ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Διπλωματική Εργασία

Στατιστικά Μοντέλα Επιβίωσης και Εφαρμογή στη Μελέτη του Καρκίνου του Πνεύμονα

Χατζίνης Γεώργιος

Επιβλέπουσα Καθηγήτρια: Καρώνη Χρυσηίς,

Επιτροπή καθηγητών:

Χ. Καρώνη,	Ι.Βόντα,	Β.Παπανικολάου,
Καθηγήτρια, ΕΜΠ,	Αν. Καθηγήτρια, ΕΜΠ	Καθηγητής, ΕΜΠ

Αθήνα, Οκτώβρης 2017

Ευχαριστίες

Η εκπόνηση της παρούσας διπλωματικής εργασίας πραγματοποιήθηκε υπό την επίβλεψη την Καθηγήτριας του Ε.Μ.Π. , κ. Χρυσηίς Καρώνη, την οποία θα ήθελα να την ευχαριστήσω θερμά για την δυνατότητα που μου έδωσε να ασχοληθώ με ένα θέμα το οποίο ανήκει στα ενδιαφέροντα μου αλλά και για την πολύτιμη βοήθεια που μου πρόσφερε καθόλη την διάρκεια εκπόνησης της εργασίας.

Επίσης θα ήθελα να ευχαριστήσω τους φίλους και τις φίλες που απέκτησα κατά την διάρκεια των φοιτητικών μου χρόνων και να τους θυμίσω τα ατελείωτα βράδια διαβάσματος που κάναμε βοηθώντας και στηρίζοντας ο ένας τον άλλον.

Τέλος το μεγαλύτερο ευχαριστώ το οφείλω στους γονείς και τον αδελφό μου για την στήριξη, την εμπιστοσύνη και την αγάπη που μου δείχνουν όλα αυτά τα χρόνια παρά τις δυσκολίες που προκύπταν.

Περίληψη

Σκοπός της παρούσας εργασίας είναι η ανάλυση δεδομένων ατόμων με καρκίνο του πνεύμονα με χρήση τεχνικών της ανάλυσης επιβίωσης. Στο πρώτο κεφάλαιο δίνονται οι βασικοί ορισμοί της ανάλυσης επιβίωσης καθώς και οι ορισμοί που αφορούν αποκομμένα δεδομένα. Στο δεύτερο κεφάλαιο παρουσιάζονται μη παραμετρικά μοντέλα ανάλυσης επιβίωσης και γίνεται αναφορά στην εκτιμήτρια Kaplan-Meier και στις μεθόδους σύγκρισης δύο καμπυλών επιβίωσης. Στο τρίτο κεφάλαιο περιγράφονται παραμετρικά μοντέλα επιβίωσης και δίνεται ιδιαίτερη έμφαση στο παραμετρικό μοντέλο Weibull. Στο τέταρτο κεφάλαιο παρουσιάζεται αναλυτικά το ημι παραμετρικό μοντέλο αναλογικής διακινδύνευσης του Cox καθώς και οι διάφορες παραλλαγές του. Επιπλέον περιγράφονται μέθοδοι που εξετάζουν αν ισχύει η υπόθεση αναλογικότητας κινδύνων όπως και με υπόλοιπα, που χρησιμοποιούνται για διάφορους ελέγχους που αφορούν την καταλληλόλητα του μοντέλου. Στο πέμπτο κεφάλαιο αναλύονται οι βασικές έννοιες των καμπυλών ROC και περιγράφεται η έννοια και η χρήση του εμβαδού κάτω από την καμπύλη ROC. Τέλος στο έκτο κεφάλαιο παρουσιάζεται το δείγμα και οι μεταβλητές που χρησιμοποιήσαμε και εφαρμόζονται οι παραπάνω τεχνικές σε δεδομένα καρκίνου του πνεύμονα.

Abstract

The aim of the present study is to analyze data corresponding to individuals with lung cancer using techniques of survival analysis. In the first chapter the basic definitions of survival analysis and censored data are presented. In the second chapter non-parametric models of survival analysis are presented and reference is made to the Kaplan-Meier estimator and methods of comparing two survival curves. The third chapter describes parametric survival models and emphasizes on the Weibull parametric model. In the fourth chapter we present the semi parametric Cox model and its various variants. Also, there is an examination of methods to consider whether the assumption of proportionality of hazards is met, as well residuals, used to check the suitability of the model. In the fifth chapter we analyze the basic concepts of the ROC curves and describe the meaning and use of the area under the ROC curve. Finally in the sixth chapter we present the sample and the variables we used and apply the techniques mentioned above to data that describe the survival time of individuals with lung cancer.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1 - ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΑΝΑΛΥΣΗ ΕΠΙΒΙΩΣΗΣ

1.1 Εισαγωγή

Η ανάλυση δεδομένων διάρκειας ζωής αποτελεί μια περιοχή της Στατιστικής με έντονη ερευνητική δραστηριότητα, με συνεχόμενη ανάπτυξη θεωρίας και μεθοδολογίας, ανταποκρινομένη έτσι στις πρακτικές απαιτήσεις διαφόρων επιστημών. Πολλές φορές μας ενδιαφέρει το πώς ένας παράγοντας κινδύνου επηρεάζει τον χρόνο σε ένα συγκεκριμένο γεγονός. Έτσι, για την ανάλυση δεδομένων τα οποία δεν μπορούν να επεξεργαστούν από τις συνηθισμένες στατιστικές μεθόδους δημιουργήθηκε η ανάλυση επιβίωσης (survival analysis).

Η ανάλυση επιβίωσης διάρκειας ζωής ασχολείται με τη μελέτη του χρόνου μέχρις ότου προκύψει ένα γεγονός και έχει εφαρμογές σε πολλές επιστημονικές περιοχές. Αρχικά αναφερόταν στο χρόνο μέχρι την αποθεραπεία ή τον θάνατο ασθενών για αυτό και πήρε αυτό το όνομα. Παρόλα αυτά έχει εφαρμογές σε πολλούς τομείς όπως μηχανολογία και οικονομία.

Σε γενικές γραμμές, η ανάλυση επιβίωσης σκοπό έχει την μοντελοποίηση του χρόνου. Οπότε, σε αυτό το πλαίσιο, ο θάνατος ή η αποτυχία θεωρείται ένα "συμβάν". Επομένως η ανάλυση επιβίωσης προσπαθεί να απαντήσει σε ερωτήματα όπως: ποιο είναι το ποσοστό του πληθυσμού που θα επιβιώσουν μετά από ένα ορισμένο χρονικό διάστημα; Ποιοι παράγοντες επηρεάζουν σημαντικά τον χρόνο την έκβαση του αποτελέσματος; Και άλλα πολλά σημαντικά ερωτήματα.

Σημαντικός ορισμός στην ανάλυση επιβίωσης είναι ο χρόνος επιβίωσης (survival time). Αναφέρεται σε μια μεταβλητή που μετράει το χρόνο (ημέρες, εβδομάδες, μήνες, κλπ.) που μεσολαβεί από την αρχή της παρακολούθησης ενός ατόμου (άνθρωπος, αντικείμενο, κλπ.), μέχρι τη στιγμή που το άτομο θα αντιμετωπίσει ένα ενδεχόμενο όπως π.χ. θάνατο, αποθεραπεία κλπ.

Ο χρόνος επιβίωσης χρήζει ιδιαίτερης μεταχείρισης για τον λόγο ότι περιορίζεται στο να είναι πάντα θετικός καθώς και στο ότι τα δεδομένα περιέχουν αποκομμένες (*censored*) παρατηρήσεις. Αυτό συνήθως συμβαίνει επειδή τα άτομα μπορεί να εισέρχονται στη μελέτη σε διαφορετικούς χρόνους, με συνέπεια ο χρόνος παρακολούθησης κάποιων ατόμων να μην είναι επαρκής ώστε να καταγραφεί ο χρόνος μέχρι την πραγματοποίηση του υπό μελέτη γεγονότος.

Αυτό που μας ενδιαφέρει για το χρόνο επιβίωσης είναι ο χαρακτηρισμός της κατανομής του χρόνου επιβίωσης, η σύγκριση αυτού του χρόνου μεταξύ διαφορετικών ομάδων καθώς και η μοντελοποίηση της σχέσης του χρόνου επιβίωσης σε σχέση με άλλες μεταβλητές. (Collett, 2003)

1.2 Αποκοπή Δεδομένων (Censoring)

Όπως αναφέρθηκε, είναι δυνατό τα δεδομένα να περιέχουν αποκομμένες παρατηρήσεις. Σε γενικές γραμμές, αποκομμένες παρατηρήσεις προκύπτουν όταν η εξαρτημένη μεταβλητή που μας αφορά αντιπροσωπεύει το χρόνο σε ένα τερματικό γεγονός, και η χρονική διάρκεια της έρευνας είναι περιορισμένη. Αν και η ιδέα αναπτύχθηκε στη βιοϊατρική έρευνα, οι αποκομμένες παρατηρήσεις μπορούν να συμβούν και σε άλλα είδη έρευνας.

Ειδικότερα, στην ανάλυση επιβίωσης, έχουμε αποκομμένα δεδομένα όταν υπάρχουν παρατηρήσεις, των οποίων οι χρόνοι επιβίωσης δεν είναι καθορισμένοι. Αν και δεν γνωρίζουμε την ακριβή διάρκεια ζωής ενός δείγματος, έχουμε την πληροφορία ότι έχει ξεπεράσει την χρονική διάρκεια κατά την οποία η μονάδα ήταν στο πείραμα. Αυτό συνήθως συμβαίνει όταν χρειαζόμαστε αποτελέσματα άμεσα για παράδειγμα σε μια μελέτη συγκεκριμένης ασθένειας δεν μπορούμε να περιμένουμε χρόνια μέχρις ότου όλα τα άτομα της μελέτης αποβιώσουν.

Είδη Αποκομμένων δεδομένων

Υπάρχουν τρία είδη αποκοπής δεδομένων. Η αποκοπή από δεξιά (right censoring), η αποκοπή από αριστερά (left censoring) και η αποκοπή κατά διάστημα (interval censoring). Οι δύο πρώτες είναι ειδικές περιπτώσεις της αποκοπής κατά διάστημα. Επομένως έχουμε:

- Αριστερά Αποκοπή

Σε αυτού του είδους παρατηρήσεις το μόνο που γνωρίζουμε είναι πως ο χρόνος επιβίωσης Τ, είναι μικρότερος από ένα χρονικό διάστημα. Ο ακριβής χρόνος επιβίωσης δεν είναι γνωστός.

- Δεξιά Αποκοπή

Είναι η πιο συνηθισμένη μορφή αποκοπής. Στην περίπτωση αυτή, ο χρόνος επιβίωσης Τ, είναι μεγαλύτερος από τον χρόνο που διαρκεί η μελέτη. Και εδώ δεν γνωρίζουμε τον ακριβή χρόνο επιβίωσης αλλά γνωρίζουμε μόνο ότι ο χρόνος επιβίωσης ανήκει σε ένα διάστημα με κατώτατο όριο το χρόνο που διαρκεί η μελέτη. Παρατηρείται σε περιπτώσεις όπου ένα άτομο χάνεται ή αποσύρεται από την παρακολούθηση ή και όταν η μελέτη τερματίζεται σε ένα προκαθορισμένο χρόνο.

- Αποκομμένο Διάστημα

Αυτά είναι τα δεδομένα για τα οποία γνωρίζουμε μόνο ότι βρίσκονται ανάμεσα σε ορισμένο ελάχιστο και μέγιστο όριο. Αποκομμένο διάστημα συνήθως προκύπτει όταν έχουμε ανακριβείς μετρήσεις ή όταν το πείραμα δεν είναι υπό συνεχή επίβλεψη.

Υπάρχουν δύο βασικοί μηχανισμοί αποκοπής παρατηρήσεων, ανεξάρτητοι της διάρκειας ζωής της μονάδας:

Αποκοπή τύπου I (Censoring Type I):

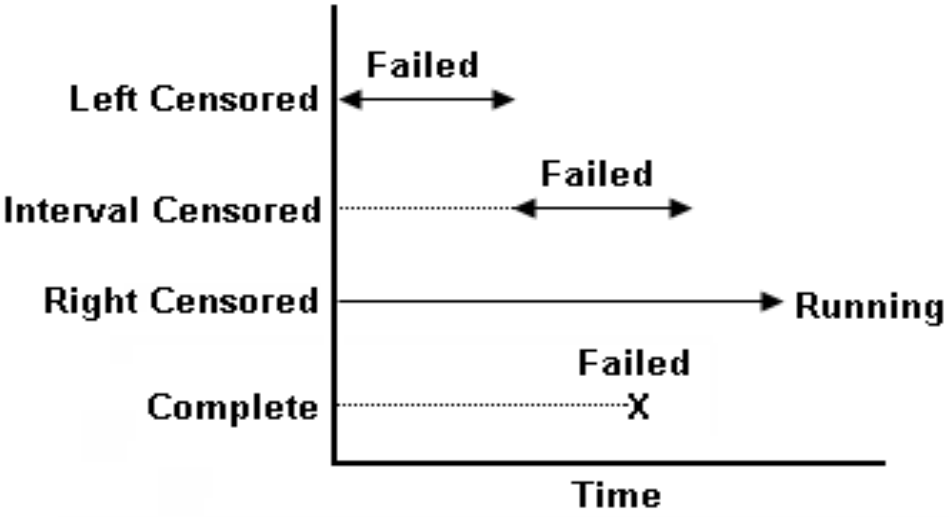
Η παρακολούθηση των μονάδων γίνεται για ένα συγκεκριμένο χρονικό διάστημα c.Γνωρίζουμε την ακριβή διάρκεια ζωής αν $T_i < c$,αλλιώς γνωρίζουμε ότι η διάρκεια ζωής έχει υπερβεί το c ($T_i > c$). Αυτό μπορεί να γενικευτεί ώστε κάθε μονάδα να έχει τον δικό της χρόνο παρακολούθησης $c_1, c_2, \ldots, c_n$. Βέβαια κάποια c_i ενδέχεται να είναι ίσα μεταξύ τους. Οι χρόνοι c_i είναι δεδομένοι, ενώ ο αριθμός των μονάδων που καταστρέφεται είναι τυχαίος .

Αποκοπή τύπου II (Censoring Type II):

Η παρακολούθηση του πειράματος διακόπτεται όταν καταστραφούν κ μονάδες. Εδώ το κ είναι προκαθορισμένο, ενώ η διάρκεια παρακολούθησης του πειράματος είναι τυχαία.

- Πλήρη Δεδομένα

Πλήρη στοιχεία ή δεδομένα είναι τα στοιχεία που η αξία της κάθε μονάδας του δείγματος παρατηρείται είτε είναι γνωστή. Τα πλήρη δεδομένα είναι πολύ πιο εύκολο να επεξεργαστούν από ότι τα αποκομμένα δεδομένα.
Το Σχήμα 1.1 συνοψίζει τις διαφορές μεταξύ πλήρη δεδομένων και τα διάφορα είδη των αποκομμένων δεδομένων.

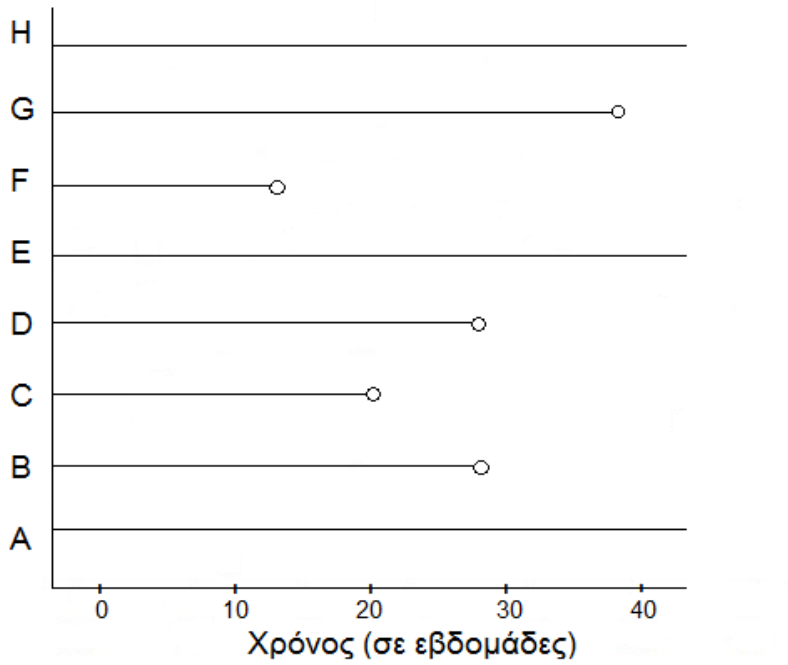


Σχήμα 1.1: Απεικόνιση όλων των ειδών αποκομμένων δεδομένων

Παράδειγμα 1

Θεωρούμε 8 ποντίκια που υποβάλλονται σε μια διαδικασία εμβολιασμού καρκινικών κυττάρων την ίδια χρονική στιγμή. Μας ενδιαφέρει ο χρόνος που απαιτείται ώστε τα ποντίκια να αναπτύξουν προκαθορισμένο όγκο. Ο ερευνητής αποφασίζει να τερματίσει το πείραμα μετά από 40 εβδομάδες. Από το Σχήμα 1.2 φαίνεται ότι οι

ποντικοί B,C,D,F,G ανέπτυξαν όγκο στους χρόνους 28, 20, 28, 15 και 38 αντίστοιχα (failure times) ενώ οι ποντικοί A, H, E δεν ανέπτυξαν όγκο που σημαίνει ότι οι χρόνοι επιβίωσης τους δεν είναι γνωστοί. Τα αποκομμένα δεδομένα στην περίπτωση αυτή είναι τύπου I.



Σχήμα 1.2: Χρόνοι αποκοπής των ποντικών για το παράδειγμα 1

1.3 Συναρτήσεις χρόνου επιβίωσης

Η κατανομή των χρόνων επιβίωσης, χαρακτηρίζεται από δύο συναρτήσεις, την συνάρτηση επιβίωσης (survival function) και την συνάρτηση κινδύνου (hazard function). Επιπλέον χρησιμοποιείται και η συνάρτηση πυκνότητας πιθανότητας (probability density function). Στην πράξη αυτές οι τρεις συναρτήσεις χρησιμοποιούνται για να επεξηγήσουν διαφορετικές όψεις των δεδομένων. Έστω T η διάρκεια ζωής, συνεχής τυχαία μεταβλητή με σ.π.π. $f(t), t \geq 0$. Ορίζουμε τη συνάρτηση κατανομής (σ.κ) ως

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

Α) Συνάρτηση επιβίωσης

Ορίζεται ως η πιθανότητα η διάρκεια ζωής να είναι μεγαλύτερη ενός χρόνου t και συμβολίζεται με $S(t)$. Άρα

$$S(t) = P[T \geq t] = 1 - F(t) = \int_0^\infty f(u)du \qquad (1.1)$$

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

Η συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής Τ βρίσκεται ως

$$f(t) = \frac{d}{dt}F(t) = -\frac{d}{dt}S(t)$$

και η καμπύλη τη $f(t)$ λέγεται καμπύλη πυκνότητας. Η αναμενόμενη διάρκεια ζωής (μέσος χρόνος επιβίωσης) βρίσκεται ως

$$\mu = E(t) = \int_0^\infty tf(t)dt = -\int_0^\infty t\frac{d}{dx}S(t)dt = \int_0^\infty S(t)dt$$

Β) Συνάρτηση Διακινδύνευσης

Η συνάρτηση διακινδύνευσης (hazard function) συμβολίζεται ως h(t) και εκφράζει την τάση προς διακοπή ενός αντικειμένου στο χρονικό διάστημα (t,t+δt) με δεδομένη την επιβίωση του μέχρι την χρονική στιγμή t.

Από τον ορισμό της δεσμευμένης πιθανότητας έχουμε ότι :

$$P[t < t \leq t + \delta t \,|\, T > t] = \frac{P[t < t \leq t + \delta t]}{P[T > t]} = \frac{S(t) - S(t + \delta t)}{S(t)}$$

Η συνάρτηση διακινδύνευσης ορίζεται ως

$$h(t) = \lim_{\delta t \to 0}\left(\frac{P[t < T \leq t + \delta t \,|\, T > t]}{\delta t}\right)$$

Άρα έχουμε εν τέλει ότι

$$h(t) = \lim_{\delta t \to 0} \left(\frac{S(t) - S(t+\delta t))/S(t)}{\delta t} \right) = \frac{f(t)}{S(t)}$$

Και εκφράζει το στιγμιαίο ρυθμό διακοπής. Επιπλέον η ποσότητα $h(t)\delta t$ είναι η υπό συνθήκη πιθανότητα της επικείμενης διακοπής μιας μονάδας δοθέντος ότι επέζησε μέχρι την συγκεκριμένη στιγμή t.

Μια ακόμα χρήσιμη συνάρτηση είναι η σωρευτική συνάρτηση διακινδύνευσης (σ.σ.δ) $H(t)$ η οποία είναι χρήσιμη για την επιλογή ενός κατάλληλου στατιστικού μοντέλου κατά την ανάλυση ενός συνόλου δεδομένων, που ορίζεται ως

$$H(t) = \int_0^t h(u)du = \int_0^t \frac{f(u)}{S(u)} du = \int_0^t \frac{S'(u)}{S(u)} du = [-\ln S(u)]_0^t = -\ln S(t)$$

Από τους ορισμούς προκύπτει ότι

$$H(t) = \int_0^t \frac{f(u)}{S(u)} du = \int_0^t \frac{S'(u)}{S(u)} du = [-\ln S(u)]_0^t = -\ln S(t) \qquad (1.2)$$

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

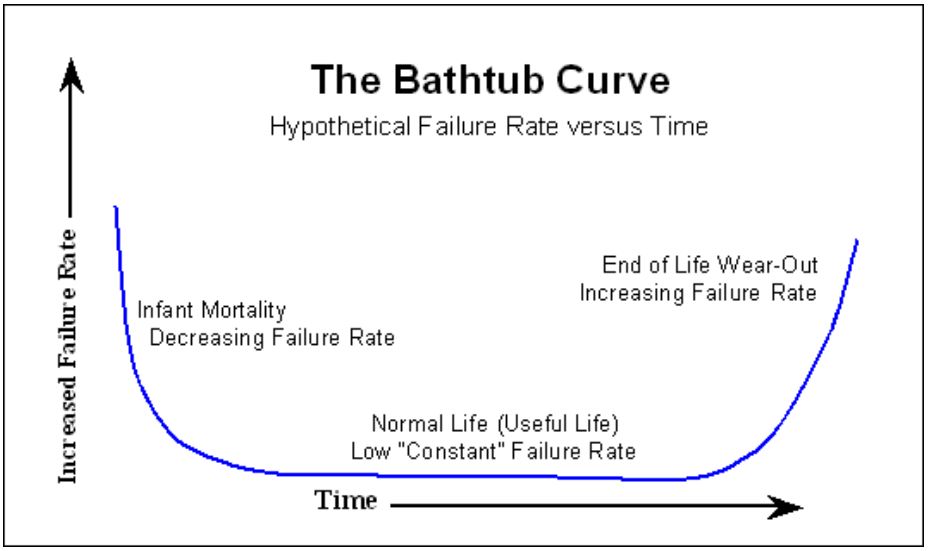
Από τους παραπάνω ορισμούς και σχέσεις είναι προφανές ότι οι συναρτήσεις $\left(f(t), h(t), F(t), S(t), H(t) \right)$ είναι μαθηματικά ισοδύναμες αφού γνωρίζοντας μια από αυτές μπορούν να βρεθούν οι υπόλοιπες τέσσερις. Μεγαλύτερη έμφαση δίνεται στην $h(t)$. Η συμπεριφορά της ανήκει σε μια από τις παρακάτω κατηγορίες.

- $h(t) = \lambda$ σταθερή $\Rightarrow H(t) = \lambda t$ και $S(t) = e^{-\lambda t}$ της Εκθετικής κατανομής με παράμετρο λ. Η ιδιότητα αυτού του μοντέλου είναι ότι ο στιγμιαίος ρυθμός διακοπής λειτουργίας μιας μονάδας είναι ανεξάρτητος της ηλικίας της ,δηλαδή δείχνει να μην γερνά.

- $h(t)$ αύξουσα συνάρτηση του t. Ο στιγμιαίος ρυθμός διακοπής αυξάνεται καθώς αυξάνεται η ηλικία της μονάδας, λογικά συνέπεια της γήρανσης.

- $h(t)$ φθίνουσα συνάρτηση του t. Εδώ μειώνεται ο στιγμιαίος ρυθμός διακοπής καθώς αυξάνεται η ηλικία. Η συμπεριφορά αυτή δεν είναι η αναμενόμενη παρά στην πρώτη φάση λειτουργίας μονάδων. Αυτό μπορεί να προκύψει όταν οι ελαττωματικές μονάδες τίθενται γρήγορα εκτός λειτουργίας, ενώ παραμένουν οι ποιοτικά καλύτερες μονάδες με αντίστοιχα χαμηλότερο ρυθμό διακοπής.

- $h(t)$ συνδέει όλες τις παραπάνω συμπεριφορές: ξεκινά με μια αρχική πτώση ως προς t, η οποία ακολουθείται από μια φάση σταθερότητας και στην συνέχεια καταλήγει σε ένα στάδιο αύξουσας διακινδύνευσης το σχήμα της καμπύλης αυτής καθιερώνει την ονομασία «διακινδύνευση μπανιέρας» (bathtub hazard). Αν και σε πολλές περιπτώσεις έχει την πιο ρεαλιστική συμπεριφορά, εν τούτοις δύσκολα μοντελοποιείται. (Χ.Καρώνη, 2009)



Σχήμα 3.1 : Συνάρτηση διακινδύνευσης «μπανιέρα»

ΚΕΦΑΛΑΙΟ 2- ΜΗ ΠΑΡΑΜΕΤΡΙΚΕΣ ΜΕΘΟΔΟΙ ΓΙΑ ΤΗΝ ΕΚΤΙΜΗΣΗ ΤΩΝ ΣΥΝΑΡΤΗΣΕΩΝ ΕΠΙΒΙΩΣΗΣ

2.1 Εκτιμήτρια Kaplan -Meier

Στο κεφάλαιο αυτό, κάνουμε λόγο για τις μεθόδους εκτίμησης των τριών συναρτήσεων επιβίωσης που αναπτύξαμε στο προηγούμενο κεφάλαιο, για αποκομμένα δεδομένα. Η χρήση αρχικά, μη παραμετρικών μεθόδων για ανάλυση δεδομένων γίνεται λόγω του ότι είναι πιο εύκολες τόσο στην εφαρμογή τους όσο και στην κατανόηση τους. Επιπλέον είναι περισσότερο αποδοτικές όταν δεν υπάρχουν γνωστές θεωρητικές κατανομές.

Από τις τρεις συναρτήσεις, χρησιμοποιείται περισσότερο η συνάρτηση επιβίωσης και η γραφική της παράσταση, η καμπύλη επιβίωσης. Επομένως η διαδικασία επιλογής ενός μοντέλου με την βέλτιστη προσαρμογή στα δεδομένα, ξεκινά από την κατασκευή των γραφικών παραστάσεων οι οποίες δείχνουν την συμπεριφορά των συναρτήσεων επιβίωσης και στην συνέχεια της διακινδύνευσης.

Στην ανάλυση επιβίωσης, όπως προαναφέραμε παρουσιάζεται πολύ συχνά αποκοπή δεδομένων, και συγκεκριμένα δεξιά αποκοπή, όταν δηλαδή π.χ. ένα άτομο ασθενής αποχωρεί από μια έρευνα. Σε αυτές τις περιπτώσεις χρησιμοποιούμε την μέθοδο Kaplan-Meier για την εκτίμηση της συνάρτησης επιβίωσης. Η μέθοδος αναπτύχθηκε από τους Kaplan και Meier το 1958, και προϋποθέτει τα εξής:

- Τα άτομα που χάθηκαν από τη διαδικασία παρακολούθηση να έχουν την ίδια πιθανότητα επιβίωσης με τα άτομα που συνεχίζουν να συμμετέχουν στη διαδικασία παρακολούθησης. Αυτό δεν μπορεί να ελεγχθεί και ελλοχεύει τον κίνδυνο να υπάρχει μεροληψία που μειώνει τη συνάρτηση S(t).
- Οι πιθανότητες επιβίωσης είναι ίδιες για άτομα που εισήλθαν στην αρχή της μελέτης με των ατόμων που εισήλθαν πιο αργά στη μελέτη.

- Το ενδεχόμενο που μελετάται (π.χ. θάνατος) συμβαίνει στον καθορισμένο χρόνο. Καθυστερημένη καταγραφή του γεγονότος, θα προκαλέσει αύξηση του S(t).
  http://www.statsdirect.com/help/survival_analysis/kaplan_meier.htm

Ένα αρχικό βήμα για την ανάλυσης ενός συνόλου δεδομένων επιβίωσης είναι να παρουσιάσουμε κάποια αριθμητικά αλλά και γραφικά αποτελέσματα για του χρόνους επιβίωσης των ατόμων που βρίσκονται σε ένα γκρουπ. Τα αποτελέσματα αυτά προκύπτουν εύκολα από τις εκτιμήσεις των συναρτήσεων επιβίωσης και διακινδύνευσης. Αυτές οι μέθοδοι μη παραμετρικοί καλούνται, αφού δεν απαιτούν συγκεκριμένες υποθέσεις σχετικά με την κατανομή που ακολουθούν οι χρόνοι επιβίωσης. Όταν βρεθεί η εκτίμηση της συνάρτησης επιβίωσης, η διάμεσος καθώς και άλλα μέτρα της κατανομής των χρόνων επιβίωσης, μπορούν να υπολογιστούν. Όταν θέλουμε να συγκρίνουμε δύο διαφορετικά γκρουπ ασθενών, μπορούμε να πραγματοποιήσουμε μια άτυπη σύγκριση της επιβίωσης για κάθε γκρουπ χρησιμοποιώντας τις εκτιμήσεις των συναρτήσεων επιβίωσης. Αυτός ο έλεγχος ονομάζεται log-rank.

Εκτίμηση της συνάρτησης επιβίωσης

Υποθέτουμε αρχικά ότι έχουμε ένα δείγμα χρόνων επιβίωση όπου καμία από τις παρατηρήσεις μας δεν είναι αποκομμένη. Η συνάρτηση επιβίωσης $S(t)$ ορίζεται από την εξίσωση (1.1) είναι η πιθανότητα ένα άτομο να επιβιώσει για ένα χρόνο μεγαλύτερο ή ίσο του t. Αυτή η συνάρτηση μπορεί να υπολογιστεί από τον παρακάτω τύπο

$$\hat{S}(t) = \frac{\alpha\rho\iota\theta\mu\acute{o}\varsigma\ \ \alpha\tau\acute{o}\mu\omega\nu\ \ \mu\varepsilon\ \ \chi\rho\acute{o}\nu o\ \ \varepsilon\pi\iota\beta\acute{\iota}\omega\sigma\eta\varsigma > t}{n} \qquad (2.1)$$

Όπου n ο αριθμός όλων των ατόμων. Ισοδύναμα $\hat{S}(t) = 1 - \hat{F}(t)$ όπου $\hat{F}(t)$ είναι η εμπειρική συνάρτηση κατανομής που ορίζεται ως ο λόγος του συνολικού αριθμού των ατόμων που είναι εν ζωή την χρονική στιγμή t προς το σύνολο των ατόμων που μελετάμε. Παρατηρούμε ότι η εμπειρική συνάρτηση κατανομής είναι ίση με μονάδα για τις τιμές του t πριν τον πρώτο θάνατο, κι μηδέν μετά τον τελευταίο

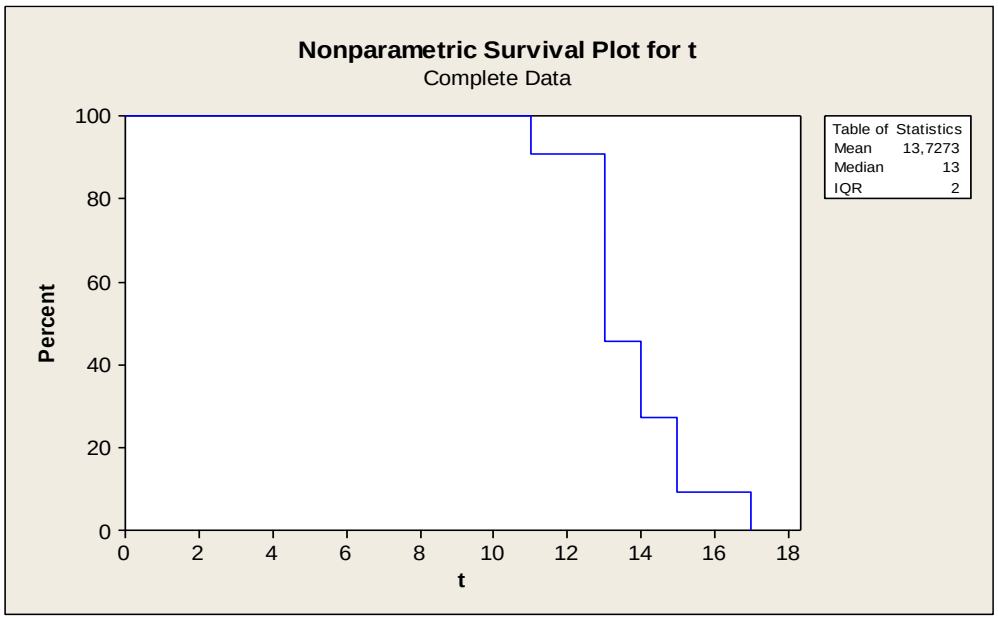
χρονικά θάνατο. Η εκτιμώμενη συνάρτηση επιβίωσης $\hat{S}(t)$ υποθέτουμε ότι είναι σταθερή ανάμεσα σε δύο συνεχόμενους θανάτους και άρα το γράφημα της σε σχέση με το t είναι μια κατά βήματα συνάρτηση η οποία φθίνει αμέσως μετά από κάθε παρατηρούμενο χρόνο επιβίωσης.

