ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Δ.Π.Μ.Σ. «ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ»

ΓΕΝΙΚΕΥΜΕΝΗ ΣΥΝΑΡΤΗΣΗ ΠΙΘΑΝΟΦΑΝΕΙΑΣ ΓΙΑ ΜΟΝΤΕΛΑ ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΥ ΚΑΙ ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΩΝ

Διπλωματική Εργασία

ΑΝΑΣΤΑΣΙΟΣ ΛΥΜΠΕΡΑΤΟΣ

Επιβλέπουσα: ΦΙΛΙΑ ΒΟΝΤΑ, Επίκουρος Καθηγήτρια Ε.Μ.Π.

Αθήνα 2013

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Δ.Π.Μ.Σ. «ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ»

ΓΕΝΙΚΕΥΜΕΝΗ ΣΥΝΑΡΤΗΣΗ ΠΙΘΑΝΟΦΑΝΕΙΑΣ ΓΙΑ ΜΟΝΤΕΛΑ ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΥ ΚΑΙ ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΩΝ

Διπλωματική Εργασία

ΑΝΑΣΤΑΣΙΟΣ ΛΥΜΠΕΡΑΤΟΣ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή στις ...../ ...../20….

.....	.....	.....
(υπογραφή)	*(υπογραφή)*	*(υπογραφή)*
ΦΙΛΙΑ ΒΟΝΤΑ	**ΧΡΥΣΗΙΣ ΚΑΡΩΝΗ**	**ΧΡΗΣΤΟΣ ΚΟΥΚΟΥΒΙΝΟΣ**
Επίκουρος Καθηγήτρια	Καθηγήτρια	Καθηγητής

……..

Αναστάσιος Λυμπεράτος

Μαθηματικός

Πτυχιούχος Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα διπλωματική εργασία εκπονήθηκε για την ολοκλήρωση του μεταπτυχιακού προγράμματος **"ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ"** της ΣΧΟΛΗΣ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ του ΕΘΝΙΚΟΥ ΜΕΤΣΟΒΙΟΥ ΠΟΛΥΤΕΧΝΕΙΟΥ.

Θα ήθελα εδώ να ευχαριστήσω θερμά την επιβλέπουσα καθηγήτρια κα. Φιλία Βόντα για την ανάθεση της εργασίας, την εμπιστοσύνη της στο πρόσωπό μου, την υπομονή της καθ' όλη τη διάρκεια της έρευνας και τις καίριες παρεμβάσεις και υποδείξεις της.

Εκφράζω επίσης θερμές ευχαριστίες στην καθηγήτρια κα. Χρύσα Καρώνη και στον καθηγητή κ. Χρήστο Κουκουβίνο για την τιμή να δεχθούν να συμμετάσχουν στην επιτροπή αξιολόγησης της εργασίας.

Κλείνοντας τέλος αυτόν τον κύκλο μεταπτυχιακών σπουδών, ευχαριστώ το Ε.Μ.Π. και τη Σ.Ε.Μ.Φ.Ε. για την ευκαιρία που μου έδωσαν να διευρύνω τους ορίζοντές μου μέσα από τα μαθήματα, τις εργασίες και την έρευνα. Ευχαριστώ επίσης τους διδάσκοντες του μεταπτυχιακού προγράμματος για τη γνώση που μου χάρισαν και για την ακούραστη καθοδήγησή τους στο αχανές πεδίο της επιστήμης των Μαθηματικών.

Η εργασία αφιερώνεται στη Χριστίνα και την Ευρυδίκη, για την απεριόριστη αγάπη, την υπομονή και τη στήριξή τους.

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΛΗΨΗ

Η παρούσα εργασία αποτελείται από τέσσερα κεφάλαια: Στο πρώτο, παρουσιάζονται οι εισαγωγικές έννοιες της Ανάλυσης Επιβίωσης και τα μοντέλα διάρκειας ζωής (παραμετρικά και μη). Στο δεύτερο, μετά από μία γενική θεώρηση της προσαρμογής των μοντέλων παλινδρόμησης στα δεδομένα διάρκειας ζωής, παρουσιάζεται εκτενώς το μοντέλο αναλογικών κινδύνων του Cox (*Cox proportional hazards model*). Στη συνέχεια, το μοντέλο του Cox εμπλουτίζεται με την εισαγωγή σε αυτό της τυχαίας μεταβλητής της ευπάθειας, της οποίας μοντέλα μελετούμε στο κεφ. 3 (*frailty models*). Στο κεφ. 4 παρουσιάζονται τα μοντέλα μετασχηματισμού (*transformation models*) ως επέκταση αυτών της ευπάθειας. Τέλος, μελετάται η δημιουργία και εκτέλεση προγραμμάτων στην R για την επιλογή μεταβλητών και την εκτίμηση παραμέτρων στα μοντέλα μετασχηματισμού, με χρήση της γενικευμένης συνάρτησης πιθανοφάνειας και των κριτηρίων AIC και BIC.

ABSTRACT

This thesis consists of four chapters: The first presents the basic concepts of Survival Analysis and the modeling of time-to-event data. In the second chapter, after an overview of the adjustment of regression models to lifetime data, we study extensively the proportional hazards model of Cox. Then the Cox model is enriched by the introduction of the frailty random variable, models of which we study in chapter 3 (frailty models). In chapter 4 we present the transformation models as an extension of the frailty models. Finally, we study the creation and execution of programs in R for the selection of variables and the estimation of parameters in transformation models, using the generalized likelihood function and the AIC and BIC criteria.

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΑΝΑΛΥΣΗ ΕΠΙΒΙΩΣΗΣ

1.1. ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ

Ανάλυση επιβίωσης είναι ο κλάδος της Στατιστικής που μελετά δεδομένα που αφορούν το χρόνο που απαιτείται μέχρι να συμβεί ένα γεγονός (*time-to-event data*), όπως π.χ. ο θάνατος ενός βιολογικού οργανισμού ή η βλάβη ενός μηχανικού συστήματος. Γενικότερα, η ανάλυση επιβίωσης μελετά μη αρνητικές τυχαίες μεταβλητές όπως π.χ. ο χρόνος μέχρι τον θάνατο, ο χρόνος μέχρι να εμφανισθεί κάποια αρρώστια, αλλά και μεταβλητές που δεν είναι χρόνος, όπως π.χ. το φορτίο που πρέπει να ασκηθεί ώστε να έχουμε θράυση κάποιου υλικού. Η ανάλυση επιβίωσης απαντάται και ως **θεωρία** ή **ανάλυση αξιοπιστίας** *(reliability theory)* όταν αφορά τα μηχανικά συστήματα.

Όπως προαναφέραμε, ο χρόνος επιβίωσης T είναι μία τυχαία μεταβλητή με μη αρνητικές τιμές ($T \geq 0$), η οποία μπορεί να είναι είτε **διακριτή** είτε **συνεχής**. Στη συνέχεια, θα θεωρήσουμε ότι ο χρόνος T είναι συνεχής τυχαία μεταβλητή (με τιμές, φυσικά στο $[0, \infty)$).

Οι βασικές συναρτήσεις που περιγράφουν τη διάρκεια ζωής T είναι:

- Η **συνάρτηση πυκνότητας πιθανότητας** (*probability density function – p.d.f.*) έστω $f(t)$, $t \geq 0$.
- Η **(αθροιστική) συνάρτηση κατανομής** (*cumulative distribution function – c.d.f.*) που συμβολίζεται με $F(t)$ και εκφράζει την πιθανότητα $P[T \leq t]$ ο χρόνος ζωής να μην υπερβεί μία συγκεκριμένη χρονική στιγμή t, δηλ.

$$\boxed{F(t) = P[T \leq t] = \int_0^t f(s)ds} \quad \textbf{(1.1)}$$

Η $F(t)$, ως συνάρτηση κατανομής είναι εξ' ορισμού αύξουσα, με $F(0) = 0$ και $F(\infty) = 1$.

- Η **συνάρτηση αξιοπιστίας** ή **συνάρτηση επιβίωσης** (*survival function*), η οποία συμβολίζεται με $S(t)$ και εκφράζει την πιθανότητα $P[T > t]$ ο χρόνος ζωής να υπερβεί τη χρονική στιγμή t, δηλ.

$$\boxed{S(t) = P[T > t] = 1 - F(t) = \int_t^{\infty} f(s)ds} \quad \textbf{(1.2)}$$

Προφανώς είναι $S(0) = 1 - F(0) = 1$.

- Η **συνάρτηση κινδύνου** (*hazard function*), συμβολίζεται με $h(t)$ και καθορίζει (βλ. [1], [2]) τον κίνδυνο διακοπής (θανάτου) μίας μονάδας αμέσως μετά τη χρονική στιγμή t, δοθέντος ότι αυτή έζησε έως εκείνη τη στιγμή, δηλ.

$$\boxed{h(t) = \lim_{dt \to 0} \frac{P[t < T \le t + dt / T > t]}{dt}} \quad \textbf{(1.3)}$$

Παρατηρούμε εδώ ότι από τον ορισμό της δεσμευμένης πιθανότητας είναι $P[t < T \le t + dt / T > t] = \frac{P[t < T \le t + dt]}{P[T > t]}$, οπότε χρησιμοποιώντας τις συναρτήσεις κατανομής και επιβίωσης, παίρνουμε $P[t < T \le t + dt / T > t] = \frac{F(t + dt) - F(t)}{S(t)} \approx \frac{f(t)dt}{S(t)}$ (βλ. [3]) και αντικαθιστώντας στον τύπο **(1.3)** ορισμού της συνάρτησης κινδύνου, προκύπτει η βασική σχέση

$$\boxed{h(t) = \frac{f(t)}{S(t)}} \quad \textbf{(1.4)}$$

που συνδέει τη συνάρτηση κινδύνου $h(t)$ με τις συναρτήσεις πυκνότητας πιθανότητας $f(t)$ και επιβίωσης $S(t)$.

- Τέλος, η **σωρευτική συνάρτηση κινδύνου** (*cumulative hazard function*) συμβολίζεται με $H(t)$ και ορίζεται από τη σχέση

$$\boxed{H(t) = \int_0^t h(s)ds} \quad \textbf{(1.5)}$$

Η σωρευτική συνάρτηση κινδύνου συνδέεται με τη συνάρτηση επιβίωσης, αφού

$$H(t) = \int_0^t h(s)ds = \int_0^t \frac{f(s)}{S(s)}ds = \int_0^t \frac{-S'(s)}{S(s)}ds = -[\ln S(s)]_0^t = \ln S(0) - \ln S(t) = -\ln S(t)$$

και επομένως έχουμε

$$\boxed{H(t) = -\ln S(t)} \quad \textbf{(1.6)}$$

άρα

$$\boxed{S(t) = e^{-H(t)}} \quad \textbf{(1.7)}$$

1.2. ΛΟΓΟΚΡΙΜΕΝΑ ΔΕΔΟΜΕΝΑ

Λογοκρισία (*censoring*) είναι η κατάσταση στην οποία η διάρκεια ζωής μίας παρατήρησης είναι μεγαλύτερη από τη διάρκεια του πειράματος. Έτσι, λογοκρισία έχουμε όταν π.χ. ένα πείραμα τελειώνει αλλά μερικές μονάδες εξακολουθούν να λειτουργούν, τότε, αν και δε γνωρίζουμε τον ακριβή χρόνο ζωής τους, έχουμε την πληροφορία ότι η ζωή των μονάδων ξεπέρασε τη διάρκεια του πειράματος.

Τα κυριότερα είδη λογοκρισίας είναι τα εξής:

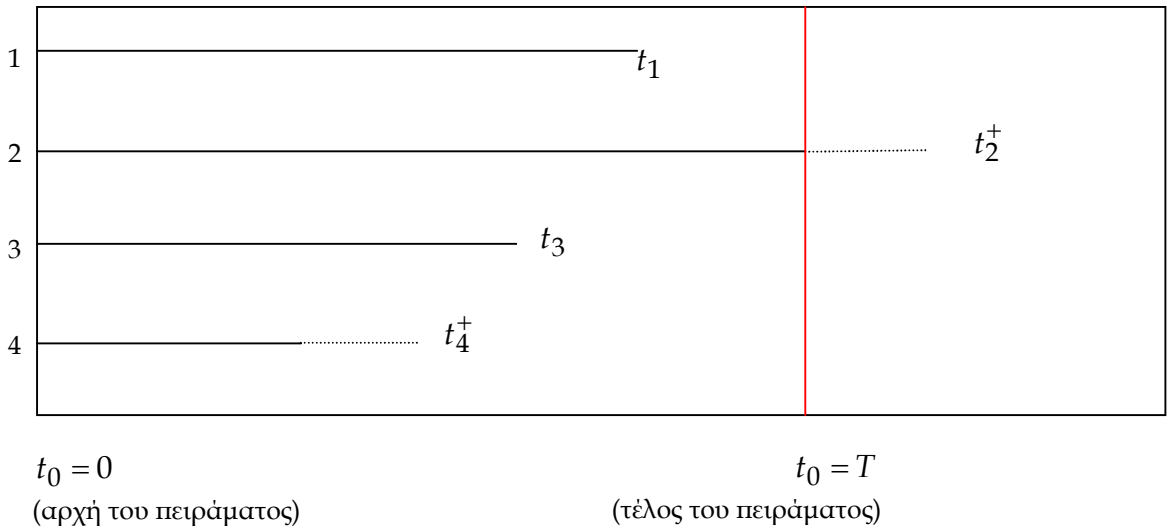
- **Δεξιά λογοκρισία** (*right censoring*): είναι η συνήθης μορή λογοκρισίας. Σε αυτήν, ο χρόνος ζωής της μονάδας είναι μεγαλύτερος από το χρόνο του πειράματος. Έτσι, αν c είναι ο χρόνος του πειράματος, τότε στη δεξιά λογοκρισία ισχύει $T > c$.
- **Αριστερή λογοκρισία** (*left censoring*): ο χρόνος ζωής της μονάδας είναι μικρότερος από το χρόνο του πειράματος, δηλ. ισχύει $T < c$. Αυτό μπορεί να συμβεί όταν η μονάδα δεν εισέρχεται στο πείραμα από την αρχή, αλλά σε κάποιο ενδιάμεσο χρονικό σημείο.
- **Λογοκρισία σε διάστημα** (*interval censoring*): για το χρόνο ζωής ισχύει $c_1 < T < c_2$.

Ένα παράδειγμα των ειδών λογοκρισίας, είναι το εξής: Στην ερώτηση "Σε ποια ηλικία κάπνισες για πρώτη φορά;" θα μπορούσαμε να έχουμε τις εξής δυνατές απαντήσεις (βλ. [4]):

1. "Δεν έχω καπνίσει" (δεξιά λογοκριμένη παρατήρηση: το άτομο μπορεί να καπνίσει μετά το τέλος της έρευνας).
2. "Στην ηλικία των …. ετών" (μη λογοκριμένη παρατήρηση).

3. "Δεν θυμάμαι πότε κάπνισα για πρώτη φορά" (αριστερά λογοκριμένη παρατήρηση).

Παραθέτουμε ακολούθως ένα παράδειγμα γραφικής απεικόνισης των δεξιά λογοκριμένων δεδομένων:



Σχ. 1: Σχηματική αναπαράσταση πλήρων και δεξιά λογοκριμένων δεδομένων

Έτσι, στο παραπάνω σχ. 1, οι μονάδες (1) και (3) εισέρχονται στο πείραμα στην αρχή αυτού ($t_0 = 0$) και οι ζωές τους διακόπτονται στους χρόνους t_1 και t_3 αντίστοιχα, οπότε δημιουργούνται πλήρεις παρατηρήσεις. Η μονάδα (2) εισέρχεται επίσης στην αρχή του πειράματος, αλλά ζει και μετά το πέρας τούτου, οπότε δημιουργεί μία δεξιά λογοκριμένη παρατήρηση, την t_2^+. Τέλος, η μονάδα (4) υφίσταται και αυτή δεξιά λογοκρισία, αφού εισέρχεται στην αρχή του πειράματος και χάνεται από αυτό στο χρόνο (δεξιάς λογοκρισίας) t_4^+.

Από τα παραπάνω, προκύπτει άμεσα η ανάγκη ώστε στην αλγεβρική αναπαράσταση των δεδομένων διάρκειας ζωής να προσδιορίζεται η ύπαρξη ή μη, λογοκρισίας στο χρόνο ζωής κάθε μονάδας. Στη συνήθη περίπτωση της δεξιάς λογοκρισίας, αυτό υλοποιείται με την παρουσίαση κάθε παρατήρησης, ως ζεύγους (X_i, D_i), όπου:

- X_i είναι ο χρόνος ζωής ή λογοκρισίας της i παρατήρησης, ($i = 1, 2, \ldots, n$, όπου n το μέγεθος του δείγματος)
- D_i είναι η δείκτρια συνάρτηση λογοκρισίας (*censoring indicator*), η οποία ορίζεται ως: $D_i = \begin{cases} 1 & \text{αν ο χρόνος } X_i \text{ είναι χρόνος διακοπής} \\ 0 & \text{αν ο χρόνος } X_i \text{ είναι χρόνος λογοκρισίας} \end{cases}$.

1.3. ΠΑΡΑΜΕΤΡΙΚΑ ΜΟΝΤΕΛΑ ΔΙΑΡΚΕΙΑΣ ΖΩΗΣ

Στα παραμετρικά μοντέλα διάρκειας ζωής, υποθέτουμε ότι η τυχαία μεταβλητή του χρόνου ζωής T, ακολουθεί γνωστή κατανομή. Στη συνέχεια, αναφέρουμε τα κυριότερα παραμετρικά μοντέλα.

1.3.1. Η Εκθετική Κατανομή (Exponential Distribution): Όταν η διάρκεια ζωής T ακολουθεί την Εκθετική Κατανομή με παράμετρο $\lambda > 0$ (συμβ. $T \sim \varepsilon(\lambda)$), τότε η συνάρτηση πυκνότητας πιθανότητας αυτής είναι

$$\boxed{f_T(t) = f(t) = \lambda e^{-\lambda t}} \textbf{ (1.8)}, \text{ όπου } t > 0 .$$

Η συνάρτηση αξιοπιστίας είναι $S(t) = \int_t^\infty f(s)ds = \int_t^\infty \lambda e^{-\lambda s} ds = \left[-e^{-\lambda s} \right]_t^\infty$ άρα

$$\boxed{S(t) = e^{-\lambda t}} \textbf{ (1.9)}$$

οπότε η συνάρτηση κινδύνου είναι $h(t) = \frac{f(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}}$, άρα

$$\boxed{h(t) = \lambda} \textbf{ (1.10)}$$

Αν και η Εκθετική Κατανομή είναι απλή ως μοντέλο, το γεγονός της σταθερής συνάρτησης κινδύνου (ανεξάρτητης από την ηλικία της μονάδας βάσει της εξίσωσης **(1.10)**) είναι το μειονέκτημα της ως ικανής να περιγράψει δεδομένα διάρκειας ζωής.

Η μέση διάρκεια ζωής εδώ, είναι

$$E[T] = \int_0^\infty t f(t) dt = \int_0^\infty \lambda t e^{-\lambda t} dt = -\left[t e^{-\lambda t} \right]_0^\infty + \int_0^\infty e^{-\lambda t} dt = -\left[\left(t + \frac{1}{\lambda} \right) e^{-\lambda t} \right]_0^\infty \text{ άρα}$$

$$\boxed{E[T]=\frac{1}{\lambda}}\ \textbf{(1.11)}.$$

Επίσης, η $E\left[T^2\right]=\int_0^\infty t^2 f(t)dt=\int_0^\infty \lambda t^2 e^{-\lambda t}dt=-\left[t^2e^{-\lambda t}\right]_0^\infty+\int_0^\infty 2te^{-\lambda t}dt=-\left[t^2e^{-\lambda t}\right]_0^\infty+\frac{2}{\lambda}E[T]$ άρα $E\left[T^2\right]=\frac{2}{\lambda^2}$. Μπορούμε τώρα να υπολογίσουμε τη διασπορά, από τον τύπο: $V[T]=E\left[T^2\right]-E^2[T]$ άρα

$$\boxed{V[T]=\frac{1}{\lambda^2}}\ \textbf{(1.12)}.$$

1.3.2. Η κατανομή Weibull (Weibull Distribution): Είναι από τις σπουδαιότερες κατανομές δεδομένων ζωής. Χρησιμοποιείται συχνότατα στην Ανάλυση Επιβίωσης λόγω της μεγάλης ευελιξίας που θα δούμε ακολούθως ότι έχει η συνάρτηση κινδύνου της. Η κατανομή, πήρε το όνομά της από τον Σουηδό μαθηματικό και μηχανικό Waloddi Weibull (1887-1979), ο οποίος την παρουσίασε το 1951 (βλ. [7] και [8]). Η τυχαία μεταβλητή T ακολουθεί την κατανομή Weibull με **παράμετρο κλίμακας** $\alpha>0$ και **παράμετρο σχήματος** $\lambda>0$ (συμβ. $T\sim W(\alpha,\lambda)$), όταν η τυχαία μεταβλητή $Y=T^\lambda\sim\varepsilon\left(\frac{1}{\alpha^\lambda}\right)$. Η συνάρτηση πυκνότητας πιθανότητας $f_T(t)$, προκύπτει έτσι, μέσω της συνάρτησης κατανομής $F_T(t)$ του ανωτέρω μετασχηματισμού, αφού $F_T(t)=P[T\le t]=P\left[T^\lambda\le t^\lambda\right]=P\left[Y\le t^\lambda\right]=F_Y\left(t^\lambda\right)$, οπότε παραγωγίζοντας έχουμε $f_T(t)=\frac{d}{dt}F_T(t)=\frac{d}{dt}F_Y\left(t^\lambda\right)=\lambda t^{\lambda-1}f_Y\left(t^\lambda\right)$. Επειδή $Y=T^\lambda\sim\varepsilon\left(\frac{1}{\alpha^\lambda}\right)$, άρα $f_Y(t)=\frac{1}{\alpha^\lambda}e^{-\frac{t}{\alpha^\lambda}}$, άρα $f_T(t)=\lambda t^{\lambda-1}f_Y\left(t^\lambda\right)=\lambda t^{\lambda-1}\frac{1}{\alpha^\lambda}e^{-\frac{t^\lambda}{\alpha^\lambda}}$, απ' όπου παίρνουμε την τελική εξίσωση της συνάρτησης πυκνότητας πιθανότητας:

$$\boxed{f_T(t)=f(t)=\lambda\alpha^{-\lambda}t^{\lambda-1}e^{-\left(\frac{t}{\alpha}\right)^\lambda}},\ t>0\ \textbf{(1.13)}.$$

Δείξαμε παραπάνω ότι η συνάρτηση κατανομής της Weibull συνδέεται με τη συνάρτηση κατανομής της Εκθετικής κατανομής $\varepsilon\left(\frac{1}{\alpha^{\lambda}}\right)$, μέσω της εξίσωσης $F_T(t)=F_Y\left(t^{\lambda}\right)$. Από εδώ, μπορούμε να υπολογίσουμε τη συνάρτηση αξιοπιστίας της Weibull, αφού είναι $S_T(t)=1-F_T(t)=$ $=1-F_Y\left(t^{\lambda}\right)=S_Y\left(t^{\lambda}\right)$ και χρησιμοποιώντας το ότι $Y=T^{\lambda}\sim\varepsilon\left(\frac{1}{\alpha^{\lambda}}\right)$, άρα

$$\boxed{S_T(t)=S(t)=e^{-\left(\frac{t}{\alpha}\right)^{\lambda}}},\ t>0 \quad \textbf{(1.14)}.$$

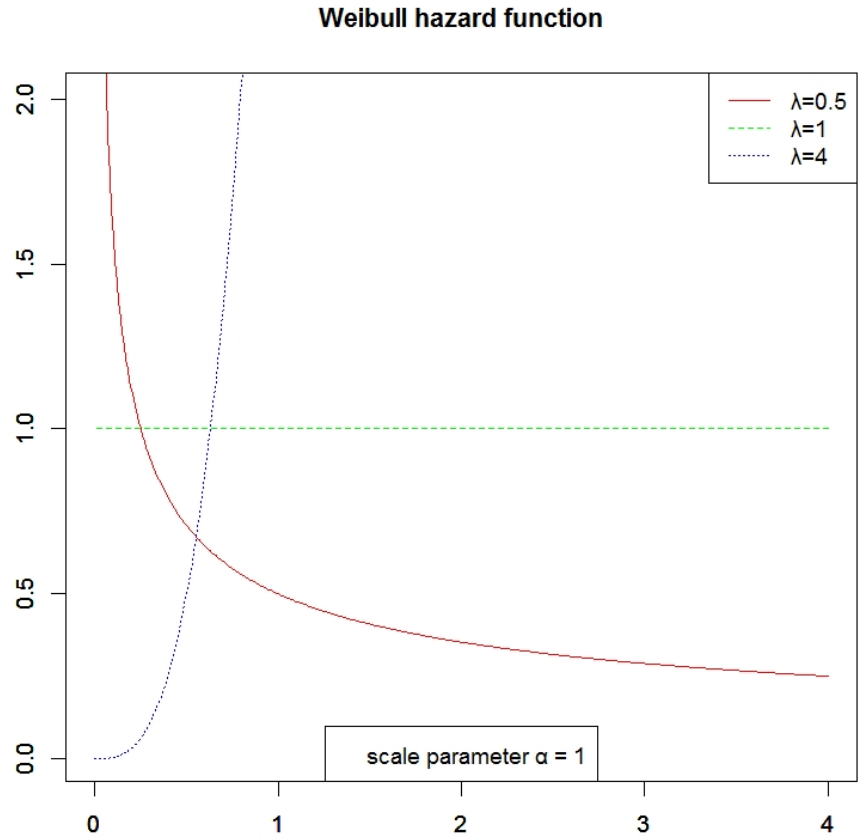
Η συνάρτηση κινδύνου της κατανομής Weibull είναι $h(t)=\frac{f(t)}{S(t)}$, άρα

$$\boxed{h(t)=\lambda\alpha^{-\lambda}t^{\lambda-1}},\ t>0 \quad \textbf{(1.15)}.$$

Παραγωγίζοντας τη συνάρτηση κινδύνου, έχουμε $\frac{d}{dt}h(t)=\lambda(\lambda-1)\alpha^{-\lambda}t^{\lambda-2}$ με $t>0$, άρα η παράγωγος $\frac{d}{dt}h(t)$ είναι ομόσημη του παράγοντα $\lambda-1$, άρα η μονοτονία της $h(t)$ είναι η ακόλουθη:

$$\text{Weibull } h(t)\rightarrow\begin{cases}\text{γνησίως αύξουσα} & \text{αν } \lambda>1\\ \text{σταθερή} & \text{αν } \lambda=1\\ \text{γνησίως φθίνουσα} & \text{αν } 0<\lambda<1\end{cases}.$$

Χρησιμοποιώντας τη γλώσσα προγραμματισμού R, μπορούμε να κατασκευάσουμε στο ίδιο σύστημα αξόνων, συναρτήσεις κινδύνου της κατανομής Weibull, για διάφορες τιμές της παραμέτρου σχήματος λ.



Σχ. 2: Συνάρτηση κινδύνου της κατανομής Weibull, για παράμετρο κλίμακας α = 1 και για διάφορες τιμές της παραμέτρου σχήματος λ

Παρατηρούμε ότι η κατανομή Weibull για παράμετρο σχήματος $\lambda = 1$ και παράμετρο κλίμακας $\frac{1}{\alpha}$, $\alpha > 0$, ταυτίζεται με την Εκθετική κατανομή $\varepsilon(\alpha)$ ή ισοδύναμα, η Εκθετική κατανομή είναι ειδική περίπτωση της κατανομής Weibull.

Η μέση τιμή και η διασπορά της κατανομής Weibull, υπολογίζονται από τη γενική ροπή $m-$τάξης περί την αρχή (βλ. [3]):

$$E\left[T^m\right] = \int_0^\infty t^m f(t)dt = \int_0^\infty t^m \lambda\alpha^{-\lambda} t^{\lambda-1} e^{-\left(\frac{t}{\alpha}\right)^\lambda} dt = \lambda\alpha^{-\lambda}\int_0^\infty t^{m+\lambda-1} e^{-\left(\frac{t}{\alpha}\right)^\lambda} dt.$$

Εφαρμόζουμε στον υπολογισμό του ολοκληρώματος την αντικατάσταση $u = \left(\frac{t}{\alpha}\right)^\lambda$ οπότε $t = \alpha u^{\frac{1}{\lambda}}$ και $dt = \frac{\alpha}{\lambda} u^{\frac{1}{\lambda}-1} du$ άρα

$$E\left[T^m\right] = \lambda\alpha^{-\lambda}\int_0^\infty \alpha^{m+\lambda-1} u^{\frac{m+\lambda-1}{\lambda}} e^{-u} \frac{\alpha}{\lambda} u^{\frac{1}{\lambda}-1} du = \alpha^m \int_0^\infty u^{\frac{m}{\lambda}} e^{-u} du.$$

Χρησιμοποιώντας στην παραπάνω σχέση τον ορισμό της συνάρτησης Γάμμα: $\Gamma(z)=\int_0^\infty u^{z-1}e^{-u}du$, παίρνουμε τελικά τη ροπή $m-$τάξης περί την αρχη:

$$E\left[T^m\right]=\alpha^m\Gamma\left(\frac{m}{\lambda}+1\right) \textbf{ (1.16)}$$

Η σχέση **(1.16)** για $m=1$ δίνει τη μέση τιμή της κατανομής Weibull:

$$\boxed{E[T]=\alpha\Gamma\left(\frac{1}{\lambda}+1\right)} \textbf{ (1.17)}$$

Επίσης, από τη γνωστή ισότητα $V[T]=E\left[T^2\right]-E^2[T]$ και τη σχέση **(1.16),** παίρνουμε τη διασπορά της κατανομής Weibull:

$$\boxed{V[T]=\alpha^2\left\{\Gamma\left(\frac{2}{\lambda}+1\right)-\Gamma^2\left(\frac{1}{\lambda}+1\right)\right\}} \textbf{ (1.18)}$$

1.3.3. Η κατανομή Gompertz: Χαρακτηρίζεται από την ιδιότητα ότι ο λογάριθμος της συνάρτησης κινδύνου είναι γραμμική συνάρτηση του χρόνου, οπότε

$$\boxed{h(t)=e^{a+bt}}, \text{ όπου } a,\ b \text{ σταθερές } \textbf{(1.19)}$$

και είναι άμεσο ότι η μονοτονία της συνάρτησης κινδύνου είναι

$$\text{Gompertz } h(t)\rightarrow\begin{cases}\text{γνησίως αύξουσα} & \text{αν } b>0\\ \text{σταθερή} & \text{αν } b=0\\ \text{γνησίως φθίνουσα} & \text{αν } b<0\end{cases}.$$

1.3.4. Η κατανομή Γάμμα (Gamma distribution): Θα λέμε ότι η τυχαία μεταβλητή T ακολουθεί την κατανομή Γάμμα με παράμετρο κλίμακας λ και παράμετρο σχήματος κ και θα συμβολίζουμε $T\sim G(\lambda,\kappa)$, όταν έχει συνάρτηση πυκνότητας πιθανότητας

$$\boxed{f_T(t)=f(t)=\frac{\lambda^\kappa}{\Gamma(\kappa)}t^{\kappa-1}e^{-\lambda t}}, \text{ με } t>0 \textbf{ (1.20)}.$$

Η συνάρτηση αξιοπιστίας δεν έχει κλειστή μορφή. Πράγματι, είναι $S(t)=\int_t^\infty f(t)dt=\frac{\lambda^\kappa}{\Gamma(\kappa)}\int_t^\infty t^{\kappa-1}e^{-\lambda t}dt$ και με την αντικατάσταση $u=\lambda t$ είναι

$S(t)=\frac{\lambda^{\kappa}}{\Gamma(\kappa)}\int_{\lambda t}^{\infty}\lambda^{-\kappa}u^{\kappa-1}e^{-u}du=\frac{1}{\Gamma(\kappa)}\int_{\lambda t}^{\infty}u^{\kappa-1}e^{-u}du$. Αν $\Gamma(\kappa,x)=\int_{x}^{\infty}u^{\kappa-1}e^{-u}du$ είναι η **άνω ατελής συνάρτηση Γάμμα** (*upper incomplete gamma function*), τότε η συνάρτηση επιβίωσης παίρνει την τελική (όχι κλειστή) μορφή:

$$\boxed{S(t)=\frac{\Gamma(\kappa,\lambda t)}{\Gamma(\kappa)}} \quad \textbf{(1.21)}$$

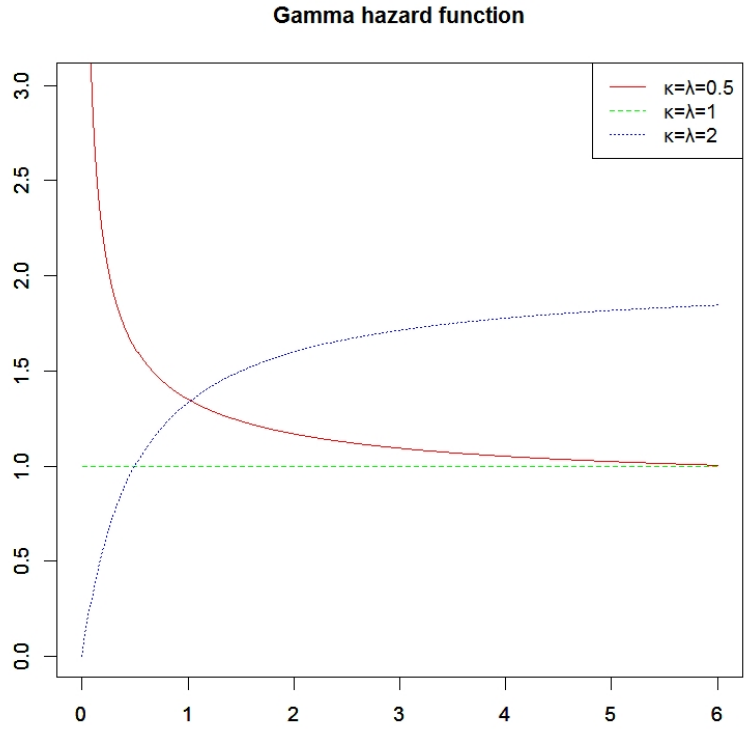
Η συνάρτηση κινδύνου της κατανομής Γάμμα είναι $h(t)=\frac{f(t)}{S(t)}$, άρα

$$\boxed{h(t)=\frac{\lambda^{\kappa}t^{\kappa-1}e^{-\lambda t}}{\Gamma(\kappa,\lambda t)}} \quad \textbf{(1.22)}$$

και έχει την ακόλουθη μονοτονία (βλ. [6]):

$$\text{Gamma } h(t)\rightarrow\begin{cases}\text{γνησίως αύξουσα} & \text{αν } \kappa>1\\ \text{σταθερή} & \text{αν } \kappa=1\\ \text{γνησίως φθίνουσα} & \text{αν } \kappa<1\end{cases}$$

Με τη βοήθεια της γλώσσας R, μπορούμε και εδώ να κατασκευάσουμε διαγράμματα της συνάρτησης κινδύνου για διάφορες τιμές των παραμέτρων σχήματος κ και κλίμακας λ.



Σχ. 3: Συνάρτηση κινδύνου της κατανομής Γάμμα, για διάφορες τιμές των παραμέτρων σχήματος κ και κλίμακας λ

Τόσο από το παραπάνω σχήμα, όσο και από τις συναρτήσεις πυκνότητας πιθανότητας, επιβίωσης και κινδύνου, είναι προφανές ότι για παράμετρο σχήματος $\kappa = 1$, η κατανομή Γάμμα ταυτίζεται με την Εκθετική κατανομή $\varepsilon(\lambda)$.

Η ροπή $m-$τάξης περί την αρχή για την κατανομή Γάμμα, είναι $E[T^m] = \int_0^\infty t^m f(t)dt = \frac{\lambda^\kappa}{\Gamma(\kappa)}\int_0^\infty t^{m+\kappa-1}e^{-\lambda t}dt = \frac{\lambda^{1-m}}{\Gamma(\kappa)}\int_0^\infty (\lambda t)^{m+\kappa-1}e^{-\lambda t}dt \overset{u=\lambda t}{=}$

$= \frac{\lambda^{-m}}{\Gamma(\kappa)}\int_0^\infty u^{m+\kappa-1}e^{-u}du$ άρα $E[T^m] = \frac{\Gamma(m+\kappa)}{\lambda^m\Gamma(\kappa)}$ και από εδώ υπολογίζονται η μέση τιμή της κατανομής Γάμμα: $E[T] = \frac{\Gamma(\kappa+1)}{\lambda\Gamma(\kappa)} = \frac{\kappa\Gamma(\kappa)}{\lambda\Gamma(\kappa)}$, άρα

$$\boxed{E[T] = \frac{\kappa}{\lambda}} \quad \textbf{(1.23)}$$

και η διασπορά: $V[T] = E[T^2] - E^2[T] = \frac{\Gamma(\kappa+2)}{\lambda^2\Gamma(\kappa)} - \frac{\kappa^2}{\lambda^2} = \frac{\kappa(\kappa+1)\Gamma(\kappa)}{\lambda^2\Gamma(\kappa)} - \frac{\kappa^2}{\lambda^2}$, άρα

$$\boxed{V[T] = \frac{\kappa}{\lambda^2}} \quad \textbf{(1.24)}$$

1.3.5. Η Λογαριθμο-κανονική κατανομή (Log-Normal distribution): Η τυχαία μεταβλητή T ακολουθεί τη Λογαριθμο-κανονική κατανομή με παραμέτρους μ και σ^2, όταν η τυχαία μεταβλητή $Y = \ln T$ ακολουθεί την Κανονική κατανομή $N(\mu, \sigma^2)$. Οι συναρτήσεις κατανομής και πυκνότητας πιθανότητας, μπορούν να υπολογισθούν μέσω του παραπάνω μετασχηματισμού, αφού $F_T(t) = P[T \le t] = P[Y \le \ln t] = P\left[\frac{Y-\mu}{\sigma} \le \frac{\ln t - \mu}{\sigma}\right] =$

$= \Phi_Z\left(\frac{\ln t - \mu}{\sigma}\right)$ όπου $\Phi_Z(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds$ είναι η συνάρτηση κατανομής τυχαίας μεταβλητής $Z \sim N(0,1)$. Παραγωγίζοντας, παίρνουμε τη συνάρτηση

πυκνότητας της Λογαριθμο-κανονικής κατανομής: $f_T(t) = f_Z\left(\frac{\ln t - \mu}{\sigma}\right) \cdot \frac{1}{\sigma t}$

και τελικά:

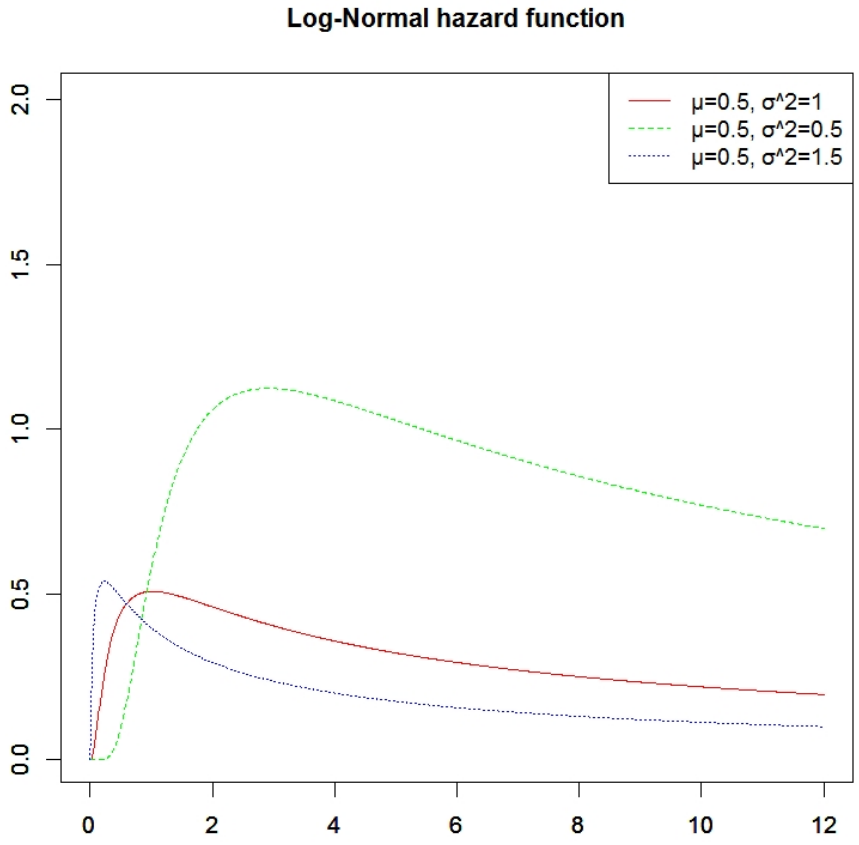
$$\boxed{f_T(t) = f(t) = \frac{1}{\sigma t\sqrt{2\pi}} e^{-\frac{(\ln t - \mu)^2}{2\sigma^2}}}, \text{ με } t > 0 \quad \textbf{(1.25)}$$

Η συνάρτηση επιβίωσης είναι τώρα:

$$\boxed{S(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_t^{\infty} \frac{1}{u} e^{-\frac{(\ln u - \mu)^2}{2\sigma^2}} du} \quad \textbf{(1.26)}$$

και εδώ είναι το μειονέκτημα της κατανομής, αφού η συνάρτηση αυτή εκφράζεται υπό μορφή ολοκληρώματος.

Εδώ, η παράμετρος σχήματος είναι σ και η παράμετρος κλίμακας είναι e^{μ}. Η συνάρτηση κινδύνου είναι αρχικά αύξουσα, φθάνει σε κάποιο μέγιστο και στη συνέχεια γίνεται φθίνουσα (βλ. [6]). Με τη βοήθεια και πάλι της R, σχεδιάσαμε ενδεικτικά τρεις συναρτήσεις κινδύνου:



Σχ. 4: Συνάρτηση κινδύνου της Λογαριθμο-κανονικής κατανομής, για διάφορες τιμές των παραμέτρων μ και σ^2

Για τον υπολογισμό της μέσης τιμής έχουμε ότι επειδή η τυχαία μεταβλητή $Y = \ln T \sim N(\mu, \sigma^2)$, άρα $E[T] = E[e^Y] = \int_{-\infty}^{\infty} e^y \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy$ και εφαρμόζοντας την αντικατάσταση $x = y - \mu$:

$$E[T] = E[e^Y] = \int_{-\infty}^{\infty} e^{x+\mu} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} dx = e^{\mu} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{x-\frac{x^2}{2\sigma^2}} dx =$$

$$= e^{\mu} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2-2\sigma^2 x}{2\sigma^2}} dx = -e^{\mu} e^{\frac{\sigma^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2-2\sigma^2 x+\sigma^4}{2\sigma^2}} dx =$$

$$= e^{\mu+\frac{\sigma^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\sigma^2)^2}{2\sigma^2}} dx = e^{\mu+\frac{\sigma^2}{2}},$$ αφού $$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\sigma^2)^2}{2\sigma^2}} dx = 1$$ ως ολοκλήρωμα της συνάρτησης πυκνότητας πιθανότητας της Κανονικής κατανομής $N(\mu = \sigma^2, \sigma^2)$. Επομένως δείξαμε ότι στη Λογαριθμο-κανονική κατανομή, η μέση τιμή είναι

$$\boxed{E[T] = e^{\mu+\frac{\sigma^2}{2}}} \quad \textbf{(1.27)}$$

Ομοίως μπορεί να αποδειχθεί (βλ. [9]) ότι $E[T^2] = e^{2\mu+2\sigma^2}$, απ' όπου προκύπτει η διασπορά της Λογαριθμο-κανονικής κατανομής

$$\boxed{V[T] = e^{2\mu+\sigma^2} \cdot \left(e^{\sigma^2} - 1\right)} \quad \textbf{(1.28)}$$

1.3.6. Η Γενικευμένη Γάμμα κατανομή (Generalized Gamma distribution): Έχει παράμετρο κλίμακας $\lambda > 0$, δύο παραμέτρους σχήματος $p > 0$, $\kappa > 0$ και συνάρτηση πυκνότητας πιθανότητας

$$\boxed{f_T(t) = f(t) = \frac{1}{\Gamma(\kappa)} \lambda p (\lambda t)^{p\kappa-1} e^{-(\lambda t)^p}} \quad \textbf{(1.29)}$$

Οι κατανομές Weibull, Εκθετική, Γάμμα και Λογαριθμο-κανονική που συναντήσαμε, αποτελούν ειδικές περιπτώσεις της Γενικευμένης Γάμμα κατανομής και ειδικότερα:

- η κατανομή Weibull προκύπτει από τη Γενικευμένη Γάμμα για $\kappa = 1$
- η Εκθετική κατανομή προκύπτει από τη Γενικευμένη Γάμμα για $\kappa = p = 1$

- η κατανομή Γάμμα προκύπτει από τη Γενικευμένη Γάμμα για $p = 1$ και
- η Λογαριθμο-κανονική κατανομή προκύπτει από τη Γενικευμένη Γάμμα

όταν $\kappa \to \infty$.

Η συνάρτηση επιβίωσης είναι

$$\boxed{S(t) = \frac{\Gamma\left(\kappa, (\lambda t)^p\right)}{\Gamma(\kappa)}} \textbf{ (1.30)}$$

όπου $\Gamma\left(\kappa, (\lambda t)^p\right) = \int_{(\lambda t)^p}^{\infty} u^{\kappa-1} e^{-u} du$ **(1.31)** είναι και εδώ η άνω ατελής συνάρτηση Γάμμα (*upper incomplete Gamma function*).

Τέλος, αποδεικνύεται ότι η μέση τιμή και η διασπορά είναι

$$\boxed{E[T] = \frac{\Gamma\left(\kappa + \frac{1}{p}\right)}{\lambda\Gamma(\kappa)}} \textbf{ (1.32)} \text{ και}$$

$$\boxed{V[T] = \frac{1}{\lambda^2}\left\{\frac{\Gamma\left(\kappa + \frac{2}{p}\right)}{\Gamma(\kappa)} - \frac{\Gamma\left(\kappa + \frac{1}{p}\right)}{\Gamma^2(\kappa)}\right\}} \textbf{ (1.33)}$$

1.3.7. Η Λογαριθμο-λογιστική κατανομή (Log-logistic distribution): Με παράμετρο κλίμακας $\lambda > 0$ και παράμετρο σχήματος $p > 0$ (βλ. [10]), έχει συνάρτηση πυκνότητας πιθανότητας

$$\boxed{f_T(t) = f(t) = \frac{\lambda p (\lambda t)^{p-1}}{\left\{1 + (\lambda t)^p\right\}^2}} \textbf{ (1.34)}$$

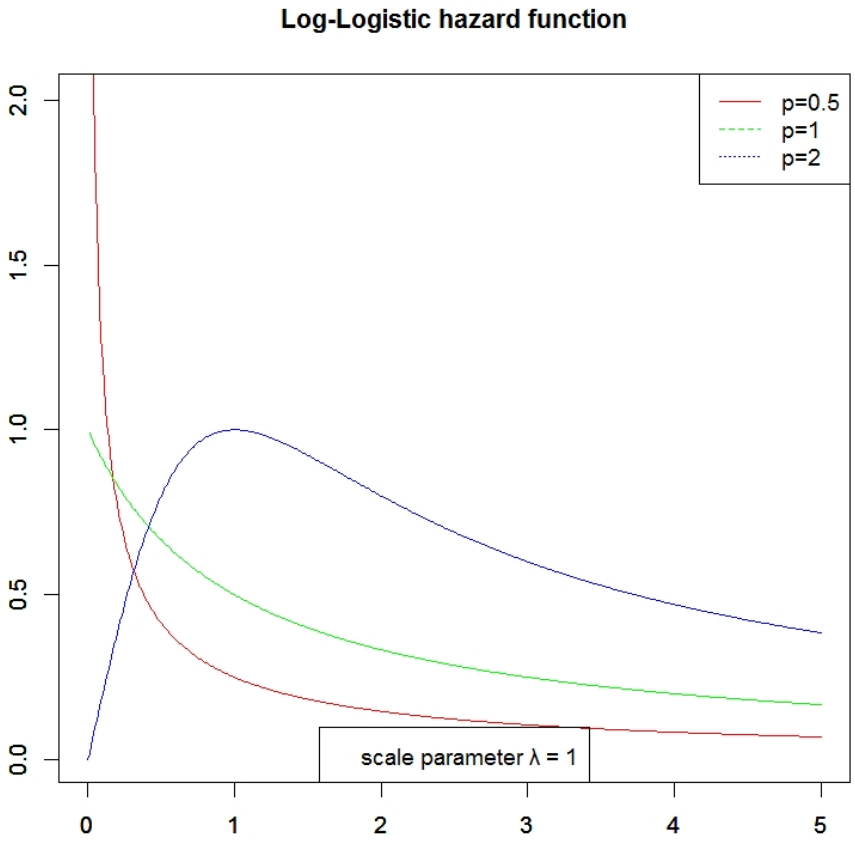
Η συνάρτηση επιβίωσης είναι $S(t) = \int_t^{\infty} f(u) du = \left[-\frac{1}{1 + (\lambda u)^p}\right]_{u=t}^{\infty}$ άρα

$$\boxed{S(t) = \frac{1}{1 + (\lambda t)^p}} \textbf{ (1.35)}$$

και επομένως, η συνάρτηση κινδύνου είναι:

$$\boxed{h(t) = \frac{\lambda p (\lambda t)^{p-1}}{1 + (\lambda t)^p}} \textbf{ (1.36)}$$

Με τη βοήθεια της R σχεδιάσαμε τρεις συναρτήσεις κινδύνου:



Σχ. 5: Συνάρτηση κινδύνου της Λογαριθμο-λογιστικής κατανομής, για παράμετρο κλίμακας $\lambda = 1$ και για διάφορες τιμές της παραμέτρου σχήματος p

Αποδεικνύεται εύκολα ότι (βλ. [6] και το παραπάνω σχ. 5) η συνάρτηση κινδύνου της Λογαριθμο-λογιστικής κατανομής:

- είναι γνησίως φθίνουσα με $\lim_{t\to 0^+} h(t) = +\infty$, όταν $p < 1$
- είναι γνησίως φθίνουσα με $\lim_{t\to 0^+} h(t) = \lambda$, όταν $p = 1$
- παρεμφερής με αυτήν της Λογαριθμο-κανονικής κατανομής, όταν $p > 1$

(βλ. σελ. 12)

1.3.8. Η αντίστροφη Γκαουσιανή κατανομή (Inverse Gaussian distribution): Έχει συνάρτηση πυκνότητας πιθανότητας

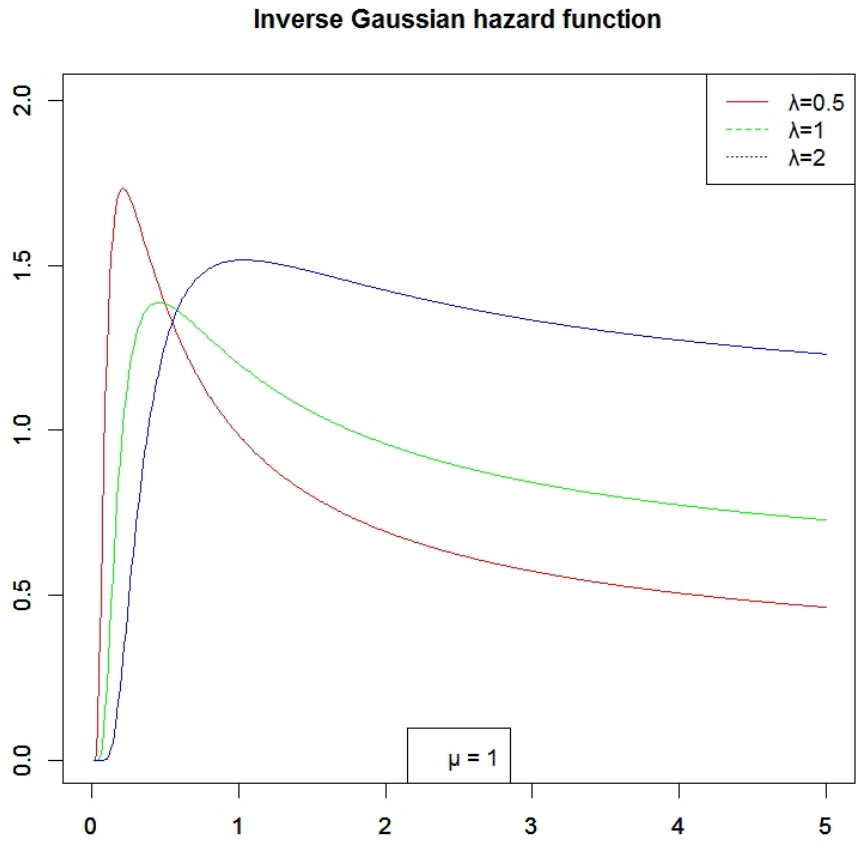
$$f_T(t) = f(t) = \sqrt{\frac{\lambda}{2\pi t^3}} \cdot e^{-\frac{\lambda(t-\mu)^2}{2\mu^2 t}}, \quad t > 0,\ \mu > 0,\ \lambda > 0 \quad \textbf{(1.37)}$$

Για εκτενή αναφορά στη μοντελοποίηση μέσω της Inverse Gaussian κατανομής παραπέμπουμε στο [11]. Εδώ, περιοριζόμαστε στο να αναφέρουμε τη συνάρτηση επιβίωσης: $$S(t)=\Phi\left(\sqrt{\frac{\lambda}{t}}\left(1-\frac{t}{\mu}\right)\right)-e^{\frac{2\lambda}{\mu}}\cdot\Phi\left(-\sqrt{\frac{\lambda}{t}}\left(1+\frac{t}{\mu}\right)\right)$$ **(1.38)**,

τη συνάρτηση κινδύνου: $$h(t)=\frac{\sqrt{\frac{\lambda}{2\pi t^3}}\cdot e^{-\frac{\lambda(t-\mu)^2}{2\mu^2 t}}}{\Phi\left(\sqrt{\frac{\lambda}{t}}\left(1-\frac{t}{\mu}\right)\right)-e^{\frac{2\lambda}{\mu}}\cdot\Phi\left(-\sqrt{\frac{\lambda}{t}}\left(1+\frac{t}{\mu}\right)\right)}$$ **(1.39)**

και τη μέση τιμή: $E[T]=\mu$ **(1.40)** και τη διασπορά: $V[T]=\frac{\mu^3}{\lambda}$ **(1.41)**

Η συνάρτηση κινδύνου της Inverse Gaussian κατανομής προσομοιάζει με αυτήν της Λογαριθμο-κανονικής κατανομής:



Σχ. 5: Συνάρτηση κινδύνου της Inverse Gaussian κατανομής, για παραμέτρο $\mu=1$ και για διάφορες τιμές της παραμέτρου λ

1.4. ΜΗ ΠΑΡΑΜΕΤΡΙΚΑ ΜΟΝΤΕΛΑ ΔΙΑΡΚΕΙΑΣ ΖΩΗΣ

Σε αντίθεση με τα παραμετρικά μοντέλα διάρκειας ζωής, στη μη-παραμετρική περίπτωση υποθέτουμε ότι ο χρόνος ζωής T δεν ακολουθεί γνωστή κατανομή. Εδώ, από τα δεδομένα του δείγματος προσπαθούμε να εκτιμήσουμε τη συνάρτηση επιβίωσης ή/και τη συνάρτηση κινδύνου, να πραγματοποιήσουμε ελέγχους υποθέσεων και να βρούμε διαστήματα εμπιστοσύνης για τις παραμέτρους.

