ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
Επιστήμη του Διαδικτύου
«Web Science»

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Βιοστατιστική: Η περίπτωση της χρήσης της γλώσσας R στην Ανάλυση Επιβίωσης
Biostatistics: The case of Using the R Language in Survival Analysis

Ποζίδου Παρθένα

Επιβλέπων: **Νικόλαος Φαρμάκης**
Αν. Καθηγητής Α.Π.Θ.

Θεσσαλονίκη, Δεκέμβριος 2013

ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
Επιστήμη του Διαδικτύου
«Web Science»

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Βιοστατιστική: Η περίπτωση της χρήσης της γλώσσας R στην Ανάλυση Επιβίωσης

Biostatistics: The case of Using the R Language in Survival Analysis

Ποζίδου Παρθένα

Επιβλέπων: **Νικόλαος Φαρμάκης**
Αν. Καθηγητής Α.Π.Θ.

Εγκρίθηκε από την Τριμελή Εξεταστική Επιτροπή την η Δεκεμβρίου 2013.

.....	.....	.....
Ν. Φαρμάκης	Ι. Αντωνίου	Φ. Κολυβά Μαχαίρα
Αν. Καθηγητής Α.Π.Θ.	Καθηγητής Α.Π.Θ.	Αν. Καθηγήτρια Α.Π.Θ.

Θεσσαλονίκη, Δεκέμβριος 2013

……..

Ποζίδου Παρθένα
Πτυχιούχος Στατιστικής Ο.Π.Α.

ΠΕΡΙΛΗΨΗ

Η ανάλυση επιβίωσης χρησιμοποιείται για την εύρεση της σχέσης μεταξύ μιας μεταβλητής που δηλώνει το χρόνο επιβίωσης κι άλλων μεταβλητών. Η χρήση της είναι συνεχής μιας κι αφορά διάφορες επιστήμες και μελέτες. Στις κοινωνικές επιστήμες ονομάζεται ανάλυση της ιστορίας του γεγονότος (event history analysis) και στη μηχανική είναι γνωστή ως θεωρία αξιοπιστίας (reliability analysis).

Η R είναι μια ελεύθερη γλώσσα ανοιχτή προς χρήση και τροποποίηση κι εξελίσσεται καθημερινά. Χρησιμοποιείται ευρέως από φοιτητές, ακαδημαϊκούς ακόμη κι από εταιρείες που σκοπό έχουν τις στατιστικές μελέτες, την έρευνα αγοράς, ιατρικές μελέτες κ.τ.λ.

Στόχος της εργασίας αυτής είναι ο εμπλουτισμός της ιστοσελίδας της R, http://cran.r-project.org με ένα εγχειρίδιο χρήσης της R σε θέματα ανάλυσης επιβίωσης, στην ελληνική γλώσσα, ώστε να ενισχυθεί η διάδοση της και η εξέλιξη της. Αξίζει να σημειωθεί ότι το εγχειρίδιο αυτό αφορά άτομα που γνωρίζουν τα στοιχειώδη της R όπως διανύσματα, πίνακες κτλ.

Συγκεκριμένα, η εργασία χωρίζεται σε τρία κεφάλαια, τα δύο πρώτα αφορούν το θεωρητικό μέρος, το τρίτο το πρακτικό με τα παραδείγματα κι ακολουθεί το παράρτημα στην R. Συγκεκριμένα, στο πρώτο κεφάλαιο γίνεται μια εισαγωγή στην ανάλυση επιβίωσης, στις έννοιες των λογοκριμένων και αποκομμένων δεδομένων, στο δεύτερο στη συνάρτηση του χρόνου επιβίωσης, καθώς επίσης στις σχέσεις μεταξύ τους. Αμέσως μετά αναπτύσσεται η μέθοδος Kaplan Meier ως μη παραμετρική μέθοδος εκτίμησης της συνάρτησης επιβίωσης. Ακολουθεί η ανάλυση για το μοντέλο του Cox, των καταλοίπων και των μεθόδων για τη σύγκριση καμπυλών επιβίωσης όπως επίσης αναπτύσσονται παραμετρικές μέθοδοι εκτίμησης των συναρτήσεων επιβίωσης. Στο τρίτο αναλύονται δεδομένα με τις προαναφερθείσες μεθόδους, που εμπεριέχονται στην R κι ακολουθεί το μέρος στο οποίο εμπεριέχονται οι συναρτήσεις της R για την ανάλυσης επιβίωσης.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ

Μέθοδος Kaplan-Meier, Μέθοδος Cox, Μέθοδος Weibull, Ανάλυση Επιβίωσης, Γλώσσα R, Residuals

ABSTRACT

Survival analysis, also called event history analysis in social science, or reliability analysis in engineering, deals with time until occurrence of an event or the time of failure.

Survival analysis is used to find the relationship between a variable that indicates the survival time and other variables. The use of it is continuous and concerns various sciences and studies.

R is a free open language and is used with great success in terms of being free to be changed, modified and evolved daily. It is widely used by students, scholars even from companies that design statistical studies, market research, medical studies etc.

The aim of this work is to enrich the site of R, http://cran.r-project.org with a manual of using R in survival analysis in the Greek language in order to strengthen the diffusion and development of R. It should be noted that this manual is applicable to people who know the basics of R as vectors, arrays, etc.

The work is divided into three sections, the first two is the theoretical part and the third includes the practical examples. The appendix in R, follows with the function used in survival analysis and in the examples presented in this task. Specifically, in the first chapter we introduce concepts such as the survival analysis, the uncensored and truncated data, the function of survival time and their interrelations. The second chapter is about the estimation of survival analysis and the methods that compare the survival curves. Following we refer to the Kaplan Meier method as a nonparametric method to estimate the survival function. Next is the Cox model, the residuals as well the parametric methods for estimating the survival functions. In the third chapter, there is the analysis of the data included in R using the methods mentioned above and the next the functions of R for survival analysis are presented.

KEY WORDS

Kaplan-Meier Method, Cox Proportional Hazard model, Weibull Parametric Model, Survival Analysis, R Language, Residuals

Θερμές ευχαριστίες

στους Καθηγητές μου και

σε αυτούς τους συμφοιτητές

που με στήριξαν

με το πάθος τους

για γνώση,

την υπομονή και

την επιμονή τους

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1-ΑΝΑΛΥΣΗ ΕΠΙΒΙΩΣΗΣ

1.1. Εισαγωγή

Η ανάλυση επιβίωσης είναι ο κλάδος της στατιστικής που ασχολείται με δεδομένα που σχετίζονται με το χρόνο μέχρι την εμφάνιση ενός γεγονότος όπως ο θάνατος ενός ασθενούς, η απόλυση, η βλάβη μιας μηχανής κ.ο.κ.

Πρωτοξεκίνησε για χρήση σε θέματα επιβίωσης ζωής σε βιολογικές και ιατρικές εφαρμογές κι ονομάστηκε ανάλυση επιβίωσης, σε τεχνικά θέματα είναι πλέον γνωστή ως θεωρία αξιοπιστίας.

Η ανάλυση επιβίωσης είναι ένα μοντέλο του χρόνου μέχρι την αποτυχία (time to failure) ή του χρόνου μέχρι την εκδήλωση του γεγονότος (time to event).

Η διαφορά της με τη γραμμική παλινδρόμηση, είναι ότι έχει διχοτομικό (δυαδικό) αποτέλεσμα και σε σχέση με τη λογιστική παλινδρόμηση, η ανάλυση επιβίωσης αναλύει το χρόνο σε ένα συμβάν λαμβάνοντας υπόψη τις λογοκριμένες παρατηρήσεις.

Χρησιμοποιείται για να συγκρίνει τα ποσοστά επιβίωσης μεταξύ δύο ή περισσοτέρων ομάδων και να αξιολογήσει τη σχέση μεταξύ των μεταβλητών και το χρόνο επιβίωσης.

Συγκεκριμένα, οι στόχοι της **ανάλυσης επιβίωσης** είναι:

- Η εκτίμηση του χρόνου ως το γεγονός time-to -event για μια ομάδα ατόμων, όπως είναι ο χρόνος μέχρι την επανεμφάνιση του καρκίνου για μια ομάδα ασθενών.
- Για να συγκρίνουμε τους χρόνους time-to-event ανάμεσα σε δύο ή περισσότερες ομάδες , όπως τα φάρμακα έναντι του placebo σε ασθενείς με μια τυχαιοποιημένη ελεγχόμενη κλινική δοκιμή.
- Να αξιολογήσει τη σχέση των μεταβλητών όπως για παράδειγμα το βάρος, η ινσουλίνη, η χοληστερίνη, η πίεση στην επίδραση του χρόνου επιβίωσης των ασθενών.

1.2. Λογοκρισία

Οι λογοκριμένες (censored) παρατηρήσεις είναι δεδομένα που παρόλο που έχουμε πληροφορίες σχετικά με το μεμονωμένο χρόνο επιβίωσης τους, εκ των υστέρων δεν γνωρίζουμε επακριβώς τον χρόνο επιβίωσης.

Παραδείγματα με censored παρατηρήσεις είναι τα εξής:
Το άτομο δεν βιώνει το γεγονός πριν το τέλος της έρευνας, χάνεται από την καταγραφή και δεν μπορούμε να το ελέγξουμε ή αποσύρεται από την έρευνα εξαιτίας κάποιου άλλου γεγονότος το οποίο δεν το ελέγχουμε ή δεν έχουμε δικαιοδοσία.

Το γεγονός της μη γνώσης του ακριβούς χρόνου επιβίωσης εκφράζεται με κάποιους τρόπους, διαφορετικούς σε καθεμία περίπτωση, όπως:

- **Right censoring** όταν δεν έχουμε γνώση σχετικά με τον ακριβή χρόνο επιβίωσης μετά το τέλος της έρευνας
- **Left censoring** όπου έχουμε εμφάνιση του γεγονότος πριν από την αρχή της έρευνας αλλά δεν ξέρουμε κατά πόσο
- **Interval censoring** στο μεσοδιάστημα.

Αναλυτικότερα:

α. **Right censoring (T>t)**

Γνωρίζουμε ότι το αντικείμενο έχει επιβιώσει τουλάχιστον ως το χρόνο *t*. Ο Θάνατος δεν οφείλεται στον λόγο που εξετάζεται. Αδυναμία συνέχισης της μελέτης δηλαδή το αντικείμενο χάνεται από την καταγραφή και δεν μπορούμε να το ελέγξουμε.

Ανάλογα με τον τρόπο που διεξάγουμε την έρευνα μπορούμε να έχουμε 3 τύπους right censoring:

- **Τύπου Ι** όταν διεξάγουμε την έρευνα που έχει έναν προκαθορισμένο χρόνο καταγραφής και όσα γεγονότα δεν έχουν συμβεί ακόμα ή έχουν

συμβεί πέραν αυτού του χρόνου θεωρούνται censored.

- **Τύπου ΙΙ** όταν η έρευνα έχει έναν προκαθορισμένο αριθμό από γεγονότα που παρατηρούνται και σταματάει όταν φτάσουν αυτό το νούμερο. Οι υπόλοιπες περιπτώσεις θεωρούνται censored γεγονότα.
- **Τύπου ΙΙΙ ή Τυχαία** όταν η κάθε παρατηρούμενη περίπτωση έχει χρόνο εισόδου στο μοντέλο που είναι ανεξάρτητο του χρόνου censored γεγονότων και οι χρόνοι επιβίωσης είναι διαφορετικοί. Αυτή η περίπτωση συμβαίνει πολύ συχνά σε επιδημιολογικά δεδομένα όπου οι ασθενείς εισέρχονται σε τυχαίο χρόνο κατά την διάρκεια της έρευνας.

β. Left censoring (T<t)

Η εκδήλωση του γεγονότος πραγματοποιήθηκε πριν από την έναρξη της μελέτης. Ο πραγματικός χρόνος επιβίωσης είναι μικρότερος από αυτόν που παρατηρήθηκε της επιβίωσης του ατόμου. Στην αριστερή λογοκρισία γνωρίζουμε ότι έγινε το συμβάν, αλλά δεν ξέρουμε πότε ακριβώς πριν την παρατήρηση π.χ. Νόσος Alzheimer, η έναρξή της είναι γενικά δύσκολο να προσδιοριστεί.

χ. Interval censoring (a<T<b)

Όταν το γεγονός συμβαίνει μεταξύ 2 χρονικών σημείων αλλά δεν γνωρίζουμε τις ακριβείς ημερομηνίες π.χ. μέτρηση καρκινικών κυττάρων σε ένα διάστημα 3-6 μηνών.

1.3.Αποκομμένα δεδομένα

Εκτός όμως από το λογοκριμένα δεδομένα υπάρχουν και τα αποκομμένα τα οποία διαφέρουν από τα λογοκριμένα στο ότι για τα δεύτερα έχουμε τουλάχιστον ατελείς παρατηρήσεις για όλα τα αντικείμενα μελέτης ενώ για τα αποκομμένα δεδομένα υπάρχει ένα ποσοστό παρατηρήσεων που μας διαφεύγει εντελώς και δεν έχουμε καθόλου πληροφορίες.

Χωρίζονται σε 3 κατηγορίες:

A. Right truncated

Στα από δεξιά αποκομμένα δεδομένα έχουμε άτομα που δεν περιλαμβάνονται στο δείγμα γιατί η ύπαρξή τους δεν είναι γνωστή. Έχουν χρόνο T μεγαλύτερο από κάποιο χρόνο t, όπως η λήξη της μελέτης, πέρα από τον οποία η παρατήρηση του πειράματος δεν είναι δυνατή.

B. Left truncated

Ομοίως όπως με τα δεξιά αποκομμένα δεδομένα, υπάρχουν άτομα που απορρίφθηκαν από την μελέτη και ο ερευνητής δεν γνωρίζει τίποτα για την ύπαρξή τους. Έστω T ο χρόνος επιβίωσης και t_0 π.χ. η έναρξη της μελέτης και πριν από τον χρόνο της έναρξης η παρατήρηση του πειράματος δεν είναι δυνατή. Τότε, μόνο τα άτομα για τα οποία $T>t_o$ μπορούν να συμμετέχουν στη μελέτη.

C. Interval truncated

Σε αυτή την κατηγορία μια παρατήρηση συμπεριλαμβάνεται στο δείγμα μόνο εάν πέσει μέσα σε ένα διάστημα στο χρόνο, όπου η παρατήρηση του φαινομένου είναι δυνατή. Δηλαδή, ο χρόνος ενδιαφέροντος T παρατηρείται μόνο υπό την συνθήκη ότι το T ανήκει στο B.

1.4.Συνάρτηση επιβίωσης και συνάρτηση κινδύνου

Έστω T η τυχαία μεταβλητή μη αρνητικών τιμών που παρουσιάζει τον χρόνο αποτυχίας failure time ενός ατόμου.

Υποθέτουμε ότι η πιθανότητα κατανομής του T περιγράφεται από την density συνάρτηση $f(t)$ δηλαδή τη Συνάρτηση πυκνότητας πιθανότητας η οποία ορίζεται ως η πιθανότητα του χρόνου αποτυχίας να συμβεί ακριβώς τη χρονική στιγμή t (από το ευρύτερο δυνατό φάσμα του χρόνου t). Προκύπτει από το όριο:

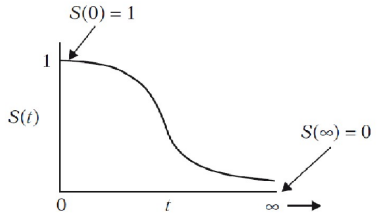
$$f(t) = \lim_{\Delta t \longrightarrow 0} \frac{P(t \le T < t + \Delta t)}{\Delta t}$$

Η συνάρτηση επιβίωσης $S(t)$ και η συνάρτηση κινδύνου χαρακτηρίζουν επιπλέον την κατανομή του T.