Παράδειγμα 2 πνευμονική μετάσταση

Ένα πρόβλημα που αντιμετωπίζουν οι ασθενείς που έχουν οστεοσάρκωμα είναι ότι συχνά ο όγκος εξαπλώνεται και στους πνεύμονες κάτι που μπορεί να οδηγήσει στον θάνατο. Σε μια σχετική μελέτη σε τέτοια άτομα καταγράφηκαν οι χρόνοι επιβίωσης σε μήνες έντεκα ασθενών.

11	**13**	**13**	**13**	**13**	**13**	**14**	**14**	**15**	**15**	**17**

Χρησιμοποιώντας την εξίσωση (2.1) οι εκτιμώμενες τιμές της συνάρτησης επιβίωσης στους παραπάνω χρόνους είναι 1.000, 0.909, 0.455, 0.273 και 0.091. Η εκτιμώμενη τιμή της συνάρτηση επιβίωσης είναι μονάδα μέχρι τους 11 πρώτους μήνες και μηδέν μετά τους 17. Στο Σχήμα 4.1 φαίνεται η εκτίμηση της συνάρτησης επιβίωσης. (Collett, 2003)



Σχήμα 4.1: Εκτίμηση συνάρτησης επιβίωσης για το Παράδειγμα 2

Αυτή η μέθοδος εκτίμησης της συνάρτηση επιβίωσης που παρουσιάστηκε παραπάνω δεν μπορεί να χρησιμοποιηθεί όταν υπάρχουν αποκομμένες παρατηρήσεις. Ο λόγος είναι ότι η μέθοδος δεν λαμβάνει πληροφορία για ένα άτομο που ο χρόνος επιβίωσης του είναι αποκομμένος πριν την χρονική στιγμή t.

Επειδή το φαινόμενο της αποκοπής δεδομένων παρουσιάζεται πολύ συχνά και επειδή η αποκοπή συνηθίζεται αν είναι από δεξιά, η εκτιμήτρια Kaplan-Meier της συνάρτησης επιβίωσης που χρησιμοποιείται σε αυτές τις περιπτώσεις έχει αποκτήσει μεγάλη σημασία. (Kaplan-Meier, 1958)

Έστω τυχαίο δείγμα n μονάδων ή ατόμων με παρατηρούμενους χρόνους επιβίωσης $t_1, t_2, ..., t_n$. Κάποιες από αυτές τις παρατηρήσεις μπορεί να είναι δεξιά αποκομμένες αλλά μπορεί επίσης περισσότεροι από ένα άτομο να έχουν τον ίδιο χρόνο επιβίωσης. Υποθέτουμε ότι συμβαίνουν r θάνατοι ανάμεσα στα άτομα, όπου φυσικά $r \le n$, και κατατάσσουμε σε αύξουσα σειρά τους χρόνους απεβίωσης $t_{(j)}$ για $j = 1, 2, ..., r$ δηλαδή $t_{(1)} < t_{(2)} < ... < t_{(r)}$. Ο αριθμός των ατόμων που θα είναι εν ζωή πριν την χρονική στιγμή $t_{(j)}$ συμπεριλαμβανομένων και αυτών που θα πεθάνουν αυτή την χρονική στιγμή θα είναι n_j για $j = 1, 2, ..., r$ και d_j ο αριθμός που απεβίωσαν εκείνη την χρονική στιγμή. Το χρονικό διάστημα από $t_{(j)} - \delta$ ως $t_{(j)}$ όπου δ πολύ μικρό, περιλαμβάνει ένα χρόνο θανάτου. Δεδομένου ότι υπάρχουν n_j άτομα που είναι εν ζωή ακριβώς πριν την χρονική στιγμή $t_{(j)}$ και d_j θάνατοι την χρονική στιγμή $t_{(j)}$, η πιθανότητα να πεθάνει κάποιο άτομο μέσα στο διάστημα $t_{(j)} - \delta$ έως $t_{(j)}$ υπολογίζεται ως d_j / n_j. Η αντίστοιχη εκτιμώμενη πιθανότητα επιβίωσής στο διάστημα αυτό είναι $(n_j - d_j) / n_j$.

Υποθέτουμε τώρα ότι οι χρόνοι που πεθαίνουν τα άτομα είναι ανεξάρτητοι μεταξύ τους. Τότε η εκτιμωμένη συνάρτηση επιβίωσης οποιαδήποτε χρονική στιγμή t, μέσα στο κ-οστό διάστημα από $t_{(\kappa)}$ έως $t_{(\kappa+1)}$ $\kappa = 1, 2, ..., r$, όπου το $t_{(r+1)}$ ορίζεται άπειρο, θα είναι η εκτιμώμενη πιθανότητα επιβίωσης μετά από το $t_{(\kappa)}$. Αυτή είναι

δηλαδή η πιθανότητα επιβίωσης μέσα στο διάστημα $t_{(\kappa)}$ έως $t_{(\kappa+1)}$ και όλων των προηγούμενων διαστημάτων που δίνει τελικά την εκτιμήτρια Kaplan-Meier

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

Για $t_{(k)} \le t < t_{(k+1)}$, $k = 1, 2, ..., r$, με $\hat{S}(t) = 1$ αν $t < t_{(1)}$ δηλαδή αν δεν έχουμε κανένα θάνατο $d_j = 0$. Άρα ο εκτιμητής Kaplan-Meier μεταβάλλεται μόνο στις χρονικές στιγμές που συμβαίνει το γεγονός, οπότε στον υπολογισμό του εκτιμητή οι χρονικές στιγμές που δεν συμβαίνει το γεγονός παραλείπονται. (Collett, 2003)

2.2 Διάμεσος χρόνος επιβίωσης

Αφού έχουμε υπολογίσει την εκτιμήτρια της συνάρτηση επιβίωσης, μπορούμε εύκολα υπολογίσουμε και τον διάμεσο χρόνο επιβίωσης. Ο διάμεσος χρόνος επιβίωσης είναι ο χρόνος στον οποίο το 50% των ατόμων του υπό μελέτη δείγματος επιβιώνει και δίνεται από την τιμή $t(50)$ το οποίο είναι $S\{t(50)\} = 0.5$. Επειδή όμως οι μη-παραμετρικές εκτιμήσεις της $S(t)$ είναι κλιμακωτές συναρτήσεις, δεν θα είναι εφικτό να βρούμε ένα χρόνο επιβίωσης που να κάνει την συνάρτηση επιβίωσης ακριβώς 0.5 για αυτό τον λόγο ορίζουμε την $\hat{t}(50)$

$$\hat{t}(50) = \min\left\{ t_i \,\middle|\, \hat{S}(t_i) < 0.5 \right\}$$

Όπου t_i είναι ο χρόνος επιβίωσης του ι-οστού ατόμου $i = 1, 2, ..., n$. Και επειδή η εκτιμώμενη συνάρτηση επιβίωσης αλλάζει μόνο όταν υπάρχει κάποιος θάνατος ισοδύναμα ορίζουμε

$$\hat{t}(50) = \min\left\{ t_{(j)} \,\middle|\, \hat{S}(t_{(j)}) < 0.5 \right\}$$

Όπου $t_{(j)}$ είναι ο j-οστός θάνατος , $j = 1, 2, ..., r$.

Στην ειδική περίπτωση όπου η εκτιμώμενη συνάρτηση επιβίωσης είναι ακριβώς ίση με 0.5 για τις τιμές του t στο διάστημα $t_{(j)}$ $t_{(j+1)}$ η διάμεσος θα είναι το μεσαίο στοιχείο του διαστήματος δηλαδή $(t_{(j)} + t_{(j+1)})/2$.

Όταν δεν υπάρχουν αποκομμένα δεδομένα, ο διάμεσος χρόνος επιβίωσης μπορεί να βρεθεί από την καμπύλη επιβίωσης, βρίσκοντας την τιμή του χρόνου για την οποία ισχύει $S(t) = 0.5$. (Collett, 2003)

2.3 Εκτιμήτρια Nelson-Aalen της σωρευτικής συνάρτησης διακινδύνευσης

Η σωρευτική συνάρτηση διακινδύνευσης μπορεί να εκτιμηθεί από την $\hat{S}(t)$ μέσω της σχέσης

$$\hat{H}(t) = -\ln \hat{S}(t) = -\sum_{j:t_{(j)} \le t} \ln\left(1 - \frac{d_j}{n_j}\right),$$

αλλά προτιμότερη εκτιμήτρια της $H(t)$ αποτελεί η εκτιμήτρια Nelson-Aalen (Nelson, 1972, Aalen, 1978)

$$\hat{H}(t) = \begin{cases} \sum_{j:t_{(j)} \le t} \frac{d_j}{n_j}, & \text{όταν } t \ge t_{(1)} \\ 0, & \text{όταν } t < t_{(1)} \end{cases}$$

Η εκτιμήτρια αυτή μπορεί να δικαιολογηθεί ως εξής.

Επειδή $H(t) = -\ln S(t)$

Έχουμε ότι $\hat{H}(t) = -\ln \hat{S}(t) = -\sum_{j:t_{(j)} \le t} \ln\left(1 - \frac{d_j}{n_j}\right)$

Και χρησιμοποιώντας την εκτιμήτρια Kaplan-Meier της $S(t)$ Και θεωρώντας το d_j / n_j μικρό, που ισχύει για τις πρώτες τουλάχιστον διακοπές παίρνουμε

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

Μια εκτίμηση της διασποράς της εκτιμήτριας Nelson-Aalen είναι η

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

Η εκτιμήτρια Nelson-Aalen είναι επίσης μια βαθμιδωτή ή κλιμακωτή συνάρτηση. Μια εναλλακτική εκτίμηση λοιπόν της συνάρτησης επιβίωσης, που βασίζεται μεμονωμένους χρόνους συμβάντων είναι η εκτιμήτρια Nelson-Aalen η οποία δίνεται από τον τύπο

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

Αυτή η εκτίμηση λαμβάνεται από την σωρευτική συνάρτηση διακινδύνευσης όπως δείξαμε από τις σχέσεις που τις συνδέουν.(Χ.Καρώνη, 2009)

2.4 Τυπικό σφάλμα και Διάστημα εμπιστοσύνης της συνάρτησης επιβίωσης

Μια σημαντική βοήθεια στην ερμηνεία μιας εκτίμησης μιας οποιαδήποτε ποσότητας είναι η ακρίβεια της εκτίμησης, δηλαδή το τυπικό σφάλμα. Ορίζεται ως η τετρα-γωνική ρίζα της εκτιμώμενης διασποράς της εκτίμηση και χρησιμοποιείται για την κατασκευή διαστημάτων εμπιστοσύνης.

Ο εκτιμητής Kaplan-Meier της συνάρτηση επιβίωσης για κάθε τιμή του χρόνου t στο διάστημα $(t_{(k)}, t_{(k-1)}]$ μπορεί να γραφτεί ως

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

για $k = 1, 2, \ldots, r$ όπου $\hat{p}_j = (n_j - d_j)/n_j$ είναι η εκτιμώμενη πιθανότητα ένα άτομο να επιβίωση στον χρόνο του διαστήματος που ξεκινά την χρονική στιγμή $t_{(j)}$, $j = 1, 2, \ldots, r$. Παίρνοντας λογαρίθμους

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

Και άρα η διασπορά του $\log \hat{S}(t)$ δίνεται από τον τύπο

$$\text{var}\left\{\log \hat{S}(t)\right\} = \sum_{j=1}^{k} \text{var}\left\{\log \hat{p}_j\right\} \qquad (2.2)$$

Θεωρούμε τώρα ότι ο αριθμός των ατόμων που επιβιώνουν στο διάστημα που ξεκινά την χρονική στιγμή $t_{(j)}$, ακολουθεί την Διωνυμική κατανομή με παραμέτρους n_j και p_j , όπου p_j η πιθανότητα επιβίωσης στο διάστημα. Ο αριθμός των ατόμων που επιβιώνουν είναι $n_j - d_j$ και χρησιμοποιώντας το αποτέλεσμα ότι η διασπορά μιας τυχαίας μεταβλητής που ακολουθεί την Διωνυμική κατανομή με παραμέτρους n, p είναι $np(1-p)$, τότε η διασπορά του $n_j - d_j$ δίνεται

$$\text{var}(n_j - d_j) = n_j p_j (1 - p_j)$$

Όπου η διασπορά του $\hat{p}_j$ είναι $\text{var}(n_j - d_j)/n_j^2$ που είναι $p_j(1-p_j)/n_j$ οπότε γράφεται τελικά

$$\hat{p}_j(1-\hat{p}_j)/n_j$$

Και χρησιμοποιώντας τον τύπο (2.3) για την διασπορά της συνάρτησης μια τυχαίας μεταβλητής $$\text{var}\{g(X)\} \approx \left\{\frac{dg(X)}{dX}\right\}^2 \text{var}(X) \qquad (2.3)$$

παίρνουμε ότι $\log \hat{p}_j = (1-\hat{p}_j)/n_j\hat{p}_j$ και αντικαθιστώντας το $\hat{p}_j$ έχουμε

$$\frac{d_j}{n_j(n_j - d_j)}$$

Και από την εξίσωση (2.2)

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

Και με χρήση της (2.3)

$$\operatorname{var}\log\left\{\hat{S}(t)\right\} \approx \frac{1}{\left[\hat{S}(t)\right]^2}\operatorname{var}\left\{\hat{S}(t)\right\}$$

Έτσι ώστε $$\operatorname{var}\left\{\hat{S}(t)\right\} \approx \frac{1}{\left[\hat{S}(t)\right]^2}\sum_{j=1}^{k}\frac{d_j}{n_j(n_j-d_j)}$$

Τελικά παίρνουμε για τυπικό σφάλμα της Kaplan-Meier

$$se\left\{\hat{S}(t)\right\} \approx \hat{S}(t)\left\{\sum_{j=1}^{k}\frac{d_j}{n_j(n_j-d_j)}\right\}^{1/2}$$

Που είναι γνωστός ως ο τύπος του Greenwood. (Collett, 2003)

Αφού υπολογίσουμε το τυπικό σφάλμα της εκτιμώμενης συνάρτηση επιβίωσης, μπορούμε να βρούμε ένα διάστημα εμπιστοσύνης για την αντίστοιχη τιμή της συνάρτηση επιβίωσης. Το διάστημα εμπιστοσύνης είναι ένα διάστημα στο οποίο υπάρχει μια πιθανότητα η τιμή της πραγματικής συνάρτησης επιβίωσης να περιέχεται σε αυτό. Αν για δεδομένη χρονική στιγμή t θεωρήσουμε ότι η $\hat{S}(t)$ ακολουθεί προσεγγιστικά την Κανονική κατανομή σε μέση τιμή $S(t)$ και διασπορά σ^2, τότε το 100(1-α)% διάστημα εμπιστοσύνης σε επίπεδο α προσδιορίζεται ως

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

Αντικαθιστώντας την τυπική απόκλιση σ με την εκτίμηση $se(\hat{S}(t))$ έχουμε

$$\hat{S}(t) \pm z_{a/2}\ se(\hat{S}(t)).$$

Μια δυσκολία με αυτή την διαδικασία είναι το γεγονός ότι τα διαστήματα εμπιστοσύνης είναι συμμετρικά. Όταν η εκτιμώμενη συνάρτηση επιβίωσης είναι κοντά στην μονάδα ή στο μηδέν, τα συμμετρικά διαστήματα δεν είναι κατάλληλα καθώς τα όρια τους θα είναι έξω από το διάστημα (0,1). Μια λύση σε αυτό το πρόβλημα είναι να αντικαταστήσουμε όποια τιμή είναι μεγαλύτερη της μονάδας με 1 και όποια τιμή μικρότερη του 0 με 0.

Μια εναλλακτική διαδικασία είναι να μετατρέψουμε την $\hat{S}(t)$ ώστε να παίρνει τιμές στο διάστημα $(-\infty,+\infty)$ και να βρούμε το διάστημα εμπιστοσύνης για την τροποποιημένη τιμή. Πιθανοί μετασχηματισμοί είναι χρησιμοποιώντας τον λογάριθμο δηλαδή $\log\left[S(t)/\{1-S(t)\}\right]$ και $\log\{-\log S(t)\}$. Χρησιμοποιώντας αυτούς τους μετασχηματισμούς παίρνουμε

$$\operatorname{var}\left\{\log[-\log\hat{S}(t)]\right\} \approx \frac{1}{\left[\log\hat{S}(t)\right]^2}\sum_{j=1}^{k}\frac{d_j}{n_j(n_j-d_j)}$$

Το τυπικό σφάλμα είναι η τετραγωνική ρίζα της ποσότητας $\log[-\log\hat{S}(t)]$. Αυτό μας δίνει την μορφή

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

Ένα επιπλέον πρόβλημα είναι ότι όταν η εκτιμώμενη συνάρτηση επιβίωσης είναι κοντά στο 1 ή στο 0, η διασπορά που παίρνουμε από τον τύπο του Greenwood μπορεί να μην είναι η ίδια με την πραγματική. Σε αυτήν την περίπτωση χρησιμοποιείται μια εναλλακτική έκφραση για το τυπικό σφάλμα της $\hat{S}(t)$ η οποία είναι

$$se\left\{\hat{S}(t)\right\} = \frac{\hat{S}(t)\sqrt{\{1-\hat{S}(t)\}}}{\sqrt{(n_k)}}$$

Για $t_{(k)} \le t < t_{(k+1)}$, $k=1,2,...,r$ όπου $\hat{S}(t)$ η εκτιμήτρια Kaplan-Meier της $S(t)$ και n_k ο αριθμός των ατόμων που βρίσκονται σε κίνδυνο την χρονική στιγμή $t_{(k)}$.

Ωστόσο η έκφραση αυτή του τυπικού σφάλματος για την $\hat{S}(t)$ είναι συντηρητική, με την έννοια ότι τα τυπικά σφάλματα θα τείνουν να είναι μεγαλύτερα από ότι πρέπει. Για αυτό τον λόγο ο τύπος του Greenwood προτιμάται για γενική χρήση. (Collett, 2003)

2.7 Σύγκριση καμπυλών επιβίωσης

Εκτός από την εκτίμηση της συνάρτησης επιβίωσης σε ένα σύνολο δεδομένων, μας ενδιαφέρει και η σύγκριση του χρόνου επιβίωσης ανάμεσα σε δύο ή περισσότερες ομάδες ατόμων που διαφέρουν ως προς κάποιο χαρακτηριστικό. Ο πιο απλός τρόπος για να συγκρίνουμε τους χρόνους επιβίωσης των ομάδων ατόμων είναι παραστίσουμε γραφικά τις αντίστοιχες συναρτήσεις επιβίωσης στου ίδιους άξονες. Υπάρχουν όμως και άλλοι μη παραμετρικοί μέθοδοι για να βρούμε διαφορο-ποιήσεις μεταξύ των ομάδων μεταξύ των οποίων είναι οι log-rank test και Wilcoxon test.

2.7.1 Έλεγχος log-rank

Έστω $t_{(1)} < t_{(2)} < \ldots < t_{(\kappa)}$ διακεκριμένες χρονικές στιγμές κατά τις οποίες παύουν να λειτουργούν μονάδες που προέρχονται από δύο ομάδες. Αμέσως πριν από την χρονική στιγμή $t_{(j)}$, θεωρούμε ότι στην ομάδα i(=1,2) υπάρχουν n_{ij} μονάδες σε κίνδυνο, εκ των οποίων d_{ij} μονάδες παύουν να λειτουργούν την χρονική στιγμή $t_{(j)}$. Ορίζουμε

$$n_j = n_{1j} + n_{2j}$$

Και $$d_j = d_{1j} + d_{2j}.$$

Τα γεγονότα της χρονικής στιγμής $t_{(j)}$ περιγράφονται περιληπτικά από τον παρακάτω Πίνακα συνάφειας 1.

		Ομάδα A	Ομάδα B	$\sum$
Διακοπή λειτουργίας	Ναι	d_{1j}	d_{2j}	d_j
	οχι	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
	$\sum$	n_{1j}	n_{2j}	n_j

Πίνακας συνάφειας 1

Στην κλασσική ανάλυση ενός πίνακα συνάφειας με το γνωστό χ^2-έλεγχο υπολογίζουμε τις αναμενόμενες συχνότητες υπο την υπόθεση ανεξαρτησίας του γεγονότος «διακοπή λειτουργίας των μονάδων» από την «ομάδα που ανήκει».

Δηλαδή οι πιθανότητες να διακοπούν οι λειτουργίες των μονάδων είναι ίδιες για τις δύο ομάδες κατά συνέπεια και οι συναρτήσεις επιβίωσης, το οποίο είναι και το ζητούμενο. Μια τέτοια αναμενόμενη συχνότητα του πρώτου κελιού του πίνακα (μονάδες της Α ομάδα των οποίων η λειτουργία διακόπηκε) είναι η

$$\mathrm{E}(d_{1j}) = n_{1j} d_j / n_j = \hat{d}_{1j}$$

Και η απόκλιση από την παρατηρούμενη d_{1j} είναι

$$u_j = d_{1j} - (n_{1j} d_j / n_j)$$

Ένας έλεγχος της υπόθεσης ανεξαρτησίας προκύπτει από την ποσότητα διαιρώντας την ποσότητα u_j με μια εκτιμήτρια του τυπικού σφάλματος ή ισοδυνάμως διαιρώντας το τετράγωνο της ποσότητας αυτής με την διασπορά της d_{1j}

$$v_j = V(d_{1j}) = n_{1j} n_{2j} d_j (n_j - d_j) / n_j^2 (n_j - 1)$$

δηλαδή,

$$\frac{\left\{d_{1j} - (n_{1j} d_j / n_j)\right\}^2}{n_{1j} n_{2j} d_j (n_j - d_j) / n_j^2 (n_j - 1)} = \frac{u_j^2}{v_j}.$$

Η τελική μορφή της ελεγχοσυνάρτησης του ελέγχου log-rank προσδιορίζεται αθροίζοντας ως προς όλες τις χρονικές στιγμές $t_{(j)}$. Πιο συγκεκριμένα αν

$$u = \sum_j u_j = \sum_j \left\{d_{1j} - (n_{1j} d_j / n_j)\right\}$$

Και θεωρώντας ότι οι πίνακες συνάφειας για κάθε $t_{(j)}$, $j = 1, 2, \ldots, \kappa$ είναι ανεξάρτητοι, τότε η διασπορά v του αθροίσματος u είναι

$$v = \sum_j v_j = \sum_j n_{1j} n_{2j} d_j - (n_j d_j / n_j) / n_j^2 (n_j - 1).$$

Αποδεικνύεται ότι υπο την $H_0 : S_1(t) = S_2(t)$, η $u/\sqrt{v}$ ακολουθεί την N(0,1) ασυμπτωτικά όταν ο αριθμός των διακοπών δεν είναι πολύ μικρός, κατά συνέπεια και η ελεγχοσυνάρτηση log-rank u^2/v ακολουθεί την X_1^2 ασυμπτωτικά.

Πρόκειται για έναν μη-παραμετρικό έλεγχο διότι η μηδενική υπόθεση αφορά την ισότητα δύο συναρτήσεων επιβίωσης, χωρίς να προσδιορίζονται μαθηματικά οι συναρτήσεις αυτές. Αυτός είναι ο βασικός λόγος για την ευρεία χρήση του ελέγχου αυτού. (Χ.Καρώνη, 2009)

2.5.2 Έλεγχος Wilcoxon

Το Wilcoxon test, που είναι γνωστό και ως Breslow test, αποτελεί ένα ακόμα τεστ για τον έλεγχο της μηδενικής υπόθεσης ότι δεν υπάρχει διαφορά στα συναρτήσεις επιβίωσης για δύο διαφορετικά γκρουπ των δεδομένων επιβίωσης. Το Wilcoxon test βασίζεται στο στατιστικό

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

Όπου d_{1j} είναι ο αριθμός των θανάτων την χρονική στιγμή $t_{(j)}$ για το πρώτο γκρουπ και $e_{1j} = n_{1j} d_j / n_j$. Ο συντελεστής στάθμισης στον έλεγχο Wilcoxon ισούται με τον αριθμό των μονάδων σε κίνδυνο αμέσως πριν από μια διακοπή, $w_j = n_j$. Βέβαια γνωρίζουμε ότι ο αριθμός των μονάδων είναι μεγαλύτερος στην αρχή του πειράματος δεδομένου ότι κατά την διάρκεια του μειώνεται. Αυτό σημαίνει ότι ο έλεγχος Wilcoxon δίνει μεγαλύτερο βάρος στις διακοπές που προκύπτουν νωρίς στο πείραμα από τι στις διαφορές που ενδέχεται να προκύψουν μεταξύ των ομάδων αργότερα, ενώ ο έλεγχος log-rank δίνει το ίδιο βάρος σε όλες τις διακοπές. Επομένως ο έλεγχος Wilcoxon είναι ισχυρότερος του log-rank για την εντόπισης διαφοροποιήσεων μεταξύ των συναρτήσεων επιβίωσης που εμφανίζονται νωρίς στο πείραμα και ενδεχομένως όχι αργότερα. Από την άλλη, αν η επιβίωση μεταξύ δύο ομάδων διαφέρει καθ' όλη την διάρκεια του πειράματος και ισχύει το μοντέλο

αναλογικών διακινδυνεύσεων αποδεικνύεται ότι ο έλεγχος log-rank έχει την μεγαλύτερη στατιστική ισχύ.

Και οι δύο έλεγχοι αυτοί είναι αποτελεσματικοί μόνο στην περίπτωση της στοχαστικής διάταξης των δύο συναρτήσεων επιβίωσης δηλαδή

Μηδενική υπόθεση $H_0 : S_1(t) = S_2(t)$ (οι δύο θεραπείες είναι το ίδιο αποτελεσματικές)

Έναντι των υποθέσεων

$H_1 : S_1(t) \neq S_2(t)$ (οι δύο θεραπείες δεν είναι το ίδιο αποτελεσματικές)

$H_2 : S_1(t) > S_2(t)$ (η θεραπεία 1 αποτελεσματικότερη της θεραπεία 2)

$H_3 : S_1(t) < S_2(t)$ (η θεραπεία 2 αποτελεσματικότερη της θεραπείας 1)

Αν αυτό δεν ισχύει τότε κανένας από τους δύο ελέγχους δε θα είναι χρήσιμος π.χ. όταν η $1^η$ ομάδα έχει μέχρι κάποιο χρονικό σημείο μεγαλύτερη επιβίωση της δεύτερης, ενώ η $2^η$ ομάδα έχει την υψηλότερη επιβίωση από το σημείο αυτό και έπειτα. Στην περίπτωση αυτή δεν υπάρχει κάποιος έλεγχος που να είναι κατάλληλος για γενική χρήση. (Χ.Καρώνη, 2009)

ΚΕΦΑΛΑΙΟ 3- ΜΟΝΤΕΛΟ ΑΝΑΛΟΓΙΚΗΣ ΔΙΑΚΙΝΔΥΝΕΥΣΗΣ ΤΟΥ COX

Υπάρχει διαφοροποίηση μεταξύ βιοιατρικών και τεχνολογικών εφαρμογών ως προς τα μοντέλα διάρκειας ζωής που χρησιμοποιούνται. Σε γενικές γραμμές η ανάλυση τεχνολογικών δεδομένων διάρκειας ζωής με συμμεταβλητές βασίζεται κυρίως σε παραμετρικά μοντέλα. Αυτό σημαίνει ότι υιοθετείται ένα συγκεκριμένο παραμετρικό μοντέλο, για παράδειγμα της κατανομής Weibull, που περιγράφει την διάρκεια ζωής. Φυσικά η επιλογή του μοντέλου θα επιβεβαιωθεί από την εξέταση των δεδομένων, σε πολλές όμως περιπτώσεις είναι σχεδόν βέβαιο ποιο μοντέλο θα κριθεί κατάλληλο, λόγω προηγούμενης εμπειρίας με παρόμοιο υλικό ή θεωριών που οδηγούν σε συγκεκριμένα μοντέλα.

Αντιθέτως σε ότι αφορά τον άνθρωπο, κάθε πληθυσμός είναι διαφορετικός και δεν υπάρχουν μαθηματικές θεωρίες, για παράδειγμα, που να περιγράφουν την πορεία ενός ασθενή είτε προς την ανάρρωση είτε προς το θάνατο. Μια άλλη σημαντική διαφορά οφείλεται στο ότι σε τεχνολογικές εφαρμογές, οι συμμεταβλητές που πρέπει να εισαχθούν σε ένα μοντέλο παλινδρόμηση, είναι συνήθως γνωστές. Στις βιοϊατρικές επιστήμες όμως, δεν είναι όλοι οι σημαντικοί παράγοντες γενικώς γνωστοί και σπανίως έχουμε να κάνουμε με εργαστηριακά δεδομένα περισυλλεγμένα υπο ελεγχόμενες συνθήκες για όλους αυτούς τους λόγους είναι δύσκολη η υιοθέτηση ενός βασικού παραμετρικού μοντέλου. Όταν έχουμε δεδομένα επιβίωσης χρησιμοποιείται κατά κύριο λόγο το μοντέλο αναλογικής διακινδύνευσης του Cox (1972) το οποίο είναι ένα από τα πιο ευρέως διαδεδομένα μοντέλα. Το πλεονέκτημα του είναι ότι δεν χρειάζεται να υποθέσουμε ότι τα δεδομένα μας ακολουθούν κάποια συγκεκριμένη κατανομή όπως συμβαίνει στα παραμετρικά μοντέλα, αλλά και το γεγονός ότι η συνάρτηση διακινδύνευσης μπορεί να πάρει οποιαδήποτε μορφή.

Το μοντέλο αναλογικού κινδύνου του Cox, παρουσιάστηκε από τον Cox το 1972. Το μοντέλο του Cox, όπως και όλα τα μοντέλα αναλογικού κινδύνου μοντελοποιούν τη συνάρτηση κινδύνου h(t). Η χρήση του σήμερα εστιάζεται στην ανάλυση απόκομμένων 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. Αποτελεί μια καλή τεχνική για να βρούμε την σχέση μεταξύ επιβίωσης του ασθενή και πολλών επεξηγηματικών μεταβλητών. Ακόμη, μας επιτρέπει να εκτιμήσουμε τον κίνδυνο θανάτου ενός ατόμου, ή άλλου γεγονότος που μας ενδιαφέρει δεδομένου των προγνωστικών τους μεταβλητών.

3.1 Ορισμός του μοντέλου

Θεωρούμε ότι έχουμε ένα καθορισμένο αριθμό ατόμων προς έρευνα έστω n και ότι το $x^{'} = (x_1, x_2, ..., x_p)$ είναι το διάνυσμα των μεταβλητών που εικάζουμε ότι επηρεάζουν το χρόνο ζωής αυτών των ατόμων. Οι μεταβλητές αυτές μπορεί να αποδίδουν διάφορα χαρακτηριστικά όπως θεραπείες ,φυσικές ιδιότητες των ατόμων (π.χ. φύλο, ηλικία, κ.ά), εξωγενείς παράγοντες.

Επιπλέον ορίζουμε το $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$, i=1,2,….,n, να είναι το διάνυσμα με τις τιμές των συμμεταβλητών που αντιστοιχεί στο i-οστό άτομο.

