ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ
ΔΠΜΣ ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ

‘Κριτήρια επιλογής μοντέλων για την κλάση μοντέλων ευπάθειας’

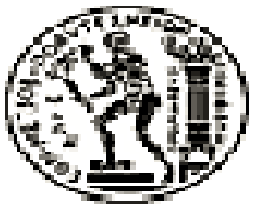
ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

-του-

-Γεώργιου Δ. Ξυνού-

Επιβλέπουσα: ‘Ιλια Βόντα
Επίκουρος Καθηγήτρια ΕΜΠ

Γεώργιος Ξυνός
Αθήνα –Ημερομηνία 2012-



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ
ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΔΠΜΣ ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ

‘Κριτήρια επιλογής μοντέλων για την κλάση μοντέλων ευπάθειας’

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

-του-

-Γεώργιου Δ. Ξυνού-

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την ΗΜΕΡΟΜΗΝΙΑ

Επ. Καθηγήτρια Ι.Βόντα Αναπλ.Καθηγήτρια Χ.Καρώνη Καθ. Χ.Κουκουβίνος

Αθήνα, Ημερομηνία
Εθνικό Μετσόβιο Πολυτεχνείο

……..

Γεώργιος Δ.Ξυνός
Πτυχιούχος Μαθηματικός Ε.Κ.Π.Α.

ΠΕΡΙΛΗΨΗ

Ο σκοπός αυτής της διπλωματικής είναι η χρήση του AIC κριτηρίου για την επιλογή των σημαντικών μεταβλητών σε μοντέλα ευπάθειας για την περίπτωση των λογοκριμένων δεδομένων. Στο κεφάλαιο 1 γίνεται εκτενής αναφορά στο πιο σημαντικό και ευρέως χρησιμοποιούμενο μοντέλο στην ανάλυση επιβίωσης, που είναι το μοντέλο του Cox. Το μοντέλο αυτό χρησιμοποιείται για να ερμηνεύσει τη σχέση μεταξύ μιας μεταβλητής που περιγράφει το χρόνο επιβίωσης ενός ατόμου με άλλες συμμεταβλητές. Στο κεφάλαιο 2 το μοντέλο του Cox επεκτείνεται και δημιουργεί μια νέα ομάδα μοντέλων, τα μοντέλα ευπάθειας. Η ευπάθεια είναι μια θετική τυχαία μεταβλητή που υπεισέρχεται στο μοντέλο του Cox με σκοπό να εξηγήσει διαφοροποιήσεις στον πληθυσμό που το μοντέλο του Cox δεν καταφέρνει να εξηγήσει. Διαφορετικές δυνατές κατανομές της ευπάθειας έχουν σαν αποτέλεσμα τον ορισμό διαφόρων μοντέλων ευπάθειας. Περιγράφονται στο κεφάλαιο αυτό βασικές κατηγορίες και ιδιότητες των μοντέλων ευπάθειας. Στο κεφάλαιο 3 αναφέρονται τα πιο γνωστά και ευρέως χρησιμοποιούμενα κριτήρια επιλογής μοντέλων.

Στο τελευταίο κεφάλαιο χρησιμοποιώντας την γλώσσα προγραμματισμού R εξετάζουμε την αποτελεσματικότητα του κριτηρίου AIC στην επιλογή των σημαντικών μεταβλητών, όταν η συνάρτηση επιβίωσης ορίζεται μέσω μιας κλάσης μοντέλων ευπάθειας και τα δεδομένα είναι λογοκριμένα από δεξιά. Η θεωρία στην οποία βασιστήκαμε αλλά και κατ' επέκταση το πρόγραμμα, είναι έτσι σχεδιασμένα έτσι ώστε με μια μικρή αλλαγή ως προς την κατανομή της ευπάθειας που θέλουμε να υποθέσουμε να μπορούμε να χειριστούμε όλα τα δυνατά μοντέλα συνεχούς ευπάθειας. Τέτοια μοντέλα είναι το μοντέλο Gamma και Inverse Gaussian. Υπολογιστικά, ερχόμαστε να καλύψουμε κενά που παρουσιάζουν στατιστικά πακέτα όπως π.χ. η R στην οποία οι κατανομές ευπάθειας που μπορεί να υποθέσει κανείς σε συνδυασμό με κριτήρια επιλογής μοντέλων είναι περιορισμένες σε αριθμό. Τα αποτελέσματά μας περιγράφονται μέσω δεδομένων προσομοίωσης.

ABSTRACT

The purpose of this dissertation is the use the AIC criterion for the selection of the significant variables in frailty models for the case of right-censored data. Chapter 1 provides an extensive review of the most important and most widely used model in survival analysis, namely, the Cox model. The purpose of this model is to examine the relationship between the survival time of individuals and other covariates. In Chapter 2 we discuss the class of frailty models which are extensions of the Cox model. The frailty is a positive random variable that is included in the Cox model in order to explain heterogeneity in the population that the Cox model fails to explain. Different distributions of the frailty result in the creation of different frailty models. In the same chapter we describe the basic properties and categories of fraiity models. In Chapter 3, the most popular and most widely used model selection criteria are discussed.

In Chapter 4, using the language R, we examine the effectiveness of the AIC criterion in the selection of important variables, in the case where the survival function is defined by a class of frailty models and the data are censored from the right. The theory on which we have relied upon and consequently the code are designed in such a way, so that by a small change in the considered frailty distribution, all possible models of continuous frailty can be treated. Such models are the Gamma and Inverse Gaussian frailty models. Our code fills a gap in statistical packages like R in which the distributions of the frailty one can consider in conjunction with model selection criteria is limited in number. Our results are described through simulated data.

Περιεχόμενα

Κεφάλαιο 4
Επιλογή μοντέλων στην κλάση των μοντέλων ευπάθειας

ΚΕΦΑΛΑΙΟ 1

Μοντέλο Αναλογικών Κινδύνων του Cox

1.1 Εισαγωγή

Στη στατιστική ένα από τα πιο βασικά και διαδεδομένα μοντέλα, με πολυάριθμες εφαρμογές είναι το μοντέλο της γραμμικής παλινδρόμησης, όπου η εξαρτημένη μεταβλητή Y σχετίζεται με επεξηγηματικές μεταβλητές x_i μέσω της εξίσωσης

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

όπου ε τα υπόλοιπα που θεωρούνται ανεξάρτητες τυχαίες μεταβλητές που συνήθως ακολουθούν κανονική κατανομή.

Το βασικό αυτό μοντέλο τροποιήθηκε ανάλογα και χρησιμοποιείται ευρέως και σε μοντέλα διάρκειας ζωής. Τρείς βασικές κατηγορίες μοντέλων παλινδρόμησης που αναπτύχθηκαν για τις ανάγκες της ανάλυσης επιβίωσης είναι το μοντέλο επιταχυνόμενης διακοπής, το μοντέλο αναλογικών λόγων συμπληρωματικών πιθανοτήτων και το μοντέλο αναλογικών κινδύνων. Στην τελευταία κατηγορία μοντέλων ανήκει το ημι-παραμετρικό μοντέλο του Cox.

Το μοντέλο αναλογικών κινδύνων του Cox (Proportional Hazards Cox model) παρουσιάστηκε από τον Cox 1972 (Cox, D.R-Regression models with life tables,JRSS Series B,34:187-220). Σκοπός του μοντέλου του Cox είναι η μοντελοποίηση της συνάρτησης κινδύνου. Εκτενής είναι η χρησιμοποίηση αυτού του μοντέλου σε ανάλυση λογοκριμένων δεδομένων επιβίωσης που αφορούν βιοϊατρικές εφαρμογές.

1.2 Το Μοντέλο του Cox

‘Εστω ότι έχουμε n το πλήθος άτομα που λαμβάνουν μέρος σε μια έρευνα και ότι $x_1, x_2, \dots, x_p$ είναι οι μεταβλητές που πιστεύουμε ότι επηρεάζουν το χρόνο ζωής των παραπάνω ατόμων. Οι συμμεταβλητές μπορεί να παριστάνουν διάφορες θεραπείες, χαρακτηριστικά των ατόμων (π.χ. φύλλο, χρώμα, ηλικία, ύψος) και εξωγενείς παράγοντες δηλαδή μεταβλητές που ελέγχονται από τον πειραματιστή (π.χ. η δόση ενός φαρμάκου, η θερμοκρασία υπό την οποία λειτουργούν τα μηχανήματα σε ένα πείραμα).

Οι επεξηγηματικές μεταβλητές $x_i, i = 1,2,.., p$ μπορούν να συνδυαστούν ώστε να εξηγήσουν επιδράσεις αλληλεπίδρασης μεταξύ τους. Επίσης μπορούν να ταξινομηθούν είτε ως σταθερές (ανεξάρτητες από το χρόνο) είτε ως εξαρτημένες από το χρόνο.

Θεωρούμε ότι οι συμμεταβλητές δεν εξαρτώνται από το χρόνο, δηλαδή οι τιμές των x_i έχουν καταγραφεί στο ξεκίνημα της μελέτης δηλαδή την χρονική στιγμή $t = 0$. Επίσης θεωρούμε ότι οι τιμές αυτές παραμένουν αμετάβλητες από την αρχή ως το τέλος της μελέτης.

Το ημι-παραμετρικό μοντέλο αναλογικών κινδύνων του Cox δίνεται από τη σχέση

$$h(t|x) = h_0(t)e^{\beta^T x}$$

όπου η $h(t|x)$ είναι η συνάρτηση κινδύνου στο χρόνο t ενώ η $h_0(t)$ ονομάζεται αναφορική συνάρτηση κινδύνου (baseline hazard function) στο χρόνο t. Το $\beta = (\beta_1, \beta_2, \dots, \beta_p)$ είναι ένα διάνυσμα p συντελεστών, οι οποίοι εκφράζουν ποσοτικά την επίδραση της κάθε μίας των συμμεταβλητών x_i.

Ο κίνδυνος εξαρτάται από δύο διαφορετικούς παράγοντες. Ο πρώτος παράγοντας, $h_0(t)$, είναι μια συνάρτηση μόνο του χρόνου και η οποία δεν

καθορίζεται, αλλά θεωρείται η ίδια για το σύνολο των ατόμων της μελέτης. Ο δεύτερος παράγοντας είναι μια ποσότητα που εξαρτάται από τις συμμεταβλητές μόνο μέσω του διανύσματος β.

Το μοντέλο του Cox καλείται ημι-παραμετρικό διότι δεν καθορίζει μια συγκεκριμένη μορφή για την αναφορική συνάρτηση κινδύνου $h_0(t)$ (μπορεί να πάρει οποιαδήποτε μορφή), υποθέτει όμως ότι οι επιδράσεις των μεταβλητών είναι σταθερές στο χρόνο και έχουν προσθετικό χαρακτήρα στην συνάρτηση κινδύνου. Η γενική μορφή ενός μοντέλου αναλογικών κινδύνων είναι

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

όπου $g(x)$ είναι μια συνάρτηση του διανύσματος των συμμεταβλητών. Ο όρος του αναλογικού κινδύνου προέρχεται από το γεγονός ότι οι συναρτήσεις κινδύνου δύο οποιονδήποτε ατόμων είναι η μια πολλαπλάσιο της άλλης. Συγκεκριμένα για το μοντέλο του Cox η συνάρτηση $g(x)$ είναι

$$e^{\beta^T x} = e^{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}$$

Στην περίπτωση που καθορίζουμε ποια είναι η $h_0(t)$ τότε παίρνουμε την παραμετρική μορφή του μοντέλου αναλογικής διακινδύνευσης.

Η αθροιστική συνάρτηση κινδύνου ορίζεται ως

$$H(t|x) = \int_0^t h(u|x)du = \int_0^t h_0(u)\, e^{\beta^T x} du = H_0(t)e^{\beta^T x}$$

και συνεπώς η συνάρτηση επιβίωσης θα είναι

$$S(t|x) = \exp(-H(t|x)) = \exp\{-H_0(t)e^{\beta^T x}\} = \{S_0(t)\}^{e^{\beta^T x}}$$

όπου $H_0(t)$ η αναφορική αθροιστική συνάρτηση κινδύνου και $S_0(t)$ η αναφορική συνάρτηση επιβίωσης.

1.3 Εκτίμηση των παραμέτρων

Έχοντας ορίσει το μοντέλο αναλογικών κινδύνων του Cox, ακολουθεί η εκτίμηση των συντελεστών παλινδρόμησης β. Ο μη παραμετρικός καθορισμός της αναφορικής συνάρτησης κινδύνου $h_0(t)$ καθιστά αδύνατη την χρησιμοποίηση της συνηθισμένης συνάρτησης πιθανοφάνειας για την εκτίμηση του β. Αντικειμενικός σκοπός είναι να εκτιμηθεί το β χρησιμοποιώντας όλη την πληροφορία από τα δεδομένα χωρίς να εμπλακεί η αναφορική συνάρτηση κινδύνου $h_0(t)$.

Θεωρούμε ότι έχουμε k διακεκριμένους χρόνους διακοπής ή αποτυχίας $t_{(1)} < t_{(2)} \dots < t_{(k)}$. Στην χρονική στιγμή $t_{(j)}$ διακόπτεται η λειτουργία μιας μονάδας με συμμεταβλητές x_j. Έστω R_j να συμβολίζει το σύνολο των μονάδων που είναι σε κίνδυνο αμέσως πριν από τη χρονική αυτή στιγμή. Η πιθανότητα να διακοπεί η λειτουργία μιας συγκεκριμένης μονάδας j δοθέντος ότι παύει να λειτουργεί μια οποιαδήποτε μονάδα του συνόλου R_j είναι

$$\frac{h(t_{(j)}|x_j)}{\sum_{i\in R_j} h(t_{(j)}|x_i)} = \frac{e^{\beta^T x_j}}{\sum_{i\in R_j} e^{\beta^T x_i}}$$

Ο Cox (1972) εισηγήθηκε την ακόλουθη συνάρτηση πιθανοφάνειας για την εκτίμηση του β.

$$\mathrm{Lik}(\beta) = \prod_{j=1}^{k} \left\{ \frac{e^{\beta^T x_j}}{\sum_{i\in R_j} e^{\beta^T x_i}} \right\}$$

την οποία ονόμασε μερική πιθανοφάνεια (partial likelihood) από την οποία προκύπτει η εκτιμήτρια μεγίστης μερικής πιθανοφάνειας $\hat{\beta}$ της β. Ο Cox έδειξε ότι η παραπάνω συνάρτηση μπορεί να χρησιμοποιηθεί σαν μια συνηθισμένη συνάρτηση πιθανοφάνειας επιτρέποντας με αυτό τον τρόπο την εκτίμηση του β με την συνηθισμένη διαδικασία. Συνεπώς ο εκτιμητής που

προκύπτει είναι αμερόληπτος, συνεπής και ασυμπτωτικά κανονικός. Ο πίνακας πληροφορίας $I(\beta)$, ο λόγος πιθανοφάνειας λ (likelihood ratio) καθώς και οι έλεγχοι υποθέσεων έχουν ακριβώς την ίδια συμπεριφορά όπως και στην περίπτωση της συνηθισμένης πιθανοφάνειας.

1.4 Έλεγχος υποθέσεων

1.4.1. Έλεγχος λόγου πιθανοφάνειας

Ο έλεγχος της υπόθεσης $H_0: \beta_i = 0$ μπορεί να πραγματοποιηθεί με έναν έλεγχο του λόγου των πιθανοφανειών. Στον έλεγχο αυτό το μοντέλο προσαρμόζεται με και χωρίς τη συμμεταβλητή x_i, η αφαίρεση της οποίας ισοδυναμεί με την επιβολή του περιορισμού $\beta_i = 0$. Τα $\hat{l}_1$ και $\hat{l}_0$ είναι οι δυο μεγιστοποιημένες τιμές του λογαρίθμου της μερικής πιθανοφάνειας με και χωρίς την x_i, αντίστοιχα. Η τιμή της παράστασης $-2(\hat{l}_0 - \hat{l}_1)$ συγκρίνεται με την ${X_1}^2$ κατανομή.

1.4.2 Έλεγχος Wald

Το Wald τεστ ελέγχει την μηδενική $H_0: \beta_0 = 0$ και βασίζεται στη στατιστική συνάρτηση $W = (\hat{\beta} - \beta_0)^T I(\hat{\beta})^{-1} (\hat{\beta} - \beta_0)$.Οι αναφερόμενοι έλεγχοι μπορεί να δώσουν τα ίδια αποτελέσματα αλλά το γεγονός αυτό δεν ισχύει γενικά. Ο παραπάνω έλεγχος του Wald συνήθως χρησιμοποιείται ως μια πρώτη ένδειξη των σημαντικών μεταβλητών όταν υπάρχουν πολλές υποψήφιες συμμεταβλητές.

1.4.3 Score tests

Θεωρώντας συνθήκες κανονικότητας, το score $\mathrm{U}(\theta) = \frac{\vartheta \mathrm{LogLik}(\theta|\mathrm{X})}{\vartheta\theta}$ όπου θ το διάνυσμα των παραμέτρων και X τα δεδομένα. Ακολουθεί μια ασυμπτωτικά κανονική κατανομή με μέση τιμή μηδέν και πίνακα διασπορών – συνδιασπορών ίσο με τον πίνακα πληροφορίας $\mathrm{U}(\theta) \sim \mathrm{N_p}(0, \mathrm{I}(\theta))$. Για τον έλεγχο της μηδενικής υπόθεσης $\mathrm{H_0}: \beta_0 = \beta_1 = \cdots = \beta_\mathrm{p}$ η τετραγωνική μορφή $\mathrm{Q} = \mathrm{U^T}(\hat{\theta})\mathrm{I}(\hat{\theta})^{-1}\mathrm{U}(\hat{\theta})$ ακολουθεί προσεγγιστικά την ${\mathrm{X_p}}^2$ κατανομή με p βαθμούς ελευθερίας. Οι τρεις έλεγχοι είναι ασυμπτωτικά ισοδύναμοι, σε μικρά δείγματα όμως η ισοδυναμία τους εξασθενεί. Όταν υπάρχει διαφοροποίηση στα αποτελέσματα ο έλεγχος του λόγου των πιθανοφανειών θεωρείται ο πιο αξιόπιστος, ενώ ο Wald έλεγχος είναι ο λιγότερος αξιόπιστος.

1.5 Επεκτάσεις του μοντέλου του Cox

Με κατάλληλες τροποποιήσεις το μοντέλο του Cox μπορεί να χρησιμοποιηθεί και σε περιπτώσεις που οι μεταβλητές έχουν χαρακτηριστικά διαφορετικά από αυτά που απαιτούνται για την εφαρμογή του. Για παράδειγμα, όταν μερικές από τις εξεταζόμενες μεταβλητές δεν είναι σταθερές αλλά εξαρτώνται από τον χρόνο. Επίσης στην περίπτωση που δεν ικανοποιείται η υπόθεση της αναλογικότητας την λύση την δίνει η στρωματοποιημένη ανάλυση. Στις περιπτώσεις αυτές το μοντέλο του Cox επιδέχεται κατάλληλες τροποποιήσεις ώστε να εφαρμόζεται αποτελεσματικά και σε αυτές.

1.5.1 Συμμεταβλητές εξαρτημένες από το χρόνο

Ως τώρα θεωρούσαμε ότι οι μεταβλητές παρέμεναν σταθερές στο χρόνο. Υπάρχουν όμως περιπτώσεις που οι τιμές των μεταβλητών μεταβάλλονται με το χρόνο. Μεταβλητή που εξαρτάται από το χρόνο είναι μια επεξηγηματική μεταβλητή που η τιμή της αλλάζει με το χρόνο, δηλαδή το i άτομο θα έχει την τιμή $x_i(t)$ την χρονική στιγμή t. Το μοντέλο του Cox έχει την δυνατότητα να επεκτείνεται και να ενσωματώνει τέτοιου είδους μεταβλητές. Ο συνηθέστερος τύπος μεταβλητής που εξαρτάται από το χρόνο είναι μια επαναλαμβανόμενη μέτρηση ή κάποια αλλαγή στη θεραπεία ενός ατόμου.

Οι εξαρτώμενες από το χρόνο μεταβλητές διακρίνονται σε δύο κατηγορίες στις εξωτερικές μεταβλητές (external covariates) και τις εσωτερικές μεταβλητές (internal covariates). Οι τιμές των εξωτερικών μεταβλητών δεν επηρεάζονται από την πορεία του ατόμου μέσα στη μελέτη, αλλά από ένα μηχανισμό που είναι εξωτερικός του ατόμου. Για μελέτες που η διάρκεια τους είναι μικρή η ηλικία του ατόμου θεωρείται σταθερή μεταβλητή. Στην περίπτωση που η διάρκεια του πειράματος είναι μεγάλη είναι αναγκαίο να οριστεί η ηλικία ως μεταβλητή εξαρτώμενη από το χρόνο. Η ηλικία θεωρείται εξωτερική μεταβλητή που μεταβάλλεται με τρόπο προβλέψιμο. Επίσης εξωτερική μεταβλητή είναι ο τρόπος που χορηγείται το φάρμακο σε έναν ασθενή, ο οποίος καθορίζεται από την αρχή της μελέτης και μεταβάλλεται κατά την διάρκεια της με προκαθορισμένο τρόπο.

Οι εσωτερικές μεταβλητές επηρεάζονται από την πορεία του ατόμου μέσα στη μελέτη. Οι εσωτερικές μεταβλητές λαμβάνουν τιμές ανάλογα με την πορεία της μονάδας. Παραδείγματα εσωτερικών μεταβλητών είναι οι κλινικές μετρήσεις (π.χ. η πίεση αίματος καθορίζεται στην αρχή της μελέτης αλλά

υπάρχει πιθανότητα να αλλάξει κατά την διάρκεια), η μόλυνση ενός ασθενή, το βάρος ενός ατόμου κτλ.

Οι μεταβλητές αυτές ενσωματώνονται εύκολα στο μοντέλο του Cox αντικαθιστώντας το κάθε x στη συνάρτηση της μερικής πιθανοφάνειας με $x(t_{(j)})$ και ο τύπος του λογαρίθμου της πιθανοφάνειας γίνεται

$$\text{loglik}(\beta) = \sum_{j=1}^{k} \beta^T x_j(t_{(j)}) - \sum_{j=1}^{k} \ln\left\{\sum_{i \in R_j} e^{\beta^T x_i(t_{(j)})}\right\}$$

Παρατηρούμε ότι σε κάθε χρόνο διακοπής πρέπει να διαθέτουμε τις τρέχουσες τιμές των συμμεταβλητών. Αυτό τις περισσότερες φορές δεν είναι εφικτό εκτός κι αν υπάρχει συνεχής παρακολούθηση των τιμών των συμμεταβλητών.

1.5.2 Στρωματοποίηση

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

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

$$h_i(t|x) = h_{oi}(t)e^{\beta^T x} \quad i = 1 \ldots N$$

i: δηλώνει το στρώμα του παράγοντα

N: το πλήθος των επιπέδων του παράγοντα

$h_{oi}(t)$:η αναφορική συνάρτηση κινδύνου του i στρώματος.

Από την παραπάνω σχέση φαίνεται ότι τα άτομα που ανήκουν στο ίδιο στρώμα έχουν τις ίδιες αναφορικές συναρτήσεις κινδύνου, ενώ τα άτομα που ανήκουν σε διαφορετικά στρώματα δεν έχουν τις ίδιες βασικές συναρτήσεις κινδύνου. Επιπλέον τα άτομα που ανήκουν στο ίδιο στρώμα έχουν συναρτήσεις κινδύνου με την αναλογική ιδιότητα. Για παράδειγμα έστω δύο άτομα με μεταβλητές x_1 και x_2 που ανήκουν στο στρώμα i, τότε ισχύει

$$\frac{h_i(t|x_1)}{h_i(t|x_2)} = \frac{h_{0i}e^{\beta^T x_1}}{h_{0i}e^{\beta^T x_2}} = e^{\beta^T(x_1 - x_2)}$$

Όταν τα άτομα προέρχονται από διαφορετικά στρώματα δεν έχουν ανάλογες συναρτήσεις κινδύνου, αφού οι βασικές συναρτήσεις κινδύνου $h_{01}, h_{02}, \ldots, h_{0N}$ κάθε στρώματος είναι αυθαίρετες συναρτήσεις του χρόνου και είναι ασυσχέτιστες.

Επίσης οι συντελεστές παλινδρόμησης είναι ίδιοι σε κάθε στρώμα ξεχωριστά. Στην περίπτωση που δεν ήταν ίδιοι σε κάθε στρώμα, τα δεδομένα κάθε στρώματος θα θεωρούνταν ως διαφορετικά σύνολα δεδομένων και θα αναλύονταν ξεχωριστά.

