ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΤΜΗΜΑ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΠΡΟΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

“ΤΟ ΜΟΝΤΕΛΟ ΑΝΑΛΟΓΙΚΩΝ ΚΙΝΔΥΝΩΝ ΤΟΥ COX ΚΑΙ ΕΦΑΡΜΟΓΗ ΣΤΗΝ R ”

ΠΡΟΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΕΥΓΕΝΙΑ-ΣΟΦΙΑ Θ. ΖΑΡΟΓΙΑΝΝΗ

ΕΠΙΒΛΕΠΟΥΣΑ : ΦΙΛΙΑ ΒΟΝΤΑ

ΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΡΙΑ ΕΜΠ

ΑΘΗΝΑ 2013

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΤΜΗΜΑ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΠΡΟΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

“ ΤΟ ΜΟΝΤΕΛΟ ΑΝΑΛΟΓΙΚΩΝ ΚΙΝΔΥΝΩΝ ΤΟΥ COX ΚΑΙ ΕΦΑΡΜΟΓΗ ΣΤΗΝ R ”

ΠΡΟΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΕΥΓΕΝΙΑ-ΣΟΦΙΑ Θ. ΖΑΡΟΓΙΑΝΝΗ

Τριμελής εξεταστική επιτροπή

Φιλία Βόντα	Καρώνη-Ρίτσαρντσον Χρυσηίς	Κουκουβίνος Χρήστος
Επικ. Καθηγήτρια	Αναπλ. Καθηγήτρια	Καθηγητής
ΕΜΠ	ΕΜΠ	ΕΜΠ

ΑΘΗΝΑ, 2013

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω θερμά την καθηγήτριά μου κ. Φιλία Βόντα για την καθοδήγηση, την πολύτιμη βοήθεια, την στήριξη και την εμπιστοσύνη που μου έδειξε κατά την διάρκεια εκπόνησης της διπλωματικής αυτής. Επίσης θα ήθελα να ευχαριστήσω τα μέλη της επιτροπής μου για τα σχόλια και τις υποδείξεις τους τα οποία βελτίωσαν σε μεγάλο βαθμό την παρουσίαση της εργασίας. Θα ήθελα ακόμη να ευχαριστήσω την οικογένειά μου για τα εφόδια που μου προσέφεραν, την συμπαράσταση και την υπομονή τους.

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΛΗΨΗ

Το μοντέλο αναλογικών κινδύνων του Cox χρησιμοποιείται ευρέως στην ανάλυση επιβίωσης για την εύρεση της σχέσης μεταξύ μιας μεταβλητής που δηλώνει τον χρόνο επιβίωσης ενός ατόμου κι άλλων συμμεταβλητών καθώς και των διαφορών στην επιβίωση που οφείλονται στο είδος της θεραπείας και σε προγνωστικούς παράγοντες σε κλινικές μελέτες. Πρόκειται για ένα ημιπαραμετρικό μοντέλο, το οποίο μοντελοποιεί τη συνάρτηση κινδύνου σε σχέση με άλλες μεταβλητές.

Στα τρία πρώτα κεφάλαια αναπτύσσεται το μοντέλο αναλογικών κινδύνων του Cox, στο κεφάλαιο 4 μέθοδοι που αφορούν την ανάλυση επιβίωσης, ενώ στο τελευταίο κεφάλαιο χρησιμοποιούνται πραγματικά δεδομένα (που έχουν βρεθεί στις σημειώσεις Φωκιανός και Χαραλάμπους 2010), τα οποία αναλύονται με τη βοήθεια της R. Συγκεκριμένα, στο πρώτο κεφάλαιο γίνεται μια εισαγωγή στην ανάλυση επιβίωσης και εισάγονται οι έννοιες των λογοκριμένων και αποκομμένων δεδομένων. Το επόμενο κεφάλαιο, αναφέρεται στις συναρτήσεις του χρόνου επιβίωσης, στις σχέσεις μεταξύ τους και αναπτύσσεται το μοντέλο του Cox. Επιπλέον, δίνονται εκτιμήσεις των συναρτήσεων επιβίωσης και αναπτύσσονται επεκτάσεις του μοντέλου του Cox. Το τρίτο κεφάλαιο ασχολείται με τους ελέγχους υποθέσεων, με μεθόδους που εξετάζουν αν ισχύει η υπόθεση αναλογικότητας των κινδύνων, με τη θεωρία των υπολοίπων που χρησιμοποιούνται για ελέγχους που αφορούν την καταλληλότητα του μοντέλου και με μεθόδους για τη σύγκριση καμπυλών επιβίωσης. Στο τέταρτο, αναπτύσσονται μη παραμετρικές μέθοδοι εκτίμησης των συναρτήσεων επιβίωσης. Στο πέμπτο κεφάλαιο, αναλύονται πραγματικά δεδομένα με την βοήθεια του στατιστικού πακέτου της R, τα οποία προσαρμόζονται στο μοντέλο αναλογικών κινδύνων του Cox και εξετάζεται κατά πόσο είναι σημαντικές οι διάφορες μεταβλητές στην πρόβλεψη του χρόνου επιβίωσης.

ABSTRACT

The Cox proportional hazards model is used widely in survival analysis to find the relationship between a variable that indicates the survival time of a person and other covariates and the differences in survival due to the type of treatment and prognostic factors in clinical studies. It is a semiparametric model, which models the hazard function in relation to other variables.

In the first three chapters we examine the theory related to the Cox proportional hazards model, in chapter 4 we refer to methods that are useful in survival analysis, while in the last chapter we consider real data (that are available in the notes Fokianos and Chalambous (2010)), that are analyzed with the help of R. In particular, in the first chapter we give an introduction to survival analysis and to the notions of cencored and truncated data. The next chapter, refers to functions of the survival time, to relations between them and the Cox model is being developed. Moreover, estimates of the survival function are given and extensions of the Cox model are being developed. The third chapter deals with hypothesis testing, methods for examining if the proportionality assumption is valid, with residual theory that is being used for tests regarding the validity of the model and methods for comparing survival curves. In the fourth, non-parametric methods for estimating survival functions are presented. In the fifth chapter, real data are being analyzed with the help of the statistical package R. We examine whether the Cox proportional hazards model fits the data well and also identify which are the important variables in predicting survival time.

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

1.1. Ανάλυση επιβίωσης

Η ανάλυση δεδομένων διάρκειας ζωής είναι ο κλάδος της στατιστικής που ασχολείται με δεδομένα που αντιπροσωπεύουν χρόνο μέχρι την εμφάνιση ενός γεγονότος όπως η υποτροπή ή ο θάνατος ενός ασθενούς, ή γενικότερα η βλάβη ενός μηχανήματος, το κάψιμο μιας λάμπας ή ακόμα πιο γενικά μια απεργία ή μια απόλυση.

Όσον αφορά τα τεχνικά συστήματα είναι γνωστή ως θεωρία αξιοπιστίας (Reliability Theory), ενώ όσον αφορά τις βιοϊατρικές εφαρμογές ως ανάλυση επιβίωσης (Survival Analysis). Αρχικά οι αναλυτές ενδιαφέρονταν για το χρόνο που μεσολαβούσε μεταξύ μιας θεραπείας μέχρι τον θάνατο του ασθενούς και για αυτό το λόγο αυτή η περιοχή πήρε και το συγκεκριμένο όνομα.

Ο *χρόνος επιβίωσης ή χρόνος αποτυχίας* (survival time ή failure time) αναφέρεται σε μία θετική τυχαία μεταβλητή που μετράει το χρόνο (σε όποια μονάδα χρόνου είναι κατάλληλη) που μεσολαβεί από την στιγμή έναρξης παρακολούθησης ενός ατόμου ή τεχνικού συστήματος, μέχρι την στιγμή που το άτομο ή το τεχνικό σύστημα θα έρθει αντιμέτωπο με κάποιο γεγονός ενδιαφέροντος.

Οι χρόνοι επιβίωσης ακολουθούν μία κατανομή η οποία διαφέρει πολύ από την κανονική και δεν είναι λοιπόν συμμετρική. Πολλές από τις συνηθισμένες στατιστικές μεθόδους προϋποθέτουν η κατανομή της μεταβλητής που εξετάζουμε να είναι κανονική, έτσι δεν είναι κατάλληλες τέτοιες μέθοδοι για την ανάλυση δεδομένων επιβίωσης.

Η ιδιαιτερότητα που εμφανίζουν τα δεδομένα επιβίωσης και δεν επιτρέπει την χρήση συνηθισμένων στατιστικών τεχνικών, είναι ότι οι χρόνοι επιβίωσης ορισμένων παρατηρήσεων είναι είτε *λογοκριμένοι* (censored) είτε *αποκομμένοι* (truncated) ή και τα δύο ταυτόχρονα. Η λογοκρισία συνήθως συμβαίνει επειδή τα άτομα μπορεί να εισέρχονται στη μελέτη σε διαφορετικούς χρόνους, με αποτέλεσμα ο χρόνος παρακολούθησης μερικών ατόμων να μην είναι επαρκής ώστε να καταγραφεί ο χρόνος μέχρι την πραγματοποίηση του υπό μελέτη γεγονότος.

1.2. Λογοκριμένα δεδομένα (censored data)

Τα *λογοκριμένα δεδομένα* είναι αυτά για τα οποία δεν είναι γνωστός ο χρόνος επιβίωσης. Το μόνο που μπορεί να σημειωθεί είναι ότι ο χρόνος επιβίωσης είναι μεγαλύτερος από την τιμή που έχει καταγραφεί.

Τα δεδομένα αυτά εμφανίζονται σε περίπτωση που ο ασθενής είναι ακόμη ζωντανός στο τέλος της κλινικής μελέτης ή για διάφορους λόγους έχει απομακρυνθεί από αυτή.

Ο όρος censoring χρησιμοποιήθηκε για πρώτη φορά από τον Hald (1949). Τα δεδομένα που δεν είναι λογοκριμένα ονομάζονται μη-λογοκριμένα ή πλήρη. Για να γίνει κατανοητή η έννοια των λογοκριμένων παρατηρήσεων, δίνεται ένα παράδειγμα όπου υπάρχουν λογοκριμένες παρατηρήσεις.

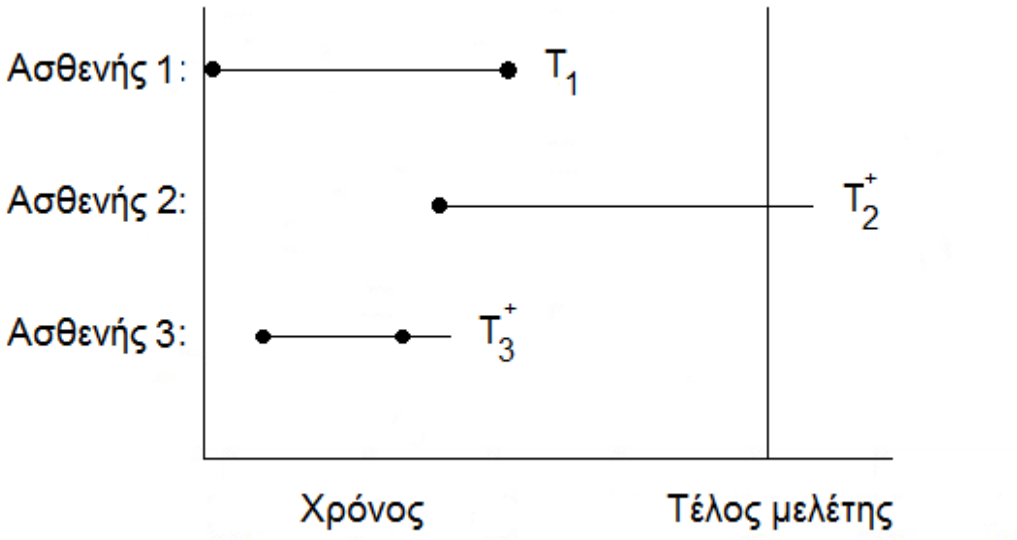
Παράδειγμα 1.1:

Σε μια έρευνα για την μελέτη της αποτελεσματικότητας μιας νέας θεραπείας για μια ασθένεια, η μεταβλητή που μας ενδιαφέρει είναι ο χρόνος που θα επιζήσει ο ασθενής, δηλαδή ο χρόνος επιβίωσης κάθε ατόμου. Οι ασθενείς μπορεί να εισέρχονται στη μελέτη σε διαφορετικούς χρόνους, ενώ η διάρκεια της μελέτης είναι προκαθορισμένη. Επομένως, για κάθε ασθενή καταγράφεται ο χρόνος από την είσοδό του στη μελέτη μέχρι τον θάνατό του. Στο τέλος της μελέτης είναι πιθανό να υπάρχουν ασθενείς που ζουν ακόμη ενώ θα υπάρχουν ασθενείς με τους οποίους χάθηκε η επαφή. Τον ακριβή χρόνο επιβίωσης των ατόμων αυτών δεν τον ξέρουμε, ξέρουμε όμως ότι είναι μεγαλύτερος από το χρόνο που παρεμβάλλεται από την είσοδό τους στη μελέτη μέχρι την ολοκλήρωση της μελέτης (στην πρώτη περίπτωση) και μέχρι την στιγμή που χάθηκε η επαφή (στην δεύτερη). Αυτές οι παρατηρήσεις είναι λογοκριμένες.

Σίγουρα, δεν πρέπει να αποκλείσουμε αυτά τα δεδομένα από την μελέτη θεωρώντας τα ως ελλιπή. Κάτι τέτοιο θα επηρέαζε τόσο την ανάλυση όσο και τα αποτελέσματα τα οποία δεν θα ήταν σωστά, αφού οι περισσότεροι από τους ασθενείς αυτούς έχουν ξεπεράσει τον χρόνο μελέτης και επομένως μας οδηγούν στο συμπέρασμα της αποτελεσματικότητας της θεραπείας.

Οπως ειπώθηκε και πιο πάνω οι λογοκριμένες παρατηρήσεις δεν προκύπτουν μόνο λόγω του χρόνου λήξης της μελέτης αλλά μπορεί να προκύψουν και όταν ο ασθενής χάνεται από την παρακολούθηση (loss to follow-up) (ο ασθενής μπορεί να αποφάσισε να μετακομίσει ή να αλλάξει γιατρό ή νοσοκομείο) ή αποσύρεται από την παρακολούθηση (drop-out) (η θεραπεία έχει πολύ κακές επιδράσεις και ο ασθενής είναι αναγκαίο να σταματήσει ή δεν θέλει να λάβει μέρος σε μια τέτοια διαδικασία μετά από ένα χρονικό διάστημα).

Στο παρακάτω σχήμα φαίνονται οι χρόνοι τριών ασθενών:



Σχήμα 1.1: Τυπική περίπτωση λογοκριμμένων παρατηρήσεων

- Ο ασθενής 1 εισέρχεται στην μελέτη τον χρόνο 0 και πεθαίνει στο T_1 , δίνοντας έτσι μια πλήρη παρατήρηση.
- Ο ασθενής 2 εισέρχεται αργότερα και μέχρι το τέλος της μελέτης παραμένει ζωντανός μέχρι το T_2 τουλάχιστον, δίνοντας έτσι μια λογοκριμένη παρατήρηση που συμβολίζεται με ${T_2}^+$.
- Ο ασθενής 3 εισέρχεται κι αυτός αργότερα αλλά χάνεται από την παρακολούθηση πριν από το τέλος της μελέτης, δίνοντας έτσι την λογοκριμένη παρατήρηση ${T_3}^+$.

1.2.1. Είδη λογοκριμένων παρατηρήσεων

Υπάρχουν τρία είδη λογοκρισίας. Η *λογοκρισία από δεξιά* (right censoring), η *λογοκρισία από αριστερά* (left censoring) και η *λογοκρισία κατά διάστημα* (interval censoring). Η *λογοκρισία από δεξιά* ή η *λογοκρισία από αριστερά* είναι ειδικές περιπτώσεις της λογοκρισίας κατά διάστημα. Επιπλέον η λογοκρισία χωρίζεται σε τρεις κατηγορίες, την *λογοκρισία τύπου I* (Type I censoring), την *λογοκρισία τύπου II* (Type II censoring) και την *τυχαία λογοκρισία* (random censoring).

Έστω ότι T είναι ο χρόνος επιβίωσης ή ο χρόνος αποτυχίας ενός ατόμου και c ο χρόνος στον οποίο σταματά η μελέτη.

- *Δεξιά λογοκρισία (right censoring)*: Στην περίπτωση αυτή ισχύει T>c. Δηλαδή ο χρόνος επιβίωσης του ατόμου είναι μεγαλύτερος από τον χρόνο τερματισμού της μελέτης. Ο ακριβής χρόνος επιβίωσης δεν είναι γνωστός είναι γνωστό μόνο ότι έχει ξεπεράσει το χρόνο τερματισμού της μελέτης. Στη στατιστική, μηχανική, την οικονομία, και την ιατρική έρευνα, λογοκρισία συμβαίνει όταν η μέτρηση ή παρατήρηση είναι γνωστή μόνο εν μέρει. Η δεξιά λογοκρισία είναι η πιο συνηθισμένη μορφή λογοκρισίας και εμφανίζεται στις περιπτώσεις όπου το άτομο αποσύρεται ή χάνεται από την μελέτη ή όταν τερματίζεται η μελέτη σε ένα προκαθορισμένο χρόνο.

Παράδειγμα 1.2:

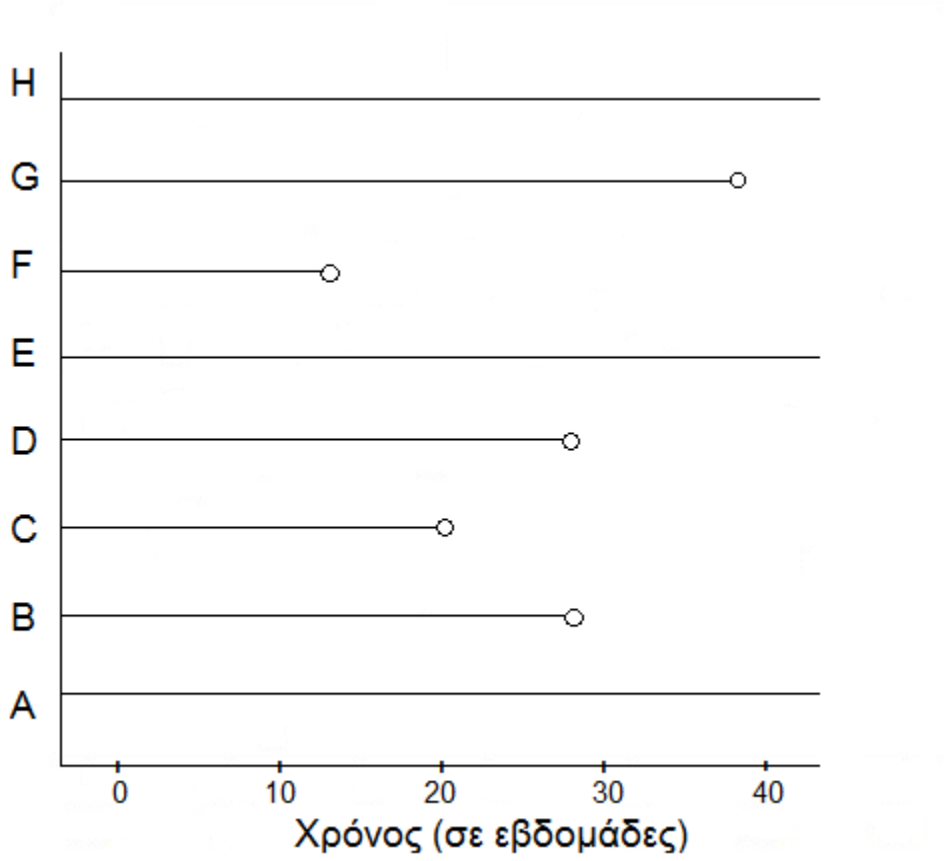
Ας υποθέσουμε ότι μια μελέτη διεξάγεται για τη μέτρηση του αντίκτυπου ενός φαρμάκου στη θνησιμότητα. Σε μια τέτοια μελέτη, μπορεί να είναι γνωστό ότι η ηλικία ενός ατόμου κατά το θάνατο είναι τουλάχιστον 75 έτη. Μια τέτοια κατάσταση μπορεί να προκύψει, εάν το άτομο αποσύρθηκε από τη μελέτη στην ηλικία των 75, ή εάν το άτομο εξακολουθεί να είναι ζωντανό στην ηλικία των 75 χρόνων.

(http://en.wikipedia.org/wiki/Censoring_%28statistics%29)

- *Λογοκρισία τύπου I (Type I censoring):* Όταν ο χρόνος διάρκειας της μελέτης είναι προκαθορισμένος από την αρχή, τότε έχουμε λογοκρισία τύπου I. Ο χρόνος c ονομάζεται χρόνος λογοκρισίας (censoring time). Καταγράφονται οι χρόνοι επιβίωσης ή αποτυχίας των ατόμων που απέτυχαν κατά την διάρκεια της μελέτης, ενώ για τα υπόλοιπα άτομα το μόνο που είναι γνωστό είναι ότι οι χρόνοι επιβίωσης είναι μεγαλύτεροι από το c.

Παράδειγμα 1.3:

Θεωρούμε 8 σωλήνες (A,B,C,D,E,F,G,H) που υποβάλλονται σε μια διαδικασία την ίδια χρονική στιγμή t=0 και καταγράφεται ο χρόνος αποτυχίας τους. Ο ερευνητής αποφασίζει να τερματίσει το πείραμα μετά από 40 εβδομάδες (c=40). Από το σχήμα 1.2, βλέπουμε ότι οι σωλήνες B, C, D, F και G καταστράφηκαν στους χρόνους 28, 20, 27, 14 και 38 αντίστοιχα (οι χρόνοι αυτοί είναι οι χρόνοι αποτυχίας), ενώ οι σωλήνες A, E και H λειτουργούσαν σε όλη τη διάρκεια της μελέτης, άρα οι χρόνοι επιβίωσης τους δεν είναι γνωστοί. Έτσι, τα δεδομένα επιβίωσης είναι 40+, 28, 20, 27, 40+, 14, 38 και 40+ εβδομάδες. Τα λογοκριμένα δεδομένα στην περίπτωση αυτή είναι τύπου I.



Σχήμα 1.2: Λογοκριμένες παρατηρήσεις τύπου I

- *Λογοκρισία τύπου II (Type II censoring)* : Στην λογοκρισία τύπου II, η μελέτη συνεχίζεται μέχρι να αποτύχουν k άτομα. Δηλαδή αν έχουμε n άτομα υπό μελέτη, τότε στο τέλος της γνωρίζουμε ότι απέτυχαν k άτομα, ενώ για τα υπόλοιπα n-k άτομα γνωρίζουμε μόνο ότι ο χρόνος επιβίωσής τους είναι μεγαλύτερος από τον χρόνο επιβίωσης των k ατόμων

που απέτυχαν. Πρέπει να σημειωθεί ότι ο αριθμός k προκαθορίζεται πριν από την έναρξη της μελέτης.

Για παράδειγμα, στο παράδειγμα με τους 8 σωλήνες, αν ο ερευνητής ήθελε να τερματίσει την έρευνα όταν 4 από τους σωλήνες καταστραφούν (δηλαδή k=4), τα δεδομένα που θα έπαιρνε θα ήταν: 27+,27+,20,27,27+,14,27+,27+.

- *Τυχαία λογοκρισία (random censoring):* Στην περίπτωση αυτή, ο χρόνος λογοκρισίας που αντιστοιχεί σε κάθε άτομο που είναι υπό παρακολούθηση δεν είναι σταθερός αλλά τυχαίος. Για παράδειγμα, σε κλινικές μελέτες ενώ οι χρονικές στιγμές έναρξης και λήξης είναι προκαθορισμένες, οι ασθενείς εισέρχονται σε διαφορετικές (τυχαίες) χρονικές στιγμές, με αποτέλεσμα οι χρόνοι λογοκρισίας να είναι τυχαίοι. Ή επίσης οι ασθενείς αποχωρούν, για διάφορους λόγους, σε άλλες χρονικές στιγμές ο καθένας δημιουργώντας μια τυχαία λογοκριμένη παρατήρηση.

- *Αριστερή λογοκρισία (left censoring):* Στις από αριστερά λογοκριμένες παρατηρήσεις το μόνο που γνωρίζουμε είναι πως ο χρόνος ενιαφέροντος T είναι μικρότερος από κάποια δεδομένη χρονική στιγμή, δηλαδή T<t_0, όπου t_0 θα μπορούσε να είναι η χρονική στιγμή της έναρξης της έρευνας. Δηλαδή ο χρόνος επιβίωσης είναι μικρότερος από ένα χρονικό διάστημα. Ο ακριβής χρόνος επιβίωσης δεν είναι γνωστός.

Παράδειγμα 1.4:

'Έστω ότι θέλουμε να μελετήσουμε τον χρόνο επιβίωσης σε σχέση με τον χρόνο της θεραπευτικής αγωγής που πήραν οι ασθενείς που έχουν μολυνθεί από κάποια ασθένεια. Στην ερώτηση "Πότε πήρες το φάρμακο για πρώτη φορά;", θα μπορούσαμε να πάρουμε τριών ειδών απαντήσεις:

1."Δεν θυμάμαι πότε ήταν η πρώτη φορά που πήρα το φάρμακο" (αριστερά λογοκριμένη παρατήρηση αφού ο ακριβής χρόνος θεραπείας δεν είναι γνωστός και είναι μικρότερος από την ηλικία του ατόμου).

2."Πήρα για πρώτη φορά το φάρμακο αυτό .... " (μη λογοκριμένη παρατήρηση).

3."Δεν πήρα ποτέ το συγκεκριμένο φάρμακο" (δεξιά λογοκριμένη παρατήρηση διότι μπορεί να αρχίσει το φάρμακο μετά το τέλος της μελέτης).

- *Λογοκρισία σε διάστημα (interval censoring):* Σε αυτή την περίπτωση λογοκρισίας οι παρατηρήσεις αποτελούν διαστήματα της μορφής $[t_1, t_2]$ αφού το μόνο που είναι γνωστό για τον χρόνο ενδιαφέροντος είναι ότι κυμαίνεται ανάμεσα στο t_1και στο t_2, δηλαδή t_1<T<t_2. Αυτού του είδους η λογοκρισία παρατηρείται όταν έχουμε περιοδική παρακολούθηση, δηλαδή το πείραμα δεν είναι υπό συνεχή επίβλεψη, ή όταν έχουμε ομαδοποιήσεις παρατηρήσεων.

Παράδειγμα 1.5:

Σε μια μελέτη, γυναίκες υποβάλλονται σε τεστ ΠΑΠ. Ο χρόνος εκδήλωσης ενδιαφέροντος κάθε ασθενούς είναι γνωστό ότι βρίσκεται σε ένα διάστημα (t_1,t_2], το οποίο αντιπροσωπεύει το χρόνο μεταξύ της πρώτης επίσκεψης και το χρόνο που εντοπίστηκε το γεγονός ενδιαφέροντος. Αν για μία ασθενή που εξετάζεται κάθε μήνα βρέθηκε τον i-μήνα που εξετάστηκε ότι εμφάνισε συμπτώματα, γνωρίζουμε μόνο ότι ο χρόνος αποτυχίας για την συγκεκριμένη ασθενή είναι μεταξύ του i-1 και i+1 μήνα, χωρίς να είναι γνωστός ο ακριβής χρόνος (Tableman & Kim (2004)).