Η γενική μορφή ενός μοντέλου αναλογικής διακινδύνευσης είναι :

$$h(t; x) = h_0(t)g(x)$$

Όπου $h_0(t)$ είναι μία βασική συνάρτηση διακινδύνευσης και $g(x)$ μια συνάρτηση του διανύσματος **x** .Το μοντέλο αναλογικού κινδύνου του Cox υποθέτει ότι η $g(x)$ είναι μια εκθετική συνάρτηση των συμμεταβλητών και ισούται με

$$g(x) = \sum_{j=1}^{p} \beta_j x_j = \exp[\beta^{'} x]$$

Αρχικά υποθέτουμε ότι οι συμμεταβλητές είναι ανεξάρτητες του χρόνου, ότι δηλαδή οι τιμές των συμμεταβλητών x_i καταγράφηκαν στην αρχή της μελέτης την χρονική στιγμή t=0 και ότι οι τιμές αυτές είναι σταθερές καθ’ όλη την διάρκεια της μελέτης. Άρα με βάση τα παραπάνω, στο μοντέλο αναλογικής διακινδύνευσης Cox (1972) , οι συμμεταβλητές **x** δρουν μέσω της σχέσης

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

Όπου $h_0(t)$ είναι μια βασική συνάρτηση διακινδύνευσης (ή αναφορική συνάρτηση κινδύνου) και β' ένα διάνυσμα p συντελεστών , οι οποίοι εκφράζουν ποσοτικά την επίδραση της κάθε μιας των συμμεταβλητών **x**. Η ανεξαρτησία της διακινδύνευσης και κατά συνέπεια και της επιβίωσης από τη συμμεταβλητή x_i σημαίνει ότι $\beta_i' = 0$. Εν τέλει ,η διακινδύνευση $h(t;x)$ εξαρτάται από το χρόνο και τις συμμεταβλητές, αλλά μέσω δύο διαφορετικών παραγόντων. Ο πρώτος παράγοντας $h_0(t)$, είναι μια συνάρτηση του χρόνου μόνο, που αφήνεται απροσδιόριστη, αλλά θεωρείται η ίδια και για τα n άτομα. Ο δεύτερος παράγοντας είναι μια ποσότητα που εξαρτάται από τις συμμεταβλητές μόνο μέσω του διανύσματος β'.

Ορίζουμε την αναλογία κινδύνου (Hazard Ratio-HR) ως το λόγο των συναρτήσεων διακινδύνευσης δύο ατόμων. Στο μοντέλο αναλογικής διακινδύνευσης του Cox παρατηρείται η εξής ιδιότητα: οι συναρτήσεις διακινδύνευσης των εκάστοτε ατόμων είναι ανάλογες μεταξύ τους. Δηλαδή, έστω ότι $[h(t|x_1)/h(t|x_2)]$ ο λόγος των συναρτήσεων διακινδύνευσης δύο ατόμων και $x_1 = (x_{11}, x_{12}, ..., x_{1p})$, $x_2 = (x_{21}, x_{22}, ..., x_{2p})$ να είναι τα διανύσματα των συμμεταβλητών.

Τότε ο λόγος αυτός είναι σταθερός (ανεξάρτητος του χρόνου):

$$HR(t) = \frac{h(t|x_1)}{h(t|x_2)} = \frac{\cancel{h_0(t)}e^{\beta' x_1}}{\cancel{h_0(t)}e^{\beta' x_2}} = e^{\beta'(x_1 - x_2)}$$

Αυτό σημαίνει ότι ο λόγος του κινδύνου θανάτου δύο ασθενών είναι ο ίδιος όσο κι αν επιζήσει ο καθένας. Γνωρίζουμε ότι η συνάρτηση είναι μαθηματικά ισοδύναμη με την συνάρτηση επιβίωσης και συνδέονται μέσω της σχέσης $h(t) = -\frac{d}{dx}\ln S(t)$ ή $S(t) = \exp[-H(t)]$ όπου $H(t)$ η σωρευτική συνάρτηση διακινδύνευσης, επομένως έχουμε:

$$H(t;x) = \int_0^t h_0(u)e^{\beta' x}du = H_0(t)e^{\beta' x}$$

Και άρα

$$S(t;x) = \exp[-H(t;x)] = \exp[-H_0(t)e^{\beta' x}] = [S_0(t)]^{e^{\beta' x}}$$

Όπου $S_0(t) = \exp[-H_0(t)]$ η αναφορική συνάρτηση επιβίωσης (baseline survival function). (Hosmer, Lemeshow and May, 2008)

3.2 Εκτίμηση των παραμέτρων του μοντέλου

Παρατηρήσαμε ότι η συνάρτηση κινδύνου $h_0(t)$ δεν καθορίζεται παραμετρικά. Επομένως δε μπορεί να χρησιμοποιηθεί η συνηθισμένη συνάρτηση πιθανοφάνειας για την εκτίμηση του διανύσματος **β**, ο Cox προτείνει μια συνάρτηση μερικής πιθανοφάνειας, υποθέτοντας ότι δεν υπάρχουν ισότιμες παρατηρήσεις. Στην πράξη όμως παρατηρούνται ισότιμοι χρόνοι επιβίωσης και έτσι η συνάρτηση μερικής πιθανοφάνειας του Cox , έχει τροποποιηθεί για να χειρίζεται ισότιμες παρατηρήσεις.

Θεωρούμε ότι διακόπτεται η λειτουργία κ μονάδων κατά τις διακεκριμένες χρονικές στιγμές $t_{(1)} < t_{(2)} < \ldots t_{(k)}$. Κατά την χρονική στιγμή R_j διακόπτεται η λειτουργία μιας μονάδας με συμμεταβλητές x_j και R_i συμβολίζει το σύνολο των ατόμων που βρίσκονται σε κίνδυνο στο χρόνο αμέσως πριν από την χρονική αυτή στιγμή. Συμβολίζουμε $x_j = (x_{(j)1}, x_{(j)2}, \ldots, x_{(j)p})$, i=1,2,…,k το διάνυσμα των συμμεταβλητών που αντιστοιχεί στο άτομο με πλήρη χρόνο ζωής $t_{(i)}$, $1 \le i \le k$. Από τη βασική θεωρία πιθανοτήτων, η πιθανότητα να διακοπεί η λειτουργία μιας συγκεκριμένης μονάδας, δοθέντος ότι παύει να λειτουργεί μια οποιαδήποτε μονάδα του συνόλου R_i είναι :

$$\frac{h(t_{(j)}; x_j)}{\sum_{i \in R_j} h(t_{(j)}; x_i)} = \frac{\exp(\beta' x_j)}{\sum_{i \in R_j} \exp(\beta' x_i)}$$

Επειδή η $h(t)dt$ εκφράζει τη στιγμιαία πιθανότητα διακοπής. Η συνάρτηση πιθανοφάνειας για το σύνολο των δεδομένων είναι

$$L(\beta) = \prod_{j=1}^{k} \left\{ \frac{\exp(\beta' x_j)}{\sum_{i \in R_j} \exp(\beta' x_i)} \right\}$$

Από την οποία προκύπτει η εκτιμήτρια μέγιστης πιθανοφάνειας $\hat{\beta}$ του β. Παρόλο που η $L(\beta)$ δεν είναι μια πιθανοφάνεια με την συνηθισμένη έννοια (αφού δεν προκύπτει από την πιθανότητα κάποιου παρατηρούμενου αποτελέσματος), έχει αποδειχτεί από τον Cox ότι μπορεί να χρησιμοποιηθεί σαν μια συνηθισμένη

συνάρτηση πιθανοφάνειας, επιτρέποντας έτσι την εκτίμηση του β με τις συνηθισμένες διαδικασίες. Συνεπώς η εκτιμήτρια $\hat{\beta}$ του β που προκύπτει είναι αμερόληπτη, συνεπής και ασυμπτωτικά κανονική, ενώ το διάνυσμα **β**, ο πίνακας πληροφορίας $I(\beta)$ (information matrix), ο λόγος πιθανοφάνειας λ (likelihood ratio) καθώς και οι έλεγχοι υποθέσεων που βασίζονται στην ποσότητα $L(\beta)$ συμπεριφέρονται ακριβώς όπως και στη περίπτωση της συνηθισμένης πιθανοφάνειας.

Οι συντελεστές παλινδρόμησης β, εκτιμώνται από τις τιμές $\hat{\beta}$ που μεγιστοποιούν την μερική πιθανοφάνεια $L(\beta)$ ή ισοδύναμα τον λογάριθμο της. Ο λογάριθμος της πιθανοφάνειας είναι:

$$l(\beta) = \log L(\beta) = \sum_{j=1}^{k} \beta' x_j - \sum_{j=1}^{k} \ln \left\{ \sum_{i \in R_j} \exp(\beta' x_i) \right\}$$

Οι πρώτες μερικοί παράγωγοι είναι

$$\frac{\partial l}{\partial \beta_r} = \sum_{j=1}^{k} x_{jr} - \sum_{j=1}^{k} \left[\sum_{i \in R_j} x_{ir} e^{\beta' x_i} \Big/ \sum_{i \in R_j} e^{\beta' x_i} \right], 1 \le r \le k$$

Και το σύστημα εξισώσεων $\frac{\partial l}{\partial \beta_r} = 0$, $r = 1, \ldots, p$

Λύνεται ως προς β με αριθμητικές επαναληπτικές μεθόδους. Σημειωτέων ότι ο όρος $\beta' x$ δεν περιλαμβάνει κάποιο σταθερό όρο καθότι αυτός θα περιέχεται στην βασική συνάρτηση διακινδύνευσης $h_0(t)$.

Εκτιμήσεις των διασπορών των εκτιμήσεων $\hat{\beta}$ προσδιορίζονται από το αντίστροφο του πίνακα παρατηρούμενης πληροφορίας με (r,s) στοιχείο $-\frac{\partial^2 l}{\partial \beta_r \partial \beta_s}\Big|\hat{\beta}$ όπου

$$-\frac{\partial^2 l}{\partial \beta_r \partial \beta_s} = \sum_{j=1}^{k} \sum_{i \in R_j} x_{ir} \left[x_{is} - \frac{\sum_{l \in R_j} x_{ls} \exp(\beta' x_l)}{\sum_{l \in R_j} \exp(\beta' x_l)} \right] \frac{\exp(\beta' x_i)}{\sum_{l \in R_j} \exp(\beta' x_l)}$$

Επισημαίνεται ότι η συνάρτηση $h_0(t)$ δεν εμφανίζεται σε αυτή την διαδικασία εξού και ο όρος «μερική πιθανοφάνεια». (C.Caroni, 2017)

3.3 Ισόπαλοι χρόνοι διακοπής

Αν οι χρόνοι διακοπής συμπίπτουν, τότε η παραπάνω συνάρτηση πιθανοφάνειας διαμορφώνεται ανάλογα με την περίπτωση. Αρχικά εξετάζουμε την περίπτωση που η μέτρηση του χρόνου είναι συνεχής. Τότε οι διακοπές που συμπίπτουν την χρονική στιγμή $t_{(j)}$ θεωρητικά θα πρόκυπταν σε διαφορετικούς χρόνους, αν οι μετρήσεις ήταν μεγαλύτερης ακρίβειας. Αυτό σημαίνει ότι υπάρχει μια συγκεκριμένη σειρά μεταξύ των διακοπών. Δεν γνωρίζουμε όμως ποια από όλες τις δυνατές διακοπές $d_j!$ έχει προκύψει. Σαν συνέπεια η συνάρτηση μερικής πιθανοφάνειας θα πρέπει να τις περιλάβει όλες και προφανώς αυτό την καθιστά πολύπλοκη. Για τον λόγο αυτό συνήθως προτιμάται η απλή προσέγγιση του Breslow (1974) κατά την οποία ο όρος

$$\frac{\exp(\beta' x_j)}{\sum_{i\in R_j}\exp(\beta' x_i)}$$

της μερικής πιθανοφάνειας για $d_j = 1$ αντικαθίσταται από το

$$\frac{\exp(\beta' z_j)}{\left\{\sum_{i\in R_j}\exp(\beta' x_i)\right\}^{d_j}}$$

Όπου $z_j = \sum_{k=1}^{d_j} x_k$ και x_k το διάνυσμα συμμεταβλητών της μονάδας κ με διακοπή την στιγμή $t_{(j)}$, $\kappa = 1,\ldots,d_j$. Η προσέγγιση θεωρείται ακριβείας, όταν η ποσότητα d_j / n_j θεωρείται μικρή. Υπάρχουν και άλλες εναλλακτικές προσεγγίσεις όπως αυτή του Elfron (1977). Στην περίπτωση όμως που το d_j / n_j δεν είναι μικρό χρησιμοποιούμε μια προσέγγιση του Cox (1972) κατά την οποία παραδεχόμαστε ότι στην πράξη τα δεδομένα μετρήθηκαν σε διακριτή αντί σε συνεχή κλίμακα. Η μέθοδος για δεδομένα τέτοιου είδους έχει ως ακολούθως: δεδομένου ότι την χρονική στιγμή $t_{(j)}$ γίνονται d_j διακοπές, τότε η πιθανότητα α προκύψει ένα οποιοδήποτε σύνολο u αποτελούμενο από d_j μονάδες, είναι

$$P(u) = \exp(\beta' z_u)$$

Όπως και πριν z_u είναι το άθροισμα των συμμεταβλητών x του συνόλου u. Τότε η υπο συνθήκη πιθανότητα του παρατηρούμενου συνόλου μονάδων u^* με διακοπή δίνεται ως

$$P(u^* \mid d_j) = \exp(\beta' z_j) / \sum_{u \in R_j} \exp(\beta' z_u)$$

Σημειώνεται ότι ο παρονομαστής αποτελείται από το άθροισμα όλων των δυνατών $\binom{n_j}{d_j}$ όρων, όπου n_j ο αριθμός των μονάδων σε κίνδυνο (μέλη του R_j) αμέσως πριν την χρονική στιγμή $t_{(j)}$. (Χ.Καρώνη, 2009)

3.4 Έλεγχοι υποθέσεων

Στο μοντέλο αναλογικής διακινδύνευσης του Cox, όταν υπάρχει ένας ικανοποιητικός αριθμός δεδομένων στο δείγμα, οι εκτιμήσεις των συντελεστών παλινδρόμησης ακολουθούν προσεγγιστικά μια κανονική κατανομή. Έτσι είναι δυνατόν να εκτελεστούν διάφοροι έλεγχοι υποθέσεων βάσει αυτών των εκτιμήσεων και των τυπικών σφαλμάτων τους στο εκάστοτε προσαρμοζόμενο μοντέλο.

3.4.1 Έλεγχος του λόγου πιθανοφανειών (likelihood ratio test)

Αποτελεί τον πιο σύνηθες τρόπο για έλεγχο υποθέσεων. Μια τέτοια υπόθεση μπορεί να είναι ότι $H_0 : \beta_i = 0$ κάτι που σημαίνει ότι η διακινδύνευση και η διάρκεια ζωής εξαρτάται από την συμμεταβλητή x_i. Το μοντέλο προσαρμόζεται με και χωρίς τις συμμεταβλητές, η αφαίρεση των οποίων ισοδυναμεί με την επιβολή του πιο πάνω περιορισμού. Έστω $\hat{l}_1$ η μεγιστοποιημένη τιμή του λογαρίθμου της μερικής πιθανοφάνειας του μοντέλου για $\beta_i \neq 0$ και $\hat{l}_0$ για $\beta_i = 0$. Έτσι γίνεται σύγκριση κατά τον έλεγχο αυτό, της τιμής $-2(\hat{l}_0 - \hat{l}_1)$ με την X_1^2 κατανομή.

Αυτό συνεχίζεται για όλο τον αριθμό των συμμεταβλητών, όπου για τον έλεγχο της μηδενικής υπόθεσης , συγκρίνεται η μεταβολή της τιμής $-2(\hat{l}_0 - \hat{l}_p)$ με την τιμή της X_p^2 κατανομής.

3.4.2 Έλεγχος του Wald

Εναλλακτικά , χρησιμοποιείται και ο έλεγχος Wald. Η ελεγχοσυνάρτηση Wald για κάθε μεταβλητή j είναι:

$$W = \left\{ \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \right\}^2$$

Για τον έλεγχο πάλι της μηδενικής υπόθεσης ότι $\beta_i = 0$ η τιμή του W συγκρίνεται με την κατανομή X_1^2 οπότε και ελέγχεται η μηδενική υπόθεση. Ισοδύναμα μπορούμε να χρησιμοποιήσουμε την τιμή της

$$\left\{ \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \right\}$$

η οποία συγκρίνεται με την κατανομή N(0,1).

Πολλές φορές οι τιμές που προκύπτουν από τους παραπάνω ελέγχους είναι πολύ κοντά μεταξύ τους κάτι που παρατηρείται συχνά όταν έχουμε μεγάλα δείγματα. Σε μικρά δείγματα είναι αρκετά διαφορετικοί και πιο αξιόπιστος θεωρείται ο έλεγχος του των πιθανοφανειών.

Η επιλογή των σημαντικών μεταβλητών από ένα σύνολο, γίνονται με τις γνωστές, από την ανάλυση παλινδρόμησης διαδικασίες κατά βήματα (stepwise) όπως την προς τα εμπρός επιλογή (forward selection) και την προς τα πίσω απαλοιφή (backward elimination). Συγκεκριμένα, δεδομένου ότι έχουμε απορρίψει την μηδενική υπόθεση ότι όλοι οι συντελεστές είναι 0, ελέγχουμε ποιες από τις υποψήφιες συμμεταβλητές πρέπει να ληφθούν υπόψιν στο μοντέλο. Οπότε είτε προστίθενται διαδοχικά στο αρχικό μοντέλο, μεταβλητές με στατιστικά σημαντικούς συντελεστές, είτε αφαιρούνται διαδοχικά μη σημαντικές μεταβλητές από το μοντέλο που περιέχει όλες τις υποψήφιες μεταβλητές.

3.5 Επεκτάσεις του μοντέλου του Cox

Το μοντέλο αναλογικού κινδύνου του Cox, αφού τροποποιηθεί κατάλληλα, μπορεί να χρησιμοποιηθεί και σε περιπτώσεις όπου οι μεταβλητές παρουσιάζουν κάποια χαρακτηριστικά, διαφορετικά από αυτά που ισχύουν στο μοντέλο του Cox. Για

παράδειγμα, στις εξαρτώμενες από το χρόνο μεταβλητές, δεν είναι όλες οι μεταβλητές σταθερές, αλλά η τιμή κάποιων μεταβλητών μεταβάλλεται με το χρόνο. Επίσης στην περίπτωση των στρωματοποιημένων μεταβλητών η υπόθεση της αναλογικότητας των κινδύνων δεν ισχύει αναγκαία. Έτσι, όταν χρειάζεται να εξεταστούν τέτοιες μεταβλητές, το μοντέλο του Cox γενικεύεται και τροποποιείται κατάλληλα ώστε να μπορεί να αντιμετωπίσει τις περιπτώσεις αυτές.

Στρωματοποίηση

Όταν θέλουμε να μελετηθεί η επίδραση ενός επιπέδου μια κατηγορικής μεταβλητής Z σε σχέση με άλλες μεταβλητές χωρίς να ενδιαφέρει η επίδραση της Z στο αποτέλεσμα, τότε χρησιμοποιείται μια επέκταση του μοντέλου αναλογικού κινδύνου του Cox. Επίσης όταν μια μεταβλητή έχει επίπεδα που δημιουργούν συναρτήσεις κινδύνου οι οποίες δεν ικανοποιούν την υπόθεση αναλογικότητας, τότε στρωματοποιούμε ως προς την μεταβλητή αυτή. Το μοντέλο που προκύπτει ονομάζεται στρωματοποιημένο μοντέλο του Cox. Και εφαρμόζεται αφού θεωρήσουμε την στρωματοποιήση των δεδομένων της Z σε υποομάδες κάθε μια από τις οποίες χαρακτηρίζεται από ένα επίπεδο του παράγοντα. Το στρωματοποιημένο μοντέλο επιτρέπει στην μορφή της συνάρτησης κινδύνου να αλλάζει ανάμεσα στα επίπεδα της στρωματοποιημένης μεταβλητής . Η μεταβλητή Z μπορεί εκτός από κατηγορική να είναι αποτέλεσμα χωρισμού μιας ποσοτικής μεταβλητής σε ομάδες.

Η συνάρτηση κινδύνου ενός ατόμου που ανήκει στο στρώμα i με διάνυσμα μεταβλητών **x**, είναι :

$$h_i(t;x) = h_{0i}(t)e^{\beta' x}, i = 1,...,I \quad \text{όπου,}$$

i : δηλώνει το στρώμα του παράγοντα

I : το πλήθος των επιπέδων του παράγοντα

$h_{0i}(t)$:η αναφορική συνάρτηση κινδύνου

Από το στρωματοποιημένο μοντέλο, φαίνεται ότι τα άτομα που ανήκουν στο ίδιο στρώμα, έχουν τις ίδιες αναφορικές συναρτήσεις κινδύνου, ενώ αντίθετα τα άτομα που ανήκουν σε διαφορετικά στρώματα έχουν διαφορετικές αναφορικές

συναρτήσεις κινδύνου. Επίσης, τα άτομα που ανήκουν στο ίδιο στρώμα έχουν συναρτήσεις κινδύνου ανάλογες μεταξύ τους, αφού για παράδειγμα για δύο άτομα με μεταβλητές x_1 και x_2 , που ανήκουν στο στρώμα i, $i = 1,...,I$ ισχύει :

$$\frac{h_i(t|x_1)}{h_i(t|x_2)} = \frac{h_{0i}(t)e^{\beta' x_1}}{h_{0i}(t)e^{\beta' x_2}} = e^{\beta'(x_1 - x_2)}$$

Αντίθετα, άτομα που ανήκουν σε διαφορετικά στρώματα δεν έχουν ανάλογες συναρτήσεις κινδύνου, αφού οι αναφορικές συναρτήσεις κινδύνου $h_{0i}(t)$ κάθε στρώματος είναι αυθαίρετες συναρτήσεις του χρόνου και αφήνονται ασυσχέτιστες.

Επιπλέον, από το στρωματοποιημένο μοντέλο φαίνεται ακόμη ότι οι συντελεστές παλινδρόμησης $\boldsymbol{\beta}$ είναι οι ίδιοι σε κάθε στρώμα. Σε αντίθετη περίπτωση, τα δεδομένα κάθε στρώματος θα θεωρούνταν ως διαφορετικά σύνολα δεδομένων και θα αναλύονταν ξεχωριστά.

Η εκτίμηση των συντελεστών παλινδρόμησης $\boldsymbol{\beta}$ προκύπτει από τη μεγιστοποίηση της συνάρτησης μερικής πιθανοφάνειας που γενικεύει την $L(\boldsymbol{\beta})$ και δίνεται από τη σχέση:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{\kappa} L_i(\boldsymbol{\beta})$$

Κάθε παράγοντας $L_i(\boldsymbol{\beta})$ είναι η μερική πιθανοφάνεια που υπολογίζεται από την σχέση $L(\boldsymbol{\beta})$ που ορίσαμε προηγουμένως για το στρώμα i και υπολογίζεται σε κάθε διακεκριμένο χρόνο αποτυχίας που παρατηρείται στο συγκεκριμένο στρώμα. (Hosmer and Lemeshow, 1998)

3.6 Έλεγχοι της Υπόθεσης Αναλογικότητας Κινδύνων

Όταν προσαρμόσαμε το μοντέλο του αναλογικής διακινδύνευσης του Cox εικάζαμε ότι ίσχυε η υπόθεση της αναλογικότητας και ότι το μοντέλο ήταν κατάλληλο για τα δεδομένα. Ωστόσο αυτή υπόθεση πρέπει να ελέγχεται. Ο έλεγχος μπορεί να πραγματοποιηθεί είτε με γραφικές μεθόδους είτε με διάφορα στατιστικά. Αφού εξετάσουμε αν ισχύει η υπόθεση επόμενο βήμα είναι να δούμε ποιες είναι οι πιο

σημαντικές μεταβλητές. Στην συνέχεια προσαρμόζουμε το μοντέλο με βάση αυτές και ελέγχουμε αν το μοντέλο είναι ικανοποιητικό ή επιδέχεται περαιτέρω βελτίωση. Στην περίπτωση που δεν ευσταθεί η υπόθεση της αναλογικότητας κινδύνων, τότε προχωράμε σε μετασχηματισμούς των δεδομένων ώστε να ικανοποιείται αυτή η υπόθεση. Ένας μεγάλος αριθμός διαγνωστικών μεθόδων έχει προταθεί μέχρι σήμερα για το μοντέλο αναλογικής διακινδύνευσης (PH) του Cox, αλλά δεν έχουν χρησιμοποιηθεί όπως θα έπρεπε. Στην συνέχεια θα παρουσιάσουμε τις πιο σύγχρονες μεθόδους που φαίνεται να έχουν μεγαλύτερη πρακτική σημασία και αξίζουν μεγαλύτερης προσοχής.

3.7.1 Γραφικός έλεγχος

Από την συνάρτηση διακινδύνευσης του μοντέλου αναλογικής διακινδύνευσης παίρνουμε την συνάρτηση επιβίωσης

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

Όπου H_0 είναι η σωρευτική συνάρτηση διακινδύνευσης της αντίστοιχης h_0. Λογαριθμίζοντας δύο φορές και τα δύο μέλη παίρνουμε

$$\ln\left\{-\ln S(t;x)\right\} = \ln H_0 + \beta' x.$$

Αυτό σημαίνει ότι κάθε συνάρτηση επιβίωσης $S(t;x)$ όταν σχεδιάζεται με το συμπληρωματική log-log κλίμακα διαφέρει από τον όρο $\ln H_0$ κατά την σταθερή ποσότητα $\beta' x$ καθόλη την διάρκεια. Ως εκ τούτου δύο οποιεσδήποτε συναρτήσεις $S(t;x_1)$ και $S(t;x_2)$ για διαφορετικές τιμές του διανύσματος συμμεταβλητών **x** θα είναι παράλληλες ή σχεδόν παράλληλες. Αυτή η τεχνική αποτελεί ένα βασικό μέσο για τον έλεγχο της υπόθεσης αναλογικότητας, ίσως η μόνη που προτάθηκε στα περισσότερα βασικά βιβλία όπως Parmar και Machin (1995):

-Λήψη των εκτιμήσεων Kaplan-Meier $\hat{S}(t;x)$ για διάφορα x.

-Οπτικός έλεγχος αν οι καμπύλες $\ln\left\{-\ln \hat{S}(t;x)\right\}$ σε σχέση με το t (ή συνάρτηση του χρόνου) είναι παράλληλες για διαφορετικά x.

Ο λόγος της χρήσης συνάρτηση του t είναι για να διευκολυνθεί η εκτίμηση του γραφήματος. Αν χρησιμοποιηθεί $\ln t$ τότε τα γραφήματα των $\ln\left\{-\ln \hat{S}(t;x)\right\}$ θα

είναι ευθείες γραμμές όταν οι χρόνοι επιβίωσης ακολουθούν την κατανομή Weibull. Είναι ευκολότερο για το μάτι να αξιολογεί ευθείες γραμμές παρά καμπύλες. Το κύριο μειονέκτημα αυτής της διαδικασίας είναι φανερό. Είναι απαραίτητο να έχουμε αξιόπιστες εκτιμήσεις των $\hat{S}$, κάτι που σημαίνει ότι χρειάζεται να έχουμε έναν μεγάλο όγκο δεδομένων για κάθε επιλεγμένο x. Αυτό μπορεί να συμβεί μόνο αν υπάρχουν λίγες συμμεταβλητές και λίγες διακριτές τιμές (ωστόσο οι συνεχείς μεταβλητές μπορούν να ομαδοποιηθούν), για παράδειγμα αν υπάρχει μόνο μία συμμεταβλητή με μικρό αριθμό επιπέδων. Γενικά όταν έχουμε πολλές μεταβλητές ή μια συμμεταβλητή έχει πολλά επίπεδα τότε είναι δύσκολο να ελέγξουμε με γραφικές μεθόδους αν ευσταθεί η υπόθεσης της αναλογικότητας. (C.Caroni, 2004)

3.6.2 Έλεγχος μέσω υπολοίπων

Ο έλεγχος μέσω υπολοίπων αποτελεί ένα πιο σημαντικό και αξιόπιστο τρόπο για να ελέγξουμε την υπόθεση αναλογικότητας, αλλά και άλλων θεμάτων όσον αφορά την καταλληλόλητα μου μοντέλου του Cox. Τα πιο διαδεδομένα υπόλοιπα για το μοντέλο του Cox είναι τα Schoenfeld, Martingale, df-beta τα οποία θα αναλύσουμε παρακάτω.

Υπόλοιπα Schoenfeld

Τα υπόλοιπα από ένα προσαρμοσμένο μοντέλο είναι πιο εύκολο να κατανοηθούν όταν εκφράζουν κατά κάποιο τρόπο την διαφορά ανάμεσα στις τιμές των παρατηρούμενων δεδομένων και των αντίστοιχων εκτιμώμενων τιμών. Αυτό μπορεί να γίνει με διάφορους τρόπους. Δεν προκαλεί έκπληξη το γεγονός ότι το πιο σύνθετο μοντέλο αναλογικής διακινδύνευσης δίνει πολλές εναλλακτικές μορφές υπολοίπων. Ένα από τα πιο σημαντικά παρουσιάστηκαν από τον Schoenfeld (1982).

Την χρονική στιγμή t_i που συμβαίνει το γεγονός (θάνατος ή αποτυχία) για ένα άτομα ή μια μονάδα i, υπάρχει ένα σετ ρίσκου R_i αποτελούμενο από αυτούς που δεν έχουν πεθάνει και βρίσκονται σε κίνδυνο να πεθάνουν εκείνη την χρονική στιγμή. Η υπο συνθήκη πιθανότητα να αποτύχει το i, δεδομένου ότι υπήρχε αποτυχία την χρονική στιγμή t_i δίνεται από τον τύπο

$$p_i = \frac{h_i(t_i)}{\sum_{j \in R_i} h_j(t_i)}$$
$$= \frac{h_0(t_i)e^{\beta' x_i}}{\sum_{j \in R_i} h_0(t_i)e^{\beta' x_j}}$$
$$= \frac{e^{\beta' x_i}}{\sum_{j \in R_i} e^{\beta' x_j}}.$$

Θεωρούμε τώρα $E(\mathbf{x}|R_i)$, την αναμενόμενη τιμή (πριν από την αποτυχία μιας μονάδας) του διανύσματος των συμμεταβλητών $\mathbf{x}$ την χρονική στιγμή της αποτυχίας t_i, δεδομένου του πακέτου που βρίσκεται σε ρίσκο

$$E(\mathbf{x}|R_i) = \sum_{k \in R_i} x_k p_k$$
$$= \frac{\sum_{k \in R_i} x_k e^{\beta' x_i}}{\sum_{j \in R_i} e^{\beta' x_j}}$$

Ορίζουμε τώρα ένα υπόλοιπο με την γνωστή μορφή της απόκλισης της παρατήρησης από την αναμενόμενη τιμή ως

$$r_i = x_i - E(\mathbf{x}|R_i)$$

Και αντικαθιστώντας τα β με τα $\hat{\beta}$, προκύπτουν τα υπόλοιπα Schoenfeld (ή μερικά υπόλοιπα -partial residuals)

$$\hat{r}_i = x_i - \hat{E}(\mathbf{x}|R_i)$$

Σε αυτό το σημείο πρέπει να λάβουμε υπόψιν ότι αυτά τα υπόλοιπα προσδιορίζονται στους χρόνους και όχι στις αποκομμένες παρατηρήσεις σε αντίθεση με τα υπόλοιπα του μοντέλου παλινδρόμησης τα υπόλοιπα Schoenfeld δεν προσδιορίζονται από τις τιμές της εξαρτημένης μεταβλητής (π.χ. από το χρόνο t) αλλά από τις συμμεταβλητές **x**. Αντιπροσωπεύει την απόκλιση μεταξύ της συμμεταβλητής της μονάδας που αποτυγχάνει την χρονική στιγμή t_i και του σταθμισμένου μέσου όρου όλων των συμμεταβλητών που βρίσκονται στο risk set, άρα ένα μεγάλο υπόλοιπο δείχνει ότι η μονάδα που αποτυγχάνει την t_i αποτελεί ακραία παρατήρηση εκείνη την χρονική στιγμή.(δεδομένου ότι το r_i είναι ένα

διάνυσμα, μπορεί να είναι μεγάλο σε ένα ή σε περισσότερα στοιχεία, και όχι απαραίτητα σε όλα). Ως εκ τούτου αυτού του είδους τα υπόλοιπα είναι χρήσιμα για να εντοπίσουμε σημεία επιρροής.

Τα υπόλοιπα Schoenfeld έχουν άμεση σχέση με την διαδικασία εκτίμησης του μοντέλου αναλογικής διακινδύνευσης του Cox. Το γινόμενο των παραπάνω πιθανοτήτων p_i στο σύνολο D όλων των ατόμων ή μονάδων που πέθαναν ή αποτύχανε αντίστοιχα, στην περίπτωση διακριτών χρόνων αποτυχίας είναι

$$L = \prod_{i \in D} p_i$$

όπου

$$l = \ln L = \sum_{i \in D} \left\{ \beta' x_i - \ln \sum_{j \in R_i} e^{\beta' x_j} \right\}$$

και

$$U = \frac{\partial l}{\partial \beta} = \sum_{i \in D} \left\{ x_i - \sum_{k \in R_j} x_k e^{\beta' x_k} / \sum_{j \in R_i} e^{\beta' x_j} \right\} = \sum_{i \in D} r_i$$

Άρα θέτοντας $U(\hat{\beta}) = 0$ από την διαδικασία εκτίμησης για $\beta = \hat{\beta}$, βλέπουμε αμέσως μια επιθυμητή ιδιότητα των υπολοίπων

$$\sum_{i \in D} \hat{r}_i = 0 .$$

Martingale residuals

Το μοντέλο αναλογικής διακινδύνευσης του Cox μπορεί να γενικευτεί με διάφορους τρόπους. Μια πολύ σημαντική βελτίωση ήταν να προσαρμοστεί σε ένα πλαίσιο από διαδικασίες καταμέτρησης (Andersen and Gill, 1982).