Η εκτίμηση των συντελεστών παλινδρόμησης προκύπτει από την μεγιστοποίηση της συνάρτησης μερικής πιθανοφάνειας. Ο λογάριθμος της μερικής πιθανοφάνειας στην στρωματοποίηση στο στρώμα m δίνεται ως

$$\text{loglik}_m(\beta) = \sum_{j=1}^{k_m} \beta^T x_{m_j} - \sum_{j=1}^{k_m} \ln\left\{\sum_{i \in R_j} e^{\beta^T x_{m_i}}\right\}$$

Η μόνη διαφορά της παραπάνω σχέσης με τον αντίστοιχη της μερικής πιθανοφάνειας για το μοντέλο του Cox είναι ότι έχει προστεθεί ο δείκτης m για να τονίσει ότι τα δεδομένα προέρχονται από το στρώμα $\mathrm{m}, \mathrm{m} = 1, \ldots, \mathrm{N}$.

1.6 Έλεγχοι της υπόθεσης αναλογικών κινδύνων στο μοντέλο του Cox

Η εφαρμογή του μοντέλου του Cox θεωρεί ότι η υπόθεση των αναλογικών κινδύνων ικανοποιείται ανάμεσα στα επίπεδα μιας μεταβλητής. Για αυτό το λόγο για να είναι σωστή η διαδικασία και τα αποτελέσματα που προκύπτουν να έχουν νόημα, πρέπει προτού εφαρμοστεί το μοντέλο του Cox να ελεγχθεί από τον μελετητή η υπόθεση των αναλογικών κινδύνων. Ο έλεγχος αυτός πραγματοποιείται είτε γραφικά είτε με στατιστικά που υπάρχουν για τον έλεγχο αυτό. Στη συνέχεια, και εφόσον υπάρχουν πολλές υποψήφιες συμμεταβλητές πρέπει να επιλεχθούν οι πιο σημαντικές και να συμπεριληφθούν στο μοντέλο. Καταλήγοντας στον αριθμό των μεταβλητών και εφαρμόζοντας το μοντέλο στα δεδομένα θα εμφανιστεί το κατά πόσο το μοντέλο είναι ικανοποιητικό ή θέλει βελτιώσεις.

Υλοποιώντας τους παραπάνω ελέγχους και εντοπίζοντας ότι η υπόθεση αναλογικών κινδύνων δεν ικανοποιείται τότε προβαίνουμε είτε σε μετασχηματισμούς των δεδομένων μας έτσι ώστε τελικώς να ικανοποιείται η συνθήκη είτε επιλέγουμε μια διαφορετική κλάση μοντέλων που να είναι κατάλληλη για τα δεδομένα μας.

1.6.1 Γραφικός έλεγχος της υπόθεσης αναλογικών κινδύνων

Εφόσον ισχύει η υπόθεση αναλογικών κινδύνων η συνάρτηση επιβίωσης ενός ατόμου με διάνυσμα συμμεταβλητών $\mathrm{x} = (\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_\mathrm{p})$ παίρνει την παρακάτω μορφή

$$\mathrm{S}(\mathrm{t}|\mathrm{x}) = \exp\left(-\mathrm{H}_0(\mathrm{t})\mathrm{e}^{\beta^\mathrm{T}\mathrm{x}}\right)$$

Λογαριθμίζοντας την παραπάνω σχέση καταλήγουμε στην σχέση.

$$\log[-\log(\mathrm{S}(\mathrm{t}|\mathrm{x}))] = \beta^\mathrm{T}\mathrm{x} + \log[(\mathrm{H}_0(\mathrm{t}))]$$

Έστω x_1 και x_2 τα διανύσματα συμμεταβλητών δύο ατόμων. Εφόσον ισχύει η συνθήκη αναλογικών κινδύνων η διαφορά των συναρτήσεων $\log[-\log(\mathrm{S}(\mathrm{t}|\mathrm{x}_1))]$ και $\log[-\log(\mathrm{S}(\mathrm{t}|\mathrm{x}_2))]$ θα είναι σταθερή. Συνεπώς

$$\log\left[-\log\left(\mathrm{S}(\mathrm{t}|\mathrm{x}_1)\right)\right] = \log[-\log(\mathrm{S}(\mathrm{t}|\mathrm{x}_1))] + (\beta^\mathrm{T}\mathrm{x}_1 - \beta^\mathrm{T}\mathrm{x}_2).$$

Σχεδιάζοντας τις δύο καμπύλες στο ίδιο γράφημα θα παρατηρήσουμε ότι οι καμπύλες είναι παράλληλες και μεταξύ τους θα απέχουν σταθερή απόσταση και ίση με $\beta^\mathrm{T}\mathrm{x}_1 - \beta^\mathrm{T}\mathrm{x}_2$. Επομένως, πραγματοποιώντας αυτή την γραφική παράσταση έχουμε ένα πρώτο έλεγχο των αναλογικών κινδύνων. Αν οι καμπύλες που θα προκύψουν είναι παράλληλες ή σχεδόν παράλληλες τότε ισχύει η συνθήκη αναλογικών κινδύνων.

Κατά τον γραφικό έλεγχο των αναλογικών κινδύνων προκύπτουν διάφορα προβλήματα, όπως το να αποφανθούμε αν οι καμπύλες είναι παράλληλες. Επίσης, όταν το επίπεδο μια μεταβλητής έχει λίγους ή καθόλου πλήρεις διακεκριμένους χρόνους. Στην περίπτωση έλλειψης πλήρων χρόνων σε μια κατηγορία είναι αδύνατον να γίνει η γραφική παράσταση για την κατηγορία αυτή και είναι αδύνατος ο γραφικός έλεγχος της αναλογικότητας.

Επιπλέον πρόβλημα δημιουργείται κατά την διαδικασία της κατηγοριοποίησης μιας ποσοτικής μεταβλητής. Προτιμότερο και πιο λειτουργικό είναι να χρησιμοποιείται μικρός αριθμός κατηγοριών.

1.6.2 Έλεγχος των αναλογικών κινδύνων χρησιμοποιώντας εξαρτώμενες από το χρόνο μεταβλητές.

Ένας ακόμα έλεγχος της αναλογικότητας των κινδύνων πραγματοποείται με τη χρήση των εξαρτώμενων από το χρόνο μεταβλητών, δηλαδή μεταβλητών που ενώ είναι σταθερές μεταβλητές με κατάλληλο μετασχηματισμό γίνονται μεταβλητές εξαρτώμενες από το χρόνο.

Έστω ότι εξετάζουμε k μεταβλητές, τις $x_1, x_2, \dots, x_k$ και θέλουμε να εξετάσουμε την περίπτωση αν η σταθερή μεταβλητή x_k ικανοποιεί την υπόθεση της αναλογικότητας των κινδύνων, στην παρουσία των υπόλοιπων $k-1$ μεταβλητών. Ορίζουμε τον μετασχηματισμό $x_k(t) = x_k g(t)$, ουσιαστικά πολλαπλασιάζουμε την x_k με μια συνάρτηση του χρόνου και με τον τρόπο αυτό η x_k από σταθερή μεταβλητή μετατρέπεται σε εξαρτώμενη από το χρόνο. Συμβολίζουμε με το x^- να είναι το διάνυσμα των υπολοίπων $k-1$ μεταβλητών και θεωρούμε το Cox να έχει την ακόλουθη μορφή

$$h(t|x) = h_0(t)\exp\left(\beta_k x_k + \gamma x_k(t) + \beta^- x^-\right)$$

Με την εισαγωγή του μετασχηματισμού η παραπάνω εξίσωση γίνεται

$$h(t|x) = h_0(t)\exp\left(\beta_k x_k + \gamma x_k g(t) + \beta^- x^-\right)$$

με την τελευταία εξίσωσή γίνεται ο έλεγχος της μηδενικής υπόθεσης $H_0: \gamma = 0$. Αν η μηδενική υπόθεση γίνει δεκτή τότε η συμμεταβλητή x_k ικανοποιεί την συνθήκη της αναλογικότητας των κινδύνων, διαφορετικά συμπεραίνουμε ότι η x_k δεν ικανοποιεί την υπόθεση των αναλογικών κινδύνων. Μια μη-μηδενική τιμή του συντελεστή γ θα σήμαινε μια αλλαγή του κινδύνου μεταξύ δύο

ατόμων με διαφορετική τιμή του x_k στο χρόνο. Η αλλαγή αυτή εξαρτάται από την επιλεγμένη μορφή της συνάρτησης $g(t)$ και επειδή στόχος μας είναι η εξέταση της υπόθεσης της αναλογικότητας των κινδύνων και όχι η μοντελοποίηση της επίδρασης του x_k στο χρόνο, η επιλογή της συνάρτησης $g(t)$ περιορίζεται συνήθως σε απλές συναρτήσεις οπώς είναι η ταυτοτική $g(t) = t$ και η λογαριθμική $g(t) = \ln t$.

Στην περίπτωση που θέλουμε να εξετάσουμε αν οι μεταβλητές $x_1, x_{2,} \dots, x_k$ ικανοποιούν ταυτόχρονα την υπόθεση αναλογικών κινδύνων σχηματίζουμε το γενικευμένο μοντέλο

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

και κάνουμε τον έλεγχο της υπόθεση $H_0: \gamma_i = 0$. Αν η μηδενική υπόθεση γίνει δεκτή συμπεραίνουμε ότι οι μεταβλητές ικανοποιούν την υπόθεση των αναλογικών κινδύνων αλλιώς δεν την ικανοποιούν.

1.6.3 Υπόλοιπα Schoenfeld

Τα υπόλοιπα Schoenfeld τα πρότεινε το 1982 ο Schoenfeld για την εξέταση του αναλογικού κινδύνου στο ημι-παραμετρικό μοντέλο του Cox. Τα υπόλοιπα Schoenfeld υπερτερούν των υπολοίπων διότι ο υπολογισμός τους δεν απαιτεί την εκτίμηση της αθροιστικής αναφορικής συνάρτησης κινδύνου. Στα υπόλοιπα Schoenfeld υπολογίζεται ένα ξεχωριστό υπόλοιπο για κάθε άτομο για κάθε μεταβλητή σε αντίθεση με τα άλλα που υπολογίζεται μόνο ένα για κάθε άτομο. Άρα αν έχουμε p μεταβλητές τότε για κάθε άτομο υπολογίζονται p Schoenfeld υπόλοιπα. Το γράφημα τους συναρτήσει του χρόνου δίνει την δυνατότητα ελέγχου της υπόθεσης των αναλογικών κινδύνων. Στην

περίπτωση που τα υπόλοιπα είναι τυχαία διασκορπισμένα στο γράφημα ικανοποιείται η υπόθεση αλλιώς δεν ικανοποιείται η υπόθεση.

Τα υπόλοιπα Schoenfeld της μεταβλητής x_k ορίζονται ως η τιμή της συμμεταβλητής x_k για το i άτομο, δηλαδή η τιμή x_{ki} με πλήρη χρόνο t_i μείον την αναμενόμενη τιμή της συμμεταβλητής για τα άτομα που είναι σε κίνδυνο. Η σχέση είναι η ακόλουθη

$$\hat{r_j} = x_j - \hat{E}(x|R_j) \text{ με } \hat{E}(x|R_j) = \frac{\sum_{k \in R_j} x_k e^{\beta^T x_k}}{\sum_{i \in R_j} e^{\beta^T x_i}}$$

Πρέπει να επισημανθεί τα υπόλοιπα Schoenfeld προσδιορίζονται στους χρόνους διακοπής και όχι στις λογοκριμένες παρατηρήσεις. Επίσης αποτελούν διανύσματα διότι κάθε μη-λογοκριμένη παρατήρηση έχει τόσα υπόλοιπα όσες είναι και οι συμμεταβλητές.

ΚΕΦΑΛΑΙΟ 2

Μοντέλα Ευπάθειας

2.1 Εισαγωγή

Η μη παρατηρούμενη προδιάθεση για θάνατο, επίσης αποκαλούμενη και ως ‘’κρυμμένη ετερογένεια’’ (hidden heterogeneity) ή ευπάθεια (frailty) είναι μια σημαντική ανησυχία στη δημογραφική ανάλυση της επιβίωσης όπου ατομικές διαφοροποιήσεις στις πιθανότητες επιβίωσης δεν μπορούν να αγνοηθούν. Για την μελέτη τέτοιων διαφορών προτάθηκαν τα μοντέλα ευπάθειας. Τα μοντέλα ευπάθειας χρησιμοποιούνται για να εξηγήσουν την αποκλίνουσα συμπεριφορά των ποσοστών θνησιμότητας σε προχωρημένες ηλικίες (Vaupel and Yashin 1965), για να διορθώσουν προκατηλειμένες εκτιμήσεις των συντελεστών παλινδρόμησης σε μοντέλα κινδύνων τύπου Cox (Chamberlain 1985) και να διαχωρίσουν συνθετικές και βιολογικές επιδράσεις στις μελέτες γήρανσης (Manton et al, 1986). Τα μοντέλα ευπάθειας έχουν σπουδαίο ρόλο στην ερμηνεία αποτελεσμάτων σε πειράματα στρες (Yashin et al. 1996a) και σε μελέτες υπεραιωνόβιων (Yashin et al. 1999).

Η έννοια της ευπάθειας παρέχει έναν εύκολο τρόπο για να εισαχθούν τυχαίες επιδράσεις σύνθεσης και ετερογένειας σε μοντέλα για δεδομένα επιβίωσης. Στην απλούστερη μορφή της, η ευπάθεια είναι ένας μη παρατηρούμενος τυχαίος αναλογικός παράγοντας που τροποποιεί τη συνάρτηση κινδύνου του ατόμου ή των συσχετιζόμενων ατόμων.

Τα μοντέλα ευπάθειας είναι επεκτάσεις του μοντέλου αναλογικών κινδύνων γνωστό ως Cox μοντέλο. Συνήθως, στις περισσότερες κλινικές εφαρμογές υποθέτουμε σιωπηρά ότι ο εξεταζόμενος πληθυσμός είναι ομοιογενής. Με

άλλα λόγια ότι όλα τα άτομα που ανήκουν στο πληθυσμό διατρέχουν τον ίδιο κίνδυνο (κίνδυνο επανεμφάνισης νόσου, κίνδυνο θανάτου κτλ). Η υπόθεση ομοιογένειας του πληθυσμού πολλές φορές είναι ακατάλληλη και απαιτείται ο πληθυσμός να θεωρηθεί ετερογενής, δηλαδή ένας πληθυσμός με άτομα με διαφορετικούς κινδύνους. Για παράδειγμα είναι αδύνατο αρκετές φορές να μετρηθούν όλοι οι σχετικοί παράγοντες που επηρεάζουν την νόσο, είτε για οικονομικούς λόγους είτε η επίδραση κάποιων παραγόντων είναι άγνωστη. Ουσιαστικά η ευπάθεια είναι ο στατιστικός όρος μοντελοποίησης της ετερογένειας που προκαλείται από τυχόν μη μετρούμενες συμμεταβλητές.

Από στατιστικής πλευράς το μοντέλο ευπάθειας είναι ένα τυχαίο μοντέλο στο οποίο η ευπάθεια έχει πολλαπλασιαστικό χαρακτήρα στην αναφορική συνάρτηση κινδύνου.

Μπορούμε να διακρίνουμε δυο μεγάλες κατηγορίες μοντέλων ευπάθειας

1) Μοντέλα με μονομεταβλητό χρόνο επιβίωσης ως τελικό σημείο.

2) Μοντέλα που περιγράφουν πολυμεταβλητά τελικά σημεία επιβίωσης (ανταγωνιστικοί κίνδυνοι, επανεμφάνιση των γεγονότων στο ίδιο άτομο, εμφάνιση μιας ασθένειας σε συγγενείς).

Στην πρώτη κατηγορία μοντέλων, μια μονομεταβλητή διάρκεια ζωής χρησιμοποιείται για να περιγράψει την επίδραση των μη παρατηρούμενων συμμεταβλητών σε ένα μοντέλο αναλογικού κινδύνου. Η μεταβλητότητα των δεδομένων επιβίωσης είναι διαιρεμένη στο μέρος που εξαρτάται από τους παράγοντες κινδύνου και ως εκ τούτου θεωρητικά προβλέψιμη και σε ένα μέρος που είναι αρχικά απρόβλεπτο, ακόμα και όταν όλες οι σχετικές πληροφορίες είναι γνωστές. Ο διαχωρισμός αυτός έχει το πλεονέκτημα η ετερογένεια να μπορεί να εξηγήσει απροσδόκητα αποτελέσματα ή να δώσει

μια εναλλακτική ερμηνεία σε διάφορα αποτελέσματα. Για παράδειγμα, η ευπάθεια μπορεί να ερμηνεύσει διασταυρό-επικαλυπτώμενες επιδράσεις ή την σύγκλιση των συναρτήσεων κινδύνου από δύο διαφορετικά σκέλη της θεραπείας (Manton and Stallard (1981)).

Στα μοντέλα με πολυμεταβλητούς χρόνους επιβίωσης ο στόχος είναι να ληφθεί υπόψη η εξάρτηση στο σύμπλεγμα των χρόνων των γεγονότων. Ένας φυσικός τρόπος μοντελοποίησης της εξάρτησης ανάμεσα στους χρόνους των γεγονότων είναι μέσω της εισαγωγής μια τυχαίας επίδρασης, της ευπάθειας. Αυτή η τυχαία επίδραση εξηγεί την εξάρτηση υπό την έννοια ότι αν γνωρίζαμε την ευπάθεια, τότε τα γεγονότα θα ήταν ανεξάρτητα. Με άλλα λόγια οι διάρκειες ζωής είναι υπό συνθήκη ανεξάρτητες, για δεδομένη ευπάθεια. Αυτή η προσέγγιση μπορεί να χρησιμοποιηθεί για χρόνους επιβίωσης συσχετιζόμενων ατόμων όπως συγγενείς ή για επαναλαμβανόμενες παρατηρήσεις για το ίδιο άτομο.

2.2 Univariate frailty models (Μονομεταβλητά μοντέλα ευπάθειας)

Η συνήθης μεθοδολογία κατά την χρησιμοποίηση της ανάλυσης επιβίωσης σε κλινικά ερευνητικά προγράμματα είναι η υπόθεση ενός ομοιογενούς πληθυσμού που ερευνάται υπό διαφορετικές συνθήκες (π.χ. πειραματική θεραπεία, καθιερωμένη θεραπεία). Το κατάλληλο μοντέλο επιβίωσης υποθέτει ότι τα δεδομένα επιβίωσης των διαφόρων ασθενών είναι ανεξάρτητα μεταξύ τους και η κατανομή του χρόνου επιβίωσης του κάθε ασθενή είναι η ίδια.

Η υπόθεση της ομοιογένειας του πληθυσμού αμφισβητείται εύκολα. Στις περισσότερες κλινικές δοκιμές παρατηρεί κανείς σε πολλές πρακτικές περιπτώσεις ότι οι ασθενείς διαφέρουν σημαντικά. Η επίδραση ενός

φαρμάκου η μιας θεραπείας μπορεί να διαφέρει σημαντικά μεταξύ των υποομάδων των ασθενών. Αυτό συμβαίνει διότι τα άτομα έχουν διαφορετικές αδυναμίες και οι πιο αδύναμοι οργανισμοί θα πεθάνουν πιο γρήγορα ή δε θα ανταπεξέλθουν στη θεραπεία τόσο καλά όσο κάποιοι δυνατοί οργανισμοί. Όταν εκτιμούνται τα ποσοστά θνησιμότητας το ενδιαφέρον έγκειται στο πώς αλλάζουν αυτά με την πάροδο του χρόνου και την ηλικία. Αρκετά συχνά παρατηρείται η συνάρτηση κινδύνου να αυξάνεται, να φτάνει ένα μέγιστο και έπειτα να μειώνεται. Όσο περισσότερο ζει ο ασθενής μετά την εμφάνιση της νόσου, βελτιώνονται οι πιθανότητές του για επιβίωση. Η πληθυσμιακή ένταση μπορεί να ξεκινήσει να μειώνεται γιατί οι υψηλού κινδύνου ασθενείς απεβίωσαν. Το ποσοστό κινδύνου ενός ατόμου όμως μπορεί κάλλιστα να συνεχίσει να αυξάνεται. Εάν οι παράγοντες κινδύνου είναι γνωστοί, τα παραπάνω μπορούν να συμπεριληφθούν στην ανάλυση με τη χρήση του μοντέλου αναλογικών κινδύνων που έχει την ακόλουθη μορφή

$$\mathrm{h}(\mathrm{t}|\mathrm{X}=\mathrm{x})=\mathrm{h}_0(\mathrm{t})\mathrm{e}^{\beta^{\mathrm{T}}\mathrm{x}}$$

όπου $\mathrm{h}_0(\mathrm{t})$ η αναφορική συνάρτηση κινδύνου η οποία υποτίθεται ότι είναι μοναδική για όλα τα άτομα του εξεταζόμενου πληθυσμού, x το διάνυσμα των παρατηρούμενων συμμεταβλητών και β των διάνυσμα των αντίστοιχων παραμέτρων παλινδρόμησης.

Δυο είναι οι βασικοί λόγοι για τους οποίους δεν μπορούμε να συμπεριλάβουμε όλους τους σημαντικούς παράγοντες στην παραπάνω ανάλυση. Πρώτον οι συμμεταβλητές που πρέπει να εξεταστούν είναι πάρα πολλές και δεύτερον ο ερευνητής μπορεί να μην ξέρει ή να μη μπορεί να μετρήσει όλες τις σχετικές μεταβλητές. Η μεταβλητότητα και στις δύο περιπτώσεις διακρίνεται στην μεταβλητότητα που μετριέται από τους

παράγοντες του ρίσκου και θεωρητικά είναι μετρήσιμη και την ετερογένεια που οφείλεται στις άγνωστες συμμεταβλητές και θεωρητικώς είναι απρόβλεπτη.

Σε ένα μοντέλο αναλογικών κινδύνων η εξαίρεση κάποιου υποσυνόλου μεταβλητών οδηγεί σε μεροληπτικές εκτιμήτριες των παραμέτρων της παλινδρόμησης και του ποσοστού κινδύνου. Ο λόγος για αυτό είναι ότι το εξαρτώμενο από το χρόνο ποσοστό κινδύνου αλλάζει καθώς η σύνθεση του πληθυσμού αλλάζει σε σχέση με τις συμμεταβλητές. Μια εκτίμηση του κινδύνου χωρίς να ληφθεί υπόψη η μη παρατηρούμενη ευπάθεια θα οδηγήσει σε υποεκτίμηση της συνάρτησης κινδύνου και η έκταση της υποτίμησης θα αυξάνεται καθώς ο χρόνος προχωρεί.

Το μονομεταβλητό μοντέλο ευπάθειας επεκτείνει το μοντέλο του Cox έτσι ώστε ο κίνδυνος του ατόμου να εξαρτάται από μια επιπλέον τυχαία μεταβλητή Z, η οποία ενεργεί πολλαπλασιαστικά στην αναφορική συνάρτηση κινδύνου.

$$h(t|z,x) = zh_0(t)e^{\beta^T X}$$

πάλι $h_0(t)$ η αναφορική συνάρτηση κινδύνου, X το διάνυσμα των συμμεταβλητών και Z η μεταβλητή ευπάθειας. Η ευπάθεια Z είναι μία τυχαία μεταβλητή που μεταβάλλεται μέσω του πληθυσμού, μειώνει τον ατομικό κίνδυνο όταν $Z < 1$ ή τον αυξάνει όταν $Z > 1$. Πρέπει να τονιστεί ότι η ευπάθεια είναι κάτι που δεν παρατηρείται. Η αντίστοιχη συνάρτηση επιβίωσης S ,που περιγράφει το σύνολο των επιζώντων ατόμων στον πληθυσμό της μελέτης δίνεται από τον ακόλουθο τύπο

$$S(t|Z = z, X = x) = e^{-ze^{\beta^T x}\int_0^t h_0(s)ds} = e^{-ze^{\beta^T x}H_0(t)}$$

όπου $S(t|z,x)$ ερμηνεύεται ως το ποσοστό των ατόμων που έχει επιζήσει μέχρι την χρονική στιγμή t και $H_0(t)$ η αθροιστική αναφορική συνάρτηση κινδύνου.