1.3. Αποκομμένα δεδομένα (truncated data)

Μια άλλη μορφή δεδομένων που συνδέεται με την ανάλυση επβίωσης ή αξιοπιστίας αποτελούν τα αποκομμένα δεδομένα. Όσον αφορά τα δεδομένα αυτά δεν γνωρίζουμε τι συμβαίνει (δεν έχουμε καθόλου πληροφορίες) για ασθενείς για τους οποίους το γεγονός ενδιαφέροντος είναι εκτός ενός παραθύρου στο χρόνο στο οποίο λαμβάνει χώρα η μελέτη, ή εκτός δυνατότητας του οργάνου μετρήσεως αν πρόκειται για τεχνικά συστήματα. Πρέπει να τονιστεί πως η διαφορά μεταξύ αποκομμένων και λογοκριμένων δεδομένων είναι πως για τα τελευταία έχουμε τουλάχιστον ημιτελείς παρατηρήσεις για όλα τα αντικείμενα μελέτης ενώ για τα αποκομμένα δεδομένα υπάρχει ένα ποσοστό παρατηρήσεων που μας διαφεύγει εντελώς και δεν έχουμε πληροφορίες για αυτά.

1.3.1. Είδη αποκομμένων παρατηρήσεων

Υπάρχουν τρία είδη αποκομμένων παρατηρήσεων. Τα από δεξιά αποκομμένα δεδομένα (right truncated), τα από αριστερά αποκομμένα δεδομένα (left truncated) και τα αποκομμένα σε διάστημα δεδομένα (interval truncated).

- *Από δεξιά αποκομμένα δεδομένα (right truncated):* Η περίπτωση αυτή σχετίζεται με άτομα που έχουν χρόνο ενδιαφέροντος T μεγαλύτερο από κάποιο χρόνο t_0 (που θα μπορούσε π.χ. να είναι η λήξη της μελέτης) πέρα από τον οποίο η παρατήρηση του πειράματος δεν είναι δυνατή. Τα άτομα αυτά δεν περιλαμβάνονται καθόλου στο δείγμα γιατί η ύπαρξή τους δεν είναι γνωστή. Τα άτομα που περιλαμβάνονται στο δείγμα είναι υπό συνθήκη αυτά για τα οποία $T \le t_0$. Μια μελέτη της θνησιμότητας με βάση τα αρχεία θανάτου είναι ένα καλό παράδειγμα.

Παράδειγμα 1.6:

Έστω ότι μας ενδιαφέρει η μελέτη του χρόνου από τη μόλυνση HIV μέχρι την εμφάνιση AIDS. Μόνο τα άτομα που έχουν αναπτύξει AIDS πριν από το τέλος της μελέτης περιλαμβάνονται στη μελέτη. Τα μολυσμένα άτομα που δεν έχουν ακόμη αναπτύξει AIDS δεν περιλαμβάνονται στο δείγμα, ως εκ τούτου, παραμένουν άγνωστα στον ερευνητή (Klein & Moeschberger (1997)).

- *Από αριστερά αποκομμένα δεδομένα (left truncated):* Στην περίπτωση αυτή, επίσης θα υπάρξουν άτομα που απορρίφθηκαν από την μελέτη και ο ερευνητής δεν γνωρίζει τίποτα για την ύπαρξή τους. Κάτι τέτοιο μπορεί να συμβεί σε περίπτωση έκθεσης σε μια συγκεκριμένη νόσο, είσοδο σε έναν οίκο ευγηρίας ή εμφάνισης ενός ενδιάμεσου συμβάντος πριν από το θάνατο. Οι παρατηρήσεις αυτές ονομάζονται αποκομμένες από αριστερά (left truncated) . Πιο συγκεκριμένα, έστω T ο χρόνος επιβίωσης ή αποτυχίας και t_0 συγκεκριμένος χρόνος, ο οποίος θα μπορούσε να είναι η έναρξη της μελέτης, και πριν από τον οποίο η παρατήρηση του πειράματος δεν είναι δυνατή. Τότε, μόνο τα άτομα υπό την συνθήκη $T \ge t_0$ παρατηρούνται. Η πιο συνηθισμένη περίπτωση αριστερών αποκομμένων δεδομένων παρατηρείται όταν τα άτομα εισέρχονται στην μελέτη σε τυχαία ηλικία.

.

Παράδειγμα 1.7:

Έστω μια κοινότητα συνταξιοδότησης. Καταγράφονται η ηλικία των μελών που πεθαίνουν ή φεύγουν από το κέντρο και η ηλικία των μελών που εισέρχονται. Τα άτομα πρέπει να επιβιώσουν μέχρι μια ορισμένη ηλικία για να εισέλθουν στην κοινότητα. Άτομα που πεθαίνουν σε νεαρή ηλικία αποκλείονται από την μελέτη. Η διάρκεια ζωής αυτών των ανθρώπων χαρακτηρίζεται σαν από αριστερά αποκομμένη παρατήρηση (Klein & Moeschberger (1997)).

- *Αποκομμένα σε διάστημα δεδομένα (interval truncated):*

Στην περίπτωση αυτή μια παρατήρηση συμπεριλαμβάνεται στο δείγμα μόνο εάν πέσει μέσα σε ένα διάστημα στο χρόνο, ας πούμε Β, όπου η παρατήρηση του φαινομένου είναι δυνατή. Δηλαδή, ο χρόνος ενδιαφέροντος Τ παρατηρείται μόνο υπό την συνθήκη $T \in B$.

Τα δεδομένα θα μπορούσαν να είναι ταυτόχρονα λογοκριμένα και αποκομμένα. Οι δυσκολίες στην ανάλυση τέτοιων δεδομένων είναι πολλές. Η πιο γενική περίπτωση λογοκριμένων και αποκομμένων δεδομένων κατά διαστήματα έχει μελετηθεί μεταξύ άλλων από τους Turnbull (1974, 1976), Alioum and Commenges (1996) και Huber and Vonta (2004).

ΚΕΦΑΛΑΙΟ 2

ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ

2.1. Συναρτήσεις του χρόνου επιβίωσης

Η κατανομή των χρόνων επιβίωσης χαρακτηρίζεται από τρεις συναρτήσεις, τη συνάρτηση πυκνότητας πιθανότητας (probability density function), τη συνάρτηση επιβίωσης ή αξιοπιστίας (survival function) και τη συνάρτηση κινδύνου (hazard function). Στην πράξη, οι τρεις αυτές συναρτήσεις μπορούν να χρησιμοποιηθούν για να επεξηγήσουν διαφορετικά είδη δεδομένων.

Έστω Τ ο χρόνος επιβίωσης ο οποίος είναι συνεχής τ.μ.

I. *Συνάρτηση πυκνότητας πιθανότητας ή συνάρτηση πυκνότητας (Probability density function ή density function) :* Περιγράφει συνεχή τυχαία μεταβλητή και συμβολίζεται με f(t) και εκτιμάται ως το όριο της πιθανότητας ένα άτομο να αποτύχει σε ένα μικρό χρονικό διάστημα (t, t+Δt) ανά μονάδα πλάτους Δt, δηλαδή:

$$\hat{f}(t)= \frac{\alpha\rho.\text{ατόμων που αποτυγχάνουν σε διάστημα } (t,t+\Delta t)}{\Delta t} \qquad (2.1)$$

ή $$\hat{f}(t) = \lim_{dt\to 0} \frac{P(t<T<t+dt)}{dt} \qquad (2.2)$$

Η καμπύλη της *f(t)* ονομάζεται *καμπύλη πυκνότητας (density curve).* Η συνάρτηση πυκνότητας πιθανότητας του χρόνου Τ έχει τις ακόλουθες ιδιότητες:

- *f(t)*≥0 για t≥0 και *f(t)*=0 για t<0
- Το εμβαδόν μεταξύ της καμπύλης πιθανότητας και του άξονα του t ισούται με 1.

Όταν δεν υπάρχουν λογοκριμένες παρατηρήσεις, η συνάρτηση πυκνότητας εκτιμάται ως η αναλογία των ατόμων που αποτυγχάνουν σε ένα διάστημα ανά μονάδα πλάτους, δηλαδή:

$$\hat{f}(\text{t})= \frac{\alpha\rho.\ \text{ατόμων που αποτυγχάνουν στο διάστημα που ξεκινά στο χρόνο } t}{(\text{συνολικός αριθμός ατόμων})*(\text{πλάτος διαστήματος})} \qquad (2.3)$$

Η συνάρτηση κατανομής F(t) ορίζεται ως F(t) = P(T≤t) = $\int_0^t f(u)du$. Εξ ορισμού η F(t) είναι αύξουσα με $\lim_{t\to 0} F(t) = 0$, $\lim_{t\to\infty} F(t) = 1$.

II. *Συνάρτηση επιβίωσης ή αξιοπιστίας (Survival function) :* Η συνάρτηση αυτή συμβολίζεται με S(t) και ορίζεται ως η πιθανότητα το άτομο να επιβιώσει για χρόνο μεγαλύτερο του t:

$$S(t) = P(T>t) \qquad (2.4)$$

Επειδή F(t) = P(T≤t) η συνάρτηση αξιοπιστίας γράφεται και S(t)=1-F(t) = P(T>t) = $\int_t^{\infty} f(u)du$.

Η S(t) είναι μια φθίνουσα συνάρτηση του t με τις εξής ιδιότητες:

- S(t)=1 για t=0
- S(t)=0 για t=∞

Δηλαδή, η πιθανότητα το άτομο να επιβιώσει τουλάχιστον στο χρόνο 0 είναι 1 και η πιθανότητα επιβίωσης σε ένα άπειρο χρόνο είναι 0.

Η γραφική παράσταση της S(t) συναρτήσει του t ονομάζεται *καμπύλη επιβίωσης (survival curve).*

Όταν δεν υπάρχουν λογοκριμένες παρατηρήσεις, η συνάρτηση επιβίωσης εκτιμάται ως η αναλογία των ασθενών που επιβιώνουν για χρόνο μεγαλύτερο του t.

$$\hat{S}(t) = \frac{αρ.\ ατόμων\ που\ επιβιώνουν\ για\ χρόνο\ μεγαλύτερο\ του\ t}{συνολικός\ αριθμός\ ατόμων} \qquad (2.5)$$

Έστω ότι υπάρχουν n παρατηρήσεις. Τις ταξινομούμε σε αύξουσα σειρά και τις ονομάζουμε $t_{(1)}$, $t_{(2)}$, … , $t_{(k)}$. Θα έχουμε k διακεκριμένες παρατηρήσεις (παρατηρήσεις με τους ίδιους χρόνους θα έχουν το ίδιο σύμβολο) , τέτοιες ώστε $t_{(1)}<t_{(2)}<\ldots<t_{(k)}$. Για κάθε χρόνο επιβίωσης υπολογίζεται η αντίστοιχη συνάρτηση επιβίωσης.

Όταν υπάρχουν λογοκριμένες παρατηρήσεις, τότε ο αριθμητής δεν μπορεί πάντα να οριστεί. Για παράδειγμα, έστω ότι έχουμε τα δεδομένα 3, 4, 5, 6+, 6+, 7, 7, 8+, 10+ και θέλουμε να υπολογίσουμε το $\hat{S}(9)$. Δεν μπορεί όμως να υπολογιστεί από την σχέση (2.5), αφού δεν γνωρίζουμε τον ακριβή αριθμό ατόμων που επιβιώνουν σε χρόνο μεγαλύτερο του 9. Μπορεί ο τέταρτος, ο πέμπτος και ο όγδοος να επιβιώνουν για χρόνο μεγαλύτερο του 9 μπορεί και όχι. Το $\hat{S}(4)$ μπορεί να υπολογιστεί, $\hat{S}(4) = \frac{7}{9}$ εφόσον 7 από τα 9 άτομα έχουν χρόνους επιβίωσης μεγαλύτερους του 4.

Επομένως, όταν έχουμε λογοκριμένα δεδομένα χρησιμοποιούμε μη παραμετρικές μεθόδους για την εκτίμηση του S(t), όπως είναι η μέθοδος Kaplan-Meier ή η μέθοδος των πινάκων επιβίωσης (life tables) (Kaplan and Meier (1958)).

III. *<u>Συνάρτηση κινδύνου (hazard function) :</u>* Είναι η πιθανότητα ένα άτομο ηλικίας t να βιώσει το γεγονός στην αμέσως επόμενη χρονική στιγμή. Ορίζεται ως εξής:

$$h(t) = \lim_{\delta t \to 0} \frac{P(t<T\leq t+\delta t \setminus T>t)}{\delta t} \quad (2.6)$$

Από τις σχέσεις (2.1) και (2.6) παίρνουμε:

$$h(t) = \lim_{\delta t \to 0} \frac{P(t<T\leq t+\delta t\ )}{P(T>t)\delta t} = \frac{f(t)}{P(T>t)} = \frac{f(t)}{1-F(t)} \quad (2.7)$$

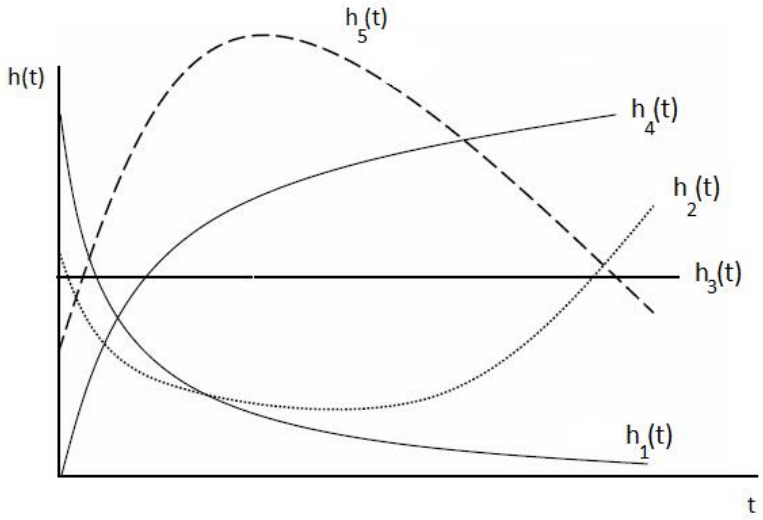
Η συνάρτηση κινδύνου είναι γνωστή και ως *στιγμιαίος ρυθμός αποτυχίας (instantaneous failure rate) ή ως δεσμευμένος ρυθμός θνησιμότητας (conditional mortality).* Εκφράζει το στιγμιαίο ρυθμό διακοπής και η h(t)δ(t) είναι η υπό συνθήκη πιθανότητα της επικείμενης διακοπής μιας μονάδας δοθέντος ότι επέζησε μέχρι την συγκεκριμένη χρονική στιγμή t.

Όταν δεν υπάρχουν λογοκριμένες παρατηρήσεις, η συνάρτηση κινδύνου εκτιμάται ως η αναλογία των ατόμων που αποτυγχάνουν σε ένα χρονικό διάστημα ανά μονάδα χρόνου, δεδομένου ότι επιβίωσαν μέχρι την αρχή του διαστήματος.

$$\hat{h}(t) = \frac{αρ.ατόμων\ που\ αποτυγχάνουν\ στο\ διάστημα\ που\ ξεκινά\ στο\ χρόνο\ t}{(αρ.ατόμων\ που\ είναι\ ζωντανά\ μέχρι\ το\ t)*(πλάτος\ χρ.διαστήματος)}$$

$$= \frac{αρ.ατόμων\ που\ αποτυγχάνουν\ ανά\ μονάδα\ χρόνου\ στο\ διάστημα\ που\ ξεκινά\ στο\ χρόνο\ t}{αρ.ατόμων\ που\ είναι\ ζωντανά\ μέχρι\ το\ t} \quad (2.8)$$

Η συμπεριφορά της *h(t)* ποικίλει. Μπορεί να αυξάνει, να μειώνεται, να μένει σταθερή ή να δηλώνει μια πιο περίπλοκη διαδικασία. Στο παρακάτω σχήμα φαίνονται διάφορες μορφές της συνάρτησης κινδύνου.



Σχήμα 2.1: Παραδείγματα συναρτήσεων κινδύνου

- Η συνάρτηση h_1(t) είναι μια φθίνουσα συνάρτηση και δεν εμφανίζεται συχνά στην πράξη. Δείχνει ότι σε αρχικούς χρόνους ο κίνδυνος είναι μεγάλος, ενώ όσο περνάει ο χρόνος ο κίνδυνος μειώνεται όπως π.χ. ο κίνδυνος μετά από μία εγχείρηση.
- Η συνάρτηση h_2(t) είναι γνωστή ως *λεκανοειδής καμπύλη* ή *«μπανιέρα» (bathtub curve)* λόγω της μορφής της και περιγράφει την εξέλιξη της ανθρώπινης ζωής. Στην διάρκεια μιας αρχικής περιόδου, ο κίνδυνος είναι μεγάλος (υψηλή βρεφική θνησιμότητα) , στη συνέχεια μέχρι μια συγκεκριμένη ηλικία, ο κίνδυνος παραμένει σταθερός ενώ σε μεγαλύτερες ηλικίες αυξάνεται ακόμη περισσότερο.
- Η συνάρτηση h_3(t) είναι μια σταθερή συνάρτηση κινδύνου, δηλαδή ο κίνδυνος παραμένει σταθερός. Για παράδειγμα, αυτό συμβαίνει όταν θέλουμε να εξετάσουμε τον κίνδυνο θανάτου υγιών ατόμων ηλικίας 18-40, των οποίων οι κύριες αιτίες θανάτου είναι τα ατυχήματα.
- Η συνάρτηση h_4(t) είναι μια αύξουσα συνάρτηση η οποία συναντάται συχνά. Με την πάροδο του χρόνου ο κίνδυνος αυξάνεται. Για παράδειγμα, ασθενείς με οξεία λευχαιμία έχουν έναν αυξανόμενο κίνδυνο θανάτου με την πάροδο του χρόνου.
- Η συνάρτηση h_5(t) αρχικά αυξάνεται και έπειτα μειώνεται. Αυτό συμβαίνει π.χ. στους ασθενείς με φυματίωση όπου αρχικά ο κίνδυνος αυξάνεται ενώ μετά από κάποια θεραπεία μειώνεται.

Τέλος, ορίζεται και η *αθροιστική ή σωρευτική συνάρτηση κινδύνου (cumulative hazard function)* ως:

$$H(t) = \int_0^t h(u)du \qquad (2.9)$$

Η συνάρτηση αυτή, είναι χρήσιμη για την επιλογή ενός κατάλληλου στατιστικού μοντέλου κατά την ανάλυση ενός συνόλου δεδομένων.
Οι περισσότερες πληροφορίες σχετικά με τα παραπάνω προέρχονται από D.Collett (2003), το άρθρο των Kaplan-Meier (1958) και την ιστοσελίδα http://esperia.iesl.forth.gr/~kafesaki/Applied-Mathematics/probabilities/p3.pdf.

2.2. Σχέσεις μεταξύ των συναρτήσεων του χρόνου επιβίωσης

Οι παραπάνω συναρτήσεις επιβίωσης είναι μαθηματικά ισοδύναμες. Από την σχέση (2.7) έχουμε:

$$h(t) = \frac{f(t)}{S(t)} \qquad (2.10)$$

Επειδή η συνάρτηση πυκνότητας πιθανότητας οποιασδήποτε κατανομής είναι ίση με την παράγωγο της συνάρτησης προκύπτει:

$$f(t) = \frac{d}{dt}(F(t)) = \frac{d}{dt}[1-S(t)] = -S'(t) \qquad (2.11)$$

Από τις (2.7) και (2.11) προκύπτει:

$$h(t) = \frac{-S'(t)}{S(t)} = -\frac{d}{dt}\ln S(t) \qquad (2.12)$$

Ολοκληρώνοντας την παραπάνω σχέση και χρησιμοποιώντας ότι $S(0)=1$ έχουμε:

$$\int_0^t h(u)du = -[\ln S(t) - \ln S(0)]$$

Συνδυάζοντάς την παραπάνω με την (2.7) έχουμε:

$$H(t) = -\ln S(t) \quad \text{ή} \quad S(t) = \exp[-H(t)] \qquad (2.13)$$

Από τις (2.7) και (2.13) τέλος προκύπτει:

$$f(t) = h(t)\exp[-H(t)] \qquad (2.14)$$

2.3. Το μοντέλο αναλογικών κινδύνων του Cox

2.3.1. Εισαγωγή στο μοντέλο αναλογικών κινδύνων του Cox

Το μοντέλο αναλογικών κινδύνων του Cox (Proportional Hazards Cox model ή PH μοντέλο του Cox), παρουσιάστηκε από τον Cox το 1972. Το μοντέλο αυτό, όπως και όλα τα μοντέλα αναλογικών κινδύνων μοντελοποιούν την συνάρτηση κινδύνου *h(t)*. Χρησιμοποιείται εκτενώς σήμερα στην ανάλυση λογοκριμένων δεδομένων επιβίωσης που αφορούν βιοϊατρικές εφαρμογές, για την εξακρίβωση των διαφορών στην επιβίωση που οφείλονται στο είδος της θεραπείας και σε προγνωστικούς παράγοντες σε κλινικές δοκιμές. Είναι επίσης μια καλή στατιστική τεχνική για την εύρεση της σχέσης μεταξύ της επιβίωσης ενός ασθενή και αρκετών επεξηγηματικών μεταβλητών. Ακόμη, μας επιτρέπει να εκτιμήσουμε τον κίνδυνο θανάτου ενός ατόμου, ή άλλου γεγονότος που μας ενδιαφέρει δεδομένων των προγνωστικών τους μεταβλητών (Cox (1972)).

2.3.2. Το μοντέλο του Cox

Έστω ότι έχουμε n άτομα στην μελέτη και ότι το $\boldsymbol{x}'=(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_p)$ είναι το διάνυσμα των μεταβλητών που πιστεύουμε ότι επηρεάζουν το χρόνο ζωής των ατόμων. Οι μεταβλητές αυτές μπορεί να παριστάνουν διάφορα χαρακτηριστικά όπως θεραπείες, φυσικές ιδιότητες των ατόμων (όπως ηλικία ή φύλο), εξωγενείς μεταβλητές. Το $\mathbf{x}_i'=(\mathbf{x}_{i1},\mathbf{x}_{i2},\ldots,\mathbf{x}_{ip})$, i=1,2,…,n είναι το διάνυσμα με τις τιμές των συμμεταβλητών που αντιστοιχεί στο i άτομο.

Οι μεταβλητές μπορούν να συνδυαστούν για να εξηγήσουν επιδράσεις αλληλεπίδρασης. Οι επεξηγηματικές μεταβλητές μπορούν να ταξινομηθούν ως σταθερές (ανεξάρτητες του χρόνου) ή εξαρτημένες από τον χρόνο.

Αρχικά υποθέτουμε ότι οι συμμεταβλητές δεν εξαρτώνται από τον χρόνο, δηλαδή θεωρούμε ότι οι τιμές των συμμεταβλητών $\mathbf{x}_i$ καταγράφηκαν στην αρχή της μελέτης, δηλαδή στο t=0, και θεωρούμε ότι παραμένουν σταθερές σε όλη την διάρκεια της μελέτης.

Το μοντέλο παλινδρόμησης του Cox δίνεται από τον τύπο:

$$h(t;\mathbf{x}) = h_0(t)e^{\boldsymbol{\beta}'\mathbf{x}} \qquad (2.15)$$

όπου η *h(t;***x**) είναι η συνάρτηση κινδύνου στο χρόνο t δεδομένων των συμμεταβλητών ενώ η $h_0(t)$ ονομάζεται *αναφορική ή βασική συνάρτηση κινδύνου (baseline hazard function)* στο χρόνο t. Το $\boldsymbol{\beta}'=(\beta_1,\beta_2,\ldots,\beta_p)$ είναι το διάνυσμα των *συντελεστών παλινδρόμησης.*

Ο κίνδυνος h(t;**x**) εξαρτάται από τον χρόνο και τις συμμεταβλητές, αλλά μέσω δύο διαφορετικών παραγόντων. Ο πρώτος παράγοντας, $h_0(t)$, είναι μια συνάρτηση του χρόνου, που αφήνεται ελεύθερη και θεωρείται η ίδια

και για τα n άτομα της μελέτης. Ο δεύτερος παράγοντας είναι μια ποσότητα που εξαρτάται από τις συμμεταβλητές μόνο μέσω του διανύσματος **β**.

Το κύριο χαρακτηριστικό του μοντέλου του Cox είναι ότι οι παραμετρικές μορφές των βασικών συναρτήσεων $h_0(t)$ και $S_0(t)$ δεν καθορίζονται. Μόνο η επίδραση των συμμεταβλητών **x** αναλύεται. Υποθέτει ότι οι επιδράσεις των μεταβλητών είναι σταθερές στο χρόνο και είναι προσθετικές σε μια συγκεκριμένη κλίμακα. Για το λόγο αυτό, το μοντέλο του Cox καλείται ημιπαραμετρικό (semiparametric) (Tableman-Kim (2004)).

Από την εξίσωση (2.15) για **x**=0 παρατηρούμε ότι προκύπτει:

$$h(t;\mathbf{0}) = h_0(t) \qquad (2.16)$$

Δηλαδή, η αναφορική συνάρτηση κινδύνου μπορεί να θεωρηθεί ως η συνάρτηση κινδύνου ενός ατόμου με τιμή όλων των συμμεταβλητών ίση με 0, $\mathbf{x}_i$=**0**, i=1,…,p.

Για να δούμε πως οι μεταβλητές είναι προσθετικές σε μια συγκεκριμένη κλίμακα, θεωρούμε για δύο οποιαδήποτε άτομα με διανύσματα μεταβλητών $\mathbf{x}_1$ και $\boldsymbol{x}_2$, το λόγο κινδύνου *H(R(t))* (hazard rate), δηλαδή το λόγο:

$$H(R(t)) = \frac{h(t;\boldsymbol{x}_1)}{h(t;\mathbf{x}_2)} = \frac{h_0(t)e^{\boldsymbol{\beta}'\mathbf{x}_1}}{h_0(t)e^{\boldsymbol{\beta}'\mathbf{x}_2}} = e^{\boldsymbol{\beta}'(\boldsymbol{x}_1-\mathbf{x}_2)} \qquad (2.17)$$

(Tableman-Kim, (2004))

Θεωρούμε πως οι μεταβλητές $\boldsymbol{x}_i$ δεν εξαρτώνται από τον χρόνο, έτσι και η ποσότητα $e^{\boldsymbol{\beta}'(\boldsymbol{x}_1-\mathbf{x}_2)}$ της σχέσης (2.17) είναι σταθερή στο χρόνο, για αυτό και το μοντέλο είναι γνωστό ως *μοντέλο αναλογικών κινδύνων.*

Η γενική μορφή ενός μοντέλου αναλογικών κινδύνων (proportional hazards model) είναι:

$$h(t;\boldsymbol{x}) = h_0(t)g(\boldsymbol{x}) \qquad (2.18)$$

όπου *g(**x**)* είναι μια συνάρτηση του διανύσματος **x**.