Συγκεκριμένα η συνάρτηση επιβίωσης $S(t)$ ορίζεται ως:

$S(t) = P(T \geq t)$ και δίνει την πιθανότητα ενός ατόμου να επιβιώσει σε χρόνο περισσότερο από το χρόνο t. Είναι ίση με $1 - F(t)$, όπου $F(t)$ η αθροιστική συνάρτηση κατανομής του χρόνου T. Δηλαδή $S(t) = 1 - F(t)$.

Οι βασικές της ιδιότητες είναι ότι είναι μονότονα αυξητική, S(0)=1 και S(∞)=0



Η συνάρτηση κινδύνου είναι ένα μέτρο της δυνατότητας ενός γεγονότος να πραγματοποιηθεί τη χρονική στιγμή t, δεδομένου ότι το γεγονός δεν έχει ακόμη συμβεί.

Μεγαλύτερες τιμές της συνάρτησης κινδύνου υποδεικνύουν μεγαλύτερη πιθανότητα να συμβεί το γεγονός.

Συγκεκριμένα η συνάρτηση κινδύνου καθορίζει τον στιγμιαίο ρυθμό αποτυχίας όταν $T = t$ δεδομένης της επιβίωσης στο χρόνο t και ορίζεται από το όριο για $\delta \downarrow 0$ του λόγου:

$$\frac{P(t \leq T < t + \delta \mid T \geq t)}{\delta} = \frac{P(t \leq T < t + \delta)}{P(T \geq t) \times \delta} = \frac{S(t) - S(t + \delta)}{\delta} \times \frac{1}{S(t)}$$

Το όριο αυτό δίνει τη συνάρτηση κινδύνου σε σχέση με τη συνάρτηση πυκνότητας πιθανότητας και της συνάρτησης επιβίωσης: $\lambda(t) = \frac{f(t)}{S(t)}$.

ΚΕΦΑΛΑΙΟ 2-ΜΕΘΟΔΟΙ ΕΚΤΙΜΗΣΗΣ ΤΩΝ ΣΥΝΑΡΤΗΣΕΩΝ ΕΠΙΒΙΩΣΗΣ

2.1. Εκτίμηση της συνάρτησης επιβίωσης

Υπάρχουν δύο μέθοδοι για τη δημιουργία μιας καμπύλης επιβίωσης.
•Η Αναλογιστική μέθοδος όπου ο άξονας x χωρίζεται σε χρονικά διαστήματα, π.χ. μήνες, χρόνια και υπολογίζεται για κάθε διάστημα η επιβίωση.
• Η Kaplan-Meier όπου η επιβίωση υπολογίζεται εκ νέου κάθε φορά που ένας ασθενής πεθαίνει. Η μέθοδος αυτή προτιμάται, εκτός εάν ο αριθμός των ασθενών είναι μεγάλος.

Αξίζει να σημειωθεί ότι οι δύο μέθοδοι ουσιαστικά είναι ίδιες. Στο μόνο που διαφέρουν είναι ότι στη μέθοδο Kaplan-Meier χρησιμοποιούμε τον κάθε χρόνο στον οποίο συμβαίνει μια αποτυχία ή έχουμε μια λογοκρισία ενώ στη μέθοδο των πινάκων επιβίωσης, οι χρόνοι επιβίωσης ομαδοποιούνται σε διαστήματα.

.

Η μέθοδος των Kaplan-Maier είναι μη-παραμετρική μέθοδος εκτίμησης της συνάρτησης επιβίωσης και παίζει καθοριστικό ρόλο στην ανάλυση των δεδομένων επιβίωσης καθώς λαμβάνει υπόψη της τη λογοκρισία. Μπορεί να θεωρηθεί μια ειδική περίπτωση της μεθόδου πινάκων επιβίωσης, όπου κάθε διάστημα περιέχει μόνο μία παρατήρηση. Είναι γνωστή και ως product-limit formula και παρουσιάζει σαν σκαλοπάτια τις καμπύλες επιβίωσης.

Για να χρησιμοποιηθεί η μέθοδος αυτή πρέπει να ισχύουν τα εξής:

- Τα άτομα που χάθηκαν από την παρακολούθηση πρέπει να έχουν την ίδια πιθανότητα επιβίωσης με τα άτομα που συνεχίζουν να παρακολουθούνται. Αυτό

δυστυχώς δε μπορεί να ελεγχθεί και μπορεί να οδηγήσει σε μεροληψία (bias) που μειώνει το S(t).

- Οι πιθανότητες επιβίωσης είναι οι ίδιες για άτομα που εισήλθαν στην αρχή της μελέτης αλλά και για τα άτομα που εισήλθαν αργότερα.
- Το γεγονός που μελετάται (π.χ. θάνατος) συμβαίνει σε καθορισμένο χρόνο. Καθυστερημένη καταγραφή του γεγονότος προκαλεί αύξηση του S(t).

Έστω ότι οι χρόνοι επιβίωσης μαζί με τις λογοκριμένες παρατηρήσεις ενός γκρουπ n ασθενών είναι οι εξής $t_1, t_2, ..., t_n$ και μάλιστα σε τάξη από τον μικρότερο στο μεγαλύτερο, $t_1 \le t_2 \le ... \le t_n$.

Για μια τιμή t βρίσκουμε τη μεγαλύτερη t_i τέτοια ώστε $t_i \le t$, η εκτίμηση της συνάρτησης επιβίωσης είναι:

$$\hat{S}(t) = \frac{r_1 - d_1}{r_1} \times \frac{r_2 - d_2}{r_2} \times ... \times \frac{r_i - d_i}{r_i}$$

Όπου r_k είναι ο αριθμός των ασθενών που είναι ζωντανοί πριν το χρόνο t_k (ο k-οστος χρόνος επιβίωσης) και d_k δείχνει τον αριθμό των ασθενών που έχουν πεθάνει στο χρόνο t_k.

Η εκτιμήτρια Kaplan Meier $\hat{S}(t)$ είναι μια συνάρτηση με βήμα η οποία αλλάζει για χρόνους t_k με θετικό d_k. Κάθε παράγοντας στην εκτιμήτρια Kaplan Meier δείχνει τη διαφορά του 1 από την εκτιμώμενη συνάρτηση κινδύνου. Για χρόνο επιβίωσης t_k λαμβάνουμε υπόψη τον αριθμό των ασθενών εν ζωή, κάποιες φορές ονομάζεται και αριθμός σε ρίσκο δηλαδή ο αριθμός r_k. Η πιθανότητα επιβίωσης δίνεται από το λόγο $(r_k - d_k)/r_k$.

Ακόμα και με βαριά λογοκρισία, η Kaplan-Meier, δίνει μια αμερόληπτη εκτίμηση της πραγματικής (πληθυσμός) καμπύλης επιβίωσης.

Δεν ακολουθεί κάποια κατανομή δηλαδή δεν υπάρχουν παραδοχές για την κατανομή του χρόνου επιβίωσης. Είναι λιγότερο αποτελεσματική από παραμετρικές μεθόδους, αν ο χρόνος επιβίωσης ακολουθεί μια θεωρητική κατανομή αλλά πιο αποτελεσματική όταν δεν είναι γνωστό αν υπάρχουν κατάλληλες θεωρητικές κατανομές. Βασίζεται σε γραφική αναπαράσταση και χρησιμοποιείται συχνά για σύγκριση 2 πληθυσμών

Το μειονέκτημα της είναι ότι απαιτεί κατηγορηματικές μεταβλητές και δεν αποδίδει όταν για παράδειγμα έχουμε πολλές συμμεταβλητές που συμβάλλουν στην επιβίωση. Όπως π.χ. Το κάπνισμα, η υπερλιπιδαιμία, ο διαβήτης, η υπέρταση, συμβάλλουν όλα στο χρόνο για το έμφραγμα του μυοκαρδίου.

Προκύπτει ανάγκη για διαφορετική προσέγγιση. Το πολυπαραγοντικό Cox μοντέλο αναλογικού κινδύνου που θα δούμε αργότερα είναι αυτό που πιο συχνά χρησιμοποιείται.

2.2.Σύγκριση των κατανομών επιβίωσης

Όταν πρόκειται για ανάλυση επιβίωσης, δεν μας ενδιαφέρει μόνο η εκτίμηση της συνάρτησης επιβίωσης, αλλά και η σύγκριση του χρόνου επιβίωσης δύο ή περισσότερων ομάδων που διαφέρουν ως προς ένα χαρακτηριστικό.

Με τις γραφικές μεθόδους σχεδιασμού των εκτιμώμενων συναρτήσεων επιβίωσης, έχουμε μια οπτική εικόνα για το αν υπάρχει διαφορά μεταξύ των συναρτήσεων επιβίωσης δύο ή περισσοτέρων ομάδων. Το πρόβλημα είναι ότι αυτές δεν αρκούν γιατί δε μπορούμε να συμπεράνουμε αν οι διαφορές μεταξύ των συναρτήσεων επιβίωσης είναι σημαντικές. Είναι απαραίτητος ένας στατιστικός έλεγχος καθώς όμως οι χρόνοι επιβίωσης δεν κατανέμονται κανονικά, πρέπει να εφαρμοστούν μη-παραμετρικοί έλεγχοι.

Διάφορες μέθοδοι χρησιμοποιούνται για τη σύγκριση των κατανομών επιβίωσης , εκ των οποίων οι βασικότερες είναι το logrank τεστ ή αλλιώς τεστ Mantel- Cox και το τεστ Gehan ή αλλιώς Wilcoxon . Πρόκειται για ειδικά τεστ λογοκριμένων

παρατηρήσεων. Διαφορετικά, μη παραμετρικά τεστ όπως το τεστ Mann – Whitney θα μπορούσαν να χρησιμοποιηθούν.

Log-Rank Τεστ

Η πιο δημοφιλής μέθοδος για να ελέγξουμε αν οι καμπύλες Kaplan-Maier για 2 ή περισσότερες ομάδες είναι στατιστικά σημαντικές διαφέρουσες, είναι το τεστ Λογαρίθμου-Βαθμού (Log-Rank Τεστ).

Δύο καμπύλες Kaplan-Maier είναι στατιστικά σημαντικές όταν δεν έχουμε αρκετά στοιχεία που αποδεικνύουν ότι οι πραγματικές καμπύλες του πληθυσμού είναι διαφορετικές. Οι όποιες διαφορές υπάρχουν είναι αποτέλεσμα της τυχαίας διακύμανσης.

Το τεστ Log-Rank ή τεστ Mantel-Haenszel ή Mantel-Cox υπολογίζει το στατιστικό p value που χρησιμοποιείται στον έλεγχο υποθέσεων για να βρούμε αν υπάρχει στατιστική σημαντικότητα στις κατανομές 2 ή περισσότερων καμπύλων επιβίωσης.

Το τεστ log-rank χρησιμοποιείται για τον έλεγχο της μηδενικής υπόθεσης της μη ύπαρξης διαφοράς μεταξύ των λειτουργιών επιβίωσης των δύο ομάδων χωρίς να δίνει καμία πληροφορία σχετικά με το μέγεθος της διαφοράς.

Συγκεκριμένα ο έλεγχος υποθέσεων είναι ο εξής:

H_0 : οι καμπύλες είναι στατιστικά ίδιες

H_1 : οι καμπύλες διαφέρουν στατιστικά

Συγκρίνει τον παρατηρούμενο αριθμό θανάτων με τον αναμενόμενο αριθμό στις δύο ομάδες. Έστω O_1 και O_2 οι παρατηρούμενοι αριθμοί θανάτων και E_1και E_2 οι αναμενόμενοι αριθμοί θανάτων στις δύο ομάδες θεραπειών 1 και 2 αντίστοιχα.

Το log-rank στατιστικό υπολογίζεται ως:

$$C^2_{Logrank} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

και ακολουθεί προσεγγιστικά X^2 κατανομή. Μια μεγάλη τιμή του $C^2 Logrank$ θα οδηγήσει στην απόρριψη της H_0.

2.3.Cox Proportional Hazards (PH) Model

Το Cox μοντέλο αναλογικού κινδύνου είναι ένα μοντέλο παλινδρόμησης του σχετικού κινδύνου. Είναι μια πολύ καλή στατιστική τεχνική για την εύρεση της σχέσης μεταξύ της επιβίωσης ενός ασθενή και πολλών επεξηγηματικών μεταβλητών. Επίσης μας βοηθά να εκτιμήσουμε τον κίνδυνο θανάτου ενός ατόμου, ή άλλου γεγονότος που μας ενδιαφέρει δεδομένων των προγνωστικών τους μεταβλητών.

Η εξαρτημένη μεταβλητή Τ μπορεί να είναι όπως ορίσαμε, ο «χρόνος αποτυχίας» ή «χρόνος για το συμβάν».Προσφέρει τις πληροφορίες που απαιτούνται για την ανάλυση επιβίωσης, όπως οι δείκτες κινδύνου και οι καμπύλες επιβίωσης, με τον ελάχιστο αριθμό των υποθέσεων και είναι ένα ισχυρό μοντέλο.

Οι τεχνικές ανάπτυξης του Cox αναλογικού μοντέλου κινδύνου (Cox proportional hazards model) αφορούν τη μοντελοποίηση του χρόνου για το συμβάν και τη σχέση του με μια σειρά από μία ή περισσότερες ερμηνευτικές μεταβλητές παρουσία λογοκρισίας.

Ένα ιδιαίτερο χαρακτηριστικό του μοντέλου Cox PH είναι ότι επικεντρώνεται στην συνάρτηση κινδύνου Hazard function**.** Η συνάρτηση κινδύνου στο μοντέλο είναι το προϊόν μιας αυθαίρετης αρχικής συνάρτησης κινδύνου (baseline hazard function) με ένα σταθερό όρο (εκθετικός όρος), η οποία είναι ανεξάρτητη από το χρόνο t. Οι παράμετροι παλινδρόμησης υπολογίζονται από τη μέθοδο μεγίστης πιθανοφάνειας, χωρίς την ανάγκη να γνωρίζουμε ή να εκτιμήσουμε τη baseline hazard function

Το μοντέλο κινδύνων του Cox είναι ένα ημιπαραμετρικό μοντέλο μιας και δεν κάνουμε καμία υπόθεση για την συνάρτηση κινδύνου και ούτε χρειάζεται να την γνωρίζουμε εκ των προτέρων. Αυτό το χαρακτηριστικό το κάνει πολύ ευσταθές μιας και όταν πχ. το σωστό παραμετρικό μοντέλο ακολουθά την Weibull ή την εκθετική κατανομή, τότε το μοντέλο του Cox τυπικά θα δώσει αποτελέσματα που πλησιάζουν σε

αυτήν.

Το μοντέλο γράφεται ως :

$h(t|X) = h_0(t) \exp(X\beta)$

όπου X είναι ένας πίνακας nxp που περιέχει σε στήλες τις p ανεξάρτητες μεταβλητές σε n παρατηρήσεις, β είναι ένα διάνυσμα διαστάσεων px1και $h_0(t)$ είναι η συνάρτηση βασικού κινδύνου (baseline hazard function) η οποία δεν χρειάζεται να καθοριστεί πλήρως και δηλώνει τον κίνδυνο που έχει οποιοδήποτε άτομο από τα δεδομένα, να υποστεί το προκαθορισμένο γεγονός σε ένα χρόνο t, ανεξάρτητα από τις τιμές της μεταβλητής X.

