ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Δ.Π.Μ.Σ. ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανάλυση Επιβίωσης με εφαρμογή σε δεδομένα σχετικά με τον καρκίνο του μαστού

Αλεξάνδρα Παπαθανασοπούλου

Α.Μ.:09419018

Επιβλέπουσα Καθηγήτρια: Καρώνη Χρυσηΐς

Επιτροπή Καθηγητών:

Χ. Καρώνη,	Β. Παπανικολάου,	Ι. Πολυράκης
Καθηγήτρια Ε.Μ.Π.,	Καθηγητής Ε.Μ.Π.,	Ομότιμος Καθηγητής Ε.Μ.Π.

Αθήνα,

Ιούλιος 2021

Ευχαριστίες

Θα ήθελα να εκφράσω τη βαθειά μου ευγνωμοσύνη στην καθηγήτρια της Σχολής Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών του Εθνικού Μετσόβιου Πολυτεχνείου, κυρία Καρώνη Χρυσηΐς για την εμπιστοσύνη που μου έδειξε αναθέτοντάς μου την εκπόνηση της παρούσας διπλωματικής εργασίας. Η υποστήριξη κι η υπομονή της με ενθάρρυναν καθ' όλη τη διάρκεια της συγγραφής, ενώ η καθοδήγησή της στάθηκε πολύτιμος αρωγός για την ολοκλήρωση της συγκεκριμένης μελέτης.

Παπαθανασοπούλου Αλεξάνδρα

Αθήνα, Ιούλιος 2021

Περίληψη

Μια εγκυμοσύνη στην οποία υπήρξε μιαν απρόβλεπτη επιπλοκή, ένας ασθενής με χρόνιο νόσημα που ξαφνικά υποτροπίασε η υγεία του ενώ βρισκόταν υπό την επήρεια κάποιας προφυλακτικής θεραπείας, ένα ανεπιθύμητο κάταγμα που συνέβη απροσδόκητα στη μέση ενός αγώνα ποδοσφαίρου ή ένα φαινομενικά αιφνίδιο έμφραγμα σε έναν άντρα μέσης ηλικίας αποτελούν γεγονότα, στα οποία θα ήταν ενδιαφέρον να μελετηθούν οι παράγοντες που συνέβαλαν στη χρονική διάρκεια από την έναρξη της κάθε κατάστασης μέχρι να προκληθεί για πρώτη φορά το γεγονός, με σκοπό την έγκαιρη πρόληψη στο μέλλον.

Η Ανάλυση Επιβίωσης είναι ο κλάδος της στατιστικής που ασχολείται με τη μελέτη του χρόνου μέχρις ότου προκύψει ένα (συνήθως ανεπιθύμητο) γεγονός. Βασίζεται σε πολλές κατανομές με ή χωρίς παραμέτρους, ενώ βρίσκει μια πληθώρα εφαρμογών σε δεδομένα διάρκειας ζωής που επηρεάζονται από διάφορες μεταβλητές. Αξιοποιείται κυρίως σε βιοϊατρικά στοιχεία μέσω της βιοστατιστικής, χωρίς όμως αυτό να την κάνει λιγότερο χρήσιμη σε άλλους επιστημονικούς τομείς, όπως τη μηχανολογία.

Στην παρούσα εργασία αναλύονται όλες οι βασικές έννοιες και μέθοδοι της Ανάλυσης Επιβίωσης μέσω βιβλιογραφικής ανασκόπησης και στη συνέχεια εφαρμόζονται πάνω σε ένα δείγμα ασθενών με καρκίνο στο μαστό για την εξαγωγή συμπερασμάτων. Συγκεκριμένα μελετάται η επιρροή διαφόρων συμμεταβλητών στην επιβίωση γυναικών με όγκο στο στήθος μέχρι να υποτροπιάσει η υγεία τους ή να αποβιώσουν. Η στατιστική ανάλυση βασίζεται στο στατιστικό πακέτο Minitab, καθώς και στη γλώσσα προγραμματισμού R.

Στα γενικά συμπεράσματα στο τέλος του εγγράφου αναφέρονται ποιες συγκεκριμένες μεταβλητές συνέβαλαν στην επιδείνωση της νόσου και ποιες δεν την επηρέασαν ιδιαίτερα, ενώ παράλληλα συστήνεται περαιτέρω διερεύνηση των αποτελεσμάτων συνυπολογίζοντας τα νέα επιτεύγματα της ιατρικής πάνω στον ογκολογικό τομέα.

Λέξεις – Κλειδιά: Ανάλυση Επιβίωσης, Καρκίνος του Μαστού, Εκτιμήτρια Kaplan – Meier, Μοντέλο Επιταχυνόμενης Διακοπής της Weibull, Μοντέλο Αναλογικής Διακινδύνευσης του Cox.

Abstract

A pregnancy in which there was an unpredictable complication, a patient with a chronic illness who suddenly relapsed while he was under the influence of some preventive treatment, an unwanted fracture that occurred unexpectedly in the middle of a football match or a sudden heart attack in a middle – aged man are events in which it would be interesting to study the factors that contributed to the time from the onset of each situation to the first occurrence of the event, with a view to timely prevention in the future.

Survival Analysis is the branch of statistics that deals with the study of time until an (usually undesirable) event occurs. It relies on many distributions with or without parameters, and finds a variety of applications in lifetime data that are affected by various variables. It is mainly used in biomedical data through biostatistics, without though making it less useful in other scientific fields, such as engineering.

This paper analyzes all the basic concepts and methods of Survival Analysis through a literature review and then applies them to a sample of breast cancer patients to draw conclusions. Specifically, it investigates the influence of various covariates on the survival of women with breast tumor until their health has relapsed or they have passed away. The statistical analysis is based on the Minitab statistical package, as well as the R programming language.

The general conclusions at the end of the paper indicate which specific variables contributed to the aggravation of the disease and which did not particularly affect it, while further investigation of the results is recommended, taking into account the new achievements of medicine in the field of oncology.

Keywords: Survival Analysis, Breast Cancer, Kaplan – Meier Estimator, Weibull Accelerated Failure Model, Cox Proportional Hazards Model

Περιεχόμενα

Κατάλογος Γραφικών Παραστάσεων

Κατάλογος Πινάκων

1 Εισαγωγή

Σκοπός αυτής της εργασίας είναι να εφαρμόσει μοντέλα διάρκειας ζωής με συμμεταβλητές, μέσω της Ανάλυσης Επιβίωσης σε δεδομένα που σχετίζονται με τον καρκίνο του μαστού σε γυναίκες για την εξαγωγή συμπερασμάτων. Η μέθοδος της στατιστικής ανάλυσης που θα ακολουθηθεί θα γίνει μέσω του στατιστικού πακέτου 'Minitab', καθώς και της γλώσσας προγραμματισμού ανοικτού κώδικα R και θα βασιστεί κυρίως στο μοντέλο αναλογικής διακινδύνευσης του Cox, καθώς και σε μοντέλα παλινδρόμησης με κυριότερο εκείνο της κατανομής Weibull.

Παρακάτω διευκρινίζονται συνοπτικά κάποιες βασικές ορολογίες.

Ανάλυση Επιβίωσης ορίζεται ως η ανάλυση δεδομένων διάρκειας ζωής σε βιο-ϊατρικές εφαρμογές, η οποία ασχολείται με τη μελέτη του χρόνου μέχρις ότου προκύψει ένα γεγονός, συνήθως δυσάρεστο, όπως η υποτροπή στην υγεία ενός ασθενούς ή ο θάνατός του.

Μοντέλα διάρκειας ζωής είναι συνεχείς και διακριτές κατανομές, οι οποίες αποδεικνύονται ιδιαίτερα χρήσιμες σε πρακτικές εφαρμογές. Στην παρούσα εργασία δίνεται ιδιαίτερη έμφαση στην κατανομή Weibull, η οποία θεωρείται ως η πιο διαδεδομένη στην Ανάλυση Επιβίωσης, μιας και αποτελεί ένα πολύ ικανοποιητικό μοντέλο για πολλές εφαρμογές.

Μοντέλα διάρκειας ζωής με συμμεταβλητές ορίζονται ως τα μοντέλα, των οποίων η διάρκεια ζωής των μονάδων τους εξαρτάται από διάφορους μετρήσιμους παράγοντες, όπως είναι η ηλικία ή τα φάρμακα που λαμβάνει κάποιος ασθενής.

Η παλινδρόμηση είναι μια τεχνική που χρησιμοποιείται για τη μοντελοποίηση και την ανάλυση αριθμητικών δεδομένων, μιας εξαρτημένης μεταβλητής και κάποιων ανεξάρτητων μεταβλητών. Το μοντέλο είναι μια συνάρτηση συσχέτισης της εξαρτημένης μεταβλητής από τις ανεξάρτητες. Η μοντελοποίηση μπορεί να γίνει χωρίς να είναι γνωστή από πριν η γνώση για τον τρόπο με τον οποίο συνδέεται η εξαρτημένη μεταβλητή από τις ανεξάρτητες. Στη γραμμική παλινδρόμηση, η απαίτηση του μοντέλου που θα παραχθεί είναι η εξαρτημένη μεταβλητή y_i να είναι ένας γραμμικός συνδυασμός των ανεξαρτήτων μεταβλητών x_{ij} για $i = 1, 2, ..., n$ και $j = 1, 2, ..., k$.

Η εύρεση της σχέσης μεταξύ μιας μεταβλητής που δηλώνει το χρόνο επιβίωσης ενός ατόμου και άλλων συμμεταβλητών επιτυγχάνεται συνήθως μέσω ενός μοντέλου παλινδρόμησης. Όταν υπάρχουν αποκομμένα δεδομένα επιβίωσης (αναλύονται στην ενότητα '2 Ανάλυση Επιβίωσης') χρησιμοποιούνται όλα τα μοντέλα για την ανάλυση δεδομένων διάρκειας ζωής, αλλά στην παρούσα εργασία θα αναλυθεί κυρίως το μοντέλο παλινδρόμησης επιταχυνόμενης διακοπής της κατανομής Weibull και το μοντέλο παλινδρόμησης του Cox (Cox regression model) ή διαφορετικά το μοντέλο αναλογικής διακινδύνευσης του Cox (Cox proportional hazard model).

Το μοντέλο του Cox, όπως και όλα τα μοντέλα αναλογικής διακινδύνευσης μοντελοποιεί τη συνάρτηση διακινδύνευσης (h(t)). Το μοντέλο αυτό χρησιμοποιείται ευρέως σήμερα στην ανάλυση αποκομμένων δεδομένων επιβίωσης για την εξακρίβωση των διαφορών στην επιβίωση που οφείλονται στο είδος της θεραπείας και σε προγνωστικούς παράγοντες σε κλινικές δοκιμές. Είναι επίσης μια καλή στατιστική τεχνική για την εύρεση της σχέσης μεταξύ της επιβίωσης ενός ασθενή και αρκετών επεξηγηματικών μεταβλητών. Επιπροσθέτως επιτρέπει να εκτιμηθεί ο κίνδυνος θανάτου ενός ατόμου ή άλλου γεγονότος δεδομένου των προγνωστικών τους μεταβλητών.

Ο όρος «καρκίνος του μαστού» αναφέρεται στην ανάπτυξη κακοήθους όγκου στην περιοχή του μαστού. Αποτελεί μία από τις συχνότερα εμφανιζόμενες μορφές καρκίνου παγκοσμίως και είναι η πρώτη σε αριθμό κρουσμάτων στο γυναικείο πληθυσμό. Προκαλείται από ανεξέλεγκτο πολλαπλασιασμό παθολογικών κυττάρων που ως αποτέλεσμα προκαλούν το σχηματισμό κακοήθους όγκου στην περιοχή του μαστού και ουσιαστικά αποτελεί κυτταρική νόσο. Τα παθολογικά αυτά κύτταρα έχουν τη δυνατότητα εξάπλωσης σε γειτονικούς ιστούς με δυσάρεστες συνέπειες για ολόκληρο τον οργανισμό. Σύμφωνα με τη Διεθνή Έκθεση για τις καρκινικές νόσους που παρουσιάστηκε στη Γενεύη της Ελβετίας τον Απρίλιο του 2003 από την IARC (International Agency for Research on Cancer) [Ferber (2003)], ο καρκίνος του μαστού αποτελεί την πιο συχνή μορφή καρκίνου μεταξύ των γυναικών, με περίπου 1.000.000 νέα κρούσματα παγκοσμίως. Στην Ελλάδα αναφέρονται 4.500 περίπου νέες περιπτώσεις το χρόνο, ενώ υπολογίζεται ότι 1 στις 8 γυναίκες παγκοσμίως θα παρουσιάσει καρκίνο μαστού σε κάποια φάση της ζωής της. Στην Ευρώπη, το 60% των κρουσμάτων καρκίνου του μαστού διαγιγνώσκεται σε πρώιμο

στάδιο. Το αντίστοιχο ποσοστό στην Ελλάδα είναι μόλις 5%. Είναι ένα από τα είδη καρκίνου που επιφέρουν τους περισσότερους θανάτους ετησίως (Παγκόσμιος Οργανισμός Υγείας) [Ferlay (2020)]. Παρατηρείται εξαιρετικά σπάνια στους άνδρες με περίπου 100 φορές μικρότερη συχνότητα συγκριτικά με τις γυναίκες και τα ίδια ποσοστά επιβίωσης (American Cancer Society). Ωστόσο, τα ποσοστά θανάτου από καρκίνο του μαστού 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. Ο αριθμός των ατόμων που έχουν αντιμετωπίσει με επιτυχία τον καρκίνο του μαστού αυξάνεται συνεχώς.

1.1 Διάρθρωση της διπλωματικής εργασίας

Η παρούσα εργασία χωρίζεται ουσιαστικά σε δύο εκτενή μέρη. Το πρώτο μέρος αφορά στο θεωρητικό υπόβαθρο κι αποτελείται από τρία κεφάλαια, ενώ το δεύτερο μέρος αναφέρεται στην πειραματική προσέγγιση και στην εξαγωγή συμπερασμάτων.

Πιο συγκεκριμένα, μετά την εισαγωγή, στην οποία τονίζονται τα πολύ βασικά στοιχεία της διπλωματικής εργασίας, στο Κεφάλαιο 2 αναλύονται τα κύρια μέρη που αφορούν στην Ανάλυση Επιβίωσης, ώστε ο αναγνώστης να εξοικειωθεί με τη γενικότερη ιδέα, στην οποία βασίζεται το θέμα προς ανάλυση.

Στο Κεφάλαιο 3 δίνονται γενικοί ορισμοί, μέσω μαθηματικών τύπων σχετικά με ελέγχους που θα υλοποιηθούν στην πειραματική διαδικασία, ώστε να εξαχθούν αρκετά συμπεράσματα και να επιτευχθεί καλύτερη σύγκριση αυτών για μεγαλύτερη αξιοπιστία.

Στο Κεφάλαιο 4 αναλύονται τα κυριότερα μοντέλα διάρκειας ζωής με συμμεταβλητές, τα οποία αποτελούν και τον ακρογωνιαίο λίθο για τη συγγραφή της συγκεκριμένης εργασίας, μιας και το μεγαλύτερο μέρος της πειραματικής προσέγγισης θα επιτευχθεί αξιοποιώντας αυτά τα μοντέλα.

Στο Κεφάλαιο 5 μπαίνουν σε εφαρμογή όλα τα παραπάνω αξιοποιώντας δεδομένα γυναικών που έπασχαν από καρκίνο του μαστού τη δεκαετία του 1980 και συλλέχθηκαν από τη ‘Γερμανική Ομάδα Μελέτης για τον Καρκίνο του Μαστού’ (the German Breast Cancer Study Group) [Sauerbrei (1999)]. Η επεξεργασία των

δεδομένων θα γίνει με τη βοήθεια του στατιστικού πακέτου Minitab, καθώς και της γλώσσας προγραμματισμού R και θα υλοποιηθεί σε τρεις φάσεις.

Στο Κεφάλαιο 6 θα συγκεντρωθούν όλα τα αποτελέσματα του προηγούμενου κεφαλαίου και θα εξαχθούν γενικά συμπεράσματα για το κατά πόσο διάφορες συμμεταβλητές, όπως η χρήση ορμονοθεραπείας μπορούν να επηρεάσουν την εξέλιξη της προαναφερθείσας ασθένειας μέχρι να συμβεί το πρώτο δυσάρεστο γεγονός (υποτροπή ή θάνατος) μέσα σε ένα ορισμένο χρονικό διάστημα.

1.2 Ερευνητική μέθοδος

Στα τρία πρώτα κεφάλαια, η προσέγγιση θα γίνει βιβλιογραφικά, ενώ στη συνέχεια η στατιστική ανάλυση των δεδομένων θα βασιστεί στη βοήθεια του στατιστικού πακέτου Minitab και της γλώσσας προγραμματισμού R.

Το Minitab είναι ένα πακέτο στατιστικών που αναπτύχθηκε στο Πανεπιστήμιο της Πενσυλβανίας από τους ερευνητές Barbara F. Ryan, Thomas A. Ryan, Jr. και Brian L. Joiner το 1972. Ξεκίνησε ως μια «ελαφριά» έκδοση του OMNITAB 80, ενός προγράμματος στατιστικής ανάλυσης από το NIST, το οποίο σχεδιάστηκε από τον Joseph Hilsenrath στα έτη 1962-1964 ως πρόγραμμα OMNITAB για το IBM 7090 [OMNITAB (1964), Peavy (1986)]. Το λογισμικό στατιστικής ανάλυσης όπως του Minitab αυτοματοποιεί τους υπολογισμούς και τη δημιουργία γραφημάτων, επιτρέποντας στο χρήστη να επικεντρωθεί περισσότερο στην ανάλυση των δεδομένων και στην ερμηνεία των αποτελεσμάτων. Εν ολίγοις, το Minitab περιέχει έτοιμες διαδικασίες στατιστικής ανάλυσης και ποιοτικού ελέγχου, ενώ ταυτόχρονα παρέχει ευδιάκριτα γραφικά που είναι πλήρως παραμετροποιήσιμα και προσφέρει για αυτόν το σκοπό μια πολύ μεγάλη ποικιλία στατιστικών μεθόδων ανάλυσης. Το Minitab διανέμεται από την Minitab, LLC, μια ιδιωτική εταιρεία με έδρα το State College της Πενσυλβανίας.

Η R είναι μια γλώσσα προγραμματισμού ανοικτού κώδικα και περιβάλλοντος που παρέχει στο χρήστη τη δυνατότητα να κάνει υπολογιστική στατιστική και γραφήματα. Υποστηρίζεται από τον οργανισμό 'R Foundation for Statistics Computing'. Η γλώσσα προγραμματισμού R χρησιμοποιείται ευρέως μεταξύ στατιστικολόγων και αναλυτών δεδομένων για την ανάπτυξη στατιστικού λογισμικού και ανάλυσης δεδομένων. Το επίσημο περιβάλλον λογισμικού της R είναι ένα πακέτο

GNU (ελεύθερου, ανοικτού λογισμικού). Είναι γραμμένο κυρίως σε γλώσσα C, Fortran, και στην ίδια την R και διατίθεται ελεύθερα βάσει της άδειας GNU General Public License. Γενικώς η R προσφέρει τα απαραίτητα εργαλεία προκειμένου να υλοποιηθεί μια στατιστική ανάλυση.

2 Ανάλυση Επιβίωσης

Η Ανάλυση Επιβίωσης (Survival Analysis) αναφέρεται στην ανάλυση δεδομένων διάρκειας ζωής που μελετούν το χρόνο που μεσολαβεί μέχρι να προκύψει κάποιο γεγονός [Καρώνη (2009)]. Αρχικά, η ανάλυση αναφερόταν στο χρόνο μεταξύ της θεραπείας ενός ασθενή μέχρι το θάνατο και για αυτό το λόγο πήρε και το συγκεκριμένο όνομα [Φωκιανός (2010)]. Η ανάλυση επιβίωσης όμως μπορεί να εφαρμοστεί σε αρκετές επιστημονικές περιοχές, όπως για παράδειγμα στη μηχανολογία για την ανάλυση του χρόνου που μεσολαβεί μέχρι την εμπλοκή/βλάβη ενός μηχανήματος ή στη γεωργία για την ανάλυση του χρόνου μέχρι τη στιγμή όπου ένα δέντρο θα βγάλει καρπούς. Στην περίπτωση της μηχανολογίας και γενικότερα στις εφαρμογές θετικών επιστημών, η ανάλυση επιβίωσης αναφέρεται και ως θεωρία αξιοπιστίας (reliability theory). Ο χρόνος επιβίωσης είναι περιορισμένος στο να είναι πάντα θετικός.

Επιπροσθέτως, τα δεδομένα που εξετάζονται περιέχουν πολλές φορές αποκομμένες παρατηρήσεις (censored data). Τα αποκομμένα δεδομένα είναι εκείνα, για τα οποία δεν είναι γνωστός ο χρόνος που συμβαίνει το γεγονός, δηλαδή ο χρόνος επιβίωσης. Υπάρχει η από δεξιά αποκοπή (right censoring), κατά την οποία υπάρχει ελλιπή γνώση για ένα άτομο (για παράδειγμα για έναν ασθενή) από ένα συγκεκριμένο χρόνο και έπειτα, η από αριστερά αποκοπή (left censoring), στην οποία το συμβάν έχει λάβει χώρα πριν από την πρώτη καταγραφή των συμβάντων, χωρίς να είναι γνωστός όμως ο ακριβής χρόνος του συμβάντος και η αποκοπή εντός διαστημάτων (interval censoring), κατά την οποία το συμβάν συμβαίνει εντός ενός συγκεκριμένου χρονικού διαστήματος, χωρίς όμως να είναι γνωστή η ακριβής χρονική στιγμή. Η πιο συνήθης μορφή αποκοπής που εμφανίζεται είναι εκείνη των δεξιά αποκομμένων παρατηρήσεων.

Στην Ανάλυση Επιβίωσης είναι πολύ σημαντικές δύο συναρτήσεις, οι οποίες περιγράφουν επαρκώς την κατανομή του χρόνου επιβίωσης: η συνάρτηση επιβίωσης ή αξιοπιστίας και η συνάρτηση διακινδύνευσης.

2.1 Συνάρτηση επιβίωσης

Συμβολίζοντας με Τ τη διάρκεια ζωής μιας υπό μελέτης μονάδας ως συνεχή τυχαία μεταβλητή με συνάρτηση πυκνότητας πιθανότητας $f(t), t \geq 0$, η συνάρτηση κατανομής ορίζεται ως

$$F(t) = P[T \leq t] = \int_0^t f(u) \cdot du .$$

Για την $F(t)$ ισχύει ότι είναι αύξουσα, $\lim_{t \to 0} F(t) = 0$ και $\lim_{t \to \infty} F(t) = 1$

Η συνάρτηση επιβίωσης (survival function) $S(t)$ [Caroni (2017), Καρώνη (2009)] ορίζεται ως η πιθανότητα επιβίωσης ενός ατόμου πέραν τη χρονική στιγμή t, δηλαδή η πιθανότητα η διάρκεια ζωής να είναι μεγαλύτερη του t και ορίζεται ως

$$S(t) = 1 - F(t) = P[T > t] = \int_t^{\infty} f(u) \cdot du .$$

Η συνάρτηση επιβίωσης είναι μη αρνητική και μη αύξουσα συνάρτηση του t με $S(0) = 1$ και $S(\infty) = 0$. Η γραφική παράσταση της $S(t)$ συναρτήσει του χρόνου t είναι γνωστή ως καμπύλη επιβίωσης και είναι πολύ σημαντική στην ανάλυση δεδομένων διάρκειας ζωής.

2.2 Συνάρτηση διακινδύνευσης

Η συνάρτηση διακινδύνευσης $h(t)$ [Caroni (2017), Καρώνη (2009)] εκφράζει την τάση προς διακοπή ενός αντικειμένου (ή του γεγονότος που εξετάζεται) στο χρονικό διάστημα $(t, t + \delta t]$ με δεδομένη την επιβίωσή του μέχρι τη χρονική στιγμή t και ορίζεται ως

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

Η συνάρτηση διακινδύνευσης (ρυθμός διακινδύνευσης – hazard function) $h(t)$ εκφράζει το ρυθμό στιγμιαίας διακοπής και η $h(t)\delta t$ είναι η πιθανότητα της

επικείμενης διακοπής μιας μονάδας, δοθέντος ότι επέζησε μέχρι τη συγκεκριμένη στιγμή t.

Η συνάρτηση διακινδύνευσης δηλαδή δίνει συνήθως ένα μέτρο του πόσο πιθανό είναι ένα άτομο να αποβιώσει ως συνάρτηση της ηλικίας του, για παράδειγμα ο κίνδυνος θανάτου ανάμεσα σε όσους είναι ζωντανοί τη συγκεκριμένη χρονική στιγμή t.

Η σωρευτική συνάρτηση διακινδύνευσης ορίζεται ως

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

και συνδέεται με τη συνάρτηση επιβίωσης μέσω των τύπων

$$H(t) = -\ln S(t)$$

και

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

2.3 Εκτιμήτρια Kaplan – Meier

Η ύπαρξη αποκομμένων δεδομένων σε ένα δείγμα καθιστά μη αποδοτική τη χρήση μιας οποιασδήποτε εκτιμήτριας της συνάρτησης επιβίωσης $S(t)$. Σε αυτή την περίπτωση και κυρίως όταν είναι γνωστοί οι παρατηρούμενοι χρόνοι επιβίωσης αποτελεσματική αποδεικνύεται η εκτιμήτρια συνάρτηση Kaplan – Meier . Η εκτιμήτρια Kaplan – Meier [Kaplan (1958)] είναι μια μη παραμετρική στατιστική εκτιμήτρια που χρησιμοποιείται για την εκτίμηση της συνάρτησης επιβίωσης σε δεδομένα διάρκειας ζωής. Στην ιατρική έρευνα χρησιμοποιείται συχνά για να υπολογίσει το πλήθος των ασθενών που ζουν για ένα ορισμένο χρονικό διάστημα μετά τη θεραπεία. Σε άλλους τομείς, οι εκτιμητές Kaplan – Meier μπορούν να χρησιμοποιηθούν για τη μέτρηση του χρόνου όπου άνθρωποι παραμένουν άνεργοι μετά από την απώλεια της εργασίας τους [Meyer (1990)], το χρονικό διάστημα έως την αποτυχία των εξαρτημάτων ενός μηχανήματος ή το κατά πόσο παραμένουν τα φρούτα στα φυτά πριν αφαιρεθούν για κατανάλωση. Ο εκτιμητής πήρε το όνομά του από τους Edward L. Kaplan και Paul Meier, οι οποίοι υπέβαλαν ξεχωριστά ο καθένας παρόμοια χειρόγραφα στην εφημερίδα της Αμερικανικής Στατιστικής Ένωσης

[Stalpers (2018)]. Ο συντάκτης της εφημερίδας, John Tukey τους έπεισε να ενώσουν το έργο τους σε ένα έγγραφο, το οποίο έχει μνημονευτεί περισσότερες από 59.000 φορές από την πρώτη του δημοσίευση το 1958 [Hevesi (2011)].

Για τον υπολογισμό της ορίζεται τυχαίο δείγμα n μονάδων και μέρος αυτού καταστρέφεται τις χρονικές στιγμές $t_1 < t_2 < ... < t_k, k \le n$. Θεωρείται ότι κατά τη χρονική στιγμή t_j καταστρέφονται d_j μονάδες, ενώ υπήρχαν n_j μονάδες σε λειτουργία αμέσως πριν από τη συγκεκριμένη στιγμή. Διευκρινίζεται ότι ο αριθμός n_j περιλαμβάνει όλες τις μονάδες που λειτουργούν εκείνη τη στιγμή, εκείνων των οποίων η λειτουργία θα διακοπεί κατά τη διάρκεια του πειράματος και εκείνων των οποίων δε θα διακοπεί, αλλά θα έχουν αποκομμένες τιμές.

Συνεπώς, η εκτιμήτρια Kaplan – Meier της συνάρτησης επιβίωσης $S(t)$ ορίζεται από τη σχέση

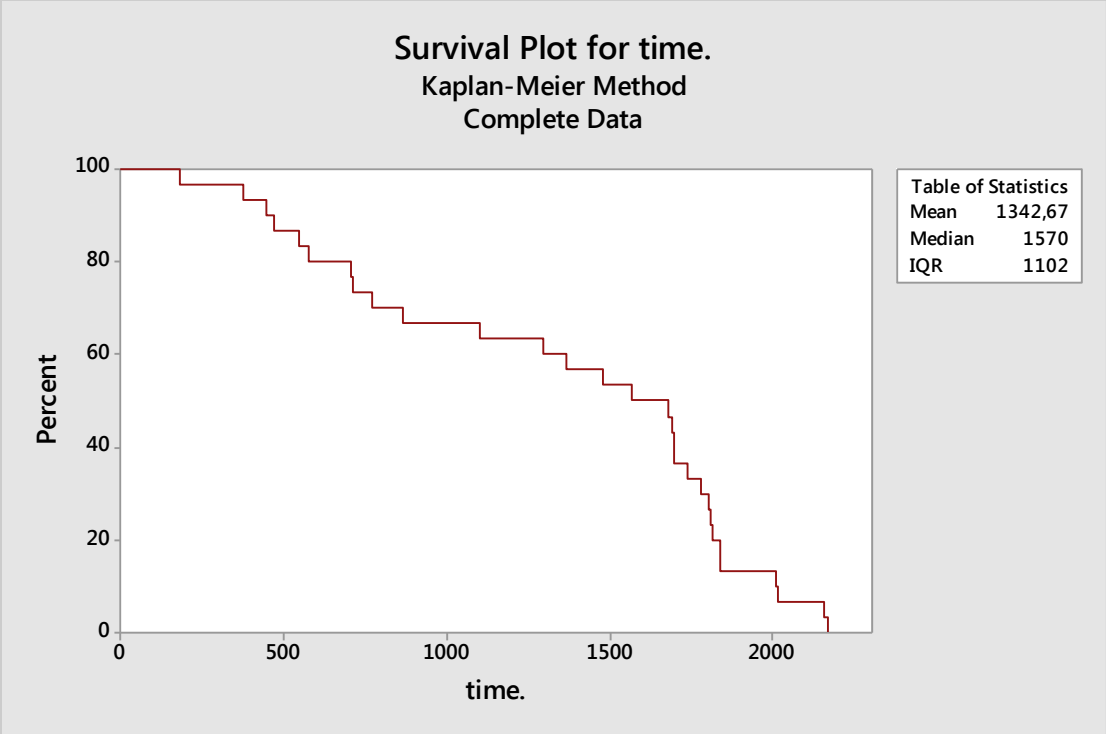
$$\hat{S}(t) = \prod_{j:t_j \le t} \frac{n_j - d_j}{n_j}, t \ge t_1$$

και $\hat{S}(t) = 1, t < t_1$.

Εφόσον αναφέρεται σε δειγματική εκτίμηση ορίζεται και το τυπικό σφάλμα της $\hat{S}(t)$ μέσω του τύπου του Greenwood

$$s.e.(\hat{S}) = \hat{S}(t) \cdot \left\{ \sum_{j:t_j \le t} \frac{d_j}{n_j (n_j - d_j)} \right\}^{\frac{1}{2}}.$$

Η εκτιμήτρια Kaplan – Meier της συνάρτησης επιβίωσης $S(t)$ θεωρείται ότι ακολουθεί προσεγγιστικά την κανονική κατανομή με μέση τιμή $S(t)$ και διασπορά σ^2, η οποία προκύπτει από την τυπική απόκλιση σ υψωμένη στο τετράγωνο, όταν η τελευταία εκτιμάται από το παραπάνω τυπικό σφάλμα. Αποτελεί μια βαθμιδωτή συνάρτηση και η γραφική της παράσταση είναι όπως στο γράφημα 2.3.1.



