Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΑΝΑΛΥΣΗ ΓΟΝΙΔΙΑΚΩΝ ΕΚΦΡΑΣΕΩΝ ΜΕΣΩ ΤΗΣ ΜΕΘΟΔΟΥ Real-Time RT PCR

Κλειώ – Μαρία Π. Νιάρου

Διπλωματική Εργασία

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

Πειραιάς
Φεβρουάριος 2011

U N I V E R S I T Y O F P I R A E U S

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN APPLIED STATISTICS

GENE EXPRESSION ANALYSIS THROUGH THE METHOD Real-Time RT PCR

By

Kleio - Maria P. Niarou

MSc Dissertation

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

Piraeus, Greece
February 2011

Στην οικογένειά μου

Ευχαριστίες

Στο σημείο αυτό θα ήθελα να εκφράσω τις ευχαριστίες μου για τους ανθρώπους που συνέβαλαν, με τον δικό του τρόπο ο καθένας, στη ολοκλήρωση της διπλωματικής αυτής εργασίας. Πρωτίστως, ευχαριστώ θερμά την κα Μαρία Κατέρη, Αναπληρώτρια Καθηγήτρια του τμήματος Μαθηματικών του Πανεπιστημίου Ιωαννίνων, για την συνεχή και αμέριστη καθοδήγησή της κατά τη διάρκεια συγγραφής της διπλωματικής αυτής εργασίας. Επίσης, ευχαριστώ τους κύριους Μάρκο Κούτρα, Καθηγητή του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς, και Κωνσταντίνο Πολίτη, Αναπληρωτή Καθηγητή του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς, οι οποίοι συμπλήρωσαν την τριμελή εξεταστική επιτροπή.

Τις θερμές μου ευχαριστίες θα ήθελα να εκφράσω προς την Ελληνική Συνεργαζόμενη Ογκολογική Ομάδα, για την παραχώρηση των δεδομένων που αναλύονται στην παρούσα εργασία, και ιδιαίτερα στην Αναστασία Ελευθεράκη για τον χρόνο και την προθυμία της να με βοηθήσει όσον αφορά τον ιδιαίτερο χειρισμό των δεδομένων γονιδιακών εκφράσεων, καθώς και για τις πολύτιμες παρατηρήσεις της. Ακόμα, ευχαριστώ τον φίλο μου Γιώργο Εμμανουήλ για την βοήθειά του στην κατανόηση των βιολογικών θεμάτων που συνοδεύουν την φύση των δεδομένων που αναλύονται και τη φίλη και συμφοιτήτριά μου Λουκία Σπινέλη για τις πολύτιμες συμβουλές και την ηθική συμπαράστασή της.

Θα ήθελα να ευχαριστήσω την οικογένειά μου για την υπομονή και υποστήριξη που έχει δείξει καθ’ όλη τη διάρκεια των σπουδών μου. Τέλος, ευχαριστώ όλους τους φίλους και τους ανθρώπους που ήταν δίπλα μου, με στήριξαν ηθικά και έδειξαν κατανόηση έως την ολοκλήρωση της εργασίας μου.

Περίληψη

Η μέθοδος *real- time RT PCR* αποτελεί μία εργαστηριακή τεχνική σύμφωνα με την οποία επιτυγχάνεται η ποσοτικοποίηση δειγμάτων mRNA και η ενίσχυσή τους όταν είναι διαθέσιμη μικρή ποσότητα δείγματος. Η εφαρμογή της μεθόδου μπορεί να εφαρμοσθεί, μεταξύ άλλων, σε γονιδιακές εκφράσεις που εντοπίζονται σε καρκινικούς ιστούς. Στην παρούσα εργασία, αναλύονται γονιδιακές εκφράσεις, οι οποίες εντοπίστηκαν στον μαστό, και αναπτύσσονται, παράλληλα, οι στατιστικές τεχνικές ανάλυσης.

Για τον σκοπό αυτό, παρουσιάζουμε μεθόδους ομαδοποίησης των γονιδιακών εκφράσεων, ώστε να μειωθεί των πλήθος των υπό εξέταση μεταβλητών. Συγκεκριμένα, αναλύουμε την θεωρία της *Ανάλυσης κατά Συστάδες* και *κατά Παράγοντες*. Επιπλέον, κάνουμε μία σύντομη αναφορά στους *χάρτες θερμότητας*, οι οποίοι αποτελούν γραφική τεχνική ομαδοποίησης ως προς τις γονιδιακές εκφράσεις και τους ασθενείς, σε ένα κοινό γράφημα, και παρουσιάζονται *δείκτες ισχύος* του αποτελέσματος της ομαδοποίησης.

Για την περαιτέρω ανάλυση βασιζόμαστε στην *Ανάλυση Επιβίωσης*, σύμφωνα με την οποία εξετάζεται η εξέλιξη της υγείας των ασθενών στο χρόνο. Αρχικά παρουσιάζεται μία σύντομη επισκόπηση της θεωρίας, ώστε να γίνουν περισσότερο κατανοητές βασικές έννοιες της μεθόδου. Στην συνέχεια, εξετάζεται ο χρόνος εφαρμόζοντας μη παραμετρική συμπερασματολογία, μέσω των εκτιμητών *Kaplan-Meier* και *Nelson-Aalen*. Τέλος, παρουσιάζεται το *μοντέλο αναλογικού κινδύνου του Cox*, ως μία ημι-παραμετρική μέθοδος μοντελοποίησης των δεδομένων και εξετάζονται τα υπόλοιπα που προκύπτουν.

Η εφαρμογή της παραπάνω ανάλυσης επιβίωσης πραγματοποιείται στον χρόνο επιβίωσης των γυναικών και στον χρόνο ελεύθερο νόσου, δεδομένου ότι οι γυναίκες θεωρούνται υγιείς την στιγμή ένταξής τους έρευνα, αφού έχει προηγηθεί χειρουργική επέμβαση απομάκρυνσης του όγκου. Τέλος, συγκρίνονται τα αποτελέσματα των δύο παραπάνω χρόνων.

Abstract

The method of real-time RT PCR is a laboratory technique, which achieves to quantify mRNA samples and amplify them when a small amount of sample is available. The method can be applied, among others, in gene expressions detected in tumor tissues. In this study, gene expressions identified in breast are analyzed, while the statistical analysis techniques applied are reviewed.

To this end, methods for clustering gene expressions in order to reduce the number of variables under consideration are presented. Specifically, the theory of Cluster Analysis and Factor Analysis is shortly reviewed. In addition, a brief reference to the heat maps is made. Heat maps are graphical clustering techniques that group gene expressions and patients in a common graph, and present indicators of validity of the resulting clustering.

Further analysis is based on Survival Analysis, which examines the evolution of the patients' health over time. Initially, a brief overview of the theory is presented to better understand its basic concepts. Then, the time to survival is examined by applying non-parametric inference, through the Kaplan-Meier and Nelson-Aalen estimators. Finally, the Cox model of proportional hazards is presented, as a semi-parametric method for modeling data and the resulting residuals are examined.

The above mentioned survival analysis is applied on the survival time of women and on their disease-free survival time, given that they are healthy at the starting time of the survey. Starting time is considered the day of surgery for removing the tumor. Finally, the results of these two time periods are compared and discussed.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Κατάλογος Συντομογραφιών

CMF	cyclophosphamide, methotrexate, fluorouracil
PCR	polymerase chain reaction (αλυσιδωτή αντίδραση πολυμεράσης)
DNA	deoxyribonucleic acid (δεοξυριβονουκλεϊκό οξύ)
RNA	ribonucleic acid **(**ριβονουκλεϊκό οξύ)
mRNA	messenger RNA (αγγελιοφόρο RNA)
rt- PCR	reverse transcriptase-PCR (αντίστροφη μεταγραφάση PCR)
cDNA	complementary DNA (συμπληρωματικό DNA)
ΑΣΚ	Αναφορική Συνάρτηση Κινδύνου
LR	Likelihood Ratio
CI	Confidence Interval

Κ Ε Φ Α Λ Α Ι Ο 1

Εισαγωγή

1.1 Καρκίνος του μαστού

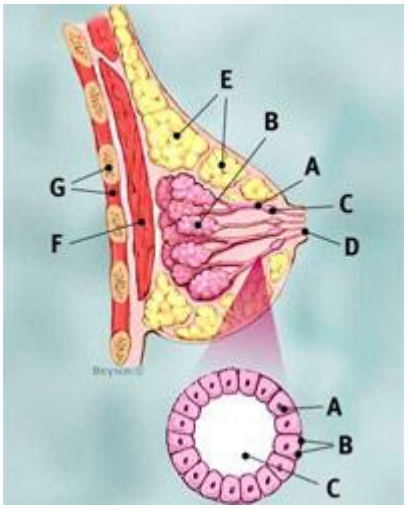
Ως καρκίνος του μαστού ορίζεται η μη ελεγχόμενη αύξηση των κυττάρων του μαστού. Για την καλύτερη κατανόησή του, είναι σημαντικό να δούμε τον τρόπο με τον οποίο αναπτύσσεται ο καρκίνος.

Ο καρκίνος δημιουργείται ως αποτέλεσμα μεταλλάξεων, ή ανώμαλων μεταβολών, των γονιδίων τα οποία είναι υπεύθυνα για την ρύθμιση του πολλαπλασιασμού των κυττάρων και την διατήρησή τους σε υγιή κατάσταση. Τα γονίδια βρίσκονται στον πυρήνα κάθε κυττάρου, ο οποίος αποτελεί τον 'χώρο ελέγχου' του κυττάρου. Υπό φυσιολογικές συνθήκες, τα κύτταρα του σώματός μας αντικαθίστανται μέσω μιας μεθοδικής διαδικασίας κυτταρικής ανάπτυξης. Ωστόσο, με την πάροδο του χρόνου, οι μεταλλάξεις είναι δυνατόν σε ένα κύτταρο να 'ενεργοποιήσουν' συγκεκριμένα γονίδια και να 'απενεργοποιήσουν' κάποια άλλα. Αυτό το μεταλλαγμένο κύτταρο έχει πλέον την ιδιότητα να συνεχίσει να πολλαπλασιάζεται χωρίς έλεγχο, παράγοντας περισσότερα όμοια κύτταρα τα οποία σχηματίζουν έναν όγκο.

Ο όγκος μπορεί να είναι καλοήθης (μη επιβλαβής στην υγεία) ή κακοήθης (έχει την ιδιότητα να είναι επικίνδυνος). Οι καλοήθεις όγκοι δεν θεωρούνται καρκινικοί. Η όψη τους πλησιάζει αυτή των φυσιολογικών κυττάρων, αναπτύσσονται αργά και δεν εισβάλουν σε κοντινούς ιστούς, ούτε εξαπλώνονται σε άλλα μέρη του σώματος. Οι κακοήθεις όγκοι είναι καρκινικοί. Σε περίπτωση που ο καρκίνος αφεθεί ανεξέλεγκτος, τα κακοήθη κύτταρα, τελικά, μπορούν να εξαπλωθούν πέραν του αρχικού όγκου σε άλλα μέρη του σώματος.

Ο όρος 'καρκίνος του μαστού' αναφέρεται σε ένα κακοήθη όγκο ο οποίος έχει αναπτυχθεί από κύτταρα του μαστού. Συνήθως, ο καρκίνος του μαστού εμφανίζεται αρχικά είτε στα κύτταρα των λοβίων (ductal carcinomas), τα οποία είναι οι αδένες παραγωγής γάλακτος, είτε των αγωγών (lobular carcinomas), οι οποίοι μεταφέρουν το γάλα από το λοβίο στη θηλή.

Σπανιότερα, καρκίνος του μαστού μπορεί να ξεκινήσει από τους στρωματικούς ιστούς, οι οποίοι περιλαμβάνουν τους λιπαρούς και ινώδεις συνδετικούς ιστούς του μαστού.



Προφίλ Μαστού:
A Γαλακτοφόροι πόροι
B Λοβία
C Γαλακτοφόροι κόλποι
D Θηλή
E Λίπος
F Μείζων θωρακικός μυς
G Θωρακικό τοίχωμα

Μεγένθυση
A Φυσιολογικά κύτταρα τωνπόρων
B Βασική μεμβράνη
C Αυλός των γαλακτοφόρων πόρων

Εικόνα 1.1: Ανατομία Μαστού (**Πηγή**: Ιστοσελίδα 15)

Με τον καιρό, τα καρκινικά κύτταρα μπορούν να εισβάλουν σε γειτονικούς υγιείς ιστούς (συνηθέστερα τους μασχαλιαίους λεμφαδένες, μικρά όργανα που λειτουργούν ως φίλτρα των ξένων ουσιών στο σώμα). Εάν τα καρκινικά κύτταρα μπουν στους λεμφαδένες, τότε έχουν διέξοδο προς άλλα μέρη του σώματος.

Ο καρκίνος του μαστού προκαλείται πάντα από μία γενετική ανωμαλία, ένα ‘λάθος’ του γενετικού υλικού. Ωστόσο, μόνο το 5-10% των περιπτώσεων καρκίνου οφείλεται σε ανωμαλία η οποία κληροδοτείται από τους γονείς. Περίπου το 90% του καρκίνων μαστού οφείλονται σε γενετικές ανωμαλίες οι οποίες παρουσιάζονται de novo ως αποτέλεσμα της διαδικασίας γήρανσης και της ‘φθοράς της ζωής’ γενικότερα (Ιστοσελίδα 11).

Η παρουσία καρκινογόνων μεταλλάξεων οδηγεί στην διαταραχή της φυσιολογίας των κυττάρων καθώς και στην διαφοροποίηση της μορφολογίας τους. Όταν τα κύτταρα διαφοροποιούνται, λαμβάνουν διάφορες μορφές και σχήματα στην προσπάθειά τους να λειτουργήσουν ως μέρος του οργανισμού. Βάσει της διαφοροποίησης αυτής των κυττάρων ορίστηκε ένα σύστημα που εφαρμόζεται ώστε να κατηγοριοποιηθούν τα καρκινικά κύτταρα ως προς το πόσο φυσιολογικά ή μη δείχνουν και πόσο γρήγορα ο όγκος είναι δυνατόν να αναπτυχθεί και να εξαπλωθεί. Τα καρκινικά κύτταρα, γενικά, ταξινομούνται ως καλώς διαφοροποιημένα (low grade-III), μέτρια διαφοροποιημένα (intermediate grade-II), χαμηλής

διαφοροποίησης (high grade-I) και μη διαφοροποιημένα (high grade - Undifferentiated). Οι καρκίνοι χαμηλής διαφοροποίησης έχουν την χειρότερη πρόγνωση (Ιστοσελίδα 12).

Η θεραπεία περιλαμβάνει χειρουργική επέμβαση, φαρμακευτική αγωγή (ορμονική θεραπεία και χημειοθεραπεία) και ακτινοβολία. Κάποιοι καρκίνοι του μαστού είναι ευαίσθητοι σε ορμόνες όπως τα οιστρογόνα και η προγεστερόνη, γεγονός που επιτρέπει την θεραπεία αναστέλλοντας την επίδραση των ορμονών αυτών στους αντίστοιχους ιστούς. Οι όγκοι που χαρακτηρίζονται από θετικούς υποδοχείς οιστρογόνων και προγεστερόνης έχουν καλύτερη πρόγνωση και απαιτούν λιγότερο επιθετική θεραπεία συγκριτικά με τους ορμονικά αρνητικούς καρκίνους.

Οι καρκίνοι του μαστού i) χωρίς ορμονικούς υποδοχείς, ii) που έχουν εξαπλωθεί στους λεμφαδένες της μασχάλης και iii) που εκφράζουν συγκεκριμένα γενετικά χαρακτηριστικά, είναι υψηλού κινδύνου και απαιτούν μία πιο επιθετική θεραπεία. Μία τυπική αγωγή, ευρέως διαδεδομένη στις ΗΠΑ, είναι η κυκλοφωσφαμίδη (cyclophosphamide) σε συνδυασμό με δοξορουβικίνη (doxorubicin – αδριαμυκίνη, Adriamycin). Τα φάρμακα αυτά καταστρέφουν το DNA στον καρκίνο, όπως επίσης και στα ταχέως αναπτυσσόμενα φυσιολογικά κύτταρα όπου προκαλούν σοβαρές παρενέργειες. Σε ορισμένες περιπτώσεις ο μιτωτικός αναστολέας *paclitaxel* προστίθεται. Μία ισοδύναμη θεραπεία, δημοφιλής στην Ευρώπη, είναι οι κυκλοφωσφαμίδη, μεθοτρεξάτη και φθοριοουρακίλη (cyclophosphamide, methotrexate, fluorouracil – CMF). Η ακτινοβολία συνήθως προστίθεται κατά την χειρουργική επέμβαση για τον έλεγχο των καρκινικών κυττάρων τα οποία χάθηκαν κατά την επέμβαση, και παρατείνει την επιβίωση, αν και η έκθεση της καρδιάς σε ακτινοβολία μπορεί να προκαλέσει βλάβη και καρδιακή ανεπάρκεια κατά τα επόμενα χρόνια.

Παγκοσμίως, ο καρκίνος του μαστού αποτελεί το 10,4% των κρουσμάτων καρκίνου για τις γυναίκες, αποτελώντας την πιο κοινή μορφή καρκίνου πέραν αυτής του δέρματος και την πέμπτη συνηθέστερη αιτία θανάτου από καρκίνο. Η εμφάνιση του καρκίνου του μαστού είναι περίπου 100 φορές πιο συχνή στις γυναίκες σε σχέση με τους άνδρες, αν και οι άνδρες τείνουν να έχουν χειρότερα αποτελέσματα εξαιτίας καθυστερήσεων στη διάγνωση (Ιστοσελίδα 1).

1.2 Ποσοτικοποίηση γονιδιακών εκφράσεων

Στη μοριακή βιολογία, η μέθοδος *αλυσιδωτής αντίδρασης πολυμεράσης* (polymerase chain reaction - PCR) είναι μία εργαστηριακή τεχνική η οποία εφαρμόζεται προκειμένου να ενισχύσει και ταυτόχρονα να ποσοτικοποιήσει ένα στοχοθετημένο μόριο DNA (Ιστοσελίδα 4). Η μέθοδος αυτή επιτρέπει την εκθετική ενίσχυση των στοχευμένων αλληλουχιών DNA (συνήθως 100 με 600 βάσεις) μέσα σε ένα μακρύτερο μόριο DNA διπλής έλικας. Σε σύγκριση με άλλες μεθόδους, η rt-PCR μπορεί να εφαρμοσθεί για την ποσοτικοποίηση επιπέδων του αγγελιοφόρου RNA (messenger RNA - mRNA) από πολύ μικρότερα δείγματα. Στην πραγματικότητα, η μέθοδος είναι αρκετά ευαίσθητη ώστε να επιτρέψει την ποσοτικοποίηση του RNA από ένα μοναδικό κύτταρο.

1.2.1 Η μέθοδος PCR

Η PCR απαιτεί την χρήση ενός ζεύγους εκκινητών (σύντομων τμημάτων μονόκλωνου DNA μήκους περίπου 20 νουκλεοτιδίων) το οποίο είναι συμπληρωματικό ως προς την υπό μελέτη αλληλουχία DNA, η οποία αναφέρεται ως αλληλουχία-στόχος και εντοπίζεται στις δύο συμπληρωματικές έλικες του DNA. Αυτοί οι εκκινητές επεκτείνονται από ένα ένζυμο (DNA πολυμεράση) έτσι ώστε να δημιουργηθεί ένα αντίγραφο από την καθορισμένη αλληλουχία. Μετά την δημιουργία αυτού του αντιγράφου, οι ίδιοι εκκινητές μπορούν να χρησιμοποιηθούν και πάλι, όχι μόνο για να δημιουργήσουν ένα επιπλέον αντίγραφο της εισερχόμενης έλικας DNA, αλλά και για το μικρό αντίγραφο που δημιουργήθηκε κατά τον πρώτο γύρο της σύνθεσης.

Εφόσον είναι απαραίτητη η αύξηση της θερμοκρασίας για τον διαχωρισμό των δύο ελίκων του DNA διπλής έλικας σε κάθε γύρο της διαδικασίας ενίσχυσης, ένα σημαντικό βήμα προόδου ήταν η ανακάλυψη μίας θερμο-σταθερής DNA πολυμεράσης (Taq polymerase), η οποία απομονώθηκε από το Thermus aquaticus, ένα βακτήριο που αναπτύσσεται σε θερμές πηγές. Ως αποτέλεσμα, δεν είναι απαραίτητο να προσθέτουμε νέα πολυμεράση σε κάθε γύρο ενίσχυσης. Μετά από αρκετούς γύρους ενίσχυσης, συνήθως σχεδόν 40, το προϊόν PCR αναλύεται σε ένα πήκτωμα αγαρόζης χρωματισμένο με βάμμα βρωμιούχου αιθιδίου. Το βρωμιούχο αιθίδιο είναι μία χρωστική ουσία η οποία συνδέεται με το DNA διπλής έλικας με παρεμβολή μεταξύ των ζευγών βάσης και κατά την σύνδεση του φωσφορίζει κατά την

έκθεση του σε υπεριώδη ακτινοβολία. Με αυτό τον τρόπο εντοπίζονται στο πύκτωμα μόνο οι αναγραφείσες δίκλωνες αλληλουχίες.

Η μέθοδος ανάλυσης αυτή είναι στην καλύτερη περίπτωση ημι-ποσοτική και σε πολλές περιπτώσεις, η ποσότητα του προϊόντος δεν σχετίζεται με την ποσότητα του αρχικά εισερχόμενου DNA, καθιστώντας το είδος αυτό PCR ένα ποιοτικό εργαλείο για την ανίχνευση της παρουσίας ή απουσίας μίας συγκεκριμένης αλληλουχίας DNA.

1.2.2 Αντίστροφη Μεταγραφάση PCR

Προκειμένου να μετρηθεί το mRNA, η μέθοδος επεκτείνεται χρησιμοποιώντας το ένζυμο αντίστροφη μεταγραφάση προκειμένου να μετατραπεί το mRNA σε συμπληρωματικό DNA (complementary DNA - cDNA), το οποίο στη συνέχεια πολλαπλασιάζεται με την μέθοδο PCR, και αναλύεται σε πήκτωμα αγαρόζης. Σε πολλές περιπτώσεις η μέθοδος αυτή έχει εφαρμοσθεί προκειμένου να μετρήσει τα επίπεδα ενός συγκεκριμένου mRNA υπό διαφορετικές συνθήκες. Ωστόσο η μέθοδος είναι στην πραγματικότητα εξίσου ημι-ποσοτική όπως η απλή PCR που εφαρμόζεται σε DNA. Η ανάλυση του mRNA μέσω της μεθόδου της αντίστροφης μεταγραφάσης (reverse transcriptase-PCR) συχνά συμβολίζεται ως ‘RT-PCR’, όμως ο συμβολισμός αυτός είναι ατυχές καθώς συχνά συγχέεται με τον όρο ‘real time PCR’.

Η παραπάνω διαδικασία σε μορφή βημάτων έχει ως εξής:

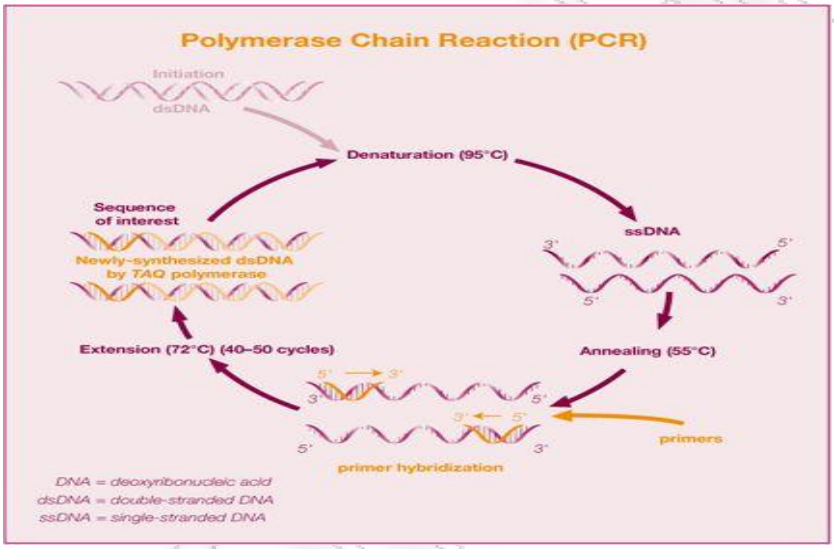
1. Το mRNA αντιγράφεται σε cDNA μέσω της αντίστροφης μεταγραφάσης. Το mRNA αφαιρείται επιτρέποντας στη δεύτερη έλικα του DNA να διαμορφωθεί. Τότε συνίσταται ένα μείγμα PCR το οποίο περιλαμβάνει μία ανθεκτική στη θερμότητα πολυμεράση, όπως η Taq polymerase, ειδικούς εκκινητές για το γονίδιο του ενδιαφέροντός μας, διοξυνουκλεοτίδια (deoxynucleotides) και ένα κατάλληλο ρυθμιστικό διάλυμα.

2. Το cDNA θερμαίνεται σε περισσότερους από 90 βαθμούς έτσι ώστε οι δύο συμπληρωματικές έλικες να διαχωρισθούν. Η θερμοκρασία κατόπιν μειώνεται στους 50 με 60 βαθμούς ώστε οι ειδικοί εκκινητές να υβριδοποιηθούν στις συμπληρωματικές τους αλληλουχίες. Τα υπό πολλαπλασιασμό τμήματα, όπως αυτά ορίζονται από τις θέσεις των εκκινητών, μπορούν να εκτείνονται σε μήκος μεγαλύτερο των 600 βάσεων, αν και συνήθως είναι περίπου 150-200.

3. Η θερμοκρασία ακολούθως αυξάνεται στους 72 βαθμούς και η ανθεκτική στη θερμότητα Taq DNA πολυμεράση επεκτείνει το DNA ξεκινώντας από τους εκκινητές. Τώρα

έχουμε τέσσερεις έλικες cDNA, δύο αντίγραφα και τις δύο αρχικές. Αυτές θερμαίνονται και πάλι στους 94 βαθμούς περίπου επαναλαμβάνοντας τα άνωθεν στάδια.

Η επαναλαμβανόμενη αυτή διαδικασία οδηγεί σε διπλασιασμό των αντιγράφων του cDNA σε κάθε κύκλο. Μετά από 30 με 40 γύρους σύνθεσης του cDNA, παράγονται πολλαπλά αντίγραφα του αρχικού στόχου DNA σε μία εκθετική βάση, τα οποία συνήθως αναλύονται σε πήκτωμα αγαρόζης χρωματισμένο με βρωμιούχο αιθίδιο (Ιστοσελίδα 6).



Εικόνα 1.2: Μέθοδος PCR (Πηγή: Ιστοσελίδα 17)

1.2.3 Η μέθοδος real-time RT-PCR

Η τυπική RT-PCR είναι μόνο ημι-ποσοτική, στην καλύτερη περίπτωση, λόγω των περιορισμών της μεθόδου κατά την ανάλυση αγαρόζης. Επομένως, η real-time PCR αναπτύχθηκε κυρίως λόγω:

1. της ανάγκης ποσοτικοποίησης των διαφορών στην έκφραση του mRNA

2. της διαθεσιμότητας μόνο μικρής ποσότητας mRNA σε κάποιες διαδικασίες όπως στη χρήση κυττάρων που λαμβάνονται μέσω λέιζερ μικρο-ανατομής, μικρής ποσότητας ιστού, πρωτογενών κυττάρων.

Σε αντίθεση με την πρακτική της rt-PCR και την ανάλυση με πηκτώματα αγαρόζης, η real-time PCR παρέχει ποσοτικά αποτελέσματα. Ένα επιπλέον πλεονέκτημα της μεθόδου είναι η

σχετική ευκολία και άνεση χρήσης της συγκριτικά με κάποιες άλλες μεθόδους, εφόσον υπάρχει πρόσβαση σε κατάλληλη συσκευή.

Κατά την real-time PCR το βήμα της ενίσχυσης συνδυάζεται με ανίχνευση. Πιο συγκεκριμένα, η αλληλουχία-στόχος απεικονίζεται σε πραγματικούς χρόνους ως σήμα φθορισμού. Το σήμα αυτό δημιουργείται σε κάθε κύκλο PCR μόνο όταν ένας συγκεκριμένος αισθητήρας ολιγονουκλεοτιδίου, συνδεδεμένος σε ένα fluorophore, υβριδοποιείται με τον συγκεκριμένο στόχο DNA. Καθώς η ποσότητα του ενισχυμένου DNA αυξάνεται με κάθε γύρο πολλαπλασιασμού, η ένταση του φθορισμού επίσης αυξάνεται. Κατά τη διάρκεια κάθε κύκλου PCR, η οπτική μονάδα μέτρησης του συστήματος πραγματικού χρόνου PCR μετρά το σήμα φθορισμού, και το λογισμικό κατασκευάζει ένα διάγραμμα της έντασης φθορισμού ως προς το πλήθος των κύκλων (Ιστοσελίδα 17).

Η real-time PCR είναι η πλέον προτεινόμενη μέθοδος για την επικύρωση αποτελεσμάτων από αναλύσεις μικρο-συστοιχιών (micro-array) και άλλες τεχνικές που αξιολογούν τις μεταβολές στις γονιδιακές εκφράσεις σε παγκόσμια κλίμακα (Ιστοσελίδα 9).

1.3 Προφίλ ασθενών

1.3.1 Σχεδιασμός της έρευνας

Σύμφωνα με την κλινική δοκιμή ΗΕ 10/97 που διεξήχθη από την Ελληνική Συνεργαζόμενη Ογκολογική Ομάδα (Hellenic Cooperative Oncology Group - HeCOG), γυναίκες οι οποίες εμφάνισαν αδενοκαρκίνωμα μαστού θεωρούνται απαλλαγμένες της νόσου κατόπιν χειρουργικής εκτομής του όγκου και ενδεχομένως να είναι υποψήφιες συστηματικής ανοσοενισχυτικής θεραπείας με στόχο την εξάλειψη μικρομεταστάσεων. Το mRNA εξήχθη από 367 ασθενείς με καρκίνωμα μαστού, ελεγχόμενο ως προς την φορμόλη και την παραφίνη, και αξιολογώντας ανοσοενισχυτική χημειοθεραπεία βασισμένη σε *epirubicin-alkylator*, με ή χωρίς *paclitaxel*. Η μέθοδος *real-time reverse transcription polymerase chain reaction* (real-time PT PCR) εφαρμόσθηκε προκειμένου να ποσοτικοποιηθούν οι γονιδιακές εκφράσεις. Οι γυναίκες που εξετάζονται θεωρούνται υγιείς την στιγμή έναρξης της έρευνας με την έννοια ότι υποβλήθηκαν σε χειρουργική επέμβαση για την απομάκρυνση του όγκου και ενδεχομένως να είναι υποψήφιες για επικουρική συστηματική θεραπεία στοχεύοντας στην εξάλειψη των μικρομεταστατικών εκθέσεων. Τα στατιστικά πακέτα που

χρησιμοποιήθηκαν για την ανάλυση είναι το *Statistical Package for Social Sciences* (SPSS 16.0) και η *R* (2.12.1).

1.3.2 Χαρακτηριστικά Ασθενών

Οι κλινικοπαθολογικοί παράγοντες των ασθενών που εξετάζονται στην παρούσα ανάλυση αφορούν στα ακόλουθα χαρακτηριστικά,

Group: δηλώνει την φαρμακευτική αγωγή που έλαβε κάθε γυναίκα. Υπάρχουν δύο εναλλακτικές αγωγές οι οποίες είναι Epirubicin – CMF (E-CMF) και Epirubicin – Taxol – CMF (E-T-CMF).

Age: δηλώνει την ηλικία της γυναίκας την στιγμή που εισάγεται στην έρευνα και είναι δίτιμη, με την τιμή 0 να αντιστοιχεί σε γυναίκες νεότερες των 50 ετών και την τιμή 1 διαφορετικά.

Menopausal: δηλώνει αν η γυναίκα βρίσκεται πριν (pre, 0) ή μετά (post, 1) την εμμηνόπαυση, την στιγμή εισαγωγής της στην έρευνα.

Grade: δηλώνει τον βαθμό διαφοροποίησης του όγκου με κατηγορίες I-II για καλή και μέτρια διαφοροποίηση (τιμή 0) και III- αδιαφοροποίητο για όγκο με χαμηλή ή καθόλου διαφοροποίηση (τιμή 1).

Size: δηλώνει το μέγεθος του όγκου που αφαιρέθηκε κατά την χειρουργική επέμβαση. Λαμβάνει την τιμή 1 αν ο όγκος είναι μικρότερος των 2cm, 2 αν το μέγεθός του είναι 2-5cm και 3 αν είναι μεγαλύτερος των 5cm.

Nodes: δηλώνει το πλήθος των θετικών μασχαλιαίων λεμφαδένων που αφαιρέθηκαν στο χειρουργείο. Κωδικοποιείται λαμβάνοντας την τιμή 0 όταν παρατηρούνται 0-3 θετικοί λεμφαδένας και την τιμή 1 αν σημειώνονται περισσότεροι των 4 θετικοί λεμφαδένες.

ER: δηλώνει εάν η πρωτεϊνική έκφραση των οιστρογονικών υποδοχέων είναι θετική (τιμή 1) ή αρνητική (τιμή 0).

PgR: δηλώνει εάν η πρωτεϊνική έκφραση των προγεστερονικών υποδοχέων είναι θετική (τιμή 1) ή αρνητική (τιμή 0).

HT: δηλώνει εάν η ασθενής έχει υποβληθεί σε ορμονοθεραπεία (τιμή 1) ή όχι (τιμή 0).

RT: δηλώνει εάν η ασθενής έχει υποβληθεί σε ραδιοθεραπεία (τιμή 1) ή όχι (τιμή 0)

Surgery: δηλώνει το είδος της επέμβασης, λαμβάνοντας την τιμή 0 για τα περιστατικά που υποβλήθηκαν σε τροποποιημένη ριζική μαστεκτομή (MRM) και την τιμή 1 για εκείνα που υποβλήθηκαν σε χειρουργική επέμβαση διατήρησης του μαστού (BCS).

Interval: χρόνος από την χειρουργική επέμβαση έως την εισαγωγή στην έρευνα με τιμές 1 για διάρκεια μικρότερη των 2 εβδομάδων, 2 για διάρκεια 2-4 εβδομάδων και 3 για διάρκεια μεγαλύτερη των 4 εβδομάδων.

Αξίζει να σημειωθεί ότι ο παράγοντας της θεραπευτικής αγωγής που χορηγήθηκε στις ασθενείς χρήζει ιδιαίτερης προσοχής. Στις ασθενείς χορηγήθηκαν είτε τέσσερεις κύκλοι epirubicin ακολουθούμενοι από τέσσερεις κύκλους εντατικοποιημένου συνδυασμού χημειοθεραπείας από cyclophosphamide, methotrexate και 5-fluorouracil (ομάδα E-CMF) είτε τρεις κύκλοι epirubicin ακολουθούμενοι από τρεις κύκλους paclitaxel και τρεις κύκλους εντατικοποιημένου CMF (ομάδα E-T-CMF).

Πέραν των παραπάνω παραγόντων, εξετάζονται παράλληλα 37 γονιδιακές εκφράσεις, οι ονομασίες των οποίων είναι:

PPIA	MLPH	ERBB2	CD3D	CXCL12	IL6ST	VEGFA	VEGFR2
UBE2c	TOP2A	TUBB3	MMP7	DHCR7	PTGER3	Herstatin	VEGFR3
MMP1	RACGAP1	MUC1	ABAT	EGFR	PVALB	VEGFB	
IGKC	CHPT1	ALCAM	AKR1C3	ERBB3	SFRP1	VEGFC	
TP53	CXCL13	SPP1	BIRC5	ERBB4	STC2	VEGFR1	

Στο Παράρτημα Α.1 δίνεται μία σύντομη περιγραφή των παραπάνω γονιδίων ως προς την ονομασία τους.

1.3.3 Περιγραφικά Στατιστικά

Στην παρούσα ενότητα παρουσιάζεται ο πίνακας συχνοτήτων των κλινικοπαθολογικών χαρακτηριστικών διαχωρίζοντάς τα ως προς τα επίπεδα του παράγοντα Group.

Παράγοντας		**E-T-CMF N (%)**	**E-CMF N (%)**	**Σύνολο N (%)**
ER(IHC)				
	Negative	30 (24,8)	45 (32,8)	75 (29,1)
	Positive	90 (74,4)	90 (65,7)	180 (69,8)
PGR(IHC)				
	Negative	40 (33,1)	53 (38,7)	93 (36)
	Positive	79 (65,3)	81 (59,1)	160 (62)
AGE				
	<50	60 (49,6)	67 (48,9)	127 (49,2)
	>=50	60 (49,6)	70 (51,1)	130 (50,4)
MENOPAUSAL				
	Pre	64 (52,9)	76 (55,5)	140 (54,3)
	Post	57 (47,1)	61 (44,5)	118 (45,7)
SURGERY				
	MRM	97 (80,2)	109 (79,6)	206 (79,8)
	BCS	24 (19,8)	28 (20,4)	52 (20,2)
INTERVAL				
	<2	14 (11,6)	19 (13,9)	33 (12,8)
	2-4	59 (48,8)	62 (45,3)	121 (46,9)
	>4	48 (39,7)	56 (40,9)	104 (40,3)
GRADE				
	I-II	51 (42,1)	79 (57,7)	130 (50,4)
	III-Undifer.	70 (57,9)	58 (42,3)	128 (49,6)
SIZE				
	<=2	34 (28,1)	41 (29,9)	75 (29,1)
	2-5	68 (56,2)	67 (48,9)	135 (52,3)
	>5	19 (15,7)	28 (20,4)	47 (18,2)
NODES				
	0-3	25 (20,7)	38 (27,7)	63 (24,4)
	>4	95 (78,5)	99 (72,3)	194 (75,2)
RT				
	No	18 (14,9)	27 (19,7)	45 (17,4)
	Yes	102 (84,3)	109 (79,6)	211 (81,8)
HT				
	No	8 (6,6)	12 (8,8)	20 (7,8)
	Yes	113 (93,4)	125 (91,2)	238 (92,2)

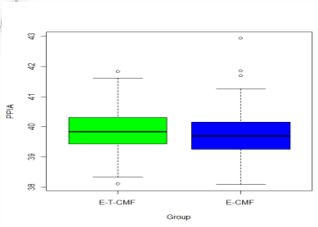
Πίνακας 1.1: Συχνότητες παραγόντων των ασθενών

Ως προς τις γονιδιακές εκφράσεις που σημειώθηκαν στις ασθενείς, παρουσιάζονται τα περιγραφικά χαρακτηριστικά των γονιδίων μέσω γραφικής αναπαράστασης θηκογράμματος (boxplot). Το είδος αυτό διαγραμμάτων έχει το πλεονέκτημα να παρουσιάζει συνοπτικά την κατανομή ποσοτικών μεταβλητών. Κάθε πλαίσιο περιλαμβάνει το $1^{ο}$ τεταρτημόριο, την διάμεσο και το $3^{ο}$ τεταρτημόριο, τα οποία ορίζουν την τιμή της μεταβλητής μέχρι την οποία σημειώνεται το 25%, 50% και 75% των παρατηρήσεων, αντίστοιχα. Οι απολήξεις δηλώνουν τα όρια των ακραίων παρατηρήσεων. Πέραν των σημείων αυτών, οι παρατηρήσεις θεωρούνται ότι λαμβάνουν ακραία τιμή και απεικονίζονται ως σημεία.

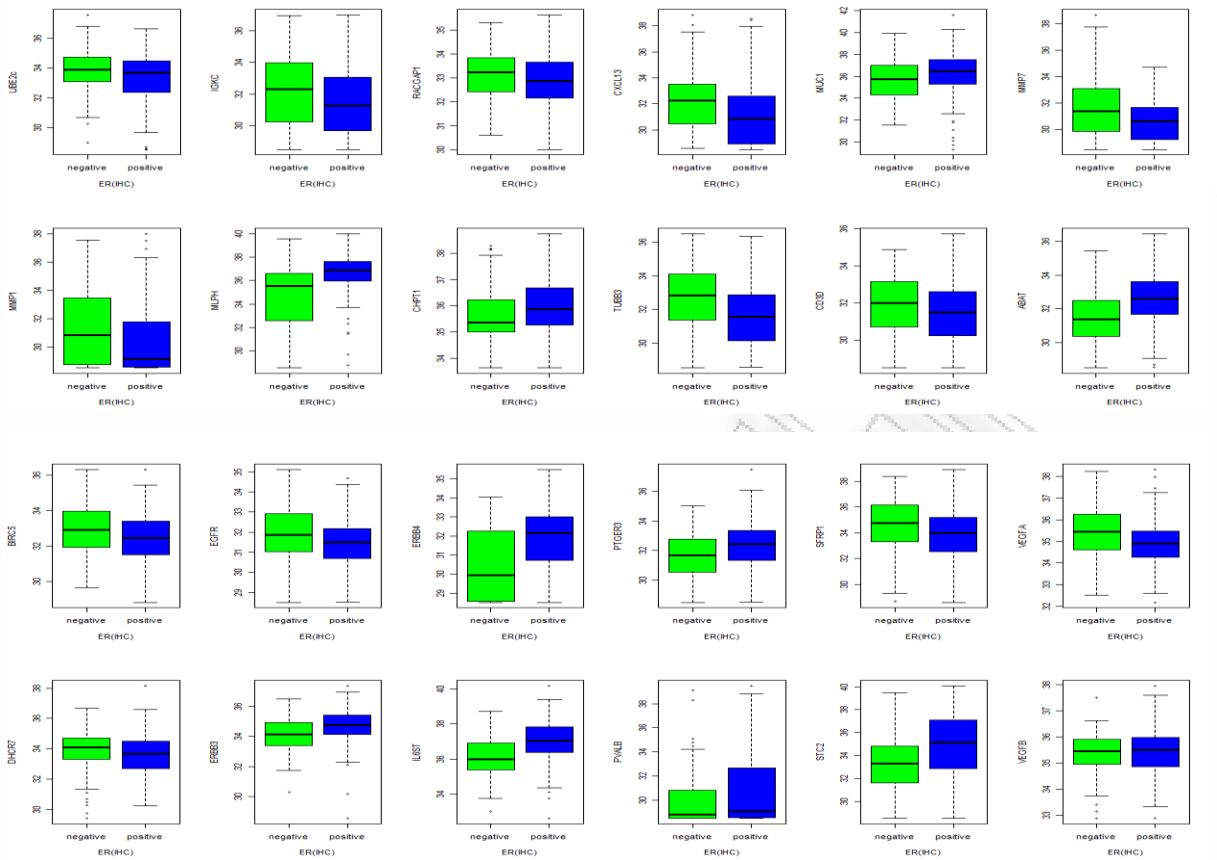
Μεγαλύτερο ερμηνευτικό ενδιαφέρον σημειώνεται για τα boxplots των γονιδίων, διαχωριζόμενα ως προς τα επίπεδα των παραγόντων, και μάλιστα όταν οι ομάδες που δημιουργούνται βάσει των παραγόντων ως προς το κάθε γονίδιο διαφέρουν σημαντικά. Για τον σκοπό αυτό εφαρμόζεται ο μη παραμετρικός έλεγχος *Mann-Whitney* για να ελεγχθεί στατιστικά, σε επίπεδο σημαντικότητας 5%, αν τα ανεξάρτητα δείγματα που δημιουργούνται από τις κατηγορίες των κλινικοπαθολογικών χαρακτηριστικών λαμβάνουν ίδιο επίπεδο τιμών.

Στην περίπτωση κατά την οποία ένας παράγοντας χαρακτηρίζεται από περισσότερες των δύο ομάδων, εφαρμόζεται ο έλεγχος *Kruskal-Wallis*, ο οποίος αποτελεί μία επέκταση του ελέγχου *Mann-Whitney* στην περίπτωση που εξετάζονται τρεις ή περισσότερες ομάδες. Εάν η υπόθεση της ισότητας των δειγμάτων απορριφθεί, κρίνεται σκόπιμη η παρουσίαση του αντίστοιχου boxplot.

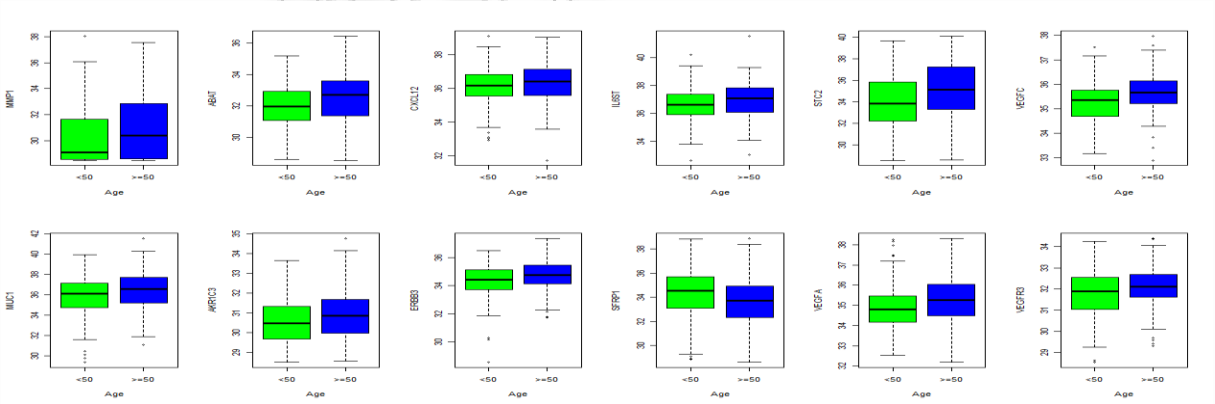
Στη συνέχεια παρουσιάζονται τα boxplots των γονιδίων για τις ομάδες των χαρακτηριστικών για τις οποίες εντοπίστηκε στατιστικά σημαντική διαφοροποίηση, σύμφωνα με το δείγμα των ασθενών που εξετάζεται.



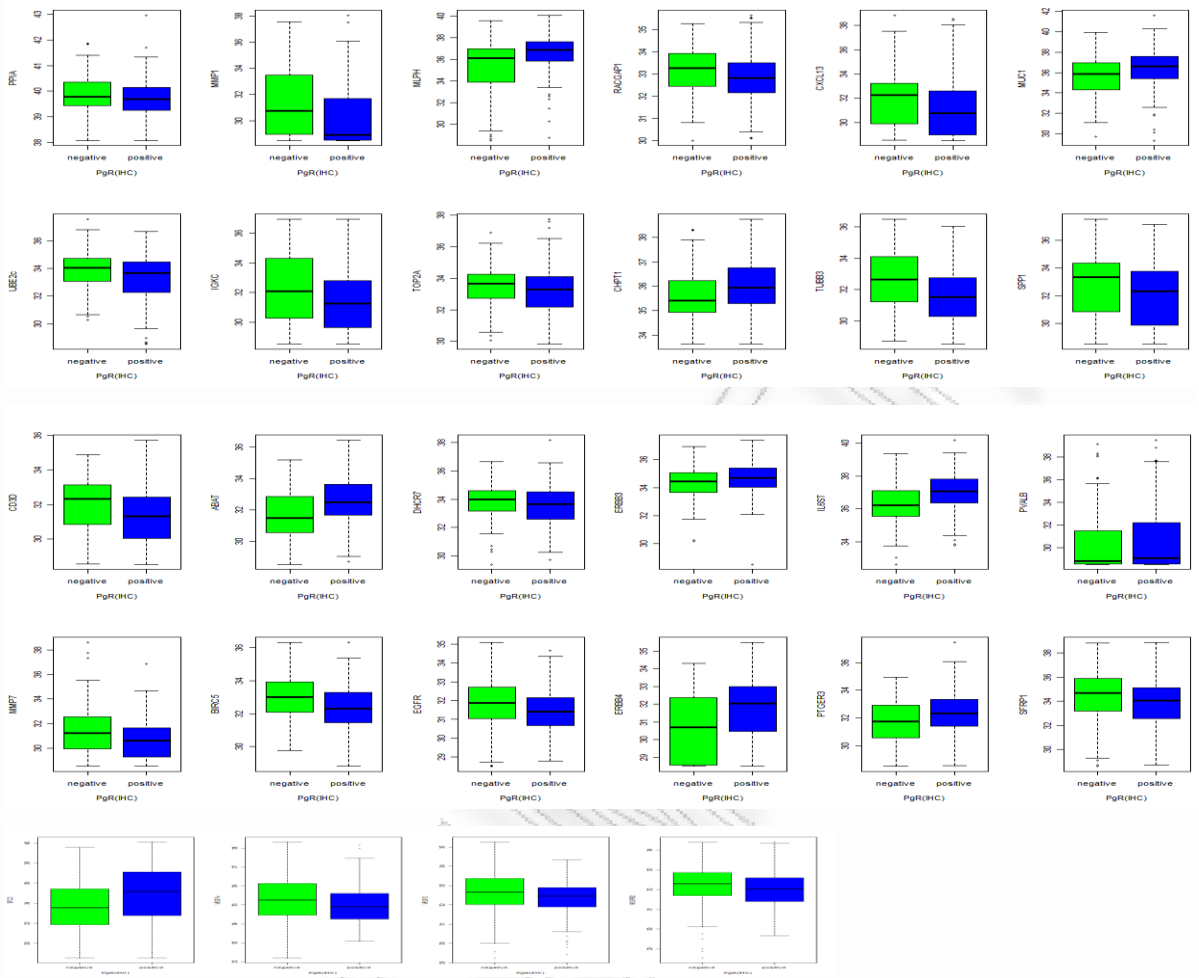
Σχήμα 1.1: Θηκόγραμμα για τις δύο ομάδες Group (αριστερά: E-T-CMF, δεξιά: E-CMF) για το γονίδιο PPIA.



Σχήμα 1.2: Θηκογράμματα για τις δύο ομάδες ER (αριστερά: Negative, δεξιά: Positive) για τα γονίδια UBE2c, RACGAP1, MUC1, BIRC5, ERBB4, SFRP1, MMP1, CHPT1, CD3D, DHCR7, IL6ST, STC2, IGKC, CXCL13, MMP7, EGFR, PTGER3, VEGFA, MLPH, TUBB3, ABAT, ERBB3, PVALB, VEGFB (ανά γραμμή).

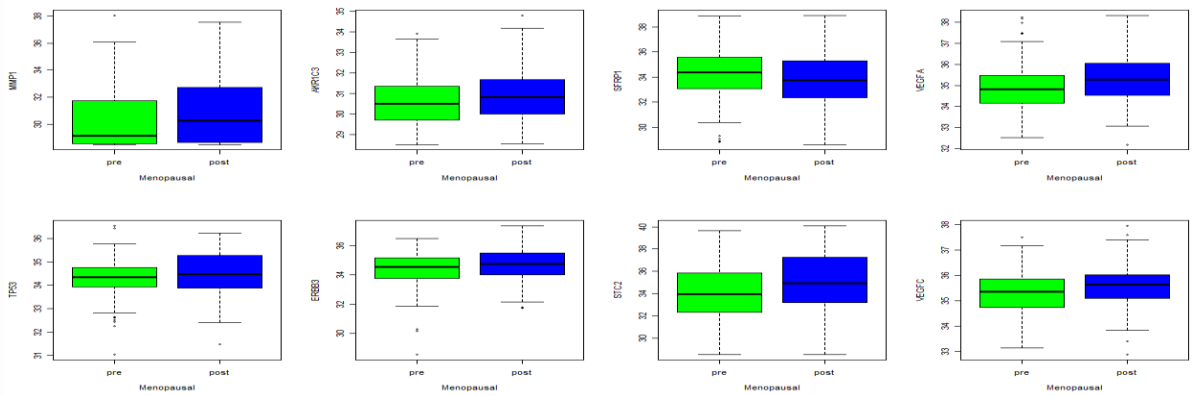


Σχήμα 1.3: Θηκογράμματα για τις δύο ομάδες Age (αριστερά: <50, δεξιά: >=50) για τα γονίδια MMP1, ACR1C3, IL6ST, VEGFA, MUC1, CXCL12, SFRP1, VEGFC, ABAT, ERBB3, STC2, VEGFR3 (ανά γραμμή).

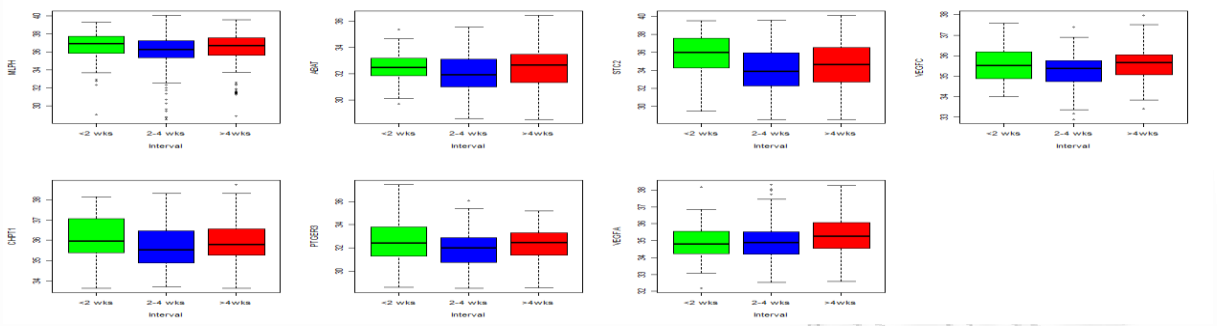


Σχήμα 1.4: Θηκογράμματα για τις δύο ομάδες PgR (αριστερά: Negative, δεξιά: Positive) για τα γονίδια PPIA, MLPH, CXCL13, CD3D, DHCR7, IL6ST, UBE2c, TOP2A, TUBB3, MMP7, EGFR, PTGER3, MMP1, RACGAP1, MUC1, ABAT, ERBB3, PVALB, IGKC, CHPT1, SPP1, BIRC5, ERBB4, SFRP1, STC2, VEGFA, VEGFC, VEGFR2 (ανά γραμμή).

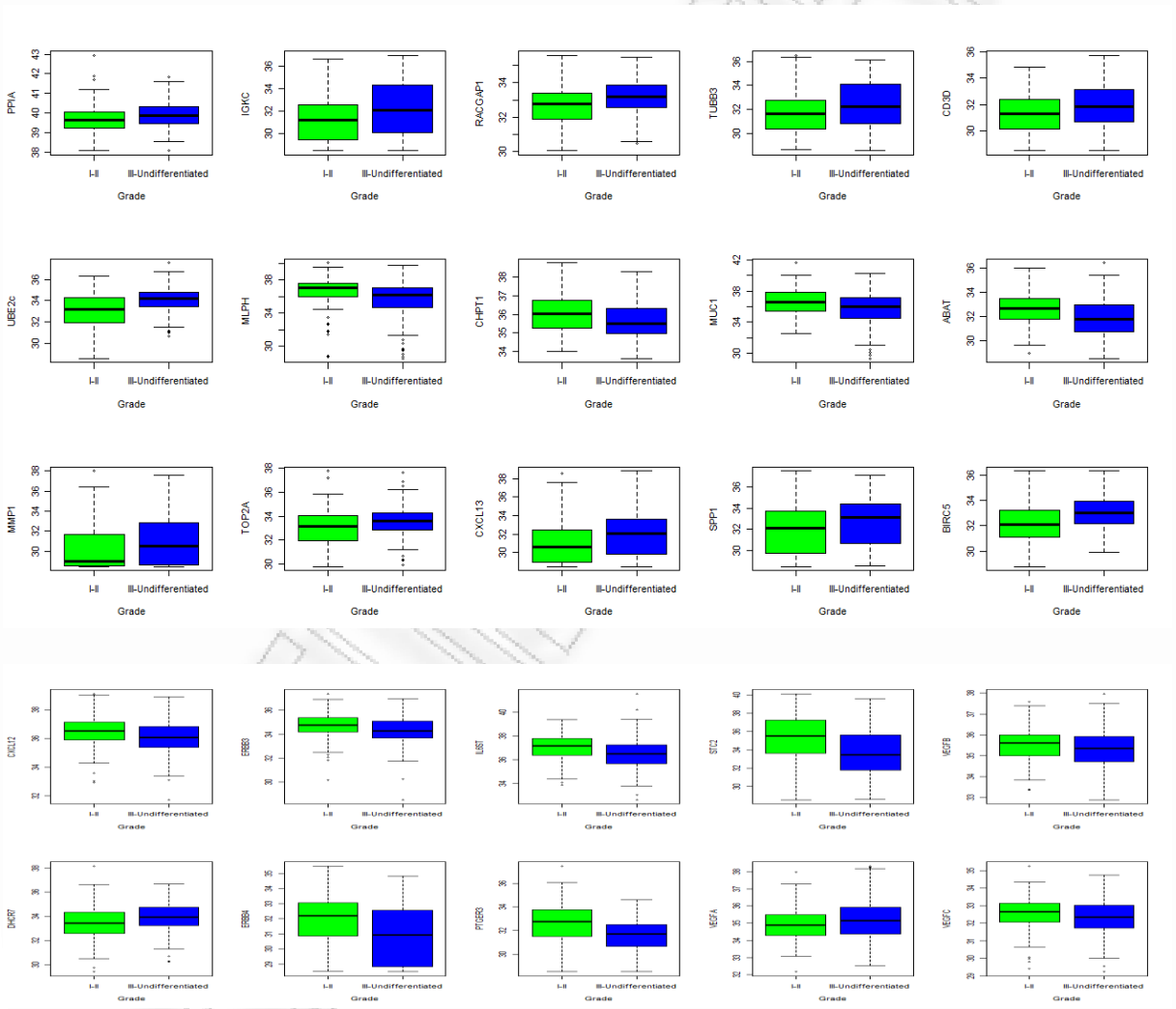
.



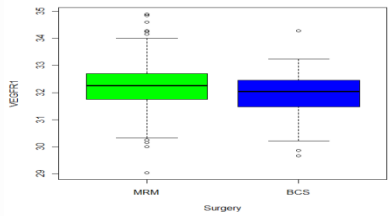
Σχήμα 1.5: Θηκογράμματα για τις δύο ομάδες Menopausal (αριστερά: pre, δεξιά: post) για τα γονίδια MMP1, ACR1C3, SFRP1, VEGFA, TP53, ERBB3, STC2, VEGFC (ανά γραμμή).



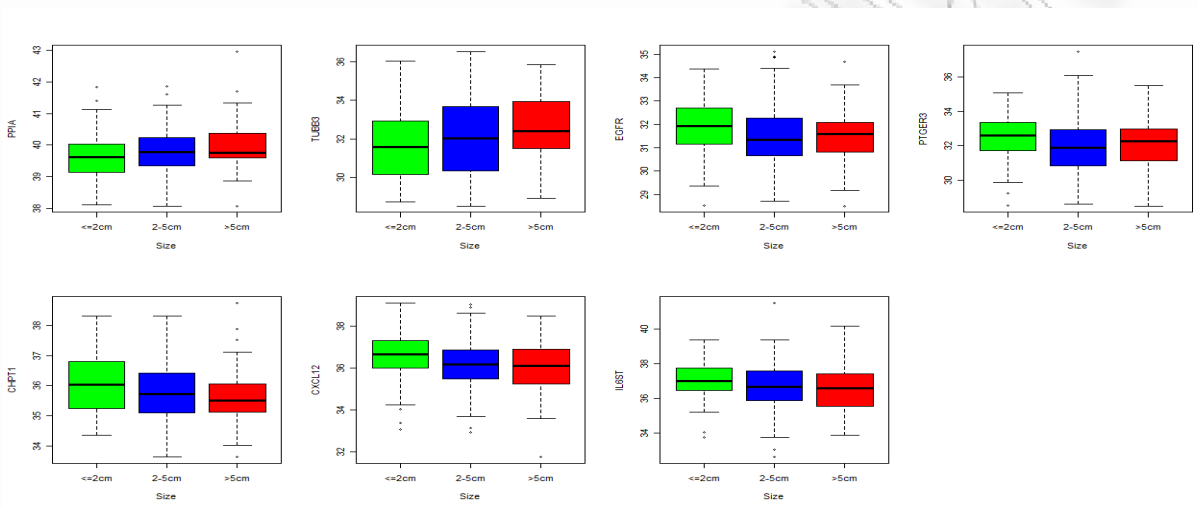
Σχήμα 1.6: Θηκογράμματα για τις τρεις ομάδες Interval (αριστερά: <2wks, κέντρο: 2-4wks, δεξιά: >4wks) για τα γονίδια MLPH, ABAT, STC2, VEGFC, CHRT1, PTGER3, VEGFA (ανά γραμμή).



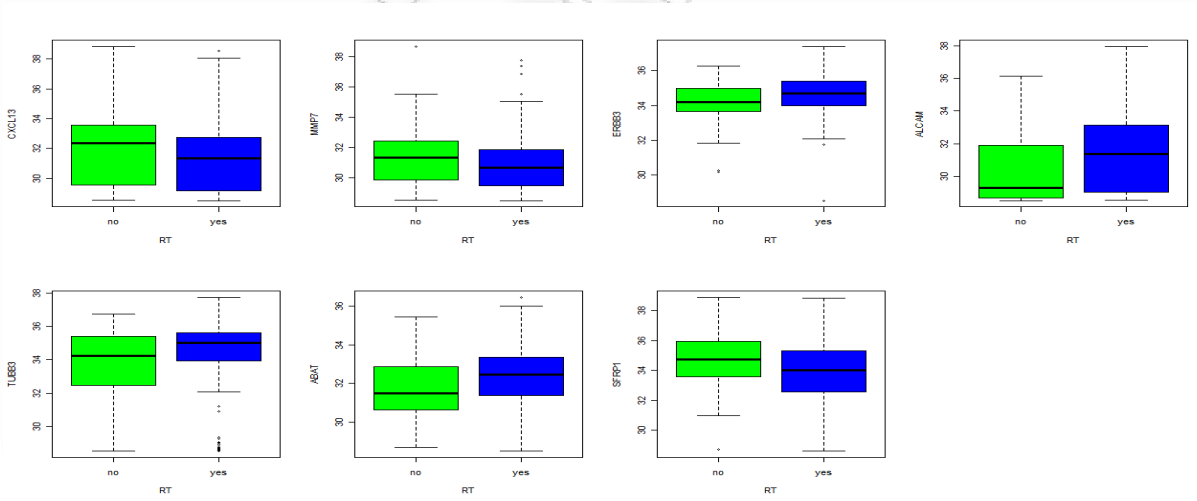
Σχήμα 1.7: Θηκογράμματα για τις δύο ομάδες Grade (αριστερά: I-II, δεξιά: III-αδιαφοροποίητο) για τα γονίδια PPIA, IGKC, RACGAP1, TUBB3, CD3D, UBE2c, MLPH, CHPT1, MUC1, ABAT, MMP1, TOP2A, CXCL13, SPP1, BIRC5, CXCL12, ERBB3, IL6ST, STC2, VEGFB, DHCR7, ERBB4, PTGER3, VEGFA, VEGFC (ανά γραμμή).



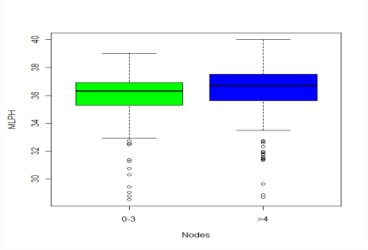
Σχήμα 1.8: Θηκόγραμμα για τις δύο ομάδες Surgery (αριστερά: MRM, δεξιά: BCS) για το γονίδιο VEGFR1.



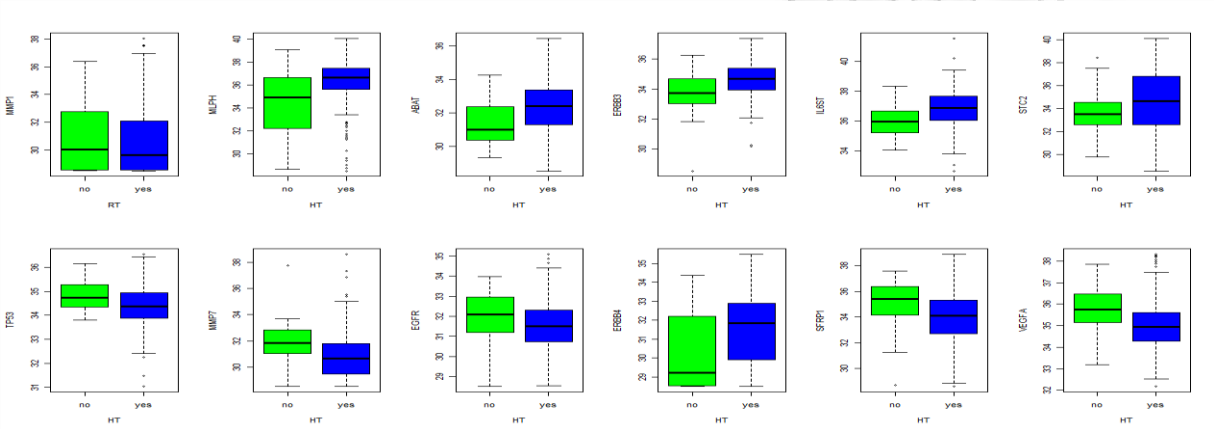
Σχήμα 1.9: Θηκογράμματα για τις τρεις ομάδες Size (αριστερά: <=2cm, κέντρο: 2-5cm, δεξιά: >5cm) για τα γονίδια PPIA, TUBB3, EGFR, PTGER3, CHPT1, CXCL12, IL6ST (ανά γραμμή).



Σχήμα 1.10: Θηκογράμματα για τις τρεις ομάδες RT (αριστερά: no, δεξιά: yes) για τα γονίδια CXCL13, MMP7, ERBB3, ALCAM, TUBB3, ABAT, SFRP1 (ανά γραμμή).



Σχήμα 1.11: Θηκόγραμμα για τις δύο ομάδες Nodes (αριστερά: 0-3, δεξιά: >4) για το γονίδιο MLPH.



Σχήμα 1.12: Θηκογράμματα για τις τρεις ομάδες HT (αριστερά: no, δεξιά: yes) για τα γονίδια MMP1, MMP7, ERBB3, SFRP1, TP53, ABAT, ERBB4, STC2, MLPH, EGFR, IL6ST, VEGFA (ανά γραμμή).

Κ Ε Φ Α Λ Α Ι Ο 2

Μέθοδοι Ομαδοποίησης Γονιδιακών Εκφράσεων

2.1 Ανάλυση κατά Συστάδες

Ο όρος *ανάλυση κατά συστάδες* ή *ομαδοποίηση δεδομένων* (χρησιμοποιήθηκε για πρώτη φορά από τον Tryon το 1939) περιλαμβάνει έναν αριθμό διαφορετικών αλγορίθμων και μεθόδων για τη συγκέντρωση αντικειμένων του ίδιου είδους σε αντίστοιχες κατηγορίες (Ιστοσελίδα 14). Το πρόβλημα συχνά αναφέρεται ως αναζήτηση των 'φυσικών ομάδων'. Πιο συγκεκριμένα, η ανάλυση κατά συστάδες είναι ένα διερευνητικό εργαλείο ανάλυσης δεδομένων το οποία αποσκοπεί στη κατάταξη διαφορετικών αντικειμένων σε ομάδες με τρόπο που ο βαθμός σύνδεσης μεταξύ δύο αντικειμένων να είναι ο μέγιστος, εφόσον ανήκουν στην ίδια ομάδα και ο ελάχιστος σε διαφορετική περίπτωση. Η τεχνική αυτή είναι πρωταρχική με την έννοια ότι δεν γίνεται καμία υπόθεση για το πλήθος ή την δομή των ομάδων (unsupervised method). Τα στοιχεία που εισάγονται είναι είτε μέτρα ομοιότητας ή απόστασης είτε δεδομένα από τα οποία μπορούν να υπολογισθούν οι ομοιότητες.

Προκειμένου να λάβει χώρα η ομαδοποίηση, είναι απαραίτητο να αναπτυχθεί κατάλληλη ποσοτική κλίμακα, σύμφωνα με την οποία αποφαίνεται αν δύο αντικείμενα (π.χ. άτομα, γονίδια κτλ) είναι όμοια ή ανόμοια μεταξύ τους. Ως τέτοιες ποσότητες θεωρούμε τις *αποστάσεις* οι οποίες λαμβάνουν μικρή τιμή αν δύο παρατηρήσεις είναι όμοιες. Σε αυτή την περίπτωση, οι παρατηρήσεις αυτές τοποθετούνται στην ίδια ομάδα. Επομένως, στόχος της μεθόδου είναι η δημιουργία ομάδων, με τις παρατηρήσεις μέσα σε κάθε ομάδα να έχουν μικρή απόσταση μεταξύ τους και τις παρατηρήσεις μεταξύ δύο διαφορετικών ομάδων να σημειώνουν μεγαλύτερη απόσταση.

Η μέθοδος αυτή είναι οικεία στους βιολόγους μέσω της εφαρμογής της στη φυλογενετική και την ανάλυση αλληλουχίας (sequence analysis). Οι σχέσεις μεταξύ αντικειμένων, όπως

γονιδίων, αναπαρίστανται μέσω ενός δενδρογράμματος του οποίου το μήκος των ‘κλαδιών’ αντικατοπτρίζει τον βαθμό ομοιότητας μεταξύ των αντικειμένων. Στην σύγκριση αλληλουχιών, οι μέθοδοι ανάλυσης κατά συστάδων εφαρμόζονται για συμπερασματολογία σχετικά με την εξελικτική ιστορία των υπό σύγκριση αλληλουχιών. Υπολογίζοντας τον βαθμό ομοιότητας μεταξύ των ομάδων σχετιζόμενων γονιδίων, η ανάλυση κατά συστάδες εφαρμόζεται προκειμένου να ταυτοποιήσει τις γονιδιακές εκφράσεις οι οποίες αντανακλούν την μοριακή ταυτότητα των ιστών από τους οποίους τα κύτταρα προέρχονται.

Ανάλογα με το αντικείμενο της μελέτης που ερευνάται, το ενδιαφέρον μπορεί να εστιάσει είτε στην αναζήτηση συστάδων γονιδίων με όμοιο πρότυπο έκφρασης στα αντιπροσωπευτικά δείγματα, είτε στην εύρεση συστάδων αντιπροσωπευτικών δειγμάτων που μοιράζονται όμοια πρότυπα έκφρασης στο σύνολο των γονιδίων.

2.1.1. Μέτρα Απόστασης - Ομοιότητας

Η ανάλυση συστάδων βασίζεται σε μέτρα απόστασης μεταξύ των αντικειμένων που πρόκειται να αποτελέσουν τις συστάδες, είτε αυτά είναι γονίδια είτε αντιπροσωπευτικά δείγματα ασθενών. Η γενική ιδέα είναι ότι τα αντικείμενα που ‘μοιράζονται’ πολλά χαρακτηριστικά θα κείτονται κοντά το ένα στο άλλο και ίσως ανήκουν στην ίδια ομάδα ή συστάδα. Σε αντίθεση, αντικείμενα που είναι ευρέως διαχωρισμένα ως προς τα χαρακτηριστικά τους είναι απίθανο να ανήκουν στην ίδια ομάδα.

Η ανάλυση κατά συστάδες μπορεί επίσης να βασισθεί σε μέτρα ομοιότητας. Ένα μέτρο ομοιότητας υποδηλώνει την δύναμη της σχέσης μεταξύ των δύο αντικειμένων και, υπό μία έννοια, είναι το αντίστροφο της απόστασης. Παρατηρήσεις που μοιάζουν πολύ μεταξύ τους δίνουν μεγάλη τιμή στο μέτρο ομοιότητας, ενώ ανόμοιες παρατηρήσεις δίνουν μικρότερες τιμές. Ως εκ τούτου, μεγαλύτερες αποστάσεις αντιστοιχούν σε μικρότερη ομοιότητα.

Προκειμένου να εφαρμοστεί η ανάλυση συστάδων σε γονίδια, τα δεδομένα μπορούν να οργανωθούν σε έναν πίνακα δεδομένων, στον οποίο κάθε γραμμή αντιστοιχεί σε ένα δείγμα (ασθενής) και κάθε στήλη σε ένα γονίδιο. Έστω ότι διαθέτουμε συνολικά N δείγματα και G γονίδια, οπότε ο πίνακας είναι διάστασης $N \times G$. Κάθε γονίδιο μπορεί να αντιπροσωπευθεί από ένα διάνυσμα-στήλη για τα N αντιπροσωπευτικά δείγματα, έστω

$$\mathbf{x}^{(g)} = (x_{g1},\ldots,x_{gN})'$$

το οποίο αντιπροσωπεύει επίπεδα έκφρασης για το γονίδιο g, δηλαδή το στοιχείο x_{gn} δηλώνει το επίπεδο έντασης του γονιδίου g που μετρήθηκε από το δείγμα n, για n=1,2,…,N. Το διάνυσμα $\bar{x}^{(g)}$ θα μπορούσε να περιγραφεί ως σημείο σε έναν N-διάστατο χώρο βιολογικών δειγμάτων. Αντίστοιχα, κάθε δείγμα μπορεί να αντιπροσωπευθεί από ένα διάνυσμα-γραμμή για τα G γονίδια, έστω

$$\mathbf{x}_n = (x_{1n},...,x_{gn},...,x_{Gn})$$

όπου το στοιχείο x_{gn} δηλώνει την ένταση του γονιδίου g που μετράται από το δείγμα n. Το διάνυσμα $\bar{x}_n$ θα μπορούσε να περιγραφεί ως ένα σημείο σε έναν G-διάστατο χώρο γονιδίων.

Οι αποστάσεις μεταξύ αντικειμένων (γονιδίων ή δειγμάτων) μπορούν να συνοψισθούν σε έναν *πίνακα αποστάσεων* **D** = (d_{ij}), όπου d_{ij} είναι η απόσταση μεταξύ των αντικειμένων *i* και *j*. Αντίστοιχα, ένας *πίνακας ομοιότητας* **S** = (s_{ij}) μπορεί να ορισθεί ως σύνοψη των ομοιοτήτων μεταξύ όλων των ζευγών αντικειμένων *i* και *j*. Είναι πάντα δυνατόν να κατασκευαστούν οι ομοιότητες από τις αποστάσεις. Για παράδειγμα, έστω $d_{g_1g_2}$ δηλώνει την απόσταση μεταξύ των γονιδίων g_1 και g_2, τότε η ομοιότητα των g_1 και g_2 μπορεί να ορισθεί από το μέτρο

$$s_{g1g2} = \frac{1}{1+d_{g1g2}}.$$

Στην συνέχεια αναφέρονται μέτρα απόστασης με ιδιαίτερη εφαρμογή στην ομαδοποίηση γονιδίων (Mei-Ling Ting Lee, 2004). Τα σχόλια που ακολουθούν αναφέρονται σε αποστάσεις μεταξύ γονιδίων, αλλά εφαρμόζονται με αντίστοιχη αλλαγή στον συμβολισμό για αποστάσεις μεταξύ δειγμάτων.

2.1.1.1 Μέτρα Απόστασης

Ένα μέτρο απόστασης συνήθως απαιτείται να ικανοποιεί ένα σύνολο λογικών ιδιοτήτων ή υποθέσεων ώστε να παράγει συνετά συμπεράσματα. Η απόσταση μεταξύ των σημείων P και Q, *d(P,Q)*, θεωρείται ως *μετρική* εάν ικανοποιεί τις ακόλουθες ιδιότητες:

(1) $d(P,Q) = d(Q,P)$ (συμμετρική ιδιότητα)

(2) $d(P,Q) > 0$ αν $P \neq Q$ και $d(P,Q) = 0$ αν $P = Q$

(3) $d(P,Q) \leq d(P,R) + d(R,Q)$, για κάθε άλλο σημείο R (τριγωνική ανισότητα)

Παρ' όλο που συνήθως προτείνεται να γίνεται χρήση μέτρων απόστασης που ικανοποιούν το παραπάνω σύνολο ιδιοτήτων, πολλοί αλγόριθμοι συστάδων αποδέχονται υποκειμενικώς προσδιορισμένα μέτρα τα οποία πιθανόν να μην ικανοποιούν όλες τις ιδιότητες απόστασης. Στην περίπτωση που το μέτρο απόστασης ικανοποιεί τις παραπάνω ιδιότητες εκτός της τριγωνικής, ονομάζεται *ημιμετρικό*. Αξίζει να επισημανθεί ότι η σχέση $d(P, Q) = 0$ δεν σημαίνει απαραίτητα ότι $P = Q$, αφού μπορεί επίσης να υποδεικνύει ότι δύο διαφορετικά αντικείμενα έχουν τις ίδιες μετρήσεις αναφορικά με την υπό μελέτη μεταβλητή.

Ευκλείδεια απόσταση (Euclidean distance) – απόσταση Pearson

$$d(\mathbf{x}^{(g_1)}, \mathbf{x}^{(g_2)}) = \left(\sum_{n=1}^{N} \left(x_{g_1 n} - x_{g_2 n} \right)^2 \right)^{1/2}$$

Η *Ευκλείδεια* απόσταση αποτελεί την πιο γνωστή και απλή απόσταση για συνεχή δεδομένα. Αξίζει να σημειώσουμε ότι η Ευκλείδεια απόσταση ικανοποιεί τις παραπάνω τρεις ιδιότητες. Το μήκος των δύο γονιδίων g_1 και g_2 προκύπτει από το μήκος της υποτείνουσας του ορθογωνίου τριγώνου που δημιουργείται από τις συντεταγμένες των σημείων αυτών, λαμβάνοντας υπόψη το Πυθαγόρειο θεώρημα. Προτιμάται για την ανάλυση κατά συστάδες αφού δεν απαιτείται προηγούμενη γνώση των διαχωριστικών ομάδων, χωρίς την οποία θα ήταν αδύνατος ο υπολογισμός ποσοτήτων όπως ο πίνακας διακυμάνσεων-συνδιακυμάνσεων. Ωστόσο, εξαρτάται από την κλίμακα μέτρησης, γεγονός που σημαίνει ότι η διάταξη των αποστάσεων δεν διατηρείται μετά από αλλαγή της κλίμακας. Επιπλέον, οι μεταβλητές με μεγάλες απόλυτες τιμές είναι αυτές που συνεισφέρουν περισσότερο συγκριτικά με μεταβλητές που έχουν μικρές απόλυτες αποστάσεις (βλ. Μ. Κούτρας, 2008). Ενδιαφέρον, επίσης, παρουσιάζει η συνολική μορφή των γονιδιακών προτύπων έκφρασης, σε σχέση με τα ατομικά μεγέθη του κάθε χαρακτηριστικού. Ένας τρόπος να παρακαμφθεί το πρόβλημα της κλίμακας μέτρησης είναι η τυποποίηση των αντικειμένων. Με αυτόν τον τρόπο, ωστόσο, χάνονται πιθανές συσχετίσεις μεταξύ τους. Αν συμβολίσουμε με $\bar{x}_g$ και s_g τον μέσο και την διακύμανση του g γονιδίου, η απόσταση έχει την ακόλουθη μορφή,

$$d(\mathbf{x}^{(g_1)}, \mathbf{x}^{(g_2)}) = \sqrt{\sum_{g=1}^{G} \frac{(x_{ig} - x_{jg})^2}{s_g^2}} = \sqrt{\sum_{g=1}^{G} \left(\frac{x_{ig} - x_{jg}}{s_g} \right)^2}$$

Σε αυτήν την περίπτωση η διακύμανση θεωρείτε ως βάρος. Αξίζει να σημειώσουμε ότι στην περίπτωση που τα γονίδια σημειώνουν κοινή διακύμανση, η Ευκλείδεια απόσταση θεωρείτε καταλληλότερη αφού όλοι οι όροι της παραπάνω σχέσης έχουν το ίδιο βάρος.

City-Block ή Manhattan απόσταση

$$d(\boldsymbol{x}^{(g1)}, \boldsymbol{x}^{(g2)}) = \sum_{n=1}^{N} | x_{g1n} - x_{g2n} |$$

Η απόσταση *City-Block* μοιάζει με την Ευκλείδεια με την διαφορά ότι αντί των τετραγωνικών χρησιμοποιούνται απόλυτες αποκλίσεις. Δίνει σχεδόν ίδια αποτελέσματα με την Ευκλείδεια, με εξαίρεση την περίπτωση που υπάρχουν έκτροπες παρατηρήσεις. Επειδή δίνει μικρότερο βάρος σε αυτές τις παρατηρήσεις, καταλήγει σε πιο ανθεκτικά αποτελέσματα. Αξίζει να σημειωθεί ότι η απόσταση αυτή αντιστοιχεί στο άθροισμα των μηκών των άλλων δύο πλευρών του τριγώνου που υποθέσαμε παραπάνω. Η χρήση της απόστασης *City-Block* συνίσταται στις περιπτώσεις κατά τις οποίες, για παράδειγμα, μια διαφορά της τάξεως του 1 στην πρώτη μεταβλητή (ή γονίδιο) και του 3 στη δεύτερη είναι ίδια με μία διαφορά της τάξεως του 2 στην πρώτη μεταβλητή και του 2 στην δεύτερη (Kaufman and Rousseeuw, 1990).

Απόσταση Maximum ή Chebyshev

$$d(\boldsymbol{x}^{(g1)}, \boldsymbol{x}^{(g2)}) = \max_{n} \left| x_{g1n} - x_{g2n} \right|$$

Χαρακτηριστικό της απόστασης αυτής είναι ότι δεν χρησιμοποιεί όλες τις αποστάσεις παρά μόνο την μεγαλύτερη απόσταση που σημειώνεται. Η απόσταση *Maximum* ορίζει ότι δύο παρατηρήσεις είναι διαφορετικές αν έχουν μεγάλες διαφορές σε μία τουλάχιστον μεταβλητή.

Αξίζει να σημειωθεί ότι οι παραπάνω τρεις αποστάσεις αποτελούν ειδική περίπτωση της οικογένειας αποστάσεων *Minkowski*, η οποία έχει την ακόλουθη γενική μορφή,

$$d(\mathbf{x}^{(g_1)}, \mathbf{x}^{(g_2)}) = \left(\sum_{n=1}^{N} | x_{g_1 n} - x_{g_2 n} |^k \right)^{1/k}$$ για κάποια παράμετρο $k \geq 1$.

Η παράμετρος *k* ονομάζεται *τάξη*: όσο μεγαλύτερη είναι η τιμή της παραμέτρου *k*, τόσο πιο σημαντική είναι η συνεισφορά των μεγαλύτερων ποσοτήτων $\left|x_{g_1 n} - x_{g_2 n}\right|$. Όπως γίνεται αντιληπτό, για *k* = 2 προκύπτει η *Ευκλείδεια* απόσταση, για *k* = 1 η απόσταση *City-Block*, ενώ η απόσταση *Maximum* προκύπτει ως περιοριστική έκφραση της οικογένειας *Minkowski* για την περίπτωση όπου $k \to \infty$. Γενικά, μεταβάλλοντας την παράμετρο *k* μεταβάλλεται και το βάρος που δίνεται σε μεγαλύτερες και μικρότερες αποστάσεις.

Απόσταση Mahalanobis

Η απόσταση *Minkowski* ανεπιφύλακτα υποθέτει ότι οι μεταβλητές είναι ασυσχέτιστες μεταξύ τους. Σε πολλές εφαρμογές αυτή η υπόθεση δεν ικανοποιείται. Επιπλέον, όταν οι συντεταγμένες των σημείων αντιπροσωπεύουν μετρήσεις οι οποίες είναι σύμφωνες με τυχαίες διακυμάνσεις διαφορετικών διαστάσεων, ή όταν οι συνιστώσες των σημείων είναι εξαρτημένες, είναι συχνά επιθυμητό να λαμβάνονται βάρη στις συνιστώσες προκειμένου να εξασφαλίζεται η συμβολή των συνδιακυμάνσεων.

Το *μέτρο απόστασης Mahalanobis* για τα γονίδια g_1 και g_2 είναι ένα μέτρο απόστασης το οποίο λαμβάνει υπόψη την συσχέτιση μεταξύ διανυσμάτων. Ορίζεται ως σταθμισμένη απόσταση στην οποία ο πίνακας των βαρών είναι ο αντίστροφος του πίνακα διακυμάνσεων-συνδιακυμάνσεων των διανυσμάτων $\boldsymbol{x}^{(g1)}$ και $\boldsymbol{x}^{(g2)}$, που συμβολίζεται ως **Σ**.

$$d(\boldsymbol{x}^{(g1)}, \boldsymbol{x}^{(g2)}) = \sqrt{\left(\mathbf{x}^{(g1)} - \mathbf{x}^{(g2)}\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}^{(g1)} - \mathbf{x}^{(g2)}\right)'}$$

Αξίζει να σημειωθεί ότι η απόσταση Mahalanobis ταυτίζεται με την Ευκλείδεια αν ο πίνακας διακυμάνσεων-συνδιακυμάνσεων αντικατασταθεί από τον μοναδιαίο πίνακα.

Στην εφαρμογή, ο πίνακας **Σ** ορίζεται έτσι ώστε να είναι ο από κοινού πίνακας διακυμάνσεων-συνδιακυμάνσεων (pooled) εντός των ομάδων (within-groups).

