ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ

ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΜΟΝΤΕΛΟ ΑΝΑΛΥΣΗΣ ΕΠΙΒΙΩΣΗΣ ΓΙΑ ΠΟΛΛΑΠΛΟΥΣ ΤΡΟΠΟΥΣ ΑΠΟΤΥΧΙΑΣ

Γιαννακόπουλος Θεόδωρος

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

Οκτώβριος 2021

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από το ΓΣΕΣ του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμόν ……. συνεδρίαση του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική.

Τα μέλη της Επιτροπής ήταν:

- Τζαβελάς Γεώργιος (Επιβλέπων)
- Γ. Βερροπούλου
- Κ. Πολίτης

Η έγκριση της Διπλωματικής Εργασίας από το τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN

APPLIED STATISTICS

A SURVIVAL MODEL FOR MULTIPLE MODE FAILURES

By

Giannakopoulos Theodoros

MSc Dissertation

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

Piraeus
October 2021

Περίληψη

Η ανάλυση επιβίωσης (survival analysis) είναι μια ομάδα μεθόδων για την ανάλυση δεδομένων όταν η υπό μελέτη μεταβλητή είναι ο χρόνος μέχρι το συμβάν που μας ενδιαφέρει (time-to-event analysis). Το συμβάν που καταγράφεται λέγεται αποτυχία. Η παρούσα διπλωματική εργασία παρουσιάζει τις βασικές μεθόδους μελέτης των περιπτώσεων εκείνων στις οποίες καταγράφονται πολλαπλοί τρόποι αποτυχίας. Στο πρώτο κεφάλαιο δίδονται οι βασικοί ορισμοί και γίνεται μια ανασκόπηση των βασικών μεθόδων στην ανάλυση επιβίωσης. Στο δεύτερο κεφάλαιο γίνεται μια περιγραφή του μοντέλου με πολλούς τρόπους αποτυχίας, επίσης τονίζονται οι ιδιαιτερότητές του και δίνεται η συνάρτηση πιθανοφάνειας. Στο τρίτο κεφάλαιο περιγράφονται οι μη παραμετρικοί μέθοδοι αντιμετώπισης του προβλήματος. Στο τέταρτο οι παραμετρικοί και στο πέμπτο κεφάλαιο οι ημιπαραμετρικοί μέθοδοι. Τα θεωρητικά ευρήματα υποστηρίζονται από αριθμητικά παραδείγματα. Στο έκτο και τελευταίο κεφάλαιο έχουμε μία εφαρμογή με την ανάλυσή της.

Abstract

Survival analysis provides a set of statistical methods for data analysis when the variable under consideration is the time up to an event of interest usually called failure (time to event analysis). This MSc thesis presents the methods for the study of the cases where multiple modes of failures are recorded (competed risks models). At the first chapter the basic definitions are given and the most important methods of survival analysis are presented. The second chapter describes the survival model with multiple modes of failure and the likelihood function is given. The third, fourth, and fifth chapter include the nonparametric, parametric and semiparametric methods respectively for studying such data. The theoretical methods are supported by numerical examples. In the sixth and last chapter we have an application with empirical data.

Ευχαριστίες

Επιθυμώ να εκφράσω τις ευχαριστίες μου σε όλους εκείνους που συνέβαλαν άμεσα ή έμμεσα στην ολοκλήρωση της διπλωματικής μου εργασίας και κατά συνέπεια των μεταπτυχιακών σπουδών μου.

Θα ήθελα να ευχαριστήσω τον επιβλέποντά μου, Αναπληρωτή Καθηγητή Τζαβελά Γεώργιο. Η υποστήριξη και διαθεσιμότητά του καθ΄ όλη τη διάρκεια της εκπόνησης της εργασίας αποτέλεσε σπουδαία βοήθεια, συμβάλλοντας ποικιλοτρόπως στην ολοκλήρωσή της, παρέχοντας μεταξύ άλλων πολύτιμες συμβουλές και καθοδήγηση όπου κρίθηκε αναγκαίο.

Επίσης, θα ήθελα εκφράσω την ευγνωμοσύνη μου σε έναν πολύ δικό μου άνθρωπο την Πόπη για την στήριξη, τη συμπαράσταση και την κατανόησή της, καθ’ όλη τη διάρκεια των σπουδών μου.

Περιεχόμενα

Κατάλογος Πινάκων…….

Κατάλογος Σχημάτων……...

Κεφάλαιο 1

1.1 Εισαγωγικά ……..1

1.2 Βασικές Γνώσεις ……..……3

1.2.1 Περικοπή και Λογοκρισία.……3

1.2.2 Τυχαία λογοκρισία (Random Censoring) ……...……..5

1.2.3 Ο εκτιμητής product-limit (Kaplan-Meyer)……..7

Κεφάλαιο 2

2.1 Εισαγωγή…….....10

2.2.1 Βασικά Χαρακτηριστικά και επιλογή Μοντέλων……...10

2.2.2 Συνάρτησης Πιθανοφάνειας στην παραμετρική περίπτωση…….……..13

Κεφάλαιο 3

Μη Παραμετρικές Μέθοδοι……..……...16

Κεφάλαιο 4

4.1 Παραμετρικές Μέθοδοι…….……20

4.2 Ομαδοποιημένα ή Διακριτά Δεδομένα……..…….25

Κεφάλαιο 5

5.1 Εκτίμηση των Συναρτήσεων Αθροιστικής Επίπτωσης ……..…...27

Κεφάλαιο 6

Ανάλυση Δεδομένων με τη χρήση του SPSS Statistics……...32

Παράρτημα…….56

Πίνακες δεδομένων……..57

Βιβλιογραφία……..61

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Κεφάλαιο 4

Πίνακας 4.1 Μη παραμετρικοί έλεγχοι υποθέσεων……...25

Κεφάλαιο 6

Πίνακας 6.1 Συχνότητες μεταβλητών……..37
Πίνακας 6.2 Crosstab…….....38
Πίνακας 6.3 Έλεγχος x^2…….....38
Πίνακας 6.4 Survival Table-event=1…….....40
Πίνακας 6.5 Survival Table-event=3…….....42
Πίνακας 6.6 Test of equality of survival distributions for the different levels of sex……....43
Πίνακας 6.7 Test of equality of survival distributions for the different levels of ulcer…….....44
Πίνακας 6.8 Test of equality of survival distributions for the different levels of thickness_groups……....46
Πίνακας 6.9 Variables in the Equation (status=1,sex)…….....47
Πίνακας 6.10 Variables in the Equation (status=3,sex)……47
Πίνακας 6.11 Variables in the Equation (status=1,age)……48
Πίνακας 6.12 Variables in the Equation (status=3,age)……49
Πίνακας 6.13 Variables in the Equation (status=1,ulcer)……....50
Πίνακας 6.14 Variables in the Equation (status=3,ulcer)……....50
Πίνακας 6.15 Variables in the Equation (status=1,thickness_group)……...……51
Πίνακας 6.16 Variables in the Equation (status=3,thickness_group)……..52
Πίνακας 6.17 Variables in the Equation (status=1,year)……53
Πίνακας 6.18 Variables in the Equation (status=3,year)……54
Πίνακας 6.19 Variables in the Equation (status=1,ulcer/thickness)……....54
Πίνακας 6.20 Variables in the Equation (status=3,ulcer/thickness)……....55
Πίνακας 6.21 Variables in the Equation (status=1,all variables)……....56
Πίνακας 6.22 Variables in the Equation (status=3,all variables)……....57
Πίνακας 6.23 Συγκεντρωτικός πίνακας επιδράσεων από τα μοντέλα που αποτελούνται κάθε φορά από μόνο μία μεταβλητή……..58
Πίνακας 6.24 Συγκεντρωτικός πίνακας επιδράσεων από το μοντέλο αποτελείται από όλες τις μεταβλητές του δείγματος……....58

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

Κεφάλαιο 1

Σχήμα 1.1 Γραφική τύπου 1 με σημείο λογοκρισίας στην τιμή L …….....4

Κεφάλαιο 4

Σχήμα 4.1 Σύγκριση συναρτήσεων επιβίωσης για αιτία θανάτου: Λέμφωμα.……...21
Σχήμα 4.2 Σύγκριση συναρτήσεων επιβίωσης για αιτία θανάτου: Σάρκωμα.……22
Σχήμα 4.3 Σύγκριση συναρτήσεων επιβίωσης για αιτία θανάτου: Άλλη αιτία..…….…22
Σχήμα 4.4 Διάγραμμα διασποράς των $(\log \hat{\Lambda}_i \ , \log t) \ \ i = 2,3$ για Λέμφωμα …….....23
Σχήμα 4.5 Διάγραμμα διασποράς των $(\log \hat{\Lambda}_i \ , \log t) \ \ i = 2,3$ για Σάρκωμα…...……24
Σχήμα 4.6 Διάγραμμα διασποράς των $(\log \hat{\Lambda}_i \ , \log t) \ \ i = 2,3$ για Άλλη αιτία……24

Κεφάλαιο 6

Σχήμα 6.1 Δήλωση μεταβλητών του dataset στο SPSS……..33
Σχήμα 6.2 Frequencies διαδικασία στο SPSS…….……33
Σχήμα 6.3 Crosstabs διαδικασία στο SPSS…….35
Σχήμα 6.4 Bar chart φύλου vs παρουσία έλκους……36
Σχήμα 6.5 Ιστόγραμμα Χρόνου Επιβίωσης Ασθενών με Μελάνωμα για κάθε ένα αποτέλεσμα (status)…..36
Σχήμα 6.6 Kaplan-Meier διαδικασία στο SPSS …….……37
Σχήμα 6.7 Καμπύλη επιβίωσης, status=1……....38
Σχήμα 6.8 Kaplan-Meier διαδικασία στο SPSS……..38
Σχήμα 6.9 Καμπύλη επιβίωσης, status=3…….……...39
Σχήμα 6.10 Σύγκριση συναρτήσεων επιβίωσης από μελάνωμα ανά φύλο……...…….40
Σχήμα 6.11 Kaplan-Meier διαδικασία στο SPSS…….……...41
Σχήμα 6.12 Σύγκριση συναρτήσεων επιβίωσης από μελάνωμα ανά έλκος…….……...41
Σχήμα 6.13 Kaplan-Meier διαδικασία στο SPSS…….……...42
Σχήμα 6.14 Σύγκριση συναρτήσεων επιβίωσης από μελάνωμα ανά μέγεθος όγκου…….……42
Σχήμα 6.15 Cox Regression διαδικασία στο SPSS ……43
Σχήμα 6.16 Cox Regression διαδικασία στο SPSS με x=age, status=1……..…45
Σχήμα 6.17 Cox Regression διαδικασία στο SPSS με x=age, status=3……..45
Σχήμα 6.18 Cox Regression διαδικασία στο SPSS με x=ulcer, status=1……...46
Σχήμα 6.19 Cox Regression διαδικασία στο SPSS με x=ulcer, status=3……....47
Σχήμα 6.20 Cox Regression διαδικασία στο SPSS με x=thickness_group, status=1……48
Σχήμα 6.21 Cox Regression διαδικασία στο SPSS με x=thickness_group, status=3…….…49
Σχήμα 6.22 Cox Regression διαδικασία στο SPSS με x=year, status=1……..…..50
Σχήμα 6.23 Cox Regression διαδικασία στο SPSS με x=year, status=3……50
Σχήμα 6.24 Cox Regression διαδικασία στο SPSS με x_1=thickness, x_2=ulcer, status=1……..….51
Σχήμα 6.25 Cox Regression διαδικασία στο SPSS με x_1=thickness, x_2=ulcer, status=3……52
Σχήμα 6.26 Cox Regression διαδικασία στο SPSS με x=όλες οι μεταβλητές, status=1……53
Σχήμα 6.27 Cox Regression διαδικασία στο SPSS με x=όλες οι μεταβλητές, status=1……..…..54

ΚΕΦΑΛΑΙΟ 1

1.1 Εισαγωγικά

Ένα άτομο σε κάθε χρονική στιγμή μπορεί να εκτίθεται σε ρίσκο διαφόρων γεγονότων. Σε τέτοιες περιπτώσεις η ανάλυση επιβίωσης πρέπει να λαμβάνει υπόψη της διαφορετικούς τρόπους αποτυχίας, κάθε ένας από τους οποίους έχει τον δικό του μηχανισμό αποτυχίας. Μελετώντας τη θνησιμότητα ενός ανθρώπινου πληθυσμού, για παράδειγμα, η εμφάνιση του θανάτου μπορεί να έρθει από διάφορες αιτίες, και κάθε αιτία μπορεί να χαρακτηρίζεται από ένα σύστημα παθολογικών και φυσιολογικών κριτηρίων (Kalbfleisch and Prentice, 2002). Η παρουσία ενός χρόνιου αναπνευστικού προβλήματος το οποίο σχετίζεται με το φύλο και την ηλικία αυξάνει τον κίνδυνο θανάτου, ενώ την ίδια στιγμή οι ίδιοι αυτοί παράγοντες επίσης σχετίζονται με τον κίνδυνο θανάτου λόγω ατυχήματος. Αυτές οι αιτίες θανάτου λέγονται ανταγωνιστικά ρίσκα, και τα στατιστικά μοντέλα, μοντέλα ανταγωνιστικού ρίσκου.

Οι πολλαπλοί τρόποι αποτυχίας και τα προβλήματα ανταγωνιστικού ρίσκου έχουν μακρά ιστορία στη δημογραφία και στον τομέα της θνησιμότητας. Οι Gail (1975) και Seal (1977), παρέχουν ιστορικές ανασκοπήσεις

Ας θεωρήσουμε ότι $\boldsymbol{T}_i = (T_{i1}.T_{i2})$ είναι το ζευγάρι χρόνου αποτυχίας. Θα μπορούσε για παράδειγμα να είναι οι χρόνοι εμφάνισης $X_{i1}.X_{i2}$ καρκίνου στο αριστερό ή στο δεξί στήθος αντίστοιχα του ατόμου i. Θα μπορούσε κάποιος από τους χρόνου ή και οι δύο να μη είναι γνωστοί λόγω τυχαίας λογοκρισίας C_i. Έτσι τελικά αυτό που παρατηρούμε είναι

$$\boldsymbol{X_i} = (X_{i1}.X_{i2}) = (T_{i1} \wedge C_i\ , T_{i2} \wedge C_i)$$

Καθώς και τη δείκτρια συνάρτηση

$$\boldsymbol{\delta_i} = (\delta_{i1}, \delta_{i2}) = \big(\ I(T_{i1} \leq C_i), \qquad I(T_{i2} \leq C_i)\big).$$

Για να αντιμετωπίσει αυτό το πρόβλημα ο Munoz (1980a) γενίκευσε τον εκτιμητή Kaplan-Meier στη δισδιάστατη περίπτωση και ο ίδιος (Munoz, 1980b) απέδειξε ότι ο εκτιμητής που προκύπτει είναι συνεπής εκτιμητής της δισδιάστατης κατανομής

$$F(t_1, t_2) = P\{T_{i1} \leq t_1, T_{i2} \leq t_2\}.$$

Η προφανής γενίκευση είναι να έχουμε ένα p −διάνυσμα χρόνων αποτυχίας

$$\boldsymbol{T}_i = (T_{i1}, \dots, T_{ip})$$

Όπου κάθε συντεταγμένη είναι μια αιτία αποτυχίας. Το άτομο παρατηρείται μέχρι την πρώτη αποτυχία. Όλοι οι άλλοι λόγοι αποτυχίας θεωρούνται λογοκριμένοι από τον χρόνο της πρώτης αποτυχίας. Παρατηρούμε λοιπόν τις ποσότητες

$$T_i = \min\{T_{i1}, \dots, T_{ip}\}$$

και

$$\boldsymbol{\delta_i} = \big(\delta_{i1}, \dots, \delta_{ip}\big) = \Big(\ I(T_{i1} \leq T_i), \dots, I\big(T_{ip} \leq T_i\big)\Big).$$

Η διανυσματική δείκτρια συνάρτηση δηλώνει τον λόγο αποτυχίας.

Η πιθανότητα

$$P\{T_{ij} \leq t, \delta_{ij} = 1\}$$

λέγεται ακατέργαστη (crude) πιθανότητα θανάτου από την αιτία j στον χρόνο t.
Μπορεί κατευθείαν να εκτιμηθεί από την στατιστική

$$\frac{1}{n}\sum_{i=1}^{n} I(T_i \le t\,, \delta_{ij} = 1).$$

Η πιθανότητα να επιβιώσει κάποιος από την αιτία j καθαρή πιθανότητα και ορίζεται σαν

$$P\{\, T_{ij} \le t\}.$$

Όταν οι αιτίες αποτυχίας είναι ανεξάρτητες τότε αυτή η πιθανότητα μπορεί να εκτιμηθεί από την Product-Limit μέθοδο δεδομένου ότι όλοι οι άλλοι χρόνοι αποτυχίας εκτός της j μπορούν να θεωρηθούν σαν λογοκρισίας.

Αξίζει να αναφερθεί ότι από θεωρητικής σκοπιάς ένα σημαντικό πρόβλημα είναι ότι στη βάση του δείγματος (T_i, $\boldsymbol{\delta}_i$), $i = 1,2, \dots, n$ είναι αδύνατον να πούμε αν οι χρόνοι αποτυχίας ($T_{i1}, \dots, T_{ip}$) είναι ανεξάρτητοι ή όχι.
Οι κύριες πηγές άντλησης του υλικού είναι από τα βιβλία Lawless (2003), Miller (1981) and Xian Liu (2012) Επειδή όπως έχει αναφερθεί σε πολλές περιπτώσεις η ανάλυση μοντέλων ανταγωνιστικών ρίσκου ανάγονται στη ανάλυση μοντέλων επιβίωσης με ένα ρίσκο και λογοκρισία, στην επόμενη ενότητα δίδονται οι βασικοί ορισμοί και μεθοδολογίες στην ανάλυση επιβίωσης με ένα ρίσκο.

1.2 ΒΑΣΙΚΕΣ ΓΝΩΣΕΙΣ

Λογοκρισία και αποκοπή (Censoring and Truncation)

Στη μελέτη αυτή περιγράφονται προχωρημένες τεχνικές ανάλυσης επιβίωσης για τον λόγο αυτό πολλές γνώσεις της ανάλυσης επιβίωσης θεωρούνται γνωστές και εκτίθενται εν συντομία σε αυτή τη ενότητα αυτή. Η ανάλυση επιβίωσης χρησιμοποιεί μη αρνητικές τυχαίες μεταβλητές ($X > 0$) οι οποίες λέγονται κατανομές ζωής (life distribution) ή κατανομές απώλειας (loss distribution). Συνήθως η μεταβλητή του ενδιαφέροντος είναι ο χρόνος. Κύριο αντικείμενο της μελέτης είναι η συμπεριφορά καθώς και η εκτίμηση της συνάρτησης επιβίωσης η οποία ορίζεται σαν

$$S(x) = P(X > x) = 1 - F(x) = \bar{F}(x).$$

1.2.1 Περικοπή και Λογοκρισία

Στην διάρκεια της δειγματοληψίας πολλές φορές οι παρατηρήσεις είναι μερικώς γνωστές. Είτε λόγω της φύσης του πειράματος είτε λόγω αδυναμίας των οργάνων μέτρησης οι παρατηρήσεις περιορίζονται σε ένα μέρος του εύρους των δυνατών τιμών της τυχαίας μεταβλητής. Ανάλογα με την φύση του περιορισμού έχουμε την περικοπή (truncation) και την λογοκρισία (censoring).

Στη στατιστική censoring (λογοκρισία) είναι μία κατάσταση κατά την οποία οι παρατηρήσεις είναι μερικώς γνωστές.

Μια εταιρεία κατασκευής ανταλλακτικών αεροπλάνου καταγράφει την διάρκεια ζωής ενός ανταλλακτικού. Αν υπάρξει πρόβλημα, το ανταλλακτικό αντικαθίσταται αμέσως και καταγράφεται η διάρκεια ζωής του μέχρι την στιγμή της αντικατάστασης. Σε διαφορετική περίπτωση αφήνεται να λειτουργήσει μέχρι κάποιες ώρες πτήσεις (T) και στη συνέχεια αντικαθίσταται. Αυτό που καταγράφεται λοιπόν είναι το $\min\{X,T\}$ όπου X είναι ο χρόνος ζωής του ανταλλακτικού είτε ο χρόνος αντικατάστασης T.

Σε μια κλινική μελέτη δοκιμής ενός νέου φαρμάκου που υπόσχεται μακροζωία στους καρκινοπαθείς, οι ασθενείς πολλές φορές αποχωρούν από το πείραμα. Και στην περίπτωση αυτή έχουμε την περίπτωση της λογοκρισίας.

Λογοκρισία (censoring) συμβαίνει επίσης όταν μια τιμή βρίσκεται εκτός του εύρους του οργάνου μετρήσεων. Για παράδειγμα, μια ζυγαριά μπάνιου μπορεί να μετρήσει μόνο έως 140 kg. Εάν ένα άτομο 160 kg ζυγίζεται με τη χρήση της ζυγαριάς, ο παρατηρητής θα γνωρίζει μόνο ότι το βάρος του ατόμου είναι τουλάχιστον 140 κιλά.

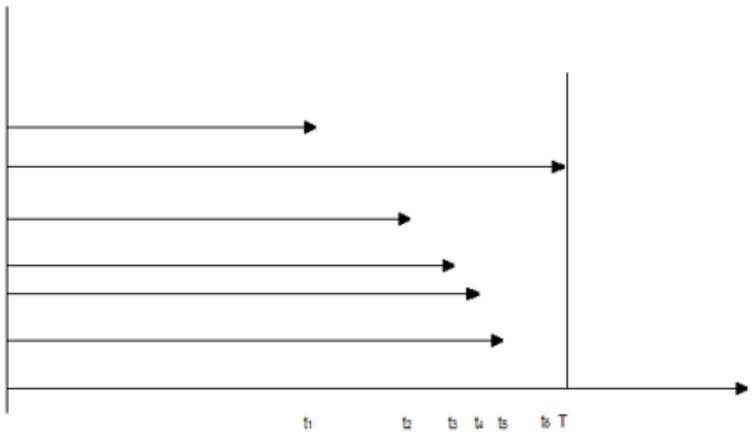
Το πρόβλημα των λογοκριμένων δεδομένων, όπου η παρατηρούμενη τιμή κάποιας μεταβλητής είναι εν μέρει γνωστή, σχετίζεται με το πρόβλημα των ελλιπών δεδομένων, όπου οι παρατηρούμενες τιμές κάποιων μεταβλητών είναι άγνωστες.

Με την περικοπή(truncation), οι παρατηρήσεις ποτέ δεν οδηγούν σε τιμές εκτός ενός συγκεκριμένου εύρους. Παρατηρήσεις που βρίσκονται εκτός αυτού του εύρους είτε δεν παρατηρούνται ή ακόμα και αν παρατηρηθούν, δεν καταγράφονται. Να σημειωθεί πως σε μία στατιστική μελέτη η περικοπή δεν ταυτίζεται με τη στρογγυλοποίηση.

Συνοψίζοντας θα λέγαμε ότι σε ένα λογοκριμένο δείγμα ένα μέρος του είναι γνωστό και στο υπόλοιπο οι παρατηρήσεις είναι μερικώς γνωστές. Στο περικομμένο δείγμα δεν καταγράφονται ούτε καν απαριθμούνται οι παρατηρήσεις εκτός του αποδεκτού εύρους τιμών.

Παρακάτω παρουσιάζονται συνοπτικά οι κυριότερες μορφές λογοκρισίας.

- **Αριστερή λογοκρισία** - ένα σημείο είναι κάτω από μια ορισμένη τιμή, αλλά είναι άγνωστο κατά πόσο.
- **Διάστημα λογοκρισίας** - ένα σημείο βρίσκεται σε ένα διάστημα μεταξύ δύο τιμών χωρίς να γνωρίζουμε την ακριβή τιμή της.
- **Δεξιά λογοκρισία** - ένα σημείο είναι πάνω από μια ορισμένη τιμή, αλλά είναι άγνωστο κατά πόσο.
- **Λογοκρισία Τύπου Ι** συμβαίνει σε ένα πείραμα που έχει ένα συγκεκριμένο αριθμό ατόμων ή αντικειμένων και σταματά σε ένα προκαθορισμένο χρονικό διάστημα. Ένα δείγμα εμφανίζει
- **λογοκρισία Τύπου Ι**, όταν τα όρια λογοκρισίας (T) είναι γνωστά. Το πλήθος των μη λογοκριμένων παρατηρήσεων θα συμβολίζεται με r και των λογοκριμένων με n-r. Το n είναι γνωστό εκ των προτέρων και προκαθορισμένο και το r άγνωστο.



Σχήμα 1.1. Γραφική ***τύπου 1*** με σημείο λογοκρισίας στην τιμή **L**

Αν Τ τμ με π.π. f(t) τότε

- η λογοκριμένη τμ Τ* από πάνω στην τιμή L είναι η

$$T^*=min\{T, L\}= T \wedge L$$

με ππ

$$f_L(t) = \begin{cases} f(t) & t < L \\ 1 - F(L) & t = L \end{cases} \quad (1.1)$$

- η λογοκριμένη τμ Τ* από κάτω στην τιμή L είναι η

$$T^*=max\{T, L\}= T \vee L$$

με ππ

$$f_L(t) = \begin{cases} F(L) & t = L \\ f(t) & t > L \end{cases}$$

- η περικομμένη από πάνω στο L τμ είναι ο περιορισμός της Τ στο διάστημα (0, L). Δηλαδή

$$T^{**}=T\, I_{(0,\, L)}$$

με π.π.

$$f_L(t) = \frac{f(t)}{F(L)} \qquad t < L$$

- η περικομμένη από κάτω στο L τμ είναι ο περιορισμός της Τ στο διάστημα (L, ∞). Δηλαδή

$$T^* = T\, I_{(L,\infty)}$$

με π.π.

$$f_L(t) = \frac{f(t)}{1 - F(L)} \qquad t > L$$

1.2.2 Τυχαία λογοκρισία (Random Censoring)

Στην λογοκρισία τύπου Ι θεωρείται γνωστός και προκαθορισμένος ο χρόνος λογοκρισίας *T*. Στην τυχαία λογοκρισία τα άτομα ξεκινούν από τυχαίες στιγμές, έτσι ώστε τόσο οι ζωές όσο και οι φορές που θα λογοκριθούν οι τιμές να είναι τυχαίες.

Ορίζονται:

T_i = η διάρκεια ζωής του *i* ατόμου

C_i = η στιγμή της λογοκρισίας του *i* ατόμου.

Επίσης γίνεται η υπόθεση:

- T_i και C_i ανεξάρτητες και τυχαίες μεταβλητές
- T_i , …, T_n είναι ανεξάρτητες παρατηρήσεις από την ίδια κατανομή με συνάρτηση πυκνότητας πιθανότητας *f(t)* και συνάρτηση επιβίωσης *S(t).*

- C_i , …, C_n είναι ανεξάρτητες παρατηρήσεις από την ίδια κατανομή με συνάρτηση πυκνότητας πιθανότητας *g(c)* και της συνάρτηση επιβίωσης *G(c)*

Αυτό σημαίνει:

$$Pr(T) = f(t)$$

$$Pr(T > t) = S(t)$$

$$Pr(C) = g(c)$$

$$Pr(C > c) = G(c).$$

Ορίζονται όπως πριν:

$$t_i = min(T_i, C_i)$$

$$\delta_i = \begin{cases} 1, & T_i \leq C_i \\ 0, & T_i > C_i \end{cases}$$

Στο πείραμά λοιπόν παρατηρούνται τα ζεύγη (t_i, δ_i) $i=1,\ldots,n.$
Η συνάρτηση πυκνότητας πιθανότητας για (t_i, δ_i) είναι:

$$Pr(t_i = t, \delta_i = 0) = Pr(C_i = t, T_i > C_i) = g(t)S(t)$$

$$Pr(t_i = t, \delta_i = 1) = Pr(T_i = t, T_i \leq C_i) = f(t)G(t)$$

Από τις δύο τελευταίες σχέσεις προκύπτει:

$$Pr(t_i = t, \delta_\iota) = [f(t)G(t)]^{\delta_i}[g(t)S(t)]^{1-\delta_i}$$

Έτσι, για ανεξάρτητες *n* και παρατηρήσεις $(t_1, \delta_1), \ldots, (t_n, \delta_n)$
η συνάρτηση πιθανότητας είναι:

$$L(\theta) = \prod_{\iota=1}^{n} [f(t_i;\theta)G(t_i;\theta)]^{\delta_\iota}[g(t_i;\theta)S(t_i;\theta)]^{1-\delta_\iota}$$

$$= \underbrace{(\prod_{i=1}^{n} \mathrm{G}(\mathrm{t}_\mathrm{i};\theta)^{\delta_\iota}\, g(t_i;\theta)^{1-\delta_\iota})}_{\text{Χρόνος λογοκρισίας}}\underbrace{(\prod_{i=1}^{n} \mathrm{f}(\mathrm{t}_\mathrm{i};\theta)^{\delta_\iota}\, S(t_i;)^{1-\delta_\iota})}_{\text{Διάρκεια ζωής}}$$

Ενδέχεται το G και το g, τα οποία εκφράζουν τον χρόνο λογοκρισίας, να μην εξαρτώνται από την άγνωστη παράμετρο θ. Σε αυτή την περίπτωση στη συνάρτηση πιθανφάνειας του τυχαία λογοκριμένου δείγματος μπορεί να αγνοηθεί η πρώτη παρένθεση που αφορά τον χρόνο λογοκρισίας και τελικά η συνάρτηση πιθανοφάνειας γίνεται

$$L(\theta) \propto \prod_{i=1}^{n} f(t_i;\theta)^{\delta_\iota} S(t_i;\theta)^{1-\delta_\iota}$$

Παρατηρείται δηλαδή ότι συμπίπτει με τη συνάρτηση πιθανότητας της λογοκρισίας Τύπου I και τύπου II. Δηλαδή η παράγωγος της λογαριθμικής συνάρτησης πιθανοφάνειας θα δώσει τις ίδιες εξισώσεις για την εκτίμηση των αγνώστων παραμέτρων.

