διακύμανσης ενός γινομένου δεν είναι εύκολη υπόθεση όπως στην περίπτωση ενός αθροίσματος, για αυτό επιλέγουμε να χρησιμοποιήσουμε την λογαριθμική μορφή της συνάρτησης επιβίωσης για να καταλήξουμε μετά από κάποιες συγκεκριμένες διαδικασίες στην εκτίμησης της διακύμανσης της $V(\hat{S}(t))$.

Έτσι ,

$$V(\log(\hat{S}(t)) = V\left(\sum_{t_j < t} \log(\hat{p}_j) \right) = \sum_{t_j < t} V(\log(1 - \hat{q}_j)),$$

αυτά ισχύουν κάτω από την υπόθεση ότι οι μεταβλητές $\log(1-\hat{q}_j)$ είναι ανεξάρτητες, δηλ. θεωρούμε ότι δεν υπάρχει κάποια συσχέτιση στην εμφάνιση των αποτυχιών.

Στην συνέχεια κάνοντας χρήση της μεθόδου Δέλτα και του αναπτύγματος του Taylor (συγκεκριμένα μόνο του πρώτου όρου) παίρνουμε ότι

$$V(\hat{S}(t)) = \left(\prod_{j:t_j < t} (1 - q_j) \right)^2 \sum_{j:t_j < t} \frac{q_j}{r_j (1 - q_j)}$$

και θέτοντας όπου q_j την εκτίμηση της $q_j = d_j / r_j$, προκύπτει ο γνωστός τύπος του **Greenwood**, που χρησιμοποιείται ευρέως για την εκτίμηση της V(S(t)),

$$\hat{V}(\hat{S}(t)) = \left(\hat{S}(t) \right)^2 \sum_{j:t_j < t} \frac{d_j}{r_j (r_j - d_j)} \qquad (3.2)$$

3.3.2 Διαστήματα εμπιστοσύνης της συνάρτησης επιβίωσης

Αφού εκτιμήσαμε την διακύμανση της S(t) και την ίδια την συνάρτηση επιβίωσης μέσω του PLE εκτιμητή, μπορούμε τώρα να φτιάξουμε διάστημα εμπιστοσύνης για την τιμή που θα πάρει η συνάρτηση επιβίωσης σε μια δοσμένη χρονική στιγμή t.

Ένα τέτοιο διάστημα για την πραγματική τιμή της S(t) σε μια συγκεκριμένη χρονική στιγμή t θα υπολογιστεί κάτω από την υπόθεση ότι η εκτιμώμενη τιμή για την S(t) ακολουθεί ασυμπτωτικά την κανονική κατανομή (προκύπτει από το γεγονός ότι ο K-M εκτιμητής της συνάρτησης επιβίωσης είναι εκτιμητής μέγιστης πιθανοφάνειας) με μέση τιμή ίση με S(t) και διακύμανση εκτιμώμενη από την σχέση (3.2).

Έτσι ένα γνωστό διάστημα εμπιστοσύνης , το οποίο είναι γνωστό ως *διάστημα εμπιστοσύνης τύπου Plain*, για την ποσότητα S(t) με συντελεστή εμπιστοσύνης 1-α δίνεται από την σχέση:

$$\hat{S}(t) \pm z_{a/2}\sqrt{\hat{V}(\hat{S}(t))}$$

Το παραπάνω διάστημα εμπιστοσύνης δεν είναι τόσο ικανοποιητικό όταν οι τιμές που παίρνει η S(t) είναι κοντά στο μηδέν ή την μονάδα, και αυτό διότι το διάστημα εμπιστοσύνης είναι συμμετρικό και έτσι υπάρχει κίνδυνος οι τιμές των ορίων εμπιστοσύνης να ξεπεράσουν το διάστημα στο οποίο παίρνει τιμές η S(t) που είναι το [0,1].

Στην περίπτωση που συμβεί αυτό μπορούμε αν το κάτω όριο του διαστήματος εμπιστοσύνης είναι μικρότερο του μηδέν να το αντικαταστήσουμε με το μηδέν και αντίστοιχα αν το άνω όριο εμπιστοσύνης είναι μεγαλύτερο της μονάδας να το αντικαταστήσουμε το 1, όμως το διάστημα που θα πάρουμε δεν θα ανταποκρίνεται στην πραγματικότητα. Για αυτό το λόγο έχουν αναπτυχθεί τα διαστήματα εμπιστοσύνης τύπου log και log- log τα οποία μας εξασφαλίσουν ότι η S(t) θα παίρνει τιμές στο διάστημα [0,1].

Στα *διαστήματα εμπιστοσύνης τύπου log* θέτουμε W(t)=log S(t) και κάνοντας χρήση της μεθόδου Δέλτα προκύπτει ότι η μέση τιμή της W(t) ισούται με:

$$\mu_{\hat{W}(t)} = \log S(t) = W(t)$$

και η διακύμανση με:

$$\sigma^2\hat{W}(t) = \hat{V}[\hat{S}(t)]\left(\frac{1}{\hat{S}(t)}\right)^2 = \sum_{j:t_j<t}\frac{d_j}{r_j(r_j-d_j)}$$

Επομένως ένα διάστημα εμπιστοσύνης για την ποσότητα W(t) με συντελεστή εμπιστοσύνης 1-α είναι το

$$\hat{W}(t) \pm z_{a/2}\hat{\sigma}\hat{W}(t)$$

Ενώ τα αντίστοιχο διάστημα εμπιστοσύνης για την ποσότητα S(t)=exp(W(t)) θα ισούται με

$$\hat{S}(t)\exp(\pm z_{a/2}\hat{\sigma}_{\hat{W}(t)})$$

Παρόμοια για να κατασκευάσουμε *διαστήματα εμπιστοσύνης τύπου log- log* θέτουμε U(t) = log(- log(S(t)) και μέσω της μεθόδου Δέλτα εκτιμούμε την μέση τιμή και διακύμανση της U(t) , οι οποίες υπολογίζονται από τις σχέσεις:

$$\mu_{\hat{U}(t)} = \log[-\log(S(t)] = U(t)$$

και

$$\sigma^2_{\hat{U}(t)} = V(\hat{S}(t))\left(\frac{1}{S(t)\log S(t)}\right)^2 = \frac{1}{[\log\hat{S}(t)]^2}\sum_{j:t_j\le t}\frac{d_j}{r_j(r_j-d_j)}$$

Έτσι, ένα διάστημα εμπιστοσύνης για την ποσότητα U(t) με συντελεστή εμπιστοσύνης (1-α) θα υπολογίζεται από την σχέση :

$$\hat{U}(t) \pm z_{a/2}\hat{\sigma}_{\hat{U}(t)}$$

Επομένως το διάστημα εμπιστοσύνης για την ποσότητα

S(t)=exp[-exp(U(t))]

με συντελεστή εμπιστοσύνης 1-α θα είναι ίσο με

$$[\hat{S}(t)]^{\exp(\pm z_{a/2}\hat{\sigma}_{\hat{U}(t)})}$$

Συνήθως στην πράξη χρησιμοποιούνται πιο πολύ τα διαστήματα εμπιστοσύνης τύπου log και log-log (που μας δίνουν καλύτερα αποτελέσματα) και όχι τόσο τα διαστήματα εμπιστοσύνης τύπου Plain λόγω του μειονεκτήματος που παρουσιάζουν, όπως προαναφέραμε παραπάνω.

(Αντζουλάκος 2009)

3.4 Σύγκριση Συναρτήσεων Επιβίωσης

Συνήθως τα δεδομένα που μελετάμε εκτός από τις βασικές μεταβλητές που δηλώνουν, το χρόνο ζωής και τη τυχαία μεταβλητή που δηλώνει αν ο χρόνος αυτός είναι λογοκριμένος ή όχι, περιλαμβάνουν και άλλες μεταβλητές οι οποίες εκφράζουν άλλα χαρακτηριστικά των ατόμων που μελετάμε. Έτσι το ενδιαφέρον των ερευνητών ύστερα από την εκτίμηση της συνάρτησης επιβίωσης, στρέφεται στην εκτίμηση της συνάρτησης επιβίωσης σε υποομάδες του δείγματος που μελετάμε. Εφόσον έχουμε λογοκριμένα δεδομένα αυτά δεν μπορεί να γίνει με κάποιο two sample t – test ή one way anova, οπότε θα εφαρμόσουμε άλλες διαδικασίες.

Αρχικά κατασκευάσουμε ένα γράφημα στο οποίο απεικονίζεται ο Kaplan-Meier εκτιμητής της συνάρτησης επιβίωσης ξεχωριστά για την κάθε ομάδα , θα έχουμε μια πρώτη εικόνα για το αν υπάρχει ή όχι κάποια διαφορά στην εκτιμώμενη συνάρτηση επιβίωσης στις δύο ομάδες , ακόμα μπορούμε να υπολογίσουμε και κάποια περιγραφικά μέτρα για τις δύο ομάδες. Στην συνέχεια θα πρέπει να ελέγξουμε αν η παρατηρούμενη διαφορά είναι στατιστικά σημαντική.

Ο έλεγχος για τη σύγκριση των συναρτήσεων επιβίωσης στις δύο ομάδες θα έχει την εξής μορφή: $H_0: S_1(t) = S_2(t)$, για κάθε t

έναντι της $H_1: S_1(t) \neq S_2(t)$, για κάποιο t

ή $H_0: h_1(t) = h_2(t)$ έναντι της $H_1: h_1(t) \neq h_2(t)$.

Κάτω από την μηδενική υπόθεση έχουμε ότι ο κίνδυνος του θανάτου είναι ο ίδιος για τις δύο ομάδες.

Εφόσον ορίσαμε τον έλεγχο υπόθεσης, το επόμενο βήμα είναι ο σχηματισμός του στατιστικού τεστ, που μετρά το βαθμό στον οποίο τα παρατηρούμενα δεδομένα απέχουν από την μηδενική υπόθεση.

Το στατιστικό ελέγχου έχει κατασκευασθεί έτσι ώστε όσο μεγαλύτερη τιμή έχει τόσο πιο πολύ είναι απέχουμε από την H_0, οπότε έχουμε ενδείξεις απόρριψης της H_0 υπέρ της H_1.

Ο υπολογισμός του στατιστικού βασίζεται σε πίνακες συνάφειας (που αφορούν τις ομάδες) και στο γεγονός ότι έχουμε λογοκριμένα δεδομένα. Μας ενδιαφέρουν μόνο οι πλήρεις χρόνοι, t_j, με $1 \leq j \leq m$, για τους οποίους όπως έχουμε ορίσει τις ποσότητες r_j και d_j που δηλώνουν αντίστοιχα το σύνολο των ατόμων που βρίσκονται σε κίνδυνο την χρονική στιγμή t_j και τον αριθμό των θανάτων, θα ορίσουμε και τις ποσότητες r_{1j},

r_{2j} και d_{1j} , d_{2j} που δηλώνουν τον αριθμό των ατόμων που βρίσκονται σε κίνδυνο την χρονική στιγμή t_j σε κάθε ομάδα καθώς και των αριθμό των θανόντων για την κάθε ομάδα.

Για κάθε πλήρη χρόνο ζωής κατασκευάσουμε τον 2x2 πίνακα:

t_j	# θανόντων την χρονική στιγμή t_j	# επιζώντων πέραν της χρονικής στιγμής t_j	# ατόμων σε κίνδυνο ακριβώς την χρονική στιγμή t_j
ομάδα 1	d_{1j}	r_{1j}-d_{1j}	r_{j1}
ομάδα 2	d_{2j}	r_{2j}-d_{2j}	r_{j2}
σύνολο	d_j	r_j-d_j	r_j

Κάτω από την H_0 υποθέτουμε ότι οι συναρτήσεις επιβίωσης των δύο υπό μελέτη ομάδων είναι ίσες, οπότε ένας τρόπος προκειμένου η υπόθεση να αποκτήσει κάποια ισχύ είναι υπολογίσουμε τις διαφορές ανάμεσα στον παρατηρούμενο αριθμό θανάτων που συμβαίνουν στις δύο ομάδες για κάθε πλήρη χρόνο και τον αναμενόμενο αριθμό θανάτων κάτω από την H_0.

Στην συνέχεια υποθέτοντας ότι τα περιθώρια αθροίσματα είναι σταθερά και ότι η μηδενική υπόθεση είναι αληθής, οι τέσσερεις εισαγωγές στον παραπάνω πίνακα μπορούν να καθοριστούν από την τιμή της d_{1j} , δηλ. από τον αριθμό των θανάτων που παρατηρούνται στην ομάδα 1την χρονική στιγμή t_j . Δοθέντος ότι οι παρατηρούμενοι θάνατοι (d_j) συμβαίνουν και στις δύο ομάδες την ίδια χρονική στιγμή t_j τότε κάτω από την μηδενική υπόθεση η d_{1j} θα ακολουθεί την υπεργεωμετρική κατανομή με παραμέτρους (r_j, r_{1j}, d_{1j}) . Η μέση τιμή μιας υπεργεωμετρικής τυχαίας μεταβλητής υπολογίζεται από την σχέση,

$$E_{1j} = r_{1j}\frac{d_j}{r_j}$$

έτσι η E_{1j} δηλώνει τον αναμενόμενο αριθμό θανάτων για την $1^{η}$ ομάδα την χρονική στιγμή t_j. Κάτω από την μηδενική υπόθεση η πιθανότητα θανάτου ενός ατόμου την χρονική στιγμή t_j δεν εξαρτάται από την ομάδα στην οποία ανήκει και ισούται με d_j/r_j.

Πολλαπλασιάζοντας όμως με το r_{1j} έχουμε τον αναμενόμενο αριθμό θανάτων για την ομάδα 1 στο χρόνο t_j. Αντίστοιχα η διακύμανση μιας τυχαίας μεταβλητής που ακολουθεί την υπεργεωμετρική κατανομή ισούται με :

$$V_{1j} = \frac{r_{1j} r_{2j} d_j (r_j - d_j)}{r_j^{\,2} (r_j - 1)}$$

Στο επόμενο βήμα για κάθε πλήρη χρόνο θα συνδυάσουμε την πληροφορία που παίρνουμε από τον παραπάνω 2x2 πίνακα έτσι ώστε να έχουμε μια συνολική εκτίμηση της μεταβλητότητας του παρατηρούμενου αριθμού θανάτων στην $1^{η}$ ομάδα από τις εκτιμώμενές τους τιμές.

Ο πιο απλός τρόπος για να το κάνουμε αυτό είναι να αθροίσουμε τις διαφορές d_{1j}-E_{1j} για όλους τους πλήρεις χρόνους , το στατιστικό που θα πάρουμε θα ισούται με :

$$U = \sum_{j=1}^{m} (d_{1j} - E_{1j})$$

όπου με m δηλώνουμε το πλήθος των χρονικών στιγμών που συμβαίνει τουλάχιστον 1 θάνατος.

Το παραπάνω στατιστικό θα έχει μέση τιμή ίσων με μηδέν αφού $E(d_{1j})=E_{1j}$. Ακόμη, εφόσον οι χρόνοι θανάτων των ατόμων είναι ανεξάρτητοι μεταξύ τους η διακύμανση του U θα ισούται με το άθροισμα των διακυμάνσεων του d_{1j} , δηλ.

$$V\,\mathrm{ar}(U) = \sum_{j=1}^{m} V_{1j}$$

Επιπλέον μπορεί να δειχθεί ότι όταν ο αριθμός των πλήρων χρόνων δεν είναι πολύ μικρός τότε η ποσότητα U μπορεί να προσεγγιστεί από την κανονική κατανομή, κατά συνέπεια η ποσότητα

$$U / Var(U)$$

θα ακολουθεί προσεγγιστικά την κανονική κατανομή με μέση τιμή 0 και διακύμανση ίση με 1.

Είναι γνωστό ότι το τετράγωνο μιας τυποποιημένης κανονικής κατανομής θα ακολουθεί μια X^2 κατανομή με 1 β.ε. , δηλ.

$$\frac{U^2}{\sqrt{\mathrm{var}(U)}} \sim X^2_1$$

Η διαδικασία αυτή (δηλ. ο συνδυασμός των πληροφοριών που παίρνουμε από 2x2 πίνακες) είχε προταθεί από τους Mantel – Haenszel το 1959 και είναι γνωστή ως Mantel – Haenszel procedure. Το τεστ που βασίζεται στο συγκεκριμένο στατιστικό φέρει διάφορες ονομασίες όπως Mantel- Cox ή Peto - Mantel – Haenszel test, όμως είναι ευρέως γνωστό με την ονομασία log- rank test (ονομασία που καθιέρωσαν οι Richard & Julian Peto) . (Για περισσότερες πληροφορίες βλ. David Collett, 2000).

Σύμφωνα με τα όσα αναφέραμε παραπάνω προκύπτει πως το log- rank test είναι ένα μη παραμετρικό τεστ για να συγκρίνουμε συναρτήσεις επιβίωσης δύο ομάδων. Το τεστ αυτό είναι κατάλληλο να χρησιμοποιηθεί όταν τα δεδομένα που μελετάμε είναι δεξιά λογοκριμένα. Το στατιστικό του κάτω από την H_0 δείξαμε πως ισούται με:

$$Q = \frac{\sum_{j=1}^{m}(d_{1j} - E_{1j})^2}{\sqrt{\sum_{j=1}^{m} V_{1j}}} \qquad (3.3)$$

και η απόφαση απόρριψης ή αποδοχής της H_0 λαμβάνεται κάθε φορά συναρτήσει των ποσοστημορίων της X^2 κατανομής με 1 βαθμό ελευθερίας για συγκεκριμένο α επίπεδο σημαντικότητας. Από τον ορισμό του στατιστικού μπορούμε να συμπεράνουμε πως περιμένουμε να απορρίψουμε την H_0 υπέρ της H_1 για μεγάλες τιμές του Q.

3.5 Άλλα τεστ σύγκρισης για δύο ομάδες

Το τεστ που περιγράψαμε παραπάνω καθώς και κάποια άλλα τα οποία θα παρουσιάσουμε στην συνέχεια ανήκουν στην ίδια οικογένεια , αυτή των ***γραμμικών βαθμωτών τεστ*** (***linear rank test***), τα οποία έχουν σαν βάση την παραπάνω στατιστική συνάρτηση (3.3) με τις διαφορές που παρατηρούνται να σταθμίζονται από βάρη (w_j).

Έτσι το στατιστικό πάνω στο οποίο βασίζονται τα συγκεκριμένα τεστ θα έχει την εξής μορφή:

$$Q = \frac{\left(\sum_{j=1}^{m} w_j (d_{1j} - E_{1j} \right)^2}{\sum_{j=1}^{m} w_j^2 V_{1j}}$$

και θα ακολουθεί προσεγγιστικά την X^2 κατανομή με 1 β.ε.

- Για $w_j=1$, καταλήγουμε στο γνωστό log- rank test, το οποίο προτάθηκε αρχικά από τον Nathan Mantel (1959) και στην συνέχεια από τους Peto- Peto το 1972 , και οι δύο οποίοι θεώρησαν ότι έχουμε ίδιο βάρος για όλες τις διαφορές.

- Για $w_j= r_j$, καταλήγουμε στο τεστ των Gehan (1965) και Breslow (1970), οι οποίοι γενίκευσαν το τεστ του Wilcoxon για λογοκριμένα ή περικομμένα δεδομένα. Στο τεστ αυτό τα βάρη συμπίπτουν με τον αριθμό των ατόμων που βρίσκονται σε κίνδυνο κάθε χρονική στιγμή $t=t_j$. Σε σύγκριση με τα υπόλοιπα τεστ θεωρείται πιο ρεαλιστικό.

- Για $w_j=\sqrt{r_j}$, καταλήγουμε στο τεστ των Tarone-Ware (1977), οι οποίοι χρησιμοποιήσανε ως βάρη την τετραγωνική ρίζα του αριθμού των ατόμων που βρίσκονται σε κίνδυνο κάθε χρονική στιγμή, πέτυχαν έτσι να δώσουν μεγαλύτερη εξομάλυνση στις παρατηρούμενες διαφορές.

Υπάρχουν και κάποια άλλα τεστ όπως είναι αυτό των Peto-Peto, Modified Peto-Peto, και Flemming-Harrington, τα οποία συναντώνται λιγότερο στην πράξη. Στα

τεστ αυτά συνήθως επιλέγονται ως βάρη κάποια εκτίμηση της συνάρτησης επιβίωσης, η οποία για το Modified Peto-Peto τεστ πολλαπλασιάζεται από την ποσότητα (r_j/ r_j+1) ενώ για το Flemming-Harrington τεστ, η εκτίμηση της συνάρτησης επιβίωσης συμπίπτει με αυτή του Kaplan-Meier εκτιμητή για το από κοινού δείγμα.

(Byuske et al 2000, Harrington & Fleming 1982)

3.5.1 Σύγκριση των τεστ log-rank & Breslow-Gehan

Οι έλεγχοι που χρησιμοποιούνται συχνότερα στην πράξη για την σύγκριση δύο συναρτήσεων επιβίωσης είναι ο έλεγχος log–rank και ο έλεγχος των Breslow-Gehan.

Συγκρίνοντας τους δύο ελέγχους παρατηρούμε ότι ο έλεγχος BG δίνει μεγαλύτερο βάρος σε μικρές τιμές των πλήρων χρόνων ζωής δηλαδή στο αριστερό άκρο της συνάρτησης επιβίωσης, όπου συγκεντρώνονται περισσότερες πληροφορίες σχετικά με τις συναρτήσεις επιβίωσης των δύο ομάδων.

Ενώ από την άλλη, έχει αποδειχθεί ότι το log–rank test έχει μεγαλύτερη ισχύ όταν ισχύει η υπόθεση αναλογικού κινδύνου (proportional hazard). Σύμφωνα με την υπόθεση αυτή θεωρούμε ότι οι συναρτήσεις κινδύνου των δύο ομάδων συνδέονται με την σχέση :

$$h_1(t) = c\, h_2(t)$$

όπου c είναι μια σταθερά.

Αντίστοιχα για τις συναρτήσεις επιβίωσης των δύο ομάδων θα ισχύει

$$S_1(t) = [S_2(t)]^c$$

Καθώς από την σχέση (2.7) έχουμε ότι

$$S_1(t) = \exp\left(-\int_0^t h_1(u)du\right) = \exp\left(-c\int_0^t h_2(u)du\right) = [S_2(t)]^c$$

Στην περίπτωση που **c=1**, δεν ισχύει η υπόθεση αναλογικού κινδύνου.

Ενώ

$$\text{για } c>1 \quad \text{έχουμε} \quad S_1(t) < S_2(t)$$
$$\text{για } c<1 \quad \text{έχουμε} \quad S_1(t) > S_2(t)$$

το οποίο δηλώνει ότι οι συναρτήσεις επιβίωσης των δύο ομάδων δεν τέμνονται.

Το γεγονός αυτό μας δίνει ένα κριτήριο επιλογής ανάμεσα στους δύο ελέγχους στην πράξη. Κατασκευάζοντας στο ίδιο γράφημα την εκτίμηση των συναρτήσεων επιβίωσης για τις δύο ομάδες με βάση την μέθοδο του Kaplan- Meier εκτιμητή, μπορούμε να αποφασίσουμε ποιο από τα δύο τεστ θα εφαρμόσουμε. Αν οι δύο γραφικές παραστάσεις δεν τέμνονται, τότε ισχύει η υπόθεση αναλογικού κινδύνου, και επομένως επιλέγουμε να εφαρμόσουμε το log–rank test για την σύγκριση των δύο συναρτήσεων επιβίωσης. Σε αντίθετη περίπτωση εφαρμόζουμε το Breslow-Gehan test.

3.5.2 Σύγκριση συναρτήσεων επιβίωσης για k (k>2) ομάδες

Τα τεστ που περιγράψαμε παραπάνω για την σύγκριση των συναρτήσεων επιβίωσης δύο ομάδων μπορούν να επεκταθούν και στη περίπτωση που θέλουμε να συγκρίνουμε τις συναρτήσεις επιβίωσης τριών ή περισσοτέρων ομάδων. Έστω λοιπόν ότι έχουμε να συγκρίνουμε k (k>2) συναρτήσεις επιβίωσης , τότε τόσο ο έλεγχος υπόθεσης που πρέπει να κάνουμε όσο και το στατιστικό ελέγχου θα πρέπει να προσαρμοστούν αντίστοιχα. Ο έλεγχος υπόθεσης θα έχει την εξής μορφή:

H_0: $S_1(t)=S_2(t)=\ldots=S_k(t)$, για κάθε t

Έναντι

H_1: τουλάχιστον ένα από τα $S_j(t)$ διαφέρει από τα υπόλοιπα για κάποιο t

Παρόμοια με παραπάνω μας ενδιαφέρουν μόνο οι πλήρεις χρόνοι ($t_1<t_2<\ldots< t_m$) για τους οποίους κατασκευάζουμε τον πίνακα:

Ομάδα	# θανάτων την χρονική στιγμή t_j	# ατόμων που επιζούν πέρα της στιγμής t_j	# ατόμων που βρίσκονται σε κίνδυνο την χρονική στιγμή t_j
1	d_{1j}	r_{1j}-d_{1j}	r_{1j}
2	d_2j	r_{2j}-d_{2j}	r_{2j}
.	.	.	.
.	.	.	.
k	d_{kj}	r_{kj}-d_{kj}	r_kj
Σύνολο	d_j	r_j-d_j	r_j

Ο αντίστοιχος ορισμός του U στατιστικού, για την σύγκριση του παρατηρούμενου αριθμού θανάτων στις ομάδες 1,2,3… k-1 σε σχέση με τις εκτιμώμενες τιμές του, θα έχει την μορφή:

$$U = \sum_{j=1}^{m} w_j \left(d_{ij} - \frac{r_{ij} d_j}{r_j} \right)$$

Όπου w_j είναι ένας διαγώνιος πίνακας, με τον οποίον δηλώνουμε τα βάρη για το κάθε τεστ. Η ποσότητα U εκφράζεται με την μορφή ενός διανύσματος που αποτελείται από k-1 στοιχεία. Επίσης είναι απαραίτητο να ορίσουμε και την κατάλληλη έκφραση για την διακύμανση της ποσότητας U και της συνδιακύμασνσης για ζευγαρωτές παρατηρήσεις. Οι τιμές αυτές έχουν συγκεντρωθεί σε ένα πίνακα διακυμάνσεων-διακυμάνσεων Σ, ο οποίος είναι ένας συμμετρικός πίνακας που έχει στα διαγώνια στοιχεία την διακύμανση της ποσότητας U και την συνδιακύμανση της στα μη - διαγώνια στοιχεία.

Έτσι η στατιστική συνάρτηση που χρησιμοποιείται για τον έλεγχο της μηδενικής υπόθεσης θα έχει την μορφή:

$$Q = U' \Sigma^{-1} U$$

που έχει αποδειχθεί πως ακολουθεί την X^2 κατανομή με k-1 β.ε όταν η μηδενική υπόθεση είναι αληθής.

(Collet 2000, Αντζουλάκος 2009)

Τέλος, να αναφέρουμε ότι υπάρχουν και άλλα τεστ για την σύγκριση συναρτήσεων επιβίωσης που εφαρμόζονται προκειμένου να ελέγξουμε κάποιες ειδικές περιπτώσεις. Ένα από τα τεστ αυτά είναι ο *έλεγχος τάσης*, που χρησιμοποιείται όταν ενδιαφερόμαστε να ελέγξουμε αν υπάρχει κάποια διάταξη σε k συναρτήσεις επιβίωσης.

Και ένα άλλο είναι τα *στρωματοποιημένα τεστ για σύγκριση συναρτήσεων επιβίωσης*. Τα τεστ αυτά χρησιμοποιούνται για την σύγκριση k συναρτήσεων επιβίωσης όταν θέλουμε να προσαρμόσουμε τα αποτελέσματα που έχουμε πάρει λαμβάνοντας υπόψη και άλλους παράγοντες οι οποίοι ενδέχεται να επηρεάζουν , ως συγχυτικοί παράγοντες, τους χρόνους ζωής στις ενδιαφερόμενες ομάδες.

ΚΕΦΑΛΑΙΟ 4^{ο}

Μοντέλο Αναλογικού Κινδύνου του COX

(Cox Proportional Hazards model)

4.1 Εισαγωγή

Όταν έχουμε να αναλύσουμε δεδομένα χρόνων επιβίωσης ενός συνόλου ατόμων (δείγματος) για τα οποία δεν έχουμε κάποια αίσθηση ούτε κάποιο στοιχείο σχετικά με την υποκείμενη κατανομή αυτών των δεδομένων και θεωρούμαι ότι η επιβίωση αυτών των ατόμων επηρεάζεται από έναν ή περισσότερους παράγοντες, τότε το πλέον κατάλληλο μοντέλο για την ανάλυση τους είναι το μοντέλο αναλογικού κινδύνου του Cox.

Το μοντέλο αυτό μπορούμε να το δούμε σαν μια γενίκευση του Kaplan-Meier εκτιμητή.

Σκοπός του κεφαλαίου αυτού είναι να μας γνωρίσει το πολύ σημαντικό και χρήσιμο για την ανάλυση επιβίωσης και όχι μόνο (έχει εφαρμογή στην μηχανική αξιοπιστία, στον αναλογισμό κτλ.) μοντέλο του Cox. Έχουμε την δυνατότητα να δούμε το πώς η ανάλυση του χρόνου επιβίωσης ενός συνόλου ατόμων εξαρτάται και επηρεάζεται από μια ή περισσότερες επεξηγηματικές μεταβλητές. Περιγράφονται σε συντομία μέθοδοι εκτίμησης των παραμέτρων του μοντέλου καθώς και οι ιδιότητες από τις οποίες αυτές συνοδεύονται. Στην συνέχεια προχωράμε στη ανάλυση επιλογής μεταβλητών για ένα μοντέλο αναλογικού κινδύνου, όπου περιγράφονται οι πιο γνωστές και προσιτές διαδικασίες επιλογής μεταβλητών. Τέλος γίνεται μια αναφορά στο στρωματοποιημένο μοντέλο του Cox και το μοντέλο του Cox με χρόνο - εξαρτώμενες μεταβλητές,

4.2 Εισαγωγή στο μοντέλο αναλογικού κινδύνου του Cox

Το μοντέλο αναλογικού κινδύνου του Cox (το οποίο παρουσιάστηκε για πρώτη φορά το 1972 σε άρθρο του περιοδικού JRSS) αποτελεί έναν επιπλέον τρόπο έκφρασης και υπολογισμού της κατανομής των χρόνων επιβίωσης ενός συνόλου ατόμων με χρήση της συνάρτησης κινδύνου.

Όπως αναφέραμε σε προηγούμενο κεφάλαιο η συνάρτηση κινδύνου εκφράζει το κίνδυνο της αποτυχίας (που τις περισσότερες φορές συμπίπτει με το θάνατο ενός ατόμου από το δείγμα που μελετάμε) την χρονική t δοθείσης της επιβίωσης μέχρι εκείνη την στιγμή, δηλ.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t} = \frac{f(t)}{S(t)}$$

Όμως στην πράξη έχει παρατηρηθεί πως είναι πιο χρήσιμο να μοντελοποιούμε την συνάρτηση κινδύνου και αντί της h(t) να χρησιμοποιούμε το λογάριθμο της , την log(h(t)).

Στο συγκεκριμένο κεφάλαιο τα δεδομένα μας γίνονται πιο περίπλοκα και αυτό διότι πέραν της μεταβλητής T_j (που δηλώνει το χρόνο ζωής του j ατόμου) και της μεταβλητής Δ_j (που δηλώνει αν ο χρόνος αυτός είναι λογοκριμένος ή όχι) τώρα μας ενδιαφέρει και η μεταβλητή Z_j η οποία γράφεται με την μορφή ενός διανύσματος $Z_j=(z_{j1},z_{j2},\ldots z_{jp})$ και μας δίνει το πλήθος των ρ ερμηνευτικών μεταβλητών που δηλώνουν διάφορά ποιοτικά ή ποσοτικά χαρακτηριστικά των υπό μελέτη ατόμων.

Επομένως τα δεδομένα μας μπορούν να γραφτούν στην μορφή:

$$(T_j, \Delta_j, Z_j) = (T_{j,} \Delta_j, z_{j1}, z_{j2}, \ldots z_{jp})$$

Ενδιαφερόμαστε λοιπόν, να εντοπίσουμε κάποια σχέση της συνάρτησης επιβίωσης με τις υπόλοιπες υπό μελέτη μεταβλητές. Προκειμένου να το πετύχουμε αυτό θα χρησιμοποιήσουμε κάποιο μοντέλο παλινδρόμησης, όμως τα γνωστά μοντέλα παλινδρόμησης δεν μπορούν να εφαρμοστούν σε δεδομένα που αφορούν χρόνους επιβίωσης λόγω της ιδιαίτερης μορφής τους. Τα μοντέλα που προκύπτουν διαφέρουν από τα γραμμικά μοντέλα παλινδρόμησης τόσο ως προς την μορφή όσο και ως προς τις ιδιότητες, παρόλα αυτά κάποιες βασικές αρχές διατηρούνται.

Με την μοντελοποίηση αρχικά της συνάρτησης κινδύνου και ύστερα της συνάρτησης επιβίωσης επιδιώκουμε α) να προσδιορίσουμε ποιος πιθανός συνδυασμός των επεξηγηματικών μεταβλητών επηρεάζει την μορφή της συνάρτησης

κινδύνου και β) την απόκτηση μιας εκτίμησης για την συνάρτηση κινδύνου ξεχωριστά για το κάθε άτομο.

Προσπαθούμε λοιπόν, να προσδιορίσουμε ένα «γραμμικό» μοντέλο για την ποσότητα logh(t) , ένα τέτοιο μοντέλο θα μπορούσε να γραφτεί ως εξής :

$$\log h_j(t) = a + \beta_1 z_{j1} + \beta_2 z_{j2} + ...\beta_p z_{jp}$$

ή ισοδύναμα

$$h_j(t) = \exp(a + \beta_1 z_{j1} + \beta_2 z_{j2} + ... + \beta_p z_{jp})$$

Το μοντέλο αυτό θα ήταν σαν ένα γραμμικό μοντέλο για την λογάριθμό της συνάρτησης κινδύνου ή σαν ένα πολλαπλασιαστικό μοντέλο για τον κίνδυνο.

Στο παραπάνω μοντέλο η σταθερά α δηλώνει κάποια λογαριθμική αναφορική συνάρτηση κινδύνου (log-baseline hazard function) αφού $\log h_j(t)=\alpha$ ή $h_j(t)=\exp(\alpha)$.

(Therneau & Grambsch 2000)

Σε αντίθεση με το μοντέλο που περιγράφθηκε παραπάνω το μοντέλο που εισήχθηκε από τον Cox (1972) αφήνει την αναφορική συνάρτηση κινδύνου(ΑΣΚ) απροσδιόριστη δηλ. $\alpha= \log h_0(t)$, οπότε το μοντέλο θα γράφεται ως εξής :

$$\log h_j(t) = \log h_0(t) + \beta_1 z_{j1} + \beta_2 z_{j2} + ... + \beta_p z_{jp}$$

ή

$$h_j(t) = h_0(t) \exp(\beta_1 z_{j1} + \beta_2 z_{j2} + ... + \beta_p z_{jp}) \qquad (4.1)$$

Για το παραπάνω μοντέλο ισχύουν οι εξής περιορισμοί και υποθέσεις:

- η $h_0(t)$ όπως αναφέραμε δηλώνει την αναφορική συνάρτηση κινδύνου (ΑΣΚ) ενός ατόμου η οποία είναι πάντα μεγαλύτερη ή ίση του μηδενός. Συνήθως ερμηνεύεται ως η συνάρτηση κινδύνου ενός ατόμου για το οποίο η τιμή της μεταβλητής Z ισούται με μηδέν.
- το διάνυσμα των παραμέτρων $\beta=(\beta_1,\beta_2,\ldots,\beta_p)$ για κάθε χρονική στιγμή είναι σταθερό και ανεξάρτητο του χρόνου, αυτό διότι στο μοντέλο κινδύνου του Cox θεωρούμε ότι ο λόγος των κινδύνων δεν εξαρτάται από το χρόνο. Κατά συνέπεια και οι μεταβλητές που εξετάσουμε δεν εξαρτώνται από το χρόνο.
- δεν κάνουμε καμία παραμετρική υπόθεση για την ΑΣΚ. Παραμετρική μορφή έχουμε μόνο για το διάνυσμα των συμμεταβλητών Z, για αυτό το λόγο το παραπάνω μοντέλο θεωρείται ένα ημι-παραμετρικό μοντέλο.

(Αντζουλάκος 2009)

Αν τώρα θεωρήσουμε δύο παρατηρήσεις j και j΄ οι οποίες διαφέρουν μόνο ως προς τις τιμές των Z_j μεταβλητών , με αντίστοιχους γραμμικούς εκτιμητές :

$$X_j = \beta_1 z_{j1} + \beta_2 z_{j2} + ... + \beta_p z_{jp}$$

και

$$X_{j'} = \beta_1 z_{j'1} + \beta_2 z_{j'2} + ... + \beta_p z_{j'p}$$

Τότε ο λόγος των δύο συναρτήσεων κινδύνου (γνωστός ως σχετικός λόγος κινδύνου, relative risk ratio) θα ισούται με :

$$\frac{h_j(t)}{h_{j'}(t)} = \frac{h_0(t)\exp X_j}{h_0(t)\exp X_{j'}} = \frac{\exp X_j}{\exp X_{j'}}$$

και είναι ανεξάρτητος του χρόνου t (σταθερός σε οποιαδήποτε χρονική στιγμή t) κατά συνέπεια το μοντέλο του Cox είναι ένα αναλογικό μοντέλο κινδύνου (proportional hazard model).

4.3 Προσαρμόζοντας το μοντέλο αναλογικού κινδύνου (Μερική πιθανοφάνεια)

Παρατηρούμε από την σχέση (4.1) ότι η συνάρτηση κινδύνου μέσω του μοντέλου αναλογικού κινδύνου του Cox αποτελείται από δύο μέρη, την αναφορική συνάρτηση κινδύνου και τους συντελεστές των επεξηγηματικών μεταβλητών που αποτελούν το γραμμικό μέρος του μοντέλου. Προκειμένου λοιπόν, να προσαρμόσουμε το μοντέλο που μας δίνει η σχέση (4.1) θα πρέπει να εκτιμήσουμε τα δύο αυτά μέρη. Τα μέρη αυτά δεν είναι αναγκαίο να εκτιμηθούν ταυτόχρονα, μπορούν να εκτιμηθούν και ξεχωριστά, συνήθως ο τρόπος που προτιμάται είναι να εκτιμήσουμε αρχικά τα β΄ (συντελεστές των $z_1, z_2, \ldots z_p$ επεξηγηματικών μεταβλητών) και στην συνέχεια να χρησιμοποιήσουμε τις εκτιμήσεις αυτές για να εκτιμήσουμε την αναφορική συνάρτηση κινδύνου. Το οποίο συνεπάγεται ότι αν θελήσουμε να υπολογίσουμε κάποιες διαφορές σχετικές με τις επιδράσεις των p επεξηγηματικών μεταβλητών ($z_1, z_2, \ldots z_p$) πάνω στο σχετικό κίνδυνο ($h_i(t)/h_0(t)$) δεν χρειάζεται να γνωρίζουμε την τιμή της αναφορικής συνάρτησης κινδύνου.

Παρόλο που η αναφορική συνάρτηση κινδύνου είναι απροσδιόριστη στο μοντέλο του Cox, αυτή μπορεί να εκτιμηθεί κάνοντας χρήση της μεθόδου μερικής πιθανοφάνειας (partial likelihood method), η οποία αναπτύχτηκε από τον Cox το 1972 , στο ίδιο άρθρο που παρουσιάστηκε για πρώτη φορά και το μοντέλο του. Όπως ο Cox και αρκετοί άλλοι έχουν δείξει ότι η μερική (log) πιθανοφάνεια μπορεί να αντιμετωπιστεί ως μια διατάξιμη (log) πιθανοφάνεια προκειμένου να προκύψουν εκτιμητές μέγιστης πιθανοφάνειας των β που θα είναι σε ισχύ.

Έτσι μπορούμε να εκτιμήσουμε λόγους κινδύνους και διαστήματα εμπιστοσύνης χρησιμοποιώντας τις ίδιες τεχνικές που χρησιμοποιούμε στην περίπτωση της μέγιστης πιθανοφάνειας με την μόνη διαφορά ότι τώρα οι εκτιμήσεις που διαθέτουμε θα βασίζονται στην μερική πιθανοφάνεια, την οποία θεωρούμε ότι είναι ολική.

Έστω, λοιπόν ότι έχουμε διαθέσιμα δεδομένα για n άτομα από το σύνολο του πληθυσμού, για τα οποία γνωρίζουμε τους r χρόνους θανάτων οπότε n-r άτομα έχουν λογοκριμένο χρόνο ζωής. Θεωρούμε ακόμη ότι δεν υπάρχουν δεσμοί δηλ. μόνο ένας θάνατος συμβαίνει σε κάθε χρόνο θανάτου. Στην συνέχεια διατάσσουμε τους r χρόνους θανάτων, $t_{(1)}<t_{(2)}<\ldots<t_{(r)}$, με την $t_{(i)}$ να δηλώνει τον i διατεταγμένο χρόνο θανάτου ενώ με $R(t_{(j)})$ συμβολίσουμε το σύνολο των ατόμων που βρίσκονται σε κίνδυνο την χρονική στιγμή $t_{(j)}$.

Οπότε η συνάρτηση πιθανοφάνειας θα υπολογιστεί από την ποσότητα:

$$L(\beta)=\prod_{i=1}^{r}\frac{\exp(\beta' Z_{(i)})}{\sum_{j\in R(t_{(i)})}\exp(\beta' Z_j)} \qquad (4.2)$$

η οποία ονομάζεται **μερική πιθανοφάνεια (partial likelihood)** και ο Cox καθώς και αρκετοί άλλοι έχουνε δείξει ότι μπορεί να χρησιμοποιηθεί για να βγάλουμε συμπεράσματα σχετικά με το διάνυσμα των παραμέτρων των β΄. Μεγιστοποιώντας την παραπάνω σχέση ως προς β προκύπτει ο ΕΜΠ $\hat{\beta}$, ο οποίος είναι ασυμπτωτικά κανονικός, συνεπής και αμερόληπτος εκτιμητής του β εφόσον ικανοποιεί τις ιδιότητες ενός ΕΜΠ.

Στην σχέση (4.2) παρατηρούμε πως ο παρανομαστής περιλαμβάνει το άθροισμα των τιμών exp(β΄Z) για όλα τα άτομα που βρίσκονται σε κίνδυνο την χρονική στιγμή $t_{(i)}$. Το γινόμενο υπολογίζεται για όλα τα άτομα για τα οποία ο χρόνος θανάτου είναι γνωστός, άρα άτομα με λογοκριμένο χρόνο επιβίωσης δεν συνεισφέρουν στον

αριθμητή για τον υπολογισμό της συνάρτησης πιθανοφάνειας ενώ λαμβάνονται υπόψη στο σύνολο των ατόμων που βρίσκονται σε κίνδυνο. Ακόμη η συνάρτηση μερικής πιθανοφάνειας εξαρτάται μόνο από την διάταξη των χρόνων θανάτου καθώς αυτή προσδιορίζει στην συνέχεια και το σύνολο των ατόμων που βρίσκονται σε κίνδυνο σε κάθε χρόνο θανάτου.

Επίσης να αναφέρουμε ότι η εξίσωση της σχέσης (4.2) θα μπορούσε να γραφτεί και ως εξής :

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\beta' z_i)}{\sum_{j \in R(t_i)} \exp(\beta' Z)} \right)^{\delta_i}$$

στην περίπτωση που θεωρήσουμε πως έχουμε μια δείκτρια μεταβλητή, δ_i, η οποία θα παίρνει την τιμή 0 αν έχουμε λογοκριμένη παρατήρηση και την τιμή 1 για πλήρη χρόνο ζωής.