2.1.1.2 Μέτρα Ομοιότητας

Μία συνάρτηση s = s(P,Q) παρέχει ένα μέτρο ομοιότητας για τις παρατηρήσεις $\mathbf{x}_i = (x_{i1}, x_{i2}, .., x_{ip})$ και $\mathbf{x}_j = (x_{j1}, x_{j2}, .., x_{jp})$ όταν ισχύουν οι ακόλουθες ιδιότητες (βλ. Κούτρας, 2008):

(1) $s_{ij} > 0$ για κάθε *i, j* (αν $i = j \Rightarrow s_{ij}=1$)

(2) $s_{ij} \leq 1$

(3) $s_{ij} = s_{ji}$ (συμμετρική ιδιότητα)

Εσωτερικό Γινόμενο (Inner ή Dot Product)

Το εσωτερικό γινόμενο είναι ένα μέτρο ομοιότητας. Έστω $\boldsymbol{x}^{(g_1)}=(x_{g_1 1},..,\ x_{g_1 N})$ και $\boldsymbol{x}^{(g_2)} = (x_{g_2 1},\ldots,x_{g_2 N})$ δηλώνουν δύο διανύσματα γονιδίων. Η απόσταση του εσωτερικού γινομένου μεταξύ των $\boldsymbol{x}^{(g_1)}$ και $\boldsymbol{x}^{(g_2)}$ ορίζεται ως,

$$\mathbf{s}_{g_1 g_2} = \left\langle \mathrm{x}^{(\mathrm{g}_1)} \cdot \mathrm{x}^{(\mathrm{g}_2)} \right\rangle = \mathrm{x}^{(\mathrm{g}_1)} \cdot \mathrm{x}^{(\mathrm{g}_2)'} = \sum_{n=1}^{N} x_{g_1 n} x_{g_2 n}$$

Συντελεστής Συσχέτισης του Pearson

Αν οι μεταβλητές $x_{g_1 n}$ και $x_{g_2 n}$ είναι τυποποιημένες, για παράδειγμα κεντραρισμένες στον μέσο τους και διαιρεμένες από την τυπική τους απόκλιση, τότε το εσωτερικό γινόμενο αντιστοιχεί στον *συντελεστή συσχέτισης του Pearson*, ο οποίος είναι ένα μέτρο ομοιότητας,

$$\mathbf{s}_{g_1 g_2} = \rho_{g_1 g_2} = \sum_{n=1}^{N} z_{g_1 n} \cdot z_{g_2 n}$$

όπου τα $z_{g_1 n}$ και $z_{g_2 n}$ είναι οι τυποποιημένες εκφράσεις των τιμών $x_{g_1 n}$ και $x_{g_2 n}$, αντίστοιχα. Το μέτρο αυτό δεν εξαρτάται από την επιλογή της μονάδας μέτρησης. Ο *συντελεστής συσχέτισης του Pearson* αντιμετωπίζει κάθε αντικείμενο ως μια τυχαία μεταβλητή με N παρατηρήσεις και μετρά την ομοιότητα μεταξύ των δύο γονιδίων υπολογίζοντας την γραμμική σχέση μεταξύ των κατανομών των δύο αντίστοιχων τυχαίων μεταβλητών. Ο συντελεστής αυτός σημειώνει ευρεία εφαρμογή και έχει αποδειχθεί αποτελεσματικός ως μέτρο ομοιότητας για ανάλυση γονιδιακών εκφράσεων (D. Jiang, C. Tang and A. Zhang, 2004; D. Jiang, J. Pei and A. Zhang, 2003; C. Tang *et al*, 2001; C. Tang and A. Zhang, 2002; J. Yang *et al*, 2002; M. B. Eisen *et al*., 1998).

Οι Eisen *et al*. (1998) σημειώνουν ότι μία καλή επιλογή μέτρου ομοιότητας για την σύγκριση μικροσυστοιχιών από δύο γονίδια πρέπει να προσαρμόζεται ικανοποιητικά στην διαισθητική βιολογική αντίληψη του τί σημαίνει για δυο γονίδια να είναι συνεκφρασμένα (co-expressed). Οι Alon *et al*. (1999), επίσης, επισημαίνουν ότι η ένταση κάθε γονιδίου στους ιστούς μπορεί να θεωρηθεί ως πρότυπο το οποίο συσχετίζεται με εκφράσεις προτύπων άλλων γονιδίων.

Ο συντελεστής συσχέτισης συλλαμβάνει την ομοιότητα στο ‘σχήμα’ του προφίλ των εκφράσεων μεταξύ δύο γονιδίων, ωστόσο, δεν δίνει έμφαση στις διαστάσεις των επιπέδων των όρων. Ο συντελεστής συσχέτισης μπορεί να είναι υψηλός μεταξύ δύο γονιδίων που

επηρεάζονται από την ίδια διαδικασία, ακόμα και στην περίπτωση που καθένα έχει διαφορετικές διαστάσεις απόκρισης στην διαδικασία. Θετική συσχέτιση μεταξύ δύο υψηλά εκφρασμένων γονιδίων είναι πολύ περισσότερο σημαντική σε σχέση με την ίδια τιμή μεταξύ δύο ελάχιστα εκφρασμένων γονιδίων. Κάνοντας χρήση των συσχετίσεων, αγνοείται αυτή η εξάρτηση της αξιοπιστίας του επιπέδου του απόλυτου όρου.

Ο *συντελεστής συσχέτισης του Pearson*, όπως κάθε άλλο *μέτρο εγγύτητας* (*proximity measure*), είναι ευαίσθητο στις έκτροπες παρατηρήσεις. Επιπλέον, ιδιαίτερη προσοχή πρέπει να σημειωθεί προκειμένου να εντοπισθούν οι έκτροπες παρατηρήσεις πριν την εφαρμογή του συντελεστή του Pearson ως μέτρου ομοιότητας. Επιπροσθέτως, ο συντελεστής συσχέτισης έχει εύρος $-1 \leq \rho_{g_1 g_2} \leq 1$, και οι αρνητικές τιμές πρέπει να χειρίζονται με προσοχή. Όταν οι συντελεστές υπολογίζονται από δείγματα και, επίσης, υποβάλλονται στην δειγματική μεταβλητότητα, τότε συντελεστές με μικρή αρνητική τιμή πρέπει να θέτονται ίσοι με μηδέν. Εναλλακτικά, αρνητικοί συντελεστές συσχέτισης πρέπει να αντικαθίστανται από την απόλυτη τιμή τους, όπως συμβαίνει προαιρετικά σε ορισμένους αλγόριθμους της ανάλυσης συστάδων. Η λογική της αντιστροφής των προσήμων είναι ότι αν δύο μεταβλητές $x_{g_1 n}$ και $x_{g_2 n}$ είναι αρνητικά συσχετισμένες τότε οι $x_{g_1 n}$ και $-x_{g_2 n}$ θα είναι θετικά συσχετισμένες με τον ίδιο απόλυτο συντελεστή. Οι δύο μεταβλητές (με την μία να έχει αντίθετο πρόσημο) είναι περισσότερο όμοιες όσο μεγαλύτερος είναι ο απόλυτος συντελεστής συσχέτισης.

Ο ερευνητής πρέπει να γνωρίζει πότε ο αλγόριθμος αντιστρέφει το πρόσημο ενός αρνητικού συντελεστή συσχέτισης διότι το γεγονός ότι δύο γονίδια έχουν αντίθετους όρους έκφρασης στα δείγματα μπορεί να είναι ένα σημαντικό χαρακτηριστικό το οποίο δεν πρέπει να χαθεί λόγω εσφαλμένων υπολογισμών. Για παράδειγμα, δύο γονίδια μπορεί να είναι υψηλά (αρνητικά) συσχετισμένα στο δείγμα διότι το ένα είναι ισχυρά συσχετισμένο με ιστούς του όγκου ενώ το άλλο γονίδιο είναι ισχυρά συσχετισμένο με υγιείς ιστούς.

Δοθέντος ενός μη-αρνητικού συντελεστή συσχέτισης $\rho_{g_1 g_2}$, ένα αντίστοιχο μέτρο απόστασης (ανομοιότητας) μπορεί να ορισθεί ως,

$$d_{g_1 g_2} = 1 - \rho_{g_1 g_2}.$$

Το μέτρο αυτό απόστασης είναι συμμετρικό και μη-αρνητικό (λαμβάνει τιμές στο διάστημα [0, 2]) και έχει την ιδιότητα ότι, αν τα διανύσματα $\mathbf{x}^{(g_1)}$ και $\mathbf{x}^{(g_2)}$ των γονιδίων g_1 και g_2 είναι γραμμικά ανεξάρτητα (ασυσχέτιστα), τότε $\rho_{g_1 g_2} = 0$ και $d_{g_1 g_2} = 1$. Αν τα παραπάνω διανύσματα είναι γραμμικά εξαρτημένα (τέλεια συσχέτιση), τότε $\rho_{g_1 g_2} = 1$ και $d_{g_1 g_2} = 0$.

Επιπλέον, το μέτρο αυτό είναι αμετάβλητο σε αλλαγές της 'τοποθεσίας' (location) ή της κλίμακας (scale) για καθένα από τα g_1 και g_2. Μεταβολές στο μέσο επίπεδο μέτρησης ή στο εύρος των μετρήσεων μεταξύ των δειγμάτων εξαλείφονται αποτελεσματικά.

Spearman Rank Συντελεστής Συσχέτισης

Όταν τα επίπεδα των όρων έχουν αρχικά μετατραπεί σε ιεραρχική σειρά, ένας συντελεστής συσχέτισης βασισμένος στις τάξεις όπως ο *συντελεστής συσχέτισης του Spearman*

$$\mathbf{s}_{g_1 g_2} = \rho^*_{g_1 g_2}$$

είναι κατάλληλο ως μέτρο ομοιότητας. Η μαθηματική έκφραση για το $\rho^*_{g_1 g_2}$ είναι ίδια με αυτή του *συντελεστή του Pearson* με την διαφορά ότι οι τυποποιημένες μεταβλητές $z_{g_1 n}$ και $z_{g_2 n}$ προκύπτουν από τις τάξεις $r_{g_1 n}$ και $r_{g_2 n}$ των μεταβλητών $x_{g_1 n}$ και $x_{g_2 n}$. Οι τάξεις υπολογίζονται από τα δείγματα για κάθε γονίδιο ως ακολούθως,

$$r_{g_1 n} = Rank_{1 \le n \le N}(x_{g_1 n}), \quad r_{g_2 n} = Rank_{1 \le n \le N}(x_{g_2 n}).$$

Το μέτρο αυτό, επίσης, δεν εξαρτάται από την επιλογή της μονάδας μέτρησης, ενώ ισχύει $-1 \le s_{g_1 g_2} \le 1$.

Η βασική διαφορά μεταξύ των δύο τελευταίων μέτρων ομοιότητας είναι ότι ο *συντελεστής του Pearson* φαίνεται ως μια γραμμική σχέση μεταξύ των γονιδίων g_1 και g_2, ενώ ο *συντελεστής του Spearman* αναζητά μια μονότονη σχέση μεταξύ τους. Επιπλέον, κανένα μέτρο ομοιότητας δεν μπορεί να χρησιμοποιηθεί άμεσα ως συντελεστής ομοιότητας δεδομένου ότι μπορεί να λάβει και αρνητικές τιμές. Προκειμένου να παρακαμφθεί αυτό το μειονέκτημα, εφαρμόζονται κατάλληλοι μετασχηματισμοί ώστε να ανήκουν στο διάστημα [0, 1]. Εάν γονίδια με έντονα αρνητική συσχέτιση θεωρούνται πολύ διαφορετικά διότι η ομοιότητά τους είναι προσανατολισμένη σε αντίθετη κατεύθυνση, τότε λαμβάνεται ο μετασχηματισμός,

$$s_{g_1 g_2} = (1 + d_{g_1 g_2}) / 2$$

σύμφωνα με τον οποίο όταν $s_{g_1 g_2} = 0$ τότε $d_{g_1 g_2} = -1$. Επιπλέον, υπάρχουν περιπτώσεις όπου γονίδια με έντονα αρνητική συσχέτιση πρέπει να ομαδοποιηθούν, διότι μετρούν κατ' ουσίαν το ίδιο αντικείμενο. Σε αυτή τη περίπτωση προτιμάται ο μετασχηματισμός,

$$s_{g_1 g_2} = \left| d_{g_1 g_2} \right|.$$

2.1.2 Συστηματικές Μέθοδοι Ομαδοποίησης

Οι μέθοδοι ομαδοποίησης διακρίνονται σε δύο βασικές κατηγορίες, σύμφωνα με τον τρόπο που οργανώνουν την διαμόρφωση των ομάδων, τις *ιεραρχικές* και τις *μη-ιεραρχικές* μεθόδους.

Στην ιεραρχική ομαδοποίηση, τα δεδομένα θέτονται σε αυστηρή ιεραρχία εμφωλευμένων υποομάδων. Οι τεχνικές της ιεραρχικής ομαδοποίησης μπορούν να είναι είτε συσσωρευτικές (bottom-up) είτε διαιρετικές (top-down). Μία μέθοδος συσσωρευτικής ομαδοποίησης αρχίζει θεωρώντας κάθε ξεχωριστό σημείο ως μία συστάδα. Η διαδικασία αρχίζει με G συστάδες (στην περίπτωση ομαδοποίησης γονιδίων) και διαδοχικά συνδυάζει τις δύο πλησιέστερες συστάδες και δι' αυτού μειώνεται το πλήθος των συστάδων κατά μία σε κάθε βήμα. Σε αντίθεση, μία μέθοδος διαιρετικής ομαδοποίησης αρχίζει με μία συστάδα η οποία περιέχει όλα τα G γονίδια και διαδοχικά χωρίζει την λιγότερο ομοιογενή συστάδα σε δύο διαδοχικές ομάδες, η κάθε μία από τις οποίες είναι πιο ομοιογενής σε σχέση με την αρχική συστάδα. Ο διαχωρισμός μπορεί να συνεχιστεί έως ότου δημιουργηθούν G ομάδες (ατομικά γονίδια). Προκειμένου να προκύψει μια λύση με τον βέλτιστο αριθμό συστάδων, πρέπει να αποφασισθεί η συγκεκριμένη βαθμίδα στην οποία θα σταματήσει η επαναληπτική διαδικασία. Οι Causton *et al.* (2003) συμπληρώνουν ότι, πέραν των παραπάνω μεθόδων, ο αλγόριθμος ομαδοποίησης είναι δυνατόν να αρχίζει με τον διαχωρισμό των δεδομένων σε ένα προκαθορισμένο πλήθος ομάδων, και στη συνέχεια να επανεξετάζει και να βελτιώνει τις καταχωρίσεις των αντικειμένων στις ομάδες, μεταβάλλοντας τα όρια. Η μέθοδος *k-means* είναι ένα παράδειγμα τέτοιου αλγορίθμου.

Η ιεραρχική ομαδοποίηση δημιουργεί μια ιεραρχική σειρά εμφωλευμένων ομάδων οι οποίες μπορούν να απεικονισθούν γραφικά μέσω ενός διαγράμματος, το οποίο καλείται *δενδρόγραμμα* (dendrogram) και απεικονίζει τις συγχωνεύσεις ή διαιρέσεις που πραγματοποιούνται στα διαδοχικά επίπεδα. Τα κλαδιά του γραφήματος δεν καταγράφουν μόνο τον σχηματισμό των συστάδων, αλλά, επιπλέον, δείχνουν την ομοιότητα μεταξύ των ομάδων. 'Κόβοντας' το δενδρόγραμμα σε κάποια επίπεδα, λαμβάνεται ένα συγκεκριμένο πλήθος ομάδων. Αναδιατάσσοντας τα δεδομένα έτσι ώστε να κλαδιά του αντίστοιχου δενδρογράμματος να μη διασταυρώνονται, το σύνολο των δεδομένων μπορεί να

διευθετηθούν με τα όμοια αντικείμενα να τοποθετούνται μαζί (D. Jiang, C. Tang and A. Zhang, 2004).

Η ιεραρχική συσταδοποίηση όχι μόνο ομαδοποιεί γονίδια με παρόμοιο πρότυπο έκφρασης, αλλά επιπλέον παρέχει ένα φυσικό τρόπο γραφικής αναπαράστασης του συνόλου των δεδομένων. Η γραφική αναπαράσταση επιτρέπει πλήρη έλεγχο του συνόλου των δεδομένων και εξασφαλίζει μια πρώτη εικόνα της κατανομής τους. Εντούτοις, η συμβατική συσωρευτική προσέγγιση χαρακτηρίζεται από έλλειψη ανθεκτικότητας, που σημαίνει ότι μια μικρή διαταραχή των στοιχείων μπορεί να μεταβάλλει σε μεγάλο βαθμό την δομή του ιεραρχικού δενδρογράμματος. Άλλο μειονέκτημα της ιεραρχικής μεθόδου είναι η υψηλή υπολογιστική της πολυπλοκότητα. Προκειμένου να κατασκευασθεί ένα 'ολοκληρωμένο' δενδρόγραμμα, η διαδικασία ομαδοποίησης απαιτεί $\frac{n^2-n}{2}$ βήματα συγχώνευσης. Επιπλέον, τόσο στις συσωρευτικές όσο και στις διαιρετικές ιεραρχικές μεθόδους, η 'άπληστη' φύση τους εμποδίζει την τελειοποίηση της προηγούμενης ομαδοποίησης. Εάν μια 'κακή' απόφαση ληφθεί κατά τα αρχικά βήματα, δεν μπορεί να διορθωθεί στα επόμενα (D. Jiang, C. Tang and A. Zhang, 2004).

Αξιοσημείωτο είναι το γεγονός ότι χρησιμοποιώντας διαφορετικές μεθόδους σύνδεσης ή κάνοντας μικρές αλλαγές στο σύνολο των δεδομένων μπορούν να προκύψουν πολύ διαφορετικά δενδρογράμματα.

Σημειώνεται ιδιαίτερη ποικιλία ως προς τις εναλλακτικές μεθόδους εφαρμογής για κάθε μία από τις παραπάνω κατηγορίες. Στην παρούσα εργασία θα παρουσιαστούν η ιεραρχική μέθοδος της *Πλήρους Συνένωσης (Complete Linkage)* και η μη ιεραρχική μέθοδος *K-Means*, οι οποίες θα εφαρμοστούν στην ανάλυση που ακολουθεί. Η επιλογή των μεθόδων αυτών έγινε με γνώμονα την βιβλιογραφία. Ενδεικτικά αναφέρονται οι ακόλουθες πηγές, Smet *et al*., 2002; Sherlock, 2000; D. Jiang *et al.*, 2004; C. Tang *et al.*, 2001; C. Tangand and A. Zhang, 2002; J. Copland *et al.*, 2003; J. Dietzsch, N. Gehlenborg and K. Nieselt, 2006.

2.1.2.1 Ιεραρχική Μέθοδος Πλήρους Συνένωσης

Η μέθοδος ομαδοποίησης της *πλήρους συνένωσης* χρησιμοποιεί το μέτρο απόστασης του μακρινότερου γείτονα για τις αποστάσεις εντός των ομάδων. Η ομαδοποίηση πραγματοποιείται ιεραρχικά. Κάθε επίπεδο προκύπτει από την συγχώνευση δύο συστάδων

του προηγούμενου επιπέδου. Όταν οι δύο συστάδες έχουν σημεία σε κοντινή απόσταση, η μέθοδος της απλής συνένωσης έχει τη τάση να συνδέει αυτές τις δύο συστάδες. Σε κάθε στάδιο, η απόσταση (ομοιότητα) μεταξύ των συστάδων ορίζεται ως η απόσταση (ομοιότητα) μεταξύ των δύο στοιχείων, ένα για κάθε συστάδα, τα οποία σημειώνουν την μεγαλύτερη απόσταση. Επομένως, οι ομάδες συνενώνονται σύμφωνα με την απόσταση μεταξύ των μακρινότερων σημείων (Johnson and Wichern, 1998). Όπως γίνεται φανερό, η πλήρης συνένωση εξασφαλίζει ότι όλα τα στοιχεία μιας ομάδας βρίσκονται μεταξύ ενός μεγίστου διαστήματος. Η τιμή της απόστασης αυτής είναι η διάμετρος της μικρότερης σφαίρας η οποία μπορεί να περικλείσει την συστάδα αυτή ως αποτέλεσμα της συνένωσης των δύο ομάδων. Από αυτήν την ιδιότητα, η μέθοδος ονομάζεται εναλλακτικά και *μέθοδος του μακριότερου γείτονα*. Ο Johnson (1967) σημειώνει ότι η ομαδοποίηση της πλήρους συνένωσης είναι αμετάβλητη σε μονότονους μετασχηματισμούς των στοιχείων του πίνακα αποστάσεων (proximity matrix). Πιο συγκεκριμένα, μια 'νέα' εκχώρηση αποστάσεων η οποία διατηρεί τις ίδιες αρχικές διατάξεις με τις αρχικές αποστάσεις δεν μεταβάλλει την διαμόρφωση των ομάδων πλήρους συνένωσης.

Στην συνέχεια παρουσιάζονται τα βήματα του αλγορίθμου συσωρευτικής ιεραρχικής ομαδοποίησης για N αντικείμενα.

1. Αρχίζουμε με N ομάδες, κάθε μία από τις οποίες περιέχει ένα αντικείμενο και έναν πίνακα αποστάσεων $\mathbf{D}_{N\times N}=\{d_{ik}\}$.
2. Εντοπίζουμε το ζεύγος των αντικειμένων που σημειώνουν την μικρότερη απόσταση στον πίνακα αποστάσεων **D**. Έστω ότι η απόσταση μεταξύ των ομάδων αυτών U και V είναι d_{UV}.
3. Συνενώνουμε τις ομάδες U και V και συμβολίζουμε την νέα ομάδα (UV). Ανανεώνουμε τον πίνακα αποστάσεων **D** διαγράφοντας τις γραμμές και τις στήλες που αντιστοιχούν στις ομάδες U και V και προσθέτοντας μια νέα γραμμή και στήλη που περιέχει τις αποστάσεις μεταξύ της νέας ομάδας (UV) και όσων παρέμειναν ως είχαν.
4. Επαναλαμβάνουμε τα βήματα 1 και 2 (N-1) φορές. Όλα τα αντικείμενα θα ανήκουν σε μία ομάδα μετά τον τερματισμό του αλγορίθμου. Καταγράφουμε τις ομάδες που έχουν συγχωνευθεί και τα επίπεδα (αποστάσεις ή ομοιότητες) στα οποία οι συγχωνεύσεις πραγματοποιήθηκαν.

Σύμφωνα με την παρούσα μέθοδο συνένωσης, στο βήμα 3 του ανωτέρω γενικού αλγορίθμου οι αποστάσεις μεταξύ της ομάδας (UV) και κάθε άλλης ομάδας W υπολογίζεται ως εξής:

$$d_{(UV)W} = \max\{d_{UW}, d_{VW}\},$$

όπου d_{UW} και d_{VW} είναι οι αποστάσεις μεταξύ των πιο μακρινών σημείων των ομάδων U και W και των ομάδων V και W, αντίστοιχα.

Με την λήψη πρόωρων αποφάσεων ως μόνιμων το πλήθος των δυνατοτήτων που χρειάζεται να εξετασθεί μειώνεται σημαντικά σε σχέση με την πλήρη απαρίθμηση. Ωστόσο, αυτή η ίδια σύμβαση αποκλείει την έγκαιρη ανακάλυψη λαθών ή την αξιοποίηση μεταγενέστερων ευκαιριών για διόρθωσή τους. Αξίζει να επισημανθεί, επίσης, ότι η μέθοδος αυτή παράγει συμπαγείς και σφαιρικές ομάδες, ωστόσο αρκετά συχνά αποτυγχάνει να ξεχωρίσει κάποιες πολύ συμπαγείς μικρές ομάδες.

Στην περίπτωση που γίνεται χρήση μέτρων ομοιότητας μεταξύ των δύο πιο ανόμοιων στοιχείων των δύο ομάδων U και V και οι δύο αυτές ομάδες συνενωθούν, τότε κάθε ζεύγος στοιχείων στην τελική συστάδα έχει ομοιότητα τουλάχιστον ίση με την ομοιότητα μεταξύ των ομάδων αυτών.

2.1.2.2 Μη Ιεραρχική Μέθοδος *k-means*

Ο MacQueen (1967) εισήγαγε μια μη-ιεραρχική μέθοδο συσταδοποίησης, η οποία καλείται *μέθοδος k-means*. Η μέθοδος αυτή ορίζει κάθε αντικείμενο στην συστάδα έχοντας το πλησιέστερο κέντρο. Ουσιώδες χαρακτηριστικό της μεθόδου αποτελεί το γεγονός ότι τα κέντρα των ομάδων υπολογίζονται επί τη βάσει των τρεχουσών στοιχείων από τα οποία οι ομάδες απαρτίζονται, μετά από κάθε μετακίνηση στοιχείου (Anderberg, 1973). Εφαρμόζοντας την μέθοδο ομαδοποίησης *k-means*, το συνολικό πλήθος των συστάδων, *k*, καθορίζεται πριν την εφαρμογή της διαδικασίας συσταδοποίησης. Δεδομένου ότι ο πίνακας αποστάσεων (proximity matrix) δεν χρειάζεται να κατασκευασθεί και τα δεδομένα δεν χρειάζεται να αποθηκεύονται από τον υπολογιστή κατά την διάρκεια της διαδικασίας, η μέθοδος μπορεί να εφαρμοσθεί σε πολύ μεγαλύτερο σύνολο δεδομένων συγκριτικά με τις ιεραρχικές μεθόδους.

Οι μη ιεραρχικές μέθοδοι, γενικά, ξεκινούν είτε από μία αρχική κατάτμηση των αντικειμένων σε ομάδες είτε από ένα σύνολο αρχικών σημείων, τα οποία θα συγκροτήσουν

τους πυρήνες των συστάδων. Ο αλγόριθμος της μεθόδου *k-mean* αντιστοιχεί κάθε στοιχείο της συστάδας λαμβάνοντας το πλησιέστερο κέντρο (μέση τιμή). Τα βασικά βήματα της μεθόδου είναι τα ακόλουθα:

1. Επιλέγουμε ένα σύνολο *k* σημείων ως μητρικά σημεία (cluster seed). Τα σημεία αυτά αντιπροσωπεύουν μια αρχική υπόθεση για τα κέντρα των *k* συστάδων.
2. Ορίζουμε κάθε μεμονωμένη παρατήρηση στην συστάδα της οποίας το κέντρο είναι πλησιέστερα. Η Ευκλείδεια απόσταση συνήθως χρησιμοποιείται ως μέτρο απόστασης με ή χωρίς τυποποιημένες παρατηρήσεις. Τα κέντρα επαναϋπολογίζονται για κάθε συστάδα που λαμβάνει ένα επιπλέον σημείο και για την αντίστοιχη συστάδα από την οποία αφαιρείται.
3. Επαναλαμβάνουμε το βήμα 2 έως ότου να μην πραγματοποιηθεί καμία αλλαγή στην σύνθεση των συστάδων.

Χρησιμοποιώντας τα *k* μητρικά σημεία και στηριζόμενοι μόνο σε μία ανακατανομή, η διαδικασία αυτή επιτυγχάνει ένα πολύ οικονομικό διαχωρισμό των στοιχείων της ομάδας. Έστω ότι επιδιώκεται η ομαδοποίηση *m* στοιχείων σε *k* ομάδες. Η συνολική διαδικασία από την αρχική διαμόρφωση έως την τελική μορφή των ομάδων απαιτεί *k(2m-k)* υπολογισμούς αποστάσεων, *(k-1)(2m-k)* συγκρίσεις αποστάσεων, και *m-k* ενημερώσεις των κέντρων των ομάδων. Η οικονομία η οποία είναι συνυφασμένη με τη μέθοδο αυτή απορρέει από την αποδοχή της αρχικής ανακατανομής των δεδομένων σε αντίθεση με τη συνέχιση της επεξεργασίας έως ότου επιτευχθεί η σύγκλιση. Προφανώς η μέθοδος δίνει καλά αποτελέσματα, επειδή οι περισσότερες μεγάλες αλλαγές στη σύνθεση των ομάδων προκύπτουν με την αρχική ανακατανομή, μεταγενέστερες ανακατανομές συνήθως οδηγούν σε σχετικά λίγες αναδιοργανώσεις.

Επιπλέον, ο αλγόριθμος *k-means* δεν παρουσιάζει ιδιαίτερες δυσκολίες με τις ελλείπουσες τιμές (missing values) διότι οι ανανεώσεις του μέσου και οι υπολογισμοί των αποστάσεων μπορούν να εφαρμοσθούν με κάποιες ελλείπουσες τιμές. Στην πράξη, γρήγοροι τύποι ανανέωσης στην *k-means* είναι σχετικά δύσκολο να τροποποιηθούν για τις ελλείπουσες τιμές. Για τον συγκεκριμένο αλγόριθμο είναι ευκολότερο να γίνεται χρήση τεκμαρτών δεδομένων.

Η τελική κατάταξη των στοιχείων στις ομάδες μπορεί να εξαρτάται, σε κάποιο βαθμό, από τον αρχικό διαχωρισμό ή την αρχική επιλογή των μητρικών σημείων. Η μέθοδος *k-means* δεν δίνει κάποια διάταξη στα αντικείμενα μιας ομάδας. Καθώς το πλήθος των συστάδων *k*

μεταβάλλεται, τα αντικείμενα των συστάδων μπορούν να αλλάξουν με αυθαίρετο τρόπο. Για παράδειγμα, το αποτέλεσμα που αντιστοιχεί σε *k*=4 συστάδες, ίσως δεν είναι εμφωλευμένο στο αποτέλεσμα για *k*=3. Προκειμένου να ελεγχθεί η σταθερότητα της ομαδοποίησης, συνιστάται η επανάληψη του αλγορίθμου με νέα μητρικά σημεία. Αφού οι ομάδες καθορισθούν, διαισθήσεις σχετικά με τις ερμηνείες τους ενισχύονται από αναδιάταξη του καταλόγου των στοιχείων, έτσι ώστε αυτά στην πρώτη ομάδα να εμφανιστούν πρώτα, τα υπαγόμενα στην δεύτερη συστάδα να εμφανιστούν στη συνέχεια, και ούτω καθεξής. Ένας πίνακας που περιέχει τα κέντρα των ομάδων και την εντός των ομάδων διακύμανση, επίσης, συμβάλει στην οριοθέτηση των διαφορών των ομάδων.

Ωστόσο, η μέθοδος έχει αρκετά μειονεκτήματα ως αλγόριθμος συσταδοποίησης βασιζόμενος σε γονίδια. Το πλήθος των ομάδων στο σύνολο δεδομένων γονιδιακών εκφράσεων δεν είναι, συνήθως, γνωστό εκ των προτέρων. Για τον εντοπισμό του βέλτιστου πλήθους συστάδων, συνήθως επαναλαμβάνεται ο αλγόριθμος με διαφορετικές τιμές του *k* και συγκρίνονται τα αποτελέσματα της ομαδοποίησης. Για ένα μεγάλο σύνολο γονιδιακών εκφράσεων, οι οποίες περιλαμβάνουν χιλιάδες γονίδια, αυτή η εκτεταμένη παράμετρος τελειοποίησης της διαδικασίας δύναται να αποβεί μη πρακτική. Επιπλέον, τα δεδομένα γονιδιακών εκφράσεων τυπικά περιλαμβάνουν μεγάλη ποσότητα θορύβου. Παρ' όλα αυτά, ο αλγόριθμος *k-means* κατανέμει κάθε γονίδιο σε μία ομάδα, οπότε μπορεί να έχει ως αποτέλεσμα ο αλγόριθμος να είναι ευαίσθητος στο θόρυβο (D. Jiang *et al.*, 2004; F.D. Smet *et al.*, 2002; G. Sherlock, 2000).

Η μέθοδος *k-means* είναι μία μη-δομημένη προσέγγιση, η οποία πραγματοποιείται σε μια τοπική διάπλαση και παράγει μια μη-οργανωμένη συλλογή συστάδων που δεν είναι πάντα ερμηνεύσιμη. Η μέθοδος αυτή ορίζει κάθε γονίδιο σε μόνο μία συστάδα. Αυτό δεν είναι απαραιτήτως βιολογικά αποδεκτό. Επιπλέον, γονίδια που ορίζονται στην ίδια συστάδα δεν έχουν απαραίτητα την ίδια δομή έκφρασης. Σημαντικό είναι να ελέγχουμε αν η τελική συστάδα έχει έννοια από βιολογική σκοπιά.

2.1.2.3 Μη Ιεραρχική Μέθοδος *Partitioning Around Medoid*

Η μέθοδος *Partitioning Around Medoid (PAM),* επίσης γνωστή ως ομαδοποίηση *k-medoid,* είναι μια παραλλαγή της *k-mean* με στόχο την ελαχιστοποίηση της εντός των συστάδων διακύμανσης (Saha and Mukhopadhyay, 2001). Η μέθοδος *PAM* διαφέρει από την μέθοδο *k-*

means στο σημείο ότι ως πυρήνας μιας συστάδας είναι ένα στοιχείο της συστάδας (*medoid*) και επιδιώκεται η ελαχιστοποίηση της απόστασης των υπόλοιπων στοιχείων από τον πυρήνα (Σημειώσεις Κ. Φωκιανός & Χ. Χαραλάμπους, 2010). Είναι πιο ανθεκτική σε θόρυβο και έκτροπες παρατηρήσεις συγκριτικά με την *k-means* διότι ελαχιστοποιεί το άθροισμα των ανομοιοτήτων αντί του αθροίσματος των τετραγώνων των Ευκλείδειων αποστάσεων. Για τον λόγο αυτό, η μέθοδος *PAM* κάποιες φορές εφαρμόζεται σε περιπτώσεις κατά τις οποίες αντιπροσωπευτικά αντικείμενα (medoids) θεωρούνται αντί των κέντρων. Δεδομένου ότι χρησιμοποιεί τα πιο κεντρικά τοποθετημένα σημεία των ομάδων, είναι λιγότερο ευαίσθητη σε έκτροπες παρατηρήσεις συγκρινόμενη με την μέθοδο *k-means*. Ωστόσο, η *PAM* έχει ένα μειονέκτημα ως προς την αποτελεσματικότητα για μεγάλα σύνολα δεδομένων ως συνέπεια της πολυπλοκότητάς της (Han *et al.*, 2001; Park *et al.*, 2006).

Το *medoid* μπορεί να ορισθεί ως το αντικείμενο της ομάδας, του οποίου η μέση ανομοιότητα με όλα τα αντικείμενα της ομάδας είναι η ελάχιστη, δηλαδή είναι το πιο κεντρικά τοποθετημένο σημείο στο δεδομένο σύνολο (Ιστοσελίδα 5).

Αρχικά επιλέγεται ένα καλό αρχικό σύνολο από *k* αντιπροσωπευτικά αντικείμενα μεταξύ αυτών του συνόλου δεδομένων (Build Phase). Όπως επικαλούνται τα ονόματά τους, τα αντικείμενα αυτά πρέπει να αντιπροσωπεύουν ποικίλες όψεις της δομής των δεδομένων. Ακολούθως ελέγχεται κατά πόσο η εναλλαγή ενός στοιχείου με ένα *medoid* θα ελαχιστοποιήσει την απόσταση μεταξύ του πυρήνα και των άλλων στοιχείων και αν συμβεί, πραγματοποιείται. Πιο συγκεκριμένα, αφού προσδιορισθεί το σύνολο των *k* αντιπροσωπευτικών αντικειμένων, οι *k* συστάδες κατασκευάζονται αντιστοιχίζοντας κάθε αντικείμενο του συνόλου των δεδομένων στο πλησιέστερο αντιπροσωπευτικό αντικείμενο (Swap Phase).

Στην συνέχεια παρατίθενται τα βήματα του αλγορίθμου της μεθόδου.

1. Τυχαία επιλογή *k* σημείων από το σύνολο *n* των δεδομένων ως *medoid*.
2. Σύνδεση κάθε σημείου στο κοντινότερο *medoid* (το ‘κοντινότερο’ εδώ ορίζεται χρησιμοποιώντας κάθε έγκυρη μετρική απόσταση, συνηθέστερα την Ευκλείδεια απόσταση, απόσταση Manhattan ή απόσταση Minkowski).
3. Για κάθε medoid *m* και για κάθε σημείο δεδομένων που δεν έχει θεωρηθεί ως *medoid* υπολογίζεται το συνολικό κόστος σχηματισμού.
4. Επιλέγεται ο σχηματισμός με το χαμηλότερο κόστος.

5. Επαναλαμβάνονται τα βήματα 2 έως 4 έως ότου να μην υπάρχει κάποια αλλαγή στα *medoids*.

Συγκρινόμενη με την μέθοδο *k-means*, η *PAM* έχει τα ακόλουθα χαρακτηριστικά (Ιστοσελίδα 16):

(α) χειρίζεται, επίσης, πίνακα ανομοιότητας των δοθέντων δεδομένων ή όταν παρουσιάζεται με ένα $n \times p$ πίνακα δεδομένων, ο αλγόριθμος αρχικά υπολογίζει έναν πίνακα ανομοιότητας.

(β) είναι πιο εύρωστη μέθοδος διότι ελαχιστοποιεί το άθροισμα ανομοιότητας αντί του αθροίσματος τετραγώνων της ευκλείδειας απόστασης.

(γ) διαθέτει πρωτότυπη γραφική απεικόνιση (silhouette plot) η οποία επιτρέπει την επιλογή του βέλτιστου πλήθους συστάδων.

(δ) επιτρέπει την επιλογή του πλήθους των συστάδων κάνοντας χρήση του μέσου.

Σε πολλά προβλήματα ομαδοποίησης το ενδιαφέρον προσανατολίζεται στον χαρακτηρισμό των ομάδων σε όρους τυπικών ή αντιπροσωπευτικών αντικειμένων. Τα αντικείμενα αυτά αντιπροσωπεύουν τις ποικίλες δομικές όψεις του συνόλου των δεδομένων υπό έρευνα. Μπορεί να υπάρχουν πολλοί λόγοι για διερεύνηση των αντιπροσωπευτικών αντικειμένων καθώς τα αντικείμενα αυτά εκτός από το γεγονός ότι παρέχουν ένα χαρακτηρισμό των ομάδων, μπορούν, επίσης, να χρησιμοποιηθούν συχνά για περαιτέρω έρευνα, ειδικότερα όταν είναι περισσότερο οικονομικό ή βολικό να γίνεται χρήση ενός μικρού συνόλου *k* αντικειμένων αντί ενός μεγάλου συνόλου. Σύμφωνα με τη μέθοδο *PAM* τα αντιπροσωπευτικά αντικείμενα μιας ομάδας είναι τα *medoids*. Η μέση ανομοιότητα μετρά τη συνοχή της ομάδας και συνεπώς, την ποιότητα της ομαδοποίησης (Kaufman and Rousseeuw, 1990).

Μια διαφορετική τυπική όψη της μεθόδου είναι ότι παρέχει ένα μεγάλο πλήθος στατιστικών με τα οποία μία πλήρης διερεύνηση των αποτελεσμάτων των ομάδων καθίσταται δυνατή. Συγκεκριμένα, άξια αναφοράς είναι τα *medoids* των ομάδων, οι διάμετροι και διαχωρισμοί των ομάδων, και επιπλέον μια γραφική αναπαράσταση των ομάδων σε όρους του λεγόμενου *διαγράμματος silhouettes*.

Αφού βρεθεί το σύνολο των *k medoids*, δημιουργούνται *k* συστάδες εκχωρώντας κάθε παρατήρηση στο πλησιέστερο *medoid*. Στόχος είναι να βρεθούν *k* αντιπροσωπευτικά αντικείμενα τα οποία ελαχιστοποιούν το άθροισμα ανομοιοτήτων των παρατηρήσεων με το πλησιέστερό τους αντιπροσωπευτικό αντικείμενο. Πιο συγκεκριμένα, το αντικείμενο *i*

καταχωρείται στην συστάδα v_i, όταν το *medoid* mv_i είναι πιο κοντά σε σχέση με κάθε άλλο medoid m_w.

$$d(i, mv_i) \leq d(i, m_w) \text{ για κάθε } w = 1,..., k.$$

Τα *k* αντιπροσωπευτικά αντικείμενα που πρέπει να ελαχιστοποιούν την αντικειμενική συνάρτηση, η οποία είναι το άθροισμα ανομοιοτήτων όλων των αντικειμένων με το πλησιέστερο *medoid*.

$$\text{Objective function} = \sum d(i, mv_i)$$

Όταν τα *medoids* δεν δίνονται, ο αλγόριθμος αρχικά αναζητάει ένα καλό αρχικό σύνολο *medoids* (κεντρικά τοποθετημένο - Build Phase). Έπειτα, αναζητείται ένα τοπικό ελάχιστο για την αντικειμενική συνάρτηση, το οποίο είναι μια λύση με χαρακτηριστικό ότι υπάρχει μη μοναδική εναλλαγή των παρατηρήσεων με ένα *medoid* το οποίο μπορεί να μειώσει την απόσταση (Swap Phase). Η εναλλαγή των σημείων επαναλαμβάνεται έως ότου να μην μπορεί να μειωθεί άλλο η συνάρτηση.

Υπάρχουν, βασικά, δύο τρόποι εισαγωγής των στοιχείων. Ο πιο κοινός τρόπος είναι μέσω ενός πίνακα μετρήσεων. Οι γραμμές του πίνακα αντιπροσωπεύουν τα αντικείμενα και οι στήλες αντιστοιχούν στις μεταβλητές οι οποίες πρέπει να είναι σε διαστηματική κλίμακα. Εναλλακτικά, η μέθοδος μπορεί να εφαρμοσθεί εισάγοντας ένα πίνακα ανομοιοτήτων μεταξύ των αντικειμένων. Τέτοιου είδους ανομοιότητες μπορούν να ληφθούν με διάφορους τρόπους. Συχνά υπολογίζονται από μεταβλητές οι οποίες δεν είναι απαραίτητα σε διαστηματική κλίμακα αλλά μπορεί να είναι δίτιμες, διατάξιμες ή ονομαστικές.

Η μέθοδος περιέχει δύο τύπους ‘απομονωμένων ομάδων’. Οι L-Cluster και L*-Cluster. Η ομάδα C είναι μια L-Cluster, εάν για κάθε αντικείμενο *i* που ανήκει στο C ισχύει:

$$\max d_{ij} < \min d_{ih},\ j \in C \qquad h \notin C$$

Η ομάδα C είναι μια L*-Cluster αν,

$$\max d_{ij} < \min d_{lh},\ i,j \in C,\ l \in C,\ h \notin C$$

Είναι φανερό ότι οι L*-Cluster είναι επίσης L-Cluster. Πρέπει να σημειωθεί ότι η ιδιότητα του να είναι απομονωμένη εξαρτάται από την εσωτερική δομή της ομάδας, όπως επίσης και από την θέση της σε σχέση με τις υπόλοιπες συστάδες.

Η διάμετρος της ομάδας C ορίζεται ως η μεγαλύτερη ανομοιότητα μεταξύ αντικειμένων που ανήκουν στην ομάδα C.

$$\text{Diameter}_C = \max d_{ij},\ i, j \in C$$

Ο διαχωρισμός της ομάδας C ορίζεται ως η ελάχιστη ανομοιότητα μεταξύ δύο αντικειμένων, ένα από τα οποία ανήκει στην ομάδα C και τα υπόλοιπα όχι.

$$\text{Separation}_C = \min d_{lh},\ l \in C,\ h \notin C$$

Εάν *j* είναι το *medoid* της ομάδας C, η απόσταση όλων των αντικειμένων της ομάδας C από το *j* υπολογίζεται ως εξής:

$$\text{Average distance}_j = \frac{\sum_{i \in C} d_{ij}}{N_j}$$

όπου N είναι το πλήθος των αντικειμένων μείων τα στοιχεία εκτός του $j^{ου}$.

Εάν *j* είναι το *medoid* της ομάδας C, η μέγιστη απόσταση των αντικειμένων της C από το *j* υπολογίζεται ως εξής,

$$\text{Maximum distance}_j = \max d_{ij},\ i \in C$$

2.1.2.3.1 Silhouette Plot

Η τεχνική *Silhouette* αναφέρεται σε μια μέθοδο ερμηνείας και επικύρωσης των ομάδων δεδομένων. Η τεχνική αυτή παρέχει μια σύντομη γραφική αναπαράσταση του πόσο καλά το κάθε αντικείμενο βρίσκεται εντός του συμπλέγματος των ομάδων. Περιγράφηκε αρχικά από τον Peter J. Rousseeuw το 1986. Κάθε ομάδα αντιπροσωπεύεται από ένα σχήμα, το οποίο δείχνει ποιο αντικείμενο κείτεται μέσα στην ομάδα και ποιο αντικείμενο απλά διατηρεί μια ενδιάμεση θέση. Η συνολική ομαδοποίηση παρατίθεται περιλαμβάνοντας σε ένα γράφημα όλα τα σχήματα, και στο οποίο μπορεί να συγκριθεί η ποιότητα των ομάδων.

Στόχος είναι η εύρεση *k medoids* τα οποία ελαχιστοποιούν το άθροισμα των αποστάσεων των παρατηρήσεων από το πλησιέστερο *medoid*. Ο Rousseeuw (1987) πρότεινε μια γραφική απεικόνιση, το *silhouette plot*, το οποίο μπορεί να χρησιμοποιηθεί (i) για την επιλογή του πλήθους των ομάδων και (ii) για την εκτίμηση του κατά πόσο κάθε παρατήρηση ομαδοποιήθηκε σωστά (S. Dudoit and R. Gentleman, 2002).

Τα σχήματα κατασκευάζονται σύμφωνα με τον ακόλουθο τρόπο. Θεωρείται κάποιο αντικείμενο *i* από το σύνολο των δεδομένων, και έστω A δηλώνει την ομάδα στην οποία έχει εκχωρηθεί, έπειτα υπολογίζεται,

a(i) = μέση ανομοιότητα του *i* ως προς όλα τα υπόλοιπα αντικείμενα του A.

Είναι δυνατόν να χρησιμοποιηθούν οποιαδήποτε μέτρα ανομοιότητας, ωστόσο τα μέτρα απόστασης είναι τα πιο συνήθη. Η ποσότητα *a(i)* μπορεί να ερμηνευθεί ως το πόσο καλά το

αντικείμενο *i* ταιριάζει στη συστάδα στην οποία έχει εκχωρηθεί (όσο πιο μικρή η τιμή, τόσο περισσότερο ταιριάζει στην αντίστοιχη συστάδα). Έπειτα θεωρούμε κάθε συστάδα C διάφορη της A και ορίζουμε,

d(i, C) = μέση ανομοιότητα του *i* ως προς όλα τα υπόλοιπα αντικείμενα του C.

Υπολογίζουμε τα *d(i, C)* για όλες τις ομάδες $C \neq A$ στις οποίες το αντικείμενο *i* δεν ανήκει, και επιλέγουμε το μικρότερο από αυτά.

$$b = min\ d\ (i, C)\ ,\quad C \neq A$$

Η συστάδα με αυτή τη μέση ανομοιότητα θεωρείται ότι είναι η γειτονική συστάδα του *i* καθώς είναι η συστάδα στην οποία ταιριάζει καλύτερα, ανεξάρτητα στην συστάδα στην οποία το *i* έχει εκχωρηθεί.

Έστω B δηλώνει την ομάδα η οποία επιτυγχάνει το ελάχιστο π.χ. *d (i, B) = b (i)*, και είναι γειτονικό του αντικειμένου *i*. Η τιμή *s(i)* μπορεί τώρα να ορισθεί ως εξής:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

και μπορεί να πάρει την ακόλουθη μορφή,

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{αν } \alpha(i) < b(i) \\ 0, & \text{αν } \alpha(i) = b(i) \\ b(i)/a(i) - 1, & \text{αν } \alpha(i) > b(i) \end{cases}$$

Είναι εύκολα φανερό ότι ισχύει $-1 < s(i) < 1$. Η τιμή του *s(i)* μπορεί να ερμηνευθεί ως ακολούθως:

s(i) = 1 (*α(i)<<b(i)*): η ανομοιότητα εντός του *α(i)* είναι πολύ μικρότερη από την ελάχιστη ανομοιότητα μεταξύ των *α(i)*. Με άλλα λόγια, το αντικείμενο *i* έχει εκχωρηθεί σε σωστή ομάδα. Η δεύτερη καλύτερη ομάδα B δεν είναι τόσο κοντά όσο η πραγματική ομάδα A.

s(i) = 0: τα *α(i)* και *b(i)* είναι σχεδόν ίσα. Επιπλέον, δεν είναι ξεκάθαρο αν το *i* θα έπρεπε να εκχωρηθεί στο A ή στο B. Θα μπορούσε να θεωρηθεί ως μια 'ενδιάμεση' περίπτωση και το αντικείμενο βρίσκεται στα σύνορα δύο φυσικών ομάδων.

s(i) = -1: το αντικείμενο *i* δεν έχει διαταχθεί ικανοποιητικά (φτωχή διάταξη). Η ανομοιότητά του με άλλα αντικείμενα της ομάδας που έχει διαταχθεί είναι πολύ μεγαλύτερη από αυτή αντικειμένων της πλησιέστερης ομάδας. Γιατί δεν είναι σε αυτή την ομάδα;

Επιπλέον, το σχήμα μιας ομάδας είναι ένα διάγραμμα των *s(i)* διατεταγμένο σε φθίνουσα διάταξη όλων των αντικειμένων *i*. Το διάγραμμα αποτελείται από μια οριζόντια γραμμή, της οποίας το μήκος είναι ανάλογο του *s(i)*, και δείχνει ποια αντικείμενα έχουν εκχωρηθεί σωστά στην ομάδα και ποια βρίσκονται κάπου μεταξύ των ομάδων. Ένα φαρδύ σχήμα δηλώνει μεγάλες τιμές *s(i)* και συνεπώς μια σωστή ομάδα. Το ύψος μιας ομάδας είναι απλά ίσο με το πλήθος των αντικειμένων της ομάδας.

Το συνολικό γράφημα *silhouette* δηλώνει το σχήμα όλων των ομάδων που βρίσκονται δίπλα η μια στην άλλη, οπότε μπορεί να συγκριθεί η ποιότητα των ομάδων. Το συνολικό μέσο πλάτος του διαγράμματος είναι ο μέσος των *s(i)* ως προς όλα τα αντικείμενα του συνόλου δεδομένων.

Το διάγραμμα *silhouette* είναι ιδιαίτερα χρήσιμο για τον προσδιορισμό του πλήθους των ομάδων. Μπορεί να εφαρμοσθεί η μέθοδος αρκετές φορές για διαφορετικό *k* και έπειτα συγκρίνονται τα αντίστοιχα διαγράμματα *silhouette*. Το μέσο πλάτος μπορεί να χρησιμοποιηθεί για την επιλογή του καλύτερου πλήθους ομάδων, επιλέγοντας αυτό το *k* το οποίο παράγει το μέγιστο πλάτος.

$$SC = \max \bar{s}(k)$$

για k = 2,3,..,n. Ο συντελεστής αυτός είναι ένα αδιάστατο μέτρο της έκτασης της δομής των ομάδων που έχουν ανακαλυφθεί από τον αλγόριθμο. Το SC ερμηνεύεται ως ακολούθως:

0.71-1.0: μια δυνατή δομή έχει βρεθεί

0.51-0.70: μια λογική δομή έχει βρεθεί

0.26-0.50: η δομή είναι ασθενής και μπορεί να είναι επίπλαστη (άλλες μέθοδοι)

$\leq$**0.25**: δεν έχει βρεθεί αξιόλογη δομή

Το μέσο *s(i)* μιας ομάδας είναι ένα μέτρο του πόσο συνεκτικώς ομαδοποιημένα είναι τα δεδομένα της ομάδας. Επομένως, το μέσο *s(i)* του συνόλου των δεδομένων είναι ένα μέτρο του πόσο κατάλληλα τα δεδομένα έχουν ομαδοποιηθεί. Εάν υπάρχουν υπερβολικά πολλές ή υπερβολικά λίγες ομάδες, όπως μπορεί να συμβεί όταν μια φτωχή επιλογή του *k* χρησιμοποιείται στον αλγόριθμο *k-means*, κάποιες από τις ομάδες τυπικά εκθέτουν πολύ στενότερα *silhouettes* σε σχέση με τα υπόλοιπα. Επομένως, τα *silhouette plots* και οι μέσοι μπορούν να χρησιμοποιηθούν για τον καθορισμό του φυσικού πλήθους των ομάδων ενός συνόλου δεδομένων.

2.2 Χάρτης θερμότητας

Ένας *χάρτης θερμότητας* (heat map) είναι μια γραφική αναπαράσταση των δεδομένων, η οποία αποκαλύπτει ταυτόχρονα την δομή της ιεραρχικής ομαδοποίησης ως προς τις γραμμές και τις στήλες ενός πίνακα δεδομένων. Το γράφημα αποτελείται από ορθογώνια πλακίδια, κάθε ένα από τα οποία είναι σκιασμένο σύμφωνα με μία χρωματική κλίμακα προκειμένου να αντιπροσωπεύσουν την τιμή του αντίστοιχου στοιχείου του πίνακα δεδομένων.

Ο Sneath (1957) παρουσίασε τα αποτελέσματα της ανάλυσης συστάδων μεταθέτοντας τις γραμμές και τις στήλες ενός πίνακα έτσι ώστε να τοποθετηθούν οι κοντινές τιμές μαζί σύμφωνα με την ομαδοποίηση. Στο οριζόντιο και κάθετο περιθώριο των πλακιδίων υπάρχου αντίστοιχα *δενδρογράμματα* (L. Wilkinson and M. Friendly (2009)). Η ιδέα της ένωσης δένδρων ως προς τις γραμμές και τις στήλες του πίνακα δεδομένων προέρχεται από τον Robert Ling (1973), ο οποίος χρησιμοποίησε χαρακτήρες εκτύπωσης προκειμένου να αναπαραστήσει διαφορετικές αποχρώσεις του γκρι, έναν χαρακτήρα πλάτους ανά pixel. Ο Leland Wilkinson (1994) ανέπτυξε το πρώτο υπολογιστικό πρόγραμμα το 1994 (SYSTAT) για την παραγωγή χαρτών θερμότητας με χρωματικά γραφικά υψηλής ανάλυσης. Αυτό το είδος γραφήματος είναι μία σύνθεση διαφορετικών γραφικών αναπαραστάσεων που αναπτύχθηκαν από τους στατιστικούς αναλυτές για περισσότερο από έναν αιώνα. Οι πρώτες πηγές εντοπίζονται σε δημοσιεύσεις στα τέλη του 19^{ου} αιώνα. Και έχουν εντοπισθεί, επίσης, ποικίλες στατιστικές μελέτες του 20^{ου} αιώνα οι οποίες παρέχουν μία βάση για αυτή την ευρέως χρησιμοποιούμενη γραφική αναπαράσταση στα πλαίσια της βιοπληροφορικής.

Μετρίου μεγέθους πίνακες δεδομένων (πολλές χιλιάδες γραμμές/στήλες) μπορούν να εμφανίζονται αποτελεσματικά σε έγχρωμες οθόνες υψηλής ανάλυσης και ακόμα μεγαλύτεροι πίνακες μπορούν να απεικονισθούν σε έντυπη μορφή ή σε μορφή εικονοστοιχείων. Τα διαγράμματα θερμότητας είναι γνωστά στις φυσικές επιστήμες και ένα από τα πιο ευρέως χρησιμοποιούμενα γραφήματα στο χώρο των βιολογικών επιστημών.

Υπάρχουν αρκετά διαφορετικά είδη θερμικών χαρτών,

-Οι *Web heatmaps* χρησιμοποιούνται για την παράθεση περιοχών μιας ιστοσελίδας που επισκέπτεται πιο συχνά από τους χρήστες.

-Οι *βιολογικοί heatmaps* χρησιμοποιούνται τυπικά στην μοριακή βιολογία για να αναπαραστήσουν το επίπεδο των εκφράσεων πολλών γονιδίων σε έναν αριθμό συγκρίσιμων

δειγμάτων (για παράδειγμα, κύτταρα σε διαφορετικές καταστάσεις, δείγματα διαφορετικών ασθενών) όπως αυτά λαμβάνονται από τις μικροσυστοιχίες DNA.

- Ο *χάρτης-δένδρο* (tree-map) είναι μια δισδιάστατη ιεραρχική κατάτμηση των δεδομένων που μοιάζει οπτικά με ένα χάρτη θερμότητας.

- Ένα *μωσαϊκό γράφημα* (mosaic plot) είναι ένας χάρτης θερμότητας με πλακίδια για την αναπαράσταση δεδομένων δύο ή περισσότερων διαστάσεων. Όπως με τους χάρτες-δένδρο, οι ορθογώνιες περιοχές ενός μωσαϊκού γραφήματος είναι ιεραρχικά οργανωμένες, που σημαίνει ότι οι περιφέρειες είναι ορθογώνιες αντί τετράγωνες (Ιστοσελίδα 3).

Όσον αφορά την εφαρμογή των *heatmap* στις γονιδιακές εκφράσεις, χρησιμοποιούνται για την εύρεση ομοιοτήτων μεταξύ γονιδίων και μεταξύ δειγμάτων. Οι χάρτες θερμότητας που παράγονται από δεδομένα μικροσυστοιχιών DNA αντανακλούν τιμές γονιδιακών εκφράσεων σε πολλές καταστάσεις. Πιο συγκεκριμένα, κάθε κελί χρωματίζεται σύμφωνα με το επίπεδο έκφρασης του συγκεκριμένου γονιδίου στο συγκεκριμένο δείγμα (Ιστοσελίδα 13). Είναι δε πιο αποτελεσματικοί εάν οι γραμμές και οι στήλες είναι διατεταγμένες έτσι ώστε να επιτρέπεται ο προσδιορισμός αυτών των προτύπων. Η ομαδοποίηση συχνά χρησιμοποιείται ώστε να αναδειχθεί αυτή η διάταξη, ταυτοποιώντας ομάδες δειγμάτων ή γονιδίων και έπειτα διευθετούνται οι ομάδες έτσι ώστε οι κοντινότερες ομάδες να είναι παρακείμενες (T. Speed, 2003). Το χρώμα κάθε κελιού είναι ιδιαίτερης σημασίας και επιλέγεται από κλίμακα χρωμάτων (συνηθέστερα από το κόκκινο στο πράσινο), όπου η σκιά του χρώματος είναι ανάλογη της έντασης (ή του log ratio) του αντίστοιχου γονιδίου του δείγματος. Οι τιμές αυτές συχνά είναι κανονικοποιημένες ή σε κλίμακα των z-scores για καλύτερο αποτέλεσμα. Είναι σύνηθες να αναπαριστώνται υψηλές τιμές ως αποχρώσεις του κόκκινου, ενδιάμεσες τιμές ως αποχρώσεις του γκρι και χαμηλές τιμές ως αποχρώσεις του πράσινου. Αξίζει να σημειωθεί ότι κάποιοι συγγραφείς αντιστοιχίζουν τα χρώματα αντίθετα (J. Copland et al., 2003).

2.3 Δείκτες Ισχύος Ομαδοποίησης

Έχει προταθεί μία σειρά μέτρων με στόχο την αποτίμηση της ισχύος των αποτελεσμάτων της ομαδοποίησης και του προσδιορισμού της μεθόδου η οποία εκτελείται καλύτερα για συγκεκριμένο πείραμα (Kerr and Churchill, 2001; Yeung *et al.*, 2001; Datta and Datta, 2003). Η ισχύς αυτή είναι δυνατό να βασισθεί αποκλειστικά στις εσωτερικές ιδιότητες των δεδομένων, σε κάποιες εξωτερικές αναφορές, αποκλειστικά στα δεδομένα γονιδιακής

έκφρασης ή σε συνδυασμό με σχετικές βιολογικές πληροφορίες (Gibbons and Roth, 2002; Gat-Viks *et al.*, 2003; Bolshakova *et al.*, 2005; Datta and Datta, 2006). Τρία μέτρα ισχύος της ομαδοποίησης διακρίνονται στις γενικές κατηγορίες τα οποία είναι τα '*internal*', '*stability*' και '*biological*'. Τα μέτρα *internal* λαμβάνουν μόνο το σύνολο των δεδομένων και τον διαχωρισμό της ομαδοποίησης ως δεδομένα και χρησιμοποιούν την εγγενή πληροφορία των δεδομένων για την ανάδειξη της ποιότητάς τους. Τα μέτρα *stability* είναι ειδική περίπτωση των εσωτερικών μέτρων. Αξιολογούν την σταθερότητα της ομαδοποίησης συγκρίνοντας τις ομάδες που προκύπτουν αφαιρώντας κάθε φορά μία στήλη με τα πλήρη δεδομένα. Τα μέτρα *biological* αξιολογούν την ικανότητα του αλγορίθμου ομαδοποίησης να παράγει βιολογικώς λογικές ομάδες. Διατίθενται μέτρα για την αξιολόγηση τόσο της βιολογικής ομοιογένειας όσο και σταθερότητας των αποτελεσμάτων ομαδοποίησης.

2.3.1 Internal Measures

Αναφορικά με την εσωτερική ισχύ, τα ακόλουθα μέτρα αντανακλούν την πυκνότητα (compactness), την συνεκτικότητα (connectedness) και τον διαχωρισμό (separation) των ομάδων. Η συνεκτικότητα συνδέεται με την έκταση στην την οποία οι παρατηρήσεις αντικαθίστανται στην ίδια ομάδα όπως οι κοντινότεροι γείτονες στον χώρο των δεδομένων, και μετράται με τον δείκτη *connectivity* (Handl *et al.*, 2005). Η πυκνότητα μετρά την ομοιογένεια της ομάδας, συνήθως εστιάζοντας στην εντός των ομάδων διακύμανση, ενώ ο διαχωρισμός ποσοτικοποιεί τον βαθμό διαχωρισμού μεταξύ ομάδων, συνήθως μετρώντας την απόσταση μεταξύ των κέντρων των ομάδων. Καθώς η πυκνότητα και ο διαχωρισμός ερευνούν αντίθετες τάσεις (η πυκνότητα αυξάνεται με το πλήθος των ομάδων ενώ ο διαχωρισμός μειώνεται), δημοφιλείς μέθοδοι συνδυάζουν τα δύο μέτρα σε ένα μοναδικό σκορ. Οι δείκτες Dunn (Dunn, 1974) και Silhouette Width (Rousseeuw, 1987) είναι δύο παραδείγματα μη γραμμικών συνδυασμών της πυκνότητας και του διαχωρισμού, και με την συνεκτικότητα αποτελούν τρία εσωτερικά μέτρα.

Connectivity

Έστω N δηλώνει το συνολικό πλήθος των παρατηρήσεων (γραμμές) στο σύνολο των δεδομένων και G δηλώνει το σύνολο των στηλών. Ορίζουμε ως $nn_{i(j)}$ το j^o κοντινότερο στοιχείο της παρατήρησης i, και έστω $x_{i,nn_{i(j)}}$ ισούται με μηδέν εάν τα i και j είναι στην ίδια

συστάδα και 1/*j* σε διαφορετική περίπτωση. Τότε, για μία συγκεκριμένη κατάτμηση $C = \{C_1,.., C_K\}$ των N παρατηρήσεων στις *k* ασύνδετες συστάδες, ο δείκτης *connectivity* ορίζεται ως

$$\text{Conn}(C) = \sum_{i=1}^{N} \sum_{j=1}^{G} x_{i,nn_{i(j)}} .$$

Ο *connectivity* λαμβάνει τιμή μεταξύ 0 και ∞ και πρέπει να είναι ελάχιστο.

Silhouette Width

Ο δείκτης Silhouette Width είναι ο μέσος της τιμής Silhouette κάθε παρατήρησης. Ο δείκτης αυτός μετρά τον βαθμό εμπιστοσύνης στην διαμόρφωση της ομαδοποίησης μιας συγκεκριμένης παρατήρησης, με τις καλώς ομαδοποιημένες παρατηρήσεις να έχουν τιμή κοντά στο 1 και τις ανεπαρκώς ομαδοποιημένες παρατηρήσεις να έχουν τιμή κοντά στο -1. Για την παρατήρηση *i*, ορίζεται ως

$$S(i) = \frac{b_i - a_i}{\max(b_i, a_i)},$$

όπου a_i είναι η μέση απόσταση μεταξύ του *i* και όλων των υπολοίπων παρατηρήσεων της ίδιας ομάδας, και b_i είναι η μέση απόσταση μεταξύ του *i* και των παρατηρήσεων της πλησιέστερης ομάδας (βλέπε παράγραφο 2.1.2.3). Η τιμή του δείκτη ανήκει στο διάστημα [-1, 1], και πρέπει να είναι μέγιστη.

Dunn Index

Ο δείκτης *Dunn* είναι ο λόγος της ελάχιστης απόστασης μεταξύ των παρατηρήσεων που δεν ανήκουν στην ίδια ομάδα προς την μεγαλύτερη απόσταση εντός της ομάδας. Ορίζεται ως

$$D(C) = \frac{\min_{C_r, C_l \in C, C_r \neq C_l} \left(\min_{i \in C_r, j \in C_l} dist(i, j) \right)}{\max_{C_m \in C} diam(C_m)},$$

όπου $diam(C_m)$ είναι η μέγιστη απόσταση μεταξύ των στοιχείων της ομάδας C_m. Ο *Dunn Index* λαμβάνει τιμή μεταξύ μηδέν και ∞, και πρέπει να είναι μέγιστο.

2.3.2 Stability Measures

Τα μέτρα σταθερότητας συγκρίνουν τα αποτελέσματα από την ομαδοποίηση που βασίζεται στο σύνολο των δεδομένων με την ομαδοποίηση που βασίζεται στην αφαίρεση μίας στήλης, κάθε φορά. Οι δείκτες αυτοί λειτουργούν ιδιαίτερα αποτελεσματικά στην περίπτωση που τα δεδομένα είναι υψηλά συσχετισμένα, κάτι που ισχύει στα γονιδιακά δεδομένα. Τα μέτρα είναι τα *average proportion of non-overlap (APN), average distance (AD), average distance between means (ADM)* και *figure of merit (FOM)* (Datta and Datta, 2003; Yeung *et al.*, 2001). Σε κάθε περίπτωση, ο μέσος λαμβάνεται για όλες τις διαγραμμένες στήλες, και όλα τα μέτρα πρέπει να ελαχιστοποιηθούν.

Average Proportion of Non-overlap (APN)

Ο δείκτης *APN* μετρά την μέση αναλογία των παρατηρήσεων που δεν τοποθετούνται στην ίδια ομάδα, ομαδοποιώντας βασιζόμενοι στα πλήρη δεδομένα και στα δεδομένα με αφαίρεση μίας στήλης. Έστω $C^{i,0}$ αντιπροσωπεύει την ομάδα που περιέχει την παρατήρηση i χρησιμοποιώντας τα αρχικά δεδομένα (σύμφωνα με όλα τα διαθέσιμα δεδομένα), και $C^{i,l}$ αντιπροσωπεύει την ομάδα που περιέχει την παρατήρηση i όταν η ομαδοποίηση βασίζεται στα δεδομένα από τα οποία έχει αφαιρεθεί η l στήλη. Τότε, θεωρώντας συνολικό πλήθος ομάδων k, ο δείκτης *APN* ορίζεται ως

$$APN(k) = \frac{1}{GN}\sum_{i=1}^{N}\sum_{j=1}^{G}\left(1 - \frac{n\left(C^{i,l} \cap C^{i,0}\right)}{n\left(C^{i,0}\right)}\right).$$

Ο *APN* ανήκει στο διάστημα [0,1], με τιμές κοντά στο 0 να αντιστοιχούν σε αποτελέσματα υψηλής σταθερότητας.

Average Distance (AD)

Ο δείκτης *AD* μετρά την μέση απόσταση μεταξύ των παρατηρήσεων που τοποθετούνται στην ίδια ομάδα, ομαδοποιώντας σύμφωνα με τα πλήρη δεδομένα και αυτά που προκύπτουν αφαιρώντας μία στήλη. Ορίζεται ως

$$AD(k) = \frac{1}{GN}\sum_{i=1}^{N}\sum_{l=1}^{G}\frac{1}{n\left(C^{i,0}\right)n\left(C^{i,l}\right)}\left[\sum_{i\in C^{i,0}, j\in C^{i,l}} dist(i,j)\right].$$

Ο δείκτης *AD* λαμβάνει τιμή μεταξύ 0 και ∞, και οι χαμηλότερες τιμές προτιμούνται.

Average Distance between Means (ADM)

Ο δείκτης *ADM* υπολογίζει την μέση απόσταση μεταξύ των κέντρων των ομάδων για τις παρατηρήσεις που τοποθετούνται στην ίδια ομάδα ομαδοποιώντας τα πλήρη δεδομένα και αυτά που προκύπτουν αφαιρώντας κάθε φορά μία στήλη. Ορίζεται ως

$$ADM(k) = \frac{1}{GN}\sum_{i=1}^{N}\sum_{j=1}^{G} dist(\bar{x}_{C^{i,l}}, \bar{x}_{C^{i,0}}),$$

όπου $\bar{x}_{C^{i,0}}$ είναι ο μέσος των παρατηρήσεων της ομάδας που περιέχει την παρατήρηση *i*, όταν η ομαδοποίηση βασίζεται στα πλήρη δεδομένα, και $\bar{x}_{C^{i,l}}$ ορίζεται αναλόγως. Ο δείκτης αυτός επίσης λαμβάνει τιμή μεταξύ 0 και ∞, και οι χαμηλότερες τιμές είναι οι επιθυμητές.

Figure of Merit (FOM)

Ο δείκτης *FOM* μετρά τη μέση διακύμανση εντός των ομάδων των παρατηρήσεων της διαγραμμένης στήλης, όπου η ομαδοποίηση βασίζεται στα δείγματα που παραμένουν. Εκτιμάται το μέσο σφάλμα χρησιμοποιώντας προβλέψεις σύμφωνα με τις μέσες τιμές των ομάδων. Για μία συγκεκριμένη στήλη που μένει εκτός ομαδοποίησης (left-out column) *l*, ο δείκτης *FOM* είναι,

$$FOM(l,k)\sqrt{\frac{1}{N}\sum_{r=1}^{k}\sum_{i\in C_r(l)} dist\left(c, \bar{x}_{C_r(l)}\right)},$$

όπου $x_{i,l}$ είναι η τιμή της $i^{ης}$ παρατήρησης της $l^{ης}$ στήλης στην ομάδα $C_r(l)$, και $\bar{x}_{C_r(l)}$ είναι ο μέσος της συστάδας $C_r(l)$. Ο δείκτης *FOM* πολλαπλασιάζεται με ένα παράγοντα προσαρμογής $\sqrt{\frac{N}{N-k}}$, για την άμβλυνση την τάση μείωσης του δείκτη καθώς το πλήθος των ομάδων αυξάνεται. Η τελική τιμή προκύπτει από τον μέσο όρο όλων των διαγεγραμμένων στηλών, και λαμβάνει τιμή μεταξύ 0 και ∞, με τις χαμηλότερες τιμές να σημαίνουν καλύτερη απόδοση.

2.3.3 Biological Measures

Η *βιολογική ισχύς* (Biological validation) αξιολογεί την ικανότητα του αλγορίθμου ομαδοποίησης να παράγει ομάδες που να έχουν λογικό νόημα από βιολογικής σκοπιάς. Συνήθης εφαρμογή των βιολογικών δεικτών είναι στα δεδομένα γονιδιακών εκφράσεων.

Υπάρχουν δύο διαθέσιμε μέτρα, τα οποία είναι τα *Biological Homogeneity Index (BHI)* και *Biological Stability Index (BSI)* (Datta and Datta, 2006).

Biological Homogeneity Index (BHI)

Όπως προκύπτει από την ονομασία του, ο *BHI* μετρά την βιολογική ομοιογένεια των ομάδων. Έστω $B = \{B_1,..,B_F\}$ είναι ένα σύνολο από *F* λειτουργικές κλάσεις, όχι απαραίτητα ξένες μεταξύ τους, και έστω *B(i)* είναι η λειτουργική κλάση που περιέχει το γονίδιο *i* (πιθανώς περισσότερες της μίας κλάσης να περιέχουν το γονίδιο *i*). Ομοίως, ορίζεται ως *B(j)* η λειτουργική κλάση που περιέχει το γονίδιο *j*, και αντιστοιχίζεται στην δείκτρια συνάρτηση *I* (*B(i) = B(j)*) η τιμή 1 αν τα *B(i)* και *B(j)* ταιριάζουν (κάθε ταίριασμα γονιδίων είναι επαρκές στην περίπτωση που συμμετέχουν σε πολλαπλές λειτουργικές κλάσεις),και 0 διαφορετικά. Διαισθητικά, επιθυμούμε τα γονίδια που τοποθετούνται μαζί στις ίδιες στατιστικές ομάδες να ανήκουν, επίσης, στις ίδιες λειτουργικές κλάσεις. Οπότε, για μία δεδομένη στατιστική ομαδοποίηση $C = \{C_1,..C_k\}$ και δεδομένο σύνολο βιολογικών κλάσεων *B*, ο δείκτης *BHI* ορίζεται ως

$$BHI(C,B) = \frac{1}{k}\sum_{r=1}^{k}\frac{1}{n_r(n_r-1)}\sum_{i\neq j\in C_r} I\big(B(i)=B(j)\big).$$

Εδώ, $n_r = n(C_r \cap B)$ είναι το πλήθος των περιγραφόμενων γονιδίων στην στατιστική ομάδα C_r. Ο *BHI* ανήκει στο διάστημα [0,1], με υψηλότερες τιμές να αντιστοιχούν σε βιολογικώς περισσότερο ομοιογενείς ομάδες.

Biological Stability Index (BSI)

Ο *BSI* είναι παρόμοιος με άλλα μέτρα σταθερότητας, και ελέγχει την συνοχή της ομαδοποίησης για τα γονίδια με όμοια βιολογική λειτουργία. Κάθε δείγμα απομακρύνεται, ένα κάθε φορά, και τα μέλη της ομάδας για τα γονίδια με όμοια λειτουργική περιγραφή συγκρίνονται με τα μέλη της ομάδας χρησιμοποιώντας τα πλήρη δεδομένα. Ο *BSI* ορίζεται ως

$$BSI(C,B) = \frac{1}{F}\sum_{r=1}^{F}\frac{1}{n(B_r)(n(B_r)-1)G}\sum_{l=1}^{G}\sum_{i\neq j\in B_k}\frac{n\left(C^{i,0}\cap C^{j,l}\right)}{n\left(C^{i,0}\right)},$$

όπου *F* είναι το συνολικό πλήθος των λειτουργικών κλάσεων, $C^{i,0}$ είναι η στατιστική ομάδα που περιέχει την παρατήρηση *i* βασιζόμενη στα πλήρη δεδομένα, και $C^{j,l}$ είναι η στατιστική ομάδα που περιέχει την παρατήρηση *j* όταν η στήλη *l* απομακρύνεται. Ο *BSI* λαμβάνει τιμές

στο διάστημα [0, 1], με τις υψηλότερες τιμές να αντιστοιχούν σε περισσότερο σταθερές ομάδες των λειτουργικά περιγραφόμενων γονιδίων.

2.4 Ανάλυση κατά παράγοντες

Μία μέθοδος για την μείωση του πλήθους των άγνωστων παραμέτρων, μέσω ενός πίνακα συνιστωσών – συνδιακυμάνσεων είναι η *Ανάλυση κατά παράγοντες*, η οποία μοντελοποιεί την δομή της συνδιακύμανσης πολυδιάστατων δεδομένων κάνοντας χρήση μικρού πλήθους λανθανουσών μεταβλητών και ονομάζονται *παράγοντες* (G. McLachlan *et al.*, 2004). Ουσιώδης σκοπός της μεθόδου είναι η περιγραφή, εάν αυτό καθίσταται δυνατόν, των σχέσεων συσχέτισης που σημειώνονται μεταξύ πολλών μεταβλητών, μέσω των παραγόντων. Με άλλα λόγια, είναι δυνατόν, για παράδειγμα, οι διακυμάνσεις τριών ή τεσσάρων *παρατηρούμενων* (observed) μεταβλητών να αντικατοπτρίζουν κυρίως τις διακυμάνσεις σε μία και μόνο *μη παρατηρούμενη* (unobserved) μεταβλητή, ή σε πλήθος μη παρατηρούμενων μεταβλητών μειωμένο συγκριτικά με το πλήθος παρατηρούμενων μεταβλητών (Johnson and Wichern, 1998).

Η *ανάλυση κατά παράγοντες* αναζητά συνδυασμένες μεταβολές σε αντιστοιχία με τις μη παρατηρούμενες λανθάνουσες μεταβλητές. Βασικά, το παραγοντικό μοντέλο έχει ως κίνητρο την εξής παραδοχή. Έστω ότι οι μεταβλητές μπορούν να ομαδοποιηθούν σύμφωνα με τις συσχετίσεις τους. Αυτό συνεπάγεται ότι όλες οι μεταβλητές που βρίσκονται σε μία ομάδα είναι υψηλά συσχετισμένες μεταξύ τους, αλλά έχουν σχετικά χαμηλές συσχετίσεις με μεταβλητές που ανήκουν σε διαφορετικές ομάδες. Συνεπώς, γίνεται αντιληπτό, ότι κάθε ομάδα μεταβλητών αντιπροσωπεύει μία μοναδική βαθύτερη δομή, ή ένα πλήθος παραγόντων, που ευθύνονται για τις παρατηρούμενες συσχετίσεις. Οι παρατηρούμενες μεταβλητές μοντελοποιούνται ως γραμμικοί συνδυασμοί των δυνητικών παραγόντων, συμπεριλαμβανομένου και ενός όρου σφάλματος. Οι πληροφορίες που έχουν προκύψει σχετικά με τις αλληλεξαρτήσεις μεταξύ μεταβλητών που παρατηρήθηκαν μπορεί να χρησιμοποιηθούν αργότερα για τη μείωση του συνόλου των μεταβλητών στο σύνολο των δεδομένων.

Η ανάλυση κατά παράγοντες εφαρμόστηκε αρχικά στον κλάδο της ψυχολογίας. Ωστόσο, η μέθοδος έγινε ιδιαίτερα δημοφιλής σε επιπρόσθετα επιστημονικά πεδία από το 1950, σε συνδυασμό με την έλευση των ηλεκτρονικών υπολογιστών. Στα πεδία αυτά περιλαμβάνονται,

μεταξύ άλλων, η μετεωρολογία και η ιατρική, η πολιτική επιστήμη και η ταξινομία, η αρχαιολογία και τα οικονομικά, καθώς και οι κοινωνικές επιστήμες (H. Harman, 1913).

Αξιοσημείωτη είναι η καταλληλότητα της παραγοντικής ανάλυσης στην ανάλυση γονιδιακών εκφράσεων. Τα γονίδια μεταγράφονται σε mRNA τα οποία στην συνέχεια μεταφράζονται σε πρωτεΐνες. Κάποιες από αυτές τις πρωτεΐνες ενεργοποιούν ή αναστέλλουν, ως μεταγραφικοί παράγοντες, την μεταγραφή πλήθους άλλων γονιδίων δημιουργώντας ένα σύνθετο δίκτυο κανονιστικών γονιδίων. Το πλήθος των μεταγραφικών παραγόντων είναι πολύ μικρότερο του πλήθους των μεταγραφόμενων γονιδίων και η πλειοψηφία των γονιδίων ρυθμίζονται μόνο από ένα μικρό πλήθος μεταγραφικών παραγόντων. Επιπλέον, ο πίνακας που περιλαμβάνει τις συνδέσεις μεταξύ των μεταγραφικών παραγόντων και των ρυθμιστικών παραγόντων είναι αραιός. Κάνοντας χρήση των μικροσυστοιχιών, τα επίπεδα mRNA χιλιάδων γονιδίων είναι δυνατόν να μετρηθούν ταυτοχρόνως, αλλά δεν λαμβάνεται άμεση πληροφορία ως προς την δραστηριότητα των μεταγραφικών παραγόντων. Στόχος είναι η αναδόμηση των μεταγραφικών παραγόντων (Pournara and Wernisch, 2007).

2.4.1 Είδη ανάλυσης κατά παράγοντες

Η *Διερευνητική ανάλυση κατά παράγοντες (Exploratory factor analysis-EFA*) χρησιμοποιείται για να αποκαλύψει τη βασική δομή ενός σχετικά μεγάλου συνόλου μεταβλητών. Η εκ των προτέρων υπόθεση του ερευνητή είναι ότι κάθε δείκτης μπορεί να σχετίζεται με κάθε παράγοντα. Αυτή είναι η πιο κοινή μορφή της ανάλυσης κατά παράγοντες. Δεν υπάρχει καμία προηγούμενη θεωρία και χρησιμοποιούνται τα φορτία των παραγόντων προκειμένου να υπάρχει κάποια διαίσθηση για τη παραγοντική δομή των δεδομένων.

Σύμφωνα με την *Επιβεβαιωτική ανάλυση κατά παράγοντες (Confirmatory factor analysis - CFA)*, ζητείται να προσδιοριστεί αν ο αριθμός των παραγόντων και τα φορτία των μετρούμενων μεταβλητών συμφωνούν με ό,τι αναμένεται επί τη βάσει προκαθορισμένης θεωρίας. Ενδεικτικές μεταβλητές επιλέγονται σύμφωνα με την προηγούμενη θεωρία και η ανάλυση παραγόντων χρησιμοποιείται για να διαπιστωθεί εάν ‘φορτώνεται’ όπως προβλέπεται από το αναμενόμενο πλήθος των παραγόντων (Ιστοσελίδα 2).

2.4.2 Ορολογία

Παραγοντικά Φορτία (factor loadings)

Τα παραγοντικά φορτία είναι οι συντελεστές συσχέτισης μεταξύ των μεταβλητών (γραμμές) και των συντελεστών (στήλες). Αντίστοιχα του συντελεστή *r* του Pearson, το τετράγωνο των παραγοντικών φορτίων είναι το ποσοστό της διασποράς σε αυτή τη μεταβλητή-δείκτη που εξηγείται από τον παράγοντα. Για να ληφθεί το ποσοστό της διασποράς από όλες τις μεταβλητές που αναλογούν σε κάθε παράγοντα, προστίθεται το άθροισμα των τετραγώνων των φορτίων για τον εν λόγω παράγοντα (στήλη) και διαιρείται με τον αριθμό των μεταβλητών. (Σημειώνεται ότι το πλήθος των μεταβλητών ισούται με το άθροισμα των διακυμάνσεών τους καθώς η διακύμανση μιας τυποποιημένης μεταβλητής είναι 1). Το παραπάνω ισοδυναμεί με τη διαίρεση της ιδιοτιμής του παράγοντα με το πλήθος των μεταβλητών.

Ερμηνεία Παραγοντικών Φορτίων (Interpreting factor loadings)

Σύμφωνα με τον κανόνα του αντίχειρα, στην επιβεβαιωτική παραγοντική ανάλυση, τα φορτία πρέπει να έχουν τιμή ίση με 0,7 ή υψηλότερη για να επιβεβαιωθεί ότι οι ανεξάρτητες μεταβλητές που προσδιορίζονται εκ των προτέρων εκπροσωπούνται από τον συγκεκριμένο παράγοντα, με το σκεπτικό ότι το επίπεδο 0,7 αντιστοιχεί στο ήμισυ περίπου της διακύμανσης του δείκτη που εξηγείται από τον παράγοντα. Ωστόσο, το 0,7 είναι ένα υψηλό πρότυπο και τα πραγματικά δεδομένα μπορεί να μην πληρούν το κριτήριο αυτό. Για το λόγο αυτό ορισμένοι ερευνητές, ιδίως για λόγους ανίχνευσης, χρησιμοποιούν ένα χαμηλότερο επίπεδο, όπως 0,4 για τον κεντρικό παράγοντα και 0,25 για άλλους παράγοντες, θεωρώντας φορτία πάνω από το 0.6 "υψηλά" και εκείνα κάτω του 0,4 "χαμηλά". Σε κάθε περίπτωση, τα παραγοντικά φορτία πρέπει να ερμηνεύονται υπό το πρίσμα της θεωρίας, παρά με αυθαίρετο επίπεδο αποκοπής.

Communality (h^2)

Το άθροισμα των τετραγώνων των φορτίων όλων των παραγόντων για μια συγκεκριμένη μεταβλητή (γραμμή) είναι η διακύμανση της εν λόγω μεταβλητής που προέρχεται από όλους τους παράγοντες, και ονομάζεται *communality*. Το *communality* μετρά το ποσοστό της διασποράς σε μια δεδομένη μεταβλητή που ερμηνεύεται από όλους τους παράγοντες από

κοινού και μπορεί να ερμηνευθεί ως η αξιοπιστία του δείκτη. Αν το *communality* υπερβαίνει το 1, υπάρχει μια ψευδεπίγραφη λύση, η οποία μπορεί να αντανακλά ένα δείγμα ως πολύ μικρό ή ο ερευνητής έχει πάρα πολλούς ή πολύ λίγους παράγοντες.

Μοναδικότητα μιας μεταβλητής (Uniqueness of a variable - $1\text{-}h^2$)

Η μοναδικότητα είναι η μεταβλητότητα της μεταβλητής μείον το αντίστοιχο *communality*.

Ιδιοτιμές / Χαρακτηριστικές ρίζες (Eigenvalues/Characteristic roots)

Η ιδιοτιμή για ένα δοσμένο παράγοντα μετρά τη διακύμανση όλων των μεταβλητών η οποία υπολογίζεται από αυτόν τον παράγοντα. Ο λόγος των ιδιοτιμών είναι ο λόγος της αιτιολογικής σπουδαιότητας των παραγόντων σε σχέση με τις μεταβλητές. Εάν ένας παράγοντας έχει χαμηλή ιδιοτιμή, τότε συμβάλλει ελάχιστα στην ερμηνεία των διακυμάνσεων στις μεταβλητές και μπορεί να αγνοηθεί ως περιττός ώστε να εστιάσουμε σε πιο σημαντικούς παράγοντες. Οι ιδιοτιμές μετρούν τη ποσότητα της μεταβλητότητας στο συνολικό δείγμα που αναλογεί σε καθένα παράγοντα.

Παραγοντικά score (factor scores)

Τα παραγοντικά *score* είναι τα αποτελέσματα από κάθε μεταβλητή (γραμμή) και για κάθε παράγοντα (στήλη). Αποτελούν τις εκτιμήσεις των κοινών παραγόντων. Τα παραγοντικά *score* δεν εκτιμούνται από άγνωστες παραμέτρους με την συνήθη έννοια. Για να υπολογιστούν τα παραγοντικά *score* για μια συγκεκριμένη περίπτωση (πχ. για ένα ασθενή) και για ένα συγκεκριμένο παράγοντα, λαμβάνεται το τυποποιημένο *score* του ατόμου (περίπτωση) για κάθε μεταβλητή, πολλαπλασιάζεται με το αντίστοιχο φορτίο της μεταβλητής για το συγκεκριμένο παράγοντα, και αθροίζονται τα γινόμενα αυτά. Ο υπολογισμός των *scores* επιτρέπει τον εντοπισμό έκτροπων παραγόντων. Οι ποσότητες αυτές συχνά χρησιμοποιούνται για διαγνωστικούς σκοπούς, όπως επίσης και ως μεταβλητές σε μεταγενέστερη ανάλυση (Ιστοσελίδα 2).