Έστω

$$N_i(t), \quad t \geq 0$$

Όπου ο δείκτης i δηλώνει των αριθμό των γεγονότων που παρατηρούνται σε χρόνο t. Υποθέτουμε ότι $N_i(t)$ έχει την συνάρτηση έντασης

$$\mathrm{Y}_i(t) e^{\beta' x} dH_0(t)$$

Όπου $\mathrm{Y}_i(t)$ είναι μια δύτιμη μεταβλητή (0-1) που υποδηλώνει αν το i-οστό αντικείμενο είναι σε κίνδυνο την χρονική στιγμή t και $x_i(t)$ είναι το διάνυσμα των

συμμεταβλητών. Για το μοντέλο του Cox το $H_0(t)$ είναι απροσδιόριστο και η $Y_i(t)$ παίρνει την τιμή 1 μέχρι να συμβεί το πρώτο γεγονός και 0 έπειτα.

Τα υπόλοιπα του μοντέλου αναλογικής διακινδύνευσης του Cox σε αυτό το πλαίσιο αναπτύχθηκαν από τον Therneau (1990), και ακολούθησαν οι Barlow and Prentice (1988) οι οποίοι εισήγαγαν τα υπόλοιπα που ορίζονται από την διαφορά ανάμεσα στην διαδικασία καταμέτρησης και του ολοκληρώματος της συνάρτησης έντασης

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) e^{\beta' x_i(s)} dH_0(s), \quad i = 1, \ldots, n$$

Και η αντίστοιχη εκτίμηση είναι

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(s) e^{\hat{\beta}' x_i(s)} d\hat{H}_0(s), \quad i = 1, \ldots, n$$

Όπου $\hat{H}_0$ η εκτιμήτρια Breslow της βασικής σωρευτικής συνάρτησης διακινδύνευσης και ονομάζεται υπόλοιπο martingale και ερμηνεύεται ως το παρατηρούμενο μείον των προβλεπόμενο αριθμό γεγονότων στο διάστημα [0.t]. Ειδικότερα για το μοντέλο του Cox χωρίς χρονοεξαρτόμενες συμμεταβλητές παίρνει την απλούστερη μορφή

$$\hat{M}_i = \delta_i(t) - \hat{H}_0(t_i) e^{\hat{\beta}' x_i}$$

Όπου δ_i είναι η τελική κατάσταση (1 για το γεγονός και 0 για αποκομμένη) και t_i ο παρατηρούμενος χρόνος για την i-οστή μονάδα. Ο δεύτερος όρος είναι μια εκτίμηση

$$H_0(t_i) e^{\beta' x_i}$$

η οποία κατανέμεται σαν εκθετική μονάδα. Ως εκ τούτου οι αντίστοιχες εκτιμήσεις μπορούν να θεωρηθούν ως υπόλοιπα του γενικού τύπου που παρουσιάστηκαν από τον Cox και Snell (1968).

Οι ακραίες παρατηρήσεις μπορούν να εντοπιστούν από το διάγραμμα των martingale υπολοίπων σε σχέση με το $\hat{\beta}' x$. Μια δυσκολία που προκύπτει άμεσα είναι ότι τα υπόλοιπα αυτά δεν έχουν καθόλου συμμετρική κατανομή αφού έχουν εύρος $(-\infty, +1)$. Αυτό έχει σαν αποτέλεσμα να δούμε μόνο τις ασυνήθιστα μεγάλες αρνητικές τιμές των αντίστοιχων ατόμων που επιβίωσαν για ένα απροσδόκητα μεγάλο χρονικό διάστημα. Ένας καλός τρόπος για να αποφύγουμε αυτό το

πρόβλημα είναι να μετασχηματίσουμε τα martingale υπόλοιπα στο να προσεγγίσουν την κανονικότητα όπως άλλα πιο γνωστά υπόλοιπα απλούστερων στατιστικών μοντέλων. Για τον σκοπό αυτό ο Therneau (1990) εισήγαγε τα deviance υπόλοιπα τα οποία έχουν την ίδια συμπεριφορά που έχουν στα γενικευμένα γραμμικά μοντέλα. Σε αυτήν της ειδική περίπτωση του μοντέλου του Cox το i-οστό deviance υπόλοιπο είναι

$$d_i = \text{sgn}(\hat{M}_i)\left[-2\left\{\hat{M}_i + \delta_i \log(\delta_i - \hat{M}_i)\right\}\right]^{1/2}$$

Ωστόσο σύμφωνα με τους Therneau και Grambsch (2000), αυτά τα υπόλοιπα δεν έχουν φανεί και πολύ χρήσιμα. Τα άτομα που πέθανα πολύ πιο νωρίς από ότι προβλεπόταν μπορεί να εμφανιστούν στα υπόλοιπα deviance , αλλά μπορεί και όχι.

Μια άλλη σημαντική χρησιμότητα που έχουν τα martingale υπόλοιπα είναι ότι μπορούν να υποδείξουν την σωστή συναρτησιακή μορφή μια συμμεταβλητής. Η ιδέα είναι ότι η συμμεταβλητή j θα μπορούσε να εισέλθει στο μοντέλο ως $\exp(f(x_j)\beta_j)$, και μετά μια πιο ομαλή γραφική παράσταση των υπολοίπων για το μοντέλο χωρίς συμμεταβλητές σε σχέση με το x_j θα δείξει την καμπύλη $f(x_j)$. Αυτή η ιδέα μπορεί να δουλέψει καλά μόνο αν οι συμμεταβλητές έχουν μικρή συσχέτιση. (C. Caroni, 2004)

<u>Έλεγχος της υπόθεσης αναλογικής διακινδύνευσης με χρήση υπολοίπων</u>

Η υπόθεση της αναλογικής διακινδύνευσης (PH) απαιτεί ο λόγος $\lambda(t;x)/\lambda_0(t)$ να είναι σταθερός όσο περνάει ο χρόνος. Άρα ένας τρόπος για να μην ισχύει η υπόθεση είναι να συμπεριλάβουμε το χρόνο και άρα θα προκύψει

$$\lambda(t;x) = \lambda_0(t)e^{\beta'(t)x}$$

Όπου τουλάχιστον μία συνιστώσα του διανύσματος $\beta(t)$ δεν είναι σταθερή. Ένα τεστ για την υπόθεση $H_0(t): \beta(t) = \beta, \forall t$ είναι και ένα τεστ για τον έλεγχο της υπόθεσης PH. Όπως παρουσίασαν οι Therneau και Grambsch (1994) τα κατάλληλα τεστ βασίζονται στα υπόλοιπα Schoenfeld και πολλά άλλα τεστ που εκδόθηκαν πριν από αυτήν την ημερομηνία, ανήκουν σε αυτήν την κατηγορία. Αναλύεται παρακάτω αυτό το τεστ που φαίνεται το οποίο είναι αρκετά ισχυρό.

Γράφουμε $\beta(t) = \beta + g(t)$ και $\hat{\beta}$ η συνηθισμένη εκτίμηση του β του κλασσικού μοντέλου PH. Μπορεί να αποδειχθεί ότι

$$E(\hat{r}_i) = V_i g(t_i)$$

Όπου V_i είναι ο πίνακας συνδιακύμανσης των υπολοίπων Schoenfeld r_i. Αντικαθιστώντας τις εκτιμήσεις έχουμε ότι

$$\beta(t) = \beta + V_i^{-1} g(t_i) = \beta + \hat{r}_i^*$$

Όπου $\hat{r}_i^*$ είναι τα κλιμακοποιημένα υπόλοιπα Schoenfeld. Ως εκ τούτου ένα γράφημα των $\hat{r}_{ij}^* + \hat{\beta}_j$ σε σχέση με το χρόνο ή μια συνάρτηση του χρόνου υποδεικνύει αποκλίσεις από την υπόθεση PH και προτείνει την σωστή συναρτησιακή μορφή του χρόνου. Στην πραγματικότητα είναι σημαντικό είναι να κάνουμε ένα scatteplot διάγραμμα για να δούμε πιο ξεκάθαρα τι γίνεται.

Οι Therneau και Grambsch (1994) προτείνουν να αντικαταστήσουν τον πίνακα $\hat{V}_i$ με τον μέσο όρο του δηλαδή

$$\bar{V} = I(\hat{\beta}) / d$$

Όπου $I(\hat{\beta})$ ο πίνακας πληροφορίας διαιρούμενος με τον αριθμό των μη αποκομμένων παρατηρήσεων d. Αυτό θα μας βοηθήσει στο να αποφύγουμε να αντιστρέψουμε όλους τους πίνακες $\hat{V}_i$ ξεχωριστά, αλλά αυτό δεν αποτελεί ιδιαίτερο πρόβλημα λόγω των υπολογιστών που μα βοηθούν. Πράγματι οι Therneau και Grambsch θεώρησαν αυτήν της αντικατάσταση ως μια βελτίωση μιας και τα τελευταία $\hat{V}_i$ θα βασίζονταν σε μικρούς αριθμούς των αντικειμένων και δεν θα είχαν καλές εκτιμήσεις. Αυτή η πρόταση εξετάστηκε πρόσφατα από τους Winnett και Sasieni (2001) οι οποίοι βρήκαν ότι σε πολλές περιπτώσεις δεν έχει πολύ μεγάλη διαφορά, αλλά σε άλλες μπορεί να οδηγήσει σε παραπλανητικές εκτιμήσεις των συντελεστών που μεταβάλλονται με τον χρόνο. Ως εκ τούτου καταλήγουν στο γεγονός ότι δεν πρέπει να χρησιμοποιηθεί αυτή η διαδικασία. Προτιμούν την χρήση των μεμονωμένων κλιμακοποιημένων υπολοίπων, ωστόσο άφησαν ανοιχτή την πιθανότητα πως το $\hat{V}_i$ θα μπορούσε να τροποποιηθεί, ίσως με εξομάλυνση.

(C. Caroni, 2004)

Σημεία επιρροής (Influence points)

Η επιρροή ενός σημείου των δεδομένων στην προσαρμογή ενός στατιστικού μοντέλου ορίζεται ως η αλλαγή στην εκτίμηση της παραμέτρου όταν το σημείο αυτό παραλείπεται. Στο μοντέλο αναλογικής διακινδύνευσης PH το πως θα επηρεάσει η παράλειψη του σημείου j, που θα έχει ως αποτέλεσμα την μεταβολή της εκτίμησης της παραμέτρου από $\hat{\beta}$ σε $\hat{\beta}_j$, μπορούμε να το καταλάβουμε μόνο εάν ξαναπροσ-αρμόσουμε το μοντέλο. Για να το αποφύγουμε αυτό χρησιμοποιούμε μια προσέγγιση. Αυτό μπορεί να γίνει εύκολα παίρνοντας τον πρώτο όρο της σειράς Taylor για την συνάρτησης που δίνει

$$\hat{\beta} - \hat{\beta}_j = I^{-1} d_j$$

Όπου $$d_j = \delta_j \hat{r}_j - \sum_{i \in D_j} \left[e^{\beta' x_j} \left\{ x_j - \hat{E}(x|R_i) \right\} / \sum_{k \in R_i} e^{\beta' x_k} \right]$$

Το D_j δηλώνει το σετ των ατόμων που απεβίωσαν την χρονική στιγμή t_j ή πιο πριν. Τα διανύσματα d_j καλούνται υπόλοιπα score. Ο πίνακας nxp του οποίου οι γραμμές αποτελούνται από τα $I^{-1}\hat{d}_j$ έχει γίνει γνωστός ως ο πίνακας των df beta τιμών τα οποία είναι διαθέσιμα σε συγκεκριμένα στατιστικά πακέτα. Ο πρώτος όρος του d_j είναι απλά τα υπόλοιπα Schoenfeld και υπάρχει μόνο αν το άτομο αυτό πέθανε. Ο δεύτερος όρος αφορά όλα τα άτομα, παριστάνει την συνεισφορά όλων των risk set R_i δηλαδή αυτών που βρίσκονται σε κίνδυνο. Αν ένα άτομο πεθάνει αρκετά νωρίς, τότε ο πρώτος όρος θα είναι πιο σημαντικός από τον δεύτερο. Όσο πιο πολύ μένει στην ζωή το άτομο τόσο μεγαλύτερη σημασία αποκτά ο δεύτερος όρος. Για μια παρατήρηση που είναι αποκομμένη στην αρχή της μελέτης ο πρώτος όρος είναι μηδέν και ο δεύτερος είναι πολύ μικρός, άρα η επιρροή είναι ελάχιστη.

Added variable plots

Στην γραμμή παλινδρόμηση, το διάγραμμα προσθετικής μεταβλητής είναι ένα πολύ χρήσιμο εργαλείο για να ελέγξουμε αν μια συγκεκριμένη μεταβλητή πρέπει να εισαχθεί στο μοντέλο αλλά και ποιες από τις παρατηρήσεις είναι σημεία επιρροής. Λόγω αυτής της χρησιμότητας έχει επεκταθεί και σε άλλα μοντέλα. Έτσι οι O'Hara Hines και Carter (1993) ανέπτυξαν αυτό το διάγραμμα και για γενικευμένα γραμμικά

μοντέλα. Η ίδια ιδέα προσαρμόστηκε στην συνέχεια και στο μοντέλο αναλογικής διακινδύνευσης(PH) του Cox από τον Hall et al. (1996). Είναι πολύ απλή η ιδέα αυτής της διαδικασιάς. Αρχικά το διάγραμμα προσθετικής μεταβλητής για μια συμ-μεταβλητή x_j στην γραμμική παλινδρόμηση αποτελείται από ένα scatterplot των υπολοίπων από τις δύο παλινδρομήσεις, την παλινδρόμησης της εξαρτημένης μεταβλητής y των συμμεταβλητών εκτός των x_j σε σχέση με τα υπόλοιπα από την παλινδρόμηση των x_j στις άλλες συμμεταβλητές. Δηλαδή

$$(\mathrm{I} - \mathrm{H}_{(j)})\mathrm{y} \qquad \text{σε σχέση με} \quad (\mathrm{I} - \mathrm{H}_{(j)})\mathrm{x}_j$$

Όπου $$\mathrm{H}_{(j)} = \mathrm{I} - \mathrm{X}_{(j)}(\mathrm{X}'_{(j)}\mathrm{X}_{(j)})^{-1}\mathrm{X}_{(j)}$$

Και $\mathrm{X}_{(j)}$ ο πίνακας των συμμεταβλητών x_j που παραλείπονται. Η κλίση της γραμμής παλινδρόμησης που περνά από τα προσαρμοσμένα σημεία ισούται με τον συντελεστή παλινδρόμησης για την συμμεταβλητή x_j αν φυσικά αυτή περιλαμβάνεται στο μοντέλο. Κατά δεύτερον παρατηρείται ότι τα γενικευμένα γραμμικά μοντέλα μπορούν να προσαρμοστούν επαναληπτικά με σταθμισμένα ελάχιστα τετράγωνα (McNullagh and Nelder, 1989). Σε αυτήν την περίπτωση η εκτιμήσεις μέγιστης πιθανοφάνειας των παραμέτρων β λαμβάνονται από την επίλυση της εξίσωσης

$$(X'WX)\hat{\beta} = X'Wz$$

Αυτά είναι πανομοιότυπα με τις εξισώσεις που εμφανίζονται στην σταθμισμένη γραμμική παλινδρόμηση, αλλά απαιτούν επανάληψη με επαναυπολογισμό των βαρών W και της εξαρτημένης μεταβλητής Z σε κάθε βήμα. (C. Caroni, 2004)

ΚΕΦΑΛΑΙΟ 4 - ΚΑΜΠΥΛΕΣ ROC

Η πραγματοποίηση προβλέψεων αποτελεί ένα από τα σημαντικότερα θέματα κάθε επιχείρησης και επιστημονικού πεδίου όσον αφορά την αναζήτηση πληροφορίας. Το γεγονός αυτό καθιστά την εξασφάλιση προγνωστικής ακρίβειας πολύ σημαντική για τον σχεδιασμό των μοντέλων , αλγορίθμων και άλλων τεχνολογιών που παράγουν προβλέψεις. Οι καμπύλες ROC (Receiver Operating Characteristic –Λειτουργικό Χαρακτηριστικό Δέκτη) αποτελούν σημαντικό εργαλείο στην εξασφάλιση της ακρίβειας στις προβλέψεις και ως εκ τούτου στην λήψη αποφάσεων. Ιστορικά οι καμπύλες ROC αναπτύχθηκαν στις αρχές της δεκαετίας του 50 στην θεωρία ανίχνευσης σημάτων και στην συνέχεια εφαρμόστηκαν σε πολλούς άλλους κλάδους όπως και αυτός της Ιατρικής. Στην ιατρική έρευνα αναπτύσσονται διαγνωστικοί έλεγχοι με σκοπό όχι μόνο τον εντοπισμό και την πρόληψη ασθενειών αλλά και την εξάλειψη της νόσου σε υγιή άτομα.

4.1 Ορισμοί

Έστω T_i ο χρόνος επιβίωσης για το άτομο i, και υποθέτουμε ότι παρατηρούμε μόνο την ελάχιστη τιμή T_i και C_i, όπου C_i παριστάνει έναν ανεξάρτητο χρόνο αποκοπής. Ορίζουμε τον χρόνο παρακολούθησης ως $X_i = \min(T_i, C_i)$, και $\Delta_i = \mathbf{1}(T_i \le C_i)$ τον δείκτη αποκοπής. Ο χρόνος επιβίωσης T_i μπορεί ακόμα να παρασταθεί μέσα από την διαδικασία καταμέτρησης $N_i^*(t) = \mathbf{1}(T_i \le t)$ ή την αντίστοιχη αύξηση $dN_i^*(t) = N_i^*(t) - N_i^*(t-)$. Σημειώνουμε εδώ ότι εστιάζουμε στην διαδικασία καταμέτρησης $N_i^*(t)$ που ορίζεται μόνο όσον αφορά τον χρόνο επιβίωσης T_i, αντί της πιο συχνής $N_i(t) = \mathbf{1}(X_i \le t, \Delta_i = 1)$ η οποία εξαρτάται από τον χρόνο αποκοπής (Fleming and Harrington, 1991). Ορίζουμε $R_i(t) = \mathbf{1}(X_i \ge t)$ ως δείκτη ρίσκου. Υποθέτουμε ακόμα ότι για κάθε i έχουμε ένα σύνολο από χρονικά αμετάβλητες συμμεταβλητές $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ip})$. Εστιάζουμε τώρα σε μεθόδους του μοντέλου του Cox για να δημιουργήσουμε ένα score για το μοντέλο αλλά και την προβλεπτική δυνατότητα του. Ωστόσο οι εκτιμώμενες μέθοδοι που προτείνονται μπορούν να

χρησιμοποιηθούν για να συνοψίσουν την ακρίβεια του προγνωστικού score που παράγεται από οποιαδήποτε παλινδρόμηση ή προγνωστική μέθοδο, και σε αυτή την περίπτωση διάφορες μέθοδοι συντελεστών (Hastie and Tibshirani, 1993) όπως σταθμισμένη μερική εκτίμηση πιθανοφάνειας (Cai and Sun, 2003), παρέχουν μια βολική προσέγγιση για την εκτίμηση ακριβών αποτελεσμάτων. Επομένως εισάγουμε εν συντομία τις σχετικές πτυχές της εκτίμησης μερικής πιθανοφάνειας. Υπο την υπόθεση της αναλογικής διακινδύνευσης

$$\lambda\left(t\middle|Z_i\right)=\lambda_0(t)\exp(Z_i^{\mathrm{T}}\boldsymbol{\beta})$$

όπου, $$\lambda\left(t\middle|Z_i\right)=\lim_{\delta\to\infty}\delta^{-1}P[T_i\in[t,t+\delta)\middle|\mathbf{Z}_i,T_i\geq t].$$

Η μερική πιθανοφάνεια των εξισώσεων μπορεί να γραφτεί ως

$$\mathbf{0}=\sum_i\Delta_i\left[\mathbf{Z}_i-\left(\sum_\kappa\pi_\kappa(\boldsymbol{\beta},X_i)\mathbf{Z}_\kappa\right)\right],$$

όπου $\pi_\kappa(\boldsymbol{\beta},t)=R_k(t)\exp(Z_k^{\mathrm{T}}\boldsymbol{\beta})/W(t)$ με $W(t)=\sum_j R_j(t)\exp(Z_j^{\mathrm{T}}\boldsymbol{\beta})$. Επιλύοντας αυτές τις εξισώσεις μας δίνουν τις εκτιμήσεις της μέγιστης πιθανοφάνειας $\hat{\boldsymbol{\beta}}$. (Cox, 1972). Έστω τώρα ότι τα αποτελέσματα μας Y_i είναι διχότομα τότε η ευαισθησία (sensitivity) ορίζεται ως $P(\hat{p}_i>c\middle|Y_i=1)$ και η ειδικότητα specificity ως $P(\hat{p}_i\leq c\middle|Y_i=0)$, όπου $\hat{p}_i$ είναι μια πρόβλεψη και c ένα κριτήριο για την ταξινόμηση της πρόβλεψης ως θετική ($\hat{p}_i>c$) ή αρνητική ($\hat{p}_i\leq c$). Κατασκευάζεται ο Πίνακας συνάφειας 2 που δείχνει πόσες προβλέψεις είναι ορθές και πόσες λάθος.

		Νόσος	Ναι	Όχι	
			Y=1	Y=0	
Αποτέλεσμα	Θετικό	Y=1	A (true positive)	B (false positive)	A+B
	Αρνητικό	Y=0	C (false negative)	D (true negative)	C+D
			A+C	B+D	N (σύνολο)

Πίνακας συνάφειας 2

Ορίζουμε ως:

- Ευαισθησία ή ποσοστό θετικών (TPR): $TPR = \frac{A}{A+C}$, που εκφράζει πόσο συχνά έχουμε ορθή πρόβλεψη για την κατάσταση Y=1.

- Ποσοστό ψευδώς θετικών (FPR): $FPR = \frac{B}{B+D}$

- Ειδικότητα ή ποσοστό αληθώς αρνητικών(TNR): $TNR = \frac{D}{B+D} = 1 - FPR$, που εκφράζει πόσο συχνά έχουμε ορθή πρόβλεψη για την κατάσταση Y=0.

- Θετική προβλεπόμενη τιμή (PPV): $PPV = \frac{A}{A+B}$, εκφράζει την πιθανότητα εμφάνισης θετικού περιστατικού μεταξύ όλων των θετικών προβλέψεων.

- Αρνητική προβλεπόμενη τιμή (NPV): $NPV = \frac{D}{C+D}$, εκφράζει την πιθανότητα εμφάνισης αρνητικού περιστατικού μεταξύ όλων των αρνητικών προβλέψεων.

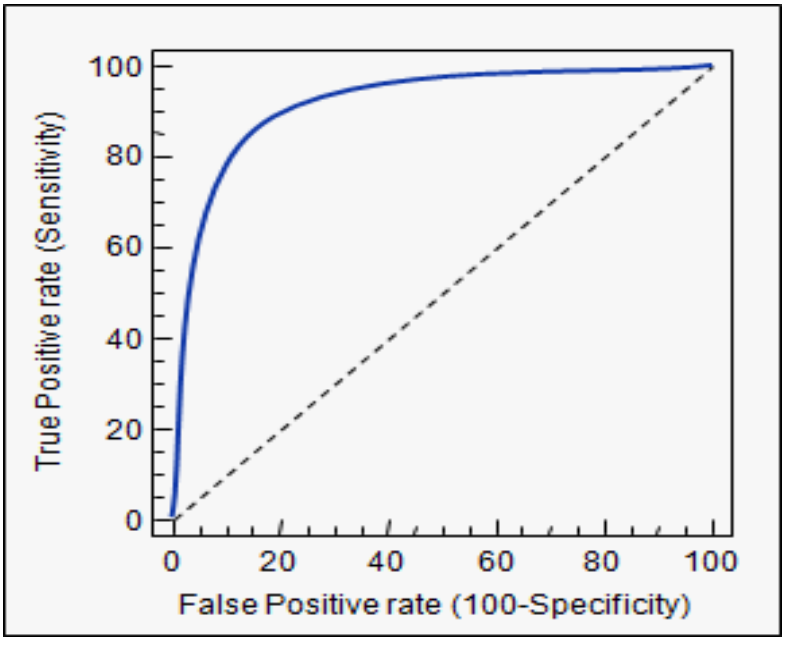
- Ακρίβεια (Accuracy): $ACC = \frac{A+D}{N}$, εκφράζει το ποσοστό των πραγματικών αποτελεσμάτων (τόσο αληθινά θετικά όσο και αληθινά αρνητικά) μεταξύ του συνολικού αριθμού των περιπτώσεων που εξετάστηκαν.

- Θετικός λόγος πιθανοφανειών (LR+): $LR+ = \frac{TPR}{FPR}$, εκφράζει πόσες φορές πιο συχνά εμφανίζεται το θετικό αποτέλεσμα σε αυτούς που έχουν το νόσημα σε σχέση με αυτούς που δεν το έχουν.

- Αρνητικός λόγος πιθανοφανειών (LR-): $LR- = \frac{FPR}{TPR}$, εκφράζει πόσες φορές πιο συχνά εμφανίζεται το αρνητικό αποτέλεσμα σε αυτούς που δεν έχουν το νόσημα σε σχέση με αυτούς που το έχουν.

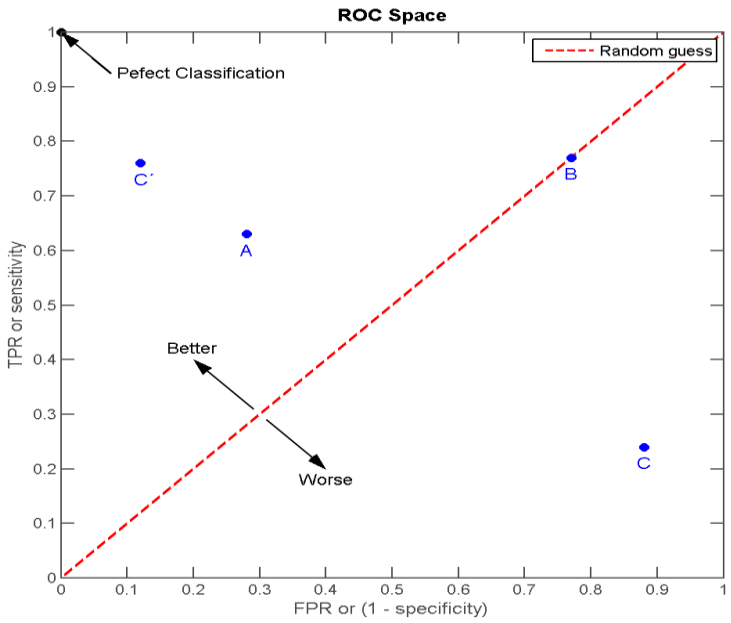
DOR (diagnostic odds ratio): $DOR = \frac{LR+}{LR-}$, αποτελεί μέτρο αποτελεσματικότητας ενός διαγνωστικού τεστ. Όταν είναι μεγαλύτερο της μονάδας ή υψηλότερα τότε είναι πολύ αποτελεσματικό.

Η καμπύλη ROC είναι μια γραφική παράσταση της ευαισθησίας (sensitivity) έναντι του 1-ειδικότητα (specificity) ή ψευδώς θετικών για ένα σύστημα με διχότομο αποτέλεσμα. Ισοδύναμα η καμπύλη ROC μπορεί να σχεδιαστεί ως το ποσοστό των αληθώς θετικών (TPR=True Positive Rate) έναντι του ποσοστού των ψευδώς θετικών(FPR=False Positive Rate). Επίσης μας παρέχει πληροφορία για όλους τους συνδυασμούς μεταξύ των αληθώς θετικών (TP) και ψευδώς θετικών (FP) τιμών. Το γράφημα της καμπύλης έχει ως άξονα ψ την ευαισθησία (TP) και ως άξονα χ τον 1-ειδικότητα(1-FPR). Η καμπύλη ορίζεται ως το μοναδιαίο τετράγωνο [0,1]x[0,1] το οποίο ξεκινά από το σημείο (0,0) και καταλήγει στο σημείο (1,1). Επιθυμητό είναι να έχουμε υψηλή ευαισθησία και χαμηλή 1-ειδικότητα άρα η καλύτερη μέθοδος πρόβλεψης θα μας έδινε ένα σημείο στην επάνω αριστερή γωνία του χώρου.



Σχήμα 4.1 : Γραφική απεικόνιση καμπύλης ROC

Η διαγώνιος χωρίζει τον χώρο ROC σχηματίζοντας 45 μοίρες με τους άξονες όπως φαίνεται από το Σχήμα 4.1. Τα σημεία πάνω από την διαγώνιο αντιπροσωπεύουν καλά αποτελέσματα ταξινόμησης ενώ τα σημεία κάτω από την διαγώνιο παρουσιάζουν κακά αποτελέσματα. Παρατηρούμε όμως ότι ένα φτωχό προγνωστικό μπορεί να αποκτήσει σημεία πάνω από την διαγώνιο, αν τα αποτελέσματα του πίνακα συνάφειας αντιστραφούν. Ας εξετάσουμε τώρα τέσσερα αποτελέσματα πρόβλεψης:



Σχήμα 4.2 : Αντικατοπτρισμός των τεσσάρων σημείων στον χώρο ROC

Από το γράφημα του Σχήματος 4.2 βλέπουμε ότι το σημείο C έχει την χειρότερη προβλεπτική ικανότητα σε σχέση με τα A,B,D. Ωστόσο αν πάρουμε το συμμετρικό του C το C' βλέπουμε ότι είναι καλύτερο και από το A. Αυτή η μέθοδος αντικατοπτρισμού αντιστρέφει τις προβλέψεις οποιαδήποτε μεθόδου. Το σημείο B ακουμπάει την διαγώνιο άρα η ακρίβεια της B είναι στο 50%. Έτσι βλέπουμε πως παρόλο που η αρχική μέθοδος έχει αρνητική προβλεπτική ικανότητα, απλά αντιστρέφοντας τις αποφάσεις της οδηγούμαστε σε μια νέα μέθοδο C' με θετική προβλεπτική ικανότητα. Όσο πιο κοντά βρίσκεται ένα σημείο στην επάνω αριστερή γωνία τόσο μεγαλύτερη δύναμη πρόβλεψης έχει αυτή η μέθοδος.

(https://en.wikipedia.org/wiki/Receiver_operating_characteristic)

4.2 Επεκτάσεις της ευαισθησίας και ειδικότητας

Ορίζουμε τώρα τις incident/dynamic εκδοχές της ευαισθησίας (sensitivity) και της ειδικότητας (specificity). Σε κάθε χρονική στιγμή t, οι ασθενείς που βρίσκονται σε ρίσκο (riskset) να συμβεί το γεγονός χωρίζεται σε δύο ξεχωριστά group, σε αυτούς που πεθαίνουν (case) και σε αυτούς που επιβιώνουν (control). Έτσι σε οποιονδήποτε χρόνο ορίζουμε

$$sensitivity^{I}(c,t) \ : \ \Pr\left\{M > c \middle| T = t\right\} = \Pr\left\{M > c \middle| dN^{*}(t) = 1\right\}$$

$$specificity^{D}(c,t) \ : \ \Pr\left\{M \le c \middle| T \ge t\right\} = \Pr\left\{M \le c \middle| N^{*}(t) = 0\right\}$$

Χρησιμοποιώντας της παραπάνω εξισώσεις μπορούμε να ορίσουμε την αντίστοιχη καμπύλη ROC $ROC(t)$ για κάθε χρονική στιγμή t. Χρησιμοποιώντας αυτή την προσέγγιση για ένα άτομο i μπορεί να παίξει ρόλο control για τον αρχικό χρόνο , $t < T_i$, αλλά μετά παίζει το ρόλο του case όταν $t = T_i$. Εδώ η ευαισθησία μετρά την το αναμενόμενο ποσοστό των ατόμων με δείκτη μεγαλύτερο του c ανάμεσα στο πληθυσμό των ατόμων που πεθαίνουν την χρονική στιγμή t, ενώ η ειδικότητα μετράει το ποσοστό των ατόμων με δείκτη μικρότερο ή ίσο του c ανάμεσα σε αυτούς που επιβιώνουν την χρονική στιγμή t.