Παρατηρήσεις:

- Ο όρος *αναλογικοί κίνδυνοι (proportional hazards)* προέρχεται από το γεγονός ότι δύο οποιαδήποτε άτομα έχουν συναρτήσεις κινδύνου που η μία είναι πολλαπλάσιο της άλλης.
- Στην περίπτωση του PH μοντέλου του Cox, η g(**x**) είναι η συνάρτηση $e^{\boldsymbol{\beta}'\boldsymbol{x}}$=$e^{\beta_1 x_1}$+$e^{\beta_2 x_2}$+…+$e^{\beta_p x_p}$
- Όταν θεωρούμε κάποια κατανομή για το $h_0(t)$, τότε έχουμε την παραμετρική μορφή του μοντέλου αναλογικού κινδύνου.

Λογαριθμίζοντας την σχέση (2.17) προκύπτει:

$$\ln[h(t;\boldsymbol{x}_1)] - \ln[h(t;\boldsymbol{x}_2)] = \boldsymbol{\beta}'(\boldsymbol{x}_1 - \boldsymbol{x}_2) \qquad (2.19)$$

Η παραπάνω σχέση δείχνει ότι το μοντέλο θεωρεί μια σταθερή διαφορά μεταξύ των λογαρίθμων των κινδύνων δύο ατόμων.

Από την (2.15) προκύπτει η παρακάτω σχέση για την συνάρτηση επιβίωσης στο χρόνο t, και από αυτή μπορεί να εκτιμηθεί η συνάρτηση επιβίωσης οποιουδήποτε ατόμου που συμμετέχει στην μελέτη:

$$\int_0^t h_i\,(s)ds = \int_0^t h_0(s)e^{\boldsymbol{\beta}'\mathbf{x}}ds$$

Χρησιμοποιώντας την (2.9) προκύπτει:

H_i(t) = $e^{\boldsymbol{\beta}'\mathbf{x}}$ H_0(t) ή $e^{-H_i(\mathrm{t})} = e^{-e^{\boldsymbol{\beta}'\mathbf{x}}H_0(\mathrm{t})} = e^{-H_0(\mathrm{t})e^{\boldsymbol{\beta}'\mathbf{x}}}$

ή από την (2.13) έχουμε:

$$S(t) = [S_0(t)]^{e^{\boldsymbol{\beta}'x}} \qquad (2.20)$$

όπου $S_0(t)=exp[-H_0(t)]$ είναι η *αναφορική ή βασική συνάρτηση επιβίωσης (baseline survival function).*

Έστω ότι έχουμε μόνο μία μεταβλητή, την *X,* που αντιπροσωπεύει το είδος της θεραπείας και έστω ότι παίρνει την τιμή 1 (x_1=1) αν το άτομο λαμβάνει την θεραπεία Α και 0 (x_2=0) αν λαμβάνει την θεραπεία Β. Τότε, η συνάρτηση κινδύνου για τα άτομα που ανήκουν στην ομάδα Α είναι $h(t,1)= h_0(t)e^{\beta}$, ενώ για τα άτομα της ομάδας Β θα είναι $h(t,0)= h_0(t)e^0= h_0(t).$ Σε αυτή την περίπτωση ο λόγος κινδύνου θα είναι $HR = \frac{h(t,1)}{h(t,0)} = e^{\beta}$ ενώ η (2.20) θα γίνει:

$$S_1(t) = [S_0(t)]^{e^{\boldsymbol{\beta}}} \qquad (2.21)$$

- Αν β>0 τότε e^{β}>1 , ο κίνδυνος ενός ατόμου που λαμβάνει την θεραπεία Α θα είναι μεγαλύτερος από τον κίνδυνο ενός ατόμου που λαμβάνει την θεραπεία Β, ενώ η πιθανότητα επιβίωσης ενός ατόμου της ομάδας Α θα είναι μικρότερη από την πιθανότητα επιβίωσης ενός ατόμου της ομάδας Β. Αυτό προκύπτει από την σχέση (2.21)

$$S_0(t) < 1 \text{ ή } S_1(t) = [S_0(t)]^{e^{\beta}} < S_0(t)$$

- Αν β=0 τότε $h(t,1)= h(t,0)$ και $S_1(t) =S_0(t)$, δηλαδή οι δύο θεραπείες θεωρούνται ισοδύναμες.
- Αν β<0 τότε 0<e^{β}<1 . Σε αυτή την περίπτωση, ο κίνδυνος ενός ατόμου που λαμβάνει την θεραπεία Α θα είναι μικρότερος από τον κίνδυνο του ατόμου που λαμβάνει την θεραπεία Β, ενώ η πιθανότητα επιβίωσης ενός ατόμου της ομάδας Α θα είναι μεγαλύτερη από την πιθανότητα επιβίωσης ενός ατόμου της ομάδας Β.

Οι περισσότερες πληροφορίες σχετικά με τα παραπάνω προέρχονται από τις ιστοσελίδες http://www.demog.berkeley.edu/213/Week14/welcome.pdf και http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-cox-regression.pdf.

2.3.3. Εκτιμητής μεγίστης μερικής πιθανοφάνειας

Αφού ορίστηκε το μοντέλο αναλογικών κινδύνων του Cox, το επόμενο βήμα είναι η εκτίμηση των συντελεστών **β** του μοντέλου. Επειδή η αναφορική συνάρτηση κινδύνου $h_0(t)$ δεν καθορίζεται παραμετρικά, το **β** θα εκτιμηθεί με βάση την πληροφορία που προκύπτει από τα παρατηρούμενα δεδομένα, χωρίς να χρειάζεται να εμπλακεί η $h_0(t)$.

Θεωρούμε ένα σύνολο N ατόμων και υποθέτουμε ότι εμφανίζονται k πλήρεις διακεκριμένοι χρόνοι. Έστω $t_{(1)}, t_{(2)}, \dots, t_{(k)}$ οι k ταξινομημένοι πλήρεις χρόνοι και $R(t_{(i)})$ ή R_i το σύνολο των ατόμων που βρίσκονται σε κίνδυνο τη χρονική στιγμή $t_{(i)}$. Συμβολίζουμε με $x_{(i)}=(x_{(i)1}, x_{(i)2}, \dots, x_{(i)p})$, $1 \le i \le k$ το διάνυσμα των συμμεταβλητών που αντιστοιχεί στο άτομο με πλήρη χρόνο ζωής $t_{(i)}$, $1 \le i \le k$.

Για την εκτίμηση του **β** στο μοντέλο (2.15) ο Cox εισήγαγε την παρακάτω συνάρτηση:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{k} \frac{e^{\boldsymbol{\beta}' \mathbf{x}_{(j)}}}{\sum_{i \in R_j} e^{\boldsymbol{\beta}' \mathbf{x}_i}} \qquad (2.22)$$

την οποία ονόμασε *μερική πιθανοφάνεια (partial likelihood)* (Cox 1972).

Η συνάρτηση αυτή μπορεί να χρησιμοποιηθεί όταν δεν υπάρχουν ισότιμες παρατηρήσεις (ties) στα δεδομένα, δηλαδή αν κάθε πλήρης χρόνος εμφανίζεται μόνο μια φορά. Αν στα δεδομένα υπάρχουν ισότιμες παρατηρήσεις, τότε η συνάρτηση μερικής πιθανοφάνειας επιδέχεται διάφορες παραλλαγές για να αντιμετωπιστούν οι ισότητες.

Ο Cox απέδειξε ότι η συνάρτηση $L(\boldsymbol{\beta})$ επιτρέπει την καλή εκτίμηση των συντελεστών **β**. Ο εκτιμητής $\hat{\boldsymbol{\beta}}$ του **β** που προκύπτει είναι αμερόληπτος και ασυμπτωτικά κανονικός.

Οι συντελεστές **β** εκτιμώνται από τις τιμές $\hat{\boldsymbol{\beta}}$ που μεγιστοποιούν τη μερική πιθανοφάνεια $L(\boldsymbol{\beta})$ ή ισοδύναμα το λογάριθμό της. Ο λογάριθμος της μερικής πιθανοφάνειας (log-partial likelihood) δίνεται από τη σχέση:

$$l(\boldsymbol{\beta}) = \ln(L(\boldsymbol{\beta})) = \sum_{j=1}^{k} \boldsymbol{\beta}' \mathbf{x}_{(j)} - \ln\left(\sum_{i \in R_j} e^{\boldsymbol{\beta}' \mathbf{x}_i}\right) \qquad (2.23)$$

Ο εκτιμητής μέγιστης μερικής πιθανοφάνειας (ΕΜΠ) $\widehat{\boldsymbol{\beta}}$, βρίσκεται από την λύση του συστήματος των εξισώσεων που προκύπτουν από τη σχέση:

$$U_s(\boldsymbol{\beta}) = \frac{d\, l(\boldsymbol{\beta})}{d\beta_s} = 0, \quad s=1,2,\ldots,p \tag{2.24}$$

Προκύπτει η παρακάτω σχέση, που είναι ένα σύστημα p εξισώσεων:

$$U_s(\boldsymbol{\beta}) = \sum_{j=1}^{k} \left(x_{(j)s} - \frac{\sum_{i \epsilon R_j} x_{is} e^{\boldsymbol{\beta}' \mathbf{x}_i}}{\sum_{i \epsilon R_j} e^{\boldsymbol{\beta}' \mathbf{x}_i}} \right) = 0 \tag{2.25}$$

Το σύστημα αυτό για να λυθεί χρειάζονται επαναληπτικές διαδικασίες, όπως η μέθοδος Newton-Raphson.

Ο πίνακας πληροφορίας $I(\boldsymbol{\beta})$ είναι ένας pxp πίνακας:

$$I(\boldsymbol{\beta}) = \begin{pmatrix} -\frac{d^2 l(\boldsymbol{\beta})}{d^2 \beta_1} & \cdots & -\frac{d^2 l(\boldsymbol{\beta})}{d\beta_1 d\beta_p} \\ \vdots & \ddots & \vdots \\ -\frac{d^2 l(\boldsymbol{\beta})}{d\beta_p d\beta_1} & \cdots & -\frac{d^2 l(\boldsymbol{\beta})}{d^2 \beta_p} \end{pmatrix}$$

Οι έλεγχοι υποθέσεων, τα διαστήματα εμπιστοσύνης και ο εκτιμητής μέγιστης μερικής πιθανοφάνειας $\widehat{\boldsymbol{\beta}}$ βασίζονται στην ασυμπτωτική τυπική κανονική κατανομή:

$$\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, I(\widehat{\boldsymbol{\beta}})^{-1})$$

(http://www.ida.liu.se/~kawah/Cox2.pdf, Tableman-Kim (2004), σελίδες 64-65, D. Collett (2003))

2.3.4. Εκτίμηση της αναφορικής συνάρτησης κινδύνου $\mathbf{h_0(t)}$

Ο Breslow εισήγαγε μια εκτίμηση για την αναφορική συνάρτηση κινδύνου, αφού θεώρησε ότι η κατανομή του χρόνου αποτυχίας έχει μια συνάρτηση κινδύνου που είναι σταθερή ανάμεσα σε κάθε ζεύγος διαδοχικών παρατηρούμενων χρόνων αποτυχίας. Η εκτιμώμενη αναφορική συνάρτηση κινδύνου δίνεται από την παρακάτω σχέση:

$$\hat{h}_0(t_{(i)})=\frac{d_i}{\sum_{j\in R_i} e^{\hat{\beta}x_j}} \qquad (2.26)$$

όπου:

d_i: το πλήθος των αποτυχιών στο χρόνο $t_{(i)}$ (Breslow (1974), (1975), Breslow and Crowley (1974)).

Η αθροιστική αναφορική συνάρτηση κινδύνου και η εκτιμώμενη αναφορική συνάρτηση επιβίωσης δίνονται από τις παρακάτω σχέσεις αντίστοιχα:

$$\hat{H}_0(t)=\sum_{i:t_{(i)}\le t}\hat{h}_0(t)=\sum_{i:t_{(i)}\le t}\frac{d_i}{\sum_{j\in R_i} e^{\hat{\beta}x_j}} \qquad (2.27)$$

$$\hat{S}_0(t)=\exp[-\hat{H}_0(t)]=\prod_{i:t_{(i)}\le t}\exp\frac{-d_j}{\sum_{j\in R_i} e^{\hat{\beta}x_j}} \qquad (2.28)$$

Για τον υπολογισμό των παραπάνω ποσοτήτων χρειάζεται η τιμή του διανύσματος **β**, για την οποία χρησιμοποιείται ο εκτιμητής μέγιστης μερικής πιθανοφάνειας του **β**, $\hat{\boldsymbol{\beta}}$ που προκύπτει από τη μεγιστοποίηση της συνάρτησης μερικής πιθανοφάνειας (2.22).

2.3.5. ΄Υπαρξη ισότιμων παρατηρήσεων και μερική πιθανοφάνεια

Όταν δεν υπάρχουν ισότιμες παρατηρήσεις (ties) στους χρόνους επιβίωσης η συνάρτηση μερικής πιθανοφάνειας υπολογίζεται από την σχέση (2.22). Σε αντίθετη περίπτωση, χρησιμοποιούνται παραλλαγές της μεθόδου μερικής πιθανοφάνειας.

Ορίζουμε με $\mathbf{z}_m$ το διάνυσμα των συμμεταβλητών του m ατόμου. Έστω $t_{(1)}, t_{(2)}, \dots, t_{(k)}$ οι k διακεκριμένοι, ταξινομημένοι πλήρεις χρόνοι. Με d_i ορίζουμε το πλήθος των αποτυχιών στο χρόνο $t_{(i)}$ και με D_i το σύνολο των ατόμων που αποτυγχάνουν στο χρόνο $t_{(i)}$. Το διάνυσμα s_i είναι το άθροισμα των διανυσμάτων $\mathbf{z}_m$ των ατόμων που αποτυγχάνουν στο χρόνο $t_{(i)}$, δηλαδή

$s_i = \sum_{m \in D_i} \mathbf{z}_m$. Με R_i ορίζεται το σύνολο των ατόμων που βρίσκονται σε κίνδυνο τη χρονική στιγμή $t_{(i)}$.

2.3.5.1. Πιθανοφάνεια του Breslow

Η πιθανοφάνεια του Breslow είναι ο απλούστερος τύπος υπολογισμού της μερικής πιθανοφάνειας. Η λύση αν και είναι η λιγότερο ακριβής, είναι γρήγορη. Δίνεται από την σχέση:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{k} \frac{e^{\boldsymbol{\beta}' s_j}}{\left(\sum_{m \in R_j} e^{\boldsymbol{\beta}' \mathbf{z}_m}\right)^{d_j}} \quad (2.29)$$

2.3.5.2. Πιθανοφάνεια του Efron

Η πιθανοφάνεια του Efron αν και φαίνεται δύσκολο να υπολογιστεί, είναι εύκολο να προγραμματιστεί όσο η πιθανοφάνεια του Breslow. Η μέθοδος αυτή, είναι αρκετά ακριβής, εφόσον η αναλογία των ισοπαλιών σε σχέση με το μέγεθος του συνόλου κινδύνου R_i είναι πολύ μικρή. Δίνεται από την σχέση:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{k} \frac{e^{\boldsymbol{\beta}' s_j}}{\prod_{i=1}^{d_j} \left[\sum_{m \in R_j} e^{\boldsymbol{\beta}' \mathbf{z}_m} - \frac{i-1}{d_j} \sum_{m \in D_j} e^{\boldsymbol{\beta}' \mathbf{z}_m}\right]} \quad (2.30)$$

2.3.5.3. Διακριτή πιθανοφάνεια

Για τον ορισμό της διακριτής πιθανοφάνειας ο Cox υπέθεσε ότι τα δεδομένα προέρχονται από μια διακριτή κατανομή χρόνου ζωής. Η μέθοδος περιλαμβάνει την καταμέτρηση των πιθανών συνόλων κινδύνου R_i, σε κάθε ισόπαλο πλήρη χρόνο και χρειάζεται αρκετό χρόνο υπολογισμού σε περίπτωση που ένας πλήρης χρόνος επαναλαμβάνεται πολλές φορές. Δίνεται από την σχέση:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{d_j} \frac{e^{\boldsymbol{\beta}' s_j}}{\sum_{q \in Q_j} e^{\boldsymbol{\beta}' s_q^*}} \quad (2.31)$$

όπου Q_j είναι το σύνολο όλων των υποσυνόλων του συνόλου κινδύνου R_j μεγέθους d_j.

Το **q**=(q_1,q_2,…,q_{d_j}) είναι το διάνυσμα των ατόμων που βρίσκονται σε κίνδυνο την χρονική στιγμή $t_{(j)}$. Το s_q^* είναι το άθροισμα των διανυσμάτων $\mathbf{z}_i$ που αντιστοιχούν στα άτομα q_i, 1≤i≤ d_j, δηλαδή $s_q^*=\sum_{l=1}^{d_j} z_{ql}$.

Όταν δεν υπάρχουν ισότιμες παρατηρήσεις στους χρόνους επιβίωσης, οι τρεις προσεγγίσεις για την εύρεση της πιθανοφάνειας είναι ίδιες. Όταν οι ισότιμες παρατηρήσεις είναι πολύ λίγες, οι τιμές των τριών προσεγγίσεων θα είναι πολύ κοντινές.

Οι περισσότερες πληροφορίες σχετικά με την πιθανοφάνεια του Breslow, Efron και την διακριτή πιθανοφάνεια προέρχονται από τους Klein & Moeschberger (2003), την ιστοσελίδα http://www.stat.cmu.edu/~acthomas/724/Efron-Morris.pdf, D.W. Hosmer & S. Lemeshow (1998) και M. Stevenson (2009).

2.3.6. Εφαρμογές στο μοντέλο του Cox

2.3.6.1. Προσαρμογή του μοντέλου του Cox σε συνεχείς μεταβλητές

Εκτός από τις κατηγορικές μεταβλητές, το μοντέλο του Cox προσαρμόζεται και σε συνεχείς μεταβλητές. Αν x μια συνεχής μεταβλητή, θεωρώντας ότι το PH μοντέλο είναι κατάλληλο, η συνάρτηση κινδύνου δίνεται από την σχέση:

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

Αν οι τιμές της μεταβλητής x δύο ατόμων είναι x_1 και x_2 αντίστοιχα, ο λόγος κινδύνου είναι:

$$H(R(t)) = \frac{h(t,x_1)}{h(t,x_2)} = \frac{h_0(t)e^{\beta' x_1}}{h_0(t)e^{\beta' x_2}} = e^{\beta'(x_1-x_2)}$$

Αν το i άτομο έχει τιμή x_l και το j άτομο x_{l-1}, τότε ο κίνδυνος θανάτου του i ατόμου είναι e^{β} φορές μεγαλύτερος από τον κίνδυνο θανάτου του j ατόμου, αν β>0, ενώ αν β<0 το j άτομο έχει e^{β} φορές μεγαλύτερο κίνδυνο θανάτου από το i άτομο.

2.3.6.2. Αλληλεπίδραση μεταβλητών

Όταν η συνάρτηση κινδύνου επηρεάζεται από μία μεταβλητή και η επίδραση αυτή είναι διαφορετική στα διαφορετικά επίπεδα μιας άλλης μεταβλητής, τότε λέμε ότι έχουμε αλληλεπίδραση μεταξύ δύο παραγόντων. Όταν θέλουμε να ελέγξουμε την αλληλεπίδραση μεταξύ δύο παραγόντων Α και Β με ν και κ επίπεδα αντίστοιχα, τότε εισάγουμε στο μοντέλο (ν-1)(κ-1) μεταβλητές αλληλεπίδρασης.

2.3.7. Επεκτάσεις του μοντέλου του Cox

Το μοντέλο αναλογικών κινδύνων του Cox μπορεί να χρησιμοποιηθεί και σε περιπτώσεις όπου οι μεταβλητές παρουσιάζουν κάποια χαρακτηριστικά, διαφορετικά από αυτά που ισχύουν στο μοντέλο του Cox. Στις εξαρτώμενες από το χρόνο μεταβλητές, για παράδειγμα, δεν είναι όλες οι μεταβλητές σταθερές, αλλά η τιμή τους μεταβάλλεται με το χρόνο. Επιπλέον, στην περίπτωση των στρωματοποιημένων μεταβλητών, η υπόθεση της αναλογικότητας των κινδύνων δεν ισχύει απαραίτητα. Έτσι, όταν πρέπει να εξεταστούν τέτοιες μεταβλητές, το μοντέλο του Cox γενικεύεται και τροποποιείται κατάλληλα ώστε να μπορεί να αντιμετωπίσει αυτές τις περιπτώσεις.

2.3.7.1. Στρωματοποιημένη ανάλυση (Stratified analysis)

Όταν μια μεταβλητή έχει επίπεδα που δημιουργούν συναρτήσεις κινδύνου που δεν ικανοποιούν την υπόθεση αναλογικότητας τότε εφαρμόζουμε την στρωματοποίηση ως προς αυτή τη μεταβλητή. Με αυτό τον τρόπο προκύπτει το στρωματοποιημένο μοντέλο του Cox (Stratified Cox model). Η επέκταση αυτή επιτρέπει στην συνάρτηση κινδύνου να διαφέρει ανάμεσα στα επίπεδα της στρωματοποιημένης μεταβλητής. Η μεταβλητή, εκτός από κατηγορική, μπορεί να είναι αποτέλεσμα χωρισμού μιας ποσοτικής μεταβλητής σε ομάδες.

Η συνάρτηση κινδύνου ενός ατόμου που ανήκει στο i στρώμα με διάνυσμα μεταβλητών **x** δίνεται από τη σχέση:

$$h_{\mathrm{i}}(\mathrm{t},\mathrm{x}) = h_{0\mathrm{i}}(\mathrm{t}) e^{\beta' \mathrm{x}} \text{ , } \mathrm{i}=1,\ldots,\mathrm{I} \qquad (2.32)$$

όπου:

i: είναι το στρώμα του παράγοντα

I: το πλήθος των επιπέδων του παράγοντα

$h_{0\mathrm{i}}(t)$: η αναφορική συνάρτηση κινδύνου στο i επίπεδο

Από την παραπάνω σχέση φαίνεται ότι τα άτομα που ανήκουν στο ίδιο στρώμα, έχουν τις ίδιες αναφορικές συναρτήσεις κινδύνου και ανάλογες τις συναρτήσεις κινδύνου. Για παράδειγμα, για δύο άτομα που ανήκουν στο στρώμα i, i=1,…,I με μεταβλητές $\mathbf{x}_1$ και $\mathbf{x}_2$ ισχύει:

$$\frac{h_{\mathrm{i}}(t,\mathbf{x}_1)}{h_{\mathrm{i}}(t,\mathbf{x}_2)}=\frac{h_{0\mathrm{i}}(t)e^{\beta'\mathbf{x}_1}}{h_{0\mathrm{i}}(t)e^{\beta'\mathbf{x}_2}}=e^{\beta'(\mathbf{x}_1-\mathbf{x}_2)} \qquad (2.33)$$

Αντίθετα, τα άτομα που ανήκουν σε διαφορετικά στρώματα δεν έχουν ούτε τις ίδιες αναφορικές συναρτήσεις κινδύνου ούτε ανάλογες τις συναρτήσεις κινδύνου. Αυτό επειδή, οι αναφορικές συναρτήσεις κινδύνου $h_{01}(t)\, h_{02}(t),\ldots, h_{0\mathrm{I}}(t)$ κάθε στρώματος είναι αυθαίρετες συναρτήσεις του χρόνου και αφήνονται ασυσχέτιστες.

Ακόμη, από τη σχέση (2.32) φαίνεται ότι οι συντελεστές **β** είναι ίδιοι σε κάθε στρώμα. Στην περίπτωση που ήταν διαφορετικοί, τα δεδομένα κάθε στρώματος θα θεωρούνταν διαφορετικά σύνολα δεδομένων και θα αναλύονταν ξεχωριστά. Η εκτίμηση των συντελεστών $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_{\mathrm{p}})$ υπολογίζεται από την μεγιστοποίηση της συνάρτησης μερικής πιθανοφάνειας και δίνεται από την παρακάτω σχέση:

$$L(\boldsymbol{\beta}) = \prod_{\mathrm{i}=1}^{\mathrm{I}} L_{\mathrm{i}}\,(\boldsymbol{\beta}) \qquad (2.34)$$

Κάθε παράγοντας $L_{\mathrm{i}}(\boldsymbol{\beta})$ είναι η μερική πιθανοφάνεια η οποία προκύπτει από τη σχέση (2.22) για το στρώμα i και υπολογίζεται σε κάθε διακεκριμένο χρόνο αποτυχίας που παρατηρείται στο συγκεκριμένο στρώμα. Η μόνη διαφορά της παραπάνω σχέσης με την αντίστοιχη της μερικής πιθανοφάνειας για το μοντέλο του Cox είναι ότι έχει προστεθεί ο δείκτης i για να τονίσει ότι τα δεδομένα προέρχονται από το στρώμα i, i=1,…,I (Hosmer & Lemeshow (1998), Therneau & Grambsch (2000)).

2.3.7.2. Μεταβλητές εξαρτώμενες από το χρόνο

Μέχρι τώρα θεωρούσαμε ότι οι τιμές όλων των μεταβλητών $\mathrm{x_i}$ που μελετάμε σε ένα μοντέλο αναλογικών κινδύνων του Cox δεν αλλάζουν κατά την διάρκεια του χρόνου παρακολούθησής τους. Υπάρχουν όμως περιπτώσεις στις οποίες η συνάρτηση κινδύνου φαίνεται να εξαρτάται από την τιμή μιας μεταβλητής, η οποία αλλάζει με την πάροδο του χρόνου. Το μοντέλο του Cox μπορεί να επεκταθεί έτσι ώστε να ενσωματώνει τέτοιες μεταβλητές. Ο πιο συνηθισμένος τύπος εξαρτημένης από το χρόνο μεταβλητής είναι μια επαλαμβανόμενη μέτρηση σε ένα άτομο ή μια αλλαγή στη θεραπεία ενός ατόμου.

Οι εξαρτημένες από το χρόνο μεταβλητές χωρίζονται σε δύο κατηγορίες, στις *εξωτερικές μεταβλητές* (external covariates) και στις *εσωτερικές μεταβλητές* (internal covariates). Οι τιμές των εξωτερικών μεταβλητών δεν επηρεάζονται από την πορεία ζωής του ατόμου μέσα στη μελέτη, αλλά από ένα μηχανισμό που είναι εξωτερικός του ατόμου. Η ηλικία ενός ατόμου θεωρείται συνήθως ως σταθερή μεταβλητή σε μελέτες των οποίων η διάρκεια είναι μικρή. Όταν η διάρκεια είναι μεγάλη, τότε απαιτείται ο ορισμός της ηλικίας ως εξαρτημένης μεταβλητής από το χρόνο.

