



ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ



Γεραντίδου Μαρία

Μέθοδοι ανάλυσης δεδομένων στην ανάλυση επιβίωσης και τη
βιοστατιστική

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΑΤΡΙΒΗ

Ιωάννινα, 2025



UNIVERSITY OF IOANNINA
Department of Mathematics



Gerantidou Maria

Statistical techniques for data analysis in survival analysis
and biostatistics

Master's Thesis

Ioannina, 2025

Αφιερώνεται στους γονείς μου.

Η παρούσα Μεταπτυχιακή Διατριβή εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική και Επιχειρησιακή Έρευνα που απονέμει το Τμήμα Μαθηματικών του Πανεπιστημίου Ιωαννίνων.

Εγκρίθηκε την 25/06/2025 από την εξεταστική επιτροπή:

Ονοματεπώνυμο	Βαθμίδα
Δημήτριος Μπάγκαβος	Αναπληρωτής Καθηγητής (Επιβλέπων)
Απόστολος Μπατσίδης	Αναπληρωτής Καθηγητής
Σταύρος Νικολακόπουλος	Επίκουρος Καθηγητής

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ

“Δηλώνω υπεύθυνα ότι η παρούσα διατριβή εκπονήθηκε κάτω από τους διεθνείς ηθικούς και ακαδημαϊκούς κανόνες δεοντολογίας και προστασίας της πνευματικής ιδιοκτησίας. Σύμφωνα με τους κανόνες αυτούς, δεν έχω προβεί σε ιδιοποίηση ξένου επιστημονικού έργου και έχω πλήρως αναφέρει τις πηγές που χρησιμοποίησα στην εργασία αυτή.”

Γεραντίδου Μαρία

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα διατριβή έχει πραγματοποιηθεί στο Τμήμα Μαθηματικών του Πανεπιστημίου Ιωαννίνων. Με την ολοκλήρωση της παρούσας διατριβής θα ήθελα να εκφράσω τις ευχαριστίες μου σε όσους συνέβαλαν στην πραγματοποίησή της.

Ευχαριστώ θερμά τον επιβλέποντα Καθηγητή μου, κ. Δημήτριο Μπάγκαβο, για τη βοήθεια, την καθοδήγηση, τις πολύτιμες συμβουλές, την ενθάρρυνση και τη συμπαράσταση που μου προσέφερε καθ' όλη τη διάρκεια της εκπόνησης της εργασίας.

Επίσης, θα ήθελα να ευχαριστήσω τον κ. Απόστολο Μπατσίδη και τον κ. Σταύρο Νικολακόπουλο, μέλη της τριμελούς εξεταστικής επιτροπής, για την πολύτιμη συμβολή τους στη διατριβή. Ευχαριστώ ακόμη όλους τους διδάσκοντες του Τμήματος Μαθηματικών για τις γνώσεις που μου προσέφεραν σε προπτυχιακό και μεταπτυχιακό επίπεδο, καθώς και για τη συνεργασία που είχαμε κατά τη διάρκεια των σπουδών μου.

Ευχαριστώ από καρδιάς την οικογένειά μου για την απέραντη υποστήριξη, την αγάπη και την κατανόηση όλα αυτά τα χρόνια. Τέλος, ευχαριστώ τους φίλους μου που πίστεψαν σε εμένα.

Ιωάννινα 2025,
Γεραντίδου Μαρία

ΠΕΡΙΛΗΨΗ

Με σκοπό τη μοντελοποίηση του χρόνου μέχρι να συμβεί ένα γεγονός, το οποίο ορίζει και το κύριο ενδιαφέρον της εκάστοτε μελέτης, αναπτύσσονται τεχνικές που χρησιμοποιούνται ευρέως στην ανάλυση επιβίωσης και τη βιοστατιστική για δεδομένα χρόνου ζωής. Στο Κεφάλαιο 1 παρουσιάζονται βασικές έννοιες για την ανάλυση επιβίωσης, όπως είναι η συνάρτηση επιβίωσης, η συνάρτηση κινδύνου και η λογοκρισία, καθώς και τα δεδομένα που θα χρησιμοποιηθούν στην παρούσα διατριβή. Στο Κεφάλαιο 2 αναπτύσσεται αναλυτικά το μοντέλο αναλογικών κινδύνων, παρατίθεται η έννοια της μεθόδου μερικής πιθανοφάνειας, βασικές στατιστικές ιδιότητες των εκτιμητών των παραμέτρων του μοντέλου και η έννοια του τεστ πηλίκου πιθανοφάνειών (LRT). Παρουσιάζεται η εκτίμηση Kaplan-Meier της συνάρτησης επιβίωσης και το Log-rank τεστ για τη σύγκριση συναρτήσεων επιβίωσης μεταξύ ομάδων. Ακόμη, δίνεται έμφαση στον γραφικό έλεγχο της υπόθεσης της αναλογικότητας των κινδύνων και παρουσιάζεται ένας τρόπος επίλυσης της παραβίασης της υπόθεσης της αναλογικότητας κινδύνων, η στρωματοποίηση. Παρουσιάζεται η έννοια του Hazard Ratio (HR), το κριτήριο AIC (Akaike's information criterion) για την επιλογή μεταβλητών, ο έλεγχος καταλληλότητας του μοντέλου μέσω υπολοίπων, καθώς και παραδείγματα υλοποιημένα στην R. Στο Κεφάλαιο 3 παρουσιάζεται η επέκταση του μοντέλου του Cox, που εφαρμόζεται στην περίπτωση που λαμβάνονται υπόψη χρονικά εξαρτημένες μεταβλητές, δίνεται η ερμηνεία του Hazard Ratio (HR) για αυτήν την περίπτωση και κάποια παραδείγματα. Στο Κεφάλαιο 4 ορίζονται τα μοντέλα επιταχυνόμενου χρόνου αποτυχίας (Accelerated failure time, A.F.T.), δίνοντας έμφαση στο μοντέλο A.F.T. με κατανομή Weibull. Εξετάζεται αν το μοντέλο αναλογικών κινδύνων και το A.F.T. μοντέλο μπορούν να ταυτίζονται, αναφέρονται κάποιοι τρόποι για την αναζήτηση του κατάλληλου μοντέλου και την αξιολόγηση του και τέλος δίνεται ένα παράδειγμα στο οποίο παρουσιάζεται η διαδικασία που ακολουθείται για την επιλογή μεταξύ μοντέλου αναλογικών κινδύνων και μοντέλου επιταχυνόμενου χρόνου αποτυχίας.

ABSTRACT

In order to analyse the time until an event occurs –which defines the main focus of the study– techniques that are widely used in survival analysis and biostatistics for time-to-event data are presented. Chapter 1 introduces fundamental concepts of survival analysis, such as the survival function, the hazard function and censoring, along with the data that will be used in this thesis. Chapter 2 provides a detailed discussion of the proportional hazards model, introducing the concept of the partial likelihood method, key statistical properties of the model’s parameter estimators, and the Likelihood Ratio Test (LRT). The Kaplan–Meier estimator of the survival function is presented, along with the log-rank test for comparing survival curves between groups. Significant focus is devoted to the graphical assessment of the proportional hazards assumption, and a method to address violations of this assumption—stratification—is presented. The chapter also introduces the concept of the Hazard Ratio (HR), the Akaike Information Criterion (AIC) for variable selection, model adequacy assessment using residuals, and includes implementation examples in R. Chapter 3 presents the extension of the Cox model for situations involving time-dependent covariates. It discusses the interpretation of the Hazard Ratio (HR) in this context and provides relevant examples. Chapter 4 introduces Accelerated Failure Time (A.F.T.) models, with particular emphasis on the Weibull A.F.T. model. It investigates whether the proportional hazards model and the A.F.T. model can coincide, outlines methods for selecting the appropriate model and assessing its performance, and provides an example that illustrates the selection process between the proportional hazards model and the accelerated failure time model.

ΠΕΡΙΕΧΟΜΕΝΑ

Περιεχόμενα	1
1 Εισαγωγή	3
1.1 Είδη λογοχρίσας και τύποι μεταβλητών	5
1.2 Σύνολα πραγματικών δεδομένων	7
2 Το μοντέλο αναλογικών κινδύνων του Cox	13
2.1 Η μέθοδος μερικής πιθανοφάνειας	14
2.2 Ιδιότητες εκτιμώμενων παραμέτρων	18
2.3 Τεστ Πηλίκου Πιθανοφανειών (LRT)	26
2.4 Εφαρμογή	32
2.5 Kaplan-Meier εκτίμηση της συνάρτησης επιβίωσης	34
2.6 Log-rank τεστ	37
2.7 Υπόθεση αναλογικότητας κινδύνων	42
2.8 Στρωματοποίηση	58
2.9 Το μοντέλο αναλογικών κινδύνων του Cox στα CNS δεδομένα . .	64
2.10 Αναλογία κινδύνου (HR)	75
2.11 Η διαδικασία AIC για επιλογή μεταβλητών	77
2.12 Τύποι υπολοίπων και έλεγχοι	87
3 Επέκταση του μοντέλου του Cox	103
3.1 Μεταβλητές που εξαρτώνται από τον χρόνο	103
3.2 Ορισμός της επέκτασης του μοντέλου του Cox	105

3.3	Η αναλογία κινδύνου για την επέκταση του μοντέλου του Cox . .	107
3.4	Εφαρμογή	109
4	Μοντέλα επιταχυνόμενου χρόνου αποτυχίας	119
4.1	Ορισμός των A.F.T. μοντέλων	119
4.2	Επιλογή συναρτήσεων των συμμεταβλητών	120
4.3	Ερμηνεία των συντελεστών παλινδρόμησης	125
4.4	Shape-scale οικογένειες κατανομών	129
4.5	Μοντέλο A.F.T. με κατανομή Weibull	130
4.5.1	Μοντέλα παλινδρόμησης με αναλογικότητα κινδύνων . . .	130
4.5.2	A.F.T. μοντέλο με παραμέτρους που εξαρτώνται από το διάνυσμα των συμμεταβλητών	132
4.5.3	Log-location-scale μοντέλα επιταχυνόμενου χρόνου απο- τυχίας	133
4.6	Αναζήτηση και αξιολόγηση μοντέλου	134
4.6.1	Αναζήτηση κατάλληλου μοντέλου	134
4.6.2	Αξιολόγηση του προσαρμοσμένου μοντέλου	140
4.6.3	Εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων	143
4.6.4	Επιλογή μεταξύ μοντέλου αναλογικών κινδύνων και μο- ντέλου επιταχυνόμενου χρόνου αποτυχίας	145
	Βιβλιογραφία	153

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

Η ανάλυση επιβίωσης ασχολείται με τη μοντελοποίηση του χρόνου μέχρι να συμβεί ένα γεγονός, το οποίο ορίζει και το κύριο ενδιαφέρον της εκάστοτε μελέτης. Για παράδειγμα, μπορεί να μας ενδιαφέρει ο χρόνος μέχρι να χαλάσει ένα ηλεκτρικό εξάρτημα, ο χρόνος μέχρι την πρώτη επανεμφάνιση μιας ασθένειας μετά από μια θεραπεία ή η διάρκεια επιβίωσης ενός ασθενούς. Ωστόσο, το γεγονός που μας ενδιαφέρει και είναι υπό μελέτη δεν αφορά απαραίτητα πάντα κάποιο θάνατο ή αποτυχία. Έτσι, θα μπορούσε να μας ενδιαφέρει ο χρόνος μέχρι την εκμάθηση μίας καινούριας δεξιότητας, η διάρκεια μίας απεργίας ή ο χρόνος μέχρι την προαγωγή ενός υπαλλήλου. Η ανάλυση επιβίωσης είναι χρήσιμη σε πολλά επιστημονικά πεδία, όπως είναι η βιοϊατρική, η μηχανική, τα οικονομικά ή οι κοινωνικές επιστήμες.

Ο χρόνος ζωής μέχρι να συμβεί το υπό μελέτη γεγονός συνήθως αναπαρίσταται με μία τυχαία μεταβλητή, έστω T . Η T είναι μια μη αρνητική και συνεχής τυχαία μεταβλητή, με πεδίο ορισμού ένα υποσύνολο, έστω M , του $[0, +\infty)$. Στο πλαίσιο αυτής της διατριβής θεωρούμε ότι η T έχει συνάρτηση πυκνότητας πιθανότητας (σ.π.π.) $f(t)$ και αθροιστική συνάρτηση κατανομής (α.σ.κ.) $F(t)$. Ισχύει ότι $F(t) = P(T \leq t) = \int_0^t f(x)dx$. Συναρτήσεις που ποσοτικοποιούν την πιθανότητα να συμβεί το υπό μελέτη γεγονός είναι η συνάρτηση επιβίωσης (survival function) και η συνάρτηση κινδύνου ή συνάρτηση ρίσκου (hazard rate function). Οι συναρτήσεις αυτές παρουσιάζονται αναλυτικά παρακάτω.

Η συνάρτηση επιβίωσης εκφράζει την πιθανότητα ένα άτομο να επιβιώσει πέραν της χρονικής στιγμής t . Ορίζεται ως:

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(x)dx, \quad \forall t \in M.$$

Οπότε

$$f(t) = -\frac{dS(t)}{dt}.$$

Κεφάλαιο 1

Η συνάρτηση επιβίωσης ικανοποιεί τις εξής ιδιότητες:

- Είναι μονότονη, μη αύξουσα συνάρτηση για κάθε $t \in M$.
- $S(0) = 1$.
- $S(\infty) = \lim_{t \rightarrow \infty} S(t) = 0$.

Η συνάρτηση κινδύνου εκφράζει τον ρυθμό με τον οποίο μεταβάλλεται η πιθανότητα να συμβεί το γεγονός που μελετάμε και ορίζεται ως εξής:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t < T < t + \Delta t | T > t)}{\Delta t}.$$

Επειδή η T είναι συνεχής τυχαία μεταβλητή

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d \ln[S(t)]}{dt}.$$

Η $h(t)\Delta t$ μπορεί να θεωρηθεί ως η προσεγγιστική πιθανότητα να συμβεί το γεγονός που μας ενδιαφέρει την αμέσως επόμενη χρονική στιγμή, δεδομένου ότι δεν έχει συμβεί μέχρι τώρα. Η συνάρτηση κινδύνου είναι μη αρνητική, δηλαδή $h(t) \geq 0$, για κάθε $t \in M$. Αθροιστικά, ο κίνδυνος από την πρώτη χρονική στιγμή της μελέτης, μέχρι την παρούσα χρονική στιγμή, που αφορά το υπό μελέτη γεγονός, εκφράζεται από την αθροιστική συνάρτηση κινδύνου που ορίζεται ως:

$$H(t) = \int_0^t h(x)dx = -\ln[S(t)].$$

Επιπλέον, η συνάρτηση επιβίωσης συνδέεται με την αθροιστική συνάρτηση κινδύνου με την σχέση:

$$S(t) = \exp[-H(t)] = \exp\left[-\int_0^t h(x)dx\right].$$

Γνωρίζοντας οποιαδήποτε από τις συναρτήσεις $F(t)$, $f(t)$, $S(t)$, $h(t)$, $H(t)$ μπορούμε να προσδιορίσουμε και τις υπόλοιπες. Έτσι, όλες δίνουν ισοδύναμο ορισμό για την κατανομή της μεταβλητής T . Εδώ θα επικεντρωθούμε σε τρόπους εκτίμησης της $h(t)$ ή ισοδύναμα της $S(t)$. Πριν από τις μεθόδους δίνουμε τις ιδιαιτερότητες των δεδομένων που συναντάμε σε τέτοιες μελέτες και τα σύνολα δεδομένων που θα χρησιμοποιήσουμε στα παραδείγματα της παρούσας διατριβής.

1.1 Είδη λογοκρισίας και τύποι μεταβλητών

Μία πιθανή δυσκολία που μπορεί να προκύψει στην ανάλυση των δεδομένων επιβίωσης είναι ότι κάποιοι χρόνοι διάρκειας ζωής της μελέτης μπορεί να μην παρατηρηθούν πλήρως. Για παράδειγμα, στο τέλος ενός πειράματος που μετρά τον χρόνο ζωής ασθενών, κάποιοι από αυτούς μπορεί να ζουν και μετά το τέλος της έρευνας. Επίσης, ίσως μετράμε τον χρόνο ζωής μέχρι ένας ασθενής να επανεμφανίσει μία συγκεκριμένη ασθένεια και ο ασθενής για κάποιο λόγο να φύγει από την έρευνα. Τέτοιες ελλειπείς παρατηρήσεις του χρόνου ζωής ονομάζονται λογοκριμένες.

Γενικά στο πλαίσιο της παρούσας διατριβής θεωρούμε ότι έχουμε διαθέσιμο προς ανάλυση ένα δείγμα $T_1, \dots, T_n$ ανεξάρτητων και ισόνομων παρατηρήσεων από πληθυσμό $T \sim f(t) = F'(t)$. Τα είδη λογοκρισίας στα οποία μπορεί να εμπίπτει το συγκεκριμένο δείγμα δίνονται αναλυτικά παρακάτω.

Στη δεξιά λογοκρισία το υπό μελέτη γεγονός παρατηρείται μόνο αν συμβεί πριν από μία χρονική στιγμή, η οποία ίσως είναι διαφορετική για κάθε άτομο της έρευνας. Οι χρονικές στιγμές λογοκρισίας μπορεί να είναι σταθερές ή να είναι τυχαίες. Έστω ότι έχουμε να μελετήσουμε τον χρόνο επιβίωσης n ατόμων. Συμβολίζουμε με C_i , $i = 1, \dots, n$ τον χρόνο λογοκρισίας που αντιστοιχεί στο i -οστό άτομο. Ο ακριβής χρόνος ζωής ενός ατόμου είναι γνωστός αν και μόνο αν είναι μικρότερος ή ίσος του αντίστοιχου χρόνου λογοκρισίας. Στην περίπτωση που ο χρόνος ζωής ενός ατόμου είναι μεγαλύτερος του αντίστοιχου χρόνου λογοκρισίας, τότε είναι λογοκριμένος. Τελικά, τα δεδομένα που παρατηρούμε στην πράξη μπορούν να αναπαρασταθούν σε ζευγάρια τυχαίων μεταβλητών (Y_i, δ_i) , όπου

$$Y_i = \min(T_i, C_i), \delta_i = \begin{cases} 1, & \text{αν } T_i \leq C_i, \\ 0, & \text{αν } T_i > C_i. \end{cases} \quad (1.1.1)$$

Ο δείκτης δ_i , $i = 1, \dots, n$ ονομάζεται δείκτης λογοκρισίας και ισούται με 1, αν παρατηρούμε τον ακριβή χρόνο ζωής του i -οστού ατόμου και με 0, αν ο χρόνος ζωής του i -οστού ατόμου είναι λογοκριμένος.

Στην αριστερή λογοκρισία το πλήρες μέγεθος της εκάστοτε χρονικής διάρκειας, παρατηρείται πλήρως, μόνο αν συμβεί μετά από μία χρονική στιγμή. Στην περίπτωση που ο χρόνος ζωής ενός ατόμου είναι αριστερά λογοκριμένος για το συγκεκριμένο άτομο το υπό μελέτη γεγονός έχει συμβεί πριν να ξεκινήσουμε να το παρατηρούμε. Τα δεδομένα αναπαρίστανται σε ζευγάρια τυχαίων μεταβλητών

(Y_i, δ_i) , όπου

$$Y_i = \max(T_i, C_i), \delta_i = \begin{cases} 1, & \text{αν } T_i \geq C_i, \\ 0, & \text{αν } T_i < C_i. \end{cases}$$

Ο δείκτης λογοκρισίας δ_i , $i = 1, \dots, n$ ισούται με 1, αν παρατηρούμε τον ακριβή χρόνο ζωής του i -οστού ατόμου και με 0, αν ο χρόνος ζωής του i -οστού ατόμου είναι λογοκριμένος.

Συχνά, όταν έχουμε αριστερή λογοκρισία, μπορεί ταυτόχρονα να εμφανίζεται και δεξιά λογοκρισία. Αυτοί οι χρόνοι ζωής θεωρούνται διπλά λογοκριμένοι, ωστόσο και πάλι τα δεδομένα αναπαρίστανται σε ζευγάρια τυχαίων μεταβλητών (Y_i, δ_i) , όπου

$$Y_i = \max\{\min(T_i, C_{ri}), C_{li}\},$$

C_{ri}, C_{li} , $i = 1, \dots, n$ είναι ο χρόνος δεξιάς και αριστερής λογοκρισίας του $i = 1, \dots, n$ ατόμου, αντίστοιχα. Ο δείκτης λογοκρισίας δ_i έχει την τιμή 1 όταν ο χρόνος ζωής του i -οστού ατόμου είναι γνωστός, 0 όταν είναι δεξιά λογοκριμένος και -1 όταν είναι αριστερά λογοκριμένος, για $i = 1, \dots, n$.

Στη λογοκρισία σε διάστημα, που είναι ένα πιο γενικό είδος λογοκρισίας, γνωρίζουμε μόνο ότι το υπό μελέτη γεγονός έχει συμβεί σε ένα διάστημα, αλλά όχι την ακριβή χρονική στιγμή που συνέβη. Έτσι, για τις παρατηρήσεις θα ισχύει

$$Y_i \in (L_i, U_i) \quad i = 1, \dots, n,$$

με L_i, U_i να είναι το κάτω και άνω άκρο, αντίστοιχα, του διαστήματος λογοκρισίας που αντιστοιχεί στο $i = 1, \dots, n$ άτομο.

Γενικά, στην ανάλυση δεδομένων χρόνου ζωής είναι σημαντικό να ληφθούν υπόψιν χαρακτηριστικά των πειραματικών μονάδων, που θεωρούμε ότι επηρεάζουν τον χρόνο μέχρι να συμβεί το υπό μελέτη γεγονός. Τα χαρακτηριστικά αυτά αναφέρονται ως συµμεταβλητές ή επεξηγηµατικές µεταβλητές. Έστω ότι έχουμε m συµμεταβλητές, η πληροφορία αυτών των συµμεταβλητών οµαδοποιείται στο διάνυσµα $\mathbf{X} = (X_1, X_2, \dots, X_m)'$. Το διάνυσµα $\mathbf{X}$ µπορεί να περιλαµβάνει και ποσοτικές µεταβλητές και ποιοτικές, κατηγορικές, µεταβλητές. Οι ποσοτικές µπορούν να είναι είτε διακριτές, όπως ο δείκτης κάθε ατόµου, το πλήθος των επεµβάσεων ή η ηλικία, αν λάβουµε υπόψιν µόνο το ακέραιο µέρος, είτε συνεχείς, όπως η πίεση του αίµατος, η θερμοκρασία ή το βάρος. Οι ποιοτικές µεταβλητές µπορούν να είναι είτε διατάξιµες, όπως είναι η κατάσταση ενός ασθενή, είτε ονοµατικές, όπως είναι το φύλο ή η οµάδα θεραπείας. Ακόµη, το διάνυσµα $\mathbf{X}$ µπορεί να περιέχει δείκτριες µεταβλητές, όπως είναι η µεταβλητή που δηλώνει αν συνέβη κάποιο γεγονός ή όχι, ή η µεταβλητή που αντιστοιχεί κάθε ασθενή σε µια µέθοδο θεραπείας. Το διάνυσµα $\mathbf{X}$ µπορεί να περιλαµβάνει και χρονικά εξαρτηµένες µεταβλητές,

σε αυτήν την περίπτωση το συμβολίζουμε με $\mathbf{X}(t)$ και τις χρονικά εξαρτημένες μεταβλητές με $X_j(t)$, $j \in \{1, \dots, m\}$. Συνήθως οι χρονικά εξαρτημένες μεταβλητές αναφέρονται στο αν έχει συμβεί ή όχι ένα ενδιαμέσο γεγονός μέχρι τη χρονική στιγμή t .

Παραπάνω αναφέραμε πως μοντελοποιούμε τη συνάρτηση επιβίωσης και τη συνάρτηση κινδύνου του χρόνου ζωής. Συχνά είναι απαραίτητο να εκτιμήσουμε τη συνάρτηση επιβίωσης ή τη συνάρτηση κινδύνου, δοθείσης της τιμής κάποιων συμμεταβλητών. Στο πλαίσιο αυτής της διατριβής θα δούμε μεθόδους εκτίμησης της συνάρτησης επιβίωσης ή της συνάρτησης κινδύνου. Αναλυτικά οι μέθοδοι αυτές παρουσιάζονται στα αντίστοιχα κεφάλαια. Για το διάνυσμα συμμεταβλητών, έστω, $\mathbf{X} = (X_1, X_2, \dots, X_m)'$ χρειάζεται να εκτιμήσουμε την υπό συνθήκη συνάρτηση συνάρτηση επιβίωσης, η οποία ορίζεται ως $S(t | \mathbf{X}) = 1 - F(t | \mathbf{X})$, ή την υπό συνθήκη συνάρτηση κινδύνου, η οποία ορίζεται ως $h(t | \mathbf{X}) = f(t | \mathbf{X})/S(t | \mathbf{X})$. Το κύριο ενδιαφέρον της μελέτης είναι η υπό συνθήκη μοντελοποίηση του χρόνου μέχρι να συμβεί ένα γεγονός. Επιπλέον λόγος που λαμβάνουμε υπόψιν τις συμμεταβλητές στη μοντελοποίηση του χρόνου ζωής, είναι η επαλήθευση της σχέσης μεταξύ του χρόνου ζωής και μιας ή περισσότερων συμμεταβλητών. Για παράδειγμα, μπορεί να θέλουμε να συγκρίνουμε τις συναρτήσεις επιβίωσης για δύο ή περισσότερους τύπους θεραπείας, να θέλουμε να βγάλουμε προγνωστικά συμπεράσματα για έναν ασθενή με συγκεκριμένα χαρακτηριστικά, ή να θέλουμε να βρούμε τους παράγοντες κινδύνου για μια ασθένεια.

Οι τεχνικές που θα παρουσιάσουμε στην παρούσα διατριβή για την μοντελοποίηση της υπό συνθήκης συνάρτησης επιβίωσης ή της υπό συνθήκης συνάρτησης κινδύνου, θα παρουσιαστούν μέσω της εφαρμογής τους στα σύνολα δεδομένων που παρατίθενται στην επόμενη ενότητα.

1.2 Σύνολα πραγματικών δεδομένων

Σύνολο δεδομένων CNS lymphoma

Τα δεδομένα προέρχονται από μια μελέτη, η οποία βασίστηκε σε παρατηρήσεις και πραγματοποιήθηκε στο Πανεπιστήμιο επιστήμης υγείας του Όρεγκον (OHSU). 58 ασθενείς χωρίς AIDS, με λέμφωμα του κεντρικού νευρικού συστήματος, μελετήθηκαν στο OHSU από τον Ιανουάριο του 1982 μέχρι τον Μάρτιο του 1992. Η πρώτη ομάδα ασθενών αποτελείται από $n = 19$ ασθενείς έλαβε κρανιακή ακτινοβολία, πριν να παραπεμφθεί σε blood-brain barrier disruption (BBBD) χημειοθεραπεία. Η ομάδα 0 με $n = 39$ ασθενείς έλαβε ως αρχική θεραπεία την BBBD χημειοθεραπεία. Κύριο ενδιαφέρον είναι ο χρόνος επιβίωσης, σε χρόνια, από την

πρώτη BBBD χημειοθεραπεία. Δεδομένου ότι το σύνολο αυτό περιέχει δεξιά λογοκριμένα δεδομένα επιβίωσης, είναι κατάλληλο για εφαρμογή του μοντέλου αναλογικών κινδύνων του Cox. Στο Κεφάλαιο 2 γίνεται η προσαρμογή του μοντέλου, η εκτίμηση των παραμέτρων με χρήση της μεθόδου μερικής πιθανοφάνειας, αξιολόγηση της υπόθεσης της αναλογικότητας κινδύνων, καθώς και ο έλεγχος καταλληλότητας του μοντέλου με χρήση υπολοίπων. Σκοπός της ανάλυσης είναι να εντοπιστούν συμμεταβλητές που επηρεάζουν σημαντικά το χρόνο ζωής των ασθενών, καθώς και να συγκριθεί η επίδραση που έχει η ομάδα που ανήκει ο ασθενής στην επιβίωση. Οι μεταβλητές του συνόλου επεξηγούνται στη συνέχεια.

- PT.NUMBER: ένας ξεχωριστός δείκτης για κάθε άτομο.
- GROUP: η ομάδα στην οποία ανήκει ο ασθενής. Στην ομάδα 0 ανήκουν όσοι ασθενείς έλαβαν ως αρχική θεραπεία την BBBD χημειοθεραπεία, ενώ στην 1 όσοι έλαβαν κρανιακή ακτινοβολία πριν από την BBBD χημειοθεραπεία.
- SEX: η μεταβλητή που δηλώνει το φύλο του ασθενή. Η τιμή 0 αντιστοιχεί σε άνδρες, ενώ η τιμή 1 σε γυναίκες.
- AGE: η ηλικία, σε έτη, κατά την οποία έγινε η πρώτη BBBD χημειοθεραπεία.
- STATUS: η μεταβλητή που δηλώνει αν ο ασθενής είναι ζωντανός ή αν έχει πεθάνει. Παίρνει την τιμή 0 για την πρώτη περίπτωση και την τιμή 1 για τη δεύτερη.
- DXTOB3: ο χρόνος, σε έτη, από τη στιγμή της διάγνωσης μέχρι τη στιγμή που γίνεται η πρώτη BBBD χημειοθεραπεία.
- DXTODeath: ο χρόνος, σε έτη, από τη διάγνωση μέχρι τον θάνατο.
- B3TODeath: ο χρόνος, σε έτη, από την πρώτη BBBD χημειοθεραπεία ως τον θάνατο.
- KPS.PRE: το σκορ Karnofsky πριν την πρώτη BBBD χημειοθεραπεία. Είναι μια αριθμητική τιμή με δυνατές τιμές από 0 ως 100.
- LESSING: η μεταβλητή που δηλώνει αν υπάρχει μόνο μια κάκωση ή πολλαπλές. Η τιμή 0 αντιστοιχεί σε μια, ενώ η 1 σε πολλαπλές κακώσεις.
- LESDEEP: η μεταβλητή που δηλώνει αν η κάκωση είναι επιφανειακή ή εκτείνεται σε βάθος. Αντίστοιχα λαμβάνει τις τιμές 0 και 1.

- LESSUP: η μεταβλητή που δηλώνει αν η κάκωση είναι πάνω, κάτω ή και πάνω και κάτω, από ένα σημείο αναφοράς. Λαμβάνει την τιμή 0, την τιμή 1 και την τιμή 2 για κάθε περίπτωση, αντίστοιχα.
- PROC: δηλώνει την ιατρική διαδικασία. Λαμβάνει την τιμή 1 για μερική χειρουργική τομή, την τιμή 2 για βιοψία και την τιμή 3 για οποιαδήποτε άλλη μέθοδο.
- RAD4000: εκφράζει αν η ακτινοβολία είναι μεγαλύτερη από 4000 rad. Αν δεν είναι λαμβάνει την τιμή 0, αλλιώς την τιμή 1.
- CHEMOPRIOR: μια μεταβλητή που εκφράζει αν έχει ληφθεί χημειοθεραπεία προηγουμένως. Λαμβάνει την τιμή 0, αν δεν έχει ληφθεί και 1 αν έχει ληφθεί.
- RESPONSE: η μεταβλητή που υποδηλώνει την ανταπόκριση του όγκου στη χημειοθεραπεία. Έχει την τιμή 1 αν υπάρχει πλήρης ανταπόκριση και την τιμή 2 αν υπάρχει μερική ανταπόκριση. Οι κενές μετρήσεις αναφέρονται σε ελλιπή δεδομένα.

Σύνολο δεδομένων Leukemia

Το σύνολο δεδομένων Leukemia προέρχεται από τη δημοσίευση των Feigl and Zelen (1965), που αφορά την ανάλυση παλινδρόμησης για δεδομένα χρόνου ζωής και περιέχει τα δεδομένα των χρόνων ζωής 33 ασθενών που πάσχουν από λευχαιμία. Σκοπός της μελέτης ήταν να παρακολουθηθεί ο χρόνος ζωής των ασθενών, προκειμένου να αναπτυχθούν μέθοδοι εκτίμησης του χρόνου ζωής και να βρεθούν παράγοντες που επηρεάζουν τη διάρκεια ζωής των ατόμων με λευχαιμία. Χρησιμοποιούμε το συγκεκριμένο σύνολο στο Κεφάλαιο 4, στην ενότητα που αφορά την αναζήτηση του κατάλληλου μοντέλου, προκειμένου να διερευνήσουμε αν υπάρχει κάποια σχέση που συνδέει τον χρόνο ζωής με το πλήθος των λευκών αιμοσφαιρίων και τη μεταβλητή ag. Οι μεταβλητές του συνόλου δεδομένων είναι οι εξής:

- t: ο χρόνος ζωής σε εβδομάδες από τη διάγνωση.
- WBC: το πλήθος των λευκών αιμοσφαιρίων κατά τη διάγνωση (white blood cell count).
- ag: μία δίτιμη μεταβλητή που υποδεικνύει το αποτέλεσμα ενός τεστ σχετικού με τα χαρακτηριστικά των λευκών αιμοσφαιρίων. Αν είναι θετικό λαμβάνει την τιμή 1, ενώ αν είναι αρνητικό λαμβάνει την τιμή 0.

Σύνολο δεδομένων Insulating Fluid Failure Times

Τα συγκεκριμένα δεδομένα προέρχονται από την εργασία του Nelson (1972). Στην εργασία αυτή περιγράφεται το αποτέλεσμα ενός πειράματος με δεδομένα χρόνου ζωής, στο οποίο δείγματα από έναν τύπο μονωτικού υγρού υποβλήθηκαν στον παράγοντα σταθερής ηλεκτρικής τάσης. Παρατηρήθηκε ο χρόνος μέχρι την στιγμή που κάθε δείγμα χάλασε. Βασικός σκοπός του πειράματος ήταν να βρεθεί η σχέση που έχει ο χρόνος ζωής με την ηλεκτρική τάση και να βρεθεί ένα μοντέλο για τον χρόνο ζωής, που μπορεί να γενικευθεί και για χαμηλότερες τιμές τάσης. Επειδή η μεταβλητή voltage επιταχύνει τον χρόνο μέχρι την εμφάνιση βλάβης, το A.F.T. μοντέλο είναι κατάλληλο. Το συγκεκριμένο σύνολο χρησιμοποιείται στο Κεφάλαιο 4 στην ενότητα που αφορά την αναζήτηση του κατάλληλου μοντέλου. Οι μεταβλητές του συνόλου δεδομένων είναι οι εξής:

- voltage: η μεταβλητή αυτή αντιστοιχεί στο επίπεδο της σταθερής ηλεκτρικής τάσης στην οποία υποβάλλεται κάθε δείγμα από έναν τύπο μονωτικού υγρού. Η ηλεκτρική τάση έχει δυνατές τιμές $v = 26, 28, 30, 32, 34, 36, 38$ kV.
- time: ο χρόνος, σε λεπτά, μέχρι την εμφάνιση βλάβης στο δείγμα.

Στον Πίνακα 1.1 δίνονται τα αποτελέσματα για 7 ομάδες δειγμάτων, που μελετήθηκαν για ηλεκτρική τάση από 26 ως 38 kilovolts (kV). Ειδικότερα 76 δείγματα μονωτικού υγρού εξετάστηκαν για ηλεκτρική τάση $v = 26, 28, 30, 32, 34, 36, 38$ kV.

Επίπεδο τάσης	n_i	Χρόνος βλάβης
26	3	5.79, 1579.52, 2323.7
28	5	68.85, 426.07, 110.29, 108.29, 1067.6
30	11	17.05, 22.66, 21.02, 175.88, 139.07, 144.12, 20.46, 43.40, 194.90, 47.30, 7.74
32	15	0.40, 82.85, 9.88, 89.29, 215.10, 2.75, 0.79, 15.93, 3.91, 0.27, 0.69, 100.58, 27.80, 13.95, 53.24
34	19	0.96, 4.15, 0.19, 0.78, 8.01, 31.75, 7.35, 6.50, 8.27, 33.91, 32.52, 3.16, 4.85, 2.78, 4.67, 1.31, 12.06, 36.71, 72.89
36	15	1.97, 0.59, 2.58, 1.69, 2.71, 25.50, 0.35, 0.99, 3.99, 3.67, 2.07, 0.96, 5.35, 2.90, 13.77
38	8	0.47, 0.73, 1.40, 0.74, 0.39, 1.13, 0.09, 2.38

Πίνακας 1.1: Χρόνος, σε λεπτά, μέχρι την εμφάνιση βλάβης για κάθε ένα από τα 7 επίπεδα της ηλεκτρικής τάσης.

Σύνολο δεδομένων oldmort

Το σύνολο δεδομένων oldmort περιλαμβάνει το ιστορικό ζωής ατόμων τρίτης ηλικίας, από 1η Ιανουαρίου 1860 έως 31η Δεκεμβρίου 1880. Η δειγματοληψία έγινε με τον εξής τρόπο: κάθε άτομο που ήταν παρόν, ζωντανό και από 60 έτη και πάνω, μεταξύ 1ης Ιανουαρίου 1860 και 31 Δεκεμβρίου 1880, παρακολούθηθηκε από την ηλικία εισόδου στην έρευνα (για τα περισσότερα άτομα αυτή είναι η ηλικία των 60 ετών) μέχρι την τελευταία φορά που το είδαν, η οποία ορίζεται ως θάνατος, μετανάστευση ή το τέλος της έρευνας (31 Δεκεμβρίου 1880). Διακρίνονται δύο τύποι ολοκλήρωσης της παρατήρησης ενός ατόμου: Είτε πρόκειται για θάνατο, είτε για κάτι άλλο, μετανάστευση ή τέλος της έρευνας. Στην πρώτη περίπτωση λέμε ότι έχει συμβεί το υπό μελέτη γεγονός, ενώ στη δεύτερη όχι. Τα συγκεκριμένα δεδομένα θα χρησιμοποιηθούν στο Κεφάλαιο 4 για την παρουσίαση ενός παραδείγματος, στο οποίο θα επιλέξουμε αν το μοντέλο αναλογικών κινδύνων ή το A.F.T. μοντέλο είναι καταλληλότερο για την ανάλυση. Τα δεδομένα περιλαμβάνουν 6508 παρατηρήσεις για 4603 άτομα και μετρήσεις για 13 μεταβλητές. Οι μεταβλητές που θα χρησιμοποιήσουμε επεξηγούνται στη συνέχεια.

- id: ένας ξεχωριστός δείκτης για κάθε άτομο.
- enter, exit: η ηλικία εισόδου στην έρευνα και η ηλικία που σταμάτησε η παρακολούθηση. Η ηλικία υπολογίζεται ως ο αριθμός ημερών από την ημέρα της γέννησης δια 365.25 για να έχουμε την ηλικία σε έτη. Ο παρονομαστής είναι η μέση διάρκεια ενός χρόνου.
- event: μια λογική μεταβλητή με τιμές TRUE ή FALSE, δηλώνει αν το τέλος της παρατήρησης είναι θάνατος, οπότε έχει την τιμή TRUE, ή όχι, οπότε έχει την τιμή FALSE.
- sex: μια κατηγορική μεταβλητή που δηλώνει το φύλο και έχει δύο επίπεδα, άνδρας και γυναίκα.
- c1n: οικογενειακή κατάσταση, είναι μια κατηγορική μεταβλητή με τρία επίπεδα, ανύπαντρος/η, παντρεμένος/η, χήρος/α.
- region: παρούσα γεωγραφική θέση κατοικίας, γίνεται ομαδοποίηση σε τρεις περιοχές: πόλη Σούντσβαλ, αγροτική και εργατική περιοχή.

Σύνολο δεδομένων addicts

Στην έρευνα των Caplehorn and Bell (1991), συγκρίνονται δυο κλινικές θεραπείες ατόμων εθισμένων στην ηρωίνη, για να γίνει εκτίμηση του χρόνου ζωής υπό

τη θεραπεία μεθαδόνης. Ο χρόνος ζωής ενός ασθενή ορίζεται ως ο χρόνος σε μέρες μέχρι να λάβει εξιτήριο από την κλινική ή μέχρι την λογοκρισία. Βασικός σκοπός της έρευνας είναι να συγκρίνει τον χρόνο παραμονής των ασθενών στις δυο κλινικές, οι οποίες διαφέρουν στις συνθήκες διαβίωσης των ασθενών. Προκειμένου να γίνει αυτή η σύγκριση, στο Κεφάλαιο 3 προσαρμόζουμε τα δεδομένα στην επέκταση του μοντέλου του Cox και ερμηνεύουμε την αναλογία κινδύνου που αντιστοιχεί στην μεταβλητή *clinic*. Οι μεταβλητές του συνόλου δεδομένων είναι οι εξής:

- *id*: ένας ξεχωριστός δείκτης για κάθε άτομο.
- *clinic*: η μεταβλητή που δηλώνει σε ποια κλινική έγινε η θεραπεία του ασθενή. Έχει δυνατές τιμές 1 και 2 που αντιστοιχούν στις δυο κλινικές.
- *status*: δηλώνει αν ο ασθενής πήρε εξιτήριο από την κλινική, οπότε παίρνει την τιμή 1, ή αν έχει λογοκριθεί, οπότε παίρνει την τιμή 0.
- *time*: ο χρόνος, σε μέρες, μέχρι ο ασθενής να λάβει εξιτήριο από την κλινική ή μέχρι να λογοκριθεί.
- *prison*: δηλώνει αν ο ασθενής έχει ιστορικό φυλάκισης, τότε λαμβάνει την τιμή 1, ή όχι, οπότε λαμβάνει την τιμή 0.
- *dose*: μια συνεχής μεταβλητή που δηλώνει τη μέγιστη δοσολογία μεθαδόνης για έναν ασθενή, η οποία μετράται σε mg ανά μέρα.

ΚΕΦΑΛΑΙΟ 2

ΤΟ ΜΟΝΤΕΛΟ ΑΝΑΛΟΓΙΚΩΝ ΚΙΝΔΥΝΩΝ ΤΟΥ COX

Το μοντέλο που χρησιμοποιείται πιο συχνά για την εκτίμηση της δεσμευμένης συνάρτησης κινδύνου ή της δεσμευμένης συνάρτησης επιβίωσης είναι το μοντέλο αναλογικών κινδύνων του Cox, το οποίο προτάθηκε στην εργασία του Cox (1972). Το μοντέλο είναι ημιπαραμετρικής φύσης, που σημαίνει ότι ένα μέρος του εκτιμάται μη παραμετρικά και το υπόλοιπο μέρος εκτιμάται με παραμετρικό τρόπο. Πλεονέκτημα του μοντέλου είναι ότι δεν έχει περιορισμό στους τύπους δεδομένων που μπορούν να συμπεριληφθούν στο μοντέλο ως συμμεταβλητές. Η δομή του μοντέλου αναλογικών κινδύνων του Cox είναι ότι η υπό συνθήκη συνάρτηση κινδύνου εκτιμάται ως το γινόμενο δυο συναρτήσεων. Η μια είναι η βασική συνάρτηση κινδύνου, $h_0(t)$ (baseline hazard function), η οποία είναι μη παραμετρικά άγνωστη και είναι μόνο συνάρτηση του χρόνου και άρα ανεξάρτητη των συμμεταβλητών. Η $h_0(t)$ εκφράζει τον κίνδυνο που έχει οποιοδήποτε άτομο της μελέτης να βιώσει το υπό μελέτη γεγονός, ανεξαρτήτως των τιμών των συμμεταβλητών. Η άλλη συνάρτηση ποσοτικοποιεί την επίδραση των συμμεταβλητών στην υπό συνθήκη συνάρτηση κινδύνου και θεωρείται ότι έχει την παραμετρική μορφή $\exp(\mathbf{X}'\boldsymbol{\beta})$. Οπότε, το τελικό αποτέλεσμα είναι η ημιπαραμετρική εκτίμηση της υπό συνθήκης συνάρτησης κινδύνου βάση του μοντέλου:

$$h(t|\mathbf{X}) = h_0(t) \exp(\mathbf{X}'\boldsymbol{\beta}) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m).$$

Η χρησιμότητα του μοντέλου αναλογικών κινδύνων του Cox στην πράξη έγκειται στο ότι οι συντελεστές $\beta_1, \dots, \beta_m$ του μοντέλου ποσοτικοποιούν τη συνεισφορά της κάθε συμμεταβλητής στον συνολικό κίνδυνο. Συγκεκριμένα:

- ένας θετικός συντελεστής αυξάνει τον κίνδυνο και έτσι μειώνει τον (μέσο) αναμενόμενο χρόνο επιβίωσης,

- ένας αρνητικός συντελεστής μειώνει τον κίνδυνο και έτσι αυξάνει τον (μέσο) αναμενόμενο χρόνο επιβίωσης.
- Στην πράξη ακόμη πιο χρήσιμος από την τιμή του κάθε συντελεστή της κάθε παραμέτρου ξεχωριστά, είναι ο λόγος των συναρτήσεων κινδύνου για δύο ομάδες. Ο λόγος αυτός μπορεί να χρησιμοποιηθεί για να εξεταστεί αν ο χρόνος επιβίωσης της πρώτης ομάδας είναι μεγαλύτερος από αυτόν της δεύτερης.

Στην πράξη τόσο η βασική συνάρτηση κινδύνου, $h_0(t)$, όσο και το διάνυσμα των συντελεστών, $\beta = (\beta_1, \dots, \beta_m)'$, είναι άγνωστα και άρα θα πρέπει να εκτιμηθούν. Όπως θα δούμε, η εκτίμηση της συνάρτησης κινδύνου ανάγεται μόνο στην εκτίμηση των συντελεστών του διανύσματος των παραμέτρων του μοντέλου $\beta = (\beta_1, \dots, \beta_m)'$, η οποία επιτυγχάνεται μέσω της τεχνικής της μερικής πιθανοφάνειας που δίνεται στην επόμενη υποενότητα.

2.1 Η μέθοδος μερικής πιθανοφάνειας

Η μέθοδος μερικής πιθανοφάνειας είναι μια επέκταση της μεθόδου μέγιστης πιθανοφάνειας στην περίπτωση των λογοκριμένων δεδομένων. Ξεκινάμε με την παρουσίαση της μεθόδου μερικής μέγιστης πιθανοφάνειας με την παράθεση της συνάρτησης πιθανοφάνειας για δεξιά λογοκριμένα δεδομένα. Θυμόμαστε ότι $T \sim f = F'$. Η συνάρτηση επιβίωσης της T συμβολίζεται με S_f . Ακόμη, C είναι ένας τυχαίος χρόνος λογοκρισίας με $C \sim g = G'$. Η συνάρτηση επιβίωσης της C συμβολίζεται με S_g . Έστω αντίστοιχα, M και S_m οι συναρτήσεις κατανομής και επιβίωσης της μεταβλητής $Y = \min(T, C)$. Έχουμε δείγμα n ανεξάρτητων και ισόνομων τυχαίων παρατηρήσεων (Y_i, δ_i) , $i = 1, \dots, n$. Οι χρόνοι T_i , C_i υποθέτουμε ότι είναι ανεξάρτητοι. Από την υπόθεση ανεξαρτησίας έχουμε ότι: $S_m(y) = P(Y > y) = P(T > y, C > y) = P(T > y)P(C > y) = S_f(y)S_g(y)$. Οπότε, $M(y) = 1 - S_f(y)S_g(y)$.

Υποθέτουμε ότι η f εξαρτάται από μια παράμετρο θ και η g από μια παράμετρο η . Οι παράμετροι θ και η είναι πολυδιάστατες. Για δοθέν δείγμα n ανεξάρτητων και ισόνομων τυχαίων ζευγαριών (Y_i, δ_i) , $i = 1, \dots, n$, όπως περιγράφεται στην εξίσωση (1.1.1) η συνάρτηση πιθανοφάνειας είναι:

$$L(\theta, \eta) = \left[\prod_{i=1}^n S_g(Y_i, \eta)^{\delta_i} \{g(Y_i, \eta)\}^{1-\delta_i} \right] \left[\prod_{i=1}^n f(Y_i, \theta)^{\delta_i} \{S_f(Y_i, \theta)\}^{1-\delta_i} \right].$$

Παρατηρούμε ότι ο πρώτος όρος στο δεξί μέλος είναι ανεξάρτητος του θ και άρα δεν επηρεάζει τη διαδικασία μεγιστοποίησης της συνάρτησης πιθανοφάνειας ως

προς θ . Άρα είναι ισοδύναμο να μεγιστοποιήσουμε τη συνάρτηση πιθανοφάνειας:

$$L(\theta) = \prod_{i=1}^n f(Y_i, \theta)^{\delta_i} \{S_f(Y_i, \theta)\}^{1-\delta_i}.$$

Για να επεκτείνουμε τη μέθοδο πιθανοφάνειας στο μοντέλο του Cox θα βασιστούμε στη γνωστή σχέση που ορίζει τη συνάρτηση πυκνότητας πιθανότητας ως το γινόμενο της συνάρτησης κινδύνου με τη συνάρτηση επιβίωσης, δηλαδή

$$f(t) = h(t) \times S(t). \quad (2.1.1)$$

Σύμφωνα με το μοντέλο του Cox η συνάρτηση κινδύνου $h(t|\mathbf{X})$, μπορεί να εκτιμηθεί βάσει της σχέσης

$$h(t|\mathbf{X}) = h_0(t) \exp(\mathbf{X}'\beta).$$

Κατά συνέπεια, με αντικατάσταση της σχέσης (2.1.1) στην ως άνωθεν εξίσωση προκύπτει ότι η σ.π.π της τυχαίας μεταβλητής T γράφεται ως:

$$f(t) = h_0(t) \exp(\mathbf{X}'\beta) \times S(t).$$

Γίνεται αμέσως αντιληπτό ότι η συνάρτηση πιθανοφάνειας εξαρτάται από τη βασική συνάρτηση κινδύνου, $h_0(t)$, η οποία είναι μη παραμετρικά άγνωστη. Αυτό το γεγονός από μόνο του εμποδίζει τη μεγιστοποίηση της συνάρτησης πιθανοφάνειας. Για να αντιμετωπιστεί το πρόβλημα αυτό στην εργασία Cox (1975), ο Cox τροποποίησε τη μέθοδο μέγιστης πιθανοφάνειας με σκοπό την εκτίμηση των άγνωστων παραμέτρων, β , χωρίς να χρειαστεί να γίνει εκτίμηση της $h_0(t)$. Συγκεκριμένα, πρότεινε τη χρήση μιας συνάρτησης πιθανοφάνειας που βασίζεται σε δεσμευμένες πιθανότητες, ανεξάρτητες της βασικής συνάρτησης κινδύνου. Οι εκτιμητές των συντελεστών $\beta = (\beta_1, \dots, \beta_m)'$ προκύπτουν από μεγιστοποίηση ως προς $\theta = \beta$ της συνάρτησης μερικής πιθανοφάνειας. Η κεντρική ιδέα της τροποποίησης είναι ότι για να είναι εφικτή η μεγιστοποίηση της συνάρτησης μερικής πιθανοφάνειας ως προς β , θα πρέπει η βασική συνάρτηση κινδύνου, $h_0(t)$, να είναι ίδια για όλα τα άτομα της μελέτης. Αυτό εισάγει τη συνθήκη της αναλογικότητας κινδύνων, η οποία όταν ισχύει, σηματοδοτεί και την ισχύ των εκτιμώμενων παραμέτρων και του μοντέλου αναλογικών κινδύνων του Cox. Η συνθήκη της αναλογικότητας κινδύνων ισχύει όταν, ισοδύναμα, η βασική συνάρτηση κινδύνου είναι σταθερή. Οπότε για να μπορεί να εφαρμοστεί το μοντέλο του Cox στην πράξη θα πρέπει πρώτα να ελέγχεται η υπόθεση της αναλογικότητας κινδύνων. Μαθηματικά η μέθοδος της μερικής πιθανοφάνειας περιγράφεται ως εξής:

Έστω t^* μια χρονική στιγμή στην οποία συνέβη το υπό μελέτη γεγονός. Έστω $R(t^*)$ το σύνολο των δεικτών των ατόμων που είναι ζωντανοί και μη λογοκριμένοι ακριβώς πριν τη χρονική στιγμή t^* . Τότε, η δεσμευμένη πιθανότητα να συμβεί μια περίπτωση του υπό μελέτη γεγονότος την αμέσως επόμενη χρονική στιγμή είναι:

$$\begin{aligned} P\{\text{Ένας θάνατος στο διάστημα } [t^*, t^* + \Delta t) | R(t^*)\} \\ &= \sum_{l \in R(t^*)} P\{T_l \in [t^*, t^* + \Delta t) | T_l \geq t^*\} \\ &\approx \sum_{l \in R(t^*)} h(t^* | \mathbf{X}_l) \Delta t^* \\ &= \sum_{l \in R(t^*)} h_0(t^*) \exp(\mathbf{X}_l' \boldsymbol{\beta}) \Delta t^*. \end{aligned}$$

Έτσι αν η πιθανότητα $P\{\text{Ένας θάνατος τη χρονική στιγμή } t^* | R(t^*)\}$ ισούται με

$$\sum_{l \in R(t^*)} P\{T_l = t^* | T_l \geq t^*\}$$

έχουμε ότι:

$$P\{\text{Ένας θάνατος τη χρονική στιγμή } t^* | R(t^*)\} = \sum_{l \in R(t^*)} h_0(t^*) \exp(\mathbf{X}_l' \boldsymbol{\beta}).$$

Έστω $t_{(1)}, \dots, t_{(r)}$ οι $r \leq n$ διατεταγμένοι μη λογοκριμένοι χρόνοι θανάτου, τότε η $t_{(j)}$ είναι η j -οστη χρονική στιγμή θανάτου. Έστω $\mathbf{X}_{(j)}$ το διάνυσμα των μεταβλητών που σχετίζονται με το άτομο που πέθανε τη χρονική στιγμή $t_{(j)}$. Ορίζουμε τα παρακάτω ενδεχόμενα:

$$A = \{\text{Το άτομο με χαρακτηριστικά } \mathbf{X}_{(j)} \text{ να πεθάνει τη χρονική στιγμή } t_{(j)}\}$$

$$B = \{\text{έχουμε ένα θάνατο στο σύνολο } R(t_{(j)}) \text{ τη χρονική στιγμή } t_{(j)}\}$$

$$\Gamma = \{\text{το άτομο ανήκει στο σύνολο } R(t_{(j)})\}$$

$$\Delta = \{\text{έχουμε ένα θάνατο τη χρονική στιγμή } t_{(j)}\}.$$

Τότε για κάθε j έχουμε:

$$\begin{aligned}
 L_j(\beta) &= P(\mathbf{A}|\mathbf{B}) \\
 &= \frac{P(\mathbf{A}|\Gamma)}{P(\Delta|R(t_{(j)}))} \\
 &= \frac{h_0(t_{(j)}) \exp(\mathbf{X}'_{(j)}\beta)}{\sum_{l \in R(t_{(j)})} h_0(t_{(j)}) \exp(\mathbf{X}'_l\beta)} \\
 &= \frac{\exp(\mathbf{X}'_{(j)}\beta)}{\sum_{l \in R(t_{(j)})} \exp(\mathbf{X}'_l\beta)}.
 \end{aligned}$$

Στην εργασία Cox (1975), ο Cox όρισε τη μερική πιθανοφάνεια ως το γινόμενο των $L_j(\beta)$, $j = 1, \dots, r$. Συνολικά η συνάρτηση μερικής πιθανοφάνειας, $L_c(\beta)$, δίνεται από τη σχέση:

$$L_c(\beta) = \prod_{j=1}^r L_j(\beta) = \prod_{j=1}^r \frac{\exp(\mathbf{X}'_{(j)}\beta)}{\sum_{l \in R(t_{(j)})} \exp(\mathbf{X}'_l\beta)}.$$

Θυμόμαστε ότι στο μοντέλο με τυχαία λογοκρισία παρατηρούμε τους χρόνους $Y_1, \dots, Y_n$ μαζί με τις αντίστοιχες δείκτριες μεταβλητές $\delta_1, \dots, \delta_n$, όπου $\delta_i = 1$ αν η Y_i παρατήρηση δεν έχει λογοκριθεί και $\delta_i = 0$ αν η Y_i παρατήρηση είναι λογοκριμένη. Μπορούμε να δώσουμε μια ισοδύναμη μορφή στη συνάρτηση μερικής πιθανοφάνειας με όρους όλων των n παρατηρούμενων χρόνων. Καταρχήν έχουμε ότι:

$$L_c(\beta) = \prod_{j=1}^n \left(\frac{\exp(\mathbf{X}'_j\beta)}{\sum_{l \in R(t_j)} \exp(\mathbf{X}'_l\beta)} \right)^{\delta_i}.$$

Αν ορίσουμε D το σύνολο των δεικτών, $i = 1, \dots, n$, που αντιστοιχούν σε χρόνους ζωής ατόμων που βίωσαν το υπό μελέτη γεγονός, δηλαδή τους μη λογοκριμένους, ένας ισοδύναμος ορισμός της συνάρτησης πιθανοφάνειας είναι ο εξής:

$$L_c(\beta) = \prod_{i \in D} \frac{\exp(\mathbf{X}'_i\beta)}{\sum_{j \in R(t_i)} \exp(\mathbf{X}'_j\beta)/n}.$$

Συνοπτικά, θα πρέπει να έχουμε υπόψη ότι η εκτίμηση των παραμέτρων β που προκύπτει από τη μεγιστοποίηση της $L_c(\beta)$ είναι χρήσιμη όταν η βασική συνάρτηση κινδύνου είναι σταθερή. Από τη μορφή της μερικής πιθανοφάνειας φαίνεται ότι για να αναλύσουμε την επίδραση των συμμεταβλητών δεν είναι ανάγκη

να εκτιμήσουμε τη βασική συνάρτηση κινδύνου $h_0(t)$. Ο Cox (1975) υποστηρίζει ότι η περισσότερη πληροφορία που σχετίζεται με τους εκτιμητές β , για παλινδρόμηση με λογοκριμένα δεδομένα, περιέχεται στη συνάρτηση μερικής πιθανοφάνειας. Επίσης, παρατηρούμε ότι η συνάρτηση μερικής πιθανοφάνειας δεν είναι πραγματική πιθανοφάνεια καθώς το ολοκλήρωμά της στο σύνολο $\{0, 1\}^n \times \mathbb{R}_+^n$ δεν είναι ίσο με 1. Η συνάρτηση μερικής πιθανοφάνειας περιέχει στον αριθμητή την πληροφορία μόνο για τα άτομα που βιώνουν το υπό μελέτη γεγονός και στον παρονομαστή την πληροφορία όλων των ατόμων που δεν έχουν υποστεί το υπό μελέτη γεγονός, ακόμη και αυτών που μπορεί να λογοκριθούν στο μέλλον. Συνεπώς οι λογοκριμένες παρατηρήσεις λαμβάνονται υπόψη μόνο έμμεσα στον παρονομαστή. Οι εκτιμήσεις των παραμέτρων β είναι συνάρτηση των συμμεταβλητών $\mathbf{X}_j$, η κατάταξη των οποίων καθορίζεται από την κατάταξη των χρονικών στιγμών που συνέβη το υπό μελέτη γεγονός. Με αυτήν την έννοια οι χρονικές στιγμές επηρεάζουν μόνο έμμεσα την εκτίμηση των παραμέτρων β του μοντέλου.

2.2 Ιδιότητες εκτιμώμενων παραμέτρων

Με σκοπό τη μελέτη των στατιστικών ιδιοτήτων των εκτιμητών των παραμέτρων του μοντέλου αναλογικών κινδύνων, στην παρούσα ενότητα βασιζόμαστε στην εργασία του Tsiatis (1981). Υποθέτουμε ότι το διάνυσμα των συμμεταβλητών είναι μονοδιάστατο, δηλαδή $\mathbf{X} = X_1 = X$, με συνάρτηση πυκνότητας πιθανότητας $z(x)$. Η παραδοχή αυτή διατηρείται καθ' όλη τη διάρκεια της ενότητας. Υποθέτουμε ότι η μεταβλητή C που εκφράζει τον χρόνο λογοκρισίας είναι φραγμένη, δηλαδή $C \leq T_0 < \infty$. Έστω $h_g(t|X)$ η υπό συνθήκη συνάρτηση κινδύνου της μεταβλητής C , δοθέντος του X . Πριν από την διατύπωση και την απόδειξη του θεωρήματος που εκφράζει τις ασυμπτωτικές ιδιότητες των εκτιμητών μερικής πιθανοφάνειας του Cox, δίνουμε τους ακόλουθους ορισμούς.