Η κατά διαστήματα λογοκριμένη κατανομή

Στην περίπτωση αυτή αντί του δείγματος $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ με σ.π.π. $f(x;\theta)$ αυτό που είναι γνωστό είναι η θέση της x_i στην ευθεία. Ποιο συγκεκριμμένα αν $(-\infty, c_1), [c_1, c_2), \dots . [c_{k-1}, c_k), (c_k, \infty)$ είναι μια διαμέριση της ευθείας τότε για κάθε i ξέρουμε σε ποιό διάστημα ανήκει x_i αλλά όχι την τιμή της δηλαδή αυτό που καταγράφουμε είναι της συχνότητες n_i

Διαστήματα	Συχνότητες
$(-\infty, c_1)$	n_1
$[c_1, c_2)$	n_2
….	…
$[c_{k-1}, c_k)$	n_k
(c_k, ∞)	n_{k+1}

Η συνάρτηση πιθανοφάνειας είναι

$$L = F^{n_1}(c_1)\big(F(c_2) - F(c_1)\big)^{n_2} \dots \big(F(c_k) - F(c_{k-1})\big)^{n_k}\big(1 - F(c_k)\big)^{n_k+1}.$$

Η εκτίμηση των παραμέτρων βρίσκεται με τη μεγιστοποίηση του $\log L$.
Για παράδειγμα αν το δείγμα ακολουθεί την κατανομή Weibull με

$$f(x; a, b) = \frac{x^{a-1}}{b^a} e^{-\left(\frac{x}{b}\right)^a}$$

Η συνάρτηση πιθανοφάνειας είναι

$$L = \left(1 - e^{-\left(\frac{c_1}{b}\right)^a}\right)^{n_1} \left(e^{-\left(\frac{c_1}{b}\right)^a} - e^{-\left(\frac{c_2}{b}\right)^a}\right)^{n_2} \dots \left(e^{-\left(\frac{c_{k-1}}{b}\right)^a} - e^{-\left(\frac{c_k}{b}\right)^a}\right)^{n_k} \left(e^{-\left(\frac{c_k}{b}\right)^a}\right)^{n_k+1}.$$

Η εύρεση του ΕΜΠ απαιτεί την μεγιστοποίηση της

$$\log L = n_1 \log\left(1 - e^{-\left(\frac{c_1}{b}\right)^a}\right) + \sum_{i=2}^{k} n_i \log\left(e^{-\left(\frac{c_{i-1}}{b}\right)^a} - e^{-\left(\frac{c_i}{b}\right)^a}\right) - (n_k + 1)\left(\frac{c_k}{b}\right)^a.$$

Είναι φανερό ότι η μεγιστοποίηση της $\log L$ γίνεται μόνο με αριθμητικές μεθόδους όπως η Newton-Rapshon ή η μέθοδος Scoring.

1.2.3 Ο εκτιμητής product-limit (Kaplan-Meyer)

Να ανακαλέσουμε ότι παρατηρούμε τα ζεύγη

$$(Y_1, \delta_1), \cdots, (Y_n, \delta_n).$$

$$\delta_\iota = \begin{cases} 1 \ \mu\eta\ \lambda o\gamma o\kappa\rho\iota\mu\mu\acute{\varepsilon}\nu\eta\ \tau\iota\mu\acute{\eta} \\ 0 \ \lambda o\gamma o\kappa\rho\iota\mu\mu\acute{\varepsilon}\nu\eta\ \tau\iota\mu\acute{\eta}. \end{cases}$$

Ας υποθέσουμε προς το παρόν ότι δεν υπάρχουν ισότιμες παρατηρήσεις ή δεσμοί (ties).

Διατάσσουμε κάτ' αρχήν τα ζεύγη σε " αύξουσα" σειρά ως προς τα Y_i δηλαδή

$$\binom{Y_{(1)}}{\delta_{(1)}}, \binom{Y_{(2)}}{\delta_{(2)}}, \quad \cdots, \quad \binom{Y_{(n)}}{\delta_{(n)}}$$

Ορίζουμε

$$n_i = \#\ \zeta\omega\nu\tau\alpha\nu\acute{\omega}\nu\ \mu\acute{\epsilon}\chi\rho\iota\ \tau o\nu\ \chi\rho\acute{o}\nu o\ Y_{(i)} -$$

$$d_i = \#\ \theta\alpha\nu\acute{\alpha}\tau\omega\nu\ \tau\eta\ \sigma\tau\iota\gamma\mu\acute{\eta}\ Y_{(i)}$$

Με τη βοήθεια των εκτιμητών

$$\hat{q}_i = \frac{d_{i,}}{n_{i,}}$$

$$\hat{p}_i = 1 - \hat{q}_i = \begin{cases} 1 - \dfrac{1}{n_i} & \alpha\nu\ \delta_{(i)} = 1 \\ 1 & \alpha\nu\ \ \delta_{(i)} = 0, \end{cases}$$

ο K-M εκτιμητής όταν δεν υπάρχουν ισότιμες παρατηρήσεις ή δεσμοί (ties) είναι:

$$\hat{S}(t) = \prod_{y_{(i)} \le t} \hat{p}_i \quad = \prod_{y_{(i)} \le t} \left(1 - \frac{1}{n_i}\right)^{\delta_{(i)}} \quad = \prod_{y_{(i)} \le t} \left(1 - \frac{1}{n - i + 1}\right)^{\delta_{(i)}}$$
$$= \prod_{y_{(i)} \le t} \left(\frac{n - i}{n - i + 1}\right)^{\delta_{(i)}} \tag{1.3}$$

<u>Στην περίπτωση που υπάρχουν ισότιμες παρατηρήσεις ή δεσμοί (ties).</u>

Ο εκτιμητής K-M είναι

$$\hat{S}(t) = \prod_{y'_{(j)} \le t} \left(1 - \frac{d_j}{n_j}\right)^{\delta'_{(j)}} \tag{1.4}$$

όπου

$$Y'_{(1)} < Y'_{(2)} < \cdots < Y'_{(r)}$$

είναι οι <u>διακριτοί</u> χρόνοι των συμβάντων.

Η συνάρτηση επιβίωσης μπορεί να εκτιμηθεί μέσω της αθροιστικής συνάρτησης κινδύνου με τη βοήθεια του εκτιμητής του **Nelson- Aalen**

$$\hat{S}(t) = \mathrm{e}^{-\hat{\Lambda}(t)}$$

όπου

$$\hat{\Lambda}(t) = \sum_{y_{(j)} \le t} \frac{\delta_{(j)}}{n_j}. \tag{1.5}$$

Η ασυμπτωτική διασπορά εκτιμάται από τις σχέσεις

$$\widehat{Var}\left[\hat{\Lambda}(t)\right] = \sum_{j:t_j \le t} \frac{d_j(n_j - d_j)}{n_j^3} \tag{1.6}$$

και

$$\widehat{Var}\left[\hat{\Lambda}(t)\right] = \sum_{j:t_j \le t} \frac{d_j}{n_j^2}. \tag{1.7}$$

Το πολλαπλασιαστικό ή log-προσθετικό μοντέλο κινδύνου (PH) ορίζεται όταν T είναι μια συνεχής μη αρνητική τυχαία μεταβλητή, $\boldsymbol{x}$ ένα διάνυσμα $p \times 1$ σταθερών συμεταβλητών, $\boldsymbol{\beta}$ ένα διάνυσμα $p \times 1$ συντελεστών, και η συνάρτηση επιβίωσης δίνεται από τη σχέση

$$h(t|x) = h_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{x}).$$

Η μέθοδος εκτίμησης είναι ανεξάρτητη της κατανομής με την έννοια ότι οι εκτιμητές που προκύπτουν δεν εξαρτώνται από την κατανομή της T. Για την εκτίμηση των $\boldsymbol{\beta}$ ο Cox(1972) εισήγαγε τη μέθοδο των partial likelihood κατά την οποία δεν απαιτείται η γνώση της $h_0(t)$.

Για δείγμα (t_i, δ_i) $i = 1,2, \dots n$ με k διαφορετικές τιμές και $n-k$ λογοκριμένοι χρόνοι, ο εκτιμητής των $\boldsymbol{\beta}$ είναι λύση του συστήματος

$$\boldsymbol{U}(\boldsymbol{\beta}) \equiv \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \delta_i[\boldsymbol{x_i} - \overline{\boldsymbol{x}}(t_i, \boldsymbol{\beta})] = \boldsymbol{0} \tag{1.8}$$

όπου

$$\overline{\boldsymbol{x}}(t_i, \boldsymbol{\beta}) = \frac{\sum_{\boldsymbol{\ell=1}}^{\boldsymbol{n}} Y_\ell(t)\boldsymbol{x_\ell} e^{\boldsymbol{\beta}'\boldsymbol{x_\ell}}}{\sum_{\boldsymbol{\ell=1}}^{\boldsymbol{n}} Y_\ell(t) e^{\boldsymbol{\beta}'\boldsymbol{x_\ell}}}$$

και $Y_i(t) = I(t_i \ge t)$.

Ο $p \times p$ πίνακας πληροφορίας του Fisher είναι

$$\boldsymbol{I}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left\{ \frac{\sum_{\boldsymbol{\ell=1}}^{\boldsymbol{n}} Y_\ell(t) e^{\boldsymbol{\beta}'\boldsymbol{x_\ell}} [\boldsymbol{x_i} - \overline{\boldsymbol{x}}(t_i, \boldsymbol{\beta})][\boldsymbol{x_i} - \overline{\boldsymbol{x}}(t_i, \boldsymbol{\beta})]'}{\sum_{\boldsymbol{\ell=1}}^{\boldsymbol{n}} Y_\ell(t) e^{\boldsymbol{\beta}'\boldsymbol{x_\ell}}} \right\}$$

Για τα ασυμπτωτικά αποτελέσματα θα χρησιμοποιήσουμε τη μέθοδο δέλτα η οποία περιγράφεται από το θεώρημα που ακολουθεί.

Θεώρημα 1.1 Έστω $\boldsymbol{T_n} = (T_{1n}, T_{2n}, \dots, T_{kn})$ τυχαίο διάνυσμα και $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$ έτσι ώστε

$$\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{\theta}) = \sqrt{n}(T_{1n} - \theta_1, T_{2n} - \theta_2, \dots, T_{kn} - \theta_k) \xrightarrow{D} N(0, \boldsymbol{\Sigma}) \quad n \uparrow \infty$$

όπου

$$\boldsymbol{\Sigma} = \left(\sigma_{ij}\right)_{k \times k}$$

ένας θετικά ορισμένος πίνακας και $g_i(x_1, \dots, x_k)$, $i = 1,2 \dots, p$ συναρτήσεις για τις οποίες υπάρχει η πρώτη παράγωγος. Τότε ισχύει ότι

$$\sqrt{n}\left(g_1(\boldsymbol{T}_n) - g_1(\theta), \dots, g_\kappa(\boldsymbol{T}_n) - g_\kappa(\theta)\right) \xrightarrow{D} N(0, \boldsymbol{G\Sigma G}') \quad n \uparrow \infty$$

όπου $G = \left(\frac{\partial g_i(\boldsymbol{\theta})}{\partial \theta_j}\right)$.

ΚΕΦΑΛΑΙΟ 2

ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΜΟΝΤΕΛΟΥ

2.1 ΕΙΣΑΓΩΓΗ

Ένα σύστημα μπορεί να αποτύχει με πολλούς τρόπους. Για παράδειγμα, ένα άτομο σε μία δημογραφική μελέτη μπορεί να καταγραφεί ως νεκρός στην ηλικία t λόγω καρκίνου, καρδιοαναπνευστικής ασθένειας ή άλλης αιτίας. Στη μετεωρολογία η οποία ενδιαφέρεται για τα ακραία καιρικά φαινόμενα σαν συμβάν μπορεί να θεωρήσει ακραίες περιπτώσεις βροχόπτωσης, χιονοθύελλας ή πολύ δυνατούς ανέμους. Οι τρόποι αποτυχίας μπορούν επίσης να καθοριστούν και με άλλους τρόπους, για παράδειγμα, να αντικατοπτρίζουν το κόστος ή τη σοβαρότητα των συνεπειών που σχετίζονται με την αποτυχία. Ο τρόπος αυτός μπορεί να αναφέρεται στο αίτιο της αποτυχίας, όπου σε αυτή την περίπτωση ορίζεται συχνά ως ανταγωνιστικό ρίσκο. Οι τρόποι αποτυχίας αναφέρονται σαν ανταγωνιστικά ρίσκα. Στην ενότητα αυτή θα περιγράψουμε τα βασικά χαρακτηριστικά των προβλημάτων των πολλαπλών τρόπων αποτυχίας καθώς και τα βασικά μοντέλα που το περιγράφουν με την βοήθεια των οποίων γίνεται στατιστική συμπερασματολογία.

2.2.1 Βασικά Χαρακτηριστικά και επιλογή Μοντέλων

Κάθε άτομο $i = 1,2, \dots n$ όπου n το μέγεθος του δείγματος, περιγράφεται από ένα ζευγάρι τυχαίων μεταβλητών (T_i, d_i) όπου T_i είναι η τυχαία μεταβλητή που περιγράφει τον χρόνος αποτυχίας του i ατόμου και d_i τον τρόπο αποτυχίας του. Προφανώς η T_i συνήθως περιγράφει χρόνο, είναι μία τυχαία συνεχής μεταβλητή με στήριγμα $(0, \infty)$ ενώ η d_i είναι μια διακριτή τυχαία μεταβλητή με τιμές από το διάστημα (1,....., k). Οι διακριτές αυτές τιμές αυτές κωδικοποιούν τους διάφορους τρόπους αποτυχίας. Το πρώτο πρόβλημα που αντιμετωπίζουμε είναι η εύρεση της από κοινού κατανομής των T και d. Αυτή μπορεί να προσδιορισθεί με πολλούς τρόπους. Ο πιο εύχρηστος τρόπος είναι αυτός με τη βοήθεια της συνάρτησης κινδύνου ο ορισμός του οποίου δόθηκε στην εισαγωγή

$$\lambda_j(t) = \lim_{\Delta t \to 0} \frac{\Pr(T < t+\Delta t,\ d = j | T \geq t)}{\Delta t} \tag{2.1}$$

Οι συναρτήσεις αυτές προσδιορίζουν πλήρως την κατανομή των *(T, d)*. Η περιθώρια συνάρτηση κινδύνου για το T είναι

$$\lambda(t) = \sum_{j=1}^{k} \lambda_j(t)$$

και η περιθώρια συνάρτηση επιβίωσης για το Τ είναι επομένως

$$S(t) = Pr(T \geq t) = e^{-\Lambda(t)}, \tag{2.2}$$

όπου $\Lambda(t) = \int_0^t \lambda(u)du$ είναι η αθροιστική συνάρτηση κινδύνου για το T. Φυσικά,

$$\Lambda(t) = \sum_{j=1}^{k} \int_0^t \lambda_j(u)du = \sum_{j=1}^{k} \Lambda_j(t),$$

αφού

$$Pr(T \in [t, t+\Delta t], d = j) = \Pr[T < t+\Delta t, d = j | T \geq t] \Pr(T \geq t),$$

συνεπάγεται ότι

$$F_j(t) = Pr(T \leq t, d = j) = \int_0^t \lambda_j(u)S(u)du \quad (2.3)$$

και

$$f_j(t) = F_j'(t) = \lambda_j(t)S(t)$$

είναι οι περιθώριες συναρτήσεις κατανομής και πυκνότητας πιθανότητας αντίστοιχα για τον j τρόπο αποτυχίας.

Προφανώς τα $F_J(t)'s$ ή $f_j(t)'s$ επίσης προσδιορίζουν την κατανομή των (T, d). Αξίζει να σημειωθεί ότι

$$\pi_j = Pr(d = j) = F_j(\infty) = \int_0^\infty f_j(t)dt\,, \quad j = 1,2,..k \quad (2.4)$$

$$F(t) = 1 - S(t) = \sum_{j=1}^k Pr(T \leq t, d = j) \quad = \quad \sum_{j=1}^k F_j(t), \quad (2.5)$$

και ότι

$$\lambda_j(t) = f_j \,/\, S(t).$$

Το πλεονέκτημα των συναρτήσεις κινδύνου $\lambda_j(t)$ είναι ευκολονόητη και μπορούν εύκολα να ερμηνευθούν. Στις μελέτες ανθρώπινης θνητότητας, για παράδειγμα, αντιπροσωπεύουν τους ρυθμούς θνητότητας από συγκεκριμένα αίτια στην ηλικία t, ανάλογα τις συνθήκες επιβίωσης έως την ηλικία t. Διάφορες άλλες πιθανότητες είναι ιδιαίτερου ενδιαφέροντος όπως είναι οι δεσμευμένης πιθανότητας

$$F_j^*(t) = Pr(T \leq t | d = j) = \frac{1}{\pi_j} F_j(t). \quad (2.6)$$

Η προσέγγιση του προβλήματος μπορεί να είναι είτε παραμετρική είτε μη παραμετρική. Τα παραμετρικά μοντέλα μπορούν να προσδιοριστούν με διάφορους τρόπους. Η πιο συνηθισμένη προσέγγιση είναι να προσδιοριστεί το $\lambda_j(t)$ παραμετρικά. Αυτό σημαίνει ότι προσδιορίζεται η συναρτησιακή μορφή της $\lambda_j(t)$ εκτός ίσως από ένα πεπερασμένο σύνολο παραμέτρων. Εναλλακτικά θα μπορούσε να προσδιοριστεί με ανάλογο τρόπο η κατανομή των $F_J(t)$'s και π_j. Αυτοί δυο τρόποι είναι ισοδύναμοι γιατί είναι γνωστό ότι η συνάρτηση κινδύνου προσδιορίζει μοναδικά η κατανομή και αντιστρόφως.

Οι πιο συνηθισμένες επιλογές για την συνάρτηση κινδύνου είναι αυτές που αντιστοιχούν στις κατανομές της

- Pareto με πυκνότητα πιθανότητας

$$F_j(x) = 1 - \left(\frac{x_0}{x}\right)^{a_j} \quad x > x_0$$

και συνάρτηση κινδύνου

$$h_j(t) = \frac{a_i}{t}$$

- Κατανομής γάμμα $G(a_j, \beta_j)$ με πυκνότητα πιθανότητας

$$f_j(x, a_j, \beta_j) = \frac{1}{\Gamma(a_j)\beta_j^a} x^{a-1} e^{-\frac{x}{\beta_j}} \quad x > 0$$

και συνάρτηση κινδύνου

$$h_i(t) = \frac{t^{\alpha_i - 1} e^{-t}}{\Gamma(\alpha_j) - \Gamma_t(a_j)} \qquad t > 0, \qquad a_j > 0.$$

Εδώ η $\Gamma_x(a_j)$ είναι η incomplete συνάρτηση Γάμμα η οποία ορίζεται ως

$$\Gamma_x(a_j) = \int_0^x t^{a_j - 1} e^{-t} dt.$$

- Κατανομής Weibull με πυκνότητα πιθανότητας

$$f_j(x, a_j, \beta_j) = \frac{1}{\Gamma(\beta_j) a_j^{\beta_j}} x^{\beta_j - 1} e^{-\left(\frac{x}{\alpha_j}\right)^{\beta_j}} \quad x > 0$$

- Και συνάρτηση κινδύνου

$$\lambda_j(t;\ \alpha_j, \beta_j) = \frac{\beta_j}{\alpha_j}\left(\frac{t}{\alpha_j}\right)^{\beta_j - 1} \qquad j = 1, \dots, k \tag{2.7}$$

Στην εργασία αυτή θα χρησιμοποιήσουμε κυρίως την κατανομή Weibull με συνάρτηση κινδύνου την (2.7).

Παράδειγμα 2.1

Η συνηθισμένη προσέγγιση είναι να επιλεγεί η κατάλληλη παραμετρικοποίηση για τις συναρτήσεις κινδύνου (2.1). Για παράδειγμα, χωρίς παρουσία συμμεταβλητών, η συνήθης παραμετρικοποίηση της Weibull δίνεται από τη σχέση (2.7). Για να ενσωματωθούν συμμεταβλητές $\boldsymbol{X}' = (X_1, X_2, \dots, X_m)$ χρησιμοποιούνται πολλαπλασιαστικές υποθέσεις κινδύνων της μορφής

$$\lambda_j(t|\boldsymbol{X}) = \lambda_{0j}(t) e^{\boldsymbol{\beta_j' x}}$$

όπου $\boldsymbol{\beta_j'} = (\beta_{j_1}, \beta_{i_2}, \dots, \beta_{i_m})$ και $\boldsymbol{x}' = (x_1, x, \dots, x_m)$ οι παρατηρήσεις των συμμεταβλητών. Μοντέλα τέτοιων ειδών χρησιμοποιούνται ευρέως και μπορούν εύκολα να ενσωματωθούν σε τυπικό λογισμικό ανάλυσης επιβίωσης. Ένα μειονέκτημα είναι ότι οι κατανομές $F_J(t)$ που δίνονται από τη σχέση (2.3) γίνονται πολύ σύνθετες γιατί ενσωματώνουν όλες τις παραμέτρους.

Μια άλλη παραμετρική προσέγγιση η οποία δεν χρησιμοποιείται συχνά γιατί είναι δύσχρηστη είναι αντί των συναρτήσεων κινδύνων να προσδιορίσουμε τις κατανομές $F_J(t)$ στην (2.6) παραμετρικά και να χειριστούμε το π_j ως επιπρόσθετη παράμετρο. Με τη βοήθεια των σχέσεων (2.5), (2.6) η συνάρτηση κινδύνου μπορεί να γραφτεί

$$\lambda_j(t) = f_j / S(t)$$

$$= \frac{\pi_j f_j^*(t)}{1 - \sum_{l=1}^{k} \pi_l F_l^*(t)} \qquad j = 1, \dots, k \tag{2.8}$$

Για παράδειγμα, μπορούμε να υιοθετήσουμε τις κατανομές Weibull με συναρτήσεις κινδύνου $h_j(t)$ ίδιας μορφής όπως το δεξιό σκέλος της (2.7), όπου σε αυτή την περίπτωση όλοι οι παράμετροι έχουν σχετικά άμεσες ερμηνείες.

Μια ενδιαφέρουσα εφαρμογή του μοντέλου που έχουμε περιγράψει είναι η εξής: Θεωρούμε ένα σύστημα με k εξαρτήματα όπου για το καθένα η τυχαία μεταβλητή T_j υποδηλώνει τη διάρκεια ή χρόνο αποτυχίας του εξαρτήματος j ($j = 1,.....,k$). Το σύστημα αποτυγχάνει όταν αποτυγχάνει το πρώτο στοιχείο, επομένως η διάρκεια είναι T = min (T_1,…,T_k). Μπορούμε ακόμα να βάλουμε στο μοντέλο μας μια ακόμα μεταβλητή η οποία να καταγράφει ποιο εξάρτημα χάλασε: Έστω d = j σημαίνει ότι χάλασε το εξάρτημα j έτσι ώστε T = T_j. Αυτό το πλαίσιο φαίνεται ενδιαφέρον, καθώς μοιάζει να μπορούμε να λάβουμε υπόψιν πολυμεταβλητά μοντέλα F (t_1,….,t_k) για την από κοινού κατανομή του (T_1,…,T_k). Σε ένα σειριακό σύστημα στο οποίο θα μπορούσε να βρει εφαρμογή το παραπάνω μοντέλο, αυτό που παρατηρούμε είναι το ζευγάρι (T, d). Από τη γνώση αυτών των δυο τυχαίων μεταβλητών δεν μπορούμε να υπολογίσουμε την από κοινού κατανομή F (t_1,….,t_k) γιατί έχουμε πρόβλημα ταυτοποίησης (identifiability problem). Αυτό συμβαίνει γιατί (T_i, C_i) ότι δύο διαφορετικές κατανομές F (t_i,….,t_k) μπορούν να αποδώσουν την ίδια κατανομή για (T, d). Είναι επίσης απίθανο να καθορίσουμε εάν τα T_1,….,T_k είναι αμοιβαία ανεξάρτητα ή όχι. Για κάθε κατανομή που περιλαμβάνει μη ανεξάρτητο T_j, υπάρχει κατανομή με ανεξάρτητο T_j που αποδίδει την ίδια κατανομή των (T, d).

2.2.2 Συνάρτησης Πιθανοφάνειας στην παραμετρική περίπτωση

Υποθέτουμε ότι οι παρατηρήσεις γίνονται από τυχαίο δείγμα n συμμετεχόντων υπό την παρουσία δεξιάς λογοκρισίας. Θεωρούμε καταρχάς το απλό μοντέλο χωρίς συμμεταβλητές. Εάν το T_i είναι λογοκρυμμένο στο t_i, τότε δεν είναι γνωστό αν έχει αποτύχει το i εξάρτημα. Συνεπώς τα δεδομένα για το άτομο (ή εξάρτημα) i αποτελούνται είτε από (T_i=t_i, d_i) είτε $T_i > t_i$. Επομένως η συνάρτηση πιθανοφάνειας υπό την παρουσία ανεξάρτητης λογοκρισίας είναι

$$L = \prod_{i=1}^{n} f_{d_i}(t_i)^{\delta_i} S(t_i)^{1-\delta_i}, \tag{2.9}$$

όπου δ_i = 1 εάν t_i ο χρόνος αποτυχίας και 0 εάν είναι ο χρόνος λογοκρισίας. Λαμβάνοντας υπόψη την (2.2) ότι δηλαδή

$$S(t) = \exp\left\{-\sum_{j=1}^{k} \Lambda_j(t)\right\}$$

$$= \prod_{j=1}^{k} G_j(t), \tag{2.10}$$

όπου $G_j(t)$ = exp{—Λ_j(t)} η συνάρτηση πιθανότητα (2.9) μπορεί να γραφτεί εκ νέου χρησιμοποιώντας το συμβολισμό δ_{ij} = I (C_i = j), ως

$$L = \prod_{i=1}^{n} \prod_{j=1}^{k} f_j(t_i)^{\delta_{ij}} S(t_i)^{1-\delta_i}$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{k} \lambda_j(t_i) S(t_i)]^{\delta_{ij}} S(t_i)^{1-\delta_i}.$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{k} \lambda_j(t_i)]^{\delta_{ij}}.$$

Παραγωγίζοντας τη σχέση $G_j(t) = \exp\{—\Lambda_j(t)\}$ έχουμε

$$g_j(t) = \lambda_j(t) G_j(t) = -G_j'(t) \tag{2.11}$$

Με τη βοήθεια της (2.11) η συνάρτηση πιθανοφάνειας γράφεται

$$L = \prod_{i=1}^{n} \prod_{j=1}^{k} g_j(t_i)^{\delta_{ij}} G_j(t_i)^{1-\delta_{ij}}, \tag{2.12}$$

Αξίζει να σημειωθεί ότι αντιστρέφοντας τη σειρά των γινομένων στην (2.12) η συνάρτηση πιθανοφάνειας γράφεται σαν

$$L = \prod_{j=1}^{k} L_j$$

όπου

$$L_j = \prod_{i=1}^{n} g_j(t_i)^{\delta_{ij}} G_j(t_i)^{1-\delta_{ij}}. \tag{2.13}$$

Παρατηρούμε ότι η L_j είναι η συνάρτηση πιθανοφάνειας που αντιστοιχεί σε ένα λογοκριμμένο δείγμα που αντιστοιχεί σε ένα συγκεκριμένο τρόπο αποτυχίας.
Η μορφή της συνάρτησης πιθανοφάνειας L, στην (2.11) δείχνει ότι τα $\lambda_j(t)$ και $\Lambda_j(t)$ υπολογίζονται από τα δεδομένα στο (T,C). Επιπροσθέτως, εάν τα $\lambda_j(t)$ περιλαμβάνουν ξεχωριστούς παραμέτρους θ_j για $j = 1,\ldots,k$ (και παρομοίως για τα $Gj\ (t)$ και $g_j(t)$), τότε η $L\ (\theta_1,\ldots\ldots\theta_k)$ παραγοντοποιείται σε ξεχωριστά μέρη $L_j(\vartheta_j)$. Αυτό είναι πολύ βολικό στην συμπερασματολογία αφού λογαριθμώντας και παραγωγίζοντας η κάθε παράγωγος έχει μόνο τη μεταβλητή θ_j καθιστώντας έτσι την εκτίμησή της εύκολη. Η συμπερασματολογία λοιπόν για το θ_j λοιπόν μπορεί να βασίζεται στην (2.13). Αυτό σημαίνει ότι για κάθε (τρόπο αποτυχίας) j, η αποτυχία του j στο t_i καταγράφεται ως αποτυχία, και η αποτυχία οποιουδήποτε άλλου τρόπου ως censoring. Η παραμετρική παρέμβαση για μοντέλα τέτοιου τύπου είναι εύκολη στην εφαρμογή της χρησιμοποιώντας μεθόδους και λογισμικά ανάλυσης επιβίωσης.
Παρομοίως, οι μη παραμετρικές και ημιπαραμετρικές μέθοδοι μπορούν να εφαρμοστούν για μοντέλα με βάση τα $\lambda_j(t)$.
Έχει ενδιαφέρον να παρατηρήσει κανείς λαμβάνοντας υπόψη τις (2.8) και (2.11) ότι όταν δεν υπάρχει λογοκρισία, η συνάρτηση πιθανοφάνειας δεν παραγοντοποιείται έτσι ώστε να διαχωρίζονται τα θ_j καθιστώντας το πρόβλημα πιο δύσκολο.