Γράφημα 2.3.1 Εκτίμηση Kaplan - Meier

2.4 Εκτιμήτρια Nelson – Aalen

Η εκτιμήτρια Nelson – Aalen είναι μια μη παραμετρική στατιστική εκτιμήτρια που χρησιμοποιείται για την εκτίμηση της σωρευτικής συνάρτησης διακινδύνευσης σε περίπτωση κυρίως αποκομμένων δεδομένων. Χρησιμοποιείται στη θεωρία Ανάλυσης Επιβίωσης, στην αξιοπιστία της μηχανικής και στην ασφάλιση ζωής για την εκτίμηση του σωρευτικού αριθμού των αναμενόμενων συμβάντων. Ένα «συμβάν» μπορεί να είναι η αποτυχία ενός μη επιδιορθώσιμου εξαρτήματος, ο θάνατος ενός ανθρώπου ή οποιοδήποτε συμβάν, για το οποίο η μονάδα του πειράματος παραμένει σε κατάσταση «αποτυχίας» (π.χ. θάνατος) από τη στιγμή που βρέθηκε σε αυτή την κατάσταση.

Για τον υπολογισμό της ορίζεται τυχαίο δείγμα n μονάδων και μέρος αυτού καταστρέφεται τις χρονικές στιγμές $t_1 < t_2 < \ldots < t_k, k \le n$. Θεωρείται ότι κατά τη χρονική στιγμή t_j καταστρέφονται d_j μονάδες, ενώ υπήρχαν n_j μονάδες σε λειτουργία αμέσως πριν από τη συγκεκριμένη στιγμή.

Τότε, η εκτιμήτρια Nelson – Aalen για τη σωρευτική συνάρτηση διακινδύνευσης $H(t)$ υπολογίζεται ως

$$\hat{H}(t) = \sum_{j:t_j \le t} \frac{d_j}{n_j}, t \ge t_1$$

και $\hat{H}(t) = 0, t < t_1$.

Επιπλέον, εκτιμήτρια της διασποράς της εκτιμήτριας Nelson – Aalen αποτελεί η

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

Η εκτιμήτρια Nelson – Aalen είναι επίσης μια βαθμιδωτή συνάρτηση.

Υπάρχουν και άλλοι τρόποι εκτίμησης της σωρευτικής συνάρτησης διακινδύνευσης, όπως για παράδειγμα μέσω της εκτιμήτριας Kaplan – Meier $\hat{S}(t)$ ή μέσω της εκτιμήτριας του Altshuler $\hat{S}(t)$, αλλά σε γενικές γραμμές προτιμάται ο εκτιμητής Nelson – Aalen.

3 Μέθοδος Μέγιστης Πιθανοφάνειας

Η εκτίμηση της μέγιστης πιθανοφάνειας είχε προταθεί, αναλυθεί και ευρέως διαδοθεί από τον Ρόναλντ Φίσερ μεταξύ 1912 και 1922 [Pfanzagl (1994)], αν και είχε χρησιμοποιηθεί νωρίτερα από τους Καρλ Φρίντριχ Γκάους, Πιερ Σιμόν Λαπλάς, Τόρβαλντ Ν. Τίλε και Φράνσις Ισίντρο Έτζγουορθ.

Στη στατιστική, η εκτίμηση μέγιστης πιθανοφάνειας (ΕΜΠ) είναι μια μέθοδος εκτίμησης των παραμέτρων μιας κατανομής πιθανότητας μεγιστοποιώντας τη συνάρτηση πιθανότητας, έτσι ώστε, σύμφωνα με το στατιστικό μοντέλο που έχει υποτεθεί, τα δεδομένα που παρατηρούνται να είναι τα πιθανότερα. Το σημείο στο χώρο παραμέτρων που μεγιστοποιεί τη συνάρτηση πιθανότητας ονομάζεται εκτίμηση μέγιστης πιθανότητας [Rossi (2018)]. Η λογική της μέγιστης πιθανότητας είναι τόσο διαισθητική όσο και ευέλικτη και ως εκ τούτου, η μέθοδος έχει γίνει κυρίαρχο μέσο για την εξαγωγή στατιστικών συμπερασμάτων [Chambers (2012), Hendry (2007), Ward (2018)].

Εάν η συνάρτηση πιθανότητας είναι διαφοροποιήσιμη μπορεί να εφαρμοστεί έλεγχος μέσω παραγώγων για τον προσδιορισμό των μέγιστων σημείων. Σε ορισμένες περιπτώσεις, οι πρώτες παράγωγοι της συνάρτησης πιθανότητας μπορούν να επιλυθούν ρητά. Για παράδειγμα, η συνηθισμένη εκτιμήτρια ελαχίστων τετραγώνων μεγιστοποιεί την πιθανότητα του μοντέλου γραμμικής παλινδρόμησης με κανονική κατανομή των σφαλμάτων [Press (1992)]. Ωστόσο, στις περισσότερες περιπτώσεις, αριθμητικές μέθοδοι είναι απαραίτητες για την εύρεση της μέγιστης συνάρτησης πιθανότητας.

Για να γίνει κατανοητή η έννοιά της, η παραπάνω μέθοδος θα εφαρμοστεί στην περίπτωση της εκθετικής κατανομής [Καρώνη (2009)], η οποία έχει συνάρτηση πυκνότητας πιθανότητας

$$f(t) = \lambda \cdot e^{-\lambda t}, t > 0 .$$

Με δεδομένο ένα απλό τυχαίο δείγμα n τιμών $t_1, t_2, ..., t_n$ από αυτή την κατανομή, η συνάρτηση πιθανοφάνειας ορίζεται ως

$$L\left(\lambda | t_1, t_2, ..., t_n\right) = \prod_{i=1}^{n} f\left(t_i\right) = \prod_{i=1}^{n} \left(\lambda \cdot e^{-\lambda t_i}\right) = \lambda^n \cdot e^{-\lambda \sum t_i}$$

και λογαριθμίζοντας

$$l = \ln L = n \cdot \ln \lambda - \lambda \sum_{i=1}^{n} t_i \, .$$

Η συνάρτηση L ή ισοδύναμα l μεγιστοποιείται ως προς την παράμετρο λ.

Παίρνοντας την πρώτη παράγωγο ως προς λ, θα ισχύει

$$\frac{\partial l}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} t_i = 0 \, ,$$

από όπου θα προκύψει

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} t_i} \, .$$

Επιπλέον παίρνοντας τη δεύτερη παράγωγο ως προς λ συμπεραίνεται ότι

$$\frac{\partial^2 l}{\partial \lambda^2} = \frac{-n}{\lambda^2} < 0 \, ,$$

οπότε η ανωτέρω τιμή αφορά στη μέγιστη τιμή της λ.

Η αρνητική τιμή της δεύτερης παραγώγου στο σημείο $\hat{\lambda}$

$$-\frac{\partial^2 l}{\partial \lambda^2} = \frac{n}{\hat{\lambda}^2}$$

ονομάζεται παρατηρούμενη πληροφορία (observed information) και το αντίστροφό της δίνει την εκτίμηση της διασποράς της εκτιμήτριας $\hat{\lambda}$

$$\hat{V}\left(\hat{\lambda}\right) = \frac{\hat{\lambda}^2}{n} \, .$$

Η ασυμπτωτική κατανομή της εκτιμήτριας μέγιστης πιθανοφάνειας $\hat{\lambda}$ είναι κανονική, κάτι το οποίο αποτελεί ένα από τα μεγαλύτερα πλεονεκτήματα αυτής της μεθόδου, μιας κι όλες τις εκτιμήτριες μέγιστης πιθανοφάνειας μπορεί κάποιος να τις προσεγγίζει πάντα μέσω της κανονικής κατανομής.

Συνεπώς, ένα διάστημα εμπιστοσύνης για τη $\hat{\lambda}$ είναι

$$\hat{\lambda} \pm z\sqrt{\hat{V}\left(\hat{\lambda}\right)},$$

με z το κατάλληλο ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής.

Στην περίπτωση των αποκομμένων δεδομένων, οι παρατηρήσεις $t_1, t_2, ..., t_n$ χωρίζονται στα υποσύνολα

$$u = \{\text{μη} - \text{αποκομμένες παρατηρήσεις}\}$$

$$c = \{\text{αποκομμένες παρατηρήσεις}\}.$$

Τότε η συνάρτηση πιθανοφάνειας L δίνεται ως

$$L\left[\left\{t_i : i = 1, ..., n\right\}\right] = L\left[\left\{t_i : i \in u\right\}\right] L\left[\left\{t_i : i \in c\right\}\right] = L_u L_c.$$

Ο όρος $L_u = \prod_{i \in u} f\left(t_i\right)$ συμπεριφέρεται όπως παραπάνω στην περίπτωση των μη αποκομμένων παρατηρήσεων. Ο όρος L_c όμως εξαρτάται από τη μορφή της αποκοπής.

Στην περίπτωση της από δεξιά αποκοπής, η τιμή t_i αντιστοιχεί με το γεγονός $T > t_i$, συνεπώς θα ισχύει

$$L_c = \prod_{i \in c} P\left[T > t_i\right] = \prod_{i \in c} S\left(t_i\right).$$

Για την από αριστερά αποκοπή θα ισχύει

$$L_c = \prod_{i \in c} \{1 - S(t_i)\},$$

ενώ για την αποκοπή εντός διαστήματος $t_{\alpha_i} < t_i < t_{\beta_i}$ η συνάρτηση πιθανοφάνειας είναι

$$L_c = \prod_{i \in c} \left\{ S\left(t_{\alpha_i}\right) - S\left(t_{\beta_i}\right) \right\}.$$

Και οι τρεις παραπάνω μορφές μπορούν να περιληφθούν στην ίδια L_c, αρκεί το γεγονός της αποκοπής να είναι τυχαίο.

3.1 P – value

Ως επίπεδο σημαντικότητας ενός ελέγχου ορίζεται το μέγιστο αποδεκτό επίπεδο της πιθανότητας σφάλματος τύπου I (λανθασμένης απόρριψης της H_0) και συμβολίζεται με a.

Έστω ότι υπάρχει ένας έλεγχος υποθέσεων για δυο κατανομές με τη μηδενική υπόθεση H_0 να υποστηρίζει ότι οι κατανομές συμπίπτουν έναντι της εναλλακτικής υπόθεσης H_1, σύμφωνα με την οποία οι κατανομές διαφέρουν. Τότε η p – τιμή (p – value) είναι το μικρότερο επίπεδο σημαντικότητας a^*, το οποίο κάποιος μπορεί να χρησιμοποιήσει ώστε να απορρίψει την H_0 (δηλαδή είναι το πρώτο a, για το οποίο είναι δυνατόν να απορριφθεί η υπόθεση H_0). Όσο μικρότερη είναι η τιμή της p – value, τόσο πιο ισχυρή είναι η απόρριψη της H_0.

Η ερμηνεία της p – value για τα διάφορα πιο συνήθη επίπεδα σημαντικότητας δίνεται στον πίνακα 1.

P – value	Ερμηνεία
< 0,10	Ένδειξη ότι η H_0 δεν αληθεύει
< 0,05	Ισχυρή ένδειξη ότι η H_0 δεν αληθεύει
< 0,01	Πολύ ισχυρή ένδειξη ότι η H_0 δεν αληθεύει
< 0,001	Υπερβολικά ισχυρή ένδειξη ότι η H_0 δεν αληθεύει

Πίνακας 1 Ερμηνεία της p – value για τα διάφορα επίπεδα σημαντικότητας

3.2 Έλεγχος Wald

Στη στατιστική, ο έλεγχος Wald (που πήρε το όνομά του από τον Abraham Wald) αξιολογεί τους περιορισμούς στις στατιστικές παραμέτρους με βάση τη σταθμισμένη απόσταση μεταξύ της εκτίμησης της μέγιστης πιθανοφάνειας και της υποθετικής τιμής της υπό την μηδενική υπόθεση, όπου το βάρος είναι η ακρίβεια της εκτίμησης [Fahrmei (2013)]. Διαισθητικά, όσο μεγαλύτερη είναι αυτή η σταθμισμένη απόσταση, τόσο λιγότερο πιθανό είναι να ισχύει ο περιορισμός. Ενώ οι πεπερασμένες κατανομές των ελέγχων Wald είναι γενικά άγνωστες [Martin (2013)], η ελεγχοσυνάρτηση μπορεί να προσεγγιστεί ασυμπτωτικά από τη X^2- κατανομή υπό την μηδενική υπόθεση, γεγονός που μπορεί να χρησιμοποιηθεί για τον προσδιορισμό της στατιστικής της σημασίας [Davidson (1993)].

Ως εκ τούτου, η ελεγχοσυνάρτηση Wald για τον έλεγχο της υπόθεσης $H_0 : \lambda = \lambda_0$ με εναλλακτική την $H_1 : \lambda \neq \lambda_0$ είναι

$$\frac{\hat{\lambda} - \lambda_0}{\sqrt{\hat{V}\left(\hat{\lambda}\right)}},$$

η οποία ακολουθεί προσεγγιστικά την τυποποιημένη κανονική κατανομή $\mathrm{N}(0,1)$.

Ισοδυνάμως, ο έλεγχος Wald εκφράζεται και ως

$$\frac{\left(\hat{\lambda}-\lambda_0\right)^2}{\hat{V}\left(\hat{\lambda}\right)},$$

λόγος ο οποίος ακολουθεί ασυμπτωτικά την X^2 - κατανομή με ένα βαθμό ελευθερίας.

Ένα πλεονέκτημα του Wald test έναντι άλλων ελέγχων είναι ότι απαιτεί μόνο την εκτίμηση της μέγιστης πιθανοφάνειας, κάτι το οποίο μειώνει το υπολογιστικό φορτίο σε σύγκριση με ελέγχους, όπως του λόγου πιθανοφανειών.

3.3 Έλεγχος του λόγου των πιθανοφανειών

Η μέθοδος μέγιστης πιθανοφάνειας οδηγεί επίσης στους ελέγχους του λόγου των πιθανοφανειών. Στη στατιστική, ένας έλεγχος του λόγου πιθανοφανειών είναι ένας στατιστικός έλεγχος που χρησιμοποιείται για τη σύγκριση της καλής προσαρμογής δύο μοντέλων, ένα εκ των οποίων (το μηδενικό μοντέλο) είναι μια ειδική περίπτωση του άλλου (το εναλλακτικό μοντέλο).

Η ελεγχοσυνάρτηση για τον έλεγχο της υπόθεσης $H_0 : \lambda = \lambda_0$ με εναλλακτική την $H_1 : \lambda \neq \lambda_0$ είναι

$$\Lambda = -2\left\{l\left(\lambda_0\right) - l\left(\hat{\lambda}\right)\right\},$$

η οποία ακολουθεί ασυμπτωτικά την X^2 - κατανομή με ένα βαθμό ελευθερίας.

$\Rightarrow$ Γενικώς, ο έλεγχος Wald απαιτεί την προσαρμογή ενός μόνο μοντέλου κι είναι χρήσιμος για μια πρώτη ένδειξη των σημαντικών μεταβλητών στο μοντέλο όταν υπάρχουν πολλές συμμεταβλητές. Ο έλεγχος του λόγου των πιθανοφανειών απαιτεί περισσότερους υπολογισμούς, αλλά είναι προτιμότερος.

3.4 Κριτήριο AIC

Ο Χιροτόγκου Ακαΐκε (5 Νοεμβρίου 1927 – 4 Αυγούστου 2009) ήταν Ιάπωνας στατιστικολόγος στη θεωρία της πληροφορίας. Στις αρχές της δεκαετίας του 1970 διατύπωσε ένα κριτήριο για την επιλογή του άριστου μοντέλου, το επονομαζόμενο Akaike Information Criterion, γνωστό πλέον με τη συντόμευση AIC, το οποίο είναι ευρέως διαδεδομένο στη στατιστική και στην οικονομετρία. Με

δεδομένο ένα σύνολο μοντέλων που ερμηνεύουν κάποια δεδομένα, το AIC υπολογίζει την ποιότητα του κάθε μοντέλου σε σχέση με τα υπόλοιπα μοντέλα. Έτσι το AIC παρέχει ένα μέσο για την επιλογή του μοντέλου που ερμηνεύει καλύτερα τα δεδομένα.

Επομένως, το κριτήριο AIC δίνεται από τον τύπο

$$AIC = -2\hat{l} + 2p,$$

όπου p ο συνολικός αριθμός παραμέτρων στο μοντέλο που εξετάζεται. Το καταλληλότερο μοντέλο θα είναι εκείνο που έχει το μικρότερο AIC.

3.5 Διαστήματα εμπιστοσύνης

Τα διαστήματα εμπιστοσύνης παρουσιάστηκαν στη στατιστική από τον Jerzy Neyman σε ένα έγγραφο που δημοσιεύτηκε το 1937 [Neyman (1937)].

Στη στατιστική, ένα διάστημα εμπιστοσύνης (ΔΕ) είναι ένας τύπος διαστήματος εκτίμησης μιας παραμέτρου του πληθυσμού. Είναι ένα διάστημα, το οποίο υπολογίζεται από τις παρατηρήσεις (διαφέρει από δείγμα σε δείγμα) και συχνά περιλαμβάνει την αξία της παρατηρούμενης παραμέτρου ενδιαφέροντος, αν επαναληφθεί το πείραμα. Το πόσο συχνά παρατηρείται το διάστημα να περιέχει την παράμετρο καθορίζεται από το επίπεδο εμπιστοσύνης ή το συντελεστή εμπιστοσύνης. Πιο συγκεκριμένα, η έννοια του όρου "επίπεδο εμπιστοσύνης" είναι ότι, αν τα διαστήματα εμπιστοσύνης είναι κατασκευασμένα σε πολλές ξεχωριστές αναλύσεις δεδομένων από επανειλημμένα (και ενδεχομένως διαφορετικά) πειράματα, το ποσοστό αυτών των διαστημάτων που περιέχουν την πραγματική τιμή της παραμέτρου θα ταιριάζει με το δεδομένο επίπεδο εμπιστοσύνης [Cox (1974), Kendall (1973), Neyman (1937)].

Τα διαστήματα εμπιστοσύνης αποτελούνται από ένα εύρος τιμών (διάστημα) που ενεργούν ως καλές εκτιμήσεις της άγνωστης παραμέτρου του πληθυσμού, ωστόσο, το χρονικό διάστημα που υπολογίζεται από ένα συγκεκριμένο δείγμα δεν περιλαμβάνει απαραίτητα την πραγματική τιμή της παραμέτρου. Όταν κάποιος υποστηρίξει ότι είναι 99% σίγουρος ότι η πραγματική τιμή της παραμέτρου είναι στο διάστημα εμπιστοσύνης που έχει σημαίνει ότι θεωρεί πως το 99% των διαστημάτων εμπιστοσύνης που υποθετικά έχει παρατηρήσει θα κρατήσει την πραγματική τιμή της

παραμέτρου. Το επιθυμητό επίπεδο εμπιστοσύνης ορίζεται από τον ερευνητή και δεν καθορίζεται από τα δεδομένα. Αν εκτελείται μια δοκιμή αντίστοιχης υπόθεσης, το επίπεδο εμπιστοσύνης είναι το συμπλήρωμα των αντίστοιχων επιπέδων σημαντικότητας, δηλαδή ένα 95% διάστημα εμπιστοσύνης αντικατοπτρίζει ένα επίπεδο σημαντικότητας 0,05. Υψηλότερα επίπεδα διακύμανσης αποδίδουν μεγαλύτερα διαστήματα εμπιστοσύνης και ως εκ τούτου λιγότερο ακριβείς εκτιμήσεις της παραμέτρου. Διαστήματα εμπιστοσύνης διαφορετικών παραμέτρων που δεν περιέχουν το μηδέν υπονοούν ότι υπάρχει στατιστικά σημαντική διαφορά μεταξύ των πληθυσμών.

Στην εφαρμοσμένη πρακτική, τα διαστήματα εμπιστοσύνης αναφέρονται συνήθως στο 95% του επιπέδου εμπιστοσύνης [Zar, (1984)]. Ορισμένοι παράγοντες μπορεί να επηρεάσουν το μέγεθος του διαστήματος εμπιστοσύνης συμπεριλαμβανομένου του μεγέθους του δείγματος, του επιπέδου εμπιστοσύνης και της πληθυσμιακής διακύμανσης. Ένα μεγαλύτερου μεγέθους δείγμα συνήθως θα οδηγήσει σε μια καλύτερη εκτίμηση της παραμέτρου του πληθυσμού.

3.6 Μέθοδος της Backward Elimination

Στη στατιστική βιβλιογραφία υπάρχουν πολλές μέθοδοι για τον προσδιορισμό του βέλτιστου κι οικονομικότερου υποσυνόλου από ένα σύνολο μεταβλητών που είναι υποψήφιες για να περιληφθούν σε ένα μοντέλο πολλαπλής παλινδρόμησης. Οι βασικότερες μέθοδοι αφορούν στη διαδικασία της προς τα πίσω απαλοιφής (Backward Elimination), στην προς τα εμπρός επιλογή (Forward Selection) και στην κατά βήματα εμπρός – πίσω επιλογή (Stepwise Selection). Η παρούσα εργασία θα εστιάσει στην πρώτη μέθοδο.

Σύμφωνα με τη μέθοδο της Backward Elimination, η διαδικασία ξεκινάει συμπεριλαμβάνοντας στο μοντέλο όλες τις διαθέσιμες μεταβλητές και τις αφαιρεί μία – μία ξεκινώντας από τη λιγότερο σημαντική, αν ο έλεγχος, που πραγματοποιείται με τη βοήθεια του F – ελέγχου είναι στατιστικά σημαντικός.

Διευκρινίζεται ότι ο έλεγχος – F είναι ο ένας εκ των δύο στατιστικών ελέγχων που χρησιμοποιούνται για την επιλογή των στατιστικά σημαντικών μεταβλητών σε ένα μοντέλο γραμμικής παλινδρόμησης. Ο έλεγχος αφορά στους συντελεστές β των

μεταβλητών στη συνάρτηση παλινδρόμησης που προσαρμόζεται στα δεδομένα του εκάστοτε προβλήματος.

Δίνεται ότι η συνάρτηση παλινδρόμησης είναι η

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_k \cdot x_k ,$$

όπου $\beta_i, i = 1, 2, ..., k$ οι συντελεστές των μεταβλητών στη συνάρτηση παλινδρόμησης.

Η ελεγχοσυνάρτηση F ορίζεται υπό τις υποθέσεις $H_0 : \beta_i = 0$ έναντι της $H_1 : \beta_i \neq 0$ για τουλάχιστον ένα i.

Ο έλεγχος βασίζεται στη διαπίστωση ότι αν μια μεταβλητή δεν είναι στατιστικά σημαντική, τότε η συνεισφορά της στη συνάρτηση παλινδρόμησης θα είναι ελάχιστη ή μηδενική. Δηλαδή, αυτό με άλλα λόγια σημαίνει ότι ο συντελεστής της συγκεκριμένης μη στατιστικά σημαντικής μεταβλητής στη συνάρτηση παλινδρόμησης, δε θα είναι στατιστικά διάφορος από το μηδέν [Καρώνη (2020)].

4 Μοντέλα Διάρκειας Ζωής με Συμμεταβλητές

Συμμεταβλητές ονομάζονται οι (μετρήσιμοι) παράγοντες, από τους οποίους εξαρτάται η διάρκεια ζωής μιας μονάδας, όπως είναι τα χαρακτηριστικά της μονάδας ή οι συνθήκες κάτω από τις οποίες λειτουργεί. Η εισαγωγή τους στα στατιστικά μοντέλα συνήθως συμβάλλει στην καλύτερη περιγραφή της διάρκειας ζωής. Υπάρχουν πολλοί έλεγχοι που αξιοποιούνται για τη σύγκριση ομάδων μονάδων με ή χωρίς επιπλέον συμμεταβλητές. Ο πιο διαδεδομένος είναι ο έλεγχος log – rank, ο οποίος εφαρμόζεται κυρίως στην περίπτωση της σύγκρισης δύο μόνο ομάδων μονάδων και εξετάζει αν διαφοροποιείται η συνάρτηση επιβίωσης μεταξύ των δυο ομάδων.

4.1 Έλεγχος Log – Rank

Ο έλεγχος προτάθηκε για πρώτη φορά από τον Nathan Mantel και ονομάστηκε έλεγχος log – rank από τους Richard και Julian Peto [Harrington (2005), Mantel (1966), Peto (1972)].

Ο έλεγχος logrank, ή το log – rank test είναι ένας έλεγχος υπόθεσης για τη σύγκριση των κατανομών επιβίωσης δύο δειγμάτων. Είναι ένας μη παραμετρικός έλεγχος και κατάλληλος για χρήση όταν τα δεδομένα είναι από δεξιά αποκομμένα. Χρησιμοποιείται ευρέως σε κλινικές δοκιμές για να διαπιστωθεί η αποτελεσματικότητα μιας νέας θεραπείας σε σύγκριση με μια θεραπεία ελέγχου όταν η μέτρηση βασίζεται στο χρόνο μέχρι να συμβεί το γεγονός (όπως για παράδειγμα το χρόνο από την αρχική θεραπεία έως την καρδιακή προσβολή). Ο έλεγχος καλείται μερικές φορές ως τεστ Mantel – Cox, παίρνοντας το όνομά του από τους Nathan Mantel και David Cox.

Γενικώς θεωρείται ένας πολύ ισχυρός έλεγχος.

Πιο συγκεκριμένα, σύμφωνα με τον έλεγχο log – rank, κάποιος μπορεί να υποθέσει ότι υπάρχουν $t_{(1)} < t_{(2)} < ... < t_{(k)}$ διακεκριμένες χρονικές στιγμές, κατά τις οποίες παύουν να λειτουργούν μονάδες που προέρχονται από δύο ομάδες 1 και 2. Θεωρείται ότι αμέσως πριν από τη χρονική στιγμή $t_{(j)}$ υπάρχουν n_{1j} μονάδες «σε κίνδυνο», δηλαδή σε λειτουργία, χωρίς να έχει προκύψει ακόμη το (αρνητικό) γεγονός για την ομάδα 1, εκ των οποίων d_{1j} μονάδες παύουν να λειτουργούν τη στιγμή $t_{(j)}$.

Παρομοίως θεωρείται ότι αμέσως πριν από τη χρονική στιγμή $t_{(j)}$ υπάρχουν n_{2j} μονάδες «σε κίνδυνο» για την ομάδα 2, εκ των οποίων d_{2j} μονάδες παύουν να λειτουργούν τη στιγμή $t_{(j)}$. Δηλαδή, τη χρονική στιγμή $t_{(j)}$ παύουν να λειτουργούν συνολικά $d_j = d_{1j} + d_{2j}$ μονάδες από τις $n_j = n_{1j} + n_{2j}$ μονάδες που είναι «σε κίνδυνο».

Τα γεγονότα της χρονικής στιγμής $t_{(j)}$ περιγράφονται περιληπτικά στον πίνακα συνάφειας 2 [Καρώνη (2009)].

Διακοπή λειτουργίας	**Ομάδα 1**	**Ομάδα 2**	**Σύνολο**
Ναι	d_{1j}	d_{2j}	d_j
Όχι	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
Σύνολο σε κίνδυνο πριν τη στιγμή $t_{(j)}$	n_{1j}	n_{2j}	n_j

Πίνακας 2 Πίνακας συνάφειας για τα γεγονότα της χρονικής στιγμής $t_{(j)}$

Η τελική μορφή της ελεγχοσυνάρτησης log – rank είναι

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

όπου $\frac{n_{1j} d_j}{n_j}$ είναι η αναμενόμενη συχνότητα του πρώτου κελιού στον πίνακα 2, δηλαδή μονάδες της ομάδας 1, των οποίων η λειτουργία διακόπηκε και $d_{1j} - \left(\frac{n_{1j} d_j}{n_j} \right)$ είναι η απόκλιση από την παρατηρούμενη d_{1j}. Η ελεγχοσυνάρτηση ακολουθεί ασυμπτωτικά την X^2- κατανομή με ένα βαθμό ελευθερίας.

4.2 Έλεγχος Wilcoxon

Γενίκευση του log – rank ελέγχου είναι ο έλεγχος Wilcoxon. Ο έλεγχος Wilcoxon είναι επίσης ένας μη παραμετρικός έλεγχος στατιστικής υπόθεσης που χρησιμοποιείται για τη σύγκριση των συναρτήσεων επιβίωσης από δύο ανεξάρτητα δείγματα.

Η τελική μορφή της ελεγχοσυνάρτησης Wilcoxon είναι

$$\frac{\left(\sum w_j u_j\right)^2}{\sum w_j^2 v_j},$$

όπου w_j είναι συντελεστής στάθμισης και ισχύει $w_j = n_j$, δηλαδή ισούται με το πλήθος των μονάδων που βρίσκονται σε κίνδυνο πριν από τη χρονική στιγμή $t_{(j)}$. Άμεσα προκύπτει το συμπέρασμα ότι ο έλεγχος log – rank είναι ένας έλεγχος Wilcoxon με συντελεστή βαρύτητας $w_j = 1$. Η διαφορά ανάμεσα στους δύο ελέγχους ερμηνεύεται ως προς το ότι ο έλεγχος log – rank δίνει ισόποσο βάρος σε όλη τη διάρκεια της διαδικασίας των διακοπών που συμβαίνουν στις μονάδες, δηλαδή είναι καταλληλότερος όταν ισχύει η υπόθεση της αναλογικής διακινδύνευσης (παράγραφος 4.3.1) ως προς τον έλεγχο για το αν υπάρχουν διαφοροποιήσεις στις συναρτήσεις επιβίωσης δύο ομάδων, ενώ ο έλεγχος Wilcoxon δίνει μεγαλύτερο βάρος στις αρχικές διακοπές που συμβαίνουν στο πείραμα σε σύγκριση με όσες θα συμβούν αργότερα, για αυτό το λόγο ο έλεγχος Wilcoxon συνήθως έχει μεγαλύτερη p – value κι είναι ισχυρότερος του ελέγχου log – rank όταν υπάρχει ενδιαφέρον για τις διαφοροποιήσεις στις αρχικές συναρτήσεις επιβίωσης [Collett (2015)].

4.3 Μοντέλα παλινδρόμησης

Όπως αναφέρθηκε στην παράγραφο 3.6, η γραμμική παλινδρόμηση, όπου η τιμή μιας εξαρτημένης μεταβλητής y σχετίζεται με τις συμμεταβλητές x_i δίνεται από τη σχέση

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_k \cdot x_k + \varepsilon = \beta' x + \varepsilon ,$$

όπου ε είναι τα υπόλοιπα διαφορετικών παρατηρήσεων και θεωρούνται ανεξάρτητα ακολουθώντας την κανονική κατανομή $N\left(0, \sigma^2\right)$. Παράλληλα, οι συμμεταβλητές x_i θεωρούνται μη – στοχαστικές, οπότε προκύπτει ότι $y \sim N\left(\mu, \sigma^2\right)$ με

$$\mu = \mu\left(x\right) = \beta' x .$$

Οι συμμεταβλητές εισάγονται στο μοντέλο με βάση την επίδρασή τους στην παράμετρο μ της κατανομής y. Στην παρούσα εργασία θα παρουσιαστούν μοντέλα παλινδρόμησης για τη διάρκεια ζωής.

4.3.1 Μοντέλο αναλογικής διακινδύνευσης

Αρχικά θα παρουσιαστεί το μοντέλο αναλογικής διακινδύνευσης (proportional hazards model (PH)). Το συγκεκριμένο μοντέλο αναφέρεται στο λόγο των συναρτήσεων διακινδύνευσης δύο μονάδων

$$\frac{h\left(t|\lambda_1\right)}{h\left(t|\lambda_2\right)} = \frac{\lambda_1}{\lambda_2} ,$$

ο οποίος είναι σταθερός και ανεξάρτητος του χρόνου t.

Ο ανωτέρω λόγος βασίζεται στη μορφή του μοντέλου

$$h\left(t|\lambda\right) = \lambda \cdot h_0\left(t\right)$$

όπου $h_0\left(t\right)$ μια βασική συνάρτηση διακινδύνευσης και η τυχαία ποσότητα $\lambda > 0$ είναι η ευπάθεια (frailty) της εκάστοτε μονάδας.