Η υπό συνθήκη πιθανότητα επιβίωσης μέχρι τη χρονική στιγμή t , χωρίς να γίνει λογοκρισία, δοθέντος του X είναι:

$$S_Y(t|X) = P(Y \geq t|X) = \exp \left[- \int_0^t \{h_0(u) \exp(\beta X) + h_g(u, X)\} du \right]. \quad (2.2.1)$$

Η πιθανότητα επιβίωσης μέχρι τη χρονική στιγμή t χωρίς να γίνει λογοκρισία και τελικά να συμβεί το υπό μελέτη γεγονός πριν τη λογοκρισία είναι:

$$Q(t) = P(Y \geq t, \delta = 1) = \int Q(t|X) z(X) dX, \quad (2.2.2)$$

όπου $z(X)$ είναι η συνάρτηση πυκνότητας πιθανότητας της X και

$$Q(t|X) = P(Y \geq t, \delta = 1|X) = \int_t^{T_0} \{h_0(u) \exp(\beta X) S_Y(u|X)\} du, \quad (2.2.3)$$

με T_0 να είναι η χρονική στιγμή κατά την οποία τελειώνει η έρευνα. Η παράγωγος του $Q(t|X)$ είναι:

$$\frac{dQ(t)}{dt} = -h_0(t) \int \{\exp(\beta X) S_Y(t|X) z(X)\} dX. \quad (2.2.4)$$

Έστω η δείκτρια συνάρτηση I_A η οποία παίρνει την τιμή 1 όταν το γεγονός A έχει συμβεί και την τιμή 0 όταν το γεγονός A δεν έχει συμβεί. Έστω $w(X)$ μια συνεχής συνάρτηση της X . Ορίζουμε

$$\begin{aligned} E\{w(X), t\} &= E\{w(X) I_{\{Y \geq t\}}\} = E\{w(X) P(Y \geq t|X)\} \\ &= \int w(X) S_Y(t|X) z(X) dX, \end{aligned} \quad (2.2.5)$$

και

$$E_1\{w(X), t\} = E\{w(X) I_{\{Y \geq t, \delta=1\}}\} = E\{w(X) Q(t|X)\}. \quad (2.2.6)$$

Η παράγωγος της $E_1\{w(X), t\}$ ισούται με

$$\frac{dE_1\{w(X), t\}}{dt} = -h_0(t) \int w(X) \exp(\beta X) S_Y(t|X) z(X) dX. \quad (2.2.7)$$

Με χρήση των (2.2.4) – (2.2.6) έχουμε ότι:

$$h_0(t) = -\frac{dQ(t)/dt}{E\{\exp(\beta X), t\}}, \quad (2.2.8)$$

και

$$\frac{dE_1\{w(X), t\}}{dt} = \frac{dQ(t)}{dt} \frac{E\{w(X) \exp(\beta X), t\}}{E\{\exp(\beta X), t\}}. \quad (2.2.9)$$

Έστω n το πλήθος των χρόνων ζωής στη μελέτη. Για κάθε χρόνο ζωής έχουμε το αντίστοιχο διάνυσμα (Y_i, δ_i, X_i) , $i = 1, \dots, n$ τα οποία θεωρούνται ανεξάρτητα και ισόνομα τυχαία διανύσματα. Συμβολίζουμε με $\hat{Q}(t)$ και $\hat{E}(\exp(\beta X), t)$ αντίστοιχα τους εμπειρικούς εκτιμητές των $Q(t)$ και $E(\exp(\beta X), t)$. Ο εκτιμητής μέγιστης μερικής πιθανοφάνειας των παραμέτρων β συμβολίζεται με $\hat{\beta}$. Έτσι,

$$\hat{Q}(t) = \sum_{i=1}^n \frac{I_{\{Y_i \geq t, \delta_i=1\}}}{n}.$$

$$\hat{E}\{\exp(\beta X), t\} = \sum_{i=1}^n \frac{\exp(\beta X_i) I_{\{Y_i \geq t\}}}{n} = \sum_{j \in R(t)} \frac{\exp(\beta X_j)}{n}.$$

Το σύνολο των δεικτών ($i = 1, \dots, n$) που αντιστοιχούν σε χρόνους ζωής ατόμων που δεν έχουν βιώσει το υπό μελέτη γεγονός, ούτε έχουν λογοκριθεί, μέχρι την χρονική στιγμή t συμβολίζεται με $R(t)$.

Ασυμπτωτικές ιδιότητες των εκτιμητών $\hat{\beta}$.

Όπως είπαμε ο εκτιμητής της παραμέτρου β προκύπτει από τη μεγιστοποίηση της συνάρτησης μερικής πιθανοφάνειας

$$L_c(\beta) = \prod_{i \in D} \frac{\exp(\beta X_i)}{\sum_{j \in R(t_i)} \exp(\beta X_j)/n}$$

ή ισοδύναμα τη μεγιστοποίηση του λογαρίθμου της μερικής πιθανοφάνειας

$$\log L_c(\beta) = \beta \sum_{i \in D} \frac{X_i}{n} - \sum_{i \in D} \frac{1}{n} \log \left\{ \sum_{j \in R(t_i)} \exp(\beta X_j)/n \right\}, \quad (2.2.10)$$

όπου D συμβολίζει το σύνολο των δεικτών, $i = 1, \dots, n$, που αντιστοιχούν σε χρόνους ζωής ατόμων που βίωσαν το υπό μελέτη γεγονός, δηλαδή τους μη λογοκριμένους. Ο εκτιμητής $\hat{\beta}$ είναι η λύση της εξίσωσης

$$\sum_{i \in D} \frac{X_i}{n} - \sum_{i \in D} \frac{1}{n} \frac{\sum_{j \in R(t_i)} X_j \exp(\beta X_j)/n}{\sum_{j \in R(t_i)} \exp(\beta X_j)/n} = 0, \quad (2.2.11)$$

ως προς β . Η σχέση (2.2.11) προκύπτει από την παραγώγιση του δεξιού μέλους της σχέσης (2.2.10) ως προς β και την εξίσωση του με μηδέν. Πριν δώσουμε τις ασυμπτωτικές ιδιότητες του $\hat{\beta}$ εκτιμητή του β , θα χρειαστούμε τις ακόλουθες υποθέσεις.

A.1 Θεωρούμε ότι για τη συμμεταβλητή X ισχύει ότι η $E[X \exp(\beta X)]^2$ είναι ομοιόμορφα φραγμένη σε μία γειτονιά του β .

A.2 Η T_0 είναι η χρονική στιγμή κατά την οποία τελειώνει η έρευνα και τέτοια ώστε $P(Y \geq T_0) > 0$.

Παρόλο που η υπόθεση A.2 δεν είναι απαραίτητη για την απόδειξη που θα κάνουμε, ισχύει στις περισσότερες μελέτες.

Παρατηρούμε ότι αν αντικαταστήσουμε στην σχέση (2.2.10) τις άγνωστες ποσότητες με τους εμπειρικούς εκτιμητές τους η εξίσωση γράφεται ως:

$$\hat{H}(\beta) = \beta \hat{E}_1(X, 0) - \int_0^{T_0} -d\hat{Q}(t) \log \hat{E}(\exp(\beta X), t), \quad (2.2.12)$$

και η εξίσωση (2.2.11) ως:

$$F(\beta) = \frac{\hat{H}(\beta)}{d\beta} = \hat{E}_1(X, 0) - \int_0^{T_0} -d\hat{Q}(t) \frac{\hat{E}\{X \exp(\beta X), t\}}{\hat{E}\{\exp(\beta X), t\}} = 0, \quad (2.2.13)$$

με $\hat{E}_1(X, 0)$ να είναι ο εμπειρικός εκτιμητής της $E_1(X, 0)$.

Η σχεδόν βέβαιη σύγκλιση του $\hat{\beta}$, που προκύπτει από την επίλυση της σχέσης (2.2.11), ως προς την πραγματική παράμετρο β , θεμελιώνεται στο παρακάτω θεώρημα, το οποίο παρουσιάστηκε και αποδείχτηκε στην εργασία Tsiatis (1981) (Theorem 3.1).

Θεώρημα 2.2.1:

Υπάρχει μια ακολουθία λύσεων $\hat{\beta}(n)$ της εξίσωσης (2.2.11) τέτοια ώστε να συγκλίνει σχεδόν σίγουρα (almost surely a.s) στο β .

Απόδειξη:

Θεωρούμε τη συνάρτηση

$$H(\beta) = \beta E_1(X, 0) - \int_0^{T_0} -dQ(t) \log E\{\exp(\beta X), t\}.$$

Με χρήση της (2.2.9) η πρώτη παράγωγος είναι:

$$\frac{dH(\beta)}{d\beta} = E_1(X, 0) - \int_0^{T_0} -dQ(t) E\{X \exp(\beta X), t\} / E\{\exp(\beta X), t\} = 0.$$

Η δεύτερη παράγωγος είναι:

$$\begin{aligned} \frac{d^2 H(\beta)}{d\beta^2} &= - \int_0^{T_0} -dQ(t) \left[\frac{E\{X^2 \exp(\beta X), t\}}{E\{\exp(\beta X), t\}} - \left(\frac{E\{X \exp(\beta X), t\}}{E\{\exp(\beta X), t\}} \right)^2 \right] \\ &= - \int_0^{T_0} -dQ(t) \left(\frac{E[X - E\{X|R(t)\}]^2 \exp(\beta X), t\}}{E\{\exp(\beta X), t\}} \right) < 0 \end{aligned}$$

με

$$E\{X|R(t)\} = \frac{E\{X \exp(\beta X), t\}}{E\{\exp(\beta X), t\}}.$$

Συνεπώς, η συνάρτηση $H(t)$ έχει τοπικό μέγιστο στο $t = \beta$. Αυτό συνεπάγεται ότι για β^* σε μια γειτονιά ε του β ($|\beta^* - \beta| \leq \varepsilon$) ισχύει

$$H(\beta) - H(\beta^*) \geq 0,$$

με την ισότητα ισχύει όταν $|\beta^* - \beta| = \varepsilon$. Με χρήση των Λημμάτων 1 και 2, που παρατίθενται στο τέλος αυτής της ενότητας, καθώς και του ισχυρού νόμου των μεγάλων αριθμών παίρνουμε ότι

$$\hat{H}(\beta) - \hat{H}(\beta^*) \xrightarrow{a.s.} H(\beta) - H(\beta^*).$$

Επομένως, για σχεδόν όλα τα β^* υπάρχει ένα n_0 τέτοιο ώστε για κάθε $n \geq n_0$

$$\hat{H}(\beta^*) < \hat{H}(\beta) \text{ για } |\beta^* - \beta| = \varepsilon.$$

Η συνάρτηση $\hat{H}(t)$ είναι συνεχής και διαφορίσιμη, οπότε στο σύνολο $|\beta^* - \beta| \leq \varepsilon$, η $\hat{H}(\beta^*)$ έχει μέγιστο που δεν είναι στο σύνορο. Αυτό σημαίνει ότι έχει τοπικό μέγιστο και η πρώτη παράγωγος μηδενίζεται, δηλαδή

$$\frac{d\hat{H}(\beta)}{d\beta} = 0,$$

που είναι ακριβώς η λύση της εξίσωσης (2.2.13). Αυτός ο ισχυρισμός μπορεί να επαναληφθεί για μπάλες μεγέθους ε που γίνονται όλο και μικρότερες. Έτσι, βρίσκουμε μια συνεπή ακολουθία $\beta^*(n) \xrightarrow{a.s.} \beta$. $\square$

Για να δείξουμε ότι η στατιστική συνάρτηση $\sqrt{n}(\hat{\beta} - \beta)$ συγκλίνει σε μια κανονική κατανομή, θα την προσεγγίσουμε με ένα άθροισμα τυχαίων μεταβλητών που συγκλίνουν από κοινού σε μια πολυδιάστατη κανονική κατανομή. Χρησιμοποιώντας το ανάπτυγμα Taylor έχουμε:

$$\begin{aligned} F(\hat{\beta}) - F(\beta) - (\hat{\beta} - \beta) \sum_{i \in D} n^{-1} \hat{Var}\{X|R(t_i)|_{\beta} \\ + \frac{(\hat{\beta} - \beta)^2}{2} \sum_{i \in D} n^{-1} \hat{E} \left([X - \hat{E}\{X|R(t_i)\}]^3 | R(t_i) \right) |_{\beta^*}, \end{aligned}$$

όπου το β^* είναι μεταξύ των $\hat{\beta}$ και β , ενώ με $\hat{Var}\{X|R(t_i)|_{\beta}$ συμβολίζουμε την εκτιμώμενη υπό συνθήκη διακύμανση της μεταβλητής X , δοθέντος του συνόλου $R(t_i)$, υπολογισμένη για την τιμή β της παραμέτρου. Αντίστοιχα, $\hat{E} \left([X - \hat{E}\{X|R(t_i)\}]^3 | R(t_i) \right) |_{\beta^*}$ είναι η ποσότητα $\hat{E} \left([X - \hat{E}\{X|R(t_i)\}]^3 | R(t_i) \right)$ υπολογισμένη στην τιμή β^* της παραμέτρου. Σημειώνουμε ότι τα αθροίσματα στο

ανάπτυγμα Taylor αντιστοιχούν στους εμπειρικούς εκτιμητές της πρώτης και της δεύτερης παραγώγου της συνάρτησης $F(\beta)$ με την αντίστοιχη σειρά. Τώρα,

$$\hat{E}\{h(X)|R(t_i)\}_{|\beta^*} = \sum_{j \in R(t_j)} \frac{h(X_j) \exp(\beta^* X_j)}{\sum_{R(t_j)} \exp(\beta^* X_j)},$$

και

$$\hat{Var}\{X|R(t_i)\}_{|\beta^*} = \sum_{R(t_i)} \frac{X_j^2 \exp(\beta^* X_j)}{\sum_{R(t_i)} \exp(\beta^* X_j)} - \left\{ \frac{\sum_{R(t_i)} X_j \exp(\beta^* X_j)}{\sum_{R(t_i)} \exp(\beta^* X_j)} \right\}^2.$$

Οπότε, με χρήση του αναπτύγματος Taylor και των παραπάνω σχέσεων, η στατιστική συνάρτηση $\sqrt{n}(\hat{\beta} - \beta)$ μπορεί να εκφραστεί ως:

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{C_n}{\int_0^{T_0} -dQ(t) Var\{X|R(t)\}} \right) - F_n,$$

με

$$\begin{aligned} C_n &= \sqrt{n} \left[\hat{E}_1(X, 0) - \int_0^{T_0} -d\hat{Q}(t) \hat{E}_1\{X|R(t)\} \right], \\ F_n &= C_n \frac{F_{2n}}{F_{3n}}, \\ F_{2n} &= \int_0^{T_0} -d\hat{Q}(t) \hat{Var}\{X|R(t)\} - \int_0^{T_0} -dQ(t) Var\{X|R(t)\} - \frac{R_n(\hat{\beta} - \beta)}{2}, \\ F_{3n} &= \left[\int_0^{T_0} -d\hat{Q}(t) \hat{Var}\{X|R(t)\} - R_n \frac{(\hat{\beta} - \beta)}{2} \right] \int_0^{T_0} -dQ(t) Var\{X|R(t)\}, \\ R_n &= \sum_{i \in D} n^{-1} \hat{E} \left([X - \hat{E}\{X|R(t_i)\}]^3 | R(t_i) \right)_{|\beta^*}. \end{aligned}$$

Χρησιμοποιώντας τη συγκεκριμένη μορφή της στατιστικής συνάρτησης $\sqrt{n}(\hat{\beta} - \beta)$ προκύπτει το επόμενο θεώρημα, το οποίο παρουσιάστηκε και αποδείχτηκε στην εργασία Tsiatis (1981).

Θεώρημα 2.2.2

Η στατιστική συνάρτηση $\sqrt{n}(\hat{\beta} - \beta)$ συγκλίνει κατά κατανομή σε μια κανονική τυχαία μεταβλητή με μέση τιμή 0 και διακύμανση ίση με $\left[\int_0^{T_0} -dQ(t) Var\{X|R(t)\} \right]^{-1}$.

Απόδειξη:

Η απόδειξη αποτελείται από i) να δείξουμε ότι το στατιστικό συγκλίνει κατά

κατανομή στην $N\left(0, \int_0^{T_0} -dQ(t) \text{Var}\{X|R(t)\}\right)$ και ii) ότι το F_n συγκλίνει κατά πιθανότητα στο μηδέν.

i) Με χρήση της σχέσης (2.2.9) το στατιστικό C_n μπορεί να γραφεί ως:

$$C_n = \sqrt{n}\{\hat{E}_1(X, 0) - E_1(X, 0)\} - \sqrt{n} \left\{ \int_0^{T_0} -d\hat{Q}(t) \frac{\hat{E}_X(t)}{\hat{E}(t)} - \int_0^{T_0} -dQ(t) \frac{E_X(t)}{E(t)} \right\},$$

όπου $E_X(t) = E\{X \exp(\beta X), t\}$ και $E(t) = E\{\exp(\beta X), t\}$. Το στατιστικό C_n , και με μεθόδους που βασίζονται στην εργασία των Breslow and Crowley (1974), μπορεί να εκφραστεί ως:

$$C_n = C_{1n} + \dots + C_{6n} + R_{1n} + \dots + R_{4n},$$

με

$$\begin{aligned} C_{1n} &= \sqrt{n}\{\hat{E}_1(X, 0) - E_1(X, 0)\}, \\ C_{2n} &= -\sqrt{n}\{\hat{Q}(0) - Q(0)\} \frac{E_X(0)}{E(0)} \\ C_{3n} &= -\int_0^{T_0} \frac{\sqrt{n}\{\hat{Q}(t) - Q(t)\}}{E(t)} dE_X(t), \\ C_{4n} &= \int_0^{T_0} \frac{\sqrt{n}\{\hat{Q}(t) - Q(t)\} E_X(t)}{E^2(t)} dE(t), \\ C_{5n} &= \int_0^{T_0} \frac{\sqrt{n}\{\hat{E}_X(t) - E_X(t)\}}{E(t)} dQ(t), \\ C_{6n} &= -\int_0^{T_0} \frac{\sqrt{n}\{\hat{E}(t) - E(t)\} E_X(t)}{E^2(t)} dQ(t), \end{aligned}$$

και

$$\begin{aligned} R_{1n} &= \int_0^{T_0} \frac{\sqrt{n}\{\hat{E}_X(t) - E_X(t)\}}{\hat{E}(t)} d\{\hat{Q}(t) - Q(t)\}, \\ R_{2n} &= -\int_0^{T_0} \frac{\sqrt{n}\{\hat{E}_X(t) - E_X(t)\}\{\hat{E}(t) - E(t)\}}{\hat{E}(t)E(t)} dQ(t), \\ R_{3n} &= -\int_0^{T_0} \frac{\sqrt{n}\{\hat{E}(t) - E(t)\} E_X(t)}{\hat{E}(t)E(t)} d\{\hat{Q}(t) - Q(t)\}, \\ R_{4n} &= \int_0^{T_0} \frac{\sqrt{n}\{\hat{E}(t) - E(t)\}^2 E_X(t)}{\hat{E}(t)E(t)} dQ(t). \end{aligned}$$

Από την απόδειξη του Θεωρήματος 5.1, Tsiatis (1981), έχουμε ότι $(C_{1n}, \dots, C_{6n})$ συγκλίνει κατά κατανομή σε μια πολυδιάστατη κανονική τυχαία μεταβλητή που συμβολίζουμε ως $(C_1, \dots, C_6)$ και τα $R_{1n}, \dots, R_{4n}$ συγκλίνουν κατά πιθανότητα στο 0. Συνεπώς η στατιστική συνάρτηση $C_n = C_{1n} + \dots + C_{6n} + R_{1n} + \dots + R_{4n}$ συγκλίνει κατά κατανομή στην κανονική τυχαία μεταβλητή $\sum_{i=1}^6 C_i$, που έχει μέση τιμή μηδέν. Από Appendix 2, Tsiatis (1981), η διακύμανση αυτής της τυχαίας μεταβλητής είναι ίση με $\int_0^{T_0} -dQ \text{Var}\{X|R(t)\}$.

ii) Με χρήση των Λημμάτων 1 και 2 που παρατίθενται παρακάτω και της υπόθεσης A.1 μπορεί ναδειχθεί ότι:

$$\int_0^{T_0} -d\hat{Q}(t)\hat{\text{Var}}\{X|R(t)\} - \int_0^{T_0} -dQ(t)\text{Var}\{X|R(t)\} \xrightarrow{a.s} 0.$$

Η παραπάνω σχέση σε συνδυασμό με την συνέπεια του $\hat{\beta}$ συνεπάγεται ότι $F_{2n} \xrightarrow{a.s} 0$. Όμοια η ποσότητα F_{3n} συγκλίνει σχεδόν σίγουρα στο $[\int_0^{T_0} -dQ \text{Var}\{X|R(t)\}]^2$. Συνεπώς με εφαρμογή του θεωρήματος του Slutsky αποδεικνύεται ότι F_n συγκλίνει κατά πιθανότητα στο 0.

Σημείωση:

Η ασυμπτωτική διακύμανση του στατιστικού $\sqrt{n}(\hat{\beta} - \beta)$, έστω $\sigma_{\hat{\beta}}^2$, δίνεται από την σχέση:

$$\sigma_{\hat{\beta}}^2 = \left[\int_0^{T_0} -dQ(t)\text{Var}\{X|R(t)\} \right]^{-1}, \quad (2.2.16)$$

και μπορεί να εκτιμηθεί με αντικατάσταση των εμπειρικών εκτιμητών σε αυτήν. Δηλαδή:

$$\hat{\sigma}_{\hat{\beta}}^2 = \left[\int_0^{T_0} -d\hat{Q}(t)\hat{\text{Var}}\{X|R(t)\}|_{\hat{\beta}} \right]^{-1} = \left[\sum_{i \in D} \frac{1}{n} \hat{\text{Var}}\{X|R(t_i)\}|_{\hat{\beta}} \right]^{-1}. \quad (2.2.17)$$

Η σχέση (2.2.17) ισούται με μείον το αντίστροφο της δεύτερης παραγώγου της συνάρτησης μερικής πιθανοφάνειας του Cox.

Με σκοπό την παρουσίαση των Λημμάτων 1 και 2 θεωρούμε ότι $\hat{E}\{g(X), t\}$ είναι ο εμπειρικός εκτιμητής της $E\{g(X), t\}$, τέτοιος ώστε:

$$\hat{E}\{g(X), t\} = \sum_{j \in R(t)} \frac{g(X_j)}{n}.$$

Λήμμα 1:

Έστω ότι η X είναι μια τυχαία μεταβλητή τέτοια ώστε η $E\{g^2(X)\}$ να είναι

πεπερασμένη, όπου $g(X)$ είναι μια συνεχής συνάρτηση, τότε

$$\sup_{0 \leq t \leq M} |\hat{E}\{g(X), t\} - E\{g(X), t\}| \xrightarrow{a.s.} 0.$$

Λήμμα 2:

Έστω $\hat{X}_n(t), \hat{Y}_n(t), \hat{Q}_n(t)$ τυχαίες συναρτήσεις στο διάστημα $[0, M]$ που συγκλίνουν σχεδόν σίγουρα κατά sup νόρμα στις συναρτήσεις $X(t), Y(t), Q(t)$. Οι συναρτήσεις $X(t), Y(t)$ είναι συνεχείς στο διάστημα $[0, M]$. Η $Q(t)$ είναι συνεχής συνάρτηση και είναι θετική φθίνουσα με $Q(0) \leq 1$ (subsurvival function). Έστω $f(x, y) : \mathbb{R}^2 \rightarrow \mathbb{R}$ μια συνεχής συνάρτηση, τέτοια ώστε οι μερικές παράγωγοι $\frac{\partial f(x, y)}{\partial x}, \frac{\partial f(x, y)}{\partial y}$ να υπάρχουν και να είναι συνεχείς στο $R_X \times R_Y$, όπου $R_X \times R_Y$ είναι το καρτεσιανό γινόμενο των συνόλων τιμών των $X(t), Y(t)$ αντίστοιχα, $t \in [0, M]$. Δηλαδή $R_X = \{X(t) | t \in [0, M]\}$, $R_Y = \{Y(t) | t \in [0, M]\}$. Τότε

$$\sup_{0 \leq t \leq M} \left| \int_0^{T_0} f\{\hat{X}_n(t), \hat{Y}_n(t)\} \{-d\hat{Q}_n(t)\} - \int_0^{T_0} f(X, Y) \{-dQ(t)\} \right| \xrightarrow{a.s.} 0.$$

Στην ακόλουθη ενότητα παρουσιάζουμε το Likelihood Ratio Test (LRT), το οποίο μας επιτρέπει να επιλέξουμε μεταξύ δυο μοντέλων, του πλήρους και του μειωμένου. Το LRT αξιολογεί την ικανότητα του πλήρους μοντέλου να εξηγήσει τα δεδομένα σε σχέση με το μειωμένο μοντέλο, χρησιμοποιώντας τις εκτιμήσεις των παραμέτρων από τα δεδομένα. Με αυτό τον τρόπο, ελέγχεται η στατιστική σημαντικότητα των επιπλέον παραμέτρων, άρα και των συμμεταβλητών, που περιλαμβάνει το πλήρες μοντέλο.

2.3 Τεστ Πηλίκου Πιθανοφανειών (LRT)

Όταν έχουμε παραπάνω από μια συμμεταβλητές συχνά καλούμαστε να επιλέξουμε από όλα τα πιθανά μοντέλα που σχηματίζονται με τον συνδυασμό των διαθέσιμων συμμεταβλητών, ποιο ταιριάζει καλύτερα στα δεδομένα. Σε αυτό το ερώτημα απαντάμε με την χρήση του τεστ πηλίκου πιθανοφανειών, Likelihood Ratio Test (LRT). Θεωρούμε ότι έχουμε ένα μοντέλο στο οποίο έχουμε προσαρμόσει ένα πλήθος συμμεταβλητών (πλήρες μοντέλο) και ένα μοντέλο που προκύπτει από το αρχικό, αφού αφαιρέσουμε μια ή περισσότερες συμμεταβλητές (μειωμένο μοντέλο). Ο σκοπός μας είναι να προσδιορίσουμε ποιο από τα δύο μοντέλα προσαρμόζεται καλύτερα στα δεδομένα, με χρήση ακριβώς όσων μεταβλητών είναι απαραίτητες.

Γενική μορφή του Τεστ Πηλίκου Πιθανοφανειών (LRT)

Για να περιγράψουμε το LRT και την θεωρία που ακολουθεί βασιζόμαστε στο σύγγραμμα των Bickel and Doksum (2001). Θεωρούμε ότι έχουμε ένα τυχαίο δείγμα $X_1, \dots, X_n$ από έναν πληθυσμό με συνάρτηση πυκνότητας πιθανότητας $f(x|\beta)$, όπου $\beta \in \Omega$, Ω είναι ο παραμετρικός χώρος του β . Η συνάρτηση πιθανοφάνειας είναι:

$$L(\beta) = L(\beta|\mathbf{X}) = \prod_{i=1}^n f(X_i|\beta),$$

με $\mathbf{X} = (X_1, \dots, X_n)'$. Έστω Ω_0 ο χώρος στον οποίο ισχύει η H_0 (null space). Τότε $\Omega = \Omega_0 \cup \Omega_0^c$.

Η επιλογή του πλήρους μοντέλου ή του μειωμένου μοντέλου αντιστοιχεί στον έλεγχο της υπόθεσης

$$\begin{aligned} H_0 : \beta &\in \Omega_0 \text{ (μειωμένο μοντέλο) έναντι της} \\ H_A : \beta &\in \Omega_0^c \text{ (πλήρες μοντέλο)} \end{aligned} \quad (2.3.1)$$

Ο έλεγχος της (2.3.1) γίνεται με τη στατιστική συνάρτηση

$$r(\mathbf{X}) = \frac{\sup_{\Omega_0} L(\beta)}{\sup_{\Omega} L(\beta)}.$$

Το τεστ μπορεί να εφαρμοστεί σε πολύ γενικές περιπτώσεις. Για παράδειγμα όταν $\beta = (\beta_1, \beta_2, \beta_3)'$ για ελέγχους της μορφής $H_0 : (\beta_1 = 0, \beta_2, \beta_3)$, έναντι $H_A : (\beta_1 \neq 0, \beta_2, \beta_3)$, όταν οι τιμές των παραμέτρων β_2, β_3 δεν αποτελούν μέρος του ελέγχου. Τις περισσότερες φορές η εύρεση των **supremum** της στατιστικής συνάρτησης, τόσο στον αριθμητή όσο και στον παρονομαστή, βασίζεται στην εύρεση του εκτιμητή μέγιστης πιθανοφάνειας (Ε.Μ.Π.) του β στα σύνολα Ω_0 και Ω αντίστοιχα και τον υπολογισμό της τιμής της συνάρτησης πιθανοφάνειας για τον εκτιμητή. Αυτό σημαίνει ότι για τον παρονομαστή βρίσκουμε τον Ε.Μ.Π. του β σε όλον τον παραμετρικό χώρο Ω , έστω $\hat{\beta}_n$ και τον αντικαθιστούμε στην $L(\beta)$. Οπότε στον παρονομαστή στην θέση του $\sup_{\Omega} L(\beta)$ θα έχουμε $L(\hat{\beta}_n)$. Αντίστοιχα για τον αριθμητή μεγιστοποιούμε την $L(\beta)$ στον χώρο Ω_0 , δηλαδή βρίσκουμε τον Ε.Μ.Π. του β στον παραμετρικό χώρο Ω_0 , έστω $\hat{\beta}_{0,n}$ και στη συνέχεια αντικαθιστούμε στον αριθμητή $L(\hat{\beta}_{0,n})$. Καθώς ισχύει πάντα $r(\mathbf{X}) \leq 1$, περιμένουμε μικρές τιμές του $r(\mathbf{X})$ να συνεπάγονται απόρριψη της H_0 . Ισοδύναμα, απορρίπτουμε την H_0 για μεγάλες τιμές της ποσότητας

$$r^* = -2 \log r(\mathbf{X}).$$

Ασυμπτωτική κατανομή της στατιστικής συνάρτησης r^* .

Υπό την $H_0 : \beta \in \Omega_0$ η κατανομή της r^* συγκλίνει στην $\chi^2_{(df)}$ για $n \rightarrow \infty$. Οι βαθμοί ελευθερίας είναι $(df) = (\text{το πλήθος των ελεύθερων παραμέτρων στον χώρο } \Omega) - (\text{το πλήθος των ελεύθερων παραμέτρων στον χώρο } \Omega_0)$. Συνεπώς,

$$r^* \sim \chi^2_{(df)}.$$

Αρχικά, θεωρούμε τα παρακάτω. Έστω $H_0 : \beta \in \Omega_0$, με Ω ανοικτό σύνολο με διάσταση r , και Ω_0 το σύνολο των $\beta \in \Omega$ που είναι τέτοια ώστε $\beta_j = \beta_{0j}$, $j = q+1, \dots, r$ με τα β_{0j} να είναι γνωστά. Ορίζουμε $d = r - q$, $\beta = (\beta_1, \beta_2)'$, $\beta_1 = (\beta_1, \dots, \beta_q)'$, $\beta_2 = (\beta_{q+1}, \dots, \beta_r)'$, $\beta_{0,2} = (\beta_{0,q+1}, \dots, \beta_{0,r})'$. Χρησιμοποιούμε τον συμβολισμό

$$l_n(\beta) = \sum_{i=1}^n \log f(X_i|\beta).$$

Επιπλέον, θεωρούμε ότι $\hat{\beta}_n$ είναι ο εκτιμητής μέγιστης πιθανοφάνειας του β στο Ω , $\hat{\beta}_n = (\hat{\beta}_{0,1}, \hat{\beta}_{0,2})'$ και $\hat{\beta}_{0,n}$ είναι ο εκτιμητής μέγιστης πιθανοφάνειας του β υπό την H_0 , $\hat{\beta}_{0,n} = (\hat{\beta}_{0,1}, \beta_{0,2})'$ με $\beta_{0,2}$ γνωστό. Στο Θεώρημα 2.3.1 που ακολουθεί χρησιμοποιούμε τον συμβολισμό

$$\lambda(\mathbf{X}) = \frac{1}{r(\mathbf{X})} = \frac{\sup_{\Omega} L(\beta)}{\sup_{\Omega_0} L(\beta)}$$

$$\implies 2 \log\{\lambda(\mathbf{X})\} = -2 \log\{r(\mathbf{X})\}.$$

Ακόμη, αν $h : O \rightarrow \mathbb{R}^p$, με O ανοικτό $\subset \mathbb{R}^d$, $h \equiv (h_1, \dots, h_p)$, η h είναι Frechet διαφορίσιμη στο $\mathbf{x} \in O$ αν και μόνο αν υπάρχει, υποχρεωτικά μοναδική, γραμμική απεικόνιση $Dh(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}^p$ τέτοια ώστε:

$$|h(\mathbf{y}) - h(\mathbf{x}) - Dh(\mathbf{x})(\mathbf{y} - \mathbf{x})| = o(|\mathbf{y} - \mathbf{x}|),$$

όπου $|\cdot|$ είναι η ευκλείδεια νόρμα. Αν $p = 1$, Dh είναι η συνολική διαφορισμότητα.

Έστω $X_1, \dots, X_n$ ανεξάρτητες και ισόνομες τυχαίες μεταβλητές που ακολουθούν κατανομή που ανήκει σε μια οικογένεια κατανομών $P \equiv \{P_\beta : \beta \in \Omega\}$, $P \in \mathcal{L}$, τέτοια ώστε:

- Το Ω είναι ανοικτό υποσύνολο του $\mathbb{R}^r$.
- Οι συναρτήσεις πυκνότητας πιθανότητας της P_β είναι $f(X|\beta)$, με $\beta \in \Omega$.

Οι παρακάτω υποθέσεις χρειάζονται για την απόδειξη του Θεωρήματος 2.3.1.

B.0 Το διάνυσμα $\Psi = (\psi_1, \dots, \psi_r)'$, με $\psi_j = \frac{\partial \log\{f(x|\beta)\}}{\partial \beta_j}$, $j = 1, \dots, r$, είναι καλά ορισμένο και

$$\frac{1}{n} \sum_{i=1}^n \Psi(X_i, \bar{\beta}_n) = \mathbf{0}.$$

B.1 Η παράμετρος $\beta(P)$ δίνεται από τη λύση της εξίσωσης

$$\int \Psi(x, \beta) dP(x) = 0,$$

η οποία είναι καλά ορισμένη στο $\mathcal{L}$, έτσι ώστε η $\beta(P)$ να είναι η μοναδική λύση της. Υποχρεωτικά $\beta(P_\beta) = \beta$, γιατί $P \subset \mathcal{L}$.

B.2 $E_P |\Psi\{X_1, \beta(P)\}|^2 < \infty$, όπου $|\cdot|$ είναι η ευκλείδεια νόρμα.

B.3 Για τα $\psi_i(x, \beta)$, $i = 1, \dots, r$, υπάρχουν οι πρώτου βαθμού παράγωγοι ως προς β και

$$E_P |D\Psi(X_1, \theta)| < \infty,$$

όπου $E_P D\Psi(X_1, \beta) = \left\| E_P \frac{\partial \psi_i(X_1, \beta)}{\partial \beta_j} \right\|_{r \times r}$ είναι αντιστρέψιμη (nonsingular).

B.4 $\sup\{|\frac{1}{n} \sum_{i=1}^n [D\Psi(X_i, t) - D\Psi\{X_i, \beta(P)\}]| : |t - \beta(P)| \leq \varepsilon_n\} \xrightarrow{P} 0$ αν $\varepsilon_n \rightarrow 0$.

B.5 $\bar{\beta}_n \xrightarrow{P} \beta(P)$, για κάθε $P \in \mathcal{L}$. Υπό τις B.0-B.5

$$\bar{\beta}_n = \beta(P) + \frac{1}{n} \sum_{i=1}^n \tilde{\Psi}\{X_i, \beta(P)\} + o_p(n^{-1/2}),$$

όπου

$$\tilde{\Psi}\{x, \beta(P)\} = -[E_P D\Psi\{X_1, \beta(P)\}]^{-1} \Psi\{x, \beta(P)\}.$$

B.6 Αν η $l(X, \beta) = \log f(X|\beta)$ είναι διαφορίσιμη ως προς β

$$\begin{aligned} E_\beta D\Psi(X_1, \beta) &= -E_\beta \Psi(X_1, \beta) D l(X_1, \beta) \\ &= -Cov_\beta\{\Psi(X_1, \beta), D l(X_1, \beta)\}. \end{aligned}$$

Τα ακόλουθα αποτελέσματα δίνουν μια γενική ασυμπτωτική συμπεριφορά των λύσεων των εκτιμώμενων εξισώσεων. Το Θεώρημα 2.3.1 εμφανίστηκε και αποδείχτηκε στο σύγγραμμα των Bickel and Doksum (2001).

Θεώρημα 2.3.1:

Υποθέτουμε ότι ισχύουν οι υποθέσεις B.0-B.6 για τις συναρτήσεις $f(X_i|\beta)$, $\beta \in \Omega$. Έστω P_0 το μοντέλο $\{P_\beta : \beta \in \Omega_0\}$, με την αντίστοιχη παραμετρικοποίηση $\beta_1 = (\beta_1, \dots, \beta_q)'$. Υποθέτουμε ότι $\hat{\beta}_{0,1}$ είναι ο εκτιμητής μέγιστης πιθανοφάνειας του β_1 υπό την H_0 και ικανοποιεί την υπόθεση B.6 για το P_0 . Έστω $\hat{\beta}_{0,n} = (\hat{\beta}_{0,1}, \beta_{0,2})'$. Τότε υπό την $H_0 : \beta \in \Omega_0$

$$\begin{aligned} 2 \log \lambda(\mathbf{X}) &\equiv 2 \left[\sum_{i=1}^n \log f(X_i|\hat{\beta}_n) - \sum_{i=1}^n \log f(X_i|\hat{\beta}_{0,n}) \right] \\ &= 2[l_n(\hat{\beta}_n) - l_n(\hat{\beta}_{0,n})] \sim \chi_{r-q}^2. \end{aligned}$$

Απόδειξη:

Για $\beta_0 \in \Omega_0$ έχουμε:

$$2 \log \lambda(\mathbf{X}) = 2[l_n(\hat{\beta}_n) - l_n(\beta_0)] - 2[l_n(\hat{\beta}_{0,n}) - l_n(\beta_0)].$$

Καθώς ισχύουν οι συνθήκες $B_0 - B_6$ για το μοντέλο με συνάρτηση πυκνότητας πιθανότητας $f(X_i|\beta)$, $\beta \in \Omega$, δηλαδή το P , συνεπάγεται ότι οι υποθέσεις $B_0 - B_5$ θα ισχύουν για το μοντέλο με συνάρτηση πυκνότητας πιθανότητας $f(X_i|\beta)$, $\beta \in \Omega_0$, δηλαδή το P_0 .

Επιπλέον ισχύει ότι:

- Αν ισχύουν οι υποθέσεις $B_0 - B_6$ για την $l(X, \beta) = \log f(X|\beta)$ ο εκτιμητής μέγιστης πιθανοφάνειας $\hat{\beta}_n$ του β ικανοποιεί την εξίσωση:

$$\hat{\beta}_n = \beta + \frac{1}{n} \sum_{i=1}^n I^{-1}(\beta) D l(X_i, \beta) + o_p(n^{-1/2}), \quad (2.3.2)$$

όπου $I^{-1}(\beta)$ είναι το μέτρο πληροφορίας του Fisher.

- Αν ισχύουν οι υποθέσεις $B_0 - B_6$ για την $l(X, \beta) = \log f(X|\beta)$, τότε υπό την $H_0 : \beta = \beta_0$ ισχύει η εξίσωση:

$$2[l_n(\hat{\beta}_n) - l_n(\beta_0)] = n(\hat{\beta}_n - \beta_0)' I_n(\beta_n^*)(\hat{\beta}_n - \beta_0), \quad (2.3.3)$$

για κάποιο β_n^* τέτοιο ώστε, $|\beta_n^* - \hat{\beta}_n| \leq |\hat{\beta}_n - \beta_0|$ και

$$I_n(\beta) = \left\| -\frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \beta_k} \frac{\partial}{\partial \beta_j} \log f(X_i|\beta) \right\|_{r \times r}.$$

Επιπλέον, καθώς

$$|\beta_n^* - \beta_0| \leq |\beta_n^* - \hat{\beta}_n| + |\hat{\beta}_n - \beta_0| \leq 2 |\hat{\beta}_n - \beta_0|,$$

ισχύει ότι $I_n(\beta_n^*) \xrightarrow{P} EI_n(\beta_0) = I_n(\beta_0)$, όπου με P συμβολίζουμε την σύγκλιση κατά πιθανότητα.

Με χρήση των σχέσεων (2.3.2) και (2.3.3) για το $\hat{\beta}_n$ και χρησιμοποιώντας το αντίστοιχο επιχείρημα για τα $\hat{\beta}_{0,1}, \hat{\beta}_{0,n}$ παίρνουμε ότι

$$2 \log \lambda(\mathbf{X}) = S(\beta_0)' I^{-1}(\beta_0) S(\beta_0) - S_1(\beta_0)' I_0^{-1}(\beta_0) S_1(\beta_0) + o_p(1), \quad (2.3.4)$$

με

$$S(\beta_0) = n^{-1/2} \sum_{i=1}^n Dl(X_i, \beta)$$

και $S = (S_1, S_2)'$ με το S_1 να έχει τις πρώτες q μεταβλητές του S . Ακόμη,

$$I_0(\beta_0) = \text{Var}_{\beta_0} S_1(\beta_0).$$

Έστω η αλλαγή μεταβλητής

$$\boldsymbol{\eta} = M(\boldsymbol{\beta} - \boldsymbol{\beta}_0),$$

όπου αγνοώντας στον συμβολισμό του M την εξάρτηση από το β_0 ,

$$M = KI^{1/2},$$

ο K είναι ένας ορθογώνιος πίνακας τέτοιος ώστε, αν $\Lambda_0 \equiv \{\boldsymbol{\beta} - \boldsymbol{\beta}_0 : \boldsymbol{\beta} \in \Omega_0\}$ $M\Lambda_0 = \{\boldsymbol{\eta} : \eta_{q+1} = \dots = \eta_r = 0, \boldsymbol{\eta} \in M\Omega\}$. Από τον τρόπο ορισμού το $\lambda(\mathbf{X})$ είναι αναλλοίωτο υπό επαναπαραμετροποίηση

$$\lambda(\mathbf{X}) = \gamma(\mathbf{X}), \quad (2.3.5)$$

όπου

$$\gamma(\mathbf{X}) = \frac{\sup_{\boldsymbol{\eta}} \{f(X_i | \beta_0 + M^{-1}\boldsymbol{\eta})\}}{\sup \{f(X_i | \beta_0 + M^{-1}\boldsymbol{\eta}) : \beta_0 + M^{-1}\boldsymbol{\eta} \in \Omega_0\}}.$$

Συμβολίζουμε με $D_{\boldsymbol{\eta}}$ την παραγώγιση ως προς $\boldsymbol{\eta}$ και με $D_{\boldsymbol{\beta}}$ την παραγώγιση ως προς $\boldsymbol{\beta}$. Με χρήση του κανόνα της αλυσίδας έχουμε:

$$D_{\boldsymbol{\eta}} l(\mathbf{X}, \beta_0 + M^{-1}\boldsymbol{\eta}) = [M^{-1}]' D_{\boldsymbol{\beta}} l(\mathbf{X}, \boldsymbol{\beta}).$$

Αν

$$T(\boldsymbol{\eta}) \equiv n^{-1/2} \sum_{i=1}^n D_{\boldsymbol{\eta}} l(X_i, \boldsymbol{\beta}_0 + M^{-1} \boldsymbol{\eta}),$$

τότε

$$\text{Var}\{T(\mathbf{0})\} = K' I^{-1/2} I I^{-1/2} K = J. \quad (2.3.6)$$

Επιπλέον, με όρους του $\boldsymbol{\eta}$, η H_0 είναι $\{\boldsymbol{\eta} \in M\Omega : \eta_{q+1} = \dots = \eta_r = 0\}$, με χρήση της σχέσης (2.3.4) για το $\gamma(\mathbf{X})$ έχουμε:

$$\begin{aligned} 2 \log \gamma(\mathbf{X}) &= T(\mathbf{0})' T(\mathbf{0}) - T_1(\mathbf{0})' T_1(\mathbf{0}) + o_p(1) \\ &= \sum_{i=1}^r T_i^2(0) - \sum_{i=1}^q T_i^2(0) + o_p(1) \\ &= \sum_{i=q+1}^r T_i^2(0) + o_p(1), \end{aligned}$$

που έχει οριακή κατανομή χ^2_{r-q} από το θεώρημα του Slutsky, καθώς, από την σχέση (2.3.6), το $T(\mathbf{0})$ έχει οριακή κατανομή $N(\mathbf{0}, J)$. Η απόδειξη του θεωρήματος ολοκληρώθηκε βάσει της σχέσης (2.3.5). $\square$

2.4 Εφαρμογή

Σε αυτήν την ενότητα παρουσιάζουμε ένα παράδειγμα εφαρμογής του μοντέλου του Cox με χρήση μιας συμμεταβλητής. Το παράδειγμα βασίζεται στα δεδομένα CNS. Σκοπός του παραδείγματος είναι να δείξουμε πως γίνεται η εκτίμηση της μερικής πιθανοφάνειας και η προσαρμογή του μοντέλου αναλογικών κινδύνων στην R. Η μεταβλητή **B3TODeath** δηλώνει είτε τον χρόνο ζωής, είτε τον χρόνο λογοκρισίας των ασθενών από την πρώτη BBBD χημειοθεραπεία. Η μεταβλητή **STATUS** αντιστοιχεί στο δείκτη λογοκρισίας κάθε ατόμου. Παίρνει την τιμή 0 αν ο χρόνος ζωής που παρατηρούμε είναι λογοκριμένος και την τιμή 1 όταν παρατηρούμε τον ακριβή χρόνο ζωής. Σύμφωνα με τον συμβολισμό που έχουμε δώσει **B3TODeath**= $Y = \min(T, C)$, και **STATUS**= C . Σκοπός μας είναι να εκτιμήσουμε πόσο πιο πιθανό είναι να επιβιώσει κάποιος ανάλογα με την ομάδα που ανήκει. Επομένως επιλέγουμε ως μοναδική μεταβλητή πρόβλεψης την **GROUP**.

Αρχικά, καλούμε τη βιβλιοθήκη **survival**, καθώς περιέχει τις συναρτήσεις **Surv** και **coxph**, οι οποίες υλοποιούν τη διαδικασία εκτίμησης των παραμέτρων του μοντέλου του Cox με τη μέθοδο μερικής πιθανοφάνειας. Γενικά στην R για να προσαρμόσουμε ένα μοντέλο αναλογικών κινδύνων χρησιμοποιούμε την εντολή **coxph**.

```
library(survival)
cns.cox<- coxph(Surv(B3TODEATH,STATUS)~GROUP)
```

Η εντολή `Surv` δέχεται ως πρώτο όρισμα τους χρόνους ζωής των ασθενών και ως δεύτερο όρισμα τη μεταβλητή που υποδηλώνει την κατάσταση του ασθενή, συνήθως η μεταβλητή αυτή παίρνει την τιμή 0 όταν ο ασθενής είναι ζωντανός και την τιμή 1 αν πέθανε, και δημιουργεί ένα αντικείμενο (object) που θα είναι το όρισμα της εντολής `coxph`. Η εντολή `coxph` προσαρμόζει το μοντέλο του Cox με μοναδική μεταβλητή πρόβλεψης την `GROUP`, οι μεταβλητές πρόβλεψης δηλώνονται δεξιά του συμβόλου `~`. Με την εντολή `coxph` γίνεται αυτόματα η διαδικασία υπολογισμού της μερικής πιθανοφάνειας και δίνονται οι εκτιμώμενοι συντελεστές. Για να δούμε την πληροφορία που περιέχεται στο αντικείμενο `cns.cox` δίνουμε την εντολή `summary(cns.cox)`, η οποία μας επιστρέφει στην οθόνη τις εξής πληροφορίες:

```
summary(cns.cox)
```

```
coxph(formula = Surv(B3TODEATH, STATUS) ~ GROUP)
```

```
n= 58, number of events= 36
```

	coef	exp(coef)	se(coef)	z	Pr(> z)
GROUP	1.0040	2.7292	0.3385	2.966	0.00301 **

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

	exp(coef)	exp(-coef)	lower .95	upper .95
GROUP	2.729	0.3664	1.406	5.298

```
Concordance= 0.614 (se = 0.042 )
```

```
Likelihood ratio test= 8.35 on 1 df, p=0.004
```

```
Wald test = 8.8 on 1 df, p=0.003
```

```
Score (logrank) test = 9.52 on 1 df, p=0.002
```

- Η εκτίμηση της παραμέτρου του μοντέλου για τη μεταβλητή `GROUP`, έστω $\hat{\beta}$ είναι 1.0040. Αυτό σημαίνει ότι οι ασθενείς της ομάδας 1, που έχουν λάβει ακτινοβολία πριν την BBBD χημειοθεραπεία, έχουν συνάρτηση κινδύνου $\exp(\hat{\beta}GROUP) = \exp(1.0040GROUP)$. Επιπλέον η βασική συνάρτηση

κινδύνου θεωρούμε ότι είναι σταθερή, δηλαδή μπορούμε να υποθέσουμε ότι $h_0(t) = 1$. Συνεπώς, οι ασθενείς της ομάδας 1 διατρέχουν μεγαλύτερο κίνδυνο, έχουν μικρότερο αναμενόμενο χρόνο ζωής σε σχέση με τους ασθενείς της ομάδας 0.

- Το τυπικό σφάλμα του εκτιμητή είναι 0.3385, που υποδεικνύει ότι έχουμε μια σχετικά ακριβή εκτίμηση του συντελεστή.
- Η επεξηγηματική μεταβλητή GROUP είναι στατιστικά σημαντική, γιατί για τον έλεγχο της υπόθεσης $H_0 : \beta = 0$ έχουμε p -τιμή = 0.00301 < 0.05, που σημαίνει ότι σε επίπεδο σημαντικότητας $\alpha = 0.05$ η $H_0 : \beta = 0$ απορρίπτεται.
- Για τη σύγκριση του μοντέλου με μοναδική μεταβλητή την GROUP με το μοντέλο με καμία επεξηγηματική μεταβλητή χρησιμοποιείται το LRT. Η τιμή του στατιστικού του LRT είναι 8.35, με 1 βαθμό ελευθερίας και p -τιμή = 0.004. Το LRT καταλήγει στην επιλογή του μοντέλου που περιέχει την GROUP ως επεξηγηματική μεταβλητή.

Τα παραπάνω συμπεράσματα καθώς και η εκτίμηση της ποσότητας $\hat{\beta}$ ισχύουν εφόσον ισχύει η υπόθεση της αναλογικότητας κινδύνων. Από μόνη της η εξέταση της υπόθεσης αναλογικότητας είναι θεμελιώδους σημασίας για την προσαρμογή του μοντέλου αναλογικών κινδύνων του Cox και για αυτό τον λόγο θα την εξετάσουμε αναλυτικά στην Ενότητα 2.7. Πριν προχωρήσουμε στον έλεγχο της υπόθεσης της αναλογικότητας κινδύνων, στην Ενότητα 2.5 παρουσιάζουμε την εκτίμηση Kaplan-Meier της συνάρτησης επιβίωσης, καθώς αυτή χρησιμοποιείται για τη γραφική αξιολόγηση της υπόθεσης. Στην Ενότητα 2.6 παρουσιάζουμε το Log-rank τεστ που χρησιμοποιείται για να ελέγξουμε αν δύο, ή και περισσότερες, συναρτήσεις επιβίωσης είναι ίδιες.

2.5 Kaplan-Meier εκτίμηση της συνάρτησης επιβίωσης

Σε αυτήν την ενότητα παρουσιάζουμε τον εκτιμητή Kaplan-Meier, ο οποίος προτάθηκε και πρωτοεμφανίστηκε στο άρθρο των Kaplan and Meier (1958) και είναι μια μη παραμετρική εκτίμηση της συνάρτησης επιβίωσης. Ο εμπειρικός εκτιμητής της συνάρτησης επιβίωσης ορίζεται ως $S_n(t) = \{\#T_i > t\}/n$, $i = 1, \dots, n$. Ο εκτιμητής Kaplan-Meier προσαρμόζει την εμπειρική εκτίμηση της συνάρτησης επιβίωσης, έτσι ώστε να λαμβάνεται υπόψη η δεξιά λογοχρισία. Η

συνάρτηση επιβίωσης, όπως έχουμε αναφέρει, ορίζεται ως $S(t) = P(T > t)$. Επιπλέον, γνωρίζουμε ότι για δύο γεγονότα A και B ισχύει $P(AB) = P(B|A)P(A)$. Επομένως, η πιθανότητα να συμβεί το υπό μελέτη γεγονός μετά την χρονική στιγμή u , δοθέντος ότι συμβαίνει μετά την χρονική στιγμή v μπορεί να γραφτεί ως:

$$\begin{aligned} P(Y \geq u) &= P(Y > u | Y > v)P(Y > v) \\ &= \{1 - P(Y \leq u | Y > v)\}P(Y > v) = S(u|v)S(v). \end{aligned} \quad (2.5.1)$$

Συνεπώς ο υπολογισμός της πιθανότητας $P(Y \geq u)$ είναι ισοδύναμος με τον υπολογισμό της $S(u|v)S(v)$ και η $S(v)$ μπορεί να υπολογιστεί αναδρομικά από την σχέση (2.5.1). Η μεταβλητή T του χρόνου ζωής λαμβάνει τιμές στο διάστημα $[0, \infty)$, όμως στην πράξη λαμβάνει τιμές στο διάστημα $[0, Y_{(n)})$ όπου $Y_{(n)}$ είναι η παρατήρηση με τη μεγαλύτερη τιμή. Μπορούμε να χωρίσουμε το διάστημα $[0, Y_{(n)})$ σε l υποδιαστήματα, $[0, t_1), [t_1, t_2), \dots, [t_{l-1}, t_l)$ με $t_l = Y_{(n)}$. Την τυχαία χρονική στιγμή $t \in [t_k, t_{k+1})$, με $k \in \{0, 1, \dots, l-1\}$, η συνάρτηση επιβίωσης είναι:

$$\begin{aligned} P(Y \geq t_k) &= P(Y > t_k | Y > t_{k-1})P(Y > t_{k-1}) \\ &= \{1 - P(Y \leq t_k | Y > t_{k-1})\}P(Y > t_{k-1}) \\ &= \{1 - P(Y \leq t_k | Y > t_{k-1})\}S(t_{k-1}) \\ &= \{1 - P(Y \leq t_k | Y > t_{k-1})\}\{1 - P(Y \leq t_{k-1} | Y > t_{k-2})\}S(t_{k-2}) \\ &= \dots \\ &= \{1 - P(Y \leq t_k | Y > t_{k-1})\} \dots \{1 - P(Y \leq t_0 | Y > t_{-1})\} \\ &= \prod_{j=1}^l S(t_j | t_{j-1}), \end{aligned} \quad (2.5.2)$$

με $\{1 - P(Y \leq t_0 | Y > t_{-1})\} = 1 - P(Y = t_0)$, με $t_0 = 0$. Για να μπορέσουμε να χρησιμοποιήσουμε την σχέση (2.5.2) πρέπει να εκτιμήσουμε τις πιθανότητες της μορφής $1 - P(Y \leq t_k | Y > t_{k-1})$ και να τις αντικαταστήσουμε στην (2.5.2). Με αυτόν το σκοπό, ορίζουμε τις ακόλουθες ποσότητες. Το πλήθος των, μη λογοκριμένων, γεγονότων στο διάστημα $(t_{k-1}, t_k]$ είναι:

$$d_k = \sum_{i=1}^n I_{\{t_{k-1} < Y_i \leq t_k, \delta_k=1\}}, \quad k = 1, \dots, l$$

Ο συνολικός αριθμός των ατόμων που έχουν χρόνο ζωής ή χρόνο λογοκρισίας μεγαλύτερο από t_{k-1} , δηλαδή ο συνολικός αριθμός των ατόμων που επιβιώνουν μετά την χρονική στιγμή t_{k-1} είναι:

$$n_k = \sum_{i=1}^n I_{\{Y_i \geq t_{k-1}\}}, \quad k = 1, \dots, l.$$

Οπότε, η πιθανότητα $1 - P(Y \leq t_k | Y > t_{k-1})$ μπορεί να εκτιμηθεί ως εξής:

$$1 - P(Y \leq t_k | Y > t_{k-1}) \simeq 1 - \frac{d_k}{n_k}. \quad (2.5.3)$$

Με επαναλαμβανόμενη αντικατάσταση της σχέσης (2.5.3) στην σχέση (2.5.2), προκύπτει ο μη παραμετρικός εκτιμητής Kaplan-Meier της συνάρτησης επιβίωσης, που ορίζεται ως:

$$S_{KM}(t) = \prod_{k:t_k \leq t} \left(1 - \frac{d_k}{n_k}\right).$$

Παράδειγμα υπολογισμού του Kaplan-Meier:

Θεωρούμε ότι έχουμε τους παρακάτω χρόνους επιβίωσης, όπου με + συμβολίζουμε τις λογοκριμένες παρατηρήσεις.

$$1, 2, 2+, 3, 3, 3+, 5+, 6+$$

Εφαρμόζουμε τον ορισμό του Kaplan-Meier, στα διαστήματα που προκύπτουν από τους χρόνους ζωής που μας δίνονται.

- $[0, 1]$, $t_1 = 1$, $d_1 = \#$ μη λογοκριμένων γεγονότων στο διάστημα $(0, 1] = 1$, $n_1 = \#$ των ατόμων που επιβιώνουν μετά την χρονική στιγμή t_0 , με $t_0 = 0 = 8$.

$$S_{KM}(t_1) = \left(1 - \frac{d_1}{n_1}\right) = \left(1 - \frac{1}{8}\right) = \frac{7}{8}.$$

- $(1, 2]$, $t_2 = 2$, $d_2 = 1$, $n_2 = 7$.

$$S_{KM}(t_2) = \left(1 - \frac{1}{8}\right) \left(1 - \frac{1}{7}\right) = \frac{6}{8}.$$

- $(2, 3]$, $t_3 = 3$, $d_3 = 2$, $n_3 = 5$.

$$S_{KM}(t_3) = \left(1 - \frac{1}{8}\right) \left(1 - \frac{1}{7}\right) \left(1 - \frac{2}{5}\right) = \frac{9}{20}.$$

- $(3, 5]$, $t_4 = 5$, $d_4 = 0$, $n_4 = 2$.

$$S_{KM}(t_4) = \left(1 - \frac{1}{8}\right) \left(1 - \frac{1}{7}\right) \left(1 - \frac{2}{5}\right) \left(1 - \frac{0}{2}\right) = \frac{9}{20}.$$

- $(5, 6], t_5 = 6, d_5 = 0, n_5 = 1.$

$$S_{KM}(t_5) = \left(1 - \frac{1}{8}\right) \left(1 - \frac{1}{7}\right) \left(1 - \frac{2}{5}\right) \left(1 - \frac{0}{2}\right) \left(1 - \frac{0}{1}\right) = \frac{9}{20}.$$

Η εκτίμηση Kaplan-Meier της συνάρτησης επιβίωσης είναι μια δεξιά συνεχής, σκαλωτή συνάρτηση, που παρουσιάζει πτώση μόνο στους χρόνους που αντιστοιχούν σε μη λογοκριμένες παρατηρήσεις. Για το παράδειγμα που παρουσιάστηκε, ο παρακάτω κώδικας δημιουργεί τη γραφική παράσταση της εκτιμώμενης συνάρτησης επιβίωσης. Η εντολή `Surv` δέχεται ως πρώτο όρισμα τους χρόνους ζωής και ως δεύτερο τους αντίστοιχους δείκτες λογοκρισίας. Το αποτέλεσμα είναι ένα αντικείμενο (object) που αποτελεί το πρώτο όρισμα της εντολής `survfit`. Η εντολή `survfit` υπολογίζει την εκτιμώμενη συνάρτηση επιβίωσης και δέχεται ως δεύτερο όρισμα το σύνολο των δεδομένων. Η εκτίμηση αυτή αποθηκεύεται στο αντικείμενο `surv`. Η εντολή `ggsurvplot` δημιουργεί τη γραφική παράσταση της εκτιμώμενης συνάρτησης επιβίωσης, χρησιμοποιώντας ως πρώτο όρισμα το `surv`, ως δεύτερο τα δεδομένα και επιπλέον δηλώνοντας ότι επιθυμούμε την εμφάνιση των διαστημάτων εμπιστοσύνης.