Ας δούμε τη μορφή της συνάρτησης πιθανοφάνειας στην περίπτωση κατά την οποία τα δεδομένα στους χρόνους ζωής είναι διακριτά ή συνεχή και ομαδοποιημένα. Η τελευταία περίπτωση αναφέρεται σαν λογοκρισία κατά διαστήματα (interval censoring). Συγκεκριμένα, υποθέτουμε ότι τα μεσοδιαστήματα $I_l = (\alpha_{i-1}, a_l]$ καθορίζονται για l = 1,….., m + 1, με $0 = a_0 < a_1 < a_m < a_{m+1} = \infty$. Αυτό που καταγράφεται είναι ο αριθμό των αποτυχιών από τους 1,....,k τρόπους σε καθένα από τα μεσοδιαστήματα I_1, ..., I_m.
Για $l = 1, \ldots . , m$ και $j = 1, \ldots . , k$ ορίζουμε τις ποσότητες

$$\pi_{jl} = \Pr\{\ \mu\iota\alpha\ \mu o\nu\acute{\alpha}\delta\alpha\ \alpha\pi o\tau\upsilon\gamma\chi\acute{\alpha}\nu\varepsilon\iota\ \sigma\tau o\ \delta\iota\acute{\alpha}\sigma\tau\eta\mu\alpha\ \ I_l\ \ \mu\varepsilon\ \tau o\nu\ \ j\ \tau\rho\acute{o}\pi o\} \tag{2.14}$$

$$= \int_{a_{l-1}}^{a_l} f_j(u)\, du$$

$$= \int_{a_{l-1}}^{a_l} f_j^*(u) du.$$

(Προφανώς στη διακριτή περίπτωση τα ολοκληρώματα αντικαθίστανται με αθροίσματα)

Η συνάρτηση πιθανοφάνειας βασιζόμενη σε n ανεξάρτητα άτομα είναι

$$L = \left\{ \prod_{l=1}^{m} \prod_{j=1}^{k} \pi_{jl}^{d_{jl}} \right\} S(a_m)^{d_{m+1}}, \tag{2.15}$$

οπου d_{jl} ο αριθμός των j αποτυχιών σε I_l και

$$d_{m+1} = n - \sum_{l=1}^{m} \sum_{j=1}^{k} d_{jl}$$

ο αριθμός των ατόμων που επιβίωσαν πέραν του χρόνου a_m. Η συνάρτηση πιθανότητας είναι πολυωνυμικής (multinomial) μορφής και εάν τα δεδομένα προέκυψαν από ομαδοποιημένους συνεχείς χρόνους ζωής, τότε τα μοντέλα στα οποία οι κατανομές $Fj(t)$ χωρίζονται παραμετρικά, είναι αρκετά χρήσιμα. Η περίπτωση διακριτών κατανομών των χρόνων καλύπτεται επίσης από τη συνάρτηση στην (2.15). Σε αυτή την περίπτωση, ο ορισμός (2.14) έχει νόημα αλλά οι δύο σειρές που τον ακολουθούν δεν πρέπει να ληφθούν υπόψη.

ΚΕΦΑΛΑΙΟ 3

ΜΗ ΠΑΡΑΜΕΤΡΙΚΕΣ ΜΕΘΟΔΟΙ

Στο κεφάλαιο αυτό γίνεται μια ανασκόπηση των μη παραμετρικών μεθόδων όπως αναπτύχθηκαν σε μια σειρά εργασιών από τους Nelson (1969), Aalen (1976) Aalen και Johansen (1978) και Fleming (1978a,b) Ο Andersen και οι συνεργάτες του (1993, Εν. 4.4) ανέπτυξαν αυτές τις ιδέες με αρκετή λεπτομέρεια. Ο Matthews (1988) δημιούργησε εμπειρικά διαστήματα εμπιστοσύνης πιθανοτήτων για συναρτήσεις υποκατανομών.

Θεωρούμε ένα λογοκριμένο δείγμα από τα (Ti, C_i), όπως περιγράφηκε στην Ενότητα 2.2. Με βάση το γεγονός ότι η συνάρτηση πιθανότητας (2.1) παραγοντοποιείται σε ξεχωριστά κομμάτια (2.13) για κάθε τρόπο αποτυχίας, η χρήση μη παραμετρικών μεθόδων είναι προφανής. Πιο συγκεκριμένα, το κομμάτι (2.13) έχει τη μαθηματική μορφή λογοκριμένων δεδομένων πιθανότητας για κατανομή των χρόνων με τις συναρτήσεις επιβίωσης, πυκνότητας και κινδύνου Gj (t) , $g_j(t)$ και $\lambda_j(t)$ αντίστοιχα. Στη συνέχεια με βάση τα δεδομένα (t_i, δ_{ij}), i = 1, ….. , n. η $G_i(t)$ εκτιμάται με τη χρήση της μεθόδου Kaplan – Meier. Αξίζει να σημειωθεί ότι η $G_i(t)$ δεν είναι η συνάρτηση επιβίωσης για κάποια τυχαία μεταβλητή. Όμως παίρνοντας το λογάριθμο αυτού έχουμε την συνάρτηση αθροιστικού κινδύνου $\Lambda_j(t)$ δηλαδή $\log G_j(t) = -\Lambda_j(t)$. Έτσι χρησιμοποιούμε την Kaplan – Meier εκτίμηση του $G_i(t)$ και στη συνέχεια λογαριθμόντας εκτιμούμε την $\Lambda_j(t)$. Όμως στην πράξη χρησιμοποιείται η εκτίμηση Nelson – Aalen για την αθροιστική συνάρτηση κινδύνου. Από την (1.5) , αυτό παίρνει τη μορφή

$$\hat{\Lambda}_j(t) = \sum_{i:t_i \le t} \frac{\delta_{ij}}{n_i}, \qquad j = 1, \ldots, k, \tag{3.1}$$

όπου $Y_i(t) = I(t_l \ge t)$ και $n_i = \sum_{l=1}^{n} Y_l(t_i)$ ο αριθμός των μη censoring και εν ζωή ατόμων ακριβώς πριν το χρόνο t_i. Η εκτίμηση της διασποράς με βάση την (1.7) είναι

$$\widehat{Var}\left[\hat{\Lambda}_j(t)\right] = \sum_{i:t_i \le t} \frac{\delta_{ij}}{{n_i}^2} \tag{3.2}$$

και συνήθως χρησιμοποιείται με την (3.1) για την κατασκευή διαστημάτων εμπιστοσύνης.
Η περιθώρια συνάρτηση επιβίωσης $S(t)$ για την T μπορεί να εκτιμηθεί εύκολα αγνοώντας τους συσχετιζόμενους τρόπους αποτυχίας και χρησιμοποιώντας την εκτίμηση Kaplan – Meier με βάση τα δεδομένα (t_i, δ_i), i = 1, …. , n. Αυτό μας δίνει

$$\hat{S}(t) = \prod_{i:t_i \le t} \left(\frac{n_i' - d_i'}{n_i'} \right) \tag{3.3}$$

όπως στην (1.4) όπου $t_{(l)} <$ $<$ $t_{(k)}$ είναι οι διακριτοί χρόνοι στους οποίους γίνεται η αποτυχία, και d_l' και n_l' οι αριθμοί των αποτυχιών και των ατόμων σε ρίσκο στο $t_{(i)}$ αντίστοιχα. Η εκτίμηση της διασποράς σε συνδυασμό με την $\hat{S}(t)$ μπορεί να χρησιμοποιηθεί για την κατασκευή διαστημάτων εμπιστοσύνης της $S(t)$. Μία εναλλακτική εκτίμηση της συνάρτησης επιβίωσης S(t) είναι

$$\hat{S}(t) = \exp\left[\hat{\Lambda}(t)\right] = \exp\left[-\sum_{j=l}^{k} \hat{\Lambda}_j(t) \right] \tag{3.4}$$

με τα $\hat{\Lambda}_j(t)$ να δίνονται από την (3.1).

Μια εκτίμηση της αθροιστικής συνάρτησης επίπτωσης $F_j(t)$ της (2.3) είναι

$$\hat{F}_j(t) = \int_0^t \hat{S}(u) d\,\hat{\Lambda}_j(u).$$

Με τη βοήθεια της (3.1) η τελευταία σχέση γράφεται

$$\hat{F}_j(t) = \sum_{i:t_i \le t} \hat{S}(t_i)\frac{\delta_{ij}}{n_i}, \qquad j = 1, \ldots\ldots, k. \tag{3.5}$$

Εάν υπάρχει μόνον ένας τρόπος αποτυχίας, τότε η (3.5) είναι ίση με 1 - $\hat{S}$ (t +), όπου η $\hat{S}(t)$ δίδεται από την (3.3). Εάν υπάρχουν $k \geq 2$ τρόποι αποτυχίας, αλλά όχι censoring, $\sum \hat{F}_j(t)$ ισούται με 1 - $\hat{S}$ (t).

Η εκτίμηση της διακύμανσης της $\hat{F}_j\ (t)$ είναι δύσκολο πρόβλημα. Ο αναγνώστης παραπέμπεται στην βιβλίο των Anderson et al. (1993 pp. 298-304) οι οποίοι με μεθόδους από την θεωρία απαρίθμησης προτείνουν έναν τρόπο εκτίμησης αυτής. Μία εναλλακτική προσέγγιση είναι να θεωρήσουμε ένα μοντέλο στο οποίο οι συναρτήσεις κινδύνου $\lambda_j(t)$ να είναι κατά τμήματα σταθερές. Στη συνέχεια με τη μέθοδο μέγιστης πιθανοφάνειας βρίσκουμε συνεπείς μη παραμετρικές εκτιμήσεις των παραμέτρων και των διακυμάνσεων αυτών Στη συνέχεια βρίσκουμε το όριο επιτρέποντας τον αριθμό των διαστημάτων να αυξάνει και συγχρόνως το μήκος τους να ελαττώνεται. Τα βήματα της μεθόδου είναι τα εξής:

Καθορίζουμε τα μεσοδιαστήματα $I_\ell = [a_{\ell-1},\ a_\ell)$ για $\ell = 1, \ldots.\ , m,$ με $0 = a_0 < a_0 <$ $< a_m$ όπου το a_m είναι μία κατάλληλα μεγάλη τιμή. Υποθέτωντας ότι τα $\lambda_j(t)$ είναι κατά τμήματα σταθερές δηλαδή,

$$\lambda_j(t) = \lambda_{j\ell}, \qquad t \in I_l \tag{3.6}$$

για j = 1, ….. , k και ℓ = 1, …. , m. Είναι δηλαδή μια κλιμακωτή συνάρτηση. Οι αθροιστικές συναρτήσεις κινδύνου είναι

$$\Lambda_j(t) = \sum_{l=1}^{m} \lambda_{j\ell}\, \Delta_\ell(t), \tag{3.7}$$

όπου $\Delta_\ell(t) = \int_{a_{\ell-1}}^{a_\ell} I(u \le t)du$ είναι το μήκος της τομής του I_ℓ και [0, t). Με άλλα λόγια $\Delta_\ell(t) = \alpha_\ell^* - \alpha_{\ell-1}^*$. Από το (2.12), η συνάρτηση πιθανότητας μπορεί να γραφτεί ως

$$L = \prod_{j=1}^{k}\prod_{i=1}^{n} \lambda_j(t_i)^{\delta_{ij}} e^{-\Lambda_j(t_i)}$$

και επομένως

$$\frac{\partial log L}{\partial \lambda_{j\ell}} = \frac{d_{j\ell}}{\lambda_{j\ell}} - \Delta_\ell,$$

όπου $\Delta_\ell = \sum_{l=1}^{n} \Delta_\ell(t_{\mathrm{i}})$ και $d_{j\ell} = \sum_{l=1}^{n} I(t_1 \epsilon\, I_\ell)\delta_{ij}$. Συμπεραίνουμε λοιπόν ότι οι εκτιμητές μέγιστης πιθανοφάνειας είναι

$$\hat{\lambda}_{j\ell} = \frac{d_{j\ell}}{\Delta_\ell} \quad j = 1, \ldots, k; \quad \ell = 1, \ldots, m\,. \tag{3.8}$$

Για να βρούμε την ασυμπτωτική κατανομή θα πάρουμε τις δεύτερες παραγώγους.

$$\frac{\partial^2 logL}{\partial \lambda_{j\ell}^2} = -\frac{d_{j\ell}}{\lambda_{j\ell}^2}$$

Αντικαθιστώντας την (3.8) έχουμε

$$\frac{\partial^2 logL}{\partial \lambda_{j\ell}^2} = -\frac{d_{j\ell}}{\Delta_\ell^2}.$$

Συμπεραίνουμε ότι

$$\widehat{Var}(\hat{\lambda}_{j\ell}) = \frac{d_{j\ell}}{\Delta_\ell^2} \tag{3.9}$$

Καθώς επίσης και ότι οι εκτιμητές είναι ασυμπτωτικά ανεξάρτητες.

Για την ασυμπτωτική διασπορά $\widehat{Var}[\hat{F}_j(t)]$ θα χρησιμοποιήσουμε την μέθοδο δέλτα που περιγράφεται από το Θεώρημα 1.1. Θα χρειαστούμε τις παραγώγους

$$w_{r\ell}^{(j)} = \frac{\partial F_j(t)}{\partial \lambda_{r\ell}}.$$

Παραγωγίζοντας κάτω από το ολοκλήρωμα στην (2.3), βρίσκουμε

$$w_{r\ell}^{(j)} = \int_0^t \{I(j=r)S(u)I(u \in I_\ell) - \Delta_\ell(u)\lambda_j(u)S(u)\}du. \tag{3.10}$$

Μια εκτίμηση της διακύμανσης για το $\hat{F}_j(t)$ μπορεί να βρεθεί υπολογίζοντας το $w_{r\ell}^{(j)}$ χρησιμοποιώντας τις εκτιμήσεις $\hat{\lambda}_{j\ell}$ αντί των $\lambda_{j\ell}$. Με το διάνυσμα $\widehat{\boldsymbol{w}}^{(j)}$ να αντιπροσωπεύει το $w_{r\ell}^{(j)}$ με κάποια προκαθορισμένη σειρά και το διάνυσμα $\hat{v}$ να αντιπροσωπεύει τις αποκλίσεις (3.9) με την ίδια σειρά, παίρνουμε

$$\widehat{Var}[\hat{F}_j(t)] = \widehat{\boldsymbol{w}}^{(j)\prime} diag(\hat{\boldsymbol{v}})\widehat{\boldsymbol{w}}^{(j)} \tag{3.11}$$

ως εκτίμηση, όπου $\widehat{\mathrm{w}}^{(\mathrm{j})\prime}$ είναι το $km \times 1$ διάνυσμα

$$\widehat{\boldsymbol{w}}^{(j)\prime} = \left(\frac{\partial F_1(t)}{\partial \lambda_{11}}, \dots, \frac{\partial F_1(t)}{\partial \lambda_{1m}}, \dots, \frac{\partial F_k(t)}{\partial \lambda_{k1}}, \dots, \frac{\partial F_k(t)}{\partial \lambda_{km}}\right)$$

και $diag(\hat{\boldsymbol{v}})$ είναι ο μπλοκ διαγώνιος πίνακας διάστασης $mk \times mk$ όπου ο διαγώνιος πίνακας στη θέση (r,r) είναι ο τετραγωνικός πίνακας διάστασης $m \times m$ ο οποίος στην ij θέση έχει το στοιχείο $\frac{d_{ir}}{\Delta_r^2}$.

Για να πάρουμε μια μη παραμετρική εκτίμηση της διασποράς της (3.5) θα πάρουμε το όριο $m \to \infty$ και τα μήκη των μεσοδιαστημάτων να πλησιάσουν το 0.

Για $\Delta_\ell^* = |a_\ell - a_{\ell-1}|$ μικρό, έχουμε $\Delta_\ell(t) = \Delta_\ell^* I(t \geq a_\ell)$ και από την (3.10),

$$w_{r\ell}^{(j)} = I(j=r)\Delta_\ell^* S(a_\ell) I(t \geq a_\ell) - \sum_{u:a_u \leq t} S(a_u)\,\Delta_l a_u \lambda_j(a_u)\Delta_u^*$$

$$= I(t \geq a_\ell)\Delta_l^* \left\{ I(j = r)S(a_\ell) - \sum_{a_u = a_\ell}^{l} f_j(a_u)\Delta_u^* \right\}$$

$$= I(t \geq a_\ell)\Delta_l^* \{ I(j = r)S(a_\ell) - [F_j(t) - F_j(a_\ell)] \}$$

Αντικαθιστώντας το τελευταίο στην (3.11) έχουμε,

$$\widehat{Var}\left[\hat{F}_j(t)\right] = \sum_{r=1}^{k} \sum_{\ell=1}^{nt} I(t \geq a_\ell)(\Delta_\ell^*)^2 \{ I(j = r)\hat{S}(a_\ell) - [\hat{F}_j(t) - \hat{F}_j(a_\ell)] \}^2 \frac{d_{j\ell}}{\Delta_\ell^2}.$$

Σημειώνοντας ότι $\Delta_l = \Delta_l^* Y_\cdot(a_l)$, όπου $Y.(u) = \sum I(t_l \geq u)$ ο αριθμός τον ατόμων σε κίνδυνο στο χρόνο u, υπολογίζουμε, μόλις πάρουμε το όριο ως $\Delta_l^* \rightarrow 0$ και m $\rightarrow \infty$, την εκτίμηση διακύμανσης

$$\widehat{Var}\left[\hat{F}_j(t)\right] = \sum_{r=1}^{k} \int_0^t \hat{S}(u)^2 \left\{ I(j = r) - \frac{\hat{F}_j(t) - \hat{F}_j(u)}{\hat{S}(u)} \right\}^2 \frac{dN_j(u)}{Y.(u)^2}, \qquad (3.12)$$

όπου $dN_j(u)$ ο αριθμός των j τρόπων αποτυχίας στο χρόνο u.
Οι παράμετροι π_j μπορούν να υπολογιστούν από την (2.4)

$$\hat{\pi}_j = \hat{F}_j(\infty) \quad j = 1, \dots, k. \qquad (3.13)$$

Όμως από την (3.5) αυτό δεν είναι εκτιμήσιμο εκτός κι αν η μέγιστη παρατηρητέα τιμή είναι ένας t χρόνος αποτυχίας, τέτοιος ώστε $\hat{S}(t) = 0$ για t > τ. Αν ο μέγιστος χρόνος είναι χρόνος λογοκρισίας, το $\hat{S}(t)$ είναι απροσδιόριστο πέρα από αυτό. Ακόμη και αν $\hat{S}(t) = 0$ για t > τ, οι εκτιμήσεις $\hat{\pi}_j$ δεν ικανοποιούν το $\sum \hat{\pi}_j = 1$, όταν υπάρχει λογοκρισία. Μία συνηθισμένη διαδικασία είναι να κανονικοποιήσουμε εκ νέου τις εκτιμήσεις ως $\hat{\pi}_j' = \hat{\pi}_j / \sum_{l=1}^{k} \hat{\pi}_l$ σε αυτή την περίπτωση.

Η μη παραμετρική εκτίμηση των αθροιστικών κινδύνων μπορεί να γίνει ακόμα και όταν οι χρόνοι ζωής είναι αριστερά περικομμένοι, ακριβώς όπως και την περίπτωση των Nelson - Aalen και Kaplan – Meier. Εάν οι χρόνοι ζωής όλων των ατόμων είναι αριστερά περικομμένοι, μόνο οι συναρτήσεις $\Lambda_j(t) - \Lambda_j(u_{min})$ είναι εκτιμήσιμες, όπου u_{min} ο ελάχιστα αριστερά-μειωμένος χρόνος στα δεδομένα.

Για τη συγκεκριμένη περίσταση ένας έλεγχος σε αυτό είναι κατά μία έννοια πιθανός, αφού ο Nelson (1970b) αναφέρθηκε σε επιπρόσθετα δεδομένα από τέστ που πραγματοποιήθηκαν μετά από συγκεκριμένες βελτιώσεις στο appliance. Αυτά τα δεδομένα δείχνουν ότι οι δύο αυτές συναρτήσεις χαρακτηριστικού κινδύνου εδώ, επηρεάστηκαν σημαντικά από αυτές τις αλλαγές.

Η παραμετρική μοντελοποίηση των $\Lambda_j(t)$ και άλλων μεγεθών είναι επίσης πιθανή. Κανένα συγκεκριμένο μοντέλο δεν προτείνεται στη βιβλιογραφία γι αυτό το πρόβλημα, παρότι ευέλικτες μορφές όπως η (9.1.7) μπορούν να επιλεχθούν. Είναι μικρής σημασίας να γίνει αυτό εδώ, αφού η κατάσταση αναφορικά με τους μηχανισμούς και τους τρόπους αποτυχίας αλλάζει σημαντικά όταν γίνονται τροποποιήσεις στο appliance.

Παρόμοιοι μέθοδοι με τους παραπάνω μπορούν επίσης να εφαρμοστούν σε ομαδοποιημένα δεδομένα. Αυτόγίνεται στην επόμενη ενότητα σε παραμετρικές μεθόδους, αφού η ομαδοποίηση των χρόνων ζωής σε πεπερασμένους αριθμούς μεσοδιαστημάτων τα κάνει μοντέλα πεπερασμένης διάστασης.

ΚΕΦΑΛΑΙΟ 4

ΕΠΙΛΥΣΗ ΠΑΡΑΜΕΤΡΙΚΟΥ ΜΟΝΤΕΛΟΥ

4.1 ΠΑΡΑΜΕΤΡΙΚΕΣ ΜΕΘΟΔΟΙ

Στο κεφάλαιο αυτό περιγράφονται με παραμετρικά μοντέλα κινδύνου με πολλούς τρόπους αποτυχίας όπως αναπτύχθηκαν σε μια σειρά εργασιών από τους Nelson (1982), Crowder (2001) Dinse (1985), David και Moeschberger (1978). Ο Seal (1977), παρέχει μία ιστορική ανασκόπηση και πολλές λεπτομέρειες μπορούν να βρεθούν σε βιβλία όπως των Elandt-Johnson και Johnson (1980), Manton και Stallard (1988) και Namboodiri και Suchindran (1987).

Όπως σχολιασθεί στο Κεφάλαιο 2, τα παραμετρικά μοντέλα για συνεχή δεδομένα χρόνου για τα οποία οι συναρτήσεις κινδύνου (2.1) προσδιορίζονται ως $\lambda_j\ (t;\theta_j)$, με $\theta_1, \ldots, \theta_k$ είναι εύκολα διαχειρήσιμα. Δηλαδή στην περίπτωση αυτή η παράμετρος θ_j εμφανίζεται μόνο στη συνάρτηση λ_j. Απο τις (2.11) και (2.12), η συνάρτηση πιθανοφάνειας γράφεται στη μορφή

$$L(\theta_1. \ldots, \theta_k) = \prod_{j=1}^{k} L_j(\theta_j) \tag{4.1}$$

με το $L_j(\theta_j)$ να δίνεται από την (2.13). Μοντέλα για τα οποία το $\lambda_j\ (t;\theta_j)$ είναι από Weibull, τα log-logistic, log-normal και μερικές άλλες συνήθεις μορφές μπορούν να ενσωματωθούν με τη χρήση τυπικών λογισμικών ανάλυσης επιβίωσης, όπως συζητήθηκε στο Κεφάλαιο 2. Για τη Weibull έχουμε

$$\lambda_j(t;\ \alpha_j, \gamma_j) = \frac{\gamma_j}{\alpha_j}\left(\frac{t}{\alpha_j}\right)^{\gamma_j - 1} . \tag{4.2}$$

Τα μοντέλα παλινδρόμησης για τα $\lambda_j(t)$ είναι εύκολα διαχειρήσιμα. Παρατηρούμε ότι για $\gamma_j < 1$ είναι φθίνουσα, για $\gamma_j = 1$ είναι σταθερά (εκθετική κατανομή) και για $\gamma_j > 1$ φθίνουσα. Έχουμε δηλαδή την ευελιξία να συνδυάσουμε συναρτήσεις κινδύνου με διάφορες συμπεριφορές. Στην περίπτωση παρουσίας συμμεταβλητών $\boldsymbol{x}$ τα $\lambda_j(t)$, $g_j(t)$, και $G_j(t)$ στην (2.11) αντικαθίστανται με $\lambda_j(t|\boldsymbol{x})$, $g_j(t|\boldsymbol{x})$, και $G_j(t|\boldsymbol{x})$ αντίστοιχα. Μια τέτοια περίπτωση είναι τα parametric accelerated failure time μοντέλα. Έχει παραμέτρους θ_j=(β_j, γ_j) όπου

$$\lambda_j(t|\boldsymbol{x};\ \theta_j) = \frac{\gamma_j}{\alpha_j(\boldsymbol{x})}\left(\frac{t}{\alpha_j(\boldsymbol{x})}\right)^{\gamma_j - 1} \tag{4.3}$$

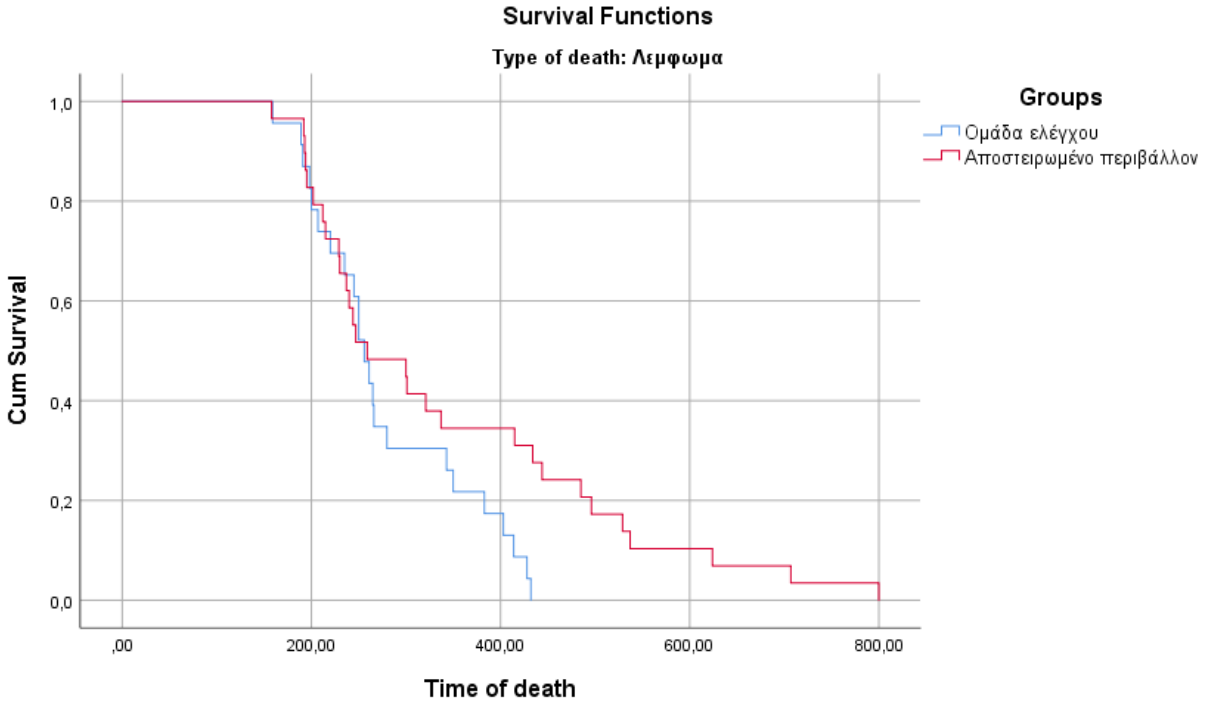
και $a_j(\boldsymbol{x}) = \exp(\boldsymbol{\beta_j'x})$.