Το μοντέλο μπορεί να γραφεί και ως εξής:

$$h_i(t) = h_0(t)e^{\beta_1 x_{i1}+\ldots+\beta_k x_{ik}}$$

χρησιμοποιείται συχνά όταν ενδιαφερόμαστε για τον αντίκτυπο των μεταβλητών i και μπορούμε να εκτιμήσουμε τα β1, . . . , βp coefficients χωρίς να χρειαστεί να εκτιμήσουμε την h_0.

Το μοντέλο αναλογικού κινδύνου, μπορεί να χειριστεί συνεχείς μεταβλητές όπως η ηλικία και κατηγορικές μεταβλητές όπως το φύλο. Δεν μπορεί να επεξεργαστεί μεταβλητές που αλλάζουν με το χρόνο αλλά ευτυχώς οι περισσότερες μεταβλητές π.χ. φύλο, εθνικότητα είναι συνεχείς.

Υπόθεση για την Cox proportional hazards είναι η αναλογικότητα, δηλαδή ο κίνδυνος για κάθε άτομο είναι μια σταθερή αναλογία του κινδύνου του κάθε ατόμου. Ο λόγος των κινδύνων είναι σταθερός και δεν εξαρτάται από το χρόνο. Το hazard ratio είναι ο κίνδυνος ενός ατόμου με χαρακτηριστικά 1 να πάθει το προκαθορισμένο γεγονός σε σχέση με το άτομο με τα χαρακτηριστικά 2 και δίνεται από τη σχέση:

$$HR = \frac{h_1(t)}{h_2(t)} = \frac{h_0(t)e^{\beta x_1}}{h_0(t)e^{\beta x_2}} = e^{\beta(x_1 - x_2)}$$

Η υπόθεση των αναλογικών κινδύνων πρέπει να αιτιολογείται πριν γίνει χρήση του μοντέλου. Για τον έλεγχο της αναλογικότητας υπάρχουν γραφικές μέθοδοι αλλά και τεστ καλής προσαρμογής.

Στην πράξη, η αναλογικότητα υπάρχει πάντα εκτός αν υπάρχουν ισχυρές ενδείξεις

για το αντίθετο όπως οι περιπτώσεις όπου 1) οι εκτιμώμενες καμπύλες επιβίωσης είναι αρκετά διαχωρισμένες και μετά συναντιούνται 2) οι εκτιμώμενες καμπύλες επιβίωσης δείχνουν μη παράλληλες με την πάροδο του χρόνου, 3) τα σταθμισμένα κατάλοιπα Schoenfeld αυξάνονται ή μειώνονται με την πάροδο του χρόνου, 4) το τεστ για την αλληλεπίδραση μεταξύ του χρόνου και των μεταβλητών βγαίνει σημαντικό.

2.3.1.Residuals-Κατάλοιπα

Τα κατάλοιπα χρησιμοποιούνται για τον έλεγχο της καταλληλότητας του μοντέλου του Cox, όπως για τον έλεγχο της υπόθεσης αναλογικότητας των κινδύνων, για τον έλεγχο της επάρκειας του μοντέλου και για την εύρεση των ακραίων τιμών. Τα πιο γνωστά κατάλοιπα στο μοντέλο του Cox είναι τα κατάλοιπα Cox-Snell, τα Martingale, τα Schoenfeld και τα Deviance.

Συγκεκριμένα έχουμε τα εξής είδη:

- Cox-Snell residuals (εξετάζει την προσαρμογή του μοντέλου «fitting»)
- Martingale residuals (εξερευνά την λειτουργία της κάθε μεταβλητής covariate και είναι χρήσιμα στο να αποκαλύπτουν τη συναρτησιακή σχέση των επεξηγηματικών μεταβλητών)
- Deviance residuals (εξετάζει την προσαρμογή του μοντέλου «fitting» και αναγνωρίζει «outliers» είναι χρήσιμα στην αναγνώριση των λιγότερο καλών εκτιμώμενων παρατηρήσεων)
- Schoenfeld residuals (εξετάζει την υπόθεση της αναλογικότητας PH assumption και είναι χρήσιμα στο να υποδεικνύουν αν το μοντέλο αναλόγων συναρτήσεων κινδύνου είναι κατάλληλο ή όχι)

Τα κατάλοιπα χρησιμοποιούνται για να ερευνήσουν την έλλειψη προσαρμογής ενός μοντέλου για ένα άτομο. Στο cox μοντέλο δεν μπορεί να μελετηθεί η διαφορά παρατηρούμενου μείον του προβλέψιμου όπως στη γραμμική παλινδρόμηση.

ι. Cox-Snell residuals

Είναι τα κατάλοιπα που πρωτοπαρουσιάστηκαν το 1968 από τον Sir David Cox και τον E. Joyce Snell, για την εκτίμηση της εγκυρότητας της συνάρτησης ανάλυσης επιβίωσης. Η τιμή της συνάρτησης επιβίωσης εξαρτάται από το χρόνο *t* και μια ή περισσότερες παράμετροι. Εκτιμάται από το διάνυσμα θ. Τα κατάλοιπα r_j, δίνονται από τον τύπο:

r_j=-ln{S(t_j; θ)}, όπου S(t_j; θ) είναι η τιμή της εκτιμώμενης ανάλυσης επιβίωσης στο χρόνο t_j. Αν το μοντέλο είναι σωστό τότε τα κατάλοιπα ακολουθούν εκθετική κατανομή με μέση τιμή 1.

ιι. Martingale residual

Προκύπτουν από τη διαφορά του ci (1 αν έχουμε το γεγονός και 0 αν πρόκειται για λογοκριμένες παρατηρήσεις) και της εκτιμώμενης αθροιστικής συνάρτησης κινδύνου για το άτομο i:

c_i -H($t_{i,Xi}$, $ß_i$) .

Για παράδειγμα για ένα άτομο που ήταν censored για 2 μήνες και η predicted cumulative hazard ήταν 30% στους 2 μήνες τα martingale κατάλοιπα είναι 0-0.30 = -0.30 για ένα άτομο που του είχε συμβεί το γεγονός στους 12 μήνες και η predicted cumulative hazard ήταν 60% τα κατάλοιπα είναι: 1-0.60 = 0.40

Δεν είναι κατανεμημένα συμμετρικά γι' αυτό και μετατρέπονται σε deviance residuals.

ιιι. Deviance Residuals

Τα κατάλοιπα απόκλισης βασίζονται στα martingale υπόλοιπα, αλλά κατανέμονται περισσότερο συμμετρικά γύρω από το μηδέν και έτσι είναι χρήσιμα για την ανίχνευση των ακραίων τιμών outliers. Ψάχνουμε για ασυνήθιστα μοτίβα patterns σε ένα διάγραμμα τους σε σχέση με τις μεταβλητές covariates και δείχνουν τις ακραίες τιμές outliers. Είναι συμμετρικά γύρω από το 0 και έχουν standard deviation 1.

Παρατηρήσεις με μεγάλα deviance residuals δεν μπορούν να προβλεφθούν σωστά από το μοντέλο.

Μοιάζουν με τα κατάλοιπα της γραμμικής παλινδρόμησης και είναι αρνητικά για παρατηρήσεις με μεγαλύτερους από τους αναμενόμενους παρατηρούμενους χρόνους επιβίωσης.

ιω. Schoenfeld residuals

Το πλεονέκτημα των κατάλοιπων Schoenfeld έναντι των άλλων είναι ότι για τον υπολογισμό τους δεν χρειάζεται η εκτίμηση της αθροιστικής αναφορικής συνάρτησης κινδύνου. Στα άλλα κατάλοιπα, υπολογίζεται ένα μόνο υπόλοιπο για κάθε άτομο. Τα κατάλοιπα Schoenfeld όμως, υπολογίζουν ένα ξεχωριστό υπόλοιπο για κάθε άτομο για κάθε μεταβλητή. Δηλαδή, εάν έχουμε p μεταβλητές, τότε για κάθε άτομο υπολογίζονται p Schoenfeld υπόλοιπα.

Με την γραφική παράσταση των υπολοίπων Schoenfeld συναρτήσει του χρόνου μπορούμε να ελέγξουμε την υπόθεση αναλογικότητας των κινδύνων (PH υπόθεση). Αν το γράφημα έχει μια τυχαία μορφή των υπολοίπων έναντι του χρόνου τότε ικανοποιείται η PH υπόθεση. Το κατάλοιπο Schoenfeld ορίζεται ως η τιμή της συμμεταβλητής για το i άτομο με πλήρη χρόνο μείον την αναμενόμενη τιμή της συμμεταβλητής για τα άτομα που βρίσκονται σε κίνδυνο την χρονική στιγμή. Προσδιορίζονται για τις μη λογοκριμένες παρατηρήσεις, για τις λογοκριμένες παρατηρήσεις δεν δίνουν δεδομένα

1.5.Παραμετρικά Μοντέλα Ανάλυσης Επιβίωσης

Τα παραμετρικά μοντέλα προϋποθέτουν ότι η κατανομή του χρόνου επιβίωσης είναι γνωστή όπως και η συνάρτηση κινδύνου είναι δεδομένη εκτός από τις τιμές των αγνώστων παραμέτρων.

Στα παραμετρικά μοντέλα δεν είναι απαραίτητες οι πολλές παρατηρήσεις για κάθε συνδυασμό των μεταβλητών αλλά το μειονέκτημα τους είναι ότι τα αποτελέσματα επηρεάζονται από το αν έχει οριστεί σωστά η baseline hazard function.

Παραδείγματα παραμετρικών μοντέλων είναι το Weibull, η εκθετική, η log-normal λογαριθμοκανονική και η loglogistic.

Στην εκθετική η συνάρτηση κινδύνου δίνεται ως σταθερή και μεγαλύτερη του 0: $\lambda(t) = \lambda$ (με $\lambda > 0$), οπότε

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

$F(t) = 1 - \exp(-\lambda t)$ και $f(t) = \lambda \exp(-\lambda t)$.

Στη weibull που είναι μια γενίκευση της εκθετικής, η συνάρτηση κινδύνου ορίζεται ως εξής: $\lambda(t) = \lambda p(\lambda t)^{p-1} = p\lambda^p \times t^{p-1}$.

Είναι μονότονα φθίνουσα για $p < 1$, μονότονα αύξουσα για $p > 1$ και μειώνεται σε σταθερά για $p = 1$.

Η συνάρτηση επιβίωσης προκύπτει από τη σχέση:

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

ΚΕΦΑΛΑΙΟ 3- Η ΓΛΩΣΣΑ R ΚΑΙ ΠΑΡΑΔΕΙΓΜΑΤΑ

3.1.Λίγα λόγια για την R

Η R είναι ένα ελεύθερο λογισμικό για στατιστικούς υπολογισμούς και γραφήματα. Ξεκίνησε το 1992 από τον Ross Ihaka και τον Robert Gentleman στο Πανεπιστήμιο του Auckland, στη Νέα Ζηλανδία.

Μπορεί να χρησιμοποιηθεί σε μια ποικιλία από πλατφόρμες όπως unix, windows, MacOs όπως επίσης μπορεί εύκολα να διασυνδεθεί με γλώσσες προγραμματισμού π.χ. C. Για να την «κατεβάσει» κανείς πηγαίνει στην ιστοσελίδα της CRAN:

http://cran.r-project.org

επιλέγει την πλατφόρμα π.χ Windows

http://cran.r-project.org/bin/windows/base/

Download R 3.0.2 for Windows

Install . . .

Η επόμενη κίνηση είναι το «κατέβασμα» των πακέτων που θα χρησιμοποιηθούν στην R

Install package(s) . . .

load package using library()

π.χ

```
install.packages("survival")
> library(survival)
Loading required package: splines
```

3.1.Παράδειγμα δεδομένων OVARIAN

Πρόκειται για ανάλυση επιβίωσης σε μια μελέτη του καρκίνου των ωοθηκών η οποία συγκρίνει 2 διαφορετικές θεραπείες. Τα δεδομένα αφορούν τον καρκίνο ωοθηκών από την μελέτη των Edmunson et al. (1979)

Οι μεταβλητές που συμμετέχουν στην ανάλυση επιβίωσης είναι:

- futime αριθμός των ημερών στην μελέτη
- fustat δείκτης θανάτου (1) ή censoring (0)
- age ηλικία του ασθενή σε μέρες/365.25
- resid.ds δείκτης της έκτασης της ασθένειας
- rx η θεραπεία που δόθηκε
- ecog.ps μέτρηση της απόδοσης βάση της μονάδας μέτρησης της Eastern Cooperative Oncology Group βλ. http://ecog.org/general/perf_stat.html

Η συνάρτηση που θα χρησιμοποιηθεί είναι η survreg.

Μπορεί να χρησιμοποιηθεί η βοήθεια help(survreg) για εκτενέστερη πληροφόρηση σχετικά με τον τρόπο χρήσης

Καλούμε αρχικά την βιβλιοθήκη «survival» και τα δεδομένα ovarian που εμπεριέχονται σε αυτή.