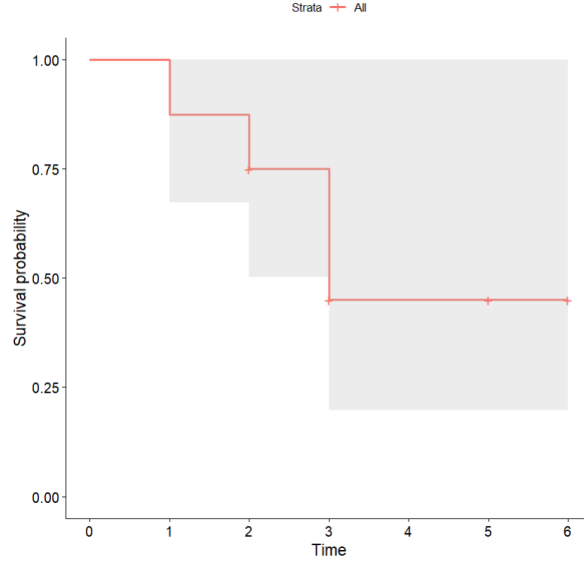
```
surv <- survfit(Surv(Time, Status) ~ 1, data = Km)
ggsurvplot(surv, data = Km, conf.int = TRUE,)
```

Αφού παρουσιάσαμε έναν τρόπο εκτίμησης της συνάρτησης επιβίωσης μέσω της μεθόδου Kaplan-Meier, μπορούμε να προχωρήσουμε στη σύγκριση των συναρτήσεων επιβίωσης μεταξύ διαφορετικών ομάδων. Για τον σκοπό αυτό, στην ακόλουθη ενότητα παρουσιάζουμε το Log-rank τεστ, το οποίο αποτελεί μια μη παραμετρική μέθοδο για το στατιστικό έλεγχο ισότητας των συναρτήσεων επιβίωσης.

2.6 Log-rank τεστ

Μια από τις πιο συνηθισμένες υποθέσεις που εξετάζονται στην ανάλυση επιβίωσης είναι αν δύο ή περισσότερες συναρτήσεις επιβίωσης είναι ίδιες. Το Log-rank τεστ χρησιμοποιείται για αυτόν τον έλεγχο. Στην παρούσα ενότητα, με βάση την εργασία του Mantel (1966), παρουσιάζουμε τη διαδικασία του ελέγχου για την περίπτωση δύο ομάδων, έστω των ομάδων 1 και 2, δηλαδή θέλουμε να ελέγξουμε την υπόθεση

$$\begin{aligned} H_0 : S_1(t) &= S_2(t), \text{ για κάθε } t, \text{ έναντι της} \\ H_A : S_1(t) &\neq S_2(t), \text{ για κάποιο } t. \end{aligned} \quad (2.6.1)$$



Σχήμα 2.1: Γραφική παράσταση της Kaplan-Meier εκτίμησης της συνάρτησης επιβίωσης του παραδείγματος.

Θεωρούμε ότι έχουμε r διακεκριμένους χρόνους $t_{(1)} < t_{(2)} < \dots < t_{(r)}$ στους οποίους συμβαίνουν τα υπό μελέτη γεγονότα, κοινούς και στις δύο ομάδες. Έστω ότι στην κάθε μια από τις r χρονικές στιγμές έχουμε d_{1j} και d_{2j} γεγονότα για την ομάδα 1 και 2, αντίστοιχα. Αντίστοιχα, έχουμε n_{1j} και n_{2j} άτομα σε κίνδυνο ακριβώς πριν τη χρονική στιγμή $t_{(j)}$, $j = 1, \dots, r$. Συνεπώς, τη χρονική στιγμή $t_{(j)}$ έχουμε $d_j = d_{1j} + d_{2j}$ γεγονότα από τα συνολικά $n_j = n_{1j} + n_{2j}$ που μπορεί να συμβούν, $j = 1, \dots, r$.

Ο πιο απλός τρόπος για να κάνουμε τον έλεγχο είναι να συγκρίνουμε τις εμπειρικές συναρτήσεις επιβίωσης για τις δύο ομάδες, ανά χρονική μονάδα. Συνδυάζουμε την πληροφορία που έχουμε ανά ομάδα και ανά χρονικό σημείο στη μορφή ενός μέτρου απόκλισης μεταξύ παρακατατηθέντων και αναμενόμενων γεγονότων στις δύο ομάδες. Ορίζουμε $e_{1j} = E(d_{1j})$. Μετράμε τη διαφορά μεταξύ παρατηρηθέντων και αναμενόμενων γεγονότων για την ομάδα 1 με τη στατιστική συνάρτηση

$$L_R = \sum_{j=1}^r (d_{1j} - e_{1j}),$$

εφόσον $E(d_{1j} - e_{1j}) = 0$ για κάθε j , συνεπάγεται $E(L_R) = 0$. Επιπλέον, επειδή τα γεγονότα είναι ανεξάρτητα μεταξύ τους η διακύμανση του L_R ισούται με το άθροισμα των διακυμάνσεων των d_{1j} . Υπό την H_0 ισχύει ότι το d_{1j} ακολουθεί

υπεργεωμετρική κατανομή με παραμέτρους $N = n_j$, $K = d_j$, $n = n_{1j}$, δηλαδή

$$P(d_{1j} = d) = \frac{\binom{K}{d} \binom{N-K}{n-d}}{\binom{N}{n}} = \frac{\binom{d_j}{d} \binom{n_j-d_j}{n_{1j}-d}}{\binom{n_j}{n_{1j}}},$$

με $d \in \mathbb{Z}$, όπου $\max(0, n + K - N) \leq d \leq \min(n, K)$. Οπότε, η μέση τιμή και η διακύμανση του d_{1j} υπό την H_0 είναι:

$$\begin{aligned} E(d_{1j}) &= n \frac{K}{N} = n_{1j} \frac{d_j}{n_j}, \\ \text{Var}(d_{1j}) &= n \frac{K}{N} \frac{(N-K)}{N} \frac{(N-n)}{N-1} = n_{1j} \frac{d_j}{n_j} \frac{(n_j - d_j)}{n_j} \frac{(n_j - n_{1j})}{n_j - 1} \\ &= \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)} = v_{1j}, \end{aligned}$$

αντίστοιχα. Τελικά η διακύμανση της στατιστικής συνάρτησης ισούται με:

$$\text{Var}(L_R) = \sum_{i=1}^r v_{1j}.$$

Ορίζοντας $Z = \frac{L_R}{\sqrt{\text{Var}(L_R)}}$, υπό την H_0 όταν το r τείνει στο άπειρο ισχύει ότι:

$$Z \sim N(0, 1) \text{ και έτσι } Z^2 = \frac{L_R^2}{\text{Var}(L_R)} \sim \chi_1^2.$$

Παράδειγμα:

Θεωρούμε ότι έχουμε τους παρακάτω χρόνους επιβίωσης ατόμων που χωρίζονται σε δύο ομάδες, με + συμβολίζουμε τις λογοκριμένες παρατηρήσεις.

Ομάδα 1 : 3.1, 6.8+, 9, 9, 11.3+, 16.2

Ομάδα 2 : 8.7, 9, 10.1+, 12.1+, 18.7, 23.1+

Θέλουμε, με χρήση του Log-rank τεστ, να ελέγξουμε την υπόθεση

$$H_0 : S_1(t) = S_2(t), \text{ για κάθε } t, \text{ έναντι της}$$

$$H_A : S_1(t) \neq S_2(t), \text{ για κάποιο } t,$$

όπου : $S_1(t)$, $S_2(t)$ είναι οι συναρτήσεις επιβίωσης των ατόμων της ομάδας 1 και 2 αντίστοιχα. Εφαρμόζουμε τη διαδικασία που αναφέραμε για την κατασκευή της στατιστικής συνάρτησης Z^2 . Έχουμε ότι $r = 5$ και $t_{(1)} = 3.1$, $t_{(2)} = 8.7$, $t_{(3)} =$

Ομάδα 1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
Ομάδα 2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Σύνολο	d_j	$n_j - d_j$	n_j

9, $t_{(4)} = 16.2$, $t_{(5)} = 18.7$. Για κάθε χρονική στιγμή μπορούμε να κατασκευάζουμε τον ακόλουθο πίνακα, ώστε να κάνουμε ευκολότερα τους υπολογισμούς μας.

- Για $t_{(1)} = 3.1$:

Ομάδα 1	1	5	6
Ομάδα 2	0	6	6
Σύνολο	1	11	12

Οπότε, $d_{11} \sim Hg(12, 1, 6)$ με μέση τιμή $E(d_{11}) = (6 \times 1)/12 = 1/2 = e_{11}$, διακύμανση $Var(d_{11}) = (6 \times 6 \times 1 \times 11)/(12^2 \times 11) = 1/4 = v_{11}$ και $(d_{11} - e_{11}) = 1 - 1/2 = 1/2$.

- Για $t_{(2)} = 8.7$:

Ομάδα 1	0	4	4
Ομάδα 2	1	5	6
Σύνολο	1	9	10

Οπότε, $d_{12} \sim Hg(10, 1, 4)$ με μέση τιμή $E(d_{12}) = 2/5 = e_{12}$, διακύμανση $Var(d_{12}) = 6/25 = v_{12}$ και $(d_{12} - e_{12}) = 0 - 2/5 = -2/5$.

- Για $t_{(3)} = 9$:

Ομάδα 1	2	2	4
Ομάδα 2	1	4	5
Σύνολο	3	6	9

Οπότε, $d_{13} \sim Hg(9, 3, 4)$ με μέση τιμή $E(d_{13}) = 4/3 = e_{13}$, διακύμανση $Var(d_{13}) = 5/9 = v_{13}$ και $(d_{13} - e_{13}) = 2 - 4/3 = 2/3$.

- Για $t_{(4)} = 16.2$:

Ομάδα 1	1	0	1
Ομάδα 2	0	2	2
Σύνολο	1	2	3

Οπότε, $d_{14} \sim Hg(3, 1, 1)$ με μέση τιμή $E(d_{14}) = 1/3 = e_{14}$, διακύμανση $Var(d_{14}) = 2/9 = v_{14}$ και $(d_{14} - e_{14}) = 1 - 1/3 = 2/3$.

- Για $t_{(5)} = 18.7$:

Ομάδα 1	0	0	0
Ομάδα 2	1	1	2
Σύνολο	1	1	2

Οπότε, $d_{15} \sim Hg(2, 1, 0)$ με μέση τιμή $E(d_{15}) = 0 = e_{15}$, διακύμανση $Var(d_{15}) = 0 = v_{15}$ και $(d_{15} - e_{15}) = 0 - 0 = 0$.

Χρησιμοποιώντας τα παραπάνω αποτελέσματα έχουμε ότι:

$$L_R = \sum_{j=1}^r (d_{1j} - e_{1j}) = \frac{1}{2} - \frac{2}{5} + \frac{2}{3} + \frac{2}{3} + 0 = \frac{43}{30} \simeq 1.43,$$

$$Var(L_R) = \sum_{i=1}^r v_{1j} = \frac{1}{4} + \frac{6}{25} + \frac{5}{9} + \frac{2}{9} + 0 = \frac{1141}{900} \simeq 1.27.$$

$$Z^2 = \frac{L_R^2}{Var(L_R)} = \frac{(43/30)^2}{1141/900} \simeq 1.62, \text{ με } Z^2 \sim \chi_1^2$$

Για επίπεδο σημαντικότητας $\alpha = 0.05$ η κρίσιμη τιμή είναι $\chi_{1,0.95}^2 = 3.841$. Επειδή ισχύει ότι $Z^2 < 3.841$ συμπεραίνουμε ότι η $H_0 : S_1(t) = S_2(t)$, για κάθε t δεν απορρίπτεται σε επίπεδο σημαντικότητας $\alpha = 0.05$.

Ο έλεγχος αυτός μπορεί να υλοποιηθεί στην R με την εντολή `survdif`. Η εντολή `Surv` δέχεται ως πρώτο όρισμα τους χρόνους ζωής και ως δεύτερο τους αντίστοιχους δείκτες λογοκρισίας, γράφοντας αριστερά του συμβόλου $\sim$ τη μεταβλητή `Group`, χωρίζουμε τα δεδομένα με βάση την ομάδα. Το αποτέλεσμα είναι ένα αντικείμενο που αποτελεί το πρώτο όρισμα της εντολής `survdif`, ενώ το δεύτερο όρισμα της είναι το σύνολο δεδομένων που χρησιμοποιούμε.

```
survdif(Surv(Time, Status) ~ Group, data = Lr)
```

Τα αποτελέσματα που λαμβάνουμε είναι τα εξής:

```
survdif(formula = Surv(Time, Status) ~ Group, data = Lr)
```

	N	Observed	Expected	(O-E) ² /E	(O-E) ² /V
Group=1	6	4	2.57	0.800	1.62
Group=2	6	3	4.43	0.463	1.62

Chisq= 1.6 on 1 degrees of freedom, p= 0.2

Παρατηρούμε ότι οι υπολογισμοί που κάναμε συμφωνούν με τα αποτελέσματα της R. Εφόσον η p -τιμή = 0.2 > 0.05 δεν απορρίπτεται η μηδενική υπόθεση. Συνεπώς, δεν υπάρχουν στατιστικά σημαντικές ενδείξεις ότι οι συναρτήσεις επιβίωσης διαφέρουν μεταξύ των δύο ομάδων.

2.7 Υπόθεση αναλογικότητας κινδύνων

Βασική υπόθεση για να είναι εφαρμόσιμο το μοντέλο αναλογικών κινδύνων του Cox, είναι η βασική συνάρτηση κινδύνου $h_0(t)$ να είναι ίδια για όλους τους ασθενείς της μελέτης (υπόθεση αναλογικότητας κινδύνων). Αν δεν ισχύει αυτή η υπόθεση το μοντέλο δεν μπορεί να ισχύει, ούτε έχει νόημα να εκτιμήσουμε τις παραμέτρους του μοντέλου με τη μέθοδο μερικής πιθανοφάνειας. Αρχικά αναφέρουμε πώς ελέγχεται η υπόθεση της αναλογικότητας κινδύνων και μετά δίνουμε τον πιο απλό τρόπο αντιμετώπισης της περίπτωσης που η υπόθεση της αναλογικότητας κινδύνων δεν ισχύει. Για τον σκοπό αυτό, η θεωρητική προσέγγιση βασίζεται κυρίως στα συγγράμματα των Kleinbaum (2012) και Tableman and Kim (2003).

Τρόποι ελέγχου της υπόθεσης αναλογικότητας κινδύνων:

Ενδιαφερόμαστε για τον έλεγχο της αναλογικότητας κινδύνων για μια δοθείσα συμμεταβλητή, έστω X_i , $i \in \{1, \dots, n\}$, αφού έχουμε προσαρμόσει όλες τις υπόλοιπες συμμεταβλητές στο μοντέλο. Θεωρούμε ότι η συμμεταβλητή για την οποία κάνουμε τον έλεγχο, δεν αλληλεπιδρά με τις υπόλοιπες προσαρμοσμένες στο μοντέλο συμμεταβλητές.

Έλεγχος με βάση τη γραφική παράσταση των συναρτήσεων κινδύνου:

Ο πρώτος τρόπος ελέγχου απαιτεί η συμμεταβλητή X_i να έχει μόνο K δυνατές τιμές. Για τον λόγο αυτόν αν είναι συνεχής την στρωματοποιούμε σε K ασυσχέτιστα στρώματα. Έπειτα, προσαρμόζουμε ένα στρωματοποιημένο Cox μοντέλο, στρωματοποιώντας ως προς τη μεταβλητή X_i . Δηλαδή προσαρμόζουμε όλες τις συμμεταβλητές στο μοντέλο του Cox, εκτός της μεταβλητής στρωματοποίησης, ενώ με βάση τις κατηγορίες, τα στρώματα, αυτής της μεταβλητής χωρίζουμε τα δεδομένα. Κάνουμε τη γραφική παράσταση των συναρτήσεων κινδύνου που αντιστοιχούν σε κάθε στρώμα.

- Αν ισχύει η υπόθεση της αναλογικότητας κινδύνων, οι καμπύλες των συναρτήσεων κινδύνου (ή των αθροιστικών συναρτήσεων κινδύνου) θα πρέπει να μην τέμνουν η μια την άλλη.
- Αν οι καμπύλες διασταυρώνονται, συμπεραίνουμε ότι δεν ισχύει η υπόθεση της αναλογικότητας κινδύνων ως προς την μεταβλητή X_i , οπότε το μοντέλο του Cox δεν μπορεί να εφαρμοστεί αυτούσιο.

Θα ελέγξουμε την υπόθεση της αναλογικότητας για τη συμμεταβλητή **GROUP** στο μοντέλο **cns.cox**, το οποίο έχουμε προσαρμόσει στα δεδομένα **CNS** στην Ενότητα 2.4. Ξεκινάμε αποθηκεύοντας τις τιμές της συμμεταβλητής **GROUP** στο αντικείμενο **t1**. Στα αντικείμενα **t1.1** και **t1.2** αποθηκεύουμε τους χρόνους ζωής ή τους χρόνους λογοκρισίας των ατόμων που ανήκουν στο **GROUP 0** και **1**, αντίστοιχα. Στα αντικείμενα **c1** και **c2** αποθηκεύουμε τους δείκτες λογοκρισίας των ασθενών που ανήκουν στο **GROUP 0** και **1** αντίστοιχα. Στη συνέχεια, δίνουμε στη συνάρτηση **Surv** ως πρώτο όρισμα τους χρόνους **t1.1** και τους δείκτες **c1**. Η συνάρτηση **coxph** δέχεται ως πρώτο όρισμα τη συνάρτηση **Surv** που προαναφέραμε, δεξιά του συμβόλου $\sim$ γράφουμε τη μονάδα, καθώς πέρα από τη μεταβλητή **GROUP** δεν είχαμε άλλες συμμεταβλητές στο μοντέλο **cns.cox**. Με αυτόν τον τρόπο αποθηκεύουμε στο αντικείμενο **h0t1.1** το μοντέλο του Cox μόνο για τα δεδομένα που αντιστοιχούν στο **GROUP 0**. Με παρόμοιο τρόπο στο αντικείμενο **h0t1.2** αποθηκεύουμε το μοντέλο του Cox μόνο για τα δεδομένα που αντιστοιχούν στο **GROUP 1**. Η συνάρτηση **basehaz** χρησιμοποιείται για να εκτιμηθεί η βασική συνάρτηση κινδύνου. Δίνοντας ως πρώτο όρισμα στη συνάρτηση **basehaz** το μοντέλο **h0t1.1**, υπολογίζεται η εκτιμώμενη βασική συνάρτηση κινδύνου για τα δεδομένα που αντιστοιχούν στην τιμή 0 για τη μεταβλητή **GROUP**. Η εκτίμηση αυτή αποθηκεύεται στο αντικείμενο **h0t1**. Όμοια, στο αντικείμενο **h0t2** αποθηκεύεται η εκτιμώμενη βασική συνάρτηση κινδύνου για τα δεδομένα που αντιστοιχούν στην τιμή 1 για τη μεταβλητή **GROUP**. Η εντολή

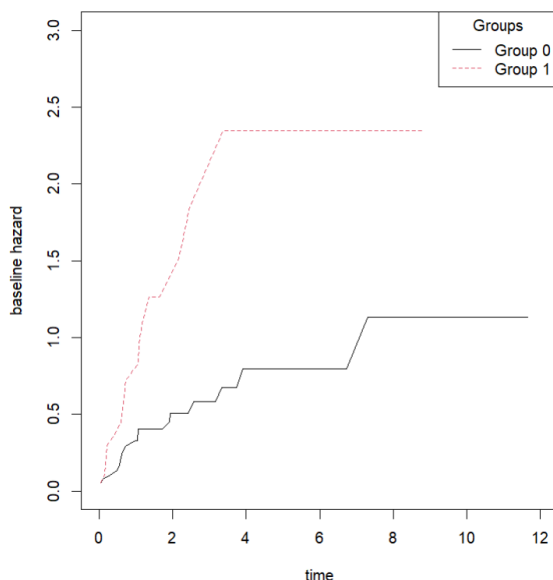
`plot`, με πρώτο όρισμα `h0t1$time` και δεύτερο `h0t1$hazard`, κάνει τη γραφική παράσταση της εκτιμώμενης βασικής συνάρτησης κινδύνου έναντι του χρόνου, που έχουμε αποθηκεύσει στο αντικείμενο `h0t1`. Η εντολή `lines`, με πρώτο όρισμα `h0t2$time` και δεύτερο `h0t2$hazard`, κάνει στην υπάρχουσα γραφική παράσταση τη γραφική της εκτιμώμενης βασικής συνάρτησης κινδύνου έναντι του χρόνου, που αποθηκεύσαμε στο αντικείμενο `h0t2`. Τέλος, η εντολή `legend` προσθέτει την κατάλληλη λεζάντα.

```
t1<-GROUP
t1.1<-B3TODEATH[which(t1==0)]
c1<-STATUS[which(t1==0)]
t1.2<-B3TODEATH[which(t1==1)]
c2<-STATUS[which(t1==1)]
h0t1.1<-coxph(Surv(t1.1, c1)~1)
h0t1.2<-coxph(Surv(t1.2, c2)~1)
h0t1<-basehaz(h0t1.1, centered=TRUE)
h0t2<-basehaz(h0t1.2, centered=TRUE)
plot(h0t1$time , h0t1$hazard , type="l", xlab="time",
     ylab="baseline hazard", xlim=c(0, 12) , ylim=c(0, 3))
lines(h0t2$time , h0t2$hazard , type="l", xlab="time",
     ylab="baseline hazard", lty=2, col=2)
legend("topright", legend=c("Group 0", "Group 1"),
     lty=c(1, 2), col=c(1, 2), title="Groups")
```

Σύμφωνα με τη γραφική παράσταση του Σχήματος 2.2 οι συναρτήσεις κινδύνου για τις κατηγορίες της μεταβλητής `GROUP` δεν τέμνονται, οπότε ισχύει η υπόθεση της αναλογικότητας κινδύνων για τη μεταβλητή `GROUP`.

Εναλλακτικά θα μπορούσαμε να κάνουμε γραφικές παραστάσεις που χρησιμοποιούν την αθροιστική συνάρτηση κινδύνου, τα συμπεράσματα που θα παίρναμε θα ήταν ανάλογα. Επιλέγουμε να χρησιμοποιήσουμε τη συνάρτηση κινδύνου έναντι της αθροιστικής συνάρτησης κινδύνου, για τον έλεγχο της υπόθεσης της αναλογικότητας κινδύνων, καθώς είναι πιο πρακτικό. Παραδείγματα γραφικών παραστάσεων για τον έλεγχο της υπόθεσης της αναλογικότητας κινδύνων με χρήση της αθροιστικής συνάρτησης κινδύνου είναι τα εξής:

- Μία γραφική παράσταση της λογαριθμισμένης αθροιστικής συνάρτησης κινδύνου έναντι του χρόνου, είναι χρήσιμη όταν X_i είναι μία μεταβλητή ομάδων. Για παράδειγμα, αν υπάρχουν δύο ομάδες θεραπείας, απεικονίζουμε και τις δύο καμπύλες στο ίδιο γράφημα και τις συγκρίνουμε. Αν οι καμπύλες είναι



Σχήμα 2.2: Γραφική παράσταση των συναρτήσεων κινδύνου για τις κατηγορίες της μεταβλητής **GROUP**, με σκοπό τον έλεγχο της υπόθεσης της αναλογικότητας κινδύνου για την μεταβλητή **GROUP** στο μοντέλο `cns.cox`.

παράλληλες με τον χρόνο, ισχύει η υπόθεση της αναλογικότητας κινδύνων. Αν διασταυρώνονται, τότε αυτό υποδηλώνει την παραβίαση της υπόθεσης.

- Η γραφική παράσταση της διαφοράς των λογαριθμισμένων αθροιστικών συναρτήσεων κινδύνου έναντι του χρόνου είναι επίσης χρήσιμη. Αυτή η γραφική παράσταση απεικονίζει τη διαφορά μεταξύ των δύο καμπυλών της προαναφερθείσας γραφικής παράστασης. Αν ισχύει η υπόθεση της αναλογικότητας κινδύνων, αυτή η γραφική είναι μια σταθερά, κατά προσέγγιση, με τον χρόνο. Αλλιώς, η παραβίαση της υπόθεσης είναι εμφανής.

Έλεγχος με βάση τη γραφική παράσταση των συναρτήσεων επιβίωσης:

Στον συγκεκριμένο τρόπο ελέγχου της υπόθεσης της αναλογικότητας κινδύνων θέλουμε και πάλι η μεταβλητή για την οποία κάνουμε τον έλεγχο, έστω X_i , να έχει μόνο K δυνατές τιμές. Επομένως, αν είναι συνεχής την στρωματοποιούμε σε K ασυσχέιστα στρώματα. Στη συνέχεια, προσαρμόζουμε το στρωματοποιη-

μένο μοντέλο του Cox, στρωματοποιώντας ως προς τη μεταβλητή X_i . Δηλαδή, προσαρμόζουμε όλες τις συμμεταβλητές στο μοντέλο του Cox εκτός της μεταβλητής στρωματοποίησης, με βάση τις κατηγορίες αυτής της μεταβλητής χωρίζουμε τα δεδομένα. Κάνουμε τη γραφική παράσταση των συναρτήσεων επιβίωσης που αντιστοιχούν σε κάθε στρώμα.

- Αν ισχύει η υπόθεση της αναλογικότητας κινδύνων, οι καμπύλες των συναρτήσεων επιβίωσης θα πρέπει να μην τέμνονται. Σε αυτήν την περίπτωση γίνεται χρήση του μοντέλου αναλογικών κινδύνων του Cox.
- Αν οι καμπύλες διασταυρώνονται, συμπεραίνουμε ότι δεν ισχύει η υπόθεση της αναλογικότητας κινδύνων ως προς τη μεταβλητή X_i .

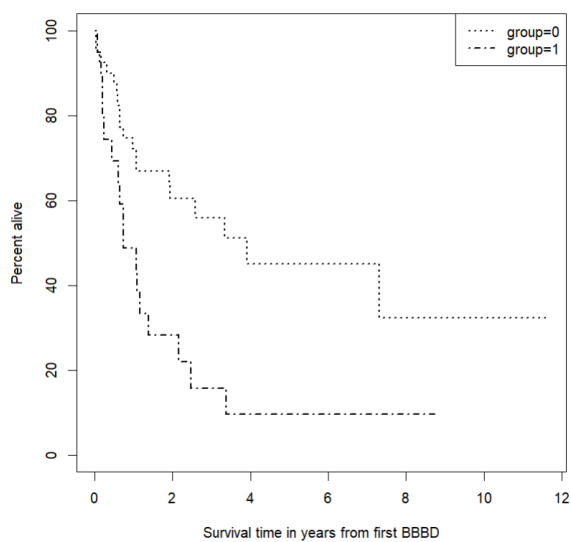
Για το μοντέλο `cns.cox`, που έχουμε ήδη προσαρμόσει, η γραφική παράσταση των συναρτήσεων επιβίωσης με μεταβλητή στρωματοποίησης την `GROUP` προκύπτει ακολουθώντας τα επόμενα βήματα. Στο αντικείμενο `cns` αποθηκεύουμε το μοντέλο του Cox που έχει μεταβλητή στρωματοποίησης την `GROUP` και δεν περιλαμβάνει άλλες μεταβλητές. Η συνάρτηση `survfit` υπολογίζει τις εκτιμώμενες συναρτήσεις επιβίωσης. Δίνοντας ως όρισμα στη συνάρτηση `survfit` το μοντέλο `cox` υπολογίζονται οι εκτιμώμενες συναρτήσεις επιβίωσης για τα δεδομένα με βάση τις κατηγορίες της μεταβλητής `GROUP`. Οι εκτιμήσεις αυτές αποθηκεύονται στο αντικείμενο `survfit`. Για να απεικονίσουμε τις γραφικές παραστάσεις των εκτιμώμενων συναρτήσεων επιβίωσης που υπολογίστηκαν, χρησιμοποιούμε τη συνάρτηση `plot` με πρώτο όρισμα το `survfit`. Η εντολή `legend` προσθέτει στο γράφημα την αντίστοιχη λεζάντα.

```
cns<- coxph(Surv(B3TODEATH,STATUS)~strata(GROUP))
survfit <- survfit(cns)
plot(survfit,col=1,lty=3:4,lwd=2,cex=3,
     xlab="Survival time in years from first BBBD",
     ylab="Percent alive",yscale=100)
legend("topright",c("group=0","group=1"),lty=3:4,lwd=2)
```

Με βάση το Σχήμα 2.3 συμπεραίνουμε ότι ισχύει η υπόθεση της αναλογικότητας κινδύνου ως προς τη μεταβλητή `GROUP`, καθώς οι δύο καμπύλες δεν διασταυρώνονται.

Έλεγχος με βάση τη γραφική παράσταση των log – log συναρτήσεων επιβίωσης:

Η log – log συνάρτηση επιβίωσης είναι ένας μετασχηματισμός της εκτιμώμενης συνάρτησης επιβίωσης, που προκύπτει εφαρμόζοντας τον φυσικό λογάριθμο στην



Σχήμα 2.3: Γραφική παράσταση των συναρτήσεων επιβίωσης για τις κατηγορίες της μεταβλητής **GROUP**, με σκοπό τον έλεγχο της υπόθεσης της αναλογικότητας κινδύνου για την μεταβλητή **GROUP** στο μοντέλο `cns.cox`.

εκτιμώμενη συνάρτηση επιβίωσης δύο φορές. Επομένως, η μαθηματική έκφραση της θα είναι:

$$\log - \log\{\hat{S}(t)\} = -\ln\{-\ln \hat{S}(t)\}.$$

Για να γίνει η αξιολόγηση της υπόθεσης της αναλογικότητας κινδύνων στο μοντέλο του Cox περιγράφουμε τη μορφή που έχει η $\log - \log$ συνάρτηση επιβίωσης στο μοντέλο του Cox. Ξεκινάμε με την μορφή της συνάρτησης κινδύνου που

$$h(t|\mathbf{X}) = h_0(t) \exp(\mathbf{X}'\boldsymbol{\beta})$$

$$= h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m)$$

είναι:

$$= h_0(t) \exp\left(\sum_{i=1}^m \beta_i X_i\right).$$

Με χρήση της σχέσης που συνδέει τη συνάρτηση κινδύνου με τη συνάρτηση επιβίωσης, δηλαδή $S(t) = \exp\left\{-\int_0^t h(u)du\right\}$, έχουμε ότι η συνάρτηση επιβίωσης για το μοντέλο αναλογικών κινδύνων του Cox είναι:

$$\begin{aligned} S(t|\mathbf{X}) &= \exp\left\{-\int_0^t h(u)du\right\} \\ &= \exp\left\{-\int_0^t h_0(u) \exp\left(\sum_{i=1}^m \beta_i X_i\right) du\right\} \\ &= \exp\left\{-\int_0^t h_0(u)du \exp\left(\sum_{i=1}^m \beta_i X_i\right)\right\} \\ &= \{S_0(t)\}^{\exp(\sum_{i=1}^m \beta_i X_i)}, \end{aligned}$$

με $S_0(t) = \exp\left\{-\int_0^t h_0(u)du\right\}$. Προκειμένου να πάρουμε τη $\log - \log$ μορφή της παραπάνω συνάρτησης επιβίωσης, λογαριθμίζουμε τη συνάρτηση επιβίωσης δύο φορές, οπότε

$$\ln\{S(t|\mathbf{X})\} = \exp\left(\sum_{i=1}^m \beta_i X_i\right) \ln\{S_0(t)\}.$$

Η ποσότητα $\ln\{S(t|\mathbf{X})\}$ είναι αρνητική, καθώς $0 \leq S(t|\mathbf{X}) \leq 1$. Έτσι, για να πάρουμε τον λογάριθμο δεύτερη φορά βάζουμε αρνητικό πρόσημο μπροστά από την παραπάνω έκφραση, άρα η ποσότητα $-\ln\{S(t|\mathbf{X})\}$ είναι θετική.

$$\begin{aligned}
\ln[-\ln\{S(t|\mathbf{X})\}] &= \ln \left[-\exp \left(\sum_{i=1}^m \beta_i X_i \right) \times \ln\{S_0(t)\} \right] \\
&= \ln \left\{ \exp \left(\sum_{i=1}^m \beta_i X_i \right) \right\} + \ln[-\ln\{S_0(t)\}] \\
&= \sum_{i=1}^m \beta_i X_i + \ln[-\ln\{S_0(t)\}].
\end{aligned}$$

Επομένως,

$$-\ln[-\ln\{S(t|\mathbf{X})\}] = -\sum_{i=1}^m \beta_i X_i - \ln[-\ln\{S_0(t)\}].$$

Η τελική ποσότητα, αφού λογαριθμίσουμε και δεύτερη φορά, είναι είτε θετική, είτε αρνητική. Για λόγους συνέπειας βάζουμε το αρνητικό πρόσημο και τη δεύτερη φορά, αν και κάποια στατιστικά πακέτα δεν χρησιμοποιούν το δεύτερο αρνητικό πρόσημο. Θεωρούμε ότι έχουμε τα διανύσματα των συμμεταβλητών $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{im})'$ και $\mathbf{X}_j = (X_{j1}, X_{j2}, \dots, X_{jm})'$ που αντιστοιχούν στα i, j άτομα της μελέτης $i, j \in \{1, \dots, n\}$. Η $\log - \log$ συνάρτηση επιβίωσης για τα i και j άτομα της μελέτης θα είναι αντίστοιχα:

$$\begin{aligned}
-\ln[-\ln\{S(t|\mathbf{X}_i)\}] &= -\sum_{k=1}^m \beta_k X_{ik} - \ln[-\ln\{S_0(t)\}], \\
-\ln[-\ln\{S(t|\mathbf{X}_j)\}] &= -\sum_{k=1}^m \beta_k X_{jk} - \ln[-\ln\{S_0(t)\}].
\end{aligned}$$

Αφαιρώντας κατά μέλη την πρώτη εξίσωση από την δεύτερη έχουμε:

$$\ln[-\ln\{S(t|\mathbf{X}_i)\}] - (-\ln[-\ln\{S(t|\mathbf{X}_j)\}]) = \sum_{k=1}^m \beta_k (X_{ik} - X_{jk}).$$

Η ποσότητα αυτή είναι ανεξάρτητη του χρόνου. Το γεγονός αυτό υποδεικνύει ότι αν χρησιμοποιήσουμε το μοντέλο αναλογικών κινδύνων του Cox και απεικονίσουμε γραφικά τις εκτιμώμενες $\log - \log$ συναρτήσεις επιβίωσης για δύο άτομα της μελέτης, στην ίδια γραφική παράσταση, οι δύο γραφικές παραστάσεις θα είναι προσεγγιστικά παράλληλες.

Συμπέρασμα:

Η παραλληλία των γραφικών παραστάσεων $\log - \log$ των συναρτήσεων επιβίωσης για το μοντέλο αναλογικών κινδύνων του Cox, μας παρέχει γραφικό τρόπο ελέγχου της υπόθεσης της αναλογικότητας κινδύνων.

- Αν το μοντέλο αναλογικών κινδύνων είναι κατάλληλο για ένα δοθέν σύνολο μεταβλητών πρόβλεψης αναμένεται οι γραφικές παραστάσεις των $\log - \log$ εμπειρικών συναρτήσεων επιβίωσης για διαφορετικά άτομα της μελέτης να είναι προσεγγιστικά παράλληλες.
- Εναλλακτικά, μπορεί να γίνει η γραφική παράσταση των $\log - \log$ εμπειρικών συναρτήσεων επιβίωσης, που προσαρμόστηκαν για τις συμμεταβλητές που υποθέτουμε ότι ικανοποιούν την υπόθεση της αναλογικότητας κινδύνων. Ωστόσο, δεν περιλαμβάνεται στο μοντέλο η μεταβλητή για την οποία γίνεται η αξιολόγηση της υπόθεσης της αναλογικότητας.

Με τον όρο εμπειρικές γραφικές παραστάσεις (empirical plots) αναφερόμαστε στη γραφική παράσταση των $\log - \log$ εμπειρικών συναρτήσεων επιβίωσης, $-\ln[-\ln\{\hat{S}(t|\mathbf{X})\}]$. Οι $\hat{S}(t)$ προκύπτουν με δύο τρόπους. Ο πρώτος τρόπος αντιστοιχεί σε συναρτήσεις επιβίωσης που βασίζονται στον εκτιμητή Kaplan-Meier (KM), ο οποίος δεν υποθέτει το Cox μοντέλο. Ο δεύτερος τρόπος αντιστοιχεί στις συναρτήσεις που προκύπτουν αν χρησιμοποιήσουμε το μοντέλο αναλογικών κινδύνων του Cox, έχοντας προσαρμόσει τις συμμεταβλητές που ικανοποιούν την υπόθεση της αναλογικότητας κινδύνων στο μοντέλο. Ωστόσο, σε αυτήν την περίπτωση, δεν συμπεριλαμβάνουμε ως συμμεταβλητή τη μεταβλητή ως προς την οποία ελέγχουμε την υπόθεση της αναλογικότητας κινδύνων, άρα θα πρέπει να τη θέσουμε ως μεταβλητή ως προς την οποία γίνεται στρωματοποίηση.

Συνοπτικά, η χρήση των συναρτήσεων επιβίωσης για τον έλεγχο της αναλογικότητας κινδύνων στο μοντέλο, μέσω των $\log - \log$ συναρτήσεων επιβίωσης, γίνεται ως εξής:

- Με τη χρήση των συναρτήσεων επιβίωσης που βασίζονται στον εκτιμητή Kaplan-Meier (KM). Αν θέλουμε να ελέγξουμε την υπόθεση αναλογικότητας κινδύνων για περισσότερες από μία συμμεταβλητές ταυτόχρονα, κατηγοριοποιούμε όλες τις συμμεταβλητές ξεχωριστά. Στη συνέχεια σχηματίζουμε όλους τους πιθανούς συνδυασμούς κατηγοριών και τέλος συγκρίνουμε τις γραφικές παραστάσεις όλων των $\log - \log$ εμπειρικών συναρτήσεων επιβίωσης σε ένα γράφημα. Κάποια μειονεκτήματα αυτής της μεθόδου είναι ότι είναι χρονοβόρα για την περίπτωση που έχουμε πολλούς συνδυασμούς κατηγοριών. Επιπλέον, το πλήθος των παρατηρήσεων που χρησιμοποιούνται για τον σχηματισμό κάθε καμπύλης τείνει να μειώνεται, όταν αυξάνεται το πλήθος των συνδυασμών κατηγοριών. Ακόμη, όταν υπάρχουν αρκετοί συνδυασμοί κατηγοριών, είναι δύσκολο να προσδιοριστεί ποια μεταβλητή ευθύνεται για τη μη παραλληλία που μπορεί να εμφανιστεί.

- Με τη χρήση του μοντέλου του Cox, όπου έχουμε προσαρμόσει όλες τις μεταβλητές ως συμμεταβλητές, εκτός από τη μεταβλητή ως προς την οποία θέλουμε να ελέγξουμε την υπόθεση της αναλογικότητας κινδύνων. Τη μεταβλητή αυτήν την ορίζουμε ως μεταβλητή στρωματοποίησης. Συγκρίνουμε τις $\log - \log$ συναρτήσεις επιβίωσης των διαφόρων στρωμάτων (οι οποίες είναι υπολογισμένες αφού έχουμε προσαρμόσει όλες τις υπόλοιπες μεταβλητές ως συμμεταβλητές). Προφανώς για να ελέγξουμε την υπόθεση της αναλογικότητας κινδύνων για όλες τις μεταβλητές, κάθε φορά χρειαζόμαστε μια μεταβλητή ως μεταβλητή στρωματοποίησης.

Παρατήρηση:

Ένα ερώτημα που προκύπτει όταν εξετάζουμε την υπόθεση της αναλογικότητας κινδύνων γραφικά είναι το πώς καθορίζεται αν οι γραφικές παραστάσεις δύο συναρτήσεων είναι επαρκώς παράλληλες. Η απόφαση αυτή μπορεί να είναι υποκειμενική, ιδιαίτερα σε περιπτώσεις που το μέγεθος του δείγματος είναι μικρό. Η συνιστώμενη προσέγγιση είναι να θεωρούμε ότι ισχύει η υπόθεση της αναλογικότητας κινδύνων, εκτός από τις περιπτώσεις που υπάρχουν ισχυρές ενδείξεις μη παραλληλίας των $\log - \log$ συναρτήσεων επιβίωσης.

Θα κάνουμε τους παραπάνω ελέγχους στο μοντέλο `cns.cox`. Σε αυτό το μοντέλο η μοναδική μεταβλητή πρόβλεψης είναι η `GROUP`. Προκειμένου να δούμε αν το μοντέλο είναι κατάλληλο επιλέγουμε τυχαία δύο άτομα της μελέτης, έστω αυτά με αύξοντα αριθμό 12 και 35. Οι γραφικές παραστάσεις των $\log - \log$ εμπειρικών συναρτήσεων επιβίωσης δίνονται από τον παρακάτω κώδικα. Στο αντικείμενο `new_data` αποθηκεύουμε όλα τα δεδομένα που αφορούν τα άτομα της μελέτης που επιλέξαμε. Με τη χρήση της συνάρτησης `survfit` υπολογίζεται η εκτιμώμενη συνάρτηση επιβίωσης για κάθε ένα από τα άτομα που επιλέχθηκαν, με βάση το μοντέλο `cns.cox` και με χρήση των δεδομένων `new_data`. Οι εκτιμήσεις αυτές αποθηκεύονται στο αντικείμενο `surv_fit`. Η εντολή `plot` δέχεται ως πρώτο όρισμα το `surv_fit`, δίνοντας στο όρισμα `fun` την τιμή "`cloglog`" εφαρμόζεται δύο φορές ο λογάριθμος στις εκτιμώμενες συναρτήσεις επιβίωσης, οπότε παίρνουμε και το ζητούμενο γράφημα. Η εντολή `legend` υποδηλώνει ποια καμπύλη αντιστοιχεί σε κάθε άτομο.

```
new_data <- data.frame(PT.NUMBER=c(12,35),GROUP=c(1,0),
                      SEX=c(0,1),AGE=c(45,60),STATUS=c(1,1),
                      DXTOB3=c(0.91666667,0.06250000),
                      DXTODEATH=c(2.0000000,1.0416667),
                      B3TODEATH=c(1.08333333,0.97916667),
                      KPS.PRE.=c(50,40),LESSING=c(1,1),
```

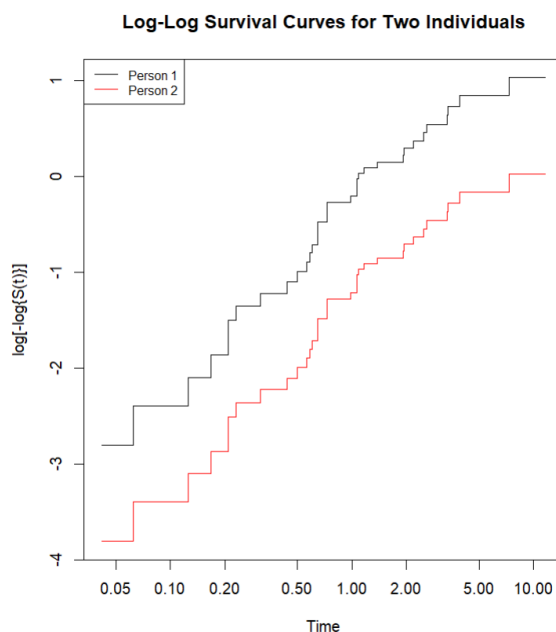
```

LESDEEP=c(1,1),LESSUP=c(2,2),
PROC=c(2,2),RAD4000=c(1,0),
CHEMOPRIOR=c(0,0),RESPONSE=c(2,1),
AGE60=c(1,1))

surv_fit <- survfit(cns.cox, newdata = new_data)
plot(surv_fit, fun = "cloglog", xlab = "Time",
     ylab = "log[-log{S(t)}]",
     main = "Log-Log Survival Curves for Two Individuals",
     col=c("black","red"))

legend("topleft", legend = c("Person 1", "Person 2"),
     col = c("black", "red"), lty = 1, cex = 0.8)

```



Σχήμα 2.4: Γραφική παράσταση των $\log - \log$ συναρτήσεων επιβίωσης, έναντι του χρόνου, για δύο τυχαία επιλεγμένα άτομα της μελέτης με σκοπό τον έλεγχο της καταλληλότητας του μοντέλου `cns.cox`.

Σύμφωνα με τη γραφική παράσταση που δίνεται στο Σχήμα 2.4 οι $\log - \log$ συναρτήσεις επιβίωσης για τα δύο τυχαία επιλεγμένα άτομα της μελέτης είναι προσεγγιστικά παράλληλες, οπότε το μοντέλο είναι κατάλληλο.

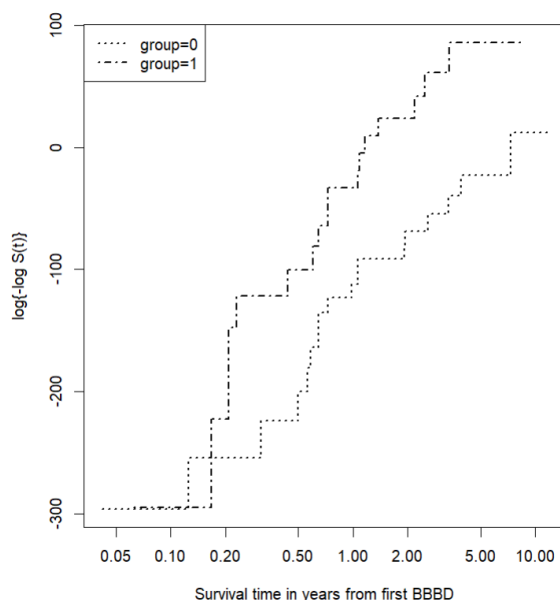
Για το ίδιο μοντέλο, δηλαδή το `cns.cox`, η γραφική παράσταση των $\log - \log$ εμπειρικών συναρτήσεων επιβίωσης, για την αξιολόγηση της υπόθεσης της αναλογικότητας της μεταβλητής `GROUP` δίνεται από τον παρακάτω κώδικα. Αρχικά, χρησιμοποιούμε το μοντέλο `cns` γιατί θέλουμε να απεικονίσουμε ξεχωριστά τις $\log - \log$ συναρτήσεις επιβίωσης για κάθε επίπεδο της συμμεταβλητής `GROUP`. Με την εντολή `survfit` υπολογίζονται οι εκτιμώμενες συναρτήσεις επιβίωσης για το μοντέλο `cns`, δηλαδή όταν μεταβλητή στρωματοποίησης είναι η `GROUP`. Η συνάρτηση `plot` κάνει τη γραφική παράσταση των εκτιμώμενων συναρτήσεων επιβίωσης σε $\log - \log$ κλίμακα. Η εντολή `legend` προσθέτει μια λεζάντα που δηλώνει σε ποια τιμή της μεταβλητής `GROUP` αντιστοιχεί κάθε καμπύλη.

```
cns<- coxph(Surv(B3TODEATH,STATUS)~strata(GROUP))
survfit <- survfit(cns)
plot(survfit,fun="cloglog", col=1,lty=3:4,lwd=2,cex=3,
     xlab="Survival time in years from first BBBD",
     ylab="log{-log S(t)}",yscale=100)
legend("topleft",c("group=0","group=1"),lty=3:4,lwd=2)
```

Με βάση τη γραφική παράσταση του Σχήματος 2.5 μπορούμε να συμπεράνουμε ότι ισχύει η υπόθεση της αναλογικότητας κινδύνων, για τη μεταβλητή `GROUP`, καθώς οι δύο γραφικές είναι προσεγγιστικά παράλληλες. Ωστόσο, υπάρχει μια επιφύλαξη καθώς για μικρούς χρόνους ζωής διασταυρώνονται.

Έλεγχος μέσω σύγκρισης των παρατηρούμενων έναντι των αναμενόμενων γραφικών παραστάσεων:

Η γραφική παράσταση των παρατηρούμενων έναντι των αναμενόμενων συναρτήσεων επιβίωσης, που χρησιμοποιείται για την αξιολόγηση της υπόθεσης της αναλογικότητας κινδύνων, είναι ένα γραφικό ανάλογο της προσέγγισης ελέγχου καλής προσαρμογής (goodness-of-fit). Είναι μια αξιολόγηση εναλλακτική του ελέγχου που βασίζεται στη γραφική παράσταση των $\log - \log$ συναρτήσεων επιβίωσης. Σε αυτόν τον τρόπο ελέγχου πρέπει να φτιάξουμε τις παρατηρούμενες και τις αναμενόμενες συναρτήσεις επιβίωσης. Ο τρόπος με τον οποίο το κάνουμε αυτό είναι ο εξής: οι παρατηρούμενες καμπύλες είναι, όπως και στην προηγούμενη περίπτωση, τα διαγράμματα των συναρτήσεων επιβίωσης. Οι παρατηρούμενες συναρτήσεις επιβίωσης μπορούν να προσδιοριστούν είτε με χρήση KM εκτίμησης, είτε με χρήση στρωματοποιημένου μοντέλου Cox, αφού έχουμε προσαρμόσει τις υπόλοιπες συμμεταβλητές. Περιγράφουμε την περίπτωση που γίνεται χρήση KM εκτίμησης για τον προσδιορισμό των παρατηρούμενων καμπυλών. Αρχικά, στρωματοποιούμε τα δεδομένα μας με βάση τις κατηγορίες της συμμεταβλητής, ως προς



Σχήμα 2.5: Γραφική παράσταση των $\log - \log$ εμπειρικών συναρτήσεων επιβίωσης, έναντι του χρόνου, για την αξιολόγηση της υπόθεσης της αναλογικότητας της μεταβλητής GROUP.

την οποία γίνεται ο έλεγχος αναλογικότητας κινδύνων. Έπειτα, προσδιορίζουμε τις παρατηρούμενες συναρτήσεις επιβίωσης με εκτίμηση KM για κάθε στρώμα ξεχωριστά. Για τον προσδιορισμό των αναμενόμενων συναρτήσεων επιβίωσης προσαρμόζουμε το μοντέλο του Cox που περιέχει ως μοναδική συμμεταβλητή, τη μεταβλητή ως προς την οποία ελέγχουμε την υπόθεση της αναλογικότητας κινδύνων. Παρατηρούμε τις αναμενόμενες καμπύλες, αντικαθιστώντας κάθε φορά στον τύπο για την εκτιμώμενη συνάρτηση επιβίωσης, την τιμή της μεταβλητής πρόβλεψης για κάθε κατηγορία. Συνεπώς, παρατηρούμε διαφορετική εκτιμώμενη συνάρτηση επιβίωσης για κάθε κατηγορία. Με σκοπό να συγκρίνουμε τις παρατηρούμενες με τις αναμενόμενες συναρτήσεις επιβίωσης, τις απεικονίζουμε στο ίδιο γράφημα. Αν για κάθε κατηγορία της μεταβλητής ως προς την οποία κάνουμε τον έλεγχο αναλογικότητας κινδύνων οι παρατηρούμενες και οι αναμενόμενες συναρτήσεις επιβίωσης είναι κοντά, μπορούμε να συμπεράνουμε ότι ισχύει η υπόθεση της αναλογικότητας κινδύνων. Ωστόσο, αν για μία ή περισσότερες κατηγορίες οι γραφικές παραστάσεις δεν είναι κοντινές, μπορούμε να θεωρήσουμε ότι παραβιάζεται η υπόθεση της αναλογικότητας κινδύνων.

Όταν γίνεται χρήση των παρατηρούμενων έναντι των αναμενόμενων γραφικών παραστάσεων για τον έλεγχο της υπόθεσης της αναλογικότητας κινδύνων για συνεχείς μεταβλητές, οι παρατηρούμενες καμπύλες παρατηρούνται όπως στην περίπτωση που έχουμε κατηγορική μεταβλητή. Δηλαδή με τη δημιουργία στρωμάτων για τη συνεχή μεταβλητή ως προς την οποία γίνεται ο έλεγχος, και έπειτα βρίσκοντας για κάθε στρώμα την KM καμπύλη. Ωστόσο, στην περίπτωση που έχουμε συνεχείς συμεταβλητές, υπάρχουν δύο διαθέσιμες επιλογές για τον υπολογισμό των αναμενόμενων καμπυλών. Η πρώτη επιλογή είναι να γίνει χρήση του μοντέλου αναλογικών κινδύνων του Cox, το οποίο θα περιέχει $k - 1$ μεταβλητές δείκτριες (dummy variables) για τις, έστω, k κατηγορίες της συνεχούς μεταβλητής. Η αναμενόμενη γραφική για μια δοθείσα κατηγορία παρατηρείται αντικαθιστώντας τις τιμές των δεικτριών που ορίζουν τη δοθείσα κατηγορία στον τύπο της εκτιμώμενης συνάρτησης επιβίωσης. Η δεύτερη επιλογή είναι να γίνει χρήση του μοντέλου αναλογικών κινδύνων του Cox, που περιέχει ως μοναδική μεταβλητή πρόβλεψης τη συνεχή μεταβλητή, για την οποία γίνεται ο έλεγχος της αναλογικότητας κινδύνων. Οι αναμενόμενες γραφικές παραστάσεις παρατηρούνται ως οι συναρτήσεις επιβίωσης που προκύπτουν αν αντικαταστήσουμε στον τύπο της συνάρτησης επιβίωσης, την τιμή της μεταβλητής που θεωρούμε αντιπροσωπευτική κάθε κατηγορίας. Για παράδειγμα, χρησιμοποιώντας τη μέση προβλεπόμενη τιμή της μεταβλητής για κάθε κατηγορία.

Παρατήρηση:

Όπως στον προηγούμενο τρόπο ελέγχου της υπόθεσης της αναλογικότητας κινδύνων, όπου προέκυπτε το ερώτημα πως καθορίζεται αν δύο γραφικές παραστάσεις είναι επαρκώς παράλληλες, έτσι και εδώ τίθεται το ερώτημα πώς αποφασίζουμε αν οι γραφικές παραστάσεις είναι επαρκώς κοντά. Προτείνεται να θεωρούμε ότι η υπόθεση της αναλογικότητας κινδύνων δεν ισχύει μόνο όταν οι παρατηρούμενες και οι αναμενόμενες συναρτήσεις επιβίωσης παρουσιάζουν έντονη απόκλιση.

Κάνουμε την παρουσίαση του ελέγχου αναλογικότητας κινδύνων με χρήση των παρατηρούμενων έναντι των αναμενόμενων συναρτήσεων επιβίωσης για την συμεταβλητή `GROUP` στο μοντέλο `cns.cox`, το οποίο έχουμε προσαρμόσει. Με τον κώδικα που ακολουθεί, παρατίθενται στο ίδιο γράφημα οι παρατηρούμενες συναρτήσεις επιβίωσης, όταν μεταβλητή στρωματοποίησης είναι η `GROUP` (για τον λόγο αυτό γίνεται χρήση του `cns` μοντέλου) και οι αναμενόμενες συναρτήσεις επιβίωσης, για τις δυο κατηγορίες της μεταβλητής `GROUP`.

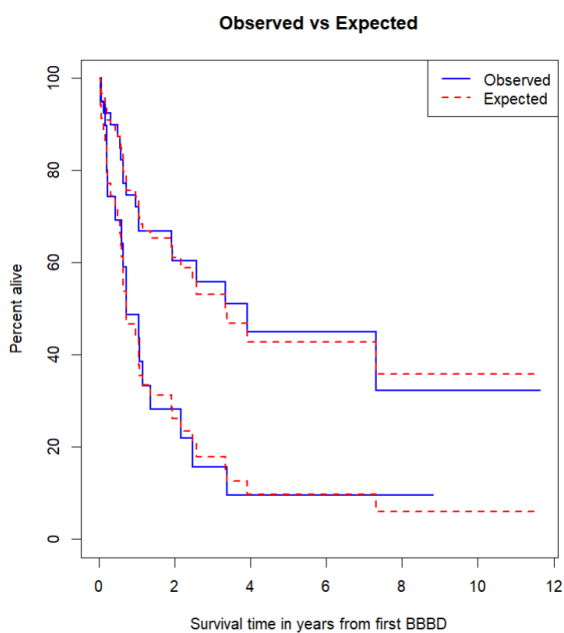
```
cns <- coxph(Surv(B3TODEATH,STATUS)~strata(GROUP))
```

```
#observed survival plots K-M
survfit <- survfit(cns)
```

```
#expected plots
cns.2 <- coxph(Surv(B3TODEATH,STATUS)~GROUP)
new_data <- data.frame(GROUP = c(0, 1))
survfit.2 <- survfit(cns.2,newdata =new_data)

#both in one graph
plot(survfit,col="blue",lty=1,lwd=2,cex=3,
xlab="Survival time in years from first BBBD",
ylab="Percent alive",yscale=100,main = "Observed vs Expected")
lines(survfit.2,col="red",lty=2,lwd=2,cex=3)

legend("topright", legend = c("Observed", "Expected"),
      col = c("blue", "red"), lwd = 2, lty = c(1, 2))
```



Σχήμα 2.6: Γραφική παράσταση των παρατηρούμενων έναντι των αναμενόμενων συναρτήσεων επιβίωσης, έναντι του χρόνου, για τον έλεγχο της υπόθεσης αναλογικότητας κινδύνων της μεταβλητής `GROUP` στο μοντέλο `cns.cox`.

Από τη γραφική παράσταση του Σχήματος 2.6 συμπεραίνουμε ότι ισχύει η υπό-

θεση της αναλογικότητας κινδύνων για τη μεταβλητή GROUP, γιατί οι παρατηρούμενες και οι αναμενόμενες συναρτήσεις επιβίωσης είναι κοντινές.