Η ηλικία θεωρείται εξωτερική μεταβλητή, η οποία μεταβάλλεται με τρόπο προβλέψιμο. Ένα άλλο παράδειγμα εξωτερικής μεταβλητής είναι ο τρόπος που χορηγείται κάποιο φάρμακο σε έναν ασθενή, ο οποίος καθορίζεται από την αρχή της μελέτης και μεταβάλλεται κατά την διάρκεια της με προκαθορισμένο τρόπο.

Οι τιμές μιας εσωτερικής μεταβλητής, σε αντίθεση με τις εξωτερικές μεταβλητές, επηρεάζονται από την πορεία ζωής του ατόμου μέσα στη μελέτη και για αυτό το λόγο το άτομο πρέπει να είναι εν ζωή σε όλη τη διάρκεια της μελέτης. Παραδείγματα εσωτερικών μεταβλητών είναι οι κλινικές ενδείξεις ενός ασθενή όπως η πίεση αίματος που μπορεί να καθοριστεί στην αρχή της μελέτης, αλλά υπάρχει η πιθανότητα να αλλάξει αργότερα, το μέγεθος του όγκου που μπορεί να επιδεινώνεται με το χρόνο, το βάρος του ατόμου, η μόλυνση ενός ασθενή (Collett (2003)).

2.3.7.3. Το γενικευμένο μοντέλο του Cox

Θεωρούμε το διάνυσμα μεταβλητών x(t)=($x_1(t)$, $x_2(t)$,…, $x_p(t)$) στη χρονική στιγμή t. Ο συμβολισμός αυτός ισχύει και για τις σταθερές μεταβλητές όπου t=0, δηλαδή για αυτές που η τιμή τους μετριέται μόνο μια φορά την t=0. Σε περίπτωση που έχουμε τουλάχιστον μια μεταβλητή που εξαρτάται από το χρόνο, το μοντέλο αναγογικών κινδύνων του Cox από τη σχέση (2.15) γενικεύεται και παίρνει την παρακάτω μορφή:

$$h(t,\mathbf{x}(t))=h_0(t)e^{\beta'\mathbf{x}(t)} \qquad (2.35)$$

όπου $\mathbf{x}(t)$ μπορεί να περιλαμβάνει εξαρτώμενες από το χρόνο αλλά και σταθερές μεταβλητές.

Η αναφορική συνάρτηση κινδύνου $h_0(t)$ είναι η συνάρτηση κινδύνου ενός ατόμου για το οποίο η τιμή του διανύσματος $\mathbf{x}(t)$ είναι 0.

Ο λόγος κινδύνου όταν υπάρχουν εξαρτημένες από το χρόνο μεταβλητές, για δύο άτομα με διανύσματα $x_1(t)$ και $x_2(t)$ αντίστοιχα είναι:

$$H(R(t)) = \frac{h(t;\boldsymbol{x}_1(\mathrm{t}))}{h(t;\mathbf{x}_2(\mathrm{t}))} = \frac{h_0(t)e^{\boldsymbol{\beta}'\mathbf{x}_1(t)}}{h_0(t)e^{\boldsymbol{\beta}'\mathbf{x}_2(t)}} = e^{\boldsymbol{\beta}'(\boldsymbol{x}_1(t)-\mathbf{x}_2(t))}$$

Παρατηρούμε ότι ο λόγος κινδύνου στην περίπτωση εξαρτημένων από το χρόνο μεταβλητών δεν είναι σταθερός σε κάθε χρονική στιγμή t αλλά μεταβάλλεται με το χρόνο.

Όταν έχουμε μόνο μια εξαρτημένη από το χρόνο μεταβλητή της μορφής της σχέσης (2.35), τότε το μοντέλο του Cox είναι:

$h(t, x_i(t)=0)=h_0(t)$ και $h(t, x_i(t)=1)=h_0(t)e^{\beta}$

(Collett (2003))

ΚΕΦΑΛΑΙΟ 3

ΕΛΕΓΧΟΙ

3.1 Έλεγχοι υποθέσεων

3.1.1. Έλεγχοι λόγου πιθανοφάνειας (Likelihood ratio tests)

Για τον έλεγχο της μηδενικής υπόθεσης H_0 : $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ μέσω του λόγου πιθανοφάνειας, θεωρούμε την στατιστική συνάρτηση:

$$L(\boldsymbol{\beta}_0) = -2\ln\frac{L(\boldsymbol{\beta}_0)}{L(\hat{\boldsymbol{\beta}})} = 2l(\hat{\beta}) - 2l(\boldsymbol{\beta}_0) \qquad (3.1)$$

που ακολουθεί ασυμπτωτικά την χ^2 κατανομή με p βαθμούς ελευθερίας.

(http://www.stern.nyu.edu/rengle/LagrangeMultipliersHandbook_of_Econ__II__Engle.pdf και Vuong (1989)).

3.1.2. Έλεγχοι Wald

Για τον έλεγχο της μηδενικής υπόθεσης H_0 : $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ το Wald τέστ βασίζεται στη στατιστική συνάρτηση:

$$W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' \, I(\hat{\boldsymbol{\beta}}) \, (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \qquad (3.2)$$

που ακολουθεί προσεγγιστικά την χ^2 κατανομή με p βαθμούς ελευθερίας

(http://www.public.iastate.edu/~vardeman/stat543/Handouts/wald-score-lrt.pdf

http://www.stern.nyu.edu/rengle/LagrangeMultipliersHandbook_of_Econ__II__Engle.pdf)

3.1.3. Score tests

Κάτω από ορισμένες συνθήκες κανονικότητας, το score από μόνο του έχει μια ασυμπτωτικά κανονική κατανομή με μέση τιμή 0 και πίνακα διασπορών-συνδιασπορών ίσο με τον πίνακα πληροφορίας, δηλαδή:

$$U(\boldsymbol{\beta}) \sim N_p(0, I(\boldsymbol{\beta}))$$

Για τον έλεγχο της μηδενικής υπόθεσης H_0 : $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ η τετραγωνική μορφή:

$$Q = U'(\boldsymbol{\beta}_0) I^{-1}(\boldsymbol{\beta_0}) \, U(\boldsymbol{\beta_0}) \qquad (3.3)$$

ακολουθεί προσεγγιστικά την χ^2 κατανομή με p βαθμούς ελευθερίας.

Οι τρεις αυτοί έλεγχοι είναι ασυμπτωτικά ισοδύναμοι, αλλά σε μικρά δείγματα μπορεί να διαφέρουν. Σε αυτή την περίπτωση, ο έλεγχος λόγου πιθανοφάνειας θεωρείται ο πιο αξιόπιστος, ενώ ο έλεγχος του Wald θεωρείται ο λιγότερο αξιόπιστος έλεγχος. Για μικρότερες τιμές του $\widehat{\boldsymbol{\beta}}$, οι τρεις έλεγχοι είναι σχεδόν οι ίδιοι, ενώ για μεγαλύτερες τιμές του $\widehat{\boldsymbol{\beta}}$ η διαφορά μεταξύ των ελέγχων μεγαλώνει.

(http://en.wikipedia.org/wiki/Score_test)

3.1.4. Κριτήρια επιλογής μοντέλων

Πολύ συχνά και σε πολλούς τομείς (βιομηχανία, ιατρική, οικονομία, κτλ) παρουσιάζεται η ανάγκη να εξηγήσουμε ένα φαινόμενο. Το φαινόμενο αυτό μπορεί να εξαρτάται από πάρα πολλές παραμέτρους οι οποίες να συνδυάζονται σ' ένα πολύπλοκο μοντέλο που δεν είναι καθόλου εύκολο να το επιβεβαιώσουμε από τις παρατηρήσεις που έχουμε.

Αναγκαστικά τις περισσότερες φορές υποθέτουμε πιθανά μοντέλα και προσπαθούμε να επιλέξουμε το καλύτερο από αυτά. Πολλά και διάφορα κριτήρια έχουν μέχρι σήμερα προταθεί τα οποία προσπαθούν να επιλέξουν το μοντέλο που εξηγεί καλύτερα τις παρατηρήσεις μας. Τα κριτήρια αυτά (μερικά από τα οποία θα αναφέρουμε στη συνέχεια) προκύπτουν ακολουθώντας διάφορες μεθόδους και προσπαθώντας να αντιμετωπίσουν διαφορετικές συνθήκες και δεδομένα οπότε τα αποτελέσματα που μας δίνουν σε συγκεκριμένα προβλήματα, δεν είναι πάντοτε τα ίδια.

Θα κάνουμε αναφορά στο κριτήριο AIC (Akaike, 1973) και στο κριτήριο BIC (Schwarz, 1978) διότι είναι εκπρόσωποι δυο κατηγοριών κριτηρίων επιλογής μοντέλων με τις ιδιότητες της ασυμπτωτικής αποδοτικότητας και της συνέπειας αντιστοίχως. Στη στατιστική ανάλυση που θα κάνουμε στο κεφάλαιο 5 θα κάνουμε χρήση μόνο του κριτηρίου AIC.

Το κριτήριο AIC ορίζεται ως:

$$AIC(\kappa) = -2\ell\left(\hat{\vartheta}_{\kappa}\right) + 2\kappa$$

=-2(μέγιστη λογαριθμική πιθανοφάνεια του μοντέλου)
+2(αριθμός των ελεύθερων παραμέτρων)

Ο λόγος για τον οποίο υπάρχει ο παράγοντας -2 που πολλαπλασιάζεται με την αμερόληπτη εκτιμήτρια $\ell\left(\hat{\vartheta}_{\kappa}\right)-\kappa$ είναι καθαρά ιστορικός. Το κριτήριο AIC(κ) επιλέγει το μοντέλο που ελαχιστοποιεί το κριτήριο.

Πρέπει να σημειώσουμε ότι η μεγιστοποίηση του λογάριθμου πιθανοφάνειας ως προς το πλήθος των ελεύθερων παραμέτρων σ' ένα μοντέλο θα οδηγούσε πάντα σε επιλογή του μοντέλου με τις περισσότερες παραμέτρους άρα προφανώς δεν μπορεί να αποτελεί σωστό κριτήριο.

Πρέπει επίσης να σημειωθεί ότι η αμεροληψία της εκτιμήτριας $\ell\left(\hat{\vartheta}_{\kappa}\right)-\kappa$ και άρα κατά αντίστοιχο τρόπο και του κριτηρίου είναι ασυμπτωτική ενώ για μικρά δείγματα έχουμε αρκετή μεροληψία.

Το κριτήριο BIC ορίζεται ως

$BIC=-2\ell\left(\hat{\vartheta}_{\kappa}\right)+\kappa\log n$, όπου κ η διάσταση του μοντέλου.

Το κριτήριο BIC(κ) επιλέγει το μοντέλο που ελαχιστοποιεί το κριτήριο.

Πρέπει να σημειωθεί ότι το κριτήριο BIC(κ) διαφέρει από το κριτήριο AIC (κ) μόνο στο δεύτερο όρο που στην περίπτωση του BIC εξαρτάται από το μέγεθος του δείγματος. Είναι φανερό ότι όσο το μέγεθος του δείγματος αυξάνει το κριτήριο BIC(κ) ευνοεί την επιλογή μοντέλων με λιγότερες παραμέτρους σε σύγκριση με το AIC(κ).

Τα δυο κριτήρια είναι φανερό ότι προκύπτουν με βάση εντελώς διαφορετική προσέγγιση του προβλήματος. Το κριτήριο AIC(κ) δέχεται ότι τα μοντέλα με τα οποία προσπαθούμε να εκτιμήσουμε, να προσεγγίσουμε το πραγματικό μοντέλο δεν είναι κατ' ανάγκη ακριβώς το ίδια με το πραγματικό και έτσι προσπαθεί να επιλέξει το μοντέλο το οποίο κατά κάποιο τρόπο «ταιριάζει» καλύτερα, κατά μέσο όρο. Η προσέγγιση με βάση την μέθοδο Bayes απ' όπου προκύπτει το BIC(κ) θεωρεί το κάθε μοντέλο ως πιθανό να είναι το πραγματικό και μετά για κάθε μοντέλο εκτιμά την πιθανότητα να είναι αυτό το πραγματικό.

Είναι φανερό ότι τα δύο κριτήρια μπορεί να οδηγήσουν σε διαφορετικά αποτελέσματα ανάλογα με το πρόβλημα επιλογής μοντέλων που έχουμε να αντιμετωπίσουμε, με το κριτήριο BIC(κ) να αντιμετωπίζει ίσως σοβαρότερα προβλήματα αν τα μοντέλα με τα οποία προσπαθούμε να προσεγγίσουμε το πραγματικό από το οποίο προκύπτουν τα δεδομένα μας δεν είναι αρκετά παρόμοια με το πραγματικό.

Άλλα κριτήρια που αναφέρονται στη βιβλιογραφία είναι το κριτήριο $FPE_a(\kappa)$ το κριτήριο C_κ και το κριτήριο GIC.

Το κριτήριο $FPE_a(\kappa)$ (Final Prediction Error) (Akaike, 1969) ορίζεται ως

$$FPE_a(\kappa) = n\hat{\sigma}_\kappa^2 + \alpha.\kappa.\frac{n\hat{\sigma}_m^2}{n-m}, a \in R$$

όπου $\hat{\sigma}_\kappa^2$ η ΕΜΠ του σ^2 στο κάθε μοντέλο με κ-παραμέτρους στο οποίο υπολογίζεται το κριτήριο και $\hat{\sigma}_m^2$ η ΕΜΠ του σ^2 του μοντέλου με όλες τις m – παραμέτρους.

Το κριτήριο C_κ (Mallows, 1973) ορίζεται ως

$$C_\kappa = \frac{n\hat{\sigma}_\kappa^2}{\hat{\sigma}_m^2} + 2\kappa$$

όπου $\hat{\sigma}_m^2$ η αμερόληπτη εκτιμήτρια του σ^2 στο μοντέλο με όλες τις m – παραμέτρους.

Το κριτήριο GIC (p) (General Information Criterion)

$$GIC(\kappa) = n\log\hat{\sigma}^2(\kappa) + a_n\kappa \qquad (3)$$

Το κριτήριο αυτό προτάθηκε από τον Atkison (1978) ως επέκταση του κριτηρίου AIC (κ). Τα δύο προηγούμενα κριτήρια είναι σχεδόν ισοδύναμα με το κριτήριο AIC(κ).

3.1.5. Διαστήματα εμπιστοσύνης

Τα διαστήματα εμπιστοσύνης δημιουργούνται συνήθως βάση του στατιστικού ελέγχου Wald. Το διάστημα εμπιστοσύνης του $e^{\widehat{\boldsymbol{\beta}}}$ είναι το:

$$(\exp(\widehat{\boldsymbol{\beta}} - z_{\alpha/2} se(\widehat{\boldsymbol{\beta}})), \exp(\widehat{\boldsymbol{\beta}} + z_{\alpha/2} se(\widehat{\boldsymbol{\beta}})) \quad (3.5)$$

Συνήθως μας ενδιαφέρει να δούμε αν ο αριθμός 1 ανήκει ή όχι στο διάστημα εμπιστοσύνης διότι αυτό αντιστοιχεί σε β ίσο με το 0, το οποίο αντίστοιχα παραπέμπει σε ισότητα των καμπύλων επιβίωσης.

(Collett 2003, Therneau & Grambsch 2000)

3.2. Έλεγχοι της υπόθεσης αναλογικότητας των κινδύνων

Για να είναι σωστή η διαδικασία και να έχουν νόημα τα αποτελέσματα που προκύπτουν, πρέπει πριν από την προσαρμογή του μοντέλου αναλογικών κινδύνων του Cox, να ελέγχουμε αν η υπόθεση της αναλογικότητας των κινδύνων ισχύει. Ο έλεγχος αυτός μπορεί να γίνει είτε γραφικά, είτε με διάφορα στατιστικά που υπάρχουν για τον έλεγχο αυτό. Το δεύτερο βήμα στην προσαρμογή του μοντέλου είναι, στην περίπτωση που υπάρχουν πολλές μεταβλητές, να ορίζουμε ένα βασικό πλάνο για την επιλογή των μεταβλητών που θα συμπεριληφθούν στο μοντέλο. Αφού βρούμε τις κατάλληλες μεταβλητές για το μοντέλο και προσαρμόσουμε στα δεδομένα αυτών το μοντέλο παλινδρόμησης του Cox, πρέπει στη συνέχεια να εξετάσουμε κατά πόσο είναι ικανοποιητικό ή αν θέλει βελτίωση.

Στην περίπτωση που καταλήξουμε πως η υπόθεση των αναλογικών κινδύνων δεν ισχύει, τότε είτε κάνουμε μετασχηματισμούς των δεδομένων έτσι ώστε να ικανοποιείται η υπόθεση της αναλογικότητας είτε θα πρέπει να επιλέξουμε μία εναλλακτική κλάση μοντέλων που να είναι πιο κατάλληλη για τα δεδομένα μας.

3.2.1. Γραφικές μέθοδοι για τον έλεγχο της αναλογικότητας των κινδύνων

Σύμφωνα με την υπόθεση αναλογικών κινδύνων, για τη συνάρτηση επιβίωσης ενός ατόμου με διάνυσμα συμμεταβλητών $\mathbf{x}=(\mathbf{x}_1, \mathbf{x}_2,\ldots, \mathbf{x}_p)$ ισχύει η (2.20):

$$S(t) = [S_0(t)]^{e^{\boldsymbol{\beta}'\mathbf{x}}}$$

από την οποία προκύπτουν οι παρακάτω σχέσεις:

$$-\log(S(t,x))=e^{\boldsymbol{\beta}'\boldsymbol{x}}[-\log S_0(t)] \quad (3.6)$$

$$\log[-\log(S(t,x))]=\boldsymbol{\beta}'\mathbf{x}+\log[-\log(S_0(t))] \quad (3.7)$$

Θεωρούμε $\mathbf{x}_1$ και $\mathbf{x}_2$ τα διανύσματα των συμμεταβλητών δύο ατόμων. Κάτω από την υπόθεση της αναλογικότητας, οι συναρτήσεις $\log[-\log(S(t,\boldsymbol{x}_1))]$ και $\log[-\log(S(t,\boldsymbol{x}_2))]$ θα έχουν μια σταθερή απόσταση, $\boldsymbol{\beta}'\mathbf{x}_1$ και $\boldsymbol{\beta}'\mathbf{x}_2$ αντίστοιχα, από τον αναφορικό αθροιστικό κίνδυνο $\log[-\log(S_0(t))]$. Για τις δύο αυτές συναρτήσεις ισχύει σύμφωνα με την (3.7):

$$\log[-\log(S(t,\boldsymbol{x}_1))]=\log[-\log(S(t,\boldsymbol{x}_2))]+\boldsymbol{\beta}'(\mathbf{x}_1-\mathbf{x}_2) \quad (3.8)$$

Επομένως, αν σχεδιάσουμε τις γραφικές παραστάσεις των $\log[-\log(S(t,x))]$ συναρτήσει του χρόνου, οι δύο καμπύλες που θα προκύψουν θα είναι παράλληλες και θα απέχουν μεταξύ τους σταθερή απόσταση ίση με $\boldsymbol{\beta}'(\mathbf{x}_1-\mathbf{x}_2)$. Έτσι, ένας πρώτος έλεγχος για τον έλεγχο της αναλογικότητας των κινδύνων, είναι αυτή η γραφική παράσταση. Αν οι καμπύλες που θα προκύψουν είναι παράλληλες ή σχεδόν παράλληλες, τότε θεωρούμε ότι ισχύει η υπόθεση της αναλογικότητας.

Σχεδιάζονται ξεχωριστά οι γραφικές παραστάσεις κάθε μεταβλητής. Στο ίδιο γράφημα, σχεδιάζονται οι καμπύλες για κάθε επίπεδο μεταβλητής. Πρέπει οι καμπύλες για κάθε επίπεδο μεταβλητής που θα προκύψουν να είναι παράλληλες. Όταν μια μεταβλητή είναι ποσοτική, τότε την κατηγοριοποιούμε και δημιουργούμε τις γραφικές παραστάσεις για κάθε κατηγορία αυτής. Αν θέλουμε να συμπεριλάβουμε στο μοντέλο αλληλεπιδράσεις κάποιων μεταβλητών, τότε σχεδιάζουμε στο ίδιο γράφημα τις γραφικές παραστάσεις των $\log[-\log(S(t,x))]$ συναρτήσει του χρόνου, για κάθε συνδυασμό των επιπέδων των μεταβλητών. Όταν μια μεταβλητή έχει πολλά επίπεδα ή έχουμε πολλές μεταβλητές, είναι δύσκολο να ελεγχθεί με γραφικές μεθόδους, αν ισχύει η υπόθεση της αναλογικότητας των κινδύνων.

Οι γραφικές παραστάσεις βασίζονται στην εκτίμηση των ποσοτήτων S(t,x) με μεθόδους που δε χρησιμοποιούν την υπόθεση της αναλογικότητας των κινδύνων. Τέτοιες είναι η εκτίμηση της συνάρτησης επιβίωσης με την

μέθοδο των Kaplan-Meier και η μέθοδος της στρωματοποιημένης διαδικασίας του Cox.

3.2.2.1. Έλεγχος της αναλογικότητας των κινδύνων στη στρωματοποιημένη ανάλυση

Οι έλεγχοι υποθέσεων στους συντελεστές **β** μπορούν να γίνουν με τους συνηθισμένους ελέγχους που βασίζονται στις υποθέσεις της συνάρτησης πιθανοφάνειας τροποποιώντας κάθε φορά αναλόγως, για κάθε στρώμα το $l(\boldsymbol{\beta})$,τον πίνακα πληροφορίας και το διάνυσμα score.

Έστω ότι p_i είναι σε πλήθος οι πλήρεις και διακεκριμένοι, ταξινομημένοι χρόνοι αποτυχίας της i ομάδας $t_{(1)}{}^i<t_{(2)}{}^i<\ldots<t_{(p_i)}{}^i$, i=1,…,I. Συμβολίζουμε με $\mathbf{x}_{(k)}{}^i=(\mathbf{x}_{(k)1}{}^i, \mathbf{x}_{(k)2}{}^i,\ldots, \mathbf{x}_{(k)p}{}^i)$, k=1,…, p_i το διάνυσμα των συμμεταβλητών που αντιστοιχεί στο άτομο του i στρώματος με πλήρη χρόνο ζωής $t_{(k)}{}^i$. Η συνάρτηση μερικής πιθανοφάνειας δίνεται από τη σχέση:

$$L_i(\boldsymbol{\beta})=\prod_{k=1}^{p_i}\frac{e^{\beta'\mathbf{x}_{(k)}{}^i}}{\sum_{j\in R_k}e^{\beta'\mathbf{x}_j{}^i}} \tag{3.9}$$

Το $l(\boldsymbol{\beta})$ γίνεται:

$$l(\boldsymbol{\beta})=\ln(L(\boldsymbol{\beta}))=\sum_{i=1}^{I}\sum_{k=1}^{p_i}\left(\beta'\mathbf{x}_{(k)}{}^i-\ln\sum_{j\in R_k}e^{\beta'\mathbf{x}_j{}^i}\right) \tag{3.10}$$

Το διάνυσμα score και τα στοιχεία του πίνακα πληροφορίας υπολογίζονται αντίστοιχα από τις παρακάτω σχέσεις:

$$U_s(\boldsymbol{\beta})=\sum_{i=1}^{I}\sum_{k=1}^{p_i}\left(\mathbf{x}_{(k)s}{}^i-\frac{\sum_{j\in R_k}\mathbf{x}_{js}{}^i e^{\beta'\mathbf{x}_j{}^i}}{\sum_{j\in R_k}e^{\beta'\mathbf{x}_j{}^i}}\right) \tag{3.11}$$

$$I_{sl}(\boldsymbol{\beta})=\sum_{i=1}^{I}\sum_{k=1}^{p_i}\left(\frac{\sum_{j\in R_k}e^{\beta'\mathbf{x}_j{}^i}\left(\sum_{j\in R_k}\mathbf{x}_{js}{}^i\mathbf{x}_{jl}{}^i e^{\beta'\mathbf{x}_j{}^i}\right)-\left(\sum_{j\in R_k}\mathbf{x}_{js}{}^i e^{\beta'\mathbf{x}_j{}^i}\right)\left(\sum_{j\in R_k}\mathbf{x}_{jl}{}^i e^{\beta'\mathbf{x}_j{}^i}\right)}{\sum_{j\in R_k}e^{\beta'\mathbf{x}_j{}^i}}\right) \tag{3.12}$$

όπου s,l=1,…,p

Αφού υπολογιστούν οι παραπάνω ποσότητες, υπολογίζονται τα στατιστικά του λόγου πιθανοφάνειας, του Wald και του score.

Όταν στρωματοποιούμε ως προς δύο ή περισσότερες μεταβλητές, το αποτέλεσμα είναι μια ξεχωριστή αναφορική συνάρτηση κινδύνου για κάθε συνδυασμό κατηγοριών. Για παράδειγμα, αν στρωματοποιήσουμε κάποια

δεδομένα ως προς δύο μεταβλητές, από τις οποίες η μια έχει τρία επίπεδα και η άλλη δύο, θα προκύψουν έξι διαφορετικά στρώματα.

Το μειονέκτημα της στρωματοποίησης είναι ότι δε δίνει κάποια εκτίμηση της επίδρασης των στρωμάτων και έτσι η ακρίβεια των εκτιμώμενων συντελεστών και η δύναμη των ελέγχων υποθέσεων μειώνεται όταν υπάρχει μεγάλος αριθμός στρωμάτων.

3.2.2.2. Έλεγχος της αναλογικότητας των κινδύνων βασισμένος στις ορισμένες εξαρτώμενες από το χρόνο μεταβλητές

Ένας έλεγχος που χρησιμοποιείται για την υπόθεση της αναλογικότητας των κινδύνων βασίζεται στις *ορισμένες* (defined) εξαρτώμενες από το χρόνο μεταβλητές. Οι μεταβλητές αυτές ονομάζονται έτσι, γιατί ενώ είναι σταθερές μεταβλητές, με ένα μετασχηματισμό γίνονται εξαρτώμενες από το χρόνο.

Έστω ότι έχουμε κ μεταβλητές $\mathbf{x}=(x_1,x_2,\ldots,x_\kappa)$ και ότι θέλουμε να εξετάσουμε αν η σταθερή μεταβλητή x_κ ικανοποιεί την υπόθεση αναλογικότητας των κινδύνων, υπό την παρουσία των υπόλοιπων κ-1 μεταβλητών. Ορίζουμε ένα μετασχηματισμό της x_κ πολλαπλασιάζοντας την x_κ με μια συνάρτηση του χρόνου g(t), δηλαδή $x_\kappa(t)=x_\kappa * g(t)$, με αποτέλεσμα η x_κ να γίνεται μεταβλητή εξαρτώμενη από το χρόνο. Συμβολίζουμε με $\mathbf{x}'$ το διάνυσμα των υπόλοιπων κ-1 μεταβλητών και θεωρούμε το μοντέλο του Cox:

$$h(t,\mathbf{x})= h_0(t)\exp(\beta_\kappa x_\kappa + \gamma x_\kappa\, g(t) + \boldsymbol{\beta}'\mathbf{x}') \qquad (3.13)$$

Βάση του παραπάνω μοντέλου γίνεται ο έλεγχος της μηδενικής υπόθεσης H_0: γ=0. Αν αυτή η υπόθεση γίνει δεκτή, τότε η μεταβλητή x_κ ικανοποιεί την υπόθεση της αναλογικότητας κινδύνων, διαφορετικά δεν την ικανοποιεί.