Αναλόγως με το αν η βασική συνάρτηση διακινδύνευσης είναι γνωστή (δηλαδή γνωστής κατανομής), το μοντέλο αναλογικής διακινδύνευσης χωρίζεται σε δύο κατηγορίες. Η πρώτη αναφέρεται στο παραμετρικό μοντέλο παλινδρόμησης, όταν η

$h_0(t)$ είναι συγκεκριμένης γνωστής κατανομής. Η δεύτερη αφορά στο ημι – παραμετρικό μοντέλο παλινδρόμησης, όταν η $h_0(t)$ είναι ακαθόριστη. Σε αυτή την κατηγορία ανήκει και το μοντέλο του Cox (αναλύεται στην παράγραφο 4.3.3) με $\lambda = e^{x'\beta}$.

Επιπλέον αναφέρεται ότι από τη συνάρτηση διακινδύνευσης

$$h(t;x) = h_0(t) \cdot \lambda(x)$$

προκύπτει η συνάρτηση επιβίωσης

$$S(t;x) = \exp\left\{-H_0(t)e^{\beta' x}\right\},$$

όπου $H_0(t)$ η σωρευτική βασική συνάρτηση διακινδύνευσης αντίστοιχη της βασικής συνάρτησης διακινδύνευσης $h_0(t)$.

Τότε λογαριθμίζοντας δύο φορές προκύπτει

$$\begin{aligned} & S(t;x) = \exp\left\{-H_0(t)e^{\beta' x}\right\} \\ & \ln S(t;x) = -H_0(t)e^{\beta' x} \\ & -\ln S(t;x) = H_0(t)e^{\beta' x} \\ & \ln\left\{-\ln S(t;x)\right\} = \ln H_0(t) + \ln e^{\beta' x} \\ & \ln\left\{-\ln S(t;x)\right\} = \ln H_0(t) + \beta' x \\ & \ln\left\{-\ln S(t;x)\right\} - \ln H_0(t) = \beta' x \end{aligned},$$

όπου $\ln H_0(t)$ σταθερό ως προς το χρόνο για όλες τις μονάδες, μιας και η H_0 είναι κοινή για όλες τις μονάδες .

Τελικά όλο αυτό σημαίνει ότι η καμπύλη $\ln\left\{-\ln S(t;x)\right\}$ είναι παράλληλη με την $\ln H_0(t)$ ως προς το χρόνο για οποιαδήποτε x. Συνεπώς, η υπόθεση της αναλογικής διακινδύνευσης ισχύει αν όλες οι καμπύλες $\ln\left\{-\ln S(t;x_i)\right\}$ για τα διάφορα x_i είναι παράλληλες. Για την υλοποίηση αυτού του ελέγχου απαραίτητες είναι οι εκτιμήσεις $\hat{S}(t;x)$, συνήθως των Kaplan – Meier για επιλεγμένες τιμές των x. Στη συνέχεια πραγματοποιείται η γραφική παράσταση $\ln\left\{-\ln \hat{S}(t;x)\right\}$ με t. Αν προκύψουν

καμπύλες παράλληλες μεταξύ τους, τότε ισχύει η υπόθεση της αναλογικής διακινδύνευσης κι αυτό εφαρμόζεται για οποιοδήποτε μοντέλο της αναλογικής διακινδύνευσης. Σημειώνεται ότι αν οι καμπύλες είναι ευθύγραμμες, τότε ισχύει η εφαρμογή της κατανομής Weibull. Τέλος τονίζεται ότι η όλη διαδικασία είναι έγκυρη μόνο εάν υπάρχει ένας αρκετά μεγάλος αριθμός μονάδων με την ίδια τιμή των συμμεταβλητών. Για το λόγο αυτό προτιμάται όταν οι συμμεταβλητές είναι λίγες και συστήνεται η ομαδοποίηση των τιμών των ποσοτικών συμμεταβλητών [Καρώνη (2009)].

4.3.2 Μοντέλο επιταχυνόμενης διακοπής

Το μοντέλο επιταχυνόμενης διακοπής (accelerated failure model (AF)) βασίζεται στο συλλογισμό ότι στα μοντέλα παλινδρόμησης που αναφέρονται σε δεδομένα διάρκειας ζωής απαιτείται η εξαρτημένη μεταβλητή y να είναι πάντα θετική. Ωστόσο, κάτι τέτοιο πολλές φορές δεν είναι εφικτό. Για αυτό το λόγο χρησιμοποιείται ισοδυνάμως ένα λογαριθμο – γραμμικό μοντέλο για μια τυχαία μεταβλητή T

$$\ln T_x = \mu + \beta' x + \sigma\varepsilon ,$$

όπου μ είναι μια παράμετρος θέσης, σ μια παράμετρος κλίμακας και ε είναι τυχαία μεταβλητή.

Τότε, η συνάρτηση επιβίωσης μετατρέπεται ως

$$\begin{aligned} S(t;x) &= P(T_x > t) \\ S(t;x) &= P(\ln T_x > \ln t) \\ S(t;x) &= P(\mu + \beta' x + \sigma\varepsilon > \ln t), \\ S(t;x) &= P\left\{\varepsilon > \frac{\ln t - \mu - \beta' x}{\sigma}\right\} \\ S(t;x) &= S_\varepsilon\left\{\frac{\ln t - \mu - \beta' x}{\sigma}\right\} \end{aligned}$$

όπου S_ε είναι η συνάρτηση αξιοπιστίας της τυχαίας μεταβλητής ε.

Συνεπώς, η συνάρτηση επιβίωσης εκφράζεται μέσω του γενικότερου τύπου

$$S(t;x)=S_0\left(t\cdot\lambda(x)\right),$$

όπου S_0 είναι μια βασική συνάρτηση επιβίωσης και $\lambda(x)$ μια θετική συνάρτηση των συμμεταβλητών x_i, όπως αναλόγως ισχύει στο μοντέλο αναλογικής διακινδύνευσης.

Γενικώς, το μοντέλο επιταχυνόμενης διακοπής δεν εφαρμόζεται σε όλες τις περιπτώσεις όπου εφαρμόζεται το μοντέλο αναλογικών διακινδυνεύσεων. Οι κατανομές με την ιδιότητα της αναλογικής διακινδύνευσης, όπως η Weibull έχουν μονότονες συναρτήσεις διακινδύνευσης. Δε μπορούν να εκφράσουν ένα φαινόμενο με κορυφή στη συνάρτηση διακινδύνευσης, αλλά μπορούν να μοντελοποιηθούν, όπως για παράδειγμα από τη λογαριθμο - λογιστική κατανομή.

Τροποποιώντας το γενικότερο τύπο της συνάρτησης επιβίωσης ως

$$\begin{aligned}
&S(t;x)=S_0\left(t\cdot\lambda(x)\right)\\
&S(t;x)=P\left(T>t\cdot\lambda(x)\right)\\
&S(t;x)=P\left(\ln T>\ln t+\ln\lambda(x)\right)\\
&S(t;x)=s_0\left(y+\ln\lambda(x)\right)
\end{aligned},$$

όπου $y=\ln t$ και s_0 η συνάρτηση αξιοπιστίας της τυχαίας μεταβλητής $y=\ln T$ μπορεί να παρατηρηθεί ότι ένα γράφημα της $S(t;x)$ με την $\ln t$ για συγκεκριμένο x πρέπει να είναι μια οριζόντια μετατόπιση της $S_0(t)$ με την $\ln t$. Συνεπώς, η υπόθεση της επιταχυνόμενης διακοπής ισχύει αν όλες οι καμπύλες $S(t;x)$ για διαφορετικές τιμές της x διαφέρουν μεταξύ τους μόνο σε οριζόντια μετατόπιση. Για την υλοποίηση αυτού του γραφικού ελέγχου απαραίτητες είναι οι εκτιμήσεις $\hat{S}(t;x)$, συνήθως των Kaplan – Meier. Επιπροσθέτως, η μέθοδος είναι εφικτή μόνο εφόσον υπάρχουν αρκετές παρατηρήσεις για την ικανοποιητική εκτίμηση της S για κάθε x, διαφορετικά οι τιμές της x ομαδοποιούνται καταλλήλως, έτσι ώστε η κάθε ομάδα να περιέχει αρκετές μονάδες [Καρώνη (2009)].

$\Rightarrow$ Διευκρινίζεται ότι μόνο για την κατανομή Weibull και τις ειδικές περιπτώσεις της, όπως την Εκθετική κατανομή, τα μοντέλα αναλογικής διακινδύνευσης και

επιταχυνόμενης διακοπής είναι εναλλακτικές αναπαραμετρήσεις του ίδιου μοντέλου, δηλαδή καταλήγουν στα ίδια αποτελέσματα.

4.3.3 Μοντέλο αναλογικής διακινδύνευσης του Cox

Το μοντέλο αναλογικής διακινδύνευσης του Cox [Cox (1972)] είναι ένα μοντέλο παλινδρόμησης που χρησιμοποιείται συνήθως στη στατιστική στην έρευνα με βιο-ιατρικές εφαρμογές για τη διερεύνηση του συσχετισμού μεταξύ του χρόνου επιβίωσης ασθενών και μίας ή περισσοτέρων συμμεταβλητών. Ουσιαστικά επεκτείνει τις μεθόδους της Ανάλυσης Επιβίωσης ώστε να εκτιμάει ταυτόχρονα την επίδραση πολλών παραγόντων κινδύνου στο χρόνο επιβίωσης.

Αναλυτικά, η εφαρμογή του μοντέλου βασίζεται στο γεγονός ότι υπάρχει ένα διάνυσμα συμμεταβλητών x περιλαμβάνοντας τόσο ποσοτικές όσο και ποιοτικές μεταβλητές, οι οποίες δρουν στη συνάρτηση διακινδύνευσης μέσω της σχέσης

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

όπου $h_0(t)$ μια βασική συνάρτηση διακινδύνευσης, ανάλογη εκείνης της παραγράφου 4.3.1 και β ένα διάνυσμα συντελεστών, οι οποίοι εκφράζουν ποσοτικά την επίδραση της καθεμιάς από τις μεταβλητές x. Διευκρινίζεται ότι το μοντέλο του Cox δεν έχει σταθερό όρο, διότι έχει απορροφηθεί στο $h_0(t)$, το οποίο δεν προσδιορίζεται με τις συνηθισμένες διαδικασίες, οπότε ο αναγνώστης δύναται να γνωρίζει ή να μπορεί να βρει μόνο το διάνυσμα των συντελεστών β. Για αυτό το λόγο, το μοντέλο αναλογικής διακινδύνευσης του Cox θεωρείται ημι – παραμετρικό.

Επιπλέον, οι συμμεταβλητές x δρουν στη συνάρτηση επιβίωσης μέσω της σχέσης

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

όπου $S_0(t)$ μια βασική συνάρτηση επιβίωσης.

Διευκρινίζεται ότι η ανεξαρτησία της διακινδύνευσης ή της επιβίωσης από τη μεταβλητή x_i σημαίνει ότι $\beta_i = 0$.

Μιας και πρόκειται για ημι – παραμετρικό μοντέλο διερευνάται τρόπος εκτίμησης των παραμέτρων. Η εκτίμηση αυτή επιτυγχάνεται μέσω της μεθόδου μερικής

πιθανοφάνειας. Σύμφωνα με το συμβολισμό της παραγράφου 4.1, κάποιος μπορεί να υποθέσει ότι υπάρχουν $t_{(1)} < t_{(2)} < ... < t_{(k)}$ διακεκριμένες χρονικές στιγμές, κατά τις οποίες παύουν να λειτουργούν k μονάδες. Ορίζεται ότι $d_j = 1$ (για κάθε j) μονάδες παύουν να λειτουργούν τη στιγμή $t_{(j)}$.

Επομένως, τη χρονική στιγμή $t_{(j)}$ διακόπτεται η λειτουργία μιας μονάδας με συμμεταβλητές $x_{(j)}$, οπότε η πιθανότητα να διακοπεί η λειτουργία μιας συγκεκριμένης μονάδας j, δοθέντος ότι διακόπτεται η λειτουργία μιας μονάδας από το σύνολο μονάδων σε κίνδυνο εκείνη τη χρονική στιγμή – έστω R_j ισούται με

$$P = \frac{h\left(t_{(j)}; x_{(j)}\right)}{\sum_{i \in R_j} h\left(t_{(j)}; x_{(i)}\right)} = \frac{h_0\left(t_{(j)}\right) e^{x_{(j)}'\beta}}{\sum_{i \in R_j} h_0\left(t_{(j)}\right) e^{x_{(i)}'\beta}} = \frac{e^{x_{(j)}'\beta}}{\sum_{i \in R_j} e^{x_{(i)}'\beta}},$$

μιας κι η συνάρτηση διακινδύνευσης εκφράζει τη στιγμιαία πιθανότητα διακοπής.

Ως εκ τούτου, η συνάρτηση μερικής πιθανοφάνειας του Cox για το σύνολο των δεδομένων είναι

$$L(\beta) = \prod_{j=1}^{k} \left\{ \frac{e^{x_{(j)}'\beta}}{\sum_{i \in R_j} e^{x_{(i)}'\beta}} \right\},$$

από την οποία προκύπτει η εκτιμήτρια μέγιστης πιθανοφάνειας $\hat{\beta}$ της β.

Λογαριθμίζοντας προκύπτει

$$l(\beta) = \sum_{j=1}^{k} x_{(j)}'\beta - \sum_{j=1}^{k} \ln \left\{ \sum_{i \in R_j} e^{x_{(i)}'\beta} \right\}.$$

Οπότε οι πρώτες μερικές παράγωγοι είναι

$$\frac{\partial l}{\partial \beta_r} = \sum_{j=1}^{k} x_{(j)r} - \sum_{j=1}^{k} \left[\frac{\sum_{i \in R_j} x_{(i)r} e^{x_{(i)}'\beta}}{\sum_{i \in R_j} e^{x_{(i)}'\beta}} \right]$$

και το σύστημα εξισώσεων

$$\frac{dl}{d\beta_r} = 0$$

λύνεται ως προς β με αριθμητικές μεθόδους.

Διευκρινίζεται ότι αξιοποιήθηκε ο εκτιμητής της βασικής συνάρτησης διακινδύνευσης του Breslow

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

ο οποίος δίνει την εκτιμήτρια Nelson – Aalen όταν δεν υπάρχουν συμμεταβλητές

$$\hat{H}_0(t) = \sum_{t_{(j)} \le t} \frac{d_j}{n_j}.$$

4.4 Έλεγχος μέσω υπολοίπων

Ο έλεγχος μέσω υπολοίπων ενός στατιστικού μοντέλου μετά από την προσαρμογή του αποτελεί έναν αξιόπιστο τρόπο εξέτασης της καταλληλότητας του μοντέλου. Ουσιαστικά πρόκειται για την εκτίμηση, μέσω της οποίας εξετάζονται κάποιες βασικές υποθέσεις που πρέπει να πληρούν τα τυχαία σφάλματα που συνήθως προκύπτουν στην προσαρμογή του γενικού γραμμικού μοντέλου παλινδρόμησης. Για αυτό το λόγο, τα υπόλοιπα στη συνήθη μορφή τους είναι

$$\hat{\varepsilon}_i = y_i - \hat{y}_i,$$

δηλαδή η διαφορά μεταξύ της παρατηρούμενης τιμής y_i και της προσαρμοσμένης $\hat{y}_i$.

Τα υπόλοιπα δείχνουν κατά πόσο τα δεδομένα συμφωνούν με τις προϋποθέσεις του μοντέλου, καθώς και με το τι προβλέπει αυτό όχι μόνο συνολικά, αλλά και

μεμονωμένα ελέγχοντας κυρίως την αναλογικότητα των κινδύνων και την ύπαρξη σημείων επιρροής, δηλαδή παρατηρήσεων που ασκούν μεγάλη επιρροή στη διαμόρφωση της εκτιμημένης συνάρτησης παλινδρόμησης υπό την έννοια ότι αν αυτές οι παρατηρήσεις είναι σημαντικές και παραληφθούν ή έχουν προστεθεί λόγω λάθους μέτρησης ενώ πρέπει να αφαιρεθούν αλλάζουν τελείως τα αποτελέσματα της εκάστοτε μελέτης. Η εξέταση αυτών των υπολοίπων γίνεται με μια πληθώρα τρόπων και συνήθως μέσω γραφικών παραστάσεων ως προς την κατανομή τους, την ύπαρξη έκτροπων παρατηρήσεων (outliers), την εξάρτησή τους με τα αμέσως προηγούμενα υπόλοιπα, τη συσχέτισή τους με τη σειρά των δεδομένων και πολλές άλλες μεθόδους. Γενικώς υπάρχουν πολλά διαφορετικά είδη υπολοίπων. Στην παρούσα εργασία θα εξεταστούν αρχικά τα υπόλοιπα Cox – Snell και τα υπόλοιπα Schoenfeld, μιας και βρίσκουν πιο άμεση και εύκολη εφαρμογή στα μοντέλα παλινδρόμησης Weibull και αναλογικής διακινδύνευσης του Cox. Στη συνέχεια, θα γίνει αναφορά στα υπόλοιπα Martingale, στα υπόλοιπα Deviance και στα σημεία επιρροής DFBETAS, ενώ η παράγραφος θα ολοκληρωθεί με την ελεγχοσυνάρτηση GLOBAL, η οποία αξιοποιείται συνήθως στην εξέταση για την υπόθεση της αναλογικής διακινδύνευσης.

Η μορφή όλων των υπολοίπων βασίζεται στο αρχικό πρότυπο

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}' x_i ,$$

δηλαδή στη διαφορά μεταξύ της παρατηρούμενης τιμής y_i και της προσαρμοσμένης $\hat{y}_i$.

4.4.1 Υπόλοιπα Cox – Snell

Τα υπόλοιπα Cox – Snell παρουσιάστηκαν το 1968 από τους Sir David Cox και E. Joyce Snell για την αξιολόγηση της εγκυρότητας μιας συνάρτησης επιβίωσης που είχε προταθεί για ένα σύνολο δεδομένων ζωής/επιβίωσης [Cox (1968)].

Ορίζονται ως

$$-\ln \hat{S}\left(t_i; x_i\right) = \hat{H}\left(t_i; x_i\right) = \hat{\varepsilon},$$

όπου $\hat{S}(\cdot)$ και $\hat{H}(\cdot)$ είναι εκτιμήσεις της συνάρτησης επιβίωσης/αξιοπιστίας και της σωρευτικής συνάρτησης διακινδύνευσης αντίστοιχα. Για να είναι κατάλληλο το μοντέλο αρκεί να εξεταστεί γραφικά εάν οι τιμές των υπολοίπων ακολουθούν την εκθετική κατανομή με παράμετρο τη μονάδα. Αν ισχύει κάτι τέτοιο, τότε το μοντέλο προσαρμόστηκε σωστά στα δεδομένα.

Σε περίπτωση που υπάρχουν από δεξιά αποκομμένες παρατηρήσεις, τότε συστήνονται τα διορθωμένα υπόλοιπα Cox – Snell, τα οποία εκφράζονται μέσω της σχέσης

$$1-\ln \hat{S}\left(t_i; x_i\right) = \hat{\varepsilon}.$$

Υπενθυμίζεται από βασικές γνώσεις της στατιστικής αναφορικά με τις κατανομές, ότι η συνάρτηση πυκνότητας πιθανότητας της κατανομής Weibull [Caroni (2017)] είναι η

$$f\left(t\right) = \eta\alpha^{-\eta}t^{\eta-1}\exp\left\{-\left(\frac{t}{\alpha}\right)^{\eta}\right\}, t > 0,$$

όπου $\alpha > 0$ είναι η παράμετρος κλίμακας και $\eta > 0$ είναι η παράμετρος σχήματος.

Επιπλέον, η συνάρτηση πυκνότητας πιθανότητας της κατανομής Gumbel είναι η

$$f\left(t\right) = \sigma^{-1}\exp\left\{\frac{t-\mu}{\sigma}\right\}S\left(t\right), -\infty < t < \infty,$$

όπου μ είναι η παράμετρος θέσης, σ είναι η παράμετρος κλίμακας και $S(t)$ είναι η συνάρτηση αξιοπιστίας της κατανομής Gumbel.

Δεδομένου ότι αυτές οι δύο κατανομές είναι άρρηκτα συνδεδεμένες μέσω της σχέσης

$$T \sim Weibull \Leftrightarrow \ln T \sim Gumbel$$

προκύπτει ότι οι παράμετροι των δύο κατανομών συνδέονται με τη σχέση [Καρώνη (2009)]

$$\begin{aligned} \mu &= \ln \alpha \\ \sigma &= \eta^{-1} \end{aligned} .$$

Οπότε, στην περίπτωση του μοντέλου παλινδρόμησης της κατανομής Weibull, τα υπόλοιπα Cox – Snell ορίζονται ως

$$H\left(t|x\right) = \exp\left(\frac{\ln t - \hat{\beta}' x}{\hat{\sigma}}\right) = \exp\left(\varepsilon\right) \sim \exp onential\left(1\right),$$

όπου

$$\varepsilon = \frac{\ln t - \beta' x}{\sigma} \sim Gumbel\left(0,1\right)$$

και

$$S\left(t|x\right) = \exp\left\{-\exp\left(\frac{\ln t - \hat{\beta}' x}{\hat{\sigma}}\right)\right\}.$$

Γενικώς, στα παραμετρικά μοντέλα τα υπόλοιπα Cox – Snell θεωρούνται πολύ χρήσιμα.

Διευκρινίζεται ότι επειδή η ποσότητα

$$\varepsilon = \frac{\ln t - \beta' x}{\sigma} \sim Gumbel\left(0,1\right)$$

είναι τυποποιημένη, αφού ουσιαστικά αφαιρείται η μέση τιμή και διαιρείται με την τυπική απόκλιση, θα αναφέρεται στα τυποποιημένα υπόλοιπα (Standardized residuals).

Στην περίπτωση του μοντέλου αναλογικής διακινδύνευσης του Cox, τα υπόλοιπα Cox – Snell ορίζονται ως

$$\hat{H}\left(t_i; x_i\right) = \hat{H}_0\left(t_i\right) e^{\hat{\beta}' x_i},$$

όπου $\hat{H}_0(\cdot)$ είναι μια μη-παραμετρική εκτίμηση της σωρευτικής βασικής συνάρτησης διακινδύνευσης.

Η ανωτέρω εκτίμηση για τα υπόλοιπα Cox – Snell αποδεικνύεται όχι και τόσο χρήσιμη για το μοντέλο του Cox, μιας και δεν είναι γνωστή η βασική συνάρτηση διακινδύνευσης $h_0(\cdot)$, συνεπώς ούτε και η $\hat{H}_0(\cdot)$ μπορεί να προσδιοριστεί. Για αυτό το λόγο, στην περίπτωση του μοντέλου του Cox συστήνονται τα υπόλοιπα Schoenfeld.

4.4.2 Υπόλοιπα Schoenfeld

Στην παράγραφο 4.3.3 αποδείχτηκε ότι η πιθανότητα να διακοπεί η λειτουργία μιας συγκεκριμένης μονάδας j, δοθέντος ότι διακόπτεται η λειτουργία μιας μονάδας από το σύνολο μονάδων σε κίνδυνο R_j εκείνη τη χρονική στιγμή ισούται με

$$P = p_j = \frac{e^{x_{(j)}'\beta}}{\sum_{i \in R_j} e^{x_{(i)}'\beta}}.$$

Ωστόσο, αν δεν είναι γνωστό ποια μονάδα από το σύνολο R_j πρόκειται να διακοπεί τη χρονική στιγμή $t_{(j)}$, τότε η τιμή των συμμεταβλητών x αυτής της μονάδας είναι τυχαία μεταβλητή με αναμενόμενη τιμή

$$E\left(x \middle| R_j\right) = \sum_{k \in R_j} x_k p_k = \frac{\sum_{k \in R_j} x_k e^{x_{(k)}'\beta}}{\sum_{i \in R_j} e^{x_{(i)}'\beta}}.$$

Συνεπώς, τα υπόλοιπα Schoenfeld ορίζονται ως

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

όπου η $\hat{E}$ δηλώνει ότι οι παράμετροι β έχουν αντικατασταθεί από τις εκτιμήσεις $\hat{\beta}$ [Schoenfeld (1982), Καρώνη (2009)]. Δηλαδή τα υπόλοιπα Schoenfeld ορίζονται ως η διαφορά ανάμεσα στην προβλεπόμενη παρατήρηση και την εκτίμηση της δεσμευμένης αναμενόμενης τιμής.

Παρατηρώντας προσεκτικά τα υπόλοιπα Schoenfeld, κάποιος μπορεί να διαπιστώσει ότι δεν προσδιορίζονται από τις τιμές της εξαρτημένης μεταβλητής, όπως για παράδειγμα το χρόνο t, αλλά από τις συμμεταβλητές x. Αυτό αποτελεί ένα χαρακτηριστικό των υπολοίπων Schoenfeld, το οποίο έρχεται σε αντίφαση με ό,τι συμβαίνει γενικά στα κλασικά μοντέλα παλινδρόμησης. Επιπροσθέτως, τα συγκεκριμένα υπόλοιπα αποτελούν διανύσματα, οπότε κάθε μη – αποκομμένη παρατήρηση έχει τόσα υπόλοιπα όσα είναι και οι συμμεταβλητές. Επιπλέον, τα υπόλοιπα Schoenfeld ορίζονται μόνο για τις μονάδες με παρατηρούμενο χρόνο διακοπής κι όχι για όλο το δείγμα. Για παράδειγμα για έναν ασθενή, για τον οποίο ο χρόνος επιβίωσής του είναι αποκομμένος, τα υπόλοιπα Schoenfeld θα πάρουν την τιμή 0. Στην περίπτωση δε που ικανοποιούνται κατάλληλες προϋποθέσεις, η κατανομή των υπολοίπων Schoenfeld ως προς το χρόνο σε μια γραφική παράσταση θα είναι τυχαία γύρω από το 0.

Στη συνέχεια ορίζονται τα κλιμακοποιημένα (scaled) υπόλοιπα, τα οποία πολλές φορές προτιμούνται για το σχεδιασμό γραφικών παραστάσεων και την εξέταση της υπόθεσης αναλογικών διακινδυνεύσεων.

Έστω $\hat{r}_j$ το διάνυσμα των υπολοίπων κατά τη χρονική στιγμή $t_{(j)}$, s το πλήθος των μη – αποκομμένων παρατηρήσεων και $\hat{V}\left(\hat{\beta}\right)$ ο πίνακας διασποράς της $\hat{\beta}$. Τότε τα κλιμακοποιημένα (scaled) υπόλοιπα ορίζονται ως

$$r_j^* = s \cdot \hat{V}\left(\hat{\beta}\right) \cdot \hat{r}_j .$$

4.4.3 Άλλα υπόλοιπα και σημεία επιρροής

Υπόλοιπα Martingale

Ένα μεγάλο μειονέκτημα που έχουν τα παραπάνω υπόλοιπα και ειδικότερα τα Cox – Snell είναι ότι δεν παρέχουν πολλές πληροφορίες για την αιτία, για την οποία οι υποθέσεις ενός μοντέλου παραβιάζονται. Θα ήταν επιθυμητό κάθε υπόλοιπο να μπορεί να πάρει μια θετική ή αρνητική τιμή που να δείχνει εάν, για παράδειγμα, ένας ασθενής επέζησε περισσότερο ή λιγότερο, σύμφωνα με την πρόβλεψη του κάθε μοντέλου και για πόσο χρονικό διάστημα. Ως εκ τούτου, σύμφωνα και με την εισαγωγή στην αρχή αυτού του υποκεφαλαίου 4.4, εάν μια διαδικασία που σχετίζεται με το χρόνο μέχρι να συμβεί ένα γεγονός αναλυθεί χρησιμοποιώντας μαθηματικά από τη θεωρία μέτρου, τότε μπορούν να προκύψουν τα ίδια υπόλοιπα ως η διαφορά μεταξύ του παρατηρούμενου και του αναμενόμενου αριθμού συμβάντων έως το χρόνο t_i που θα συμβούν (ενσωματώνοντας το χρονικό διάστημα στο οποίο ο ασθενής βρισκόταν σε κίνδυνο) ως

$$\hat{M}_i = \delta_i - \hat{H}_0\left(t_i\right)e^{x_i'\hat{\beta}},$$

τα οποία ονομάζονται υπόλοιπα Martingale [Caroni (2004)] και χρησιμοποιούνται για την ανίχνευση σημείων επιρροής και κυρίως για τον προσδιορισμό της συναρτησιακής μορφής που θα πάρει μία μεταβλητή που πρόκειται να εισαχθεί στο εκάστοτε μοντέλο.

Αν $\delta_i = 0$, τότε η παραπάνω σχέση θα πάρει αρνητική τιμή που συνεπάγεται ότι ο ασθενής θα ζήσει περισσότερο από ό,τι αναμενόταν (ή ότι θα αποτελεί αποκομμένη παρατήρηση).
Αν $\delta_i = 1$, τότε συνεπάγεται ότι συνέβη το γεγονός, δηλαδή ότι ο ασθενής υποτροπίασε ή απεβίωσε νωρίτερα από ό,τι αναμενόταν (σύμφωνα πάντα με το μοντέλο).

Σε γενικές γραμμές, τα υπόλοιπα Martingale είναι πολύ χρήσιμα και μπορούν να χρησιμοποιηθούν για πολλούς από τους συνήθεις σκοπούς, για τους οποίους κάποιος χρησιμοποιεί τα υπόλοιπα (residuals) σε διάφορα μοντέλα, όπως για παράδειγμα για

τον προσδιορισμό έκτροπων παρατηρήσεων ή για την ανίχνευση της μη γραμμικότητας .

Ωστόσο, το κύριο μειονέκτημα στα υπόλοιπα Martingale είναι η σαφής ασυμμετρία τους, αφού το άνω όριο τους είναι η μονάδα, αλλά δεν έχουν κανένα χαμηλότερο όριο, δηλαδή

$$\hat{M}_i \in (-\infty, +1),$$

όπου οι αρνητικές τιμές τους αντιστοιχούν σε υπόλοιπα για αποκομμένες παρατηρήσεις.

Αυτή η ασυμμετρία μπορεί να κάνει την εξαγωγή σωστών συμπερασμάτων και ερμηνειών δύσκολη. Ένας μετασχηματισμός για να αποφευχθεί αυτό το πρόβλημα οδηγεί στα υπόλοιπα Deviance.

Υπόλοιπα Deviance

Μια τεχνική για τη δημιουργία συμμετρικών, κανονικοποιημένων υπολοίπων που χρησιμοποιείται ευρέως στη γενικευμένη γραμμική μοντελοποίηση είναι η κατασκευή των υπολοίπων Deviance.

Η ιδέα πίσω από τα υπόλοιπα Deviance είναι να εξεταστεί η διαφορά μεταξύ της λογαριθμοποιημένης πιθανοφάνειας για ένα άτομο (μια παρατήρηση) i υπό ένα δεδομένο μοντέλο και της μέγιστης δυνατής λογαριθμοποιημένης πιθανοφάνειας για αυτό το άτομο

$$2\left\{\hat{l}_i - l_i\right\}.$$

Συνεπώς, δεδομένου ότι ουσιαστικά πρόκειται για έναν έλεγχο του λόγου των πιθανοφανειών, η παραπάνω ποσότητα θα πρέπει να ακολουθεί τη X_1^2- κατανομή και για να μετατραπεί σε μια ποσότητα που ακολουθεί προσεγγιστικά την κανονική κατανομή, κάποιος θα χρησιμοποιήσει τον τύπο

$$d_i = \text{sgn}\left(\hat{M}_i\right)\left[-2\left\{l_i - \hat{l}_i\right\}\right]^{\frac{1}{2}}$$
$$d_i = \text{sgn}\left(\hat{M}_i\right)\sqrt{\left[-2\left\{\hat{M}_i + \delta_i \ln\left(\delta_i - \hat{M}_i\right)\right\}\right]},$$

ο οποίος αντιπροσωπεύει τα υπόλοιπα Deviance. Με άλλα λόγια τα υπόλοιπα Deviance δίνουν το μέτρο της απόκλισης που συνεισφέρεται από κάθε παρατήρηση σε ένα δείγμα.