>library(survival)

Loading required package: splines

> data(ovarian)

> ovarian (εδώ βλέπουμε αναλυτικά τα δεδομένα)

futime fustat age resid.ds rx ecog.ps

1	59	1	72.3315	2 1	1
2	115	1	74.4932	2 1	1
3	156	1	66.4658	2 1	2
4	421	0	53.3644	2 2	1
5	431	1	50.3397	2 1	1
6	448	0	56.4301	1 1	2
7	464	1	56.9370	2 2	2
8	475	1	59.8548	2 2	2
9	477	0	64.1753	2 1	1
10	563	1	55.1781	1 2	2
11	638	1	56.7562	1 1	2
12	744	0	50.1096	1 2	1
13	769	0	59.6301	2 2	2
14	770	0	57.0521	2 2	1
15	803	0	39.2712	1 1	1
16	855	0	43.1233	1 1	2
17	1040	0	38.8932	2 1	2
18	1106	0	44.6000	1 1	1
19	1129	0	53.9068	1 2	1
20	1206	0	44.2055	2 2	1
21	1227	0	59.5890	1 2	2

22 268 1 74.5041 2 1 2

23 329 1 43.1370 2 1 1

24 353 1 63.2192 1 2 2

25 365 1 64.4247 2 2 1

26 377 0 58.3096 1 2 1

> names(ovarian) (δίνει τα ονόματα των μεταβλητών)

[1] "futime" "fustat" "age" "resid.ds" "rx" "ecog.ps"

> summary(ovarian) (περίληψη των βασικών σημείων των δεδομένων)

futime fustat age resid.ds

Min. : 59.0 Min. :0.0000 Min. :38.89 Min. :1.000

1st Qu.: 368.0 1st Qu.:0.0000 1st Qu.:50.17 1st Qu.:1.000

Median : 476.0 Median :0.0000 Median :56.85 Median :2.000

Mean : 599.5 Mean :0.4615 Mean :56.17 Mean :1.577

3rd Qu.: 794.8 3rd Qu.:1.0000 3rd Qu.:62.38 3rd Qu.:2.000

Max. :1227.0 Max. :1.0000 Max. :74.50 Max. :2.000

rx ecog.ps

Min. :1.0 Min. :1.000

1st Qu.:1.0 1st Qu.:1.000

Median :1.5 Median :1.000

Mean :1.5 Mean :1.462

3rd Qu.:2.0 3rd Qu.:2.000

Max. :2.0 Max. :2.000

>summary(survfit(Surv(futime, fustat)~1, data=ovarian))

Call: survfit(formula = Surv(futime, fustat) ~ 1, data = ovarian)

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
59	26	1	0.962	0.0377	0.890	1.000
115	25	1	0.923	0.0523	0.826	1.000
156	24	1	0.885	0.0627	0.770	1.000
268	23	1	0.846	0.0708	0.718	0.997
329	22	1	0.808	0.0773	0.670	0.974
353	21	1	0.769	0.0826	0.623	0.949
365	20	1	0.731	0.0870	0.579	0.923
431	17	1	0.688	0.0919	0.529	0.894
464	15	1	0.642	0.0965	0.478	0.862
475	14	1	0.596	0.0999	0.429	0.828
563	12	1	0.546	0.1032	0.377	0.791
638	11	1	0.497	0.1051	0.328	0.752

summ<-survfit(Surv(futime, fustat)~1, data=ovarian)

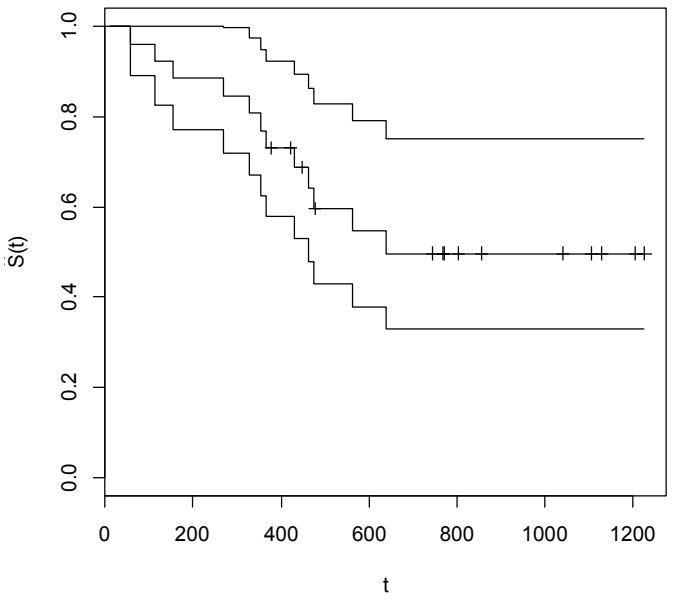
> summ

Call: survfit(formula = Surv(futime, fustat) ~ 1, data = ovarian)

records n.max n.start events median 0.95LCL 0.95UCL

26 26 26 12 638 464 NA

plot(summ,xlab="t",ylab=expression(hat(S)*"(t)"))



Προσαρμόζουμε το μοντέλο Cox:

coxfit <- coxph(Surv(futime, fustat) ~ age + resid.ds +rx+ecog.ps,data=ovarian)

> coxfit

Call:

coxph(formula = Surv(futime, fustat) ~ age + resid.ds + rx +

ecog.ps, data = ovarian)

coef exp(coef) se(coef) z p

age 0.125 1.133 0.0469 2.662 0.0078

resid.ds 0.826 2.285 0.7896 1.046 0.3000

rx -0.914 0.401 0.6533 -1.400 0.1600

ecog.ps 0.336 1.400 0.6439 0.522 0.6000

Likelihood ratio test=17 on 4 df, p=0.0019 n= 26, number of events= 12

> summary(full.fit) δίνει την περίληψη του μοντέλου

Call:

coxph(formula = Surv(futime, fustat) ~ age + resid.ds + rx +

ecog.ps, data = ovarian)

n= 26, number of events= 12

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

age 0.12481 1.13294 0.04689 2.662 0.00777 **

resid.ds 0.82619 2.28459 0.78961 1.046 0.29541

rx -0.91450 0.40072 0.65332 -1.400 0.16158

ecog.ps 0.33621 1.39964 0.64392 0.522 0.60158

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

exp(coef) exp(-coef) lower .95 upper .95

age 1.1329 0.8827 1.0335 1.242

resid.ds 2.2846 0.4377 0.4861 10.738

rx 0.4007 2.4955 0.1114 1.442

ecog.ps 1.3996 0.7145 0.3962 4.945

Concordance= 0.807 (se = 0.091)

Rsquare= 0.481 (max possible= 0.932)

Likelihood ratio test= 17.04 on 4 df, p=0.001896

Wald test = 14.25 on 4 df, p=0.006538

Score (logrank) test = 20.81 on 4 df, p=0.0003449

> prop.coxfit <- cox.zph(full.fit)

> prop.coxfit

rho chisq p

age -0.0399 0.0262 0.871

resid.ds -0.1417 0.2463 0.620

rx 0.1325 0.2001 0.655

ecog.ps 0.4845 1.8819 0.170

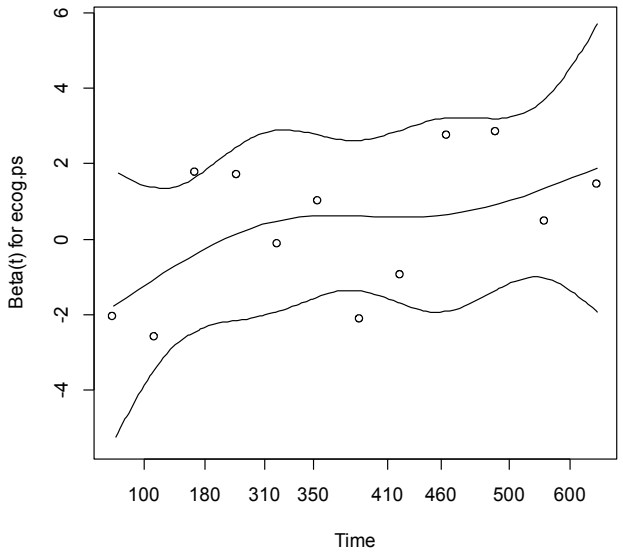
GLOBAL NA 3.3609 0.499

plot(prop.coxfit [1])

plot(prop.coxfit [2])

plot(prop.coxfit [3])

plot(prop.coxfit [4])



Cox για τις 3 μεταβλητές:

> cox.fit1 <- coxph(Surv(futime, fustat) ~ age + resid.ds +rx, data=ovarian)

> cox.fit1

Call:

coxph(formula = Surv(futime, fustat) ~ age + resid.ds + rx, data = ovarian)

	coef	exp(coef)	se(coef)	z	p
age	0.129	1.137	0.0473	2.718	0.0066
resid.ds	0.696	2.006	0.7585	0.918	0.3600
rx	-0.849	0.428	0.6392	-1.328	0.1800

Likelihood ratio test=16.8 on 3 df, p=0.000789 n= 26, number of events= 12

> summary(coxfit1)

Call:

coxph(formula = Surv(futime, fustat) ~ age + resid.ds + rx, data = ovarian)

n= 26, number of events= 12

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
age	0.1285	1.1372	0.0473	2.718	0.00657	**
resid.ds	0.6964	2.0065	0.7585	0.918	0.35858	
rx	-0.8489	0.4279	0.6392	-1.328	0.18416	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

	exp(coef)	exp(-coef)	lower .95	upper .95
age	1.1372	0.8794	1.0365	1.248
resid.ds	2.0065	0.4984	0.4537	8.874

Για τις 2 μεταβλητές age και rx:

> coxfit2 <- coxph(Surv(futime, fustat) ~ age+rx,data=ovarian)

> coxfit2

Call:

coxph(formula = Surv(futime, fustat) ~ age + rx, data = ovarian)

	coef	exp(coef)	se(coef)	z	p
age	0.147	1.159	0.0461	3.19	0.0014
rx	-0.804	0.448	0.6320	-1.27	0.2000

Likelihood ratio test=15.9 on 2 df, p=0.000355 n= 26, number of events= 12

> summary(coxfit2)

Call:

coxph(formula = Surv(futime, fustat) ~ age + rx, data = ovarian)

  n= 26, number of events= 12

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

age 0.14733   1.15873 0.04615 3.193 0.00141 **

rx -0.80397   0.44755 0.63205 -1.272 0.20337

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    exp(coef) exp(-coef) lower .95 upper .95

age   1.1587   0.863   1.0585   1.268

rx   0.4475   2.234   0.1297   1.545

Concordance= 0.798 (se = 0.091)

Rsquare= 0.457   (max possible= 0.932)

Likelihood ratio test= 15.89 on 2 df,   p=0.0003551

Wald test   = 13.47 on 2 df,   p=0.00119

Score (logrank) test = 18.56 on 2 df,   p=9.341e-05

Μπορούμε να αφαιρέσουμε από το μοντέλο τη μεταβλητή resid.ds

> coxfit3 <- coxph(Surv(futime, fustat) ~ rx,data=ovarian)

> coxfit3

Call:

coxph(formula = Surv(futime, fustat) ~ rx, data = ovarian)

coef exp(coef) se(coef) z p

rx -0.596 0.551 0.587 -1.02 0.31

Likelihood ratio test=1.05 on 1 df, p=0.305 n= 26, number of events= 12

> summary(coxfit3)

Call:

coxph(formula = Surv(futime, fustat) ~ rx, data = ovarian)

n= 26, number of events= 12

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

rx -0.5964 0.5508 0.5870 -1.016 0.31

exp(coef) exp(-coef) lower .95 upper .95

rx 0.5508 1.816 0.1743 1.74

Concordance= 0.608 (se = 0.078)

Rsquare= 0.04 (max possible= 0.932)

Likelihood ratio test= 1.05 on 1 df, p=0.3052

Wald test = 1.03 on 1 df, p=0.3096

Score (logrank) test = 1.06 on 1 df, p=0.3026

Διαπιστώνουμε ότι δεν μπορούμε να αφαιρέσουμε τη μεταβλητή age

>pchisq(15.9 - 14.3, df = 1, lower.tail = FALSE)

[1] 0.2059032

Ελέγχουμε την αλληλεπίδραση μεταξύ των μεταβλητών age και rx:

> inter<-coxph(Surv(futime,fustat)~ age * rx, data = ovarian)

> inter

Call:

coxph(formula = Surv(futime, fustat) ~ age * rx, data = ovarian)

coef exp(coef) se(coef) z p

age -0.0122 9.88e-01 0.165 -0.0738 0.94

rx -9.5043 7.45e-05 9.008 -1.0551 0.29

age:rx 0.1464 1.16e+00 0.150 0.9775 0.33

Likelihood ratio test=16.9 on 3 df, p=0.000723 n= 26, number of events= 12

Επομένως δεν υπάρχει αλληλεπίδραση μεταξύ των μεταβλητών.

Το τελικό μοντέλο cox επομένως είναι αυτό που περιλαμβάνει τις 2 μεταβλητές age, rx:

> final.coxfit <- coxph(Surv(futime, fustat) ~age + rx,data=ovarian)

> final.coxfit

Call:

coxph(formula = Surv(futime, fustat) ~ age + rx, data = ovarian)

coef exp(coef) se(coef) z p

age 0.147 1.159 0.0461 3.19 0.0014

rx -0.804 0.448 0.6320 -1.27 0.2000

Likelihood ratio test=15.9 on 2 df, p=0.000355 n= 26, number of events= 12

> summary(final.coxfit)

Call:

coxph(formula = Surv(futime, fustat) ~ age + rx, data = ovarian)

n= 26, number of events= 12

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
age	0.14733	1.15873	0.04615	3.193	0.00141	**
rx	-0.80397	0.44755	0.63205	-1.272	0.20337	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

	exp(coef)	exp(-coef)	lower .95	upper .95
age	1.1587	0.863	1.0585	1.268
rx	0.4475	2.234	0.1297	1.545

Concordance= 0.798 (se = 0.091)

Rsquare= 0.457   (max possible= 0.932)

Likelihood ratio test= 15.89 on 2 df,   p=0.0003551

Wald test   = 13.47 on 2 df,   p=0.00119

Score (logrank) test = 18.56 on 2 df,   p=9.341e-05

Προσαρμόζουμε ένα **Weibull** μοντέλο με όλες τις μεταβλητές αρχικά. Με την Backward μέθοδο καταλήγουμε στις μεταβλητές που είναι στατιστικά σημαντικές στο μοντέλο μας. Το κριτήριο επιλογής είναι το chi-square wald test.

weibullfit <- survreg(Surv(futime, fustat) ~ age + resid.ds + rx + ecog.ps, data = ovarian, dist = "weibull")

> weibullfit

Call:

survreg(formula = Surv(futime, fustat) ~ age + resid.ds + rx +

ecog.ps, data = ovarian, dist = "weibull")

Coefficients:

(Intercept) age resid.ds rx ecog.ps

10.63195736 -0.06503959 -0.52095807 0.52060393 -0.06676905

Scale= 0.5195874

Loglik(model)= -87.8 Loglik(intercept only)= -98

Chisq= 20.21 on 4 degrees of freedom, p= 0.00045

n= 26

> summary(weibullfit)

Call:

survreg(formula = Surv(futime, fustat) ~ age + resid.ds + rx +

ecog.ps, data = ovarian, dist = "weibull")

Value Std. Error z p

(Intercept) 10.6320 1.4237 7.468 8.14e-14

age -0.0650 0.0194 -3.356 7.90e-04

resid.ds -0.5210 0.3921 -1.329 1.84e-01

rx 0.5206 0.3244 1.605 1.08e-01

ecog.ps -0.0668 0.3234 -0.206 8.36e-01

Log(scale) -0.6547 0.2399 -2.729 6.35e-03

Scale= 0.52

Weibull distribution

Loglik(model)= -87.8 Loglik(intercept only)= -98

Chisq= 20.21 on 4 degrees of freedom, p= 0.00045

Number of Newton-Raphson Iterations: 6

n= 26

Για τη σύγκριση και την επιλογή μοντέλων μπορεί να χρησιμοποιηθεί η συνάρτηση **anova.** Κάθε φορά προστίθεται μια μεταβλητή στο μοντέλο αρχίζοντας από το μικρότερο δυνατό μοντέλο αυτό που συνήθως περιλαμβάνει μόνο το intercept.