Η εκτίμηση ποσοτήτων όπως οι συναρτήσεις κατανομής $F_j(t)$ ή $F_j^*(t)$ είναι εύκολη λόγω του ότι η Weibull έχει συνάρτηση κατανομής σε κλειστή μορφή. Οι εκτιμήσεις όμως των παραμέτρων δεν είναι εύκολη υπόθεση. Επειδή οι εξισώσεις που καλούμαστε να λύσουμε είναι πολύπλοκες, η εκτίμηση των $\theta_1, \ldots, \theta_k$ γίνεται με τη βοήθεια επαναληπτικών μεθόδων προσέγγισης όπως Newton-Rapshon ή μέθοδος scoring. Οι εκτιμήσεις διακύμανσης μπορούν να υπολογιστούν από μία εφαρμογή ενός Θεωρήματος. Η μεθοδολογία bootstrap είναι μία εναλλακτική προσέγγιση για τον υπολογισμό εκτιμήσεων διακύμανσης ή διαστημάτων εμπιστοσύνης (confidence intervals).

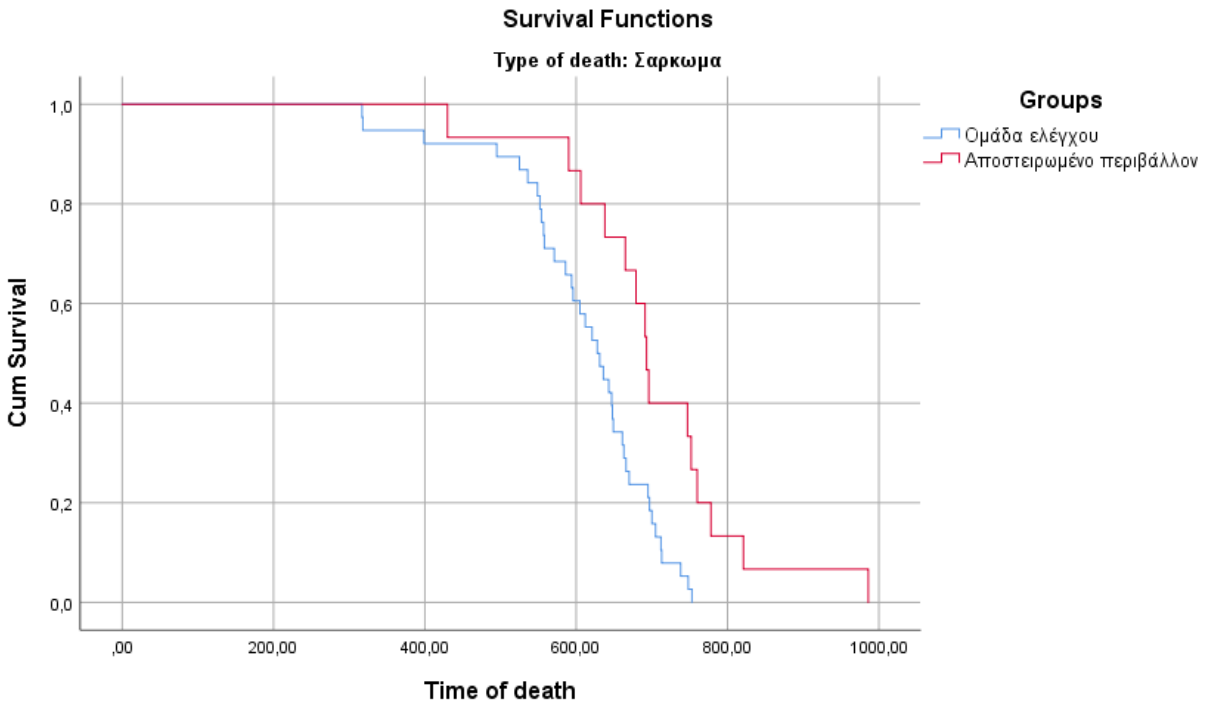
Παράδειγμα 4.1

Τα παρακάτω στοιχεία παραθέτονται στον Appendix (Παράρτημα) και είναι από το βιβλίο Lawless(2002) καθώς και από το άρθρο Hoel .
Δείχνουν τους χρόνους επιβίωσης για δύο ομάδες εργαστηριακών ποντικιών, τα οποία όλα εκτέθηκαν σε συγκεκριμένη δόση ραδιενέργειας σε ηλικία 5 με 6 εβδομάδων. Η πρώτη ομάδα ποντικιών ζούσε σε ένα παραδοσιακό χώρο εργαστηρίου και η δεύτερη ομάδα κρατήθηκε σε αποστειρωμένο περιβάλλον. Το αίτιο θανάτου για κάθε ποντίκι αντιστοιχήθηκε μετά από αυτοψία σε ένα από τα τρία: λέμφωμα, σάρκωμα (C_2) και άλλα αίτια (C_3). Όλα τα ποντίκια απεβίωσαν μέχρι το τέλος του πειράματος, οπότε δεν υπάρχει λογοκρισία.
Σκοπός μας είναι να συγκρίνουμε τις θνησιμότητες των τριών αιτιών θανάτου στα δυο περιβάλλοντα.

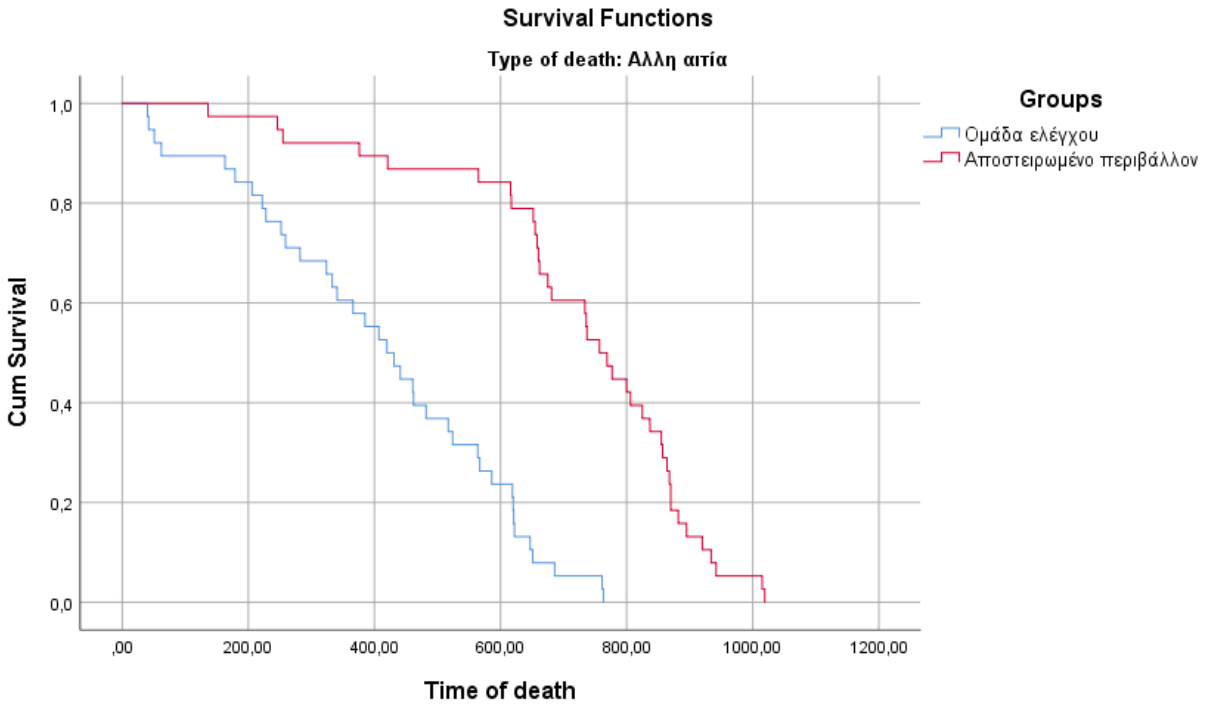
Μια πρώτη σύγκριση των συναρτήσεων επιβίωσης γίνεται στα σχήματα που ακολουθούν.



Σχήμα 4.1



Σχήμα 4.2



Σχήμα 4.3

Οι παρακάτω μη παραμετρικοί έλεγχοι υποθέσεων δηλώνουν ότι οι συναρτήσεις κινδύνου δεν διαφέρουν στα δυο περιβάλλοντα για την αιτία θανάτου λέμφωμα ενώ διαφοροποιούνται για τις δυο άλλες αιτίες θανάτου.

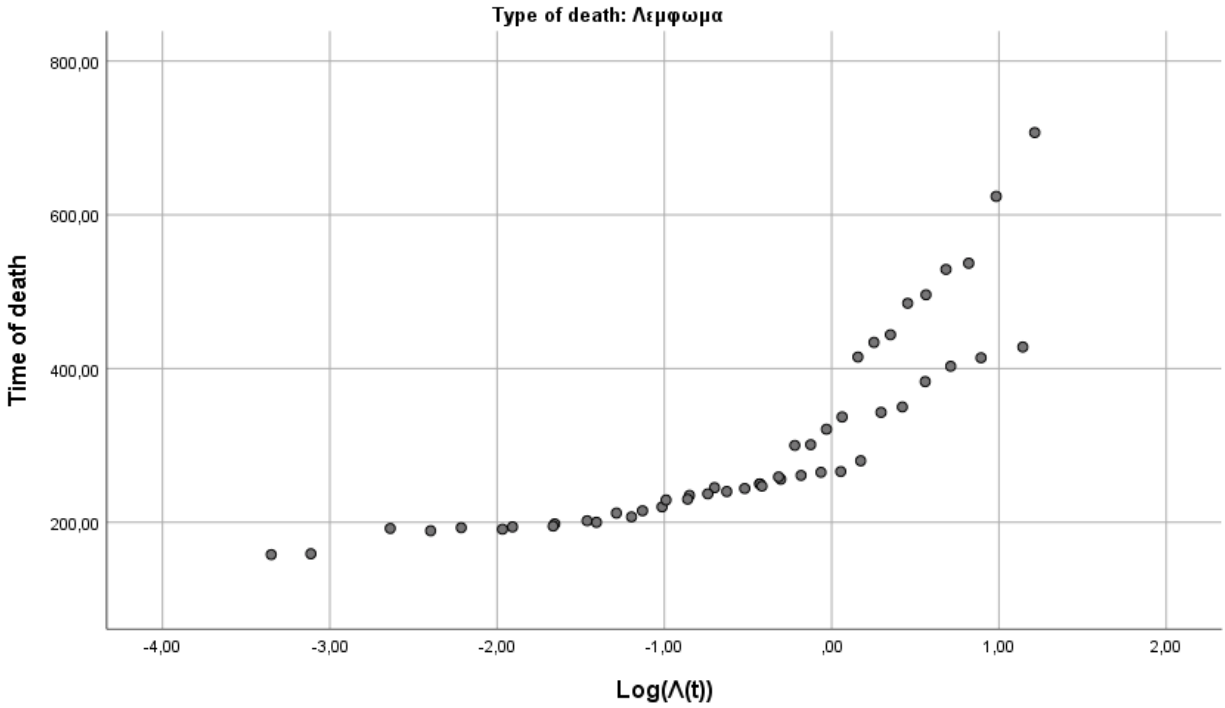
Overall Comparisons

Type of death		Chi-Square	df	Sig.
Λεμφωμα	Log Rank (Mantel-Cox)	3,399	1	,065
	Breslow (Generalized Wilcoxon)	,739	1	,390
	Tarone-Ware	1,635	1	,201
Σαρκωμα	Log Rank (Mantel-Cox)	9,294	1	,002
	Breslow (Generalized Wilcoxon)	6,635	1	,010
	Tarone-Ware	7,857	1	,005
Αλλη αιτία	Log Rank (Mantel-Cox)	43,515	1	,000
	Breslow (Generalized Wilcoxon)	34,805	1	,000
	Tarone-Ware	39,327	1	,000

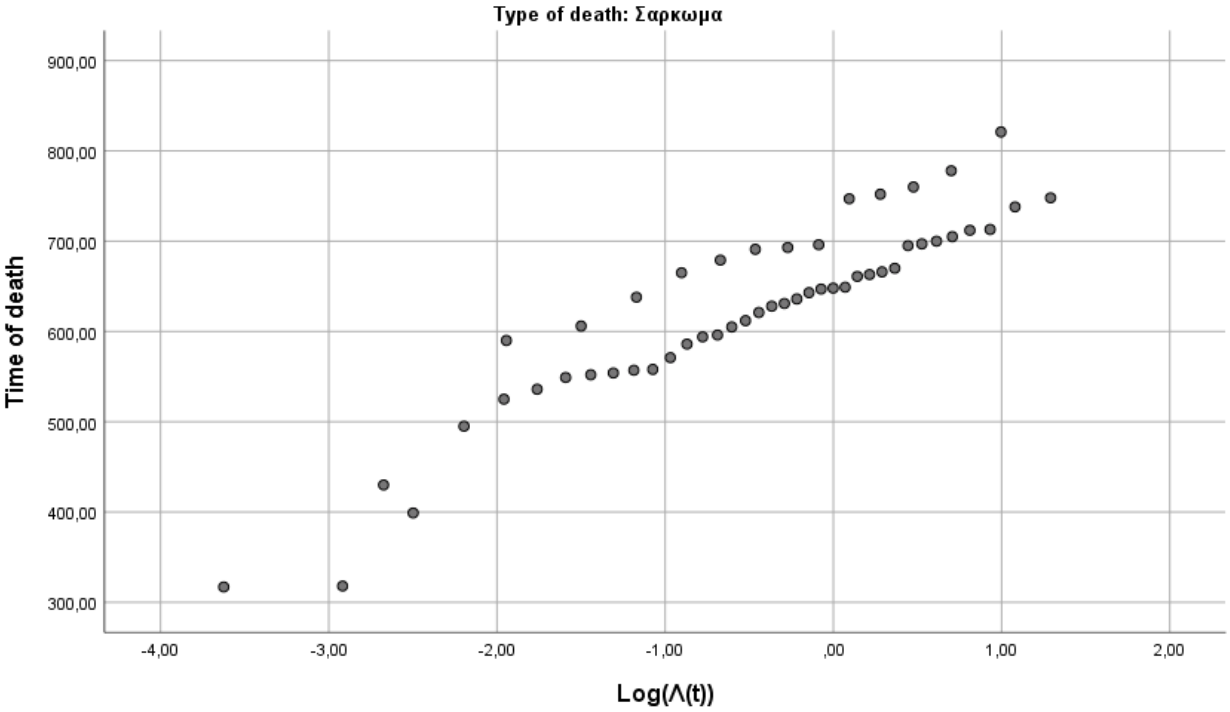
The vector of trend weights is -1, 1. This is the default.

Πίνακας 4.1

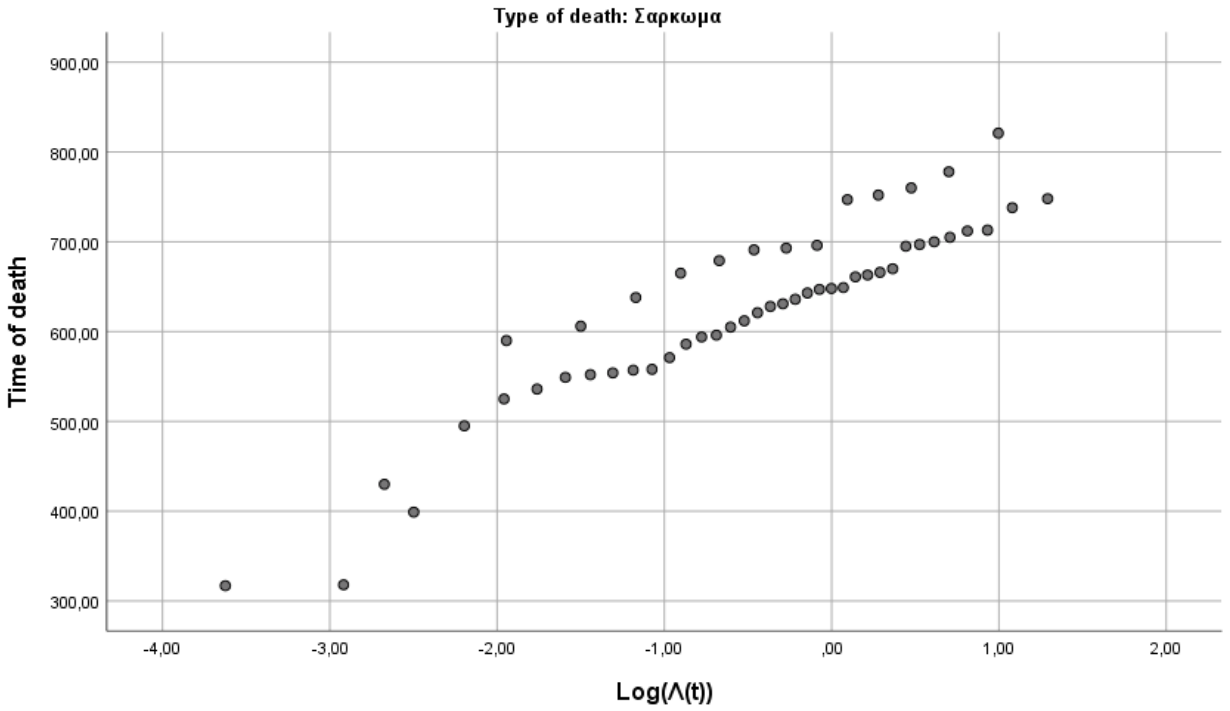
Θα μελετήσουμε τις δυο τελευταίες αιτίες θανάτου υιοθετώντας ένα παραμετρικό μοντέλο.
Κάνοντας το διάγραμμα διασποράς των $(\log \hat{\Lambda}_i\ , \log t)$ $i = 2,3$ παρατηρούμε μια σχετική γραμμική σχέση η οποία δεν υπάρχει στο $(\log \hat{\Lambda}_1\ , \log t)$.



Σχήμα 4.4



Σχήμα 4.5



Σχήμα 4.6

Αυτό μας είναι ισχυρή ένδειξη ότι ότι το παραμετρικό μοντέλο Weibull είναι ένα κατάλληλο μοντέλο. Η (4.2) με τη βοήθεια της (3.13) για την δεύτερη αιτία θανάτου γίνεται

$$L_2(a_2, \gamma_2) = \prod_{\iota=1}^{n} \frac{\gamma_2}{a_2}\left(\frac{t_i}{a_2}\right)^{\gamma_2 - 1} e^{-\left(\frac{t_i}{a_2}\right)^{\gamma_2}}$$

Λογαριθμώντας και παραγωγίζοντας βρίσκουμε το σύστημα

$$\frac{\partial L_2(a_2,\gamma_2)}{\partial \gamma_2} = \frac{n}{\gamma_2} + \sum_{i=1}^{n} \log\frac{t_i}{a_2} - \sum_{i=1}^{n} \left(\frac{t_i}{a_2}\right)^{\gamma_2} \log\frac{t_i}{a_2}$$

$$\frac{\partial L_2(a_2,\gamma_2)}{\partial \alpha_2} = -\frac{n\alpha_2}{\gamma_2} + \frac{n\alpha_2}{\gamma_2} \sum_{i=1}^{n} \left(\frac{t_i}{a_2}\right)^{\gamma_2}.$$

Με τη βοήθεια του Mathematica για την δεύτερη αιτία θανάτου η μέθοδος Newton-Rapson δίνει για την ομάδα ελέγχου $\hat{a}_2 = 678{,}57$ και $\hat{\gamma}_2 = 0{,}806$ και για το αποστειρωμένο περιβάλλον και $\hat{a}_2^0 = 1012{,}32$, $\hat{\gamma}_2^0 = 5{,}181$.

Ο έλεγχος λόγου πιθανοφάνειας απορρίπτει την υπόθεση $\alpha_2 = \alpha_2^0$ και $\gamma_2 = \gamma_2^0$ σε επίπεδο σημαντικότητας 0.05 αφού η τιμή του λόγου πιθανοφάνειας είναι 61 με δυο βαθμούς ελευθερίας.

Για την τρίτη αιτία θανάτου τα αποτελέσματα είναι ανάλογα αφού οι εκτιμήσεις είναι για την ομάδα ελέγχου $\hat{a}_3 = 518{,}07$ και $\hat{\gamma}_2 = 1{,}116$ και για το αποστειρωμένο περιβάλλον και $\hat{a}_2^0 = 881{,}22$, $\hat{\gamma}_2^0 = 4{,}143$ και η υπόθεση της ισότητας των παραμέτρων απορρίπτεται.

4.2 Ομαδοποιημένα ή Διακριτά Δεδομένα

Τα δεδομένα με πολλαπλούς τρόπους αποτυχίας, σε πολλές επιστήμες όπως η Δημογραφία ή οι Ιατρικές επιστήμες, είναι ομαδοποιημένα σε μορφή πίνακα κατά τομείς. Με απουσία συμεταβλητών και λογοκριμμένων δεδομένων εκτός του τελευταίου μεσοδιαστήματος, η συνάρτηση πιθανοφάνειας δίνεται από την (2.15),

$$L = \left\{ \prod_{\ell=1}^{m} \prod_{j=1}^{k} \pi_{j\ell}^{d_{j\ell}} \right\} S(a_m)^{d_{m+1}}, \tag{4.7}$$

όπου οι χρόνοι ζωής ομοδοποιούνται σε μεσοδιαστήματα $I_\ell = [a_{\ell-1}, a_\ell)$, $\ell = 1, \ldots, m+1$ με $a_0 = 0$ και $a_{m+1} = \infty$. Η παράμετρος $\pi_{j\ell}$ είναι η πιθανότητα ένα άτομο να αποτύχει στο $[a_{\ell-1}, a_\ell)$ από τον τρόπο j και $d_{j\ell}$ ο αριθμός των ατόμων που απέτυχαν στο $[a_{\ell-1}, a_\ell)$ από τον j. Οι τρόποι αποτυχίας για τα d_{m+1} άτομα που είναι ακόμη εν ζωή στο χρόνο a_m είναι άγνωστοι. Σημειώστε ότι $S(a_m) = 1 - \sum_{j,\ell} \pi_{j\ell}$.

Ένας εναλλακτικός τρόπος να γραφεί η (4.7) είναι με τους όρους των δεσμευμένων πιθανοτήτων

$$q_{j\ell} = \Pr(\eta\ \mu o \nu \acute{a} \delta a\ \ j\ \nu a\ a\pi o \tau \acute{\upsilon} \chi \varepsilon \iota\ \sigma \tau o \ \delta \iota \acute{a} \sigma \tau \eta \mu a\ I_\ell | \ \varepsilon \nu\ \zeta \omega \acute{\eta}\ \ \sigma \tau o\ a_{\ell-1}) = \pi_{j\ell}/S(a_{\ell-1}). \tag{4.8}$$

Με τη βοήθεια της (4.8), η (4.7) μπορεί να εκφραστεί ως

$$L = \prod_{\ell=1}^{m} \prod_{j=1}^{k} q_{j\ell}^{d_{j\ell}} (1 - q_{\cdot\ell})^{n_\ell - d_{\cdot\ell}}, \tag{4.9}$$

Όπου

$q_{\cdot\ell} = \sum_j q_{j\ell}$, $d_{\cdot\ell} = \sum_j d_{j\ell}$

και

$n_\ell = n - (d_1 + \cdots + d_{\ell-1})$ ο αριθμός των ατόμων που είναι ζωντανή στο $a_{\ell-1}$.

Με μια τροποποίηση επί των n_ℓ η παραπάνω συνάρτηση πιθανοφάνειας ισχύει επίσης και όταν υπάρχει λογοκρισία στο δεξί άκρο των διαστημάτων. Πιο συγκεκριμένα σε αυτή την περίπτωση, το n_ℓ είναι ο αριθμός των ατόμων εν ζωή χωρίς να καταγράφουμε τη λογοκρισία στο $a_{\ell-1}$ και επομένως σε ρίσκο αποτυχίας στο $[a_{\ell-1}, a_\ell)$. Η συνάρτηση πιθανοφάνειας (4.9) είναι σε αναλογία με τη συνάρτηση πιθανοφάνειας που χρησιμοποιείται για πίνακες δεδομένων ζωής, όταν υπάρχει μόνο ένας τρόπος αποτυχίας. (Miller 1981, Τζαβελάς 2018) . Η περίπτωση της λογοκρισίας μέσα στα μεσοδιαστήματα είναι δύσκολα διαχειρίσιμο και απαιτεί περαιτέρω υποθέσεις.

Για ένα παραμετρικό μοντέλο συνεχούς χρόνου, το π_{jl} ή το q_{jl} στην (4.7) ή στην (4.8) είναι όπως δίνονται από την (2.14). Η συνάρτηση πιθανοφάνειας (4.7) είναι πολυωνυμικής μορφής (multinomial form) και η (4.9) είναι γινόμενο πολυωνυμικής μορφής και μπορεί να μεγιστοποιηθεί με τυπικό λογισμικό βελτιστοποίησης (Newton-Rapshon, Fisher). Στην περίπτωση που δεν υιοθετείται κάποιο παραμετρικό μοντέλο, υπάρχουν $k \cdot m$ παράμετροι $q_{j\ell}$ όπου πληρούν τις προϋποθέσεις

$0 \leq q_{j\ell} \leq 1$

και

$0 \leq q_{1l} + + q_{k\ell} \leq 1$ για κάθε ℓ .

Εύκολα παρατηρείται ότι η (4.8) είναι μεγιστοποιημένη στο χώρο αυτής της παραμέτρου για

$$\hat{q}_{j\ell} = \frac{d_{j\ell}}{n_\ell} \quad j = 1, \dots, k; \quad \ell = 1, \dots, m. \tag{4.10}$$

Έτσι ο ΕΜΠ του S (a_ℓ) είναι

$$\hat{S}(a_\ell) = \prod_{u=1}^{\ell-1}(1 - \hat{q}_{.u}),$$

και οι εκτιμήσεις άλλων ποσοτήτων είναι επίσης εύκολα υπολογίσιμες. Για την κατασκευή διαστημάτων εμπιστοσύνης απαιτείται η γνώση των διασπορών των εκτιμητών η οποία εν γένει δεν είναι εύκολο να βρεθεί. Με τη βοήθεια της ασυμπτωτικής διασποράς μπορούν να κατασκευασθούν ασυμπτωτικά διαστήματα εμπιστοσύνης. Αντιστρέφοντας τον παρατηρούμενο πίνακα πληροφορίας με βάση την (4.9) έχουμε ότι οι ασυμπτωτικές εκτιμήσεις διακύμανσης του $\hat{q}_{j_l}$ είναι

$$\widehat{Var}(\hat{q}_{j\ell}) = \hat{q}_{j\ell}(1 - \hat{q}_{j\ell})/n_\ell \qquad j = 1, \dots, k; \quad \ell = 1, \dots, m$$

$$\widehat{Cov}(\hat{q}_{j\ell}, \hat{q}_{r\ell}) = -\frac{\hat{q}_{j\ell}\hat{q}_{r\ell}}{n_\ell} \quad j \neq r; \quad \ell = 1,, \dots, m$$

$$\widehat{Cov}(\hat{q}_{j\ell}, \hat{q}_{ru}) = 0 \qquad \ell \neq u. \tag{4.11}$$

ΚΕΦΑΛΑΙΟ 5

ΗΜΙΠΑΡΑΜΕΤΡΙΚΕΣ ΜΕΘΟΔΟΙ ΕΠΙΛΥΣΗΣ ΓΙΑ ΠΟΛΛΑΠΛΑΣΙΑΣΤΙΚΑ ΜΟΝΤΕΛΑ ΕΠΙΚΙΝΔΥΝΟΤΗΤΑΣ

Στο κεφάλαιο αυτό περιγράφονται οι βασικές μέθοδοι ημιπαραμετρικών μοντέλων όπως αυτές αναπτύχθηκαν από τους Kalbfleisch και Prentice (1980, Εν. 7.2), Benichou και Gail (1990) Cheng et. al. 1998). Πληροφορίες αυτής της μεθοδολογίας για ιατρικά δεδομένα δίδονται από Kay (1986), Gaynor at. al. 1993), Lunn και McNeil (1995).

Όπως σημειώθηκε στο Κεφάλαιο 4, η μοντελοποίηση παλινδρόμησης για ανταγωνιστικούς τρόπους αποτυχίας είναι εύκολη για κατάλληλη επιλογή των $\lambda_j(t|\boldsymbol{x})$ όπου $\boldsymbol{x}$ ένα διάνυσμα συμμεταβλητών. Για πολλαπλασιαστικά μοντέλα μπορούν να χρησιμοποιηθούν ημιπαραμετρικές μέθοδοι ανάλογες με αυτές που χρησιμοποιούνται στο μοντέλο του Cox. Ποιο συγκεκριμένα τα μοντέλα της μορφής

$$\lambda_j(t|x) = \lambda_{0j}(t)e^{\boldsymbol{\beta}_j'\boldsymbol{x}} \quad j = 1, \ldots., k \tag{5.1}$$

μπορούν να μελετηθούν μελετώντας κάθε τρόπο αποτυχίας χωριστά με τη μέθοδο της μερικής πιθανοφάνειας (partial likelihood).