Στατιστικός τρόπος ελέγχου: τεστ των Grambsch and Thernau

Οι Grambsch and Thernau (1994), βασιζόμενοι στην προσέγγιση των Fleming and Harrington (1991), ανέπτυξαν το ακόλουθο τεστ. Θεωρούμε κάθε άτομο ως μια ανεξάρτητη διαδικασία καταμέτρησης (independent counting process) $\{N_i(t), t \geq 0, i = 1, \dots, n\}$, με $N_i(t) = I_{\{Y \leq t, \delta=1\}}$ και I_A να είναι η δείκτρια συνάρτηση η οποία παίρνει την τιμή 1 όταν το γεγονός A έχει συμβεί και την τιμή 0 όταν το γεγονός A δεν έχει συμβεί. Οπότε η συνάρτηση $N_i(t)$ εκφράζει αν το i -οστό άτομο έχει βιώσει το υπό μελέτη γεγονός τη χρονική στιγμή t . Η συνάρτηση έντασης (intensity function) είναι της μορφής

$$\lambda_i(t) = K_i(t) \exp\{\beta' \mathbf{X}_i(t)\} dH_0,$$

με $K_i(t)$ να έχει δυνατές τιμές 0 και 1 και δηλώνει αν το i -οστό άτομο είναι σε κίνδυνο τη χρονική στιγμή t , $i = 1, \dots, n$, β είναι το διάνυσμα των συντελεστών, $\mathbf{X}_i(t)$ είναι ένα m -διάστατο διάνυσμα συμμεταβλητών και dH_0 είναι μια απροσδιόριστη συνάρτηση κινδύνου. Η συνάρτηση έντασης εκφράζει τον στιγμιαίο ρυθμό εμφάνισης του υπό μελέτη γεγονότος τη χρονική στιγμή t για το i -οστό άτομο. Όταν οι συντελεστές β_j , $j = 1, \dots, m$ είναι χρονικά εξαρτημένοι δεν ισχύει πλέον η υπόθεση της αναλογικότητας κινδύνων στο μοντέλο αναλογικών κινδύνων του Cox. Έστω $\beta_j(t) = \beta_j + \gamma_j g_j(t)$ με $g_j(t)$ να είναι μια προβλέψιμη διαδικασία και θεωρούμε ότι η συνάρτηση έντασης είναι της μορφής

$$\lambda_i(t) = K_i(t) \exp \left\{ \sum_{j=1}^m \beta_j(t) X_{ij}(t) \right\} dH_0.$$

Υποθέτουμε ότι η $g(t)$ παίρνει τιμές γύρω από το 0, χωρίς ωστόσο να είναι απαραίτητη προϋπόθεση. Οι Grambsch and Thernau έδειξαν ότι $E\{r_{Sk}(\beta)\} \approx V(\beta, t_k)G(t_k)\gamma$, με r_{Sk} να είναι το k -οστό Schoenfeld υπόλοιπο, $G(t)$ είναι μια προσδιορισμένη συνάρτηση του χρόνου, είναι ένας $m \times m$ διαγώνιος πίνακας για τον οποίο ισχύει $G_{jj}(t) = g_j(t)$. Το τεστ για τον έλεγχο της $H_0 : \beta(t) = \beta$ είναι ισοδύναμο με ένα γενικευμένο τεστ ελαχίστων τετραγώνων για τα Schoenfeld υπόλοιπα και χρησιμοποιείται για να ελεγχθεί η υπόθεση της αναλογικότητας κινδύνων. Έστω $r_{Sk}^* = r_{Sk}^*(\beta) = V^{-1}(\beta, t_k)r_{Sk}(\beta)$ να είναι τα scaled Schoenfeld υπόλοιπα και $\hat{\beta}$ να είναι ο εκτιμητής μέγιστης μερικής πιθανοφάνειας του β υπό την H_0 . Συμβολίζουμε με $\hat{V}_k$ την ποσότητα $V(\hat{\beta}, t_k)$ και με $\hat{r}_{Sk}$ την ποσότητα

$r_{Sk}(\hat{\beta})$. Οι γενικευμένοι εκτιμητές ελαχίστων τετραγώνων μας δίνουν

$$\hat{\gamma} = D^{-1} \sum_{k=1}^n G_k \hat{r}_{Sk},$$

με

$$D = \sum_{k=1}^n G_k \hat{V}_k G_k' - \left(\sum_{k=1}^n G_k \hat{V}_k \right) \left(\sum_{k=1}^n \hat{V}_k \right)^{-1} \left(\sum_{k=1}^n G_k \hat{V}_k \right)'.$$

Λόγω αυτού, οδηγούμαστε σε μια στατιστική συνάρτηση που έχει ασυμπτωτικά μια $\chi^2_{(m)}$ κατανομή, η οποία ορίζεται ως:

$$S = \left(\sum_{k=1}^n G_k \hat{r}_{Sk} \right)' D^{-1} \left(\sum_{k=1}^n G_k \hat{r}_{Sk} \right).$$

Η στατιστική συνάρτηση S χρησιμοποιείται για τον έλεγχο της υπόθεσης $H_0 : (\beta, \gamma) = (\hat{\beta}, \mathbf{0})$. Όταν ισχύει η μπορούμε να συμπεράνουμε ότι ισχύει η υπόθεση της αναλογικότητας κινδύνων.

Για το μοντέλο `cns.cox` που έχουμε προσαρμόσει κάνουμε τον παραπάνω έλεγχο της υπόθεσης της αναλογικότητας κινδύνων για την συμμεταβλητή `GROUP` με χρήση της συνάρτησης `cox.zph`. Δίνουμε στη συνάρτηση `cox.zph` ως όρισμα το μοντέλο `cns.cox` και λαμβάνουμε τα αποτελέσματα που ακολουθούν.

```
PH.test <- cox.zph(cns.cox)
PH.test
      chisq df    p
GROUP  0.199  1 0.66
GLOBAL 0.199  1 0.66
```

Παρατηρούμε ότι η p -τιμή ισούται με $0.66 > 0.05$, συνεπώς ισχύει η υπόθεση της αναλογικότητας κινδύνων για τη συμμεταβλητή `GROUP`. Η p -τιμή που αναφέρεται στο συνολικό μοντέλο (`GLOBAL`) εξετάζει αν ισχύει η υπόθεση της αναλογικότητας κινδύνων συνολικά στο μοντέλο ή αν παραβιάζεται για οποιαδήποτε από τις συμμεταβλητές. Στο συγκεκριμένο μοντέλο οι δύο p -τιμές ταυτίζονται, γιατί έχουμε μόνο μια συμμεταβλητή.

2.8 Στρωματοποίηση

Μέχρι στιγμής στα παραδείγματα που παρουσιάσαμε ίσχυε ο έλεγχος της υπόθεσης της αναλογικότητας κινδύνων. Σε αυτή την ενότητα θα παρουσιάσουμε

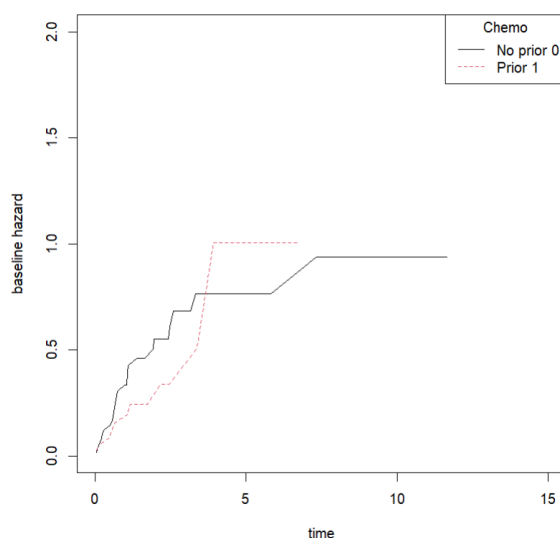
ένα παράδειγμα για το οποίο δεν ισχύει η υπόθεση της αναλογικότητας κινδύνων και έπειτα θα παραθέσουμε έναν τρόπο επίλυσης της παραβίασης της υπόθεσης, τη στρωματοποίηση. Έστω ότι έχουμε το μοντέλο του Cox με συμμεταβλητές την **GROUP** και την **CHEMOPRIOR**.

```
cns.cox1<- coxph(Surv(B3TODEATH,STATUS)~CHEMOPRIOR+GROUP)
```

Ξεκινάμε με τον έλεγχο της υπόθεσης της αναλογικότητας κινδύνων για τη μεταβλητή **CHEMOPRIOR**, με χρήση της γραφικής παράστασης των συναρτήσεων κινδύνου για τις δύο κατηγορίες της. Για την υλοποίηση του ελέγχου αναλογικότητας στην R ακολουθούμε τα επόμενα βήματα. Αρχικά, χωρίζουμε όλα τα δεδομένα σε δύο υποσύνολα δεδομένων. Το πρώτο σύνολο δεδομένων, **group_0**, περιέχει όλα τα δεδομένα για τα οποία η μεταβλητή **CHEMOPRIOR** έχει την τιμή 0, ενώ το δεύτερο, **group_1**, αυτά που η τιμή της μεταβλητής **CHEMOPRIOR** είναι 1. Ο λόγος που χωρίζουμε τα δεδομένα είναι γιατί θέλουμε να προσαρμόσουμε δύο μοντέλα του Cox, ένα για κάθε κατηγορία της **CHEMOPRIOR**, αλλά καθώς σε αυτά τα μοντέλα θα συμπεριλάβουμε την μεταβλητή **GROUP** θέλουμε οι τιμές της να είναι και αυτές χωρισμένες με βάση την τιμή που έχει η μεταβλητή **CHEMOPRIOR** σε κάθε μέτρηση. Στο αντικείμενο **h0t1.1** αποθηκεύουμε το μοντέλο του Cox με μεταβλητή πρόβλεψης την **GROUP** και με χρήση των δεδομένων **group_0**, ενώ στο αντικείμενο **h0t1.2** αποθηκεύουμε το μοντέλο του Cox με μεταβλητή πρόβλεψης την **GROUP** και με χρήση των δεδομένων **group_1**. Η συνάρτηση **basehaz** υπολογίζει την εκτιμώμενη βασική συνάρτηση κινδύνου για το μοντέλο του Cox που δηλώνεται ως πρώτο όρισμα της. Οπότε στο αντικείμενο **h0t1** αποθηκεύεται η εκτιμώμενη βασική συνάρτηση κινδύνου για το μοντέλο **h0t1.1**, ενώ στο **h0t2** για το μοντέλο **h0t1.2**. Η εντολή **plot** κάνει τη γραφική παράσταση της εκτιμώμενης βασικής συνάρτησης κινδύνου, που αποθηκεύσαμε στο αντικείμενο **h0t1**, έναντι του χρόνου. Η εντολή **lines** κάνει στην υπάρχουσα γραφική παράσταση τη γραφική της εκτιμώμενης βασικής συνάρτησης κινδύνου, που αποθηκεύσαμε στο αντικείμενο **h0t2**, έναντι του χρόνου. Η εντολή **legend** προσθέτει την κατάλληλη λεζάντα.

```
group_0 <- cns2[cns2$CHEMOPRIOR == 0, ]
group_1 <- cns2[cns2$CHEMOPRIOR == 1, ]
h0t1.1 <- coxph(Surv(B3TODEATH, STATUS) ~ GROUP,
  data = group_0)
h0t1.2 <- coxph(Surv(B3TODEATH, STATUS) ~ GROUP,
  data = group_1)
h0t1 <- basehaz(h0t1.1, centered = TRUE)
h0t2 <- basehaz(h0t1.2, centered = TRUE)
```

```
plot(h0t1$time, h0t1$hazard, type = "l", col = 1, xlab="time",
     ,ylab="baseline hazard",xlim = c(0, 15), ylim = c(0, 2))
lines(h0t2$time, h0t2$hazard, type = "l", xlab="time",
     ,ylab="baseline hazard",lty = 2, col = 2)
legend("topright", legend = c("No prior 0", "Prior 1"),
     col = c(1, 2), lty = c(1, 2),title="Chemo")
```



Σχήμα 2.7: Έλεγχος της υπόθεσης της αναλογικότητας κινδύνων για την μεταβλητή CHEMOPRIOR στο μοντέλο `cns.cox1`. Ο έλεγχος γίνεται μέσω της γραφικής παράστασης των συναρτήσεων κινδύνου, έναντι του χρόνου, για της κατηγορίες της μεταβλητής CHEMOPRIOR.

Από την γραφική παράσταση του Σχήματος 2.7 συμπεραίνουμε ότι δεν ισχύει η υπόθεση της αναλογικότητας κινδύνων για τη μεταβλητή CHEMOPRIOR, καθώς οι συναρτήσεις κινδύνου που αντιστοιχούν στις δύο κατηγορίες της διασταυρώνονται. Ο πιο απλός τρόπος διόρθωσης της παραβίασης της υπόθεσης της αναλογικότητας κινδύνων είναι μέσω στρωματοποίησης, η οποία παρουσιάζεται στη συνέχεια.

Στρωματοποίηση

Υπάρχουν περιπτώσεις που τα δεδομένα που μελετάμε προέρχονται από διάφορα στρώματα του ίδιου πληθυσμού. Σε αυτές τις περιπτώσεις το μοντέλο αναλογικών

κινδύνων του Cox επεκτείνεται σε ένα μοντέλο για όλο τον πληθυσμό, το οποίο έχει διαφορετική βασική συνάρτηση κινδύνου σε κάθε στρώμα και ίδιους εκτιμητές των παραμέτρων για κάθε στρώμα. Ένας σημαντικός λόγος που κάνουμε στρωματοποίηση είναι, γιατί είναι ένας τρόπος να αντιμετωπίσουμε την παραβίαση της υπόθεσης της αναλογικότητας κινδύνων. Στρωματοποιούμε ως προς την μεταβλητή για την οποία δεν ισχύει η υπόθεση της αναλογικότητας κινδύνων, έχοντας προσαρμοσμένες τις υπόλοιπες συμμεταβλητές. Η μεταβλητή ως προς την οποία γίνεται η στρωματοποίηση είναι μία κατηγορική μεταβλητή, όπως είναι το φύλο ή η ομάδα, ωστόσο δεν αντιπροσωπεύει απαραίτητα κάποιο κατηγορικό φαινόμενο, καθώς μπορεί να επιλέξουμε να χωρίσουμε μια συνεχή μεταβλητή σε ομάδες. Με αυτόν τον τρόπο θεωρούμε ότι το μοντέλο αναλογικών κινδύνων είναι κατάλληλο για το εσωτερικό κάθε στρώματος και περιλαμβάνει σε κάθε στρώμα όλες τις συμμεταβλητές, εκτός από τη συμμεταβλητή ως προς την οποία γίνεται η στρωματοποίηση. Με τη στρωματοποίηση αποκτούμε μη παραμετρική εκτίμηση των συναρτήσεων κινδύνου για τα διάφορα επίπεδα, έχοντας προσαρμόσει τις υπόλοιπες συμμεταβλητές. Απεικονίζοντας γραφικά τις συναρτήσεις κινδύνου, παρατηρούμε την εκτιμώμενη επίδραση της κατηγορικής μεταβλητής, αφού έχουμε προσαρμόσει την επίδραση των υπόλοιπων συμμεταβλητών.

- Αν οι καμπύλες διασταυρώνονται συμπεραίνουμε ότι δεν ισχύει η υπόθεση της αναλογικότητας κινδύνων ως προς τη μεταβλητή στρωματοποίησης. Αυτό συμβαίνει λόγω της ύπαρξης ενός παράγοντα αλληλεπίδρασης που δεν υπάρχει στο μοντέλο. Για να επιλυθεί η παραβίαση της υπόθεσης της αναλογικότητας κινδύνων, είτε ψάχνουμε τους κατάλληλους όρους ή όρο αλληλεπίδρασης που πρέπει να συμπεριληφθούν στο μοντέλο, είτε συνεχίζουμε με το στρωματοποιημένο μοντέλο.
- Αν οι καμπύλες δεν τέμνονται, συμπεραίνουμε ότι οι όροι αλληλεπίδρασης επεξηγούνται καλά από το μοντέλο που έχει οριστεί και ισχύει το μοντέλο αναλογικών κινδύνων του Cox. Σε αυτήν την περίπτωση γίνεται χρήση του μοντέλου αναλογικών κινδύνων του Cox, χωρίς στρωματοποίηση. Το μειονέκτημα, όταν στρωματοποιούμε και ισχύει η υπόθεση της αναλογικότητας, είναι ότι δεν μπορούμε να παρατηρήσουμε τους εκτιμώμενους συντελεστές της κατηγορικής μεταβλητής.

Συνοπτικά οι λόγοι για τους οποίους στρωματοποιούμε είναι οι εξής:

1. Είναι ένας τρόπος να αντιμετωπίσουμε την παραβίαση της υπόθεσης της αναλογικότητας κινδύνων. Στρωματοποιούμε ως προς τη μεταβλητή για την οποία δεν ισχύει η υπόθεση της αναλογικότητας κινδύνων.

2. Μπορούμε να τη χρησιμοποιήσουμε για να κάνουμε γραφικούς ελέγχους της υπόθεσης της αναλογικότητας κινδύνων. Οι τρόποι αυτοί έχουν παρουσιαστεί στην προηγούμενη ενότητα.
3. Μπορούμε να προσαρμόσουμε έναν παράγοντα στο μοντέλο, χωρίς να εκτιμήσουμε τη συνεισφορά του.

Παρόλο που η στρωματοποίηση είναι ο πιο απλός τρόπος επίλυσης της παραβίασης της υπόθεσης της αναλογικότητας κινδύνων, υπάρχουν διάφοροι τρόποι ανάλυσης. Αυτοί είναι οι εξής:

1. Ανάλυση με στρωματοποίηση ως προς τη μεταβλητή ενδιαφέροντος, όπου δεν προσαρμόζουμε κάποιο μοντέλο παλινδρόμησης, αλλά παρατηρούμε την καμπύλη Kaplan Meier για κάθε ομάδα ξεχωριστά. Αν υπάρχουν και άλλες συμμεταβλητές στο μοντέλο χρησιμοποιούμε το Cox μοντέλο με στρωματοποίηση ως προς τη μεταβλητή ενδιαφέροντος.
2. Ξεκινάμε την ανάλυση μετά τη χρονική στιγμή που διασταυρώνονται οι καμπύλες των συναρτήσεων κινδύνου και χρησιμοποιούμε Cox PH μοντέλο.
3. Προσαρμόζουμε ένα Cox PH μοντέλο που αντιστοιχεί στη χρονική περίοδο πριν από τη διασταύρωση των καμπυλών των συναρτήσεων κινδύνου και ένα διαφορετικό Cox PH μοντέλο για μετά τη διασταύρωση των καμπυλών των συναρτήσεων κινδύνου. Έτσι λαμβάνουμε δύο διαφορετικούς εκτιμητές του HR, έναν για κάθε χρονική περίοδο.
4. Προσαρμόζουμε ένα Extended Cox PH μοντέλο. Σε αυτήν την περίπτωση, επειδή για κάποιες από τις συμμεταβλητές δεν ισχύει η υπόθεση της αναλογικότητας κινδύνων, θα πρέπει οι συγκεκριμένες συμμεταβλητές να αλληλεπιδρούν με μία συνάρτηση του χρόνου ή κάθε μία. Οπότε προσαρμόζουμε ένα Extended Cox PH μοντέλο, το οποίο λαμβάνει υπόψη τις συναρτήσεις αλληλεπίδρασης των συμμεταβλητών με τον χρόνο.
5. Χρησιμοποιούμε censored regression quantile προσέγγιση, η οποία είναι μια μη παραμετρική προσέγγιση και δεν απαιτεί να ισχύει η υπόθεση της αναλογικότητας κινδύνων για να είναι εφαρμόσιμη.

Σύμφωνα με τα παραπάνω για να συνεχίσουμε με το μοντέλο `cns.cox1`, στο οποίο είδαμε ότι παραβιάζεται η υπόθεση της αναλογικότητας κινδύνων για τη μεταβλητή `CHEMOPRIOR`, κάνουμε στρωματοποίηση ως προς τη συμμεταβλητή `CHEMOPRIOR` έχοντας προσαρμόσμενη στο μοντέλο τη συμμεταβλητή `GROUP`.

```
cns.cox1.1<- coxph(Surv(B3TODEATH,STATUS)~strata(CHEMOPRIOR)
                  +GROUP)
summary(cns.cox1.1)
```

Τα αποτελέσματα που λαμβάνουμε είναι τα εξής:

```
      coef      exp(coef) se(coef) z      Pr(>|z|)
GROUP 1.1211      3.0683   0.3585  3.127  0.00177 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

      exp(coef) exp(-coef) lower .95 upper .95
GROUP      3.068      0.3259      1.52      6.196

Concordance= 0.618 (se = 0.042 )
Likelihood ratio test= 9.29 on 1 df, p=0.002
Wald test              = 9.78 on 1 df, p=0.002
Score (logrank) test = 10.51 on 1 df, p=0.001
```

Παρατηρήσεις:

- Η εκτίμηση της παραμέτρου του μοντέλου για τη μεταβλητή **GROUP** είναι 1.1211. Αυτό σημαίνει ότι ασθενείς που ανήκουν στην ομάδα 1, δηλαδή έχουν λάβει ακτινοβολία πριν την BBBD χημειοθεραπεία, έχουν συνάρτηση κινδύνου $\exp(\hat{\beta}\text{GROUP}) = \exp(1.1211\text{GROUP})$, ενώ η βασική συνάρτηση κινδύνου θεωρούμε ότι είναι σταθερή και μπορούμε να υποθέσουμε ότι $h_0(t) = 1$. Επομένως, διατρέχουν μεγαλύτερο κίνδυνο, έχουν μικρότερο αναμενόμενο χρόνο ζωής σε σχέση με τους ασθενείς της ομάδας 0.
- Το τυπικό σφάλμα του εκτιμητή της παραμέτρου είναι 0.3585.
- Η επεξηγηματική μεταβλητή **GROUP** είναι στατιστικά σημαντική, γιατί για τον έλεγχο της υπόθεσης $H_0 : \beta = 0$ έχουμε p -τιμή = 0.00177 < 0.05, που σημαίνει ότι σε επίπεδο σημαντικότητας $\alpha = 0.05$ η $H_0 : \beta = 0$ απορρίπτεται.
- Για τη σύγκριση του μοντέλου που έχουμε προσαρμόσει με το μοντέλο χωρίς επεξηγηματικές μεταβλητές χρησιμοποιείται το LRT. Η τιμή του στατιστικού του LRT είναι 9.29 με 1 βαθμό ελευθερίας και p -τιμή=0.002.

Κεφάλαιο 2 2.9. Το μοντέλο αναλογικών κινδύνων του Cox στα CNS δεδομένα

Καθώς η p -τιμή $= 0.002 < 0.05$, σύμφωνα με το LRT επιλέγουμε το μοντέλο που προσαρμόσαμε.

2.9 Το μοντέλο αναλογικών κινδύνων του Cox στα CNS δεδομένα

Σε αυτήν την ενότητα παρατίθεται ένα πιο σύνθετο παράδειγμα. Σκοπός του παραδείγματος είναι να δείξουμε πως ελέγχεται η υπόθεση της αναλογικότητας κινδύνων για το μοντέλο του Cox που περιέχει περισσότερες από μία συμεταβλητές. Γίνεται χρήση των δεδομένων CNS, τα οποία έχουν παρουσιαστεί στην Ενότητα 1.2. Έστω ότι θέλουμε να συμπεριλάβουμε στο μοντέλο μια μεταβλητή δείκτη την AGE60, που ορίζεται ως $AGE60=1$ αν $AGE \leq 60$ και 0 αλλιώς. Η δημιουργία αυτής της μεταβλητής γίνεται με τη συνάρτηση `factor` με τον παρακάτω κώδικα.

```
cns2$AGE60<-as.integer(cns$AGE<=60) .
```

Θεωρούμε ότι μπορούμε να έχουμε μέχρι δευτέρου βαθμού αλληλεπίδραση στο μοντέλο. Δηλαδή, θεωρούμε ότι η επίδραση μιας συμεταβλητής στον χρόνο ζωής επηρεάζεται από τη σχέση που έχει με μια άλλη, το πολύ, συμεταβλητή. Έστω ότι θέλουμε να προσαρμόσουμε το Cox μοντέλο με συμεταβλητές τις `KPS.PRE.`, `GROUP`, `SEX`, `AGE60` και `SEX:AGE60`, με `SEX:AGE60` να είναι η συμεταβλητή που αντιστοιχεί στον όρο αλληλεπίδρασης μεταξύ των συμεταβλητών `SEX` και `AGE60`. Οι εντολές που χρησιμοποιούμε για την προσαρμογή αυτού του μοντέλου είναι οι παρακάτω. Με την εντολή `coxph` προσαρμόζουμε το μοντέλο του Cox και το αποθηκεύουμε στο αντικείμενο (object) `cns2.coxint6`.

```
cns2.coxint6 <- coxph(Surv(B3TODEATH,STATUS)~KPS.PRE.+GROUP  
+SEX+AGE60+SEX:AGE60)
```

Έστω ότι θέλουμε να ελέγξουμε την υπόθεση αναλογικότητας κινδύνων ως προς τη μεταβλητή `GROUP`, στο μοντέλο που έχουμε προσαρμόσει. Η μεταβλητή `GROUP` είναι κατηγορική και έχει δύο πιθανές τιμές, $K = 2$. Ξεκινάμε κάνοντας τον έλεγχο με βάση τις συναρτήσεις επιβίωσης. Για να στρωματοποιήσουμε ως προς τη μεταβλητή `GROUP` έχοντας προσαρμόσει τις μεταβλητές `KPS.PRE.`, `SEX`, `AGE60` και `SEX:AGE60` ως συμεταβλητές χρησιμοποιούμε την εντολή `coxph`. Δίνουμε ως πρώτο όρισμα στην εντολή `coxph` το `Surv`, που περιέχει τους χρόνους ζωής ή τους χρόνους λογοκρισίας και τους δείκτες λογοκρισίας των ασθενών. Δεξιά

Κεφάλαιο 2 2.9. Το μοντέλο αναλογικών κινδύνων του Cox στα CNS δεδομένα

του συμβόλου ~, στο εσωτερικό της εντολής `coxph`, με τον όρο `strata(GROUP)` ορίζουμε ως μεταβλητή στρωματοποίησης την `GROUP` και δηλώνουμε ποιες συμ-
μεταβλητές θα περιέχει το μοντέλο. Στη συνέχεια με την εντολή `cns2.coxint7`
λαμβάνουμε την πληροφορία που περιέχεται στο μοντέλο που προσαρμόσαμε.

```
cns2.coxint7 <-coxph(Surv(B3TODEATH,STATUS)~strata(GROUP)+  
  KPS.PRE.+SEX+AGE60+SEX:AGE60)  
cns2.coxint7
```

	coef	exp(coef)	se(coef)	z	p
KPS.PRE.	-0.03260	0.96792	0.01078	-3.025	0.00248
SEX	-2.20284	0.11049	0.71952	-3.062	0.00220
AGE60	-1.12783	0.32373	0.47778	-2.361	0.01825
SEX:AGE60	2.25760	9.56016	0.87850	2.570	0.01017

Likelihood ratio test=20.31 on 4 df, p=0.0004329
n= 58, number of events= 36

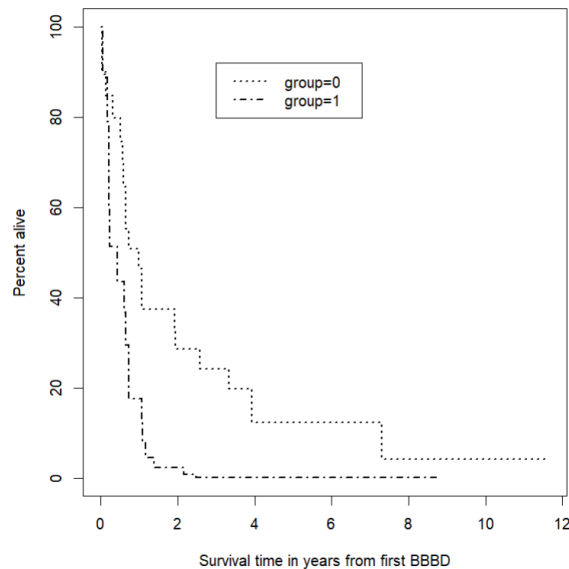
Προκειμένου να κάνουμε τη γραφική παράσταση των συναρτήσεων επιβίωσης για το στρωματοποιημένο μοντέλο του Cox, `cns2.coxint7` χρησιμοποιούμε τον κώδικα που ακολουθεί. Η εντολή `survfit` υπολογίζει τις εκτιμώμενες συναρτήσεις επιβίωσης για το μοντέλο `cns2.coxint7`. Η εκτίμηση αυτή αποθηκεύεται στο αντικείμενο `survfit.int7`. Η εντολή `plot` κάνει τη γραφική παράσταση των εκτιμώμενων συναρτήσεων επιβίωσης και δέχεται ως πρώτο όρισμα το αντικείμενο `survfit.int7`. Στα ορίσματα `xlab=""`, `ylab=""` εσωτερικά των εισαγωγικών δίνεται η λεζάντα που εμφανίζεται στον x και y άξονα του γραφήματος, αντίστοιχα. Τα υπόλοιπα ορίσματα αφορούν οπτικά χαρακτηριστικά της γραφικής παράστασης (χρώμα, πάχος γραμμής) και δεν σχολιάζονται. Η συνάρτηση `legend` προσθέτει στη γραφική παράσταση μια λεζάντα η οποία υποδεικνύει ποια συνάρτηση επιβίωσης αντιστοιχεί σε ποιο επίπεδο της μεταβλητής `GROUP`.

```
survfit.int7 <- survfit(cns2.coxint7)  
plot(survfit.int7,col=1,lty=3:4,lwd=2,cex=3,  
  xlab="Survival time in years from first BBBD",  
  ylab="Percent alive",yscale=100)  
legend(3.0,.92,c("group=0","group=1"),lty=3:4,lwd=2)
```

Συμπέρασμα:

Από τη γραφική παράσταση του Σχήματος 2.8 συμπεραίνουμε ότι ισχύει η υπόθεση της αναλογικότητας κινδύνου ως προς τη μεταβλητή `GROUP`, για το μοντέλο `cns2.coxint6`, καθώς οι δύο καμπύλες δεν διασταυρώνονται.

Κεφάλαιο 2 2.9. Το μοντέλο αναλογικών κινδύνων του Cox στα CNS δεδομένα



Σχήμα 2.8: Γραφική παράσταση των εκτιμώμενων συναρτήσεων επιβίωσης για τις κατηγορίες της μεταβλητής **GROUP**, έναντι του χρόνου. Σκοπός είναι ο έλεγχος της υπόθεσης της αναλογικότητας κινδύνων για την μεταβλητή **GROUP** στο μοντέλο `cns2.coxint7`.

Συνεχίζουμε κάνοντας τον έλεγχο της υπόθεσης αναλογικότητας κινδύνων, για τη μεταβλητή **GROUP** στο μοντέλο `cns2.coxint6`, με βάση τη γραφική παράσταση των συναρτήσεων κινδύνου για τις κατηγορίες της μεταβλητής **GROUP**. Στο αντικείμενο `cns2` είναι αποθηκευμένα τα δεδομένα μας. Ξεκινάμε αποθηκεύοντας όλα τα δεδομένα που αντιστοιχούν σε τιμή 0 για τη μεταβλητή **GROUP** σε ένα νέο σύνολο δεδομένων το `group_0` και όλα τα δεδομένα που αντιστοιχούν σε τιμή 1 για τη μεταβλητή **GROUP** σε ένα νέο σύνολο δεδομένων το `group_1`. Ο λόγος που το κάνουμε αυτό είναι γιατί θέλουμε να προσαρμόσουμε δύο μοντέλα του Cox, ένα για κάθε κατηγορία της **GROUP**, αλλά καθώς σε αυτά τα μοντέλα θα συμπεριλάβουμε συμμεταβλητές θέλουμε οι τιμές των συμμεταβλητών να είναι και αυτές χωρισμένες με βάση την τιμή που έχει η μεταβλητή **GROUP** σε κάθε μέτρηση. Στη συνάρτηση `Surv` δίνουμε ως πρώτο όρισμα τους χρόνους ζωής ή λογοχρισίας και ως δεύτερο τους δείκτες λογοχρισίας των ασθενών. Η εντολή `coxph` θα προσαρμόσει το μοντέλο του Cox για τα δυο σύνολα δεδομένων που ορίσαμε. Δέχεται ως πρώτο όρισμα τη συνάρτηση `Surv`, δεξιά του συμβόλου `~` γράφουμε τις συμ-

Κεφάλαιο 2 2.9. Το μοντέλο αναλογικών κινδύνων του Cox στα CNS δεδομένα

μεταβλητές που θέλουμε στο μοντέλο, εκτός από την `GROUP`, στο όρισμα `data` ορίζουμε το σύνολο των δεδομένων από το οποίο θα ληφθούν οι τιμές για την προσαρμογή του μοντέλου. Συνεπώς, στο αντικείμενο `h0t1.1` προσαρμόζεται το μοντέλο του Cox για τα δεδομένα `group_0` και στο `h0t1.2` προσαρμόζεται το μοντέλο του Cox για τα δεδομένα `group_1`. Η συνάρτηση `basehaz` υπολογίζει την εκτιμώμενη βασική συνάρτηση κινδύνου για το μοντέλο του Cox που δηλώνεται ως πρώτο όρισμα της. Οπότε, στο αντικείμενο `h0t1` αποθηκεύεται η εκτιμώμενη βασική συνάρτηση κινδύνου για το μοντέλο `h0t1.1`, ενώ στο `h0t2` για το μοντέλο `h0t1.2`. Η εντολή `plot` κάνει τη γραφική παράσταση της εκτιμώμενης βασικής συνάρτησης κινδύνου, που αποθηκεύσαμε στο αντικείμενο `h0t1`, έναντι του χρόνου. Η εντολή `lines` κάνει στην υπάρχουσα γραφική παράσταση τη γραφική της εκτιμώμενης βασικής συνάρτησης κινδύνου, που αποθηκεύσαμε στο αντικείμενο `h0t2`, έναντι του χρόνου. Τέλος, η εντολή `legend` προσθέτει την κατάλληλη λεζάντα.

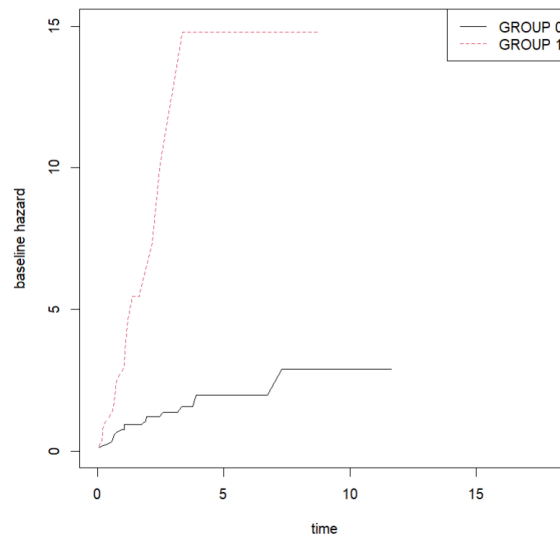
```
group_0 <- cns2[cns2$GROUP == 0, ]
group_1 <- cns2[cns2$GROUP == 1, ]
h0t1.1 <- coxph(Surv(B3TODEATH, STATUS) ~ KPS.PRE. + SEX
               + AGE60 + SEX:AGE60, data = group_0)
h0t1.2 <- coxph(Surv(B3TODEATH, STATUS) ~ KPS.PRE. + SEX
               + AGE60 + SEX:AGE60, data = group_1)
h0t1 <- basehaz(h0t1.1, centered = TRUE)
h0t2 <- basehaz(h0t1.2, centered = TRUE)
plot(h0t1$time, h0t1$hazard, type = "l", col = 1,
     xlab="time", ylab="baseline hazard",
     xlim = c(0, 18), ylim = c(0, 15))
lines(h0t2$time, h0t2$hazard, type = "l",
     xlab="time", ylab="baseline hazard", lty = 2, col = 2)
legend("topright", legend = c("GROUP 0", "GROUP 1"),
     col = c(1, 2), lty = c(1, 2))
```

Συμπέρασμα:

Με βάση τις γραφικές παραστάσεις των βασικών συναρτήσεων κινδύνου για τις δύο κατηγορίες της μεταβλητής `GROUP`, όπως φαίνονται στο Σχήμα 2.9, παρατηρούμε ότι οι καμπύλες δεν διασταυρώνονται. Επομένως, θεωρούμε ότι ισχύει η υπόθεση της αναλογικότητας κινδύνων για τη μεταβλητή `GROUP` στο μοντέλο `cns2.coxint6`.

Σκοπός μας είναι ο έλεγχος της υπόθεσης της αναλογικότητας κινδύνων για τη συμμεταβλητή `GROUP` στο μοντέλο `cns2.coxint6`. Ο κώδικας που ακολουθεί

Κεφάλαιο 2 2.9. Το μοντέλο αναλογικών κινδύνων του Cox στα CNS δεδομένα



Σχήμα 2.9: Γραφική παράσταση των συναρτήσεων κινδύνου για τις κατηγορίες της μεταβλητής `GROUP`, έναντι του χρόνου. Σκοπός είναι ο έλεγχος της υπόθεσης της αναλογικότητας κινδύνων για την μεταβλητή `GROUP` στο μοντέλο `cns2.coxint6`.

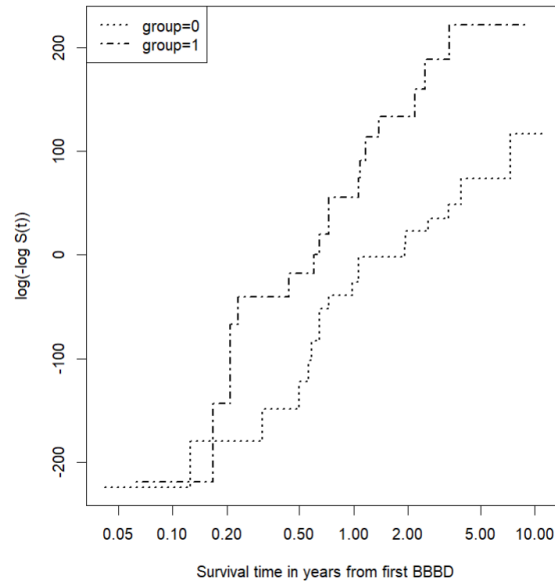
δημιουργεί τη γραφική παράσταση των $\log - \log$ συναρτήσεων επιβίωσης έναντι του χρόνου, για το μοντέλο `cns2.coxint7`, όπου μεταβλητή στρωματοποίησης είναι η `GROUP`. Επισημαίνεται ότι πάλι χρησιμοποιείται η εντολή `survfit` για τον υπολογισμό των εκτιμώμενων συναρτήσεων επιβίωσης. Οι εκτιμήσεις αυτές αποθηκεύονται στο αντικείμενο `survfit.int7`. Το πρώτο όρισμα της εντολής `plot` περιέχει τις εκτιμώμενες συναρτήσεις επιβίωσης. Με το όρισμα `fun="cloglog"` στην εντολή `plot` εφαρμόζεται ο φυσικός λογάριθμος δύο φορές στη συνάρτηση επιβίωσης, αρνητικό πρόσημο εφαρμόζεται μόνο στον εσωτερικό λογάριθμο.

```
survfit.int7 <- survfit(cns2.coxint7)
plot(survfit.int7, fun="cloglog", col=1, lty=3:4, lwd=2, cex=3,
     xlab="Survival time in years from first BBBD",
     ylab="log-log Survival", yscale=100)
legend("topleft", c("group=0", "group=1"), lty=3:4, lwd=2)
```

Συμπέρασμα:

Σύμφωνα με το γράφημα του Σχήματος 2.10 μπορούμε να συμπεράνουμε ότι ισχύει η υπόθεση της αναλογικότητας κινδύνων για τη μεταβλητή `GROUP` στο μοντέλο `cns2.coxint6`, καθώς οι δύο γραφικές είναι προσεγγιστικά παράλληλες.

Κεφάλαιο 2 2.9. Το μοντέλο αναλογικών κινδύνων του Cox στα CNS δεδομένα



Σχήμα 2.10: Γραφική παράσταση των $\log - \log$ συναρτήσεων επιβίωσης για τις κατηγορίες της μεταβλητής **GROUP**, με σκοπό τον έλεγχο της υπόθεσης της αναλογικότητας κινδύνων της μεταβλητής **GROUP** στο μοντέλο `cns2.coxint6`.

Ωστόσο, υπάρχει μια επιφύλαξη καθώς για μικρούς χρόνους ζωής οι δύο καμπύλες διασταυρώνονται.

Όπως έχει προαναφερθεί στην Ενότητα 2.7, αν το μοντέλο αναλογικών κινδύνων είναι κατάλληλο για ένα δοθέν σύνολο μεταβλητών πρόβλεψης, αναμένεται οι γραφικές παραστάσεις των $\log - \log$ εμπειρικών συναρτήσεων επιβίωσης για διαφορετικά άτομα της μελέτης να είναι προσεγγιστικά παράλληλες. Έστω ότι θέλουμε να ελέγξουμε αν το μοντέλο `cns2.coxint6` είναι κατάλληλο. Επιλέγουμε τυχαία δύο άτομα της μελέτης, έστω αυτά με αύξοντα αριθμό 12 και 35 και κάνουμε τις γραφικές παραστάσεις των $\log - \log$ εμπειρικών συναρτήσεων επιβίωσης για τα άτομα που επιλέξαμε. Αρχικά, στο αντικείμενο `new_data` αποθηκεύουμε όλα τα δεδομένα που αφορούν τα άτομα με αύξοντα αριθμό 12 και 35. Στην συνέχεια με την εντολή `survfit` υπολογίζονται οι εκτιμώμενες συναρτήσεις επιβίωσης για το μοντέλο `cns2.coxint6`. Δίνοντας ως δεύτερο όρισμα στην εντολή `survfit` το `new_data` γίνεται η εκτίμηση των συναρτήσεων επιβίωσης για τα δεδομένα `new_data`, δηλαδή μόνο για τα δυο άτομα που έχουν επιλεγεί. Ομοίως με πριν, η εντολή `plot` κάνει τις γραφικές παραστάσεις που θέλουμε.

Κεφάλαιο 2 2.9. Το μοντέλο αναλογικών κινδύνων του Cox στα CNS δεδομένα

```
new_data <- data.frame(PT.NUMBER=c(12,35),GROUP=c(1,0),
  SEX=c(0,1),AGE=c(45,60),STATUS=c(1,1),
  DXTOB3=c(0.91666667,0.06250000),
  DXTODEATH=c(2.0000000,1.0416667),
  B3TODEATH=c(1.08333333,0.97916667),KPS.PRE.=c(50,40),
  LESSING=c(1,1),LESDEEP=c(1,1),LESSUP=c(2,2),PROC=c(2,2),
  RAD4000=c(1,0),CHEMOPRIOR=c(0,0),RESPONSE=c(2,1),
  AGE60=c(1,1))

surv_fit <- survfit(cns2.coxint6, newdata = new_data)
plot(surv_fit, fun = "cloglog",
  xlab = "Time",
  ylab = "log(-log(S(t)))",
  main = "Log-Log Survival Curves for Two Individuals",
  col=c("black","red"))

legend("topright", legend = c("Person 1", "Person 2"),
  col = c("black","red"), lty = 1, cex = 0.8)
```

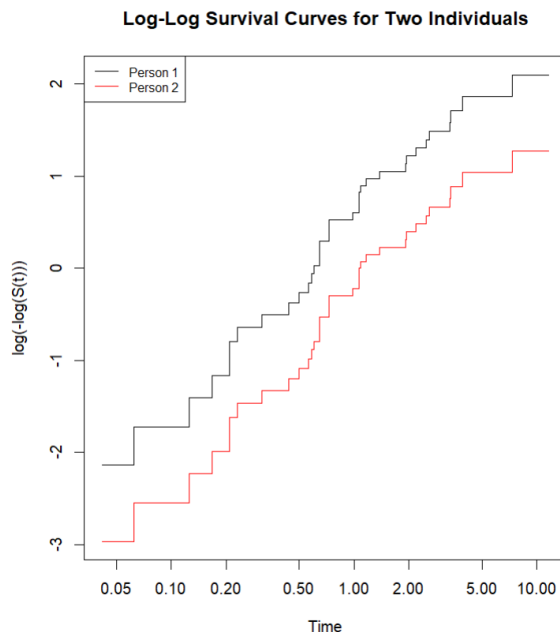
Συμπέρασμα:

Σύμφωνα με το Σχήμα 2.11, παρατηρούμε ότι οι δύο συναρτήσεις είναι προσεγγιστικά παράλληλες, οπότε το μοντέλο είναι κατάλληλο.

Στη συνέχεια, ελέγχουμε την υπόθεση της αναλογικότητας κινδύνου για τη μεταβλητή GROUP στο μοντέλο `cns2.coxint6`, χρησιμοποιώντας τον έλεγχο μέσω σύγκρισης των παρατηρούμενων έναντι των αναμενόμενων γραφικών παραστάσεων των συναρτήσεων κινδύνου. Αρχικά, προσδιορίζουμε τις παρατηρούμενες συναρτήσεις επιβίωσης. Με την εντολή `coxph` προσαρμόζουμε το στρωματοποιημένο μοντέλο του Cox με μεταβλητή στρωματοποίησης τη μεταβλητή GROUP, έχοντας προσαρμόσει τις υπόλοιπες συμμεταβλητές στο μοντέλο. Με την εντολή `survfit` υπολογίζονται οι παρατηρούμενες συναρτήσεις επιβίωσης με χρήση της εκτίμησης Kaplan-Meier για κάθε στρώμα ξεχωριστά. Οι παρατηρούμενες συναρτήσεις επιβίωσης αποθηκεύονται στο αντικείμενο `survfit.int7`. Τέλος, με την εντολή `plot` γίνεται η γραφική παράσταση των παρατηρούμενων καμπυλών που προσδιορίσαμε.

```
#observed survival plots (K-M)
cns2.coxint7 <- coxph(Surv(B3TODEATH,STATUS)~strata(GROUP)+
  KPS.PRE.+SEX+AGE60+SEX:AGE60)
survfit.int7 <- survfit(cns2.coxint7)
```

Κεφάλαιο 2 2.9. Το μοντέλο αναλογικών κινδύνων του Cox στα CNS δεδομένα

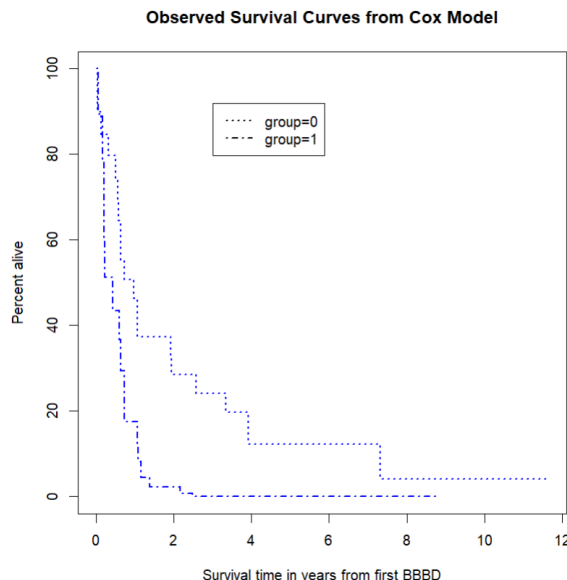


Σχήμα 2.11: Γραφικές παραστάσεις των $\log - \log$ συναρτήσεων επιβίωσης για δυο τυχαία επιλεγμένα άτομα της μελέτης, έναντι του χρόνου. Σκοπός είναι ο έλεγχος της καταλληλότητας του μοντέλου `cns2.coxint6`.

```
plot(survfit.int7,col="blue",lty=3:4,lwd=2,cex=3,  
     xlab="Survival time in years from first BBBD",  
     ylab="Percent alive",yscale=100,main = "Observed Survival  
     Curves from Cox Model")  
legend(3.0,.92,c("group=0","group=1"),lty=3:4,lwd=2)
```

Έπειτα, προσδιορίζουμε τις αναμενόμενες συναρτήσεις επιβίωσης. Με την εντολή `coxph` προσαρμόζουμε το μοντέλο του Cox που περιέχει ως μοναδική μεταβλητή τη συμεταβλητή ως προς την οποία ελέγχουμε την υπόθεση της αναλογικότητας κινδύνων, δηλαδή την `GROUP`. Στο αντικείμενο `new_data` αποθηκεύουμε τις δυνατές τιμές της μεταβλητής `GROUP`. Με την εντολή `survfit` υπολογίζονται οι εκτιμώμενες συναρτήσεις επιβίωσης για το μοντέλο `cns2.coxint7.2`. Δίνοντας ως δεύτερο όρισμα στην εντολή `survfit` το αντικείμενο `new_data`, γίνεται αντικατάσταση των δυνατών τιμών της μεταβλητής `GROUP` στις εκτιμώμενες συναρτήσεις επιβίωσης. Με αυτόν τον τρόπο παίρνουμε για κάθε στρώμα την αντίστοιχη αναμενόμενη συνάρτηση επιβίωσης. Οι συναρτήσεις αυτές αποθηκεύονται

Κεφάλαιο 2 2.9. Το μοντέλο αναλογικών κινδύνων του Cox στα CNS δεδομένα



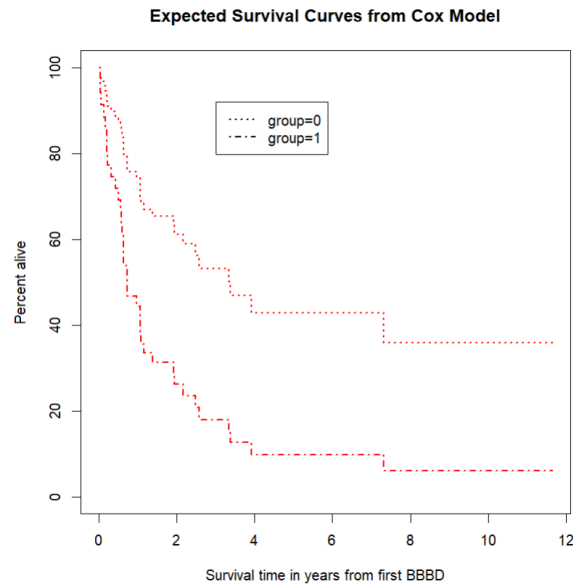
Σχήμα 2.12: Γραφικές παραστάσεις των παρατηρούμενων συναρτήσεων επιβίωσης για το μοντέλο `cns2.coxint7`, όπου μεταβλητή στρωματοποίησης είναι η `GROUP`.

στο αντικείμενο `survfit.int7.2`. Η εντολή `plot` κάνει τη γραφική παράσταση των αναμενόμενων συναρτήσεων επιβίωσης. Η εντολή `legend` προσθέτει την αντίστοιχη λεζάντα στο γράφημα.

```
#expected plots
cns2.coxint7.2 <- coxph(Surv(B3TODEATH,STATUS)~GROUP)
new_data <- data.frame(GROUP = c(0, 1))
survfit.int7.2 <- survfit(cns2.coxint7.2,newdata =new_data)
plot(survfit.int7.2,col="red",lty=3:4,lwd=2,cex=3,
     xlab="Survival time in years from first BBBD",
     ylab="Percent alive",yscale=100,
     main = "Expected Survival Curves from Cox Model")
legend(3.0,.92,c("group=0","group=1"),lty=3:4,lwd=2)
```

Τέλος, προκειμένου να βγάλουμε συμπέρασμα, απεικονίζουμε τις δύο προηγούμενες γραφικές παραστάσεις σε μία. Με την εντολή `plot`, δίνοντας ως πρώτο όρισμα το αντικείμενο `survfit.int7`, γίνεται η γραφική παράσταση των παρατηρούμενων συναρτήσεων επιβίωσης. Με τη χρήση της εντολής `lines` γίνεται στο ίδιο το γράφημα που παράχθηκε από την `plot`, η γραφική παράσταση των ανα-

Κεφάλαιο 2 2.9. Το μοντέλο αναλογικών κινδύνων του Cox στα CNS δεδομένα



Σχήμα 2.13: Γραφικές παραστάσεις των αναμενόμενων συναρτήσεων επιβίωσης για τον έλεγχο της υπόθεσης της αναλογικότητας κινδύνων της μεταβλητής GROUP.

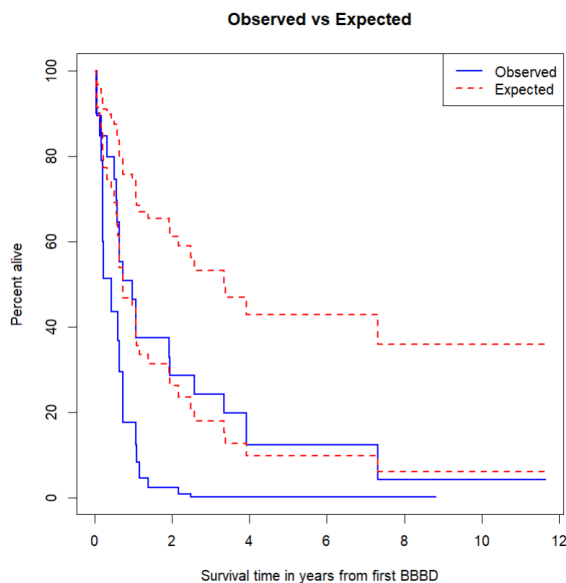
μενόμενων συναρτήσεων επιβίωσης. Αυτό γίνεται δίνοντας ως πρώτο όρισμα στην εντολή `lines` το αντικείμενο `survfit.int7.2`, όπου έχουν αποθηκευτεί οι αναμενόμενες συναρτήσεις επιβίωσης. Με την εντολή `legend` προστίθεται η αντίστοιχη λεζάντα.

```
plot(survfit.int7,col="blue",lty=1,lwd=2,cex=3,
     xlab="Survival time in years from first BBBD",
     ylab="Percent alive",yscale=100,main = "Observed
     vs Expected")
lines(survfit.int7.2,col="red",lty=2,lwd=2,cex=3)
legend("topright", legend = c("Observed", "Expected"),
     col = c("blue", "red"), lwd = 2, lty = c(1, 2))
```

Συμπέρασμα:

Σύμφωνα με το Σχήμα 2.14 παρατηρούμε ότι οι παρατηρούμενες και αναμενόμενες καμπύλες επιβίωσης δεν ταυτίζονται προσεγγιστικά. Οπότε με αυτήν την μέθοδο ελέγχου προκύπτει το συμπέρασμα ότι δεν ισχύει η υπόθεση της αναλογικότητας κινδύνου για τη μεταβλητή GROUP.

Κεφάλαιο 2 2.9. Το μοντέλο αναλογικών κινδύνων του Cox στα CNS δεδομένα



Σχήμα 2.14: Γραφική παράσταση των παρατηρούμενων έναντι των αναμενόμενων συναρτήσεων επιβίωσης με σκοπό τον έλεγχο της υπόθεσης της αναλογικότητας κινδύνων για την μεταβλητή GROUP στο μοντέλο `cns2.coxint6`.

Στη συνέχεια κάνουμε τον έλεγχο της υπόθεσης της αναλογικότητας κινδύνων για τη συμμεταβλητή GROUP στο μοντέλο `cns2.coxint6` με χρήση του ελέγχου Grambsch and Therneau. Χρησιμοποιούμε την εντολή `cox.zph` δίνοντας της ως όρισμα το μοντέλο που έχουμε προσαρμόσει. Τα αποτελέσματα που λαμβάνουμε είναι τα εξής:

```
PH.test <- cox.zph(cns2.coxint6)
PH.test
```

	chisq	df	p
KPS.PRE.	0.208	1	0.65
GROUP	1.064	1	0.30
SEX	2.104	1	0.15
AGE60	0.142	1	0.71
SEX:AGE60	1.650	1	0.20
GLOBAL	3.511	5	0.62

Συμπέρασμα:

Τόσο για την συμμεταβλητή GROUP, όσο και για τις υπόλοιπες συμμεταβλητές και

το συνολικό μοντέλο, συμπεραίνουμε ότι ισχύει η υπόθεση της αναλογικότητας κινδύνων, καθώς όλες οι p -τιμές είναι μεγαλύτερες από το στατιστικό επίπεδο σημαντικότητας $\alpha = 0.0.5$.

2.10 Αναλογία κινδύνου (HR)

Το hazard ratio (HR) ή αναλογία κινδύνου είναι ένα αριθμητικό μέτρο και ορίζεται ως ο κίνδυνος ενός ατόμου με χαρακτηριστικά $\mathbf{X}^*$ ως προς τον κίνδυνο ενός ατόμου με χαρακτηριστικά $\mathbf{X}$.

$$HR(t|\mathbf{X}^*, \mathbf{X}) = \frac{h(t|\mathbf{X}^*)}{h(t|\mathbf{X})}.$$

Το HR ποσοτικοποιεί το πόσο μεγαλύτερο κίνδυνο έχει ένα άτομο με χαρακτηριστικά $\mathbf{X}^*$, σε σχέση με ένα άτομο με χαρακτηριστικά $\mathbf{X}$ να υποστεί το υπό μελέτη γεγονός, στην περίπτωση που θεωρήσουμε ότι υπάρχουν συµμεταβλητές που επηρεάζουν τον κίνδυνο. Για αυτόν τον λόγο είναι πολύ χρήσιμο στην πράξη. Προκειμένου να δούμε τη μορφή που έχει το HR για το μοντέλο αναλογικών κινδύνων του Cox, υπενθυμίζουμε τη μορφή της συνάρτησης κινδύνων στο μοντέλο αναλογικών κινδύνων του Cox.

$$\begin{aligned} h(t|\mathbf{X}) &= h_0(t) \exp(\mathbf{X}'\boldsymbol{\beta}) \\ &= h_0(t) \exp(\beta_1 X_1) \exp(\beta_2 X_2) \cdots \exp(\beta_m X_m). \end{aligned}$$

Για δύο διαφορετικά διανύσματα συµμεταβλητών $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{im})'$, $\mathbf{X}_j = (X_{j1}, X_{j2}, \dots, X_{jm})'$ που αντιστοιχούν στα i, j άτομα της μελέτης $i, j \in \{1, \dots, n\}$, ο λόγος

$$\frac{h(t|\mathbf{X}_i)}{h(t|\mathbf{X}_j)} = \frac{\exp(\mathbf{X}_i'\boldsymbol{\beta})}{\exp(\mathbf{X}_j'\boldsymbol{\beta})} = \exp\{(\mathbf{X}_i' - \mathbf{X}_j')\boldsymbol{\beta}\}$$

είναι ο λόγος των συναρτήσεων κινδύνου (HR), που είναι σταθερός με τον χρόνο. Δηλαδή η $h_0(t)$ είναι ίδια για όλα τα άτομα της μελέτης, στην περίπτωση που υποθέτουμε ότι οι συµμεταβλητές και το διάνυσμα των συντελεστών είναι ανεξάρτητα του χρόνου. Επομένως, προκειμένου να ερµηνεύσουμε το HR πρέπει να ισχύει η υπόθεση της αναλογικότητας κινδύνων. Σε περίπτωση που δεν ισχύει η υπόθεση της αναλογικότητας κινδύνων το HR ισχύει μόνο προσεγγιστικά, δηλαδή µπορούµε να πούµε ότι αντιπροσωπεύει την αναλογία κινδύνου κατά µέσο όρο στη διάρκεια της μελέτης. Για να γίνει πιο κατανοητή η έννοια του HR θα δώσουμε την ερµηνεία του για τα μοντέλα `cns.cox` και `cns.cox1` που έχουμε προσαρµόσει.

Αρχικά για το μοντέλο `cns.cox`, σύμφωνα με τα αποτελέσματα που μας έδωσε η R και τα οποία παρατίθενται στην Ενότητα 2.4, έχουμε ότι:

- η αναλογία κινδύνου των ατόμων που ανήκουν στην ομάδα 1 και των ατόμων που ανήκουν στην ομάδα 0 της μεταβλητής `GROUP`, δηλαδή για ασθενείς με διανύσματα συμμεταβλητών $\mathbf{X}^* = X^* = 1$ και $\mathbf{X} = X = 0$ αντίστοιχα είναι:

$$HR(t|\mathbf{X}^* = 1, \mathbf{X} = 0) = \frac{\exp(1 \times 1.004)}{\exp(0)} = \exp(1.004) = 2.7292.$$

Αυτό σημαίνει ότι ασθενείς που ανήκουν στην ομάδα 1 έχουν 2.7292 φορές μεγαλύτερο κίνδυνο σε σχέση με τους ασθενείς που ανήκουν στην ομάδα 0. Δηλαδή, είναι 2.7292 φορές πιο πιθανό να έχουν μικρότερο χρόνο επιβίωσης σε σχέση με τους ασθενείς της ομάδας 0.

Για το μοντέλο `cns.cox1`, το οποίο προσαρμόσαμε στην Ενότητα 2.8, είδαμε ότι παραβιάζεται η υπόθεση της αναλογικότητας κινδύνων για την συμμεταβλητή `CHEMOPRIOR`. Ωστόσο, τα αποτελέσματα που παίρνουμε είναι τα εξής:

	coef	exp(coef)	se(coef)	z	Pr(> z)
CHEMOPRIOR	-0.1600	0.8522	0.3927	-0.407	0.68373
GROUP	1.0479	2.8518	0.3542	2.959	0.00309 **

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

	exp(coef)	exp(-coef)	lower .95	upper .95
CHEMOPRIOR	0.8522	1.1735	0.3947	1.840
GROUP	2.8518	0.3507	1.4244	5.709

Concordance= 0.625 (se = 0.049)
 Likelihood ratio test= 8.52 on 2 df, p=0.01
 Wald test = 8.99 on 2 df, p=0.01
 Score (logrank) test = 9.71 on 2 df, p=0.008

Καθώς δεν ισχύει η υπόθεση της αναλογικότητας κινδύνων για τη μεταβλητή `CHEMOPRIOR` η εκτίμηση του συντελεστή της δεν είναι σωστή και το HR δεν μπορούμε να θεωρήσουμε ότι ισχύει γενικά, καθώς μπορεί να μην είναι σταθερό με τον χρόνο. Ωστόσο, μπορούμε να πούμε το εξής:

- σταθεροποιώντας την συμμεταβλητή `GROUP`, κατά μέσο όρο κατά την διάρκεια της μελέτης η αναλογία κινδύνου των ασθενών που έχουν λάβει

χημειοθεραπεία προηγουμένως και των ασθενών που δεν έχουν λάβει είναι 0.8522. Δηλαδή, κατά μέσο όρο ένας ασθενής που έχει λάβει χημειοθεραπεία προηγουμένως έχει 85,22% του κινδύνου που έχει ένα ασθενής που δεν έχει λάβει χημειοθεραπεία προηγουμένως.

