ΜΠΣ ΒΙΟΣΤΑΤΙΣΤΙΚΗ

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΙΑΤΡΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

Διπλωματική Εργασία

Μπεϋζιανή Συμπερασματολογία Για Μοντέλα Ανάλυσης Επιβίωσης

ΧΡΗΣΤΟΣ ΘΩΜΑΔΑΚΗΣ

Αθήνα, 2014

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στη

ΒΙΟΣΤΑΤΙΣΤΙΚΗ

που απονέμει η Ιατρική Σχολή και το Τμήμα Μαθηματικών του Εθνικού & Καποδιστριακού Πανεπιστημίου Αθηνών και το Τμήμα Μαθηματικών του Πανεπιστημίου Ιωαννίνων.

Εγκρίθηκε την _____ από την εξεταστική επιτροπή:

ΟΝ/ΜΟ	**ΒΑΘΜΙΔΑ**	**ΥΠΟΓΡΑΦΗ**
Λ. Μελιγκοτσίδου (Επιβλέπουσα)	Επ. Καθηγήτρια	_____
Π. Τουλούμη	Αν. Καθηγήτρια	_____
Φ. Σιάννης	Επ. Καθηγητής	_____

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω την κα Μελιγκοτσίδου για την επίβλεψη της διπλωματικής εργασίας και όλους τους διδάσκοντες του μεταπτυχιακού για τις πολύτιμες γνώσεις που μου προσέφεραν.

Περιεχόμενα

Κατάλογος πινάκων

Κατάλογος σχημάτων

Κεφάλαιο 1

Εισαγωγή

Η στατιστική ανάλυση δεδομένων επιβίωσης (time to event data) έχει ευρεία εφαρμογή σε αρκετές επιστημονικές περιοχές συμπεριλαμβανομένης της ιατρικής, της επιδημιολογίας, της βιολογίας και άλλων πεδίων. Η πιο κοινή υπόθεση η οποία γίνεται στην ανάλυση δεδομένων επιβίωσης είναι ότι οι χρόνοι επιβίωσης, δοθέντων κάποιων επεξηγηματικών μεταβλητών (covariates), είναι ανεξάρτητοι. Συχνά, δεν μπορούμε να υποθέσουμε ότι οι χρόνοι μέχρι την εμφάνιση του γεγονότος (event) είναι ανεξάρτητοι σε κάποιες ομάδες του πληθυσμού, αφού τα άτομα της ίδιας ομάδας μπορεί να μοιράζονται κοινά χαρακτηριστικά τα οποία είναι μη παρατηρήσιμα σε εμάς. Τα κοινά αυτά χαρακτηριστικά ονομάζονται frailties στην ανάλυση επιβίωσης. Υποθέτουμε ότι οι χρόνοι επιβίωσης ατόμων της ίδιας ομάδας είναι ανεξάρτητοι, δεδομένης μιας μη παρατηρήσιμης μεταβλητής (frailty). Επομένως, οι χρόνοι επιβίωσης των ατόμων της ίδιας ομάδας είναι συσχετισμένοι, εξαιτίας της frailty μεταβλητής.

Στην παρούσα διπλωματική εργασία εξετάζονται παραμετρικά μοντέλα ανάλυσης επιβίωσης, βασισμένα στην Weibull και στην κατά τμήματα εκθετική κατανομή, υπό το πρίσμα της Μπεϋζιανής στατιστικής, κατάλληλα για δεδομένα με δεξιά λογοκρισία (right censoring). Ως κατανομές για τα frailties χρησιμοποιούνται οι Gamma και Lognormal κατανομές. Η στατιστική συμπερασματολογία εξάγεται μέσω μεθόδων Markov Chain Monte Carlo (MCMC). Οι MCMC μέθοδοι περιλαμβάνουν τον αλγόριθμο Metropolis–Hastings ο οποίος βασίζεται στην προτείνουσα σταθμισμένων ελαχίστων τετραγώνων (Gamerman 1997) και το δειγματολήπτη Gibbs (adaptive rejection sampling των Gilks & Wild 1992).

Η εφαρμογή των μεθόδων αξιολογείται μέσω προσομοιωμένων δεδομένων. Τα μοντέλα συγκρίνονται με το Deviance Information Criterion (DIC). Στην Gamma κατανομή, η περιθώρια πιθανοφάνεια (ως προς τα frailties) μπορεί να βρεθεί αναλυτικά. Στην Lognormal κατανομή, επειδή το ολοκλήρωμα δεν είναι διαχειρίσιμο, χρησιμοποιούμε adaptive Gauss-Hermite quadrature μέθοδο για τον υπολογισμό του. Επιπλέον, οι MCMC αλγόριθμοι εφαρμόζονται σε δεδομένα νεφροπαθών ασθενών των McGilchrist & Aisbett (1991).

Η παρούσα διπλωματική εργασία χωρίζεται σε 7 κεφάλαια. Στο κεφάλαιο 2 παρουσιάζονται οι βασικές έννοιες της ανάλυσης επιβίωσης και γίνεται εισαγωγή σε μοντέλα frailty. Στο κεφάλαιο 3 παρουσιάζονται οι βασικές αρχές της Μπεϋζιανής στατιστικής και η γενική μορφή των αλγορίθμων MCMC. Επίσης, περιγράφεται η μέθοδος προσομοίωσης adaptive rejection sampling. Στο κεφάλαιο 4 αναπτύσσονται λεπτομερώς οι διαθέσιμοι αλγόριθμοι MCMC για shared frailty μοντέλα. Στο κεφάλαιο 5 γίνεται εφαρμογή των αλγορίθμων σε προσομοιωμένα δεδομένα. Οι εκτιμήσεις μέγιστης πιθανοφάνειας παρουσιάζονται για σύγκριση. Στο κεφάλαιο 6 γίνεται εφαρμογή των μεθόδων σε πραγματικά δεδομένα νεφροπαθών ασθενών. Στο κεφάλαιο 7 γίνεται συζήτηση σχετικά με τα αποτελέσματα των προσομοιωμένων δεδομένων.

Κεφάλαιο 2

Ανάλυση Επιβίωσης

2.1 Βασικές Έννοιες στην Ανάλυση Επιβίωσης

Στην παρούσα ενότητα θα παρουσιαστούν οι βασικές πτυχές της ανάλυσης επιβίωσης και οι ορισμοί των εννοιών που θα χρησιμοποιηθούν στα επόμενα κεφάλαια. Η ανάγκη για την ανάπτυξη ξεχωριστής στατιστικής μεθοδολογίας για την ανάλυση επιβίωσης δημιουργείται διότι η εξαρτημένη μεταβλητή είναι ο χρόνος, αλλά ο χρόνος δεν μπορεί να μετρηθεί όπως οι άλλες μεταβλητές.

Θεωρούμε μια μη-αρνητική τυχαία μεταβλητή T, η οποία δηλώνει το χρόνο από ένα καλά ορισμένο αρχικό σημείο μέχρι την εμφάνιση ενός γεγονότος. Σε αρκετές περιπτώσεις το γεγονός (event) θα είναι ο θάνατος, επομένως η τυχαία μεταβλητή T θα συμβολίζει το χρόνο επιβίωσης του ατόμου (δίνοντας το όνομα Ανάλυση Επιβίωσης στη σχετική μεθοδολογία). Δεν είναι απαραίτητο ο χρόνος να μετράται μέχρι το θάνατο του ατόμου. Για παράδειγμα, στην ιατρική έρευνα είναι σύνηθες ο χρόνος να μετράται από την ένταξη (recruitment) ενός ασθενούς σε μια πειραματική μελέτη, η οποία γίνεται για να συγκριθούν δυο ή περισσότερες θεραπείες (treatments). Αυτό συχνά συμπίπτει με τη διάγνωση μιας κατάστασης, η οποία κάνει τους ασθενείς κατάλληλους για ενταχθούν στη μελέτη. Το τελικό γεγονός μπορεί να είναι ο θάνατος, η υποτροπή μιας ασθένειας ή η υποχώρηση του πόνου του ασθενούς. Με τον όρο δεδομένα επιβίωσης (time to event data, survival data) θα εννοούμε δεδομένα που προκύπτουν από κάθε, τέτοιου είδους, καταληκτικό συμβάν.

Σε πραγματικά προβλήματα, η μελέτη κάποια στιγμή θα ολοκληρωθεί και ορισμένοι ασθενείς δεν θα έχουν εμφανίσει το γεγονός. Επομένως, για ορισμένα άτομα δεν έχουμε γνώση για το πραγματικό χρόνο επιβίωσής τους. Αυτό το φαινόμενο ονομάζεται λογοκρισία (censoring), και αποτελεί την αιτία διαχωρισμού της ανάλυσης επιβίωσης από τα υπόλοιπα πεδία της στατιστικής. Υπάρχουν διαφορετικοί τύποι λογοκρισίας. Το αποτέλεσμα της λογοκρισίας είναι ότι τα προσβάσιμα σε εμάς δεδομένα θα είναι μια μείξη από διακριτά και συνεχή δεδομένα. Ίδιας φύσεως

δεδομένα μπορούν να προκύψουν από διάφορα επιστημονικά πεδία, αλλά η έμφαση της παρούσας διπλωματικής εργασίας θα είναι σε δεδομένα επιβίωσης που προέκυψαν από την ιατρική έρευνα.

Υπάρχουν διάφοροι τρόποι για να χαρακτηριστεί η κατανομή μιας τυχαίας μεταβλητής. Στην ανάλυση επιβίωσης η έμφαση δίνεται στη συνάρτηση επιβιωσης (survival function), συνάρτηση κινδύνου (hazard function) και στην αθροιστική συνάρτηση κινδύνου (cumulative hazard function). Οι σχέσεις μεταξύ των παραπάνω συναρτήσεων θα δοθούν για τις περιπτώσεις συνεχούς και διακριτής τυχαίας μεταβλητής, T.

Η συνάρτηση επιβίωσης ορίζεται για συνεχείς και διακριτές κατανομές ως η πιθανότητα η T να ξεπεράσει κάποιο χρόνο t, δηλαδή

$$S(t) = \Pr(T > t),\ 0 < t < \infty.$$

Η $S(t)$ είναι φθίνουσα συνάρτηση του t, με $S(0) = 1$ και $\lim_{t\to\infty} S(t) = 0$. Το διάστημα $(0, \infty)$ ερμηνεύεται ως το εύρος για όλες τις πιθανές παραμετροποιήσεις του χρόνου.