Μέχρι τώρα το μοντέλο έχει περιγραφεί σε ατομικό επίπεδο. Το ατομικό αυτό μοντέλο δεν παρατηρείται. Συνεπώς, πρέπει το μοντέλο να εξεταστεί σε επίπεδο ολόκληρου του εξεταζόμενου πληθυσμού. Η συνάρτηση επιβίωσης όλου του πληθυσμού είναι ο μέσος όρος των ατομικών συναρτήσεων επιβίωσης. Η συνάρτηση κινδύνου όλου του πληθυσμού μπορεί να διαφέρει από την ατομική συνάρτηση κινδύνου. Η συνάρτηση επιβίωσης του πληθυσμού λαμβάνεται από την $S(t|z,x)$ ολοκληρώνοντας ως προς την ευπάθεια.

$$S(t|x) = \int_0^\infty e^{-ze^{\beta^T x}H_0(t)}dF_z(z) \equiv e^{-G(e^{\beta^T x}H_0(t))}$$

όπου F_z η συνάρτηση κατανομής της ευπάθειας. Για να ισχύει η ισότητα πρέπει το $G(w) = -\ln\left(\int_0^\infty e^{-wz}dF_z(z)\right)$. Για διαφορετική κατανομή ευπάθειας παράγεται διαφορετική G. Η συνάρτηση πιθανοφάνειας για δεξιά λογοκριμένα δεδομένα γράφεται

$$\text{Likelihood} = \prod_{\text{uncensored}} f(x_i) \prod_{\text{censored}} S(x_i)$$

Δηλαδή η συνάρτηση πιθανοφάνειας χωρίζεται στα δύο, στα δεδομένα που δεν είναι λογοκριμένα και σε αυτά που είναι λογοκριμένα. Για τα δεδομένα έχουμε ότι T_i είναι οι χρόνοι αποτυχίας, C_i οι λογοκριμένοι χρόνοι και τα $W_i = \min(T_i, C_i), i = 1, \dots, n$. θεωρούμε την δείκτρια συνάρτηση

$$\Delta_i = \begin{cases} 1, & T_i \le C_i \\ 0, & T_i > C_i \end{cases}$$

Εισάγοντας την δείκτρια συνάρτηση στην πιθανοφάνεια έχουμε

$$\mathrm{Lik} = \prod_{i=1}^{n} f(W_i|X_i)^{\Delta_i} S(W_i|X_i)^{1-\Delta_i} = \prod_{i=1}^{n} \left(\frac{f(W_i|X_i)}{S(W_i|X_i)}\right)^{\Delta_i} S(W_i|X_i)$$

Παραγωγίζοντας την συνάρτηση επιβίωσης μεταβαίνουμε στην συνάρτηση πυκνότητας πιθανότητας.

$$S(t|x) = e^{-G\left(e^{\beta^T x} H_0(t)\right)} \Rightarrow f(t|x) = e^{-G\left(e^{\beta^T x} H_0(t)\right)} G'(e^{\beta^T x} H_0(t)) e^{\beta^T X} h_0(t)$$

Επομένως η πιθανοφάνεια γίνεται

$$\mathrm{Lik} \quad = \prod_{i=1}^{n} (G'(\, e^{\beta^T X_i} H_0(W_i)) e^{\beta^T X_i} h_0(W_i))^{\Delta_i} \, e^{-G\left(e^{\beta^T X_i} H_0(W_i)\right)}$$

Τελικώς έχουμε την συνάρτηση πιθανοφάνειας συναρτήσει της συνάρτησης G και των δεδομένων $\Delta_{i,} W_{i,} X_i$ και άγνωστες παραμέτρους τα β και τα $h_0(t)$. Υποθέτοντας μια κατανομή για την ευπάθεια και κατά συνέπεια μια συγκεκριμένη συνάρτηση G μπορούμε με το συνήθη τρόπο να βρούμε τις εκτιμήσεις των παραμέτρων εφόσον έχουμε μια αναλυτική μορφή. Στην περίπτωση που αυτό δεν καθίσταται εφικτό εφαρμόζονται πιο εξελιγμένες προσεγγίσεις όπως αριθμητική ολοκλήρωση ή Monte Carlo μέθοδοι. Οι πιο συχνά χρησιμοποιούμενες κατανομές ευπάθειας είναι η Γάμμα κατανομή η Λογαριθμοκανονική κατανομή και η αντίστροφη Gaussian κατανομή.

2.2.1 Γάμμα μοντέλο ευπάθειας

Η Γάμμα κατανομή ως μεικτή κατανομή υπολογιστικά και αναλυτικά ταιριάζει πολύ καλά σε δεδομένα αποτυχίας. Εύκολα αντλούνται οι κλειστές μορφές της συνάρτησης κινδύνου, της αθροιστικής και της μη δεσμευμένης συνάρτησης επιβίωσης που οφείλεται στην απλότητα του μετασχηματισμού Laplace. Η Γάμμα κατανομή $g(k, \lambda)$ είναι μία ευέλικτη κατανομή που παίρνει μια ποικιλία από σχήματα καθώς η παράμετρος της k μεταβάλλεται π.χ. αν

$k = 1$ είναι πανομοιότυπη με την εκθετική, ενώ όταν το k είναι αρκετά μεγάλο το σχήμα της θυμίζει κανονική κατανομή.

Παρ' όλα τα πλεονεκτήματα αυτά δεν συντρέχει κάποιος βιολογικός λόγος που κάνει την κατανομή Γάμμα προτιμότερη από κάποιες άλλες κατανομές ευπάθειας. Μαθηματικοί και υπολογιστικοί είναι οι λόγοι που την καθιστούν ευρέως χρησιμοποιούμενη. Το άρθρο των Abbring και Van den Berg (2007) εκλογικεύει την χρήση της κατανομής για τις ευπάθειες. Οι συγγραφείς έδειξαν ότι σε μια μεγάλη κλάση των μονομεταβλητών μοντέλων ευπάθειας, η κατανομή ευπάθειας ανάμεσα στους επιζώντες συγκλίνει προς μια κατανομή Γάμμα καθώς ο χρόνος πηγαίνει στο άπειρο κάτω από ήπιες προϋποθέσεις κανονικότητας.

Η ευπάθεια δεν μπορεί να είναι αρνητική, η Γάμμα και η Λογαριθμό-κανονική κατανομή είναι οι πιο συνηθισμένες κατανομές που μοντελοποιούν μεταβλητές μη αρνητικές. Η συνάρτηση πυκνότητας πιθανότητας της Γάμμα δίνεται από τον τύπο

$$f(x) = \frac{1}{\Gamma(k)}\lambda^k \chi^{k-1} e^{-\lambda x} \;\; x \geq 0, k > 0, \lambda > 0$$

στο απλοποιημένο μοντέλο $h(t|z) = zh_0(t)$ η συνάρτηση επιβίωσης είναι $S(t|z) = e^{-zH_0(t)}$ και η συνάρτηση επιβίωσης του πληθυσμού είναι ο μέσο-σταθμικός μέσος των δεσμευμένων συναρτήσεων επιβίωσης

$$S(t) = E(S(t|z)) = E\left(e^{-zH_0(t)}\right) = L(H_0(t))$$

όπου L συμβολίζει τον μετασχηματισμό Laplace ο οποίος για τη Γάμμα κατανομή δίνεται από

$$L(u) = \frac{1}{\Gamma(k)}\lambda^k \int e^{-ux} x^{k-1} e^{-\lambda x}\, dx$$

$$= \frac{\lambda^k}{(\lambda+u)^k}(\lambda+u)^k \frac{1}{\Gamma(k)} \int x^{k-1} e^{-(\lambda+u)x} dx$$

$$= (1 + \frac{u}{\lambda})^{-k}$$

Άρα η συνάρτηση επιβίωσης θα δίνεται από την σχέση

$$S(t) = L\big(H_0(t)\big) = (1 + \frac{H_0(t)}{\lambda})^{-k}$$

και η συνάρτηση κινδύνου

$$h(t) = k\frac{h_0(t)}{\lambda + H_0(t)}$$

Επεκτείνοντας το μοντέλο εισάγοντας τον όρο του Cox ($e^{\beta^T X}$) η ευπάθεια παραμένει Γάμμα με διαφορετικές όμως παραμέτρους. Για το μοντέλο $h(t|z,x) = zh_0(t)e^{\beta^T X}$

η συνάρτηση πιθανοφάνειας στην παραμετρική περίπτωση είναι της μορφής

$$\mathrm{Lik}(\beta, \theta, Z_i) = \prod_{i=1}^{n} (z_i h_0(t_i|\theta) e^{\beta^T X_i})^{\Delta_i}\, e^{-z_i H_0(t_i|\theta)} e^{\beta^T X_i}$$

όπου θ το διάνυσμα των παραμέτρων της αναφορικής συνάρτησης.

Η παραπάνω πιθανοφάνεια εξαρτάται από τις άγνωστες ευπάθειες των ατόμων. Θεωρώντας για την αναφορική συνάρτηση κινδύνου ότι ακολουθεί Γάμμα κατανομή με παραμέτρους $(k, \lambda) = (\frac{1}{\sigma^2}, \frac{1}{\sigma^2})$ παίρνουμε την παραμετρική μορφή του Γάμμα μοντέλου ευπάθειας και η συνάρτηση πιθανοφάνειας παίρνει την ακόλουθη μορφή

$$\mathrm{Lik}(\beta, \theta, \sigma^2) = \prod_{i=1}^{n} \left(\frac{h_0(t_i|\theta) e^{\beta^T X_i}}{1 + \sigma^2 H_0(t_i|\theta) e^{\beta^T X_i}} \right)^{\Delta_i} (1 + \sigma^2 H_0(t_i|\theta) e^{\beta^T X_i})^{-\frac{1}{\sigma^2}}$$

Στην περίπτωση που δεν γίνεται καμία υπόθεση για την κατανομή της αναφορικής συνάρτησης έχουμε την ημι-παραμετρική μορφή του Γάμμα μοντέλου ευπάθειας. Στο μοντέλο αυτό η συνάρτηση $h_0(t)$ θεωρείται ως άγνωστη μεταβλητή. Ο EM (expectation-maximization) αλγόριθμος

χρησιμοποιείται για τον υπολογισμό των εκτιμητριών στο ημιπαραμετρικό Γάμμα μοντέλο ευπάθειας. Ο αλγόριθμος αυτός προτάθηκε από τον Dempster et al.(1970) και συχνά χρησιμοποιείται όταν υπάρχουν μη παρατηρούμενα δεδομένα. Πρώτος ο Nielsen et al.(1992) υιοθέτησε τον αλγόριθμο για την εκτίμηση παραμέτρων σε μοντέλα ευπάθειας. Ο αλγόριθμος έχει δυο βασικά στάδια. Στο πρώτο στάδιο γίνεται η εκτίμηση των μη παρατηρούμενων ευπαθειών βασισμένη στα παρατηρούμενα δεδομένα. Οι εκτιμήσεις του πρώτου σταδίου χρησιμοποιούνται στο δεύτερο στάδιο μεγιστοποίησης για να πάρουμε νέες εκτιμήσεις, δεδομένου των εκτιμώμενων ευπαθειών. Για το συγκεκριμένο Γάμμα μοντέλο υπάρχουν οι κλειστές μορφές των δεσμευμένων αναμενόμενων ευπαθειών.

Αρχικώς θεωρούμε την πλήρη συνάρτηση πιθανοφάνειας και τις μεταβλητές ευπάθειας τις υποθέτουμε ως παρατηρούμενες τυχαίες μεταβλητές συναφείς με τους χρόνους των γεγονότων. Η πλήρης πιθανοφάνεια έχει την μορφή

$$\begin{aligned}\mathrm{Lik}(\beta,\sigma^2|\mathrm{z}) &= \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{f}(\Delta_{\mathrm{i}},\mathrm{t}_{\mathrm{i}},\mathrm{z}_{\mathrm{i}},\beta,\sigma^2)\\ &= \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{f}(\Delta_{\mathrm{i}},\mathrm{t}_{\mathrm{i}},\mathrm{z}_{\mathrm{i}}|\beta)\prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{f}(\mathrm{z}_{\mathrm{i}}|\sigma^2)\\ &= \mathrm{Lik}_1(\beta|\mathrm{z})\mathrm{Lik}_2(\sigma^2|\mathrm{Z})\end{aligned}$$

με $\mathrm{Z} = (\mathrm{Z}_1, \mathrm{Z}_2 \ldots \mathrm{Z}_{\mathrm{n}})$ και

$$\mathrm{Lik}_1(\beta|\mathrm{z}) = \prod_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{z}_{\mathrm{i}}\mathrm{h}_0(\mathrm{t}_{\mathrm{i}})\mathrm{e}^{\beta^{\mathrm{T}}\mathrm{X}_{\mathrm{i}}})^{\Delta_{\mathrm{i}}}\, \mathrm{e}^{-\mathrm{z}_{\mathrm{i}}\mathrm{H}_0(\mathrm{t}_{\mathrm{i}})\mathrm{e}^{\beta^{\mathrm{T}}\mathrm{X}_{\mathrm{i}}}}$$

$$\mathrm{Lik}_2 = \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{f}(\mathrm{z}_{\mathrm{i}}|\sigma^2)$$

Με τον τρόπο αυτό παίρνουμε εκτιμήτριες των ευπαθειών που χρησιμοποιούνται στο στάδιο της μεγιστοποίησης για να πάρουμε τις εκτιμήσεις των παραμέτρων της παλινδρόμησης. Η μερική πιθανοφάνεια για τις παραμέτρους της παλινδρόμησης στο μοντέλο ευπάθειας είναι της μορφής

$$\mathrm{Lik}(\beta|z) = \prod_{i=1}^{n} \left(\frac{e^{\beta^T X_i + \log(z_i)}}{\sum_{j \in R_{(t_i)}} z_j e^{\beta^T X_i}}\right)^{\Delta_i}$$

Οι άγνωστες τυχαίες μεταβλητές αντικαθίστανται με τις πρόσφατες μέσες τιμές στην k-οστή επανάληψη $E_{(k)}z_i$ και $E_{(k)}\log(z_i)$

$$\log \mathrm{Lik}(\beta, \sigma^2) = \sum_{i=1}^{n} \Delta_i \left[\beta^T X_i + E_{(k)}(\log(z_i)) - \log\left(\sum_{j \in R(t_i)} E_{(k)}(z_j) e^{\beta^T X_j}\right)\right]$$

Από την παραπάνω έκφραση αυτή παίρνουμε καινούργιες εκτιμήσεις $\beta_{(k)}$. Επίσης καινούργια εκτίμηση για την παράμετρο της ευπάθειας $\sigma^2_{(k)}$ παίρνουμε μεγιστοποιώντας την $\mathrm{Lik}_2(\sigma^2|z)$ έχοντας πρώτα αντικαταστήσει τις άγνωστες μεταβλητές Z_i με τις πρόσφατες μέσες τιμές του επαναληπτικού βήματος k.

Η μη παρατηρούμενη ευπάθεια Z_i $(i = 1, .., n)$ του κάθε ατόμου μπορεί να εκτιμηθεί από τον τύπο

$$E_{(k+1)}(Z_i) = \frac{\Delta_i + \frac{1}{\sigma^2_{(k)}}}{\frac{1}{\sigma^2_{(k)}} + H_{(k)}(t_i) e^{\beta^T_{(k)} X_i}}$$

όπου $H_{(k)}(.)$ είναι μια μη παραμετρική εκτιμήτρια της αθροιστικής συνάρτησης αναφοράς κινδύνου βασισμένη στην πρόσφατη εκτίμηση της επανάληψης k.

Η αποτελεσματικότητα του EM αλγορίθμου ερευνήθηκε από τους Barker and Henderson (2005) στην εφαρμογή του στο μονομεταβλητό Γάμμα μοντέλο ευπάθειας. Παρατήρησαν μεγάλη υποεκτίμηση της διασποράς της ευπάθειας. Οι αντίστοιχοι συντελεστές παλινδρόμησης υποεκτιμούνται σε απόλυτες τιμές. Για την διόρθωση αυτής της υποεκτίμησης οι Barker and Henderson πρότειναν μια προσαρμογή, αντικατέστησαν την μη παραμετρική εκτιμήτρια

του Breslow της συνάρτησης κινδύνου με μία τοπική πιθανοφάνεια για την αναφορική συνάρτηση κινδύνου που επιτρέπει στους χρόνους επιβίωσης να εισέλθουν στους όρους εκτίμησης.

2.2.2 Inverse Gaussian frailty model

Η αντίστροφη κανονική ή Gaussian κατανομή εισήχθη ως μια κατανομή ευπάθειας από τον Hougaard (1984) και χρησιμοποιήθηκε από αρκετούς ερευνητές όπως από τον Manton et al. (1986), Vonta (2004), Economou and Caroni (2005), Kheiriet al. (2007), Duchateau and Janssen (2008) και Androulakis et al. (2011). Υπάρχει η κλειστή μορφή της μη δεσμευμένης συνάρτησης επιβίωσης και κινδύνου, που κάνει το μοντέλο αρκετά ελκυστικό. Η συνάρτηση πυκνότητας πιθανότητας της αντίστροφης Gaussian τυχαίας μεταβλητής με παραμέτρους $\mu > 0$ και $\lambda > 0$ δίνεται από τη σχέση.

$$f(z) = \frac{\sqrt{\lambda}}{\sqrt{2\pi z^3}} \exp\left(-\frac{\lambda}{2\mu^2 z}(z-\mu)^2\right)$$

Ο μετασχηματισμός Laplace της αντίστροφης Gaussian κατανομής δίνεται

$$L(u) = Ee^{-uZ}$$

$$= \int \frac{\sqrt{\lambda}}{\sqrt{2\pi z^3}} e^{-uz} \exp\left(-\frac{\lambda}{2z\mu^2}(z-\mu)^2\right) dz$$

$$= \int \frac{\sqrt{\lambda}}{\sqrt{2\pi z^3}} \exp\left(-\frac{(\lambda+2\mu^2 u)z^2 - 2\mu\lambda z + \lambda\mu^2}{2\mu^2 z}\right) dz$$

$$= \exp\left(-\frac{\lambda\sqrt{1+\frac{2\mu^2 u}{\lambda}}}{\mu} + \frac{\lambda}{\mu}\right) \times \int \frac{\sqrt{\lambda}}{\sqrt{2\pi z^3}} \exp\left(-\frac{\lambda z}{2} * \frac{1+\frac{2\mu^2 u}{\lambda}}{\mu^2} + \frac{\lambda\sqrt{1+\frac{2\mu^2 u}{\lambda}}}{\mu} - \frac{\lambda}{2z}\right) dz$$

Με διάφορες απλοποιήσεις παίρνουμε την απλοποιημένη μορφή του μετασχηματισμού Laplace

$$L(u) = \exp\left(-\frac{\lambda\sqrt{1+\frac{2\mu^2 u}{\lambda}}}{\mu}+\frac{\lambda}{\mu}\right) = \exp\left(\frac{\lambda}{\mu}\left(1-\sqrt{1+\frac{2\mu^2 u}{\lambda}}\right)\right)$$

Η μη δεσμευμένη συνάρτηση επιβίωσης και κινδύνου δίνονται από τις ακόλουθες συναρτήσεις.

$$S(t) = e^{\frac{1}{\sigma^2}(1-\sqrt{1+2\sigma^2 H_0(t)})}$$

$$h(t) = \frac{h_0(t)}{(1+2\sigma^2 H_0(t))^{1/2}}$$

Στόχος μας είναι να αξιολογήσουμε την επίδραση της μη παρατηρούμενης ετερογένειας των παραμέτρων παλινδρόμησης στο αντίστροφο Gaussian μοντέλο ευπάθειας. Έστω οι οριακοί κίνδυνοι δύο ατόμων με μία δίτιμη συμμεταβλητή στο μοντέλο. Ο λόγος κινδύνου δύο ατόμων με τιμές της συμμεταβλητής 0 και 1 είναι της μορφής.

$$\frac{h(t|X=1)}{h(t|X=0)} = \frac{(1+2\sigma^2 H_0(t))^{1/2}}{(1+2\sigma^2 H_0(t)e^{\beta})^{1/2}}e^{\beta}$$

Την χρονική στιγμή $t = 0$ ο παραπάνω λόγος παίρνει την τιμή e^{β} και παίρνει την τιμή $e^{\beta/2}$ καθώς ο χρόνος απειρίζεται $t \to \infty$. Η επίδραση των συμμεταβλητών καθώς ο χρόνος περνάει εξασθενεί. Επίσης η επίδραση της μη παρατηρούμενης ετερογένειας είναι εντονότερη σε περιπτώσεις με μεγάλες τιμές για τα σ^2,β,$H_0(t)$. Η συνάρτηση πυκνότητας πιθανότητας της ευπάθειας ανάμεσα στους επιζώντες την χρονική στιγμή t μπορεί να γραφτεί στην μορφή

$$f(z|x, T > t) = \frac{S(t|x,z)f(z)}{S(t|x)} = \frac{\exp\left(-zH_0(t)e^{\beta^T x}\right)\exp\left(-\frac{(z-1)^2}{2\sigma^2 z}\right)}{\sqrt{2\pi^2 z^3}\exp\left(\frac{1}{\sigma^2}\left(1-\sqrt{1+2\sigma^2 H_0(t)e^{\beta^T x}}\right)\right.}$$

$$= \frac{1}{\sqrt{2\pi^2 z^3}}\exp\left(-\frac{\left(z-\left(1+2\sigma^2 H_0(t)e^{\beta^T x}\right)^{-1/2}\right)^2}{\frac{2\sigma^2 z}{1+2\sigma^2 H_0(t)e^{\beta^T x}}}\right)$$

Η μέση τιμή ανάμεσα στους επιζώντες είναι

$$E(Z|x, T > t) = \frac{1}{\sqrt{1+\sigma^2 H_0(t)e^{\beta^T x}}}$$

και η διασπορά είναι

$$V(Z|x, T > t) = \frac{\sigma^2}{(1+\sigma^2 H_0(t)e^{\beta^T x})^2}$$

2.3 Multivariate frailty models (Πολυμεταβλητά μοντέλα ευπάθειας)

Τα μοντέλα ευπάθειας βρίσκουν σημαντική εφαρμογή στο πεδίο των πολυμεταβλητών δεδομένων επιβίωσης. Τέτοιου είδους δεδομένα θεωρούνται ότι είναι διάρκειες ζωής (οι στιγμές από την εμφάνιση μιας ασθένειας) συγγενών (δίδυμα, γονείς–παιδιά) ή επαναλαμβανόμενα γεγονότα όπως λοιμώξεις στο ίδιο άτομο. Σε τέτοιες περιπτώσεις η ανεξαρτησία μεταξύ των ομαδοποιημένων δεδομένων είναι αδύνατο να ισχύει. Τα πολυμεταβλητά μοντέλα έχουν την επεξηγηματική δυνατότητα για να αποσαφηνίσουν την παρουσία της εξάρτησης μεταξύ των χρόνων των γεγονότων. Η εξάρτηση στα πολυμεταβλητά μοντέλα εκφράζεται από μια λανθάνουσα μεταβλητή (latent variable) στα δεσμευμένα μοντέλα για πολλαπλούς παρατηρούμενους

χρόνους επιβίωσης, για παράδειγμα $S(t_1|z, x_1)$ και $S(t_2|z, x_2)$ είναι οι δεσμευμένες συναρτήσεις επιβίωσης δύο συσχετιζόμενων ατόμων με διαφορετικά διανύσματα παρατηρούμενων μεταβλητών X_1 και X_2 αντίστοιχα. Υποθέτοντας μια κατανομή (Γάμμα,Λογαριθμο-κανονική) για την λανθάνουσα μεταβλητή και ολοκληρώνοντας τότε επάγεται ένα πολυμεταβλητό μοντέλο για τα παρατηρούμενα δεδομένα. Στην περίπτωση των παρατηρήσεων σε ζεύγη η δισδιάστατη συνάρτηση επιβίωσης είναι της μορφής

$$S(t_{1,}t_2) = \int_0^\infty S(t_1|z, x_1)S(t_2|z, x_2)g(z)dz$$

όπου το g ορίζει την πυκνότητα της ευπάθειας Z. Στην περίπτωση των δίδυμων η $S(t_{1,}t_2)$ δηλώνει το κλάσμα των δίδυμων όπου το πρώτο επιζεί στο χρόνο $t_{1,}$ ενώ το δεύτερο στο χρόνο t_2.

Τα μοντέλα ευπάθειας για πολυμεταβλητά δεδομένα επιβίωσης προέρχονται υπό την προϋπόθεση της δεσμευμένης ανεξαρτησίας καθορίζοντας λανθάνουσες μεταβλητές που δρουν πολλαπλασιαστικά στην συνάρτηση αναφοράς.

2.4 The shared frailty model

Το από κοινού (shared) μοντέλο ευπάθειας έχει σχέση με χρόνους γεγονότων συσχετιζόμενων ατόμων και με επαναλαμβανόμενες παρατηρήσεις. Το μοντέλο αυτό παίρνει την ονομασία του από το γεγονός ότι τα άτομα σε μια ομάδα υποτίθεται ότι μοιράζονται την ίδια ευπάθεια Z. Εισήχθη από τον Clayton (1978) και διεξοδικά μελετήθηκε από τον Hougaard (2000). Το από κοινού μοντέλο ευπάθειας υποθέτει ότι όλοι οι χρόνοι αποτυχίας σε μια ομάδα είναι υπό όρους ανεξάρτητοι δεδομένης της ευπάθειας. Η τιμή του όρου της

ευπάθειας είναι σταθερή κατά την διάρκεια του χρόνου και κοινή για όλα τα άτομα στην ομάδα, και έτσι είναι υπεύθυνη για την δημιουργία εξάρτησης μεταξύ των χρόνων των γεγονότων σε μια ομάδα.