2.11 Η διαδικασία AIC για επιλογή μεταβλητών

Όσες φορές προσαρμόσαμε το μοντέλο του Cox ως τώρα θεωρήσαμε αυθαίρετα ποιες συμμεταβλητές θα συμπεριλάβουμε. Θα μπορούσαμε να διαλέξουμε οποιεσδήποτε από τις διαθέσιμες συμμεταβλητές και αντίστοιχα θα μπορούσαμε να προσαρμόσουμε πλήθος Cox μοντέλων. Στη συνέχεια παρουσιάζουμε το κριτήριο πληροφορίας του Akaike (Akaike's information criterion AIC), το οποίο μας δίνει ένα τρόπο επιλογής των συμμεταβλητών που θα συμπεριλάβουμε στο μοντέλο του Cox. Το κριτήριο αυτό προτάθηκε στην εργασία του Akaike (1974). Το κριτήριο πληροφορίας του Akaike (AIC) στο πλαίσιο του Cox PH μοντέλου παλινδρόμησης αποτελεί ένα κριτήριο με βάση το οποίο μπορούμε να κάνουμε συγκρίσεις μεταξύ πλήθους πιθανών μοντέλων. Ορίζεται ως:

$$\text{AIC} = -2 \times \log(\text{maximum likelihood}) + 2 \times b,$$

όπου b είναι το πλήθος των συντελεστών β και **maximum likelihood** είναι η μεγιστοποιημένη συνάρτηση πιθανοφάνειας του μοντέλου στο οποίο αντιστοιχεί το AIC. Η μεγιστοποιημένη συνάρτηση πιθανοφάνειας αντικαθίσταται από τη μεγιστοποιημένη συνάρτηση μερικής πιθανοφάνειας. Όσο μικρότερη είναι η τιμή του AIC, τόσο καλύτερο είναι το μοντέλο. Για την ανάπτυξη του παραδείγματος που ακολουθεί, βασιζόμαστε στο σύγγραμμα των Tableman and Kim (2003). Στην πράξη εφαρμόζουμε μια αυτοματοποιημένη διαδικασία επιλογής μοντέλου με βάση το κριτήριο AIC. Αυτό γίνεται μέσω της **stepAIC** συνάρτησης, που περιλαμβάνεται στη βιβλιοθήκη **MASS**. Η συνάρτηση **stepAIC** χρειάζεται ένα όρισμα που αναπαριστά ένα Cox μοντέλο. Χρήσιμα προαιρετικά όρια είναι τα **scope** και **direction**. Στο όρισμα **scope** καθορίζεται το εύρος των μοντέλων που εξετάζονται στη βηματική αναζήτηση. Το όρισμα **direction** μπορεί να έχει τις τιμές **"both"**, **"backward"**, **"forward"** με προεπιλογή το **"both"**. Ανάλογα με την τιμή του, καθορίζεται ο τρόπος επιλογής των συμμεταβλητών στο μοντέλο. Όταν έχει την τιμή **"backward"**, η διαδικασία ξεκινά από το μοντέλο που περιέχει όλες τις συμμεταβλητές και σε κάθε βήμα αφαιρείται μία συμμεταβλητή. Όταν έχει την τιμή **"forward"**, η διαδικασία ξεκινά από το μοντέλο χωρίς συμμεταβλητές σε κάθε βήμα προστίθεται μία συμμεταβλητή. Τέλος, όταν η τιμή είναι **"both"**, η διαδικασία μπορεί να ξεκινήσει είτε από το μοντέλο με όλες τις πιθανές συμμεταβλητές είτε από το μοντέλο χωρίς συμμεταβλητές, και σε κάθε βήμα εξετάζεται

αν πρέπει να προστεθεί ή να αφαιρεθεί μία συµμεταβλητή, µε στόχο τη µείωση της τιµής του AIC. Στη συνέχεια, µέσω ενός παραδείγματος παρουσιάζουµε τη διαδικασία AIC για την επιλογή μεταβλητών στο µοντέλο του Cox. Παρουσιάζεται η διαδικασία AIC σε συνδυασµό µε το likelihood ratio test (LRT) για την επιλογή του καλύτερου µοντέλου. Το LRT παρουσιάστηκε αναλυτικά στην Ενότητα 2.3.

Για το παράδειγµα µας προσαρµόζουµε τα CNS δεδοµένα σε ένα Cox PH µοντέλο. Υπενθυµίζουµε τον τρόπο µε τον οποίο ορίσαµε την δείκτρια μεταβλητή AGE, η οποία θα συµπεριληφθεί στο ακόλουθο παράδειγµα. Η μεταβλητή AGE60 ορίζεται ως $AGE60=1$ αν $AGE \leq 60$ και 0 αλλιώς.

```
cns2$AGE60<-as.integer(cns$AGE<=60) .
```

Θεωρούµε µοντέλο µε µέχρι δευτέρου βαθµού αλληλεπίδραση, που µπορεί εύκολα να ενσωµατωθεί στην **stepAIC**. Μοντέλα µε µεγαλύτερο βαθµό αλληλεπίδρασης µπορούν να θεωρηθούν, όµως η ερµηνεία του αποτελέσµατος της αλληλεπίδρασης δεν είναι εύκολη.

Αρχικά, προσαρµόζουµε το µοντέλο αναλογικών κινδύνων του Cox που περιέχει όλες τις μεταβλητές που µπορεί να µας είναι χρήσιµες. Το µοντέλο αυτό περιέχει τις μεταβλητές **KPS.PRE.**, **GROUP**, **SEX**, **AGE60**, **LESSING**, **LESSDEEP**, **factor(LESSUP)**, **factor(PROC)** και **CHEMOPRIOR** χωρίς τις αλληλεπιδράσεις. Οι μεταβλητές **LESSUP** και **PROC** είναι κατηγορικές, µε τρία επίπεδα η κάθε µία. Για τον λόγο αυτό τις συµπεριλαµβάνουµε στο µοντέλο µε χρήση της συνάρτησης **factor**. Το παραπάνω µοντέλο προσαρµόζεται µε τον ακόλουθο τρόπο. Πρώτα καλούµε την εντολή **Surv** που δέχεται ως πρώτο όρισμα τους χρόνους ζωής ή τους χρόνους λογοκρισίας των ασθενών και ως δεύτερο όρισμα τους δείκτες λογοκρισίας και δηµιουργεί ένα αντικείµενο που θα είναι το όρισμα της εντολής **coxph**. Η εντολή **coxph** προσαρµόζει το µοντέλο του Cox. Η εντολή **coxph** δέχεται ως πρώτο όρισμα το αντικείµενο **Surv** και δεξιά του συµβόλου $\sim$ δηλώνονται οι συµμεταβλητές που θα συµπεριληφθούν στο µοντέλο. Το µοντέλο αυτό αποθηκεύεται στο αντικείµενο **cns2.coxint**. Η εντολή **stepAIC** δέχεται ως πρώτο όρισμα το µοντέλο του Cox που προσαρµόσαµε. Στο δεύτερο όρισμα της **stepAIC**, το **scope**, δηλώνουµε ότι το µοντέλο πρέπει να είναι µε µέχρι δευτέρου βαθµού αλληλεπίδραση. Στο όρισμα **direction** η επιλογή είναι **“both”**, που είναι η προεπιλογή. Το µοντέλο του Cox που προκύπτει από την διαδικασία επιλογής μεταβλητών µέσω του κριτηρίου AIC αποθηκεύεται στο αντικείµενο **cns2.coxint1**. Με την εντολή **cns2.coxint1\$anova** παρουσιάζεται η διαδικασία που ακολούθησε η **stepAIC** καθώς και οι αντίστοιχες τιµές του AIC για κάθε βήµα. Η πρώτη τιµή του AIC βασίζεται στο αρχικό µοντέλο που περιέχει όλες τις μεταβλητές και όχι τις αλληλεπιδράσεις τους. Με **“+”** συµβολίζουµε

ότι η αντίστοιχη συμμεταβλητή προστέθηκε σε αυτό το βήμα, ενώ με “-” ότι αφαιρέθηκε σε αυτό το βήμα. Το τελικό μοντέλο που προέκυψε με χρήση της εντολής `stepAIC` περιέχει τις ακόλουθες συμμεταβλητές `KPS.PRE.`, `GROUP`, `SEX`, `AGE60`, `LESSING`, `CHEMOPRIOR`, `SEX:AGE60`, `AGE60:LESSING` και `GROUP:AGE60`.

Βήμα 1: stepAIC συνάρτηση για την επιλογή του καλύτερου μοντέλου σύμφωνα με το AIC στατιστικό.

```
library(MASS)
library(survival)
attach(cns2)
cns2$AGE60 <- as.integer(cns2$AGE<=60)
cns2.coxint<- coxph(Surv(B3TODEATH,STATUS)~ KPS.PRE.+GROUP+SEX
                    +AGE60+LESSING+LESDEEP+factor(LESSUP)
                    +factor(PROC)+CHEMOPRIOR)
#Αυτό είναι το αρχικό μοντέλο
cns2.coxint1 <- stepAIC(cns2.coxint,~.^2)
cns2.coxint1$anova
```

	Step	Df	Deviance	Resid.	Df	Resid.	Dev	AIC
1					25	224.0864	246.0864	
2	+ SEX:AGE60	1	8.75265472		24	215.3337	239.3337	
3	- factor(PROC)	2	1.41352590		26	216.7472	236.7472	
4	- LESDEEP	1	0.02918476		27	216.7764	234.7764	
5	- factor(LESSUP)	2	2.36994524		29	219.1464	233.1464	
6	+ AGE60:LESSING	1	2.30031305		28	216.8460	232.8460	
7	+ GROUP:AGE60	1	2.19494887		27	214.6511	232.6511	

Στο Βήμα 2 χρησιμοποιούμε το LRT για να δούμε αν είναι εφικτό να μειώσουμε παραπάνω το πλήθος των συμμεταβλητών που θα συμπεριληφθούν στο μοντέλο του Cox.

Βήμα 2: LRT για περαιτέρω μείωση

Παρακάτω παρουσιάζονται οι p -τιμές για τις μεταβλητές που επιλέχθηκαν από την `stepAIC`. Η μεταβλητή `AGE60` έχει μεγάλη p -τιμή, 0.56, ενώ οι όροι αλληλεπίδρασής της με τις `SEX` και `LESSING` έχουν μικρές p -τιμές, 0.0019 και 0.0590 αντίστοιχα.

```
cns2.coxint1
```

	coef	exp(coef)	se(coef)	z	p
KPS.PRE.	-0.04714	0.95395	0.01402	-3.362	0.000774
GROUP	2.01391	7.49252	0.70671	2.850	0.004376
SEX	-3.30889	0.03656	0.88581	-3.735	0.000187
AGE60	-0.40369	0.66785	0.68631	-0.588	0.556400
LESSING	1.64703	5.19154	0.67049	2.456	0.014032
CHEMOPRIOR	1.01016	2.74603	0.53856	1.876	0.060704
SEX:AGE60	2.86675	17.57986	0.92099	3.113	0.001854
AGE60:LESSING	-1.58596	0.20475	0.83848	-1.891	0.058562
GROUP:AGE60	-1.25752	0.28436	0.83829	-1.500	0.133588

Μία βασική αρχή στη στατιστική μοντελοποίηση είναι ότι ένας όρος αλληλεπίδρασης πρέπει να συμπεριληφθεί σε ένα μοντέλο μόνο αν οι αντίστοιχες μεταβλητές είναι στο μοντέλο. Βλέπουμε αν μπορούμε να αφαιρέσουμε τη μεταβλητή AGE60 και τους όρους αλληλεπίδρασής της με τις άλλες μεταβλητές. Χρησιμοποιούμε το LRT. Το LRT τεστ δίνει ισχυρή ένδειξη κατά του μειωμένου μοντέλου και έτσι διατηρούμε το μοντέλο που επιλέχθηκε από την stepAIC, $p\text{-τιμή} \approx 0.00938 < \alpha$.

```
cns2.coxint2<- coxph(Surv(B3TODEATH,STATUS)~ KPS.PRE.+GROUP
+SEX+LESSING+CHEMOPRIOR)
-2*cns2.coxint2$loglik[2] + 2*cns2.coxint1$loglik[2]
[1] 13.42442
1 - pchisq(13.42442,4)
[1] 0.009377846
```

Τώρα ξεκινάμε τη διαδικασία αφαίρεσης μιας μεταβλητής τη φορά (ενώ πριν αφαιρέσαμε την AGE60 και τους όρους αλληλεπίδρασής της με άλλες μεταβλητές μαζί). Αυτό μπορεί να γίνει, είτε αν βασιστούμε στις p -τιμές, είτε στο LRT. Ασυμπτωτικά παίρνουμε το ίδιο αποτέλεσμα. Εφόσον η μεταβλητή GROUP:AGE60 έχει σχετικά μεγάλη p -τιμή $=0.133588 > 0.05$, την αφαιρούμε. Το ακόλουθο LRT τεστ δεν δίνει ένδειξη κατά του μειωμένου μοντέλου ($p\text{-τιμή} \approx 0.138 > \alpha$), έτσι υιοθετούμε το μειωμένο μοντέλο.

```
cns2.coxint3 <- coxph(Surv(B3TODEATH,STATUS)~KPS.PRE.+GROUP
+SEX+AGE60+LESSING+CHEMOPRIOR+SEX:AGE60
+AGE60:LESSING)
-2*cns2.coxint3$loglik[2] + 2*cns2.coxint1$loglik[2]
[1] 2.194949
```

```
1 - pchisq(2.194949,1)
[1] 0.1384638
cns2.coxint3
```

	coef	exp(coef)	se(coef)	z	p
KPS.PRE.	-0.04363	0.95731	0.01342	-3.251	0.00115
GROUP	1.12765	3.08839	0.43512	2.592	0.00955
SEX	-2.75199	0.06380	0.76128	-3.615	0.00030
AGE60	-0.92089	0.39816	0.59906	-1.537	0.12424
LESSING	1.36094	3.89984	0.63329	2.149	0.03164
CHEMOPRIOR	0.86697	2.37969	0.52595	1.648	0.09927
SEX:AGE60	2.45623	11.66073	0.87882	2.795	0.00519
AGE60:LESSING	-1.23102	0.29199	0.80588	-1.528	0.12663

Από αυτό το σημείο και έπειτα χρησιμοποιούμε τη μέθοδο της p -τιμής για να αφαιρούμε έναν όρο τη φορά. Καθώς η AGE60:LESSING έχει μεγάλη p -τιμή=0.12663, την αφαιρούμε.

```
cns2.coxint4 <- coxph(Surv(B3TODEATH,STATUS)~KPS.PRE.+GROUP
+SEX+AGE60+LESSING+CHEMOPRIOR+SEX:AGE60)
```

```
cns2.coxint4
```

	coef	exp(coef)	se(coef)	z	p
KPS.PRE.	-0.03711	0.96357	0.01236	-3.002	0.002678
GROUP	1.15239	3.16575	0.43315	2.661	0.007802
SEX	-2.59653	0.07453	0.76478	-3.395	0.000686
AGE60	-1.37994	0.25159	0.51294	-2.690	0.007140
LESSING	0.57095	1.76995	0.40370	1.414	0.157274
CHEMOPRIOR	0.85550	2.35255	0.51786	1.652	0.098537
SEX:AGE60	2.34798	10.46437	0.87653	2.679	0.007390

Αφαιρούμε την μεταβλητή LESSING, p -τιμή=0.157274. Πρώτα αφαιρέσαμε τους όρους αλληλεπίδρασης στους οποίους ήταν και η LESSING σύμφωνα με τη βασική αρχή που αναφέρθηκε προηγουμένως.

```
cns2.coxint5 <- coxph(Surv(B3TODEATH,STATUS)~KPS.PRE.+GROUP
+SEX+AGE60+CHEMOPRIOR+SEX:AGE60)
```

```
cns2.coxint5
```

	coef	exp(coef)	se(coef)	z	p
KPS.PRE.	-0.04020	0.96060	0.01214	-3.311	0.000929
GROUP	0.96951	2.63665	0.40910	2.370	0.017795
SEX	-2.47426	0.08423	0.76764	-3.223	0.001268

AGE60	-1.11088	0.32927	0.47291	-2.349	0.018822
CHEMOPRIOR	0.79534	2.21519	0.51054	1.558	0.119271
SEX:AGE60	2.18445	8.88578	0.87127	2.507	0.012169

Αφαιρούμε τη μεταβλητή CHEMOPRIOR, p -τιμή=0.119271. Καθώς όλες οι τιμές στο προσαρμοσμένο μειωμένο μοντέλο είναι επαρκώς μικρές, σε επίπεδο σημαντικότητας $\alpha=0.05$, σταματάμε εδώ και διατηρούμε το μοντέλο με τις πέντε ακόλουθες συμμεταβλητές: KPS.PRE., GROUP, SEX, AGE60 και SEX:AGE60. Το μοντέλο αυτό το αποθηκεύουμε στο αντικείμενο `cns2.coxint6`.

```
cns2.coxint6 <- coxph(Surv(B3TODEATH,STATUS)~KPS.PRE.+GROUP
+SEX+AGE60+SEX:AGE60)

cns2.coxint6
      coef exp(coef) se(coef)      z      p
KPS.PRE. -0.03067  0.96980  0.01024 -2.994 0.00276
GROUP      1.15915  3.18723  0.37941  3.055 0.00225
SEX        -2.11129  0.12108  0.70112 -3.011 0.00260
AGE60      -1.05384  0.34860  0.45722 -2.305 0.02117
SEX:AGE60  2.14004  8.49977  0.85404  2.506 0.01222
```

Ωστόσο, είναι σημαντικό να συγκρίνουμε το μοντέλο που διατηρήσαμε με το μοντέλο που επιλέχθηκε από την `stepAIC` στο Βήμα 1. Δηλαδή συγκρίνουμε τα μοντέλα `cns2.coxint1` και `cns2.coxint6`. Η p -τιμή που βασίζεται στο LRT είναι $0.06512354 > \alpha = 0.05$, οπότε επιλέγουμε το μειωμένο μοντέλο.

```
-2*cns2.coxint6$loglik[2] + 2*cns2.coxint1$loglik[2]
[1] 8.843838
1 - pchisq(8.843838,4)
[1] 0.06512354
```

Στο ακόλουθο output, που βασίζεται στο μοντέλο που περιλαμβάνει τις μεταβλητές KPS.PRE., GROUP, SEX, AGE60 και SEX:AGE60, παρατηρούμε ότι τα τεστ LRT, Wald και Efficient score test υποδεικνύουν ότι υπάρχει μια στατιστικά σημαντική σχέση μεταξύ αυτής της ομάδας συμμεταβλητών και του χρόνου ζωής.

```
summary(cns2.coxint6)
Concordance= 0.765 (se = 0.037 )
Likelihood ratio test= 27.62 on 5 df, p=4e-05
Wald test               = 24.63 on 5 df, p=2e-04
Score (logrank) test = 28.46 on 5 df, p=3e-05
```

Παρατηρήσεις:

1. Η διαδικασία επιλογής μεταβλητών μπορεί να εξαρτάται από τον σκοπό της έρευνας. Σε κάποιες μελέτες ίσως υπάρχουν κάποιες μεταβλητές ενδιαφέροντος, δηλαδή μεταβλητές που θέλουμε να συμπεριληφθούν στο μοντέλο. Σε αυτήν την περίπτωση μπορούμε και πάλι να εφαρμόσουμε τα βήματα 1 και 2. Στο Βήμα 1 επιλέγουμε την καλύτερη ομάδα μεταβλητών με βάση τη μικρότερη τιμή που παίρνει το AIC στατιστικό. Αν στο σύνολο των συμμεταβλητών που επιλέχθηκαν περιλαμβάνονται όλες οι μεταβλητές ενδιαφέροντος, στο Βήμα 2 εξετάζουμε αν από το μοντέλο μπορούν να αφαιρεθούν περαιτέρω μεταβλητές. Σε αντίθετη περίπτωση συμπεριλαμβάνουμε στο επιλεγμένο μοντέλο τις μεταβλητές ενδιαφέροντος και συνεχίζουμε με το Βήμα 2.
2. Είναι σημαντικό να συμπεριληφθούν όροι αλληλεπίδρασης στη διαδικασία επιλογής μοντέλου, εκτός αν οι ερευνητές επιβάλουν να μη συμπεριληφθούν όροι αλληλεπίδρασης. Είναι πιθανό να οδηγηθούμε σε διαφορετικό μοντέλο όταν λαμβάνονται υπόψη μόνο οι κύριες επιδράσεις και όχι οι αλληλεπιδράσεις.

Επαναλαμβάνουμε τη διαδικασία επιλογής μεταβλητών για τα CNS δεδομένα, για την περίπτωση που δεν θέλουμε να λάβουμε υπόψη όρους αλληλεπίδρασης. Δίνοντας ως δεύτερο όρισμα της εντολής `stepAIC` το `~`, δηλώνουμε ότι δεν πρέπει να συμπεριληφθούν όροι αλληλεπίδρασης στο μοντέλο. Το μοντέλο που προκύπτει με την διαδικασία AIC αποθηκεύεται στο αντικείμενο `cns2.cox1`. Με την εντολή `cns2.coxint6$anova` παρουσιάζεται η διαδικασία που ακολούθησε η `stepAIC` καθώς και οι αντίστοιχες τιμές του AIC για κάθε βήμα. Η αρχική τιμή του AIC αναφέρεται στο αρχικό μοντέλο που περιλαμβάνει όλες τις μεταβλητές. Το τελικό μοντέλο επιλέγεται εφαρμόζοντας “προς τα πίσω” διαδικασία αφαίρεσης μεταβλητών. Το εύρος των δυνατών μοντέλων που μπορούν να προκύψουν είναι από το αρχικό μοντέλο που περιλαμβάνει όλες τις μεταβλητές μέχρι το μικρότερο δυνατό μειωμένο μοντέλο που δεν περιλαμβάνει συμμεταβλητές (intercept only μοντέλο). Το μοντέλο που προκύπτει περιλαμβάνει τις συμμεταβλητές `KPS.PRE.`, `GROUP`, `SEX`, και `CHEMOPRIOR`.

Βήμα 1: stepAIC συνάρτηση για την επιλογή του καλύτερου μοντέλου σύμφωνα με το AIC στατιστικό

```
cns2.cox <-coxph(Surv(B3TODEATH,STATUS)~KPS.PRE.+GROUP+SEX
```

```

+AGE60+LESSING+LESDEEP+factor(LESSUP)
+factor(PROC)+CHEMOPRIOR)
cns2.cox1 <- stepAIC(cns2.cox,~.)
cns2.cox1$anova

```

	Step	Df	Deviance	Resid.	Df	Resid. Dev	AIC
1					25	224.0864	246.0864
2	- factor(PROC)	2	0.190265427		27	224.2766	242.2766
3	- LESDEEP	1	0.003879781		28	224.2805	240.2805
4	- AGE60	1	0.452246097		29	224.7327	238.7327
5	- factor(LESSUP)	2	3.342766505		31	228.0755	238.0755
6	- LESSING	1	0.479306267		32	228.5548	236.5548

Βήμα 2: LRT για περαιτέρω μείωση

Στα παρακάτω αποτελέσματα παρουσιάζονται οι p -τιμές που αντιστοιχούν στις μεταβλητές που επιλέχθηκαν από την `stepAIC`. Οι p -τιμές που αντιστοιχούν στις μεταβλητές `GROUP` και `CHEMOPRIOR` είναι πολύ κοντινές. Συμπεραίνουμε ότι η επίδραση τους στον χρόνο ζωής, προσαρμοσμένη για τις υπόλοιπες μεταβλητές, είναι περίπου ίδια.

```

cns2.cox1

```

	coef	exp(coef)	se(coef)	z	p
KPS.PRE.	-0.04324	0.95768	0.01165	-3.711	0.000206
GROUP	0.55641	1.74439	0.38815	1.433	0.151721
SEX	-1.07213	0.34228	0.45509	-2.356	0.018480
CHEMOPRIOR	0.72588	2.06655	0.47720	1.521	0.128230

Αρχικά αφαιρούμε τη μεταβλητή `GROUP`. Εφόσον όλες οι p -τιμές στο μειωμένο μοντέλο είναι επαρκώς μικρές, σε επίπεδο σημαντικότητας 0.05, σταματάμε και διατηρούμε τις μεταβλητές `KPS.PRE.`, `SEX`, και `CHEMOPRIOR`.

```

cns2.cox2 <- coxph(Surv(B3TODEATH,STATUS)~KPS.PRE.+SEX
+CHEMOPRIOR)
cns2.cox2

```

	coef	exp(coef)	se(coef)	z	p
KPS.PRE.	-0.04913	0.95206	0.01102	-4.459	8.22e-06
SEX	-1.20025	0.30112	0.44556	-2.694	0.00706
CHEMOPRIOR	1.00914	2.74323	0.43966	2.295	0.02172

Παρουσιάζουμε την περίπτωση που αφαιρούμε πρώτα τη μεταβλητή `CHEMOPRIOR`, αντί για την `GROUP`. Όλες οι p -τιμές στο μειωμένο μοντέλο είναι επαρκώς μικρές,

σε επίπεδο σημαντικότητας 0.05, οπότε σταματάμε και διατηρούμε τις μεταβλητές KPS.PRE., SEX και GROUP.

```
cns2.cox3 <- coxph(Surv(B3TODEATH,STATUS)~KPS.PRE.+GROUP+SEX)
cns2.cox3
```

	coef	exp(coef)	se(coef)	z	p
KPS.PRE.	-0.03465	0.96594	0.01005	-3.448	0.000564
GROUP	0.77848	2.17815	0.35424	2.198	0.027976
SEX	-0.79679	0.45078	0.40993	-1.944	0.051932

Συνεπώς, ανάλογα με τη σειρά αφαίρεσης μεταβλητών διατηρούμε είτε τις SEX, KPS.PRE. και CHEMOPRIOR, είτε τις KPS.PRE., GROUP και SEX. Έτσι έχουμε δύο διαφορετικά μοντέλα, το ένα περιέχει τη μεταβλητή CHEMOPRIOR, ενώ το άλλο τη μεταβλητή GROUP, αντίστοιχα. Παρατηρούμε ότι κανένα από τα δύο μοντέλα δεν περιλαμβάνει τη μεταβλητή AGE60 που είναι ένας σημαντικός προ-γνωστικός παράγοντας σε αυτήν την μελέτη, όπως φαίνεται από το αποτέλεσμα της αλληλεπίδρασης της AGE60 με την SEX στο μοντέλο cns2.coxint6 (p -τιμή=0.01222 < α =0.05). Επιπλέον, η σημαντικότητα της επίδρασης της μεταβλητής GROUP στο μοντέλο με τις αλληλεπιδράσεις, cns2.coxint6, είναι μεγαλύτερη (p -τιμή=0.00225 έναντι 0.027976). Η μεταβλητή GROUP είναι η μεταβλητή κύριου ενδιαφέροντος της μελέτης. Οπότε επιλέγουμε το μοντέλο cns2.coxint6 που περιέχει αλληλεπιδράσεις.

Παρατηρήσεις:

1. Οι μεταβλητές KPS.PRE., GROUP, SEX, AGE60 και SEX:AGE60 έχουν σημαντική επίδραση στο χρόνο ζωής. Όλες οι p -τιμές για τον έλεγχο της υπόθεσης ότι ο συντελεστής της αντίστοιχης μεταβλητής είναι μηδέν, υποδεικνύουν ότι απορρίπτεται αυτή η υπόθεση σε επίπεδο σημαντικότητας $\alpha = 0.05$ (p -τιμές < 0.05). Υπάρχει σημαντική διαφορά μεταξύ των καμπυλών επιβίωσης των δύο ομάδων (GROUP 1 ασθενείς που έλαβαν κρανιακή ακτινοβολία πριν να παραπεμφθούν σε BBBD χημειοθεραπεία, GROUP 0 αρχική θεραπεία είναι η BBBD χημειοθεραπεία).
2. Η εκτίμηση της παραμέτρου του μοντέλου για τη μεταβλητή KPS.PRE είναι -0.3067 με p -τιμή=0.00276. Έτσι, με σταθερές τις υπόλοιπες συμμεταβλητές, ασθενείς με μεγαλύτερες τιμές για τη μεταβλητή KPS.PRE διατρέχουν μικρότερο κίνδυνο, άρα έχουν μεγαλύτερο αναμενόμενο χρόνο ζωής από αυτούς που έχουν χαμηλές τιμές στη μεταβλητή KPS.PRE..
3. Η εκτίμηση της παραμέτρου του μοντέλου για τη μεταβλητή GROUP είναι

1.15915 με p -τιμή=0.00225. Έτσι, με τις υπόλοιπες συμμεταβλητές σταθερές, ασθενείς της ομάδας 1, δηλαδή που έχουν λάβει κρανιακή ακτινοβολία πριν την BBBD χημειοθεραπεία, διατρέχουν μεγαλύτερο κίνδυνο. Οπότε, ασθενείς της ομάδας 1 έχουν μικρότερο αναμενόμενο χρόνο ζωής σε σχέση με τους ασθενείς της ομάδας 0.

4. Σταθεροποιώντας τις υπόλοιπες συμμεταβλητές, ο λόγος κινδύνου μεταξύ των ομάδων 1 και 0 είναι:

$$\frac{\exp(1.15915)}{\exp(0)} = 3.18723.$$

Αυτό σημαίνει ότι, με σταθερές τις υπόλοιπες συμμεταβλητές, ασθενείς που έχουν δεχτεί κρανιακή ακτινοβολία πριν την πρώτη BBBD χημειοθεραπεία είναι 3.18723 φορές πιο πιθανό να έχουν μικρότερο χρόνο επιβίωσης από αυτούς που δεν έχουν δεχτεί.

5. Σταθεροποιώντας τις υπόλοιπες συμμεταβλητές, αν ένας ασθενής της ομάδας 1 έχει κατά 10 μονάδες μεγαλύτερη τιμή στη μεταβλητή KPS.PRE από έναν ασθενή της ομάδας 0, ο λόγος των συναρτήσεων κινδύνου είναι:

$$\frac{\exp(1.15915)\exp(-0.03067 \times (k + 10))}{\exp(0)\exp(-0.03067 \times k)} = 2.345,$$

όπου k ένας τυχαίος αριθμός. Οπότε, με τις υπόλοιπες συμμεταβλητές σταθερές, ένας ασθενής της ομάδας 1 με 10 μονάδες μεγαλύτερη τιμή στη μεταβλητή KPS.PRE από έναν ασθενή της ομάδας 0, είναι 2.345 φορές πιο πιθανό να έχει μικρότερο χρόνο ζωής. Συνοψίζοντας, είναι πιο σημαντικό αν ένας ασθενής έχει λάβει κρανιακή ακτινοβολία πριν την πρώτη BBBD χημειοθεραπεία, από το πόσο μεγάλη τιμή έχει στη μεταβλητή KPS.PRE..

6. Υπάρχει σημαντική αλληλεπίδραση των μεταβλητών AGE60 και SEX. Η εκτίμηση της παραμέτρου του μοντέλου για την μεταβλητή SEX:AGE60 είναι 2.14004 με p -τιμή=0.01222. Σταθεροποιώντας τις υπόλοιπες συμμεταβλητές, ένας άνδρας ασθενής που είναι νεότερος από 60 έτη, έχει 34,86% του κινδύνου που έχει ένας άνδρας μεγαλύτερος των 60 ετών να υποκύψει στην ασθένεια.

$$\frac{\exp(-2.11129 \times 0 - 1.05384 \times 1 + 2.14 \times 0)}{\exp(-2.11129 \times 0 - 1.05384 \times 0 + 2.14 \times 0)} = \exp(-1.05384) = 0.3486.$$

Μια γυναίκα ασθενής, που είναι νεότερη από 60 έτη, έχει 2.963 φορές τον

κίνδυνο που έχει μια γυναίκα μεγαλύτερη των 60 ετών να υποκύψει στην ασθένεια.

$$\frac{\exp(-2.11129 \times 1 - 1.05384 \times 1 + 2.14 \times 1)}{\exp(-2.11129 \times 1 - 1.05384 \times 0 + 2.14 \times 0)} = \exp(1.0862) = 2.963.$$

2.12 Τύποι υπολοίπων και έλεγχοι

Τύποι υπολοίπων για τα Cox μοντέλα

Αφού έχουμε προσαρμόσει το μοντέλο του Cox στα διαθέσιμα δεδομένα μας, προκειμένου να ερμηνεύσουμε το μοντέλο και να βγάλουμε συμπεράσματα, θα πρέπει να ελέγξουμε ότι το μοντέλο που προσαρμόσαμε είναι κατάλληλο. Ένας άλλος έλεγχος που γίνεται, αφορά την εύρεση της καλύτερης συναρτησιακής μορφής μιας συμμεταβλητής, για την περιγραφή της επίδρασης αυτής της συμμεταβλητής στον χρόνο επιβίωσης, όσο οι υπόλοιπες συμμεταβλητές είναι προσαρμοσμένες. Ακόμη, εξετάζεται η συνολική ορθότητα του μοντέλου στο να προβλέπει τον χρόνο επιβίωσης των ατόμων της έρευνας. Σε αυτήν την περίπτωση ενδιαφερόμαστε για άτομα που βίωσαν το υπό μελέτη γεγονός είτε πολύ αργά, είτε πολύ νωρίς σε σχέση με την πρόβλεψη του μοντέλου, οπότε βρίσκουμε και ποια άτομα αποτελούν πιθανές ακραίες τιμές. Επιπλέον, εξετάζεται και η επιρροή που έχει κάθε άτομο στην προσαρμογή του μοντέλου, από όπου και πάλι βρίσκουμε πιθανές ακραίες τιμές. Όλοι οι έλεγχοι που αναφέρθηκαν μπορούν να γίνουν γραφικά με χρήση υπολοίπων. Έχουν προταθεί διάφορα είδη υπολοίπων και το κάθε ένα χρησιμοποιείται για συγκεκριμένο είδος ελέγχου. Παρακάτω ορίζουμε κάποιους τύπους υπολοίπων για Cox PH μοντέλα. Για την ανάλυση αυτή βασιζόμαστε στο σύγγραμμα των Tableman and Kim (2003).

Cox-Snell υπόλοιπα για την αξιολόγηση της συνολικής προσαρμογής ενός PH μοντέλου:

Ισχύει ότι:

$$H(t) = -\log\{S(t)\} = -\log\{1 - F(t)\},$$

όπου $F(t)$ είναι η πραγματική αθροιστική συνάρτηση κατανομής (α.σ.κ) του χρόνου επιβίωσης T τη χρονική στιγμή t και $H(t)$ συμβολίζει την πραγματική αθροιστική συνάρτηση κινδύνου τη χρονική στιγμή t (cumulative hazard rate). Ανεξαρτήτως της μορφής της $F(t)$, η τυχαία μεταβλητή $F(T)$ είναι ομοιόμορφα κατανομημένη στο διάστημα $(0,1)$. Συνεπώς, η τυχαία μεταβλητή $H(T)$ είναι εκθετικά κατανομημένη με ρυθμό κινδύνου, παράμετρο, $\lambda = 1$. Η απόδειξη γίνεται

βρίσκοντας την α.σ.κ. της $H(T)$. Δηλαδή,

$$\begin{aligned} P(H(T) \leq t) &= P(T \leq H^{-1}(t)) \\ &= F(H^{-1}(t)) \\ &= 1 - \exp\{-H(H^{-1}(t))\} \\ &= 1 - \exp(-t), \end{aligned}$$

που αντιστοιχεί στην αθροιστική συνάρτηση κατανομής της τυπικής εκθετικής κατανομής. Έστω $\mathbf{X}_i$ το διάνυσμα των συμμεταβλητών του $i = 1, \dots, n$ από-μου. Τότε για δοθέν $\mathbf{X}_i = (X_{i1}, \dots, X_{im})'$, $H(t|\mathbf{X}_i)$ συμβολίζει την πραγματική αθροιστική συνάρτηση κινδύνου, για ένα άτομο με διάνυσμα συμμεταβλητών $\mathbf{X}_i$. Ισχύει ότι:

$$H(T_i|\mathbf{X}_i) \sim \text{Exp}(\lambda = 1).$$

Οπότε, αν το μοντέλο Cox PH είναι σωστό για δοθέν $\mathbf{X}_i = (X_{i1}, \dots, X_{im})$ θα ισχύει ότι:

$$H(T_i|\mathbf{X}_i) = H_0(T_i) \times \exp\left(\sum_{j=1}^m \beta_j X_{ij}\right) \sim \text{Exp}(\lambda = 1).$$

Τα Cox-Snell υπόλοιπα, τα οποία προτάθηκαν στην εργασία των Cox and Snell (1968), ορίζονται ως:

$$r_{C_i} = \hat{H}_0(Y_i) \times \exp\left(\sum_{j=1}^m \hat{\beta}_j X_{ij}\right), i = 1, \dots, n,$$

όπου $Y_i = \min\{T_i, C_i\}$. Τα $\hat{\beta}_j$ είναι οι εκτιμητές μέγιστης μερικής πιθανοφάνειας των συντελεστών β_j , $j = 1, \dots, m$, τους οποίους αποκοτούμε από μεγιστοποίηση της συνάρτησης μερικής πιθανοφάνειας του Cox. Με $\hat{H}_0(t)$ συμβολίζουμε τον εμπειρικό εκτιμητή της βασικής αθροιστικής συνάρτησης κινδύνου τη χρονική στιγμή t . Αυτός είναι συνήθως είτε ο Breslow, είτε ο Nelson-Aalen εκτιμητής, τους οποίους παρουσιάζουμε παρακάτω. Τα Cox-Snell υπόλοιπα είναι πάντα μη αρνητικά. Αν το τελικό μοντέλο είναι σωστό, τότε τα $\hat{\beta}_j$ είναι κοντά στις πραγματικές τιμές των β_j και τα r_{C_i} πρέπει να μοιάζουν με ένα λογοκριμένο δείγμα από την εκθετική κατανομή με παράμετρο 1. Έστω ότι $H_E(t)$ είναι η αθροιστική συνάρτηση κινδύνου από την εκθετική κατανομή με παράμετρο 1, τότε $H_E(t) = t$. Έστω $\hat{H}_{r_C}(t)$ ένας συνεπής εκτιμητής της αθροιστικής συνάρτησης κινδύνου των r_{C_i} . Τότε ο $\hat{H}_{r_C}(t)$ πρέπει να είναι κοντά στην $H_E(t) = t$. Συνεπώς, για κάθε μη λογοκριμένο r_{C_i} , $\hat{H}_{r_C}(r_{C_i}) \approx r_{C_i}$. Για να γίνει ο έλεγχος ότι τα r_{C_i} αποτελούν ένα λογοκριμένο δείγμα από την εκθετική με παράμετρο 1, κάνουμε τη γραφική

παράσταση των $\hat{H}_{r_C}(r_{C_i})$ έναντι των r_{C_i} και θα πρέπει να αναπαριστούν μια ευθεία με κλίση 45^0 που διέρχεται από την αρχή των αξόνων.

Παρακάτω δίνεται ο ορισμός των εμπειρικών εκτιμητών Breslow και Nelson-Aalen, με τους οποίους μπορούμε να υπολογίσουμε τον εμπειρικό εκτιμητή της βασικής αθροιστικής συνάρτησης κινδύνου.

Ορισμός των εμπειρικών εκτιμητών Breslow και Nelson-Aalen

Η τυχαία μεταβλητή Y έχει πεδίο ορισμού $[0, \infty)$, στην πράξη παίρνει τιμές στο $[0, Y_{(n)})$, όπου $Y_{(n)}$ είναι ο μεγαλύτερος χρόνος ζωής. Μπορούμε να χωρίσουμε το διάστημα $[0, Y_{(n)})$ σε p υποδιαστήματα $[0, t_1), [t_1, t_2), \dots, [t_{p-1}, t_p)$, όπου $t_p = Y_{(n)}$. Το πλήθος των υπό μελέτη γεγονότων μέχρι τη χρονική στιγμή t_k είναι:

$$d_k = \sum_{i=1}^n I_{\{t_{k-1} < Y_i \leq t_k, \delta_k=1\}}, \quad k = 1, 2, \dots.$$

Ο συνολικός αριθμός των ατόμων που επιβιώνουν μετά τη χρονική στιγμή t_k , δηλαδή δεν τους έχει συμβεί το υπό μελέτη γεγονός, ούτε είναι λογοκριμένοι, είναι:

$$n_k = \sum_{i=1}^n I_{\{Y_i \geq t_{k-1}\}}.$$

Ο εκτιμητής Nelson-Aalen (Nelson, 1972; Aalen, 1978) της αθροιστικής συνάρτησης κινδύνου ορίζεται ως:

$$H_{NA}(t) = \sum_{t_j \leq t} \frac{d_j}{n_j}.$$

Ο εκτιμητής Breslow (Breslow, 1972) της αθροιστικής συνάρτησης κινδύνου ορίζεται ως:

$$H_B(t) = \sum_{t_j \leq t} \frac{d_j}{\sum_{l \in R(t_j)} \exp(\mathbf{X}'_l \hat{\beta})},$$

με $R(t_j)$ το σύνολο των δεικτών των ατόμων που είναι ζωντανοί και μη λογοκριμένοι ακριβώς πριν τη χρονική στιγμή t_j .

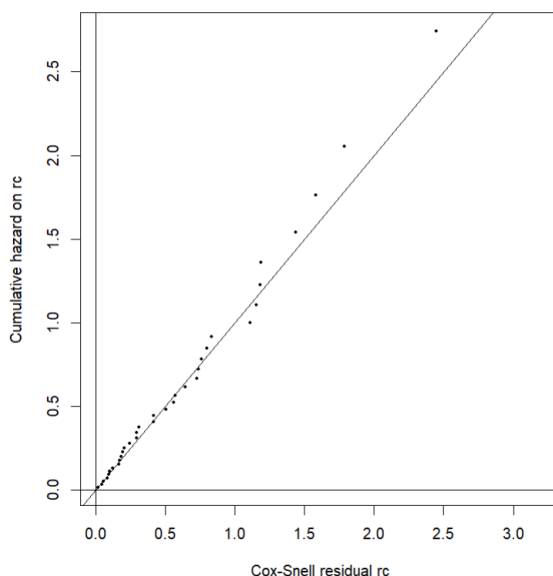
Παρατηρήσεις:

- Τα Cox-Snell υπόλοιπα είναι χρήσιμα για τον έλεγχο της συνολικής προσαρμογής ενός μοντέλου. Μια αδυναμία τους είναι ότι δεν υποδεικνύουν το είδος της απόκλισης από το εξεταζόμενο μοντέλο, όταν η γραφική παράσταση της εκτιμώμενης αθροιστικής συνάρτησης κινδύνου δεν είναι ευθεία.

- Ιδανικά, η γραφική παράσταση των $\hat{H}_{r_{C_i}}(r_{C_i})$ έναντι των r_{C_i} πρέπει να περιλαμβάνει ένα διάστημα εμπιστοσύνης, έτσι ώστε να μπορούμε να αναφερθούμε σε επίπεδο σημαντικότητας. Δυστυχώς, τα r_{C_i} δεν είναι ακριβώς ένα λογοκριμένο δείγμα από μια κατανομή. Οπότε, αυτή η γραφική παράσταση χρησιμοποιείται ως πρόχειρος έλεγχος.
- Η εγγύτητα της κατανομής των r_{C_i} με την εκθετική κατανομή με παράμετρο 1, εξαρτάται σημαντικά από την υπόθεση ότι όταν τα β και $H_0(t)$ αντικατασταθούν από τους εκτιμητές τους, ο μετασχηματισμός $F(T)$ δίνει και πάλι μεταβλητές ομοιόμορφα κατανεμημένες στο $(0,1)$. Αυτή η προσέγγιση ενδέχεται να μην ισχύει για μικρά δείγματα. Επιπλέον, αποκλίσεις από την εκθετική κατανομή με παράμετρο 1 μπορεί να οφείλονται εν μέρει, στην αβεβαιότητα των εκτιμήσεων των β και $H_0(t)$. Αυτή η αβεβαιότητα είναι μεγαλύτερη στη δεξιά ουρά της κατανομής και για μικρά δείγματα.

Τη γραφική παράσταση των Cox-Snell υπολοίπων έναντι της αθροιστικής συνάρτησης κινδύνου των Cox-Snell υπολοίπων, για το μοντέλο `cns2.coxint6` που έχουμε προσαρμόσει, τη δίνει ο ακόλουθος κώδικας. Τα Cox-Snell υπόλοιπα τα παίρνουμε μέσω των Martingale υπολοίπων, που είναι τα προκαθορισμένα υπόλοιπα στην R και τα οποία ορίζουμε παρακάτω. Στο αντικείμενο `rc` αποθηκεύονται τα Cox-Snell υπόλοιπα. Στο αντικείμενο `km.rc` αποθηκεύεται η εκτιμώμενη συνάρτηση επιβίωσης για τα Cox-Snell υπόλοιπα, ενώ στο αντικείμενο `summary.km.rc` αποθηκεύονται με λεπτομέρεια οι πληροφορίες για την εκτιμώμενη συνάρτηση επιβίωσης. Μέσω του αντικειμένου `summary.km.rc` αποθηκεύουμε τα Cox-Snell υπόλοιπα που αντιστοιχούν σε μη λογοκριμένους χρόνους ζωής στο αντικείμενο `rcu`. Ομοίως, στο `surv.rc` αποθηκεύουμε τις τιμές της εκτιμώμενης συνάρτησης επιβίωσης. Με την εντολή `plot` κάνουμε τη γραφική παράσταση των Cox-Snell υπολοίπων, έναντι της αθροιστικής συνάρτησης κινδύνου των Cox-Snell υπολοίπων. Μέσω του μετασχηματισμού $-\log(\text{surv.rc})$ παίρνουμε την εκτιμώμενη αθροιστική συνάρτηση κινδύνου των Cox-Snell υπολοίπων.

```
rc <- abs(STATUS - cns2.coxint6$residuals)
km.rc <- survfit(Surv(rc,STATUS) ~ 1)
summary.km.rc <- summary(km.rc)
rcu <- summary.km.rc$time
surv.rc <- summary.km.rc$surv
plot(rcu,-log(surv.rc),type="p",pch=16,cex=0.4,
      xlab="Cox-Snell residual rc",ylab="Cumulative
      hazard of rc")
abline(a=0,b=1); abline(v=0); abline(h=0)
```



Σχήμα 2.15: Γραφική παράσταση των Cox-Snell υπολοίπων έναντι της αθροιστικής συνάρτησης κινδύνου των Cox-Snell υπολοίπων για το μοντέλο `cns2.coxint6`.

Συμπέρασμα:

Παρατηρούμε από τη γραφική παράσταση του Σχήματος 2.15 ότι το τελικό μοντέλο προσαρμόζεται ικανοποιητικά στα δεδομένα. Συνολικά, τα υπόλοιπα βρίσκονται σε μια ευθεία που διέρχεται από την αρχή των αξόνων και έχει κλίση 1. Δεν υπάρχουν μεγάλες αποκλίσεις από αυτήν την ευθεία, ούτε μεγάλη διακύμανση στη δεξιά ουρά της κατανομής.

Martingale υπόλοιπα για τον προσδιορισμό της καλύτερης συν-αρτησιακής μορφής μιας συμμεταβλητής:

Τα martingale υπόλοιπα είναι μια μικρή τροποποίηση των Cox-Snell υπολοίπων. Τα υπόλοιπα αυτά προτάθηκαν στην εργασία των Therneau, Grambsch and Fleming (1990). Όταν τα δεδομένα είναι δεξιά λογοκριμένα και οι συμμεταβλητές είναι ανεξάρτητες του χρόνου, τα martingale υπόλοιπα που συμβολίζονται ως $\widehat{M}_i$

ορίζονται ως:

$$\widehat{M}_i = \delta_i - \hat{H}_0(Y_i) \times \exp\left(\sum_{j=1}^m \hat{\beta}_j X_{ij}\right) = \delta_i - r_{Ci}, i = 1, \dots, n,$$

με r_{Ci} να είναι τα Cox-Snell υπόλοιπα. Τα martingale υπόλοιπα χρησιμοποιούνται για να εξεταστεί η καλύτερη συναρτησιακή μορφή για μια δοθείσα συμμεταβλητή, υποθέτοντας ένα Cox PH μοντέλο για τις υπόλοιπες συμμεταβλητές. Θεωρούμε ότι το διάνυσμα των συμμεταβλητών $\mathbf{X}$ χωρίζεται σε ένα διάνυσμα $\mathbf{X}_*$, για το οποίο ξέρουμε τη συναρτησιακή μορφή και σε μια μοναδική συνεχή συμμεταβλητή X_1 , για την οποία δεν είμαστε βέβαιοι για το τι συναρτησιακή μορφή να χρησιμοποιήσουμε. Υποθέτουμε ότι η X_1 είναι ανεξάρτητη του διανύσματος $\mathbf{X}_*$. Έστω ότι $g(\cdot)$ είναι η συνάρτηση που περιγράφει την επίδραση της X_1 στον χρόνο ζωής με τον καλύτερο τρόπο. Το Cox PH μοντέλο είναι

$$H(T_i|\mathbf{X}_*, X_1) = H_0(t) \times \exp(\mathbf{X}_*'\boldsymbol{\beta}_*) \times \exp\{g(X_1)\},$$

όπου $\boldsymbol{\beta}_*$ είναι ένα $(m-1)$ -διάστατο διάνυσμα συντελεστών των συμμεταβλητών. Για την εύρεση της $g(\cdot)$ προσαρμόζουμε ένα Cox PH μοντέλο για τα δεδομένα που βασίζονται στο $\mathbf{X}_*$ και υπολογίζουμε τα martingale υπόλοιπα, $\widehat{M}_i, i = 1, \dots, n$. Κάνουμε γραφική παράσταση αυτών των υπολοίπων έναντι των $X_{i1}, i = 1, \dots, n$. Συνήθως χρησιμοποιείται μια εξομαλυσμένη γραφική παράσταση (smoothed fit of the scatter plot). Η καμπύλη αυτή δίνει μια ένδειξη της συνάρτησης $g(\cdot)$. Αν η γραφική είναι γραμμική συνάρτηση, τότε δεν χρειάζεται μετατροπή της X_1 . Αν παρατηρηθεί ότι υπάρχει ένα σημείο, τέτοιο ώστε η γραφική παράσταση να έχει διαφορετική μορφή πριν από αυτό και διαφορετική μετά από αυτό (threshold), τότε προτείνεται μια διακριτή μορφή της συμμεταβλητής.

Παρατηρήσεις:

- Τα martingale υπόλοιπα μπορούν να παραχθούν εύκολα από τα Cox-Snell υπόλοιπα.
- Είναι σύνηθες σε πολλές ιατρικές μελέτες να γίνεται διακριτοποίηση συνεχών μεταβλητών. Τα martingale υπόλοιπα είναι χρήσιμα για τον προσδιορισμό του διαχωριστικού σημείου για τέτοιου είδους μεταβλητές.
- Τα martingale υπόλοιπα για ένα άτομο μπορούν να ερμηνευθούν ως η διαφορά του δείκτη λογοκρισίας, δηλαδή του δείκτη που υποδεικνύει αν παρατηρήθηκε το υπό μελέτη γεγονός ή όχι, μείον του αναμενόμενου αριθμού γεγονότων που προβλέπει το Cox μοντέλο που έχουμε υποθέσει.

Στον γενικό ορισμό τους, τα martingale υπόλοιπα εκφράζουν τη διαφορά μεταξύ του παρατηρούμενου και του αναμενόμενου αριθμού γεγονότων για κάθε άτομο. Ωστόσο, δεν θα επεκταθούμε στον γενικό τους ορισμό, καθώς θεωρούμε ότι κάθε άτομο της μελέτης μπορεί να βιώσει το υπό μελέτη γεγονός μία φορά.

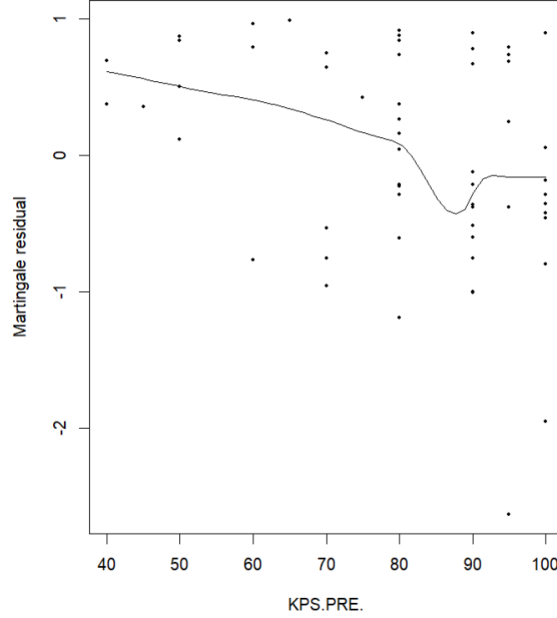
- Τα martingale υπόλοιπα αθροίζουν στο μηδέν, $\sum_{i=1}^n \widehat{M}_i = 0$. Για μεγάλες τιμές του n τα $\widehat{M}_i$ είναι ένα ασυσχέτιστο δείγμα από πληθυσμό με μέση τιμή μηδέν. Ωστόσο, δεν είναι συμμετρικά κατανομημένα γύρω από το μηδέν, αφού τα υπόλοιπα παίρνουν τιμές από $-\infty$ ως 1, η κατανομή τους είναι λοξή.
- Εξαιτίας της λοξής τους κατανομής τα martingale υπόλοιπα δεν είναι χρήσιμα στην ανίχνευση ακραίων τιμών.
- Καθώς τα martingale υπόλοιπα αντιμετωπίζουν διαφορετικά τις περιπτώσεις όπου υπάρχει λογοκρισία σε σχέση με αυτές που δεν υπάρχει, στις γραφικές παραστάσεις τους τείνει να υπάρχει διαχωρισμός μεταξύ των περιπτώσεων.

Συνεχίζοντας με το παράδειγμά μας, έστω ότι θέλουμε να βρούμε την καλύτερη συναρτησιακή μορφή για τη συμεταβλητή KPS.PRE. με χρήση των martingale υπολοίπων. Κάνουμε τη γραφική παράσταση που έχει στον y -άξονα τα martingale υπόλοιπα για το Cox μοντέλο με συμεταβλητές τις GROUP, SEX, AGE60, SEX:AGE60 (όλες εκτός της KPS.PRE.) και στον x -άξονα τη μεταβλητή KPS.PRE.. Το μοντέλο του Cox χωρίς την μεταβλητή KPS.PRE. αποθηκεύεται στο αντικείμενο `fit`, ενώ με την εντολή `scatter.smooth` παίρνουμε τη ζητούμενη γραφική παράσταση.

```
fit <- coxph(Surv(B3TODEATH,STATUS) ~ GROUP+SEX+AGE60
             +SEX:AGE60)
scatter.smooth(cns2$KPS.PRE.,resid(fit),type="p",pch=16,
               cex=0.4,xlab="KPS.PRE.",ylab="Martingale residual")
```

Συμπέρασμα:

Στη γραφική παράσταση του Σχήματος 2.16 παρατηρούμε ότι υπάρχει ένα εξόγκωμα για τη μεταβλητή KPS.PRE. μεταξύ των τιμών 80 και 90. Ωστόσο, οι ευθείες πριν και μετά το εξόγκωμα σχεδόν συμπίπτουν. Συμπεραίνουμε ότι η γραμμική συνάρτηση είναι κατάλληλη για τη μεταβλητή KPS.PRE. και δεν χρειάζεται μετατροπή της KPS.PRE.. Υπάρχουν περιπτώσεις όπου η διακριτοποίηση μίας συνεχούς μεταβλητής είναι καταλληλότερη και δίνει περισσότερες πληροφορίες.



Σχήμα 2.16: Γραφική παράσταση των martingale υπολοίπων για την εύρεση της καλύτερης συναρτησιακής μορφής της συµµεταβλητής KPS.PRE. στο µοντέλο `cns2.coxint6`.

Υπόλοιπα αποκλίσεων για τον εντοπισμό πιθανών ακραίων τιμών:

Όπως αναφέραµε προηγουµένως, τα martingale υπόλοιπα για ένα άτοµο είναι η διαφορά του δείκτη, που υποδεικνύει αν συνέβη το υπό µελέτη γεγονός ή όχι, µείον του αναµενόµενου αριθµού γεγονότων που προβλέπει το Cox µοντέλο που έχουµε υποθέσει. Καθώς όµως, η κατανοµή τους είναι ιδιαίτερα λοξή, δεν µπορούν να χρησιµοποιηθούν για τον εντοπισµό ακραίων τιµών. Τα υπόλοιπα αποκλίσεων χρησιµοποιούνται για να πάρουµε υπόλοιπα που είναι πιο συµµετρικά, κανονικά κατανεµηµένα από τα martingale υπόλοιπα. Τα υπόλοιπα αποκλίσεων προτάθηκαν στην ίδια εργασία µε τα martingale υπόλοιπα, δηλαδή την εργασία των Therneau, Grambsch and Fleming (1990). Τα υπόλοιπα αποκλίσεων ορίζονται ως:

$$D_i = \text{sign}(\widehat{M}_i) \times \sqrt{-2 \times \{\widehat{M}_i + \delta_i \log(\delta_i - \widehat{M}_i)\}}, \quad i = 1, \dots, n,$$

µε $\widehat{M}_i$ να είναι τα martingale υπόλοιπα. Σε όλες τις γραφικές παραστάσεις πιθανές ακραίες τιμές αντιστοιχούν σε υπόλοιπα αποκλίσεων µε μεγάλες απόλυτες

τιμές. Για την αξιολόγηση της επίδρασης ενός δοθέντος ατόμου στο μοντέλο, προτείνεται η κατασκευή της γραφικής παράστασης των υπολοίπων αποκλίσεων D_i , $i = 1, \dots, n$, έναντι των risk scores $\sum_{j=1}^m \beta_j^* X_{ij}$, $i = 1, \dots, n$, όπου παρατηρούμε το σημείο της γραφικής που αναφέρεται στο δοθέν άτομο.

Παρατηρήσεις:

- Όταν το ποσοστό λογοχρισίας είναι μικρό, μικρότερο ή ίσο από 25%, τα υπόλοιπα είναι συμμετρικά κατανεμημένα γύρω από το μηδέν και έχουν τη μορφή ενός δείγματος από κανονική κατανομή. Για λογοχρισία μεγαλύτερη από 40% μια μεγάλη σφαίρα από σημεία με υπόλοιπα κοντά στο μηδέν χαλάει την κανονική προσέγγιση.
- Ο μετασχηματισμός που δίνει τα υπόλοιπα αποκλίσεων είναι χρήσιμος για να γίνει συμμετρικό ένα σύνολο υπολοίπων.
- Προφανώς, τα υπόλοιπα αποκλίσεων δεν αθροίζουν αναγκαστικά στο μηδέν.
- Πιθανές ακραίες τιμές έχουν υπόλοιπα αποκλίσεων με μεγάλες απόλυτες τιμές.