Συνεχής τυχαία μεταβλητή T

Συνήθως θεωρούμε ότι ο χρόνος επιβίωσης, T, ακολουθεί κάποια συνεχή κατανομή. Θα συμβολίσουμε τη συνάρτηση πυκνότητας πιθανότητας (probability density function) με f. Ισχύει ότι:

$$f(t) = -\frac{\partial S(t)}{\partial t} = -\lim_{h\to 0} \frac{S(t+h) - S(t)}{h} = \lim_{h\to 0} \frac{\Pr(t < T \leq t + h)}{h}.$$

Η κατανομή του χρόνου επιβίωσης προσδιορίζεται πλήρως και μοναδικά από τη συνάρτηση πυκνότητας f. Η $f(t)$ είναι η πυκνότητα της πιθανότητας αποτυχίας στη χρονική στιγμή t, και για μικρό h ισχύει ότι:

$$f(t)h \simeq \Pr(t < T \leq t + h).$$

Στην ανάλυση επιβίωσης, το ενδιαφέρον επικεντρώνεται στην πιθανότητα επιβίωσης μέτα από ένα χρονικό διάστημα t, η οποία δίνεται από τη συνάρτηση επιβίωσης:

$$S(t) = \Pr(T > t) = \int_t^\infty f(s)ds.$$

Συνήθως, οι μελέτες επιβίωσης διαρκούν αρκετό χρονικό διάστημα, άρα, είναι λογικό ότι ένα μέρος του υπό μελέτη πληθυσμού πεθαίνει. Επομένως, το ενδιαφέρον εστιάζεται στον κίνδυνο που διατρέχει ο υπό μελέτη πληθυσμός ο οποίος βρίσκεται εν ζωή. Μια σημαντική συνάρτηση στην ανάλυση επιβίωσης είναι η συνάρτηση κινδύνου (hazard function), την οποία συμβολί-

ζουμε με $\lambda(t)$. Η συνάρτηση κινδύνου ορίζεται από τη δεσμευμένη πιθανότητα αποτυχίας στο χρόνο $(t, t+h]$, δεδομένου ότι το άτομο βρίσκεται υπό κίνδυνο (at-risk) τη χρονική στιγμή t. Για να καταλήξουμε σε ρυθμό, διαιρούμε τη δεσμευμένη πιθανότητα με το μήκος του χρονικού διαστήματος h. Η συνάρτηση κινδύνου είναι το όριο του πηλίκου για πολύ μικρό χρονικό διάστημα h

$$\begin{aligned}\lambda(t) &= \lim_{h\to 0} \frac{\Pr(t < T \leq t+h | T > t)}{h} = \lim_{h\to 0} \frac{\Pr(t < T \leq t+h)}{h} \times \frac{1}{S(t)} \\ &= \frac{f(t)}{S(t)}. \end{aligned} \tag{2.1}$$

Μια παρόμοια συνάρτηση είναι η συνάρτηση αθροιστικού κινδύνου η οποία ορίζεται ως το ολοκλήρωμα της συνάρτησης κινδύνου, δηλαδή $\Lambda(t) = \int_0^t \lambda(s)ds$. Ο αθροιστικός κίνδυνος δεν έχει φυσική ερμηνεία, αλλά είναι ιδιαίτερα χρήσιμη ποσότητα σε ορισμένες περιπτώσεις. Απο τη σχέση (2.1) προκύπτει ότι:

$$\lambda(t) = -\frac{\partial \log(S(t))}{\partial t}.$$

Αν ολοκληρώσουμε ως προς t και χρησιμοποιήσουμε ότι $S(0) = 1$ έχουμε ότι

$$\Lambda(t) = \int_0^t \lambda(s)ds = -\log(S(t)) + \log(S(0)) = -\log(S(t)),$$

και συνεπώς

$$S(t) = \exp\{-\Lambda(t)\} = \exp\left\{-\int_0^t \lambda(s)ds\right\}. \tag{2.2}$$

Η εξίσωση (2.2) αποτελεί μια σημαντική σχέση στην ανάλυση επιβίωσης. Περιγράφεται η εξάρτηση της συνολικής πιθανότητας επιβίωσης από την συνάρτηση στιγμιαίου κινδύνου. Γενικότερα, η συνάρτηση κινδύνου έχει ελκυστική ερμηνεία με όρους πιθανοτήτων και οδηγεί σε απλούστερες εκφράσεις της συνάρτησης πιθανοφάνειας σε σχέση με τη συνάρτηση πυκνότητας. Περιγράφει τον τρόπο με τον οποίο αλλάζει η στιγμιαία (δεσμευμένη) πιθανότητα αποτυχίας για ένα άτομο στην πάροδο του χρόνου. Πιο συνηθισμένη περίπτωση είναι να θεωρήσουμε ότι ο χρόνος αποτυχίας ακολουθεί συνεχή κατανομή. Συχνά, σε πρακτικές εφαρμογές έχουμε εκ των προτέρων πληροφορία για τη μορφή της συνάρτησης κινδύνου, το οποίο μας βοηθάει να διαλέξουμε ένα κατάλληλο μοντέλο.

Η μορφή της συνάρτησης κινδύνου μπορεί να πάρει διάφορες μορφές: μπορεί να είναι αύξουσα (Weibull με $\gamma > 1$), φθίνουσα (Weibull με $\gamma < 1$), σταθερή (Weibull με $\gamma = 1$) ή U-shaped. Για παράδειγμα, σε μελέτες που περιλαμβάνουν άτομα μεγάλης ηλικίας είναι λογικό να υποθέσουμε ότι ο κίνδυνος αυξάνει με το χρόνο. Επίσης, σε μελέτες της επιβίωσης

από τη γέννηση του ανθρώπου μέχρι το θάνατο, μια U-shaped μορφή της συνάρτησης κινδύνου θα μπορούσε να είναι κατάλληλη. Μετά από μια περίοδο από τη γέννηση του ανθρώπου στην οποία η θνησιμότητα οφείλεται κυρίως σε βρεφικές ασθένειες και γενετικές ανωμαλίες, η θνησιμότητα πέφτει και παραμένει περίπου σταθερή μέχρι την ηλικία των 30, όπου αυξάνεται πάλι. Σε ορισμένες περιπτώσεις, μοντέλα με φθίνουσα συνάρτηση κινδύνου μπορεί να είναι χρήσιμα. Για παράδειγμα, σε ασθενείς οι οποίοι υποβλήθηκαν σε μια επικίνδυνη εγχείρηση, ο κίνδυνος μπορεί να μειώνεται μετά από ένα χρονικό διάστημα. Η εξήγηση είναι η εξής: οι ασθενείς οι οποίοι βίωσαν επιπλοκές είναι πολύ πιθανό να πέθαιναν σε πολύ μικρό χρονικό διάστημα από την εγχειρηση, αφήνοντας τους ασθενείς χωρίς επιπλοκές στον υπό μελέτη πληθυσμό, οι οποίοι, πιθανώς, έλαβαν την ευεργητική επίδραση της εγχειρησης (Wienke 2010, σελ. 18). Σε κάθε περίπτωση, η εκ των προτέρων γνώση της συνάρτησης κινδύνου μας βοηθάει να διαλέξουμε ένα μοντέλο το οποίο εφαρμόζει καλά στα δεδομένα και έχει λογική ερμηνεία.

Διακριτή τυχαία μεταβλητή T

Αν θεωρήσουμε ότι ο χρόνος αποτυχίας, T, είναι διακριτή τυχαία μεταβλητή, η οποία παίρνει τις τιμές $\alpha_1 < \alpha_2 < \ldots$, η συνάρτηση πιθανότητας και η συνάρτηση επιβίωσης θα είναι:

$$f(\alpha_j) = \Pr(T = \alpha_j), \quad S(t) = \Pr(T > t) = \sum_{j|\alpha_j > t} f(\alpha_j) \quad j = 1, 2, \ldots$$

Η συνάρτηση κινδύνου (hazard function) στο χρόνο α_j ορίζεται ως η δεσμευμένη πιθανότητα αποτυχίας στο α_j δεδομένου ότι το άτομο έχει επιβιώσει μέχρι το α_j, δηλαδή

$$\lambda_j = \Pr(T = \alpha_j | T \geq \alpha_j) = \frac{\Pr(T = \alpha_j | T \geq \alpha_j)}{\Pr(T \geq \alpha_j)} = \frac{f(\alpha_j)}{S(\alpha_j^-)} \quad i = 1, 2, \ldots, \tag{2.3}$$

όπου $S(\alpha_j^-) = \lim_{t \to \alpha_j^-} S(t)$. Αντίστοιχα με την περίπτωση συνεχούς τυχαίας μεταβλητής, η συνάρτηση πιθανότητας δίνεται από την

$$f(\alpha_i) = \lambda_i S(\alpha_i^-) = \lambda_i \prod_{j=1}^{i-1} (1 - \lambda_j). \tag{2.4}$$

Έστω ότι $t \in [\alpha_\kappa, \alpha_{\kappa+1})$, η συνάρτηση επιβίωσης δίνεται από την

$$\begin{aligned} S(t) =& \Pr(T > t) = \Pr(T > \alpha_\kappa) = \Pr(T \geq \alpha_1, T \geq \alpha_2, \ldots, T \geq \alpha_\kappa, T \geq \alpha_{\kappa+1}) \\ =& \Pr(T \geq \alpha_1) \times \Pr(T \geq \alpha_2 | T \geq \alpha_1) \times \ldots \times \Pr(T \geq \alpha_{\kappa+1} | T \geq \alpha_\kappa, \ldots, T \geq \alpha_1) \\ =& 1 \times \prod_{j=1}^{\kappa} \Pr(T \geq \alpha_{j+1} | T \geq \alpha_j) = \prod_{j=1}^{\kappa} \{1 - \Pr(T = \alpha_j | T \geq \alpha_j)\} = \prod_{j|\alpha_j \leq t} (1 - \lambda_j). \end{aligned} \tag{2.5}$$

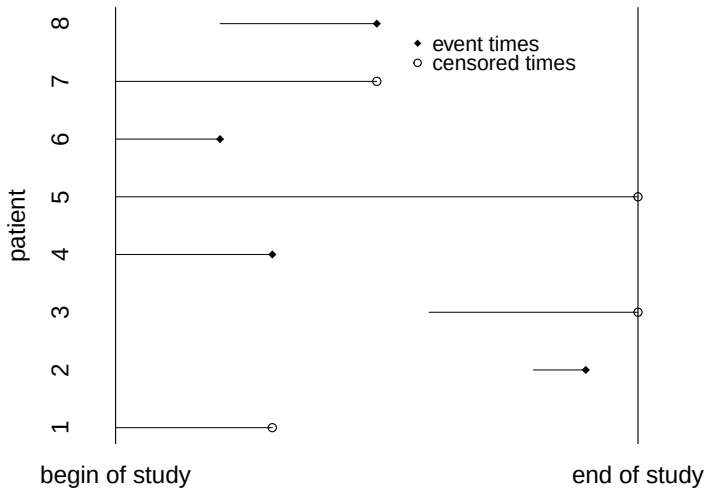
Οι σχέσεις (2.4) και (2.5) μπορούν να ερμηνευτούν, θεωρώντας μια διαδικασία δοκιμών σε όλα τα χρονικά διαστήματα κατά την οποία το άτομο θα αποτύχει ή θα επιβιώσει σε κάθε χρονικό διάστημα (Kalbfleisch & Prentice 2002, σελ. 9) . Για παράδειγμα, η (2.4) μπορεί να εξηγηθεί ως εξής: ένα άτομο θα αποτύχει στο χρόνο α_i αν και μόνο αν επιβιώσει στα προηγούμενα χρονικά διαστήμα με (δεσμευμένες) πιθανότητες $1-\lambda_1, 1-\lambda_2, \ldots, 1-\lambda_{i-1}$, και εν συνεχεία, έχοντας επιβιώσει μέχρι το α_i, θα αποτύχει στο α_i με (δεσμευμένη) πιθανότητα λ_i.