4.3 Χρονοεξαρτώμενες καμπύλες ROC

Αφού ορίσαμε τις incident ευαισθησία και dynamic ειδικότητα, μπορούμε να σχεδιάσουμε καμπύλες ROC. Εστιάζουμε στις (I/D) καμπύλες ROC που ορίζονται ως $ROC_t^{I/D}(p)$, όπου p δηλώνει το ποσοστό των dynamic ψευδώς θετικών (1-ειδικότητα), και $ROC_t^{I/D}(p)$ δηλώνει το αντίστοιχο ποσοστό των αληθών θετικών. Ειδικότερα ας θεωρήσουμε ότι το c^p είναι η κατώτερη τιμή που αποφέρει ένα ποσοστό ψευδών θετικών $p : P(M_i > c^p \,|T_i > t) = 1 - specificity^{D}(c^p, t) = p$. Το ποσοστό των αληθών θετικών, $ROC_t^{I/D}(p)$, είναι η ευαισθησία που παίρνουμε χρησιμοποιώντας το συγκεκριμένο κατώτατο όριο δηλαδή $ROC_t^{I/D}(p) = sensitivity^{I}(c^p, t) = P(M_i > c^p \,|T_i = t)$.

Χρησιμοποιώντας τις συναρτήσεις των αληθών και ψευδών θετικών ποσοστών $TP_t^I(c) = sensitivity^I(c,t)$ και $FP_t^D(c) = 1 - specificity^D(c,t)$ η καμπύλη ROC μπορεί να εκφραστεί ως η σύνθεση της $TP_t^I(c)$ και της αντίστροφης $[FP_t^D]^{-1}(p) = c^p$:

$$ROC_t^{I/D}(p) = TP_t^I(c)\left\{[FP_t^D]^{-1}(p)\right\}$$

για $p \in [0,1]$. Για το εμβαδό AUC (Area Under Curve) της I/D ROC καμπύλης χρησιμοποιούμε τον τύπο

$$AUC(t) = \int_0^1 ROC_t^{I/D}(p)dp$$

Η περιοχή κάτω από την καμπύλη ROC που ονομάζεται AUC (Area Under Curve) παριστάνει ένα μέτρο συμφωνίας ανάμεσα στην τιμή του δείκτη και στον δείκτη κατάστασης της νόσου (Hanley and McNeil, 1982). Ειδικότερα το AUC μετράει την πιθανότητα η τιμή του δείκτη, για μια τυχαία επιλεγμένη περίπτωση (case), να ξεπεράσει την τιμή του δείκτη για ένα τυχαία επιλεγμένο έλεγχο (control).

4.4 Χρονοεξαρτώμενα AUC

Στην ακριβώς προηγούμενη ενότητα είδαμε πώς οι καμπύλες ROC μπορούν να χρησιμοποιηθούν για να χαρακτηρίσουν την ικανότητα του δείκτη (marker) να ξεχωρίζει αυτούς που πεθαίνουν στο χρόνο t (cases) από αυτούς που πεθαίνουν στο χρόνο t (controls). Ωστόσο σε πολλές εφαρμογές δεν γνωρίζουμε εκ των προτέρων των χρόνο . Θα δούμε λοιπόν πώς οι χρονοεξαρτώμενες καμπύλες ROC σχετίζονται με ένα τυποποιημένο αποτέλεσμα «συμφωνίας» (“concordance”). Το γενικό αποτέλεσμα που χρησιμοποιούμε είναι

$$C = P[M_j > M_k \,|\, T_j > T_k]$$

που εκφράζει την πιθανότητα το άτομο που πέθανε στην αρχή να έχει μεγαλύτερη τιμή δείκτη. Για να καταλάβουμε την σχέση ανάμεσα αυτής της διάκρισης και της καμπύλης ROC, υποθέτουμε ότι οι παρατηρήσεις μας (M_j, T_j) και (M_k, T_k) είναι ανεξάρτητες και επιπλέον ότι ο χρόνος T_j είναι συνεχής έτσι ώστε $P(T_k, T_j) = 0$. Αυτές οι υποθέσεις

οδηγούν στο ότι το αποτέλεσμα συμφωνίας (concordance) είναι ένας σταθμισμένος μέσος όρος του χώρου κάτω από την χρονοεξαρτώμενη καμπύλη ROC,

$$\begin{aligned} & P[M_j > M_k \,|\, T_j < T_k] \\ = \; & 2\int_t P[\{M_j > M_k\} | T_j = t\} \cap \{t < T_k\}] \\ & \times P[\{T_j = t\} \cap \{t < T_k\}]dt \\ = \; & \int_t AUC(t) \times w(t)dt \\ = \; & E_T[AUC(T) \times 2 \times S(T)] \end{aligned}$$

Με $w^\tau(t) = 2 \times f(t) \times S(t) / W^\tau = \int_0^\tau 2 \times f(t) \times S(t)dt = 1 - S(\tau^2)$. Σε αυτήν την έκφραση το $AUC(t)$ βασίζεται στο I/D ορισμό τη ευαισθησίας και της ειδικότητας, $AUC(t) = P(M_j > M_k \,|\, T_j = t, T_k > t)$.

Στην πράξη θα δώσουμε περισσότερο έμφαση σε ένα σταθερό χρονικό διάστημα (0,τ). έτσι το concordance θα τροποποιηθεί για τον πεπερασμένο χρόνο ως

$$C^\tau = \int_0^\tau AUC(t) \times w^\tau(t)dt$$

όπου $w^\tau(t) = 2 \times f(t) \times S(t) / W^\tau$, $W^\tau = \int_0^\tau 2 \times f(t) \times S(t)dt = 1 - S(\tau^2)$.

Η περιορισμένη αυτή μορφή της concordance παραμένει ο σταθμισμένος μέσος όρος των AUCs για την συγκεκριμένη χρονική στιγμή με τα βάρη τροποποιημένα ώστε να πλησιάσουν το 1.0 στο διάστημα (0,τ). Το C^τ είναι αποτελεί μια μικρή τροποποίηση του C, $C^\tau = P[M_j > M_k \,|\, T_j < T_k, T_j < \tau]$. Το C^τ είναι η πιθανότητα ότι οι προβλέψεις για ένα ζευγάρι αντικειμένων, έρχονται σε συμφωνία με τα αποτελέσματα τους, δεδομένου ότι η μικρότερη χρονική στιγμή του συμβάντος πραγματοποιείται στο διάστημα (0,τ). (Heagerty and Zheng, 2005)

ΚΕΦΑΛΑΙΟ 5 – ΠΑΡΑΜΕΤΡΙΚΗ ΑΝΑΛΥΣΗ ΕΠΙΒΙΩΣΗΣ

Όταν προσαρμόσαμε το μοντέλο του Cox για την ανάλυση των δεδομένων επιβίωσης δεν χρειαζόταν να υποθέσουμε ότι υπήρχε κάποια συγκεκριμένη κατανομή που περιέγραφε τους χρόνους επιβίωσης. Αυτό είχε ως αποτέλεσμα η συνάρτηση διακινδύνευσης δεν είχε κάποια συγκεκριμένη συναρτησιακή μορφή και το μοντέλο είχε μια ευελιξία και ευρεία εφαρμογή. Από την άλλη μεριά αν η υπόθεση μια συγκεκριμένης κατανομής είναι έγκυρη, τα συμπεράσματα βάσει αυτής της υπόθεσης θα είναι πιο ακριβή. Πιο συγκεκριμένα οι εκτιμήσεις των ποσοτήτων όπως των σχετικών κινδύνων και των μέσων διάρκειας ζωής θα τείνουν να έχουν πιο μικρά τυπικά σφάλματα από ότι θα είχαν με την απουσία της υπόθεσης της κατανομής. Αυτά τα μοντέλα στα οποία υποθέτουμε ότι ακολουθούν μια συγκεκριμένη κατανομή για τους χρόνου επιβίωσης , ονομάζονται παραμετρικά μοντέλα. Μια κατανομή που παίζει σπουδαίο ρόλο στην ανάλυση δεδομένων επιβίωσης είναι η κατανομή Weibull που παρουσιάστηκε από τον W.Weibull το 1951 στο πλαίσιο των δοκιμών βιομηχανικής αξιοπιστίας. Πράγματι αυτή η κατανομή είναι τόσο σημαντική για την παραμετρική ανάλυση δεδομένων διάρκειας ζωής όσο είναι η κανονική κατανομή στα γραμμικά μοντέλα. Από την ευρεία γκάμα κατανομών που υπάρχουν για τους χρόνους επιβίωσης , εμείς θα επικεντρωθούμε στην Weibull λόγω της ευελιξίας που μπορεί να έχει στην μορφή της για διαφορετικές τιμές των παραμέτρων της. (Collett, 2003)

5.1 Εκθετική κατανομή

Το πιο απλό μοντέλο για την συνάρτηση διακινδύνευσης είναι να υποθέσουμε ότι είναι σταθερή στο πέρασμα του χρόνου. Ο κίνδυνος θανάτου σε οποιαδήποτε χρονική στιγμή μετά από την αρχή της μελέτης είναι ο ίδιος ανεξάρτητα από το χρονικό διάστημα που έχει περάσει. Σύμφωνα με αυτό το μοντέλο η συνάρτηση διακινδύνευσης μπορεί να γραφτεί ως

$$h(t) = \lambda,$$

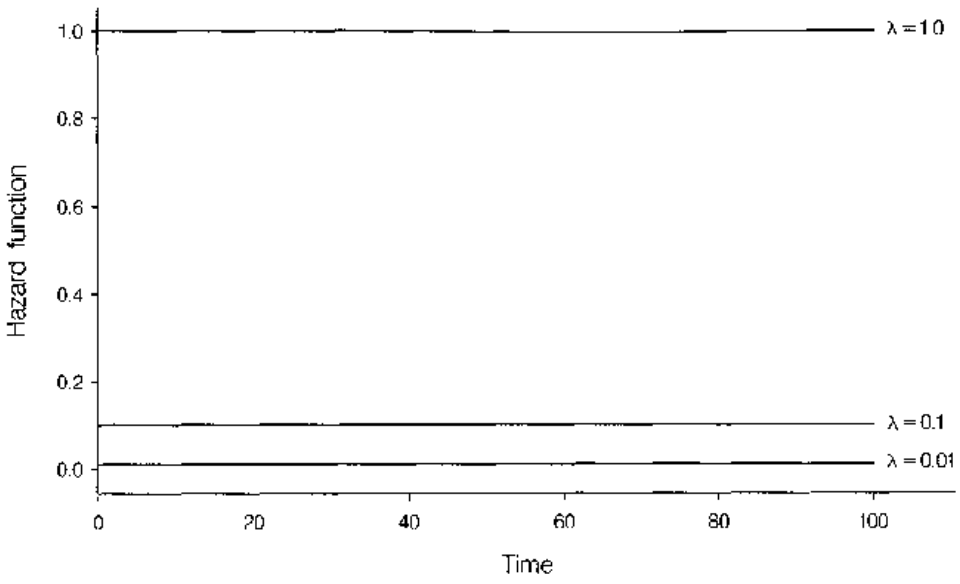
για $0 \le t < \infty$. Η παράμετρος λ είναι μια θετική σταθερά η οποία εκτιμάται από την προσαρμογή του μοντέλου στα παρατηρούμενα δεδομένα. Η αντίστοιχη συνάρτηση επιβίωσης είναι

$$S(t) = \exp\left\{-\int_0^t \lambda du\right\} = e^{-\lambda t}$$

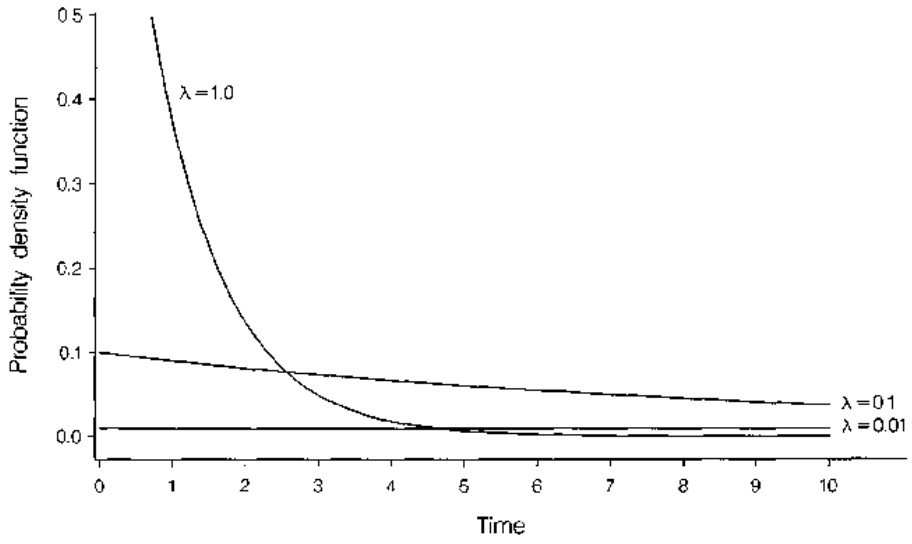
Και άρα η συνάρτηση πυκνότητας πιθανότητας για τους χρόνου επιβίωσης είναι

$$f(t) = \lambda e^{-\lambda t}$$

για $0 \le t < \infty$. Αυτή είναι η συνάρτηση πυκνότητας πιθανότητας για μια τυχαία μεταβλητή T που ακολουθεί την εκθετική κατανομή με μέση τιμή λ^{-1}. Είναι πιο βολικό κάποιες φορές να γράφουμε $\mu = \lambda^{-1}$, ώστε η συνάρτηση διακινδύνευσης να ισούται με μ^{-1} έτσι ώστε η κατανομή του χρόνου επιβίωσης να έχει μέση τιμή μ. Στα γραφήματα 5.1 και 5.2 παρουσιάζονται γραφικά οι συνάρτηση διακινδύνευσης και η αντίστοιχη συνάρτηση πυκνότητας πιθανότητας για την εκθετική κατανομή.



Γράφημα 5.1: Συνάρτηση διακινδύνευσης για την εκθετική κατανομή



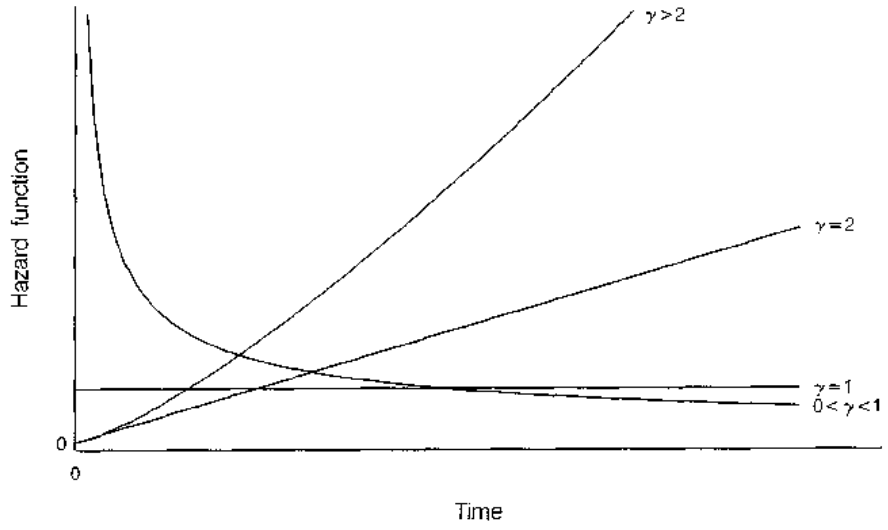
Γράφημα 5.2: Συνάρτηση πυκνότητας πιθανότητας για την εκθετική κατανομή

5.2 Weibull κατανομή

Στην πράξη η υπόθεση μια σταθερής συνάρτησης διακινδύνευσης δηλαδή η χρήση της εκθετικής κατανομής δεν ενδείκνυται. Μια πιο γενική μορφή της συνάρτησης διακινδύνευσης είναι

$$h(t) = \lambda\gamma t^{\gamma-1} \qquad (5.1)$$

για $0 \le t < \infty$, μια συνάρτηση που εξαρτάται από δύο παραμέτρους λ και γ οι οποίες είναι και οι δύο μεγαλύτερες του μηδενός. Στην ειδική περίπτωση που το $\gamma = 1$ η συνάρτηση διακινδύνευσης παίρνει την σταθερή τιμή λ και οι χρόνοι επιβίωσης ακολουθούν την εκθετική κατανομή. Για άλλες τιμές του γ η συνάρτηση διακινδύνευσης αυξάνεται ή μειώνεται μονοτονικά δηλαδή δεν αλλάζει κατεύθυνση. Το σχήμα της συνάρτησης διακινδύνευσης εξαρτάται κυρίως από την τιμή του γ για αυτό και το γ είναι γνωστό ως παράμετρος σχήματος (shape parameter) , ενώ το λ είναι η παράμετρος κλίμακας (scale parameter). Στο γράφημα 5.3 φαίνεται η γενική μορφή της συνάρτησης διακινδύνευσης για τις διάφορες τιμές του γ.



Γράφημα 5.3: Συνάρτηση διακινδύνευσης για την Weibull κατανομή για τις διάφορες τιμές του γ.

Για αυτήν την συγκεκριμένη επιλογή της συνάρτησης διακινδύνευσης, η συνάρτηση επιβίωσης δίνεται από τον τύπο

$$S(t) = \exp\left\{-\int_0^t \lambda\gamma u^{\gamma-1} du\right\} = \exp(-\lambda t^{\gamma}) \qquad (5.2)$$

Η αντίστοιχη συνάρτηση πυκνότητας πιθανότητας είναι

$$f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^{\gamma})$$

για $0 \le t < \infty$, που είναι η πυκνότητα μια τυχαίας μεταβλητής που ακολουθεί την κατανομή Weibull με παράμετρο κλίμακας λ και παράμετρο σχήματος γ.

Η αναμενόμενη τιμή μιας τυχαίας μεταβλητής Τ που ακολουθεί την $W(\lambda,\gamma)$ κατανομή δίνεται ως

$$\mathrm{E(T)} = \lambda^{-1/\gamma}\Gamma(\gamma^{-1}+1),$$

Όπου $\Gamma(x)$ είναι συνάρτηση Γ που ορίζεται από το ολοκλήρωμα

$$\Gamma(x) = \int_0^{\infty} u^{x-1}e^{-u}du$$

Η τιμή αυτού του ολοκληρώματος είναι (x-1)! Και για τις ακέραιες τιμές του x μπορεί πολύ εύκολα να υπολογιστεί. Για να υπολογίσουμε την μέση τιμή για τις μη ακέραιες τιμές του x θα χρειαστεί να χρησιμοποιήσουμε τους πίνακες της συνάρτησης Γ ή κατάλληλο υπολογιστικό πρόγραμμα. Δεδομένου ότι η κατανομή Weibull είναι λοξή μια πιο χρήσιμη έννοια για την κατανομή είναι η διάμεσος. Αυτή είναι η τιμή t(50) τέτοια ώστε $S\{t(50)\} = 0.5$, άρα

$$\exp\left\{-\lambda[t(50)]^{\gamma}\right\} = 0.5$$

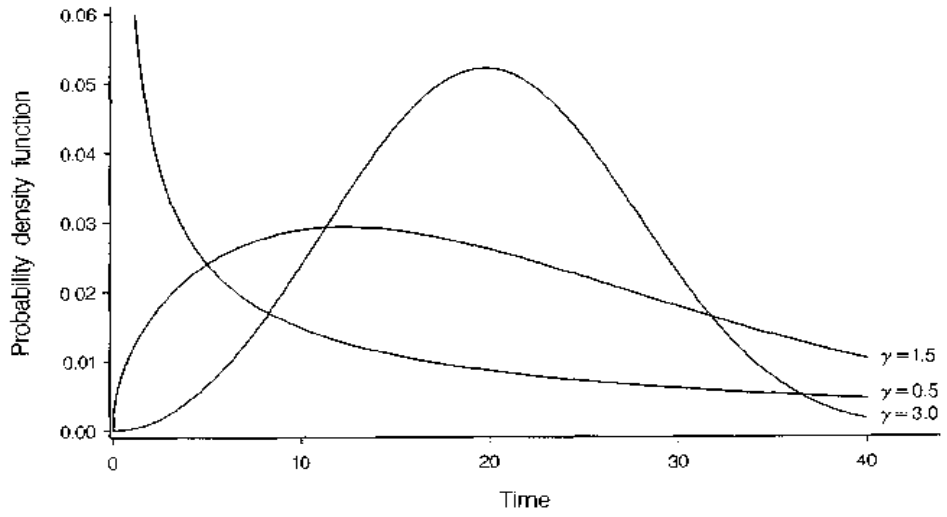
και

$$t(50) = \left\{\frac{1}{\lambda}\log 2\right\}^{1/\gamma}$$

Γενικά για οποιοδήποτε p ποσοστό έχουμε,

$$t(p) = \left\{\frac{1}{\lambda}\log\left(\frac{100}{100-p}\right)\right\}^{1/\gamma} \tag{5.3}$$

Η διάμεσος και άλλα μέτρα της Weibull κατανομής είναι πιο απλά στον υπολογισμό τους σε σχέση με την μέση τιμή. Στο γράφημα 5.4 απεικονίζεται η συνάρτηση πυκνότητας πιθανότητας για διάφορες τιμές του γ.



Γράφημα 5.4: Συνάρτηση πυκνότητας πιθανότητας για την Weibull κατανομή για τις διάφορες τιμές του γ.

Δεδομένου ότι η συνάρτηση διακινδύνευσης της κατανομής Weibull μπορεί να πάρει πολλές μορφές, ανάλογα την τιμή της παραμέτρου γ και κατάλληλα στατιστικά στοιχεία μπορούν εύκολα να βρεθούν, κάνει αυτήν την κατανομή να έχει ευρεία χρήση στην παραμετρική ανάλυση δεδομένων επιβίωσης. (Collett, 2003)

5.3 Αξιολόγηση της καταλληλόλητας ενός παραμετρικού μοντέλου

Πριν την προσαρμογή του μοντέλου με βάση μια υποτιθέμενη παραμετρική μορφή της συνάρτησης διακινδύνευσης , είναι απαραίτητη μια προκαταρκτική μελέτη για την εγκυρότητα της υπόθεσης. Ένας τρόπος θα ήταν να υπολογίσουμε την συνάρτηση διακινδύνευσης χρησιμοποιώντας κατάλληλους τύπους. Αν η συνάρτηση δείχνει να είναι σταθερή τότε αυτό υποδηλώνει ότι η εκθετική κατανομή είναι η κατάλληλη για τα δεδομένα. Από την άλλη μεριά αν η συνάρτηση αυξάνεται ή μειώνεται μονοτονικά με αυξανόμενο χρόνο επιβίωσης ,ένα μοντέλο της κατανομής Weibull θα ήταν το πιο κατάλληλο.

Ένα πιο χρήσιμος τρόπος για να υποθέσουμε ποια κατανομή είναι η καλύτερη για τους χρόνους επιβίωσης, είναι να συγκρίνουμε την συνάρτηση επιβίωσης των δεδομένων του μοντέλου που επιλέξαμε. Αν το γράφημα που προκύπτει απεικονίζει μια ευθεία γραμμή τότε το μοντέλο που επιλέξαμε είναι το κατάλληλο.

Υποθέτουμε ότι έχουμε ένα δείγμα από δεδομένα επιβίωσης και θέλουμε να μελετήσουμε την κατανομή Weibull για τους χρόνους επιβίωσης. Η συνάρτηση επιβίωσης για την κατανομή μας με παράμετρο κλίμακας λ και παράμετρο σχήματος γ δίνεται από τον τύπο

$$S(t) = \exp\left\{-\lambda t^{\gamma}\right\},$$

Παίρνοντας τον λογάριθμο της $S(t)$ και πολλαπλασιάζοντας επί -1 και λογαριθμίζοντας για δεύτερη φορά παίρνουμε

$$\log\left\{-\log S(t)\right\} = \log\lambda + \gamma \log t\,. \qquad (5.4)$$

Αντικαθιστούμε τώρα την εκτιμήτρια Kaplan-Meier της συνάρτησης επιβίωσης $\hat{S}(t)$ στην εξίσωση. Αν η κατανομή Weibull είναι η κατάλληλη, τότε η εκτιμήτρια $\hat{S}(t)$ θα

βρίσκεται κοντά στην $S(t)$, και η γραφική παράσταση της $\log\left\{-\log \hat{S}(t)\right\}$ με την $\log t$ θα είναι περίπου μια ευθεία γραμμή. Από τις σχέσεις που συνδέονται η συνάρτηση επιβίωσης με την σωρευτική συνάρτηση διακινδύνευσης , η $H(t)$ θα είναι ίση με $-\log S(t)$ και άρα η $H(t)=\log\left\{-\log S(t)\right\}$ είναι η λογαριθμοποιημένη σωρευτική συνάρτηση διακινδύνευσης . Αν το γράφημα της αυτής της συνάρτησης δίνει μια ευθεία γραμμή , τότε το γράφημα αυτό μας παρέχει μια πρόβλεψη για τις δύο παραμέτρους της κατανομής Weibull. Πιο συγκεκριμένα από την παραπάνω εξίσωση η σταθερά και η κλίση της ευθείας γραμμής θα είναι το logλ και γ αντίστοιχα. Έτσι η κλίση της γραμμής του γραφήματος της λογαριθμοποιημένης σωρευτικής συνάρτησης διακινδύνευσης δίνει μια εκτίμηση της παραμέτρου σχήματος και ο εκθέτης της σταθεράς μας δίνει την εκτίμηση της παραμέτρου κλίμακας. (Collett, 2003)

5.4 Προσαρμογή του παραμετρικού μοντέλου

Τα παραμετρικά μοντέλα μπορούν να προσαρμοστούν σε ένα σύνολο δεδομένων επιβίωσης με την χρήση της μεθόδου της μέγιστης πιθανοφάνειας. Ας θεωρήσουμε αρχικά την κατάσταση όπου έχουν παρατηρηθεί πραγματικοί χρόνοι επιβίωσης για n άτομα και δεν υπάρχουν λογοκριμένες παρατηρήσεις. Αν η συνάρτηση πυκνότητας πιθανότητας μιας τυχαίας μεταβλητής που συνδέεται με τον χρόνο επιβίωσης είναι $f(t)$, η πιθανοφάνεια των n παρατηρήσεων $t_1, t_2, ..., t_n$ είναι το γινόμενο

$$\prod_{i=1}^{n} f(t_i)$$

Η πιθανοφάνεια θα είναι μια συνάρτηση των άγνωστων παραμέτρων στην συνάρτηση πυκνότητας πιθανότητας και η μέγιστη πιθανοφάνεια εκτιμά αυτές τις παραμέτρους όπου, από αυτές τις τιμές, η συνάρτηση πιθανοφάνειας θα μεγιστοποιείται. Στην πράξη γενικότερα είναι πιο βολικό να δουλεύουμε με τον λογάριθμο της συνάρτηση πιθανοφάνειας. Οι τιμές των άγνωστων παραμέτρων της συνάρτηση πυκνότητας που μεγιστοποιούν την λογαριθμοποιημένη πιθανοφάνεια είναι φυσικά οι ίδιες τιμές που μεγιστοποιούν την ίδια την συνάρτηση πιθανοφάνειας.

Θεωρούμε τώρα μια πιο συνηθισμένη κατάσταση όπου τα δεδομένα επιβίωσης περιλαμβάνουν λογοκριμένες παρατηρήσεις. Ειδικότερα υποθέτουμε ότι r το πλήθος από το συνολικό δείγμα n πεθαίνουν τις χρονικές στιγμές $t_1, t_2, ..., t_r$ και οι χρονικές στιγμές αυτών που απομένουν n-r, $t_1^*, t_2^*, ..., t_{n-r}^*$ είναι δεξιά αποκομμένες. Οι r θάνατοι συνεισφέρουν με τον όρο

$$\prod_{j=1}^{r} f(t_j)$$

στην συνολική συνάρτηση πιθανοφάνειας. Φυσικά δεν μπορούμε να αγνοήσουμε την πληροφορία επιβίωσης για τα n-r για τους οποίους έχει καταγραφεί αποκοπή. Αν ένας χρόνος επιβίωσης είναι αποκομμένος την χρονική στιγμή t^*, γνωρίζουμε ότι η πιθανότητα να συμβεί το γεγονός είναι $P(T \geq t^*)$ το οποίο είναι $S(t^*)$. Έτσι κάθε αποκομμένη παρατήρηση συνεισφέρει σαν όρος στην πιθανοφάνεια για το σύνολο των παρατηρήσεων. Άρα η συνολική συνάρτηση πιθανοφάνειας είναι

$$\prod_{j=1}^{r} f(t_j) \prod_{i=1}^{n-r} S(t_i^*) \tag{5.5}$$

Όπου το πρώτο γινόμενο αφορά τους χρόνους θανάτου για τους r και το δεύτερο γινόμενο για τους n-r αποκομμένους χρόνους επιβίωσης. Πιο αναλυτικά υποθέτουμε ότι τα δεδομένα θεωρούνται ως n ζεύγη παρατηρήσεων, όπου το ζευγάρι του i-οστού ατόμου είναι $(t_i, \delta_i), i = 1, 2, ..., n.$ Σε αυτό το ζεύγος το δ_i είναι μια δείκτρια μεταβλητή που παίρνει την τιμή 0 όταν ο χρόνος επιβίωσης t_i είναι αποκομμένος και μονάδα όταν δεν είναι. Έτσι η συνάρτηση πιθανοφάνειας παίρνει την μορφή

$$\prod_{i=1}^{n} \{ f(t_i) \}^{\delta_i} \{ S(t_i) \}^{1-\delta_i} \tag{5.6}$$

Αυτή η συνάρτηση η οποία είναι ισοδύναμη με την (5.5) μπορεί να μεγιστοποιηθεί σε σχέση με τις άγνωστες παραμέτρους στις συναρτήσεις πυκνότητας και επιβίωσης. Μια εναλλακτική έκφραση για την συνάρτηση πιθανοφάνειας μπορεί να είναι

$$\log L(\lambda,\gamma) = r\log(\lambda\gamma) + (\gamma-1)\sum_{i=1}^{n}\delta_i \log t_i - \lambda\sum_{i=1}^{n} t_i^{\gamma} \ ,$$

Και χρησιμοποιώντας την σχέση που συνδέει την συνάρτηση πυκνότητας πιθανότητας με την συνάρτηση διακινδύνευσης παίρνουμε

$$\prod_{i=1}^{n}\{h(t_i)\}^{\delta_i} S(t_i)\,. \qquad (5.7)$$

Αυτή η μορφή της συνάρτησης πιθανοφάνειας είναι ιδιαίτερα χρήσιμη όταν η συνάρτηση πυκνότητας πιθανότητας έχει περίπλοκη μορφή, που συνήθως έχει. Οι εκτιμήσεις των άγνωστων παραμέτρων σε αυτήν την συνάρτηση πιθανοφάνειας μπορούν να βρεθούν μεγιστοποιώντας τον λογάριθμο της συνάρτησης πιθανοφάνειας. (Collett, 2003)

<u>Προσαρμογή του παραμετρικού μοντέλου της κατανομής Weibull</u>

Για την προσαρμογή του παραμετρικού μοντέλου της κατανομής Weibull με παράμετρο κλίμακας λ και παράμετρο σχήματος γ, θεωρούμε ότι οι χρόνοι επιβίωσης n ατόμων είναι ένα δείγμα που έχει αποκομμένες παρατηρήσεις. Υποθέτουμε ότι έχουμε r θανάτους στο σύνολο n των ατόμων και n-r είναι οι δεξιά αποκομμένοι χρόνοι επιβίωσης. Θα χρησιμοποιήσουμε την μέθοδο μέγιστης πιθανοφάνειας στο δείγμα μας. Η συνάρτηση πυκνότητας πιθανότητας, επιβίωσης και διακινδύνευσης για την $W(\lambda,\gamma)$ κατανομή δίνονται από τους τύπους

$$f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^{\gamma})\,,\ \ S(t) = \exp(-\lambda t^{\gamma})\,,\ \ h(t) = \lambda\gamma t^{\gamma-1}$$

Και αντικαθιστώντας στην σχέση (5.6) παίρνουμε

$$\prod_{i=1}^{n}\left\{\lambda\gamma t_i^{\gamma-1}\exp(-\lambda t^{\gamma})\right\}^{\delta_i}\left\{\exp(-\lambda t_i^{\gamma})\right\}^{1-\delta_i} \ ,$$

Όπου δ_i είναι μηδέν αν η i-οστή παρατήρηση είναι αποκομμένη και μονάδα αν δεν είναι. Ισοδύναμα από την έκφραση (5.7) παίρνουμε

$$\prod_{i=1}^{n}\left\{\lambda\gamma t_i^{\gamma-1}\right\}^{\delta_i}\exp(-\lambda t_i^{\gamma})$$

Αυτή θεωρείται ως μια συνάρτηση των λ και γ δηλαδή των άγνωστων παραμέτρων της κατανομής Weibull και άρα μπορεί να γραφτεί ως $L(\lambda,\gamma)$. Η αντίστοιχη συνάρτηση πιθανοφάνειας είναι

$$\log L(\lambda,\gamma)=\sum_{i=1}^{n}\delta_i\log(\lambda\gamma)+(\gamma-1)\sum_{i=1}^{n}\delta_i\log t_i-\lambda\sum_{i=1}^{n}t_i^{\gamma}\ ,$$

Όπου $\sum_{i=1}^{n}\delta_i=r$ και άρα η λογαριθμοποιημένη συνάρτηση γίνεται

$$\log L(\lambda,\gamma)=r\log(\lambda\gamma)+(\gamma-1)\sum_{i=1}^{n}\delta_i\log t_i-\lambda\sum_{i=1}^{n}t_i^{\gamma} \qquad (5.8)$$

Η μέγιστη πιθανοφάνεια εκτιμά τις παραμέτρους λ και γ κάνοντας μερική παραγώγιση ως προς λ και γ της εξίσωσης (5.8) και εξισώνοντας τον κάθε όρο με το 0 παίρνουμε τις εκτιμήτριες $\hat{\lambda}$ και $\hat{\gamma}$. Τα αποτελέσματα των εξισώσεων που προκύπτουν είναι

$$\frac{r}{\hat{\lambda}}-\sum_{i=1}^{n}t_i^{\hat{\gamma}}=0 \qquad \text{και} \qquad \frac{r}{\hat{\gamma}}+\sum_{i=1}^{n}\delta_i\log t_i-\hat{\lambda}\sum_{i=1}^{n}t_i^{\hat{\gamma}}\log t_i=0$$

Από την πρώτη παίρνουμε ότι $\hat{\lambda}=r/\sum_{i=1}^{n}t_i^{\hat{\gamma}}$ και αντικαθιστώντας το $\hat{\lambda}$ στην δεύτερη παίρνουμε την εξίσωση $\frac{r}{\hat{\gamma}}+\sum_{i=1}^{n}\delta_i\log t_i-\frac{r}{\sum_{i=1}^{n}t_i^{\hat{\gamma}}}\sum_{i=1}^{n}t_i^{\hat{\gamma}}\log t_i=0$

Αυτή είναι μια μη γραμμική με άγνωστο το $\hat{\gamma}$ η οποία μπορεί να λυθεί μόνο κάποια επαναληπτική αριθμητική μέθοδο. Όταν βρεθεί η εκτιμήτρια $\hat{\gamma}$ που ικανοποιεί την εξίσωση τότε μπορούμε πολύ εύκολα να βρούμε και το $\hat{\lambda}$. Στην πράξη μια επαναληπτική διαδικασία Newton-Raphson χρησιμοποιείται για να βρεθούν οι τιμές των $\hat{\gamma}$ και $\hat{\lambda}$ οι οποίες μεγιστοποιούν την συνάρτηση πιθανοφάνειας ταυτόχρονα.