Να σημειώσουμε ότι η εξίσωση μερικής πιθανοφάνειας έχει ισχύ και είναι εφικτή να υπολογιστεί όταν δεν υπάρχουν δεσμοί στο σετ των δεδομένων που μελετάμε, δηλ. όταν δεν συμβαίνουν περισσότεροι του ενός θανάτου την ίδια χρονική στιγμή t. Στην περίπτωση που υπάρχουν δεσμοί τότε οι εκτιμητές μερικής πιθανοφάνειας δεν υπολογίζονται εύκολα, απαιτούν μια χρονοβόρα διαδικασία καθώς απαιτείται κάποιο υπολογιστικό πρόγραμμα για την εύρεσή τους. Την λύση στο πρόβλημα αυτό έχουν δώσει τεχνικές που έχουν αναπτυχθεί από τους Breslow και Efron, οι οποίες επιδιώκουν να προσεγγίσουμε τους εκτιμητές μερικής πιθανοφάνειας.

(Οι περισσότερες πληροφορίες για την συγκεκριμένη ενότητα και την επόμενη προέρχονται από το βιβλίο του D. Collett, 2005, και κάποιες από το Boist 515 Lecture 17)

4.4 Έλεγχοι υποθέσεων του διανύσματος των παραμέτρων β

Όπως αναφέραμε και παραπάνω ο εμπ των β παραμέτρων επιτυγχάνεται με την μεγιστοποίηση της παρακάτω ποσότητας:

$$l(\beta) = \log L(\beta) = \sum_{i=1}^{d}\sum_{k=1}^{p}\beta_k Z_{(i)k} - \sum_{i=1}^{d}\log(\sum_{j\in Rt(i)}\exp(\sum_{k=1}^{p}\beta_k Z_{jk}))$$

Συγκεκριμένα προκύπτει λύνοντας το παρακάτω σύστημα μερικών παραγώγων ως προς β, τις οποίες έχουμε εξισώσει με μηδέν.

$$U(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{d}(Z_{(i)} - \frac{\sum_{j\in R_{t(i)}} Z_j \exp(\beta' Z_j)}{\sum_{j\in Rt_{(i)}}\exp(\beta' Z_j)} \qquad (4.3)$$

με $U(\beta)=(U_1(\beta),U_2(\beta),\ldots U_p(\beta))'$ δηλώνουμε το διάνυσμα των σκορ (score vector).

Η εξίσωση (4.3) λύνεται με χρήση της μεθόδου Newton – Raphson, για να βρούμε τον εμπ β αρκεί ο πίνακας των δεύτερων μερικών παραγώγων ως προς β να είναι αρνητικά ορισμένος δηλ.

$$\frac{\partial^2}{\partial \beta^2} l(\beta) = -I_0(\beta)$$

Για να ελέγξουμε υποθέσεις (τις περισσότερες φορές η υπόθεση που επιδιώκουμε να ελέγξουμε έχει την εξής μηδενική υπόθεση: H_0: $\beta=\beta_0$) σχετικές με το διάνυσμα των παραμέτρων β χρησιμοποιούμε συνήθως τις τρεις παρακάτω προσεγγίσεις:

- **Likelihood ratio test,** σύμφωνα με το τεστ αυτό για δύο εμφωλευμένα μοντέλα υπολογίσουμε το διπλάσιο της διαφοράς μεταξύ της log μερικής πιθανοφάνειας στην θέση $\hat{\beta}$ και της log μερικής πιθανοφάνειας στην θέση $\hat{\beta}_0$ δηλ,

$$X^2_{LR} = -2\log(\frac{L(\hat{\beta}_0)}{L(\hat{\beta})} = 2[l(\hat{\beta}) - l(\hat{\beta}_0)] \sim X^2_p$$

- **Wald test,** η στατιστική συνάρτηση για την εκτιμώμενη τιμή της παραμέτρου $\hat{\beta}$ ακλουθεί προσεγγιστικά την πολυδιάστατη κανονική κατανομή με μέση τιμή β και πίνακα διακυμάνσεων- συνδιακυμάνσεων $I^{-}_0(\beta)$. Επομένως το στατιστικό του Wald test θα έχει την μορφή:

$$X^2_W = (\hat{\beta} - \hat{\beta}_0)' I_0 (\hat{\beta} - \hat{\beta}_0) \sim X^2_p$$

Να αναφέρουμε ότι το Wald test δίνει αξιόπιστα αποτελέσματα όταν το δείγμα των χρόνων ζωής που μελετάμε είναι μεγάλο.

- **Score test,** σε σχέση με τα δύο παραπάνω τεστ το Score test είναι το μόνο που δεν χρησιμοποιεί τον εμπ β για τον υπολογισμό του στατιστικού. Σε αντίθεση με τα άλλα δύο χρησιμοποιεί το διάνυσμα των σκορ U(β), το οποίο έχει αποδειχθεί ότι ακολουθεί προσεγγιστικά με πολυδιάστατη κανονική κατανομή με μέση τιμή 0 και πίνακα διακυμάνσεων – συνδιακυμάνσεων $I_0(\beta)$. Έτσι η στατιστική συνάρτηση για το συγκεκριμένο τεστ εκτιμάται από την σχέση:

$$X^2_{Sc} = (U(\hat{\beta}_0) - 0)' I^{-1}_0 (U(\hat{\beta}_0) - 0) \sim X^2_p$$

Ασυμπτωτικά οι τρείς παραπάνω έλεγχοι είναι ισοδύναμοι.

(Rodriguez 2005)

4.5 Διάστημα εμπιστοσύνης για την παράμετρο β του μοντέλου αναλογικού κινδύνου του Cox

Για να υπολογίσουμε ένα διάστημα εμπιστοσύνης για την παράμετρο β, βασιζόμαστε στο στατιστικό

$$\frac{\hat{\beta}}{s.e(\hat{\beta})}$$

το οποίο υποθέτουμε ότι ακολουθεί την τυποποιημένη κανονική κατανομή.
Επομένως ένα 100(1-α)% διάστημα εμπιστοσύνης για την παράμετρο β θα υπολογίζεται από την σχέση:

$$\hat{\beta} \pm z_{a/2} s.e(\hat{\beta})$$

Οι παραπάνω ποσότητες δίνονται από όλα τα στατιστικά προγράμματα ανάλυσης δεδομένων χρόνων επιβίωσης και έτσι το παραπάνω δ.ε. υπολογίζεται εύκολα.

Το δ.ε. χρησιμοποιείται και για ένα στατιστικό έλεγχο προκειμένου να ελέγξουμε αν η παράμετρος β=0, αφού αν το δ.ε για το β με συντελεστή εμπιστοσύνης (1-α) δεν περιλαμβάνει την τιμή 0 τότε έχουμε ένδειξη ότι η τιμή της παραμέτρου είναι

διαφορετική του μηδέν. Πιο συγκεκριμένα η μορφή ενός στατιστικού ελέγχου για την παράμετρο β θα ήταν η εξής:

H_0: β=0 έναντι της

H_1: β≠0

Στην συνέχεια αφού έχουμε υπολογίσει την τιμή του στατιστικού θα χρησιμοποιήσουμε τις ποσοστιαίες τιμές της τυποποιημένης κανονικής κατανομής για συγκεκριμένο ε.σ, α, προκειμένου να υπολογίσουμε την τιμή p-value κάτω από την οποία απορρίπτουμε την μηδενική υπόθεση. Αποδοχή της H_0 σημαίνει ότι η ερμηνευτική μεταβλητή δεν προσθέτει κάποια πληροφορία στο μοντέλο και επομένως μπορούμε να μην την λάβουμε υπόψη στην ανάλυσή μας.

Στο μοντέλο του Cox όμως συνήθως περιλαμβάνονται περισσότερες από μια επεξηγηματικές μεταβλητές, σε αυτή την περίπτωση θα πρέπει να είμαστε προσεκτικοί διότι αν επιθυμούσαμε να ελέγξουμε αν για κάποια X_i ερμηνευτική μεταβλητή η παράμετρος της $\beta_i=0$ τότε αυτό θα το κάναμε δοθέντος ότι όλες οι υπόλοιπες μεταβλητές περιλαμβάνονται στο μοντέλο.

Το τεστ αυτό, η στατιστική συνάρτηση του οποίου συμπίπτει με αυτή του Wald test για τοπικούς ελέγχους, δεν μας βοηθάει τόσο πολύ στο να συμπεράνουμε ποιες μεταβλητές είναι σημαντικές για την ανάλυση των δεδομένων μας σύμφωνα με το μοντέλο παλινδρόμησης του Cox. Αυτό γιατί οι εκτιμήσεις των β που παίρνουμε δεν είναι αναγκαίο να είναι όλες ανεξάρτητες μεταξύ τους και κατά συνέπεια κάποιες από αυτές μπορεί να σχετίζονται. Έτσι αν καταλήξουμε στο συμπέρασμα ότι κάποια μεταβλητή δεν χρειάζεται στο μοντέλο, σύμφωνα με το Wald test, και την αφαιρέσουμε τότε μπορεί κάποια άλλη μεταβλητή που ήταν σημαντική πριν να παύει να είναι, ισχύει και το αντίστροφο. Για αυτό το λόγο έχουν αναπτυχθεί άλλα πιο ικανοποιητικά τεστ για τον έλεγχο σημαντικότητας των επεξηγηματικών μεταβλητών ενός μοντέλου αναλογικού κινδύνου.

Ιδιαίτερο ενδιαφέρον όμως παρουσιάζει και το δ.ε. για το λόγο κινδύνου (έστω ψ=exp(β)) το οποίο υπολογίζεται εκθετοποιώντας το δ.ε που πήραμε για την παράμετρο β. Συγκεκριμένα ένα 100(1-α)% δ.ε. για τον πραγματικό λόγο κινδύνου θα υπολογίζεται από την σχέση:

$$\hat{\psi} \pm z_{a/2} s.e.(\hat{\psi})$$

ή

$$\exp(\hat{\beta}) \pm z_{a/2} \exp(\hat{\beta}) s.e(\hat{\beta})$$

Στην πράξη σχεδόν πάντα εκθετοποιούμε το δ.ε της παραμέτρου β αφού μια λογαριθμική εκτίμηση προσεγγίσει καλύτερα μια κανονική κατανομή από ότι ο ίδιος ο λόγος κινδύνου.

(Για περισσότερες πληροφορίες βλέπε, Collet 2000, Αντζουλάκος 2009 και Therneau & Grambsch 2000)

4.6 Υπολογισμός της συνάρτησης πιθανοφάνειας στην περίπτωση ύπαρξης δεσμών

Όπως αναφέραμε και παραπάνω στην περίπτωση που στα δεδομένα μας υπάρχουν δεσμοί, η εξίσωση (4.2) δεν μπορεί να χρησιμοποιηθεί για να υπολογίσουμε την συνάρτηση πιθανοφάνειας, για αυτό είναι αναγκαίο να μοντελοποιήσουμε την εξίσωση αυτή ώστε να λαμβάνει υπόψη την ύπαρξη δεσμών στο σετ των δεδομένων που μελετάμε.

Να αναφέρουμε ότι δεσμοί εμφανίζονται όχι μόνο στην περίπτωση που συμβαίνουν περισσότεροι του ενός θανάτου την ίδια χρονική στιγμή αλλά και όταν εμφανίζονται μια ή περισσότερες λογοκριμένες παρατηρήσεις σε ένα χρόνο θανάτου. Για αυτό προκειμένου να αντιμετωπίσουμε δυσκολίες που παρουσιάζονται στον υπολογισμό της συνάρτησης πιθανοφάνειας υποθέτουμε ότι, όταν παρατηρούνται ταυτόχρονα λογοκριμένοι χρόνοι επιβίωσης και θάνατοι, η λογοκρισία συμβαίνει πάντα ύστερα από όλους τους θανάτους.

Από όλες τις εξισώσεις που έχουν προταθεί για τον υπολογισμό της συνάρτησης πιθανοφάνειας στην περίπτωση που στα δεδομένα μας υπάρχουν δεσμοί, η πιο κατάλληλη (με την έννοια ότι μας δίνει την πιο ακριβή προσέγγιση) είναι αυτή που πρότειναν οι Kalbfleisch και Prentice το 2002 , όμως η μορφή της είναι πολύπλοκη και για αυτό δεν και θα την αναφέρουμε καθόλου.

Στην πράξη για την εκτίμηση της συνάρτησης πιθανοφάνειας στην περίπτωση δεσμών χρησιμοποιούνται οι προσεγγίσεις που έχουν προταθεί από τους Breslow και Efron. Προτού όμως εισάγουμε τις εξισώσεις αυτές θα πρέπει να ορίσουμε κάποιες νέες ποσότητες.

Έστω ότι με S_i συμβολίζουμε το άθροισμα των διανυσμάτων Z_j που αντιστοιχούν στα άτομα που πεθαίνουν την χρονική στιγμή $t_{(i)}$ (δηλ.

$$S_i = \sum_{j \in D(t_{(i)})} Z_j \ ,1 \le i \le r$$

άθροισμα συμμεταβλητών εκείνων των ατόμων που ανήκουν στο διάστημα (1,r)), με $D(t_{(i)})$ δηλώνουμε το πλήθος των ατόμων που πεθαίνουν την χρονική στιγμή $t_{(i)}$ και όπως αναφέραμε και παραπάνω με $R(t_{(i)})$ συμβολίζουμε το σύνολο των ατόμων που βρίσκονται σε κίνδυνο την χρονική στιγμή $t_{(i)}$.

- **Προσέγγιση Breslow**

Η προσέγγιση αυτή για την συνάρτηση πιθανοφάνειας στην περίπτωση δεσμών είναι πιο απλή και πιο κατανοητή από όλες και δίνεται από την εξίσωση:

$$\prod_{i=1}^{r} \frac{\exp(\beta' S_i)}{\left(\sum_{j \in R(t_{(i)})} \exp(\beta' Zj) \right)^{d_i}}$$

Σύμφωνα με την προσέγγιση αυτή οι (d_i) το πλήθος των θάνατοι θεωρούνται διαφορετικοί και ότι συμβαίνουν διαδοχικά. Η προσέγγιση αυτή είναι παρόμοια με αυτήν που είχε προτείνει και ο Peto το 1972.

- **Προσέγγιση Efron**

Η εξίσωση που πρότεινε ο Efron για την συνάρτηση πιθανοφάνειας είναι η εξής:

$$\prod_{i=1}^{r} \frac{\exp(\beta' S_i)}{\prod_{k=1}^{d_i} \left(\sum_{j \in R(t_{(i)})} \exp(\beta' Z_j) - \frac{k-1}{d_i} \sum_{j \in D(t_{(i)})} \exp(\beta' Z_i) \right)}$$

Στην περίπτωση που έχουμε μικρό αριθμό δεσμών οι δύο προσεγγίσεις δίνουν παρόμοια αποτελέσματα , σε αντίθετη περίπτωση η προσέγγιση του Efron προτιμάται έναντι αυτής του Breslow αφού οδηγεί σε καλύτερα αποτελέσματα.

- **Προσέγγιση Cox**

Η προσέγγιση αυτή προτάθηκε από τον Cox το 1972 και είναι γνωστή ως **διακριτή πιθανοφάνεια,** αφού υποθέτει ότι τα δεδομένα που μελετάμε προέρχονται από μια διακριτή κατανομή χρόνων ζωής, και υπολογίζεται από την εξίσωση:

$$\prod_{i=1}^{r} \frac{\exp(\beta' S_i)}{\sum_{j \in R(t_{(i)}; d_i)} \exp(\beta' S_j)}$$

όπου με $R(t_{(i)};d_i)$ δηλώνουμε το πλήθος των ατόμων (d_i) που έχουν απομακρυνθεί από το σύνολο $R(t_{(i)})$, που είναι το σύνολο των ατόμων που βρίσκονται σε κίνδυνο την χρονική στιγμή $t_{(i)}$.

(Για περισσότερα βλέπε D.W.Hosmer, JR. & S. Lemeshow, 1999, M. Stevenson, 2009, Αντζουλάκος, 2009)

4.7 Εκτίμηση της συνάρτησης επιβίωσης με βάση το μοντέλο αναλογικού κινδύνου του Cox

Γνωρίζουμε ότι κάτω από την υπόθεση του αναλογικού κινδύνου η συνάρτηση επιβίωσης υπολογίζεται από την σχέση:

$$S(t \mid z) = [S_0(t)]^{\exp(\beta' z)} \qquad (4.3)$$

Σύμφωνα με την σχέση αυτή για να εκτιμήσουμε την συνάρτηση επιβίωσης μιας υποομάδας που μας ενδιαφέρει από το σύνολο των δεδομένων χρειάζεται μόνο να προσδιορίσουμε τους συντελεστές τις παλινδρόμησης (β) και την αναφορική συνάρτηση επιβίωσης. Οι συντελεστές αυτοί όπως προαναφέραμε και σε προηγούμενο κεφάλαιο εκτιμώνται με βάση την μέθοδο της μερικής πιθανοφάνειας, εκείνο που μας δυσκολεύει λίγο είναι ο προσδιορισμός της αναφορικής συνάρτησης επιβίωσης. Μέσω της μεθόδου της μέγιστης πιθανοφάνειας αποκτάμε και μια πρώτη εκτίμηση για την αναφορική συνάρτηση επιβίωσης, η οποία βλέπουμε ότι είναι ασυνεχής στις διατεταγμένες χρονικές στιγμές, $t_{(i)}$, που συμβαίνουν οι θάνατοι.

Καθοριστικής σημασίας για την εκτίμηση της αποτελεί ο προσδιορισμός της ποσότητας:

$$\hat{a}_i = 1 - \frac{d_i}{n_i}$$

με την οποία υπολογίζουμε έναν εκτιμητή για την υπό συνθήκη πιθανότητα επιβίωσης στους διατεταγμένους χρόνους επιβίωσης, $t_{(i)}$.

Όπως είδαμε στο 2^{ο} κεφάλαιο ο Kaplan – Meier εκτιμητής δηλώνει ακριβώς το γινόμενο μιας σειράς υπό συνθήκη ατομικών πιθανοτήτων επιβίωσης. Έτσι η έκφραση για την υπό συνθήκη πιθανότητα επιβίωσης που οδηγεί στο συγκεκριμένο εκτιμητή θα ισούται με:

$$a_i = \frac{S(t_{(i)})}{S(t_{(i-1)})}$$

Αντίστοιχα για το αναλογικό μοντέλο κινδύνου ορίζουμε ότι η υπό συνθήκη αναφορική συνάρτηση επιβίωσης θα εκτιμάται μέσω της σχέσης:

$$a_i = S_0(t_{(i)}) / S_0(t_{(i-1)})$$

Επομένως η υπό συνθήκη πιθανότητα επιβίωσης θα ισούται με:

$$\frac{S(t_{(i)} \mid z)}{S(t_{(i-1)} \mid z)} = \frac{[S_0(t_{(i)})]^{\exp(z'\beta)}}{[S_0(t_{(i-1)})]^{\exp(z'\beta)}} = \left\{ \frac{S_0(t_{(i)})}{S_0(t_{(i-1)})} \right\}^{\exp(z'\beta)} = a_i^{\exp(z'\beta)}$$

Για να εκτιμήσουμε την αναφορική υπό συνθήκη πιθανότητα επιβίωσης θα πρέπει να λύσουμε ως προς α_i την παρακάτω εξίσωση :

$$\sum_{j \in D(t_{(i)})} \frac{\exp(z'\hat{\beta})}{1 - a_i^{\exp(z'\hat{\beta})}} = \sum_{j \in R(t_{(i)})} \exp(z'\hat{\beta}) \qquad (4.4)$$

όπου με $R(t_{(i)})$ συμβολίσουμε το σύνολο των ατόμων που βρίσκονται σε κίνδυνο την χρονική στιγμή $t_{(i)}$, με $D(t_{(i)})$ δηλώνουμε το σύνολο των ατόμων με χρόνο ζωής ίσο με $t_{(i)}$ και με $\hat{\beta}$ τους εκτιμητές των συντελεστών παλινδρόμησης που έχουν προκύψει από την μερική πιθανοφάνεια.

Στην περίπτωση που στο σετ των δεδομένων που μελετάμε δεν υπάρχουν δεσμοί τότε μια λύση της εξίσωσης (4.4) είναι η εξής:

$$\hat{a}_i = \left[1 - \frac{\exp(z'\hat{\beta})}{\sum_{j \in R(t_{(i)})} \exp(z'\hat{\beta})} \right]^{[\exp(z'\hat{\beta})]-1}$$

Στην περίπτωση όμως που υπάρχουν δεσμοί στα δεδομένα που μελετάμε η λύση της εξίσωσης (4.4) είναι δυνατή μέσω επαναληπτικών μεθόδων. Ωστόσο μια προσεγγιστική λύση μπορεί να βρεθεί, η προσέγγιση αυτή οφείλεται στον Breslow (1974) και προκύπτει αν αντικαταστήσουμε το αριστερό μέρος της εξίσωσης (4.4) με την προσέγγισή της δηλ:

$$a_i^{\exp(z'\hat{\beta})} \approx 1 + \exp(z'\hat{\beta}) \ln(a_i)$$

τότε η λύση που παίρνουμε για την σχέση (4.4) είναι ότι

$$\hat{a}_i = \exp\left(- \frac{d_i}{\sum_{j \in R(t_{(i)})} \exp(z'\hat{\beta})} \right)$$

Βλέπουμε πως ο εκτιμητής για την αναφορική συνάρτηση επιβίωσης είναι και πάλι το γινόμενο ατομικών υπό συνθήκη πιθανοτήτων επιβίωσης, αφού

$$\hat{S}_0(t) = \prod_{t_{(i)} < t} \hat{a}_i$$

Όσο αφορά την αναφορική συνάρτηση κινδύνου, αυτή εφόσον είναι συνάρτηση του εκτιμητή της υπό συνθήκη πιθανότητας επιβίωσης θα ισούται με:

$$\hat{h}_0(t) = 1 - \hat{a}_i$$

Όμως στην πράξη συνήθως αποφεύγουμε να χρησιμοποιήσουμε την εκτίμηση της αναφορικής συνάρτησης κινδύνου διότι έχει αποδειχθεί ότι είναι μια πολλών διαστάσεων συνάρτηση, μας δίνει μια υπερεκτίμηση της πραγματικής τιμή της (Royston and Parmar, 2002) και αυτό γιατί η εκτίμηση της αναφορικής συνάρτησης κινδύνου εξαρτάται άμεσα από τα παρατηρούμενα δεδομένα και έτσι λειτουργεί σαν μια ενοχλητική συνάρτηση. Εξάλλου όπως έχουμε δει σύμφωνα με το μοντέλο του Cox η σχέση μεταξύ των συμμεταβλητών και του ρυθμού κινδύνου μπορεί να προσδιοριστεί χωρίς να χρειάζεται να κάνουμε κάποια υπόθεση για την αναφορική συνάρτηση κινδύνου.

Για τους παραπάνω λόγους επιλέγουμε να χρησιμοποιούμε την εκτίμηση για την αθροιστική αναφορική συνάρτηση κινδύνου (cumulative baseline hazard rate) η οποία λαμβάνεται συναρτήσει της σχέσης (2.9) που ισχύει για την συνάρτηση επιβίωσης. Επομένως η εκτίμηση της αναφορικής αθροιστικής

συνάρτησης κινδύνου και της αναφορικής συνάρτησης επιβίωσης θα δίνονται αντίστοιχα από τις σχέσεις:

$$\hat{H}_0(t) = -\ln[\hat{S}_0(t)] = -\sum_{i:t_{(i)}<t} \ln \hat{a}_i \quad \text{και} \quad \hat{S}_0(t) = \exp(-\hat{H}_0(t))$$

Τέλος, ένας εκτιμητής για την αθροιστική συνάρτηση κινδύνου θα υπολογίζεται από την σχέση:

$$\hat{H}(t \mid z) = -\ln[\hat{S}(t \mid z)] = -\exp(z'\hat{\beta}) \ln[\hat{S}_0(t)]$$

4.8Στρωματοποιημένο μοντέλο του Cox (Stratified Cox model)

Πολλές φορές είναι αναγκαίο να συγκρίνουμε δύο ή περισσότερα σετ δεδομένων χρόνων επιβίωσης, αφού λάβουμε υπόψη κάποιες επιπλέον μεταβλητές που αφορούν το κάθε άτομο του δείγματος. Για τις μεταβλητές αυτές συνήθως γνωρίζουμε ότι ασκούν κάποια επίδραση στο τελικό μοντέλο αναλογικού κινδύνου στο οποίο έχουμε καταλήξει σύμφωνα με μεθόδους επιλογής μεταβλητοτήτων που θα περιγράψαμε στην συνέχεια, απλώς θεωρούμε ότι οι επιδράσεις αυτές ορισμένες φορές είναι

δευτερεύουσας σημασίας σε σχέση με αυτές που ασκούν οι άλλες μεταβλητές, όμως προκειμένου να ερμηνεύσουμε και να κατανοήσουμε καλύτερα τα υπό μελέτη δεδομένα θα πρέπει να εξετάσουμε και τις δευτερεύουσες αυτές επιδράσεις. Η στρωματοποίηση διαιρεί τα δεδομένα του δείγματος σε διαφορετικές ομάδες, στρώματα όπως ονομάζονται, για κάθε μια από τις οποίες έχουμε μια ξεχωριστή αναφορική συνάρτηση κινδύνου και κοινές τιμές για τους συντελεστές του διανύσματος β. Υποθέτουμε δηλαδή ότι τα άτομα i=1,2,…,n_1 θα ανήκουν στο 1^o στρώμα ενώ τα άτομα n_{1+1},…, n_1+n_2 θα ανήκουν στο 2^o στρώμα κτλ. Έτσι για το i άτομο που ανήκει στο k στρώμα ο κίνδυνος θα ισούται με:

$$h_i(t) = h_{k0}(t)\exp(z_i\beta)$$

Η μορφή της εξίσωσης μερικής πιθανοφάνειας για την εκτίμηση της παραμέτρου β θα είναι παρόμοια με αυτήν που αναφέραμε στο μοντέλο αναλογικού κινδύνου του Cox με την μόνη διαφορά ότι τώρα θα περιλαμβάνει το δείκτη k, που δηλώνει το στρώμα στο οποίο αναφερόμαστε. Επομένως η συνεισφορά του k στρώματος στην εξίσωση μερικής πιθανοφάνειας θα ισούται με:

$$l_{kp}(\beta) = \prod_{k=1}^{n_k}\left(\frac{\exp(z'_{ki}\beta)}{\sum_{j\in R(t_{ki})}\exp(z'_{kj}\beta)}\right)^{\delta_{ki}}$$

όπου με n_k ορίζουμε τον αριθμό των ατόμων που περιλαμβάνονται στο k στρώμα, με t_{ki} δηλώνουμε την i παρατηρούμενη τιμή του χρόνου στο k στρώμα, με δ_{ki} δηλώνουμε τη λογοκριμένη τιμή στο χρόνο t_{ki}, με $R(t_{ki})$ συμβολίζουμε τον αριθμό των που βρίσκονται σε κίνδυνο στο χρόνο t_{ki}, και με z_{ki} συμβολίζουμε το διάνυσμα των p συμμεταβλητών.

Έτσι η ολική (στρωματοποιημένη) εξίσωση μερικής πιθανοφάνειας θα ισούται με:

$$l_{kp}(\beta) = \prod_{k=1}^{K} l_{kp}(\beta)$$

Στην μέθοδο στρωματοποίησης τόσο το διάνυσμα των σκορ όσο και ο πίνακας πληροφορίας των z_k συμμεταβλητών έχουν παρόμοια αθροίσματα με αυτά του αναλογικού μοντέλου κινδύνου, εκείνο που αλλάζει είναι ότι πλέον ο μέσος των συμμεταβλητών, $\bar{z}_k(t)$, θα είναι ο σταθμισμένος μέσος για όλα τα άτομα που ανήκουν στο k στρώμα και βρίσκονται σε κίνδυνο στο χρόνο t και η διακύμανση, $V_k(t)$, θα μας δίνει τον σταθμισμένο πίνακα διακύμανσης των ατόμων στο k στρώμα.

Οι μεταβλητές ως προς τις οποίες στρωματοποιούμε συνήθως αντιμετωπίζονται ως κατηγορικές μεταβλητές, αυτό λόγω της ιδιότητας της μεθόδου να χωρίζει το υπό μελέτη δείγματα σε διαφορετικές ομάδες. Μέσω της στρωματοποίησης λοιπόν, αποκτάμε μια απλή λύση για ένα αναλογικό ή μη αναλογικό μοντέλο κινδύνου πάνω σε μια ονομαστική κλίμακα. Επίσης με την στρωματοποίηση δεν χρειάζεται να εκτιμήσουμε τις επιδράσεις κάποιου παράγοντα όχλησης (nuisance factor), το γεγονός αυτό αποτελεί και το μεγαλύτερο πλεονέκτημα της μεθόδου. Ενώ κάποια από τα μειονεκτήματά της είναι ότι δεν μας δίνει απευθείας κάποια εκτίμηση σχετική με την σημαντικότητα της επίδρασης του στρώματος, δηλ. δεν μας δίνει κάποια τιμή p-value. Ακόμη η ακρίβεια των εκτιμώμενων συντελεστών και η ισχύς της υπόθεσης που εξετάσουμε μπορεί να είναι μειωμένη εάν υπάρχει ένας μεγάλος αριθμός στρωμάτων. Από την άλλη πλευρά μικρός αριθμός ατόμων σε κάθε στρώμα θα είχε σαν συνέπεια η εκτίμηση για την αναφορική συνάρτηση επιβίωσης να είχε βασιστεί σε μια μεγαλύτερη εκτίμηση της διακύμανσης από αυτήν που προκύπτει από κάποιο στρώμα με πιο πολλά δεδομένα. Το πρόβλημα αυτό δεν είναι τόσο σοβαρό καθώς η διακύμανση που εκτιμάται για τους εκτιμώμενους συντελεστές του μοντέλου είναι συνάρτηση τόσο του μεγέθους του δείγματος όσο και του συνολικού αριθμού των χρόνων επιβίωσης, χαρακτηριστικό που ισχύ και σε ένα στρωματοποιημένο μοντέλο.

Τέλος να αναφέρουμε ότι στο στρωματοποιημένο μοντέλο του Cox μπορούμε να προσθέσουμε και όρους αλληλεπίδρασης και χρόνο- εξαρτημένες μεταβλητές (θα τις περιγράψουμε αναλυτικά στην επόμενη παράγραφο) αν είναι αναγκαίο, κάνοντας τις απαραίτητες αλλαγές και ακολουθώντας μια συγκεκριμένη διαδικασία.

Για περισσότερα βλέπε: Hosmer & Lemeshow 1999, Dhananjay Kumar & Bengt Klefsjo 1993 και Therneau & Grambsch 2000)

4.9 Γενικεύσεις του μοντέλου αναλογικού κινδύνου (Extensions of the proportional hazards model)

Το μοντέλο αναλογικού κινδύνου του Cox όπως έχουμε αναφέρει και παραπάνω αποτελεί μια από τις πιο γνωστές μεθόδους ανάλυσης δεδομένων επιβίωσης. Σε βραχυπρόθεσμες follow - up μελέτες η υπόθεση του σταθερού λόγο κινδύνου είναι πολύ λογική, όμως σε μακράς διάρκειας follow - up μελέτες είναι πιο λογικό να υποθέσουμε ότι ο χρόνος επηρεάζει κατά κάποιο τρόπο το λόγο κινδύνου.

Όταν η υπόθεση της αναλογικότητας παραβιάζεται τα αποτελέσματα που παίρνουμε από ένα μοντέλο παλινδρόμησης του Cox δεν είναι πλέον αξιόπιστα, στην περίπτωση αυτή θα πρέπει να θεωρήσουμε εναλλακτικά μοντέλα.

Στο αναλογικό μοντέλο κινδύνου (σχέση 4.1) η παράλειψη μιας σημαντικής μεταβλητής καταλήγει σε μεροληπτικές εκτιμήσεις τόσο των συντελεστών παλινδρόμησης (β_i) όσο και του ρυθμού κινδύνου. Η αιτία της εν λόγου μεροληψίας βασίζεται στο γεγονός ότι ο χρονο – εξαρτώμενος ρυθμός κινδύνου επιφέρει αλλαγές στην σύνθεση του υπό μελέτη πληθυσμού στην διάρκεια του χρόνου.

Έχουν αναπτυχθεί αρκετά μοντέλα προκειμένου να αντιμετωπίσουν την περίπτωση μη αναλογικού κινδύνου π.χ. μοντέλα με χρονο – εξαρτώμενες επιδράσεις των μεταβλητών (προκύπτουν αν πάρουμε αλληλεπιδράση των μεταβλητών που μελετάμε με κάποια συναρτησιακή μορφή του χρόνου), μοντέλα εξασθένησης (frailty models), cure mixture models κτλ.

Στην εργασία αυτή πιο πολύ θα ασχοληθούμε με τα δύο πρώτα μοντέλα.

4.9.1 Χρόνο – εξαρτημένες μεταβλητές (time – depended variables)

Σε όσα αναφερθήκαμε παραπάνω είχαμε υποθέσει ότι οι τιμές όλων των μεταβλητών που μελετάμε σε ένα μοντέλο αναλογικού κινδύνου του Cox καθορίζονται την χρονική στιγμή έναρξης της μελέτης τους, την στιγμή μηδέν όπως λέγεται, και οι τιμές τους δεν αλλάζουν κατά την διάρκεια του χρόνου παρακολούθησης τους. Όμως παρουσιάστηκαν περιπτώσεις στις οποίες η συνάρτηση κινδύνου φαινόταν να εξαρτάται πιο πολύ από την τρέχουσα τιμή μιας μεταβλητής από ότι την τιμή που αυτή είχε την χρονική στιγμή μηδέν. Έτσι το ενδιαφέρον

μελέτης της συνάρτησης κινδύνου στην περίπτωση που στο μοντέλο μας περιλαμβάνονται χρόνο –εξαρτημένες ήταν μεγάλο. Βέβαια θα πρέπει να είμαστε σίγουροι για την αναγκαιότητά τους καθώς έχει αποδειχτεί ότι οι χρόνο – εξαρτώμενες μεταβλητές οδηγούν συνήθως σε ένα πιο πολύπλοκο μοντέλο, τόσο από πλευράς εφαρμογής όσο και ερμηνείας. Θα πρέπει λοιπόν να εξετάσουμε πολύ προσεκτικά την εισαγωγή τους στο μοντέλο.

Οι χρόνο – εξαρτώμενες μεταβλητές συνήθως διακρίνονται σε δύο κατηγορίες εσωτερικές ή εξωτερικές (internal or external).

Οι εσωτερικές χρόνο – εξαρτώμενες μεταβλητές αναφέρονται συγκεκριμένα σε ένα άτομο και μπορούν να μετρηθούν μόνο αν το άτομο είναι ζωντανό. Δεδομένα τέτοιας μορφής συνήθως εμφανίζονται σε επαναλαμβανόμενες μετρήσεις όπου μας ενδιαφέρει για τον κάθε ασθενή το πώς μεταβάλλονται οι τιμές συγκεκριμένων χαρακτηριστικών στην διάρκεια του χρόνου παρακολούθησης τους.

Από την άλλη οι εξωτερικές χρόνο – εξαρτώμενες μεταβλητές δεν απαιτούν την επιβίωση του ασθενή για μέτρησή τους. Μια εξωτερική μεταβλητή είναι μια μεταβλητή που αλλάζει κατά τέτοιο τρόπο ώστε η τιμή της να είναι γνωστή πριν από ένα μελλοντικό χρόνο, παραδείγματος χάριν η μεταβλητή που δηλώνει την ηλικία των ατόμων που συμμετέχουν σε μια μελέτη, η ηλικία ενός ασθενή θα είναι γνωστή για οποιαδήποτε μελλοντική χρονική στιγμή.

Ακόμη να αναφέρουμε ότι χρόνο – εξαρτώμενες μεταβλητές εμφανίζονται και στην περίπτωση που οι συντελεστές μιας σταθερής επεξηγηματικής μεταβλητής είναι συνάρτηση του χρόνου. Στο μοντέλο του Cox είχαμε αναφέρει ότι ο λογάριθμος του λόγου κινδύνου είναι μια σταθερή ποσότητα ανεξάρτητη του χρόνου, αν όμως ο λόγος αυτός ήταν συνάρτηση του χρόνου τότε ο συντελεστής της επεξηγηματικής μεταβλητής που αλλάζει τιμή με το χρόνο θα αναφέρεται ως χρόνο – εξαρτώμενος συντελεστής (time varying coefficient). Όπως είναι αναμενόμενο στην περίπτωση αυτή δεν ισχύει η υπόθεση αναλογικού κινδύνου. Από δω και πέρα το μοντέλο που περιλαμβάνει χρόνο – εξαρτώμενες μεταβλητές θα αποκαλείται *γενικευμένο μοντέλο παλινδρόμησης του Cox.*

(Collet 2000)

4.9.1.1 Το γενικευμένο μοντέλο παλινδρόμησης του Cox (Cox regression Model)

Σύμφωνα με το μοντέλο αναλογικού κινδύνου, ο κίνδυνος του θανάτου του i ατόμου στο χρόνο t μπορεί να γραφτεί στην μορφή:

$$h(t) = h_0(t)[\exp(\sum_{j=1}^{p} \beta_j z_{ji})]$$

όπου η $h_0(t)$ δηλώνει την αναφορική συνάρτηση κινδύνου και η z_{ji} δηλώνει την παρατηρούμενη τιμή για την j (j=1,2,…p) επεξηγηματική μεταβλητή για το i άτομο (i=1,2,…n).

Επεκτείνοντας το μοντέλο αυτό στην περίπτωση που έχουμε χρόνο – εξαρτώμενες μεταβλητές, θα αποκτήσει την εξής μορφή:

$$h_i(t) = h_0(t)[\exp(\sum_{j=1}^{p} \beta_j z_{ji}(t))]$$

Στο μοντέλο αυτό υποθέτουμε ότι η αναφορική συνάρτηση κινδύνου αναφέρεται σε ένα άτομο για τον οποίον όλες οι μεταβλητές ισούνται με μηδέν στο χρόνο έναρξης της μελέτης. Ακόμη όπως μπορούμε να δούμε οι τιμές της μεταβλητής $z_{ji}(t)$ εξαρτώνται από το χρόνο, t, έτσι και ο σχετικός κίνδυνος $h_i(t)/h_0(t)$ θα εξαρτάται από το χρόνο t, επομένως ο κίνδυνος του θανάτου δεν θα είναι αναλογικός του αναφορικού κινδύνου, δεν θα ισχύει το μοντέλο αναλογικού κινδύνου όπως προαναφέραμε. Οπότε για δύο οποιοδήποτε άτομα, έστω το i και το k, ο λόγος κινδύνου θα υπολογίσετε από την σχέση:

$$\frac{h_i(t)}{h_k(t)} = \exp[\beta_1(z_{i1}(t) - z_{ki}(t)) + ... + \beta_p(z_{ip}(t) - z_{kp(}t))]$$

Στην περίπτωση των χρόνο – εξαρτώμενων μεταβλητών η συνάρτηση μερικής πιθανοφάνειας που θα πρέπει νε μεγιστοποιήσουμε προκειμένου να εκτιμήσουμε τις β παραμέτρους θα δίνεται από την σχέση:

$$\sum_{i=1}^{n} \delta_i \left\{ \sum_{j=1}^{p} \beta_j z_{ji}(t_i) - \log \sum_{l \in R(t_i)} \exp(\sum_{j=1}^{p} \beta_j z j_p(t_i) \right\}$$ (Collett 2000)

4.9.2 Μοντέλα εξασθένησης (Frailty models)

Τα μοντέλα εξασθένησης χρησιμοποιούνται στην ανάλυση επιβίωσης για να περιγράψουν την ετερογένεια ενός δείγματος. Η ιδέα είναι ότι κάθε άτομο έχει τα δικά του χαρακτηριστικά και ελαττώματα, διαφορετικά άτομα παρουσιάζουν διαφορετικές αδυναμίες, έτσι άτομα που είναι πιο αδύναμα από άλλα θα τείνουν να γευτούν το συμβάν νωρίτερα από άλλους (δηλ. αδύναμοι ασθενείς περιμένουμε να πεθάνουν νωρίτερα). Ο όρος εξασθένηση (frailty) χρησιμοποιήθηκε στην ανάλυση δεικτών θνησιμότητας από τον Vapuel et al. (1979) για να δηλώσει μια μη παρατηρήσιμη τυχαία επίδραση την οποία μοιράζονται άτομα με παρόμοιους (μη μετρήσιμους) κινδύνους.

Η έννοια των μοντέλων εξασθένησης (frailty models) αποτελεί ένα βολικό τρόπο για την εισαγωγή τυχαίων επιδράσεων σε ένα μοντέλο, καθώς η ετερογένεια ενός ατόμου είναι μη παρατηρούμενη είναι αναγκαία η εισαγωγή ενός τυχαίου όρου, που μοντελοποιεί τη συνάρτηση κινδύνου ενός ατόμου ή των σχετιζόμενων ατόμων. Πρόκειται στην ουσία για μια τυχαία μεταβλητή η οποία δηλώνει τον επιπλέον κίνδυνο (excess risk).

Η πιο απλή μορφή ενός frailty model είναι η εξής:

$$h(t \mid X, Z) = Z h_0(t) \exp(X\beta)$$

όπου με Z δηλώνουμε την τυχαία επίδραση. Η πιο συνήθης επιλογή για την τυχαία επίδραση Z είναι ότι ακολουθεί μια Γάμμα κατανομή, $Z \sim \Gamma(1/\theta, 1/\theta)$, με μέση τιμή E(Z)=1 και διακύμανση Var(Z)=θ (ωστόσο η καμπύλη εξασθένησης μπορεί να ακολουθεί και άλλες κατανομές όπως την compound Poisson, log-normal, διάφορες τριπαραμετρικές κατανομές κτλ). Εφόσον η αδυναμία δεν είναι παρατηρήσιμη κάποιος θα μπορούσε να χρησιμοποίηση τον περιθώριο κίνδυνο που προκύπτει από την παραπάνω σχέση προκειμένου να πάρει την αναμενόμενη τιμή της τυχαίας επίδρασης.

Έστω ότι με $H_0(t)$ δηλώνουμε την αθροιστική αναφορική συνάρτηση κινδύνου τότε ο περιθώριος κίνδυνος θα υπολογίζεται από την παρακάτω σχέση:

$$h(t \mid X) = \frac{h_0(t)\exp(X\beta)}{1 + \theta H_0(t)\exp(X\beta)}$$

η συγκεκριμένη σχέση είναι γνωστή ως το μοντέλο του Burr.

Η αντίστοιχη συνάρτηση επιβίωσης (S(t)) υπολογίζεται από την σχέση:

$$S(t \mid Z, X) = \exp(-Z\int_0^t h_0(s)ds \exp(\beta' \mathrm{X}))$$

όπου με S(t|Z, X) συμβολίζουμε το κλάσμα της επιβίωσης των ατόμων στο χρόνο t δοθέντος του διανύσματος των παρατηρούμενων συμμεταβλητών X και της frailty Z.

Τα μοντέλα εξασθένησης μπορούν να χρησιμοποιηθούν για να υπολογίσουν την παράλειψη μιας σημαντικής μεταβλητής από το μοντέλο, καθώς επίσης να προσαρμοστούν για τυχόν έλλειψη προσαρμογής του μοντέλου ή απομάκρυνση από την αναλογικότητα.