>anova(weibullfit, test = "Chi")

Df Deviance Resid. Df -2*LL Pr(>Chi)

NULL NA NA 24 195.9078 NA

age 1 15.90533616 23 180.0025 6.659071e-05

resid.ds 1 1.96247801 22 178.0400 1.612485e-01

rx 1 2.30273540 21 175.7373 1.291464e-01

ecog.ps 1 0.04307911 20 175.6942 8.355764e-01

Διαπιστώνουμε ότι η μεταβλητή ecog.ps έχει τη μεγαλύτερη p τιμή άρα τη μικρότερη στατιστική σημαντικότητα κι άρα μπορεί να εξαιρεθεί από το μοντέλο. Μοντελοποιούμε με τις υπόλοιπες μεταβλητές κι έχουμε:

weibullfit2 <- survreg(Surv(futime, fustat) ~ age + resid.ds + rx, data = ovarian, dist = "weibull")

> summary(weibullfit2)

Call:

survreg(formula = Surv(futime, fustat) ~ age + resid.ds + rx,

data = ovarian, dist = "weibull")

	Value	Std. Error	z	p
(Intercept)	10.5634	1.381	7.65	2.02e-14
age	-0.0661	0.019	-3.48	4.96e-04
resid.ds	-0.5002	0.380	-1.32	1.88e-01
rx	0.5152	0.324	1.59	1.11e-01
Log(scale)	-0.6577	0.238	-2.76	5.80e-03

Scale= 0.518

Weibull distribution

Loglik(model)= -87.9 Loglik(intercept only)= -98

Chisq= 20.17 on 3 degrees of freedom, p= 0.00016

Number of Newton-Raphson Iterations: 6

Ομοίως προκύπτει ότι η επόμενη μεταβλητή που μπορεί να εξαιρεθεί από το μοντέλο είναι η resid.ds. Το μοντέλο πλέον είναι το εξής:

weibullfit3 <- survreg(Surv(futime,fustat) ~ age + rx, data=ovarian, dist='weibull')

> summary(weibullfit3)

Call:

survreg(formula = Surv(futime, fustat) ~ age + rx, data = ovarian,

dist = "weibull")

	Value	Std. Error	z	p
(Intercept)	10.4626	1.4427	7.25	4.10e-13
age	-0.0791	0.0198	-4.00	6.41e-05
rx	0.5673	0.3403	1.67	9.55e-02
Log(scale)	-0.5967	0.2352	-2.54	1.12e-02

Scale= 0.551

Weibull distribution

Loglik(model)= -88.8 Loglik(intercept only)= -98

Chisq= 18.38 on 2 degrees of freedom, p= 1e-04

Number of Newton-Raphson Iterations: 5

Επαναφέρουμε την μεταβλητή ecog.ps για δοκιμή:

weibullfit4 <- survreg(Surv(futime,fustat) ~ age + rx + ecog.ps, data=ovarian, dist='weibull')

> summart(weibullfit4)

Call:

survreg(formula = Surv(futime, fustat) ~ age + rx + ecog.ps,

data = ovarian, dist = "weibull")

```
              Value Std. Error      z        p
(Intercept) 10.4085      1.463  7.113 1.14e-12
age         -0.0797      0.020 -3.984 6.77e-05
rx           0.5611      0.341  1.646 9.97e-02
ecog.ps      0.0602      0.331  0.182 8.56e-01
Log(scale)  -0.6030      0.237 -2.545 1.09e-02

Scale= 0.547

Weibull distribution
Loglik(model)= -88.7   Loglik(intercept only)= -98
       Chisq= 18.42 on 3 degrees of freedom, p= 0.00036
Number of Newton-Raphson Iterations: 5
n= 26
```

Καταλήγουμε στη μεμονωμένη μεταβλητή age.

```
weibullfit5<- survreg(Surv(futime,fustat) ~ age, data=ovarian, dist='weibull')
> summary(weibullfit5)
Call:
survreg(formula = Surv(futime, fustat) ~ age, data = ovarian,
   dist = "weibull")
              Value Std. Error     z        p
```

(Intercept) 12.3970 1.4821 8.36 6.05e-17

age -0.0962 0.0237 -4.06 4.88e-05

Log(scale) -0.4919 0.2304 -2.14 3.27e-02

Scale= 0.611

Weibull distribution

Loglik(model)= -90 Loglik(intercept only)= -98

Chisq= 15.91 on 1 degrees of freedom, p= 6.7e-05

Number of Newton-Raphson Iterations: 5

Βλέπουμε ανά δύο τις μεταβλητές με σίγουρη την ηλικία:

weibullfit6<- survreg(Surv(futime,fustat) ~ age + resid.ds, data=ovarian, dist='weibull')

> summary(weibullfit6)

Call:

survreg(formula = Surv(futime, fustat) ~ age + resid.ds, data = ovarian,

dist = "weibull")

Value Std. Error z p

(Intercept) 12.3409 1.3822 8.93 4.31e-19

age -0.0804 0.0223 -3.60 3.18e-04

resid.ds -0.5570 0.4056 -1.37 1.70e-01

Log(scale) -0.5640 0.2366 -2.38 1.71e-02

Scale= 0.569

Weibull distribution

Loglik(model)= -89   Loglik(intercept only)= -98

Chisq= 17.87 on 2 degrees of freedom, p= 0.00013

Number of Newton-Raphson Iterations: 5

weibullfit7<- survreg(Surv(futime,fustat) ~ age + ecog.ps, data=ovarian, dist='weibull')

> summary(weibullfit7)

Call:

survreg(formula = Surv(futime, fustat) ~ age + ecog.ps, data = ovarian,

dist = "weibull")

Value Std. Error z p

(Intercept) 12.2850 1.5015 8.182 2.80e-16

age -0.0970 0.0235 -4.127 3.67e-05

ecog.ps 0.0998 0.3657 0.273 7.85e-01

Log(scale) -0.5054 0.2351 -2.149 3.16e-02

Scale= 0.603

Weibull distribution

Loglik(model)= -90   Loglik(intercept only)= -98

Chisq= 15.98 on 2 degrees of freedom, p= 0.00034

Number of Newton-Raphson Iterations: 5

Καταλήγουμε στο ότι η ηλικία είναι η μόνη σημαντική μεταβλητή στα δεδομένα ovarian κι άρα το μοντέλο το ιδανικό είναι το 5.

> summary(weibullfit5)

Call:

survreg(formula = Surv(futime, fustat) ~ age, data =

ovarian, dist = "weibull")

	Value	Std. Error	z	p
(Intercept)	12.3970	1.4821	8.36	6.05e-017
age	-0.0962	0.0237	-4.06	4.88e-005
Log(scale)	-0.4919	0.2304	-2.14	3.27e-002

Scale= 0.611

Weibull distribution

Loglik(model)= -90 Loglik(intercept only)= -98

Chisq= 15.91 on 1 degrees of freedom, p= 0.000067

Number of Newton-Raphson Iterations: 5

n= 26

Correlation of Coefficients:

	(Intercept)	age
age	-0.992	
Log(scale)	0.363	-0.340

Για τη σύγκριση παραμετρικών μοντέλων το τεστ log-likelihood score βοηθά να βρούμε ποιο είναι το κατάλληλο σετ μεταβλητών στην παραμετρική παλινδρόμηση ανάλυσης επιβίωσης.

Εξετάζουμε το μοντέλο για loglnormal κατανομή ανάλυσης επιβίωσης και loglogistic.

Η anova θα δώσει τις τιμές για το log likelihood score.

lonor <- survreg(Surv(futime,fustat) ~ age, data=ovarian, dist='lognormal')

> lonor

Call:

survreg(formula = Surv(futime, fustat) ~ age, data = ovarian,

dist = "lognormal")

Coefficients:

(Intercept) age

11.37071212 -0.08382014

Scale= 0.8063526

Loglik(model)= -89.7 Loglik(intercept only)= -97.1

Chisq= 14.77 on 1 degrees of freedom, p= 0.00012

lolog <- survreg(Surv(futime,fustat) ~ age, data=ovarian, dist='loglogistic')

> lolog

Call:

survreg(formula = Surv(futime, fustat) ~ age, data = ovarian,

dist = "loglogistic")

Coefficients:

(Intercept) age

11.6494663 -0.0887475

Scale= 0.4492142

Loglik(model)= -89.6 Loglik(intercept only)= -97.4

Chisq= 15.61 on 1 degrees of freedom, p= 7.8e-05

n= 26

anova(weibullfit5, lonor, lolog, test ="Chisq")

Terms	Resid. Df	-2*LL	Test Df	Deviance	Pr(>Chi)
1 age	23	180.0025	NA	NA	NA
2 age	23	179.4698	= 0	0.5326176	NA
3 age	23	179.1017	= 0	0.3681373	NA

Όπως παρατηρείται το τελευταίο μοντέλο lolog, της Log Logistic κατανομής είναι καλύτερο από τα άλλα αφού έχει το μικρότερο γινόμενο -2*LogLik το οποίο αντιστοιχεί στο μεγαλύτερο LogLik.

summary(lolog)

Call:

survreg(formula = Surv(futime, fustat) ~ age, data = ovarian,

dist = "loglogistic")

	Value	Std. Error	z	p
(Intercept)	11.6495	1.3767	8.46	2.63e-17
age	-0.0887	0.0225	-3.95	7.88e-05
Log(scale)	-0.8003	0.2428	-3.30	9.79e-04

Scale= 0.449

Log logistic distribution

Loglik(model)= -89.6 Loglik(intercept only)= -97.4

Chisq= 15.61 on 1 degrees of freedom, p= 7.8e-05

Number of Newton-Raphson Iterations: 5

Το Akaike information criterion ή αλλιώς AIC ορίζεται ως:

AIC = -2LogL + 2p= -2(μέγιστη λογαριθμική πιθανοφάνεια του μοντέλου) +2(αριθμός των παραμέτρων).

Το AIC επιλέγει το μοντέλο που ελαχιστοποιεί το κριτήριο.

Η συνάρτηση survreg δεν δίνει απευθείας την τιμή AIC αλλά μπορεί να βρεθεί από το log likelihood και τις παραμέτρους. Για το μοντέλο weibullfit, το AIC μπορεί να βρεθεί με την εντολή:

-2*weibullfit$loglik[2]+2*(length(weibullfit$coef)-1)

[1] 183.6942

> -2*weibullfit2$loglik[2]+2*(length(weibullfit2$coef)-1)

[1] 181.7373

> -2*weibullfit3$loglik[2]+2*(length(weibullfit3$coef)-1)

[1] 181.5234

-2*weibullfit4$loglik[2]+2*(length(weibullfit4$coef)-1)

[1] 183.4909

-2*weibullfit5$loglik[2]+2*(length(weibullfit5$coef)-1)

[1] 182.0025

-2*weibullfit6$loglik[2]+2*(length(weibullfit6$coef)-1)

[1] 182.04

> -2*weibullfit7$loglik[2]+2*(length(weibullfit7$coef)-1)

[1] 183.9304

Βάση του κριτηρίου αυτού, προτιμούμε το μοντέλο με το μικρότερο AIC δηλαδή το τέταρτο σε σειρά. Με βάση την εντολή extractAIC έχουμε:

> extractAIC(weibullfit)

[1] 6.0000 187.6942

> extractAIC(weibullfit2)

[1] 5.0000 185.7373

> extractAIC(weibullfit3)

[1] 4.0000 185.5234

> extractAIC(weibullfit4)

[1] 5.0000 187.4909

extractAIC(weibullfit5)

[1] 3.0000 186.0025

extractAIC(weibullfit6)

[1] 4.00 186.04

extractAIC(weibullfit7)

[1] 4.0000 187.9304

για το εκθετικό μοντέλο έχουμε:

> exponential1 <- survreg(Surv(futime,fustat) ~ 1, data=ovarian, dist='exponential')

> exponential1

Call:

survreg(formula = Surv(futime, fustat) ~ 1, data = ovarian, dist = "exponential")

Coefficients:

(Intercept)

7.16935

Scale fixed at 1

Loglik(model)= -98 Loglik(intercept only)= -98

Επειδή για την εκθετική στην ανάλυση επιβίωσης ισχύουν: $h(t) = \lambda$ και $S(t) = \exp(-\lambda t)$, έχουμε $\lambda = \exp(-(\text{Intercept})) = \exp(-7.17)$ και $S(t) = \exp(-(\exp(-7.17)t)$.

3.3.Παράδειγμα- Veteran's Administration lung cancer trial

Οι μεταβλητές που συμμετέχουν στα δεδομένα:

- stime

  χρόνος επιβίωσης

- status

  θάνατος ή λογοκριμένη τιμή

- treat

θεραπεία standard or test.

- age

  ηλικία ασθενούς σε έτη

- Karn

  Τιμή Karnofsky από την απόδοση του ασθενούς σε κλίμακα από 0 ως 100.

- diag.time

  χρόνος μέχρι τη διάγνωση σε μήνες στην είσοδο στην κλινική δοκιμή.

- cell

  ένας από τους 4 τύπους κυττάρων.

- prior

  προηγούμενη θεραπεία

Καλούμε τη βιβλιοθήκη survival και τα δεδομένα που εμπεριέχονται στην R με το όνομα αρχείου VA.

> library(MASS)

> library(survival)

Loading required package: splines

> data(VA)

> VA

	stime	status	treat	age	Karn	diag.time	cell	prior
1	72	1	1	69	60	7	1	0
2	411	1	1	64	70	5	1	10
3	228	1	1	38	60	3	1	0
4	126	1	1	63	60	9	1	10
5	118	1	1	65	70	11	1	10
6	10	1	1	49	20	5	1	0
7	82	1	1	69	40	10	1	10
8	110	1	1	68	80	29	1	0

9 314 1 1 43 50 18 1 0

10 100 0 1 70 70 6 1 0

11 42 1 1 81 60 4 1 0

……..

135 231 1 2 67 70 18 4 10

136 378 1 2 65 80 4 4 0

137 49 1 2 37 30 3 4 0

> attach(VA)

> fit1 <- survfit(Surv(stime,status) ~ treat)

> summary(fit1)

Call: survfit(formula = Surv(stime, status) ~ treat)

treat=1

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
3	69	1	0.9855	0.0144	0.95771	1.000
4	68	1	0.9710	0.0202	0.93223	1.000
7	67	1	0.9565	0.0246	0.90959	1.000
8	66	2	0.9275	0.0312	0.86834	0.991
10	64	2	0.8986	0.0363	0.83006	0.973
11	62	1	0.8841	0.0385	0.81165	0.963
12	61	2	0.8551	0.0424	0.77592	0.942
13	59	1	0.8406	0.0441	0.75849	0.932
16	58	1	0.8261	0.0456	0.74132	0.921
18	57	2	0.7971	0.0484	0.70764	0.898
20	55	1	0.7826	0.0497	0.69109	0.886
21	54	1	0.7681	0.0508	0.67472	0.874
22	53	1	0.7536	0.0519	0.65851	0.862
27	51	1	0.7388	0.0529	0.64208	0.850
30	50	1	0.7241	0.0539	0.62580	0.838
31	49	1	0.7093	0.0548	0.60967	0.825
35	48	1	0.6945	0.0556	0.59368	0.812

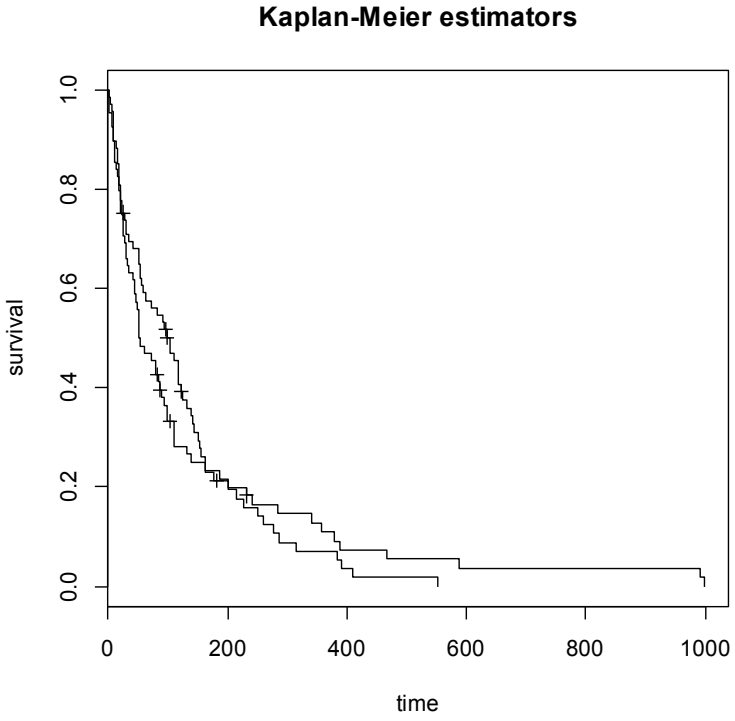
42	47	1	0.6797	0.0563	0.57782	0.800
……						
411	2	1	0.0177	0.0175	0.00256	0.123
553	1	1	0.0000	NaN	NA	NA

treat=2

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	68	2	0.9706	0.0205	0.93125	1.000
2	66	1	0.9559	0.0249	0.90830	1.000
7	65	2	0.9265	0.0317	0.86647	0.991
8	63	2	0.8971	0.0369	0.82766	0.972
13	61	1	0.8824	0.0391	0.80900	0.962
15	60	2	0.8529	0.0429	0.77278	0.941
……						
389	5	1	0.0732	0.0344	0.02912	0.184
467	4	1	0.0549	0.0303	0.01861	0.162
587	3	1	0.0366	0.0251	0.00953	0.140
991	2	1	0.0183	0.0180	0.00265	0.126
999	1	1	0.0000	NaN	NA	NA