Συνοψίζοντας, τα υπόλοιπα Deviance αξιοποιούνται για τον έλεγχο ύπαρξης σημείων επιρροής ή άτυπων (έκτροπων) παρατηρήσεων, για αυτό το λόγο στη γραφική παράσταση αναμένεται αυτά τα υπόλοιπα να κατανέμονται συμμετρικά γύρω από το μηδέν, ενώ οι πολύ μεγάλες κατά απόλυτη τιμή παρατηρήσεις των υπολοίπων Deviance δηλώνουν την ύπαρξη πιθανών έκτροπων παρατηρήσεων. Παράλληλα με αυτά, χρήσιμα αποδεικνύονται για τους ίδιους ελέγχους κι ο τιμές των DFBETAS.

Σημεία επιρροής DFBETAS

Η ιδέα πίσω από τα υπόλοιπα DFBETAS είναι πολύ απλή. Έστω $\hat{\beta}_{j(i)}$ η εκτίμηση του συντελεστή β_j όταν η παρατήρηση i παραλείπεται από το μοντέλο και $\hat{\beta}_j$ η εκτίμηση του συντελεστή β_j. Τότε, τα υπόλοιπα DFBETAS για τη συμμεταβλητη j δεδομένης της παράλειψης i ορίζονται ως

$$DFBETAS_{ji} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{S_{(i)}^2 c_{jj}}},\ i = 1, 2, ..., n,\ \ j = 0, 1, 2, ..., k,$$

όπου n το μέγεθος του δείγματος, k ο αριθμός των επεξηγηματικών μεταβλητών στο μοντέλο, $S_{(i)}^2$ η διασπορά των υπολοίπων, όταν η παρατήρηση i έχει αφεθεί εκτός

μοντέλου και c_{jj} ο αριθμός των επαναλήψεων της παρατήρησης x_j ή αλλιώς το j-οστό διαγώνιο στοιχείο του πίνακα $\left(X'X\right)^{-1}$.

Ισχύει ότι εάν

$$\left|DFBETAS_{ji}\right| > \frac{2}{\sqrt{n}},$$

τότε υπάρχει μιαν ένδειξη ότι η παρατήρηση i μπορεί να ασκεί επιρροή στην εκτίμηση του συντελεστή β_j.

Οι γραφικές παραστάσεις σχετικά με τα υπόλοιπα DFBETAS παρουσιάζουν πάντα μεγάλο ενδιαφέρον, επειδή συμβάλουν στην ουσιαστική κατανόηση της δομής ακόμη και των πιο περίπλοκων μοντέλων. Επιπλέον μπορούν να δείξουν εάν η εκτίμηση ενός συντελεστή κυριαρχείται από λίγα άτομα, κάτι που θα πρέπει ξεκάθαρα να προβληματίσει τον παρατηρητή.

Ελεγχοσυνάρτηση GLOBAL

Η ελεγχοσυνάρτηση GLOBAL χρησιμοποιείται ώστε να ελεγχθεί εάν ισχύει η υπόθεση της αναλογικής διακινδύνευσης σε κάποιο μοντέλο. Ο τύπος της δίνεται μέσω της σχέσης

$$T = \frac{\left(g-\bar{g}\right)'S^{*}I\left(\hat{\beta}\right)S^{*}'\left(g-\bar{g}\right)}{k\sum_{j=1}^{k}\left(g_j-\bar{g}\right)^2} \sim X_p^2,$$

όπου k είναι ο αριθμός των γεγονότων, $g-\bar{g}$ είναι το διαστατικό διάνυσμα k με j-οστό στοιχείο το $g_j-\bar{g}$, S^{*} είναι ο $k\times p$ πίνακας των κλιμακοποιημένων (scaled) υπολοίπων Schoenfeld, p ο αριθμός των μεταβλητών στο μοντέλο και $I^{-1}\left(\hat{\beta}\right)=\hat{V}\left(\hat{\beta}\right)$.

Ισχύει ότι απορρίπτεται η υπόθεση της αναλογικής διακινδύνευσης για μεγάλες τιμές της ελεγχοσυνάρτησης T.

5 Πειραματική Προσέγγιση

5.1 Εισαγωγή

Η σχετική θεωρία θα εφαρμοστεί σε πραγματικά δεδομένα που προέρχονται από τη ‘Γερμανική Ομάδα Μελέτης για τον Καρκίνο του Μαστού’ (the German Breast Cancer Study Group). Συγκεκριμένα, από τον Ιούλιο του 1984 έως τον Δεκέμβριο του 1989, η ‘Γερμανική Ομάδα Μελέτης για τον Καρκίνο του Μαστού’ [Sauerbrei (1999)] ανέλαβε 720 ασθενείς σε πρώιμο στάδιο καρκίνου του μαστού στην Ολοκληρωμένη Μελέτη Κοόρτης (Comprehensive Cohort Study). Τυχαιοποιημένοι και μη τυχαιοποιημένοι ασθενείς ήταν επιλέξιμοι και περίπου τα δύο τρίτα εισήχθησαν στο τυχαιοποιημένο μέρος. Η αποτελεσματικότητα τριών έναντι έξι κύκλων χημειοθεραπείας και επιπρόσθετης ορμονικής θεραπείας με ταμοξιφαΐνη διερευνήθηκαν και μετά από κάποιο χρόνο παρακολούθησης σχεδόν 5 ετών διαπιστώθηκε ότι 312 ασθενείς είχαν τουλάχιστον μιαν υποτροπή της νόσου ή είχαν πεθάνει.

Σε αυτό το έγγραφο αναλύεται ο χρόνος επιβίωσης χωρίς υποτροπή των 686 ασθενών (με 299 συμβάντα) που είχαν πλήρη δεδομένα για τους τυπικούς παράγοντες της ηλικίας, του μεγέθους του όγκου, του αριθμού των θετικών λεμφαδένων, της κατάστασης του υποδοχέα προγεστερόνης και οιστρογόνου, της κατάστασης εμμηνόπαυσης, καθώς και του βαθμού του όγκου. Στο τυχαιοποιημένο μέρος της μελέτης, ο αριθμός των κύκλων χημειοθεραπείας δεν έχει καμία επίδραση στην επιβίωση χωρίς υποτροπή και όλες οι αναλύσεις προσαρμόζονται με χρήση ή όχι ορμονικής θεραπείας [Sauerbrei (1999)].

Στον πίνακα 3 παρουσιάζονται οι συμμεταβλητές [Sauerbrei (1999)] που θα ληφθούν υπόψη για τη διεκπεραίωση της ανάλυσης και σχετίζονται με το μέγεθος του όγκου υπολογισμένο σε χιλιοστά (mm) και το βαθμό του όγκου, ο οποίος παίρνει τιμές από 1 που δηλώνει καλοήθη κατάσταση μέχρι και την τιμή 3 που μαρτυράει επικίνδυνη κατάσταση για κακοήθη καρκίνο. Επίσης, μεγάλο ρόλο διαδραματίζει η ηλικία. Ο καρκίνος του μαστού μπορεί να προκύψει σε οποιαδήποτε ηλικία μετά την εφηβεία, αλλά τα ποσοστά αυξάνονται όσο αυξάνονται και οι ηλικιακές κλίμακες. Οι περισσότερες περιπτώσεις παρουσιάζονται μετά από την ηλικία των 50 ετών, ενώ είναι σπάνιος σε γυναίκες ηλικίας κάτω των 35 ετών (5% των περιπτώσεων), με εξαίρεση τις γυναίκες που έχουν κληρονομική προδιάθεση. Επιπροσθέτως, κάποιος

μπορεί να δει τις μεταβλητές του αριθμού των θετικών λεμφαδένων στη μασχάλη, οι οποίοι αποτελούν βασικό προγνωστικό παράγοντα για την εξέλιξη της νόσου (όσο περισσότεροι θετικοί λεμφαδένες υπάρχουν τόσο αυξάνεται το στάδιο του καρκίνου και μειώνεται η επιβίωση των ασθενών), καθώς και της ένδειξης ορμονοθεραπείας με την τιμή 0 αν δεν υπήρχε κάποια αγωγή και 1 αν γινόταν θεραπεία με ορμόνες. Γενικώς, εάν στον όγκο υπάρχουν ορμονικοί υποδοχείς μπορεί να χρησιμοποιηθεί ορμονοθεραπεία μόνη ή σε συνδυασμό με χημειοθεραπεία ή/και ακτινοθεραπεία. Συνήθως, όπως συνέβη και στην παρούσα μελέτη χορηγείται ταμοξιφαΐνη (tamoxifen).

Στη συνέχεια, άλλη μεταβλητή που παρατηρείται είναι εκείνη της κατάστασης εμμηνόπαυσης με τιμές 0 αν η γυναίκα βρισκόταν πριν από αυτή τη φάση και 1 αν είχε περάσει τη διαδικασία λόγω ηλικίας, ενώ οι δύο τελευταίες συμμεταβλητές που θα συνυπολογιστούν για την εξαγωγή συμπερασμάτων είναι εκείνες της κατάστασης του υποδοχέα προγεστερόνης και οιστρογόνου υπολογισμένα σε fmol. Στοιχεία υποδηλώνουν πως γυναίκες με καθυστερημένη εμμηνόπαυση (μετά τα 55) αντιμετωπίζουν αυξημένο κίνδυνο εμφάνισης καρκίνου στο μαστό, ενώ 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. Δηλώνεται ότι οι υποδοχείς οιστρογόνου και προγεστερόνης είναι πρωτεΐνες που βρίσκονται στην επιφάνεια των κυττάρων του μαστού. Το οιστρογόνο και η προγεστερόνη συνδέονται με αυτούς τους υποδοχείς για να σηματοδοτήσουν το κύτταρο ώστε να αναπτυχθεί και να χωριστεί. Όλα τα κύτταρα του μαστού έχουν αυτούς τους υποδοχείς, αλλά βρίσκονται σε πολύ μεγαλύτερο αριθμό σε κύτταρα καρκίνου του μαστού τα οποία θεωρούνται θετικά.

Τέλος, ο πίνακας 4 παρουσιάζει ένα δείγμα από τα πλήρη δεδομένα των 686 ασθενών από την παραπάνω μελέτη σχετικά με το χρόνο επιβίωσης χωρίς υποτροπή συμπεριλαμβάνοντας και τις αποκομμένες τιμές.

Τα δεδομένα αυτά θα εισαχθούν στο στατιστικό πακέτο Minitab και τη γλώσσα προγραμματισμού R, ώστε να εφαρμοστούν όλες οι προαναφερθείσες μέθοδοι των ανωτέρω κεφαλαίων και να εξαχθούν συμπεράσματα για το κατά πόσο οι διάφορες συμμεταβλητές μπόρεσαν να επηρεάσουν το χρόνο επιβίωσης σε ημέρες των

ασθενών γυναικών που είχαν διαγνωστεί με καρκίνο στο μαστό μέχρι να υποτροπιάσουν ή να αποβιώσουν.

ΜΕΤΑΒΛΗΤΕΣ

Id. Ο αριθμός των ασθενών σε αύξουσα σειρά

Age. Η ηλικία των ασθενών σε έτη

Tsize. Μέγεθος όγκου σε mm

Pnodes. Αριθμός θετικών λεμφαδένων

Progrec. Κατάσταση του υποδοχέα προγεστερόνης σε fmol

Estrec. Κατάσταση του υποδοχέα οιστρογόνου σε fmol

Hormone. Ένδειξη ορμονοθεραπείας, η οποία παίρνει τη τιμή no = 0 ή yes = 1

Meno. Κατάσταση εμμηνόπαυσης, η οποία παίρνει τη τιμή premenopausal = 0 ή postmenopausal = 1

Tgrad. Βαθμός όγκου με τιμές 1, 2, 3 (το 1 υποδηλώνει πιο καλοήθη, ενώ το 3 πιο κακοήθη όγκο)

Time. Χρόνος επιβίωσης σε ημέρες μέχρι την υποτροπή

Cens. Δείκτης αποκοπής με 0:censored (αποκομμένη τιμή), 1:event (συμβάν)

Πίνακας 3 Συμμεταβλητές της έρευνας

id	age	tsize	pnodes	progrec	estrec	hormone	meno	tgrad	time	cens
1	70	21	3	48	66	0	1	2	1814	1
2	56	12	7	61	77	1	1	2	2018	1
3	58	35	9	52	271	1	1	2	712	1
4	59	17	4	60	29	1	1	2	1807	1
5	73	35	1	26	65	0	1	2	772	1
6	32	57	24	0	13	0	0	3	448	1
7	59	8	2	181	0	1	1	2	2172	0
8	65	16	1	192	25	0	1	2	2161	0

Πίνακας 4 Δείγμα από τα δεδομένα των ασθενών σχετικά με τις συμμεταβλητές του πίνακα 3

Ως υποθέσεις ελέγχου σε σχέση με τους ελέγχους log – rank, Wilcoxon, Wald κι όσους άλλους αναλύθηκαν στα κεφάλαια 3 και 4 θα θεωρηθούν οι κάτωθι.

$H_0 =$ οι κατανομές ταυτίζονται

$H_1 =$ οι κατανομές διαφέρουν μεταξύ τους

5.2 Μέρος Α

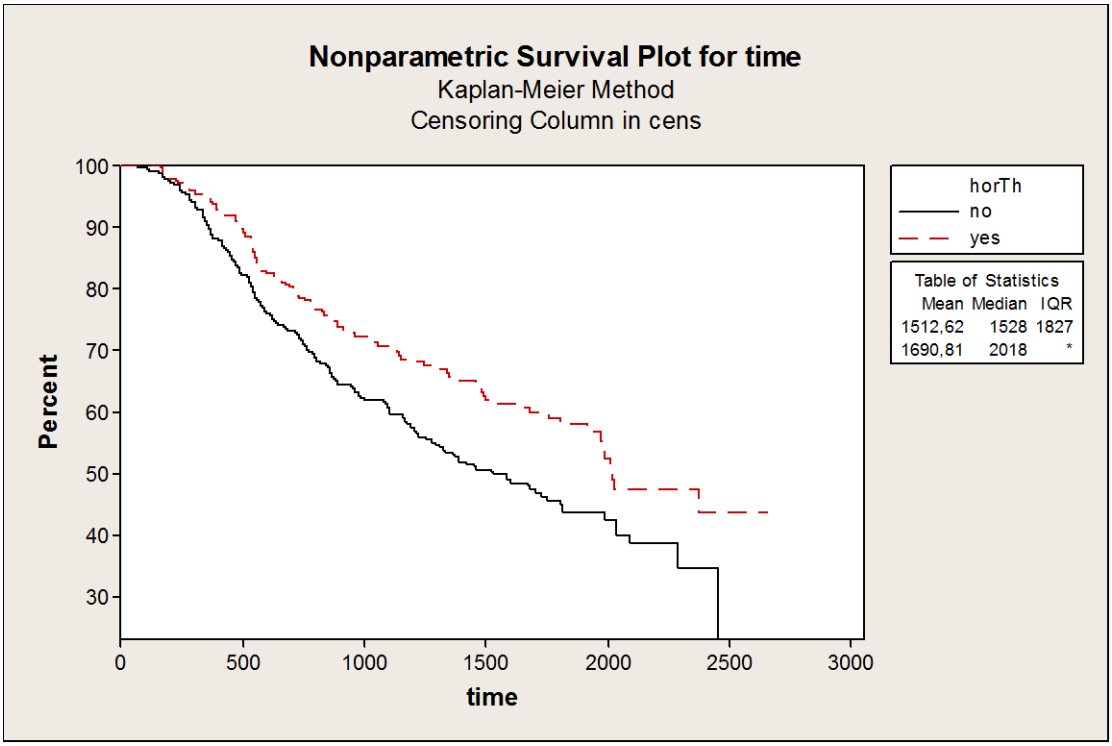
5.2.1 Εξέταση δεδομένων με βάση μία μόνο συμμεταβλητή

Σε αυτό το κεφάλαιο, αρχικά, μέσω του Minitab θα εξαχθούν οι εκτιμήσεις και η γραφική παράσταση της Kaplan – Meier, καθώς και οι έλεγχοι log – rank και Wilcoxon, ώστε να γίνει σύγκριση ανάμεσα στις πιθανότητες επιβίωσης ασθενών γυναικών που υποβλήθηκαν σε ορμονοθεραπεία ή όχι. Κατόπιν θα γίνουν οι ίδιοι έλεγχοι, ώστε να εξαχθούν αντίστοιχα αποτελέσματα σχετικά με το βαθμό του όγκου που είχαν ή την κατάσταση της εμμηνόπαυσής τους (αν είχε προέλθει αυτό το στάδιο στη ζωή τους ή όχι). Τα αποτελέσματα που προκύπτουν ως προς την Kaplan – Meier για την ορμονοθεραπεία, το βαθμό του όγκου και την κατάσταση της εμμηνόπαυσης παρατίθενται στο παράρτημα Π.1. Τα αποτελέσματα για τους ελέγχους log – rank και Wilcoxon φαίνονται αντίστοιχα στους πίνακες 5, 6 και 7. Οι γραφικές παραστάσεις της Kaplan – Meier φαίνονται στα γραφήματα 5.2.1, 5.2.2 και 5.2.3.

Για την Ορμονοθεραπεία

Distribution Analysis: time by horTh			
Test Statistics			
Method	Chi-Square	DF	**P-Value**
Log-Rank	8.56478	1	0.003
Wilcoxon	8.36141	1	0.004

Πίνακας 5 Αποτελέσματα για τους ελέγχους log – rank και Wilcoxon για τη σύγκριση ανάμεσα στις πιθανότητες επιβίωσης ασθενών γυναικών που υποβλήθηκαν σε ορμονοθεραπεία ή όχι



Γράφημα 5.2.1 Εκτίμηση Kaplan - Meier για την ορμονοθεραπεία

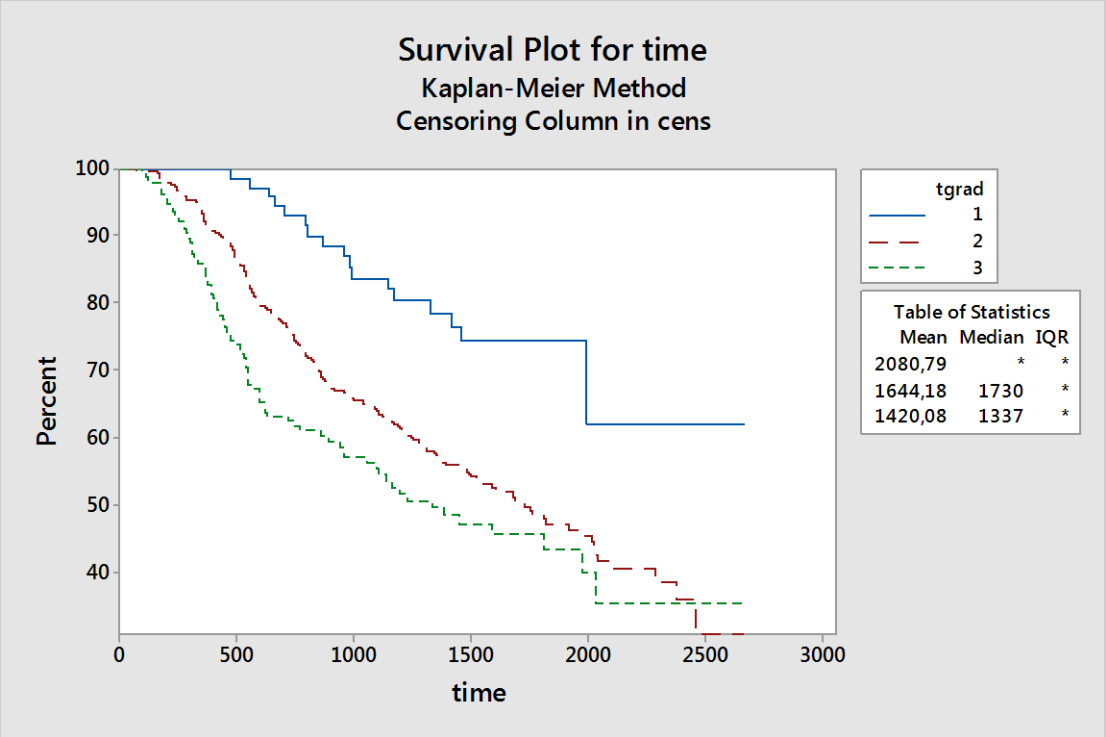
Για το Μέγεθος του Όγκου

Distribution Analysis: time by tgrad

Test Statistics

Method	Chi-Square	DF	P-Value
Log-Rank	21,0944	2	0,000
Wilcoxon	27,2049	2	0,000

Πίνακας 6 Αποτελέσματα για τους ελέγχους log – rank και Wilcoxon για τη σύγκριση ανάμεσα στις πιθανότητες επιβίωσης ασθενών γυναικών αναλόγως με το βαθμό του όγκου που είχαν



Γράφημα 5.2.2 Εκτίμηση Kaplan - Meier για το βαθμό του όγκου

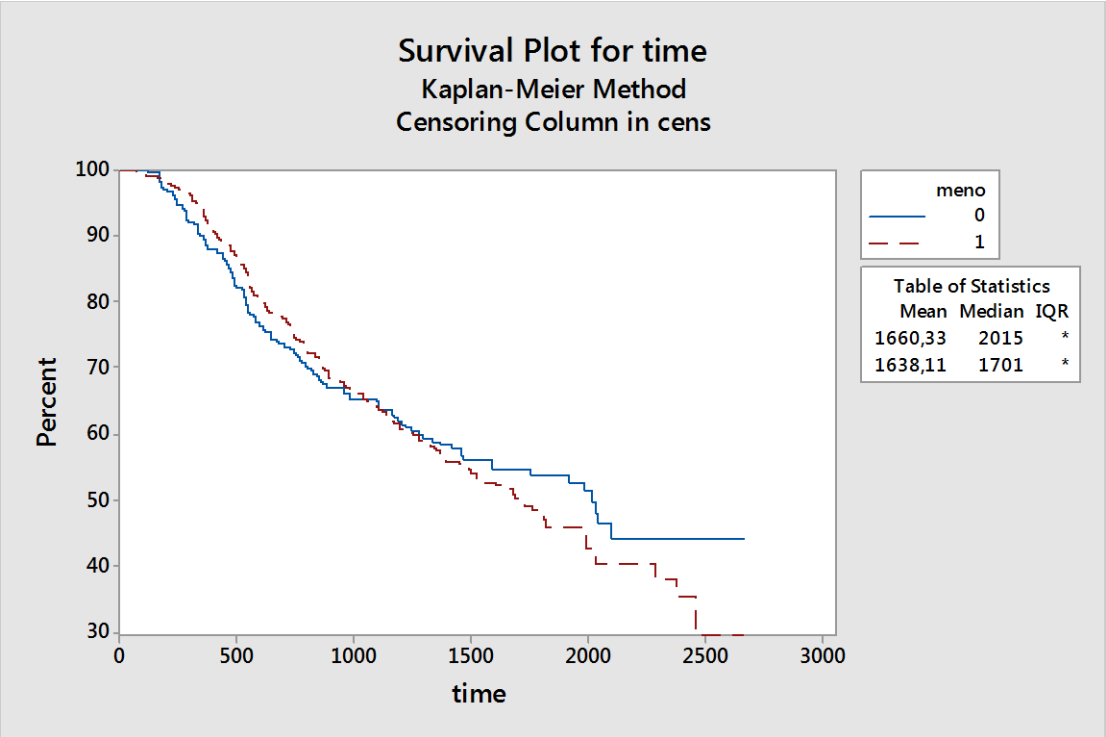
Για την Εμμηνόπαυση

Distribution Analysis: time by meno

Test Statistics

Method	Chi-Square	DF	P-Value
Log-Rank	0,280190	1	0,597
Wilcoxon	0,117589	1	0,732

Πίνακας 7 Αποτελέσματα για τους ελέγχους log – rank και Wilcoxon για τη σύγκριση ανάμεσα στις πιθανότητες επιβίωσης ασθενών γυναικών σχετικά με την κατάσταση της εμμηνόπαυσης



Γράφημα 5.2.3 Εκτίμηση Kaplan - Meier για την εμμηνόπαυση

5.2.2 Συμπεράσματα

Στην παράγραφο 5.2.1 εφαρμόστηκαν οι έλεγχοι log – rank και Wilcoxon, καθώς επίσης δημιουργήθηκαν οι γραφικές παραστάσεις των Kaplan – Meier, οι οποίες παρέχουν μια πιο άμεση (οπτική) εικόνα των αποτελεσμάτων που σχετίζονται με τις πιθανότητες επιβίωσης των γυναικών όταν υπεισέρχονται στο πείραμα οι συμμεταβλητές της ορμονοθεραπείας, του μεγέθους του όγκου που είχαν στο στήθος και της κατάστασης της εμμηνόπαυσης. Ως εκ τούτου εξάγονται κάποια συμπεράσματα.

Σε ό,τι αφορά στη συμμεταβλητή της ορμονοθεραπείας, κάποιος παρατηρεί από τη γραφική παράσταση της Kaplan – Meier στο γράφημα 5.2.1 ότι στην αρχή του άξονα του χρόνου και συγκεκριμένα τις πρώτες 300 περίπου ημέρες, οι εκτιμήτριες παίρνουν σχεδόν τις ίδιες τιμές, δηλαδή στην έναρξη των συμπτωμάτων φαίνεται σα να μην έπαιζε τόσο ρόλο η πρόσληψη ή όχι κάποιας θεραπείας ως προς την έκβαση της ασθένειας. Ωστόσο, αυτή η ένδειξη ίσως να μην είναι τόσο ρεαλιστική, αφού αφενός η συνάρτηση επιβίωσης παίρνει πάντα την ίδια τιμή στην έναρξη των μετρήσεων και αφετέρου συνήθως χρειάζεται ένα επαρκές χρονικό διάστημα (κατά μέσο όρο 3 με 6 μήνες) ώστε βαριές θεραπείες, όπως οι συσχετιζόμενες με τον καρκίνο να αποφέρουν αποτελέσματα. Ως εκ τούτου, σε μεγαλύτερους χρόνους και μετά τον πρώτο χρόνο είναι ξεκάθαρο ότι υπήρχαν μεγαλύτερες πιθανότητες επιβίωσης για όσες γυναίκες είχαν ακολουθήσει κάποια θεραπεία έναντι όσων δεν είχαν πάρει κανένα προληπτικό μέτρο. Συγκεκριμένα, το διάγραμμα δείχνει ότι πριν περάσουν 2500 μέρες, δηλαδή περίπου μέσα σε 6 χρόνια δεν επιβίωσε καμία γυναίκα χωρίς ορμονοθεραπεία. Από την άλλη πλευρά, στο ίδιο χρονικό διάστημα, πάνω από το 40% των γυναικών που ακολουθούσαν κάποια θεραπεία όχι μόνο έζησαν, αλλά φαίνεται σαν η κατάστασή τους να άρχιζε να σταθεροποιείται.

Τα αποτελέσματα αυτά επισφραγίζονται και από τους ελέγχους log – rank και Wilcoxon, στους οποίους η p-value είναι πολύ μικρή ($< 0,05$), οπότε απορρίπτεται η αρχική υπόθεση H_0 ότι οι κατανομές ταυτίζονται και δεχόμαστε ότι διαφέρουν σημαντικά μεταξύ τους.

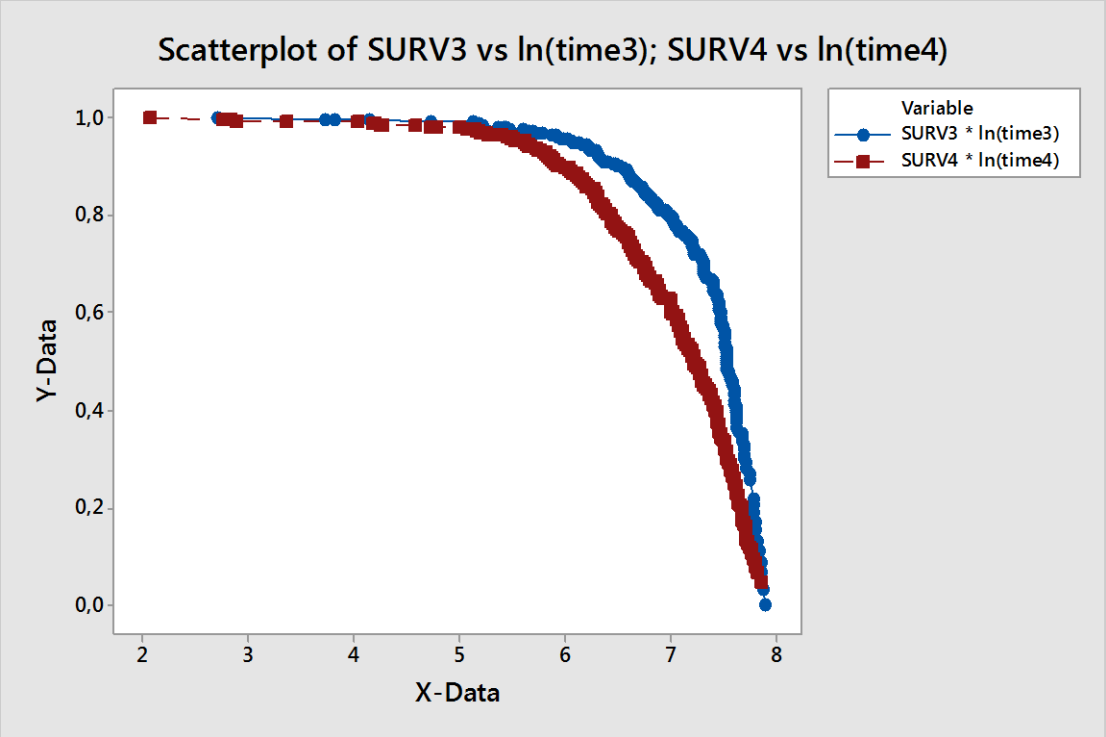
Σχετικά με τη συμμεταβλητή του βαθμού του όγκου, κάποιος μπορεί να παρατηρήσει από τη γραφική παράσταση της Kaplan – Meier στο γράφημα 5.2.2 ότι οι πιθανότητες επιβίωσης διαφοροποιούνται ανάλογα με το πόσο σοβαρός είναι ο όγκος, δηλαδή εκείνες που είχαν όγκο βαθμού 1, συνεπώς πιο καλοήθη νεοπλασία επιβίωσαν πολύ περισσότερο σε σχέση με όσες είχαν βαθμού 2 ή 3 που είναι πιο κακοήθεις όγκοι.

Τα αποτελέσματα αυτά επιβεβαιώνονται και από τους ελέγχους log –rank και Wilcoxon, στους οποίους η p – value είναι πάρα πολύ μικρή (< 0,001), οπότε οι κατανομές διαφέρουν πολύ μεταξύ τους.