Αποτελούν στην ουσία μια γενίκευση του αναλογικού μοντέλου κινδύνου του Cox έτσι ώστε ο κίνδυνος ενός ατόμου να εξαρτάται επιπλέον και από μια μη παρατηρούμενη τυχαία μεταβλητή Z, η οποία δρα πολλαπλασιαστικά στην αναφορική συνάρτηση κινδύνου. Χρησιμοποιούνται πιο πολύ στην ανάλυση επιβίωσης λόγο της φύσης των δεδομένων, όπου στις περισσότερες εφαρμογές υποθέτουμε (σιωπηρά) μια ομοιογένεια του υπό εξέταση πληθυσμού, το οποίο σημαίνει ότι όλα τα άτομα του δείγματος υπόκεινται στον ίδιο κίνδυνο (κίνδυνος θανάτου ή επανεμφάνισης της ασθένειας). Σε κάποιες εφαρμογές όμως δεν μπορεί να θεωρήσουμε ότι ο πληθυσμός είναι ομοιογενής θα πρέπει να λάβουμε υπόψη κάποια ετερογένεια, το δείγμα μας μπορεί να αποτελείται από άτομα με διαφορετικό κίνδυνο. Υπάρχουν δύο κυρίως λόγοι για τους οποίους δεν μπορούμε να συμπεριλάβουμε όλους τους σημαντικούς παράγοντες στην ανάλυση επιβίωσης ενός ατόμου. Πρώτος λόγος, οικονομικό κόστος, είναι αδύνατον να μετρηθούν όλες οι σχετικές συμμεταβλητές κάποιας ασθένειας θα χρειαζόταν να λάβουμε υπόψη πάρα πολλές μεταβλητές. Δεύτερος λόγος, κάποιες φορές η σημαντικότητα ορισμένων μεταβλητών είναι άγνωστη.

Μέσω ενός μοντέλου εξασθένησης στοχεύουμε να μετρήσουμε την ετερογένεια που οφείλεται σε μη μετρήσιμες συμμεταβλητές. Σε στατιστικούς όρους ένα frailty model εκφράζει μια τυχαία επίδραση που εφαρμόζεται σε μοντέλα της μορφής time - to – event και ασκεί μια πολλαπλασιαστική επίδραση πάνω στην αναφορική συνάρτηση κινδύνου.

Διακρίνουμε δύο ευρείες κατηγορίες των frailty model:

1. Μοντέλα με μονομεταβλητό χρόνο επιβίωσης ως χρόνο τερματισμού του συμβάντος
2. Μοντέλα που περιγράφουν πολυμεταβλητά τερματικά σημεία επιβίωσης (π.χ. ανταγωνιστικών κινδύνων, επανεμφάνισης της ασθένειας στο ίδιο άτομο, επανεμφάνιση της ασθένειας γενικά κτλ.)

Στην $1^{η}$ περίπτωση χρησιμοποιούμε μοντέλα με μονομεταβλητό (ανεξάρτητο) χρόνο επιβίωσης για να περιγράψουμε την επίδραση μη παρατηρούμενων συμμεταβλητών σε ένα μοντέλο αναλογικού κινδύνου.

Στα μοντέλα αυτά η μεταβλητότητα των δεδομένων επιβίωσης χωρίζεται σε ένα μέρος που εξαρτάται από τους παράγοντες κινδύνου, το οποίο είναι προβλέψιμο θεωρητικά και σε ένα άλλο μέρος που είναι αρχικά μη προβλέψιμο ακόμα και όταν όλη η σχετική πληροφορία είναι γνωστή.

Πιο ενδιαφέρουσα είναι η $2^{η}$ κατηγορία μοντέλων, όπου υποθέτουμε πολυμεταβλητό χρόνο επιβίωσης. Τα μοντέλα αυτά στοχεύουν στο να μετρήσουν την εξάρτηση των αθροιστικών χρόνων αποτυχίας μέσω ενός frailty model, εφόσον γνωρίζουμε την αδυναμία του συμβάν, οι χρόνοι επιβίωσης θα είναι υπό συνθήκη ανεξάρτητοι δοθείσης της εξασθένησης (frailty).

(Για περισσότερα βλέπε: Perperoglou et al 2006 , Andreas Wienke 2003 και David D. Hanagal 2011)

4.10 Επιλογή μεταβλητών

Το μοντέλο που προσαρμόσουμε όταν αναλύουμε δεδομένα που αναφέρονται σε χρόνους επιβίωσης εξαρτάται από την συνάρτηση κινδύνου πάνω σε μια ή περισσότερες επεξηγηματικές μεταβλητές. Επομένως είναι φυσικό να ενδιαφερόμαστε για την ανάπτυξη διαδικασιών οι οποίες θα μας βοηθήσουν να εντοπίσουμε ποιες από όλες τις επεξηγηματικές μεταβλητές που μπορούν να συμπεριληφθούν σε ένα μοντέλο αναλογικού κινδύνου είναι πράγματι σημαντικές,

δηλ. μας δίνουν περισσότερες πληροφορίες από τα δεδομένα και καθιστούν την ανάλυση τους εύκολη και κατανοητή. Από την μια πλευρά η ύπαρξη πολλών μεταβλητών σε ένα μοντέλο σημαίνει μεγάλη μεταβλητότητα, διογκωμένα διαστήματα εμπιστοσύνης και λιγότερο ισχυρούς ελέγχους. Ενώ από την άλλη πλευρά ένα μοντέλο με λίγες ερμηνευτικές μεταβλητές θα μπορούσε να δημιουργήσει σοβαρό πρόβλημα μεροληψίας και την παράλειψη σημαντικών μεταβλητών στο μοντέλο.

Για να συγκρίνουμε εναλλακτικά μοντέλα επιβίωσης βασιζόμαστε στην συνάρτηση πιθανοφάνειας (L). Αυτό διότι η συνάρτηση πιθανοφάνειας είναι ακριβώς η ποσότητα που αθροίζει τις πληροφορίες που περιέχουν τα δεδομένα σχετικά με τις άγνωστες παραμέτρους ενός μοντέλου. Στην περίπτωση αυτή οι παράμετροι του μοντέλου αντικαθιστούνται από τους εκτιμητές μέγιστης πιθανοφάνειας και έχει δειχθεί ότι όσο μεγαλύτερη είναι η τιμή που μεγιστοποιεί την συνάρτηση πιθανοφάνειας τόσο καλύτερη είναι και η συμφωνία ανάμεσα στο μοντέλο και στα δεδομένα.

Στην πράξη για να συγκρίνουμε εναλλακτικά μοντέλα δεν χρησιμοποιούμε απευθείας την συνάρτηση πιθανοφάνειας αλλά την ποσότητα -2log(L), η οποία είναι πάντοτε θετική, αφού ο όρος L δηλώνει το γινόμενο μιας σειράς υπό συνθήκη πιθανοτήτων (τιμή <1). Ακόμη όσο μικρότερη είναι η τιμή της ποσότητας -2log(L) τόσο καλύτερο είναι το συγκεκριμένο μοντέλο σε σχέση με τα υπόλοιπα μοντέλα που εξετάσουμε.

Όμως η ποσότητα αυτή δεν μπορεί από μόνη της να κρίνει αν ένα μοντέλο είναι επαρκές εφόσον εξαρτάται από τον αριθμό των παρατηρήσεων. Επομένως θα πρέπει να χρησιμοποιούμε την συγκεκριμένη ποσότητα μόνο όταν κάνουμε συγκρίσεις ανάμεσα σε μοντέλα που προσαρμόζονται πάνω στα ίδια δεδομένα.

Θα προσπαθήσουμε να εξηγήσουμε πως λειτουργούν όλα τα παραπάνω μέσω ενός παραδείγματος. Έστω ότι έχουμε δύο μοντέλα (M1 και M2) να συγκρίνουμε, και έστω ότι το μοντέλο M1 αποτελείται από p μεταβλητές ενώ το M2 από p + q μεταβλητές, με αντίστοιχες συναρτήσεις πιθανοφάνειας τις L_1 και L_2. Αφού υπολογίσουμε τις ποσότητες -2log(L1) και -2log(L2) , στην συνέχεια θα υπολογίσουμε την διαφορά τους, $(-2\log(\hat{L}1)-(-2\log(\hat{L}2)))$, και αν η διαφορά αυτή είναι μεγάλη τότε θα καταλήξουμε στο συμπέρασμα ότι οι q επιπλέον μεταβλητές από τις οποίες αποτελείται το μοντέλο M2 βελτιώνουν την προσαρμογή του μοντέλου. Βέβαια το πόσο μεταβάλλεται η ποσότητα -2log(L) όταν προστίθενται

άλλοι όροι στο μοντέλο εξαρτάται κάθε φορά από τους όρους που ήδη περιλαμβάνονται στο μοντέλο. Το στατιστικό του ελέγχου μπορεί να γραφτεί και στην μορφή:

$$-2\log(\hat{L}1/\hat{L}2)$$

η οποία αντιστοιχεί στον λογάριθμο του λόγου πιθανοφάνειας για τον έλεγχο της υπόθεσης ότι οι q επιπλέον παράμετροι του μοντέλου M2 είναι όλα ίσες με μηδέν.

Το στατιστικό αυτό σύμφωνα με την θεωρία του λόγου πιθανοφάνειας ακολουθεί ασυμπτωτικά την X^2 κατανομή με β.ε ίσους με την διαφορά του αριθμού των β ανεξάρτητων παραμέτρων που προσαρμόζονται στα δύο μοντέλα, δηλ.

p+q- p=q. Αν η διαφορά των δύο στατιστικών δεν είναι μεγάλη τότε πιθανότατα και τα δύο μοντέλα δεν διαφέρουν σημαντικά για την ανάλυση των δεδομένων , σε αυτή την περίπτωση θα επιλέξουμε το καλύτερο μοντέλο λαμβάνοντας υπόψη και άλλα χαρακτηριστικά όπως το πλήθος των μεταβλητών, συνήθως προτιμούμε μοντέλο με λιγότερες ερμηνευτικές μεταβλητές καθώς είναι πιο κατανοητά και ερμηνεύονται εύκολα.

Βέβαια από την διαδικασία επιλογής μοντέλου δεν περιμένουμε να καταλήξουμε μόνο σε ένα και μοναδικό, καλύτερο, μοντέλο για τα δεδομένα μας. Δοκιμάζουμε όλους τους δυνατούς συνδυασμούς των επεξηγηματικών μεταβλητών, συμπεριλαμβάνοντας και τους όρους αλληλεπίδρασης και ύστερα επιλέγουμε ένα με βάση κάποια λογικά κριτήρια και αν αυτό είναι επαρκές. Τις συνθήκες που προσδιορίζουν αν ένα μοντέλο είναι επαρκές ή όχι θα τις περιγράψουμε αναλυτικά στο επόμενο κεφάλαιο.

Να σημειώσουμε ότι στην περίπτωση που επιλέγουμε να συμπεριλάβουμε όρους αλληλεπίδρασης στο μοντέλο θα πρέπει επίσης να συμπεριλάβουμε και τις κύριες επιδράσεις των μεταβλητών αυτών, ακόμα και στην περίπτωση που δεν είναι στατιστικά σημαντικές. Σε αντίθετη περίπτωση το μοντέλο μας δεν θα είναι ιεραρχικό και η ανάλυση και ερμηνεία ενός μη ιεραρχικού μοντέλου δεν είναι καθόλου εύκολη υπόθεση.

(Collet 2000 και Αντζουλάκος 2009)

4.10.1 Διαδικασίες επιλογής μεταβλητών

Ορισμένες φορές, όπως αναφέραμε και παραπάνω τα δεδομένα μας μπορεί να αποτελούνται από μεγάλο αριθμό ερμηνευτικών μεταβλητών, όμως μπορεί για το σκοπό της ανάλυσης μόνο κάποιες από αυτές να είναι απαραίτητες. Στόχος μας είναι να εντοπίσουμε αυτές τις μεταβλητές από τις οποίες εξαρτάται η συνάρτηση κινδύνου. Έχουν αναπτυχθεί πολλές και διάφορες τεχνικές για την επιλογή σημαντικών μεταβλητών τόσο μέσω κριτηρίων όσο και μέσω στατιστικών πακέτων.

Στην συνέχεια θα προσπαθήσουμε να περιγράψουμε εν συντομία κάποιες από τις πιο γνωστές τεχνικές.

Στην περίπτωση που έχουμε μικρό αριθμό ερμηνευτικών μεταβλητών (μαζί με όρους αλληλεπίδρασης και μη γραμμικούς όρους) η επιλογή μοντέλου μπορεί να γίνει με χρήση της ποσότητας $-2\log(\hat{L})$. Συγκρίνοντας κάθε φορά τις τιμές που παίρνει ο λογάριθμος του λόγου πιθανοφάνειας όταν προσθέτουμε ή αφαιρούμε όρους από ένα μοντέλο, τηρώντας πάντα τις αρχές που πρέπει να ικανοποιεί ένα ιεραρχικό μοντέλο. Ως καλύτερο επιλέγεται το μοντέλο που εμφανίζει την μικρότερη τιμή για το παραπάνω στατιστικό.

Ένας άλλος τρόπος επιλογής μεταβλητών και κατά συνέπεια κατάλληλου μοντέλου για τα δεδομένα μας είναι μέσω του *κριτήριου πληροφορίας του Akaike (Akaike's information criterion),* η ποσότητα αυτή υπολογίζεται από την σχέση:

$$AIC = -2\log\hat{L} + aq$$

όπου με q δηλώνουμε τον αριθμό των άγνωστων β παραμέτρων στο μοντέλο και με a συμβολίσουμε μια προκαθορισμένη σταθερά την οποία συνήθως θέτουμε ίση με 3. Έχει αποδειχθεί ότι όσο μικρότερη είναι η τιμή του κριτηρίου τόσο καλύτερο είναι το μοντέλο. Επίσης η τιμή του κριτηρίου αυξάνεται όταν μη – απαραίτητοι όροι προστίθενται στο μοντέλο. Επομένως είτε επιλέξουμε ένα μοντέλο με βάση την τιμή του Akaike είτε με βάση την ποσότητα $-2\log(\hat{L})$ θα επιλέξουμε ως καλύτερο εκείνο που έχει την μικρότερη τιμή για τις δύο ποσότητες. Βέβαια μπορεί να καταλήξουμε σε περισσότερα από ένα και μοναδικό καλύτερο μοντέλο για τα δεδομένα μας, στην περίπτωση αυτή θα επιλέξουμε το καλύτερο εξετάζοντας αν ικανοποιούνται κάποιες επιπλέον προϋποθέσεις.

Η μέθοδος αυτή όπως προαναφέραμε μπορεί να χρησιμοποιηθεί όταν το πλήθος των μεταβλητών που εξετάζουμε είναι μικρό, στην περίπτωση όμως που το πλήθος των p επεξηγηματικών μεταβλητών ισούται με 10, τότε όπως είναι γνωστό υπάρχουν 2^{10} δυνατοί συνδυασμοί, πάρα πολλοί, και όπως είναι φυσικό δεν μπορούμε να τους ελέγξουμε όλους ένα προς ένα και να υπολογίσουμε τις ποσότητες που αναφέραμε παραπάνω. Έτσι καταφεύγουμε στην χρήση *έτοιμων ρουτινών* που επιλέγουν αυτόματα το κατάλληλο πλήθος ερμηνευτικών μεταβλητών για τα δεδομένα μας και είναι διαθέσιμα από τα περισσότερα στατιστικά πακέτα.

Οι τρείς πιο γνωστές και αποδεκτές ρουτίνες επιλογής μεταβλητών είναι:

- *Forward Selection*

Σύμφωνα με την συγκεκριμένη ρουτίνα οι μεταβλητές εισάγονται μία - μία σε ένα μοντέλο που αποτελείται αρχικά μόνο από την αναφορική συνάρτηση κινδύνου. Η μεταβλητή που ικανοποιεί το κριτήριο που έχουμε θέσει συμπεριλαμβάνεται στο μοντέλο, η ρουτίνα κατά το ίδιο τρόπο ελέγχει και τις υπόλοιπες, σταματά όταν δεν μπορούμε να προσθέσουμε κάποια άλλη μεταβλητή στο μοντέλο.

- *Backward Elimination*

Η μέθοδος αυτή δουλεύει αντίθετα της Forward, ξεκινάει από το πλήρες μοντέλο και αφαιρεί μια - μια τις μεταβλητές που σύμφωνα με το κριτήριο που έχουμε θέση δεν συμβάλλουν στην ερμηνεία των δεδομένων που μελετάμε. Επαναλαμβάνει την διαδικασία αυτή για όλες τις μεταβλητές και σταματά όταν δεν μπορούμε να διώξουμε κάποια άλλη μεταβλητή.

- *Stepwise Procedure*

Η μέθοδος αυτή αποτελεί ένα συνδυασμό των δύο παραπάνω μεθόδων και αναπτύχτηκε προκειμένου να καλύψει ένα σημαντικό μειονέκτημα τους. Οι Forward και Backward διαδικασίες δεν ελέγχουν σε κάθε τους βήμα αν μπορεί να συμπεριληφθεί στο μοντέλο κάποια μεταβλητή που αρχικά έχει αφαιρεθεί. Έτσι το πρώτο βήμα της Stepwise procedure δουλεύει ακριβώς όπως η Forward selection, ελέγχει ποια μεταβλητή μεγιστοποιεί το κριτήριο που έχουμε θέσει, η συγκεκριμένη μεταβλητή εισάγεται πρώτη στο μοντέλο. Επαναλαμβάνουμε το ίδιο βήμα όσες φορές χρειαστεί με την μόνη διαφορά ότι τώρα σε κάθε βήμα ελέγχουμε αν η εισαγωγή μιας νέας μεταβλητής στο μοντέλο καθιστά μη απαραίτητη κάποια μεταβλητή που έχει εισαχθεί νωρίτερα. Η διαδικασία αυτή σταματά όταν η τιμή εισαγωγής μιας μεταβλητής στο μοντέλο είναι μεγαλύτερη ή ίση της τιμής απομάκρυνσής της από το μοντέλο.

Μπορεί οι ρουτίνες αυτές να μας γλυτώνουν πολύ χρόνο, μη αναγκαίους υπολογισμούς και προβληματισμούς σχετικά με το ποιό μοντέλο να επιλέξουμε κάθε φορά, όμως συγχρόνως τα προτερήματά τους αποτελούν και τα μειονεκτήματά τους. Καθώς οι ρουτίνες αυτές καταλήγουν σε ένα μόνο καλύτερο μοντέλο και εφόσον οδηγούνται στο μοντέλο αυτό μέσω διαφορετικών διαδικασιών μπορεί να είναι διαφορετικό και για τις τρείς μεθόδους. Επιπλέον οι δύο πρώτες διαδικασίες μπορεί να παραλείψουν κάποια σημαντική μεταβλητή εφόσον δεν έχουν την δυνατότητα να γυρίσουν πίσω και να δούνε αν κάποια μεταβλητή που έχει ήδη διωχθεί μπορεί να συμπεριλήφθη τώρα στο μοντέλο δοθέντος κάποιας άλλης μεταβλητής.

Για τους παραπάνω λόγους συνήθως προτιμούμε να χρησιμοποιήσουμε την ακόλουθη διαδικασία προκειμένου να επιλέξουμε τις κατάλληλες μεταβλητές για την ερμηνεία των δεδομένων μας.

1^{o} Βήμα: Προσαρμόζουμε μια – μια τις μεταβλητές στο μοντέλο (προσαρμόζουμε ένα μονομεταβλητό μοντέλο για κάθε μεταβλητή), οι τιμές που παίρνουμε για το στατιστικό $-2\log(\hat{L})$ συγκρίνονται με αυτή του μηδενικού μοντέλου (δηλ. του μοντέλου που αποτελείται μόνο από την αναφορική συνάρτηση κινδύνου) προκειμένου να προσδιορίσουμε τις μεταβλητές που από μόνες τους μειώνουν σημαντικά την τιμή του στατιστικού. Σε αυτό το βήμα χρησιμοποιούμε ένα προκαθορισμένο επίπεδο αναφοράς 20% προκειμένου να μην παραλείψουμε κάποια σημαντική μεταβλητή από την αρχή.

2^{o} Βήμα: Στην συνέχεια οι μεταβλητές που αποδείχτηκαν σημαντικές στο 1^{o} βήμα προσαρμόζονται μαζί σε ένα μοντέλο. Παρουσία κάποιων μεταβλητών μπορεί κάποιες άλλες να πάψουν να είναι σημαντικές, κατά συνέπεια μεταβλητές οι οποίες δεν αυξάνουν σημαντικά την τιμή της $-2\log(\hat{L})$ όταν της απομακρύνουμε από το μοντέλο μπορούμε να τις αφαιρέσουμε. Υπολογίσουμε την διαφορά στην τιμή των στατιστικών για κάθε μεταβλητή που παραλείπεται από το μοντέλο. Βέβαια θα πρέπει κάθε φορά που διώχνουμε μια μεταβλητή να ελέγξουμε την επίδραση της σε αυτές που παραμένουν. Στην ουσία χρησιμοποιούμε backward elimination για να απομακρύνουμε τις μη σημαντικές μεταβλητές.

3^{o} Βήμα: Στο βήμα αυτό εξετάζουμε την περίπτωση οι μεταβλητές που δεν ήταν σημαντικές από μόνες τους και κατά συνέπεια δεν ληφθήκαν υπόψη στο 2^{o} βήμα να έχουν τώρα κάποια σημαντική επίδραση παρουσία των άλλων μεταβλητών. Έτσι στο μοντέλο που έχει προκύψει από το 2^{o} βήμα προστίθενται μια – μια οι μεταβλητές

αυτές και εκείνες που μειώνουν την ποσότητα του στατιστικού συμπεριλαμβάνονται στο μοντέλο. Εφαρμόζουμε δηλαδή forward selection για να εξετάσουμε αν θα πρέπει οι μεταβλητές που αφαιρέσαμε στο $1^{ο}$ βήμα να συμπεριληφθούν στο μοντέλο Μ2.

$4^{ο}$ Βήμα: Κάνουμε ένα τελευταίο έλεγχο ώστε να διαπιστώσουμε ότι καμία μεταβλητή που αυξάνει την τιμή της ποσότητας $-2\log(\hat{L})$ δεν περιλαμβάνεται στο μοντέλο, ενώ από την άλλη δεν θα πρέπει να έχουμε απομακρύνει κάποια μεταβλητή που μειώνει σημαντικά την ποσότητα $-2\log(\hat{L})$. Εφαρμόζουμε stepwise regression. Στο βήμα αυτό εξετάζουμε αν μπορούν να συμπεριληφθούν στο μοντέλο αλληλεπιδράσεις που ορίζονται μεταξύ κυρίων επιδράσεων που υπάρχουν ήδη στο μοντέλο.

Να σημειώσουμε ότι σε όλη την παραπάνω διαδικασία συνίσταται να χρησιμοποιούμε κάποιο ε.σ. που να μην είναι ούτε πολύ μικρό ούτε πολύ μεγάλο, συνήθως επιλέγουμε α=10%.

(Collett 2000, Αντζουλάκος 2009)

Ακόμη, μια άλλη διαδικασία επιλογής μεταβλητών αφορά την περίπτωση που έχουμε προαποφασίσει ποιες μεταβλητές θα πρέπει οπωσδήποτε να εισάγουμε στο μοντέλο. Εφαρμόζουμε τα βήματα 1 – 4 για να εντοπίσουμε το κατάλληλο μοντέλο για τα δεδομένα μας και στην συνέχεια αφού εισάγουμε τις προεπιλεγμένες μεταβλητές εξετάζουμε την εισαγωγή αλληλεπιδράσεων στο μοντέλο, σύμφωνα βέβαια με την ιεραρχική αρχή.

(Αντζουλάκος 2009).

Παρατηρήσεις

Σχετικά με τα όσα αναφέραμε στην παραπάνω ενότητα ισχύουν κάποιες γενικές παρατηρήσεις σύμφωνα με τις σημειώσεις του Αντζουλάκου (2009):

- Όταν ένας παράγοντας Α με a επίπεδα αλληλεπιδρά με ένα παράγοντα Ν με n επίπεδα, τότε στο μοντέλο εισάγουμε συνολικά (a-1)(n-1) αλληλεπιδράσεις και όχι a*n, αυτό διότι για κάθε μεταβλητή θεωρούμε και ένα επίπεδο αναφοράς (συνήθως το πρώτο).
- Τόσο το μέγεθος εκτίμησης των συντελεστών παλινδρόμησης όσο και το μέγεθος του τυπικού τους σφάλματος καθορίζουν αν ένα μοντέλο είναι επαρκές ή πλεονάζων ως προς τις μεταβλητές που περιέχει. Συγκεκριμένα, ένας συντελεστής παλινδρόμησης με ιδιαίτερα μεγάλη εκτίμηση κατά απόλυτη τιμή και μεγάλο τυπικό σφάλμα αποτελεί ένδειξη ότι στο μοντέλο έχει γίνει υπερβολική προσαρμογή (overfitting) και κατά συνέπεια η εν λόγω μεταβλητή δεν θα πρέπει να εισαχθεί στο μοντέλο.
- Υπάρχει περίπτωση σε ένα μοντέλο να συναντήσουμε ισχυρά συσχετισμένες μεταβλητές. Στην περίπτωση αυτή θα πρέπει να είμαστε πολύ προσεκτικοί, καθώς μεταβλητές που φαίνονται μη σημαντικές όταν ελέγχονται χωριστά εμφανίζονται σημαντικές όταν ελέγχονται ταυτόχρονα στο μοντέλο.

ΚΕΦΑΛΑΙΟ 5ο

Αξιολόγηση του μοντέλου παλινδρόμησης του Cox

5.1 Εισαγωγή

Στο κεφάλαιο αυτό περιγράφονται τρόποι ελέγχου της υπόθεσης αναλογικού κινδύνου τόσο μέσω γραφικών μεθόδων όσο και μέσο στατιστικών ελέγχων. Περιγράφονται κάποια τεστ καλής προσαρμογής, από τα οποία τα περισσότερα βασίζονται σε γραφικές μεθόδους και ανάλυση υπολοίπων (residuals). Τα είδη υπολοίπων που αναφέρουμε παρουσιάζουν κάποιες ομοιότητες με τα υπόλοιπα που χρησιμοποιούνται στην ανάλυση διαφορετικών δεδομένων όπως στην γραμμική ή λογιστική παλινδρόμηση όμως έχουν διαφορετικές ιδιότητες και ερμηνείες. Στο τέλος γίνεται μια αναφορά στις γενικεύσεις του μοντέλου αναλογικού κινδύνου

5.2 Τα είδη υπολοίπων για το μοντέλο παλινδρόμησης του Cox

Τα υπόλοιπα αποτελούν ένα χρήσιμο εργαλείο προκειμένου να ερευνήσουμε την έλλειψη προσαρμογής που μπορεί να παρουσιάζει κάποιο μοντέλο για ένα δοσμένο υποκείμενο. Τα υπόλοιπα που χρησιμοποιούνται στην γραμμική παλινδρόμηση ορίζονται και ερμηνεύονται εύκολα μέσω της διαφοράς της παρατηρούμενης τιμής μιας μεταβλητής μείον την αναμενόμενη τιμή της, κάτι τέτοιο όμως δεν ισχύει για το μοντέλο παλινδρόμησης του Cox. Οι σχέσεις ορισμού και οι ερμηνείες τους είναι λίγο πιο πολύπλοκα και λιγότερο κατανοητά (διαισθητικά).

Για την αξιολόγηση ενός μοντέλου παλινδρόμησης του Cox χρησιμοποιούνται ευρέως τέσσερα είδη υπολοίπων: τα Martingale, τα Deviance, τα Schoenfeld και τα Score residuals. Τα υπόλοιπα αυτά συνοδεύονται κι από δύο είδη υπολοίπων τα bfdelta και τα scaled Schoenfeld residuals, που προέρχονται από τα παραπάνω είδη. Όμως τα πρώτα είδη υπολοίπων που χρησιμοποιήθηκαν για το μοντέλο αναλογικού κινδύνου του Cox είναι τα Cox – Snell υπόλοιπα κι τα Modified Cox – Snell υπόλοιπα, τα οποία δημιουργήθηκαν αρχικά από τον ίδιο τον Cox και στην πορεία εξελίχτηκαν.

(Για περισσότερα πάνω στα είδη των υπολοίπων που περιγράφονται παρακάτω βλέπε Collett 2000, Winnett and Sasieni 2001, K. Sainani PH.D , Gillespie 2006, Αντζουλάκος 2009)

5.2.1 Cox – Snell residuals

Τα Cox – Snell υπόλοιπα είναι ιδιαίτερα γνωστά για την χρήση τους στην ανάλυση δεδομένων που αφορούν χρόνους επιβίωσης, πήραν το όνομα τους λόγω των δημιουργών τους, Cox & Snell (1982), και υπολογίζονται από την εξίσωση:

$$r_{Ci} = \exp(\hat{\beta}' z_i)\hat{H}_0(t_i)$$

όπου i=1,2,…n και $\hat{H}_0(t_i)$ η εκτίμηση της αθροιστικής συνάρτησης κινδύνου στο χρόνο t_i. Όπως είναι γνωστό τα υπόλοιπα αποτελούν ένα μαθηματικό αποτέλεσμα για την απόκτηση της οποίας στηριζόμαστε στην συνάρτηση κατανομής μιας τυχαίας μεταβλητής. Επομένως αν θεωρήσουμε ότι T είναι μια τυχαία μεταβλητή που περιγράφει το χρόνο επιβίωσης ενός ατόμου, με αντίστοιχη συνάρτηση επιβίωσης S(t), τότε η F(T) ~ U(0,1) άρα η ποσότητα $Y = -\log(S(t))$ θα ακολουθεί μια εκθετική κατανομή με μοναδιαία μέση τιμή.

Τα Cox – Snell υπόλοιπα αποτελούν ένα μέτρο αξιολόγησης του μοντέλου που έχουμε επιλέξει να προσαρμόσουμε, με την έννοια ότι αν το προσαρμοσμένο μοντέλο είναι το σωστό τότε θα πρέπει το δείγμα των υπολοίπων του να προέρχεται από μια εκθετική κατανομή με παράμετρο 1. Οι ιδιότητες που έχουν τα υπόλοιπα αυτά διαφέρουν σημαντικά από αυτές τις γραμμικής παλινδρόμησης, πιο συγκεκριμένα τα υπόλοιπα αυτά δεν κατανέμονται συμμετρικά γύρω από το μηδέν και έτσι δεν μπορούν να είναι αρνητικά.

5.2.2 Modified Cox – Snell residuals

Τα Modified Cox – Snell υπόλοιπα αναπτύχτηκαν προκειμένου να βελτιώσουν τα Cox – Snell υπόλοιπα κατά τέτοιο τρόπο ώστε να λαμβάνουν υπόψη την λογοκρισία. Έστω ότι η $i^{η}$ παρατήρηση δηλώνει το λογοκριμένο χρόνο επιβίωσης ενός ατόμου, t_i*, ισχύει ότι $t_i > t_i$* οπότε τα Cox – Snell υπόλοιπα για το άτομο αυτό θα υπολογίζονται από την σχέση:

$$r_{Ci} = \hat{H}_i(t_i^*) = -\log \hat{S}_i(t_i^*)$$

όπου με $\hat{H}_i(t_i^*)$ και $\hat{S}_i(t_i^*)$ δηλώνουμε αντίστοιχα την εκτίμηση για την αθροιστική συνάρτηση κινδύνου και την συνάρτηση επιβίωσης στο χρόνο λογοκρισίας t_i^*.

Έτσι σύμφωνα με το χαρακτηριστικό που εμφανίζουν τα Cox – Snell υπόλοιπα αν το προσαρμοσμένο μοντέλο είναι το κατάλληλο τότε τα υπόλοιπα θα ακολουθούν μια μοναδιαία εκθετική κατανομή. Η αθροιστική συνάρτηση κινδύνου της συγκεκριμένης κατανομής θα αυξάνει γραμμικά με το χρόνο επιβίωσης των ατόμων στο δείγμα.

Η τροποποίηση των Cox – Snell υπολοίπων προκύπτει με την προσθήκη μιας σταθεράς θετικής ποσότητας, έστω Δ, η οποία στην συνέχεια προσδιορίζεται με βάση την ιδιότητα έλλειψης μνήμης της εκθετικής κατανομής, έτσι Δ=1 και τα Modified Cox – Snell υπόλοιπα θα ισούνται με:

$$r_{Ci}' = \begin{cases} r_{Ci}, & \text{για μη – λογοκριμένες παρατηρήσεις} \\ r_{Ci}+1, & \text{για λογοκριμένες παρατηρήσεις} \end{cases}$$

Ένας άλλος τρόπος ορισμού των Modified Cox – Snell υπολοίπων προκύπτει αν θεωρήσουμε μια δείκτρια μεταβλητή που παίρνει τις τιμές 0 και 1, όπου η τιμή 0 δηλώνει λογοκριμένη παρατήρηση και αντίστοιχα η τιμή 1 μη λογοκριμένη παρατήρηση.

Επομένως,

$$r_{Ci}' = 1 - \delta_i + r_{Ci}$$

5.2.3 Martingale residuals

Προκύπτουν από τα Modified Cox – Snell υπόλοιπα αν αφαιρέσουμε την μονάδα, έτσι για μη – λογοκριμένες παρατηρήσεις τα υπόλοιπα αυτά έχουν μηδενική μέση τιμή. Υπολογίζονται από την σχέση:

$$r_{Mi} = \delta_i - r_{Ci}$$

Τα martingale υπόλοιπα παίρνουν τιμές στο διάστημα $(-\infty, 1)$, για λογοκριμένες παρατηρήσεις είναι αρνητικά, αθροίζουν στο μηδέν και για μεγάλα μεγέθους δείγματα τα υπόλοιπα αυτά είναι ασυσχέτιστα το ένα με το άλλο.

Έχουν παρόμοιες ιδιότητες με τα υπόλοιπα της γραμμικής παλινδρόμησης αφού τα Martingale υπόλοιπα εκφράζουν στην ουσία την διαφορά ανάμεσα στο παρατηρούμενο αριθμό θανάτων για το i^o άτομο στο διάστημα (0, t) και τον αναμενόμενο αριθμό θανάτων κάτω από το προσαρμοσμένο μοντέλο.

Παρά τις ομοιότητες που παρουσιάζουν τα martingale υπόλοιπα με τα υπόλοιπα της γραμμικής παλινδρόμησης αυτά δεν μπορούν να χρησιμοποιηθούν προκειμένου να αξιολογήσουν την ολική προσαρμογή του μοντέλου. Για τα γραμμικά μοντέλα παλινδρόμησης ισχύει ότι το μοντέλο που μας δίνει το μικρότερο άθροισμα τετραγώνων των υπολοίπων προτιμάται έναντι των άλλων υποψηφίων μοντέλων, καθώς το στατιστικό αυτό μας παρέχει ένα συνολικό μέτρο καλής προσαρμογής. Να σημειώσουμε ότι τα martingale υπόλοιπα δεν κατανέμονται συμμετρικά ακόμα και αν το προσαρμοσμένο μοντέλο είναι το σωστό, αυτός είναι και ο λόγος που τα μετατρέπουμε σε deviance υπόλοιπα.

(Rodriguez 2001)

5.2.4 Deviance residuals

Τα deviance υπόλοιπα επιδιώκουν να βελτιώσουν τα martingale υπόλοιπα έτσι ώστε να λαμβάνουν υπόψη την ιδιότητα που έχουν τα υπόλοιπα που συναντάμε σε άλλους τομείς της ανάλυσης δεδομένων, αυτή της συμμετρικής κατανομής γύρω από την τιμή μηδέν. Η έλλειψη της συγκεκριμένης ιδιότητας δυσκολεύει την σωστή ερμηνεία των υπολοίπων. Για αυτό τα deviance υπόλοιπα αποτελούν μια κανονική μετατροπή των martingale υπολοίπων, καθώς κατανέμονται περισσότερο συμμετρικά γύρω από το μηδέν.

Υπολογίζονται από την εξίσωση:

$$r_{Di} = \text{sgn}(r_{Mi})[-2[r_{Mi} + \delta_i \cdot \log(\delta_i - r_{Mi})]]^{1/2}$$

όπου η sgn( ) δηλώνει ότι τα deviance υπόλοιπα θα έχουν το ίδιο πρόσημο με τα martingale υπόλοιπα.

Στα deviance υπόλοιπα βασίζεται το στατιστικό Deviance με το οποίο ελέγχουμε πόσο ικανοποιητικό είναι το προσαρμοσμένο μοντέλο έναντι του πλήρους μοντέλου. Το στατιστικό αυτό ορίζεται ως εξής:

$$D = -2[\log \hat{L}_i - \log \hat{L}_f]$$

όπου η $\log \hat{L}_i$ μας δίνει την εκτίμηση της μερική πιθανοφάνειας για το προσαρμοσμένο μοντέλο, αντίστοιχα η $\log \hat{L}_f$ μας δίνει την εκτίμηση της μερικής πιθανοφάνειας για το πλήρες μοντέλο. Όσο μικρότερη είναι η τιμή του στατιστικού τόσο καλύτερο είναι το προσαρμοσμένο έναντι του πλήρους μοντέλου.

Τα deviance υπόλοιπα χρησιμοποιούνται επίσης και σαν ένα μέτρο προκειμένου να προσδιορίσουμε ποια παρατήρηση επηρεάζει την προσαρμογή του μοντέλου, στην ουσία χρησιμοποιούμε το άθροισμα των τετραγώνων των deviance υπολοίπων ($\sum r_{Di}^{\,2}$) . Το μέτρο αυτό είναι αντίστοιχο με το τετραγωνικό άθροισμα των υπολοίπων που χρησιμοποιούνται στην ανάλυση κανονικών δεδομένων.

Παρατηρήσεις με μεγάλες τιμές για τα deviance υπόλοιπα είναι υπεύθυνες για την μη καλή προσαρμογή του μοντέλου.

5.2.5 Schoenfeld residuals

Όλα τα είδη των υπολοίπων που περιγράψαμε παραπάνω έχουν δύο μειονεκτήματα:

$1^{ον}$ εξαρτώνται ισχυρά από το παρατηρούμενο χρόνο ζωής και

$2^{ον}$ απαιτούν μια εκτίμηση για την αθροιστική συνάρτηση κινδύνου.

Τα μειονεκτήματα αυτά αντιμετωπίζονται από τα Schoenfeld υπόλοιπα (1982), τα οποία αρχικά ονομάστηκαν Partial residuals. Μια άλλη σημαντική διαφορά που έχουν τα υπόλοιπα αυτά σε σχέση με τα προηγούμενα είδη είναι ότι δίνουν τιμή υπολοίπου σε κάθε άτομο και για κάθε μεταβλητή που περιλαμβάνεται στο μοντέλο αναλογικού κινδύνου.

Υπολογίζονται από την σχέση:

$$r_{pji} = \delta_i \{x_{ji} - \hat{a}_{ji}\}$$

όπου η x_j δηλώνει την j επεξηγηματική μεταβλητή, ενώ η $\hat{a}_{ji} = \dfrac{\sum_{l \in R_{(ti)}} x_{jl} \exp(\hat{\beta}' x_l)}{\sum_{l \in R(t_i)} \exp(\hat{\beta}' x_l)}$

και η $R(t_i)$ μας δίνει το σύνολο των ατόμων που βρίσκονται σε κίνδυνο την χρονική στιγμή t_i. Για τα Schoenfeld υπόλοιπα έχουμε μη μηδενικές τιμές μόνο για τις μη λογοκριμένες παρατηρήσεις, έτσι αν για ένα δείγμα χρόνων επιβίωσης η τελευταία παρατήρηση είναι πλήρης χρόνος τότε η ποσότητα $\hat{a}_{ji}$ σύμφωνα με την σχέση υπολογισμού της θα ισούται με x_{ji} και κατά συνέπεια το Schoenfeld υπόλοιπο για το άτομο αυτό θα ισούται με μηδέν, $r_{pji}=0$. Προκειμένου να διακρίνουμε τις παρατηρήσεις που είναι πραγματικά μηδέν από αυτές που δημιουργούνται από τις λογοκριμένες παρατηρήσεις, δηλώνουμε τις τελευταίες ως ελλιπής παρατηρήσεις.

Μια εκτίμηση για τα Schoenfeld υπόλοιπα, που οφείλεται στους Hosmer & Lemeshow, δίνεται από την σχέση:

$$\hat{r}_{Sik} = x_{ik} - \hat{\bar{x}}_{wik}$$

όπου η x_{ik} μας δίνει την τιμή για την $k^{η}$ μεταβλητή για το i άτομο, ενώ η $\hat{\bar{x}}_{wik}$ μας δίνει τον σταθμισμένο μέσο για τις τιμές της μεταβλητής για τα άτομα που βρίσκονται σε κίνδυνο στο δοσμένο πλήρη χρόνο. Μια θετική τιμή του $\hat{r}_{Sik}$ μας δείχνει ότι η x τιμή είναι υψηλότερη από τον αναμενόμενο χρόνο θανάτου.

5.2.6 Scaled Schoenfeld residuals

Οι Grambsch & Therneau (1994) έδειξαν ότι τα Scaled Schoenfeld υπόλοιπα παίζουν πολύ σημαντικό ρόλο στην διάγνωση ενός μοντέλου παλινδρόμησης του Cox και συγκεκριμένα στην αξιολόγηση της υπόθεσης αναλογικού κινδύνου. Τα Scaled Schoenfeld υπόλοιπα προκύπτουν από τα Schoenfeld υπόλοιπα προσαρμοσμένα για τον αντίστροφο του πίνακα συνδιακυμάνσεων των Schoenfeld υπολοίπων.

Υπολογίζονται από την ακόλουθη σχέση:

$$r^*_{Pi} = mVar(\hat{\beta})r_{Pi}$$

όπου με m συμβολίσουμε τον αριθμό των αποτυχιών (θανάτων) για τα n άτομα, με $Var(\hat{\beta})$ δηλώνουμε τον πίνακα διακυμάνσεων - συνδιακυμάνσεων για τις εκτιμώμενες παραμέτρους του μοντέλου παλινδρόμησης του Cox και με $r_{Pi} = (r_{P1i}, r_{P2i}, ..., r_{Ppi})$ δηλώνουμε το διάνυσμα των Schoenfeld υπολοίπων για το i άτομο.

(Για περισσότερες λεπτομέρειες σχετικά με τα Scaled Schoenfeld υπόλοιπα βλέπε διπλωματική: Ali Mohamed Ali, Analysis of Vaccine Efficacy under Time Dependent, 2008)

5.2.7 Score residuals

Υπολογίζονται από τη $1^{η}$ παράγωγο του λογαρίθμου της εξίσωσης μερικής πιθανοφάνειας για την β_j παράμετρο, είναι τα σκορ απόδοσης για την β_j και για αυτό τα υπόλοιπα αυτά είναι γνωστά ως score residuals, δηλ:

$$\frac{\theta \log L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \left\{ \delta_i (x_{ji} - a_{ji}) + \exp(\beta' x_i) \sum_{t_r \le t_i} \frac{(a_{ji} - x_{ji})\delta_r}{\sum_{l \in R(t_r)} \exp(\beta' x_l)} \right\}$$

Στην παραπάνω εξίσωση η συνεισφορά της i παρατήρησης στην παράγωγο εξαρτάται μόνο από την πληροφορία που έχουμε μέχρι την χρονική στιγμή t_i, π.χ. αν

μια μελέτη είχε ολοκληρωθεί την χρονική στιγμή t_i τότε η i συνιστώσα της παραγώγου δεν θα είχε επηρεαστεί.

Το i score υπόλοιπο για την j επεξηγηματική μεταβλητή δίνεται από την σχέση:

$$r_{sji} = \delta_i (x_{ji} - \hat{a}_{ji}) + \exp(\hat{\beta}' x_i) \sum_{t_r \le t_i} \frac{(\hat{a}_{jr} - x_{ji})\delta_r}{\sum_{l \in R(t_i)} \exp(\hat{\beta}' x_l)}$$

Μια εναλλακτική σχέση υπολογισμού των score υπολοίπων είναι η εξής:

$$r_{sji} = r_{pji} + \exp(\hat{\beta}' x_i) \sum_{t_r \le t_i} \frac{(\hat{a}_{ji} - x_{ji})\delta_r}{\sum_{l \in R(t_i)} \exp(\hat{\beta}' x_l)}$$

Από όπου προκύπτει ότι τα score υπόλοιπα αποτελούν μια μοντελοποίηση των Schoenfeld υπολοίπων.

5.3 Τα διαγράμματα των υπολοίπων

Τα διαγράμματα των υπολοίπων χρησιμοποιούνται προκειμένου να ελέγξουμε αν το προσαρμοσμένο μοντέλο είναι επαρκές.