Γενικά μια μη μηδενική τιμή του γ σημαίνει αλλαγή της κινδυνότητας μεταξύ δύο ατόμων με διαφορετική τιμή του x_κ στο χρόνο. Η αλλαγή αυτή εξαρτάται από τη μορφή που θα έχει η συναρτήση g(t). Επειδή σκοπός μας είναι να εξετάσουμε την υπόθεση της αναλογικότητας των κινδύνων και όχι να μοντελοποιήσουμε την επίδραση του x_κ στο χρόνο, η επιλογή της g(t) τις περισσότερες φορές περιορίζεται σε κάποιες απλές συναρτήσεις του χρόνου. Μερικές συνηθισμένες επιλογές για τη g(t) είναι η ταυτοτική συνάρτηση g(t)=t και η $g(t)=\ln t$.

Η κινδυνότητα δύο ατόμων με τιμές της μεταβλητής 1 και 0 αντίστοιχα, και με τις ίδιες τιμές στις υπόλοιπες μεταβλητές $\mathbf{x}'$, είναι:

$$HR=\exp[\beta_\kappa + \gamma g(t)]$$

Στην περίπτωση που έχουμε γ>0 και η g(t) αύξουσα συνάρτηση του χρόνου, τότε η κινδυνότητα είναι αύξουσα συνάρτηση του χρόνου, ενώ για γ<0 η κινδυνότητα είναι φθίνουσα συνάρτηση του χρόνου.

Αν θέλουμε να εξετάσουμε την περίπτωση όλες οι μεταβλητές $x_1,x_2,\ldots,x_\kappa$ να ικανοποιούν ταυτόχρονα την υπόθεση αναλογικότητας των κινδύνων, τότε σχηματίζουμε το παρακάτω γενικευμένο μοντέλο:

$$h(t,\mathbf{x})= h_0(t)\exp\left(\sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \gamma_i [x_i g(t)]\right) \qquad (3.14)$$

Έπειτα κάνουμε τον έλεγχο της μηδενικής υπόθεσης H_0: γ=0. Αν η υπόθεση αυτή γίνει δεκτή, τότε συμπεραίνουμε ότι οι μεταβλητές $x_1,x_2,\ldots,x_\kappa$ ικανοποιούν την υπόθεση της αναλογικότητας των κινδύνων, διαφορετικά οι μεταβλητές $x_1,x_2,\ldots,x_\kappa$ δεν ικανοποιούν ταυτόχρονα την υπόθεση της αναλογικότητας των κινδύνων.

3.2.2.3. Έλεγχος της αναλογικότητας των κινδύνων στο γενικευμένο μοντέλο του Cox

Για τον υπολογισμό των συντελεστών **β** του γενικευμένου μοντέλου του Cox (2.36) χρησιμοποιούμε τη συνάρτηση μερικής πιθανοφάνειας από τη σχέση (2.22), αφού τροποποιηθεί κατάλληλα. Για τον υπολογισμό της μερικής πιθανοφάνειας είναι αρκετό να γνωρίζουμε τις τιμές του διανύσματος $x(t)=(x_1(t), x_2(t),\ldots, x_p(t))$ μόνο στις χρονικές στιγμές $t_{(j)}$, για τα άτομα που βρίσκονται σε κίνδυνο την χρονική αυτή στιγμή. Η συνάρτηση μερικής πιθανοφάνειας όταν έχουμε εξαρτημένες από το χρόνο μεταβλητές και δεν έχουμε ισοπαλίες δίνεται από την παρακάτω σχέση:

$$L(\boldsymbol{\beta}) =\prod_{j=1}^{k} \frac{e^{\boldsymbol{\beta}' x_{(j)}(t_{(j)})}}{\sum_{i\in R_j} e^{\boldsymbol{\beta}' x_i(t_{(j)})}} \qquad (3.15)$$

Όταν υπάρχουν ισότιμες παρατηρήσεις χρησιμοποιούνται οι τύποι των σχέσεων της παραγράφου (2.3.5.) με τις ανάλογες τροποποιήσεις.

3.3. ΥΠΟΛΟΙΠΑ

Τα υπόλοιπα χρησιμοποιούνται για τον έλεγχο διαφορετικών θεμάτων που αφορούν την καταλληλότητα του μοντέλου του Cox, όπως για τον έλεγχο της υπόθεσης αναλογικότητας των κινδύνων, για τον έλεγχο της ολικής επάρκειας του μοντέλου καθώς και για την εύρεση των outliers. Τα πιο γνωστά υπόλοιπα στο μοντέλο του Cox είναι τα υπόλοιπα Cox-Snell, τα martingale υπόλοιπα, τα Schoenfeld υπόλοιπα και τα υπόλοιπα απόκλισης

(deviance residuals). Κάθε ένα από αυτά χρησιμοποιείται για να ελέγξει κάποιο από τα θέματα που αναφέρθηκαν.

3.3.1. Τα υπόλοιπα Cox-Snell

Τα υπόλοιπα Cox-Snell είναι οι ποσότητες

$$\hat{r}_{ci} = \hat{H}_0(t_i)\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_i)=\hat{H}(t_i|\mathbf{x}_i),\ \ i=1,2,\ldots,n$$

όπου $\mathbf{x}_i = (x_{1i}, x_{2i},\ldots, x_{pi})'$ είναι το διάνυσμα των συμμεταβλητών του i ατόμου, $\hat{\boldsymbol{\beta}}' = (\hat{\beta}_1, \hat{\beta}_2,..., \hat{\beta}_p)$ οι εκτιμώμενοι συντελεστές παλινδρόμησης, n το πλήθος των παρατηρήσεων και t_i είναι ο πλήρης ή ο λογοκριμένος χρόνου του i ατόμου. Η εκτιμώμενη αθροιστική αναφορική συνάρτηση κινδύνου είναι: $\hat{H}_0(t)=\sum_{i:t_{(i)}\le t}\frac{d_i}{\sum_{j\in R_i}\exp[\hat{\boldsymbol{\beta}}'\mathbf{x}_j]}$.

Με d_i ορίζουμε το πλήθος των αποτυχιών στο χρόνο t_i και R_i το μέγεθος του συνόλου κινδύνου στο χρόνο t_i.

Η ποσότητα $\hat{r}_{ci}$ ονομάζεται Cox-Snell υπόλοιπο του i ατόμου, $1\le i\le n$ και οι ποσότητες αυτές υπολογίζονται μετά την προσαρμογή του μοντέλου στα δεδομένα για τους πλήρεις και διακεκριμένους χρόνους.

Για να είναι το μοντέλο κατάλληλο για την περιγραφή της τ.μ. T, τα υπόλοιπα αυτά θα πρέπει να κατανέμονται προσεγγιστικά σαν λογοκριμένα δεδομένα από μια εκθετική κατανομή με παράμετρο 1. Για την εκθετική κατανομή με παράμετρο 1 έχουμε ότι:

$$H(t) = -\ln S(t) = -\ln\left(e^{-\lambda t}\right) = \lambda t = t,\ \ t>0 \qquad (3.16)$$

και

$$\ln[H(t)] = \ln[-\ln S(t)] = \ln(\lambda t)= \ln\lambda + \ln t = \ln t,\ \ t>0 \qquad (3.17)$$

Έτσι, γραφική παράσταση της ποσότητας $\hat{H}(\hat{r}_{ci})$ συναρτήσει του $\hat{r}_{ci}$ θα πρέπει να είναι προσεγγιστικά μια ευθεία που περνάει από την αρχή των αξόνων και να έχει κλίση 1 για να είναι το μοντέλο επαρκές. Θα συγκρίνουμε δηλαδή με την ευθεία y=x λόγω της σχέσης (3.16). Μια άλλη γραφική παράσταση που μπορούμε να κάνουμε είναι της ποσότητας $\ln[\hat{H}(\hat{r}_{ci})]$ συναρτήσει του $\ln\hat{r}_{ci}$, που θα συγκριθεί με την ευθεία y=x λόγω της (3.17). Για την ποσότητα $\hat{H}(\hat{r}_{ci})$ μπορεί να χρησιμοποιηθεί είτε ο εκτιμητής αθροιστικής συνάρτησης κινδύνου Kaplan-Meier ή ο Nelson-Aalen εκτιμητής (Tableman & Kim (2004)).

3.3.2. Τροποποιημένα Cox-Snell υπόλοιπα

Τα Cox-Snell υπόλοιπα που αντιστοιχούν σε λογοκριμένους χρόνους δεν λαμβάνονται υπόψην στις γραφικές παραστάσεις για τον έλεγχο της ολικής επάρκειας του μοντέλου αναλογικού κινδύνου. Τα τροποποιημένα Cox-Snell υπόλοιπα λαμβάνουν υπόψην και τους λογοκριμένους χρόνους. Ορίζονται ως:

$$\hat{r}_{c\mathrm{i}}{}' = \begin{cases} \hat{r}_{c\mathrm{i}}, & \alpha\nu\ \mathrm{t_i}\ \pi\lambda\acute{\eta}\rho\eta\varsigma\ \chi\rho\acute{o}\nu o\varsigma \\ \hat{r}_{c\mathrm{i}} + \widehat{D}_{\mathrm{i}}, & \alpha\nu\ \mathrm{t_i}\ \ \lambda o\gamma o\kappa\rho\iota\mu\acute{\epsilon}\nu o\varsigma\ \chi\rho\acute{o}\nu o\varsigma \end{cases} \quad (3.18)$$

Αν υποθέσουμε ότι στο i άτομο αντιστοιχεί ο λογοκριμένος χρόνος $t_{\mathrm{i}}{}^{*}$ και ότι t_{i} είναι ο ακριβής, άγνωστος χρόνος επιβίωσης, τότε θα ισχύει $t_{\mathrm{i}} = t_{\mathrm{i}}{}^{*} + \widehat{D}_{\mathrm{i}}$, δηλαδή το $\widehat{D}_{\mathrm{i}}$ συμβολίζει τον υπόλοιπο χρόνο ζωής του i-στού ατόμου. Το $\widehat{D}_{\mathrm{i}}$ πρέπει να ακολουθεί εκθετική κατανομή με παράμετρο 1, για να είναι επαρκές το μοντέλο. Ως $\widehat{D}_{\mathrm{i}}$ παίρνουμε την μέση τιμή του $\widehat{D}_{\mathrm{i}}$. Όμως, $E(\widehat{D}_{\mathrm{i}})=1$, έτσι τα τροποποιημένα υπόλοιπα Cox-Snell υπολογίζονται από την παρακάτω σχέση:

$$\hat{r}_{c\mathrm{i}}{}' = \begin{cases} \hat{r}_{c\mathrm{i}}, & \alpha\nu\ \mathrm{t_i}\ \pi\lambda\acute{\eta}\rho\eta\varsigma\ \chi\rho\acute{o}\nu o\varsigma \\ \hat{r}_{c\mathrm{i}} + 1, & \alpha\nu\ \mathrm{t_i}\ \ \lambda o\gamma o\kappa\rho\iota\mu\acute{\epsilon}\nu o\varsigma\ \chi\rho\acute{o}\nu o\varsigma \end{cases} \quad (3.19)$$

3.3.3. Υπόλοιπα Schoenfeld

Ο Schoenfeld (1982) πρότεινε τα υπόλοιπα αυτά για το μοντέλο αναλογικών κινδύνων του Cox. Το πλεονέκτημα των υπολοίπων Schoenfeld έναντι των άλλων υπολοίπων είναι ότι για τον υπολογισμό τους δεν χρειάζεται η εκτίμηση της αθροιστικής αναφορικής συνάρτησης κινδύνου. Στα άλλα υπόλοιπα, υπολογίζεται ένα μόνο υπόλοιπο για κάθε άτομο. Τα υπόλοιπα Schoenfeld όμως, υπολογίζουν ένα ξεχωριστό υπόλοιπο για κάθε άτομο για κάθε μεταβλητή. Δηλαδή, εάν έχουμε p μεταβλητές, τότε για κάθε άτομο υπολογίζονται p Schoenfeld υπόλοιπα. Με την γραφική παράσταση των υπολοίπων Schoenfeld συναρτήσει του χρόνου μπορούμε να ελέγξουμε την υπόθεση αναλογικότητας των κινδύνων (PH υπόθεση). Αν το γράφημα έχει μια τυχαία μορφή των υπολοίπων έναντι του χρόνου τότε ικανοποιείται η PH υπόθεση.

Αν $\mathrm{x_k}$ είναι μεταβλητή για την οποία θέλουμε να υπολογίσουμε τα υπόλοιπα Schoenfeld, τότε το υπόλοιπο Schoenfeld ορίζεται να είναι η τιμή της συμμεταβλητής $\mathrm{x_k}$ για το i άτομο με πλήρη χρόνο $\mathrm{t_i}$ μείον την αναμενόμενη τιμή της συμμεταβλητής για τα άτομα που βρίσκονται σε κίνδυνο την χρονική στιγμή $\mathrm{t_i}$. Υπολογίζονται από την παρακάτω σχέση:

$$r_{\mathrm{ki}}=\delta_{\mathrm{i}}\left(\mathrm{x}_{\mathrm{ki}}-\frac{\sum_{l\in \mathrm{R}(\mathrm{t_i})}\mathrm{x}_{\mathrm{k}l}\exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l)}{\sum_{l\in \mathrm{R}(\mathrm{t_i})}exp(\hat{\boldsymbol{\beta}}'\mathbf{x}_l)}\right),\ \mathrm{i}=1,2,\ldots,\mathrm{n},\mathrm{k}=1,2,\ldots,\mathrm{p} \qquad (3.20)$$

όπου n το πλήθος των ατόμων και p το πλήθος των μεταβλητών, δ_{i}=0 αν $\mathrm{t_i}$ λογοκριμένος και δ_{i}=1 αν $\mathrm{t_i}$ πλήρης χρόνος.

Από την παραπάνω σχέση παρατηρούμε ότι αν δ_{i}=0 τότε η τιμή των υπολοίπων Schoenfeld είναι 0 για όλες τις μεταβλητές. Άρα τα υπόλοιπα Schoenfeld προσδιορίζονται στους μη λογοκριμένους χρόνους.

3.3.4. Υπόλοιπα martingale

Τα martingale υπόλοιπα χρησιμοποιούνται για την εύρεση των outliers, αλλά κυρίως για την εύρεση της συναρτησιακής μορφής μιας μεταβλητής που πρόκειται να εισαχθεί στο μοντέλο του Cox. Η πιο απλή μέθοδος για να βρούμε τη συναρτησιακή μορφή μιας μεταβλητής είναι να σχεδιάσουμε τα υπόλοιπα martingale από ένα μηδενικό μοντέλο (χωρίς μεταβλητές), δηλαδή ένα μοντέλο με $\hat{\beta}$=0, συναρτήσει κάθε μεταβλητής ξεχωριστά. Τα υπόλοιπα martingale υπολογίζονται από την εξής σχέση:

$$\hat{r}_{\mathrm{M_i}}=\delta_{\mathrm{i}}-\hat{r}_{c\mathrm{i}}\ ,\ 1\leq \mathrm{i}\leq \mathrm{n} \qquad (3.21)$$

Παίρνουν τιμές στο διάστημα (-∞,1), για λογοκριμένες παρατηρήσεις είναι αρνητικά και για δείγματα μεγάλα σε μέγεθος τα υπόλοιπα αυτά είναι ασυσχέτιστα το ένα με το άλλο. Εκφράζουν την διαφορά ανάμεσα στον παρατηρούμενο αριθμό θανάτων για το i άτομο στο διάστημα $(0,t_i)$ και τον αναμενόμενο αριθμό θανάτων κάτω από το προσαρμοσμένο μοντέλο.

Τα martingale υπόλοιπα δεν κατανέμονται συμμετρικά γύρω από το μηδέν, ακόμα και όταν το προσαρμοσμένο μοντέλο είναι σωστό. Αν κάποιο martingale υπόλοιπο έχει μια ψηλή αρνητική τιμή, τότε η αντίστοιχη παρατήρηση είναι outlier και δεν ερμηνεύεται καλά από το μοντέλο (Tableman & Kim (2004))

3.2.5. Υπόλοιπα απόκλισης (deviance residuals)

Τα υπόλοιπα απόκλισης βασίζονται στα martingale υπόλοιπα, αλλά κατανέμονται περισσότερο συμμετρικά γύρω από το μηδέν και έτσι είναι χρήσιμα για την ανίχνευση των outliers. Υπολογίζονται από την σχέση:

$$\hat{r}_{D_{\mathrm{i}}}=\mathrm{sgn}(\hat{r}_{\mathrm{M_i}})\sqrt{-2[\hat{r}_{\mathrm{M_i}}+\delta_{\mathrm{i}}\log(\delta_{\mathrm{i}}-\hat{r}_{\mathrm{M_i}})]}\ ,\ 1\leq \mathrm{i}\leq \mathrm{n} \qquad (3.22)$$

όπου $\mathrm{sgn}(\hat{r}_{\mathrm{M_i}})$ = 1 αν $\hat{r}_{\mathrm{M}_i}$ > 0 και $\mathrm{sgn}(\hat{r}_{\mathrm{M_i}})$ = 0 αν $\hat{r}_{\mathrm{M_i}}$ < 0.

Το $\hat{r}_{M_i}$ ορίζεται από τη σχέση (3.21).

Οι πληροφορίες για τα υπόλοιπα προέρχονται και από τις ιστοσελίδες :

http://ciser.cornell.edu/sasdoc/saspdf/stat/chap49.pdf και

http://www.meduniwien.ac.at/imc/biometrie/publikationen/Separata/Nardi_Schemper_1999_Biometrics.pdf) καθώς και από τους Tableman & Kim (2004).

3.4. Σύγκριση κατανομών επιβίωσης

3.4.1. Σύγκριση δύο κατανομών επιβίωσης

Τα δεδομένα που μελετάμε εκτός από τις βασικές μεταβλητές που δηλώνουν τον χρόνο ζωής και την τυχαία μεταβλητή που δηλώνει αν ο χρόνος είναι λογοκριμένος ή όχι, περιλαμβάνουν και άλλες μεταβλητές οι οποίες εκφράζουν χαρακτηριστικά των ατόμων που μελετάμε. Αυτό είχε σαν αποτέλεσμα, το ενδιαφέρον των ερευνητών να στρέφεται στην εκτίμηση της συνάρτησης επιβίωσης σε υποομάδες.

Προσαρμόζουμε στα δεδομένα το PH μοντέλο του Cox, σε σχέση με μια δίτιμη δείκτρια μεταβλητή X, η οποία παίρνει τις τιμές 1 και 0:

$$X=\begin{cases}1 \text{ αν το άτομο ανήκει στην κατηγορία } 1, \text{με κατανομή χρόνου επιβίωσης } S_1(t)\\ 0 \text{ αν το άτομο ανήκει στην κατηγορία } 2, \text{με κατανομή χρόνου επιβίωσης } S_2(t)\end{cases}$$

Τότε, το PH μοντέλο του Cox γίνεται:

$$h(t,x) = h_0(t)e^{\beta x} \text{ και } h(t,1) = h_0(t)e^{\beta} \qquad (3.23)$$

Επειδή όμως η αναφορική συνάρτηση $h_0(t)$ προκύπτει από τη συνάρτηση κινδύνου όλων των συμμεταβλητών για τιμή ίση με 0, έχουμε ότι $h_0(t) = h_2(t)$ ή $S_0(t) = S_2(t)$ και από την σχέση (2.21) έχουμε:

$$S_1(t) = [S_2(t)]^{e^{\beta}} \qquad (3.24)$$

Επομένως, ο έλεγχος της μηδενικής υπόθεσης $H_0 : S_1(t) = S_2(t)$, είναι ισοδύναμος με τον έλεγχο $H_0 : \beta = 0$.

Ο έλεγχος $H_0 : \beta = 0$ μπορεί να γίνει και με έναν από τους ελέγχους λόγου πιθανοφάνειας, Wald ή Score tests.

Στην περίπτωση που έχουμε μία μόνο μεταβλητή που παίρνει τις τιμές 0 και 1, το score test οδηγεί σε ένα απλό έλεγχο, αφού το διάνυσμα score θα είναι 1x1, δηλαδή θα είναι ένας αριθμός όπως και ο πίνακας πληροφορίας.

Έχουμε p=1 και έστω x_i, i=1,2,…,n η τιμή της μεταβλητής X που αντιστοιχεί στο άτομο με πλήρη ή λογοκριμένο χρόνο t_i, ενώ $x_{(i)}$, i=1,2,…,k είναι η τιμή της μεταβλητής X που αντιστοιχεί στο άτομο με πλήρη χρόνο t_i με $t_{(1)} < t_{(2)} <...< t_{(k)}$ να είναι οι πλήρεις, ταξινομημένοι χρόνοι. Τότε η σχέση (3.4) για το διάνυσμα score γίνεται:

$$U(\boldsymbol{\beta}) = \sum_{j=1}^{k}\left(x_{(j)} - \frac{\sum_{i\epsilon R_j} x_i e^{\boldsymbol{\beta}' \mathbf{x}_i}}{\sum_{i\epsilon R_j} e^{\boldsymbol{\beta}' \mathbf{x}_i}}\right) \quad (3.25)$$

Ταξινομούμε λογοκριμένους και πλήρεις χρόνους σε αύξουσα σειρά και στη συνέχεια υπολογίζουμε τις ποσότητες d_{1j} , d_j , n_{1j} και n_j όπου:

d_{1j} = I(t_j είναι ένας πλήρης χρόνος της ομάδας 1 με συνάρτηση επιβίωσης $S_1(t)$)

d_j = I(t_j είναι ένας πλήρης χρόνος)

n_{1j} : το πλήθος των ατόμων της ομάδας 1 με συνάρτηση επιβίωσης $S_1(t)$ που είναι σε κίνδυνο στον πλήρη χρόνο t_j

n_{2j} : το πλήθος των ατόμων της ομάδας 2 με συνάρτηση επιβίωσης $S_2(t)$ που είναι σε κίνδυνο στον πλήρη χρόνο t_j

Στην (3.26) όταν x_i=0, δηλαδή το άτομο ανήκει στη δεύτερη ομάδα ο όρος $x_i e^{\boldsymbol{\beta}' \mathbf{x}_i}$ δεν συνεισφέρει στο άθροισμα του αριθμητή. Ο παρονομαστής μπορεί να γραφεί ως $n_{1j}e^{\beta}+n_{2j}$, δηλαδή ως άθροισμα των όρων του αριθμητή συν το πλήθος των ατόμων της δεύτερης ομάδας που βρίσκονται σε κίνδυνο την χρονική στιγμή t_j. Αυτό συμβαίνει γιατί για x_i=1 έχουμε το άθροισμα των όρων $e^{\boldsymbol{\beta}' \mathbf{x}_i}$, ενώ για x_i=0 έχουμε $e^{\boldsymbol{\beta}' \mathbf{x}_i}$=1.

Σύμφωνα με τους παραπάνω συμβολισμούς, το διάνυσμα score στην σχέση (3.26) όταν έχουμε μία μόνο μεταβλητή με τιμές 0 και 1 γίνεται:

$$U(\boldsymbol{\beta}) = \sum_{j=1}^{k}\left(d_{1j} - \frac{d_j n_{1j} e^{\boldsymbol{\beta}' \mathbf{x}_i}}{n_{1j} e^{\beta}+n_{2j}}\right) \quad (3.26)$$

Ο πίνακας πληροφορίας είναι:

$$I(\boldsymbol{\beta}) = \sum_{j=1}^{n} \frac{d_j n_{1j} n_{2j} e^{\beta}}{(n_{1j}e^{\beta}+n_{2j})^2}$$

Για τον έλεγχο της μηδενικής υπόθεσης H_0 : β = 0, η (3.3) γίνεται για β_0=0:

$$Q=\frac{U(\mathbf{0})^2}{I(\mathbf{0})} \quad (3.27)$$

όπου $U(\mathbf{0}) = \sum_{j=1}^{n}(d_j - \frac{d_j n_{1j}}{n_{1j}+n_{2j}})$ και $I(\mathbf{0}) = \sum_{j=1}^{n} \frac{d_j n_{1j} n_{2j}}{(n_{1j}+n_{2j})^2}$

Η Q ακολουθεί προσεγγιστικά την $\chi_{(1)}{}^2$ κατανομή, έτσι βρίσκουμε τη σημαντικότητα του ελέγχου (p-value), για ένα βαθμό ελευθερίας (p-value=1-pchisq(Q,1)) και αν η τιμή αυτή είναι μικρότερη του 0.05 η μηδενική υπόθεση της ισότητας των δύο κατανομών απορρίπτεται.

Η διαδικασία αυτή είχε προταθεί από τους Mantel-Haenszel το 1959 και είναι γνωστή ως Mantel-Haenszel procedure (Collett 2003).

3.4.2. Σύγκριση r κατανομών επιβίωσης

Όταν μια κατηγορική μεταβλητή είναι χωρισμένη σε r κατηγορίες και χρειάζεται να γίνει η σύγκριση των συναρτήσεων επιβίωσης τους, $S_1(t)$,..., $S_r(t)$, τότε ακολουθούμε την παρακάτω διαδικασία:

Αρχικά, επιλέγουμε ένα επίπεδο της μεταβλητής που θα καθορίσει την αναφορική συνάρτηση κινδύνου. Το αποτέλεσμα θα είναι το ίδιο, όποιο επίπεδο και αν επιλεγεί. Έστω ότι επιλέγουμε το επίπεδο 1. Τότε ορίζουμε ένα διάνυσμα με r-1 βωβές μεταβλητές **z**=(z_2,…, z_r), των οποίων οι τιμές καθορίζονται από την παρακάτω σχέση:

z_{jq}= I(το άτομο που αντιστοιχεί σε χρόνο $t_{(j)}$ ανήκει στο επίπεδο q) , q=2,…,r, j=1,..,k

Εφόσον θεωρήσαμε ότι το πρώτο επίπεδο της μεταβλητής ορίζει την αναφορική συνάρτηση κινδύνου, προκύπτει ότι $h(t,0)= h_0(t)$, δηλαδή $h_1(t,\mathbf{z})=h_0(t)$ επομένως $S_1(t)=S_0(t)$. Αν προσαρμόσουμε ένα PH μοντέλο του Cox, προκύπτει:

$$S_q(t)= [S_0(t)]^{e^{\beta' z_q}} \text{ ή } S_q(t) = [S_0(t)]^{e^{\beta_q}} \text{ ,q=2,…,r} \quad (3.28)$$

όπου β_q αντιστοιχεί στον συντελεστή β στο επίπεδο q.