Πιο συγκεκριμένα μπορούμε να γράψουμε την μερική πιθανοφάνεια για τα $\boldsymbol{\beta_j}$ θεωρώντας την πιθανότητα του ενδεχόμενου ένα συγκεκριμένο άτομο να αποτύχει με τον τρόπο *j* στο χρόνο *t*, δεδομένου ότι ένα άτομο σε κίνδυνο στο χρόνο t αποτυγχάνει από τον τρόπο *j*. Αυτό δίνει μία μερική πιθανοφάνεια

$$L(\boldsymbol{\beta_1}. \ldots, \boldsymbol{\beta_k}) = \prod_{i=1}^{n}\prod_{j=1}^{k}\left(\frac{e^{\boldsymbol{\beta}_j'\boldsymbol{x_i}}}{\sum_{l=1}^{n} Y_l(t_i)e^{\boldsymbol{\beta}_j'\boldsymbol{x_l}}}\right)^{\delta_{ij}} \tag{5.2}$$

όπου $Y_l(t)$ = I $(t_l \geq t)$. Αυτό όπως έχουμε ξαναδεί παραγοντοποιείται σε γινόμενο παραγόντων με όρους $L_j(\boldsymbol{\beta_j})$ για j = 1, …., k, με

$$L_j(\beta_j) = \prod_{i=1}^{n}\left(\frac{e^{\boldsymbol{\beta}_j'\boldsymbol{x_i}}}{\sum_{l=1}^{n} Y_l(t_i)e^{\boldsymbol{\beta}_j'\boldsymbol{x_l}}}\right)^{\delta_{ij}} \tag{5.3}$$

Δηλαδή

$$L(\boldsymbol{\beta_1}. \ldots, \boldsymbol{\beta_k}) = \prod_{\mathrm{j}=1}^{k} L_j(\boldsymbol{\beta_j}).$$

Η τελευταία σχέση μας επιτρέπει να εκτιμήσουμε τα $\boldsymbol{\beta_1}. \ldots, \boldsymbol{\beta_k}$ χωριστά δηλαδή τα $\boldsymbol{\beta_j}$ από τα $L_j(\beta_j)$

Η Ομοίως, η γενικευμένη εκτίμηση Nélson – Aalen,

$$\hat{\Lambda}_{0j}(t) = \sum_{i:t_i \leq t}\left(\frac{\delta_{ij}}{\sum_{l=1}^{n} Y_l(t_i)e^{\boldsymbol{\beta}_j'\boldsymbol{x_l}}}\right) \tag{5.4}$$

Μπορεί να χρησιμοποιηθεί για την εκτίμηση της αθροιστικής συνάρτησης κινδύνου για τον τρόπο αποτυχίας *j*. Παρατηρούμε λοιπόν ότι οι διαδικασίες για εκτιμήσεις παραγόντων, ελέγχους υποθέσεων ή κατασκευή διαστημάτων εμπιστοσύνης ανάγονται στην περίπτωση ενός τρόπου αποτυχίας και αντιμετωπίζονται με τον

γνωστό τρόπο. Για τα υποθετικά τέστ ή την εκτίμηση διαστήματος για τα $\beta_j, \Lambda_{o_j}(t)$ ή $\Lambda_j(t|x)$ η μεθοδολογία επεκτείνεται για να χειριστεί τις χρονικά μεταβαλλόμενες συμμεταβλητές x(t) στη θέση του x στην (9.4.1). Εν τέλει, σημειώστε ότι δεν είναι αναγκαίο να συμπεριλάβουμε τις ίδες συμμεταβλητές στα μοντέλα για διαφερετικούς τρόπους αποτυχίας. Συγκεκριμένα στοιχεία του β_j στην (9.4.1) μπορεί να είναι περιορισμένα να ισούνται με το μηδέν, οπότε οι αντίστοιχοι όροι αποσύρονται από το μοντέλο.

5.1 Εκτίμηση των Συναρτήσεων Αθροιστικής Επίπτωσης

Έχουμε δει ότι για το πολλαπλασιαστικό μοντέλο οι συναρτήσεις κινδύνου περιγράφονται από τη σχέση (5.1). Χρησιμοποιώντας τους εκτιμητές $\hat{\beta}_j'$ μπορούμε να εκτιμήσουμε τη συνάρτηση επιβίωσης $S(t|X)$ ως

$$\hat{S}(t|X) = \exp\left\{-\sum_{j=1}^{k} \hat{\Lambda}_{0j}(t) e^{\hat{\boldsymbol{\beta}}_j' \boldsymbol{x}}\right\}, \tag{5.5}$$

Οι εκτιμήσεις διακύμανσης για τις ποσότητες

$$\hat{\Lambda}_j(t|x) = \hat{\Lambda}_{0j}(t) e^{\hat{\boldsymbol{\beta}}_j' \boldsymbol{x}} \tag{5.6}$$

Μπορούν να βρεθούν εφαρμόζοντας τη μέθοδο δέλτα στην διασπορά των $\hat{\beta}_j'$. Η κατασκευή ασυμπτωτικών διαστημάτων εμπιστοσύνης για την $S(t|X)$ γίνεται μέσω του ασυμπτωτικού αποτελέσματος

$$\widehat{Var}[\log[-\log \hat{S}(t|X)]] = \hat{H}_0(t)^{-2} \left\{ \sum_{i:t_i \le t} \frac{\delta_i}{S^{(0)}\left(t_i, \hat{\boldsymbol{\beta}}\right)^2} \bar{\boldsymbol{x}}\left(t_i, \hat{\boldsymbol{\beta}}\right) + \left[\widehat{\boldsymbol{W}}(t) - \hat{H}_0(t)\boldsymbol{x}\right]' I\left(\hat{\boldsymbol{\beta}}\right)\left[\widehat{\boldsymbol{W}}(t) - \hat{H}_0(t)\boldsymbol{x}\right] \right\}$$

Όπου

$$\hat{H}_0(t) = \sum_{i:t_i \le t} \frac{\delta_i}{\sum_{\ell=1}^{n} Y_\ell(t_i) e^{\hat{\boldsymbol{\beta}}_j' \boldsymbol{x}}}$$

(Beslow ή γενικευμένος εκτιμητής Nelson-Aalen),

$$S^{(0)}(t) = \sum\nolimits_{i=1}^{n} Y_i(t) e^{\hat{\boldsymbol{\beta}}_j' \boldsymbol{x}}$$

και

$$W(t) = \sum_{i:t_i \le t} \frac{\delta_i}{S^{(0)}\left(t_i, \hat{\boldsymbol{\beta}}\right)} \bar{\boldsymbol{x}}\left(t_i, \hat{\boldsymbol{\beta}}\right).$$

Δηλαδή αν $[\alpha, \beta]$ είναι το ασυμπτωτικό διάστημα εμπιστοσύνης για το $\widehat{Var}[\log[-\log \hat{S}(t|X)]]$ το α.δ.ε. για το $\hat{S}(t|X)$ είναι το

$$[\exp[-\exp(b)], \exp[-\exp(a)].$$

Η συνάρτηση αθροιστικής επίπτωσης $F_j(t|\boldsymbol{x})$ μπορεί να υπολογιστεί ως εξής

$$\hat{F}_j(t|\boldsymbol{x}) = \int_0^t \hat{S}(u|\boldsymbol{x}) d\hat{\Lambda}_j(u|\boldsymbol{x})$$

$$= \int_0^t \exp\left\{-\sum_{\ell=1}^{k} \hat{\Lambda}_{0\ell}(u) e^{\hat{\boldsymbol{\beta}}_\ell' \boldsymbol{x}}\right\} e^{\hat{\boldsymbol{\beta}}_j' \boldsymbol{x}} d\hat{\Lambda}_{0j}(u)$$

$$= \sum_{i:t_i \le t} \delta_{ij} \exp\left\{-\sum_{\ell=1}^{k} \hat{\Lambda}_{0\ell}(t_i) e^{\hat{\boldsymbol{\beta}}_\ell' \boldsymbol{x}}\right\} \frac{e^{\hat{\boldsymbol{\beta}}_j' \boldsymbol{x}}}{\sum_{\ell=1}^{n} Y_\ell(t_i) e^{\boldsymbol{\beta}_j' \boldsymbol{x}_\ell}} . \tag{5.7}$$

Anderson et. al.(1993, 7.2.3) μέσα σε ένα γενικότερο πλαίσιο δίνει μια διαφορετική προσέγγιση της εκτίμησης της $S(t|\boldsymbol{x})$ στην περίπτωση συμμεταβλητών αλλά η (5.5) είναι πιο απλή και γι' αυτό συνήθως υιοθετείται στην πράξη.

Η εκτίμηση διακύμανσης για το $\hat{F}_j(t)$ στην περίπτωση μη-συμμεταβλητών δίνεται στο Κεφάλαιο 2. Η εκτίμηση της $\hat{F}_j(t)$ στην περίπτωση συμμεταβλητών είναι ιδιαίτερα πολύπλοκη (Benichou και Gail 1990). Η προσέγγιση που θα υιοθετήσουμε μοιάζει με αυτή του Κεφαλαίου 2. Θα προσεγγίσουμε το μοντέλο μας με ένα απλούστερο στο οποίο η συνάρτηση κινδύνου είναι κλιμακωτή (κατά τμήματα σταθερή). Οι συναρτήσεις κινδύνου $\lambda_{0j}(t)$ είναι της μορφής της (4.5), η οποία σε σχέση με τις παραμέτρους λ_{jl}

$$\lambda_{0j}(t) = \lambda_{j\ell} \qquad t \in I_{\ell,} \tag{5.8}$$

όπου $I_\ell = [a_{\ell-1}, a_\ell)$, $\ell = 1, \ldots, m$, με όρια $0 = a_0 < a_\ell < \cdots < a_m$. Οι αντίστοιχες αθροιστικές συναρτήσεις κινδύνου είναι

$$\Lambda_{0j}(t) = \sum_{l=1}^{m} \lambda_{jl} \Delta_l(t), \tag{5.9}$$

όπου $\Delta_\ell(t)$ είναι το μήκος της τομής των I_ℓ και [0, t).

Με αυτόν τον τρόπο η συνάρτηση πιθανοφάνειας (2.11) ακόμα και στην περίπτωση παρουσίας συμμεταβλητών παραγοντοποιείται σε ξεχωριστά κομμάτια για κάθε τρόπο αποτυχίας. συμπερασματικές διαδικασίες για τα $\lambda_j = (\lambda_{jt}, \ldots, \lambda_{jm})$ και β_j υπό το μοντέλο αναλογικών κινδύνων (9.4.1), ακολουθούν ακριβώς το Παράδειγμα 6.5.1. Συγκεκριμένα, η m.l. εκτίμηση για το β_j μπορεί να βρεθεί με το να μεγιστοποιηθεί την profil λογαριθμική συνάρτηση πιθανοφάνεια

$$l_{pj}(\boldsymbol{\beta_j}) = \sum_{i=1}^{n} \delta_{ij} \left\{ \boldsymbol{\beta_j' x_i} - \log \left[\sum_{l=1}^{m} I(t_i \in I_l) \sum_{r=1}^{n} \Delta_l(t_r) e^{\boldsymbol{\beta_j' x_r}} \right] \right\}.$$

και το λ_{jt} εκτιμώνται από την

$$\hat{\lambda}_{j\ell} = \frac{d_{j\ell}}{\sum_{i=1}^{n} \Delta_\ell(t_i) e^{\boldsymbol{\beta_j' x_i}}}, \qquad \ell = 1, \ldots, m, \tag{5.10}$$

όπου $d_{j\ell}$ ο αριθμός των ατόμων που απέτυχαν στο μεσοδιάστημα I_ℓ με τρόπο j. Οι εκτιμητές μέγιστης πιθανοφάνειας των $\boldsymbol{\beta_j}$ και $\Lambda_{0j}(t)$ υπό αυτό το μοντέλο, προσεγγίζουν τις εκτιμήσεις των $\boldsymbol{\beta_j}$ που υπολογίστηκαν από την Cox μερική συνάρτηση πιθανότητας (5.3) και τις γενικευμένες εκτιμήσεις Nelson – Aalen (5.4), όπου το m αυξάνεται και τα μήκη των μεσοδιαστημάτων $|a_l - a_{l-1}|$ προσεγγίζουν το μηδέν.

Η εκτίμηση διακύμανσης του εκτιμητή $\hat{F}_j(t|x)$ στο μοντέλο που έχουμε υιοθετήσει με την κατά διαστήματα σταθερή συνάρτηση κινδύνου ακολουθεί τα εξής βήματα. Από τον πίνακα των δευτέρων παραγώγων της λογαριθμικής συνάρτησης πιθανοφάνειας βρίσκουμε ότι ο πληροφοριακός πίνακας για τις παραμέτρους $(\boldsymbol{\lambda_j}, \boldsymbol{\beta_j})$ διάστασης $m + p$ μπορεί να γραφεί σε block μορφή

$$\boldsymbol{I_j}(\boldsymbol{\lambda_j}, \boldsymbol{\beta_j}) = \begin{pmatrix} \boldsymbol{D_j} & \boldsymbol{C_j} \\ \boldsymbol{C_j} & \boldsymbol{B_j} \end{pmatrix}, \tag{5.11}$$

όπου

$$\boldsymbol{D_j} = Diag(d_{jl}/\lambda_{jl}^2)$$

είναι ένας διαγώνιος πίνακας διαστάσεων $m \times m$, $\boldsymbol{C_j}$ ένας πίνακας διαστάσεων $m \times p$, με το ℓq στοιχείο

$$(C_j)_{\ell q} = \sum_{i=1}^{n} \Delta_\ell(t_i) x_{iq} e^{\boldsymbol{\beta_j' x_i}}$$

και $\boldsymbol{B_j}$ ο p x p πίνακας,

$$\boldsymbol{B_j} = \sum_{i=1}^{n} \boldsymbol{x_i x_l'} \Lambda_{0j}(t_i) e^{\boldsymbol{\beta_j' x_i}}$$

Η εκτίμηση m.l. $\hat{F}_j(t|\boldsymbol{x})$ περιλαμβάνει ($\hat{\lambda}_r$, $\hat{\beta}_r$) για όλα τα r = 1, … , k. Από την ασυμπτωματική ανεξαρτησία των ($\hat{\lambda}_r$, $\hat{\beta}_r$) και με εφαρμογή του Θεωρήματος Β2 (Παράρτημα Β), συνεπάγεται ότι

$$\widehat{Var}\left[\hat{F}_j(t|x)\right] = \sum_{r=1}^{k} \widehat{w}_r^{(j)\prime} I_r\left(\hat{\lambda}_r, \hat{\beta}_r\right)^{-1} \widehat{w}_r^{(j)}, \tag{5.12}$$

όπου $\boldsymbol{w_r^{(j)}}$ ένα διάνυσμα (m = p) x 1,

$$\boldsymbol{w_r^{(j)}} = \left(\frac{\partial F_j(t|x)}{\partial \lambda_{rl}} . \dots , \frac{\partial F_j(t|x)}{\partial \lambda_{rm}}, \frac{\partial F_j(t|x)}{\partial \beta_{rl}}, \dots , \frac{\partial F_j(t|x)}{\partial \beta_{rp}} \right)'$$

Τα στοιχεία αυτού του διανύσματος βρίσκονται παραγωγίζοντας κάτω από το ολοκλήρωμα την

$$F_j(t|X) = \int_0^t S(u|\boldsymbol{X}) \lambda_{0j}(u) e^{\boldsymbol{\beta_j' x}} du,$$

που δίνει

$$\frac{\partial F_j(t|\boldsymbol{x})}{\partial \lambda_{rl}} = e^{\beta_j' x} \int_0^t S(u|\boldsymbol{x}) [I(r = j) I(u \in I_l) - e^{\beta_r' x} \Delta_\lambda(u) \lambda_{0j}(u)] du \tag{5.13}$$

και

$$\frac{\partial F_j(t|X)}{\partial \beta_r} = \boldsymbol{x} e^{\boldsymbol{\beta}_j' \boldsymbol{x}} \int_0^t S(u|\boldsymbol{x}) \left[I(r = j) - \Lambda_{0r}(u) e^{\boldsymbol{\beta}_r' \boldsymbol{x}} \right] \lambda_{0j}(u) du \tag{5.14}$$

Η διαδικασία λοιπόν είναι να προσεγγίσουμε την $\hat{F}_j(t|\boldsymbol{x})$ με βάση τις ημιπαραμετρικές εκτιμήσεις χρησιμοποιώντας την (5.12) για μεγάλο m και το a_ℓ να επιλεγεί ώστε κάθε διάστημα $[a_{\ell-1} - a_\ell)$ να έχει τουλάχιστον μερικές αποτυχίες. Αντί να υπολογίσουμε τις εκτιμήσεις των β_j και $\Lambda_{Oj}(t)$ με το αποσπασματικά συνεχές μοντέλο, μπορούμε αντ' αυτού να χρησιμοποιήσουμε τις ημιπαραμετρικές εκτιμήσεις $\hat{\beta}_j$ και $\hat{\Lambda}_{Oj}(t)$ και να προσεγγίσουμε τις εισόδους στις (9.4.11), (9.4.13) και (9.4.14) με το να αντικαταστήσουμε το λ_{j_l} με

$$\hat{\lambda}_{jl} = \frac{\hat{\Lambda}_{0j}(a_l) - \hat{\Lambda}_{0j}(a_l - 1)}{a_l - a_{l-1}},$$

$\lambda_{0j}(u)du$ με $\mathrm{d}\hat{\Lambda}_{Oj}(u)$ και άλλες ποσότητες με τις ημιπαραμετρικές εκτιμήσεις τους.

ΚΕΦΑΛΑΙΟ 6

ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

Το σύνολο δεδομένων αποτελείται από 205 ασθενείς και για κάθε έναν υπάρχουν 7 μεταβλητές. Οι μεταβλητές αυτές αφορούν μετρήσεις που έγιναν σε ασθενείς με κακοήθες μελάνωμα. Σε κάθε ασθενή αφαιρέθηκε ο όγκος με χειρουργική επέμβαση. Η χειρουργική επέμβαση περιλαμβάνει την πλήρη απομάκρυνση του όγκου μαζί με περίπου 2.5 εκατοστά του περιβάλλοντος δέρματος. Μεταξύ των μετρήσεων που ελήφθησαν ήταν το πάχος του όγκου και εάν υπήρχε δερματικό έλκος ή όχι (απώλεια δέρματος). Αυτές θεωρούνται σημαντικές μεταβλητές για πρόγνωση διότι ασθενείς με παχύ ή/και ελκόμενο όγκο έχουν μεγαλύτερη πιθανότητα θανάτου από μελάνωμα.
Πηγή dataset: https://vincentarelbundock.github.io/Rdatasets/articles/data.html

Μεταβλητές:

-time: Χρόνος επιβίωσης σε ημέρες από την επέμβαση του ασθενούς.

- **status**: Η κατάσταση των ασθενών στο τέλος της μελέτης. Το 1 δείχνει ότι ο

ασθενής πέθανε από μελάνωμα, το 2 δείχνει ότι ήταν ακόμα ζωντανός μέχρι το τέλος

της μελέτηςκαι το 3 ότι πέθανε από αιτία που δεν σχετίζεται με το μελάνωμα.

- **sex**: Το φύλο του ασθενούς, όπου το 1 αντιστοιχεί στους άντρες και το 0 στις γυναίκες.

- **age**: Ηλικία του ασθενούς σε έτη κατά τη στιγμή της επέμβασης.

- **year**: ΄Ετος επέμβασης.

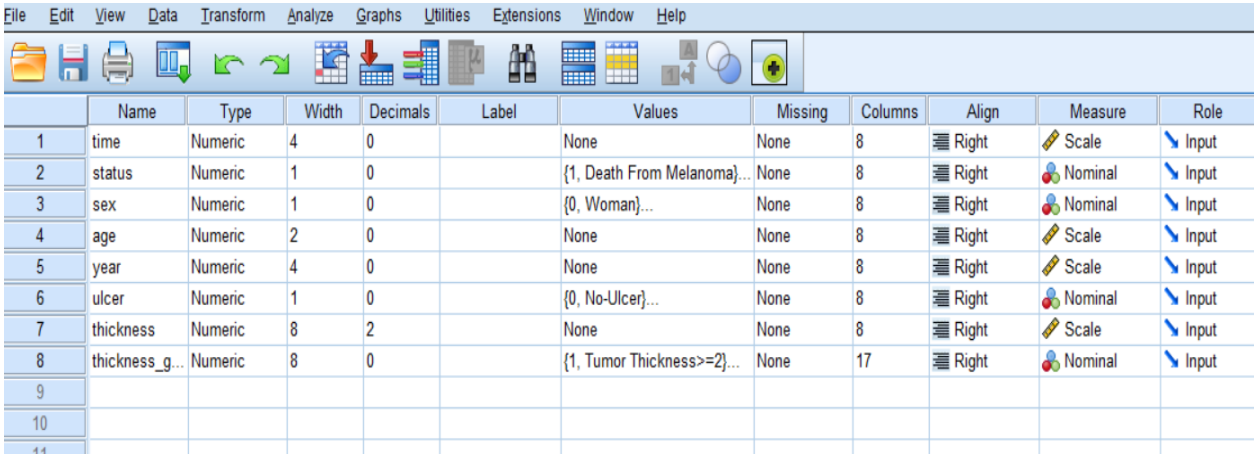
-thickness: Πάχος όγκου σε mm.

- **ulcer**: Δείκτης έλκους, όπου το 1 συμβολίζει την ύπαρξη έλκους, ενώ το 0 την

απουσία έλκους.

Περνάμε το dataset στο SPSS όπως φαίνεται παρακάτω στο Σχήμα 1. Αρχικά πάμε στην καρτέλα Variable View και στην επιλογή Values δηλώνουμε για τις μεταβλητές "status", "sex" και "ulcer" τι συμβολίζει το 0,1,2 αντίστοιχα. Επιπλέον μέσω της επιλογής Transform-Recode into Different Variables θα φτιάξουμε μια ακόμη μεταβλητή που θα ομαδοποιεί ουσιαστικά τη μεταβλητή Thickness. Θα φτιάξουμε τις εξής κατηγορίες:

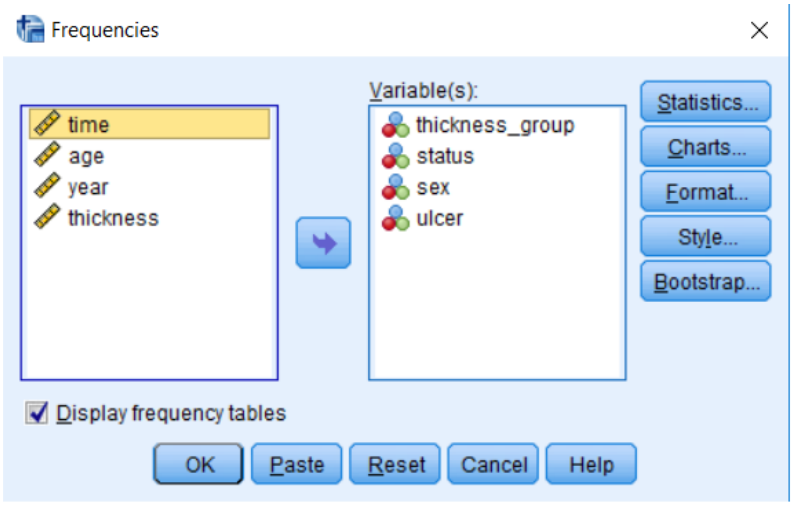
		Tumor Thickness	≤	2
2	<	Tumor Thickness	≤	5
5	<	Tumor Thickness		



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	time	Numeric	4	0		None	None	8	Right	Scale	Input
2	status	Numeric	1	0		{1, Death From Melanoma}...	None	8	Right	Nominal	Input
3	sex	Numeric	1	0		{0, Woman}...	None	8	Right	Nominal	Input
4	age	Numeric	2	0		None	None	8	Right	Scale	Input
5	year	Numeric	4	0		None	None	8	Right	Scale	Input
6	ulcer	Numeric	1	0		{0, No-Ulcer}...	None	8	Right	Nominal	Input
7	thickness	Numeric	8	2		None	None	8	Right	Scale	Input
8	thickness_g...	Numeric	8	0		{1, Tumor Thickness>=2}...	None	17	Right	Nominal	Input
9											
10											

Σχήμα 6.1

Αρχικά θα ξεκινήσουμε με μια περιγραφική ανάλυση των κατηγορικών μας μεταβλητών. Στο SPSS τα βήματα που θα ακολουθήσουμε είναι τα εξής: Analyze-Descriptive Statistics-Frequencies



Σχήμα 6.2

thickness_group

		Frequency	Percent
Valid	Tumor Thickness<=200	109	53,2
	200>Tumor Thickness<=500	64	31,2
	Tumor Thickness>500	32	15,6
	Total	205	100,0

status

		Frequency	Percent
Valid	Death From Melanoma	57	27,8
	Alive	134	65,4
	Death From Another Cause	14	6,8
	Total	205	100,0

sex

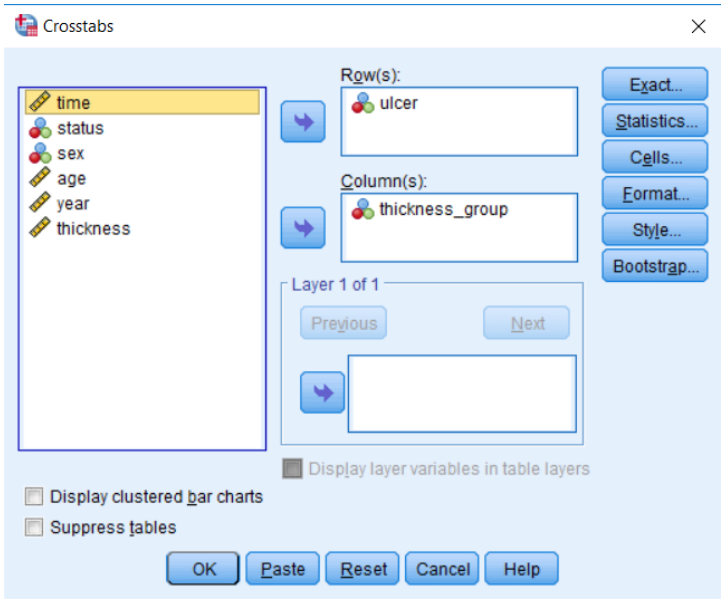
		Frequency	Percent
Valid	Woman	126	61,5
	Man	79	38,5
	Total	205	100,0

ulcer

		Frequency	Percent
Valid	No-Ulcer	115	56,1
	Ulcer	90	43,9
	Total	205	100,0

Πίνακας 6.1

Από τον Πίνακα 6.1 παρατηρούμε ότι από το σύνολο των περιπτώσεων οι 71 πέθαναν πριν το τέλος της ανάλυσης. όμως από αυτές οι 57 ήταν λόγω του όγκου (μελάνωμα), ενώ οι υπόλοιπες 14 περιπτώσεις πέθαναν από άλλη αιτία. Παρ΄ όλα αυτά, το μεγαλύτερο
ποσοστό των ασθενών -134 άτομα- παρέμεινε ζωντανό μέχρι το τέλος της έρευνας (status). Επιπρόσθετα, παρατηρούμε
ότι το μεγαλύτερο ποσοστό των ασθενών του δείγματος είναι γυναίκες -126 άτομα- (sex).
Απο τη μεταβλητή thickness_group παρατηρούμε ότι το μεγαλύτερο ποσοστό έχει μέγεθος όγκου μικρότερο από 200 mm. Τέλος, παρουσιάζεται επίσης και ο αριθμός των ασθενών οι οποίοι έχουν έλκος και ο αριθμός των ασθενών που δεν έχουν έλκος. Από τους 205 ασθενείς του δείγματος οι 90 παρουσίασαν έλκος, ενώ οι 115 δεν παρουσίασαν. Ακόμη μπορούμε να βρούμε ανά κατηγορία μεγέθους όγκου τον αριθμό ασθενών που παρουσίασαν ή όχι έλκος. Αυτό μπορούμε να το κάνουμε στο SPSS από την επιλογή Analyze-Descriptive Statistics-Crosstab



Σχήμα 6.3

ulcer * thickness_group Crosstabulation

Count

		thickness_group			
		Tumor Thickness<= 200	200>Tumor Thickness<= 500	Tumor Thickness>500	Total
ulcer	No-Ulcer	87	21	7	115
	Ulcer	22	43	25	90
Total		109	64	32	205

Πίνακας 6.2

Παρατηρείται ότι, στην περίπτωση όπου ο ασθενής έχει μέγεθος όγκου μεγαλύτερο από 500 mm υπάρχει και μεγαλύτερη πιθανότητα εμφάνισης έλκους, ενώ στην περίπτωση όπου το μέγεθος του όγκου για τον ασθενή είναι μικρότερο από 200 mm ο ασθενής έχει μεγαλύτερη πιθανότητα να μην εμφανίσει έλκος. Στη συνέχεια κάνοντας έναν χ^2 έλεγχο ανεξαρτησίας προκύπτει μία $p - value = 0.000$, επομένως απορρίπτουμε τη μηδενική υπόθεση ότι αυτές οι δύο μεταβλητές είναι ανεξάρτητες, δηλαδή το πάχος του όγκου θα επηρεάσει την παρουσία ή απουσία έλκους (Πίνακας 6.3).