Οι χρόνοι των γεγονότων μεταξύ των ομάδων θεωρούνται ότι είναι ανεξάρτητοι. Ο αριθμός των παρατηρήσεων σε μια ομάδα υποτίθεται γνωστός. Τυπικό παράδειγμα διμεταβλητών δεδομένων αποτελούν μελέτες για δίδυμα. Ένα ακόμα παράδειγμα είναι οι χρόνοι αποτυχίας για παρόμοια ανθρώπινα όργανα όπως ο χρόνος τύφλωσης του δεξιού και του αριστερού ματιού σε μελέτες διαβητικής αμφιβληστροειδοπάθειας. Συνήθως εμφανίζεται σε οικογενειακές μελέτες το φαινόμενο των μικρών και άνισων ομάδων. Παράδειγμα μέτριων και μεγάλου μεγέθους ομάδων είναι χρόνοι γεγονότων σε πολυκλινικές δοκιμές.

Έστω ότι υπάρχουν n ομάδες και η i ομάδα έχει n_i παρατηρήσεις με μη παρατηρούμενες ευπάθειες Z_i $(1 \le i \le n)$.Το διάνυσμα X_{ij} $(1 \le i \le n, 1 \le j \le n_i)$ περιέχει την πληροφορία των συμμεταβλητών των χρόνων των γεγονότων T_{ij} της j th παρατήρησης στην i th ομάδα. Δεδομένου του όρου της ευπάθειας Z_i οι χρόνοι επιβίωσης στην i $(1 \le i \le n)$ ομάδα υποτίθενται ότι είναι ανεξάρτητοι και η συνάρτηση κινδύνου τους είναι της ακόλουθης μορφής

$$h\left(t\middle|x_{ij}, z_i\right) = z_i h_0(t) e^{\beta^T x_{ij}}$$

όπου $h_0(t)$ ορίζει την αναφορική συνάρτηση κινδύνου και β είναι ένα διάνυσμα παραμέτρων για εκτίμηση. Οι ευπάθειες Z_i $(i = 1, \dots, n)$ θεωρούνται ότι είναι ανεξάρτητες και ταυτοτικά κατανεμημένες τυχαίες μεταβλητές με συνάρτηση πυκνότητας αυτή που ορίζει ο ερευνητής κάθε φορά.

Η βασική υπόθεση στα από κοινού μοντέλα ευπάθειας είναι ότι τα άτομα της ίδιας ομάδας μοιράζονται την ίδια ευπάθεια. Το γεγονός αυτό είναι και η αιτία

της ύπαρξης εξάρτησης μεταξύ των χρόνων των γεγονότων μεταξύ των ατόμων σε μια ομάδα. Η ανεξαρτησία στις διάρκειες ζωής ανάμεσα στις ομάδες αντιστοιχεί σε μια εκφυλισμένη κατανομή ευπάθειας (μη ύπαρξης μεταβλητότητας στην ευπάθεια). Στις υπόλοιπες περιπτώσεις η ευπάθεια είναι θετική. Υποτίθεται ότι υπάρχει ανεξαρτησία μεταξύ των χρόνων των γεγονότων από διαφορετικές ομάδες.

Η από κοινού δεσμευμένη πολυμεταβλητή συνάρτηση επιβίωσης των ατόμων της ith ομάδας δεδομένου της Z_i ευπάθειας που μοιράζονται τα άτομα στην ith ομάδα είναι

$$S(t_{i1,}t_{i2},\dots,t_{in_i}|x_{i,}z_i) = S(t_{i1}|x_{i1},z_i)\dots S(t_{in_i}|x_{in_i},z_i)$$

$$= \exp\left(-z_i\sum_{j=1}^{n_i} H_0(t_{ij})e^{\beta^T x_{ij}}\right)$$

όπου $H_0(t) = \int_0^t h_0(s)ds$ η αθροιστική αναφορική συνάρτηση κινδύνου και $X_i = (X_{i1},\dots,X_{in_i})$ ο πίνακας συμμεταβλητών των ατόμων της ith ομάδας. Η μη δεσμευμένη από κοινού συνάρτηση επιβίωσης θα δίνεται

$$S(t_{i1,}t_{i2},\dots,t_{in_i}|x_{i,}) = ES(t_{i1,}t_{i2},\dots,t_{in_i}|x_{i,}z_i)$$

$$E\exp\left(-z_i\sum_{j=1}^{n_i} H_0(t_{ij})e^{\beta^T x_{ij}}\right) = L\left(\sum_{j=1}^{n_i} H_0(t_{ij})e^{\beta^T x_{ij}}\right)$$

όπου L ο μετασχηματισμός Laplace της ευπάθειας. Η από κοινού συνάρτηση επιβίωσης όλων των δεδομένων είναι το γινόμενο των συναρτήσεων επιβίωσης των ομάδων διότι αυτές έχουν θεωρηθεί ανεξάρτητες μεταξύ τους.

$$S(t_{11}, \dots, t_{nn_n} | x_1, \dots, x_n) = \prod_{i=1}^{n} L(\sum_{j=1}^{n_i} H_0(t_{ij}) e^{\beta^T x_{ij}})$$

Η μονομεταβλητή μη δεσμευμένη συνάρτηση επιβίωσης μπορεί να εκφραστεί μέσω του μετασχηματισμού Laplace

$$S(t_{ij}|x_{ij}) = ES(t_{ij}|x_{ij}, z_i) = E\exp\left(-z_i H_0(t_{ij}) e^{\beta^T x_{ij}}\right) = L(H_0(t_{ij}) e^{\beta^T x_{ij}})$$

Αν L^{-1} ορίζει τον αντίστροφο μετασχηματισμό Laplace τότε

$L^{-1}(S(t_{ij}|x_{ij}) = H_0(t_{ij}) e^{\beta^T x_{ij}}$ και η μη δεσμευμένη συνάρτηση επιβίωσης της ith ομάδας δίνεται

$$S(t_{i1,\dots} t_{in_i} | x_i) = L(L^{-1}(S(t_{i1}|x_{i1}) + \cdots + L(L^{-1}\left(S(t_{in_i}|x_{in_i})\right))$$

2.4.1 Shared Gamma Frailty Model

Θεωρώντας ότι η ευπάθεια του μοντέλου ακολουθεί την Γάμμα κατανομή παίρνουμε το από κοινού Γάμμα μοντέλο ευπάθειας. Η ευρεία χρήση της Γάμμα κατανομής οφείλεται στις πολύ καλές μαθηματικές τις ιδιότητες, κυρίως στην απλοϊκή μορφή του μετασχηματισμού Laplace. Υποθέτοντας ότι η ευπάθεια της ith ομάδας ακολουθεί Γάμμα κατανομή $Z\_i \sim G(1, \sigma^2)$ η πολυμεταβλητή συνάρτηση επιβίωσης για την ith ομάδα θα δίνεται από την ακόλουθη σχέση

$$S(t_{i1}, \dots, t_{in_i} | x_i) = L(\sum_{j=1}^{n_i} H_0(t_{ij}) e^{\beta^T x_{ij}}) = (1 + \sigma^2 \sum_{j=1}^{n_i} H_0(t_{ij}) e^{\beta^T x_{ij}})^{-\frac{1}{\sigma^2}}$$

$$= (\sum_{j=1}^{n_i} S(t_{ij}|x_{ij})^{-\sigma^2} - (n_i - 1))^{-\frac{1}{\sigma^2}}$$

Στην περίπτωση που υποθέσουμε μια συνάρτηση για την αναφορική συνάρτηση κινδύνου $h_0(t)$ παίρνουμε την παραμετρική μορφή του από κοινού Γάμμα μοντέλου ευπάθειας. Μπορούμε εύκολα να πάρουμε την μη δεσμευμένη συνάρτηση πιθανοφάνειας. Οι παράμετροι παλινδρόμησης β, το

διάνυσμα των παραμέτρων της αναφορικής συνάρτησης κινδύνου h_0 και η διασπορά της ευπάθειας είναι οι παράμετροι που εκτιμούνται από την συνάρτηση πιθανοφάνειας. Έστω η δεσμευμένη πιθανοφάνεια στην περίπτωση των n ομάδων μεγέθους n_i $(i = 1, \dots, n)$

$$\mathrm{Lik}(\beta, \theta, \sigma^2) = \prod_{i=1}^{n} \int_0^{\infty} \prod_{j=1}^{n_i} \left(z_i h_0(t_{ij}|\theta) e^{\beta^T X_{ij}} \right)^{\Delta_{ij}} e^{-z_i H_0(t_{ij}|\Theta) e^{\beta^T X_{ij}}} f(z_i|\sigma^2) dz_i$$

$$\text{με} \quad f(z_i|\sigma^2) = \frac{z_i^{\frac{1}{\sigma^2}-1} e^{-\frac{z_i}{\sigma^2}}}{\sigma^{\frac{2}{\sigma^2}} \Gamma(\frac{1}{\sigma^2})}$$

η συνάρτηση πυκνότητας πιθανότητας της Γάμμα κατανομής με μέση τιμή 1 και διασπορά σ^2. Χρησιμοποιώντας την παρακάτω απλοποίηση για τα y_i

$$y_i = \frac{1}{\sigma^2} + \sum_{j=1}^{n_i} H_0(t_{ij}|\theta) e^{\beta^T X_{ij}}$$

η προηγουμένη συνάρτηση πιθανοφάνειας γράφεται

$$\mathrm{Lik}(\beta, \theta, \sigma^2) = \prod_{i=1}^{n} \frac{\prod_{j=1}^{n_i} (h_0(t_{ij}|\theta) e^{\beta^T X_{ij}})^{\Delta_{ij}}}{y_\iota^{d_i + 1/\sigma^2} \sigma^{2/\sigma^2} \Gamma(1/\sigma^2)} \int_0^{\infty} (y_i z_i)^{1/\sigma^2 + d_i - 1} e^{-y_i z_i} y_i dz_i$$

όπου $d_i = \sum_{j=1}^{n_i} \Delta_{ij}$ ο αριθμός των γεγονότων σε μια ομάδα. Το ολοκλήρωμα ουσιαστικά είναι η Γάμμα κατανομή και η προηγούμενη πιθανοφάνεια γίνεται

$$\mathrm{Lik}(\beta, \theta, \sigma^2) = \prod_{i=1}^{n} \frac{\Gamma(1/\sigma^2 + d_i) \prod_{j=1}^{n_i} (h_0(t_{ij}|\theta) e^{\beta^T X_{ij}})^{\Delta_{ij}}}{(\frac{1}{\sigma^2} + \sum_{j=1}^{n_i} H_0(t_{ij}|\theta) e^{\beta^T X_{ij}})^{d_i + 1/\sigma^2} \sigma^{2/\sigma^2} \Gamma(1/\sigma^2)}$$

Λογαριθμίζοντας παίρνουμε την λογαριθμό-πιθανοφάνεια του από κοινού Γάμμα μοντέλου ευπάθειας (Klein1992, Duchateau and Janssen 2008):

$$\log\mathrm{Lik}(\beta, \theta, \sigma^2) = \sum_{i=1}^{n} [d_i \log \sigma^2 + \log\Gamma\left(1/\sigma^2 + d_i\right) - \log\Gamma\left(1/\sigma^2\right) - \left(1/\sigma^2 + d_i\right)$$

$$\log(1+\sigma^2\sum_{j=1}^{n_i} H(t_{ij}|\theta)e^{\beta^T X_{ij}}) + \sum_{j=1}^{n_\iota} \Delta_{ij}(\beta^T X_{ij} + \log h_0(t_{ij}|\theta))]$$

Οι μη παρατηρούμενες ευπάθειες Z_i $i = 1,2,\dots,n$ στο από κοινού μοντέλο Γάμμα ευπάθειας μπορούν να εκτιμηθούν από την ακόλουθη σχέση

$$\hat{Z}_i = \frac{1/\hat{\sigma}^2 + \sum_{j=1}^{n_i} \Delta_{ij}}{\frac{1}{\hat{\sigma}^2} + \sum_{j=1}^{n_i} H_0(t_{ij}|\hat{\theta})e^{\hat{\beta}^T X_{ij}}}$$

όπου $\hat{\sigma}^2$ είναι η εκτίμηση της διασποράς της ευπάθειας, $\hat{\theta}$ το διάνυσμα των εκτιμώμενων παραμέτρων της αθροιστικής συνάρτησης κινδύνου και $\hat{\beta}$ το διάνυσμα των εκτιμώμενων συντελεστών παλινδρόμησης.

2.5 Correlated frailty model

Το συσχετιζόμενο μοντέλο ευπάθειας αναπτύχθηκε για την ανάλυση δεδομένων διμεταβλητών χρόνων αποτυχίας, στο οποίο δύο τυχαίες μεταβλητές που συνδέονται χρησιμοποιούνται για τον χαρακτηρισμό της ευπάθειας στο κάθε ζευγάρι. Για παράδειγμα, μια τυχαία μεταβλητή έχει εκχωρηθεί στον σύντροφο 1 και μια για τον σύντροφο 2 έτσι ώστε να μην είναι περιορισμένοι από κάποια κοινή ευπάθεια. Οι δύο αυτές συνδέονται και έχουν από κοινού συνάρτηση κατανομής. Γνωρίζοντας την μια από τις δύο δε σημαίνει κατ' ανάγκην ότι είναι γνωστή και η δεύτερη. Δεν υπάρχει περιορισμός σχετικά με το είδος της συσχέτισης. Μπορεί αυτές οι δυο μεταβλητές να συσχετίζονται αρνητικά, γεγονός που θα προκαλέσει μια αρνητική συσχέτιση μεταξύ των χρόνων επιβίωσης. Υποθέτοντας Γάμμα κατανομή για τις ευπάθειες οι Yashin and Iachine (1995) χρησιμοποιώντας το συσχετιζόμενο μοντέλο ευπάθειας κατέληξαν σε μια διμεταβλητή κατανομή επιβίωσης της μορφής

$$S(t_1, t_2) = \frac{S_1(t_1)^{1-\rho} S_2(t_2)^{1-\rho}}{(S_1(t_1)^{-\sigma^2} S_2(t_2)^{-\sigma^2} - 1)^{\frac{\rho}{\sigma^2}}}$$

ΚΕΦΑΛΑΙΟ 3

Κριτήρια Επιλογής Μοντέλων

3.1 Εισαγωγή

Έχοντας ένα σύνολο δεδομένων μπορούμε εύκολα να προσαρμόσουμε ένα μεγάλο πλήθος μοντέλων σε αυτά. Αυτομάτως τίθενται ερωτήματα όπως ποιο μοντέλο είναι το καταλληλότερο, με ποιόν τρόπο μπορεί να επιλεχθεί και πως γίνεται ανάμεσα τους η κατάταξη για την εύρεση του βέλτιστου μοντέλου.

Η επιλογή μοντέλων προσπαθεί να συγκεράσει δυο αντικρουόμενες πλευρές. Την πρώτη, την καλύτερη προσαρμογή του μοντέλου στα δεδομένα που συνήθως επιτυγχάνεται με την εισαγωγή μεταβλητών στο μοντέλο και τη δεύτερη που είναι η μείωση της πολυπλοκότητας του μοντέλου έτσι ώστε να είναι εύκολα παρουσιάσιμο και ερμηνεύσιμο, το οποίο επιτυγχάνεται με την συγκρατημένη χρήση μεταβλητών.

Χρησιμοποιώντας μεγάλο αριθμό μεταβλητών στο μοντέλο αυξάνουμε την επεξηγηματική δυνατότητα του μοντέλου μας, την ίδια στιγμή όμως αυξάνεται η διασπορά της μεταβλητής απόκρισης και μειώνεται η μεροληψία των μεταβλητών μας (modeling bias). Αντίθετα, με την χρήση λίγων μεταβλητών μειώνουμε τη διασπορά της μεταβλητής απόκρισης στο μοντέλο παλινδρόμησης αλλά αυξάνεται η μεροληψία των μεταβλητών. Η αυξημένη μεροληψία καθιστά το μοντέλο μας πιο δυσλειτουργικό και καθιστά καθόλου εύκολη τη γενίκευση του σε όλο το πληθυσμό. Από την άλλη πλευρά η μεγάλη διασπορά δεν βοηθά στην εξαγωγή ασφαλών συμπερασμάτων. Τελικώς το

επιλεγόμενο μοντέλο προσπαθεί να ανακαλύψει τον βέλτιστο συνδυασμό μεροληψίας-διασποράς.

Η επιλογή του σωστού μοντέλου μεταφράζεται ως μια ισορροπία όπου επιλέγεται το μοντέλο με τις ακριβώς απαραίτητες μεταβλητές αποφεύγοντας έτσι την εισαγωγή πλεοναζόντων μεταβλητών που περιπλέκουν τον ερευνητή και οδηγούν σε μη ασφαλή συμπεράσματα.

Για την εύρεση του καταλληλότερου μοντέλου δημιουργήθηκαν διάφορα κριτήρια δηλαδή μέτρα που αξιολογούν την προσαρμογή των μοντέλων στα δεδομένα. Τέτοια κριτήρια είναι το Akaike Information Criterion (AIC), διάφορες βελτιώσεις του όπως το AICc και TIC, το τελικό σφάλμα πρόβλεψης (FPE), το προγνωστικό άθροισμα τετραγώνων (PRESS) , το Μπεϋζιανό κριτήριο πληροφορίας (BIC) και το κριτήριο Mallows.

3.2 Akaike Information Criterion (AIC)

Το AIC θεωρείται ως το πρώτο κριτήριο επιλογής μοντέλων. Επίσης είναι το πιο ευρέως γνωστό και χρησιμοποιούμενο εργαλείο επιλογής μοντέλων ανάμεσα στους επαγγελματίες. Το AIC το εισήγαγε για πρώτη φορά ο Hirotugu Akaike to 1973 στο άρθρο με τίτλο “Information Theory and an Extension of the maximum Likelihood Principle” (in: B. N. Petrov and F. Csaki, eds., 2nd International Symposium on Information Theory, Akademia Kiado, Budapest, pp. 267-281).

Η γενικευμένη μορφή του AIC κριτηρίου είναι

$$\mathrm{AIC} = -2\log(\text{maximized likelihood}) + 2\mathrm{p}$$

όπου p ο αριθμός μεταβλητών του μοντέλου.

Ο Akaike συνδύασε την καλή προσαρμογή του μοντέλου (goodness of fit) μέσα από τον υπολογισμό της πιθανοφάνειας του, με την όσο το δυνατόν μικρότερη πολυπλοκότητα, που εκφράζεται από τον όρο ποινικοποίησης $2p$.
Με τον τρόπο αυτό επιλέγεται το απλούστερο μοντέλο, δηλαδή αυτό με τον μικρότερο αριθμό μεταβλητών, που έχει την μέγιστη πιθανοφάνεια. Έτσι το μοντέλο που προκύπτει προσαρμόζεται ικανοποιητικά στα δεδομένα και έχει τον ελάχιστο αριθμό μεταβλητών που το κάνει πιο εύχρηστο.

Τα πλεονεκτήματα του AIC είναι

- Δεν απαιτεί την υπόθεση ότι κάποιο από τα εξεταζόμενα μοντέλα πρέπει να είναι το σωστό μοντέλο
- Μπορεί να χρησιμοποιηθεί να συγκρίνει non-nested models
- Συγκρίνει μοντέλα που είναι βασισμένα σε διαφορετικές κατανομές πιθανότητας.

Τα μειονεκτήματα του AIC είναι

- Εάν η κλάση των υποψήφιων μοντέλων είναι μεγάλη, οι AIC τιμές για αρκετά μοντέλα μπορεί να είναι κοντά στην ελάχιστη AIC τιμή, με συνέπεια το βέλτιστο μοντέλο να εντοπίζεται δύσκολα
- Η επιτυχής εφαρμογή του AIC απαιτεί μεγάλα δείγματα.

3.3 Βελτιώσεις του AIC

Η προσπάθεια για την διόρθωση κάποιων αδυναμιών του AIC όπως την επιλογή όλο και πιο πολύπλοκων μοντέλων καθώς το μέγεθος του δείγματος αυξάνεται οδήγησε στην δημιουργία του AICc. Η μέγιστη πιθανοφάνεια ενός μοντέλου αυξάνει γραμμικά με την αύξηση του μεγέθους του δείγματος ενώ ο

παράγοντας ποινικοποίησης της εξαρτάται μόνο από το μέγεθος του παραμετρικού χώρου.

Το διορθωμένο Akaike κριτήριο πληροφορίας (AICc) είναι το AIC με μία διόρθωση για πεπερασμένο δείγμα.

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2\mathrm{k}(\mathrm{k}+1)}{\mathrm{n}-\mathrm{k}-1}$$

όπου το k δηλώνει τον αριθμό των παραμέτρων του εξεταζόμενου μοντέλου. Επομένως το AICc είναι το AIC με μεγαλύτερη ποινή για τις επιπλέον μεταβλητές. Το AICc καθώς το n μεγαλώνει συγκλίνει στο αρχικό AIC. Οι Burnham & Anderson (2002) συνιστούν την χρήση του AICc έναντι του AIC εάν το μέγεθος του δείγματος n είναι μικρό ή ο αριθμός των παραμέτρων μεγάλος. Στην περίπτωση που χρησιμοποιούμε το AIC αντί την διόρθωση του όταν το μέγεθος του δείγματος n δεν είναι αρκετά μεγαλύτερο από το k^2 αυξάνεται η πιθανότητα της επιλογής μοντέλων που έχουν πολλές παραμέτρους.

Οι Brockwell & Davis προτείνουν την χρήση του AICc ως το πρωταρχικό κριτήριο επιλογής της τάξης ενός αυτοπαλινδρονούμενου κινούμενου μέσου όρου (ARMA) μοντέλου χρονοσειρών. Επίσης οι McQuarrie & Tsai το χρησιμοποιούν εκτεταμένα στην παλινδρόμηση και τις χρονοσειρές.

3.4 Κριτήριο BIC

Το μπεϋζιανό κριτήριο πληροφορίας (BIC) το εισήγαγε ο Schwartz το 1978. Η γενική μορφή του BIC για ένα μοντέλο M με μέγεθος δείγματος n και διάσταση παραμετρικού χώρου k είναι

$$\mathrm{BIC} = -2\log(\text{maximized likelihood}) + \mathrm{k} * \log\mathrm{n}$$

Παρατηρούμε τρομερές ομοιότητες με τον τύπο του AIC. Αυτό οφείλεται στο γεγονός ότι και τα δύο μοντέλα χρησιμοποιούν σαν κύριο εργαλείο τους τις μέγιστες λογαριθμικές πιθανοφάνειες του εξεταζόμενου μοντέλου και με όρους ποινικοποίησης το $2 * \mathrm{k}$ στο AIC και στο BIC το $\mathrm{k} * \ln(\mathrm{n})$, για να εντοπίσουν το μοντέλο που προσαρμόζεται καλύτερα στα δεδομένα.

Οι όροι ποινικοποίησης και των δύο κριτηρίων εξαρτώνται από τον αριθμό των παραμέτρων του μοντέλου, ενώ στο BIC εξαρτάται και από τον λογάριθμο του αριθμού των παρατηρήσεων.

Ως καταλληλότερο μοντέλο, το κριτήριο BIC θεωρεί αυτό που έχει την μεγαλύτερη τιμή. Συνήθως η τιμή του BIC είναι αρνητική για αυτό το προτεινόμενο μοντέλο είναι αυτό με την μικρότερη απόλυτη τιμή.

3.5 C_p Mallows

Ένα μέτρο για την αξιολόγηση της καταλληλότητας του μοντέλου είναι η στατιστική συνάρτηση $\mathrm{C_p} - \mathrm{Mallows}$ που δίνεται από τον ακόλουθο τύπο

$$\mathrm{C_p} = \frac{\mathrm{SSE_\kappa}}{\mathrm{S^2}} - \mathrm{n} + 2\mathrm{p}$$

όπου p ο αριθμός των επεξηγηματικών μεταβλητών στο μοντέλο, n ο αριθμός των παρατηρήσεων. Ο όρος $\mathrm{SSE_\kappa}$ συμβολίζει το άθροισμα τετραγώνων των υπολοίπων για το μοντέλο που περιλαμβάνει k μεταβλητές από αυτές που διαθέτουμε. Ο όρος $\mathrm{S^2}$ είναι το μέσο τετραγωνικό υπόλοιπο όταν χρησιμοποιούνται και οι p μεταβλητές. Ως καταλληλότερο μοντέλο διαλέγουμε το μοντέλο με $\mathrm{C_p} \cong \mathrm{p}$.Το $\mathrm{C_p}$ και το AIC είναι ασυμπωτικά ισοδύναμα. Σε μεγάλα δείγματα και τα δύο κριτήρια θα διαλέξουν το ίδιο προσαρμοσμένο μοντέλο.