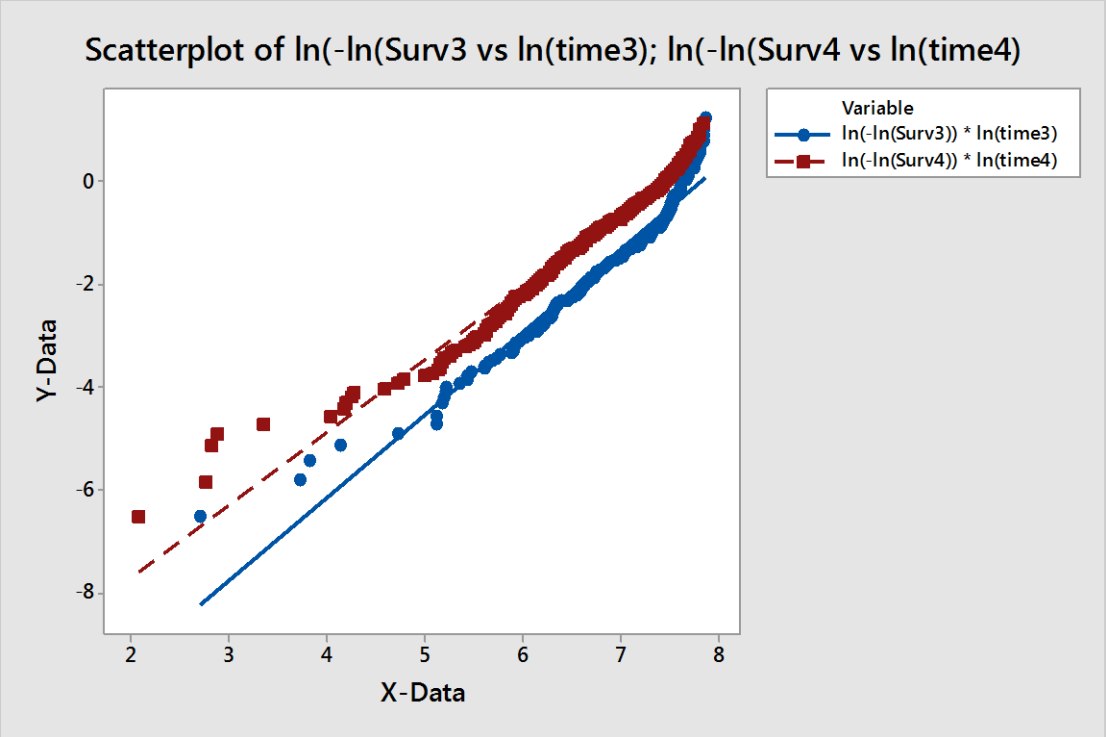
Τέλος, αναφορικά με τη συμμεταβλητή της εμμηνόπαυσης, δε φαίνεται να διαφέρουν και τόσο οι καμπύλες στη γραφική παράσταση της Kaplan – Meier στο γράφημα 5.2.3, με εξαίρεση προς το τέλος της έρευνας που δείχνει να επιβίωσαν περισσότερο όσες γυναίκες βρίσκονταν πριν την εμμηνόπαυση. Αναλόγως, η p – value στους ελέγχους log – rank και Wilcoxon είναι αρκετά μεγάλη, οπότε κάποιος δέχεται την αρχική υπόθεση H_0, δηλαδή ότι οι κατανομές ουσιαστικά ταυτίζονται.

5.2.3 Μοντέλα παλινδρόμησης με βάση μία μόνο συμμεταβλητή

Στη συνέχεια, με βάση τις συμμεταβλητές ‘ένδειξη ορμονοθεραπείας’ (hormone) και ‘κατάσταση εμμηνόπαυσης’ (meno) θα εξεταστεί γραφικά, ξεχωριστά για την κάθε μία, αν στα δεδομένα ταιριάζει ένα μοντέλο παλινδρόμησης της επιταχυνόμενης διακοπής (AF) ή ένα μοντέλο της αναλογικής διακινδύνευσης (PH) με ταυτόχρονο έλεγχο για την κατανομή Weibull. Η θεωρία βασίζεται στα Κεφάλαια 4.3.1 και 4.3.2. Από το Minitab εξάγονται οι γραφικές παραστάσεις στα γραφήματα 5.2.4, 5.2.5, 5.2.6 και 5.2.7. Για την ορμονοθεραπεία, στα γραφήματα 5.2.4 και 5.2.5 οι μπλε γραμμές αναφέρονται στο ‘όχι θεραπεία ορμονών’, ενώ οι κόκκινες γραμμές στο ‘ναι θεραπεία ορμονών’. Για την εμμηνόπαυση, στα γραφήματα 5.2.6 και 5.2.7 οι μπλε γραμμές αφορούν στην προ – εμμηνόπαυση περίοδο, ενώ οι κόκκινες στη μετά – εμμηνόπαυση περίοδο.

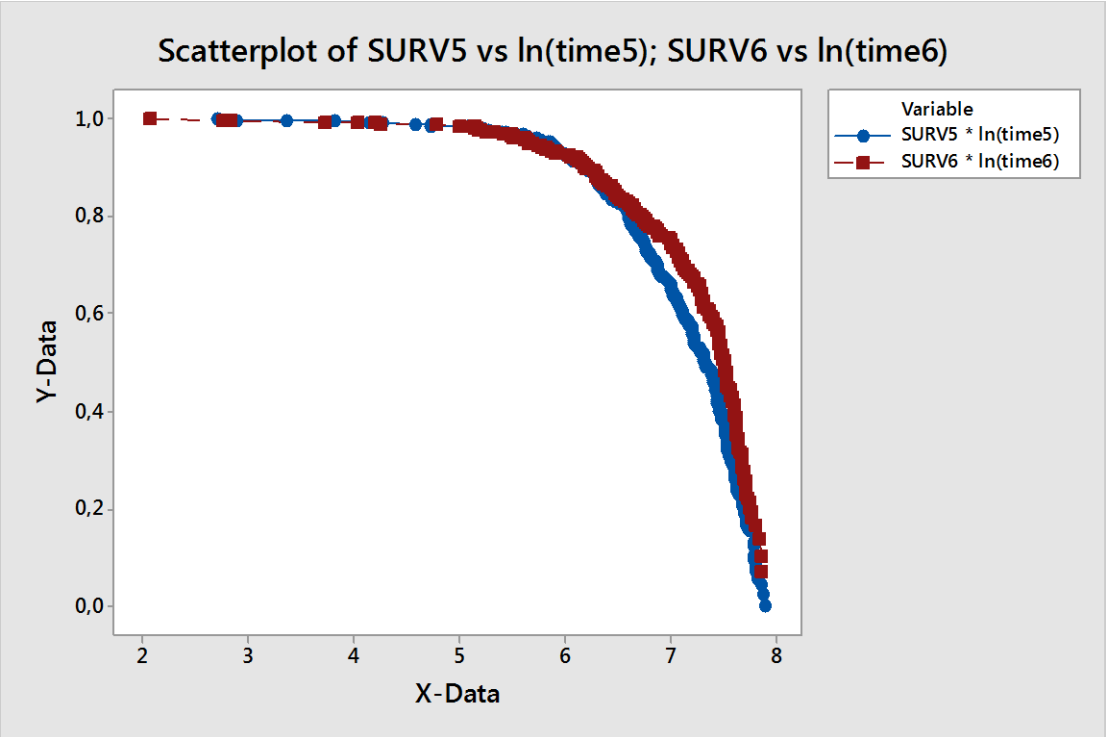


Γράφημα 5.2.4 Προσαρμοσμένη καμπύλη επιβίωσης της κατανομής Weibull για τη συμμεταβλητή της ορμόνης

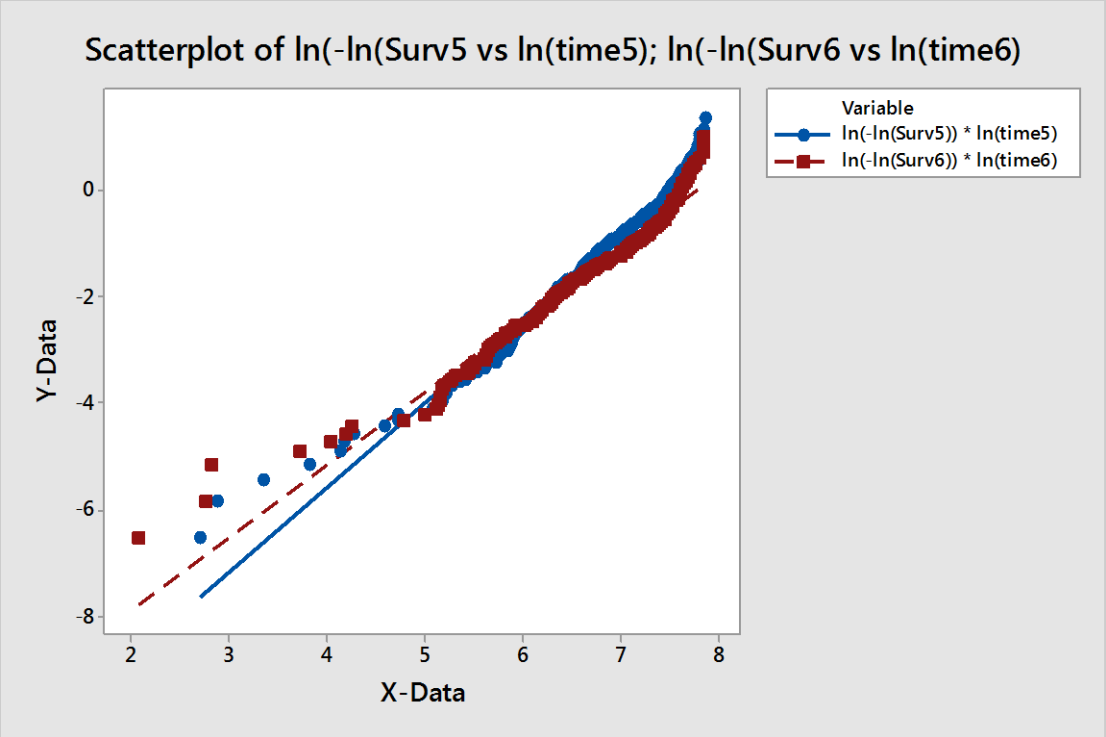


Γράφημα 5.2.5 Προσαρμογή της κατανομής Weibull με τη μέθοδο της μέγιστης πιθανοφάνειας για τη συμμεταβλητή της ορμόνης

Μοντέλο παλινδρόμησης της Επιταχυνόμενης Διακοπής (AF) για τη συμμεταβλητή της εμμηνόπαυσης



Γράφημα 5.2.6 Προσαρμοσμένη καμπύλη επιβίωσης της κατανομής Weibull για τη συμμεταβλητή της εμμηνόπαυσης



Γράφημα 5.2.7 Προσαρμογή της κατανομής Weibull με τη μέθοδο της μέγιστης πιθανοφάνειας για τη συμμεταβλητή της εμμηνόπαυσης

5.2.4 Συμπεράσματα

Στην παράγραφο 5.2.3 εξετάστηκε γραφικά αν στα δεδομένα, με βάση τις συμμεταβλητές της ορμονοθεραπείας και της κατάστασης της εμμηνόπαυσης ταιριάζει ένα μοντέλο παλινδρόμησης της επιταχυνόμενης διακοπής (AF) ή ένα μοντέλο της αναλογικής διακινδύνευσης (PH) με ταυτόχρονο έλεγχο για την κατανομή Weibull.

Τα συμπεράσματα που προέκυψαν παρατίθενται παρακάτω.

Αρχικά από το γράφημα 5.2.4, το οποίο δείχνει την προσαρμογή της επιταχυνόμενης διακοπής ως προς τη μεταβλητή της ορμονοθεραπείας ή εναλλακτικά την προσαρμοσμένη καμπύλη επιβίωσης της κατανομής Weibull για την ίδια συμμεταβλητή φαίνεται ότι υπάρχει οριζόντια μετατόπιση της μιας ευθείας ως προς την άλλη στους μεσαίους χρόνους για τη συνάρτηση επιβίωσης, ενώ στους αρχικούς και τελικούς χρόνους παρουσιάζεται σαν οι συναρτήσεις να είναι πολύ κοντά η μία στην άλλη, δηλαδή σα να ανάγονται σε μία. Αν θεωρητικά εξαιρεθούν αυτά τα

κομμάτια, τότε ευσταθεί η υπόθεση της επιταχυνόμενης διακοπής (AF), το συγκεκριμένο μοντέλο παλινδρόμησης φαίνεται να ταιριάζει στα δεδομένα και αυτά τα συμπεράσματα «καλής προσαρμογής» μπορεί να αποδειχτούν χρήσιμα στη συνέχεια της ανάλυσης για την προσαρμογή της κατανομής Weibull.

Σε ό,τι αφορά στο γράφημα 5.2.5, το οποίο αναφέρεται στην προσαρμογή της αναλογικής διακινδύνευσης ως προς τη μεταβλητή της ορμονοθεραπείας παρατηρείται ότι οι καμπύλες σχηματίζουν παράλληλες ευθείες στο μεγαλύτερο τμήμα, κάτι που δείχνει μιαν αρκετά καλή προσαρμογή του μοντέλου της αναλογικής διακινδύνευσης (PH) κι ειδικότερα της ειδικής περίπτωσης που είναι η Weibull, μιας κι οι καμπύλες είναι ευθύγραμμες.

Επιπλέον, στις γραφικές παραστάσεις στα γραφήματα 5.2.6 και 5.2.7 για την προσαρμογή της επιταχυνόμενης διακοπής ή της αναλογικής διακινδύνευσης αντιστοίχως ως προς τη μεταβλητή της εμμηνόπαυσης, κάποιος παρατηρεί ότι οι καμπύλες ουσιαστικά ταυτίζονται στο μεγαλύτερο μέρος τους, άρα στα υπάρχοντα δεδομένα, σε σχέση με τη συμμεταβλητή της εμμηνόπαυσης, δε φαίνεται να ταιριάζει ιδιαίτερα κανένα από τα παραπάνω μοντέλα.

5.3 Μέρος Β

Στην παράγραφο 5.2.3 εξετάστηκε γραφικά αν ταιριάζει στα δεδομένα ένα μοντέλο παλινδρόμησης επιταχυνόμενης διακοπής ξεχωριστά ως προς τις μεταβλητές της ορμονοθεραπείας και της εμμηνόπαυσης. Σε αυτό το κεφάλαιο, θα προσαρμοστεί ένα μοντέλο παλινδρόμησης επιταχυνόμενης διακοπής της κατανομής Weibull αξιοποιώντας όλες τις συμμεταβλητές που υπάρχουν και εφαρμόζοντας τους ελέγχους Wald, του λόγου των πιθανοφανειών, το κριτήριο AIC και τη μέθοδο της Backward Elimination, όπως ορίστηκαν στο κεφάλαιο 3 για την εύρεση του βέλτιστου μοντέλου ως προς το είδος και το πλήθος των συμμεταβλητών. Επιπλέον, θα κατασκευαστούν 0,95 – διαστήματα εμπιστοσύνης για τους συντελεστές των συμμεταβλητών του τελικού μοντέλου που θα έχει βρεθεί. Εν τέλει, θα κατασκευαστούν οι γραφικές παραστάσεις των υπολοίπων Cox-Snell και Standardized για την εξαγωγή ακόμη περισσότερων συμπερασμάτων, όπως ορίστηκαν στο κεφάλαιο 4. Σκοπός όλων αυτών είναι να εξεταστεί πως η επιβίωση των γυναικών με καρκίνο στο μαστό εξαρτιόταν από διάφορους παράγοντες και ειδικότερα στην παρούσα εργασία από τις

συμμεταβλητές που έχουν καταγραφεί στον πίνακα 3. Η ανάλυση θα βασιστεί στη σύγκριση του μοντέλου που περιέχει όλες τις συμμεταβλητές με εκείνο που δεν περιέχει καμία. Όσες μεταβλητές αποδειχτούν στατιστικά μη σημαντικές θα αφαιρεθούν και σε επόμενη εφαρμογή θα πραγματοποιηθεί παρόμοια σύγκριση μεταξύ του αρχικού μοντέλου με όλες τις συμμεταβλητές κι εκείνου με όσες μεταβλητές θα έχουν μείνει μετά από την αφαίρεση όσων συμμεταβλητών στατιστικά δε συνεισέφεραν κάτι. Η διαδικασία θα ολοκληρωθεί όταν παραμείνουν στο μοντέλο μόνο στατιστικά σημαντικές μεταβλητές (βέλτιστο μοντέλο), οι οποίες θα είναι κι εκείνες που κυρίως συσχετίζονταν με την επιβίωση των ασθενών.

5.3.1 Εξέταση δεδομένων με βάση όλες τις συμμεταβλητές

Εισάγοντας τα δεδομένα στην R προκύπτουν συνοπτικά τα αποτελέσματα στον πίνακα 8, ο οποίος, μέσω της προσαρμογής του μοντέλου παλινδρόμησης επιταχυνόμενης διακοπής της κατανομής Weibull δείχνει τα αποτελέσματα από τη σύγκριση του μοντέλου που περιέχει όλες τις 8 συμμεταβλητές (age, tsize, pnodes, progrec, estrec, hormonef1, menostatf1, tgradef2 – tgradef3) με αυτό που δεν περιέχει καμία, δηλαδή τις τιμές των συντελεστών των συμμεταβλητών, το τυπικό σφάλμα, τον έλεγχο Wald (μέσω της μεταβλητής z) και τη τιμή στον έλεγχο του λόγου των πιθανοφανειών. Στη συνέχεια, στον πίνακα 9 φαίνονται τα αποτελέσματα της R σχετικά με το κριτήριο AIC μέσω της μεθόδου της Backward Elimination για την εύρεση του βέλτιστου μοντέλου, καθώς και η τιμή του λόγου των πιθανοφανειών (μέσω της μεταβλητής LRT). Εν τέλει, στον πίνακα 10 φαίνονται συνοπτικά τα επακόλουθα από την προσαρμογή του μοντέλου παλινδρόμησης επιταχυνόμενης διακοπής της κατανομής Weibull στο τελικό μοντέλο, δηλαδή οι τιμές των συντελεστών των μεταβλητών, το τυπικό σφάλμα και ο έλεγχος Wald (μέσω της z). Πλήρης καταγραφή όλων των στοιχείων που έκδωσε η R παρατίθενται στο παράρτημα Π.2.

Διευκρινίζεται ότι οι μεταβλητές της ορμόνης (hormonef1) και της εμμηνόπαυσης (menostatf1) εισήχθησαν στην R ως κατηγορικές – δυαδικές λαμβάνοντας υπόψη την περίπτωση όπου υπάρχει θετική ένδειξη για ορμονοθεραπεία, καθώς και την περίπτωση όπου η γυναίκα έχει βιώσει την περίοδο της εμμηνόπαυσης. Επίσης θεωρήθηκε ως κατηγορική κι η συμμεταβλητή που αναφέρεται στο βαθμό του όγκου,

όπου η R την διαίρεσε στις δύο μόνο υποκατηγορίες, σε αυτή με βαθμό όγκου 2 (tgradef2) και σε εκείνη με βαθμό όγκου 3 (tgradef3).

	Value of coefficients	Std. Error	**z**	**p**
(Intercept)	8.015220	0.350300	22.88	< 2e-16
age	0.006819	0.006644	1.03	0.3048
tsize	-0.005765	0.002817	-2.05	0.0407
pnodes	-0.037979	0.005383	-7.06	1.7e-12
progrec	0.001643	0.000417	3.94	8.2e-05
estrec	-0.000179	0.000324	-0.55	0.5807
hormonef1	0.268358	0.092826	2.89	0.0038
menostatf1	-0.194856	0.131185	-1.49	0.1375
tgradef2	-0.471977	0.180262	-2.62	0.0088
tgradef3	-0.582663	0.193759	-3.01	0.0026

Loglik(model) = -2579.7 **Loglik(intercept only)** = -2637.3

Chisq = 115.16 on 9 degrees of freedom, **p** = 1.3e-20

Πίνακας 8 Αποτελέσματα από τη σύγκριση του μοντέλου που περιέχει όλες τις συμμεταβλητές με εκείνο που δεν περιέχει καμία αξιοποιώντας το μοντέλο παλινδρόμησης επιταχυνόμενης διακοπής της κατανομής Weibull

Start: AIC=5181.39

Step: AIC=5179.68

Step: AIC=5178.58

Step: AIC=5177.99

Scale = 0.7201373

	Df	AIC	LRT	Pr(>Chi)
<none>		5178.0		
- tsize	1	5179.6	3.628	0.056821 .
- hormonef	1	5183.9	7.955	0.004797
- tgradef	2	5184.9	10.888	0.004323
- pnodes	1	5212.9	36.922	1.229e-09
- progrec	1	5199.2	23.252	1.421e-06

Πίνακας 9 Αποτελέσματα ως προς το κριτήριο AIC μέσω της μεθόδου της Backward Elimination για την εύρεση του βέλτιστου μοντέλου στο μοντέλο επιταχυνόμενης διακοπής της κατανομής Weibull

	Value of coefficients	Std. Error	z	p
(Intercept)	8.249659	0.198482	41.56	< 2e-16
tsize	-0.005434	0.002784	-1.95	0.0510
hormonef1	0.250590	0.090707	2.76	0.0057
tgradef2	-0.477478	0.180369	-2.65	0.0081
tgradef3	-0.587401	0.193933	-3.03	0.0025
pnodes	-0.038199	0.005391	-7.09	1.4e-12
progrec	0.001637	0.000404	4.05	5.1e-05

Loglik(model) = -2581 **Loglik(intercept only)** = -2637.3

Chisq = 112.56 on 6 degrees of freedom, **p** = < 2e-16

Πίνακας 10 Προσαρμογή του μοντέλου παλινδρόμησης επιταχυνόμενης διακοπής της κατανομής Weibull στο βέλτιστο μοντέλο.

5.3.2 Συμπεράσματα

Στην παράγραφο 5.3.1 υλοποιήθηκε η προσαρμογή του μοντέλου παλινδρόμησης επιταχυνόμενης διακοπής της κατανομής Weibull μέσω του ελέγχου Wald, του ελέγχου του λόγου των πιθανοφανειών και της μεθόδου της Backward Elimination, η οποία βασίστηκε στην εύρεση του μοντέλου με το μικρότερο AIC, ώστε να επιτευχθεί το βέλτιστο μοντέλο συμμεταβλητών για τη διεξαγωγή της έρευνας. Ως επακόλουθο προκύπτουν κάποια συμπεράσματα.

Παρατηρώντας κάποιος στον πίνακα 8 τον έλεγχο Wald, ο οποίος στην R μεταφέρεται μέσω της μεταβλητής z και των αντίστοιχων p – values διακρίνει εμφανώς ότι υπάρχουν μεταβλητές με p – value > 0,05. Συγκεκριμένα, οι συμμεταβλητές της ηλικίας (age), της εμμηνόπαυσης (menostatf1) και της κατάστασης του υποδοχέα οιστρογόνου (estrec) είναι στατιστικά μη σημαντικές, αφού οδηγούν σε αποδοχή της μηδενικής υπόθεσης, συνθήκη που προτιμάται να αποφεύγεται. Από την άλλη πλευρά, ο έλεγχος του λόγου των πιθανοφανειών ανάμεσα στο μοντέλο με όλες τις συμμεταβλητές και του αντίστοιχου που έχει μόνο το σταθερό και καμία άλλη μεταβλητή (στην R μεταφέρονται μέσω των εντολών Loglik(model) και Loglik(intercept only) ή της μεταβλητής LRT) δίνει p – value < 0,05, το οποίο οδηγεί σε απόρριψη της μηδενικής υπόθεσης. Επομένως πρόκειται για ένα μοντέλο, το οποίο προβληματίζει κατά πόσο είναι στατιστικά σημαντικό, αφού υπάρχει αντίφαση ανάμεσα στους δύο ελέγχους κι είναι επιθυμητό να εφαρμοστεί η μέθοδος της Backward Elimination, ώστε μέσω του κριτηρίου AIC να βρεθεί πιο ξεκάθαρα το βέλτιστο μοντέλο για τα υπάρχοντα δεδομένα.

Παρατηρώντας τα αποτελέσματα στον πίνακα 9 φαίνεται ότι το κριτήριο AIC του βέλτιστου μοντέλου με τις λιγότερες δυνατές μεταβλητές, καθώς κι ο έλεγχος του λόγου των πιθανοφανειών δε διαφέρουν και τόσο από τα αντίστοιχα του αρχικού μοντέλου με όλες τις συμμεταβλητές, σημάδι που μαρτυράει ότι τελικά στο αρχικό μοντέλο χρειάζονταν μικρές βελτιώσεις. Συνυπολογίζοντας και τον πίνακα 10, κάποιος καταλήγει στο ότι το βέλτιστο μοντέλο είναι εκείνο με τις συμμεταβλητές του μεγέθους του όγκου (tsize), της ένδειξης ορμονοθεραπείας (hormonef), του βαθμού του όγκου (tgradef), του αριθμού των θετικών λεμφαδένων (pnodes) και της κατάστασης του υποδοχέα προγεστερόνης (progrec). Πράγματι, από τους ελέγχους Wald και του λόγου των πιθανοφανειών φαίνεται ότι όλες σχεδόν οι

προαναφερθείσες συμμεταβλητές έχουν p – value < 0,05, δηλαδή είναι στατιστικά σημαντικές για την επιβίωση των γυναικών με καρκίνο στο μαστό. Παρατηρείται βέβαια ότι η μεταβλητή του μεγέθους του όγκου (tsize) έχει οριακή τιμή κοντά στην επιθυμητή p – value, αλλά γίνεται αποδεκτή για το τελικό μοντέλο.

5.3.3 Διαστήματα εμπιστοσύνης και ερμηνείες

Μέσω της R προκύπτει ο πίνακας 11 για τα διαστήματα εμπιστοσύνης 95% ως προς τους συντελεστές των συμμεταβλητών του αρχικού μοντέλου με όλες τις συμμεταβλητές. Εύκολα κάποιος μπορεί να διαπιστώσει ότι κάποιοι από τους συντελεστές (age, estrec, menostatf1) είναι στατιστικά μη σημαντικοί, αφού τα αντίστοιχα διαστήματα εμπιστοσύνης τους εμπεριέχουν το μηδέν. Παρατηρείται μάλιστα πως αυτές είναι οι ίδιες συμμεταβλητές που αφαιρέθηκαν πρώτες στην αναζήτηση του βέλτιστου μοντέλου, σύμφωνα με τα συμπεράσματα στην παράγραφο 5.3.2. Αυτός είναι κι ο λόγος, για τον οποίο κάποιος θα ενδιαφερόταν να αναλύσει και να ερμηνεύσει τα διαστήματα εμπιστοσύνης 95% ως προς τους συντελεστές των συμμεταβλητών του τελικού μόνο μοντέλου που βρέθηκε στην παράγραφο 5.3.2. Στον πίνακα 12 φαίνονται τα αποτελέσματα για τους συντελεστές των συμμεταβλητών του βέλτιστου μοντέλου μέσω της R. Παρατηρείται ότι η μεταβλητή του μεγέθους του όγκου (tsize) είναι στατιστικά μη σημαντική, αφού το μηδέν ανήκει στο διάστημα εμπιστοσύνης της, αλλά όπως σχολιάστηκε και στην παράγραφο 5.3.2 εγκρίνεται οριακά ως σημαντική μεταβλητή για το βέλτιστο μοντέλο.

	2.5 %	97.5 %
(Intercept)	7.3286455944	8.7017952019
age	-0.0062036463	0.0198411624
tsize	-0.0112854860	-0.0002444975
pnodes	-0.0485291169	-0.0274296347
progrec	0.0008257403	0.0024611795
estrec	-0.0008127757	0.0004553901
hormonef1	0.0864235231	0.4502930456
menostatf1	-0.4519745657	0.0622626173
tgradef2	-0.8252848078	-0.1186689431
tgradef3	-0.9624239505	-0.2029029455

Πίνακας 11 Διαστήματα εμπιστοσύνης 95% για τους συντελεστές των συμμεταβλητών του αρχικού μοντέλου με όλες τις συμμεταβλητές

	2.5 %	97.5 %
(Intercept)	7.8606411962	8.638677e+00
tsize	-0.0108914698	2.284198e-05
hormonef1	0.0728067278	4.283733e-01
tgradef2	-0.8309941487	-1.239626e-01
tgradef3	-0.9675022285	-2.073004e-01
pnodes	-0.0487651085	-2.763245e-02
progrec	0.0008449267	2.428241e-03

Πίνακας 12 Διαστήματα εμπιστοσύνης 95% για τους συντελεστές των συμμεταβλητών του τελικού μοντέλου

Ερμηνείες. Από τα διαστήματα εμπιστοσύνης του πίνακα 12 προκύπτουν κάποιες πολύ χρήσιμες ερμηνείες για τους συντελεστές των συμμεταβλητών του βέλτιστου μοντέλου.

Σύμφωνα με τη θεωρία στην παράγραφο 4.4.1 αναφορικά με τις κατανομές Weibull και Gumbel, καθώς και από τα αποτελέσματα στην R, όπου $\sigma = scale = 0.7201373$ (πίνακας 9), θα ισχύει

$$\sigma = \frac{1}{\eta} \Leftrightarrow \eta = \frac{1}{\sigma} \Leftrightarrow \eta = \frac{1}{0.7201373} \Leftrightarrow \eta \simeq 1,39 .$$

Συμπερασματικά, για την ερμηνεία των συντελεστών $\hat{\beta}_i, i = 1,2,3,4,5,6$ (πίνακας 10) των συμμεταβλητών του τελικού μοντέλου, όπου

$$\begin{aligned} i &= 1 \rightarrow \text{tsize} \\ i &= 2 \rightarrow \text{hormonef1} \\ i &= 3 \rightarrow \text{tgradef2} \\ i &= 4 \rightarrow \text{tgradef3} \\ i &= 5 \rightarrow \text{pnodes} \\ i &= 6 \rightarrow \text{progrec} \end{aligned} \quad ,$$

ισχύουν τα κάτωθι παίρνοντας προσεγγιστικά τις αναγραφόμενες τιμές.

- Αν αυξανόταν η συμμεταβλητή tsize, δηλαδή το μέγεθος του όγκου κατά μία μονάδα, τότε

  $$\exp\left\{-\hat{\beta}_1 \cdot \hat{\eta}\right\} = \exp\left\{-(-0,005)\cdot 1,39\right\} = \exp\left\{0,007\right\} = 1,007 > 1 .$$

  Αυτό σημαίνει ότι η πιθανότητα επιβίωσης υψωμένη σε αυτή την ποσότητα θα μειωνόταν λίγο, δηλαδή το γεγονός (υποτροπή ή θάνατος) δε θα αργούσε και τόσο να συμβεί, κάτι το οποίο είναι λογικό αφού θα είχε χειροτερέψει η κατάσταση του καρκίνου.

- Αν αυξανόταν η συμμεταβλητή hormonef1, δηλαδή η ένδειξη της ορμονοθεραπείας κατά μία μονάδα, τότε

  $$\exp\left\{-\hat{\beta}_2 \cdot \hat{\eta}\right\} = \exp\left\{-0,251\cdot 1,39\right\} = \exp\left\{-0,349\right\} = 0,705 < 1 .$$

  Αυτό σημαίνει ότι η πιθανότητα επιβίωσης υψωμένη σε αυτή την ποσότητα θα αυξανόταν, δηλαδή το γεγονός (υποτροπή ή θάνατος) θα αργούσε να συμβεί, το οποίο είναι λογικό αφού αύξηση από 0 σε 1 σημαίνει ότι κάποια ασθενής άρχισε θεραπεία με ορμόνες, οι οποίες καθυστερούν την εξέλιξη της νόσου.

- Αν αυξανόταν η συμμεταβλητή tgradef2, δηλαδή ο βαθμός του όγκου από 2 σε 3, τότε

$$\exp\left\{-\hat{\beta}_3 \cdot \hat{\eta}\right\} = \exp\left\{-\left(-0,447\right)\cdot 1,39\right\} = \exp\left\{0,621\right\} = 1,861 > 1.$$

Αυτό σημαίνει ότι η πιθανότητα επιβίωσης υψωμένη σε αυτή την ποσότητα θα μειωνόταν, δηλαδή το γεγονός (υποτροπή ή θάνατος) θα συνέβαινε ταχύτερα, αφού ο όγκος θα γινόταν ακόμη πιο κακοήθης από ό,τι ήταν.