Έχοντας βρει τις εκτιμήσεις των παραμέτρων των λ και γ από την προσαρμογή της Weibull κατανομής στα δεδομένα, μπορούμε να υπολογίσουμε ποσοστά του χρόνου επιβίωσης χρησιμοποιώντας την παρακάτω εξίσωση

$$\hat{t}(p) = \left\{ \frac{1}{\hat{\lambda}} \log\left(\frac{100}{100-p} \right) \right\}^{1/\hat{\gamma}}$$

Και άρα η εκτιμώμενη μέση διάρκεια ζωής είναι

$$\hat{t}(50) = \left\{ \frac{1}{\hat{\lambda}} \log 2 \right\}^{1/\hat{\gamma}}$$

Το τυπικό σφάλμα του εκτιμώμενου ποσοστού λαμβάνεται χρησιμοποιώντας μια γενίκευση της εξίσωσης $\operatorname{var}\left\{ g(\hat{\lambda}) \right\} \approx \left\{ \frac{dg(\hat{\lambda})}{d\hat{\lambda}} \right\} \operatorname{var}(\hat{\lambda})$ στην περίπτωση που η κατά προσέγγιση διακύμανση μιας συνάρτηση δύο παραμέτρων, δηλαδή

$$se\left\{ \hat{t}(p) \right\} = \frac{\hat{t}(p)}{\hat{\lambda}\hat{\gamma}^2} \left\{ \hat{\gamma}^2 \operatorname{var}(\hat{\lambda}) + \hat{\lambda}^2 (c_p - \log\hat{\lambda})^2 \operatorname{var}\hat{\gamma} + 2\hat{\lambda}\hat{\gamma}(c_p - \log\hat{\lambda}) \operatorname{cov}(\hat{\lambda}, \hat{\gamma}) \right\}^{\frac{1}{2}}$$

όπου $c_p = \log\log\left(\frac{100}{100-p} \right)$.

Οι διακυμάνσεις των $\hat{\lambda}$ και $\hat{\gamma}$ βρίσκονται από τον πίνακα διακύμανσης-συνδιακύμανσης των εκτιμητριών. (Collett, 2003)

Τα διαστήματα εμπιστοσύνης των πραγματικών τιμών του ποσοστιαίου σημείου, $t(p)$, είναι προτιμότερο να το βρούμε από το αντίστοιχο διάστημα για το $\log t(p)$.Το τυπικό σφάλμα του $\log\hat{t}(p)$ είναι

$$se\left\{ \log\hat{t}(p) \right\} = \frac{1}{\hat{t}(p)} se\left\{ \hat{t}(p) \right\}$$

και τα $100(1-a)\%$ όρια του διαστήματος είναι

$$\log\hat{t}(p) \pm z_{a/2} se\left\{ \log\hat{t}(p) \right\}.$$

ΚΕΦΑΛΑΙΟ 6 - ΠΑΡΟΥΣΙΑΣΗ ΔΕΙΓΜΑΤΟΣ ΚΑΙ ΜΕΤΑΒΛΗΤΩΝ - ΑΝΑΛΥΣΗ ΕΠΙΒΙΩΣΗΣ ΣΕ ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ

6.1 Παρουσίαση δείγματος και μεταβλητών

Το δείγμα μας αποτελείται από 75 ασθενείς που νοσηλεύτηκαν στο νοσοκομείο Ευγενίδειο Θεραπευτήριο, τη χρονική περίοδο Μάιο 2005 –Σεπτέμβριο 2015 και οι οποίοι υποβλήθηκαν σε χειρουργική εκτομή καρκίνου του πνεύμονα. Τα δεδομένα των ασθενών προήλθαν από τις ιστολογικές εξετάσεις που πραγματοποιήθηκαν από το παθολογοανατομικό εργαστήριο του Ευγενιδίου Θεραπευτηρίου του Πανεπιστημίου Αθηνών και ένα μικρό μέρος τους παρουσιάζεται στον Πίνακα 6.1.

ασθενής	t	c	Ηλικία	Φύλο	καπνιστής	είδος εκτομής	λεμφαδένες	διάμετρος	Στάδια	Υπο-στάδια	Ιστ/κός τύπος	βαθμός διαφ/σης
1	43	1	58	2	1	1	8	2.5	1	2	1	2
2	66	1	78	1	1	1	6	2.0	1	1	1	2
3	54	1	74	2	1	1	4	7.0	2	4	1	2
4	136	0	72	1	1	1	6	5.0	2	3	2	3
5	135	0	75	2	1	1	4	7.0	2	4	1	2

Πίνακας 6.1

Η εξαρτημένη μεταβλητή ως προς τις οποίες εξετάσαμε το δείγμα μας είναι η θνητότητα όπως φαίνεται από την Πίνακα 6.2

Μεταβλητή	Κωδικοποίηση	
Θνητότητα	0	1
	Επιβίωση	θάνατος

Πίνακας 6.2

Οι ανεξάρτητες ή επεξηγηματικές μεταβλητές είναι οι εξής:

Συνεχείς μεταβλητές

Οι συνεχείς μεταβλητές είναι οι Ηλικία , αριθμός λεμφαδένων , διάμετρος όγκου όπως φαίνεται στον Πίνακα 6.3.

Μεταβλητές	Μέση τιμή	Διάμεσος	Ελάχιστο	Μέγιστο
Ηλικία	67,2	69	45	87
Αριθμός λεμφαδένων	6	6	2	12
Διάμετρος όγκου	4,2	4	0.3	7

Πίνακας 6.3

Κατηγορικές μεταβλητές

Στις κατηγορικές μεταβλητές έχουν γίνει οι παρακάτω κωδικοποιήσεις :

- Στη μεταβλητή φύλο έχει γίνει η κωδικοποίηση 1: άνδρας 2: γυναίκα
- Στη μεταβλητή καπνιστής έχει γίνει η κωδικοποίηση 0: όχι 1: ναι
- Στη μεταβλητή είδος εκτομής έχει γίνει η κωδικοποίηση 1: λοβεκτομή
  2: τμηματοεκτομή

Αποτελούν δύο βασικές τεχνικές αφαίρεσης ενός όγκου. Στην Τμηματική εκτομή αφαιρείται ένα μεγαλύτερο τμήμα του πνεύμονα, αλλά όχι όλος ο λοβός. Στην Λοβεκτομή αφαιρείται όλος ο λοβός του ενός πνεύμονα.

- Στη μεταβλητή ιστολογικός τύπος έχει γίνει η κωδικοποίηση
  1: αδενοκαρκίνωμα
  2: πλακώδες

Δύο είναι οι κύριοι τύποι του καρκίνου του πνεύμονα: ο μη μικροκυτταρικός και ο μικροκυτταρικός. Ο όρος «μικροκυτταρικός» αναφέρεται στο μέγεθος και το σχήμα των κυττάρων όπως αυτά φαίνονται στο μικροσκόπιο. Πιο συχνός είναι ο μη μικροκυτταρικός τύπος, ο οποίος προέρχεται από τα επιθηλιακά κύτταρα του πνεύμονα ενώ ο μικροκυτταρικός τύπος προέρχεται από νευρικά και νευρο-ενδοκρινικά κύτταρα. Τα παραπάνω χαρακτηριστικά περιγράφονται και καταγράφονται από ειδικευμένο παθολογο-ανατόμο κατά την ιστολογική εξέταση του καρκινικού ιστού ή κατά την κυτταρολογική εξέταση. Η διάκριση

μεταξύ μη μικροκυτταρικού και μικροκυτταρικού καρκίνου του πνεύμονα είναι απαραίτητη, γιατί ανάλογα με τον τύπο καθορίζεται διαφορετικό είδος θεραπείας. Υπάρχουν διάφοροι τύποι, καθένας από τους οποίους χαρακτηρίζεται από διαφορετικά είδη κυττάρων. Τα κύτταρα αυτά πολλαπλασιάζονται και εξαπλώνονται με διαφορετικό τρόπο. Με βάση τις ιδιότητές τους και τη μορφολογία τους, καθορίζεται ο ιστολογικός τύπος του μη μικροκυτταρικού καρκίνου του πνεύμονα. Οι κυριότεροι τύποι είναι:

- ✓ Πλακώδες καρκίνωμα, το οποίο απαρτίζεται από πλακώδη κύτταρα
- ✓ Αδενοκαρκίνωμα, το οποίο απαρτίζεται από κύτταρα που έχουν εκκριτικές ιδιότητες.
- ✓ Μεγαλοκυτταρικό καρκίνωμα, το οποίο απαρτίζεται από μεγάλα κύτταρα.
- ✓ Αδενοπλακώδες καρκίνωμα, το οποίο απαρτίζεται από κύτταρα πλακώδους μορφολογίας, τα οποία όμως έχουν εκκριτικές ιδιότητες.

Στην περίπτωση μας συναντήσαμε τους δύο πρώτους από τους παραπάνω τύπους μη μικροκυτταρικκού καρκίνου.

- Στη μεταβλητή <u>στάδια</u> έχει γίνει η κωδικοποίηση 1: αν ο όγκος είναι ≤ 3 cm(T1)
  2: αν ο όγκος είναι > 3 cm(T2)

Το σύστημα σταδιοποίησης του καρκίνου του πνεύμονα βασίζεται στην κλινική ταξινόμηση των παραγόντων **TNM** (TUMOR NODE METASTASIS) όπου:

T- Η έκταση του πρωτογενούς όγκου
N- παρουσία ή απουσία καθώς και έκταση λεμφαδενικών μεταστάσεων
M- παρουσία ή απουσία απομακρυσμένων μεταστάσεων

Η σταδιοποίηση αποτελεί μία μέθοδο προσδιορισμού του μεγέθους και της επέκτασης της νόσου σε άλλα μέρη του σώματος, παράγοντες καθοριστικής σημασίας για την πρόγνωση και την επιλογή της καλύτερης δυνατής θεραπείας.

- Στη μεταβλητή <u>υπο-στάδια</u> έχει γίνει η κωδικοποίηση :

1: αν ο όγκος είναι $t \leq 2$ cm (στάδιο Τ1α)

2: αν ο όγκος είναι $2 < t \leq 3$cm (στάδιοΤ1β)

3: αν ο όγκος είναι $3 < t \leq 5$ cm (στάδιοΤ2α)

4: αν ο όγκος είναι $5 < t \leq 7$ cm (στάδιοΤ2β)

Τα υπο-στάδια Τ1α, Τ1β ανήκουν στο στάδιο Τ1, ενώ τα Τ2α ,Τ2β στο στάδιο Τ2.

- Στη μεταβλητή <u>βαθμός διαφοροποίησης</u> έχει γίνει η κωδικοποίηση : 1: υψηλός

2: μέσος

3: χαμηλός

✓ **Βαθμός 1** (Grade I) = Καλά διαφοροποιημένο (τα κύτταρα φαίνονται πιο όμοια με τα φυσιολογικά και δεν αυξάνεται ο όγκος με ταχείς ρυθμούς)

✓ **Βαθμού 2** (Grade II) = Μέτρια διαφοροποιημένα (τα κύτταρα μοιάζουν κάπως διαφορετικά από τα φυσιολογικά)

✓ **Βαθμός 3** (Grade III) = Ελάχιστα διαφοροποιημένα (κύτταρα ακανόνιστα παθολογικά που μπορούν να μεγαλώνουν και να εξαπλώνονται πιο επιθετικά από ό, τι στους άλλους βαθμούς).

Ο βαθμός διαφοροποίησης αντανακλά την επιθετικότητα του όγκου, για αυτό όσο πιο υψηλός είναι ο βαθμός διαφοροποίησης, τόσο πιο επιθετικός είναι ο όγκος.

Το δείγμα μας αποτελείται απο 75 ασθενείς εκ των οποίων οι 55 (73%) είναι άντρες και οι 20 (27%) είναι γυναίκες. Η διάμεση ηλικία των ασθενών είναι 69 ετών. Οι καπνίζοντες ανέρχονται στους 62 (83%) και οι μη-καπνιστές μόλις στους 13(17%). Όσον αφορά το είδος εκτομής, οι 59 (79%) υποβλήθηκαν σε λοβεκτομή και οι υπόλοιποι 16 (21%) σε τμηματοεκτομή. Μέγεθος όγκου >3 cm (Τ2) είχαν οι 31 (41%) ενώ ≤ 3cm (Τ1)είχαν οι 44 (59%). Πιο συγκεκριμένα έχουμε την κατηγοριοποίηση σε υπο-στάδια: στο στάδιο Τ1α(<=2cm) ήταν οι 24(32%) ,στο στάδιο Τ1β (2-3cm) οι 20(27%), στο στάδιο Τ2α (3-5cm) οι 19 (25%) και τέλος στο στάδιο Τ2β (5-7cm) οι 12 (16%). Επιπλέον για τον ιστολογικό τύπο του όγκου, οι 50 (67%) διαγνώστηκαν με αδενοκαρκίνωμα και οι υπόλοιποι 25(33%) με πλακώδες. Υψηλό βαθμό

διαφοροποίησης είχαν οι 13 (17%), μέσο οι 39 (52%) και χαμηλό οι 23 (31%). Τέλος το ποσοστό επιβίωσης φτάνει το 63%. Στον Πίνακα 6.4 παρουσιάζονται συγκεντρωτικά τα στατιστικά.

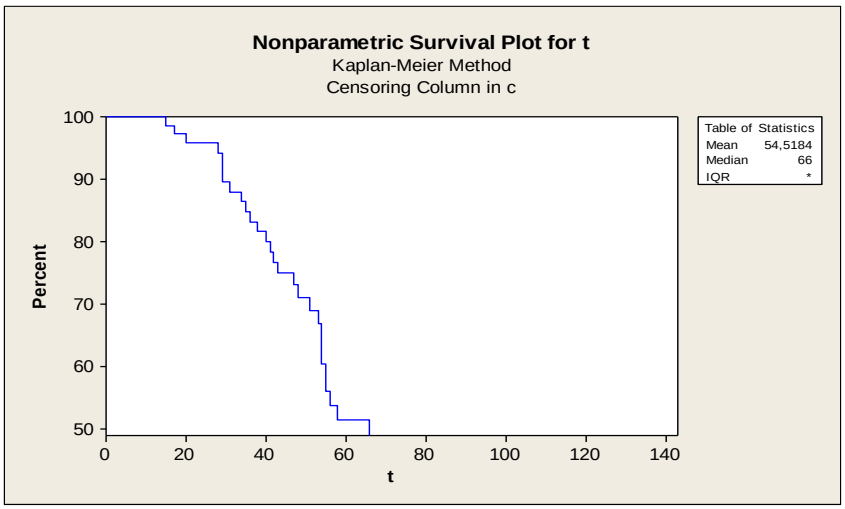
Χαρακτηριστικά ασθενών	N(%)		
Φύλο		Στάδιο	
Άντρες	55(73%)	T1(<=3 cm)	44(59%)
Γυναίκες	20(27%)	T2(>3 cm)	31(41%)
Καπνιστής		Υποστάδια	
		T1α(<2 cm)	24(32%)
Ναι	62(83%)		
		T1β(2-3 cm)	20(27%)
Οχι	13(17%)		
		T2α(3-5 cm)	19(25%)
		T2β(5-7 cm)	12(16%)
Είδος εκτομής		Βαθμός διαφοροποίησης	
Λοβεκτομή	59(79%)	Υψηλό	13(17%)
Τμηματοεκτομή	16(21%)	Μέσο	39(52%)
		Χαμηλό	23(31%)
Ιστολογικός τύπος			
Αδενοκαρκίνωμα	50(67%)		
Πλακώδες	25(33%)		

Πίνακας 6.4: Δημογραφικά και κλινικά χαρακτηριστικά ασθενών

6.2 Ανάλυση δεδομένων καρκίνου του πνεύμονα

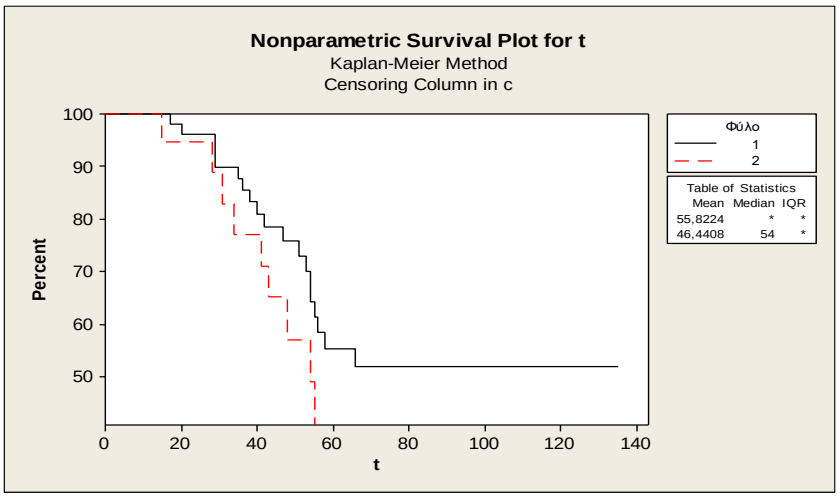
Ένα αρχικό βήμα για την ανάλυσης ενός συνόλου δεδομένων επιβίωσης είναι να παρουσιάσουμε κάποια γραφικά αποτελέσματα για του χρόνους επιβίωσης των ατόμων που βρίσκονται σε ένα γκρουπ με την βοήθεια των εκτιμήσεων της Kaplan-Meier συνάρτησης επιβίωσης. Τα αποτελέσματα αυτά προκύπτουν εύκολα από τις εκτιμήσεις των συναρτήσεων επιβίωσης. Στην συνέχεια θα συγκρίνουμε δύο διαφορετικές ομάδες ασθενών. Μπορούμε να πραγματοποιήσουμε μια άτυπη

σύγκριση της επιβίωσης για κάθε ομάδα χρησιμοποιώντας τις εκτιμήσεις των συναρτήσεων επιβίωσης με την βοήθεια του ελέγχου log-rank. Στο Σχήμα 6.1 παρουσιάζεται η καμπύλη Kaplan-Meier για το σύνολο των δεδομένων. Αυτό που παρατηρούμε είναι ότι τους πρώτους 15 μήνες αλλά και μετά τους 65 περίπου μήνες δεν προκύπτει κάποιος θάνατος.



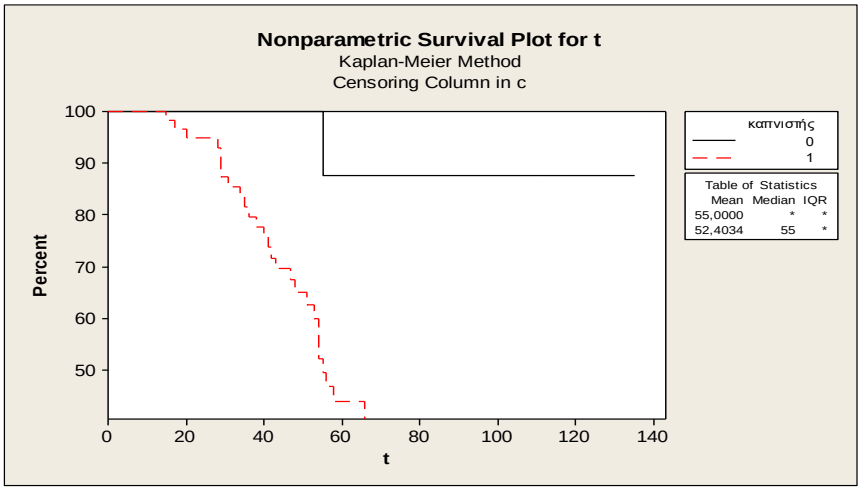
Σχήμα 6.1: Εκτίμηση της Kaplan-Meier της συνάρτησης επιβίωσης $S(t)$

Αρχικά ελέγχθηκε αν το φύλο επιδρά στην επιβίωση και οι καμπύλες Kaplan-Meier παρουσιάζονται από το διάγραμμα του Σχήματος 6.2. Από τον έλεγχο log-rank προέκυψε ότι δεν υπάρχει στατιστικά σημαντική διαφορά στην επιβίωση ανάμεσα σε άνδρες και γυναίκες. $(\mathrm{X}^2_{(1)} = 1.128,\ p = 0.288)$



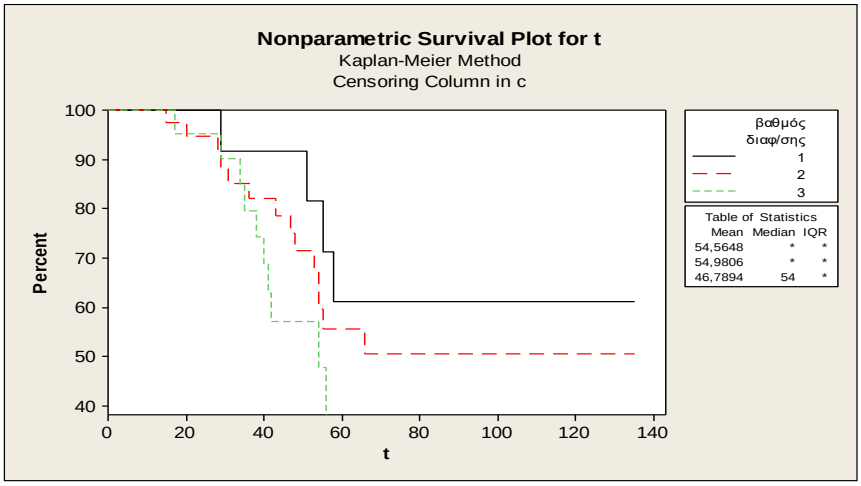
Σχήμα 6.2: Σύγκριση των εκτιμήσεων Kaplan-Meier μεταξύ ανδρών και γυναικών

Στην συνέχεια ελέγχθηκε αν το κάπνισμα επιδρά στην επιβίωση των ατόμων και οι καμπύλες Kaplan-Meier παρουσιάζονται στο διάγραμμα του Σχήματος 6.3. Από τον έλεγχο log-rank προέκυψε ότι υπάρχει διαφοροποίηση μεταξύ καπνιστών και μη καπνιστών $(X^2_{(1)} = 5.977, p = 0.014)$ καθώς και ότι η επιβίωση είναι καλύτερη για τους μη καπνιστές αφού το ποσοστό δεν πέφτει κάτω από το 85%.



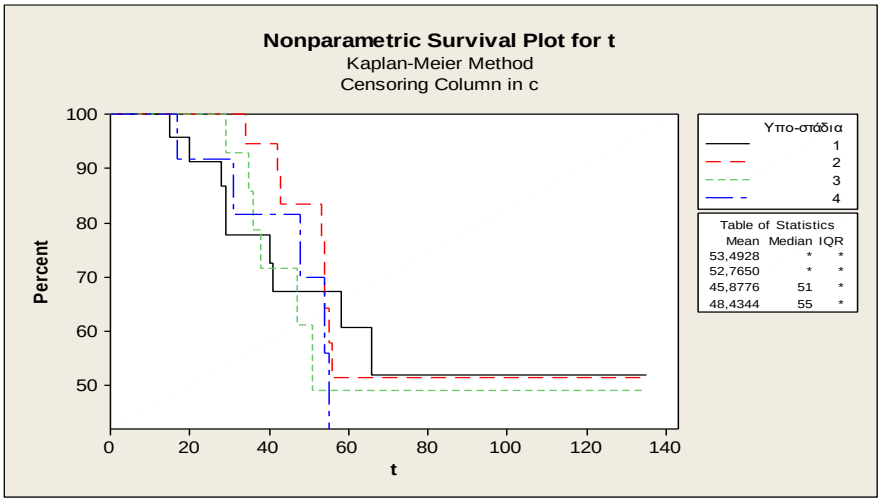
Σχήμα 6.3: Σύγκριση των εκτιμήσεων Kaplan-Meier μεταξύ καπνιστών και μη-καπνιστών

Επίσης ελέγχθηκε ο βαθμός διαφοροποίησης επιδρά στην επιβίωση των ατόμων και οι καμπύλες Kaplan-Meier παρουσιάζονται στο διάγραμμα του Σχήματος 6.4. Από τον έλεγχο log-rank προέκυψε ότι δεν υπάρχει διαφοροποίηση μεταξύ των τριών βαθμών διαφοροποίησης. $(X^2_{(2)} = 2.543, p = 0.280)$



Σχήμα 6.4: Σύγκριση των εκτιμήσεων Kaplan-Meier για τον βαθμό διαφοροποίησης

Επίσης ελέγχθηκε αν τα υποστάδια επιδρούν στην επιβίωση των ατόμων και οι καμπύλες Kaplan-Meier παρουσιάζονται στο διάγραμμα του Σχήματος 6.5. Από τον έλεγχο log-rank προέκυψε ότι δεν υπάρχει απολύτως καμία διαφοροποίηση μεταξύ των τεσσάρων υποσταδίων. $(X^2_{(3)} = 0.500, p = 0.919)$



Σχήμα 6.5: Σύγκριση των εκτιμήσεων Kaplan-Meier για την Υποστάδια

Εφαρμογή του μοντέλου αναλογικής διακινδύνευσης του Cox

Σκοπός μας είναι να προσαρμόσουμε τα δεδομένα μας με το μοντέλο του Cox, χρησιμοποιώντας το στατιστικό πακέτο survival της R, και να επιλέξουμε το στατιστικά καταλληλότερο μοντέλο για την περιγραφή του προβλήματος. Έχοντας ορίσει τις επεξηγηματικές και κατηγορικές μεταβλητές όπως και την μεταβλητή απόκρισης αλλά και τον χρόνο επιβίωσης, μπορούμε να προσαρμόσουμε αναλογικής διακινδύνευσης .

Πριν την προσαρμογή του μοντέλου, θα ελέγξουμε την συσχέτιση των μεταβλητών αριθμητικά. Όσο πιο κοντά στο απόλυτο 1 είναι το συντελεστής συσχέτισης, τόσο πιο έντονη θα είναι η συσχέτιση μεταξύ τους. Στην περίπτωση που είναι 0 τότε θα είναι ασυσχέτιστες.

	Φύλλο	καπνιστής	Είδος εκτομής	Στάδια	Υπο-στάδια	Ιστ/κός τύπος	Βαθμός διαφ/σης
Φύλο	1	-0.04248	-0.09322	-0.07755	0.02625	-0.18584	0.10347
καπνιστής	-0.04248	1	0.06649	0.24129	0.23964	0.11151	0.08980
Είδος εκτομής	-0.09322	0.06649	1	-0.17272	-0.18378	-0.03738	0.08936
Στάδια	-0.07755	0.24129	-0.17272	1	0.88774	0.12820	-0.08496
Υπο-στάδια	0.02625	0.23964	-0.18378	0.88774	1	0.06307	-0.06464
Ιστ/κός τύπος	-0.18584	0.11151	-0.03738	0.12820	0.06307	1	0.06318
Βαθμός διαφ/σης	0.10347	0.08980	0.08936	-0.08496	-0.06464	0.06318	1

Πίνακας 6.5: Συσχέτιση των συμμεταβλητών

Από τα αποτελέσματα του Πίνακα 6.5 φαίνεται ξεκάθαρα ότι υπάρχει μεγάλη συσχέτιση μεταξύ των μεταβλητών Στάδια και Υποστάδια καθώς η τιμή μεταξύ τους είναι 0,87743, πολύ κοντά στην μονάδα. Το φαινόμενο αυτό προκύπτει όταν η μια μεταβλητή προβλέπεται σε μεγάλο βαθμό από την άλλη, ώστε κατ' ουσίαν να παρέχουν τις ίδιες πληροφορίες και ο διαχωρισμός των επιδράσεων τους να είναι δύσκολος. Θα επιλέξουμε λοιπόν να αφαιρέσουμε την μεταβλητή Στάδια.

Στην συνέχεια προσαρμόζουμε το μοντέλο του Cox χρησιμοποιώντας το στατιστικό πακέτο R, με την συνάρτηση επιβίωσης να εξαρτάται από τον χρόνο επιβίωσης και την κατάσταση του ασθενούς μετά το χειρουργείο σαν κατηγορική μεταβλητή.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
ΗΛΙΚΙΑ	-0.02503	0.97528	0.02494	-1.004	0.31554
ΦΥΛΟ	0.46096	1.58559	0.52022	0.886	0.37558
ΚΑΠΝΙΣΤΗΣ	3.22095	25.05184	1.11670	2.884	0.00392 **
ΕΙΔΟΣ_ΕΚΤΟΜΗΣ	0.21995	1.24602	0.50320	0.437	0.66203
ΛΕΜΦΑΔΕΝΕΣ	0.14354	1.15435	0.09014	1.592	0.11131
ΔΙΑΜΕΤΡΟΣ	-0.97058	0.37886	0.46378	-2.093	0.03637 *
(ΥΠΟΣΤΑΔΙΑf)2	0.07240	1.07508	0.72042	0.100	0.91995
(ΥΠΟΣΤΑΔΙΑf)3	2.16162	8.68522	1.39211	1.553	0.12048

(ΥΠΟΣΤΑΔΙΑf)4	3.85492	47.22491	2.23058	1.728	0.08395 .
ΙΣΤΟΛΟΓΙΚΟΣ_ΤΥΠΟΣ	-0.88205	0.41393	0.54428	-1.621	0.10511
(ΒΑΘΜΟΣΔΙΑΦΟΡf)2	0.52312	1.68728	0.69739	0.750	0.45319
(ΒΑΘΜΟΣΔΙΑΦΟΡf)3	0.87767	2.40530	0.68632	1.279	1.279
Likelihood ratio test=25.4 on 12 df, p=0.0131 n= 75, number of events					AIC=208.89

Αποτελέσματα 6.1: Από την προσαρμογή του μοντέλου Cox

Παρατηρώντας τα Αποτελέσματα 6.1, του προσαρμοσμένου μοντέλου με όλες τις μεταβλητές, εκτός από την μεταβλητή στάδια που όπως προαναφέραμε δεν θα την βάλουμε στο μοντέλο, βλέπουμε από τις p-τιμές των ελέγχων Wald ότι η πιο σημαντική μεταβλητή είναι η μεταβλητή ΚΑΠΝΙΣΤΗΣ με p-τιμή 0,00392 και ακολουθούν οι λιγότερο σημαντικές ΔΙΑΜΕΤΡΟΣ και ΥΠΟΣΤΑΔΙΑ.