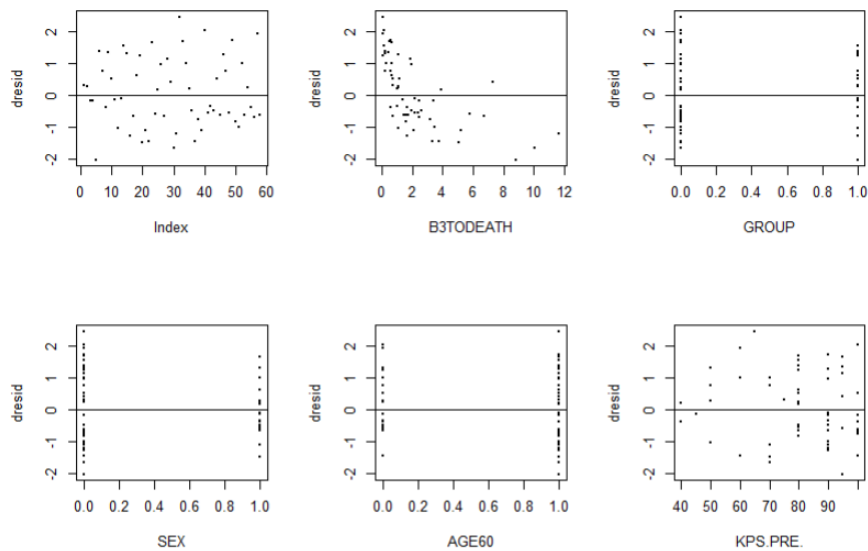
Με τον ακόλουθο κώδικα φτιάχνουμε τις γραφικές παραστάσεις των υπολοίπων αποκλίσεων έναντι του δείκτη i και των υπολοίπων αποκλίσεων έναντι κάθε συμμεταβλητής ξεχωριστά. Μέσω της εντολής **par** χωρίζεται η οθόνη σε ένα 3×3 πίνακα. Η εντολή **resid** δέχεται ως πρώτο όρισμα το μοντέλο για το οποίο θέλουμε να υπολογίσουμε τα υπόλοιπα και στο όρισμα **type** το είδος των υπολοίπων που θέλουμε. Τα υπόλοιπα αποκλίσεων αποθηκεύονται στο αντικείμενο **dresid**. Στην εντολή **plot** δίνουμε ως πρώτο όρισμα τη μεταβλητή που θέλουμε να είναι στον άξονα x και ως δεύτερο τα υπόλοιπα αποκλίσεων. Όταν δίνουμε ως όρισμα τα υπόλοιπα αποκλίσεων και όχι κάποια συμμεταβλητή γίνεται η γραφική παράσταση αυτών των υπολοίπων έναντι του δείκτη i κάθε ατόμου. Με την εντολή **abline(h=0)** προστίθεται στο γράφημα η οριζόντια ευθεία που διέρχεται από το 0.

```
par(mfrow=c(3,3))
dresid <- resid(cns2.coxint6,type="deviance")
plot(dresid,type="p",pch=16,cex=0.4)
abline(h=0)
plot(B3TODEATH,dresid,type="p",pch=16,cex=0.4)
abline(h=0)
plot(GROUP,dresid,type="p",pch=16,cex=0.4)
```

```

abline(h=0)
plot(SEX,dresid,type="p",pch=16,cex=0.4)
abline(h=0)
plot(AGE60,dresid,type="p",pch=16,cex=0.4)
abline(h=0)
plot(KPS.PRE.,dresid,type="p",pch=16,cex=0.4)
abline(h=0)

```



Σχήμα 2.17: Γραφική παράσταση των υπολοίπων αποκλίσεων έναντι του δείκτη i και των υπολοίπων αποκλίσεων έναντι κάθε συμμεταβλητής ξεχωριστά, για το μοντέλο `cns2.coxint6`.

Συμπέρασμα:

Σύμφωνα με τις γραφικές παραστάσεις του Σχήματος 2.17, παρατηρούμε ότι υπάρχει μια μικρή τάση οι μεγάλοι χρόνοι διάρκειας ζωής να έχουν αρνητικά υπολοίπα αποκλίσεων. Το γεγονός αυτό υποδεικνύει ότι το μοντέλο υπερεκτιμά την πιθανότητα να συμβεί το υπό μελέτη γεγονός σε μεγάλους χρόνους διάρκειας ζωής. Υπάρχει μόνο μία πιθανή ακραία τιμή για τον μικρότερο χρόνο διάρκειας ζωής, η οποία πιθανώς να μην επηρεάσει την καταλληλότητα του μοντέλου. Ακόμη, παρατηρούμε ότι τα υπολοίπα αποκλίσεων είναι συμμετρικά κατανομημένα γύρω από το μηδέν.

Schoenfeld υπόλοιπα για την εξέταση της προσαρμογής του μοντέλου:

Το k -οστό Schoenfeld υπόλοιπο, σύμφωνα με την εργασία του Schoenfeld (1982), ορισμένο για το k -οστό άτομο και την j -οστή επεξηγηματική μεταβλητή X_j ορίζεται ως:

$$r_{Skj} = \delta_k \{X_{kj} - a_{kj}\},$$

με δ_k να είναι ο δείκτης λογοκρισίας του k -οστού ατόμου, X_{kj} να είναι η τιμή της j -οστής επεξηγηματικής μεταβλητής για το k άτομο της μελέτης,

$$a_{kj} = \frac{\sum_{m \in R(Y_k)} \exp(\mathbf{X}'_m \hat{\beta}) X_{mj}}{\sum_{m \in R(Y_k)} \exp(\mathbf{X}'_m \hat{\beta})}$$

και $R(Y_k)$ είναι το σύνολο των δεικτών των ατόμων που είναι ζωντανοί και μη λογοκριμένοι ακριβώς πριν τη χρονική στιγμή Y_k (σύνολο κινδύνου τη χρονική στιγμή Y_k). Ο εκτιμητής $\hat{\beta}$ μέγιστης πιθανοφάνειας παρατηρείται από μεγιστοποίηση της συνάρτησης μερικής πιθανοφάνειας του Cox, $L_c(\beta)$. Παρατηρούμε ότι μη μηδενικά υπόλοιπα εμφανίζονται μόνο για μη λογοκριμένες παρατηρήσεις. Τα Schoenfeld υπόλοιπα είναι η διαφορά μεταξύ X_{kj} και του μέσου, πολλαπλασιασμένο με μια συνάρτηση βάρους, των τιμών των επεξηγηματικών μεταβλητών για τα άτομα που βρίσκονται σε κίνδυνο τη χρονική στιγμή Y_k . Η συνάρτηση βάρους για το m -οστό άτομο στο σύνολο $R(Y_k)$ είναι

$$\frac{\exp(\mathbf{X}'_m \hat{\beta})}{\sum_{m \in R(Y_k)} \exp(\mathbf{X}'_m \hat{\beta})},$$

που είναι η συνεισφορά αυτού του ατόμου στη μεγιστοποιημένη συνάρτηση μερικής πιθανοφάνειας. Επιπλέον, καθώς ο εκτιμητής μέγιστης πιθανοφάνειας του β , $\hat{\beta}$ ικανοποιεί τη σχέση

$$\left. \frac{\partial \log(L_c(\beta))}{\partial \beta_j} \right|_{\hat{\beta}} = 0,$$

τα Schoenfeld υπόλοιπα για κάθε μεταβλητή πρόβλεψης X_j αθροίζουν στο μηδέν. Τα Schoenfeld υπόλοιπα έχουν την ιδιότητα ότι για μεγάλα σε μέγεθος δείγματα η αναμενόμενη τιμή του r_{Skj} είναι μηδέν και είναι ασυσχέτιστα μεταξύ τους. Επιπρόσθετα, αν υποθέσουμε ότι Y_k είναι ένας μικρός χρόνος θανάτου σε σχέση με τους υπόλοιπους, τότε το σύνολο κινδύνου του, $R(Y_k)$, είναι μεγάλο. Έτσι, γενικά, όχι μόνο τα στοιχεία στο σύνολο κινδύνου έχουν μεγάλο εύρος τιμών για τις συμμεταβλητές, αλλά επιπλέον η συνάρτηση βάρους που αντιστοιχεί σε κάθε συμμεταβλητή που σχετίζεται με το σύνολο κινδύνου, είναι μικρή. Οπότε,

άτομα με μεγάλες τιμές στις συμμεταβλητές που έχουν μικρούς χρόνους θανάτου, θα έχουν μεγάλα θετικά Schoenfeld υπόλοιπα. Αυτό φαίνεται πιο εύκολα αν γράψουμε το r_{Skj} στη μορφή

$$X_{kj} \left(1 - \frac{\exp(\mathbf{X}'_m \hat{\beta})}{\sum_{m \in R(Y_k)} \exp(\mathbf{X}'_m \hat{\beta})} \right) - \sum_{l \in R(Y_k); l \neq k} \left(X_{lj} \frac{\exp(\mathbf{X}'_l \hat{\beta})}{\sum_{m \in R(Y_k)} \exp(\mathbf{X}'_m \hat{\beta})} \right).$$

Είναι ξεκάθαρο από την παραπάνω έκφραση ότι ο πρώτος όρος είναι μεγάλος και ο δεύτερος είναι μικρός σε σχέση με τον πρώτο. Ομοίως, άτομα με μικρές τιμές στις συμμεταβλητές, που πεθαίνουν σε μικρές χρονικές στιγμές, θα έχουν μεγάλα αρνητικά Schoenfeld υπόλοιπα. Συνεπώς, μερικά υπόλοιπα με σχετικά μεγάλη απόλυτη τιμή για μικρές χρονικές στιγμές θανάτου, μπορεί να μην προκαλούν ιδιαίτερη ανησυχία. Έτσι, αυτά τα υπόλοιπα είναι χρήσιμα για τον εντοπισμό ακραίων τιμών των συμμεταβλητών, για μικρές χρονικές στιγμές θανάτου. Ωστόσο, αν ισχύει η υπόθεση της αναλογικότητας κινδύνων, Schoenfeld υπόλοιπα με μεγάλες τιμές δεν είναι αναμενόμενα για μεγάλες χρονικές στιγμές θανάτων. Συνεπώς, πρέπει να ελέγξουμε τα υπόλοιπα για μεγάλες χρονικές στιγμές θανάτων.

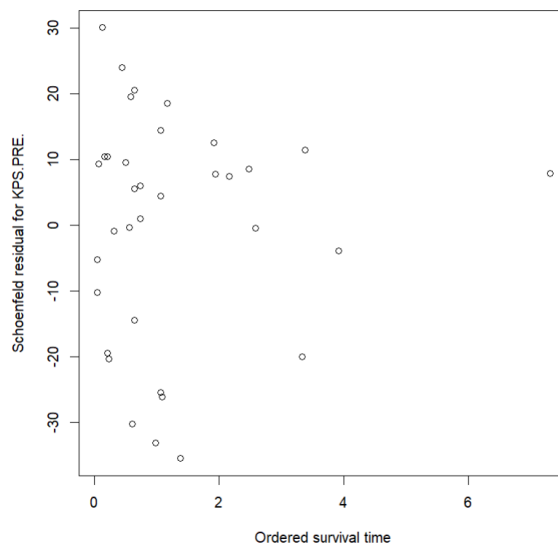
Παρατηρήσεις:

- Ο Schoenfeld (1982) αποκαλεί αυτά τα υπόλοιπα μερικά υπόλοιπα, καθώς προκύπτουν από μεγιστοποίηση της συνάρτησης μερικής πιθανοφάνειας. Ο Collet (2003), μεταξύ άλλων, αποκαλεί αυτά τα υπόλοιπα score υπόλοιπα.
- Αν δεν ισχύει η υπόθεση αναλογικότητας κινδύνων, η γραφική παράσταση των Schoenfeld υπολοίπων, έναντι των ταξινομημένων χρόνων θανάτων, θα εμφανίσει μεγάλα υπόλοιπα σε κάποιες χρονικές στιγμές.
- Τα Schoenfeld υπόλοιπα υπολογίζονται μόνο για τις μη λογοκριμένες παρατηρήσεις. Μπορούν να χρησιμοποιηθούν για συμμεταβλητές που δεν εξαρτώνται από τον χρόνο.
- Τα Schoenfeld υπόλοιπα είναι η διαφορά μεταξύ της παρατηρούμενης τιμής μείον της αναμενόμενης τιμής της συμμεταβλητής X_{kj} , επί τον αντίστοιχο δείκτη λογοκρίσιμης, δ_k , $k = 1, \dots, n$, $j = 1, \dots, m$. Η διαφορά αυτή αναμένεται να είναι προσεγγιστικά μηδέν, αν ισχύει η υπόθεση της αναλογικότητας κινδύνων.

Για τη γραφική παράσταση των Schoenfeld υπολοίπων της συμμεταβλητής KPS.PRE. έναντι των ταξινομημένων μη λογοκριμένων χρόνων επιβίωσης χρησιμοποιούμε τον παρακάτω κώδικα. Στο αντικείμενο `detail` μέσω της εντολής

`coxph.detail` αποθηκεύουμε λεπτομερή πληροφορία για το μοντέλο `cns2.coxint6`. Με χρήση του αντικειμένου `detail` αποθηκεύουμε τους ταξινομημένους χρόνους ζωής, λογοκριμένους και μη, στο αντικείμενο `time`. Όμοια, στο αντικείμενο `status` αποθηκεύουμε τους δείκτες λογοκρισίας. Με την εντολή `resid` παίρνουμε τα Schoenfeld υπόλοιπα για το μοντέλο `cns2.coxint6` και τα αποθηκεύουμε στο αντικείμενο `sch`. Δίνουμε ως πρώτο όρισμα στην `plot` τους μη λογοκριμένους χρόνους ζωής και ως δεύτερο όρισμα την πρώτη στήλη του αντικειμένου `sch` που περιλαμβάνει τα Schoenfeld υπόλοιπα για τη μεταβλητή `KPS.PRE.`. Έτσι γίνεται η ζητούμενη γραφική παράσταση.

```
detail <- coxph.detail(cns2.coxint6)
time <- detail$y[,2] # ταξινομημένοι χρόνοι, περιλαμβάνονται
                     και όσοι αντιστοιχούν σε λογοκριμένους χρόνους ζωής
status <- detail$y[,3]
sch <- resid(cns2.coxint6, type="schoenfeld") # Schoenfeld
                                             υπόλοιπα
plot(time[status==1], sch[,1], xlab="Ordered survival time",
     ylab="Schoenfeld residual for KPS.PRE.")
```



Σχήμα 2.18: Γραφική παράσταση των Schoenfeld υπολοίπων για την μεταβλητή `KPS.PRE.` έναντι των ταξινομημένων χρόνων θανάτων για το μοντέλο `cns2.coxint6`.

Συμπέρασμα:

Σύμφωνα με τη γραφική παράσταση των Schoenfeld υπολοίπων της συμμεταβλητής KPS.PRE. έναντι των ταξινομημένων χρόνων θανάτων που δίνεται στο Σχήμα 2.18, ακραίες τιμές παρατηρούνται για τους χρόνους ζωής 0.125, 0.604, 0.979 και 1.375. Οι συγκεκριμένοι χρόνοι ζωής αντιστοιχούν είτε στις μεγαλύτερες, είτε στις μικρότερες τιμές για τη συμμεταβλητή KPS.PRE.. Συνεπώς, αυτά τα υπόλοιπα δεν μας δημιουργούν ιδιαίτερη ανησυχία. Οι γραφικές παραστάσεις που αντιστοιχούν στις υπόλοιπες συμμεταβλητές παράγονται ομοίως και σε καμία από αυτές δεν εμφανίζονται μεγάλες τιμές των υπολοίπων. Οπότε, η υπόθεση της αναλογικότητας των κινδύνων φαίνεται να ισχύει.

dfbeta για την αξιολόγηση της επιρροής κάθε παρατήρησης:

Έστω ότι θέλουμε να ελέγξουμε την επιρροή κάθε παρατήρησης στον εκτιμητή $\hat{\beta}$ του β . Έστω $\hat{\beta}_k$ το εκτιμώμενο διάνυσμα συντελεστών με βάση το δείγμα, διαγράφοντας την k -οστή παρατήρηση. Έπειτα, ελέγχουμε ποιες συνιστώσες του διανύσματος $\hat{\beta} - \hat{\beta}_k$ έχουν υπερβολικά μεγάλες απόλυτες τιμές. Επαναλαμβάνουμε αυτήν την διαδικασία για κάθε μία από τις n παρατηρήσεις. Η διαδικασία αυτή περιλαμβάνει την προσαρμογή $(n - 1)$ μοντέλων παλινδρόμησης Cox. Προφανώς έχει μεγάλο υπολογιστικό κόστος, εκτός αν το μέγεθος δείγματος είναι μικρό. Ευτυχώς, υπάρχει μια προσέγγιση βασισμένη στην προσαρμογή ενός Cox PH μοντέλου, που την παίρνουμε από τα πλήρη δεδομένα και μπορεί να χρησιμοποιηθεί για να μειώσει το υπολογιστικό κόστος. Το k -οστό dfbeta, σύμφωνα με το σύγγραμμα των Belsley, Kuh and Welsch (1980), ορίζεται ως:

$$dfbeta_k = I(\hat{\beta})^{-1}(r_{Sk1}^*, \dots, r_{Skm}^*)', k = 1, \dots, n,$$

όπου $I(\hat{\beta})^{-1}$ είναι ο αντίστροφος του πίνακα πληροφορίας του Fisher και για $j = 1, \dots, m$,

$$r_{Skj}^* = \delta_k \{X_{kj} - a_{kj}\} - \exp(\mathbf{X}_k' \hat{\beta}) \sum_{t_i \leq Y_k} \frac{\{X_{kj} - a_{ij}\}}{\sum_{l \in R(t_i)} \exp(\mathbf{X}_l' \hat{\beta})}.$$

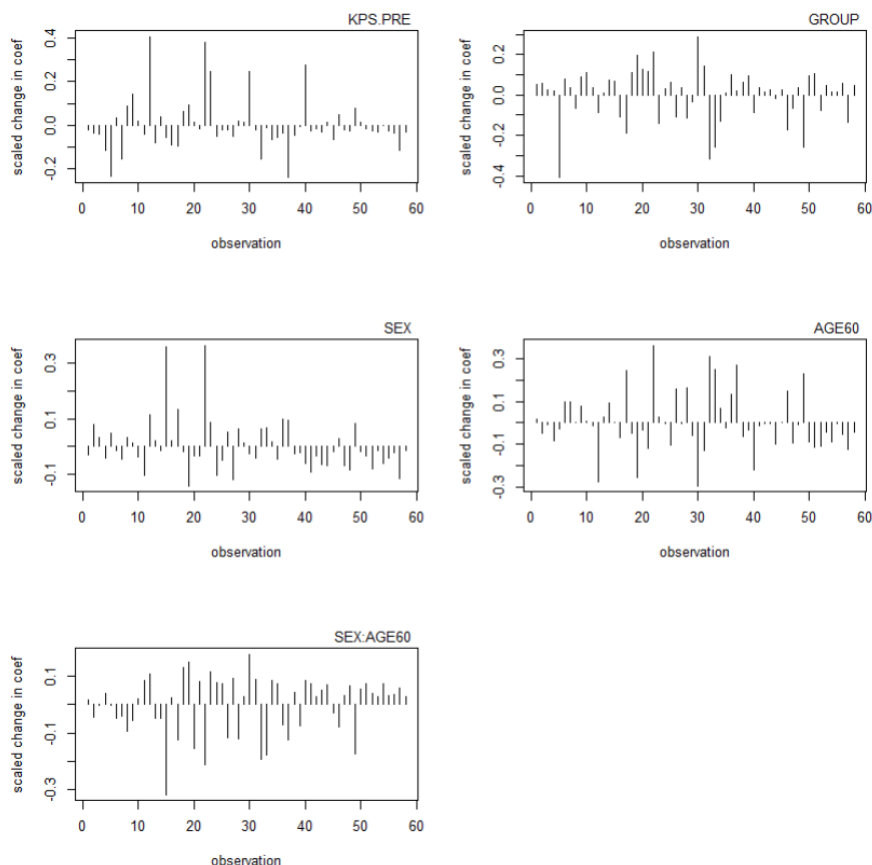
Ο πρώτος όρος είναι το k -οστό Schoenfeld υπόλοιπο. Ο δεύτερος όρος μετρά τη συνδυαστική επίδραση όλων των παρατηρήσεων που περιέχονται σε σύνολα κινδύνου που περιέχουν την k -οστή παρατήρηση. Αυτή η έκφραση, που προτάθηκε από τους Cain και Lange (1984), προσεγγίζει καλά τη διαφορά $\hat{\beta} - \hat{\beta}_k$ για $k = 1, \dots, n$. Οι δύο παραπάνω όροι έχουν γενικά αντίθετα πρόσημα. Ο δεύτερος όρος αυξάνεται κατά απόλυτη τιμή με την αύξηση του Y_k , γιατί είναι το άθροισμα ενός αυξανόμενου πλήθους όρων. Έτσι, για μικρές χρονικές στιγμές θανάτου ο πρώτος όρος υπερिशύει, ενώ για μεγάλες ο δεύτερος όρος είναι συνήθως μεγαλύτερου

μεγέθους. Αυτό σημαίνει ότι για ασθενείς που πεθαίνουν αργά, το γεγονός ότι ο ασθενής έζησε μεγάλο χρονικό διάστημα, και έτσι συμπεριλήφθηκε σε πολλά σύνολα δεικτών των ατόμων που βρίσκονται σε κίνδυνο, έχει περισσότερη επιρροή στο $\hat{\beta}$, από ότι το γεγονός ότι ο ασθενής πέθανε ή λογοκρίθηκε. Οι γραφικές παραστάσεις αυτών των ποσοτήτων έναντι του δείκτη (index case number) ή έναντι της αντίστοιχης συμμεταβλητής X_{kj} , χρησιμοποιούνται για τη μέτρηση της επιρροής της k -οστής παρατήρησης στον j -οστό συντελεστή.

Στο παράδειγμά μας κάνουμε τις γραφικές παραστάσεις των dfbeta υπολογισμένα για κάθε συμμεταβλητή αντίστοιχα, έναντι του δείκτη κάθε ατόμου της μελέτης. Με την εντολή `resid` αποθηκεύουμε στο αντικείμενο `bresid` τα dfbeta υπολογισμένα με βάση το μοντέλο `cns2.coxint6`. Έτσι, η πρώτη στήλη του αντικειμένου περιέχει τα dfbeta που αντιστοιχούν στη συμμεταβλητή `KPS.PRE`, ενώ η δεύτερη στήλη αυτά που αντιστοιχούν στη συμμεταβλητή `GROUP`. Ομοίως, η τρίτη, η τέταρτη και η πέμπτη στήλη περιέχει τα dfbeta για τις συμμεταβλητές `SEX`, `AGE60` και `SEX:AGE60`, αντίστοιχα. Στο αντικείμενο `index` αποθηκεύουμε τον δείκτη που αντιστοιχεί σε κάθε παρατήρηση. Με την εντολή `plot`, δίνοντας ως πρώτο όρισμα τους δείκτες που ορίσαμε και ως δεύτερο όρισμα τα dfbeta που αντιστοιχούν σε μια συμμεταβλητή παίρνουμε και το ανάλογο γράφημα. Με την εντολή `legend` προσθέτουμε τις αντίστοιχες λεζάντες. Έστω ότι έχουμε το πρώτο γράφημα, που έχει στον άξονα x τους δείκτες των παρατηρήσεων και στον άξονα y τα dfbeta που αντιστοιχούν στη συμμεταβλητή `KPS.PRE`. Το γράφημα αυτό απεικονίζει πως επηρεάζει τον εκτιμώμενο συντελεστή της συμμεταβλητής `KPS.PRE` κάθε παρατήρηση.

```
par(mfrow=c(3,2))
bresid <- resid(cns2.coxint6,type="dfbetas")
index <- seq(1:58)
plot(index,bresid[,1],type="h",ylab="scaled change in coef",
      xlab="observation")
legend(x="topright",legend=c("KPS.PRE"),bty = "n")
plot(index,bresid[,2],type="h",ylab="scaled change in coef",
      xlab="observation")
legend(x="topright",legend=c("GROUP"),bty = "n")
plot(index,bresid[,3],type="h",ylab="scaled change in coef",
      xlab="observation")
legend(x="topright",legend=c("SEX"),bty= "n")
plot(index,bresid[,4],type="h",ylab="scaled change in coef",
      xlab="observation")
legend(x="topright",legend=c("AGE60"),bty = "n")
```

```
plot(index,bresid[,5],type="h",ylab="scaled change in coef",
      xlab="observation")
legend(x="topright",legend=c("SEX:AGE60"),bty="n")
```



Σχήμα 2.19: Γραφικές παραστάσεις των dfbeta υπολογισμένα για κάθε συμ-μεταβλητή αντίστοιχα, έναντι του δείκτη κάθε παρατήρησης για το μοντέλο `cns2.coxint6`.

Συμπέρασμα:

Σύμφωνα με τις γραφικές παραστάσεις που δίνονται στο Σχήμα 2.19 η μεταβολή στους συντελεστές είναι μικρή και μπορούμε να συμπεράνουμε ότι δεν υπάρχουν άτομα, παρατηρήσεις, που επηρεάζουν σημαντικά τον εκτιμητή των συντελεστών.

ΚΕΦΑΛΑΙΟ 3

ΕΠΕΚΤΑΣΗ ΤΟΥ ΜΟΝΤΕΛΟΥ ΤΟΥ COX

Σε αυτό το Κεφάλαιο παρουσιάζουμε την επέκταση του μοντέλου του Cox, έτσι ώστε να εφαρμόζεται για χρονικά εξαρτημένες συμμεταβλητές. Για την παρουσίαση που ακολουθεί, βασιζόμαστε στο σύγγραμμα των Kleinbaum και Klein (2012). Η γενική μορφή του μοντέλου αναλογικών κινδύνων, όπως έχει αναφερθεί στην εισαγωγή του Κεφαλαίου 2, είναι:

$$h(t | \mathbf{X}) = h_0(t) \exp(\mathbf{X}'\boldsymbol{\beta}) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m).$$

Το μοντέλο αυτό δίνει τον κίνδυνο που έχει ένα άτομο να βιώσει το υπό μελέτη γεγονός τη χρονική στιγμή t , δοθέντος του διανύσματος των συμμεταβλητών $\mathbf{X} = (X_1, X_2, \dots, X_m)'$. Στη συνέχεια, παραθέτουμε πως ορίζονται οι χρονικά εξαρτημένες μεταβλητές, τις κατηγορίες στις οποίες μπορεί να ανήκουν και δίνουμε κάποια παραδείγματα.

3.1 Μεταβλητές που εξαρτώνται από τον χρόνο

Μια μεταβλητή που εξαρτάται από τον χρόνο είναι μια οποιαδήποτε μεταβλητή η τιμή της οποίας, για ένα δοθέν άτομο της μελέτης, αλλάζει με το πέρασμα του χρόνου. Για παράδειγμα, αν έχουμε τη μεταβλητή, έστω *Smoker*, που θεωρούμε ότι είναι σταθερή με τον χρόνο και την πολλαπλασιάσουμε με τον χρόνο, t , παίρνουμε μια μεταβλητή που εξαρτάται από τον χρόνο, $Smoker \times t$. Τέτοιου είδους μεταβλητές καλούνται καθορισμένες (defined) χρονικά εξαρτημένες μεταβλητές. Οι περισσότερες καθορισμένες χρονικά εξαρτημένες μεταβλητές είναι της μορφής γινομένου μιας μεταβλητής που είναι σταθερή με τον χρόνο επί τον χρόνο ή μια συνάρτηση του χρόνου. Ένα δεύτερο παράδειγμα αυτής της κατηγορίας είναι η μεταβλητή $E \times g(t)$. Η E είναι μια μεταβλητή η τιμή της οποίας καθορίστηκε

στην αρχή της έρευνας, για παράδειγμα η μεταβλητή που δηλώνει αν ένας ασθενής λαμβάνει θεραπεία ή όχι με δυνατές τιμές 0 και 1, ενώ η $g(t)$ ορίζεται ως:

$$g(t) = \begin{cases} 1, & \text{αν } t \geq t_0, \\ 0, & \text{αν } t < t_0, \end{cases}$$

όπου t_0 να είναι μια συγκεκριμένη χρονική στιγμή.

Ένα άλλο είδος χρονικά εξαρτημένων μεταβλητών ονομάζονται εσωτερικά (internal) χρονικά εξαρτημένες μεταβλητές. Παραδείγματα τέτοιων μεταβλητών είναι το αν καπνίζει ή όχι το άτομο της μελέτης τη χρονική στιγμή t , η κατάσταση της θεραπείας του ατόμου τη χρονική στιγμή t ή ο δείκτης παχυσαρκίας του ατόμου τη χρονική στιγμή t . Για τις εσωτερικά χρονικά μεταβαλλόμενες μεταβλητές ο λόγος αλλαγής της τιμής της μεταβλητής στο χρόνο, εξαρτάται από “εσωτερικά” χαρακτηριστικά ή συμπεριφορά που είναι συγκεκριμένη για κάθε άτομο της μελέτης.

Αντίθετα, μια χρονικά εξαρτημένη μεταβλητή ονομάζεται ancillary αν η αλλαγή της τιμής της προκαλείται κυρίως από εξωτερικά χαρακτηριστικά του περιβάλλοντος, τα οποία μπορούν να επηρεάζουν πολλαπλά άτομα της μελέτης ταυτόχρονα. Παραδείγματα τέτοιων μεταβλητών είναι ο δείκτης μόλυνσης του αέρα σε μια γεωγραφική περιοχή τη χρονική στιγμή t , ή η εργασιακή κατάσταση ενός ατόμου τη χρονική στιγμή t , αν υποθέσουμε ότι αν εργάζεται κάποιος ή όχι εξαρτάται κυρίως από τις οικονομικές συνθήκες, αντί των χαρακτηριστικών του ατόμου.

Υπάρχουν και μεταβλητές που μπορούν να θεωρηθούν μερικώς εσωτερικές και μερικώς ancillary χρονικά εξαρτημένες μεταβλητές. Για παράδειγμα, θεωρούμε τη μεταβλητή $KT(t)$ που υποδηλώνει την κατάσταση μεταμόσχευσης νεφρού τη χρονική στιγμή t για έναν ασθενή. Η μεταβλητή $KT(t)$ παίρνει την τιμή 1 τη χρονική στιγμή t , αν ο ασθενής έχει κάνει τη μεταμόσχευση σε μια χρονική στιγμή, έστω t_0 , πριν την t . Η μεταβλητή $KT(t)$ παίρνει την τιμή 0 τη χρονική στιγμή t , αν ο ασθενής δεν έχει κάνει τη μεταμόσχευση μέχρι τη χρονική στιγμή t . Σημειώνουμε ότι αν ο ασθενής κάνει τη μεταμόσχευση τη χρονική στιγμή t_0 , η τιμή της μεταβλητής $KT(t)$ παραμένει 1 για όλες τις ακόλουθες στιγμές. Έτσι, η τιμή της μεταβλητής $KT(t)$ είναι 0 μέχρι τη χρονική στιγμή t_0 και 1 για όλες τις χρονικές στιγμές μετά την t_0 . Αν ένας ασθενής δεν κάνει τη μεταμόσχευση, σε όλη τη διάρκεια της μελέτης η τιμή της μεταβλητής $KT(t)$ θα είναι 0. Η μεταβλητή $KT(t)$ μπορεί να θεωρηθεί εσωτερική, καθώς τα χαρακτηριστικά του ασθενούς είναι καθοριστικά για το αν θα γίνει η επέμβαση ή όχι. Ωστόσο, μπορεί να θεωρηθεί και ancillary, καθώς η διαθεσιμότητα δότη είναι εξωτερικός παράγοντας.

Ο βασικός λόγος που έγινε διαχωρισμός μεταξύ καθορισμένων, εσωτερικών και ancillary χρονικά εξαρτημένων μεταβλητών είναι επειδή οι υπολογιστικές

εντολές που απαιτούνται για τον ορισμό των μεταβλητών που θα χρησιμοποιηθούν στην επέκταση του μοντέλου του Cox, είναι κάπως διαφορετικές για τα διάφορα είδη των μεταβλητών, ανάλογα με το υπολογιστικό πρόγραμμα που χρησιμοποιείται. Ωστόσο, η μορφή της επέκτασης του μοντέλου του Cox παραμένει ίδια ανεξαρτήτως του τύπου των μεταβλητών. Οι διαδικασίες εύρεσης των εκτιμητών των συντελεστών παλινδρόμησης και άλλων παραμέτρων και εύρεσης στατιστικών συμπερασμάτων παραμένουν ίδιες, ανεξαρτήτως του τύπου των μεταβλητών. Αφού δώσαμε τον ορισμό των χρονικά εξαρτημένων μεταβλητών μπορούμε να παραθέσουμε τον ορισμό της επέκτασης του μοντέλου του Cox.

3.2 Ορισμός της επέκτασης του μοντέλου του Cox

Έστω ότι έχουμε ένα σύνολο δεδομένων επιβίωσης στο οποίο περιέχονται και συμμεταβλητές που είναι ανεξάρτητες του χρόνου και που εξαρτώνται από τον χρόνο, έστω δηλαδή ότι έχουμε το διάνυσμα των συμμεταβλητών $\mathbf{X}(t) = (X_1, \dots, X_{m_1}, X_1(t), \dots, X_{m_2}(t))'$, όπου οι πρώτες m_1 συμμεταβλητές είναι ανεξάρτητες του χρόνου και οι υπόλοιπες m_2 συμμεταβλητές εξαρτώνται από τον χρόνο, με $m_1 + m_2 = m$. Οι χρονικά ανεξάρτητες μεταβλητές συμβολίζονται με $X_i, i = 1, \dots, m_1$, ενώ οι χρονικά εξαρτημένες μεταβλητές με $X_j(t), j = 1, \dots, m_2$. Αντίστοιχα, το διάνυσμα των συντελεστών θα είναι $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{m_1}, \delta_1, \dots, \delta_{m_2})'$. Η επέκταση του μοντέλου του Cox ορίζεται ως εξής:

$$\begin{aligned} h(t | \mathbf{X}(t)) &= h_0(t) \exp(\mathbf{X}(t)' \boldsymbol{\beta}) \\ &= h_0(t) \exp(\beta_1 X_1 + \dots + \beta_{m_1} X_{m_1} + \delta_1 X_1(t) + \dots + \delta_{m_2} X_{m_2}(t)) \\ &= h_0(t) \exp \left\{ \sum_{i=1}^{m_1} \beta_i X_i + \sum_{j=1}^{m_2} \delta_j X_j(t) \right\}. \end{aligned}$$

Όπως και στο μοντέλο αναλογικών κινδύνων του Cox, η επέκταση του μοντέλου του Cox περιέχει μια βασική συνάρτηση κινδύνου, $h_0(t)$, η οποία πολλαπλασιάζεται με μια εκθετική συνάρτηση.

Όμοια με το μοντέλο αναλογικών κινδύνων του Cox στην επέκταση του μοντέλου του Cox οι συντελεστές των συμμεταβλητών παρατηρούνται μεγιστοποιώντας τη συνάρτηση μερικής πιθανοφάνειας. Οι υπολογισμοί είναι πιο περίπλοκοι από ότι είναι αυτοί στο μοντέλο αναλογικών κινδύνων του Cox. Ο λόγος είναι ότι το σύνολο των δεικτών των ατόμων που είναι ζωντανοί και μη λογοκριμένοι ακριβώς πριν τη χρονική στιγμή t , το οποίο χρησιμοποιείται για να υπολογιστεί η συνάρτηση μερικής πιθανοφάνειας, υπολογίζεται πιο δύσκολα όταν λαμβάνουμε υπόψη μεταβλητές που εξαρτώνται από τον χρόνο.

Μια σημαντική υπόθεση της επέκτασης του μοντέλου του Cox είναι ότι, η επίδραση μιας χρονικά εξαρτημένης συμμεταβλητής, έστω $X_j(t)$, στην πιθανότητα επιβίωσης τη χρονική στιγμή t , εξαρτάται από την τιμή της συμμεταβλητής τη χρονική στιγμή t και όχι από την τιμή που έχει σε προηγούμενες ή επόμενες χρονικές στιγμές. Σημειώνουμε ότι παρόλο που η τιμή της συμμεταβλητής μπορεί να αλλάζει με τον χρόνο, το μοντέλο για τη συνάρτηση κινδύνου δίνει μόνο έναν συντελεστή για κάθε χρονικά μεταβαλλόμενη συμμεταβλητή στο μοντέλο. Έτσι, τη χρονική στιγμή t υπάρχει μόνο μια τιμή της μεταβλητής $X_j(t)$ που επιδρά στον κίνδυνο, η τιμή της την χρονική στιγμή t .

Ωστόσο, είναι δυνατό να τροποποιήσουμε τη μορφή των χρονικά εξαρτημένων συμμεταβλητών, έτσι ώστε να λάβουμε υπόψη την επίδραση του χρόνου καθυστέρησης (lag-time effect). Για παράδειγμα, υποθέτουμε ότι έχουμε τη μεταβλητή $EMP(t)$ που εκφράζει την εργασιακή κατάσταση του ατόμου την εβδομάδα t . Η επέκταση του μοντέλου του Cox που δεν λαμβάνει υπόψη την επίδραση του χρόνου καθυστέρησης, υποθέτει ότι η επίδραση της εργασιακής κατάστασης στην πιθανότητα επιβίωσης την εβδομάδα t εξαρτάται από την παρατηρούμενη τιμή της συμμεταβλητής $EMP(t)$ για αυτήν την εβδομάδα και όχι, για παράδειγμα, από την τιμή της την προηγούμενη εβδομάδα, δηλαδή

$$h\{t \mid \mathbf{X}(t)\} = h_0(t) \exp \{\delta EMP(t)\}.$$

Ωστόσο, αν επιτρέψουμε έναν χρόνο καθυστέρησης, έστω μια εβδομάδα, η μεταβλητή $EMP(t)$ τροποποιείται έτσι ώστε η συνάρτηση κινδύνου τη χρονική στιγμή t να υπολογίζεται με βάση την τιμή της μεταβλητής την $t - 1$ εβδομάδα. Οπότε, η μεταβλητή $EMP(t)$ αντικαθίσταται στο μοντέλο από την $EMP(t - 1)$, δηλαδή

$$h\{t \mid \mathbf{X}(t)\} = h_0(t) \exp \{\delta EMP(t - 1)\}.$$

Γενικότερα, το εκτεταμένο μοντέλο του Cox μπορεί εναλλακτικά να γραφεί στην ακόλουθη μορφή, ώστε να λαμβάνεται υπόψη η επίδραση του χρόνου καθυστέρησης οποιασδήποτε χρονικά εξαρτημένης συμμεταβλητής.

$$h\{t \mid \mathbf{X}(t)\} = h_0(t) \exp \left\{ \sum_{i=1}^{m_1} \beta_i X_i + \sum_{j=1}^{m_2} \delta_j X_j(t - L_j) \right\},$$

όπου L_j είναι ο χρόνος καθυστέρησης ορισμένος για την j -οστή χρονικά εξαρτημένη συμμεταβλητή.

Κεφάλαιο 3.3.3. Η αναλογία κινδύνου για την επέκταση του μοντέλου του Cox

3.3 Η αναλογία κινδύνου για την επέκταση του μοντέλου του Cox

Το hazard ratio (HR) ή αναλογία κινδύνου είναι ένα αριθμητικό μέτρο και ορίζεται ως ο κίνδυνος ενός ατόμου με χαρακτηριστικά $\mathbf{X}^*(t)$ ως προς τον κίνδυνο ενός ατόμου με χαρακτηριστικά $\mathbf{X}(t)$. Ο γενικός τύπος του HR, για την επέκταση του μοντέλου του Cox, είναι ο εξής:

$$\begin{aligned} HR\{t|\mathbf{X}^*(t), \mathbf{X}(t)\} &= \frac{h\{t|\mathbf{X}^*(t)\}}{h\{t|\mathbf{X}(t)\}} \\ &= \exp \left[\sum_{i=1}^{m_1} \beta_i (X_i^* - X_i) + \sum_{j=1}^{m_2} \delta_j \{X_j^*(t) - X_j(t)\} \right]. \end{aligned}$$

Συνεπάγεται ότι γενικός τύπος του εκτιμώμενου HR, για την επέκταση του μοντέλου του Cox, θα είναι:

$$\begin{aligned} \widehat{HR}\{t|\mathbf{X}^*(t), \mathbf{X}(t)\} &= \frac{\hat{h}\{t|\mathbf{X}^*(t)\}}{\hat{h}\{t|\mathbf{X}(t)\}} \\ &= \exp \left[\sum_{i=1}^{m_1} \hat{\beta}_i (X_i^* - X_i) + \sum_{j=1}^{m_2} \hat{\delta}_j \{X_j^*(t) - X_j(t)\} \right]. \end{aligned}$$

Ο τύπος αυτός περιγράφει τον λόγο των συναρτήσεων κινδύνου τη χρονική στιγμή t , δύο ατόμων με διανύσματα συμμεταβλητών $\mathbf{X}^*(t)$ και $\mathbf{X}(t)$. Τα δύο διανύσματα συμμεταβλητών $\mathbf{X}^*(t)$ και $\mathbf{X}(t)$ περιέχουν τις τιμές των συμμεταβλητών που είναι ανεξάρτητες του χρόνου και τις τιμές των συμμεταβλητών που εξαρτώνται από τον χρόνο, μετρημένες τη χρονική στιγμή t , δηλαδή

$$\begin{aligned} \mathbf{X}^*(t) &= (X_1^*, \dots, X_{m_1}^*, X_1^*(t), \dots, X_{m_2}^*(t))', \\ \mathbf{X}(t) &= (X_1, \dots, X_{m_1}, X_1(t), \dots, X_{m_2}(t))'. \end{aligned}$$

Παράδειγμα 1:

Έστω ότι έχουμε ένα μοντέλο που περιέχει μόνο μια συμμεταβλητή ανεξάρτητη του χρόνου, έστω την κατάσταση θεραπείας E με δυνατές τιμές 0 και 1, και μια συμμεταβλητή που εξαρτάται από τον χρόνο, έστω την $E \times t$.

$$E = \begin{cases} 1, & \text{αν λαμβάνεται θεραπεία,} \\ 0, & \text{αν δεν λαμβάνεται θεραπεία.} \end{cases}$$

Η επέκταση του μοντέλου του Cox είναι $h\{t | \mathbf{X}(t)\} = h_0(t) \exp\{\beta E + \delta(E \times t)\}$. Αν $\mathbf{X}^*(t) = (E = 1, E \times t = t)$, $\mathbf{X}(t) = (E = 0, E \times t = 0)$, η εκτιμώμενη αναλογία κινδύνου που συγκρίνει άτομα που δέχτηκαν θεραπεία με άτομα που δεν

Κεφάλαιο 3.3.3. Η αναλογία κινδύνου για την επέκταση του μοντέλου του Cox

δέχτηκαν θεραπεία τη χρονική στιγμή t είναι:

$$\begin{aligned}\widehat{HR}\{t|\mathbf{X}^*(t), \mathbf{X}(t)\} &= \frac{\hat{h}\{t | \mathbf{X}^*(t) = (1, t)\}}{\hat{h}\{t | \mathbf{X}(t) = (0, 0)\}} \\ &= \exp[\hat{\beta}(X_1^* - X_1) + \hat{\delta}\{X_1^*(t) - X_1(t)\}] \\ &= \exp\{\hat{\beta}(1 - 0) + \hat{\delta}(1 \times t - 0 \times t)\} \\ &= \exp(\hat{\beta} + \hat{\delta}t).\end{aligned}$$

Το HR είναι συνάρτηση του χρόνου. Γενικά, αν ο συντελεστής $\hat{\delta}$ είναι θετικός, τότε η αναλογία κινδύνου αυξάνεται εκθετικά με την αύξηση του χρόνου. Συνεπώς, σε αυτό το παράδειγμα το HR δεν είναι σταθερό με τον χρόνο, οπότε δεν ισχύει η υπόθεση της αναλογικότητας κινδύνου για το συγκεκριμένο μοντέλο.

Γενικότερα, η επέκταση του μοντέλου αναλογικών κινδύνων του Cox δεν ικανοποιεί την υπόθεση της αναλογικότητας κινδύνων όταν έστω ένας συντελεστής $\hat{\delta}_j$ δεν είναι ίσος με μηδέν. Ο συντελεστής $\hat{\delta}_j$ στον τύπο του HR δεν εξαρτάται από τον χρόνο. Ο συντελεστής $\hat{\delta}_j$ εκφράζει τη συνολική επίδραση της αντίστοιχης χρονικά εξαρτημένης συμμεταβλητής, δηλαδή της $X_j(t)$, λαμβάνοντας υπόψη όλες τις χρονικές στιγμές στις οποίες μετρήθηκε η συμμεταβλητή κατά τη διάρκεια της μελέτης.

Παράδειγμα 2:

Θεωρούμε ότι έχουμε την επέκταση του μοντέλου του Cox με μόνο μια μεταβλητή, έστω την εβδομαδιαία μέτρηση της κατάστασης που αφορά την έκθεση σε χημικές ουσίες. Συμβολίζουμε τη μεταβλητή με $E(t)$ με δυνατές τιμές 0 και 1.

$$\begin{aligned}E(t) &= \text{η κατάσταση έκθεσης σε χημικές ουσίες τη χρονική στιγμή } t \\ &\quad (\text{εβδομαδιαία}) \\ &= \begin{cases} 1, & \text{αν υπάρχει έκθεση σε χημικές ουσίες το χρόνο } t, \\ 0, & \text{αν δεν υπάρχει έκθεση σε χημικές ουσίες το χρόνο } t. \end{cases}\end{aligned}$$

Σημειώνουμε ότι, στο συγκεκριμένο παράδειγμα δεν θεωρούμε δύο ξεχωριστές ομάδες ατόμων όπου η μια ομάδα είναι συνεχώς εκτεθειμένη σε χημικές ουσίες, ενώ η άλλη όχι. Αν είχαμε μια τέτοια περίπτωση θα χρειαζόταν μια μεταβλητή ανεξάρτητη του χρόνου που θα λάμβανε τις τιμές 0 και 1. Στο παράδειγμα μας η τιμή της μεταβλητής μπορεί να αλλάζει κάθε εβδομάδα, για κάθε άτομο της μελέτης. Η επέκταση του μοντέλου του Cox που περιέχει μόνο τη μεταβλητή $E(t)$ είναι:

$$h\{t | \mathbf{X}(t)\} = h_0(t) \exp\{\delta E(t)\}.$$

Στο συγκεκριμένο μοντέλο οι τιμές της μεταβλητής μπορούν να αλλάζουν στον χρόνο για τα άτομα της μελέτης, ωστόσο υπάρχει μόνο ένας συντελεστής, δ , που αντιστοιχεί στη μεταβλητή του μοντέλου. Συνεπώς, ο συντελεστής δ αντιπροσωπεύει τη συνολική επίδραση της μεταβλητής $E(t)$ στον χρόνο επιβίωσης. Επιπλέον, η αναλογία κινδύνου που συγκρίνει ένα άτομο της μελέτης που εκτέθηκε σε χημικές ουσίες, με ένα άτομο που δεν εκτέθηκε, για τη χρονική στιγμή t_1 , ισούται με $\exp(\delta)$, καθώς

$$\widehat{HR}(t_1) = \frac{\hat{h}\{t_1 \mid E(t_1) = 1\}}{\hat{h}\{t_1 \mid E(t_1) = 0\}} = \exp\{\delta(1 - 0)\} = \exp(\delta).$$

Παρόλο που το αποτέλεσμα είναι ένας σταθερός αριθμός δεν ισχύει η υπόθεση της αναλογικότητας κινδύνου. Ο σταθερός αριθμός $\exp(\delta)$, είναι η αναλογία κινδύνου για μια δοθείσα χρονική στιγμή t_1 , θεωρώντας ότι για αυτήν τη χρονική στιγμή στον αριθμητή έχουμε ένα άτομο που εκτέθηκε σε χημικές ουσίες και στον παρονομαστή ένα που δεν εκτέθηκε. Συνεπώς, το HR εξαρτάται από τον χρόνο, καθώς η κατάσταση έκθεσης εξαρτάται από τον χρόνο, παρόλο που ο τύπος δίνει έναν σταθερό αριθμό. Για να γίνει πιο κατανοητό υπολογίζουμε το HR για τα παραπάνω άτομα σε μια χρονική στιγμή t_2 της μελέτης κατά την οποία και τα δύο έχουν εκτεθεί σε χημικές ουσίες.

$$\widehat{HR}(t_2) = \frac{\hat{h}\{t_2 \mid E(t_2) = 1\}}{\hat{h}\{t_2 \mid E(t_2) = 1\}} = \exp\{\delta(1 - 1)\} = \exp(0) = 1.$$

Παρατηρούμε ότι, το HR που προκύπτει είναι και πάλι ένας σταθερός αριθμός αλλά διαφορετικός από αυτόν που είχαμε τη χρονική στιγμή t_1 για τα ίδια άτομα της μελέτης.

3.4 Εφαρμογή

Στο παράδειγμα που θα παρουσιαστεί γίνεται χρήση των δεδομένων `addicts`. Τα δεδομένα αυτά αφορούν τη θεραπεία ατόμων εθισμένων στην ηρωίνη, τα άτομα αυτά κατανέμονται σε δύο κλινικές. Ο χρόνος ζωής ενός ασθενή ορίζεται ως ο χρόνος σε μέρες μέχρι να λάβει εξιτήριο από την κλινική ή μέχρι τη λογοκρισία. Βασικός σκοπός της έρευνας είναι να συγκρίνει τον χρόνο παραμονής των ασθενών στις δύο κλινικές. Αρχικά, καλούμε τη βιβλιοθήκη `survival` καθώς περιέχει τις συναρτήσεις `Surv` και `coxph`. Η εντολή `Surv` δέχεται ως πρώτο όρισμα τον χρόνο, σε μέρες, μέχρι ο ασθενής να λάβει εξιτήριο από την κλινική ή μέχρι να λογοκριθεί και ως δεύτερο όρισμα τους δείκτες λογοκρισίας. Η εντολή αυτή

δημιουργεί ένα αντικείμενο (object) που θα είναι το όρισμα της εντολής `coxph`. Η εντολή `coxph` προσαρμόζει το μοντέλο του Cox με συμμεταβλητές τις `clinic`, `prison` και `dose`. Επισημαίνουμε ότι συμμεταβλητή κύριου ενδιαφέροντος είναι η `clinic`. Το μοντέλο αποθηκεύεται στο αντικείμενο `fit1`.

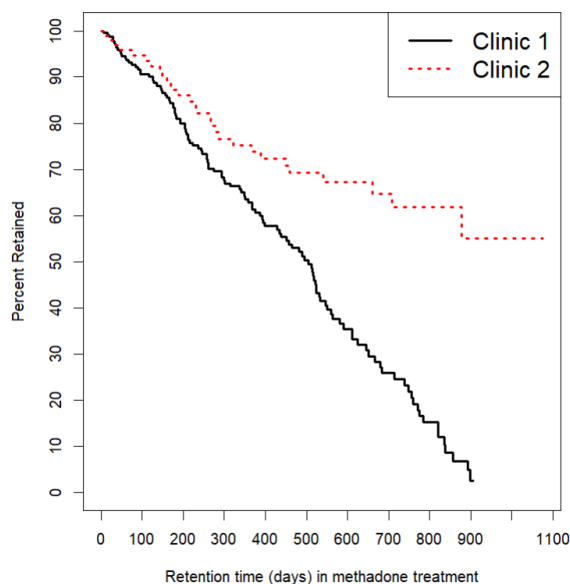
```
library(survival)
attach(addicts)
fit1 <- coxph(Surv(time,status) ~clinic+prison+dose,
             data =addicts,x =T)
```

Στη συνέχεια, θέλουμε να ελέγξουμε αν ισχύει η υπόθεση της αναλογικότητας κινδύνων για τη μεταβλητή `clinic`. Για τον έλεγχο αυτό κάνουμε τις γραφικές παραστάσεις των συναρτήσεων επιβίωσης για τις δύο κατηγορίες της μεταβλητής `clinic`. Αυτό επιτυγχάνεται στρωματοποιώντας ως προς τη μεταβλητή `clinic` και έχοντας προσαρμόσει ως συμμεταβλητές τις `prison` και `dose`. Στο αντικείμενο `addph` αποθηκεύουμε το μοντέλο του Cox με μεταβλητή στρωματοποίησης την `clinic` και συμμεταβλητές τις `prison` και `dose`. Με την εντολή `survfit` υπολογίζεται η εκτιμώμενη συνάρτηση επιβίωσης για το μοντέλο `addph` και αποθηκεύεται στο αντικείμενο `addict.fit`. Δίνοντας ως πρώτο όρισμα το `addict.fit` στην εντολή `plot`, γίνεται η γραφική παράσταση των εκτιμώμενων συναρτήσεων επιβίωσης όταν μεταβλητή στρωματοποίησης είναι η `clinic`. Η εντολή `legend` προσθέτει την αντίστοιχη λεζάντα. Το αποτέλεσμα φαίνεται στο Σχήμα 3.1.

```
addph<-coxph(Surv(time,status)~strata(clinic)+prison+dose,
             data= addicts)
addict.fit <-survfit(addph)
plot(addict.fit, lwd = 3, col = c("black", "red"),
     lty =c(1,3),cex = 2, lab = c(10, 10, 7),
     xlab = "Retention time (days) in methadone treatment",
     ylab = "Percent Retained", yscale = 100)
legend("topright", legend = c("Clinic 1", "Clinic 2"),
     col = c("black", "red"), lty = c(1, 3), lwd = 3, cex =
     1.5)
```

Συμπεράσματα:

- Από το γράφημα του Σχήματος 3.1 βλέπουμε ότι οι δύο συναρτήσεις επιβίωσης είναι πολύ κοντά για μικρούς χρόνους, περίπου μέχρι τον πρώτο χρόνο (365 μέρες), ωστόσο οι δύο καμπύλες αποκλίνουν σημαντικά από εκεί και



Σχήμα 3.1: Γραφική παράσταση των εκτιμώμενων συναρτήσεων επιβίωσης για τις κατηγορίες της μεταβλητής `clinic`, με σκοπό τον έλεγχο της υπόθεσης της αναλογικότητας κινδύνου για την μεταβλητή `clinic` στο μοντέλο `fit1`.

έπειτα. Συνεπώς, δεν ισχύει η υπόθεση της αναλογικότητας κινδύνων για τη μεταβλητή `clinic`. Επιπλέον, το HR θα είναι κοντά στη μονάδα για μικρούς χρόνους και από εκεί και έπειτα θα διαφέρει.

- Παρατηρούμε ότι η συνάρτηση επιβίωσης που αντιστοιχεί σε ασθενείς της κλινικής 2 είναι συνεχώς πάνω από αυτήν για ασθενείς της κλινικής 1. Αυτή είναι μια ένδειξη ότι η κλινική 2 διατηρεί περισσότερο χρόνο τους ασθενείς, υπό την θεραπεία μεθαδόνης. Επιπλέον, καθώς η απόκλιση των δύο καμπυλών αυξάνεται μετά τον πρώτο χρόνο, προκύπτει ότι η κλινική 2 διατηρεί τους ασθενείς σε θεραπεία για σημαντικά μεγαλύτερο χρονικό διάστημα μετά το πέρας του πρώτου έτους.

Καθώς η μεταβλητή `clinic` είναι η μεταβλητή στρωματοποίησης στην ανάλυση δεν μπορούμε να παρατηρήσουμε το HR που αφορά την επίδραση της, μετά την προσαρμογή των συμμεταβλητών `prison` και `dose` στο μοντέλο. Ωστόσο, μπορούμε να παρατηρήσουμε το HR κάνοντας ανάλυση με χρήση της επέκτασης του μοντέλου του Cox, το οποίο θα περιέχει μια συνάρτηση $g(t)$ μαζί με τη

μεταβλητή `clinic`, δηλαδή θα είναι:

$$h\{t \mid \mathbf{X}(t)\} = h_0(t) \exp\{\beta_1 \text{clinic} + \beta_2 \text{prison} + \beta_3 \text{dose} + \delta \text{clinic} \times g(t)\},$$

όπου

$$g(t) = \begin{cases} 1, & \text{αν } t \geq 365 \text{ μέρες,} \\ 0, & \text{αν } t < 365 \text{ μέρες.} \end{cases}$$

Η συνάρτηση $g(t)$ ορίστηκε κατά αυτόν τον τρόπο καθώς με βάση τη γραφική παράσταση των συναρτήσεων επιβίωσης που προαναφέρθηκε, είναι εύλογο να ορίσουμε ως σημείο αλλαγής της τιμής της συνάρτησης τον ένα χρόνο. Το μοντέλο αυτό δίνει δυο HR, το ένα θα είναι σταθερό για $t < 365$ και το άλλο για $t \geq 365$ μέρες. Επομένως, ισχύει ότι

$$\begin{aligned} \text{αν } t \geq 365 \text{ μέρες: } HR &= \exp(\beta_1 + \delta), \\ \text{ενώ αν } t < 365 \text{ μέρες: } HR &= \exp(\beta_1). \end{aligned}$$

Επιπλέον, κωδικοποιούμε τη μεταβλητή `clinic` ως εξής:

$$\text{clinic} = \begin{cases} 1, & \text{για την κλινική 1,} \\ 0, & \text{για την κλινική 2,} \end{cases}$$

για να έχουμε το κατάλληλο output. Ένας ισοδύναμος τρόπος να γράψουμε το μοντέλο είναι με χρήση δύο συναρτήσεων $g_1(t)$ και $g_2(t)$ που ορίζονται ως:

$$g_1(t) = \begin{cases} 1, & \text{αν } t < 365 \text{ μέρες,} \\ 0, & \text{αν } t \geq 365 \text{ μέρες,} \end{cases}$$

$$g_2(t) = \begin{cases} 1, & \text{αν } t \geq 365 \text{ μέρες,} \\ 0, & \text{αν } t < 365 \text{ μέρες.} \end{cases}$$

Σε αυτήν την περίπτωση η επέκταση του μοντέλου του Cox είναι:

$$h\{t \mid \mathbf{X}(t)\} = h_0(t) \exp\{\beta_1 \text{prison} + \beta_2 \text{dose} + \delta_1 \text{clinic} \times g_1(t) + \delta_2 \text{clinic} \times g_2(t)\}.$$

Αντίστοιχα με την προηγούμενη περίπτωση, η επίδραση της μεταβλητής `clinic` περιγράφεται με δύο HR, ένα για χρόνο $t < 365$ και ένα για $t \geq 365$ μέρες. Η αναλογία κινδύνων των ατόμων που ανήκουν στην κλινική 1 προς των ατόμων που ανήκουν στην κλινική 2 είναι:

- Για $t < 365$ μέρες

$$\begin{aligned} HR(t) &= \frac{h_0(t) \exp\{\beta_1 \text{prison} + \beta_2 \text{dose} + \delta_1 \times 1 \times 1 + \delta_2 \times 1 \times 0\}}{h_0(t) \exp\{\beta_1 \text{prison} + \beta_2 \text{dose} + \delta_1 \times 0 \times 1 + \delta_2 \times 0 \times 0\}} \\ &= \frac{\exp(\delta_1)}{\exp(0)} = \exp(\delta_1) \end{aligned}$$

- Για $t \geq 365$ μέρες

$$\begin{aligned} HR(t) &= \frac{h_0(t) \exp\{\beta_1 \text{prison} + \beta_2 \text{dose} + \delta_1 \times 1 \times 0 + \delta_2 \times 1 \times 1\}}{h_0(t) \exp\{\beta_1 \text{prison} + \beta_2 \text{dose} + \delta_1 \times 0 \times 0 + \delta_2 \times 0 \times 1\}} \\ &= \frac{\exp(\delta_2)}{\exp(0)} = \exp(\delta_2) \end{aligned}$$

Συνεπώς, τα εκτιμώμενα HR θα είναι:

$$\begin{aligned} \text{για } t < 365 \text{ μέρες: } \widehat{HR}(t) &= \exp(\hat{\delta}_1), \\ \text{ενώ για } t \geq 365 \text{ μέρες: } \widehat{HR}(t) &= \exp(\hat{\delta}_2). \end{aligned}$$

Παρακάτω παρουσιάζουμε την προσαρμογή αυτού του μοντέλου στην R. Αρχικά, με την εντολή `as.integer` επανακωδικοποιούμε τη μεταβλητή `clinic` έτσι ώστε να έχει την τιμή 1 για την κλινική 1 και την τιμή 0 για την κλινική 2. Δηλαδή, κρατάμε την τιμή 1 όπου υπήρχε και αντικαθιστούμε την τιμή 2 με την τιμή 0.

```
clinic <- as.integer(addicts$clinic==1)
```

Προκειμένου να προσαρμόσουμε το μοντέλο που θέλουμε, το σύνολο δεδομένων `addicts` πρέπει να τροποποιηθεί έτσι ώστε να έχει τη μορφή `counting process`. Για να το κάνουμε αυτό φτιάχνουμε τη συνάρτηση `extcox.1ET`. Η συνάρτηση αυτή δέχεται στο πρώτο όρισμα το σύνολο δεδομένων, στο δεύτερο όρισμα τους χρόνους ζωής ή λογοχρισίας, στο τρίτο όρισμα τους δείκτες λογοχρισίας, στο τέταρτο όρισμα τη μεταβλητή με τις δύο κατηγορίες που θα πολλαπλασιαστεί με τις συναρτήσεις $g_1(t)$ και $g_2(t)$ και ως πέμπτο όρισμα το χρονικό σημείο αλλαγής.

```
extcox.1ET<-function(data,end,status,trt,cut){
  data<-cbind(data,end,status,trt)
  AG<-survSplit(Surv(end, status) ~ ., data, cut=cut,
    start="start")
}
```

```

AG$ET1<-AG$trt
AG$ET1[AG$start!=0] <-0
AG$ET2<-AG$trt
AG$ET2[AG$start==0] <-0
return(AG)
}

```

Με χρήση αυτής της συνάρτησης επανακωδικοποιούμε τα δεδομένα μας και τα αποθηκεύουμε στο αντικείμενο `out`. Στα αντικείμενα `Clinicg1t` και `Clinicg2t` αποθηκεύουμε τις στήλες `ET1` και `ET2` των δεδομένων, αντίστοιχα, οι οποίες περιέχουν τις τιμές $\text{clinic} \times g_1(t)$ και $\text{clinic} \times g_2(t)$, αντίστοιχα.

```

out<-extcox.1Et(addicts,addicts$time,addicts$status,
               clinic,365)
Clinicg1t <- out$ET1
Clinicg2t <- out$ET2

```

Με χρήση της εντολής `coxph` προσαρμόζουμε την επέκταση του μοντέλου του Cox με συμμεταβλητές τις `prison`, `dose`, `Clinicg1t` και `Clinicg2t`. Το μοντέλο αυτό αποθηκεύεται στο αντικείμενο `fit3`. Οι πληροφορίες που περιέχει το μοντέλο εμφανίζονται στην οθόνη όταν καλούμε το αντικείμενο `fit3` και δίνονται παρακάτω.

```

fit3 <- coxph(Surv(start,end,status) ~ prison+dose+
              Clinicg1t+Clinicg2t,data=out)
fit3

```

```

coxph(formula = Surv(start, end, status) ~ prison + dose
      + Clinicg1t + Clinicg2t, data = out)

```

	coef	exp(coef)	se(coef)	z	p
prison	0.377951	1.459291	0.168415	2.244	0.0248
dose	-0.035480	0.965142	0.006435	-5.514	3.52e-08
Clinicg1t	0.459373	1.583081	0.255290	1.799	0.0720
Clinicg2t	1.830517	6.237113	0.385954	4.743	2.11e-06

```

Likelihood ratio test=74.25 on 4 df, p=2.868e-15
n= 360, number of events= 150

```

Με τον ακόλουθο κώδικα υπολογίζουμε το εκτιμώμενο HR της συµμεταβλητής `clinic`, μεταξύ ατόμων από τις δύο κλινικές με ίδιες τιμές στις συµμεταβλητές `prison` και `dose`, για $t < 365$ μέρες και για $t \geq 365$ μέρες, καθώς και τα αντίστοιχα 95% διαστήματα εμπιστοσύνης. Στο αντικείμενο `ci` αποθηκεύουμε τα εκτιμώμενα HR και 95% διαστήματα εμπιστοσύνης όλων των συµμεταβλητών και στην συνέχεια εκτυπώνουμε μόνο αυτά που αντιστοιχούν στις μεταβλητές `Clinicg1t` και `Clinicg2t`.

```
ci <- exp(cbind(HR = coef(fit3),
               Lower = coef(fit3) - 1.96 * sqrt(diag(vcov(fit3))),
               Upper = coef(fit3) + 1.96 * sqrt(diag(vcov(fit3))))))

print(ci["Clinicg1t", ])
      HR      Lower      Upper
1.5830814 0.9598336 2.6110221

print(ci["Clinicg2t", ])
      HR      Lower      Upper
6.237113  2.927202 13.289681
```

Σύμφωνα με τα παραπάνω αποτελέσματα έχουμε ότι:

- Για $t < 365$ μέρες το εκτιμώμενο HR της συµμεταβλητής `clinic` ισούται με 1.5830814. Ωστόσο, αυτό το HR είναι μη στατιστικά σημαντικό σε επίπεδο σημαντικότητας $\alpha = 0.05$, γιατί η επεξηγηματική μεταβλητή `Clinicg1t` είναι μη στατιστικά σημαντική. Η επεξηγηματική μεταβλητή `Clinicg1t`, αντιστοιχεί στη μεταβλητή `clinic` για $t < 365$ μέρες. Για τον έλεγχο της υπόθεσης $H_0 : \delta_1 = 0$ έχουμε p -τιμή = 0.0720 > 0.05, που σημαίνει ότι σε επίπεδο σημαντικότητας $\alpha = 0.05$ η $H_0 : \delta_1 = 0$ δεν απορρίπτεται.
- Το 95% διάστημα εμπιστοσύνης για την επίδραση της μεταβλητής `clinic` στον χρόνο ζωής και για $t < 365$ είναι από 0.9598336 ως 2.6110221, περιέχει την τιμή 1 που αντιστοιχεί στο ότι οι δύο κλινικές επιδρούν το ίδιο στον χρόνο ζωής.
- Για $t \geq 365$ μέρες το εκτιμώμενο HR της συµμεταβλητής `clinic` ισούται με 6.237113. Η μεταβλητή `Clinicg2t`, η οποία αντιπροσωπεύει τη μεταβλητή `clinic` για $t \geq 365$ μέρες, είναι στατιστικά σημαντική σε επίπεδο σημαντικότητας $\alpha = 0.05$. Για τον έλεγχο της υπόθεσης $H_0 : \delta_2 = 0$ έχουμε p -τιμή < 0.05, συνεπώς η $H_0 : \delta_2 = 0$ απορρίπτεται.