2.2 Λογοκρισία

Στην ανάλυση δεδομένων επιβίωσης είναι σημαντικό να υπάρχει ξεκάθαρος ορισμός της αρχής του χρόνου από την οποία μετράται η επιβίωση. Σε ορισμένες περιπτώσεις, η χρονική κλίμακα είναι η ηλικία, επομένως η αρχή του χρόνου είναι η γέννηση του ατόμου. Σε άλλες περιπτώσεις, ως αρχή του χρόνου ορίζεται η εμφάνιση κάποιου γεγονότος, όπως η διάγνωση μιας ασθένειας ή η τυχαιοποίηση ενός ασθενούς σε μια κλινική δοκιμή. Είναι εξίσου σημαντικό να υπάρχει σαφής ορισμός του τι συνιστά αποτυχία. Για παράδειγμα, σε μια κλινική δοκιμή η οποία συγκρίνει θεραπείες καρδιακών νοσημάτων, η καταγεγραμμένη εμφάνιση καρδιακής προσβολής μπορεί να θεωρηθεί ως ένα από τα κριτήρια εισαγωγής ενός ασθενούς στη μελέτη. Η αρχή του χρόνου θα μπορούσε να οριστεί ως η στιγμή της τυχαιοποίησης του ατόμου, και ως αποτυχία η επανεμφάνιση καρδιακής προσβολής με σαφή ιατρικά κριτήρια.

Όπως έχουμε δει, πολλές φορές οι μελέτες τελειώνουν και αρκετοί ασθενές δεν έχουν εμφανίσει το γεγονός το οποίο μελετάται. Τέτοιου είδους δεδομένα λέμε ότι είναι δεξιά λογοκριμένα (right censored). Μια λογοκριμένη (censored) παρατήρηση είναι μια ημιτελής παρατήρηση καθώς περιέχει μερική πληροφορια για το γεγονός. Σε κάποιες περιπτώσεις, δεξιά λογοκριμένα δεδομένα προκύπτουν γιατι οι ασθενείς δεν έχουν αποτύχει στο χρονικό διάστημα της μελέτης. Σε άλλες περιπτώσεις, οι ασθενείς έχουν φύγει από την περιοχή της μελέτης για λόγους ασυσχέτιστους με την κατάσταση της υγείας τους, με αποτέλεσμα η επαφή να χαθεί. Όμως είναι πιθανό η επαφή να χαθεί λόγω αλλαγής της πρόγνωσης του ασθενούς. Επίσης, είναι δυνατόν το γενονός που μας ενδιαφέρει να μην παρατηρηθεί λόγω ενός ανταγωνιστικού κινδύνου (competing risk), όπως για παράδειγμα ο θάνατος από ατύχημα. Είναι λογικό ότι κάποιοι μηχανισμοί λογοκρισίας δεν είναι ανεξάρτητοι με το καταληκτικό συμβάν και μπορούν να εισάγουν μεροληψία στις εκτιμήσεις.

Στο γράφημα (2.1) βλέπουμε ένα παράδειγμα λογοκριμένων δεδομένων. Παρατηρούμε ότι οι ασθενείς 2, 4, 6, 8 παρακολουθήθηκαν μέχρι να αναπτύξουν το γεγονός. Αντίθετα, οι ασθενείς 1 και 7 χάθηκαν από τη μελέτη για κάποιους λόγους. Οι χρόνοι των ασθενών 3 και 5 είναι λογοκριμένοι εξαιτίας της λήξης της μελέτης. Έστω ότι $T_1, T_2, \ldots, T_n$ οι πραγματικοί χρόνοι επιβίωσης n ανεξάρτητων ατόμων με συνάρτηση κατανομής F και $C_1, C_2, \ldots, C_n$



Σχήμα 2.1: Χρόνοι επιβίωσης και δεξιά λογοκριμένοι χρόνοι ασθενών σε μια κλινική δοκιμή

οι χρόνοι λογοκρισίας με συνάρτηση κατανομής G. Θα θεωρήσουμε ότι οι F και G είναι συνεχείς κατανομές. Επίσης, έστω f και g οι συνάρτησεις πυκνότητας των F και G, αντίστοιχα. Θα εξετάσουμε την περίπτωση της δεξιάς λογοκρισίας, όπου ο ασθενής παρακολουθείται για κάποια χρονική περίοδο αλλά δεν έχει συμβεί αποτυχία. Είμαστε σε θέση να παρατηρήσουμε τα δεδομένα $(Y_1, \Delta_1), (Y_2, \Delta_2), \ldots, (Y_n, \Delta_n)$, όπου $Y_i = \min\{T_i, C_i\}$ και $\Delta_i = \mathbb{1}(T_i \leq C_i),\ i = 1, 2, \ldots, n$. Σε κάθε άτομο είναι γνωστό ένα διάνυσμα από p επεξηγηματικές μεταβλητές, $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \ldots, x_{ip})$, οι οποίες έχουν μετρηθεί στην αρχή του χρόνου. Θα εξετάσουμε την αρκετά περιοριστική υπόθεση της ανεξάρτητης λογοκρισίας. Συγκεκριμένα, υποθέτουμε ότι οι χρόνοι λογοκρίσιας $C_1, C_2, \ldots, C_n$, δοθέντων των επεξηγηματικών μεταβλητών, είναι ανεξάρτητες μεταξύ τους τυχαίες μεταβλητές και ανεξάρτητες από τους χρόνους επιβίωσης $T_1, T_2, \ldots, T_n$. Έστω H_0 και H_1 οι συναρτήσεις υπο-κατανομής του χρόνου παρατήρησης Y και h_0 και h_1 οι υπο-πυκνότητες. Ισχύει ότι (Wienke 2010, σελ. 20)

$$\begin{aligned} H_1(y_i|\mathbf{x}_i) &= \Pr(Y_i \leq y_i, \Delta_i = 1|\mathbf{x}_i) = \Pr(T_i \leq y_i, T_i \leq C_i|\mathbf{x}_i) \\ &= \int_0^{y_i} \int_y^{\infty} f(y|\mathbf{x}_i) g(c|\mathbf{x}_i)\, dcdy = \int_0^{y_i} f(y|\mathbf{x}_i)\,(1 - G(y|\mathbf{x}_i))\, dy. \end{aligned}$$

Επίσης,

$$\begin{aligned} H_0(y_i|\mathbf{x}_i) &= \Pr(Y_i \leq y_i, \Delta_i = 0|\mathbf{x}_i) = \Pr(C_i \leq y_i, C_i < T_i|\mathbf{x}_i) \\ &= \int_0^{y_i} \int_c^{\infty} f(y|\mathbf{x}_i) g(c|\mathbf{x}_i)\, dydc = \int_0^{y_i} g(c|\mathbf{x}_i)\,(1 - F(c|\mathbf{x}_i))\, dc. \end{aligned}$$

Επομένως, οι υπο-πυκνότητες θα είναι $h_1(y_i|\mathbf{x}_i) = \frac{\partial H_1(y_i|\mathbf{x}_i)}{\partial y_i} = f(y_i|\mathbf{x}_i)(1 - G(y_i|\mathbf{x}_i))$ και $h_0(y_i|\mathbf{x}_i) = \frac{\partial H_0(y_i|\mathbf{x}_i)}{\partial y_i} = g(y_i|\mathbf{x}_i)\,(1 - F(y_i|\mathbf{x}_i))$, αντίστοιχα. Άρα η συνάρτηση πυκνότητας για τα δεδομένα (Y_i, Δ_i) θα είναι

$$\begin{aligned} f(y_i, \delta_i) &= (h_1(y_i|\mathbf{x}_i))^{\delta_i} \times (h_0(y_i|\mathbf{x}_i))^{1-\delta_i} \\ &= \{f(y_i|\mathbf{x}_i)(1 - G(y_i|\mathbf{x}_i))\}^{\delta_i} \times \{g(y_i|\mathbf{x}_i)\,(1 - F(y_i|\mathbf{x}_i))\}^{1-\delta_i}. \end{aligned}$$

Αν επιπλέον υποθέσουμε ότι η κατανομή του χρόνου λογοκρισίας δεν περιέχει καμία πληροφορία για τις παραμέτρους οι οποίες σχετίζονται με τη συνάρτηση επιβίωσης, τότε οι παράγοντες $(1 - G(y_i|\mathbf{x}_i))^{\delta_i}$ και $(g(y_i|\mathbf{x}_i))^{1-\delta_i}$ δεν παρέχουν πληροφορία για τη συνάρτηση επιβίωσης και μπορούν να παραληφθούν από την πιθανοφάνεια. Επομένως, η συνάρτηση πιθανοφάνειας για n ανεξάρτητα άτομα με δεξιά λογοκρισία, κάτω από την υπόθεση της μη-πληροφοριακης και ανεξάρτητης λογοκρισίας, θα είναι

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} (f(y_i|\mathbf{x}_i))^{\delta_i}\,(S(y_i|\mathbf{x}_i))^{1-\delta_i} = \prod_{i=1}^{n} (\lambda(y_i|\mathbf{x}_i))^{\delta_i}\,S(y_i|\mathbf{x}_i). \tag{2.6}$$

Μπορεί να αποδειχθεί ότι η συνάρτηση πιθανοφάνειας (2.6) ισχύει και στην περίπτωση όπου οι χρόνοι λογοκρισίας δεν είναι τυχαίες μεταβλητές αλλά σταθερές (Duchateau & Janssen 2007, σελ. 20). Όταν ο χρόνος αποτυχίας ενός ατόμου είναι γνωστό ότι βρίσκεται στο διάστημα (l_i, r_i) (interval censoring), τότε το άτομο συνεισφέρει στην πιθανοφάνεια με τον όρο $S(l_i) - S(r_i)$.