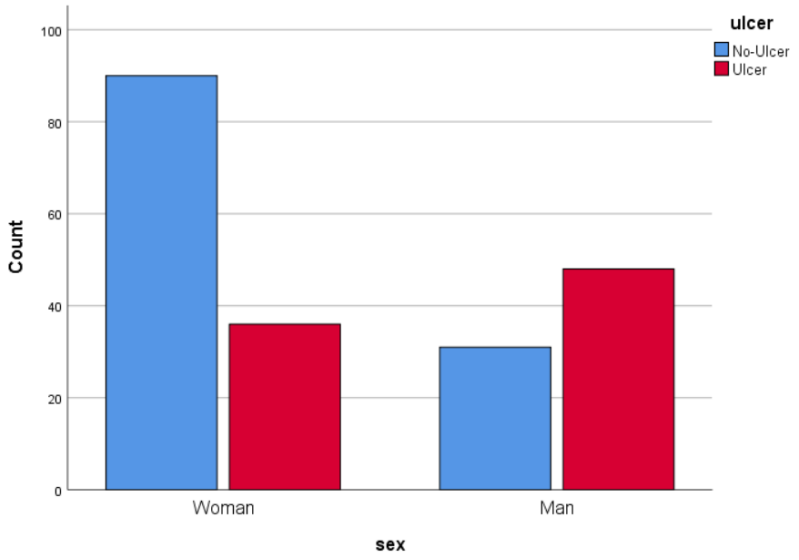
Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	54,206[a]	2	,000
Likelihood Ratio	56,870	2	,000
Linear-by-Linear Association	48,835	1	,000
N of Valid Cases	205		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 14,05.

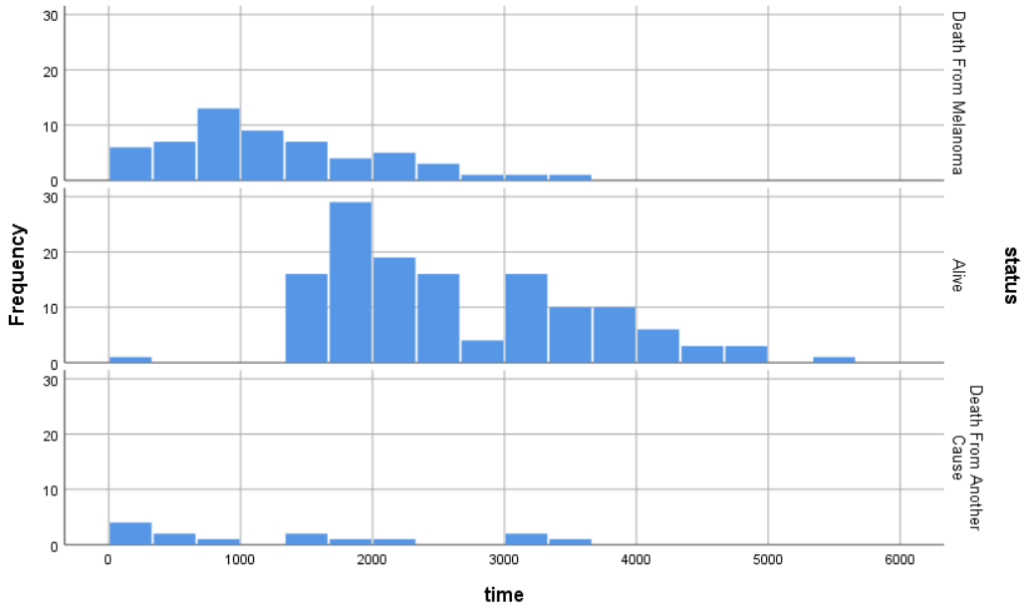
Πίνακας 6.3

Μπορούμε επίσης από ένα bar chart του φύλου με την παρουσία έλκους, να συμπαιράνουμε ότι οι περισσότεροι άντρες εμφανίζουν έλκος. Πιο συγκεκριμένα, φαίνεται ότι από τις γυναίκες σχεδόν το 37% έχει έλκος, ενώ για τους άντρες το ποσοστό αυτό είναι σχεδόν 55% (Σχήμα 6.4).



Σχήμα 6.4

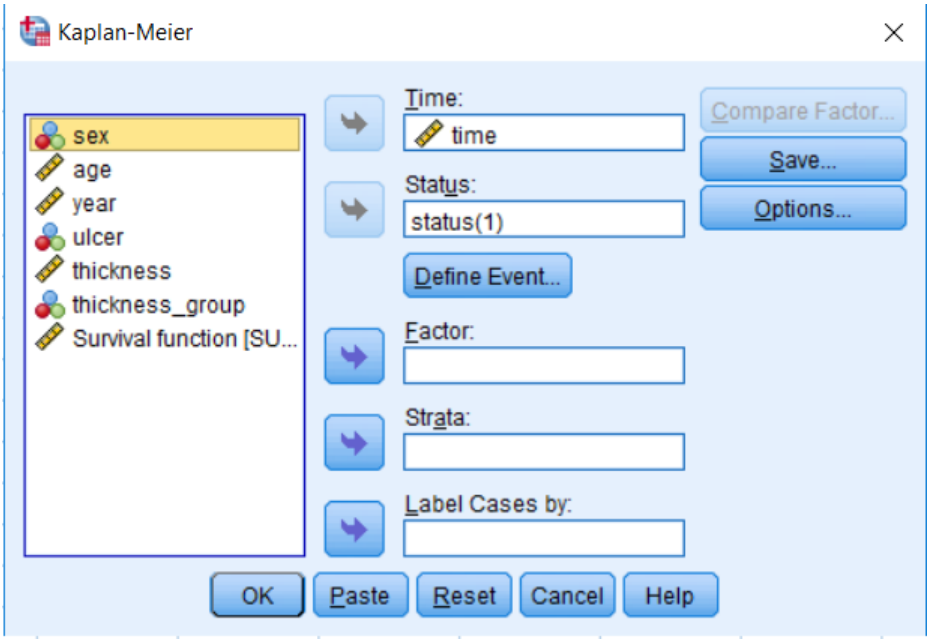
Ιστόγραμμα Χρόνου Επιβίωσης Ασθενών με Μελάνωμα για κάθε ένα αποτέλεσμα (status)



Σχήμα 6.5

Με τα παραπάνω 3 ιστογράμματα αντιλαμβανόμαστε ότι για τους περισσότερους ασθενείς που πέθαναν από μελάνωμα, ο θάνατος τους καταγράφηκε τα πρώτα 5 χρόνια ύστερα από την επέμβαση. Επιπλέον, βλέπουμε ότι για τους περισσότερους ασθενείς που παρέμειναν ζωντανοί μέχρι το τέλος της μελέτης, η κατάσταση τους είναι γνωστή πέραν των 5 ετών, ενώ για τους ασθενείς οι οποίοι πέθαναν από κάποια ανταγωνιστική αιτία, φαίνεται αρχικά ότι είναι λιγότεροι αλλά και ότι και η πλειοψηφία από αυτούς τους θανάτους καταγράφηκε νωρίτερα από τα 5 έτη.

Συνεχίζουμε με survival analysis
Στο SPSS ακολουθούμε τα εξής βήματα:
Analyze-Survival -Kaplan-Meier



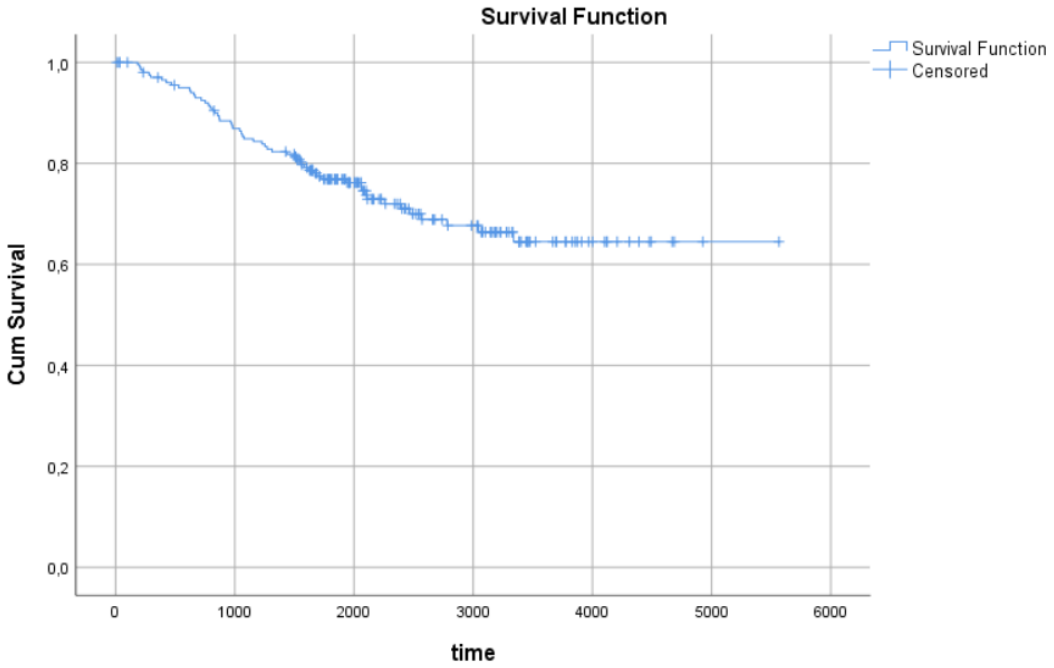
Σχήμα 6.6

Survival Table

	Time	Status	Cumulative Proportion Surviving at the Time: Estimate	Cumulative Proportion Surviving at the Time: Std. Error	N of Cumulative Events	N of Remaining Cases
1	10,000	Death From Another Cause	.	.	0	204
2	30,000	Death From Another Cause	.	.	0	203
3	35,000	Alive	.	.	0	202
4	99,000	Death From Another Cause	.	.	0	201
5	185,000	Death From Melanoma	,995	,005	1	200
6	204,000	Death From Melanoma	,990	,007	2	199
7	210,000	Death From Melanoma	,985	,009	3	198
8	232,000	Death From Melanoma	,980	,010	4	197
9	232,000	Death From Another Cause	.	.	4	196
10	279,000	Death From Melanoma	,975	,011	5	195
11	295,000	Death From Melanoma	,970	,012	6	194

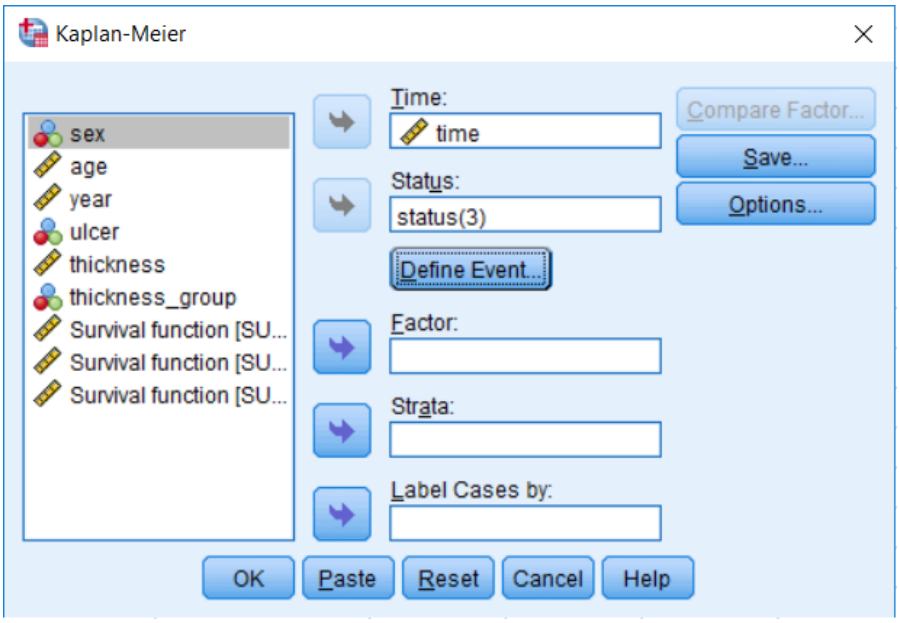
Πίνακας 6.4

Από τον Πίνακα 6.4 βλέπουμε για παράδειγμα ότι στις 185 ημέρες μετά την επέμβαση, 200 άτομα βρίσκονται σε κίνδυνο ενώ αποβιώνει 1 άτομο από μελάνωμα και η εκτίμηση της συνάρτησης επιβίωσης είναι Sˆ(t) = 0.995, δηλαδή η πιθανότητα να είναι ο χρόνος ζωής ενός ατόμου, μεγαλύτερος
του χρόνου t=185 είναι 99% και το τυπικό σφάλμα ισούται με sd = 0.005. Έτσι λοιπόν με βάση τον Πίνακα 6.4 αλλά και το Σχήμα 6.7 μπορούμε να συμπεράνουμε ότι όσο αυξάνεται ο χρόνος η πιθανότητα να επιβιώσει ένα άτομο μειώνεται, ενώ παράλληλα βλέπουμε πως με το πέρασμα του χρόνου, κάποιοι από τους ασθενείς πεθαίνουν λόγω του μελανώματος, αλλά η καμπύλη επιβίωσης μετά τις 3.500 ημέρες σταθεροποιείται στο 65%, δηλαδή η πιθανότητα να επιβιώσει ένα άτομο μετά από αυτή τη χρονική στιγμή είναι 65%.



Σχήμα 6.7

Καθορίζοντας event-->3



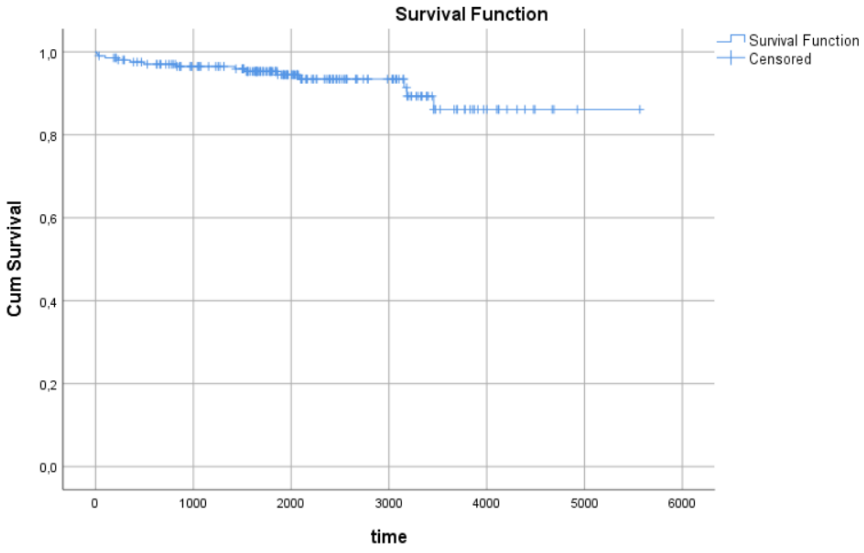
Σχήμα 6.8

Survival Table

	Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
			Estimate	Std. Error		
1	10,000	Death From Another Cause	,995	,005	1	204
2	30,000	Death From Another Cause	,990	,007	2	203
3	35,000	Alive	.	.	2	202
4	99,000	Death From Another Cause	,985	,008	3	201
5	185,000	Death From Melanoma	.	.	3	200
6	204,000	Death From Melanoma	.	.	3	199
7	210,000	Death From Melanoma	.	.	3	198
8	232,000	Death From Another Cause	,980	,010	4	197
9	232,000	Death From Melanoma	.	.	4	196

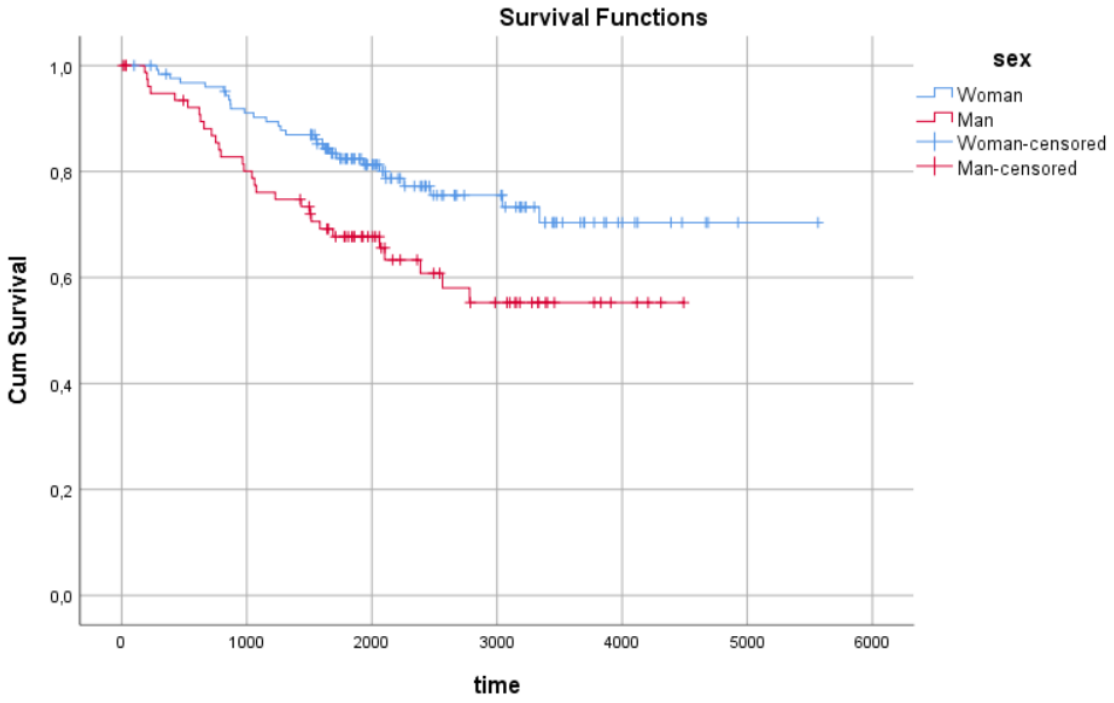
Πίνακας 6.5

Από τον Πίνακα 6.5 παρατηρούμε ότι στις 10 ημέρες μετά την επέμβαση 204 άτομα βρίσκονται σε κίνδυνο θανάτου λόγω άλλης αιτίας ενώ 1 άτομο αποβιώνει λόγω άλλης αιτίας και η εκτίμηση της συνάρτησης επιβίωσης Sˆ(t) = 0.995, δείχνει την πιθανότητα να είναι ο χρόνος ζωής ενός ατόμου μεγαλύτερος του χρόνου t=10 ημέρες. Το τυπικό σφάλμα ισούται με sd = 0.005. Συμπερασματικά λοιπόν, με βάση τον Πίνακα 5 αλλά και το Σχήμα 6.9 μπορούμε να καταλήξουμε στο συμπέρασμα ότι όσο αυξάνεται ο χρόνος η πιθανότητα να επιβιώσει ένα άτομο μειώνεται, αλλά τις πρώτες 2000 ημέρες μετά την επέμβαση, η πιθανότητα επιβίωσης των ασθενών από άλλη αιτία, είναι αρκετά υψηλή και μετά τις 3.500 ημέρες η πιθανότητα επιβίωσης σταθεροποιείται περίπου στο 86%.



Σχήμα 6.9

Στη συνέχεια μπορούμε να εξετάσουμε και εάν η επιβίωση από το μελάνωμα διαφέρει για το φύλο των ατόμων-ασθενών, θεωρώντας λογοκριμένους του ασθενείς που βιώνουν ένα ανταγωνιστικό γεγονός. Αυτό στο SPSS μπορούμε να το πραγματοποιήσουμε ως εξής:
Analyze-Survival-Kaplan Meier



Σχήμα 6.10

Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	6,468	1	,011

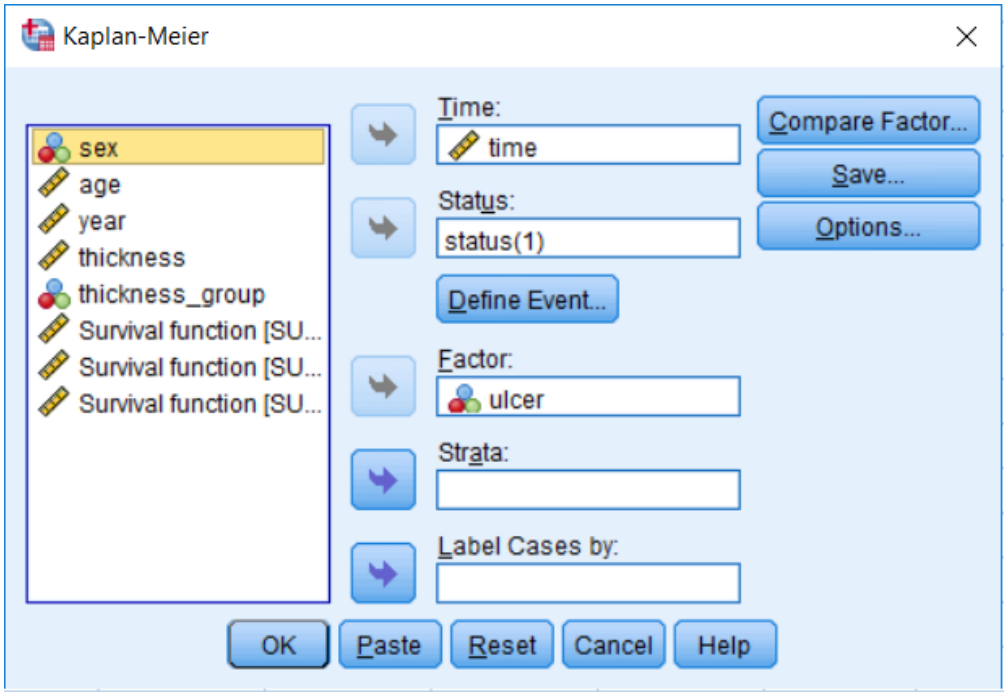
Test of equality of survival distributions for the different levels of sex.

Πίνακας 6.6

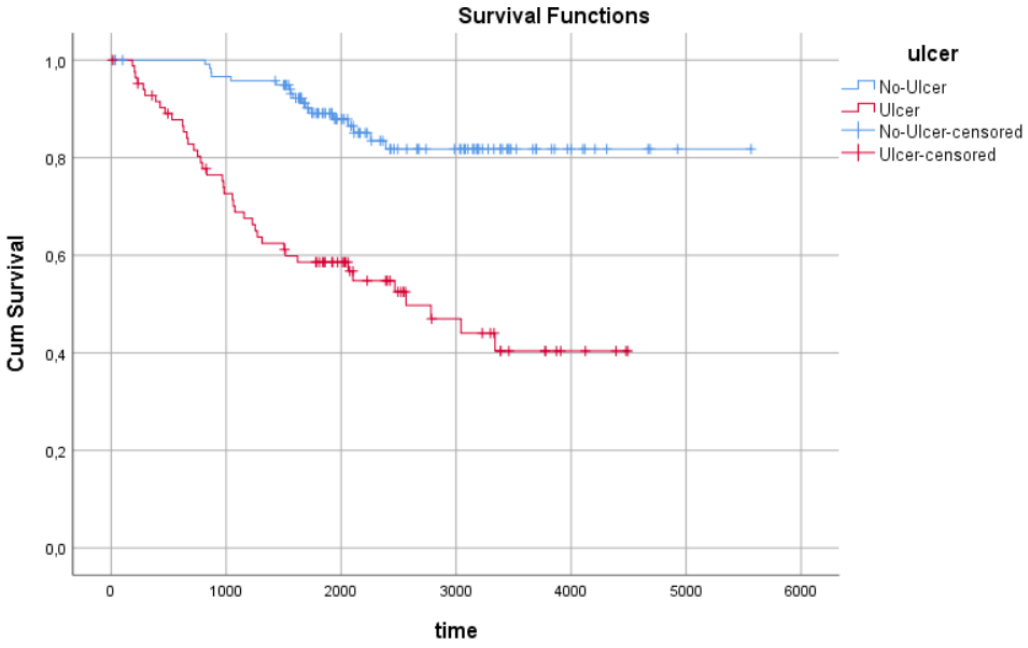
Παρατηρώντας τα παραπάνω αποτελέσματα, προκύπτει ότι σε επίπεδο σημαντικότητας α=0.05 απορρίπτουμε τη μηδενική υπόθεση (αφού p-value=0.011) έτσι, η επιβίωση των ασθενών από το μελάνωμα διαφέρει για τους άντρες και τις γυναίκες, το οποίο φαίνεται και γραφικά στο Σχήμα 6.10. Κοιτάζοντας το Σχήμα 6.9,μπορούμε να δούμε ότι οι γυναίκες έχουν μεγαλύτερη πιθανότητα επιβίωσης από το μελάνωμα σε σχέση με τους άντρες. Φαίνεται για παράδειγμα ότι 3000 ημέρες μετά την επέμβαση που υποβλήθηκαν οι ασθενείς, οι γυναίκες έχουν 75% πιθανότητα επιβίωσης, ενώ οι άντρες 57%.

Αντίστοιχα μπορούμε να δουλέψουμε και για τη μεταβλητή που αντιπροσωπεύει το έλκος.

Η διαδικασία στο SPSS:



Σχήμα 6.11



Σχήμα 6.12

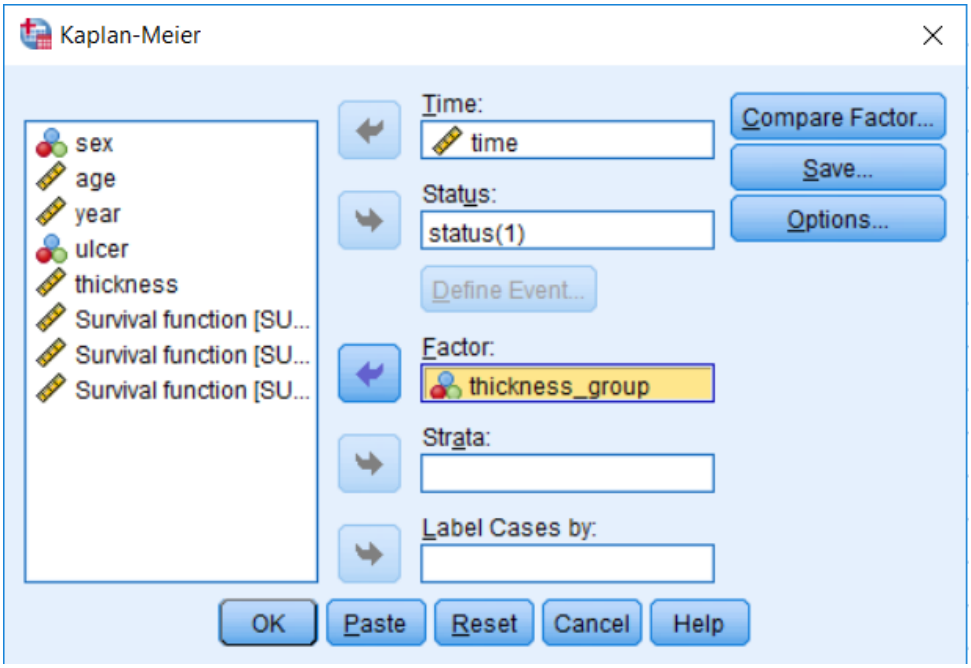
Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	32,572	1	,000

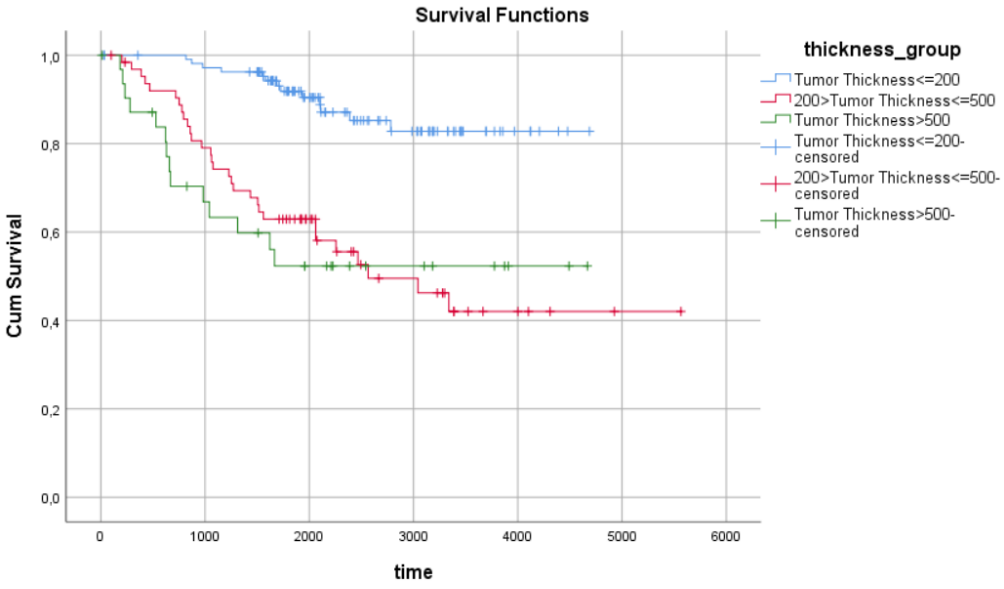
Test of equality of survival distributions for the different levels of ulcer.