1.4.1. Η εκτιμήτρια Kaplan-Meier για τη μη παραμετρική εκτίμηση της συνάρτησης επιβίωσης: Πήρε το όνομά της από τους Edward Kaplan και Paul Meier που την παρουσίασαν το 1958 (βλ. [13]) και χρησιμοποιείται πολύ συχνά για την εκτίμηση της συνάρτησης επιβίωσης από δεδομένα διάρκειας ζωής που είναι δεξιά λογοκριμένα (*right censored data*).

Υποθέτουμε ότι έχουμε ένα τυχαίο δείγμα μονάδων, μεγέθους n, κάποιες εκ των οποίων διακόπτουν τη λειτουργία τους στις διακεκριμένες χρονικές στιγμές (χρόνοι διακοπής ή αποτυχίας) $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$, όπου $k \le n$. Υποθέτουμε επίσης ότι κατά τη χρονική στιγμή $t_{(j)}$, $j = 1, 2, \ldots, k$ σταματούν να λειτουργούν d_j το πλήθος μονάδες, ενώ αμέσως πριν τη χρονική στιγμή $t_{(j)}$ λειτουργούσαν (και άρα βρίσκονταν σε κίνδυνο) r_j το πλήθος μονάδες. Η εκτιμήτρια Kaplan-Meier της συνάρτησης επιβίωσης ορίζεται (βλ. [14], [15]) ως $\hat{S}(t) = \hat{S}\left(t^{-}\right) \cdot \hat{P}[T > t / T \ge t]$. Στην ουσία, αυτή είναι μία εκτιμήτρια που προκύπτει από τον ορισμό της συνάρτησης επιβίωσης και την εφαρμογή του πολλαπλασιαστικού νόμου των πιθανοτήτων $P(A \cap B) = P(A) \cdot P(B / A)$ (βλ. [3]).

Συγκεκριμένα, επειδή $P\left[T > t_{(j)}\right] = P\left[\bigcap_{k=1}^{j} \left\{T > t_{(k)}\right\}\right]$, έχουμε ότι : $S\left(t_{(j)}\right) = P\left[T > t_{(j)}\right] = P\left[T > t_{(1)}\right] \cdot P\left[T > t_{(2)} / T > t_{(1)}\right] \cdot \ldots \cdot P\left[T > t_{(j)} / T > t_{(j-1)}\right]$.

Μία εκτιμήτρια της πιθανότητας $P\left[T > t_{(1)}\right]$ είναι τώρα η $\hat{P}\left[T > t_{(1)}\right] = 1 - \hat{P}\left[T \le t_{(1)}\right] = 1 - \frac{d_1}{r_1} = \frac{r_1 - d_1}{r_1}$ οπότε $P\left[T > t_{(2)} / T > t_{(j)}\right] = \frac{r_2 - d_2}{r_2}$.

Αντικαθιστώντας στον παραπάνω πολλαπλασιαστικό τύπο, η εκτιμήτρια Kaplan-Meier τελικά είναι:

$$\boxed{\hat{S}(t) = \prod_{j: t \ge t_{(j)}} \frac{r_j - d_j}{r_j}} \textbf{ (1.42)} \text{ για } t \ge t_{(1)}$$

ενώ προφανώς είναι $\hat{S}(t) = 1$ για $t < t_{(1)}$.

Η διασπορά της εκτιμήτριας Kaplan-Meier είναι ίση με $V\left[\hat{S}(t)\right] = V\left[\prod_{j: t \ge t_{(j)}} \frac{r_j - d_j}{r_j}\right] = V\left[\prod_{j: t \ge t_{(j)}} p_{(j)}\right]$, όπου $p_{(j)} = \frac{r_j - d_j}{r_j}$, $j = 1,\ 2,\ \ldots,\ k$.

Μας συμφέρει όμως να χρησιμοποιήσουμε τη διασπορά αθροίσματος ανεξάρτητων τυχαίων μεταβλητών, η οποία ισούται με το άθροισμα των διασπορών τους. Έτσι, υποθέτοντας ότι οι μονάδες διακόπτουν τη λειτουργία τους ανεξάρτητα, παίρνουμε τη διασπορά του λογαρίθμου της εκτιμήτριας Kaplan – Meier: $V\left[\ln \hat{S}(t)\right] = V\left[\sum_{j: t \ge t_{(j)}} p_{(j)}\right]$. Στη συνέχεια, με τη βοήθεια μίας τεχνικής που ονομάζεται "μέθοδος Δέλτα", με την οποία η διασπορά μίας εκτιμήτριας υπολογίζεται από την προσέγγιση μέσω αναπτύγματος Taylor γύρω από τη μέση τιμή, μπορεί να υπολογισθεί ο τελικός τύπος της διασποράς της εκτιμήτριας Kaplan-Meier:

$$\boxed{V\left[\hat{S}(t)\right] = \hat{S}^2(t) \cdot \sum_{j: t \ge t_{(j)}} \frac{d_j}{r_j (r_j - d_j)}} \textbf{ (1.43)} \text{ (τύπος του Greenwood)}$$

(για λεπτομέρειες σχετικά με τον υπολογισμό της διασποράς μέσω της μεθόδου δέλτα και τον τύπο Greenwood, παραπέμπουμε στα [15]) και [16]).

ΠΑΡΑΔΕΙΓΜΑ 1: Με τη βοήθεια της R, δημιουργήσαμε δείγμα από τις ακόλουθες 20 παρατηρήσεις:

0,147759	1,010824*	0,317879*	4,841124*	0,009327
0,223461	3,999043	0,21502*	0,473134	0,009027
3,396098	0,01316*	9,33835*	9,178013*	0,477899*
0,010846*	9,738259*	1,764082*	0,272073	4,766219

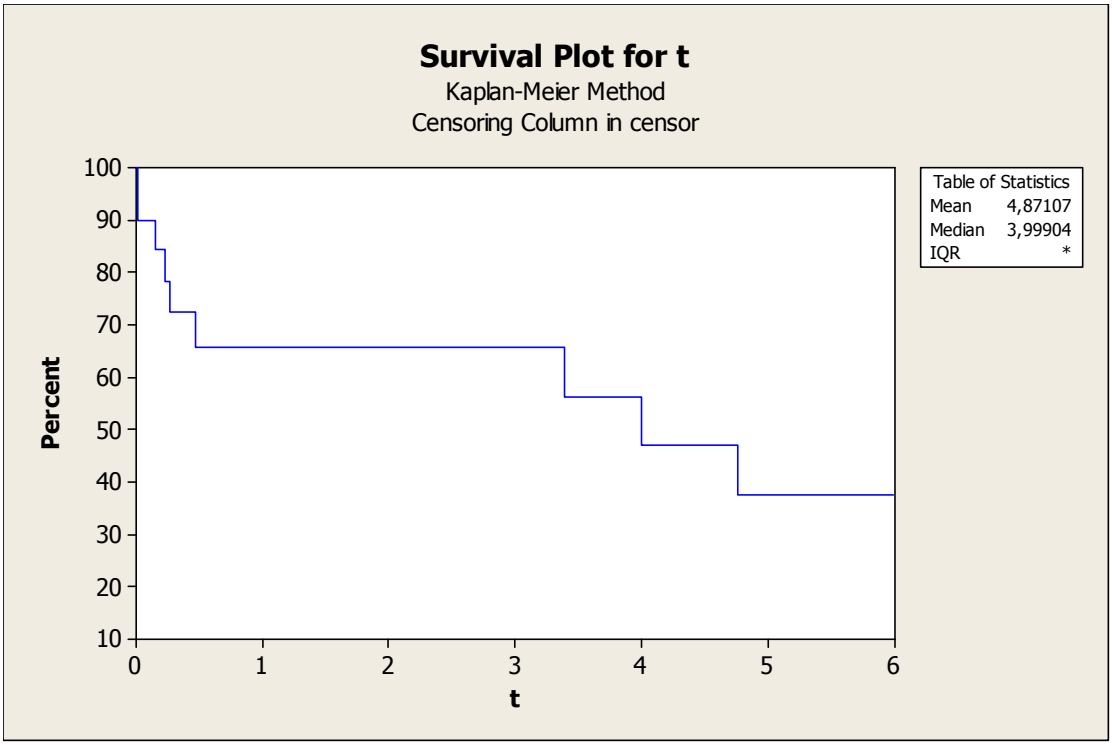
όπου με (*) σημειώνονται οι λογοκριμένες παρατηρήσεις. Για να κατασκευάσουμε την εκτιμήτρια Kaplan-Meier της συνάρτησης επιβίωσης, ταξινομούμε τους χρόνους αποτυχίας σε αύξουσα σειρά και καταρτίζουμε πίνακα υπολογισμών.

Time	Censor id	rj	Dj	(rj-dj)/rj	S_hat
0.009027	1	20	1	0.95	0.95
0.009327	1	19	1	0.947368	0.9
0.010846	0	18	0	1	0.9
0.01316	0	17	0	1	0.9
0.147759	1	16	1	0.9375	0.84375
0.21502	0	15	0	1	0.84375
0.223461	1	14	1	0.928571	0.783482
0.272073	1	13	1	0.923077	0.723214
0.317879	0	12	0	1	0.723214
0.473134	1	11	1	0.909091	0.657468
0.477899	0	10	0	1	0.657468
1.010824	0	9	0	1	0.657468
1.764082	0	8	0	1	0.657468
3.396098	1	7	1	0.857143	0.563544
3.999043	1	6	1	0.833333	0.46962
4.766219	1	5	1	0.8	0.375696
4.841124	0	4	0	1	0.375696
9.178013	0	3	0	1	0.375696
9.338354	0	2	0	1	0.375696
9.738259	0	1	0	1	0.375696

Επομένως η εκτιμήτρια Kaplan-Meier για τη συνάρτηση επιβίωσης του παραδείγματος, είναι:

$$\hat{S}(t)=\begin{cases}1 & \text{αν } t<0.00903\\ 0.95 & \text{αν } 0.00927\le t<0.009327\\ 0.9 & \text{αν } 0.009327\le t<0.147759\\ 0.84375 & \text{αν } 0.147759\le t<0.223461\\ 0.783482 & \text{αν } 0.223461\le t<0.272073\\ 0.723214 & \text{αν } 0.272073\le t<0.473134\\ 0.657468 & \text{αν } 0.473134\le t<3.396098\\ 0.563544 & \text{αν } 3.396098\le t<3.999043\\ 0.46962 & \text{αν } 3.999043\le t<4.766219\\ 0.375696 & \text{αν } t\ge 4.766219\end{cases}$$

Ένα διάγραμμα της παραπάνω κλιμακωτής (*step-function*) και φθίνουσας εκτιμηθείσας συνάρτησης επιβίωσης μπορούμε να κατασκευάσουμε με τη βοήθεια του στατιστικού προγράμματος Minitab:



Σχ. 6: Διάγραμμα της εκτιμηθείσας μέσω Kaplan-Meier συνάρτησης επιβίωσης του προηγούμενου παραδείγματος

1.4.2. Μη παραμετρική εκτίμηση της σωρευτικής συνάρτησης κινδύνου – Η εκτιμήτρια Nelson-Aalen: Με βάση τη γνωστή σχέση $H(t)=-\ln S(t)$ που συνδέει τη σωρευτική συνάρτηση κινδύνου με τη συνάρτηση επιβίωσης και χρησιμοποιώντας ως $S(t)$ την Kaplan-Meier εκτίμησή της, μπορούμε να έχουμε μία εκτίμηση για την $H(t)$. Έτσι, θα είναι

$$\hat{H}(t)=-\ln\hat{S}(t)=-\ln\prod_{j:t\geq t_{(j)}}\frac{r_j-d_j}{r_j}=-\sum_{j:t\geq t_{(j)}}\ln\left(1-\frac{d_j}{r_j}\right) \quad \textbf{(1.44).}$$

Από τη Μαθηματική Ανάλυση, γνωρίζουμε ότι $\ln x\leq x-1$ για κάθε $x>0$ και το ίσον ισχύει για $x=1$. Θέτοντας όπου x το $1-x$, παίρνουμε $\ln(1-x)\leq -x$ με την ισότητα τώρα να ισχύει για $x=0$. Δηλ. μπορούμε να γράψουμε

$\ln(1-x) \approx -x$ για πολύ μικρά x. Έτσι, υποθέτοντας ότι $\frac{d_j}{r_j}$ πολύ μικρό (πράγμα που είναι λογικό να συμβαίνει στους πρώτους χρόνους αποτυχίας), η σχέση **(1.41)** μπορεί να γραφεί

$$\boxed{\hat{H}(t) \approx \sum_{j:\, t \geq t_{(j)}} \frac{d_j}{r_j}} \quad \textbf{(1.45)}$$

Η **(1.45)** είναι η εκτιμήτρια Nelson-Aalen της σωρευτικής συνάρτησης κινδύνου, είναι επίσης κλιμακωτή συνάρτηση και αποδεικνύεται ότι έχει διασπορά:

$$\boxed{V\left[\hat{H}\right] = \sum_{j:\, t \geq t_{(j)}} \frac{d_j}{r_j^2}} \quad \textbf{(1.46)}$$

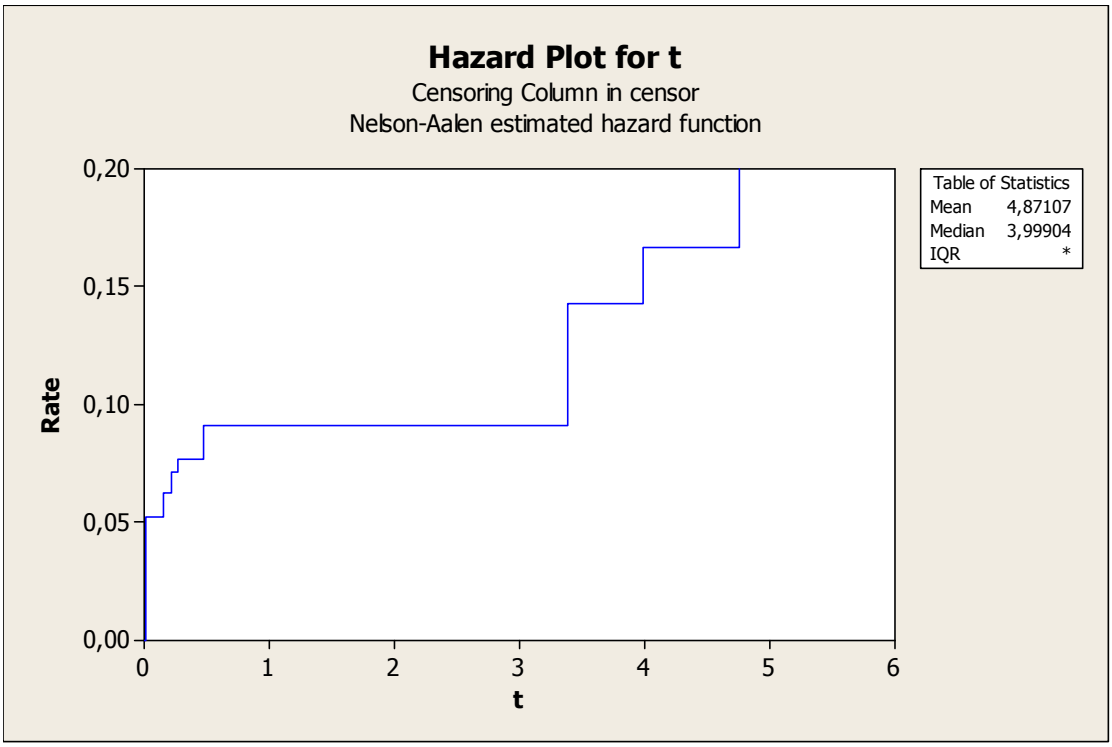
Η εκτιμήτρια Nelson-Aalen, που εισήχθη αρχικά από τον Nelson το 1972 και ξαναπροτάθηκε από τον Odd Aalen το 1978, χρησιμοποιείται ευρέως στη Βιοστατιστική ως μη παραμετρική εκτιμήτρια της σωρευτικής συνάρτησης κινδύνου με βάση δεξιά λογοκριμένα (*right censored*) δεδομένα (βλ. [17]).

ΠΑΡΑΔΕΙΓΜΑ 2: Συνεχίζοντας το προηγούμενο παράδειγμα 1, μπορούμε στον πίνακα υπολογισμών να συμπληρώσουμε μία στήλη, με τις τιμές της εκτιμήτριας Nelson-Aalen:

Time	Censor id	rj	dj	(rj-dj)/rj	S_hat	H_hat
0.009027	1	20	1	0.95	0.95	0.05
0.009327	1	19	1	0.947368	0.9	0.052632
0.010846	0	18	0	1	0.9	0.052632
0.01316	0	17	0	1	0.9	0.052632
0.147759	1	16	1	0.9375	0.84375	0.0625
0.21502	0	15	0	1	0.84375	0.0625
0.223461	1	14	1	0.928571	0.783482	0.071429
0.272073	1	13	1	0.923077	0.723214	0.076923
0.317879	0	12	0	1	0.723214	0.076923
0.473134	1	11	1	0.909091	0.657468	0.090909
0.477899	0	10	0	1	0.657468	0.090909
1.010824	0	9	0	1	0.657468	0.090909
1.764082	0	8	0	1	0.657468	0.090909
3.396098	1	7	1	0.857143	0.563544	0.142857
3.999043	1	6	1	0.833333	0.46962	0.166667
4.766219	1	5	1	0.8	0.375696	0.2
4.841124	0	4	0	1	0.375696	0.2

9.178013	0	3	0	1	0.375696	0.2
9.338354	0	2	0	1	0.375696	0.2
9.738259	0	1	0	1	0.375696	0.2

και χρησιμοποιώντας το πρόγραμμα Minitab έχουμε το διάγραμμά της:



Σχ. 7: Διάγραμμα της εκτιμηθείσας μέσω Nelson-Aalen συνάρτησης κινδύνου του προηγούμενου παραδείγματος

ΚΕΦΑΛΑΙΟ 2

ΤΟ ΜΟΝΤΕΛΟ ΑΝΑΛΟΓΙΚΩΝ ΚΙΝΔΥΝΩΝ ΤΟΥ COX

2.1. ΜΟΝΤΕΛΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΣΤΗΝ ΑΝΑΛΥΣΗ ΕΠΙΒΙΩΣΗΣ

2.1.1. Μοντέλο Γραμμικής Παλινδρόμησης: Ένα σύνηθες θέμα προς αντιμετώπιση στη Στατιστική είναι το κατά πόσο διάφορες μεταβλητές επηρεάζουν μεταβλητές που μελετούμε. Έτσι, στην Ανάλυση Επιβίωσης, μας ενδιαφέρει το αν κάποιες μεταβλητές επηρεάζουν τη διάρκεια ζωής (π.χ. αν με τον όρο "διάρκεια ζωής" εννοούμε τη διάρκεια της ζωής του ανθρώπου, τότε μπορεί να ενδιαφερόμαστε για μεταβλητές όπως η ηλικία, οι τιμές λιπιδίων στο αίμα, το κάπνισμα κ.λ.π.). Ο κλάδος της Στατιστικής που μελετά γενικά την εξάρτηση μεταξύ μεταβλητών, είναι η **Ανάλυση Παλινδρόμησης** (*Regression Analysis*).

Στην Ανάλυση Παλινδρόμησης, καθορίζουμε δύο είδη μεταβλητών: τις **ανεξάρτητες** ή **επεξηγηματικές μεταβλητές** (*predictor variables*) ή **συμμεταβλητές** (*covariates*) και τις **εξαρτημένες μεταβλητές** ή **μεταβλητές απόκρισης** (*response variables*). Η ανάλυσή μας, συνίσταται στο να εξετάσουμε αν αλλαγές στις επεξηγηματικές μεταβλητές επηρεάζουν τις τιμές των μεταβλητών απόκρισης.

Κυρίαρχη θέση στην Ανάλυση Παλινδρόμησης έχει η **Γραμμική Παλινδρόμηση** (*Linear Regression*), στην οποία η σχέση των μεταβλητών είναι της μορφής:

(Μεταβλητή απόκρισης) = (Γραμμική συνάρτηση των επεξηγηματικών μεταβλητών) + (τυχαίο σφάλμα)

(βλ. [18]). Έτσι, αν συμβολίσουμε με y τη μεταβλητή απόκρισης και με $z_1, z_2, \ldots, z_k$ τις επεξηγηματικές μεταβλητές, τότε θα ισχύει:

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_k z_k + \varepsilon \quad \textbf{(2.1)}$$

όπου $\beta_0, \beta_1, \ldots, \beta_k$ παράμετροι (συντελεστές) προς προσδιορισμό και ε το τυχαίο σφάλμα, το οποίο υποθέτουμε ότι ακολουθεί Κανονική κατανομή $N\left(0, \sigma^2\right)$. Σημειώνουμε εδώ ότι το μοντέλο παλινδρόμησης ονομάζεται γραμμικό, επειδή είναι γραμμικό ως προς τις παραμέτρους β_i, $i = 1, 2, \ldots, k$ (και όχι ως προς τις συμμεταβλητές z_i, $i = 1, 2, \ldots, k$).

Η εξίσωση **(2.1)** μπορεί τώρα να γραφεί υπό μορφή εξίσωσης πινάκων, ως εξής: αν έχουμε ένα δείγμα μεγέθους n από παρατηρήσεις και για κάθε παρατήρηση $i \in \{1, 2, \ldots, n\}$ ονομάσουμε $\boldsymbol{z}_i = [1, z_{i1}, z_{i2}, \ldots, z_{ik}]^{\mathrm{T}}$ το διάνυσμα των συμμεταβλητών, $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \ldots, \beta_k]^{\mathrm{T}}$ το διάνυσμα των συντελεστών, y_i την τιμή της μεταβλητής απόκρισης και ε_i την τιμή του τυχαίου σφάλματος, τότε για κάθε $i = 1, 2, \ldots, n$, η εξίσωση **(2.1)** γράφεται $y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \ldots + \beta_k z_{ik} + \varepsilon_i$, ή ισοδύναμα $y_i = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z_i} + \varepsilon_i$.

Περαιτέρω, αν $\boldsymbol{z} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_{11} & z_{21} & \ldots & z_{n1} \\ z_{12} & z_{22} & \ldots & z_{n2} \\ \vdots & \vdots & \ldots & \vdots \\ z_{1k} & z_{2k} & \cdots & z_{nk} \end{bmatrix}$ είναι ο $k \times n$-πίνακας του οποίου στήλες είναι οι τιμές των συμμεταβλητών για κάθε παρατήρηση $i \in \{1, 2, \ldots, n\}$, $\mathbf{y} = [y_1, y_2, \ldots, y_n]^{\mathrm{T}}$ είναι το διάνυσμα των τιμών της μεταβλητής απόκρισης και $\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n]^{\mathrm{T}}$ είναι το διάνυσμα των σφαλμάτων, τότε το μοντέλο παλινδρόμησης που περιγράψαμε με την εξίσωση **(2.1)**, παίρνει τη μορφή εξίσωσης πινάκων:

$$\boxed{\mathbf{y} = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z} + \boldsymbol{\varepsilon}} \quad \textbf{(2.2)}$$

Θεωρώντας τα τυχαία σφάλματα ε_i ως ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν την Κανονική κατανομή $N\left(0, \sigma^2\right)$, προκύπτει ότι η διανυσματική τυχαία μεταβλητή $\boldsymbol{\varepsilon}$ ακολουθεί την n-μεταβλητή Κανονική κατανομή με μέση τιμή το μηδενικό διάνυσμα $\mathbf{0}$ και πίνακα διακύμανσης-συνδιακύμανσης τον $\sigma^2 \mathbf{I_n}$, δηλ. $\boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0}, \sigma^2 \mathbf{I_n}\right)$ και άρα η τυχαία μεταβλητή $\mathbf{y}$ ακολουθεί επίσης την πολυμεταβλητή Κανονική κατανομή με μέση τιμή

$\boldsymbol{\mu} = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}$ και πίνακα διακύμανσης-συνδιακύμανσης τον $\sigma^2\mathbf{I_n}$, δηλ. $\mathbf{y} \sim N_n\left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z},\ \sigma^2\mathbf{I_n}\right)$ (για λεπτομέρειες ανατρέξτε στο [19], σελ. 57-61).

2.1.2. Γενικευμένα Γραμμικά Μοντέλα Παλινδρόμησης: Το μοντέλο Γραμμικής Παλινδρόμησης που περιγράψαμε πριν, οφείλει από την κατασκευή του να έχει τα ακόλουθα χαρακτηριστικά (βλ. [20]):

- Τα τυχαία σφάλματα $\varepsilon_i,\ i = 1, 2, \ldots, n$, είναι ανεξάρτητες τυχαίες μεταβλητές και ακολουθούν Κανονική κατανομή με μέση τιμή 0 και σταθερή διασπορά σ^2, απ' όπου προκύπτει ότι οι τιμές y_i της μεταβλητής απόκρισης ακολουθούν Κανονικές κατανομές με μέση τιμή $E[y_i] = \mu_i$ και σταθερή διασπορά σ^2, δηλ. $y_i \overset{\text{iid}}{\sim} N\left(\mu_i, \sigma^2\right)$ (κανονικότητα του στοχαστικού μέρους).
- Οι συμμεταβλητές $z_1, z_2, \ldots, z_k$ συνδυάζονται γραμμικά με τους συντελεστές $\beta_1, \beta_2, \ldots, \beta_k$ προκειμένου να δημιουργηθεί η **γραμμική προβλέπουσα** (*linear predictor*) $\eta_i = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_i,\ i = 1, 2, \ldots, n$.
- Ο στοχαστικός παράγοντας $E[y_i] = \mu_i$ και η γραμμική προβλέπουσα $\eta_i = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_i$ συνδέονται μέσω μίας συνάρτησης g, που ονομάζεται **συνάρτηση σύνδεσης** (*link function*), ώστε $g(\mu_i) = \eta_i,\ i = 1, 2, \ldots, n$. Στην περίπτωση της Γραμμικής Παλινδρόμησης είναι προφανές ότι η συνάρτηση g είναι η ταυτοτική, δηλ. $g(\mu_i) = \mu_i = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_i,\ i = 1, 2, \ldots, n$.

Επέκταση των παραπάνω, αποτελούν τα **Γενικευμένα Γραμμικά Μοντέλα** (*Generalized Linear Models – GLM*), στα οποία αφενός το στοχαστικό μέρος μπορεί να ακολουθεί και άλλες κατανομές πέραν της Κανονικής, αφετέρου η συνάρτηση σύνδεσης μπορεί να μην είναι η ταυτοτική.

Έτσι, τα Γενικευμένα Γραμμικά Μοντέλα ικανοποιούν τα ακόλουθα:

- Η μεταβλητή απόκρισης y ακολουθεί κατανομή που ανήκει στην **Εκθετική Οικογένεια Κατανομών** (*Exponential Family*) και ως εκ τούτου έχει συνάρτηση πυκνότητας πιθανότητας (για συνεχείς τυχαίες μεταβλητές) ή συνάρτηση μάζας πιθανότητας (για διακριτές τυχαίες μεταβλητές) της μορφής

$f(y;\theta,\varphi)=\exp\left\{\frac{y\theta-b(\theta)}{a(\varphi)}+c(y,\varphi)\right\}$ με τις συναρτήσεις $a(\varphi)$, $b(\theta)$ και $c(y,\varphi)$ να είναι γνωστές και το στήριγμα $S=\{y\in\Re / f(y)>0\}$ να είναι ανεξάρτητο των παραμέτρων θ και φ.

- Η συνάρτηση σύνδεσης g είναι αντιστρέψιμη (1-1), άρα η σχέση $g(\mu_i)=\eta_i=\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_i$ του ορισμού, μπορεί να λυθεί αντίστροφα ως $\mu_i=g^{-1}\left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_i\right)$ δίνοντας τη μέση τιμή της μεταβλητής απόκρισης

Συνήθεις επιλογές για τη συνάρτηση σύνδεσης είναι οι:

- $g(\mu_i)=\mu_i$ (ταυτοτική συνάρτηση) όταν οι y_i ακολουθούν την Κανονική κατανομή
- $g(\mu_i)=\ln\mu_i$ όταν οι y_i ακολουθούν την κατανομή Poisson
- $g(\mu_i)=\ln\frac{\mu_i}{1-\mu_i}$ όταν οι y_i ακολουθούν την κατανομή Bernoulli.

Για περισσότερες πληροφορίες πάνω στα Γενικευμένα Γραμμικά Μοντέλα, παραπέμπουμε στα [19] (σελ. 335-417) και [20].

Στο σημείο αυτό αναφέρουμε ότι στην παρούσα εργασία εξετάζουμε μοντέλα διάρκειας ζωής στα οποία οι συμμεταβλητές δεν είναι εξαρτώμενες από τον χρόνο.

2.1.3. Το Μοντέλο Επιταχυνόμενης Διακοπής (Accelerated Failure Time model - AFT) για δεδομένα διάρκειας ζωής: Με βάση δεδομένα διάρκειας ζωής, θέλουμε να δημιουργήσουμε ένα μοντέλο παλινδρόμησης με μεταβλητή απόκρισης y τη μεταβλητή T του χρόνου.

Έστω λοιπόν ένα δείγμα δεδομένων διάρκειας ζωής μεγέθους n. Θεωρώντας για κάθε $i\in\{1,2,\ldots,n\}$ ως διάνυσμα συμμεταβλητών το $\boldsymbol{z}_i=[z_{i1},z_{i2},\ldots,z_{ik}]^{\mathrm{T}}$ και προσαρμόζοντας ένα μοντέλο Γραμμικής Παλινδρόμησης στα δεδομένα, η μεταβλητή απόκρισης T θα ικανοποιεί τις εξισώσεις $T_i=\beta_1 z_{i1}+\beta_2 z_{i2}+\ldots+\beta_k z_{ik}+\varepsilon_i$, $i=1,2,\ldots,n$. Έτσι όμως,

οδηγούμαστε σε εσφαλμένο μοντέλο, αφού το δεύτερο μέλος της παραπάνω σχέσης παίρνει τιμές σε όλο το $\Re$, ενώ η μεταβλητή T παίρνει μη αρνητικές τιμές [21].

Το πρόβλημα λύνεται, αν προσαρμόσουμε στα δεδομένα μας ένα Γραμμικό Μοντέλο με μεταβλητή απόκρισης το λογάριθμο $\ln T_i$:. Έτσι, θα έχουμε το μοντέλο (βλ. [22])

$$\boxed{\ln T_i = \beta_1 z_{i1} + \beta_2 z_{i2} + \ldots + \beta_k z_{ik} + \varepsilon_i}, \quad i = 1, 2, \ldots, n \quad \textbf{(2.3)}$$ ή

$$\boxed{\ln T_i = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i + \varepsilon_i}, \quad i = 1, 2, \ldots, n \quad \textbf{(2.4)}$$

όπου $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_k]^{\mathrm{T}}$ το διάνυσμα των παραμέτρων.

Το μοντέλο που περιγράφεται από την εξίσωση **(2.3)** ή ισοδύναμα την **(2.4)** είναι το **Μοντέλο Επιταχυνόμενης Διακοπής** (*Accelerated Failure Time Model*). Ονομάζεται έτσι γιατί αύξηση κατά 1 μονάδα μίας συμμεταβλητής z_{ij}, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, k$ οδηγεί σε αύξηση (αν $\beta_j > 0$) ή μείωση (αν $\beta_j < 0$) του λογαρίθμου $\ln T_i$, άρα σε e^{β_j} φορές επιτάχυνση (αν $\beta_j > 0$) ή επιβράδυνση (αν $\beta_j < 0$) του χρόνου ζωής T_i της μονάδας i.

2.1.4. Το μοντέλο αναλογικών κινδύνων (Proportional Hazards Model - PH) για δεδομένα διάρκειας ζωής: Θεωρούμε όπως και πριν δείγμα μεγέθους n και για κάθε μονάδα $i \in \{1, 2, \ldots, n\}$ το διάνυσμα $\boldsymbol{z}_i = [z_{i1}, z_{i2}, \ldots, z_{ik}]^{\mathrm{T}}$ των συμμεταβλητών. Το **μοντέλο αναλογικών κινδύνων** (*Proportional Hazards Model*) ορίζεται (βλ. [3] σελ. 94) από την εξίσωση

$$h(t \mid \boldsymbol{z}_i) = h_0(t) \cdot \varphi(\boldsymbol{z}_i), \quad i = 1, 2, \ldots, n \quad \textbf{(2.5)}$$

όπου $h(t \mid \boldsymbol{z}_i)$ είναι η συνάρτηση κινδύνου μίας μονάδας, $h_0(t)$ είναι η λεγόμενη **βασική συνάρτηση κινδύνου** (*baseline hazard*) και φ μία θετική συνάρτηση.

Το μοντέλο ονομάζεται έτσι, γιατί για δύο διανύσματα $\mathbf{z_i}$ και $\mathbf{z_j}$ συμμεταβλητών, ισχύει $h(t\,|\,\mathbf{z}_i) \propto h(t\,|\,\mathbf{z}_j)$, αφού ο λόγος των τιμών της συνάρτησης κινδύνου είναι $\dfrac{h(t\,|\,\mathbf{z}_i)}{h(t\,|\,\mathbf{z}_j)} = \dfrac{\varphi(\mathbf{z}_i)}{\varphi(\mathbf{z}_j)}$ **(2.6)** (ανεξάρτητος του χρόνου).

Στο μοντέλο αναλογικών κινδύνων, η βασική συνάρτηση κινδύνου (*baseline hazard*) $h_0(t)$ εκφράζει τη συνάρτηση κινδύνου μίας μονάδας όταν όλοι οι συντελεστές των συμμεταβλητών που συμμετέχουν στο μοντέλο είναι ίσοι με 0 (δηλ. το μοντέλο δεν εξαρτάται από τις συμμεταβλητές) (βλ. [23] σελ. 14).

2.2. ΤΟ ΜΟΝΤΕΛΟ ΑΝΑΛΟΓΙΚΩΝ ΚΙΝΔΥΝΩΝ ΤΟΥ COX

2.2.1. Γενικά περί του μοντέλου αναλογικών κινδύνων του Cox: Πρόκειται για το σπουδαιότερο μοντέλο αναλογικών κινδύνων. Παρουσιάστηκε (βλ. [24]) το 1972 από τον Sir David Cox στην εργασία του "Regression Models and Life Tables" (Journal of the Royal Statistical Society, Series B (Methodological), Vol. 34, No. 2 (1972), pp. 187-220).

Το **μοντέλο αναλογικών κινδύνων του Cox** ορίζεται από την εξίσωση

$$\boxed{h(t\,|\,\mathbf{z}_i) = h_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}}, \quad i = 1, 2, \ldots, n \quad \textbf{(2.7)}$$

δηλ. αποτελεί ειδική περίπτωση του μοντέλου αναλογικών κινδύνων που δόθηκε στην εξίσωση **(2.5)** για $\varphi(\mathbf{z_i}) = e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}$. Ειδικότερα, ο λόγος των τιμών της συνάρτησης κινδύνου που εκφράσθηκε με την εξίσωση **(2.6)** είναι εδώ ίσος με

$$\boxed{\frac{h(t\,|\,\mathbf{z}_i)}{h(t\,|\,\mathbf{z}_j)} = e^{\boldsymbol{\beta}^{\mathrm{T}}(\mathbf{z}_i - \mathbf{z}_j)}} \quad \textbf{(2.8)}$$

Παρατηρούμε επίσης ότι αν λογαριθμίσουμε την **(2.7)**, έχουμε

$$\ln h(t\,|\,\mathbf{z}_i) = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i + \ln h_0(t) \quad \textbf{(2.9)}$$

και η εξίσωση **(2.9)** περιγράφει ένα Γενικευμένο Γραμμικό Μοντέλο Παλινδρόμησης όπως αναπτύχθηκε στην παράγραφο **(2.1.2)**, με $\mu_i = h(t \mid \mathbf{z}_i)$ και συνάρτηση σύνδεσης $g(\mu_i) = \ln \mu_i = \ln h(t \mid \mathbf{z}_i)$.

Η σωρευτική συνάρτηση κινδύνου για το μοντέλο του Cox είναι $H(t \mid \mathbf{z}_i) = \int_0^t h(u \mid \mathbf{z}_i) du = \int_0^t h_0(u) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i} du = e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i} \cdot \int_0^t h_0(u) du$ και άρα

$$\boxed{H(t \mid \mathbf{z}_i) = e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i} \cdot H_0(t)}, \quad i = 1, 2, \ldots, n \quad \textbf{(2.10)}$$

όπου

$$\boxed{H_0(t) = \int_0^t h_0(u) du} \quad \textbf{(2.11)}$$

είναι η βασική σωρευτική συνάρτηση κινδύνου.

Η συνάρτηση επιβίωσης είναι $S(t \mid \mathbf{z}_i) = e^{-H(t \mid \mathbf{z}_i)}$ άρα

$$\boxed{S(t \mid \mathbf{z}_i) = e^{-e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i} \cdot H_0(t)}} \quad \textbf{(2.12)}$$

και ισοδύναμα γράφεται $S(t \mid \mathbf{z_i}) = \left\{ e^{-H_0(t)} \right\}^{e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i}}$ άρα:

$$\boxed{S(t \mid \mathbf{z_i}) = \left\{ S_0(t) \right\}^{e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i}}} \quad \textbf{(2.13)}$$

όπου $S_0(t) = e^{-H_0(t \mid \mathbf{z}_i)}$ **(2.14)** η βασική συνάρτηση επιβίωσης.

Βασικό χαρακτηριστικό στο μοντέλο του Cox είναι ότι <u>η βασική συνάρτηση κινδύνου</u> $h_0(t)$ (άρα και η βασική συνάρτηση επιβίωσης $S_0(t)$) <u>δεν</u> (είναι απαραίτητο να) <u>προσδιορίζεται, αλλά θεωρείται ως άγνωστη παράμετρος απείρου διαστάσεως, που πρέπει και αυτή να εκτιμηθεί</u>. Έτσι, το μοντέλο του Cox θεωρείται **ημιπαραμετρικό** (*semiparametric*), με την έννοια του ότι εκτιμώνται οι παράμετροι $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_k]^{\mathrm{T}}$ (συντελεστές των συμμεταβλητών) με παραμετρικές μεθόδους και η άγνωστη συνάρτηση κινδύνου $h_0(t)$ με μη παραμετρικές μεθόδους.

2.2.2. Εκτίμηση των παραμέτρων στο μοντέλο αναλογικών κινδύνων του Cox, με τη μέθοδο της μερικής πιθανοφάνειας (partial likelihood): Στην εργασία του 1972 ([24]), ο Sir David Cox προσάρμοσε το μοντέλο του,

μεγιστοποιώντας την **μερική πιθανοφάνεια** (partial likelihood) ως εξής (βλ. [24] και [6]): Έστω δείγμα μονάδων $\{1, 2, \ldots, n\}$ και έστω $t_{(1)} < t_{(2)} < \ldots < t_{(m)}$ οι διατεταγμένοι διακεκριμένοι χρόνοι αποτυχίας (χρόνοι θανάτου), όπου $m \le n$. Έστω επίσης $\Re_{(j)}$ το σύνολο των μονάδων που βρίσκονται σε κίνδυνο αμέσως πριν τη χρονική στιγμή $t_{(j)}$, $j = 1, 2, \ldots, m$ (προφανώς $\Re_{(1)} = \{1, 2, \ldots, n\}$) και έστω ότι σε κάθε χρονική στιγμή $t_{(j)}$, $j = 1, 2, \ldots, m$ έχουμε μόνο μία αποτυχία (θάνατο), δηλ. ότι $d_j = 1$ κατά τον συμβολισμό της παραγράφου **(1.4.1)**.

Η πιθανότητα μία συγκεκριμένη μονάδα του συνόλου $\Re_{(j)}$ με διάνυσμα συμμεταβλητών $\mathbf{z}_j$ να διακόψει τη λειτουργία της τη χρονική στιγμή $t_{(j)}$ με δεδομένο το σύνολο $\Re_{(j)}$ των υποψήφιων προς διακοπή μονάδων, είναι ίση με $\dfrac{h\left(t_{(j)} \mid \mathbf{z}_j\right)dt}{\sum_{i \in \Re_{(j)}} h\left(t_{(j)} \mid \mathbf{z}_i\right)dt} = \dfrac{h_0\left(t_{(j)}\right) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_j}}{\sum_{i \in \Re_{(j)}} h_0\left(t_{(j)}\right) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i}} = \dfrac{e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_j}}{\sum_{i \in \Re_{(j)}} e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i}}$, δηλ. η πιθανότητα αυτή δεν εξαρτάται από τη βασική συνάρτηση κινδύνου $h_0(t)$.

Στη συνέχεια, ο Cox πολλαπλασίασε τις πιθανότητες για όλους τους χρόνους αποτυχίας και θεώρησε το γινόμενο

$$L = L(\boldsymbol{\beta}) = \prod_{j=1}^{m} \frac{e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_j}}{\sum_{i \in \Re_{(j)}} e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i}} \quad \textbf{(2.15)}$$

ως μία συνηθισμένη πιθανοφάνεια, την οποία ονόμασε "υπό συνθήκη πιθανοφάνεια" (*conditional likelihood*) (βλ. [24] σελ. 190-191) επειδή είναι γινόμενο υπό συνθήκη πιθανοτήτων, ενώ το 1975 τη μετονόμασε σε **μερική πιθανοφάνεια** (*partial likelihood*) (για την ιστορία βλ. [7] σελ. 9).

Οι εκτιμήσεις των παραμέτρων $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_k]^{\mathrm{T}}$ έγιναν από τον Cox, με μεγιστοποίηση του λογαρίθμου της μερικής πιθανοφάνειας. Έτσι, ο λογάριθμος της μερικής πιθανοφάνειας είναι:

$$\ell = \ln L(\boldsymbol{\beta}) = \sum_{j=1}^{m} \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_j - \sum_{j=1}^{m} \ln\left\{ \sum_{i \in \Re_{(j)}} e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i} \right\} \quad \textbf{(2.16)}.$$

Επειδή $\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_j = \beta_0 + \beta_1 z_{j1} + \beta_2 z_{j2} + \ldots + \beta_k z_{jk}$ άρα για κάθε $j = 1, 2, \ldots, m$ και $\xi = 0,1, \ldots, k$ θα είναι $\dfrac{\partial\left(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_j\right)}{\partial \beta_\xi} = z_{j\xi}$, οπότε παραγωγίζοντας τη μερική πιθανοφάνεια ως προς τον συντελεστή β_ξ, $\xi = 0,1, \ldots, k$, έχουμε

$$\frac{\partial \ell}{\partial \beta_\xi} = \sum_{j=1}^{m} z_{j\xi} - \sum_{j=1}^{m} \left\{ \frac{\sum_{i \in \Re_{(j)}} z_{i\xi} \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i}}{\sum_{i \in \Re_{(j)}} e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i}} \right\}.$$

Θέτοντας

$$A_{j\xi}(\boldsymbol{\beta}) = \frac{\sum_{i \in \Re_{(j)}} z_{i\xi} \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i}}{\sum_{i \in \Re_{(j)}} e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i}} \quad \textbf{(2.17)}$$

η προηγούμενη μερική παράγωγος γίνεται

$$\frac{\partial \ell}{\partial \beta_\xi} = \sum_{j=1}^{m} \left(z_{j\xi} - A_{j\xi}(\boldsymbol{\beta}) \right) \text{ για } \xi = 0,1, \ldots, k \quad \textbf{(2.18)}$$

Εξισώνοντας τις **(2.17)** με 0 για $\xi = 0,1, \ldots, k$ και λύνοντας το σύστημα αυτό (με αριθμητικές μεθόδους), παίρνουμε τους εκτιμητές $\hat{\boldsymbol{\beta}}$ των συντελεστών, ενώ οι διασπορές των συντελεστών και οι συνδιακυμάνσεις μεταξύ αυτών υπολογίζονται από τον **πίνακα παρατηρούμενης πληροφορίας** (*observed information matrix*), το $(\xi, \eta)-$στοιχείο του οποίου είναι ίσο με

$$-\frac{\partial^2 \ell}{\partial \beta_\xi \partial \beta_\eta} = -\frac{\partial}{\partial \beta_\eta}\left\{\sum_{j=1}^{m}\left(z_{j\xi} - A_{j\xi}(\boldsymbol{\beta})\right)\right\} = \sum_{j=1}^{m}\frac{\partial A_{j\xi}(\boldsymbol{\beta})}{\partial \beta_\eta} =$$

$$= \sum_{j=1}^{m}\frac{\left\{\sum_{i\in\Re_{(j)}} z_{i\xi}z_{i\eta}\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}\right\}\cdot\left\{\sum_{i\in\Re_{(j)}} e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}\right\} - \left\{\sum_{i\in\Re_{(j)}} z_{i\xi}\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}\right\}\cdot\left\{\sum_{i\in\Re_{(j)}} z_{i\eta}e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}\right\}}{\left(\sum_{i\in\Re_{(j)}} e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}\right)^2} =$$

$$= \sum_{j=1}^{m}\left(\frac{\left\{\sum_{i\in\Re_{(j)}} z_{i\xi}z_{i\eta}\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}\right\}}{\sum_{i\in\Re_{(j)}} e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}} - \frac{\left\{\sum_{i\in\Re_{(j)}} z_{i\xi}\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}\right\}\cdot\left\{\sum_{i\in\Re_{(j)}} z_{i\eta}e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}\right\}}{\left(\sum_{i\in\Re_{(j)}} e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}\right)^2}\right) =$$

$$= \sum_{j=1}^{m}\left(\frac{\sum_{i\in\Re_{(j)}} z_{i\xi}z_{i\eta}\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}}{\sum_{i\in\Re_{(j)}} e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}} - A_{j\xi}(\boldsymbol{\beta})\cdot A_{j\eta}(\boldsymbol{\beta})\right)$$

και τελικά

$$\boxed{-\frac{\partial^2 \ell}{\partial \beta_\xi \partial \beta_\eta} = \sum_{j=1}^{m} C_{j\xi\eta}} \quad \textbf{(2.19)}$$

όπου

$$\boxed{C_{j\xi\eta} = \frac{\sum_{i\in\Re_{(j)}} z_{i\xi}z_{i\eta}\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}}{\sum_{i\in\Re_{(j)}} e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}} - A_{j\xi}(\boldsymbol{\beta})\cdot A_{j\eta}(\boldsymbol{\beta})} \quad \textbf{(2.20)}$$

Σημειωτέον ότι στην εργασία του 1972, ο Cox (βλ. [24] σελ. 191) επισημαίνει ότι η ποσότητα $A_{j\xi}(\boldsymbol{\beta})$ που ορίζεται στην εξίσωση **(2.17)**, αποτελεί το σταθμισμένο μέσο όρο των $z_{i\xi}$ πάνω στον πληθυσμό $\Re_{(j)}$, χρησιμοποιώντας τους εκθετικούς συντελεστές βαρύτητας $e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}$.

2.2.3. Οι ισόπαλοι χρόνοι στο μοντέλο του Cox: Στην προηγούμενη μοντελοποίηση, θεωρήσαμε ότι σε κάθε χρονική στιγμή $t_{(j)}$, $j = 1, 2, \ldots, m$ έχουμε μόνο μία αποτυχία (θάνατο), δηλ. ότι $d_j = 1$. Στην πράξη όμως μπορεί να εμφανίζεται κατά τη χρονική στιγμή $t_{(j)}$, πλήθος αποτυχιών $d_j > 1$, $j = 1, 2, \ldots, m$ **(ισόπαλοι χρόνοι αποτυχίας** (*ties*)**)**. Αυτό μπορεί να συμβεί για λόγους όπως (βλ. [12]):

- τα δεδομένα είναι διακριτά και έτσι έχουμε θετική πιθανότητα αποτυχίας κατά την χρονική στιγμή $t_{(j)}$, $j = 1, 2, \ldots, m$
- τα δεδομένα είναι συνεχή, αλλά είναι ομαδοποιημένα και έτσι, ο αριθμός d_j μετράει το πλήθος αποτυχιών σε κάποιο διάστημα (κλάση) γύρω από τη χρονική στιγμή $t_{(j)}$, $j = 1, 2, \ldots, m$
- τα δεδομένα είναι συνεχή και όχι ομαδοποιημένα, αλλά οι παρατηρούμενες ισοπαλίες προκύπτουν από σφάλματα μέτρησης των χρόνων αποτυχίας.

Οι κύριες προσεγγίσεις της μερικής πιθανοφάνειας **(2.15)** στους ισόπαλους χρόνους διακοπής είναι αυτές των Breslow (1972) και Efron (1977).

Συγκεκριμένα, έστω $t_{(1)}$, $t_{(2)}$, …, $t_{(n)}$ οι χρόνοι αποτυχίας των n μονάδων και έστω $T_1 = T_2$ δύο ισόπαλοι χρόνοι διακοπής (όπου $T_1, T_2 \in \{t_{(1)}, t_{(2)}, \ldots, t_{(n)}\}$. Θέτουμε $P_j(t) = \begin{cases} e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_j} & \text{αν } t_{(j)} \geq t \\ 0 & \text{αλλιώς} \end{cases}$ για $j = 1, 2, \ldots, m$, δηλ. $P_j(t) = I\left(t_{(j)} \geq t\right) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_j}$, $j = 1, 2, \ldots, m$ (όπου με $I\left(t_{(j)} \geq t\right)$ είναι η δείκτρια συνάρτηση του ενδεχομένου $\{t_{(j)} \geq t\}$) (βλ. [26] σελ. 400 και [27] σελ. 15-17). Αν οι χρόνοι T_1 και T_2 δεν ήταν ισόπαλοι, τότε η μερική πιθανοφάνεια στο μοντέλο του Cox θα προέκυπτε από την εξίσωση **(2.15)** και τη διάταξη των χρόνων T_1 και T_2. Έτσι, για $T_1 < T_2$ η μερική πιθανοφάνεια θα ήταν ίση με

$$\frac{e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_1}}{\sum_{i\in\Re_{(T_1)}} e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}}\cdot\frac{e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_2}}{\sum_{i\in\Re_{(T_2)}} e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}}=\frac{P_1(T_1)}{P_1(T_1)+P_2(T_1)+\ldots+P_n(T_1)}\cdot\frac{P_2(T_2)}{P_2(T_2)+P_3(T_2)+\ldots+P_n(T_2)}$$

ενώ για $T_1 > T_2$ θα ήταν

$$\frac{e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_2}}{\sum_{i\in\Re_{(T_2)}} e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}}\cdot\frac{e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_1}}{\sum_{i\in\Re_{(T_1)}} e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}}=\frac{P_2(T_2)}{P_1(T_2)+P_2(T_2)+\ldots+P_n(T_2)}\cdot\frac{P_1(T_1)}{P_1(T_1)+P_3(T_1)+\ldots+P_n(T_1)}$$

Η προσέγγιση του Breslow χρησιμοποιεί το άθροισμα $\sum_{i=1}^{n} P_i(T_2)$ και στους δύο παρονομαστές και δίνει τελικά την εκτίμηση

$$\frac{P_1(T_1)\cdot P_2(T_2)}{\left\{\sum_{i=1}^{n} P_i(T_2)\right\}^2} \quad \textbf{(2.21)}$$

Αντίθετα, η προσέγγιση του Efron χρησιμοποιεί την εκτίμηση

$$\frac{P_1(T_1)}{P_1(T_2)+P_2(T_2)+\ldots+P_n(T_2)}\cdot\frac{P_2(T_2)}{0.5P_1(T_2)+0.5P_2(T_2)+P_3(T_2)+\ldots+P_n(T_2)} \quad \textbf{(2.22)}$$

δηλ. στον δεύτερο παρονομαστή χρησιμοποιεί τον μέσο όρο για τους $P_1(T_2)$ και $P_2(T_2)$.

Στη γενική περίπτωση όπου έχουμε k το πλήθος ισοπαλίες ($T_1 = T_2 = \ldots = T_k$, όπου $T_1, T_2, \ldots, T_k \in \{t_{(1)}, t_{(2)}, \ldots, t_{(n)}\}$ με $k \le n$), οι προσεγγίσεις **(2.21)** του Breslow και **(2.22)** του Efron διαμορφώνονται ως εξής (βλ. [26]):

$$\boxed{\prod_{i=1}^{k}\frac{P_i(T_1)}{\sum_{i=1}^{n} P_i(T_1)}=\frac{\prod_{i=1}^{k} P_i(T_1)}{\left\{\sum_{i=1}^{n} P_i(T_1)\right\}^k}} \quad \textbf{(τύπος του Breslow) (2.23)}$$

και

$$\prod_{i=1}^{k} \frac{P_i(T_1)}{\frac{k-i+1}{k}\sum_{j=1}^{k} P_j(T_1) + \sum_{j=k+1}^{n} P_j(T_1)}$$ **(τύπος του Efron) (2.24)**.

2.2.4. Το στρωματοποιημένο μοντέλο του Cox: Το μοντέλο του Cox, έχει εξ ορισμού την ιδιότητα του αναλογικών κινδύνων, η οποία περιγράφηκε στην εξίσωση **(2.5)** (για τον έλεγχο της ιδιότητας αυτής αναφερόμαστε στη συνέχεια της εργασίας). Στην περίπτωση που κάποια μεταβλητή παραβιάζει την ιδιότητα της αναλογικότητας, είναι δυνατόν αυτή να χωρισθεί σε **στρώματα** (ομάδες) και να προσαρμοσθεί στα δεδομένα το **στρωματοποιημένο μοντέλο του Cox** (*stratified Cox model*). Π.χ. (βλ. [44]) ας υποθέσουμε ότι μελετούμε το χρόνο ανάρρωσης από μία ασθένεια σε ένα δείγμα ατόμων στα οποία έχει δοθεί είτε ένα φάρμακο είτε ένα εικονικό φάρμακο (*placebo*). Αν υποψιαζόμαστε ότι η ιδιότητα της αναλογικότητας παραβιάζεται στη συμμεταβλητή "φάρμακο – εικονικό φάρμακο" και στη συμμεταβλητή "ηλικία (κάτω των 40 – άνω των 40)", τότε εφαρμόζουμε τη στρωματοποιημένη ανάλυση με βάση τον ακόλουθο πίνακα:

	Κάτω των 40	**Άνω των 40**
Φάρμακο	1	2
Εικονικό φάρμακο	3	4

Συγκεκριμένα, ορίζουμε 4 το πλήθος ομάδες (στρώματα) με βάση την αρίθμηση των κελιών του παραπάνω πίνακα και προσαρμόζουμε το μοντέλο του Cox με ίδιους συντελεστές σε όλα τα στρώματα, αλλά διαφορετική βασική συνάρτηση κινδύνου (*baseline hazard*) για κάθε στρώμα. Έτσι, για κάθε στρώμα $k \in \{1, 2, 3, 4\}$ το οποίο περιέχει n_k το πλήθος μονάδες του δείγματος, θέτουμε

$$h_k(t \mid \mathbf{z}_m) = h_{0k}(t) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_m}$$

όπου $h_{0k}(t)$, $k = 1, 2, 3, 4$ είναι η βασική συνάρτηση κινδύνου του $k-$στρώματος και $m = 1, 2, \ldots, n_k$ είναι ο δείκτης της μονάδας του $k-$στρώματος στην οποία αναφερόμαστε. Δηλ. έχουμε ίδιους συντελεστές $\boldsymbol{\beta}$ για όλα τα στρώματα, αλλά διαφορετική συνάρτηση κινδύνου για κάθε

στρώμα. Στη συνέχεια, εφαρμόζουμε τη μέθοδο της μέγιστης μερικής πιθανοφάνειας σε κάθε στρώμα βάσει της εξίσωσης **(2.16)** και έχουμε

$$\ell_k = \ln L_k(\boldsymbol{\beta}) = \sum_{m=1}^{n_k} \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_{km} - \sum_{m=1}^{n_k} \ln\left\{ \sum_{i \in \Re_{(m)}} e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_{ki}} \right\}$$

οπότε η μερική λογαριθμοποιημένη πιθανοφάνεια για όλα τα στρώματα είναι

$$\ell = \sum_k \ell_k$$

2.3. ΠΡΟΣΑΡΜΟΓΗ ΜΟΝΤΕΛΟΥ ΑΝΑΛΟΓΙΚΩΝ ΚΙΝΔΥΝΩΝ

2.3.1. Γραφικός έλεγχος της υπόθεσης αναλογικών κινδύνων: Η προσαρμογή ενός μοντέλου αναλογικών κινδύνων σε δεδομένα διάρκειας ζωής, βασίζεται στην αναλογικότητα του κινδύνου, δηλ. στο ότι ο λόγος των τιμών της συνάρτησης κινδύνου για μία δεδομένη χρονική στιγμή εξαρτάται μόνο από τις τιμές των συμμεταβλητών και όχι από τη χρονική στιγμή (όπως περιγράφηκε με τη γενική εξίσωση **(2.6)** και την ειδική για το μοντέλο του Cox εξίσωση **(2.8)**). Γι' αυτό, το πρώτο πράγμα που πρέπει να γίνει όσον αφορά την προσαρμογή του μοντέλου, είναι ο έλεγχος της υπόθεσης αναλογικών κινδύνων.