Η μηδενική υπόθεση H_0 : $S_1(t)= S_2(t)=...= S_r(t)$ από την παραπάνω σχέση γίνεται H_0 : **β**=0, **β**=(β_2,..., β_q). Ο έλεγχος αυτός μπορεί να γίνει με έναν από τους ελέγχους λόγου πιθανοφάνειας, Wald, Score tests ή διαστημάτων εμπιστοσύνης.

Ορίζουμε για q=2,…,r :

d_{qj} = I(t_j είναι ένας πλήρης χρόνος της ομάδας q με συνάρτηση επιβίωσης $S_q(t)$)

n_{1j} : το πλήθος των ατόμων της ομάδας 1 με συνάρτηση επιβίωσης $S_1(t)$ που είναι σε κίνδυνο στον πλήρη χρόνο t_j

δ_{sl}= I(s=l)

Επίσης $d_j=\sum_{q=1}^{r} d_{qj}$ και $n_j=\sum_{q=1}^{r} n_{qj}$. Τα στοιχεία του διανύσματος Score $U_q(\boldsymbol{\beta})$ και τα στοιχεία του πίνακα πληροφορίας $I_{sl}(\boldsymbol{\beta})$ γράφονται αντίστοιχα για $\boldsymbol{\beta}_0$=0:

$$U_q(0)=\sum_{j=1}^{n}\left(d_{qj}-\frac{d_j n_{qj}}{n_j}\right),\ q=2,\ldots,r$$

$$I_{sl}(0)=\sum_{j=1}^{n}\frac{d_j n_{sj}}{n_j}\left(\delta_{sl}-\frac{n_{lj}}{n_j}\right),\ s,l=2,\ldots,r$$

Χρησιμοποιώντας το στατιστικό score βρίσκουμε την τιμή του Q από την σχέση (3.3) για $\boldsymbol{\beta}_0$=0, η οποία ακολουθεί την χ^2_{r-1}.

ΚΕΦΑΛΑΙΟ 4

ΜΗ ΠΑΡΑΜΕΤΡΙΚΕΣ ΜΕΘΟΔΟΙ ΣΤΗΝ ΑΝΑΛΥΣΗ ΕΠΙΒΙΩΣΗΣ

4.1. Εκτιμητής Kaplan-Meier

Όταν υπάρχουν λογοκριμένα δεδομένα, τότε χρησιμοποιείται η μέθοδος γινομένου-ορίου (Product-Limit method) ή μέθοδος Kaplan-Meier (Kaplan & Meier, 1958), η οποία είναι διαφορετική από την μέθοδο εκτίμησης της συνάρτησης επιβίωσης της παραγράφου (2.1). Ο εκτιμητής S(t) που προκύπτει ονομάζεται εκτιμητής Kaplan-Meier ή εκτιμητής γινομένου-ορίου (PLE).

Για να χρησιμοποιηθεί η μέθοδος πρέπει να ισχύουν οι παρακάτω προϋποθέσεις:

- Τα άτομα που χάθηκαν από την παρακολούθηση έχουν την ίδια πιθανότητα επιβίωσης με τα άτομα που συνεχίζουν να παρακολουθούνται. Κάτι τέτοιο δε μπορεί να ελεγχθεί και μπορεί να οδηγήσει σε μεροληψία που μειώνει το S(t).
- Οι πιθανότητες επιβίωσης είναι οι ίδιες για άτομα που εισήλθαν στην αρχή της μελέτης και για άτομα που εισήλθαν αργότερα.
- Το γεγονός που μελετάται (π.χ. θάνατος) συμβαίνει σε καθορισμένο χρόνο. Καθυστερημένη καταγραφή του γεγονότος προκαλεί αύξηση του S(t).

Ο εκτιμητής γινομένου-ορίου είναι ο εκτιμητής της πιθανότητας οι ασθενείς να επιβιώσουν για μια συγκεκριμένη διάρκεια χρόνου. Έστω:

p_1: η πιθανότητα επιβίωσης για τουλάχιστον έναν χρόνο

p_2: η πιθανότητα επιβίωσης το δεύτερο χρόνο, δεδομένου ότι οι ασθενείς επιβίωσαν τον πρώτο χρόνο

p_j: η πιθανότητα επιβίωσης τον χρόνο j, δεδομένου ότι οι ασθενείς επιβίωσαν τα προηγούμενα j-1 χρόνια.

Υποθέτουμε ότι είναι γνωστοί οι πλήρεις και λογοκριμένοι χρόνοι N ατόμων και έστω $t_{(1)}<t_{(2)}<\ldots<t_{(k)}$, $k\leq N$ οι διακεκριμένοι, ταξιμονημένοι χρόνοι αποτυχίας.

Για j=2,…,k έστω:

n_j: ο αριθμός των ατόμων που είναι σε κίνδυνο στο χρόνο $t_{(j)}$.

d_j: ο αριθμός των θανάτων στο χρόνο $t_{(j)}$.

c_j: ο αριθμός των διαφυγών στο διάστημα $[t_{(j)}, t_{(j+1)})$.

Η ποσότητα n_j-d_j είναι ο αριθμός των ασθενών που επιβιώνουν στο χρόνο $t_{(j)}$ και η n_j-d_j-c_j ο αριθμός των ασθενών που βρίσκονται σε κίνδυνο στο χρόνο $t_{(j+1)}$, δηλαδή ισχύει:

$$n_{(j+1)}=n_j\text{-}d_j\text{-}c_j$$

Σύμφωνα με τους συμβολισμούς αυτούς, η σχετική συχνότητα αυτών που επιβιώνουν στο διάστημα $[t_{(j)}, t_{(j+1)})$ είναι $p_j=p(t_{(j)})$ και δίνεται από την παρακάτω σχέση:

$$p_j = \frac{n_j-d_j}{n_j} = 1 - \frac{d_j}{n_j} \qquad (4.1)$$

Ο εκτιμητής Kaplan-Meier του S(t) υπολογίζεται ως:

$$\hat{S}(t)=\prod_{j:t_{(j)}\le t}\left(\frac{n_j-d_j}{n_j}\right) \qquad (4.2)$$

Θεωρούμε ότι όλοι οι ασθενείς στο χρόνο 0 είναι ζωντανοί, δηλαδή $\hat{S}(0)=1$.

Παρατηρούμε στη σχέση (4.1) ότι όταν δεν πεθαίνει κανένα άτομο, δηλαδή για $d_j=0$, έχουμε $p_j=1$. Ο εκτιμητής Kaplan-Meier της πιθανότητας επιβίωσης στο χρόνο t αλλάζει μόνο σε χρόνους στους οποίους πεθαίνει τουλάχιστον ένα άτομο. Αυτό έχει σαν αποτέλεσμα να μπορούμε να παραλείψουμε τους χρόνους στους οποίους δεν παρατηρούνται θάνατοι, δηλαδή στους λογοκριμένους χρόνους για τον υπολογισμό του εκτιμητή Kaplan-Meier.

4.2. Καμπύλη επιβίωσης

Η γραφική παράσταση του $\hat{S}(t)$ συναρτήσει του t μας δίνει τον εκτιμητή Kaplan-Meier της καμπύλης επιβίωσης και πρόκειται για μια καλή προσέγγιση των δεδομένων. Η $\hat{S}(t)$ είναι μια βαθμιδωτή φθίνουσα συνάρτηση, συνεχής από αριστερά. Η τιμή της δεν αλλάζει, παρά μόνο στα σημεία όπου παρατηρούνται πλήρεις χρόνοι. Κάθε επόμενο βήμα προς τα κάτω θα είναι λίγο μεγαλύτερο και η τιμή της μειώνεται κατά $\frac{n_j-d_j}{n_j}$ αμέσως μετά τον j πλήρη χρόνο, $t_{(j)}$.

Επειδή η λογοκρισία του ασθενή μειώνει τον αριθμό των ασθενών που συνεισφέρουν στην καμπύλη, κάθε θάνατος από αυτό το σημείο παριστάνει μια μεγαλύτερη αναλογία του υπόλοιπου πληθυσμού από την αναλογία που θα είχαμε αν γνωρίζαμε τους πλήρεις χρόνους. Έτσι, η λογοκρισία επηρεάζει την καμπύλη επιβίωσης.

Όταν η μεγαλύτερη παρατήρηση είναι μη-λογοκριμένη, ο εκτιμητής Kaplan-Meier στο σημείο αυτό είναι 0, αφού θα έχουμε $n_k=d_k$ και $c_k=0$ οπότε από τη σχέση (4.1) θα έχουμε $p_k=0$ και $\hat{S}(k)=0$. Σε αυτή την περίπτωση, στην καμπύλη επιβίωσης θα έχουμε μια κάθετη γραμμή στο t_k, από τον προτελευταίο πλήρη χρόνο που θα κατεβαίνει κάθετα στο t_k. Δηλαδή $\hat{S}(t)=0$ για όλα τα $t \geq t_k$. Το αποτέλεσμα αυτό βασίζεται μόνο σε ένα ασθενή και είναι λάθος να συμπεράνουμε ότι η πιθανότητα είναι 0 ένας ασθενής να επιβιώσει περισσότερο από t_k.

Όταν η μεγαλύτερη παρατήρηση είναι λογοκριμένη, τότε η καμπύλη επιβίωσης δεν είναι 0 μετά το t_{max}, αφού σε αυτή την περίπτωση $n_{max} \neq d_{max}$ και $\hat{S}(max) \neq 0$. Δηλαδή, η καμπύλη μετά το t_{max}, θα συνεχίζει παράλληλα με τον άξονα των t χωρίς να κατεβαίνει προς τα κάτω (Kaplan, E. L. and Meier, P. (1958)).

4.3. Ο εκτιμητής Nelson-Aalen της αθροιστικής συνάρτησης κινδύνου

Ο εκτιμητής Nelson-Aalen (Nelson & Aalen, 1969, 1976) είναι ένας απλός μη παραμετρικός εκτιμητής της αθροιστικής συνάρτησης κινδύνου για δεξιά λογοκριμένα δεδομένα. Υπολογίζεται από τη σχέση:

$$\hat{H}(t)=\sum_{j:t_{(j)}\leq t}\frac{d_j}{n_j} \qquad (4.3)$$

Το γράφημα της $\hat{H}(t)$ δίνει χρήσιμες πληροφορίες για το σχήμα της συνάρτησης κινδύνου h(t). Για παράδειγμα, η $\hat{H}(t)$ είναι γραμμική αν η h(t) είναι σταθερή.

Από τη σχέση S(t)=exp[-H(t)] μια εναλλακτική εκτίμηση για τη συνάρτηση επιβίωσης γνωστή με το όνομα Nelson-Aalen εκτιμητής της συνάρτησης επιβίωσης S(t) είναι:

$$\hat{S}(t) =\exp[-\hat{H}(t)] \qquad (4.4)$$

http://data.princeton.edu/pop509/NonParametricSurvival.pdf) και Tableman & Kim (2004).

4.4. Μη παραμετρικές μέθοδοι για τη σύγκριση καμπυλών επιβίωσης

Όταν έχουμε δεδομένα επιβίωσης, δεν μας ενδιαφέρει μόνο η εκτίμηση της συνάρτησης επιβίωσης, αλλά και η σύγκριση του χρόνου επιβίωσης δύο ή περισσότερων ομάδων που διαφέρουν ως προς ένα χαρακτηριστικό. Εάν έχουμε ένα σύνολο ατόμων με την ίδια ασθένεια, μπορεί να ενδιαφερόμαστε να συγκρίνουμε την ικανότητα δύο ή περισσότερων

θεραπειών να παρατείνουν την ζωή των ασθενών. Με το σχεδιασμό των εκτιμώμενων συναρτήσεων επιβίωσης, μπορεί να έχουμε μια οπτική εικόνα για το αν υπάρχει διαφορά μεταξύ των συναρτήσεων επιβίωσης διαφορετικών ομάδων. Αυτό μας δίνει μια γενική ιδέα για τους χρόνους επιβίωσης και δε μπορούμε να συμπεράνουμε αν οι διαφορές είναι σημαντικές. Είναι απαραίτητος ένας στατιστικός έλεγχος και επειδή οι χρόνοι επιβίωσης δεν κατανέμονται κανονικά, πρέπει να εφαρμοστούν μη-παραμετρικοί έλεγχοι που θα βασίζονται στην ταξινόμηση των χρόνων επιβίωσης.

Το Logrank test

Έστω ότι υπάρχουν n άτομα τα οποία χωρίζονται τυχαία σε δύο ομάδες. Στην πρώτη ομάδα τοποθετούνται n_1 άτομα που λαμβάνουν την θεραπεία Α και στη δεύτερη n_2 άτομα με θεραπεία Β. Έστω $x_1, \dots, x_{r_1}$ οι r_1 πλήρεις χρόνοι και $x^+_{r_1+1}, \dots, x^+_{n_1}$ οι $n_1 - r_1$ λογοκριμένοι χρόνοι της πρώτης ομάδας. Αντίστοιχα, για την δεύτερη ομάδα ορίζονται οι $y_1, \dots, y_{r_2}$ οι r_2 πλήρεις χρόνοι και $y^+_{r_2+1}, \dots, y^+_{n_2}$ οι $n_2 - r_2$ λογοκριμένοι χρόνοι. Δηλαδή στο τέλος της μελέτης $n_1 - r_1$ άτομα που λαμβάνουν την θεραπεία Α και $n_2 - r_2$ άτομα που λαμβάνουν την θεραπεία Β θα είναι ζωντανοί. Υποθέτουμε επίσης ότι $S_A(t)$ και $S_B(t)$ είναι οι συναρτήσεις επιβίωσης των ατόμων με θεραπείες Α και Β αντίστοιχα. Ο έλεγχος είναι:

$H_0: S_A(t)=S_B(t)$ έναντι των εναλλακτικών:

$H_1: S_A(t) \neq S_B(t)$ ή

$H_2: S_A(t) > S_B(t)$ ή

$H_3: S_A(t) < S_B(t)$

Το logrank test συγκρίνει τον παρατηρούμενο αριθμό θανάτων με τον αναμενόμενο αριθμό στις δύο ομάδες. Έστω O_1 και O_2 οι παρατηρούμενοι αριθμοί θανάτων και E_1, E_2 οι αναμενόμενοι αριθμοί θανάτων στις δύο ομάδες θεραπειών Α και Β αντίστοιχα.

Το logrank στατιστικό υπολογίζεται από τη σχέση:

$$C^2_{\text{logrank}} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \qquad (4.5)$$

Ακολουθεί προσεγγιστικά την χ^2 κατανομή με 1 βαθμό ελευθερίας. Μια μεγάλη τιμή του C^2_{logrank} θα οδηγήσει σε απόρριψη της H_0. Γενικά, για m ομάδες θεραπειών, το C^2_{logrank} συγκρίνεται με το χ^2 με m-1 βαθμούς ελευθερίας.

Για τον υπολογισμό των E_1 και E_2, ταξινομούμε τους πλήρεις χρόνους και των δύο ομάδων μαζί σε αύξουσα σειρά. Υπολογίζουμε τους αναμενόμενους χρόνους θανάτου σε κάθε μη-λογοκριμένο χρόνο για κάθε ομάδα ξεχωριστά και τους αθροίζουμε για να προκύψουν τα E_1 και E_2. Έστω d_{1t} και d_{2t} το πλήθος των θανάτων των δύο ομάδων στο χρόνο t και n_{1t}, n_{2t} το πλήθος των ασθενών που βρίσκονται σε κίνδυνο μέχρι και το χρόνο t στις ομάδες Α και Β αντίστοιχα. Οι αναμενόμενοι αριθμοί θανάτων e_{1t} και e_{2t} στο χρόνο t είναι:

$$e_{1t} = \frac{n_{1t}}{n_{1t}+ n_{2t}} d_{1t} \quad \text{και} \quad e_{2t} = \frac{n_{2t}}{n_{1t}+ n_{2t}} d_{2t} \tag{4.6}$$

και

$$E_1 = \sum_t e_{1t} \quad \text{και} \quad E_2 = \sum_t e_{2t} \tag{4.7}$$

Για τη σύγκριση συναρτήσεων επιβίωσης περισσότερων των δύο ομάδων, το Logrank test υπολογίζεται με την παρακάτω διαδικασία.

Έστω $O_1,...,O_k$ οι παρατηρούμενοι αριθμοί θανάτων και $E_1,..,E_k$ οι αναμενόμενοι αριθμοί θανάτων στις k ομάδες θεραπειών $A_1,...,A_k$ αντίστοιχα.

Το logrank στατιστικό υπολογίζεται από τη σχέση:

$$C^2_{\text{logrank}} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{4.8}$$

και ακολουθεί προσεγγιστικά την χ^2 κατανομή με k-1 βαθμούς ελευθερίας.

Για τον υπολογισμό των E_j, ταξινομούμε τους πλήρεις χρόνους και των k ομάδων μαζί σε αύξουσα σειρά. Υπολογίζουμε τους αναμενόμενους αριθμούς θανάτων σε κάθε μη-λογοκριμένο χρόνο για κάθε ομάδα ξεχωριστά και τους αθροίζουμε για να προκύψουν τα E_j. Έστω d_{jt}, j=1,..,k το πλήθος των θανάτων των k ομάδων στο χρόνο t και n_{jt} το πλήθος των ασθενών που βρίσκονται σε κίνδυνο μέχρι και το χρόνο t στις ομάδες $A_1,...,A_k$ αντίστοιχα. Οι αναμενόμενοι χρόνοι θανάτου e_{jt} στο χρόνο t είναι:

$$e_{jt} = \frac{n_{jt}}{\sum_{i=1}^{k} n_{it}} d_{jt} \text{ , j=1,..,k} \tag{4.9}$$

και

$$E_j = \sum_t e_{jt} \text{ , } \quad \text{j=1,..,k} \tag{4.10}$$

(Mantel-Haenszel (1959))

ΚΕΦΑΛΑΙΟ 5

ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΧΡΗΣΙΜΟΠΟΙΩΝΤΑΣ ΤΗΝ R

5.1. Πραγματικά δεδομένα

Τα δεδομένα που θα χρησιμοποιήσουμε στη συνέχεια αφορούν 51 ασθενείς οι οποίοι πάσχουν από οξεία μυελοβλαστική λευχαιμία και δεν έχουν δεχτεί μέχρι τώρα κάποια θεραπεία. Τα δεδομένα βρίσκονται στις σημειώσεις Κ. Φωκιανός & Χ. Χαραλάμπους, (2010) Εισαγωγή στην R – Πρόχειρες σημειώσεις και είναι διαθέσιμα στο διαδίκτυο. Οι ασθενείς υποβάλλονται σε μια θεραπεία, στο τέλος της οποίας εξετάζονται αν έχουν ανταποκριθεί ή όχι. Πριν τη θεραπεία καταγράφονται οι παρακάτω μεταβλητές:

1. *Age*: η ηλικία διάγνωσης
2. *Smear*: το ποσοστό επίστρωσης των βλαστοκυττάρων
3. *Infil*: το ποσοστό των κυττάρων από τη λευχαιμία που εισήλθαν στο μυελό των οστών
4. *Index*: το ποσοστό των κυττάρων που προήλθαν από το μυελό των οστών
5. *Blasts*: τα απόλυτα βλαστοκύτταρα

Ακόμη, καταγράφονται ο χρόνος επιβίωσης του ατόμου, Time, και η ανταπόκρισή του στη θεραπεία, Resp. Η μεταβλητή Status δείχνει αν οι παρατηρήσεις ενός ατόμου είναι λογοκριμένες ή όχι. Συγκεκριμένα, το ποσοστό λογοκρισίας είναι περίπου 11,8%, ποσοστό όχι μεγάλο. Στα δεδομένα περιλαμβάνεται και μία έκτη μεταβλητή, η θερμοκρασία (*Temp*) αλλά στην ανάλυση που ακολουθεί δεν έχει χρησιμοποιηθεί.

Θα εξεταστεί κατά πόσο είναι σημαντικές για την πρόβλεψη του χρόνου επιβίωσης οι πέντε μεταβλητές που καταγράφηκαν πριν την θεραπεία.

Τα παραπάνω δεδομένα αποθηκεύονται στην R με το όνομα cancer.dat.

5.2. Περιγραφή των συναρτήσεων coxph(), Surv() και survdiff() στην R

Στην R οι πιο σημαντικές συναρτήσεις για την ανάλυση επιβίωσης βρίσκονται στη βιβλιοθήκη survival. Η συνάρτηση coxph() είναι η βασικότερη εντολή της όσον αφορά το μοντέλο αναλογικών κινδύνων του Cox, αφού μέσω της εντολής αυτής γίνεται η προσαρμογή του μοντέλου στα δεδομένα.

Η συνάρτηση Surv() -δημιουργεί ένα αντικείμενο επιβίωσης- παράγει έναν εκτιμητή της συνάρτησης επιβίωσης. Στην περίπτωση που τα δεδομένα είναι δεξιά λογοκριμένα, η συνάρτηση Surv() έχει τη μορφή Surv(time,event), όπου time είναι ο χρόνος μέχρι το γεγονός ή χρόνος λογοκρισίας και event είναι μια μεταβλητή με τιμή ίση με 1 αν το γεγονός παρατηρείται ή ίση με 0 αν η παρατήρηση είναι λογοκριμένη.

Τέλος, η συνάρτηση survdiff() ελέγχει αν υπάρχει διαφορά μεταξύ δύο ή περισσότερων καμπυλών επιβίωσης.

5.3. Εφαρμογή του μοντέλου αναλογικών κινδύνων του Cox

Η στατιστική ανάλυση των δεδομένων αυτού του κεφαλαίου θα ακολουθήσει τα βασικά βήματα των Tableman και Kim (2004) αλλά και των σημειώσεων Φωκιανός και Χαραλάμπους (2010). Εφαρμόζοντας το μοντέλο αναλογικών κινδύνων με τις 5 επεξηγηματικές μεταβλητές παίρνουμε τα παρακάτω αποτελέσματα:

```
> library(survival)

> attach(cancer.dat)

>cancer.cox<-
coxph(Surv(time,status)~Age+Smear+Infil+Index+Blasts,data=cancer.dat)

> summary(cancer.cox)

coxph(formula = Surv(time, status) ~ Age + Smear + Infil + Index + Blasts, data =
cancer.dat)

 n= 51, number of events= 45

              coef         exp(coef)      se(coef)        z      Pr(>|z|)

Age       0.035355       1.035988      0.010181    3.473      0.000516 ***
```

Smear	0.009154	1.009196	0.014512	0.631	0.528181
Infil	-0.018349	0.981818	0.012469	-1.472	0.141119
Index	-0.089552	0.914341	0.044822	-1.998	0.045724 *
Blasts	0.002853	1.002857	0.009733	0.293	0.769434

	exp(coef)	exp(-coef)	lower .95	upper .95
Age	1.0360	0.9653	1.0155	1.0569
Smear	1.0092	0.9909	0.9809	1.0383
Infil	0.9818	1.0185	0.9581	1.0061
Index	0.9143	1.0937	0.8374	0.9983
Blasts	1.0029	0.9972	0.9839	1.0222

Likelihood ratio test=	19.06	on 5 df,	p=0.001877
Wald test =	17.59	on 5 df,	p=0.003505
Score (logrank) test =	18.94	on 5 df,	p=0.001972

Από τα παραπάνω αποτελέσματα συμπεραίνουμε ότι η μηδενική υπόθεση H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ απορρίπτεται σε επίπεδο σημαντικότητας 0.05. Ακόμη, με ε.σ. 0.05 η μεταβλητή Age είναι σημαντική αφού έχει p-value<0.05, οι μεταβλητές Smear, Infil και Blasts δεν είναι σημαντικές ενώ για την μεταβλητή Index δεν μπορούμε προς το παρόν να αποφασίσουμε γιατί έχει p-value πολύ κοντά στο ε.σ. Να σημειωθεί πως ο αριθμός 1 ανήκει στα διαστήματα εμπιστοσύνης των μεταβλητών Smear, Infil και Blasts.

Και οι τρεις έλεγχοι δείχνουν πως υπάρχει σημαντική σχέση μεταξύ των μεταβλητών Age και Index και του χρόνου επιβίωσης. Οι τιμές κάθε κριτηρίου-ελεγχοσυνάρτησης συγκρίνονται με την $\chi^2{}_5$ και δίνουν την p-value της τρίτης στήλης. Εφόσον και οι τρεις έλεγχοι έχουν p-value μικρότερη από 0.05 συμφωνούν πως το μοντέλο που περιέχει όλες τις μεταβλητές με κύριες τις Age και Index ταιριάζει στα δεδομένα.

Για να επιλέξουμε ποιο είναι το καλύτερο μοντέλο που ταιριάζει στα δεδομένα θα εφαρμοστεί το AIC κριτήριο στο αρχικό μοντέλο, το οποίο θα συμπεριλαμβάνει και αλληλεπιδράσεις μέχρι τάξεως 2.

> *library(MASS)*

>*cancer.cox1<-stepAIC(cancer.cox,~.^2)*

> *cancer.cox1$anova*

Initial Model:

Surv(time, status) ~ Age + Smear + Infil + Index + Blasts

Final Model:

Surv(time, status) ~ Age + Infil + Index

Step	Df	Deviance Resid.	Df	Resid. Dev	AIC
1			46	266.8247	276.8247
2 - Blasts	1	0.08236656	47	266.9071	274.9071
3 - Smear	1	0.44359927	48	267.3507	273.3507

Καλύτερο μοντέλο είναι αυτό με την μικρότερη τιμή του AIC κριτηρίου. Σύμφωνα με τα παραπάνω, οι μεταβλητές Smear, Blasts δεν επηρεάζουν το μοντέλο και οι αλληλεπιδράσεις δεν είναι σημαντικές.

> *cancer.cox1*

coxph(formula = Surv(time, status) ~ Age + Infil + Index, data = cancer.dat)

	coef	exp(coef)	se(coef)	z	p
Age	0.0339	1.034	0.00977	3.47	0.00052
Infil	-0.0108	0.989	0.00699	-1.54	0.12000
Index	-0.0831	0.920	0.03990	-2.08	0.03700

Likelihood ratio test=18.5 on 3 df, p=0.000342 n= 51, number of events= 45

Από το μοντέλο αφαιρείται η Infil καθώς έχει μεγάλη p-value=0.12. Παρακάτω προσαρμόζονται τα δεδομένα στο μοντέλο αναλογικών κινδύνων με μεταβλητές τις Age και Index.

> *cancer.cox2<-coxph(Surv(time,status)~Age+Index,data=cancer.dat)*

> *cancer.cox2*

coxph(formula = Surv(time, status) ~ Age + Index, data = cancer.dat)

	coef	exp(coef)	se(coef)	z	p
Age	0.0340	1.035	0.00959	3.55	0.00039
Index	-0.0778	0.925	0.03862	-2.02	0.04400

Likelihood ratio test=16.1 on 2 df, p=0.000312 n= 51, number of events= 45

> *-2*cancer.cox2$loglik[2]+2*cancer.cox1$loglik[2]*

[1] 2.381104

> *1-pchisq(2.381104,1)*

[1] 0.1228107

Έπειτα, ελέγχουμε αν μπορούμε πράγματι να θεωρήσουμε το απλούστερο μοντέλο αφαιρώντας τις τρεις μεταβλητές που δεν είναι σημαντικές. Ο έλεγχος λόγου πιθανοφάνειας για τη σύγκριση του μοντέλου που περιλαμβάνει τις μεταβλητές Age, Infill και Index και του μειωμένου μοντέλου που περιλαμβάνει τις μεταβλητές Age και Index δίνει τιμή *2.381104.* Η τιμή αυτή συγκρίνεται με την χ^2_1. Εφόσον *1-pchisq(2.381104,1)= 0.1228107 > 0.05* δεχόμαστε το μειωμένο μοντέλο.