- Το 95% διάστημα εμπιστοσύνης για την επίδραση της μεταβλητής `clinic` στον χρόνο ζωής και για $t \geq 365$ είναι από 2.927202 ως 13.289681. Το μεγάλο εύρος του διαστήματος δηλώνει την έλλειψη ακρίβειας στην τιμή του HR για $t \geq 365$ μέρες.

Συμπέρασμα:

Τα παραπάνω αποτελέσματα υποστηρίζουν ότι παρατηρήσαμε από τη γραφική παράσταση των συναρτήσεων επιβίωσης για τις κατηγορίες της μεταβλητής `clinic` και με προσαρμοσμένες τις μεταβλητές `prison` και `dose` στο μοντέλο. Δηλαδή, ότι υπάρχει μια μεγάλη διαφορά στην επίδραση που έχει η κάθε κλινική για χρόνους $t \geq 365$ μέρες, σε αντίθεση με τη μικρή διαφορά για $t < 365$ μέρες, με την κλινική 2 να διατηρεί περισσότερο χρονικό διάστημα τους ασθενείς της σε σχέση την κλινική 1 σε κάθε χρονική στιγμή.

Παρουσιάζουμε μία ακόμη προσέγγιση, όπου γίνεται χρήση χρονικά εξαρτημένων μεταβλητών. Αυτή η προσέγγιση λαμβάνει υπόψη ότι στη γραφική παράσταση των συναρτήσεων επιβίωσης για τις κατηγορίες της μεταβλητής `clinic`, με προσαρμοσμένες τις μεταβλητές `prison` και `dose`, οι δύο συναρτήσεις συνεχίζουν να αποκλίνουν η μια από την άλλη ακόμη και μετά την πάροδο ενός χρόνου. Είναι εύλογο να θεωρήσουμε την επέκταση του μοντέλου του Cox που επιτρέπει αυτήν την απόκλιση, αντί για ένα μοντέλο που θεωρεί ότι το HR είναι σταθερό πριν και μετά τον ένα χρόνο. Ένας τρόπος να ορίσουμε την επέκταση του μοντέλου του Cox που επιτρέπει αυτή την απόκλιση είναι ο εξής:

$$h\{t \mid \mathbf{X}(t)\} = h_0(t) \exp(\beta_1 \text{clinic} + \beta_2 \text{prison} + \beta_3 \text{dose} + \delta \text{clinic} \times t).$$

Το συγκεκριμένο μοντέλο περιλαμβάνει και την μεταβλητή `clinic` και το γινόμενο της με τον χρόνο, `clinic`× t . Συμπεριλαμβάνοντας τον όρο του γινομένου μπορούμε να εκτιμήσουμε την επίδραση της μεταβλητής `clinic` στον χρόνο και έτσι και το HR για οποιαδήποτε συγκεκριμένη χρονική στιγμή. Για να παρουσιάσουμε πως το HR αλλάζει με τον χρόνο σε αυτό το μοντέλο θεωρούμε το μοντέλο και τα αντίστοιχα HR για διάφορες χρονικές στιγμές. Έστω ότι ενδιαφερόμαστε για την επίδραση της μεταβλητής `clinic` στον χρόνο επιβίωσης την ημέρα $t = 91$.

$$h\{t \mid \mathbf{X}(t)\} = h_0(t) \exp(\beta_1 \text{clinic} + \beta_2 \text{prison} + \beta_3 \text{dose} + \delta \text{clinic} \times 91).$$

Το εκτιμώμενο HR των ατόμων που ανήκουν στην κλινική 1 προς των ατόμων που ανήκουν στην κλινική 2 και έχουν τις ίδιες τιμές για τις συμμεταβλητές `prison` και `dose` για $t = 91$ είναι:

$$\begin{aligned}\widehat{HR}(91) &= \frac{h_0(91) \exp\{\hat{\beta}_1 \times 1 + \hat{\beta}_2 \mathbf{prison} + \hat{\beta}_3 \mathbf{dose} + \hat{\delta}(1 \times 91)\}}{h_0(91) \exp\{\hat{\beta}_1 \times 0 + \hat{\beta}_2 \mathbf{prison} + \hat{\beta}_3 \mathbf{dose} + \hat{\delta}(0 \times 91)\}} \\ &= \exp(\hat{\beta}_1 + 91\hat{\delta}).\end{aligned}$$

Ομοίως, για $t = 274$ και για $t = 639$ θα έχουμε αντίστοιχα

$$\widehat{HR}(74) = \exp(\hat{\beta}_1 + 274\hat{\delta}), \quad \widehat{HR}(639) = \exp(\hat{\beta}_1 + 639\hat{\delta}).$$

Παρατηρούμε ότι το HR εξαρτάται από τους εκτιμητές $\hat{\beta}_1$, $\hat{\delta}$ και αν $\hat{\delta} > 0$ το εκτιμώμενο HR αυξάνεται με τον χρόνο.

ΚΕΦΑΛΑΙΟ 4

ΜΟΝΤΕΛΑ ΕΠΙΤΑΧΥΝΟΜΕΝΟΥ ΧΡΟΝΟΥ ΑΠΟΤΥΧΙΑΣ

Τα μοντέλα επιταχυνόμενου χρόνου αποτυχίας (Accelerated failure time A.F.-T.) είναι πολύ διαδεδομένα παραμετρικά μοντέλα που λειτουργούν με διαφορετικό τρόπο από το μοντέλο του Cox. Το χαρακτηριστικό των A.F.T. μοντέλων είναι ότι οι συμμεταβλητές δρουν έτσι ώστε να επιταχύνουν ή να επιβραδύνουν τον χρόνο μέχρι να συμβεί το υπό μελέτη γεγονός, δηλαδή μειώνουν ή αυξάνουν τον χρόνο μέχρι να συμβεί το υπό μελέτη γεγονός. Οι συμμεταβλητές μπορεί να εξαρτώνται από τον χρόνο. Τα συγκεκριμένα μοντέλα παρουσιάζονται αναλυτικά στη συνέχεια. Το παρόν κεφάλαιο βασίστηκε κυρίως στα συγγράμματα των Bagdonavicius and Nikulin (2001) και Lawless (2003).

4.1 Ορισμός των A.F.T. μοντέλων

Αρχικά για το διάνυσμα των συμμεταβλητών χρησιμοποιούμε τον συμβολισμό $\mathbf{X}(t) = \{X_0(t), X_1(t), X_2(t), \dots, X_m(t)\}'$ όταν εξαρτάται από τον χρόνο και τον συμβολισμό $\mathbf{X} = (X_0, X_1, X_2, \dots, X_m)'$ όταν δεν εξαρτάται από τον χρόνο. Θεωρούμε ότι $X_0(t) \equiv 1$ ή αντίστοιχα $X_0 \equiv 1$ για να έχουμε σταθερό όρο στο διάνυσμα των άγνωστων παραμέτρων, β_0 . Αν δεν θέλουμε σταθερό όρο αγνοούμε την X_0 . Η συνάρτηση επιβίωσης για την μεταβλητή του χρόνου ζωής, δοθέντος του διανύσματος $\mathbf{X}(t)$ θα είναι της μορφής

$$S\{t|\mathbf{X}(t)\} = S_0 \left[\int_0^t r\{\mathbf{X}(u)\} du \right], \quad (4.1.1)$$

όπου η $S_0(t)$ θεωρούμε ότι ανήκει σε μια προσδιορισμένη κλάση συναρτήσεων επιβίωσης. Αν οι επεξηγηματικές μεταβλητές είναι σταθερές με τον χρόνο τότε

η σχέση (4.1.1) γράφεται ως:

$$S(t|\mathbf{X}) = S_0\{r(\mathbf{X})t\}. \quad (4.1.2)$$

Η $r(\mathbf{X})$ παραμετροποιείται ως:

$$r(\mathbf{X}) = \exp\{-\boldsymbol{\beta}'\mathbf{Z}(\mathbf{X})\},$$

με $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_m)'$ να είναι το διάνυσμα των άγνωστων παραμέτρων και $\mathbf{Z}(\mathbf{X}) = \{Z_0(\mathbf{X}), Z_1(\mathbf{X}), \dots, Z_m(\mathbf{X})\}' = \{\varphi_0(\mathbf{X}), \varphi_1(\mathbf{X}), \dots, \varphi_m(\mathbf{X})\}'$ το διάνυσμα κάποιων ορισμένων συναρτήσεων φ_i , με $\varphi_0(\mathbf{X}) \equiv 1$.

Συνεπώς, υπό το παραμετροποιημένο A.F.T. μοντέλο η συνάρτηση επιβίωσης για τη μεταβλητή του χρόνου ζωής, δοθέντος του διανύσματος $\mathbf{X}(t)$ και με $r\{\mathbf{X}(t)\} = \exp\{-\boldsymbol{\beta}'\mathbf{X}(t)\}$ είναι:

$$S\{t|\mathbf{X}(t)\} = S_0 \left[\int_0^t \exp\{-\boldsymbol{\beta}'\mathbf{X}(u)\} du \right]. \quad (4.1.3)$$

Τα $X_j(t)$ $j = 1, \dots, m$ δεν είναι απαραίτητα οι παρατηρούμενες επεξηγηματικές μεταβλητές, μπορεί να είναι κάποιες συναρτήσεις τους $\varphi_j\{\mathbf{X}(t)\}$, ωστόσο χρησιμοποιούμε τον ίδιο συμβολισμό τόσο για $X_j(t)$ όσο και για $\varphi_j\{\mathbf{X}(t)\}$.

Αν οι επεξηγηματικές μεταβλητές είναι σταθερές με τον χρόνο τότε η σχέση (4.1.3) γράφεται ως:

$$S(t|\mathbf{X}) = S_0\{\exp(-\boldsymbol{\beta}'\mathbf{X})t\}, \quad (4.1.4)$$

οπότε, ο λογάριθμος της μεταβλητής του χρόνου ζωής δοθέντος του διανύσματος $\mathbf{X}$ γράφεται ως:

$$\log(T|\mathbf{X}) = \boldsymbol{\beta}'\mathbf{X} + \varepsilon,$$

με τη συνάρτηση επιβίωσης της τυχαίας μεταβλητής ε να μην εξαρτάται από το διάνυσμα $\mathbf{X}$ και $S(t) = S_0\{\log(t)\}$. Στην περίπτωση που ο λογάριθμος της μεταβλητής του χρόνου ζωής έχει κανονική κατανομή, η κατανομή της μεταβλητής ε είναι κανονική (standard multiple linear regression model). Οι δυνατές επιλογές των συναρτήσεων φ_j , $j = 1, \dots, m$, που μπορούμε να συναντήσουμε στην πράξη παρουσιάζονται στη συνέχεια.

4.2 Επιλογή συναρτήσεων των συµµεταβλητών

Αρχικά θεωρούμε ότι οι επεξηγηματικές μεταβλητές είναι συνεχείς. Έστω ακόμη ότι οι επεξηγηματικές μεταβλητές είναι σταθερές με τον χρόνο, άρα ισχύει

το μοντέλο (4.1.2) και έστω ότι ισχύει στο σύνολο E_0 . Τότε για κάθε $\mathbf{X}_i, \mathbf{X}_j \in E_0$, $i, j \in \{1, \dots, m\}$,

$$S(t|\mathbf{X}_j) = S\{\rho(\mathbf{X}_i, \mathbf{X}_j)t|\mathbf{X}_i\}, \quad (4.2.1)$$

με τη συνάρτηση $\rho(\mathbf{X}_i, \mathbf{X}_j) = r(\mathbf{X}_j)/r(\mathbf{X}_i)$ να υποδηλώνει την αλλαγή του ρυθμού του κινδύνου όταν μεταβούμε από την $\mathbf{X}_i$ στην $\mathbf{X}_j$. Δηλαδή, εκφράζει κατά πόσο ο χρόνος επιβίωσης επηρεάζεται από την αλλαγή των συμμεταβλητών, από τη μετάβαση από την $\mathbf{X}_i$ στην $\mathbf{X}_j$. Υποθέτουμε αρχικά ότι το διάνυσμα των συμμεταβλητών είναι μονοδιάστατο, δηλαδή $\mathbf{X} = X$. Ο ρυθμός με τον οποίο μεταβάλλεται ο κίνδυνος να συμβεί το υπό μελέτη γεγονός ορίζεται με τη συνάρτηση (infinitesimal characteristic)

$$\delta(X) = \lim_{\Delta X \rightarrow 0} \frac{\rho(X, X + \Delta X) - \rho(X, X)}{\Delta X} = \log\{r(X)\}. \quad (4.2.2)$$

Δηλαδή, η ταχύτητα με την οποία αλλάζει ο κίνδυνος με την αλλαγή της τιμής της συμμεταβλητής X ισούται με $\log\{r(X)\}$. Για κάθε $X \in E_0$ η συνάρτηση $r(X)$ δίνεται από τη σχέση

$$r(X) = r(X_0) \exp \left\{ \int_{X_0}^X \delta(u) du \right\}, \quad (4.2.3)$$

όπου $X_0 \in E_0$ είναι μια σταθερή προκαθορισμένη μεταβλητή. Η σχέση αυτή εκφράζει πως μεταβάλλεται ο ρυθμός κινδύνου καθώς μεταβάλλεται η τιμή της μεταβλητής X ξεκινώντας από την προκαθορισμένη τιμή $X_0 \in E_0$. Υποθέτουμε ότι η $\delta(X)$ είναι ανάλογη με μία προκαθορισμένη συνάρτηση $u(X)$, δηλαδή

$$\delta(X) = \alpha u(X).$$

Σε αυτήν την περίπτωση

$$r(X) = \exp\{-\beta_0 - \beta_1 \varphi_1(X)\}, \quad (4.2.4)$$

η $\varphi_1(X)$ είναι η αρχική της $u(X)$ και β_0, β_1 είναι άγνωστες παράμετροι. Παραθέτουμε κάποια παραδείγματα για διάφορες συναρτήσεις $\varphi_1(\mathbf{X})$ που μπορούμε να εφαρμόσουμε στη σχέση (4.2.4).

Παράδειγμα 1:

Έστω $\delta(X) = \alpha$, $\varphi_1(X) = X$. Ο ρυθμός αλλαγής της κλίμακας είναι σταθερός, καθώς $\delta(X) = \alpha$. Τότε

$$r(X) = \exp(-\beta_0 - \beta_1 X).$$

Το µοντέλο αυτό ονοµάζεται log-linear µοντέλο, γιατί ο λογάριθµος του είναι µια γραµµική συνάρτηση του X .

Παράδειγμα 2:

Έστω $\delta(X) = \alpha/X$, $\varphi_1(X) = \log X$. Τότε

$$r(X) = \exp(-\beta_0 - \beta_1 \log X) = \alpha_1 X^{\beta_1}.$$

Το µοντέλο αυτό ονοµάζεται power-rule µοντέλο.

Παράδειγμα 3:

Έστω $\delta(X) = \alpha/X^2$, $\varphi_1(X) = -1/X$. Τότε

$$r(X) = \exp\left(-\beta_0 - \frac{\beta_1}{X}\right) = \alpha_1 \exp\left(-\frac{\beta_1}{X}\right).$$

Το µοντέλο αυτό είναι γνωστό ως µοντέλο Arrhenius.

Παράδειγμα 4:

Έστω $\delta(X) = \alpha/X(1-X)$, $\varphi_1(X) = \log(\frac{X}{1-X})$. Τότε

$$r(X) = \exp\left(-\beta_0 - \beta_1 \frac{X}{1-X}\right) = \alpha_1 \left(\frac{X}{1-X}\right)^{-\beta_1}, \quad 0 < X < 1,$$

που είναι το Meeker-Luvalle µοντέλο.

Το µοντέλο Arrhenius χρησιµοποιείται για τη µοντελοποίηση χρόνου ζωής προϊόντων όταν η επεξηγηµατική µεταβλητή είναι η θερµοκρασία, το power-rule µοντέλο χρησιµοποιείται όταν η επεξηγηµατική µεταβλητή είναι η τάση του ρεύµατος (ή µηχανικές φορτίσεις), το log-linear µοντέλο εφαρµόζεται στην ανάλυση δεδοµένων αντοχής και εξάντλησης, κατά τον έλεγχο ηλεκτρονικών εξαρτηµάτων. Το µοντέλο Meeker-Luvalle χρησιµοποιείται όταν η επεξηγηµατική µεταβλητή είναι το ποσοστό υγρασίας. Αν δεν είναι ξεκάθαρο ποιο από τα τρία πρώτα µοντέλα να επιλεγεί, µπορεί να επιλεχθεί µια µεγαλύτερη κλάση µοντέλων. Για παράδειγμα, αυτά τα µοντέλα είναι ειδικές περιπτώσεις της κλάσης µοντέλων που ορίζεται από την

$$\delta(X) = \begin{cases} \exp\left(-\beta_0 - \beta_1 \frac{X^\varepsilon - 1}{\varepsilon}\right), & \text{αν } \varepsilon \neq 0, \\ \exp(-\beta_0 - \beta_1 \log X), & \text{αν } \varepsilon = 0. \end{cases}$$

Σε αυτήν την περίπτωση η παράµετρος ε θα πρέπει να εκτιµηθεί.

Το µοντέλο (4.2.1) µπορεί να γενικευθεί. Για παράδειγμα, µπορεί να γίνει η υπόθεση ότι η $\delta(X)$ είναι γραμμικός συνδυασµός κάποιων συγκεκριµένων συναρτήσεων της επεξηγηµατικής µεταβλητής, δηλαδή είναι της µορφής

$$\delta(X) = \sum_{i=1}^k a_i u_i(X).$$

Σε αυτήν την περίπτωση

$$r(X) = \exp\{-\beta_0 - \sum_{i=1}^k \beta_i z_i(X)\},$$

µε z_i να είναι συγκεκριµένες συναρτήσεις της επεξηγηµατικής µεταβλητής και $\beta_0, \beta_1, \dots, \beta_k$ είναι άγνωστες (πιθανώς όχι όλες) παράµετροι.

Παράδειγμα 5:

Έστω $\delta(X) = 1/X + \alpha/X^2$. Τότε

$$r(X) = \exp\left(-\beta_0 - \beta_1 \log X - \frac{\beta_2}{X}\right) = \alpha_1 X \exp\left(-\frac{\beta_2}{X}\right),$$

που είναι το Eyring µοντέλο και χρησιµοποιείται όταν η επεξηγηµατική µεταβλητή είναι η θερμοκρασία.

Παράδειγμα 6:

Έστω $\delta(X) = \sum_{i=1}^k \frac{a_i}{X^i}$. Τότε

$$r(X) = \exp\left(-\beta_0 - \beta_1 \log X - \sum_{i=1}^{k-1} \frac{\beta_i}{X^i}\right).$$

Αυτό είναι το γενικευµένο Eyring µοντέλο.

Υποθέτουμε ότι έχουµε παραπάνω από μία επεξηγηµατική µεταβλητή, δηλαδή $\mathbf{X} = (X_1, X_2, \dots, X_m)$. Αν δεν υπάρχουν αλληλεπιδράσεις µεταξύ των συµμεταβλητών µπορούµε να χρησιµοποιήσουµε

$$r(\mathbf{X}) = \exp\left\{-\beta_0 - \sum_{i=1}^m \sum_{j=1}^{k_i} \beta_{ij} z_{ij}(X_i)\right\},$$

µε $z_{ij}(X_i)$ να είναι συγκεκριµένες συναρτήσεις και β_{ij} οι άγνωστες παράµετροι, $i = 1, \dots, m$, $j = 1, \dots, k_i$, k_i είναι το πλήθος των συναρτήσεων $z_{ij}(X_i)$ που σχετίζονται µε την επεξηγηµατική µεταβλητή X_i .

Παράδειγμα 7:

Αν η επιρροή της πρώτης συµµεταβλητής εκφράζεται από το power-rule μοντέλο και η επιρροή της δεύτερης από το Arrhenius έχουμε το μοντέλο

$$r(X_1, X_2) = \exp \left(-\beta_0 - \beta_1 \log X_1 - \frac{\beta_2}{X_2} \right).$$

Έτσι σε αυτήν την περίπτωση $k_1 = k_2 = 1$. Αν υπάρχουν αλληλεπιδράσεις μεταξύ των συµµεταβλητών επιπλέον όροι πρέπει να ληφθούν υπόψιν.

Παράδειγμα 8:

Αν στο αμέσως προηγούμενο παράδειγμα θεωρήσουμε ότι υπάρχει αλληλεπίδραση μεταξύ των μεταβλητών X_1, X_2 , τότε θα μπορούσαμε να θεωρήσουμε το μοντέλο

$$r(X_1, X_2) = \exp \left\{ -\beta_0 - \beta_1 \log X_1 - \frac{\beta_2}{X_2} - \beta_3 \frac{(\log X_1)}{X_2} \right\}.$$

Αν θεωρήσουμε ότι οι επεξηγηματικές μεταβλητές είναι διακριτές, η μορφή των συναρτήσεων είναι ίδια με αυτή στην περίπτωση που οι επεξηγηματικές μεταβλητές είναι συνεχείς, για παράδειγμα η φ_j ίσως είναι $\varphi_j(\mathbf{X}) = \mathbf{X}, \log(\mathbf{X})$ ή $\frac{1}{\mathbf{X}}$.

Έστω τώρα ότι η j -οστή επεξηγηματική μεταβλητή είναι κατηγορική και έχει k_j το πλήθος δυνατές τιμές. Η $X_j = \mathbf{X}_j$ θεωρείται ως ένα $(k_j - 1)$ -διάστατο διάνυσμα, καθώς αναπαριστούμε τις κατηγορίες της με $(k_j - 1)$ δείκτριες μεταβλητές, απορρίπτοντας μια κατηγορία την οποία θεωρούμε βάση αναφοράς.

$$X_j = (X_{j(1)}, X_{j(2)}, \dots, X_{j(k_j-1)})',$$

με k_j το πλήθος δυνατές τιμές, που είναι οι εξής:

$$(0, 0, \dots, 0)', (1, 0, \dots, 0)', (0, 1, \dots, 0)', \dots, (0, 0, \dots, 1)'.$$

Έτσι αν η j -οστή επεξηγηματική μεταβλητή είναι κατηγορική και οι υπόλοιπες συνεχείς ή διακριτές

$$\beta' \mathbf{X} = \beta_0 + \beta_1 X_1 + \dots + \beta_{j-1} X_{j-1} + \sum_{l=1}^{k_j-1} \beta_{jl} X_{j(l)} + \beta_{j+1} X_{j+1} + \dots + \beta_m X_m. \quad (4.2.5)$$

Το παρατηρούμενο μοντέλο είναι ισοδύναμο με αυτό της σχέσης (4.1.4) με $m + k_j - 1$ μονοδιάστατες επεξηγηματικές μεταβλητές. Αν $k_j = 2$ η επεξηγηματική μεταβλητή είναι διχοτομική με δυνατές τιμές 0 ή 1. Στην επόμενη υποενότητα παραθέτουμε την ερμηνεία των συντελεστών παλινδρόμησης.

4.3 Ερμηνεία των συντελεστών παλινδρόμησης

Υποθέτουμε ότι οι επεξηγηματικές μεταβλητές είναι σταθερές με τον χρόνο. Θα δώσουμε τον ορισμό του p -ποσοστημορίου της μεταβλητής του χρόνου ζωής, T , δοθέντος $\mathbf{X}$ και του μέσου χρόνου ζωής δοθέντος του διανύσματος των συμμεταβλητών $\mathbf{X}$. Οι ποσότητες αυτές σχετίζονται και ο ορισμός του μέσου χρόνου ζωής μας βοηθά να ερμηνεύσουμε τους συντελεστές παλινδρόμησης. Υπό το A.F.T. μοντέλο της σχέσης (4.1.4) το p -ποσοστημόριο της μεταβλητής του χρόνου ζωής, T , δοθέντος $\mathbf{X}$ είναι:

$$t_p(\mathbf{X}) = \exp(\beta' \mathbf{X}) S_0^{-1}(1 - p),$$

και ο λογάριθμος του θα είναι:

$$\log\{t_p(\mathbf{X})\} = \beta' \mathbf{X} + c_p,$$

που είναι γραμμική συνάρτηση των παραμέτρων παλινδρόμησης, όπου $c_p = \log\{S_0^{-1}(1 - p)\}$. Αυτό συμβαίνει καθώς για το p -ποσοστημόριο ισχύει ότι $P(T > t_p) = 1 - p$, οπότε

$$\begin{aligned} P(T > t_p | \mathbf{X}) &= 1 - p \\ S(t_p | \mathbf{X}) &= 1 - p \\ S_0\{\exp(-\beta' \mathbf{X}) t_p\} &= 1 - p \\ \exp(-\beta' \mathbf{X}) t_p &= S_0^{-1}(1 - p) \\ t_p(\mathbf{X}) &= \exp(\beta' \mathbf{X}) S_0^{-1}(1 - p). \end{aligned}$$

Έστω $m(\mathbf{X})$ ο μέσος χρόνος ζωής δοθέντος του διανύσματος των συμμεταβλητών $\mathbf{X}$. Τότε

$$m(\mathbf{X}) = \exp(\beta' \mathbf{X}) \int_0^\infty S_0(u) du, \quad (4.3.1)$$

και ο λογάριθμος του είναι:

$$\log\{m(\mathbf{X})\} = \beta' \mathbf{X} + c, \quad (4.3.2)$$

που είναι επίσης μια γραμμική συνάρτηση των παραμέτρων παλινδρόμησης, με $c = \log\{\int_0^\infty S_0(u) du\}$. Συμβολίζουμε με

$$MR(\mathbf{X}, \mathbf{Y}) = \frac{m(\mathbf{Y})}{m(\mathbf{X})}, \quad QR(\mathbf{X}, \mathbf{Y}) = \frac{t_p(\mathbf{Y})}{t_p(\mathbf{X})}$$

τον λόγο των μέσων τιμών του χρόνου ζωής και τον λόγο των ποσοστημορίων, για δύο διανύσματα συμμεταβλητών $\mathbf{X}, \mathbf{Y}$, αντίστοιχα. Για το A.F.T. μοντέλο

$$MR(\mathbf{X}, \mathbf{Y}) = QR(\mathbf{X}, \mathbf{Y}) = \exp\{\boldsymbol{\beta}'(\mathbf{Y} - \mathbf{X})\}. \quad (4.3.3)$$

Επομένως, $\exp\{\boldsymbol{\beta}'(\mathbf{Y} - \mathbf{X})\}$ είναι ο λόγος των μέσων τιμών του χρόνου ζωής, που αντιστοιχούν στα διανύσματα συμμεταβλητών $\mathbf{X}$ και $\mathbf{Y}$. Θα δώσουμε την ερμηνεία των παραμέτρων β_j υπό το μοντέλο της σχέσης (4.1.4). Παρακάτω παρατίθεται η ερμηνεία των παραμέτρων β_j , αρχικά ανάλογα με το αν υπάρχει ή όχι αλληλεπίδραση μεταξύ των επεξηγηματικών συμμεταβλητών και έπειτα, για την περίπτωση που υπάρχει αλληλεπίδραση, ανάλογα με το είδος των συμμεταβλητών.

Πρώτα θα παρουσιάσουμε την περίπτωση που έχουμε μοντέλο χωρίς αλληλεπιδράσεις. Ξεκινάμε με την περίπτωση που το μοντέλο περιέχει συνεχείς ή διακριτές επεξηγηματικές μεταβλητές. Υποθέτουμε ότι η j -οστή επεξηγηματική μεταβλητή X_j είναι συνεχής ή διακριτή. Η ποσότητα

$$\exp(\beta_j) = \frac{\exp\{\beta_1 X_1 + \cdots + \beta_j(X_j + 1) + \cdots + \beta_m X_m\}}{\exp(\beta_1 X_1 + \cdots + \beta_j X_j + \cdots + \beta_m X_m)} = MR_j$$

είναι ο λόγος των μέσων τιμών που αντιστοιχούν σε αλλαγή της X_j κατά μια μονάδα, όσο οι υπόλοιπες συμμεταβλητές παραμένουν σταθερές.

Έστω τώρα ότι το μοντέλο, χωρίς αλληλεπιδράσεις, περιλαμβάνει κατηγορικές επεξηγηματικές μεταβλητές. Υποθέτουμε ότι η $X_j = \mathbf{X}_j = (X_{j(1)}, \dots, X_{j(k_j-1)})'$ είναι κατηγορική, οπότε θα έχει αντίστοιχο διάνυσμα παραμέτρων $\boldsymbol{\beta}_j = (\beta_{j1}, \dots, \beta_{ji}, \dots, \beta_{j(k_j-1)})'$. Η πρώτη της τιμή είναι $(0, 0, \dots, 0)'$ και η $(i+1)$ τιμή είναι $(0, \dots, 0, 1, 0, \dots, 0)'$ με μονάδα στην i -οστή θέση. Η ποσότητα

$$\exp(\beta_{ji}) = \frac{\exp(\beta_1 X_1 + \cdots + \beta_{ji} \times 1 + \cdots + \beta_m X_m)}{\exp(\beta_1 X_1 + \cdots + \beta_{ji} \times 0 + \cdots + \beta_m X_m)} = MR_{ji}$$

είναι ο λόγος των μέσων τιμών που αντιστοιχούν σε αλλαγή της X_j από την πρώτη δυνατή τιμή της, στην $(i+1)$ -οστή δυνατή τιμή της, όσο οι υπόλοιπες συμμεταβλητές παραμένουν σταθερές.

Συνεχίζουμε με την περίπτωση που έχουμε μοντέλο με αλληλεπιδράσεις. Αν η επιρροή της επεξηγηματικής μεταβλητής στον μέσο χρόνο ζωής είναι διαφορετική για διάφορες τιμές άλλων επεξηγηματικών μεταβλητών, τότε υπάρχει αλληλεπίδραση μεταξύ των επεξηγηματικών μεταβλητών και το μοντέλο πρέπει να τροποποιηθεί. Αρχικά, παρουσιάζουμε την περίπτωση που υπάρχει αλληλεπίδραση μεταξύ συνεχών ή διακριτών επεξηγηματικών μεταβλητών. Αν υπάρχουν δύο

συνεχείς ή διακριτές επεξηγηματικές μεταβλητές που αλληλεπιδρούν, έστω οι X_1 και X_2 , τότε

$$\beta' \mathbf{X} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2.$$

Για τρεις επεξηγηματικές μεταβλητές, έστω τις X_1 , X_2 και X_3 , έχουμε

$$\beta' \mathbf{X} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \beta_6 X_2 X_3 + \beta_7 X_1 X_2 X_3.$$

Ομοίως και για περισσότερες μεταβλητές. Στην περίπτωση των δύο επεξηγηματικών μεταβλητών ο λόγος των μέσων τιμών

$$MR_2(X_1) = \frac{m(X_1, X_2 + 1)}{m(X_1, X_2)} = \exp(\beta_2 + \beta_3 X_1),$$

εξαρτάται από την τιμή της X_1 . Οπότε

$$\exp(\beta_2 + \beta_3 X_1)$$

είναι ο λόγος των μέσων τιμών που αντιστοιχούν σε αλλαγή της X_2 κατά μία μονάδα όταν η άλλη επεξηγηματική μεταβλητή με την οποία υπάρχει αλληλεπίδραση είναι σταθερή, με τιμή X_1 . Συνεχίζουμε με την περίπτωση που υπάρχει αλληλεπίδραση μεταξύ συνεχών ή διακριτών και κατηγορικών επεξηγηματικών μεταβλητών. Υποθέτουμε ότι έχουμε δύο επεξηγηματικές μεταβλητές, έστω X_1 , X_2 , η X_1 είναι συνεχής ή διακριτή και η X_2 είναι κατηγορική με k_2 το πλήθος δυνατές τιμές. Τότε

$$\beta' \mathbf{X} = \beta_1 X_1 + \sum_{i=1}^{k_2-1} \beta_{2i} X_{2(i)} + \sum_{i=1}^{k_2-1} \beta_{12i} X_1 X_{2(i)}.$$

Ο λόγος

$$MR_{2i}(X_1) = \frac{\exp(\beta_1 X_1 + \beta_{2i} \times 1 + \beta_{12i} X_1 \times 1)}{\exp(\beta_1 X_1 + \beta_{2i} \times 0 + \beta_{12i} X_1 \times 0)} = \exp(\beta_{2i} + \beta_{12i} X_1),$$

εξαρτάται από την τιμή X_1 . Σε αυτό το παράδειγμα $\exp(\beta_{2i} + \beta_{12i} X_1)$ είναι ο λόγος των μέσων που αντιστοιχεί στην X_2 να λαμβάνει την πρώτη δυνατή τιμή της ως προς το να λαμβάνει την $(i+1)$, $i = 1, \dots, k_2 - 1$ δυνατή τιμή της, όταν η άλλη επεξηγηματική μεταβλητή είναι σταθερή και ίση με X_1 . Η πρώτη δυνατή τιμή της X_2 είναι η $(0, \dots, 0, 0, 0, \dots, 0)'$, ενώ η $(i+1)$ δυνατή τιμή της X_2 είναι η $(0, \dots, 0, 1, 0, \dots, 0)'$, όπου η μονάδα είναι στην i θέση, $i = 1, \dots, k_2 - 1$.

Ας εξετάσουμε τώρα την περίπτωση όπου υπάρχει αλληλεπίδραση μεταξύ κατηγορικών επεξηγηματικών μεταβλητών. Υποθέτουμε ότι τόσο η X_1 όσο και η X_2

είναι κατηγορικές με τρεις πιθανές τιμές η κάθε μια, τις οποίες αναπαριστούμε, όπως και προηγουμένως, με 2 δείκτριες. Έτσι οι

$$\mathbf{X}_1 = (X_{11}, X_{12})', \quad \mathbf{X}_2 = (X_{21}, X_{22})'.$$

έχει η κάθε μία τις εξής δυνατές τιμές: $(0, 0)'$, $(1, 0)'$, $(0, 1)'$. Ακόμη, έχουμε ότι

$$\begin{aligned} \beta' \mathbf{X} &= \beta_{11}X_{11} + \beta_{12}X_{12} + \beta_{21}X_{21} + \beta_{22}X_{22} \\ &+ \beta_{1121}X_{11}X_{21} + \beta_{1122}X_{11}X_{22} + \beta_{1221}X_{12}X_{21} + \beta_{1222}X_{12}X_{22}. \end{aligned}$$

Σε αυτήν την περίπτωση, ο λόγος

$$\begin{aligned} MR_{22}(X_1) &= \frac{\exp(\beta'_1 X_1 + \beta_{21} \times 1 + \beta_{1121}X_{11} \times 1 + \beta_{1221}X_{12} \times 1)}{\exp(\beta'_1 X_1)} \\ &= \exp(\beta_{21} + \beta_{1121}X_{11} + \beta_{1221}X_{12}), \end{aligned}$$

είναι ο λόγος για την περίπτωση που η X_1 είναι σταθερή και η X_2 έχει στον αριθμητή την τιμή $(1, 0)'$, ενώ στον παρονομαστή $(0, 0)'$. Ο λόγος αυτός εξαρτάται από την τιμή της $X_1 = (X_{11}, X_{12})'$, και της παραμέτρου $\beta_1 = (\beta_{11}, \beta_{12})'$. Συνεπώς, $\exp(\beta_{21} + \beta_{1121}X_{11} + \beta_{1221}X_{12})$ είναι ο λόγος των μέσων τιμών που αντιστοιχεί σε αλλαγή της X_2 από την πρώτη στην δεύτερη τιμή της, όταν η επεξηγηματική μεταβλητή $\mathbf{X}_1 = (X_{11}, X_{12})'$ παραμένει σταθερή. Η γενίκευση αυτής της διαδικασίας είναι ανάλογη για περιπτώσεις όπου οι επεξηγηματικές μεταβλητές έχουν τρεις ή περισσότερες δυνατές τιμές. Παρακάτω αναφέρουμε συνοπτικά τη μορφή που παίρνει το A.F.T. μοντέλο στην περίπτωση που το διάνυσμα των παραμέτρων, $\beta = (\beta_0, \beta_1, \dots, \beta_m)'$, εξαρτάται από τον χρόνο.

Θεωρούμε το A.F.T. μοντέλο με επεξηγηματικές μεταβλητές που εξαρτώνται από τον χρόνο. Επιπλέον, υποθέτουμε ότι το διάνυσμα των άγνωστων παραμέτρων $\beta = (\beta_0, \beta_1, \dots, \beta_m)'$, εξαρτάται από τον χρόνο. Το A.F.T. μοντέλο μπορεί να παραμετροποιηθεί ως εξής:

$$S\{t|\mathbf{X}(t)\} = S_0 \left[\int_0^t \exp\{-\beta'(u)\mathbf{X}(u)\} du \right]. \quad (4.3.4)$$

Θεωρούμε ότι οι συντελεστές $\beta_i(t)$ έχουν τη μορφή:

$$\beta_i(t) = \beta_i + \gamma_i g_i(t), \quad i = 1, \dots, m,$$

όπου $g_i(t)$ είναι κάποιες προσδιορισμένες ντετερμινιστικές συναρτήσεις ή υλοποιήσεις προβλέψιμων διαδικασιών. Σε αυτήν την περίπτωση το μοντέλο με συντελεστές παλινδρόμησης που εξαρτώνται από τον χρόνο και σταθερές συμμεταβλητές, ή συμμεταβλητές που εξαρτώνται από τον χρόνο, μπορεί να γραφτεί στη

μορφή της σχέσης (4.1.3) με διαφορετική ερμηνεία για τις επεξηγηματικές μεταβλητές. Για να το φέρουμε σε αυτήν την μορφή θέτουμε:

$$\begin{aligned}\boldsymbol{\theta} &= (\theta_0, \theta_1, \dots, \theta_{2m})' = (\beta_0, \beta_1, \dots, \beta_m, \gamma_1, \gamma_2, \dots, \gamma_m)', \\ \mathbf{z}(t) &= \{z_0(t), z_1(t), \dots, z_{2m}(t)\}' \\ &= \{1, X_1(t), \dots, X_m(t), X_1(t)g_1(t), \dots, X_m(t)g_m(t)\}'.\end{aligned}\quad (4.3.5)$$

Οπότε,

$$\boldsymbol{\beta}'(u)\mathbf{X}(u) = \beta_0 + \sum_{i=1}^m \{\beta_i + \gamma_i g_i(u)\} X_i(t) = \boldsymbol{\theta}'\mathbf{z}(u).$$

Επομένως, το A.F.T. μοντέλο μπορεί να γραφεί στη μορφή:

$$S\{t|\mathbf{z}(t)\} = S_0 \left[\int_0^t \exp\{-\boldsymbol{\theta}'\mathbf{z}(u)\} du \right]. \quad (4.3.6)$$

4.4 Shape-scale οικογένειες κατανομών

Θεωρούμε το A.F.T. μοντέλο

$$S\{t|\mathbf{X}(t)\} = S_0 \left[\int_0^t \exp\{-\boldsymbol{\beta}'\mathbf{X}(u)\} du \right], \quad (4.4.1)$$

ή το A.F.T. μοντέλο της σχέσης (4.3.4) με συντελεστές παλινδρόμησης που εξαρτώνται από τον χρόνο, και υποθέτουμε ότι η $S_0(t)$ ανήκει σε μια προσδιορισμένη shape-scale κλάση συναρτήσεων επιβίωσης. Δηλαδή είναι της μορφής

$$S_0(t) = G_0 \left\{ \left(\frac{t}{\alpha} \right)^\delta \right\}, \text{ με } (\alpha, \delta > 0).$$

Η $G_0(t)$ είναι μία καθορισμένη συνάρτηση. Για παράδειγμα, αν για $t > 0$

$$G_0(t) = \exp(-t), \quad G_0(t) = (1+t)^{-1}, \quad G_0(t) = 1 - \Phi(\log t),$$

τότε προκύπτουν οι οικογένειες κατανομών Weibull, loglogistic και lognormal, αντίστοιχα. Η Φ είναι η συνάρτηση κατανομής της τυπικής κανονικής κατανομής. Στη συνέχεια θα ασχοληθούμε με την κατανομή Weibull.

4.5 Μοντέλο A.F.T. με κατανομή Weibull

Σε αυτήν την ενότητα θεωρούμε ότι το διάνυσμα των συμμεταβλητών είναι ανεξάρτητο του χρόνου. Θεωρούμε την κατανομή Weibull με παράμετρο θέσης $\alpha > 0$ και παράμετρο σχήματος $\delta > 0$. Η συνάρτηση πυκνότητας πιθανότητας της είναι:

$$f(t, \alpha, \delta) = \frac{\delta}{\alpha} \left(\frac{t}{\alpha} \right)^{\delta-1} \exp[-(t/\alpha)^\delta], \quad t \geq 0.$$

Η συνάρτηση επιβίωσης της μεταβλητής T είναι:

$$S(t) = \exp\{-(t/\alpha)^\delta\}, \quad t \geq 0,$$

και η συνάρτηση κινδύνου είναι:

$$h(t) = \frac{\delta}{\alpha} \left(\frac{t}{\alpha} \right)^{\delta-1} = (\delta t^{\delta-1}) \alpha^{-\delta}, \quad t \geq 0.$$

4.5.1 Μοντέλα παλινδρόμησης με αναλογικότητα κινδύνων

Υπάρχουν δύο βασικές προσεγγίσεις για τη μοντελοποίηση του χρόνου ζωής με μοντέλα παλινδρόμησης. Η μία προσέγγιση χρησιμοποιεί μετασχηματισμούς του χρόνου ζωής, θεωρώντας ότι η επίδραση των συμμεταβλητών ισοδυναμεί με το να τροποποιήσουμε τον ρυθμό με τον οποίο περνά ο χρόνος, τα A.F.T. είναι τέτοια μοντέλα. Η δεύτερη προσέγγιση υιοθετεί συγκεκριμένες πληροφορίες για τον τρόπο με τον οποίο οι συμμεταβλητές επιδρούν στη συνάρτηση κινδύνου του χρόνου ζωής. Το πιο κοινό μοντέλο αυτής της κατηγορίας είναι το μοντέλο αναλογικών κινδύνων του Cox. Για το μοντέλο αναλογικών κινδύνων του Cox η συνάρτηση κινδύνου για την T δοθέντος $\mathbf{X}$ είναι της μορφής

$$h(t|\mathbf{X}) = h_0(t)w(\mathbf{X}), \quad (4.5.1)$$

με $w(\mathbf{X}), h_0(t)$ να είναι θετικά ορισμένες συναρτήσεις. Η $w(\mathbf{X})$ συνήθως ορίζεται ως $\exp(\boldsymbol{\beta}'\mathbf{X})$. Στα πλήρως παραμετρικά μοντέλα αναλογικών κινδύνων οι συναρτήσεις $h_0(t, \boldsymbol{\alpha}), w(\mathbf{X}, \boldsymbol{\beta})$ καθορίζονται παραμετρικά. Από τη σχέση (4.5.1) συνεπάγεται ότι η συνάρτηση επιβίωσης της T δοθέντος $\mathbf{X}$ είναι της μορφής

$$S(t|\mathbf{X}) = S_0(t)^{w(\mathbf{X})}, \quad (4.5.2)$$

με $S_0(t) = \exp[-H_0(t)]$ να είναι η βασική συνάρτηση επιβίωσης (baseline survival function). Για δύο διαφορετικά διανύσματα $\mathbf{X}_1, \mathbf{X}_2$ η μία συνάρτηση επιβίωσης πρέπει να είναι μεγαλύτερη από την άλλη. Ένα χαρακτηριστικό του μοντέλου αναλογικών κινδύνων είναι ότι αν η $S_0(t, \boldsymbol{\alpha})$ ανήκει σε κάποια οικογένεια

παραμετρικών μοντέλων, τότε η $S(t|\mathbf{X})$ δεν ανήκει γενικά στην ίδια οικογένεια. Αυτή είναι μια αντίθεση με τα A.F.T. μοντέλα και αυτός είναι ένας λόγος που τα πλήρως παραμετρικά μοντέλα χρησιμοποιούνται λιγότερο από τα ημιπαραμετρικά μοντέλα, στα οποία η βασική συνάρτηση κινδύνου δεν προσδιορίζεται. Ένας ακόμη λόγος είναι η ύπαρξη πιο εύκολων μεθόδων για ημιπαραμετρικά μοντέλα αναλογικών κινδύνων.

Υπάρχει περίπτωση που το μοντέλο αναλογικών κινδύνων και το A.F.T. μοντέλο ταυτίζονται; Την απόδειξη που ακολουθεί ανέπτυξε ο Rodríguez (2010) στις σημειώσεις του για παραμετρικά μοντέλα επιβίωσης. Αν ξεκινήσουμε με μια συνάρτηση κινδύνου την οποία πολλαπλασιάσουμε με ένα σχετικό ρίσκο και με μια άλλη συνάρτηση κινδύνου για την οποία αυξάνουμε (ή μειώνουμε) τον χρόνο μέχρι το υπό μελέτη γεγονός, υπάρχει περίπτωση να καταλήξουμε στην ίδια κατανομή; Δηλαδή, θέλουμε να ελέγξουμε αν η συνάρτηση κινδύνου ενός μοντέλου αναλογικών κινδύνων μπορεί να ισούται με τη συνάρτηση κινδύνου ενός A.F.T. μοντέλου. Από τη σχέση (4.1.4) για τα A.F.T. μοντέλα και με χρήση της σχέσης $h(t) = -d \ln[S(t)]/dt$ έχουμε ότι η συνάρτηση κινδύνου των A.F.T. μοντέλων μπορεί να γραφεί στη μορφή $h(t|\mathbf{X}) = h_0\{t \exp(-\beta'\mathbf{X})\} \exp(-\beta'\mathbf{X})$. Σύμφωνα με τα παραπάνω έχουμε την ακόλουθη ισότητα

$$h_0(t) \exp(\beta'\mathbf{X}) = h_0^*\{t \exp(-\beta^*\mathbf{X})\} \exp(-\beta^*\mathbf{X}), \quad (4.5.3)$$

που πρέπει να ισχύει για κάθε $\mathbf{X}, t$. Το αριστερό μέλος της ισότητας αντιστοιχεί στη συνάρτηση κινδύνου ενός μοντέλου αναλογικών κινδύνων, ενώ το δεξί μέλος της ισότητας αντιστοιχεί στη συνάρτηση κινδύνου ενός A.F.T. μοντέλου. Οι αστερίσκοι συμβολίζουν ότι δεν ξεκινάμε απαραίτητα με την ίδια βασική συνάρτηση κινδύνου ή τις ίδιες παραμέτρους. Αν η παραπάνω σχέση ισχύει για κάθε $\mathbf{X}$ θα ισχύει και για $\mathbf{X} = \mathbf{0}$, οπότε συνεπάγεται

$$h_0(t) = h_0^*(t),$$

δηλαδή οι βασικές συναρτήσεις κινδύνου πρέπει να ταυτίζονται. Ας προσπαθήσουμε να βρούμε τη βασική συνάρτηση κινδύνου και κατά συνέπεια τη συνάρτηση κινδύνου. Θεωρούμε το διάνυσμα μεταβλητών $\mathbf{X} = \left(\frac{\log t}{\beta_0^*}, 0, \dots, 0\right)'$. Οπότε, πολλαπλασιάζοντας με β^* έχουμε $\beta^*\mathbf{X} = \log t$, ενώ πολλαπλασιάζοντας με $\beta = (\beta_0, \beta_1, \dots, \beta_m)'$ έχουμε $\beta'\mathbf{X} = \log(t)\beta_0/\beta_0^*$. Αντικαθιστώντας στη σχέση

(4.5.3) έχουμε:

$$\begin{aligned} h_0(t) \exp\left(\log t \frac{\beta_0}{\beta_0^*}\right) &= h_0^*\{t \exp(-\log t)\} \exp(-\log t) \\ \implies h_0(t) t^{\frac{\beta_0}{\beta_0^*}} &= h_0^*(1) \frac{1}{t} \\ \implies h_0(t) &= h_0^*(1) t^{-\frac{\beta_0}{\beta_0^*}-1}. \end{aligned}$$

Επαναλαμβάνουμε αυτήν τη διαδικασία για διανύσματα συμμεταβλητών $\mathbf{X}$ που έχουν στην i -οστή θέση την τιμή $\frac{\log t}{\beta_i^*}$ και σε όλες τις υπόλοιπες 0, για κάθε $i = 1, \dots, m$, συνεπώς $\beta^{*\prime} \mathbf{X} = \log t$ και $\beta' \mathbf{X} = \log t \frac{\beta_i}{\beta_i^*}$. Προκύπτει $h_0(t) = h_0^*(1) t^{-\frac{\beta_i}{\beta_i^*}-1}$. Αν η συνθήκη ισχύει για κάθε $\mathbf{X}$ πρέπει ο λόγος των συντελεστών να είναι σταθερός, έστω

$$-\frac{\beta_i}{\beta_i^*} = -\frac{\beta}{\beta^*} = \delta.$$

Οδηγούμαστε στη λύση $h_0(t) = h_0^*(1) t^{\delta-1}$, που είναι στη μορφή της συνάρτησης κινδύνου της κατανομής Weibull. Για να γίνει πιο εμφανές μπορούμε να θεωρήσουμε $h_0^*(1) = \delta \alpha^{-\delta}$, από όπου προκύπτει η Weibull(α, δ). Το παραπάνω αποτέλεσμα αποδεικνύει ότι η κατανομή Weibull, είναι η μοναδική κατανομή, σε κλειστή μορφή, που ανήκει ταυτόχρονα και στα μοντέλα αναλογικών κινδύνων και στα A.F.T. μοντέλα. Σημειώνουμε ότι οι παράμετροι παλινδρόμησης β και β^* για το μοντέλο αναλογικών κινδύνων και το A.F.T. μοντέλο αντίστοιχα, όταν έχουμε κατανομή Weibull, είναι ανάλογες μεταξύ τους. Στη συνέχεια παραθέτουμε την περίπτωση που μια από τις παραμέτρους της κατανομής Weibull εξαρτάται από το διάνυσμα των συμμεταβλητών $\mathbf{X}$.

4.5.2 A.F.T. μοντέλο με παραμέτρους που εξαρτώνται από το διάνυσμα των συμμεταβλητών

Έστω ότι έχουμε το A.F.T. μοντέλο με χρόνους επιβίωσης από την κατανομή Weibull(α, δ). Μπορούμε να θεωρήσουμε μοντέλα παλινδρόμησης όπου είτε η παράμετρος α είτε η δ εξαρτάται από το διάνυσμα $\mathbf{X}$. Καθώς και οι δύο είναι θετικές μια κατάλληλη μορφή είναι: $\alpha(\mathbf{X}) = \exp(\beta' \mathbf{X})$ και $\delta(\mathbf{X}) = \exp(\gamma' \mathbf{X})$, όπου β και γ είναι τα διανύσματα των συντελεστών για το μοντέλο παλινδρόμησης και είναι ίδιας διάστασης με το $\mathbf{X}$. Σε αυτήν την περίπτωση ισχύει $\alpha(\mathbf{X}) > 0$ και $\delta(\mathbf{X}) > 0$, χωρίς κανένα περιορισμό για τα β και γ . Το μοντέλο παλινδρόμησης όπου μόνο η παράμετρος α εξαρτάται από το διάνυσμα των συμμεταβλητών, $\mathbf{X}$, εί-

ναι χρήσιμο σε πολλές περιπτώσεις. Σε αυτήν την περίπτωση η συνάρτηση επιβίωσης της μεταβλητής T δοθέντος του διανύσματος των συμμεταβλητών, $\mathbf{X}$, είναι:

$$S(t|\mathbf{X}) = \exp[-\{t/\alpha(\mathbf{X})\}^\delta], t \geq 0. \quad (4.5.4)$$

Για τον λογάριθμο της μεταβλητής του χρόνου ζωής, έστω $Y = \log T$, η συνάρτηση επιβίωσης είναι:

$$S(y|\mathbf{X}) = \exp \left[-\exp \left\{ \frac{y - u(\mathbf{X})}{b} \right\} \right], -\infty < y < \infty, \quad (4.5.5)$$

με $u(\mathbf{X}) = \log\{\alpha(\mathbf{X})\}$ και $b = \delta^{-1}$. Αυτή η κατανομή ανήκει στην οικογένεια κατανομών location-scale, για ακραίες τιμές, με $u = u(\mathbf{X})$. Η μορφή της συνάρτησης επιβίωσης των location-scale κατανομών αναφέρεται παρακάτω. Το μοντέλο της σχέσης (4.5.4) αναφέρεται ως log-location-scale ή μοντέλο επιταχυνόμενου χρόνου αποτυχίας (accelerated failure time model A.F.T.). Αυτού του είδους τα μοντέλα είναι από τα πιο ευρέως χρησιμοποιούμενα παραμετρικά μοντέλα παλινδρόμησης και αναλύονται στην συνέχεια.