Στην ανάλυση γονιδιακών εκφράσεων που λαμβάνει χώρα στην παρούσα εργασία, πραγματοποιείται η ανάλυση κατά παράγοντες εφαρμόζοντας την μέθοδο *Principal Component* και περιστρέφοντας τους παράγοντες στην τελική λύση μέσω της ορθογώνιας περιστροφής *Quartimax* και ικανοποιώντας το κριτήριο του *Kaiser*. Οι παραπάνω επιλογές προκύπτουν από την βιβλιογραφία καθώς έχουν εφαρμοσθεί σε προηγούμενες έρευνες

ανάλογων δεδομένων και περιγράφονται στην συνέχεια. Ενδεικτικά αναφέρονται οι εξής πηγές, Peterson (2002), Wang *et al.* (2010), Wang *et al.* (2006), Zitzmann *et al.* (2003), Schmidt *et al.* (1998), Figueroa *et al.* (2003).

2.4.3 Ορισμός του μοντέλου

Ας υποθέσουμε ότι έχουμε ένα σύνολο από p παρατηρήσιμες τυχαίες μεταβλητές, X_1, X_2,..,X_p με μέσους μ_1, μ_2,..,μ_p, αντίστοιχα. Ας υποθέσουμε ότι για κάποιες άγνωστες σταθερές l_{ij}, οι οποίες αποτελούν τα φορτία, και k μη παρατηρούμενες τυχαίες μεταβλητές F_j (*κοινοί παράγοντες*), όπου $i \in 1,..,p$ και $j \in 1,..,k$, όπου k < p, έχουμε,

$$X_i - \mu_i = l_{i1}F_1 + \ldots + l_{ik}F_k + \varepsilon_i.$$

Στην *ανάλυση κατά παράγοντες*, κάθε μεταβλητή μοντελοποιείται ως γραμμικός συνδυασμός k < p μη παρατηρηθέντων κοινών παραγόντων συν ένα μη παρατηρηθέν ειδικό σφάλμα, το οποίο μπορεί να περιέχει εκτός από το σφάλμα παρατήρησης και άλλους ειδικούς όρους που αφορούν μόνο στην συγκεκριμένη μεταβλητή. Εδώ, τα ε_i είναι ανεξάρτητα κατανεμημένοι όροι σφάλματος (*ειδικός παράγοντας*) με μηδενική μέση τιμή και πεπερασμένη διακύμανση, η οποία δεν μπορεί να είναι κοινή για όλα τα i. Σημειώνεται, επίσης, ότι ο $i^{ος}$ ειδικός όρος ε_i συνδέεται μόνο με την $i^{η}$ απόκριση X_i. Έστω $Var(\varepsilon_i) = \psi_i$, έτσι ώστε να έχουμε,

$$Cov(\varepsilon) = Diag(\psi_1,..,\psi_p) = \psi \text{ και } E(e) = 0.$$

Σε όρους μητρών έχουμε,

$$\underset{(p\times 1)}{\mathbf{X}} - \boldsymbol{\mu} = \underset{(p\times k)}{\mathbf{L}}\ \underset{(k\times 1)}{\mathbf{F}} + \underset{(p\times 1)}{\boldsymbol{\varepsilon}}\ .$$

Η μέθοδος αυτή υποθέτει ότι υπάρχουν κάποιοι κοινοί παράγοντες που ερμηνεύουν τα δεδομένα. Συνεπώς κρίνεται απαραίτητο οι μεταβλητές να έχουν μεγάλες συσχετίσεις και κάθε ζεύγος μεταβλητών να έχει μικρή συσχέτιση όταν σταθεροποιηθούν οι υπόλοιπες, δηλαδή μικρό μερικό συντελεστή συσχέτισης, έτσι ώστε όταν εξασφαλίζεται η επίδραση των υπόλοιπων μεταβλητών να εξασφαλίζεται και η επίδραση των κοινών παραγόντων.

Επιπλέον, θέτονται οι εξής παραδοχές σχετικά με τις **F**.

1. **F** και **ε** είναι ανεξάρτητα, οπότε Cov(**F**, **ε**) = E(**εF**΄) = **0**.
2. E (**F**) = **0** και Cov(**F**) = **I**.
3. E(**ε**) = **0** και Cov (**ε**) = E[**εε**΄] = **Ψ** (όπου **Ψ** διαγώνιος πίνακας)

Οποιαδήποτε λύση του παραπάνω συνόλου εξισώσεων ικανοποιεί τους περιορισμούς για την **F**, ορίζεται ως παράγοντας, και η **L** ως η μήτρα φορτίων. Ας υποθέσουμε ότι Cov (**X**) = **Σ**. Από τις παραπάνω υποθέσεις που τέθηκαν για τις **F** προκύπτει,

$$\mathbf{\Sigma} = \mathbf{LL}' + \mathbf{\Psi}.$$

Από την τελευταία σχέση, η ποσότητα **LL΄** είναι το μέρος της διασποράς που οφείλεται στους κοινούς παράγοντες (*communality* - $\hat{h}_i$) και **Ψ** είναι το μέρος της διασποράς που οφείλεται στον ειδικό παράγοντα. Το παραπάνω μοντέλο είναι γραμμικό ως προς τους κοινούς παράγοντες. Στην περίπτωση κατά την οποία οι *p* αποκρίσεις **X** συνδέονται με τους βασικούς παράγοντες, αλλά η σχέση δεν είναι γραμμική, τότε η δομή των συσχετίσεων **Σ** = **LL΄** + **Ψ** δεν είναι επαρκής. Η πολύ σημαντική υπόθεση της γραμμικότητας είναι ενυπάρχουσα με την διατύπωση του κλασικού παραγοντικού μοντέλου.

Το παραγοντικό μοντέλο υποθέτει ότι οι $p + p(p - 1)/2 = p(p + 1)/2$ διακυμάνσεις και συνδιακυμάνσεις για τον **X** μπορούν να αναπαραχθούν από τα *pm* φορτία l_{ij} και τις *p* ειδικές διακυμάνσεις ψ_i. Όταν $m = p$, κάθε πίνακας συσχετίσεων **Σ** μπορεί να αναπαραχθεί ως **LL΄**, οπότε το **Ψ** είναι πλέον ο μηδενικός πίνακας. Παρ' όλα αυτά, στην περίπτωση όπου το *m* είναι μικρό σε σχέση με το *p* αυτή η μέθοδος είναι περισσότερο επωφελής. Αυτό προκύπτει από το γεγονός ότι το μοντέλο αυτό παρέχει μια 'απλή' ερμηνεία των συσχετίσεων των **X** με λιγότερες παραμέτρους από τις $p(p + 1)/2$ παραμέτρους του **Σ** (Johnson and Wichern, 1998).

2.4.4 Μέθοδοι ανάλυσης κατά παράγοντες

Δεδομένων των παρατηρήσεων $\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_n$ των *p* συσχετισμένων μεταβλητών, η ανάλυση κατά παράγοντες επιδιώκει να απαντήσει στο ερώτημα εάν το παραπάνω παραγοντικό μοντέλο, με μικρό πλήθος παραγόντων, αντιπροσωπεύει επαρκώς τα δεδομένα. Κατ' ουσίαν, αυτό το πρόβλημα μοντελοποίησης αντιμετωπίζεται προσπαθώντας να επαληθευθούν οι σχέσεις συσχέτισης **Σ = LL΄+ Ψ** και Cov(**X, Ψ**) = **L**. Ο δειγματικός πίνακας συνδιακύμανσης **S** είναι ένας εκτιμητής του πίνακα συνδιακύμανσης **Σ** του άγνωστου πληθυσμού. Εάν τα μη διαγώνια στοιχεία του **S** είναι μικρά ή εκείνα του δειγματικού πίνακα συσχέτισης **R** είναι σχεδόν μηδέν, οι μεταβλητές δεν σχετίζονται, και η παραγοντική ανάλυση δεν επαληθεύεται ως χρήσιμη. Υπό αυτές τις συνθήκες, ο ειδικός παράγοντας κατέχει τον κυρίαρχο ρόλο, ενώ ο κύριος στόχος της παραγοντικής ανάλυσης είναι ο προσδιορισμός κάποιων σημαντικών κοινών παραγόντων.

Εάν ο **Σ** αποκλίνει σημαντικά από ένα διαγώνιο πίνακα, τότε ένα παραγοντικό μοντέλο μπορεί να εφαρμοσθεί, και το αρχικό πρόβλημα είναι αυτό της εκτίμησης των παραγοντικών φορτίων l_{ij} και των ειδικών διακυμάνσεων ψ_i. Έχουν προταθεί ποικίλες μέθοδοι παραγοντικής ανάλυσης, κάποιες από τις οποίες είναι οι *Principal component analysis, Canonical factor analysis, Alpha factoring, Image factoring* και *Common factor analysis.*

2.4.4.1 Μέθοδος Principal Component

Έστω ότι ο πίνακας **Σ** έχει τα ζεύγη ιδιοτιμών-ιδιοδιανυσμάτων $(\lambda_i, \mathbf{e}_i)$ με $\lambda_1 \geq \lambda_2 \geq .. \geq \lambda_p \geq 0$. Ο πίνακας αυτός ορίζεται για το παραγοντικό μοντέλο το οποίο έχει τόσους παράγοντες όσες μεταβλητές ($k = p$) και ειδική διακύμανση $\psi_i = 0$ για κάθε *i*. Για τον πίνακα φορτίων η j^{η} στήλη αποτελείται από τα $\sqrt{\lambda_j}\,\mathbf{e}_j$. Εκτός του παράγοντα κλίμακας $\sqrt{\lambda_j}$, τα φορτία του $j^{ου}$ παράγοντα είναι οι συντελεστές της $j^{ης}$ κύριας συνιστώσας του πληθυσμού.

Παρ' όλο που ο πίνακας **Σ** είναι ακριβής, δεν είναι ιδιαίτερα χρήσιμος, αφού χρησιμοποιεί τόσους κοινούς παράγοντες όσες οι μεταβλητές και δεν επιτρέπει κάθε μεταβολή των ειδικών παραγόντων **ε** του παραγοντικού μοντέλου. Προτιμούνται μοντέλα τα οποία ερμηνεύουν την δομή της συνδιακύμανσης σε όρους περιορισμένου πλήθους κοινών παραγόντων. Μία προσέγγιση, όταν οι τελευταίες *p - m* ιδιοτιμές είναι μικρές, είναι η αγνόηση της συνεισφοράς των $\lambda_{m+1}\mathbf{e}_{m+1}e'_{m+1} + ... + \lambda_p \mathbf{e}_p e'_p$ στον **Σ**. Αγνοώντας αυτή τη συνεισφορά, λαμβάνεται η προσέγγιση

$$\mathbf{\Sigma} = \begin{bmatrix} \sqrt{\lambda_1}\mathbf{e}_1 & \sqrt{\lambda_2}\mathbf{e}_2 & \dots & \sqrt{\lambda_m}\mathbf{e}_m \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1}\mathbf{e}'_1 \\ \sqrt{\lambda_2}\mathbf{e}'_2 \\ \vdots \\ \sqrt{\lambda_m}\mathbf{e}'_m \end{bmatrix} = \underset{(p\times m)}{\mathbf{L}}\,\underset{(m\times p)}{\mathbf{L}'}$$

Η παραπάνω σχέση υποθέτει ότι οι ειδικοί παράγοντες **ε** είναι αμελητέας σημαντικότητας και μπορούν επίσης να αγνοηθούν στην παραγοντοποίηση του **Σ**. Εάν οι ειδικοί παράγοντες περιλαμβάνονται στο μοντέλο, οι διακυμάνσεις τους μπορούν να θεωρηθούν ότι αντιστοιχούν στα διαγώνια στοιχεία του **Σ – LL΄**, όπου **LL΄** ορίστηκε παραπάνω.

Καθιστώντας δυνατούς τους ειδικούς παράγοντες, η προσέγγιση γίνεται **Σ = LL΄+ Ψ**, όπου

$$\psi_i = \sigma_{ii} - \sum_{j=1}^{m} l_{ij}^2 \text{ για } i = 1, .., p.$$

Για να εφαρμοσθεί η παραπάνω προσέγγιση σε ένα σύνολο δεδομένων $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$, συνηθίζεται αρχικά να 'κεντράρονται' οι παρατηρήσεις αφαιρώντας τον δειγματικό μέσο $\bar{\mathbf{x}}$. Οι κεντραρισμένες παρατηρήσεις έχουν τον ίδιο δειγματικό πίνακα συνδιακύμανσης **S** με τις αρχικές παρατηρήσεις. Στην περίπτωση κατά την οποία οι μεταβλητές δεν είναι ανάλογης τάξης μεγέθους και κλίμακας, είναι συνήθως επιθυμητό να ληφθούν οι τυποποιημένες τιμές

$$\mathbf{z}_j = \begin{bmatrix} \frac{x_{j1} - \bar{x}}{\sqrt{s_{11}}_1} \\ \frac{x_{j2} - \bar{x}_2}{\sqrt{s_{22}}} \\ \vdots \\ \frac{x_{jp} - \bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix}, \quad j = 1, 2, .., n$$

των οποίων ο δειγματικός πίνακας συνδιακύμανσης είναι ο δειγματικός πίνακας συσχέτισης **R** των παρατηρήσεων $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$. Τυποποιώντας, αποφεύγεται το πρόβλημα ύπαρξης μεταβλητής με υψηλή διακύμανση επηρεάζοντας υπερβολικά τον καθορισμό των παραγοντικών φορτίων.

Η προσέγγιση αυτή, όταν εφαρμόζεται στον δειγματικό πίνακα συνδιακύμανσης **S** ή στον δειγματικό πίνακα συσχέτισης **R**, είναι γνωστή ως *λύση κύριων συνιστωσών* (principal component solution). Η ονομασία αυτή προκύπτει από το γεγονός ότι τα παραγοντικά φορτία είναι οι κλιμακωτοί συντελεστές κάποιων από τις πρώτες δειγματικές κύριες συνιστώσες.

Σύμφωνα με την μέθοδο των κύριων συνιστωσών, τα εκτιμώμενα φορτία για δεδομένους παράγοντες δεν μεταβάλλονται καθώς το πλήθος των παραγόντων αυξάνεται. Από τον ορισμό του $\tilde{\psi}_i$, τα διαγώνια στοιχεία του **S** είναι ισοδύναμα με τα διαγώνια στοιχεία του πίνακα $\tilde{\mathbf{L}}\tilde{\mathbf{L}}' + \tilde{\mathbf{\Psi}}$. Ωστόσο, τα μη διαγώνια στοιχεία του **S** συνήθως δεν αναπαράγονται από τον $\tilde{\mathbf{L}}\tilde{\mathbf{L}}' + \tilde{\mathbf{\Psi}}$. Σε αυτή την περίπτωση προβληματισμό αποτελεί η επιλογή του πλήθους των παραγόντων *m*.

Στην περίπτωση κατά την οποία το πλήθος των κοινών παραγόντων δεν καθορίζεται εκ των προτέρων, από την θεωρία ή από προηγούμενες έρευνες, η επιλογή του *m* μπορεί να βασιστεί στις εκτιμώμενες ιδιοτιμές. Έστω ο πίνακας υπολοίπων

$$\mathbf{S} - \left(\tilde{\mathbf{L}}\tilde{\mathbf{L}}' + \tilde{\mathbf{\Psi}}\right)$$

απορρέει από την προσέγγιση του **S** με την λύση κύριων συνιστωσών. Τα διαγώνια στοιχεία είναι μηδενικά, και εφόσον τα υπόλοιπα στοιχεία είναι επίσης μικρά, επιλέγεται υποκειμενικά το μοντέλο *m* παραγόντων ως καταλληλότερο. Πιο αναλυτικά,

$$\text{Sum of squared entries of } \mathbf{S}-\left(\tilde{\mathbf{L}}\tilde{\mathbf{L}}'+\tilde{\mathbf{\Psi}}\right)\le \hat{\lambda}_{m+1}^2+\ldots+\hat{\lambda}_p^2 .$$

Επομένως, μια μικρή τιμή του αθροίσματος των τετραγώνων των ιδιοτιμών που έχουν αγνοηθεί υποδηλώνει μικρή τιμή του αθροίσματος των τετραγώνων των σφαλμάτων της προσέγγισης.

Ιδανικά, οι συνεισφορές των πρώτων παραγόντων στις δειγματικές διακυμάνσεις των μεταβλητών πρέπει να είναι μεγάλες. Η συνεισφορά στην δειγματική διακύμανση s_{ii} των πρώτων κοινών παραγόντων είναι $\tilde{l}_{i1}^{\,2}$. Η συνεισφορά στην συνολική δειγματική διακύμανση, $s_{11}+s_{22}+\ldots+s_{pp}=tr(\mathbf{S})$, των πρώτων κοινών παραγόντων είναι

$$\tilde{l}_{11}^{\,2}+\tilde{l}_{21}^{\,2}+\ldots+\tilde{l}_{p1}^{\,2}=\left(\sqrt{\hat{\lambda}_1}\,\hat{\mathbf{e}}_1\right)'\left(\sqrt{\hat{\lambda}_1}\,\hat{\mathbf{e}}_1\right)=\hat{\lambda}_1$$

εφόσον το ιδιοδιάνυσμα $\hat{\mathbf{e}}_1$ έχει μοναδιαίο μήκος. Γενικά,

$$\begin{pmatrix}\text{Αναλογία συνολικής}\\ \text{δειγματικής διακύμανσης}\\ \text{του j}^{\text{ου}}\text{ παράγοντα}\end{pmatrix}=\begin{cases}\dfrac{\hat{\lambda}_j}{s_{11}+s_{22}+\ldots+s_{pp}}, \text{για παραγοντική ανάλυση του } \mathbf{S}\\ \dfrac{\hat{\lambda}_j}{p}, \text{για παραγοντική ανάλυση του } \mathbf{R}\end{cases}.$$

Το παραπάνω κριτήριο συχνά χρησιμοποιείται ως ευρετικός μηχανισμός για τον καθορισμό του κατάλληλου πλήθους κοινών παραγόντων. Το πλήθος των κοινών παραγόντων που παραμένουν στο μοντέλο αυξάνεται έως ότου μια 'κατάλληλη αναλογία' της συνολικής δειγματικής μεταβλητότητας να ερμηνευθεί.

Μία άλλη σύμβαση είναι να τεθεί το *m* ίσο με το πλήθος των ιδιοτιμών του **R** που ξεπερνούν την μονάδα, εάν παραγοντοποιείται ο δειγματικός πίνακας συσχετίσεων, ή ίσο με το πλήθος των θετικών ιδιοτιμών του S, εάν παραγοντοποιείται ο δειγματικός πίνακας συνδιακύμανσης. Επιπλέον μέθοδοι επιλογής του πλήθους των συνιστωσών που τελικά επιλέγονται, παρουσιάζονται στην παράγραφο 2.4.6.

2.4.5 Μέθοδοι περιστροφής των παραγόντων

Η ανάλυση του μοντέλου πραγματοποιείται θέτοντας υποθέσεις οι οποίες επιτρέπουν την μοναδική εκτίμηση των **L** και **Ψ**. Ο πίνακας των φορτίων περιστρέφεται (πολλαπλασιαζόμενος από έναν ορθογώνιο πίνακα), όπου η περιστροφή καθορίζεται από κάποιο κριτήριο που εξασφαλίζει ευκολία στην ερμηνεία. Επομένως, η *περιστροφή* (rotation) χρησιμεύει για να καταστήσει το αποτέλεσμα της παραγοντικής ανάλυσης πιο κατανοητό και είναι συνήθως απαραίτητη για την καλύτερη ερμηνεία των παραγόντων. Όταν τα αρχικά φορτία δεν είναι άμεσα ερμηνεύσιμα, είναι σύνηθες να αντιμετωπίζεται με περιστροφή έως ότου μια 'απλούστερη' δομή να επιτευχθεί.

Η *Varimax* περιστροφή (ή *Normal Varimax*) είναι μια ορθογώνια περιστροφή των αξόνων των παραγόντων για να μεγιστοποιηθεί η διακύμανση των τετραγώνων των φορτίων ενός παράγοντα (στήλη) για όλες τις μεταβλητές (γραμμές) σε έναν πίνακα παραγόντων, η οποία έχει ως αποτέλεσμα τη διαφοροποίηση των αρχικών μεταβλητών από τους εξαγόμενους παράγοντες. Είναι σύνηθες (M. Lewis - Beck, 1994) να χρησιμοποιούνται κανονικοποιημένα φορτία κατά την περιστροφή προκειμένου να ελαχιστοποιηθεί η ανεπιθύμητη υπεροχή των υψηλών αρχικών φορτίων στην τελική λύση. Έστω $\tilde{l}_{ij}^{*} = \hat{l}_{ij}^{*} / \hat{h}_i$ οι περιστρεφόμενοι συντελεστές κλιμακωμένοι από την τετραγωνική ρίζα των *communalities*. Η μέθοδος *Varimax* επιλέγει τον ορθογώνιο μετασχηματισμό **Q** ο οποίος μεγιστοποιεί την ποσότητα,

$$V = \frac{1}{p}\sum_{j=1}^{k}\left[\sum_{i=1}^{p}\tilde{l}_{ij}^{*4} - \left(\sum_{i=1}^{p}\tilde{l}_{ij}^{*2}\right)^2 \Big/ p\right]$$

Κλιμακώνοντας τους περιστρεμμένους συντελεστές $\hat{l}_{ij}^{*}$ έχουμε ως αποτέλεσμα να δίνουμε στις μεταβλητές με μικρά *communalities* σχετικά μεγαλύτερο βάρος προκειμένου να εξασφαλισθεί απλή δομή. Μετά τον καθορισμό του μετασχηματισμού του **Q** τα φορτία $\tilde{l}_{ij}^{*}$ πολλαπλασιάζονται με το $\hat{h}_i$ ώστε να προκύψουν τα αρχικά *communalities*. Παρ' όλο που η παραπάνω σχέση φαίνεται απαγορευτική, εξασφαλίζει απλή ερμηνεία.

Κάθε παράγοντας τείνει να έχει είτε μικρά είτε μεγάλα φορτία της συγκεκριμένης μεταβλητής. Μια λύση *Varimax* παράγει αποτελέσματα που ελαχιστοποιούν το πλήθος των παραγόντων με υψηλά φορτία, οπότε καθιστούν κατά το δυνατόν ευκολότερη την αναγνώριση κάθε μεταβλητής με ένα μόνο παράγοντα. Αυτή είναι η πιο συνηθισμένη λύση

περιστροφής. Χρησιμοποιώντας την μέθοδο αυτή, ή οποιαδήποτε άλλη μέθοδο που χρησιμοποιεί ορθογώνιο πίνακα, οι παράγοντες παραμένουν ασυσχέτιστοι μεταξύ τους.

Η *Quartimax* περιστροφή είναι μία εναλλακτική ορθογώνια λύση που ελαχιστοποιεί τον αριθμό των παραγόντων που απαιτούνται για να ερμηνευθεί κάθε μεταβλητή. Αυτό το είδος περιστροφής δημιουργεί συχνά ένα γενικό παράγοντα στον οποίο οι περισσότερες μεταβλητές φορτώνονται σε υψηλό ή μεσαίο βαθμό. Μια τέτοια δομή παράγοντα συνήθως δεν είναι χρήσιμη για το σκοπό της έρευνας, ωστόσο απλοποιεί την ερμηνεία των παρατηρούμενων μεταβλητών. Πιο συγκεκριμένα, η ερμηνεία μιας μεταβλητής γίνεται απλούστερη καθώς λιγότεροι κοινοί παράγοντες εμπλέκονται, ενώ η ερμηνεία ενός παράγοντα γίνεται απλούστερη όταν ένας σχετικά μικρός αριθμός μεταβλητών έχει υψηλά φορτία στον παράγοντα, ενώ οι υπόλοιπες μεταβλητές έχουν μηδενικά φορτία στον παράγοντα. Η περιστροφή σύμφωνα με την μέθοδο *Ouartimax* ορίζεται ως,

$$Q = \sum_{i=1}^{p} \frac{\sum_{j=1}^{k} \tilde{l}_{ij}^{*4} - \left(\sum_{j=1}^{k} \tilde{l}_{ij}^{*2} \right)^2}{k^2}$$

Η *Equimax* περιστροφή είναι ένας συμβιβασμός μεταξύ των κριτηρίων *Varimax*, το οποίο απλοποιεί τους παράγοντες, και *Quartimax*, το οποίο απλοποιεί τις μεταβλητές. Το πλήθος των μεταβλητών που σημειώνουν υψηλά φορτία σε έναν παράγοντα και το πλήθος των παραγόντων που απαιτείται για την ερμηνεία μιας μεταβλητής ελαχιστοποιούνται.

Στη *πλάγια περιστροφή* (oblique rotation), θεωρούνται μία πρότυπη μήτρα και μια δομική μήτρα. Η δομική μήτρα είναι απλά ο πίνακας των φορτίων όπως στην ορθογώνια περιστροφή, που αντιπροσωπεύει τη διακύμανση σε μια μετρίσιμη μεταβλητή που ερμηνεύεται από έναν παράγοντα υπό τη βάση μοναδικών και κοινών συνεισφορών. Η πρότυπη μήτρα, αντιθέτως, περιέχει συντελεστές που μόλις αντιπροσωπεύουν μοναδικές συνεισφορές. Όσο περισσότεροι παράγοντες, τόσο λιγότερο λαμβάνονται οι πρότυποι συντελεστές ως κανόνας καθώς θα υπάρχουν περισσότερες κοινές συνεισφορές με την διακύμανση που ερμηνεύεται. Για τη πλάγια περιστροφή, ο ερευνητής εξετάζει τόσο τους δομικούς όσο και τους πρότυπους συντελεστές κατά τη χορήγησή μιας ετικέτας σε ένα συντελεστή.

Η *Direct oblimin* περιστροφή είναι η τυπική μέθοδος, όταν επιδιώκεται μια μη ορθογώνια (πλάγια) λύση - δηλαδή, οι παράγοντες μπορούν να συσχετιστούν. Το γεγονός αυτό οδηγεί σε υψηλότερες ιδιοτιμές, αλλά μειώνεται η επεξηγηματικότητα των παραγόντων. Η πλάγια περιστροφή σε μια απλή δομή αντιστοιχεί σε μια μη άκαμπτη περιστροφή του συστήματος

συντεταγμένων όπως οι περιστρεφόμενοι άξονες διέρχονται τις ομάδες. Η μέθοδος αυτή αναζητά να εκφράσει κάθε μεταβλητή σε όρους ενός ελάχιστου πλήθους παραγόντων – κατά προτίμησιν, ενός μόνο παράγοντα. Το πλεονέκτημα της πλάγιας περιστροφής ως μη ορθογώνια μέθοδος, έγκειται στο γεγονός ότι μετά την εφαρμογή της, εάν οι παράγοντες που λαμβάνονται είναι ορθογώνιοι, τότε η ορθογωνιότητα που προκύπτει δεν είναι 'τεχνούργημα' της μεθόδου περιστροφής.

Η *Promax* περιστροφή είναι μια εναλλακτική μη ορθογώνια (πλάγια) μέθοδος περιστροφής που επιτρέπει στους παράγοντες να συσχετίζονται και είναι υπολογιστικά πιο γρήγορα από τη μέθοδο *Direct oblimin* και ως εκ τούτου χρησιμοποιείται συχνά για πολύ μεγάλα σύνολα δεδομένων.

2.4.6 Κριτήρια καθορισμού του πλήθους των παραγόντων

Κριτήριο Kaiser

Ο κανόνας *Kaiser* είναι να αφαιρεθούν όλοι οι παράγοντες με ιδιοτιμές μικρότερες του 1. Το κριτήριο *Kaiser* δεν συνιστάται, όταν χρησιμοποιείται ως το μοναδικό κριτήριο αποκοπής για την εκτίμηση του αριθμού των παραγόντων. Ωστόσο το κριτήριο αυτό εφαρμόζεται ιδιαίτερα σε ανάλυση γονιδιακών εκφράσεων (Wang *et al.*, 2006; Wang *et al.*, 2010).

Scree Plot

Το *Cattell scree test* απεικονίζει τις συνιστώσες στον άξονα X και τις αντίστοιχες ιδιοτιμές στον άξονας Y. Καθώς κινείται προς τα δεξιά, προς τις επόμενες συνιστώσες, οι ιδιοτιμές σημειώνουν πτώση. Όταν η πτώση παύει και η καμπύλη κάνει ένα 'αγκώνα' σημειώνοντας λιγότερο απότομη πτώση, το *Cattell scree plot* υποδεικνύει να απαλειφθούν όλες οι επιπλέον συνιστώσες που έπονται της μεταβολής αυτής, συμπεριλαμβανομένης και της συνιστώσας στην οποία παρατηρείται η αλλαγή. Ο κανόνας αυτός επιδέχεται συχνά κριτική διότι επιτρέπει ελεγχόμενες από τον ερευνητή παραποιήσεις, με την έννοια ότι η επιλογή του κρίσιμου σημείου μπορεί να είναι υποκειμενική.

Παράλληλη Ανάλυση του Horn (Horn's Parallel Analysis - PA)

Πρόκειται για μια μέθοδος προσομοίωσης βασιζόμενη στο Monte-Carlo η οποία συγκρίνει τις παρατηρούμενες ιδιοτιμές με εκείνες που προκύπτουν από ασυσχέτιστες κανονικές

μεταβλητές. Ένας παράγοντα διατηρείται εάν η σχετική ιδιοτιμή είναι μεγαλύτερη από την 95η της κατανομής των ιδιοτιμών που προέρχονται από τα τυχαία δεδομένα.

Κριτήριο ερμηνευτικής διακύμανσης (Variance explained criteria)

Μερικοί ερευνητές χρησιμοποιούν απλά τον κανόνα της διατήρησης αρκετών στοιχείων ώστε να αντιπροσωπεύεται το 90% (μερικές φορές 80%) της μεταβλητότητας. Στη περίπτωση που ο στόχος του ερευνητή εστιάζει στην οικονομία (ερμηνεύοντας την διακύμανση με όσο το δυνατόν λιγότερους παράγοντες), το κριτήριο θα μπορούσε να είναι τόσο χαμηλό όσο το 50%. Ωστόσο, πριν από τη θεώρηση ενός παράγοντα κάτω του cut-off, ο ερευνητής πρέπει να ελέγξει τον συσχετισμό του με την εξαρτημένη μεταβλητή. Ένας πολύ μικρός παράγοντας μπορεί να έχει σημαντική συσχέτιση με την εξαρτημένη μεταβλητή, οπότε δεν πρέπει να παραληφθεί (Ιστοσελίδα 2).

Κριτήριο Κατανόησης (Comprehensibility)

Αν και δεν είναι αυστηρά μαθηματικό κριτήριο, υπάρχουν πολλά να ειπωθούν για τον περιορισμό του αριθμού των παραγόντων σε αυτούς των οποίων η διάσταση που έχει κάποιο νόημα είναι εύκολα κατανοητή. Συχνά πρόκειται για τις πρώτες δύο ή τρεις συνιστώσες. Χρησιμοποιώντας μία ή περισσότερες από τις παραπάνω μεθόδους, ο ερευνητής καθορίζει ένα κατάλληλο εύρος λύσεων για την έρευνα. Για παράδειγμα, το κριτήριο Kaiser μπορεί να προτείνει τρεις παράγοντες και το scree test μπορεί να προτείνει 5, επομένως ο ερευνητής καλείται να αποφασίσει μεταξύ των 3, 4 και 5 παραγόντων και να επιλέξει τη λύση που παράγει την πιο λογική δομή των παραγόντων.

2.4.7 Στατιστική εγκυρότητα των δειγμάτων ανάλυσης κατά παράγοντες

Αναπόφευκτα υπάρχει ανησυχία όταν γίνεται επιλογή μεταβλητών προς ανάλυση, για την περίπτωση κατά την οποία οι πίνακες συσχετίσεων δεν είναι κατάλληλοι για ανάλυση κατά παράγοντες. Για μεγάλα δείγματα ο έλεγχος του Bartlett προσεγγίζει μία X^2 κατανομή (Ιστοσελίδα 8). Η μηδενική υπόθεση του ελέγχου δηλώνει ότι ο πίνακας συσχετίσεων είναι ο μοναδιαίος πίνακας, δηλαδή έχει στην διαγώνιο μονάδες και στα υπόλοιπα κελιά μηδενικά, πιο συγκεκριμένα έχει την μορφή,

H_0: ο πίνακας συσχετίσεων είναι ο μοναδιαίος πίνακας.

Κατά συνέπεια, συνήθως θεωρείται ότι η δειγματική συσχέτιση προέρχεται από έναν πολυμεταβλητό κανονικό πληθυσμό με τις υπό μελέτη μεταβλητές να είναι ανεξάρτητες. Εάν δεν ισχύει η παραπάνω παραδοχή, τότε τα δεδομένα είναι κατάλληλα για ανάλυση κατά παράγοντες. Ο έλεγχος του Bartlett, ωστόσο, δεν είναι το ίδιο αξιόπιστος για μικρά δείγματα. Πολύ μικρές τιμές σημαντικότητας (κάτω του 0,05) δηλώνουν υψηλή πιθανότητα να υπάρχουν σημαντικές σχέσεις μεταξύ των μεταβλητών, ενώ υψηλότερες τιμές (μεγαλύτερες του 0,1) δηλώνουν ότι τα δεδομένα είναι ακατάλληλα για ανάλυση κατά παράγοντες. Συνεπώς, επιθυμητή είναι η απόρριψη του ελέγχου.

Το μέτρο δειγματικής επάρκειας *Kaiser-Meyer-Olkin* (KMO) παρέχει ένα δείκτη (μεταξύ 0 και 1) του ποσοστού της διακύμανσης μεταξύ των μεταβλητών που θα μπορούσε να αντιστοιχεί σε κοινές διακυμάνσεις (Ιστοσελίδα 8). Εάν η μερική διακύμανση είναι 0, οι μεταβλητές μοιράζονται ένα κοινό παράγοντα οπότε ισχύει KMO=1, και αντίστροφα (Ιστοσελίδα 7). Το περιγραφικό αυτό μέτρο υπολογίζεται ως εξής,

$$KMO = \frac{\sum_{\substack{i,j=1 \\ i \neq j}}^{p} r_{ij}^2}{\sum_{\substack{i,j=1 \\ i \neq j}}^{p} r_{ij}^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^{p} \alpha_{ij}^2},$$

όπου r_{ij} ο δειγματικός συντελεστής συσχέτισης και $\alpha_{ij}^2 :=$ Corr(X_i, Xj|$\mathbf{X}\backslash\{X_i,X_j\}$) ο μερικός συντελεστής συσχέτισης των X_i, X_j. Ο δείκτης δηλώνει αν τα δεδομένα υποστηρίζουν ικανοποιητικά την εφαρμογή της ανάλυσης κατά παράγοντες και προτείνει αν τα δεδομένα είναι δυνατό να ομαδοποιηθούν σε ένα μικρότερο σύνολο παραγόντων. Οι τιμές που πλησιάζουν στη μονάδα είναι οι πιο αποδεκτές, σύμφωνα με αυτό το κριτήριο. Ο Kaiser (1974) πρότεινε την εξής κλίμακα. Όταν η τιμή του δείκτη είναι άνω του 0,90 είναι ‘θαυμάσιο’ (marvelous), μεταξύ 0,80 και 0,90 ‘αξιέπαινο’ (meritorious), μεταξύ 0,60 και 0,80 ‘μέτριο’ (middling, mediocre), μεταξύ 0,50 και 0,60 ‘άθλιο’ (miserable) και κάτω του 0,50 ‘μη αποδεκτό’ (unacceptable).

Λαμβάνοντας υπόψη τους παραπάνω ελέγχους από κοινού, παρέχουν ελάχιστα προαπαιτούμενα τα οποία πρέπει να ισχύουν προτού προβούμε στην εφαρμογή της ανάλυσης κατά παράγοντες (Ιστοσελίδα 10).

Κ Ε Φ Α Λ Α Ι Ο 3

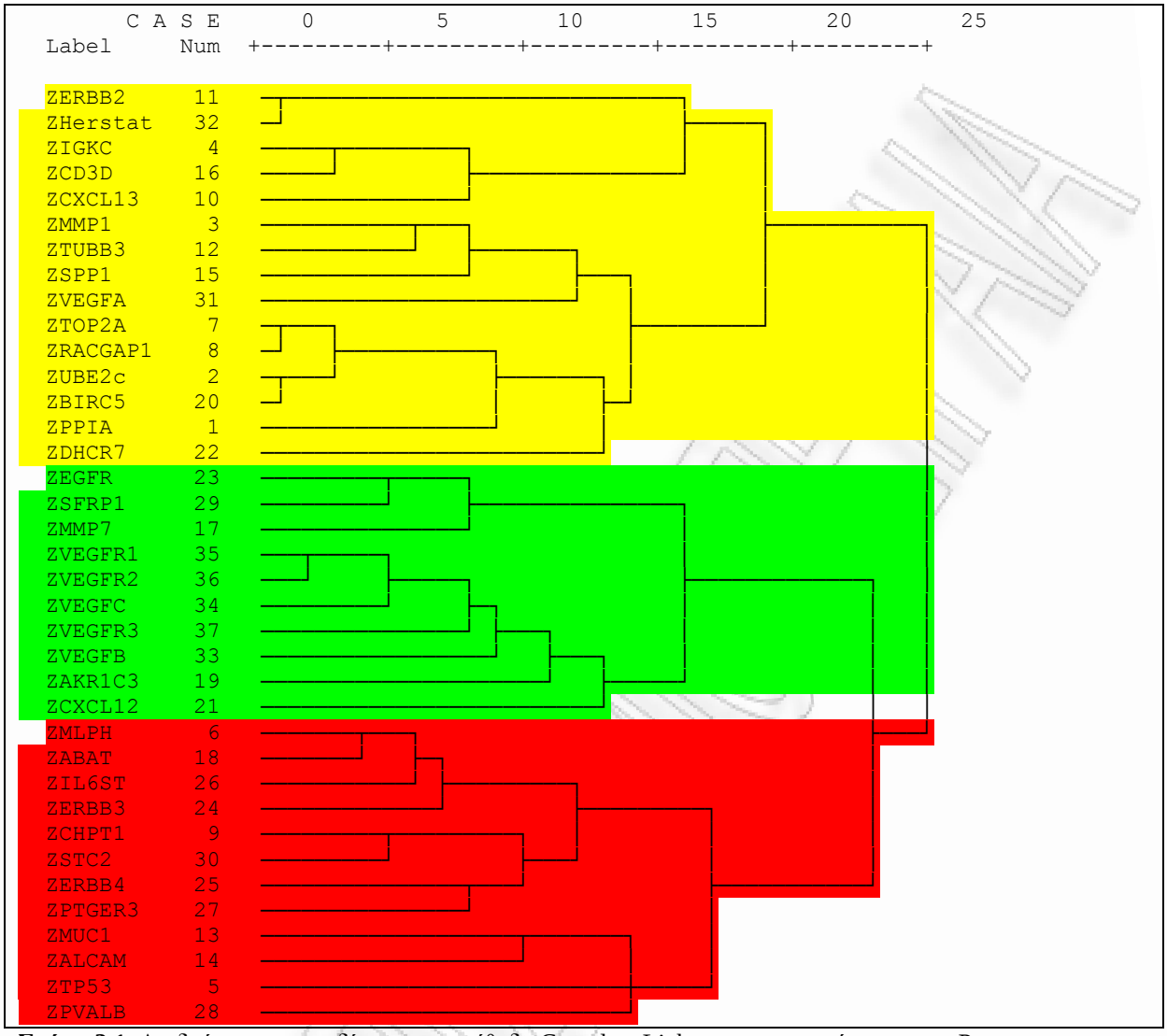
Ομαδοποίηση Γονιδιακών Εκφράσεων για Καρκίνο του Μαστού

Στο παρόν κεφάλαιο, επιχειρείται η ομαδοποίηση των γονιδιακών εκφράσεων σε ομάδες ομοιογενείς ως προς κοινά χαρακτηριστικά, εφαρμόζοντας την ιεραρχική μέθοδο του *μακρινότερου γείτονα*, θεωρώντας την απόσταση του Pearson ως μέτρο ανομοιότητας, ώστε να προσδιορισθεί το κατάλληλο πλήθος ομάδων για τον διαχωρισμό των γονιδίων. Έπειτα εφαρμόζεται η μη ιεραρχική μέθοδος *k-means* ώστε να διαμορφωθούν οι ομάδες των γονιδίων και η μέθοδος *Partitioning around Medoid*. Τέλος, πραγματοποιείται *ανάλυση κατά παράγοντες* έτσι ώστε να εξάγουμε ένα αρχικό σύνολο παραγόντων σύμφωνα με την μέθοδο *principal component* και περιστρέφοντας τους παράγοντες στην τελική λύση μέσω της ορθογώνιας περιστροφής *Quartimax*, ικανοποιώντας, παράλληλα, το *κριτήριο του Kaiser*. Τα δεδομένα είναι τυποποιημένα ως προς την μέση τιμή.

3.1 Ομαδοποίηση σύμφωνα με τη μέθοδο *Complete Linkage*

Αρχικά, κατασκευάζεται ένα δενδρόγραμμα σύμφωνα με μια ιεραρχική μέθοδο ομαδοποίησης ώστε να αποφανεί το κατάλληλο πλήθος ομάδων με βάση το οποίο θα πραγματοποιηθεί ο διαχωρισμός των γονιδίων σε συστάδες. Η ιεραρχική μέθοδος που εφαρμόζεται για την *cluster analysis* είναι η *complete linkage* (μέθοδος του μακρινότερου γείτονα). Αξίζει να επισημανθεί ότι η Ευκλείδεια απόσταση που εφαρμόζεται σε τυποποιημένα δεδομένα καταλήγει στην ίδια ομαδοποίηση με αυτή της απόστασης του Pearson. Αφού τα δεδομένα έχουν ήδη τυποποιηθεί όποια από τις δύο αποστάσεις εφαρμοσθεί έχει τα ίδια αποτελέσματα.

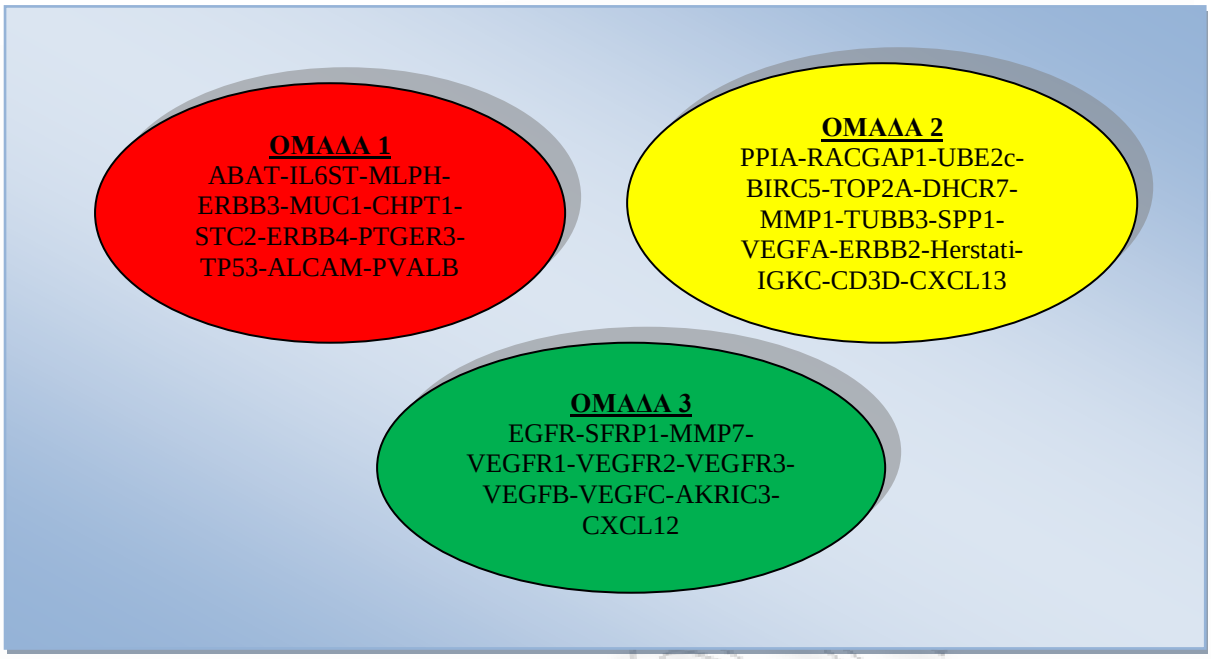
Το δενδρόγραμμα που ακολουθεί κατασκευάστηκε κάνοντας χρήση της απόστασης Pearson (βλέπε παραγράφους 2.1.1.1. και 2.1.2.1).



Σχήμα 3.1: Δενδρόγραμμα γονιδίων με την μέθοδο Complete Linkage και την απόσταση του Pearson.

Από την μορφή του δενδρογράμματος προκύπτει ότι το πλήθος των ομάδων είναι τρεις. Στο συμπέρασμα αυτό καταλήγουμε παρατηρώντας τις αποστάσεις των διαδοχικών κάθετων γραμμών συνένωσης των γονιδιακών ομάδων. Η μεγαλύτερη απόσταση μεταξύ των κάθετων αυτών γραμμών βρίσκεται στο δεύτερο επίπεδο (θεωρώντας ως αρχικό επίπεδο την μία ομάδα που περιέχει το σύνολο των γονιδίων) και θεωρώντας μια νοητή κάθετη γραμμή, προκύπτει ότι 'κόβει' τρεις γραμμές συνένωσης. Οπότε καταλήγουμε στο συμπέρασμα ότι το πλήθος των ομάδων είναι τρεις.

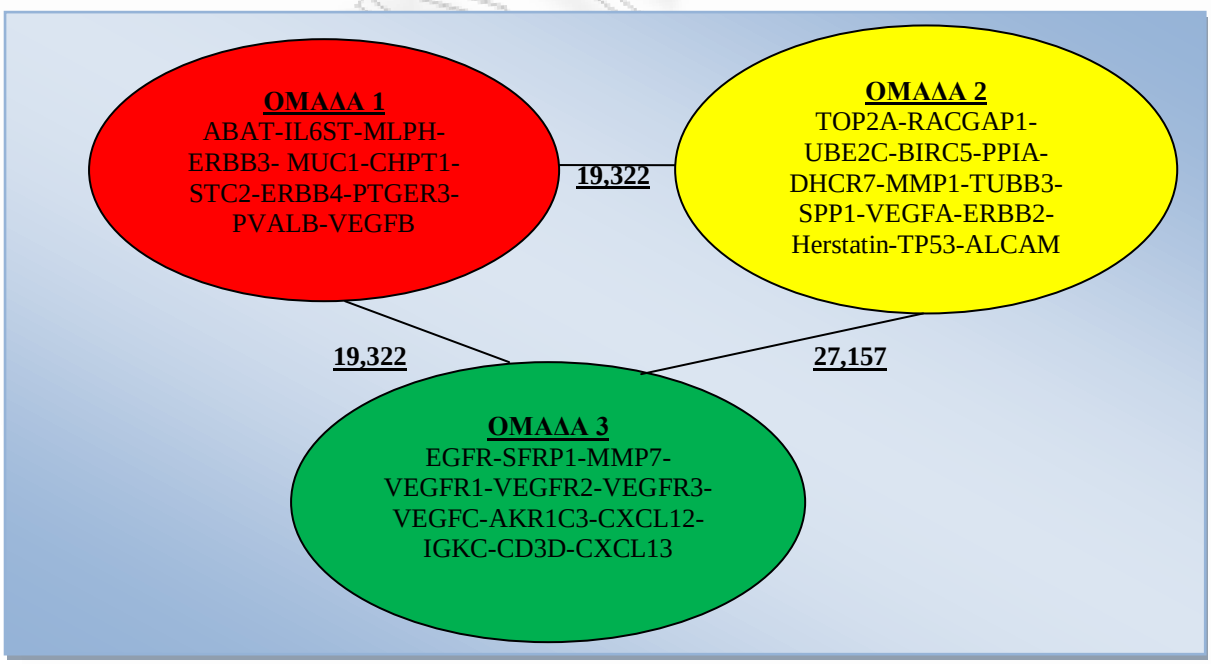
Στη συνέχεια παρουσιάζονται σχηματικά οι ομάδες των δεδομένων:



Σχήμα 3.2: Ομαδοποίηση γονιδίων σύμφωνα με την μέθοδο Complete Linkage.

3.2 Μέθοδος *k-means*

Σύμφωνα με τα αποτελέσματα της μη ιεραρχικής μεθόδου *k-means*, προκύπτει η ακόλουθη ομαδοποίηση (βλέπε παράγραφο 2.1.2.2):



Σχήμα 3.3: Ομαδοποίηση γονιδίων σύμφωνα με την μέθοδο *k-means* και απεικόνιση των αποστάσεων των ομάδων.

Συγκρίνοντας τις δύο παραπάνω ομαδοποιήσεις, παρατηρούμε ότι υπάρχουν συνολικά 6 γονίδια (16,216%) τα οποία εκχωρούνται σε διαφορετικές ομάδες. Αναλυτικά παρουσιάζονται στον ακόλουθο πίνακα:

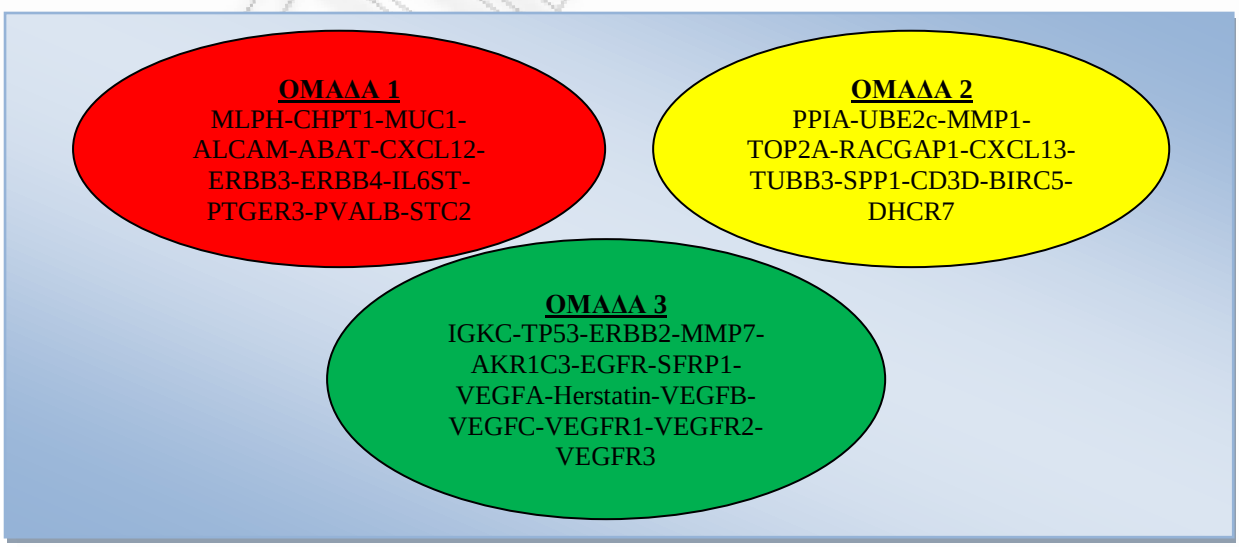
Γονίδιο	Complete Linkage	k-means
IGKC	2	3
CD3D	2	3
CXCL13	2	3
TP53	1	2
ALCAM	1	2
VEGFB	3	1

Πίνακας 3.1: Σύγκριση μεθόδων *Complete Linkage* και *k-means*.

Παρατηρείται ότι οι ομάδες 2 και 3 είναι πιο 'μακριά' μεταξύ τους συγκρίνοντας τα δυνατά ζεύγη αποστάσεων των τριών ομάδων. Σε αντίθεση, οι ομάδες 1 και 2 έχουν σχεδόν την ίδια απόσταση όσο και οι ομάδες 1 και 3.

3.3 Μέθοδος *Partitioning Around Medoid*

Στην συνέχεια εφαρμόζεται η μέθοδος *PAM* (βλέπε παράγραφο 2.1.2.3.) για το πλήθος των τριών ομάδων, όπως προέκυψε παραπάνω, κάνοντας χρήση της Ευκλείδειας απόστασης. Τα medoids των τριών ομάδων είναι τα γονίδια UBE2c, VEGFR1 και MLPH και η ομαδοποίηση έχει ως εξής:



Σχήμα 3.4: Ομαδοποίηση γονιδίων σύμφωνα με την μέθοδο PAM.

Τα αποτελέσματα των μεθόδων *k-means* και *PAM* φαίνεται να σημειώνουν κάποια διαφοροποίηση δεδομένου ότι τα 9 γονίδια (24,324%) εκχωρούνται σε διαφορετικές ομάδες. Συγκεκριμένα, έχουμε:

Γονίδιο	PAM	k-means
VEGFB	3	1
VEGFA	3	2
ERBB2	3	2
Herstati	3	2
TP53	3	2
ALCAM	1	2
CXCL12	1	3
CD3D	2	3
CXCL13	2	3

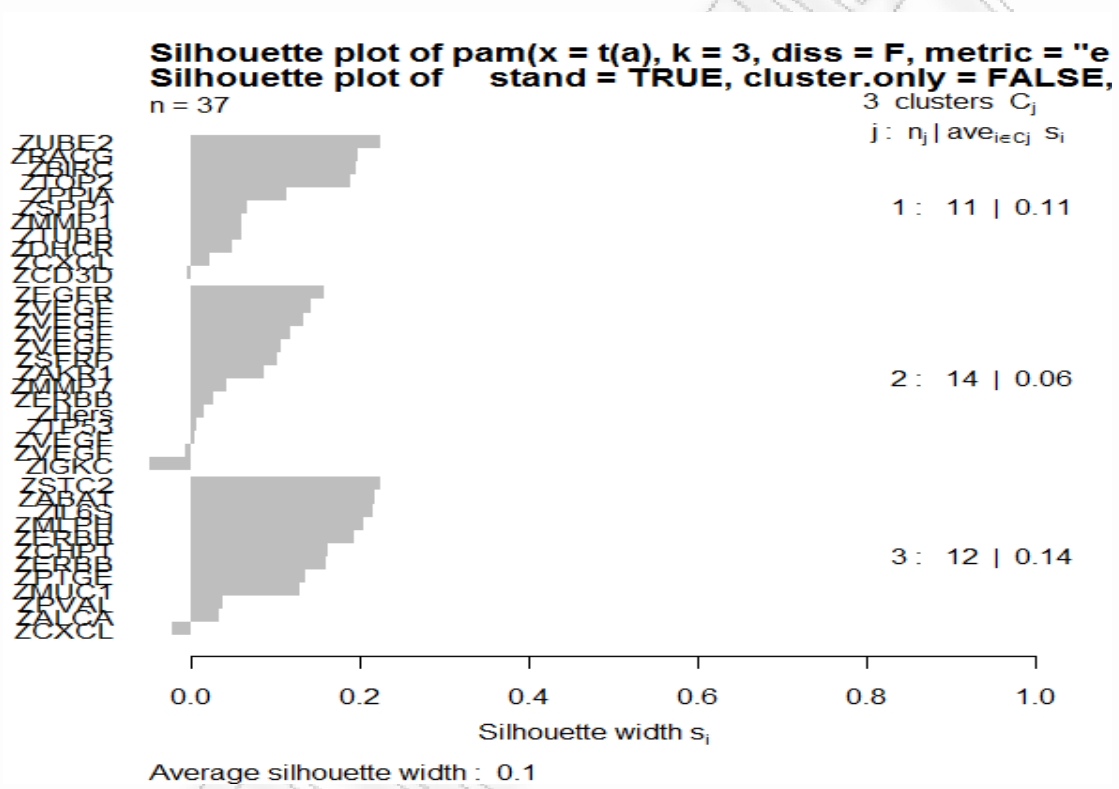
Πίνακας 3.2: Σύγκριση μεθόδων *PAM* και *k-means*.

Παρατηρούμε ότι τα αποτελέσματα δεν είναι πολύ κοντινά με αυτά της *k-means*. Από την παραπάνω σύγκριση, τα αποτελέσματα που προκύπτουν δεν φαίνεται να είναι ικανοποιητικά.

gene	cluster	neighbor	sil_width	gene	cluster	neighbor	sil_width
UBE2c	1	2	0.202513096	ERBB2	2	1	0.027403784
RACGAP1	1	2	0.174217670	Herstatin	2	1	0.017259136
BIRC5	1	2	0.168317347	TP53	2	3	0.011068408
TOP2A	1	2	0.167898225	VEGFA	2	1	0.001872804
PPIA	1	2	0.088371808	VEGFB	2	3	-0.02896175
SPP1	1	2	0.05357800	PVALB	2	3	-0.04306069
MMP1	1	2	0.052549072	IGKC	2	1	-0.04404643
TUBB3	1	2	0.048790423	ALCAM	2	3	-0.05351143
DHCR7	1	2	0.029402019	IL6ST.24	3	2	0.244023941
CXCL13	1	2	0.011200914	ABAT	3	2	0.232094718
CD3D	1	2	-0.00775544	STC2	3	2	0.228527219
EGFR	2	3	0.122173738	MLPH	3	2	0.19722824
VEGFR1	2	3	0.121433963	ERBB4	3	2	0.188988419
VEGFC	2	3	0.098933416	CHPT1	3	2	0.175915773
VEGFR2	2	3	0.095099489	ERBB3	3	2	0.163288825
VEGFR3	2	3	0.074574337	PTGER3	3	2	0.16164688
SFRP1	2	3	0.067353818	MUC1	3	2	0.129849227
AKR1C3	2	3	0.059722144	CXCL12	3	2	0.010607647
MMP7	2	1	0.033335779				

Πίνακας 3.3: Πληροφορία ομαδοποίησης *Silhouette*.

Σύμφωνα με τον πίνακα πληροφορίας του Silhouette plot, το 43,243% των γονιδίων έχουν *s(i)* κοντά στο μηδέν, γεγονός που σημαίνει ότι τα *α(i)* και *b(i)* είναι σχεδόν ίσα. Επομένως, δεν είναι ξεκάθαρο αν το γονίδιο *i* θα έπρεπε να εκχωρηθεί στην ομάδα αυτή ή την αντίστοιχη πλησιέστερη. Θα μπορούσε να θεωρηθεί ως μια 'ενδιάμεση' περίπτωση. Το γεγονός αυτό είναι μια αρχική ένδειξη ότι η ομαδοποίηση με την μέθοδο αυτή δεν είναι ικανοποιητική.

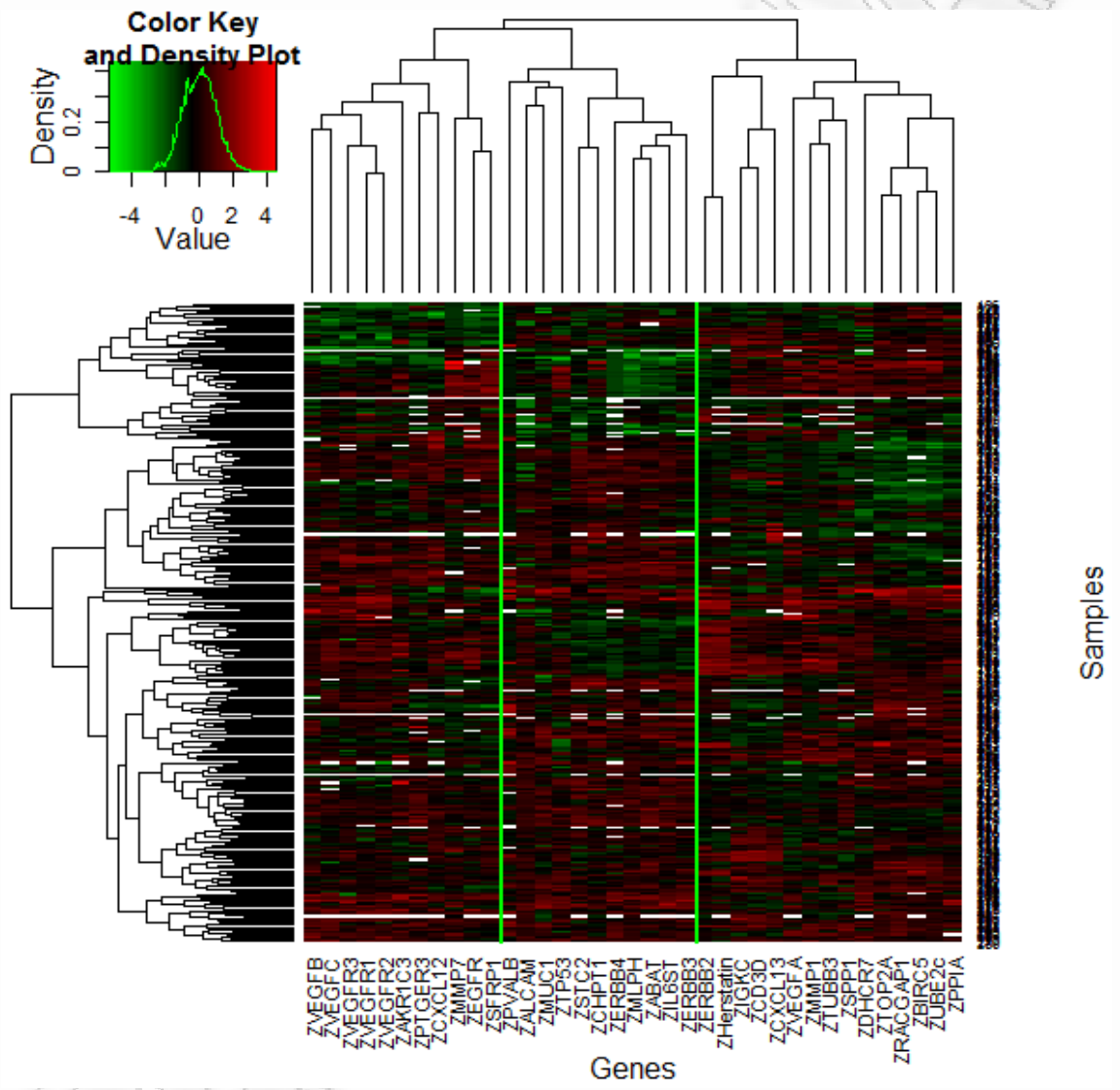


Σχήμα 3.5: Silhouette Plot.

Σύμφωνα με το παραπάνω γράφημα (βλέπε παράγραφο 2.1.2.3.1) τα σχήματα των ομάδων απέχουν αισθητά από την μονάδα, γεγονός που σημαίνει ότι δεν έχει βρεθεί αξιόλογη δομή, οπότε πρέπει να βασιστούμε σε άλλες μεθόδους. Το πλάτος που αντιστοιχεί σε k=3 δεν φαίνεται να είναι καθόλου ικανοποιητικό αφού το μέσο πλάτος προκύπτει ως SC = 0,1. Οπότε επιβεβαιώνεται η ένδειξη που αναφέρθηκε παραπάνω.

3.4 Χάρτης Θερμότητας

Στη συνέχεια παρουσιάζεται ο χάρτης θερμότητας του συνόλου των δεδομένων (βλέπε παράγραφο 2.2), στον οποίο περιλαμβάνεται δενδρόγραμμα για τις γονιδιακές εκφράσεις και τους ίδιους τους ασθενείς.



Σχήμα 3.6: Heat map.

Όπως προκύπτει από το δενδρόγραμμα που αντιστοιχεί στους ασθενείς, μπορούμε να τους διακρίνουμε σε τρεις κατηγορίες. Αξίζει να επισημάνουμε ότι οι αποχρώσεις του πράσινου αντιστοιχούν σε αρνητικά επίπεδα έκφρασης των γονιδίων, ενώ οι κόκκινες σε θετικά. Η πρώτη ομάδα των ασθενών, χαρακτηρίζεται από αρνητική έκφραση των

γονιδιακών εκφράσεων που απαρτίζουν τις ομάδες 2 και 3, ενώ σημειώνουν θετική έκφραση για τα γονίδια της πρώτης ομάδας. Στην δεύτερη ομάδα ασθενών, παρατηρείται αντίστροφη έκφραση, δηλαδή για την πρώτη ομάδας γονιδιακών εκφράσεων προκύπτει αρνητική έκφραση, ενώ για τις υπόλοιπες δύο θετική έκφραση. Τέλος, στην τρίτη ομάδα ασθενών, η οποία συγκεντρώνει την πλειοψηφία αυτών, για κάθε ομάδα γονιδιακών εκφράσεων παρατηρείται θετική έκφραση.

3.5 Δείκτες ισχύος της ομαδοποίησης

Σύμφωνα με τα μέτρα που αναφέραμε (βλέπε παράγραφο 2.3) για τον προσδιορισμό της ισχύος της ομαδοποίησης στην οποία καταλήξαμε, προκύπτουν τα αποτελέσματα που παρουσιάζονται συγκεντρωτικά στον Πίνακα 3.4.

Measure	Index	
Internal		
	Connectivity	17,4726
	Silhouette Width	0,125
	Dunn Index	0,7162
Stability		
	APN	0,0048
	AD	19,0468
	ADM	0,1098
	FOM	0,7032
Biological		
	BHI	0,2284
	BSI	0,3372

Πίνακας 3.4: Δείκτες Ισχύος Ομαδοποίησης

Όσον αφορά τα *Internal Measures*, μία ομαδοποίηση είναι ικανοποιητική όταν ο δείκτης *Connectivity* είναι ελάχιστος, δηλαδή λαμβάνει τιμή κοντά στο μηδέν, ενώ οι δείκτες *Silhouette Width* και *Dunn* είναι μέγιστοι, με τον πρώτο να λαμβάνει τιμή κοντά στη μονάδα. Όπως προκύπτει από τον παραπάνω πίνακα, τα εσωτερικά μέτρα δεν είναι ικανοποιητικά. Πιο

συγκεκριμένα, φαίνεται να υπάρχει έντονη αντικατάσταση (*Conn* = 17,47) των παρατηρήσεων σε κάθε ομάδα. Κατά μέσο όρο, η εμπιστοσύνη της διαμορφούμενης ομαδοποίησης δεν είναι καθόλου ικανοποιητική, αφού *S* = 0,125 με τις αποδεκτές τιμές να βρίσκονται κοντά στην μονάδα. Τέλος, η ομοιογένεια της ομαδοποίησης, σύμφωνα με την πυκνότητα που προκύπτει *D* = 0,72 δεν είναι, επίσης, ικανοποιητική.

Λαμβάνοντας υπόψη τα *Stability Measures,* επιθυμητές τιμές είναι αυτές κοντά στο μηδέν. Σύμφωνα με τον δείκτη *APN*, η μέση αναλογία των παρατηρήσεων που δεν τοποθετούνται στην ίδια ομάδα (*APN* = 0,0048) είναι αρκετά ικανοποιητική, όπως, επίσης, η μέση απόσταση μεταξύ των κέντρων των ομάδων (*ADM* = 0,11) και η μέση διακύμανση εντός των ομάδων (*FOM* = 0,70), αφού οι τρεις παραπάνω δείκτες λαμβάνουν τιμές πολύ κοντά στο μηδέν. Παρατηρείται ότι η μέση απόσταση μεταξύ των ομάδων (*AD* = 19,05) δεν λαμβάνει τιμή πολύ κοντά στο μηδέν, όπως οι υπόλοιποι δείκτες της ομάδας αυτής, ωστόσο φαίνεται να είναι ικανοποιητικός. Συγκεντρωτικά, η ομαδοποίηση δεν εμφανίζει κάποιο πρόβλημα ως προς τα μέτρα σταθερότητας.

Τέλος, τα *Biological Measures* πρέπει να λαμβάνουν τιμή κοντά στην μονάδα, ώστε να εξασφαλίζεται βιολογική ομοιογένεια και σταθερότητα. Προκύπτουν, αντίστοιχα, *BHI* = 0,23 και *BSI* = 0,34, οπότε η ομαδοποίηση δεν φαίνεται ικανοποιητική από βιολογικής σκοπιάς. Σημειώνεται ότι για τα συγκεκριμένα μέτρα πραγματοποιήθηκε σύγκριση με ήδη γνωστή προηγούμενη ομαδοποίηση.

Τα αποτελέσματα των μέτρων *Internal* και *Biological* δεν φαίνονται να είναι ικανοποιητικά ως προς την ομαδοποίηση των γονιδιακών εκφράσεων. Ωστόσο, ο D'haeseleer (2005), επισημαίνει ότι η '*χρήση αυτών των θεμελιωδών εργαλείων ομαδοποίησης γονιδιακών εκφράσεων είναι ακόμα εκπληκτικά ανεπαρκώς διατυπωμένα, (...) αλλά δεν υπάρχει αμφιβολία ότι πάντα θα υπάρχει περιθώριο να κυνηγήσουμε για αυτά τα απροσδόκητα πρότυπα τα οποία ενδέχεται να διαφανούν αν μόνο εξετάσουμε τα δεδομένα από την σωστή γωνία*'.

3.6 Ανάλυση γονιδιακών εκφράσεων κατά παράγοντες

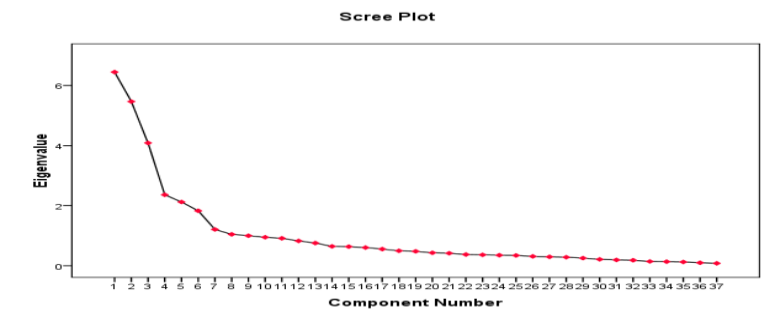
Πριν την εφαρμογή της *factor analysis,* κρίνεται σκόπιμο να ερευνηθεί αν τα δεδομένα των γονιδιακών εκφράσεων που διαθέτουμε είναι κατάλληλα (βλέπε παράγραφο 2.4.7). Για τον σκοπό αυτό παρατίθεται ο πίνακας με το μέτρο *KMO* και τον έλεγχο του *Bartlette*.

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,790
Bartlett's Test of Sphericity	Approx. Chi-Square	5699,403
	df	666
	Sig.	,000

Πίνακας 3.5: Αποτελέσματα μέτρου KMO και Bartlett's test

Το μέτρο *KMO* λαμβάνει υψηλή τιμή (*KMO* = 0,790), οπότε προκύπτει ότι τα δεδομένα υπό εξέταση είναι ικανοποιητικά για την εφαρμογή της *factor analysis*. Σε επίπεδο σημαντικότητας 5%, απορρίπτεται η μηδενική υπόθεση αφού προκύπτει p-value 'πρακτικά' μηδέν. Συνεπώς, ο πίνακας συσχετίσεων είναι διαφορετικός από τον μοναδιαίο και η ανάλυση κατά παράγοντες μπορεί να εφαρμοσθεί αφού υπάρχει υψηλή πιθανότητα να υπάρχουν σχέσεις μεταξύ των μεταβλητών.

Το επόμενο ζήτημα που πρέπει να δοθεί μία απάντηση από τον ερευνητή είναι σχετικά με το κατάλληλο πλήθος παραγόντων που πρέπει να λάβει υπόψη. Σύμφωνα με το *κριτήριο του Kaiser*, προκύπτουν συνολικά οκτώ παράγοντες (βλ. Παράρτημα Α2). Λαμβάνοντας το scree plot (Σχήμα 3.7) που προκύπτει από τα δεδομένα, παρατηρείται ότι αρκούν οι εφτά συνιστώσες. Τέλος, από το κριτήριο της ερμηνευτικής διακύμανσης και για λόγους οικονομίας, θεωρούμε ότι η αρκεί να ερμηνεύεται το 70% της μεταβλητότητας, οπότε καταλήγουμε στους δέκα παράγοντες. Λαμβάνοντας υπόψη ότι σύμφωνα με τα δύο πρώτα κριτήρια ερμηνεύεται το 66,392% και το 63,568% της μεταβλητότητας, αντίστοιχα, προκύπτει ότι ιδιαίτερη βαρύτητα πρέπει να δοθεί στο τρίτο κριτήριο. Συγκρίνοντας την ερμηνευτική μεταβλητότητα των εννέα και δέκα συνιστωσών, παρατηρούμε ότι η διαφορά ως προς την διακύμανση που ερμηνεύεται είναι αμελητέα (69,092% και 71,654%, αντίστοιχα). Οπότε, τελικά, καταλήγουμε στους εννέα παράγοντες για την περαιτέρω ανάλυση.



Σχήμα 3.7: Scree Plot

Θεωρώντας, λοιπόν, τις εννέα συνιστώσες, προκύπτει ο πίνακας του Παραρτήματος Α.3, στον οποίο διαχωρίζονται τα γονίδια ως προς τα φορτία τους στις συνιστώσες.

Σε αντιστοιχία με τα αποτελέσματα της μεθόδου *k-mean*, παρατηρούμε ότι η $1^{η}$ ομάδα αποτελείται από τη πρώτη συνιστώσα, η $2^{η}$ από τις 2, 4, 7 και 9 και η $3^{η}$ από τις 3, 5, και 6, με την $8^{η}$ συνιστώσα να μην είναι σαφές αν ανήκει στην $1^{η}$ ή την $2^{η}$ ομάδα, ωστόσο, θα μπορούσαμε να πούμε ότι αντιστοιχεί στην $1^{η}$ ομάδα δεδομένου ότι σε αυτή αντιστοιχεί το γονίδιο με το υψηλότερο loading. Παρατηρούμε ότι συνολικά 4 γονίδια ομαδοποιούνται σε διαφορετικές συστάδες, σύμφωνα με τις συνιστώσες, και αντιστοιχούν στο 10,81% του συνόλου των γονιδίων. Ιδιαίτερη σημασία έχει ο εντοπισμός της ηγούσα μεταβλητής (sarogate variable) σε κάθε συνιστώσα, η οποία αποτελεί το γονίδιο με το υψηλότερο loading στο προφίλ ομάδας. Αναλυτικά παρουσιάζονται στον Πίνακα 3.7.

Συνιστώσα	Sarogate	Γονίδια
1	ABAT	IL6ST-MLPH-ERBB3-MUC1-CHPT1-STC2-ERBB4-PTGER3-ALCAM
2	RACGAP1	TOP2A-UBE2c-BIRC5-CXCL12
3	VEGFR2	VEGFR1-VEGFR3-VEGFC-VEGFB
4	SPP1	MMP1-TUBB3-VEGFA
5	CD3D	IGKC-CXCL13
6	SFRP1	EGFR-MMP7-TP53
7	ERBB2	Herstatin
8	PVALB	PPIA
9	DHCR7	

Πίνακας 3.6: Αντιστοίχηση συνιστωσών με γονιδιακές εκφράσεις και ηγούσες μεταβλητές.

Κ Ε Φ Α Λ Α Ι Ο 4

Στοιχεία Ανάλυσης Επιβίωσης: Επισκόπηση της Θεωρίας

4.1 Εισαγωγή

Η *Ανάλυση Επιβίωσης* (*Survival Analysis*) ως έννοια χρησιμοποιείται για να περιγράψει την ανάλυση δεδομένων που αντιστοιχούν στον χρόνο από μία καλώς ορισμένη χρονική έναρξη έως την πραγματοποίηση ενός συγκεκριμένου γεγονότος ή ως ένα *τερματικό σημείο* (end point), σε συνδυασμό με την μελέτη δίτιμων αποτελεσμάτων για την επέλευση ή μη ενός γεγονότος. Ο όρος αυτός αντανακλά την προέλευση των μεθόδων αυτών από δημογραφικές μελέτες του προσδόκιμου ζωής. Στην ιατρική έρευνα, η χρονική αρχή συχνά αντιστοιχεί στην εισαγωγή ενός ατόμου σε μία πειραματική μελέτη, όπως μία κλινική δοκιμή. Αυτό με την σειρά του μπορεί να συμπίπτει με τη διάγνωση μιας συγκεκριμένης κατάστασης, την έναρξη ενός θεραπευτικού καθεστώτος ή την πραγματοποίηση κάποιου δυσμενούς γεγονότος. Εάν το τερματικό σημείο είναι ο θάνατος ενός ασθενή, τα τελικά αποτελέσματα είναι κυριολεκτικά χρόνοι επιβίωσης. Ωστόσο, δεδομένα παρόμοιας μορφής μπορούν να ληφθούν όταν το τερματικό σημείο δεν είναι μοιραίο, όπως η ανακούφιση από τον πόνο ή η επανεμφάνιση των συμπτωμάτων. Η μεθοδολογία της ανάλυσης επιβίωσης μπορεί να εφαρμοσθεί και σε άλλες περιοχές ενδιαφέροντος, όπως οι χρόνοι επιβίωσης των ζώων σε μία πειραματική μελέτη, ο χρόνος που απαιτείται από ένα άτομο να ολοκληρώσει ένα έργο σε ένα ψυχολογικό πείραμα ή ο χρόνος ζωής βιομηχανικών ή ηλεκτρολογικών εξαρτημάτων.

Στη συνέχεια του παρόντος κεφαλαίου περιγράφεται ο τρόπος εκτίμησης των συναρτήσεων επιβίωσης και κινδύνου των δεδομένων χρόνων επιβίωσης, καθώς επίσης παρουσιάζονται και τα ειδικά χαρακτηριστικά τω χρόνων αυτών. Ενδιαφέρον σημειώνεται στα μη συμμετρικά δεδομένα, στους λογοκριμένους χρόνους καθώς, επίσης, και στα πολλαπλά γεγονότα. Τέλος, γίνεται αναφορά στα διάφορα είδη λογοκρισίας.

4.2 Συνάρτηση επιβίωσης και συνάρτηση κινδύνου

Συνοψίζοντας τα δεδομένα επιβίωσης, υπάρχουν δύο συναρτήσεις κεντρικού ενδιαφέροντος, γνωστές ως *συνάρτηση επιβίωσης* (survival function) και *συνάρτηση κινδύνου* (hazard function).

Ο πραγματικός χρόνος επιβίωσης ενός ατόμου, *t*, μπορεί να θεωρηθεί ως η τιμή της μεταβλητής *T*, η οποία μπορεί να λαμβάνει μη αρνητικές τιμές. Οι διαφορετικές τιμές που μπορεί να λαμβάνει η *t* έχουν μία κατανομή πιθανότητας, και καλούμε την *T* τη τυχαία μεταβλητή που συνδέεται με τον χρόνο επιβίωσης. Έστω ότι η τυχαία μεταβλητή *T* έχει μία κατανομή πιθανότητας με μία βασική συνάρτηση πυκνότητας πιθανότητας *f(t)*. Η *συνάρτηση κατανομής* της *T* δίνεται από την σχέση

$$F(t) = P(T < t) = \int_0^t f(u)du ,$$

και αντιπροσωπεύει την πιθανότητα ο χρόνος επιβίωσης να είναι μικρότερος από κάποια τιμή *t*.

Η *συνάρτηση επιβίωσης*, *S(t)*, ορίζεται ως η πιθανότητα ότι ο χρόνος επιβίωσης είναι μεγαλύτερος ή ίσος του *t*, και επομένως

$$S(t) = P(T \geq t) = 1 - F(t).$$

Η *συνάρτηση επιβίωσης*, επομένως, μπορεί να αντιπροσωπεύσει την πιθανότητα ένα άτομο να επιζήσει από τον χρόνο έναρξης έως κάποιο χρόνο πέραν του *t*. Η συνάρτηση *S(t)* είναι φθίνουσα συνάρτηση, συνεχής, και ισχύουν *S(0)*=1 και $\lim_{t\to\infty} S(t) = 0$.

Η *συνάρτηση κινδύνου* είναι η πιθανότητα ένα άτομο να πεθάνει τον χρόνο *t*, δεδομένου ότι έχει επιζήσει έως αυτή τη χρονική στιγμή. Επομένως, η συνάρτηση αυτή δηλώνει την πιθανότητα στιγμιαίου θανάτου για ένα άτομο που έχει επιζήσει έως τον χρόνο *t*. Προκειμένου να λάβουμε ένα πιο επίσημο ορισμό της *συνάρτησης κινδύνου*, θεωρούμε την πιθανότητα η τυχαία μεταβλητή που σχετίζεται με τον χρόνο επιβίωσης του ατόμου, *T*, να λαμβάνει τιμές μεταξύ των *t* και *t*+*δt*, δεδομένου ότι η *T* είναι μεγαλύτερη ή ίση του *t*, $P(t \leq T \leq t + \delta t \mid T \geq t)$. Η συνάρτηση κινδύνου *h(t)* είναι το όριο της πιθανότητας αυτής διαιρούμενη με το χρονικό διάστημα *δt*, καθώς το *δt* τείνει στο μηδέν, δηλαδή,

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T \leq t + \delta t \mid T \geq t)}{\delta t} \right\}.$$

Από τον παραπάνω ορισμό, μπορούμε να λάβουμε κάποιες χρήσιμες σχέσεις μεταξύ των συναρτήσεων επιβίωσης και κινδύνου. Σύμφωνα με ένα τυπικό αποτέλεσμα της θεωρίας πιθανοτήτων, η πιθανότητα ενός γεγονότος Α, δεσμευμένο ως προς την πραγματοποίηση ενός γεγονότος Β δίνεται από την σχέση $P(A \mid B) = P(AB)/P(B)$, όπου $P(AB)$ είναι η πιθανότητα να πραγματοποιηθούν από κοινού τα ενδεχόμενα Α και Β. Θεωρώντας αυτό το αποτέλεσμα, η δεσμευμένη πιθανότητα στον ορισμό της συνάρτησης κινδύνου είναι

$$\frac{P(t \le T \le t + \delta t)}{P(T \ge t)} = \frac{F(t+\delta t) - F(t)}{S(t)},$$

όπου *F(t)* είναι η συνάρτηση κατανομής της *T*. Οπότε,

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

Η ποσότητα $\lim_{\delta t \to 0} \left\{ \frac{F(t+\delta t) - F(t)}{\delta t} \right\}$ είναι ο ορισμός της παραγώγου της *F(t)* ως προς *t*, η οποία είναι η *f(t)*, οπότε έχουμε

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

Επομένως, έχουμε

$$h(t) = -\tfrac{d}{dt} \{\log S(t)\},$$

και

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

όπου

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

Η συνάρτηση *H(t)* χαρακτηρίζει ευρέως την ανάλυση επιβίωσης, καλείται *αθροιστικός κίνδυνος* και ορίζεται ως

$$H(t) = -\log S(t).$$

Στην ανάλυση των δεδομένων επιβίωσης, οι συναρτήσεις επιβίωσης και κινδύνου εκτιμούνται από τους παρατηρούμενους χρόνους επιβίωσης. Η ανάλυση χρόνων επιβίωσης γίνεται με *παραμετρικές* και *μη-παραμετρικές* μεθόδους. Οι παραμετρικές βασίζονται στην υπόθεση κάποιας κατανομής του χρόνου επιβίωσης ενώ οι μη-παραμετρικές δεν απαιτούν τον προσδιορισμό της συνάρτησης πυκνότητας πιθανότητας της *T*.

4.3 Ειδικά χαρακτηριστικά των δεδομένων επιβίωσης

Τα δεδομένα επιβίωσης παρουσιάζουν κάποιες ιδιαιτερότητες που καθιστούν τις κλασικές στατιστικές διαδικασίες ακατάλληλες. Οι κυριότερες από αυτές είναι (i) η μη-συμμετρικότητα των δεδομένων, (ii) η λογοκρισία των δεδομένων και (iii) η εμφάνιση πολλαπλών γεγονότων. Στις υπο-ενότητες που ακολουθούν, θα αναφερθούμε σύντομα σε αυτές.

4.3.1 Μη-συμμετρικά δεδομένα

Ο πρώτος λόγος είναι ότι τα δεδομένα επιβίωσης δεν είναι συμμετρικά κατανεμημένα. Τυπικά, ένα ιστόγραμμα χρόνων επιβίωσης τείνει να είναι λοξό από τα δεξιά, που σημαίνει ότι το ιστόγραμμα έχει βαρύτερη ‘ουρά’ στο δεξί διάστημα που περιέχει τον μεγαλύτερο αριθμό παρατηρήσεων. Ως αποτέλεσμα, δεν θα ήταν εύλογο να υποθέσουμε ότι δεδομένα αυτού του τύπου ακολουθούν την κανονική κατανομή. Η δυσκολία αυτή θα μπορούσε να αντιμετωπισθεί μετασχηματίζοντας, αρχικά, τα δεδομένα ώστε να έχουν μία πιο συμμετρική κατανομή. Ένας τρόπος να επιτευχθεί αυτό είναι να λογαριθμίσουμε. Ωστόσο, μια πιο ικανοποιητική προσέγγιση είναι να υιοθετήσουμε ένα εναλλακτικό μοντέλο για τα αρχικά δεδομένα.

4.3.2 Λογοκριμένα δεδομένα

Ένα δεύτερο χαρακτηριστικό των δεδομένων επιβίωσης, το οποίο καθιστά τις κλασικές μεθόδους ακατάλληλες, είναι ότι οι χρόνοι των δεδομένων επιβίωσης είναι συχνά λογοκριμένοι. Ο χρόνος επιβίωσης ενός ατόμου θεωρείται λογοκριμένος όταν το τελικό σημείο του ενδιαφέροντός μας δεν έχει παρατηρηθεί για το άτομο αυτό. Το γεγονός αυτό προκύπτει στις εξής καταστάσεις:

(α) Ο χρόνος επιβίωσης του εν λόγω ατόμου ξεπέρασε το χρονικό ορίζοντα της έρευνας. Συνεπώς, κατά τον χρόνο αξιολόγησης η μόνη γνωστή πληροφορία είναι ότι ακόμα δεν έχει επέλθει το γεγονός.

(β) Το άτομο αυτό «χάθηκε» κατά τον χρόνο παρακολούθησης (περίοδος follow-up), πχ. αποχώρησε πριν τη λήξη της έρευνας, σταμάτησε τη θεραπευτική αγωγή λόγω παρενεργειών ή αρνήθηκε να συνεχίσει, (Collet, 1994).

Σημείο-κλειδί αποτελεί το γεγονός ότι η ανάλυση πρέπει να περιλαμβάνει όρους που να αντιστοιχούν σε ό,τι έχουμε παρατηρήσει. Αυτό σημαίνει ότι εάν ένας ακριβής χρόνος ζωής έχει παρατηρηθεί, θα συνεισφέρει με την πυκνότητά του, και εάν ο χρόνος αυτός είναι λογοκριμένος, θα συνεισφέρει με την πιθανότητα ότι ο χρόνος αυτός ζωής ξεπερνά το χρόνο λήξης της έρευνας.