Ένας πολύ απλός τρόπος του ελέγχου καταλληλότητας του μοντέλου, είναι ο γραφικός έλεγχος. Π.χ. λογαριθμίζοντας την εξίσωση **(2.12)** που δίνει τη συνάρτηση επιβίωσης στο μοντέλο του Cox, παίρνουμε $-\ln S(t \mid \boldsymbol{z}_i) = e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i} \cdot H_0(t)$ και λογαριθμίζοντας ξανά:

$$\ln(-\ln S(t \mid \boldsymbol{z}_i)) = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i + \ln H_0(t) \quad \textbf{(2.25)}$$

Η εξίσωση **(2.25)** είναι της μορφής $\varphi(t \mid \boldsymbol{z}_i) = c + \psi(t)$, δηλ. περιγράφει δύο συναρτήσεις (τις $\varphi(t \mid \boldsymbol{z}_i) = \ln(-\ln S(t \mid \boldsymbol{z}_i))$ και $\psi(t) = \ln H_0(t)$) που διαφέρουν κατά μία σταθερά (την $c = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i$), άρα οι γραφικές τους παραστάσεις έχουν μεταξύ τους σχέση κατακόρυφης μετατόπισης (είναι τρόπον τινά

"παράλληλες"). Τότε όμως, οι γραφικές παραστάσεις των συναρτήσεων $\varphi(t\,|\,\mathbf{z}_i)$ για τις διάφορες τιμές των $\mathbf{z}_i$ θα έχουν μεταξύ τους σχέση οριζόντιας μετατόπισης (είναι και αυτές μεταξύ τους "παράλληλες", με σχέση οριζόντιας μετατόπισης).

Έτσι, βάσει της εξίσωσης **(2.25)** προκύπτει ο γραφικός έλεγχος της αναλογικότητας, ο οποίος συνίσταται (βλ. [3] σελ. 101) στον χωρισμό των δεδομένων σε ομάδες που αντιστοιχούν σε επιλεγμένες τιμές $\mathbf{z}_k$, στη συνέχεια στην εκτίμηση $\hat{S}(t\,|\,\mathbf{z}_k)$ της συνάρτησης κινδύνου μέσω της εκτιμήτριας Kaplan-Meier για κάθε ομάδα και τέλος στην κατασκευή των γραφικών παραστάσεων των συναρτήσεων $\varphi(t\,|\,\mathbf{z}_k) = \ln\left(-\ln \hat{S}(t\,|\,\mathbf{z}_k)\right)$ ως προς t. Η ύπαρξη "παραλληλίας" μεταξύ αυτών των γραφικών παραστάσεων, επιβεβαιώνει την υπόθεση του αναλογικών κινδύνων.

2.3.2. Έλεγχος της υπόθεσης αναλογικών κινδύνων μέσω των υπολοίπων: Ένας άλλος τρόπος ελέγχου της υπόθεσης αναλογικών κινδύνων, είναι η χρήση των υπολοίπων του μοντέλου (στη γενική περίπτωση της Γραμμικής Παλινδρόμησης, ο όρος **υπόλοιπα** (*residuals*) δηλώνει τις διαφορές $\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{z}_i$ μεταξύ των παρατηρούμενων τιμών y_i και των προσαρμοσμένων τιμών $\hat{y}_i$).

Για μία πρώτη ανάγνωση σε διάφορες μεθόδους χρήσης των υπολοίπων σε δεδομένα διάρκειας ζωής παραπέμπουμε στο [29]. Χαρακτηριστικά, αναφέρουμε εδώ τα υπόλοιπα Cox & Snell και τα υπόλοιπα Schoenfeld.

- Τα υπόλοιπα Cox & Snell (Sir David Cox & E. Joyce Snell, 1968) βασίζονται στην ακόλουθη ιδιότητα της Θεωρίας Πιθανοτήτων (βλ. [3] σελ. 111): «Αν Y είναι μία τυχαία μεταβλητή με αθροιστική συνάρτηση κατανομής $F(\cdot)$, τότε η τυχαία μεταβλητή $V = F(Y)$ ακολουθεί την Ομοιόμορφη κατανομή $U(0,1)$ και η τυχαία μεταβλητή $W = -\ln(1-F(Y))$ ακολουθεί την Εκθετική κατανομή $\varepsilon(1)$». Έτσι, αν $F(t)$ είναι η αθροιστική συνάρτηση κατανομής των δεδομένων διάρκειας ζωής και $S(t)$ η συνάρτηση

επιβίωσης, τότε η τυχαία μεταβλητή $F(t\,|\,\boldsymbol{z}_i) \sim U(0,1)$, άρα η τυχαία μεταβλητή $-\ln S(t\,|\,\boldsymbol{z}_i) = -\ln(1-F(t\,|\,\boldsymbol{z}_i)) \sim \varepsilon(1)$. Χρησιμοποιώντας λοιπόν πάλι την εκτιμήτρια Kaplan-Meier $\hat{S}(t\,|\,\boldsymbol{z}_i)$, υπολογίζουμε ως υπόλοιπα τις τιμές

$$\hat{r}_i = -\ln \hat{S}(t_i\,|\,\boldsymbol{z}_i) \quad \textbf{(υπόλοιπα Cox-Snell) (2.26)}$$

και εξετάζουμε γραφικά αν $r_i \sim \varepsilon(1)$ οπότε αποδεχόμαστε, ή αλλιώς απορρίπτουμε το μοντέλο. Για την περίπτωση των δεξιά αποκομμένων παρατηρήσεων χρησιμοποιούμε ως υπόλοιπο το $r_i = 1 - \ln \hat{S}(t_i\,|\,\boldsymbol{z}_i)$ (διορθωμένο υπόλοιπο – βλ. [3] σελ. 111-112).

- Τα υπόλοιπα Schoenfeld (David Schoenfeld, 1982, βλ. [31]) αφορούν το μοντέλο του Cox. Δεν ορίζονται για λογοκριμένα δεδομένα, αλλά υπερτερούν αφενός γιατί δεν χρειάζονται εκτίμηση της συνάρτησης επιβίωσης, αφετέρου γιατί αντί για τον απλό υπολογισμό ενός υπολοίπου για κάθε μονάδα, υπολογίζεται ένα υπόλοιπο για κάθε μονάδα και για κάθε συμμεταβλητή. Συγκεκριμένα, για την $i-$μονάδα και την $u-$συμμεταβλητή ορίζουμε το υπόλοιπο Schoenfeld ως τη διαφορά μεταξύ της τιμής $\boldsymbol{z}_{iu}$ της $u-$συμμεταβλητής μείον την αναμενόμενη τιμής $E[\boldsymbol{z}_{iu}\,|\,\Re_i]$ αυτής, όπου με $\Re_i$ συμβολίζουμε το σύνολο των μονάδων που βρίσκονται σε κίνδυνο αμέσως πριν τη χρονική στιγμή t_i, δηλ:

$$\boxed{\hat{r}_{iu} = \boldsymbol{z}_{iu} - E[\boldsymbol{z}_{iu}\,|\,\Re_i]} \quad \textbf{(υπόλοιπα Schoenfeld) (2.27)}$$

Υπολογίζουμε τώρα την αναμενόμενη τιμή $E[\boldsymbol{z}_{iu}\,|\,\Re_i] = \sum_{j\in\Re_i} \boldsymbol{z}_{ju} p_j$, όπου p_j είναι η πιθανότητα διακοπής μίας μονάδας $j \in \Re_i$. Γνωρίζουμε ότι $p_j = \dfrac{e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_{ju}}}{\sum_{m\in\Re_i} e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_{mu}}}$, άρα:

$$E[\boldsymbol{z}_{iu}\,|\,\Re_i] = \sum_{j\in\Re_i} \boldsymbol{z}_{ju} \frac{e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_{ju}}}{\sum_{m\in\Re_i} e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_{mu}}} = \frac{\sum_{j\in\Re_i} \boldsymbol{z}_{ju} e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_{ju}}}{\sum_{m\in\Re_i} e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_{mu}}} \quad \textbf{(2.28)}$$

Αντικαθιστώντας στην (2.27) παίρνουμε τελικά:

$$\hat{r}_{iu} = \mathbf{z}_{iu} - \frac{\sum_{j \in \Re_i} \mathbf{z}_{ju} e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_{ju}}}{\sum_{m \in \Re_i} e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_{mu}}}$$ **(υπόλοιπα Schoenfeld) (2.29)**

Η υπόθεση των αναλογικών κινδύνων ικανοποιείται για την $u-$συμμεταβλητή, όταν η γραφική παράσταση $\hat{r}_{iu} + \hat{\beta}_u$ ως προς τον χρόνο δείχνει μία οριζόντια γραμμή (για ανάλυση αυτού, βλ. [35]).

2.3.3. Προσαρμογή παραμετρικού μοντέλου αναλογικών κινδύνων με τη μέθοδο μεγίστης πιθανοφάνειας: Θεωρούμε ένα δείγμα δεδομένων διάρκειας ζωής μεγέθους n και σε αυτό θέλουμε να προσαρμόσουμε ένα παραμετρικό μοντέλο αναλογικών κινδύνων, δηλ. μοντέλο στο οποίο η τυχαία μεταβλητή του χρόνου T ακολουθεί γνωστή κατανομή (**προσαρμογή μοντέλου**: η περιγραφή του μοντέλου μέσω εξισώσεων και η εκτίμηση των παραμέτρων του μοντέλου).

Υποθέτουμε ότι το δείγμα αποτελείται από τις παρατηρήσεις $\{t_i : i = 1, 2, \ldots, \mathrm{n}\}$. Στην περίπτωση όπου όλα τα t_i, $i = 1, 2, \ldots, \mathrm{n}$ είναι χρόνοι αποτυχίας (μη λογοκριμένα δεδομένα), η συνάρτηση πιθανοφάνειας είναι $L = \prod_{i=1}^{n} f(t_i)$, όπου $f(\cdot)$ η συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής του χρόνου. Αν τώρα στο δείγμα υπάρχουν αριστερά και δεξιά λογοκριμένες παρατηρήσεις, τότε η συνάρτηση πιθανοφάνειας υπολογίζεται ως εξής (βλ. [3] σελ. 78-79 και [33] σελ. 52-53):

Γράφουμε το σύνολο $\mathrm{A} = \{t_i : i = 1, 2, \ldots, \mathrm{n}\}$ των παρατηρήσεων ως ένωση υποσυνόλων του: $\mathrm{A} = V \cup C_R \cup C_L$, όπου:

- $V = \{t_i \in \mathrm{A} / t_i \text{ μη λογοκριμένη}\}$
- $C_R = \{t_i \in \mathrm{A} / t_i \text{ δεξιά λογοκριμένη}\}$ και
- $C_L = \{t_i \in \mathrm{A} / t_i \text{ αριστερά λογοκριμένη}\}$.

Η συνάρτηση πιθανοφάνειας θα είναι τώρα το γινόμενο των συναρτήσεων πιθανοφανειών των τριών υποσυνόλων, δηλ. θα ισχύει $L = L_V \cdot L_{C_R} \cdot L_{C_L}$. Αρκεί λοιπόν τώρα να υπολογίσουμε τις τρεις επιμέρους συναρτήσεις $\mathrm{L_V}$, $\mathrm{L_{C_R}}$ και $\mathrm{L_{C_L}}$.

- Προφανώς για τις παρατηρήσεις του συνόλου V θα ισχύει ό,τι και πριν, δηλ. $L_V = \prod_{i \in V} f(t_i)$.
- Για τις δεξιά λογοκριμένες παρατηρήσεις του συνόλου C_R, θα έχουμε ότι $L_{C_R} = \prod_{i \in C_R} P[T > t_i] = \prod_{i \in C_R} S(t_i)$
- Για τις αριστερά λογοκριμένες παρατηρήσεις του συνόλου C_L, θα έχουμε ότι $L_{C_L} = \prod_{i \in C_L} P[T \le t_i] = \prod_{i \in C_L} [1 - S(t_i)]$

Από τα παραπάνω αποτελέσματα, προκύπτει η συνάρτηση πιθανοφάνειας ενός δείγματος στο οποίο υπάρχουν δεξιά λογοκριμένα και αριστερά λογοκριμένα δεδομένα:

$$\boxed{L = \prod_{i \in V} f(t_i) \cdot \prod_{i \in C_R} S(t_i) \cdot \prod_{i \in C_L} [1 - S(t_i)]} \quad \textbf{(2.30)}$$

Θεωρούμε στη συνέχεια τις δείκτριες συναρτήσεις:

$D_{1i} = \begin{cases} 1 & \text{αν η παρατήρηση είναι μη λογοκριμένη} \\ 0 & \text{αλλιώς} \end{cases}$ και

$D_{2i} = \begin{cases} 1 & \text{αν η παρατήρηση είναι δεξιά λογοκριμένη} \\ 0 & \text{αλλιώς} \end{cases}$, για $i = 1, 2, \ldots, \mathrm{n}$. Τότε, η τιμή κάθε ζεύγους $(\delta_{1i}, \delta_{2i})$ δηλώνει την κατάσταση της παρατήρησης i:

$(D_{1i}, D_{2i}) =$	Η παρατήρηση i είναι:
$(1, 0)$	μη λογοκριμένη
$(0, 1)$	δεξιά λογοκριμένη
$(0, 0)$	αριστερά λογοκριμένη

(περίπτωση $(D_{1i}, \mathrm{D}_{2i}) = (1, 1)$ προφανώς δεν υφίσταται).

Χρησιμοποιώντας τις παραπάνω δείκτριες συναρτήσεις δ_{1i} και δ_{2i}, η συνάρτηση πιθανοφάνειας γράφεται::

$$L = \prod_{i=1}^{n} \left(f(t_i)\right)^{D_{1i}} \left(S(t_i)\right)^{D_{2i}} \left[1 - S(t_i)\right]^{(1-D_{1i})(1-D_{2i})} \quad \textbf{(2.31)}$$

Λογαριθμίζοντας την **(2.31)**, παίρνουμε την λογαριθμοποιημένη πιθανοφάνεια $\ell = \ln L$, δηλ:

$$\ell = \sum_{i=1}^{n} D_{1i} \ln f(t_i) + \sum_{i=1}^{n} D_{2i} \ln S(t_i) + \sum_{i=1}^{n} (1 - D_{1i})(1 - D_{2i}) \ln\left[1 - S(t_i)\right] \quad \textbf{(2.32)}$$

της οποίας η μεγιστοποίηση μας δίνει την εκτίμηση των παραμέτρων του μοντέλου.

Στην περίπτωση που το δείγμα μας αποτελείται μόνο από χρόνους αποτυχίας και δεξιά λογοκριμένους χρόνους, τα δεδομένα περιγράφονται από το σύνολο $A = \{(X_i, D_i) : i = 1, 2, \ldots, n\}$, όπου:

- X_i είναι ο χρόνος αποτυχίας ή λογοκρισίας, $i = 1, 2, \ldots, n$
- $D_i = \begin{cases} 1 & \text{αν ο χρόνος } X_i \text{ είναι χρόνος διακοπής} \\ 0 & \text{αν ο χρόνος } X_i \text{ είναι χρόνος λογοκρισίας} \end{cases}$ είναι η δείκτρια συνάρτηση λογοκρισίας.

Οι συναρτήσεις πιθανοφάνειας και λογαριθμοποιημένης πιθανοφάνειας, όπως περιγράφηκαν στις εξισώσεις **(2.31)** και **(2.32)** αντίστοιχα, μετασχηματίζονται τώρα σε:

$$L = \prod_{i=1}^{n} \left(f(X_i)\right)^{D_i} \left(S(X_i)\right)^{1-D_i} \quad \textbf{(2.33)}$$

και

$$\ell = \sum_{i=1}^{n} D_i \ln f(X_i) + \sum_{i=1}^{n} (1 - D_i) \ln S(X_i) \quad \textbf{(2.34)}$$

Εναλλακτικά, η εξίσωση **(2.33)** γράφεται $L = \prod_{i=1}^{n} \left(\frac{f(X_i)}{S(X_i)} \right)^{D_i} S(X_i)$ ή ισοδύναμα:

$$\boxed{L = \prod_{i=1}^{n} \left(h(X_i)\right)^{D_i} \cdot S(X_i)} \quad \textbf{(2.35)}$$

και λογαριθμίζοντας οδηγούμαστε στην:

$$\boxed{\ell = \sum_{i=1}^{n} \{D_i \ln h(X_i) + \ln S(X_i)\}} \quad \textbf{(2.36)}$$

Ειδικά για το μοντέλο του Cox, επειδή $h(X_i \mid \mathbf{z}_i) = h_0(X_i) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i}$ και $S(X_i \mid \mathbf{z}_i) = e^{-e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i} \cdot H_0(X_i)}$, η εξίσωση **(2.36)** μας δίνει:

$$\boxed{\ell(X_i, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ D_i \left[\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i + \ln h_0(X_i) \right] - e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i} \cdot H_0(X_i) \right\}} \quad \textbf{(2.37)}$$

Αν μάλιστα επικεντρωθούμε στο παραμετρικό μοντέλο του Cox και θεωρήσουμε και το διάνυσμα $\boldsymbol{\theta}$ των παραμέτρων της κατανομής του χρόνου T, τότε η **(2.37)** ισοδυναμεί με:

$$\boxed{\ell(X_i, \boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \left\{ D_i \left[\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i + \ln h_0(X_i, \boldsymbol{\theta}) \right] - e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i} \cdot H_0(X_i, \boldsymbol{\theta}) \right\}} \quad \textbf{(2.38)}$$

Λύνοντας (με αριθμητικές μεθόδους όπως π.χ. Newton-Raphson) το σύστημα $\frac{\partial \ell(X_i, \boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = 0$ και $\frac{\partial \ell(X_i, \boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$, παίρνουμε την εκτίμηση $\hat{\boldsymbol{\beta}}$ των συντελεστών των συμμεταβλητών και την εκτίμηση $\hat{\boldsymbol{\theta}}$ των παραμέτρων της κατανομής του T.

ΠΑΡΑΔΕΙΓΜΑ: (βάσει αντίστοιχου παραδείγματος του [34] σελ. 9-10) Έστω δείγμα δεδομένων διάρκειας ζωής $\mathrm{A} = \{(X_i, D_i) : i = 1, 2, \ldots, \mathrm{n}\}$, όπου i, X_i, D_i, $i = 1, 2, \ldots, \mathrm{n}$ όπως στη θεωρία παραπάνω και έστω το διάνυσμα συμμεταβλητών $\mathbf{z}_i = [\, z_{i1}, z_{i2}, \ldots, z_{ik} \,]^{\mathrm{T}}$ για κάθε μονάδα i, $i = 1, 2, \ldots, \mathrm{n}$. Έστω r το πλήθος των λογοκριμένων παρατηρήσεων στο δείγμα ($r \le n$), δηλ. $\sum_{i=1}^{n} D_i = n - r$. Προσαρμόζουμε στο δείγμα το μοντέλο του Cox, με διάνυσμα

παραμέτρων το $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_k]^{\mathrm{T}}$, υποθέτοντας ως βασική κατανομή την Εκθετική $\varepsilon(\lambda)$, όπου $\lambda > 0$ παράμετρος προς εκτίμηση. Τότε οι βασικές συναρτήσεις κινδύνου και σωρευτικού κινδύνου μίας μονάδας, είναι (βλ. παράγραφο **(1.3.1.)**) $f(t) = \lambda e^{-\lambda t}$, $S(t) = e^{-\lambda t}$, $h(t) = \lambda$ και $H(t) = \lambda t$ αντίστοιχα. Χρησιμοποιώντας την εξίσωση **(2.38)**, η συνάρτηση λογαριθμοποιημένης πιθανοφάνειας είναι

$$\ell(X_i, \boldsymbol{\beta}, \lambda) = \sum_{i=1}^{n} \left\{ D_i \left[\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i + \ln h_0(X_i) \right] - e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i} \cdot H_0(X_i) \right\} \Leftrightarrow$$

$$\Leftrightarrow \ell(X_i, \boldsymbol{\beta}, \lambda) = \sum_{i=1}^{n} \left\{ D_i \left[\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i + \ln \lambda \right] - e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i} \cdot \lambda X_i \right\} \Leftrightarrow$$

$$\Leftrightarrow \ell(X_i, \boldsymbol{\beta}, \lambda) = \sum_{i=1}^{n} D_i \left(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i \right) + (n - r) \ln \lambda - \lambda \sum_{i=1}^{n} e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i} \cdot X_i .$$

Για την εκτίμηση των παραμέτρων $\boldsymbol{\beta}$ και λ έχουμε

$$\frac{\partial \ell(X_i, \boldsymbol{\beta}, \lambda)}{\partial \lambda} = \frac{n - r}{\lambda} - \sum_{i=1}^{n} e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i} \cdot X_i$$

και

$$\frac{\partial \ell(X_i, \boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \beta_\xi} = \sum_{i=1}^{n} D_i z_{i\xi} - \lambda \sum_{i=1}^{n} z_{i\xi} \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i} \cdot X_i \quad \text{για } \xi = 1, 2, \ldots, k .$$

Λύνουμε το σύστημα εξισώσεων $\begin{cases} \dfrac{\partial \ell(X_i, \boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \lambda} = 0 \\ \text{και} \\ \dfrac{\partial \ell(X_i, \boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \beta_\xi} = 0 \end{cases}$, $\xi = 1, 2, \ldots, k$ και παίρνουμε τις εκτιμήτριες μεγίστης πιθανοφάνειας $\hat{\lambda} = \dfrac{n - r}{\sum_{i=1}^{n} e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i} \cdot X_i}$ και $\hat{\boldsymbol{\beta}}$

(προφανώς το σύστημα των εξισώσεων $\dfrac{\partial \ell(X_i, \boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \beta_\xi} = 0$, $\xi = 1, 2, \ldots, k$ που δίνει την εκτιμήτρια $\hat{\boldsymbol{\beta}}$, λύνεται με αριθμητικές μεθόδους).

2.4. ΚΡΙΤΗΡΙΑ ΕΠΙΛΟΓΗΣ ΜΟΝΤΕΛΟΥ

2.4.1. Έλεγχος υποθέσεων στο μοντέλο αναλογικών κινδύνων: Μετά την προσαρμογή του μοντέλου αναλογικών κινδύνων στα δεδομένα, μπορούμε να εκτελέσουμε στατιστικούς ελέγχους υποθέσεων και να κατασκευάσουμε διαστήματα εμπιστοσύνης για τις παραμέτρους και τους εκτιμώμενους συντελεστές των συμμεταβλητών.

- Ο **έλεγχος λόγου πιθανοφανειών** (*likelihood ratio test*): χρησιμοποιείται για τη σύγκριση δύο μοντέλων, όταν το ένα είναι εμφωλευμένο (*nested*) στο άλλο. Τέτοια περίπτωση έχουμε όταν π.χ. έχουμε προσαρμόσει ένα μοντέλο κατανομής Weibull και θέλουμε να ελέγξουμε αν το μοντέλο είναι της Εκθετικής κατανομής (αφού η Εκθετική κατανομή είναι ειδική περίπτωση κατανομής Weibull με παράμετρο σχήματος ίση με 1) ή όταν θέλουμε να ελέγξουμε τη σημαντικότητα μίας συμμεταβλητής, οπότε εφαρμόζουμε τον έλεγχο λόγου πιθανοφανειών μεταξύ του μοντέλου χωρίς και του μοντέλου με τη συμμεταβλητή (το πρώτο είναι εμφωλευμένο στο δεύτερο). Υπενθυμίζουμε ότι αν $\hat{\ell}_0$ είναι η τιμή της λογαριθμοποιημένης πιθανοφάνειας για το εμφωλευμένο μοντέλο (με τις λιγότερες παραμέτρους) και $\hat{\ell}_1$ η αντίστοιχη τιμή για το μεγαλύτερο μοντέλο (με τις περισσότερες παραμέτρους), τότε η ελεγχοσυνάρτηση του λόγου πιθανοφανειών είναι $\Lambda = -2\left(\hat{\ell}_0 - \hat{\ell}_1\right)$ και ακολουθεί την χ^2 κατανομή με αριθμό βαθμών ελευθερίας τη διαφορά του πλήθους των παραμέτρων.
- Ο **έλεγχος Wald** (*Wald test*) χρησιμοποιείται για τον έλεγχο της μηδενικής υπόθεσης $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ όπου $\boldsymbol{\theta}$ είναι (διανυσματική) παράμετρος προς εκτίμηση και $\boldsymbol{\theta}_0$ είναι μία προς έλεγχο τιμή της. Εδώ, αν $\hat{\boldsymbol{\theta}}$ είναι η εκτιμώμενη τιμή μέσω της μεθόδου μεγίστης πιθανοφάνειας και $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathrm{var}(\boldsymbol{\theta})$ είναι ο πίνακας διακύμανσης-συνδιακύμανσης, τότε η ελεγχοσυνάρτηση είναι $Z = (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}} \cdot \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ και ακολουθεί την χ^2 κατανομή με αριθμό βαθμών ελευθερίας τη διάσταση της παραμέτρου $\boldsymbol{\theta}$. Στην ειδική περίπτωση

που η παράμετρος είναι μονοδιάστατη, η ελεγχοσυνάρτηση είναι $Z = \frac{(\theta - \theta_0)^2}{\mathrm{Var}(\theta)} \sim \chi_1^2$ ή εναλλακτικά $Z = \frac{\theta - \theta_0}{\mathrm{se}(\theta)} \sim N(0,1)$.

- Το **Score Test** του C. R. Rao (βλ. [36] και [6] σελ. 12) μπορεί να χρησιμοποιηθεί κατά την προσαρμογή του μοντέλου του Cox για τον έλεγχο της υπόθεσης $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$. Συγκεκριμένα, μετά την προσαρμογή του μοντέλου μέσω της συνάρτησης ℓ μερικής πιθανοφάνειας (*partial likelihood*), η **συνάρτηση score** (*score function*) είναι η $\mathbf{U}(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ και ακολουθεί πολυδιάστατη Κανονική κατανομή με μέση τιμή $\mathbf{0}$ και πίνακα διακύμανσης-συνδιακύμανσης τον $\boldsymbol{\Sigma} = I^{-1}(\boldsymbol{\beta})$ (δηλ. τον αντίστροφο του πίνακα παρατηρούμενης πληροφορίας) . Η ελεγχοσυνάρτηση score είναι η $\mathbf{U}(\boldsymbol{\beta}_0)^{\mathrm{T}} \cdot \boldsymbol{\Sigma} \cdot \mathbf{U}(\boldsymbol{\beta}_0)$ και ακολουθεί την χ^2 κατανομή με αριθμό βαθμών ελευθερίας τη διάσταση της παραμέτρου $\boldsymbol{\beta}$.

2.4.2. Κριτήρια επιλογής συμμεταβλητών: Το επόμενο βήμα μετά την προσαρμογή ενός μοντέλου παλινδρόμησης είναι η αναζήτηση των σημαντικών συμμεταβλητών άρα η εύρεση του βέλτιστου μοντέλου που προσαρμόζεται στα δεδομένα (για εκτενή ανάλυση της θεωρίας ανάπτυξης του βέλτιστου μοντέλου, παραπέμπουμε στο [19], κεφ. 5).

- Με τον έλεγχο του λόγου πιθανοφανειών στον οποίο αναφερθήκαμε, μπορούμε να ελέγξουμε την σημαντικότητα μίας μεταβλητής z_ξ, $\xi = 0, 1, 2, \ldots, k$, αφού με αυτόν ελέγχουμε τη μηδενική υπόθεση $H_0 : \beta_\xi = 0$. Έτσι, υπολογίζουμε τη μεγιστοποιημένη λογαριθμοποιημένη πιθανοφάνεια $\hat{\ell}_0$ του μοντέλου που δεν περιέχει την εν λόγω συμμεταβλητή, στη συνέχεια την αντίστοιχη πιθανοφάνεια $\hat{\ell}_1$ του μοντέλου που περιέχει τη συμμεταβλητή και χρησιμοποιούμε την ελεγχοσυνάρτηση $\Lambda = -2\left(\hat{\ell}_0 - \hat{\ell}_1\right)$ που ακολουθεί την κατανομή χ^2 με 1 βαθμό ελευθερίας.

- Εναλλακτικά μπορούμε να χρησιμοποιήσουμε τους ελέγχους Wald ή F για τον έλεγχο της παραπάνω μηδενικής υπόθεσης (βλ. [37] για γενικά στοιχεία περί του ελέγχου F και [38] για ειδικά περί του ελέγχου F στη γραμμική παλινδρόμηση).
- Το **κριτήριο AIC** (*Akaike Information Criterion*): Πρόκειται για ένα εξαιρετικά σημαντικό κριτήριο επιλογής μεταβλητών. Δημιουργήθηκε από τον Hirotugu Akaike το 1971, ο οποίος του έδωσε την ονομασία AIC ως αρχικά του *An Information Criterion* και στη συνέχεια προτάθηκε από τον ίδιο ως κριτήριο επιλογής μεταβλητών στο [40] (για την ιστορία του κριτηρίου βλ. [39]). Για ένα μοντέλο με p το πλήθος προσαρμοσμένες παραμέτρους, στο οποίο έχει υπολογισθεί η τιμή $\hat{\ell}$ μεγίστης λογαριθμοποιημένης πιθανοφάνειας , το κριτήριο AIC ορίζεται από τη σχέση $\mathrm{AIC} = -2\hat{\ell} + 2p$. Καλύτερο μοντέλο είναι αυτό με τη μικρότερη τιμή του AIC.
- Το **κριτήριο BIC** (*Bayesian Information Criterion*): Εισήχθη από τον Gideon Schwarz το 1978 στο [42], ως εναλλακτικό κριτήριο του AIC. Για ένα μοντέλο που βασίζεται σε δείγμα μεγέθους n με p το πλήθος προσαρμοσμένες παραμέτρους, στο οποίο έχει υπολογισθεί η τιμή $\hat{\ell}$ μεγίστης λογαριθμοποιημένης πιθανοφάνειας , το κριτήριο BIC ορίζεται από τη σχέση $\mathrm{BIC} = -2\hat{\ell} + p\ln n$. Καλύτερο μοντέλο είναι αυτό με τη μικρότερη τιμή του BIC. Παρατηρούμε εδώ ότι για το ίδιο υποψήφιο μοντέλο, το κριτήριο BIC είναι πιο αυστηρό από το AIC, αφού για $n \geq 8$ ισχύει $\ln n > 2$ άρα $\mathrm{BIC} > \mathrm{AIC}$. Η ονομασία του κριτηρίου προέρχεται από το γεγονός ότι επιλέγοντας το μοντέλο με το μικρότερο BIC, επιλέγεται αυτό με την υψηλότερη **Μπεϋζιανή εκ των υστέρων πιθανότητα** (*Bayesian posterior probability*) (βλ. [41]).

Το **κριτήριο C_p-Mallows** που παρουσίασε ο Colin Lingwood Mallows το 1973 στο [43], βασίζεται στο μέσο τετραγωνικό σφάλμα της εκτίμησης μίας παραμέτρου. Έτσι, για ένα μοντέλο γραμμικής παλινδρόμησης με p παραμέτρους, προσαρμοσμένο βάσει δείγματος μεγέθους n, στο οποίο οι τιμές της εξαρτημένης μεταβλητής από το δείγμα είναι y_i και οι αντίστοιχες

εκτιμήσεις τους είναι οι $\hat{y}_i$, $i = 1, 2, \ldots, \mathrm{n}$, η συνάρτηση C_p-Mallows ορίζεται ως $C_p = \frac{SSE}{\hat{s}^2} + 2p - n$, όπου $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ είναι το **άθροισμα τετραγώνων των υπολοίπων** και $\hat{s}^2$ η εκτίμηση της διασποράς των σφαλμάτων του μοντέλου. Το καλύτερο μοντέλο είναι αυτό στο οποίο $C_p \approx p$.

ΚΕΦΑΛΑΙΟ 3

ΜΟΝΤΕΛΑ ΜΟΝΟΜΕΤΑΒΛΗΤΗΣ ΕΥΠΑΘΕΙΑΣ

3.1. ΜΟΝΤΕΛΑ ΕΥΠΑΘΕΙΑΣ

3.1.1. Εισαγωγικό παράδειγμα: (Πηγή: [45]) Υποθέτουμε έναν πληθυσμό που κατηγοριοποιείται σε 3 ομάδες και θεωρούμε δείγμα δεδομένων διάρκειας ζωής όσον αφορά τη λήψη ενός φαρμάκου ή ενός εικονικού φαρμάκου (*placebo*), σύμφωνα με τον ακόλουθο πίνακα.

Ομάδα	**Ποσοστό επί του πληθυσμού**	**Τιμή της συνάρτησης κινδύνου με το εικονικό φάρμακο**	**Τιμή της συνάρτησης κινδύνου με το πραγματικό φάρμακο**
1	40%	1	0.5
2	40%	2	1
3	20%	10	8

Από τα παραπάνω δεδομένα είναι προφανές ότι το πραγματικό φάρμακο δρα ευεργετικά και στις 3 ομάδες του πληθυσμού.

Με τη βοήθεια της R, στα δεδομένα του παραπάνω πίνακα:

- προσομοιώνεται δείγμα 100 ατόμων σε 3 ομάδες με τα αντίστοιχα ποσοστά του πίνακα
- σε καθένα από τα άτομα αυτά χορηγείται τυχαία πραγματικό ή εικονικό φάρμακο (συμμεταβλητή "treat") (δηλ. η επιλογή φαρμάκου ακολουθεί τη διωνυμική κατανομή με πιθανότητα επιτυχίας 0.5).
- Στη συνέχεια, για κάθε άτομο του δείγματος, προσομοιώνεται στην R ένας τυχαίος χρόνος διακοπής (*censoring indicator* = 1), που ακολουθεί την Εκθετική κατανομή με παράμετρο τέτοια ώστε η τιμή της συνάρτησης κινδύνου για το άτομο και το αν πήρε πραγματικό ή εικονικό φάρμακο να είναι αυτή του πίνακα.
- Η προσομοίωση του χρόνου διακοπής για το ίδιο άτομο επαναλαμβάνεται έως ότου το άθροισμα όλων των χρόνων υπερβεί το 1,

οπότε στην τελευταία προσομοίωση θεωρούμε το χρόνο δεξιά λογοκριμένο (*censoring indicator* = 0) και προχωρούμε στο επόμενο άτομο.

Έτσι, ο πίνακας αποτελεσμάτων που προκύπτει, έχει την εξής μορφή:

	id	group	treat	time	status
[1,]	1	1	0	0.77	1
[2,]	1	1	0	0.19	1
[3,]	1	1	0	0.81	0
[4,]	2	1	1	1.00	0
[5,]	3	3	1	0.32	1
[6,]	3	3	1	0.21	1
[7,]	3	3	1	0.03	1
[8,]	3	3	1	0.05	1
[9,]	3	3	1	0.06	1
[10,]	3	3	1	0.08	1

Προσαρμόζουμε στην R το μοντέλο του Cox στα δεδομένα μας, χρησιμοποιώντας τη συνάρτηση coxph:

```
> myfit1
Call:
coxph(formula = Surv(time, status ~ treat)

       coef      exp(coef)      se(coef)    z     p
treat  -0.0619   0.94           0.117    -0.531  0.6

Likelihood ratio test=0.28  on 1 df, p=0.595  n= 397, number of events= 297
```

Από τα αποτελέσματα παρατηρούμε ότι ο έλεγχος του λόγου πιθανοφανειών για τη συμμεταβλητή "treat" δίνει p-τιμή ίση με 0.595 , άρα η "treat" δε θεωρείται σημαντική μεταβλητή, πράγμα που έρχεται σε αντίθεση με αυτό που ήδη γνωρίζουμε, ότι δηλ. το πραγματικό φάρμακο δίνει αποτελέσματα και στις 3 ομάδες του πληθυσμού.

Η παραπάνω διένεξη προκύπτει από το γεγονός ότι στα δεδομένα υπάρχει **κρυμμένη ετερογένεια** (*hidden heterogeneity*) η οποία επηρεάζει τη διάρκεια ζωής. Δηλ. οι μονάδες του δείγματος έχουν διαφορετική **ευπάθεια** (*frailty*), με αποτέλεσμα οι πιο ευπαθείς μονάδες (π.χ. ηλικιωμένοι) να

πεθαίνουν νωρίτερα, αφήνοντας έτσι στο δείγμα τους λιγότερο ευπαθείς και επηρεάζοντας κατ' αυτό τον τρόπο τη συνάρτηση κινδύνου (βλ. [2] pp. 55).

Τα μοντέλα ευπάθειας (*frailty models*) δημιουργήθηκαν ακριβώς για να περιγράψουν τους συσχετισμούς και την κρυμμένη ετερογένεια στα δεδομένα διάρκειας ζωής (βλ. [2] – preface pp. XIX). Ο όρος **ευπάθεια** (*frailty*) εισήχθη το 1979 (βλ. [2] – preface pp. XX) από τους Vaupel et al. στο [46], προκειμένου να μελετηθούν τα σφάλματα που προέκυπταν από τα συνήθη μοντέλα επιβίωσης ιδίως στις μεγάλες ηλικίες και στην εργασία αυτή μελετήθηκε η επίδραση της ευπάθειας στην πιθανότητα επιβίωσης ενός ατόμου σε μία συγκεκριμένη ηλικία (βλ. [46] pp. 440-441).

Υπάρχουν δύο μεγάλες κατηγορίες μοντέλων ευπάθειας: τα **μονομεταβλητά μοντέλα ευπάθειας** (*univariate frailty models*) και τα **πολυμεταβλητά μοντέλα ευπάθειας** (*multivariate frailty models*). Η διαφορά τους έγκειται στο αν ο χρόνος ζωής είναι μονομεταβλητός ή πολυμεταβλητός (πολυμεταβλητό χρόνο έχουμε π.χ. στη μελέτη χρόνου ζωής σχετιζόμενων ατόμων, όπως ζευγαριών, ή στη μελέτη επαναλαμβανόμενων γεγονότων).

3.1.2. Η μονομεταβλητή ευπάθεια: Το μονομεταβλητό μοντέλο ευπάθειας επεκτείνει το μοντέλο αναλογικών κινδύνων του Cox, ώστε να περιλάβει μέσα σε αυτό την κρυμμένη ετερογένεια. Συγκεκριμένα, θεωρούμε την **ευπάθεια**, η οποία είναι μία θετική τυχαία μεταβλητή U που δρα πολλαπλασιαστικά στη βασική συνάρτηση κινδύνου $h_0(t)$. Έτσι, η συνάρτηση κινδύνου μίας μονάδας δοθείσης της ευπάθειας $U = u$ και δοθέντος του διανύσματος $\boldsymbol{z} = [z_1, z_2, \ldots, z_k]^{\mathrm{T}}$ των συμμεταβλητών, ανάγεται σε:

$$\boxed{h(t \mid u, \boldsymbol{z}) = u \cdot h_0(t) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}}} \quad \textbf{(3.1)}$$

όπου $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_k]^{\mathrm{T}}$ το διάνυσμα των συντελεστών.

Στην ουσία, η ευπάθεια εκπροσωπεί μία επιπλέον συμμεταβλητή z_{k+1} με αντίστοιχο συντελεστή β_{k+1}, τα οποία προστιθέμενα στο μοντέλο $h(t \mid \boldsymbol{z}) = h_0(t) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}}$ αναλογικών κινδύνων δίνουν

$h(t\,|\,\mathbf{z}) = h_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}+\beta_{k+1}z_{k+1}} = h_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}}\cdot e^{\beta_{k+1}z_{k+1}} = u\cdot h_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}}$, όπου $u = e^{\beta_{k+1}z_{k+1}}$.

Η τυχαία μεταβλητή της ευπάθειας οφείλει να είναι θετική επειδή η συνάρτηση κινδύνου είναι θετική και θεωρείται να έχει μέση τιμή $E[U]=1$, ενώ η διασπορά αυτής $\sigma^2 = V[U]$ (αν υπάρχει) είναι δείγμα της ετερογένειας, αφού μικρή τιμή της σ^2 δίνει τιμές της ευπάθειας κοντά στην μέση τιμή $E[U]=1$ και άρα συνάρτηση κινδύνου που δεν είναι εξαρτώμενη από την ευπάθεια (δηλ. μικρή τιμή της σ^2 υποδηλώνει ομοιογένεια στις μονάδες του δείγματος και άρα το μοντέλο του Cox είναι αρκετό για την περιγραφή του μοντέλου), το αντίθετο δε συμβαίνει για μεγάλες τιμές του σ^2 (βλ. [2] σελ. 57). Στα επόμενα, για απλούστευση θα θεωρούμε την ευπάθεια ως απολύτως συνεχή τυχαία μεταβλητή.

Από τη δεσμευμένη συνάρτηση κινδύνου **(3.1)** μπορούμε να υπολογίσουμε τη δεσμευμένη σωρευτική συνάρτηση κινδύνου:

$$H(t\,|\,u,\mathbf{z}) = \int_0^t h(s\,|\,\mathrm{u},\mathbf{z})ds = u\cdot H_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}} \quad \textbf{(3.2)}$$

και κατ' επέκταση τη δεσμευμένη συνάρτηση επιβίωσης:

$$S(t\,|\,u,\mathbf{z}) = e^{-H(t|\,\mathrm{u},\mathbf{z})} = e^{-u\cdot H_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}}} \quad \textbf{(3.3)}$$

- Έστω τώρα δείγμα δεδομένων διάρκειας ζωής μεγέθους n, στο οποίο προσαρμόζουμε το μοντέλο ευπάθειας. Επεκτείνοντας τους συμβολισμούς των προηγούμενων κεφαλαίων, τα δεδομένα μας είναι οι τριάδες $\mathrm{A} = \{(X_i, D_i, \mathrm{u}_i) : i = 1, 2, \ldots, \mathrm{n}\}$, όπου:
- X_i είναι ο χρόνος αποτυχίας ή λογοκρισίας, $i = 1, 2, \ldots, \mathrm{n}$
- $D_i = \begin{cases} 1 & \text{αν ο χρόνος } X_i \text{ είναι χρόνος διακοπής} \\ 0 & \text{αν ο χρόνος } X_i \text{ είναι χρόνος λογοκρισίας} \end{cases}$ είναι η δείκτρια συνάρτηση λογοκρισίας
- u_i είναι η τιμή της ευπάθειας για την $i-$μονάδα, $i = 1, 2, \ldots, \mathrm{n}$

Η συνάρτηση πιθανοφάνειας που προκύπτει από τα δεδομένα, κατ' αντιστοιχία με την εξίσωση **(2.35)** είναι:

$$L = \prod_{i=1}^{n} \left(h(X_i \mid \mathrm{u}_i, \boldsymbol{z}_i)\right)^{D_i} \cdot S(X_i \mid \mathrm{u}_i, \boldsymbol{z}_i) \Leftrightarrow$$

$$\Leftrightarrow L = \prod_{i=1}^{n} \left\{ \left(\mathrm{u}_i \cdot h_0(X_i) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i} \right)^{D_i} \cdot e^{-u_i \cdot H_0(X_i) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i}} \right\} \quad \textbf{(3.4)}$$

Προκειμένου να πάρουμε την πιθανοφάνεια **(3.4)** απαλλαγμένη από τις ευπάθειες u_i, μπορούμε να υπολογίσουμε τη μη δεσμευμένη από την ευπάθεια συνάρτηση επιβίωσης, αφού αυτή θα είναι

$$S(t \mid \boldsymbol{z}) = \int_0^{\infty} S(t \mid \mathrm{u}, \boldsymbol{z}) \cdot f_U(u) du \quad \textbf{(3.5)}$$

όπου $f_U(u)$ είναι η συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής U της ευπάθειας. Από την (3.5) υπολογίζονται οι μη δεσμευμένες από την ευπάθεια συναρτήσεις πυκνότητας πιθανότητας $f(t \mid \boldsymbol{z}) = -\frac{d}{dt} S(t \mid \boldsymbol{z})$ και κινδύνου $h(t \mid \boldsymbol{z}) = \frac{f(t \mid \boldsymbol{z})}{S(t \mid \boldsymbol{z})}$, άρα και η πιθανοφάνεια L χωρίς τις ευπάθειες.

3.1.3. Η ευπάθεια και ο μετασχηματισμός Laplace: (Πηγή: [2] σελ. 57-60) Ας θεωρήσουμε την απλούστερη περίπτωση ενός μοντέλου ευπάθειας χωρίς συμμεταβλητές. Τότε η δεσμευμένη συνάρτηση κινδύνου θα είναι

$$h(t \mid u) = u \cdot h_0(t) \quad \textbf{(3.6)}$$

όπου U η τυχαία μεταβλητή της ευπάθειας και $h_0(t)$ η βασική συνάρτηση κινδύνου. Η δεσμευμένη σωρευτική συνάρτηση κινδύνου θα είναι

$$H(t \mid U) = \int_0^t h(s \mid u) ds = u \cdot H_0(t) \quad \textbf{(3.7)}$$

όπου $H_0(t) = \int_0^t h_0(s) ds$ η βασική σωρευτική συνάρτηση κινδύνου. Επίσης, η δεσμευμένη συνάρτηση επιβίωσης είναι

$$S(t \mid u) = e^{-H(t \mid u)} = e^{-u \cdot H_0(t)} \quad \textbf{(3.8)}$$

Έτσι, αν $f_U(u)$ είναι η συνάρτηση πυκνότητας πιθανότητας της ευπάθειας, η μη δεσμευμένη συνάρτηση επιβίωσης είναι

$$S(t)=\int_0^{\infty} S(t\mid u)\cdot f_U(u)du \quad \textbf{(3.9)}$$

Αν αντικαταστήσουμε την **(3.8)** στην **(3.9)** έχουμε

$$S(t)=\int_0^{\infty} e^{-u\cdot H_0(t)}\cdot f_U(u)du \quad \textbf{(3.10)}$$

Παρατηρούμε ότι η εξίσωση **(3.10)** εκφράζει τη μέση τιμή της τυχαίας μεταβλητής $e^{-u\cdot H_0(t)}$, δηλ. ότι

$$S(t)=E\left[e^{-u\cdot H_0(t)}\right]$$

Επιπλέον, γνωρίζουμε ότι ο μετασχηματισμός Laplace (*Laplace transform*) μίας μη αρνητικής τυχαίας μεταβλητής X ορίζεται ως $\mathbf{L}_X(s)=\mathbf{L}(s)=E\left[e^{-sX}\right]=\int_0^{\infty} e^{-st}\cdot f_X(t)dt$ όπου $f_X(\cdot)$ η συνάρτηση πυκνότητας πιθανότητας της X. Έτσι, οδηγούμαστε στο ότι η μη δεσμευμένη συνάρτηση επιβίωσης είναι ο μετασχηματισμός Laplace της ευπάθειας, δηλ:

$$S(t)=\mathbf{L}_U\{H_0(t)\}=\mathbf{L}\{H_0(t)\} \quad \textbf{(3.11)}$$

(για εισαγωγικά περί του μετασχηματισμού Laplace, παραπέμπουμε στο [47]).

Οι παράγωγοι του μετασχηματισμού Laplace μπορούν να χρησιμοποιηθούν στο να εξάγουμε γενικά αποτελέσματα για τις μη δεσμευμένες συναρτήσεις πυκνότητας και κινδύνου, καθώς και για τη μέση τιμή και διασπορά (εφόσον υφίστανται) της ευπάθειας.

ΠΡΟΤΑΣΗ: Η μη δεσμευμένη συνάρτηση πυκνότητας πιθανότητας του χρόνου T είναι:

$$f(t)=-h_0(t)\cdot\mathbf{L}'\{H_0(t)\} \quad \textbf{(3.12)}$$

ΑΠΟΔΕΙΞΗ: Από την **(3.11)** έχουμε ισοδύναμα $e^{-H(t)}=\mathbf{L}\{H_0(t)\}$ και παραγωγίζοντας:

$$-h(t)e^{-H(t)}=\mathbf{L}'\{H_0(t)\}\cdot h_0(t)\Leftrightarrow h(t)S(t)=-h_0(t)\mathbf{L}'\{H_0(t)\}\Leftrightarrow f(t)=-h_0(t)\mathbf{L}'\{H_0(t)\}$$

ΠΡΟΤΑΣΗ: Η μη δεσμευμένη συνάρτηση κινδύνου είναι:

$$h(t)=-h_0(t)\cdot\frac{\mathbf{L}'\{H_0(t)\}}{\mathbf{L}\{H_0(t)\}} \quad \textbf{(3.13)}$$

ΑΠΟΔΕΙΞΗ: Λογαριθμίζουμε την **(3.11)** οπότε $H(t)=\ln \mathbf{L}\{H_0(t)\}$ και παραγωγίζοντας κατά μέλη παίρνουμε το ζητούμενο.

ΠΡΟΤΑΣΗ: Η μέση τιμή και η διασπορά της ευπάθειας (εφ' όσον υπάρχουν) είναι:

$$E[U]=-\mathbf{L}'(0) \quad \textbf{(3.14)}$$

$$V[U]=\mathbf{L}''(0)-(\mathbf{L}'(0))^2 \quad \textbf{(3.15)}$$

ΑΠΟΔΕΙΞΗ: $\mathbf{L}_U(s)=\mathbf{L}(s)=E\left[e^{-sU}\right]$ άρα $\mathbf{L}'(s)=E\left[-Ue^{-sU}\right]=-E\left[Ue^{-sU}\right]$ και για $s=0$ προκύπτει η **(3.14)**. Επίσης $\mathbf{L}''(s)=E\left[U^2e^{-sU}\right]$, άρα $\mathbf{L}''(0)=E\left[U^2\right]$ οπότε από τη γνωστή ιδιότητα $V[U]=E\left[U^2\right]-E^2[U]$ προκύπτει η **(3.15)**.

3.1.4. Η συνάρτηση κινδύνου του πληθυσμού και η αναμενόμενη ευπάθεια των επιζώντων: (Πηγές: [2] σελ. 60-61 και [48] σελ. 2-3) Για τη δεσμευμένη συνάρτηση κινδύνου του πληθυσμού έχουμε ότι:

$$h(t\,|\,u)=\frac{f(t\,|\,u)}{S(t\,|\,u)}\overset{(3.7)}{\Leftrightarrow} u\cdot h_0(t)=\frac{f(t\,|\,u)}{S(t\,|\,u)}\Leftrightarrow f(t\,|\,u)=u\cdot h_0(t)\cdot S(t\,|\,u)$$

Έτσι, εισάγοντας την από κοινού συνάρτηση πυκνότητας πιθανότητας των T, U, έχουμε:

$$\frac{f(t,u)}{f_U(u)}=u\cdot h_0(t)\cdot S(t\,|\,u),$$

άρα

$$f(t,u)=u\cdot h_0(t)\cdot S(t\,|\,u)\cdot f_U(u)$$

Η συνάρτηση πυκνότητας πιθανότητας της διάρκειας ζωής υπολογίζεται τώρα ως

$$f_T(t)=\int_0^\infty f(t,u)du=h_0(t)\cdot\int_0^\infty u\cdot S(t\,|\,u)\cdot f_U(u)du\,.$$

Η μη δεσμευμένη συνάρτηση κινδύνου του πληθυσμού είναι τώρα

$$h(t)=\frac{f_T(t)}{S(t)}=\frac{h_0(t)\cdot\int_0^\infty u\cdot S(t\,|\,u)\cdot f_U(u)du}{S(t)} \quad \textbf{(3.16)}$$

Η πυκνότητα της ευπάθειας των επιζώντων τη χρονική στιγμή t είναι

$$f(u \mid T > t) = \frac{P[T > t \mid u] \cdot f_U(u)}{P[T > t]} = \frac{S(t \mid u) \cdot f_U(u)}{S(t)} \quad \textbf{(3.17)}$$

άρα η αναμενόμενη ευπάθεια των επιζώντων είναι

$$E[U \mid T > t] = \int_0^\infty u \cdot f(u \mid T > t) du = \frac{1}{S(t)} \int_0^\infty u \cdot S(t \mid u) \cdot f_U(u) du \quad \textbf{(3.18)}$$

Συνδυάζοντας τις **(3.16)** και **(3.18)** παίρνουμε ότι

$$\boxed{h(t) = h_0(t) \cdot E[U \mid T > t]} \quad \textbf{(3.19)}$$

και αποδείξαμε ότι ***η μη δεσμευμένη συνάρτηση κινδύνου του πληθυσμού σε κάθε χρονική στιγμή, ισούται με τη βασική συνάρτηση κινδύνου του πληθυσμού πολλαπλασιασμένη με την αναμενόμενη ευπάθεια των επιζώντων εκείνη τη χρονική στιγμή***.