2.3 Παραμετρικά Μοντέλα

Στη παρούσα ενότητα θα παρουσιαστούν μερικές κατανομές πιθανότητας οι οποίες είναι ιδιαίτερα χρήσιμες στην ανάλυση επιβίωσης. Γενικά, κάθε κατανομή η οποία παίρνει μη αρνητικές τιμές μπορεί να περιγράψει χρόνικά διαστήματα μέχρι κάποιο γεγονός. Οι περισσότερες κατανομές μοντελοποιούν τη συνάρτηση επιβίωσης με ένα μικρό αριθμό παραμέτρων. Εξαίρεση αποτελεί η κατά τμήματα (piecewise) εκθετική κατανομή, η οποία μπορεί να περιλαμβάνει πολλές παραμέτρους. Σηνήθη παραμετρικά μοντέλα τα οποία εφαρμόζονται στην ανάλυση επιβίωσης, βασίζονται στην Εκθετική, στην Weibull και στην Gompertz κατανομή. Σε αυτές τις κατανομές μπορουν να βρεθούν οι συναρτήσεις επιβίωσης, πυκνότητας και κινδύνου (hazard function) σε κλείστη μορφή. Αντίθετα, οι κατανομές Gamma and Lognormal παρουσιάζουν υπολογιστικές δυσκολίες, αλλά χρησιμοποιούνται αρκετά συχνά στην πράξη. Σε κάθε περίπτωση, μπορούμε να αξιολογήσουμε την προσαρμογή του μοντέλου συγκρίνοντας τα αποτελέσματα με τον απαραμετρικό εκτιμητή Kaplan–Meier (Kaplan & Meier 1958). Αν το παραμετρικό μοντέλο έχει ικανοποιητική προσαρμογή στα δεδομένα, αναμένουμε ότι τα τυπικά σφάλματα των εκτιμητών της συνάρτησης επιβίωσης θα είναι μικρότερα σε σύγκριση με τις απαραμετρικές μεθόδους.

Στις επομένες ενότητες θα παρουσιαστούν παραμετρικά μοντέλα για ομοιογενείς πληθυσμους. Οι ιδιότητες των κατανομών θα συζυτηθούν μερικώς. Συμβολίζουμε με T τη μεταβλητή η οποία δηλώνει το χρόνο μέχρι κάποιο γεγονός, για τον οποίο θέλουμε να βγάλουμε συμπεράσματα. Θα περιοριστούμε στην περίπτωση ανεξάρτητων και ισόνομων χρόνων $T_1, T_2, \ldots, T_n$. Ως συνήθως, είμαστε σε θέση να παρατηρήσουμε τα δεδομένα $(Y_1, \Delta_1), (Y_2, \Delta_2) \ldots, (Y_n, \Delta_n)$, όπου $Y_i = \min\{T_i, C_i\}$ και $\Delta_i = \mathbb{1}(T_i \leq C_i)$. Συμβολίζουμε τα παρατηρούμενα δεδομένα με (y_i, δ_i), $i = 1, 2, \ldots, n$.

2.3.1 Εκθετική Κατανομή

Η εκθετική κατανομή ($T \sim \exp(\lambda)$, $\lambda > 0$) είναι το πιο απλό παραμετρικό μοντέλο, καθώς περιέχει μόνο μια άγνωστη παράμετρο. Το μοντέλο υποθέτει ότι κάθε άτομο έχει σταθερό κίνδυνο να εμφάνισει το γεγονός, ανεξάρτητα από το πόσος χρόνος έχει περάσει κατα τον οποίο το άτομο ήταν υπο παρακολούθηση και δεν είχε εμφανίσει το γεγονός που μας ενδιαφέρει. Η ιδιότητα αυτή της εκθετικής συχνά αποκαλείται 'έλλειψη μνήμης'. Αυτό σημαίνει ότι η κατανομή του $T - t | T \geq t$ παραμένει ίδια με την αρχική κατανομή του T. Εύκολα διαπιστώνουμε ότι:

$$\Pr(t < T \leq t + \epsilon | T > t) = \Pr(T \leq \epsilon),$$

για κάθε θετικό ϵ. Επομένως, η πιθανότητα αποτυχίας σε ένα χρονικό διάστημα εξάρταται από το μήκος του διαστήματος και όχι από τη θέση του. Η ισχυρή υπόθεση η οποία επιφέρει το μοντέλο της εκθετικής κατανομής, έχει ως αποτέλεσμα τη σπάνια χρησιμοποίηση της στη μελέτη της

ανθρώπινης επιβίωσης (εκτός αν μελετάμε πολύ μικρά χρονικά διαστήματα). Για το εκθετικό μοντέλο ισχύουν ακόλουθες σχέσεις:

Συνάρτηση πυκνότητας	$f(t) = \lambda \exp\{-\lambda t\}$
Συνάρτηση επιβίωσης	$S(t) = \exp\{-\lambda t\}$
Συνάρτηση κινδύνου	$\lambda(t) = \lambda$
Αθροιστική συνάρτηση κινδύνου	$\Lambda(t) = \lambda t$
Μέση τιμή	$\mathrm{E}(T) = \frac{1}{\lambda}$
Διακύμανση	$\mathrm{Var}(T) = \frac{1}{\lambda^2}$

Η κατανομή του cT, όπου c θετική σταθερά, είναι επίσης εκθετική με παράμετρο λ/c. Επιπλέον, η ελάχιστη παρατήρηση από n ανεξάρτητες εκθετικές τυχαίες μεταβλητές με παράμετρο λ είναι πάλι εκθετική με παράμετρο $n\lambda$, δηλαδή

$$\Pr\left(\min\{T_1, T_2, \ldots, T_n\} > t\right) = \prod_{i=1}^{n} \Pr(T_i > t) = \exp\{-\lambda n t\}.$$

Η εκθετική κατανομή αποτελεί υποπερίπτωση της Weibull και piecewise εκθετικής κατανομής, οι οποίες θα εξεταστούν στις επόμενες ενότητες. Ένας τρόπος να αξιολογήσουμε την υπόθεση είναι να θεωρήσουμε το εκθετικό μοντέλο ως εμφωλευμένο στα 2 προηγούμενα. Επίσης, αν η εκθετική κατανομή εφαρμόζει καλά στα δεδομένα και $\hat{S}(t)$ είναι μια απαραμετρική εκτίμηση για τη συνάρτηση επιβίωσης $S(t)$, τότε αναμένουμε ότι ένα γράφημα του $-\log\{\hat{S}(t)\}$ σε σχέση με το χρόνο t, θα είναι προσεγγιστικά μια ευθεία γραμμή.

Παραδειγμα 2.1.

Έστω $T_1, T_2, \ldots, T_n$ ανεξάρτητες και ισόνομες παρατηρήσεις από την $\exp(\lambda)$. Κάτω από την υπόθεση της μη πληροφοριακής λογοκρισίας, η συνάρτηση πιθανοφάνειας για τα παρατηρούμενα δεδομένα (y_i, δ_i) $i = 1, 2, \ldots, n$, θα είναι:

$$L(\lambda) = \prod_{i=1}^{n} \lambda(y_i)^{\delta_i} S(y_i) = \lambda^{\sum_{i=1}^{n} \delta_i} \exp\left\{ -\lambda \sum_{i=1}^{n} y_i \right\}$$

Αν λογαριθμήσουμε την πιθανοφάνεια και θέσουμε την πρώτη παράγωγο, ως προς λ, ίση με μηδέν, βρίσκουμε ότι:

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} y_i}. \tag{2.7}$$

Ο εκτιμητής (2.7) είναι το πηλίκο του συνολικού αριθμού γεγονότων δια το συνολικό χρόνο παρακολούθησης. Μπορούμε να παρατηρήσουμε την άμεση σχέση του εκτιμητή μέγιστης πιθα-

νοφάνειας (2.7) με τις μεθόδους επιδημιολογίας που στηρίζονται στον ανθρωποχρόνο.

2.3.2 Μοντέλο Σταθερού κατά Τμήματα Κινδύνου

Η εκθετική κατανομή έχει μια παράμετρο, άρα, είναι αρκετά σπάνιο να συνθέτει ικανοποιητικά τα παρατηρούμενα δεδομένα. Μια ενδιαφέρουσα επέκταση του μοντέλου της εκθετικής κατανομής είναι να υποθέσουμε ότι ο κίνδυνος παραμένει σταθερός σε προκαθορισμένα χρονικά διαστήματα, δηλαδή στην ουσία να υποθέσουμε ένα ξεχωριστό εκθετικό μοντέλο για κάθε χρονικό διάστημα. Το κατά τμήματα εκθετικό μοντέλο είναι ιδιαίτερο χρήσιμο στην πράξη γιατι είναι σχετικά απλό και και μπορεί να αποδώσει αρκετές μορφές (shapes) στην συνάρτηση κινδύνου, υπό την προϋπόθεση ότι έχουμε χωρίσει το χρόνο σε ικανοποιητικό αριθμό διαστημάτων. Βέβαια, για κάθε ένα επιπλέον χρονικό διάστημα πρέπει να εκτιμήσουμε μια επιπλέον παράμετρο. Το κύριο μειονέκτημα του μοντέλου είναι οτι η συνάρτηση κινδύνου δεν είναι συνεχής, αφού παρουσιάζει άλματα στο τέλος κάθε διαστήματος.