Για το διάγραμμα των εκτιμητριών ΚΜ, με την ανάλυση επιβίωσης και το χρόνο δίνουμε την εντολή:

> plot(fit1, xlab="time", ylab="survival", main="Kaplan-Meier estimators")



Εξετάζουμε το τεστ log rank για τις καμπύλες επιβίωσης:

> survdiff(Surv(stime, status)~treat)

Call:

survdiff(formula = Surv(stime, status) ~ treat)

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
treat=1	69	64	64.5	0.00388	0.00823
treat=2	68	64	63.5	0.00394	0.00823

Chisq= 0 on 1 degrees of freedom, p= 0.928

Το Cox μοντέλο αναλογικού κινδύνου για όλες τις μεταβλητές προκύπτει ως εξής:

va.fitcox1 <- coxph(Surv(stime, status) ~ treat + age + Karn + diag.time +

+ cell+prior, data=VA)

> va.fitcox1

Call:
coxph(formula = Surv(stime, status) ~ treat + age + Karn + diag.time +
cell + prior, data = VA)

	coef	exp(coef)	se(coef)	z	p
treat2	2.95e-01	1.343	0.20755	1.4194	1.6e-01
age	-8.71e-03	0.991	0.00930	-0.9361	3.5e-01
Karn	-3.28e-02	0.968	0.00551	-5.9580	2.6e-09
diag.time	8.13e-05	1.000	0.00914	0.0089	9.9e-01
cell2	8.62e-01	2.367	0.27528	3.1297	1.7e-03
cell3	1.20e+00	3.307	0.30092	3.9747	7.0e-05
cell4	4.01e-01	1.494	0.28269	1.4196	1.6e-01
prior10	7.16e-02	1.074	0.23231	0.3082	7.6e-01

Likelihood ratio test=62.1 on 8 df, p=1.8e-10 n= 137, number of events= 128

Βρίσκουμε την τιμή του κριτηρίου AIC για το μοντέλο

> extractAIC(va.fitcox1)

[1] 8.0000 964.7942

Οι μεταβλητές Karn and Cell δείχνουν να είναι στατιστικά σημαντικές αλλά η μεταβλητή cell είναι κατηγορική οπότε ας δοκιμάσουμε την διαγραφή Single term deletion με την εντολή drop.

> drop1(va.fitcox1)

Single term deletions

Model:
Surv(stime, status) ~ treat + age + Karn + diag.time + cell +
prior

	Df	AIC
<none>		964.79

treat 1 964.81

age 1 963.65

Karn 1 997.78

diag.time 1 962.79

cell 3 977.63

prior 1 962.89

Η διαγραφή των Karn και Cell οδηγεί σε αύξηση του AIC ενώ η διαγραφή άλλων μεταβλητών δεν την αλλάζει πολύ άρα οι Karn και Cell είναι πιθανόν οι πιο σημαντικές μεταβλητές.

Ας δοκιμάσουμε τη Stepwise μέθοδο επιλογής μοντέλου

> step(va.fitcox1)

Start: AIC=964.79

Surv(stime, status) ~ treat + age + Karn + diag.time + cell +

prior

	Df	AIC
- diag.time	1	962.79
- prior	1	962.89
- age	1	963.65
<none>		964.79
- treat	1	964.81
- cell	3	977.63
- Karn	1	997.78

Step: AIC=962.79

Surv(stime, status) ~ treat + age + Karn + cell + prior

	Df	AIC
- prior	1	960.92
- age	1	961.67

<none> 962.79
- treat 1 962.84
- cell 3 975.67
- Karn 1 996.82

Step: AIC=960.92
Surv(stime, status) ~ treat + age + Karn + cell

      Df AIC
- age 1 959.83
<none> 960.92
- treat 1 961.09
- cell 3 973.76
- Karn 1 994.86

Step: AIC=959.83
Surv(stime, status) ~ treat + Karn + cell

      Df AIC
- treat 1 959.53
<none> 959.83
- cell 3 971.93
- Karn 1 993.04

Step: AIC=959.53
Surv(stime, status) ~ Karn + cell

     Df AIC
<none> 959.53
- cell 3 970.87
- Karn 1 992.05

Call:
coxph(formula = Surv(stime, status) ~ Karn + cell, data = VA)

	coef	exp(coef)	se(coef)	z	p
Karn	-0.0311	0.969	0.00518	-6.00	2.0e-09
cell2	0.7153	2.045	0.25269	2.83	4.6e-03
cell3	1.1577	3.183	0.29294	3.95	7.7e-05
cell4	0.3256	1.385	0.27668	1.18	2.4e-01

Likelihood ratio test=59.4 on 4 df, p=3.93e-12 n= 137, number of events= 128

Άρα και η μέθοδος stepwise επιλέγει τις μεταβλητές Karn and Cell.
Το μοντέλο γίνεται:

va.fitcox2 <- coxph(Surv(stime, status) ~ Karn + cell, data=VA)
> va.fitcox2
Call:
coxph(formula = Surv(stime, status) ~ Karn + cell, data = VA)

	coef	exp(coef)	se(coef)	z	p
Karn	-0.0311	0.969	0.00518	-6.00	2.0e-09
cell2	0.7153	2.045	0.25269	2.83	4.6e-03
cell3	1.1577	3.183	0.29294	3.95	7.7e-05
cell4	0.3256	1.385	0.27668	1.18	2.4e-01

Likelihood ratio test=59.4 on 4 df, p=3.93e-12 n= 137, number of events= 128
Συγκρίνουμε με anova τα δυο προαναφερθέντα μοντέλα:
anova(va.fitcox2, va.fitcox1)
Analysis of Deviance Table
Cox model: response is Surv(stime, status)

Model 1: ~ Karn + cell

Model 2: ~ treat + age + Karn + diag.time + cell + prior

  loglik Chisq Df P(>|Chi|)

1 -475.76

2 -474.40 2.7322 4   0.6036

> extractAIC(va.fitcox2)

[1]   4.0000 959.5264

> va.fitcox3 <- stepAIC(va.fitcox2, ~.^2)

Start: AIC=959.53

Surv(stime, status) ~ Karn + cell

           Df   AIC

<none>   959.53

+ Karn:cell 3 962.51

- cell   3 970.87

- Karn   1 992.05

> va.fitcox3$anova

Stepwise Model Path

Analysis of Deviance Table

Initial Model:

Surv(stime, status) ~ Karn + cell

Final Model:

Surv(stime, status) ~ Karn + cell

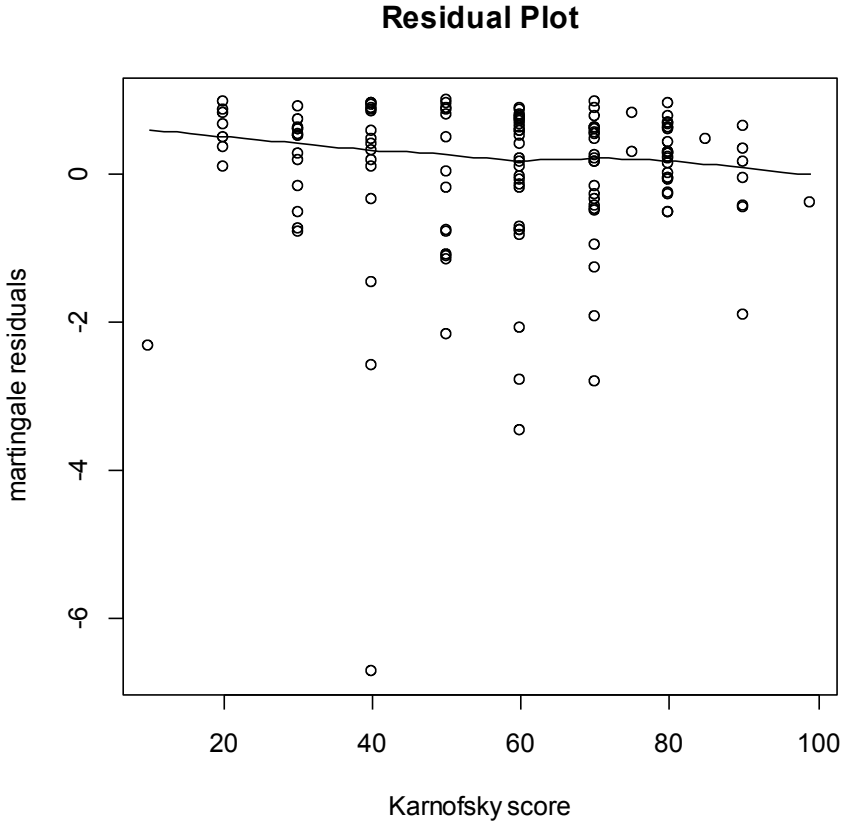
 Step Df Deviance Resid. Df Resid. Dev   AIC

1 133 951.5264 959.5264

Έλεγχος του μοντέλου με martingale residuals

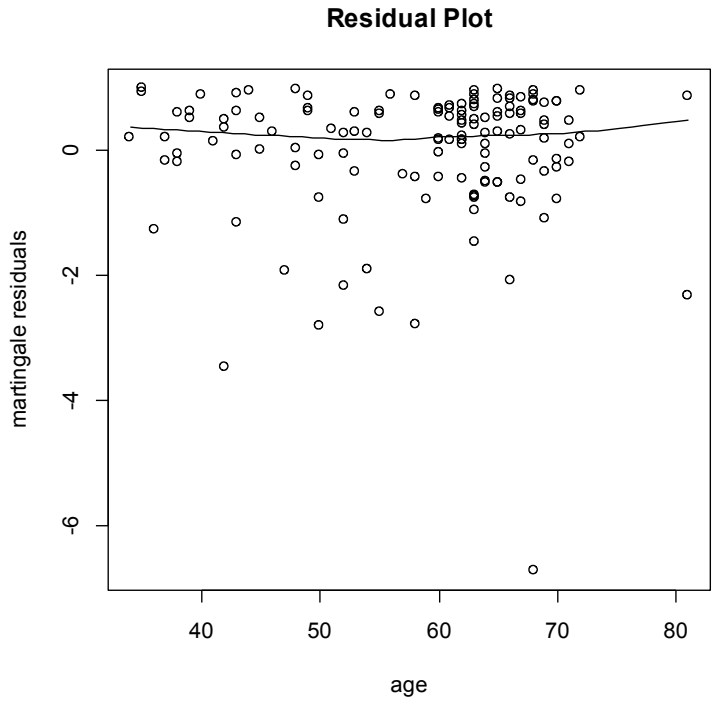
>scatter.smooth(VA$Karn, residuals(va.fitcox1), main="Residual Plot",

+ xlab="Karnofsky score", ylab="martingale residuals")



>scatter.smooth(VA$age, residuals(va.fitcox1), main="Residual Plot",

ylab="martingale residuals", xlab="age")



Φτιάχνουμε το μοντέλο Cox αναλογικού κινδύνου:

> coxfit2 <- coxph(Surv(stime) ~ age + Karn + diag.time + cell)

> summary(coxfit2)

Call:

coxph(formula = Surv(stime) ~ age + Karn + diag.time + cell)

n= 137, number of events= 137

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
age	-0.006505	0.993516	0.008785	-0.740	0.45904	
Karn	-0.029734	0.970704	0.005277	-5.635	1.75e-08	***
diag.time	0.001855	1.001856	0.008292	0.224	0.82300	
cell2	0.680321	1.974511	0.240232	2.832	0.00463	**
cell3	1.104426	3.017491	0.282743	3.906	9.38e-05	***
cell4	0.202646	1.224639	0.265202	0.764	0.44480	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

	exp(coef)	exp(-coef)	lower .95	upper .95
age	0.9935	1.0065	0.9766	1.0108
Karn	0.9707	1.0302	0.9607	0.9808
diag.time	1.0019	0.9981	0.9857	1.0183
cell2	1.9745	0.5065	1.2330	3.1619
cell3	3.0175	0.3314	1.7337	5.2519
cell4	1.2246	0.8166	0.7282	2.0594

Concordance= 0.721 (se = 0.029)

Rsquare= 0.34 (max possible= 1)

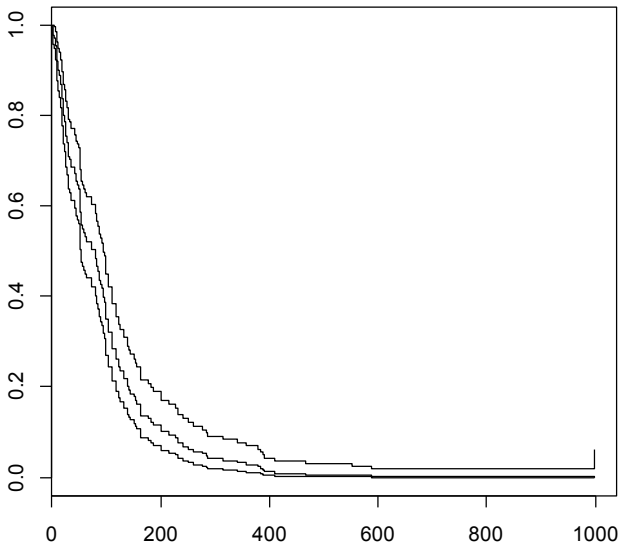
Likelihood ratio test= 57.01 on 6 df, p=1.82e-10

Wald test = 57.45 on 6 df, p=1.479e-10

Score (logrank) test = 60.72 on 6 df, p=3.219e-11

Το διάγραμμα για το cox που προκύπτει είναι το εξής:

>plot(survfit(coxfit2))



Μοντελοποιούμε τα δεδομένα χρησιμοποιώντας μοντέλα διαφορετικά και συγκρίνουμε τα αποτελέσματα για τις 2 στατιστικά σημαντικές μεταβλητές όπως αυτές ορίστηκαν νωρίτερα:

Αρχικά δοκιμάζουμε το cox proportional hazards

> va.fitcox <- coxph(Surv(stime, status) ~ Karn + cell, data=VA)

> summary(va.fitcox)

Call:

coxph(formula = Surv(stime, status) ~ Karn + cell, data = VA)

n= 137, number of events= 128

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
Karn	-0.031057	0.969421	0.005177	-5.999	1.99e-09	***
cell2	0.715334	2.044870	0.252686	2.831	0.00464	**
cell3	1.157733	3.182711	0.292937	3.952	7.74e-05	***
cell4	0.325645	1.384923	0.276680	1.177	0.23921	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

	exp(coef)	exp(-coef)	lower .95	upper .95
Karn	0.9694	1.0315	0.9596	0.9793
cell2	2.0449	0.4890	1.2462	3.3555
cell3	3.1827	0.3142	1.7925	5.6513
cell4	1.3849	0.7221	0.8052	2.3820