Η μεταβλητή Index έχει p-value πολύ κοντά στο ε.σ. 0.05 με αποτέλεσμα να πρέπει να εξεταστεί περαιτέρω αν είναι στατιστικά σημαντική μεταβλητή ή όχι.

> *cancer.cox3<-coxph(Surv(time,status)~Age,data=cancer.dat)*

> *cancer.cox3*

coxph(formula = Surv(time, status) ~ Age, data = cancer.dat)

```
        coef  exp(coef)  se(coef)     z        p
Age   0.0324       1.03   0.00952   3.4  0.00067

Likelihood ratio test=11.8  on 1 df , p=0.000577  n= 51, number of events= 45
```

*> -2*cancer.cox3$loglik[2]+2*cancer.cox1$loglik[2]*

```
[1] 6.680512
```

> 1-pchisq(6.680512,2)

```
[1] 0.03542789
```

Από τα παραπάνω αποτελέσματα, εφόσον *1-pchisq(6.680512,2)= 0.03542789<0.05* απορρίπτουμε το μοντέλο με μόνη μεταβλητή την Age σε επίπεδο σημαντικότητας 5%. Άρα το μοντέλο που ταιριάζει καλύτερα στα δεδομένα είναι πιθανόν το cancer.cox2 το οποίο περιλαμβάνει τις μεταβλητές Age και Index.

> summary(cancer.cox2)

```
coxph(formula = Surv(time, status) ~ Age + Index, data = cancer.dat)

 n= 51, number of events= 45

              coef     exp(coef)      se(coef)          z      Pr(>|z|)
Age       0.034016      1.034601      0.009588      3.548      0.000388 ***
Index    -0.077847      0.925106      0.038623     -2.016      0.043845 *


        exp(coef)     exp(-coef)    lower .95    upper .95
Age        1.0346         0.9666       1.0153       1.0542
Index      0.9251         1.0810       0.8577       0.9979

Likelihood ratio test=  16.15   on 2 df,    p=0.0003115
Wald test             =  15.51  on 2 df,    p=0.0004277
Score (logrank) test =  16.42  on 2 df,    p=0.0002719
```

Σύμφωνα με τα παραπάνω, οι μεταβλητές Age και Index είναι στατιστικά σημαντικές. Συγκεκριμένα, η Age έχει p-value πολύ μικρότερη από το ε.σ. ενώ η Index αν και έχει p-τιμή πολύ κοντά στο 0.05, θεωρείται σημαντική. Και οι τρεις έλεγχοι, έλεγχος λόγου πιθανοφάνειας, Wald και Score test δίνουν μικρές p-value και συμφωνούν στο γεγονός ότι υπάρχει σημαντική σχέση μεταξύ των μεταβλητών Age και Index και του χρόνου επιβίωσης.

Το κριτήριο AIC και η αρχική πρόβλεψη, δηλαδή το μοντέλο cancer.cox, δίνουν τελικά το ίδιο αποτέλεσμα για το ποιες μεταβλητές είναι σημαντικές και ποιες όχι. Το cancer.cox δίνει τις Age και Index σημαντικές, ενώ το AIC κριτήριο, επιλέγει τις Age, Index και Infil αλλά η τελευταία απορρίπτεται μέσω στατιστικού ελέγχου αφού έχει p-value μεγαλύτερη του 0.05.

Συμπεράσματα

- Οι μεταβλητές Age και Index συνεισφέρουν στην καλή ερμηνεία του μοντέλου. Η ερμηνεία του εκτιμώμενου συντελεστή του μοντέλου είναι ότι κάθε επιπρόσθετος χρόνος ζωής αυξάνει το λογάριθμο της συνάρτησης κινδύνου κατά 0.0340, ενώ αύξηση κατά ένα βαθμό του ποσοστού των κυττάρων που προήλθαν από το μυελό των οστών μειώνει το λογάριθμο της συνάρτησης κινδύνου κατά 0.077847.
- Βρίσκοντας πρώτα την εκθετική συνάρτηση του συντελεστή μπορεί να γίνει μια πιο σωστή προσέγγιση. Για κάθε αύξηση κατά μία μονάδα της μεταβλητής Age, η συνάρτηση κινδύνου πολλαπλασιάζεται με τον εκθετικό συντελεστή exp(*coef)*. Ο ορος

$$\exp(coef)\text{-}1 = 1.034\text{-}1 = 0{,}034$$

δίνει την ποσοστιαία αλλαγή στη συνάρτηση κινδύνου για κάθε μοναδιαία αύξηση στην μεταβλητή. Άρα, αύξηση ενός χρόνου στην ηλικία διάγνωσης οδηγεί σε αύξηση 3.4% της συνάρτησης κινδύνου. Αντίστοιχα, για αύξηση κατά μία μονάδα της μεταβλητής Index η ποσοστιαία αλλαγή της συνάρτησης κινδύνου είναι:

$$\exp(coef)\text{-}1 = 0.925\text{-}1 = -0.075$$

Δηλαδή αύξηση μιας μονάδας επί του ποσοστού των κυττάρων που προήλθαν από το μυελό των οστών οδηγεί σε μείωση κατά 7.5% της συνάρτησης κινδύνου.

- Για παράδειγμα, θεωρούμε τη μικρότερη ηλικία διάγνωσης που είναι 20 ετών και τη μεγαλύτερη που είναι 80. Η κινδυνότητα μεταξύ δύο ατόμων που οι ηλικίες διάγνωσης της ασθένειας είναι 80 και 20 αντίστοιχα είναι:

$$HR=e^{\hat{\beta}_1(X_1-X_2)}=1.034^{60} \cong 7$$

Δηλαδή, ο κίνδυνος να μην ανταποκριθεί κάποιος στη θεραπεία που η ασθένεια διαγνώστηκε σε ηλικία 80 ετών είναι 7 φορές μεγαλύτερος από τον κίνδυνο να μην ανταποκριθεί κάποιος, στον οποίο η ασθένεια διαγνώστηκε σε ηλικία 20 ετών. Ακόμη, συγκρίνοντας το μικρότερο ποσοστό των κυττάρων που προήλθαν από το μυελό των οστών 2 και το μεγαλύτερο που είναι 19, η κινδυνότητα μεταξύ δύο ατόμων με αντίστοιχα ποσοστά κυττάρων είναι:

$$HR=e^{\hat{\beta}_2(X_1-X_2)}= 0.925^{17} \cong 0.27$$

Ο κίνδυνος να μην ανταποκριθεί κάποιος στη θεραπεία, στον οποίο το ποσοστό των κυττάρων είναι 19 είναι κατά 0.27 φορές μεγαλύτερος από τον κίνδυνο να μην ανταποκριθεί κάποιος με ποσοστό των κυττάρων 2.

Τέλος, συμπεραίνουμε ότι όσο μικρότερη είναι η ηλικία διάγνωσης της ασθένειας τόσο πιθανότερη είναι η ανταπόκριση του ασθενή στη θεραπεία. Δηλαδή, οι ασθενείς με μεγαλύτερη ηλικία έχουν αυξημένο κίνδυνο και ως εκ τούτου, μικρότερο χρόνο επιβίωσης από τους ασθενείς με μικρότερη ηλικία. Όσον αφορά το ποσοστό των κυττάρων που προήλθαν από το μυελό των οστών, άτομα με μεγαλύτερο ποσοστό κυττάρων έχουν μειωμένο κίνδυνο, με αποτέλεσμα μεγαλύτερο χρόνο επιβίωσης σε σύγκριση με αυτούς που εμφανίζουν μικρές τιμές της μεταβλητής Index.

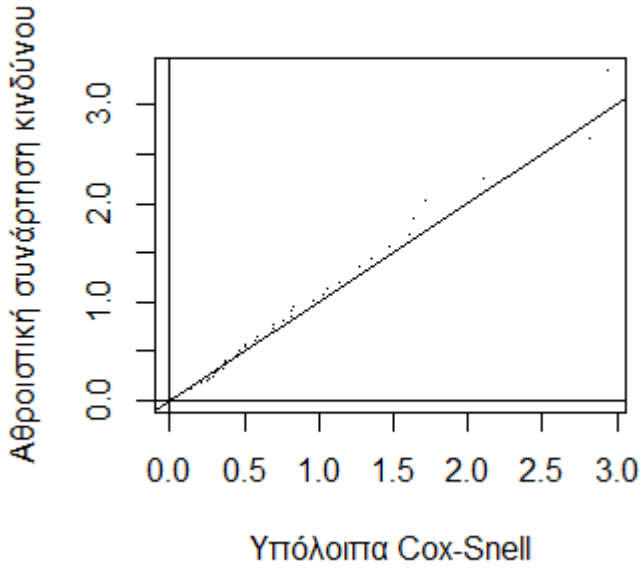
Το επόμενο στάδιο είναι η ανάλυση των υπολοίπων του μοντέλου με τη βοήθεια διαγνωστικών γραφημάτων για τον έλεγχο της υπόθεσης της αναλογικότητας, της ολικής επάρκειας του μοντέλου, για την εύρεση των outliers και τη συναρτησιακή μορφή μιας μεταβλητής.

Αρχικά, κατασκευάζεται το γράφημα των υπολοίπων Cox-Snell σε συνάρτηση με την αθροιστική συνάρτηση κινδύνου H(t) χρησιμοποιώντας τις παρακάτω εντολές:

```
> attach(cancer.cox2)

> rc<-abs(status-cancer.cox2$residuals)

> km.rc<-survfit(Surv(rc,status)~1)

> summary.km.rc<-summary(km.rc)

> rcu<-summary.km.rc$time

> surv.rc<-summary.km.rc$surv
```

> plot(rcu,-log(surv.rc),type="p",pch=".",xlab="Υπόλοιπα Cox-Snell",ylab="Αθροιστική συνάρτηση κινδύνου")

> abline(a=0,b=1);abline(v=0);abline(h=0)



Σχήμα 5.1: Γραφικός έλεγχος του μοντέλου για τα Cox-Snell υπόλοιπα

Από τα παραπάνω γράφημα συμπεραίνουμε ότι το τελικό μοντέλο, *cancer.cox2*, ταιριάζει στα δεδομένα. Τα δεδομένα βρίσκονται κοντά στην ευθεία που περνάει από την αρχή των αξόνων, γεγονός που μας οδηγεί στο συμπέρασμα ότι έχουμε καλή προσαρμογή.

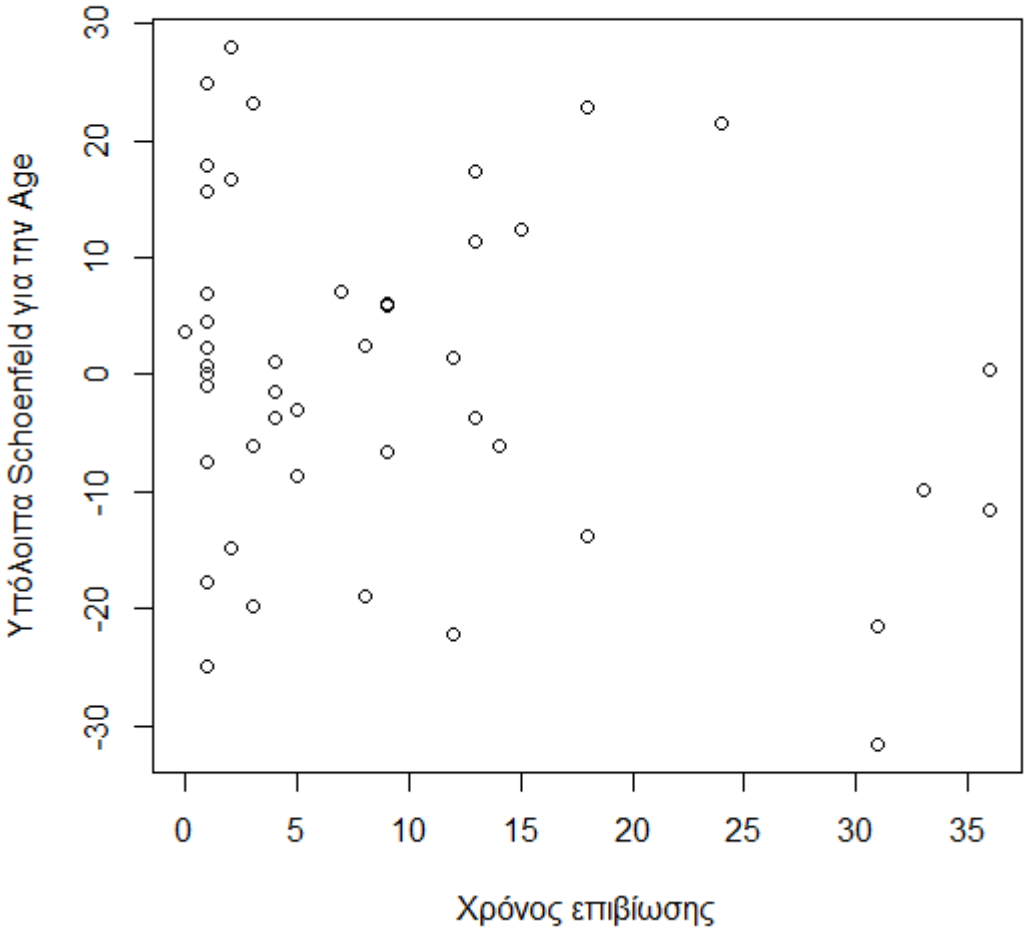
Στη συνέχεια, κατασκευάζεται το γράφημα των Schoenfeld υπολοίπων συναρτήσει του χρόνου επιβίωσης ξεχωριστά για κάθε μεταβλητή με τις παρακάτω εντολές:

>detail<-coxph.detail(cancer.cox2)

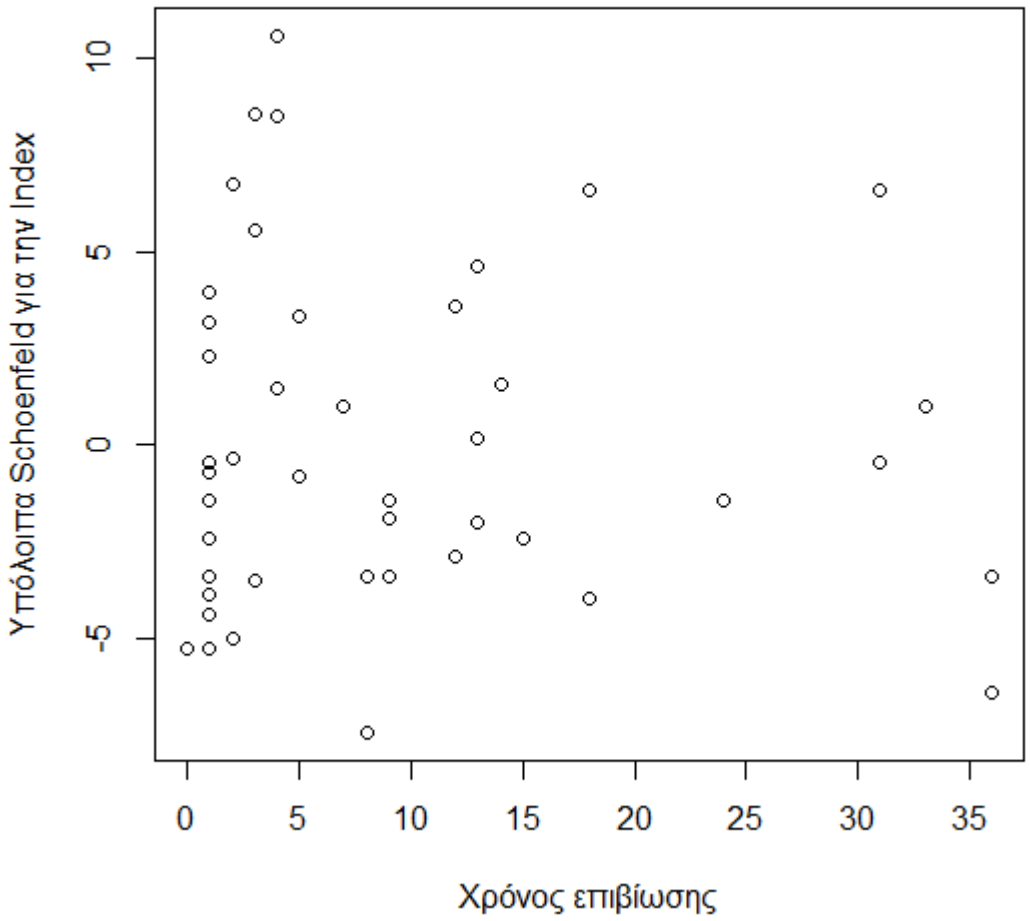
> sch<-resid(cancer.cox2,type="schoenfeld")

> plot(time[status==1],sch[,1],xlab="Χρόνος επιβίωσης", ylab="Υπόλοιπα Schoenfeld για την Age")

> plot(time[status==1],sch[,2],xlab="Χρόνος επιβίωσης", ylab="Υπόλοιπα Schoenfeld για την Index")



Σχήμα 5.2: Χρόνος επιβίωσης συναρτήσει των υπολοίπων Schoenfeld για την Age



Σχήμα 5.3: Χρόνος επιβίωσης συναρτήσει των υπολοίπων Schoenfeld για την Index

Από τα παραπάνω γραφήματα φαίνεται πως τα υπόλοιπα και των δύο μεταβλητών έχουν τυχαία μορφή σε συνάρτηση με το χρόνο. Άρα η υπόθεση της αναλογικότητας φαίνεται κι εδώ να ισχύει. Πρέπει να σημειώσουμε όμως ότι και στα δύο γραφήματα παρουσιάζεται ένα σημείο με μεγάλο χρόνο επιβίωσης το οποίο έχει μεγάλο σε απόλυτη τιμή residual, κάτι το οποίο δεν περιμένουμε να συμβαίνει αν η υπόθεση της αναλογικότητας ισχύει. Επειδή όμως πρόκειται για ενα μόνο σημείο σε κάθε γράφημα δεν θεωρούμε ότι παραβιάζεται η υπόθεση της αναλογικότητας.

Έπειτα, κατασκευάζονται τα γραφήματα των υπολοίπων martingale και deviance συναρτήσει της ηλικίας διάγνωσης και του ποσοστού των

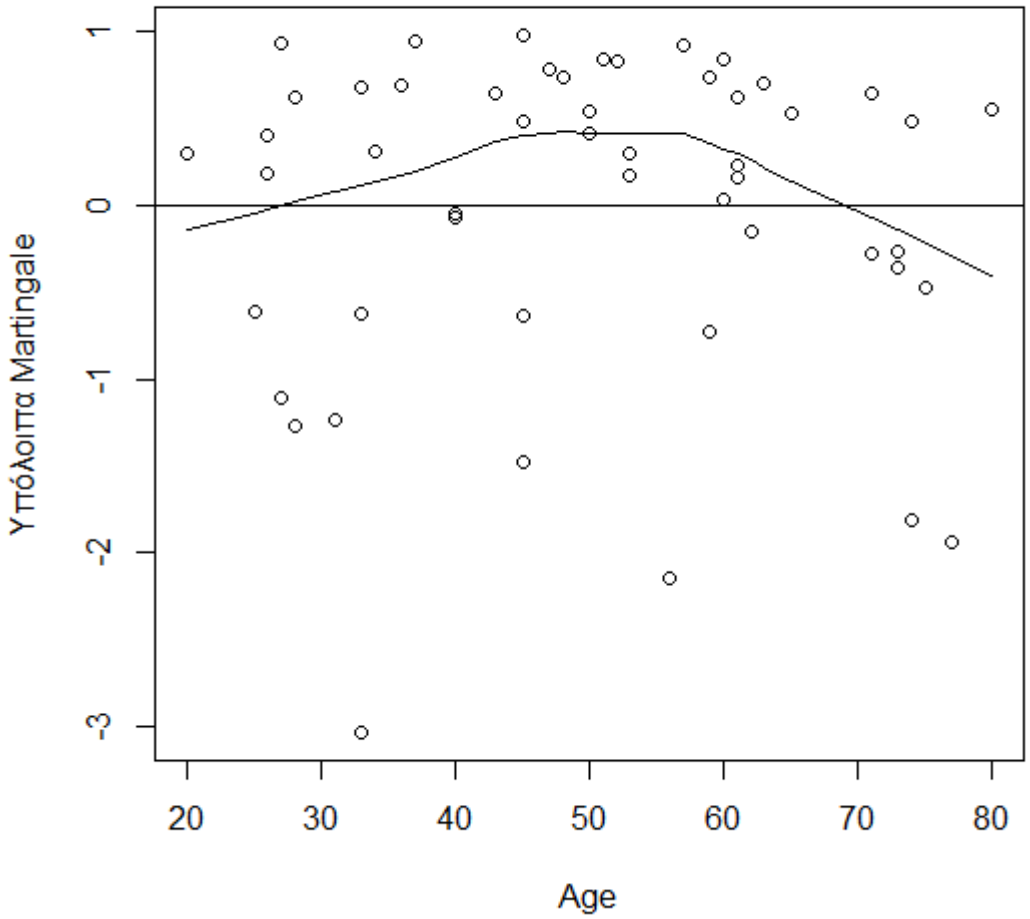
κυττάρων για την εύρεση της συναρτησιακής μορφής κάθε μεταβλητής, χρησιμοποιώντας τις παρακάτω εντολές αντίστοιχα:

> cancer.cox2.mart<-residuals(cancer.cox2,type="martingale")

> plot(Age,cancer.cox2.mart,ylab="Υπόλοιπα Martingale")

> abline(h=0)

>lines(lowess(Age,cancer.cox2.mart))

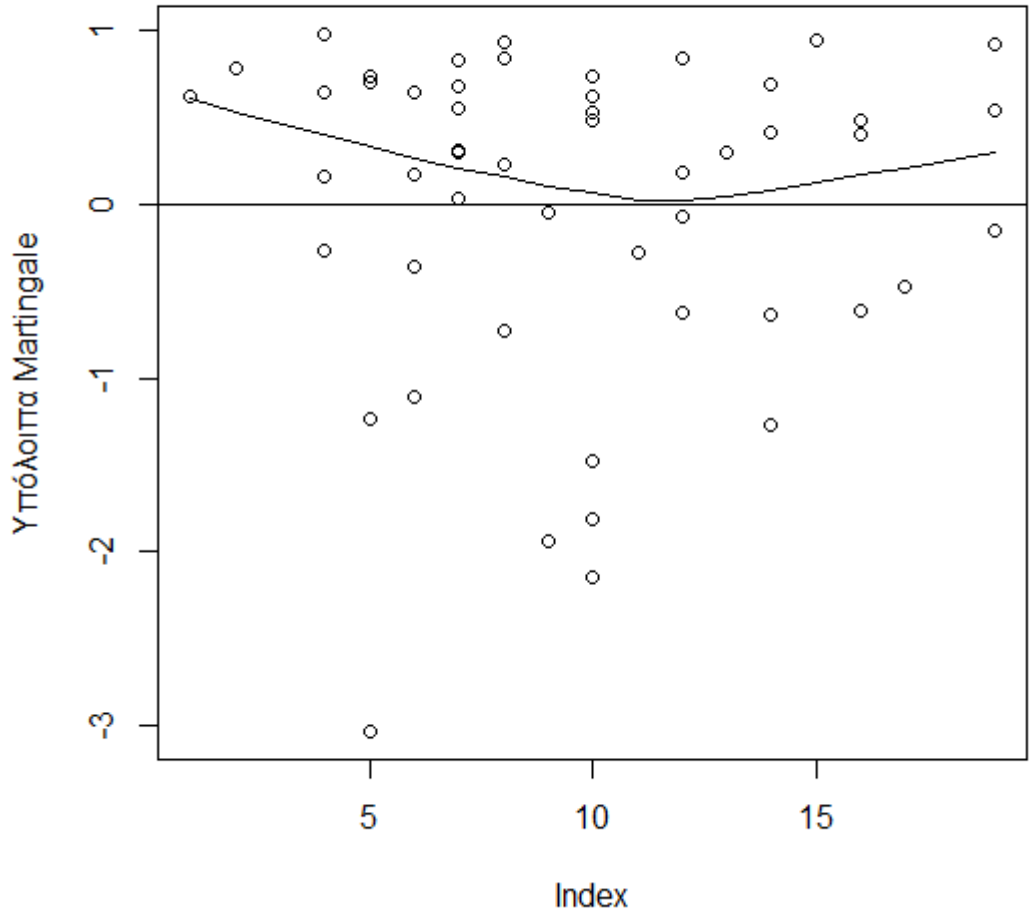


Σχήμα 5.4: Υπόλοιπα martingale για την μεταβλητή Age

> plot(Index,cancer.cox2.mart,ylab="Υπόλοιπα Martingale")

> abline(h=0)

> lines(lowess(Index,cancer.cox2.mart))



Σχήμα 5.5: Υπόλοιπα martingale για την μεταβλητή Index

Από τις γραφικές παραστάσεις 5.4 και 5.5, τα υπόλοιπα martingale και των δύο μεταβλητών δεν κατανέμονται συμμετρικά γύρω από το μηδέν. Μια γραμμική μορφή φαίνεται κατάλληλη για την κάθε μεταβλητή αφού οι καμπύλες είναι πολύ κοντά στην οριζόντια γραμμή στο 0.