Πίνακας 6.7

Παρατηρούμε ότι σε επίπεδο σημαντικότητας α=0.05 απορρίπτουμε την μηδενική υπόθεση, άρα η επιβίωση των ασθενών από μελάνωμα δεν είναι ίδια για ασθενείς που παρουσίασαν έλκος και για ασθενείς που δεν παρουσίασαν έλκος, το οποίο φαίνεται και γραφικά στο Σχήμα 6.12, από το οποίο προκύπτει ότι οι ασθενείς που δεν παρουσίασαν έλκος έχουν αρκετά μεγαλύτερη πιθανότητα επιβίωσης από τους ασθενείς που παρουσίασαν έλκος. Ομοίως, μπορεί να εξεταστεί εάν η πιθανότητα επιβίωσης για τους ασθενείς που απεβίωσαν από μελάνωμα διαφέρει για τις τρεις ομάδες μεγέθους όγκου που δημιουργήσαμε .



Σχήμα 6.13



Σχήμα 6.14

Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	31,582	2	,000

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

Πίνακας 6.8

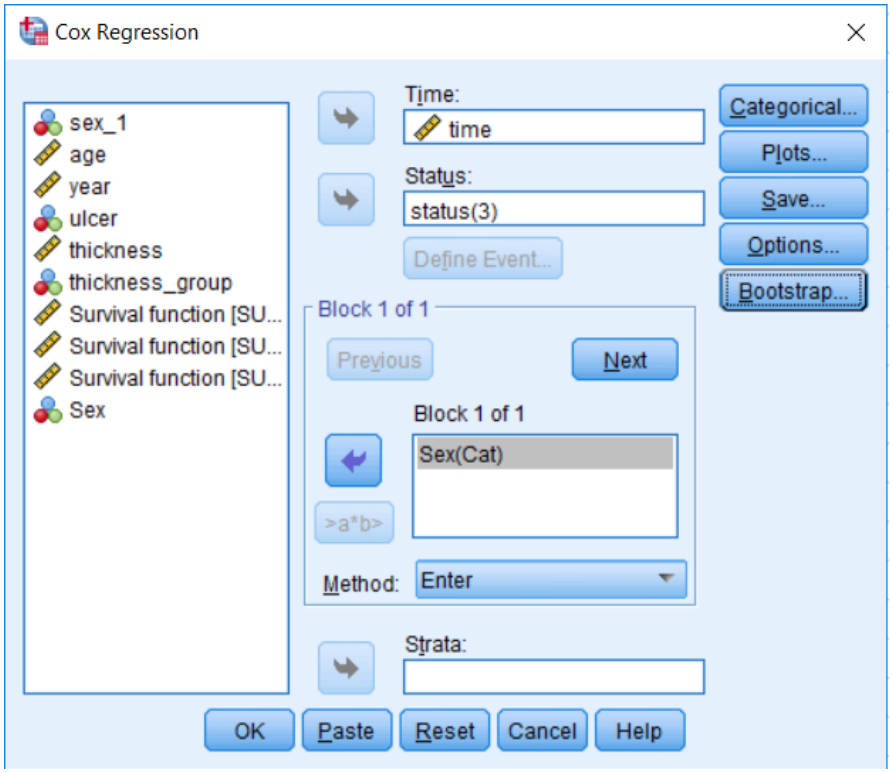
Επομένως, σε επίπεδο σημαντικότητας 0.05 απορρίπτουμε την μηδενική υπόθεση, δηλαδή η πιθανότητα επιβίωσης των ασθενών διαφέρει για τα τρία μεγέθη όγκου, το οποίο φαίνεται και στο Σχήμα 6.14. Με βάση το Σχήμα 6.14 προκύπτει ότι οι ασθενείς που απεβίωσαν από μελάνωμα και είχαν μέγεθος όγκου μικρότερο από 200 mm είχαν μεγαλύτερη πιθανότητα επιβίωσης από τους υπόλοιπους ασθενείς, ενώ για τους ασθενείς με μέγεθος όγκου >500, φαίνεται η πιθανότητα επιβίωσης μέχρι τις 2000 ημέρες, να είναι μικρότερη από την πιθανότητα επιβίωσης των ασθενών με μέγεθος όγκου (200, 500]. Μάλιστα παρατηρείται ότι η πιθανότητα επιβίωσης των ασθενών με μέγεθος όγκου (500, ∞] μετά τις 1800 ημέρες σταθεροποιείται στο 55%. Αντίθετα για τους ασθενείς που ανήκουν στην ομάδα με μέγεθος όγκου (200, 500] η πιθανότητα επιβίωσης εξακολουθεί να μειώνεται.

Cox Regression

Θα ξεκινήσουμε ένα κατάλληλο Cox μοντέλο αναλογικών κινδύνων με μοναδική συμμεταβλητή το φύλο και χρησιμοποιώντας το status=1 (θάνατος από μελάνωμα), ως event, θεωρώντας λογοκριμένους τους θανάτους από άλλη αιτία, και στη συνέχεια, χρησιμοποιώντας το status=3 (θάνατοι από άλλη αιτία) ως event, με λογοκριμένους τους θανάτους από μελάνωμα.

SPSS:

Analyze-Survival-Cox Regression



Σχήμα 6.15

Output Status=1

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
sex	,662	,265	6,238	1	,013	1,939

Πίνακας 6.9

Output Status=3

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
sex	,630	,536	1,385	1	,239	1,878

Πίνακας 6.10

Παρατηρώντας τους παραπάνω Πίνακες "Variables in the Equation" βλέπουμε ότι ο εκτιμώμενος συντελεστής για το φύλο των ασθενών όσον αφορά τον θάνατο από μελάνωμα υπολογίζεται ότι είναι B=0.662 με SE = 0.2651, με λόγο κινδύνου Exp(B) = 1.939 και p−value = 0.013. Αντίθετα, η επίδραση του φύλου πάνω στα δεδομένα ανταγωνιστικών κινδύνων δεν είναι σημαντική (B = 0.63 με SE = 0.536, Exp(B) = 1.878 και p−value = 0.239). Οπότε, το φύλο των ασθενών επηρεάζει το θάνατο από μελάνωμα, αλλά όχι το θάνατο από άλλες αιτίες οπότε καταλήγουμε στο συμπέρασμα ότι ο παράγοντας αυτός δεν χρειάζεται στο μοντέλο αφού και ο έλεγχος Wald test για την υπόθεση

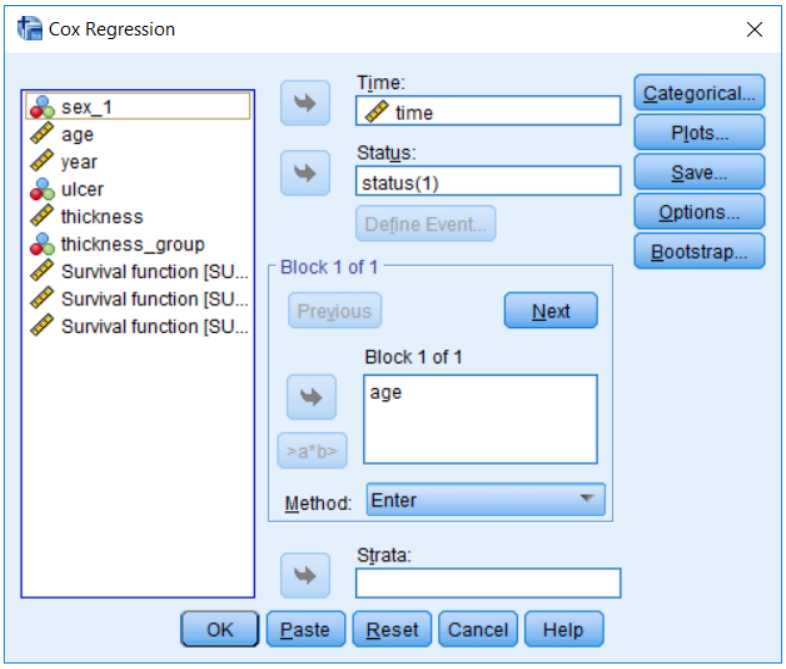
$$H_0 : \beta=0 \text{ έχει } p-value > 0.05.$$

Συμπέρασμα:

Άρα λοιπόν αφού στο δείγμα έχουν συμβολιστεί με 0 οι γυναίκες και με 1 οι άντρες και ο λόγος κινδύνου για το θάνατο από μελάνωμα ισούται με 1.93, φαίνεται ότι ένας άντρας ασθενής έχει κατά 93% φορές μεγαλύτερη πιθανότητα να πεθάνει από μελάνωμα σε σχέση με μια γυναίκα. Δηλαδή, ο κίνδυνος θανάτου που αντιμετωπίζει κάθε άντρας κάθε χρονική στιγμή ισούται με το 1.93 του κινδύνου που αντιμετωπίζει κάθε γυναίκα κάθε χρονική στιγμή. Συνεχίζοντας την ανάλυση, θα προσαρμόσουμε ένα Cox Regression μοντέλο για το θάνατο από μελάνωμα και το θάνατο από άλλα αίτια και την επίδραση που έχει σε αυτόν η ηλικία των ατόμων-ασθενών.

Status=1

Στο SPSS:



Σχήμα 6.16

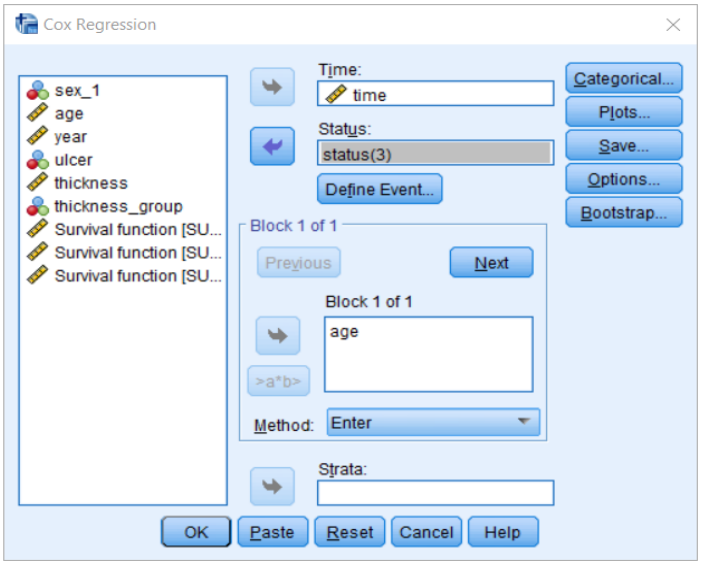
Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
age	,019	,009	4,804	1	,028	1,019

Πίνακας 6.11

Status=3 :

Στο SPSS



Σχήμα 6.17

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
age	,078	,022	13,227	1	,000	1,081

Πίνακας 6.12

Παρατηρώντας λοιπόν τους παραπάνω Πίνακες καταλήγουμε στο ότι η ηλικία επηρεάζει το θάνατο από μελάνωμα αλλά και το θάνατο από άλλη αιτία. Στην περίπτωση θανάτου από μελάνωμα, ο λόγος ισούται με Exp(B) = 1.019, δηλαδή κάθε φορά που προστίθεται ένας χρόνος στην ηλικία του ασθενούς συνδέεται με μία αύξηση κατά 1.019. Αν πάρουμε ένα παράδειγμα έστω δύο ασθενείς, ένας με ηλικία 75 και ένας με ηλικία 55 έτη. Εφόσον, ο λόγος κινδύνου για τους ασθενείς που πέθαναν από μελάνωμα ισούται με 1.019 προκύπτει ότι για την επικινδυνότητα μεταξύ των δύο αυτών ασθενών θα ισχύει το εξής:

$$HR = \exp(B^*(x_1 - x_2)) = 1{,}4622$$

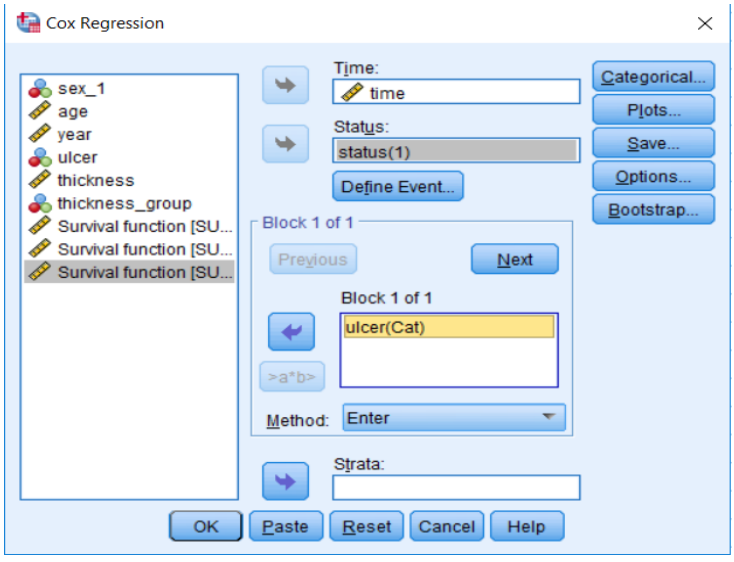
Δηλαδή, ο κίνδυνος να αποβιώσει ένας ασθενής λόγω του μελανώματος με ηλικία τα 75 ετών είναι 46% φορές μεγαλύτερος από τον κίνδυνο που αντιμετωπίζει κάθε χρονική στιγμή ένας ασθενής με ηλικία 55. Αντίστοιχα μπορούμε να δούμε και για τις περιπτώσεις θανάτου από άλλη αιτία. Έστω πάλι οι ασθενείς με ηλικίες 75 και 55 έτη αντίστοιχα. Στην περίπτωση θανάτου από άλλα αίτια, για την κινδυνότητα μεταξύ των δύο αυτών ασθενών ισχύει:

$$HR = \exp(B^*(x_1 - x_2)) = 4{,}7588$$

Συμπερασματικά λοιπόν, ο κίνδυνος να αποβιώσει ένας ασθενής εξαιτίας άλλων αιτίων με ηλικία τα 75 έτη είναι 4.75 φορές μεγαλύτερος (αύξηση 475%) από τον κίνδυνο που αντιμετωπίζει κάθε χρονική στιγμή ένας ασθενής με ηλικία 55 και παρατηρείται ότι στην περίπτωση θανάτου από άλλη αιτία οι ασθενείς με μεγαλύτερη ηλικία αντιμετωπίζουν μεγαλύτερο κίνδυνο κάθε χρονική στιγμή σε σχέση με τους ασθενείς που απεβίωσαν από μελάνωμα.

Συνεχίζουμε προσαρμόζοντας ένα Cox Regression μοντέλο για το θάνατο από μελάνωμα και το θάνατο από άλλα αίτια και την επίδραση που έχει σε αυτόν το έλκος.

Status = 1 :



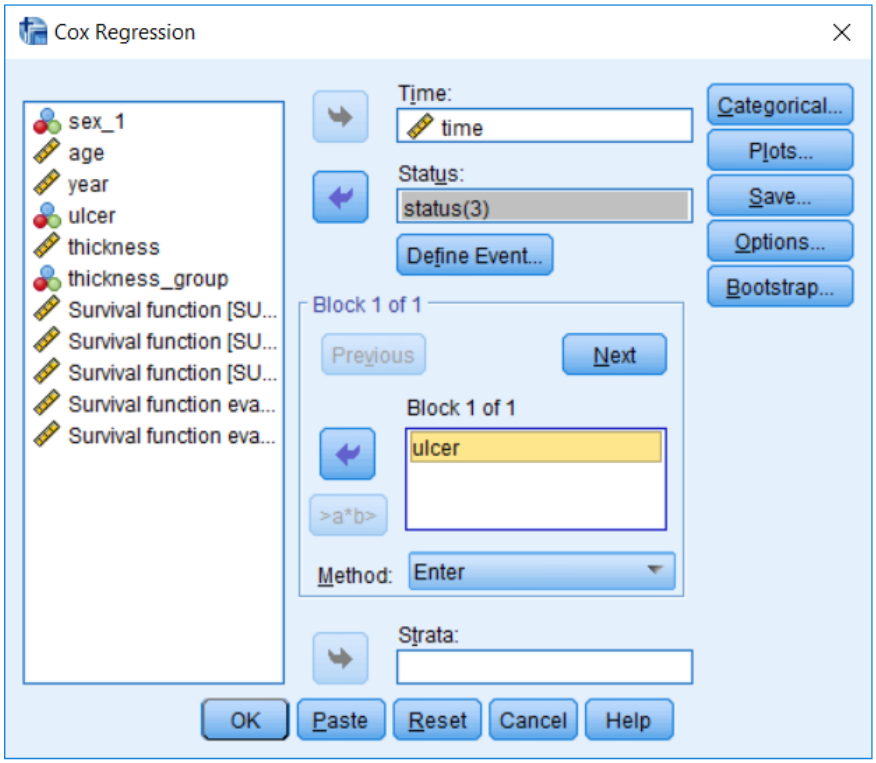
Σχήμα 6.18

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
ulcer	1,510	,290	27,092	1	,000	4,525

Πίνακας 6.13

Status = 3 :



Σχήμα 6.19

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
ulcer	,300	,542	,307	1	,580	1,350

Πίνακας 6.14

Από τους παραπάνω Πίνακες προκύπτει ότι η παρουσία ή η απουσία έλκους επηρεάζει το θάνατο του ασθενούς από μελάνωμα αλλά όχι το θάνατο του ασθενούς λόγω των άλλων αιτιών, αφού p−value =0.580, έτσι λοιπόν η παρουσία έλκους φαίνεται να μην επιδρά στη συνάρτηση κινδύνου που αφορά τα άλλα αίτια.

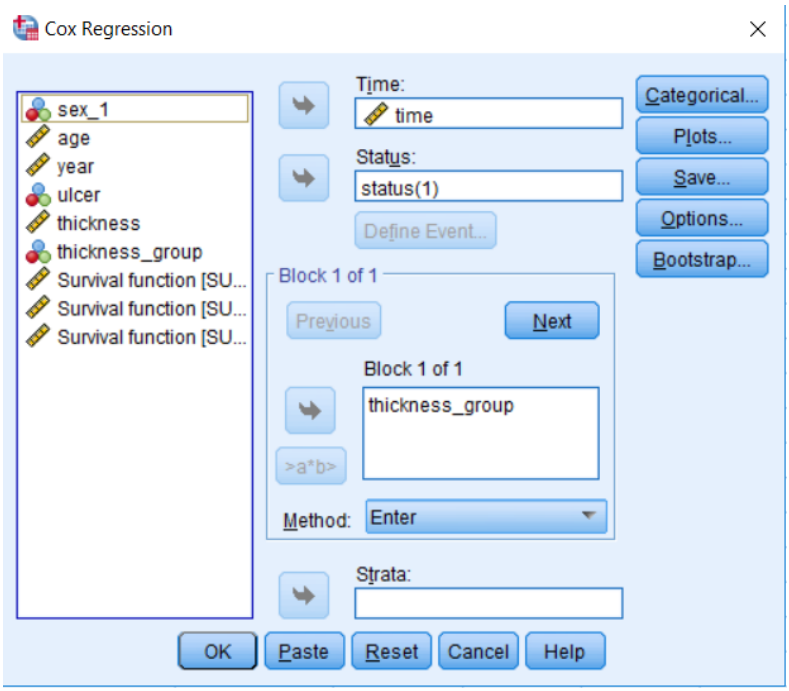
Στην περίπτωση θανάτου από μελάνωμα ο λόγος κινδύνου ισούται με Exp(B) = 4.525, το οποίο σημαίνει ότι για τους ασθενείς που παρουσίασαν έλκος, ο κίνδυνος θανάτου από μελάνωμα που αντιμετωπίζουν κάθε χρονική στιγμή οι ασθενείς είναι 4.52 φορές μεγαλύτερος (αύξηση 452%) από τον κίνδυνο που αντιμετωπίζουν κάθε χρονική στιγμή οι ασθενείς που δεν παρουσίασαν έλκος. Αντίστοιχα, στην περίπτωση θανάτου λόγω άλλων αιτιών, ο λόγος κινδύνου ισούται με 1.35, και η ύπαρξη έλκους αυξάνει τον κίνδυνο

θανάτου από άλλη αιτία κατά 0.300 φορές, ωστόσο η μεταβλητή ulcer σε αυτήν την περίπτωση δεν είναι στατιστικά σημαντική.

Συνεχίζουμε ελέγχοντας και τον παράγοντα πάχος όγκου και την επιρροή του στο θάνατο από μελάνωμα και στο θάνατο λόγω των ανταγωνιστικών αιτιών.

Status=1

Στο SPSS:



Σχήμα 6.20

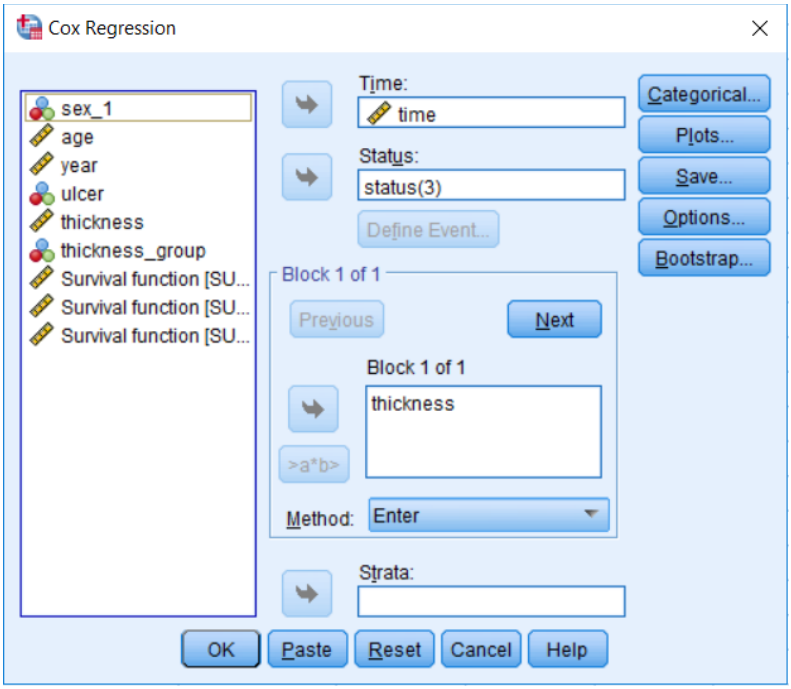
Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
thickness	,160	,031	26,273	1	,000	1,174

Πίνακας 6.15

Status=3

Στο SPSS:



Σχήμα 6.21

Variables in the Equation

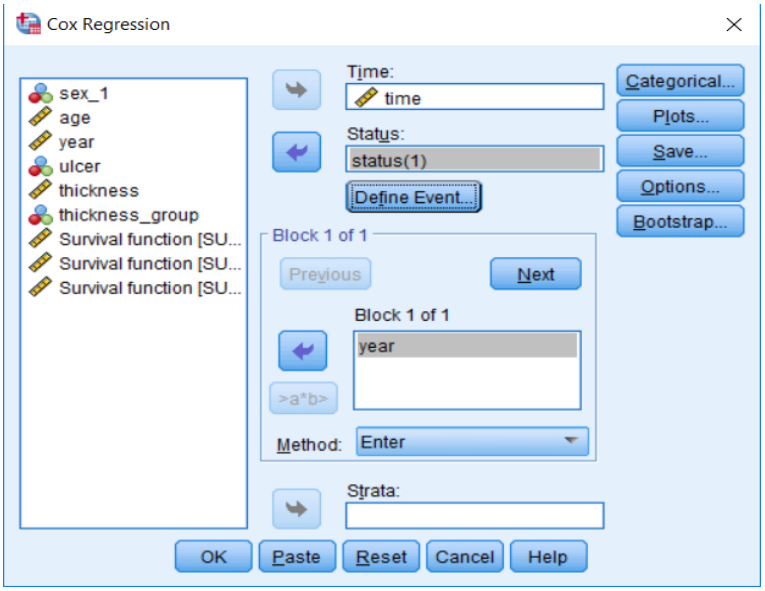
	B	SE	Wald	df	Sig.	Exp(B)
thickness	,109	,073	2,218	1	,136	1,115

Πίνακας 6.16

Σύμφωνα με τα παραπάνω παρατηρούμε ότι το μέγεθος του όγκου με συντελεστή $\hat{\beta}$= 0.160 στην περίπτωση θανάτου από μελάνωμα, επηρεάζει τη συνάρτηση κινδύνου των ασθενών. Αντίθετα, η συνάρτηση κινδύνου που αφορά τα ανταγωνιστικά αίτια φαίνεται να μην επηρεάζεται από τη μεταβλητή thickness εφόσον έχει p−value =0136. Θα ελέγξουμε και τη μεταβλητή που εκφράζει το έτος της επέμβασης κατα πόσο επιδρά στο θάνατο από μελάνωμα και στο θάνατο από ανταγωνιστικές αιτίες.

Status=1

Στο SPSS



Σχήμα 6.22

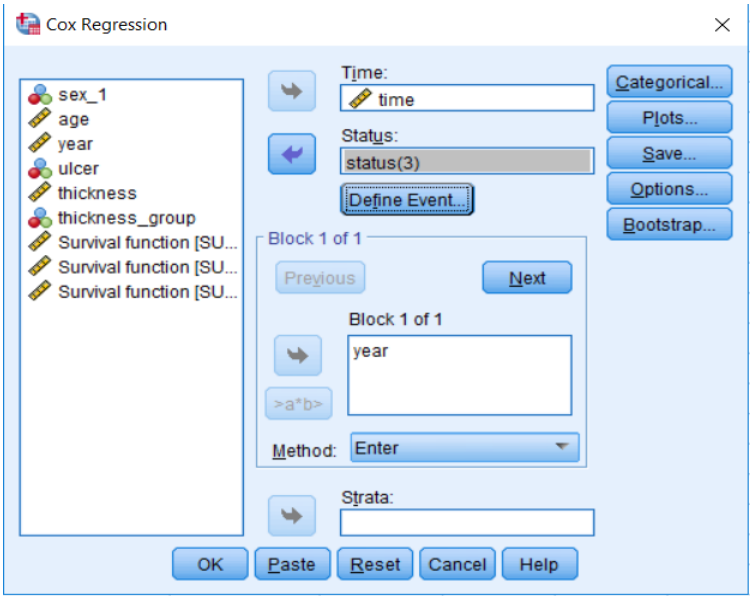
Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
year	-,069	,053	1,687	1	,194	,934

Πίνακας 6.17

Status=3

Στο SPSS:



Σχήμα 6.23

Variables in the Equation

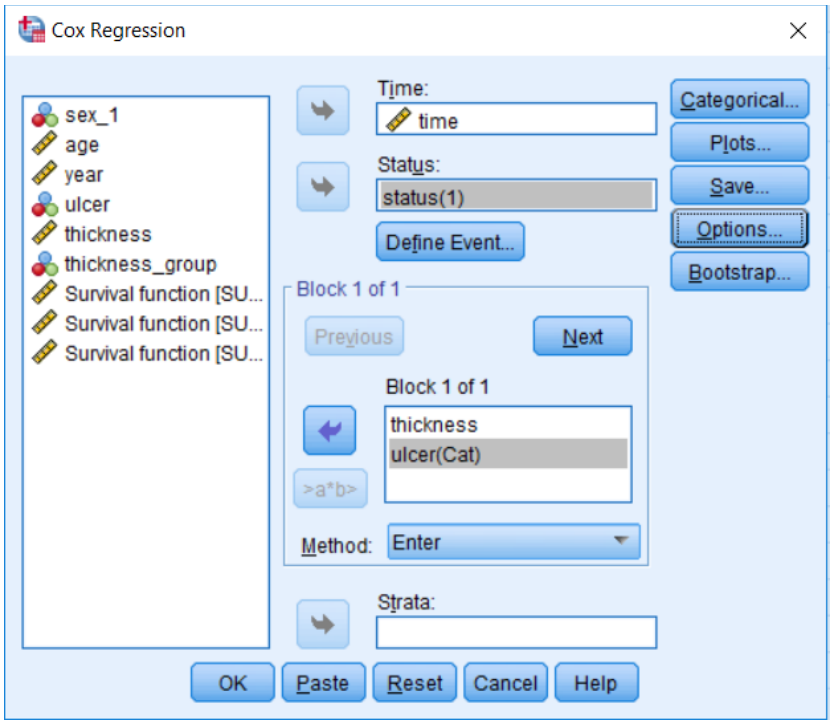
	B	SE	Wald	df	Sig.	Exp(B)
year	,080	,123	,425	1	,514	1,084

Πίνακας 6.18

Βλέποντας τα αποτελέσματα καταλήγουμε στο συμπέρασμα ότι το έτος που πραγματοποιήθηκε η επέμβαση για αφαίρεση του μελανώματος, δεν είναι στατιστικά σημαντική μεταβλητή και επομένως δεν επηρεάζει ούτε το κίνδυνο θανάτου από μελάνωμα ούτε το κίνδυνο θανάτου λόγω άλλων αιτιών.
Ας προχωρήσουμε ελέγχοντας τους παράγοντες έλκος (ulcer) και μέγεθος του όγκου (thickness) και την επιρροή τους στο θάνατο από μελάνωμα και στο θάνατο από ανταγωνιστικά αίτια χρησιμοποιούμε το παρακάτω μοντέλο Cox.