- Αν αυξανόταν η συμμεταβλητή tgradef3 κατά μία μονάδα, δηλαδή ο βαθμός του όγκου έφτανε την πιο βαριά κατάσταση κακοήθειας, τότε

$$\exp\left\{-\hat{\beta}_4 \cdot \hat{\eta}\right\} = \exp\left\{-\left(-0,587\right)\cdot 1,39\right\} = \exp\left\{0,816\right\} = 2,261 > 1.$$

Αυτό σημαίνει ότι η πιθανότητα επιβίωσης υψωμένη σε αυτή την ποσότητα θα μειωνόταν αισθητά και το γεγονός θα συνέβαινε πολύ γρηγορότερα, αφού η ασθενής θα βρισκόταν στο χειρότερο στάδιο του καρκίνου.

- Αν αυξανόταν η συμμεταβλητή pnodes, δηλαδή ο αριθμός των θετικών λεμφαδένων κατά μία μονάδα, τότε

$$\exp\left\{-\hat{\beta}_5 \cdot \hat{\eta}\right\} = \exp\left\{-\left(-0,038\right)\cdot 1,39\right\} = \exp\left\{0,053\right\} = 1,054 > 1.$$

Αυτό σημαίνει ότι η πιθανότητα επιβίωσης υψωμένη σε αυτή την ποσότητα θα μειωνόταν, δηλαδή το γεγονός δε θα αργούσε να συμβεί. Γενικώς, η ύπαρξη θετικών λεμφαδένων γύρω από την περιοχή της μασχάλης είναι θετική ένδειξη για επιδείνωση της νόσου.

- Αν αυξανόταν η συμμεταβλητή progrec, δηλαδή η κατάσταση του υποδοχέα προγεστερόνης κατά μία μονάδα, τότε

$$\exp\left\{-\hat{\beta}_6 \cdot \hat{\eta}\right\} = \exp\left\{-0,002\cdot 1,39\right\} = \exp\left\{-0,003\right\} = 0,997 < 1.$$

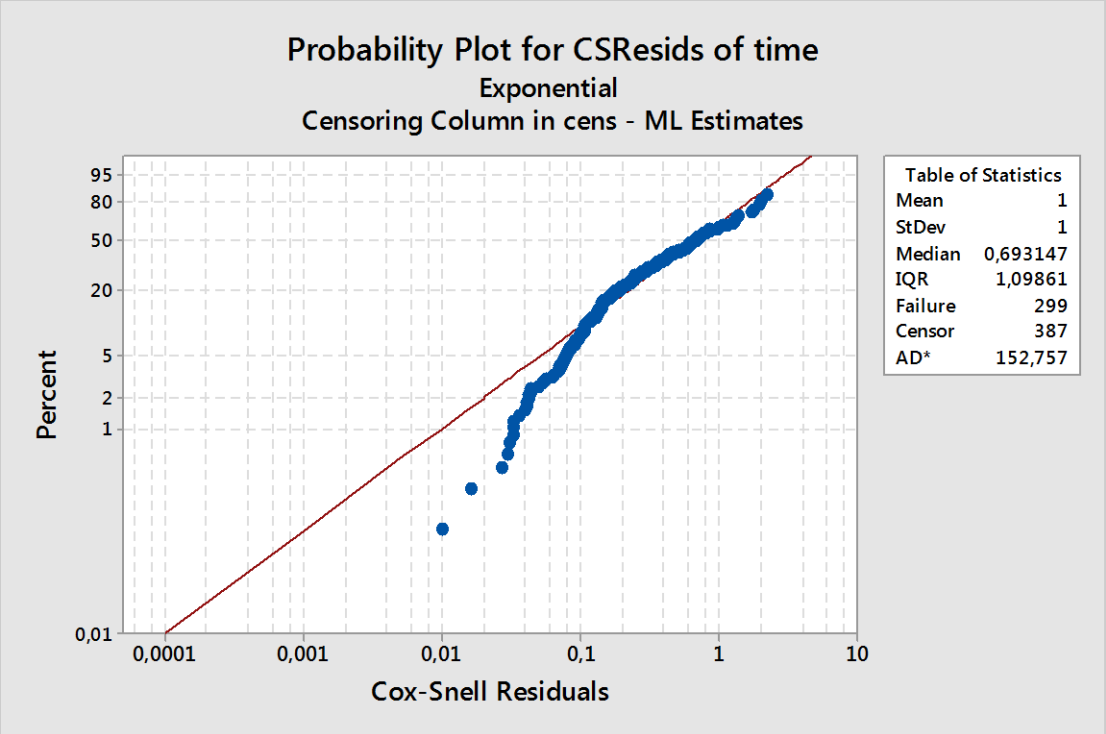
Αυτό σημαίνει ότι η πιθανότητα επιβίωσης υψωμένη σε αυτή την ποσότητα θα αυξανόταν λίγο και το γεγονός θα αργούσε να συμβεί, αν κι η επίδραση φαίνεται να είναι μικρή, μιας κι η παραπάνω ποσότητα προσεγγίζει τη μονάδα σε πολύ μεγάλο βαθμό.

5.3.4 Εφαρμογή υπολοίπων και συμπεράσματα

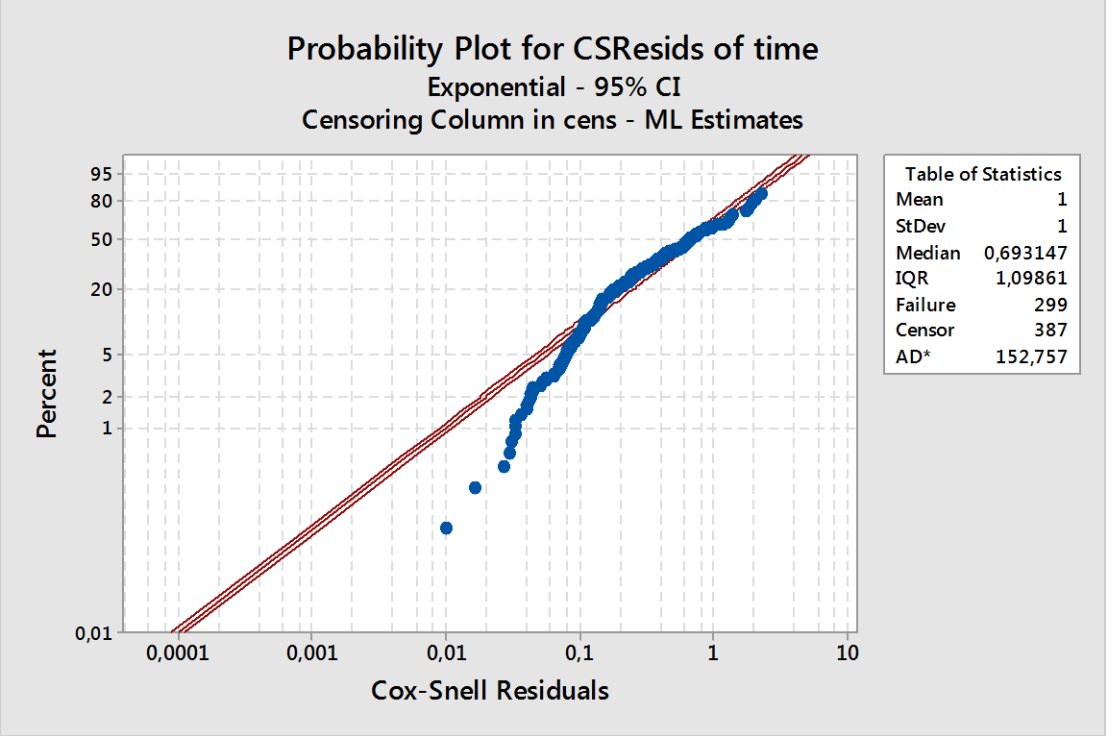
Για τον έλεγχο της καταλληλότητας του βέλτιστου μοντέλου μέσω των υπολοίπων Cox – Snell και Standardized, τα δεδομένα εισάγονται στο Minitab και εφαρμόζεται η μέθοδος μέγιστης πιθανοφάνειας ως προς την κατανομή Weibull. Τα αποτελέσματα φαίνονται στον πίνακα 13. Στη συνέχεια, το Minitab εξάγει τις αντίστοιχες γραφικές παραστάσεις. Στο γράφημα 5.3.1 φαίνεται το διάγραμμα των υπολοίπων Cox – Snell, ενώ στο γράφημα 5.3.2 φαίνεται το διάγραμμα των υπολοίπων Cox – Snell μέσα στο διάστημα εμπιστοσύνης 95% που εξετάστηκε στην παράγραφο 5.3.3. Παρομοίως, τα γραφήματα 5.3.3 και 5.3.4 δείχνουν τις γραφικές παραστάσεις των Standardized υπολοίπων και Standardized υπολοίπων σε διάστημα εμπιστοσύνης 95% αντιστοίχως.

Standardized Residuals = 152,757
Cox – Snell Residuals = 152,757

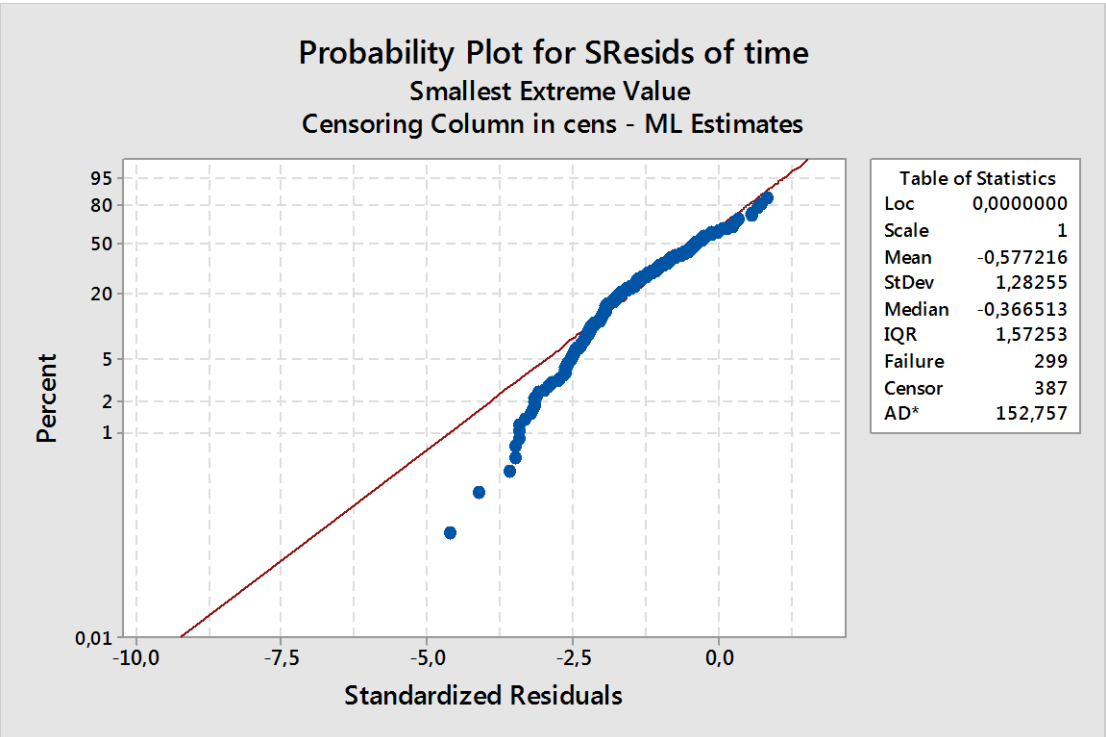
Πίνακας 13 Υπόλοιπα Cox – Snell και Standardized



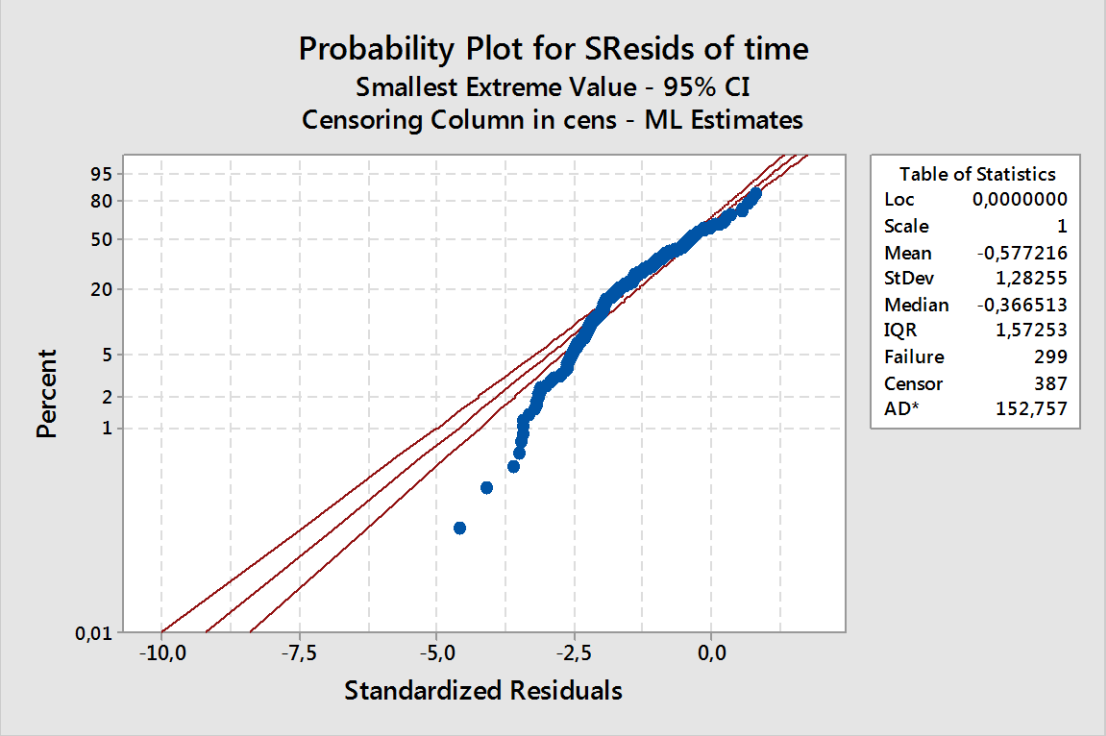
Γράφημα 5.3.1 Διάγραμμα υπολοίπων Cox – Snell



Γράφημα 5.3.2 Διάγραμμα υπολοίπων Cox – Snell σε διάστημα εμπιστοσύνης 95%



Γράφημα 5.3.3 Διάγραμμα υπολοίπων Standardized



Γράφημα 5.3.4 Διάγραμμα υπολοίπων Standardized σε διάστημα εμπιστοσύνης 95%

Από τα γραφήματα 5.3.1 και 5.3.3 παρατηρείται ότι οι παρατηρήσεις δεν έχουν τόσο καλή προσαρμογή στην ευθεία, αλλά υποστηρίζουν μια μέτρια προς καλή κατάσταση. Ομοίως, στα γραφήματα 5.3.2 και 5.3.4 υπάρχουν πολλές παρατηρήσεις που ξεφεύγουν από τα όρια των διαστημάτων εμπιστοσύνης. Συνεπώς προκύπτει ότι σύμφωνα με τον έλεγχο μέσω υπολοίπων, το μοντέλο δε φαίνεται τόσο κατάλληλο για τα υπάρχοντα δεδομένα. Αυτό βέβαια μπορεί να επηρεάζεται από το γεγονός ότι η συμμεταβλητή tsize (μέγεθος όγκου) δεν ταιριάζει και τόσο στο βέλτιστο μοντέλο.

5.4 Μέρος Γ

Στο κεφάλαιο 5.3 προσαρμόστηκε το μοντέλο παλινδρόμησης επιταχυνόμενης διακοπής της κατανομής Weibull αξιοποιώντας όλες τις συμμεταβλητές που υπάρχουν και εφαρμόζοντας τους ελέγχους Wald, του λόγου των πιθανοφανειών, το κριτήριο AIC και τη μέθοδο της Backward Elimination. Επιπλέον κατασκευάστηκαν 0,95 – διαστήματα εμπιστοσύνης για τους συντελεστές των συμμεταβλητών του βέλτιστου μοντέλου που βρέθηκε στην παράγραφο 5.3.2. Τελικώς κατασκευάστηκαν οι γραφικές παραστάσεις των υπολοίπων Cox – Snell και Standardized.

Σε αυτό το κεφάλαιο θα προσαρμοστεί το μοντέλο αναλογικής διακινδύνευσης του Cox και θα εφαρμοστούν περίπου οι ίδιες μέθοδοι όπως στο προηγούμενο κεφάλαιο. Εκτός αυτών, θα εξεταστεί αν το τελικό μοντέλο πληροί τις προϋποθέσεις της αναλογικότητας της διακινδύνευσης μέσω στατιστικών ελέγχων, καθώς και με τη χρήση των υπολοίπων Schoenfeld. Στο τελευταίο βήμα, θα δημιουργηθούν οι γραφικές παραστάσεις των υπολοίπων DFBETAS, Martingale και Deviance για ακόμη μεγαλύτερη αξιοπιστία. Σκοπός είναι πάλι να εξεταστεί η συσχέτιση της επιβίωσης των γυναικών που είχαν καρκίνο στο μαστό με τις διάφορες συμμεταβλητές που έχουν καταγραφεί στον πίνακα 3.

5.4.1 Προσαρμογή του μοντέλου του Cox

Εισάγοντας τα δεδομένα στην R προκύπτουν συνοπτικά τα αποτελέσματα στον πίνακα 14, ο οποίος, μέσω της προσαρμογής του μοντέλου αναλογικής διακινδύνευσης του Cox δείχνει τα αποτελέσματα από τη σύγκριση του μοντέλου που περιέχει όλες τις 8 συμμεταβλητές (age, tsize, pnodes, progrec, estrec, hormonef1, menostatf1, tgradef2 – tgradef3) με αυτό που δεν περιέχει καμία. Συγκεκριμένα, κάποιος μπορεί να παρατηρήσει τις τιμές των συντελεστών των συμμεταβλητών, καθώς και την εκθετική συνάρτηση υψωμένη σε αυτούς τους συντελεστές, κάτι το οποίο προτιμάται στο μοντέλο του Cox, το τυπικό σφάλμα, τον έλεγχο Wald (μέσω της z) και την τιμή στον έλεγχο του λόγου των πιθανοφανειών για την κάθε μεταβλητή ξεχωριστά. Επιπροσθέτως, κάποιος μπορεί να αντιληφθεί το concordance, το οποίο δείχνει την προβλεπτική ικανότητα του μοντέλου (για παράδειγμα ποιος θα ζήσει περισσότερο), δηλαδή εξετάζει αν οι προβλέψεις που γίνονται από έναν ερευνητή βρίσκονται σε συμφωνία με τις μεταβλητές εστιάζοντας στη διάταξη κι όχι στην τιμή της διάρκειας ζωής. Παίρνει τιμές από μηδέν που δηλώνει καμία συμφωνία μέχρι τη μονάδα που δηλώνει απόλυτη συμφωνία. Εκτός αυτών, κάποιος μπορεί επίσης να διαπιστώσει τα αποτελέσματα από τον έλεγχο του λόγου των πιθανοφανειών, του ελέγχου Wald και του ελέγχου log – rank για όλο το μοντέλο συνολικά.

Στη συνέχεια, στον πίνακα 15 φαίνονται τα αποτελέσματα της R σχετικά με το κριτήριο AIC μέσω της μεθόδου της Backward Elimination για την εύρεση του βέλτιστου μοντέλου στο μοντέλο του Cox, καθώς και η τιμή του λόγου των πιθανοφανειών (μέσω της LRT). Εν τέλει, στον πίνακα 16 φαίνονται συνοπτικά τα

αποτελέσματα από την προσαρμογή του μοντέλου αναλογικής διακινδύνευσης του Cox στο τελικό μοντέλο, δηλαδή οι τιμές των συντελεστών των μεταβλητών, η εκθετική συνάρτηση υψωμένη σε αυτούς τους συντελεστές στη θετική, αλλά και στην αρνητική εκδοχή τους, το τυπικό σφάλμα και ο έλεγχος Wald (μέσω της z) για την κάθε μεταβλητή ξεχωριστά. Επιπλέον, κάποιος μπορεί να εντοπίσει το concordance, καθώς και τα αποτελέσματα από τον έλεγχο του λόγου των πιθανοφανειών, του ελέγχου Wald και του ελέγχου log – rank για όλο το μοντέλο συνολικά. Πλήρης καταγραφή όλων των στοιχείων που έκδωσε η R παρατίθενται στο παράρτημα Π.3.

	coef	exp(coef)	se(coef)	z	p
age	-0.0094592	0.9905854	0.0093006	-1.017	0.309126
tsize	0.0077961	1.0078266	0.0039390	1.979	0.047794
pnodes	0.0487886	1.0499984	0.0074471	6.551	5.7e-11
progrec	-0.0022172	0.9977852	0.0005735	-3.866	0.000111
estrec	0.0001973	1.0001973	0.0004504	0.438	0.661307
hormonef1	-0.3462784	0.7073155	0.1290747	-2.683	0.007301
menostatf1	0.2584448	1.2949147	0.1834765	1.409	0.158954
tgradef2	0.6361117	1.8891211	0.2492025	2.553	0.010693
tgradef3	0.7796542	2.1807181	0.2684801	2.904	0.003685

Likelihood ratio test = 104.8 on 9 df, **p** =< 2.2e-16

	exp(coef)	exp(-coef)	lower .95	upper .95
age	0.9906	1.0095	0.9727	1.0088
tsize	1.0078	0.9922	1.0001	1.0156
pnodes	1.0500	0.9524	1.0348	1.0654
progrec	0.9978	1.0022	0.9967	0.9989
estrec	1.0002	0.9998	0.9993	1.0011
hormonef1	0.7073	1.4138	0.5492	0.9109
menostatf1	1.2949	0.7723	0.9038	1.8553
tgradef2	1.8891	0.5293	1.1591	3.0788
tgradef3	2.1807	0.4586	1.2885	3.6909

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

Πίνακας 1410 Αποτελέσματα από τη σύγκριση του μοντέλου που περιέχει όλες τις συμμεταβλητές με αυτό που δεν περιέχει καμία αξιοποιώντας το μοντέλο αναλογικής διακινδύνευσης του Cox

Start: AIC=3489.46

Step: AIC=3487.65

Step: AIC=3486.56

Step: AIC=3485.68

	Df	AIC	LRT	Pr(>Chi)
<none>		3485.7		
- tsize	1	3487.0	3.364	0.066621 .
- hormonef	1	3490.5	6.847	0.008878
- tgradef	2	3491.8	10.121	0.006343
- progrec	1	3506.1	22.446	2.162e-06
- pnodes	1	3515.9	32.185	1.402e-08

Πίνακας 15 Αποτελέσματα ως προς το κριτήριο AIC μέσω της μεθόδου της Backward Elimination για την εύρεση του βέλτιστου μοντέλου στο μοντέλο του Cox

	coef	exp(coef)	se(coef)	z	p
tsize	0.0073151	1.0073420	0.0038898	1.881	0.06003
hormonef1	-0.3235059	0.7236077	0.1258226	-2.571	0.01014
tgradef2	0.6438880	1.9038688	0.2490175	2.586	0.00972
tgradef3	0.7879212	2.1988208	0.2682588	2.937	0.00331
pnodes	0.0489911	1.0502110	0.0074538	6.573	4.94e-11
progrec	-0.0022168	0.9977856	0.0005538	-4.003	6.25e-05

Likelihood ratio test=102.5 on 6 df, p=< 2.2e-16

	exp(coef)	exp(-coef)	lower .95	upper .95
tsize	1.0073	0.9927	0.9997	1.0151
hormonef1	0.7236	1.3820	0.5655	0.9260
tgradef2	1.9039	0.5252	1.1686	3.1017
tgradef3	2.1988	0.4548	1.2997	3.7199
pnodes	1.0502	0.9522	1.0350	1.0657
progrec	0.9978	1.0022	0.9967	0.9989

Concordance = 0.689 (se = 0.015)

Likelihood ratio test = 102.5 on 6 df, p =< 2e-16

Wald test = 112.1 on 6 df, p =< 2e-16

Score (logrank) test = 118.5 on 6 df, p =< 2e-16

Πίνακας 16 Προσαρμογή του μοντέλου αναλογικής διακινδύνευσης του Cox στο τελικό μοντέλο

5.4.2 Συμπεράσματα

Στην παράγραφο 5.4.1 υλοποιήθηκε η προσαρμογή του μοντέλου αναλογικής διακινδύνευσης του Cox αξιοποιώντας τον έλεγχο Wald, τον έλεγχο του λόγου των πιθανοφανειών και της μεθόδου της Backward Elimination, η οποία βασίστηκε στην εύρεση του μοντέλου με το μικρότερο AIC, ώστε να επιτευχθεί το βέλτιστο μοντέλο συμμεταβλητών για τη διεξαγωγή της έρευνας. Ως εκ τούτου εξάγονται κάποια συμπεράσματα.

Παρατηρώντας κάποιος στον πίνακα 14 τον έλεγχο Wald (μέσω της μεταβλητής z) και των αντίστοιχων p – values διακρίνει εμφανώς ότι υπάρχουν μεταβλητές με p – value > 0,05. Συγκεκριμένα, οι συμμεταβλητές της ηλικίας (age), της εμμηνόπαυσης (menostatf1) και της κατάστασης του υποδοχέα οιστρογόνου (estrec) είναι στατιστικά μη σημαντικές, αφού οδηγούν σε αποδοχή της μηδενικής υπόθεσης, συνθήκη που προτιμάται να αποφεύγεται. Από την άλλη πλευρά βέβαια, ο έλεγχος του λόγου των πιθανοφανειών (likelihood ratio test), ο έλεγχος log – rank, αλλά κι ο έλεγχος Wald συλλογικά για όλο το μοντέλο δίνουν μια p – value < 0,05, συνθήκη η οποία οδηγεί σε απόρριψη της μηδενικής υπόθεσης. Παράλληλα, η τιμή του concordance είναι 0.692 που μαρτυράει μια σχετικά καλή προβλεπτική ικανότητα. Επομένως πρόκειται για ένα μοντέλο το οποίο προβληματίζει ως προς το κατά πόσο είναι στατιστικά σημαντικό, αφού υπάρχει αντίφαση ανάμεσα στους ελέγχους κι είναι επιθυμητό να εφαρμοστεί η μέθοδος της Backward Elimination, ώστε μέσω του κριτηρίου AIC να βρεθεί πιο ξεκάθαρα το βέλτιστο μοντέλο για τα υπάρχοντα δεδομένα.

Παρατηρώντας τα αποτελέσματα στον πίνακα 15 φαίνεται ότι το κριτήριο AIC του βέλτιστου μοντέλου με τις λιγότερες δυνατές μεταβλητές, καθώς κι ο έλεγχος των λόγων πιθανοφανειών δε διαφέρουν και τόσο από τα αντίστοιχα του αρχικού μοντέλου με όλες τις συμμεταβλητές, δείγμα που μαρτυράει ότι τελικά στο αρχικό μοντέλο χρειάζονταν μικρές βελτιώσεις, όπως αποδείχτηκε και από την εφαρμογή των ελέγχων στο αρχικό μοντέλο. Συνυπολογίζοντας τον πίνακα 16, κάποιος καταλήγει στο ότι το βέλτιστο μοντέλο είναι εκείνο με τις συμμεταβλητές του μεγέθους του όγκου (tsize), της ένδειξης ορμονοθεραπείας (hormonef), του βαθμού του όγκου (tgradef), του αριθμού των θετικών λεμφαδένων (pnodes) και της κατάστασης του υποδοχέα προγεστερόνης (progrec). Πράγματι, από τους ελέγχους Wald, του λόγου των πιθανοφανειών και του log – rank test φαίνεται ότι όλες σχεδόν

οι προαναφερθείσες συμμεταβλητές έχουν p – value < 0,05, δηλαδή είναι στατιστικά σημαντικές για την επιβίωση των γυναικών με καρκίνο στο μαστό. Εξαίρεση αποτελεί ξανά, όπως και στο μοντέλο παλινδρόμησης Weibull η μεταβλητή του μεγέθους του όγκου (tsize), η οποία έχει οριακή τιμή κοντά στην επιθυμητή p – value, αλλά γίνεται αποδεκτή για το τελικό μοντέλο. Επιπλέον, η τιμή του concordance είναι 0.689, σχεδόν ίδια με του αρχικού μοντέλου, άρα υπάρχει μια σχετικά καλή προβλεπτική ικανότητα.

5.4.3 Διαστήματα εμπιστοσύνης και ερμηνείες

Στον πίνακα 17 κάποιος δύνανται να δει τα διαστήματα εμπιστοσύνης 95% ως προς τους συντελεστές των συμμεταβλητών του τελικού μοντέλου, ενώ στον πίνακα 18 παρατηρεί τα διαστήματα εμπιστοσύνης 95% ως προς τους συντελεστές των συμμεταβλητών του τελικού μοντέλου υψωμένους ως εκθέτες στην εκθετική συνάρτηση (τα οποία είχαν βρεθεί και στην προσαρμογή του μοντέλου του Cox στην παράγραφο 5.4.1 στον πίνακα 16), μιας και στο μοντέλο του Cox λόγω της μορφής του $(h(t;x)=h_0(t)e^{x'\beta})$ ενδείκνυται να χρησιμοποιούνται τα εκθετικά έναντι των απλών συντελεστών.

	2.5 %	97.5 %
tsize	-0.0003087705	0.014939070
hormonef1	-0.5701136269	-0.076898081
tgradef2	0.1558228041	1.131953284
tgradef3	0.2621436352	1.313698774
pnodes	0.0343819115	0.063600205
progrec	-0.0033021709	-0.001131442

Πίνακας 17 Διαστήματα εμπιστοσύνης 95% για τους συντελεστές των συμμεταβλητών του τελικού μοντέλου

	2.5 %	97.5 %
tsize	0.9996913	1.0150512
hormonef1	0.5654612	0.9259842
tgradef2	1.1686191	3.1017091
tgradef3	1.2997132	3.7199074
pnodes	1.0349798	1.0656663
progrec	0.9967033	0.9988692

Πίνακας 18 Διαστήματα εμπιστοσύνης 95% για τα εκθετικά των συντελεστών των συμμεταβλητών του τελικού μοντέλου

Ερμηνείες. Εστιάζοντας στα εκθετικά των συντελεστών στον πίνακα 16 και στα διαστήματα εμπιστοσύνης τους (πίνακας 18) προκύπτουν ενδιαφέρουσες ερμηνείες ως προς αυτούς. Υπενθυμίζεται ότι τα εκθετικά των συντελεστών $\hat{\beta}_i, i = 1,2,3,4,5,6$ (πίνακας 16) των συμμεταβλητών του τελικού μοντέλου βασίζονται στην αντιστοιχία

$$i = 1 \rightarrow \text{tsize}$$
$$i = 2 \rightarrow \text{hormonef1}$$
$$i = 3 \rightarrow \text{tgradef2}$$
$$i = 4 \rightarrow \text{tgradef3}$$
$$i = 5 \rightarrow \text{pnodes}$$
$$i = 6 \rightarrow \text{progrec}$$

- Αν αυξανόταν η συμμεταβλητή του μεγέθους του όγκου (tsize) κατά μια μονάδα, τότε η συνάρτηση διακινδύνευσης θα αυξανόταν κατά 1,0073 (δηλαδή κατά exp(coef of tsize), δηλαδή θα συνέβαινε πιο γρήγορα το γεγονός (υποτροπή ή θάνατος).

- Αν αυξανόταν η συμμεταβλητή της ένδειξης για ορμονοθεραπεία (hormonef1) κατά μια μονάδα, τότε η συνάρτηση διακινδύνευσης θα μειωνόταν κατά 0,7236 (δηλαδή κατά exp (coef of hormonef1), δηλαδή θα μειωνόταν ο κίνδυνος για επιδείνωση της νόσου και το γεγονός θα συνέβαινε πιο αργά.

- Αν αυξανόταν η συμμεταβλητή του βαθμού του όγκου (tgradef2) κατά μια μονάδα, τότε η συνάρτηση διακινδύνευσης θα αυξανόταν κατά 1,9039, δηλαδή θα αυξανόταν ο κίνδυνος για επιδείνωση του καρκίνου και το συμβάν θα προκαλούνταν ταχύτερα.