Concordance= 0.734 (se = 0.03)

Rsquare= 0.352 (max possible= 0.999)

Likelihood ratio test= 59.37 on 4 df, p=3.931e-12

Wald test = 61.26 on 4 df, p=1.577e-12

Score (logrank) test = 63.94 on 4 df, p=4.305e-13

Δοκιμάζουμε το Weibull (AFT μοντέλο)

> va.fitwei <- survreg(Surv(stime, status) ~ Karn + cell, data=VA, dist='weibull')

> summary(va.fitwei)

Call:

survreg(formula = Surv(stime, status) ~ Karn + cell, data = VA,

dist = "weibull")

	Value	Std. Error	z	p
(Intercept)	3.4806	0.34048	10.223	1.57e-24
Karn	0.0292	0.00462	6.310	2.78e-10
cell2	-0.7082	0.22609	-3.132	1.74e-03
cell3	-1.1085	0.25269	-4.387	1.15e-05
cell4	-0.3220	0.25002	-1.288	1.98e-01
Log(scale)	-0.0642	0.06596	-0.973	3.30e-01

Scale= 0.938

Weibull distribution

Loglik(model)= -716.5 Loglik(intercept only)= -748.1

Chisq= 63.15 on 4 degrees of freedom, p= 6.3e-13

Number of Newton-Raphson Iterations: 5

n= 137

Το lognormal δίνει:

> va.lnorm <- survreg(Surv(stime, status) ~ Karn + cell, data=VA, dist='lognormal')

> summary(va.lnorm)

Call:

survreg(formula = Surv(stime, status) ~ Karn + cell, data = VA,

dist = "lognormal")

```
             Value Std. Error      z        p
(Intercept)  2.2613     0.3430  6.592 4.33e-11
Karn         0.0374     0.0048  7.801 6.16e-15
cell2       -0.5334     0.2455 -2.173 2.98e-02
cell3       -0.6543     0.2805 -2.332 1.97e-02
cell4        0.1126     0.2808  0.401 6.88e-01
Log(scale)   0.0707     0.0626  1.130 2.59e-01

Scale= 1.07

Log Normal distribution
Loglik(model)= -716.2   Loglik(intercept only)= -749.5
       Chisq= 66.62 on 4 degrees of freedom, p= 1.2e-13
Number of Newton-Raphson Iterations: 4
n= 137
```

Δοκιμάζουμε το εκθετικό μοντέλο:

> va.expon <- survreg(Surv(stime, status) ~ Karn + cell, data=VA, dist='exponential')

> summary(va.expon)

```
Call:
survreg(formula = Surv(stime, status) ~ Karn + cell, data = VA,
   dist = "exponential")
             Value Std. Error     z        p
(Intercept)  3.4222    0.35463  9.65 4.92e-22
Karn         0.0297    0.00486  6.11 9.97e-10
cell2       -0.7102    0.24061 -2.95 3.16e-03
cell3       -1.0933    0.26863 -4.07 4.70e-05
```

cell4 -0.3113 0.26635 -1.17 2.43e-01

Scale fixed at 1

Exponential distribution
Loglik(model)= -717 Loglik(intercept only)= -751.2
Chisq= 68.5 on 4 degrees of freedom, p= 4.7e-14
Number of Newton-Raphson Iterations: 5
n= 137

Συγκρίνοντας τα αποτελέσματα βλέπουμε ότι για το Weibull μοντέλο η εκτιμώμενη τιμή scale είναι κοντά στο 1 άρα το εκθετικό μοντέλο που ούτως ή άλλως εξετάστηκε παραπάνω μπορεί να μοντελοποιήσει τα δεδομένα.
Το Weibull και το lognormal δίνουν διαφορετικές εκτιμήσεις αλλά σε γενικές γραμμές τα αποτελέσματα δεν διαφέρουν πολύ.

ΕΠΙΛΟΓΟΣ

Η Ανάλυση επιβίωσης χρησιμοποιείται για να εξερευνηθεί η εμφάνιση κάποιου γεγονότος όπως π.χ. ο θάνατος μετά από χρήση ενός φαρμάκου ή η ίαση της ασθένειας.
Στην εργασία αυτή παρουσιάστηκαν διάφορα κλασικά μοντέλα ανάλυσης επιβίωσης όπως το Kaplan-Meier, το Cox μοντέλο αναλογικού κινδύνου καθώς και παραμετρικά όπως το Weibull.
Οι μέθοδοι κλασικής ανάλυσης επιβίωσης έχουν προϋποθέσεις οι οποίες τα κάνουν «δύστροπα» στην λειτουργία τους αν κι εξακολουθούν να είναι αποτελεσματικά αν ξεπεραστούν τα εμπόδια των προϋποθέσεων αυτών.
Για την αντιμετώπιση των προβλημάτων που προκαλούν οι προϋποθέσεις για τη χρήση διαφόρων μεθόδων όπως π.χ. η αναλογικότητα για το μοντέλο Cox, μπορούν να χρησιμοποιηθούν τα νευρωνικά δίκτυα και να δώσουν προβλέψεις για τους χρόνους επιβίωσης και κινδύνου όπως και η Bayes ανάλυση επιβίωσης (Bayesian Proportional-Hazards Model).

Το νευρωνικό δίκτυο ANN μπορεί να χρησιμοποιηθεί για να προβλέψει άμεσα τους χρόνους επιβίωσης ή να προβλέψει την επιβίωση και τον κινδύνο. Μπορεί επίσης να χρησιμοποιηθεί για να επεκτείνει το μοντέλο PH του Cox.

Επιγγραματικά να αναφέρουμε ότι το κύριο πλεονέκτημα του νευρωνικού δικτύου είναι ότι μπορεί να χρησιμοποιήσουν δεδομένα χωρίς περιορισμούς σχετικά με τις ιδιότητες τους και θεωρείται ως ένα από τα πιο ευέλικτα μοντέλα και κατάλληλο για μη γραμμικά πολυμεταβλητά προβλήματα. Ωστόσο η τυχαιότητα των αρχικών τιμών που χρησιμοποιούνται για να εκπαιδευτεί το νευρικό δίκτυο καθιστά την μέθοδο ασταθή μερικές φορές κι έχει οδηγήσει σε αμφισβήτηση των αποτελεσμάτων της.

Θεωρητικά , η μέθοδος ANN είναι πιο αποτελεσματική στα πολύπλοκα δεδομένα με μεγάλο βαθμό αλληλεπίδρασης στις συμμεταβλητές και στις μεταβλητές που εξαρτώνται

από το χρόνο σε σχέση με τις παραδοσιακές μεθόδους παλινδρόμησης ή τις κλασικές μεθόδους ανάλυσης επιβίωσης.
Ωστόσο, δεν δείχνουν όλες οι μελέτες ότι οι μέθοδοι με νευρωνικά δίκτυα είναι ανώτερες από τις παραδοσιακές και θα είχε ενδιαφέρον μια έρευνα σύγκρισης μεταξύ των νευρωνικών δικτύων και της κλασικής ανάλυσης επιβίωσης που εξετάστηκε στην εργασία αυτή.

Όσον αφορά στη γλώσσα R, επιβεβαιώθηκε ότι έχει πολλές δυνατότητες στο θέμα των γραφημάτων και των εντολών της και είναι αναλογικά εύκολη στη χρήση. Για τα νευρωνικά δίκτυα στη γλώσσα R, μπορεί να χρησιμοποιηθεί η εντολή nnet για να εκπαιδεύσει ένα νευρωνικό δίκτυο Partial Logistic Artificial Neural Network (PLANN) και να χρησιμοποιηθεί στην ανάλυση επιβίωσης για την πρόβλεψη των χρόνων επιβίωσης και κινδύνου. Για τη μπεϋζιανή ανάλυση επιβίωσης αντίστοιχα μπορεί να χρησιμοποιηθεί π.χ. η εντολή survBayes η οποία δημιουργεί ένα proportional hazards model για δεξιά και διαστήματος λογοκριμμένες παρατηρήσεις.

.

ΠΑΡΑΡΤΗΜΑ ΣΤΗΝ R

Στο παράρτημα παρουσιάζονται συναρτήσεις κι εντολές στην R, που χρησιμοποιήθηκαν στο κύριο μέρος στην αγγλική γλώσσα και προκύπτουν από το εγχειρίδιο της R και τη βιβλιοθήκη: http://stat.ethz.ch/R-manual/R-patched/library/survival/html/

- coxph (from survival) fits Cox proportional hazards regression models.
- cox.zph (from survival) when applied to coxph model objects tests the proportional hazards assumption.
- plot.survfit Plot method for survfit
- strata (from survival) defines a stratum variable in a survival regression model. With coxph it leads to separate baseline hazards being assumed for the different strata. With survreg and dist=Weibull it causes a different Weibull shape parameter to be estimated in each stratum.
- Surv (from survival) is used to create a survival object for use with other survival package functions.
- survdiff (from survival) is carries out the log-rank test and its variations. Test Survival Curve Differences, performs tests for differences in lifetime distributions
- survfit (from survival) generates a Kaplan-Meier estimate of the survivor function for different groups using either raw data or a coxph model object. Compute a survival Curve for Censored Data, estimates (nonparametrically) the survival function
- survreg (from survival) estimates parametric regression models for survival data. Fits an Accelerated Failure Time Model

- survreg.control (from survival) is used to adjust the optimization settings of the survreg function.

Συνάρτηση coxph στην R

Fits a Cox proportional hazards regression model. Time dependent variables, time dependent strata, multiple events per subject, and other extensions are incorporated using the counting process formulation of Andersen and Gill.

Usage

```
coxph(formula, data=, weights, subset,
    na.action, init, control,
    ties=c("efron","breslow","exact"),
    singular.ok=TRUE, robust=FALSE,
    model=FALSE, x=FALSE, y=TRUE, tt, method, ...)
```

Arguments

formula
: a formula object, with the response on the left of a ~ operator, and the terms on the right. The response must be a survival object as returned by the Surv function.

data
: a data.frame in which to interpret the variables named in the formula, or in the subset and the weights argument.

weights
: vector of case weights. If weights is a vector of integers, then the estimated coefficients are equivalent to estimating the model from data with the individual cases replicated as many times as indicated by weights.

subset
: expression indicating which subset of the rows of data should be used in the

fit. All observations are included by default.

na.action
: a missing-data filter function. This is applied to the model.frame after any subset argument has been used. Default is options()\$na.action.

init
: vector of initial values of the iteration. Default initial value is zero for all variables.

control
: Object of class coxph.control specifying iteration limit and other control options. Default is coxph.control(...).

ties
: a character string specifying the method for tie handling. If there are no tied death times all the methods are equivalent. Nearly all Cox regression programs use the Breslow method by default, but not this one. The Efron approximation is used as the default here, it is more accurate when dealing with tied death times, and is as efficient computationally. The "exact partial likelihood" is equivalent to a conditional logistic model, and is appropriate when the times are a small set of discrete values. If there are a large number of ties and (start, stop) style survival data the computational time will be excessive.

singular.ok
: logical value indicating how to handle collinearity in the model matrix. If TRUE, the program will automatically skip over columns of the X matrix that are linear combinations of earlier columns. In this case the coefficients for such columns will be NA, and the variance matrix will contain zeros. For ancillary calculations, such as the linear predictor, the missing coefficients are treated as zeros.

robust
: this argument has been deprecated, use a cluster term in the model instead.

model
: logical value: if TRUE, the model frame is returned in component model.

X
: logical value: if TRUE, the x matrix is returned in component x.

Y
: logical value: if TRUE, the response vector is returned in component y.

tt
: optional list of time-transform functions.

method
: alternate name for the ties argument.

...
: Other arguments will be passed to coxph.control

The proportional hazards model is usually expressed in terms of a single survival time value for each person, with possible censoring. Andersen and Gill reformulated the same problem as a counting process; as time marches onward we observe the events for a subject, rather like watching a Geiger counter. The data for a subject is presented as multiple rows or "observations", each of which applies to an interval of observation (start, stop].

The routine internally scales and centers data to avoid overflow in the argument to the exponential function. These actions do not change the result, but lead to more numerical stability. However, arguments to offset are not scaled since there are situations where a large offset value is a purposefully used. Users should not use normally allow large numeric offset values.

Survreg

Fit a parametric survival regression model. These are location-scale models for an arbitrary transform of the time variable; the most common cases use a log transformation, leading to accelerated failure time models.

Usage

```
survreg(formula, data, weights, subset,
    na.action, dist="weibull", init=NULL, scale=0,
    control,parms=NULL,model=FALSE, x=FALSE,
    y=TRUE, robust=FALSE, score=FALSE, ...)
```

Arguments

formula
: a formula expression as for other regression models. The response is usually a survival object as returned by the Surv function. See the documentation for Surv, lm and formula for details.

data
: a data frame in which to interpret the variables named in the formula, weights or the subset arguments.

weights
: optional vector of case weights

subset
: subset of the observations to be used in the fit

na.action
: a missing-data filter function, applied to the model.frame, after any subset argument has been used. Default is options()\$na.action.

dist
: assumed distribution for y variable. If the argument is a character string, then

it is assumed to name an element from survreg.distributions. These include "weibull", "exponential", "gaussian", "logistic","lognormal" and "loglogistic". Otherwise, it is assumed to be a user defined list conforming to the format described in survreg.distributions.

parms
: a list of fixed parameters. For the t-distribution for instance this is the degrees of freedom; most of the distributions have no parameters.

init
: optional vector of initial values for the parameters.

scale
: optional fixed value for the scale. If set to <=0 then the scale is estimated.

control
: a list of control values, in the format produced by survreg.control. The default value is survreg.control()

model,x,y
: flags to control what is returned. If any of these is true, then the model frame, the model matrix, and/or the vector of response times will be returned as components of the final result, with the same names as the flag arguments.

score
: return the score vector. (This is expected to be zero upon successful convergence.)

robust
: Use robust 'sandwich' standard errors, based on independence of individuals if there is no cluster() term in the formula, based on independence of clusters if there is.

...
: other arguments which will be passed to survreg.control.

Residuals.coxph

Calculates martingale, deviance, score or Schoenfeld residuals for a Cox proportional hazards model.

Usage

```
## S3 method for class 'coxph'
residuals(object,
    type=c("martingale", "deviance", "score", "schoenfeld",
           "dfbeta", "dfbetas", "scaledsch","partial"),
    collapse=FALSE, weighted=FALSE, ...)
## S3 method for class 'coxph.null'
residuals(object,
    type=c("martingale", "deviance","score","schoenfeld"),
    collapse=FALSE, weighted=FALSE, ...)
```

Arguments

object	an object inheriting from class coxph, representing a fitted Cox regression model. Typically this is the output from the coxph function.
type	character string indicating the type of residual desired. Possible values are "martingale", "deviance", "score", "schoenfeld", "dfbeta"', "dfbetas", and "scaledsch". Only enough of the string to determine a unique match is required.
collapse	vector indicating which rows to collapse (sum) over. In time-dependent models more than one row data can pertain to a single individual. If there were 4 individuals represented by 3, 1, 2 and 4 rows of data respectively, then collapse=c(1,1,1, 2, 3,3, 4,4,4,4) could be used to obtain per subject rather than per observation residuals.
weighted	if TRUE and the model was fit with case weights, then the weighted residuals are returned.