Το κάθε είδος υπολοίπων όταν απεικονίζεται έναντι των παρατηρήσεων ή συνεχών μεταβλητών ή του χρόνου μας δίνει μια πρώτη εικόνα για το αν το προσαρμοσμένο μοντέλο είναι το κατάλληλο, αν υπάρχουν έκτροπες παρατηρήσεις, αν η συναρτησιακή μορφή των συνεχών μεταβλητών είναι η σωστή, αν ικανοποιείται η υπόθεση αναλογικού κινδύνου κτλ.

Τα διαγράμματα που χρησιμοποιούνται συνήθως στην πράξη είναι:

- **Για τα Cox – Snell υπόλοιπα**

Όπως αναφέραμε παραπάνω τα Cox – Snell υπόλοιπα προέρχονται από μια εκθετική κατανομή με μοναδιαία παράμετρο, όταν το προσαρμοσμένο μοντέλο είναι το σωστό, κατά συνέπεια τόσο η μέση τιμή όσο και η διακύμανση αυτών των υπολοίπων θα ισούται με 1 και σύμφωνα με ιδιότητα της εκθετικής κατανομής δεν θα κατανέμονται συμμετρικά γύρω από τον μέσο. Επομένως αν απεικονίσουμε τα υπόλοιπα αυτά σε κάποιο διάγραμμα έναντι του αριθμού των παρατηρήσεων (γνωστό ως index plot) δεν θα έπρεπε να παρατηρήσουμε κάποια συμμετρική κίνηση γύρω από τον μέσο, ακόμη εφόσον τα υπόλοιπα αυτά σχετίζονται με τον χρόνο ούτε κάποια διάγραμμα των υπολοίπων με τα τεταρτημόρια θα ήταν χρήσιμο. Για αυτό καταλήγουμε στο συμπέρασμα ότι ο μόνος τρόπος για να ελέγξουμε αν το

προσαρμοσμένο μοντέλο είναι το κατάλληλο είναι να διαπιστώσουμε ότι πράγματι τα Cox – Snell υπόλοιπα προέρχονται από μια μοναδιαία εκθετική κατανομή. Αυτό που κάνουμε είναι να απεικονίσουμε σε ένα γράφημα τον λογάριθμο της αθροιστικής συνάρτησης κινδύνου έναντι του χρόνου. Το αποτέλεσμα θα πρέπει να είναι μια ευθεία γραμμή που ξεκινάει από την αρχή των αξόνων και έχει μοναδιαία κλίση.

Αρχικά υπολογίσουμε τα Cox – Snell υπόλοιπα (r_{Ci}) έπειτα υπολογίσουμε τον K-M εκτιμητή για τα r_{Ci} και τον χρόνο επιβίωσης για τα δεδομένα μας, τώρα τα r_{Ci} υπόλοιπα εφόσον έχουν προέλθει από λογοκριμένους χρόνους θα είναι επίσης λογοκριμένα. Επομένως το προσαρμοσμένο μοντέλο θα είναι το κατάλληλο μοντέλο για τα δεδομένα μας αν το γράφημα της ποσότητας $\hat{H}(r_{Ci}) = -\log \hat{S}(r_{Ci})$ έναντι των r_{Ci} υπολοίπων είναι μια ευθεία γραμμή, με μοναδιαία κλίση και μηδενική σταθερά. Αντίθετα αν στο διάγραμμα των παραπάνω ποσοτήτων παρατηρήσουμε κάποια συμμετρική απομάκρυνση από την ευθεία γραμμή ή δεν υπάρχει μηδενική σταθερά ή δεν έχουμε μοναδιαία κλίση τότε το προσαρμοσμένο μοντέλο δεν είναι το σωστό και επιδέχεται τροποποίησης.

- **Για τα Martingale υπόλοιπα**

Τα martingale υπόλοιπα εκφράζουν διαφορές ανάμεσα στον παρατηρούμενο και αναμενόμενο αριθμό θανάτων σε κάποιο χρονικό διάστημα $(0,t_i)$ για το i^o άτομο. Στην ουσία μας δείχνουν σύμφωνα με το κάθε φορά προσαρμοσμένο μοντέλο ποια άτομα έχουν πεθάνει πολύ νωρίς ή έχουν ζήσει για πολύ. Συγκεκριμένα μεγάλες αρνητικές τιμές των υπολοίπων αντιστοιχούν σε άτομα με μεγάλο χρόνο επιβίωσης, για τα οποία οι τιμές των μεταβλητών στο μοντέλο δείχνουν ότι θα έπρεπε να είχαν πεθάνει νωρίτερα. Αντίθετα ένα martingale υπόλοιπο με τιμή κοντά στην μονάδα δηλώνει άτομο με μικρό χρόνο ζωής.

Τα martingale υπόλοιπα χρησιμοποιούνται προκειμένου να εξετάσουμε αν η συναρτησιακή μορφή των μεταβλητών που περιλαμβάνονται στο μοντέλο είναι η κατάλληλη. Για να το διαπιστώσουμε αυτό απεικονίσουμε σε ένα γράφημα τα martingale υπόλοιπα αρχικά έναντι του μηδενικού μοντέλου και στην συνέχεια για κάθε επεξηγηματική μεταβλητή. Στην περίπτωση που το γράφημα των υπολοίπων με κάποια επεξηγηματική μεταβλητή είναι μια ευθεία γραμμή τότε αυτό συνεπάγει ότι θα πρέπει να προσθέσουμε έναν γραμμικό όρο στο μοντέλο, καθώς η συναρτησιακή μορφή της συγκεκριμένης μεταβλητής χαλάει την προσαρμογή του μοντέλου.

Ακόμα θα μπορούσαμε αν γνωρίζαμε την συναρτησιακή μορφή που θα έπρεπε να έχουν κάποιες μεταβλητές να τις προσαρμόσουμε σε ένα μοντέλο παλινδρόμησης του Cox και να υπολογίσουμε τα martingale υπόλοιπά τους. Στην συνέχεια τα υπόλοιπα αυτά θα πρέπει να τα απεικονίσουμε έναντι των μεταβλητών που η συναρτησιακή μορφή χρειάζεται να προσδιοριστεί. Βέβαια το γράφημα που παίρνουμε συνήθως είναι πολύπλοκο και δύσκολο να ερμηνευτεί για αυτό χρησιμοποιούμε κάποιο smoothers με γνωστή την lowess ή loess (η οποία προτάθηκε από το Cleveland το 1979).

- **Για τα Deviance υπόλοιπα**

Τα deviance υπόλοιπα όπως αναφέραμε και παραπάνω είναι μια κανονική μετατροπή των martingale υπολοίπων ώστε τα δεύτερα να κατανέμονται πιο συμμετρικά γύρω από το μηδέν. Θετικές τιμές των deviance υπολοίπων αντιστοιχούν σε άτομα που πέθαναν πολύ νωρίς από ότι αναμενόταν, ενώ αρνητικές τιμές δηλώνουν άτομα που έζησαν πολύ. Πολύ μεγάλες θετικές ή αρνητικές τιμές των deviance υπολοίπων αντιστοιχούν σε έκτροπες παρατηρήσεις, για τις οποίες το προσαρμοσμένο μοντέλο δεν είναι ικανοποιητικό. Τέτοιες παρατηρήσεις μας ωθούν στο να κάνουμε περισσότερους ελέγχους προκειμένου να αποφασίσουμε αν είναι αναγκαίο ή όχι να συμπεριλάβουμε τις παρατηρήσεις αυτές στο μοντέλο. Ένα διάγραμμα των deviance υπολοίπων έναντι του αριθμού των παρατηρήσεων μας πληροφορεί για την ύπαρξη ή μη έκτροπων παρατηρήσεων, είναι το πιο κατάλληλο διάγραμμα για αυτό το σκοπώ.

- **Για τα Schoenfeld υπόλοιπα**

Τα Schoenfeld υπόλοιπα αποτελούν ένα γραφικό τρόπο για να ελέγξουμε αν το εξεταζόμενο κάθε φορά μοντέλο ικανοποιεί την υπόθεση αναλογικού κινδύνου. Έτσι όπως ορίζονται τα υπόλοιπα αυτά είναι ανεξάρτητα του χρόνου, το οποίο σημαίνει ότι αν τα απεικονίσουμε σε ένα γράφημα έναντι του χρόνου και παρατηρήσουμε κάποια μη τυχαία πορεία έχουμε ένδειξη παραβίασης της υπόθεσης αναλογικού κινδύνου.

(Για την παραπάνω παράγραφο βλέπε: Collett 2000, Hess 1995, Gillespie 2006)

5.4 Γραφικές μέθοδοι ελέγχου της υπόθεσης αναλογικού κινδύνου

Όπως έχουμε αναφέρει στο προηγούμενο κεφάλαιο το μοντέλο που εισήγαγε ο Cox είναι ένα αναλογικό μοντέλο κινδύνου, αφού ο σχετικός κίνδυνος δύο οποιονδήποτε ατόμων είναι μια σταθερή ποσότητα ανεξάρτητη από το χρόνο, δηλ. ισχύει

$$\frac{h_0(t)\exp(z_i\beta)}{h_0(t)\exp(z_j\beta)} = \exp\beta(z_i - z_j)$$

Στην περίπτωση όμως των χρόνο – εξαρτώμενων μεταβλητών κάτι τέτοιο δεν ισχύει καθώς ο λόγος κινδύνου εξαρτάται από το χρόνο,

$$\frac{\exp(z_i(t)\beta)}{\exp(z_j(t)\beta)}$$

Μια πρώτη εικόνα για το αν ισχύει ή όχι η υπόθεση αναλογικού κινδύνου μπορούμε να έχουμε από το γράφημα της συνάρτησης επιβίωσης στην περίπτωση που τα δεδομένα μας αφορούν χρονικά σταθερές μεταβλητές και με μικρό αριθμό επιπέδων. Αν ισχύει η υπόθεση αναλογικού κινδύνου τότε θα πρέπει το γράφημα της λογαριθμικής συνάρτησης επιβίωσης να δείχνει μια σταθερά ανισορροπία μεταξύ της S(t) και του χρόνου ζωής t, αφού για το μοντέλο του Cox ισχύει

$$S_i(t) = \exp(-h_0(t)\beta z_i) \text{ και} \log(-\log S_i(t)) = \log(h_0(t)) - z_i\beta$$

Έτσι αν το μοντέλο που έχουμε προσαρμόσει είναι το κατάλληλο τότε θα πρέπει το γράφημα του Kaplan – Meier εκτιμητή για κάθε επίπεδο των μεταβλητών να απεικονίζει παράλληλες γραμμές στην περίπτωση που παραστήσουμε τον εκτιμητή σε κλίμακα log – log. Η απόσταση ανάμεσα σε κάθε γραμμή θα πρέπει να είναι σχεδόν ίση με την εκτίμηση του συντελεστή β, που έχει προκύψει από το μοντέλο αναλογικού κινδύνου που έχουμε προσαρμόσει.

Όταν όμως οι υπό εξέταση μεταβλητές αποτελούνται από πολλά επίπεδα και είναι και συνεχείς τότε το γράφημα του Kaplan – Meier εκτιμητή θα είναι πολύπλοκο, δύσκολο να ερμηνευτεί και έτσι δεν θα μπορέσουμε να πάρουμε κάποια ένδειξη για το αν ισχύει ή όχι η υπόθεση αναλογικού κινδύνου.

Μια άλλη γραφική μέθοδος ελέγχου της υπόθεσης αναλογικού κινδύνου είναι μέσω των συγκεντρωτικών αθροισμάτων των Schoenfeld υπολοίπων. Σύμφωνα με την

οποία κάτω από την υπόθεση του αναλογικού κινδύνου θα έπρεπε το γράφημα των υπολοίπων για το προσαρμοσμένο μοντέλο να έχει την μορφή ενός Brownian bridge. Τα διαγράμματα αυτά είναι δύσκολο να γίνουν κατανοητά και να ερμηνευτούν καθώς δεν είναι εύκολο να φανταστεί κάποιος την πραγματική πορεία τους και κατά συνέπεια να πάρει μια απόφαση για την παραβίαση ή μη της υπόθεσης αναλογικού κινδύνου.

(Βλέπε Therneau & Grambsch 2000)

Να αναφέρουμε ότι αν προσαρμόσουμε είτε γραφικά είτε μέσω ενός μοντέλου παλινδρόμησης τα Schoenfeld υπόλοιπα με το χρόνο μπορούμε να ελέγξουμε την ανεξαρτησία ή όχι μεταξύ τους. Η υπόθεση αναλογικού κινδύνου ενισχύεται από μια μη σημαντική σχέση ανάμεσα τους και αντίστοιχα παραβιάζεται από μια σημαντική σχέση μεταξύ τους.

Στην περίπτωση των χρόνο – εξαρτώμενων μεταβλητών, αν για κάποια μεταβλητή ισχύει η υπόθεση αναλογικού κινδύνου τότε θα πρέπει το διάγραμμα των Scaled Schoenfeld υπολοίπων και του Smooth του να μην εμφανίζουν κάποια τάση στην διάρκεια του χρόνου. Θα μπορούσε παραδείγματος χάρη να ισχύει η αναλογικότητα για κάποια συνεχή μεταβλητή ενώ για κάποια δίτιμη μεταβλητή η συνάρτηση κινδύνου να μην είναι αναλογική. Στην περίπτωση αυτή το διάγραμμα των Scaled Schoenfeld υπολοίπων θα έδειχνε μια ισχυρή αρχικά θετική κλίση η οποία θα μειώνεται καθώς προχωράμε προς τα δεξιά του διαγράμματος. Ένα τέτοιο διάγραμμα συνήθως σημαίνει ότι η δίτιμη μεταβλητή μπορεί να ήταν ένας καθοριστικός παράγοντας για την επιβίωσης αρχικά όμως καθώς προχωράμε στο χρόνο κάτι τέτοιο παύει να ισχύ. Αντίστοιχα το πολύγωνο του Smooth των υπολοίπων για την συνεχή μεταβλητή θα έπρεπε να έχει μηδενική κλίση, γεγονός που ενισχύει την υπόθεση της μη σημαντικότητας του όρο αλληλεπίδρασης με το χρόνο.

Επομένως η διαδικασία ελέγχου της υπόθεσης αναλογικού κινδύνου συνοπτικά θα μπορούσε να περιγραφεί μέσω δύο βημάτων όπου:

$1^{ο}$ βήμα: Προσθέτουμε στο μοντέλο όλες τις μεταβλητές μαζί και τις αλληλεπιδράσεις τους με το λογάριθμο του χρόνου (z* log(t)). Αποφασίζουμε για την σημαντικότητα τους με βάση το τεστ λόγου πιθανοφανειών, ή το score ή το Wald test ή γραφικά μέσω απεικόνισης του Kaplan –Meier εκτιμητή με το χρόνο σε κλίμακα

log- log. Η υπόθεση αναλογικού κινδύνου ενισχύεται όταν οι γραμμές για κάθε επίπεδο των μεταβλητών στο απεικονιζόμενο γράφημα είναι παράλληλες.

2^{ο} βήμα: Απεικονίζουμε σε γράφημα τα Scaled Schoenfeld υπόλοιπα καθώς και το πολύγωνο του Smooth τους, χωρίς να συμπεριλάβουμε τους όρους αλληλεπίδρασης . Για να ισχύει η υπόθεση αναλογικού κινδύνου θα πρέπει τα υπόλοιπα αυτά να εμφανίζουν μια τυχαία πορεία σε αντίθετη περίπτωση έχουμε ένδειξη παραβίασης της υπόθεσης αναλογικού κινδύνου.

Βέβαια τα αποτελέσματα που έχουμε από τα δύο βήματα θα πρέπει να ενισχύουν η μια την άλλη.

(Hess 1995, Sainani Ph.D, Gillespie 2006)

5.4.1 Τα Score Residuals σε ρόλο Leverage για το μοντέλο παλινδρόμησης του Cox

Στο μοντέλο αναλογικού κινδύνου τα score υπόλοιπα έχουν αντίστοιχο ρόλο με αυτό των leverage στην γραμμική και λογιστική παλινδρόμηση. Τα leverage εκφράζουν

ένα διαγνωστικό στατιστικό το οποίο μετράει το πόσο ασύνηθες είναι οι τιμές κάποιας μεταβλητής για κάποιο άτομο. Στην γραμμική και λογιστική παλινδρόμηση τα leverage υπολογίζονται σαν μια απόσταση ανάμεσα στις τιμές των μεταβλητών για κάποιο άτομο και τον συνολικό μέσο των μεταβλητών, είναι ανάλογα της ποσότητας $(x-\bar{x})^2$. Τα leverage είναι χρήσιμα λόγω των σημαντικών ιδιοτήτων που διαθέτουν, όπως το χαρακτηριστικό ότι είναι πάντα θετικά και αθροίζουν το δείγμα με τον αριθμό των παραμέτρων στο μοντέλο. Όμως για το μοντέλο παλινδρόμησης του Cox τα leverage δεν είναι εύκολο να οριστούν ούτε έχουν και τις ίδιες ιδιότητες και αυτό διότι τα ίδια άτομα μπορεί να εμφανιστούν σε περισσότερα του ενός σετ κινδύνου και έτσι θα έχουμε πολλαπλούς όρους στην εξίσωση μερικής πιθανοφάνειας.

Για αυτό χρησιμοποιούμε τα score υπόλοιπα τα οποία ορίζονται ως εξής:

$$\hat{rs}_{ik} = \delta_i (x_{ik} - \hat{\bar{x}}_{wik}) - x_{ik} \cdot \hat{H}(t_i, x, \hat{\beta}) + e^{x_i'\hat{\beta}} \sum_{t_j \le t_i} \hat{\bar{x}}_{wjk} \frac{\delta_j}{\sum_{l \in R_j} e x_l' \hat{\beta}}$$

Με αντίστοιχο διάνυσμα για τα score υπόλοιπα το:

$$\hat{rs}_i = (\hat{rs}_{i1}, \hat{rs}_{i2}, ..., \hat{rs}_{ip})$$

τα οποία σχηματίζουν τον πυρήνα των διαγνωστικών του αναλογικού μοντέλου κινδύνου. Από όπου προκύπτει πως το score υπόλοιπο για το άτομο i πάνω στην k μεταβλητή είναι ένας σταθμισμένος μέσος της απόστασης της x_{ik} τιμής με το μέσο του σετ κινδύνου, x_{wjk}, όπου τα βάρη δηλώνουν αλλαγή στο martingale υπόλοιπο ($d(M_i(t_j))$. Επομένως για μια συνεχή μεταβλητή το score υπόλοιπο (ή όπως αποκαλείται ορισμένες φορές το partial leverage residual) θα δηλώνει, ότι και η μόχλευση στην γραμμική παλινδρόμηση, και όσο μεγαλύτερη θα είναι η διαφορά στην τιμή της μεταβλητής από το μέσο τόσο μεγαλύτερο θα είναι κα το score υπόλοιπο, βέβαια στο μοντέλο παλινδρόμησης του Cox η μεγαλύτερη αυτή τιμή μπορεί να είναι είτε θετική είτε αρνητική.

Τις περισσότερες φορές όμως δεν μας ενδιαφέρει το πόσο μεγάλη είναι η τιμή του υπολοίπου αλλά ποιά επίδραση έχει η τιμή της συγκεκριμένης μεταβλητής πάνω στην εκτίμηση του συντελεστή προσδιορισμού. Στην γραμμική παλινδρόμηση για να εκτιμήσουμε την επίδραση αυτή χρησιμοποιούμε την απόσταση του Cook. Σκοπός της απόστασης του Cook είναι να υπολογίζει ένα στατιστικό που να προσεγγίζει την αλλαγή στην εκτιμώμενη τιμή του συντελεστή όταν αφαιρέσουμε την παρατήρηση με το μεγαλύτερο leverage . Για την γραμμική παλινδρόμηση το στατιστικό αυτό έχει την μορφή:

$$\Delta\hat{\beta}_{ki} = \hat{\beta}_k - \hat{\beta}_{(k-i)} \qquad (5.1)$$

όπου $\hat{\beta}_k \rightarrow$ εκτιμητής μερικής πιθανοφάνειας όταν χρησιμοποιούμε όλο το δείγμα

$\hat{\beta}_{(k-i)} \rightarrow$ εκτιμητής μερικής πιθανοφάνειας όταν έχουμε αφαιρέσει το i^o άτομο.

Παρόμοια απόσταση για το μοντέλο αναλογικού κινδύνου αναπτύχθηκε από τους Cain & Lange (1984), η απόσταση αυτή μας δίνει ένα προσεγγιστικό εκτιμητή για την σχέση (5.1) και ορίζεται ως εξής:

$$\Delta\hat{\beta}_i = (\hat{\beta} - \hat{\beta}_{(-i)}) = Var(\hat{\beta}) \cdot r\hat{s}_i$$

όπου $r\hat{s}_i$ είναι το διάνυσμα των score υπολοίπων και $Var(\hat{\beta})$ είναι ο εκτιμώμενος πίνακας συνδιακύμανσης των εκτιμωμένων συντελεστών.

(Βλέπε Hosmer & Lemeshow 1999)

ΚΕΦΑΛΑΙΟ 6^{Ο}

Ανάλυση δεδομένων – Εφαρμογή Kaplan – Meier εκτιμητή

6.1 Εισαγωγή

Τα δεδομένα που έχουμε να αναλύσουμε αναφέρονται στους χρόνους επιβίωσης 870 γυναικών με καρκίνο του μαστού που χειρουργήθηκαν στο νοσοκομείο ΙΑΣΩ την χρονική περίοδο 1980 – 2006.

Για την συγκεκριμένη περίοδο για την κάθε ασθενή καταγράφτηκαν οι εξής μεταβλητές:

- CODE, η μεταβλητή αυτή δηλώνει τον κωδικό της ασθενούς και δεν θα χρησιμοποιηθεί καθόλου για την ανάλυση των χρόνων επιβίωσής τους
- MONTH, δηλώνει το χρόνο παρακολούθησης των ασθενών (σε μήνες) από το χειρουργείο μέχρι το θάνατο ή τη λήξη του χρόνου παρακολούθησης των ασθενών.
- DEATH, δίτιμη τυχαία μεταβλητή που δηλώνει αν ο χρόνος ζωής της ασθενούς είναι πλήρης (death = 1) ή λογοκριμένος (death = 0).
- AGE, συνεχής μεταβλητή, η οποία μας δίνει την ηλικία της ασθενούς όταν έγινε το χειρουργείο.
- GRADE, εκφράζει την διαφοροποίηση του όγκου όταν αυτό μετρήθηκε μετά από το χειρουργείο. Παίρνει τις τιμές 1,2 3 ανάλογα με το αν παρατηρείται υψηλή, μεσαία ή χαμηλή διαφοροποίηση και την τιμή 0 στην περίπτωση που δεν γνωρίζουμε ποια είναι η διαφοροποίηση του όγκου.
- POSITIVE, δηλώνει τον αριθμό των διηθημένων λεμφαδένων. Ο συγκεκριμένος παράγοντας σχετίζεται αρνητικά με την επιβίωσης των ασθενών, όσο μεγαλύτερος είναι ο αριθμός των διηθημένων λεμφαδένων τόσο μεγαλύτερη είναι και η πιθανότητα υποτροπής της νόσου δηλ. η πιθανότητα επανεμφάνισης ή μετάστασης του καρκίνου. Δεν έχει προσδιοριστεί πόσο θα πρέπει να είναι ο αριθμός των διηθημένων λεμφαδένων, όμως μια κατηγοριοποίηση που συναντάμε συχνά (Perperoglou et al 2005) είναι η εξής: $positive = \begin{pmatrix} 0-3 \\ 4-8 \\ \geq 9 \end{pmatrix}$

- SIZE, συνεχής μεταβλητή η οποία μας δίνει το μέγεθος του όγκου σε χιλιοστά, μια συχνή κατηγοριοποίηση είναι η ακόλουθη: $size = \begin{pmatrix} 0-20\,mm \\ 21-50\,mm \\ \geq 50\,mm \end{pmatrix}$

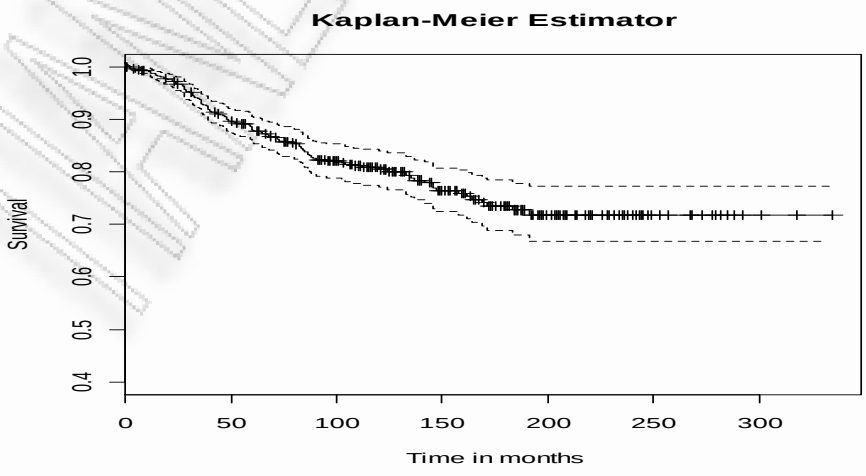
Η μελέτη των δεδομένων θα βασιστεί στο μοντέλο αναλογικού κινδύνου του Cox . Αρχικά όμως προκειμένου να κατανοήσουμε και να εξοικειωθούμε καλύτερα με τα δεδομένα που έχουμε να αναλύσουμε θα εφαρμόσουμε τον Kaplan – Meier εκτιμητή συνοδευόμενο από τις γραφικές απεικονίσεις και τα τεστ που περιγράψαμε στο $2^{ο}$ και $3^{ο}$ κεφάλαιο.

6.2 Εφαρμογή του Kaplan – Meier εκτιμητή

Ο Kaplan – Meier εκτιμητής όπως αναφέραμε και στην θεωρία χρησιμοποιείται προκειμένου να εκτιμήσουμε την S(t) όταν στο υπό εξέταση δείγμα περιλαμβάνονται λογοκριμένοι χρόνοι.

Στο γράφημα του εκτιμητή (γράφημα 6.1) δεν παρατηρούμε κάτι διαφορετικό από αυτό που αναμέναμε με βάση την θεωρία που αναφέραμε σχετικά με τον συγκεκριμένο εκτιμητή. Βλέπουμε πως η τιμή του εκτιμητή ξεκινάει από την μονάδα, αλλάζει τιμή μόνο στα σημεία που έχουμε αποτυχίες, δηλ. σε αυτά που ο χρόνος ζωής της ασθενούς είναι γνωστός, ακόμη η τιμή του εκτιμητή μειώνεται κάθε φορά συναρτήσει της ποσότητας $(r_j-d_j)/r_j$. Οι λογοκριμένες παρατηρήσεις δηλώνονται με κάθετες γραμμές πάνω στο γράφημα και εφόσον η τελευταία παρατήρηση είναι λογοκριμένος χρόνος ο K-M εκτιμητής θα υπολογιστεί μέχρι εκείνη την στιγμή.

Γράφημα 6.1



Ο K-M

εκτιμητής μπορεί επίσης να χρησιμοποιηθεί και για την εκτίμηση ποσοστιαίων σημείων του χρόνου επιβίωσης μιας υπό εξέταση ομάδας. Ένα p ποσοστιαίο σημείο θα πρέπει να ικανοποιεί τις σχέσεις:

$$S(t_p) \geq 1-p \quad \text{και } S(t_p+) \leq 1-p$$

οι οποίες προκύπτουν ως συνέπεια των σχέσεων

$$F(t_p-) = P(T < t_p) \leq p \quad \text{και} \quad F(t_p) = P(T \leq t_p) \geq p$$

Έτσι, μια πιθανή εκτίμηση $\hat{t}_p$ του t_p ποσοστιαίου σημείου προκύπτει από την σχέση:

$$\hat{t}_p = \min\{t_j : \hat{S}(t_j+) \leq 1-p\} = \max\{t_j : \hat{S}(t_j) \geq 1-p\}$$

(Αντζουλάκος, 2009)

Έναν γραφικό τρόπο εκτίμησης της διαμέσου του χρόνου ζωής, αποκτάμε παρατηρώντας απλώς το γράφημα του K – M εκτιμητή και φέροντας μια ευθεία παράλληλη με τον οριζόντιο άξονα από το σημείο 0.5, και όπου η ευθεία αυτή τέμνει την γραφική παράσταση της S(t) εκεί εντοπίζεται η διάμεσος.

Ενώ μέσω του στατιστικού πακέτου R ο διάμεσος χρόνος ζωής υπολογίζεται εισάγοντας την παρακάτω εντολή:

```
afit<-survfit(Surv(MONTH,DEATH)~1,data=a)
```

records	n.max	n.start	events	median	0.95LCL	0.95UCL
870	870	870	125	NA	NA	NA

Τα αποτελέσματα που παίρνουμε συμφωνούν και με το διάγραμμα, αφού η ευθεία που φέρνουμε δεν τέμνει πουθενά το γράφημα της S(t). Άρα ο διάμεσος χρόνος ζωής για τους συγκεκριμένους χρόνους επιβίωσης δεν μπορεί να εκτιμηθεί, αφού πάνω από τους μισούς ασθενείς βρίσκονται εν ζωή την χρονική στιγμή που τελειώνει η έρευνα. Ακόμη από το διάγραμμα βλέπουμε πως δεν ορίζεται ούτε το ποσοστιαίο σημείο $t_{0.75}$ ενώ για το ποσοστιαίο σημείο $t_{0.25}$ διακρίνουμε ότι το 95% διάστημα εμπιστοσύνης είναι κοντά στο (80,160). Λόγω των αρκετών λογοκριμένων παρατηρήσεων η ακριβής τιμή δεν προσδιορίζεται εύκολα από το διάγραμμα, φαίνεται ότι η εκτίμηση για το τρίτο τεταρτημόριο είναι περίπου ίση με 130.

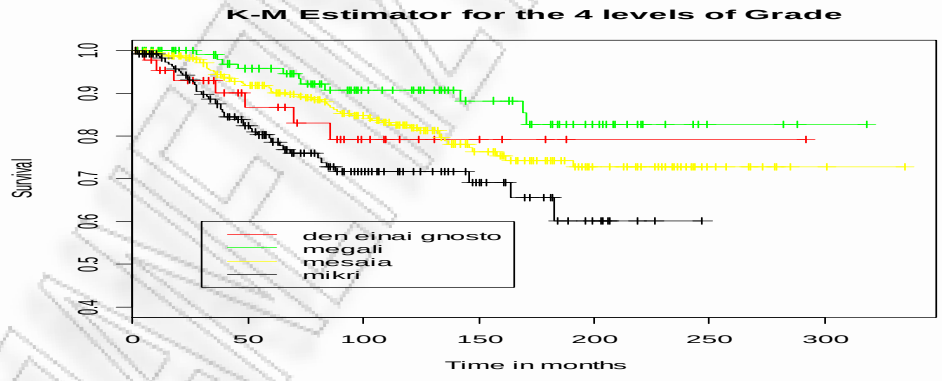
6.2.1 Εφαρμογή του Kaplan – Meier εκτιμητή στην περίπτωση συμμεταβλητών

Στην περίπτωση που θέλουμε να μελετήσουμε την επιβίωση για κάθε επίπεδο μιας συμμεταβλητής τότε θα πρέπει αντί της μονάδας να εισάγουμε το όνομα της συμμεταβλητής.

Η μεταβλητή GRADE (Για τον πίνακα επιβίωσης της μεταβλητής Grade βλέπε παράρτημα), η οποία εκφράζει την διαφοροποίηση του όγκου στις μετρήσεις που ακολούθησαν μετά την επέμβαση , όπως αναφέραμε στην αρχή του κεφαλαίου, έχει τέσσερα επίπεδα, το GRADE=0 όταν δεν είναι γνωστή η διαφοροποίηση, το GRADE=1 για μεγάλη διαφοροποίηση, το GRADE=2 για μεσαία, και το GRADE=3 για μικρή διαφοροποίηση.

Αν ενδιαφερόμαστε να δούμε αν υπάρχει κάποια διαφορά στην συνάρτηση επιβίωσης μεταξύ των τεσσάρων επιπέδων της μεταβλητής θα πρέπει αρχικά να απεικονίσουμε τις συναρτήσεις επιβίωσης των επιπέδων της μεταβλητής σε ένα κοινό γράφημα και στην συνέχεια να αποφασίσουμε ποιο έλεγχο ισότητας συναρτήσεων επιβίωσης θα εφαρμόσουμε.

Γράφημα 6.2



Δεν φαίνεται να υπάρχει κάποια σημαντική διαφορά στις συναρτήσεις επιβίωσης των τεσσάρων ομάδων. Όμως μπορούμε να δούμε πως η καμπύλη της συνάρτησης επιβίωσης για την μικρή ομάδα δεν τέμνεται με αυτή της μεσαίας και τις υψηλής ομάδας, άρα οι συναρτήσεις επιβίωσης για τις ομάδες αυτές περιμένουμε να διαφέρουν ενώ τέμνεται με την καμπύλη επιβίωσης της ομάδας που δεν είναι γνωστή η διαφοροποίηση του όγκου. Η ομάδα που παρουσιάζει την μεγαλύτερη

διαφοροποίηση έχει και την υψηλότερη πιθανότητα επιβίωσης έναντι των υπολοίπων, πράγμα που είναι αναμενόμενο διότι η μεγαλύτερη διαφοροποίηση του όγκου σημαίνει μικρότερη συρρίκνωση. Την μικρότερη πιθανότητα επιβίωσης έχουν οι ασθενείς με την μικρότερη διαφοροποίηση του όγκου.

Για να συγκρίνουμε όμως με μεγαλύτερη αξιοπιστία αν διαφέρουν οι συναρτήσεις επιβίωσης για τα τέσσερα επίπεδα της μεταβλητής GRADE θα εφαρμόσουμε στατιστικό έλεγχο. Η μορφή του στατιστικού ελέγχου θα είναι η εξής:

H_0: $S_1(t)=S_2(t)=S_3(t)=S_4(t)$ για κάποιο $t\leq\tau$

έναντι της

H_1: τουλάχιστον ένα από τα $S_i(t)$ να διαφέρει από τα υπόλοιπα για όλα τα $t\leq\tau$, όπου τ είναι ο μέγιστος πλήρης χρόνος.

Τα τεστ που χρησιμοποιούνται πιο συχνά στην πράξη για τον έλεγχο της παραπάνω υπόθεσης είναι γνωστά ως log–rank test και Breslow-Gehan test. (βλ. 3^{ο} κεφάλαιο)

Η επιλογή μεταξύ του log–rank και Breslow-Gehan test θα εξαρτάται από το αν οι καμπύλες επιβίωσης του γραφήματος 6.2 είναι παράλληλες ή όχι. Στην περίπτωση που οι καμπύλες του διαγράμματος είναι παράλληλες τότε θα εφαρμόσουμε το log – rank test, το οποίο έχει μεγαλύτερη ισχύ όταν ισχύει η υπόθεση αναλογικού κινδύνου (proportional hazard) καθώς δίνει ίδιο βάρος και στα δύο άκρα των πλήρων χρόνων ζωής. Σε αντίθετη περίπτωση (οι καμπύλες του διαγράμματος τέμνονται) δεν ικανοποιείται η υπόθεση αναλογικού κινδύνου και θα εφαρμόσουμε το Breslow-Gehan test, το οποίο δίνει βάρος σε μικρές τιμές των πλήρων χρόνων ζωής δηλαδή στο αριστερό άκρο της συνάρτησης επιβίωσης, όπου συγκεντρώνονται περισσότερες πληροφορίες σχετικά με τις συναρτήσεις επιβίωσης των υπό εξέταση ομάδων.

Στο γράφημα 6.2 οι καμπύλες επιβίωσης για τα τέσσερα επίπεδα της μεταβλητής GRADE φαίνονται να τέμνονται οπότε θα εφαρμόσουμε το Breslow-Gehan test.

Τα αποτελέσματα από την εκτέλεση του συγκεκριμένου τεστ δίνονται στο πίνακα που ακολουθεί:

Πίνακας 6.1

Breslow-Gehan test for Grade covariate

	Chi-Square	df	Sig.
Breslow (Generalized Wilcoxon)	23,677	3	,000

Test of equality of survival distributions for the different levels of GRADE.

Εφόσον η τιμή p-value είναι σχεδόν μηδέν, απορρίπτουμε την μηδενική υπόθεση υπέρ της εναλλακτικής για επίπεδο σημαντικότητας α=5%, συμπεραίνουμε ότι η συνολική επιβίωση μεταξύ των ασθενών διαφέρει για τα τέσσερα επίπεδα της μεταβλητής GRADE.

6.2.2 Μετατροπή συνεχών μεταβλητών σε κατηγορικές – Εφαρμογή του Kaplan – Meier εκτιμητή

Στα δεδομένα επιβίωσης η μεταβλητή που δηλώνει την ηλικία παρουσιάζει ενδιαφέρον όταν εκφράζεται ως παράγοντας (factor), καθώς επιδιώκουμε να καταλήγουμε σε συμπεράσματα που αφορούν το γενικό σύνολο του πληθυσμού και όχι μεμονωμένα άτομα, για αυτό θεωρώ αναγκαίο να διακρίνω την μεταβλητή που δηλώνει τις ηλικίες των ασθενών σε τρείς ομάδες. Οι δύο πρώτες ομάδες έχουν το ίδιο εύρος, συγκεκριμένα στην $1^{η}$ ομάδα περιλαμβάνονται γυναίκες ηλικίας, 21 έως και 40 χρονών, στην $2^{η}$ ομάδα έχουμε γυναίκες ηλικίας 41 έως και 60 χρονών, στην $3^{η}$ ομάδα έχουμε γυναίκες ηλικίας 61 χρονών και άνω.

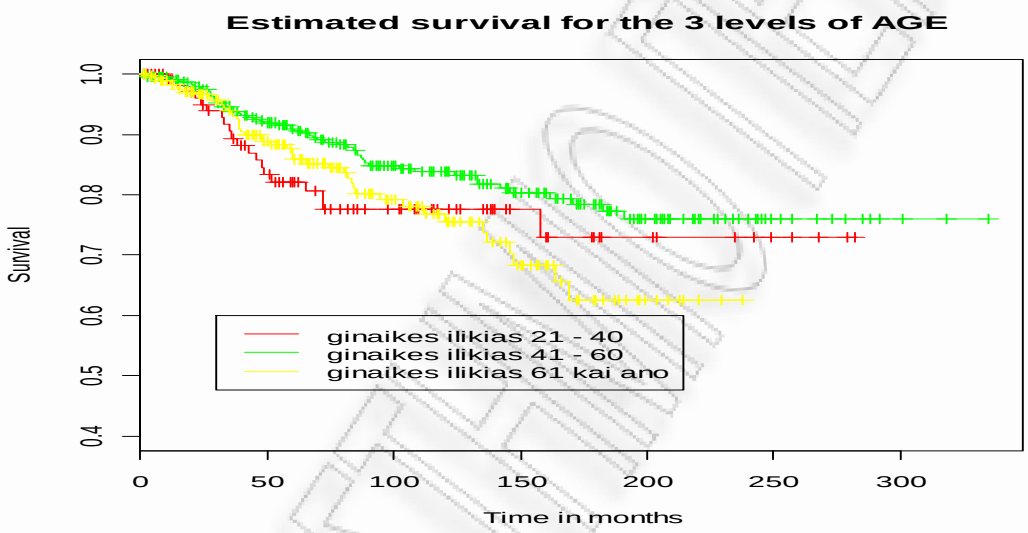
Για να μετατρέψουμε την μεταβλητή age σε μια νέα μεταβλητή, την newage, η οποία παίρνει τρείς τιμές ανάλογα με την ηλικία της ασθενούς κάθε φορά, εκτελούμε την παρακάτω ρουτίνα στο R. Στο $1^{ο}$ βήμα του κώδικα που ακολουθεί ορίζουμε να πάρει την μεταβλητή age από το a, όπως έχουμε ονομάσει το αρχείο που περιέχει τα δεδομένα, να το αντιστοιχίσει με την newage η οποία για κάποιο i από το 1 έως το μήκος της age (δηλ. 870) θα παίρνει την τιμή 1 αν η ηλικία της ασθενούς είναι μεγαλύτερη των 20 χρονών και μικρότερη ή ίση των 40 χρονών, την τιμή 2 για ασθενείς ηλικίας 41 έως 60 χρονών και την τιμή 3 αν η ηλικία της ασθενούς την στιγμή της επέμβασης είναι μεγαλύτερη των 60 χρονών.

```
AGE=a$AGE
NEWAGE=a$AGE
for(i in 1:length(AGE)) {
if (AGE[i]>20 & AGE[i]<=40) {NEWAGE[i]=1} else
```

```
if (AGE[i]>40 & AGE[i]<=60) {NEWAGE[i]=2} else
if (AGE[i]>60) {NEWAGE[i]=3}
}
```

Εφαρμόζουμε τον Kaplan – Meier εκτιμητή για την συμμεταβλητή newage και κατασκευάσουμε το διάγραμμα των καμπύλων επιβίωσης για κάθε επίπεδο της μεταβλητής newage:

Γράφημα 6.3



Παρατηρώντας το παραπάνω γράφημα βλέπουμε πως οι καμπύλες επιβίωσης για τα τρία επίπεδα της μεταβλητής newage για τους 50 πρώτους μήνες έχουν την τάση να συμπέσουν μεταξύ τους , είναι σαν μια, στην συνέχεια διακρίνονται σε τρείς. Οι συναρτήσεις επιβίωσης που αφορούν της ασθενείς ηλικίας 21 – 40 χρονών και 41 – 60 χρονών, έχουν υψηλές πιθανότητες επιβίωσης και ακολουθούν την ίδια πορεία, με την πιθανότητα επιβίωσης για γυναίκες ηλικίας 41 – 60 χρονών να είναι υψηλότερη στο μεγαλύτερο μέρος σε σχέση με τις γυναίκες ηλικίας 21 – 40 χρονών. Η χαμηλότερη πιθανότητα επιβίωσης εμφανίζεται σε γυναίκες ηλικίας μεγαλύτερης των 60 χρονών, πράγμα το οποίο και περιμέναμε καθώς λόγω της ηλικίας τους ο οργανισμός είναι πιο αδύναμος, δεν ανταποκρίνεται γρήγορα στην θεραπεία και ακόμη οι γυναίκες αυτές έχουν και μεγαλύτερη πιθανότητα να πεθάνουν από άλλα αίτια εκτός του καρκίνου.

Για να ελέγξουμε την ισότητα των συναρτήσεων επιβίωσης για τα τρία επίπεδα της μεταβλητής newage θα εφαρμόσουμε το τεστ των Gehan – Breslow αφού οι καμπύλες φαίνονται να τέμνονται όχι μόνο στους πρώτους χρόνους αλλά και στην συνέχεια.

Πίνακας 6.2
Case Summary

New_age	Total N	N of Events	Censored	
	N	Percent	N	Percent
1,00	123	20	103	83,7%
2,00	448	58	390	87,1%
3,00	299	47	252	84,3%
Overall	870	125	745	85,6%

Πίνακας 6.3

Gehan-Breslow test for newage covariate

	Chi-Square	df	Sig.
Breslow (Generalized Wilcoxon)	5,235	2	,073

Test of equality of survival distributions for the different levels of New_age.

Σύμφωνα με τον παραπάνω πίνακα στην $1^{η}$ ομάδα έχουμε 123 ασθενείς, η ομάδα αυτή περιμέναμε να έχει τον μικρότερο αριθμό ασθενών, αυτό λόγω της ηλικίας των γυναικών αφού συνήθως ο καρκίνος του μαστού εμφανίζεται σε γυναίκες ηλικίας άνω των 35 ετών. Ακόμα είναι γνωστό ότι η πιθανότητα να εμφανίσει κάποια γυναίκα καρκίνο στο μαστό αυξάνεται όσο μεγαλώνει η ηλικία της. Στην δεύτερη ομάδα περιλαμβάνονται 448 ασθενείς, και 299 στην τελευταία ομάδα, επομένως οι περισσότερες γυναίκες που εμφάνισαν καρκίνο σε κάποιον ή και στους δύο μαστούς είναι ηλικίας 41 έως 60 χρονών, και η ομάδα αυτή είδαμε ότι είχε και την υψηλότερη επιβίωση.