Για να δημιουργήσουμε το μοντέλο θεωρούμε μια διαμέριση του χρονικού άξονα σε $J-1$ σημεία, $0 = s_0 < s_1 < s_2 < \ldots < s_{J-1} < s_J$, όπου $s_J = \infty$. Σε κάθε ένα χρονικό διάστημα υποθέτουμε διαφορετικό κίνδυνο, άρα για το j-οστό διάστημα έχουμε ότι $\lambda(t) = \lambda_j$ για $t \in [s_{j-1}, s_j)$, $j = 1, 2, \ldots, J$. Επομένως, έχουμε J χρονικά διαστήματα

$$\underbrace{[0, s_1)}_{\lambda_1} \; \underbrace{[s_1, s_2)}_{\lambda_2} \; \ldots \; \underbrace{[s_{i-1}, s_i)}_{\lambda_i} \; \ldots \; \underbrace{[s_{J-1}, s_J)}_{\lambda_J}$$

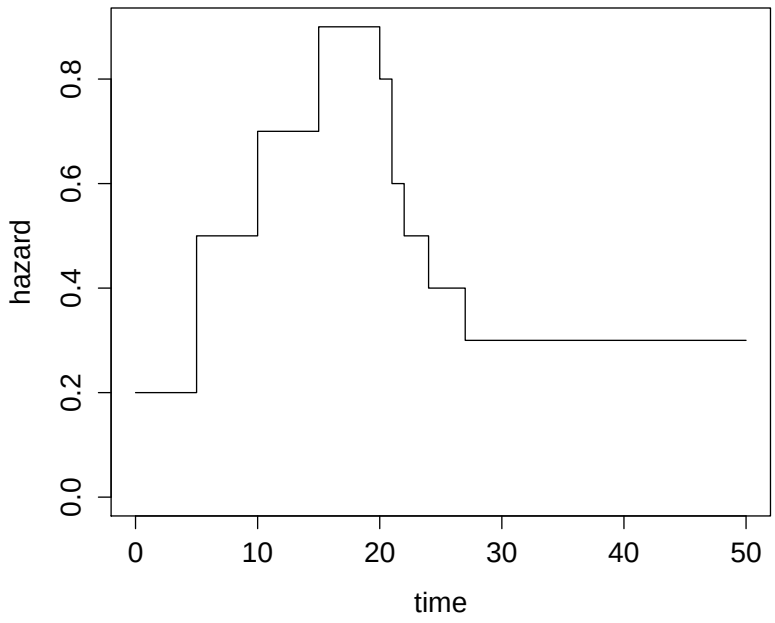
Ορίζουμε τη συνάρτηση κινδύνου ως εξής:

$$\lambda(t) = \sum_{j=1}^{J} \lambda_j I_j(t), \text{ όπου } I_j(t) = \mathbb{1}\left(s_{j-1} \leq t < s_j\right).$$

Η πιθανότητα επιβίωσης πέραν του χρόνου t, αν το t ανήκει στο i-οστό χρονικό διάστημα $(s_{i-1} \leq t < s_i)$ θα είναι:

$$\begin{aligned} S(t) &= \exp\left\{ -\int_0^t \lambda(s)ds \right\} = \exp\left\{ -\sum_{j=1}^{J} \lambda_j \int_0^t I_j(s)ds \right\} \\ &= \exp\left\{ -\sum_{j=1}^{i-1} \lambda_j \int_0^t I_j(s)ds - \lambda_i \int_0^t I_i(s)ds - \sum_{j=i+1}^{J} \lambda_j \int_0^t I_j(s)ds \right\} \\ &= \exp\left\{ -\lambda_i \left(t - s_{i-1}\right) - \sum_{j=1}^{i-1} \lambda_j \left(s_j - s_{j-1}\right) \right\}. \end{aligned}$$

Στο γράφημα 2.2 δίνεται ένα παράδειγμα σταθερού κατα τμήματα κινδύνου. Ο χρονικός άξονας έχει διαμεριστεί σε 9 χρονικά διαστήματα.



Σχήμα 2.2: Μοντέλο σταθερού κατά τμήματα κινδύνου με διαμέριση του χρόνου και ρυθμούς αποτυχίας $\{0, 5, 10, 15, 20, 21, 22, 24, 27, \infty\}$ και $\{0.2, 0.5, 0.7, 0.9, 0.8, 0.6, 0.5, 0.4, 0.3\}$, αντίστοιχα.

2.3.3 Κατανομή Weibull

Η Weibull κατανομή, η οποία προτάθηκε από τον Σουηδό φυσικό Waloddi Weibull το 1939, αποτελεί γενίκευση της εκθετικής κατανομής με δυο θετικές παραμέτρους. Η δεύτερη παράμετρος γ διαχειρίζεται τη μορφή της συνάρτησης κινδύνου και είναι ικανή να περιγράψει διάφορα σχήματα. Το κύριο πλεονέκτημα της κατανομής είναι η ευελιξία σε συνδιασμό με κλειστές εκφράσεις για τις συναρτήσεις επιβίωσης και κινδύνου.

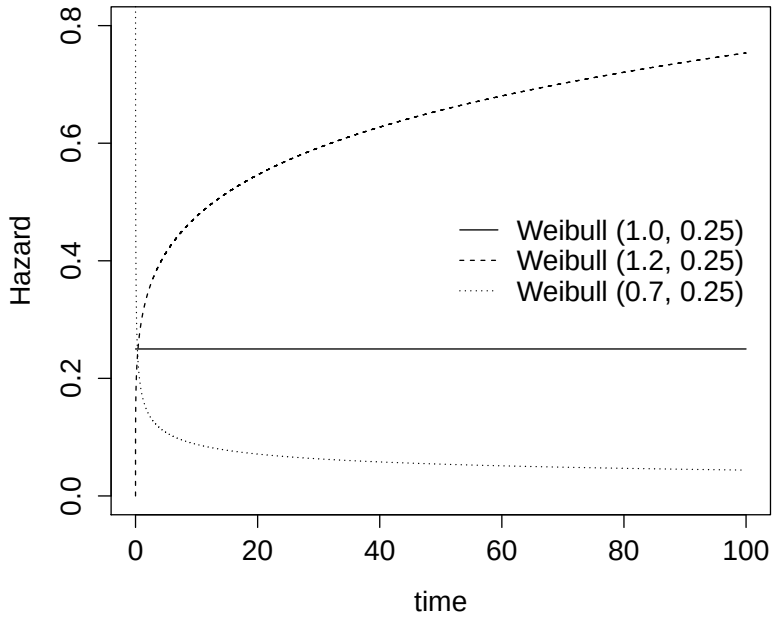
Συνάρτηση πυκνότητας	$f(t) = \lambda\gamma t^{\gamma-1}\exp\{-\lambda t^{\gamma}\} \quad (\lambda, \gamma > 0)$
Συνάρτηση επιβίωσης	$S(t) = \exp\{-\lambda t^{\gamma}\}$
Συνάρτηση κινδύνου	$\lambda(t) = \lambda\gamma t^{\gamma-1}$
Αθροιστική συνάρτηση κινδύνου	$\Lambda(t) = \lambda t^{\gamma}$
Μέση τιμή	$\mathrm{E}(T) = \lambda^{-\frac{1}{\gamma}}\Gamma(\frac{1}{\gamma}+1)$
Διακύμανση	$\mathrm{Var}(T) = \lambda^{-\frac{2}{\gamma}}\left(\Gamma(\frac{2}{\gamma}+1) - \left(\Gamma(\frac{1}{\gamma}+1)\right)^2\right)$

όπου με Γ συμβολίζουμε την γαμμα συνάρτηση $\Gamma(k) = \int_0^{\infty} s^{k-1}\exp\{-s\}ds$. Αν $\gamma = 1$, τότε καταλήγουμε στην εκθετική κατανομή. Αν $\gamma < 1$, τότε ο κίνδυνος είναι άπειρος στο χρόνο μηδέν και μειώνεται μονότονα στο μηδέν σε άπειρο χρόνο. Αντίθετα, άν $\gamma > 1$, τότε ο κίνδυνος είναι μηδενικός στο χρόνο μηδέν και αυξάνεται μονότονα στο άπειρο, σε άπειρο χρόνο. Σε περίπτωση

που $\gamma = 1$, ο κινδύνος παραμένει σταθερός σε όλη τη διάρκεια του χρόνου.

Η Weibull κατανομή μπορεί να δημιουργηθεί ως η οριακή κατανομή της ελάχιστης παρατήρησης από ένα δείγμα από συνεχή κατανομή με στήριγμα $[0, u)$, οπου $0 < u < \infty$. Εξαιτίας της ιδιότητα αυτής, η Weibull κατανομή μπορεί να θεωρηθεί (αρχικά) κατάλληλη για να μοντελοποιήσει την κατανομή του χρόνου θανάτου ενός ατόμου. Πολλές αιτίες θανάτου ανταγωνίζονται μεταξύ τους, αυτή η οποία θα έρθει πιο γρήγορα θα επιφέρει το θάνατο. Μια αναλυτική παρουσίαση της Weibull κατανομής δίνεται από τους Murthy, Xie & Jiang (2004).

Ενας γραφικός τρόπος να ελέγξουμε την καταλληλότητα της Weibull κατανομής είναι ο εξής: αν $\hat{S}(t)$ μια απαραμετρική εκτίμηση για τη συνάρτηση επιβίωσης και το Weibull μοντέλο είναι ικανοποιητικό για να συνθέσει τα δεδομένα, τότε ένα γράφημα του $\log[-\log(\hat{S}(t))]$ με το λογάριθμο του χρόνου $\log(t)$, θα πρέπει να προσεγγίζεται από μια ευθεία γραμμή. Οι εκτιμητές ελαχίστων τετραγώνων $(\hat{\beta}_0, \hat{\beta}_1)$, ενός γραμμικού μοντέλου μεταξύ του $\log[-\log(\hat{S}(t))]$ και $\log(t)$, μας δίνουν μια πρόχειρη εκτίμηση για το $\log(\lambda)$ και γ, αντίστοιχα. Στο γράφημα 2.3 βλέπουμε τη συνάρτηση κινδύνου της Weibull κατανομής, για διάφορες τιμές του γ. Ένα μειονέκτημα της Weibull κατανομής είναι ότι μπορεί να μοντελοποιήσει μόνο μονότονες συναρτήσεις κινδύνου.



Σχήμα 2.3: Συναρτήσεις κινδύνου Weibull κατανομής για διάφορες τιμές του γ

2.4 Εκτιμητής Kaplan–Meier

Για να εφαρμόσουμε παραμετρικά μοντέλα, πρέπει να υποθέσουμε κάποια κατανομή πιθανότητας για το χρόνο επιβίωσης T. Αν υπάρχει ισχυρή προηγούμενη γνώση για τη μορφή της συνάρτησης επιβίωσης, είναι εύλογο ότι μπορούμε να χρησιμοποιήσουμε ένα κατάλληλο παραμετρικό μοντέλο. Ωστόσο, το πλεονέκτημα των απαραμετρικών μεθόδων είναι ότι δίνουν καλά αποτελέσματα, οποιαδήποτε κατανομή και αν ακολουθεί ο χρόνος T. Ένας τρόπος να διερευνήσουμε τη μορφή της καμπύλης επιβίωσης σε ένα μεμονωμένο group ατόμων, είναι ένα γράφημα με την εμπειρική συνάρτηση επιβίωσης. Αν δεν υπάρχει λογοκρισία (censoring), τότε η εμπειρική συνάρτηση επιβίωσης ορίζεται ως το ποσοστό των ατόμων οι οποίοι έχουν επιβιώσει πέραν του χρόνου t. Σε πρακτικές εφαρμογές, όμως, θα έχουμε censoring, άρα χρειαζόμαστε μια διαφορετική μεθοδολογία. Ας υποθέσουμε ότι $y_1 < y_2 < \ldots < y_D$ είναι οι παρατηρούμενοι χρόνοι αποτυχίας σε δείγμα ατόμων μεγέθους n από ενα πληθυσμό με συνάρτηση επιβίωσης S. Επίσης, d_j άτομα αποτυγχάνουν τη χρονική στιγμή y_j και m_j είναι δεξιά λογοκριμένα στο χρονικό διάστημα $[y_j, y_{j+1})$ τις χρονικές στιγμές $y_{j1} \leq y_{j2} \leq \ldots \leq y_{jm_j}$, $j = 0, 1, \ldots, D$, όπου $y_0 = 0$ και $y_{D+1} = \infty$. Έστω ότι

$$r_j = (m_j + d_j) + \cdots + (m_D + d_D) = \sum_{i=j}^{D} (m_i + d_i),$$

όπου r_j είναι ο αριθμός των ατόμων σε κίνδυνο ελάχιστα πριν την j-οστή αποτυχία. Η συνεισφορά στην πιθανοφάνεια για ένα άτομο το οποίο απέτυχε τη χρονική στιγμή y_j είναι:

$$f(y_j) = -\left.\frac{\partial S(y)}{\partial y}\right|_{y=y_j},$$

ενώ κάτω από την υπόθεση της μη πληροφοριακής λογοκρισίας (non-informative censoring), η συνεισφορά στην πιθανοφάνεια για μια λογοκριμένη παρατήρηση στο χρόνο y_{ji} είναι:

$$S(y_{ji}) = \Pr(T > y_{ji}).$$

Επομένως, η πιθανοφάνεια των δεδομένων θα είναι η εξής:

$$L = \prod_{j=0}^{D} \left\{ f(y_j)^{d_j} \prod_{i=1}^{m_j} S(y_{ji}) \right\}.$$