- Παρομοίως, αν αυξανόταν κι άλλο η συμμεταβλητή του βαθμού του όγκου (tgradef3) κατά μια μονάδα, τότε η συνάρτηση διακινδύνευσης θα αυξανόταν κατά 2,1988, δηλαδή θα μεγάλωνε πολύ περισσότερο ο κίνδυνος και το γεγονός (να υποτροπιάσει ή να αποβιώσει) θα επιταχυνόταν.

- Αν αυξανόταν η συμμεταβλητή του αριθμού των θετικών λεμφαδένων (pnodes) κατά μια μονάδα, τότε η συνάρτηση διακινδύνευσης θα αυξανόταν κατά 1,0502, δηλαδή θα αυξανόταν ο κίνδυνος για την εκάστοτε γυναίκα, μιας κι όπως έχει προαναφερθεί, πολλαπλασιασμός των θετικών λεμφαδένων συμβάλλει στην μετάσταση του καρκίνου.

- Αν αυξανόταν η συμμεταβλητή της κατάστασης του υποδοχέα της προγεστερόνης (progrec) κατά μια μονάδα, τότε η συνάρτηση διακινδύνευσης θα μειωνόταν λίγο κατά 0,9978, δηλαδή ο κίνδυνος θα μειωνόταν σε μικρό ποσοστό για την εκάστοτε ασθενή.

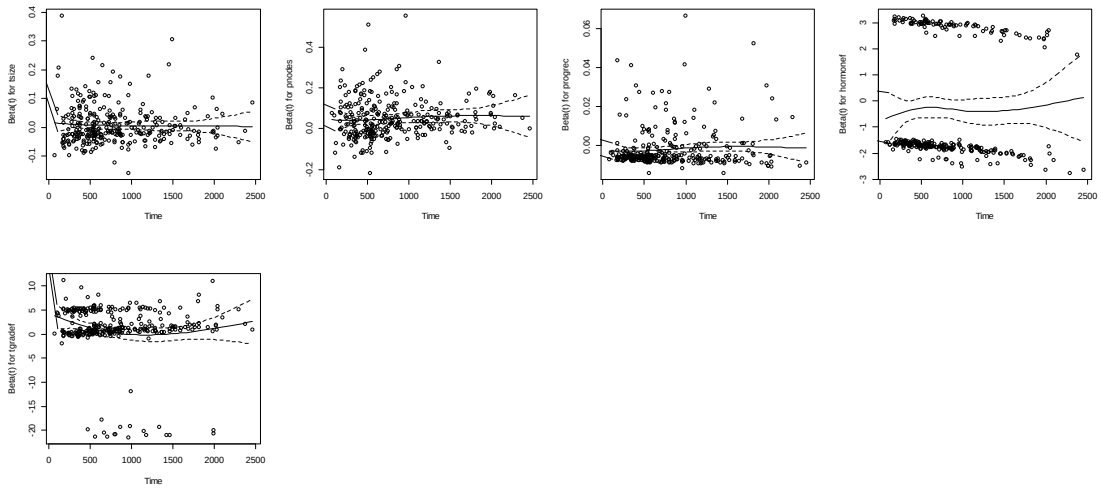
5.4.4 Προϋποθέσεις αναλογικότητας διακινδύνευσης

Για να εξεταστεί αν το τελικό μοντέλο πληροί τις προϋποθέσεις της αναλογικότητας της διακινδύνευσης μέσω στατιστικών ελέγχων, καθώς και με τη χρήση γραφικών παραστάσεων των υπολοίπων Schoenfeld, θα αξιοποιηθούν ο πίνακας 19 και τα γραφήματα 5.4.1 και 5.4.2. Στον πίνακα 19 φαίνεται ο στατιστικός X^2 – έλεγχος καλής προσαρμογής του μοντέλου ως προς τους συντελεστές της κάθε μεταβλητής ξεχωριστά και μετά ως προς έναν καθολικό έλεγχο ως προς όλους τους συντελεστές (μέσω του ελέγχου GLOBAL), ο οποίος έχει τόσους βαθμούς ελευθερίας όσες είναι και οι μεταβλητές, δηλαδή 6 βαθμούς ελευθερίας. Διευκρινίζεται ότι στον πίνακα 19 αρχικά φαίνεται η προσαρμογή του X^2 – ελέγχου καλής προσαρμογής στα δεδομένα παίρνοντας τη συμμεταβλητή tgradef ως μια ενιαία μεταβλητή, ενώ στη συνέχεια θεωρώντας την ως δυο μεταβλητές tgradef2 και tgradef3.

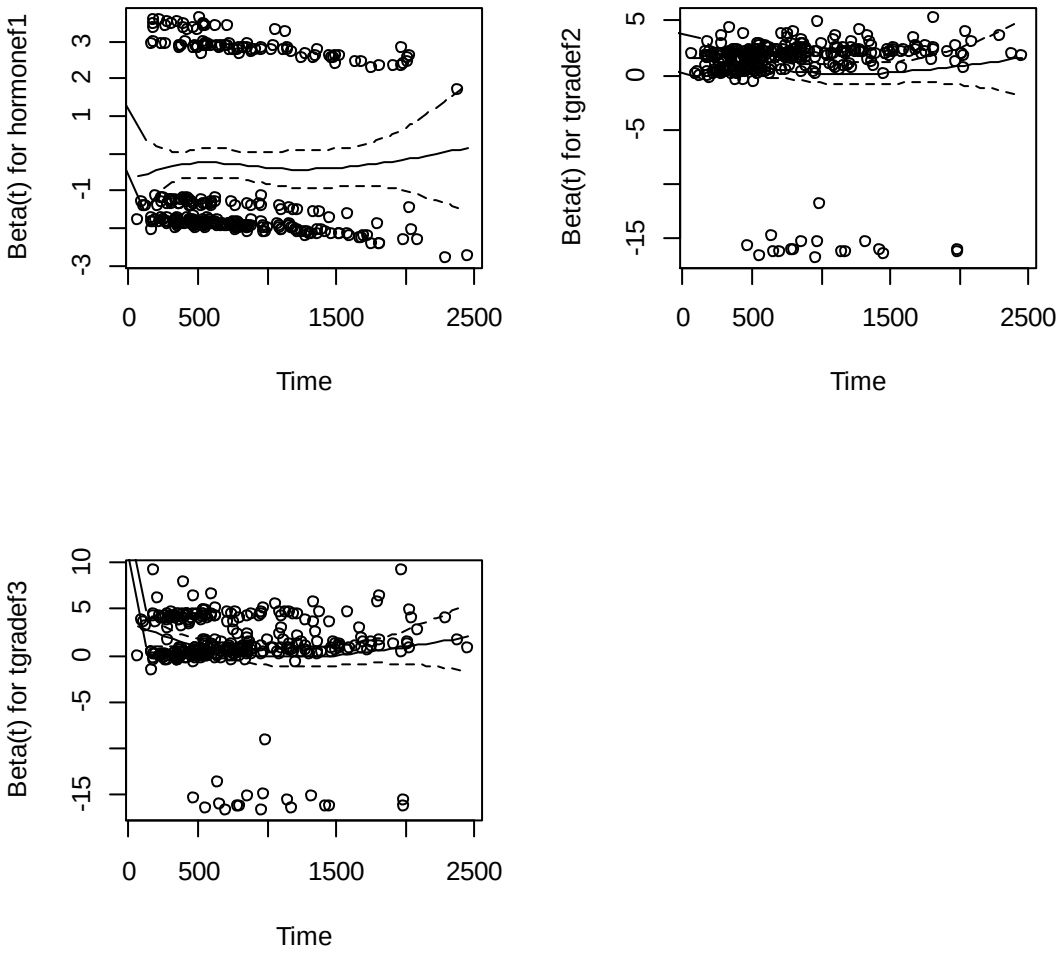
Στο γράφημα 5.4.1 φαίνεται ο γραφικός έλεγχος των υπολοίπων Schoenfeld για τα δεδομένα παίρνοντας τη συμμεταβλητή tgradef ως μια ενιαία μεταβλητή, ενώ στο γράφημα 5.4.2 φαίνεται ο γραφικός έλεγχος των υπολοίπων Schoenfeld για τα δεδομένα παίρνοντας τη συμμεταβλητή tgradef ως δυο μεταβλητές tgradef2 και tgradef3.

	chisq	df	p
tsize	0.103	1	0.749
hormonef	0.170	1	0.680
tgradef	7.512	2	0.023
pnodes	1.281	1	0.258
progrec	3.188	1	0.074
GLOBAL	11.610	6	0.071
hormonef1	0.170	1	0.680
tgradef2	2.529	1	0.112
tgradef3	6.576	1	0.010

Πίνακας 19 Στατιστικός X^2 – έλεγχος καλής προσαρμογής του βέλτιστου μοντέλου



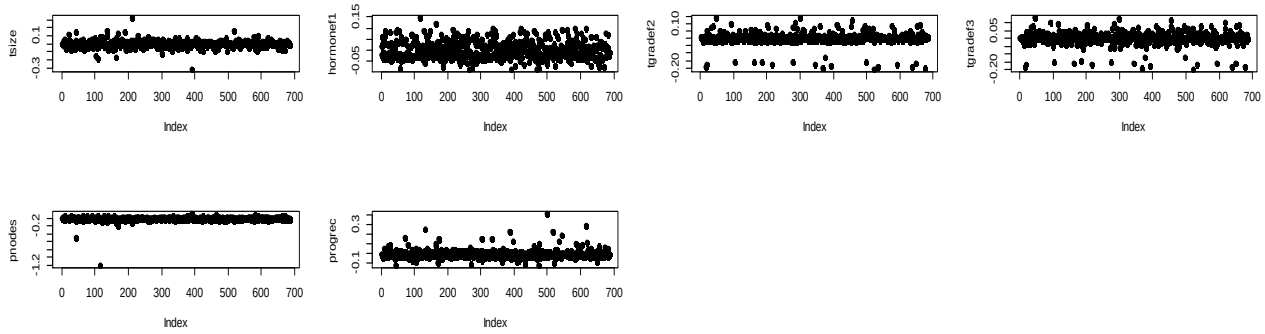
Γράφημα 5.4.1 Γραφικός έλεγχος της υπόθεσης αναλογικότητας της διακινδύνευσης μέσω των υπολοίπων Schoenfeld θεωρώντας τη συμμεταβλητή tgradef ως μια ενιαία μεταβλητή



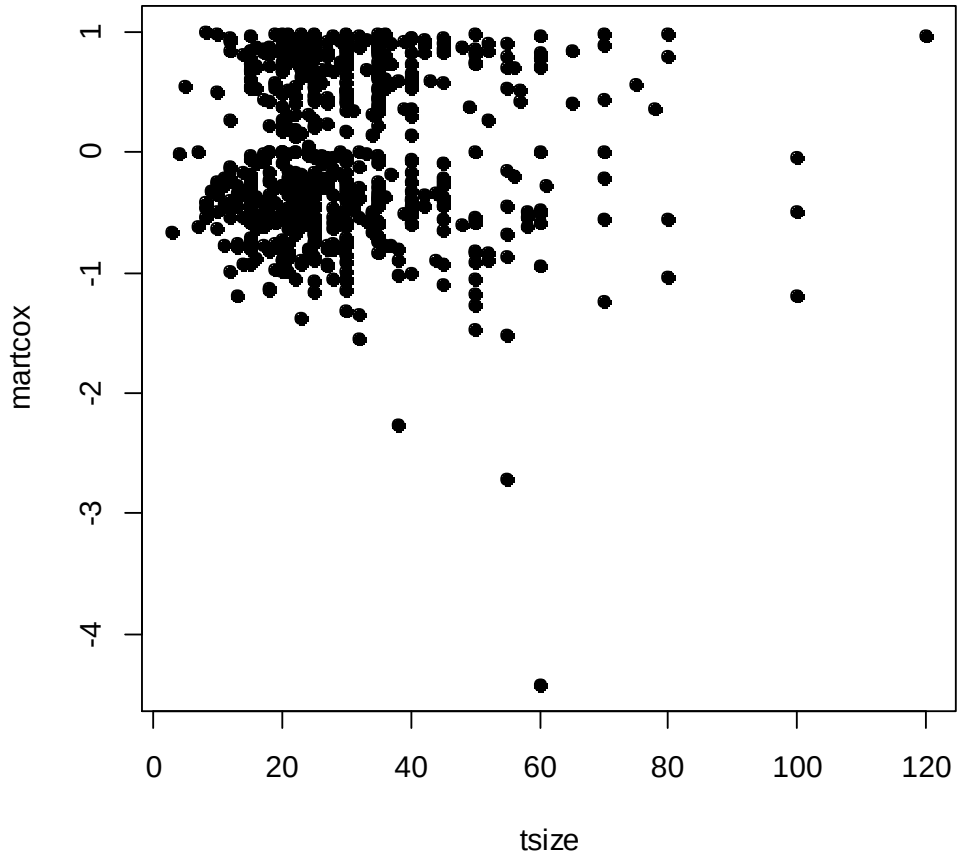
Γράφημα 5.4.2 Γραφικός έλεγχος της υπόθεσης αναλογικότητας της διακινδύνευσης μέσω των υπολοίπων Schoenfeld θεωρώντας τη συμμεταβλητή tgradef ως δύο μεταβλητές

5.4.5 Γραφικές παραστάσεις και συμπεράσματα

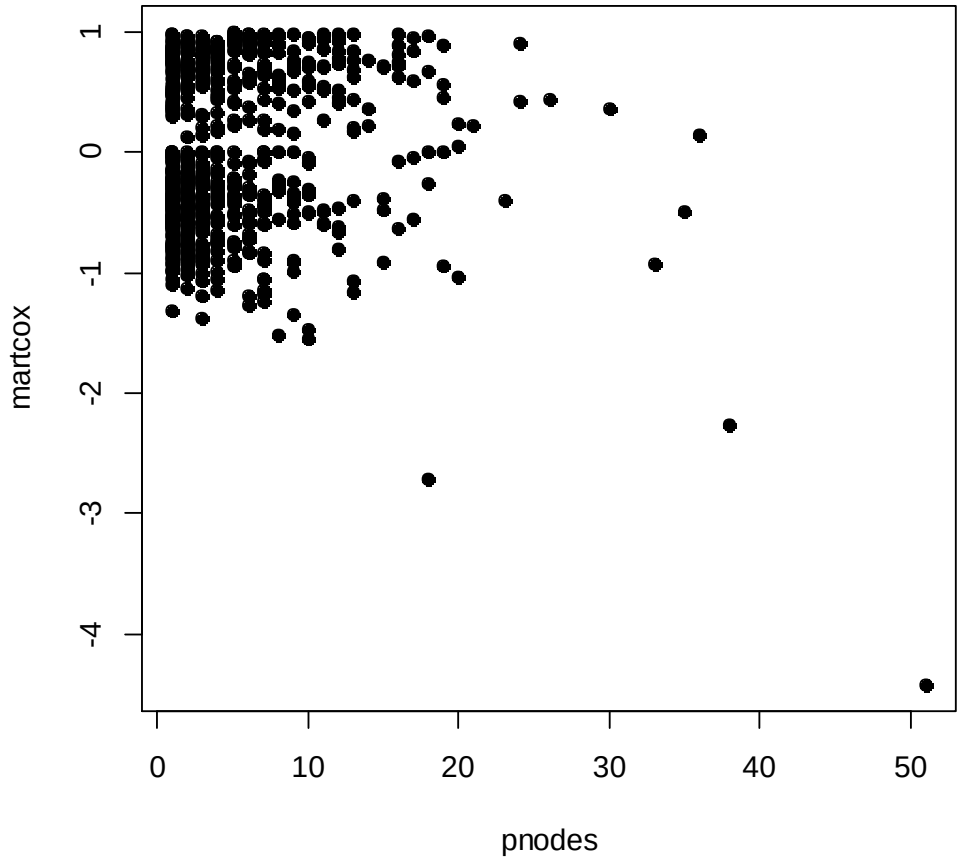
Συμπληρώνοντας την παράγραφο 5.4.4 παρατίθενται κάποια επιπλέον διαγράμματα για μεγαλύτερη διασφάλιση των σωστών συμπερασμάτων. Συγκεκριμένα, στο γράφημα 5.4.3 κάποιος μπορεί να παρατηρήσει τα υπόλοιπα DFBETAS, στα γραφήματα 5.4.4, 5.4.5 και 5.4.6 τα υπόλοιπα Martingale ως προς τις συμμεταβλητές του μεγέθους του όγκου, του αριθμού θετικών λεμφαδένων και της κατάστασης του υποδοχέα προγεστερόνης αντίστοιχα, ενώ τέλος στο γράφημα 5.4.7 μπορεί να δει τα υπόλοιπα Deviance.



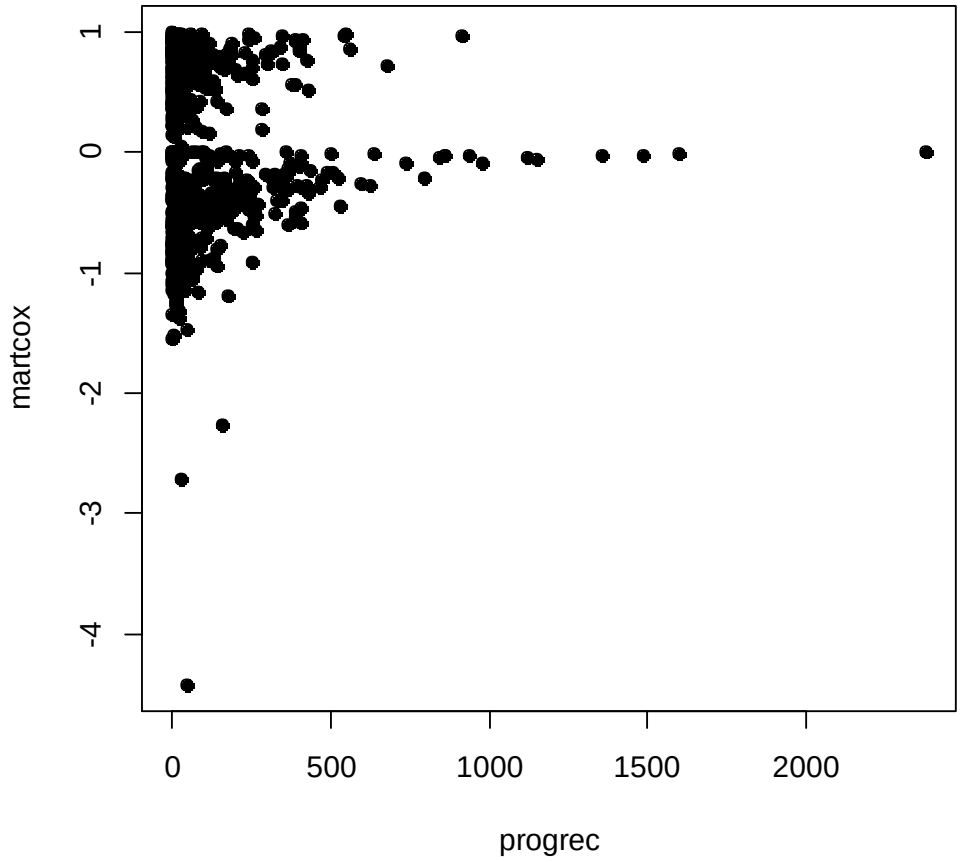
Γράφημα 5.4.3 Υπόλοιπα DFBETAS



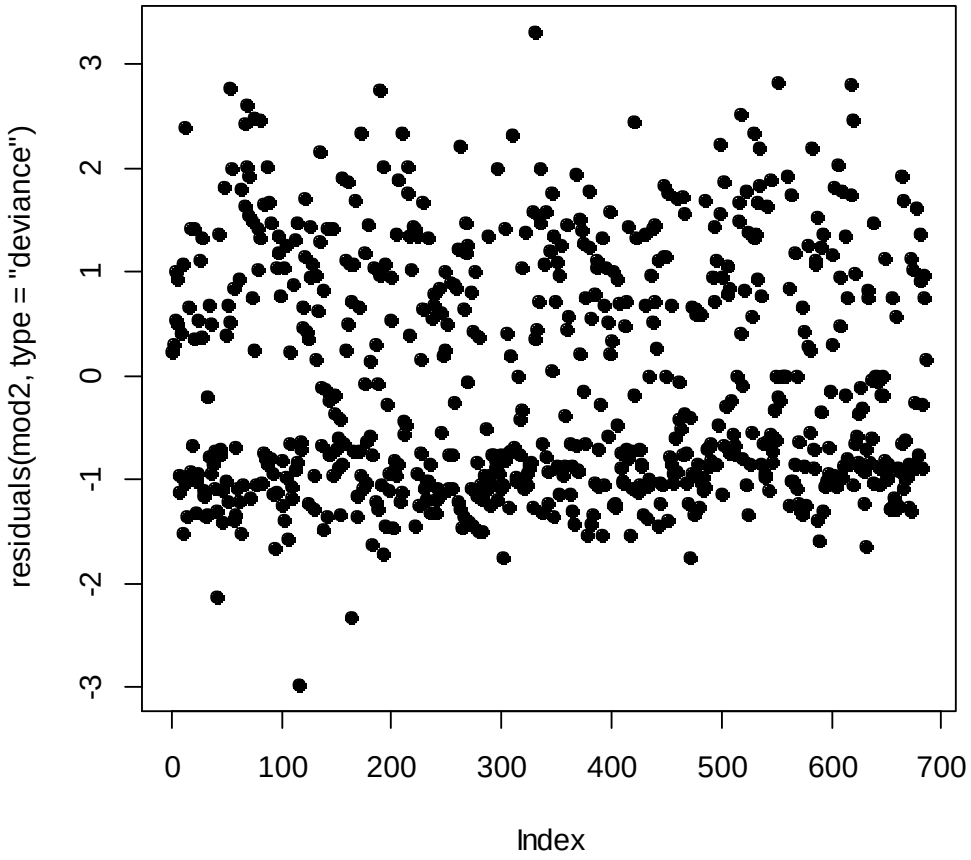
Γράφημα 5.4.4 Υπόλοιπα Martingale για τη μεταβλητή του μεγέθους του όγκου (tsize)



Γράφημα 5.4.5 Υπόλοιπα Martingale για τη μεταβλητή του αριθμού θετικών λεμφαδένων (pnodes)



Γράφημα 5.4.6 Υπόλοιπα Martingale για τη μεταβλητή της κατάστασης του υποδοχέα προγεστερόνης (progrec)



Γράφημα 5.4.7 Υπόλοιπα Deviance

Συμπεράσματα

Παρατηρώντας κάποιος στον πίνακα 19 τα αποτελέσματα από το στατιστικό X^2 – έλεγχο καλής προσαρμογής του βέλτιστου μοντέλου διαπιστώνει ότι σχεδόν για όλους τους συντελεστές των συμμεταβλητών ξεχωριστά, αλλά και ως ενιαίο σύνολο (μέσω του ελέγχου GLOBAL), οι τιμές της p – value βγαίνουν σχετικά μεγάλες, δηλαδή είναι άνω της τιμής 0,05, γεγονός που μαρτυράει ότι τα δεδομένα είναι στατιστικά μη σημαντικά κι επομένως δεν υπάρχει χρονική εξάρτηση στους συντελεστές των μεταβλητών του μοντέλου. Συνεπώς πληρούνται οι προϋποθέσεις της αναλογικής διακινδύνευσης. Εξαίρεση αποτελεί η μεταβλητή tgradef3, η οποία αποδεικνύεται στατιστικά σημαντική, άρα ότι σχετίζεται σε κάποιο βαθμό με το χρόνο, αλλά δε θα ληφθεί υπόψη, αφού ο καθολικός έλεγχος για το σύνολο των συμμεταβλητών βγαίνει στατιστικά μη σημαντικός.

Στη συνέχεια, στα διαγράμματα 5.4.1 και 5.4.2 παρατηρείται ότι η R έχει προσαρμόσει τα δεδομένα μέσω κάποιων γραμμών σε διαστήματα εμπιστοσύνης. Εύκολα κάποιος είναι σε θέση να αντιληφθεί ότι αυτές οι γραμμές σχεδόν στο σύνολό τους είναι παράλληλες στον οριζόντιο άξονα, γεγονός που ενισχύει την ισχύ της πληρότητας των προδιαγραφών της αναλογικής διακινδύνευσης του Cox. Παραβίαση της παραλληλότητας συμβαίνει πάλι από τη μεταβλητή tgradef3, η οποία έχει μια ελαφριά καμπυλότητα, αλλά σε τόσο μικρό βαθμό που δε μπορεί να επηρεάσει στο σύνολο τη συσχέτιση των συντελεστών των μεταβλητών με το χρόνο. Διευκρινίζεται ότι στη συμμεταβλητή της ένδειξης ορμονοθεραπείας (hormonef) τα σημεία στη γραφική παράσταση φαίνονται να συγκεντρώνονται σε δύο μόνο παράλληλες γραμμές, επειδή η συγκεκριμένη συμμεταβλητή είναι δυαδική και παίρνει μόνο δύο τιμές.

Συνοψίζοντας, κάποιος καταλήγει στο συμπέρασμα ότι το τελικό μοντέλο πληροί τις προδιαγραφές της αναλογικότητας της διακινδύνευσης, δηλαδή οι συντελεστές των μεταβλητών δε σχετίζονται με το χρόνο.

Παράλληλα, παρατηρώντας κάποιος το διάγραμμα 5.4.3 με τα υπόλοιπα DFBETAS και συνυπολογίζοντας τη σχέση

$$\left|DFBETAS_{ji}\right| > \frac{2}{\sqrt{n}}$$

για $n = 686$ καταλήγει ότι υπάρχουν ενδείξεις για πολλά σημεία επιρροής για όλους τους συντελεστές των συμμεταβλητών του βέλτιστου μοντέλου. Ωστόσο εστιάζοντας την προσοχή στα διαγράμματα 5.4.4, 5.4.5 και 5.4.6 με τα υπόλοιπα Martingale, κάποιος θα προσέξει ότι και για τις τρεις ουσιαστικά συμμεταβλητές σχηματίζονται ευθείες γραμμές (με πιο έντονο το φαινόμενο να εμφανίζεται στη συμμεταβλητή του αριθμού θετικών λεμφαδένων (pnodes)), κάτι το οποίο υποδηλώνει ότι η συναρτησιακή μορφή των στατιστικά σημαντικών ποσοτικών μεταβλητών του βέλτιστου μοντέλου είναι κατάλληλη και δε χρειάζεται κάποιος μετασχηματισμός. Υπενθυμίζεται ότι οι αρνητικές τιμές αντιστοιχούν σε υπόλοιπα για αποκομμένες παρατηρήσεις. Τέλος, στο γράφημα 5.4.7, όπου απεικονίζονται τα υπόλοιπα Deviance είναι φανερό ότι τα υπόλοιπα κατανέμονται αρκετά συμμετρικά γύρω από το μηδέν, χωρίς να υπάρχουν πολύ μεγάλα κατά απόλυτη τιμή υπόλοιπα, τα οποία θα πρόδιδαν την ύπαρξη πιθανών άτυπων (έκτροπων) παρατηρήσεων.

Σε γενικές γραμμές προκύπτει κι ότι μέσα από τον έλεγχο διαφόρων ειδών υπολοίπων, το βέλτιστο μοντέλο ταιριάζει σε αρκετά ικανοποιητικό βαθμό στα δεδομένα του προβλήματος.

6 Γενικά Συμπεράσματα

Στο κεφάλαιο 5 πραγματοποιήθηκε πειραματική προσέγγιση, ώστε να διερευνηθεί κατά πόσο διάφορες συμμεταβλητές σε ένα δείγμα δεδομένων (συνήθως ασθενών) επηρεάζουν την επιβίωσή τους μέχρι να συμβεί το πρώτο γεγονός, το οποίο συνήθως είναι υποτροπή της ασθένειας ή θάνατος.

Στην παρούσα εργασία αξιοποιήθηκαν πραγματικά δεδομένα από 686 ασθενείς γυναίκες που είχαν διαγνωστεί σε πρώιμο στάδιο καρκίνου του μαστού από τον Ιούλιο του 1984 έως τον Δεκέμβριο του 1989. Τα στοιχεία προήλθαν από τη 'Γερμανική Ομάδα Μελέτης για τον Καρκίνο του Μαστού' (the German Breast Cancer Study Group). Στα δεδομένα υπήρχαν αποκομμένες παρατηρήσεις κι όλες οι αναλύσεις προσαρμόστηκαν με χρήση ή όχι ορμονικής θεραπείας μέσω ταμοξιφαΐνης. Οι συμμεταβλητές που λήφθηκαν υπόψη σχετίζονταν με την ηλικία των ασθενών, το μέγεθος και το βαθμό του όγκου, τον αριθμό των θετικών λεμφαδένων, την κατάσταση του υποδοχέα προγεστερόνης και οιστρογόνου, καθώς και την κατάσταση της εμμηνόπαυσης. Στη στατιστική ανάλυση που υλοποιήθηκε εφαρμόστηκαν στατιστικοί έλεγχοι και μέθοδοι της Ανάλυσης Επιβίωσης. Σκοπός ήταν η ανάλυση του χρόνου επιβίωσης χωρίς υποτροπή των 686 ασθενών σχετικά με τις προαναφερθείσες μεταβλητές μέχρι να συμβεί το πρώτο δυσάρεστο γεγονός (υποτροπή ή θάνατος).

Εισάγοντας τα δεδομένα στο στατιστικό πακέτο Minitab, καθώς και στη γλώσσα προγραμματισμού ανοιχτού κώδικα R, η ανάλυση προσεγγίστηκε με μια πληθώρα μεθόδων.

Συγκεκριμένα κατασκευάστηκαν γραφικές παραστάσεις της εκτιμήτριας Kaplan – Meier κι εξετάστηκε γραφικά αν στα δεδομένα ταιριάζει ένα μοντέλο παλινδρόμησης της επιταχυνόμενης διακοπής ή ένα μοντέλο της αναλογικής διακινδύνευσης.

Επιπροσθέτως προσαρμόστηκαν το μοντέλο παλινδρόμησης επιταχυνόμενης διακοπής της κατανομής Weibull, όπως επίσης και το μοντέλο αναλογικής διακινδύνευσης του Cox εφαρμόζοντας κυρίως ελέγχους Wald, ελέγχους του λόγου των πιθανοφανειών και το κριτήριο AIC μέσω της Backward Elimination τεχνικής με σκοπό την εύρεση του βέλτιστου μοντέλου. Επιπλέον κατασκευάστηκαν κι

ερμηνεύτηκαν 0.95 – διαστήματα εμπιστοσύνης για τους συντελεστές των συμμεταβλητών του τελικού (βέλτιστου) μοντέλου και στο τέλος, μέσω γραφικών κυρίως ελέγχων των υπολοίπων εξετάστηκε η ισχύς όλων των παραπάνω.

Η εφαρμογή όλων αυτών οδήγησε στην εξαγωγή χρήσιμων συμπερασμάτων που παρουσιάζουν αρκετό ενδιαφέρον.

Αρχικά αποδείχτηκε ότι η χρήση ορμονοθεραπείας με ταμοξιφαΐνη σίγουρα συνέβαλε θετικά στην επιβίωση των γυναικών με όγκο στο μαστό. Τουναντίον, η ηλικία και η κατάσταση της εμμηνόπαυσης φάνηκαν σα να μην επηρέασαν ιδιαίτερα την εξέλιξη της νόσου. Από την άλλη πλευρά, ο αριθμός των θετικών λεμφαδένων και ο βαθμός του όγκου έπαιξαν σημαντικό ρόλο για την επιδείνωση του καρκίνου, ενώ παραδόξως το μέγεθος του όγκου φάνηκε να επηρεάζει σε μικρότερο βαθμό από ό,τι θα προβλεπόταν την εξέλιξη της. ασθένειας. Εν τέλει, η κατάσταση του υποδοχέα της προγεστερόνης αποδείχτηκε να ασκεί κάποιαν επιρροή στην επιβίωση των γυναικών συμβάλλοντας θετικά στην καθυστέρηση του συμβάντος, δηλαδή της υποτροπής ή του θανάτου, ενώ, αντιθέτως, η κατάσταση του υποδοχέα του οιστρογόνου δεν έδειξε να επηρεάζει και τόσο την έκβαση της νόσου.