Σύμφωνα με την τιμή p-value του ελέγχου που ισούται με 0.073 > 0.05, δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση υπέρ της εναλλακτικής, άρα οι συναρτήσεις επιβίωσης για τις τρείς ομάδες ηλικιών δεν διαφέρουν.

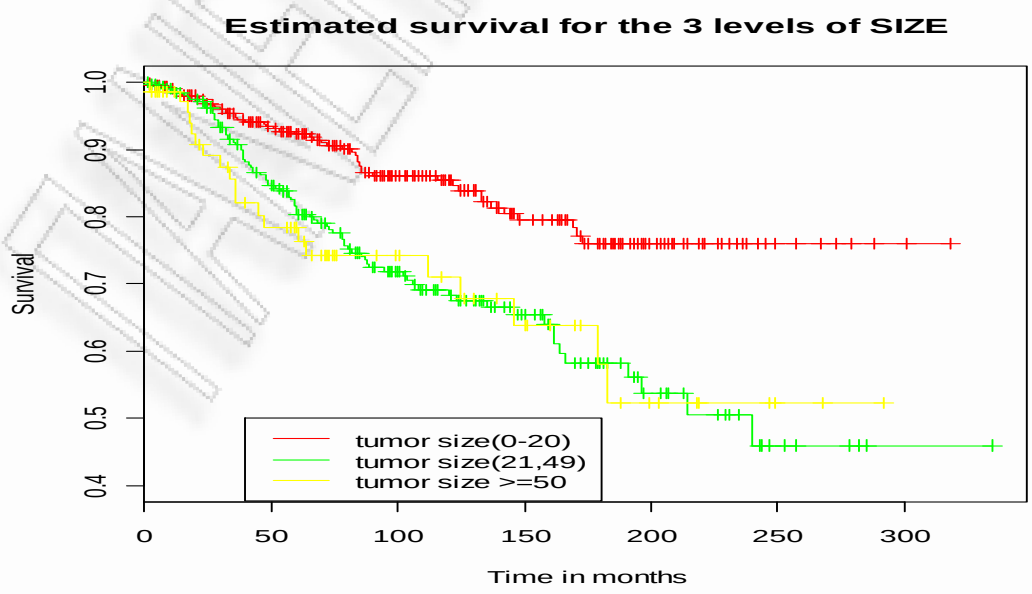
Από τις άλλες μεταβλητές που έχουμε προς εξέταση (size και positive), είναι μια συνεχής και μια διακριτή, μια πιθανή κατηγοριοποίηση τους προκειμένου να δούμε πώς συμπεριφέρονται οι καμπύλες επιβίωσης είναι αυτή που αναφέρουμε στην εισαγωγή του κεφαλαίου, σύμφωνα με παρόμοια δεδομένα που έχουν αναλυθεί από τους Perperoglou et al (2005) σε άρθρο τους, Η κατηγοριοποίηση τους στο συγκεκριμένο κεφάλαιο γίνεται μόνο για ερμηνευτικούς λόγους, όταν θα εφαρμόσουμε το μοντέλο του Cox για τα δεδομένα μας, θα χρησιμοποιήσουμε την αρχική μορφή των μεταβλητών.

Χρησιμοποιώντας την παρακάτω ρουτίνα κατηγοριοποιούμε την μεταβλητή size σε τρείς κατηγορίες, οι οποίες διακρίνονται στον παρακάτω κώδικα.

```
SIZE=a$SIZE
 NEWSIZE=a$SIZE
 for(i in 1:length(SIZE)){
 if(SIZE[i]>=0&SIZE[i]<=20) {NEWSIZE[i]=1} else
 if(SIZE[i]>=21&SIZE[i]<50) {NEWSIZE[i]=2} else
 if(SIZE[i]>=50) {NEWSIZE[i]=3}
 }
```

Εκτιμούμε τη συνάρτηση επιβίωσης για τα τρία επίπεδα της μεταβλητής size και στη συνέχεια κατασκευάσουμε το διάγραμμα των συναρτήσεων επιβίωσης.

Γράφημα 6.4



Από το παραπάνω διάγραμμα βλέπουμε πως η εκτιμώμενη επιβίωση για την ομάδα με μέγεθος όγκου (0 – 20 mm) βρίσκεται υψηλότερα σε σχέση με την επιβίωση για τις άλλες δύο ομάδες, οι καμπύλες των οποίων τέμνονται και οι πιθανότητες επιβίωσης κινούνται στα ίδια επίπεδα. Περιμένουμε λοιπόν, η επιβίωση ενός ατόμου να εξαρτάται από το μέγεθος του όγκου και μάλιστα όσο μικρότερος είναι τόσο μεγαλύτερη είναι και η πιθανότητα επιβίωσης του ατόμου, μπορούμε να δούμε όμως ότι στην περίπτωση που το μέγεθος του όγκου είναι μεγαλύτερο των 20 mm δεν περιμένουμε να παρατηρήσουμε σημαντικές διαφορές στις πιθανότητες επιβίωσης είτε ο όγκος βρίσκεται στο διάστημα (21 -50) είτε είναι μεγαλύτερος των 50 mm. Θα μπορούσαν αυτές οι δύο υπο-ομάδες να αποτελέσουν μια κοινή ομάδα, αφού η πιθανότητα συνολικής επιβίωσης δεν περιμένουμε να διαφοροποιείται σημαντικά μεταξύ των ασθενών με μέγεθος όγκου από 21-50 και μεγαλύτερο του 50.

Όλα τα παραπάνω αποτελούν απλώς ενδείξεις. Για να είμαστε βέβαιοι ότι οι συναρτήσεις επιβίωσης διαφέρουν για κάθε ομάδα της μεταβλητής size εκτελούμε τα γνωστά τεστ ισότητας συναρτήσεων επιβίωσης, και εφόσον δύο από τις τρείς καμπύλες τέμνονται θα εφαρμόσουμε το τεστ του Gehan με το οποίο δίνουμε μεγαλύτερο βάρος στο αριστερό άκρο των χρόνων επιβίωσης, που είναι συγκεντρωμένες οι περισσότερες πληροφορίες.

Πίνακας 6.4

Gehan-Breslow test for newsize covariate

	Chi-Square	df	Sig.
Breslow (Generalized Wilcoxon)	16,949	2	,000

Test of equality of survival distributions for the different levels of new_size.

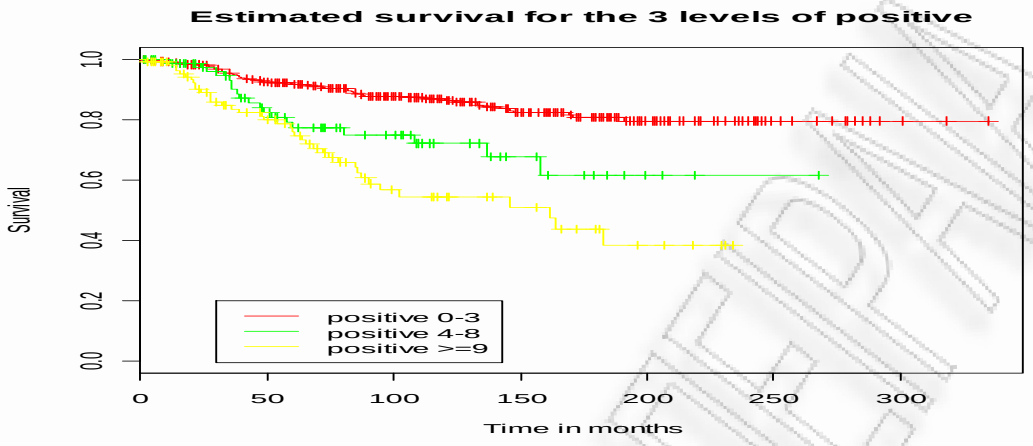
Σύμφωνα με το τεστ του Gehan απορρίπτουμε την υπόθεση ισότητας των συναρτήσεων επιβίωσης για τις τρεις ομάδες της μεταβλητής size υπέρ της εναλλακτικής υπόθεσης για επίπεδο σημαντικότητας 5%. Άρα τουλάχιστον μια από τις τρεις συναρτήσεις διαφέρει των υπολοίπων.

Στην κατηγοριοποίηση της μεταβλητής positive παρουσιάστηκε το εξής πρόβλημα. Επειδή η μεταβλητή αυτή είναι διακριτή και έχει 10 missing values δεν μπορούμε να εφαρμόσουμε κάποιο αντίστοιχο κώδικα με τον παραπάνω, χωρίς να δηλώσουμε πρώτα τα missing values, για αυτό εκτιμούμε αρχικά τις ελλείπουσες τιμές μέσω του προγράμματος mi.categorical που υπάρχει στην έκδοση του R 2.12.1.

Το πρόγραμμα αυτό εκτιμά ελλείπουσες τιμές στην περίπτωση μιας διακριτής μεταβλητής με τον εξής τρόπο, εφαρμόζει ένα loglinear μοντέλο με εξαρτημένη μεταβλητή την διακριτή, στην συγκεκριμένη περίπτωση την positive, και ανεξάρτητες μεταβλητές τις υπόλοιπες μεταβλητές που εξετάσουμε οι οποίες όμως δεν θα πρέπει να περιέχουν missing values, θωρεί ότι οι missing values οφείλονται είτε στην επίδραση κάποιων ανεξάρτητων μεταβλητών είτε ότι έχουν δημιουργηθεί τυχαία. Αφού εγκαταστήσουμε το πρόγραμμα {mi} και εκτελέσουμε τις κατάλληλες εντολές. (Παράρτημα 1), η positive παίρνει τρεις τιμές, την τιμή 1 για αριθμό διηθημένων λεμφαδένων από 0 έως 3, την τιμή 2 για αριθμό διηθημένων λεμφαδένων από 4 έως 8 και την 3 για αριθμό διηθημένων λεμφαδένων μεγαλύτερο ή ίσων του 9.

Περιμένουμε οι καμπύλες επιβίωσης για τις τρεις ομάδες να είναι διατεταγμένες, αφού όπως αναφέραμε στην εισαγωγή του κεφαλαίου η positive σχετίζεται αρνητικά με την επιβίωση των ασθενών, όσο μεγαλύτερος είναι ο αριθμός των διηθημένων λεμφαδένων τόσο μεγαλύτερη είναι και η πιθανότητα επανεμφάνισης ή μετάστασης του καρκίνου, άρα η πρώτη ομάδα που έχει και το μικρότερο αριθμό διηθημένων λεμφαδένων θα έχει και την μεγαλύτερη πιθανότητα επιβίωσης, η καμπύλη της θα βρίσκεται ψηλότερα των υπολοίπων, θα ακολουθεί η $2^{η}$ και στην συνέχεια η $3^{η}$ ομάδα με την χαμηλότερη επιβίωση. Όλα αυτά είναι υποθετικά, για αυτό κατασκευάσουμε το παρακάτω διάγραμμα για να έχουμε μια πρώτη ένδειξη για τα παραπάνω και ύστερα θα εφαρμόσουμε κάποιο τεστ ελέγχου ισότητας συναρτήσεων επιβίωσης και αν έχουμε ενδείξεις διάταξης των συναρτήσεων επιβίωσης θα κάνουμε έναν έλεγχο τάσης (trend test) .

Γράφημα 6.5



Οι καμπύλες επιβίωσης για τις τρεις ομάδες της positive τέμνονται μόνο τους πρώτους μήνες και στην συνέχεια οι αποκλίσεις μεταξύ τους μεγαλώνουν και καθώς προχωράμε προς τα δεξιά οι διαφορές γίνονται πιο έντονες. Εφόσον οι συναρτήσεις επιβίωσης δεν τέμνονται παρά μόνο στην αρχή, για τον έλεγχο ισότητας συναρτήσεων επιβίωσης θα εφαρμόσουμε το Logrank test (LR) που δίνει ίσο βάρος και στα δύο άκρα των πλήρων χρόνων ζωής.

Πίνακας 6.5

Logrank test for newpositive covariate

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
newpositive=1	657	67	97.7	9.63	44.24
newpositive=2	93	19	12.5	3.38	3.77
newpositive=3	120	39	14.8	39.36	44.82
Chisq= 52.6 on 2 degrees of freedom, p= 3.77e-12					

Το δείγμα μας απότελείται από 870 γυναίκες (10 παρατηρήσεις που ήταν ελλιπείς τις αντικαταστήσαμε μέσω του προγράμματος mi προκειμένου να είναι δυνατή η κατηγοριοποίηση της). Στα παραπάνω αποτελέσματα μας δίνεται ο αριθμός των ασθενών (N) για κάθε επίπεδο της μεταβλητής newpositive, η ποσότητα (O-E)^2/E που είναι ουσιαστικά ο κλασικός X^2 έλεγχος, ο οποίος προσεγγίζει την ποσότητα (O-

E)^2/V (δηλ. το στατιστικό $\frac{U^2}{\sqrt{\text{var}(U)}}$ όπως το συμβολίσαμε στο 3^{ο} κεφάλαιο) όταν το μέγεθος του δείγματος είναι αρκετά μεγάλο.

Η τιμή του στατιστικού ισούται με 52.6, για δύο βαθμούς ελευθερίας και αντιστοιχεί τιμή p-value=0.00000000000377, περίπου μηδέν, επομένως σχεδόν για κάθε επίπεδο σημαντικότητας απορρίπτουμε την μηδενική υπόθεση υπέρ της εναλλακτικής. Άρα η συνολική επιβίωση των ασθενών διαφέρει μεταξύ των τριών επιπέδων της μεταβλητής positive.

Ακόμη από το παραπάνω γράφημα υποψιαζόμαστε ότι οι συναρτήσεις επιβίωσης ικανοποιούν την σχέση:

$S_1(t) \geq S_2(t) \geq S_3(t)$, για $t \leq \tau$ με μια τουλάχιστον αυστηρή (>) ανισότητα.

Ο έλεγχος που θα εφαρμόσουμε προκειμένου να ελέγξουμε αν για την συγκεκριμένη μεταβλητή οι συναρτήσεις επιβίωσης είναι διατεταγμένες εναλλακτικές υποθέσεις (δηλ. ικανοποιείται η παραπάνω σχέση των συναρτήσεων επιβίωσης) ονομάζεται έλεγχος τάσης (trend test). Για να εφαρμόσουμε το συγκεκριμένο τεστ στην R στηριζόμαστε στην υπόθεση ότι το LR τεστ είναι ισοδύναμο με ένα Cox model με μια κατηγορική μεταβλητή ως προβλεπτικό παράγοντα. Έτσι για να κάνουμε έναν έλεγχο τάσης θα πρέπει απλώς να προσαρμόσουμε ένα Cox model για μια μεταβλητή η οποία θα έχει ως σκορ το 1,2,3 ή θα πρέπει να δηλώσουμε την μεταβλητή ως factor και να κάνουμε ένα post hoc trend test. Επομένως εκτελούμε τις παρακάτω εντολές:

```
fit1<-coxph(Surv(MONTH,DEATH)~newpositive,data=a); fit1
                coef     exp(coef)  se(coef)     z        p
newpositive 0.68     1.97        0.0994     6.84    8e-12
Likelihood ratio test=40.8  on 1 df, p=1.70e-10  n= 870

 fit2<-coxph(Surv(MONTH,DEATH)~factor(newpositive),data=a); fit2
                                coef exp(coef)   se(coef)     z       p
factor(newpositive)2   0.80      2.23         0.260      3.08   2.1e-03
factor(newpositive)3  1.349      3.86         0.202      6.69  2.3e-11
Likelihood ratio test=41  on 2 df, p=1.23e-09  n= 870
```

```
zz<-c(1,2)
test.num<-zz%*%coef(fit2)
test.var<-zz%*%fit2$var%*%zz
test.num/sqrt(test.var)
      [,1]
[1,] 6.489265
```

Το παραπάνω στατιστικό ακολουθεί προσεγγιστικά μια τυποποιημένη κανονική κατανομή όταν η μηδενική υπόθεση είναι αληθής. Έτσι, θα απορρίπτουμε την μηδενική υπόθεση σε επίπεδο σημαντικότητας α αν ισχύει ότι :

$$\text{test.num/sqrt(test.var)} > z_{\alpha}$$

Η τιμή p-value για το παραπάνω στατιστικό υπολογίζεται ως εξής:

```
pvalue<-2*(1-pnorm(test.num/sqrt(test.var)))
 pvalue
          [,1]
[1,] 8.625634e-11
```

Η τιμή αυτή είναι πολύ κοντά στο μηδέν, σχεδόν για κάθε επίπεδο σημαντικότητας απορρίπτουμε την μηδενική υπόθεση υπέρ της εναλλακτικής. Άρα οι συναρτήσεις επιβίωσης για τις τρεις ομάδες της positive αφορούν πράγματι διατεταγμένες εναλλακτικές υποθέσεις. Έτσι η πιθανότητα συνολικής επιβίωσης διαφοροποιείται σημαντικά μεταξύ των τριών ομάδων της μεταβλητής positive και μάλιστα όσο αυξάνεται ο αριθμός των διηθημένων λεμφαδένων τόσο μικρότερη είναι και η πιθανότητα συνολικής επιβίωσης.

6.3 Εφαρμογή για στρωματοποιημένα τεστ

Όταν θέλουμε να μελετήσουμε τον χρόνο επιβίωσης στα επίπεδα μιας συμμεταβλητής λαμβάνοντας υπόψη και κάποιον άλλον παράγοντα που μπορεί να επηρεάζει τους χρόνους ζωής των ασθενών στις ομάδες, χρησιμοποιούμε τα **στρωματοποιημένα τεστ**. Καθώς ενδιαφερόμαστε να συγκρίνουμε τα επίπεδα της συμμεταβλητής σε κάθε στρώμα του παράγοντα, με τα τεστ αυτά μας δίνεται η δυνατότητα να απομονώσουμε κάποια ομάδα, ακόμη στα τεστ αυτά επιτρέπουμε στις

συναρτήσεις επιβίωσης να διαφέρουν ανάμεσα στα στρώματα όμως μέσα στο στρώμα θεωρούμε ότι είναι ίσες.

Για να επιλέξουμε το σετ των δεδομένων που αντιστοιχούν στο επίπεδο της συμμεταβλητής που μας ενδιαφέρει χρησιμοποιούμε την παράμετρο **subset.**

6.3.1 Στρωματοποίηση της συμμεταβλητής grade για τα τρία επίπεδα της newage

Θα επαναλάβουμε στην συνέχεια κάποια από τα τεστ που παρουσιάσαμε παραπάνω με την μόνη διαφορά ότι τώρα επιδιώκουμε να διαπιστώσουμε το πώς συμπεριφέρεται ο παράγοντας που δηλώνει την διαφοροποίηση του όγκου ξεχωριστά για την κάθε ομάδα ηλικιών. Αν με i=(1,2,3,4) δηλώνουμε την μεταβλητή grade και με l=(1,2,3) την μεταβλητή New_age που δηλώνει τα στρώματα (L = 3), και με $S_{il}(t)$ συμβολίσουμε την συνάρτηση επιβίωσης που αντιστοιχεί στις ασθενείς της ομάδας i που ανήκουν στο l στρώμα. Τότε η υπόθεση που θέλουμε να ελέγξουμε είναι ότι στο l στρώμα δεν υπάρχει διαφορά μεταξύ των συναρτήσεων επιβίωσης των τεσσάρων ομάδων, και η μορφή της θα είναι η εξής:

$$H_0: S_{1l}(t)=S_{2l}(t)=S_{3l}(t)=S_{4l}(t)\ ,\quad t\leq\tau$$

Για να εκτελέσουμε το παραπάνω τεστ στην R θα υπολογίσουμε αρχικά τις συναρτήσεις επιβίωσης για κάθε επίπεδο της συμμεταβλητής newage, θα τις απεικονίσουμε σε ένα κοινό γράφημα και θα εφαρμόσουμε κάποιο έλεγχο ισότητας συναρτήσεων επιβίωσης, αφού ενδιαφερόμαστε να συγκρίνουμε την επιβίωση των ασθενών μεταξύ των επιπέδων της μεταβλητής grade για κάθε επίπεδο της συμμεταβλητής newage. Ουσιαστικά μας ενδιαφέρει ο τοπικός έλεγχος της υπόθεσης:

$H_0(t): S_{1l}(t) = S_{2l}(t) = S_{3l}(t) = S_{4l}(t)$ έναντι

$H_1(t): S_{il}(t) \neq S_{jl}(t)$ για κάποια $1 \leq i, j \leq 4$ και για το $l\ (1 \leq l \leq 3)$ στρώμα της *newage*

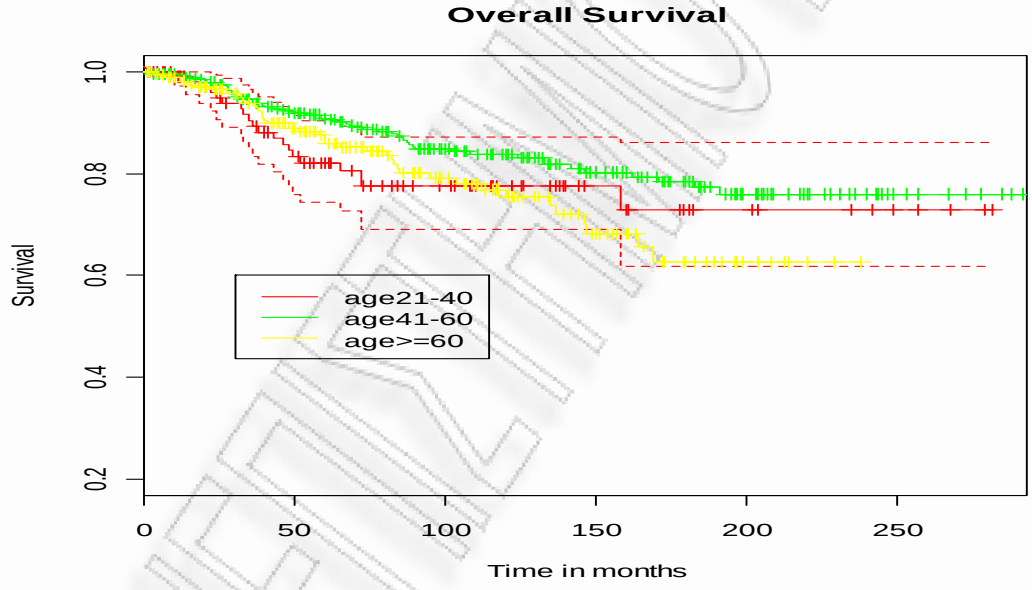
Εκτιμούμε τις συναρτήσεις επιβίωσης για κάθε επίπεδο της newage μέσω των εντολών:

```
fitage1<-survfit(Surv(MONTH,DEATH)~1,subset=NEWAGE==1,data=a)
fitage2<-survfit(Surv(MONTH,DEATH)~1,subset=NEWAGE==2,data=a)
fitage3<-survfit(Surv(MONTH,DEATH)~1,subset=NEWAGE==3,data=a)
```

Χρησιμοποώντας τις εντολές που ακολουθούν απεικονίζουμε γραφικά τις τρεις παραπάνω εκτιμήσεις:

```
plot(fitage1,xlab="Time in months",ylab="Survival",main="Overall Survival",
ylim=c(0.2,1), lty=1,col="red")
lines(fitage2,lty=1,col="green")
lines(fitage3,lty=1,col="yellow")
legend(30,0.6,c("age21-40","age41-60","age>=60"),
lty=1:1, col=c("red","green","yellow"))
```

Γράφημα 6.6



Εφόσον οι παραπάνω συναρτήσεις επιβίωσης τέμνονται θα εφαρμόσουμε το τεστ των Gehan – Breslow προκειμένου να ελέγξουμε αν οι συναρτήσεις επιβίωσης σε κάθε ($1 \leq l \leq 3$) στρώμα είναι ίσες.

Τα αποτελέσματα που παίρνουμε είναι:

Πίνακας 6.6

Στρωματοποιημένο τεστ της grade στα επίπεδα της newage

New_age		Chi-Square	df	Sig.
1,00	Breslow (Generalized Wilcoxon)	13,434	3	,004
2,00	Breslow (Generalized Wilcoxon)	23,111	3	,000
3,00	Breslow (Generalized Wilcoxon)	3,225	3	,358

Test of equality of survival distributions for the different levels of GRADE.

Σύμφωνα με τα αποτελέσματα των παραπάνω ελέγχων απορρίπτουμε την μηδενική υπόθεση για τα δύο πρώτα στρώματα, αφού έχουμε τιμή p – value μικρότερη του επιπέδου σημαντικότητας. Συμπεραίνουμε λοιπόν, ότι οι συναρτήσεις επιβίωσης για τα τέσσερα επίπεδα της μεταβλητής GRADE διαφέρουν όταν μελετάμε μόνο τις ασθενείς ηλικίας (21-40) ή τις ασθενείς ηλικίας (41-60).

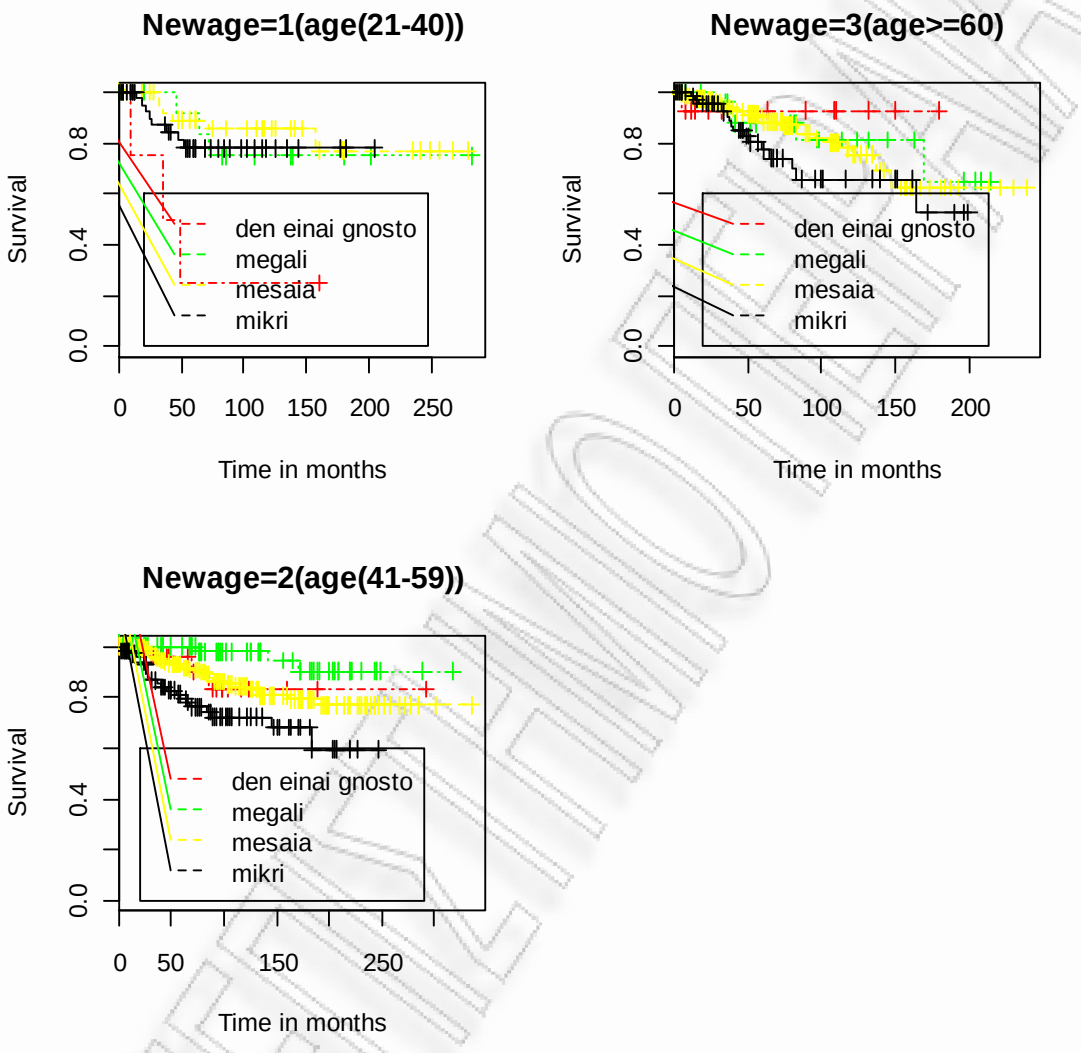
Ενώ για τις ασθενείς ηλικίας >= 60 χρονών οι συναρτήσεις επιβίωσης για τα τέσσερα επίπεδα της GRADE δεν διαφέρουν, αφού σε ε.σ 5% δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση υπέρ της εναλλακτικής εφόσον η τιμή p – value του ελέγχου ισούται με $0.358 > 0.05$.

(Για τους πίνακες επιβίωσης των ασθενών για κάθε ομάδα της newage για τα τέσσερα επίπεδα της GRADE βλέπε παράρτημα 1)

Οι τοπικοί στρωματοποιημένοι έλεγχοι για κάθε επίπεδο της newage αποδεικνύονται και γραφικά (για την κατασκευή του παρακάτω γραφήματος βλέπε παράρτημα) :

Γράφημα 6.7

Γραφήματα τοπικών ελέγχων στρωματοποίησης της grade για κάθε επίπεδο της newage



Όπως περιμένουμε για τα δύο πρώτα στρώματα τουλάχιστον μια από τις τέσσερεις συναρτήσεις επιβίωσης θα έπρεπε να διαφέρει των υπόλοιπων τριών, γεγονός που επιβεβαιώνεται από τα παραπάνω γραφήματα. Απ' όπου μπορούμε να δούμε ότι η καμπύλη επιβίωσης για την ομάδα που δεν είναι γνωστή η διαφοροποίηση του όγκου (Grade=0) δεν τέμνει τις υπόλοιπες τρεις καμπύλες επιβίωσης και βρίσκεται χαμηλότερα από αυτές. Περιμένουμε λοιπόν γυναίκες ηλικίας 21 – 40 για τις οποίες δεν είναι γνωστή η διαφοροποίηση του όγκου να πεθάνουν νωρίτερα έναντι γυναικών μεγαλύτερης ηλικίας, ακόμα για την συγκεκριμένη ομάδα όλες οι παρατηρήσεις είναι πλήρεις χρόνοι, όλες έχουν αποβιώσει στο υπό μελέτη χρονικό διάστημα. Για το

στρώμα που έχουμε γυναίκες ηλικίας 41 – 60 χρονών βλέπουμε ότι η συνάρτηση επιβίωσης που διαφέρει των υπολοίπων είναι αυτή που αφορά ασθενείς στις οποίες παρατηρήθηκε μικρή διαφοροποίηση του όγκου. Και στην περίπτωση αυτή η καμπύλη επιβίωσης κινείται σε χαμηλότερο επίπεδο έναντι των υπολοίπων.

Ενώ για το τελευταίο στρώμα οι συναρτήσεις επιβίωσης για τις τέσσερεις ομάδες του παράγοντα grade τέμνονται και ακολουθούν σχεδόν την ίδια πορεία, οπότε και γραφικά έχουμε ενδείξεις ισότητας των συναρτήσεων επιβίωσης για το συγκεκριμένο στρώμα. Επομένως γυναίκες άνω των 61 χρονών παρουσιάζουν την ίδια συμπεριφορά ως προς την εκτιμώμενη επιβίωση και για τις τέσσερεις κατηγορίες του παράγοντα grade.

Τέλος, εφαρμόσουμε τον ολικό στρωματοποιημένο έλεγχο προκειμένου να ελέγξουμε αν οι συναρτήσεις επιβίωσης διαφέρουν στα στρώματα λαμβάνοντας υπόψη την μεταβλητή newage. Η υπόθεση που ελέγχουμε είναι η :

$$H_0 : S_{1l}(t) = S_{2l}(t) = S_{3l}(t) = S_{4l}(t),\ t \le \tau,\ 1 \le l \le 3$$

Πίνακας 6.7

Ολικός έλεγχος για τα 4 επίπεδα της grade λαμβάνοντας υπόψη την newage

	Chi-Square	df	Sig.
Breslow (Generalized Wilcoxon)	25,954	3	,000

Test of equality of survival distributions for the different levels of GRADE.
a Adjusted for New_age.

Έχουμε τιμή p – value μικρότερη του 5% ε.σ, επομένως απορρίπτουμε την μηδενική υπόθεση υπέρ της εναλλακτικής της, έτσι οι συναρτήσεις επιβίωσης των τεσσάρων ομάδων λαμβάνοντας υπόψη την μεταβλητή newage διαφέρουν.

Παραπάνω είδαμε ότι για τα δύο πρώτα στρώματα οι τέσσερεις συναρτήσεις επιβίωσης διαφέρουν ενώ για το τρίτο όχι. Ακόμα και όταν εξετάσαμε την ισότητα των συναρτήσεων επιβίωσης για τις τέσσερεις κατηγορίες του παράγοντα grade χωρίς να λάβουμε υπόψη την μεταβλητή New_age είδαμε ότι αυτές διαφέρουν. Επομένως μπορούμε να πούμε ότι ο ολικός έλεγχος καταγράφει αυτή την διαφορά, βέβαια υπάρχουν και αρκετές περιπτώσεις που οι συναρτήσεις επιβίωσης διαφέρουν στα στρώματα και ο ολικός έλεγχος δεν καταγράφει αυτή την διαφορά.

6.3.2 Στρωματοποίηση της συμμεταβλητής newpositive στα επίπεδα της newsize

Έστω ότι τώρα θέλουμε να μελετήσουμε τον χρόνο επιβίωσης των ασθενών στα επίπεδα της μεταβλητής newsize λαμβάνοντας υπόψη την newpositive, ενδιαφερόμαστε λοιπόν να συγκρίνουμε τα επίπεδα της newsize σε κάθε στρώμα της newpositive, κάνουμε το εξής τοπικό έλεγχο υπόθεσης:

$H_0 : S_{1h}(t) = S_{2h}(t), t \le \tau$ *έναντι* $H_1 : S_{1h}(t) \ne S_{2h}(t) \ne S_{3h}(t), t \le \tau$ *για το* h $(1 \le h \le 3)$ στρώμα της newpositive.

Πίνακας 6.8

Breslow- Gehan test της newsize στα επίπεδα της newpositive

newpositive		Chi-Square	df	Sig.
1,00	Breslow (Generalized Wilcoxon)	10,222	2	,006
2,00	Breslow (Generalized Wilcoxon)	,733	2	,693
3,00	Breslow (Generalized Wilcoxon)	,465	2	,793

Test of equality of survival distributions for the different levels of new_size.

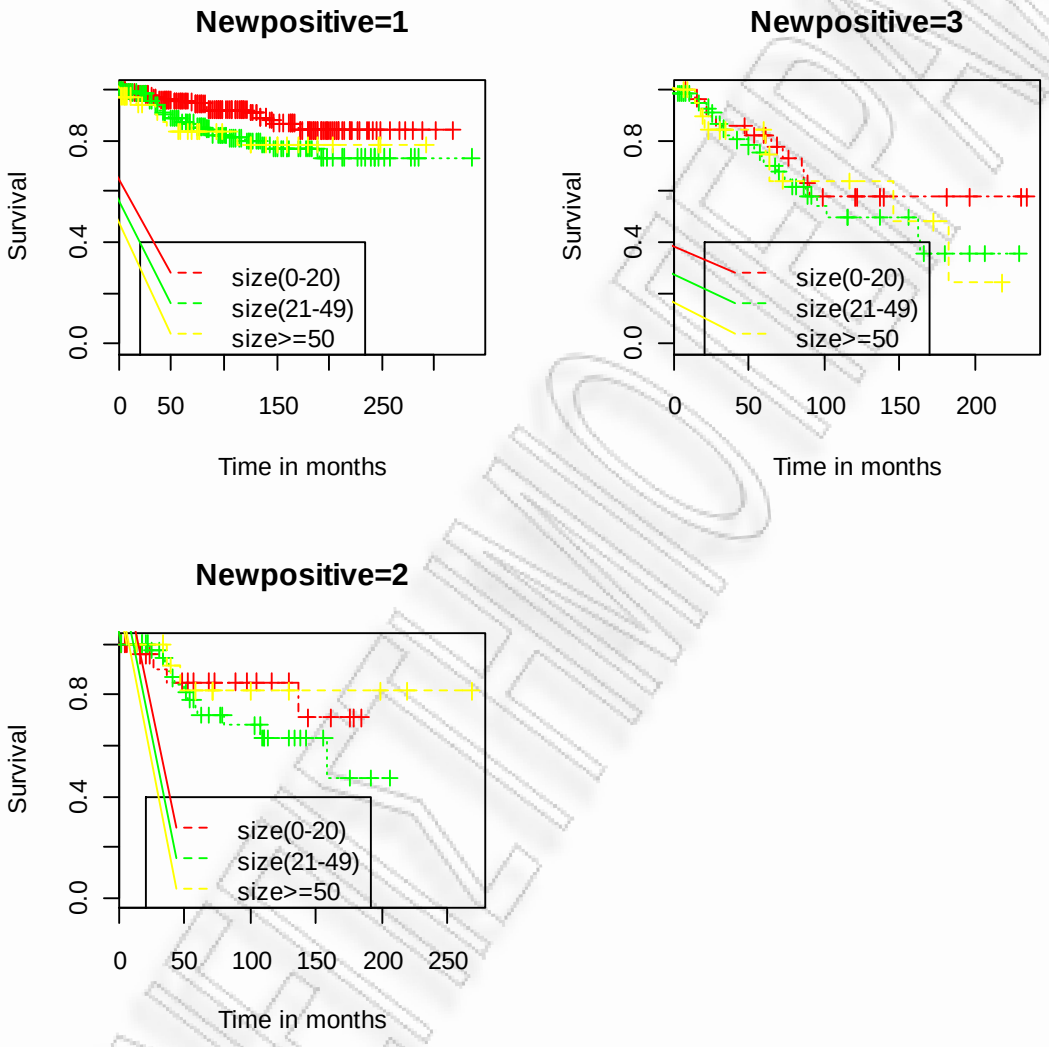
Μόνο στο $1^{ο}$ στρώμα απορρίπτουμε την μηδενική υπόθεση υπέρ της εναλλακτικής, αφού το p-value=0.006 μικρότερο του επιπέδου σημαντικότητας 5% , ενώ για τα δύο άλλα στρώματα αποδεχόμαστε την μηδενική υπόθεση αφού το p-value είναι σημαντικά μεγαλύτερο του ε.σ.. Συμπεραίνουμε λοιπόν ότι η πιθανότητα συνολικής επιβίωσης των ασθενών διαφέρει σημαντικά για τα τρία επίπεδα της μεταβλητής newsize όταν ο αριθμός των διηθημένων λεμφαδένων είναι από (0 – 3), ενώ δεν διαφέρει όταν ο αριθμός των διηθημένων λεμφαδένων είναι από >=4.

(Για τους πίνακες επιβίωσης των ασθενών στα τρία επίπεδα της newsize για κάθε επίπεδο της newpositive βλέπε παράρτημα)

Γραφικά από τους στρωματοποιημένους τοπικούς ελέγχους έχουμε:

Γράφημα 6.8

Γραφήματα τοπικών ελέγχων στρωματοποίησης της newpositive για κάθε επίπεδο της newsize



Από τα παραπάνω γραφήματα έχουμε ενδείξης ότι οι συναρτήσεις επιβίωσης των τρίων ομάδων διαφέρουν στο στρώμα newpositive=1, ενώ δεν φαίνεται να διαφέρουν στα δύο άλλα στρώματα.

Αφού οι καμπύλες τέμνονται και στα τρία γραφήματα εφαρμόσαμε τον έλεγχο των Breslow – Gehan.

Στην συνέχεια για τον ολικό στρωματοποιημένο έλεγχο δεν συγκρίνουμε τις συναρτήσεις επιβίωσης των τριών ομάδων για κάθε επίπεδο τις newpositive αλλά λαμβάνοντας υπόψη την newpositive.

Πίνακας 6.9

Ολικός έλεγχος για τα 3 επίπεδα της newsize λαμβάνοντας υπόψη την newpositive

	Chi-Square	df	Sig.
Breslow (Generalized Wilcoxon)	10,586	2	,005

Test of equality of survival distributions for the different levels of new_size.
a Adjusted for newpositive.

Με βάση την τιμή p-value συμπεραίνουμε ότι λαμβάνοντας υπόψη την newpositive οι συναρτήσεις επιβίωσης των τριών ομάδων διαφέρουν σημαντικά. Στους τοπικούς ελέγχους όμως δεχτήκαμε ότι οι συναρτήσεις επιβίωσης δεν διαφέρουν για newpositive=2 και 3 ενώ διαφέρουν για την newpositive=1, δεν είναι αναγκαίο όμως τα αποτελέσματα των τοπικών ελέγχων να συμπίπτουν με αυτά του ολικού ελέγχου, καθώς όπως είναι γνωστό σε πολλές περιπτώσεις οι συναρτήσεις επιβίωσης διαφέρουν στα επίπεδα του παράγοντα αλλά ο ολικός έλεγχος δεν καταγράφει αυτή την διαφορά.

Κεφάλαιο 7^{ο}

Ανάλυση δεδομένων με χρήση του μοντέλου αναλογικού κινδύνου του Cox

7.1 Εισαγωγή

Στο συγκεκριμένο κεφάλαιο επιδιώκουμε να συγκρίνουμε συναρτήσεις επιβίωσης στην περίπτωση που υπάρχουν περισσότερες από μια συμμεταβλητές, οπότε και δεν μπορούμε να εφαρμόσουμε τις σχέσεις που περιγράψαμε στο προηγούμενο κεφάλαιο, ακόμα παρουσιάζεται η ανάγκη να διαπιστώσουμε την σχέση μεταξύ της μεταβλητής που δηλώνει τον χρόνο ζωής και μιας ή περισσότερων συμμεταβλητών. Χρησιμοποιούμε για το σκοπό αυτό ένα μοντέλο παλινδρόμησης, το γνωστό μοντέλο παλινδρόμησης του Cox που έχει ευρεία εφαρμογή σε δεδομένα ανάλυσης χρόνου επιβίωσης. Όπως έχουμε αναφέρει και στην θεωρία του τέταρτου κεφαλαίου, προσπαθούμε να καθορίσουμε τις μεταβλητές που επηρεάζουν την συνάρτηση κινδύνου, να την εκτιμήσουμε και μέσω της εκτίμησης αυτής να οδηγηθούμε στην εκτίμηση της συνάρτησης επιβίωσης. Στην συνέχεια θα εφαρμόσουμε κάποια από τις τεχνικές επιλογής μεταβλητών προκειμένου να καταλήξουμε σε κάποιο βέλτιστο μοντέλο για τα δεδομένα μας . Τέλος, θα αξιολογήσουμε την υπόθεση αναλογικού κινδύνου τόσο γραφικά όσο και μέσω στατιστικών ελέγχων.