Για να βελτιώσουμε το μοντέλο μας θα χρησιμοποιήσουμε την διαδικασία διαδοχικής αφαίρεσης η οποία ξεκινάει εισάγοντας όλες τις μεταβλητές του μοντέλου και αφαιρεί μια μια τις μεταβλητές ξεκινώντας από την λιγότερη σημαντική και με βάση τον έλεγχο Wald, καταλήγουμε στο τελικό μοντέλο παίρνοντας τα Αποτελέσματα 6.2.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
ΚΑΠΝΙΣΤΗΣ	2.2963	9.9378	1.0269	2.236	0.025*
ΛΕΜΦΑΔΕΝΕΣ	0.1117	1.1182	0.0718	1.555	0.120
ΙΣΤΟΛΟΓΙΚΟΣ ΤΥΠΟΣ	-0.9179	0.3993	0.0478	-1.917	0.055 .
Likelihood (max possible=0.939)			0.003		
Wald test			0.031		
Score test			0.017		

Αποτελέσματα 6.2: για το μοντέλο Cox μετά την διαδικασία διαδοχικής αφαίρεσης

Από τα p-values των Αποτελεσμάτων 6.2 φαίνεται η μεταβλητή καπνιστής είναι η πιο σημαντική και ακολουθούν οι λεμφαδένες και ο ιστολογικός τύπος. Εφαρμόζοντας και το κριτίριο AIC βλέπουμε ότι το μοντέλο βελτιώθηκε (AIC=202.75). Για τη σύγκριση μεταξύ των συμμεταβλητών και πώς αυτές δρουν πολλαπλασιαστικά πάνω στον κίνδυνο για θάνατο, θα χρησιμοποιήσουμε τις τιμές exp(coef) οι

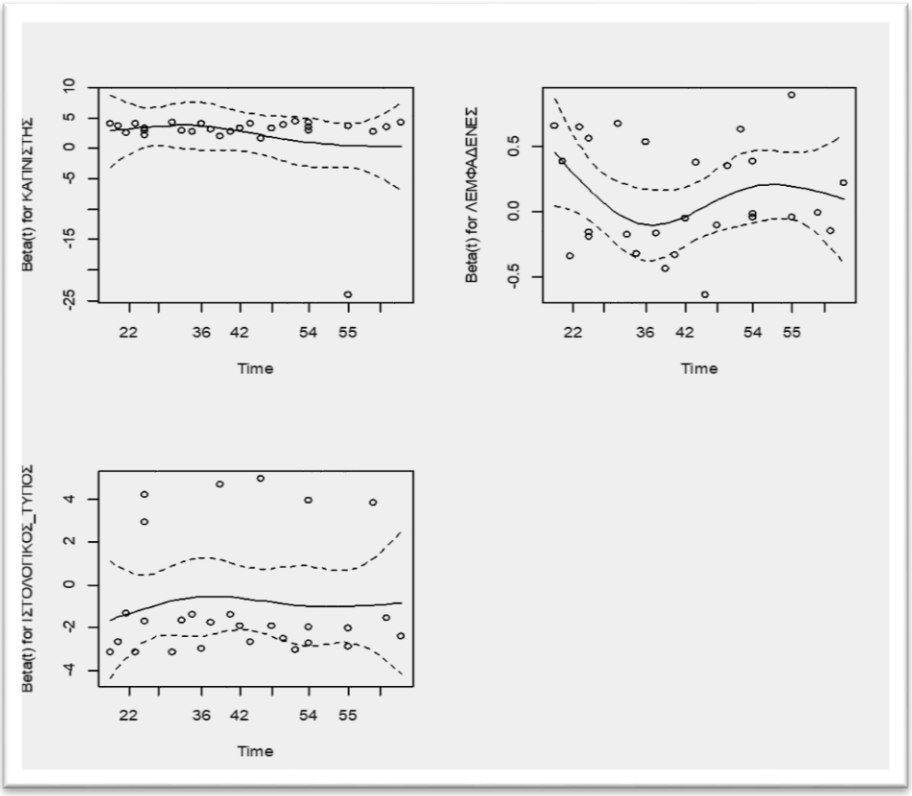
οποίες δείχνουν κατά πόσο πολλαπλασιάζεται η συνάρτηση διακινδύνευσης $h(t;x) = ho(t) * e^{\beta' x}$, δηλαδή κατά πόσο μια συμμεταβλητή επιδρά στην διάρκεια ζωής, όταν οι άλλες συμμεταβλητές του τελικού μοντέλου θεωρούνται σταθερές. Εδώ έχουμε, για έναν ασθενή που είναι καπνιστής αυξάνεται κατά 9.93 σε σχέση με έναν που δεν είναι $(h(t;x) = ho(t) * 9.93)$. Κάθε επιπλέον λεμφαδένας αυξάνει την συνάρτηση διακινδύνευσης κατά 1,11 $(h(t;x) = ho(t) * 1{,}11)$ για έναν ασθενή που έχει αδενοκαρκίνωμα μειώνεται κατά 0,39 σε σχέση με έναν που έχει πλακώδες $(h(t;x) = ho(t) * 0.39)$.

Το ημιπαραμετρικό μοντέλο διακινδύνευσης του Cox δίνεται από την σχέση $h(t;x) = ho(t) * e^{\beta' x}$ με το β να είναι ένα διάνυσμα το οποίο εκφράζει την επίδραση της καθεμιάς των συμμεταβλητών **x**. Η μηδενική υπόθεση λέει ότι όλα τα $\beta_j = 0$ δηλαδή ότι υπάρχει ανεξαρτησία της διάρκειας ζωής από την συμμεταβλητή x_j. Ο έλεγχος του λόγου των πιθανοφανειών φαίνεται να είναι πιο ξεκάθαρος ως προς την απόρριψη τη μηδενικής υπόθεσης σε σχέση με τους άλλους δύο που έχουν μεγαλύτερες p-τιμές.

Θα εξετάσουμε τώρα την υπόθεση της αναλογικότητας στο μοντέλο του Cox. Στον παρακάτω πίνακα βλέπουμε τα Αποτελέσματα 6.3 που πήραμε για τον έλεγχο αυτό. Από την στήλη με τις p-τιμές μπορούμε να πούμε ότι δεχόμαστε την υπόθεση της αναλογικότητας για όλες τις συμμεταβλητές, αλλά και για ολόκληρο το μοντέλο (GLOBAL).

	Rho	Chisq	P
ΚΑΠΝΙΣΤΗΣ	-0.2200	1.2133	0.271
ΛΕΜΦΑΔΕΝΕΣ	0.00819	0.0575	0.811
ΙΣΤΟΛΟΓΙΚΟΣ ΤΥΠΟΣ	0.0330	0.0342	0.853
GLOBAL	NA	1.2321	0.745

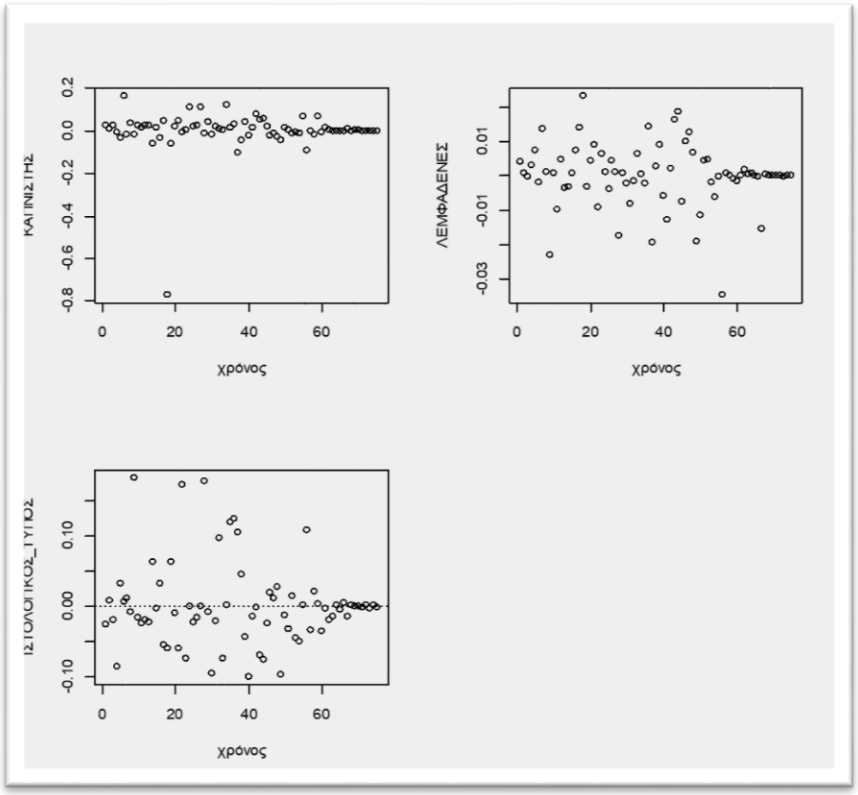
Αποτελέσματα 6.3: Για την υπόθεση αναλογικότητας



Σχήμα 6.6: Schoenfeld για τις συμμεταβλητές του μοντέλου

Στο Σχήμα 6.6 εμφανίζονται τα υπόλοιπα Schoenfeld συναρτήσει του χρόνου για τις συμμεταβλητές του μοντέλου. Η κλίμακα του χρόνου στα γραφήματα προέκυψε με βάση την προκαθορισμένη επιλογή της R, δηλαδή την Kaplan-Meier. Τα παραπάνω σχήματα αποτελούν ουσιαστικά μια οπτική επιβεβαίωση των παραπάνω ελέγχων για την υπόθεση της αναλογικής διακινδύνευσης. Τα γραφήματα θα πρέπει να αποτελούνται από ευθεία γραμμή ή μια εξομαλυμένη καμπύλη, κάτι που ισχύει για όλες τις συμμεταβλητές μας. Συνεπώς υπάρχει συμφωνία των αριθμητικών αποτελεσμάτων και των γραφικών παραστάσεων. Οι διακεκομμένες γραμμές συμβολίζουν $\pm$ 2standard-error γύρω από την προσαρμογή.

Για την αξιολόγηση της επάρκειας του μοντέλου είναι σημαντικό να ελέγξουμε αν κάποια από τις παρατηρήσεις μας έχει μεγάλη επίδραση στο μοντέλο που προσαρμόστηκε για το σύνολο των δεδομένων. Είναι λοιπόν σημαντικό να εξετάσουμε ποιες από τις παρατηρήσεις μας είναι σημεία επιρροής.

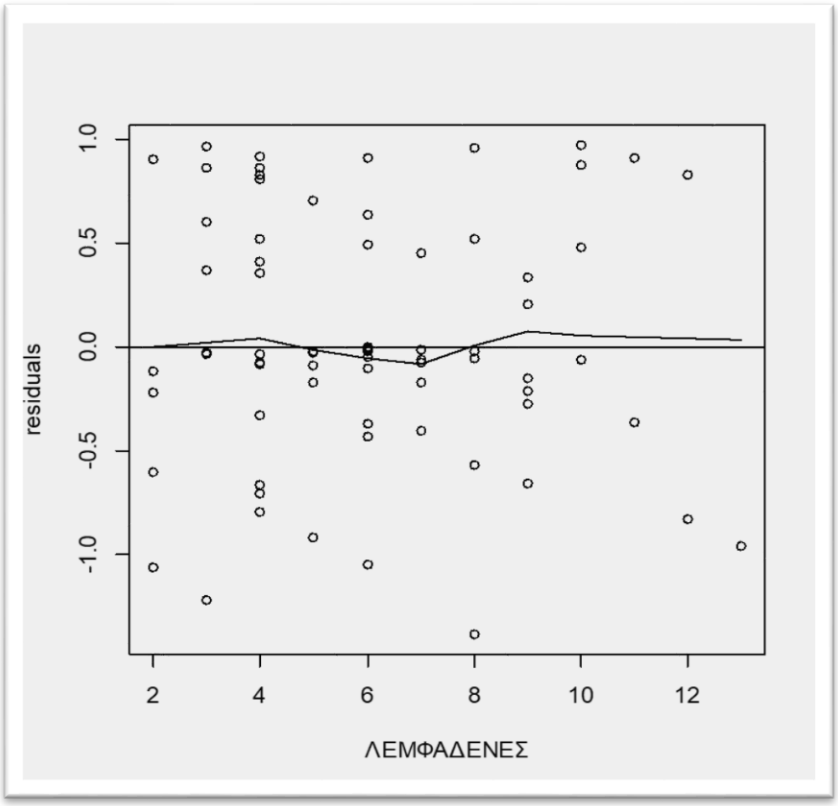


Σχήμα 6.7: Υπόλοιπα df-beta σε σχέση με τον χρόνο

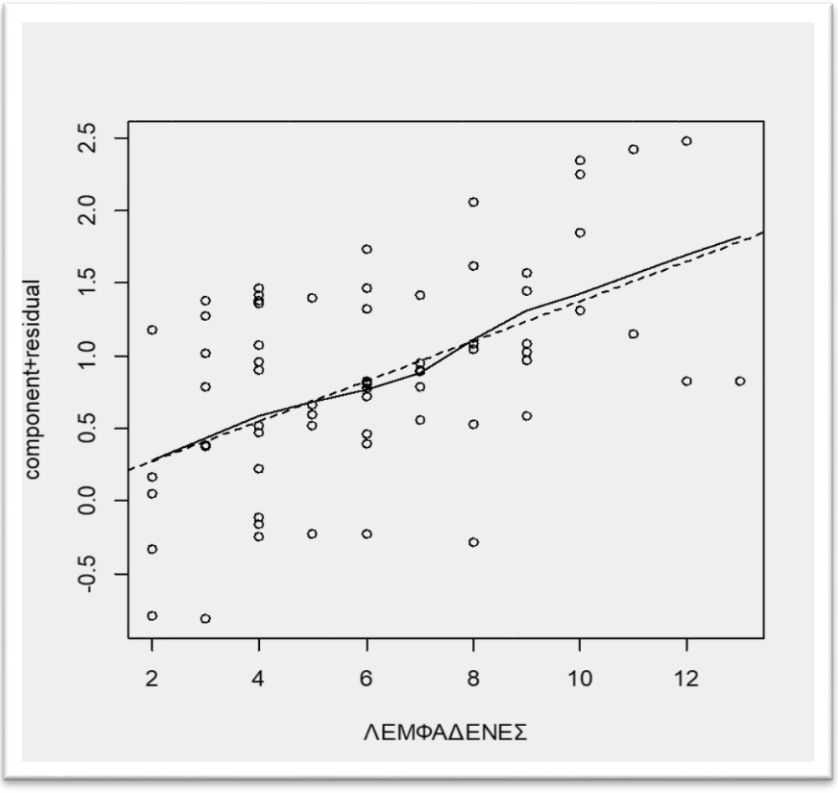
Τιμές μεγαλύτερες του $2/\sqrt{n}$ =0,23 κατ' απόλυτη τιμή θεωρούνται ότι είναι σημεία επιρροής. Εδώ παρατηρούμε για την μεταβλητή ΚΑΠΝΙΣΤΗΣ υπάρχει ένα σημείο που ξεφεύγει από τα όρια και είναι η 18η παρατήρηση. Ο λόγος για τον οποίο το σημείο αυτό ξεφεύγει είναι ότι ο συγκεκριμένος ασθενής πεθαίνει πιο γρήγορα (στους 55 μήνες) σε σχέση με τους άλλους μη καπνιστές.

Η μη γραμμικότητα αποτελεί ένα πιθανό πρόβλημα στο μοντέλο αναλογικής δια-κινδύνευσης του Cox όπως συμβαίνει και γενικευμένα γραμμικά μοντέλα. Τα υπόλοιπα martingale μπορούν να χρησιμοποιηθούν σε αυτήν την περίπτωση για να ανιχνεύσουμε την μη-γραμμικότητα. Στις διχότομες μεταβλητές δεν έχουμε θέμα μη γραμμικότητας, οπότε θα ελέγξουμε μόνο σε σχέση με την συμμεταβλητή ΛΕΜΦΑΔΕΝΕΣ και όπως βλέπουμε στο Σχήμα 6.8 δείχνει ομαλή.

Τα Martingale residual μπορούν επιπλέον να χρησιμοποιηθούν για να σχηματίσουν τα component+residual ή partial residuals όπου και εδώ τα θα προκύψουν γραφήματα για να ελέγξουμε την γραμμικότητα.



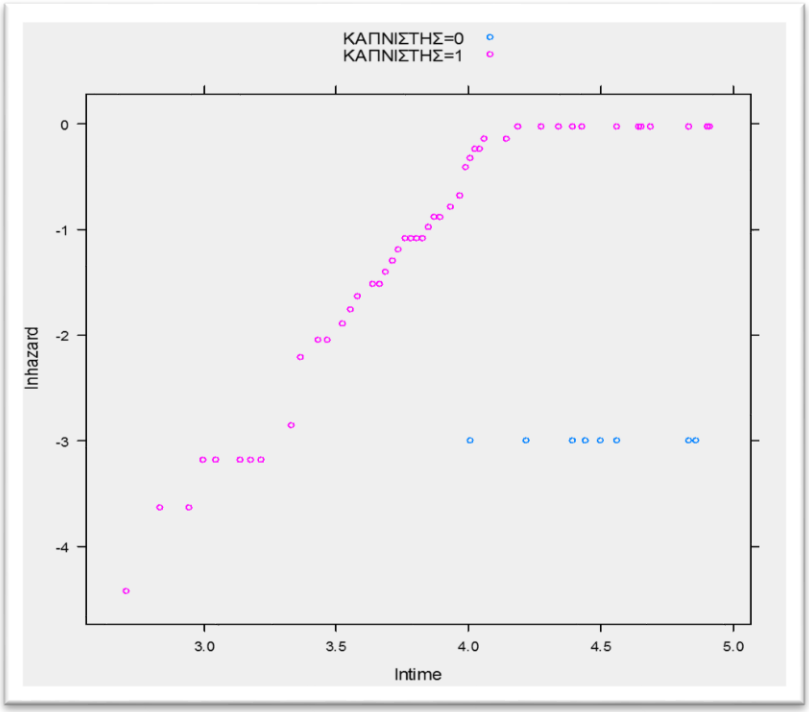
Σχήμα 6.8: Martingale residual γράφημα για την συμμεταβλητή ΛΕΜΦΑΔΕΝΕΣ



Σχήμα 6.9: Component+residual γράφημα για την συμμεταβλητή ΛΕΜΦΑΔΕΝΕΣ

Η γραμικότητα στο Σχήμα 6.9 φαίνεται να είναι ικανοποιητική αφού δείχνει κάπως ομαλή. Η ευθεία προέκυψε από γραμμική παλινδρόμηση με την βοήθεια της συνάρτησης lowess. Η διακεκομμένη ευθεία προσαρμόστηκε με γραμμικά ελάχιστα τετράγωνα.

Μια σημαντική επέκταση του βασικού μοντέλου του Cox είναι η δυνατότητα εφαρμογής μιας στρωματοποιημένης ανάλυσης (stratified Cox model). Αυτή συνήθως πραγματοποιείται όταν εικάζεται ότι οι συναρτήσεις διακινδύνευσης μεταξύ δύο ή περισσοτέρων κατηγοριών δεν βρίσκονται σε αναλογία μεταξύ τους. Στην περίπτωση μας χωρίζουμε την μεταβλητή ΚΑΠΝΙΣΤΗΣ σε δύο στρώματα ,τους μη-καπνιστές και τους καπνιστές. Οπότε η συνάρτηση διακινδύνευσης ορίζεται ως $h_j(t;x) = h_{0j}(t)\exp(\beta' x)$, όπου $j = 0,1$ δηλώνει το στρώμα του παράγοντα.



Σχήμα 6.10: Στρωματοποίηση για την συμμεταβλητή καπνιστής

Από το Σχήμα 6.10 φαίνεται να μην υπάρχει αναλογία ανάμεσα στις δύο κατηγορίες κάτι που μας οδηγεί στο να κάνουμε την στρωματοποίηση ως προς την συμμεταβλητή καπνιστής. Άρα θα ξαναπροσαρμόσουμε το μοντέλο θεωρώντας την συμμεταβλητή καπνιστής ως στρωματοποιημένη μεταβλητή και από τα Αποτελέσματα 6.4 παίρνουμε:

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
ΔΙΑΜΕΤΡΟΣ	-0.883	0.414	0.4593	0.399	0.027
ΥΠΟΣΤΑΔΙΑf2	0.517	1.678	0.677	0.764	0.445
ΥΠΟΣΤΑΔΙΑf3	2.103	8.189	1.197	1.76	0.079
ΥΠΟΣΤΑΔΙΑf4	4.212	67.463	2.053	2.05	0.040
Likelihood ratio test				0.209	
Wald test				0.201	
Score test				0.179	

Αποτελέσματα 6.4: R για το στρωματοποιημένο μοντέλο

Τα άτομα που ανήκουν στο ίδιο στρώμα, έχουν τις ίδιες αναφορικές συναρτήσεις κινδύνου, ενώ αντίθετα τα άτομα που ανήκουν σε διαφορετικά στρώματα έχουν διαφορετικές αναφορικές συναρτήσεις κινδύνου δηλαδή:

$$h_0(t;x) = h_{00}(t)\exp(\beta_1 \Delta\text{ΙΑΜΕΤΡΟΣ} + \beta_2 \text{ΥΠΟΣΤΑΔΙΑ}f)$$

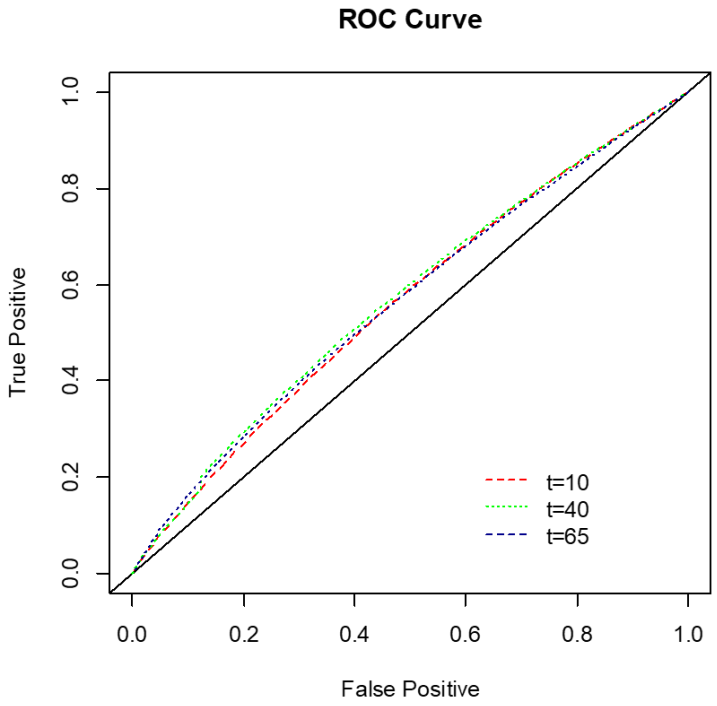
$$h_1(t;x) = h_{01}(t)\exp(\beta_1 \Delta\text{ΙΑΜΕΤΡΟΣ} + \beta_2 \text{ΥΠΟΣΤΑΔΙΑ}f)$$

Επίσης, τα άτομα που ανήκουν στο ίδιο στρώμα έχουν συναρτήσεις κινδύνου ανάλογες μεταξύ τους, αφού για παράδειγμα για δύο άτομα με μεταβλητές x_1 και x_2, που ανήκουν στο στρώμα $i=0$,

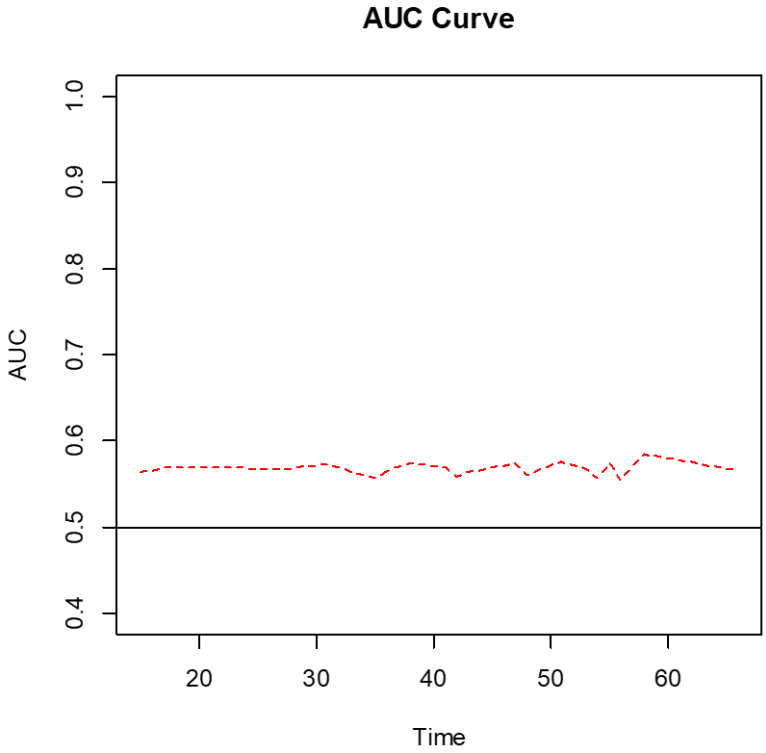
$$\frac{h_0(t|x_1)}{h_0(t|x_2)} = \frac{h_{00}(t)e^{\beta' x_1}}{h_{00}(t)e^{\beta' x_2}} = e^{\beta'(x_1 - x_2)}$$

Επιπλέον, από το στρωματοποιημένο μοντέλο φαίνεται ακόμη ότι οι συντελεστές παλινδρόμησης β είναι οι ίδιοι σε κάθε στρώμα.

Στην συνέχεια αφού κάναμε και την στρωματοποίηση και καταλήξαμε στο τελικό μοντέλο, θα εφαρμόσουμε την καμπύλη ROC (Receiver Operating Characteristic) για να εξετάσουμε την προβλεπτική του ικανότητα. Ο άξονας True Positive είναι η ευαισθησία (sensitivity) και ο άξονας False Positive είναι 1-ειδικότητα. Επιθυμητό είναι να έχουμε υψηλή ευαισθησία και χαμηλή 1-ειδικότητα. Επιπλέον αυτό που μας ενδιαφέρει είναι και το εμβαδόν (AUC=Area Under Curve) που σχηματίζεται κάτω από κάθε καμπύλη για τις χρονικές στιγμές που επιλέξαμε. Όσο μεγαλύτερη είναι η τιμή του τόσο καλύτερη είναι η πρόβλεψη. Για την εφαρμογή των καμπυλών ROC χρησιμοποιήσαμε το πακέτο risketROC της R.



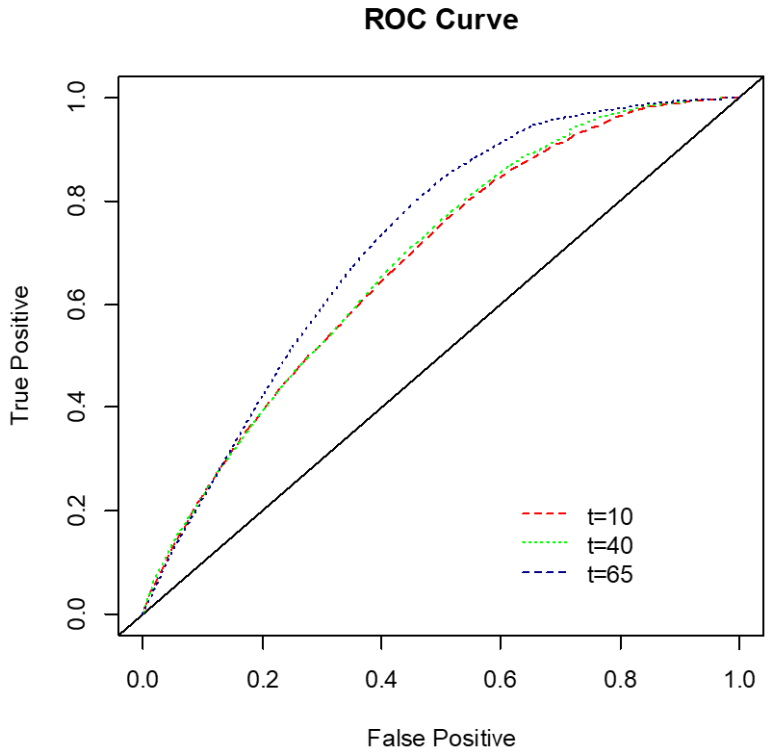
Σχήμα 6.11: Καμπύλες ROC για το στρωματοποιημένο μοντέλο για τις χρονικές στιγμές t=10, t=40, t=65



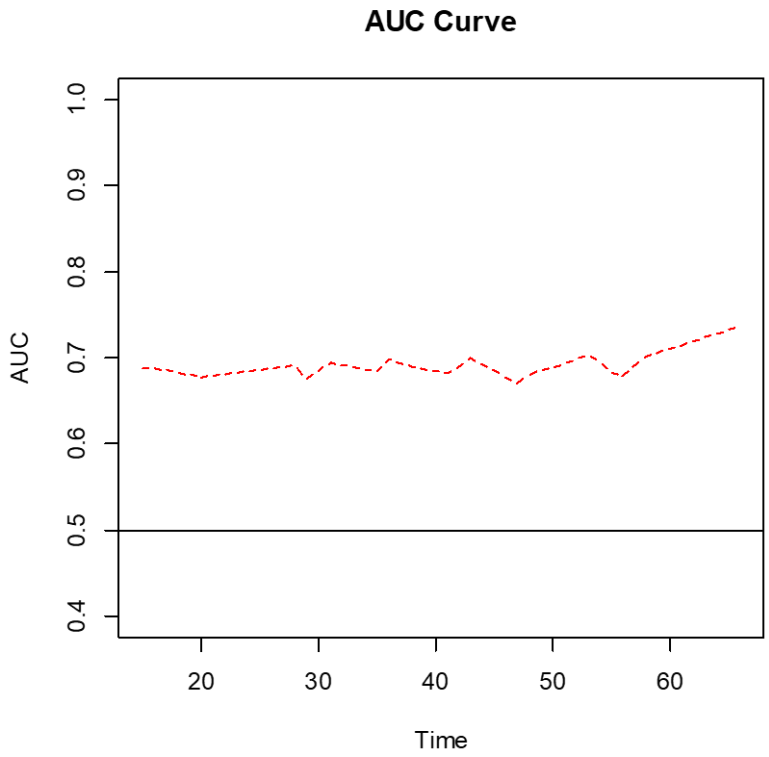
Σχήμα 6.12: Καμπύλη AUC για το στρωματοποιημένο μοντέλο

Από το διάγραμμα του Σχήματος 6.11 παρατηρούμε ότι για 20% false-positive rate οδηγεί σε ευαισθησία του γεγονότος στο 25% στους 10 μήνες και παραμένει σχεδόν το ίδιο ποσοστό στους 40 μήνες. Επίσης οι καμπύλες ROC μας παρέχουν πληροφορία για την σχέση μεταξύ ευαισθησίας και ειδικότητας. Από το διάγραμμα του Σχήματος 6.12 παίρνουμε τις τιμές των εμβαδών για τις χρονικές στιγμές που εξετάσαμε είναι AUC(t=10)=0.562, AUC(t=40)=0.571, AUC(t=65)=0.565. Παρατηρούμε τόσο από τις καμπύλες, όσο και από τα αντίστοιχα εμβαδά ότι η προβλεπτική ικανότητα του μοντέλου δεν είναι ιδιαίτερα καλή αφού και οι τρεις καμπύλες για τις χρονικές στιγμές που επιλέξαμε, βρίσκονται πολύ κοντά στην διαγώνιο.

Θα εφαρμόσουμε τώρα τις καμπύλες ROC και AUC για το μοντέλο που προκύπτει από τα Αποτελέσματα 6.2, δηλαδή σε αυτό που καταλήξαμε μετά την διαδικασία διαδοχικής αφαίρεσης.



Σχήμα 6.13: Καμπύλες ROC για το μοντέλο μετά την διαδικασία διαδοχικής αφαίρεσης για τις χρονικές στιγμές t=10, t=40, t=65



Σχήμα 6.14: Καμπύλη AUC για το μοντέλο μετά την διαδικασία διαδοχικής αφαίρεσης

Από το διάγραμμα του Σχήματος 6.13 παρατηρούμε ότι για 20% false-positive rate οδηγεί σε ευαισθησία του γεγονότος στο 35% στους 10 μήνες και αυξάνεται ελαφρώς στο 40% στους 65 μήνες. Από το διάγραμμα του Σχήματος 6.14 παίρνουμε τις τιμές των εμβαδών για τις χρονικές στιγμές που εξετάσαμε είναι AUC(t=10)=0.678, AUC(t=40)=0.5683, AUC(t=65)=0.718. Παρατηρούμε τόσο από τις καμπύλες, όσο και από τα αντίστοιχα εμβαδά ότι η προβλεπτική ικανότητα του μοντέλου είναι σαφώς καλύτερη σε σχέση με αυτή του στρωματοποιημένου μοντέλου.