Τελικά, ως βέλτιστο μοντέλο αναδείχτηκε εκείνο με τις συμμεταβλητές της ορμονοθεραπείας, του μεγέθους και του βαθμού του όγκου, του αριθμού των θετικών λεμφαδένων και της κατάστασης του υποδοχέα προγεστερόνης. Με άλλα λόγια δηλαδή, αυτές αποτελούν τις συμμεταβλητές που μπορούσαν να επηρεάσουν σε μεγαλύτερο ή μικρότερο βαθμό την εξέλιξη της νόσου, άρα και το χρόνο επιβίωσης των γυναικών.

Σε γενικές γραμμές, τόσο το μοντέλο παλινδρόμησης επιταχυνόμενης διακοπής της κατανομής Weibull, όσο και το μοντέλο αναλογικής διακινδύνευσης του Cox, όταν εφαρμόστηκαν αξιοποιώντας περίπου τους ίδιους ελέγχους και τεχνικές κατέληξαν στα ίδια ακριβώς αποτελέσματα. Παράλληλα όμως αποδείχτηκε ότι ευσταθεί η υπόθεση της επιταχυνόμενης διακοπής ως προς τη μεταβλητή της ορμονοθεραπείας, δηλαδή ότι η προσαρμοσμένη καμπύλη επιβίωσης της κατανομής Weibull για τη συμμεταβλητή της ορμόνης ταιριάζει στα δεδομένα, όπως κι ότι το μοντέλο της αναλογικής διακινδύνευσης στην ειδική περίπτωση της κατανομής Weibull, δηλαδή η

προσαρμογή της κατανομής Weibull με τη μέθοδο της μέγιστης πιθανοφάνειας για τη συμμεταβλητή της ορμόνης έχει πολύ καλή προσαρμογή σε αυτά τα δεδομένα. Προφανώς, τα προαναφερθέντα αποτελούν ιδιαίτερα χρήσιμες πληροφορίες, αφού όλες οι αναλύσεις προσαρμόστηκαν με χρήση ή όχι ορμονικής θεραπείας. Από την άλλη πλευρά, σε ό,τι αφορά στο μοντέλο του Cox υλοποιήθηκαν αρκετοί γραφικοί έλεγχοι, οι οποίοι στην πλειοψηφία τους, με εξαίρεση των υπολοίπων DFBETAS απέδειξαν ότι το βέλτιστο μοντέλο που βρέθηκε ταιριάζει σε μεγάλο βαθμό στα δεδομένα του προβλήματος, χωρίς την ύπαρξη πιθανών άτυπων παρατηρήσεων. Κι οι δύο δηλαδή αναλύσεις συνέκλιναν στα ίδια αποτελέσματα είτε ακολούθησαν την ίδια προσέγγιση είτε διαφορετική, γνώση πολύτιμη για την αξιοπιστία και την εγκυρότητα των συμπερασμάτων.

Τα τελευταία χρόνια ωστόσο, η ιατρική έχει κάνει άλματα στην ανακάλυψη νέων φαρμάκων και θεραπειών για πολλές ασθένειες, με τον καρκίνο να πρωτοστατεί. Ως εκ τούτου, η περαιτέρω επεξεργασία των δεδομένων με πολύ περισσότερους τρόπους λαμβάνοντας υπόψη τις όποιες καινούριες τάσεις γύρω από τη θεραπεία του καρκίνου του μαστού, καθώς και η προσθήκη περισσότερων συμμεταβλητών στην έρευνα, όπως ο παράγοντας της κληρονομικότητας θα αποτελούσαν σημαντικό κομμάτι για μελλοντική βελτίωση.

Βιβλιογραφία

Ξενόγλωσση

Caroni, C. (2004), Diagnostics for Cox's proportional hazards model. In M.S. Nikulin, N. Balakrishnan, M. Mesbah and N. Limnios (eds) *Parametric and Semiparametric Models with Applications to Reliability, Survival Analysis, and Quality of Life* in honour of Prof. Catherine Huber, Birkhauser, Boston, pp. 27 – 38.

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

Chambers, R.L., Steel, D.G., Wang, S. and Welsh, A. (2012), *Maximum Likelihood Estimation for Sample Surveys*, CRC Press.

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

Cox, D.R. and Hinkley, D.V. (1974), *Theoretical Statistics*, Chapman & Hall.

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

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

Davidson, R. and MacKinnon, J.G. (1993), *Estimation and Inference in Econometrics*, Oxford University Pres.

Fahrmei, L., Kneib, T., Lang, S. and Marx, B. (2013), *Regression: Models, Methods and Applications.*

Ferber, D. (2003), Carcinogens. Lashed by critics, WHO's cancer agency begins a new regime, *Science*, **301**, pp. 36 – 37.

Ferlay, J., Ervik, M., Lam, F., Colombet, M., Mery, L. and Piñeros, M. (2020), Global Cancer Observatory, International Agency for Research on Cancer.

Harrington, D. (2005), Linear rank tests in survival analysis, *Encyclopedia of Biostatistics*, Wiley Interscience.

Hendry, D.F. and Nielsen, B. (2007), *Econometric Modeling: A Likelihood Approach*, Princeton University Press.

Hevesi, D. and Meier, P. (2011), Statistician who revolutionized medical trials, dies at 87, *The New York Times.*

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

Kendall, M.G. and Stuart, D.G. (1973), *The Advanced Theory of Statistics,* **2**, Inference and Relationship, Griffin, London Section 20.4.

Mantel, N. (1966), Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer Chemotherapy Reports,* **50** (3), pp. 163 – 70.

Martin, V., Hurn, S. and Harris, D. (2013), *Econometric Modeling with Time Series: Specification, Estimation and Testing,* Cambridge University Press.

Meyer, B.D. (1990), Unemployment insurance and unemployment spells, *Econometrica* **58** (4), pp. 757 – 782.

Neyman, J. (1937), Outline of a theory of statistical estimation based on the classical theory of probability, *Philosophical Transactions of the Royal Society of London, Series A,* **236**, pp. 333 – 380.

OMNITAB (1964), *Digital Computer Newsletter*, **16** (1), pp. 4 – 6.

Peavy, S.T. (1986), OMNITAB 80, *NBS Special Publication*, **701**, pp. 1 – 2.

Peto, R. and Peto, J. (1972), Asymptotically efficient rank invariant test procedures, *Journal of the Royal Statistical Society, Series A*, **135** (2), pp. 185 – 207.

Pfanzagl, J. and Hamböker, R. (1994), *Parametric Statistical Theory,* Walter de Gruyter, Berlin, DE.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1992), Least squares as a maximum likelihood estimator, *Numerical Recipes in Fortran 77: The Art of Scientific Computing (2nd ed.),* Cambridge University Press **1**, pp. 651 – 655.

Rossi, R.J. (2018), *Mathematical Statistics: An Introduction to Likelihood Based Inference,* John Wiley & Sons.

Sauerbrei, W. and Royston, P. (1999), Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials, *Journal of the Royal Statistics Society, Series A*, **162** (1), pp. 71 – 94.

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

Stalpers, L.J.A. and Kaplan, E.L. (2018), Edward L. Kaplan and the Kaplan – Meier survival curve, *Journal of the British Society for the History of Mathematics,* **33** (2), pp. 109 – 135.

Ward, M.D. and Ahlquist, J.S. (2018), *Maximum Likelihood for Social Science: Strategies for Analysis,* Cambridge University Press.

Zar, J.H. (1984), *Biostatistical Analysis,* Prentice Hall International, New Jersey.

Ελληνική

Καρώνη, Χ. και Οικονόμου, Π. (2020), *Στατιστικά Μοντέλα Παλινδρόμησης, 2η Έκδοση*, Αθήνα, Εκδόσεις Συμεών.

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

Φωκιανός, Κ. και Χαραλάμπους, Χ. (2010), *Εισαγωγή στην R: Πρόχειρες Σημειώσεις, 2η Έκδοση*, Πανεπιστήμιο Κύπρου.

Παραρτήματα

Π.1 Αποτελέσματα Minitab για την Kaplan-Meier

Π.2 Αποτελέσματα R για την προσαρμογή της κατανομής Weibull

Π.3 Αποτελέσματα R για την προσαρμογή του μοντέλου του Cox

ΙΙ.1 Αποτελέσματα Minitab για την Kaplan – Meier

Εκτιμήσεις Kaplan – Meier

Distribution Analysis: time by horTh

Variable: time

horTh = no

Censoring Information Count
Uncensored value 205
Right censored value 235

Censoring value: cens = 0

Nonparametric Estimates

Characteristics of Variable

	Standard	95.0% Normal CI	
Mean(MTTF)	Error	Lower	Upper
1512.62	46.1176	1422.23	1603.01

Median = 1528

IQR = 1827 Q1 = 629 Q3 = 2456

Kaplan-Meier Estimates

Time	Number at Risk	Number Failed	Survival Probability	Standard Error	95.0% Normal CI Lower	Upper
72	430	1	0.997674	0.002323	0.993122	1.00000
98	429	1	0.995349	0.003281	0.988918	1.00000
113	428	1	0.993023	0.004014	0.985156	1.00000
…						
2093	27	1	0.387073	0.035411	0.317668	0.45648
2286	10	1	0.348366	0.048622	0.253068	0.44366
2456	3	1	0.232244	0.100201	0.035853	0.42863

Distribution Analysis: time by horTh

Variable: time

horTh = yes

Censoring Information Count
Uncensored value 94
Right censored value 152

Censoring value: cens = 0

Nonparametric Estimates

Characteristics of Variable

	Standard	95.0% Normal CI	
Mean(MTTF)	Error	Lower	Upper
1690.81	55.0708	1582.87	1798.74

Median = 2018

IQR = * Q1 = 859 Q3 = *

Kaplan-Meier Estimates

Time	Number at Risk	Number Failed	Survival Probability	Standard Error	95.0% Normal CI Lower	Upper
169	240	1	0.995833	0.0041580	0.987684	1.00000
177	239	2	0.987500	0.0071716	0.973444	1.00000
180	237	1	0.983333	0.0082636	0.967137	0.99953
…						
2018	30	1	0.492647	0.0455092	0.403451	0.58184
2030	27	1	0.474401	0.0473404	0.381616	0.56719
2372	13	1	0.437909	0.0560254	0.328101	0.54772

Distribution Analysis: time by tgrad

Variable: time

tgrad = 1

Censoring Information Count

Uncensored value 18

Right censored value 63

Censoring value: cens = 0

Nonparametric Estimates

Characteristics of Variable

	Standard	95,0% Normal CI	
Mean(MTTF)	Error	Lower	Upper
2080,79	97,6198	1889,46	2272,12

Median = *

IQR = * Q1 = 1459 Q3 = *

Kaplan-Meier Estimates

Time	Number at Risk	Number Failed	Survival Probability	Standard Error	95,0% Normal CI Lower	Upper
476	73	1	0,986301	0,0136045	0,959637	1,00000
559	71	1	0,972410	0,0192396	0,934701	1,00000
637	70	1	0,958518	0,0234495	0,912558	1,00000
…						
1459	37	1	0,744235	0,0562727	0,633943	0,85453
1989	12	1	0,682216	0,0786558	0,528053	0,83638
1990	11	1	0,620196	0,0927888	0,438333	0,80206

Distribution Analysis: time by tgrad

Variable: time

tgrad = 2

Censoring Information Count

Uncensored value 202

Right censored value 242

Censoring value: cens = 0

Nonparametric Estimates

Characteristics of Variable

	Standard	95,0% Normal CI	
Mean(MTTF)	Error	Lower	Upper
1644,18	49,8033	1546,57	1741,79

Median = 1730

IQR = * Q1 = 745 Q3 = *

Kaplan-Meier Estimates

Time	Number at Risk	Number Failed	Survival Probability	Standard Error	95,0% Normal CI Lower	Upper
72	434	1	0,997696	0,0023015	0,993185	1,00000
160	433	1	0,995392	0,0032510	0,989020	1,00000
169	431	1	0,993082	0,0039802	0,985281	1,00000
…						
2286	19	1	0,383820	0,0380144	0,309313	0,45833
2372	15	1	0,358232	0,0432427	0,273478	0,44299
2456	7	1	0,307056	0,0601554	0,189154	0,42496

Distribution Analysis: time by tgrad

Variable: time

tgrad = 3

Censoring Information Count

Uncensored value 79
Right censored value 82
Censoring value: cens = 0
Nonparametric Estimates

Characteristics of Variable

	Standard	95,0% Normal CI	
Mean(MTTF)	Error	Lower	Upper
1420,08	81,5418	1260,26	1579,90

Median = 1337
IQR = * Q1 = 476 Q3 = *

Kaplan-Meier Estimates

Time	Number at Risk	Number Failed	Survival Probability	Standard Error	95,0% Normal CI Lower	Upper
98	158	1	0,993671	0,0063091	0,981305	1,00000
113	157	1	0,987342	0,0088939	0,969910	1,00000
120	156	1	0,981013	0,0108578	0,959732	1,00000
…						
1806	20	1	0,434707	0,0484247	0,339797	0,52962
1975	12	1	0,398482	0,0563325	0,288072	0,50889
2034	9	1	0,354206	0,0651910	0,226434	0,48198

Distribution Analysis: time by meno

Variable: time

meno = 0

Censoring Information Count
Uncensored value 119
Right censored value 171
Censoring value: cens = 0
Nonparametric Estimates

Characteristics of Variable

	Standard	95,0% Normal CI	
Mean(MTTF)	Error	Lower	Upper
1660,33	61,3937	1540,00	1780,66

Median = 2015
IQR = * Q1 = 648 Q3 = *

Kaplan-Meier Estimates

Time	Number at Risk	Number Failed	Survival Probability	Standard Error	95,0% Normal CI Lower	Upper
120	282	1	0,996454	0,0035398	0,989516	1,00000
169	280	1	0,992895	0,0050060	0,983083	1,00000
171	279	1	0,989336	0,0061238	0,977334	1,00000
…						
2034	29	1	0,480016	0,0409763	0,399703	0,56033
2039	28	1	0,462872	0,0429496	0,378693	0,54705
2093	22	1	0,441832	0,0458620	0,351945	0,53172

Distribution Analysis: time by meno

Variable: time

meno = 1

Censoring Information Count
Uncensored value 180
Right censored value 216
Censoring value: cens = 0
Nonparametric Estimates

Characteristics of Variable

Mean(MTTF)	Standard Error	95,0% Normal CI Lower	Upper
1638,11	52,8319	1534,56	1741,66

Median = 1701

IQR = * Q1 = 745 Q3 = *

Kaplan-Meier Estimates

Time	Number at Risk	Number Failed	Survival Probability	Standard Error	95,0% Normal CI Lower	Upper
72	390	1	0,997436	0,0025608	0,992417	1,00000
98	389	1	0,994872	0,0036169	0,987783	1,00000
113	388	1	0,992308	0,0044240	0,983637	1,00000
…						
2286	17	1	0,378468	0,0409127	0,298281	0,45866
2372	15	1	0,353237	0,0453021	0,264446	0,44203
2456	6	1	0,294364	0,0656775	0,165639	0,42309

Πίνακας 20 Αποτελέσματα Minitab για την Kaplan – Meier

Π.2 Αποτελέσματα R για την προσαρμογή της κατανομής Weibull

Προσαρμογή της Κατανομής Weibull

```
> library(splines)
> library(survival)
> cdat<-read.table("c:/breastcan.txt ",header=TRUE)
> attach(cdat)
> cdat
     id   age  tsize pnodes progrec estrec hormone meno tgrad time   cens
1     1    70    21     3       48     66       0     1     2  1814    1
2     2    56    12     7       61     77       1     1     2  2018    1
3     3    58    35     9       52    271       1     1     2   712    1
…
684 684  51    25     5       43      0        0     0     3   769    1
685 685  52    23     3       15     34        0     1     2   727    1
686 686  55    23     9      116     15        0     1     2  1701    1
> hormonef<-factor(hormone)
> is.factor(hormonef)
[1] TRUE
> menostatf<-factor(meno)
> is.factor(menostatf)
[1] TRUE
> tgradef<-factor(tgrad)
> is.factor(tgradef)
[1] TRUE
>mod<-
survreg(Surv(time,cens)~age+tsize+pnodes+progrec+estrec+hormonef+menostatf+tgradef,
data=cdat,dist="weibull")
> mod
Call:
survreg(formula = Surv(time, cens) ~ age + tsize + pnodes + progrec +
    estrec + hormonef + menostatf + tgradef, data = cdat, dist = "weibull")
Coefficients:
  (Intercept)           age         tsize        pnodes       progrec
 8.0152203982  0.0068187581 -0.0057649918 -0.0379793758  0.0016434599
       estrec     hormonef1    menostatf1      tgradef2     tgradef3
-0.0001786928  0.2683582844 -0.1948559742 -0.4719768755 -0.5826634480
Scale = 0.7192839
Loglik(model) = -2579.7   Loglik(intercept only) = -2637.3
       Chisq = 115.16 on 9 degrees of freedom, p = < 2e-16
n = 686
> summary(mod)
Call:
survreg(formula = Surv(time, cens) ~ age + tsize + pnodes + progrec +
    estrec + hormonef + menostatf + tgradef, data = cdat, dist = "weibull")
                Value     Std. Error  z        p
(Intercept)  8.015220   0.350300 22.88   < 2e-16
age          0.006819   0.006644  1.03    0.3048
tsize       -0.005765   0.002817 -2.05    0.0407
pnodes      -0.037979   0.005383 -7.06    1.7e-12
progrec      0.001643   0.000417  3.94    8.2e-05
estrec      -0.000179   0.000324 -0.55    0.5807
hormonef1    0.268358   0.092826  2.89    0.0038
menostatf1 -0.194856   0.131185 -1.49    0.1375
tgradef2    -0.471977    0.180262 -2.62    0.0088
tgradef3    -0.582663   0.193759 -3.01    0.0026
Log(scale)  -0.329499   0.048892 -6.74    1.6e-11
Scale = 0.719
Weibull distribution
Loglik(model) = -2579.7   Loglik(intercept only) = -2637.3
```

```
      Chisq = 115.16 on 9 degrees of freedom, p = 1.3e-20
Number of Newton-Raphson Iterations: 7
n = 686
> step(mod, direction="backward", test="Chisq")
Start:  AIC = 5181.39
Surv(time, cens) ~ age + tsize + pnodes + progrec + estrec +
    hormonef + menostatf + tgradef
                Df    AIC     LRT    Pr(>Chi)
<none>              5181.4
- estrec         1   5179.7   0.295   0.587160
- age            1   5180.4   1.052   0.304935
- menostatf      1   5181.6   2.208   0.137255
- tsize          1   5183.4   3.991   0.045755 *
- tgradef        2   5188.1  10.689   0.004775 **
- hormonef       1   5188.1   8.706   0.003172 **
- progrec        1   5200.4  21.026  4.531e-06 ***
- pnodes         1   5216.1  36.686  1.388e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Step:  AIC = 5179.68
Surv(time, cens) ~ age + tsize + pnodes + progrec + hormonef +
    menostatf + tgradef
                Df    AIC     LRT      Pr(>Chi)
<none>              5179.7
- age            1   5178.6   0.892    0.344813
- menostatf      1   5179.9   2.235    0.134935
- tsize          1   5181.5   3.851    0.049716 *
- hormonef       1   5186.3   8.573    0.003412 **
- tgradef        2   5186.3  10.620    0.004941 **
- progrec        1   5199.5  21.832    2.976e-06 ***
- pnodes         1   5214.3  36.641    1.420e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Step:  AIC = 5178.58
Surv(time, cens) ~ tsize + pnodes + progrec + hormonef + menostatf +
    tgradef
                Df    AIC     LRT   Pr(>Chi)
<none>              5178.6
- menostatf      1  5178.0   1.418   0.233725
- tsize          1  5180.4   3.798   0.051304 .
- tgradef        2  5185.4  10.837   0.004435 **
- hormonef       1  5185.6   9.054   0.002621 **
- progrec        1  5199.4  22.854  1.748e-06 ***
- pnodes         1  5213.1  36.500  1.526e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Step:  AIC = 5177.99
Surv(time, cens) ~ tsize + pnodes + progrec + hormonef + tgradef
                Df    AIC    LRT   Pr(>Chi)
<none>              5178.0
- tsize          1   5179.6  3.628   0.056821 .
- hormonef       1   5183.9  7.955   0.004797 **
- tgradef        2   5184.9 10.888   0.004323 **
- progrec        1   5199.2 23.252  1.421e-06 ***
- pnodes         1   5212.9 36.922  1.229e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Call:
survreg(formula = Surv(time, cens) ~ tsize + pnodes + progrec +
    hormonef + tgradef, data = cdat, dist = "weibull")
```

```
Coefficients:
 (Intercept)       tsize       pnodes      progrec    hormonef1     tgradef2
 8.249658908 -0.005434314 -0.038198779  0.001636584  0.250590000 -0.477478381
    tgradef3
-0.587401338
Scale = 0.7201373
Loglik(model) = -2581   Loglik(intercept only) = -2637.3
       Chisq = 112.56 on 6 degrees of freedom, p = < 2e-16
n = 686
Call:
survreg(formula = Surv(time, cens) ~ tsize + hormonef + tgradef +
    pnodes + progrec, data = cdat, dist = "weibull")
                  Value     Std. Error    z         p
(Intercept)    8.249659   0.198482 41.56 < 2e-16
tsize         -0.005434   0.002784 -1.95  0.0510
hormonef1      0.250590   0.090707  2.76  0.0057
tgradef2      -0.477478   0.180369 -2.65  0.0081
tgradef3      -0.587401   0.193933 -3.03  0.0025
pnodes        -0.038199   0.005391 -7.09 1.4e-12
progrec        0.001637   0.000404  4.05 5.1e-05
```

Πίνακας 21 Αποτελέσματα R για την προσαρμογή της κατανομής Weibull

Π.3 Αποτελέσματα R για την προσαρμογή του μοντέλου του Cox

Προσαρμογή του Μοντέλου του Cox

```
> rm(list=ls())
> library(splines)
> library(survival)
> library(lattice)
> cdat<-read.table("C:/breastcan.txt",header=TRUE)
> attach(cdat)
> cdat
     id   age  tsize pnodes progrec estrec hormone meno tgrad time  cens
1     1    70    21     3       48     66       0       1     2  1814    1
2     2    56    12     7       61     77       1       1     2  2018    1
3     3    58    35     9       52    271       1       1     2   712    1
…
684 684  51    25     5       43      0       0       0     3   769    1
685 685  52    23     3       15     34       0       1     2   727    1
686 686  55    23     9      116     15       0       1     2  1701    1
> hormonef<-factor(hormone)
> is.factor(hormonef)
[1] TRUE
> menostatf<-factor(meno)
> is.factor(menostatf)
[1] TRUE
> tgradef<-factor(tgrad)
> is.factor(tgradef)
[1] TRUE
> mod1 <-coxph(Surv(time,cens)~age+tsize+pnodes+progrec +estrec+ hormonef+ menostatf+
tgradef)
> mod1
Call:
coxph(formula = Surv(time, cens) ~ age + tsize + pnodes + progrec +
    estrec + hormonef + menostatf + tgradef)
                   coef      exp(coef)     se(coef)     z          p
age          -0.0094592  0.9905854  0.0093006 -1.017  0.309126
tsize         0.0077961  1.0078266  0.0039390  1.979  0.047794
pnodes        0.0487886  1.0499984  0.0074471  6.551  5.7e-11
progrec      -0.0022172  0.9977852  0.0005735 -3.866  0.000111
estrec        0.0001973  1.0001973  0.0004504  0.438  0.661307
hormonef1    -0.3462784  0.7073155  0.1290747 -2.683  0.007301
menostatf1    0.2584448  1.2949147  0.1834765  1.409  0.158954
tgradef2      0.6361117  1.8891211  0.2492025  2.553  0.010693
tgradef3      0.7796542  2.1807181  0.2684801  2.904  0.003685
Likelihood ratio test = 104.8  on 9 df, p =<  2.2e-16
n = 686, number of events = 299
> summary(mod1)
Call:
coxph(formula = Surv(time, cens) ~ age + tsize + pnodes + progrec +
    estrec + hormonef + menostatf + tgradef)
  n = 686, number of events = 299
                   coef      exp(coef)     se(coef)     z        Pr(>|z|)
age          -0.0094592  0.9905854  0.0093006 -1.017  0.309126
tsize         0.0077961  1.0078266  0.0039390  1.979  0.047794 *
```

```
pnodes       0.0487886   1.0499984  0.0074471   6.551  5.7e-11 ***
progrec     -0.0022172   0.9977852  0.0005735 -3.866  0.000111 ***
estrec       0.0001973   1.0001973  0.0004504  0.438   0.661307
hormonef1   -0.3462784   0.7073155  0.1290747  -2.683  0.007301 **
menostatf1   0.2584448   1.2949147  0.1834765  1.409   0.158954
tgradef2     0.6361117   1.8891211  0.2492025  2.553   0.010693 *
tgradef3     0.7796542   2.1807181  0.2684801  2.904   0.003685 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
             exp(coef) exp(-coef) lower .95 upper .95
age          0.9906     1.0095    0.9727    1.0088
tsize        1.0078     0.9922    1.0001    1.0156
pnodes       1.0500     0.9524    1.0348    1.0654
progrec      0.9978     1.0022    0.9967    0.9989
estrec       1.0002     0.9998    0.9993    1.0011
hormonef10.7073     1.4138    0.5492    0.9109
menostatf11.2949     0.7723    0.9038    1.8553
tgradef2     1.8891     0.5293    1.1591    3.0788
tgradef3     2.1807     0.4586    1.2885    3.6909
Concordance = 0.692  (se = 0.015 )
Likelihood ratio test = 104.8  on 9 df,   p =< 2e-16
Wald test            = 114.8  on 9 df,   p =< 2e-16
Score (logrank) test = 120.7  on 9 df,   p =< 2e-16
> mod2<-step(mod1, direction="backward", test="Chisq")
Start:  AIC = 3489.46
Surv(time, cens) ~ age + tsize + pnodes + progrec + estrec +
    hormonef + menostatf + tgradef
               Df   AIC       LRT      Pr(>Chi)
<none>              3489.5
- estrec        1   3487.7     0.187    0.665503
- age           1   3488.5     1.031    0.309934
- menostatf     1   3489.4     1.977    0.159741
- tsize         1   3491.2     3.729    0.053473 .
- hormonef      1   3494.9     7.436    0.006392 **
- tgradef       2   3495.3     9.855    0.007246 **
- progrec       1   3507.5    20.005   7.724e-06 ***
- pnodes        1   3519.5    31.987   1.552e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Step:  AIC = 3487.65
Surv(time, cens) ~ age + tsize + pnodes + progrec + hormonef +
    menostatf + tgradef
               Df    AIC      LRT     Pr(>Chi)
<none>              3487.7
- age           1   3486.6    0.911   0.339981
- menostatf     1   3487.7    1.999   0.157389
- tsize         1   3489.3    3.627   0.056844 .
- hormonef      1   3493.0    7.335   0.006761 **
- tgradef       2   3493.5    9.812   0.007402 **
- progrec       1   3506.8   21.120   4.314e-06 ***
- pnodes        1   3517.6   31.943   1.588e-08 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Step:  AIC = 3486.56
Surv(time, cens) ~ tsize + pnodes + progrec + hormonef + menostatf +
    tgradef
               Df    AIC     LRT     Pr(>Chi)
<none>           3486.6
- menostatf  1   3485.7   1.119   0.290143
- tsize          1   3488.1   3.538   0.059961 .
- hormonef   1   3492.3   7.763   0.005333 **
- tgradef       2   3492.6  10.061   0.006534 **
- progrec       1   3506.6  22.086   2.607e-06 ***
- pnodes        1   3516.4  31.822   1.690e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Step:  AIC = 3485.68
Surv(time, cens) ~ tsize + pnodes + progrec + hormonef + tgradef
               Df    AIC     LRT      Pr(>Chi)
<none>           3485.7
- tsize          1   3487.0   3.364     0.066621 .
- hormonef   1   3490.5   6.847     0.008878 **
- tgradef       2   3491.8  10.121     0.006343 **
- progrec       1   3506.1  22.446   2.162e-06 ***
- pnodes        1   3515.9   32.185  1.402e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> mod2<-coxph(Surv(time,cens)~ tsize + hormonef+ tgradef+pnodes+progrec, ties=c("breslow"))
> mod2
Call:
coxph(formula = Surv(time, cens) ~ tsize + hormonef + tgradef +
    pnodes + progrec, ties = c("breslow"))
                    coef       exp(coef)     se(coef)      z          p
tsize          0.0073151   1.0073420  0.0038898  1.881   0.06003
hormonef1 -0.3235059   0.7236077  0.1258226 -2.571  0.01014
tgradef2       0.6438880  1.9038688  0.2490175  2.586   0.00972
tgradef3       0.7879212  2.1988208  0.2682588  2.937   0.00331
pnodes         0.0489911  1.0502110  0.0074538  6.573   4.94e-11
progrec       -0.0022168  0.9977856  0.0005538 -4.003   6.25e-05
Likelihood ratio test = 102.5  on 6 df, p =< 2.2e-16
n = 686, number of events = 299
> summary(mod2)
Call:
coxph(formula = Surv(time, cens) ~ tsize + hormonef + tgradef +
    pnodes + progrec, ties = c("breslow"))
  n = 686, number of events = 299
                    coef       exp(coef)     se(coef)      z       Pr(>|z|)
tsize          0.0073151  1.0073420   0.0038898   1.881    0.06003 .
hormonef1 -0.3235059   0.7236077  0.1258226   -2.571    0.01014 *
tgradef2       0.6438880   1.9038688  0.2490175   2.586    0.00972 **
tgradef3       0.7879212   2.1988208  0.2682588   2.937    0.00331 **
pnodes         0.0489911   1.0502110   0.0074538   6.573    4.94e-11 ***
progrec       -0.0022168   0.9977856  0.0005538  -4.003    6.25e-05 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
              exp(coef)  exp(-coef) lower .95 upper .95
tsize          1.0073     0.9927    0.9997    1.0151
hormonef1   0.7236     1.3820    0.5655    0.9260
tgradef2       1.9039     0.5252    1.1686    3.1017
tgradef3       2.1988     0.4548    1.2997    3.7199
pnodes         1.0502     0.9522    1.0350    1.0657
progrec        0.9978     1.0022    0.9967    0.9989
Concordance = 0.689  (se = 0.015 )
Likelihood ratio test = 102.5  on 6 df,   p =< 2e-16
Wald test           = 112.1  on 6 df,   p =< 2e-16
Score (logrank) test = 118.5  on 6 df,   p =< 2e-16
> sresid<-resid(mod2,type="scaledsch")
> sresid
              [,1]            [,2]          [,3]              [,4]             [,5]              [,6]
72   -0.0938357367 -1.757349   1.854930082  -0.052325548  0.0693720955 -8.127888e-03
98    0.0606855088 -1.295771   0.145699619   3.674933385  0.0787847483 -2.658110e-03
113   0.1752737879 -1.423661   0.300775277   3.571385032 -0.1056915606 -2.522315e-03
…
2286 -0.0522351291 -2.794684   3.666649660   4.016485251  0.1610959252 1.425355e-02
2372 -0.0121045544  1.716351   1.900181293   1.618269635  0.0506464909 -1.082527e-02
2456  0.0874146453 -2.741320   1.683306086   0.683123362 -0.0033560831 -9.388799e-03
```

Πίνακας 22 Αποτελέσματα R για την προσαρμογή του μοντέλου του Cox