Ένας πραγματικός χρόνος επιβίωσης μπορεί ακόμη να θεωρηθεί ως λογοκριμένος όταν ο θάνατος οφείλεται σε ένα γεγονός που είναι γνωστό ότι δεν σχετίζεται με το αντικείμενο που εξετάζεται. Σε πολλές περιπτώσεις, είναι δύσκολο να είμαστε σίγουροι πως ο θάνατος δεν οφείλεται, για παράδειγμα, σε μία συγκεκριμένη θεραπεία στην οποία υποβλήθηκε ο ασθενής.

Λόγω της σημαντικότητας της λογοκρισίας σε μελέτες επιβίωσης, στοιχεία λογοκρισίας θα δοθούν σε ανεξάρτητη ενότητα.

4.3.3 Πολλαπλά γεγονότα

Προκειμένου να εφαρμόσουμε διατάξιμες μεθόδους σε δεδομένα επιβίωσης, είναι απαραίτητος ο ορισμός κάποιων τυχαίων μεταβλητών. Για πραγματικούς χρόνους ζωής, αυτό είναι εύκολο, διότι γνωρίζουμε ότι ο χρόνος ζωής είναι καλώς ορισμένος, ακόμα και στην περίπτωση που δεν έχει παρατηρηθεί, λόγω της παρουσίας λογοκριμένων δεδομένων. Σταδιακά όλοι οι άνθρωποι θα αποβιώσουν. Ωστόσο, για πολλά άλλα γεγονότα, δεν μπορούμε να είμαστε σίγουροι ότι το γεγονός που μελετάται θα συμβεί. Στην περίπτωση που μελετάμε κάποια ασθένεια, ένα άτομο είναι πιθανό να εμφανίσει ή όχι την ασθένεια υπό εξέταση, ακόμα και σε παρακολούθηση μακράς διάρκειας. Προκειμένου να ορισθεί μία τυχαία μεταβλητή αντίστοιχη της ασθένειας, πρέπει να επιτρέψουμε την τιμή του απείρου. Αυτό δεν είναι κατάλληλο για τα τυπικά μοντέλα. Για άλλες ασθένειες, ένα άτομο μπορεί να σημειώσει πολλαπλή προσβολή της ίδιας ασθένειας, καθιστώντας αδύνατη την προσαρμογή σε ένα κλασικό πρότυπο σχεδιασμού του πειράματος. Το πρότυπο των δεδομένων επιβίωσης που βασίζεται στον κίνδυνο επιτρέπει την πιθανότητα ύπαρξης μηδενικών ή πολλαπλών γεγονότων για κάθε άτομο, (Hougaard, 2001).

4.4 Γενικά χαρακτηριστικά και είδη λογοκρισίας

Σε μία τυπική μελέτη, οι ασθενείς δεν εισάγονται την ίδια χρονική στιγμή, αλλά συσσωρεύονται κατά μία χρονική περίοδο μηνών, ή ακόμα ετών. Μετά την εισαγωγή τους οι ασθενείς παρακολουθούνται έως να αποβιώσουν, ή έως ένα ημερολογιακό σημείο το οποίο σηματοδοτεί την λήξη της έρευνας, κατά την ανάλυση των δεδομένων. Η ημερολογιακή χρονική περίοδος κατά την οποία ένα άτομο βρίσκεται στην έρευνα είναι γνωστή ως ο *χρόνος έρευνας* (study time). Η χρονική περίοδος την οποία ένας ασθενής περνά εντός της έρευνας, υπολογιζόμενη από την εισαγωγή του στην έρευνα, αναφέρεται συνήθως ως *χρόνος ασθενούς* (patient time). Συνεπώς, η περίοδος που μετράται από την εισαγωγή στην μελέτη έως τον θάνατο του ασθενούς, αντιστοιχεί στον *χρόνο επιβίωσης*.

Αρχικά, πρέπει να γίνουν κάποιες υποθέσεις σχετικά με τον μηχανισμό της λογοκρισίας. Η λογοκρισία πρέπει να είναι ασυσχέτιστη με την μελλοντική διάρκεια ζωής, και αυτή η απαίτηση πρέπει να είναι ρητή. Επιπλέον, μία σημαντική υπόθεση η οποία λαμβάνεται στην ανάλυση λογοκριμένων δεδομένων επιβίωσης είναι ότι ο πραγματικός χρόνος επιβίωσης ενός ατόμου, t, είναι ανεξάρτητος από κάθε μηχανισμό που προκαλεί τη λογοκρισία των δεδομένων επιβίωσης σε χρόνο t_c, όπου $t_c < t$. Αυτό σημαίνει πως αν θεωρήσουμε μία ομάδα ατόμων, στην οποία όλα τα άτομα έχουν τις ίδιες τιμές ως προς τους σχετικούς προγνωστικούς παράγοντες, τότε ένα άτομο, του οποίου ο χρόνος λογοκρίνεται τη χρονική στιγμή t_c, πρέπει να αντιπροσωπεύεται από τα υπόλοιπα άτομα της ομάδας που έχουν επιζήσει σε αυτή τη χρονική στιγμή. Ένα άτομο του οποίου ο χρόνος επιβίωσης είναι λογοκριμένος θα αντιπροσωπεύεται από εκείνους που βρίσκονται σε κίνδυνο την στιγμή λογοκρισίας, εάν η διαδικασία λογοκρισίας πραγματοποιείται με τυχαίο τρόπο. Ομοίως, όταν τα δεδομένα πρόκειται να αναλυθούν σε ένα προκαθορισμένο σημείο στον ημερολογιακό χρόνο, ή σε σταθερό χρονικό διάστημα μετά τον χρόνο εισαγωγής κάθε ασθενή, η πρόγνωση των ατόμων που βρίσκονται ακόμα εν ζωή μπορούν να θεωρηθούν ανεξάρτητα ως προς την λογοκρισία, εφόσον ο χρόνος ανάλυσης είναι καθορισμένος πριν την εξέταση των δεδομένων. Ωστόσο, η υπόθεση αυτή δεν μπορεί να απαιτείται εάν, για παράδειγμα, ο χρόνος επιβίωσης ενός ατόμου είναι λογοκριμένος αφού κριθεί απαραίτητη η απόσυρση της θεραπείας ως αποτέλεσμα της υποβάθμισης της φυσικής κατάστασης του ασθενή. Αυτό το είδος λογοκρισίας είναι γνωστό ως *ειδοποιητήρια λογοκρισία* (informative censoring). Ιδιαίτερη προσοχή πρέπει να δίνεται προκειμένου να επιβεβαιωθεί ότι κάθε λογοκρισία που

σημειώνεται δεν είναι ειδοποιητήρια, διαφορετικά οι μέθοδοι ανάλυσης που ακολουθούν δεν μπορούν να εφαρμοσθούν.

4.4.1 Δεξιά λογοκρισία

Σε κάθε περίπτωση, ένας ασθενής εισάγεται στην έρευνας σε χρόνο t_0 και επέρχεται το γεγονός σε χρόνο t_0+t. Ωστόσο το χρονικό διάστημα t είναι άγνωστο, είτε επειδή ο ασθενής βρίσκεται ακόμα εν ζωή, είτε επειδή έχει χαθεί κατά την περίοδο follow-up. Εάν είναι γνωστό ότι το άτομο ζούσε τη χρονική στιγμή t_0+t_c, ο χρόνος t_c καλείται λογοκριμένος χρόνος επιβίωσης. Αυτή η λογοκρισία δημιουργείται αφού το άτομο εισαχθεί στην έρευνα, και αναφέρεται στον τελευταίο γνωστό χρόνο επιβίωσης (από δεξιά), και επομένως είναι γνωστή ως *δεξιά λογοκρισία*. Ο δεξιά λογοκριμένος χρόνος επιβίωσης είναι, επομένως, μικρότερος από τον πραγματικό, αλλά άγνωστος. Όσον αφορά την δεξιά λογοκρισία είναι δυνατόν να θεωρήσουμε δύο διαφορετικούς τύπους αυτής.

4.4.1.1 Δεξιά λογοκρισία τύπου I

Η *δεξιά λογοκρισία τύπου I* είναι αυτή κατά την οποία γνωρίζουμε ότι ο χρόνος ζωής του ατόμου είναι μεγαλύτερος του χρόνου λογοκρισίας $t_0+ t_c$ και είναι αυτή που περιγράφηκε παραπάνω. Μια πιο σύνθετη περίπτωση *λογοκρισίας τύπου I* μπορεί να προκύψει όταν η διάρκεια παρακολούθησης των ατόμων αν και είναι γνωστή για κάθε άτομο δεν είναι η ίδια για όλα τα άτομα. Σε αυτές τις περιπτώσεις ο χρόνος παρακολούθησης κάθε ατόμου δεν συμπίπτει αναγκαστικά με τον χρόνο έναρξης της έρευνας.

4.4.1.2 Δεξιά λογοκρισία τύπου II

Στην περίπτωση της *λογοκρισίας τύπου II*, αποφασίζεται από την αρχή της έρευνας ότι αυτή θα τερματισθεί όταν αποτύχουν συνολικά r άτομα. Επομένως, τα δεδομένα αποτελούνται από τους πλήρεις χρόνους ζωής των πρώτων r ατόμων που απέτυχαν, ενώ για τα υπόλοιπα n-r άτομα γνωρίζουμε ότι ο χρόνος ζωής τους είναι μεγαλύτερος του μέγιστου χρόνου ζωής των r ατόμων. Το σημαντικότερο μειονέκτημα της δεξιάς λογοκρισίας τύπου II

είναι ότι ο συνολικός χρόνος διάρκειας της έρευνας δεν είναι γνωστός (Σημειώσεις Δ. Αντζουλάκου, 2009).

4.4.2 Αριστερή και διπλή λογοκρισία

Ένας άλλος τύπος λογοκρισίας είναι η *αριστερή λογοκρισία*, η οποία παρουσιάζεται όταν ο πραγματικός χρόνος επιβίωσης είναι μικρότερος αυτού που παρατηρήθηκε. Έστω ότι το ενδιαφέρον μιας έρευνας επικεντρώνεται στην επανεμφάνιση συγκεκριμένου τύπου καρκίνου, μετά από επέμβαση αφαίρεσης του αρχικού όγκου. Τρεις μήνες μετά την επέμβαση εξετάζεται η επανεμφάνιση του καρκίνου. Αυτή τη χρονική στιγμή κάποιοι από τους ασθενείς πιθανώς να έχουν επανεμφανίσει τον όγκο. Για αυτούς τους ασθενείς ο πραγματικός χρόνος επανεμφάνισης είναι μικρότερος των τριών μηνών, οπότε, ορίζουμε ότι ο χρόνος επανεμφάνισης είναι *αριστερά λογοκριμένος*. Στην ειδική περίπτωση που η έρευνα έχει προκαθορισμένο χρόνο λήξης, μπορεί να έχουμε ταυτόχρονα λογοκριμένα δεδομένα και από δεξιά και από αριστερά, οπότε ομιλούμε για *διπλή λογοκρισία* (double censored).

4.4.3 Διαστηματική λογοκρισία και περικομμένα δεδομένα

Ένας επιπλέον τύπος λογοκρισίας είναι η *διαστηματική*. Σε αυτή τη περίπτωση, γνωρίζουμε πως τα άτομα έχουν βιώσει μία αποτυχία εντός ενός χρονικού διαστήματος. Στο προηγούμενο παράδειγμα, εάν ένας ασθενής δεν εμφανίσει καρκίνο μετά τους πρώτους τρεις μήνες της επέμβασης και σε μία επόμενη εξέταση διαπιστωθεί η επανεμφάνιση του όγκου, ο πραγματικός χρόνος επανεμφάνισης είναι μεταξύ τριών και έξι μηνών. Τότε, ο παρατηρούμενος χρόνος επανεμφάνισης είναι *λογοκριμένος σε διάστημα*. Τέλος, σημειώνονται στην πράξη τα *περικομμένα δεδομένα* (truncated data). Η *περικοπή* ορίζεται ως μία συνθήκη την οποία πρέπει να ικανοποιούν τα άτομα που θα συμμετάσχουν στην έρευνα. Τα άτομα που δεν ικανοποιούν την συνθήκη δεν τελούν υπό παρακολούθηση και ο ερευνητής αγνοεί την ύπαρξή τους.

Η λογοκρισία δεν αποτελεί πρόβλημα αποκλειστικά για τα δεδομένα επιβίωσης. Όπως κάθε μηχανισμός μέτρησης, έχει ένα εύρος μέσα στο οποίο μπορεί να λάβει τιμές και εκτός αυτού μπορούμε μόνο να καταλήξουμε στο συμπέρασμα ότι βρίσκεται εκτός του διαστήματος αυτού. Για παράδειγμα, ένα θερμόμετρο μπορεί να λάβει τιμές μόνο σε ένα

συγκεκριμένο εύρος. Αυτό είναι επίσης ένα κοινό πρόβλημα σε χημικές μεθόδους μετρήσεων της συγκέντρωσης κάποιας ουσίας στο αίμα, όπου υπάρχει κάποιο όριο ανίχνευσης, κάτω από το οποίο αυτό που γνωρίζουμε είναι ότι η συγκέντρωση είναι χαμηλή.

Κ Ε Φ Α Λ Α Ι Ο 5

Μη Παραμετρική Συμπερασματολογία Χρόνων Επιβίωσης

5.1 Εισαγωγή

Η μη παραμετρική εκτίμηση πραγματοποιείται μέσω των εκτιμητών *Kaplan-Meier* και *Nelson-Aalen* οι οποίοι παρουσιάζονται στο παρόν κεφάλαιο. Παρατίθενται η εκτίμηση των τυπικών σφαλμάτων και διαστημάτων εμπιστοσύνης για την συνάρτηση επιβίωσης, καθώς επίσης και η εκτίμηση της συνάρτησης κινδύνου και της αθροιστικής συνάρτησης κινδύνου. Τέλος, παρουσιάζεται η εκτίμηση της διαμέσου και των ποσοστιαίων σημείων, καθώς και τα αντίστοιχα διαστήματα εμπιστοσύνης.

5.2 Εκτίμηση της συνάρτησης επιβίωσης

Η απλούστερη μη παραμετρική εκτίμηση της συνάρτησης κατανομής είναι η εμπειρική κατανομή, που σημαίνει ότι η κατανομή θεωρείται ισοδύναμη με την παρατηρούμενη. Αυτή είναι η μη παραμετρική εκτίμηση για το σύνολο των παρατηρήσεων. Επομένως, παρ' όλο που υποθέτουμε ότι η πραγματική κατανομή είναι συνεχής, την εκτιμούμε μέσω μιας διακριτής κατανομής. Αρχικά υποθέτουμε ότι έχουμε στην διάθεσή μας ένα δείγμα χρόνων επιβίωσης, στο οποίο καμία από τις παρατηρήσεις δεν είναι λογοκριμένες. Η συνάρτηση επιβίωσης, *S(t)*, όπως ορίστηκε παραπάνω, είναι η πιθανότητα ενός ατόμου να επιζήσει για χρόνο μεγαλύτερο ή ίσο του *t*. Η συνάρτηση αυτή μπορεί να εκτιμηθεί από την *εμπειρική συνάρτηση επιβίωσης* (empirical survival function), η οποία δίνεται από την σχέση,

$$\hat{S}(t) = \frac{\Pi\lambda\acute{\eta}\theta o\varsigma\ \alpha\tau\acute{o}\mu\omega\nu\ \mu\varepsilon\ \chi\rho\acute{o}\nu o\upsilon\varsigma\ \varepsilon\pi\iota\beta\acute{\iota}\omega\sigma\eta\varsigma \geq t}{\Pi\lambda\acute{\eta}\theta o\varsigma\ \alpha\tau\acute{o}\mu\omega\nu\ \sigma\tau o\ \sigma\acute{\upsilon}\nu o\lambda o\ \delta\varepsilon\delta o\mu\acute{\varepsilon}\nu\omega\nu}. \qquad (5.1)$$

Ισοδύναμα, έχουμε $\hat{S}(t)=1-\hat{F}(t)$, όπου $\hat{F}(t)$ είναι η εμπειρική συνάρτηση κατανομής, που σημαίνει, το πλήθος των ατόμων εν ζωή στον χρόνο *t* προς το σύνολο των ατόμων στην έρευνα. Σημειώνεται ότι η εμπειρική συνάρτηση επιβίωσης ισούται με την μονάδα για χρόνο *t* μικρότερο του χρόνου που αντιστοιχεί στον πρώτο θάνατο, και με το μηδέν μετά τον τελευταίο χρόνο θανάτου.

Η εκτιμώμενη συνάρτηση επιβίωσης, $\hat{S}(t)$, υποθέτουμε πως είναι σταθερή μεταξύ δύο διαδοχικών χρόνων θανάτου, και επομένως ένα διάγραμμα της $\hat{S}(t)$ ως προς το χρόνο *t* είναι σκαλωτό. Η συνάρτηση μειώνεται αμέσως μετά από κάθε παρατηρούμενο χρόνο επιβίωσης. Οι Kaplan and Meier (1958) πρότειναν μία προέκταση των λογοκριμένων δεδομένων, διατυπωμένη σε όρους της συνάρτησης επιβίωσης. Με μία παραλλαγή προκύπτει ένας ελαφρώς διαφορετικός εκτιμητής, ο Nelson-Aalen.

5.3 Εκτιμητής Kaplan-Meier της συνάρτησης επιβίωσης

Για να καθορίσουμε την εκτίμηση *Kaplan-Meier* ή εκτίμηση *γινομένου-ορίου* (product limit-PLE) για ένα δείγμα λογοκριμένων δεδομένων επιβίωσης, απαιτείται μία σειρά χρονικών διαστημάτων. Ωστόσο, κάθε ένα από αυτά τα διαστήματα κατασκευάζεται έτσι ώστε να περιλαμβάνει ένα χρόνο θανάτου, και αυτός ο χρόνος θανάτου να θεωρείται ότι βρίσκεται στην αρχή του διαστήματος.

Γενικά, έστω ότι υπάρχουν *n* άτομα με παρατηρούμενους χρόνους επιβίωσης t_1, t_2,..,t_n. Κάποιες από αυτές τις παρατηρήσεις μπορεί να είναι *δεξιά λογοκριμένες*, και ίσως να υπάρχουν περισσότερα του ενός άτομα με τον ίδιο χρόνο επιβίωσης. Επομένως, μπορούμε να υποθέσουμε ότι υπάρχουν *r* χρόνοι θανάτου μεταξύ των ατόμων, όπου $r \le n$. Αφού διαταχθούν οι χρόνοι αυτοί σε αύξουσα τάξη μεγέθους, ο $j^{ος}$ χρόνος συμβολίζεται ως $t_{(j)}$, για j=1, 2,.., *r*, οπότε, οι *r* διατεταγμένοι χρόνοι είναι $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$. Το πλήθος των ατόμων που είναι εν ζωή ακριβώς πριν τον χρόνο $t_{(j)}$, συμπεριλαμβανομένων των ατόμων που πρόκειται να πεθάνουν στο χρονικό αυτό σημείο, συμβολίζονται ως n_j, για j=1, 2,.., *r*, και με d_j συμβολίζουμε αυτούς που πεθαίνουν στον ίδιο χρόνο. Το χρονικό διάστημα από $t_{(j)}$-δ έως $t_{(j)}$, όπου δ είναι ένα απειροστό χρονικό διάστημα, περιλαμβάνει, επομένως, ένα χρόνο θανάτου. Εφόσον υπάρχουν n_j άτομα που βρίσκονται εν ζωή ακριβώς πριν τον χρόνο $t_{(j)}$ και d_j θάνατοι ακριβώς στον χρόνο $t_{(j)}$, η πιθανότητα ότι το άτομο πεθαίνει κατά την διάρκεια του

διαστήματος $t_{(j)}$-δ έως $t_{(j)}$ εκτιμάται ως d_j / n_j. Η αντίστοιχη εκτιμώμενη πιθανότητα επιβίωσης μέσω αυτού του διαστήματος είναι $(n_j - d_j) / n_j$.

Κάποιες φορές συμβαίνει να υπάρχουν λογοκριμένοι χρόνοι επιβίωσης οι οποίοι πραγματοποιούνται στον ίδιο χρόνο με έναν ή περισσότερους θανάτους, έτσι ώστε ένας χρόνος θανάτου και ένας λογοκριμένος χρόνος θανάτου να πραγματοποιούνται ταυτόχρονα. Σε αυτή τη περίπτωση, ο λογοκριμένος χρόνος επιβίωσης θεωρούμε πως πραγματοποιείται αμέσως μετά τον χρόνο θανάτου κατά τον υπολογισμό των n_j.

Από τον τρόπο που κατασκευάζονται τα διαστήματα χρόνου, το διάστημα από $t_{(j)}$ έως $t_{(j+1)}$-δ, που είναι ο χρόνος ακριβώς πριν τον επόμενο χρόνο θανάτου, δεν περιέχει κανένα θάνατο. Η πιθανότητα να επιβιώσει στο διάστημα από $t_{(j)}$ έως $t_{(j+1)}$-δ είναι, επομένως, η μονάδα, και η από κοινού πιθανότητα επιβίωσης από $t_{(j)}$-δ έως $t_{(j)}$ και $t_{(j)}$ έως $t_{(j+1)}$-δ μπορεί να εκτιμηθεί ως $(n_j - d_j) / n_j$. Θεωρώντας το όριο, καθώς το δ τείνει στο μηδέν, η ποσότητα $(n_j - d_j) / n_j$ είναι η εκτίμηση της πιθανότητας επιβίωσης από $t_{(j)}$ έως $t_{(j+1)}$.

Στην συνέχεια, υποθέτουμε πως οι θάνατοι των ατόμων στο δείγμα πραγματοποιούνται ανεξάρτητα ο ένας από τον άλλο. Τότε, η εκτιμώμενη συνάρτηση επιβίωσης σε κάθε χρονικό σημείο του $k^{ου}$ κατασκευαζόμενου χρονικού διαστήματος από $t_{(k)}$ έως $t_{(k+1)}$, k=1, 2, .., r, όπου $t_{(r+1)}$ ορίζεται να είναι ∞, θα είναι η εκτιμώμενη πιθανότητα επιβίωσης πέραν του $t_{(k)}$. Αυτή είναι, στην πραγματικότητα, η πιθανότητα επιβίωσης κατά τη διάρκεια του διαστήματος από $t_{(k)}$ έως $t_{(k+1)}$, και όλα τα προηγούμενα διαστήματα. Αυτός είναι ο εκτιμητής *Kaplan-Meier* της συνάρτησης επιβίωσης, ο οποίος δίνεται από την σχέση

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right), \tag{5.2}$$

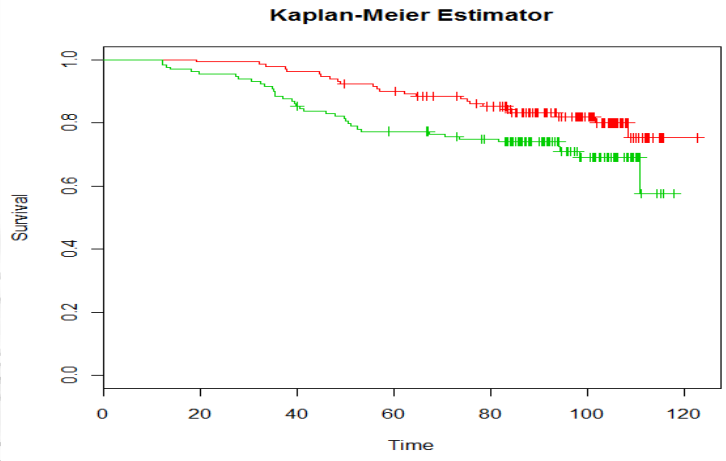
για $t_{(k)} \le t \le t_{(k+1)}$, k=1, 2,.., r, με $\hat{S}(t)$= 1 για $t<t_{(1)}$ και $t_{(r+1)}$ ορίζεται να είναι ∞. Στην περίπτωση που η μεγαλύτερη παρατήρηση χρόνου επιβίωσης, $t_{(r)}$, είναι λογοκριμένος χρόνος ζωής, έστω t^*, η $\hat{S}(t)$ δεν ορίζεται για $t>t^*$. Από την άλλη μεριά, εάν η μεγαλύτερη παρατήρηση χρόνου επιβίωσης, $t_{(r)}$, είναι μη λογοκριμένη παρατήρηση, $n_r=d_r$, τότε η $\hat{S}(t)$ είναι μηδενική για $t \ge t_{(r)}$. Μία συνέπεια της πρώτης περίπτωσης είναι ότι η μέση διάρκεια ζωής δεν μπορεί να εκτιμηθεί. Μία λύση είναι να υπολογίσουμε τον μέσο περιοριζόμενοι στην παρατηρούμενη περίοδο, η οποία λαμβάνεται υποθέτοντας ότι η συνάρτηση επιβίωσης είναι μηδενική μετά τον μέγιστο αυτό χρόνο. Αυτό είναι το αντίστοιχο κάτω φράγμα των παρατηρήσεων, και προφανώς χαρακτηρίζεται από μεροληψία. Το άνω φράγμα της μέσης

διάρκειας ζωής θα είναι άπειρο. Μία καλύτερη λύση είναι η εκτίμηση της διαμέσου, η οποία μπορεί να καθορισθεί όταν το μήκος της παρατηρούμενης περιόδου είναι αρκετά μακρύ ώστε να ξεπερνά το ½. Ένα γράφημα του εκτιμητή *Kaplan-Meier* της συνάρτησης επιβίωσης είναι μία σκαλωτή συνάρτηση, στην οποία οι εκτιμώμενες πιθανότητες επιβίωσης είναι σταθερές μεταξύ διαδοχικών χρόνων θανάτου και μειώνονται σε κάθε χρόνο θανάτου.

Ο εκτιμητής *Kaplan-Meier* είναι, επίσης, γνωστός ως *εκτιμητής γινομένου-ορίου* (*product-limit estimation*) της συνάρτησης επιβίωσης. Εάν δεν υπάρχουν λογοκριμένοι χρόνοι στα δεδομένα, n_j-d_j=n_{j+1}, j=1, 2,…, k, στην τελευταία σχέση, και επεκτείνοντας το γινόμενο, λαμβάνουμε

$$\hat{S}(t)=\frac{n_2}{n_1}\times\frac{n_3}{n_2}\times\ldots\times\frac{n_{k+1}}{n_k}.$$

Το παραπάνω γινόμενο απλοποιείται σε n_{k+1}/n_1, για k=1, 2, .., r-1, με $\hat{S}(t)$=1 για $t<t_{(1)}$ και $\hat{S}(t)$=0 για $t\geq t_{(r)}$. Ως n_1 ορίζουμε το πλήθος των ατόμων που βρίσκονται σε κίνδυνο πριν τον πρώτο χρόνο θανάτου, το οποίο είναι το πλήθος των ατόμων στο δείγμα, και n_{k+1} είναι το πλήθος των ατόμων με χρόνους επιβίωσης μεγαλύτερους ή ίσους του $t_{(k+1)}$. Επομένως, με την απουσία λογοκρισίας, η $\hat{S}(t)$ είναι απλώς η εμπειρική συνάρτηση επιβίωσης όπως ορίστηκε στη σχέση (5.1).



Σχήμα 5.1: Γραφική παρουσίαση του εκτιμητή *Kaplan-Meier* δύο ομάδων.

Η γραφική αναπαράσταση του εκτιμητή Kaplan-Meier παρέχει αμεσότερη ερμηνεία. Οι γραμμές που αντιπροσωπεύουν τις συναρτήσεις ‘πέφτουν’ όταν σημειώνεται επέλευση του γεγονότος, και παραμένουν σταθερές στις ενδιάμεσες περιόδους. Ως αποτέλεσμα, μπορούν να χαρακτηριστούν περίοδοι υψηλού κινδύνου, όταν η καμπύλη επιβίωσης μειώνεται ραγδαία,

καθώς και περιόδους χαμηλού κινδύνου, όταν η καμπύλη παραμένει σχετικά ευθεία. Επιπλέον, μπορεί να γίνει σύγκριση δύο ή περισσότερων συναρτήσεων που αντιστοιχούν σε διαφορετικές ομάδες κάποιου κλινικού χαρακτηριστικού. Η γραμμή που βρίσκεται υψηλότερα αντιστοιχεί σε μεγαλύτερη πιθανότητα επιβίωσης (Vittinghoff *et al.*, 2005).

5.3.1 Τυπικό σφάλμα του εκτιμητή *Kaplan-Meier*

Ο υπολογισμός του τυπικού σφάλματος είναι σημαντικός για την ασυμπτωτική συμπερασματολογία αφού απαιτείται για τον υπολογισμό διαστημάτων εμπιστοσύνης (βλ. παραγράφους 7.3 και 7.5). Ο εκτιμητής Kaplan-Meier της συνάρτησης επιβίωσης για κάθε τιμή του χρόνου *t* εντός του διαστήματος από $t_{(k)}$ έως $t_{(k+1)}$ μπορεί να γραφεί ως

$$\hat{S}(t) = \prod_{j=1}^{k} \hat{p}_j ,$$

για $k = 1, 2,\ldots, r$, όπου $p_j = (n_j - d_j)/n_j$ είναι η εκτιμούμενη πιθανότητα ένα άτομο να επιβιώσει κατά το χρονικό διάστημα με έναρξη την στιγμή $t_{(j)}$, $j=1, 2,\ldots, r$. Λαμβάνοντας λογαρίθμους και θεωρώντας την αντίστοιχη διακύμανση έχουμε,

$$\log \hat{S}(t) = \sum_{j=1}^{k} \log \hat{p}_j \quad \text{και} \quad Var\left\{ \log \hat{S}(t) \right\} = \sum_{j=1}^{k} Var\left\{ \log \hat{p}_j \right\}.$$

Μπορούμε να υποθέσουμε πως το πλήθος των ατόμων που επιβιώνουν κατά το διάστημα με έναρξη στο χρόνο $t_{(j)}$ ακολουθεί διωνυμική κατανομή με παραμέτρους n_j και p_j, όπου p_j είναι η πραγματική πιθανότητα επιβίωσης σε αυτό το διάστημα. Το πραγματικό πλήθος επιζώντων είναι n_j-d_j, και χρησιμοποιώντας το γνωστό αποτέλεσμα ότι η διακύμανση μιας τυχαίας διωνυμικής κατανομής με παραμέτρους *n* και *p* είναι *np(1-p)*, η διακύμανση της ποσότητας n_j-d_j δίνεται από την σχέση

$$Var(n_j - d_j) = n_j p_j (1 - p_j).$$

Εφόσον $\hat{p}_j = (n_j - d_j)/n_j$, έχουμε

$$Var\,\hat{p}_j = \tfrac{1}{n_j^2} Var(n_j - d_j) = p_j(1 - p_j)/n_j .$$

Προκειμένου να έχουμε την διακύμανση του λογαρίθμου του $\hat{p}_j$, χρησιμοποιούμε το γενικό αποτέλεσμα για την προσέγγιση της διακύμανσης μιας συνάρτησης της τυχαίας μεταβλητής. Σύμφωνα με αυτό, η διακύμανση μιας συνάρτησης g(X) της τυχαίας μεταβλητής X δίνεται από την σχέση,

$$Var\{g(X)\} \approx \left\{\frac{dg(X)}{dX}\right\}^2 Var(X). \tag{5.3}$$

Οπότε,

$$Var \log \hat{p}_j = Var\left(\hat{p}_j\right) / \hat{p}_j^2 = \left(1-\hat{p}_j\right)/n_j\hat{p}_j ,$$

και δεδομένης της σχέσης $\hat{p}_j = \left(n_j - d_j\right)/n_j$, γίνεται

$$Var \log \hat{p}_j = \frac{d_j}{n_j\left(n_j - d_j\right)}.$$

Επιστρέφοντας στην συνάρτηση επιβίωσης, έχουμε,

$$Var\left\{\log \hat{S}(t)\right\} \approx \sum_{j=1}^{k} \frac{d_j}{n_j\left(n_j - d_j\right)},$$

Εφαρμόζοντας και πάλι το παραπάνω γενικό αποτέλεσμα, λαμβάνουμε,

$$Var\left\{\log \hat{S}(t)\right\} \approx \frac{1}{\left[\hat{S}(t)\right]^2} Var\left\{\hat{S}(t)\right\} \Leftrightarrow Var\left\{\hat{S}(t)\right\} \approx \left[\hat{S}(t)\right]^2 \sum_{j=1}^{k} \frac{d_j}{n_j\left(n_j - d_j\right)}. \tag{5.4}$$

Τέλος, το τυπικό σφάλμα του εκτιμητή *Kaplan-Meier* της συνάρτησης επιβίωσης, όπως ορίζεται από την τετραγωνική ρίζα της εκτιμώμενης διακύμανσης του εκτιμητή, δίνεται από τη σχέση,

$$s.e.\left\{\hat{S}(t)\right\} \approx \left[\hat{S}(t)\right]\left\{\sum_{j=1}^{k} \frac{d_j}{n_j\left(n_j - d_j\right)}\right\}^{\frac{1}{2}},$$

για $t_{(k)} \le t \le t_{(k+1)}$. Το αποτέλεσμα αυτό είναι γνωστό ως *τύπος του Greenwood*. Μία συνέπεια του τύπου αυτού είναι το γεγονός ότι σε ένα διάστημα χωρίς θανάτους, όχι μόνο η εκτιμώμενη πιθανότητα θανάτου είναι μηδενική, αλλά και η εκτιμώμενη διακύμανση είναι επίσης μηδέν, σημαίνοντας ότι είμαστε απολύτως σίγουροι πως σε αυτό το διάστημα δεν συμβαίνουν θάνατοι. Αυτό το αποτέλεσμα είναι φυσικά αντίθετο με την κοινή λογική, και πρέπει να αντιμετωπισθεί ως μειονέκτημα της χρησιμοποιούμενης μεθόδου (Hougaard, 2001).

5.4 Εκτιμητής Nelson-Aalen της συνάρτησης επιβίωσης

Αυτή είναι μία εκτίμηση της ολοκληρωμένης συνάρτησης κινδύνου. Η προέλευσή του είναι όμοια με αυτή του *Kaplan-Meier*. Ανεπίσημα μπορούμε να θεωρήσουμε ότι αποτελεί

μία οριακή εκτίμηση του πιθανοτικού μοντέλου $h(t)=\lim_{\Delta t\downarrow 0}\frac{P(t\le T\le t+\Delta t\mid T\ge t)}{\Delta t}, t\ge 0.$ Η πιθανότητα παρατήρησης θανάτου εκτιμάται ως μηδέν, όταν δεν υπάρχουν θάνατοι, και ως *1/n(t)*, εάν σημειώνεται θάνατος την στιγμή *t*. Οπότε καταλήγουμε σε ένα μοντέλο Poisson, έτσι ώστε σε κάθε απειροελάχιστο διάστημα (*j*), και για κάποιο άτομο (*i*) σε κίνδυνο, υπάρχει μία μετρίσιμη μεταβλητή, d_{ij}, η οποία ακολουθεί την κατανομή Poisson με παράμετρο λ_j, έτσι ώστε συγκεκριμένα οι πιθανότητες για 0 ή 1 συμβάν να είναι

$$P(d_{ij}=1)=\lambda_j \exp(-\lambda_j),\ \ P(d_{ij}=0)=\exp(-\lambda_j),$$

και, επομένως, υπάρχει μια μικρή διαφοροποίηση από τον εκτιμητή *Kaplan-Meier*. Επιπλέον, αυτό το μοντέλο επιτρέπει να ισχύει $d_{ij}>1$, παρ' όλο που είναι αδύνατον κάποιος να πεθάνει περισσότερες από μία φορές. Όταν συνδυάζονται περισσότερα του ενός άτομα, κάτι τέτοιο δεν είναι πλέον αδύνατο να συμβεί, αν και οι δεσμοί (πολλαπλά συμβάντα την ίδια χρονική στιγμή) συμβαίνουν με πιθανότητα 0, καθώς η κατανομή που έχουμε υποθέσει είναι συνεχής. Η συνολική πιθανοφάνεια γίνεται,

$$\prod_i \prod_j \left\{\lambda_j^{d_{ij}}\ exp(-\lambda_j)\right\}^{n_{ij}}.$$

Αντιστρέφοντας την τάξη των γινομένων, προκύπτει $\hat{\lambda}_j = d_{.j}/n_{.j}$, χρησιμοποιώντας την σχέση $d_{ij}=\hat{\lambda}_j \cdot n_{ij}$. Επομένως, είναι σαφές ότι υπάρχουν συνεισφορές μόνο κατά τους χρόνους αποβίωσης, και συνεπώς η ολοκληρωμένη συνάρτηση κινδύνου εκτιμάται ως

$$\hat{H}_{NA}(t)=\sum_{r=1}^{q} N_r / n(t_r),$$

όπου ο συμβολισμός μεταβάλλεται με το N να εκτιμάται την στιγμή που πραγματοποιείται το γεγονός, όπου το *d* μπορεί να ορισθεί οποιαδήποτε στιγμή. Μόνο οι χρόνοι θανάτου $t_1,.., t_q$ πριν τον χρόνο *t* και συμπεριλαμβανομένου του χρόνου αυτού συνεισφέρουν στην εκτίμηση. Στην περίπτωση που υπάρχουν δεσμοί, η εκτίμηση είναι ελαφρώς μικρότερη της τιμής που θα προέκυπτε εάν οι τιμές λαμβάνονταν αμέσως η μία μετά την άλλη. Αυτό μπορεί να επιβεβαιωθεί από τα αποτελέσματα στην περίπτωση που υπάρχουν δύο ταυτόχρονα γεγονότα. Η προσέγγιση *Nelson-Aalen* δίνει μία συνεισφορά της τάξης του *2/n(t)*, ενώ αν ένα δεύτερο γεγονός συμβεί αμέσως μετά τον χρόνο *t*, η συνολική συνεισφορά θα ήταν *1/n(t) + 1/{n(t)-1}*. Η μορφή της συνάρτησης επιβίωσης, οπότε, ορίζεται ως,

$$\hat{S}_{NA}(t) = exp\left[-\sum_{r=1}^{q} N_r / n(t_r) \right].$$

Για την διακύμανση των εκτιμητών $\hat{H}_{NA}(t)$ και $\hat{S}_{NA}(t)$ έχουμε τις εκτιμήσεις:

Greenwood (1926): $Var\left[\hat{H}_{NA}(t)\right] = \sum_{r=1}^{q} \frac{N_r}{n(t_r)(n(t_r) - N_r)}$ και

$$Var\left[\hat{S}_{NA}(t)\right] = \left[\hat{S}_{NA}(t)\right]^2 \sum_{r=1}^{q} \frac{N_r}{n(t_r)(n(t_r) - N_r)},$$

Tsiatis (1978): $Var\left[\hat{H}_{NA}(t)\right] = \sum_{r=1}^{q} \frac{N_r}{n(t_r)^2}$ και

$$Var\left[\hat{S}_{NA}(t)\right] = \left[\hat{S}_{NA}(t)\right]^2 \sum_{r=1}^{q} \frac{N_r}{n(t_r)^2},$$

Klein (1997): $Var\left[\hat{H}_{NA}(t)\right] = \sum_{r=1}^{q} \frac{N_r (n(t_r) - N_r)}{n(t_r)^3}$ και

$$Var\left[\hat{S}_{NA}(t)\right] = \left[\hat{S}_{NA}(t)\right]^2 \sum_{r=1}^{q} \frac{N_r (n(t_r) - N_r)}{n(t_r)^3}.$$

Στην πράξη για την εκτίμηση της ποσότητας $Var\left[\hat{H}_{NA}(t)\right]$ χρησιμοποιείται ο τύπος του Tsiatis ενώ για την εκτίμηση της ποσότητας $Var\left[\hat{S}_{NA}(t)\right]$ ο τύπος του Greenwood.

Ο εκτιμώμενος ολοκληρωμένος κίνδυνος είναι ελαφρώς χαμηλότερος του *–logS(t)* που εκτιμάται σύμφωνα με τον *Kaplan-Meier* εκτιμητή. Η διαφορά μπορεί να δικαιολογηθεί από την μικρή διαφοροποίηση των δύο εκτιμητών. Ο *Nelson-Aalen* εκτιμητής της συνάρτησης *H(t)* χαρακτηρίζεται από καλύτερη συμπεριφορά και ιδιότητες από τον αντίστοιχο *Kaplan-Meier* εκτιμητή όταν το μέγεθος του δείγματος είναι μικρό (βλ. Αντζουλάκος, 2009).

5.5 Διαστήματα εμπιστοσύνης για την συνάρτηση επιβίωσης

Αφού υπολογισθεί το τυπικό σφάλμα της $\hat{S}(t)$ είναι δυνατή η εύρεση αντίστοιχων διαστημάτων εμπιστοσύνης. Ένα διάστημα εμπιστοσύνης είναι η διαστηματική εκτίμηση της συνάρτησης επιβίωσης και είναι το διάστημα στο οποίο βρίσκεται μία ορισμένη πιθανότητα ότι η τιμή της πραγματικής συνάρτησης επιβίωσης περιλαμβάνεται σε αυτό.

Ένα διάστημα εμπιστοσύνης για την πραγματική τιμή της συνάρτησης επιβίωσης στον χρόνο *t* εκτιμάται υποθέτοντας ότι η εκτιμώμενη τιμή της συνάρτησης επιβίωσης στον χρόνο *t* κατανέμεται κανονικά με μέσο *S(t)* και εκτιμώμενη διακύμανση όπως ορίζεται από την σχέση (5.4). Έτσι, το 100(1-α)% διάστημα εμπιστοσύνης για την *S(t)*, για μία δεδομένη τιμή *t*, είναι το διάστημα

$$\hat{S}(t) \pm z_{a/2} s.e.\{\hat{S}(t)\},$$

όπου $z_{\alpha/2}$ αντιστοιχεί στο άνω (μονόπλευρο) ποσοστιαίο σημείο της τυπικής κανονικής κατανομής (ισχύει: $P(Z>z_{\alpha/2}) = \alpha/2$). Το διάστημα αυτό είναι γνωστό ως *διάστημα εμπιστοσύνης τύπου plain* και είναι συμμετρικό γύρω από την $\hat{S}(t)$.

Μία δυσκολία της παραπάνω διαδικασίας προκύπτει από το γεγονός ότι τα διαστήματα εμπιστοσύνης είναι συμμετρικά. Όταν η εκτιμώμενη συνάρτηση επιβίωσης είναι κοντά στο μηδέν ή την μονάδα, τα συμμετρικά διαστήματα είναι ακατάλληλα, αφού μπορεί να οδηγήσουν σε όρια εμπιστοσύνης για την συνάρτηση επιβίωσης τα οποία βρίσκονται εκτός του διαστήματος [0, 1]. Μία λύση του προβλήματος αυτού είναι η αντικατάσταση των ορίων που είναι μεγαλύτερα της μονάδας με 1 και αυτών που είναι μικρότερα του μηδενός με 0.

Μία εναλλακτική διαδικασία είναι ο μετασχηματισμός της $\hat{S}(t)$ σε μία τιμή που να ανήκει στο $(-\infty, \infty)$, οπότε λαμβάνουμε διάστημα για την μετασχηματισμένη τιμή. Τα τελικά όρια εμπιστοσύνης μετασχηματίζονται αντίστροφα για να δώσουν το διάστημα εμπιστοσύνης της *S(t)*. Πιθανοί μετασχηματισμοί είναι ο λογιστικός, *log{-log S(t)}*, ο οποίος αντιστοιχεί στον λογάριθμο της αθροιστικής συνάρτησης κινδύνου. Σε κάθε περίπτωση, το τυπικό σφάλμα της μετασχηματισμένης τιμής της $\hat{S}(t)$ μπορεί να βρεθεί χρησιμοποιώντας τη σχέση (5.3). Το διάστημα αυτό αναφέρεται ως *διάστημα εμπιστοσύνης τύπου log-log*, τα άκρα του έχουν τιμή στο [0, 1], και έχει την μορφή

$$\left[\hat{S}(t)\right]^{exp\left(\pm z_{a/2} s.e.\{\hat{S}(t)\}\right)}.$$

Τέλος, συνηθίζεται η κατασκευή διαστημάτων εμπιστοσύνης για τον λογάριθμο του χρόνου, και ένα αντίστοιχο διάστημα για τον λογάριθμο της συνάρτησης επιβίωσης καλείται *διάστημα εμπιστοσύνης τύπου log* και ορίζεται ως

$$\left[\hat{S}(t)\right] \cdot exp\left(\pm z_{a/2} s.e.\{\hat{S}(t)\}\right).$$

Ένα επιπλέον πρόβλημα είναι ότι στις ουρές της κατανομής των χρόνων επιβίωσης, στις οποίες η $\hat{S}(t)$ λαμβάνει τιμές κοντά στο μηδέν ή τη μονάδα, η διακύμανση της $\hat{S}(t)$ που

λαμβάνουμε από τον τύπο του Greenwood υποεκτιμά την πραγματική διακύμανση. Υπό αυτές τις συνθήκες, μία εναλλακτική έκφραση του τυπικού σφάλματος της $\hat{S}(t)$ μπορεί να χρησιμοποιηθεί. Οι Peto *et al.* (1977) πρότειναν ότι το τυπικό σφάλμα της $\hat{S}(t)$ πρέπει να λαμβάνεται από την σχέση,

$$s.e.\{\hat{S}(t)\} = \frac{\hat{S}(t)\sqrt{\{1-\hat{S}(t)\}}}{\sqrt{n_k}},$$

για $t_{(k)} \le t < t_{(k+1)}$, $k = 1, 2,.., r$, όπου $\hat{S}(t)$ είναι η εκτίμηση *Kaplan-Meier* της *S(t)* και n_k είναι το πλήθος τον ατόμων σε κίνδυνο την χρονική στιγμή $t_{(k)}$, δηλαδή στην αρχή του $k^{ου}$ χρονικού διαστήματος.

Τα διαστήματα τύπου *log-log* και τύπου *log* μπορούν να εφαρμοσθούν και στην περίπτωση του εκτιμητή *Nelson-Aalen*. Το διάστημα τύπου *plain* δεν χρησιμοποιείται συχνά στη πράξη.

5.6 Εκτίμηση της συνάρτησης κινδύνου

Ένα απλό δείγμα δεδομένων επιβίωσης μπορεί, επίσης, να εξετασθεί μέσω της συνάρτησης κινδύνου, η οποία εκφράζει την εξέλιξη του στιγμιαίου κινδύνου θανάτου στον χρόνο. Ένας φυσικός τρόπος εκτίμησης της συνάρτησης κινδύνου για μη ομαδοποιημένα δεδομένα είναι να θεωρήσουμε τον λόγο του πλήθους των θανάτων σε ένα δεδομένο χρόνο θανάτου προς το πλήθος των ατόμων σε κίνδυνο την ίδια χρονική στιγμή. Εάν η συνάρτηση κινδύνου υποθέτουμε να είναι σταθερή μεταξύ διαδοχικών χρόνων θανάτου ο κίνδυνος ανά χρονική στιγμή μπορεί να βρεθεί διαιρώντας επιπλέον με το χρονικό διάστημα. Επομένως, εάν υπάρχουν d_j θάνατοι στον j^o χρόνο θανάτου $t_{(j)}$, j=1, 2,…, r, και n_j άτομα σε κίνδυνο στον χρόνο $t_{(j)}$, η συνάρτηση κινδύνου στο διάστημα από $t_{(j)}$ έως $t_{(j+1)}$ εκτιμάται από την σχέση

$$\hat{h}(t) = \frac{d_j}{n_j \tau_j},$$

για $t_{(j)} \le t < t_{(j+1)}$, όπου $\tau_j = t_{(j+1)} - t_{(j)}$. Αξίζει να σημειώσουμε ότι δεν είναι δυνατή η χρήση της παραπάνω σχέσης για την εκτίμηση του κινδύνου σε διάστημα το οποίο ξεκινά με τον τελικό χρόνο κινδύνου, αφού το διάστημα είναι ανοιχτό από τα δεξιά.

5.7 Εκτίμηση της αθροιστικής συνάρτησης κινδύνου

Η αθροιστική συνάρτηση κινδύνου είναι σημαντική για την ταυτοποίηση των μοντέλων σε δεδομένα επιβίωση. Επομένως, είναι σημαντικό να δούμε μεθόδους εκτίμησης αυτής. Ο αθροιστικός κίνδυνος στο χρόνο *t*, *H(t)*, ορίστηκε παραπάνω να είναι το ολοκλήρωμα της συνάρτησης κινδύνου. Ωστόσο, είναι πιο βολικό να υπολογίζεται από την σχέση $H(t) = -\log S(t)$, και επομένως, εάν $\hat{S}(t)$ είναι ο *Kaplan-Meier* εκτιμητής της συνάρτησης επιβίωσης, $\hat{H}(t) = -\log \hat{S}(t)$ είναι ο κατάλληλος εκτιμητής του αθροιστικού κινδύνου στο χρόνο *t*.

Κάνοντας χρήση της σχέσης (5.2), έχουμε

$$\hat{H}(t) = -\sum_{j=1}^{k} \log\left(\frac{n_j - d_j}{n_j}\right).$$

Επιπλέον, δεδομένης της σειράς *log(1-x)* η οποία γίνεται $-x-x^2/2+\ldots$, έχουμε

$$\log\left(\frac{n_j - d_j}{n_j}\right) = \log\left(1 - \frac{d_j}{n_j}\right) \approx -\frac{d_j}{n_j},$$

αγνοώντας τους όρους δεύτερης τάξης και άνω. Συνεπώς, προκύπτει η σχέση

$$\hat{H}(t) \approx \sum_{j=1}^{k} \frac{d_j}{n_j},$$

η οποία είναι το άθροισμα των εκτιμώμενων πιθανοτήτων θανάτου από το πρώτο έως το $k^{\text{ο}}$ διάστημα χρόνου, για *j*=1, 2,..., *k*. Η ποσότητα, αυτή, συνεπώς, έχει άμεση διαισθητική επίκληση ως εκτιμητής αθροιστικού κινδύνου και δηλώνει την πιθανότητα ότι το γεγονός έχει πραγματοποιηθεί έως τον χρόνο *t*, ή ισοδύναμα, την πιθανότητα ότι ο χρόνος επιβίωσης είναι μικρότερος ή ίσος του *t*. Η αθροιστική συνάρτηση κινδύνου και η συνάρτηση επιβίωσης συνδέονται με την σχέση *H(t)=1–S(t)*. Συνεπώς, η αθροιστική συνάρτηση επίπτωσης εκτιμάται ως το συμπλήρωμα του εκτιμητή *Kaplan-Meier* της συνάρτησης επιβίωσης, δηλαδή $\hat{H}(t) = 1 - \hat{S}(t)$.

5.8 Εκτίμηση της διαμέσου και των ποσοστημορίων των χρόνων επιβίωσης

Καθώς η κατανομή των χρόνων επιβίωσης τείνει να είναι θετικά λοξή, η διάμεσος είναι το επιθυμητό μέτρο που συνοψίζει την τοποθεσία της κατανομής. Όταν εκτιμάται η συνάρτηση επιβίωσης, στην συνέχεια μπορεί να εκτιμηθεί και ο *διάμεσος χρόνος επιβίωσης* (median survival time). Η εκτίμηση της διαμέσου αντιστοιχεί στο χρονικό σημείο πέραν του οποίου το 50% των ατόμων του πληθυσμού υπό εξέταση αναμένεται να επιζήσουν, και δίνεται από την τιμή *t(50),* η οποία ορίζεται έτσι ώστε $S\{t(50)\} = 0{,}5$.

Δεδομένου ότι η μη παραμετρική εκτίμηση της *S(t)* είναι κλιμακωτή, δεν είναι συνήθως δυνατό να αντιληφθούμε μία εκτίμηση του χρόνου επιβίωσης που να κάνει την συνάρτηση επιβίωσης ακριβώς ίση με 0,5. Αντί αυτού, ο εκτιμώμενος διάμεσος χρόνος επιβίωσης, $\hat{t}(50)$, ορίζεται να είναι ο ελάχιστος παρατηρούμενος χρόνος επιβίωσης για τον οποίο η τιμή της εκτιμώμενης συνάρτησης επιβίωσης είναι μικρότερη του 0,5.

Με μαθηματικούς όρους,

$$\hat{t}(50) = \min\{t_i \mid \hat{S}(t_i) \le 0{,}5\},$$

όπου t_i είναι ο παρατηρούμενος χρόνος επιβίωσης για το $i^{\text{ο}}$ άτομο, i=1, 2,..., n. Εφόσον η εκτιμώμενη συνάρτηση επιβίωσης μεταβάλλεται μόνο στον χρόνο θανάτου, η διάμεσος είναι ισοδύναμη με τον ορισμό

$$\hat{t}(50) = \min\{t_{(j)} \mid \hat{S}(t_{(j)}) \le 0{,}5\},$$

όπου $t_{(j)}$ είναι ο $j^{\text{ος}}$ διατεταγμένος χρόνος θανάτου, j=1, 2, .., r.

Σε κάποιες περιπτώσεις συμβαίνει η εκτιμώμενη συνάρτηση επιβίωσης να είναι μεγαλύτερη του 0,5 για κάθε τιμή του t. Σε αυτές τις περιπτώσεις, ο διάμεσος χρόνος επιβίωσης δεν μπορεί να εκτιμηθεί, και είναι φυσικό να συνοψίζουμε τα δεδομένα σε όρους των εκτιμώμενων πιθανοτήτων επιβίωσης σε συγκεκριμένα χρονικά σημεία.

Μία παρόμοια διαδικασία με αυτή που περιγράφηκε παραπάνω μπορεί να εφαρμοσθεί για τον υπολογισμό άλλων *ποσοστιαίων σημείων* (percentiles) της κατανομής των χρόνων επιβίωσης. Το $p^{\text{ο}}$ ποσοστημόριο της κατανομής ορίζεται ως η τιμή *t(p)* η οποία είναι τέτοια ώστε $F\{t(p)\} = p/100$. Σε όρους της συνάρτησης επιβίωσης, το *t(p)* ορίζεται έτσι ώστε να ισχύει $S\{t(p)\} = 1-(p/100)$, και δηλώνει τον μικρότερο χρόνο στον οποίο η καμπύλη *Kaplan-Meier* 'πέφτει' κάτω του *1-p*. Έναν περιορισμό του εκτιμητή Kaplan-Meier αποτελεί

το γεγονός ότι εάν η καμπύλη δεν φτάνει το *1-p*, το p^o ποσοστημόριο δεν μπορεί να εκτιμηθεί.

Εκτιμήσεις της διασποράς ενός δείγματος δεδομένων επιβίωσης δεν χρησιμοποιούνται ευρέως, όμως όταν απαιτείται μία τέτοια εκτίμηση, μπορεί να υπολογισθεί το *ημι-ενδοτεταρτημοριακό εύρος* (semi-interquartile range - SIQR). Αυτό ορίζεται ως το μισό της διαφοράς του $75^{ου}$ και $25^{ου}$ τεταρτημορίου της κατανομής των χρόνων επιβίωσης. Δηλαδή,

$$SIQR = \frac{1}{2}\{t(75) - t(25)\},$$

όπου *t(25)* και *t(75)* είναι το 25^o και 75^o ποσοστημόριο της κατανομής του χρόνου. Τα δύο παραπάνω ποσοστημόρια είναι γνωστά ως το πρώτο και το τρίτο τεταρτημόριο, αντίστοιχα. Όπως στην περίπτωση της διακύμανσης, όσο μεγαλύτερη η τιμή του SIQR, τόσο πιο διεσπαρμένη είναι η κατανομή.

Αξίζει να σημειωθεί ότι ενώ μπορούμε να εκτιμήσουμε την διάμεσο και άλλα ποσοστημόρια της κατανομής των χρόνων επιβίωσης χρησιμοποιώντας τα αποτελέσματα *Kaplan-Meier*, δεν είναι δυνατόν να εκτιμήσουμε τη μέση τιμή της κατανομής στην τυπική περίπτωση κατά την οποία ο μέγιστος follow-up χρόνος είναι λογοκριμένος.

5.8.1 Διαστήματα εμπιστοσύνης για την διάμεσο και τα ποσοστημόρια

Προσεγγιστικά διαστήματα εμπιστοσύνης για την διάμεσο και άλλα ποσοστημόρια μιας κατανομής χρόνων επιβίωσης μπορούν να βρεθούν όταν είναι γνωστή η διακύμανση των εκτιμώμενων ποσοστημορίων. Μία έκφραση της προσεγγιστικής διακύμανσης ενός ποσοστημορίου μπορεί να προέλθει από άμεση εφαρμογή του γενικού αποτελέσματος της σχέσης (5.3) όπως αυτό ορίστηκε παραπάνω, και προκύπτει

$$Var\left[\hat{S}\{t(p)\}\right] = \left(\frac{d\hat{S}\{t(p)\}}{dt(p)}\right)^2 Var\{t(p)\},$$

όπου *t(p)* είναι το p^o ποσοστημόριο της κατανομής και $\hat{S}\{t(p)\}$ είναι ο *Kaplan-Meier* εκτιμητής της συνάρτησης επιβίωσης στο *t(p)*. Επίσης,

$$\frac{d\hat{S}\{t(p)\}}{dt(p)} = -\hat{f}\{t(p)\},$$

είναι μία εκτίμηση της συνάρτηση πυκνότητας πιθανότητας των χρόνων επιβίωσης στο *t(p)*, οπότε σύμφωνα με τις δύο τελευταίες σχέσεις, έχουμε,

$$Var\{t(p)\} = \left(\frac{1}{\hat{f}\{t(p)\}} \right)^2 Var\left[\hat{S}\{t(p)\}\right].$$

Το τυπικό σφάλμα του $\hat{t}(p)$, για το p^o εκτιμώμενο ποσοστημόριο, δίνεται, επομένως, από την σχέση,

$$s.e.\{\hat{t}(p)\} = \frac{1}{\hat{f}\{\hat{t}(p)\}} s.e.\left[\hat{S}\{\hat{t}(p)\}\right].$$

Το τυπικό σφάλμα της $\hat{S}\{\hat{t}(p)\}$ δίνεται από τον τύπο του Greenwood για το τυπικό σφάλμα της εκτίμησης *Kaplan-Meier* της συνάρτησης επιβίωσης, ενώ η εκτίμηση της συνάρτησης πυκνότητας πιθανότητας στο $\hat{t}(p)$ είναι

$$\hat{f}\{\hat{t}(p)\} = \frac{\hat{S}\{\hat{u}(p)\} - \hat{S}\{\hat{l}(p)\}}{\hat{l}(p) - \hat{u}(p)},$$

όπου

$$\hat{u}(p) = \max\left\{ t_{(j)} \mid \hat{S}(t_{(j)}) \geq 1 - \frac{p}{100} + \varepsilon \right\},$$

και

$$\hat{l}(p) = \min\left\{ t_{(j)} \mid \hat{S}(t_{(j)}) \leq 1 - \frac{p}{100} - \varepsilon \right\},$$

για j=1, 2,..., r, και μικρές τιμές του ε. Σε πολλές περιπτώσεις η τιμή ε = 0,05 είναι ικανοποιητική, αλλά μια μεγαλύτερη τιμή του ε απαιτείται όταν τα $\hat{u}(p)$ και $\hat{l}(p)$ είναι ίσα. Συγκεκριμένα, για την περίπτωση της διαμέσου, το τυπικό σφάλμα είναι

$$s.e.\{\hat{t}(50)\} = \frac{1}{\hat{f}\{\hat{t}(50)\}} s.e.\left[\hat{S}\{\hat{t}(50)\}\right],$$

όπου η $\hat{f}\{\hat{t}(50)\}$ μπορεί να βρεθεί ως

$$\hat{f}\{\hat{t}(50)\} = \frac{\hat{S}\{\hat{u}(50)\} - \hat{S}\{\hat{l}(50)\}}{\hat{l}(50) - \hat{u}(50)}.$$

Σε αυτή την έκφραση, το $\hat{u}(50)$ είναι ο μεγαλύτερος χρόνος επιβίωσης για τον οποίο η εκτίμηση *Kaplan-Meier* της συνάρτησης επιβίωσης ξεπερνά την τιμή 0,55, και το $\hat{l}(50)$ είναι ο μικρότερος χρόνος επιβίωσης για τον οποίο η συνάρτηση επιβίωσης είναι μικρότερη ή ίση του 0,45.

Όταν βρεθεί η εκτίμηση του $p^{ου}$ ποσοστημορίου, ένα 100(1-α)% διάστημα εμπιστοσύνης για το *t(p)* έχει όρια

$$\hat{t}(p) \pm z_{\alpha/2}\, s.e.\{\hat{t}(p)\},$$

όπου $z_{\alpha/2}$ είναι το άνω (μονόπλευρο) α/2-σημείο της τυπικής κανονικής κατανομής.

Αυτή η εκτίμηση διαστήματος είναι μόνο προσεγγιστική, με την έννοια ότι η πιθανότητα το διάστημα να περιέχει το πραγματικό ποσοστημόριο δεν είναι ακριβώς *1-α*. Ένα πλήθος μεθόδων έχει προταθεί για την κατασκευή διαστήματος εμπιστοσύνης με τις παραπάνω ιδιότητες, παρ' όλο που είναι υπολογιστικά πιο δύσκολες σε σχέση με την μέθοδο που παρουσιάστηκε παραπάνω.

5.9 Σύγκριση δύο ομάδων σε δεδομένα επιβίωσης

Ο απλούστερος τρόπος σύγκρισης των εκτιμώμενων χρόνων επιβίωσης δύο ομάδων ατόμων είναι η γραφική παρουσίαση των αντίστοιχων εκτιμήσεων των δύο συναρτήσεων επιβίωσης σε κοινούς άξονες. Υπάρχουν δύο δυνατές ερμηνείες για την παρατηρούμενη διαφορά μεταξύ των χρόνων επιβίωσης δύο εκτιμώμενων συναρτήσεων επιβίωσης. Η μία ερμηνεία είναι ότι υπάρχει μία πραγματική διαφορά μεταξύ των χρόνων επιβίωσης των δύο ομάδων ατόμων. Μία εναλλακτική ερμηνεία είναι ότι δεν υπάρχουν πραγματικές διαφορές μεταξύ των χρόνων επιβίωσης σε κάθε ομάδα, και ότι η διαφορά που παρατηρείται είναι απλώς το αποτέλεσμα της τυχαίας διακύμανσης. Για να διαχωρίσουμε τις δύο αυτές πιθανές ερμηνείες καταφεύγουμε στον κατάλληλο *έλεγχος υπόθεσης*.

Ο *έλεγχος υπόθεσης* είναι μία διαδικασία η οποία μας καθιστά ικανούς να εκτιμήσουμε την έκταση στην οποία ένα παρατηρούμενο σύνολο δεδομένων είναι συμβατό με μία συγκεκριμένη υπόθεση, γνωστή ως *μηδενική υπόθεση* (*null hypothesis*) ή όχι, οπότε απορρίπτεται η μηδενική υπόθεση για χάρη της *εναλλακτικής υπόθεσης* (*alternative hypothesis*). Η εναλλακτική υπόθεση μπορεί να είναι δίπλευρη ή μονόπλευρη. Αφού προσδιοριστεί το ζεύγος υποθέσεων που επιθυμούμε να ελέγξουμε, διατυπώνεται η στατιστική συνάρτηση που μετρά την έκταση στην οποία τα παρατηρούμενα δεδομένα αποκλίνουν από την μηδενική υπόθεση. Στη συνέχεια, υπολογίζεται η πιθανότητα να εκτιμήσουμε κάτω από τη μηδενική υπόθεση μία τιμή τόσο ακραία ή και ακόμα περισσότερο από την παρατηρούμενη τιμή, προς την κατεύθυνση της εναλλακτικής, δηλαδή την *probability value* ή *p-value* για συντομία. Η *p-value* συνοψίζει την ισχύ των αποδεικτικών

στοιχείων στο δείγμα κατά της μηδενικής υπόθεσης. Παραδοσιακά, *p-value* με τιμές 0,05 ή 0,01 χρησιμοποιούνται προκειμένου να καταλήξουμε σε μία απόφαση σχετικά με το αν η μηδενική υπόθεση πρέπει να απορριφθεί ή όχι.

Η στατιστική πληροφορία που συνοψίζεται στην *p-value* για έναν έλεγχο υπόθεσης αποτελεί μόνο ένα μέρος της διαδικασίας λήψης απόφασης. Επιπρόσθετα των στατιστικών στοιχείων, υπάρχουν επίσης και επιστημονικά στοιχεία που πρέπει να αξιολογούμε κάθε φορά και να λαμβάνουμε υπόψη.

5.9.1 Έλεγχος log-rank

Ο εκτιμητής *Kaplan-Meier* παρέχει μια ερμηνεύσιμη περιγραφή της επιβίωσης δύο ομάδων που διαφοροποιούνται ως προς κάποιο χαρακτηριστικό (π.χ. διαφορετική φαρμακευτική αγωγή ή ομάδα ηλικιών). Το βασικό εργαλείο σύγκρισης της επιβίωσης δύο ή περισσότερων ομάδων είναι το *log-rank test*. Προκειμένου να κατασκευάσουμε τον έλεγχο, αρχικά θεωρούμε χωριστά κάθε χρόνο θανάτου στις δύο ομάδες των χρόνων επιβίωσης. Οι δύο αυτές ομάδες συμβολίζονται ως Ομάδα I και Ομάδα II. Υποθέτουμε ότι υπάρχουν r διακριτοί χρόνου θανάτου, $t_{(1)}<t_{(2)}<\ldots<t_{(r)}$, για τις δύο ομάδες (*pooled sample*), και ότι στον χρόνο $t_{(j)}$, d_{1j} άτομα στην Ομάδα I και d_{2j} άτομα στην Ομάδα II πεθαίνουν, για j=1, 2,…, r. Εάν δύο ή περισσότερα άτομα σε μία ομάδα δεν σημειώνουν τους ίδιους χρόνους θανάτου, οι τιμές d_{1j} και d_{2j} θα είναι είτε μηδέν είτε μονάδα. Υποθέτουμε, επίσης, ότι υπάρχουν n_{1j} άτομα σε κίνδυνο θανάτου στην πρώτη ομάδα ακριβώς πριν τον χρόνο $t_{(j)}$, και n_{2j} άτομα στην δεύτερη ομάδα. Επομένως, στον χρόνο $t_{(j)}$ υπάρχουν $d_j=d_{1j}+d_{2j}$ θάνατοι συνολικά από τα $n_j=n_{1j}+n_{2j}$ άτομα σε κίνδυνο. Τα παραπάνω συνοψίζονται στον πίνακα 5.1.

Ομάδα	**Πλήθος θανάτων στο $t_{(j)}$**	**Πλήθος επιζώντων πέραν του $t_{(j)}$**	**Πλήθος σε κίνδυνο πριν το $t_{(j)}$**
I	d_{1j}	$n_{1j}-d_{1j}$	n_{1j}
II	d_{2j}	$n_{2j}-d_{2j}$	n_{2j}
Σύνολο	d_j	n_j-d_j	n_j

Πίνακας 5.1: Η δομή των δεδομένων.

Στην συνέχεια θεωρούμε την μηδενική υπόθεση ότι δεν υπάρχουν διαφορές στις συναρτήσεις επιβίωσης των ατόμων στις δύο ομάδες. Ένας τρόπος να εκτιμήσουμε την ισχύ

αυτής της υπόθεσης είναι να θεωρήσουμε την έκταση της διαφοράς μεταξύ των παρατηρούμενων τιμών των ατόμων στις δύο ομάδες που πεθαίνουν σε κάθε χρονική στιγμή θανάτου, και των αναμενόμενων τιμών υπό την μηδενική υπόθεση. Η πληροφορία σχετικά με την έκταση αυτών των πληροφοριών μπορούν στη συνέχεια να συνδυαστούν για κάθε χρόνο θανάτου.

Εάν τα περιθώρια αθροίσματα του πίνακα 5.1 θεωρηθούν σταθερά, και η μηδενική υπόθεση ότι η επιβίωση είναι ανεξάρτητη της ομάδας είναι αληθής, οι τέσσερεις τιμές που εισάγουμε σε αυτόν τον πίνακα καθορίζονται αποκλειστικά από την τιμή του d_{1j}. Μπορούμε, επομένως, να θεωρήσουμε το d_{1j} ως μία τυχαία μεταβλητή, η οποία μπορεί να λάβει οποιαδήποτε τιμή στο εύρος από 0 έως d_j και n_{1j}. Στην πραγματικότητα η d_{1j} έχει μία κατανομή γνωστή ως *υπεργεωμετρική κατανομή* (hypergeometric distribution), σύμφωνα με την οποία η πιθανότητα ότι η τυχαία μεταβλητή που σχετίζεται με το πλήθος των θανάτων στην πρώτη ομάδα λαμβάνει την τιμή d_{1j} είναι

$$\frac{\binom{d_j}{d_{1j}}\binom{n_j - d_j}{n_{1j} - d_{1j}}}{\binom{n_j}{n_{1j}}}.$$

Ο μέσος της υπεργεωμετρικής τυχαίας μεταβλητής d_{1j} δίνεται από τη σχέση $e_{1j} = n_{1j}\, d_j / n_j$. Επομένως, η τιμή e_{1j} είναι το αναμενόμενο πλήθος ατόμων που πεθαίνουν στον χρόνο $t_{(j)}$ στην Ομάδα Ι. Η τιμή αυτή είναι διαισθητικά άμεση, εφόσον υπό την μηδενική υπόθεση ότι η πιθανότητα θανάτου στον χρόνο $t_{(j)}$ δεν εξαρτάται από την ομάδα στην οποία ανήκει το άτομο, η πιθανότητα θανάτου στο $t_{(j)}$ είναι d_j/n_j. Πολλαπλασιάζοντάς την με n_{1j}, δίνει την e_{1j} ως το αναμενόμενο πλήθος θανάτων της Ομάδας Ι στον χρόνο $t_{(j)}$.

Το επόμενο βήμα είναι να συνδυάσουμε την πληροφορία από τον 2×2 πίνακα που αντιστοιχεί σε κάθε χρόνο θανάτου προκειμένου να έχουμε ένα ολικό μέτρο της απόκλισης των παρατηρούμενων τιμών της d_{1j} από τις αναμενόμενες τιμές τους. Ο αμεσότερος τρόπος για αυτό είναι να αθροίσουμε τις διαφορές d_{1j} - e_{1j} για τον συνολικό αριθμό των χρόνων θανάτου, r, στις δύο ομάδες. Η τιμή αυτή δίνεται από την σχέση

$$U_L = \sum_{j=1}^{r} \left(d_{1j} - e_{1j} \right). \tag{5.5}$$

Παρατηρούμε ότι γίνεται $\sum d_{1j} - \sum e_{1j}$, η οποία είναι η διαφορά των συνολικών παρατηρούμενων και αναμενόμενων τιμών του πλήθους θανάτων στην Ομάδα Ι. Αυτό το

στατιστικό θα έχει μηδενικό μέσο, εφόσον $E(d_{1j}) = e_{1j}$. Επιπλέον, εφόσον οι χρόνοι θανάτου είναι μεταξύ τους ανεξάρτητοι, η διακύμανση του U_L είναι απλώς το άθροισμα των διακυμάνσεων των d_{1j}. Δεδομένου, λοιπόν, ότι το d_{1j} έχει μία υπεργεωμετρική κατανομή, η διακύμανση του d_{1j} δίνεται από την σχέση

$$u_{1j} = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}, \tag{5.6}$$

οπότε η διακύμανση του U_L είναι

$$Var(U_L) = \sum_{j=1}^{r} u_{1j} = V_L .$$

Επιπλέον, μπορεί να δειχθεί ότι η U_L έχει μία προσεγγιστικά κανονική κατανομή, όπου το πλήθος των χρόνων θανάτου δεν είναι πολύ μικρό. Συνεπώς προκύπτει ότι η $U_L / \sqrt{V_L}$ ακολουθεί κανονική κατανομή με μηδενικό μέσο και μοναδιαία διακύμανση, που συμβολίζεται ως N(0,1). Επομένως, γράφουμε

$$\frac{U_L}{\sqrt{V_L}} \sim N(0,1),$$

Το τετράγωνο της τυπικής κανονικής τυχαίας μεταβλητής ακολουθεί την X^2 κατανομή με ένα βαθμό ελευθερίας (συμβολίζεται ως X_1^2), και επομένως έχουμε ότι

$$\frac{U_L^2}{V_L} \sim X_1^2 .$$

Το στατιστικό $W_L = U_L^2 / V_L$ συνοψίζει την έκταση στην οποία οι παρατηρούμενοι χρόνοι επιβίωσης στις δύο ομάδες δεδομένων διαχωρίζονται από τους αναμενόμενους υπό την μηδενική υπόθεση μη διαφοροποίησης των ομάδων. Όσο μεγαλύτερη η τιμή αυτού του στατιστικού, τόσο μεγαλύτερη η ένδειξη κατά της μηδενικής υπόθεσης. Επειδή η κατανομή της *W* υπό την μηδενική υπόθεση είναι προσεγγιστικά X^2 με ένα βαθμό ελευθερίας, η *p-value* μπορεί να εκτιμηθεί από την συνάρτηση κατανομής μίας X^2 τυχαίας μεταβλητής. Εναλλακτικά, ποσοστιαία σημεία της κατανομής X^2 μπορούν να χρησιμοποιηθούν προκειμένου να προσδιορισθεί ένα εύρος τιμών στο οποίο μπορεί να βρίσκεται η *p-value*. Όταν η μηδενική υπόθεση απορρίπτεται, η μελέτη των διαγραμμάτων *Kaplan-Meier* μπορεί να βοηθήσει στον εντοπισμό της σημαντικής αυτής διαφοράς. Σε κάθε περίπτωση, οι μη παραμετρικές προσεγγίσεις, όπως η παρούσα, μειώνουν την ανάγκη περιορισμών και κάποιες φορές υποθέσεων δύσκολων να επαληθευτούν.

5.9.2 Έλεγχος Wilcoxon ή Breslow ή Gehan's

Ο έλεγχος Wilcoxon (ή Breslow ή Gehan's) για την μηδενική υπόθεση ότι δεν υπάρχει διαφορά στις συναρτήσεις επιβίωσης δύο ομάδων των δεδομένων επιβίωσης είναι παρόμοιος με τον έλεγχο log-rank. Ο έλεγχος βασίζεται στο στατιστικό

$$U_W = \sum_{j=1}^{r} n_j \left(d_{1j} - e_{1j}\right),$$

όπου, όπως και στην προηγούμενη ενότητα, d_{1j} είναι το πλήθος των θανάτων στον χρόνο $t_{(j)}$ στην πρώτη ομάδα και e_{1j} είναι το αναμενόμενο πλήθος ατόμων που πεθαίνουν στον χρόνο $t_{(j)}$ στην Ομάδα Ι. Η διαφορά των στατιστικών U_W και U_L είναι ότι στον έλεγχο Wilcoxon, κάθε διαφορά $d_{1j} - e_{1j}$ είναι σταθμισμένη για n_j, το οποίο δηλώνει το συνολικό πλήθος των ατόμων σε κίνδυνο στον χρόνο $t_{(j)}$. Η επίδραση αυτού είναι ότι δίνει μικρότερο βάρος στις διαφορές των d_{1j} και e_{1j} στους χρόνους όταν το συνολικό πλήθος των ατόμων που είναι ακόμα εν ζωή είναι μικρό, που σημαίνει στους μεγαλύτερους χρόνους επιβίωσης. Αυτό το στατιστικό είναι, επομένως, λιγότερο ευαίσθητο από το log-rank σε αποκλίσεις του d_{1j} από το e_{1j} στην ουρά της κατανομής των χρόνων επιβίωσης.

Η διακύμανση του στατιστικού Wilcoxon U_W δίνεται από την σχέση

$$V_W = \sum_{j=1}^{r} n_j^2 u_{1j},$$

όπου u_{1j} δίνεται από την σχέση (5.6), και επομένως, το στατιστικό του ελέγχου γίνεται

$$W_W = U_W^2 / V_W,$$

το οποίο ακολουθεί X^2 κατανομή με ένα βαθμό ελευθερίας όταν ισχύει η μηδενική υπόθεση. Ο έλεγχος Wilcoxon, επομένως, διεξάγεται όπως και το log-rank test.

5.9.3 Σύγκριση των ελέγχων log-rank και Wilcoxon

Από τους δύο παραπάνω ελέγχους, ο έλεγχος log-rank είναι περισσότερο κατάλληλος όταν η εναλλακτική της μηδενικής υπόθεσης για μη ύπαρξη διαφοράς στις συναρτήσεις επιβίωσης δύο ομάδων είναι ότι ο κίνδυνος θανάτου σε κάθε δεδομένη στιγμή για ένα άτομο μιας ομάδας είναι ανάλογος του κινδύνου σε αυτόν τον χρόνο για ένα άτομο με αντίστοιχα κοινά χαρακτηριστικά της άλλης ομάδας. Αυτή είναι η *υπόθεση του αναλογικού κινδύνου*

(proportional hazard), η οποία αποτελεί την βάση μιας σειράς μεθόδων για την ανάλυση δεδομένων επιβίωσης. Για άλλους τύπους απόκλισης από την μηδενική υπόθεση, ο έλεγχος Wilcoxon είναι πιο κατάλληλος από τον έλεγχο log-rank για την σύγκριση δύο συναρτήσεων επιβίωσης.

Προκειμένου να καταλήξουμε σε μία απόφαση σχετικά με το ποιος έλεγχος είναι πιο κατάλληλος για μία δεδομένη κατάσταση, κάνουμε χρήση του αποτελέσματος ότι εάν οι συναρτήσεις κινδύνου είναι ανάλογες, τότε αυτές που αντιστοιχούν στις δύο ομάδες των δεδομένων δεν διασταυρώνονται μεταξύ τους. Για να δείξουμε την παραπάνω πρόταση, υποθέτουμε ότι $h_1(t)$ είναι ο κίνδυνος θανάτου στο χρόνο t για ένα άτομο της Ομάδας I, και $h_2(t)$ είναι ο κίνδυνος θανάτου στο χρόνο t για ένα άτομο της Ομάδας II. Εάν οι δύο αυτοί κίνδυνοι είναι ανάλογοι, τότε ισχύει $h_1(t) = \psi\, h_2(t)$, όπου ψ είναι μία σταθερά η οποία δεν εξαρτάται από τον χρόνο t. Λαμβάνοντας τα ολοκληρώματα των δύο μερών της σχέσης αυτής, πολλαπλασιάζοντας με -1 και λαμβάνοντας το εκθετικό για κάθε μέρος καταλήγουμε στη σχέση

$$\exp\left\{-\int_0^t h_1(t)du\right\} = \exp\left\{-\int_0^t \psi h_2(t)du\right\}.$$

Όπως ορίστηκε παραπάνω η συνάρτηση επιβίωσης, προκύπτει ως

$$S(t) = \exp\left\{-\int_0^t h(u)du\right\},$$

οπότε εάν οι $S_1(t)$ και $S_2(t)$ είναι οι συναρτήσεις επιβίωσης των δύο ομάδων των δεδομένων επιβίωσης, καταλήγουμε στην σχέση

$$S_1(t) = [S_2(t)]^{\psi}.$$

Δεδομένου ότι η συνάρτηση επιβίωσης λαμβάνει τιμές μεταξύ του μηδενός και της μονάδας, το αποτέλεσμα αυτό δείχνει ότι η $S_1(t)$ είναι μεγαλύτερη ή ίση της $S_2(t)$, σύμφωνα με το εάν η ψ είναι μικρότερη ή μεγαλύτερη της μονάδας, στον χρόνο t. Αυτό σημαίνει πως εάν δύο συναρτήσεις επιβίωσης είναι ανάλογες, οι πραγματικές συναρτήσεις επιβίωσης δεν διασταυρώνονται. Αυτή είναι μία αναγκαία, αλλά όχι ικανή συνθήκη για αναλογικούς κινδύνους.

Μία ‘ανεπίσημη’ εκτίμηση για πιθανή ισχύ της υπόθεσης αναλογικού κινδύνου μπορεί να λάβει χώρα από ένα διάγραμμα των εκτιμώμενων συναρτήσεων επιβίωσης για δύο ομάδες των δεδομένων επιβίωσης. Εάν οι δύο εκτιμώμενες συναρτήσεις επιβίωσης δεν διασταυρώνονται, η υπόθεση αναλογικού κινδύνου ίσως ικανοποιείται, και τότε ο έλεγχος

log-rank είναι κατάλληλος. Φυσικά, εκτιμήσεις της συνάρτησης επιβίωσης που στηρίζονται σε δείγμα, μπορεί να διασταυρώνονται ακόμα και αν οι αντίστοιχες πραγματικές συναρτήσεις κινδύνου είναι ανάλογες, οπότε ιδιαίτερη προσοχή απαιτείται στην ερμηνεία τέτοιων διαγραμμάτων.

5.9.4 Άλλοι έλεγχοι για δύο ομάδες

Όπως αναφέρθηκε παραπάνω, ο έλεγχος Wilcoxon προκύπτει θέτοντας κάποιο βάρος στην στατιστική ελέγχου του log-rank test. Ωστόσο έχουν προταθεί κι άλλα βάρη διαμορφώνοντας αντίστοιχους ελέγχους. Κάποια από αυτά είναι τα *Tarone-Ware* $\left(w_j = \sqrt{n_j}\right)$, *Peto-Peto* $\left(w_j = \tilde{S}\left(t_j+\right)\right)$, *Modified Peto-Peto* $\left(w_j = \tilde{S}\left(t_j+\right)\cdot\frac{n_j}{n_j+1}\right)$ και *Flemming-Harrington* $\left(w_j = \left[\hat{S}\left(t_{j-1}+\right)\right]^{\rho}\left[1-\hat{S}\left(t_{j-1}+\right)\right]^{\pi}\right)$, για $\rho,\pi \geq 0$. Η ποσότητα $\tilde{S}\left(t_j\right)$ ορίζεται από τη σχέση $\tilde{S}\left(t_j\right)=\prod_{j:t_j<t}\left(1-\frac{d_j}{n_j+1}\right)$, ενώ η ποσότητα $\hat{S}(t)$ αποτελεί τον εκτιμητή *Kaplan-Meier*. Στο τεστ *Flemming-Harrington* για (ρ,π)=(0,0) προκύπτει ο έλεγχος log-rank, ενώ για (ρ,π)=(1,0) μια προσέγγιση του ελέγχου *Peto-Peto* (βλ. Αντζουλάκος, 2009).

5.10 Σύγκριση τριών ή περισσοτέρων ομάδων

Οι έλεγχοι log-rank και Wilcoxon μπορούν να επεκταθούν έτσι ώστε να επιτρέπουν την σύγκριση τριών ή περισσοτέρων ομάδων των δεδομένων επιβίωσης. Έστω ότι πρόκειται να συγκριθούν οι συναρτήσεις επιβίωσης g ομάδων, για $g \geq 2$. Στην συνέχεια ορίζουμε ανάλογα τα *U*-στατιστικά για την σύγκριση των παρατηρούμενων αριθμών θανάτων στις ομάδες 1, 2, …, *g-1* με τις αναμενόμενες τιμές τους.

Σε αντιστοιχία με την περίπτωση των δύο ομάδων, έχουμε

$$U_{Lk} = \sum_{j=1}^{r}\left(d_{kj} - \frac{n_{kj}d_j}{n_j}\right) \text{ και } U_{Wk} = \sum_{j=1}^{r} n_j\left(d_{kj} - \frac{n_{kj}d_j}{n_j}\right),$$

για k = 1, 2, … *g-1*. Οι ποσότητες αυτές εκφράζονται σε μορφή διανύσματος με *(g-1)* στοιχεία, τα οποία συμβολίζουμε με $\boldsymbol{U}_L$ και $\boldsymbol{U}_W$.

Επίσης χρειαζόμαστε εκφράσεις για τις διακυμάνσεις των U_{Lk} και U_{Wk}, καθώς και για την συνδιακύμανση μεταξύ ζευγών. Πιο συγκεκριμένα, η συνδιακύμανση μεταξύ U_{Lk} και $U_{Lk'}$ δίνεται από την σχέση

$$V_{Lkk'} = \sum_{j=1}^{r} \frac{n_{kj} d_j \left(n_j - d_j\right)}{n_j \left(n_j - 1\right)} \left(\delta_{kk'} - \frac{n_{k'j}}{n_j} \right),$$

για $k, k' = 1, 2, \ldots, g\text{-}1$, όπου $\delta_{kk'}$ είναι τέτοιο ώστε

$$\delta_{kk'} = \begin{cases} 1, & \alpha\nu\ k = k' \\ 0, & \delta\iota\alpha\phi o\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{cases}$$

Οι παραπάνω όροι στην συνέχεια συναθροίζονται στα πλαίσια του τύπου ενός πίνακα διακυμάνσεων-συνδιακυμάνσεων, $\boldsymbol{V}_L$, ο οποίος είναι ένας συμμετρικός πίνακας με τις διακυμάνσεις των U_{Lk} στην κύρια διαγώνιο, και τους όρους συνδιακυμάνσεων εκτός της διαγωνίου.

Ομοίως, ο πίνακας διακυμάνσεων-συνδιακυμάνσεων για το στατιστικό Wilcoxon είναι ο πίνακας $\boldsymbol{V}_W$, του οποίο το (k, k') στοιχείο ορίζεται ως

$$V_{Wkk'} = \sum_{j=1}^{r} n_j^2 \frac{n_{kj} d_j \left(n_j - d_j\right)}{n_j \left(n_j - 1\right)} \left(\delta_{kk'} - \frac{n_{k'j}}{n_j} \right),$$

για $k, k' = 1, 2, \ldots, g\text{-}1$.

Τέλος, προκειμένου να ελέγξουμε την μηδενική υπόθεση μη ύπαρξης διαφοράς στις ομάδες, κάνουμε χρήση του αποτελέσματος ότι το στατιστικό ελέγχου $\mathbf{U}_L' \mathbf{V}_L^{-1} \mathbf{U}_L$, ή $\mathbf{U}_W' \mathbf{V}_W^{-1} \mathbf{U}_W$, ακολουθεί X^2 κατανομή με $(g\text{-}1)$ βαθμούς ελευθερίας, όταν η μηδενική υπόθεση είναι αληθής.