... other unused arguments

For martingale and deviance residuals, the returned object is a vector with one element for each subject (without collapse). For score residuals it is a matrix with one row per subject and one column per variable. The row order will match the input data for the original fit. For Schoenfeld residuals, the returned object is a matrix with one row for each event and one column per variable. The rows are ordered by time within strata, and an attribute strata is attached that contains the number of observations in each strata. The scaled Schoenfeld residuals are used in the cox.zph function.

The score residuals are each individual's contribution to the score vector. Two transformations of this are often more useful: dfbeta is the approximate change in the coefficient vector if that observation were dropped, and dfbetas is the approximate change in the coefficients, scaled by the standard error for the coefficients.

dropterm

Try fitting all models that differ from the current model by dropping a single term, maintaining marginality.

This function is generic; there exist methods for classes lm and glm and the default method will work for many other classes.

Usage

```
dropterm (object, ...)

## Default S3 method:
dropterm(object, scope, scale = 0, test = c("none", "Chisq"),
    k = 2, sorted = FALSE, trace = FALSE, ...)

## S3 method for class 'lm'
dropterm(object, scope, scale = 0, test = c("none", "Chisq", "F"),
    k = 2, sorted = FALSE, ...)
```

S3 method for class 'glm'

dropterm(object, scope, scale = 0, test = c("none", "Chisq", "F"),

k = 2, sorted = FALSE, trace = FALSE, ...)

Arguments

object	A object fitted by some model-fitting function.
scope	a formula giving terms which might be dropped. By default, the model formula. Only terms that can be dropped and maintain marginality are actually tried.
scale	used in the definition of the AIC statistic for selecting the models, currently only for lm, aov and glm models. Specifying scale asserts that the residual standard error or dispersion is known.
Test	should the results include a test statistic relative to the original model? The F test is only appropriate for lm and aov models, and perhaps for some over-dispersed glm models. The Chisq test can be an exact test (lm models with known scale) or a likelihood-ratio test depending on the method.
K	the multiple of the number of degrees of freedom used for the penalty. Only k = 2 gives the genuine AIC: k = log(n) is sometimes referred to as BIC or SBC.
sorted	should the results be sorted on the value of AIC?
trace	if TRUE additional information may be given on the fits as they are tried.
...	arguments passed to or from other methods.

The definition of AIC is only up to an additive constant: when appropriate (lm models with specified scale) the constant is taken to be that used in Mallows' Cp statistic and the results are labelled accordingly.

Value

A table of class "anova" containing at least columns for the change in degrees of freedom and AIC (or Cp) for the models. Some methods will give further information, for example sums of squares, deviances, log-likelihoods and test statistics.

AIC- Akaike's Information Criterion

Generic function calculating Akaike's 'An Information Criterion' for one or several fitted model objects for which a log-likelihood value can be obtained, according to the formula *-2*log-likelihood + k*npar*, where *npar* represents the number of parameters in the fitted model, and $k = 2$ for the usual AIC.

Usage

```
AIC(object, ..., k = 2)
```

Arguments

object	a fitted model object for which there exists a logLik method to extract the corresponding log-likelihood, or an object inheriting from class logLik.
...	optionally more fitted model objects.
K	numeric, the *penalty* per parameter to be used; the default k = 2 is the classical AIC.

However methods should be defined for the log-likelihood function logLik rather than these functions: the action of their default methods is to call logLik on all the supplied objects and assemble the results.

When comparing fitted objects, the smaller the AIC, the better the fit.

The log-likelihood and hence the AIC/BIC is only defined up to an additive constant. Different constants have conventionally be used for different purposes and so extractAIC and AIC may give different values. Particular care is needed when comparing fits of different classes (with, for example, a comparison of a Poisson and gamma GLM being meaningless since one has a discrete response, the other continuous).

step

Select a formula-based model by AIC.

Usage

```
step(object, scope, scale = 0,
    direction = c("both", "backward", "forward"),
    trace = 1, keep = NULL, steps = 1000, k = 2, ...)
```

Arguments

object	an object representing a model of an appropriate class (mainly "lm" and "glm"). This is used as the initial model in the stepwise search.
scope	defines the range of models examined in the stepwise search. This should be either a single formula, or a list containing components upper and lower, both formulae. See the details for how to specify the formulae and how they are used.
Scale	used in the definition of the AIC statistic for selecting the models, currently only for lm, aov and glm models. The default value, 0, indicates the scale should be estimated: see extractAIC.
direction	the mode of stepwise search, can be one of "both", "backward", or "forward", with a default of "both". If the scope argument is missing the default for direction is "backward".
Trace	if positive, information is printed during the running of step. Larger values may give more detailed information.
Keep	a filter function whose input is a fitted model object and the associated AIC statistic, and whose output is arbitrary. Typically keep will select a subset of the components of the object and return them. The default is not to keep anything.
Steps	the maximum number of steps to be considered. The default is 1000 (essentially as many as required). It is typically used to stop the process early.
K	the multiple of the number of degrees of freedom used for the penalty. Only k = 2 gives the genuine AIC: k = log(n) is sometimes referred to as BIC or SBC.
...	any additional arguments to extractAIC.

step uses add1 and drop1 repeatedly; it will work for any method for which they work, and that is determined by having a valid method for extractAIC. When the additive constant can be chosen so that AIC is equal to Mallows' *Cp*, this is done and the tables are labelled appropriately.

The set of models searched is determined by the scope argument. The right-hand-side of its lower component is always included in the model, and right-hand-side of the model is included in the upper component. If scope is a single formula, it specifies the upper component, and the lower model is empty. If scope is missing, the initial model is used as the upper model.

Models specified by scope can be templates to update object as used by update.formula. So using . in a scope formula means 'what is already there', with .^2 indicating all interactions of existing terms.

There is a potential problem in using glm fits with a variable scale, as in that case the deviance is not simply related to the maximized log-likelihood. The "glm" method for function extractAIC makes the appropriate adjustment for a gaussian family, but may need to be amended for other cases. (The binomial and poisson families have fixed scale by default and do not correspond to a particular maximum-likelihood problem for variable scale.)

Value

the stepwise-selected model is returned, with up to two additional components. There is an "anova" component corresponding to the steps taken in the search, as well as a "keep" component if the keep= argument was supplied in the call. The "Resid. Dev" column of the analysis of deviance table refers to a constant minus twice the maximized log likelihood: it will be a deviance only in cases where a saturated model is well-defined (thus excluding lm, aov and survreg fits, for example).

stepAIC

Performs stepwise model selection by AIC.

Usage

```
stepAIC(object, scope, scale = 0,
    direction = c("both", "backward", "forward"),
    trace = 1, keep = NULL, steps = 1000, use.start = FALSE,
    k = 2, ...)
```

Arguments

object	an object representing a model of an appropriate class. This is used as the initial model in the stepwise search.
scope	defines the range of models examined in the stepwise search. This should be either a single formula, or a list containing components upper and lower, both formulae. See the details for how to specify the formulae and how they are used.
Scale	used in the definition of the AIC statistic for selecting the models, currently only for lm and aov models (see extractAIC for details).
direction	the mode of stepwise search, can be one of "both", "backward", or "forward", with a default of "both". If the scope argument is missing the default for direction is "backward".
Trace	if positive, information is printed during the running of stepAIC. Larger values may give more information on the fitting process.
Keep	a filter function whose input is a fitted model object and the associated AIC statistic, and whose output is arbitrary. Typically keep will select a subset of the components of the object and return them. The default is not to keep anything.
Steps	the maximum number of steps to be considered. The default is 1000 (essentially as many as required). It is typically used to stop the process early.
use.start	if true the updated fits are done starting at the linear predictor for the currently

selected model. This may speed up the iterative calculations for glm (and other fits), but it can also slow them down. **Not used** in R.

K — the multiple of the number of degrees of freedom used for the penalty. Only k = 2 gives the genuine AIC: k = log(n) is sometimes referred to as BIC or SBC.

... — any additional arguments to extractAIC. (None are currently used.)

The set of models searched is determined by the scope argument. The right-hand-side of its lower component is always included in the model, and right-hand-side of the model is included in the upper component. If scope is a single formula, it specifies the upper component, and the lower model is empty. If scope is missing, the initial model is used as the upper model.

Models specified by scope can be templates to update object as used by update.formula.

There is a potential problem in using glm fits with a variable scale, as in that case the deviance is not simply related to the maximized log-likelihood. The glm method for extractAIC makes the appropriate adjustment for a gaussian family, but may need to be amended for other cases. (The binomial and poisson families have fixed scale by default and do not correspond to a particular maximum-likelihood problem for variable scale.)

Where a conventional deviance exists (e.g. for lm, aov and glm fits) this is quoted in the analysis of variance table: it is the *unscaled* deviance.

Value

the stepwise-selected model is returned, with up to two additional components. There is an "anova" component corresponding to the steps taken in the search, as well as a "keep" component if the keep= argument was supplied in the call. The "Resid. Dev" column of the analysis of deviance table refers to a constant minus twice the maximized log likelihood: it will be a deviance only in cases where a saturated model is well-defined (thus excluding lm, aov and survreg fits, for example).

anova.coxph

Compute an analysis of deviance table for one or more Cox model fits.

Usage

```
## S3 method for class 'coxph'
anova(object, ...,  test = 'Chisq')
```

Arguments

object	An object of class coxph
...	Further coxph objects
test	a character string. The appropriate test is a chisquare, all other choices result in no test being done.

Details

Specifying a single object gives a sequential analysis of deviance table for that fit. That is, the reductions in the model log-likelihood as each term of the formula is added in turn are given in as the rows of a table, plus the log-likelihoods themselves. A robust variance estimate is normally used in situations where the model may be mis-specified, e.g., multiple events per subject. In this case a comparison of partial-likelihood values does not make sense, and anova will refuse to print results.

If more than one object is specified, the table has a row for the degrees of freedom and loglikelihood for each model. For all but the first model, the change in degrees of freedom and loglik is also given. (This only make statistical sense if the models are nested.) It is conventional to list the models from smallest to largest, but this is up to the user.

The table will optionally contain test statistics (and P values) comparing the reduction in loglik for each row.

Value

An object of class "anova" inheriting from class "data.frame".

Βοήθεια στην R

help.search("topic") or "topic" (depends on the installed packages)

RSiteSearch("topic") (requires internet connection)

help() or invoke the on-line help file for the specified function

checking the FAQ

ΒΙΒΛΙΟΓΡΑΦΙΑ-ΔΙΑΔΙΚΤΥΑΚΕΣ ΠΗΓΕΣ

ΙΣΤΟΣΕΛΙΔΕΣ

http://www.math.wustl.edu/~jmding/math434/lec3_ho.pdf

http://stat.ethz.ch/R-manual/R-patched/library/survival/html/ovarian.html

http://stat.ethz.ch/R-manual/R-devel/library/MASS/html/VA.html

http://cran.r-project.org/web/packages/ISwR/ISwR.pdf

http://cran.r-project.org/web/packages/survival/survival.pdf

http://www.uni-kiel.de/psychologie/rexrepos/posts/survivalCoxPH.html

http://rwiki.sciviews.org/doku.php

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

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

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

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

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

http://www.inside-r.org/r-doc

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

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

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

http://cran.r-project.org/web/views/

http://www.r4stats.com/

https://stat.ethz.ch/mailman/listinfo/r-help

http://www.rseek.org/

http://sundoc.bibliothek.uni-halle.de/habil-online/07/07H056/t3.pdf

http://people.umass.edu/biep640w/pdf/6.%20%20Survival%20Analysis%202011.pdf

http://isites.harvard.edu/fs/docs/icb.topic79671.files/survival-lec2.pdf

http://www.nickfieller.staff.shef.ac.uk/sheff-only/pas6012-pas361.html

http://nickfieller.staff.shef.ac.uk/tampere/

http://www.ics.uci.edu/~vqnguyen/stat255/

http://www.stanford.edu/~kcobb

http://statweb.stanford.edu/~jtaylo/courses/stats262/spring.2004/

http://statistics.ats.ucla.edu/stat/r/examples/asa/default.htm

http://www.itl.nist.gov/div898/handbook/eda/section3/eda3669.htm

http://www.itl.nist.gov/div898/handbook/eda/section3/eda3669.htm

http://en.wikipedia.org/wiki/Log-logistic_distribution

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

http://www.math.wustl.edu/~jmding/math434/h434.html

https://openaccess.leidenuniv.nl/bitstream/handle/1887/11456/01.pdf?sequence=6

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

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

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

http://www.unipr.it/arpa/facvet/annali/2008/01%2017_42.pdf

ΕΛΛΗΝΙΚΗ

[1] ΦΑΡΜΑΚΗΣ Ν (2001) *«ΣΤΑΤΙΣΤΙΚΗ, Περιληπτική Θεωρία-Ασκήσεις»*, Εκδόσεις Α & Π Χριστοδουλίδη, Θεσσαλονίκη.

[2] ΦΑΡΜΑΚΗΣ Ν (2009) *«ΕΙΣΑΓΩΓΗ στη ΔΕΙΓΜΑΤΟΛΗΨΙΑ»*, Εκδόσεις Α & Π Χριστοδουλίδη, Θεσσαλονίκη.

[3] ΝΤΖΟΥΦΡΑΣ Ι., ΠΕΡΠΕΡΟΓΛΟΥ Α. (2009), «Εισαγωγή στη Βιοστατιστική και την Επιδημιολογία», Πανεπιστήμιο Αιγαίου

[4] ΠΑΥΛΟΥ Ε.(2006) «Το μοντέλο αναλογικού κινδύνου του Cox στην Ανάλυση Επιβίωσης», Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

[5] ΧΑΡΑΛΑΜΠΟΥΣ Χ., ΦΩΚΙΑΝΟΣ Κ. (2010) Εισαγωγή στην R, Πρόχειρες Σημειώσεις, Πανεπιστήμιο Κύπρου

ΞΕΝΟΓΛΩΣΣΗ

[1] FARMAKIS N (2003) “Estimation of Coefficient of Variation: Scaling of Symmetric Coninuous Distributions”, *Statistics in Transition*, Vol. 6 (1), pp 83-96.

[2] D. MACHIN, M.CAMPBELL, S. WALTERS (2007) “Medical Statistics”, John Wiley & Sons Ltd

[3] YANYING YANG (2009-2010), Neural Network Survival Analysis, University of Gent

[4] M. AKRAM, M. AMAN ULLAH AND R. TAJ “Survival Analysis of Cancer Patients using Parametric and Non-Parametric Approaches”, Bahauddin Zakariya University Multan, Pakistan

[5] LARRY WINNER (2004), “Introduction to Biostatistics”, University of Florida

[6] KLEINBAUM D. KLEIN M. (2012) “Survival Analysis A self-learning text”, 3rd edition, Editions Springer

[7] KIRKWOOD B. STERNE J. “Essential Medical Statistics”, 2nd edition, Blackwell Publishing

[8] CHONGSUVIVATWONG V. “Analysis of epidemiological data using R and Epicalc”, Prince of Songkla University Thailand

[9] VAN BELLE G., FISHER L.D, HEAGERTY P., LUMLEY T. (2004) “Biostatistics A Methodology for the Health Sciences, Wiley Interscience, John Wiley & Sons

[10] DALGAARD P. (2002), “Introductory Statistics with R”, Springer