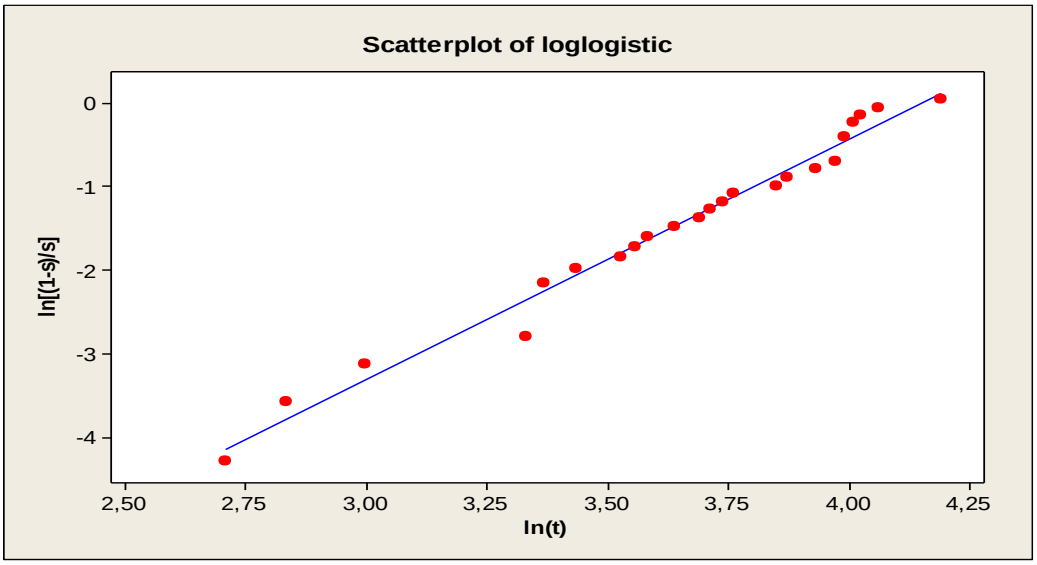
Εφαρμογή του παραμετρικού μοντέλου Weibull

Σκοπός μας είναι να προσαρμόσουμε ένα παραμετρικό μοντέλο παλινδρόμησης για τα δεδομένα μας. Με αυτόν τον τρόπο μπορούμε να δούμε αν ένας ή περισσότεροι παράγοντες επηρεάζουν τον χρόνο ζωής τους ασθενούς. Όταν προσαρμόσαμε το

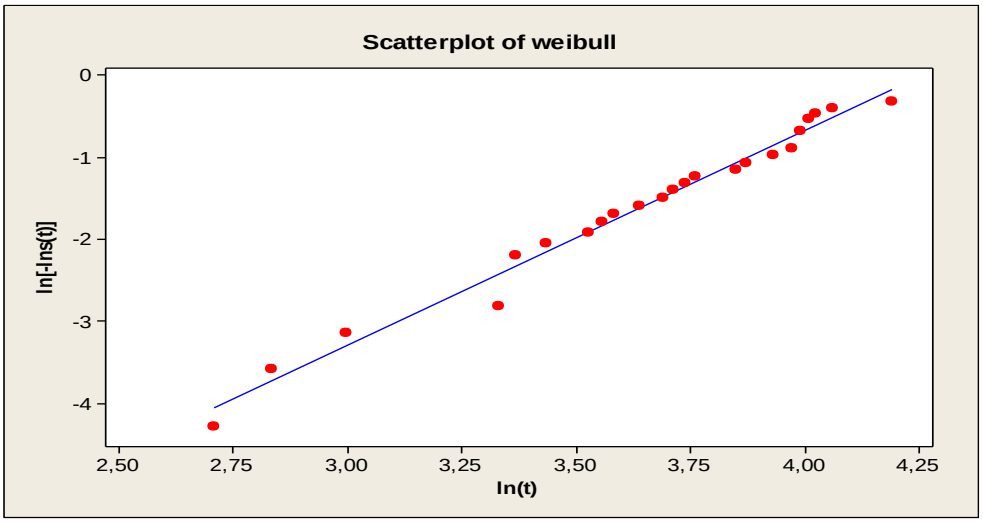
μοντέλο του Cox για την ανάλυση των δεδομένων μας δεν χρειαζόταν να υποθέσουμε κάποια συγκεκριμένη κατανομή για τους χρόνους επιβίωσης. Αυτό έχει ως αποτέλεσμα η συνάρτηση διακινδύνευσης να μην έχει κάποια συγκεκριμένη μορφή και το μοντέλο έχει μεγαλύτερη ευελιξία. Όμως όταν υποθέτουμε ότι τα δεδομένα μας ακολουθούν κάποια κατανομή τότε τα συμπεράσματα είναι πιο ακριβή. Αρχικά θα εκτελέσουμε κάποιους γραφικούς ελέγχους για να δούμε ποια κατανομή περιγράφει καλύτερα τα δεδομένα μας και στην συνέχεια θα προσαρμόσουμε το παραμετρικό μοντέλο της Weibull αφού αυτό ενδείκνυται πιο πολύ για δεδομένα διάρκειας ζωής.

Γραφικοί έλεγχοι κατανομής χρόνου

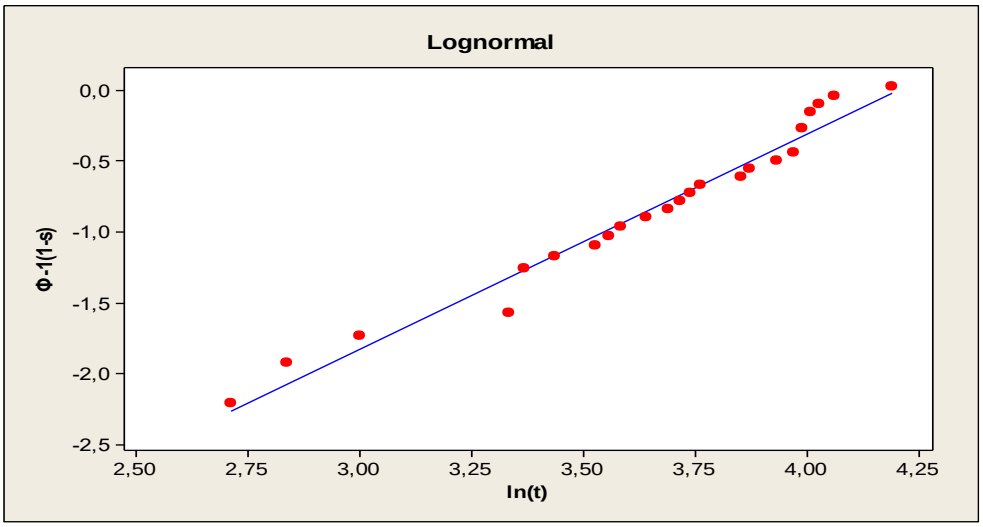
Θα ελέγξουμε με την βοήθεια της εκτιμήτριας Kaplan-Meier αν κάποιο παραμετρικό μοντέλο περιγράφει επαρκώς τα δεδομένα μας. Σαν πρώτο βήμα θα βρούμε μέσω των γραφικών παραστάσεων την κατανομή που ακολουθούν οι χρόνοι μέχρις ότου συμβεί το γεγονός.



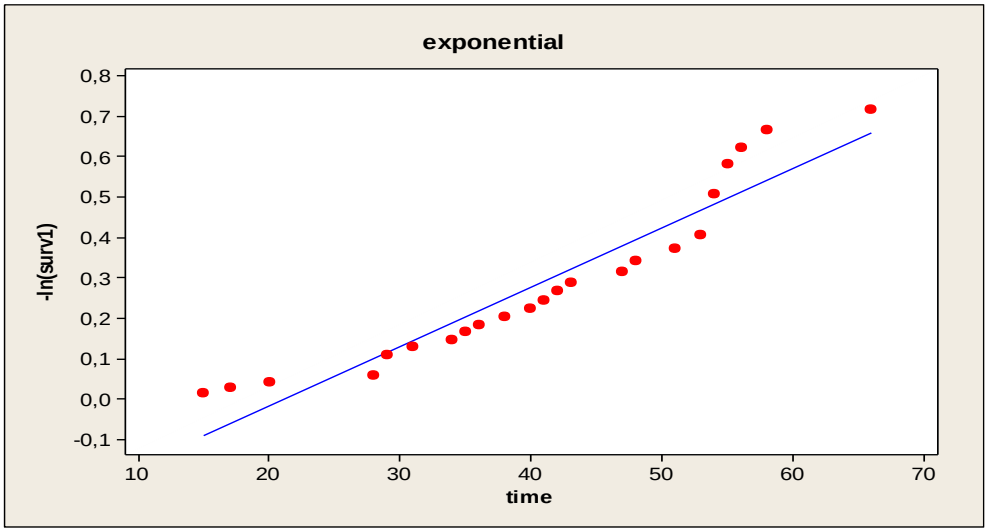
Σχήμα 6.15: Έλεγχος της Λογαριθμο-λογιστικής κατανομής



Σχήμα 6.16: Έλεγχος της Weibull κατανομής



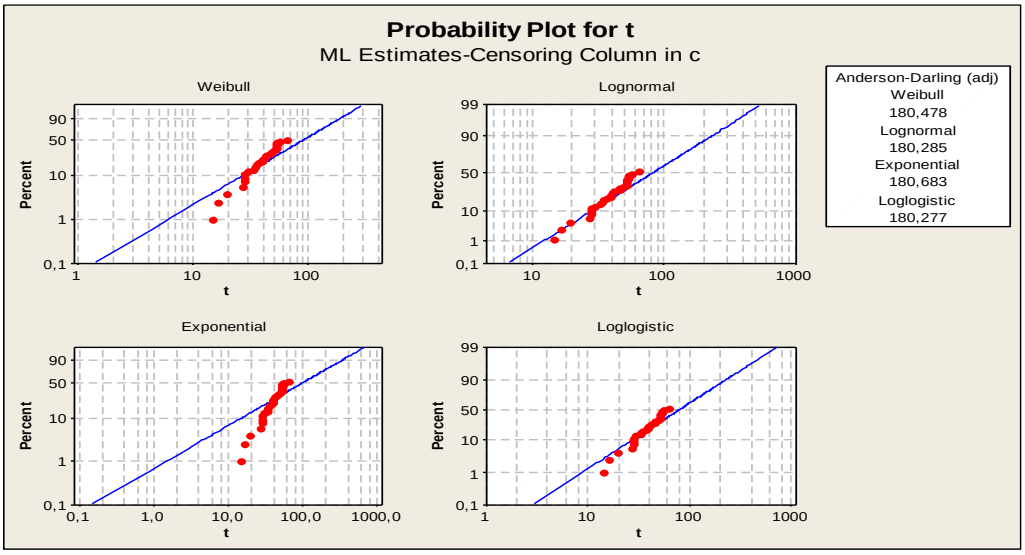
Σχήμα 6.17: Έλεγχος της Λογαριθμο-κανονικής κατανομής



Σχήμα 6.18: Έλεγχος της εκθετικής κατανομής

Όσο πιο κοντά πέφτουν τα σημεία στην προσαρμοσμένη ευθεία τόσο καλύτερη η προσαρμογή. Παρατηρούμε ότι η καλύτερη περιγραφή των δεδομένων είναι με την Λογαριθμολογιστική Σχήμα 6.15 και την Weibull Σχήμα 6.16. Οι δύο αυτές κατανομές είναι πολύ κοντά μεταξύ τους με την Weibull να υπερέχει με πολύ μικρή διαφορά. Εν συνεχεία θα εξετάσουμε τα Cox-Snell υπόλοιπα και τα Standardized υπόλοιπα.

Επιπλέον θα χρησιμοποιήσουμε ID plot, δηλαδή ένα γράφημα πιθανότητας για να ελέγξουμε και εδώ ποια κατανομή ταιριάζει καλύτερα. Το γράφημα αυτό μας παρέχει και έναν έλεγχο καλής προσαρμογής Anderson-Darling ο οποίος μας μετράει πόσο μακριά πέφτουν τα σημεία από την προσαρμοσμένη ευθεία. Μικρή τιμή του στατιστικού δηλώνει ότι η κατανομή περιγράφει καλύτερα τα δεδομένα.



Σχήμα 6.19: Distribution ID για τις Weibull, Lognormal, Exponential, Loglogistic

Στο Σχήμα 6.19 παρατηρούμε ότι η καλύτερη κατανομή είναι η Log-logistic με πολύ μικρή διαφορά από τις Lognormal και Weibull κάτι που επιβεβαιώνεται και από τον έλεγχο Anderson-Darling. Ωστόσο είναι πολύ μικρές οι διαφορές μεταξύ τους για να είμαστε εντελώς βέβαιοι.

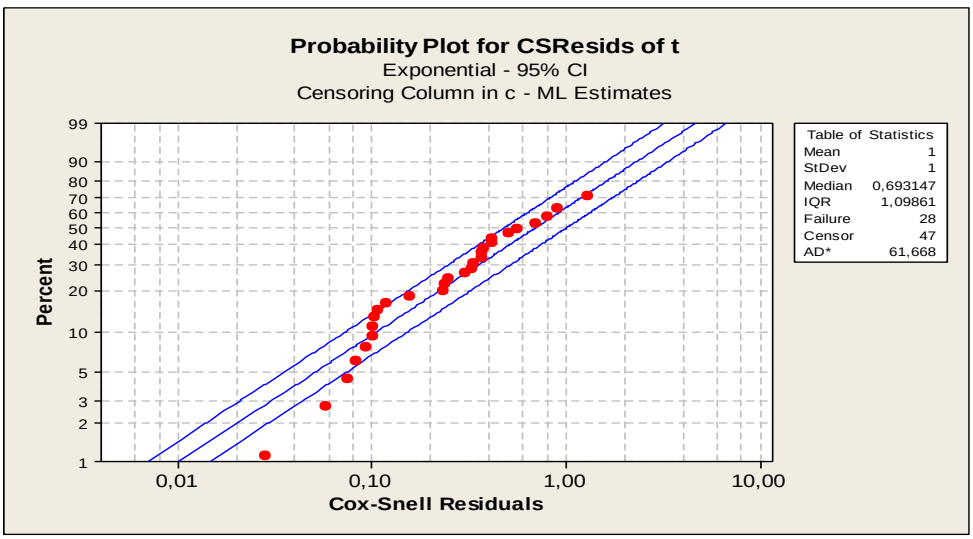
Προσαρμόζουμε τώρα το μοντέλο παλινδρόμησης για την κατανομή Weibull με όλες τις συμμεταβλητές εκτός της συμμεταβλητής ΣΤΑΔΙΑ λόγω υψηλής συσχέτισης της με την ΥΠΟΣΤΑΔΙΑ, όπως προαναφέραμε, και θα πάρουμε τα Αποτελέσματα 6.5:

	Value	Std. Error	z	Pr(>|z|)
ΗΛΙΚΙΑ	0.0148	1.0102	1.448	0.148
ΦΥΛΟ	-0.1825	0.2165	-0.543	0.399
ΕΙΔΟΣ_ΕΚΤΟΜΗΣ	-0.1818	0.2074	-0.877	0.381
ΚΑΠΝΙΣΤΗΣ	-1.7254	0.4900	-3.521	<0.001
ΛΕΜΦΑΔΕΝΕΣ	-0.0830	0.0367	-2.259	0.023
ΔΙΑΜΕΤΡΟΣ	0.5480	0.1953	2.806	0.005
ΥΠΟΣΤΑΔΙΑf2	-0.0923	0.3021	-0.306	0.760
ΥΠΟΣΤΑΔΙΑf3	-1.1367	0.5856	-1.941	0.052
ΥΠΟΣΤΑΔΙΑf4	-2.0700	0.9460	-2.188	0.028
ΙΣΤΟΛΟΓΙΚΟΣ_ΤΥΠΟΣ	0.3763	0.2275	1.654	0.098
ΒΑΘΜΟΣΔΙΑΦΟΡf2	-0.2688	0.2968	-0.906	0.365
ΒΑΘΜΟΣΔΙΑΦΟΡf3	-0.4267	0.2931	-1.456	0.145
Scale= 0.424				
Weibull distribution				
Loglik(model)= -145.6 Loglik(intercept only)= -163.8				
Chisq= 36.33 on 12 degrees of freedom, p= 0.00029				
AIC=319.177				

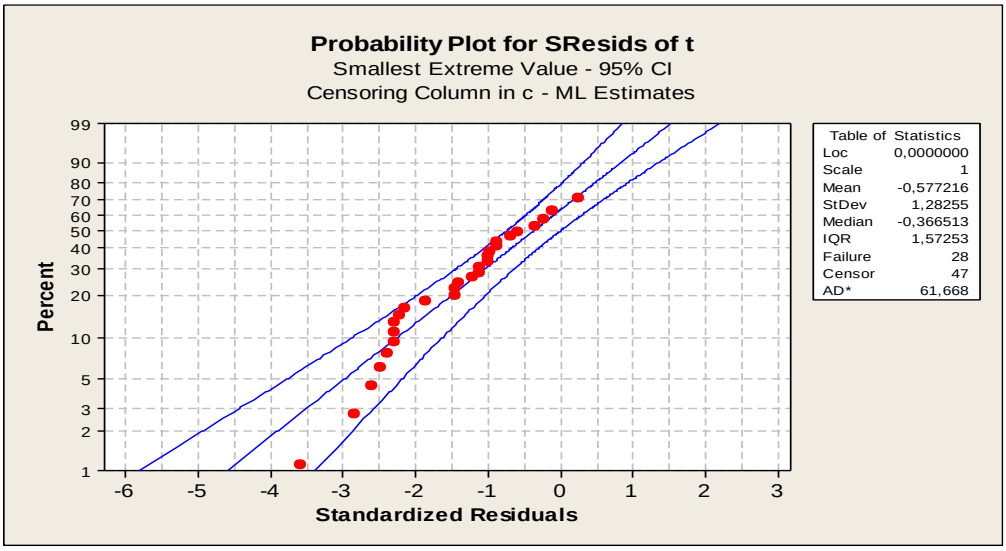
Αποτελέσματα 6.5 για το μοντέλο Weibull

Από τις p-τιμές των Αποτελεσμάτων 6.5 παρατηρούμε ότι οι σημαντικότερες συμ-μεταβλητές είναι οι ΚΑΠΝΙΣΤΗΣ, ΛΕΜΦΑΔΕΝΕΣ, ΔΙΑΜΕΤΡΟΣ και ΥΠΟΣΤΑΔΙΑ.

Θα εξετάσουμε τώρα τα Cox-Snell και τα Standardized υπόλοιπα για να ανιχνεύσουμε τυχόν ακραίες παρατηρήσεις αλλά και για να δούμε αν η κατανομή που επιλέξαμε είναι η κατάλληλη.



Σχήμα 6.20: Cox-Snell υπόλοιπα για το μοντέλο της Weibull



Σχήμα 6.21: Standardized υπόλοιπα για το μοντέλο της Weibull

Από τα Σχήματα 6.20 και 6.21 των υπολοίπων δεν φαίνεται να υπάρχει κάποια ακραία παρατήρηση αλλά και η κατανομή που επιλέξαμε δείχνει ικανοποιητική.

Για να βελτιώσουμε τώρα το μοντέλο θα εφαρμόσουμε την διαδικασία διαδοχικής αφαίρεσης (stepwise), με την βοήθεια R, η οποία γίνεται με βάση το κριτήριο AIC.

	Value	Std. Error	z	Pr(>|z|)
ΗΛΙΚΙΑ	0.0217	0.0100	2.160	0.003
ΚΑΠΝΙΣΤΗΣ	-1.7794	0.5117	-3.478	<0.001
ΛΕΜΦΑΔΕΝΕΣ	-0.0686	0.0379	-1.811	0.007
ΔΙΑΜΕΤΡΟΣ	0.5986	0.1855	3.226	0.001
ΥΠΟΣΤΑΔΙΑf2	-0.1512	0.3004	-0.503	0.615
ΥΠΟΣΤΑΔΙΑf3	-1.0951	0.5485	-1.997	0.045
ΥΠΟΣΤΑΔΙΑf4	-2.4457	0.9046	-2.704	0.006
Scale= 0.453				
Weibull distribution				
Loglik(model)= -149.2 Loglik(intercept only)= -163.8				
Chisq= 29.18 on 7 degrees of freedom, p= 0.00013				
AIC=316.3261				

Αποτελέσματα 6.6 για το τελικό μοντέλο Weibull

Συμπεράσματα

Στην παρούσα εργασία έγινε μια βιβλιογραφική ανασκόπηση των μεθόδων που χρησιμοποιούνται στην ανάλυση επιβίωσης καθώς και εφαρμογή τους σε δεδομένα καρκίνου του πνεύμονα με σκοπό να βρεθούν ποιες μεταβλητές επηρεάζουν σημαντικά τον χρόνο επιβίωσης των ασθενών. Πιο συγκεκριμένα παρουσιάστηκε το μη παραμετρικό μοντέλο Kaplan-Meier, το ημιπαραμετρικό μοντέλο του Cox, στο οποίο έγινε εφαρμογή της καμπύλης ROC και το παραμετρικό μοντέλο της Weibull. Σε γενικές γραμμές όλες οι μέθοδοι συμφώνησαν ως προς το ποιες μεταβλητές επηρεάζουν τον χρόνο ζωής, ωστόσο υπήρχαν μικρές διαφοροποιήσεις.

Από την εφαρμογή της Kaplan-Meier προέκυψε ότι ο χρόνος επιβίωσης των ασθενών επηρεάζεται από το αν ο ασθενής ήταν καπνιστής ή όχι, καθώς υπήρχε μεγάλη διαφοροποίηση στις καμπύλες επιβίωσης. Στην συνέχεια εφαρμόσαμε το μοντέλο αναλογικής διακινδύνευσης του Cox όπου προέκυψε ότι στο χρόνο επιβίωσης επιδρούν τρεις μεταβλητές. Η πρώτη και πιο σημαντική ήταν ο καπνιστής και ακολούθησαν οι λεμφαδένες και ο ιστολογικός τύπος. Είδαμε ότι για έναν ασθενή που είναι καπνιστής η συνάρτηση διακινδύνευσης αυξάνεται κατά 9,93 σε σχέση με έναν που δεν είναι. Κάθε επιπλέον λεμφαδένας αυξάνει την συνάρτηση διακινδύνευσης κατά 1,11. Για έναν ασθενή που έχει αδενοκαρκίνωμα μειώνεται κατά 0,39 σε σχέση με έναν που έχει πλακώδες. Στο μοντέλο αυτό χρειάστηκε να εφαρμόσουμε στρωματοποίηση ως προς την μεταβλητή καπνιστής καθώς όπως είδαμε και από Σχήμα 6.10 δεν υπήρχε αναλογία μεταξύ των κατηγοριών, κάτι που προϋποθέτει το μοντέλο του Cox. Σημαντικότεροι παράγοντες προέκυψαν οι μεταβλητές υποστάδια και διάμετρος του όγκου. Επιπλέον εφαρμόσαμε και μια καμπύλη ROC στο στρωματοποιημένο μοντέλο όπου εξετάσαμε την προβλεπτική του ικανότητα, η οποία δεν ήταν πολύ ικανοποιητική. Αντιθέτως το μοντέλο του Cox που προέκυψε από τη διαδικασία διαδοχικής αφαίρεσης όπως είδαμε από τα γραφήματα έχει καλύτερη προβλεπτική ικανότητα.

Τέλος, προσαρμόσαμε και το παραμετρικό μοντέλο της Weibull αφού πρώτα κάναμε κάποιους γραφικούς ελέγχους για να ελέγξουμε την καταλληλόλητα του.

Εφαρμόζοντας και την διαδικασία διαδοχικής αφαίρεσης με την βοήθεια της R καταλήξαμε στο τελικό μοντέλο με τις συμμεταβλητές καπνιστής, διάμετρος, λεμφαδένες, υποστάδια και ηλικία.

Βιβλιογραφία

[1] Aalen, O.O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist*., **6,** 701-726.

[2] Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist*., **10,** 1100-1120.

[3] Barlow, W.E. and Prentice, R.L. (1988). Residuals for relative risk regression. *Biometrika*, **75,** 65-74.

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

[5] Cai, Z. and Sun, Y. (2003). Local linear estimation for time-dependent coefficients in Cox's regression models. *Scandinavian Journal of Statistics* **30,** 93-111.

[6] C. Caroni (2004). Diagnostics for Cox 's Proportional Hazards Model. *"Parametric and Semiparametric Models with Applications to Reliability, Survival Analysis, and Quality of Life"* (pp. 27-38). Birkhauser, Boston: M. S. Nikulin, N. Balakrishnan, M. Mesbah and N. Limnios, Eds.

[7] C. Caroni (2017). *First Hitting Time Regression Models: Lifetime Data Analysis Based on Underlying Stochastic Processes*, Wiley-ISTE, London.

[8] Collet, D. (2003). *Modelling Survival Data in Medical Research,* Chapman and Hall, London.

[9] Cox, D. R. (1972). Regression models and life tables (with Discussion). *Journal Royal Statistical Society, A*, **34,** 187-220.

[10] Cox, D.R. and Snell, E.J. (1968) A general definition of residuals (with discussion). *J. R. Statist. Soc. B*, **30,** 248-275.

[11] Fleming, T.R. and Harrington, D.P (1991), *Counting Processes and Survival Analysis.* New York: John Wiley & Sons.

[12] Grambsch, P.M. and Therneau, T.M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81,** 515-526.

[13] Hanley, J.A. and McNeil, B. (1982). The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology* **143,** 29-36.

[14] Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B* **55,** 757-796.

[15] Heagerty, P.J. and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics,* **61,** 92-105.

[16] Hosmer, D.W., Lemeshow, S. and May S. (2008). *Applied Survival Analysis: Reggression Modeling of Time-to-Event Data.* John Wiley & Sons, New Jersey.

[17] Kaplan, E.L. and Meier, P. (1958). Nonparametric Estimation from incomplete observations, *Journal of the J.Amer.Stat. Assoc*., **53,** 457-481.

[18] Nelson, W.B. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics,***14,** 945-966.

[19] O' Hara Hines, R.J. and Carter, E.M. (1993). Improved added variable and partial residual plots for the detection of influential observations in generalized linear models. *Appl. Statist*., **42,** 3-20.

[20] Parmar, M.K.B. and Machin, D. (1995). *Survival Analysis: A Practical Approach*, John Wiley, Chichester.

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

[22] Therneau, T.M. and Grambsch, P.M. (2000). *Modeling Survival Data: Extending the Cox Model*., Springer-Verlag, New York.

[23] Therneau, T.M., Grambsch, P.M. and Fleming, T.R. (1990). Martingale- based residuals for survival models. *Biometrika*, **77**, 147-160.

[24] Winnett, A. and Sasieni, P. (2001). A note on scaled Schoenfeld residuals for the proportional hazards model. *Biometrika*, **88,** 565-571.

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

Ιστότοποι:

https://en.wikipedia.org/wiki/Receiver_operating_characteristic

http://www.lungcancer.gr/portal/content/karkinos/oz_20071017291.php3

https://en.wikipedia.org/wiki/Survival_analysis

http://www.statsdirect.com/help/survival_analysis/kaplan_meier.htm

Παράρτημα

Εντολές **R:**

Εισαγωγή των δεδομένων:

```
library(splines)
library(survival)
data<-read.csv(file.choose(),header=TRUE,sep=";")
data
c<-data[,3]
t<-data[,2]
ΗΛΙΚΙΑ<-data[,4]
ΦΥΛΛΟ<-data[,5]
ΚΑΠΝΙΣΤΗΣ<-data[,6]
ΕΙΔΟΣ_ΕΚΤΟΜΗΣ<-data[,7]
ΛΕΜΦΑΔΕΝΕΣ<-data[,8]
ΔΙΑΜΕΤΡΟΣ<-data[,9]
ΣΤΑΔΙΑ<-data[,10]
ΥΠΟΣΤΑΔΙΑ<-data[,11]
ΙΣΤΟΛΟΓΙΚΟΣ_ΤΥΠΟΣ<-data[,12]
ΒΑΘΜΟΣΔΙΑΦΟΡ<-data[,13]
ΒΑΘΜΟΣΔΙΑΦΟΡf[1:75]
```

Προσαρμογή του μοντέλου του Cox:

```
mod1<-coxph(Surv(t,c)~ΗΛΙΚΙΑ+ΦΥΛΛΟ+ΚΑΠΝΙΣΤΗΣ+ΕΙΔΟΣ_ΕΚΤΟΜΗΣ+ΛΕΜΦΑΔΕΝΕΣ+ΔΙΑΜΕΤΡΟΣ+
+ factor(ΥΠΟΣΤΑΔΙΑf)+ΙΣΤΟΛΟΓΙΚΟΣ_ΤΥΠΟΣ+factor(ΒΑΘΜΟΣΔΙΑΦΟΡf))
```

Διαδικασία διαδοχικής αφαίρεσης:

```
mod2<-step(mod1, direction="backward")
mod3<-coxph(Surv(t,c)~ΚΑΠΝΙΣΤΗΣ+ΛΕΜΦΑΔΕΝΕΣ+ΙΣΤΟΛΟΓΙΚΟΣ_ΤΥΠΟΣ)
```

'Ελεγχος για την υπόθεση αναλογικότητας:

```
cox.zph(mod3)
par(mfrow=c(2,2))
```

Schoenfeld residuals:

```
plot(cox.zph(mod3))
cox.zph(mod1,transform='identity')
cox.zph(mod2,transform='identity')
```

```
sresid<-resid(mod3,type="schoenfeld")

sresid<-resid(mod3,type="scaledsch")
```

df beta residuals:

```
par(mfrow=c(2,2))

for(j in 1:4)

+    plot(sresid[,j],xlab="χρόνος",ylab=names(coef(mod3))[j])

abline(h=0,lty=3)

dfbeta<-residuals(mod3,type='dfbeta')

par(mfrow=c(2,2))

for(j in 1:4)

+ plot(dfbeta[,j],xlab="χρόνος",ylab=names(coef(mod2))[j])
```

Martingale residuals:

```
abline(h=0,lty=3)

par(mfrow=c(1,1))

 res<-residuals(mod2,type=('martingale'))

X<-as.matrix(data[,c("λεμφαδένες")])

par(mfrow=c(1,1))

for(j in 1:1){

+   plot(X[,j],res,xlab="ΛΕΜΦΑΔΕΝΕΣ"[j],ylab="residuals")

+  abline(h=0 ,lty=1)

+  lines(lowess(X[,j],res,iter=0))

+  }
```

Component+residual:

```
b<-coef(mod3)[c(3)]

for ( j in 1:1){

+   plot(X[,j],b[j]*X[,j]+res,xlab="λεμφαδένες"[j],ylab="component+residual")

+   abline(lm(b[j]*X[,j]+res ~X[,j]),lty=2)

+   lines(lowess(X[,j],b[j]*X[,j]+res,iter=0))

+ }
```

Για την στρωματοποίηση:

```
library(lattice)

mods<-
coxph(Surv(t,c)~ΗΛΙΚΙΑ+ΦΥΛΛΟ+strata(ΚΑΠΝΙΣΤΗΣ)+ΕΙΔΟΣ_ΕΚΤΟΜΗΣ+ΛΕΜΦΑΔΕΝΕΣ+ΔΙΑΜΕΤΡΟΣ+factor(ΥΠΟΣΤΑΔΙΑf)+ΙΣΤΟΛΟΓΙΚΟΣ_ΤΥΠΟΣ+factor(ΒΑΘΜΟΣΔΙΑΦΟΡf),data=data,method="breslow")

bh<-basehaz(mods,centered=TRUE)

lnhazard<-log(bh[,1])
```

```
Intime<-log(bh[,2])

xyplot(Inhazard~Intime,group=strata,auto.key=TRUE,data=bh)

library(risksetROC)

modf<-coxph(Surv(t,c)~strata(ΚΑΠΝΙΣΤΗΣ)+ΔΙΑΜΕΤΡΟΣ+factor(ΥΠΟΣΤΑΔΙΑf),data=data,method="breslow")
```

Για την καμπύλη ROC:

```
eta<-modf$linear.predictor

ROC10=ROC10=risksetROC(Stime=t,status=c,marker=eta,predict.time=10,method="Cox",main="ROC
Curve",lty=2,col="red",ylab="True Positive",xlab="False Positive")

ROC50=ROC50=risksetROC(Stime=t,status=c,marker=eta,predict.time=50,method="Cox",plot=FALSE)

lines(ROC50$FP,ROC50$TP,lty=3,col="green")

legend(.6,.25,lty=c(2,3),col=c("red","green"),legend=c("t=10","t=50"),bty="n")
```

Για το γράφημα AUC:

```
risksetAUC(Stime=t,status=c,marker=eta,method="Cox",tmax=65,main="AUC Curve",lty=2,col="red")
```

Προσαρμογή του μοντέλο Weibull:

```
mod <-survreg(Surv(t,c)~ΗΛΙΚΙΑ+ΦΥΛΛΟ+ΚΑΠΝΙΣΤΗΣ+ΕΙΔΟΣ_ΕΚΤΟΜΗΣ+ΛΕΜΦΑΔΕΝΕΣ+ΔΙΑΜΕΤΡΟΣ
+ΥΠΟΣΤΑΔΙΑf+ΙΣΤΟΛΟΓΙΚΟΣ_ΤΥΠΟΣ+ΒΑΘΜΟΣΔΙΑΦΟΡf,dist="weibull")
```

Διαδικασία διαδοχικής αφαίρεσης:

```
step(mod, direction="backward")
```