ΚΕΦΑΛΑΙΟ 4

Επιλογή μοντέλων στην κλάση των μοντέλων ευπάθειας

4.1 Ορισμός συνάρτησης πιθανοφάνειας

Σε αυτή την ενότητα περιγράφεται συνοπτικά το θεωρητικό υπόβαθρο για την εκτίμηση των συντελεστών παλινδρόμησης στα μοντέλα ευπάθειας με την μέθοδο μεγίστης πιθανοφάνειας για την περίπτωση των λογοκριμένων από δεξιά παρατηρήσεων. Τα μοντέλα ευπάθειας είναι μια επέκταση του ημι-παραμετρικού μοντέλου του Cox. Όπως ήδη έχουμε αναφέρει η συνάρτηση κινδύνου του μοντέλου του Cox είναι η

$$\mathrm{h}(\mathrm{t}|\mathrm{x}) = \mathrm{h}_0(\mathrm{t})\mathrm{e}^{\beta^{\mathrm{T}}\mathrm{x}}$$

ενώ η συνάρτηση κινδύνου των μοντέλων ευπάθειας έχει παρόμοια μορφή με την προσθήκη επιπλέον μιας μεταβλητής της ευπάθειας z η οποία υπεισέρχεται πολλαπλασιαστικά στη συνάρτηση κινδύνου του Cox ως

$$\mathrm{h}(\mathrm{t}|\mathrm{x},\mathrm{z}) = \mathrm{z}\mathrm{h}_0(\mathrm{t})\mathrm{e}^{\beta^{\mathrm{T}}\mathrm{x}}$$

Η υπό συνθήκη συνάρτηση επιβίωσης του μοντέλου ευπάθειας δίνεται από τον τύπο

$$\mathrm{S}(\mathrm{t}|\mathrm{Z}=\mathrm{z},\mathrm{X}=\mathrm{x}) = \mathrm{e}^{-\mathrm{z}\mathrm{e}^{\beta^{\mathrm{T}}\mathrm{x}}\,\mathrm{H}_0(\mathrm{t})}$$

όπου Z η ευπάθεια, $\mathrm{H}_0(\mathrm{t})$ η αθροιστική αναφορική συνάρτηση κινδύνου, β το διάνυσμα των συντελεστών παλινδρόμησης και X το διάνυσμα των συμμεταβλητών.

Οπως έχουμε δει στο κεφάλαιο 2, η συνάρτηση επιβίωσης του πληθυσμού δεδομένων μόνο των συμμεταβλητών λαμβάνεται από την $S(t|z,x)$ ολοκληρώνοντας ως προς την ευπάθεια.

$$S(t|x) = \int_0^\infty e^{-ze^{\beta^T x} H_0(t)} dF_z(z) \equiv e^{-G(e^{\beta^T x} H_0(t))} \quad (1)$$

όπου F_z η συνάρτηση κατανομής της ευπάθειας. Προφανώς η συνάρτηση G ορίζεται ως $G(w) = -\ln\left(\int_0^\infty e^{-wz} dF_z(z)\right)$. Για διαφορετική κατανομή ευπάθειας F_z παράγεται διαφορετική συνάρτηση G. Ευρέως χρησιμοποιούμενη κατανομή ευπάθειας είναι η Γάμμα λόγω της μαθηματικής ευκολίας που παρουσιάζει. Σε αυτή την εργασία θα χρησιμοποιήσουμε τη Γάμμα και την Inverse Gaussian σαν κατανομές ευπάθειας. Η συνάρτηση G είναι ουσιαστικά ίση με τον –λογάριθμο του μετασχηματισμού Laplace της κατανομής της ευπάθειας και δίνεται στην περίπτωση που η ευπάθεια ακολουθεί Γαμμα(k,λ) κατανομή από την σχέση

$$G(x,k,\lambda) = k * \ln\left(1 + \frac{x}{\lambda}\right)$$

και στην περίπτωση που ακολουθεί Inverse Gaussian(μ, b) η G παίρνει την μορφή

$$G(u,\mu,b) = -\frac{b}{\mu}\left(1 - \sqrt{1 + \left(\frac{2\mu^2 u}{b}\right)}\right)$$

Για σκοπούς αναγνωρισιμότητας (identifiability) η μέση τιμή της ευπάθειας ορίζεται πάντα ίση με 1, οπότε στην πρώτη περίπτωση θεωρούμε μια κατανομή Γάμμα(α,1/α) με μέση τιμή 1 και διασπορά 1/α και στη δεύτερη μια Inverse Gaussian(1,b) κατανομή με μέση τιμή 1 και διασπορά 1/b.

Στην περίπτωση των λογοκριμένων δεδομένων οι παρατηρήσεις χωρίζονται σε αυτές που είναι μη λογοκριμένες και ανήκουν στο σύνολο U και στις

λογοκριμένες που ανήκουν στο σύνολο C. Η συνάρτηση πιθανοφάνειας τότε παίρνει την μορφή

$$L = \prod_{i\in U} f(t_i) \prod_{i\in C} S(t_i) = \prod_i \{f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}\} \quad (2)$$

όπου $$\delta_i = \begin{cases} 1, & \mu\eta\ \lambda o\gamma o\kappa\rho\iota\mu\acute{\epsilon}\nu\epsilon\varsigma \\ 0, & \lambda o\gamma o\kappa\rho\iota\mu\acute{\epsilon}\nu\epsilon\varsigma \end{cases}$$

‘Εστω τα δεδομένα μας να δίνονται απο τις τριάδες τυχαίων μεταβλητών (W_i, X_i, δ_i) όπου $W_i = \min(T_i, C_i)$ ο μικρότερος από τους χρόνους επιβίωσης και λογοκρισίας, X_i το διάνυσμα των συμμεταβλητών και δ_i η δείκτρια συνάρτηση λογοκρισίας. Αντικαθιστώντας τη συνάρτηση πυκνότητας πιθανότητας και τη συνάρτηση επιβίωσης από τον παραπάνω τύπο (1) των μοντέλων ευπάθειας στον τύπο (2) της συνάρτησης πιθανοφάνειας και λογαριθμίζοντας παίρνουμε την λογαριθμοποιημένη πιθανοφάνεια

$$\text{loglik} = \sum_{i=1}^{n} \delta_i \left(\log\left(G'\left(e^{\beta^T X_i} H_0(W_i)\right)\right) + \log\left(e^{\beta^T X_i} h_0(W_i)\right)\right) - G\left(e^{\beta^T X_i} H_0(W_i)\right)$$

Ο τελευταίος τύπος της loglik μεγιστοποιείται για την εύρεση των εκτιμητριών μεγίστης πιθανοφάνειας των συντελεστών παλινδρόμησης β. Αυτή η συνάρτηση πιθανοφάνειας όμως εξαρτάται από την άγνωστη οχληρά παράμετρο $h_0(t)$ ή κατ’ επέκταση και από τη συνάρτηση $H_0(t)$ η οποία είναι απείρου διαστάσεως παράμετρος. Στα μοντέλα ευπάθειας δεν έχουμε τη δυνατότητα να εκτιμήσουμε ξεχωριστά την ενδιαφέρουσα παράμετρο β και την οχληρά παράμετρο $H_0(t)$ όπως γίνεται στην περίπτωση του μοντέλου του Cox μέσω της μερικής συνάρτησης πιθανοφάνειας. Για να ευκολύνουμε το πρόβλημα θα θεωρήσουμε από εδώ και πέρα το παραμετρικό μοντέλο ευπάθειας. Υποθέτουμε λοιπόν ότι η αναφορική συνάρτηση κινδύνου είναι γνωστής συναρτησιακής μορφής εκτός από μια πεπερασμένης διαστάσεως οχληρά παράμετρο την οποία θα συμβολίσουμε λ, δηλαδή

$$h_0(t) = h_0(t, \lambda)$$

και

$$H_0(t) = H_0(t, \lambda)$$

Αν παραδείγματος χάριν υποθέσουμε για τους χρόνους επιβίωσης ότι ακολουθούν την Εκθετική κατανομή με παράμετρο λ τότε

$$h_0(t) = \lambda$$

και

$$H_0(t) = \lambda t$$

και συνεπώς η loglik παίρνει τη μορφή

$$\text{loglik} = \sum_{i=1}^{n} \delta_i (\log\left(G'\left(e^{\beta^T X_i} \lambda W_i\right)\right) + \log\left(e^{\beta^T X_i} \lambda\right)) - G\left(e^{\beta^T X_i} \lambda W_i\right) \quad (3)$$

Αν υποθέσουμε ότι οι χρόνοι επιβίωσης ακολουθούν την κατανομή Weibull τότε η αναφορική συνάρτηση κινδύνου είναι

$$h_0(t) = \lambda k^{-\lambda} t^{\lambda - 1}$$

και

$$H_0(t) = k^{-\lambda} t^{\lambda}$$

και συνεπώς η loglik παίρνει τη μορφή

$$\text{loglik} = \sum_{i=1}^{n} \delta_i (\log\left(G'\left(e^{\beta^T X_i} k^{-\lambda} W_i^{\lambda}\right)\right) + \log\left(e^{\beta^T X_i} \lambda k^{-\lambda} W_i^{\lambda - 1}\right)) - G\left(e^{\beta^T X_i} k^{-\lambda} W_i^{\lambda}\right) \quad (4)$$

Η μεγιστοποίηση της loglik (3) θα γίνει με βάση την καινούργια υπόθεση, ως προς τις παραμέτρους β και λ ενώ η μεγιστοποίηση της loglik (4) θα γίνει ως προς τις παραμέτρους β, k και λ.

4.2 Επιλογή μοντέλων με βάση το AIC κριτήριο

Η επιλογή των σημαντικών μεταβλητών σε ένα μοντέλο, στα δικά μας πλαίσια η επιλογή των μεταβλητών που επηρεάζουν την επιβίωση, γίνεται μέσω κριτηρίων επιλογής μοντέλων. Στην εργασία αυτή θα χρησιμοποιηθεί κυρίως το κριτήριο AIC αλλά έχει εξεταστεί επίσης και το κριτήριο AICc (AIC corrected) όπως επίσης μπορεί να γίνει γενίκευση σε οποιοδήποτε άλλο κριτήριο. Θεωρώντας όλους τους δυνατούς συνδυασμούς των συμμεταβλητών που έχουμε στη διαθεσή μας σε μια μελέτη, υπολογίζουμε για όλους την μέγιστη τιμή της loglik από τον τύπο (3) π.χ., με την παράμετρο ενδιαφέροντος και την οχληρά παράμετρο να εχουν αντικατασταθεί από τους αντίστοιχους εκτιμητές μεγίστης πιθανοφάνειας. Στη συνέχεια υπολογίζουμε την τιμή του AIC για κάθε ένα από τα δυνατά μοντέλα από τον τύπο

$$\mathrm{AIC} = -2\log(\text{maximized likelihood}) + 2\mathrm{p}$$

που έχουμε δει στο κεφάλαιο 3. Το μοντέλο που επιλέγεται είναι αυτό που ελαχιστοποιεί την τιμή του κριτηρίου.

Η εφαρμογή του κριτηρίου σε μοντέλα ευπάθειας αλλά και η αποτελεσματικότητά του εξετάζεται με βάση προσομοιωμένα δεδομένα που περιγράφονται αναλυτικά στην επόμενη ενότητα του κεφαλαίου αυτού, όπου δίνονται επίσης τα αποτελέσματα της εργασίας και τα συμπεράσματά μας.

Πρέπει να τονίσουμε εδώ ότι κύριος σκοπός της εργασίας αυτής είναι να καλύψουμε κενά που παρουσιάζουν στατιστικά πακέτα όπως π.χ. η R στην οποία οι κατανομές ευπάθειας που μπορεί να υποθέσει κανείς σε συνδυασμό με κριτήρια επιλογής μοντέλων είναι περιορισμένες σε αριθμό. Συγκεκριμένα στην R οι κατανομές ευπάθειας που μπορεί κάποιος να υποθέσει είναι οι t, Γάμμα και Gaussian. Με τη μεθοδολογία που έχουμε εμείς εφαρμόσει

μπορούμε να υποθέσουμε οποιαδήποτε συνεχή κατανομή ευπάθειας και στις προσομοιώσεις μας στη συνέχεια θα θεωρήσουμε τη Γάμμα και την Inverse Gaussian κατανομή.

4.3 Προσομοιώσεις

Για την εφαρμογή των παραπάνω δημιουργούμε τα δεδομένα μας με βάση το ακόλουθο μοντέλο προσομοίωσης. Θεωρούμε το διάνυσμα των συμμεταβλητών $X = (X_1, X_2, X_3, X_4)$ οι οποίες ακολουθούν πολυμεταβλητή κανονική κατανομή με μέση τιμή μηδέν και πίνακα διασπορών συνδιασπορών τον ακόλουθο

Πίνακας Διασπορών-Συνδιασπορών			
1	0.5	0.25	0.125
0.5	1	0.5	0.25
0.25	0.5	1	0.5
0.125	0.25	0.5	1

Οι χρόνοι αποτυχίας $t_i, i = 1, \dots, n$ δεδομένων των συμμεταβλητών και της ευπάθειας, θεωρούμε ότι ακολουθούν ένα μοντέλο ευπάθειας με συνάρτηση κινδύνου την

$$h(t|z, x) = z \exp(\beta^T x)\, \lambda$$

Η Z δηλώνει την ευπάθεια και στην πρώτη περίπτωση των προσομοιωμένων δεδομένων υποθέσαμε ότι ακολουθεί μια Γάμμα κατανομή με μέση τιμή 1 και διασπορά 0.25. Στην δεύτερη περίπτωση υποθέσαμε ότι η ευπάθεια ακολουθεί Inverse Gaussian επίσης με μέση τιμή 1 και διασπορά 0.25. Οι χρόνοι αποτυχίας t_i κατασκευάζονται συνεπώς σαν τυχαίες παρατηρήσεις από εκθετική κατανομή με μέση τιμή $z \exp(\beta_0^T x)\, \lambda_0$. Το β_0 είναι η πραγματική

τιμή των συντελεστών και ισούται με το διάνυσμα (0.8,0,0,1). Το λ_0 είναι η πραγματική τιμή της παραμέτρου της εκθετικής αναφορικής συνάρτησης κινδύνου και ισούται με 1. Για την δημιουργία των λογοκριμένων παρατηρήσεων θα χρησιμοποιήσουμε την μεταβλητή $\mathrm{U}\sim\mathrm{Uniform}(1,\mathrm{c})$ όπου το c επιλέγεται ανάλογα έτσι ώστε να παίρνουμε στο κάθε σύνολο δεδομένων το ποσοστό λογοκριμένων παρατηρήσεων που επιθυμούμε. Οι χρόνοι λογοκρισίας κατασκευάζονται σαν τυχαίες παρατηρήσεις από την εκθετική κατανομή με μέση τιμή την $\mathrm{U}\exp(\beta_o{}^{\tau}\mathrm{x})\,\lambda_0$. Συνεπώς τα δεδομένα μας αποτελούνται από τα $\mathrm{W_i}$ με $\mathrm{W_i} = \min(\mathrm{T_i},\mathrm{C_i}), \mathrm{i}=1,\dots,\mathrm{n}$ μαζί με τα $\mathrm{X_i}$, και δ_i, $\mathrm{i}=1,\dots\mathrm{n}$.

Λόγω της τιμής του β_0 μόνο οι συμμεταβλητές X_1 και X_4 είναι σημαντικές. Επομένως για το κάθε σύνολο δεδομένων και αφού εξετάσουμε όλους τους συνδυασμούς των μοντέλων που παράγονται από τις 4 συμμεταβλητές, το AIC κριτήριο θα πρέπει να παίρνει την μικρότερη τιμή του στο μοντέλο που έχει μόνο τις συμμεταβλητές X_1 και X_4. Δηλαδή θα πρέπει να επιλέγει ως βέλτιστο μοντέλο τις περισσότερες φορές το μοντέλο που περιλαμβάνει μόνο τις συμμεταβλητές X_1 και X_4 που έχουν μη μηδενικούς συντελεστές παλινδρόμησης στο διάνυσμα β_0.

4.4. Αποτελέσματα

Τα αποτελέσματά μας που δίνονται κατωτέρω στους πίνακες 1-4 βασίζονται σε 100 προσωμοιωμένα data sets κάθε φορά με την ευπάθεια να ακολουθεί Γάμμα κατανομή (Πϊνακες 1-2) ή Inverse Gaussian κατανομή (Πϊνακες 3-4). Εξετάσαμε μικρά και μεγάλα μεγέθη δείγματος, δηλαδή πήραμε σύνολα δεδομένων (data sets) να αποτελούνται από 50 παρατηρήσεις ή από 100

παρατηρήσεις. Υποθέσαμε επίσης διάφορα ποσοστά λογοκρισίας και πιο συγκεκριμένα ποσοστά λογοκρισίας περίπου 15%, 35% και 55%. Σκοπός μας είναι να δούμε πως το μέγεθος του δείγματος ή το ποσοστό λογοκρισίας επηρεάζει τα αποτελέσματα.

Στον πίνακα αποτελεσμάτων 1 η πρώτη στήλη δείχνει το ποσοστό των λογοκριμένων παρατηρήσεων στο κάθε σύνολο δεδομένων. Η δεύτερη στήλη δείχνει πόσες φορές επιλέχτηκε το σωστό μοντέλο δηλαδή το μοντέλο με συμμεταβλητές τις X_1και X_4. Στην τρίτη στήλη είναι το Error1 δηλαδή το ποσοστό που επιλέγεται λανθασμένα η X_2 ή η X_3 ή και οι δύο μαζί. Στην τέταρτη στήλη είναι το Error2 που δηλώνει το ποσοστό να μην επιλεγεί λανθασμένα η X_1 ή η X_4 ή και οι δύο μαζί. Στις στήλες 5 και 6 παίρνουμε τις μέσες τιμές των εκτιμητριών των β_1 και β_4. Στις στήλες 7 και 8 παίρνουμε τις τυπικές αποκλίσεις των εκτιμητριών των συντελεστών β_1 και β_4 αντίστοιχα. Η ένατη στήλη δίνει την αρνητική μέση τιμή του λογαρίθμου της εκτίμησης του λ_0. Η διαφορά στον πίνακα 1 μας δίνει το ποσοστό που τα κριτήρια AIC και AICc (AIC corrected) δίνουν διαφορετικές επιλογές στην επιλογή του βέλτιστου μοντέλου.

Οι στήλες CoxMβ1 και CoxMβ4 δίνουν τις μέσες τιμές των εκτιμώμενων β_1 και β_4 αντίστοιχα χρησιμοποιώντας την έτοιμη συνάρτηση της R coxph με την ευπάθεια να ακολουθεί Γάμμα κατανομή. Οι επόμενες δυο στήλες δίνουν την τυπική απόκλιση των β_1 και β_4 αντίστοιχα μέσω της coxph. Το percent1 είναι το ποσοστό που έχει επιλεχθεί το σωστό μοντέλο δηλαδή αυτό με συμμεταβλητές X_1 και X_4 χρησιμοποιώντας την coxph για την εξαγωγή των εκτιμώμενων συντελεστών και την extractAIC συνάρτηση για τον υπολογισμό της τιμής του κριτηρίου AIC. Τα CoxError1 και CoxError2 είναι ανάλογα του

Error1 και Error2 με τη δική μας μέθοδο. Θα πρέπει να σημειώσουμε εδώ βέβαια ότι η coxph υποθέτει και δουλεύει με το ημιπαραμετρικό μοντέλο Cox όπου δηλαδή η οχληρά παράμετρος είναι συνάρτηση γι' αυτό και δεν υπάρχει εκτίμηση του λ για σκοπούς σύγκρισης.

Το MSurλ είναι η μέση τιμή της εκτίμησης της παραμέτρου (-logλ) μέσω της έτοιμης ρουτίνας της R, survreg. Στις υπόλοιπες στήλες του Πίνακα 1 ακολουθούν οι μέσες τιμές και οι τυπικές αποκλίσεις των εκτιμητών των συντελεστών β_1 και β_4 αντίστοιχα, το ποσοστό των φορών που επιλέχθηκε το σωστό μοντέλο με την survreg και τα SurError1 και SurError2 που είναι αντίστοιχα των Error1 και Error2 της δικής μας μεθόδου. Θα πρέπει να τονίσουμε εδώ ότι η survreg υποθέτει και δουλεύει με το παραμετρικό μοντέλο Cox όπως και εμείς αλλά δεν υποθέτει ευπάθεια.

Στη συνέχεια θα εξηγήσουμε τη διαφορά προσήμου ανάμεσα στους δικούς μας συντελεστές παλινδρόμησης (και της διαδικασίας coxph) και της διαδικασίας survreg η οποία υποθέτει ότι ο λογάριθμος του χρόνου επιβίωσης συνδέεται γραμμικά με τις συμμεταβλητές. Θεωρώντας τον ορισμό μας για τη συνάρτηση επιβίωσης με εκθετική συνάρτηση αθροιστικού αναφορικού κινδύνου και κάνοντας πράξεις έχουμε ότι

$$\mathrm{S(t|x)} = \mathrm{e}^{-\mathrm{G}\left(\lambda \mathrm{t e}^{\beta^{\mathrm{T}}\mathrm{x}}\right)} <=> ln\ (\mathrm{S(t|x)}) = -\mathrm{G}\left(\mathrm{e}^{\beta^{\mathrm{T}}\mathrm{x}}\lambda \mathrm{t}\right) <=>$$

$$\mathrm{G}^{-1}\ (-\ln\ (\mathrm{S(t|x)}) = \mathrm{e}^{\beta^{\mathrm{T}}\mathrm{x}}\lambda \mathrm{t} <=>$$

$$\ln\ (\mathrm{G}^{-1}(-\ln\left(\mathrm{S(t|x)}\right) = \beta^{\mathrm{T}}\mathrm{x} + \ln(\lambda) + \ln(\mathrm{t}) <=>$$

$$\ln(\mathrm{t}) = \ln\ (\mathrm{G}^{-1}(-\ln\left(\mathrm{S(t|x)}\right) - \beta^{\mathrm{T}}\mathrm{x} - \ln(\lambda)$$

Δηλαδή, σε μορφή γραμμικού μοντέλου η πιο πάνω σχέση γράφεται σαν

$$\ln(\mathrm{T}) = -\beta \mathrm{X} - \ln(\lambda) + \varepsilon$$

όπου το σφάλμα e^{ε} ακολουθεί Pareto κατανομή. Από την παραπάνω ανάλυση εξηγούνται οι αρνητικές τιμές των συντελεστών β στη survreg που οφείλονται στο (-) που βρίσκεται μπροστά στο συντελεστή β^{T}. Εξηγείται επίσης γιατί χρειάζεται να συγκρίνουμε τις τιμές του $-\ln(\lambda)$ αντί του λ.

Στον Πίνακα 1 για όλα τα ποσοστά λογοκρισίας το καλύτερο ποσοστό επιλογής του σωστού μοντέλου το παίρνουμε από την μέθοδο μας. Οι εκτιμήσεις μας για τα β_1 και β_4 είναι καλές και πολύ κοντά με αυτές που παίρνουμε από την έτοιμη συνάρτηση της R την survreg σε απόλυτη τιμή. Η εκτίμηση της παραμέτρου της εκθετικής αναφορικής συνάρτησης κινδύνου υπερεκτιμάται και στις τρεις μεθόδους που δίνουν τιμές από 2,4 μέχρι 2,6 που είναι αρκετά μεγαλύτερες από την πραγματική τιμή 1. Στις εκτιμήσεις των παραμέτρων με την μέθοδο μας και την coxph παίρνουμε τιμές κοντά στις πραγματικές σε απόλυτη τιμή.

Καθώς το ποσοστό των λογοκριμένων παρατηρήσεων αυξάνει όλο και περισσότερο αποκλίνουν οι εκτιμήσεις μας από τις πραγματικές τιμές. Επίσης η τυπική απόκλιση των εκτιμητών των β αυξάνεται. Σημαντική αύξηση έχει το Error2 που από 2% αυξάνεται στο 20% για ποσοστό λογοκρισίας 55%.

Καθώς αυξάνεται ο αριθμός παρατηρήσεων του κάθε συνόλου δεδομένων από 50 σε 100 παρατηρήσεις, οι εκτιμήσεις των β όπως φαίνεται στον Πίνακα 2 είναι αισθητά βελτιωμένες από αυτές του Πίνακα 1. Οι τυπικές αποκλίσεις των εκτιμώμενων β είναι μικρότερες. Οι εκτιμήσεις των β με την μέθοδο μας είναι σχεδόν ίδιες με τις εκτιμήσεις που παίρνουμε από την έτοιμη συνάρτηση της R την survreg, γεγονός που αποδεικνύει ότι η μέθοδος μας δίνει σωστά αποτελέσματα. Όταν ο αρθμός των παρατηρήσεων στα δείγματα αυξάνεται τα κριτήρια AIC και AICc τείνουν να επιλέγουν το ίδιο μοντέλο.

ΠΙΝΑΚΑΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ 1 – Γάμμα κατανομή

Sample Size = 50								
Censor	Percent	Error1	Error2	Mβ1	Mβ4	Sd(β1)	Sd(β4)	M(-logλ)
15%	0,72	0,28	0,02	-0,7889	-1,0390	0,1818	0,2027	2,4872
35%	0,72	0,26	0,07	-0,8325	-1,0166	0,1854	0,2472	2,5198
55%	0,57	0,38	0,20	-0,8527	-1,0418	0,2471	0,3181	2,5599
Censor	Διαφορά	CoxMβ1	CoxMβ4	Sd(Cβ1)	Sd(Cβ4)	Percent1	CoxError1	CoxError2
15%	0,08	-0,7513	-1,0210	0,2184	0,3173	0,56	0,40	0,12
35%	0,04	-0,8138	-0,9895	0,2376	0,2733	0,59	0,40	0,16
55%	0,07	-0,8648	-1,0715	0,3007	0,3803	0,56	0,40	0,26
Censor	Msurλ	MSurβ1	MSurβ4	Sd(Sβ1)	Sd(Sβ4)	Percent2	SurError1	SurError2
15%	2,6896	0,7904	1,0280	0,1889	0,1990	0,64	0,36	0,02
35%	2,6685	0,8361	1,0276	0,1855	0,2329	0,69	0,31	0,04
55%	2,6636	0,8499	1,0837	0,2402	0,3000	0,54	0,43	0,16

ΠΙΝΑΚΑΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ 2 – Γάμμα κατανομή

Sample Size = 100								
Censor	Percent	Error1	Error2	Mβ1	Mβ4	Sd(β1)	Sd(β4)	M(-logλ)
15%	0,71	0,29	0	-0,7919	-1,0205	0,1423	0,1311	2,5160
35%	0,65	0,35	0	-0,8075	-1,0090	0,1881	0,1697	2,5144
55%	0,64	0,36	0,01	-0,8386	-1,0199	0,2161	0,2473	2,5328
Censor	Διαφορά	CoxMβ1	CoxMβ4	Sd(Cβ1)	Sd(Cβ4)	Percent1	CoxError1	CoxError2
15%	0,06	-0,7470	-0,9602	0,1558	0,1719	0,63	0,37	0,03
35%	0,03	-0,7572	-0,9483	0,1910	0,1832	0,58	0,42	0,03
55%	0,02	-0,8144	-1,0000	0,2139	0,2513	0,60	0,40	0,03
Censor	Msurλ	MSurβ1	MSurβ4	Sd(Sβ1)	Sd(Sβ4)	Percent2	SurError1	SurError2
15%	2,7220	0,7968	1,0233	0,1418	0,1263	0,60	0,40	0
35%	2,6813	0,8089	1,0042	0,1818	0,1684	0,54	0,46	0
55%	2,6300	0,8347	1,0139	0,2098	0,2130	0,60	0,40	0,01

Οι Πίνακες 3 και 4 δίνουν τα αποτελέσματα της μεθόδου μας για την περίπτωση που η ευπάθεια ακολουθεί την Inverse Gaussian κατανομή, για πλήθος παρατηρήσεων 50 και 100 στο κάθε σύνολο δεδομένων αντίστοιχα. Το ποσοστό επιλογής του σωστού μοντέλου είναι, για τις τρείς περιπτώσεις ποσοστού λογοκρισίας, γύρω από το ικανοποιητικό ποσοστό του 67%. Η εκτίμηση του β_1 και στις τρείς περιπτώσεις είναι πολύ καλή (υπερεκτιμά λίγο την πραγματική τιμή) όπως εξαιρετική είναι και η εκτίμηση του β_4 που ισούται σχεδόν με το 1 δηλαδή την πραγματική τιμή του β_4. Όμοια αποτελέσματα παίρνουμε με την έτοιμη συνάρτηση survreg γεγονός που επιβεβαιώνει τα αποτελέσματα της μεθόδου μας. Στον Πίνακα 4 και για $\mathrm{n} = 100$ οι εκτιμήσεις μας βελτιώνονται αισθητά και είναι ίδιες με τις πραγματικές που έχουμε θέσει από την αρχή της διαδικασίας. Την καλύτερη εκτίμηση των β την έχουμε για το μικρότερο ποσοστό λογοκριμένων παρατηρήσεων, ακόμα οι τυπικές αποκλίσεις των εκτιμώμενων β μειώνονται αρκετά με την αύξηση του αριθμού των παρατηρήσεων από 50 σε 100. Η αύξηση του πλήθους των παρατηρήσεων μειώνει και εκμηδενίζει την διαφορά μεταξύ των κριτηρίων AIC και AICc. Στον Πίνακα 4 βλέπουμε ότι το Error2 είναι ίδιο με το SurError2 που ισούται με μηδέν και στις τρείς περιπτώσεις των διαφορετικών ποσοστών λογοκριμένων παρατηρήσεων γεγονός που δηλώνει ότι οι μεταβλητές X_1και X_4 έχουν επιλεγεί όλες τις φορές.

Θα πρέπει να σημειώσουμε εδώ ότι σύγκριση με την διαδικασία coxph δεν μπορεί να γίνει στην περίπτωση της Inverse Gaussian frailty γιατί η κατανομή αυτή δεν δίνεται σαν επιλογή στη διαδικασία coxph. Πιο συγκεκριμένα, η coxph δέχεται σαν κατανομές ευπάθειας την Γάμμα, την Κανονική (Normal) και την t.

ΠΙΝΑΚΑΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ 3 - Inverse Gaussian

Sample Size = 50

Censor	Percent	Error1	Error2	Mβ1	Mβ4	Sd(β1)	Sd(β4)	M(-logλ)
15%	0,69	0,31	0,02	-0,8393	-1,0000	0,1996	0,2081	-0,2759
35%	0,67	0,32	0,04	-0,8226	-1,0430	0,2422	0,2612	-0,2575
55%	0,65	0,29	0,14	-0,8134	-0,9994	0,2148	0,2492	-0,1649

Censor	Διαφορά	MSurλ	MSurβ1	MSurβ4	Sd(Sβ1)	Sd(Sβ4)	Percent2	SurError1
15%	0,11	-0,1034	0,8484	1,0010	0,2025	0,2098	0,65	0,35
35%	0,09	-0,1243	0,8240	1,0409	0,2373	0,2615	0,63	0,36
55%	0,08	-0,0710	0,8139	0,9978	0,2086	0,2414	0,66	0,31

Censor	SurError2
15%	0,01
35%	0,04
55%	0,11

ΠΙΝΑΚΑΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ 4 - Inverse Gaussian

Sample Size = 100

Censor	Percent	Error1	Error2	Mβ1	Mβ4	Sd(β1)	Sd(β4)	M(-logλ)
15%	0,72	0,28	0	-0,7965	-1,0010	0,1448	0,1253	-0,2358
35%	0,70	0,30	0	-0,8175	-0,9876	0,1588	0,1505	-0,2159
55%	0,60	0,40	0	-0,8032	-0,9772	0,1848	0,2372	-0,1871

Censor	Διαφορά	MSurλ	MSurβ1	MSurβ4	Sd(Sβ1)	Sd(Sβ4)	Percent2	SurError1
15%	0,03	-0,0524	0,8077	0,9986	0,1433	0,1256	0,68	0,32
35%	0,06	-0,0799	0,8145	0,9873	0,1613	0,1504	0,66	0,34
55%	0,10	-0,0823	0,8150	0,9829	0,1771	0,2277	0,50	0,50

Censor	SurError2
15%	0
35%	0
55%	0

Συνεπώς η διαδικασία που ακολουθούμε λειτουργεί ικανοποιητικά και όσο αυξάνεται το πλήθος των παρατηρήσεων δίνει εκτιμητές με μεγαλύτερη ακρίβεια. Το ποσοστό επιλογής σωστού μοντέλου με το AIC κυμαίνεται στα ικανοποιητικά επίπεδα του 70%. Η μέθοδός μας διευρύνει το πλήθος των συνεχών κατανομών που μπορούμε να χρησιμοποιήσουμε ως κατανομές ευπάθειας σε σύγκριση με τις περιορισμένες δυνατοτήτες που έχει μέχρι τώρα η γλώσσα προγραμματισμού R.