5.11 Στρωματοποιημένοι έλεγχοι

Στην περίπτωση κατά την οποία η υπόθεση αναλογικού κινδύνου δεν ισχύει για κάποιον παράγοντα (confounding), μπορούμε να 'συμβιβάσουμε' την παραβίαση αυτή προσαρμόζοντας έναν *στρωματοποιημένο έλεγχο*. Ένα παράδειγμα είναι μία πολυκεντρική κλινική δοκιμή, στην οποία κάθε κέντρο αποτελεί ένα στρώμα. Αξίζει να σημειωθεί ότι καμία υπόθεση δεν γίνεται σχετικά με τις σχέσεις των στρωματοποιημένων συναρτήσεων κινδύνου για τα διάφορα επίπεδα του παράγοντα. Δεδομένου ότι η τελική προσαρμογή της μεταβλητής στρωματοποίησης είναι χωρίς περιορισμούς, αυτός είναι ένας αποτελεσματικός τρόπος να

αποκλειστεί η σύγχυση (confounding) του υπό εξέταση παράγοντα από μία συμμεταβλητή η οποία παραβιάζει την υπόθεση αναλογικού κινδύνου. Ωστόσο, επειδή δεν λαμβάνονται εκτιμήσεις, διαστήματα εμπιστοσύνης ή *p-value* για την μεταβλητή στρωματοποίησης, αυτή η προσέγγιση είναι λιγότερο χρήσιμη για κάθε παράγοντα άμεσου ενδιαφέροντος.

Μεμονωμένοι έλεγχοι *log-rank* ή *Wilcoxon* που βασίζονται στα δεδομένα από κάθε στρώμα θα παρείχαν αξιοσημείωτες πληροφορίες, ωστόσο ένας έλεγχος που συνδυάζει πληροφορίες από κάθε στρώμα θα παρείχε μια πιο ακριβή σύνοψη του αποτελέσματος. Σε τέτοιου είδους περιπτώσεις, μια στρωματοποιημένη εκδοχή των ελέγχων *log-rank* ή *Wilcoxon* μπορεί να εφαρμοσθεί. Κατ' ουσίαν, αυτό συνεπάγεται τον υπολογισμό των *U* και *V* στατιστικών τιμών για κάθε στρώμα, και στην συνέχεια τον συνδυασμό των τιμών αυτών για τα στρώματα. Στην συνέχεια περιγράφεται ο στρωματοποιημένος έλεγχος *log-rank*, αλλά μία στρωματοποιημένη εκδοχή του ελέγχου *Wilcoxon* μπορεί να ληφθεί με παρόμοιο τρόπο.

Έστω U_{Lk} η τιμή του στατιστικού *log-rank* για την σύγκριση δύο ομάδων, υπολογιζόμενο για το k^{o} στρώμα όπως ορίστηκε στην σχέση (5.5). Επίσης, θεωρούμε την διακύμανση του στατιστικού για το k^{o} στρώμα ως V_{Lk}, όπου V_{Lk} υπολογίζεται για κάθε στρώμα μέσω της σχέσης (5.6). Ο στρωματοποιημένος έλεγχος *log-rank* βασίζεται τότε στο στατιστικό

$$W_S = \frac{\left(\sum_{k=1}^{s} U_{Lk}\right)^2}{\sum_{k=1}^{S} V_{Lk}},$$

το οποίο ακολουθεί την X^2 κατανομή με ένα βαθμό ελευθερίας υπό την μηδενική υπόθεση ότι δεν υπάρχουν διαφορές στις συναρτήσεις επιβίωσης. Τέλος, συγκρίνουμε τις παρατηρούμενες τιμές αυτού του στατιστικού με τα ποσοστιαία σημεία της X^2 κατανομής προκειμένου να εξετασθεί η υπόθεση του ελέγχου σε κάθε στρώμα. Αθροίζοντας τις παραπάνω ποσότητες για κάθε στρώμα, προκύπτει το αντίστοιχο στατιστικό του ολικού ελέγχου, με μηδενική υπόθεση ότι δεν υπάρχει διαφοροποίηση μεταξύ των συναρτήσεων επιβίωσης για το σύνολο των στρωμάτων. Το στατιστικό αυτό ακολουθεί κατανομή X^2 με *k-1* βαθμούς ελευθερίας.

Αξίζει να σημειωθεί ότι η μέθοδος αυτή μπορεί να εφαρμοσθεί σε μία συνεχή μεταβλητή αφού πρώτα κατηγοριοποιηθεί. Η χρήση περισσότερων στρωμάτων περιορίζει πιο αποτελεσματικά το *confounding*, αλλά η ακρίβεια και η ισχύς μπορεί να επηρεάζονται αρνητικά εάν ο συγχέων παράγοντας στρωματοποιείται με πολλά επίπεδα, διότι η ισχύς δεν δανείζεται ανάμεσα στα στρώματα. Πέντε με έξι στρώματα γενικά επαρκούν, αλλά πρέπει να περιλαμβάνονται τουλάχιστον 5-7 γεγονότα σε κάθε στρώμα.

5.12 Έλεγχος τάσης

Θεωρούμε την περίπτωση κατά την οποία τρεις ή περισσότερες ομάδες δεδομένων επιβίωσης πρόκειται να συγκριθούν, και αυτές οι ομάδες είναι διατάξιμες με κάποιο τρόπο. Για τη σύγκριση αυτών των ομάδων με την χρήση του ελέγχου log-rank που περιγράφηκε παραπάνω, μπορεί να συμβεί η ανάλυση να μην οδηγήσει σε σημαντική διαφορά μεταξύ των ομάδων, παρ' όλο που ο κίνδυνος θανάτου αυξάνεται ή μειώνεται κατά μήκος αυτών. Πράγματι, ένας έλεγχος που χρησιμοποιεί πληροφορία σχετικά με την διάταξη των ομάδων είναι πιο πιθανό να αναγνωρίσει την ύπαρξη τάσης ως σημαντική σε σύγκριση με τον τυπικό έλεγχο log-rank. Ο έλεγχος τάσης εξετάζει την μηδενική υπόθεση,

$$H_0: S_1(t)=S_2(t)=\ldots=S_k(t)$$
$$H_1: S_1(t)\geq S_2(t)\geq\ldots\geq S_k(t)$$

Ο έλεγχος log-rank για τάση κατά μήκος των g διατεταγμένων ομάδων βασίζεται στο στατιστικό

$$U_T=\sum_{k=1}^{g} w_k\left(d_{k.}-e_{k.}\right),$$

όπου w_k είναι ένας βάρος το οποίο αντιστοιχίζεται στη k^{η} ομάδα, k=1,2, …, g, και

$$d_{k.}=\sum_{j=1}^{r_k} d_{kj}\,, \qquad e_{k.}=\sum_{j=1}^{r_k} e_{kj}\,,$$

είναι τα παρατηρούμενα και τα αναμενόμενα πλήθη θανάτων στη k^{η} ομάδα, όπου η άθροιση είναι πέραν των r_k χρόνων θανάτου στην ομάδα. Σημειώνουμε ότι η τελεία που περιέχεται στον συμβολισμό $d_{k.}$ και $e_{k.}$ δηλώνει την άθροιση ως προς το δείκτη που η τελεία αντιπροσωπεύει. Οι κωδικοί συνήθως λαμβάνονται έτσι ώστε να ισαπέχουν σε αντιστοιχία μιας γραμμικής τάσης κατά μήκος των ομάδων. Για παράδειγμα, εάν υπάρχουν τρεις ομάδες, οι κωδικοί μπορούν να ληφθούν ως 1, 2 και 3, αν και η ισοδύναμη επιλογή των -1, 0 και 1 συντελεί στην απλοποίησης των υπολογισμών. Η διακύμανση του U_T δίνεται από την σχέση

$$V_T=\sum_{k=1}^{g}\left(w_k-\overline{w}\right)^2 e_{k.}\,,$$

όπου $\overline{w}$ είναι το σταθμισμένο άθροισμα των ποσοτήτων w_k, στο οποίο τα αναμενόμενα πλήθη θανάτων, $e_{k.}$, είναι τα βάρη, και είναι

$$\overline{w} = \frac{\sum_{k=1}^{g} w_k e_{k.}}{\sum_{k=1}^{g} e_{k.}}.$$

Το στατιστικό $W_T = U_T^2 / V_T$ έχει μία X^2 κατανομή με 1 βαθμό ελευθερίας υπό την μηδενική υπόθεση μη ύπαρξης τάσης στις g ομάδες.

5.13 Διαστήματα Εμπιστοσύνης Bootstrap

Οι διαδικασίες Bootstrap (Efron and Tibshirani, 1994) προσεγγίζουν την δειγματική κατανομή των στατιστικών στοιχείων που μας ενδιαφέρουν μέσω διαδικασίας επαναδειγματοληψίας. Η τεχνική Bootstrap είναι ευρέως εφαρμόσιμη για την εύρεση τυπικών σφαλμάτων και διαστημάτων εμπιστοσύνης σε περιπτώσεις όπου οι ασυμπτωτικές μέθοδοι για τον υπολογισμό βάσιμων διαστημάτων εμπιστοσύνης δεν είναι αξιόπιστες λόγω κυρίως περιορισμένου μεγέθους δείγματος.

Γενικά, τα τυπικά σφάλματα και τα διαστήματα εμπιστοσύνης αντανακλούν την δειγματική κατανομή των υπό εξέταση στατιστικών. Η μέθοδος Bootstrap μπορεί να εφαρμοσθεί, για παράδειγμα, στην εκτίμηση των συντελεστών παλινδρόμησης. Η εκτίμηση αυτή προκύπτει από την σχετική συχνότητα τραβώντας επανειλημμένα ανεξάρτητα δείγματα ίδιου μεγέθους από τον πληθυσμό, και επαναϋπολογίζοντας τα στατιστικά σε κάθε νέο δείγμα. Σε κλασσικά προβλήματα, όπως είναι η γραμμική παλινδρόμηση, η δειγματική κατανομή των εκτιμήσεων των συντελεστών παλινδρόμησης είναι γνωστή σε θεωρητική βάση, ώστε να πληρούνται οι απαιτούμενες υποθέσεις.

Πιο συγκεκριμένα, το πραγματικό δείγμα αντιμετωπίζεται ως η πηγή του πληθυσμού, και τα δείγματα Bootstrap λαμβάνονται επαναλαμβανόμενα από αυτόν. Δείγματα Bootstrap ίδιου μεγέθους με αυτό του πραγματικού δείγματος – ένας καθοριστικός παράγοντας ακριβείας – επιτυγχάνονται λαμβάνοντας δείγματα με επανάθεση, έτσι ώστε σε ένα δεδομένο δείγμα Bootstrap κάποιες παρατηρήσεις εμφανίζονται περισσότερες από μία φορές, κάποιες μόνο μία και κάποιες καθόλου. Στην συνέχεια, σε κάθε ένα από τα μεγάλου πλήθους δείγματα Bootstrap, υπολογίζονται οι στατιστικές ποσότητες του ενδιαφέροντός μας.

Στο πλαίσιο της ανάλυσης δεδομένων χρόνου επιβίωσης, έχουν προταθεί μέθοδοι για την ανάλυση δεξιά λογοκριμένων χρόνων. Μία τέτοια μέθοδος χαρακτηρίζεται ως περιθώρια.

Αρχικά εκτιμούνται οι συντελεστές παλινδρόμησης, υποθέτοντας ανεξαρτησία, και έπειτα εφαρμόζεται η μέθοδος επαναδειγματοληψίας Bootstrap ώστε να διορθωθεί η διακύμανσή τους. Για συνολικά B δείγματα Bootstrap, το τυπικό μοντέλο του Cox με p συμμεταβλητές παράγει ένα πίνακα διαστάσεως B×p συντελεστών παλινδρόμησης (βλ. Κεφάλαιο 6 για την μοντελοποίηση χρόνων επιβίωσης). Για το λογάριθμο κάθε αναλογίας κινδύνου $\hat{\beta}_i$, i=1,…,p, τα τυπικά σφάλματα σύμφωνα με τα δείγματα Bootstrap εκτιμούνται ως η εμπειρική τυπική απόκλιση των B αντίστοιχων εκτιμήσεων $\hat{\beta}_i^{*1},...,\hat{\beta}_i^{*B}$,

$$\hat{\sigma}_B = \left[\left(\sum_{b=1}^{B}\left(\hat{\beta}_i^{*b} - \hat{\beta}_i^{*\bullet}\right)^2\right)\Big/(B-1)\right]^{\frac{1}{2}}, \quad \hat{\beta}_i^{*\bullet} = \frac{\sum_{b=1}^{B}\hat{\beta}_i^{*b}}{B}.$$

Υποθέτοντας κανονικότητα, το 95% διάστημα εμπιστοσύνης κατασκευάζεται ως $\hat{\beta}_i \pm 1{,}96 \cdot \hat{\sigma}_B$, όπου $\hat{\beta}_i$ είναι η εκτίμηση του μοντέλου του Cox από το πραγματικό δείγμα (Xiao and Abrahamowicz, 2009).

Τα δείγματα Bootstrap εφαρμόζονται, επιπλέον, στην επιλογή μεταβλητών του μοντέλου, όταν το πλήθος τους είναι μεγάλο και είναι απαραίτητη η μείωσή του ώστε να καταλήξουμε σε πιο απλό και οικονομικό μοντέλο. Χρησιμοποιώντας σύνολα δεδομένων Bootstrap, η σταθερότητα αναπαραγωγής μπορεί να ερευνηθεί για τα μοντέλα παλινδρόμησης. Για δεδομένα επιβίωσης, τα δεδομένα ασθενών με πλήρεις παρατηρήσεις (προγνωστικοί παράγοντες, χρονικό αποτέλεσμα και δείκτης λογοκρισίας) μπορούν να χρησιμοποιηθούν ως δειγματικές μονάδες. Γενικά, σημαντικές μεταβλητές πρέπει να περιλαμβάνεται στη πλειοψηφία των ασθενών, και οι συμπεριλαμβανόμενες συχνότητες μπορούν να χρησιμοποιηθούν ως κριτήριο της σημαντικότητας των μεταβλητών.

Λαμβάνοντας τα δείγματα Bootstrap, ένας στατιστικός αναλυτής αναγνωρίζει την αστάθεια του επιλεγμένου μοντέλου, ειδικότερα όταν επιλέγεται ένα σύνθετο μοντέλο το οποίο περιέχει πολλούς αδύναμους παράγοντες. Ένας αδύναμος παράγοντας έχει μικρή ισχύ να εισέλθει στο τελικό μοντέλο και περιλαμβάνεται σε κάποιες μόνο των επαναλήψεων. Από συσχετιζόμενους αδύναμους παράγοντες συνήθως μόνο ένας επιλέγεται ώστε να αντιπροσωπεύσει την αντίστοιχη επίδραση των μεταβλητών της ‘συσχετισμένης ομάδας’. των μεταβλητών αυτών. Εάν οι επιδράσεις είναι όμοιου μεγέθους, οι συγκεκριμένοι ‘αντιπρόσωποι’ που επιλέχθηκαν σε μία επανάληψη Bootstrap είναι τυχαίοι. Επιπλέον, οι μεταβλητές χωρίς κάποια επίδραση στο αποτέλεσμα επιλέγονται με πιθανότητα εξαρτώμενη στο επίπεδο σημαντικότητας. Εάν θεωρούνται ποικίλα τέτοια επίπεδα στην έρευνα, η

πιθανότητα τουλάχιστον μία από αυτές τις μεταβλητές να περιλαμβάνεται στο τελικό μοντέλο είναι υψηλή. Η σταθερότητα του επιλεγμένου μοντέλου μειώνεται με την αύξηση του πλήθους των υποψήφιων μεταβλητών (Sauerbrei and Schumacher, 2000).

Κ Ε Φ Α Λ Α Ι Ο 6

Μοντελοποίηση Δεδομένων Επιβίωσης

6.1 Εισαγωγή

Οι μη παραμετρικές μέθοδοι μπορούν να φανούν χρήσιμες στην ανάλυση ενός απλού δείγματος δεδομένων επιβίωσης, ή στην σύγκριση δύο ή περισσότερων ομάδων χρόνων επιβίωσης. Ωστόσο, στη πλειοψηφία των ιατρικών ερευνών, οι οποίες δημιουργούν τα δεδομένα επιβίωσης, συμπληρωματικές πληροφορίες μπορούν επίσης να καταγραφούν για κάθε άτομο, όπως μεταβλητές που αντιστοιχούν σε κάποια δημογραφικά χαρακτηριστικά. Επομένως οι τιμές αυτών των μεταβλητών, οι οποίες αναφέρονται ως *επεξηγηματικές μεταβλητές* (explanatory variables), θεωρούνται ως η έναρξη της έρευνας. Το τελικό σύνολο δεδομένων είναι, σε αυτή τη περίπτωση, πιο πολύπλοκο από αυτό που παρουσιάστηκε στο Κεφάλαιο 5, επομένως, οι μέθοδοι αυτοί θα ήταν γενικά ακατάλληλες να εφαρμοσθούν εδώ.

6.2 Μοντελοποίηση της συνάρτησης κινδύνου

Μέσω της προσέγγισης μοντελοποίησης στην ανάλυση των δεδομένων επιβίωσης, μπορούμε να εξετάσουμε τον τρόπο με τον οποίο η επιβίωση μιας ομάδας ασθενών εξαρτάται από τις τιμές μίας ή περισσότερων επεξηγηματικών μεταβλητών, των οποίων οι τιμές καταγράφονται για κάθε ασθενή στον χρόνο έναρξης. Αυτές μπορεί να είναι βασικές μεταβλητές, όπως η ηλικία (όταν δεν χρησιμοποιείται ως κλίμακα χρόνου) και το φύλο. Μπορεί να είναι παράγοντες συγκεκριμένου ενδιαφέροντος, όπως η θερμοκρασία σε μία δοκιμή φαρμάκου, ή μπορεί να είναι ‘ενοχλητικές μεταβλητές’, οι οποίες συμπεριλαμβάνονται βοηθητικά προκειμένου να περιγραφεί ο κίνδυνος των γεγονότων με την μεγαλύτερη δυνατή ακρίβεια, αλλά δεν ενδιαφερόμαστε άμεσα για την επίδρασή τους. Στην ανάλυση των δεδομένων επιβίωσης, το ενδιαφέρον εστιάζεται στον κίνδυνο θανάτου σε

κάθε χρονική στιγμή μετά τον χρόνο έναρξης της έρευνας. Ως αποτέλεσμα, στην ανάλυση επιβίωσης η συνάρτηση κινδύνου μοντελοποιείται άμεσα.

Υπάρχουν δύο ευρείς λόγοι για την μοντελοποίηση των δεδομένων επιβίωσης. Ένας στόχος της διαδικασίας μοντελοποίησης είναι να καθορίσουμε ποιος συνδυασμός ενδεχόμενων επεξηγηματικών μεταβλητών επηρεάζει την μορφή της συνάρτησης κινδύνου. Ο δεύτερος λόγος αναφέρεται στην λήψη εκτιμήσεων της συνάρτησης κινδύνου αυτής κάθε αυτής για ένα άτομο. Κάτι τέτοιο μπορεί να είναι αντικείμενο ενδιαφέροντος, αλλά μπορεί επιπλέον να γίνει εκτίμηση της συνάρτησης επιβίωσης μέσω της σχέσης που συνδέει τις δύο συναρτήσεις. Αυτό, με την σειρά του, οδηγεί σε εκτίμηση άλλων ποσοτήτων όπως ο διάμεσος χρόνος επιβίωσης, ο οποίος θα είναι μία συνάρτηση των επεξηγηματικών μεταβλητών του μοντέλου.

Το βασικό μοντέλο των δεδομένων επιβίωσης είναι το *αναλογικό μοντέλο επιβίωσης* (proportional hazards model). Το μοντέλο αυτό προτάθηκε από τον Cox (1972) και είναι επίσης γνωστό ως *μοντέλο παλινδρόμησης του Cox* (Cox regression model). Παρ' όλο που το μοντέλο αυτό στηρίζεται στην υπόθεση αναλογικού κινδύνου, δεν υποθέτουμε καμία συγκεκριμένη μορφή της κατανομής πιθανότητας για τους χρόνους επιβίωσης. Επομένως, το μοντέλο αναφέρεται ως *ημι-παραμετρικό* (semi-parametric model). Ωστόσο, προκειμένου να θεωρήσουμε σωστά τις έννοιες, είναι απαραίτητο να αντιμετωπίσουμε το μοντέλο ως δύο μέρη, το μοντέλο πιθανότητας και την μέθοδο εκτίμησης. Η όλη διαδικασία είναι σχεδιασμένη να εκτιμά διαφορές στον κίνδυνο συνδεδεμένες με τις συμμεταβλητές. Αυτό σημαίνει ότι ο απόλυτος κίνδυνος θεωρείται λιγότερο σχετικός, αν και σε πολλές περιπτώσεις, ο απόλυτος κίνδυνος είναι πράγματι σημαντικός.

6.3 Μοντέλο αναλογικού κινδύνου του Cox

Έστω ότι ασθενείς τυχαιοποιούνται έτσι ώστε να λάβουν είτε την τυπική είτε μία νέα θεραπεία, και έστω $h_S(t)$ και $h_N(t)$ είναι οι κίνδυνοι θανάτου στον χρόνο t για ασθενείς της τυπικής (Standard) και της νέας (New) θεραπείας, αντίστοιχα. Σύμφωνα με ένα απλό μοντέλο για τους χρόνους επιβίωσης των δύο ομάδων ασθενών, ο κίνδυνος στον χρόνο t για έναν ασθενή της νέας θεραπείας είναι ανάλογος του κινδύνου στον ίδιο χρόνο για ένα ασθενή της τυπικής θεραπείας. Αυτό το μοντέλο αναλογικού κινδύνου μπορεί να εκφρασθεί ως

$$h_N(t) = \psi\, h_S(t)$$

για κάθε μη αρνητική τιμή του t, όπου ψ είναι μία σταθερά. Είναι ένα ημι-παραμετρικό μοντέλο αφού μόνο το διάνυσμα των συμμεταβλητών **X** εισέρχεται στο μοντέλο με παραμετρική μορφή ενώ για την συνάρτηση κινδύνου δεν γίνεται καμία παραμετρική υπόθεση. Μία επίπτωση αυτής της υπόθεσης είναι ότι οι αντίστοιχες πραγματικές συναρτήσεις επιβίωσης για άτομα των δύο θεραπειών δεν διασταυρώνονται. Στόχος αυτού του μοντέλου είναι ο καθορισμός των συμμεταβλητών που επηρεάζουν την συνάρτηση κινδύνου και η εκτίμηση της συνάρτησης κινδύνου, και συνεπώς η εκτίμηση της συνάρτησης επιβίωσης.

Η τιμή της ψ είναι η αναλογία (ratio) των κινδύνων θανάτου σε κάθε χρόνο για ένα άτομο της νέας ως προς την τυπική θεραπείας, και επομένως η ψ είναι γνωστή ως *σχετικός κίνδυνος* (relative hazard ή hazard ratio). Εάν $\psi<1$, ο κίνδυνος θανάτου στο t είναι μικρότερος για ένα άτομο της νέας θεραπείας σε σχέση με ένα άτομο της τυπικής θεραπείας. Η νέα θεραπεία τότε αποτελεί μία βελτίωση της τυπικής. Εάν, όμως, $\psi>1$, ο κίνδυνος θανάτου στο t είναι μεγαλύτερος για ένα άτομο της νέας θεραπείας, και η τυπική θεραπεία είναι καλύτερης αποτελεσματικότητας.

Ένας εναλλακτικός τρόπος έκφρασης του μοντέλου οδηγεί σε ένα εύκολα γενικεύσιμο μοντέλο. Έστω ότι τα δεδομένα επιβίωσης είναι διαθέσιμα για n άτομα και συμβολίζουμε την συνάρτηση κινδύνου του $i^{ου}$ ατόμου ως $h_i(t)$, i=1, 2,..., n. Επίσης, θεωρούμε ως $h_0(t)$ την συνάρτηση κινδύνου ενός ατόμου που λαμβάνει την τυπική θεραπεία. Η συνάρτηση κινδύνου για ένα άτομο της νέας θεραπείας ορίζεται τότε ως $\psi\, h_0(t)$. Ο σχετικός κίνδυνος ψ δεν μπορεί να είναι αρνητικός, επομένως είναι βολικό να θέσουμε $\psi=exp(\beta)$. Η παράμετρος β είναι ο λογάριθμος του λόγου κινδύνου, δηλαδή $\beta=log\psi$, και κάθε τιμή του β στο εύρος $(-\infty,\infty)$ οδηγεί σε μία θετική τιμή της ψ. Σημειώνεται πως θετικές τιμές της β λαμβάνονται όταν ο λόγος κινδύνου, ψ, είναι μεγαλύτερος της μονάδας, που σημαίνει πως η νέα θεραπεία είναι κατώτερη της τυπικής.

Έστω X μία δείκτρια μεταβλητή η οποία λαμβάνει την τιμή μηδέν εάν το άτομο είναι στην τυπική θεραπεία και την μονάδα εάν το άτομο λαμβάνει την νέα θεραπεία. Εάν x_i είναι η τιμή της X για το $i^{ο}$ άτομο στην έρευνα, i=1, 2,…, n, η συνάρτηση κινδύνου για το άτομο αυτό μπορεί να γραφεί ως

$$h_i(t) = h_0(t)e^{\beta x_i}.$$

Αυτό είναι το αναλογικό μοντέλο κινδύνου για την σύγκριση δύο ομάδων θεραπειών.

Το μοντέλο αναλογικού κινδύνου μπορεί να γενικευτεί για την περίπτωση κατά την οποία ο κίνδυνος θανάτου σε ένα συγκεκριμένο χρόνο εξαρτάται από τις τιμές $x_1, x_2, ..., x_p$ των p επεξηγηματικών μεταβλητών $X_1, X_2, ..., X_p$. Οι τιμές αυτών των μεταβλητών μπορούμε να υποθέσουμε ότι καταγράφηκαν στον χρόνο έναρξης της έρευνας. Το σύνολο των τιμών των επεξηγηματικών μεταβλητών στο μοντέλο αναλογικού κινδύνου αντιπροσωπεύεται από το διάνυσμα **x**, οπότε $\mathbf{x}=(x_1, x_2, \ldots, x_p)'$. Έστω $h_0(t)$ είναι η συνάρτηση κινδύνου για ένα άτομο του οποίου οι τιμές του συνόλου των επεξηγηματικών μεταβλητών που συντελούν το διάνυσμα **x** είναι μηδέν. Η συνάρτηση $h_0(t)$ καλείται *Αναφορική Συνάρτηση Κινδύνου* (baseline hazard function).

Στη γενική του μορφή το μοντέλο αναλογικού κινδύνου είναι

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

Η ποσότητα η οποία αποτελεί την δύναμη στην οποία υψώνεται το εκθετικό καλείται *γραμμική συνιστώσα* του μοντέλου (linear component), αλλά είναι επίσης γνωστή και ως *σκορ κινδύνου* (risk score) ή *προγνωστικός δείκτης* (prognostic index) για το *i* άτομο. Το μοντέλο, επίσης, μπορεί να επανεκφρασθεί στην μορφή

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

και τότε θεωρείται ως ένα γραμμικό μοντέλο για τον λογάριθμο του λόγου κινδύνου.

Αξίζει να σημειώσουμε πως δεν υπάρχει σταθερός όρος στην γραμμική συνιστώσα του μοντέλου. Εάν ένας σταθερός όρος, έστω β_0, συμπεριληφθεί, η ΑΣΚ θα μπορούσε να επαναπροσδιορισθεί ως προς την κλίμακά της διαιρώντας την $h_0(t)$ με $exp(\beta_0)$ και ο σταθερός όρος θα απαλειφόταν. Επομένως, δεν κάνουμε καμία υπόθεση ως προς την πραγματική μορφή της ΑΣΚ $h_0(t)$.

6.3.1 Η γραμμική συνιστώσα του μοντέλου αναλογικού κινδύνου

Υπάρχουν δύο είδη μεταβλητών από τις οποίες μπορεί να εξαρτάται η συνάρτηση κινδύνου, οι ονομαζόμενες *μεταβλητές* (variates) και *παράγοντες* (factors). Η μεταβλητή λαμβάνει αριθμητικές τιμές οι οποίες είναι συχνά σε συνεχή κλίμακα μέτρησης. Ένας παράγοντας λαμβάνει ένα περιορισμένο σύνολο τιμών, το οποίο είναι γνωστό ως επίπεδα του παράγοντα.

Οι μεταβλητές, είτε μόνες είτε σε συνδυασμό, ενσωματώνονται εύκολα σε ένα μοντέλο αναλογικού κινδύνου. Κάθε μεταβλητή εμφανίζεται στο μοντέλο με έναν αντίστοιχο συντελεστή β. Σε μοντέλα που περιλαμβάνουν μόνο μεταβλητές, η ΑΣΚ $h_0(t)$ είναι η συνάρτηση κινδύνου για ένα άτομο για το οποίο όλες οι μεταβλητές που περιλαμβάνονται στο μοντέλο λαμβάνουν την τιμή μηδέν.

Για την περίπτωση στην οποία περιλαμβάνεται παράγοντας στο μοντέλο, υποθέτουμε ότι η εξάρτηση της συνάρτησης κινδύνου από έναν παράγοντα Α πρόκειται να μοντελοποιηθεί, όπου ο Α αποτελείται από α επίπεδα. Το μοντέλο για ένα άτομο για το οποίο το επίπεδο του Α είναι j χρειάζεται να ενσωματώσει τον όρο α_j ο οποίος αντιπροσωπεύει την επίδραση από το j επίπεδο του παράγοντα. Οι όροι α_1, α_2, …, α_α είναι γνωστοί ως *κύριες επιδράσεις* (main effects) του παράγοντα. Σύμφωνα με το μοντέλο αναλογικού κινδύνου, η συνάρτηση κινδύνου για ένα άτομο με παράγοντα Α στο επίπεδο j είναι $h_0(t)\ exp(\alpha_j)$. Η ΑΣΚ $h_0(t)$ ορίζεται να είναι ο κίνδυνος για ένα άτομο με τιμές όλων των επεξηγηματικών μεταβλητών ίσων με μηδέν. Προκειμένου να είμαστε συνεπείς με αυτόν τον ορισμό, ένα από τα επίπεδα α_j πρέπει να θεωρείται ότι είναι το επίπεδο αναφοράς. Μία επιλογή είναι να υιοθετήσουμε τον περιορισμό $\alpha_1=0$, ο οποίος αντιστοιχεί με το να θεωρήσουμε ότι ο αναφορικός κίνδυνος είναι ο κίνδυνος για ένα άτομο για το οποίο ο παράγοντας Α είναι στο πρώτο επίπεδο.

Τα μοντέλα τα οποία περιλαμβάνουν όρους που αντιστοιχούν σε παράγοντες μπορούν να εκφρασθούν ως γραμμικοί συνδυασμοί των ερμηνευτικών μεταβλητών ορίζοντας *δείκτριες μεταβλητές* (indicator ή dummy variables) για κάθε παράγοντα. Εάν υιοθετείται ο περιορισμός $\alpha_1=0$, ο όρος α_j περιλαμβάνεται στο μοντέλο ορίζοντας α-1 δείκτριες μεταβλητές Z_2, Z_3, ..., Z_α.

Όταν όροι που αντιστοιχούν σε περισσότερους του ενός παράγοντες πρόκειται να περιληφθούν στο μοντέλο, μπορούμε να ορίσουμε σύνολα δεικτριών μεταβλητών για κάθε παράγοντα. Σε αυτή την περίπτωση, ίσως είναι κατάλληλο να περιλάβουμε στο μοντέλο έναν όρο ο οποίος αντιστοιχεί στις ατομικές επιδράσεις για κάθε συνδυασμό επιπέδων δύο ή περισσοτέρων παραγόντων. Τέτοιες επιδράσεις είναι γνωστές ως *αλληλεπιδράσεις* (interactions).

Γενικά, εάν Α και Β είναι δύο παράγοντες, και ο κίνδυνος θανάτου εξαρτάται από τον συνδυασμό των επιπέδων των Α και Β, τότε τα Α και Β αλληλεπιδρούν. Εάν τα Α και Β έχουν α και b επίπεδα, αντίστοιχα, ο όρος που αντιπροσωπεύει μία αλληλεπίδραση μεταξύ των δύο αυτών παραγόντων ορίζεται ως $(\alpha\beta)_{jk}$, για j=1, 2, …,α και k=1, 2, …,b.

Στην στατιστική μοντελοποίηση, μία σημαντική αρχή είναι ότι ένας όρος αλληλεπίδρασης πρέπει να περιλαμβάνεται στο μοντέλο μόνο όταν περιλαμβάνονται οι αντίστοιχες κύριες επιδράσεις. Προκειμένου να περιλάβουμε τον όρο $(\alpha\beta)_{jk}$ στο μοντέλο, υπολογίζονται τα γινόμενα των δεικτριών μεταβλητών που σχετίζονται με τις κύριες επιδράσεις. Αξίζει να επισημάνουμε ότι υπάρχουν δύο παράμετροι που συνδέονται με την αλληλεπίδραση μεταξύ των Α και Β. Γενικά, εάν τα Α και Β έχουν a και b επίπεδα αντίστοιχα, η αλληλεπίδραση δύο παραγόντων έχει $(a-1)(b-1)$ παραμέτρους να συνδέονται μαζί της, ή με άλλα λόγια η ΑΒ έχει $(a-1)(b-1)$ βαθμούς ελευθερίας. Τέλος, ο όρος $(\alpha\beta)_{jk}$ ισούται με μηδέν όταν είτε ο Α είτε ο Β είναι στο πρώτο επίπεδο, δηλαδή όταν $j=1$ ή $k=1$.

6.3.2 Προσαρμογή μοντέλου αναλογικού κινδύνου

Η προσαρμογή του μοντέλου αναλογικού κινδύνου σε ένα παρατηρούμενο σύνολο δεδομένων επιβίωσης συνεπάγεται την εκτίμηση των αγνώστων παραμέτρων των επεξηγηματικών μεταβλητών $X_1, X_2, \ldots, X_p$ στην γραμμική συνιστώσα του μοντέλου, $\beta_1, \beta_2, \ldots, \beta_p$. Η ΑΣΚ $h_0(t)$ ενδέχεται να χρειάζεται, επίσης, να εκτιμηθεί. Οι δύο αυτοί όροι του μοντέλου μπορούν να εκτιμηθούν χωριστά. Αρχικά εκτιμούνται τα β και οι εκτιμήσεις αυτές χρησιμοποιούνται στην συνέχεια για να κατασκευάσουμε μία εκτίμηση για την ΑΣΚ. Αυτό είναι ένα σημαντικό αποτέλεσμα, αφού σημαίνει ότι προκειμένου να κάνουμε αναφορές για τις επιδράσεις των p ερμηνευτικών μεταβλητών $X_1, X_2, \ldots, X_p$ στον σχετικό κίνδυνο, $h_i(t)/h_0(t)$, δεν απαιτείται η εκτίμηση της $h_0(t)$.

Οι συντελεστές β στο μοντέλο αναλογικού κινδύνου, οι οποίοι είναι οι άγνωστες παράμετροι του μοντέλου, μπορούν να εκτιμηθούν μέσω της *μεθόδου μέγιστης πιθανοφάνειας* (method of maximum likelihood). Για να εφαρμόσουμε την μέθοδο αυτή, αρχικά, λαμβάνουμε την πιθανοφάνεια του δείγματος. Αυτή είναι η από κοινού πιθανότητα των παρατηρούμενων δεδομένων, θεωρούμενη ως συνάρτηση των άγνωστων παραμέτρων στο μοντέλο που υποθέτουμε. Για το μοντέλο αναλογικού κινδύνου, αυτή είναι μία συνάρτηση των παρατηρούμενων χρόνων επιβίωσης και των άγνωστων παραμέτρων β της γραμμικής συνιστώσας του μοντέλου. Οι εκτιμήσεις των β είναι εκείνες οι τιμές οι οποίες είναι οι πιο πιθανές στην βάση των παρατηρούμενων δεδομένων. Οι εκτιμήσεις μέγιστης πιθανοφάνειας είναι, επομένως, οι τιμές οι οποίες μεγιστοποιούν την συνάρτηση πιθανοφάνειας. Από υπολογιστικής άποψης, είναι πιο βολική η μεγιστοποίηση του λογαρίθμου της συνάρτησης

πιθανοφάνειας. Επιπλέον, εκτιμήσεις της διακύμανσης των εκτιμήσεων μέγιστης πιθανοφάνειας λαμβάνονται από την δεύτερη παράγωγο του λογαρίθμου της συνάρτησης πιθανοφάνειας.

Έστω ότι υπάρχουν διαθέσιμα δεδομένα για *n* άτομα, ανάμεσα στα οποία υπάρχουν *r* διακριτοί χρόνοι θανάτου και *n-r* δεξιά λογοκριμένοι χρόνοι επιβίωσης. Αρχικά υποθέτουμε πως αντιστοιχεί μόνο ένας θάνατος σε κάθε χρόνο θανάτου, οπότε δεν υπάρχουν *δεσμοί* (ties) στα δεδομένα. Ο τρόπος χειρισμού των δεσμών θα παρουσιαστεί στην συνέχεια. Οι *r* διατεταγμένοι χρόνοι θανάτου συμβολίζονται ως $t_{(1)}<t_{(2)}<\ldots<t_{(r)}$, οπότε $t_{(j)}$ είναι ο $j^{ος}$ διατεταγμένος χρόνος θανάτου. Το σύνολο των ατόμων που βρίσκονται σε κίνδυνο στο χρόνο $t_{(j)}$ συμβολίζεται με $R(t_{(j)})$, οπότε $R(t_{(j)})$ είναι το σύνολο των ατόμων που είναι εν ζωή και μη λογοκριμένα την χρονική στιγμή ακριβώς πριν τον χρόνο $t_{(j)}$. Η ποσότητα $R(t_{(j)})$ καλείται *σύνολο κινδύνου* (risk set).

Ο Cox (1972) έδειξε πως η σχετική συνάρτηση πιθανοφάνειας για το μοντέλο αναλογικού κινδύνου δίνεται από την σχέση

$$L(\beta)=\prod_{j=1}^{r}\frac{\exp\left(\beta'\mathbf{x}_{(j)}\right)}{\sum_{l\in R(t_{(j)})}\exp\left(\beta' x_l\right)}, \tag{6.1}$$

στην οποία το $\mathbf{x}_{(j)}$ είναι το διάνυσμα συμμεταβλητών για το άτομο που πεθαίνει στον j^{o} διατεταγμένο χρόνο θανάτου, $t_{(j)}$. Η άθροιση στον παρονομαστή αυτής της συνάρτησης πιθανοφάνειας είναι το άθροισμα των τιμών των ποσοτήτων $\exp\left(\beta' x\right)$ όλων των ατόμων που βρίσκονται σε κίνδυνο στον χρόνο $t_{(j)}$. Παρατηρούμε ότι το γινόμενο λαμβάνεται για τα άτομα των οποίων οι χρόνοι θανάτου έχουν καταγραφεί. Τα άτομα για τα οποία ο χρόνος επιβίωσης είναι λογοκριμένος δεν συνεισφέρουν στον αριθμητή της συνάρτησης log-likelihood, αν και συμπεριλαμβάνονται στον υπολογισμό μέσω του συνόλου κινδύνου στους χρόνους θανάτου που πραγματοποιούνται πριν από ένα λογοκριμένο χρόνο. Επιπλέον, η συνάρτηση πιθανοφάνειας εξαρτάται μόνο από τους διατεταγμένους χρόνους θανάτου, εφόσον αυτό καθορίζει το σύνολο κινδύνου σε κάθε χρόνο θανάτου. Επομένως, αναφορές για την επίδραση των ερμηνευτικών μεταβλητών στην συνάρτηση κινδύνου βασίζονται αποκλειστικά στην τάξη διάταξης των χρόνων επιβίωσης.

Αναφορικά με το μοντέλο που εξετάζουμε, η βάση της παραπάνω συνάρτησης πιθανοφάνειας είναι ότι διαστήματα μεταξύ διαδοχικών χρόνων θανάτου δεν παρέχουν πληροφορίες για την επίδραση των ερμηνευτικών μεταβλητών στον κίνδυνο θανάτου. Αυτό

οφείλεται στο γεγονός ότι η ΑΣΚ έχει αυθαίρετη μορφή, οπότε η $h_0(t)$, και επομένως η $h(t)$, είναι μηδέν σε αυτά τα χρονικά διαστήματα, στα οποία δεν υπάρχουν θάνατοι. Ως αποτέλεσμα, τα διαστήματα αυτά δεν παρέχουν κάποια πληροφορία για τις τιμές των παραμέτρων β. Θεωρούμε, λοιπόν, την πιθανότητα το $i^{ο}$ άτομο να πεθαίνει σε κάποια χρονική στιγμή $t_{(j)}$, δεσμεύοντας ως προς το $t_{(j)}$ να είναι ένας από τους r παρατηρούμενους χρόνους θανάτου $t_{(1)}, t_{(2)}, \ldots, t_{(r)}$. Εάν το διάνυσμα των επεξηγηματικών μεταβλητών για το άτομο που πεθαίνει στο χρόνο $t_{(j)}$ είναι $\mathbf{x}_{(j)}$, τότε η πιθανότητα είναι

$$\text{P[άτομο με μεταβλητές } \mathbf{x}_{(j)} \text{ πεθαίνει στο χρόνο } t_{(j)} \mid \text{συμβαίνει ένας θάνατος στο χρόνο } t_{(j)}] =$$

$$\frac{\text{P}[\textit{άτομο με μεταβλητές } x_{(j)} \textit{ πεθαίνει στο } t_{(j)}]}{P[\textit{συμβαίνει ένας θάνατος στο χρόνο t(j)}]}.$$

Εάν πεθαίνει το $i^{ο}$ άτομο στο χρόνο $t_{(j)}$, η συνάρτηση κινδύνου γράφεται ως $h_i(t_{(j)})$. Ο παρονομαστής είναι το άθροισμα των κινδύνων θανάτου στο χρόνο $t_{(j)}$ ως προς όλα τα άτομα που βρίσκονται σε κίνδυνο θανάτου σε αυτό το χρόνο, $R(t_{(j)})$. Επομένως, η δεσμευμένη πιθανότητα γίνεται

$$\frac{h_i\left(t_{(j)}\right)}{\sum_{l\in R(t_{(j)})} h_l\left(t_{(j)}\right)} = \frac{\exp\left(\beta'\mathbf{x}_{(j)}\right)}{\sum_{l\in R(t_{(j)})} \exp\left(\beta'\mathbf{x}_l\right)},$$

Τελικά, λαμβάνοντας το γινόμενο των δεσμευμένων πιθανοτήτων για τους r χρόνους θανάτου λαμβάνουμε την συνάρτηση πιθανοφάνειας της σχέσης (6.1). Η συνάρτηση πιθανοφάνειας που έχουμε λάβει δεν είναι η πραγματική πιθανοφάνεια, δεδομένου ότι δεν γίνεται άμεση χρήση των πραγματικών λογοκριμένων και μη λογοκριμένων χρόνων επιβίωσης. Για τον λόγο αυτό αναφέρεται ως *μερική συνάρτηση πιθανοφάνειας* (partial likelihood function).

Στην συνέχεια υποθέτουμε ότι τα δεδομένα αποτελούνται από n παρατηρούμενους χρόνους επιβίωσης, που συμβολίζονται με $t_1, t_2, \ldots, t_n$ και ότι δ_i είναι ένας δείκτης λογοκρισίας ο οποίος είναι μηδέν εάν ο $i^{ος}$ χρόνος επιβίωσης t_i, i=1, 2, n, είναι δεξιά λογοκριμένος, και μονάδα σε διαφορετική περίπτωση. Η συνάρτηση πιθανοφάνειας τότε μπορεί να λάβει την μορφή

$$\prod_{i=1}^{n}\left[\frac{\exp\left(\beta'\mathbf{x}_i\right)}{\sum_{l\in R(t_i)}\exp\left(\beta'\mathbf{x}_l\right)}\right],$$

όπου $R(t_i)$ είναι το σύνολο κινδύνου στον χρόνο t_i. Η αντίστοιχη συνάρτηση log-likelihood δίνεται από

$$\log L(\beta) = \sum_{i=1}^{n} \delta_i \left\{ \beta' \mathbf{x}_i - \log \sum_{l \in R(t_i)} \exp(\beta' \mathbf{x}_l) \right\}.$$

Οι εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων β στο μοντέλο αναλογικού κινδύνου μπορούν να βρεθούν μεγιστοποιώντας την συνάρτηση log-likelihood μέσω αριθμητικών μεθόδων. Η μεγιστοποίηση αυτή, γενικά, επιτυγχάνεται χρησιμοποιώντας την *διαδικασία Newton-Raphson*.

6.3.3 Δεσμοί

Το μοντέλο αναλογικού κινδύνου για δεδομένα επιβίωσης υποθέτει ότι η συνάρτηση κινδύνου είναι συνεχής, και υπό αυτή την υπόθεση, η ύπαρξη δεσμών στους χρόνους επιβίωσης δεν είναι δυνατή. Φυσικά, οι χρόνοι επιβίωσης συνήθως καταγράφονται στην πλησιέστερη ημέρα, μήνα ή έτος, και έτσι οι χρόνοι επιβίωσης με δεσμούς αναπτύσσονται ως αποτέλεσμα αυτής της διαδικασίας στρογγυλοποίησης.

Πέραν της πιθανότητας περισσότερων του ενός θανάτων σε δεδομένο χρόνο, μπορούν να υπάρξουν, επιπλέον, μία ή περισσότερες λογοκριμένες παρατηρήσεις σε ένα χρόνο θανάτου. Όταν υπάρχουν θάνατοι και λογοκριμένοι χρόνοι σε ένα δεδομένο χρόνο, η λογοκρισία υποθέτουμε πως συμβαίνει μετά από όλους τους θανάτους. Οπότε η δυναμική ασάφεια ως προς το ποιά άτομα πρέπει να συμπεριληφθούν στο σύνολο κινδύνου στο συγκεκριμένο χρόνο θανάτου επιλύεται, και οι λογοκριμένες παρατηρήσεις με δεσμούς δεν παρουσιάζουν πλέον δυσκολίες στον υπολογισμό της συνάρτησης πιθανοφάνειας.

Η κατάλληλη συνάρτηση πιθανοφάνειας στην παρουσία δεσμών έχει δοθεί από τους Kalbfleisch and Prentice (1980). Ωστόσο, η πιθανοφάνεια αυτή έχει πολύπλοκη μορφή. Επιπλέον, ο υπολογισμός της μπορεί να είναι πολύ χρονοβόρος, ειδικά όταν υπάρχει ένα σχετικά μεγάλο πλήθος δεσμών σε έναν ή περισσότερους χρόνους θανάτου. Προς διευκόλυνσή μας, υπάρχει ένα πλήθος προσεγγίσεων της συνάρτησης πιθανοφάνειας οι οποίες χαρακτηρίζονται από υπολογιστικά πλεονεκτήματα σε σχέση με την ακριβή μέθοδο.

Έστω $\mathbf{s}_j$ είναι το διάνυσμα αθροισμάτων για κάθε μία από τις p συμμεταβλητές για αυτά τα άτομα που πεθαίνουν στον $j^{\text{ο}}$ χρόνο θανάτου, $t_{(j)}$, j=1, 2,…, r. Εάν υπάρχουν d_j θάνατοι στον χρόνο $t_{(j)}$, το $h^{\text{ο}}$ στοιχείο του $\mathbf{s}_j$ είναι $s_{hj} = \sum_{k=1}^{d_j} x_{hjk}$, όπου x_{hjk} είναι η τιμή της $h^{\text{ης}}$

επεξηγηματικής μεταβλητής, h=1, 2,…,p, για το k^{o} από τα d_j άτομα, k=1, 2,…, d_j που πέθαναν στον j^{o} χρόνο θανάτου, j=1, 2, …, r.

Η απλούστερη προσέγγιση της συνάρτησης πιθανοφάνειας είναι αυτή του Breslow (1974), ο οποίος πρότεινε

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{exp(\beta' \mathbf{s}_j)}{\left[\sum_{l \in R(t_{(j)})} exp(\beta' x_l)\right]^{d_j}},$$

Σε αυτή τη προσέγγιση, οι d_j θάνατοι στον χρόνο $t_{(j)}$ θεωρούνται ότι είναι διακριτοί και ότι συμβαίνουν διαδοχικά. Οι πιθανότητες όλων των δυνατών διαδοχών θανάτου αθροίζονται προκειμένου να δώσουν την παραπάνω πιθανοφάνεια. Εκτός από μία σταθερά αναλογίας, αυτή είναι, επίσης, η προσέγγιση που πρότεινε ο Peto (1972). Η πιθανοφάνεια αυτή είναι αρκετά απλή να υπολογισθεί, και είναι μία επαρκής προσέγγιση όταν το πλήθος των παρατηρήσεων με δεσμούς σε κάθε χρόνο θανάτου δεν είναι πολύ μεγάλο.

Ο Efron (1977) πρότεινε

$$\prod_{j=1}^{r} \frac{\exp(\beta' \mathbf{s}_j)}{\prod_{k=1}^{d_j}\left[\sum_{l \in R(t_{(j)})} \exp(\beta' \mathbf{s}_l) - (k-1) d_k^{-1} \sum_{l \in D(t_{(j)})} \exp(\beta' \mathbf{x}_l)\right]},$$

ως μια προσέγγιση της πιθανοφάνειας για το μοντέλο αναλογικού κινδύνου, όπου $D(t_{(j)})$ είναι το σύνολο των ατόμων που πεθαίνουν στον χρόνο $t_{(j)}$. Αυτή είναι πλησιέστερη προσέγγιση της κατάλληλης συνάρτησης πιθανοφάνειας από αυτή του Breslow, παρ' όλο που στη πράξη, και οι δύο προσεγγίσεις συχνά καταλήγουν σε παρόμοια αποτελέσματα.

Ο Cox (1972) πρότεινε την προσέγγιση

$$\prod_{j=1}^{r} \frac{\exp(\beta' \mathbf{s}_j)}{\sum_{l \in R(t_{(j)}; d_j)} \exp(\beta' \mathbf{s}_l)},$$

όπου $R(t_{(j)}; d_j)$ δηλώνει ένα σύνολο από d_j άτομα τα οποία προέρχονται από το σύνολο κινδύνου στο $t_{(j)}$, $R(t_{(j)})$. Η άθροιση στον παρονομαστή είναι το άθροισμα όλων των δυνατών συνόλων των d_j ατόμων που επιλέχθηκαν ως δείγμα από το σύνολο κινδύνου χωρίς επανάθεση. Η προσέγγιση αυτή βασίζεται το μοντέλο για την περίπτωση όπου η κλίμακα του χρόνου είναι διακριτή, οπότε υπό αυτό το μοντέλο, επιτρέπονται οι παρατηρήσεις με δεσμούς. Για τον λόγο αυτό είναι γνωστή ως *διακριτή πιθανοφάνεια* ή *exact*.

6.4 Διαστήματα εμπιστοσύνης και έλεγχοι υποθέσεων για τα β

Στο μοντέλο του Cox, οι εκτιμώμενοι συντελεστές ακολουθούν την κανονική κατανομή όταν υπάρχει επαρκές πλήθος γεγονότων στο δείγμα. Η κανονική προσέγγιση είναι καλύτερη για τις εκτιμήσεις των συντελεστών σε σύγκριση με την αναλογία κινδύνου, οπότε οι έλεγχοι υποθέσεων και τα διαστήματα εμπιστοσύνης βασίζονται στους υπολογισμούς που αφορούν τους συντελεστές και τα τυπικά τους σφάλματα. Εάν υπάρχουν λιγότερα των 15-25 γεγονότων, η κανονική προσέγγιση δεν είναι αξιόπιστη, οπότε διαστήματα εμπιστοσύνης μέση της μεθόδου Bootstrap ίσως είναι καλύτερα. Για κάθε προγνωστικό παράγοντα ελέγχεται η μηδενική υπόθεση H_0: $\beta = 0$, ή ισοδύναμα, ότι η αναλογία κινδύνου ισούται με την μονάδα. Υπό την μηδενική υπόθεση, ο λόγος του εκτιμώμενου συντελεστή προς το τυπικό του σφάλμα τείνει σε τυπική κανονική, ή Z, κατανομή με μέσο 0 και τυπική απόκλιση 1.

Όταν είναι διαθέσιμα τα τυπικά σφάλματα των άγνωστων παραμέτρων β, είναι δυνατός ο υπολογισμός διαστημάτων εμπιστοσύνης. Ένα 100(1-α)% διάστημα εμπιστοσύνης για μία παράμετρο β είναι το διάστημα με όρια $\hat{\beta} \pm z_{a/2} s.e.(\hat{\beta})$, όπου $\hat{\beta}$ είναι η εκτίμηση του β και $z_{\alpha/2}$ είναι το άνω $\alpha/2$ σημείο της τυπικής κανονικής κατανομής.

Εάν το 100(1-α)% διάστημα εμπιστοσύνης για το β δεν περιέχει το μηδέν, αυτό είναι ένδειξη ότι η τιμή του β είναι μη μηδενική. Πιο συγκεκριμένα, η μηδενική υπόθεση β=0 μπορεί να ελεγχθεί υπολογίζοντας την τιμή του στατιστικού $\hat{\beta}/s.e(\hat{\beta})$. Η παρατηρούμενη τιμή του στατιστικού συγκρίνεται με το ποσοστιαίο σημείο της τυπικής κανονικής κατανομής προκειμένου να λάβουμε την αντίστοιχη *p-value*. Ισοδύναμα, το τετράγωνο αυτού του στατιστικού μπορεί να συγκριθεί με τα ποσοστιαία σημεία της X^2 κατανομής με ένα βαθμό ελευθερίας. Η διαδικασία αυτή είναι γνωστή ως *έλεγχος του Wald* (Wald test).

Ο έλεγχος του Wald εφαρμόζεται για να εξετάσουμε την μηδενική υπόθεση για κάθε παράγοντα ξεχωριστά, ενώ αν εξετάζεται μόνο ένας παράγοντας, τότε ο έλεγχος μπορεί να πραγματοποιηθεί και μέσω του ελέγχου Likelihood Ratio (LR). Σε διαφορετική περίπτωση, το LR test αντιστοιχεί στον ολικό έλεγχο, με μηδενική υπόθεση ότι όλοι οι συντελεστές ισούνται ταυτόχρονα με το μηδέν. Στην πλειοψηφία των περιπτώσεων τα αποτελέσματα των ελέγχων Wald και LR είναι όμοια αλλά όχι ακριβώς τα ίδια. Τα αποτελέσματα αυτά είναι πιο κοντινά όσο το μέγεθος του δείγματος είναι μεγαλύτερο ή η εκτιμώμενη αναλογία κινδύνου είναι κοντά στην μονάδα. Ωστόσο, σε σύνολα δεδομένων με μικρό πλήθος γεγονότων, ο

έλεγχος LR δίνει πιο ακριβή *p-value*, οπότε συνιστάται σε αυτές τις περιπτώσεις. Ποιοτική ασυμφωνία μεταξύ των αποτελεσμάτων των δύο ελέγχων μπορεί να δηλώνει ότι το μοντέλο περιλαμβάνει πολλούς προγνωστικούς παράγοντες για το πλήθος των γεγονότων που σημειώνονται.

Επιχειρώντας να ερμηνεύσουμε την *p-value* για μία δεδομένη παράμετρο, έστω β_j, είναι απαραίτητο να αναγνωρίσουμε ότι η υπόθεση που εξετάζουμε είναι β_j=0, δεδομένης της παρουσίας όλων των υπόλοιπων όρων που συμπεριλαμβάνονται στο μοντέλο. Γενικά, οι παράμετροι β_1, β_2,… δεν είναι όλες ανεξάρτητες μεταξύ τους. Αυτό σημαίνει ότι τα αποτελέσματα από τον ξεχωριστό έλεγχο των υποθέσεων για τις παραμέτρους β σε ένα μοντέλο ίσως να μην είναι εύκολα στην ερμηνεία τους. Λόγω αυτού απαιτούνται εναλλακτικές μέθοδοι για την σύγκριση διαφορετικών μοντέλων αναλογικού κινδύνου.

6.5 Προγνωστικοί παράγοντες

6.5.1 Κατηγορικοί παράγοντες

Η ερμηνεία δίτιμων παραγόντων απλοποιείται εφαρμόζοντας την κλίμακα 0/1 ως κωδικοποίηση. Οι εκθετικοί συντελεστές (exponantiated coefficients) δίνουν την αναλογία κινδύνου για την κατηγορία 1 έναντι της κατηγορίας 0, και μάλιστα παραμένει η κυριολεκτική ερμηνεία ως η αναλογία κινδύνου για μία μονάδα αύξησης του παράγοντα. Εναλλακτικές κωδικοποιήσεις, όπως 1-2, δίνουν τα ίδια αποτελέσματα, ωστόσο καθιστούν πολύπλοκη την ερμηνεία σε τυχόν παρουσίαση αλληλεπίδρασης. Το γεγονός αυτό θα έκανε και την ερμηνεία της ΑΣΚ δυσκολότερη.

Στην περίπτωση παραγόντων περισσότερων των δύο επιπέδων, η κωδικοποίησης γίνεται ως 1, 2, 3,… Συνήθως, το επίπεδο που αντιστοιχεί στο χαμηλότερο σκορ θεωρείται ως επίπεδο αναφοράς. Επιπλέον, είναι δυνατόν να πραγματοποιηθούν ανά δύο συγκρίσεις μεταξύ των επιπέδων του παράγοντα.

Σε κάποια σύνολα δεδομένων είναι δυνατόν να συναντήσουμε κατηγορίες στις οποίες δεν σημειώνεται κανένα γεγονός διότι η ομάδα είναι μικρή ή ο αθροιστικός κίνδυνος είναι χαμηλός. Η αναλογία κινδύνου που αντιστοιχεί σε κατηγορία αναφοράς χωρίς να έχει σημειωθεί σε αυτή κάποιο γεγονός είναι άπειρη, με την έννοια ότι δεν ορίζεται, και οι αντίστοιχοι έλεγχοι καθώς και τα διαστήματα εμπιστοσύνης είναι δύσκολο να ερμηνευθούν.

Σε αυτή τη περίπτωση η επιλογή κάποιας εναλλακτικής κατηγορίας αναφοράς μπορεί να διορθώσει το πρόβλημα, ωστόσο η αναλογία κινδύνου, το *Wald test* και για την κατηγορία χωρίς κανένα γεγονός, ως προς την νέα κατηγορία αναφοράς, παραμένουν δύσκολα στην ερμηνεία τους.

Ο ολικός έλεγχος για την συνολική επίδραση των κατηγορικών παραγόντων μπορεί να πραγματοποιηθεί μέσω των ελέγχων *Wald* ή *Likelihood Ratio*, με βαθμούς ελευθερίας ίσους με το πλήθος των επιπέδων μείον ένα. Η στατιστική σημαντικότητα των ζευγαρωτών συγκρίσεων πρέπει να ερμηνεύεται με προσοχή, ειδικά όταν ο ολικός έλεγχος δεν είναι στατιστικά σημαντικός. Για μεγάλο πλήθος κατηγοριών, οι πολλαπλές συγκρίσεις μπορούν να οδηγήσουν σε αύξηση του σφάλματος τύπου Ι. Επιπλέον, κάποιες συγκρίσεις είναι δυνατόν να σημειώνουν έλλειψη δύναμης λόγω μικρού πλήθους δεδομένων σε οποιαδήποτε υπό σύγκριση κατηγορία.

6.5.2 Συνεχείς παράγοντες

Η αναλογία κινδύνου για τους συνεχείς προγνωστικούς παράγοντες επηρεάζεται από την κλίμακα των μετρήσεων, και μία μονάδα αύξησης μπορεί να μην έχει κάποια αξιοσημείωτη ερμηνεία. Εάν το εύρος τιμών ενός παράγοντα είναι μεγάλο, δεν έχει κάποιο ιδιαίτερο νόημα να εξετάζουμε την αναλογία κινδύνου σε μία μονάδα αύξησης του παράγοντα, με την έννοια ότι θεωρώντας μία πολλαπλάσια αύξηση του παράγοντα, η αναλογία κινδύνου μπορεί να παρέχει καλύτερη ερμηνεία. Αν, για παράδειγμα, θεωρήσουμε την αναλογία κινδύνου για διαφορά *k* μονάδων του προγνωστικού παράγοντα, τότε έχουμε,

$$\frac{h_0(t)e^{\beta(x+k)}}{h_0(t)e^{\beta x}} = \frac{e^{\beta(x+k)}}{e^{\beta x}} = e^{\beta(x+k)-\beta k} = e^{\beta k}.$$

Επομένως, μία αλλαγή *k* μονάδων στον προγνωστικό παράγοντα πολλαπλασιάζει τον κίνδυνο κατά $e^{\beta k}$, χωρίς να έχει σημασία ποια είναι η τιμή αναφοράς *x* του παράγοντα. Ισοδύναμα , θα μπορούσαμε να υψώσουμε την εκτίμηση της αναλογίας κινδύνου για αύξηση μίας μονάδας του παράγοντα στην k^{η} δύναμη, δηλαδή $[\exp(\beta)]^k$, και να εκτιμήσουμε αντίστοιχα τα όρια εμπιστοσύνης. Αξίζει να σημειωθεί ότι μεταβολές στην κλίμακα μιας συνεχούς μεταβλητής δεν επηρεάζουν τους ελέγχους αυτούς.

6.6 Τυπικά σφάλματα και διαστήματα εμπιστοσύνης για την αναλογία κινδύνου

Έχουμε δει ότι σε καταστάσεις όπου υπάρχουν δύο ομάδες δεδομένων επιβίωσης, η παράμετρος β είναι ο λογάριθμος της αναλογίας κινδύνου θανάτου στο χρόνο *t* για τα άτομα της μίας ομάδας ως προς την άλλη. Επιπλέον, η αναλογία κινδύνου είναι $\psi=e^{\beta}$. Η αντίστοιχη εκτίμηση της αναλογίας κινδύνου είναι $\hat{\psi} = \exp(\hat{\beta})$, και το τυπικό σφάλμα του $\hat{\psi}$ μπορεί να εκτιμηθεί από το τυπικό σφάλμα του $\hat{\beta}$. Συνεπώς, η προσεγγιστική διακύμανση του $\hat{\psi}$, ως συνάρτηση του $\hat{\beta}$, είναι

$$\left\{\exp(\hat{\beta})\right\}^2 Var\left(\hat{\beta}\right),$$

που είναι $\hat{\psi}^2 Var\left(\hat{\beta}\right)$, και επομένως το τυπικό σφάλμα του $\hat{\psi}$ δίνεται ως

$$s.e.\left(\hat{\psi}\right) = \hat{\psi} s.e.\left(\hat{\beta}\right).$$

Γενικά, ένα διάστημα εμπιστοσύνης για την πραγματική αναλογία κινδύνου παρέχει περισσότερες πληροφορίες από το τυπικό σφάλμα της εκτιμώμενης αναλογίας κινδύνου. Ένα 100(1-α)% διάστημα για την πραγματική αναλογία κινδύνου, ψ, μπορεί να βρεθεί απλά, λαμβάνοντας το εκθετικό των ορίων εμπιστοσύνης για το β. Μία εκτίμηση διαστήματος που λαμβάνεται με αυτόν τον τρόπο είναι προτιμότερη από αυτή που προκύπτει μέσω του $\hat{\psi} \pm z_{a/2} s.e.\left(\hat{\psi}\right)$. Αυτό συμβαίνει διότι η κατανομή του λογαρίθμου της εκτιμώμενης αναλογίας κινδύνου προσεγγίζεται αμεσότερα από την κανονική κατανομή από ό,τι μέσω της αναλογίας κινδύνου αυτή κάθε αυτή.

6.7 Σύγκριση εναλλακτικών μοντέλων

Σε μία προσέγγιση μοντελοποίησης στην ανάλυση δεδομένων επιβίωσης, ένα μοντέλο αναπτύσσεται για την εξάρτηση της συνάρτησης κινδύνου από μία ή περισσότερες επεξηγηματικές μεταβλητές. Σε αυτή τη διαδικασία, προσαρμόζονται μοντέλα αναλογικού κινδύνου με γραμμικές συνιστώσες που περιέχουν διαφορετικά σύνολα όρων, και πραγματοποιούνται συγκρίσεις μεταξύ τους.

Έστω ότι δύο μοντέλα εξετάζονται για ένα συγκεκριμένο σύνολο δεδομένων, Μοντέλο (1) και Μοντέλο (2), όπου το Μοντέλο (1) περιλαμβάνει ένα υποσύνολο των όρων του Μοντέλου (2). Το Μοντέλο (1) τότε λέγεται ότι είναι *παραμετρικά εμφωλευμένο* (*parametrically nested*)

στο Μοντέλο (2). Πιο συγκεκριμένα, έστω ότι οι p επεξηγηματικές μεταβλητές $X_1, X_2, \ldots, X_p$ προσαρμόζονται στο Μοντέλο (1), οπότε η συνάρτηση κινδύνου υπό το μοντέλο αυτό είναι

$$h_0(t)\, exp\{\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p\}.$$

Επίσης, υποθέτουμε ότι οι $p+q$ επεξηγηματικές μεταβλητές $X_1, X_2, \ldots, X_p, X_{p+1},\ldots, X_{p+q}$ προσαρμόζονται στο Μοντέλο (2), οπότε η συνάρτηση κινδύνου υπό το μοντέλο αυτό είναι

$$h_0(t)\, exp\{\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p+\beta_{p+1} x_{p+1}+\ldots+\beta_{p+q} x_{p+q}\}.$$

Δεδομένου ότι το Μοντέλο (2) έχει μεγαλύτερο πλήθος όρων από το Μοντέλο (1), πρέπει να έχει καλύτερη προσαρμογή στα παρατηρούμενα δεδομένα. Το στατιστικό πρόβλημα ανάγεται στον προσδιορισμό εάν οι επιπρόσθετοι q όροι του μοντέλου (2) βελτιώνουν σημαντικά την επεξηγηματική ισχύ του μοντέλου. Εάν αυτό δεν συμβαίνει, τότε μπορούν να παραληφθούν και το Μοντέλο (1) θεωρείται ως επαρκές. Τέλος, αξίζει να σημειώσουμε ότι η επίδραση κάθε δεδομένου όρου εξαρτάται από τους υπόλοιπους όρους που περιλαμβάνονται στο μοντέλο.

6.7.1 Το στατιστικό $-2\log\hat{L}$ και σύγκριση εμφωλευμένων μοντέλων

Προκειμένου να συγκρίνουμε εναλλακτικά μοντέλα προσαρμοσμένα σε ένα παρατηρούμενο σύνολο δεδομένων, είναι απαραίτητο ένα στατιστικό το οποίο μετρά την έκταση στην οποία τα δεδομένα προσαρμόζονται σε ένα συγκεκριμένο μοντέλο. Εφόσον η συνάρτηση πιθανοφάνειας συνοψίζει την πληροφορία των άγνωστων παραμέτρων, ένα κατάλληλο συνοπτικό στατιστικό είναι η τιμή της συνάρτησης πιθανοφάνειας. Για ένα συγκεκριμένο σύνολο δεδομένων, όσο μεγαλύτερη η τιμή της μέγιστης πιθανοφάνειας, τόσο καλύτερη η συμφωνία μεταξύ του μοντέλου και των παρατηρούμενων δεδομένων.

Για την σύγκριση εναλλακτικών μοντέλων είναι πιο βολικό να χρησιμοποιούμε μείον δύο φορές τον λογάριθμο της μέγιστης πιθανοφάνειας. Εάν η μέγιστη πιθανοφάνεια για ένα δεδομένο μοντέλο συμβολίζεται με $\hat{L}$, το μέτρο συμφωνίας μεταξύ του μοντέλου και των δεδομένων είναι $-2\log\hat{L}$. Η $\hat{L}$ είναι μικρότερη της μονάδας, επομένως, η ποσότητα $-2\log\hat{L}$ θα είναι πάντα θετική, και για ένα δεδομένο σύνολο δεδομένων, όσο πιο μικρή η τιμή του $-2\log\hat{L}$ τόσο καλύτερο το μοντέλο.

Έστω και πάλι τα μοντέλα (1) και (2) όπως ορίστηκαν παραπάνω, και έστω η τιμή του λογαρίθμου της συνάρτησης μέγιστης πιθανοφάνειας για κάθε μοντέλο ορίζεται ως $\hat{L}(1)$ και

$\hat{L}(2)$, αντίστοιχα. Τα δύο μοντέλα μπορούν να συγκριθούν στη βάση της διαφοράς μεταξύ των τιμών των $-2\log\hat{L}$. Μία μεγάλη τιμή της διαφοράς αυτής οδηγεί στο συμπέρασμα ότι οι *q* μεταβλητές του Μοντέλου (2) οι οποίες είναι επιπρόσθετες σε αυτές του Μοντέλου (1) πράγματι βελτιώνουν την επάρκεια του μοντέλου. Φυσικά, η ποσότητα με την οποία η τιμή $-2\log\hat{L}$ μεταβάλλεται όταν προστίθενται όροι στο μοντέλο εξαρτάται από το ποίοι όροι βρίσκονται ήδη στο μοντέλο.

Το στατιστικό που αναφέρθηκε παραπάνω μπορεί να γραφεί ως

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

και ακολουθεί μία X^2 κατανομή, υπό την μηδενική υπόθεση ότι οι παράμετροι των επιπλέον μεταβλητών είναι μηδενικές. Οι βαθμοί ελευθερίας είναι ίσοι με την διαφορά μεταξύ των ανεξάρτητων παραμέτρων β που προσαρμόζονται στα δύο μοντέλα. Εάν η παρατηρούμενη τιμή του στατιστικού δεν είναι σημαντικά υψηλή, τα δύο μοντέλα κρίνονται να είναι ισοδύναμα κατάλληλα για τα δεδομένα. Σε αυτή τη περίπτωση, προτιμάται το απλούστερο μοντέλο, το οποίο είναι αυτό με τους λιγότερους όρους. Από την άλλη μεριά, εάν οι τιμές $-2\log\hat{L}$ των δύο μοντέλων διαφέρουν σημαντικά, υποστηρίζουμε ότι οι επιπρόσθετοι όροι είναι απαραίτητοι και υιοθετείται το πιο σύνθετο μοντέλο.

6.7.2 Στρατηγική επιλογής μοντέλου

Ένα πρώτο βήμα στην διαδικασία επιλογής μοντέλου είναι ο προσδιορισμός ενός συνόλου ερμηνευτικών μεταβλητών οι οποίες είναι δυνατόν να περιληφθούν στην γραμμική συνιστώσα του μοντέλου αναλογικού κινδύνου. Στην συνέχεια πρέπει να καθορισθεί ο συνδυασμός των μεταβλητών οι οποίες πρόκειται να χρησιμοποιηθούν στην μοντελοποίηση της συνάρτησης κινδύνου.

Μία σημαντική αρχή της στατιστικής μοντελοποίησης είναι ότι όταν περιλαμβάνεται στο μοντέλο όρος αλληλεπίδρασης, οι αντίστοιχοι όροι χαμηλότερης τάξης πρέπει, επίσης, να περιλαμβάνονται. Ο κανόνας αυτός είναι γνωστός ως *ιεραρχική αρχή* (hierarchical principle), και σημαίνει πως οι αλληλεπιδράσεις δεν πρέπει να προσαρμόζονται αν δεν συμπεριλαμβάνονται οι αντίστοιχες κύριες επιδράσεις. Μοντέλα που δεν είναι ιεραρχικά είναι δύσκολο να ερμηνευθούν.

6.7.3 Διαδικασία επιλογής μεταβλητών

Αρχικά θεωρούμε την περίπτωση κατά την οποία όλες οι επεξηγηματικές μεταβλητές βρίσκονται σε ισοδύναμη βάση, και στόχος είναι να προσδιορίσουμε υποσύνολα μεταβλητών από τις οποίες εξαρτάται η συνάρτηση κινδύνου. Όταν το πλήθος των ενδεχόμενων μεταβλητών, περιλαμβανομένων αλληλεπιδράσεων, μη γραμμικών όρων και ούτω καθεξής, δεν είναι πολύ μεγάλο, είναι δυνατόν να προσαρμόσουμε όλους τους δυνατούς συνδυασμούς όρων λαμβάνοντας δεόντως υπόψη την ιεραρχική αρχή. Εναλλακτικά εμφωλευμένα μοντέλα μπορούν να συγκριθούν εξετάζοντας την μεταβολή στην τιμή της ποσότητας $-2\log\hat{L}$ προσθέτοντας ή διαγράφοντας όρους στο μοντέλο.

Συγκρίσεις ενός πλήθους δυνατών μοντέλων, τα οποία δεν είναι απαραίτητο να είναι εμφωλευμένα μπορούν, επίσης, να πραγματοποιηθούν στην βάση του στατιστικού

$$AIC = -2\log\hat{L} + \alpha q\,,$$

στο οποίο q είναι το πλήθος των άγνωστων παραμέτρων β του μοντέλου και α είναι μία προκαθορισμένη σταθερά. Η τιμή της α συνήθως λαμβάνεται να είναι μεταξύ του 2 και 6. Η επιλογή $\alpha = 3$ είναι περίπου ισοδύναμη με την χρήση ενός 5% επιπέδου σημαντικότητας για να ελέγξουμε την διαφορά μεταξύ των τιμών $-2\log\hat{L}$ για δύο εμφωλευμένα μοντέλα τα οποία διαφέρουν από μία έως τρεις παραμέτρους. Το στατιστικό αυτό είναι γνωστό ως *κριτήριο πληροφορίας του Akaike* (Akaike's information criterion). Όσο μικρότερη είναι η τιμή του στατιστικού, τόσο καλύτερο το μοντέλο. Το ιδιαίτερο χαρακτηριστικό του στατιστικού είναι ότι εάν η μόνη διαφορά μεταξύ των δύο μοντέλων είναι ότι το ένα περιέχει μη αναγκαίες μεταβλητές, η τιμή του *AIC* για τα δύο μοντέλα δεν θα είναι πολύ διαφορετική.

Όταν το πλήθος των μεταβλητών είναι σχετικά μεγάλο, το πλήθος των δυνατών μοντέλων που χρειάζεται να προσαρμοσθεί είναι υπολογιστικά δαπανηρό. Πιο συγκεκριμένα, εάν υπάρχουν p ενδεχόμενες επεξηγηματικές μεταβλητές, υπάρχουν 2^p δυνατοί συνδυασμοί όρων, οπότε εάν $p>10$, υπάρχουν περισσότεροι από χίλιους δυνατούς συνδυασμούς. Σε αυτή την περίπτωση, αυτόματες ρουτίνες για επιλογή μεταβλητών που είναι διαθέσιμες στα στατιστικά πακέτα φαίνονται ως μία ελκυστική προοπτική. Οι ρουτίνες αυτές βασίζονται στις διαδικασίες *forward selection, backward elimination* ή σε ένα συνδυασμό των δύο γνωστό ως *stepwise procedure*.

Στην επιλογή της *forward selection*, οι μεταβλητές προστίθενται στο μοντέλο μία κάθε φορά. Σε κάθε στάδιο, η μεταβλητή που προστίθεται είναι αυτή η οποία δίνει την μεγαλύτερη μείωση στην τιμή $-2\log\hat{L}$ με την ένταξή της στο μοντέλο. Η διαδικασία τερματίζεται όταν η επόμενη υποψήφια για ένταξη δεν μειώνει την τιμή αυτή για περισσότερο από μία προκαθορισμένη ποσότητα, η οποία είναι γνωστή ως *κανόνας διακοπής* (stopping rule). Ο κανόνας αυτός συχνά διατυπώνεται σε όρους επιπέδου σημαντικότητας της διαφοράς των τιμών $-2\log\hat{L}$ όταν μία μεταβλητή προστίθεται στο μοντέλο.

Στην περίπτωση της *backward elimination*, αρχικά εφαρμόζεται ένα μοντέλο το οποίο περιλαμβάνει το μέγιστο πλήθος μεταβλητών υπό εξέταση. Οι μεταβλητές στη συνέχεια εξάγονται μία κάθε φορά. Σε κάθε στάδιο η μεταβλητή που αφαιρείται είναι αυτή η οποία αυξάνει την τιμή $-2\log\hat{L}$ κατά την μικρότερη. Η διαδικασία τερματίζεται όταν η επόμενη υποψήφια προς εξαγωγή μεταβλητή αυξάνει την τιμή $-2\log\hat{L}$ περισσότερο από μία προκαθορισμένη ποσότητα.

Η διαδικασία *stepwise* λειτουργεί με τον ίδιο τρόπο όπως η *forward selection*. Ωστόσο, μία μεταβλητή η οποία έχει εισαχθεί στο μοντέλο εξετάζεται για να εξαχθεί από αυτό σε επόμενο στάδιο. Επομένως, μετά την εισαγωγή μιας μεταβλητής στο μοντέλο, η διαδικασία ελέγχει εάν κάποια προηγούμενα εισηγμένη μεταβλητή μπορεί να διαγραφεί.

Οι παραπάνω αυτόματες ρουτίνες έχουν και κάποια μειονεκτήματα. Τυπικά, καταλήγουν στην διαμόρφωση ενός συγκεκριμένου υποσυνόλου μεταβλητών, παρά σε πλήθος ισοδύναμων υποσυνόλων. Τα υποσύνολα που βρίσκονται με αυτές τις ρουτίνες συχνά εξαρτώνται από την διαδικασία επιλογής μεταβλητών που έχει εφαρμοσθεί, και συχνά τείνει να μην λαμβάνει υπόψη την *ιεραρχική αρχή*. Επίσης, εξαρτώνται από τον κανόνα διακοπής που χρησιμοποιείται προκειμένου να καθορισθεί αν ένας όρος θα συμπεριληφθεί στο μοντέλο ή θα διαγραφεί.

Αντί την χρήση αυτόματων διαδικασιών επιλογής μεταβλητών, προτείνεται η ακόλουθη στρατηγική επιλογής μεταβλητών.

1. Το πρώτο βήμα είναι να προσαρμόσουμε μοντέλα κάθε ένα από τα οποία περιέχει μία μόνο από τις μεταβλητές που εξετάζουμε. Οι τιμές των $-2\log\hat{L}$ για τα μοντέλα αυτά συγκρίνονται με αυτή του μηδενικού μοντέλου.
2. Οι μεταβλητές που προκύπτουν να είναι σημαντικές από το Βήμα 1 προσαρμόζονται από κοινού σε ένα μοντέλο. Υπολογίζουμε την μεταβολή στις

τιμές των $-2\log\hat{L}$ όταν κάθε μεταβλητή παραλείπεται από το σύνολο. Μόνο αυτές που οδηγούν σε σημαντική αύξηση της τιμής $-2\log\hat{L}$ παραμένουν στο μοντέλο. Όταν μία μεταβλητή απορρίπτεται, η επίδραση που προκύπτει διαγράφοντας κάθε μία από τις εναπομένουσες μεταβλητές πρέπει να εξετασθεί.

3. Οι μεταβλητές οι οποίες δεν ήταν σημαντικές από μόνες τους, και, επομένως, δεν λήφθηκαν υπόψη στο Βήμα 2, μπορεί να είναι σημαντικές με την παρουσία των υπολοίπων. Αυτές οι μεταβλητές, επομένως, προστίθενται στο μοντέλο που προέκυψε στο Βήμα 2, μία κάθε φορά, και όποια μειώνει σημαντικά το $-2\log\hat{L}$ επιστρέφει στο μοντέλο.
4. Ένας τελικός έλεγχος πραγματοποιείται για να επιβεβαιώσουμε ότι κανένας όρος του μοντέλου δεν μπορεί να παραληφθεί χωρίς να αυξάνει σημαντικά την τιμή $-2\log\hat{L}$, και ότι κανένας όρος που δε συμπεριλαμβάνεται δεν μπορεί να μειώσει το $-2\log\hat{L}$ σημαντικά.

Όταν χρησιμοποιείται αυτή η διαδικασία επιλογής, αυστηρή εφαρμογή ενός συγκεκριμένου επιπέδου σημαντικότητας πρέπει να αποφεύγεται. Προκειμένου να προσδιορισθούν αποφάσεις για το αν θα συμπεριλάβουμε ή θα απορρίψουμε έναν όρο, το επίπεδο σημαντικότητας δεν πρέπει να είναι πολύ μικρό, συνίσταται ένα επίπεδο περίπου 10%.

Σε κάποιες εφαρμογές, ένα μικρό πλήθος αλληλεπιδράσεων και άλλων όρων υψηλότερης τάξης, όπως δυνάμεις συγκεκριμένων μεταβλητών, μπορεί να απαιτείται να εισαχθούν στο μοντέλο. Τέτοιοι όροι εισάγονται στο μοντέλο που προσδιορίστηκε στο Βήμα 3 παραπάνω, αφού διαπιστώσουμε ότι κάθε όρος που απαιτείται σύμφωνα με την ιεραρχική αρχή υπάρχει ήδη στο μοντέλο. Εάν ένας όρος μεγαλύτερης τάξης οδηγήσει σε σημαντική μείωση του $-2\log\hat{L}$, ο όρος αυτός πρέπει να συμπεριληφθεί στο μοντέλο.

Μία δεύτερη διαδικασία επιλογής μεταβλητών αφορά στην περίπτωση που κάποιες ερμηνευτικές μεταβλητές πρέπει οπωσδήποτε να εισαχθούν στο μοντέλο, όπως για παράδειγμα η μεταβλητή που δηλώνει τη θεραπευτική αγωγή. Η επιλογή των μεταβλητών γίνεται ως εξής:

1. Όταν το πλήθος των μεταβλητών είναι μικρό μελετούμε όλα τα μοντέλα που περιέχουν όλες τις ερμηνευτικές μεταβλητές εκτός εκείνων που έχει προαποφασισθεί ότι η επιλογή τους, έτσι ώστε να επιλέξουμε μερικές από αυτές.

Όταν ο αριθμός των μεταβλητών είναι μεγάλος η επιλογή τους γίνεται εφαρμόζοντας τα βήματα που αναφέρθηκαν παραπάνω.

2. Σχηματίζουμε το μοντέλο εισάγοντας τις προεπιλεγμένες μεταβλητές και αυτές που προέκυψαν από το Βήμα 1, και στη συνέχεια εξετάζουμε την περίπτωση αλληλεπιδράσεων.

6.8 Ερμηνεία εκτιμώμενων παραμέτρων

Όταν το μοντέλο αναλογικού κινδύνου εφαρμόζεται στην ανάλυση δεδομένων επιβίωσης, οι συντελεστές των ερμηνευτικών μεταβλητών μπορούν να ερμηνευτούν ως λογάριθμοι της αναλογίας του κινδύνου θανάτου. Αυτό σημαίνει πως εκτιμήσεις της αναλογίας κινδύνου, και αντίστοιχα διαστήματα εμπιστοσύνης, μπορούν εύκολα να βρεθούν από το προσαρμοσμένο μοντέλο. Ένα μοντέλο μπορεί να περιέχει όρους που αντιστοιχούν σε ένα πλήθος μεταβλητών, παραγόντων ή συνδυασμό των δύο. Με κατάλληλη κωδικοποίηση των δεικτριών μεταβλητών, που αντιστοιχούν στους παράγοντες του μοντέλου, οι εκτιμώμενες παράμετροι μπορούν και πάλι να ερμηνευθούν ως λογάριθμοι των αναλογιών κινδύνου. Όταν ένα μοντέλο περιέχει περισσότερες από μία μεταβλητές, η εκτίμηση της παραμέτρου που συνδέεται με μία συγκεκριμένη επίδραση είναι προσαρμοσμένη ως προς τις υπόλοιπες μεταβλητές του μοντέλου. Η ερμηνεία των παραμέτρων που αντιστοιχούν σε διαφορετικούς τύπους όρων του μοντέλου περιγράφονται στην συνέχεια.

Έστω ότι το μοντέλο αναλογικού κινδύνου περιέχει μία μοναδική συνεχή μεταβλητή X, οπότε η συνάρτηση κινδύνου του $i^{ου}$ από τα n άτομα, για το οποίο η X λαμβάνει την τιμή x_i, είναι

$$h_i(t) = h_0(t)e^{\beta x_i}.$$

Ο συντελεστής του x_i μπορεί να ερμηνευθεί ως ο λογάριθμος της αναλογίας κινδύνου. Στη συνέχεια θεωρούμε την αναλογία κινδύνου θανάτου ενός ατόμου για την οποία η τιμή x+1 καταγράφεται στην X, σε σχέση με αυτή για την οποία λαμβάνεται η τιμή x. Αυτό ορίζεται ως

$$\frac{\exp\{\beta(x+1)\}}{\exp(\beta x)} = e^{\beta},$$

οπότε η $\hat{\beta}$ στο προσαρμοσμένο μοντέλο αναλογικού κινδύνου είναι η εκτιμώμενη μεταβολή στον λογάριθμο της αναλογίας κινδύνου όταν η τιμή της X αυξάνεται κατά μία μονάδα.

Σε αναλογία των παραπάνω, η εκτιμώμενη μεταβολή στο log-hazard ratio όταν η τιμή της μεταβλητής X μεταβάλλεται κατά r μονάδες είναι $r\hat{\beta}$, και η αντίστοιχη εκτίμηση του hazard ratio είναι $\exp(r\hat{\beta})$. Όπως φαίνεται, η αναλογία κινδύνου στην αύξηση της X κατά r μονάδες δεν εξαρτάται από την πραγματική τιμή της X. Το χαρακτηριστικό αυτό είναι ένα άμεσο αποτέλεσμα της προσαρμογής της X ως ένας γραμμικός όρος του μοντέλου αναλογικού κινδύνου.