Αν επικεντρωθούμε στη μεγιστοποίηση της πιθανοφάνειας ως προς τη συνάρτηση επιβίωσης, S, έτσι ωστε η S να αποδίδει θετική πιθανότητα μόνο στους παρατηρούμενους χρόνους αποτυχίας, τότε καταλήγουμε σε μια διακριτή συνάρτηση επιβίωσης η οποία είναι ασυνεχής στους χρόνους αποτυχίας. Επομένως, $S(y) = S(y_j)$ για $y \in [y_j, y_{j+1})$. Αφού οι λογοκριμένοι χρόνοι $y_{ji} \in$

$[y_j, y_{j+1})$, ισχύει ότι $S(y_{ji}) = S(y_j) = \Pr(T > y_j)$. Επίσης, η συνάρτηση πιθανότητας f μπορεί να γραφτεί ως: $f(y_j) = \Pr(T = y_j) = \Pr(T \geq y_j) - \Pr(T > y_j) = S(y_{j-1}) - S(y_j)$. Άρα, η πιθανοφάνεια των δεδομένων ως συνάρτηση μιας διακριτής συνάρτησης επιβίωσης S θα είναι:

$$L(S) = \prod_{j=1}^{D} \left\{ (S(y_{j-1}) - S(y_j))^{d_j} S(y_j)^{m_j} \right\}.$$

Θα κάνουμε αναπαραμέτρηση για να γράψουμε την πιθανοφάνεια ως συνάρτηση της διακριτής συνάρτησης κινδύνου (discrete hazard function) $\lambda_j = \Pr(T = y_j | T \geq y_j)$. Από τις σχέσεις (2.4) και (2.5) έχουμε ότι:

$$f(y_j) = \lambda_j \prod_{i=1}^{j-1}(1 - \lambda_i), \quad S(y_j) = \prod_{i=1}^{j}(1 - \lambda_i).$$

Επομένως, η πιθανοφάνεια, ως συνάρτηση του $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_D)$, μπορεί να γραφτεί ως εξής

$$\begin{aligned}
L(\boldsymbol{\lambda}) &= \prod_{j=1}^{D} \left\{ \lambda_j^{d_j} \prod_{i=1}^{j-1}(1-\lambda_i)^{d_j} \prod_{i=1}^{j}(1-\lambda_i)^{m_j} \right\} = \prod_{j=1}^{D} \left\{ \frac{\lambda_j^{d_j}}{(1-\lambda_j)^{d_j}} \prod_{i=1}^{j}(1-\lambda_i)^{d_j+m_j} \right\} \\
&= \prod_{j=1}^{D} \left\{ \frac{\lambda_j^{d_j}}{(1-\lambda_j)^{d_j}} \right\} \times \prod_{j=1}^{D}\prod_{i=1}^{j} \left\{ (1-\lambda_i)^{d_j+m_j} \right\} \\
&= \prod_{j=1}^{D} \left\{ \frac{\lambda_j^{d_j}}{(1-\lambda_j)^{d_j}} \right\} \times (1-\lambda_1)^{(d_1+m_1)+(d_2+m_2)+\ldots+(d_D+m_D)} \\
&\times (1-\lambda_2)^{(d_2+m_2)+(d_3+m_3)+\ldots+(d_D+m_D)} \times (1-\lambda_3)^{(d_3+m_3)+(d_4+m_4)+\ldots+(d_D+m_D)} \times \ldots \\
&= \prod_{j=1}^{D} \left\{ \frac{\lambda_j^{d_j}}{(1-\lambda_j)^{d_j}} \right\} \prod_{j=1}^{D} \left\{ (1-\lambda_j)^{\sum_{i=j}^{D}(d_i+m_i)} \right\} = \prod_{j=1}^{D} \left\{ \lambda_j^{d_j} (1-\lambda_j)^{r_j-d_j} \right\}.
\end{aligned}$$

Παρατηρούμε ότι καταλήγουμε στην πιθανοφάνεια από D ανεξάρτητες διωνυμικές κατανομές με r_j δοκιμές, d_j αποτυχίες και πιθανότητα αποτυχίας λ_j ($d_j \sim \text{Binomial}(r_j, \lambda_j)$, $j = 1, 2, \ldots, D$). Αν λογαριθμίσουμε την πιθανοφάνεια και θέσουμε τις μερικές παραγώγους ως προς λ_j ίσες με 0, βρίσκουμε εύκολα ότι ο εκτιμητής μέγιστης πιθανοφάνειας του λ_j, είναι $\hat{\lambda}_j = \dfrac{d_j}{r_j}$, δηλαδή η αναλογία των αποτυχιών οι οποίες έγιναν στο j-οστό διάστημα $[y_j, y_{j+1})$, σε σχεση με το πλήθος των ατόμων οι οποίοι βρίσκονταν σε κίνδυνο ακριβώς πριν την έναρξη του διαστήματος. Αντικαθιστώντας το $\hat{\lambda}_j$ στη συνάρτηση επιβίωσης, καταλήγουμε στον εκτιμητή που προτάθηκε απο τους Kaplan & Meier (1958):

$$\hat{S}_{KM}(t) = \prod_{j|y_j \leq t} \left(1 - \hat{\lambda}_j\right) = \prod_{j|y_j \leq t} \left(1 - \frac{d_j}{r_j}\right)$$

Ο εκτιμητής Kaplan–Meier είναι μια φθίνουσα κλιμακωτή συνάρτηση (step-function), η

οποία μειώνεται μόνο στους χρόνους στους οποίους παρατηρούμε αποτυχίες. Ένα προβληματικό σημείο είναι ότι ο εκτιμητής $\hat{S}_{KM}$ δεν φθάνει ποτέ το μηδέν αν υπάρχουν λογοκριμένοι χρόνοι οι οποίοι είναι μεγαλύτεροι ή ίσοι από το μεγαλύτερο χρόνο αποτυχίας (δηλαδή $m_D > 0$). Σε αυτήν την περίπτωση, ο μεγαλύτερος χρόνος παρακολούθησης είναι λογοκριμένος και είναι σύνηθες να μην ορίζουμε το $\hat{S}_{KM}(t)$ για $t > y_{Dm_D}$.

2.5 Μοντέλο Αναλογικών Κινδύνων

Ο εκτιμητής Kaplan–Meier μπορεί να χρησιμοποιηθεί μόνο στην περίπτωση ανεξάρτητων και ισόνομων χρόνων επιβίωσης, δηλαδή σε ομοιογενείς πληθυσμούς. Σε πολλά πρακτικά προβλήματα τα άτομα θα διαφέρουν σημαντικά ως προς την ηλικία, το φύλο, τις καπνιστικές συνήθειες και άλλους παράγοντες. Κάποιες από αυτές τις μεταβλητές μπορεί να είναι ειδικού ενδιαφέροντος, όπως η θεραπεία σε μια κλινική δοκιμή, ή συγχυτικοί παράγοντες οι οποίοι πρέπει να ληφθούν υπόψη. Έστω $\lambda(t|\mathbf{x})$ η συνάρτηση κινδύνου για ένα άτομο σε χρόνο t (από την αρχή του χρόνου), με ένα διάνυσμα επεξηγηματικών μεταβλητών $\mathbf{x}^T = (x_1, x_2, \ldots, x_p)$. Ένα μοντέλο αναλογικών κινδύνων υποθέτει ότι:

$$\lambda(t|\mathbf{x}) = \lambda_0(t)\psi(\mathbf{x}, \boldsymbol{\beta}),$$

όπου $\lambda_0(t)$ είναι μια βασική (baseline) συνάρτηση κινδύνου και $\psi(\cdot)$ μια θετική συνάρτηση η οποία εξαρτάται από τις επεξηματικές μεταβλητές και το αντίστοιχο διάνυσμα παραμέτρων $\boldsymbol{\beta}^T = (\beta_1, \beta_2, \ldots, \beta_p)$. Η συνήθης επιλογή είναι $\psi(\mathbf{x}, \boldsymbol{\beta}) = \exp\{\mathbf{x}^T\boldsymbol{\beta}\}$, άρα

$$\lambda(t|\mathbf{x}) = \lambda_0(t)\exp\{\mathbf{x}^T\boldsymbol{\beta}\}.$$

Η εκθετική μορφή εξασφαλίζει ότι η συνάρτηση κινδύνου είναι θετική. Το μοντέλο υποθέτει ότι υπάρχει μια βασική, χρονοεξαρτώμενη, συνάρτηση κινδύνου, κοινή για όλα τα άτομα. Οι επεξηγηματικές μεταβλητές δρουν πολλαπλασιαστικά στην βασική συνάρτηση κινδύνου, αυξάνοντας ή μειώνοντας τον κίνδυνο ανάλογα με την πληροφορία που είναι διαθέσιμη για κάθε άτομο. Το μοντέλο ξεχωρίζει από τη μια μεριά την επίδραση του χρόνου στην βασική συνάρτηση κινδύνου ($\lambda_0(t)$) και από την άλλη την επίδραση των επεξηγηματικών μεταβλητών ($\exp\{\mathbf{x}^T\boldsymbol{\beta}\}$). Το πηλίκο των κινδύνων σε κάθε χρονική στιγμή t ανάμεσα σε δυο άτομα με επεξηγηματικές μεταβλητές $\mathbf{x}_1$ και $\mathbf{x}_2$, αντίστοιχα, θα είναι πάντα ίσο με $\lambda(t|\mathbf{x}_1)/\lambda(t|\mathbf{x}_2) = \exp\{(\mathbf{x}_1 - \mathbf{x}_2)^T\boldsymbol{\beta}\}$, δηλαδή θα είναι ανεξάρτητο του χρόνου. Επομένως, η συνάρτηση κινδύνου μπορεί να έχει οποιαδήποτε μορφή, αλλά οι κίνδυνοι μεταξύ δυο διαφορετικών πληθυσμών θα είναι, σε κάθε χρονική στιγμή, ανάλογοι. Η βασική συνάρτηση κινδύνου $\lambda_0(t)$ μπορεί να θεωρηθεί ότι έχει συγκεκριμένη παραμετρική μορφή ή να αφεθεί απροσδιόριστη.