Τέλος, κατασκευάζονται τα υπόλοιπα deviance για την κάθε μεταβλητή ξεχωριστά, τα οποία κατανέμονται περισσότερο συμμετρικά από τα martingale γύρω από το μηδέν και είναι χρήσιμα για την ανίχνευση outliers. Για την κατασκευή τους χρησιμοποιούνται οι παρακάτω εντολές:

> cancer.cox2.dev<-residuals(cancer.cox2,type="deviance")

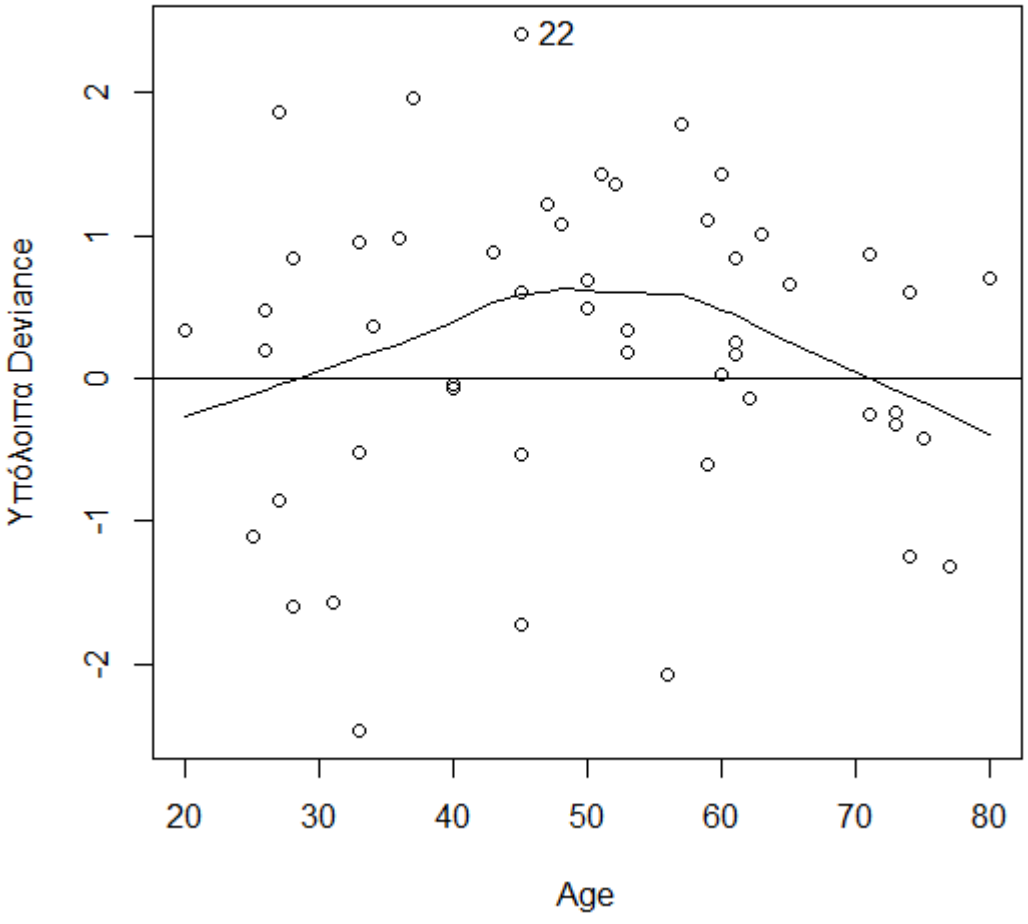
> plot(Age,cancer.cox2.dev,ylab="Υπόλοιπα Deviance")

> abline(h=0)

> lines(lowess(Age,cancer.cox2.dev))

> identify(Age,cancer.cox2.dev,n=1)

[1] 22



Σχήμα 5.6: Υπόλοιπα deviance για την μεταβλητή Age για την εύρεση outliers

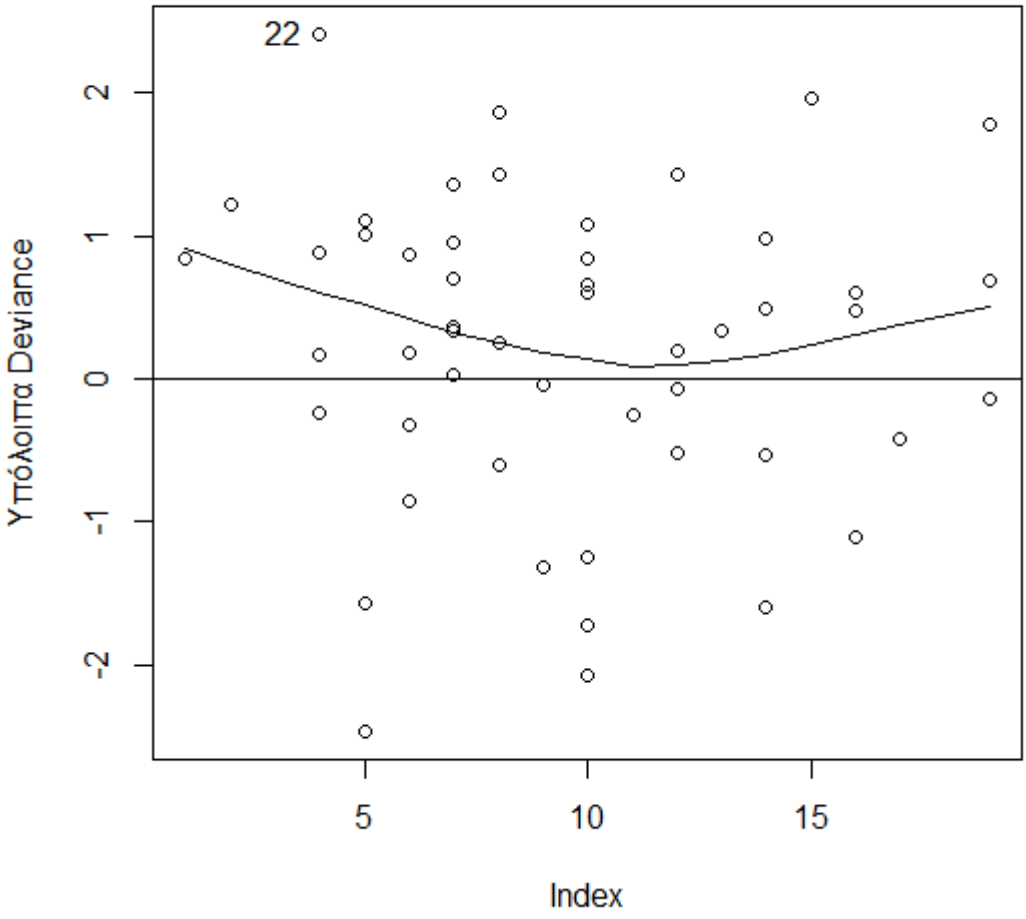
>plot(Index,cancer.cox2.dev,ylab="Υπόλοιπα Deviance")

> abline(h=0)

> lines(lowess(Index,cancer.cox2.dev))

> identify(Index,cancer.cox2.dev,n=1)

[1] 22



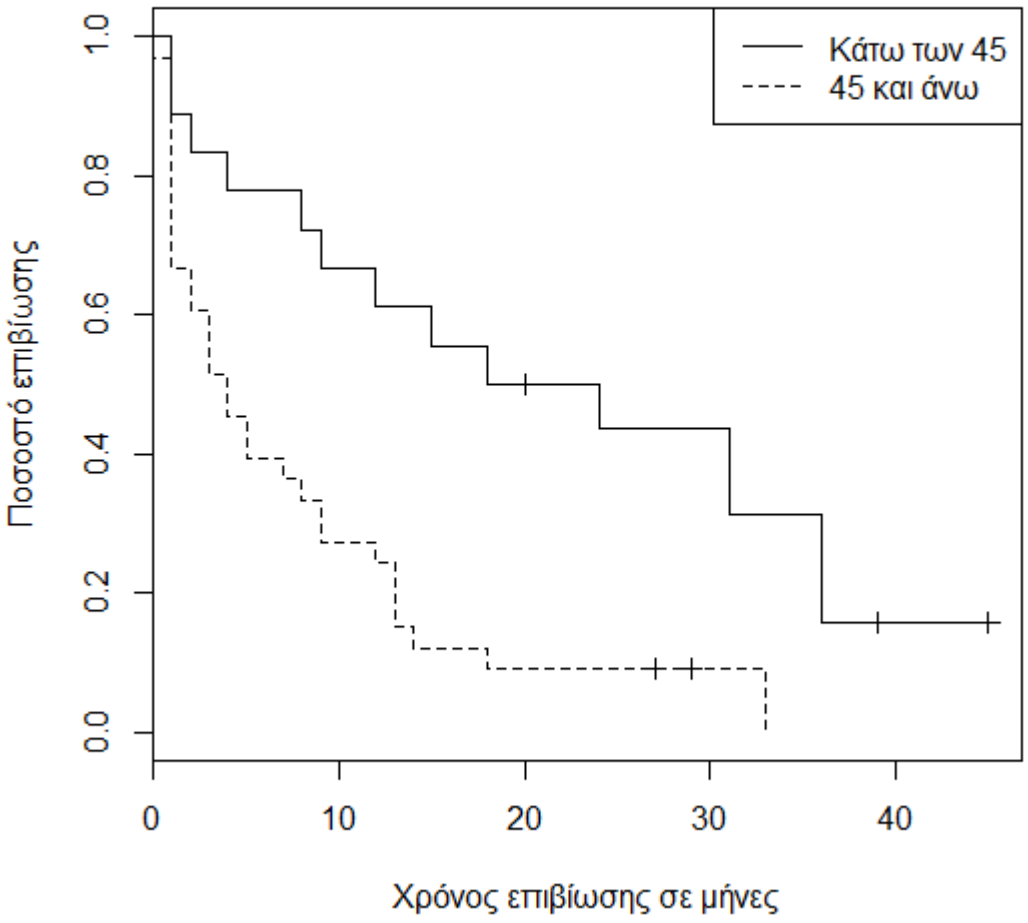
Σχήμα 5.7: Υπόλοιπα deviance για την μεταβλητή Index για την εύρεση outliers

Τα υπόλοιπα που αντιστοιχούν στην 22η παρατήρηση αντιστοιχούν σε απομακρυσμένη τιμή και στις δύο μεταβλητές. Πράγματι, το άτομο αυτό

έχει μηδενικό χρόνο επιβίωσης. Η παρατήρηση αυτή, είναι το μόνο πιθανό outlier με αποτέλεσμα να μην προκαλεί ανησυχία για την καταλληλότητα του μοντέλου. Όπως φαίνεται από τις γραφικές παραστάσεις, τα υπόλοιπα και στις δύο μεταβλητές κατανέμονται συμμετρικά γύρω από το μηδέν.

Σύμφωνα με την παραπάνω ανάλυση υπολοίπων, θα γίνει ανάλυση των δεδομένων εξετάζοντας τους ασθενείς βάσει της ηλικίας τους, χωρίζοντάς τους σε δύο κατηγορίες, μικρότερους και μεγαλύτερους από 45 ετών και έπειτα, βάσει του ποσοστού των κυττάρων τους, χωρίζοντάς τους σε αυτούς με ποσοστά μικρότερα και μεγαλύτερα από 10. Στο ίδιο γράφημα κατασκευάζεται το γράφημα των καμπυλών διαβίωσης των παραπάνω κατηγορίων με τις εξής εντολές:

```
> group<-cancer.dat[,"Age"]-45

> group[group>=0]<-1

> group[group<0]<-0

> plot(survfit(Surv(time,status)~ group),xlab="Χρόνος επιβίωσης σε μήνες",ylab="Ποσοστό επιβίωσης",lty=1:2)

> legend("topright",c("Κάτω των 45","45 και άνω"),lty=1:2)
```



Σχήμα 5.8: Καμπύλη επιβίωσης για τα άτομα με ηλικία διάγνωσης άνω και κάτω των 45

Από το γράφημα, φαίνεται πως υπάρχει σημαντική διαφορά στις καμπύλες των δύο κατηγοριών. Αυτό εξετάζεται με την παρακάτω εντολή:

> survdiff(Surv(time,status)~ group)

survdiff(formula = Surv(time, status) ~ group)

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
group=0	18	14	23.8	4.05	11

group=1 33 31 21.2 4.55 11

Chisq= 11 on 1 degrees of freedom, p= 0.000926

Η μηδενική υπόθεση H_0: $S_1(t) = S_2(t)$ απορρίπτεται καθώς η p-value είναι πολύ μικρότερη από το επίπεδο σημαντικότητας. Δηλαδή, οι δύο κατηγορίες δεν έχουν ίδια συνάρτηση επιβίωσης. Συγκεκριμένα, άτομα μικρότερης ηλικίας έχουν μεγαλύτερο χρόνο επιβίωσης. Επιπλέον, μπορούμε να εξετάσουμε ξεχωριστά την κάθε ομάδα με τις εντολές:

> c1<-coxph(Surv(time,status)~group+Index,data=cancer.dat)

> c1

coxph(formula = Surv(time, status) ~ group + Index, data = cancer.dat)

	coef	exp(coef)	se(coef)	z	p
group	1.2795	3.595	0.3687	3.47	0.00052
Index	-0.0925	0.912	0.0398	-2.32	0.02000

Likelihood ratio test=17 on 2 df , p=0.000206 n= 51, number of events= 45

Η κινδυνότητα μεταξύ των ατόμων με ηλικία μεγαλύτερη και μικρότερη των 45 είναι:

$$HR=\frac{\exp(1.2795*1)}{\exp(1.2975*0)}=3.595\cong 3.6$$

Αυτό σημαίνει πως ασθενείς μεγαλύτεροι των 45 έχουν περίπου 3.6 φορές μεγαλύτερο κίνδυνο, να έχουν μικρότερο χρόνο επιβίωσης, από εκείνους που είναι μικρότεροι των 45.

Για την Index έχουμε:

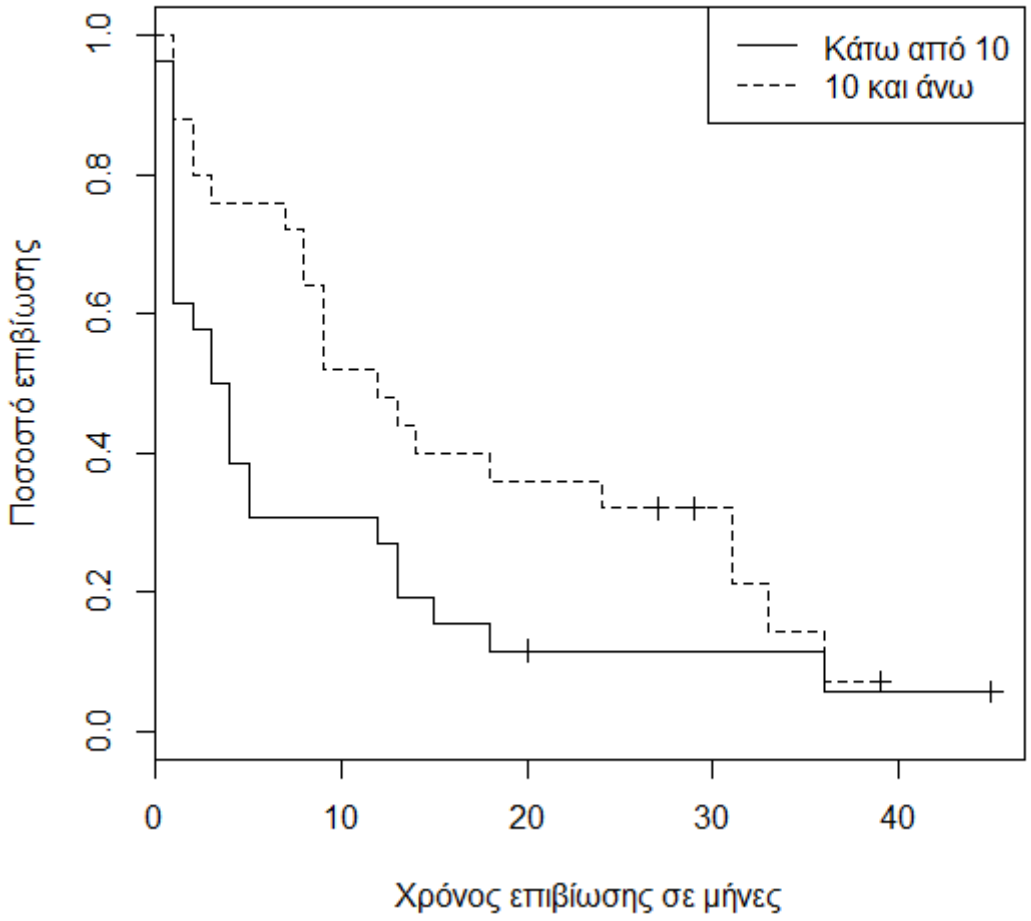
> group<-cancer.dat[,"Index"]-10

> group[group>=0]<-1

> group[group<0]<-0

> plot(survfit(Surv(time,status)~ group),xlab="Χρόνος επιβίωσης σε μήνες",ylab="Ποσοστό επιβίωσης",lty=1:2)

> legend("topright",c("Κάτω από 10","10 και άνω"),lty=1:2)



Σχήμα 5.9: Καμπύλη επιβίωσης για τα άτομα με ποσοστό κυττάρων άνω και κάτω του 10

Από το γράφημα, φαίνεται πως υπάρχει κάποια διαφορά στις καμπύλες των δύο κατηγοριών. Αυτό εξετάζεται με την εντολή:

> survdiff(Surv(time,status)~ group)

survdiff(formula = Surv(time, status) ~ group)

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
group=0	26	24	18	1.97	3.84
group=1	25	21	27	1.32	3.84

Chisq = 3.8 on 1 degrees of freedom, p= 0.0499

Ο έλεγχος δίνει τιμή για την ελεγχοσυνάρτηση ίση με 3.8 με 1 βαθμό ελευθερίας. Η p-value αν και είναι πολύ κοντά στο επίπεδο σημαντικότητας, είναι μικρότερη από αυτό με αποτέλεσμα η μηδενική υπόθεση H_0: $S_1(t) = S_2(t)$ να απορρίπτεται και να γίνεται δεκτό ότι οι δύο καμπύλες διαφέρουν. Ακόμη, έχουμε:

> c2<-coxph(Surv(time,status)~Age+group,data=cancer.dat)

> c2

coxph(formula = Surv(time, status) ~ Age + group, data = cancer.dat)

	coef	exp(coef)	se(coef)	z	p
Age	0.033	1.034	0.00946	3.49	0.00049
Group	-0.645	0.525	0.30694	-2.10	0.03600

Likelihood ratio test= 16.3 on 2 df , p=0.000294 n= 51, number of events= 45

Η κινδυνότητα μεταξύ των ατόμων με ποσοστό κυττάρων μικρότερο και μεγαλύτερο του 10 είναι:

$$HR=\frac{\exp(-0.645*0)}{\exp(-0.645*1)}=\frac{1}{0,525}\cong 2$$

Δηλαδή, ασθενείς με ποσοστό κυττάρων μικρότερο από 10 έχουν περίπου 2 φορές μεγαλύτερο κίνδυνο, να έχουν μικρότερο χρόνο επιβίωσης, από εκείνους με ποσοστό κυττάρων μεγαλύτερο από 10. Γενικά, άτομα με μεγαλύτερο ποσοστό κυττάρων έχουν μεγαλύτερο χρόνο επιβίωσης, κάτι το οποίο φαίνεται και από το γράφημα 5.9.

ΠΑΡΑΡΤΗΜΑ

Στο κεφάλαιο 5 χρησιμοποιήθηκαν τα παρακάτω δεδομένα για την στατιστική ανάλυση:

	Age	Smear	Infil	Index	Blasts	Resp	Time	Status
1	20	78	39	7	0.6	1	18	1
2	25	64	61	16	35.0	1	31	0
3	26	61	55	12	7.5	1	31	1
4	26	64	64	16	21.0	1	31	1
5	27	95	95	6	7.5	1	36	1
6	27	80	64	8	0.6	0	1	1
7	28	88	88	10	4.8	1	9	1
8	28	70	70	14	10.0	1	39	0
9	31	72	72	5	2.3	1	20	0
10	33	58	58	7	5.7	0	4	1
11	33	92	92	5	2.6	1	45	0
12	33	42	38	12	2.5	1	36	1
13	34	26	26	7	7.0	0	12	1
14	36	55	55	14	4.5	1	8	1
15	37	71	71	15	4.4	0	1	1
16	40	91	91	9	35.0	1	15	1
17	40	52	49	12	2.1	1	24	1
18	43	74	63	4	0.1	0	2	1
19	45	78	47	14	4.2	1	33	1
20	45	60	36	10	0.6	1	29	0
21	45	82	32	10	28.1	0	7	1
22	45	79	79	4	1.1	0	0	1

23	47	56	28	2	0.9	0	1	1
24	48	60	54	10	2.2	0	2	1
25	50	83	66	19	11.6	1	12	1
26	50	36	32	14	4.5	1	9	1
27	51	88	70	8	0.5	0	1	1
28	52	87	87	7	10.3	0	1	1
29	53	75	68	13	2.3	1	9	1
30	53	65	65	6	2.3	0	5	1
31	56	97	92	10	16.0	1	27	0
32	57	87	83	19	21.6	0	1	1
33	59	45	45	8	1.1	0	13	1
34	59	36	34	5	0.0	0	1	1
35	60	39	33	7	0.9	0	5	1
36	60	76	53	12	0.4	0	1	1
37	61	46	37	4	1.4	0	3	1
38	61	39	8	8	0.3	0	4	1
39	61	90	90	1	9.9	0	1	1
40	62	84	84	19	115.0	1	18	1
41	63	42	27	5	0.3	0	1	1
42	65	75	75	10	20.0	0	2	1
43	71	44	22	6	0.3	0	1	1
44	71	63	63	11	10.0	1	8	1
45	73	33	33	4	0.5	0	3	1
46	73	93	84	6	38.0	0	4	1
47	74	58	58	10	2.4	1	14	1
48	74	32	30	16	6.7	0	3	1
49	75	60	60	17	8.2	1	13	1

50	77	69	69	9	1.5	1	13	1
51	80	73	73	7	1.5	0	1	1

ΒΙΒΛΙΟΓΡΑΦΙΑ

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

Akaike, H. (1969), Fitting autoregressive models for prediction, *Ann. Inst. Statist. Math*., **21**, 243-247.

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle. Proc. 2nd Intern. Symp. on Information Theory (B. N. Petrov and F. Csaki, eds.), Akademia Kiado, Budapest, 267-281.

Akaike, H. (1974), A New Look at the Statistical Model Identification, *IEEE Trans. Automat. Contr*., AC-**19**, 716-723.

Alioum, A. and Commenges, D. (1996), A Proportional Hazards Model for Arbitrarily Censored and Truncated Data, *Biometrics* **52**, 512-524.

Allison, P. D. (1995), Survival Analysis Using SAS - A Practical Guide, SAS Publishing.

Atkinson, A.C. (1978), Simple Bayesian formula is misleading, *Biometrika*,**65**, 39-48.

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

Breslow, N. E. (1975), Analysis of survival data under the proportional hazards model, *Internat. Statist. Review*, **43**, 45-58.

Breslow, N. E. and Crowley, J.J. (1974), A large sample study of the life table and product limit estimates under random censorship, *Ann. Statist*., **2**, 437-453.

Collett, David. (2003), Modelling survival data in medical research (second edition), CRC Press.

Cox, D. R. (1972), Regression Models And Life Tables (with discussion), *Journal of the Royal Statistical Society*, Series B **34**, 187-220.

Cox, D R. (1968), A general definition of residuals (with discussion), *Journal of the Royal Statistical Society*, Series A **30**, 248-275.

Cox, D. R. & D. Oakes. (1984), Analysis of Survival Data. London: Chapman and Hall.

Efron, B. (1977), The efficiency of Cox's likelihood function for censored data, *J. Amer. Statist. Assoc*., **72**, 557-565.

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

Everitt, B. S. (2006), The Cambridge dictionary of statistics, Cambridge University Press.

Hald, A. (1949), Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point, *Scandinavian Actuarial Journal*, **32**, 119-134.

Helsel, D. (2010), Much ado about next to Nothing: Incorporating Nondetects in Science, *Ann. Occup. Hyg*., Vol. **54**, No. 3, pp. 257-262.

Hosmer, D. W. Jr. and Lemeshow, S. (1998), Applied Survival Analysis: Regression Modeling of Time to Event Data, Wiley Series in Probability and Statistics.

Hougaard, P., (2000), Analysis of Multivariate Survival Data, Springer.

Huber-Carol, C. and Vonta F. (2004), Frailty Models for Arbitrarily Censored And Truncated Data. *Lifetime Data Analysis*, **10**, 369-388.

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

Kaplan, E. L. and Meier, P. (1958), Nonparametric Estimation from Incomplete Observations, Journal of the American Statistical Association, **53**, 457-481.

Klein, J. P. and Moeschberger, M. L. (2003), Survival Analysis: Techniques for Censored and Truncated Data, 2nd edition, Springer.

Kleinbaum, D. G. (2005), Survival Analysis, A Self-Learning Text 2ed, Springer.

Lawless, J.F. (2003), Statistical Models and Methods for Lifetime Data, John Wiley & Sons.

Lee, E. T. (1992), Statistical Methods for Survival Data Analysis, John Wiley & Sons.

Mallows, C. L. (1973), Some comments on C_p, *Technometrics*, **15**, 661-676.

Mantel, N. and Haenszel, W. (1959), Statistical aspects of the analysis of data from retrospective studies of disease, *J. National Cancer Institute*, **22**, 719-322.

Marubini, E. and Valsecchi. M.G. (1995), Analysing Survival Data from Clinical Trials and Observational Studies, John Wiley & Sons.

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

O'Quigley, J. (2008), Proportional Hazards Regression, Springer.

Parmar, M. K. B. and Machin, D. (1995), Survival Analysis, A Practical Approach, John Wiley & Sons.

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

Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Statist*., **6**, 461-464.

Stevenson, M. (2009), Dependency Pattern Models for Information Extraction, Springer.

Tableman, M., Kim, J. S. and Portnoy, S. (2004). Survival Analysis Using S: Analysis of time to Event Data. Chapman & Hall/CRC.

Therneau, Terry M, Grambsch, Patricia M. (2000), Modelling Survival Data: Extending the Cox Model, Springer.

Turnbull, B. W. (1974), Nonparametric estimation of a survivorship function with doubly censored data, *J. Amer. Statist. Ass*., **69**, 169-173.

Turnbull, B.W. (1976), The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data, *Journal of the Royal Statistical Society*, Series B **38**, 290-295.

Vuong, Q. H. (1989), Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica.* (1989), Vol. **57**, No.2 ,307-333.

Κ. Φωκιανός & Χ. Χαραλάμπους. (2010), Εισαγωγή στην R-Πρόχειρες σημειώσεις, Τμήμα μαθηματικών & Στατιστικής Πανεπιστήμιο Κύπρου .

Χ.Καρώνη. (2009), Μοντέλα αξιοπιστίας και επιβίωσης, ΣΥΜΕΩΝ.

http://courses.washington.edu/b515/l17.pdf

http://www.biostat.sdu.dk/courses/e02/basalebegreber/bb_sur_e01sm.pdf

http://www.ats.ucla.edu/stat/examples/asa/test_proportionality.html

http://en.wikipedia.org/wiki/Censoring_%28statistics%29

http://esperia.iesl.forth.gr/~kafesaki/Applied-Mathematics/probabilities/p3.pdf

http://www.demog.berkeley.edu/213/Week14/welcome.pdf

http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-cox-regression.pdf

http://www.ida.liu.se/~kawah/Cox2.pdf

http://www.stat.cmu.edu/~acthomas/724/Efron-Morris.pdf

http://www.stern.nyu.edu/rengle/LagrangeMultipliersHandbook_of_Econ__II__Engle.pdf

http://www.public.iastate.edu/~vardeman/stat543/Handouts/wald-score-lrt.pdf

http://www.stern.nyu.edu/rengle/LagrangeMultipliersHandbook_of_Econ__II__Engle.pdf

http://en.wikipedia.org/wiki/Score_test

http://data.princeton.edu/pop509/NonParametricSurvival.pdf

http://ciser.cornell.edu/sasdoc/saspdf/stat/chap49.pdf

http://www.meduniwien.ac.at/imc/biometrie/publikationen/Separata/Nardi_Schemper_1999_Biometrics.pdf