Στην περίπτωση των μοντέλων με ένα διακριτό παράγοντα, όταν τα άτομα ανήκουν σε μία από τις m ομάδες, $m \geq 2$, οι οποίες αντιστοιχούν στις κατηγορίες μιας ερμηνευτικής μεταβλητής, οι ομάδες μπορούν να αναπροσαρμόζονται ως προς τα επίπεδα του παράγοντα. Υπό το μοντέλο αναλογικού κινδύνου, η συνάρτηση κινδύνου για ένα άτομο της $j^{ης}$ ομάδας, j=1, 2, …, m, δίνεται από

$$h_j(t) = h_0(t)e^{\gamma_j},$$

όπου γ_j είναι η επίδραση του $j^{ου}$ επιπέδου του παράγοντα, και $h_0(t)$ είναι η ΑΣΚ. Το μοντέλο αυτό είναι υπερ-παραμετροποιημένο, οπότε λαμβάνουμε γ_1=0. Τότε η ΑΣΚ αντιστοιχεί στον κίνδυνο θανάτου τον χρόνο t για ένα άτομο της πρώτης ομάδας. Η αναλογία κινδύνου στο χρόνο t για ένα άτομο της $j^{ης}$ ομάδας, $j \geq 2$, σε σχέση με ένα άτομο της πρώτης ομάδας είναι $\exp(\gamma_j)$. Ένα μοντέλο το οποίο περιέχει τους όρους γ_j, j=1, 2, …, m, με γ_1=0, μπορεί να προσαρμοσθεί από $m-1$ δείκτριες μεταβλητές X_2, X_3,…, X_m. Η προσαρμογή του μοντέλου οδηγεί σε εκτιμήσεις των $\hat{\gamma}_2, \hat{\gamma}_3, ..., \hat{\gamma}_m$, και των τυπικών τους σφαλμάτων.

Σε κάποιες εφαρμογές, μπορεί να απαιτείται η αναλογία κινδύνου ως προς ένα επίπεδο του παράγοντα διαφορετικό από το πρώτο. Υπό αυτές τις συνθήκες, τα επίπεδα του παράγοντα, και οι σχετικές δείκτριες μεταβλητές, μπορούν να επαναπροσδιορισθούν έτσι ώστε κάποιο άλλο επίπεδο του παράγοντα να αντιστοιχεί στο απαιτούμενο επίπεδο αναφοράς, και το μοντέλο επαναπροσαρμόζεται. Οι συναρτήσεις κινδύνου για άτομα των επιπέδων j και j' του παράγοντα είναι αντίστοιχα, $h_0(t)\exp(\alpha_j)$ και $h_0(t)\exp(\alpha_{j'})$, οπότε η αναλογία κινδύνου για ένα άτομο του επιπέδου j, ως προς ένα του επιπέδου j', είναι $\exp(\alpha_j - \alpha_{j'})$.

6.9 Εκτίμηση των συναρτήσεων κινδύνων και επιβίωσης

Μέχρι τώρα, έχουμε προσεγγίσει μόνο την εκτίμηση των παραμέτρων β στην γραμμική συνιστώσα του μοντέλου αναλογικού κινδύνου. Αφού έχει προσδιορισθεί ένα κατάλληλο μοντέλο για το σύνολο των δεδομένων επιβίωσης, η συνάρτηση κινδύνου και η αντίστοιχη συνάρτηση επιβίωσης μπορούν να εκτιμηθούν. Στη συνέχεια αυτές οι εκτιμήσεις μπορούν να συνοψίσουν την εμπειρία επιβίωσης των ατόμων της έρευνας.

Έστω ότι η γραμμική συνιστώσα του μοντέλου αναλογικού κινδύνου περιέχει p μεταβλητές, X_1, X_2, …, X_p, και ότι οι εκτιμώμενοι συντελεστές αυτών των μεταβλητών είναι $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p$. Η συνάρτηση κινδύνου ενός ατόμου μπορεί να εκτιμηθεί εφόσον μπορεί να βρεθεί η εκτίμηση της $\hat{h}_0(t)$. Οι σχέσεις μεταξύ των συναρτήσεων κινδύνου, αθροιστικού κινδύνου και επιβίωσης μπορούν τότε να χρησιμοποιηθούν για να δώσουν εκτιμήσεις της αθροιστικής συνάρτησης κινδύνου και της συνάρτησης επιβίωσης.

Μία εκτίμηση της ΑΣΚ προήλθε από τους Kalbfleisch and Prentice (1973) χρησιμοποιώντας την προσέγγιση που βασίζεται στην μέθοδο μέγιστης πιθανοφάνειας. Έστω ότι υπάρχουν r διακριτοί χρόνοι θανάτου, οι οποίοι, όταν διατάσσονται σε αύξουσα τάξη, είναι $t_{(1)}<t_{(2)}<...<t_{(r)}$, και ότι υπάρχουν d_j θάνατοι και n_j άτομα σε κίνδυνο στον χρόνο $t_{(j)}$. Η εκτιμώμενη ΑΣΚ στο χρόνο $t_{(j)}$ τότε δίνεται από

$$\hat{h}_0\left(t_{(j)}\right) = 1 - \hat{\xi}_j,$$

όπου $\hat{\xi}_j$ είναι η λύση της σχέσης

$$\sum_{l \in D(t_{(j)})} \frac{exp\left(\hat{\beta}'\mathbf{x}_l\right)}{1 - \hat{\xi}_j^{exp(\hat{\beta}'\mathbf{x}_l)}} = \sum_{l \in R(t_{(j)})} exp\left(\hat{\beta}'\mathbf{x}_l\right), \tag{6.2}$$

για j=1, 2, …, r. Επίσης, $D\left(t_{(j)}\right)$ είναι το σύνολο των d_j ατόμων οι οποίοι πεθαίνουν στον j^{o} διατεταγμένο χρόνο θανάτου, $t_{(j)}$, και $R\left(t_{(j)}\right)$ είναι το σύνολο των n_j ατόμων σε κίνδυνο στο χρόνο $t_{(j)}$. Οι εκτιμήσεις των β οι οποίες αποτελούν το διάνυσμα $\hat{\boldsymbol{\beta}}$ είναι αυτές που μεγιστοποιούν την συνάρτηση πιθανοφάνειας της σχέσης (6.1).

Στην ειδική περίπτωση όπου δεν υπάρχουν δεσμοί στους χρόνους θανάτου, έχουμε ότι, όπου d_j=1 για j=1, 2, …, r, το αριστερό μέλος της σχέσης (6.2) αποτελείται από ένα μοναδικό όρο. Η σχέση αυτή, οπότε, μπορεί να λυθεί ώστε να δώσει

$$\hat{\xi}_j = \left(1 - \frac{exp\left(\hat{\beta}'\mathbf{x}_{(j)}\right)}{\sum_{l \in R(t_{(j)})} exp\left(\hat{\beta}'\mathbf{x}_l\right)}\right)^{exp\left(-\hat{\beta}'\mathbf{x}_{(j)}\right)},$$

όπου $\mathbf{x}_{(j)}$ είναι το διάνυσμα των επεξηγηματικών μεταβλητών για το άτομο που πεθαίνει στο χρόνο $t_{(j)}$.

Όταν υπάρχουν δεσμοί, που σημαίνει ότι, όταν ένα ή περισσότερα από τα d_j είναι μεγαλύτερα της μονάδας, η άθροιση του αριστερού μέλους της σχέσης (6.2) είναι το άθροισμα μιας σειράς κλασμάτων στα οποία η ποσότητα $\hat{\xi}_j$ εμφανίζεται στον παρονομαστή, υψωμένη σε διαφορετικές δυνάμεις. Η σχέση (6.2) τότε δεν λύνεται άμεσα και μία επαναληπτική διαδικασία απαιτείται.

Στην συνέχεια, υποθέτουμε ότι ο κίνδυνος θανάτου είναι σταθερός μεταξύ διαδοχικών χρόνων θανάτου. Τότε, το $\hat{\xi}_j$ μπορεί να θεωρηθεί ως εκτίμηση της πιθανότητας ότι το άτομο επιβιώνει μεταξύ του διαστήματος από $t_{(j)}$ έως $t_{(j+1)}$. Η αναφορική συνάρτησης επιβίωσης μπορεί να εκτιμηθεί από την σχέση

$$\hat{S}_0(t) = \prod_{j=1}^{k} \hat{\xi}_j, \tag{6.3}$$

για $t_{(k)} \le t \le t_{(k+1)}$, k=1, 2, …, r-1, και επομένως η εκτίμηση αυτή είναι μία κλιμακωτή συνάρτηση. Η εκτιμώμενη τιμή της αναφορικής συνάρτησης επιβίωσης είναι μηδέν για $t \ge t_{(r)}$ εάν δεν υπάρχουν λογοκριμένοι χρόνοι επιβίωσης μεγαλύτεροι από $t_{(r)}$. Σε αυτή τη περίπτωση, η $\hat{S}_0(t)$ δεν ορίζεται πέραν του $t_{(r)}$. Η αθροιστική αναφορική συνάρτηση κινδύνου δίνεται από την σχέση $H_0(t) = - \log S_0(t)$, επομένως, μία εκτίμηση της συνάρτησης αυτής είναι

$$\hat{H}_0(t) = -log\, \hat{S}_0(t) = -\sum_{j=1}^{k} log\, \hat{\xi}_j,$$

για $t_{(k)} \le t \le t_{(k+1)}$, k=1, 2, …, r-1.

Οι εκτιμήσεις των συναρτήσεων αναφορικού κινδύνου, επιβίωσης και αθροιστικού κινδύνου που ορίστηκαν παραπάνω μπορούν να χρησιμοποιηθούν ώστε να εκτιμηθούν οι αντίστοιχες τιμές ενός ατόμου με διάνυσμα επεξηγηματικών μεταβλητών $\mathbf{x}_i$. Η εκτιμώμενη αθροιστική συνάρτηση κινδύνου για το i^o άτομο δίνεται από τη σχέση

$$\hat{H}_i(t) = \hat{H}_0(t)\, exp\left(\hat{\beta}'\mathbf{x}_i\right). \tag{6.4}$$

Από την σχέση αυτή προκύπτει και η εκτιμώμενη συνάρτηση επιβίωσης για το i^{o} άτομο ως

$$\hat{S}_i(t) = \left[\hat{S}_0(t)\right]^{exp\left(\hat{\beta}'\mathbf{x}_i\right)}, \tag{6.5}$$

για $t_{(k)} \le t \le t_{(k+1)}$, k=1, 2, …, *r-1*. Σημειώνεται ότι, όταν εκτιμηθεί η συνάρτηση επιβίωσης, $\hat{S}_i(t)$, η εκτίμηση της αθροιστικής συνάρτησης κινδύνου είναι $-\log \hat{S}_i(t)$. Στην ειδική περίπτωση όταν δεν υπάρχουν συμμεταβλητές, η αντίστοιχη εκτίμηση της συνάρτησης επιβίωσης είναι

$$\prod_{j=1}^{k} \frac{n_j - d_j}{n_j},$$

η οποία είναι ο εκτιμητής *Kaplan-Meier* της συνάρτησης επιβίωσης. Από το αποτέλεσμα αυτό συμπεραίνουμε πως η εκτίμηση της συνάρτησης επιβίωσης της σχέσης (6.3) είναι μία γενίκευση του εκτιμητή *Kaplan-Meier* στην περίπτωση κατά την οποία η συνάρτηση κινδύνου εξαρτάται από επεξηγηματικές μεταβλητές.

6.9.1 Προσεγγιστική διαδικασία για την περίπτωση δεσμών

Όταν έχουμε χρόνους επιβίωσης με δεσμούς, ο εκτιμώμενος κίνδυνος αναφοράς μπορεί να βρεθεί μόνο μέσω επαναληπτικής μεθόδου για την λύση της σχέσης (6.2). Αυτή η επαναληπτική διαδικασία μπορεί να αποφευχθεί χρησιμοποιώντας μια προσέγγιση στην άθροιση του αριστερού μέλους της σχέσης αυτής.

Ο όρος

$$\hat{\xi}_j^{exp\left(\hat{\beta}'\mathbf{x}_l\right)}$$

στον παρονομαστή του αριστερού μέλους της σχέσης μπορεί να γραφεί ως

$$exp\left[e^{\hat{\beta}'\mathbf{x}_l}\, log\, \hat{\xi}_j\right],$$

και η αντίστοιχη εκτίμηση της συνάρτησης επιβίωσης, μετά από υπολογισμούς δίνεται από τη σχέση

$$S_0^*(t) = \prod_{j=1}^{k} exp\left(\frac{-d_j}{\sum_{l \in R(t_{(j)})} exp\left(\hat{\beta}'\mathbf{x}_l\right)}\right),$$

για $t_{(k)} \le t \le t_{(k+1)}$, k=1, 2, …, *r-1*. Όταν δεν υπάρχουν συμμεταβλητές, η εκτίμηση αυτή γίνεται

$$\prod_{j=1}^{k} exp\left(-d_j/n_j\right),$$

η οποία είναι γνωστή ως *εκτίμηση του Altshuler* (*Altshuler's estimate*) για την συνάρτηση επιβίωσης.

Μία αντίστοιχη εκτίμηση της αναφορικής αθροιστικής συνάρτησης κινδύνου είναι

$$H_0^*(t) = -log\ S_0^*(t) = \sum_{j=1}^{k} \frac{d_j}{\sum_{l \in R(t_{(j)})} exp\left(\hat{\beta}'\mathbf{x}_l\right)},$$

για $t_{(k)} \leq t \leq t_{(k+1)}$, k=1, 2, …, *r-1*.

Μία επιπλέον προσέγγιση του ξ_j ως $\tilde{\xi}_j$ ορίζεται

$$\tilde{\xi}_j = 1 - \frac{d_j}{\sum_{l \in R(t_{(j)})} exp\left(\hat{\beta}'\mathbf{x}_l\right)}.$$

Η εκτιμώμενη ΑΣΚ στον χρόνο $t_{(j)}$ δίνεται, τότε, από την σχέση

$$\tilde{h}_0\left(t_{(j)}\right) = \frac{d_j}{\sum_{l \in R(t_{(j)})} exp\left(\hat{\beta}'\mathbf{x}_l\right)}.$$

Η αθροιστική συνάρτηση κινδύνου βρίσκεται αθροίζοντας τους ατομικούς κινδύνους ώστε να προκύψει

$$\tilde{H}_0(t) = \sum_{j=1}^{k} \frac{d_j}{\sum_{l \in R(t_{(j)})} exp\left(\hat{\beta}'\mathbf{x}_l\right)},$$

για $t_{(k)} \leq t \leq t_{(k+1)}$, k=1, 2, …, *r-1*, η οποία είναι ίδια με την $H_0^*(t)$. Η αντίστοιχη εκτιμώμενη αναφορική συνάρτηση επιβίωσης είναι

$$\tilde{S}_0(t) = \prod_{j=1}^{k} \left(1 - \frac{d_j}{\sum_{l \in R(t_{(j)})} exp\left(\hat{\beta}'\mathbf{x}_l\right)}\right).$$

Στην συγκεκριμένη περίπτωση όπου δεν υπάρχουν συμμεταβλητές, οι εκτιμήσεις $\tilde{h}_0(t)$, $\tilde{H}_0(t)$ και $\tilde{S}_0(t)$ είναι ίδιες με αυτές που παρουσιάστηκαν παραπάνω και προκύπτουν από την *εκτίμηση του Altshuler*. Σχέσεις όμοιες με τις (6.4) και (6.5) μπορούν να χρησιμοποιηθούν για την εκτίμηση των συναρτήσεων αθροιστικού κινδύνου και επιβίωσης για ένα άτομο με διάνυσμα επεξηγηματικών μεταβλητών $\mathbf{x}_i$.

6.10 Ανάλυση υπολοίπων στο μοντέλο αναλογικού κινδύνου

Στην παράγραφο αυτή παρατίθεται ένα πλήθος υπολοίπων (residuals) τα οποία προκύπτουν από το μοντέλο αναλογικού κινδύνου με σταθερούς παράγοντες και σημειώνουν ενδιαφέρουσες διαγνωστικές ιδιότητες για το μοντέλο.

6.10.1 Cox-Snell υπόλοιπα

Τα υπόλοιπα Cox-Snell ορίζονται ως

$$\hat{r}_{C_j} = -log\ \hat{S}(t_j \mid \mathbf{X}_j) = \hat{H}(t_j \mid \mathbf{X}_j) = \hat{H}_0(t) exp\left(\sum_{k=1}^{p} \hat{\beta}_k X_{jk}\right).$$

Η ποσότητα $\hat{H}_0(t)$ υπολογίζεται μέσω της εκτίμησης κατά Breslow της αναφορικής αθροιστικής συνάρτησης κινδύνου, δηλαδή από τη σχέση,

$$\hat{H}_0(t) = \sum_{i:t_{(i)} \le t} \frac{d_i}{\sum_{j \in R(t_{(i)})} exp\left(\hat{\boldsymbol{\beta}}' \mathbf{X}_j\right)}.$$

Τα υπόλοιπα αυτά χρησιμοποιούνται για τον έλεγχο της ολικής επάρκειας του μοντέλου. Στην περίπτωση που το μοντέλο είναι ικανοποιητικό, τα υπόλοιπα πρέπει να προέρχονται από εκθετική κατανομή με παράμετρο 1. Επομένως, η γραφική απεικόνιση της ποσότητας $log\left[-log\left(\hat{S}\left(\hat{r}_{C_j}\right)\right)\right] = log\left[\hat{H}\left(\hat{r}_{C_j}\right)\right]$ έναντι της $log\left(\hat{r}_{C_j}\right)$ πρέπει να προσεγγίζει την γραφική παράσταση μίας ευθείας που περνά από την αρχή των αξόνων με κλίση μονάδα ($x = y$). Μειονέκτημα των υπολοίπων Cox-Snell αποτελεί το γεγονός ότι στην γραφική παράσταση δεν λαμβάνονται υπόψη αυτά που αντιστοιχούν σε λογοκριμένα δεδομένα.

6.10.2 Martingales και Deviance υπόλοιπα

Έστω δ_j δηλώνει την κατάσταση λογοκρισίας του j ατόμου λαμβάνοντας την τιμή 1 αν ο χρόνος του ατόμου είναι πλήρης και την τιμή 0 αν είναι λογοκριμένος. Τότε, τα Martingales υπόλοιπα ορίζονται ως

$$\hat{r}_{M_j} = \delta_j - \hat{r}_{C_j} = \delta_j - \hat{H}_0(t) exp\left(\sum_{k=1}^{p} \hat{\beta}_k X_{jk}\right).$$

Τα υπόλοιπα παίρνουν τιμές στο διάστημα $(-\infty,1]$, με αυτά που αντιστοιχούν σε λογοκριμένους χρόνους να λαμβάνουν αρνητική τιμή. Για μεγάλο πλήθος δειγμάτων τα Martingales υπόλοιπα είναι ασυσχέτιστα και έχουν μέση τιμή μηδέν, χωρίς να κατανέμονται συμμετρικά γύρω από αυτήν. Τα παραπάνω υπόλοιπα εντοπίζουν την ύπαρξη έκτροπων παρατηρήσεων (outliers) και την συναρτησιακή μορφή μίας ερμηνευτικής μεταβλητής. Η ανίχνευση αυτή πραγματοποιείται μέσω του διαγράμματος αυτών ως προς τις τιμές της νέας μεταβλητής της οποίας αναζητούμε την συναρτησιακή μορφή.

Δεδομένου του μειονεκτήματος της μη συμμετρικότητας γύρω από την τιμή 0 των Martingales residuals, η ανίχνευση των outliers είναι δύσκολη. Προς αποφυγήν αυτού, καταφεύγουμε στα Deviance residuals τα οποία κατανέμονται περισσότερο συμμετρικά γύρω από την τιμή μηδέν, επιτρέποντας έτσι την ευκολότερη ανίχνευση των outliers, και ορίζονται ως

$$\hat{r}_{D_j} = sgn\left(\hat{r}_{M_j}\right)\sqrt{-2\left[\hat{r}_{M_j} + \delta_j \, log\left(\delta_j - \hat{r}_{M_j}\right)\right]}.$$

Εάν το $\hat{r}_{M_j}$ λαμβάνει τιμές στο διάστημα $(-\infty, 0)$, τότε το αντίστοιχο $\hat{r}_{D_j}$ παίρνει τιμή κοντά στο μηδέν, ενώ εάν το $\hat{r}_{M_j}$ λαμβάνει τιμές στο διάστημα (0, 1), τότε το αντίστοιχο $\hat{r}_{D_j}$ παίρνει τιμή προς την κατεύθυνση του $+\infty$.

6.10.3 Schoenfeld, scaled Schoenfeld και rescaled Schoenfeld υπόλοιπα

Τα υπόλοιπα που παρουσιάστηκαν παραπάνω χαρακτηρίζονται από το μειονέκτημα ότι εξαρτώνται από την αθροιστική συνάρτηση κινδύνου η οποία πρέπει να εκτιμηθεί. Επιπλέον, υπάρχει μοναδική τιμή υπολοίπου σε κάθε άτομο ανεξαρτήτως του πλήθους των συμμεταβλητών. Προς αποφυγήν αυτών, εκτιμώνται τα Schoenfeld υπόλοιπα, τα οποία δίνουν τιμή σε κάθε άτομο και για κάθε μεταβλητή χωριστά. Τα υπόλοιπα αυτά για το άτομο *i* και την συμμεταβλητή X_g δίνονται από τη σχέση,

$$\hat{r}_{ig} = \delta_i \left(X_{ig} - \frac{\sum_{j \in R(t_i)} X_{jg} \, exp\left(\hat{\boldsymbol{\beta}}' \mathbf{X}_j\right)}{\sum_{j \in R(t_i)} exp\left(\hat{\boldsymbol{\beta}}' \mathbf{X}_j\right)} \right).$$

Τα Schoenfeld υπόλοιπα ισούνται με μηδέν όταν ο χρόνος ζωής του αντίστοιχου ατόμου είναι λογοκριμένος, για κάθε μεταβλητή.

Οι Grambsch and Therneau (1994) όρισαν τα scaled Schoenfeld υπόλοιπα, με κατάλληλη τυποποίηση των τελευταίων, ως

$$\hat{\mathbf{r}}^{*}_{(i)} \cong d \cdot \mathbf{I}_0^{-1}\left(\hat{\boldsymbol{\beta}}\right) \cdot \hat{\mathbf{r}}_{(i)},$$

όπου *d* δηλώνει το πλήθος των πλήρων χρόνων ζωής και

$$\mathbf{r}_{(i)} = \mathbf{r}_{(i)}\left(\boldsymbol{\beta}\right) = \left(\frac{\partial}{\partial \beta_g} log\, L_{(i)}\left(\boldsymbol{\beta}\right) \right)_{p\times 1},$$

και τα rescaled Schoenfeld υπόλοιπα, τα οποία ορίζονται ως

$$\hat{\mathbf{R}}_{(i)} \cong \hat{\boldsymbol{\beta}} + d \cdot \mathbf{I}_0^{-1}\left(\hat{\boldsymbol{\beta}}\right) \cdot \hat{\mathbf{r}}_{(i)}.$$

Από τα τελευταία μπορούμε να δούμε την ισχύ της υπόθεσης αναλογικού κινδύνου για κάθε μεταβλητή. Το αντίστοιχο διάγραμμα διασποράς δηλώνει πώς μεταβάλλεται η ποσότητα β_g στο χρόνο. Ελέγχοντας την υπόθεση μηδενικής κλίσης σε μία προσαρμοζόμενη ευθεία παρέχεται μία γραφική ένδειξη της ισχύος της υπόθεσης αναλογικού κινδύνου για κάθε μεταβλητή του μοντέλου. Μία μη μηδενική κλίση δηλώνει ότι ο β_g εξαρτάται από τον χρόνο, οπότε η υπόθεση δεν ισχύει.

6.10.4 Scored και scaled Scored υπόλοιπα

Τα Scored υπόλοιπα ορίζονται ως

$$L_{ig} = \delta_i\left(Z_{ig} - \bar{\alpha}_{ig}\right) - X_{ig} H\left(t_i \mid \mathbf{X}_i\right) + exp\left(\boldsymbol{\beta}'\mathbf{X}_i\right) \sum_{j:t_j \le t_i} \bar{\alpha}_{jg} \frac{\delta_j}{\sum_{l\in R(t_j)} exp\left(\beta' X_l\right)},$$

όπου

$$\bar{\alpha}_{ig} = \frac{\sum_{j\in R(t_i)} X_{jg}\, exp\left(\boldsymbol{\beta}'\mathbf{X}_j\right)}{\sum_{j\in R(t_i)} exp\left(\boldsymbol{\beta}'\mathbf{X}_j\right)},$$

Τα scored υπόλοιπα χρησιμοποιούνται για τον συνολικό (global) έλεγχο της υπόθεσης αναλογικού κινδύνου του μοντέλου.

Επιπλέον, από τα υπόλοιπα αυτά μπορούμε να δούμε την επιρροή κάθε παρατήρησης στην εκτίμηση των παραμέτρων **β** του μοντέλου (μόχλευση - leverage). Η διαδικασία αυτή απαιτεί την σύγκριση των εκτιμώμενων παραμέτρων πλην αυτών από τα οποία αφαιρείται η αντίστοιχη παρατήρηση. Η διαφορά αυτή αποτελεί τα scaled score υπόλοιπα. Αν η τιμή ενός

υπολοίπου είναι κοντά στο μηδέν, τότε η επιρροή της αντίστοιχης παρατήρησης είναι αμελητέα, (βλ. Αντζουλάκος, 2009).

Κ Ε Φ Α Λ Α Ι Ο 7

Ανάλυση Επιβίωσης των Δεδομένων Καρκίνου Μαστού

7.1 Εισαγωγή

Η ανάλυση επιβίωσης των γονιδιακών εκφράσεων πραγματοποιείται εφαρμόζοντας τον εκτιμητή *Kaplan-Meier* και το μοντέλο αναλογικού κινδύνου του Cox (βλέπε Κεφάλαια 5 και 6). Το σύνολο των δεδομένων που εξετάζουμε χαρακτηρίζεται από δεξιά λογοκρισία τύπου I, αφού κάθε γυναίκα εισήχθη στην έρευνα κατά την τυχαιοποίηση και πραγματοποιήθηκαν οι απαιτούμενες μετρήσεις έως την λήξη αυτής, εκτός εάν διακόπηκε η επικοινωνία μαζί της για οποιοδήποτε λόγο πέραν του θανάτου, για τον χρόνο επιβίωσης, ή και την επανεμφάνιση της νόσου, για την περίπτωση του χρόνου ελευθέρου νόσου (Disease Free Survival-DFS). Επίσης, στην περίπτωση που μελετάμε τον χρόνο Survival δεν σημειώθηκε κανένας δεσμός, ενώ για τον χρόνο DFS σημειώθηκαν δεσμοί κατά τις χρονικές στιγμές 8,72 και 10,39, κάθε ένας από τους οποίους σημειώνει δύο παρατηρήσεις (βλέπε πίνακα 7.1).

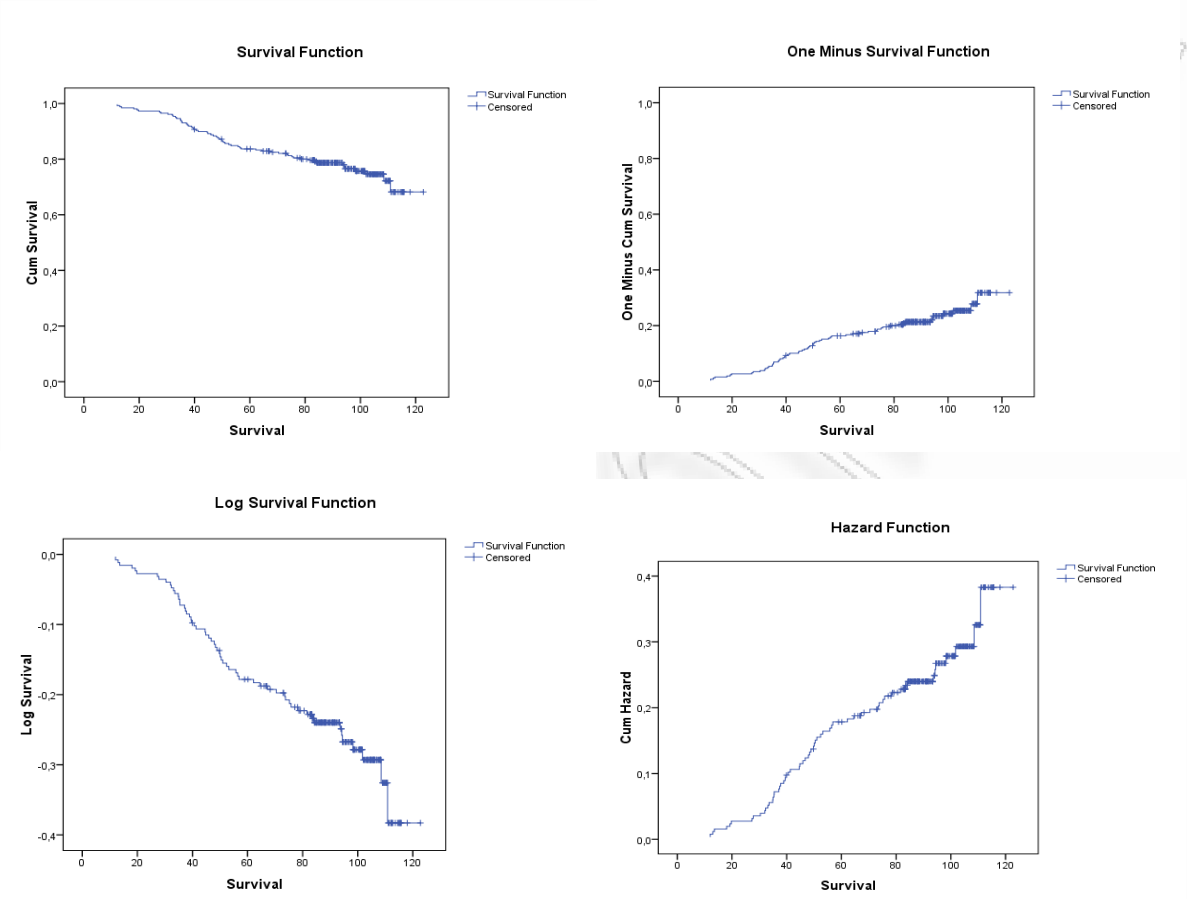
Οι έλεγχοι υποθέσεων πραγματοποιούνται σε επίπεδα σημαντικότητας 5% ή 10%, όπως αναφέρεται κατά περίπτωση.

7.2 Μη παραμετρική εκτίμηση του χρόνος επιβίωσης

7.2.1 Εκτίμηση *Kaplan-Meier* της συνάρτησης επιβίωσης

Ο εκτιμητής Kaplan-Meier είναι δυνατόν να εφαρμοσθεί σε σύνολα δεδομένων, στα οποία περιέχονται και λογοκριμένοι χρόνοι ζωής, γεγονός που ισχύει στα δεδομένα που εξετάζουμε. Συγκεκριμένα, το 76,4% των ασθενών σημείωσαν λογοκρισία, είτε επειδή ο θάνατος δεν συνέβη έως το χρονικό σημείο λήξης της έρευνας είτε επειδή χάθηκε η επικοινωνία μαζί

τους. Στο Παράρτημα Α.4 δίνεται ο πίνακας των εκτιμήσεων *Kaplan-Meier*, και στη συνέχεια παρουσιάζονται τα αντίστοιχα γραφήματα.



Σχήμα 7.1: Διαγράμματα επιβίωσης, 1-S(t), Log(S(t)) και κινδύνου της εκτίμησης *Kaplan-Meier* για τον χρόνο επιβίωσης.

Η καμπύλη που αναπαριστά την συνάρτηση επιβίωσης είναι σκαλωτή, φθίνουσα συνάρτηση με μέγιστη τιμή την μονάδα και συνεχής από αριστερά. Έως την χρονική διάρκεια των 40 εβδομάδων η καμπύλη της συνάρτησης επιβίωσης παραμένει επίπεδη σχεδόν, γεγονός που σημαίνει ότι ο κίνδυνος θανάτου παραμένει χαμηλός. Μετά το χρονικό αυτό σημείο, η κλίση της συνάρτησης γίνεται εντονότερη, με αποτέλεσμα ο κίνδυνος να επέλθει το γεγονός να είναι υψηλότερος.

Η εκτιμώμενη μέση τιμή είναι $\hat{\mu} = 104{,}582$ (95% CI: 100,459 – 108,704). Στο σημείο αυτό πρέπει να επισημάνουμε ότι ο τελευταίος χρόνος επιβίωσης είναι λογοκριμένος, οπότε η αντίστοιχη μέση τιμή δεν μπορεί να εκτιμηθεί. Ωστόσο, τα στατιστικά πακέτα, σε αυτή τη περίπτωση, θεωρούν το τελευταίο χρονικό σημείο ως χρόνο θανάτου και πραγματοποιείται μία μελέτη προσομοίωσης προκειμένου να διαφανεί πώς συμπεριφέρεται αυτή η διόρθωση

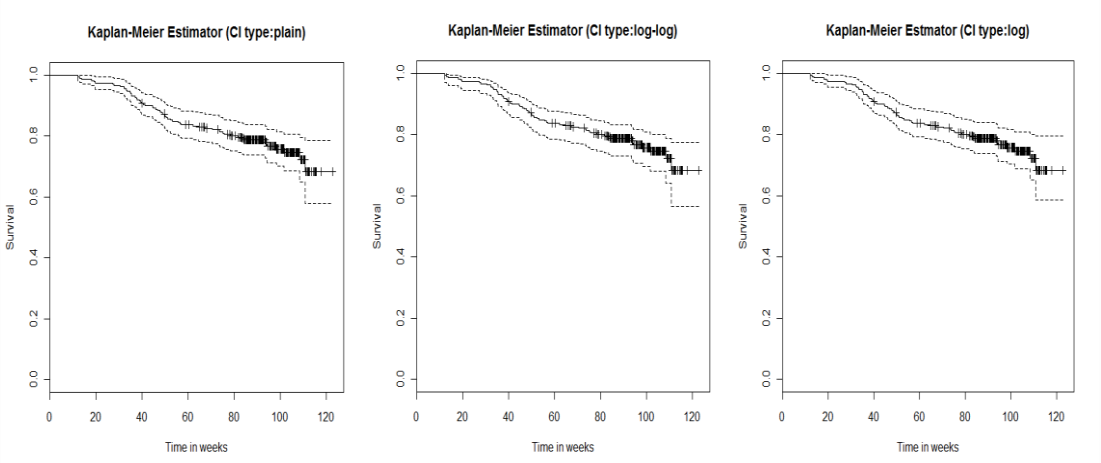
και να εκτιμηθεί, τελικά ο μέσος του χρόνου επιβίωσης (Zhong and Hess, 2009; Barker, 2009). Ωστόσο, στην εκτίμηση αυτή αντιστοιχεί υψηλή μεροληψία. Το 25% ποσοστιαίο σημείο $\hat{t}_{0.25} = 101{,}738$ δηλώνει την μικρότερη χρονική στιγμή στην οποία η καμπύλη μειώνεται κάτω του 1-0,25=75%. Το αντίστοιχο 95% διάστημα εμπιστοσύνης έχει ως εξής,

$$\hat{t}_{0,25} \pm z_{a/2}\sqrt{\hat{V}(\hat{t}_{0,25})} = \hat{t}_{0,25} \pm 1{,}96\sqrt{\hat{V}(\hat{t}_{0,25})} = 101{,}738 \pm 1{,}96 \cdot 8{,}363 = 94{,}197 \pm 16{,}391 =$$

$$(77.806\ ,\ 110.588).$$

Όπως γίνεται φανερό από το γράφημα της συνάρτησης επιβίωσης, κατά το χρονικό διάστημα των αρχικών 81,6 εβδομάδων δεν παρατηρείται μεγάλη συχνότητα λογοκριμένων δεδομένων, ενώ μετά του χρονικού αυτού διαστήματος, η πυκνότητα των παρατηρήσεων αυτών γίνεται ιδιαίτερα έντονη. Η συνολική διάρκεια επιβίωσης είναι προσεγγιστικά 110,9 εβδομάδες, και η αθροιστική αναλογία επιβίωσης δεν φαίνεται να ξεπερνά το 68,2%. Τέλος, αξίζει να σημειωθεί ότι, δεδομένου αυτού, δεν μπορεί να ορισθεί η διάμεσος και το 75% ποσοστιαίο σημείο. Σημειώνεται ότι με + συμβολίζεται λογοκριμένο δεδομένο.

Στην συνέχεια παρουσιάζονται τα διαγράμματα της συνάρτησης επιβίωσης, συμπεριλαμβανομένων και των διαστημάτων εμπιστοσύνης τύπου plain, log-log και log, αντίστοιχα.



Σχήμα 7.2: Εκτίμηση *Kaplan-Meier* με διαστήματα εμπιστοσύνης τύπου *plain*, *log-log* και *log*, αντίστοιχα.

7.2.2 Σύγκριση συναρτήσεων επιβίωσης ως προς τα κλινικά χαρακτηριστικά

Στην παρούσα παράγραφο παρουσιάζοντα τα αποτελέσματα των ελέγχων για την ισότητα των συναρτήσεων επιβίωσης, ως προς κάθε έναν από τους παράγοντες κλινικών χαρακτηριστικών. Οι έλεγχοι που πραγματοποιούνται αναφέρονται σε βάρη τύπου log-rank, Wilcoxon, Tarone-Ware και Flemming-Harrington για παραμέτρους ρ=1, π=0 (βλέπε παράγραφο 5.9). Η μηδενική υπόθεση που για έναν δίτιμο παράγοντα διαμορφώνεται ως

$$H_0: S_1(t) = S_2(t) \text{ έναντι της εναλλακτικής } H_1: S_1 \neq S_2.$$

Για μεταβλητές περισσότερων των δύο κατηγοριών ο έλεγχος διαμορφώνεται ανάλογα. Στο σημείο αυτό επισημαίνεται ότι στο Παράρτημα Α.5 παρατίθενται για κάθε παράγοντα οι έλεγχοι με τα διαφορετικά βάρη ώστε να διαμορφωθεί μια γενική εικόνα, ωστόσο, η απόφαση για την ισότητα ή μη των συναρτήσεων επιβίωσης λαμβάνεται σύμφωνα με την *υπόθεση αναλογικού κινδύνου* (βλέπε παράγραφο 5.9.3).

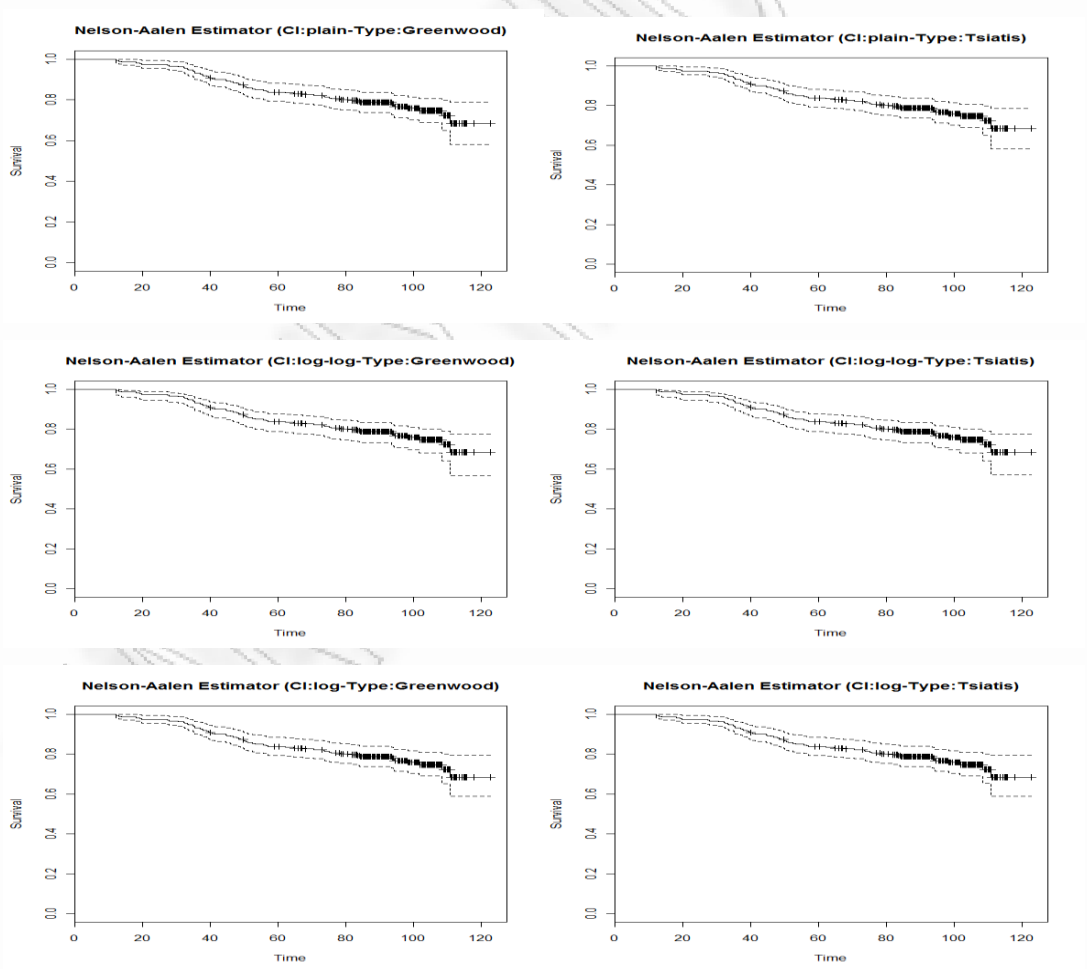
Για το συγκεκριμένο σύνολο δεδομένων και λαμβάνοντας υπόψη τον εκτιμητή *Kaplan-Meier*, καθώς και τους ελέγχους *log-rank* και *Wilcoxon*, προκύπτει, σύμφωνα με τις γραφικές ενδείξεις, ότι η υπόθεση αναλογικού κινδύνου δεν ισχύει για τους κλινικοπαθολογικούς παράγοντες ‘Group’, ‘Age’, ‘Menopausal’, ‘Size’, ‘RT’, ‘Surgery’ και ‘Interval’. Για τα συγκεκριμένα χαρακτηριστικά πιο αξιόπιστος είναι ο έλεγχος Wilcoxon, από τον οποίο προκύπτει ότι σε επίπεδο σημαντικότητας 5% για κανένα από αυτούς τους παράγοντες δεν μπορεί να απορριφθεί η μηδενική υπόθεση. Συνεπώς, οι ομάδες ασθενών που διαμορφώθηκαν από τα επίπεδα της θεραπευτικής αγωγής, της ηλικιακής ομάδας, της κατάστασης εμμηνόπαυσης, του μεγέθους του όγκου που αφαιρέθηκε, από την λήψη ή όχι ραδιοθεραπείας, από το είδος της επέμβασης και το χρόνο έως την εισαγωγή στην έρευνα δεν σημειώνουν διαφοροποίηση ως προς την εκτίμηση της συνάρτησης επιβίωσης. Ωστόσο, για τον παράγοντα του μεγέθους του όγκου, αν χαλαρώσουμε το επίπεδο σημαντικότητας στο 10%, τότε προκύπτει ότι είναι στατιστικά σημαντικός.

Για τους υπόλοιπους παράγοντες ‘Grade’, ‘Nodes’, ‘ER’, ‘PgR’ και ‘HT’ η υπόθεση αναλογικού κινδύνου φαίνεται να ικανοποιείται αφού οι αντίστοιχες γραμμές ακολουθούν παράλληλη πορεία. Θεωρώντας, λοιπόν, τον έλεγχο *log-rank* καταλήγουμε στο συμπέρασμα ότι σε επίπεδο σημαντικότητας 5%, μόνο οι παράγοντες ‘Grade’ και ‘Nodes’ προκύπτουν ως στατιστικά σημαντικοί. Όσον αφορά τον παράγοντα διαφοροποίησης του όγκου, η ομάδα με βαθμό διαφοροποίησης του όγκου ‘I–II’, αντιστοιχεί σε υψηλότερη πιθανότητα επιβίωσης, σε

σχέση με αυτή της ομάδας με βαθμό διαφοροποίησης ‘III – Αδιαφοροποίητο’. Αντίστοιχα, για το πλήθος αφαιρούμενων λεμφαδένων, παρατηρείται ότι η ομάδα που αντιστοιχεί σε πλήθος θετικών λεμφαδένων μεγαλύτερο των τεσσάρων αντιστοιχεί σε μικρότερη πιθανότητα επιβίωσης, με την διαφορά αυτή να γίνεται εντονότερη με το πέρασμα του χρόνου. Η γραμμή που αντιστοιχεί σε πλήθος 0 – 3 θετικών λεμφαδένων φαίνεται να ακολουθεί σταθερή πορεία, και μάλιστα σε υψηλά επίπεδα πιθανότητας επιβίωσης.

7.2.3 Nelson-Aalen εκτιμητής της συνάρτησης επιβίωσης και διαστήματα εμπιστοσύνης

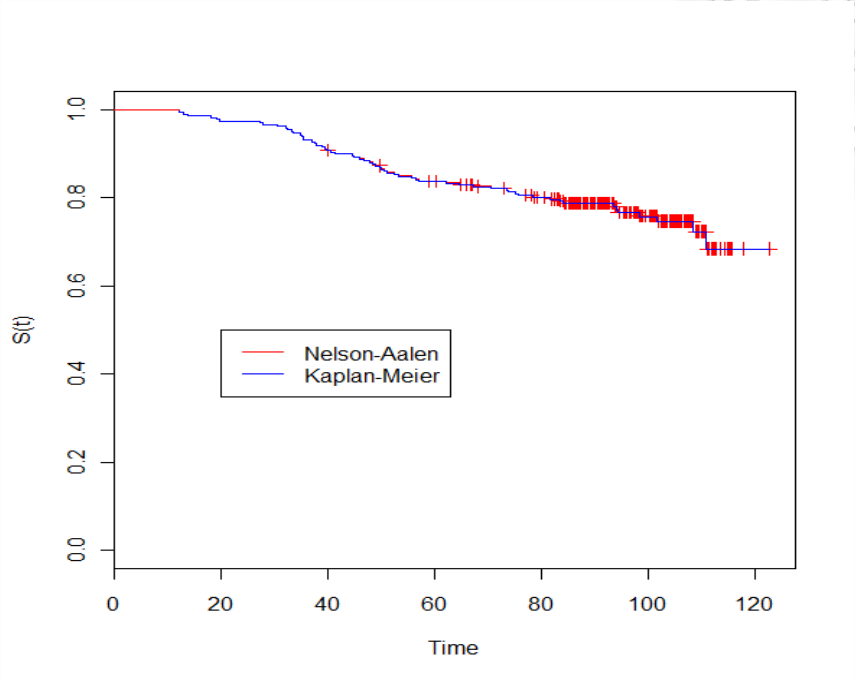
Στην συνέχεια παρουσιάζονται τα γραφήματα της Nelson-Aalen εκτίμησης της συνάρτησης επιβίωσης (βλέπε παράγραφο 5.4) με τα αντίστοιχα διαστήματα τύπου plain, log-log και log, με τους τύπους των Greenwood (1926) και Tsiatis (1978).



Σχήμα 7.3: Εκτιμητής Nelson-Aalen με τους τύπους των Greenwood και Tsiatis με διαστήματα εμπιστοσύνης τύπου plain, log-log και log.

7.2.4 Σύγκριση συναρτήσεων επιβίωσης Kaplan-Meier και Nelson-Aalen

Συγκρίνοντας τις δύο μεθόδους εκτίμησης των συναρτήσεων επιβίωσης, λαμβάνουμε το ακόλουθο γράφημα, από το οποίο προκύπτει γραφική ένδειξη ότι οι δύο γραμμές ταυτίζονται. Οπότε οι εκτιμητές καταλήγουν στα ίδια αποτελέσματα.



Σχήμα 7.4: Σύγκριση συναρτήσεων Kaplan-Meier και Nelson-Aalen.

7.2.5 Στρωματοποιημένοι έλεγχοι και έλεγχοι τάσης

Λαμβάνοντας υπόψη τους δύο παράγοντες ‘Grade’ και ‘Nodes’ ως προς τους οποίους οι συναρτήσεις επιβίωσης διαφέρουν, κρίνεται απαραίτητη η περαιτέρω διερεύνησή τους ώστε να διαπιστωθεί αν το αποτέλεσμα αυτό αντικατοπτρίζει πράγματι την διαφορετικότητα των συναρτήσεων επιβίωσης των επιπέδων τους ή αντανακλά την διαφοροποίηση των συναρτήσεων ως προς κάποιον άλλο υπό μελέτη παράγοντα. Για τον σκοπό αυτό πραγματοποιούνται στρωματοποιημένοι έλεγχοι, θεωρώντας ως στρώματα τους δύο αυτούς παράγοντες. Στην συνέχεια παρατίθενται τα αποτελέσματα των ‘αυτόνομων’ ελέγχων, εξετάζοντας την περίπτωση της ισότητας των συναρτήσεων επιβίωσης σε κάθε στρώμα χωριστά και των ολικών ελέγχων για την ταυτόχρονη σύγκριση των συναρτήσεων επιβίωσης (βλέπε παραγράφους 5.11 και 5.12).

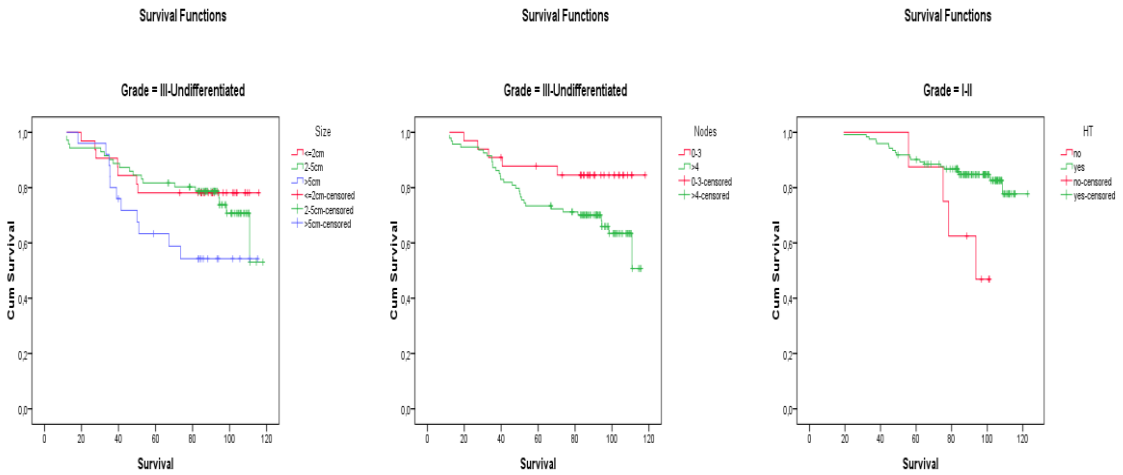
Παράγοντας	Επίπεδο Παράγοντα Grade	Αυτόνομοι Έλεγχοι				Ολικοί Έλεγχοι			
		Έλεγχος	Chi-Square	df	Sig.	Έλεγχος	Chi-Square	df	Sig.
ER	**I-II**	Log Rank	1,556	1	0,212	Log Rank	0,345	1	0,557
	III-Undiffer.	Wilcoxon	0,009	1	0,923	Wilcoxon	0,624	1	0,43
PgR	**I-II**	Wilcoxon	1,563	1	0,211	Wilcoxon	0,492	1	0,483
	III-Undiffer.	Wilcoxon	0,009	1	0,925				
Group	**I-II**	Wilcoxon	0,611	1	0,434	Wilcoxon	0,386	1	0,535
	III-Undiffer.	Log Rank	0,238	1	0,626	Log Rank	0,653	1	0,419
Age	**I-II**	Log Rank	0,602	1	0,438	Log Rank	0,229	1	0,632
	III-Undiffer.	Wilcoxon	0,034	1	0,855	Wilcoxon	0,26	1	0,61
Menopausal	**I-II**	Wilcoxon	0,002	1	0,964	Wilcoxon	0,065	1	0,799
	III-Undiffer.	Wilcoxon	0,136	1	0,712				
Surgery	**I-II**	Wilcoxon	0,018	1	0,894	Wilcoxon	0,172	1	0,678
	III-Undiffer.	Wilcoxon	0,205	1	0,651				
Interval	**I-II**	Wilcoxon	0,547	2	0,761	Wilcoxon	0,078	2	0,962
	III-Undiffer.	Wilcoxon	0,324	2	0,851				
Size	**I-II**	Wilcoxon	1,836	2	0,399	Wilcoxon	4,264	2	0,119
	III-Undiffer.	Wilcoxon	4,641	2	0,098				
Nodes	**I-II**	Log Rank	1,836	1	0,175	Log Rank	5,159	1	0,023
	III-Undiffer.	Log Rank	3,325	1	0,068				
RT	**I-II**	Log Rank	0,484	1	0,486	Log Rank	1,704	1	0,192
	III-Undiffer.	Wilcoxon	0,695	1	0,404	Wilcoxon	1,319	1	0,251
HT	**I-II**	Wilcoxon	3,299	1	0,069	Log Rank	3,042	1	0,081
	III-Undiffer.	Log Rank	0,419	1	0,518	Wilcoxon	2,506	1	0,113

Πίνακας 7.1: Στρωματοποίηση ως προς τον Παράγοντα Grade.

Θεωρώντας την στρωματοποίηση ως προς τον παράγοντα διαφοροποίησης του μεγέθους του όγκου που αφαιρέθηκε από τις ασθενείς (Grade), προκύπτει, σύμφωνα με τους στρωματοποιημένους ‘αυτόνομους’ ελέγχους, ότι δεν υπάρχει διαφοροποίηση των συναρτήσεων επιβίωσης ως προς τα επίπεδα των παραγόντων σε επίπεδο σημαντικότητας 5%. Ωστόσο, εάν χαλαρώσουμε το επίπεδο σημαντικότητας στο 10%, τότε προκύπτει ότι οι παράγοντες ‘Size’ και ‘Nodes’ προκαλούν διαφοροποίηση των συναρτήσεων επιβίωσης (με τιμές του στατιστικού 4,641 και 3,315, αντίστοιχα), όταν ο βαθμός διαφοροποίησης του όγκου είναι III ή ‘Αδιαφοροποίητο’. Αντίστοιχα, λαμβάνοντας τον παράγοντα ‘HT’ σημειώνεται διαφοροποίηση των συναρτήσεων επιβίωσης, όταν ο βαθμός διαφοροποίησης του όγκου είναι I ή II (X^2=3,299).

Λαμβάνοντας υπόψη τους ολικούς ελέγχους, επιβεβαιώνεται ότι υπάρχουν ενδείξεις ότι οι συναρτήσεις επιβίωσης των ομάδων που δημιουργούνται από τους παράγοντες ‘Nodes’ και ‘HT’ διαφέρουν για τα στρώματα ‘III-Undifferentiated’ και ‘I-II’, αντίστοιχα, σε επίπεδο σημαντικότητας 10%. Αξιοσημείωτο είναι το γεγονός ότι ο ολικός έλεγχος ως προς τον παράγοντα του μεγέθους του όγκου δεν καταγράφει την διαφορά των συναρτήσεων που διαφαίνεται μέσω των ελέγχων που αναφέρονται σε κάθε στρώμα χωριστά.

Στην συνέχεια παρουσιάζονται τα διαγράμματα που διαγράφουν την πορεία των συναρτήσεων επιβίωσης στον χρόνο για τις τρεις παραπάνω περιπτώσεις.



Σχήμα 7.5: Διάγραμμα επιβίωσης για στρώμα ‘III-Αδιαφοροποίητο’ του παράγοντα Grade για Size και Nodes και στρώμα ‘I-II’ του παράγοντα Grade για HT.

Ως προς τον παράγοντα του μεγέθους του όγκου που αφαιρέθηκε κατά την χειρουργική επέμβαση, παρατηρείται να μην υπάρχει κάποια έντονη διαφορά στις συναρτήσεις που αντιστοιχούν σε μέγεθος έως 5cm, ενώ η πιθανότητα επιβίωσης που αντιστοιχεί σε μέγεθος μεγαλύτερο αυτού διαγράφει έντονα φθίνουσα πορεία, μετά το πέρας των 50 εβδομάδων περίπου, στην περίπτωση που η διαφοροποίηση του όγκου ανήκει στην ομάδα ‘III-Αδιαφοροποίητο’. Απότομη πτώση φαίνεται να διαγράφει και η ομάδα μεγέθους 2-5cm, στο τέλος της έρευνας, και να φτάνει στα επίπεδα επιβίωσης της ομάδας ‘>5cm’.

Αναφορικά με το πλήθος των θετικών λεμφαδένων, στις γυναίκες που βρέθηκαν περισσότερα των τεσσάρων σημειώνεται έντονα φθίνουσα πιθανότητα επιβίωσης, συγκριτικά με τις γυναίκες στις οποίες σημειώθηκαν έως τρεις θετικοί λεμφαδένες, , στην περίπτωση που η διαφοροποίηση του όγκου ανήκει στην ομάδα ‘III-Αδιαφοροποίητο’. Η γραμμή που αντιστοιχεί στην πρώτη ομάδα (0-3) βρίσκεται κάτω από αυτή της δεύτερης (>4) καθ’ όλη τη

διάρκεια της έρευνας, , ενώ μετά την 40η εβδομάδα περίπου, η πιθανότητα επιβίωσης της ομάδας με 0-3 θετικούς λεμφαδένες φαίνεται να σταθεροποιείται.

Τέλος, οι ασθενείς που δεν υπεβλήθησαν σε ορμονοθεραπεία και ανήκουν σε βαθμό διαφοροποίησης I και II, ενώ αρχικά φαίνεται να διατηρούν υψηλή πιθανότητα επιβίωσης, και μάλιστα υψηλότερη των γυναικών που δεν υπεβλήθησαν σε αντίστοιχη θεραπεία, μετά την 50η περίπου βδομάδα σημειώνουν έντονα πτωτική πορεία. Σε αντίθεση, οι γυναίκες που υπεβλήθησαν σε τέτοιου είδους θεραπεία, σημειώνουν πιο σταθερή πορεία ως προς την επιβίωσή τους.

Στην συνέχεια, παρουσιάζονται οι αντίστοιχοι έλεγχοι τάσης για τις μεταβλητές που διαθέτουν το χαρακτηριστικό της διαταξιμότητας, ώστε να επιβεβαιωθούν και στατιστικώς τα αποτελέσματα που προέκυψαν από τις ερμηνείες των παραπάνω γραφημάτων.

	Grade		Chi-Square	df	Sig.
Size	2 III-Undifferentiated	Log Rank	4,467	2	0,107
		Breslow	4,641	2	0,098
		Tarone-Ware	4,756	2	0,093
	Pooled	Log Rank	4,35	1	0,037
		Breslow	3,933	1	0,047
		Tarone-Ware	4,162	1	0,041
Nodes	2 III-Undifferentiated	Log Rank	3,325	1	0,068
		Breslow	2,574	1	0,109
		Tarone-Ware	2,884	1	0,089
	Pooled	Log Rank	5,159	1	0,023
		Breslow	3,979	1	0,046
		Tarone-Ware	4,438	1	0,035

Πίνακας 7.2: Έλεγχοι τάσης των παραγόντων Size και Nodes ως προς στρωματοποίηση του Grade.

Θεωρώντας ως επίπεδο σημαντικότητας το 10%, απορρίπτεται η μηδενική υπόθεση όσον αφορά τους παράγοντες Size και Nodes. Συμπερασματικά, καταλήγουμε ότι καθώς αυξάνεται το μέγεθος του όγκου που αφαιρέθηκε κατά την χειρουργική επέμβαση ή το πλήθος των θετικών λεμφαδένων που εντοπίστηκαν, μειώνεται η πιθανότητα επιβίωσης των ασθενών, όταν ο βαθμός διαφοροποίησης του όγκου ανήκει στη κατηγορία III – Αδιαφοροποίητο. Παρατηρούμε ότι στα ίδια αποτελέσματα καταλήγουμε λαμβάνοντας τον ολικό έλεγχο, και μάλιστα θεωρώντας επίπεδο σημαντικότητας 5%.

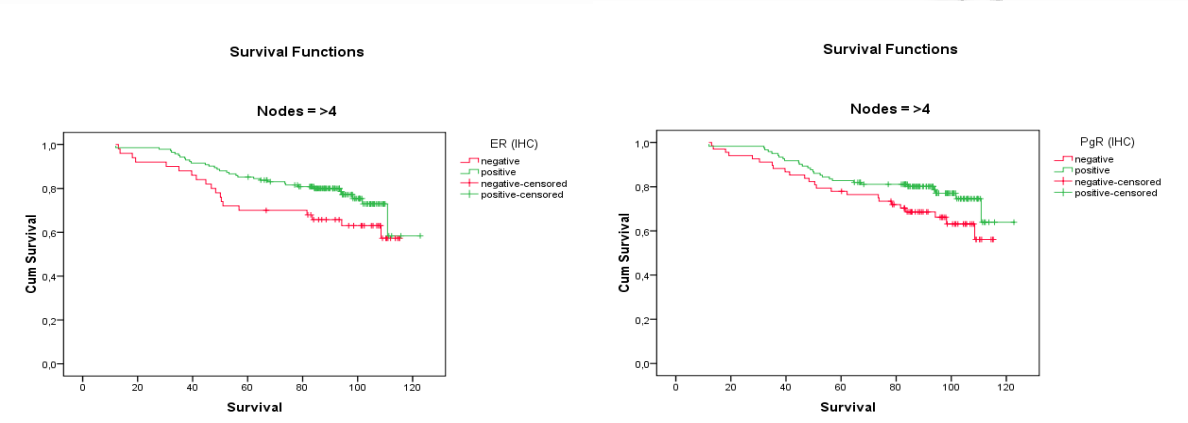
Παράγοντας	Επίπεδο Παράγοντα Nodes	Αυτόνομοι Έλεγχοι				Ολικοί Έλεγχοι			
		Έλεγχος	Chi-Square	df	Sig.	Έλεγχος	Chi-Square	df	Sig.
	0-3	Log Rank	0,939	1	0,333				
ER	**>4**	Log Rank	3,019	1	0,082	Log Rank	1,424	1	0,233
	0-3	Log Rank	0,762	1	0,383				
PgR	**>4**	Log Rank	3	1	0,083	Log Rank	1,636	1	0,201
	0-3	Log Rank	0,449	1	0,503	Log Rank	0,287	1	0,592
Group	**>4**	Wilcoxon	0,924	1	0,336	Wilcoxon	0,72	1	0,396
	0-3	Log Rank	0,662	1	0,416				
Age	**>4**	Log Rank	1,176	1	0,278	Log Rank	1,654	1	0,198
	0-3	Log Rank	0,348	1	0,555	Log Rank	0,181	1	0,67
Menopausal	**>4**	Wilcoxon	0,002	1	0,964	Wilcoxon	0	1	0,982
	0-3	Wilcoxon	0,181	1	0,67				
Surgery	**>4**	Wilcoxon	0,129	1	0,719	Wilcoxon	0,09	1	0,764
	0-3	Log Rank	3,033	2	0,219	Log Rank	0,618	2	0,734
Interval	**>4**	Wilcoxon	1,127	2	0,569	Wilcoxon	0,709	2	0,702
	0-3	Log Rank	0,49	1	0,484	Log Rank	5,397	1	0,02
Grade	**>4**	Log Rank	4,93	1	0,026				
	0-3	Wilcoxon	0,204	2	0,903				
Size	**>4**	Wilcoxon	6,255	2	0,044	Wilcoxon	5,863	2	0,053
	0-3	Wilcoxon	1,739	1	0,187				
RT	**>4**	Wilcoxon	0,297	1	0,586	Wilcoxon	0,038	1	0,846
	0-3	Log Rank	0,046	1	0,83	Log Rank	3,898	1	0,048
HT	**>4**	Wilcoxon	3,572	1	0,059	Wilcoxon	3,535	1	0,06

Πίνακας 7.3: Στρωματοποίηση ως προς τον Παράγοντα Nodes.

Ο δεύτερος παράγοντας που φαίνεται να σημειώνει διαφοροποίηση των συναρτήσεων επιβίωσης στα επίπεδά του είναι αυτός του πλήθους των θετικών λεμφαδένων. Σύμφωνα με τον ‘αυτόνομο’ έλεγχο, και σε επίπεδο σημαντικότητας 5%, προκύπτει ότι οι παράγοντες Grade, Size και HT (οριακά) σημειώνουν διαφορά, για τις ασθενείς με πλήθος θετικών λεμφαδένων μεγαλύτερο του 4. Εάν θεωρήσουμε επίπεδο σημαντικότητας 10%, τότε και οι παράγοντες ER και PgR προκύπτουν σημαντικοί, για το ίδιο στρώμα ασθενών.

Οι ολικοί έλεγχοι, επιβεβαιώνουν ότι έστω μία των συναρτήσεων διαφέρει από τις υπόλοιπες της αντίστοιχης ομάδας ασθενών, όσον αφορά τους παράγοντες Grade, Size και RT, σε επίπεδο σημαντικότητας 5%. Ωστόσο, αδυνατούν να καταγράψουν τις διαφορές στους παράγοντες ER και PgR.

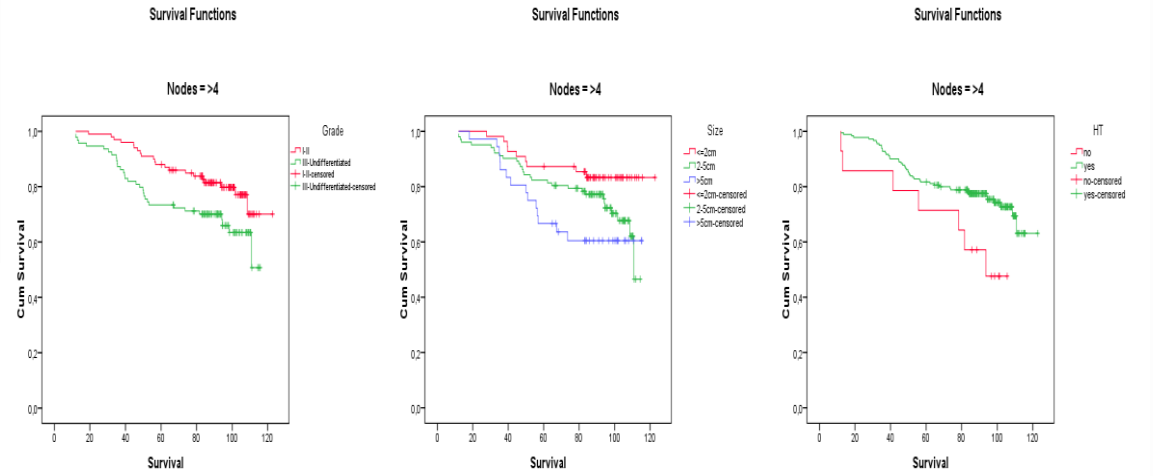
Στην συνέχεια παρουσιάζονται τα διαγράμματα που διαγράφουν την πορεία των συναρτήσεων επιβίωσης στον χρόνο για τις πέντε παραπάνω περιπτώσεις.



Σχήμα 7.6: Διάγραμμα επιβίωσης για στρώμα '>4' του παράγοντα Nodes για ER και PgR.

Λαμβάνοντας υπόψη τον παράγοντα ER για τις γυναίκες στις οποίες βρέθηκαν άνω των τεσσάρων θετικοί λεμφαδένες, φαίνεται ότι στις γυναίκες με θετικό ER αντιστοιχεί υψηλότερη πιθανότητα επιβίωσης, καθ' όλη την διάρκεια της έρευνας. Ωστόσο, αξίζει να επισημάνουμε ότι οι δύο γραμμές καταλήγουν στα ίδια επίπεδα επιβίωσης, με την κόκκινη γραμμή των αρνητικών ER ασθενών να πλησιάζει νωρίτερα.

Ως προς τον παράγοντα PgR, παρατηρείται η ίδια συμπεριφορά με αυτή του ER, με την διαφοροποίηση ότι η διαφορά των δύο γραμμών είναι σταθερή κατά τις πρώτες 70 εβδομάδες περίπου. Στην συνέχεια, η διαφορά αυτή γίνεται εντονότερη καθώς η γραμμή των ασθενών που είναι θετικές στο PgR διαγράφουν πιο σταθερή πορεία.



Σχήμα 7.7: Διάγραμμα επιβίωσης για στρώμα '>4' του παράγοντα Nodes για Grade, Size και HT.

Για τις γυναίκες με πλήθος θετικών λεμφαδένων μεγαλύτερο του 4, διαφορά των συναρτήσεων επιβίωσης σημειώνεται και ως προς τον βαθμό διαφοροποίησης των όγκων. Η κατηγορία με βαθμό I-II διατηρεί υψηλότερη πιθανότητα επιβίωσης, σε όλη την διάρκεια της έρευνας.

Αντίστοιχα, η επιβίωση των γυναικών με περισσότερων των τεσσάρων θετικών λεμφαδένων, φαίνεται να είναι μικρότερη για την κατηγορία των γυναικών από τις οποίες ο όγκος που αφαιρέθηκε ήταν μεγαλύτερος των 5cm. Οι γυναίκες με μέγεθος όγκου μικρότερο των 2cm διατηρούν την ψηλότερη πιθανότητα επιβίωσης, ενώ οι γυναίκες με μέγεθος μεταξύ 2 και 5cm σημειώνουν έντονα πτωτική πορεία μετά το πέρας της 90^{ης} εβδομάδας περίπου.

Τέλος, οι γυναίκες που υπεβλήθησαν σε ορμονοθεραπεία, σημειώνουν υψηλότερη πιθανότητα να επιβιώσουν, καθώς η αντίστοιχη γραμμή βρίσκεται συνεχώς πάνω αυτής που αντιστοιχεί στις γυναίκες που δεν έλαβαν τέτοιου είδους θεραπεία. Η τελευταία, μάλιστα, σημειώνει έντονα καθοδική πορεία και καταλήγει σε πολύ χαμηλότερα επίπεδα επιβίωσης.

	Nodes		Chi-Square	df	Sig.
Grade	1 >4	Log Rank	4,93	1	0,026
		Breslow	5,756	1	0,016
		Tarone-Ware	5,405	1	0,02
	Pooled[a]	Log Rank	5,397	1	0,02
		Breslow	6,2	1	0,013
		Tarone-Ware	5,943	1	0,015
Size	1 >4	Log Rank	6,074	2	0,048
		Breslow	6,255	2	0,044
		Tarone-Ware	6,234	2	0,044
	Pooled[a]	Log Rank	4,586	1	0,032
		Breslow	5,596	1	0,018
		Tarone-Ware	5,312	1	0,021

Πίνακας 7.4: Έλεγχοι τάσης των παραγόντων Grade και Size ως προς στρωματοποίηση του Nodes.

Τέλος, σύμφωνα με τους ελέγχους τάσης τόσο για τα συγκεκριμένα στρώματα των παραγόντων Grade και Size, οι οποίοι είναι διατάξιμοι, όσο και από τους ολικούς ελέγχους, και θεωρώντας επίπεδο σημαντικότητας 5%, επιβεβαιώνονται και στατιστικώς τα παραπάνω συμπεράσματα. Πιο συγκεκριμένα, προκύπτει ότι όσο μεγαλύτερος ο βαθμός διαφοροποίησης του όγκου ή το μέγεθος του αφαιρούμενου όγκου, τόσο μικρότερη η πιθανότητα επιβίωσης της ασθενούς.

7.3 Μοντέλο αναλογικού κινδύνου του χρόνου επιβίωσης

Στα δεδομένα που εξετάζονται στην παρούσα ανάλυση περιέχονται συνολικά 21 μεταβλητές που εξετάζονται ως προς την συμβολή τους στην βιωσιμότητα των ασθενών με καρκίνο του μαστού. Από τις 21 αυτές μεταβλητές, οι 12 παράγοντες αναφέρονται σε κλινικοπαθολογικά χαρακτηριστικά των ασθενών και οι 9 αποτελούν τους παράγοντες που αντιπροσωπεύουν τις γονιδιακές εκφράσεις. Δεδομένου, λοιπόν, του μεγάλου πλήθους μεταβλητών που πρέπει να εξετασθούν, η ανάλυση σύμφωνα με τον εκτιμητή *Kaplan-Meier*, γίνεται αρκετά περίπλοκη, όταν στο μοντέλο πρέπει να περιλαμβάνονται περισσότερες των δύο μεταβλητών. Για τον λόγο αυτό καταφεύγουμε στην μοντελοποίηση της συνάρτησης κινδύνου, μέσω παλινδρόμησης. Συγκεκριμένα, γίνεται χρήση του *μοντέλου αναλογικού κινδύνο*υ (proportional hazard model - PH) ή *μοντέλο παλινδρόμησης του Cox*.

7.3.1 Αναζήτηση μοντέλου

Αρχικά παρουσιάζονται τα αποτελέσματα της εφαρμογής του παραπάνω μοντέλου, θεωρώντας τις κύριες επιδράσεις των διαφόρων επιπέδων των παραγόντων, με την μέθοδο *Breslow*. Στο Παράρτημα Α.6 παρατίθεται ο πίνακα των βωβών μεταβλητών που απαιτούνται και θεωρούμε ότι η ΑΣΚ αποτελεί το $1^{ο}$ επίπεδο, για τους κατηγορικούς παράγοντες.

Εφαρμόζοντας Backward Model Selection θεωρώντας τις παραπάνω μεταβλητές και τους παράγοντες που αντιπροσωπεύουν τις γονιδιακές εκφράσεις, κατά το τελευταίο βήμα της διαδικασίας ($17^{ο}$) προκύπτουν σημαντικοί οι παράγοντες ‘Nodes’, ‘Factor1’, ‘Factor2’, ‘Factor5’ και ‘Factor7’. Στον ακόλουθο πίνακα παρουσιάζονται οι εκτιμήσεις των αντίστοιχων παραμέτρων.

		B	SE	Wald	df	Sig.	Exp(B)
Step 17	Nodes	1,047	,406	6,648	1	,010	2,849
	Factor1	-,371	,125	8,874	1	,003	,690
	Factor2	,514	,146	12,428	1	,000	1,672
	Factor5	-,316	,127	6,194	1	,013	,729
	Factor7	,217	,130	2,789	1	,095	1,242

Πίνακας 7.5: Εκτιμήσεις των παραμέτρων του μοντέλου αναλογικού κινδύνου σύμφωνα με τη διαδικασία Backward Selection.

Δεδομένου ότι ο παράγοντας που δηλώνει τη θεραπευτική αγωγή στην οποία υποβλήθηκε η ασθενής (‘Group’) έχει αξία για τον σχεδιασμό της μελέτης, δημιουργούμε το μοντέλο PH στο οποίο περιέχονται οι παραπάνω σημαντικοί παράγοντες συμπεριλαμβανομένης της ‘Group’. Το μοντέλο που προκύπτει έχει ως εξής:

							95,0% CI for Exp(B)	
	B	SE	Wald	df	Sig.	Exp(B)	Lower	Upper
Group	,057	,259	,049	1	,825	1,059	,638	1,758
Nodes	1,016	,384	7,008	1	,008	2,762	1,302	5,860
Factor1	-,375	,122	9,438	1	,002	,687	,541	,873
Factor2	,482	,141	11,722	1	,001	1,619	1,229	2,134
Factor5	-,283	,125	5,127	1	,024	,753	,590	,963
Factor7	,264	,126	4,374	1	,037	1,302	1,017	1,667

Πίνακας 7.6: Εκτιμήσεις των παραμέτρων του τελικού μοντέλου αναλογικού κινδύνου.

.

Σε αυτό το σημείο πρέπει να αναφέρουμε ότι από βιολογικής σκοπιάς, δεν παρουσιάζουν ενδιαφέρον οι αλληλεπιδράσεις δεύτερης ή μεγαλύτερης τάξης, παρά μόνον αυτές της μεταβλητής ‘Group’ με τα γονίδια. Εφαρμόζοντας Backward Selection στο μοντέλο που περιέχει τους παραπάνω σημαντικούς παράγοντες κλινικών χαρακτηριστικών και τους παράγοντες των γονιδίων με τις αντίστοιχες αλληλεπιδράσεις αυτών με την Group, προκύπτουν ως σημαντικές οι αλληλεπιδράσεις των παραγόντων ‘Factor4’ και ‘Factor7’, σε επίπεδο σημαντικότητας 5%. Λαμβάνοντας το αντίστοιχο μοντέλο PH προκύπτει ότι καμία από τις αλληλεπιδράσεις δεν είναι στατιστικώς σημαντική, οπότε καταλήγουμε στο ακόλουθο μοντέλο, σύμφωνα με τον Πίνακα 7.7.

$$h(t) = h_0(t)\exp(0{,}057\cdot \text{Group} + 1{,}016\cdot \text{Nodes} - 0{,}375\cdot \text{Factor1} + 0{,}482\cdot \text{Factor2} - 0{,}283\cdot \text{Factor5} + 0{,}264\cdot \text{Factor7})$$

Προκειμένου να εξετασθεί η σταθερότητα του παραπάνω μοντέλου, ακολουθείται η ακόλουθη διαδικασία. Αρχικά, εφαρμόζουμε το μοντέλο PH μόνο για τους παράγοντες που αποτελούν τα κλινικά χαρακτηριστικά των ασθενών, και ξεχωριστά το μοντέλο PH για τους παράγοντες των γονιδίων, οι σημαντικοί παράγοντες από το πρώτο είναι οι ‘Grade’, ‘Nodes’ και ‘HT’, ενώ από το δεύτερο οι ‘Factor1’, ‘Factor2’, ‘Factor5’ και ‘Factor7’. Τα αποτελέσματα της διαδικασίας επιλογής παρουσιάζονται στους πίνακες του Παραρτήματος

Α.7. Τέλος, εφαρμόζουμε Backward Selection στο σύνολο των παραπάνω σημαντικών παραγόντων και καταλήγουμε στο ίδιο μοντέλο, εάν συμπεριλάβουμε, επιπλέον, και τον παράγοντα της θεραπείας, όπως παραπάνω. Συνεπώς, το μοντέλο στο οποίο καταλήξαμε φαίνεται να χαρακτηρίζεται από σταθερότητα.

Από τους ολικούς ελέγχους *score* και *log-rank*, ελέγχεται η μηδενική υπόθεση που ορίζει ότι οι συντελεστές των παραγόντων ισούνται με μηδέν, με αποτέλεσμα κανένας να μην επηρεάζει την συνάρτηση κινδύνου, έναντι της εναλλακτικής ότι τουλάχιστον ένας από αυτούς τους παράγοντες είναι σημαντικός.

	Overall (score)			Change From Previous Step			Change From Previous Block		
-2 Log Likelihood	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
609,589	32,090	6	,000	33,922	6	,000	33,922	6	,000

Πίνακας 7.7: Έλεγχοι *score* και *log-rank*.

Λαμβάνοντας τους δύο παραπάνω ελέγχους, καταλήγουμε στα στατιστικά X^2= 32,090 και X^2=33,922, αντίστοιχα, με 6 β.ε. για καθένα από αυτά. Σε επίπεδο σημαντικότητας 5%, καταλήγουμε και με τους δύο ελέγχους σε κοινό συμπέρασμα, ότι θα πρέπει να απορριφθεί η μηδενική υπόθεση και συνεπώς να δεχθούμε ότι η συνάρτηση κινδύνου εξαρτάται τουλάχιστον από μία από τις υπό εξέταση μεταβλητές. Από τους τοπικούς ελέγχους που παρέχονται από το Wald test όπως παρουσιάζονται στον Πίνακα 7.7, παρατηρούμε ότι σε επίπεδο σημαντικότητας 5%, η μηδενική υπόθεση απορρίπτεται για το σύνολο των παραγόντων εκτός της Group, η οποία, όμως, δεν μπορεί να εξαχθεί από το μοντέλο.

7.3.2 Ερμηνεία μοντέλου

Η ΑΣΚ του μοντέλου που μελετάμε αντιστοιχεί σε ασθενείς οι οποίες ανήκουν στην θεραπευτική αγωγή E-T-CMF, έχουν αφαιρέσει 0-3 μασχαλιαίους λεμφαδένες και στους παράγοντες που αντιστοιχούν στις ομάδες των γονιδίων έχουν μηδενικές τιμές.

Από την παραπάνω ανάλυση προκύπτει ότι $\hat{\beta}_1$=0,057 και $s\hat{e}\left(\hat{\beta}_1\right)$=0,259. Ένα προσεγγιστικό διάστημα εμπιστοσύνης (CI) για την παράμετρο β_1 με επίπεδο σημαντικότητας 5% είναι το (-0.449, 0.564). Ο ΕΜΠ για τον λόγο κινδύνου ως προς τον παράγοντα Group δεν έχει νόημα να ερμηνευθεί, δεδομένου ότι η αντίστοιχη παράμετρος δεν προκύπτει σημαντική σύμφωνα με το Wald test. Επιπλέον, τα διαστήματα εμπιστοσύνης των $\hat{\beta}_1$ και

$exp(\hat{\beta}_1)$ περιέχουν τις τιμές 0 και 1, αντίστοιχα, οπότε έχουμε δύο επιπλέον ενδείξεις ότι δεν υπάρχει σημαντική διαφοροποίηση μεταξύ των δύο θεραπευτικών αγωγών.

Για τη μεταβλητή του πλήθους των λεμφαδένων προκύπτει $\hat{\beta}_2$=1,016 και $s\hat{e}(\hat{\beta}_2)$=0,259. Το θετικό πρόσημο του ΕΜΠ $\hat{\beta}_2$ δηλώνει ότι μία ασθενής με πλήθος αφαιρούμενων θετικών λεμφαδένων μεγαλύτερου των τεσσάρων διατρέχει μεγαλύτερο κίνδυνο να πεθάνει σε σχέση με μία ασθενή με αντίστοιχο πλήθος μικρότερο των τεσσάρων. Ένα προσεγγιστικό διάστημα εμπιστοσύνης με συντελεστή εμπιστοσύνης 95% για την παράμετρο $\hat{\beta}_2$ είναι το (0.264, 1.768). Ο λόγος κινδύνου, όταν οι υπόλοιποι παράγοντες παραμένουν σταθεροί, προκύπτει ως

$$\hat{HR}(t) = \frac{h(t \mid Group, Nodes = 1, Factor1, Factor2, Factor5, Factor7)}{h(t \mid Group, Nodes = 0, Factor1, Factor2, Factor5, Factor7)} =$$

$$\frac{h_0(t)exp(\beta_1 + \beta_2 \cdot 1 + \beta_3 + \beta_4 + \beta_5 + \beta_6)}{h_0(t)exp(\beta_1 + \beta_2 \cdot 0 + \beta_3 + \beta_4 + \beta_5 + \beta_6)} = e^{\hat{\beta}_2} = 2{,}762,$$

οπότε συμπεραίνουμε ότι μία ασθενής στην οποία αφαιρέθηκαν περισσότεροι των τεσσάρων θετικών λεμφαδένων έχει 2,76 φορές μεγαλύτερο κίνδυνο να πεθάνει εν συγκρίσει με μία ασθενή στην οποία αφαιρέθηκαν έως τρεις θετικοί λεμφαδένες, όταν οι υπόλοιποι παράγοντες δεν διαφοροποιούνται. Ένα προσεγγιστικό διάστημα εμπιστοσύνης για τον λόγο κινδύνου με συντελεστή εμπιστοσύνης 95% είναι το (1.302, 5.860). Επειδή το διάστημα αυτό δεν περιέχει τη μονάδα, είναι ένδειξη ότι το πλήθος των αφαιρούμενων θετικών λεμφαδένων είναι σημαντικός προγνωστικός παράγοντας.

Όσον αφορά τους παράγοντες που αναφέρονται στις ομάδες των γονιδίων, παρατηρούμε ότι οι ΕΜΠ $\hat{\beta}_3$ $\left(-0.375, s\hat{e}(\hat{\beta}_3) = 0.122\right)$ και $\hat{\beta}_5$ $\left(-0.283, s\hat{e}(\hat{\beta}_5) = 0.125\right)$ των 'Factor1' και 'Factor5', οι οποίοι αντιπροσωπεύουν τις ομάδες 1 και 3 αντίστοιχα, έχουν αρνητικό πρόσημο που σημαίνει ότι όσο αυξάνεται η τιμή των παραγόντων αυτών, μειώνεται ο κίνδυνος θανάτου για τις ασθενείς. Πιο συγκεκριμένα, ο λόγος κινδύνου για μία ασθενή με τιμή *x*+1 του 'Factor1', ως προς μια ασθενή με αντίστοιχη τιμή *x* ορίζεται ως,

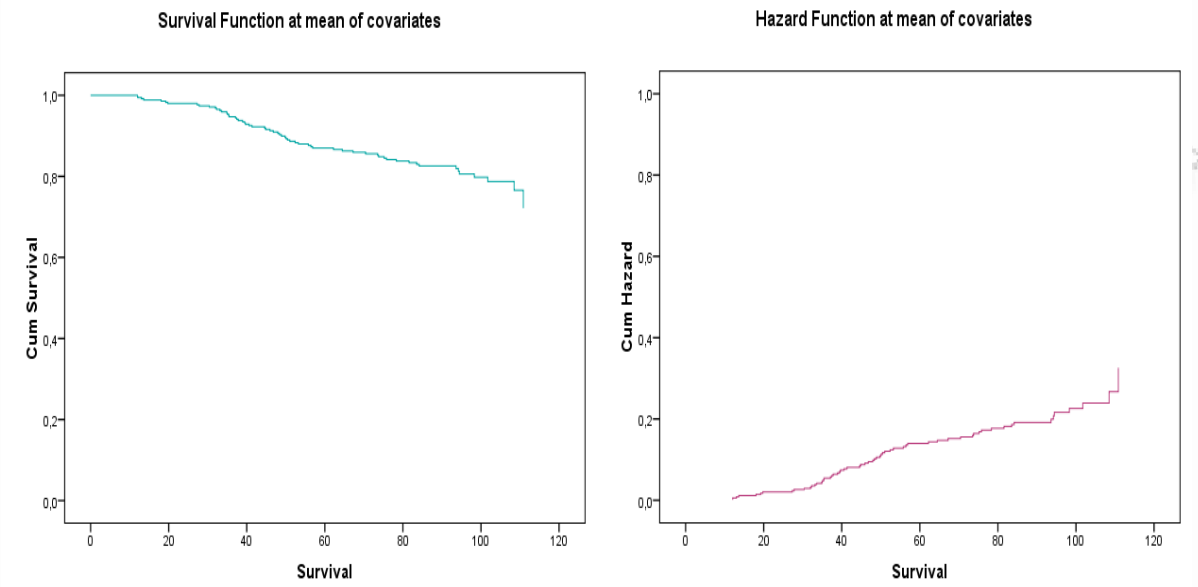
$$\hat{HR}(t) = \frac{h(t \mid Group, Nodes, Factor1 \cdot (x+1), Factor2, Factor5, Factor7)}{h(t \mid Group, Nodes, Factor1 \cdot x, Factor2, Factor5, Factor7)} =$$

$$\frac{h_0(t)exp(\beta_1+\beta_2+\beta_3(x+1)+\beta_4+\beta_5+\beta_6)}{h_0(t)exp(\beta_1+\beta_2+\beta_3 x+\beta_4+\beta_5+\beta_6)}=e^{\hat{\beta}_3}=0{,}687.$$

Η εκτιμώμενη μεταβολή στο λόγο κινδύνου όταν η τιμή του 'Factor1' αυξηθεί κατά μία μονάδα είναι 0,687, και συμπεραίνουμε ότι μία γυναίκα που σημειώνει τιμή του 'Factor1' αυξημένη κατά μία μονάδα έχει σχεδόν 1,5 φορές (1/0,687 = 1,456) χαμηλότερο κίνδυνο θανάτου, όταν οι υπόλοιπες μεταβλητές του μοντέλου παραμένουν σταθερές. Ένα 95% προσεγγιστικό διάστημα εμπιστοσύνης για τον λόγο κινδύνου είναι το (0.541, 0.873), το οποίο δεν περιλαμβάνει την μονάδα. Ομοίως, θεωρώντας τον 'Factor5', προκύπτει ότι ο λόγος κινδύνου για μία ασθενή με τιμή x+1 του 'Factor5', ως προς μια ασθενή με αντίστοιχη τιμή x είναι 0,753, οπότε μία γυναίκα με αύξηση του 'Factor5' κατά 1, έχει 1,3 φορές χαμηλότερο κίνδυνο να πεθάνει, δεδομένου ότι οι υπόλοιπες μεταβλητές διατηρούνται σταθερές.