4.5.3 Log-location-scale μοντέλα επιταχυνόμενου χρόνου αποτυχίας

Οι location-scale κατανομές έχουν συνάρτηση επιβίωσης της μορφής

$$S(y, u, b) = S_0 \left(\frac{y - u}{b} \right), -\infty < y < \infty, \quad (4.5.6)$$

η $u \in \mathbb{R}$ είναι η παράμετρος θέσης, η $b > 0$ είναι η παράμετρος σχήματος και η S_0 είναι μια πλήρως ορισμένη συνάρτηση επιβίωσης. Τα location-scale μοντέλα παλινδρόμησης θεωρούν την κατανομή της Y δοθέντος $\mathbf{X}$ να είναι της παραπάνω μορφής, με $u(\mathbf{X}) = u$ και

$$S(y|\mathbf{X}) = S_0 \left\{ \frac{y - u(\mathbf{X})}{b} \right\}, -\infty < y < \infty, \quad (4.5.7)$$

με την S_0 να είναι ανεξάρτητη του $\mathbf{X}$. Ένας άλλος τρόπος να εκφραστεί αυτό είναι ο εξής:

$$Y = u(\mathbf{X}) + bZ, \quad (4.5.8)$$

με Z να είναι μια τυχαία μεταβλητή με συνάρτηση επιβίωσης $S_0(z)$. Η οικογένεια των μοντέλων για τα οποία η Z έχει τυπική κανονική κατανομή είναι συχνά η βάση για την ανάλυση παλινδρόμησης με δεδομένα χρόνου ζωής. Η χρήση

άλλων κατανομών για την Z όπως η extreme value logistic είναι επίσης κοινή. Η συνάρτηση επιβίωσης για την T δοθέντος $\mathbf{X}$, που αντιστοιχεί στη σχέση (4.5.7), είναι της μορφής

$$S(t|\mathbf{X}) = S_0^* \left[\left\{ \frac{t}{\alpha(\mathbf{X})} \right\}^\delta \right], t \geq 0, \quad (4.5.9)$$

με $\alpha(\mathbf{X}) = \exp\{u(\mathbf{X})\}$, $\delta = b^{-1}$ και $S_0^*(t) = S\{\log(t)\}$. Οι συμμεταβλητές αλλάζουν ουσιαστικά την κλίμακα του χρόνου και η σχέση (4.5.9) καλείται συχνά ως μοντέλο επιταχυνόμενου χρόνου αποτυχίας. Γενικά, αν $\alpha(\mathbf{X}) > 1$ η επίδραση του διανύσματος των συμμεταβλητών είναι ότι επιβραδύνει τον χρόνο, ενώ αν $\alpha(\mathbf{X}) < 1$ το αποτέλεσμα είναι ότι επιταχύνει τον χρόνο. Διαφορετικά διανύσματα συμμεταβλητών, έστω $\mathbf{X}_1$ $\mathbf{X}_2$, δίνουν συναρτήσεις πυκνότητας πιθανότητας και συναρτήσεις επιβίωσης, για την μεταβλητή του λογαρίθμου του χρόνου ζωής, που είναι παράλληλη μετατόπιση η μία της άλλης. Οι συναρτήσεις αυτές έχουν το ίδιο σχήμα αλλά χωρίζονται από μία απόσταση $u(\mathbf{X}_1) - u(\mathbf{X}_2)$. Αρκετά είναι τα μοντέλα στον τομέα της μηχανικής, στα οποία ο χρόνος μέχρι το υπό μελέτη γεγονός επιταχύνεται με τη θερμότητα, με την τάση ή άλλους παράγοντες, που είναι αυτού του τύπου και έχουν γραμμικές προδιαγραφές $u(\mathbf{X}) = \beta' \mathbf{X}$. Για παράδειγμα, το μοντέλο με $u(\mathbf{X}) = \beta_0 + \beta_1 X$, με $X = \log(\text{παράγοντα})$ αναφέρεται συχνά ως αντίστροφο power law μοντέλο. Το μοντέλο αυτό καθώς το β_1 είναι συνήθως αρνητικό, χρησιμοποιείται όταν ο παράγοντας είναι η υψηλή τάση του ρεύματος. Η ακόλουθη ενότητα παρουσιάζει κάποιους γενικούς τρόπους αναζήτησης του κατάλληλου μοντέλου και αξιολόγησης του.

4.6 Αναζήτηση και αξιολόγηση μοντέλου

4.6.1 Αναζήτηση κατάλληλου μοντέλου

Μια προσέγγιση για την επιλογή του κατάλληλου μοντέλου για τα διαθέσιμα δεδομένα είναι η προσαρμογή κάποιων συνηθισμένων μοντέλων και έπειτα η αξιολόγηση της προσαρμογής τους. Αυτή η προσέγγιση είναι εύλογη όταν υπάρχει πληροφορία από προηγούμενες μελέτες ή όταν υπάρχει κάποια θεωρία που υποδεικνύει συγκεκριμένο τύπο μοντέλων για τα δεδομένα. Ωστόσο, είναι απαραίτητο να γίνεται διερευνητική ανάλυση στην οποία λαμβάνουμε υπόψιν μια ευρεία γκάμα κατανομών για την κατανομή της T , δοθέντος του διανύσματος των συμμεταβλητών $\mathbf{X}$. Συνήθως παρατηρούμε ανεξάρτητες παρατηρήσεις $(T_i, \delta_i, \mathbf{X}_i)$ για ένα τυχαίο δείγμα μεγέθους n , με T_i να είναι είτε ο χρόνος ζωής είτε ο χρόνος λογοχρισίας του i -οστού ατόμου, δ_i ο δείκτης λογοχρισίας και $\mathbf{X}_i$ το διάνυσμα των συμμεταβλητών, $i = 1, \dots, n$.

Σε περιπτώσεις που υπάρχει μόνο μια ποσοτική συμμεταβλητή X οι γραφικές παραστάσεις του χρόνου, T_i , ή του λογαρίθμου του χρόνου, $y_i = \log(T_i)$, έναντι της συμμεταβλητής X_i ή μιας συνάρτησης της συμμεταβλητής μπορεί να υποδεικνύουν τη μορφή της σχέσης μεταξύ T και X . Η παρουσία λογοκρισίας αποτελεί ένα πρόβλημα. Όμως, αν το ποσοστό των λογοκριμένων χρόνων είναι μικρό, τότε είναι συνήθως ικανοποιητικό να σημειώσουμε ποια σημεία του γραφήματος αντιστοιχούν στους λογοκριμένους χρόνους. Αν το ποσοστό λογοκρισίας είναι μεγάλο ίσως να μην μπορούμε να βγάλουμε συμπέρασμα για την σχέση μεταξύ T και X . Όταν υπάρχουν περισσότερες από μία ποσοτικές συμμεταβλητές και μικρό ποσοστό λογοκρισίας, οι γραφικές παραστάσεις του χρόνου T_i , $i = 1, \dots, n$, ή του λογαρίθμου του χρόνου $y_i = \log(T_i)$, έναντι της κάθε συμμεταβλητής ξεχωριστά είναι συνήθως χρήσιμες, εκτός αν οι συμμεταβλητές είναι σημαντικά συσχετισμένες, τότε οι γραφικές παραστάσεις ίσως να μην υποδεικνύουν την κατάλληλη σχέση.

Μια ακόμη χρήσιμη τεχνική είναι να ομαδοποιήσουμε τα άτομα της μελέτης έτσι ώστε μέσα στις ομάδες οι τιμές των συμμεταβλητών που είναι μεγαλύτερου ενδιαφέροντος να είναι κοντινές. Εκτός από την περίπτωση που οι συμμεταβλητές είναι διακριτές με λίγες το πλήθος πιθανές τιμές, όπου έτσι χάνεται πληροφορία. Ο υπολογισμός της μέσης τιμής ή της διαμέσου των χρόνων ζωής ή των λογαριθμισμένων χρόνων ζωής για κάθε ομάδα είναι χρήσιμος. Η χρήση της διαμέσου είναι προτιμότερη τις περισσότερες φορές, ειδικά στις περιπτώσεις που υπάρχει λογοκρισία. Ένας βασικός λόγος που προτιμάμε να υπολογίσουμε την διάμεσο είναι γιατί μπορεί να εκτιμηθεί αξιόπιστα χωρίς να χρειάζεται να γνωρίζουμε όλους τους ακριβείς χρόνους ζωής. Ωστόσο, αυτό ισχύει μόνο όταν το ποσοστό των λογοκριμένων χρόνων ζωής είναι μικρότερο από 50%, διαφορετικά η εκτίμηση της δεν είναι εφικτή. Επιπλέον η διάμεσος δεν επηρεάζεται από ακραίες παρατηρήσεις και έχει σαφή ορισμό: το 50% των ατόμων αναμένεται να έχει βιώσει το υπό μελέτη γεγονός μέχρι το συγκεκριμένο χρονικό σημείο. Γραφικά εργαλεία, όπως τα θηκογράμματα (boxplots) είναι επίσης πολύτιμα, αλλά χρειάζονται τροποποίηση έτσι ώστε να λαμβάνουν υπόψιν την λογοκρισία. Η κατανομή του χρόνου ζωής για το εσωτερικό των ομάδων, που δημιουργήθηκαν όπως προαναφέρθηκε, μπορεί να εξεταστεί με μεγαλύτερη λεπτομέρεια μέσω μη παραμετρικών εκτιμήσεων της συνάρτησης κινδύνου, της συνάρτησης πυκνότητας πιθανότητας ή της συνάρτησης επιβίωσης, εφόσον έχουμε αρκετά δεδομένα. Οι γραφικές παραστάσεις της εκτιμώμενης Kaplan Meier συνάρτησης επιβίωσης $\hat{S}_j(t)$ για κάθε μια από τις, έστω J , ομάδες είναι χρήσιμη. Μια καλή διαδικασία είναι να κάνουμε τη γραφική παράσταση των Weibull probability plots των $\log[-\log\{\hat{S}_j(t)\}]$ έναντι $\log T$, για κάθε ομάδα. Με αυτόν τον τρόπο ελέγχουμε κατά πόσο τα δεδομένα προέρχονται από την κατανομή Weibull. Εναλλακτικά,

μπορούν να χρησιμοποιηθούν οι εκτιμητές των Nelson Aalen για να γίνει η γραφική παράσταση των $\log[\hat{H}_j(t)]$ έναντι $\log T$. Σημειώνουμε τα ακόλουθα:

1. Για το μοντέλο επιταχυνόμενου χρόνου αποτυχίας, προκύπτει από τη σχέση (4.5.8) ότι:

$$S(t|\mathbf{X}) = S_0 \left\{ \frac{\log t - u(\mathbf{X})}{b} \right\},$$

και έτσι

$$\log\{-\log S(t|\mathbf{X})\} = \log\left(-\log\left[S_0 \left\{ \frac{\log t - u(\mathbf{X})}{b} \right\}\right]\right).$$

Οπότε, αν η $u(\mathbf{X})$ είναι προσεγγιστικά σταθερή για άτομα στο εσωτερικό της ίδιας ομάδας, για κάθε ομάδα $j = 1, \dots, J$, και αν το μοντέλο A.F.T. είναι κατάλληλο, τότε η γραφική παράσταση των $\log\{-\log\{\hat{S}_j(t)\}\}$ έναντι $\log t$ θα πρέπει να είναι περίπου παράλληλη στον οριζόντιο άξονα. Η γραφική παράσταση των $\hat{S}_j(t)$ έναντι $\log t$, θα είναι επίσης παράλληλη στον οριζόντιο άξονα, αλλά λόγω των παρατηρήσεων που ακολουθούν επιλέγουμε να κάνουμε τη γραφική παράσταση των $\log\{-\log\{\hat{S}_j(t)\}\}$ έναντι $\log t$.

2. Για το μοντέλο αναλογικών κινδύνων και με χρήση της σχέσης (4.5.2) συνεπάγεται ότι:

$$\log\{-\log\{S(t|\mathbf{X})\}\} = \log\{-\log\{S_0(t)\}\} + \log\{w(\mathbf{X})\}.$$

Έτσι, αν η $w(\mathbf{X})$ είναι προσεγγιστικά σταθερή στο εσωτερικό κάθε ομάδας, $j = 1, \dots, J$, η γραφική παράσταση των $\log\{-\log\{\hat{S}_j(t)\}\}$ έναντι $\log t$ θα πρέπει να είναι περίπου παράλληλη στον οριζόντιο άξονα.

3. Αν η γραφική παράσταση των $\log\{-\log\{\hat{S}_j(t)\}\}$ έναντι $\log t$ είναι περίπου γραμμική, τότε προτείνεται η χρήση μοντέλων για δεδομένα χρόνου ζωής με κατανομή Weibull. Αν ακόμη η γραφική παράσταση είναι προσεγγιστικά παράλληλη στον οριζόντιο άξονα τότε προτείνεται η χρήση του μοντέλου της σχέσης (4.5.4) με παράμετρο σχήματος δ σταθερά. Τότε, αυτό το μοντέλο είναι ταυτόχρονα και μοντέλο επιταχυνόμενου χρόνου αποτυχίας και μοντέλο αναλογικών κινδύνων.

Η ανάλυση των δεδομένων περιλαμβάνει τη διερεύνηση, την προσαρμογή και την αξιολόγηση του μοντέλου. Στη συνέχεια, παρουσιάζουμε κάποια παραδείγματα γραφικής διερεύνησης μοντέλου.

Παράδειγμα 1:

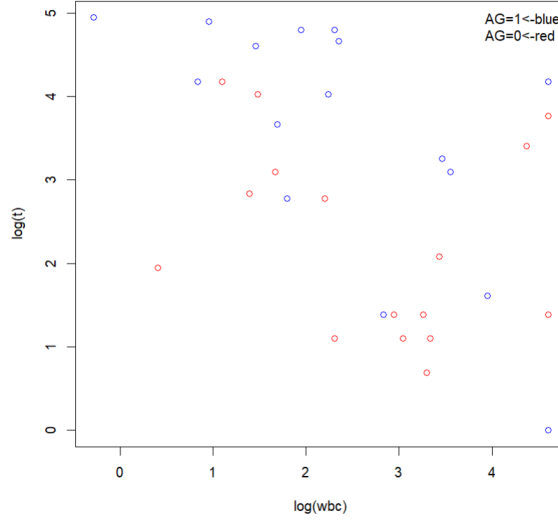
Για το παράδειγμα αυτό χρησιμοποιούμε τα δεδομένα *Leukemia*, που παρουσιάσαμε στην ενότητα 1.2. Τα αρχικά δεδομένα δεν έχουν λογοκριμένους χρόνους, ωστόσο αντικαταστήσαμε τρεις χρόνους ζωής με χρόνους λογοκρισίας για να παρουσιάσουμε το παράδειγμα που ακολουθεί. Η γραφική παράσταση του $\log(T)$ έναντι $\log(wbc)$, με $wbc = WBC/1000$ παρατίθεται στο Σχήμα 4.1. Για την υλοποίηση της στην R θα ακολουθήσουμε τα επόμενα βήματα. Αρχικά, καλούμε τη βιβλιοθήκη *stats*, γιατί περιέχει τη συνάρτηση *plot*. Με τη συνάρτηση *as.factor* δηλώνουμε τη συμμεταβλητή *ag* ως κατηγορική με δυνατές τιμές 0 και 1. Στο αντικείμενο *color_easy* αποθηκεύουμε την αντιστοίχιση χρωμάτων με βάση τις τιμές της *ag*: το κόκκινο χρώμα αντιστοιχεί στην τιμή 0, ενώ το μπλε στην τιμή 1. Με την εντολή *plot* και δίνοντας τα κατάλληλα ορίσματα γίνεται η γραφική παράσταση του $\log(T)$ έναντι $\log(wbc)$ και κάθε σημείο έχει το αντίστοιχο χρώμα ανάλογα με την τιμή που έχει η συμμεταβλητή *ag* για το συγκεκριμένο άτομο. Η εντολή *legend* προσθέτει τη λεζάντα που υποδεικνύει την τιμή της συμμεταβλητής *ag* για κάθε άτομο.

```
library(stats)
ag<-as.factor(ag)
color_easy=c("red","blue")[ag]
plot(log(wbc),log(t),col=color_easy)
legend(x="topright",legend=c("AG=1<-blue","AG=0<-red"),
      bty = "n")
```

Δύο άτομα με $ag = 1$ έχουν $wbc = 100$ και $t = 1$, οπότε δύο σημεία της γραφικής ταυτίζονται. Κάναμε τη γραφική παράσταση του $\log(T)$ έναντι $\log(wbc)$, λόγω προηγούμενων γνώσεων που υποδεικνύουν ότι βιολογικού τύπου μετρήσεις αντιμετωπίζονται καλύτερα ως συμμεταβλητές σε λογαριθμική κλίμακα. Η γραφική παράσταση του Σχήματος 4.1 υποδεικνύει σημαντική διακύμανση στους χρόνους ζωής για άτομα με κοντινές τιμές στις συμμεταβλητές. Επίσης, υποδηλώνει ότι οι χρόνοι ζωής τείνουν να είναι μικρότεροι για άτομα με υψηλές τιμές για τη συμμεταβλητή *WBC* και αρνητικό *ag* τεστ, δηλαδή $ag = 0$. Ο βαθμός συσχέτισης μεταξύ των *WBC* και *ag* δεν είναι μεγάλος.

Παράδειγμα 2:

Στο συγκεκριμένο παράδειγμα χρησιμοποιούμε τα δεδομένα *Insulating Fluid Failure Times*, που παρουσιάσαμε στην ενότητα 1.2. Όπως φαίνεται και από τα δεδομένα, όσο μεγαλύτερη είναι η τιμή της τάσης τόσο μικρότερος είναι ο χρόνος ζωής, γεγονός που υποδεικνύει ότι η συμμεταβλητή επιταχύνει τον χρόνο μέχρι



Σχήμα 4.1: Γραφική παράσταση του $\log(T)$ έναντι $\log(wbc)$ για τα δεδομένα χρόνου ζωής ατόμων που πάσχουν από λευχαιμία.

να συμβεί το υπό μελέτη γεγονός. Για το λόγο αυτό το A.F.T. μοντέλο είναι κατάλληλο. Ο τομέας της μηχανικής προτείνει ένα power law μοντέλο για αυτό το πρόβλημα, καθώς το μοντέλο αυτό εφαρμόζεται συχνά όταν επεξηγηματική μεταβλητή είναι η τάση του ρεύματος. Το μοντέλο αυτό είναι ένα A.F.T. μοντέλο όπου η παράμετρος κλίμακας α σχετίζεται με την τάση v με την σχέση $\alpha(v) = cv^p$. Αντίστοιχα, αν γίνει χρήση του λογαρίθμου του χρόνου ζωής και του ισοδύναμου location-scale μοντέλου έχουμε, $u(v) = \beta_0 + \beta_1 \log v$ με $\beta_0 = \log c$ και $\beta_1 = p$. Ο τομέας της μηχανικής προτείνει ότι το A.F.T. μοντέλο με κατανομή Weibull είναι κατάλληλο.

Η γραφική παράσταση των $\log[-\log\{\hat{S}_j(t)\}]$ έναντι $\log t$, με $\hat{S}_1(t), \dots, \hat{S}_7(t)$, να είναι η εκτίμηση Kaplan-Meier της συνάρτησης επιβίωσης, για τάση $v = 26, 28, 30, 32, 34, 36, 38$ kV αντίστοιχα δίνεται στο Σχήμα 4.2. Για να είναι πιο ξεκάθαρη η γραφική παράσταση παρουσιάζονται τα επίπεδα τάσης $v = 26, 30, 34, 38$. Η μόνη πιθανή ακραία τιμή είναι $v = 26$ για μικρό χρόνο ζωής. Εκτός αυτού οι γραφικές παραστάσεις δεν είναι παράλληλες, αλλά είναι προσεγγιστικά παράλληλες και προσεγγιστικά γραμμικές. Λαμβάνοντας υπόψιν ότι για κάθε γραφική έχουμε διαθέσιμο ένα μικρό δείγμα χρόνων ζωής θεωρούμε τη γραφική κατάλληλη για ένα Weibull A.F.T. μοντέλο. Παραθέτουμε τον κώδικα που χρησιμοποιήθηκε για τη δημιουργία της γραφικής παράστασης των $\log[-\log\{\hat{S}_j(t)\}]$ έναντι $\log t$.

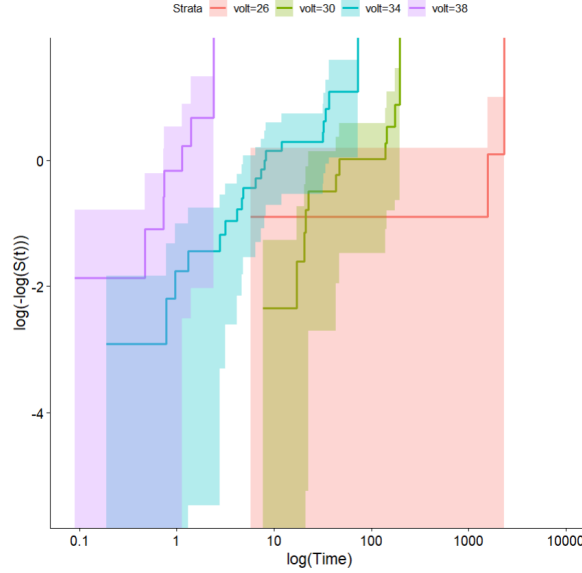
Αρχικά, καλούμε τις βιβλιοθήκες `survival` και `survminer`, καθώς η πρώτη περιέχει τη συνάρτηση `survfit`, ενώ η δεύτερη τη συνάρτηση `ggsurvplot`, που θα χρησιμοποιήσουμε στη συνέχεια.

```
library(survival)
library(survminer)
```

Επειδή θέλουμε να κάνουμε τη γραφική παράσταση μόνο για επίπεδα τάσης $v = 26, 30, 34, 38$ παρέχουμε τις αντίστοιχες παρατηρήσεις των δεδομένων στην **R**. Παρόλο που τα δεδομένα μας είναι μη λογοκριμένα, προκειμένου να πάρουμε την Kaplan-Meier εκτίμηση της συνάρτησης επιβίωσης, δημιουργούμε μια επιπλέον μεταβλητή για τα δεδομένα, την `status`. Η μεταβλητή αυτή περιέχει τον δείκτη λογοκρισίας για κάθε παρατήρηση, οπότε περιέχει την τιμή 1 για κάθε παρατήρηση. Η συνάρτηση `Surv` δέχεται ως πρώτο όρισμα τους χρόνους ζωής και ως δεύτερο όρισμα τους δείκτες λογοκρισίας. Η συνάρτηση `survfit` υπολογίζει την εκτίμηση Kaplan-Meier συνάρτησης επιβίωσης. Δέχεται ως πρώτο όρισμα το αντικείμενο `Surv`, δεξιά του συμβόλου $\sim$ γράφουμε τη μεταβλητή `volt`, καθώς ως προς τις κατηγορίες αυτής της μεταβλητής θέλουμε να χωριστούν τα δεδομένα και για κάθε κατηγορία να πάρουμε την εκτιμώμενη συνάρτηση επιβίωσης. Το δεύτερο όρισμα της εντολής `survfit` υποδεικνύει ποια δεδομένα πρέπει να χρησιμοποιηθούν. Οι εκτιμώμενες συναρτήσεις επιβίωσης αποθηκεύονται στο αντικείμενο `km`. Η εντολή `ggsurvplot` χρησιμοποιείται για να γίνει η γραφική παράσταση που θέλουμε. Πρώτο όρισμα της είναι το αντικείμενο `km`, στο δεύτερο όρισμα δηλώνουμε ότι στη γραφική παράσταση θέλουμε να περιέχονται διαστήματα εμπιστοσύνης και με το όρισμα `fun="cloglog"` εφαρμόζεται δύο φορές ο φυσικός λογάριθμος στη συνάρτηση επιβίωσης, με αρνητικό πρόσημο εφαρμόζεται μόνο στον εσωτερικό λογάριθμο.

```
km<-survfit(Surv(time,status)~volt,data=voltage)
ggsurvplot(km,conf.int=TRUE,fun ="cloglog")
```

Η γραφική παράσταση $\log t$ έναντι $\log v$ του Σχήματος 4.3, υποδεικνύει ότι για μικρό μέγεθος δείγματος είναι δύσκολο να βγάλουμε συμπεράσματα από τις γραφικές παραστάσεις για το αν τα δεδομένα είναι κατάλληλα για ένα συγκεκριμένο μοντέλο παλινδρόμησης, ακόμη και στην περίπτωση που έχουμε μόνο μια συμμεταβλητή. Όταν τα δεδομένα είναι στρωματοποιημένα σε ομάδες με βάση μια συμμεταβλητή, υπάρχει σημαντική διακύμανση στις μη παραμετρικές εκτιμήσεις, όπως είναι οι $\hat{S}_j(t)$, και είναι δύσκολο να έχουμε ξεκάθαρη εικόνα. Ωστόσο, είναι ξεκάθαρο ότι υπάρχει σχέση μεταξύ της συμμεταβλητής και του χρόνου ζωής. Για παράδειγμα, από την γραφική παράσταση του Σχήματος 4.2 έχουμε ότι παρόλο



Σχήμα 4.2: Γραφική παράσταση των $\log[-\log\{\hat{S}_j(t)\}]$ έναντι $\log t$.

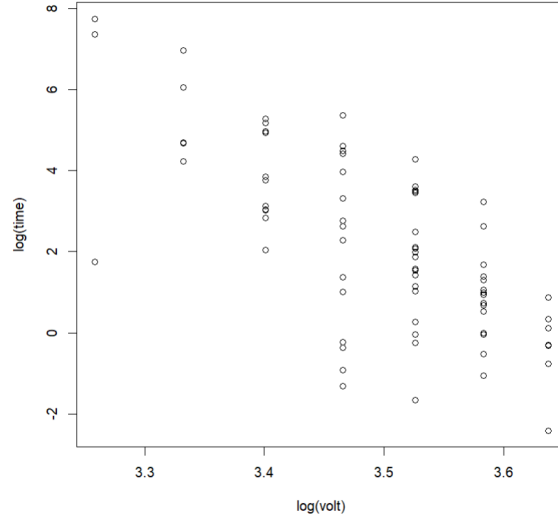
που οι γραφικές παραστάσεις της κατανομής του χρόνου ζωής για τα διάφορα επίπεδα της συµµεταβλητής επικαλύπτονται, η διάµεσος του χρόνου ζωής ή του λογαρίθµου του χρόνου ζωής µειώνεται καθώς αυξάνεται η τάση.

Η εντολή που χρησιµοποιήθηκε για να γίνει η γραφική παράσταση των $\log t$ έναντι $\log v$ είναι η `plot`, µε πρώτο όρισμα `log(volt)` και δεύτερο `log(time)`.

```
plot(log(volt),log(time))
```

4.6.2 Αξιολόγηση του προσαρµοσµένου μοντέλου

Αφού προσαρµόσουµε ένα µοντέλο παλινδρόµησης είναι σηµαντικό να ελέγξουµε τις υποθέσεις του µοντέλου για τα παρατηρούµενα δεδοµένα και την ευαισθησία των συµπερασµάτων σε αλλαγές του µοντέλου ή των δεδοµένων. Οι γραφικές παραστάσεις που προαναφέρθηκαν δεν χρησιµοποιούνται µόνο για να αξιολογηθεί αν το συγκεκριµένο µοντέλο που προσαρµόσαµε είναι κατάλληλο, αλλά και για την εύρεση πιθανών ακραίων τιµών ή επηρεάζουσων παρατηρήσεων. Η επέκταση του µοντέλου που έχει προσαρµοστεί είναι σηµαντική και περιλαµβά-νει προσθήκη παραµέτρων που αντιπροσωπεύουν συγκεκριµένου είδους απόκλιση

Σχήμα 4.3: Γραφική παράσταση $\log t$ έναντι $\log v$.

από το προσαρμοσμένο μοντέλο. Η ανάγκη για την προσθήκη αυτών των επιπλέον παραμέτρων μπορεί να αξιολογηθεί μέσω ελέγχων υποθέσεων. Κάποιοι τρόποι με τους οποίους επεκτείνουμε ένα μοντέλο είναι οι εξής: με την προσθήκη συμμεταβλητών που αναπαριστούν όρους αλληλεπίδρασης ή μη γραμμικούς όρους, με σκοπό να γίνει έλεγχος ενός γραμμικού μοντέλου, επιτρέποντας στο b σε ένα location-scale μοντέλο να εξαρτάται από το διάνυσμα των συμμεταβλητών, με σκοπό τον έλεγχο της συνέπειας του b , με τη δημιουργία όρων αλληλεπίδρασης των συμμεταβλητών με τον χρόνο για τον έλεγχο της αναλογικότητας στο μοντέλο αναλογικών κινδύνων, μέσω της διεύρυνσης της οικογένειας μοντέλων $S_0(z)$ για την Z στη σχέση (4.5.8), ως έλεγχος της κατανομής σφάλματος που έχουμε θεωρήσει.

Θεωρούμε το A.F.T. μοντέλο ή ισοδύναμα location-scale μοντέλα, που δίνονται στις σχέσεις (4.5.7) και (4.5.8), με την $u(\mathbf{X})$ να εξαρτάται από το διάνυσμα των παραμέτρων β . Έστω ότι $(y_i, \delta_i, \mathbf{X}_i)$, $i = 1, \dots, n$ αναπαριστά ένα τυχαίο λογοκριμένο δείγμα από τις κατανομές των Y_i δοθέντος $\mathbf{X}_i$, με το $\mathbf{X}_i$ να είναι είτε σταθερό είτε τυχαίο ανάλογα με το είδος της μελέτης. Συμβολίζουμε με u_i το $u(\mathbf{X}_i, \beta)$, για τη σχέση (4.5.7), και έστω $\hat{\beta}, \hat{b}$ οι εκτιμητές μέγιστης πιθανοφάνειας που προκύπτουν αν προσαρμόσουμε τα δεδομένα, $(y_i, \delta_i, \mathbf{X}_i)$, στην σχέση (4.5.7). Οι τυποποιημένες μεταβλητές $Z_i = (Y_i - u_i)/b$ είναι ανεξάρτητες και ισόνομες με συνάρτηση επιβίωσης $S_0(z)$ υπό την σχέση (4.5.7). Ο συνήθης τρόπος να

οριστούν τα υπόλοιπα είναι ως:

$$z_i = \frac{y_i - \hat{u}_i}{\hat{b}}, i = 1, \dots, n,$$

με $\hat{u}_i = u(\mathbf{X}_i, \hat{\beta})$. Αν δεν υπάρχει λογοκρισία τα $\hat{z}_i$, τουλάχιστον για μεγάλο μέγεθος δείγματος, θα πρέπει να αποτελούν τυχαίο δείγμα από την $S_0(z)$ και να είναι ανεξάρτητα των $\mathbf{X}_i$. Ωστόσο, επειδή τα $\hat{z}_i$ εξαρτώνται από τα $\hat{\beta}, \hat{b}$ δεν είναι ακριβώς ανεξάρτητα και ισόνομα κατανεμημένα με κατανομή $S_0(z)$. Τροποποιήσεις για τα $\hat{z}_i$ μπορούν να θεωρηθούν, έτσι ώστε να μοιάζουν περισσότερο με δείγμα από την $S_0(z)$. Αυτό ίσως αποτελεί λόγο ανησυχίας για περιπτώσεις που το μέγεθος δείγματος είναι μικρό και είναι δύσκολη μια ακριβής αξιολόγηση του μοντέλου.

Οι γραφικές παραστάσεις των $\hat{z}_i$ έναντι των συμμεταβλητών ή άλλων παραγόντων, όπως οι προσαρμοσμένες τιμές $\hat{u}_i$ μπορούν να χρησιμοποιηθούν για τον έλεγχο της συνέπειας του b στη σχέση (4.5.7) ή για την αναζήτηση συστηματικών αποκλίσεων από την $u(\mathbf{X}, \beta)$ που έχουμε θεωρήσει. Probability plots για τα $\hat{z}_i$ μπορούν να χρησιμοποιηθούν για την αξιολόγηση της κατανομής της βασικής συνάρτησης επιβίωσης $S_0(z)$. Αποκλίσεις από τη μορφή της location-scale οικογένειας είναι πιο δύσκολο να εντοπιστούν, δύο τρόποι είναι μέσω επέκτασης του μοντέλου ή με γραφικές μεθόδους.

Αν το μοντέλο είναι σωστό τα $\hat{z}_i$ πρέπει να μοιάζουν με ένα λογοκριμένο τυχαίο δείγμα από την κατανομή $S_0(z)$. Όταν το ποσοστό λογοκρισίας είναι μικρό οι γραφικές παραστάσεις των $\hat{z}_i$ έναντι των συμμεταβλητών είναι και πάλι χρήσιμες, αλλά θα πρέπει τα υπόλοιπα για τους λογοκριμένους και μη λογοκριμένους χρόνους ζωής να συμβολιστούν διαφορετικά. Όμως, όταν το ποσοστό λογοκρισίας είναι μεγάλο αυτές οι γραφικές παραστάσεις είναι χρήσιμες υπό αυστηρούς συμβιβασμούς. Το βασικό πρόβλημα είναι ότι με τη λογοκρισία η κατανομή των y_i δεν δίνεται από τη σχέση (4.5.7), αλλά εξαρτάται από τη διαδικασία λογοκρισίας. Για παράδειγμα, αν τα Y_i έχουν την κατανομή της σχέσης (4.5.7) δοθέντος $\mathbf{X}_i$ και $\log C_i$ είναι ένας σταθερός χρόνος λογοκρισίας, τότε $Y_i^* = \min\{Y_i, \log C_i\}$ έχει συνάρτηση κατανομής

$$F^*(y_i) = \begin{cases} F\left(\frac{y_i - u_i}{b}\right), & \text{αν } y_i \leq \log C_i, \\ 1, & \text{αν } y_i > \log C_i, \end{cases}$$

με $F_0(z) = 1 - S_0(z)$. Έτσι, είναι πιθανό τα y_i να είναι υλοποιήσεις της Y_i^* που δεν έχουν την κατανομή της σχέσης (4.5.7). Μια προσέγγιση που χρησιμοποιείται για τις γραφικές παραστάσεις των υπολοίπων έναντι των συμμεταβλητών όταν

υπάρχει σημαντικό ποσοστό λογοκρισίας αναφέρεται στη συνέχεια. Στην προσέγγιση αυτή χρησιμοποιούνται τροποποιημένα υπόλοιπα, για τα οποία τα $\hat{z}_i$ που αντιστοιχούν σε λογοκριμένες παρατηρήσεις προσαρμόζονται-στρογγυλοποιούνται προς τα πάνω. Συνήθως για τα location-scale μοντέλα ορίζονται ως:

$$\hat{z}_i^{adj} = \delta_i \hat{z}_i + (1 - \delta_i)E(Z_i | Z_i \geq \hat{z}_i), \quad (4.6.1)$$

με $E(Z_i | Z_i \geq \hat{z}_i)$ να είναι η αναμενόμενη τιμή των Z_i βασισμένη στην κατανομή $S_0(z)$. Τα $\hat{z}_i$ για τους λογοκριμένους χρόνους ζωής είναι ένα κάτω φράγμα των πραγματικών υπολοίπων. Σημειώνουμε ότι αν $Z_i = (Y_i - u_i)/b$ και $R_i = (\log C_i - u_i)/b$, τότε

$$Z_i^{adj} = \delta_i Z_i + (1 - \delta_i)E(Z_i | Z_i \geq R_i).$$

Η αναμενόμενη τιμή του είναι $E(Z_i^{adj}) = E(Z_i)$. Συνεπώς, τα τροποποιημένα υπόλοιπα, $\hat{z}_i^{adj}$, αναμένεται να συμπεριφέρονται περίπου όπως τα Z_i . Επομένως, οι εξομαλυμένες (smooths) ή οι μέσες μη παραμετρικές καμπύλες παλινδρόμησης στη γραφική παράσταση των $\hat{z}_i^{adj}$ έναντι των συμμεταβλητών, πρέπει να είναι περίπου σταθερές και ίσες με $E(Z_i)$ αν το μοντέλο είναι σωστό. Ωστόσο, είναι σημαντικό να σημειώσουμε ότι η κατανομή των τροποποιημένων υπολοίπων εξαρτάται από τους χρόνους λογοκρισίας και ίσως διαφέρει από την κατανομή των Z_i . Ένα ακόμη πρόβλημα είναι ότι τα τροποποιημένα υπόλοιπα υπολογίζονται χρησιμοποιώντας το μοντέλο που θέλουμε να ελέγξουμε και έτσι αν το ποσοστό λογοκρισίας είναι μεγάλο η δυνατότητα να εντοπίσουμε αδυναμίες του μοντέλου μέσω γραφικών παραστάσεων είναι περιορισμένη.

4.6.3 Εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων

Θεωρούμε ότι έχουμε διαθέσιμο το τυχαίο δείγμα δεδομένων $(y_i, \delta_i, \mathbf{X}_i), i = 1, \dots, n$, $y_i = \log T_i$ είναι είτε ο λογαριθμισμένος χρόνος ζωής είτε ο λογαριθμισμένος χρόνος λογοκρισίας. Θα περιγράψουμε την εκτίμηση παραμέτρων μέσω μέγιστης πιθανοφάνειας και τα συμπεράσματα που προκύπτουν για το location-scale μοντέλο για την Y με την μορφή της σχέσης (4.5.7) ή (4.5.8), αλλά με

$$u(\mathbf{X}_i, \boldsymbol{\beta}) = \boldsymbol{\beta}' \mathbf{X}_i, \quad (4.6.2)$$

όπου $\boldsymbol{\beta}$ και $\mathbf{X}_i$ είναι διανύσματα διάστασης $m \times 1$. Τα περισσότερα μοντέλα παλινδρόμησης έχουν σταθερό όρο, έτσι $\boldsymbol{\beta}' \mathbf{X}_i = \beta_0 + \beta_1 X_{i,1} + \dots + \beta_{m-1} X_{i,m-1}$. Είναι καλύτερο, για πιο ακριβείς υπολογισμούς, να κεντράρουμε τις ποσοτικές συμμεταβλητές (να επιλέξουμε την αρχή), έτσι ώστε οι μέσες τιμές τους να είναι κοντά στο μηδέν. Η επιλογή της αρχής για μία συμμεταβλητή επηρεάζει μόνο τον

σταθερό όρο β_0 και όχι τα $\beta_1, \dots, \beta_{m-1}$. Ο σταθερός όρος είναι μια πιο ασήμαντη παράμετρος όταν οι συμμεταβλητές είναι κεντραρισμένες, και οι συσχετίσεις μεταξύ του β_0 και των υπόλοιπων εκτιμώμενων συντελεστών παλινδρόμησης μειώνονται. Κάποιες φορές κεντράρουμε τις συμμεταβλητές, αλλά όταν η μέση τους τιμή είναι αρκετά μικρή τις διατηρούμε όπως είναι για λόγους απλότητας.

Υπό το μοντέλο της σχέσης (4.5.6) η συνάρτηση πιθανοφάνειας είναι της μορφής

$$L\{u(\mathbf{X}_i, \boldsymbol{\beta}), b\} = \prod_{i=1}^n \left[\frac{1}{b} f_0 \left\{ \frac{y_i - u(\mathbf{X}_i, \boldsymbol{\beta})}{b} \right\} \right]^{\delta_i} S_0 \left\{ \frac{y_i - u(\mathbf{X}_i, \boldsymbol{\beta})}{b} \right\}^{1-\delta_i},$$

με $f_0(z) = -S'_0(z)$ να είναι η συνάρτηση πυκνότητας πιθανότητας για την $Z_i = (Y_i - u_i)/b$. Η λογαριθμισμένη συνάρτηση πιθανοφάνειας συμβολίζεται με $l(\boldsymbol{\beta}, b)$ όπου $z_i = \{y_i - u_i(\mathbf{X}_i, \boldsymbol{\beta})\}/b$, δηλαδή το z_i είναι συνάρτηση του $\boldsymbol{\beta}$ και συνάρτηση του b , $u_i = u(\mathbf{X}_i, \boldsymbol{\beta})$, $f_0(z) = -S'_0(z)$ είναι η συνάρτηση πυκνότητας πιθανότητας για την $Z = (Y_i - u_i)/b$ και $r = \sum_{i=1}^n \delta_i$ το πλήθος των μη λογοκριμένων χρόνων ζωής, δηλαδή

$$l(\boldsymbol{\beta}, b) = -r \log b + \sum_{i=1}^n [\delta_i \log f_0\{z_i(\boldsymbol{\beta}, b)\} + (1 - \delta_i) \log S_0\{z_i(\boldsymbol{\beta}, b)\}]. \quad (4.6.3)$$

Έστω $\mathbf{X}_i = (X_{i1}, \dots, X_{im})'$ και $\mathbf{X}$ ο $n \times m$ πίνακας με στοιχεία X_{ij} . Τότε $\frac{\partial z_i(\boldsymbol{\beta}, b)}{\partial \beta_j} = -X_{ij}b^{-1}$ και $\frac{\partial z_i(\boldsymbol{\beta}, b)}{\partial b} = -z_i(\boldsymbol{\beta}, b)b^{-1}$, καθώς

$$\begin{aligned} \frac{\partial z_i(\boldsymbol{\beta}, b)}{\partial \beta_j} &= \frac{\partial \{y_i - u(\mathbf{X}_i, \boldsymbol{\beta})\}/b}{\partial \beta_j} \\ &= \frac{\partial \{y_i - (\beta_0 + \beta_1 X_{i,1} + \dots + \beta_{m-1} X_{i,m-1})\}/b}{\partial \beta_j} \\ &= -X_{ij}b^{-1} \\ \frac{\partial z_i(\boldsymbol{\beta}, b)}{\partial b} &= \frac{\partial (y_i - u_i)/b}{\partial b} = \frac{y_i - u_i}{b^2} \\ &= z_i(\boldsymbol{\beta}, b)b^{-1}. \end{aligned}$$

Στις παραγώγους για λόγους ευκολίας συμβολίζουμε το $z_i(\boldsymbol{\beta}, b)$ με z_i . Οι πρώτες παράγωγοι της $l(\boldsymbol{\beta}, b)$ είναι:

$$\begin{aligned} \frac{\partial l(\boldsymbol{\beta}, b)}{\partial \beta_i} &= \sum_{i=1}^n \left\{ \delta_i \frac{\partial \log f_0(z_i)}{\partial z_i} + (1 - \delta_i) \frac{\partial \log S_0(z_i)}{\partial z_i} \right\} \frac{\partial z_i}{\partial \beta_j} \\ &= -\frac{1}{b} \sum_{i=1}^n \left\{ \delta_i \frac{\partial \log f_0(z_i)}{\partial z_i} + (1 - \delta_i) \frac{\partial \log S_0(z_i)}{\partial z_i} \right\} X_{ij}, \end{aligned}$$

και

$$\frac{\partial l(\beta, b)}{\partial b} = -\frac{r}{b} - \frac{1}{b} \sum_{i=1}^n \left\{ \delta_i \frac{\partial \log f_0(z_i)}{\partial z_i} + (1 - \delta_i) \frac{\partial \log S_0(z_i)}{\partial z_i} \right\} z_i$$

Οι δεύτερες παράγωγοι είναι:

$$\frac{\partial^2 l(\beta, b)}{\partial \beta_j \partial \beta_k} = \frac{1}{b^2} \sum_{i=1}^n \left\{ \delta_i \frac{\partial^2 \log f_0(z_i)}{\partial z_i^2} + (1 - \delta_i) \frac{\partial^2 \log S_0(z_i)}{\partial z_i^2} \right\} X_{ij} X_{ik},$$

$$\begin{aligned} \frac{\partial l(\beta, b)}{\partial b^2} &= \frac{r}{b^2} + \frac{2}{b^2} \sum_{i=1}^n \left\{ \delta_i \frac{\partial \log f_0(z_i)}{\partial z_i} + (1 - \delta_i) \frac{\partial \log S_0(z_i)}{\partial z_i} \right\} z_i \\ &\quad + \frac{1}{b^2} \sum_{i=1}^n \left\{ \delta_i \frac{\partial^2 \log f_0(z_i)}{\partial z_i^2} + (1 - \delta_i) \frac{\partial^2 \log S_0(z_i)}{\partial z_i^2} \right\} z_i^2, \end{aligned}$$

και

$$\begin{aligned} \frac{\partial^2 l(\beta, b)}{\partial \beta_j \partial b} &= \frac{1}{b^2} \sum_{i=1}^n \left\{ \delta_i \frac{\partial \log f_0(z_i)}{\partial z_i} + (1 - \delta_i) \frac{\partial \log S_0(z_i)}{\partial z_i} \right\} X_{ij} \\ &\quad + \frac{1}{b^2} \sum_{i=1}^n \left\{ \delta_i \frac{\partial^2 \log f_0(z_i)}{\partial z_i^2} + (1 - \delta_i) \frac{\partial^2 \log S_0(z_i)}{\partial z_i^2} \right\} z_i X_{ij}. \end{aligned}$$

Οι εκτιμητές μέγιστης πιθανοφάνειας $\hat{\beta}$ και $\hat{b}$ παρατηρούνται από την επίλυση των εξισώσεων $\frac{\partial l(\beta, b)}{\partial \beta} = 0$, $\frac{\partial l(\beta, b)}{\partial b} = 0$ ή από απευθείας μεγιστοποίηση της $l(\beta, b)$.

4.6.4 Επιλογή μεταξύ μοντέλου αναλογικών κινδύνων και μοντέλου επιταχυνόμενου χρόνου αποτυχίας

Στην ενότητα αυτή παρουσιάζουμε ένα παράδειγμα στο οποίο καλούμαστε να επιλέξουμε μεταξύ του μοντέλου αναλογικών κινδύνων και του A.F.T. μοντέλου για ένα συγκεκριμένο σύνολο δεδομένων χρόνων ζωής. Το σύνολο δεδομένων που θα χρησιμοποιήσουμε είναι το `oldmort`, που παρουσιάστηκε στην Ενότητα 1.2. Υπενθυμίζουμε ότι το σύνολο δεδομένων `oldmort` περιλαμβάνει το ιστορικό ζωής ατόμων τρίτης ηλικίας, τα οποία ήταν από 60 έτη και πάνω στην αρχή της έρευνας και παρακολουθήθηκαν μέχρι να τελειώσει η έρευνα ή μέχρι τον θάνατο. Στην περίπτωση που το τέλος της παρατήρησης είναι το τέλος της έρευνας η παρατήρηση είναι λογοκριμένη. Το παράδειγμα βασίζεται στην εργασία του Broström (2024). Αρχικά, θέλουμε να προσαρμόσουμε τα δεδομένα στο μοντέλο

αναλογικών κινδύνων, ωστόσο υποθέτουμε ότι η βασική συνάρτηση κινδύνου είναι παραμετρική. Σε αυτήν την περίπτωση το μοντέλο αναλογικών κινδύνων έχει τα ίδια χαρακτηριστικά με το μοντέλο αναλογικών κινδύνων του Cox, μόνο που η βασική συνάρτηση κινδύνου εκτιμάται αναλυτικά μαζί με τις παραμέτρους του μοντέλου. Για να μοντελοποιήσουμε τη βασική συνάρτηση κινδύνου υποθέτουμε ότι έχει κατανομή Gompertz. Η επιλογή αυτή βασίζεται στο γεγονός ότι πρόκειται για μια συνήθη επιλογή για τη βασική συνάρτηση κινδύνου, που χρησιμοποιείται ευρέως για δεδομένα καθώς περιγράφει καταστάσεις όπου ο κίνδυνος αυξάνεται εκθετικά με την πάροδο του χρόνου. Συνεπώς η κατανομή Gompertz είναι κατάλληλη για τα συγκεκριμένα δεδομένα. Η συνάρτηση πυκνότητας πιθανότητας της κατανομής είναι:

$$f(x, \eta, b) = b\eta \exp\{\eta + bx - \eta \exp(bx)\}, \text{ με } x \geq 0, b > 0, \eta > 0.$$

Για τη μοντελοποίηση στην R ακολουθούμε τα επόμενα βήματα.

Αρχικά, καλούμε τις βιβλιοθήκες `survival` και `eha`. Η `survival` περιέχει την συνάρτηση `Surv`. Η βιβλιοθήκη `eha` περιέχει το σύνολο δεδομένων `oldmort`, καθώς και τις συναρτήσεις `phreg`, `coxreg`, `check.dist` και `aftreg` που χρησιμοποιούμε στη συνέχεια. Ξεκινάμε προσαρμόζοντας τα δεδομένα στο μοντέλο αναλογικών κινδύνων με παραμετρική βασική συνάρτηση κινδύνου. Η συνάρτηση `Surv` δέχεται ως πρώτο όρισμα την ηλικία εισόδου στην έρευνα, ως δεύτερο όρισμα την ηλικία που σταμάτησε η παρακολούθηση και ως τρίτο όρισμα τον αντίστοιχο δείκτη λογοκρισίας κάθε ατόμου. Επιλέγουμε να αφαιρέσουμε 60 έτη από την ηλικία εισόδου και από την ηλικία που σταμάτησε η παρακολούθηση, καθώς όσα άτομα συμμετείχαν στην έρευνα ήταν από 60 έτη και πάνω. Η συνάρτηση `phreg` προσαρμόζει το μοντέλο αναλογικών κινδύνων με παραμετρική βασική συνάρτηση κινδύνου. Δέχεται ως πρώτο όρισμα τη συνάρτηση `Surv` και δεξιά του συμβόλου `~` ορίζουμε τις συμμεταβλητές που θα περιέχονται στο μοντέλο, στη δική μας περίπτωση αυτές είναι οι `sex`, `civ` και `region`. Στο όρισμα `dist` της συνάρτησης `phreg` δηλώνουμε την κατανομή που υποθέτουμε για τη βασική συνάρτηση κινδύνου, ενώ στο όρισμα `data` το σύνολο των δεδομένων που χρησιμοποιούμε. Το μοντέλο αυτό αποθηκεύεται στο αντικείμενο `fit.g`. Με την εντολή `summary` λαμβάνουμε τα αποτελέσματα από την προσαρμογή του μοντέλου, που παρατίθεται παρακάτω.

```
library(survival)
library(eha)
fit.g <- phreg(Surv(enter - 60, exit - 60, event) ~ sex
               + civ + region, dist = "gompertz", data = oldmort)
summary(fit.g)
```

Covariate	Mean	Coef	Rel.Risk	S.E.	LR p
sex					0.0000
male	0.406	0	1 (reference)		
female	0.594	-0.239	0.788	0.047	
civ					0.0000
unmarried	0.080	0	1 (reference)		
married	0.530	-0.416	0.660	0.082	
widow	0.390	-0.262	0.770	0.080	
region					0.0008
town	0.111	0	1 (reference)		
industry	0.326	0.265	1.303	0.086	
rural	0.563	0.119	1.127	0.084	
Events	1971				
Total time at risk	37824				
Max. log. likelihood	-7268				
LR test statistic	56.99				
Degrees of freedom	5				
Overall p-value	5.08754e-11				

Στη συνέχεια, προσαρμόζουμε τα δεδομένα στο μοντέλο αναλογικών κινδύνων του Cox με χρήση της συνάρτησης `coxreg`, σε αυτήν την περίπτωση η βασική συνάρτηση κινδύνου είναι μη παραμετρική. Η συνάρτηση `coxreg` δέχεται ως πρώτο όρισμα τη συνάρτηση `Surv`, η οποία δέχεται τα ίδια όρια με την προηγούμενη περίπτωση. Δεξιά του συμβόλου `~` ορίζουμε τις συμμεταβλητές που θα περιέχονται στο μοντέλο και στο όρισμα `data` το σύνολο των δεδομένων που χρησιμοποιούμε. Αποθηκεύουμε το μοντέλο που μόλις προσαρμόσαμε στο αντικείμενο `fit.c`. Η εντολή `summary` μας παρέχει τα αποτελέσματα από την προσαρμογή του μοντέλου `fit.c`.

```
fit.c <- coxreg(Surv(enter - 60, exit - 60, event) ~ sex
               + civ + region, data = oldmort)
summary(fit.c)
```

Covariate	Mean	Coef	Rel.Risk	S.E.	LR p
sex					0.000
male	0.406	0	1 (reference)		
female	0.594	-0.235	0.791	0.047	
civ					0.000

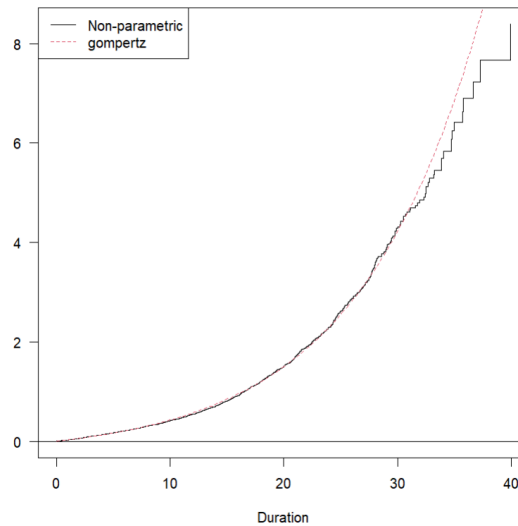
unmarried	0.080	0	1 (reference)	
married	0.530	-0.410	0.664	0.082
widow	0.390	-0.262	0.770	0.080
region				0.001
town	0.111	0	1 (reference)	
industry	0.326	0.268	1.308	0.086
rural	0.563	0.122	1.130	0.084
Events	1971			
Total time at risk	37824			
Max. log. likelihood	-13551			
LR test statistic	55.71			
Degrees of freedom	5			
Overall p-value	9.32227e-11			

Για τα δύο μοντέλα που προσαρμόσαμε, `fit.g` και `fit.c`, συγκρίνουμε τις εκτιμώμενες βασικές συναρτήσεις κινδύνου κάνοντας τις γραφικές παραστάσεις τους. Η σύγκριση αυτή γίνεται με χρήση της συνάρτησης `check.dist`, που δέχεται ως πρώτο όρισμα το ένα μοντέλο και ως δεύτερο το άλλο μοντέλο.

```
check.dist(fit.c, fit.g)
```

Με βάση τα όσα δίνονται στο Σχήμα 4.4 συμπεραίνουμε ότι η βασική συνάρτηση κινδύνου, για την οποία υποθέσαμε κατανομή Gompertz, είναι πολύ κοντά με την μη παραμετρική βασική συνάρτηση κινδύνου και υποδεικνύουν ότι η θνησιμότητα στην τρίτη ηλικία αυξάνεται εκθετικά με τα έτη. Ωστόσο, τα τελευταία 10 χρόνια, δηλαδή για 90 έτη και πάνω, ο ρυθμός αύξησης της θνησιμότητας μειώνεται.

Επαναλαμβάνουμε την διαδικασία ανάλυσης των δεδομένων `oldmort`, αυτή την φορά προσαρμόζουμε το μοντέλο επιταχυνόμενου χρόνου αποτυχίας. Υποθέτουμε και πάλι ότι η βασική συνάρτηση κινδύνου είναι παραμετρική με κατανομή Gompertz. Η συνάρτηση `aftreg` μας επιτρέπει να προσαρμόσουμε τα δεδομένα στο μοντέλο επιταχυνόμενου χρόνου αποτυχίας με παραμετρική βασική συνάρτηση κινδύνου. Δέχεται ως πρώτο όρισμα την συνάρτηση `Surv`, δεξιά του συμβόλου `~` γράφουμε τις συμμεταβλητές που θέλουμε στο μοντέλο, στο όρισμα `data` το σύνολο δεδομένων που χρησιμοποιούμε, στο όρισμα `id` τον δείκτη που ξεχωρίζει κάθε άτομο και τέλος στο όρισμα `dist` την κατανομή που υποθέτουμε για την βασική συνάρτηση κινδύνου. Η συνάρτηση `Surv` δέχεται τα ίδια ορίσματα με πριν. Το A.F.T. που προσαρμόσαμε αποθηκεύεται στο αντικείμενο `fit.g1`. Η εντολή `summary` μας δίνει τα αποτελέσματα από την προσαρμογή του μοντέλου.



Σχήμα 4.4: Σύγκριση των βασικών συναρτήσεων κινδύνου που παρατηρούμε αν προσαρμόσουμε τα δεδομένα `oldmort` στο μοντέλο αναλογικών κινδύνων, υποθέτοντας αρχικά ότι η βασική συνάρτηση κινδύνου είναι παραμετρική με κατανομή Gompertz και έπειτα υποθέτοντας ότι είναι μη παραμετρική.

```
fit.g1 <- aftreg(Surv(enter - 60, exit - 60, event) ~ sex
+ civ+ region, data = oldmort, id = id,
dist = "gompertz")
summary(fit.g1)
```

Covariate	W.mean	Coef	Time-Accn	se(Coef)	LR p
sex					0.0001
male	0.406	0	1	(reference)	
female	0.594	-0.079	0.924	0.020	
civ					0.0011
unmarried	0.080	0	1	(reference)	
married	0.530	-0.132	0.876	0.036	
widow	0.390	-0.079	0.924	0.034	
region					0.0045
town	0.111	0	1	(reference)	
industry	0.326	0.115	1.122	0.040	
rural	0.563	0.072	1.075	0.040	

Events	1971
Total time at risk	37824
Max. log. likelihood	-7281.5
LR test statistic	35.52
Degrees of freedom	5
Overall p-value	1.18508e-06

Συμπέρασμα:

Το πρόβλημα της επιλογής μεταξύ του μοντέλου αναλογικών κινδύνων και του μοντέλου επιταχυνόμενου χρόνου αποτυχίας μπορεί να επιλυθεί συγκρίνοντας τα AIC των δύο μοντέλων. Καθώς ο αριθμός των παραμέτρων είναι ίδιος και στις δύο περιπτώσεις αυτό ισοδυναμεί με το να συγκρίνουμε τις μεγιστοποιημένες συναρτήσεις πιθανοφάνειας. Για το παράδειγμα μας συγκρίνοντας τα αποτελέσματα για το μοντέλο αναλογικών κινδύνων και για το μοντέλο επιταχυνόμενου χρόνου αποτυχίας, υποθέτοντας κατανομή Gompertz για τη βασική συνάρτηση κινδύνου, η μεγιστοποιημένη συνάρτηση πιθανοφάνειας έχει την τιμή -7268 και -7281.5, αντίστοιχα. Συμπεραίνουμε ότι το μοντέλο αναλογικών κινδύνων προσαρμόζεται καλύτερα στα δεδομένα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- Aalen, O. (1978). Nonparametric Inference for a Family of Counting Processes, *The Annals of Statistics* **6**(4): 701 – 726.
- Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control* **19**(6): 716–723.
- Bagdonavicius, V. and Nikulin, M. (2001). *Accelerated Life Models: Modeling and Statistical Analysis*, 1st edn, Chapman and Hall/CRC.
- Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, New York.
- Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship, *The Annals of Statistics* **2**(3): 437–453.
- Breslow, N. E. (1972). Discussion of professor cox’s paper, *Journal of the Royal Statistical Society: Series B (Methodological)* **34**(2): 216–217.
- Broström, G. (2024). Parametric survival models, Available at: <https://cran.r-project.org/web/packages/eha/vignettes/parametric.html>.
- Cain, K. C. and Lange, N. T. (1984). Approximate case influence for the proportional hazards regression model with censored data, *Biometrics* **40**(2): 40,493–499.
- Caplehorn, J. R. and Bell, J. (1991). Methadone dosage and retention of patients in maintenance treatment, *Medical Journal of Australia* **154**(3): 195–199.

- Collett, D. (2003). *Modelling Survival Data in Medical Research, Second Edition*, Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis.
- Cox, D. and Oakes, D. (1984). *Analysis of Survival Data*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Taylor & Francis.
- Cox, D. R. (1972). Regression models and life-tables, *Journal of the Royal Statistical Society. Series B (Methodological)* **34**(2): 187–220.
- Cox, D. R. (1975). Partial likelihood, *Biometrika* **62**(2): 269–276.
- Cox, D. R. and Snell, E. J. (1968). A general definition of residuals, *Journal of the Royal Statistical Society. Series B (Methodological)* **30**(2): 248–275.
- Efron, B. (1977). The efficiency of Cox’s likelihood function for censored data, *Journal of the American Statistical Association* **72**(359): 557–565.
- Feigl, P. and Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information, *Biometrics* **21**(4): 826–838.
- Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons, New York.
- Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals, *Biometrika* **81**(3): 515–526.
- J.Bickel, P. and A.Doksum, K. (2001). *Mathematical Statistics: Basic Ideas and Selected Topics, Volume I*, 2nd edn, Prentice Hall, Upper Saddle River, NJ.
- Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**(282): 457–481.
- Klein, J. and Moeschberger, M. (2003). *Survival Analysis: Techniques for Censored and Truncated Data (2nd edition)*, Springer-Verlag, New York.
- Kleinbaum, D. G. and Klein, M. (2012). *Survival Analysis : A Self-Learning Text, Third Edition*, Statistics for Biology and Health, 3rd ed. 2012. edn, Springer New York, New York, NY.
- Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*, 2 edn, John Wiley & Sons, Hoboken, NJ.

- Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer Chemother Rep* **50**(3): 163–170.
- Nelson, W. (1972a). Graphical analysis of accelerated life test data with the inverse power law model, *IEEE Transactions on Reliability* **1**: 2–11.
- Nelson, W. (1972b). Theory and applications of hazard plotting for censored failure data, *Technometrics* **14**(4): 945–966.
- Ng’andu, N. (1997). An empirical comparison of statistical tests for assessing the proportional hazards assumption of cox’s model, *Statistics in Medicine* **16**(6): 611–626.
- Rodríguez, G. (2010). Parametric survival models, Available at: <https://data.princeton.edu/pop509/ParametricSurvival>. Originally written Spring 2001; revised Spring 2005 and Summer 2010.
- Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model, *Biometrika* **69**(1): 69,239–241.
- Stanford University (n.d.). Unit 6 logrank test 6.1 introduction, Available at: <https://web.stanford.edu/~lutian/coursepdf/unit6.pdf>.
- Tableman, M. and Kim, J. S. (2003). *Survival Analysis Using S: Analysis of Time-to-Event Data*, Chapman & Hall/CRC, Boca Raton, FL.
- Therneau, T. M., Grambsch, P. M. and Fleming, T. R. (1990). Martingale-based residuals for survival models, *Biometrika* **77**(1): 147–160.
- Tsiatis, A. A. (1981). A large sample study of Cox’s regression model, *The Annals of Statistics* **9**(1): 93–108.