3.2. ΤΟ ΜΟΝΤΕΛΟ ΓΑΜΜΑ ΕΥΠΑΘΕΙΑΣ

(Πηγή: [2] σελ. 72-95)

3.2.1. Εισαγωγή στο μοντέλο Γάμμα ευπάθειας: Πρόκειται για το πιο διαδεδομένο μοντέλο ευπάθειας. Σε αυτό, η τυχαία μεταβλητή U της ευπάθειας ακολουθεί την Γάμμα κατανομή με παράμετρο κλίμακας λ και παράμετρο σχήματος κ δηλ. $U \sim G(\lambda, \kappa)$ και ως εκ τούτου, η U έχει συνάρτηση πυκνότητας πιθανότητας

$$f_U(u) = f(u) = \frac{\lambda^\kappa}{\Gamma(\kappa)} u^{\kappa-1} e^{-\lambda u} \quad \textbf{(3.20)}$$

Ο μετασχηματισμός Laplace της U είναι

$$\mathbf{L}_U(s) = \mathbf{L}(s) = E\left[e^{-sU}\right] = \int_0^\infty e^{-su} \cdot f_U(u) du = \frac{\lambda^\kappa}{\Gamma(\kappa)} \int_0^\infty u^{\kappa-1} e^{-(s+\lambda)u} du \overset{x=(s+\lambda)u}{=}$$

$$= \frac{\lambda^\kappa}{(s+\lambda)^\kappa \Gamma(\kappa)} \int_0^\infty x^{\kappa-1} e^{-x} dx = \frac{\lambda^\kappa}{(s+\lambda)^\kappa \Gamma(\kappa)} \cdot \Gamma(\kappa) = \frac{\lambda^\kappa}{(s+\lambda)^\kappa} = \left(\frac{s+\lambda}{\lambda}\right)^{-\kappa}$$

και τελικά

$$\mathbf{L}(s) = \left(1 + \frac{s}{\lambda}\right)^{-\kappa} \quad \textbf{(3.21)}$$

Η αναμενόμενη τιμή και η διασπορά της ευπάθειας προκύπτουν από τις παραγώγους του μετασχηματισμού Laplace στη θέση $s=0$. Έτσι, είναι:

$\mathbf{L}'(s)=-\frac{\kappa}{\lambda}\left(1+\frac{s}{\lambda}\right)^{-\kappa-1}$ και $\mathbf{L}''(s)=\frac{\kappa\cdot(\kappa+1)}{\lambda^2}\left(1+\frac{s}{\lambda}\right)^{-\kappa-2}$. Εφαρμόζοντας τα αποτελέσματα αυτά στις εξισώσεις **(3.14)** και **(3.15)**, παίρνουμε $E[U]=\frac{\kappa}{\lambda}$ και $V[U]=\frac{\kappa}{\lambda^2}$. Λόγω του ότι για την ευπάθεια θέλουμε αναμενόμενη τιμή $E[U]=1$, επιλέγουμε τις παραμέτρους της να είναι $\kappa=\lambda=\frac{1}{\sigma^2}$ δηλ. $U\sim G\left(\frac{1}{\sigma^2},\frac{1}{\sigma^2}\right)$. Τότε από τις προηγούμενες εξισώσεις προκύπτει ως συνάρτηση πυκνότητας πιθανότητας της ευπάθειας η $f_U(u)=f(u)=\frac{\left(\frac{1}{\sigma^2}\right)^{\frac{1}{\sigma^2}}}{\Gamma\left(\frac{1}{\sigma^2}\right)}u^{\frac{1}{\sigma^2}-1}e^{-\frac{u}{\sigma^2}}$, ενώ η αναμενόμενη τιμή και η διασπορά:

$$E[U]=1 \quad \textbf{(αναμενόμενη τιμή της Γάμμα ευπάθειας) (3.22)}$$

και

$$V[U]=\sigma^2 \quad \textbf{(διασπορά της Γάμμα ευπάθειας) (3.23)}$$

Από την εξίσωση **(3.11)** και για $\kappa=\lambda=\frac{1}{\sigma^2}$, παίρνουμε τη μη δεσμευμένη συνάρτηση επιβίωσης:

$$\boxed{S(t)=\mathbf{L}\{H_0(t)\}=\frac{1}{\left(1+\sigma^2H_0(t)\right)^{\frac{1}{\sigma^2}}}} \quad \textbf{(3.24)}$$

Η συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής της διάρκειας ζωής είναι $f(t)=-S'(t)=\frac{1}{\sigma^2}\left(\frac{1}{1+\sigma^2H_0(t)}\right)^{\frac{1}{\sigma^2}-1}\cdot\frac{\sigma^2h_0(t)}{\left(1+\sigma^2H_0(t)\right)^2}$ άρα

$$\boxed{f(t)=\frac{h_0(t)}{\left(1+\sigma^2H_0(t)\right)^{\frac{1}{\sigma^2}+1}}} \quad \textbf{(3.25)}$$

Από τη διαίρεση των **(3.25)** και **(3.24)** προκύπτει η συνάρτηση κινδύνου:

$$\boxed{h(t)=\frac{h_0(t)}{1+\sigma^2 H_0(t)}} \quad \textbf{(3.26)}$$

3.2.2. Το μοντέλο του Cox και η Γάμμα ευπάθεια: Εισάγουμε στο μοντέλο του Cox την ευπάθεια $U \sim G\left(\frac{1}{\sigma^2},\frac{1}{\sigma^2}\right)$. Τότε η δεσμευμένη συνάρτηση κινδύνου περιγράφεται από την εξίσωση **(3.1)**. Η δεσμευμένη συνάρτηση πυκνότητας πιθανότητας της ευπάθειας των ατόμων που πεθαίνουν τη χρονική στιγμή t, δοθέντων των συμμεταβλητών $\mathbf{z}=[z_1,z_2,\ldots,z_k]^{\mathrm{T}}$ είναι

$$f_U(u\mid T=t,\mathbf{z})=\frac{f_T(t\mid \mathrm{u},\mathbf{z})\cdot f_U(u)}{f_T(t\mid \mathbf{z})} \quad \textbf{(3.27)}$$

Όμως

$$h(t\mid u,\mathbf{z})=u\cdot h_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}} \quad \textbf{(3.28)}$$

άρα

$$S(t\mid u,\mathbf{z})=e^{-u\cdot H_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}}} \quad \textbf{(3.29)}$$

και παραγωγίζοντας

$$f(t\mid u,\mathbf{z})=-S'(t\mid u,\mathbf{z})=u\cdot h_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}}\cdot e^{-u\cdot H_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}}} \quad \textbf{(3.30)}$$

Επίσης, από την **(3.25)** είναι

$$f(t\mid \mathbf{z})=\frac{h_0(t\mid \mathbf{z})}{\left(1+\sigma^2 H_0(t\mid \mathbf{z})\right)^{\frac{1}{\sigma^2}+1}}=\frac{h_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}}}{\left(1+\sigma^2 H_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}}\right)^{\frac{1}{\sigma^2}+1}} \quad \textbf{(3.31)}$$

Αντικαθιστούμε στην **(3.27)** άρα

$$f_U(u\mid T=t,\mathbf{z})=\frac{u\cdot h_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}}\cdot e^{-u\cdot H_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}}}}{\dfrac{h_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}}}{\left(1+\sigma^2 H_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}}\right)^{\frac{1}{\sigma^2}+1}}}\cdot\frac{\left(\frac{1}{\sigma^2}\right)^{\frac{1}{\sigma^2}}}{\Gamma\left(\frac{1}{\sigma^2}\right)}u^{\frac{1}{\sigma^2}-1}e^{-\frac{u}{\sigma^2}} \Leftrightarrow$$

$$\Leftrightarrow f_U(u \mid T=t, \boldsymbol{z}) = \left(1+\sigma^2 H_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}}\right)^{\frac{1}{\sigma^2}+1} e^{-u\cdot H_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}}} \cdot \frac{\left(\frac{1}{\sigma^2}\right)^{\frac{1}{\sigma^2}+1}}{\frac{1}{\sigma^2}\Gamma\left(\frac{1}{\sigma^2}\right)} u^{\frac{1}{\sigma^2}} e^{-\frac{u}{\sigma^2}} \Leftrightarrow$$

$$\boxed{f_U(u \mid T=t, \boldsymbol{z}) = \frac{\left(\frac{1}{\sigma^2}+H_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}}\right)^{\frac{1}{\sigma^2}+1}}{\Gamma\left(\frac{1}{\sigma^2}+1\right)} u^{\frac{1}{\sigma^2}} e^{-u\left(\frac{1}{\sigma^2}+H_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}}\right)}} \quad \textbf{(3.32)}$$

δηλ. η ευπάθεια των ατόμων που πεθαίνουν τη χρονική στιγμή t, δοθέντων των συμμεταβλητών $\boldsymbol{z}$, ακολουθεί επίσης κατανομή Γάμμα με παράμετρο κλίμακας $\lambda_1 = \frac{1}{\sigma^2} + H_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}}$ και παράμετρο σχήματος $\kappa_1 = \frac{1}{\sigma^2}+1$.

Από την άλλη, η πυκνότητα της ευπάθειας των επιζώντων τη χρονική στιγμή t, δοθέντων των συμμεταβλητών $\boldsymbol{z} = [z_1, z_2, \ldots, z_k]^{\mathrm{T}}$ είναι

$$f_U(u \mid T>t, \boldsymbol{z}) = \frac{P[T>\mathrm{t} \mid u, \boldsymbol{z}]\cdot f_U(u)}{P[T>\mathrm{t} \mid \boldsymbol{z}]} = \frac{S(\mathrm{t} \mid u, \boldsymbol{z})\cdot f_U(u)}{S(\mathrm{t} \mid \boldsymbol{z})} \quad \textbf{(3.33)}$$

Χρησιμοποιώντας τις **(3.29)** και **(3.24)** έχουμε διαδοχικά:

$$f_U(u \mid T>t, \boldsymbol{z}) = \frac{e^{-u\cdot H_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}}}}{\frac{1}{\left(1+\sigma^2 H_0(t)\right)^{\frac{1}{\sigma^2}}}} \cdot \frac{\left(\frac{1}{\sigma^2}\right)^{\frac{1}{\sigma^2}}}{\Gamma\left(\frac{1}{\sigma^2}\right)} u^{\frac{1}{\sigma^2}-1} e^{-\frac{u}{\sigma^2}} \Leftrightarrow$$

$$\boxed{f_U(u \mid T>t, \boldsymbol{z}) = \frac{\left(\frac{1}{\sigma^2}+H_0(t)\right)^{\frac{1}{\sigma^2}}}{\Gamma\left(\frac{1}{\sigma^2}\right)} u^{\frac{1}{\sigma^2}-1} e^{-u\left(\frac{1}{\sigma^2}+H_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}}\right)}} \quad \textbf{(3.34)}$$

άρα η ευπάθεια των επιζώντων τη χρονική στιγμή t, δοθέντων των συμμεταβλητών $\boldsymbol{z}$, ακολουθεί και αυτή κατανομή Γάμμα με παράμετρο κλίμακας $\lambda_2 = \frac{1}{\sigma^2} + H_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}}$ και παράμετρο σχήματος $\kappa_2 = \frac{1}{\sigma^2}$.

Από τα συμπεράσματα των εξισώσεων **(3.32)** και **(3.34)** μπορούμε να υπολογίσουμε τις αναμενόμενες ευπάθειες των ατόμων που πεθαίνουν τη

χρονική στιγμή t και των επιζώντων την ίδια χρονική στιγμή:

$E[u \mid T = t, \mathbf{z}] = \frac{\kappa_1}{\lambda_1} = \frac{\frac{1}{\sigma^2} + 1}{\frac{1}{\sigma^2} + H_0(t) \cdot e^{\boldsymbol{\beta}^T \mathbf{z}}}$, άρα:

$$\boxed{E[u \mid T = t, \mathbf{z}] = \frac{1 + \sigma^2}{1 + \sigma^2 H_0(t) \cdot e^{\boldsymbol{\beta}^T \mathbf{z}}}} \quad \textbf{(3.35)}$$

και $E[u \mid T > t, \mathbf{z}] = \frac{\kappa_2}{\lambda_2} = \frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma^2} + H_0(t) \cdot e^{\boldsymbol{\beta}^T \mathbf{z}}}$, άρα:

$$\boxed{E[u \mid T > t, \mathbf{z}] = \frac{1}{1 + \sigma^2 H_0(t) \cdot e^{\boldsymbol{\beta}^T \mathbf{z}}}} \quad \textbf{(3.36)}$$

απ' όπου προκύπτει ότι $E[u \mid T = t, \mathbf{z}] > E[u \mid T > t, \mathbf{z}]$ δηλ. **οι μονάδες που διακόπτουν τη λειτουργία τους σε μία χρονική στιγμή έχουνε μεγαλύτερη αναμενόμενη ευπάθεια από αυτές που συνεχίζουν να ζουν την ίδια χρονική στιγμή**.

Αντίστοιχα για τις διασπορές των παραπάνω ομάδων έχουμε ότι

$V[u \mid T = t, \mathbf{z}] = \frac{\kappa_1}{\lambda_1^2} = \frac{\frac{1}{\sigma^2} + 1}{\left(\frac{1}{\sigma^2} + H_0(t) \cdot e^{\boldsymbol{\beta}^T \mathbf{z}}\right)^2}$ άρα

$$\boxed{V[u \mid T = t, \mathbf{z}] = \frac{\sigma^2\left(1 + \sigma^2\right)}{\left(1 + \sigma^2 H_0(t) \cdot e^{\boldsymbol{\beta}^T \mathbf{z}}\right)^2}} \quad \textbf{(3.37)}$$

και όμοια

$$\boxed{V[u \mid T > t, \mathbf{z}] = \frac{\sigma^2}{\left(1 + \sigma^2 H_0(t) \cdot e^{\boldsymbol{\beta}^T \mathbf{z}}\right)^2}} \quad \textbf{(3.38)}$$

απ' όπου προκύπτει ότι **η διασπορά της ευπάθειας μειώνεται κατά τη διάρκεια του χρόνου και έτσι, ο πληθυσμός τείνει στο πέρασμα του χρόνου να γίνει πιο ομοιογενής (όσον αφορά την ευπάθεια)**.

Τέλος, οι συντελεστές μεταβλητότητας των δύο ομάδων είναι

$$CV\left[u \mid T=t, \boldsymbol{z}\right]=\frac{\sqrt{V\left[u \mid T=t, \boldsymbol{z}\right]}}{E\left[u \mid T=t, \boldsymbol{z}\right]}=\frac{\sigma\sqrt{1+\sigma^2}}{1+\sigma^2}=\frac{\sigma}{\sqrt{1+\sigma^2}} \quad \textbf{(3.39)}$$

και

$$CV\left[u \mid T>t, \boldsymbol{z}\right]=\frac{\sqrt{V\left[u \mid T>t, \boldsymbol{z}\right]}}{E\left[u \mid T>t, \boldsymbol{z}\right]}=\sigma \quad \textbf{(3.40)}$$

δηλ. παραμένουν σταθεροί στο πέρασμα του χρόνου και έτσι η ευπάθεια έχει σταθερή ομοιογένεια ως προς τον μέσο όρο της.

3.2.3. Εκτίμηση παραμέτρων στο παραμετρικό μοντέλο Γάμμα ευπάθειας: Στην περίπτωση που προσαρμόζουμε σε δείγμα $\mathrm{A}=\left\{\left(X_i, D_i, \mathrm{u}_i\right): i=1,2,\ldots,n\right\}$ μεγέθους n (όπου για την $i-$μονάδα είναι όπως πάντα $\left(X_i, D_i\right)$ το ζεύγος χρόνου αποτυχίας ή λογοκρισίας και η δείκτρια συνάρτηση λογοκρισίας, $i=1,2,\ldots,\mathrm{n}$) ένα παραμετρικό μοντέλο ευπάθειας, με διάνυσμα παραμέτρων της βασικής συνάρτησης κινδύνου έστω $\boldsymbol{\theta}$, διάνυσμα συντελεστών των συμμεταβλητών $\boldsymbol{\beta}$, και τιμή της ευπάθειας για την $i-$μονάδα του δείγματος την u_i, $i=1,2,\ldots,n$, η συνάρτηση πιθανοφάνειας προκύπτει από την **(3.4)** και είναι:

$$L\left(\boldsymbol{\beta}, \boldsymbol{\theta} \mid u_1, u_2, \ldots, u_n\right)=\prod_{i=1}^{n}\left(h\left(X_i \mid \mathrm{u}_i, \boldsymbol{z}_i\right)\right)^{D_i} \cdot S\left(X_i \mid \mathrm{u}_i, \boldsymbol{z}_i\right)$$

και με τη βοήθεια των **(3.24)**, **(3.26)**, **(3.28)** και **(3.29)** παίρνουμε τη συνάρτηση πιθανοφάνειας για το παραμετρικό μοντέλο Γάμμα ευπάθειας:

$$\boxed{L\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2\right)=\prod_{i=1}^{n}\left(\frac{h_0\left(X_i, \boldsymbol{\theta}\right) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}}}{1+\sigma^2 H_0\left(X_i, \boldsymbol{\theta}\right) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}}}\right)^{D_i} \cdot\left(1+\sigma^2 H_0\left(X_i, \boldsymbol{\theta}\right) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}}\right)^{-\frac{1}{\sigma^2}}} \quad \textbf{(3.41)}$$

Η μεγιστοποίηση της συνάρτησης $\ell\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2\right)=\ln L\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2\right)$ μας δίνει τη ζητούμενη εκτίμηση των παραμέτρων.

3.2.4. Εκτίμηση παραμέτρων στο ημι-παραμετρικό μοντέλο Γάμμα ευπάθειας: Στην αντίστοιχη περίπτωση της προσαρμογής ημι-παραμετρικού μοντέλου Γάμμα ευπάθειας, όπου δεν υπάρχει παραμετρική μορφή της

συνάρτησης κινδύνου, θεωρούμε την από κοινού συνάρτηση πυκνότητας πιθανότητας f των $(X_i, D_i, \mathrm{u}_i): i = 1, 2, \ldots, \mathrm{n}$. Η πιθανοφάνεια γράφεται αντίστοιχα ως:

$$L(\boldsymbol{\beta}, \sigma^2 \mid u_1, u_2, \ldots, u_n) = \prod_{i=1}^{n} f(X_i, D_i, \mathrm{u}_i; \boldsymbol{\beta}, \sigma^2) \Leftrightarrow$$

$$\Leftrightarrow L(\boldsymbol{\beta}, \sigma^2 \mid u_1, u_2, \ldots, u_n) = \prod_{i=1}^{n} f(X_i, D_i, \mathrm{u}_i; \boldsymbol{\beta}) \cdot \prod_{i=1}^{n} f(\mathrm{u}_i; \sigma^2) \Leftrightarrow$$

$$\Leftrightarrow L(\boldsymbol{\beta}, \sigma^2 \mid u_1, u_2, \ldots, u_n) = L_1(\boldsymbol{\beta} \mid u_1, u_2, \ldots, u_n) \cdot L_2(\sigma^2 \mid u_1, u_2, \ldots, u_n)$$

όπου η $L_1(\boldsymbol{\beta} \mid u_1, u_2, \ldots, u_n)$ προκύπτει από την εξίσωση **(3.4)** και είναι ίση με

$$L_1(\boldsymbol{\beta} \mid u_1, u_2, \ldots, u_n) = \prod_{i=1}^{n} \left\{ \left(\mathrm{u}_i \cdot h_0(X_i) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i} \right)^{D_i} \cdot e^{-u_i \cdot H_0(X_i) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i}} \right\} \quad \textbf{(3.42)}$$

και η

$$L_2(\sigma^2 \mid u_1, u_2, \ldots, u_n) = \prod_{i=1}^{n} f(\mathrm{u}_i; \sigma^2) \quad \textbf{(3.43)}$$

είναι το γινόμενο των τιμών της συνάρτησης πυκνότητας πιθανότητας της ευπάθειας στα σημεία u_i, $i = 1, 2, \ldots, \mathrm{n}$.

Ο υπολογισμός της L_1 στην εξίσωση (3.42) γίνεται αν επεκτείνουμε την ιδέα της μερικής πιθανοφάνειας του μοντέλου του Cox. Έτσι, αν $t_{(1)} < t_{(2)} < \ldots < t_{(m)}$ οι διατεταγμένοι διακεκριμένοι χρόνοι αποτυχίας (χρόνοι θανάτου) μονάδων του δείγματος (δηλ. X_j, $j \in \{1, 2, \ldots, n\}$ με $D_j = 1$), και με τη βοήθεια της εξίσωσης **(2.15)**: $L = L(\boldsymbol{\beta}) = \prod_{j=1}^{m} \frac{e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_j}}{\sum_{i \in \Re_{(j)}} e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i}}$ που δίνει τη μερική πιθανοφάνεια στο μοντέλο του Cox, η εξίσωση **(3.42)** γίνεται:

$$L_1(\boldsymbol{\beta} \mid u_1, u_2, \ldots, u_n) = \prod_{j=1}^{m} \frac{u_j \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_j}}{\sum_{i \in \Re_{(j)}} u_i \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i}} \quad \textbf{(3.44)}$$

Στην εξίσωση **(3.44)**, η άγνωστη ευπάθεια u_j οποιασδήποτε μονάδας $j \in \{1, 2, \ldots, n\}$ του δείγματος, μπορεί να εκτιμηθεί από την αναμενόμενη τιμή της. Με τη βοήθεια των εξισώσεων **(3.35)** και **(3.36)**, η εκτίμηση είναι:

$$\hat{u}_j = \frac{\frac{1}{\hat{\sigma}^2} + D_j}{\frac{1}{\hat{\sigma}^2} + \hat{H}_0(t) \cdot e^{\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{z}_j}} \text{ για } j \in \{1, 2, \ldots, n\} \quad \textbf{(3.45)}$$

όπου $\hat{\sigma}^2$ είναι μη παραμετρική εκτίμηση της διασποράς της ευπάθειας και $\hat{H}_0(t)$ είναι μη παραμετρική εκτίμηση της σωρευτικής συνάρτησης κινδύνου (π.χ. εκτιμήτρια Nelson-Aalen).

3.3. ΤΟ ΜΟΝΤΕΛΟ ΛΟΓΑΡΙΘΜΟΚΑΝΟΝΙΚΗΣ ΕΥΠΑΘΕΙΑΣ

(Πηγή: [2] σελ. 96-100)

3.3.1. Εισαγωγή στο μοντέλο Λογαριθμοκανονικής ευπάθειας: Εδώ, ακολουθώντας τη θεωρία της Λογαριθμοκανονικής κατανομής (βλ. παράγραφο **1.3.5**), θεωρούμε μία τυχαία μεταβλητή $W \sim N\left(\mu_w, \sigma_w^2\right)$ και ορίζουμε την ευπάθεια ως $U = e^W$. Τότε η ευπάθεια ακολουθεί τη Λογαριθμοκανονική κατανομή με μέση τιμή και διασπορά (βλ. παράγραφο **1.3.5**) τις $E[U] = e^{\mu_w + \frac{\sigma_w^2}{2}}$ και $V[U] = e^{2\mu_w + \sigma_w^2} \cdot \left(e^{\sigma_w^2} - 1\right)$ αντίστοιχα.

Προκειμένου να επιτύχουμε τον περιορισμό $E[U] = 1$, από τις παραπάνω σχέσεις παίρνουμε ότι $\mu_w = -\frac{\sigma_w^2}{2}$ και $\sigma_w^2 = \ln\left(1 + \sigma^2\right)$. Ωστόσο, μία συνήθης επιλογή είναι η $\mu_w = 0$ και $\sigma_w^2 = \sigma^2$, έτσι ώστε $W = \ln U \sim N\left(0, \sigma^2\right)$. Σ' αυτή την περίπτωση, η συνάρτηση πυκνότητας πιθανότητας της ευπάθειας είναι:

$$f_U(u) = f(u) = \frac{1}{\sigma u \sqrt{2\pi}} e^{-\frac{\ln^2 u}{2\sigma^2}} \quad \textbf{(3.46)}$$

3.3.2. Εκτίμηση παραμέτρων στο παραμετρικό μοντέλο Λογαριθμοκανονικής ευπάθειας: Το μειονέκτημα της Λογαριθμοκανονικής ευπάθειας είναι ότι ο μετασχηματισμός Laplace

$$\mathbf{L}_U(s) = \mathbf{L}(s) = E\left[e^{-sU}\right] = \int_0^\infty e^{-su} \cdot f_U(u)du = \frac{1}{\sigma\sqrt{2\pi}} \int_0^\infty \frac{e^{-su-\frac{\ln^2 u}{2\sigma^2}}}{u} du$$

εκφράζεται μόνο υπό μορφή ολοκληρώματος (όχι κλειστή μορφή) με αποτέλεσμα το ίδιο να συμβαίνει με τις μη δεσμευμένες συναρτήσεις επιβίωσης και κινδύνου. Αποτέλεσμα τούτου, είναι η συνάρτηση πιθανοφάνειας για δεδομένα (X_i, D_i, u_i), $i = 1, 2, \ldots, n$ να είναι:

$$L\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2\right) = \prod_{i=1}^{n} \int \left(h_0\left(X_i; \boldsymbol{\theta}\right) \cdot e^{w_i + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i} \right)^{D_i} \cdot \cdot e^{-H_0\left(X_i; \boldsymbol{\theta}\right) \cdot e^{w_i + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i}} d\Phi\left(w_i\right) \quad \textbf{(3.47)}$$

όπου $w_i = \ln u_i$ και $\Phi(\cdot)$ η αθροιστική συνάρτηση κατανομής της $N\left(0, \sigma^2\right)$ και η μεγιστοποίησή της να εναπόκειται είτε σε αριθμητικές μεθόδους είτε σε μεθόδους όπως π.χ. οι Monte Carlo αλυσίδες Markov (*Monte Carlo Markov Chains – MCMC*).

3.3.3. Εκτίμηση παραμέτρων στο ημι-παραμετρικό μοντέλο Λογαριθμοκανονικής ευπάθειας: Για το ημι-παραμετρικό μοντέλο, χρησιμοποιείται η μέθοδος της επί ποινή μερικής πιθανοφάνειας (*penalized partial likelihood - PPL*). Σε αυτήν, θεωρούμε την τυχαία μεταβλητή $W = \ln U$ του λογαρίθμου της ευπάθειας U (οπότε $W = \ln U \sim N\left(0, \sigma^2\right)$) και γράφουμε τη βασική εξίσωση $h(t \mid u, \boldsymbol{z}) = u \cdot h_0(t) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}}$ στην ισοδύναμη μορφή $h(t \mid w, \boldsymbol{z}) = h_0(t) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z} + w}$. Για τα δεδομένα (X_i, D_i, u_i), $i = 1, 2, \ldots, n$, θεωρούμε ως επιπλέον παραμέτρους προς εκτίμηση τις $w_i = \ln u_i$, $i = 1, 2, \ldots, n$ και τότε, η (επί ποινή) συνάρτηση μερικής πιθανοφάνειας είναι:

$$L_{ppl}\left(\boldsymbol{\beta}, \sigma^2 \mid w_1, w_2, \ldots, w_n\right) = \prod_{i=1}^{n} \left(\frac{e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i + w_i}}{\sum_{j \in \Re_{(i)}} e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_j + w_j}} \right)^{D_i} \cdot f_W\left(w_i; \sigma^2\right)$$

όπου f_W είναι η συνάρτηση πυκνότητας της τυχαίας μεταβλητής W. Λογαριθμίζοντας, παίρνουμε:

$$\ell_{ppl}\left(\boldsymbol{\beta}, \sigma^2 \mid w_1, w_2, \ldots, w_n\right) = \sum_{i=1}^{n} D_i \left\{ \boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i + w_i - \ln\left(\sum_{j \in \Re_{(i)}} e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_j + w_j} \right) \right\} + \sum_{i=1}^{n} \ln f_W\left(w_i; \sigma^2\right)$$

Το πρώτο άθροισμα αφορά το κομμάτι της μερικής πιθανοφάνειας και είναι ίδιο για οποιοδήποτε μοντέλο ευπάθειας επιλέξουμε. Συμβολίζουμε:

$$\boxed{\ell_{partial}\left(\boldsymbol{\beta} \mid w_1, w_2, \ldots, w_n\right) = \sum_{i=1}^{n} D_i \left\{ \boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i + w_i - \ln\left(\sum_{j \in \Re_{(i)}} e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_j + w_j} \right) \right\}} \quad \textbf{(3.48)}$$

Το δεύτερο άθροισμα είναι ο παράγοντας της ποινής (που οφείλεται στη διασπορά σ^2 της ευπάθειας). Συμβολίζουμε:

$$\boxed{\ell_{pen}\left(\sigma^2 \mid w_1, w_2, \ldots, w_n\right) = \sum_{i=1}^{n} \ln f_W\left(w_i; \sigma^2\right)} \quad \textbf{(3.49)}$$

Στο μοντέλο της Λογαριθμοκανονικής ευπάθειας, επειδή $W = \ln U \sim N\left(0, \sigma^2\right)$,

άρα $f_W\left(w_i\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{w_i^2}{2\sigma^2}}$ και άρα η **(3.49)** για δίνει

$$\boxed{\ell_{pen}\left(\sigma^2 \mid w_1, w_2, \ldots, w_n\right) = -n \ln\left(\sigma\sqrt{2\pi}\right) - \frac{1}{2\sigma^2} \sum_{j=1}^{n} w_j^2} \quad \textbf{(3.50)}$$

Τελικά:

$$\boxed{\ell_{ppl}\left(\boldsymbol{\beta}, \sigma^2 \mid w_1, \ldots, w_n\right) = \ell_{partial}\left(\boldsymbol{\beta} \mid w_1, \ldots, w_n\right) + \ell_{pen}\left(\sigma^2 \mid w_1, \ldots, w_n\right)} \quad \textbf{(3.51)}$$

και η μεγιστοποίηση της $\ell_{ppl}\left(\boldsymbol{\beta}, \sigma^2 \mid w_1, \ldots, w_n\right)$ γίνεται με αριθμητικές μεθόδους.

3.4. ΤΟ ΜΟΝΤΕΛΟ ΑΝΤΙΣΤΡΟΦΗΣ ΓΚΑΟΥΣΙΑΝΗΣ ΕΥΠΑΘΕΙΑΣ

(Πηγή: [2] σελ. 101-104)

3.4.1. Εισαγωγή στο μοντέλο Αντίστροφης Γκαουσιανής ευπάθειας: Αν η τυχαία μεταβλητή U της ευπάθειας ακολουθεί την Αντίστροφη Γκαουσιανή κατανομή (*Inverse Gaussian*) τότε έχει συνάρτηση πυκνότητας πιθανότητας

$$f_U(u) = f(u) = \sqrt{\frac{\lambda}{2\pi u^3}} \cdot e^{-\frac{\lambda(u-\mu)^2}{2\mu^2 u}}\ ,\ t>0\ ,\ \mu>0\ ,\ \lambda>0\ .\ \textbf{(3.52)}$$

Ο μετασχηματισμός Laplace της ευπάθειας είναι

$$\mathbf{L}_U(s) = \mathbf{L}(s) = E\left[e^{-sU}\right] = \int_0^\infty e^{-su}\cdot f_U(u)du = \int_0^\infty \sqrt{\frac{\lambda}{2\pi u^3}}\cdot e^{-su-\frac{\lambda(u-\mu)^2}{2\mu^2 u}}\,du\ .$$

Η ποσότητα που εμφανίζεται στον εκθέτη μέσα στο ολοκλήρωμα, μετασχηματίζεται σε

$$-su-\frac{\lambda(u-\mu)^2}{2\mu^2 u} = -\frac{\left(2\mu^2 s+\lambda\right)u^2-2\lambda\mu u+\lambda\mu^2}{2\mu^2 u} = -\frac{\left(2\mu^2 s+\lambda\right)u}{2\mu^2}+\frac{\lambda}{\mu}-\frac{\lambda}{2u} =$$

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

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

και άρα έχουμε

$$\mathbf{L}(s) = \exp\left\{-\frac{\lambda\sqrt{\frac{2\mu^2 s}{\lambda}+1}}{\mu}+\frac{\lambda}{\mu}\right\}\cdot\int_0^\infty\sqrt{\frac{\lambda}{2\pi u^3}}\cdot\exp\left\{\frac{\lambda\sqrt{\frac{2\mu^2 s}{\lambda}+1}}{\mu}-\frac{\lambda\left(\frac{2\mu^2 s}{\lambda}+1\right)u}{2\mu^2}-\frac{\lambda}{2u}\right\}du =$$

Επίσης είναι:

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

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

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

Θέτουμε $M=\frac{\mu}{\sqrt{\frac{2\mu^2 s}{\lambda}+1}}$ άρα ο μετασχηματισμός Laplace της ευπάθειας γίνεται:

$$\mathbf{L}(s)=\exp\left\{-\frac{\lambda\sqrt{\frac{2\mu^2 s}{\lambda}+1}}{\mu}+\frac{\lambda}{\mu}\right\}\cdot\int_0^\infty\sqrt{\frac{\lambda}{2\pi u^3}}\cdot\exp\left\{-\frac{\lambda(u-M)^2}{2M^2 u}\right\}du$$

Όμως $\int_0^\infty\sqrt{\frac{\lambda}{2\pi u^3}}\cdot\exp\left\{-\frac{\lambda(u-M)^2}{2M^2 u}\right\}du=1$ ως ολοκλήρωμα της συνάρτησης πυκνότητας πιθανότητας της Inverse Gaussian κατανομής με παραμέτρους λ και $M=\frac{\mu}{\sqrt{\frac{2\mu^2 s}{\lambda}+1}}$. Τελικά:

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

Παραγωγίζουμε δύο φορές:

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

και

$$\mathbf{L}''(s)=\frac{\mu^3}{\lambda\left(\sqrt{\frac{2\mu^2 s}{\lambda}+1}\right)^3}\exp\left\{-\frac{\lambda\sqrt{\frac{2\mu^2 s}{\lambda}+1}}{\mu}+\frac{\lambda}{\mu}\right\}+\frac{\mu^2}{\frac{2\mu^2 s}{\lambda}+1}\exp\left\{-\frac{\lambda\sqrt{\frac{2\mu^2 s}{\lambda}+1}}{\mu}+\frac{\lambda}{\mu}\right\}$$

και για $s=0$, έχουμε $\mathbf{L}'(0)=-\mu$ και $\mathbf{L}''(0)=\frac{\mu^3}{\lambda}+\mu^2$, άρα η αναμενόμενη τιμή και η διασπορά της ευπάθειας είναι:

$$E[U]=-\mathbf{L}'(0)=\mu$$

και

$$V[U]=\mathbf{L}''(0)-(\mathbf{L}'(0))^2=\frac{\mu^3}{\lambda}$$

Επιλέγουμε $\mu=1$ και $\sigma^2=\frac{1}{\lambda}$, προκειμένου να έχουμε $E[U]=1$ και $V[U]=\sigma^2$ και καταλήγουμε στη μορφή:

$$\boxed{\mathbf{L}(s)=\exp\left\{\frac{1}{\sigma^2}\left(1-\sqrt{2\sigma^2 s+1}\right)\right\}} \textbf{ (3.53)}$$

απ' όπου προκύπτουν οι μη δεσμευμένες συναρτήσεις επιβίωσης και κινδύνου

$$\boxed{S(t)=\exp\left\{\frac{1}{\sigma^2}\left(1-\sqrt{2\sigma^2 H_0(t)+1}\right)\right\}} \textbf{ (3.54)} \text{ και}$$

$$\boxed{h(t)=\frac{h_0(t)}{\sqrt{2\sigma^2 H_0(t)+1}}} \textbf{ (3.55)}$$

Εργαζόμενοι τώρα όπως και στην περίπτωση της Γάμμα ευπάθειας, μπορούμε να υπολογίσουμε την αναμενόμενη ευπάθεια των επιζώντων σε μία χρονική στιγμή t στο μοντέλο του Cox και τη διασπορά της:

$$\boxed{E[u\mid T>t,\mathbf{z}]=\frac{1}{\sqrt{\sigma^2 H_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}}+1}}} \textbf{ (3.56)}$$

$$\boxed{V[u\mid T>t,\mathbf{z}]=\frac{\sigma^2}{\left(\sigma^2 H_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}}+1\right)^2}} \textbf{ (3.57)}$$

3.5. ΤΟ ΜΟΝΤΕΛΟ ΟΜΟΙΟΜΟΡΦΗΣ ΕΥΠΑΘΕΙΑΣ

(Πηγή: [49])

3.5.1. Εισαγωγή στο μοντέλο Ομοιόμορφης ευπάθειας: Εδώ, η τυχαία μεταβλητή U της ευπάθειας ακολουθεί την Ομοιόμορφη κατανομή σε διάστημα $[a,b]$, $0<a<b$, με συνάρτηση πυκνότητας πιθανότητας

$$f_U(u)=\frac{1}{b-a},\ u\in[a,b].$$

Η αναμενόμενη τιμή και η διασπορά της κατανομής είναι $E[U]=\frac{a+b}{2}$ και $V[U]=\frac{(b-a)^2}{12}$. Προκειμένου να έχουμε $E[U]=1$ και $V[U]=\sigma^2$, λύνουμε το σύστημα $\begin{cases}\frac{a+b}{2}=1\\ \frac{(b-a)^2}{12}=\sigma^2\end{cases}$, άρα $\begin{cases}a+b=2\\ b-a=2\sigma\sqrt{3}\end{cases}$, άρα $\begin{cases}a=1-\sigma\sqrt{3}\\ b=1+\sigma\sqrt{3}\end{cases}$.

Ο μετασχηματισμός Laplace της ευπάθειας είναι

$$\mathbf{L}_U(s)=\mathbf{L}(s)=E\left[e^{-sU}\right]=\int_a^b e^{-su}\cdot f_U(u)du=\int_a^b\frac{1}{b-a}\cdot e^{-su}du=\frac{1}{b-a}\cdot\left[\frac{e^{-su}}{-s}\right]_a^b,$$

άρα

$$\boxed{\mathbf{L}(s)=\frac{e^{-sa}-e^{-sb}}{s(b-a)}}\ \textbf{(3.58)}$$

Έτσι, η μη δεσμευμένη συνάρτηση επιβίωσης είναι:

$$\boxed{S(t)=\mathbf{L}\{H_0(t)\}=\frac{e^{-aH_0(t)}-e^{-bH_0(t)}}{H_0(t)(b-a)}}\ \textbf{(3.59)}$$

Επίσης είναι $\ln\mathbf{L}(s)=\ln\left(e^{-sa}-e^{-sb}\right)-\ln s-\ln(b-a)$. Παραγωγίζουμε, άρα $\frac{\mathbf{L}'(s)}{\mathbf{L}(s)}=\frac{be^{-sb}-ae^{-sa}}{e^{-sa}-e^{-sb}}-\frac{1}{s}$, άρα η μη δεσμευμένη συνάρτηση κινδύνου είναι

$$h(t)=-h_0(t)\cdot\frac{\mathbf{L}'\{H_0(t)\}}{\mathbf{L}\{H_0(t)\}}\Leftrightarrow$$

$$\boxed{h(t)=h_0(t)\cdot\left(\frac{1}{H_0(t)}-\frac{be^{-bH_0(t)}-ae^{-aH_0(t)}}{e^{-aH_0(t)}-e^{-bH_0(t)}}\right)}\ \textbf{(3.60)}$$

3.5.2. Εκτίμηση παραμέτρων στο παραμετρικό μοντέλο Ομοιόμορφης ευπάθειας: Κατά τα γνωστά, σε δείγμα δεδομένων $\mathrm{A} = \{(X_i, D_i, \mathrm{u}_i) : i = 1, 2, \ldots, n\}$ μεγέθους n προσαρμόζουμε ένα παραμετρικό μοντέλο ευπάθειας, με διάνυσμα παραμέτρων της βασικής συνάρτησης κινδύνου έστω $\boldsymbol{\theta}$, διάνυσμα συντελεστών των συμμεταβλητών $\boldsymbol{\beta}$, και τιμή της ευπάθειας για την $i-$μονάδα του δείγματος την u_i, $i = 1, 2, \ldots, n$. Η συνάρτηση πιθανοφάνειας απαλλαγμένη από την ευπάθεια είναι:

$$L\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2\right) = \prod_{i=1}^{n} \left(h\left(X_i \mid \boldsymbol{z}_i\right)\right)^{D_i} \cdot S\left(X_i \mid \boldsymbol{z}_i\right),$$

όπου οι $h\left(X_i \mid \boldsymbol{z}_i\right)$ και $S\left(X_i \mid \boldsymbol{z}_i\right)$ υπολογίζονται με τη βοήθεια των **(3.60)** και **(3.59)** αντίστοιχα, δηλ:

$$h\left(X_i \mid \boldsymbol{z}_i\right) = h_0\left(X_i \mid \boldsymbol{z}_i\right) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}} \left(\frac{1}{H_0\left(X_i \mid \boldsymbol{z}_i\right) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}}} - \frac{b e^{-b H_0\left(X_i \mid \boldsymbol{z}_i\right) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}}} - a e^{-a H_0\left(X_i \mid \boldsymbol{z}_i\right) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}}}}{e^{-a H_0\left(X_i \mid \boldsymbol{z}_i\right) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}}} - e^{-b\left(X_i \mid \boldsymbol{z}_i\right) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}}}} \right)$$

και

$$S\left(X_i \mid \boldsymbol{z}_i\right) = \frac{e^{-a\left(X_i \mid \boldsymbol{z}_i\right) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}}} - e^{-b\left(X_i \mid \boldsymbol{z}_i\right) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}}}}{H_0\left(X_i \mid \boldsymbol{z}_i\right) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}} \cdot (b - a)}$$

3.5.3. Εκτίμηση παραμέτρων στο ημι-παραμετρικό μοντέλο Ομοιόμορφης ευπάθειας: Ακολουθώντας την ιδέα της επί ποινή μερικής πιθανοφάνειας (*penalized partial likelihood*), την οποία περιγράψαμε στην παράγραφο **3.3.3.**, θεωρούμε την τυχαία μεταβλητή $W = \ln U$ και έχουμε ότι η επί ποινή λογαριθμοποιημένη πιθανοφάνεια είναι:

$$\ell_{ppl}\left(\boldsymbol{\beta}, \sigma^2 \mid w_1, \ldots, w_n\right) = \sum_{i=1}^{n} D_i \left\{ \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i + w_i - \ln\left(\sum_{j \in \Re_{(i)}} e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_j + w_j} \right) \right\} + \sum_{i=1}^{n} \ln f_W\left(w_i ; \sigma^2\right)$$

(3.61)

όπου το δεύτερο άθροισμα είναι ο παράγοντας της ποινής:

$$\ell_{pen}\left(\sigma^2 \mid w_1, w_2, \ldots, w_n\right) = \sum_{i=1}^{n} \ln f_W\left(w_i ; \sigma^2\right)$$

Ο υπολογισμός της $f_W\left(w;\sigma^2\right)$ γίνεται κατά τα γνωστά, αν ξεκινήσουμε από την αντίστοιχη αθροιστική συνάρτηση κατανομής:

$$F_W(w)=P[W\le w]=P[\ln U\le w]=P\left[U\le e^w\right]=F_U\left(e^w\right)$$

Παραγωγίζοντας παίρνουμε $f_W(w)=f_U\left(e^w\right)\cdot e^w$, άρα:

$$f_W(w)=\frac{e^w}{b-a}$$

άρα ο παράγοντας ποινής είναι:

$$\ell_{pen}\left(\sigma^2\mid w_1,w_2,\ldots,w_n\right)=\sum_{i=1}^{n}w_i-n\ln(b-a) \quad \textbf{(3.62)}$$

ΚΕΦΑΛΑΙΟ 4

ΜΟΝΤΕΛΑ ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΥ – ΠΡΟΣΟΜΟΙΩΣΕΙΣ & ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ

4.1. ΜΟΝΤΕΛΑ ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΥ

4.1.1. Εισαγωγή στα μοντέλα μετασχηματισμού: (Πηγές: [49]-[55]) Θεωρούμε το μοντέλο ευπάθειας που περιγράψαμε με την εξίσωση **(3.1)**:

$$h(t \mid u, \mathbf{z}) = u \cdot h_0(t) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}}$$

όπου $h(\cdot)$ η συνάρτηση κινδύνου, $h_0(\cdot)$ η βασική συνάρτηση κινδύνου (*baseline hazard*), $\mathbf{z}$ το διάνυσμα των συμμεταβλητών, $\boldsymbol{\beta}$ το διάνυσμα των προς εκτίμηση παραμέτρων και u η τυχαία μεταβλητή της ευπάθειας, η οποία, όπως είδαμε, χρησιμοποιείται για να περιγράψει την κρυμμένη ετερογένεια, δηλ. πιθανή ύπαρξη συμμεταβλητών που δεν έχουν εισέλθει στο διάνυσμα $\mathbf{z}$.

Η συνάρτηση επιβίωσης του παραπάνω μοντέλου προκύπτει από την εξίσωση

$$S(t \mid u, \mathbf{z}) = e^{-H(t \mid u, \mathbf{z})} = e^{-u \cdot H_0(t) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}}}$$

όπου $H(\cdot)$ η σωρευτική συνάρτηση κινδύνου και $H_0(\cdot)$ η βασική σωρευτική συνάρτηση κινδύνου αντίστοιχα.

Είδαμε επίσης ότι η μη-δεσμευμένη συνάρτηση επιβίωσης προκύπτει αν ολοκληρώσουμε την παραπάνω εξίσωση ως προς την ευπάθεια. Συγκεκριμένα, αν $F_U(u)$ είναι η αθροιστική συνάρτηση κατανομής της ευπάθειας U, τότε:

$$S(t \mid \mathbf{z}) = \int_0^{\infty} e^{-u \cdot H_0(t) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}}} dF_U(u) \quad \textbf{(4.1)}$$

ή ισοδύναμα, αν $f_U(u)$ είναι η συνάρτηση πυκνότητας πιθανότητας της U, τότε:

$$S(t \mid \boldsymbol{z}) = \int_0^{\infty} e^{-u \cdot H_0(t) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}}} f_U(u) du \quad \textbf{(4.2)}$$

Η εξίσωση **(4.2)** μπορεί να γραφεί στη μορφή:

$$S(t \mid \boldsymbol{z}) = e^{\ln\left(\int_0^{\infty} e^{-u \cdot H_0(t) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}}} f_U(u) du \right)} \quad \textbf{(4.3)}$$

ή ισοδύναμα:

$$S(t \mid \boldsymbol{z}) = e^{-G\left(H_0(t) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}} \right)} \quad \textbf{(4.4)}$$

όπου

$$G(x) = -\ln\left(\int_0^{\infty} e^{-u \cdot x} f_U(u) du \right) \quad \textbf{(4.5)}$$

Για κάθε επιλογή της κατανομής της ευπάθειας U, δημιουργείται από τις εξισώσεις **(4.4)** και **(4.5)**, μία διαφορετική συνάρτηση $G(\cdot)$. Επιπλέον, από την εξίσωση **(4.5)** προκύπτει ότι η μη-δεσμευμένη αθροιστική συνάρτηση κατανομής του χρόνου επιβίωσης είναι

$$F(t \mid u, \boldsymbol{z}) = 1 - e^{-G\left(H_0(t) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}} \right)} \quad \textbf{(4.6)}$$

Η εξίσωση **(4.6)** περιγράφει ένα μοντέλο που, αν και προέκυψε από το μοντέλο ευπάθειας (βάσει του ορισμού της συνάρτησης $G(\cdot)$ μέσω της εξίσωσης **(4.5)**), ωστόσο αποτελεί "μία κλάση παραέξω" από αυτό, αν θεωρηθεί ότι η συνάρτηση $G(\cdot)$ ανήκει σε μία κλάση συναρτήσεων που ικανοποιούν κάποιες απαιτούμενες συνθήκες τις οποίες θα δούμε παρακάτω (χωρίς αυτό να σημαίνει ότι γενικά η συνάρτηση $G(\cdot)$ πρέπει να οριστεί απαραίτητα μέσω του μετασχηματισμού Laplace μιας μεταβλητής ευπάθειας).

Η κλάση που περιέχει μοντέλα όπως αυτό της εξίσωσης **(4.6)** αποτελεί την κλάση των **μοντέλων μετασχηματισμού** (*transformation models*).

Στην παραμετρική περίπτωση, η σωρευτική συνάρτηση κινδύνου ενός μοντέλου μετασχηματισμού με διάνυσμα συμμεταβλητών $\boldsymbol{z}$ και διάνυσμα συντελεστών $\boldsymbol{\beta}$, είναι της μορφής:

$$\boxed{H(t, \boldsymbol{\mu} \mid \mathbf{z}) = G\left(\Lambda(t, \boldsymbol{\mu}) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}} \right)} \quad \textbf{(4.7)}$$

όπου $\boldsymbol{\mu}$ είναι διανυσματική "οχληρά παράμετρος"(*nuisance parameter*)[1], $G(\cdot)$ είναι η αύξουσα, τρεις φορές συνεχώς παραγωγίσιμη κοίλη συνάρτηση μετασχηματισμού με τις ιδιότητες $G(0) = 0$ και $G(\infty) = \infty$ και $\Lambda(t, \boldsymbol{\mu})$ είναι γνησίως αύξουσα συνάρτηση του χρόνου με τις ιδιότητες $\Lambda(0, \boldsymbol{\mu}) = 0$ και $\Lambda(\infty, \boldsymbol{\mu}) = \infty$ (βλ. [54] σελ. 1 και [50]).

Προφανώς η επιλογή $G(x) = x$ και $\Lambda(t, \boldsymbol{\mu}) = H_0(t)$ δίνει το μοντέλο του Cox χωρίς ευπάθεια, ενώ στην περίπτωση του μοντέλου ευπάθειας, παρατηρούμε από την **(4.5)** ότι η συνάρτηση $G(x)$ είναι η αντίθετη του λογαρίθμου του μετασχηματισμού Laplace της ευπάθειας, δηλ. $G(x) = -\ln \mathbf{L}(x)$ με $\mathbf{L}(x) = \mathbf{L}_U(x) = E\left[e^{-Ux}\right] = \int_0^\infty e^{-u \cdot x} f_U(u) du$.

Το πλεονέκτημα των μοντέλων μετασχηματισμού είναι ότι μπορούν να περιγράψουν και μοντέλα στα οποία παραβιάζεται η υπόθεση του αναλογικών κινδύνων, όπως π.χ. συχνά συμβαίνει σε μοντέλα πολυμεταβλητού χρόνου (βλ. [51] σελ. 1 και 2).

4.1.2. Εκτίμηση παραμέτρων στα μοντέλα μετασχηματισμού: Θα εφαρμόσουμε τη μέθοδο μεγίστης πιθανοφάνειας, ώστε στο μοντέλο μετασχηματισμού που περιγράφεται από την εξίσωση **(4.7)**, να εκτιμήσουμε τις (βασικές) παραμέτρους παλινδρόμησης $\boldsymbol{\beta}$.

Από την εξίσωση **(4.7)** και υποθέτοντας ότι η συνάρτηση $\Lambda(t, \boldsymbol{\mu})$ είναι παραγωγίσιμη, η συνάρτηση κινδύνου του μοντέλου είναι

$$h(t, \boldsymbol{\beta}, \boldsymbol{\mu} \mid \mathbf{z}) = G'\left(\Lambda(t, \boldsymbol{\mu}) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}} \right) \cdot \lambda(t, \boldsymbol{\mu}) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}} \quad \textbf{(4.8)}$$

όπου $\lambda(t, \boldsymbol{\mu}) = \Lambda'(t, \boldsymbol{\mu})$ και η συνάρτηση επιβίωσης

$$S(t, \boldsymbol{\beta}, \boldsymbol{\mu} \mid \mathbf{z}) = e^{-G\left(\Lambda(t, \boldsymbol{\mu}) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}} \right)} \quad \textbf{(4.9)}$$

[1] Οχληρά παράμετρος: παράμετρος όχι άμεσου ενδιαφέροντος, η οποία όμως χρησιμοποιείται στην εκτίμηση παραμέτρων που ενδιαφέρουν (βλ. [56])

Θεωρώντας κατά τα συνήθη, δείγμα δεδομένων $\mathrm{A}=\{(X_i, D_i): i=1,2,\ldots,\mathrm{n}\}$ και με τη βοήθεια των εξισώσεων **(4.8)** και **(4.9)**, η συνάρτηση πιθανοφάνειας είναι

$$L=\prod_{i=1}^{n}\left(h(X_i\,|\,\mathbf{z}_i)\right)^{D_i}\cdot S(X_i\,|\,\mathbf{z}_i)\Leftrightarrow$$

$$L(\boldsymbol{\mu},\boldsymbol{\beta})=\prod_{i=1}^{n}\left\{\left(G'\left(\Lambda(X_i,\boldsymbol{\mu})\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}\right)\cdot\lambda(t,\boldsymbol{\mu})\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}\right)^{D_i}\cdot e^{-G\left(\Lambda(X_i,\boldsymbol{\mu})\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}\right)}\right\}$$

(4.10)

Η εκτίμηση των συντελεστών $\boldsymbol{\beta}$, καθώς και άλλων παραμέτρων όπως π.χ. παραμέτρων $\boldsymbol{\theta}$ που υπεισέρχονται στη συνάρτηση λ αν αυτή έχει παραμετρική μορφή, γίνεται με τη μεγιστοποίηση της λογαριθμοποιημένης πιθανοφάνειας $\ell=\ln L$.

4.1.3. Εκτίμηση παραμέτρων στο μοντέλο ευπάθειας, αντιμετωπίζοντας αυτό ως ειδική περίπτωση του μοντέλου μετασχηματισμού: Θα εκτιμήσουμε τις παραμέτρους μοντέλου ευπάθειας, κάνοντας χρήση της προηγούμενης θεωρίας που αφορά τα μοντέλα μετασχηματισμού. Θεωρούμε το μοντέλο ευπάθειας

$$h(t\,|\,u,\mathbf{z})=u\cdot h_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}}$$

του οποίου η σωρευτική συνάρτηση κινδύνου είναι

$$H(t\,|\,u,\mathbf{z})=u\cdot H_0(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}}$$

Το μοντέλο ευπάθειας μετατρέπεται σε μοντέλο μετασχηματισμού μέσω της εξίσωσης **(4.7)**, αφού

$$H(t\,|\,\mathbf{z})=G\left(\Lambda(t)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}}\right)$$

όπου $\Lambda(t)=H_0(t)$ και όπου $G(x)=-\ln\left(\int_0^{\infty}e^{-u\cdot x}f_U(u)du\right)$ η συνάρτηση μετασχηματισμού που ορίζεται από την **(4.5)**.

Η συνάρτηση πιθανοφάνειας **(4.10)**, μετασχηματίζεται τώρα σε:

$$L=\prod_{i=1}^{n}\left\{\left(G'\left(H_0(X_i\,|\,\mathbf{z}_i)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}\right)\cdot h_0(X_i\,|\,\mathbf{z}_i)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}\right)^{D_i}\cdot e^{-G\left(H_0(X_i|\,\mathbf{z}_i)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}\right)}\right\}$$

(4.11)

και άρα η λογαριθμοποιημένη πιθανοφάνεια είναι:

$$\ell=\ln L=\sum_{i=1}^{n}\left\{D_i\left[\ln G'\left(H_0(X_i\,|\,\mathbf{z}_i)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}\right)+\ln h_0(X_i\,|\,\mathbf{z}_i)+\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i\right]-\right.$$
$$\left.-G\left(H_0(X_i\,|\,\mathbf{z}_i)\cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i}\right)\right\}$$

(4.12)

Στη συνέχεια, έχουμε δύο δυνατές περιπτώσεις:

- Αν η βασική συνάρτηση κινδύνου θεωρηθεί παραμετρική (με γνωστή κατανομή) με διάνυσμα παραμέτρων $\boldsymbol{\theta}$, τότε οι L και ℓ είναι συναρτήσεις των $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, δηλ. $L(\boldsymbol{\beta},\boldsymbol{\theta})$ και $\ell(\boldsymbol{\beta},\boldsymbol{\theta})$.
- Αν η βασική συνάρτηση κινδύνου είναι μη παραμετρική (οπότε έχουμε ημι-παραμετρικό μοντέλο ευπάθειας), τότε η βασική συνάρτηση κινδύνου $h_0(t)$ ή η σωρευτική βασική συνάρτηση κινδύνου $H_0(t)$ είναι οχληρά παράμετρος και οι πιθανοφάνειες είναι συναρτήσεις π.χ. των $\boldsymbol{\beta}$, $h_0(t)$ δηλ. $L(\boldsymbol{\beta},h_0(t))$ και $\ell(\boldsymbol{\beta},h_0(t))$.

4.2. ΠΡΟΣΟΜΟΙΩΣΕΙΣ ΓΙΑ ΤΗΝ ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ ΚΑΙ ΤΗΝ ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΣΤΟ ΠΑΡΑΜΕΤΡΙΚΟ ΜΟΝΤΕΛΟ ΓΑΜΜΑ ΕΥΠΑΘΕΙΑΣ ΜΕΣΩ ΤΟΥ ΜΟΝΤΕΛΟΥ ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΥ

4.2.1. Εισαγωγή: Με τη βοήθεια του στατιστικού πακέτου R, θα προσομοιώσουμε μοντέλα ευπάθειας με παραμετρική βασική συνάρτηση κινδύνου και θα εκτιμήσουμε τις παραμέτρους αυτής, καθώς και τους συντελεστές των συμμεταβλητών.