2.5.1 Ημι-παραμετρικό Μοντέλο Αναλογικών Κινδύνων

Μπορούμε να αφήσουμε τη μορφή του $\lambda_0(t)$ ως αδιευκρίνιστη. Επείδη με αυτή την υπόθεση η μόνη παράμετρος προς εκτίμηση είναι το $\boldsymbol{\beta}$, λέμε ότι το μοντέλο είναι ημι-παραμετρικό. Εφόσον τα $\lambda(t|\mathbf{x})$ και $S(t|\mathbf{x})$ είναι συναρτήσεις του $\lambda_0(t)$ η πιθανοφάνεια (2.6) δεν μπορεί να χρησιμοποιηθεί για δεξια λογοκριμένα δεδομένα. Χρειαζόμαστε μια τροποποίηση της πιθανοφάνειας η οποία δεν θα περιέχει το $\lambda_0(t)$, αλλά θα συγκρατεί επαρκή πληροφορία για να εκτιμήσουμε το $\boldsymbol{\beta}$ με συνέπεια. Ο Cox (1972) πρότεινε τη μέθοδο της μερική πιθανοφάνειας (partial likelihood) η οποία χρησιμοποιείται ευρέως μέχρι σήμερα. Η μερική πιθανοφάνεια μπορεί να εξαχθεί ως profile likelihood, δηλαδή θεωρώντας το $\boldsymbol{\beta}$ ως σταθερά, μεγιστοποιούμε την πιθανοφάνεια ως προς $\lambda_0(t)$ για να βρούμε εκτιμητές του $\lambda_0(t)$ οι οποίοι θα περιέχουν το $\boldsymbol{\beta}$, και στη συνέχεια αντικαθιστούμε στην πιθανοφάνεια τους εκτιμητές του $\lambda_0(t)$. Η τελική πιθανοφάνεια θα περιέχει μόνο το $\boldsymbol{\beta}$.

Έστω $y_{(1)} < y_{(2)} < \ldots < y_{(D)}$ οι παρατηρούμενοι χρόνοι αποτυχίας (θεωρώντας ότι είναι μοναδικοί) n ανεξάρτητων ατόμων με αντίστοιχες επεξηγηματικές μεταβλητές $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(D)}$. Οι χρόνοι παρακολούθησης και τα αντίστοιχα διανύσματα επεξηγηματικών μεταβλητών $(y_j, \mathbf{x}_j)$, $j = 1, 2, \ldots, n$, δεν είναι διατεταγμένα, όπως προηγουμένως, και δεν ταυτίζονται απαραίτητα με τα $(y_{(i)}, \mathbf{x}_{(i)}), i = 1, 2, \ldots, D$. Η συνάρτηση πιθανοφάνειας γράφεται ως εξής:

$$\prod_{i=1}^{D} \lambda(y_{(i)}|\mathbf{x}_{(i)}) \prod_{j=1}^{n} S(y_y|\mathbf{x}_j) = \prod_{i=1}^{D} \lambda_0(y_{(i)}) \exp\left\{\mathbf{x}_{(i)}^T \boldsymbol{\beta}\right\} \prod_{j=1}^{n} \exp\left\{-\Lambda_0(y_j) \exp\{\mathbf{x}_j^T \boldsymbol{\beta}\}\right\}. \quad (2.8)$$

Θα θεωρήσουμε τη διακριτή περίπτωση του αθροιστικού κινδύνου, υπονοώντας ότι η συνάρτηση κινδύνου είναι μηδέν παντού εκτός από τους παρατηρούμενους χρόνους αποτυχίας, άρα $\Lambda_0(y_j) = \sum_{i|y_{(i)} \le y_j} \lambda_0(y_{(i)})$. Αν αναδιατάξουμε τους δείκτες των ατόμων έχουμε ότι

$$\begin{aligned} q &= \prod_{j=1}^{n} \exp\left\{-\Lambda_0(y_j) \exp\{\mathbf{x}_j^T \boldsymbol{\beta}\}\right\} = \prod_{j=1}^{n} \exp\left\{-\sum_{i|y_{(i)} \le y_j} \lambda_0(y_{(i)}) \exp\{\mathbf{x}_j^T \boldsymbol{\beta}\}\right\} \\ &= \exp\left\{-\lambda_0(y_{(1)}) \sum_{j|y_j \ge y_{(1)}} \exp\{\mathbf{x}_j^T \boldsymbol{\beta}\} - \ldots - \lambda_0(y_{(D)}) \sum_{j|y_j \ge y_{(D)}} \exp\{\mathbf{x}_j^T \boldsymbol{\beta}\}\right\} \\ &= \exp\left\{-\sum_{i=1}^{D} \lambda_0(y_{(i)}) \sum_{j \in R(y_{(i)})} \exp\{\mathbf{x}_j^T \boldsymbol{\beta}\}\right\}, \end{aligned} \quad (2.9)$$

όπου με $R(y_{(i)})$ συμβολίζουμε το σύνολο κινδύνου (risk set) στο χρόνο $y_{(i)}$, το οποίο περιλαμβάνει όλα τα άτομα (στην πραγματικότητα τους δείκτες των ατόμων) τα οποία ήταν σε κίνδυνο να αναπτύξουν το γεγονός τη χρονική στιγμή $y_{(i)}$ (συμπεριλαμβανομένου και αυτού που απέτυχε). Επίσης, τα άτομα τα οποία λογοκρίθηκαν τη χρονική στιγμή $y_{(i)}$, θα θεωρήσουμε ότι ανήκουν στο $R(y_{(i)})$. Επομένως, η πιθανοφάνεια (2.8) ως συνάρτηση του $\lambda_0(y_{(i)})$, για $i = 1, 2, \ldots, D$,

με το $\boldsymbol{\beta}$ να θεωρείται ως σταθερά, μπορεί να γραφτεί μέσω της (2.9) ως εξής:

$$L(\boldsymbol{\lambda}_0|\boldsymbol{\beta}) = \prod_{i=1}^{D} \lambda_0(y_{(i)}) \exp\{\mathbf{x}_{(i)}^T \boldsymbol{\beta}\} \times \exp\left\{ -\sum_{i=1}^{D} \lambda_0(y_{(i)}) \sum_{j \in R(y_{(i)})} \exp\{\mathbf{x}_j^T \boldsymbol{\beta}\} \right\}. \tag{2.10}$$

Αν λογαριθμίσουμε την πιθανοφάνεια (2.10) ως προς $\lambda_0(y_{(i)})$, θεωρώντας το $\boldsymbol{\beta}$ ως σταθερά, βρίσκουμε ότι (Duchateau & Janssen 2007, σελ. 25):

$$\lambda_0(y_{(i)}) = \frac{1}{\sum_{j \in R(y_{(i)})} \exp\{\mathbf{x}_j^T \boldsymbol{\beta}\}}, \ i = 1, 2, \ldots, D.$$

Αντικαθιστώντας τη λύση στην σχέση (2.10), καταλήγουμε στην μερική πιθανοφάνεια του Cox (ο όρος e^{-D} παραλείπεται)

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp\{\mathbf{x}_{(i)}^T \boldsymbol{\beta}\}}{\sum_{j \in R(y_{(i)})} \exp\{\mathbf{x}_j^T \boldsymbol{\beta}\}},$$

η οποία προτείνεται να χρησιμοποιείται ως μια συνηθισμένη πιθανοφάνεια, δηλαδή μπορούμε να εκτιμήσουμε το $\boldsymbol{\beta}$ μεγιστοποιώντας την $L(\boldsymbol{\beta})$. Οι στατιστικές ιδιότητες (συνέπεια, ασυμπτωτική κανονικότητα) της μερικής πιθανοφάνειας έχουν αναλυθεί εκτενώς (Gill 1984, Fleming & Harrington 1991). Είναι πιο βολικό να χρησιμοποιήσουμε για τη μεγιστοποίηση τη λογαριθμισμένη πιθανοφάνεια.

$$\ell(\boldsymbol{\beta}) = \log(L(\boldsymbol{\beta})) = \sum_{i=1}^{D} \left[\mathbf{x}_{(i)}^T \boldsymbol{\beta} - \log\left(\sum_{j \in R(y_{(i)})} \exp\{\mathbf{x}_j^T \boldsymbol{\beta}\} \right) \right].$$

Θέλουμε να βρούμε την παράγωγο του $\ell(\boldsymbol{\beta})$ ως προς $\boldsymbol{\beta}$, $\mathcal{U}(\boldsymbol{\beta}) = \dfrac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ (score vector). Αρχικά θα παραγωγίσουμε ως προς την k συνιστώσα του διανύσματος $\boldsymbol{\beta}$, $k = 1, 2 \ldots, p$.

$$\mathcal{U}_k(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_k} = \sum_{i=1}^{D} \left[x_{(i)k} - \frac{\sum_{j \in R(y_{(i)})} x_{jk} \exp\{\mathbf{x}_j^T \boldsymbol{\beta}\}}{\sum_{l \in R(y_{(i)})} \exp\{\mathbf{x}_l^T \boldsymbol{\beta}\}} \right] = \sum_{i=1}^{D} \left[x_{(i)k} - \bar{x}_{(i)k} \right] \ k = 1, 2, \ldots, p$$

όπου $\bar{x}_{(i)k} = \sum_{j \in R(y_{(i)})} x_{jk} \omega_{(i)j}(\boldsymbol{\beta})$ είναι ένας σταθμισμένος μέσος της μεταβλητής X_k ανάμεσα σε όλα τα άτομα του συνόλου κινδύνου του χρόνου $y_{(i)}$, με βάρη:

$$\omega_{(i)j}(\boldsymbol{\beta}) = \frac{\exp\{\mathbf{x}_j^T \boldsymbol{\beta}\}}{\sum_{l \in R(y_{(i)})} \exp\{\mathbf{x}_l^T \boldsymbol{\beta}\}}, \ i = 1, 2, \ldots, D, \text{ και } j \in R(y_{(i)}),$$

επομένως, το $\mathcal{U}(\boldsymbol{\beta})$ (score vector) μπορεί να γραφτεί πιο απλά ως $\mathcal{U}(\boldsymbol{\beta}) = \sum_{i=1}^{D} \left[\mathbf{x}_{(i)} - \bar{\mathbf{x}}_{(i)} \right]$ με $\bar{\mathbf{x}}_{(i)} = (\bar{x}_{(i)1}, \bar{x}_{(i)2}, \ldots, \bar{x}_{(i)p})^T$.