7.2 Εφαρμογή του μοντέλου αναλογικού κινδύνου του Cox

Για να εφαρμόσουμε το μοντέλο παλινδρόμησης του Cox χρησιμοποιούμε την εντολή **coxph().** Έτσι αν θέλουμε να ελέγξουμε αν η μεταβλητή Grade επηρεάζει και πόσο την συνάρτηση κινδύνου ή αντίστοιχα την συνάρτηση επιβίωσης, εφαρμόζουμε την παρακάτω εντολή:

```
cox1<-coxph(Surv(MONTH,DEATH)~factor(GRADE),method="breslow",data=a)
summary(cox1)
 n=842 (28 observations deleted due to missingness)
```

Πίνακας 7.1

Συντελεστές παλινδρόμησης του μοντέλου (7.1)

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
factor(GRADE)1	-0.7462	0.4742	0.4841	-1.541	0.123
factor(GRADE)2	-0.1750	0.8394	0.3996	-0.438	0.661
factor(GRADE)3	0.4669	1.5950	0.4054	1.152.	0.249
	exp(coef)	exp(-coef)	lower .95	upper .95	
factor(GRADE)1	0.4742	2.109	0.1836	1.225	
factor(GRADE)2	0.8394	1.191	0.3836	1.837	
factor(GRADE)3	1.5950	0.627	0.7206	3.530	
Rsquare= 0.021 (max possible= 0.836)					
Likelihood ratio test= 18.22 on 3 df, p=0.0003957					
Wald test = 17.89 on 3 df, p=0.000464					
Score (logrank) test = 19.01 on 3 df, p=0.0002720					

Στο παραπάνω μοντέλο η μεταβλητή που δηλώνει την διαφοροποίηση του όγκου είναι κατηγορική, παίρνει τέσσερεις τιμές, και αφού την δηλώσουμε ως factor το πρώτο επίπεδο (συνήθως) μπαίνει ως επίπεδο αναφοράς (reference category). Με βάση τα παραπάνω αποτελέσματα η μεταβλητή Grade ή τουλάχιστον μια από τις κατηγορίες της φαίνεται να επηρεάζει την συνάρτηση συνολικού κινδύνου, αφού οι ολικοί έλεγχοι (Likelihood ratio test, Wald test, Score test) έχουν τιμή p – value πολύ μικρότερη του επιπέδου σημαντικότητας α=5%, με αποτέλεσμα να απορρίπτουμε την μηδενική υπόθεση $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, για το μοντέλο:

$$h(t \mid GRADE) = h_0(t)\exp(\beta_1 grade1 + \beta_2 grade2 + \beta_3 grade3), \ t \geq 0 \qquad (7.1)$$

Όμως για τους τοπικούς ελέγχους $H_0 : \beta_1 = 0$, *ή* $H_0 : \beta_2 = 0$, *ή* $H_0 : \beta_3 = 0$ δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση, εφόσον το Wald test για την κάθε κατηγορία της Grade έχει τιμή p – value αρκετά μεγαλύτερη του επιπέδου σημαντικότητας 5%, επίσης και το 95% διάστημα εμπιστοσύνης για την παράμετρο $\exp(\beta_i)$ περιέχει την τιμή 1. Τα συμπεράσματα αυτά έρχονται σε αντίθεση με το αποτέλεσμα των ολικών ελέγχων, για τους οποίους οι όροι αυτοί ήταν σημαντικοί.

Αν θέλαμε να υπολογίσουμε τον κίνδυνο θανάτου ή μετάστασης του καρκίνου καθώς μετακινούμαστε από μια άγνωστη διαφοροποίηση του όγκου προς μια μεγάλη διαφοροποίηση, αυτός θα ισούται με :

$$HR(t) = \frac{\hat{h}(t \mid grade = 1)}{\hat{h}(t \mid grade = 0)} = e^{\hat{\beta}_1} = 0.4742$$

δηλαδή ασθενείς με άγνωστη διαφοροποίηση του όγκου έχουν 2.11 φορές μεγαλύτερο κίνδυνο θανάτου ή μετάστασης του καρκίνου από τους ασθενείς με μεγάλη διαφοροποίηση του όγκου.

Αντίστοιχα ο κίνδυνος θανάτου ασθενών με μεγάλη διαφοροποίηση του όγκου σε σχέση με ασθενείς με μεσαία διαφοροποίηση θα ισούται με

$$HR(t) = \frac{\hat{h}(t \mid grade = 2)}{\hat{h}(t \mid grade = 1)} = \mathrm{e}(\hat{\beta}_2 - \hat{\beta}_1) = e^{0.5712} = 1.77$$

δηλ. ασθενείς με μεσαία διαφοροποίηση όγκου έχουν 1.77 φορές μεγαλύτερο κίνδυνο θανάτου από ότι ασθενείς με μεγάλη διαφοροποίηση.

Για ασθενείς με χαμηλή διαφοροποίηση του όγκου ο κίνδυνος θανάτου θα είναι

$$HR(t) = e^{\hat{\beta}_3 - \hat{\beta}_2} = 1.90$$

1.90 φορές μεγαλύτερος σε σχέση με ασθενείς με μεσαία διαφοροποίηση.

Ασθενείς με χαμηλή διαφοροποίηση του όγκου έχουν τριπλάσιο κίνδυνο θανάτου

$$HR(t) = e^{\hat{\beta}_3 - \hat{\beta}_1} = 3.36$$

από τους ασθενείς με μεγάλη διαφοροποίηση.

Συμπεραίνουμε, όπως και περιμέναμε, ότι ασθενείς με μεγάλη διαφοροποίηση του όγκου, δηλ. μικρότερη συρρίκνωση, έχουν μικρότερο κίνδυνο θανάτου σε σχέση με τους ασθενείς που ανήκουν στις άλλες κατηγορίες της συμμεταβλητής Grade. Χειρότεροι πρόγνωση έχνουν οι ασθενείς με χαμηλή διαφοροποίηση του όγκου έναντι όλων των επιπέδων τις grade.

Για την μεταβλητή newage μελετάμε το παρακάτω μοντέλο αναλογικού κινδύνου:

$$h(t \mid Z) = h_0(t)\exp(\beta_1 newage1 + \beta_2 newage2 + \beta_3 newage3) \qquad (7.2)$$

Πίνακας 7.2

Συντελεστές παλινδρόμησης του μοντέλου (7.2)

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
factor(NEWAGE)2	-0.41330	0.66146	0.25960	-1.592	0.111
factor(NEWAGE)3	0.01374	1.01383	0.26716	0.051	0.959
	exp(coef)	exp(-coef)	lower .95	upper .95	
factor(NEWAGE)2	0.6615	1.5118	0.3977	1.100	
factor(NEWAGE)3	1.0138	0.9864	0.6006	1.711	
Rsquare= 0.006 (max possible= 0.827)					
Likelihood ratio test= 5.52 on 2 df, p=0.0632					
Wald test = 5.51 on 2 df, p=0.0637					
Score (logrank) test = 5.59 on 2 df, p=0.0612					

Με βάση τους ολικούς ελέγχους (LR, Wald test, score test) δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση $H_0 : \beta_2 = \beta_3 = 0$ για επίπεδο σημαντικότητας α=5% αφού η τιμή p – value για το κάθε τεστ είναι μεγαλύτερη του 0.05, άρα η μεταβλητή newage δεν επηρεάζει την συνάρτηση συνολικού κινδύνου. Τα συμπεράσματα αυτά συμφωνούν και με τα αποτελέσματα των τοπικών ελέγχων, σύμφωνα με τα οποία τα δύο επίπεδα της μεταβλητής newage δεν είναι στατιστικά σημαντικά για α=5%.

Αν θέλαμε να υπολογίσουμε τον κίνδυνο θανάτου για ασθενείς ηλικίας 41-60 σε σχέση με ασθενείς ηλικίας 21-40, αυτός θα ισούταν με:

$$HR(t) = \frac{\hat{h}(t \mid newage = 2)}{\hat{h}(t \mid newage = 1)} = e^{\hat{\beta}_2} = 0.66146$$

δηλ. ασθενείς ηλικίας 21-40 χρονών έχουν 1.51 φορές μεγαλύτερο κίνδυνο θανάτου από ασθενείς ηλικίας 41-60 χρονών.

Αντίστοιχα ο κίνδυνος θανάτου για ασθενείς ηλικίας >= 60 χρονών είναι

$$HR(t) = \frac{\hat{h}(t \mid newage = 3)}{\hat{h}(t \mid newage = 2)} = e^{(\hat{\beta}_3 - \hat{\beta}_2)} = 1.53$$

φορές μεγαλύτερος από ότι για ασθενείς ηλικίας 41-60 χρονών. Παρόμοια μπορούμε να υπολογίσουμε και άλλους λόγους κινδύνων.

Αντίστοιχα για την μεταβλητή size εφαρμόζουμε το μοντέλο:

$$h(t \mid Z) = h_0(t)\exp(\beta \cdot size) \qquad (7.3)$$

Πίνακας 7.3

Συντελεστές παλινδρόμησης του μοντέλου (7.3)

coef exp(coef) se(coef) z Pr(>|z|) exp(-coef) lower .95 upper .95
SIZE 0.022119 1.022365 0.004608 4.8 1.59e-06 *** 0.9781 1.013 1.032
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Rsquare= 0.022 (max possible= 0.827)
Likelihood ratio test = 19.13 on 1 df, p=1.222e-05
Wald test = 23.04 on 1 df, p=1.586e-06
Score (logrank) test = 23.26 on 1 df, p=1.417e-06

Από τα αποτελέσματα για την μεταβλητή size, βλέπουμε πως η μεταβλητή αυτή επηρεάζει την συνάρτηση συνολικού κινδύνου, διότι και τα τρία p – value για τους ολικούς ελέγχους είναι πολύ μικρότερα του επιπέδου σημαντικότητας, α=5%. Επομένως απορρίπτουμε την μηδενική υπόθεση $H_0 : \beta = 0$ για το μοντέλο (7.3).

Ακόμη, παρατηρώντας το πρόσημο του συντελεστή παλινδρόμησης $\hat{\beta}$ συμπεραίνουμε ότι το μέγεθος του όγκου επηρεάζει θετικά τον κίνδυνο θνησιμότητας, ακόμη εφόσον η τιμή αυτού του συντελεστή είναι μεγαλύτερη του μηδενός, αν αυξήσουμε την μεταβλητή size κατά 1mm περιμένουμε να αυξάνεται σημαντικά ο κίνδυνος θανάτου της ασθενούς. Πράγματι

$$\frac{h(t \mid size+1)}{h(t \mid size)} = e^{\hat{\beta}} = 1.022 > 1$$

όσο μεγαλύτερο είναι το μέγεθος του όγκου τόσο μεγαλύτερο κίνδυνο διατρέχει η ασθενής.

Αντίστοιχα είναι και τα συμπεράσματα για την μεταβλητή που δηλώνει τον αριθμό των διηθημένων λεμφαδένων (βλέπε παράρτημα) .

Έστω τώρα ότι θέλουμε να μελετήσουμε την επίδραση της διαφοροποίησης του όγκου στη συνάρτηση κινδύνου λαμβάνοντας υπόψη και την επίδραση του μεγέθους του όγκου.

Μελετάμε τώρα το παρακάτω μοντέλο παλινδρόμησης στο οποίο έχουμε συμπεριλάβει δύο μεταβλητές την grade και την size :

$$h(t \mid Z) = h_0(t)\exp(\beta_1 grade1 + \beta_2 grade2 + \beta_3 grade3 + \beta_4 size),\ t \geq 0,$$
$$Z = (\,factor(grade), size) \qquad (7.4)$$

Πίνακας 7.4

Συντελεστές παλινδρόμησης για το μοντέλο (7.4)

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
factor(grade)1	-0.452271	0.636182	0.492333	-0.919	0.358290
factor(grade)2	0.029403	1.029839	0.404946	0.073	0.942117
factor(grade)3	0.520049	1.682109	0.406043	1.281	0.200273
size	0.017953	1.018115	0.004969	3.613	0.000302 ***
	exp(coef)	exp(-coef)	lower .95	upper .95	
factor(grade)1	0.6362	1.5719	0.2424	1.670	
factor(grade)2	1.0298	0.9710	0.4657	2.278	
factor(grade)3	1.6821	0.5945	0.7590	3.728	
size	1.0181	0.9822	1.0082	1.028	
Rsquare= 0.035 (max possible= 0.836)					
Likelihood ratio test = 29.63 on 4 df, p=5.83e-06					
Wald test = 31.96 on 4 df, p=1.946e-06					
Score (logrank) test = 33.43 on 4 df, p=9.743e-07					

Από τις τιμές των τριών p – value για τους ολικούς ελέγχους απορρίπτουμε την μηδενική υπόθεση $H_0 : (\beta_1, \beta_2)' = 0$ σχεδόν για κάθε επίπεδο σημαντικότητας, τουλάχιστον μια από τις συμμεταβλητές του μοντέλου επηρεάζει την συνάρτηση κινδύνου. Από τους τοπικούς ελέγχους βλέπουμε πως μόνο το μέγεθος του όγκου (Wald p – value=0.000302 < 0.05) επηρεάζει σημαντικά την συνάρτηση κινδύνου. Σύμφωνα με το σχετικό λόγο κινδύνων υπάρχει η τάση να μειώνεται ο κίνδυνος θανάτου στις ασθενείς με μεγάλη διαφοροποίηση του όγκου σε σχέση με τις ασθενείς με άγνωστη διαφοροποίηση, αλλά η τάση αυτή δεν είναι στατιστικά σημαντική

(HR=0.64, 95% CI: 0.24 – 1.67, Wald p-value= 0.358). Αντίθετα είναι τα συμπεράσματα για τις ασθενείς με μέτρια διαφοροποίηση και για τις ασθενείς με χαμηλή διαφοροποίηση, για τις οποίες παρατηρείται η τάση να αυξάνεται ο κίνδυνος θανάτου σε σχέση με τις ασθενείς με άγνωστη διαφοροποίηση, η τάση αυτή και πάλι δεν είναι (στατιστικά) σημαντική (HR=1.03, CI: 0.46 – 2.28, Wald p-value=0.942, και HR=1.68, CI: 0.75 – 3.73, Wald p-value=0.2).

Στην περίπτωση που λάβουμε υπόψη και το πώς επηρεάζει η διαφοροποίηση του όγκου την συνάρτηση κινδύνου για κάθε μεταβολή του μεγέθους του όγκου, τότε θα πρέπει να μελετήσουμε το μοντέλο παλινδρόμησης του Cox με όρο αλληλεπίδρασης μεταξύ του grade και size:

$$h(t \mid Z) = h_0(t)\exp(\beta_1 grade1 + \beta_2 grade2 + \beta_3 grade3 + \beta_4 size + \beta_5 grade1 : size + \beta_6 grade2 : size + \beta_7 grade3 : size) \quad t \geq 0, \qquad (7.5)$$
$$Z = [\, factor(grade), size, factor(grade) : size]$$

Πίνακας 7.5

Συντελεστές παλινδρόμησης για το μοντέλο (7.5)

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
factor(grade)1	-1.42315	0.24095	0.94794	-1.501	0.1333
factor(grade)2	-1.33095	0.26423	0.81679	-1.629	0.1032
factor(grade)3	-0.55897	0.57180	0.83092	-0.673	0.5011
size	-0.01577	0.98436	0.02381	-0.662	0.5079
factor(grade)1:size	0.02580	1.02613	0.03256	0.792	0.4282
factor(grade)2:size	0.04172	1.04260	0.02496	1.671	0.0947
factor(grade)3:size	0.03220	1.03272	0.02489	1.294	0.1958
	exp(coef)	exp(-coef)	lower .95	upper .95	
factor(grade)1	0.2410	4.1502	0.03759	1.545	
factor(grade)2	0.2642	3.7846	0.05330	1.310	
factor(grade)3	0.5718	1.7489	0.11219	2.914	
size	0.9844	1.0159	0.93947	1.031	
factor(grade)1:size	1.0261	0.9745	0.96269	1.094	

factor(grade)2:size	1.0426	0.9591	0.99281	1.095
factor(grade)3:size	1.0327	0.9683	0.98355	1.084
Rsquare= 0.039 (max possible= 0.836)				
Likelihood ratio test = 33.36 on 7 df, p=2.268e-05				
Wald test = 35.62 on 7 df, p=8.535e-06				
Score (logrank) test = 38.83 on 7 df, p=2.108e-06				

Με βάση τους ολικούς ελέγχους απορρίπτουμε την μηδενική $H_0 : (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7)' = 0$, ενώ οι τοπικοί έλεγχοι αποδέχονται την μηδενική υπόθεση, επομένως ούτε η μεταβλητή grade ούτε η size, άλλα ούτε και η αλληλεπίδρασή τους φαίνεται να επηρεάζει την συνάρτηση κινδύνου. Από τους ολικούς ελέγχους προκύπτει ότι τουλάχιστον μια από τις συμμεταβλητές επηρεάζει την συνάρτηση κινδύνου ενώ σύμφωνα με τους τοπικούς ελέγχους οι όροι αυτοί είναι στατιστικά ασήμαντοι, αυτό αποτελεί ένδειξη πιθανής συσχέτισης μεταξύ τους.

Η γραφική απεικόνιση του παραπάνω μοντέλου δεν θα μας έδινε κάποια ξεκάθαρη εικόνα αφού η μεταβλητή που δηλώνει το μέγεθος του όγκου είναι συνεχής, θα είχαμε απλώς ένα γράφημα με πολλές γραμμές σε τέσσερα διαφορετικά χρώματα ένα για το κάθε επίπεδο της grade.

Ενδεικτικά υπολογίζουμε τον κίνδυνο θανάτου για μια γυναίκα με μέγεθος όγκου έστω c, στην οποία παρατηρήθηκε μεγάλη διαφοροποίηση του όγκου σε σχέση με μια γυναίκα με ίδιο μέγεθος όγκου άλλα που παρουσίασε μεσαία διαφοροποίηση.

Συγκεκριμένα, έχουμε

$$HR(t) = \frac{\hat{h}(t \mid grade = 1, size = c, grade1 : size)}{\hat{h}(t \mid grade = 2, size = c, grade2 : size)} = e^{(\hat{\beta}_1 + \hat{\beta}_5 - \beta_2 - \beta_6)} = 0.8975$$

Δηλαδή, ο κίνδυνος θανάτου για μια γυναίκα που έχει size c και grade1 είναι 0.90

φορές μικρότερος σε σχέση με τον κίνδυνο θανάτου για μια γυναίκα που έχει το ίδιο size, αλλά grade2.

7.3 Εφαρμογή τεχνικών επιλογής μεταβλητών

Προκειμένου να διαπιστώσουμε ποιες από τις υπό εξέταση μεταβλητές είναι πράγματι σημαντικές θα εφαρμόσουμε μια από τις τεχνικές επιλογής μεταβλητών που περιγράψαμε στην θεωρία (Κεφ. $4^{ο}$ παράγ. 4.10.1).

Σκοπός μας είναι να εντοπίσουμε το καλύτερο μοντέλο για τα δεδομένα μας, που αφορούν 870 γυναίκες που εμφάνισαν καρκίνο σε κάποιο ή και στους δύο μαστούς και υποβλήθηκαν σε εγχείρηση, μας ενδιαφέρει να δούμε ποιες μεταβλητές μετά από την επέμβαση επηρεάζουν το χρόνο ζωής τους μέχρι την επανεμφάνιση της νόσου ή το θάνατο τους από το καρκίνο. Οι μεταβλητές που έχουμε στην διάθεση μας είναι οι age, grade, positive, size και other.

Στο **πρώτο βήμα,** εξετάζουμε μια – μια τις μεταβλητές για να δούμε αν είναι αναγκαίο να μπούνε στο μοντέλο. Όπως έχουμε αναφέρει στην θεωρία ο έλεγχος βασίζεται στις διαφορές $Y=-2\log L_0-(-2\log L)$, οι οποίες κάτω από την μηδενική υπόθεση ακολουθούν μια X^2 κατανομή με q β.ε.. Στο πρώτο βήμα ορίζουμε το επίπεδο σημαντικότητας να ισούται με 20%, και αυτό διότι δεν επιθυμούμε να αποκλείσουμε κάποια σημαντική μεταβλητή από την αρχή. Η απόφαση για την επιλογή ή όχι της μεταβλητής στο μοντέλο λαμβάνεται κάθε φορά με βάση την τιμή του p-value, η οποία υπολογίζεται μέσω του R με χρήση της εντολής: p_value<- 1-pchisq(Y,q), αν η τιμή του p-value < α τότε η μεταβλητή θα πρέπει να συμπεριληφθεί στο μοντέλο.

Η τιμή για την ποσότητα -2logL υπολογίζεται κάθε φορά με χρήση της εντολής -2*objectname$loglik ως εξής:

Εκτιμούμε το κατάλληλο κάθε φορά μοντέλο παλινδρόμησης c0<-coxph(Surv(MONTH,DEATH)~1,method="breslow",data=a) και ύστερα εκτιμούμε την ζητούμενη ποσότητα -2*c0$loglik, στην συνέχεια υπολογίζουμε τη διαφορά Y, τους β.ε και το αντίστοιχο p-value, τα οποία δίνονται συνοπτικά στο παρακάτω πίνακα:

Πίνακας 7.6

Αποτελέσματα 1ου βήματος της διαδικασιας επιλογής μεταβλητών

Μοντέλο	Μεταβλητές	-2logL	Y	df	p-value	Επιλέγεται
0	καμία	1526,568				
1	age	1524,554	2,14	1	0.1558539	age
2	grade	1512,831	13.737	1	0.0002102711	grade
3	positive	1487,203	39.365	1	3.515419e-10	positive
4	size	1507,440	19.128	1	1.222384e-05	size

Με βάση το παραπάνω πίνακα κανένα από τα p-value δεν είναι μεγαλύτερο του 20% οπότε όλες οι μεταβλητές είναι σημαντικές. Το μοντέλο μας στο τέλος του 1ου βήματος έχει την εξής μορφή:

$$h(t \mid z) = h_0(t)\exp(\beta_1 age + \beta_2 grade + \beta_3 positive + \beta_4 size)$$

Στο **δεύτερο βήμα** της διαδικασίας επιλογής μοντέλου ελέγχουμε αν οι μεταβλητές που περιλαμβάνονται στο μοντέλο του 1ου βήματος είναι όλες σημαντικές, το κάνουμε αυτό αφαιρώντας κάθε φορά μια μεταβλητή από το αρχικό μοντέλο. Για το συγκεκριμένο βήμα και για τα υπόλοιπα που ακολουθούν ορίζουμε το ε.σ. να ισούται με 10%.

Πίνακας 7.7

Αποτελέσματα 2ου βήματος της διαδικασιας επιλογής μεταβλητών

Μοντέλο	Μεταβλητές	-2logL	Y	df	p-value	Επιλέγεται
0	age,grade,positive,size	1472.260				
1	grade,positive,size	1473.015	0.755	1	0.3848978	
2	positive,size	1478.349	6.089	1	0.01360261	grade
3	grade,size	1497.885	25.625	1	4.146326e-07	positive
4	grade,positive	1480.472	8.212	1	0.004161426	size

Από την παραπάνω διαδικασία προκύπτει πως από το αρχικό μοντέλο θα πρέπει να απομακρυνθεί η μεταβλητή που δηλώνει την ηλικία της ασθενούς την στιγμή της επέμβασης, αφού η τιμή p-value για την συγκεκριμένη μεταβλητή ισούται με 0.3848978, μεγαλύτερο από το 0.1 ε.σ., επομένως η παράμετρος της μεταβλητής AGE μπορούμε να δεχτούμε ότι δεν διαφέρει από το μηδέν με πιθανότητα σφάλματος 10%.

Η μορφή του μοντέλου μας θα είναι:

$$h(t \mid z) = h_0(t)\exp(\beta_1 grade + \beta_2 positive + \beta_3 size)$$

Εφόσον το μοντέλο μας άλλαξε θα πρέπει να επαναλάβουμε το 2ο βήμα προκειμένου να επιβεβαιώσουμε αν χρειάζονται μόνο αυτές οι μεταβλητές στο μοντέλο.

Πίνακας 7.8

Αποτελέσματα επανάληψης του 2ου βήματος

Μοντέλο	Μεταβλητές	-2logL	Y	df	p-value	Επιλέγεται
0	grade,positive,size	1473.015				
1	positive,size	1478.349	5.334	1	0.02091333	grade
2	grade,size	1497.885	24.87	1	6.132944e-07	positive
3	grade,positive	1480.472	7.457	1	0.006319023	size

Και οι τρείς μεταβλητές που περιλαμβάνονται στο 2ο μοντέλο είναι σημαντικές, το μοντέλο μας παραμένει το ίδιο δεν περιέχει κάποια επιπλέον μεταβλητή.

Στο **τρίτο βήμα** ελέγχουμε αν κάποιες από τις μεταβλητές που αφαιρέσαμε στο 1ο βήμα θα μπορούσαν να συμπεριληφθούν στο 2ο μοντέλο που προέκυψε από το 2ο βήμα. Επειδή όμως στο 1ο βήμα δεν αφαιρέσαμε καμία μεταβλητή, αφού όλες ήταν σημαντικές για την ανάλυση των δεδομένων μας, δεν κάνουμε κάτι σε αυτό το βήμα.

Προχωράμε στο **τέταρτο βήμα,** στο βήμα αυτό προσθέτουμε στο 2ο μοντέλο και κάποιους όρους αλληλεπίδρασης. Γνωρίζουμε ότι εισάγουμε αλληλεπιδράσεις μεταβλητών μόνο όταν οι κύριες επιδράσεις των μεταβλητών αυτών υπάρχουν ήδη στο μοντέλο. Το μοντέλο που πρόεκυψε από το 2ο βήμα αποτελείται από τις τρείς μεταβλητές grade, positive, και size, από τις οποίες δύο είναι συνεχείς, οι positive και size, για αυτές δεν έχει νόημα να πάρουμε αλληλεπίδραση. Επομένως οι όροι αλληλεπίδρασης που πρέπει να εξετάσουμε ως προς την σημαντικότητα τους είναι: grade : positive, και grade : size.

Πίνακας 7.9

Αποτελέσματα 4ου βήματος της διαδικασιας επιλογής μεταβλητών

Μοντέλο	Μεταβλητές	-2logL	Y	df	p-value	Επιλέγεται
0	grade,positive,size	1473.015				
1	grade,positive,size,grade:positive	1471.354	1.661	1	0.1974684	
2	grade,positive,size,grade:size	1472.819	0.196	1	0.6579691	

Κανένας όρος αλληλεπίδρασης δεν αποδείχθηκε σημαντικός, το μοντέλο μας παραμένει ως έχει.

Άρα το πλέον κατάλληλο μοντέλο για τα δεδομένα μας είναι το:

$$h(t \mid z) = h_0(t)\exp(\beta_1 grade + \beta_2 positive + \beta_3 size) \qquad (7.6)$$

Μια επιπλέον πληροφορία για τις μεταβλητές που θα πρέπει να συμπεριλάβουμε σε ένα μοντέλο αναλογικού κινδύνου μας δίνει το **κριτήριο πληροφορίας του Akaike,** το οποίο όπως αναφέραμε στην θεωρεία υπολογίζεται από την σχέση:

$$AIC = -2\log \hat{L} + aq$$

με την μικρότερη τιμή του AIC να αντιστοιχεί στο καλύτερο μοντέλο για τα δεδομένα μας.

Πίνακας 7.10

Αποτελέσματα από την εφαρμογή του κριτηρίου του AIC

Μοντέλο	Μεταβλητές	-2logL	AIC
0	Καμιά	1526.568	1526.568
1	age	1524.554	1527.544
2	grade	1512.831	1515.831
3	positive	1487.203	1490.203
4	size	1510.44	1513.44
5	age+grade	1510.629	1516.629
6	age+positive	1486.305	1492.305
7	age+size	1505.777	1511.777
8	grade+positive	1480.472	1486.472
9	grade+size	1497.885	1503.885
10	positive+size	1478.349	1484.349
11	age+grade+positive	1479.558	1488.558
12	age+grade+size	1496.084	1505.084
13	**grade+positive+size**	**1473.015**	**1482.015**
14	age+grade+positive+size	1472.260	1484.26

Επομένως το καλύτερο μοντέλο είναι το 13^{ο} για το οποίο παρατηρείται η μικρότερη τιμή του κριτηρίου του Akaike, το μοντέλο αυτό συμπίπτει με αυτό της παραπάνω διαδικασίας επιλογής μεταβλητών. Δεν έχει νόημα ο έλεγχος σημαντικότητας όρων αλληλεπίδρασης αφού θα ήταν μια επανάληψη του 4^{ου} βήματος της διαδικασίας που περιγράψαμε πιο πάνω, από την οποία αποφασίσαμε ως μη αναγκαία την εισαγωγή όρων αλληλεπίδρασης στο μοντέλο.

7.3.1 Εφαρμογή του μοντέλου

Αρχικά θα προσαρμόσουμε το μοντέλο (7.6) που προέκυψε από την προηγούμενη παράγραφο, το οποίο θεωρούμε ως το πλέον κατάλληλο για την ερμηνεία και αξιολόγηση των δεδομένων μας.

Πίνακας 7.11

Συντελεστές παλινδρόμησης για το μοντέλο (7.6)

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
factor(grade)1	-0.473364	0.622903	0.492140	-0.962	0.3361
factor(grade)2	-0.064043	0.937965	0.403728	-0.159	0.8740
factor(grade)3	0.292124	1.339270	0.409848	0.713	0.4760
positive	0.045229	1.046267	0.008399	5.385	7.24e-08 ***
size	0.013637	1.013731	0.005478	2.489	0.0128 *
	exp(coef)	exp(-coef)	lower .95		upper .95
factor(grade)1	0.6229	1.6054	0.2374		1.634
factor(grade)2	0.9380	1.0661	0.4251		2.069
factor(grade)3	1.3393	0.7467	0.5998		2.990
positive	1.0463	0.9558	1.0292		1.064
size	1.0137	0.9865	1.0029		1.025
Rsquare= 0.062 (max possible= 0.84)					
Likelihood ratio test = 53.22 on 5 df, p=3.033e-10					
Wald test = 69.67 on 5 df, p=1.198e-13					
Score (logrank) test = 77.7 on 5 df, p=2.554e-15					

Από τα p – value των ολικών ελέγχων που είναι σχεδόν ίσα με μηδέν απορρίπτουμε την μηδενική υπόθεση $H_0 : (\beta_1, ...\beta_5)' = 0$, τουλάχιστον μια από τις παραμέτρους του μοντέλου είναι διαφορετική του μηδενός, επομένως τουλάχιστον μια από τις συμμεταβλητές επηρεάζει τη συνάρτηση κινδύνου.

Παρατηρώντας τον πίνακα των αποτελεσμάτων και σύμφωνα με τους τοπικούς ελέγχους μόνο οι συμμεταβλητές size και positive επηρεάζουν σημαντικά την συνάρτηση κινδύνου (έχουν την τάση να αυξάνουν το κίνδυνο θανάτου). Τα επίπεδα

της μεταβλητής grade δεν φαίνονται να έχουν καμία προγνωστική δύναμη, αφού όλα τα p-value για το κάθε επίπεδο έχουν τιμή μεγαλύτερη του επιπέδου σημαντικότητας 5%.

Ακόμη βλέπουμε ότι αν αυξηθεί το μέγεθος του όγκου για κάποια ασθενή κατά 1mm, τότε ο κίνδυνος θανάτου αυξάνεται σε σχέση με κάποια άλλη ασθενή για την οποία το μέγεθος του όγκου έχει παραμείνει σταθερό, όλα αυτά εφόσον κρατάμε σταθερές τις υπόλοιπες συμμεταβλητές του μοντέλου.

$$HR(t) = \frac{h(t \mid grade1, grade2, grade3, size = i+1, positive)}{h(t \mid grade1, grade2, grade3, size = i, positive)} = e^{\hat{\beta}5} = 1.0137$$

Τα ίδια ισχύουν και για την μεταβλητή positive.

7.4 Αξιολόγηση της υπόθεσης αναλογικού κινδύνου - Εφαρμογή

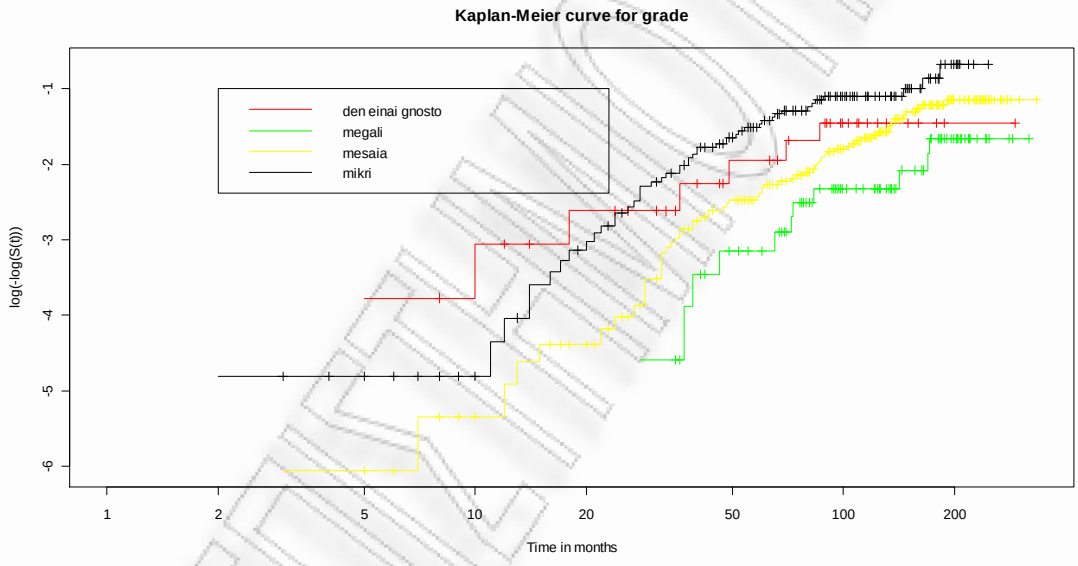
Στην ενότητα αυτή εφαρμόζουμε κάποιες γραφικές μεθόδους αξιολόγησης της υπόθεσης PH (αναλογικού κινδύνου). Οι μέθοδοι αυτές μας δίνουν κάποιες ενδείξεις για την ισχύ ή όχι της υπόθεσης PH, βέβαια η απόφαση μας περιέχει στοιχεία υποκειμενικότητας, και απορρίπτουμε την υπόθεση PH μόνο στην περίπτωση που υπάρχουν σοβαρές ενδείξεις για αυτό.

Υπολογίζουμε τις συναρτήσεις επιβίωσης για δύο άτομα με διανύσματα συμμεταβλητών $S(t \mid Z) = [S_0(t)]\exp(\beta' Z)$, $S(t \mid Z^*) = [S_0(t)]\exp(\beta' Z^*)$, στην συνέχεια απεικονίζουμε σε ένα κοινό γράφημα τις ποσότητες $\log[-\log S(t \mid Z)]$ *και* $\log[-\log S(t \mid Z^*)]$, οι δύο καμπύλες θα πρέπει να έχουν απόσταση μεταξύ τους ίση με $\beta'(Z - Z^*)$ σταθερή για κάθε t, δηλαδή οι δύο καμπύλες θα πρέπει να είναι σχεδόν παράλληλες. Επομένως αν ισχύει η υπόθεση αναλογικού κινδύνου για ένα σύνολο συμμεταβλητών περιμένουμε οι log-log γραφικές παραστάσεις για δύο οποιαδήποτε άτομα να είναι προσεγγιστικά παράλληλες.

Ενδιαφερόμαστε να αξιολογήσουμε την υπόθεση αναλογικού κινδύνου για την μεταβλητή Grade. Για να κατασκευάσουμε τις log-log γραφικές παραστάσεις των εκτιμήσεων των συναρτήσεων επιβίωσης των τεσσάρων ομάδων με την μέθοδο Kaplan – Meier εφαρμόζουμε τις παρακάτω εντολές στην R:

```
fit1<-survfit(Surv(MONTH,DEATH)~GRADE,data=a)
plot(fit1,fun="cloglog",xlab="Time in months",ylab="log(-log(S(t)))",main="Kaplan-Meier curve for grade",lty=1:1,col=c("red","green","yellow","black") )
legend(2,-1,c("den einai gnosto","megali","mesaia","mikri"),
  lty=1:1,col=c("red","green","yellow","black")
```

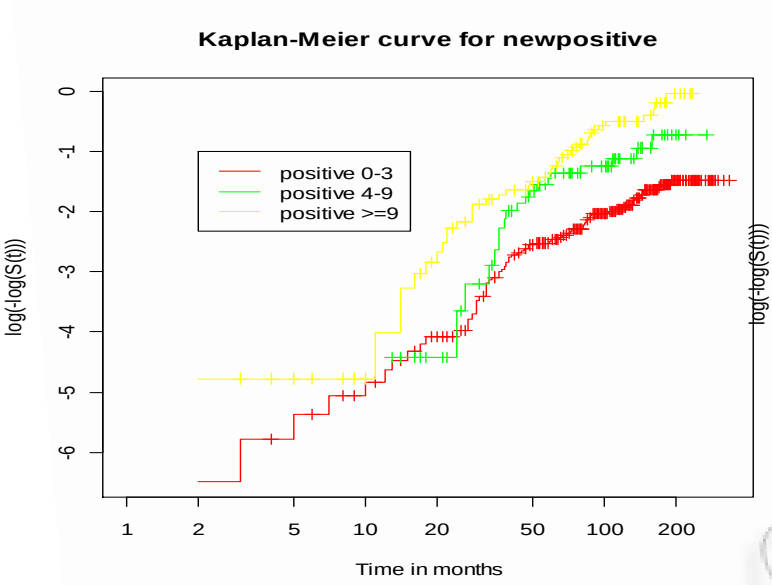
Γράφημα 7.1

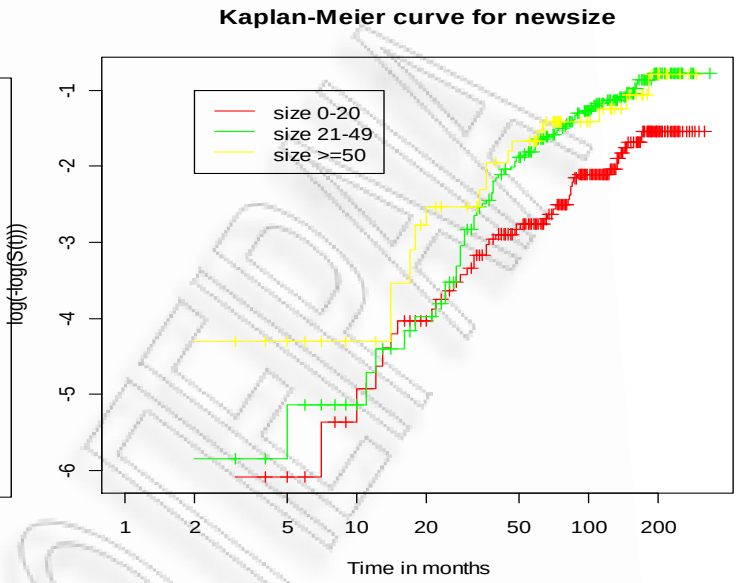


Σύμφωνα με την παραπάνω γραφική παράσταση η υπόθεση αναλογικού κινδύνου δε φαίνεται να ισχύει για τη μεταβλητή grade, τουλάχιστον όχι για όλα τα επίπεδά της, φαίνεται να ικανοποιείται για τα επίπεδα grade=1, grade=2 και grade=3 τα οποία θα απεικόνιζαν σχεδόν "παράλληλες" ευθείες αν δεν περιλαμβανόταν το επίπεδο grade=0 (που δηλώνει την άγνωστη διαφοροποίηση του όγκου) .

Αν θέλαμε να εξετάσουμε γραφικά αν ισχύει ή όχι η υπόθεση αναλογικού κινδύνου για τις μεταβλητές positive και size θα έπρεπε να τις κατηγοριοποιήσουμε κατάλληλα πρώτα. Χρησιμοποιώντας τις κατηγοριοποιήσεις που αναφέραμε στο προηγούμενο κεφάλαιο και συγκεκριμένα στην παράγραφο 6.3.1, έχουμε τις εξής γραφικές παραστάσεις:

Γράφημα 7.2 **Γράφημα 7.3**





Παρατηρούμε πως η υπόθεση αναλογικού κινδύνου φαίνεται να παραβιάζεται τόσο για την μεταβλητή positive όσο και για την μεταβλητή size (για την συγκεκριμένη μεταβλητή το γράφημα είναι πιο ξεκάθαρο), εφόσον οι καμπύλες τέμνονται.

Βέβαια τα αποτελέσματα που παίρνουμε από την παραπάνω μέθοδο είναι υποκειμενικά, μας δίνουν ωστόσο κάποιες ενδείξεις. Για την καλύτερη τεκμηρίωση της απόφασης για το αν κάποια μεταβλητή ικανοποιεί ή όχι την υπόθεση αναλογικού κινδύνου, καλό θα είναι να χρησιμοποιήσουμε και κάποιο στατιστικό έλεγχο, όπως αυτός που περιγράφεται στην επόμενη ενότητα

7.4.1 Αξιολόγηση της υπόθεσης αναλογικού κινδύνου μέσω υπολοίπων

Μπορεί από τα παραπάνω διαγράμματα να πήραμε μια πρώτη αίσθηση για την ισχύ ή όχι της υπόθεσης αναλογικού κινδύνου, όμως για να είμαστε βέβαιοι θα εκτιμήσουμε τα παρακάτω υπόλοιπα και θα αποφασίσουμε αν ισχύει η υπόθεση αναλογικού κινδύνου για την κάθε μεταβλητή ξεχωριστά αλλά και ταυτοχρόνως. Ταυτόχρονα με την υπόθεση αναλογικού κινδύνου μπορούμε παράλληλα να συμπερασματολογήσουμε σχετικά με την ολική επάρκεια του μοντέλου (overall model fit).

Αρχικά εκτιμούμε τα Schoenfeld υπόλοιπα, τα οποία δεν εξαρτώνται από την αθροιστική συνάρτηση κινδύνου και δίνουν τιμή υπολοίπου σε κάθε άτομο για κάθε συμμεταβλητή που υπάρχει στο μοντέλο.

Για τον υπολογισμό των Schoenfeld υπολοίπων εφαρμόζουμε την εντολή residuals() εισάγοντας ως αντικείμενο το μοντέλο παλινδρόμησης που ορίζει η ρουτίνα coxph() (για τις τιμές των συγκεκριμένων υπολοίπων βλέπε παράρτημα).

Εκτελώντας την παραπάνω ρουτίνα παίρνουμε τα Schoenfeld υπόλοιπα για την κάθε μεταβλητή σε κάθε πλήρη χρόνο, διότι όπως είναι γνωστό στους λογοκριμένους χρόνους τα Schoenfeld υπόλοιπα ισούνται με μηδέν.

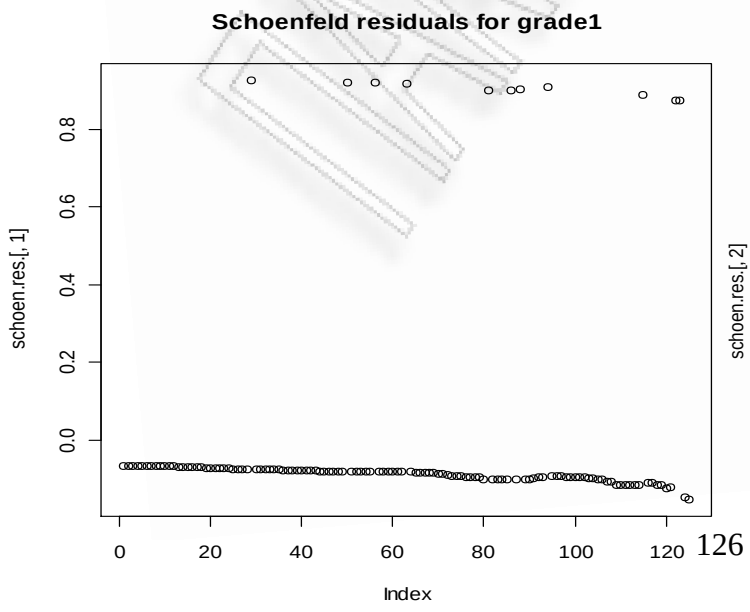
Για την κάθε μεταβλητή που μας ενδιαφέρει απεικονίζουμε τα Schoenfeld υπόλοιπα της έναντι του χρόνου μέσων την εντολής plot(schoen.res.[,i]), προκειμένου να έχουμε μια πρώτη ένδειξη για το αν η υπό εξέταση μεταβλητή ικανοποιεί την υπόθεση αναλογικού κινδύνου, αν ισχύει κάτι τέτοιο θα πρέπει στο γράφημα των συγκεκριμένων υπολοίπων έναντι του χρόνου να παρατηρήσουμε μια τυχαία συμπεριφορά.