Προγενέστερη εργασία υπάρχει στο [57], όπου προσομοιώθηκαν μοντέλα ευπάθειας με βασική συνάρτηση κινδύνου της Εκθετικής κατανομής

ε(1), ευπάθεια που ακολουθεί Γάμμα και Inverse Gaussian κατανομή και $k=4$ το πλήθος συμμεταβλητές. Ακολούθως, θα μελετήσουμε μοντέλο που έχει βασική συνάρτηση κινδύνου της Weibull κατανομής, Γάμμα ευπάθεια και οσεσδήποτε το πλήθος συμμεταβλητές.

4.2.2. Υποθέσεις του μοντέλου: Θεωρούμε το μοντέλο ευπάθειας

$$h(t \mid u, \mathbf{z}) = u \cdot h_0(t) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}}$$

Για απλούστευση, εργαζόμαστε αρχικά με τις υποθέσεις του [57] ως προς τις συμμεταβλητές, δηλ. υποθέτουμε ότι:

- Υπάρχουν k το πλήθος συμμεταβλητές, άρα $\mathbf{z}^T = (Z_1, Z_2, \ldots, Z_k)$
- Το διάνυσμα $\mathbf{z}^T = (Z_1, Z_2, \ldots, Z_k)$ ακολουθεί την πολυδιάστατη Κανονική κατανομή $N(\mathbf{0}, \boldsymbol{\Sigma})$ με μέση τιμή το μηδενικό διάνυσμα και πίνακα διακύμανσης-συνδιακύμανσης τον $\boldsymbol{\Sigma} = \left[\left(\frac{1}{2} \right)^{|i-j|} \right]_{i,\, j=1,2,\ldots,k}$ δηλ. ισχύει $Cov(Z_i, Z_j) = \left(\frac{1}{2} \right)^{|i-j|}$ για κάθε $i, j = 1, 2, \ldots, k$ (*Π.χ. για* $k=4$ *είναι* $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 & 0.25 & 0.125 \\ 0.5 & 1 & 0.5 & 0.25 \\ 0.25 & 0.5 & 1 & 0.5 \\ 0.125 & 0.25 & 0.5 & 1 \end{bmatrix}$)
- Το διάνυσμα των συντελεστών του μοντέλου είναι γνωστό και ίσο με $\boldsymbol{\beta}_0^T = (\beta_1, \beta_2, \ldots, \beta_k)$. Από το διάνυσμα $\boldsymbol{\beta}_0^T$, φαίνονται οι σημαντικές συμμεταβλητές z_i που είναι αυτές με συντελεστή $\beta_i \neq 0$. (*Π.χ. για* $k=4$ *και* $\boldsymbol{\beta}_0^T = (0.8, 0, 0, 1)$, *μόνο οι συμμεταβλητές* Z_1 *και* Z_4 *είναι σημαντικές*).

Επίσης υποθέτουμε ότι:

- Η βασική συνάρτηση κινδύνου είναι της κατανομής Weibull, με παράμετρο κλίμακας $\alpha > 0$ και παράμετρο σχήματος $\lambda > 0$, άρα

$$h_0(t) = \lambda\alpha^{-\lambda}t^{\lambda-1},\ t>0$$

και

$$H_0(t) = \left(\frac{t}{\alpha}\right)^{\lambda},\ t>0$$

Για τις ανάγκες της προσομοίωσης θα θεωρήσουμε παράμετρο κλίμακας $\alpha = 1$ *και παράμετρο σχήματος* $\lambda = 2$.

- Η τυχαία μεταβλητή U της ευπάθειας, ακολουθεί την κατανομή Γάμμα με παράμετρο κλίμακας $\frac{1}{\sigma^2}$ και παράμετρο σχήματος $\frac{1}{\sigma^2}$, οπότε έχει αναμενόμενη τιμή $E[U]=1$ και αναμενόμενη διασπορά $V[U]=\sigma^2$, ενώ η συνάρτηση πυκνότητας πιθανότητας της U είναι

$$f_U(u) = f(u) = \frac{\left(\frac{1}{\sigma^2}\right)^{\frac{1}{\sigma^2}}}{\Gamma\left(\frac{1}{\sigma^2}\right)} u^{\frac{1}{\sigma^2}-1} e^{-\frac{u}{\sigma^2}}$$

Για τις ανάγκες της προσομοίωσης χρησιμοποιούμε $\sigma^2 = 0.25$.

- Σύμφωνα με τα παραπάνω, το μοντέλο ευπάθειας περιγράφεται τώρα από την εξίσωση

$$h(t \mid u, \boldsymbol{z}) = u \cdot \lambda\alpha^{-\lambda}t^{\lambda-1} \cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}}$$

η οποία ισοδύναμα γράφεται:

$$h(t \mid u, \boldsymbol{z}) = \lambda \cdot \left\{\alpha \cdot \left(u \cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}}\right)^{-\frac{1}{\lambda}}\right\}^{-\lambda} t^{\lambda-1}$$

και άρα *το μοντέλο ευπάθειας είναι επίσης της κατανομής Weibull με παράμετρο σχήματος* λ *και παράμετρο κλίμακας* $A = \alpha \cdot \left(u \cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}}\right)^{-\frac{1}{\lambda}}$.

- Ο μετασχηματισμός Laplace της ευπάθειας έχει υπολογισθεί στην παράγραφο **3.2.1.** και συγκεκριμένα στην εξίσωση **(3.21)** και είναι

$$\mathbf{L}(s) = \left(1+\sigma^2 s\right)^{-\frac{1}{\sigma^2}}$$

Επομένως, η συνάρτηση μετασχηματισμού είναι $G(x) = -\ln \mathbf{L}(x)$, άρα:

$$G(x)=\frac{1}{\sigma^2}\ln\left(1+\sigma^2 x\right)$$

με πρώτη παράγωγο

$$G'(x)=\frac{1}{1+\sigma^2 x}$$

4.2.3. Ζητούμενο του προβλήματος: Έχοντας τις παραπάνω υποθέσεις, στόχος μας είναι αρχικά, να δημιουργήσουμε δείγματα μεγέθους n (το n επιλέγεται κατά βούληση), με δεδομένα που θα αποτελούνται από τις γνωστές τριάδες $\mathrm{A}=\{(X_i, D_i, \mathrm{u}_i): i=1,2,\ldots,n\}$, όπου:

- X_i είναι ο χρόνος αποτυχίας ή λογοκρισίας, $i=1,2,\ldots,\mathrm{n}$
- $D_i=\begin{cases}1 & \text{αν ο χρόνος } X_i \text{ είναι χρόνος διακοπής}\\ 0 & \text{αν ο χρόνος } X_i \text{ είναι χρόνος λογοκρισίας}\end{cases}$ είναι η δείκτρια συνάρτηση λογοκρισίας, $i=1,2,\ldots,\mathrm{n}$
- u_i είναι η τιμή της ευπάθειας της $i-$παρατήρησης, $i=1,2,\ldots,\mathrm{n}$

Στα δεδομένα αυτά, θα προσαρμόσουμε όλα τα δυνατά μοντέλα ευπάθειας με $p=1,2,\ldots,k$ το πλήθος συμμεταβλητές. Για καθένα από αυτά τα μοντέλα, θα υπολογίσουμε τις εκτιμήτριες $\hat{\boldsymbol{\beta}}, \hat{\alpha}, \hat{\lambda}$ των συντελεστών $\boldsymbol{\beta}$ των συμμεταβλητών και των παραμέτρων κλίμακας α και σχήματος λ της βασικής κατανομής Weibull που μεγιστοποιούν τη συνάρτηση λογαριθμοποιημένης πιθανοφάνειας $\ell(\boldsymbol{\beta}, \alpha, \lambda)$ όπως αυτή δημιουργείται μέσω του μοντέλου μετασχηματισμού της ευπάθειας (εξίσωση **(4.12)**) για δεδομένο σ^2, δηλ. για γνωστή συνάρτηση $G(\cdot)$.

Στη συνέχεια, για κάθε μοντέλο θα υπολογίσουμε τις τιμές των κριτηρίων AIC και BIC, δηλ. $\mathrm{AIC}=-2\cdot\hat{\ell}\left(\hat{\boldsymbol{\beta}}, \hat{\alpha}, \hat{\lambda}\right)+2p$ και $\mathrm{BIC}=-2\cdot\hat{\ell}\left(\hat{\boldsymbol{\beta}}, \hat{\alpha}, \hat{\lambda}\right)+p\ln n$ και θα ζητήσουμε να επιλεγεί το βέλτιστο μοντέλο (αυτό με το μικρότερο AIC ή BIC), αναμένοντας ότι στις περισσότερες περιπτώσεις θα επιλεγεί το σωστό μοντέλο (αυτό με τις σημαντικές συμμεταβλητές).

Παράλληλα, προκειμένου να έχουμε μία βάση αναφοράς, οι εκτιμήσεις των παραμέτρων και η επιλογή του μοντέλου θα πραγματοποιηθούν και με χρήση της συνάρτησης coxph (έτοιμη συνάρτηση της R για προσαρμογή του μοντέλου του Cox σε δεδομένα).

Τέλος, όλα τα παραπάνω θα επαναληφθούν για όσες φορές επιλέξουμε, προκειμένου να συγκρίνουμε τα ποσοστά επιτυχίας αλλά και τις εκτιμήσεις των παραμέτρων $\boldsymbol{\beta}, \alpha, \lambda$, μεταξύ της μεθόδου μας και της coxph.

Το υπολογιστικό πλεονέκτημα της μεθόδου μας είναι οτι με μια μικρή αλλαγή στον κώδικα ως προς την συνάρτηση $G(\cdot)$ αντιμετωπίζονται ταυτόχρονα όλα τα μοντέλα μετασχηματισμού.

4.2.4. Κατασκευή του κώδικα: Αρχικοποίηση: Στις πρώτες γραμμές του κώδικα γίνεται η αρχικοποίηση (*initialization*) όπου δίνονται:

- το διάνυσμα των συντελεστών $\boldsymbol{\beta}_0^T$ του μοντέλου
- ο αριθμός *niter* των επαναλήψεων όλης της διαδικασίας
- το πλήθος *k* των συμμεταβλητών (που ισούται με τη διάσταση του $\boldsymbol{\beta}_0^T$)
- το μέγεθος *n* του δείγματος των προσομοιωμένων δεδομένων που θα κατασκευάσουμε
- η παράμετρος κλίμακας α της κατανομής Weibull που αφορά στη βασική συνάρτηση κινδύνου (μεταβλητή *w_scale_init*)
- η παράμετρος σχήματος λ της κατανομής Weibull που αφορά στη βασική συνάρτηση κινδύνου (μεταβλητή *w_shape*)
- η διασπορά σ^2 της ευπάθειας (μεταβλητή *sqr_sigma*)
- η βασική συνάρτηση κινδύνου $h_0(t) = \lambda\alpha^{-\lambda}t^{\lambda-1}$, η οποία στον κώδικα γράφεται στην ισοδύναμη μορφή $h_0(t) = \lambda e^{-\lambda\ln\alpha}e^{(\lambda-1)\ln t} = \lambda e^{\lambda\ln\left(\frac{t}{\alpha}\right)-\ln t}$
- η βασική σωρευτική συνάρτηση κινδύνου $H_0(t) = \left(\frac{t}{\alpha}\right)^{\lambda}$, η οποία στον κώδικα γράφεται στην ισοδύναμη μορφή $H_0(t) = e^{\lambda\ln\left(\frac{t}{\alpha}\right)}$

- η συνάρτηση μετασχηματισμού $G(x)=\frac{1}{\sigma^2}\ln\left(1+\sigma^2 x\right)$
- η παράγωγος $G'(x)=\frac{1}{1+\sigma^2 x}$ της συνάρτησης μετασχηματισμού (μεταβλητή *dG*)
- το διάνυσμα λογοκρισίας *censor_vector*: πρόκειται για διάνυσμα διάστασης ίσης με το πλήθος *niter* των επαναλήψεων, στις συντεταγμένες του οποίου αποθηκεύονται τα ποσοστά λογοκριμένων παρατηρήσεων του δείγματος κάθε επανάληψης.
- οι παράμετροι *lower_censor*, *upper_censor* και *uniform_mean* του ποσοστού λογοκρισίας, που καθορίζουν στη συνέχεια του κώδικα μεταξύ ποιων τιμών κυμαίνεται το ποσοστό των λογοκριμένων παρατηρήσεων.
- ο πίνακας αποτελεσμάτων *our_AIC_iter_results* βάσει της μεθόδου μας και του κριτηρίου AIC: είναι πίνακας διάστασης $(k+7)\times niter$, στου οποίου τις *niter* το πλήθος στήλες αποθηκεύονται τα αποτελέσματα για το καλύτερο μοντέλο που επιλέγει η μέθοδός μας βάσει του κριτηρίου AIC. Η δομή μίας στήλης αυτού του πίνακα είναι:

$$\begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \\ \hat{\alpha} \\ \hat{\lambda} \\ AIC \\ Error1 \\ Error2 \\ Error1\&2 \\ PerCentCensor \end{bmatrix}$$

όπου $\hat{\beta}_1$ είναι η εκτίμηση της παραμέτρου β_1 κ.ο.κ., ενώ τα σφάλματα $Error1$, $Error2$ και $Error1\&2$ ορίζονται ως:

$$Error1=\begin{cases}1 & \text{αν επιλεγεί τουλάχιστον μία ΜΗ - σημαντική μεταβλητή}\\ 0 & \text{αλλιώς}\end{cases}$$

$$Error2 = \begin{cases} 1 & \text{αν ΔΕΝ επιλεγεί τουλάχιστον μία σημαντική μεταβλητή} \\ 0 & \text{αλλιώς} \end{cases}$$

και $Error1\&2 = \begin{cases} 1 & \text{αν } Error1 = 1 \text{ και } Error2 = 1 \\ 0 & \text{αλλιώς} \end{cases}$ δηλ. το $Error1\&2$ παίρνει την τιμή 1 όταν τα $Error1$ και $Error2$ πάρουν ταυτόχρονα την τιμή 1, δηλ. όταν επιλεγεί μοντέλο που περιέχει τουλάχιστον μία μη σημαντική μεταβλητή και δεν περιέχει τουλάχιστον μία σημαντική μεταβλητή.

- ο πίνακας αποτελεσμάτων *our_BIC_iter_results* βάσει της μεθόδου μας και του κριτηρίου BIC: ο αντίστοιχος πίνακας διάστασης $(k+7)\times niter$ για τα καλύτερα μοντέλα που επιλέγει η μέθοδός μας βάσει του κριτηρίου BIC.
- ο πίνακας αποτελεσμάτων *coxph_AIC_iter_results* βάσει της συνάρτησης *coxph* και του κριτηρίου AIC: είναι πίνακας διάστασης $(k+5)\times niter$, στου οποίου τις *niter* το πλήθος στήλες αποθηκεύονται τα αποτελέσματα για το καλύτερο μοντέλο που επιλέγει η συνάρτηση *coxph* βάσει του κριτηρίου AIC. Η δομή μίας στήλης αυτού του πίνακα είναι:

$$\begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \\ AIC \\ Error1 \\ Error2 \\ Error1\&2 \\ PerCentCensor \end{bmatrix}$$

κατ' αντιστοιχία με τα προηγούμενα (εδώ δεν εκτιμώνται οι παράμετροι κλίμακας και σχήματος της συνάρτησης Weibull, αφού το μοντέλο είναι ημιπαραμετρικό).

- ο πίνακας αποτελεσμάτων *coxph_BIC_iter_results* βάσει της συνάρτησης *coxph* και του κριτηρίου BIC: ο αντίστοιχος πίνακας διάστασης $(k+5)\times niter$ για τα καλύτερα μοντέλα που επιλέγει η *coxph* βάσει του κριτηρίου BIC.

Ο κώδικας μέχρι εδώ και για $\boldsymbol{\beta}_0^T = (0.8, 0, 0, 1)$, $niter = 10$, $n = 50$, $\alpha = 1$, $\lambda = 2$, $\sigma^2 = 0.25$, $lower\_censor = 0.43$, $upper\_censor = 0.48$ και $uniform\_mean = 1$, είναι ο ακόλουθος:

```
library(nlme)
library(mvtnorm)        #R library for the Multivariate Normal Distribution
library(splines)
library(survival)
library(combinat)       #R library for the Combinatorics

#=============================================
#Initialization

#Parameter vector named "b" is given manually by the user
#"b" is the vector of real coefficients in our model
#if b[i]==0, then a not important covariate is indicated

#b<-matrix(c(2, 0, -1.3), nrow=1)              #Vector of real coefficients - 3 Covariates
b<-matrix(c(0.8, 0, 0, 1), nrow=1)             #Vector of real coefficients - 4 Covariates
#b<-matrix(c(0, 2.5, 4, 0, -1), nrow=1)        #Vector of real coefficients - 5 Covariates
#b<-matrix(c(1, 0, -3, -6.8, 0,0), nrow=1)     #Vector of real coefficients - 5 Covariates

niter<-10                  #Number of iterations

k<-ncol(b)                 #Number of covariates, k>=2

n<-50                      #Sample size

w_scale_init<-1#Weibull scale parameter (alpha)
w_shape<-2     #Weibull shape parameter (lambda)
sqr_sigma<-0.25            #Variance of frailty (Gamma) distribution
h<-function(x,alpha,lambda){lambda*exp(lambda*log(x/alpha)-log(x))}   #Baseline hazard
                                                                      #function
h(t)
```

```
H<-function(x,alpha,lambda){exp(lambda*log(x/alpha))} #Baseline cumulative hazard
                                                      #function H(t)
G<-function(x) {(1/sqr_sigma)*log(1+sqr_sigma*x)}      #G function
dG<-function(x)          {1/(1+sqr_sigma*x)}      #Derivative of G function
censor_vector<-numeric(niter)   #in this vector we store all censoring proportions


#Choose the desired censoring proportion interval, or create yours

lower_censor<-0.43       #lower censoring proportion in every iteration
upper_censor<-0.48       #upper censoring proportion in every iteration
uniform_mean<-1
#(for censoring proportion between 0.43 and 0.45 set uniform_mean<-1)

#lower_censor<-0.05     #lower censoring proportion in every iteration
#upper_censor<-0.15     #upper censoring proportion in every iteration
#uniform_mean<-15
#(for censoring proportion between 0.05 and 0.15 set uniform_mean<-15)

#lower_censor<-0.75     #lower censoring proportion in every iteration
#upper_censor<-0.85     #upper censoring proportion in every iteration
#uniform_mean<-0.2
#(for censoring proportion between 0.75 and 0.85 set uniform_mean<-0.2)


#In the following matrices, we store the best models

our_AIC_iter_results<-matrix(nrow=k+7, ncol=niter)
                #for storage of the best (due to AIC) fitted model of our method
our_BIC_iter_results<-matrix(nrow=k+7, ncol=niter)
                #for storage of the best (due to BIC) fitted model of our method
coxph_AIC_iter_results<-matrix(nrow=k+5, ncol=niter)
                #for storage of the best (due to AIC) fitted model of coxph method
coxph_BIC_iter_results<-matrix(nrow=k+5, ncol=niter)
                #for storage of the best (due to BIC) fitted model of coxph method
```

4.2.5. Κατασκευή του κώδικα: Προσομοίωση των δεδομένων: Σε μία μεγάλη επανάληψη *for (iter in 1:niter){...}* αρχίζει το κύριο μέρος που αφορά την προσομοίωση των δεδομένων και την επιλογή του βέλτιστου μοντέλου. Το κύριο μέρος εκτελείται *niter* το πλήθος φορές και τα αποτελέσματα αποθηκεύονται στους πίνακες που ορίσαμε παραπάνω.

Όσον αφορά το κομμάτι της προσομοίωσης, η δομή του κώδικα είναι η εξής:

- Κατασκευάζουμε τον πίνακα διακύμανσης-συνδιακύμανσης $\mathbf{\Sigma} = \left[\left(\frac{1}{2} \right)^{|i-j|} \right]_{i,\, j=1,2,\ldots,k}$. Ο πίνακας συμβολίζεται με *Cov_Matrix*.
- Κατασκευάζουμε τον πίνακα Z του οποίου στοιχεία είναι οι τιμές των συμμεταβλητών, ως τυχαίο πίνακα που ακολουθεί την πολυδιάστατη Κανονική κατανομή $N(\mathbf{0}, \mathbf{\Sigma})$. Ο πίνακας Z είναι διάστασης $n \times k$ και η δομή του είναι $Z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1k} \\ z_{21} & z_{22} & \cdots & z_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ z_{n1} & z_{n2} & \cdots & z_{nk} \end{bmatrix}$ δηλ. $Z = \left\lfloor z_{ij} \right\rfloor_{n \times k}$
- Κατασκευάζουμε το διάνυσμα *frailty*, διάστασης n, συντεταγμένες του οποίου είναι οι τιμές $u_1, u_2, \ldots, u_n$ της ευπάθειας. Αυτές, δημιουργούνται ως τυχαίες παρατηρήσεις από την κατανομή που ακολουθεί η ευπάθεια (εν προκειμένω την Γάμμα με παράμετρο κλίμακας $\lambda = \frac{1}{\sigma^2}$ και παράμετρο σχήματος $\kappa = \frac{1}{\sigma^2}$).
- Κατασκευάζουμε δείγμα από n το πλήθος χρόνους διακοπής (θανάτου) $\{T_1, T_2, \ldots, T_n\}$, οι οποίοι ακολουθούν το μοντέλο ευπάθειας της παραγράφου **4.2.2.**, δηλ. είναι τυχαίες παρατηρήσεις από την κατανομή Weibull με παράμετρο σχήματος λ και παράμετρο κλίμακας $A_i = \alpha \cdot \left(u_i \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}} \right)^{-\frac{1}{\lambda}}$,

$i = 1, 2, \ldots, \text{n}$, όπου α, λ οι παράμετροι κλίμακας και σχήματος αντίστοιχα της κατανομής Weibull την οποία ακολουθεί η βασική συνάρτηση κινδύνου.

- Κατασκευάζουμε δείγμα n το πλήθος χρόνων λογοκρισίας $\{C_1, C_2, \ldots, C_n\}$, ως τυχαίες παρατηρήσεις που ακολουθούν το ίδιο μοντέλο ευπάθειας, με τη διαφορά όμως ότι οι ευπάθειες u_i, $i = 1, 2, \ldots, \text{n}$, ακολουθούν τώρα Ομοιόμορφη κατανομή $U((0,\ 2 * uniform\_mean))$. Όπως είδαμε πριν, η παράμετρος *uniform_mean* επιλέγεται κατά τέτοιο τρόπο, ώστε το ποσοστό των λογοκριμένων παρατηρήσεων που θα προκύψει στη συνέχεια, να έχει συγκεκριμένο (επιθυμητό από εμάς) εύρος.
- Για κάθε $i = 1, 2, \ldots, \text{n}$, ορίζουμε ως παρατήρηση X_i τον ελάχιστο χρόνο από τους T_i και C_i. Δηλ. $X_i = \min\{T_i, C_i\}$, $i = 1, 2, \ldots, \text{n}$. Ταυτόχρονα ορίζεται η δείκτρια λογοκρισίας $D_i = \begin{cases} 1 & \text{αν } T_i \le C_i \\ 0 & \text{αλλιώς} \end{cases}$, $i = 1, 2, \ldots, \text{n}$.
- Ονομάζουμε D το <u>διάνυσμα λογοκρισίας</u> διάστασης n, στοιχεία του οποίου είναι οι δείκτριες D_i, $i = 1, 2, \ldots, \text{n}$. Προφανώς το <u>ποσοστό λογοκρισίας</u> *censoring proportion* είναι ο μέσος όρος των στοιχείων του διανύσματος D. Στη συνέχεια, η τιμή *censoring proportion* καθίσταται συντεταγμένη στο διάνυσμα *censor_vector*
- Η ανωτέρω διαδικασία προσομοίωσης επαναλαμβάνεται μέχρι να πετύχουμε *censoring proportion* εντός των ορίων που έχουμε θέσει με τις παραμέτρους *lower_censor* και *upper_censor*.
- Τα αποτελέσματα της προσομοίωσης μπορούν να παρατεθούν, αν επιθυμούμε, στον <u>πίνακα αποτελεσμάτων</u> *Results Matrix*.

Παραθέτουμε ακολούθως το κομμάτι του κώδικα που αφορά την προσομοίωση.

```
#Start of Iterations
#Model simulation and variable selection are repeated for niter times

for (iter in 1:niter)
{
```

```
#=============================================
#Variance-Covariance matrix of the real model
Cov_Matrix<-mat.or.vec(k, k)
        for (i in 1:k)
        {
                for (j in 1:k)
                {Cov_Matrix[i, j]<-exp(-abs(i-j)*log(2))}
                                                                        }


#=============================================
#Generation of Random matrix Z (nxk size), which follows
#Multivariate Normal Distribution N(0, Cov_Matrix)
#Matrix columns are the covariates Z[i]

Z<-rmvnorm(n, mean = rep(0, nrow(Cov_Matrix)), sigma = Cov_Matrix)

#=============================================
#Generation of frailty data

frailty<-numeric(n)
for (i in 1:n) {frailty[i]<-rgamma(1, shape=1/sqr_sigma, scale=sqr_sigma)}

#=============================================
#Generation of failure times T[i] given the frailty and given the covariates

#(Generation of failure times and censoring times will be made
#with respect to censoring proportion

per_cent_censor<-0
while (per_cent_censor<lower_censor|per_cent_censor>upper_censor ) {
        #start of desired per cent censoring proportion loop
          T<-numeric(n)
        for (i in 1:n)
        {
```

```
                w_scale<-w_scale_init*exp(-(1/w_shape)*log(frailty[i]*exp(b%*%Z[i,])))
                T[i]<-rweibull(1, w_shape, w_scale)
                                                                        }
        w_scale<-w_scale_init

#=============================================
#Generation of censoring times C[i]

        C<-numeric(n)
        for (i in 1:n)
        {
                u<-runif(1, min=0, max=2*uniform_mean)
                w_scale<-w_scale_init*exp(-(1/w_shape)*log(u*exp(b%*%Z[i,])))
                C[i]<-rweibull(1, w_shape, w_scale)
                                                                        }
        w_scale<-w_scale_init

#=============================================
#Generation of X[i]=min{T[i], C[i]}

        X<-numeric(n)
        for (i in 1:n)
        {
                X[i]<-min(T[i], C[i])
                                                        }

#=============================================
#Generation of the censoring indicator D[i]

        D<-numeric(n)
        for (i in 1:n)
        {if (T[i]<=C[i]) {D[i]<-1}
                else {D[i]<-0}
                                }
```

```
#=============================================
#Calculation of censoring proportion

      per_cent_censor<-1-mean(D)}     #end of desired per cent censoring proportion
                                      #loop

censor_vector[iter]<-per_cent_censor

#=============================================
#Results Matrix

Res_Matrix<-cbind(T, C, X, D, Z)
Res_Matrix
...}
```

4.2.6. Κατασκευή του κώδικα: Κατασκευή της συνάρτησης πιθανοφάνειας μέσω της συνάρτησης μετασχηματισμού: Για την κατασκευή της εξίσωσης **(4.12)**

$$\ell = \ln L = \sum_{i=1}^{n} \left\{ D_i \left[\ln G' \left(H_0(X_i \mid \mathbf{z}_i) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i} \right) + \ln h_0(X_i \mid \mathbf{z}_i) + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i \right] - G\left(H_0(X_i \mid \mathbf{z}_i) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i} \right) \right\}$$

που δίνει τη λογαριθμοποιημένη πιθανοφάνεια ως συνάρτηση των $\boldsymbol{\beta}$, α, λ, εργαζόμαστε ως εξής:

- Ονομάζουμε *param* το διάνυσμα των παραμέτρων – μεταβλητών της συνάρτησης ℓ. Αν το μοντέλο που εξετάζουμε έχει p το πλήθος συμμεταβλητές, τότε το διάνυσμα *param* είναι διάστασης $p+2$, οι πρώτες p συντεταγμένες του είναι οι συντελεστές $\boldsymbol{\beta}$ των συμμεταβλητών, η συντεταγμένη τάξης $p+1$ είναι η παράμετρος κλίμακας α της Weibull και η συντεταγμένη τάξης $p+2$ είναι η παράμετρος σχήματος λ της Weibull δηλ. $param = (\beta_1, \ldots, \beta_p, \alpha, \lambda)$.
- Στη συνέχεια, με μία επαναληπτική διαδικασία, κατασκευάζουμε τους όρους του αθροίσματος της εξίσωσης **(4.12)** και κάθε όρο τον προσθέτουμε στο άθροισμα των προηγούμενων, έως ότου υπολογισθεί η τιμή της συνάρτησης ℓ

(μεταβλητή *log_lik* στον κώδικα). Σε κάθε επανάληψη $i = 1, 2, \ldots, n$ αυτής της διαδικασίας, εμφανίζεται μία βοηθητική μεταβλητή *prod*, στην οποία αποθηκεύεται το εσωτερικό γινόμενο $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_i$, καθώς επίσης και μία βοηθητική μεταβλητή *m* στην οποία αποθηκεύεται η τιμή $H_0(X_i \mid \boldsymbol{z}_i) \cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_i}$.

- Σημειωτέον εδώ ότι, επειδή για τη μεγιστοποίηση της συνάρτησης ℓ θα χρησιμοποιήσουμε την έτοιμη συνάρτηση *constr_Optim* της R, η οποία όμως, ελαχιστοποιεί συναρτήσεις, ζητάμε από τον κώδικά μας να μας επιστρέψει ως αποτέλεσμα την αντίθετη τιμή: – *log_lik* (μεγιστοποίηση της ℓ ισοδυναμεί με ελαχιστοποίηση της $-\ell$).

```
{...
#=============================================
#Making of the log-likelihood function via G function

log_likelihood<-function(param)
{       beta<-param[1:p]
        scale<-param[p+1]
        shape<-param[p+2]

        log_lik<-0
                for (i in 1:n)
                {       prod<-0
                        for (j in 1:p)      {prod<-prod+beta[j]*zeta[i,j]}
                        m<-exp(prod)*H(X[i],scale, shape)
                        v<-D[i]*(prod+log(dG(m))+log(h(X[i],scale, shape)))-G(m)
                        log_lik<-log_lik+v                          }
        return(-log_lik)
                                }
...}
```

4.2.7. Κατασκευή του κώδικα: Προετοιμασία της βελτιστοποίησης και της επιλογής μοντέλου: Έχοντας τις *k* το πλήθος συμμεταβλητές, θα

προσαρμόσουμε και θα πάρουμε αποτελέσματα από όλα τα δυνατά μοντέλα που μπορούμε να σχηματίσουμε με αυτές.

- Χρησιμοποιώντας απλή Συνδυαστική, το πλήθος των δυνατών μοντέλων που μπορούμε να κατασκευάσουμε με p από τις k συμμεταβλητές είναι ίσο με $\binom{k}{p}$, $p = 1, 2, \ldots, k$, άρα το πλήθος όλων των δυνατών μοντέλων είναι ίσο με $\sum_{p=1}^{k}\binom{k}{p} = 2^k - 1$. Για να αποθηκεύσουμε τα αποτελέσματα, κατασκευάζουμε έναν πίνακα, διάστασης $(2k+10)\times\left(2^k-1\right)$, στις στήλες του οποίου θα αποθηκεύονται τα αποτελέσματα από κάθε μοντέλο. Ο πίνακας αυτός, έχει στον κώδικα το όνομα *results* και η δομή μίας στήλης του είναι η εξής:

$$\begin{bmatrix} \hat{\beta}_1 \text{ (our method)} \\ \vdots \\ \hat{\beta}_k \text{ (our method)} \\ \hat{\alpha} \text{ (our method)} \\ \hat{\lambda} \text{ (our method)} \\ AIC \text{ (our method)} \\ BIC \text{ (our method)} \\ Error1 \text{ (our method)} \\ Error2 \text{ (our method)} \\ \hat{\beta}_1 \text{ (coxph)} \\ \vdots \\ \hat{\beta}_k \text{ (coxph)} \\ AIC \text{ (coxph)} \\ BIC \text{ (coxph)} \\ Error1 \text{ (coxph)} \\ Error2 \text{ (coxph)} \end{bmatrix}$$

Επίσης, οι στήλες του πίνακα *results* εμφανίζουν τα μοντέλα σε ομάδες με ίδιο αριθμό συμμεταβλητών και σε κάθε ομάδα κατά σειρά συμμεταβλητών, δηλ. οι πρώτες $\binom{k}{1}$ το πλήθος στήλες εμφανίζουν την ομάδα μοντέλων $(Z_1, Z_2, \ldots, Z_k)$ με μία συμμεταβλητή, οι επόμενες $\binom{k}{2}$ στήλες εμφανίζουν την

ομάδα μοντέλων $(Z_1Z_2, Z_1Z_3, \ldots, Z_1Z_k, Z_2Z_3, \ldots, Z_2Z_k, \ldots, Z_{k-1}Z_k)$ με δύο συμμεταβλητές κ.ο.κ. Αν p το πλήθος συμμεταβλητών μίας τέτοιας ομάδας ($p = 1, 2, \ldots, k$), η αρχική στήλη *lower_column* μίας τέτοιας ομάδας καθορίζεται από τον αναδρομικό τύπο $(lower\_column)_1 = 1$ και $(lower\_column)_p = (lower\_column)_{p-1} + \binom{k}{p-1}$ για $p = 2, 3, \ldots, k$. Π.χ. για $k = 4$ συνολικό πλήθος συμμεταβλητών, η ομάδα μοντέλων με $p = 2$ το πλήθος συμμεταβλητές έχει ως αρχική στήλη την $(lower\_column)_2 = (lower\_column)_1 + \binom{4}{1} = 1 + 4 = 5$ και τελική στήλη την $(lower\_column)_3 - 1 = (lower\_column)_2 + \binom{4}{2} - 1 = 5 + 6 - 1 = 10$, άρα καταλαμβάνει τις στήλες 5 έως 10 του πίνακα *results* κατά τη μορφή:

$$\begin{array}{c} \text{στήλη} \\ \text{συμμεταβλητές} \end{array} \begin{bmatrix} \ldots & 5 & 6 & 7 & 8 & 9 & 10 & \ldots \\ \ldots & Z_1Z_2 & Z_1Z_3 & Z_1Z_4 & Z_2Z_3 & Z_2Z_4 & Z_3Z_4 & \ldots \end{bmatrix}$$

- Έχοντας κατασκευάσει τον πίνακα *results* αποθήκευσης των αποτελεσμάτων, ξεκινούμε τις προσαρμογές, ζητώντας από το πρόγραμμα να προσαρμόσει πρώτα τα μοντέλα με 1 συμμεταβλητή ($Z_1, Z_2, \ldots, Z_k$), μετά με 2 συμμεταβλητές $(Z_1Z_2, Z_1Z_3, \ldots, Z_1Z_k, Z_2Z_3, \ldots, Z_2Z_k, \ldots, Z_{k-1}Z_k)$ κ.λ.π. Έστω ότι θέλουμε να προσαρμόσουμε τα μοντέλα με p συμμεταβλητές. Για να γίνει η επιλογή αυτών, δημιουργούμε τον πίνακα *idx* στις στήλες του οποίου εμφανίζονται οι συνδυασμοί των k το πλήθος συμμεταβλητών ανά p, ώστε επιλέγοντας τα στοιχεία κάθε στήλης, αυτά να μας δίνουν το μοντέλο που θα προσαρμοστεί. Π.χ. για $k = 4$ συνολικό πλήθος συμμεταβλητών και $p = 2$, ο πίνακας *idx* είναι:

$$idx = \begin{bmatrix} 1 & 1 & 1 & 2 & 2 & 3 \\ 2 & 3 & 4 & 3 & 4 & 4 \end{bmatrix}$$

και οι στήλες του δηλώνουν τα προς προσαρμογή μοντέλα που είναι τα $Z_1Z_2, Z_1Z_3, Z_1Z_4, Z_2Z_3, Z_2Z_4, Z_3Z_4$. Επομένως, κατά τον υπολογισμό του εσωτερικού γινομένου $\boldsymbol{\beta}^T\boldsymbol{z}$, ως $\boldsymbol{z}$ θα είναι ο πίνακας που θα δημιουργείται

από τις στήλες του ήδη κατασκευασμένου τυχαίου πίνακα Z, τις οποίες υποδεικνύει ο πίνακας *idx*.

- Η μεγιστοποίηση της λογαριθμοποιημένης πιθανοφάνειας (ισοδύναμα ελαχιστοποίηση της αντίθετής της), γίνεται με τη βοήθεια της έτοιμης συνάρτησης *constr_Optim* που παρέχει η R για την ελαχιστοποίηση συναρτήσεων υπό περιορισμούς. Εδώ, επειδή η λογαριθμοποιημένη πιθανοφάνεια είναι η $\ell(\boldsymbol{\beta}, \alpha, \lambda)$, οι περιορισμοί που πρέπει να τεθούν αφορούν τις παραμέτρους της Weibull και είναι $\alpha > 0$ και $\lambda > 0$. Επίσης, επειδή η ελαχιστοποίηση της $-\ell$ θα γίνει με αριθμητικές μεθόδους, οφείλουμε να υποδείξουμε το διάνυσμα *initial_vector* των αρχικών τιμών που θα δοθούν στις μεταβλητές $\boldsymbol{\beta}, \alpha, \lambda$. Προς τούτο, επιλέξαμε σε κάθε συντεταγμένη του διανύσματος *initial_vector*, να δώσουμε μία τυχαία τιμή από την Ομοιόμορφη κατανομή με μέση τιμή την πραγματική τιμή της συντεταγμένης αυτής. Π.χ. αν η πραγματική τιμή της παραμέτρου σχήματος είναι $\lambda = 2$, τότε στην $(p+2)-$ συντεταγμένη του *initial_vector*, δίνουμε μία τυχαία τιμή από την κατανομή $U((0,\ 4))$.
- Οι περιορισμοί $\alpha > 0$ και $\lambda > 0$ που αφορούν τις παραμέτρους της Weibull, δηλώνονται στη συνάρτηση *constrOptim* μέσω των πινάκων *ui* και *ci*, υπό μορφή γραμμικού συστήματος ανισώσεων. Συγκεκριμένα, αν $(\boldsymbol{\beta}, \alpha, \lambda) = (\beta_1, \ldots, \beta_p, \alpha, \lambda)$ είναι το διάνυσμα των προς εκτίμηση παραμέτρων, οι περιορισμοί που θέλουμε αντιστοιχούν στο σύστημα γραμμικών ανισώσεων $\begin{cases} 0\cdot\beta_1 + \ldots + 0\cdot\beta_p + 1\cdot\alpha + 0\cdot\lambda > 0 \\ 0\cdot\beta_1 + \ldots + 0\cdot\beta_p + 0\cdot\alpha + 1\cdot\lambda > 0 \end{cases}$ και αρκεί να δηλώσουμε ότι *ui* είναι ο $2\times(p+2)$ πίνακας των συντελεστών $ui = \begin{bmatrix} 0 & \ldots & 0 & 1 & 0 \\ 0 & \ldots & 0 & 0 & 1 \end{bmatrix}$ και *ci* το διάνυσμα των σταθερών όρων του β′ μέλους $ci = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ (η συνάρτηση *constrOptim* θεωρεί πάντα ότι μεταξύ α′ και β′ μέλους υπάρχει ανισότητα “>”).

- Τέλος, δημιουργούμε και ένα διάνυσμα με όνομα *row*, του οποίου τα στοιχεία θα δηλώνουν σε ποια γραμμή του πίνακα *results* θα τοποθετείται η εκτιμημένη παράμετρος

```
{...
#================================================
#Optimization of likelihood - Selection of Variables

results<-matrix(nrow=2*k+10, ncol=2^k-1)      #Storage of coefficients, AIC, BIC of the
                                              #2^k-1 models

for (i in 1:k)      {results[i,]<-0
                     results[k+6+i,]<-0}

#Start of Variable Selection

for (p in 1:k)      #p is the number of covariates of fitted model
{

        #Column-group of results matrix, where estimations will be stored
        lower_column<-0
        for (u in 0:(p-1)) {lower_column<-lower_column+choose(k,u)}

        initial_vector<-numeric(p+2)    #Initial vector with starting values for minimization

        idx<-matrix(combn(k,p),nrow=p)          #elements of the matrix are the combinations
                                                #of k by p
        row<-numeric(p)          #vector which shows the rows of the results matrix
        formula_constr<-numeric(p) #Will be used for creatimg the coxph string formula
        id<-seq(1:n)
        zeta<-matrix(nrow=n, ncol=p)  #Matrix which contains values of covariates that are
                                      #used

        for (j in 1:ncol(idx))
                {       results_column_number<-lower_column+j-1
                        for (i in 1:p)
```

```
                    {element<-idx[i,j]      #the (i,j)-element of idx matrix
                    zeta[,i]<-Z[,element]   #i-th column of zeta is the element-
                                            #column of Z
                    row[i]<-element
                    initial_vector[i]<-runif(1, min=2*min(0,b[,element]),
                                            max=2*max(0,b[,element]))
                    formula_constr[i]<-element}     #end for (i in 1:p)

        initial_vector[p+1]<-runif(1, min=0, max=2*w_scale_init)
        initial_vector[p+2]<-runif(1, min=0, max= 2*w_shape)

        null_matrix<-matrix(rep(0, 2*p), nrow=2, ncol=p)
        ui<-matrix(nrow=2, ncol=p+2)
        ui<-cbind(null_matrix, diag(2))
        ci<-c(0, 0)
        …}
...}
...}
```

4.2.8. Κατασκευή του κώδικα: Βελτιστοποίηση και εκτίμηση των παραμέτρων μοντέλου: Μπορούμε τώρα να εκτιμήσουμε τις παραμέτρους των μοντέλων, να αποθηκεύσουμε τα αποτελέσματα των εκτιμήσεων στον πίνακα *results* και να επιλέξουμε το καλύτερο μοντέλο.

- Η εφαρμογή της συνάρτησης *constrOptim* με αρχικό διάνυσμα το *initial_vector*, μας επιστρέφει τις εκτιμήσεις των παραμέτρων $\boldsymbol{\beta}, \alpha, \lambda$. Από αυτές, μπορούμε να υπολογίσουμε την τιμή της λογαριθμοποιημένης πιθανοφάνειας $\hat{\ell}\left(\hat{\boldsymbol{\beta}}, \hat{\alpha}, \hat{\lambda}\right)$ και τις τιμές $\mathrm{AIC} = -2 \cdot \hat{\ell}\left(\hat{\boldsymbol{\beta}}, \hat{\alpha}, \hat{\lambda}\right) + 2p$ και $\mathrm{BIC} = -2 \cdot \hat{\ell}\left(\hat{\boldsymbol{\beta}}, \hat{\alpha}, \hat{\lambda}\right) + p \cdot \ln n$ των κριτηρίων AIC και BIC αντίστοιχα. Αμέσως μετά, οι εκτιμημένες τιμές των $\boldsymbol{\beta}, \alpha, \lambda$, AIC, BIC, αποθηκεύονται στα στοιχεία του πίνακα *results* που βρίσκονται στη γραμμή που υποδεικνύεται

από το διάνυσμα *row* και στη στήλη που βρίσκεται μεταξύ $(lower\_column)_p$ και $(lower\_column)_p + \binom{k}{p} - 1$ και "φιλοξενεί" το προς εκτίμηση μοντέλο.

- Στη συνέχεια, το πρόγραμμα δημιουργεί τις μεταβλητές *important* και *not_important* με τις οποίες αναγνωρίζονται οι σημαντικές και μη σημαντικές μεταβλητές αντίστοιχα του μοντέλου, καθώς και τις μεταβλητές *error1_string* και *error2_string*, που εξετάζονται ως προς το αληθές ή το ψευδές για να δοθεί στις παραμέτρους *Error*1 (Our Method) και *Error*2 (Our Method) η τιμή 1 ή 0 αντίστοιχα.
- Ακολουθεί η προσαρμογή του μοντέλου με τη συνάρτηση coxph, και η εκτίμηση μέσω αυτής των συντελεστών του μοντέλου και των σφαλμάτων *Error*1 (Coxph Method) και *Error*2 (Coxph Method).
- Κατόπιν όλων των ανωτέρω, συμπληρώνεται ο πίνακας *results* και ζητάμε από το πρόγραμμα την εκτύπωσή του.

```
{...
{...
        {...
                opt<-constrOptim(initial_vector, log_likelihood, grad=NULL, ui=ui,
                                ci=ci,method="Nelder-Mead",control=list(maxit=500))
                l<-log_likelihood(opt$par)
                AIC<-2*l+2*p
                BIC<-2*l+p*log(n)

                for (m in 1:p){results[row[m],results_column_number]<-opt$par[m]}
                results[k+1,results_column_number]<-opt$par[p+1]
                results[k+2,results_column_number]<-opt$par[p+2]
                results[k+3,results_column_number]<-AIC
                results[k+4,results_column_number]<-BIC

                #Calculation of Error 1(: Not important variables are selected in model) in
                #Our Method
                #and Calculation of Error 2(: At least one of important variables is not
                #selected in model) in Our Method
```

```
important<-which(b[1,]!=0,arr.ind=TRUE)       #vector which contains
                                                #indexes of important variables
not_important<-which(b[1,]==0,arr.ind=TRUE)          #vector which
                                #contains indexes of NOT important variables

error1_string<-
      as.vector(paste("results[",not_important[1],",results_column_number]!=0"))
                if (length(not_important)!=1){
                        for (i in 2:length(not_important))
                                {error1_string<-
as.vector(paste(error1_string," | results[",not_important[i],",results_column_number]!=0"))
}#end for
                }          #end if

error2_string<-
paste("results[",important[1],",results_column_number]==0")
                if (length(important)!=1){
                        for (i in 2:length(important))
                                {error2_string<-
paste(error2_string," | results[",important[i],",results_column_number]==0")}#end for
                }                  #end if

if (eval(parse(text=error1_string)))
        {results[k+5,results_column_number]<-1} else
        {results[k+5,results_column_number]<-0} #Error1

if (eval(parse(text=error2_string)))
        {results[k+6,results_column_number]<-1} else
        {results[k+6,results_column_number]<-0} #Error2


#Fit models with coxph function

formula_string_initial<-paste("Surv(X,D)~Z[,", formula_constr[1],"]")
```

```
data_string_initial<-paste("list(X,D,")
{if (p==1){formula_string<-paste(formula_string_initial,"+frailty(id,sparce=TRUE,method='em')")}
if (p==2){formula_string<-paste(formula_string_initial,"+Z[,",formula_constr[2],"]+frailty(id,sparce=TRUE,method='em')")}
if (p>2)
{ formula_string<-formula_string_initial
{for (u in 2:(p-1)) {formula_string<-paste(formula_string,"+Z[,",formula_constr[u],"]")}
for (u in (p:p)) {formula_string<-paste(formula_string,"+Z[,", formula_constr[u], "]+frailty(id,sparce=TRUE,method='em',)")}}
} #end if (p>2)

data_string<-data_string_initial
for (u in 1:(k-1)) {data_string<-paste(data_string,"Z[,", u, "],")}
for (u in k:k) {data_string<-paste(data_string,"Z[,", u, "],id)")}}

coxph_formula<-as.formula(formula_string)
coxph_data<-as.data.frame(data_string)

cox_fit<-coxph(coxph_formula,data=coxph_data)

for (m in 1:p){results[k+6+row[m],results_column_number]<-cox_fit$coef[m]}
results[2*k+7,results_column_number]<-extractAIC(cox_fit)[2]
results[2*k+8,results_column_number]<-extractAIC(cox_fit, k = log(n))[2]

#Calculation of Error 1(: Not important variables are selected in model) in
#Coxph Method
#and Calculation of Error 2(: At least one of important variables is not
#selected in model) in Coxph Method

error1_string<-
as.vector(paste("results[",k+6+not_important[1],",results_column_number]!=0"))
```

```
                if (length(not_important)!=1){
                        for (i in 2:length(not_important))
                        {error1_string<-
as.vector(paste(error1_string," | results[",k+6+not_important[i],",results_column_number]!
=0"))}#end for
                        }       #end if

                error2_string<-
paste("results[",k+6+important[1],",results_column_number]==0")
                if (length(important)!=1){
                        for (i in 2:length(important))
                        {error2_string<-
paste(error2_string," | results[",k+6+important[i],",results_column_number]==0")}#end for
                        }       #end if

                if (eval(parse(text=error1_string)))
                        {results[2*k+9,results_column_number]<-1} else
                        {results[2*k+9,results_column_number]<-0} #Error1

                if (eval(parse(text=error2_string)))
                        {results[2*k+10,results_column_number]<-1} else
                        {results[2*k+10,results_column_number]<-0} #Error2


                results_name<-1:(2*k+10)
                for (i in 1:k)
                        {results_name[i]<-paste("est. coef. β(", i,") (Our Method)")
                        results_name[k+6+i]<-paste("est. coef. β(", i,") (Coxph Method)")}
                        results_name[k+1]<-"est. w_scale (Our Method)"
                        results_name[k+2]<-"est. w_shape (Our Method)"
                        results_name[k+3]<-"est. AIC (Our Method)"
                        results_name[k+4]<-"est. BIC (Our Method)"
                        results_name[k+5]<-"Error 1 (Boolean 0-1) (Our Method)"
                        results_name[k+6]<-"Error 2 (Boolean 0-1) (Our Method)"
                        results_name[2*k+7]<-paste("est. AIC (Coxph Method)")
```

```
                    results_name[2*k+8]<-paste("est. BIC (Coxph Method)")
                    results_name[2*k+9]<-"Error 1 (Boolean 0-1) (Coxph Method)"
                    results_name[2*k+10]<-"Error 2 (Boolean 0-1) (Coxph Method)"



            rownames(results)<-results_name
        } #end for (j in 1:ncol(idx))

}       #End of Variable Selection

cat("=========================================================="
, "\n")
cat("Iteration no.", iter, "- Censoring proportion:", per_cent_censor, "\n")
cat("Iteration no.", iter," - Matrix of ALL subsets", "\n")
print(results)
cat(" ","\n")
...}
```

4.2.9. Κατασκευή του κώδικα: Επιλογή μοντέλου: Από τον πίνακα *results* είμαστε πλέον σε θέση να επιλέξουμε το βέλτιστο μοντέλο βάσει του κριτηρίου AIC ή BIC και της μεθόδου μας ή της coxph. Π.χ. για να επιλέξουμε το βέλτιστο μοντέλο με τη μέθοδό μας και το κριτήριο AIC, αρκεί να δηλώσουμε ποιο είναι το ελάχιστο στοιχείο της $(k+3)-$ γραμμής του πίνακα *results* (αφού στην $(k+3)-$ γραμμή αποθηκεύεται η τιμή του AIC).

- Ονομάζοντας *our_AIC_index* το ζεύγος που δίνει τη γραμμή και στήλη στην οποία βρίσκεται το μικρότερο AIC, μπορούμε να πάρουμε τα στοιχεία αυτής της στήλης που αφορούν στη μέθοδό μας και να τα αντιγράψουμε στη στήλη της επανάληψης *iter* του πίνακα *our_AIC_iter_results*.
- Τα ίδια επαναλαμβάνονται για τη μέθοδό μας και το κριτήριο BIC, οπότε συμπληρώνεται η *iter*-στήλη του πίνακα *our_BIC_iter_results* και ακολούθως για την *coxph* και τα δύο κριτήρια, οπότε συμπληρώνονται οι *iter*-στήλες των πινάκων *coxph_AIC_iter_results* και *coxph_BIC_iter_results*.

- Στο σημείο αυτό ολοκληρώνεται η τρέχουσα $(iter)-$επανάληψη και το πρόγραμμα προχωρά στην επόμενη $((iter+1)-)$επανάληψη, ενώ ζητούμε και την εκτύπωση των 4 πινάκων με τα βέλτιστα μοντέλα.

```
{...
#Fill the matrices with the best models of our method

our_AIC_index<-which(results == min(results[k+3,]), arr.ind = TRUE) #Finds the model with the min. AIC
        for (i in 1:(k+3)){our_AIC_iter_results[i,iter]<-results[i,our_AIC_index[2]]}
        our_AIC_iter_results[k+4,iter]<-results[k+5,our_AIC_index[2]]
        our_AIC_iter_results[k+5,iter]<-results[k+6,our_AIC_index[2]]
        if (isTRUE(our_AIC_iter_results[k+4,iter]==1 && our_AIC_iter_results[k+5,iter]==1)) {
                our_AIC_iter_results[k+6,iter]<-1}
        else {our_AIC_iter_results[k+6,iter]<-0}
        our_AIC_iter_results[k+7,iter]<-per_cent_censor

our_BIC_index<-which(results == min(results[k+4,]), arr.ind = TRUE) #Finds the model with the min. BIC
        for (i in 1:(k+2)) {our_BIC_iter_results[i,iter]<-results[i,our_BIC_index[2]]}
        our_BIC_iter_results[k+3,iter]<-results[k+4,our_BIC_index[2]]
        our_BIC_iter_results[k+4,iter]<-results[k+5,our_BIC_index[2]]
        our_BIC_iter_results[k+5,iter]<-results[k+6,our_BIC_index[2]]
        if (isTRUE(our_BIC_iter_results[k+4,iter]==1 && our_BIC_iter_results[k+5,iter]==1)) {
                our_BIC_iter_results[k+6,iter]<-1}
        else {our_BIC_iter_results[k+6,iter]<-0}
        our_BIC_iter_results[k+7,iter]<-per_cent_censor

#Row names of the matrices above
our_AIC_iter_name<-1:(k+7)
        for (i in 1:k){our_AIC_iter_name[i]<-paste("est. coef. β(", i,") (Our Method)")}
        our_AIC_iter_name[k+1]<-"est. w_scale (Our Method)"
        our_AIC_iter_name[k+2]<-"est. w_shape (Our Method)"
```

```
        our_AIC_iter_name[k+3]<-"est. AIC (Our Method)"
        our_AIC_iter_name[k+4]<-"Error 1 (Boolean 0-1) (Our Method)"
        our_AIC_iter_name[k+5]<-"Error 2 (Boolean 0-1) (Our Method)"
        our_AIC_iter_name[k+6]<-"Error 1 & 2 (Boolean 0-1) (Our Method)"
        our_AIC_iter_name[k+7]<-"Per Cent Censor"

our_BIC_iter_name<-1:(k+7)
        for (i in 1:k){our_BIC_iter_name[i]<-paste("est. coef. β(", i,") (Our Method)")}
        our_BIC_iter_name[k+1]<-"est. w_scale (Our Method)"
        our_BIC_iter_name[k+2]<-"est. w_shape (Our Method)"
        our_BIC_iter_name[k+3]<-"est. BIC (Our Method)"
        our_BIC_iter_name[k+4]<-"Error 1 (Boolean 0-1) (Our Method)"
        our_BIC_iter_name[k+5]<-"Error 2 (Boolean 0-1) (Our Method)"
        our_BIC_iter_name[k+6]<-"Error 1 & 2 (Boolean 0-1) (Our Method)"
        our_BIC_iter_name[k+7]<-"Per Cent Censor"

        rownames(our_AIC_iter_results)<-our_AIC_iter_name
        rownames(our_BIC_iter_results)<-our_BIC_iter_name



#Fill the matrices with the best models of coxph method

coxph_AIC_index<-which(results == min(results[2*k+7,]), arr.ind = TRUE)     #Finds the
model with the min. AIC
for (i in 1:(k+1)) {coxph_AIC_iter_results[i,iter]<-results[k+6+i,coxph_AIC_index[2]]}
coxph_AIC_iter_results[k+2,iter]<-results[2*k+9,coxph_AIC_index[2]]
coxph_AIC_iter_results[k+3,iter]<-results[2*k+10,coxph_AIC_index[2]]
if                    (isTRUE(coxph_AIC_iter_results[k+2,iter]==1                    &&
coxph_AIC_iter_results[k+3,iter]==1)) {
                coxph_AIC_iter_results[k+4,iter]<-1}
        else {coxph_AIC_iter_results[k+4,iter]<-0}
coxph_AIC_iter_results[k+5,iter]<-per_cent_censor

coxph_BIC_index<-which(results == min(results[2*k+8,]), arr.ind = TRUE)     #Finds the
model with the min. BIC
```

```
for (i in 1:k) {coxph_BIC_iter_results[i,iter]<-results[k+6+i,coxph_BIC_index[2]]}
coxph_BIC_iter_results[k+1,iter]<-results[2*k+8,coxph_BIC_index[2]]
coxph_BIC_iter_results[k+2,iter]<-results[2*k+9,coxph_BIC_index[2]]
coxph_BIC_iter_results[k+3,iter]<-results[2*k+10,coxph_BIC_index[2]]
if                    (isTRUE(coxph_BIC_iter_results[k+2,iter]==1                    &&
coxph_BIC_iter_results[k+3,iter]==1)) {
                coxph_BIC_iter_results[k+4,iter]<-1}
        else {coxph_BIC_iter_results[k+4,iter]<-0}
coxph_BIC_iter_results[k+5,iter]<-per_cent_censor

coxph_AIC_iter_name<-1:(k+5)
        for (i in 1:k){coxph_AIC_iter_name[i]<-paste("est. coef. β(", i,") (Coxph Method)")}
        coxph_AIC_iter_name[k+1]<-"est. AIC (Coxph Method)"
        coxph_AIC_iter_name[k+2]<-"Error 1 (Boolean 0-1) (Coxph Method)"
        coxph_AIC_iter_name[k+3]<-"Error 2 (Boolean 0-1) (Coxph Method)"
        coxph_AIC_iter_name[k+4]<-"Error 1 & 2 (Boolean 0-1) (Coxph Method)"
        coxph_AIC_iter_name[k+5]<-"Per Cent Censor"

coxph_BIC_iter_name<-1:(k+5)
        for (i in 1:k){coxph_BIC_iter_name[i]<-paste("est. coef. β(", i,") (Coxph Method)")}
        coxph_BIC_iter_name[k+1]<-"est. BIC (Coxph Method)"
        coxph_BIC_iter_name[k+2]<-"Error 1 (Boolean 0-1) (Coxph Method)"
        coxph_BIC_iter_name[k+3]<-"Error 2 (Boolean 0-1) (Coxph Method)"
        coxph_BIC_iter_name[k+4]<-"Error 1 & 2 (Boolean 0-1) (Coxph Method)"
        coxph_BIC_iter_name[k+5]<-"Per Cent Censor"

        rownames(coxph_AIC_iter_results)<-coxph_AIC_iter_name
        rownames(coxph_BIC_iter_results)<-coxph_BIC_iter_name

}       #End of Iterations

cat("============================================================"
, "\n")
cat("Matrices of BEST subsets", "\n")
```

our_AIC_iter_results

our_BIC_iter_results

coxph_AIC_iter_results

coxph_BIC_iter_results

4.2.10. Κατασκευή του κώδικα: Πίνακες αποτελεσμάτων: Απομένει τώρα να κατασκευάσουμε τον τελικό πίνακα αποτελεσμάτων.