Τέλος, οι 'Factor2' και 'Factor7' αντιστοιχούν στη 2^{η} ομάδα των γονιδιακών εκφράσεων και τα θετικά τους πρόσημα $\left(\hat{\beta}_4=0.482, s\hat{e}\left(\hat{\beta}_4\right)=0.141, \hat{\beta}_6=0.264, s\hat{e}\left(\hat{\beta}_6\right)=0.126\right)$ δηλώνουν ότι αύξηση της τιμής τους συνεπάγεται αύξηση και του κινδύνου θανάτου της γυναίκας. Ο λόγος κινδύνου που εκτιμάται από το παραπάνω μοντέλο, όταν στην ασθενή σημειώνεται η τιμή του 'Factor2' αυξημένη κατά μία μονάδα έναντι μιας γυναίκας που η τιμή του παράγοντα δεν έχει την μοναδιαία αύξηση, και οι υπόλοιπες μεταβλητές παραμένουν σταθερές είναι 1,6, με αντίστοιχο 95% CI (1.229, 2.134). Για τον 'Factor7' ο αντίστοιχος λόγος κινδύνου είναι 1,3 και ένα 95% CI (1.017, 1.667). Δεδομένου ότι οι δύο αυτοί παράγοντες αντιπροσωπεύουν από κοινού την 2^{η} ομάδα των γονιδίων, αξίζει να εξετάσουμε την περίπτωση του λόγου κινδύνου όταν και οι δύο παράγοντες αυξάνονται κατά μία μονάδα και οι υπόλοιπες μεταβλητές παραμένουν σταθερές. Ο λόγος κινδύνου σε αυτή τη περίπτωση είναι 2,109 και δηλώνει ότι η ταυτόχρονη μοναδιαία αύξηση των 'Factor2' και 'Factor7' κατά μία μονάδα αντιστοιχεί σε 2,1 φορές υψηλότερο κίνδυνο θνησιμότητας, δεδομένου ότι οι υπόλοιπες μεταβλητές διατηρούνται σταθερές.

Στη συνέχεια παρατίθενται τα διαγράμματα επιβίωσης και κινδύνου, αντίστοιχα, στο μέσο των συμμεταβλητών.



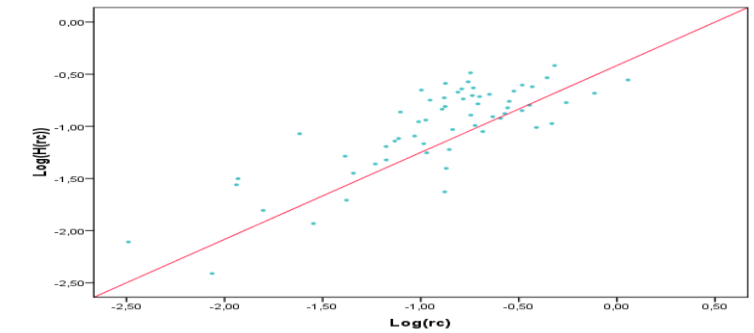
Σχήμα 7.8: Διαγράμματα των συναρτήσεων επιβίωσης και κινδύνου.

Κατά τις πρώτες 25 περίπου εβδομάδες, η πιθανότητα επιβίωσης διατηρείται σε υψηλά επίπεδα και πέραν αυτής της χρονικής στιγμής σημειώνει περισσότερο απότομη μείωση. Επιπλέον, η συνάρτηση επιβίωσης δεν φαίνεται να ξεπερνά την πιθανότητα 0,70, αφού στο σημείο αυτό σημειώνεται ο τελευταίος παρατηρούμενος θάνατος, γεγονός που σημαίνει ότι η πιθανότητα επιβίωσης διατηρείται σε υψηλά επίπεδα. Γενικά, μπορούμε να πούμε ότι η επιβίωση δεν διαγράφει έντονα φθίνουσα πορεία.

7.3.3 Ανάλυση υπολοίπων

Στην συνέχεια γίνεται η ανάλυση υπολοίπων του *Cox* μοντέλου, όπως αυτά παρουσιάστηκαν στην παράγραφο 6.10. Από την ερμηνεία των παρακάτω υπολοίπων καταλήγουμε σε ουσιώδη συμπεράσματα σχετικά με την ολική επάρκεια του μοντέλου, τον έλεγχο της υπόθεσης αναλογικού κινδύνου, την ανίχνευση έκτροπων παρατηρήσεων (outliers) και την επιρροή κάθε παρατήρησης στην εκτίμηση των ΕΜΠ κάθε μεταβλητής και την εύρεση της συναρτησιακής μορφής κάθε μεταβλητής που εισάγεται στο μοντέλο.

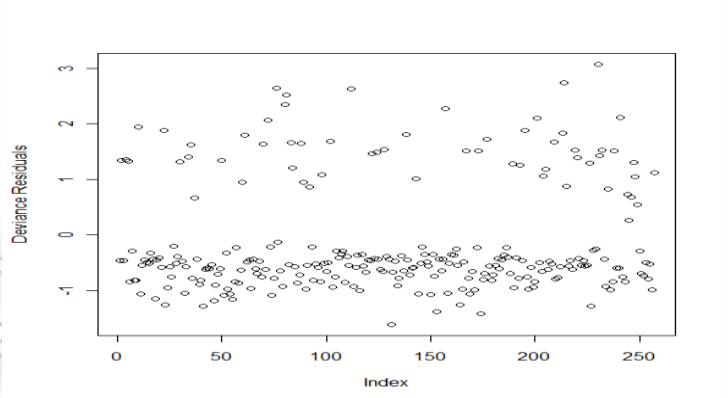
Cox-Snell Υπόλοιπα



Σχήμα 7.9: Διάγραμμα *Cox-Snell* υπολοίπων.

Τα υπόλοιπα *Cox-Snell* εφαρμόζονται για τον *έλεγχο της ολικής επάρκειας του μοντέλου αναλογικού κινδύνου* (check overall (global) model fit). Από το παραπάνω γράφημα φαίνεται τα σημεία να 'ακολουθούν' την ευθεία *y=x* αρκετά ικανοποιητικά με κάποια να απέχουν αισθητά από την ευθεία. Σε γενικές γραμμές, μπορούμε να πούμε ότι η ολική επάρκεια του μοντέλου είναι ικανοποιητική, και συνεπώς το μοντέλο PH προσαρμόζεται ικανοποιητικά στα δεδομένα.

Υπόλοιπα Deviance

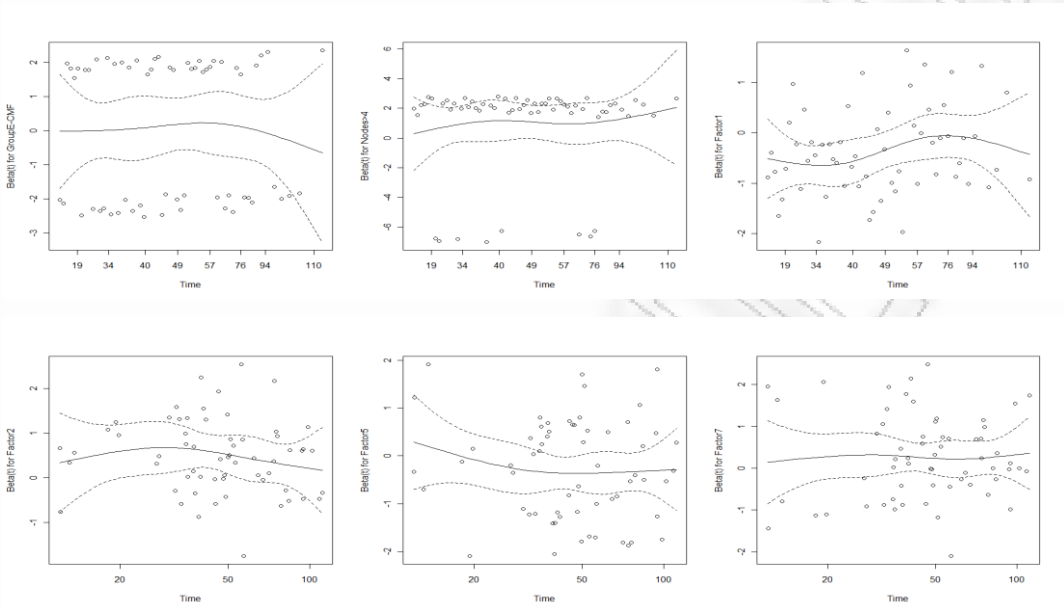


Σχήμα 7.10: Διάγραμμα *Deviance* υπολοίπων.

Από το διάγραμμα των *Deviance* υπολοίπων, οι τιμές που είναι κοντά στο μηδέν αντιστοιχούν σε λογοκριμένους χρόνους ζωής, ενώ αυτές που αντιστοιχούν σε πλήρεις χρόνους απομακρύνονται από το μηδέν και εκτείνονται έως την τιμή 3, για το σύνολο των δεδομένων που εξετάζουμε. Τα σημεία που αντιστοιχούν σε πλήρεις χρόνους διασκορπίζονται τυχαία στο χώρο και υπάρχει ένα μικρό πλήθος αυτών (τρεις παρατηρήσεις) οι οποίες λαμβάνουν θετική τιμή αλλά πολύ κοντά στο μηδέν, με αποτέλεσμα να μην είναι

σαφές σε ποια κατηγορία χρόνων ανήκουν και να ‘ξεφεύγουν’ από το ‘σύννεφο’ που δημιουργείται από τις υπόλοιπες παρατηρήσεις.

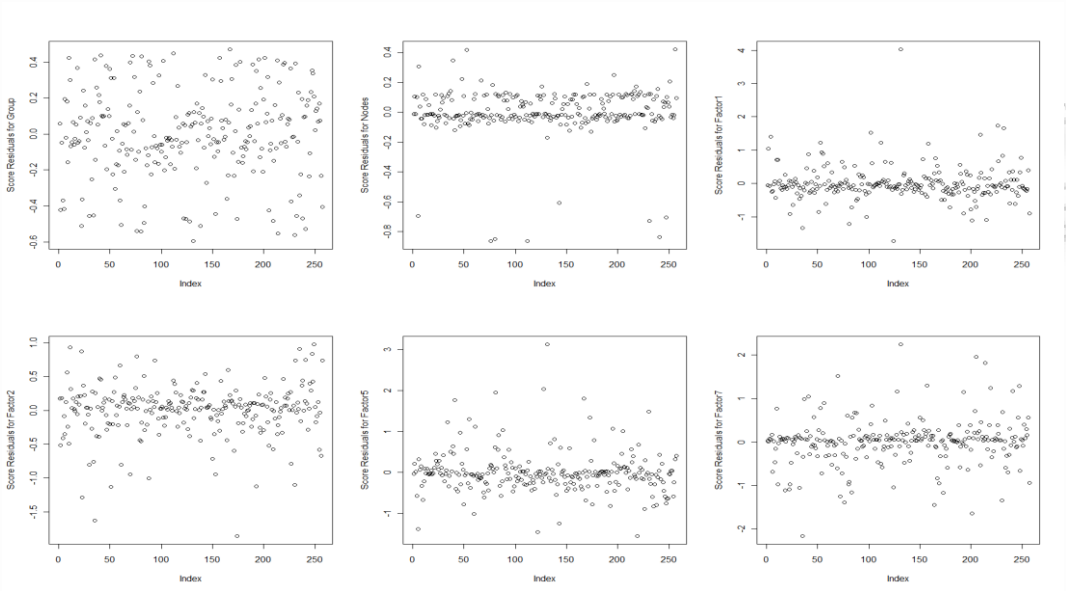
Schoenfeld Υπόλοιπα



Σχήμα 7.11: Διαγράμματα *Schoenfeld* υπολοίπων για κάθε παράγοντα.

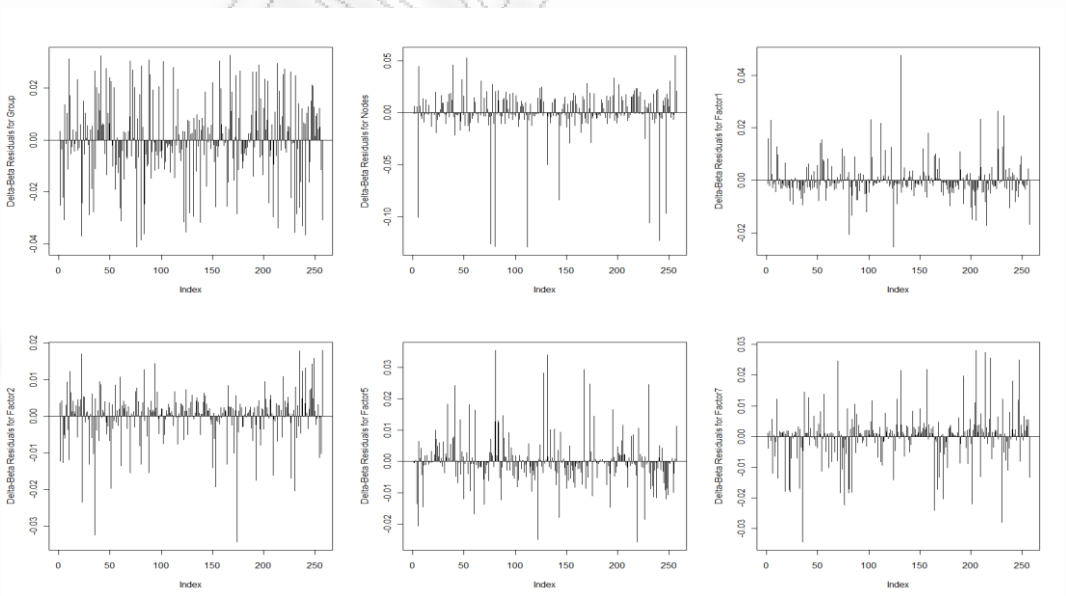
Από τα *rescaled Schoenfeld* υπόλοιπα που παρουσιάζονται στα παραπάνω διαγράμματα σημείων για κάθε παράγοντα χωριστά, φαίνεται να ισχύει η μηδενική κλίση ως προς την ευθεία $g(t) = t$. Το αποτέλεσμα αυτό επιβεβαιώνεται και μέσω των ελέγχων μηδενικής κλίσης της smoothing curve. Όπως παρατηρούμε στο πίνακα του Παραρτήματος Α.8, τόσο οι ατομικοί έλεγχοι όσο και ο ολικός καταλήγουν στο συμπέρασμα ότι σε επίπεδο σημαντικότητας 5% δεν απορρίπτουν την μηδενική υπόθεση. Συνεπώς, για κάθε μεταβλητή ισχύει η υπόθεση αναλογικού κινδύνου, όπως επίσης, και για το τελικό μοντέλο.

Score και scaled score (Δέλτα-Βήτα) Υπόλοιπα



Σχήμα 7.12: Διαγράμματα *Score* υπολοίπων για κάθε παράγοντα.

Τα παραπάνω διαγράμματα αντιστοιχούν στα *Score Residuals* για κάθε έναν από τους παράγοντες που αποτελούν το *Cox* μοντέλο. Παρατηρούμε ότι στο σύνολο των διαγραμμάτων οι παρατηρήσεις κατανέμονται τυχαία γύρω από το μηδέν, χωρίς να σημειώνουν όλες την ίδια διασπορά. Το γεγονός αυτό παρέχει μία γραφική ένδειξη της ικανοποίησης της υπόθεσης αναλογικού κινδύνου για το μοντέλο από το οποίο παράγονται τα υπόλοιπα αυτά.



Σχήμα 7.13: Διαγράμματα *scaled score* υπολοίπων για κάθε παράγοντα.

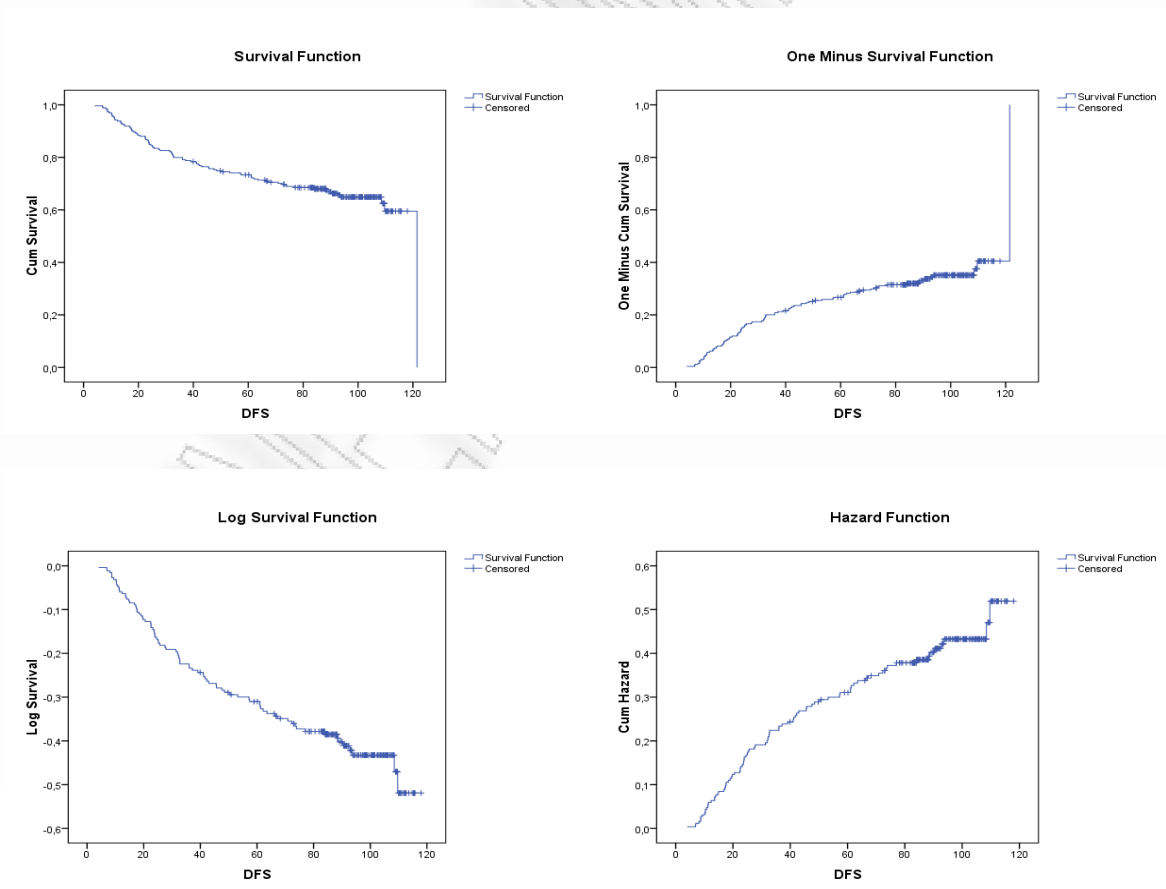
Τέλος, από τα παραπάνω διαγράμματα των υπολοίπων *Δέλτα-Βήτα* εντοπίζεται η επιρροή κάθε ασθενούς στη εκτίμηση των παραμέτρων που απαρτίζουν το μοντέλο του *Cox*. Οι ασθενείς που έχουν τιμές κοντά στο μηδέν ασκούν αμελητέα επιρροή, ενώ οι μεγαλύτερες αποκλίσεις υποδεικνύουν ότι οι αντίστοιχες ασθενείς έχουν μεγάλη επιρροή.

7.4 Μη παραμετρική εκτίμηση του ελεύθερου νόσου χρόνου

7.4.1 Εκτίμηση *Kaplan-Meier* της συνάρτησης επιβίωσης

Όπως και στη περίπτωση του χρόνου επιβίωσης, εξετάζουμε τις ίδιες μεταβλητές θεωρώντας τις κύριες επιδράσεις των διαφόρων επιπέδων των παραγόντων, με την μέθοδο Breslow. Οι βωβές μεταβλητές διαμορφώνονται όπως στη περίπτωση του χρόνου επιβίωσης.

Το ποσοστό λογοκριμένων δεδομένων είναι αρκετά υψηλό αφού ξεπερνά το 65,1% των ασθενών. Στο Παράρτημα Α.9 δίνεται ο πίνακας των εκτιμήσεων *Kaplan-Meier*, και στη συνέχεια παρουσιάζονται τα αντίστοιχα γραφήματα.



Σχήμα 7.14: Διαγράμματα επιβίωσης, 1-S(t), Log(S(t)) και κινδύνου της εκτίμησης Kaplan-Meier για το χρόνο DFS.

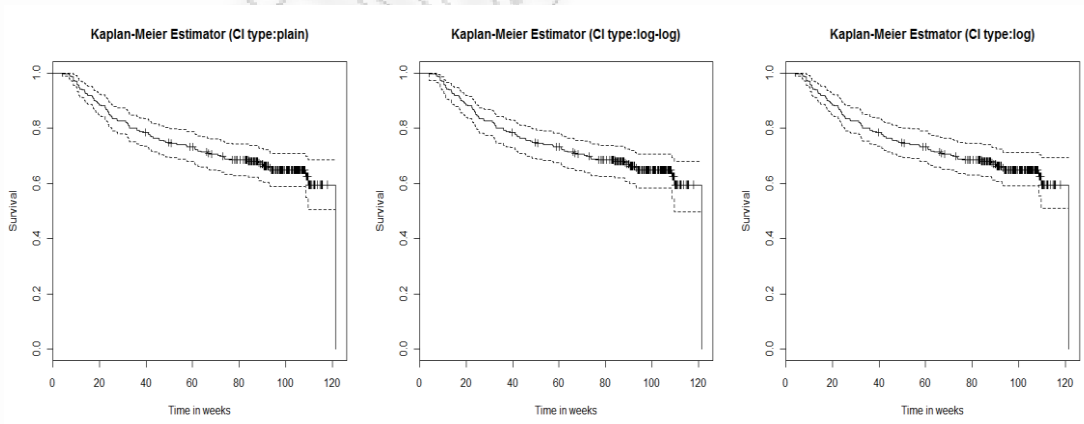
Η καμπύλη που αναπαριστά την συνάρτηση επιβίωσης ακολουθεί έντονα καθοδική πορεία, γεγονός που σημαίνει ότι ο κίνδυνος θανάτου αυξάνεται με γρήγορο ρυθμό. Έως την 72η εβδομάδα, περίπου, αυτή η μείωση είναι εντονότερη σε σχέση με την πορεία της επιβίωσης πέραν αυτού του χρονικού σημείου. Πέραν αυτού, επομένως, η κλίση της συνάρτησης φαίνεται να σταθεροποιείται. Αξιοσημείωτο είναι, επίσης, το γεγονός ότι έως την 50η εβδομάδα σημειώνονται ελάχιστες λογοκριμένες παρατηρήσεις, ενώ μετά του σημείου αυτού εντοπίζεται ο κύριος όγκος λογοκρισίας.

Η εκτιμώμενη μέση τιμή είναι $\hat{\mu} = 91{,}266$ (95% CI: 85,950 – 96,582). Το 25% ποσοστιαίο σημείο $\hat{t}_{0.25} = 48{,}492$ δηλώνει την μικρότερη χρονική στιγμή στην οποία η καμπύλη μειώνεται κάτω του 75%. Το αντίστοιχο 95% διάστημα εμπιστοσύνης ορίζεται ως,

$$\hat{t}_{0,25} \pm z_{a/2}\sqrt{\hat{V}(\hat{t}_{0,25})} = \hat{t}_{0,25} \pm 1{,}96\sqrt{\hat{V}(\hat{t}_{0,25})} = 48{,}492 \pm 1{,}96 \cdot 10{,}074 = 48{,}492 \pm 19{,}745 =$$

$$(28.747\ ,\ 68.237).$$

Η συνολική διάρκεια επιβίωσης είναι προσεγγιστικά 121,5 εβδομάδες, και η αθροιστική αναλογία επιβίωσης δεν φαίνεται να ξεπερνά το 59,4%, που σημαίνει ότι η ελάχιστη πιθανότητα μία ασθενής να επιζήσει πέραν της παραπάνω χρονικής διάρκειας είναι 59,4%. Τέλος, αξίζει να σημειωθεί ότι, δεδομένου αυτού, δεν μπορεί να ορισθεί η διάμεσος και το 75% ποσοστιαίο σημείο, όπως και στην περίπτωση του χρόνου επιβίωσης. Στην συνέχεια παρουσιάζονται τα διαγράμματα της συνάρτησης επιβίωσης, συμπεριλαμβανομένων και των διαστημάτων εμπιστοσύνης τύπου plain, log-log και log, αντίστοιχα.



Σχήμα 7.15: Εκτίμηση *Kaplan-Meier* με διαστήματα εμπιστοσύνης τύπου plain, log-log και log.

7.4.2 Σύγκριση συναρτήσεων επιβίωσης ως προς τα κλινικά χαρακτηριστικά

Στην παρούσα παράγραφο παρουσιάζοντα τα αποτελέσματα των ελέγχων για την ισότητα των συναρτήσεων επιβίωσης, ως προς τα κλινικά χαρακτηριστικά. Οι έλεγχοι που πραγματοποιούνται αναφέρονται στα βάρη όπως αυτά παρουσιάστηκαν στην περίπτωση της ανάλυσης του χρόνου επιβίωσης. Υπενθυμίζεται ότι η μηδενική υπόθεση, στην περίπτωση δίτιμου παράγοντα, έχει την μορφή,

$$H_0: S_1(t) = S_2(t) \text{ έναντι της εναλλακτικής } H_1: S_1 \neq S_2.$$

Για μεταβλητές περισσότερων των δύο κατηγοριών ο έλεγχος διαμορφώνεται ανάλογα.

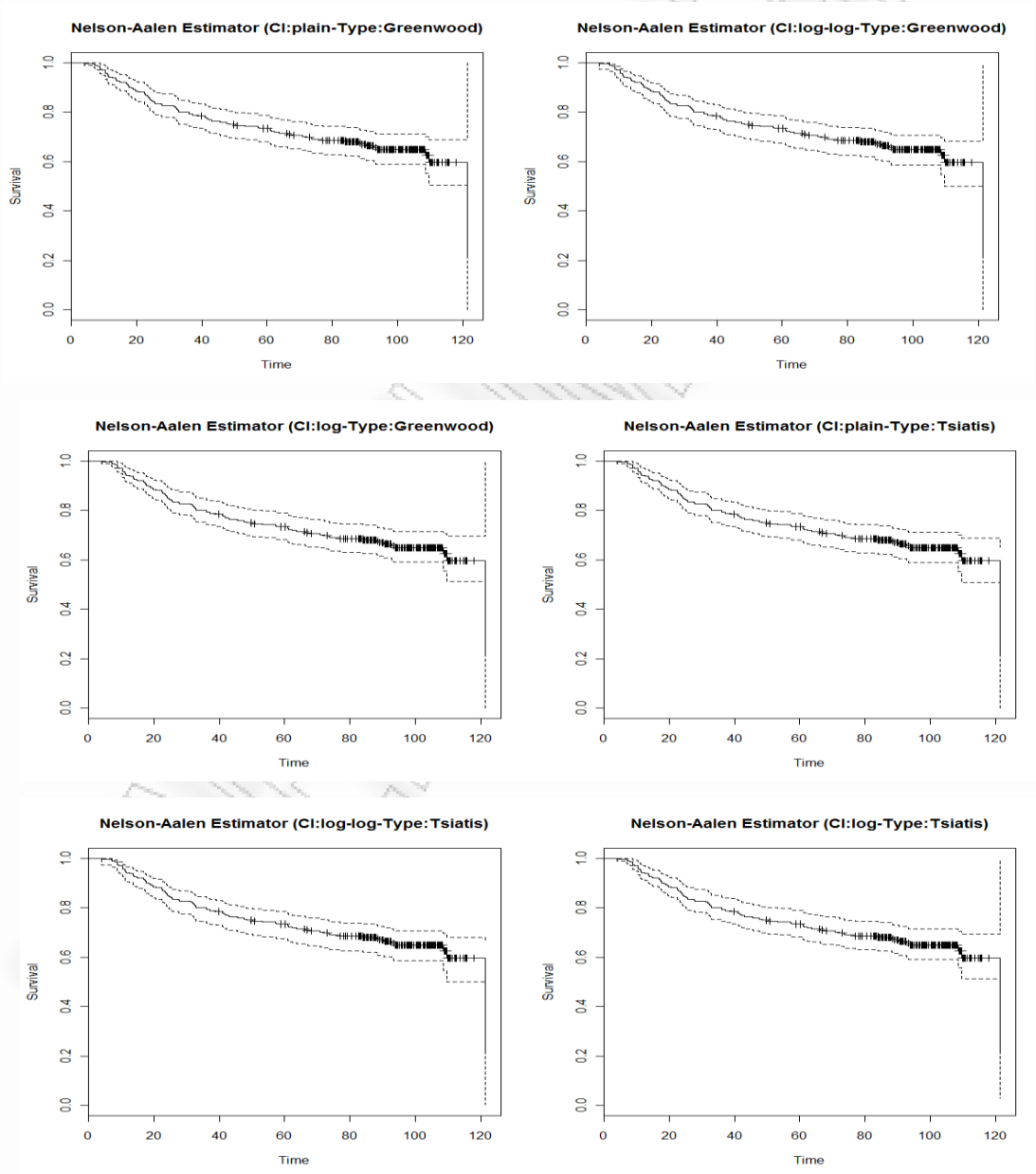
Για το συγκεκριμένο σύνολο δεδομένων και θεωρώντας τον ελεύθερο νόσου χρόνο, προκύπτει ότι η υπόθεση αναλογικού κινδύνου δεν ισχύει για τους κλινικοπαθολογικούς παράγοντες ‘Group’, ‘Age’, ‘Menopausal’, ‘Size’, ‘PgR’, ‘Surgery’ και ‘Interval’ (τα αντίστοιχα αποτελέσματα παρουσιάζονται στο Παράρτημα Α.10). Για τα συγκεκριμένα χαρακτηριστικά πιο αξιόπιστος είναι ο έλεγχος Wilcoxon, από τον οποίο προκύπτει ότι σε επίπεδο σημαντικότητας 5% για κανένας από αυτούς τους παράγοντες δεν μπορεί να απορριφθεί η μηδενική υπόθεση. Συνεπώς, οι ομάδες ασθενών που διαμορφώθηκαν από τα επίπεδα των παραπάνω παραγόντων δεν σημειώνουν διαφοροποίηση ως προς την εκτίμηση της συνάρτησης επιβίωσης. Ωστόσο, για τον παράγοντα του μεγέθους του όγκου, αν χαλαρώσουμε το επίπεδο σημαντικότητας στο 10%, τότε προκύπτει ότι είναι στατιστικά σημαντικός.

Για τους υπόλοιπους παράγοντες ‘Grade’, ‘Nodes’, ‘ER’, ‘HT’ και ‘RT’ υπάρχει γραφική ένδειξη ότι η υπόθεση αναλογικού κινδύνου ικανοποιείται αφού οι αντίστοιχες γραμμές φαίνεται να είναι παράλληλες. Θεωρώντας, λοιπόν, τον έλεγχο *log-rank* καταλήγουμε στο συμπέρασμα ότι σε επίπεδο σημαντικότητας 5%, μόνο οι παράγοντες ‘Grade’ και ‘Nodes’ προκύπτουν ως στατιστικά σημαντικοί. Όσον αφορά τον παράγοντα διαφοροποίησης του όγκου, η ομάδα με βαθμό διαφοροποίησης ‘I–II’, αντιστοιχεί σε υψηλότερη πιθανότητα επιβίωσης, σε σχέση με αυτή της ομάδας ‘III – Αδιαφοροποίητο’. Αντίστοιχα, για το πλήθος αφαιρούμενων λεμφαδένων, παρατηρείται ότι η ομάδα που αντιστοιχεί σε πλήθος θετικών λεμφαδένων μεγαλύτερο των τεσσάρων αντιστοιχεί σε μικρότερη πιθανότητα επιβίωσης. Η γραφική εικόνα που λαμβάνουμε είναι παρόμοια με αυτή του χρόνου επιβίωσης, δηλαδή η διαφορά των δύο ομάδων γίνεται εντονότερη με το πέρασμα του χρόνου και η γραμμή που αντιστοιχεί σε πλήθος 0–3 θετικών λεμφαδένων φαίνεται να ακολουθεί σταθερή πορεία σε

υψηλά επίπεδα επιβίωσης. Τέλος, αν χαλαρώσουμε το επίπεδο σημαντικότητας στο 10%, τότε ο παράγοντας 'RT' προκύπτει ως στατιστικά σημαντικός.

7.4.3 Nelson-Aalen εκτιμητής της συνάρτησης επιβίωσης και διαστήματα εμπιστοσύνης

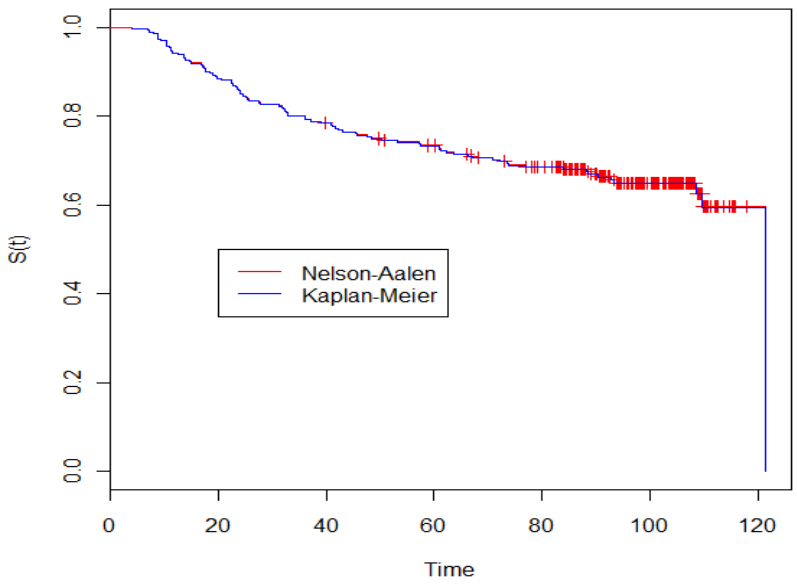
Στην συνέχεια παρουσιάζονται τα γραφήματα της Nelson-Aalen εκτίμησης της συνάρτησης επιβίωσης με τα αντίστοιχα διαστήματα τύπου plain, log-log και log.



Σχήμα 7.16: Εκτιμητής Nelson-Aalen με τους τύπους των Greenwood και Tsiatis με διαστήματα εμπιστοσύνης τύπου plain, log-log και log.

7.4.4 Σύγκριση συναρτήσεων επιβίωσης Kaplan-Meier και Nelson-Aalen

Συγκρίνοντας τους δύο παραπάνω εκτιμητές των συναρτήσεων επιβίωσης, προκύπτει το ακόλουθο γράφημα, από το οποίο είναι εμφανές ότι οι δύο γραμμές ταυτίζονται.



Σχήμα 7.17: Σύγκριση συναρτήσεων Kaplan-Meier και Nelson-Aalen.

7.4.5 Στρωματοποιημένοι έλεγχοι και έλεγχοι τάσης

Για την περεταίρω διερεύνηση των παραγόντων ‘Grade’ και ‘Nodes’, σύμφωνα με τα επίπεδα των οποίων η συνάρτηση του χρόνου DFS διαφοροποιείται, συνεχίζεται η ανάλυση σε στρωματοποιημένους ελέγχους και ελέγχους τάσης, προκειμένου να διαπιστωθεί αν η διαφοροποίηση αυτή αντανακλά την διαφορετικότητα των συναρτήσεων ως προς τον συγκεκριμένο παράγοντα. Ακολούθως παρουσιάζονται οι αυτόνομοι και ολικοί έλεγχοι στρωματοποίησης για κάθε ένα από τους προαναφερόμενους παράγοντες, αντιστοίχως.

ΠαράγοντΓας	Επίπεδο Παράγοντα Grade	Αυτόνομοι Έλεγχοι				Ολικοί Έλεγχοι			
		Έλεγχος	Chi-Square	df	Sig.	Έλεγχος	Chi-Square	df	Sig.
ER	**I-II**	Log Rank	1,576	1	0,209	Log Rank	0,425	1	0,515
	III-Undiffer.	Wilcoxon	0,006	1	0,941	Wilcoxon	0,562	1	0,453
PgR	**I-II**	Wilcoxon	0,286	1	0,593	Wilcoxon	0,056	1	0,813
	III-Undiffer.	Wilcoxon	0,017	1	0,895				
Group	**I-II**	Wilcoxon	0,35	1	0,554	Wilcoxon	0,232	1	0,63
	III-Undiffer.	Wilcoxon	0,16	1	0,899				
Age	**I-II**	Wilcoxon	0,151	1	0,698	Wilcoxon	0,057	1	0,738
	III-Undiffer.	Wilcoxon	0,437	1	0,509				
Menopausal	**I-II**	Wilcoxon	0,215	1	0,643	Wilcoxon	0,016	1	0,9
	III-Undiffer.	Wilcoxon	0,322	1	0,57				
Surgery	**I-II**	Log Rank	0,449	1	0,503	Log Rank	0,103	1	0,748
	III-Undiffer.	Wilcoxon	0,008	1	0,93	Wilcoxon	0,11	1	0,74
Interval	**I-II**	Wilcoxon	2,057	2	0,358	Wilcoxon	1,394	2	0,498
	III-Undiffer.	Wilcoxon	0,655	2	0,721				
Size	**I-II**	Wilcoxon	2,253	2	0,324	Wilcoxon	3,632	2	0,163
	III-Undiffer.	Wilcoxon	2,594	2	0,273				
Nodes	**I-II**	Log Rank	2,24	1	0,134	Log Rank	9,711	1	0,002
	III-Undiffer.	Log Rank	7,77	1	0,005				
RT	**I-II**	Log Rank	1,526	1	0,217	Log Rank	3,605	1	0,058
	III-Undiffer.	Log Rank	2,08	1	0,149				
HT	**I-II**	Wilcoxon	1,006	1	0,316	Wilcoxon	1,195	1	0,274
	III-Undiffer.	Wilcoxon	0,366	1	0,545				

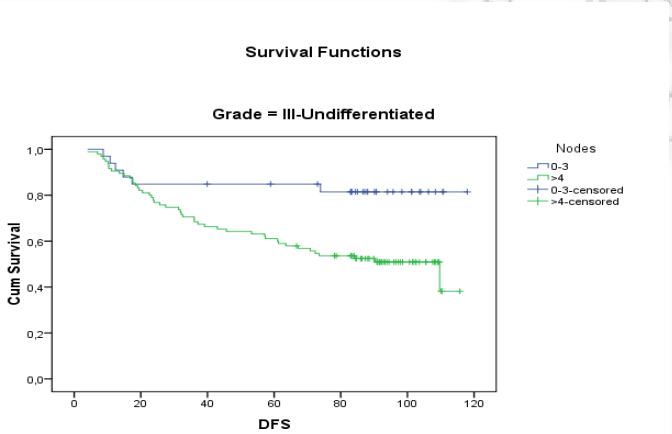
Πίνακας 7.8: Στρωματοποίηση ως προς τον Παράγοντα Grade.

Λαμβάνοντας τη στρωματοποίηση ως προς τον παράγοντα του βαθμού διαφοροποίησης του όγκου, προκύπτει ότι, σε επίπεδο σημαντικότητας 5% και σύμφωνα με τους 'αυτόνομους' ελέγχους, υπάρχει διαφοροποίηση του χρόνου υποτροπής, επανεμφάνισης ή θανάτου για το στρώμα 'III ή Αδιαφοροποίητο' ως προς τον παράγοντα του πλήθους των λεμφαδένων που αφαιρέθηκαν κατά την χειρουργική επέμβαση, με αντίστοιχη τιμή στατιστικού X^2=7,77 και 1 β.ε. Επομένως, φαίνεται ότι το πλήθος των αφαιρούμενων λεμφαδένων είναι σημαντικό για τον χρόνο DFS στην περίπτωση που ο βαθμός διαφοροποίησης του όγκου ανήκει στις κατηγορίες 'III ή Αδιαφοροποίητο'.

Θεωρώντας τους αντίστοιχους ολικούς ελέγχους, επιβεβαιώνεται το παραπάνω αποτέλεσμα σε επίπεδο σημαντικότητας 5%, αφού προκύπτει p-value=0,002 για την

αντίστοιχη μεταβλητή, οπότε απορρίπτεται η μηδενική υπόθεση ισότητας των συναρτήσεων του χρόνου DFS για την μεταβλητή ‘Nodes’ στο επίπεδο ‘III ή Αδιαφοροποίητο’ του παράγοντα ‘Grade’. Χαλαρώνοντας το επίπεδο σημαντικότητας σε α=0,10, προκύπτει ότι απορρίπτεται η μηδενική υπόθεση του ολικού ελέγχου και για τον παράγοντα RT, χωρίς αυτή η διαφορά να καταγράφεται από τους αυτόνομους ελέγχους κάθε επιπέδου του παράγοντα στρωματοποίησης.

Στην συνέχεια παρουσιάζεται η γραφική απεικόνιση του παραπάνω αποτελέσματος για τον παράγοντα ‘Nodes’.



Σχήμα 7.18: Διάγραμμα επιβίωσης για στρώμα ‘ III ή Αδιαφορ.’ του παράγοντα Grade για Nodes.

Ενώ κατά τις αρχικές 20 εβδομάδες παρακολούθησης των γυναικών δεν φαίνεται να παρουσιάζεται κάποια διαφοροποίηση στις γραμμές που αντιστοιχούν στα επίπεδα του πλήθους αφαιρούμενων λεμφαδένων, πέραν αυτού του χρονικού σημείου η πιθανότητα χρόνου ελευθέρου νόσου για την ομάδα των γυναικών στις οποίες αφαιρέθηκαν έως 3 μασχαλιαίοι λεμφαδένες σταθεροποιείται στο επίπεδο της τάξεως του 80%. Αντιθέτως, η αντίστοιχη πιθανότητα για τις γυναίκες με αφαίρεση άνω των τεσσάρων λεμφαδένων, φαίνεται να διαγράφει έντονα καθοδική πορεία.

	Grade		Chi-Square	df	Sig.
Nodes	2 III-Undifferentiated	Log Rank	7,770	1	,005
		Breslow	6,017	1	,014
		Tarone-Ware	6,794	1	,009
	Pooled	Log Rank	9,711	1	,002
		Breslow	7,758	1	,005
		Tarone-Ware	8,631	1	,003

Πίνακας 7.9: Έλεγχος τάσης του παράγονται Nodes ως προς στρωματοποίηση του Grade.

Σύμφωνα με τον έλεγχο τάσης του παράγοντα ‘Nodes’ και θεωρώντας ως επίπεδο σημαντικότητας το 5%, απορρίπτεται η μηδενική υπόθεση. Επομένως, καθώς αυξάνεται το πλήθος των θετικών λεμφαδένων που εντοπίστηκαν, μειώνεται η πιθανότητα ελευθέρας νόσου επιβίωσης των ασθενών, όταν ο βαθμός διαφοροποίησης του όγκου ανήκει στη κατηγορία ‘III–Αδιαφοροποίητο’. Παρατηρούμε ότι στα ίδια αποτελέσματα καταλήγουμε λαμβάνοντας το από κοινού δείγμα.

Παράγοντας	Επίπεδο Παράγοντα Nodes	Αυτόνομοι Έλεγχοι				Ολικοί Έλεγχοι			
		Έλεγχος	Chi-Square	df	Sig.	Έλεγχος	Chi-Square	df	Sig.
ER	**0-3**	Log Rank	0,733	1	0,392	Log Rank	2,785	1	0,095
	>4	Log Rank	4,783	1	0,029				
PgR	**0-3**	Breslow	0,131	1	0,718	Log Rank	1,355	1	0,244
	>4	Log Rank	1,711	1	0,191	Breslow	1,249	1	0,264
Group	**0-3**	Breslow	0,237	1	0,627	Breslow	0,315	1	0,575
	>4	Breslow	0,397	1	0,529				
Age	**0-3**	Log Rank	1,59	1	0,207	Log Rank	0,383	1	0,536
	>4	Breslow	0,063	1	0,802	Breslow	0,146	1	0,702
Menopausal	**0-3**	Log Rank	0,895	1	0,344	Log Rank	0,229	1	0,633
	>4	Breslow	0,065	1	0,799	Breslow	0,119	1	0,73
Surgery	**0-3**	Log Rank	0,609	1	0,435	Log Rank	0,353	1	0,553
	>4	Breslow	0,149	1	0,7	Breslow	0,216	1	0,642
Interval	**0-3**	Log Rank	1,938	2	0,379	Log Rank	1,9	2	0,387
	>4	Log Rank	2,391	2	0,303				
Grade	**0-3**	Log Rank	0,072	1	0,788	Log Rank	7,203	1	0,007
	>4	Log Rank	7,657	1	0,006				
Size	**0-3**	Breslow	0,697	2	0,706	Breslow	4,882	2	0,087
	>4	Breslow	4,957	2	0,084				
RT	**0-3**	Log Rank	0,809	1	0,369	Log Rank	0,291	1	0,59
	>4	Breslow	0,027	1	0,87	Breslow	0,018	1	0,895
HT	**0-3**	Breslow	0,021	1	0,886	Log Rank	1,949	1	0,163
	>4	Log Rank	2,464	1	0,116	Breslow	2,65	1	0,104

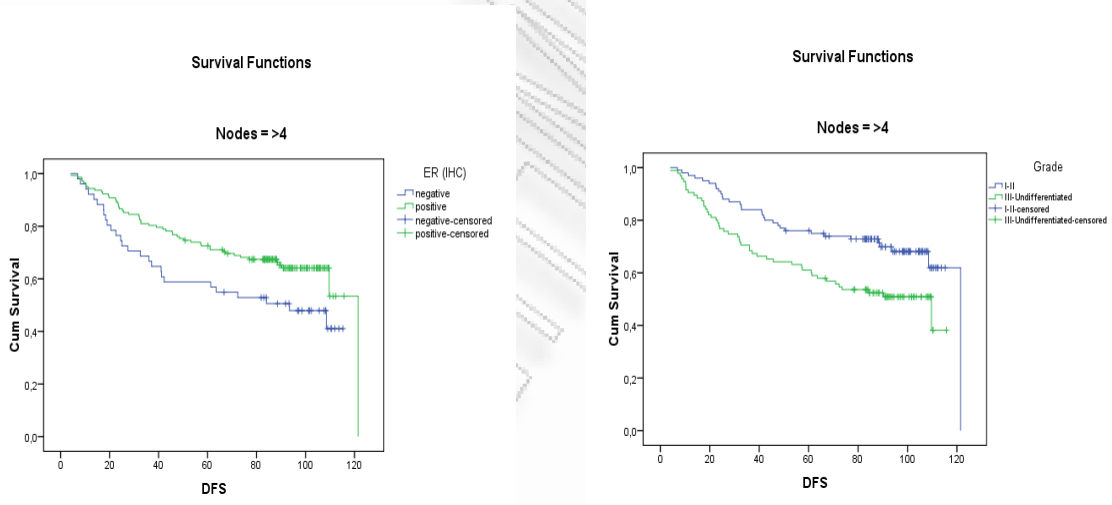
Πίνακας 7.10: Στρωματοποίηση ως προς τον Παράγοντα Nodes.

Θεωρώντας ως παράγοντα στρωματοποίησης το πλήθος των αφαιρούμενων κατά την επέμβαση λεμφαδένων και λαμβάνοντας τους ‘αυτόνομους’ ελέγχους για κάθε στρώμα χωριστά, παρατηρούμε ότι για τους παράγοντες ‘ER’ και ‘Grade’ απορρίπτεται η μηδενική υπόθεση των ‘αυτόνομων ελέγχων’ σε επίπεδο σημαντικότητας 5%, όσον αφορά το επίπεδο λεμφαδένων περισσότερων των τεσσάρων, με αντίστοιχες τιμές στατιστικού X^2=4,783 και

X^2=7,657 και 1 β.ε. για κάθε έλεγχο. Όπως προκύπτει από τους παραπάνω ελέγχους, η θετική ή αρνητική πρωτεϊνική έκφραση των οιστρογονικών υποδοχέων και το πλήθος των αφαιρούμενων λεμφαδένων είναι σημαντικοί παράγοντες προσδιορισμού του χρόνου DFS στην περίπτωση που αφαιρούνται περισσότεροι των τεσσάρων λεμφαδένων.

Σύμφωνα με τους ολικούς ελέγχους, παρατηρούμε ότι επιβεβαιώνεται το αποτέλεσμα για τον παράγοντα 'Grade', σε επίπεδο σημαντικότητας 5%, δεδομένου ότι απορρίπτεται η μηδενική υπόθεση με αντίστοιχη p-value=0,007. Χαλαρώνοντας το επίπεδο σημαντικότητας σε 10%, επιβεβαιώνεται το αντίστοιχο αποτέλεσμα και για τον παράγοντα 'ER' (p-value=0,095). Επιπλέον, προκύπτει απόρριψη της υπόθεσης ισότητας των συναρτήσεων επιβίωσης και για τον παράγοντα του μεγέθους του όγκου, με στατιστικό X^2=4,882 και 1 β.ε., κάτι που δεν εντοπίστηκε θεωρώντας ελέγχους για κάθε στρώμα του παράγοντα 'Nodes'.

Στην συνέχεια παρουσιάζονται οι γραφικές αναπαραστάσεις των συναρτήσεων επιβίωσης που σημειώνουν διαφοροποίηση, όπως αυτές εντοπίστηκαν από τους αυτόνομους ελέγχους.



Σχήμα 7.19: Διάγραμμα επιβίωσης για στρώμα '>4' του παράγοντα Nodes για ER και Grade.

Ως προς τον παράγοντα έκφρασης των οιστρογονικών υποδοχέων, παρατηρούμε οι δύο γραμμές που αντιστοιχούν σε θετική και αρνητική έκφραση των υποδοχέων να διαγράφουν σχεδόν παράλληλη πορεία, διατηρώντας ίδιο ρυθμό μείωσης της συνάρτησης επιβίωσης. Η δε πιθανότητα επιβίωσης σημειώνει γρηγορότερη μείωση για τις γυναίκες στις οποίες σημειώνεται αρνητική έκφραση των συγκεκριμένων υποδοχέων, όταν αφαιρούνται άνω των τεσσάρων λεμφαδένων.

Αναφορικά με τον παράγοντα του βαθμού διαφοροποίησης του όγκου, παρατηρούμε τη γραμμή που αντιστοιχεί σε βαθμό I ή II να είναι πάνω από την γραμμή της ομάδας ασθενών

με βαθμό διαφοροποίησης ‘III ή Αδιαφοροποίητο’. Το γεγονός αυτό σημαίνει ότι για την πρώτη ομάδα η πιθανότητα επιβίωσης μειώνεται με πιο αργό ρυθμό, οπότε στις γυναίκες αυτής της κατηγορίας το ενδεχόμενο επιβίωσης χωρίς υποτροπή ή επανεμφάνιση του όγκου είναι μεγαλύτερο.

	Nodes		Chi-Square	df	Sig.
Grade	1 >4	Log Rank	7,657	1	,006
		Breslow	8,286	1	,004
		Tarone-Ware	8,034	1	,005
	Pooled	Log Rank	7,203	1	,007
		Breslow	8,488	1	,004
		Tarone-Ware	8,106	1	,004

Πίνακας 7.11: Έλεγχος τάσης του παράγοντα Grade ως προς στρωματοποίηση του Nodes.

Τέλος, σύμφωνα με τον έλεγχο τάσης για τον παράγοντα ‘Grade’, ο οποίος σημειώνει διαταξιμότητα, απορρίπτεται η μηδενική υπόθεση σε επίπεδο σημαντικότητας 5%, τόσο για το στρώμα των ατόμων με περισσότερους από τέσσερις αφαιρούμενους λεμφαδένες, όσο και στην περίπτωση του από κοινού δείγματος. Επομένως, όταν αυξάνεται ο βαθμός διαφοροποίησης του όγκου, ως προς τις αντίστοιχες κατηγορίες, μειώνεται η πιθανότητα επιβίωσης χωρίς εμφάνιση νόσου, με μορφή υποτροπής ή μετάστασης, για την ασθενή στην οποία χρειάστηκε να αφαιρεθούν άνω των τεσσάρων λεμφαδένες.

7.5 Μοντέλο αναλογικού κινδύνου του ελεύθερου νόσου χρόνου

Αντίστοιχα με τον χρόνο επιβίωσης, εφαρμόζεται το *μοντέλο αναλογικού κινδύνου του Cox*, λαμβάνοντας υπόψη τους ίδιους κλινικοπαθολογικούς παράγοντες και τις συνιστώσες των γονιδιακών εκφράσεων, προκειμένου να εξετασθούν από κοινού.

7.5.1 Αναζήτηση μοντέλου

Οι βωβές μεταβλητές που απαιτούνται προκειμένου να εξετασθεί χωριστά κάθε επίπεδο των κατηγορικών παραγόντων συμπίπτουν με αυτές που μελετήσαμε προηγούμενα για την περίπτωση του χρόνου επιβίωσης (βλέπε Παράρτημα Α.6) και θεωρούμε ότι η ΑΣΚ ορίζεται

έτσι να αντιστοιχεί στο πρώτο επίπεδο των κατηγορικών παραγόντων, ενώ για τους συνεχείς παράγοντες των γονιδιακών εκφράσεων λαμβάνει μηδενική τιμή.

Αρχικά εφαρμόζουμε Backward Model Selection και μετά από 16 βήματα καταλήγουμε στο ακόλουθο μοντέλο:

								95,0% CI for Exp(B)	
		B	SE	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 16	Grade	,588	,232	6,439	1	,011	1,800	1,143	2,834
	Nodes	1,080	,339	10,156	1	,001	2,945	1,516	5,723
	Factor2	,233	,114	4,172	1	,041	1,262	1,009	1,579
	Factor5	-,369	,113	10,645	1	,001	,691	,554	,863
	Factor7	,277	,114	5,929	1	,015	1,319	1,056	1,649
	Factor8	,188	,103	3,357	1	,067	1,207	,987	1,476

Πίνακας 7.12: Εκτιμήσεις των παραμέτρων του μοντέλου αναλογικού κινδύνου σύμφωνα με τη διαδικασία Backward Selection.

Σε σύγκριση με το προηγούμενο μοντέλο, στη περίπτωση του χρόνου DFS προκύπτει σημαντικός και ο παράγοντας που δηλώνει την διαφοροποίηση του όγκου (Grade). Ως προς τους παράγοντες που αντιπροσωπεύουν τα γονίδια, παρατηρούμε ότι ο Factor1, ο οποίος αντιστοιχεί στην ομάδα 1, δεν είναι σημαντικός, και αντί αυτού στο μοντέλο περιλαμβάνεται ο Factor8, ο οποίος αντιστοιχεί όμως στην ομάδα 2. Η μεταβλητή της θεραπευτικής αγωγής δεν προκύπτει σημαντική και πάλι, ωστόσο δεν μπορεί να παραληφθεί. Εξετάζοντας και τις αλληλεπιδράσεις της μεταβλητής 'Group' με τους παράγοντες των γονιδιακών εκφράσεων, καμία αλληλεπίδραση δεν προκύπτει ως σημαντική, και επιπλέον απορρίπτεται και ο 'Factor8'. Προκειμένου να επιβεβαιώσουμε το παραπάνω αποτέλεσμα, εφαρμόζουμε Backward Model Selection στους παράγοντες που προέκυψαν αρχικά ως σημαντικοί και εισάγοντας και τον παράγοντα 'Group', ο οποίος δεν μπορεί να μείνει εκτός μοντέλου λόγω της αξίας του στο σχεδιασμό της μελέτης, τελικά λαμβάνουμε τα ακόλουθα αποτελέσματα.

							95,0% CI for Exp(B)	
	B	SE	Wald	df	Sig.	Exp(B)	Lower	Upper
Group	,088	,215	,166	1	,684	1,092	,716	1,664
Grade	,593	,230	6,643	1	,010	1,809	1,153	2,839
Nodes	1,043	,324	10,344	1	,001	2,838	1,503	5,360
Factor2	,202	,112	3,258	1	,071	1,224	,983	1,524
Factor5	-,365	,112	10,604	1	,001	,694	,558	,865
Factor7	,303	,113	7,232	1	,007	1,354	1,086	1,689

Πίνακας 7.13: Εκτιμήσεις των παραμέτρων του τελικού μοντέλου αναλογικού κινδύνου.

Τελικά, καταλήγουμε στο ακόλουθο μοντέλο, σύμφωνα με τον Πίνακα 7.14:

$$h(t) = h_0(t)\ \exp(0{,}088 \cdot \text{Group} + 0{,}593 \cdot \text{Grade} + 1{,}043 \cdot \text{Nodes} + 0{,}202 \cdot \text{Factor2} - 0{,}365 \cdot \text{Factor5} + 0{,}303 \cdot \text{Factor7})$$

Προκειμένου να εξετασθεί η σταθερότητα του μοντέλου, εφαρμόζουμε Backward Cox Regression στους παράγοντες των κλινικών χαρακτηριστικών και των γονιδιακών εκφράσεων χωριστά (Παράρτημα Α.11), και έπειτα στους παράγοντες που προκύπτουν σημαντικοί από τη διαδικασία αυτή. Τελικά, καταλήγουμε στο ίδιο μοντέλο, οπότε το μοντέλο στο οποίο βασίζουμε την συμπερασματολογία χαρακτηρίζεται από σταθερότητα. Αξίζει να σημειώσουμε, ωστόσο, ότι πλέον δεν υπάρχει κάποιος παράγοντας στο μοντέλο ο οποίος να αντιπροσωπεύει την Ομάδα 1, όπως αυτή διαμορφώθηκε κατά την ομαδοποίηση των γονιδίων.

Από τους ολικούς ελέγχους *score* και *log-rank*, προκύπτουν τα ακόλουθα αποτελέσματα,

	Overall (score)			Change From Previous Step			Change From Previous Block		
-2 Log Likelihood	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
902,834	33,788	6	,000	36,528	6	,000	36,528	6	,000

Πίνακας 7.14: Έλεγχοι *score* και *log-rank*.

Σε επίπεδο σημαντικότητας 5%, οι παραπάνω έλεγχοι δίνουν X^2 = 33,788 και X^2=36,528, αντίστοιχα, με 6 β.ε. ο καθένας και p-value πρακτικά μηδέν. Καταλήγουμε, λοιπόν, σε κοινό συμπέρασμα, να δεχθούμε ότι η συνάρτηση κινδύνου εξαρτάται τουλάχιστον από μία από τις υπό εξέταση μεταβλητές. Από τους τοπικούς ελέγχους που παρέχονται από το Wald test (Πίνακας 7.13), παρατηρούμε ότι σε επίπεδο σημαντικότητας 5%, η μηδενική υπόθεση απορρίπτεται για το σύνολο των παραγόντων, με εξαίρεση τη 'Group'.

7.5.2 Ερμηνεία μοντέλου

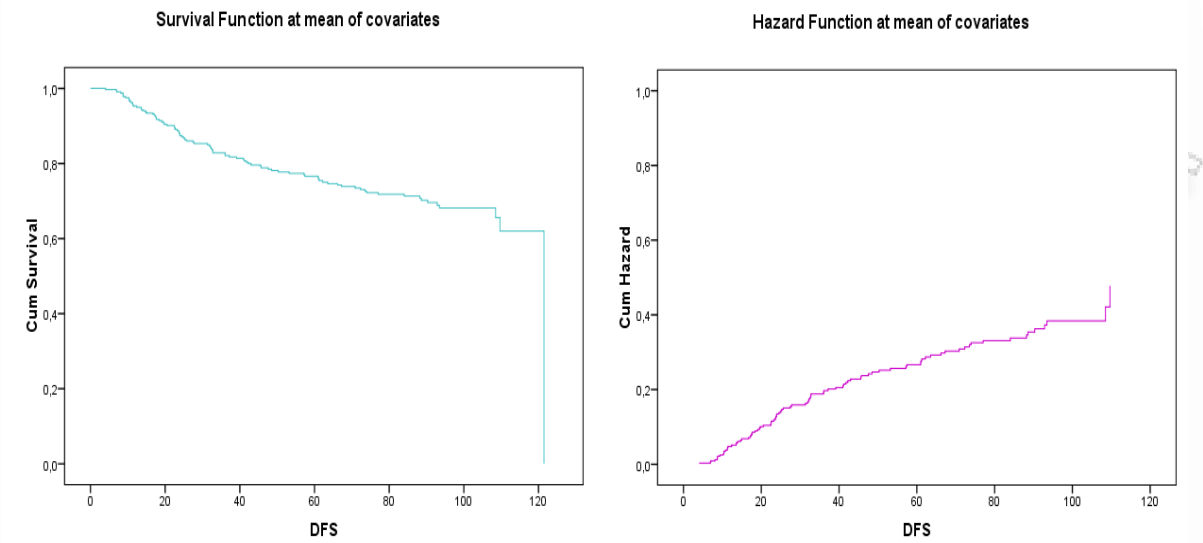
Η ΑΣΚ του μοντέλου, λαμβάνοντας υπόψη το ελεύθερο νόσου χρονικό διάστημα, αντιστοιχεί σε ασθενείς οι οποίες ανήκουν στην θεραπευτική αγωγή E-T-CMF, σημειώθηκε βαθμός διαφοροποίησης I ή II, έχουν αφαιρέσει 0-3 μασχαλιαίους λεμφαδένες και στους παράγοντες που αντιστοιχούν στις ομάδες των γονιδίων έχουν μηδενικές τιμές.

Ως προς την παράμετρο β_1 της μεταβλητής Group, προκύπτουν οι εκτιμήσεις $\hat{\beta}_1$= 0,88 και $s\hat{e}\left(\hat{\beta}_1\right)$ = 0,215. Όπως προκύπτει τόσο από τον έλεγχο του Wald (X^2=0,166 με 1 β.ε., p-value=0,684) όσο και από το διάστημα εμπιστοσύνης, η παράμετρος δεν είναι στατιστικά σημαντική σε επίπεδο σημαντικότητας 5%, οπότε περαιτέρω ερμηνεία της μεταβλητής δεν σημειώνει στατιστικό ή ερευνητικό ενδιαφέρον.

Στη μεταβλητή που δηλώνει την διαφοροποίηση του όγκου της ασθενούς (Grade) αντιστοιχεί εκτίμηση παραμέτρου $\hat{\beta}_2$=0,593 με τυπικό σφάλμα $s\hat{e}\left(\hat{\beta}_2\right)$=0,230. Ο λόγος κινδύνου, θεωρώντας ότι οι υπόλοιποι παράγοντες του μοντέλου παραμένουν σταθεροί, είναι 1,809, οπότε μία γυναίκα με βαθμό διαφοροποίησης του όγκου 'III ή Αδιαφοροποίητο' διατρέχει κατά 1,8 φορές μεγαλύτερο κίνδυνο επέλευσης του κινδύνου, σε σχέση με μία γυναίκα στην οποία σημειώθηκε διαφοροποίηση I ή II. Αντίστοιχα, για τη μεταβλητή του πλήθους των αφαιρούμενων λεμφαδένων (Nodes) προκύπτουν $\hat{\beta}_3$=1,043 και $s\hat{e}\left(\hat{\beta}_3\right)$=0,324. Ο αντίστοιχος λόγος κινδύνου είναι 2,838, οπότε μία γυναίκα στην οποία αφαιρέθηκαν περισσότεροι των τεσσάρων λεμφαδένες έχει 2,8 φορές μεγαλύτερο κίνδυνο θανάτου σε σχέση με την περίπτωση να αφαιρούνταν έως τρεις λεμφαδένες, όταν οι υπόλοιποι παράγοντες παραμένουν αμετάβλητοι.

Από τους παράγοντες που αντιστοιχούν στις γονιδιακές εκφράσεις, οι 'Factor2' και 'Factor7' αντιστοιχούν στην ομάδα 2, ενώ ο 'Factor5' αντιστοιχεί στην ομάδα 3. Τα θετικά πρόσημα των δύο πρώτων παραγόντων δηλώνουν ότι αύξηση στην τιμή τους συνεπάγεται μεγαλύτερο κίνδυνο υποτροπής, μετάστασης ή θανάτου. Αντίθετα, υψηλότερες τιμές του 'Factor5' συντελούν σε μεγαλύτερο χρονικό διάστημα χωρίς εμφάνιση της νόσου. Οι λόγοι κινδύνου για τους 'Factor2', 'Factor5'και 'Factor7' είναι 1.224, 0.694 και 1.354, αντιστοίχως. Σύμφωνα με τις παραπάνω εκτιμήσεις των λόγων κινδύνου, όταν οι παράγοντες 2 και 7 αυξάνονται κατά μία μονάδα (Ομάδα γονιδίων 2), ο κίνδυνος των ασθενών αυξάνεται κατά 1,224 και 1,354, αντίστοιχα. Αντιθέτως, στην μοναδιαία αύξηση του παράγοντα 5 (Ομάδα γονιδίων 3), ο κίνδυνος επανεμφάνισης, υποτροπής ή θανάτου μειώνεται κατά 1,5 φορές (1/0,694 = 1,441). Η ταυτόχρονη μοναδιαία αύξηση των παραγόντων 2 και 7 δηλώνει αύξηση του κινδύνου κατά 1,657 φορές.

Στη συνέχεια παρατίθενται τα διαγράμματα επιβίωσης και κινδύνου, αντίστοιχα, στο μέσο των συμμεταβλητών.



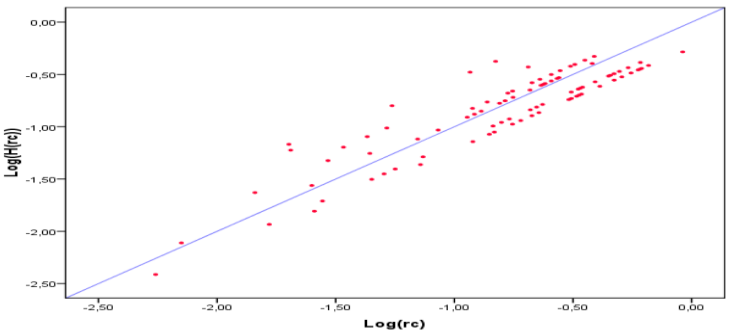
Σχήμα 7.20: Διαγράμματα των συναρτήσεων επιβίωσης και κινδύνου.

Ο ελεύθερος νόσου χρόνος για μία γυναίκα φαίνεται να διαγράφει έντονα πτωτική πορεία από τις πρώτες κιόλας βδομάδες παρακολούθησης, στα πλαίσια της έρευνας. Η πιθανότητα επιβίωσης από ενδεχόμενη επανεμφάνιση της νόσου, είτε ως υποτροπή είτε ως εντοπισμός άλλου ογκώδους σχηματισμού, ή ο θάνατος από αυτή, δεν φαίνεται να ξεπερνά το 60%. Επιπλέον, η τελευταία παρατήρηση δεν είναι λογοκριμένη. Παρατηρούμε ότι η πιθανότητα επιβίωσης που αντιστοιχεί στον χρόνο DFS είναι μικρότερη κατά δέκα ποσοστιαίες μονάδες σε σχέση με αυτή του χρόνου επιβίωσης από το θάνατο.

7.5.3 Ανάλυση υπολοίπων

Στη συνέχεια παρατίθενται τα διαγράμματα που αντιστοιχούν στα υπόλοιπα του μοντέλου. Επισημαίνουμε και πάλι ότι από την ερμηνεία των παρακάτω υπολοίπων καταλήγουμε σε ουσιώδη συμπεράσματα σχετικά με την ολική επάρκεια του μοντέλου, τον έλεγχο της υπόθεσης αναλογικού κινδύνου, την ανίχνευση έκτροπων παρατηρήσεων (outliers) και την επιρροή κάθε παρατήρησης στην εκτίμηση των ΕΜΠ κάθε μεταβλητής και την εύρεση της συναρτησιακής μορφής κάθε μεταβλητής που εισάγεται στο μοντέλο.

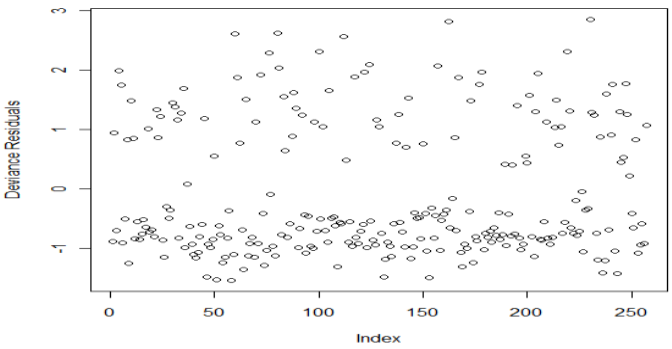
Cox-Snell Υπόλοιπα



Σχήμα 7.21: Διάγραμμα *Cox-Snell* υπολοίπων.

Όπως προκύπτει από την παραπάνω γραφική ένδειξη, τα *Cox-Snell* υπόλοιπα ακολουθούν αρκετά ικανοποιητικά την ευθεία $y=x$, γεγονός που σημαίνει ότι η ολική επάρκεια του μοντέλου είναι ικανοποιητική και, συνεπώς, επαληθεύεται η υπόθεση αναλογικού κινδύνου για το μοντέλο του Cox το οποίο αναφέρεται στην χρόνο DFS.

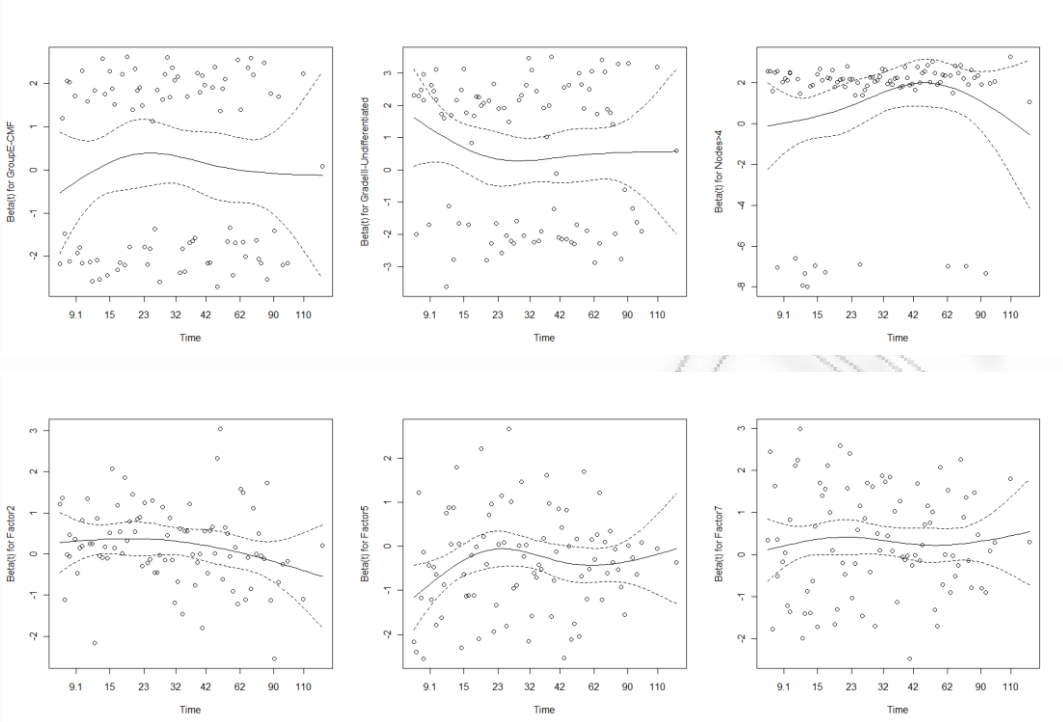
Υπόλοιπα Martingale και Deviance



Σχήμα 7.22: Διάγραμμα *Deviance* υπολοίπων.

Σύμφωνα με το διάγραμμα των *Deviance* υπολοίπων, δεν υπάρχει σαφής ένδειξη διαχωρισμού των πλήρων από τους λογοκριμένους χρόνους DFS, καθώς οι παρατηρήσεις φαίνεται να έχουν ‘κατέβει’ χαμηλότερα σε σχέση με τον χρόνο Survival που εξετάστηκε παραπάνω. Η διασπορά των σημείων είναι μεγαλύτερη, με αποτέλεσμα να μην διαφαίνονται ξεκάθαρα τυχόν έκτροπες παρατηρήσεις.

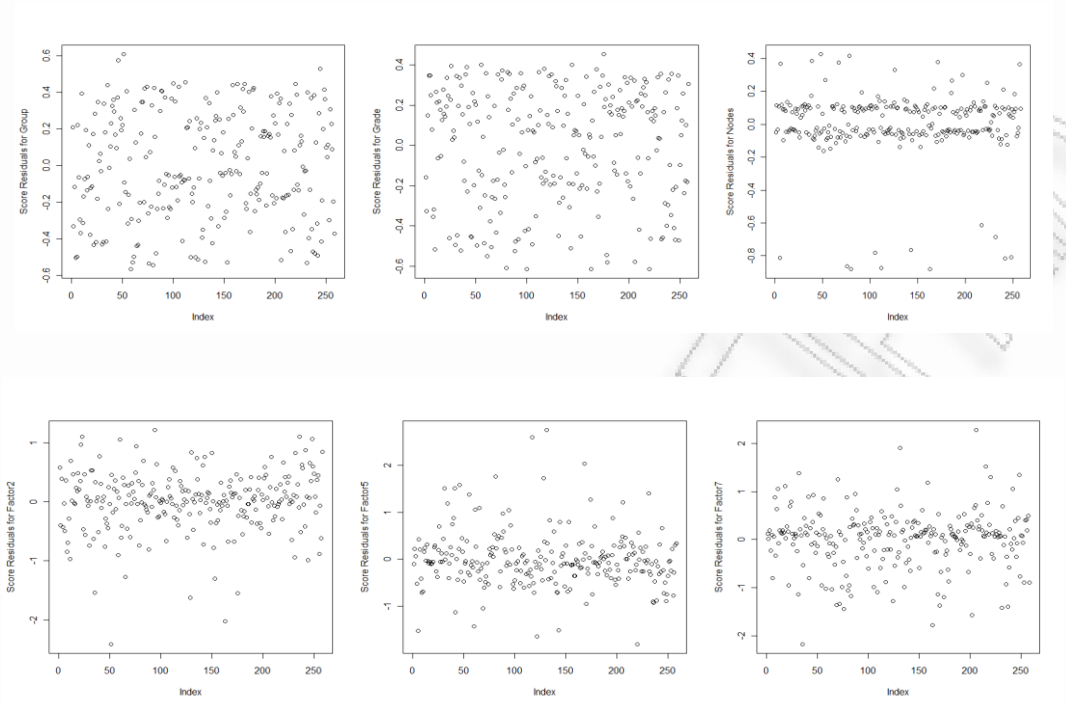
Schoenfeld Υπόλοιπα



Σχήμα 7.23: Διαγράμματα *Schoenfeld* υπολοίπων για κάθε παράγοντα.

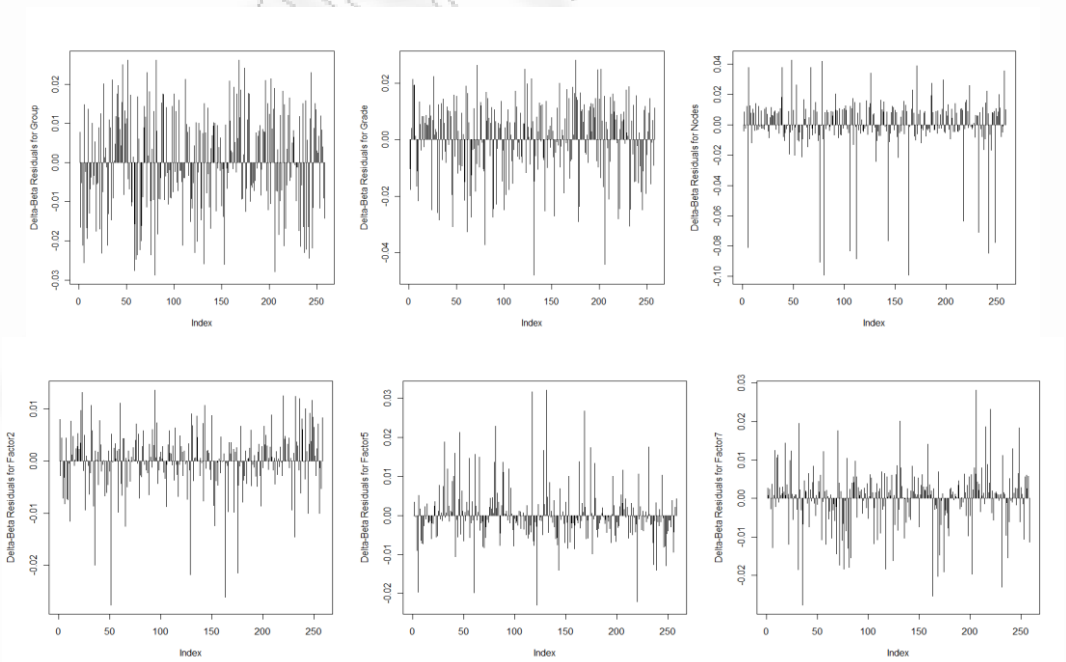
Από τα rescaled Schoenfeld υπόλοιπα για κάθε παράγοντα του μοντέλου, φαίνεται να ισχύει η μηδενική κλίση ως προς την ευθεία $g(t) = t$, γεγονός που επιβεβαιώνεται και μέσω των ελέγχων μηδενικής κλίσης της smoothing curve. Τόσο οι ατομικοί έλεγχοι όσο και ο ολικός (Παράρτημα Α.12) καταλήγουν στο συμπέρασμα ότι σε επίπεδο σημαντικότητας 5% δεν απορρίπτουν την μηδενική υπόθεση, οπότε για κάθε μεταβλητή, όπως επίσης και για το τελικό μοντέλο, ισχύει η υπόθεση αναλογικού κινδύνου.

Score και scaled score (Δέλτα-Βήτα) Υπόλοιπα



Σχήμα 7.24: Διαγράμματα *Score* υπολοίπων για κάθε παράγοντα.

Σύμφωνα με τις παραπάνω γραφικές απεικονίσεις, τα score residuals κατανέμονται με τυχαίο τρόπο γύρω από το μηδέν, οπότε υπάρχει ένδειξη ότι ικανοποιείται συνολικά η υπόθεση αναλογικού κινδύνου. Ωστόσο, αξίζει να σημειωθεί ότι για τους παράγοντες ‘Nodes’, ‘Factor2’ και ‘Factor5’ παρατηρούνται κάποιες αρκετά απομακρυσμένες τιμές.



Σχήμα 7.25: Διαγράμματα scaled score υπολοίπων για κάθε παράγοντα.

Τέλος, θεωρώντας τα διαγράμματα των υπολοίπων Δέλτα-Βήτα, διαχωρίζονται εκείνες οι ασθενείς οι οποίες συνεισφέρουν περισσότερο στην εκτίμηση του ΕΜΠ ενός συγκεκριμένου παράγοντα, και συγκεκριμένα είναι αυτές που αντιστοιχούν σε τιμή με μεγάλη απόκλιση από το μηδέν.

7.6 Συμπερασματολογία

7.6.1 Χρόνος επιβίωσης

Συνοψίζοντας τα αποτελέσματα που αφορούν στον χρόνο επιβίωσης των ασθενών με καρκίνου του μαστού, καταλήγουμε στα ακόλουθα συμπεράσματα για το συγκεκριμένο δείγμα ασθενών που εξετάζουμε.

Θεωρώντας τη μη παραμετρική εκτίμηση *Kaplan-Meier*, και σύμφωνα με τους ελέγχους log-rank και Wilcoxon, ανάλογα με την ισχύ ή μη της υπόθεσης αναλογικού κινδύνου, δεν υπάρχει στατιστικώς σημαντική διαφορά στην πορεία των συναρτήσεων επιβίωσης για τις ομάδες που διαμορφώνονται ως προς τις κατηγορίες των παραγόντων ‘Group’, ‘Age’, ‘Menopausal’, ‘Size’, ‘RT’, ‘Surgery’, ‘Interval’ ‘ER’, ‘PgR’ και ‘HT’. Αντιθέτως, οι ομάδες που διαμορφώνονται για τα επίπεδα των παραγόντων ‘Grade’ και ‘Nodes’ διαφέρουν από στατιστικής σκοπιάς, σε επίπεδο σημαντικότητας 5%.

Για περαιτέρω διερεύνηση των σημαντικών παραγόντων, εξετάζονται ως προς τη στρωματοποίηση και την τάση, καταλήγοντας στα εξής αποτελέσματα:

- ✓ Για το στρώμα ‘I-II’ του παράγοντα ‘Grade’ σημειώνεται διαφορά στις συναρτήσεις επιβίωσης του παράγοντα ‘HT’, με καλύτερη πρόγνωση επιβίωσης να αντιστοιχεί στις ασθενείς που έχουν υποβληθεί σε χημειοθεραπεία.
- ✓ Για τις γυναίκες που ανήκουν στην κατηγορία ‘III-Αδιαφοροποίητο’ του παράγοντα διαφοροποίησης του όγκου, με την αύξηση του μεγέθους του αφαιρούμενου όγκου ή του πλήθους των λεμφαδένων που αφαιρέθηκαν κατά την χειρουργική επέμβαση, μειώνεται η πιθανότητα επιβίωσης.
- ✓ Όταν σε ασθενή αφαιρέθηκαν περισσότεροι των τεσσάρων μασχαλιαίοι λεμφαδένες, η συνάρτηση επιβίωσης διαφοροποιείται για τους παράγοντες ‘Grade’, ‘Size’, ‘HT’ (α=5%) και ‘ER’, ‘PgR’ (α=10%). Όταν αυξάνεται η κατηγορία διαφοροποίησης του όγκου ή το μέγεθος αυτού, μειώνεται η πιθανότητα επιβίωσης,

ενώ καλύτερη πρόγνωση σημειώνεται στις γυναίκες που έχουν υποβληθεί σε χημειοθεραπεία, και είναι θετικές σε οιστρογονικούς και προγεστερονικούς υποδοχείς.

Σύμφωνα με τους παράγοντες που εξετάζονται για το συγκεκριμένο σύνολο δεδομένων, αυτοί που απαρτίζουν το *μοντέλο αναλογικού κινδύνου του Cox* είναι οι 'Group', 'Nodes', 'Factor1', 'Factor2', 'Factor5' και 'Factor7'. Από τις συνιστώσες που αντιστοιχούν στις γονιδιακές εκφράσεις, οι 'Factor1' και 'Factor5' αντιστοιχούν στις ομάδες 1 και 3 των γονιδιακών εκφράσεων, ενώ οι 'Factor2' και 'Factor7' αντιπροσωπεύουν την $2^{η}$ ομάδα. Οι υπόλοιποι υπό εξέταση παράγοντες δεν προκύπτουν στατιστικά σημαντικοί, σύμφωνα με την μέθοδο επιλογής μοντέλου Backward Selection.

Σύμφωνα με τις εκτιμήσεις του μοντέλου, καταλήγουμε στα ακόλουθα συμπεράσματα:

- ✓ Ο παράγοντας της φαρμακευτικής αγωγής στην οποία υποβλήθηκαν οι ασθενείς δεν έχει ερμηνευτική αξία παρά μόνο για τον σχεδιασμό της μελέτης.
- ✓ Μία ασθενής στην οποία αφαιρέθηκαν περισσότεροι των τεσσάρων θετικών λεμφαδένων έχει 2,76 φορές μεγαλύτερο κίνδυνο να πεθάνει όταν οι υπόλοιποι παράγοντες του μοντέλου παραμένουν σταθεροί.
- ✓ Όταν οι 'Factor1' και 'Factor5' αυξηθούν κατά μία μονάδα, ο κίνδυνος θανάτου μειώνεται κατά 1,5 και 1,3 φορές, αντίστοιχα. Συνεπώς, υψηλά επίπεδα έκφρασης των γονιδιακών εκφράσεων των ομάδων 1 και 3 αντιστοιχούν σε χαμηλότερα επίπεδα κινδύνου.
- ✓ Όταν οι 'Factor2' και 'Factor7' αυξηθούν κατά μία μονάδα, τότε ο κίνδυνος θανάτου αυξάνεται κατά 1,6 και 1,3 φορές, αντίστοιχα. Στην ταυτόχρονη αύξηση των δύο συνιστωσών, δεδομένου ότι αντιπροσωπεύουν από κοινού την $2^{η}$ ομάδα γονιδιακών εκφράσεων, ο κίνδυνος αυξάνεται κατά 2,1 φορές. Συνεπώς, υψηλά επίπεδα έκφρασης των γονιδιακών εκφράσεων της $2^{ης}$ ομάδας αντιστοιχούν σε υψηλή αναλογία κινδύνου.