Παράρτημα Α

```
#Gamma distribution
#Likelihood with exponential data
library(MASS)
library(nlme)
library(mvtnorm)
library(splines)
library(splines)
library(survival)

data<-function(){
beta_0<-c(0.8,0,0,1)      #Αρχικό διάνυσμα
lamda_0<-1                 # Αρχική τιμή για το λ
n<-100  #sample size of data set

sigma_0<-mat.or.vec(4,4)          #Μηδενικός πίνακας  4x4
rho<-0.5
for (i in 1:4){
for (j in 1:4){
sigma_0[i,j]<-rho^(abs(i-j))
}
}
sigma_0                      #Πινακάς συνδιασπορών

#Δημιουργεί έναν πίνακα  nx4 από κανονική κατανομή με πίνακα συνδιασπορών το sigma_0
#Η κάθε στήλη του πίνακα Z είναι μια συμμεταβλητή
Z <- rmvnorm(n,mean = rep(0, nrow(sigma_0)),sigma_0)
Z

alpha<-4
eta_i<-rgamma(n,alpha,1/alpha)    #Ευπάθεια ακολουθεί την γάμμα κατανομή
eta_i

#Δημιουργεί την  T_i από την εκθετική κατανομή  με μέση τιμή
#eta_i*exp(beta_0*Z_i)*lamda_0
mean<-rep(0,n)
for (i in 1:n){
mean[i]<-eta_i[i]*exp(beta_0%*%Z[i,])*lamda_0
}
mean
T_i<-rexp(n,(1/mean))
T_i

#Δημιουργεί τους χρόνους αποκοπής από την εκθετική κατανομή με μέση τιμή την
#u_i e^{beta_0 ‘ Z_i) lambda_0

U_i<-runif(n,1,23)
m<-rep(0,n)
for(j in 1:n){
m[j]<-U_i[j]*exp(beta_0%*%Z[j,])*lamda_0
}
m
C_i<-rexp(n,(1/m))
C_i            #χρόνοι αποκοπής

X<-pmin(T_i,C_i)          #min(T_i,C_i)
X
#delta=1 για μη αποκομμένες και 0 για αποκομμένες
```

```
delta<-rep(0,n)
for(i in 1:n){
if(T_i[i]-C_i[i]<0)
delta[i]<-1
else if (T_i[i]-C_i[i]>0)
delta[i]<-0
}
delta
percent<-sum(delta)/n
percent    #ποσοστό μη αποκομμένων παρατηρήσεων στο καθέ σύνολο δεδομένων
#=========================================================================
#G function
G<-function(x,alpha){
out.G<-alpha*log(1+(1/alpha)*x)
return(out.G)
}
#Η παράγωγος της G
dg<-function(x,alpha){
dg<- alpha/(alpha+x)
return(dg)
}
#=========================================================================
#Ορισμός της συνάρτησης πιθανοφάνειας
p<-rep(0,n)   #διάνυσμα 1Xn
q<-rep(0,n)     # διάνυσμα 1xn
w<-rep(0,n)    # διάνυσμα 1Xn
like.out<-rep(0,n)     #αποθηκέυονται οι τιμές της συνάρτησης πιθανοφάνειας
#Εχουμέ την συνάρτηση like  με μεταβλητές το αρχικό διάνυσμα των συντελεστών β και τον
#πίνακα τιμών των συμμεταβλητών

like<-function(theta,Z){
k<-ncol(Z)
n<-nrow(Z)
beta<-theta[1:k]
lamda<-theta[k+1]
for(i in 1:n){
p[i]<-exp(beta%*%Z[i,])*lamda*(X[i])
q[i]<-exp(beta%*%Z[i,])*lamda
like.out[i]<-delta[i]*( log(dg(p[i],alpha))+ log(q[i])) -G(p[i],alpha)
w[i]<-like.out[i]
}
return(-sum(w))
}
#Συνάρτηση πιθανοφάνειας για μονομεταβλητά μοντέλα
likem<-function(theta,Z){
beta<-theta[1]
lamda<-theta[2]
for(i in 1:n){
p[i]<-exp(beta%*%Z[i])*lamda*(X[i])
q[i]<-exp(beta%*%Z[i])*lamda
like.out[i]<-delta[i]*( log(dg(p[i],alpha))+log(q[i])) -G(p[i],alpha)
w[i]<-like.out[i]
}
return(-sum(w))
}
```

```
#=====================================================================
#Τα 15 μοντέλα που δημιουργούν οι 4 συμμεταβλητές χωρισμένα σε υποκατηγορίες.

#Μοντέλα με μια συμμεταβλητή
mmodela<-array(dim=c(n,1,4))
mmodela[,,1]<-Z[,1]
mmodela[,,2]<-Z[,2]
mmodela[,,3]<-Z[,3]
mmodela[,,4]<-Z[,4]

#Μοντελά με δύο συμμεταβλητές
dmodela<-array(dim=c(n,2,6))
dmodela[,,1]<-Z[,1:2]
dmodela[,,2]<-Z[,2:3]
dmodela[,,3]<-Z[,3:4]
dmodela[,,4]<-cbind(Z[,1],Z[,3])
dmodela[,,5]<-cbind(Z[,1],Z[,4])
dmodela[,,6]<-cbind(Z[,2],Z[,4])

#Μοντέλα με τρείς συμμεταβλητλές
tmodela<-array(dim=c(n,3,4))
tmodela[,,1]<-Z[,1:3]
tmodela[,,2]<-Z[,2:4]
tmodela[,,3]<-cbind(Z[,1],Z[,2],Z[,4])
tmodela[,,4]<-cbind(Z[,1],Z[,3],Z[,4])

fmodela<-Z     #Το μοντέλο με όλες τις συμμεταβλητές
#=====================================================================
store<-matrix(0,15,17)    #πινακάς αποθήκευσης αποτελεσμάτων
#Μονομεταβλητά μοντέλα Z_1,Z_2,Z_3,Z_4
Akaikem<-rep(0,6)  #Διάνυσμα αποθήκευσης των τιμών του AIC
maximm<-rep(0,4)
Akaim<-rep(0,6)   #Διάνυσμα αποθήκευσης των τιμών του AICc
estimm<-matrix(rep(0,8),ncol=2)       #Διάνυσμα αποθήκευσης των εκτιμώμενων συντελεστών
beta1m<-c(0.8,lamda_0)
beta2m<-c(0,lamda_0)
beta3m<-c(0,lamda_0)
beta4m<-c(1,lamda_0)
thetam<-rbind(beta1m,beta2m,beta3m,beta4m)

for(i in 1:4){
estimm[i,]<-constrOptim(thetam[i,],likem,NULL,Z=mmodela[,,i],ui=rbind(c(0,1)),ci=c(0))$par
maximm[i]<-likem(estimm[i,],mmodela[,,i])
Akaikem[i]<- 2*maximm[i]+2*2
Akaim[i]<-Akaikem[i]+(2*2*(2+1))/(n-2-1)
store[i,1]<-Akaikem[i]
store[i,17]<-Akaim[i]
store[i,2:3]<-estimm[i,]
store[i,7]<-estimm[i,2]
}
#=====================================================================
#Μοντέλα με δύο μεταβλητές Z_1-Z_2,Z_2-Z_3,Z_3_Z_4,Z_1-Z_3,Z_1-Z_4,Z_2-Z_4
Akaike<-rep(0,6)
Akaid<-rep(0,6)
maxi<-rep(0,6)
estim1<-matrix(rep(0,18),ncol=3)

beta1<-c(0.8,0,lamda_0)
beta2<-c(0,0,lamda_0)
beta3<-c(0,1,lamda_0)
```

```
beta4<-c(0.8,0,lamda_0)
beta5<-c(0.8,1,lamda_0)
beta6<-c(0,1,lamda_0)
theta<-rbind(beta1,beta2,beta3,beta4,beta5,beta6)

for( i in 1:6){
estim1[i,]<-constrOptim(theta[i,],like,NULL,Z=dmodela[,,i],ui=rbind(c(0,0,1)),ci=(0))$par
maxi[i]<-like(estim1[i,],dmodela[,,i])
Akaike[i]<- 2*maxi[i]+2*3
Akaid[i]<-Akaike[i]+(2*3*(3+1))/(n-3-1)
store[4+i,17]<-Akaid[i]
store[4+i,1]<-Akaike[i]
store[4+i,2:4]<-estim1[i,]
store[4+i,7]<-estim1[i,3]
}
#=========================================================================
#Μοντέλα με τρείς συμμεταβλητές
#Z_1-Z_2-Z_3, Z_2-Z_3-Z_4, Z_1-Z_2-Z_4, Z_1-Z_3-Z_4
Akaike3<-rep(0,6)
Akait<-rep(0,6)
maxi3<-rep(0,4)
estim3<-matrix(rep(0,16),ncol=4)

#Καθορίζονται τα διανύσματα βήτα για το κάθε μοντέλο με 3 μεταβλητές
betaa<-c(0.8,0,0,lamda_0)
betab<-c(0,0,1,lamda_0)
betac<-c(0.8,0,1,lamda_0)
betad<-c(0.8,0,1,lamda_0)

theta3<-rbind(betaa,betab,betac,betad)
for( i in 1:4){
estim3[i,]<-constrOptim(theta3[i,],like,NULL,Z=tmodela[,,i],ui=rbind(c(0,0,0,1)),ci=c(0))$par
maxi3[i]<-like(estim3[i,],tmodela[,,i])
Akaike3[i]<- 2*maxi3[i]+2*4
Akait[i]<-Akaike3[i]+(2*4*(4+1))/(n-4-1)
store[10+i,17]<-Akait[i]
store[10+i,1]<-Akaike3[i]
store[10+i,2:5]<-estim3[i,]
store[10+i,7]<-estim3[i,4]
}
#=========================================================================
#Το μοντέλο με 4 συμμεταβλητές
Akaike4<-rep(0,6)
Akaif<-rep(0,6)
maxi4<-rep(0)
estim4<-rep(0,5)
#estim14<-rep(0,5)
theta4<-c(0.8,0,0,1,lamda_0)
estim4<-constrOptim(theta4,like,NULL,Z=fmodela,ui=rbind(c(0,0,0,0,1)),ci=c(0))$par
maxi4<-like(estim4,fmodela)
Akaike4<- 2*maxi4+2*5
Akaif<-Akaike4+(2*5*(5+1))/(n-5-1)
store[15,17]<-Akaif
store[15,1]<-Akaike4
store[15,2:6]<-estim4
store[15,7]<-estim4[5]
```

```
#====================================================================
#Υπολογισμός των συντελεστών παλινδρόμησης με τις έτοιμες συνάρτησεις της R την
#survreg και την coxph και χρήση της exactAIC για τον  υπολογισμό της τιμής του AIC .

group1<-Z[,1]
group2<-Z[,2]
group3<-Z[,3]
group4<-Z[,4]
status<-delta
time<-X
id<-seq(1:n)
group<-list(time,status,group1,group2,group3,group4,id)
mode<-matrix(0,15,4)
mod<-matrix(0,15,5)

mod[15,]<survreg(Surv(time,status)~group1+group2+group3+group4,data=group,dist='expon
ential')$coef
mod15<-
survreg(Surv(time,status)~group1+group2+group3+group4,data=group,dist='exponential')
value15<-extractAIC(mod15)
mod[14,1:4]<-
survreg(Surv(time,status)~group1+group3+group4,data=group,dist='exponential')$coef
mod14<-survreg(Surv(time,status)~group1+group3+group4,data=group,dist='exponential')
value14<-extractAIC(mod14)
mod[13,1:4]<-
survreg(Surv(time,status)~group1+group2+group4,data=group,dist='exponential')$coef
mod13<-survreg(Surv(time,status)~group1+group2+group4,data=group,dist='exponential')
value13<-extractAIC(mod13)
mod[12,1:4]<-
survreg(Surv(time,status)~group2+group3+group4,data=group,dist='exponential')$coef
mod12<-survreg(Surv(time,status)~group2+group3+group4,data=group,dist='exponential')
value12<-extractAIC(mod12)
mod[11,1:4]<-
survreg(Surv(time,status)~group1+group2+group3,data=group,dist='exponential')$coef
mod11<-survreg(Surv(time,status)~group1+group2+group3,data=group,dist='exponential')
value11<-extractAIC(mod11)
mod[10,1:3]<-survreg(Surv(time,status)~group2+group4,data=group,dist='exponential')$coef
mod10<-survreg(Surv(time,status)~group2+group4,data=group,dist='exponential')
value10<-extractAIC(mod10)
mod[9,1:3]<-survreg(Surv(time,status)~group1+group4,data=group,dist='exponential')$coef
mod9<-survreg(Surv(time,status)~group1+group4,data=group,dist='exponential')
value9<-extractAIC(mod9)
mod[8,1:3]<-survreg(Surv(time,status)~group1+group3,data=group,dist='exponential')$coef
mod8<-survreg(Surv(time,status)~group1+group3,data=group,dist='exponential')
value8<-extractAIC(mod8)
mod[7,1:3]<-survreg(Surv(time,status)~group3+group4,data=group,dist='exponential')$coef
mod7<-survreg(Surv(time,status)~group3+group4,data=group,dist='exponential')
value7<-extractAIC(mod7)
mod[6,1:3]<-survreg(Surv(time,status)~group2+group3,data=group,dist='exponential')$coef
mod6<-survreg(Surv(time,status)~group2+group3,data=group,dist='exponential')
value6<-extractAIC(mod6)
mod[5,1:3]<-survreg(Surv(time,status)~group1+group2,data=group,dist='exponential')$coef
mod5<-survreg(Surv(time,status)~group1+group2,data=group,dist='exponential')
value5<-extractAIC(mod5)
mod[4,1:2]<-survreg(Surv(time,status)~group4,data=group,dist='exponential')$coef
mod4<-survreg(Surv(time,status)~group4,data=group,dist='exponential')
value4<-extractAIC(mod4)
mod[3,1:2]<-survreg(Surv(time,status)~group3,data=group,dist='exponential')$coef
mod3<-survreg(Surv(time,status)~group3,data=group,dist='exponential')
value3<-extractAIC(mod3)
```

```
mod[2,1:2]<-survreg(Surv(time,status)~group2,data=group,dist='exponential')$coef
mod2<-survreg(Surv(time,status)~group2,data=group,dist='exponential')
value2<-extractAIC(mod2)
mod[1,1:2]<-survreg(Surv(time,status)~group1,data=group,dist='exponential')$coef
mod1<-survreg(Surv(time,status)~group1,data=group,dist='exponential')
value1<-extractAIC(mod1)

mode[15,]<-
coxph(Surv(time,status)~group1+group2+group3+group4+frailty(id,dist="gamma",sparce=TR
UE,method='em'),data=group)$coef
mode15<-
coxph(Surv(time,status)~group1+group2+group3+group4+frailty(id,dist="gamma",sparce=TR
UE,method='em'),data=group)
valu15<-extractAIC(mode15)
mode[14,1:3]<-
coxph(Surv(time,status)~group1+group3+group4+frailty(id,dist="gamma",sparce=TRUE,meth
od='em'),data=group)$coef
mode14<-
coxph(Surv(time,status)~group1+group3+group4+frailty(id,dist="gamma",sparce=TRUE,meth
od='em'),data=group)
valu14<-extractAIC(mode14)
mode[13,1:3]<-
coxph(Surv(time,status)~group1+group2+group4+frailty(id,dist="gamma",sparce=TRUE,meth
od='em'),data=group)$coef
mode13<-
coxph(Surv(time,status)~group1+group2+group4+frailty(id,dist="gamma",sparce=TRUE,meth
od='em'),data=group)
valu13<-extractAIC(mode13)
mode[12,1:3]<-
coxph(Surv(time,status)~group2+group3+group4+frailty(id,dist="gamma",sparce=TRUE,meth
od='em'),data=group)$coef
mode12<-
coxph(Surv(time,status)~group2+group3+group4+frailty(id,dist="gamma",sparce=TRUE,meth
od='em'),data=group)
valu12<-extractAIC(mode12)
mode[11,1:3]<-
coxph(Surv(time,status)~group1+group2+group3+frailty(id,dist="gamma",sparce=TRUE,meth
od='em'),data=group)$coef
mode11<-
coxph(Surv(time,status)~group1+group2+group3+frailty(id,dist="gamma",sparce=TRUE,meth
od='em'),data=group)
valu11<-extractAIC(mode11)
mode[10,1:2]<-
coxph(Surv(time,status)~group2+group4+frailty(id,dist="gamma",sparce=TRUE,method='em')
,data=group)$coef
mode10<-
coxph(Surv(time,status)~group2+group4+frailty(id,dist="gamma",sparce=TRUE,method='em')
,data=group)
valu10<-extractAIC(mode10)
mode[9,1:2]<-
coxph(Surv(time,status)~group1+group4+frailty(id,dist="gamma",sparce=TRUE,method='em')
,data=group)$coef
mode9<-
coxph(Surv(time,status)~group1+group4+frailty(id,dist="gamma",sparce=TRUE,method='em')
,data=group)
valu9<-extractAIC(mode9)
mode[8,1:2]<-
coxph(Surv(time,status)~group1+group3+frailty(id,dist="gamma",sparce=TRUE,method='em')
,data=group)$coef
```

```
mode8<-
coxph(Surv(time,status)~group1+group3+frailty(id,dist="gamma",sparce=TRUE,method='em')
,data=group)
valu8<-extractAIC(mode8)
mode[7,1:2]<-
coxph(Surv(time,status)~group3+group4+frailty(id,dist="gamma",sparce=TRUE,method='em')
,data=group)$coef
mode7<-
coxph(Surv(time,status)~group3+group4+frailty(id,dist="gamma",sparce=TRUE,method='em')
,data=group)
valu7<-extractAIC(mode7)
mode[6,1:2]<-
coxph(Surv(time,status)~group2+group3+frailty(id,dist="gamma",sparce=TRUE,method='em')
,data=group)$coef
mode6<-
coxph(Surv(time,status)~group2+group3+frailty(id,dist="gamma",sparce=TRUE,method='em')
,data=group)
valu6<-extractAIC(mode6)
mode[5,1:2]<-
coxph(Surv(time,status)~group1+group2+frailty(id,dist="gamma",sparce=TRUE,method='em')
,data=group)$coef
mode5<-
coxph(Surv(time,status)~group1+group2+frailty(id,dist="gamma",sparce=TRUE,method='em')
,data=group)
valu5<-extractAIC(mode5)
mode[4,1]<-
coxph(Surv(time,status)~group4+frailty(id,dist="gamma",sparce=TRUE,method='em'),data=gr
oup)$coef
mode4<-
coxph(Surv(time,status)~group4+frailty(id,dist="gamma",sparce=TRUE,method='em'),data=gr
oup)
valu4<-extractAIC(mode4)
mode[3,1]<-
coxph(Surv(time,status)~group3+frailty(id,dist="gamma",sparce=TRUE,method='em'),data=gr
oup)$coef
mode3<-
coxph(Surv(time,status)~group3+frailty(id,dist="gamma",sparce=TRUE,method='em'),data=gr
oup)
valu3<-extractAIC(mode3)
mode[2,1]<-
coxph(Surv(time,status)~group2+frailty(id,dist="gamma",sparce=TRUE,method='em'),data=gr
oup)$coef
mode2<-
coxph(Surv(time,status)~group2+frailty(id,dist="gamma",sparce=TRUE,method='em'),data=gr
oup)
valu2<-extractAIC(mode2)
mode[1,1]<-
coxph(Surv(time,status)~group1+frailty(id,dist="gamma",sparce=TRUE,method='em'),data=gr
oup)$coef
mode1<-
coxph(Surv(time,status)~group1+frailty(id,dist="gamma",sparce=TRUE,method='em'),data=gr
oup)
valu1<-extractAIC(mode1)


store[1:4,8]=mode[1:4,1]
store[5:10,8:9]=mode[5:10,1:2]
store[11:14,8:10]=mode[11:14,1:3]
store[15,8:11]=mode[15,1:4]
store[,12]=percent
```

```
p<-21
store1<-matrix(0,15,p)
store1[1,1:p]<-
c(store[1,1],store[1,2],0,0,0,store[1,3],log(store[1,3]),percent,0,mode[1,1],0,0,0,store[1,17],valu1[2],value1[2],mod[1,2],0,0,0,mod[1,1])
store1[2,1:p]<-c(store[2,1],0,store[2,2],0,0,store[2,3],-
log(store[2,3]),percent,0,0,mode[2,1],0,0,store[2,17],valu2[2],value2[2],0,mod[2,2],0,0,mod[2,1])
store1[3,1:p]<-c(store[3,1],0,0,store[3,3],0,store[3,3],-
log(store[3,3]),percent,0,0,0,mode[3,1],0,store[3,17],valu3[2],value3[2],0,0,mod[3,2],0,mod[3,1])
store1[4,1:p]<-c(store[4,1],0,0,0,store[4,2],store[4,3],-
log(store[4,3]),percent,0,0,0,0,mode[4,1],store[4,17],valu4[2],value4[2],0,0,0,mod[4,2],mod[4,1])
store1[5,1:p]<-c(store[5,1],store[5,2],store[5,3],0,0,store[5,4],-
log(store[5,4]),percent,0,mode[5,1],mode[5,2],0,0,store[5,17],valu5[2],value5[2],mod[5,2],mod[5,3],0,0,mod[5,1])
store1[6,1:p]<-c(store[6,1],0,store[6,2],store[5,3],0,store[6,4],-
log(store[6,4]),percent,0,0,mode[6,1],mode[6,2],0,store[6,17],valu6[2],value6[2],0,mod[6,2],mod[6,3],0,mod[6,1])
store1[7,1:p]<-c(store[7,1],0,0,store[7,2],store[7,3],store[7,4],-
log(store[7,4]),percent,0,0,0,mode[7,1],mode[7,2],store[7,17],valu7[2],value7[2],0,0,mod[7,2],
mod[7,3],mod[7,1])
store1[8,1:p]<-c(store[8,1],store[8,2],0,store[8,3],0,store[8,4],-
log(store[8,4]),percent,0,mode[8,1],0,mode[8,2],0,store[8,17],valu8[2],value8[2],mod[8,2],0,mod[8,3],0,mod[8,1])
store1[9,1:p]<-c(store[9,1],store[9,2],0,0,store[9,3],store[9,4],-
log(store[9,4]),percent,1,mode[9,1],0,0,mode[9,2],store[9,17],valu9[2],value9[2],mod[9,2],0,0,
mod[9,3],mod[9,1])
store1[10,1:p]<-c(store[10,1],0,store[10,2],0,store[10,3],store[10,4],-
log(store[10,4]),percent,0,0,mode[10,1],0,mode[10,2],store[10,17],valu10[2],value10[2],0,mod[10,2],0,mod[10,3],mod[10,1])
store1[11,1:p]<-c(store[11,1],store[11,2],store[11,3],store[11,4],0,store[11,5],-
log(store[11,5]),percent,0,mode[11,1],mode[11,2],mode[11,3],0,store[11,17],
valu11[2],value11[2],mod[11,2],mod[11,3],mod[11,4],0,mod[11,1])
store1[12,1:p]<-c(store[12,1],0,store[11,2],store[11,3],store[11,4],store[12,5],-
log(store[12,5]),percent,0,0,mode[12,1],mode[12,2],mode[12,1],store[12,17],
valu12[2],value12[2],0,mod[12,2],mod[12,3],mod[12,4],mod[12,1])
store1[13,1:p]<-c(store[13,1],store[13,2],store[13,3],0,store[13,4],store[13,5],-
log(store[13,5]),percent,0,mode[13,1],mode[13,2],mode[13,3],
0,store[13,17],valu13[2],value13[2],mod[13,2],mod[13,3],0,mod[13,4],mod[13,1])
store1[14,1:p]<-c(store[14,1],store[14,2],0,store[14,3],store[14,4],store[14,5],-
log(store[14,5]),percent,0,mode[14,1],0,mode[14,2],mode[14,3],store[14,17],
valu14[2],value14[2],mod[14,2],0,mod[14,3],mod[14,4],mod[14,1])
store1[15,1:p]<-c(store[15,1],store[15,2],store[15,3],store[15,4],store[15,5],store[15,6],-
log(store[15,6]),percent,0,
mode[15,1],mode[15,2],mode[15,3],mode[15,4],store[15,17],valu15[2],value15[2],mod[15,2],mod[15,3],mod[15,4],mod[15,5],mod[15,1])
return(store1)
}
#============================================================
#Δημιουργώ τον πίνακα  results me diastaseis 15*21 και με 3 διάστασή τον αριθμό
#που θέλουμε να τρέξουμε το προγραμμά.Σε κάθε υποπίνακα του res  αποθηκέυεται ο
#εκάστωτε πίνακας store1

k<-100   # πόσες φόρες θα τρέξει το μοντέλο  k>=2
res<-array(dim=c(15,21,k))
for(i  in 1:k){
res[,,i]<-data()
}
```

```
res

minaic<-rep(0,k)
minAICc<-rep(0,k)
minCoxaic<-rep(0,k)
minSurvaic<-rep(0,k)
inds<-matrix(0,k,2)
inds2<-matrix(0,k,2)
inds3<-matrix(0,k,2)
inds4<-matrix(0,k,2)
par<-matrix(0,k,21)
parCox<-matrix(0,k,21)
parSur<-matrix(0,k,21)
#Πινακάς των μοντέλων
W<-c("Z_1","Z_2","Z_3","Z_4","Z_1-Z_2","Z_2-Z_3","Z_3-Z_4","Z_1-Z_3","Z_1-Z_4","Z_2-
Z_4","Z_1-Z_2-Z_3","Z_2-Z_3-Z_4","Z_1-Z_2-Z_4","Z_1-Z_3-Z_4","Z_1-Z_2-Z_3-Z_4")

x<-rep(0,k)
y<-rep(0,k)
z<-rep(0,k)
w<-rep(0,k)
h<-rep(0,k)
for(i in 1:k){
minaic[i]<-min(res[,1,i])
minAICc[i]<-min(res[,14,i])
minCoxaic[i]<-min(res[,15,i])
minSurvaic[i]<-min(res[,16,i])
inds[i,]<-which( res[,1,i]== min(res[,1,i]),arr.ind=TRUE)
inds2[i,]<-which( res[,14,i]== min(res[,14,i]),arr.ind=TRUE)
inds3[i,]<-which( res[,15,i]== min(res[,15,i]),arr.ind=TRUE)
inds4[i,]<-which( res[,16,i]== min(res[,16,i]),arr.ind=TRUE)
x[i]<-W[inds[i,1]]
y[i]<-W[inds2[i,1]]
z[i]<-inds[i,1]-inds2[i,1]
w[i]<-W[inds3[i,1]]
h[i]<-W[inds4[i,1]]
par[i,]<-res[inds[i,1],,i]
parCox[i,]<-res[inds3[i,1],,i]
parSur[i,]<-res[inds4[i,1],,i]
}
x
y
z          #Μη μηδενικά στοιχεία του  z δείχνουν ότι το  AICc και το  AIC δεν
           #έχουν επιλέξει το ίδιο μοντέλο
w           #ποιά μοντέλα επιλέγει  η coxph
parCox   #ο πίνακάς με εκτιμήσεις των συντελεστων από την  coxph
h           #ποιά μοντέλα διάλεξε η surv
parSur   # ο πίνακάς με εκτιμήσεις των συντελεστων από την  surv
v<-rep(0,k)
for(i in 1:k){
if(z[i]!=0)
v[i]<-1
else
v[i]<-0
}
diafora<-sum(v)/k         #το ποσοστό που AICc και AIC δε δίνουν το ίδιο μοντέλο
dimnames(par) = list(NULL,c("AICvalue","beta1","beta2","beta3","beta4",
"Lamda","-
log(lamda)","UnCensor","CorrectModel","Cbeta1","Cbeta2","Cbeta3","Cbeta4","AICc",
"AICcoxvalue","AICsurv","Surbeta1","Surbeta2","Surbeta3","Surbeta4","Surlamda"))
```

```
par
minaic
minAICc

#Επιλέγεται το μοντέλο  Z1-Z4,Το κάθε στοιχείο της στήλης 9 είναι 1 εφόσον έχει επιλεχθεί το
#μοντέλο Z1-Z4 διαφορετικά μηδέν
percent<-sum(par[,9])/k
percent
rest<-1-percent
rest
percent1<-sum(parCox[,9])/k
percent2<-sum(parSur[,9])/k
percent1
percent2

number1<-length(par[par[,2]==0,2])
number1
number2<-length(par[par[,2]!=0,2])
number2
mbeta_1<- sum(par[,2])/(k-number1) #μέση τιμής του β1
mbeta_1
sd_beta_1<-sd(par[par[,2]!=0,2]) #τυπική απόκλιση του β1

number3<-length(par[par[,5]==0,5])
number3
number4<-length(par[par[,5]!=0,5])
number4
mbeta_4<-sum(par[,5])/(k-number3)  #μέση τιμής του β4
mbeta_4
sd_beta_4<-sd(par[par[,5]!=0,5]) #τυπική απόκλιση του β4

mlamda<-mean(par[,6])
mlamda
m_log_lamda<-mean(par[,7])
#=====================================================================
#Βρίσκει  πόσες φορές παρατηρούντε οι συμμεταβλητές Z2,Z3

value<-rep(0,k)
for( i in 1:k){
if (par[i,3]!=0 | par[i,4]!=0 )
value[i]<-1
else
value[i]<-0
}
value
error1<-sum(value)/k  #Να επιλεγεί η  Z2 ή  Z3 ή και οι δύο μαζί
error1
#=====================================================================
#Υπολογισμός error2
zeros<-matrix(0,k,4)
ermat<-cbind(par[1:k,2:5],zeros)
for( i in 1:k){
if(  ermat[i,1]==0)    #Να είναι μηδέν το  β1 αποθηκέυει την τιμή 1 στην στήλη  6
ermat[i,6]<-1
else
ermat[i,6]<-0
}
ermat

for( i in 1:k){
```

```
if( ermat[i,4]==0) #Να είναι μηδέν το  β4 αποθηκέυει την τιμή 1 στην στήλη  5
ermat[i,5]<-1
else
ermat[i,5]<-0
}
ermat
#Να είναι μηδέν το β1  και το β4 αποθηκέυει την τιμή 1 στην στήλη 7
for( i in 1:k){
if(ermat[i,1]==0 && ermat[i,4]==0)
ermat[i,7]<-1
else
ermat[i,7]<-0
}
ermat
for(i in 1:k){
ermat[i,8]<-sum(ermat[i,5:7])  #Αθροισμά των στηλών 5,6,7 στην 8 στήλη ανα γραμμή
}
error2<-sum(ermat[,8])/k
error2
#=========================================================================
#Υπολογισμός Coxerror1 kai Coxerror2
cvalue<-rep(0,k)
for( i in 1:k){
if (parCox[i,3]!=0 | parCox[i,4]!=0 )
cvalue[i]<-1
else
cvalue[i]<-0
}
cvalue
Coxerror1<-sum(cvalue)/k  #Να επιλεγεί το Z2 ή το Z3 ή και τα δύο μαζί
Coxerror1

#Υπολογισμός Coxerror2
zeros<-matrix(0,k,4)
Coxermat<-cbind(parCox[1:k,2:5],zeros)
for ( i in 1:k){
if (Coxermat[i,1]==0)   #Να είναι μηδέν το β1 απόθηκέυει την τιμή 1 στην στήλη 6
Coxermat[i,6]<-1
else
Coxermat[i,6]<-0
}
Coxermat

for( i in 1:k){
if (Coxermat[i,4]==0) #Να είναι μηδέν το β4 απόθηκέυει την τιμή 1 στην στήλη 5
Coxermat[i,5]<-1
else
Coxermat[i,5]<-0
}
Coxermat

for (i in 1:k){
if(Coxermat[i,1]==0 && Coxermat[i,4]==0)
#Να είναι μηδέν το β1 και το β4  αποθηκέυει την τιμή 1 στην στήλη 7
Coxermat[i,7]<-1
else
Coxermat[i,7]<-0
}
Coxermat
for (i in 1:k){
```

```
Coxermat[i,8]<-sum(Coxermat[i,5:7]) #Αθροίζει τις στήλες  5,6,7 στην  8 ανά γραμμή
}
Coxerror2<-sum(Coxermat[,8])/k
Coxerror2
#=========================================================================
#Υπολογισμός Surerror1 και Surerror2
svalue<-rep(0,k)
for( i in 1:k){
if (parSur[i,3]!=0 | parSur[i,4]!=0 )
svalue[i]<-1
else
svalue[i]<-0
}
svalue
Surerror1<-sum(svalue)/k #Να επιλέγει το Z2 ή το Z3 ή και τα δύο μαζί
Surerror1

#Υπολογισμός Surerror2
zeros<-matrix(0,k,4)
Surermat<-cbind(parSur[1:k,2:5],zeros)
for (i in 1:k){
if (Surermat[i,1]==0)   #Να είναι μηδέν το  β1 αποθηκέυει την τιμή 1 στην στήλη 6
Surermat[i,6]<-1
else
Surermat[i,6]<-0
}
Surermat

for( i in 1:k){
if (Surermat[i,4]==0) #Να είναι μηδέν το  β4 αποθηκέυει την τιμή 1 στην στήλη  5
Surermat[i,5]<-1
else
Surermat[i,5]<-0
}
Surermat

for( i in 1:k){
if(Surermat[i,1]==0 && Surermat[i,4]==0
#Να είναι μηδέν το  β1και β4  αποθηκέυει την τιμή 1 στην στήλη  7
Surermat[i,7]<-1
else
Surermat[i,7]<-0
}
Surermat
for (i in 1:k){
Surermat[i,8]<-sum(Surermat[i,5:7]) #Αθροίζει τις στήλες  5,6,7 στην 8 ανά  γραμμή
}
Surerror2<-sum(Surermat[,8])/k
Surerror2
#=========================================================================
number5<-length(par[par[,10]==0,10])
number5
number6<-length(par[par[,10]!=0,10])
number6
cmbeta_1<- sum(par[,10])/(k-number5)
cmbeta_1
sd_cbeta_1<-sd(par[par[,10]!=0,10])
number7<-length(par[par[,13]==0,13])
number7
number8<-length(par[par[,13]!=0,13])
```

```
number8
cmbeta_4<-sum(par[,13])/(k-number7)
cmbeta_4
sd_cbeta_4<-sd(par[par[,13]!=0,13])
#=====================================================================
number9<-length(par[par[,17]==0,17])
number9
number10<-length(par[par[,17]!=0,17])
number10
smbeta_1<- sum(par[,17])/(k-number9)
smbeta_1
sd_sbeta_1<-sd(par[par[,17]!=0,17])
sd_sbeta_1

number11<-length(par[par[,20]==0,20])
number11
number12<-length(par[par[,20]!=0,20])
number12
smbeta_4<- sum(par[,20])/(k-number11)
smbeta_4
sd_sbeta_4<-sd(par[par[,20]!=0,20])
sd_sbeta_4
#=====================================================================
#Πίνακας Αποτελεσμάτων
resu2<-matrix(0,1,26)
resu2[1,1]<-mean(par[,8])
resu2[1,2]<-percent
resu2[1,3]<-rest
resu2[1,4]<-error1
resu2[1,5]<-error2
resu2[1,6]<-mbeta_1
resu2[1,7]<-mbeta_4
resu2[1,8]<-sd_beta_1
resu2[1,9]<-sd_beta_4
resu2[1,10]<-m_log_lamda
resu2[1,11]<-diafora
resu2[1,12]<-cmbeta_1
resu2[1,13]<-cmbeta_4
resu2[1,14]<-sd_cbeta_1
resu2[1,15]<-sd_cbeta_4
resu2[1,16]<-percent1
resu2[1,17]<-Coxerror1
resu2[1,18]<-Coxerror2
resu2[1,19]<-mean(parSur[,21])
resu2[1,20]<-smbeta_1
resu2[1,21]<-smbeta_4
resu2[1,22]<-percent2
resu2[1,23]<-sd_sbeta_1
resu2[1,24]<-sd_sbeta_4
resu2[1,25]<-Surerror1
resu2[1,26]<-Surerror2
dimnames(resu2) = list(NULL,c("UnCensor","Percent","1-Percent","Error1","Error2","Mbeta_1"
,"Mbeta_4","Sd(beta_1)","Sd(beta_4)","MlogL","diafora","CMbeta_1"
,"CMbeta_4","Sd(cbeta_1)","Sd(cbeta_4)","Percent1","Coxerror1","Coxerror2"
,"SurLamda","SurBeta_1","Surbeta_4","Percent2","Sd_sbeta_1","Sd_xbeta_4","Surerror1","S
urerror2"))

resu2
```