- Από κάθε έναν από τους 4 πίνακες *our_AIC_iter_results*, *our_BIC_iter_results*, *coxph_AIC_iter_results*, *coxph_BIC_iter_results* μπορούμε να εξάγουμε το μέσο όρο και την τυπική απόκλιση των εκτιμητριών για κάθε εκτιμώμενη παράμετρο. Ο μέσος όρος του συντελεστή μίας σημαντικής μεταβλητής υπολογίζεται από τα αποτελέσματα του εν λόγω πίνακα, στα οποία ο συντελεστής πήρε τιμή $\neq 0$. Π.χ., αν η μεταβλητή z_1 είναι σημαντική και σε $niter = 3$ επαναλήψεις, η αντίστοιχη γραμμή του πίνακα *our_AIC_iter_results* εμφάνισε τα αποτελέσματα $[2, 0, 4]$, τότε ο μέσος όρος του συντελεστή β_1 είναι $\beta_1 = \frac{2+4}{2} = 3$ και αντίστοιχα για την τυπική απόκλιση. Αντίθετα, για μη-σημαντική μεταβλητή, ο μέσος όρος και η τυπική απόκλιση υπολογίζονται από τα αποτελέσματα όλων των επαναλήψεων, π.χ. αν η προηγούμενη μεταβλητή z_1 δεν είναι σημαντική, τότε ως μέσος όρος του συντελεστή β_1 είναι ο $\beta_1 = \frac{2+4}{3} = 2$.
- Ο μέσος όρος και η τυπική απόκλιση των παραμέτρων κλίμακας α και σχήματος λ της Weibull, υπολογίζονται αντίστοιχα επί του συνόλου *niter* των επαναλήψεων.
- Ο τελικός πίνακας αποτελεσμάτων *final_matrix* έχει διάσταση $5 \times (2k + 11)$. Στις 2 πρώτες γραμμές εμφανίζονται αντίστοιχα τα αποτελέσματα από τη μέθοδό μας μέσω κριτηρίου AIC και κριτηρίου BIC, στις επόμενες 2 γραμμές τα αποτελέσματα από την coxph μέσω κριτηρίου AIC και κριτηρίου BIC και στην 5η γραμμή οι πραγματικές τιμές των συντελεστών. Στη στήλη 1 εμφανίζεται το μέγεθος n του δείγματος, στη στήλη 2 ο μέσος

όρος του ποσοστού λογοκρισίας, στη στήλη 3 ο αριθμός επιτυχιών (ορθή επιλογή του μοντέλου), στη στήλη 4 το ποσοστό επιτυχιών (επί του συνόλου επαναλήψεων), στις στήλες 5, 6 και 7 οι μέσοι όροι των $Error1$, $Error2$ και $Error1\&2$ αντίστοιχα, στις επόμενες πρώτες $2k$ το πλήθος στήλες εμφανίζονται ανά συμμεταβλητή οι μέσοι όροι και οι τυπικές αποκλίσεις των συντελεστών $\boldsymbol{\beta}$, και στις τελευταίες 4 στήλες οι μέσοι όροι και τυπικές αποκλίσεις των παραμέτρων κλίμακας και σχήματος της Weibull (μόνο για τη μέθοδό μας).

- Δίνεται και η δυνατότητα να εξάγουμε τον πίνακα *final_matrix* σε μορφή αρχείου κειμένου (*.txt*) ώστε να επεξεργαστούμε τα αποτελέσματα σε άλλο πρόγραμμα όπως σε πρόγραμμα λογιστικών φύλλων, επεξεργασίας κειμένου κ.λ.π.

```
#Calculation of means and standard deviations of coefficients in matrices of results
#(function with variables the coefficient and the matrix)

coef_mean_sd<-function(coef, matr)
{sum1<-0
sum2<-0

if (b[1,coef]!=0)          #coef refers to important variable
{counter<-0
 for (j in 1:ncol(matr))
        {       if (matr[coef,j]!=0)
                {sum1<-sum1+as.double(matr[coef,j])
                sum2<-sum2+as.double(matr[coef,j]^2)
                counter<-counter+1}
                                                }

coeff_mean<-sum1/counter
coeff_st_dev<-as.double(sqrt((sum2-sum1^2/counter)/(counter-1)))}

else                                #coef refers to NOT important variable
{for (j in 1:ncol(matr))
```

```
                    {sum1<-sum1+as.double(matr[coef,j])
                    sum2<-sum2+as.double(matr[coef,j]^2)}

coeff_mean<-sum1/ncol(matr)
coeff_st_dev<-as.double(sqrt((sum2-sum1^2/ncol(matr))/(ncol(matr)-1)))}

return(as.numeric(list(coeff_mean,coeff_st_dev)))
                                                    }



#Calculation of means & standard deviations of Weibull parameters
#(function with variables the parameter and the matrix)

weibull_mean_sd<-function(coef, matr)
{sum1<-0
sum2<-0

for (j in 1:ncol(matr))
                    {sum1<-sum1+as.double(matr[coef,j])
                    sum2<-sum2+as.double(matr[coef,j]^2)}

coeff_mean<-sum1/ncol(matr)
coeff_st_dev<-as.double(sqrt((sum2-sum1^2/ncol(matr))/(ncol(matr)-1)))

return(as.numeric(list(coeff_mean,coeff_st_dev)))
                                                    }



#Creation of final matrix (contains all the results)

final_matrix<-matrix(nrow=5, ncol=2*k+11)

for (i in 1:4)      {final_matrix[i,1]<-n                                   #Sample size
                            final_matrix[i,2]<-mean(censor_vector)}         #average   censoring
proportion
```

```
#Row 1 of final matrix contains the results of AIC criterion in Our Method
final_matrix[1,3]<-niter-sum(our_AIC_iter_results[k+4,])-
sum(our_AIC_iter_results[k+5,])+sum(our_AIC_iter_results[k+6,])      #number of success
(correct selection of model)
final_matrix[1,4]<-final_matrix[1,3]/niter #percentage of success
final_matrix[1,5]<-mean(our_AIC_iter_results[k+4,])   #Error1
final_matrix[1,6]<-mean(our_AIC_iter_results[k+5,])   #Error2
final_matrix[1,7]<-mean(our_AIC_iter_results[k+6,])   #Error1&2
for (i in 1:k)                                        #Means      &      Standard
Deviations of coef. β1, β2,...βk
{final_matrix[1,2*i+6]<-coef_mean_sd(i, our_AIC_iter_results)[1]
final_matrix[1,2*i+7]<-coef_mean_sd(i, our_AIC_iter_results)[2]}
for (i in (k+1):(k+2))                                #Means      &      Standard
Deviations of Weibull parameters
{final_matrix[1,2*i+6]<-weibull_mean_sd(i, our_AIC_iter_results)[1]
final_matrix[1,2*i+7]<-weibull_mean_sd(i, our_AIC_iter_results)[2]}


#Row 2 of final matrix contains the results of BIC criterion in Our Method
final_matrix[2,3]<-niter-sum(our_BIC_iter_results[k+4,])-
sum(our_BIC_iter_results[k+5,])+sum(our_BIC_iter_results[k+6,])      #number of success
(correct selection of model)
final_matrix[2,4]<-final_matrix[2,3]/niter #percentage of success
final_matrix[2,5]<-mean(our_BIC_iter_results[k+4,])   #Error1
final_matrix[2,6]<-mean(our_BIC_iter_results[k+5,])   #Error2
final_matrix[2,7]<-mean(our_BIC_iter_results[k+6,])   #Error1&2
for (i in 1:k)                                        #Means      &      Standard
Deviations of coef. β1, β2,...βk
{final_matrix[2,2*i+6]<-coef_mean_sd(i, our_BIC_iter_results)[1]
final_matrix[2,2*i+7]<-coef_mean_sd(i, our_BIC_iter_results)[2]}
for (i in (k+1):(k+2))                                #Means      &      Standard
Deviations of Weibull parameters
{final_matrix[2,2*i+6]<-weibull_mean_sd(i, our_BIC_iter_results)[1]
```

```
final_matrix[2,2*i+7]<-weibull_mean_sd(i, our_BIC_iter_results)[2]}



#Row 3 of final matrix contains the results of AIC criterion in Coxph Method
#(Weibull parameters are not estimated here)
final_matrix[3,3]<-niter-sum(coxph_AIC_iter_results[k+2,])-
sum(coxph_AIC_iter_results[k+3,])+sum(coxph_AIC_iter_results[k+4,])      #number of
success (correct selection of model)
final_matrix[3,4]<-final_matrix[3,3]/niter #percentage of success
final_matrix[3,5]<-mean(coxph_AIC_iter_results[k+2,])      #Error1
final_matrix[3,6]<-mean(coxph_AIC_iter_results[k+3,])      #Error2
final_matrix[3,7]<-mean(coxph_AIC_iter_results[k+4,])      #Error1&2
for (i in 1:k)                                             #Means & Standard
Deviations of coef. β1, β2,...βk
{final_matrix[3,2*i+6]<-coef_mean_sd(i, coxph_AIC_iter_results)[1]
final_matrix[3,2*i+7]<-coef_mean_sd(i, coxph_AIC_iter_results)[2]}

#Row 4 of final matrix contains the results of BIC criterion in Coxph Method
#(Weibull parameters are not estimated here)
final_matrix[4,3]<-niter-sum(coxph_BIC_iter_results[k+2,])-
sum(coxph_BIC_iter_results[k+3,])+sum(coxph_BIC_iter_results[k+4,])      #number of
success (correct selection of model)
final_matrix[4,4]<-final_matrix[4,3]/niter #percentage of success
final_matrix[4,5]<-mean(coxph_BIC_iter_results[k+2,])      #Error1
final_matrix[4,6]<-mean(coxph_BIC_iter_results[k+3,])      #Error2
final_matrix[4,7]<-mean(coxph_BIC_iter_results[k+4,])      #Error1&2
for (i in 1:k)                                             #Means &
Standard Deviations of coef. β1, β2,...βk
{final_matrix[4,2*i+6]<-coef_mean_sd(i, coxph_BIC_iter_results)[1]
final_matrix[4,2*i+7]<-coef_mean_sd(i, coxph_BIC_iter_results)[2]}

#Row 5 of final matrix contains the real values of coefficients
for (i in 1:k)
        {final_matrix[5,2*i+6]<-b[1,i]
        final_matrix[5,2*i+7]<-0}
```

```
final_matrix[5,2*k+8]<-w_scale_init
final_matrix[5,2*k+9]<-0.0
final_matrix[5,2*k+10]<-w_shape
final_matrix[5,2*k+11]<-0.0

#Rownames and Colnames of final_matrix
coef_labels<-1:(2*k)
for (i in 1:k) {coef_labels[2*i-1]<-paste("Mean of β(",i, ")")}
for (i in 1:k) {coef_labels[2*i]<-paste("St. Dev. of β(",i, ")")}
#coef_labels

colnames(final_matrix)<-c("Sample size (n)", "Average Censor (Percentage)",
                          "Num. of Success","Percentage of Success", "Error 1",
                          "Error 2","Error 1&2", coef_labels,"Est. Weibull scale param.",
                          " St. Dev. of Weibull scale param.", "Est. Weibull shape param.",
                          " St. Dev. of Weibull shape param.")
rownames(final_matrix)<-c("Our Method - AIC Criterion", "Our Method - BIC Criterion",
                                    "Coxph Method - AIC Criterion","Coxph Method -
BIC Criterion","Real values of coef.")

cat("censor vector","\n")
censor_vector
cat(" minimum censor proportion", min(censor_vector),"\n",
"maximum censor proportion", max(censor_vector),"\n",
"average censor proportion", mean(censor_vector),"\n",
"number of iterations", niter,"\n")

final_matrix

#Set working directory & title of file which will be exported as .txt
setwd("C:/Users/user/ Documents /Results")
title<-paste("Weibull results - niter =", niter,", censor =",mean(censor_vector), ".txt")

#Export to .txt file
write.table(final_matrix, file=title, sep=",")
```

4.2.11. Εκτέλεση του προγράμματος - Αποτελέσματα: Εκτελέσαμε το πρόγραμμα προσομοιώνοντας μοντέλα με 3, 4 και 5 συμμεταβλητές. Η προσομοίωση κάθε μοντέλου έγινε ώστε να πάρουμε ποσοστό λογοκρισίας χαμηλό ($\approx 10\%$), μεσαίο ($\approx 46\%$) και υψηλό ($\approx 80\%$) και κάθε φορά προσομοιώσαμε μικρό ($n = 50$) και μεγάλο δείγμα ($n = 100$). Υπενθυμίζουμε ότι $Error1$ είναι το σφάλμα να επιλεγούν μη σημαντικές μεταβλητές, $Error2$ είναι το σφάλμα να μην επιλεγούν σημαντικές μεταβλητές και $Error1\&2$ είναι το σφάλμα του να συμβούν τα $Error1$ και $Error2$ ταυτόχρονα .

Για ποσοστό λογοκρισίας χαμηλό, έχουμε υψηλά ποσοστά επιτυχίας και καλές εκτιμήσεις των συντελεστών. Συγκρίνοντας τα ποσοστά επιτυχίας (*Perc. of Success*) στο ακόλουθο πίνακα 1, βλέπουμε τη μέθοδό μας να υπερτερεί της *coxph* και τα καλύτερα αποτελέσματα να προκύπτουν όταν η μέθοδός μας συνδυάζεται με το κριτήριο BIC.

ΠΙΝΑΚΑΣ 1: ΠΟΣΟΣΤΑ ΕΠΙΤΥΧΙΑΣ – Parametric Case with Gamma Frailty and Censoring Proportion ≈ 10%									
Number of Covariates	**Sample Size**	**Our Method - AIC**				**Coxph - AIC**			
		Perc. of Success	**Error1**	**Error2**	**Error1&2**	**Perc. of Success**	**Error1**	**Error2**	**Error1&2**
$k = 3$	$n = 50$	0.88	0.12	0.00	0.00	0.78	0.22	0.00	0.00
	$n = 100$	0.88	0.12	0.00	0.00	0.78	0.22	0.00	0.00
$k = 4$	$n = 50$	0.62	0.36	0.02	0.00	0.56	0.04	0.12	0.08
	$n = 100$	0.76	0.24	0.00	0.00	0.78	0.22	0.02	0.02
$k = 5$	$n = 50$	0.66	0.34	0.00	0.00	0.64	0.34	0.02	0.00
	$n = 100$	0.72	0.28	0.00	0.00	0.68	0.32	0.00	0.00
Number of Covariates	**Sample Size**	**Our Method – BIC**				**Coxph – BIC**			
		Perc. of Success	**Error1**	**Error2**	**Error1&2**	**Perc. of Success**	**Error1**	**Error2**	**Error1&2**
$k = 3$	$n = 50$	0.94	0.06	0.00	0.00	0.86	0.12	0.02	0.00
	$n = 100$	0.98	0.02	0.00	0.00	0.84	0.16	0.00	0.00
$k = 4$	$n = 50$	0.78	0.16	0.06	0.00	0.64	0.26	0.12	0.02
	$n = 100$	0.92	0.08	0.00	0.00	0.70	0.30	0.08	0.08
$k = 5$	$n = 50$	0.84	0.14	0.04	0.02	0.82	0.16	0.04	0.02
	$n = 100$	0.96	0.04	0.00	0.00	0.92	0.08	0.00	0.00

Στην περίπτωση του μεσαίου ποσοστού λογοκρισίας (πίνακας 2), τα αποτελέσματα είναι και πάλι καλά, ενώ επιβεβαιώνεται η υπεροχή του κριτηρίου BIC σε συνδυασμό με τη μέθοδό μας.

ΠΙΝΑΚΑΣ 2: ΠΟΣΟΣΤΑ ΕΠΙΤΥΧΙΑΣ – Parametric Case with Gamma Frailty and Censoring Proportion ≈ 46%									
Number of Covariates	**Sample Size**	**Our Method - AIC**				**Coxph - AIC**			
		Perc. of Success	**Error1**	**Error2**	**Error1&2**	**Perc. of Success**	**Error1**	**Error2**	**Error1&2**
$k=3$	$n=50$	0.86	0.14	0.00	0.00	0.78	0.18	0.04	0.00
	$n=100$	0.88	0.12	0.00	0.00	0.86	0.14	0.00	0.00
$k=4$	$n=50$	0.66	0.32	0.04	0.02	0.62	0.34	0.12	0.08
	$n=100$	0.72	0.28	0.00	0.00	0.64	0.34	0.02	0.00
$k=5$	$n=50$	0.48	0.52	0.02	0.02	0.66	0.32	0.04	0.02
	$n=100$	0.54	0.46	0.00	0.00	0.54	0.46	0.02	0.02
Number of Covariates	**Sample Size**	**Our Method – BIC**				**Coxph – BIC**			
		Perc. of Success	**Error1**	**Error2**	**Error1&2**	**Perc. of Success**	**Error1**	**Error2**	**Error1&2**
$k=3$	$n=50$	0.90	0.08	0.02	0.00	0.88	0.10	0.02	0.00
	$n=100$	0.94	0.06	0.00	0.00	0.90	0.08	0.02	0.00
$k=4$	$n=50$	0.74	0.14	0.18	0.06	0.68	0.14	0.22	0.04
	$n=100$	0.94	0.04	0.02	0.00	0.78	0.16	0.08	0.02
$k=5$	$n=50$	0.74	0.24	0.06	0.04	0.70	0.26	0.08	0.04
	$n=100$	0.88	0.12	0.00	0.00	0.86	0.14	0.00	0.00

Στην περίπτωση του υψηλού ποσοστού λογοκρισίας (πίνακας 3) και όπως είναι αναμενόμενο, τα ποσοστά επιτυχίας είναι μειωμένα, ιδίως στην περίπτωση το μικρού δείγματος $n=50$ (βλ. αφενός τα εξαιρετικά χαμηλά ποσοστά επιτυχίας που έδωσαν τόσο η μέθοδός μας όσο και η *coxph* για $k=4$ και $n=50$, αφετέρου το ότι για $n=50$ τα ποσοστά επιτυχίας δεν ξεπέρασαν το 66%).

ΠΙΝΑΚΑΣ 3: ΠΟΣΟΣΤΑ ΕΠΙΤΥΧΙΑΣ – Parametric Case with Gamma Frailty and Censoring Proportion ≈ 80%									
Number of Covariates	**Sample Size**	**Our Method - AIC**				**Coxph - AIC**			
		Perc. of Success	**Error1**	**Error2**	**Error1&2**	**Perc. of Success**	**Error1**	**Error2**	**Error1&2**
$k=3$	$n=50$	0.60	0.30	0.18	0.08	0.66	0.24	0.22	0.12
	$n=100$	0.80	0.20	0.00	0.00	0.86	0.14	0.00	0.00
$k=4$	$n=50$	0.38	0.36	0.44	0.18	0.36	0.34	0.46	0.16
	$n=100$	0.62	0.32	0.18	0.12	0.62	0.32	0.20	0.14
$k=5$	$n=50$	0.42	0.46	0.22	0.10	0.42	0.34	0.32	0.08
	$n=100$	0.60	0.40	0.06	0.06	0.62	0.34	0.08	0.04
Number of Covariates	**Sample Size**	**Our Method - BIC**				**Coxph - BIC**			
		Perc. of Success	**Error1**	**Error2**	**Error1&2**	**Perc. of Success**	**Error1**	**Error2**	**Error1&2**
$k=3$	$n=50$	0.60	0.16	0.32	0.08	0.50	0.18	0.42	0.10
	$n=100$	0.92	0.06	0.02	0.00	0.88	0.06	0.08	0.02
$k=4$	$n=50$	0.22	0.22	0.72	0.16	0.22	0.18	0.70	0.10
	$n=100$	0.56	0.16	0.40	0.12	0.46	0.18	0.52	0.16
$k=5$	$n=50$	0.62	0.16	0.32	0.10	0.38	0.18	0.50	0.06
	$n=100$	0.70	0.16	0.20	0.06	0.62	0.16	0.28	0.06

Δεδομένου ότι χρησιμοποιήσαμε την coxph ως βάση αναφοράς για να συγκρίνουμε τη μέθοδό μας και "διαβάζοντας" τις γραμμές σε καθέναν από τους παραπάνω πίνακες 1, 2, 3, βλέπουμε ότι η μέθοδός μας δεν υστερεί της coxph, αντιθέτως δίνει σχετικά καλύτερα αποτελέσματα (υψηλότερο ποσοστό επιτυχίας με χαμηλότερα σφάλματα), ιδίως όταν συνδυάζεται με το κριτήριο BIC.

Όσον αφορά την εκτίμηση των παραμέτρων (συντελεστών και παραμέτρων της Weibull), στον πίνακα 4 έχουμε τα αποτελέσματα (τους μέσους όρους και στις παρενθέσεις τις τυπικές αποκλίσεις) από το μοντέλο με $k=3$ το πλήθος συμμεταβλητές. Στο πραγματικό μοντέλο θεωρήσαμε ως διάνυσμα συντελεστών το $\boldsymbol{\beta}=(2,0,-1.3)$ και παραμέτρους της Weibull τις $\alpha = w\_scale = 1$ και $\lambda = w\_shape = 2$.

ΠΙΝΑΚΑΣ 4: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – Parametric Case with Gamma Frailty, 3 Covariates & Censoring Proportion ≈ 10%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 2)	2.065 (0.359)	2.074 (0.359)	1.890 (0.380)	1.867 (0.360)
	β_2 (real value = 0)	0.034 (0.176)	0.007 (0.141)	0.003 (0.193)	0.026 (0.145)
	β_3 (real value = –1.3)	–1.337 (0.284)	–1.325 (0.277)	–1.210 (0.318)	–1.205 (0.331)
	$\alpha = w\_scale$ (real value = 1)	0.965 (0.087)	0.964 (0.086)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.073 (0.254)	2.069 (0.253)	-	-
$n = 100$	β_1 (real value = 2)	2.076 (0.225)	2.084 (0.222)	1.914 (0.290)	1.894 (0.280)
	β_2 (real value = 0)	0.021 (0.109)	–0.008 (0.060)	0.009 (0.109)	0.008 (0.010)
	β_3 (real value = –1.3)	–1.334 (0.173)	–1.319 (0.179)	–1.223 (0.205)	–1.206 (0.191)
	$\alpha = w\_scale$ (real value = 1)	0.977 (0.061)	0.978 (0.060)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.082 (0.183)	2.082 (0.183)	-	-

Από τον πίνακα 4 βλέπουμε και πάλι ότι στην περίπτωση του χαμηλού ποσοστού λογοκρισίας, έχουμε καλή εκτίμηση των συντελεστών ακόμα και με μικρό μέγεθος δείγματος.

Στον επόμενο πίνακα 5 έχουμε την εκτίμηση του ίδιου μοντέλου (με 3 συμμεταβλητές) και μεσαίο ποσοστό λογοκρισίας, όπου και πάλι επιτυγχάνουμε καλά αποτελέσματα.

ΠΙΝΑΚΑΣ 5: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – Parametric Case with Gamma Frailty, 3 Covariates & Censoring Proportion ≈ 46%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 2)	2.201 (0.464)	2.208 (0.459)	2.083 (0.505)	2.055 (0.508)
	β_2 (real value = 0)	0.022 (0.286)	–0.006 (0.262)	0.024 (0.259)	0.013 (0.230)
	β_3 (real value = –1.3)	–1.509 (0.351)	–1.519 (0.329)	–1.449 (0.338)	–1.408 (0.339)
	$\alpha = w\_scale$ (real value = 1)	0.994 (0.110)	0.999 (0.120)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.237 (0.378)	2.231 (0.388)	-	-
$n = 100$	β_1 (real value = 2)	2.148 (0.346)	2.144 (0.337)	1.984 (0.346)	1.931 (0.356)
	β_2 (real value = 0)	–0.042 (0.168)	–0.036 (0.146)	–0.028 (0.150)	–0.034 (0.123)
	β_3 (real value = –1.3)	–1.353 (0.254)	–1.355 (0.259)	–1.267 (0.268)	–1.243 (0.232)
	$\alpha = w\_scale$ (real value = 1)	0.989 (0.074)	0.989 (0.073)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.120 (0.225)	2.118 (0.224)	-	-

Τέλος, στον πίνακα 6 έχουμε τα αντίστοιχα αποτελέσματα του μοντέλου 3 συμμεταβλητών, με υψηλό ποσοστό λογοκρισίας, όπου βλέπουμε ότι αν και μειώθηκαν τα ποσοστά επιτυχίας, ωστόσο οι εκτιμήσεις είναι γενικά καλές .

ΠΙΝΑΚΑΣ 6: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – Parametric Case with Gamma Frailty, 3 Covariates & Censoring Proportion ≈ 80%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 2)	2.423 (0.909)	2.388 (0.917)	2.340 (0.895)	2.274 (0.976)
	β_2 (real value = 0)	–0.092 (0.734)	–0.127 (0.654)	0.022 (0.867)	–0.013 (0.813)
	β_3 (real value = –1.3)	–1.596 (0.591)	–1.617 (0.611)	–1.595 (0.847)	–1.676 (0.910)
	$\alpha = w\_scale$ (real value = 1)	0.889 (0.358)	0.972 (0.471)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.428 (0.704)	2.349 (0.742)	-	-
$n = 100$	β_1 (real value = 2)	2.196 (0.553)	2.148 (0.527)	2.043 (0.569)	2.000 (0.583)
	β_2 (real value = 0)	–0.068 (0.370)	–0.025 (0.253)	–0.025 (0.312)	–0.055 (0.225)
	β_3 (real value = –1.3)	–1.406 (0.374)	–1.420 (0.340)	–1.337 (0.431)	–1.351 (0.371)
	$\alpha = w\_scale$ (real value = 1)	0.874 (0.155)	0.876 (0.155)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.141 (0.478)	2.120 (0.455)	-	-

Στους πίνακες 7, 8, 9 παρουσιάζουμε τις αντίστοιχες εκτιμήσεις των παραμέτρων του μοντέλου με 4 συμμεταβλητές. Εδώ, οι συντελεστές του πραγματικού μοντέλου είναι $\boldsymbol{\beta} = (0.8, 0, 0, 1)$ και οι παράμετροι της Weibull τις $\alpha = w\_scale = 1$ και $\lambda = w\_shape = 2$. Τα αποτελέσματα δείχνουν καλή εκτίμηση των συντελεστών και από τις δύο μεθόδους, με τη μεγαλύτερη απόκλιση από τις πραγματικές τιμές να εμφανίζεται για υψηλό ποσοστό λογοκρισίας με μικρό μέγεθος δείγματος.

ΠΙΝΑΚΑΣ 7: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – Parametric Case with Gamma Frailty, 4 Covariates & Censoring Proportion ≈ 10%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 0.8)	0.846 (0.251)	0.872 (0.227)	0.821 (0.255)	0.799 (0.222)
	β_2 (real value = 0)	–0.017 (0.254)	–0.027 (0.203)	–0.016 (0.265)	–0.008 (0.229)
	β_3 (real value = 0)	–0.001 (0.234)	–0.023 (0.197)	–0.002 (0.251)	–0.027 (0.212)
	β_4 (real value = 1)	1.139 (0.325)	1.144 (0.303)	1.069 (0.327)	1.028 (0.314)
	$\alpha = w\_scale$ (real value = 1)	0.939 (0.072)	0.942 (0.076)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.222 (0.285)	2.202 (0.281)	-	-
$n = 100$	β_1 (real value = 0.8)	0.837 (0.176)	0.832 (0.167)	0.762 (0.180)	0.741 (0.182)
	β_2 (real value = 0)	–0.001 (0.127)	0.007 (0.051)	0.003 (0.089)	0.022 (0.124)
	β_3 (real value = 0)	0.016 (0.124)	0.010 (0.102)	0.034 (0.091)	0.024 (0.137)
	β_4 (real value = 1)	1.088 (0.180)	1.084 (0.169)	0.982 (0.183)	0.956 (0.173)
	$\alpha = w\_scale$ (real value = 1)	0.970 (0.056)	0.970 (0.056)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.118 (0.202)	2.109 (0.198)	-	-

ΠΙΝΑΚΑΣ 8: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – Parametric Case with Gamma Frailty, 4 Covariates & Censoring Proportion ≈ 46%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 0.8)	0.871 (0.298)	0.920 (0.272)	0.830 (0.296)	0.857 (0.241)
	β_2 (real value = 0)	0.036 (0.240)	0.032 (0.244)	–0.045 (0.300)	0.023 (0.230)
	β_3 (real value = 0)	–0.049 (0.335)	–0.007 (0.287)	0.014 (0.353)	–0.020 (0.243)
	β_4 (real value = 1)	1.163 (0.388)	1.158 (0.385)	1.068 (0.359)	1.048 (0.369)
	$\alpha = w\_scale$ (real value = 1)	0.971 (0.097)	0.988 (0.109)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.164 (0.380)	2.123 (0.387)	-	-
$n = 100$	β_1 (real value = 0.8)	0.867 (0.216)	0.855 (0.205)	0.839 (0.261)	0.778 (0.177)
	β_2 (real value = 0)	–0.054 (0.163)	–0.018 (0.088)	–0.053 (0.161)	–0.030 (0.106)
	β_3 (real value = 0)	–0.001 (0.098)	0.00 (0.000)	–0.020 (0.131)	0.022 (0.078)
	β_4 (real value = 1)	1.082 (0.214)	1.072 (0.208)	1.053 (0.282)	0.968 (0.173)
	$\alpha = w\_scale$ (real value = 1)	0.992 (0.056)	0.994 (0.062)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.077 (0.239)	2.066 (0.242)	-	-

ΠΙΝΑΚΑΣ 9: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – Parametric Case with Gamma Frailty, 4 Covariates & Censoring Proportion ≈ 80%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 0.8)	1.358 (0.843)	1.528 (0.984)	1.233 (0.667)	1.355 (0.816)
	β_2 (real value = 0)	–0.095 (0.837)	–0.109 (0.754)	–0.037 (0.554)	–0.057 (0.677)
	β_3 (real value = 0)	0.002 (0.900)	0.051 (0.864)	–0.027 (0.744)	0.058 (0.709)
	β_4 (real value = 1)	1.381 (0.656)	1.392 (0.601)	1.365 (0.583)	1.313 (0.553)
	$\alpha = w\_scale$ (real value = 1)	0.941 (0.368)	1.042 (0.436)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.269 (0.794)	2.102 (0.753)	-	-
$n = 100$	β_1 (real value = 0.8)	0.902 (0.250)	0.928 (0.225)	0.868 (0.230)	0.922 (0.237)
	β_2 (real value = 0)	0.060 (0.284)	0.047 (0.273)	0.046 (0.288)	0.041 (0.261)
	β_3 (real value = 0)	–0.035 (0.306)	0.050 (0.201)	–0.026 (0.343)	0.047 (0.254)
	β_4 (real value = 1)	1.012 (0.316)	1.041 (0.251)	0.999 (0.331)	1.000 (0.295)
	$\alpha = w\_scale$ (real value = 1)	0.867 (0.161)	0.917 (0.205)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.042 (0.400)	1.974 (0.426)	-	-

Τέλος, στους πίνακες 10, 11, 12 έχουμε τα αποτελέσματα των εκτιμήσεων για το μοντέλο των 5 συμμεταβλητών, με πραγματικές παραμέτρους $\boldsymbol{\beta} = (0, 2.5, 4, 0, -1)$, $\alpha = w\_scale = 1$ και $\lambda = w\_shape = 2$. Επιβεβαιώνεται εδώ το ότι μεγαλύτερο σφάλμα έχουμε αυξανόμενου του ποσοστού λογοκρισίας και μειούμενου του μεγέθους του δείγματος.

ΠΙΝΑΚΑΣ 10: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – Parametric Case with Gamma Frailty, 5 Covariates & Censoring Proportion ≈ 10%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 0)	0.023 (0.192)	0.029 (0.164)	–0.029 (0.172)	0.004 (0.108)
	β_2 (real value = 2.5)	2.930 (0.493)	2.903 (0.490)	2.645 (0.509)	2.600 (0.501)
	β_3 (real value = 4)	4.575 (0.701)	4.537 (0.684)	4.140 (0.728)	4.071 (0.707)
	β_4 (real value = 0)	–0.017 (0.250)	–0.022 (0.200)	–0.038 (0.244)	–0.034 (0.253)
	β_5 (real value = –1)	–1.074 (0.330)	–1.094 (0.310)	–0.978 (0.300)	–0.981 (0.297)
	$\alpha = w\_scale$ (real value = 1)	0.959 (0.082)	0.959 (0.086)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.308 (0.295)	2.29 (0.297)	-	-
$n = 100$	β_1 (real value = 0)	0.011 (0.114)	0.006 (0.044)	0.005 (0.121)	0.005 (0.080)
	β_2 (real value = 2.5)	2.539 (0.260)	2.530 (0.256)	2.282 (0.285)	2.268 (0.270)
	β_3 (real value = 4)	4.160 (0.436)	4.133 (0.416)	3.704 (0.462)	3.676 (0.446)
	β_4 (real value = 0)	–0.004 (0.091)	0.007 (0.048)	–0.011 (0.109)	0.006 (0.045)
	β_5 (real value = –1)	–0.994 (0.156)	–0.993 (0.154)	–0.888 (0.160)	–0.887 (0.160)
	$\alpha = w\_scale$ (real value = 1)	0.956 (0.065)	0.957 (0.065)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.065 (0.190)	2.055 (0.184)	-	-
ΠΙΝΑΚΑΣ 11: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – Parametric Case with Gamma Frailty, 5 Covariates & Censoring Proportion ≈ 46%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 0)	0.018 (0.324)	0.023 (0.227)	0.050 (0.273)	0.038 (0.245)
	β_2 (real value = 2.5)	2.761 (0.544)	2.689 (0.514)	2.532 (0.505)	2.518 (0.510)
	β_3 (real value = 4)	4.507 (0.725)	4.437 (0.739)	4.151 (0.778)	4.129 (0.770)
	β_4 (real value = 0)	0.023 (0.309)	0.018 (0.286)	0.038 (0.276)	0.016 (0.273)
	β_5 (real value = –1)	–1.185 (0.342)	–1.199 (0.316)	–1.142 (0.332)	–1.152 (0.291)
	$\alpha = w\_scale$ (real value = 1)	0.962 (0.095)	0.969 (0.104)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.254 (0.333)	2.214 (0.334)	-	-
$n = 100$	β_1 (real value = 0)	0.011 (0.201)	0.024 (0.153)	0.020 (0.179)	0.025 (0.147)
	β_2 (real value = 2.5)	2.622 (0.336)	2.605 (0.323)	2.436 (0.351)	2.413 (0.344)
	β_3 (real value = 4)	4.182 (0.509)	4.156 (0.492)	3.932 (0.668)	3.885 (0.642)
	β_4 (real value = 0)	–0.025 (0.200)	–0.020 (0.099)	–0.028 (0.237)	–0.010 (0.117)
	β_5 (real value = –1)	–1.07 (0.218)	–1.065 (0.224)	–1.007 (0.227)	–1.002 (0.227)
	$\alpha = w\_scale$ (real value = 1)	0.974 (0.071)	0.974 (0.068)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.093 (0.210)	2.082 (0.208)	-	-

ΠΙΝΑΚΑΣ 12: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – Parametric Case with Gamma Frailty, 5 Covariates & Censoring Proportion ≈ 80%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 0)	–0.066 (0.391)	–0.029 (0.204)	–0.082 (0.650)	–0.028 (0.365)
	β_2 (real value = 2.5)	2.869 (1.162)	2.800 (1.079)	3.215 (1.638)	3.142 (1.307)
	β_3 (real value = 4)	4.693 (1.870)	4.503 (1.852)	5.031 (2.332)	4.832 (2.261)
	β_4 (real value = 0)	–0.025 (0.817)	–0.085 (0.703)	0.071 (1.079)	–0.064 (0.98)
	β_5 (real value = –1)	–1.531 (0.773)	–1.517 (0.800)	–1.833 (1.083)	–1.992 (1.092)
	$\alpha = w\_scale$ (real value = 1)	0.934 (0.332)	0.986 (0.368)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.284 (0.806)	2.181 (0.802)	-	-
$n = 100$	β_1 (real value = 0)	–0.019 (0.275)	–0.011 (0.184)	0.018 (0.387)	–0.007 (0.324)
	β_2 (real value = 2.5)	2.812 (0.884)	2.756 (0.908)	2.747 (0.888)	2.649 (0.855)
	β_3 (real value = 4)	4.423 (1.383)	4.316 (1.429)	4.360 (1.369)	4.183 (1.346)
	β_4 (real value = 0)	–0.029 (0.430)	–0.039 (0.362)	0.045 (0.390)	–0.020 (0.274)
	β_5 (real value = –1)	–1.186 (0.438)	–1.271 (0.396)	–1.240 (0.430)	–1.318 (0.351)
	$\alpha = w\_scale$ (real value = 1)	0.890 (0.199)	0.926 (0.297)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.208 (0.642)	2.162 (0.662)	-	-

4.3. ΠΡΟΣΟΜΟΙΩΣΕΙΣ ΓΙΑ ΤΗΝ ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ ΚΑΙ ΤΗΝ ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΣΤΟ ΠΑΡΑΜΕΤΡΙΚΟ ΜΟΝΤΕΛΟ INVERSE GAUSSIAN ΕΥΠΑΘΕΙΑΣ ΜΕΣΩ ΤΟΥ ΜΟΝΤΕΛΟΥ ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΥ

4.3.1. Εισαγωγή: Με τη βοήθεια της R και κατ' αντιστοιχία με τα όσα είδαμε στην προηγούμενη παράγραφο, θα προσομοιώσουμε τώρα το μοντέλο με βασική συνάρτηση κινδύνου της Weibull κατανομής, ευπάθεια που ακολουθεί την Inverse Gaussian κατανομή και οσεσδήποτε το πλήθος συμμεταβλητές. Στη συνέχεια, θα επιλέξουμε το βέλτιστο μοντέλο με τη βοήθεια των κριτηρίων AIC και BIC και θα συγκρίνουμε τα αποτελέσματα με την επιλογή μοντέλου που θα μας δώσει η έτοιμη συνάρτηση *coxph*. Έτσι:

- Η βασική συνάρτηση κινδύνου είναι και πάλι της κατανομής Weibull, με παράμετρο κλίμακας $\alpha > 0$ και παράμετρο σχήματος $\lambda > 0$, άρα

$$h_0(t) = \lambda\alpha^{-\lambda}t^{\lambda-1},\ t > 0$$

και

$$H_0(t) = \left(\frac{t}{\alpha}\right)^{\lambda},\ t > 0$$

Για τις ανάγκες της προσομοίωσης θα θεωρήσουμε παράμετρο κλίμακας $\alpha = 1$ *και παράμετρο σχήματος* $\lambda = 2$.

Η τυχαία μεταβλητή U της ευπάθειας, ακολουθεί την Inverse Gaussian κατανομή όπως την παρουσιάσαμε στην παράγραφο **3.4**, με παραμέτρους $\mu = 1$ και $\lambda = \frac{1}{\sigma^2}$, $\sigma > 0$, οπότε η συνάρτηση πυκνότητας πιθανότητας είναι

$$f_U(u) = f(u) = \frac{1}{\sigma\sqrt{2\pi u^3}} \cdot e^{-\frac{(u-1)^2}{2\sigma^2 u}},\ t > 0,$$

οπότε η αναμενόμενη τιμή της ευπάθειας είναι $E[U] = 1$ και αναμενόμενη διασπορά αυτής είναι $V[U] = \sigma^2$.

Για τις ανάγκες της προσομοίωσης χρησιμοποιούμε $\sigma^2 = 0.25$.

- Όπως και στην περίπτωση της Γάμμα ευπάθειας (βλ. παράγραφο **4.2.**, εδάφιο **4.2.2.**), το μοντέλο μας περιγράφεται από την εξίσωση

$$h(t \mid u, \mathbf{z}) = \lambda \cdot \left\{\alpha \cdot \left(u \cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}}\right)^{-\frac{1}{\lambda}}\right\}^{-\lambda} t^{\lambda-1}$$

οπότε και πάλι *το μοντέλο είναι της κατανομής Weibull με παράμετρο σχήματος* λ *και παράμετρο κλίμακας* $A = \alpha \cdot \left(u \cdot e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}}\right)^{-\frac{1}{\lambda}}$.

- Ο μετασχηματισμός Laplace της Inverse Gaussian ευπάθειας έχει υπολογισθεί στην παράγραφο **3.4.** και συγκεκριμένα στην εξίσωση **(3.53)** και είναι

$$\mathbf{L}(s) = \exp\left\{\frac{1}{\sigma^2}\left(1 - \sqrt{2\sigma^2 s + 1}\right)\right\}$$

Επομένως, η συνάρτηση μετασχηματισμού είναι $G(x) = -\ln \mathbf{L}(x)$, άρα:

$$G(x) = \frac{1}{\sigma^2}\left(\sqrt{2\sigma^2 x + 1} - 1\right)$$

με πρώτη παράγωγο

$$G'(x) = \frac{1}{\sqrt{2\sigma^2 x + 1}}$$

4.3.2. Κατασκευή του κώδικα: Η κατασκευή του κώδικα είναι η ίδια με πριν, διαφοροποιούμαστε μόνο στα εξής σημεία:

- Στην αρχικοποίηση, αλλάζουν η συνάρτηση μετασχηματισμού G και η παράγωγός της G'.

```
.....
.....
#============================================
#Initialization
.....
.....
w_scale_init<-1#Weibull scale parameter (alpha)
w_shape<-2      #Weibull shape parameter (lambda)
sqr_sigma<-0.25          #Variance of frailty (Gamma) distribution
h<-function(x,alpha,lambda){lambda*exp(lambda*log(x/alpha)-log(x))}   #Baseline      hazard
                                                                      #function h(t)
H<-function(x,alpha,lambda){exp(lambda*log(x/alpha))} #Baseline cumulative hazard
                                                      #function H(t)
G<-function(x) {(sqrt(2*sqr_sigma*x+1)-1)/sqr_sigma}  #G function
dG<-function(x)          {1/sqrt(2*sqr_sigma*x+1)}        #Derivative of G function
.....
.....
```

- Στο κομμάτι της προσομοίωσης των χρόνων διακοπής, οι ευπάθειες των μονάδων παράγονται ως τυχαίες παρατηρήσεις της Inverse Gaussian με παραμέτρους $\mu = 1$ και $\lambda = \frac{1}{\sigma^2}$, $\sigma > 0$. Προς τούτο, αξιοποιούμε τη συνάρτηση *rinvgauss* της βιβλιοθήκης *statmod* της R:

```
.....
.....
library(statmod)
.....
.....
#=============================================
#Generation of frailty data

frailty<-numeric(n)
for (i in 1:n) {frailty[i]<-rinvgauss(1, mu=1, lambda=1/sqr_sigma)}
.....
```

Τα υπόλοιπα μέρη του κώδικα παραμένουν ίδια, με μόνο μειονέκτημα ως προς τη σύγκριση το ότι κατά την εκτέλεση της *coxph* αφαιρέθηκε η ευπάθεια, αφού η συνάρτηση *frailty* της R δεν έχει την επιλογή ευπάθειας που ακολουθεί την Inverse Gaussian κατανομή (οι δυνατές κατανομές ευπάθειας στη συνάρτηση *frailty* είναι Γάμμα, Gaussian και t – βλ. τεκμηρίωση της συνάρτησης *frailty* στο [58]) .

4.3.3. Εκτέλεση του προγράμματος - Αποτελέσματα: Εκτελέσαμε το πρόγραμμα προσομοιώνοντας και εδώ μοντέλα με 3, 4 και 5 συμμεταβλητές. Οι αντίστοιχοι πίνακες ποσοστών επιτυχίας, από τους οποίους προκύπτει και πάλι η υπεροχή του κριτηρίου BIC, έχουν ως ακολούθως:

ΠΙΝΑΚΑΣ 13: ΠΟΣΟΣΤΑ ΕΠΙΤΥΧΙΑΣ – Parametric Case with Inverse Gaussian Frailty and Censoring Proportion ≈ 10%									
Number of Covariates	Sample Size	Our Method - AIC				Coxph - AIC			
		Perc. of Success	Error1	Error2	Error1&2	Perc. of Success	Error1	Error2	Error1&2
$k=3$	$n=50$	0.78	0.22	0.00	0.00	0.76	0.24	0.00	0.00
	$n=100$	0.72	0.28	0.00	0.00	0.68	0.32	0.00	0.00
$k=4$	$n=50$	0.68	0.32	0.00	0.00	0.64	0.36	0.00	0.00
	$n=100$	0.74	0.26	0.00	0.00	0.78	0.22	0.00	0.00
$k=5$	$n=50$	0.68	0.32	0.00	0.00	0.64	0.36	0.02	0.00
	$n=100$	0.76	0.24	0.00	0.00	0.76	0.24	0.00	0.00
Number of Covariates	Sample Size	Our Method - BIC				Coxph – BIC			
		Perc. of Success	Error1	Error2	Error1&2	Perc. of Success	Error1	Error2	Error1&2
$k=3$	$n=50$	0.88	0.12	0.00	0.00	0.94	0.06	0.00	0.00
	$n=100$	0.96	0.04	0.00	0.00	0.96	0.04	0.00	0.00
$k=4$	$n=50$	0.84	0.16	0.00	0.00	0.88	0.12	0.00	0.00
	$n=100$	0.92	0.08	0.02	0.02	0.92	0.08	0.02	0.02
$k=5$	$n=50$	0.84	0.16	0.00	0.00	0.86	0.14	0.02	0.02
	$n=100$	0.98	0.02	0.00	0.00	0.96	0.04	0.00	0.00
ΠΙΝΑΚΑΣ 14: ΠΟΣΟΣΤΑ ΕΠΙΤΥΧΙΑΣ – Parametric Case with Inverse Gaussian Frailty and Censoring Proportion ≈ 46%									
Number of Covariates	Sample Size	Our Method - AIC				Coxph - AIC			
		Perc. of Success	Error1	Error2	Error1&2	Perc. of Success	Error1	Error2	Error1&2
$k=3$	$n=50$	0.88	0.12	0.00	0.00	0.84	0.16	0.00	0.00
	$n=100$	0.86	0.14	0.00	0.00	0.88	0.12	0.00	0.00
$k=4$	$n=50$	0.68	0.32	0.00	0.00	0.70	0.28	0.02	0.00
	$n=100$	0.62	0.38	0.02	0.02	0.58	0.42	0.02	0.02
$k=5$	$n=50$	0.70	0.28	0.04	0.02	0.64	0.30	0.08	0.02
	$n=100$	0.74	0.26	0.00	0.00	0.66	0.34	0.00	0.00
Number of Covariates	Sample Size	Our Method - BIC				Coxph – BIC			
		Perc. of Success	Error1	Error2	Error1&2	Perc. of Success	Error1	Error2	Error1&2
$k=3$	$n=50$	0.96	0.04	0.00	0.00	0.96	0.04	0.00	0.00
	$n=100$	0.98	0.02	0.00	0.00	1.00	0.00	0.00	0.00
$k=4$	$n=50$	0.84	0.10	0.06	0.00	0.80	0.12	0.08	0.00
	$n=100$	0.92	0.08	0.04	0.04	0.84	0.16	0.04	0.04
$k=5$	$n=50$	0.84	0.08	0.10	0.02	0.78	0.10	0.14	0.02
	$n=100$	0.98	0.02	0.00	0.00	0.98	0.02	0.00	0.00

ΠΙΝΑΚΑΣ 15: ΠΟΣΟΣΤΑ ΕΠΙΤΥΧΙΑΣ – Parametric Case with Inverse Gaussian Frailty and Censoring Proportion ≈ 80%									
Number of Covariates	**Sample Size**	**Our Method - AIC**				**Coxph - AIC**			
		Perc. of Success	**Error1**	**Error2**	**Error1&2**	**Perc. of Success**	**Error1**	**Error2**	**Error1&2**
$k=3$	$n=50$	0.66	0.26	0.22	0.14	0.66	0.26	0.22	0.14
	$n=100$	0.74	0.26	0.02	0.02	0.76	0.21	0.02	0.02
$k=4$	$n=50$	0.34	0.50	0.52	0.36	0.40	0.44	0.50	0.34
	$n=100$	0.56	0.34	0.24	0.14	0.58	0.34	0.24	0.16
$k=5$	$n=50$	0.36	0.48	0.36	0.20	0.38	0.54	0.28	0.20
	$n=100$	0.68	0.28	0.06	0.02	0.60	0.34	0.12	0.06
Number of Covariates	**Sample Size**	**Our Method - BIC**				**Coxph – BIC**			
		Perc. of Success	**Error1**	**Error2**	**Error1&2**	**Perc. of Success**	**Error1**	**Error2**	**Error1&2**
$k=3$	$n=50$	0.74	0.12	0.24	0.10	0.68	0.16	0.28	0.12
	$n=100$	0.92	0.08	0.02	0.02	0.88	0.12	0.04	0.04
$k=4$	$n=50$	0.28	0.32	0.68	0.28	0.24	0.28	0.74	0.26
	$n=100$	0.52	0.16	0.46	0.14	0.56	0.08	0.42	0.06
$k=5$	$n=50$	0.46	0.22	0.44	0.12	0.38	0.24	0.58	0.20
	$n=100$	0.66	0.12	0.24	0.02	0.56	0.10	0.40	0.06

Οι πίνακες εκτιμήσεων των παραμέτρων είναι εδώ οι ακόλουθοι:

ΠΙΝΑΚΑΣ 16: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – Parametric Case w. Inverse Gaussian Frailty, 3 Covariates & Censoring Proportion ≈ 10%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 2)	2.172 (0.365)	2.17 (0.373)	1.932 (0.344)	1.911 (0.338)
	β_2 (real value = 0)	–0.062 (0.236)	–0.072 (0.198)	–0.062 (0.205)	–0.034 (0.138)
	β_3 (real value = –1.3)	–1.396 (0.302)	–1.392 (0.294)	–1.257 (0.292)	–1.265 (0.283)
	$\alpha = w\_scale$ (real value = 1)	0.963 (0.092)	0.963 (0.092)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.133 (0.251)	2.126 (0.246)	-	-
$n = 100$	β_1 (real value = 2)	2.057 (0.27)	2.046 (0.241)	1.842 (0.243)	1.830 (0.222)
	β_2 (real value = 0)	0.020 (0.168)	0.017 (0.085)	0.016 (0.156)	0.015 (0.074)
	β_3 (real value = –1.3)	–1.366 (0.187)	–1.360 (0.189)	–1.226 (0.182)	–1.218 (0.179)
	$\alpha = w\_scale$ (real value = 1)	0.977 (0.060)	0.980 (0.059)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.063 (0.200)	2.053 (0.197)	-	-

ΠΙΝΑΚΑΣ 17: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – Parametric Case w. Inverse Gaussian Frailty, 3 Covariates & Censoring Proportion ≈ 46%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 2)	2.096 (0.479)	2.097 (0.482)	1.991 (0.493)	1.974 (0.479)
	β_2 (real value = 0)	0.027 (0.259)	0.001 (0.193)	0.006 (0.235)	0.005 (0.168)
	β_3 (real value = –1.3)	–1.464 (0.416)	–1.445 (0.400)	–1.351 (0.373)	–1.34 (0.369)
	$\alpha = w\_scale$ (real value = 1)	0.981 (0.084)	0.982 (0.084)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.135 (0.345)	2.126 (0.342)	-	-
$n = 100$	β_1 (real value = 2)	2.085 (0.313)	2.074 (0.307)	1.935 (0.314)	1.934 (0.302)
	β_2 (real value = 0)	–0.005 (0.144)	0.009 (0.062)	0.014 (0.126)	0.000 (0.000)
	β_3 (real value = –1.3)	–1.381 (0.221)	–1.384 (0.205)	–1.292 (0.200)	–1.282 (0.208)
	$\alpha = w\_scale$ (real value = 1)	0.977 (0.053)	0.977 (0.054)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.104 (0.243)	2.100 (0.243)	-	-

ΠΙΝΑΚΑΣ 18: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – Parametric Case w. Inverse Gaussian Frailty, 3 Covariates & Censoring Proportion ≈ 80%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 2)	2.544 (0.96)	2.443 (0.931)	2.512 (1.145)	2.442 (1.137)
	β_2 (real value = 0)	–0.176 (0.719)	–0.037 (0.635)	–0.225 (0.795)	–0.127 (0.764)
	β_3 (real value = –1.3)	–1.570 (0.449)	–1.612 (0.458)	–1.549 (0.607)	–1.577 (0.612)
	$\alpha = w\_scale$ (real value = 1)	0.950 (0.514)	0.955 (0.509)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.364 (0.782)	2.332 (0.782)	-	-
$n = 100$	β_1 (real value = 2)	2.235 (0.509)	2.211 (0.507)	2.077 (0.481)	2.08 (0.502)
	β_2 (real value = 0)	–0.020 (0.392)	0.000 (0.295)	0.018 (0.399)	–0.040 (0.351)
	β_3 (real value = –1.3)	–1.439 (0.374)	–1.44 (0.364)	–1.380 (0.448)	–1.371 (0.437)
	$\alpha = w\_scale$ (real value = 1)	0.823 (0.144)	0.821 (0.142)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.219 (0.409)	2.208 (0.408)	-	-