Status=1

Στο SPSS:



Σχήμα 6.24

Variables in the Equation

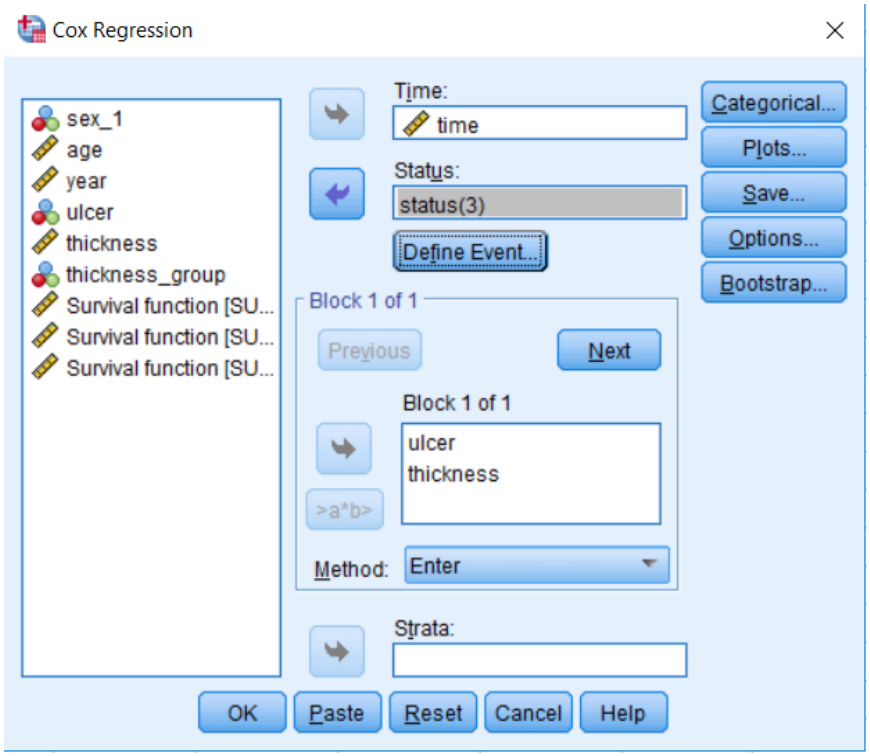
	B	SE	Wald	df	Sig.	Exp(B)
ulcer	1,279	,302	17,978	1	,000	3,595
thickness	,114	,035	10,537	1	,001	1,121

Πίνακας 6.19

Από τον Πίνακα 6.19 παρατηρούμε ότι το έλκος και το μέγεθος του όγκου με συντελεστές $\beta_1 = 1.279$ και $\beta_2 = 0.114$ αντίστοιχα επηρεάζουν τη συνάρτηση κινδύνου των ασθενών που αφορά το μελάνωμα και το θετικό πρόσημο δηλώνει ότι ένας ασθενής με μεγάλο μέγεθος όγκου ή με παρουσία έλκους διατρέχει μεγαλύτερο κίνδυνο να αποβιώσει λόγω του μελανώματος σε σχέση με έναν ασθενή με μικρότερο όγκο και χωρίς έλκος.

Status=3

Στο SPSS:



Σχήμα 6.25

Variables in the Equation

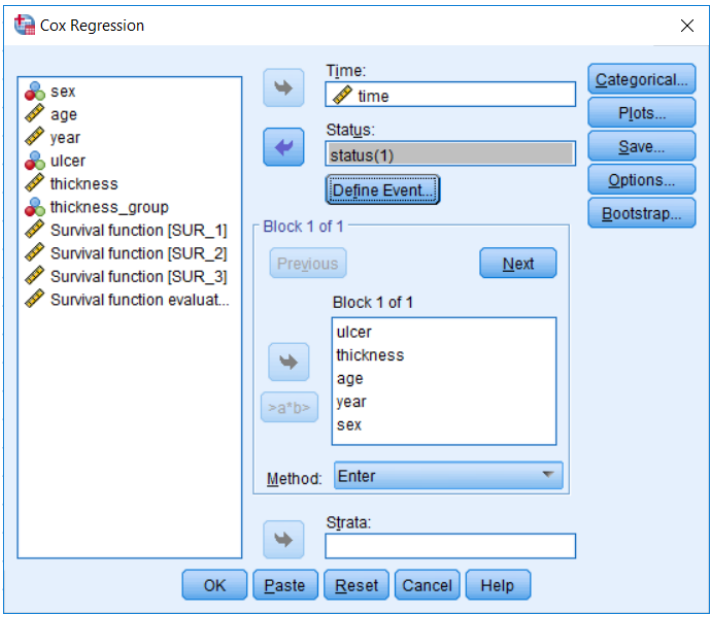
	B	SE	Wald	df	Sig.	Exp(B)
ulcer	,054	,582	,009	1	,926	1,056
thickness	,106	,079	1,824	1	,177	1,112

Πίνακας 6.20

Αντίθετα, από τον Πίνακα 6.20 βλέπουμε ότι στους ανταγωνιστικούς θανάτους φαίνεται ο δείκτης έλκους και το μέγεθος του όγκου να μην επηρεάζουν τη συνάρτηση κινδύνου και επομένως οι συντελεστές $\beta_1 = 0.054$, $\beta_2 = 0.106$ δεν είναι στατιστικά σημαντικοί, $p-value > 0.05$, το οποίο σημαίνει ότι οι παράγοντες αυτοί δεν χρειάζονται στο μοντέλο. Άρα η συνάρτηση κινδύνου δεν εξαρτάται από τις μεταβλητές ulcer, thickness. Ας δούμε και την επίδραση όλων των παραγόντων στο θάνατο από μελάνωμα και στο θάνατο από ανταγωνιστικά αιτία.

Status=1

Στο SPSS:



Σχήμα 6.26

Variables in the Equation

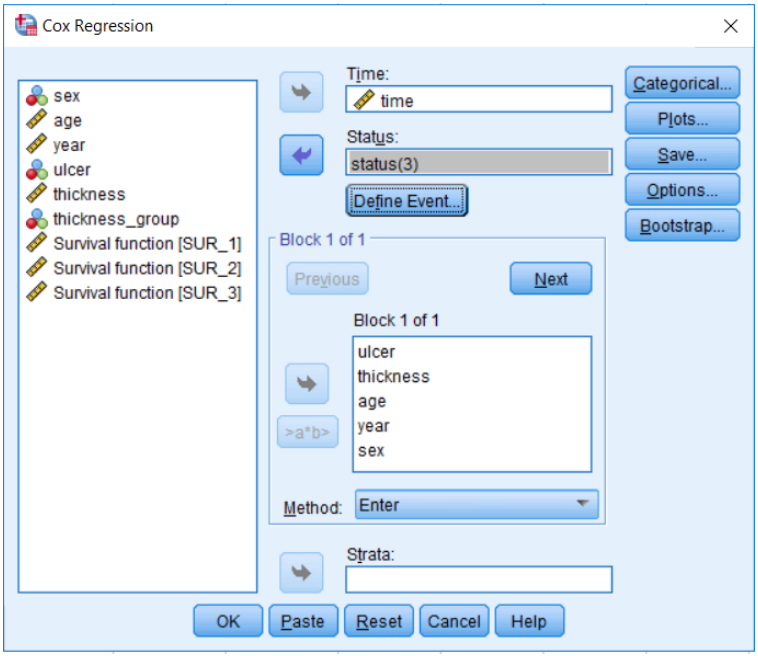
	B	SE	Wald	df	Sig.	Exp(B)
ulcer	1,155	,310	13,866	1	,000	3,174
thickness	,109	,036	9,069	1	,003	1,115
age	,014	,009	2,619	1	,106	1,014
year	-,075	,061	1,507	1	,220	,928
sex	,268	,275	,954	1	,329	1,308

Πίνακας 6.21

Από τον παραπάνω πίνακα και σύμφωνα με τον έλεγχο Wald προκύπτει το συμπέρασμα ότι πρέπει να απορρίψουμε την μηδενική υπόθεση $H_0 : \beta_{sex} = \beta_{age} = \beta_{year} = \beta_{ulcer} = \beta_{thickness} = 0$ και άρα να δεχτούμε ότι η συνάρτηση κινδύνου εξαρτάται από τουλάχιστον έναν από τους παράγοντες sex, age, year, ulcer, thickness. Παρατηρούμε ότι το φύλο, η ηλικία, ο δείκτης έλκους και το πάχος του όγκου αυξάνουν το κίνδυνο θανάτου από μελάνωμα ενώ το έτος της επέμβασης τον μειώνει. Το έτος της επέμβασης, η ηλικία και το φύλο των ασθενών δεν θεωρούνται στατιστικά σημαντικοί και επομένως δεν χρειάζονται στο μοντέλο. Επιπλέον, σύμφωνα με τα παραπάνω αποτελέσματα προκύπτει ότι η ύπαρξη έλκους αυξάνει τον κίνδυνο θανάτου από μελάνωμα κατά 1.155 φορές και είναι ένας από τους σημαντικούς παράγοντες για το μοντέλο μαζί με το πάχος του όγκου που αυξάνει τον κίνδυνο κατά 0.109 φορές.

Status=3

Στο SPSS:



Σχήμα 6.27

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
ulcer	-,308	,614	,251	1	,616	,735
thickness	,068	,084	,642	1	,423	1,070
age	,075	,022	11,169	1	,001	1,077
year	,011	,139	,007	1	,935	1,011
sex	,446	,564	,627	1	,428	1,563

Πίνακας 6.22

Από τον πίνακα αυτού του μοντέλου Cox συμπεραίνουμε ότι η επίδραση όλων των παραγόντων είναι διαφορετική στην συνάρτηση κινδύνου που αφορά το θάνατο από ανταγωνιστικά αιτία. Κοιτώντας τα αποτελέσματα βλέπουμε ότι σχεδόν όλοι οι παράγοντες αυξάνουν τον κίνδυνο θανάτου από άλλα αιτία ωστόσο σε αυτό το μοντέλο μόνο η μεταβλητή που αφορά την ηλικία είναι στατιστικά σημαντική. Σύμφωνα με τον έλεγχο Wald, απορρίπτουμε την H_0 : $\beta_{sex} = \beta_{age} = \beta_{year} = \beta_{ulcer} = \beta_{thickness} = 0$, και άρα μπορούμε να δεχτούμε ότι η συνάρτηση κινδύνου που αφορά τα ανταγωνιστικά αίτια εξαρτάται από τουλάχιστον έναν από τους παράγοντες sex, age, year, ulcer, thickness.

Μεταβλητές	Θάνατος από μελάνωμα		Θάνατος από ανταγωνιστικές αιτίες	
	Παράγοντας	p-value	Παράγοντας	p-value
sex	0,662	0,013	-0,63	0,239
age	0,019	0,028	0,078	0,00
ulcer	1,51	0,00	0,3	0,58
thickness	0,16	0,00	0,109	0,136
year	-0,069	0,194	0,08	0,514

Πίνακας 6.23

Μεταβλητές	Θάνατος από μελάνωμα		Θάνατος από ανταγωνιστικές αιτίες	
	Παράγοντας	p-value	Παράγοντας	p-value
sex	0,268	0,329	0,446	0,428
age	0,014	0,106	0,075	0,001
ulcer	1,155	0,000	-0,308	0,616
thickness	0,109	0,003	0,068	0,423
year	-0,075	0,220	0,011	0,935

Πίνακας 6.24

Ο Πίνακας 6.23 περιέχει συγκεντρωτικά τις επιδράσεις των μεταβλητών στο θάνατο από μελάνωμα και στο θάνατο από ανταγωνιστικές αιτίες από τα μοντέλα που αποτελούνται κάθε φορά από μόνο μία μεταβλητή. Ο Πίνακας 6.24 δείχνει την επίδραση των μεταβλητών που προέκυψε από το μοντέλο αναλογικών κινδύνων του Cox το οποίο αποτελείται από όλες τις μεταβλητές του δείγματος.

Συμπέρασμα :

Τα δύο μοντέλα που αφορούν το θάνατο από μελάνωμα διαφέρουν μεταξύ τους. Το μοντέλο που αποτελείται από όλες τις διαθέσιμες μεταβλητές η μεταβλητή που αφορά το φύλο των ασθενών δεν είναι στατιστικά σημαντική, σε αντίθεση με το μοντέλο του Πίνακα 22 όπου είναι στατιστικά σημαντική. Παρ΄ όλα αυτά και στα δύο μοντέλα προέκυψε ότι ο πιο σημαντικός προγνωστικός παράγοντας είναι η μεταβλητή που αφορά το έλκος ενώ ο δεύτερος πιο σημαντικός παράγοντας είναι το πάχος του όγκου. Αντίθετα, στο θάνατο από ανταγωνιστικές αιτίες συμπεραίνεται ότι η μόνη στατιστικά σημαντική μεταβλητή που επηρεάζει το θάνατο είναι η ηλικία του ασθενούς.

ΠΑΡΑΡΤΗΜΑ

Δεδομένα παραδείγματος Κεφάλαιο 4

Τα παραπάνω δεδομένα είναι από το βιβλίο Lawless(2002) καθώς και από το άρθρο Hoel (1972) και αναφέρονται στο **παράδειγμα 4.1**

Λέμφωμα		Σάρκωμα			Άλλες αιτίες		
Περιβάλλον εργαστηρίου	Αποστειρωμένο περιβάλλον	Περιβάλλον εργαστηρίου		Αποστειρωμένο περιβάλλον	Περιβάλλον εργαστηρίου		Αποστειρωμένο περιβάλλον
159	192	317	665	430	40	421	136
189	193	318	679	590	42	565	246
191	194	399	691	606	51	616	255
198	195	495	693	638	62	617	376
200	202	525	696	665	163	652	421
207	212	536	747	679	179	655	565
220	215	549	752	691	206	658	616
235	229	552	760	693	222	660	617
245	230	554	778	696	228	662	652
250	237	557	821	747	252	675	655
250	240	558	986	752	259	681	658
256	244	571		760	282	734	660
261	247	586		778	324	736	662
265	259	594		821	333	737	675
266	300	596		986	341	757	681
280	301	605			366	769	734
343	321	612			385	777	736
350	337	621			407	800	737
383	415	628			420	806	757
403	434	631			431	825	769
414	444	636			441	855	777
428	485	643			461	857	800
432	496	647			462	864	806
	529	648			482	868	825
	537	649			517	870	855
	624	661			524	870	857
	707	663			564	837	864
	800	666			567	882	868
		670			586	895	870
		695			619	920	870
		697			620	934	837
		700			621	942	882
		705			622	1015	895
		712			647	1019	920
		713			651		934
		738			686		942
		748			761		1015
		753			763		1019
		430			136		
		590			246		
		606			255		
		638			376		

Δεδομένα Κεφάλαιο 6

time	status	sex	age	year	ulcer	thickness	thickness_group
10	3	1	76	1972	1	6.76	3
30	3	1	56	1968	0	0.65	1
35	2	1	41	1977	0	1.34	1
99	3	0	71	1968	0	2.9	2
185	1	1	52	1965	1	12.08	3
204	1	1	28	1971	1	4.84	2
210	1	1	77	1972	1	5.16	3
232	3	0	60	1974	1	3.22	2
232	1	1	49	1968	1	12.88	3
279	1	0	68	1971	1	7.41	3
295	1	0	53	1969	1	4.19	2
355	3	0	64	1972	1	0.16	1
386	1	0	68	1965	1	3.87	2
426	1	1	63	1970	1	4.84	2
469	1	0	14	1969	1	2.42	2
493	3	1	72	1971	1	12.56	3
529	1	1	46	1971	1	5.8	3
621	1	1	72	1972	1	7.06	3
629	1	1	95	1968	1	5.48	3
659	1	1	54	1972	1	7.73	3
667	1	0	89	1968	1	13.85	3
718	1	1	25	1967	1	2.34	2
752	1	1	37	1973	1	4.19	2
779	1	1	43	1967	1	4.04	2
793	1	1	68	1970	1	4.84	2
817	1	0	67	1966	0	0.32	1
826	3	0	86	1965	1	8.54	3
833	1	0	56	1971	1	2.58	2
858	1	0	16	1967	0	3.56	2
869	1	0	42	1965	0	3.54	2
872	1	0	65	1968	0	0.97	1
967	1	1	52	1970	1	4.83	2
977	1	1	58	1967	1	1.62	1
982	1	0	60	1970	1	6.44	3
1041	1	1	68	1967	0	14.66	3
1055	1	0	75	1967	1	2.58	2
1062	1	1	19	1966	1	3.87	2
1075	1	1	66	1971	1	3.54	2
1156	1	0	56	1970	1	1.34	1
1228	1	1	46	1973	1	2.24	2
1252	1	0	58	1971	1	3.87	2
1271	1	0	74	1971	1	3.54	2
1312	1	0	65	1970	1	17.42	3
1427	3	1	64	1972	0	1.29	1
1435	1	1	27	1969	0	3.22	2
1499	2	1	73	1973	0	1.29	1
1506	1	1	56	1970	1	4.51	2
1508	2	1	63	1973	1	8.38	3

1510	2	0	69	1973	0	1.94	1
1512	2	0	77	1973	0	0.16	1
1516	1	1	80	1968	1	2.58	2
1525	3	0	76	1970	0	1.29	1
1542	2	0	65	1973	0	0.16	1
1548	1	0	61	1972	0	1.62	1
1557	2	0	26	1973	0	1.29	1
1560	1	0	57	1973	0	2.1	2
1563	2	0	45	1973	0	0.32	1
1584	1	1	31	1970	0	0.81	1
1605	2	0	36	1973	0	1.13	1
1621	1	0	46	1972	1	5.16	3
1627	2	0	43	1973	0	1.62	1
1634	2	0	68	1973	0	1.37	1
1641	2	1	57	1973	0	0.24	1
1641	2	0	57	1973	0	0.81	1
1648	2	0	55	1973	0	1.29	1
1652	2	0	58	1973	0	1.29	1
1654	2	1	20	1973	0	0.97	1
1654	2	0	67	1973	0	1.13	1
1667	1	0	44	1971	0	5.8	3
1678	2	0	59	1973	0	1.29	1
1685	2	0	32	1973	0	0.48	1
1690	1	1	83	1971	0	1.62	1
1710	2	0	55	1973	0	2.26	2
1710	2	1	15	1973	0	0.58	1
1726	1	0	58	1970	0	0.97	1
1745	2	0	47	1973	0	2.58	2
1762	2	0	54	1973	0	0.81	1
1779	2	1	55	1973	1	3.54	2
1787	2	1	38	1973	0	0.97	1
1787	2	0	41	1973	1	1.78	1
1793	2	0	56	1973	0	1.94	1
1804	2	0	48	1973	0	1.29	1
1812	2	1	44	1973	1	3.22	2
1836	2	0	70	1972	0	1.53	1
1839	2	0	40	1972	0	1.29	1
1839	2	1	53	1972	1	1.62	1
1854	2	0	65	1972	1	1.62	1
1856	2	1	54	1972	0	0.32	1
1860	3	1	71	1969	1	4.84	2
1864	2	0	49	1972	0	1.29	1
1899	2	0	55	1972	0	0.97	1
1914	2	0	69	1972	0	3.06	2
1919	2	1	83	1972	0	3.54	2
1920	2	1	60	1972	1	1.62	1
1927	2	1	40	1972	1	2.58	2
1933	1	0	77	1972	0	1.94	1
1942	2	0	35	1972	0	0.81	1
1955	2	0	46	1972	0	7.73	3
1956	2	0	34	1972	0	0.97	1
1958	2	0	69	1972	0	12.88	3
1963	2	0	60	1972	0	2.58	2

1970	2	1	84	1972	1	4.09	2
2005	2	0	66	1972	0	0.64	1
2007	2	1	56	1972	0	0.97	1
2011	2	0	75	1972	1	3.22	2
2024	2	0	36	1972	0	1.62	1
2028	2	1	52	1972	1	3.87	2
2038	2	0	58	1972	1	0.32	1
2056	2	0	39	1972	0	0.32	1
2059	2	1	68	1972	1	3.22	2
2061	1	1	71	1968	1	2.26	2
2062	1	0	52	1965	0	3.06	2
2075	2	1	55	1972	1	2.58	2
2085	3	0	66	1970	0	0.65	1
2102	2	1	35	1972	1	1.13	1
2103	1	1	44	1966	1	0.81	1
2104	2	0	72	1972	0	0.97	1
2108	1	0	58	1969	0	1.76	1
2112	2	0	54	1972	0	1.94	1
2150	2	0	33	1972	0	0.65	1
2156	2	0	45	1972	0	0.97	1
2165	2	1	62	1972	0	5.64	3
2209	2	0	72	1971	0	9.66	3
2227	2	0	51	1971	0	0.1	1
2227	2	1	77	1971	1	5.48	3
2256	1	0	43	1971	0	2.26	2
2264	2	0	65	1971	0	4.83	2
2339	2	0	63	1971	0	0.97	1
2361	2	1	60	1971	0	0.97	1
2387	2	0	50	1971	1	5.16	3
2388	1	1	40	1966	0	0.81	1
2403	2	0	67	1971	1	2.9	2
2426	2	0	69	1971	0	3.87	2
2426	2	0	74	1971	1	1.94	1
2431	2	0	49	1971	0	0.16	1
2460	2	0	47	1971	0	0.64	1
2467	1	0	42	1965	1	2.26	2
2492	2	0	54	1971	0	1.45	1
2493	2	1	72	1971	1	4.82	2
2521	2	0	45	1971	1	1.29	1
2542	2	1	67	1971	1	7.89	3
2559	2	0	48	1970	1	0.81	1
2565	1	1	34	1970	1	3.54	2
2570	2	0	44	1970	0	1.29	1
2660	2	0	31	1970	0	0.64	1
2666	2	0	42	1970	0	3.22	2
2676	2	0	24	1970	0	1.45	1
2738	2	0	58	1970	0	0.48	1
2782	1	1	78	1969	1	1.94	1
2787	2	1	62	1970	1	0.16	1
2984	2	1	70	1969	0	0.16	1
3032	2	0	35	1969	0	1.29	1
3040	2	0	61	1969	0	1.94	1
3042	1	0	54	1967	1	3.54	2

3067	2	0	29	1969	0	0.81	1
3079	2	1	64	1969	0	0.65	1
3101	2	1	47	1969	0	7.09	3
3144	2	1	62	1969	0	0.16	1
3152	2	0	32	1969	0	1.62	1
3154	3	1	49	1969	0	1.62	1
3180	2	0	25	1969	0	1.29	1
3182	3	1	49	1966	0	6.12	3
3185	2	0	64	1969	0	0.48	1
3199	2	0	36	1969	0	0.64	1
3228	2	0	58	1969	1	3.22	2
3229	2	0	37	1969	0	1.94	1
3278	2	1	54	1969	0	2.58	2
3297	2	0	61	1968	1	2.58	2
3328	2	1	31	1968	0	0.81	1
3330	2	1	61	1968	1	0.81	1
3338	1	0	60	1967	1	3.22	2
3383	2	0	43	1968	0	0.32	1
3384	2	0	68	1968	1	3.22	2
3385	2	0	4	1968	0	2.74	2
3388	2	1	60	1968	1	4.84	2
3402	2	1	50	1968	0	1.62	1
3441	2	0	20	1968	0	0.65	1
3458	3	0	54	1967	0	1.45	1
3459	2	0	29	1968	0	0.65	1
3459	2	1	56	1968	1	1.29	1
3476	2	0	60	1968	0	1.62	1
3523	2	0	46	1968	0	3.54	2
3667	2	0	42	1967	0	3.22	2
3695	2	0	34	1967	0	0.65	1
3695	2	0	56	1967	0	1.03	1
3776	2	1	12	1967	1	7.09	3
3776	2	0	21	1967	1	1.29	1
3830	2	1	46	1967	0	0.65	1
3856	2	0	49	1967	0	1.78	1
3872	2	0	35	1967	1	12.24	3
3909	2	1	42	1967	1	8.06	3
3968	2	0	47	1967	0	0.81	1
4001	2	0	69	1967	0	2.1	2
4103	2	0	52	1966	0	3.87	2
4119	2	1	52	1966	0	0.65	1
4124	2	0	30	1966	1	1.94	1
4207	2	1	22	1966	0	0.65	1
4310	2	1	55	1966	0	2.1	2
4390	2	0	26	1965	1	1.94	1
4479	2	0	19	1965	1	1.13	1
4492	2	1	29	1965	1	7.06	3
4668	2	0	40	1965	0	6.12	3
4688	2	0	42	1965	0	0.48	1
4926	2	0	50	1964	0	2.26	2
5565	2	0	41	1962	0	2.9	2

ΒΙΒΛΙΟΓΡΑΦΙΑ

Aalen, O. (1976). Nonparametric inference in connection with multiple decrement models. *Scand. J. Stat*., **3**,15-27.

Aalen, O. and Johansen, S. (1978). An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. *Scand. J. Stat.*, **5**,141-150.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer-Verlag, New York.

Benichou, J. and Gail, M. H. (1990). Estimates of absolute cause-specific risk in cohort studies. *Biometrics*, **46**, 813-826.

Crowder, M, J. (2001). *Classical Competing Risks*. Chapman & Hall/CRC, Boca Raton.

David, H. A. and Moeschbergcr, M. L. (1978). *Theory of Competing Risks*, Griffin, London.

Dinse, G. E. (1986). Nonparametric prevalence and mortality estimators for animal experiments with incomplete cause-of-death. *Amer. Stat. Assoc.*, **81**, 328-336.

Elandt-Johnson, R. C. and Johnson, N. L. (1980). *Survival Models and Data Analysis*. John Wiley & Sons, New York.

Fleming, T. R. (1978a). Nonparametric estimation for non-time-homogeneous Markov processes in the problem of competing risks. *Ann. Stat*., **6**, 1057-10.

Fleming, T. R. (1978b). Asymptotic distribution results in competing risks estimation. *Ann. Stat*., **6**,1071-1079.

Gaynor, J. J., Feuer, E. J., Tan, C. C., et al. (1993). On the use of cause-specific failure and conditional probabilities: Examples from clinical oncology data. J. *Amer. Stat. Assoc*., **88**, 400-409.

Kalbfleisch, J, D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, New York.

Kay, R. (1986). Treatment effects in Competing Risks analysis of prostate cancer data. *Biometrics*, **42**, 203-211.

Lawless J. F. (2003). *Statistical Models and Methods for lifetime data*. Second Edition New York.

Lunn, M. and McNeil, D. (1995). Applying Cox regression to competing risks. *Biometrics*, **51**, 524-532.

Manton, K. G. and Stallard, E. (1988). *Chronic Disease Modelling*. Griffin, London.

Matthews, D. E. (1988). Likelihood-based confidence intervals for functions of many parameters. *Biotnetrika*, **75**, 139-144.

Miller R. G. Jr (1981). Survival Analysis John Willey & sons inc. New York.

Munoz, A. (1980a). *Nonparametric estimation from censored bivariate observations. Technical Report No 60*. Division of Biostatistics, Stanford University, Stanford, California.

Munoz, A. (1980b). *Consistency of the self-consistent estimator of the distribution function from censored observations. Technical Report No 61*. Division of Biostatistics, Stanford University, Stanford, California.

Namboodiri, K. and Suchindran, C. M. (1987). *Life Table Techniques and Their Applications*. Academic Press, Orlando.

Nelson, W. B. (1969). Hazard plotting for incomplete failure data. *J.Quai. Technol*., **1**, 27-52.

Nelson, W. B. (1982). Applied Life Data Analysis. John Wiley & Sons, New York.

Seal, H. L. (1977). Studies in the history of probability and statistics XXXV. Multiple decrements or competing risks. *Biometrika*, **64**, 429-439.

Xian Liu (2012). Survival Analysis Models and Applications John Willey & sons inc. United Kingdom.

Link 1. https://vincentarelbundock.github.io/Rdatasets/articles/data.html