Έτσι για την μεταβλητή grade (και για όλες τις υπόλοιπες εφαρμόζουμε την παρακάτω εντολή κάνοντας τις αναγκαίες αλλαγές) έχουμε τα παρακάτω γραφήματα:

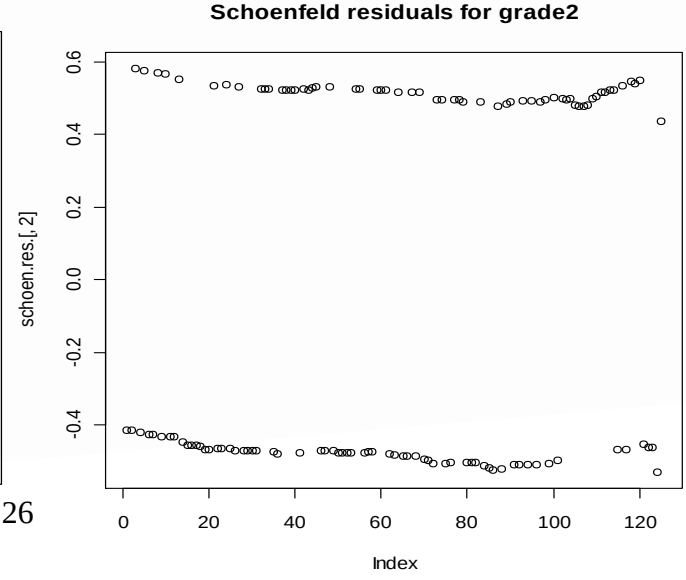
plot(schoen.res.[,1], main="Schoenfeld residuals for grade1")

Σύμφωνα με τα παρακάτω γραφήματα υπάρχουν ενδείξεις ότι η υπόθεση αναλογικού κινδύνου ισχύει για της μεταβλητές positive και size, ενώ υπάρχουν ενδείξεις παραβίασης της υπόθεσης αναλογικού κινδύνου για τις υπόλοιπες μεταβλητές.

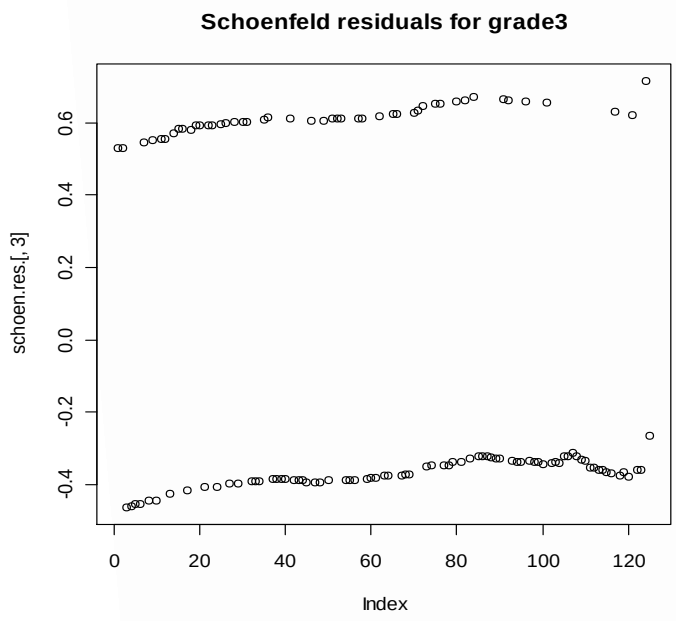
Γράφημα 7.4



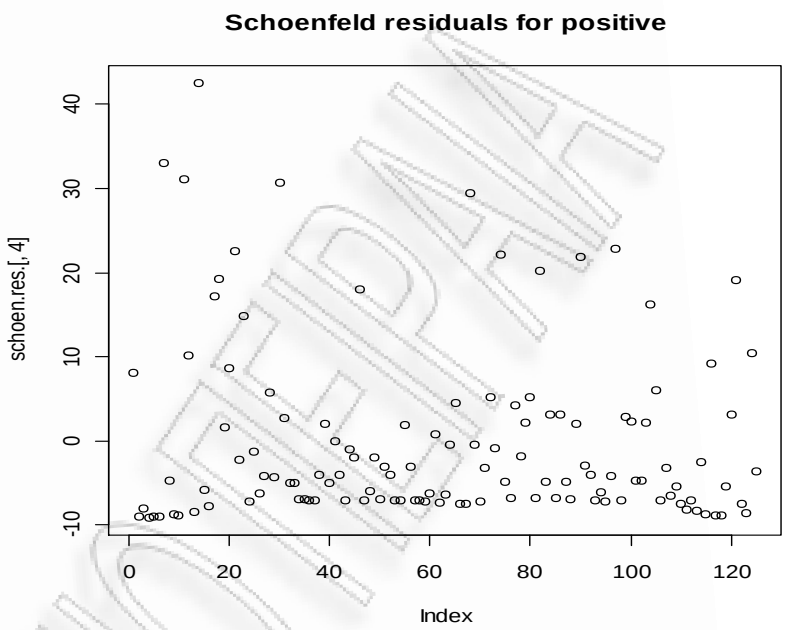
Γράφημα 7.5



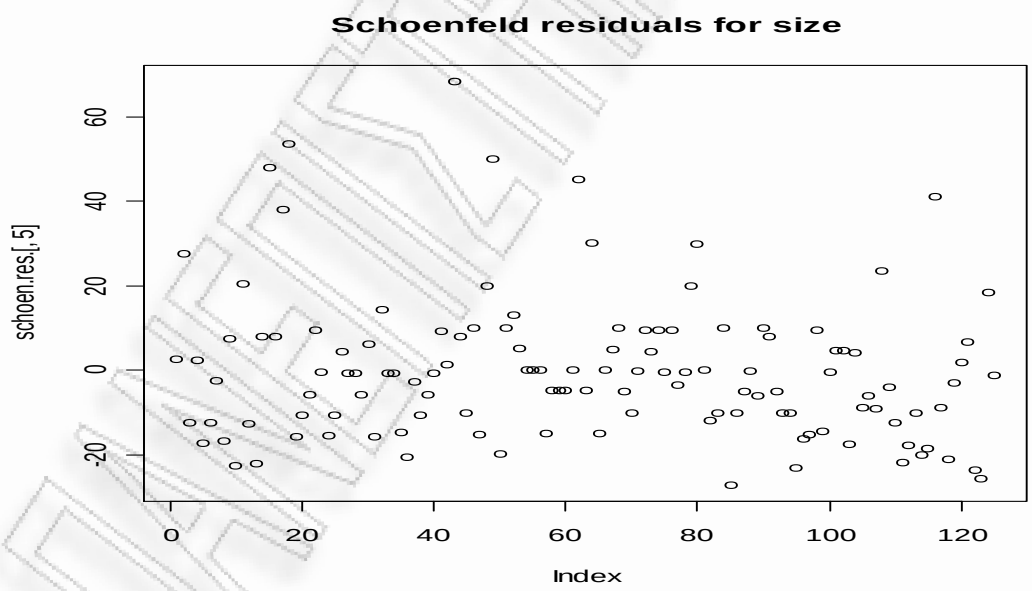
Γράφημα 7.6



Γράφημα 7.7



Γράφημα 7.8



Στην πράξη χρησιμοποιούνται πιο πολύ τα Scaled Schoenfeld υπόλοιπα και αυτό λόγω των σημαντικών ιδιοτήτων που έχουν. Μας δίνουν την δυνατότητα να εφαρμόσουμε στατιστικούς ελέγχους και την προσαρμογή μια καμπύλης εξομάλυνσης πάνω στο γράφημα. Υπολογίζονται μέσω της εντολής cox.zph(), χρησιμοποιώντας την ως εξής:

test.ph<-cox.zph(cox8,transform="identity");test.ph

Πίνακας 7.12

Ολικός έλεγχος και τοπικοί έλεγχοι για το μοντέλο (7.6)

	rho	chisq	p-value
factor(grade)1	0.12542	2.01081	0.156
factor(grade)2	0.12174	1.93755	0.164
factor(grade)3	0.03882	0.19819	0.656
positive	-0.00734	0.00617	0.937
size	-0.09238	1.07757	0.299
GLOBAL	NA	8.85418	0.115

Παρατηρώντας την στήλη με τα p – value και για επίπεδο σημαντικότητας α=5% βλέπουμε πως όλες οι μεταβλητές ξεχωριστά ικανοποιούν την υπόθεση αναλογικού κινδύνου. Ακόμη, συμπεραίνουμε ότι το μοντέλο μας είναι επαρκές εφόσον για τον ολικό έλεγχο δεν μπορούμε να απορρίψουμε την υπόθεση $H_0 : \beta_1(t) = \beta_1, \beta_2(t) = \beta_2, ...\beta_5(t) = \beta_5$ αφού έχουμε τιμή p-value = 0.115 > 0.05, επομένως οι μεταβλητές του μοντέλου ικανοποιούν ταυτόχρονα την υπόθεση αναλογικού κινδύνου.

Να σημειώσουμε ότι τα παραπάνω αποτελέσματα έρχονται σε αντίθεση με την εικόνα που πήραμε από το log – log γραφήματα, σύμφωνα με τα οποία οι μεταβλητές grade, positive και size δεν ικανοποιούσαν την υπόθεση αναλογικού κινδύνου.

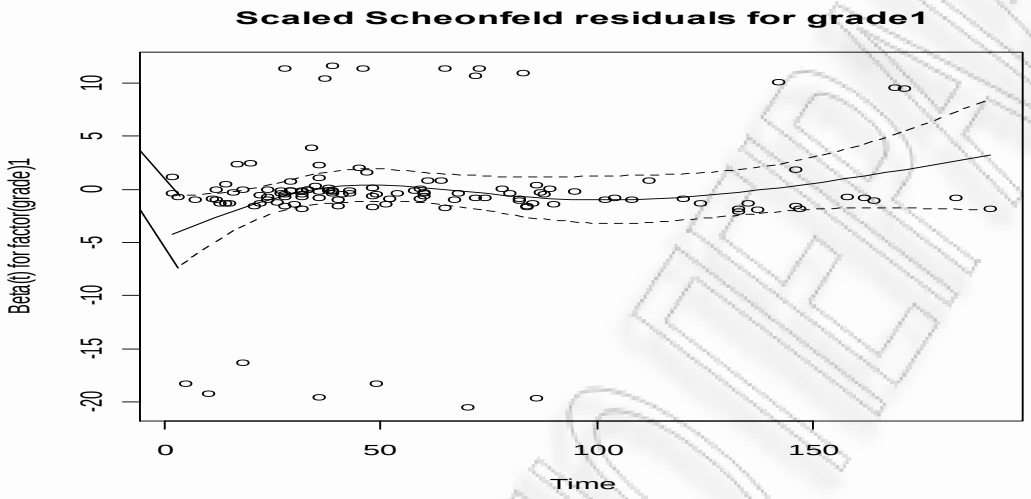
Για τον υπολογισμό και την γραφική απεικόνιση των Scaled Schoenfeld υπολοίπων για την κάθε συμμεταβλητή ακολουθούμε παρόμοια διαδικασία με αυτή που εφαρμόσαμε για τα Schoenfeld υπόλοιπα.

Τα Scaled Schoenfeld υπόλοιπα για την κάθε μεταβλητή δίνονται στο παράρτημα.

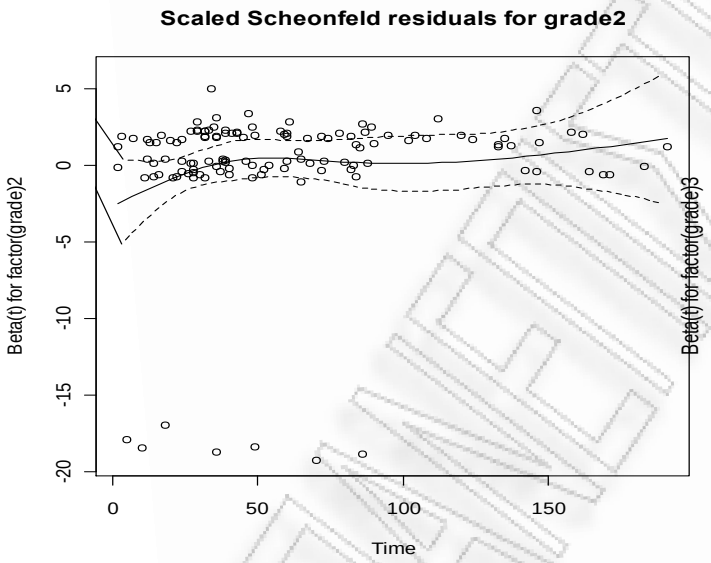
Ενώ για την γραφική απεικόνιση έχουμε:

plot(test.ph[1 ή 2 ή 3], main="Scaled Scheonfeld residuals for grade1ή 2ή 3")

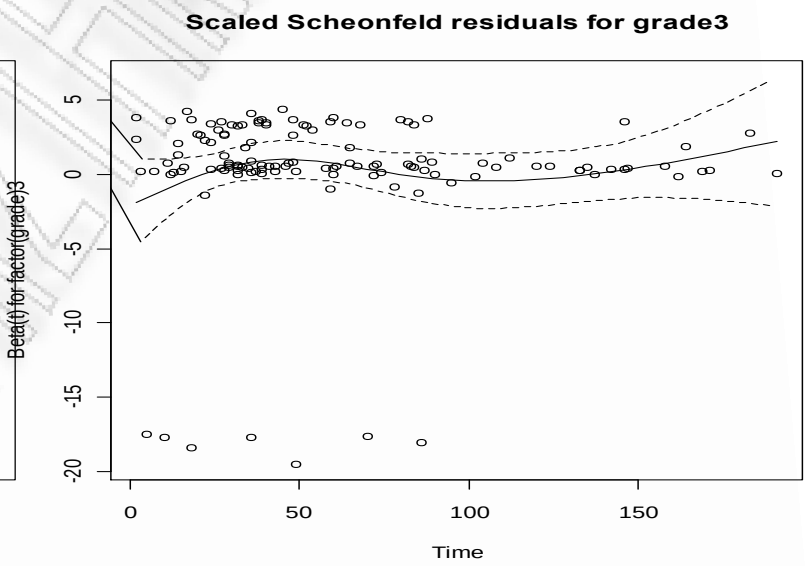
Γράφημα 7.9



Γράφημα 7.10

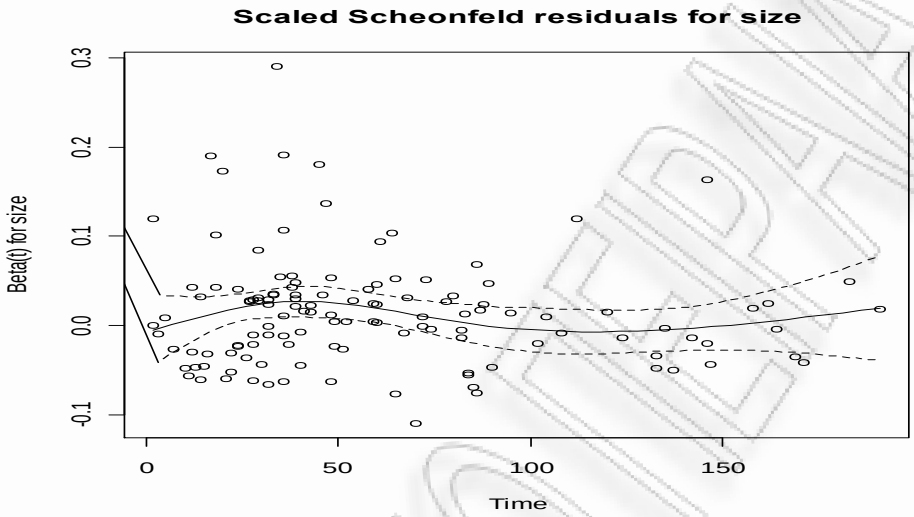


Γράφημα 7.11

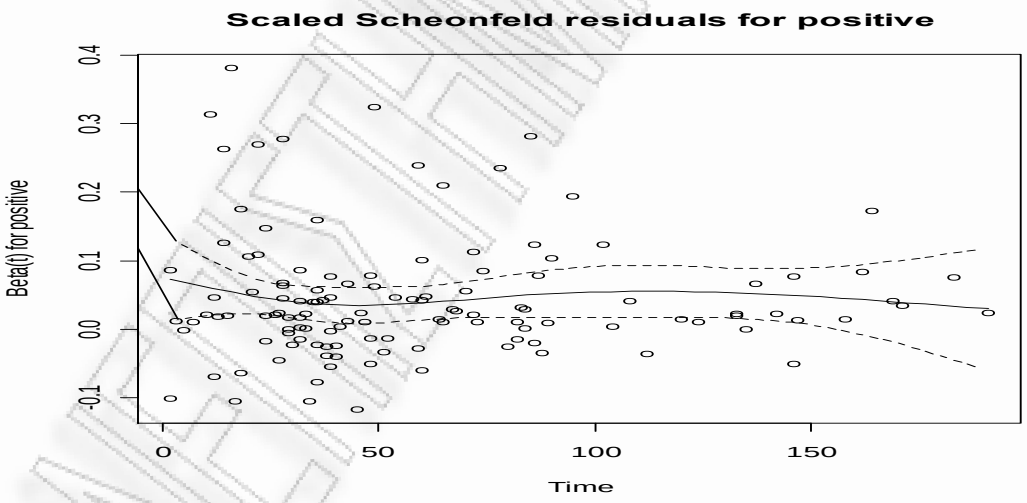


plot(test.ph[4 ή 5], main="Scaled Scheonfeld residuals for positive ή size")

Γράφημα 7.12



Γράφημα 7.13



Στα παραπάνω γραφήματα παρατηρούμε πως η smoothing curve έχει σχεδόν μηδενική κλίση, πράγμα το οποίο περιμέναμε αφού στους τοπικούς ελέγχους για την κάθε συμμεταβλητή απορρίψαμε την μηδενική υπόθεση $H_0 : \beta_i(t) = \beta_i$. Μάλιστα όσο πιο μεγάλη είναι η τιμή p-value τόσο περισσότερο θα τείνει η καμπύλη εξομάλυνσης να έχει μηδενική κλίση. Άρα οι συμμεταβλητές grade, positive και size ικανοποιούν την υπόθεση αναλογικού κινδύνου, δηλ. για την κάθε ασθενή η τιμή τους παραμένει σταθερή με την πάροδο του χρόνου.

Στην συνέχεια θα υπολογίσουμε τις διαφορές Δέλτα – Βήτα ή scaled score residuals προκειμένου να προσδιορίσουμε τις παρατηρήσεις που ασκούν την μεγαλύτερη επιρροή στην εκτίμηση του διανύσματος των παραμέτρων β, δηλαδή μετράμε την επιρροή κάθε ασθενούς στο μοντέλο (7.6). Όπως αναφέραμε στην θεωρία η μέθοδος αυτή είναι γνωστή με τον όρο μόχλευση (leverage).

Χρησιμοποιούμε και πάλι την εντολή **residuals() (ή resid()),** (για τις εκτιμήσεις των υπολοίπων βλέπε παράρτημα)

Για την γραφική αναπαράσταση των scaled score υπολοίπων ακολουθούμε την ίδια διαδικασία με αυτή των Scheonfeld και scaled υπολοίπων με την μόνη διαφορά ότι τώρα στην εντολή **plot()** προσθέτουμε την παράμετρο **type="h"** , προκειμένου τα σημεία να εμφανίζονται με την μορφή ευθειών γραμμών παράλληλων στον άξονα των y.

Αρχικά εκτιμούμε τα scaled score υπόλοιπα (παράρτημα) και στην συνέχεια παίρνουμε την γραφική απεικόνιση για κάθε συμμεταβλητή.

Για το $2^{ο}$ επίπεδο της μεταβλητής grade τα scaled score υπόλοιπα κυμαίνονται στο διάστημα (-0.15 , 0.1). Παρατηρούμε ότι υπάρχει αρκετή διαφορετικότητα στην μόχλευση (leverage) κάθε ασθενούς καθώς υπάρχουν ασθενείς με μεγάλη μόχλευση, μέτρια ακόμα και μηδενική. Στο παραπάνω διάστημα περίπου 13 ασθενείς έχουν scaled score υπόλοιπα μεγαλύτερο του 0.5 και έξι ασθενείς έχουν τιμή υπολοίπου μικρότερη του -.10, αυτές οι 6 παρατηρήσεις φαίνεται να επηρεάζουν παρόμοια το διάνυσμα των παραμέτρων β, παρόλα αυτά μπορούμε να διακρίνουμε πως την μεγαλύτερη επιρροή ασκεί η $600^{η}$ παρατήρηση με τιμή scaled score υπολοίπου ίση με -0.1490.

Για το $3^{ο}$ επίπεδο τις μεταβλητής grade τα scaled score υπόλοιπα βρίσκονται στο διάστημα (-0.15 ,0.1). Η εικόνα είναι παρόμοια με αυτή του $1^{ου}$ επιπέδου όσον αφορά τα αρνητικά υπόλοιπα, διακρίνουμε τις 6 παρατηρήσεις που επηρεάζουν περισσότερο το διανύσμα των παραμέτρων β και η $600^{η}$ παρατήρηση φαίνεται και για αυτό το επίπεδο να επηρεάζει συγκριτικά περισσότερο το μοντέλο, το scaled score υπόλοιπο της ισούται με -0.1465. Μόνο δύο παρατηρήσεις έχουν τιμή υπολοίπου μεγαλύτερη του 0.05.

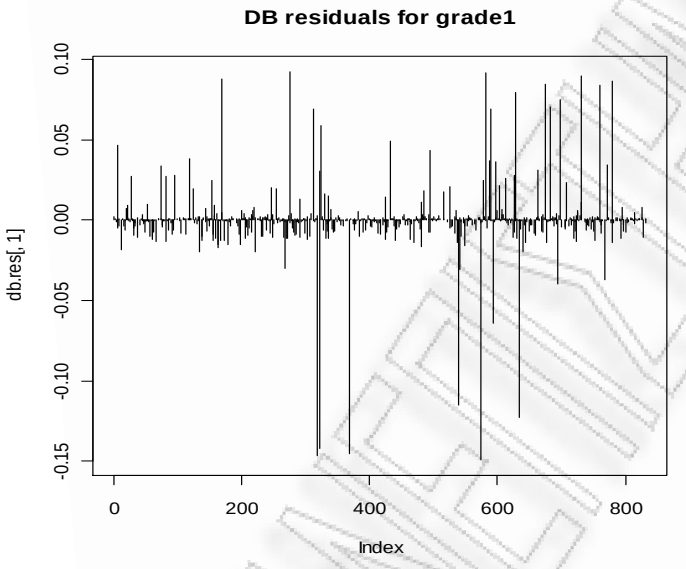
Για το $4^{ο}$ επίπεδο της grade το γράφημα των scaled score υπολοίπων είναι παρόμοιο με το γράφημα του $3^{ου}$ επιπέδου. Μόνο δύο ασθενείς έχουν τιμή υπολοίπου

μεγαλύτερη του 0.05 και στα αρνητικά υπόλοιπα πάλι φαίνονται οι 6 παρατηρήσεις που επηρεάζουν περισσότερο το μοντέλο, με την 600η παρατήρηση να ασκεί ελάχιστη περισσότερη επιρροή.

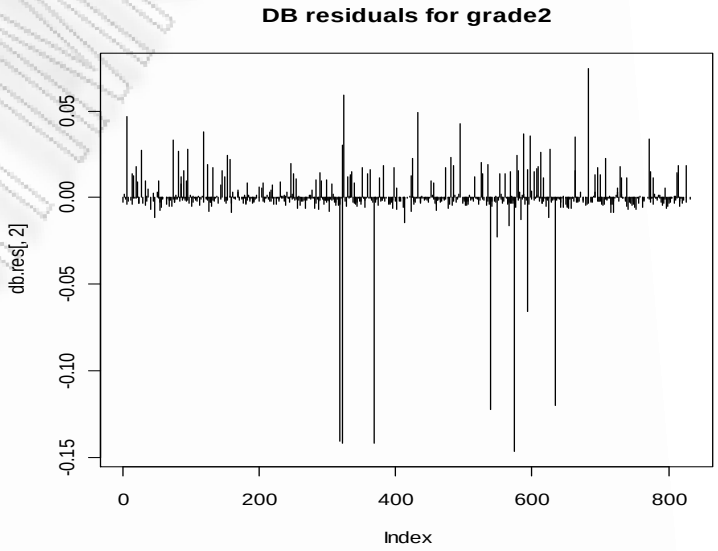
Για την μεταβλητή positive οι ασθενείς φαίνεται να ασκούν την ίδια επιρροή στο μοντέλο, μόνο μια παρατήρηση φαίνεται να επηρεάζει σημαντικά περισσότερο, η 279η με τιμή υπολοίπου -0.00811 ενώ όλα τα άλλα υπόλοιπα κυμαίνονται στο διάστημα (-0.002 – 0.002).

Τέλος για την μεταβλητή size παρατηρούμε πως όλες οι παρατηρήσεις ακούν κάποια επιρροή στο μοντέλο, οι τιμές των scaled score υπολοίπων κυμαίνονται στο διάστημα (- 0.001 – 0.0015), μόνο μια παρατήρηση η 690η έχει τιμή υπολοίπου μεγαλύτερη του 0.0015 (συγκεκριμένα το υπόλοιπό της ισούται με 0.001939) κατά συνέπεια ασκεί και την μεγαλύτερη επιρροή.

Γράφημα 7.14

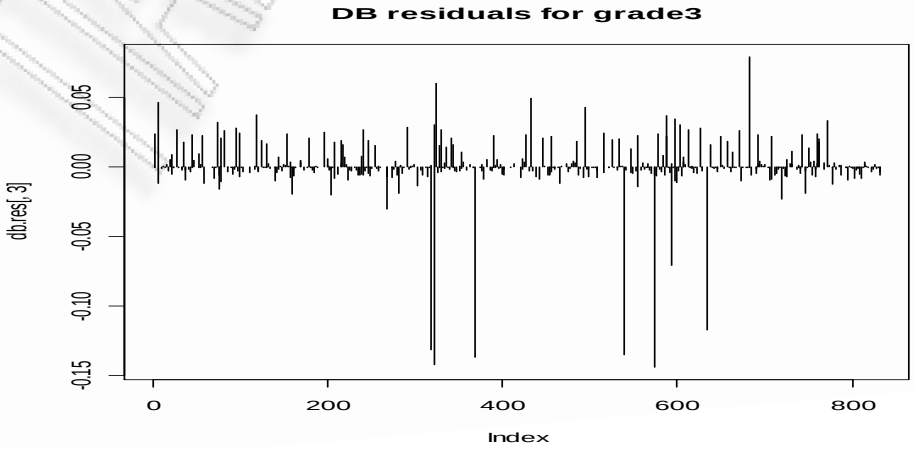


Γράφημα 7.15

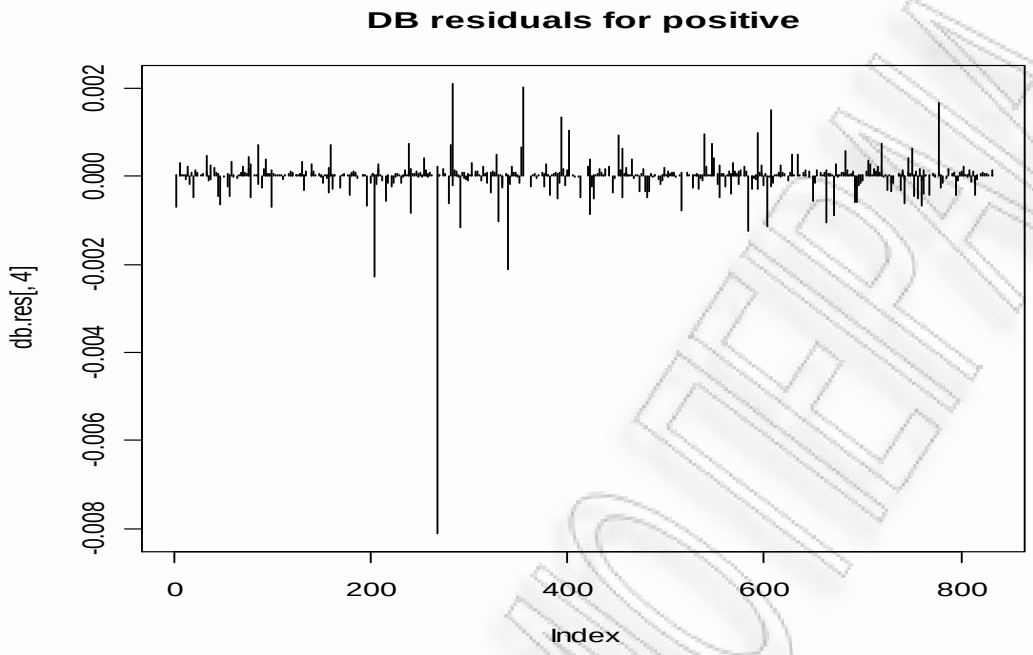


.

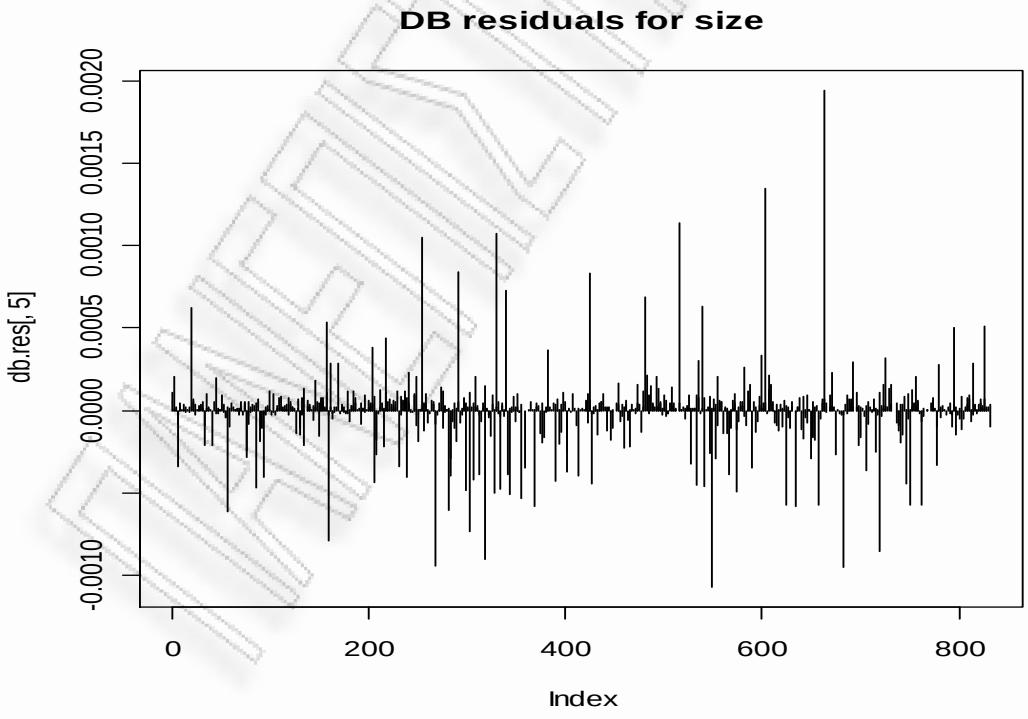
Γράφημα 7.16



Γράφημα 7.17



Γράφημα 7.18



7.5 Martingale residuals, Deviance residuals - εφαρμογή

Όπως αναφέραμε και στην θεωρεία τα martingale υπόλοιπα χρησιμοποιούνται πιο πολύ για την εύρεση της συναρτησιακής μορφής μιας ερμηνευτικής μεταβλητής που πρόκειται να εισάγουμε σε ένα μοντέλο παλινδρόμησης.

Έστω ότι έχουμε το παρακάτω μοντέλο παλινδρόμησης,

$$h(t \mid Z) = h_0(t)\exp(\beta_1 SIZE)$$

και θέλουμε να προσθέσουμε σε αυτό μια νέα ερμηνευτική μεταβλητή την positive με συναρτησιακή μορφή $f(positive)$, τότε το διευρυμένο μοντέλο θα έχει την μορφή,

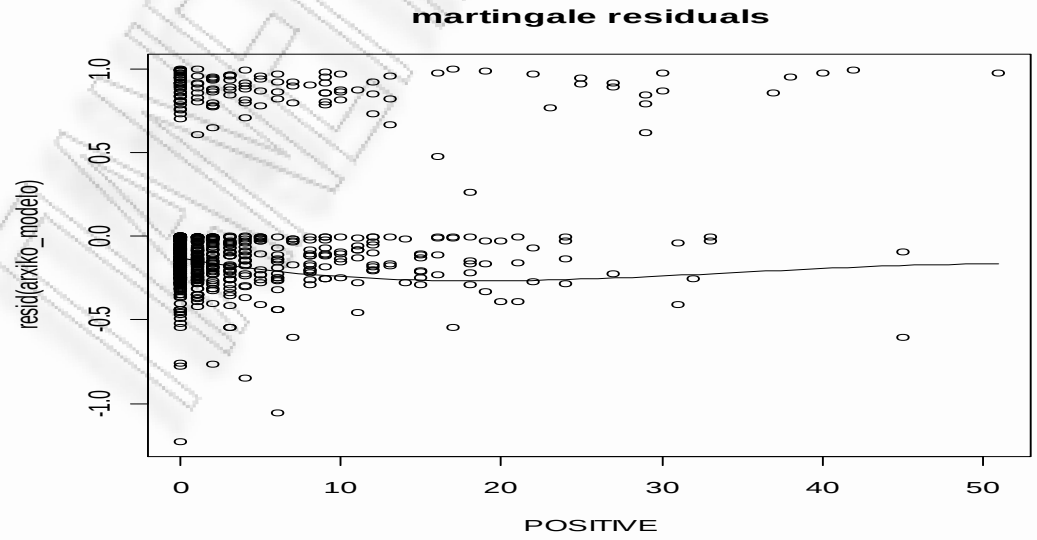
$$h(t \mid Z^*) = h_0(t)\exp(\beta_1 SIZE + f(positive))$$

Για να εντοπίσουμε την μορφή της $f(positive)$ θα απεικονίσουμε σε ένα διάγραμμα τα martingale υπόλοιπα του αρχικού μοντέλου (y άξονας) ως προς τις τιμές της μεταβλητής positive (x άξονας). Στο διάγραμμα αυτό εφαρμόζουμε και μια smoothing curve, η οποία μας βοηθάει στο προσδιορισμό της συναρτησιακής μορφής της positive για την είσοδο της στο μοντέλο όταν σε αυτό υπάρχει ήδη η size.

Εφαρμόζουμε τις ακόλουθες ρουτίνες στην R:

```
arxiko_modelo<-coxph(Surv(MONTH,DEATH)~SIZE, data=a)
scatter.smooth(POSITIVE,resid(arxiko_modelo), main="martingale residuals")
```

Γράφημα 7.19



Στο παραπάνω γράφημα βλέπουμε πως η smoothing curve προσεγγίζει τη γραφική παράσταση μιας ευθείας (δεν χρειάζεται να κάνουμε κάποια μετατροπή), οπότε το διευρυμένο μοντέλο θα έχει την μορφή:

$$h(t \mid Z) = h_0(t)\exp(\beta_1 size + \beta_2 positive)$$

Η κατηγοριοποίηση αυτή βλέπουμε πως δεν συμπίπτει καθόλου με αυτήν που είχαμε προτείνει για την μεταβλητή positive στο $6^{ο}$ κεφάλαιο.

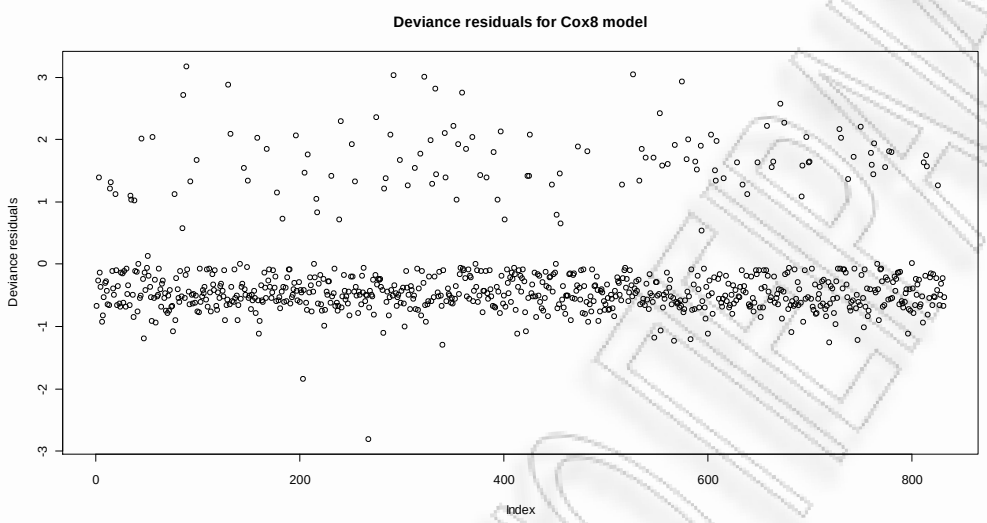
Τώρα όσο αφορά τα **deviance υπόλοιπα**, τα οποία όπως έχουμε αναφέρει στην θεωρεία κατανέμονται περισσότερο συμμετρικά γύρω από το μηδέν σε σύγκριση με τα martingale υπόλοιπα επιτρέποντας έτσι την ευκολότερη ανίχνευση έκτροπων παρατηρήσεων, υπολογίζονται κάνοντας χρήση της εντολής residuals():

(Υπολογίσουμε τα deviance υπόλοιπα για το μοντέλο cox8 , το οποίο προέκυψε ως το καλύτερο μοντέλο για την ανάλυση των δεδομένων μας ως αποτέλεσμα της τεχνικής επιλογής μεταβλητών.)

```
dev.res.<-residuals(cox8,type="deviance");dev.res.
        1          2          3          4          5          6
-0.66545073 -0.25567141  1.39742291 -0.12939898 -0.36446549 -0.92026717
        7          8          9         10         11         12
-0.81608096 -0.52251694 -0.30267397 -0.27704391 -0.42379289 -0.64203491
.........  ..........  .........  ..........  ..........  ..........
```

```
plot(dev.res.,ylab="Deviance residuals",main="Deviance residuals for Cox8 model")
```

Γράφημα 7.20



Μια μεγάλη τιμή αυτών των υπολοίπων αποτελεί ένδειξη ύπαρξης έκτροπων παρατηρήσεων. Σύμφωνα με το γράφημα αρκετές παρατηρήσεις έχουν deviance υπόλοιπο μεγαλύτερο του δύο (περίπου 13). Η μεγαλύτερη τιμή υπολοίπου ισούται με 3.17292853 η οποία θεωρείται έκτροπη παρατήρηση όπως και η αμέσως επόμενη που έχει τιμή υπολοίπου ίση με 3.03977435, έπειτα ακολουθεί η $600^{η}$ παρατήρηση με τιμή υπολοίπου 2.93103846, αυτή και οι υπόλοιπες παρατηρήσεις με deviance υπόλοιπο >= 2 και < 3 είναι υποψήφιες έκτροπες παρατηρήσεις.

7.6 Εφαρμογή μοντέλων ευπάθειας (Frailty models)

Προκειμένου να εφαρμόσουμε ένα μονομεταβλητό μοντέλο ευπάθειας στην R θα πρέπει στο σύνολο των δεδομένων μας να συμπεριλάβουμε και μια μεταβλητή η οποία στην ουσία μετράει των αριθμό των παρατηρήσεων του δείγματος (έστω ID), την συγκεκριμένη μεταβλητή βάζουμε ως όρο ευπάθειας. Με αυτό το τρόπο επιδιώκουμε να διαπιστώσουμε αν στο μοντέλο που επιλέξαμε παραπάνω ως το καλύτερο (μοντέλο 7.6) χρειάζεται να λάβουμε υπόψη κάποια τυχαία επίδραση, έτσι ώστε να καταλήξουμε σε ένα καλύτερο μοντέλο.

Τα μοντέλα ευπάθειας, μας δίνουν την δυνατότητα να εξετάσουμε αν η έλλειψη της αναλογικότητας οφείλεται στην παράλειψη κάποιας σημαντικής μεταβλητής, την οποία εκφράζουμε ως μια τυχαία μη παρατηρούμενη ποσότητα.

Για να εφαρμόζουμε ένα μοντέλο ευπάθειας στην R, εκτελούμε την παρακάτω εντολή:

```
frail.1<-
coxph(Surv(MONTH,DEATH)~factor(GRADE)+POSITIVE+SIZE+frailty(ID),data=
a)
```
(7.6α)

Αν δεν προσδιορίσουμε κάποια κατανομή για την frailty τότε τα αποτελέσματα που παίρνουμε αφορούν μια γάμμα κατανομή (με μέση τιμή 1 και διακύμανση θ), έχουμε:

Πίνακας 7.13

Αποτελέσματα για το μοντέλο 7.6α (με frailty την γάμμα κατανομή)

	coef	se(coef)	se2	Chisq	DF	p
factor(grade)1	-0.5224	0.58970	0.50648	0.78	1	3.8e-01
factor(grade)2	-0.1172	0.49955	0.41782	0.06	1	8.1e-01
factor(grade)3	0.4506	0.51009	0.42348	0.78	1	3.8e-01
positive	0.0667	0.01271	0.00985	27.57	1	1.5e-07
size	0.0185	0.00702	0.00574	6.94	1	8.4e-03
frailty(id)				170.69	159	2.4e-01
Iterations: 6 outer, 32 Newton-Raphson						
Variance of random effect= 1.62 I-likelihood = -732.1						
Degrees of freedom for terms= 2.2 0.6 0.7 158.8						
Likelihood ratio test=301 on 162 df, p=2.33e-10						

Ενώ για ένα μοντέλο ευπάθειας με κανονική κατανομή θα έπρεπε μέσα στον όρο ευπάθειας να ορίσουμε ως κατανομή την κανονική, δηλαδή:

```
frail.2<-
coxph(Surv(MONTH,DEATH)~factor(GRADE)+POSITIVE+SIZE+frailty(ID,dist="
gauss"),data=a)
```
(7.6β)

Πίνακας 7.14

Απότελέσματα για το μοντέλο 7.6β (με frailty την Gaussian distribution)

	coef	se(coef)	se2	Chisq	DF	p
factor(grade)1	-0.4911	0.53146	0.49442	0.85	1	3.6e-01
factor(grade)2	-0.0912	0.44191	0.40534	0.04	1	8.4e-01
factor(grade)3	0.3592	0.44914	0.41014	0.64	1	4.2e-01
positive	0.0551	0.01038	0.00881	28.17	1	1.1e-07
size	0.0155	0.00609	0.00548	6.52	1	1.1e-02
frailty(id, dist = "gauss				89.73	75.7	1.3e-01
Iterations: 5 outer, 19 Newton-Raphson						
Variance of random effect= 0.745						
Degrees of freedom for terms= 2.6 0.7 0.8 75.7						
Likelihood ratio test=216 on 79.8 df, p=1.67e-14						

Στην ερμηνεία των μοντέλων ευπάθειας θα χρειαστούμε επιπλέον κάποιες τιμές των λόγο πιθανοφανειών προκειμένου να αξιολογήσουμε την σημαντικότητα του όρου ευπάθειας για αυτό υπολογίσουμε και της ποσότητες:

frail.1$loglik

[1] -761.8827 -611.5169

cox8$loglik
-762.0015 -735.3921

Ερμηνεία για γάμμα frailty

Προκειμένου να ελέγξουμε την αναγκαιότητα του όρου ευπάθειας στο μοντέλο (7.6) θα εφαρμόσουμε έναν έλεγχο λόγου πιθανοφανειών.