ΠΙΝΑΚΑΣ 19: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – Parametric Case w. Inverse Gaussian Frailty, 4 Covariates & Censoring Proportion ≈ 10%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 0.8)	0.862 (0.234)	0.85 (0.226)	0.778 (0.224)	0.764 (0.213)
	β_2 (real value = 0)	–0.047 (0.217)	–0.025 (0.191)	–0.034 (0.182)	–0.035 (0.142)
	β_3 (real value = 0)	0.031 (0.134)	0.021 (0.104)	0.024 (0.152)	0.025 (0.100)
	β_4 (real value = 1)	1.07 (0.240)	1.064 (0.235)	0.958 (0.228)	0.951 (0.224)
	$\alpha = w\_scale$ (real value = 1)	0.953 (0.099)	0.953 (0.098)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.164 (0.255)	2.156 (0.256)	-	-
$n = 100$	β_1 (real value = 0.8)	0.852 (0.196)	0.864 (0.176)	0.757 (0.174)	0.768 (0.154)
	β_2 (real value = 0)	0.001 (0.132)	–0.008 (0.09)	–0.008 (0.127)	–0.008 (0.078)
	β_3 (real value = 0)	–0.008 (0.089)	–0.007 (0.048)	0.002 (0.084)	–0.007 (0.051)
	β_4 (real value = 1)	1.085 (0.223)	1.081 (0.223)	0.963 (0.189)	0.963 (0.191)
	$\alpha = w\_scale$ (real value = 1)	0.974 (0.06)	0.976 (0.059)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.130 (0.235)	2.122 (0.233)	-	-

ΠΙΝΑΚΑΣ 20: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – Parametric Case w. Inverse Gaussian Frailty, 4 Covariates & Censoring Proportion ≈ 46%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 0.8)	0.867 (0.268)	0.886 (0.242)	0.792 (0.237)	0.825 (0.226)
	β_2 (real value = 0)	0.006 (0.234)	–0.017 (0.122)	0.010 (0.175)	–0.018 (0.129)
	β_3 (real value = 0)	–0.081 (0.299)	–0.062 (0.213)	–0.064 (0.260)	–0.072 (0.220)
	β_4 (real value = 1)	1.127 (0.310)	1.118 (0.274)	1.037 (0.299)	1.035 (0.302)
	$\alpha = w\_scale$ (real value = 1)	1.002 (0.099)	1.014 (0.111)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.140 (0.337)	2.110 (0.346)	-	-
$n = 100$	β_1 (real value = 0.8)	0.802 (0.201)	0.821 (0.169)	0.744 (0.198)	0.760 (0.174)
	β_2 (real value = 0)	0.051 (0.193)	0.020 (0.097)	0.059 (0.200)	0.026 (0.132)
	β_3 (real value = 0)	–0.024 (0.211)	0.001 (0.094)	–0.020 (0.203)	0.010 (0.107)
	β_4 (real value = 1)	1.058 (0.221)	1.038 (0.19)	0.978 (0.197)	0.958 (0.17)
	$\alpha = w\_scale$ (real value = 1)	0.976 (0.065)	0.980 (0.067)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.082 (0.226)	2.062 (0.223)	-	-

ΠΙΝΑΚΑΣ 21: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – Parametric Case w. Inverse Gaussian Frailty, 4 Covariates & Censoring Proportion ≈ 80%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 0.8)	1.188 (0.415)	1.203 (0.346)	1.116 (0.414)	1.211 (0.34)
	β_2 (real value = 0)	–0.007 (0.644)	0.063 (0.526)	–0.010 (0.568)	0.125 (0.503)
	β_3 (real value = 0)	0.114 (0.551)	0.116 (0.495)	0.116 (0.557)	0.117 (0.535)
	β_4 (real value = 1)	1.404 (0.661)	1.451 (0.604)	1.393 (0.642)	1.338 (0.643)
	$\alpha = w\_scale$ (real value = 1)	0.922 (0.233)	0.967 (0.243)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.263 (0.565)	2.148 (0.559)	-	-
$n = 100$	β_1 (real value = 0.8)	0.953 (0.328)	1.000 (0.301)	0.912 (0.293)	0.943 (0.267)
	β_2 (real value = 0)	0.075 (0.304)	0.064 (0.259)	0.034 (0.278)	0.000 (0.000)
	β_3 (real value = 0)	0.081 (0.355)	0.088 (0.273)	0.052 (0.316)	0.053 (0.188)
	β_4 (real value = 1)	1.107 (0.344)	1.155 (0.321)	1.098 (0.354)	1.122 (0.324)
	$\alpha = w\_scale$ (real value = 1)	0.856 (0.147)	0.886 (0.164)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.144 (0.417)	2.077 (0.433)	-	-

ΠΙΝΑΚΑΣ 22: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – Parametric Case w. Inverse Gaussian Frailty, 5 Covariates & Censoring Proportion ≈ 10%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 0)	0.050 (0.198)	0.036 (0.172)	0.039 (0.194)	0.028 (0.172)
	β_2 (real value = 2.5)	2.718 (0.392)	2.701 (0.384)	2.448 (0.383)	2.419 (0.350)
	β_3 (real value = 4)	4.328 (0.608)	4.298 (0.577)	3.944 (0.634)	3.878 (0.604)
	β_4 (real value = 0)	0.012 (0.254)	0.016 (0.223)	–0.031 (0.335)	0.016 (0.274)
	β_5 (real value = –1)	–1.089 (0.264)	–1.082 (0.259)	–0.989 (0.266)	–0.991 (0.256)
	$\alpha = w\_scale$ (real value = 1)	0.957 (0.087)	0.959 (0.087)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.162 (0.270)	2.146 (0.260)	-	-
$n = 100$	β_1 (real value = 0)	–0.011 (0.069)	0.000 (0.000)	–0.013 (0.055)	–0.006 (0.042)
	β_2 (real value = 2.5)	2.667 (0.252)	2.652 (0.246)	2.399 (0.242)	2.387 (0.245)
	β_3 (real value = 4)	4.209 (0.372)	4.195 (0.365)	3.828 (0.361)	3.826 (0.361)
	β_4 (real value = 0)	0.018 (0.128)	0.009 (0.064)	0.029 (0.119)	–0.007 (0.047)
	β_5 (real value = –1)	–1.070 (0.203)	–1.060 (0.199)	–0.982 (0.184)	–0.961 (0.175)
	$\alpha = w\_scale$ (real value = 1)	0.954 (0.053)	0.954 (0.052)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.119 (0.185)	2.111 (0.182)	-	-
ΠΙΝΑΚΑΣ 23: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – Parametric Case w. Inverse Gaussian Frailty, 5 Covariates & Censoring Proportion ≈ 46%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 0)	0.050 (0.211)	0.032 (0.173)	–0.015 (0.136)	0.000 (0.000)
	β_2 (real value = 2.5)	2.815 (0.629)	2.78 (0.597)	2.757 (0.750)	2.699 (0.701)
	β_3 (real value = 4)	4.556 (0.703)	4.477 (0.716)	4.343 (0.792)	4.256 (0.779)
	β_4 (real value = 0)	–0.054 (0.255)	–0.031 (0.155)	–0.055 (0.345)	–0.080 (0.249)
	β_5 (real value = –1)	–1.137 (0.366)	–1.175 (0.322)	–1.099 (0.382)	–1.110 (0.322)
	$\alpha = w\_scale$ (real value = 1)	1.006 (0.112)	1.008 (0.118)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.251 (0.334)	2.217 (0.324)	-	-
$n = 100$	β_1 (real value = 0)	–0.025 (0.155)	–0.01 (0.070)	–0.031 (0.158)	–0.010 (0.072)
	β_2 (real value = 2.5)	2.637 (0.420)	2.612 (0.388)	2.526 (0.473)	2.494 (0.459)
	β_3 (real value = 4)	4.227 (0.619)	4.203 (0.599)	4.012 (0.663)	3.988 (0.662)
	β_4 (real value = 0)	–0.004 (0.141)	0.000 (0.000)	0.022 (0.152)	0.000 (0.000)
	β_5 (real value = –1)	–1.063 (0.204)	–1.058 (0.202)	–1.017 (0.243)	–0.998 (0.223)
	$\alpha = w\_scale$ (real value = 1)	0.982 (0.059)	0.984 (0.061)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.110 (0.284)	2.098 (0.275)	-	-

ΠΙΝΑΚΑΣ 24: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – Parametric Case w. Inverse Gaussian Frailty, 5 Covariates & Censoring Proportion ≈ 80%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 0)	0.009 (0.575)	0.009 (0.426)	0.180 (1.250)	0.207 (0.865)
	β_2 (real value = 2.5)	3.085 (1.045)	2.993 (1.084)	4.155 (2.437)	3.694 (2.054)
	β_3 (real value = 4)	4.755 (1.794)	4.577 (1.830)	6.144 (3.215)	5.540 (3.178)
	β_4 (real value = 0)	–0.078 (0.627)	–0.086 (0.545)	–0.075 (1.180)	–0.179 (0.889)
	β_5 (real value = –1)	–1.609 (0.465)	–1.623 (0.424)	–1.710 (1.699)	–2.225 (1.008)
	$\alpha = w\_scale$ (real value = 1)	0.859 (0.220)	0.893 (0.251)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.409 (0.755)	2.322 (0.777)	-	-
$n = 100$	β_1 (real value = 0)	–0.048 (0.303)	–0.041 (0.209)	–0.051 (0.390)	0.034 (0.319)
	β_2 (real value = 2.5)	2.812 (0.757)	2.741 (0.800)	2.950 (0.937)	2.762 (0.866)
	β_3 (real value = 4)	4.347 (0.899)	4.230 (0.959)	4.476 (1.157)	4.164 (1.112)
	β_4 (real value = 0)	–0.040 (0.377)	–0.057 (0.354)	–0.033 (0.506)	–0.071 (0.358)
	β_5 (real value = –1)	–1.022 (0.348)	–1.102 (0.327)	–1.079 (0.409)	–1.187 (0.339)
	$\alpha = w\_scale$ (real value = 1)	0.871 (0.130)	0.912 (0.192)	-	-
	$\lambda = w\_shape$ (real value = 2)	2.176 (0.379)	2.125 (0.415)	-	-

4.4. ΠΡΟΣΟΜΟΙΩΣΕΙΣ ΓΙΑ ΤΗΝ ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ ΚΑΙ ΤΗΝ ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΤΩΝ ΗΜΙΠΑΡΑΜΕΤΡΙΚΩΝ ΜΟΝΤΕΛΩΝ ΓΑΜΜΑ ΚΑΙ INVERSE GAUSSIAN ΕΥΠΑΘΕΙΑΣ ΜΕΣΩ ΤΟΥ ΜΟΝΤΕΛΟΥ ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΥ

4.4.1. Εισαγωγή: Θα επαναλάβουμε την προσομοίωση και επιλογή των μεταβλητών, αλλά τώρα στα ημιπαραμετρικά μοντέλα Γάμμα και Inverse Gaussian ευπάθειας. Εδώ, η βασική συνάρτηση κινδύνου $h_0(t)$ θεωρείται άγνωστη "οχληρά παράμετρος" άπειρης διάστασης και δεν ακολουθεί γνωστή κατανομή.

Προκειμένου να αντιμετωπίσουμε το πρόβλημα, ακολουθούμε την ιδέα του Breslow που περιγράφεται στο [59] (pp. 80), δηλ. από το δείγμα $\mathrm{A} = \{(X_i, D_i, \mathrm{u}_i) : i = 1, 2, \ldots, n\}$, θεωρούμε τους διατεταγμένους χρόνους

αποτυχίας $t_{(1)} < t_{(2)} < \ldots < t_{(m)}$, όπου $t_{(i)} \in \{X_1, X_2, \ldots, X_n\}$ με $D_i = 1$ για $i = 1, 2, \ldots, m$ και θεωρούμε τη συνάρτηση κινδύνου ως σταθερή στα επιμέρους διαστήματα μεταξύ των χρόνων αποτυχίας, δηλ:

$$h_0(t) = \begin{cases} 0 & \text{αν } t < t_{(1)} \\ \lambda_1 & \text{αν } t_{(1)} \le t < t_{(2)} \\ \lambda_2 & \text{αν } t_{(2)} \le t < t_{(3)} \\ \ldots & \ldots \\ \lambda_m & \text{αν } t_{(m)} \le t \end{cases} \quad \textbf{(4.13)}$$

Η σωρευτική συνάρτηση κινδύνου $H_0(t)$, θα είναι τότε αύξουσα, κατά διαστήματα σταθερή και θα κάνει άλματα στους χρόνους διακοπής:

$$H_0(t) = \begin{cases} 0 & \text{αν } t < t_{(1)} \\ \lambda_1 & \text{αν } t_{(1)} \le t < t_{(2)} \\ \lambda_1 + \lambda_2 & \text{αν } t_{(2)} \le t < t_{(3)} \\ \ldots & \ldots \\ \lambda_1 + \lambda_2 + \ldots + \lambda_m & \text{αν } t_{(m)} \le t \end{cases} \quad \textbf{(4.14)}$$

ή ισοδύναμα

$$H_0(t) = \begin{cases} 0 & \text{αν } t < t_{(1)} \\ \sum_{j: t_{(j)} \le t} h_0\left(t_{(j)}\right) & \text{αν } t \ge t_{(1)} \end{cases} \quad \textbf{(4.15)}$$

4.4.2. Κατασκευή του κώδικα: Κρατώντας την αρχικοποίηση που είχαμε δώσει στον κώδικα των παραμετρικών μοντέλων ευπάθειας (αφαιρώντας μόνο τους τύπους των παραμετρικών συναρτήσεων $h_0(t)$ και $H_0(t)$), κατασκευάζουμε και πάλι τις ευπάθειες και τους χρόνους αποτυχίας ή λογοκρισίας των μονάδων.

Ακολούθως, γράφουμε το κομμάτι του κώδικα που περιγράφει τη βασική συνάρτηση κινδύνου και τη βασική σωρευτική συνάρτηση κινδύνου, υλοποιώντας έτσι τις εξισώσεις **(4.13)** και **(4.15)**.

```
#=============================================
#hazard function


h<-function(t, time, haz)
```

```
{       hazard<-0
        for (i in 1:failure_num)
        { if (t>=time[i]) {hazard<-haz[i]}       }
        return(hazard)
} #end function

#================================================
#Cumulative hazard function

H<-function(t,time, haz)
{
        cum_haz<-0
        for (k in 1:failure_num)
        {if (t>=time[k]) {cum_haz<-cum_haz+h(time[k], time, haz)}}
        return(cum_haz)
}       #end function
```

Η λογαριθμοποιημένη πιθανοφάνεια είναι τώρα συνάρτηση με μεταβλητές τόσο τους συντελεστές $\boldsymbol{\beta}$, όσο και τις τιμές $\mathbf{h} = (h(t_{(1)}), \ldots, h(t_{(m)}))$ των διατεταγμένων διακεκριμένων χρόνων αποτυχίας, δηλ. ισχύει ότι $\ell = \ell(\boldsymbol{\beta}, \mathbf{h})$. Το μειονέκτημα εδώ είναι ότι πρόκειται για συνάρτηση μεγάλου πλήθους μεταβλητών και αυτό δημιουργεί πρόβλημα στην απευθείας βελτιστοποίησή της (δηλ. στην εύρεση ταυτόχρονα και των συντελεστών $\boldsymbol{\beta}$ και των τιμών $\mathbf{h} = (h(t_{(1)}), \ldots, h(t_{(m)}))$ της συνάρτησης κινδύνου που μεγιστοποιούν τη συνάρτηση ℓ).

Προς τούτο, προχωρήσαμε στη βελτιστοποίηση τμηματικά, σε δύο βήματα (*2-step method*) ως εξής:

- Προσαρμόζουμε στα δεδομένα το μοντέλο του Cox, χρησιμοποιώντας τη συνάρτηση *coxph*. Από αυτήν, εξάγουμε τις εκτιμημένες τιμές της βασικής συνάρτησης κινδύνου $\hat{h}_0(t_{(i)})$, $i = 1, 2, \ldots, m$ και τις εκτιμήσεις $\hat{\boldsymbol{\beta}}_{coxph}$ των συντελεστών $\boldsymbol{\beta}$.

- Θεωρούμε τη λογαριθμοποιημένη πιθανοφάνεια ως συνάρτηση με μεταβλητή το διάνυσμα $\mathbf{h} = (h(t_{(1)}), \ldots, h(t_{(m)}))$ των τιμών της συνάρτησης κινδύνου, δηλ. $\ell_{hazard} = \ell(\mathbf{h})$. Το διάνυσμα των συντελεστών $\boldsymbol{\beta}$ θεωρείται εδώ ως σταθερά με τιμή $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{coxph}$.

- Θέτοντας ως αρχικές τιμές τις εκτιμημένες από την *coxph* τιμές $\hat{h}_0(t_{(i)})$, $i = 1, 2, \ldots, m$, της βασικής συνάρτησης κινδύνου, υπολογίζουμε το διάνυσμα $\hat{\mathbf{h}} = (\hat{h}(t_{(1)}), \ldots, \hat{h}(t_{(m)}))$ των τιμών της συνάρτησης κινδύνου που βελτιστοποιεί την ℓ_{hazard} και έτσι ολοκληρώνεται το πρώτο βήμα.

- Θεωρούμε στη συνέχεια τη λογαριθμοποιημένη πιθανοφάνεια ως συνάρτηση με μεταβλητή το διάνυσμα $\boldsymbol{\beta}$ των συντελεστών, δηλ. $\ell_{coef} = \ell(\boldsymbol{\beta})$. Τώρα, το διάνυσμα των τιμών της συνάρτησης κινδύνου θεωρείται σταθερά με τιμή την εκτίμηση $\hat{\mathbf{h}} = (\hat{h}(t_{(1)}), \ldots, \hat{h}(t_{(m)}))$ που προέκυψε στο πρώτο βήμα της βελτιστοποίησης.

- Θέτοντας ως αρχικές τιμές τις εξαχθείσες από την *coxph* εκτιμήσεις $\hat{\boldsymbol{\beta}}_{coxph}$ των συντελεστών, υπολογίζουμε το διάνυσμα $\hat{\boldsymbol{\beta}}$ των συντελεστών που βελτιστοποιεί την ℓ_{coef} και ολοκληρώνεται και το δεύτερο βήμα.

- Υπολογίζουμε την τιμή $\ell(\hat{\boldsymbol{\beta}}, \hat{\mathbf{h}})$ της πιθανοφάνειας και από αυτήν, τις τιμές των AIC και BIC.

Για να υλοποιηθούν τα παραπάνω, χρειάζεται αρχικά να γράψουμε τη λογαριθμοποιημένη πιθανοφάνεια υπό τις μορφές $\ell = \ell_{hazard} = \ell(\mathbf{h})$, $\ell = \ell_{coef} = \ell(\boldsymbol{\beta})$ και $\ell = \ell(\boldsymbol{\beta}, \mathbf{h})$

```
#=============================================
#Making of the log-likelihood function via G function

log_likelihood_hazard<-function(hazard)
{       log_lik<-0
                for (i in 1:n)
                {       prod<-0
                        for (j in 1:p)      {prod<-prod+beta[j]*zeta[i,j]}
```

```
                    m<-exp(prod)*H(X[i],time_vector, hazard)
                    v<-D[i]*(prod+log(dG(m))+log(h(X[i],time_vector, hazard)))-G(m)
                    log_lik<-log_lik+v        }
        return(-log_lik)

                                                                    }


log_likelihood_beta<-function(beta)
{       log_lik<-0
                for (i in 1:n)
                {       prod<-0
                        for (j in 1:p)      {prod<-prod+beta[j]*zeta[i,j]}
                        m<-exp(prod)*H(X[i],time_vector, hz)
                        v<-D[i]*(prod+log(dG(m))+log(h(X[i],time_vector, hz)))-G(m)
                        log_lik<-log_lik+v        }
        return(-log_lik)

                                                                    }

log_likelihood<-function(param)
{       beta<-param[1:p]
        hz<-param[(p+1):(p+failure_num)]
        log_lik<-0
                for (i in 1:n)
                {       prod<-0
                        for (j in 1:p)      {prod<-prod+beta[j]*zeta[i,j]}
                        m<-exp(prod)*H(X[i],time_vector, hz)
                        v<-D[i]*(prod+log(dG(m))+log(h(X[i],time_vector, hz)))-G(m)
                        log_lik<-log_lik+v        }
        return(-log_lik)

                                                                    }
```

Ακολούθως, χρειάζεται να καλέσουμε την *coxph* για να προσαρμόσουμε με αυτήν το μοντέλο του Cox και να εξάγουμε την εκτίμηση $\hat{h}_0(t)$ της βασικής συνάρτησης κινδύνου και την εκτίμηση $\hat{\boldsymbol{\beta}}_{coxph}$ των συντελεστών. Στο σημείο αυτό, πρέπει να παρατηρήσουμε τα εξής:

- Κατά την εκτέλεση της *coxph* δεν χρησιμοποιούμε την επιλογή *frailty*, προκειμένου η προσαρμογή του μοντέλου του Cox να ανταποκρίνεται σε ημιπαραμετρική προσέγγιση πραγματικών δεδομένων (στα οποία προφανώς δε θα γνωρίζουμε την ύπαρξη ή μη της ευπάθειας).
- Μετά την προσαρμογή του μοντέλου του Cox μέσω της *coxph*, καλούμε τη συνάρτηση *basehaz* της R, με την οποία εξάγεται η σωρευτική βασική συνάρτηση κινδύνου $H_0(t)$ του προσαρμοσμένου μοντέλου. Στη συνέχεια, η βασική συνάρτηση κινδύνου $h_0(t)$ υπολογίζεται από τον τύπο $h_0(t_{(1)}) = H_0(t_{(1)})$ και $h_0(t_{(i)}) = H_0(t_{(i)}) - H_0(t_{(i-1)})$ για $i = 2, 3, \ldots, m$.

Προκειμένου να παρακολουθήσουμε πλήρως την υλοποίηση της προσαρμογής του μοντέλου του Cox και της βελτιστοποίησης σε 2 βήματα, παρουσιάζουμε ακολούθως όλο το κομμάτι του κώδικα που αφορά τα σημεία αυτά. Τα επίμαχα σημεία εμφανίζονται με σκούρο κόκκινο χρώμα.

```
#=============================================
#Optimization of likelihood - Selection of Variables

results<-matrix(nrow=2*k+8, ncol=2^k-1)      #Storage of coefficients, AIC, BIC of the 2^k-1 models

for (i in 1:k)
{                        results[i,]<-0
                         results[k+4+i,]<-0}

for (p in 1:k)     #p is the number of covariates of fitted model
{

#Column-group of results matrix, where estimations will be stored
lower_column<-0
```

```
for (u in 0:(p-1)) {lower_column<-lower_column+choose(k,u)}

#par_vector<-numeric(p+failure_num)  #vector with the estimated parameters
idx<-matrix(combn(k,p),nrow=p)          #elements of the matrix are the combinations of k by
p
#print(idx)

row<-numeric(p)          #vector which shows the rows of the results matrix
formula_constr<-numeric(p) #Will be used for creating the cox(Survreg) string formula
id<-seq(1:n)
name_col<-numeric(p)  #elements of this vector show which covariates are used
initial_beta<-numeric(p)

zeta<-matrix(nrow=n, ncol=p)  #Matrix which contains values of covariates that are used

for (j in 1:ncol(idx))

        {       results_column_number<-lower_column+j-1
                for (u in 1:p)
                        {element<-idx[u,j]        #the (u,j)-element of idx matrix
                        zeta[,u]<-Z[,element]    #u-th column of zeta is the element-column
of Z
                        row[u]<-element
                        formula_constr[u]<-element
                        name_col[u]<- paste("Z",element)}

        #fit models with coxph function
        formula_string_initial<-paste("Surv(X,D)~Z[", formula_constr[1],"]")
        data_string_initial<-paste("list(X,D,")
        {if (p==1){formula_string<-paste(formula_string_initial)}
        if                                                              (p==2){formula_string<-
paste(formula_string_initial,"+Z[",formula_constr[2],"]")}
        if (p>2)
        {       formula_string<-formula_string_initial
```

```
                {for (u in 2:(p-1)) {formula_string<-
paste(formula_string,"+Z[,",formula_constr[u],"]")}
                 for (u in (p:p)) {formula_string<-paste(formula_string,"+Z[,",
formula_constr[u], "]")}}
        }

        data_string<-data_string_initial
                for (u in 1:(k-1)) {data_string<-paste(data_string,"Z[,", u, "],")}
                for (u in k:k) {data_string<-paste(data_string,"Z[,", u, "],id)")}}

        coxph_formula<-as.formula(formula_string)
        coxph_data<-as.data.frame(data_string)

        cox_fit<-coxph(coxph_formula,data=coxph_data)

        for (y in 1:p){
                                results[k+4+row[y],results_column_number]<-
cox_fit$coef[y]
                                initial_beta[y]<-cox_fit$coef[y]
        }
        results[2*k+5,results_column_number]<-extractAIC(cox_fit)[2]
        results[2*k+6,results_column_number]<-extractAIC(cox_fit, k = log(n))[2]

                cum_basehaz<-basehaz(cox_fit,centered=FALSE)
                cum_basehaz<-cum_basehaz[!duplicated(cum_basehaz[,1]),]
                cum_basehaz<-cum_basehaz[is.finite(cum_basehaz[,1]),]
                cum_hazard_vector<-cum_basehaz[,1]

                failure_num<-nrow(cum_basehaz)

                baseline_hazard<-numeric(failure_num)
                baseline_hazard[1]<-cum_hazard_vector[1]
                for (i in 2:failure_num) {baseline_hazard[i]<-cum_hazard_vector[i]-
cum_hazard_vector[i-1]}
```

```
                time_vector<-cum_basehaz[,2]



                for (i in 1:failure_num) {
                if  (baseline_hazard[failure_num+1-i]==0)  {baseline_hazard[failure_num+1-
i]<-baseline_hazard[failure_num+2-i]/2}
                if (baseline_hazard[i]==Inf) {baseline_hazard[i]<-baseline_hazard[i-1]+1}
                                                        }



                initial_vector<-numeric(p+failure_num)                  #Initial  vector  for
minimization

                initial_vector<-c(initial_beta, baseline_hazard)

        ui<-matrix(nrow=failure_num, ncol=failure_num)
        ui<-diag(failure_num)
        ci<-numeric(failure_num)
        for (i in 1:failure_num)  {ci[i]<-0}

        beta<-initial_beta
        opt_hazard<-constrOptim(baseline_hazard, log_likelihood_hazard, gr=NULL, ui=ui,
ci=ci,method="SANN",control=list(maxit=150))

        hz<-numeric(failure_num)
        hz<-opt_hazard$par

        opt<-nlm(log_likelihood_beta, initial_beta, iterlim = 50)
        beta_est<-opt$est

        l<-log_likelihood(c(beta_est, hz))
        AIC<-2*l+2*p
        BIC<-2*l+p*log(n)

        for (y in 1:p){results[row[y],results_column_number]<-round(beta_est[y],8)}
```

```
    results[k+1,results_column_number]<-AIC
    results[k+2,results_column_number]<-BIC
```

Στο παραπάνω απόσπασμα του κώδικα, οι εντολές

for (i in 1:failure_num) {

if (hazard_param[failure_num+1-i]==0) {hazard_param[failure_num+1-i]<-hazard_param[failure_num+2-i]/2}

if (hazard_param[i]==Inf) {hazard_param[i]<-hazard_param[i-1]+1}

}

αποσκοπούν στο να αποφευχθεί να δοθεί στο αρχικό διάνυσμα της βελτιστοποίησης, συντεταγμένη ίση με 0 ή με ∞, δεδομένου ότι τότε, κατά την εκτέλεση της συνάρτησης *log_likelihood*, ο λογάριθμος *log(h(X[i],time_vector, hz))* θα δώσει μη επιτρεπτές τιμές.

Όλα τα υπόλοιπα κομμάτια του κώδικα παραμένουν μέχρι το τέλος ίδια με αυτά της παραμετρικής περίπτωσης.

4.4.3. Αποτελέσματα στην ημιπαραμετρική περίπτωση της Γάμμα ευπάθειας: Εκτελέσαμε τον κώδικα για το ημιπαραμετρικό μοντέλο Γάμμα ευπάθειας με *niter* = 100 το πλήθος επαναλήψεις. Για τα ποσοστά επιτυχίας έχουμε τους ακόλουθους 3 πίνακες:

ΠΙΝΑΚΑΣ 25: ΠΟΣΟΣΤΑ ΕΠΙΤΥΧΙΑΣ – SemiParametric Case with Gamma Frailty and Censoring Proportion ≈ 10%									
Number of Covariates	Sample Size	Our Method - AIC				Coxph - AIC			
		Perc. of Success	Error1	Error2	Error1&2	Perc. of Success	Error1	Error2	Error1&2
$k=3$	$n=50$	0.84	0.16	0	0	0.83	0.17	0	0
	$n=100$	0.78	0.22	0	0	0.77	0.23	0	0
$k=4$	$n=50$	0.67	0.33	0.05	0.05	0.65	0.35	0.06	0.06
	$n=100$	0.68	0.32	0	0	0.69	0.31	0	0
$k=5$	$n=50$	0.53	0.32	0.27	0.12	0.66	0.34	0.01	0.01
	$n=100$	0.68	0.28	0.15	0.11	0.77	0.23	0	0
Number of Covariates	Sample Size	Our Method - BIC				Coxph - BIC			
		Perc. of Success	Error1	Error2	Error1&2	Perc. of Success	Error1	Error2	Error1&2
$k=3$	$n=50$	0.92	0.08	0	0	0.92	0.08	0	0
	$n=100$	0.97	0.03	0	0	0.77	0.23	0	0
$k=4$	$n=50$	0.86	0.10	0.09	0.05	0.84	0.12	0.09	0.05
	$n=100$	0.88	0.12	0	0	0.88	0.12	0	0
$k=5$	$n=50$	0.69	0.08	0.29	0.06	0.85	0.15	0.01	0.01
	$n=100$	0.82	0.11	0.16	0.09	0.94	0.06	0	0
ΠΙΝΑΚΑΣ 26: ΠΟΣΟΣΤΑ ΕΠΙΤΥΧΙΑΣ – SemiParametric Case with Gamma Frailty and Censoring Proportion ≈ 46%									
Number of Covariates	Sample Size	Our Method - AIC				Coxph - AIC			
		Perc. of Success	Error1	Error2	Error1&2	Perc. of Success	Error1	Error2	Error1&2
$k=3$	$n=50$	0.88	0.12	0	0	0.86	0.14	0	0
	$n=100$	0.84	0.16	0	0	0.81	0.19	0	0
$k=4$	$n=50$	0.67	0.30	0.07	0.04	0.7	0.27	0.07	0.04
	$n=100$	0.73	0.27	0	0	0.74	0.26	0	0
$k=5$	$n=50$	0.33	0.36	0.53	0.22	0.66	0.33	0.03	0.02
	$n=100$	0.63	0.28	0.22	0.13	0.67	0.33	0	0
Number of Covariates	Sample Size	Our Method - BIC				Coxph - BIC			
		Perc. of Success	Error1	Error2	Error1&2	Perc. of Success	Error1	Error2	Error1&2
$k=3$	$n=50$	0.97	0.02	0.01	0	0.98	0.01	0.01	0
	$n=100$	0.95	0.05	0	0	0.95	0.05	0	0
$k=4$	$n=50$	0.78	0.14	0.17	0.09	0.8	0.12	0.16	0.08
	$n=100$	0.95	0.04	0.01	0	0.93	0.06	0.01	0
$k=5$	$n=50$	0.33	0.20	0.63	0.16	0.80	0.11	0.12	0.03
	$n=100$	0.71	0.09	0.27	0.07	0.94	0.06	0.01	0.01

ΠΙΝΑΚΑΣ 27: ΠΟΣΟΣΤΑ ΕΠΙΤΥΧΙΑΣ – SemiParametric Case with Gamma Frailty and Censoring Proportion ≈ 80%									
Number of Covariates	**Sample Size**	**Our Method - AIC**				**Coxph - AIC**			
		Perc. of Success	**Error1**	**Error2**	**Error1&2**	**Perc. of Success**	**Error1**	**Error2**	**Error1&2**
$k=3$	$n=50$	0.65	0.13	0.27	0.05	0.73	0.15	0.17	0.05
	$n=100$	0.77	0.23	0.04	0.04	0.78	0.22	0.03	0.03
$k=4$	$n=50$	0.24	0.45	0.59	0.28	0.27	0.48	0.53	0.28
	$n=100$	0.47	0.33	0.36	0.16	0.47	0.34	0.34	0.15
$k=5$	$n=50$	0.06	0.40	0.90	0.36	0.32	0.47	0.4	0.19
	$n=100$	0.25	0.48	0.61	0.34	0.53	0.44	0.12	0.09
Number of Covariates	**Sample Size**	**Our Method – BIC**				**Coxph - BIC**			
		Perc. of Success	**Error1**	**Error2**	**Error1&2**	**Perc. of Success**	**Error1**	**Error2**	**Error1&2**
$k=3$	$n=50$	0.53	0.06	0.45	0.04	0.68	0.07	0.31	0.06
	$n=100$	0.89	0.07	0.07	0.03	0.89	0.07	0.05	0.01
$k=4$	$n=50$	0.17	0.24	0.79	0.20	0.22	0.24	0.73	0.19
	$n=100$	0.41	0.15	0.55	0.11	0.42	0.14	0.54	0.10
$k=5$	$n=50$	0.05	0.20	0.92	0.17	0.31	0.21	0.58	0.10
	$n=100$	0.23	0.21	0.74	0.18	0.65	0.17	0.28	0.10

Από τους πίνακες 25, 26, 27 παρατηρούμε ότι για να έχουμε υψηλό ποσοστό επιτυχίας με τη μέθοδό μας, πρέπει τα δεδομένα να δίνουν "αρκετή πληροφορία", δηλ. να έχουμε μειωμένο ποσοστό λογοκρισίας ή αυξημένο μέγεθος δείγματος ή και τα δύο. Όταν αυτό δε συμβαίνει, τα ποσοστά επιτυχίας της μεθόδου μας επηρεάζονται βασικά από αύξηση του $Error2$ χωρίς να υπάρχει αντίστοιχη αύξηση του $Error1\&2$, δηλ. από επιλογές μοντέλων που περιέχουν μόνο σημαντικές μεταβλητές αλλά όχι όλες (π.χ. στον πίνακα 27, για $k=5$ συμμεταβλητές και $n=50$ έχουμε ποσοστό επιτυχίας μόλις 0.05, με $Error1=0.20$, $Error2=0.92$ και $Error1\&2=0.17$, επομένως, από ποσοστό $1-0.05=0.95$ των εσφαλμένων μοντέλων, το $0.92-0.17=0.75$ αφορά μοντέλα που είχαν μόνο σημαντικές μεταβλητές αλλά όχι όλες, ενώ μόλις το $0.20-0.17=0.03$ αφορά μοντέλα που δεν είχαν καθόλου σημαντικές μεταβλητές).

Η εκτίμηση των συντελεστών φαίνεται τώρα στους ακόλουθους πίνακες.

ΠΙΝΑΚΑΣ 28: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – SemiParametric Case with Gamma Frailty, 3 Covariates & Censoring Proportion ≈ 10%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 2)	1.852 (0.31)	1.844 (0.298)	1.864 (0.313)	1.856 (0.302)
	β_2 (real value = 0)	-0.018 (0.195)	-0.012 (0.164)	-0.014 (0.194)	-0.011 (0.157)
	β_3 (real value = –1.3)	-1.161 (0.28)	-1.162 (0.277)	-1.168 (0.278)	-1.166 (0.279)
$n = 100$	β_1 (real value = 2)	1.823 (0.221)	1.825 (0.206)	1.82 (0.224)	1.823 (0.209)
	β_2 (real value = 0)	0.023 (0.139)	0.003 (0.067)	0.033 (0.137)	0.01 (0.079)
	β_3 (real value = –1.3)	-1.208 (0.176)	-1.195 (0.172)	-1.214 (0.174)	-1.201 (0.174)
ΠΙΝΑΚΑΣ 29: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – SemiParametric Case with Gamma Frailty, 3 Covariates & Censoring Proportion ≈ 46%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 2)	1.895 (0.391)	1.873 (0.398)	1.927 (0.388)	1.905 (0.394)
	β_2 (real value = 0)	0.003 (0.236)	0.022 (0.171)	0.003 (0.236)	0.015 (0.151)
	β_3 (real value = –1.3)	-1.281 (0.346)	-1.289 (0.337)	-1.292 (0.35)	-1.293 (0.341)
$n = 100$	β_1 (real value = 2)	1.869 (0.27)	1.861 (0.264)	1.885 (0.269)	1.876 (0.262)
	β_2 (real value = 0)	-0.009 (0.185)	0.003 (0.139)	-0.006 (0.185)	0.003 (0.135)
	β_3 (real value = –1.3)	-1.223 (0.212)	-1.225 (0.209)	-1.237 (0.21)	-1.235 (0.205)
ΠΙΝΑΚΑΣ 30: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – SemiParametric Case with Gamma Frailty, 3 Covariates & Censoring Proportion ≈ 80%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 2)	1.883 (0.801)	1.776 (0.775)	2.22 (0.902)	2.209 (0.874)
	β_2 (real value = 0)	-0.013 (0.537)	0.003 (0.398)	-0.025 (0.661)	-0.071 (0.518)
	β_3 (real value = –1.3)	-1.541 (0.729)	-1.567 (0.749)	-1.679 (0.739)	-1.694 (0.724)
$n = 100$	β_1 (real value = 2)	1.928 (0.624)	1.898 (0.623)	2.078 (0.691)	2.052 (0.684)
	β_2 (real value = 0)	-0.024 (0.369)	-0.017 (0.262)	-0.026 (0.351)	-0.012 (0.246)
	β_3 (real value = –1.3)	-1.268 (0.403)	-1.279 (0.377)	-1.373 (0.448)	-1.384 (0.433)

ΠΙΝΑΚΑΣ 31: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – SemiParametric Case with Gamma Frailty, 4 Covariates & Censoring Proportion ≈ 10%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 0.8)	0.757 (0.267)	0.768 (0.245)	0.772 (0.261)	0.776 (0.238)
	β_2 (real value = 0)	0.014 (0.254)	0.021 (0.191)	0.005 (0.262)	0.02 (0.201)
	β_3 (real value = 0)	0.019 (0.184)	0.008 (0.133)	0.018 (0.179)	0.007 (0.13)
	β_4 (real value = 1)	0.961 (0.27)	0.954 (0.258)	0.958 (0.264)	0.949 (0.256)
$n = 100$	β_1 (real value = 0.8)	0.699 (0.177)	0.695 (0.167)	0.695 (0.175)	0.695 (0.16)
	β_2 (real value = 0)	-0.003 (0.147)	-0.003 (0.103)	0.002 (0.136)	-0.006 (0.099)
	β_3 (real value = 0)	0.007 (0.131)	0.009 (0.094)	0.003 (0.126)	0.005 (0.085)
	β_4 (real value = 1)	0.934 (0.165)	0.928 (0.158)	0.923 (0.161)	0.92 (0.151)
ΠΙΝΑΚΑΣ 32: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – SemiParametric Case with Gamma Frailty, 4 Covariates & Censoring Proportion ≈ 46%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 0.8)	0.799 (0.281)	0.833 (0.246)	0.812 (0.281)	0.847 (0.24)
	β_2 (real value = 0)	0.03 (0.181)	0.037 (0.151)	0.03 (0.192)	0.031 (0.14)
	β_3 (real value = 0)	-0.012 (0.264)	-0.025 (0.21)	-0.013 (0.246)	-0.029 (0.196)
	β_4 (real value = 1)	0.987 (0.341)	0.987 (0.338)	0.996 (0.338)	0.992 (0.333)
$n = 100$	β_1 (real value = 0.8)	0.739 (0.196)	0.743 (0.187)	0.74 (0.19)	0.744 (0.179)
	β_2 (real value = 0)	-0.004 (0.119)	-0.011 (0.075)	0.004 (0.099)	-0.001 (0.064)
	β_3 (real value = 0)	0.008 (0.159)	0 (0.056)	0.002 (0.152)	0 (0.075)
	β_4 (real value = 1)	0.939 (0.231)	0.94 (0.223)	0.942 (0.22)	0.941 (0.219)
ΠΙΝΑΚΑΣ 33: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – SemiParametric Case with Gamma Frailty, 4 Covariates & Censoring Proportion ≈ 80%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 0.8)	1.254 (0.737)	1.277 (0.743)	1.355 (1.017)	1.4 (1.031)
	β_2 (real value = 0)	0.009 (0.687)	-0.014 (0.585)	0.048 (0.736)	0 (0.614)
	β_3 (real value = 0)	-0.007 (0.528)	0.098 (0.429)	-0.044 (0.656)	0.09 (0.507)
	β_4 (real value = 1)	1.286 (0.555)	1.225 (0.525)	1.493 (1.141)	1.409 (1.18)
$n = 100$	β_1 (real value = 0.8)	0.9 (0.314)	0.914 (0.275)	0.938 (0.33)	0.94 (0.276)
	β_2 (real value = 0)	-0.011 (0.319)	0.005 (0.217)	-0.012 (0.324)	-0.005 (0.192)
	β_3 (real value = 0)	0.035 (0.353)	0.036 (0.284)	0.032 (0.354)	0.051 (0.261)
	β_4 (real value = 1)	0.995 (0.328)	1.022 (0.31)	1.03 (0.333)	1.052 (0.314)

ΠΙΝΑΚΑΣ 34: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – SemiParametric Case with Gamma Frailty, 5 Covariates & Censoring Proportion ≈ 10%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 0)	-0.009 (0.135)	-0.009 (0.087)	0.005 (0.153)	0.008 (0.107)
	β_2 (real value = 2.5)	2.112 (0.365)	2.087 (0.369)	2.384 (0.416)	2.363 (0.399)
	β_3 (real value = 4)	3.337 (0.552)	3.299 (0.553)	3.824 (0.636)	3.785 (0.61)
	β_4 (real value = 0)	-0.042 (0.244)	-0.04 (0.165)	-0.028 (0.258)	-0.014 (0.216)
	β_5 (real value = –1)	-0.959 (0.256)	-0.942 (0.248)	-0.968 (0.267)	-0.962 (0.262)
$n = 100$	β_1 (real value = 0)	-0.008 (0.11)	0.006 (0.058)	0.019 (0.13)	0.012 (0.093)
	β_2 (real value = 2.5)	2.1 (0.283)	2.081 (0.287)	2.291 (0.281)	2.287 (0.278)
	β_3 (real value = 4)	3.4 (0.388)	3.379 (0.39)	3.691 (0.38)	3.676 (0.377)
	β_4 (real value = 0)	-0.056 (0.167)	-0.05 (0.155)	-0.005 (0.096)	-0.004 (0.04)
	β_5 (real value = –1)	-0.882 (0.173)	-0.879 (0.172)	-0.91 (0.167)	-0.905 (0.162)
ΠΙΝΑΚΑΣ 35: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – SemiParametric Case with Gamma Frailty, 5 Covariates & Censoring Proportion ≈ 46%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 0)	0.08 (0.289)	0.053 (0.236)	0.029 (0.267)	0.019 (0.209)
	β_2 (real value = 2.5)	2.065 (0.542)	2.01 (0.497)	2.535 (0.668)	2.489 (0.612)
	β_3 (real value = 4)	3.308 (0.674)	3.185 (0.704)	4.235 (1.063)	4.12 (0.975)
	β_4 (real value = 0)	-0.109 (0.331)	-0.083 (0.281)	-0.036 (0.298)	-0.039 (0.211)
	β_5 (real value = –1)	-0.9 (0.34)	-0.961 (0.263)	-1.026 (0.34)	-1.056 (0.273)
$n = 100$	β_1 (real value = 0)	-0.017 (0.137)	0.001 (0.061)	-0.001 (0.153)	-0.004 (0.044)
	β_2 (real value = 2.5)	2.132 (0.343)	2.093 (0.349)	2.434 (0.387)	2.42 (0.371)
	β_3 (real value = 4)	3.382 (0.497)	3.323 (0.551)	3.837 (0.576)	3.805 (0.572)
	β_4 (real value = 0)	-0.036 (0.191)	-0.033 (0.141)	-0.008 (0.167)	-0.005 (0.125)
	β_5 (real value = –1)	-0.908 (0.217)	-0.897 (0.208)	-0.942 (0.231)	-0.947 (0.213)

ΠΙΝΑΚΑΣ 36: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – SemiParametric Case with Gamma Frailty, 5 Covariates & Censoring Proportion ≈ 80%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 0)	0.15 (0.685)	0.098 (0.607)	0.133 (0.982)	0.04 (0.733)
	β_2 (real value = 2.5)	2.473 (1.409)	2.416 (1.417)	4.298 (2.86)	4.046 (2.443)
	β_3 (real value = 4)	3.076 (1.388)	2.979 (1.326)	6.389 (3.993)	5.614 (3.264)
	β_4 (real value = 0)	-0.077 (0.59)	-0.008 (0.39)	-0.231 (0.99)	-0.163 (0.638)
	β_5 (real value = –1)	-1 (1.105)	-1.618 (0.738)	-2.098 (1.279)	-2.421 (1.118)
$n = 100$	β_1 (real value = 0)	0.07 (0.381)	0.062 (0.366)	0.049 (0.419)	0.047 (0.251)
	β_2 (real value = 2.5)	2.088 (0.513)	2.039 (0.501)	2.63 (0.696)	2.552 (0.687)
	β_3 (real value = 4)	3.281 (0.955)	3.078 (0.941)	4.506 (1.17)	4.31 (1.206)
	β_4 (real value = 0)	-0.199 (0.493)	-0.15 (0.426)	-0.154 (0.458)	-0.141 (0.398)
	β_5 (real value = –1)	-1.048 (0.308)	-1.125 (0.284)	-1.147 (0.424)	-1.24 (0.373)

Από τους πίνακες εκτίμησης παραμέτρων προκύπτει ότι η μέθοδός μας έχει γενικά την τάση να δίνει χαμηλότερες εκτιμήσεις για τους συντελεστές από την *coxph*. Επίσης, στις περιπτώσεις που η "πληροφορία" δεν είναι αρκετή, βλέπουμε την τάση να υποεκτιμώνται κάποιοι από τους συντελεστές, απόρροια του υψηλού $Error2$. Ένα τέτοιο παράδειγμα έχουμε στον πίνακα 36 για μέγεθος δείγματος $n = 50$, όπου η μέθοδός μας με το κριτήριο AIC έδωσε (βλ. σχετικά αποτελέσματα στον πίνακα 27) ποσοστό επιτυχίας 0.06 και σφάλματα $Error1 = 0.40$, $Error2 = 0.90$ και $Error1\&2 = 0.36$. Εξαιτίας του υψηλού $Error2$ και του υψηλού ποσοστού $0.90 - 0.36 = 0.54$ της επιλογής μοντέλων μόνο με σημαντικές μεταβλητές αλλά όχι με όλες, έχουμε υποεκτίμηση των συντελεστών $\beta_2 = 2.473$ και $\beta_3 = 3.076$ (αντί του ορθού 2.5 και 4 αντίστοιχα), αν και ο συντελεστής β_5 εκτιμήθηκε σωστά ως -1.

4.4.4. Αποτελέσματα στην ημιπαραμετρική περίπτωση της Inverse Gaussian ευπάθειας: Εργαστήκαμε και για το ημιπαραμετρικό μοντέλο Inverse Gaussian ευπάθειας, πάλι με $niter = 100$ το πλήθος επαναλήψεις και παρουσιάζουμε ακολούθως τους πίνακες αποτελεσμάτων – αρχικά για τα ποσοστά επιτυχίας.