7.6.2 Ελεύθερος νόσου χρόνος

Όσον αφορά στον χρόνο DFS καταλήγουμε, συγκεντρωτικά, στα ακόλουθα συμπεράσματα για το συγκεκριμένο δείγμα ασθενών που εξετάζουμε.

Θεωρώντας τη μη παραμετρική εκτίμηση Kaplan-Meier, και θεωρώντας ως κριτήριο επιλογής ελέγχου την υπόθεση αναλογικού κινδύνου, δεν υπάρχει στατιστικώς σημαντική διαφορά στην πορεία των συναρτήσεων επιβίωσης για τις ομάδες που διαμορφώνονται ως προς τις κατηγορίες των παραγόντων 'Group', 'Age', 'Menopausal', 'Size', 'Surgery', 'Interval' 'ER', 'PgR', 'RT', και 'HT'. Αντιθέτως, οι ομάδες που διαμορφώνονται για τα επίπεδα των παραγόντων 'Grade' και 'Nodes' σημειώνουν στατιστική διαφοροποίηση της συνάρτησης επιβίωσης, σε επίπεδο σημαντικότητας 5%.

Εστιάζοντας στους δύο παράγοντες για τους οποίους εντοπίστηκε διαφορά στην επιβίωση των κατηγοριών τους, και εφαρμόζοντας ελέγχους ως προς την στρωματοποίηση και την τάση, προκύπτουν τα εξής συμπεράσματα:

- ✓ Για τις γυναίκες που ανήκουν στην κατηγορία 'III-Αδιαφοροποίητο' του παράγοντα διαφοροποίησης του όγκου, με την αύξηση του πλήθους των λεμφαδένων που αφαιρέθηκαν κατά την χειρουργική επέμβαση, μειώνεται η πιθανότητα επιβίωσης.
- ✓ Όταν σε ασθενή αφαιρέθηκαν περισσότεροι των τεσσάρων μασχαλιαίοι λεμφαδένες, η συνάρτηση επιβίωσης διαφοροποιείται για τους παράγοντες 'Grade' και 'ER'. Όταν αυξάνεται η κατηγορία διαφοροποίησης του όγκου, μειώνεται η πιθανότητα επιβίωσης, ενώ καλύτερη πρόγνωση σημειώνεται στις γυναίκες που είναι θετικές σε οιστρογονικούς υποδοχείς.

Επιπλέον, θεωρώντας το *μοντέλο αναλογικού κινδύνου του Cox*, οι παράγοντες οι οποίοι συμπεριλαμβάνονται στο μοντέλο είναι οι 'Group', 'Grade', 'Nodes', 'Factor2', 'Factor5' και 'Factor7'. Από τις συνιστώσες που αντιστοιχούν στις γονιδιακές εκφράσεις, ο 'Factor5' αντιστοιχεί στην ομάδα 3 των γονιδιακών εκφράσεων, ενώ οι 'Factor2' και 'Factor7' αντιπροσωπεύουν την $2^{η}$ ομάδα. Οι υπόλοιποι υπό εξέταση παράγοντες δεν προκύπτουν στατιστικά σημαντικοί, σύμφωνα με την μέθοδο επιλογής μοντέλου Backward Selection και για το συγκεκριμένο σύνολο δεδομένων. Αξιοσημείωτο είναι το γεγονός ότι στο μοντέλο του χρόνου DFS δεν προκύπτει σημαντικός παράγοντας ο οποίος να αντιστοιχεί στην $1^{η}$ ομάδα των γονιδιακών εκφράσεων.

Σύμφωνα με τις εκτιμήσεις του μοντέλου, καταλήγουμε στα ακόλουθα συμπεράσματα:

- ✓ Η αξία του παράγοντα της φαρμακευτικής αγωγής εντοπίζεται μόνο στο σχεδιασμό της μελέτης, και περεταίρω συμπεράσματα δεν έχει νόημα να εξαχθούν.

- ✓ Μία γυναίκα η οποία ανήκει στη κατηγορία διαφοροποίησης του όγκου 'III-Αδιαφοροποίητο' διατρέχει κατά 1,8 φορές μεγαλύτερο κίνδυνο επανεμφάνισης της νόσου ή θανάτου από αυτή σε σχέση με μία γυναίκα με διαφοροποίηση 'I-II', δεδομένου ότι οι υπόλοιποι παράγοντες παραμένουν σταθεροί.
- ✓ Μία ασθενής στην οποία αφαιρέθηκαν περισσότεροι των τεσσάρων θετικών λεμφαδένων έχει 2,8 φορές μεγαλύτερο κίνδυνο επανεμφάνισης του όγκου, μετάστασης ή θανάτου, όταν οι υπόλοιποι παράγοντες του μοντέλου παραμένουν σταθεροί.
- ✓ Όταν ο 'Factor5' αυξηθεί κατά μία μονάδα, τότε ο κίνδυνος επανεμφάνισης καρκίνου του μαστού ή ο θάνατος από τη νόσο μειώνεται κατά 1,5 φορές. Συνεπώς, υψηλά επίπεδα έκφρασης των γονιδιακών εκφράσεων της ομάδας 3 αντιστοιχούν σε χαμηλότερα επίπεδα κινδύνου, και συνεπώς μεγαλύτερο χρονικό διάστημα DFS.
- ✓ Όταν οι 'Factor2' και 'Factor7' αυξηθούν κατά μία μονάδα, τότε ο κίνδυνος υποτροπής ή θανάτου αυξάνεται κατά 1,22 και 1,35 φορές, αντίστοιχα. Στην ταυτόχρονη αύξηση των δύο συνιστωσών, δεδομένου ότι αντιπροσωπεύουν από κοινού την $2^{η}$ ομάδα γονιδιακών εκφράσεων, ο κίνδυνος υποτροπής ή θανάτου αυξάνεται κατά 1,66 φορές. Συνεπώς, υψηλά επίπεδα έκφρασης των γονιδιακών εκφράσεων της $2^{ης}$ ομάδας αντιστοιχούν σε υψηλή αναλογία κινδύνου.

7.6.3 Σύγκριση αποτελεσμάτων

Όπως προκύπτει από την προηγούμενη ανάλυση, οι δύο χρόνοι που μελετήθηκαν, επιβίωσης και DFS, παρουσιάζουν παρόμοια συμπεριφορά.

- ✓ Ως προς την μη παραμετρική μελέτη, μέσω του εκτιμητή *Kaplan-Meier*, και στις δύο περιπτώσεις προκύπτουν ως σημαντικοί οι παράγοντες του βαθμού διαφοροποίησης του όγκου και του πλήθους των αφαιρούμενων μασχαλιαίων λεμφαδένων.
- ✓ Για τους δύο χρόνους, η διαφοροποίηση της συνάρτησης επιβίωσης ως προς τον το πλήθος των αφαιρούμενων λεμφαδένων φαίνεται να οφείλεται στην ομάδα των γυναικών στις οποίες αφαιρέθηκαν περισσότεροι των τεσσάρων λεμφαδένων. Επιρροή ασκείται και από τον παράγοντα 'ER', για τον χρόνο επιβίωσης σε επίπεδο σημαντικότητας 10%, ενώ για τον χρόνο DFS σε 5%.

- ✓ Επιπλέον, ο χρόνος επιβίωσης επηρεάζεται, επιπρόσθετα των προαναφερόμενων, και από τους παράγοντες του μεγέθους του αφαιρούμενου όγκου και την λήψη ή μη χημειοθεραπείας (α=5%) καθώς και από τον παράγοντα 'PgR'.
- ✓ Τα αποτελέσματα του παράγοντα διαφοροποίησης του όγκου δεν συμπίπτουν για τους δύο χρόνους ως προς τους παράγοντες που φαίνεται να οφείλονται για αυτή τη διαφοροποίηση. Για τον χρόνο επιβίωσης, διαφοροποίηση σημειώνεται για την ομάδα 'I-II' από τον παράγοντα 'HT' και για την ομάδα 'III-Αδιαφοροποίητο' από το μέγεθος του όγκου, ενώ για τον χρόνο DFS διαφοροποίηση σημειώνεται μόνο για την κατηγορία 'III-Αδιαφοροποίητο' από το πλήθος αφαιρούμενων λεμφαδένων.
- ✓ Λαμβάνοντας υπόψη το μοντέλο PH, τα μοντέλα που αντιστοιχούν στους δύο χρόνους έχουν κοινούς τους παράγοντες 'Group', 'Nodes', 'Factor2', 'Factor5' και 'Factor7'.
- ✓ Το μοντέλο του χρόνου επιβίωσης περιλαμβάνει επιπλέον τον παράγοντα 'Factor1', ο οποίος αντιπροσωπεύει την $1^{η}$ ομάδα γονιδιακών εκφράσεων, ενώ το μοντέλο του χρόνου DFS δεν περιλαμβάνει κάποιο αντίστοιχο παράγοντα για την συγκεκριμένη ομάδα.
- ✓ Το μοντέλο του χρόνου DFS περιλαμβάνει, επιπλέον, τον παράγοντα του βαθμού διαφοροποίησης του όγκου ο οποίος προκύπτει ως σημαντικός και μέση της μη παραμετρικής μελέτης. Ωστόσο, το μοντέλο του χρόνου επιβίωσης δεν 'πιάνει' αυτόν τον παράγοντα.
- ✓ Παρατηρώντας τις αναλογίες κινδύνου για κάθε μοντέλο, είναι εμφανές ότι οι εκτιμήσεις των παραγόντων που εμφανίζονται και στα δύο μοντέλα είναι παρόμοιες, με τον 'Factor2' να έχει ίσως πιο έντονη επίδραση στη περίπτωση που μελετάμε τον χρόνο επιβίωσης.

Τέλος, επισημαίνουμε ότι τα αποτελέσματα και τα συμπεράσματα που εξήχθησαν από την παραπάνω ανάλυση αφορούν το συγκεκριμένο δείγμα ασθενών, γεγονός που σημαίνει ότι τα αποτελέσματα αυτά μπορεί να διαφέρουν σε σχέση με κάποιο άλλο σύνολο δεδομένων.

Π Α Ρ Α Ρ Τ Η Μ Α Τ Α

Α. Διαγράμματα και Πίνακες

Α.1 Στον πίνακα που ακολουθεί παρουσιάζεται συνοπτική περιγραφή των γονιδιακών εκφράσεων που μελετώνται στην παρούσα εργασία βάσει της ονομασίας τους.

Ονομασία	Περιγραφή
PPIA	peptidylprolyl isomerase A (cyclophilin A)
UBE2c	ubiquitin-conjugating enzyme E2C
ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)
TUBB3	tubulin, beta 3
MUC1	mucin 1, cell surface associated
ALCAM	activated leukocyte cell adhesion molecule
SPP1	secreted phosphoprotein 1
CD3D	CD3d molecule, delta (CD3-TCR complex)
MMP7	matrix metallopeptidase 7 (matrilysin, uterine)
ABAT	4-aminobutyrate aminotransferase
AKR1C3	aldo-keto reductase family 1, member C3 (3-alpha hydroxysteroid dehydrogenase, type II)
BIRC5	baculoviral IAP repeat containing 5
CXCL12	chemokine (C-X-C motif) ligand 12
DHCR7	7-dehydrocholesterol reductase
EGFR	epidermal growth factor receptor
ERBB3	v-erb-b2 erythroblastic leukemia viral oncogene homolog 3 (avian)
ERBB4	v-erb-a erythroblastic leukemia viral oncogene homolog 4 (avian)
IL6ST	interleukin 6 signal transducer (gp130, oncostatin M receptor
PTGER3	prostaglandin E receptor 3 (subtype EP3)
PVALB	myoglobin
SFRP1	secreted frizzled-related protein 1
STC2	stanniocalcin 2
VEGFA	vascular endothelial growth factor A
Herstatin	---
VEGFB	vascular endothelial growth factor B
VEGFC	vascular endothelial growth factor C
VEGFR1	fms-related tyrosine kinase 1 (vascular endothelial growth factor/vascular permeability factor receptor)
VEGFR2	kinase insert domain receptor (a type III receptor tyrosine kinase))

Ονομασία	Περιγραφή
VEGFR3	fms-related tyrosine kinase 4
MMP1	matrix metallopeptidase 1 (interstitial collagenase)
IGKC	immunoglobulin kappa constant
TP53	Tumor protein p53 (Li-Fraumeni syndrome)
MLPH	melanophilin
TOP2A	topoisomerase (DNA) II alpha 170kDa
RACGAP1	Rac GTPase activating protein 1
CHPT1	choline phosphotransferase 1
CXCL13	chemokine (C-X-C motif) ligand 13

A.2 Ο ακόλουθος πίνακας αντιστοιχεί στην ολική μεταβλητότητα που ερμηνεύεται από το μοντέλο που προκύπτει από την ανάλυση των γονιδιακών εκφράσεων κατά παράγοντες.

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6,448	17,426	17,426	6,448	17,426	17,426	5,167	13,965	13,965
2	5,467	14,775	32,200	5,467	14,775	32,200	4,195	11,338	25,303
3	4,087	11,047	43,248	4,087	11,047	43,248	3,879	10,483	35,787
4	2,361	6,382	49,630	2,361	6,382	49,630	2,753	7,442	43,228
5	2,122	5,734	55,364	2,122	5,734	55,364	2,374	6,417	49,645
6	1,828	4,940	60,305	1,828	4,940	60,305	2,360	6,379	56,025
7	1,207	3,263	63,568	1,207	3,263	63,568	2,343	6,333	62,357
8	1,045	2,824	66,392	1,045	2,824	66,392	1,493	4,035	66,392
9	,999	2,700	69,092						
10	,948	2,562	71,654						
11	,913	2,467	74,121						

A.3 Ο πίνακας που ακολουθεί περιλαμβάνει τις περιστρεφόμενες συνιστώσες της ανάλυσης κατά παράγοντας, που προκύπτουν μέσω της μεθόδου περιστροφής Quartimax.

	Συνιστώσες								
	1	2	3	4	5	6	7	8	9
PPIA	-,008	,449	,093	,257	,138	-,098	,051	,523	,151
UBE2c	-,147	,886	-,033	,124	,103	-,002	,082	,060	,055
MMP1	-,236	,124	,089	,753	,096	-,032	,151	,022	,004
IGKC	-,199	,164	,094	-,015	,784	,170	,083	-,081	,001
TP53	,296	,238	,143	,283	,145	,393	-,014	,240	-,259

	Συνιστώσες								
	1	2	3	4	5	6	7	8	9
MLPH	,715	-,099	,127	-,187	-,106	-,212	,332	,188	-,041
TOP2A	-,005	,885	-,023	,068	,053	,016	,065	,060	-,037
RACGAP1	,040	,894	,045	,099	,115	,028	,003	,121	,019
CHPT1	,630	-,006	,011	-,040	-,165	,153	-,279	-,103	-,056
CXCL13	-,199	,002	-,057	,040	,711	-,040	,006	,148	-,025
ERBB2	-,056	,116	,172	,185	,099	-,054	,885	,017	-,050
TUBB3	-,252	,152	,065	,688	-,017	,129	,254	,117	-,002
MUC1	,590	-,156	-,069	,213	,005	-,009	,164	,350	-,066
ALCAM	,359	,106	-,165	,222	,225	-,027	,197	,279	,233
SPP1	,004	,248	-,024	,773	-,019	,081	-,096	-,157	,168
CD3D	-,120	,145	,165	,007	,863	,047	,090	-,048	-,051
MMP7	-,106	,079	-,004	,305	,099	,728	-,044	-,040	,078
ABAT	,828	,015	,100	-,018	-,020	-,145	,072	-,047	-,176
AKR1C3	,161	,022	,434	-,026	,024	,429	,103	-,115	,350
BIRC5	-,150	,866	,035	,131	,035	-,076	-,010	-,034	,114
CXCL12	,280	-,443	,407	-,071	,219	,258	,007	,039	-,030
DHCR7	-,027	,322	,151	,141	-,100	,102	,114	,098	,692
EGFR	-,205	-,261	,414	,106	,011	,676	-,034	-,004	,084
ERBB3	,659	,250	,102	-,180	-,221	,017	,272	,021	,130
ERBB4	,678	-,092	,091	-,220	-,110	-,127	-,113	,109	,117
IL6ST	,795	-,031	,222	-,176	-,011	-,010	-,044	-,046	-,122
PTGER3	,607	-,273	,238	,098	-,006	,082	-,143	-,159	,344
PVALB	,196	,204	,085	-,106	-,030	-,117	-,073	,718	-,009
SFRP1	-,139	-,107	,207	-,214	,018	,774	-,183	-,152	-,043
STC2	,726	-,131	,009	-,008	-,115	-,013	-,316	,061	,074
VEGFA	-,106	,267	,413	,455	-,250	,054	,168	,118	-,186
Herstatin	-,087	,062	,195	,134	,082	-,126	,848	-,058	,161
VEGFB	,400	,039	,595	,172	,091	-,025	,001	-,188	,074
VEGFC	,143	-,139	,719	,246	,084	,091	,100	,146	,228
VEGFR1	,120	,130	,838	,114	,014	,051	,083	-,038	-,127
VEGFR2	,083	,052	,862	-,043	-,016	,044	,090	,069	,023
VEGFR3	,089	-,075	,796	-,182	,045	,111	,021	,064	,005

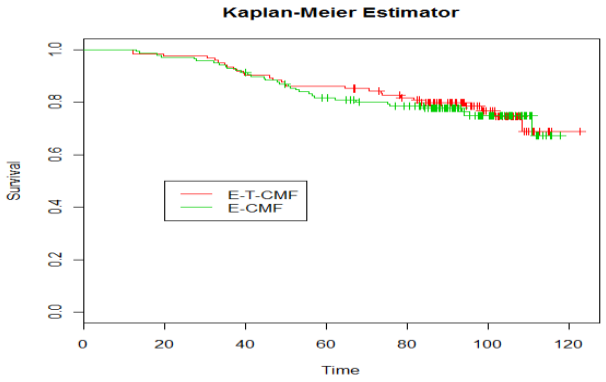
A.4 Ο ακόλουθος πίνακας αντιστοιχεί στις εκτιμήσεις *Kaplan-Meier* του χρόνου επιβίωσης.

time	n.risk	n.event	survival	std.err	time	n.risk	n.event	survival	std.err
12.0	258	1	0.996	0.00387	48.4	226	1	0.876	0.02054
12.1	257	1	0.992	0.00546	48.9	225	1	0.872	0.02082
13.0	256	1	0.988	0.00667	49.8	223	1	0.868	0.02109
13.6	255	1	0.984	0.00769	50.1	222	1	0.864	0.02135
18.1	254	1	0.981	0.00858	50.4	221	1	0.860	0.02161
19.2	253	1	0.977	0.00938	51.0	220	1	0.856	0.02186

time	n.risk	n.event	survival	std.err	time	n.risk	n.event	survival	std.err
19.8	252	1	0.973	0.01011	52.4	219	1	0.852	0.02211
27.3	251	1	0.969	0.01079	53.2	218	1	0.848	0.02235
27.8	250	1	0.965	0.01142	55.8	217	1	0.845	0.02259
30.4	249	1	0.961	0.01202	56.4	216	1	0.841	0.02282
32.0	248	1	0.957	0.01258	56.9	215	1	0.837	0.02305
32.3	247	1	0.953	0.01311	62.2	212	1	0.833	0.02327
33.1	246	1	0.950	0.01362	64.5	211	1	0.829	0.02349
33.6	245	1	0.946	0.01410	67.2	205	1	0.825	0.02373
34.9	244	1	0.942	0.01457	70.4	203	1	0.821	0.02395
35.0	243	1	0.938	0.01502	73.5	200	1	0.817	0.02418
35.4	242	1	0.934	0.01545	73.8	199	1	0.813	0.02441
35.4	241	1	0.930	0.01586	75.2	198	1	0.808	0.02463
37.1	240	1	0.926	0.01626	75.8	197	1	0.804	0.02484
37.4	239	1	0.922	0.01665	78.3	194	1	0.800	0.02506
37.8	238	1	0.919	0.01702	81.6	190	1	0.796	0.02528
38.9	237	1	0.915	0.01739	83.5	175	1	0.791	0.02554
39.5	236	1	0.911	0.01774	84.2	170	1	0.787	0.02581
39.6	235	1	0.907	0.01808	93.6	112	1	0.780	0.02652
40.6	233	1	0.903	0.01842	94.2	108	1	0.773	0.02724
41.3	232	1	0.899	0.01875	94.5	107	1	0.765	0.02792
44.6	231	1	0.895	0.01907	98.3	92	1	0.757	0.02883
44.8	230	1	0.891	0.01938	101.7	69	1	0.746	0.03043
45.9	229	1	0.888	0.01968	108.5	31	1	0.722	0.03778
46.8	228	1	0.884	0.01997	110.9	18	1	0.682	0.05285
47.9	227	1	0.880	0.02026					

A.5 Στη συνέχεια παρατίθενται για κάθε παράγοντα οι έλεγχοι για την ισότητα των συναρτήσεων επιβίωσης που διαμορφώνονται από τις ομάδες των επιπέδων τους ως προς τον χρόνο επιβίωσης, καθώς και οι αντίστοιχες γραφικές παραστάσεις.

Παράγοντας Group

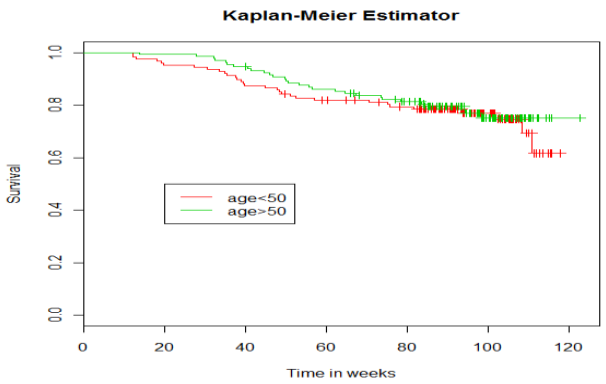


Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	,077	1	,781
Breslow (Generalized Wilcoxon)	,191	1	,662
Tarone-Ware	,143	1	,706
Flemming-Harrington(ρ=1,π=0)	,095	1	,758

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

Παράγοντας Age

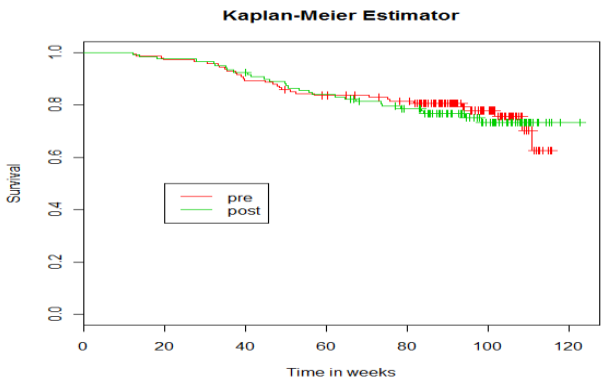


Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	,297	1	,586
Breslow (Generalized Wilcoxon)	,289	1	,591
Tarone-Ware	,237	1	,626
Flemming-Harrington(ρ=1,π=0)	,413	1	,520

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

Παράγοντας Menopausal

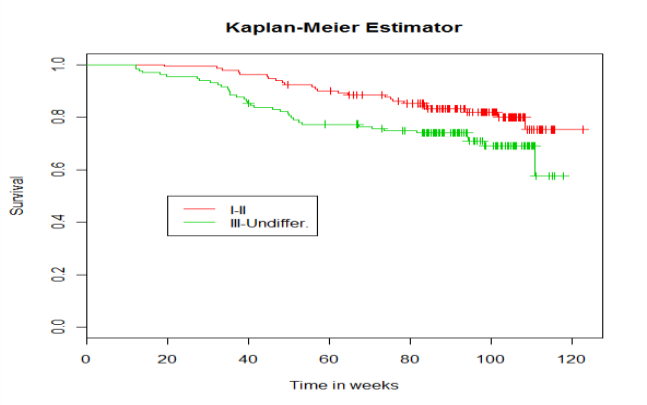


Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	,052	1	,820
Breslow (Generalized Wilcoxon)	,175	1	,675
Tarone-Ware	,156	1	,693
Flemming-Harrington(ρ=1,π=0)	,044	1	,833

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

Παράγοντας Grade

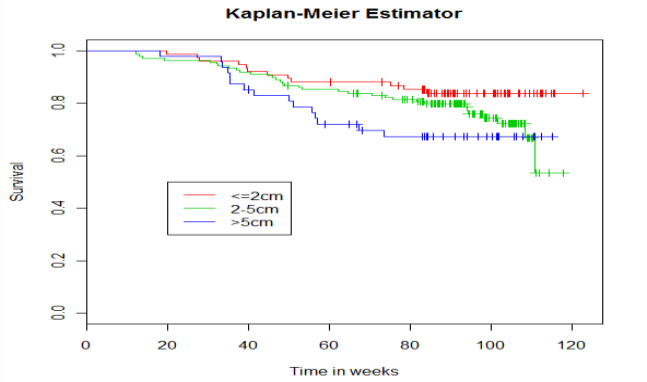


Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	5,007	1	,025
Breslow (Generalized Wilcoxon)	5,822	1	,016
Tarone-Ware	5,424	1	,020
Flemming-Harrington(ρ=1,π=0)	5,74	1	,016

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

Παράγοντας Size

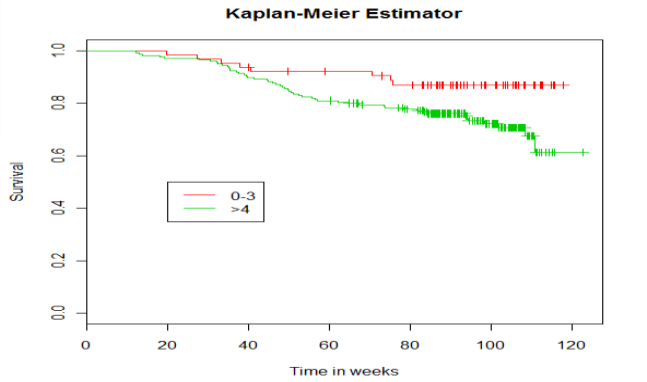


Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	5,138	2	,077
Breslow (Generalized Wilcoxon)	5,084	2	,079
Tarone-Ware	5,095	2	,078
Flemming-Harrington(ρ=1,π=0)	5,000	2	,081

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

Παράγοντας Nodes

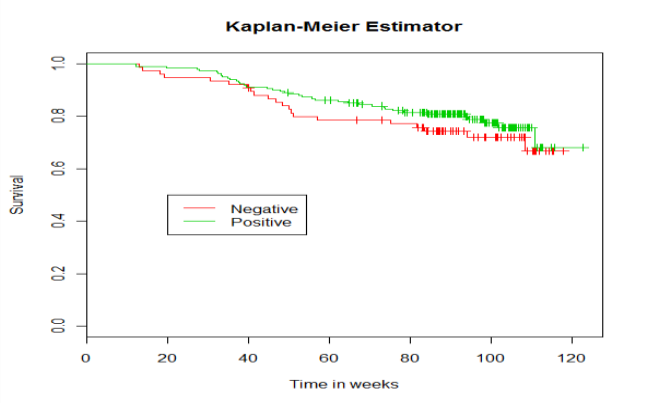


Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	5,035	1	,025
Breslow (Generalized Wilcoxon)	3,780	1	,052
Tarone-Ware	4,241	1	,039
Flemming-Harrington(ρ=1,π=0)	4,760	1	,029

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

Παράγοντας ERIhc

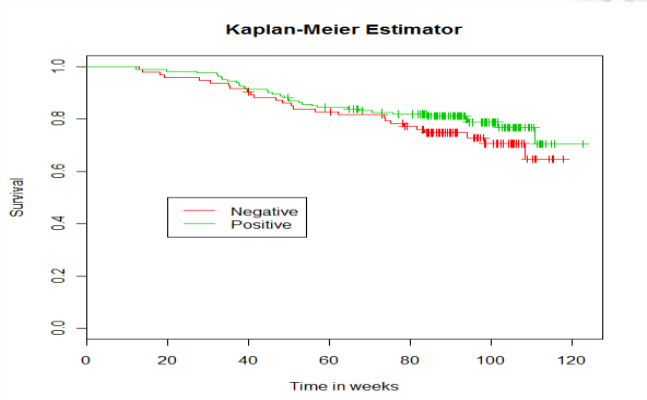


Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	,928	1	,335
Breslow (Generalized Wilcoxon)	1,269	1	,260
Tarone-Ware	1,162	1	,281
Flemming-Harrington(ρ=1,π=0)	1,030	1	,309

Test of equality of survival distributions for the different levels of ER (IHC).

Παράγοντας PgRIhc

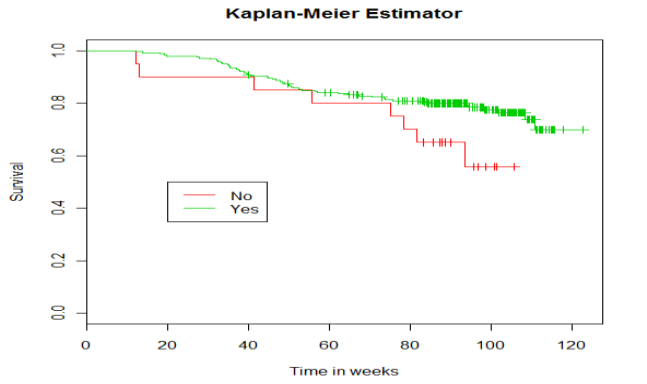


Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	1,296	1	,255
Breslow (Generalized Wilcoxon)	1,089	1	,297
Tarone-Ware	1,203	1	,273
Flemming-Harrington(ρ=1,π=0)	1,220	1	,270

Test of equality of survival distributions for the different levels of PgR (IHC).

Παράγοντας HT

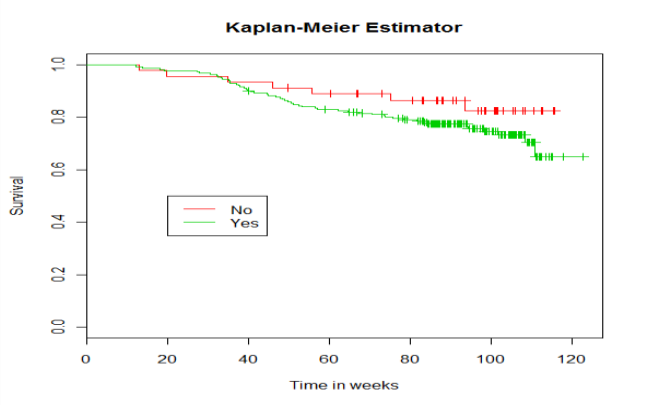


Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	3,472	1	,062
Breslow (Generalized Wilcoxon)	2,534	1	,111
Tarone-Ware	2,955	1	,086
Flemming-Harrington(ρ=1,π=0)	3,100	1	,078

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

Παράγοντας RT

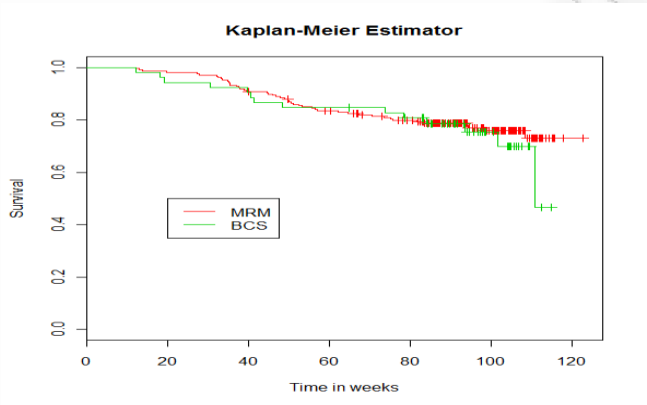


Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	1,709	1	,191
Breslow (Generalized Wilcoxon)	1,372	1	,241
Tarone-Ware	1,485	1	,223
Flemming-Harrington(ρ=1,π=0)	1,590	1	,207

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

Παράγοντας Surgery

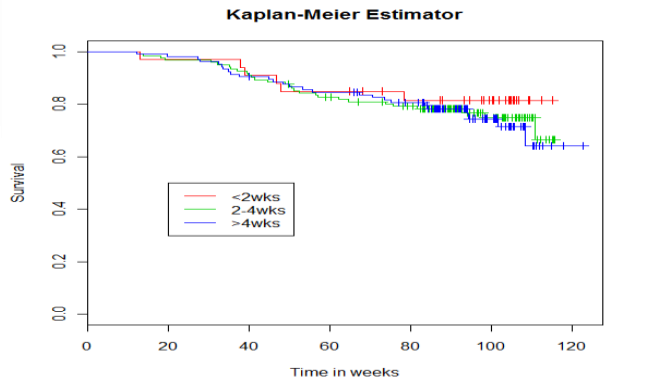


Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	,317	1	,573
Breslow (Generalized Wilcoxon)	,030	1	,862
Tarone-Ware	,083	1	,774
Flemming-Harrington(ρ=1,π=0)	,267	1	,606

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

Παράγοντας Interval



Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	,744	2	,689
Breslow (Generalized Wilcoxon)	,292	2	,864
Tarone-Ware	,433	2	,805
Flemming-Harrington(ρ=1,π=0)	,600	2	,734

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

A.6 Στον ακόλουθο πίνακα παρουσιάζεται η κωδικοποίηση που εφαρμόζεται προκειμένου να δημιουργηθούν οι βωβές μεταβλητές για τις διάφορες κατηγορίες των κλινικοπαθολογικών παραγόντων που εξετάζονται.

		Frequency	(1)	(2)
ERIhc	0=negative	73	0	
	1=positive	174	1	
PgRIhc	0=negative	91	0	
	1=positive	156	1	
Group	0=E-T-CMF	116	0	
	1=E-CMF	131	1	
Age	0=<50	124	0	
	1=>=50	123	1	
Menopausal	0=pre	136	0	
	1=post	111	1	
Surgery	1=MRM	197	0	
	2=BCS	50	1	
Interval	1=<2 wks	32	0	0
	2=2-4 wks	115	1	0
	3=>4wks	100	0	1
Grade	1=I-II	128	0	
	2=III-Undifferentiated	119	1	
Size	1=<=2cm	72	0	0
	2=2-5cm	132	1	0
	3=>5cm	43	0	1
Nodes	0=0-3	59	0	
	1=>4	188	1	
RT	0=no	42	0	
	1=yes	205	1	
HT	0=no	20	0	
	1=yes	227	1	

A.7 Οι ακόλουθοι δύο πίνακες αντιστοιχούν στο τελικό βήμα κατά την διαδικασία επιλογής μοντέλου, θεωρώντας τους κλινικοπαθολογικούς παράγοντες και τις συνιστώσες των γονιδιακών εκφράσεων του μοντέλου, αντίστοιχα, ως μία τεχνική εξέτασης της σταθερότητας του μοντέλου.

Variables in the Equation

		B	SE	Wald	df	Sig.	Exp(B)
Step 10	Grade	,544	,267	4,154	1	,042	1,723
	Nodes	,950	,405	5,498	1	,019	2,585
	HT	-,763	,384	3,948	1	,047	,466

Variables in the Equation

		B	SE	Wald	df	Sig.	Exp(B)
Step 6	Factor1	-,334	,124	7,330	1	,007	,716
	Factor2	,432	,136	10,031	1	,002	1,540
	Factor5	-,271	,127	4,571	1	,033	,763
	Factor7	,267	,123	4,732	1	,030	1,306

A.8 Ακολούθως παρουσιάζονται οι ατομικοί έλεγχοι για κάθε παράγοντα του μοντέλου PH, καθώς και ο ολικός έλεγχος, για την ισχύ της υπόθεσης αναλογικού κινδύνου.

	rho	chisq	p
GroupE-CMF	-0.0125	0.00972	0.921
Nodes>4	0.0744	0.33930	0.560
Factor1	0.2173	2.29187	0.130
Factor2	-0.1474	0.74630	0.388
Factor5	-0.0748	0.38593	0.534
Factor7	0,0128	0.01062	0.918
GLOBAL	NA	3.43581	0.752

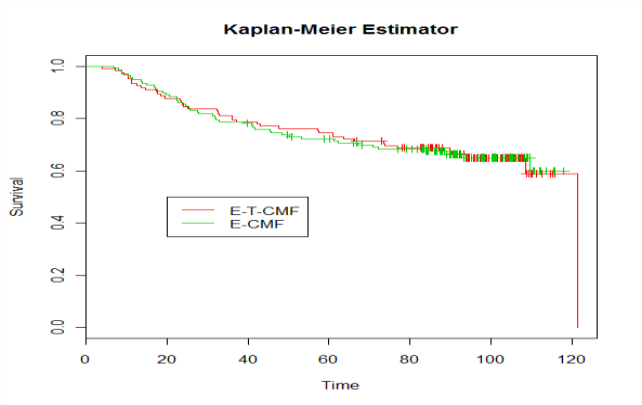
A.9 Ο ακόλουθος πίνακας αντιστοιχεί στις εκτιμήσεις *Kaplan-Meier* του DFS.

time	**n.risk**	**n.event**	**survival**	**std.err**	**time**	**n.risk**	**n.event**	**survival**	**std.err**
4.03	258	1	0,996	0,00387	31,9	212	1	0,818	0,02403
6,92	257	1	0,992	0,00546	32,13	211	1	0,814	0,02423
7,05	256	1	0,988	0,00667	32,3	210	1	0,81	0,02442
8,1	255	1	0,984	0,00769	32,59	209	1	0,806	0,02461
8,72	254	2	0,977	0,00938	32,75	208	1	0,802	0,02479
8,75	252	1	0,973	0,01011	32,79	207	1	0,798	0,02498
9,38	251	1	0,969	0,01079	36,03	206	1	0,795	0,02515
10,26	250	1	0,965	0,01142	36,07	205	1	0,791	0,02533

time	n.risk	n.event	survival	std.err	time	n.risk	n.event	survival	std.err
10,39	249	2	0,957	0,01258	37,18	204	1	0,787	0,0255
10,79	247	1	0,953	0,01311	39,15	203	1	0,783	0,02566
11,25	246	1	0,95	0,01362	41,02	201	1	0,779	0,02583
11,31	245	1	0,946	0,0141	41,18	200	1	0,775	0,02599
11,44	244	1	0,942	0,01457	41,77	199	1	0,771	0,02615
12,39	243	1	0,938	0,01502	42,26	198	1	0,767	0,02631
13,57	242	1	0,934	0,01545	42,98	197	1	0,763	0,02646
13,61	241	1	0,93	0,01586	45,57	196	1	0,76	0,02661
13,97	240	1	0,926	0,01626	45,7	195	1	0,756	0,02676
14,69	239	1	0,922	0,01665	47,57	194	1	0,752	0,0269
14,98	238	1	0,919	0,01702	48,49	193	1	0,748	0,02705
16,75	237	1	0,915	0,01739	50,3	191	1	0,744	0,02719
17,18	236	1	0,911	0,01774	53,21	189	1	0,74	0,02733
17,38	235	1	0,907	0,01808	57,18	188	1	0,736	0,02746
17,61	234	1	0,903	0,01842	57,41	187	1	0,732	0,0276
17,64	233	1	0,899	0,01874	61,02	184	1	0,728	0,02773
18,36	232	1	0,895	0,01906	61,05	183	1	0,724	0,02786
19,02	231	1	0,891	0,01936	61,28	182	1	0,72	0,02799
19,54	230	1	0,888	0,01966	62,2	181	1	0,716	0,02812
19,74	229	1	0,884	0,01996	63,54	180	1	0,712	0,02824
20,52	228	1	0,88	0,02024	66,2	178	1	0,708	0,02837
22,46	227	1	0,876	0,02052	67,25	175	1	0,704	0,02849
22,56	226	1	0,872	0,02079	70,89	173	1	0,7	0,02862
22,59	225	1	0,868	0,02106	72,33	172	1	0,696	0,02874
23,21	224	1	0,864	0,02132	73,51	169	1	0,692	0,02886
23,61	223	1	0,86	0,02157	73,93	168	1	0,688	0,02898
23,74	222	1	0,857	0,02182	77,05	167	1	0,684	0,0291
23,9	221	1	0,853	0,02206	84,03	144	1	0,679	0,02928
23,93	220	1	0,849	0,0223	88,2	119	1	0,673	0,02959
24,56	219	1	0,845	0,02253	88,59	116	1	0,667	0,0299
25,02	218	1	0,841	0,02276	90,33	111	1	0,661	0,03023
25,25	217	1	0,837	0,02298	92,85	96	1	0,655	0,03069
25,67	216	1	0,833	0,0232	93,44	90	1	0,647	0,0312
27,48	215	1	0,829	0,02342	108,49	27	1	0,623	0,03816
27,8	214	1	0,826	0,02362	109,67	21	1	0,594	0,04647
31,34	213	1	0,822	0,02383	121,48	1	1	0	-

A.10 Στη συνέχεια παρατίθενται για κάθε παράγοντα οι έλεγχοι για την ισότητα των συναρτήσεων επιβίωσης που διαμορφώνονται από τις ομάδες των επιπέδων τους ως προς τον χρόνο DFS, καθώς και οι αντίστοιχες γραφικές παραστάσεις.

Παράγοντας Group

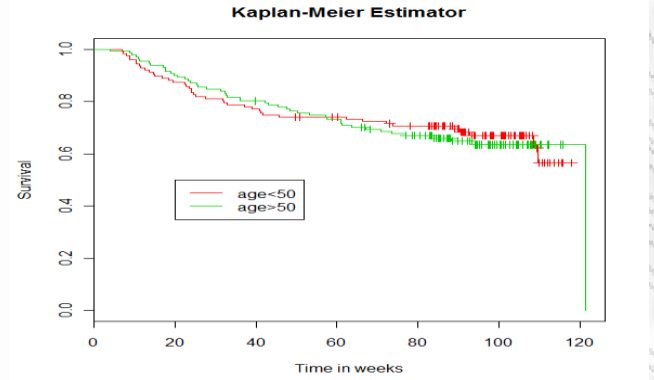


Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	,001	1	,973
Breslow (Generalized Wilcoxon)	,013	1	,908
Tarone-Ware	,008	1	,929
Flemming-Harrington(ρ=1,π=0)	,003	1	,959

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

Παράγοντας Age

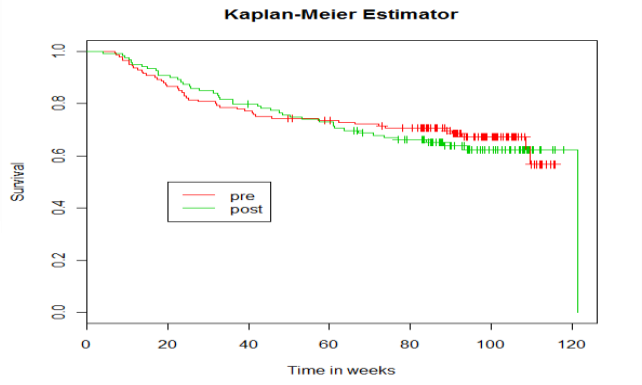


Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	,044	1	,834
Breslow (Generalized Wilcoxon)	,037	1	,847
Tarone-Ware	,063	1	,802
Flemming-Harrington(ρ=1,π=0)	,004	1	,950

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

Παράγοντας Menopausal

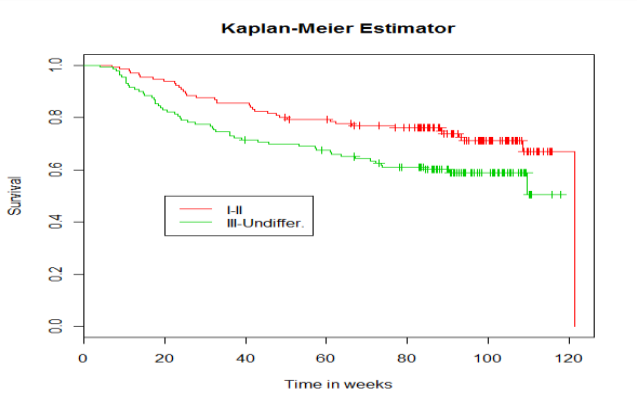


Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	,136	1	,712
Breslow (Generalized Wilcoxon)	,098	1	,754
Tarone-Ware	,151	1	,697
Flemming-Harrington(ρ=1,π=0)	,040	1	,841

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

Παράγοντας Grade

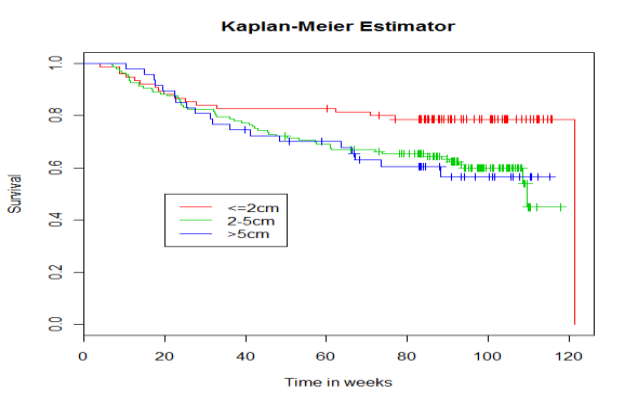


Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	6,158	1	,013
Breslow (Generalized Wilcoxon)	7,089	1	,008
Tarone-Ware	6,690	1	,010
Flemming-Harrington(ρ=1,π=0)	6,780	1	,009

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

Παράγοντας Size

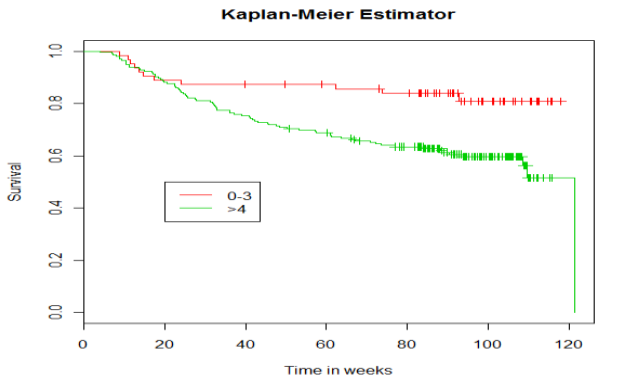


Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	7,231	2	,027
Breslow (Generalized Wilcoxon)	4,827	2	,090
Tarone-Ware	5,805	2	,055
Flemming-Harrington(ρ=1,π=0)	5,900	2	,053

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

Παράγοντας Nodes

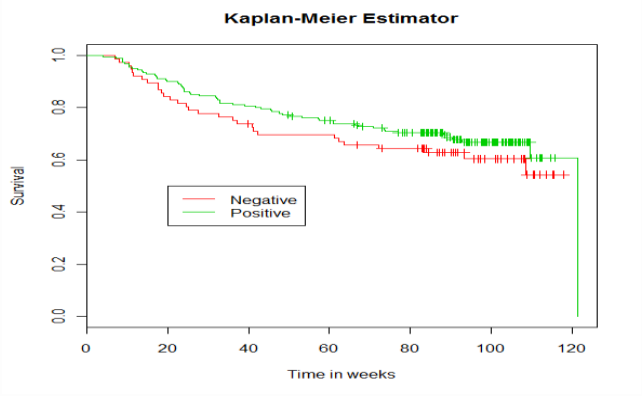


Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	9,108	1	,003
Breslow (Generalized Wilcoxon)	7,514	1	,006
Tarone-Ware	8,242	1	,004
Flemming-Harrington(ρ=1,π=0)	8,050	1	,005

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

Παράγοντας ERIhc

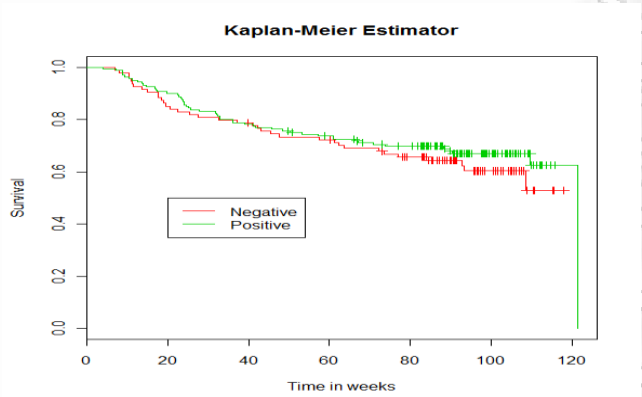


Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	1,262	1	,261
Breslow (Generalized Wilcoxon)	1,412	1	,235
Tarone-Ware	1,348	1	,246
Flemming-Harrington(ρ=1,π=0)	1,420	1	,233

Test of equality of survival distributions for the different levels of ER (IHC).

Παράγοντας PgRIhc

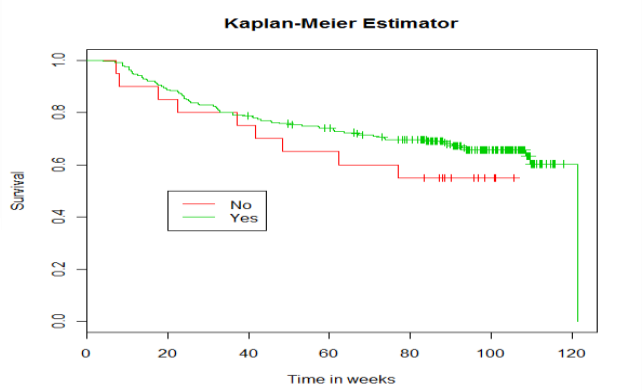


Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	,841	1	,359
Breslow (Generalized Wilcoxon)	,512	1	,474
Tarone-Ware	,629	1	,428
Flemming-Harrington(ρ=1,π=0)	,705	1	,401

Test of equality of survival distributions for the different levels of PgR (IHC).

Παράγοντας HT

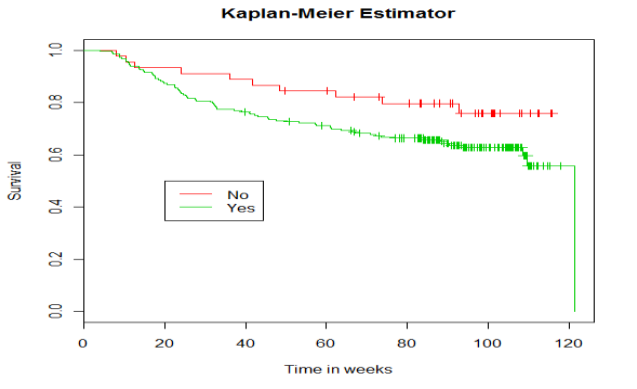


Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	1,262	1	,261
Breslow (Generalized Wilcoxon)	1,304	1	,253
Tarone-Ware	1,303	1	,254
Flemming-Harrington(ρ=1,π=0)	1,210	1	,271

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

Παράγοντας RT

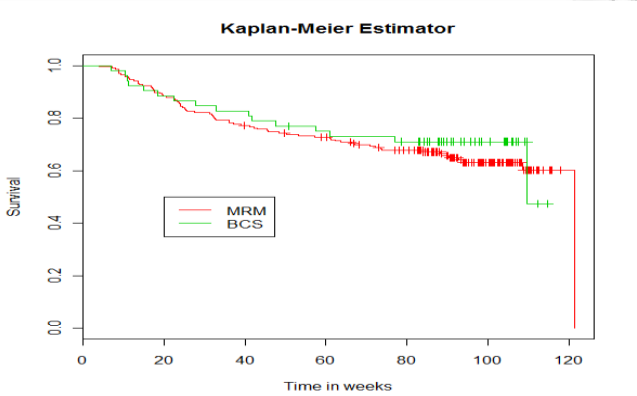


Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	3,422	1	,064
Breslow (Generalized Wilcoxon)	3,264	1	,071
Tarone-Ware	3,317	1	,069
Flemming-Harrington(ρ=1,π=0)	3,510	1	,061

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

Παράγοντας Surgery

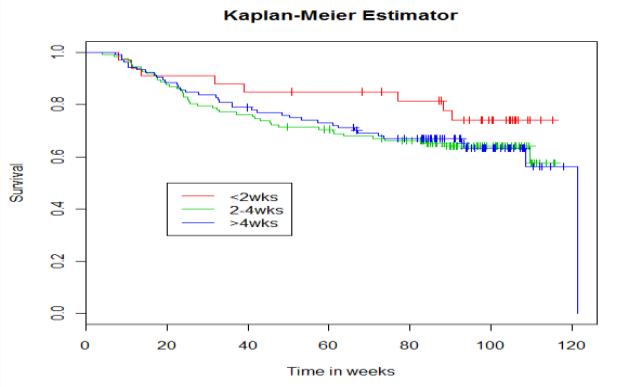


Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	,338	1	,561
Breslow (Generalized Wilcoxon)	,344	1	,557
Tarone-Ware	,386	1	,535
Flemming-Harrington(ρ=1,π=0)	,300	1	,578

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

Παράγοντας Interval



Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	1,882	2	,390
Breslow (Generalized Wilcoxon)	2,042	2	,360
Tarone-Ware	1,970	2	,373
Flemming-Harrington(ρ=1,π=0)	1,900	2	,386

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

A.11 Οι ακόλουθοι δύο πίνακες αντιστοιχούν στο τελικό βήμα κατά την διαδικασία επιλογής μοντέλου, θεωρώντας τους κλινικοπαθολογικούς παράγοντες και τις συνιστώσες των γονιδιακών εκφράσεων του μοντέλου, αντίστοιχα, ως μία τεχνική εξέτασης της σταθερότητας του μοντέλου, για την περίπτωση του χρόνου DFS.

Variables in the Equation

								95,0% CI for Exp(B)	
		B	SE	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 10	Grade	,490	,221	4,931	1	,026	1,633	1,059	2,518
	Size			4,591	2	,101			
	Size(1)	,604	,287	4,414	1	,036	1,829	1,041	3,212
	Size(2)	,579	,354	2,667	1	,102	1,784	,891	3,573
	Nodes	,979	,337	8,442	1	,004	2,663	1,375	5,155

Variables in the Equation

								95,0% CI for Exp(B)	
		B	SE	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 5	Factor1	-,173	,103	2,853	1	,091	,841	,687	1,028
	Factor2	,256	,107	5,678	1	,017	1,292	1,046	1,594
	Factor4	,204	,105	3,807	1	,051	1,227	,999	1,506
	Factor5	-,282	,108	6,744	1	,009	,755	,610	,933
	Factor7	,326	,104	9,781	1	,002	1,386	1,130	1,701

A.12 Ακολούθως παρουσιάζονται οι ατομικοί έλεγχοι για κάθε παράγοντα του μοντέλου PH, καθώς και ο ολικός έλεγχος, για την ισχύ της υπόθεσης αναλογικού κινδύνου.

	rho	chisq	p
GroupE-CMF	0.00781	0.00539	0.941
GradeIII-Undiffer.	-0.09327	0.79893	0.371
Nodes>4	0.15388	2.17082	0.141
Factor2	-0.18143	2.35150	0.125
Factor5	0.08975	0.80603	0.369
Factor7	0.00428	0.00226	0.962
GLOBAL	NA	7.22034	0.301

B. Ρουτίνες της R

Στο παρόν παράρτημα παρατίθενται οι ρουτίνες τις R που εφαρμόστηκαν στην παρούσα ανάλυση προκειμένου να εξαχθούν τα αντίστοιχα αποτελέσματα. Σημειώνεται ότι όσον αφορά την ανάλυση επιβίωσης για τους χρόνους επιβίωσης και DFS παρουσιάζονται οι ρουτίνες που αντιστοιχούν στους πρώτους, δεδομένου ότι με κατάλληλη αντικατάσταση των μεταβλητών που αντιστοιχούν στο χρόνο και στην ένδειξη λογοκρισίας προκύπτουν άμεσα τα αποτελέσματα του χρόνου DFS.

Τα δεδομένα 'διαβάζονται' από αρχείο SPSS, φορτώνοντας το πακέτο '*foreign*' (library(foreign)) και μέσω της εντολής '*read.spss*'. Στις ρουτίνες που ακολουθούν με α συμβολίζεται ο πίνακας των δεδομένων.

B.1 Shilhouette Plot

```
library(cluster)

pamdata=pam(t(a), k=3, metric="euclidean", medoids=NULL, diss=F, stand=TRUE,
cluster.only=FALSE, do.swap=TRUE);pamdata

summary(pamdata)
plot(pamdata)
```

B.2 Heatmap

```
a = a[rowSums(!is.na(a))!=0, colSums(!is.na(a))!=0]

library(caTools)
library(bitops)
library(grid)
library(gplots)

heatmap.2(as.matrix(a),Rowv=TRUE,Colv=TRUE,distfun=dist,
hclustfun=hclust,dedrogram="both",symm = FALSE,
scale="none",revC=TRUE,symbreaks=FALSE,trace="none",
col=greenred(75),
colsep=c(11,22),sepcolor="green",sepwidth=0.1,key=TRUE,keysize =1.5,
density.info="density",denscol="green",xlab="Genes",
ylab="Samples")
```

B.3 Δείκτες Ισχύος Ομαδοποίησης

```
library(mclust)
library(cluster)
library(kohonen)
library(class)
library(clValid)

library(Hmisk)
library(survival)
library(splines)

b<-impute(a,what="mean")

inter<-clValid(t(b), 3:3, clMethods = "kmeans",
validation ="internal", maxitems = 37, metric = "euclidean",neighbSize=10,)
summary(inter)
inter<-clValid(t(b), 3:3, clMethods = "kmeans",
validation ="internal", maxitems = 37, metric = "euclidean",neighbSize=10,)
summary(inter)

stab<-clValid(t(b), 3:3, clMethods = "kmeans",
validation ="stability", maxitems = 37, metric = "euclidean")
summary(stab)
b<-as.data.frame(b)
gene<-names(b)   ###krataei mono ta onomata twn gonidiwn#####
group=c("kinases","misc","misc","cyto/chemokines","misc","cyto/chemokines",
"kinases","cyto/chemokines","kinases","cyto/chemokines","kinases","misc",
"misc","cyto/chemokines","cyto/chemokines","cyto/chemokines","misc","misc","misc",
"growth factor","cyto/chemokines","misc","kinases","kinases","kinases",
"cyto/chemokines","cyto/chemokines","misc","growth factor","growth factor",
"growth factor","growth factor","growth factor","growth factor","kinases",
"kinases","kinases")    ###gnwsth omadopoihsh####
group=as.data.frame(group)
fc<-tapply(gene,group,c); ####syndeei to gonidio me thn omada tou####
biol<-clValid(t(b), 3:3, clMethods = "kmeans",
validation ="biological", metric = "euclidean",
annotation = fc)
summary(biol)
```

B.4 Εκτίμηση Kaplan-Meier

```
library(splines)
library(survival)

brcan<-survfit(Surv(Survival,Death)~1,data=a,conf.type="none")
summary(brcan)
```

```
plot(brcan,xlab='Time in weeks',ylab='Survival',main='Kaplan-Meier Estimator')
```

Διάμεσος χρόνος

```
med<-survfit(Surv(Survival,Death)~1,data=a,conf.int=0.95,conf.type="plain")
med
```

Συνάρτηση επιβίωσης και διαστήματα εμπιστοσύνης

```
brcan1<-survfit(Surv(Survival,Death)~1,data=a,conf.int=0.95,conf.type="plain")
summary(brcan1)
plot(brcan1,xlab='Time in weeks',ylab='Survival',main='Kaplan-Meier Estimator (CI
type:plain)')

brcan2<-survfit(Surv(Survival,Death)~1,data=a,conf.int=0.95,conf.type="log-log")
summary(brcan2)
plot(brcan2,xlab='Time in weeks',ylab='Survival',main='Kaplan-Meier Estimator (CI
type:log-log)')

brcan3<-survfit(Surv(Survival,Death)~1,data=a,conf.int=0.95,conf.type="log")
summary(brcan3)
plot(brcan3,xlab='Time in weeks',ylab='Survival',main='Kaplan-Meier Estmator (CI
type:log)')
```

B.5 Σύγκριση συναρτήσεων επιβίωσης ως προς τα κλινικά χαρακτηριστικά

Παράγοντας Group

```
brcan_group<-survfit(Surv(Survival,Death)~Group,data=a,conf.type="none")

summary(brcan_group)

plot(brcan_group,xlab='Time',ylab='Survival',main='Kaplan-Meier Estimator',
col=2:3)
legend(20,0.5,c("E-T-CMF","E-CMF"),lty=1:1,col=2:3)

br1<-survdiff(Surv(Survival,Death)~Group,data=a);br1

#test Flemming-Harrington (ρ=1,π=0) - προσέγγιση Peto Peto#
fh1<-survdiff(Surv(Survival,Death)~Group,data=a,rho=1);fh1
```

Παράγοντας Age

```
brcan_age<-survfit(Surv(Survival,Death)~Age,data=a,conf.type="none")

summary(brcan_age)
```

```
plot(brcan_age,xlab='Time in weeks',ylab='Survival',main='Kaplan-Meier Estimator',
col=2:3)
legend(20,0.5,c("age<50","age>50"),lty=1:1,col=2:3)

br2<-survdiff(Surv(Survival,Death)~Age,data=a);br2

#test Flemming-Harrington (ρ=1,π=0) - προσέγγιση Peto Peto#
fh2<-survdiff(Surv(Survival,Death)~Age,data=a,rho=1);fh2
```

Παράγοντας Menopausal

```
brcan_Menopausal<-survfit(Surv(Survival,Death)~Menopausal,data=a,conf.type="none")

summary(brcan_Menopausal)

plot(brcan_Menopausal,xlab='Time in weeks',ylab='Survival',main='Kaplan-Meier Estimator',
col=2:3)
legend(20,0.5,c("pre","post"),lty=1:1,col=2:3)

br3<-survdiff(Surv(Survival,Death)~Menopausal,data=a);br3

#test Flemming-Harrington (ρ=1,π=0) - προσέγγιση Peto Peto#
fh3<-survdiff(Surv(Survival,Death)~Menopausal,data=a,rho=1);fh3
```

Παράγοντας Grade

```
brcan_Grade<-survfit(Surv(Survival,Death)~Grade,data=a,conf.type="none")

summary(brcan_Grade)

plot(brcan_Grade,xlab='Time in weeks',ylab='Survival',main='Kaplan-Meier Estimator',
col=2:3)
legend(20,0.5,c("I-II","III-Undiffer."),lty=1:1,col=2:3)

br4<-survdiff(Surv(Survival,Death)~Grade,data=a);br4

#test Flemming-Harrington (ρ=1,π=0) - προσέγγιση Peto Peto#
fh4<-survdiff(Surv(Survival,Death)~Grade,data=a,rho=1);fh4
```

Παράγοντας Size

```
brcan_Size<-survfit(Surv(Survival,Death)~Size,data=a,conf.type="none")

summary(brcan_Size)

plot(brcan_Size,xlab='Time in weeks',ylab='Survival',main='Kaplan-Meier Estimator',
```

```
col=2:4)
legend(20,0.5,c("<=2cm","2-5cm",">5cm"),lty=1:1,col=2:4)

br5<-survdiff(Surv(Survival,Death)~Size,data=a);br5

#test Flemming-Harrington (ρ=1,π=0) - προσέγγιση Peto Peto#
fh5<-survdiff(Surv(Survival,Death)~Size,data=a,rho=1);fh5
```

Παράγοντας Nodes

```
brcan_Nodes<-survfit(Surv(Survival,Death)~Nodes,data=a,conf.type="none")

summary(brcan_Nodes)

plot(brcan_Nodes,xlab='Time in weeks',ylab='Survival',main='Kaplan-Meier Estimator',
col=2:3)
legend(20,0.5,c("0-3",">4"),lty=1:1,col=2:3)

br6<-survdiff(Surv(Survival,Death)~Nodes,data=a);br6

#test Flemming-Harrington (ρ=1,π=0) - προσέγγιση Peto Peto#
fh6<-survdiff(Surv(Survival,Death)~Nodes,data=a,rho=1);fh6
```

Παράγοντας ERIhc

```
brcan_ERIhc<-survfit(Surv(Survival,Death)~ERIhc,data=a,conf.type="none")

summary(brcan_ERIhc)

plot(brcan_ERIhc,xlab='Time in weeks',ylab='Survival',main='Kaplan-Meier Estimator',
col=2:3)
legend(20,0.5,c("Negative","Positive"),lty=1:1,col=2:3)

br7<-survdiff(Surv(Survival,Death)~ERIhc,data=a);br7

#test Flemming-Harrington (ρ=1,π=0) - προσέγγιση Peto Peto#
fh7<-survdiff(Surv(Survival,Death)~ERIhc,data=a,rho=1);fh7
```

Παράγοντας PgRIhc

```
brcan_PgRIhc<-survfit(Surv(Survival,Death)~PgRIhc,data=a,conf.type="none")

summary(brcan_PgRIhc)

plot(brcan_PgRIhc,xlab='Time in weeks',ylab='Survival',main='Kaplan-Meier Estimator',
col=2:3)
legend(20,0.5,c("Negative","Positive"),lty=1:1,col=2:3)

br8<-survdiff(Surv(Survival,Death)~PgRIhc,data=a);br8
```

```
#test Flemming-Harrington (ρ=1,π=0) - προσέγγιση Peto Peto#
fh8<-survdiff(Surv(Survival,Death)~PgRIhc,data=a,rho=1);fh8
```

Παράγοντας HT

```
brcan_HT<-survfit(Surv(Survival,Death)~HT,data=a,conf.type="none")

summary(brcan_HT)

plot(brcan_HT,xlab='Time in weeks',ylab='Survival',main='Kaplan-Meier Estimator',
col=2:3)
legend(20,0.5,c("No","Yes"),lty=1:1,col=2:3)

br9<-survdiff(Surv(Survival,Death)~HT,data=a);br9

#test Flemming-Harrington (ρ=1,π=0) - προσέγγιση Peto Peto#
fh9<-survdiff(Surv(Survival,Death)~HT,data=a,rho=1);fh9
```

Παράγοντας RT

```
brcan_RT<-survfit(Surv(Survival,Death)~RT,data=a,conf.type="none")

summary(brcan_RT)

plot(brcan_RT,xlab='Time in weeks',ylab='Survival',main='Kaplan-Meier Estimator',
col=2:3)
legend(20,0.5,c("No","Yes"),lty=1:1,col=2:3)

br10<-survdiff(Surv(Survival,Death)~RT,data=a);br10

#test Flemming-Harrington (ρ=1,π=0) - προσέγγιση Peto Peto#
fh10<-survdiff(Surv(Survival,Death)~RT,data=a,rho=1);fh10
```

Παράγοντας Surgery

```
brcan_Surgery<-survfit(Surv(Survival,Death)~Interval,data=a,conf.type="none")

summary(brcan_Surgery)

plot(brcan_Surgery,xlab='Time in weeks',ylab='Survival',main='Kaplan-Meier Estimator',
col=2:4)
legend(20,0.5,c("MRM","BCS"),lty=1:1,col=2:4)

br11<-survdiff(Surv(Interval,Death)~Surgery,data=a);br11

#test Flemming-Harrington (ρ=1,π=0) - προσέγγιση Peto Peto#
```

```
fh11<-survdiff(Surv(Survival,Death)~Surgery,data=a,rho=1);fh11
```

Παράγοντας Interval

```
brcan_Interval<-survfit(Surv(Survival,Death)~Interval,data=a,conf.type="none")

summary(brcan_Interval)

plot(brcan_Interval,xlab='Time in weeks',ylab='Survival',main='Kaplan-Meier Estimator',
col=2:4)
legend(20,0.5,c("<2wks","2-4wks",">4wks"),lty=1:1,col=2:4)

br12<-survdiff(Surv(Interval,Death)~Surgery,data=a);br12

#test Flemming-Harrington (ρ=1,π=0) - προσέγγιση Peto Peto#
fh12<-survdiff(Surv(Survival,Death)~Interval,data=a,rho=1);fh12
```

B.6 Nelson-Aalen εκτιμητής της συνάρτησης επιβίωσης και διαστήματα εμπιστοσύνης

```
na1<-survfit(Surv(Survival,Death)~1,data=a,conf.int=0.95,conf.type="plain",
type="fleming-harrington",error="greenwood")
summary(na1)
plot(na1,xlab='Time',ylab='Survival',
main='Nelson-Aalen Estimator (CI:plain-Type:Greenwood)')

na2<-survfit(Surv(Survival,Death)~1,data=a,conf.int=0.95,conf.type="log-log",
type="fleming-harrington",error="greenwood")
summary(na2)
plot(na2,xlab='Time',ylab='Survival',
main='Nelson-Aalen Estimator (CI:log-log-Type:Greenwood)')

na3<-survfit(Surv(Survival,Death)~1,data=a,conf.int=0.95,conf.type="log",
type="fleming-harrington",error="greenwood")
summary(na3)
plot(na3,xlab='Time',ylab='Survival',
main='Nelson-Aalen Estimator (CI:log-Type:Greenwood)')

na4<-survfit(Surv(Survival,Death)~1,data=a,conf.int=0.95,conf.type="plain",
type="fleming-harrington",error="tsiatis")
summary(na4)
plot(na4,xlab='Time',ylab='Survival',
main='Nelson-Aalen Estimator (CI:plain-Type:Tsiatis)')

na5<-survfit(Surv(Survival,Death)~1,data=a,conf.int=0.95,conf.type="log-log",
```

```
type="fleming-harrington",error="tsiatis")
summary(na5)
plot(na5,xlab='Time',ylab='Survival',
main='Nelson-Aalen Estimator (CI:log-log-Type:Tsiatis)')
ylab='Survival',main='Nelson-Aalen Estimator (log-log,Tsiatis)')

na6<-survfit(Surv(Survival,Death)~1,data=a,conf.int=0.95,conf.type="log",
type="fleming-harrington",error="tsiatis")
summary(na6)
plot(na6,xlab='Time',ylab='Survival',
main='Nelson-Aalen Estimator (CI:log-Type:Tsiatis)')
```

B.7 Σύγκριση συναρτήσεων επιβίωσης Kaplan-Meier και Nelson-Aalen

```
nelson <- survfit(Surv(Survival,Death)~1,data=a, conf.type = "none",
type="fleming-harrington")

km <- survfit(Surv(Survival,Death)~1,data=a, conf.type = "none")
nelson$time
km$surv

plot(nelson, xlab = "Time", ylab = " S(t) ", col = 2)
points(nelson$time, km$surv, type = "s", col = 4)
legend(20, 0.5, c("Nelson-Aalen", "Kaplan-Meier"), lty = c(1, 1),
col = c(2, 4))
```

B.8 Ανάλυση υπολοίπων

```
fit<-coxph(Surv(Survival,Death)~Group+Nodes+
Factor1+Factor2+Factor5+Factor7,data=a)
```

Υπόλοιπα Deviance

```
dev<-resid(fit,type='deviance')
plot(dev,ylab='Deviance Residuals')
```

Schoenfeld Υπόλοιπα

```
scho<-resid(fit,type='schoenfeld')
scho
resca<-resid(fit,type='scaledsch')
resca
phfit<-cox.zph(fit,transform='log')
phfit
```

```
plot(phfit[1])
plot(phfit[2])
plot(phfit[3])
plot(phfit[4])
plot(phfit[5])
plot(phfit[6])
```

Score και scaled score (Δέλτα-Βήτα) Υπόλοιπα

```
score<-resid(fit,type='score')
score
plot(score[,1],ylab='Score Residuals for Group')
plot(score[,2],ylab='Score Residuals for Nodes')
plot(score[,3],ylab='Score Residuals for Factor1')
plot(score[,4],ylab='Score Residuals for Factor2')
plot(score[,5],ylab='Score Residuals for Factor5')
plot(score[,6],ylab='Score Residuals for Factor7')

db<-resid(fit,type='dfbeta')
db
plot(db[,1],type='h',ylab='Delta-Beta Residuals for Group')
abline(h=0)
plot(db[,2],type='h',ylab='Delta-Beta Residuals for Nodes')
abline(h=0)
plot(db[,3],type='h',ylab='Delta-Beta Residuals for Factor1')
abline(h=0)
plot(db[,4],type='h',ylab='Delta-Beta Residuals for Factor2')
abline(h=0)
plot(db[,5],type='h',ylab='Delta-Beta Residuals for Factor5')
abline(h=0)
plot(db[,6],type='h',ylab='Delta-Beta Residuals for Factor7')
abline(h=0)
```

Β.9 Θηκογράμματα

Στη συνέχεια παρατίθεται η ρουτίνα από την οποία κατασκευάστηκε το θηκόγραμμα του παράγοντα ‘ER’ ως προς την γονιδιακή έκφραση ‘UBE2c’. Τα υπόλοιπα διαγράμματα προκύπτουν ανάλογα, αντικαθιστώντας τα αντίστοιχα πεδία.

```
library(graphics)
library(gtools)
library(gdata)
library(caTools)
library(bitops)
```

```
library(grid)
library(gplots)
par(mfcol=c(2,5))

boxplot(UBE2c~ERIhc, data = a, na.action = NULL,border = par("fg"),
col =c('green','blue'),ylab='UBE2c',xlab='ER(IHC)')
```

Β Ι Β Λ Ι Ο Γ Ρ Α Φ Ι Α

Ελληνική

Αντζουλάκος Δ. (2009). '*Ανάλυση Επιβίωσης*', Πανεπιστημιακές Σημειώσεις για το ΠΜΣ στην "Εφαρμοσμένη Στατιστική", Πανεπιστήμιο Πειραιά.

Κούτρας Μ. (2008). '*Ανάλυση Παλινδρόμησης και Ανάλυση Διακύμανσης*', Πανεπιστημιακές Σημειώσεις για το ΠΜΣ στην "Εφαρμοσμένη Στατιστική", Πανεπιστήμιο Πειραιά

Φωκιανός Κ. και Χαραλάμπους Χ. (2010). '*Εισαγωγή στην R - Πρόχειρες Σημειώσεις*', Πανεπιστήμιο Κύπρου.

Ξένη

Anderberg, M. (1973). *Cluster analysis for applications*, Academic press, New York.

Barger, C. (2009). 'The Mean, Median, and Confidence Intervals of the Kaplan-Meier Survival Estimate-Computations and Applications', *The American Statistician*, **63**:1, pp. 78-80.

Bolshakova, N., Azuaje, F. and Cunningham, P. (2005). 'A knowledge-driven approach to cluster validity assessment', *Bioinformatics*, **21**, pp. 2546–2547.

Breslow, N. (1974). 'Covariance Analysis of Censored Survival Data'. *Biometrics*, **30**, pp. 89-100.

Causton, H., Quackenbush, J, and Brazma, A. (2003). *A Beginner's Guide: Microarray Gene Expression Data Analysis*, Blackwell Publishing, United Kingdom.

Collet, D. (1994). '*Modelling Survival Data in Medical Research*', Chapman & Hall, London and New York.

Copland, J., Davies, P., Shipley, G., Wood, C., Luxon B. and Urban, R. (2003). 'The Use of DNA Microarrays to Assess Clinical Samples: The Transition from Bedside to Bench to Bedside ', *Recent Progress in Hormone Research*, **58**, pp. 25-53.

Cox, D. (1972). 'Regression Models and Life Tables (with discussion)'. *Journal of the Royal Statistical Society*, B, **74**, pp.187-220.

D'haeseleer, P. (2005). 'How does gene expression clustering work?', *Nature Biotechnology*, **23**:12, pp. 1499-1501.

Datta, S. and Datta, S. (2003). 'Comparisons and validation of statistical clustering techniques for microarray gene expression data', *Bioinformatics*, **19**, pp.459–466.

Datta, S. and Datta, S. (2006). 'Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes', *BMC Bioinformatics*, **7**, pp. 397.

Dietzsch, J., Gehlenborg, N. and Nieselt, K. (2006). 'Mayday-a microarray data analysis workbench', *Bioinformatics*, **22**, pp. 1010–1012.

Dudoit, S. and Gentleman, R. (2002). '*Cluster Analysis in DNA Microarray Experiments*', Bioconductor Short Course.

Dunn, J. (1974). 'Well separated clusters and fuzzy partitions', *Journal on Cybernetics*, **4**, pp. 95–104.

Efron, B. (1977). 'The Efficiency of Cox's Likelihood Function of Censored Data'. *Journal of the American Statistical Association*, **72**, pp. 557-565.

Efron, B., and Tibshirani, R. J. (1994). '*An Introduction to the Bootstrap*'. Boca Raton, FL: CRC Press.

Eisen, M., Spellman, P., Brown, P. and Botstein, D. (1998). "Cluster analysis and display of genome-wide expression patterns", *Genetics*, **95**, pp. 14863–14868.

Figueroa, A., Borneman, J. and Jiang, T. (2003). 'Clustering binary fingerprint vectors with missing values for DNA array data analysis', *Proc IEEE Comput Soc Bioinform Conf.* ; **2**, 38-47.

Gat-Viks, I., Sharan, R. and Shamir, R. (2003). 'Scoring clustering solutions by their biological relevance', *Bioinformatics*, **19**, pp. 2381–2389.

Gibbons, F. and Roth, F., (2002). 'Judging the quality of gene expression-based clustering methods using gene annotation', *Genome Res*, **12**, pp. 1574–1581.

Glitten, D. and Vittinghoff, E. (2004). 'Modelling clustering survival data from multicentre clinical trials', *Statistics in Medicine*, **23**, pp. 369-388.

Grambsch, P. and Therneau, T. (1994). 'Proportional Hazards Tests and Diagnostics based on Weighted Residuals'. *Biometrika*, **81**, pp. 515-526.

Han, J., Kamber, M. and Tung, A. K. H. (2005). 'Spatial Clustering Methods in Data Mining: A Survey', In *Geographic Data Mining and Knowledge Discovery* [Miller, H.J, and Han, J., Eds]. London: Taylor & Francis Inc., pp. 5-25.

Handl, J., Knowles, J. and Kell, D. (2005). 'Computational cluster validation in postgenomic data analysis', *Bioinformatics*, **21**, pp. 3201–3212.

Harman, H. (1913). '*Modern Factor Analysis'*, The University of Chicago Press, Chicago and London.

Hougaard, P. (2001). '*Analysis of Multivariate Survival Data*', Springer, New York.

Jiang, D., Pei, J. and Zhang, A. (2003). "DHC: A Density-Based Hierarchical Clustering Method for Time-Series Gene Expression Data", *IEEE Symp. on Bioinf. and Bioeng.*

Jiang, D., Tang, C. and Zhang, A. (2004). "Cluster Analysis for Gene Expression Data: A Survey", *IEEE Transactions on Knowledge and Data*, **16**, pp 1370-1386.

Johnson and Wichern (1998), *Applied Multivariate Statistical Analysis*, Prentice Hall, Upper Saddle River, NJ.

Johnson, S. (1967). "Hierarchical Clustering Schemes", *Psychometrika*, **32**, pp. 241-254.

Kaiser, H. (1974). 'An index of factorial simplicity', *Psychometrika*, **39**, pp. 31-36.

Kalbfleisch, J. and Prentice, R. (1980). '*The Statistical Analysis of Failure Time Data*', Wiley, New York.

Kaplan, E. and Meier, P. (1958). 'Non-parametric estimation from incomplete observations', *Journal of the American Statistical Association*, **53**, pp. 457-481, 562-563.

Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data*, Wiley, United States of America.

Kerr, M. and Churchill, G. (2001). 'Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments', *Proc Natl Acad Sci USA*, **98**, pp. 8961–8965.

Klein, J. and Moeschberger, M. (1997). '*Survival Analysis: Techniques for Censored and Truncated Data*', Springer, New York.

Lee, M.L.T., (2004). *Analysis of Microarray Gene Expression Data*, Kluwer Academic Publishers, United States of America.

Lewis-Beck, M. (1994). *Factor Analysis and Related Techniques*, **5**, SAGE Publications-Toppan Publishing.

Ling, R. (1973). "A computer generated aid for cluster analysis". *Communications of the ACM*, **16**, pp. 355–361.

MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations", *Proceeding of 5th Berkeley Symposium on Mathematical Statistics and Probability*, **1**, pp. 281-297.

McLachlan, G., Do, K. and Ambroise, C. (2004). *Analyzing Microarray Gene Expression Data*, Wiley, New Jersey.

Park, H., Lee, J. and Jun, C. (2006). 'A K-means-like Algorithm for K-medoids Clustering and Its Performance', *Proceedings of the 36th CIE Conference on Computers & Industrial Engineering*, pp.1222-1231.

Peterson, L. (2002). 'Factor analysis of cluster-specific gene expression levels from cDNA microarrays', *Computer Methods and Programs in Biomedicine*, **69**, pp. 179–188.

Peto, R. (1972). 'Contribution to the Discussion of a Paper by D.R. Cox', *Journal of the Royal Statistical Society*, B, **34**, pp. 205-207.

Pournara, I. and Wernisch, L. (2007). 'Factor analysis for gene regulatory networks and transcription factor activity profiles', *BMC Bioinformatics*, **8**: 61, doi:10.1186/1471-2105-8-61

Rousseeuw, P. (1987). 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', *Journal of Computational and Applied Mathematics*, **20**, pp.53–65.

Saha, I. and Mukhopadhyay, A. (2001). 'Improved Crisp and Fuzzy Clustering Techniques for Categorical Data', *IAENG International Journal of Computer Science*, **35**:4 (13 pages).

Sauerbrei, W. and Schumacher, M. (2000). 'Bootstrap and Cross-Validation to Assess Complexity of Data-Driven Regression Models', Springer, **1933**, pp. 234–241.

Schmidt, L., Harms, H., Kuhn, S. and Rommelspacher, H. and Sander, T. (1998). 'Modification of Alcohol Withdrawal by the A9 Allele of the Dopamine Transporter Gene', *Am J Psychiatry*, **155**, pp. 474-478.

Schoenfeld, D. (1980). Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika*, **67**, pp.145-153.

Sherlock, G. (2000). “Analysis of Large-Scale Gene Expression Data”, Current Opinion in Immunology, **12**, pp. 201-205.

Smet, F., Mathys, J., Marchal, K., Thijs, G., Moor, M., Bart, D., and Moreau Y. (2002). Adaptive Quality-Based Clustering of Gene Expression Profiles,” *Bioinformatics*, **18**, pp. 735-746.

Sneath, P. (1957). "The application of computers to taxonomy". *Journal of General Microbiology* , **17**, pp. 201–226.

Speed, T. (2003). *Statistical Analysis of Gene Expression Microarray Data*, Chapman & Hall/CRC, Florida.

Tang, C. and Zhang, A. (2002). ‘An iterative strategy for pattern discovery in high-dimensional data sets’, In Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM 02), *ACM Press,* pp. 10-17.

Tang, C., Zhang, L., Zhang, A. and Ramanathan, M. (2001). “Interrelated Two-Way Clustering: An Unsupervised Approach for Gene Expression Data Analysis”, *Proc. BIBE2001: Second IEEE Int’l Symp. Bioinformatics and Bioeng.*, pp. 41-48.

Vittinghoff, E., Glidden, D., Shiboski, S., and McCulloch, C.E. (2005). ‘*Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*’, Springer, USA.

Wang, S., Gui, J. and Li, X. (2010). ‘Factor Analysis for Cross-Platform Tumor Classification Based on Gene Expression Profiles’, *World Scientific*, **19**, pp. 243-258.

Wang, S., Wang, J., Chen, H., and Tang, W. (2006). ‘The Classification of Tumor Using Gene Expression Profile Based on Support Vector Machines and Factor Analysis’, *Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications (ISDA'06)*

Wilkinson, L. (1994). *Advanced Applications: Systat for DOS* Version 6. SYSTAT Inc.

Wilkinson, L. and Friendly, M. (2009). ‘The History of the Cluster Heat Map’, *The American Statistician*, **63**, pp. 179-184.

Xiao, Y. and Abrahamowicz, M. (2009). ’ Bootstrap-based methods for estimating standard errors in Cox’s regression analyses of clustered event times’, *Wiley*, **95**, pp. 915-923.

Yang, J., Wang, W., Wang, H. and Yu, P. (2002), “δ-Cluster: Capturing Subspace Correlation in a Large Data Set”, *Proc. 18th Int’l Conf. Data Eng. (ICDE 2002)*, pp. 517-528.

Yeung, K., Haynor, D. and Ruzzo, W. (2001). ‘Validating clustering for gene expression data’, *Bioinformatics*, **17**, pp. 309–318.

Zhong, M. and Hess, K. (2009). ‘Mean survival time for right censored data’, *The Berkeley Electronic Press*, **66**.

Zitzmann, M., Gromoll, J., von Eckardstein, A. and Nieschlag, E. (2003). ‘The CAG repeat polymorphism in the androgen receptor gene modulates body fat mass and serum concentrations of leptin and insulin in men’, *Diabetologia*, **46**, pp. 31-39.

Ιστοσελίδες

1. http://en.wikipedia.org/wiki/Breast_cancer#Risk_factors
2. http://en.wikipedia.org/wiki/Factor_analysis.
3. http://en.wikipedia.org/wiki/Heat_map
4. http://en.wikipedia.org/wiki/Real-time_polymerase_chain_reaction
5. http://en.wikipedia.org/wiki/Silhouette_(clustering).
6. http://pathmicro.med.sc.edu/pcr/realtime-home.htm\
7. http://peoplelearn.homestead.com/Topic20-FACTORanalysis3a.html
8. http://www.aliquote.org/articles/tech/multvar/22_Appendix_6.pdf
9. http://www.ambion.com/techlib/basics/rtpcr/index.html
10. http://www.ats.ucla.edu/stat/SPSS/output/principal_components.htm
11. http://www.breastcancer.org/symptoms/understand_bc/what_is_bc.jsp
12. http://www.cancer.gov/cancertopics/factsheet/Detection/tumor-grade
13. http://www.labescape.com/info/articles/what-is-a-heat-map.html
14. http://www.statsoft.com/textbook/cluster-analysis/?button=1
15. http://www.surgeon.gr/110/3342.aspx
16. http://www.unesco.org/webworld/idams/advguide/Chapt7_1_1.htm.
17. http://www3.bio-rad.com/B2B/BioRad/product/br_category.jsp?BV_SessionID=@@@@0336872906.1296491814@@@@&BV_EngineID=ccciademjhjedlhcfngcfkmdhkkdfll.0&divName=Food+%7c+Animal+%7c+Environment+Testing&loggedIn=false&serviceLevel=Lit+Request&lang=Chinese&csel=CN&catLevel=4&catOID=-2725&isPA=false&categoryPath=Catalogs%2fFood+%7c+Animal+%7c+Environment+Testing%2fFood+Testing%2fReal-Time+PCR

Ρουτίνες της R

ClValid: an R package for cluster validation

Guy Brock, Vasyl Pihur, Susmita Datta, and Somnath Datta

Department of Bioinformatics and Biostatistics, University of Louisville