Παρατηρούμε πως ο λογάριθμος της μερικής πιθανοφάνειας (log partial likelihood) όταν το διάνυσμα των παραμέτρων β, ισούται με 0 είναι -762.0015. Για το σύνηθες μοντέλο του Cox από την τελευταία προσαρμογή η ποσότητα log partial likelihood ισούται με -735.3921. Στο γάμμα μοντέλο ευπάθειας (gamma frailty model) η διακύμανση της τυχαίας επίδρασης είναι ίση με 1.62

Ο έλεγχος λόγου πιθανοφανειών για ένα μοντέλο ευπάθειας ορίζεται ώς το διπλάσιο της διαφοράς ανάμεσα στην log partial likelihood του όρου ευπάθειας (στο output της R η ποσότητα αυτή αναφέρεται ως I- likelihood) και της loglikelihood του μοντέλου χωρίς τον όρο. Επομένως έχουμε:

2*(735.3921-732.1) = 6.5842, η τιμή του στατιστικού μας με αντίστοιχη τιμή p-value για α=5% την

```
> pvalue<-1-pchisq((6.5842),1); pvalue
[1] 0.01028878
```

Άρα εφόσον η τιμή p-value = 0.0103 < 0.05 ε.σ, ο όρος ευπάθειας (id) είναι στατιστικά σημαντικός σε σχέση με το χρόνο μέχρι τον θάνατο από καρκίνου του μαστού ή μετάστασής του .

Προκειμένου να εξετάσουμε αν το παραπάνω μοντέλο είναι καλύτερο από το (7.6) εκτιμούμε και τους πίνακες anova, ξεχωριστά για το κάθε μοντέλο.

Πίνακας 7.15

Πίνακας anova για το μοντέλο 7.6

	Df	Deviance	Resid. Df	Resid. Dev
NULL			832	1524.00
factor(grade)	3	17.61	829	1506.40
positive	1	29.93	828	1476.47
size	1	5.68	827	1470.78

Πίνακας 7.16

Πίνακας anova για το μοντέλο 7.6α

	Df	Deviance	Resid. Df	Resid. Dev
NULL			832	1523.77
factor(grade)	3	17.64	829	1506.12
positive	1	30.00	828	1476.12
size	1	5.69	827	1470.43
frailty(id)	0	247.39	827	1223.03

Το μοντέλο (7.6) με βάση το λόγο πιθανοφανειών έχει τιμή για το $\mathbf{X^2}_{827;0.95}$=**895 < 1470.78,** έτσι το μοντέλο (7.6) κρίνεται συνολικά ανεπαρκές.

Προσπαθήσαμε λοιπόν να βελτιώσουμε την προσαρμογή του μοντέλου εισάγοντας ένα τυχαίο παράγοντα όμως και πάλι το συμπέρασμα είναι το ίδιο, το μοντέλο διαφέρει σημαντικά από το κορεσμένο μοντέλο, και το μοντέλο (7.6α) είναι ανεπαρκές αφού $\mathbf{X^2}_{827;0.95}$ **= 895 < 1223.03**, μπορεί η τιμή του στατιστικού να μειώθηκε, αλλά η μείωση αυτή δεν είναι στατιστικά σημαντική. Ακόμη μπορούμε να δούμε πως η σημαντικότητα των μεταβλητών δεν αλλάζει έτσι όπως και στο μοντέλο (7.6) μόνο οι μεταβλητές positive και size ήταν σημαντικές το ίδιο ισχύει και για το μοντέλο (7.6α), παρόλο που παρατηρούμε μεταβολή στην απόκλιση των μεταβλητών positive και size από το ένα μοντέλο στο άλλο.

Ερμηνεία του μοντέλου με όρο ευπάθειας την κανονική κατανομή

Για την δεύτερη εφαρμογή ο όρος ευπάθειας ακολουθεί κανονική κατανομή, με την οποία επιλέγουμε να εκτιμήσουμε την τυχαία επίδραση χρησιμοποιώντας ένα REML κριτήριο. Η προσαρμογή αυτή συνήθως οδηγεί σε μεγαλύτερη εκτίμηση της διακύμανσης του όρου ευπάθειας (όμως κάτι τέτοιο δεν παρατηρείται για τα δεδομένα μας καθώς όπως μπορούμε να το δούμε από την εφαρμογή για τη κανονική

κατανομή η εκτίμηση της διακύμανσης του όρου ευπάθειας ισούται με 0.745 σε σχέση με αυτή της γάμμα κατανομής που είναι μεγαλύτερη και ισούται με 1.62).

Επίσης στο μοντέλο που χρησιμοποιούμε την κανονική κατανομή για την εκτίμηση του όρου πιθανοφάνειας οι τυπικές αποκλίσεις των μεταβλητών αυξάνονται αντί να μειώνονται σε σχέση με την γάμμα frailty, περιμένουμε το μοντέλο αυτό να μην είναι ικανοποιητικό για τα δεδομένα μας.

Αν υπάρχει σημαντική μεταβλητότητα (ετερογένεια), τότε το μοντέλο σταθερών επιδράσεων δίνει πιο μικρά διαστήματα εμπιστοσύνης σε σχέση με την προσέγγιση των τυχαίων επιδράσεων, για αυτό μια εναλλακτική λύση είναι η εφαρμογή της μεθόδου περιορισμένης μέγιστης πιθανοφάνειας (REML).

Η μέθοδος περιορισμένης μέγιστης πιθανοφάνειας (REML) χρησιμοποιεί την πιθανοφάνεια όλων των γραμμικών ανεξάρτητων αντιθέσεων των οποίων η μέση τιμή είναι μηδέν. Η μέθοδος αυτή είναι επαναληπτική αφού πρώτα εκτιμώνται οι επιδράσεις του σταθερού κομματιού του μοντέλου (δηλαδή η παράμετρος β και μετά οι διακυμάνσεις).

Οι διαφορές μεταξύ της μεθόδου μέγιστης πιθανοφάνειας και της REML εντοπίζονται στην μικρότερη μεροληψία της REML στην εκτίμηση των διακυμάνσεων σε σχέση με την ML. Παρόλα αυτά όσο το μέγεθος του δείγματος μεγαλώνει οι διαφορές μεταξύ των δύο μεθόδων είναι λιγότερο σημαντικές. Ακόμη, η REML μέθοδος δεν μπορεί να χρησιμοποιηθεί για ελέγχους της μορφής $H_0 : \beta = 0$, αυτό διότι η REML μέθοδος χρησιμοποιεί την εκτίμηση των παραμέτρων διακύμανσης και υποθέτει πάντοτε τον ίδιο αριθμό παραμέτρων β.

(Σημειώσεις Β. Βασδέκης)

Εφόσον όσο μεγαλύτερο είναι το μέγεθος του δείγματος τόσο μικρότερες είναι και οι διαφορές μεταξύ των δύο όρων ευπάθειας δεν αναμένουμε να παρατηρήσουμε κάτι διαφορετικό στον παρακάτω πίνακα anova για το μοντέλο (7.6β). Πράγματι και το μοντέλο (7.6β) είναι ανεπαρκές και διαφορετικά από ότι ισχυεί συνήθως εκτιμώντας το όρο ευπάθειας με το RELM κριτήριο καταλήξαμε σε αύξηση στην τιμή του στατιστικού από αυτή που πήραμε από μια gamma frailty. Πιθανότατα η εκτίμηση του όρου ευπάθειας χρησιμοποιώντας την κανονική κατανομή ίσως να μην ήταν κατάλληλη για τα δεδομένα μας.

Πίνακας 7.17

Πίνακας anova για το μοντέλο (7.6β)

	Df	Deviance	Resid. Df	Resid. Dev
NULL			832	1523.77
factor(grade)	3	17.64	829	1506.12
positive	1	30.00	828	1476.12
size	1	5.69	827	1470.43
frailty(id, dist = "gauss")	0	162.70	827	1307.72

ΚΕΦΑΛΑΙΟ 8^{Ο}

Συμπεράσματα

Στο τελευταίο αυτό κεφάλαιο παρουσιάζονται συνοπτικά τα πιο σημαντικά αποτελέσματα (του 6^{ου} και 7^{ου} κεφαλαίου) που προέκυψαν από την εφαρμογή της ανάλυσης επιβίωσης πάνω στα δεδομένα που έχουμε στην διάθεσή μας.

Αρχικά αναφέρουμε αποτελέσματα από το κομμάτι της ανάλυσης επιβίωσης που αφορά τον Kaplan – Meier εκτιμητή καθώς και διάφορων τεστ ελέγχου συναρτήσεων επιβίωσης.

Από το γράφημα του Kaplan – Meier εκτιμητή (Γράφημα 6.1) παρατηρούμε ότι πάνω από τους μισούς ασθενείς βρίσκονται εν ζωή την χρονική στιγμή που τελειώνει η έρευνα, το ποσοστό επιβίωσης για κάποια ασθενή εκτιμάται να είναι γενικά μεγαλύτερο του 70%.

Στην συνέχεια εξετάσαμε τόσο γραφικά όσο και στατιστικά αν υπάρχει κάποια διαφορά στην συνάρτηση επιβίωσης για τα διάφορα επίπεδα των υπό εξέταση μεταβλητών.

Απεικονίζοντας τα τέσσερα επίπεδα της μεταβλητής Grade σε ένα κοινό γράφημα (γράφημα 6.2), φαίνεται να υπάρχει σημαντική διαφορά στις συναρτήσεις επιβίωσης των τεσσάρων ομάδων. Συγκεκριμένα ασθενείς με μεγάλη διαφοροποίηση του όγκου έχουν και την μεγαλύτερη πιθανότητα επιβίωσης, ενώ την μικρότερη πιθανότητα επιβίωσης παρουσιάζουν ασθενείς με την χαμηλότερη διαφοροποίηση του όγκου. Τα συμπεράσματα που προκύπτουν από το γράφημα συμφωνούν και με τα αποτελέσματα του ελέγχου των Breslow – Gehan, σύμφωνα με το οποίο απορρίπτουμε την μηδενική υπόθεση υπέρ της εναλλακτικής, άρα η συνολική επιβίωση των ασθενών διαφέρει στα τέσσερα επίπεδα της μεταβλητής Grade.

Εφόσον οι υπόλοιπες διαθέσιμες μεταβλητές είναι είτε συνεχείς είτε διακριτές προχωράμε στην κατηγοριοποίηση τους προκειμένου να έχουμε μια καλύτερη εικόνα και κατανόηση της επίδρασής τους πάνω στην συνάρτηση επιβίωσης.

Έτσι την μεταβλητή που δηλώνει την ηλικία την διακρίνουμε σε τρείς κατάλληλες ομάδες, απεικονίζοντας την σε ένα γράφημα δεν παρατηρούμε σημαντικές διαφορές στις συναρτήσεις επιβίωσης των τριών ομάδων. Αρχικά οι τρείς καμπύλες σχεδόν συμπίπτουν και στην συνέχεια διαφοροποιούνται, δεν περιμένουμε λοιπόν η

μεταβλητή που δηλώνει την ηλικία της ασθενούς την στιγμή της εξέτασης να είναι στατιστικά σημαντική.

Επίσης και την μεταβλητή που δηλώνει το μέγεθος του όγκου την διακρίνουμε σε τρείς ομάδες, από το γράφημα της συνάρτησης επιβίωσής της (γράφημα 6.5) συμπεραίνουμε ότι ίσως μια καλύτερη κατηγοριοποίηση της μεταβλητής θα ήταν τις ομάδες που περιλαμβάνουν ασθενείς με μέγεθος όγκου (21 – 49) και (>=50) να τις συγχωνεύσουμε σε μια, καθώς οι καμπύλες επιβίωσης για τις ομάδες αυτές συμπίπτουν σχεδόν για κάθε χρονική στιγμή. Ασθενείς με μέγεθος όγκου (0 -20 mm) έχουν μεγαλύτερη πιθανότητα επιβίωσης σε σχέση με τις άλλες δύο ομάδες, περιμένουμε λοιπόν η επιβίωση ενός ασθενή να εξαρτάται από το μέγεθος του όγκου.

Οι συναρτήσεις επιβίωσης για τις τρείς ομάδες της μεταβλητής positive φαίνονται να διαφέρουν, και μάλιστα όσο αυξάνεται ο αριθμός των διηθημένων λεμφαδένων τόσο μικραίνει και η πιθανότητα επιβίωσης ή αντίστοιχα αυξάνεται η πιθανότητα θνησιμότητας ή μετάστασης του καρκίνου. Στην ουσία πρόκειται για τρεις διατεταγμένες συναρτήσεις επιβίωσης.

Στην συνέχεια εφαρμόζουμε στρωματοποιημένα τεστ ώστε να μελετήσουμε τον χρόνο επιβίωσης στα επίπεδα μιας συμμεταβλητής λαμβάνοντας υπόψη κάποιον άλλο παράγοντα. Έτσι εξετάζουμε πώς συμπεριφέρεται ο παράγοντας που δηλώνει το μέγεθος του όγκου ξεχωριστά για την κάθε ομάδα διηθημένων λεμφαδένων. Συμπεραίνουμε ότι οι συναρτήσεις επιβίωσης για τα τρία επίπεδα της size διαφέρουν για αριθμό διηθημένων λεμφαδένων >=3 ενώ δεν διαφέρουν για >=4.

Τέλος εφαρμόζουμε το ολικό στρωματοποιημένο τεστ με το οποίο συγκρίνουμε τις συναρτήσεις επιβίωσης των τριών ομάδων λαμβάνοντας υπόψη την μεταβλητή newpositive. Συμπεραίνουμε ότι οι συναρτήσεις επιβίωσης των 3 ομάδων διαφέρουν στατιστικά λαμβάνοντας υπόψη την newpositive. Άρα τα αποτελέσματα των τοπικών ελέγχων δεν συμπίπτουν με αυτά του ολικού ελέγχου.

Στο $7^{ο}$ κεφάλαιο προσαρμόσαμε στα δεδομένα το μοντέλο αναλογικού κινδύνου. Αρχικά θεωρούμε το απλό μοντέλο αναλογικού κινδύνου με μια συμμεταβλητή την grade. Με βάση τα αποτελέσματα των ολικών ελέγχων η μεταβλητή grade είναι στατιστικά σημαντική, αν και από τους τοπικούς ελέγχους οι συντελεστές για τα διάφορα επίπεδα της μεταβλητής δε φαίνεται να διαφοροποιούνται σημαντικά από το μηδέν.

Ακόμη εκτιμώντας κάποιους λόγους κινδύνων συμπεραίνουμε ότι ασθενείς με μεγάλη διαφοροποίηση του όγκου έχουν μικρότερο κίνδυνο θανάτου, σε σχέση με τους ασθενείς που παρουσίασαν μεσαία, μικρή ή άγνωστη διαφοροποίηση του όγκου.

Η μεταβλητή size φαίνεται να επηρεάζει σημαντικά την συνάρτηση κινδύνου και μάλιστα παρατηρώντας το πρόσημο του συντελεστή παλινδρόμησης ($\hat{\beta}$) συμπεραίνουμε ότι το μέγεθος του όγκου επηρεάζει θετικά τον κίνδυνο και εφόσον η τιμή του συντελεστή είναι μεγαλύτερη της μονάδας, αν αυξήσουμε το μέγεθος του όγκου κατά 1 mm περιμένουμε να αυξάνεται σημαντικά και ο κίνδυνος θανάτου.

Παρόμοια είναι και τα συμπεράσματα και για την μεταβλητή positive.

Στην συνέχεια εφαρμόσαμε κάποιες τεχνικές επιλογής μεταβλητών επιλογής μεταβλητών προκειμένου να προσδιορίσουμε τις μεταβλητές καθώς και τις αλληλεπιδράσεις που πράγματι επηρεάζουν την συνάρτηση κινδύνου.

Ως καλύτερο μοντέλο επιλέγεται το:

$$h(t \mid Z) = h_0(t)\exp(\beta_1 grade1 + \beta_2 grade2 + \beta_3 grade3 + \beta_4 positive + \beta_5 size)$$

Καμία αλληλεπίδραση δεν ήταν στατιστικά σημαντική.

Εφαρμόζουμε το παραπάνω μοντέλο, από τους ολικούς ελέγχους έχουμε ότι τουλάχιστον μια από τις παραμέτρους του μοντέλου είναι διαφορετική του μηδενός. Ενώ σύμφωνα με τους τοπικούς ελέγχους μόνο οι μεταβλητές positive και size επηρεάζουν σημαντικά την συνάρτηση κινδύνου, τα επίπεδα της μεταβλητής grade δεν φαίνεται να έχουν κάποια προγνωστική δύναμη.

Για να αξιολογήσουμε αν ισχύει η υπόθεση αναλογικού κινδύνου για το παραπάνω μοντέλο εφαρμόζουμε:

1. **Γραφικές μεθόδους ,**

1.α) log[-log(S(t|Z))] γραφικές παραστάσεις, βέβαια τα αποτελέσματα που παίρνουμε δεν είναι και τόσο αξιόπιστα, μας δίνουν απλώς τις πρώτες ενδείξεις. Για τα συγκεκριμένα δεδομένα τα αποτελέσματα από τα παραπάνω γραφήματα έρχονται σε αντίθεση με αυτά των στατιστικών ελέγχων.

1.β) Schoenfeld residuals, η απεικόνιση των συγκεκριμένων υπολοίπων συναρτήσει του χρόνου, μια τυχαία συμπεριφορά των υπολοίπων αποτελεί ένδειξη ικανοποίηση της υπόθεσης αναλογικού κινδύνου. Για τα διαθέσιμα δεδομένα έχουμε ενδείξεις ότι μόνο οι μεταβλητές positive και size ικανοποιούν την υπόθεση αναλογικού κινδύνου.

1.γ) Scaled Schoenfeld residuals + smoothing curve, η μηδενική κλίση της οποίας αποτελεί ένδειξη ικανοποίησης της υπόθεσης αναλογικού κινδύνου.

2. Στατιστικούς ελέγχους, τα Scaled Schoenfeld residuals μας δίνουν την δυνατότητα να εφαρμόσουμε στατιστικό τεστ προκειμένου να ελέγξουμε, τόσο ξεχωριστά για την κάθε μεταβλητή όσο και ταυτοχρόνως, αν οι υπό μελέτη μεταβλητές ικανοποιούν την υπόθεση αναλογικού κινδύνου.
Όλες οι μεταβλητές που περιλαβάνονται στο μοντέλο ικανοποιούν την υπόθεση αναλογικού κινδύνου τόσο ξεχωριστά όσο και ταυτοχρόνος.

Τέλος, εφαρμόζουμε τα μοντέλα ευπάθειας, προσθέτουμε ένα τυχαίο όρο στο αρχικό μοντέλο με σκοπό να αντιμετωπίσουμε την:

1) Έλλειψη προσαρμογής του μοντέλου
2) Απομάκρυνση από την αναλογικότητα
3) Παράλειψη κάποιας σημαντικής μεταβλητής

Όμως εφαρμόζοντας είτε γάμμα frailty model είτε κανονικό frailty model, καταλήγουμε στο συμπέρασμα ότι το μοντέλο μας είναι ανεπαρκές. Η προσθήκη λοιπόν του τυχαίου όρου δεν επιφέρει στατιστικά σημαντικές αλλαγές στην προσαρμογή του μοντέλου.

ΠΑΡΑΡΤΗΜΑ

Εφαρμογή στην Ανάλυση Επιβίωσης με εντολές του προγράμματος R

Φορτώνουμε τα δεδομένα στην κονσόλα του R μέσω της εντολής:

a<-read.table("C:/Users/Alexandra/Desktop/iaso.txt",header=TRUE).

Για την εκτίμηση του Kaplan – Meier εκτιμητή χωρίς να λάβουμε υπόψη κάποια συμμεταβλητή θα χρησιμοποιήσουμε την εντολή **survfit(),** αφού πρώτα φορτώσουμε τα πακέτα **splines** και **survival,** τα αποτελέσματα που παίρνουμε ακολουθούν:

afit<-survfit(Surv(MONTH,DEATH)~1,data=a)

summary(afit)

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
2	861	2	0.998	0.00164	0.994	1.000
3	849	1	0.997	0.00202	0.993	1.000
5	830	1	0.995	0.00234	0.991	1.000
7	806	1	0.994	0.00265	0.989	0.999
. . .	….	….	…	…	…	…
169	122	1	0.741	0.02362	0.697	0.789
171	119	1	0.735	0.02423	0.689	0.784
183	94	1	0.727	0.02520	0.680	0.779
191	81	1	0.718	0.02644	0.668	0.772

Η παραπάνω διαδικασία μας δίνει με την σειρά που αναφέρονται, τους διαφορετικούς πλήρους χρόνους διατεταγμένους , το σύνολο των ατόμων που βρίσκονται σε κίνδυνο τη χρονική στιγμή t_j, το πλήθος των αποτυχιών την στιγμή t_j, την πιθανότητα επιβίωσης πέραν της χρονικής στιγμής t_j, την τυπική απόκλιση της επιβίωσης υπολογισμένη με βάση το τύπο του Greenwood (βλ. σχέση (3.2)) και τέλος έχουμε το 95% διάστημα εμπιστοσύνης για την συνάρτηση επιβίωσης για κάθε $t \in (t_j, t_{j+1}]$.

Στην περίπτωση που θέλουμε να μελετήσουμε τον πίνακα επιβίωσης για κάθε επίπεδο μιας συμμεταβλητής τότε αντί της μονάδας εισάγουμε το όνομα της συμμεταβλητής.

Έτσι ο πίνακας επιβίωσης για κάθε επίπεδο της μεταβλητής GRADE θα προκύπτει ως εξής:

```
afit1<-survfit(Surv(MONTH,DEATH)~GRADE,conf.type="none" ,data=a)
summary(afit1)
 28 observations deleted due to missingness
                GRADE=0
 time    n.risk  n.event  survival std.err
   5       44       1         0.977    0.0225
  10       42       1         0.954    0.0318
  18       39       1         0.930    0.0393
……
                GRADE=1
 time   n.risk       n.event     survival    std.err
  28     99             1          0.990     0.0100
  37     95             1          0.979     0.0144
  39     93             1          0.969     0.0177
……
               GRADE=2
 time   n.risk   n.event    survival      std.err
   3     426       1          0.998       0.00234
   7     411       1          0.995       0.00337
  12     394       1          0.993       0.00420
……
               GRADE=3
 Time  n.risk   n.event    survival      std.err
   2     246        2         0.992       0.00573
  11     213        1         0.987       0.00735
  12      209       1         0.982       0.00870
……..
```

Για την μεταβλητή newage έχουμε:

kmnewage<-survfit(Surv(MONTH,DEATH)~NEWAGE,data=a)

summary(kmnewage)

NEWAGE=1

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
10	109	1	0.991	0.00913	0.973	1.000
14	101	1	0.981	0.01331	0.955	1.000
18	96	1	0.971	0.01663	0.939	1.000
22	92	1	0.960	0.01952	0.923	0.999
24	90	1	0.950	0.02202	0.907	0.994
26	88	1	0.939	0.02427	0.892	0.988
….	...	...	....	....	....	....

NEWAGE=2

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
2	441	2	0.995	0.00320	0.989	1.000
11	397	1	0.99	0.00406	0.985	1.000
13	394	1	0.990	0.00477	0.981	1.000
16	388	1	0.988	0.00539	0.977	0.999
18	380	1	0.985	0.00597	0.974	0.997
...	...	...	....	....	....	....

NEWAGE=3

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
3	296	1	0.997	0.00337	0.990	1.000
5	289	1	0.993	0.00481	0.984	1.000
7	278	1	0.990	0.00597	0.978	1.000
12	267	2	0.982	0.00790	0.967	0.998
14	257	1	0.978	0.00875	0.961	0.996
....	....	....	....	.....	....	....

Ακολουθεί ο κώδικας που χρειάζεται να εφαρμόσουμε στην R προκειμένου να πάρουμε τις εκτιμήσεις για τις ελλειπούσες τιμές της μεταβλητής positive, τις οποίες στην συνέχεια αντικαθιστούμε στην μεταβλητή και εφαρμόζουμε από εκεί και πέρα τις κατάλληλες εντολές για να κατηγοριοποιήσουμε την positive, με παρόμοιο τρόπο όπως κάναμε και για τις μεταβλητές size και age.

```
  x=cbind(a$month,a$age,a$death,a$size)
y=a$positive
  dat.xy<-data.frame(x,y)
imp=mi.categorical(formula=y~x,data=dat.xy);imp
```

παίρνουμε τις εξής εκτιμήσεις για τις ελλείπουσες τιμές του y, όπου το y = a$positive.

```
y[43]=20;y[75]=0;y[240]=0;y[468]=0;y[493]=0;y[576]=4;y[617]=12;y[622]=6;
y[750]=4;y[852]=9
  newpositive<-y
  for(i in 1:length(y)){
   if(y[i]>=0&y[i]<=3){newpositive[i]=1} else
   if(y[i]>=4&y[i]<9){newpositive[i]=2} else
   if(y[i]>=9) {newpositive[i]=3}
  }
```

Στην συνέχεια θα κατασκευάσουμε για κάθε ομάδα της newage τους πίνακες επιβίωσης των ασθενών για τα τέσσερα επίπεδα της GRADE.

```
  sfit1<-survfit (Surv(MONTH,DEATH)~GRADE ,subset=NEWAGE==1,
  conf.type="none",data=a); summary(sfit1)
```

GRADE=0

time	n.risk	n.event	survival	std.err
10	4	1	0.75	0.217
36	3	1	0.50	0.250
…	…	…	…	…

GRADE=1

time	n.risk	n.event	survival	std.err
46	12	1	0.917	0.0798
65	11	1	0.833	0.1076
…	…	…	…	…

GRADE=2

time	n.risk	n.event	survival	std.err
32	38	1	0.974	0.0260

33 37 1 0.947 0.0362

… … … … …

GRADE=3

time n.risk n.event survival std.err

14 43 1 0.977 0.0230

18 38 1 0.951 0.0338

… … … … …

sfit2<-survfit (Surv(MONTH,DEATH)~GRADE,subset=NEWAGE==2, conf.type =”none”,data=a);summary(sfit2)

GRADE=0

time n.risk n.event survival std.err

18 26 1 0.962 0.0377

70 16 1 0.901 0.0681

… … … …. ….

GRADE=1

time n.risk n.event survival std.err

73 48 1 0.979 0.0206

142 27 1 0.943 0.0407

… … … … …

GRADE=2

time n.risk n.event survival std.err

13 202 1 0.995 0.00494

24 191 1 0.990 0.00715

… … … …. …

GRADE=3

time n.risk n.event survival std.err

2 110 2 0.982 0.0127

11 94 1 0.971 0.0163

… … … … ….

sfit3<-survfit(Surv(MONTH,DEATH)~GRADE,subset=NEWAGE==3, conf.type=”none”, data=a); summary(sfit3)

GRADE=0

time	n.risk	n.event	survival	std.err
5	14	1	0.9286	0.0688

GRADE=1

time	n.risk	n.event	survival	std.err
28	26	1	0.962	0.0377
37	23	1	0.920	0.0545
…	…	…	…	…

GRADE=2

time	n.risk	n.event	survival	std.err
3	161	1	0.994	0.00619
7	154	1	0.987	0.00890
…	….	…	….	….

GRADE=3

time	n.risk	n.event	survival	std.err
12	72	1	0.986	0.0138
14	69	1	0.972	0.0196
…	…	…	…	…

Οι εντολές που χρειάζεται να εφαρμόσουμε στην R για να αποδείξουμε και γραφικά τα αποτελέσματα της στρωματοποίησης της μεταβλητής grade για το κάθε στρώμα της newage είναι οι εξής:

```
par(mfcol=c(2,2))
plot(sfit1,xlab=”Time in months”,ylab=”Survival”,main=”Newage=1(age(21-40))”,
lty=4:1,col=c(“red”,”green”,”yellow”,”black”))
 legend(20,0.6,c(“den einai gnosto”,”megali”,”mesaia”,”mikri”),lty=2:2,
col=c(“red”,”green”,”yellow”,”black”))
plot(sfit2,xlab=”Time in months”,ylab=”Survival”,main=”Newage=2(age(41-59))”,
lty=4:1,col=c(“red”,”green”,”yellow”,”black”))
legend(20,0.6,c(“den einai gnosto”,”megali”,”mesaia”,”mikri”),lty=2:2,
col=c(“red”,”green”,”yellow”,”black”))
plot(sfit3,xlab=”Time in months”,ylab=”Survival”,main=”Newage=3(age >=60)”,
```

lty=4:1,col=c(“red”,”green”,”yellow”,”black”))

legend(20,0.6,c(“den einai gnosto”,”megali”,”mesaia”,”mikri”),

lty=2:2, col=c(“red”,”green”,”yellow”,”black”))

par(mfcol=c(1,1))

Έχει ενδιαφέρον να δούμε και μια ακόμα στρωματοποίηση της συμμεταβλητής newpositive για τα τρία επίπεδα της μεταβλητής newsize. Αρχικά εκτελούμε τα Breslow – Gehan test χρησιμοποιώντας το πρόγραμμα spss, στα επίπεδα της newage για κάθε στρώμα της newpositive.

Στην συνέχεια κατασκευάζουμε για κάθε επίπεδο της newpositive τους πίνακες επιβίωσης των ασθενών στα τρία επίπεδα της newsize:

k.m1<-

survfit(Surv(MONTH,DEATH)~NEWSIZE,subset=newpositive==1,data=a);

summary(k.m1)

NEWSIZE=1

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
3	378	1	0.997	0.00264	0.992	1.000
7	359	1	0.995	0.00383	0.987	1.000
10	343	1	0.992	0.00479	0.982	1.000
13	333	1	0.989	0.00562	0.978	1.000
...	....	...	....	....	.....	....

NEWSIZE=2

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
5	226	1	0.996	0.00441	0.987	1.000
12	214	1	0.991	0.00639	0.978	1.000
18	207	1	0.986	0.00795	0.971	1.000
27	189	2	0.976	0.01076	0.955	0.997
28	187	1	0.970	0.01190	0.947	0.994
...	...	....	....	....	.....	....

NEWSIZE=3

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
2	38	1	0.974	0.0260	0.924	1.000
17	31	1	0.942	0.0398	0.867	1.000
34	27	1	0.907	0.0514	0.812	1.000

36	26	1	0.872	0.0601	0.762	0.999
45	25	1	0.838	0.0671	0.716	0.980
...	....	...	....	....	.....	....

Αντίστοιχα αποτελέσματα έχουμε και για newpositive==2 και newpositive==3.

Αποτελέσματα για την μεταβλητή που δηλώνει τον αριθμό των διηθημένων λεμφαδένων.

```
cox5<-coxph(Surv(MONTH,DEATH)~POSITIVE, method="breslow",data=a)
summary(cox5)
```

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	exp(-coef)	l .95	u.95
POSITIVE	0.058012	1.059727	0.007712	7.522	5.4e-14 ***	0.9436	1.0444	1.076

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

Rsquare= 0.044 (max possible= 0.83)

Likelihood ratio test = 38.76 on 1 df, p=4.799e-10

Wald test = 56.58 on 1 df, p=5.396e-14

Score (logrank) test = 63.81 on 1 df, p=1.332e-15

Παρατηρούμε ότι υπάρχει η τάση να αυξάνεται ο κίνδυνος θανάτου όταν αυξήσουμε τον αριθμό των διηθημένων λενφαδένων, η τάση αυτή είναι στατιστικά σημαντική αφού το p-value του Wald test είναι σχεδόν μηδέν.

Εφαρμογή του μοντέλου που επιλέχτηκε ως το πλέον κατάλληλο για τα δεδομένα μας:

```
cox8<-coxph(Surv(MONTH,DEATH)~factor(GRADE)+SIZE+POSITIVE,
method="breslow",data=a)
summary(cox8)
```

n=832 (38 observations deleted due to missingness

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
factor(grade)1	-0.473364	0.622903	0.492140	-0.962	0.3361
factor(grade)2	-0.064043	0.937965	0.403728	-0.159	0.8740
factor(grade)3	0.292124	1.339270	0.409848	0.713	0.4760

size 0.013637 1.013731 0.005478 2.489 0.0128 *

positive 0.045229 1.046267 0.008399 5.385 7.24e-08 ***

	exp(coef)	exp(-coef)	lower .95	upper .95
factor(grade)1	0.6229	1.6054	0.2374	1.634
factor(grade)2	0.9380	1.0661	0.4251	2.069
factor(grade)3	1.3393	0.7467	0.5998	2.990
size	1.0137	0.9865	1.0029	1.025
positive	1.0463	0.9558	1.0292	1.06

Rsquare= 0.062 (max possible= 0.84)

Likelihood ratio test= 53.22 on 5 df, p=3.033e-10

Wald test = 69.67 on 5 df, p=1.198e-13

Score (logrank) test = 77.7 on 5 df, p=2.554e-15

Με βάση τα p-value των Wald ελέγχων για την σημαντικότητα της κάθε μεταβλητής στο μοντέλο βλέπουμε, πως κανένα από τα επίπεδα της μεταβλητής grade για ε.σ. 5% δεν είναι στατιστικά σημαντικό, αντίθετα οι μεταβλητές positive και size είναι σημαντκές. Η επίδραση των μεταβλητών positive και size πάνω στην συνάρτησηη κινδύνου είναι θετική, ενώ η επίδραση της μεταβλητής grade είναι αρνητική και αυτό είναι αναμενόμενο διότι μεγαλύτερη συρρίκνωση του όγκου σημαίνει μικρότερο κινδύνο θανάτου ή μετάσταση του καρκίνου.

Τιμές Schoenfeld υπολοίπων:

```
cox8<-coxph(Surv(MONTH,DEATH)~factor(GRADE)+POSITIVE+SIZE,
method="breslow",data=a)
schoen.res.<-residuals(cox8,type="schoenfeld"); schoen.res.
```

	factor(GRADE)1	factor(GRADE)2	factor(GRADE)3	POSITIVE	SIZE
2	-0.06471380	-0.4137160	0.5315636	8.06759686	2.67243300
2	-0.06471380	-0.4137160	0.5315636	-8.93240314	27.67243300

3	-0.06507276	0.5807107	-0.4626698	-7.98366558	-12.37631325
5	-0.06520759	-0.4201077	-0.4607660	-9.07744676	2.50433313
7	-0.06520337	0.5745411	-0.4547007	-8.97146759	-17.32130772
10	-0.06530911	-0.4265141	-0.4531167	-8.97991935	-12.41889251
11	-0.06586525	-0.4260006	0.5464994	32.97310114	-2.51380480
…	….	….	…	…	…..

Εκτιμήσεις των Scaled Schoenfeld υπολοίπων τα οποία χρησιμοποιούνται για την αξιολόγηση της υπόθεσης αναλογικού κινδύνου από την κάθε μεταβλητή του μοντέλου αλλά και ταυτοχρόνως, συντελούν και στον έλεγχο της ολικής επάρκειας του μοντέλου.

```
cox8<-coxph(Surv(MONTH,DEATH)~factor(GRADE)+POSITIVE+SIZE,
method="breslow",data=a)
scaled.res.<-residuals(cox8,type="scaled");scaled.res.
```

	[,1]	[,2]	[,3]	[,4]	[,5]
2	-0.406978159	-0.13888995	2.352517914	0.0870173185	0.0002156517
2	1.182185590	1.17943812	3.848797855	-0.1003808308	0.1195043313
3	-0.718783706	1.86806262	0.188888145	0.0119315429	-0.0094190516
5	-18.314687978	-17.87860931	-17.527088951	-0.0010738991	0.0082689480
7	-0.983315179	1.70603943	0.221375426	0.0102552351	-0.0266064127
10	-19.190659896	-18.46819888	-17.742112630	0.0218134185	-0.0479920744
11	-0.850411402	-0.82187607	0.762252385	0.3136921505	-0.0569087302
12	-0.962361029	1.65368254	0.005208132	0.0459158979	-0.0303001163
...	....	.....	.....	.....	.....

Τιμές των Scaled score υπολοίπων, τα οποία χρησιμοποιούμε προκειμένου να εντοπίσουμε τις μεταβλητές που συντελούν περισσότερο στην διαμορφώση των συντελεστών παλινδρόμησης, β_i.

```
db.res.<-residuals(cox8,type="dfbeta"); db.res.
```

	[,1]	[,2]	[,3]	[,4]	[,5]
1	2.212120e-03	-2.421978e-03	1.627191e-04	6.395136e-06	1.121056e-04
2	1.110798e-04	4.394894e-05	-8.058366e-04	2.338439e-05	6.607870e-06
3	1.199454e-03	2.048986e-03	2.380292e-02	-7.026718e-04	2.105059e-04
4	-7.332397e-04	2.074733e-05	1.251669e-05	7.840358e-07	2.667028e-06
5	2.838949e-04	-9.278782e-04	-3.051364e-05	1.414154e-05	2.006875e-05

6	-5.156906e-03	-3.711315e-03	-1.195506e-02	3.176840e-04	-3.399429e-04
7	4.687356e-02	4.672251e-02	4.638571e-02	1.471501e-04	-8.404073e-05
8	6.993281e-04	-1.851985e-03	-9.476224e-05	1.901372e-05	4.328588e-05
9	-3.840997e-03	2.268841e-04	6.981498e-05	-1.850845e-06	2.305738e-05
10	4.038138e-05	-3.567393e-06	-9.535116e-04	2.697664e-05	2.424025e-06
11	6.268520e-05	-1.458526e-03	-9.932745e-05	2.134659e-05	6.673613e-06
12	-1.817300e-02	4.270920e-04	6.177691e-04	-8.272972e-05	3.628259e-05
	...	....	....	....	....

Τέλος δίνονται τα Martingale υπόλοιπα που χρησιμοποιούνται προκειμένου να προσδιορίσουμε την συναρτησιακή μορφή μιας μεταβλητής, οι στηλές δηλώνουν τις μεταβλητές του μοντέλου ενώ οι γραμμές αντιστοιχούν στις ασθενείς.

Score<-residuals(cox8,type=”score”); score

	factor(GRADE)1	factor(GRADE)2	factor(GRADE)3	POSITIVE	SIZE
1	0.0221729458	-0.1135885819	0.0795708304	1.232760e+00	4.721385e+00
2	0.0022385547	0.0145111425	-0.0185659061	2.690186e-01	2.218422e-01
3	-0.0812578353	-0.4263943251	0.5547865862	-5.581660e+00	8.099339e+00
4	-0.0078117796	0.0036360712	0.0037111729	7.239626e-02	1.869668e-01
5	0.0050402344	-0.0354706987	0.0266581696	4.954396e-01	1.042620e+00
6	0.0360667203	0.2054873928	-0.2659481114	9.370017e-01	-1.250823e+01
7	0.0326034710	0.1610362629	0.1210213243	2.547909e+00	-4.552322e+00
8	0.0116644746	-0.0701766037	0.0506471964	8.474708e-01	2.111477e+00
9	-0.0422131938	0.0216217236	0.0180046882	3.378848e-01	1.265380e+00
10	0.0026387307	0.0170748432	-0.0218477534	2.757696e-01	6.635297e-02
11	0.0069650726	-0.0475683679	0.0355313895	5.755624e-01	6.859832e-01

Βιβλιογραφία

Ελληνική Βιβλιογραφία

1. Δημήτριος Αντζουλάκος, Ανάλυση Επιβίωσης , Σημειώσεις παραδόσεων, Πρόγραμμα μεταπτυχιακών σπουδών στην Εφαρμοσμένη Στατιστική, 2009
2. Γιάννης Π. Φύσσας , Ο μαστός και οι παθήσεις του, εκδόσεις ΛΙΒΑΝΗ, 2006
3. Β. Βασδέκης , Πανεπιστημιακές σημειώσεις στις επαναλαμβανόμενες μετρήσεις, Τμήμα στατιστικής Οικονομικό Πανεπιστήμιο Αθηνών.
4. Κ. Φωκιανός , Χ. Χαραλάμπους , University of Cyprus, Κεφάλαιο 19^{ο} Ανάλυση Επιβίωσης.
5. Αλέξανδρος Α. Ξυραφά, Διπλωματική εργασία με θέμα την Διαχρονική μεταβολή επιβίωσης παρουσία συγχυτικών παραγόντων: Στατιστικές μέθοδοι και εφαρμογή τους στον προχωρημένο καρκίνο του μαστού.
6. Παπαδάκης, Μ. και Κ. Τσίμπος (2004) Δημογραφική Ανάλυση, Αρχές, Μέθοδοι, Υποδείγματα, Αθήνα, εκδ. Α. Σταμούλης.

Ξένη Βιβλιογραφία

7. David Collett, Modelling Survival Data in Medical Research (second edition), 2005
8. Terry M. Therneau, Patricia M. Grambsch, Modeling Survival Data Extending the Cox Model , New York Springer, 2000.
9. Boist 515 lecture 17, Cox proportional hazards model, March 2004.
10. Eric V. Slud, Actuarial mathematics and life table statistics, University of Maryland, College Park, March 2009.
11. Kristin Sainani Ph.D, Cox Regression 2, http://www.stanford.edu/~kcobb, Stanford University, Department of Health Research and Policy.
12. Brenda Gillespie Ph.D, Cox Proportional Hazards Regression Model, October 2006, University of Michigan.

13. Kenneth R. Hess, Graphical Methods for Assessing Violations oh the Proportional Hazards Assumption in Cox Regression, Statistics In Medicine, Vol.14, p. 1707-1723, 1995.

14. John D. Kalbfleisch and Ross L. Prentice, The Statistical Analysis of Failure Time Data, Second Edition, Wiley series in probability and statistics, 2002.

15. Mark Stevenson, An Introduction to Survival Analysis, EpiCentre, IVABS, Massey University, June 4, 2009.

16. Steven Buyske, Richard Fagerstrom and Zhiliang Ying, A Class of Weighted Log – Rank tests for Survival Data When the Event is Rare, Journal of the American Statistical Association, Vol. 95, No. 449 (Mar., 2000), pp. 249-258

17. David P. Harrington and Thomas R. Fleming, A Class of Rank Procedures for Censored Survival Data, Biometrika, Vol.69, No.3 (Dec., 1982), pp. 553- 566

18. A. Perperoglou, Saskia le Cessie, H.C. van Houwelingen, A fast routine for fitting Cox models with time varying effects of the covariates, 2005, journal ELSEVIER.

19. Daowen Zhang, Right Censoring and Kaplan-Meier Estimator (Chapter 2), St745.

20. Alex Cook, Kaplan-Meier estimate of S(t), ST3242, Introduction to Survival Analysis, August 2008.

21. E.L. Kaplan & P. Meier, Nonparametric Estimation from Incomplete Observations, Jasa 1958.

22. David. D. Hanagal, Modelling Survival Data Using Frailty Models, 2011, CRC press Taylor & Francis group, A chapman & Hall book.

23. DW Hosmer, Jr., S Lemeshow, Applied Survival Analysis: Regression Modeling of Time to Event Data, New York: John Wiley, 1999

24. Aris Perperoglou1, Antonis Keramopoullos and Hans C. van Houwelingen, Approaches in modelling long-term survival: An application to breast cancer, STATISTICS IN MEDICINE *Statist. Med.* 2007; **26**:2666–2685
Published online 30 October 2006 in Wiley InterScience (www. interscience. wiley. com)

25. Andreas Wienke, **Frailty Models,** MPIDR WORKING PAPER WP 2003-032 SEPTEMBER2003

Ιστοσελίδες

1. **http://www.bestrong.org.gr/el/learncancer**
2. http://www.karkinos24.gr/index.php/giatonkarkino
3. **http://www.medinfo.gr/?cat_id=343&article_id=1844**
4. http://www.mastology.gr
5. **http://www.karkinos24.gr/index.php/karkinostoumastou**
6. **http://www.labtestsonline.gr/condition/Condition_BreastCancer.html**
7. http://www.labtestsonline.gr/condition/Condition_BreastCancer.html
8. http://www.gyn.gr/index.php?option=com_content&view=article&id=437:2011-03-03-20-53-10&catid=23:2010-06-11-16-48-52&Itemid=157
9. http://www.teleiosgamos.gr/component/content/1482?task=view
10. **http://www.medinfo.gr/?cat_id=343&article_id=3009**
11. **http://www.mastologos.gr/breast-cancer/male-breast-cancer**