ΠΙΝΑΚΑΣ 37: ΠΟΣΟΣΤΑ ΕΠΙΤΥΧΙΑΣ – SemiParametric Case with Inverse Gaussian Frailty and Censoring Proportion ≈ 10%									
Number of Covariates	Sample Size	Our Method - AIC				Coxph - AIC			
		Perc. of Success	Error1	Error2	Error1&2	Perc. of Success	Error1	Error2	Error1&2
$k=3$	$n=50$	0.79	0.21	0	0	0.78	0.22	0	0
	$n=100$	0.83	0.17	0	0	0.82	0.18	0	0
$k=4$	$n=50$	0.64	0.35	0.01	0	0.63	0.36	0.01	0
	$n=100$	0.69	0.31	0	0	0.71	0.29	0	0
$k=5$	$n=50$	0.44	0.35	0.43	0.22	0.66	0.34	0	0
	$n=100$	0.65	0.31	0.11	0.07	0.67	0.33	0	0
Number of Covariates	Sample Size	Our Method - BIC				Coxph - BIC			
		Perc. of Success	Error1	Error2	Error1&2	Perc. of Success	Error1	Error2	Error1&2
$k=3$	$n=50$	0.95	0.05	0	0	0.94	0.06	0	0
	$n=100$	0.95	0.05	0	0	0.94	0.06	0	0
$k=4$	$n=50$	0.89	0.09	0.05	0.03	0.89	0.09	0.04	0.02
	$n=100$	0.90	0.10	0.01	0.01	0.90	0.10	0.01	0.01
$k=5$	$n=50$	0.47	0.20	0.48	0.15	0.91	0.09	0	0
	$n=100$	0.81	0.10	0.14	0.05	0.93	0.07	0	0

ΠΙΝΑΚΑΣ 38: ΠΟΣΟΣΤΑ ΕΠΙΤΥΧΙΑΣ – SemiParametric Case with Inverse Gaussian Frailty and Censoring Proportion ≈ 46%									
Number of Covariates	Sample Size	Our Method - AIC				Coxph - AIC			
		Perc. of Success	Error1	Error2	Error1&2	Perc. of Success	Error1	Error2	Error1&2
$k=3$	$n=50$	0.83	0.17	0	0	0.83	0.17	0	0
	$n=100$	0.80	0.20	0	0	0.79	0.21	0	0
$k=4$	$n=50$	0.60	0.37	0.12	0.09	0.55	0.41	0.13	0.09
	$n=100$	0.64	0.36	0.02	0.02	0.63	0.37	0.02	0.02
$k=5$	$n=50$	0.45	0.32	0.41	0.18	0.67	0.32	0.03	0.02
	$n=100$	0.54	0.36	0.23	0.13	0.7	0.3	0	0
Number of Covariates	Sample Size	Our Method - BIC				Coxph - BIC			
		Perc. of Success	Error1	Error2	Error1&2	Perc. of Success	Error1	Error2	Error1&2
$k=3$	$n=50$	0.93	0.07	0	0	0.92	0.08	0	0
	$n=100$	0.95	0.05	0	0	0.96	0.04	0	0
$k=4$	$n=50$	0.71	0.19	0.2	0.10	0.67	0.22	0.22	0.11
	$n=100$	0.82	0.15	0.04	0.01	0.82	0.15	0.04	0.01
$k=5$	$n=50$	0.45	0.12	0.53	0.10	0.86	0.09	0.06	0.01
	$n=100$	0.71	0.12	0.25	0.08	0.95	0.05	0	0

ΠΙΝΑΚΑΣ 39: ΠΟΣΟΣΤΑ ΕΠΙΤΥΧΙΑΣ – SemiParametric Case with Inverse Gaussian Frailty and Censoring Proportion ≈ 80%									
Number of Covariates	Sample Size	Our Method - AIC				Coxph - AIC			
		Perc. of Success	Error1	Error2	Error1&2	Perc. of Success	Error1	Error2	Error1&2
$k=3$	$n=50$	0.59	0.24	0.29	0.12	0.62	0.27	0.19	0.08
	$n=100$	0.85	0.15	0	0	0.85	0.15	0	0
$k=4$	$n=50$	0.19	0.43	0.71	0.33	0.26	0.42	0.61	0.29
	$n=100$	0.51	0.35	0.25	0.11	0.46	0.4	0.25	0.11
$k=5$	$n=50$	0.07	0.39	0.88	0.34	0.39	0.42	0.35	0.16
	$n=100$	0.27	0.46	0.61	0.34	0.54	0.45	0.15	0.14
Number of Covariates	Sample Size	Our Method – BIC				Coxph - BIC			
		Perc. of Success	Error1	Error2	Error1&2	Perc. of Success	Error1	Error2	Error1&2
$k=3$	$n=50$	0.5	0.16	0.49	0.15	0.59	0.15	0.38	0.12
	$n=100$	0.93	0.04	0.03	0	0.94	0.04	0.02	0
$k=4$	$n=50$	0.10	0.26	0.88	0.24	0.16	0.26	0.82	0.24
	$n=100$	0.50	0.12	0.48	0.1	0.52	0.12	0.46	0.1
$k=5$	$n=50$	0.03	0.12	0.97	0.12	0.33	0.24	0.6	0.17
	$n=100$	0.22	0.23	0.76	0.21	0.63	0.17	0.32	0.12

Για τις εκτιμήσεις των παραμέτρων:

ΠΙΝΑΚΑΣ 40: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – SemiPar. Case w. InvGaussian Frailty, 3 Covariates & Censoring Proportion ≈ 10%					
Sample Size	Parameter	Our Method – AIC mean (sd)	Our Method – BIC mean (sd)	Coxph – AIC mean (sd)	Coxph – BIC mean (sd)
$n=50$	β_1 (real value = 2)	1.877 (0.338)	1.862 (0.331)	1.886 (0.336)	1.872 (0.33)
	β_2 (real value = 0)	0.009 (0.22)	0.011 (0.148)	0.01 (0.222)	0.006 (0.156)
	β_3 (real value = –1.3)	-1.258 (0.262)	-1.248 (0.243)	-1.252 (0.259)	-1.24 (0.241)
$n=100$	β_1 (real value = 2)	1.833 (0.209)	1.828 (0.206)	1.833 (0.208)	1.828 (0.205)
	β_2 (real value = 0)	0.008 (0.129)	0.003 (0.092)	0.012 (0.126)	0.007 (0.089)
	β_3 (real value = –1.3)	-1.196 (0.184)	-1.191 (0.183)	-1.202 (0.183)	-1.196 (0.18)

ΠΙΝΑΚΑΣ 41: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – SemiPar. Case w. InvGaussian Frailty, 3 Covariates & Censoring Proportion ≈ 46%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 2)	1.955 (0.438)	1.942 (0.431)	1.989 (0.437)	1.982 (0.431)
	β_2 (real value = 0)	0.042 (0.268)	0.054 (0.204)	0.04 (0.261)	0.045 (0.212)
	β_3 (real value = –1.3)	-1.313 (0.391)	-1.311 (0.363)	-1.341 (0.398)	-1.337 (0.373)
$n = 100$	β_1 (real value = 2)	1.927 (0.311)	1.925 (0.308)	1.933 (0.308)	1.93 (0.307)
	β_2 (real value = 0)	0.006 (0.18)	0 (0.116)	0.012 (0.172)	0 (0.107)
	β_3 (real value = –1.3)	-1.237 (0.242)	-1.23 (0.22)	-1.25 (0.241)	-1.24 (0.218)

ΠΙΝΑΚΑΣ 42: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – SemiPar. Case w. InvGaussian Frailty, 3 Covariates & Censoring Proportion ≈ 80%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 2)	1.85 (0.812)	1.858 (0.848)	2.353 (1.709)	2.325 (1.685)
	β_2 (real value = 0)	-0.016 (0.682)	-0.05 (0.62)	-0.093 (0.819)	-0.133 (0.74)
	β_3 (real value = –1.3)	-1.374 (0.552)	-1.447 (0.528)	-1.595 (0.818)	-1.639 (0.829)
$n = 100$	β_1 (real value = 2)	2.015 (0.509)	1.988 (0.503)	2.16 (0.562)	2.131 (0.551)
	β_2 (real value = 0)	0.002 (0.267)	0.016 (0.174)	-0.005 (0.282)	0.016 (0.174)
	β_3 (real value = –1.3)	-1.326 (0.374)	-1.346 (0.341)	-1.403 (0.388)	-1.421 (0.358)

ΠΙΝΑΚΑΣ 43: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – SemiPar. Case w. InvGaussian Frailty, 4 Covariates & Censoring Proportion ≈ 10%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 0.8)	0.748 (0.206)	0.751 (0.189)	0.752 (0.202)	0.756 (0.19)
	β_2 (real value = 0)	0.013 (0.179)	0.021 (0.106)	0.015 (0.175)	0.019 (0.095)
	β_3 (real value = 0)	0.005 (0.156)	0.015 (0.111)	-0.001 (0.171)	0.007 (0.112)
	β_4 (real value = 1)	0.978 (0.217)	0.96 (0.213)	0.979 (0.208)	0.96 (0.206)
$n = 100$	β_1 (real value = 0.8)	0.75 (0.162)	0.751 (0.141)	0.75 (0.157)	0.749 (0.135)
	β_2 (real value = 0)	-0.011 (0.147)	-0.002 (0.109)	-0.009 (0.142)	-0.001 (0.108)
	β_3 (real value = 0)	0.015 (0.111)	-0.002 (0.07)	0.009 (0.107)	-0.003 (0.068)
	β_4 (real value = 1)	0.946 (0.149)	0.947 (0.145)	0.947 (0.146)	0.945 (0.145)

ΠΙΝΑΚΑΣ 44: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – SemiPar. Case w. InvGaussian Frailty, 4 Covariates & Censoring Proportion ≈ 46%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 0.8)	0.833 (0.276)	0.852 (0.256)	0.838 (0.282)	0.868 (0.261)
	β_2 (real value = 0)	0.047 (0.311)	0.04 (0.27)	0.051 (0.307)	0.042 (0.261)
	β_3 (real value = 0)	0.003 (0.248)	0.012 (0.175)	0.011 (0.25)	0.011 (0.186)
	β_4 (real value = 1)	1.036 (0.345)	1.039 (0.311)	1.066 (0.336)	1.069 (0.326)
$n = 100$	β_1 (real value = 0.8)	0.79 (0.225)	0.783 (0.209)	0.79 (0.219)	0.784 (0.201)
	β_2 (real value = 0)	-0.024 (0.229)	0.001 (0.178)	-0.021 (0.217)	-0.005 (0.162)
	β_3 (real value = 0)	0.022 (0.152)	0.015 (0.084)	0.021 (0.145)	0.018 (0.089)
	β_4 (real value = 1)	0.941 (0.228)	0.932 (0.216)	0.941 (0.222)	0.93 (0.21)

ΠΙΝΑΚΑΣ 45: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – SemiPar. Case w. InvGaussian Frailty, 4 Covariates & Censoring Proportion ≈ 80%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 0.8)	1.056 (0.425)	1.074 (0.427)	1.176 (0.523)	1.213 (0.515)
	β_2 (real value = 0)	0.132 (0.602)	0.125 (0.524)	0.128 (0.63)	0.095 (0.486)
	β_3 (real value = 0)	0.058 (0.718)	0.111 (0.676)	0.062 (0.711)	0.164 (0.457)
	β_4 (real value = 1)	1.324 (0.569)	1.298 (0.515)	1.513 (0.803)	1.515 (0.706)
$n = 100$	β_1 (real value = 0.8)	0.912 (0.357)	0.933 (0.33)	0.953 (0.37)	0.966 (0.339)
	β_2 (real value = 0)	0.011 (0.368)	0.034 (0.284)	0.001 (0.366)	0.026 (0.273)
	β_3 (real value = 0)	-0.041 (0.23)	0 (0)	-0.047 (0.25)	0.011 (0.109)
	β_4 (real value = 1)	1.025 (0.371)	1.043 (0.307)	1.078 (0.394)	1.08 (0.327)

ΠΙΝΑΚΑΣ 46: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – SemiPar. Case w. InvGaussian Frailty, 5 Covariates & Censoring Proportion ≈ 10%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 0)	0.004 (0.166)	-0.005 (0.147)	0.017 (0.2)	0.003 (0.133)
	β_2 (real value = 2.5)	2.093 (0.442)	2.067 (0.438)	2.519 (0.447)	2.492 (0.417)
	β_3 (real value = 4)	3.446 (0.662)	3.378 (0.678)	4.135 (0.684)	4.076 (0.673)
	β_4 (real value = 0)	-0.101 (0.277)	-0.069 (0.23)	0 (0.253)	0.006 (0.184)
	β_5 (real value = –1)	-1.023 (0.287)	-1.047 (0.27)	-1.056 (0.256)	-1.045 (0.246)
$n = 100$	β_1 (real value = 0)	0.002 (0.088)	-0.007 (0.041)	0.01 (0.112)	0.011 (0.065)
	β_2 (real value = 2.5)	2.204 (0.309)	2.183 (0.298)	2.402 (0.316)	2.384 (0.303)
	β_3 (real value = 4)	3.532 (0.403)	3.486 (0.425)	3.873 (0.434)	3.85 (0.42)
	β_4 (real value = 0)	-0.026 (0.156)	-0.022 (0.109)	-0.002 (0.135)	0 (0.086)
	β_5 (real value = –1)	-0.915 (0.173)	-0.908 (0.17)	-0.954 (0.187)	-0.95 (0.178)

ΠΙΝΑΚΑΣ 47: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – SemiPar. Case w. InvGaussian Frailty, 5 Covariates & Censoring Proportion ≈ 46%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n = 50$	β_1 (real value = 0)	0.029 (0.24)	0.032 (0.189)	-0.019 (0.224)	-0.025 (0.146)
	β_2 (real value = 2.5)	2.081 (0.512)	2.012 (0.51)	2.59 (0.671)	2.547 (0.649)
	β_3 (real value = 4)	3.306 (0.715)	3.173 (0.758)	4.189 (1.017)	4.132 (1.017)
	β_4 (real value = 0)	-0.029 (0.259)	-0.03 (0.169)	-0.006 (0.334)	-0.011 (0.276)
	β_5 (real value = –1)	-0.942 (0.27)	-0.918 (0.274)	-1.068 (0.374)	-1.063 (0.335)
$n = 100$	β_1 (real value = 0)	0.008 (0.158)	0.022 (0.11)	-0.003 (0.142)	0.01 (0.075)
	β_2 (real value = 2.5)	2.173 (0.342)	2.145 (0.33)	2.427 (0.335)	2.404 (0.308)
	β_3 (real value = 4)	3.435 (0.564)	3.392 (0.56)	3.909 (0.46)	3.886 (0.432)
	β_4 (real value = 0)	-0.028 (0.214)	-0.037 (0.173)	0.006 (0.16)	0.003 (0.086)
	β_5 (real value = –1)	-0.942 (0.185)	-0.923 (0.176)	-0.994 (0.2)	-0.984 (0.192)

ΠΙΝΑΚΑΣ 48: ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ – SemiPar. Case w. InvGaussian Frailty, 5 Covariates & Censoring Proportion ≈ 80%					
Sample Size	Parameter	**Our Method – AIC mean (sd)**	**Our Method – BIC mean (sd)**	**Coxph – AIC mean (sd)**	**Coxph – BIC mean (sd)**
$n=50$	β_1 (real value = 0)	0.013 (0.589)	0.005 (0.444)	-0.192 (1.038)	-0.099 (0.85)
	β_2 (real value = 2.5)	2.248 (1.024)	2.188 (1.031)	3.911 (2.63)	3.689 (1.927)
	β_3 (real value = 4)	3.018 (1.301)	2.937 (1.229)	6.1 (4.407)	5.537 (3.81)
	β_4 (real value = 0)	0.015 (0.573)	-0.016 (0.421)	-0.242 (1.526)	-0.256 (1.355)
	β_5 (real value = –1)	-1.044 (0.703)	-1.43 (0.573)	-1.868 (1.16)	-2.116 (0.946)
$n=100$	β_1 (real value = 0)	0.117 (0.4)	0.092 (0.326)	0.039 (0.383)	0.036 (0.231)
	β_2 (real value = 2.5)	2.156 (0.592)	2.119 (0.506)	2.659 (0.816)	2.564 (0.746)
	β_3 (real value = 4)	3.089 (0.832)	2.926 (0.828)	4.256 (1.131)	4.072 (1.153)
	β_4 (real value = 0)	-0.053 (0.432)	-0.091 (0.346)	-0.076 (0.487)	-0.108 (0.386)
	β_5 (real value = –1)	-1.007 (0.342)	-1.049 (0.348)	-1.148 (0.468)	-1.193 (0.409)

4.4.5. Σχεδίαση της εκτιμηθείσας σωρευτικής βασικής συνάρτησης κινδύνου: Προκειμένου να δούμε τη σχεδίαση της εκτιμηθείσας σωρευτικής βασικής συνάρτησης κινδύνου (*cumulative baseline hazard*) όπως αυτή υπολογίζεται από την εξίσωση **(4.14)** και να τη συγκρίνουμε με την πραγματική σωρευτική συνάρτηση κινδύνου $H_0(t)=\left(\frac{t}{\alpha}\right)^{\lambda}$, $t>0$ της Weibull, εκτελέσαμε μία ακόμα φορά τον κώδικα, με $k=4$ συμμεταβλητές, αριθμό επαναλήψεων $niter=5$, μέγεθος δείγματος $n=100$, ποσοστό λογοκρισίας ≈ 46%, πραγματικό διάνυσμα $\boldsymbol{\beta}=(0.8,0,0,1)$ και ευπάθεια που ακολουθεί την Inverse Gaussian κατανομή. Ζητήσαμε τη σχεδίαση της σωρευτικής βασικής συνάρτησης κινδύνου για όλα τα εκτιμηθέντα μοντέλα της 5ης επανάληψης ταυτόχρονα με τη σχεδίαση της εκτιμηθείσας $\hat{H}_0(t)$, προκειμένου να γίνει η σύγκριση:

```
if (iter==niter){
        H0<-function(x){exp(w_shape*log(x/w_scale_init))}
        title_string_initial<-paste("X~Z[,", formula_constr[1],"]")
        if (p==1){title_string<-paste(title_string_initial)}
```

```
        if (p==2){title_string<-paste(title_string_initial,"+Z[,",formula_constr[2],"]")}
        if (p>2)
        {       title_string<-title_string_initial
                for (u in 2:p) {title_string<-paste(title_string,"+Z[,",formula_constr[u],"]")}
        }

        hazard_est<-numeric(failure_num)
        for (i in 1:failure_num) {hazard_est[i]<-hz[i]}
        cum_hazard_est<-numeric(failure_num)
        for (i in 1:failure_num) {cum_hazard_est[i]<-
H(time_vector[i],time_vector,hazard_est)}
        coxph_hazard_est<-numeric(failure_num)
        for (i in 1:failure_num) {coxph_hazard_est[i]<-cum_hazard_vector[i]}
        hazard<-cbind(time_vector, hazard_est, cum_hazard_est,H0(cum_hazard_est))
        png(filename=paste(title_string,".png"), width=800, height=800, pointsize=24)
        plot(x=time_vector, y=cum_hazard_est, type="s", xlim=c(0,
time_vector[failure_num]), xlab="failure times", ylab="H_hat & H0")
        par(new=TRUE)
        curve(H0, type="l", lty=3, lwd=2, xlim=c(0, time_vector[failure_num]), xaxt='n',
yaxt='n', xlab="", ylab="")
        title("Estimated cumulative baseline hazard function")
        legend("bottomright",c(title_string), col="blue")
        dev.off()
} #end if
```

Οι πίνακες αποτελεσμάτων *our_AIC_iter_results*, *our_BIC_iter_results*, *coxph_AIC_iter_results* και *coxph_BIC_iter_results*, στους οποίους φαίνεται ποια μοντέλα επελέγησαν στην 5η επανάληψη, είναι οι εξής:

our_AIC_iter_results	iter1	iter 2	iter 3	iter 4	iter 5
est. coef. β(1) (Our Method)	0.7014178	0.83764328	1.1402584	0.80556073	0.897
est. coef. β(2) (Our Method)	0	0	-0.40461745	0	0
est. coef. β(3) (Our Method)	0.250519	0	0	0	0
est. coef. β(4) (Our Method)	1.0709534	0.92434559	0.97413253	0.77477442	0.818
est. AIC (Our Method)	430.24027	431.6339233	469.0613094	462.3444177	473.9
Error 1 (Boolean 0-1) (Our Method)	1	0	1	0	0
Error 2 (Boolean 0-1) (Our Method)	0	0	0	0	0
Error 1 & 2 (Boolean 0-1) (Our Method)	0	0	0	0	0
Per Cent Censor	0.47	0.47	0.43	0.45	0.43

our_BIC_iter_results	iter1	iter 2	iter 3	iter 4	iter 5
est. coef. β(1) (Our Method)	0.7149788	0.83764328	1.1402584	0.80556073	0.897
est. coef. β(2) (Our Method)	0	0	-0.40461745	0	0
est. coef. β(3) (Our Method)	0	0	0	0	0
est. coef. β(4) (Our Method)	1.1886216	0.92434559	0.97413253	0.77477442	0.818
est. BIC (Our Method)	435.74809	436.8442637	476.8768199	467.5547581	479.1
Error 1 (Boolean 0-1) (Our Method)	0	0	1	0	0
Error 2 (Boolean 0-1) (Our Method)	0	0	0	0	0
Error 1 & 2 (Boolean 0-1) (Our Method)	0	0	0	0	0
Per Cent Censor	0.47	0.47	0.43	0.45	0.43

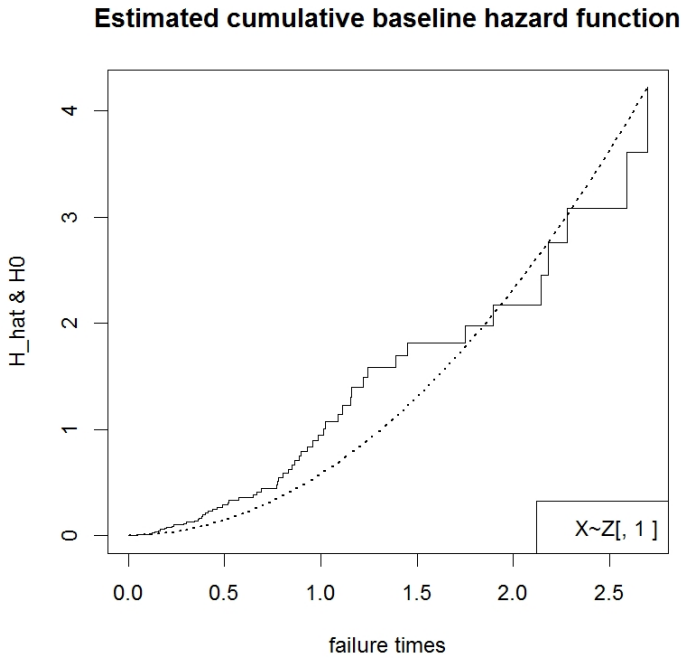
coxph_AIC_iter_results	iter1	iter 2	iter 3	iter 4	iter 5
est. coef. β(1) (Our Method)	0.7060383	0.853986276	1.149466096	0.82476073	0.852
est. coef. β(2) (Our Method)	0	0	-0.388438119	0	0
est. coef. β(3) (Our Method)	0.2686892	0	0	0	0
est. coef. β(4) (Our Method)	1.0892795	0.915563201	0.982974423	0.767287904	0.826
est. BIC (Our Method)	319.87937	321.5429428	350.1419142	349.2036399	357.7
Error 1 (Boolean 0-1) (Our Method)	1	0	1	0	0
Error 2 (Boolean 0-1) (Our Method)	0	0	0	0	0
Error 1 & 2 (Boolean 0-1) (Our Method)	0	0	0	0	0
Per Cent Censor	0.47	0.47	0.43	0.45	0.43

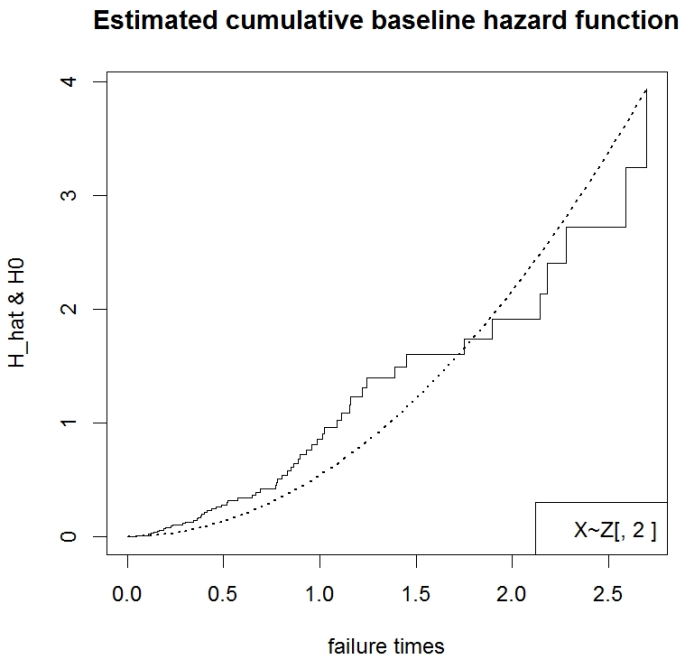
coxph_BIC_iter_results	iter1	iter 2	iter 3	iter 4	iter 5
est. coef. β(1) (Our Method)	0.7147049	0.853986276	1.149466096	0.82476073	0.852
est. coef. β(2) (Our Method)	0	0	-0.388438119	0	0
est. coef. β(3) (Our Method)	0	0	0	0	0
est. coef. β(4) (Our Method)	1.2222623	0.915563201	0.982974423	0.767287904	0.826
est. BIC (Our Method)	325.56524	326.7532832	357.9574248	354.4139803	362.9
Error 1 (Boolean 0-1) (Our Method)	0	0	1	0	0
Error 2 (Boolean 0-1) (Our Method)	0	0	0	0	0
Error 1 & 2 (Boolean 0-1) (Our Method)	0	0	0	0	0
Per Cent Censor	0.47	0.47	0.43	0.45	0.43

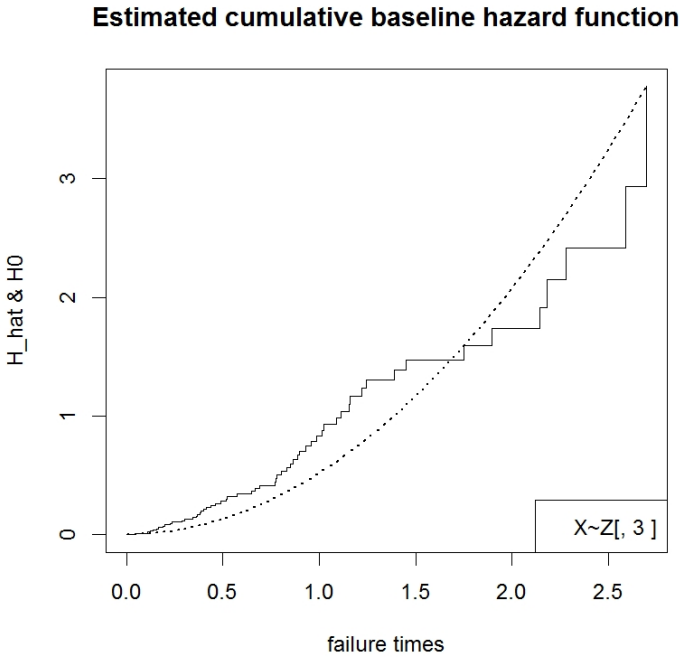
Ο τελικός πίνακας *final_matrix* με τα συγκεντρωτικά στοιχεία:

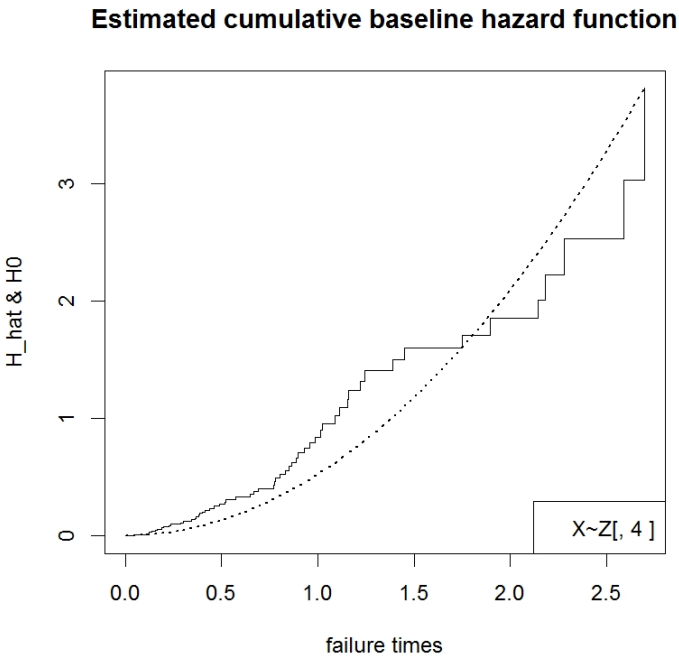
	Average Censor (Percentage)	**Num. of Success**	**Percentage of Success**	**Error 1**	**Error 2**	**Error 1&2**	**Mean of β1**	**St. Dev. of β1**	**Mean of β2**	**St. Dev. of β2**	**Mean of β3**	**St. Dev. of β3**	**Mean of β4**	**St. Dev. of β4**
Our Method – AIC	0.45	3	0.6	0.4	0	0	0.876	0.164	-0.081	0.181	0.050	0.112	0.912	0.119
Our Method – BIC	0.45	4	0.8	0.2	0	0	0.879	0.160	-0.081	0.181	0	0	0.936	0.162
Coxph Method – AIC	0.45	3	0.6	0.4	0	0	0.877	0.164	-0.078	0.174	0.054	0.120	0.916	0.127
Coxph Method – BIC	0.45	4	0.8	0.2	0	0	0.879	0.162	-0.078	0.174	0	0	0.943	0.177
Real values of coef.	***NA***	***NA***	***NA***	***NA***	***NA***	***NA***	***0,8***	***0***	***0***	***0***	***0***	***0***	***1***	***0***

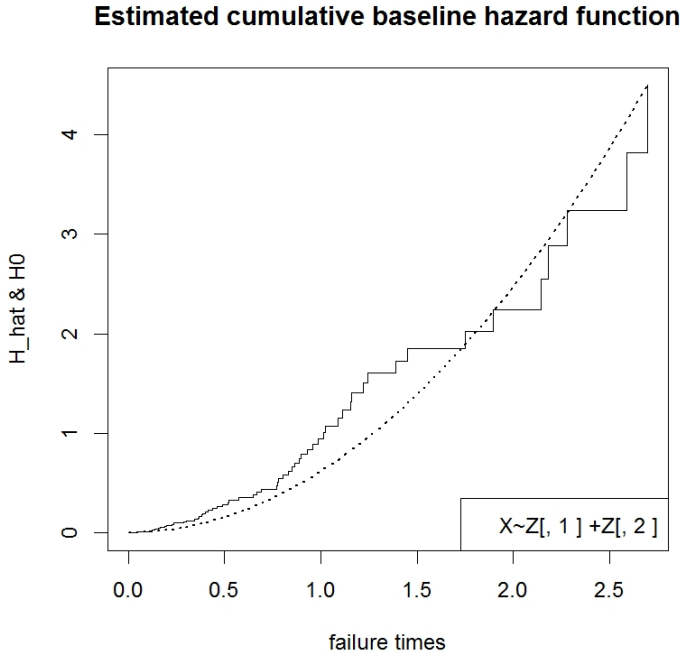
και η σχεδίαση όλων των σωρευτικών βασικών συναρτήσεων κινδύνου που προέκυψαν από τα 15 μοντέλα της 5ης επανάληψης μαζί με την καμπύλη της σωρευτικής συνάρτησης κινδύνου της κατανομής Weibull (η οποία στα ακόλουθα σχήματα εμφανίζεται με διακεκομμένη γραμμή):

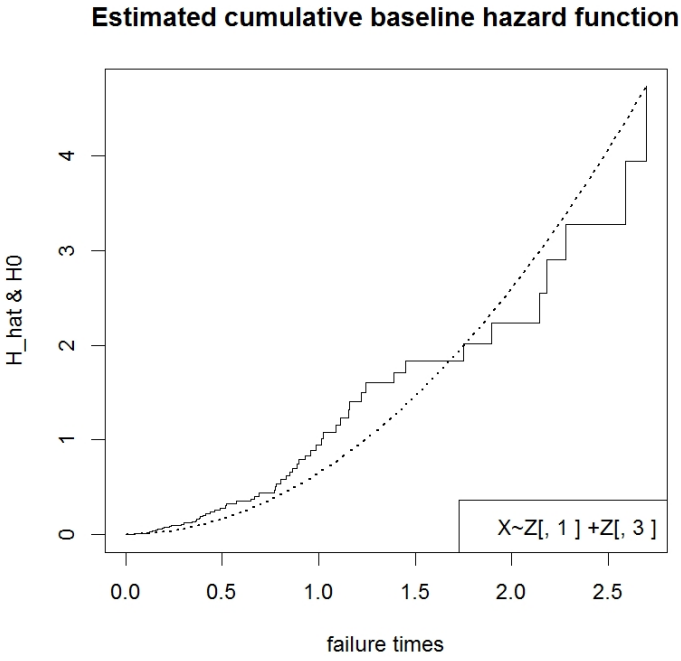


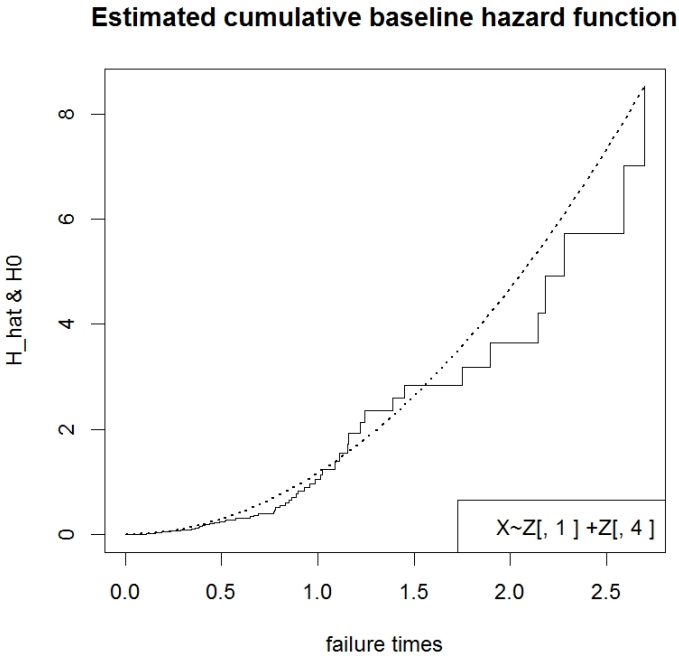


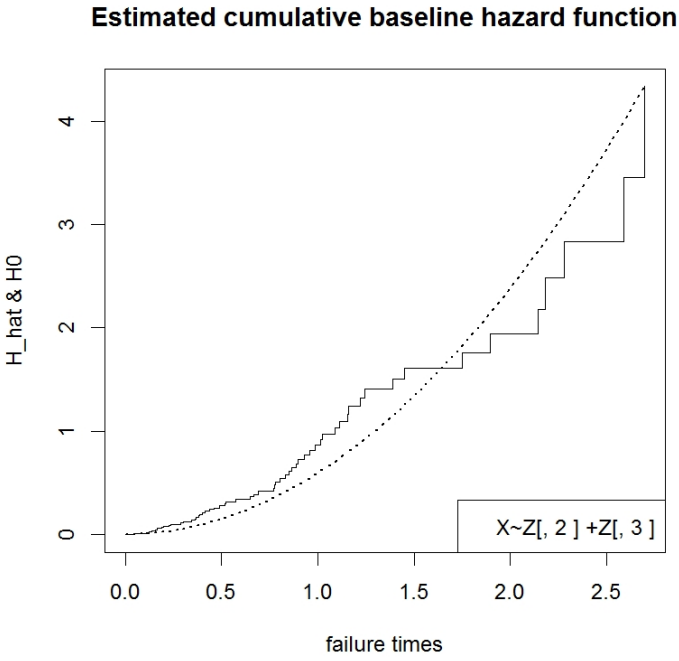

Estimated cumulative baseline hazard function
H_hat & H0
failure times
X~Z[, 3]
0.0
0.5
1.0
1.5
2.0
2.5
0
1
2
3

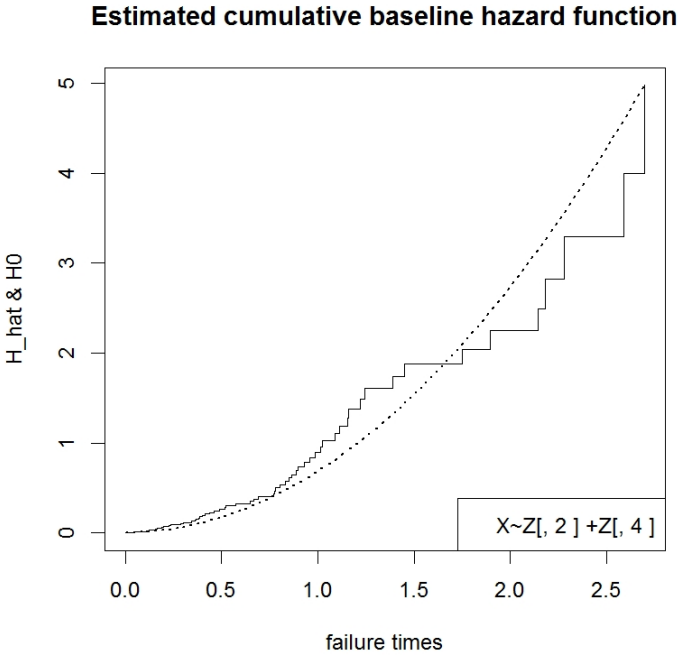

Estimated cumulative baseline hazard function
H_hat & H0
failure times
X~Z[, 4]
0.0
0.5
1.0
1.5
2.0
2.5
0
1
2
3

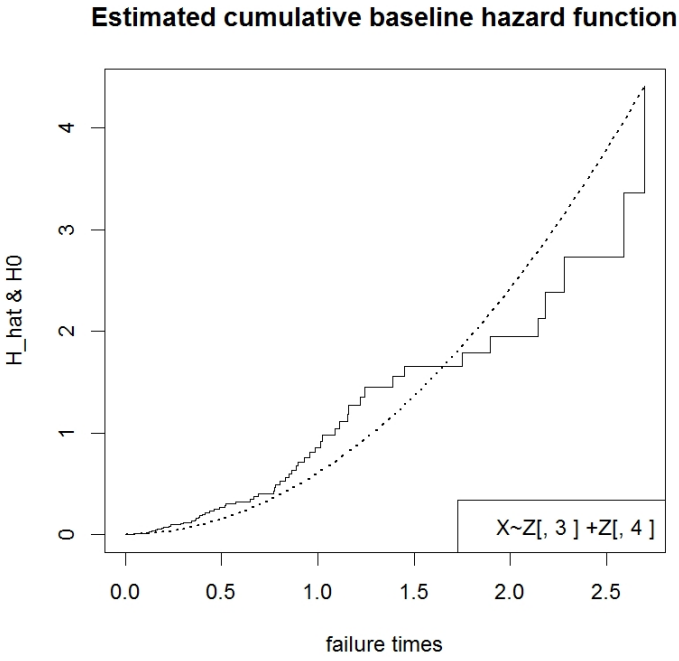

Estimated cumulative baseline hazard function
H_hat & H0
failure times
X~Z[, 1] +Z[, 2]
0.0
0.5
1.0
1.5
2.0
2.5
0
1
2
3
4

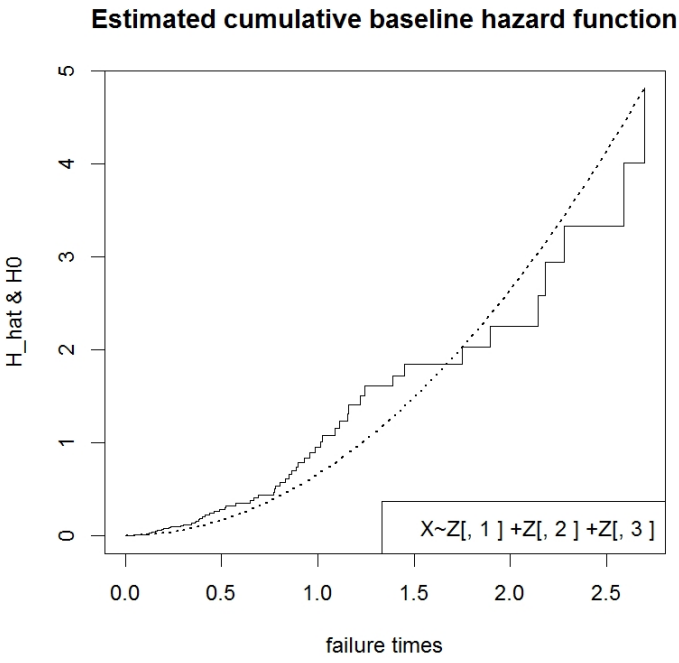

Estimated cumulative baseline hazard function
H_hat & H0
failure times
X~Z[, 1] +Z[, 3]
0.0
0.5
1.0
1.5
2.0
2.5
0
1
2
3
4

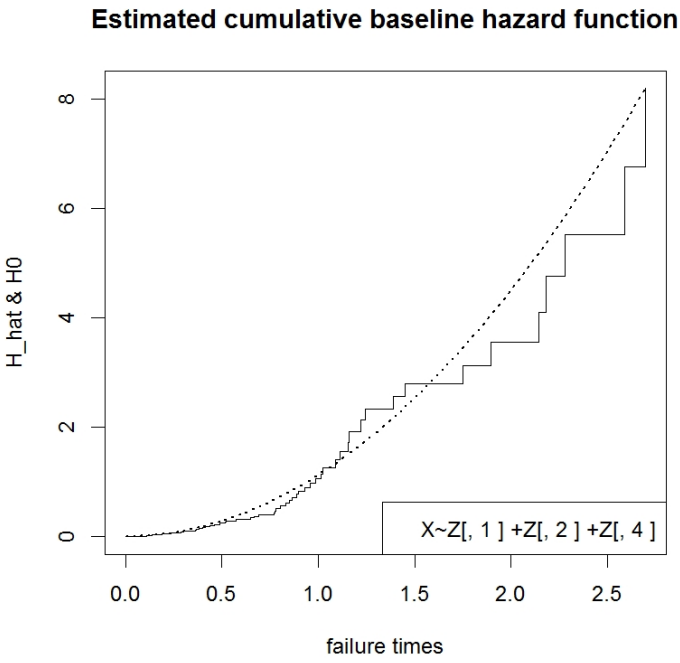

Estimated cumulative baseline hazard function
H_hat & H0
failure times
X~Z[, 1] +Z[, 4]

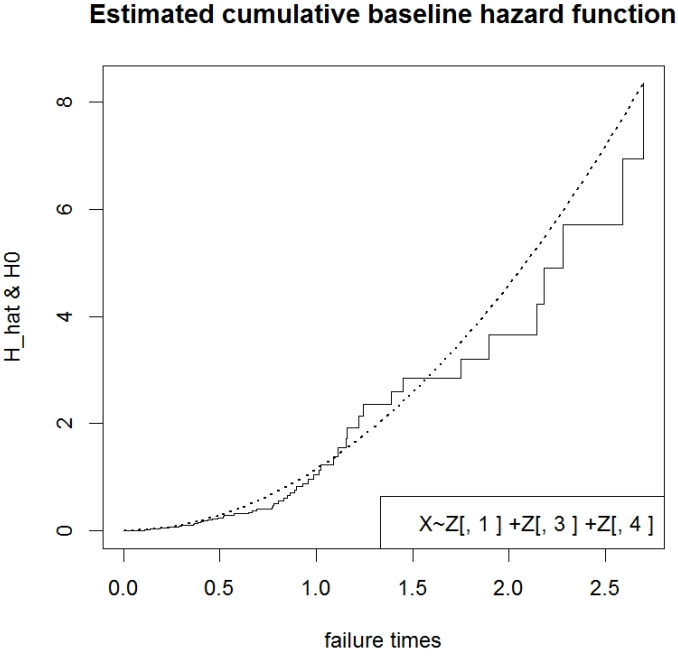

Estimated cumulative baseline hazard function
H_hat & H0
failure times
X~Z[, 2] +Z[, 3]

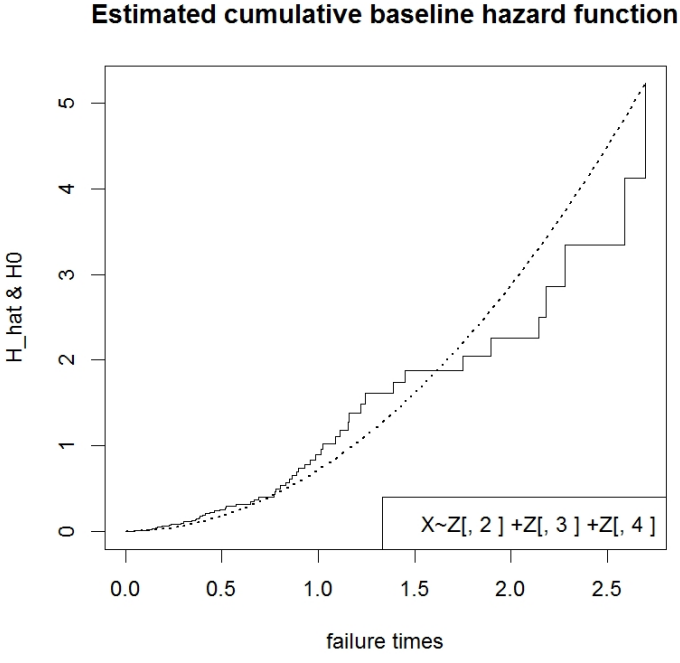

Estimated cumulative baseline hazard function
H_hat & H0
failure times
X~Z[, 2] +Z[, 4]

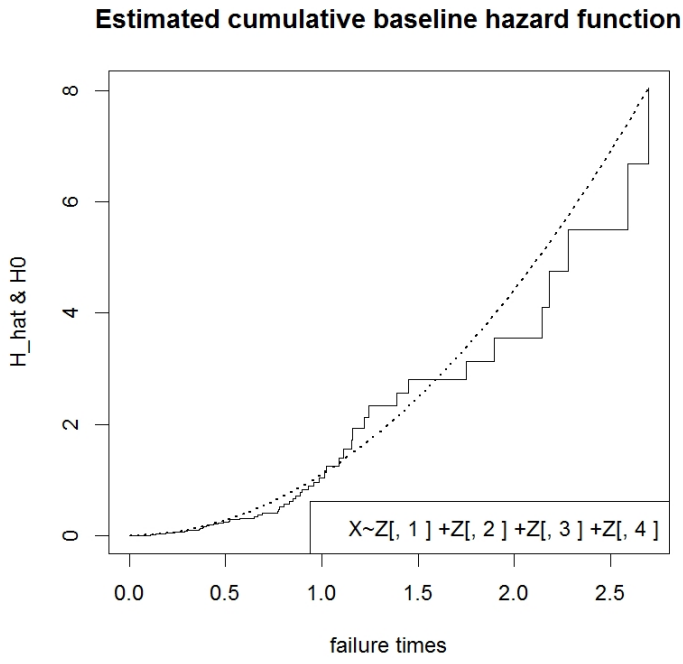

Estimated cumulative baseline hazard function
H_hat & H0
failure times
X~Z[, 3] +Z[, 4]


Estimated cumulative baseline hazard function
H_hat & H0
failure times
X~Z[, 1] +Z[, 2] +Z[, 3]


Estimated cumulative baseline hazard function
H_hat & H0
failure times
X~Z[, 1] +Z[, 2] +Z[, 4]


Estimated cumulative baseline hazard function
H_hat & H0
failure times
X~Z[, 1] +Z[, 3] +Z[, 4]


Estimated cumulative baseline hazard function
H_hat & H0
failure times
X~Z[, 2] +Z[, 3] +Z[, 4]



4.4.6. Περαιτέρω συζήτηση: Κλείνουμε την παρούσα εργασία με λίγα σχόλια για τη μέθοδό μας πάνω στο ημιπαραμετρικό μοντέλο:

- Η δυσκολία στη μεθόδό μας έγκειται βασικά στο ότι επιχειρούμε να βελτιστοποιήσουμε μία συνάρτηση με μεγάλο πλήθος μεταβλητών, αφού η προς εκτίμηση συνάρτηση κινδύνου είναι οχληρά παράμετρος άπειρης διάστασης.
- Η προσέγγιση που δόθηκε εδώ ήταν η βελτιστοποίηση να γίνει σε δύο βήματα, ώστε αφενός στο πρώτο βήμα να απαλλαγούμε από την οχληρά παράμετρο, αφετέρου στο δεύτερο να βελτιστοποιήσουμε μόνο ως προς τις βασικές μεταβλητές που είναι οι συντελεστές $\boldsymbol{\beta}$.
- Θεωρήσαμε αναγκαία την παραπάνω διάσπαση του προβλήματος στα δύο βήματα, επειδή η συνάρτηση *constrOptim* της R έδειξε να μην ανταποκρίνεται καλά στην απευθείας βελτιστοποίηση. Μάλιστα, επειδή η *constrOptim* εκτελεί βελτιστοποίηση υπό περιορισμούς (*Constrained Optimization*), χρησιμοποιήθηκε μόνο για την εκτίμηση της συνάρτησης κινδύνου η οποία παίρνει μη αρνητικές τιμές.

- Προτείνουμε να διερευνηθεί το πώς μπορεί να επιτυχθεί καλή βελτιστοποίηση σε ένα βήμα, προκειμένου να μειωθεί ο χρόνος επεξεργασίας και αριθμητικών υπολογισμών ο οποίος είναι υψηλός εξαιτίας των επαναλαμβανόμενων υπολογισμών της *log_likelihood*.
- Σημαντικό σημείο της βελτιστοποίησης, είναι η καλή επιλογή των αρχικών διανυσμάτων $\mathbf{h}$ και $\boldsymbol{\beta}$ για τα οποία χρησιμοποιήσαμε την εκτίμησή τους από την *coxph*. Εισηγούμαστε τη μελέτη τρόπου εκτίμησης των αρχικών τιμών της συνάρτησης κινδύνου χωρίς χρήση της *coxph* (π.χ. με εφαρμογή της εκτιμήτριας Nelson-Aalen στα προσομοιωμένα δεδομένα)
- Επαναλαμβάνουμε και θεωρούμε σημαντικό το υπολογιστικό πλεονέκτημα της μεθόδου μας, που είναι ότι με μία μικρή αλλαγή στον κώδικα ως προς την συνάρτηση $G(\cdot)$ αντιμετωπίζονται ταυτόχρονα ολα τα μοντέλα μετασχηματισμού.
- Τέλος, για τη μαθηματική θεμελίωση (ύπαρξη και ιδιότητες) της εκτιμήτριας μεγίστης πιθανοφάνειας του ημιπαραμετρικού μοντέλου μετασχηματισμού καθώς και την παρουσίαση αριθμητικού αλγορίθμου εκτίμησης της συνάρτησης κινδύνου, παραπέμπουμε στο [60] που αποτελεί μία ολοκληρωμένη μελέτη πάνω στο θέμα.

ΠΗΓΕΣ – ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Joseph Ibrahim: Survival Analysis: Introduction (American Statistical Association - Northeastern Illinois Chapter - Summer 2005 Workshop – http://www.amstat.org/chapters/northeasternillinois/pastevents/presentations/summer05_Ibrahim_I.pdf)

[2] Andreas Wienke: Frailty Models in Survival Analysis (Chapman & Hall/CRC Biostatistics Series)

[3] Χρύσα Καρώνη: Μοντέλα αξιοπιστίας και επιβίωσης (Εκδόσεις Συμεών - Αθήνα 2009)

[4] Alex Cook: Censoring and Truncation (National University of Singapore – http://courses.nus.edu.sg/course/stacar/internet/st3242/handouts/ch1l1.pdf)

[5] Ευγενία Σοφία Θ. Ζαρογιάννη: Το μοντέλο αναλογικών κινδύνων του Cox και εφαρμογή στην R (Προπτυχιακή Διπλωματική Εργασία - Ε.Μ.Π./Σ.Ε.Μ.Φ.Ε. – http://dspace.lib.ntua.gr/bitstream/123456789/8009/3/zarogiannye_Cox.pdf)

[6] German Rodriguez: Non-Parametric Estimation in Survival Models (Princeton University - Spring, 2001; revised Spring 2005, Summer 2010 - http://data.princeton.edu/pop509/ParametricSurvival.pdf)

[7] Wikipedia: Wallodi Weibull (http://en.wikipedia.org/wiki/Waloddi_Weibull)

[8] Wikipedia: Weibull distribution (http://en.wikipedia.org/wiki/Weibull_distribution)

[9] John Norstad: The Normal and Lognormal Distributions (John Nordstad, February 2, 1999 - Updated: November 3, 2011 - http://www.norstad.org/finance/normdist.pdf)

[10] Josef Brüederl & Andreas Diekmann: The Log-Logistic Rate Model (Sociological Methods & Research, Vol. 24 No. 2, November 1995 p. 158-186)

[11] V.V. Rykov et al. (editors): Mathematical and Statistical Models and Methods in Reliability - Chapter 33: Lemeshko et al.: Inverse Gaussian Model and Its Applications (Springer Science+Business Media, LLC 2010)

[12] German Rodriguez: Non-Parametric Estimation in Survival Models (Princeton University - Spring, 2001; revised Spring 2005 - http://data.princeton.edu/pop509/NonParametricSurvival.pdf)

[13] Edward L. Kaplan και Paul Meier: Non-Parametric Estimation from Incomplete Observations (Journal of the American Statistical Association - Vol. 53, No. 282, June, 1958, p. 457-481)

[14] Alex Cook: The Kaplan-Meier estimate of the survival function (presentation) (National University of Singapore - http://courses.nus.edu.sg/course/stacar/internet/st3242/handouts/ch2l1.pdf)

[15] Alex Cook: The Kaplan-Meier estimate of the survival function (notes) (National University of Singapore - http://courses.nus.edu.sg/course/stacar/internet/st3242/handouts/notes2.pdf)

[16] Stanley Sawyer: The Greenwood and Exponential Greenwood Confidence Intervals in Survival Analysis (Washington University - http://www.math.wustl.edu/~sawyer/handouts/greenwood.pdf)

[17] Ana M. Pérez-Marín: Empirical comparison between the Nelson-Aalen Estimator and the Naive Local Constant Estimator (Institut d'Estadística de Catalunya - http://www.idescat.cat/sort/sort321/32.1.4.perez.pdf)

[18] K. F. Turkman: Linear Regression (http://docentes.deio.fc.ul.pt/kfturkman/regression.pdf)

[19] Π. Οικονόμου - Χ. Καρώνη: Στατιστικά Μοντέλα Παλινδρόμησης (Εκδόσεις Συμεών - Αθήνα 2010)

[20] Simon Jackman: Generalized Linear Models (Stanford University - http://jackman.stanford.edu/papers/glm.pdf)

[21] Alex Cook: The Cox Proportional Hazards Model (National University of Singapore - http://courses.nus.edu.sg/course/stacar/internet/st3242/handouts/notes3.pdf)

[22] L. J. Wei: The Accelerated Failure Time Model: A useful alternative to the Cox Regression Model in Survival Analysis (Journal: Statistics in Medicine, Vol. 11 pp. 1871-1879 - Pubished by John Wiley & Sons, 1992)

[23] Jiezhi Qi: Comparison of Proportional Hazards and Accelerated Failure Time Models (Thesis Submitted for the Degree of the Master of Science) (University of Saskatchewan, 2009 - http://ecommons.usask.ca/bitstream/handle/10388/etd-03302009-140638/JiezhiQiThesis.pdf)

[24] Sir David Cox: Regression Models and Life Tables (Journal of the Royal Statistical Society, Series B (Methodological), Vol. 34, No. 2 (1972), pp. 187-220)

[25] Jianqing Fan & Jiancheng Jiang: Non- and Semi- Parametrical Modelling in Survival Analysis (http://www.researchgate.net/publication/228574743_Non-and_Semi-Parametric_Modeling_in_Survival_Analysis)

[26] Thomas H. Scheike & Yanqing Sun: Maximum likelihood estimation for tied survival data under Cox regression model via EM-algorithm (Lifetime Data Anal (2007) 13, pp. 399-420)

[27] Terry Therneau & Patricia Grambsch: Modeling Survival Data: Extending the Cox Model (Springer Verlag (2000))

[28] Brenda Gillespie: Checking Assumptions in the Cox Proportional Hazards Regression Model (Presented at the 2006 Midwest SAS Users Group (MWSUG), Dearborn, Michigan, October 22-24, 2006) (Center for Statistical Consultation and Research, University of Michigan (2006) - http://www.mwsug.org/proceedings/2006/stats/MWSUG-2006-SD08.pdf)

[29] Inger Persson: Essays on the Assumption of Proportional hazards in Cox Regression (Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences 110 - Acta Universitatis Upsaliensis, Uppsala (2002) – http://uu.divaportal.org/smash/get/diva2:161225/FULLTEXT01.pdf)

[30] Kristin Sainani: Cox Regression II (Powerpoint presentation) (Stanford University - Department of Health Research and Policy - http://www.pitt.edu/~super4/33011-34001/33111.ppt)

[31] David Schoenfeld: Residuals for the proportional hazards regression model (Biometrika, 1982, 69(1):239-241)

[32] Anastasios Tsiatis & Daowen Zhang: Analysis of Survival Data (Lecture Notes) (North Carolina State University - Department of Statistics (2005) - http://www4.stat.ncsu.edu/~dzhang2/st745/chap1.pdf)

[33] Anastasios Tsiatis & Daowen Zhang: Analysis of Survival Data (Lecture Notes) - Chapter 3: Likelihood and Censored (or Truncated) Survival Data (North Carolina State University - Department of Statistics (2005) – http://www4.stat.ncsu.edu/~dzhang2/st745/chap3.pdf)

[34] Yakov Ben-Haim: Lecture Notes on Censoring and Estimation in Statistical Sampling (Israel Institute of Technology - http://www.technion.ac.il/yakov/intrel/censor01.pdf)

[35] Patricia Grambsch & Terry Therneau: Proportional Hazards Test and Diagnostics Based on Weighted Residuals (Biometrika, 1994, 81(3):515-526)

[36] Wikipedia: Rao's Score Test (http://en.wikipedia.org/wiki/Score_test)

[37] Wikipedia: F-Test (http://en.wikipedia.org/wiki/F-test)

[38] Matt Blackwell: Multiple Hypothesis Testing - The F-test (http://www.mattblackwell.org/files/teaching/ftests.pdf)

[39] Citation Classic Commentaries 1981: Akaike H. A new look at the statistical model identification (Eugene Garfield's Citacion Classic Commentaries electronic library - http://www.garfield.library.upenn.edu/classics1981/A1981MS54100001.pdf)

[40] Hirotugu Akaike: A new look at the statistical model identification (IEEE Trans. Automat. Contr. AC-19:716-723, 1974)

[41] Joseph E. Cavanaugh: Model Selection - The BIC Criterion (Lecture Notes) (The University of Iowa: Department of Statistics and Actuarial Science - Department of Biostatistics – http://myweb.uiowa.edu/cavaaugh/ms_lec_5_ho.pdf)

[42] Gideon Schwarz: Estimating the Dimension of a Model (The Annals of Statistics, Vol. 6, No. 2. (Mar., 1978), pp. 461-464)

[43] Colin Lingwood Mallows: Some comments on Cp (Technometrics, Vol. 15, No. 4. (Nov., 1973), pp. 661-675)

[44] Lisa Borsi, Marc Lickes & Lovro Soldo: The startified Cox Procedure (Presentation) (Swiss Federal Institute of Technology, Zurich – http://stat.ethz.ch/education/semesters/ss2011/seminar/contents/presentation_5.pdf)

[45] Benjamin Hall: Introduction to frailty models (STA635 Project) (http://www.ms.uky.edu/~mai/sta635/bhall_intro_to_frailty_models.ppt)

[46] James W. Vaupel; Kenneth G. Manton; Eric Stallard The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality (Demography 16 (3): 439–454)

[47] C. T. J. Dodson: Introduction to Laplace Transforms for Engineers (School of Mathematics - Manchester University - http://isites.harvard.edu/fs/docs/icb.topic924979.files/laplace.pdf)

[48] German Rodriguez: Unobserved Heterogeneity (Princeton University - Spring, 2001; revised Spring 2005 - http://data.princeton.edu/pop509/UnobservedHeterogeneity.pdf)

[49] Emmanouil Androulakis, Christos Koukouvinos and Filia Vonta: On the Uniform Frailty Model with Penalized Likelihood and Clustered Data (Journal of Reliability and Statistical Studies; ISSN (Print): 0974-8024, (Online):2229-5666, Vol. 5, Issue Special (2012): 97-106)

[50] Donglin Zeng & D. Y. Lin: Semiparametric Transformation Models With Random Effects for Recurrent Events (Journal of the American Statistical Association, March 2007, Vol. 102, No. 477, Theory and Methods, DOI 10.1198/016214506000001239)

[51] Leonard Kiti Alii: General Box-Cox Transformation Model for Recurrent Events Data (J Biomet Biostat, Volume 2 - Issue 1 - 1000110, ISSN:2155-6180 (JBMBS, an open access journal))

[52] Dorota M. Dabrowska: Estimation in a class of semiparametric transformation models (IMS Lecture Notes–Monograph Series, 2nd Lehmann Symposium – Optimality Vol. 49 (2006) 131–169)

[53] Donglin Zeng, D. Y. Lin & Xihong Lin: Semiparametric Transformation Models with Random Effects for Clustered Failure Time Data (Stat Sin. 2008 January 1; 18(1): 355–377)

[54] Filia Vonta: Efficient estimation in a non-proportional hazards model in survival analysis (Scand. Journal of Statistics, Vol. 23, 1996, pp. 49–61)

[55] F. Vonta & C. Huber: On the Estimation of Structural Parameters in Frailty Models for Interval Censored and Truncated Data (Computer Modelling and New Technologies, 2010, Vol.14, No.4, 40–49 (Transport and Telecommunication Institute, Lomonosova 1, LV-1019, Riga, Latvia))

[56] Wikipedia : Nuisance Parameter
(*en.wikipedia.org* - http://en.wikipedia.org/wiki/Nuisance_parameter)

[57] Γεώργιος Ξυνός: Κριτήρια επιλογής μοντέλων για την κλάση μοντέλων ευπάθειας (Διπλωματική εργασία για το Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών "Εφαρμοσμένες Μαθηματικές Επιστήμες" της Σχολής Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών του Ε. Μ. Π. - http://dspace.lib.ntua.gr/bitstream/123456789/6384/3/xynosg_model.pdf)

[58] R manual: Random effects terms (frailty function) (http://stat.ethz.ch/R-manual/R-devel/library/survival/html/frailty.html)

[59] J. Fan & R. Li: Variable Selection for Cox's Proportional Hazards (The Annals of Statistics 2002, Vol. 30, No. 1, 74–99)

[60] E. V. Slud & F. Vonta: Consistency of the NPML Estimator in the Right-Censored Transformation Model (Board of the Foundation of the Scandinavian Journal of Statistics 2004. Published by Blackwell Publishing Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA Vol 31: 21–41, 2004)