Παράρτημα Β

```
#Inverse Gaussian
#Likelihood with exponential data
library(MASS)
library(nlme)
library(mvtnorm)
library(splines)
library(splines)
library(survival)
library(statmod)

data<-function(){
beta_0<-c(0.8,0,0,1)      #Αρχικό διάνυσμα
lamda_0<-1                          # Αρχική τιμή για το λ
n<-100  #μέγεθος του data set

sigma_0<-mat.or.vec(4,4)
rho<-0.5
for (i in 1:4){
for (j in 1:4){
sigma_0[i,j]<-rho^(abs(i-j))
}
}
sigma_0                     #πινακάς διασπορών -συνδιασπορών

#Πινακάς τιμών των συμμεταβλητών
Z <- rmvnorm(n,mean = rep(0, nrow(sigma_0)),sigma_0)
Z

eta_i<-eta_i<-rinvgauss(n,1,3.4)    #Ευπάθεια ακολουθεί inverse gaussian
eta_i

#Δημιουργεί χρόνους  T_i  απο εκθετική κατανομή με μέση τιμή
#eta_i*exp(beta_0*Z_i)*lamda_0
mean<-rep(0,n)
for (i in 1:n){
mean[i]<-eta_i[i]*exp(beta_0%*%Z[i,])*lamda_0
}
mean
T_i<-rexp(n,(1/mean))
T_i

#Δημιουργεί τους χρόνους αποκοπής από εκθετική κατανομή  με μέση τιμη την
#u_i e^{beta_0 ‘ Z_i) lambda_0

U_i<-runif(n,1,14.5)
m<-rep(0,n)
for(j in 1:n){
m[j]<-U_i[j]*exp(beta_0%*%Z[j,])*lamda_0
}
m
C_i<-rexp(n,(1/m))
C_i   #χρόνοι αποκοπης

X<-pmin(T_i,C_i)
X
#delta=1  για μη αποκομένες και 0 για αποκομένες
delta<-rep(0,n)
```

```
for(i in 1:n){
if(T_i[i]-C_i[i]<0)
delta[i]<-1
else if (T_i[i]-C_i[i]>0)
delta[i]<-0
}
delta
percent<-sum(delta)/n
percent
#=========================================================================
#Η συνάρτηση G και η παραγωγός της για την Inverse Gaussian
G<-function(u,b,mi)
{
G.out<- -(b/mi)*(1-sqrt(1+((2*mi^2*u)/b)))
G.out
}

dg<-function(u,b,mi)
{
dg.out<- mi/(sqrt(1+(2*mi^2*u)/b))
dg.out
}
#=========================================================================
#Ορισμός των συναρτήσεων πιθανοφανειών
mi<-1
b<-4
p<-rep(0,n)   #dianisma 1Xn
q<-rep(0,n)     #dianisma 1xn
w<-rep(0,n)    #dianisma 1Xn
like.out<-rep(0,n

like<-function(theta,Z){
k<-ncol(Z)
n<-nrow(Z)
beta<-theta[1:k]
lamda<-theta[k+1]
for(i in 1:n){
p[i]<-exp(beta%*%Z[i,])*lamda*(X[i])
q[i]<-exp(beta%*%Z[i,])*lamda
like.out[i]<-delta[i]*( log(dg(p[i],b,mi))+ log(q[i])) -G(p[i],b,mi)
w[i]<-like.out[i]
}
return(-sum(w))
}

#Συνάρτηση πιθανοφάνειας για μονομεταβλητά μοντέλα
likem<-function(theta,Z){
beta<-theta[1]
lamda<-theta[2]
for(i in 1:n){
p[i]<-exp(beta%*%Z[i])*lamda*(X[i])
q[i]<-exp(beta%*%Z[i])*lamda
like.out[i]<-delta[i]*( log(dg(p[i],b,mi))+log(q[i])) -G(p[i],b,mi)
w[i]<-like.out[i]
}
return(-sum(w))
}

#=========================================================================
```

```
#Τα 15 μοντέλα

#Μοντέλα με μια συμμεταβλητή
mmodela<-array(dim=c(n,1,4))
mmodela[,,1]<-Z[,1]
mmodela[,,2]<-Z[,2]
mmodela[,,3]<-Z[,3]
mmodela[,,4]<-Z[,4]

#Μοντέλα με δύο συμμεταβλητές
dmodela<-array(dim=c(n,2,6))
dmodela[,,1]<-Z[,1:2]
dmodela[,,2]<-Z[,2:3]
dmodela[,,3]<-Z[,3:4]
dmodela[,,4]<-cbind(Z[,1],Z[,3])
dmodela[,,5]<-cbind(Z[,1],Z[,4])
dmodela[,,6]<-cbind(Z[,2],Z[,4])

#Μοντέλα με τρείς συμμεταβλητές
tmodela<-array(dim=c(n,3,4))
tmodela[,,1]<-Z[,1:3]
tmodela[,,2]<-Z[,2:4]
tmodela[,,3]<-cbind(Z[,1],Z[,2],Z[,4])
tmodela[,,4]<-cbind(Z[,1],Z[,3],Z[,4])

fmodela<-Z     #Το μοντέλο με 4 συμμεταβλητές
#=========================================================================
store<-matrix(0,15,17)    #Πινακάς αποθήκευσης αποτελεσμάτων
#Για τα μονομεταβλήτα μοντέλα  Z_1,Z_2,Z_3,Z_4
Akaikem<-rep(0,6)
maximm<-rep(0,4)
Akaim<-rep(0,6)
estimm<-matrix(rep(0,8),ncol=2) #Αποθηκέυονται οι εκτιμητές των παραμέτρων
beta1m<-c(0.8,lamda_0)
beta2m<-c(0,lamda_0)
beta3m<-c(0,lamda_0)
beta4m<-c(1,lamda_0)
thetam<-rbind(beta1m,beta2m,beta3m,beta4m)

for(i in 1:4){
estimm[i,]<-constrOptim(thetam[i,],likem,NULL,Z=mmodela[,,i],ui=rbind(c(0,1)),ci=c(0))$par
maximm[i]<-likem(estimm[i,],mmodela[,,i])
Akaikem[i]<- 2*maximm[i]+2*2
Akaim[i]<-Akaikem[i]+(2*2*(2+1))/(n-2-1)
store[i,1]<-Akaikem[i]
store[i,17]<-Akaim[i]
store[i,2:3]<-estimm[i,]
store[i,7]<-estimm[i,2]
}

#=========================================================================
#Τα μοντέλα  Z_1-Z_2,Z_2-Z_3,Z_3_Z_4,Z_1-Z_3,Z_1-Z_4,Z_2-Z_4
Akaike<-rep(0,6)
Akaid<-rep(0,6)
maxi<-rep(0,6)
estim1<-matrix(rep(0,18),ncol=3)
#Το διανυσμα β για το κάθε μοντέλο
beta1<-c(0.8,0,lamda_0)
beta2<-c(0,0,lamda_0)
beta3<-c(0,1,lamda_0)
```

```
beta4<-c(0.8,0,lamda_0)
beta5<-c(0.8,1,lamda_0)
beta6<-c(0,1,lamda_0)
theta<-rbind(beta1,beta2,beta3,beta4,beta5,beta6)

for( i in 1:6){
estim1[i,]<-constrOptim(theta[i,],like,NULL,Z=dmodela[,,i],ui=rbind(c(0,0,1)),ci=(0))$par
maxi[i]<-like(estim1[i,],dmodela[,,i])
Akaike[i]<- 2*maxi[i]+2*3
Akaid[i]<-Akaike[i]+(2*3*(3+1))/(n-3-1)
store[4+i,17]<-Akaid[i]
store[4+i,1]<-Akaike[i]
store[4+i,2:4]<-estim1[i,]
store[4+i,7]<-estim1[i,3]
}
#=========================================================================
#Τα μοντέλα με τρείς συμμεταβλητές
#Z_1-Z_2-Z_3, Z_2-Z_3-Z_4, Z_1-Z_2-Z_4, Z_1-Z_3-Z_4
Akaike3<-rep(0,6)
Akait<-rep(0,6)
maxi3<-rep(0,4)
estim3<-matrix(rep(0,16),ncol=4)

#Ορίζεται το δίανυσμα β για κάθε από τα παραπάνω μοντέλα τριών συμμεταβλητών
 betaa<-c(0.8,0,0,lamda_0)
betab<-c(0,0,1,lamda_0)
betac<-c(0.8,0,1,lamda_0)
betad<-c(0.8,0,1,lamda_0)

theta3<-rbind(betaa,betab,betac,betad)
for( i in 1:4){
estim3[i,]<-constrOptim(theta3[i,],like,NULL,Z=tmodela[,,i],ui=rbind(c(0,0,0,1)),ci=c(0))$par
maxi3[i]<-like(estim3[i,],tmodela[,,i])
Akaike3[i]<- 2*maxi3[i]+2*4
Akait[i]<-Akaike3[i]+(2*4*(4+1))/(n-4-1)
store[10+i,17]<-Akait[i]
store[10+i,1]<-Akaike3[i]
store[10+i,2:5]<-estim3[i,]
store[10+i,7]<-estim3[i,4]
}
#=========================================================================
#Το μοντέλο με όλες τις συμμεταβλητές
Akaike4<-rep(0,6)
Akaif<-rep(0,6)
maxi4<-rep(0)
estim4<-rep(0,5)
theta4<-c(0.8,0,0,1,lamda_0)
estim4<-constrOptim(theta4,like,NULL,Z=fmodela,ui=rbind(c(0,0,0,0,1)),ci=c(0))$par
maxi4<-like(estim4,fmodela)
Akaike4<- 2*maxi4+2*5
Akaif<-Akaike4+(2*5*(5+1))/(n-5-1)
store[15,17]<-Akaif
store[15,1]<-Akaike4
store[15,2:6]<-estim4
store[15,7]<-estim4[5]
#=========================================================================
group1<-Z[,1]
group2<-Z[,2]
group3<-Z[,3]
group4<-Z[,4]
```

```
status<-delta
time<-X
id<-seq(1:n)
group<-list(time,status,group1,group2,group3,group4,id)

mod<-matrix(0,15,5)

mod[15,1:5]<survreg(Surv(time,status)~group1+group2+group3+group4,data=group,dist='exp
onential')$coef
mod15<-
survreg(Surv(time,status)~group1+group2+group3+group4,data=group,dist='exponential')
value15<-extractAIC(mod15)
mod[14,1:4]<-
survreg(Surv(time,status)~group1+group3+group4,data=group,dist='exponential')$coef
mod14<-survreg(Surv(time,status)~group1+group3+group4,data=group,dist='exponential')
value14<-extractAIC(mod14)
mod[13,1:4]<-
survreg(Surv(time,status)~group1+group2+group4,data=group,dist='exponential')$coef
mod13<-survreg(Surv(time,status)~group1+group2+group4,data=group,dist='exponential')
value13<-extractAIC(mod13)
mod[12,1:4]<-
survreg(Surv(time,status)~group2+group3+group4,data=group,dist='exponential')$coef
mod12<-survreg(Surv(time,status)~group2+group3+group4,data=group,dist='exponential')
value12<-extractAIC(mod12)
mod[11,1:4]<-
survreg(Surv(time,status)~group1+group2+group3,data=group,dist='exponential')$coef
mod11<-survreg(Surv(time,status)~group1+group2+group3,data=group,dist='exponential')
value11<-extractAIC(mod11)
mod[10,1:3]<-survreg(Surv(time,status)~group2+group4,data=group,dist='exponential')$coef
mod10<-survreg(Surv(time,status)~group2+group4,data=group,dist='exponential')
value10<-extractAIC(mod10)
mod[9,1:3]<-survreg(Surv(time,status)~group1+group4,data=group,dist='exponential')$coef
mod9<-survreg(Surv(time,status)~group1+group4,data=group,dist='exponential')
value9<-extractAIC(mod9)
mod[8,1:3]<-survreg(Surv(time,status)~group1+group3,data=group,dist='exponential')$coef
mod8<-survreg(Surv(time,status)~group1+group3,data=group,dist='exponential')
value8<-extractAIC(mod8)
mod[7,1:3]<-survreg(Surv(time,status)~group3+group4,data=group,dist='exponential')$coef
mod7<-survreg(Surv(time,status)~group3+group4,data=group,dist='exponential')
value7<-extractAIC(mod7)
mod[6,1:3]<-survreg(Surv(time,status)~group2+group3,data=group,dist='exponential')$coef
mod6<-survreg(Surv(time,status)~group2+group3,data=group,dist='exponential')
value6<-extractAIC(mod6)
mod[5,1:3]<-survreg(Surv(time,status)~group1+group2,data=group,dist='exponential')$coef
mod5<-survreg(Surv(time,status)~group1+group2,data=group,dist='exponential')
value5<-extractAIC(mod5)
mod[4,1:2]<-survreg(Surv(time,status)~group4,data=group,dist='exponential')$coef
mod4<-survreg(Surv(time,status)~group4,data=group,dist='exponential')
value4<-extractAIC(mod4)
mod[3,1:2]<-survreg(Surv(time,status)~group3,data=group,dist='exponential')$coef
mod3<-survreg(Surv(time,status)~group3,data=group,dist='exponential')
value3<-extractAIC(mod3)
mod[2,1:2]<-survreg(Surv(time,status)~group2,data=group,dist='exponential')$coef
mod2<-survreg(Surv(time,status)~group2,data=group,dist='exponential')
value2<-extractAIC(mod2)
mod[1,1:2]<-survreg(Surv(time,status)~group1,data=group,dist='exponential')$coef
mod1<-survreg(Surv(time,status)~group1,data=group,dist='exponential')
value1<-extractAIC(mod1)
```

```
p<-16
store1<-matrix(0,15,p)
store1[1,1:p]<-c(store[1,1],store[1,2],0,0,0,store[1,3],-
log(store[1,3]),percent,0,value1[2],mod[1,2],0,0,0,mod[1,1],store[1,17])
store1[2,1:p]<-c(store[2,1],0,store[2,2],0,0,store[2,3],-
log(store[2,3]),percent,0,value2[2],0,mod[2,2],0,0,mod[2,1],store[2,17])
store1[3,1:p]<-c(store[3,1],0,0,store[3,3],0,store[3,3],-
log(store[3,3]),percent,0,value3[2],0,0,mod[3,2],0,mod[3,1],store[3,17])
store1[4,1:p]<-c(store[4,1],0,0,0,store[4,2],store[4,3],-
log(store[4,3]),percent,0,value4[2],0,0,0,mod[4,2],mod[4,1],store[4,17])
store1[5,1:p]<-c(store[5,1],store[5,2],store[5,3],0,0,store[5,4],-
log(store[5,4]),percent,0,value5[2],mod[5,2],mod[5,3],0,0,mod[5,1],store[5,17])
store1[6,1:p]<-c(store[6,1],0,store[6,2],store[5,3],0,store[6,4],-
log(store[6,4]),percent,0,value6[2],0,mod[6,2],mod[6,3],0,mod[6,1],store[6,17])
store1[7,1:p]<-c(store[7,1],0,0,store[7,2],store[7,3],store[7,4],-
log(store[7,4]),percent,0,value7[2],0,0,mod[7,2],mod[7,3],mod[7,1],store[7,17])
store1[8,1:p]<-c(store[8,1],store[8,2],0,store[8,3],0,store[8,4],-
log(store[8,4]),percent,0,value8[2],mod[8,2],0,mod[8,3],0,mod[8,1],store[8,17])
store1[9,1:p]<-c(store[9,1],store[9,2],0,0,store[9,3],store[9,4],-
log(store[9,4]),percent,1,value9[2],mod[9,2],0,0,mod[9,3],mod[9,1],store[9,17])
store1[10,1:p]<-c(store[10,1],0,store[10,2],0,store[10,3],store[10,4],-
log(store[10,4]),percent,0,value10[2],0,mod[10,2],0,mod[10,3],mod[10,1],store[10,17])
store1[11,1:p]<-c(store[11,1],store[11,2],store[11,3],store[11,4],0,store[11,5],-
log(store[11,5]),percent,0,value11[2],mod[11,2],mod[11,3],mod[11,4],0,mod[11,1],store[11,17]
)
store1[12,1:p]<-c(store[12,1],0,store[11,2],store[11,3],store[11,4],store[12,5],-
log(store[12,5]),percent,0,value12[2],0,mod[12,2],mod[12,3],mod[12,4],mod[12,1],store[12,17]
)
store1[13,1:p]<-c(store[13,1],store[13,2],store[13,3],0,store[13,4],store[13,5],-
log(store[13,5]),percent,0,value13[2],mod[13,2],mod[13,3],0,mod[13,4],mod[13,1],store[13,17]
)
store1[14,1:p]<-c(store[14,1],store[14,2],0,store[14,3],store[14,4],store[14,5],-
log(store[14,5]),percent,0,value14[2],mod[14,2],0,mod[14,3],mod[14,4],mod[14,1],store[14,17]
)
store1[15,1:p]<-c(store[15,1],store[15,2],store[15,3],store[15,4],store[15,5],store[15,6],-
log(store[15,6]),percent,0,value15[2],mod[15,2],mod[15,3],mod[15,4],mod[15,5],mod[15,1],sto
re[15,17])

return(store1)
}
#=====================================================================
#Δημιουργούμε τον πίνακα  res με διαστάσεις 15*16 και την 3 διάσταση καθορίζεται αναλογά
#με τον αριθμό που θέλουμε να τρέξουμε το μοντέλο .Καλούμε την συνάρτηση data  και το
#αποτέλεσμα της το αποθηκέυουμαι σε έναν υποπίνακα του  πίνακα res

k<-100  #Πόσες φορές θα τρέξει το μοντέλο k>2
res<-array(dim=c(15,16,k))
for(i  in 1:k){
res[,,i]<-data()
}
res

minaic<-rep(0,k)
minAICc<-rep(0,k)
minSurvaic<-rep(0,k)
inds<-matrix(0,k,2)
inds2<-matrix(0,k,2)
inds4<-matrix(0,k,2)
par<-matrix(0,k,16)
parSur<-matrix(0,k,16)
```

```
#Πίνακας των μοντέλων.
W<-c("Z_1","Z_2","Z_3","Z_4","Z_1-Z_2","Z_2-Z_3","Z_3-Z_4","Z_1-Z_3","Z_1-Z_4","Z_2-
Z_4","Z_1-Z_2-Z_3","Z_2-Z_3-Z_4","Z_1-Z_2-Z_4","Z_1-Z_3-Z_4","Z_1-Z_2-Z_3-Z_4")

x<-rep(0,k)
y<-rep(0,k)
z<-rep(0,k)
w<-rep(0,k)
h<-rep(0,k)

for(i in 1:k){
minaic[i]<-min(res[,1,i])
minAICc[i]<-min(res[,16,i])
minSurvaic[i]<-min(res[,10,i])
inds[i,]<-which( res[,1,i]== min(res[,1,i]),arr.ind=TRUE)
inds2[i,]<-which( res[,16,i]== min(res[,16,i]),arr.ind=TRUE)
inds4[i,]<-which( res[,10,i]== min(res[,10,i]),arr.ind=TRUE)
x[i]<-W[inds[i,1]]
y[i]<-W[inds2[i,1]]
z[i]<-inds[i,1]-inds2[i,1]
h[i]<-W[inds4[i,1]]
par[i,]<-res[inds[i,1],,i]
parSur[i,]<-res[inds4[i,1],,i]
}
x
y
z  #Μη μηδενικά στοιχεία του z δείχνουν ότι το  AICc και το AIC δεν έχουν επιλέξει το ίδιο
   #μοντέλο
h       #ποιά μοντέλα διαλέγει  η survreg
parSur
v<-rep(0,k)
for(i in 1:k){
if(z[i]!=0)
v[i]<-1
else
v[i]<-0
}
diafora<-sum(v)/k
dimnames(par) = list(NULL,c("AICvalue","beta1","beta2","beta3","beta4",
"Lamda","-
log(lamda)","UnCensor","CorrectModel","AICsurv","Surbeta1","Surbeta2","Surbeta3","Surbeta
4","Surlamda","AICc"))

par #Πίνακας που κάθε γραμμή του είναι η γραμμή του εκάστωτε store1 πινάκα με την
#μικρότερη τιμή για το AIC κριτήριο
minaic
minAICc

#Επιλέγεται το  Z1-Z4
percent<-sum(par[,9])/k
percent
rest<-1-percent
rest

percent2<-sum(parSur[,9])/k
percent2

number1<-length(par[par[,2]==0,2])
number1
```

```
number2<-length(par[par[,2]!=0,2])
number2
mbeta_1<- sum(par[,2])/(k-number1)
mbeta_1
sd_beta_1<-sd(par[par[,2]!=0,2])

number3<-length(par[par[,5]==0,5])
number3
number4<-length(par[par[,5]!=0,5])
number4
mbeta_4<-sum(par[,5])/(k-number3)
mbeta_4
sd_beta_4<-sd(par[par[,5]!=0,5])

mlamda<-mean(par[,6])
mlamda
m_log_lamda<-mean(par[,7])
#=========================================================================
#Βρίσκει τον αριθμό που εμφανίζονται οι Z2,Z3
value<-rep(0,k)
for( i in 1:k){
if (par[i,3]!=0 | par[i,4]!=0 )
value[i]<-1
else
value[i]<-0
}
value
error1<-sum(value)/k #Να επιλέγει το Z2 ή το Z3 ή και τα δύο μαζί
error1
#=========================================================================
#Υπολογισμός error2
zeros<-matrix(0,k,4)
ermat<-cbind(par[1:k,2:5],zeros)
for( i in 1:k){
if(  ermat[i,1]==0)   #Να είναι το β1 μήδεν αποθηκέυει την τιμή 1 στην στήλη 6
ermat[i,6]<-1
else
ermat[i,6]<-0
}
ermat

for( i in 1:k){
if( ermat[i,4]==0   #Να είναι το β4 μήδεν αποθηκέυει την τιμή 1 στην στήλη 5
ermat[i,5]<-1
else
ermat[i,5]<-0
}
ermat

for( i in 1:k){
if(ermat[i,1]==0 && ermat[i,4]==0)  #Να είναι μηδέν το  β1 και το β4και να αποθηκέυει την τιμή
#στην  7 στήλη
ermat[i,7]<-1
else
ermat[i,7]<-0
}
ermat
for(i in 1:k){
ermat[i,8]<-sum(ermat[i,5:7])  #Αθροίζει την στήλη 5,6,7 στην 8 ανα γραμμή
}
```

```
error2<-sum(ermat[,8])/k
error2
#=====================================================================
#Υπολογισμός Surerror1 kai Surerror2
svalue<-rep(0,k)
for( i in 1:k){
if (parSur[i,3]!=0 | parSur[i,4]!=0 )
svalue[i]<-1
else
svalue[i]<-0
}
svalue
Surerror1<-sum(svalue)/k  #Να επιλεγεταί ή  Z2 ή   Z3 ή και οι δύο μαζί
Surerror1

#Υπολογισμός Surerror2
zeros<-matrix(0,k,4)
Surermat<-cbind(parSur[1:k,2:5],zeros)
for (i in 1:k){
if (Surermat[i,1]==0)   #Να είναι το β1 μήδεν αποθηκέυει την τιμή 1 στην στήλη 6
Surermat[i,6]<-1
else
Surermat[i,6]<-0
}
Surermat

for( i in 1:k){
if (Surermat[i,4]==0)  #Να είναι το β4 μήδεν αποθηκέυει την τιμή 1 στην στήλη 5
Surermat[i,5]<-1
else
Surermat[i,5]<-0
}
Surermat

for( i in 1:k){
if(Surermat[i,1]==0 && Surermat[i,4]==0)
# Να είναι μηδέν το β1  και το β4 αποθηκευεί την τιμή 1 στην στήλη 7
Surermat[i,7]<-1
else
Surermat[i,7]<-0
}
Surermat
for (i in 1:k){
Surermat[i,8]<-sum(Surermat[i,5:7]) #Αθροίζει τι στήλες 5,6,7 στην  8 ανα γραμμή
}
Surerror2<-sum(Surermat[,8])/k
Surerror2

#=====================================================================
number9<-length(par[par[,11]==0,11])
number9
number10<-length(par[par[,11]!=0,11])
number10
smbeta_1<- sum(par[,11])/(k-number9)
smbeta_1
sd_sbeta_1<-sd(par[par[,11]!=0,11])
sd_sbeta_1

number11<-length(par[par[,14]==0,14])
number11
```

```
number12<-length(par[par[,14]!=0,14])
number12
smbeta_4<- sum(par[,14])/(k-number11)
smbeta_4
sd_sbeta_4<-sd(par[par[,14]!=0,14])
sd_sbeta_4
#=====================================================================
#Πίνακας Αποτελεσμάτων
resu2<-matrix(0,1,19)
resu2[1,1]<-mean(par[,8])
resu2[1,2]<-percent
resu2[1,3]<-rest
resu2[1,4]<-error1
resu2[1,5]<-error2
resu2[1,6]<-mbeta_1
resu2[1,7]<-mbeta_4
resu2[1,8]<-sd_beta_1
resu2[1,9]<-sd_beta_4
resu2[1,10]<-m_log_lamda
resu2[1,11]<-diafora
resu2[1,12]<-mean(parSur[,15])
resu2[1,13]<-smbeta_1
resu2[1,14]<-smbeta_4
resu2[1,15]<-percent2
resu2[1,16]<-sd_sbeta_1
resu2[1,17]<-sd_sbeta_4
resu2[1,18]<-Surerror1
resu2[1,19]<-Surerror2
dimnames(resu2) = list(NULL,c("UnCensor","Percent","1-Percent","Error1","Error2","Mbeta_1"
,"Mbeta_4","Sd(beta_1)","Sd(beta_4)","MlogL","diafora","SurLamda","SurBeta_1","Surbeta_4"
,"Percent2","Sd_sbeta_1","Sd_xbeta_4","Surerror1","Surerror2"))

resu2
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

Androulakis, E., Koukouvinos, C. and Vonta, F. (2012). Estimation and variable selection via the penalized frailty model, Statistics in Medicine, (to appear).

Barker, P.,Henderson, R. (2004). Modeling converging hazards in survival Analysis, Lifetime Data Analysis, 10, 263–281.

Barker, P., Henderson, R. (2005). Small sample bias in the gamma frailty model for univariate survival, Lifetime Data Analysis, 11, 265–284.

Brockwell, P.J., and Davis, R.A. (2009). Time Series: Theory and Methods, 2nd ed. Springer.

Burnham,K.P., and Anderson, D.R (2004) "Multimodel Inference": understanding AIC and BIC in Model Selection" Sociological Methods and Research",33: 261-304.

Clayton, D.G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, Biometrika, 65, 141–151.

Caroni, C. (2009). Reliability and Survival models, Simeon Editions.

Duchateau, L., Janssen, P. (2008). The Frailty Model, Springer, New York.

Economou, P., Caroni, C. (2005) Graphical tests for the assumption of gamma and inverse Gaussian frailty distributions, Lifetime Data Analysis, 11, 565–582.

Economou, P., Caroni, C. (2008) Graphical tests for the frailty distribution in the shared frailty model. 11, 565–582.

Fokianos, K and Xaralampos, X. (2008). Introduction to R. 1-130.

Huber-Carol, C and Vonta, F. (2004). Frailty models for arbitrarily censored and truncated data, Lifetime Data Analysis, 10(4), 369-388.

Kheiri, S., Kimber, A., Meshkani, M.R. (2007) Bayesian analysis of an inverse Gaussian correlated frailty model. Computational Statistics and Data Analysis 51, 5317–5326.

Manton, K., Stallard, E., Vaupel, J. (1981). Methods for comparing mortality experience of heterogeneous populations, Demography, 18, 389–410.

Manton, K., Stallard, E., Vaupel, J. (1986). Alternative models for

heterogeneity of mortality risks among the aged, Journal of the American Statistical Association 81, 635–644.

Macquarie, A. D. R., and Tsai, C.-L. (1998). Regression and Time Series Model Selection.World Scientific.

Nielsen, G.G., Gill, R.D., Andersen, P.K., Sørensen, T.I.A. (1992). A counting process approach to maximum likelihood estimation in frailty models, Scandinavian Journal of Statistics, 19-25.

Schwarz, Gideon E. (1978)."Estimating the dimension of a model". Annals of Statistics 6 (2): 461–464.

Yashin, A.I., Iachine, I.A. (1995a). Genetic analysis of durations: Correlated frailty model applied to survival of Danish twins, Genetic Epidemiology, 12, 529–538.

Yashin, A.I., Iachine, I.A. (1995b). Survival of related individuals: an extension of some fundamental results of heterogeneity analysis, Mathematical Population Studies, 5,321-39.

Yashin, A.I., Iachine, I.A. (1996) Random effect models of bivariate survival: quadratic hazard as a new alternative. In: Transactions of Symposium i Anvendt Statistik. G. Kristensen (ed.), Institute of Economy, Odense University, pp. 87–101.

Vaupel, J., Manton, K., Stallard, E. (1979).The impact of heterogeneity in individual frailty on the dynamics of mortality, Demography, 16,439–54.

Vaupel, J.W., Yashin, A.I. (1985). Heterogeneity's ruses: some surprising effects of selection on population dynamics, The American Statistician 39, 176–185.