Οι εκτιμητές μέγιστης πιθανοφάνειας μπορούν να βρέθουν λύνοντας το συστημα $\mathcal{U}(\boldsymbol{\beta}) = \mathbf{0}$. Για να εκτιμήσουμε τη διακύμανση χρειαζόμαστε τον πίνακα πληροφορίας της μερικής πιθανοφάνειας $\mathcal{I}(\boldsymbol{\beta}) = -\dfrac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}$. Ισχύει ότι

$$\begin{aligned}
\mathcal{I}_{kh}(\boldsymbol{\beta}) = -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_k \partial \beta_h} &= \sum_{i=1}^{D} \sum_{j \in R(y_{(i)})} x_{jk} \frac{\partial}{\partial \beta_h} \left(\frac{\exp\{\mathbf{x}_j^T \boldsymbol{\beta}\}}{\sum_{l \in R(y_{(i)})} \exp\{\mathbf{x}_l^T \boldsymbol{\beta}\}} \right) \\
&= \sum_{i=1}^{D} \sum_{j \in R(y_{(i)})} x_{jk} \frac{x_{jh} e^{\mathbf{x}_j^T \boldsymbol{\beta}} \sum_{l \in R(y_{(i)})} e^{\mathbf{x}_l^T \boldsymbol{\beta}} - e^{\mathbf{x}_j^T \boldsymbol{\beta}} \sum_{l \in R(y_{(i)})} x_{lh} e^{\mathbf{x}_l^T \boldsymbol{\beta}}}{(\sum_{l \in R(y_{(i)})} e^{\mathbf{x}_l^T \boldsymbol{\beta}})^2} \\
&= \sum_{i=1}^{D} \sum_{j \in R(y_{(i)})} x_{jk} x_{jh} \frac{e^{\mathbf{x}_j^T \boldsymbol{\beta}}}{\sum_{l \in R(y_{(i)})} e^{\mathbf{x}_l^T \boldsymbol{\beta}}} \\
&- \sum_{i=1}^{D} \left\{ \sum_{j \in R(y_{(i)})} x_{jk} \frac{e^{\mathbf{x}_j^T \boldsymbol{\beta}}}{\sum_{l \in R(y_{(i)})} e^{\mathbf{x}_l^T \boldsymbol{\beta}}} \times \sum_{l \in R(y_{(i)})} x_{lh} \frac{e^{\mathbf{x}_l^T \boldsymbol{\beta}}}{\sum_{l \in R(y_{(i)})} e^{\mathbf{x}_l^T \boldsymbol{\beta}}} \right\} \\
&= \sum_{i=1}^{D} \sum_{j \in R(y_{(i)})} x_{jk} x_{jh} \omega_{(i)j}(\boldsymbol{\beta}) - \sum_{i=1}^{D} \bar{x}_{(i)k} \bar{x}_{(i)h}.
\end{aligned}$$

Αν συμβολίσουμε ως $\mathbf{X}_{R(y_{(i)})} = \begin{pmatrix} \vdots \\ \mathbf{x}_j^T \\ \vdots \end{pmatrix}_{j \in R(y_{(i)})}$ και $\bar{\mathbf{X}}_{(\cdot)} = \begin{pmatrix} \bar{\mathbf{x}}_{(1)}^T \\ \vdots \\ \bar{\mathbf{x}}_{(D)}^T \end{pmatrix}_{D \times p}$ τον πίνακα σχεδιασμού για τα άτομα που ανήκουν στο σύνολο κινδύνου τη χρονική στιγμή $y_{(i)}$ και τον πίνακα με τους σταθμισμένους μέσους των μεταβλητών ανά σύνολο κινδύνου, αντίστοιχα, τότε χρησιμοποιώντας άλγεβρα πινάκων μπορούμε να δούμε ότι ο πίνακας πληροφορίας γράφεται ως εξής:

$$\mathcal{I}(\boldsymbol{\beta}) = -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \sum_{i=1}^{D} \mathbf{X}_{R(y_{(i)})}^T \mathbf{W}_{(i)} \mathbf{X}_{R(y_{(i)})} - \bar{\mathbf{X}}_{(\cdot)}^T \bar{\mathbf{X}}_{(\cdot)},$$

όπου $\mathbf{W}_{(i)} = \mathrm{diag}\{\omega_{(i)j}(\boldsymbol{\beta}) | j \in R(y_{(i)})\}$ ο διαγώνιος πίνακας βαρών για το i-οστο σύνολο κινδύνου. Οι εξισώσεις της μερικής πιθανοφάνειας δεν μπορούν να λυθούν αναλυτικά. Ενας αριθμητικός τρόπος επίλυσης ο οποίος χρησιμοποιείται συνήθως από τα στατιστικά πακέτα είναι ο αλγόριθμος Newton-Raphson. Ξεκινώντας από μια πρόχειρη αλλά βολική τιμή ανανεώνουμε επαναληπτικά την εκτίμηση για το $\boldsymbol{\beta}$. Αν η τρέχουσα τιμή του αλγορίθμου είναι η $\boldsymbol{\beta}^i$, τότε το επόμενο βήμα θα είναι:

$$\boldsymbol{\beta}^{i+1} = \boldsymbol{\beta}^i + \mathcal{I}(\boldsymbol{\beta}^i)^{-1} \mathcal{U}(\boldsymbol{\beta}^i).$$

Όταν η διαφορά μεταξύ δυο διαχοδικών τιμών για το $\boldsymbol{\beta}$ είναι μικρότερη από κάποιο προαποφασμένο μικρό (κατ΄ απόλυτη τιμή) αριθμό, θεωρούμε ότι ο αλγόριθμος έχει βρει την κορυφή της πιθανοφάνειας.

2.6 Μονοδιάστατα Frailty Μοντέλα

Στην παρούσα ενότητα θα ασχοληθούμε με την ανάλυση μονοδιάστατων δεδομένων, για παράδειγμα, χρόνων επιβίωσης ανεξάρτητων ατόμων. Μια συχνή προσέγγιση είναι να θεωρήσουμε ότι ο υπό μελέτη πληθυσμός, δοθέντων των επεξηγηματικών μεταβλητών, είναι ομοιογενής. Όμως, έχει παρατηρηθεί ότι μια θεραπεία ή ένα φάρμακο μπορεί να επιδρά διαφορετικά σε άτομα με ίδιες τιμές των γνωστών παραγόντων κινδύνου. Αυτού του είδους η ετερογένεια είναι δύσκολο να αποδοθεί αποκλειστικά σε παρατηρήσιμους παράγοντες κινδύνου, αφού είναι σχεδόν αδύνατο να συμπεριλάβουμε όλους τους παράγοντες κινδύνου σε ένα μοντέλο. Συχνά, οι ερευνητές δεν έχουν γνώση όλων των χαρακτηριστικών του ατόμου τα οποία συνθέτουν την ευπάθεια του ασθενούς (γενετικοί παράγοντες ή προσωπικές συνήθειες).

Η βασική ιδέα των frailty μοντέλων είναι ότι κάθε άτομο έχει κάποια προσωπικά χαρακτηριστικά τα οποία επηρεάζουν την επιβίωση του και δεν είναι παρατηρήσιμα, επομένως, οι πιο ευπαθείς (frail) θα τείνουν να πεθαίνουν νωρίτερα από τους λιγότερο ευπαθείς. Άρα, όσο περνά ο χρόνος γίνεται συστηματική επιλογή των πιο ανθεκτικών ατόμων. Σε ορισμένες περιπτώσεις ενδιαφερόμαστε για τον τρόπο με τον οποίο μεταβάλλεται η θνησιμότητα με την πάροδο του χρόνου. Συχνά, ειδικά σε ασθενείς με καρκίνο, ο κίνδυνος στην αρχή του χρόνου παρακολούθησης αυξάνεται, φθάνει στην μέγιστη τιμή του, και στη συνέχεια μειώνεται ή εξισορροπείται (μονοκόρυφη συνάρτηση κινδύνου). Αυτή η μορφή συνάρτησης κινδύνου μπορεί να βρεθεί σε ασθενείς με κοινές τιμές παραγόντων κινδύνου. Μια ερμηνεία που μπορεί να δοθεί είναι ότι όσο περνά ο χρόνος από τη διάγνωση ή τη θεραπεία, τόσο μειώνεται ο κίνδυνος του ατόμου. Είναι όμως αμφίβολο αν η μορφή της πληθυσμιακής συνάρτησης κινδύνου αντανακλάται και σε ατομικό επίπεδο. Είναι πιθανό η πληθυσμιακή συνάρτηση κινδύνου να μειώνεται γιατί οι ασθενείς υψηλού κινδύνου έχουν ήδη πεθάνει, αλλά ο κίνδυνος ενός μεμονωμένου ατόμου μπορεί να συνεχίζει να αυξάνεται (Wienke 2010, σελ. 55). Για να λάβουμε υπόψη την επιλογή των πιο ανθεκτικών ατόμων μπορούμε να χρησιμοποιήσουμε μεικτά μοντέλα. Θεωρούμε ότι ο υπό μελέτη πληθυσμός είναι μια μείξη από άτομα με μερικώς άγνωστους κινδύνους. Τα προσωπικά χαρακτηριστικά του κάθε ατόμου ενσωματώνονται σε μια τυχαία μεταβλητή η οποία ονομάζεται *frailty* στην ανάλυση επιβίωσης. Η σχέση μεταξύ ατομικού και πληθυσμιακού κινδύνου εξαρτάται από την κατανομή των frailties ανάμεσα στα άτομα. Οι Beard (1959), Vaupel, Manton & Stallard (1979), Lancaster (1979) πρότειναν ένα μοντέλο τυχαίων επιδράσεων για να αξιολογήσουν την ετερογένεια λόγω μη παρατηρούμενων επεξηγηματικών μεταβλητών. Οι Vaupel et al. (1979)

εισήγαγαν τον όρο frailty στην βιοστατιστική, εφαρμόζοντας το μοντέλο σε πληθυσμιακά δεδομένα επιβίωσης. Τα frailty μοντέλα λαμβάνουν υπόψη τη μη παρατηρούμενη ετερογένεια, η οποία υπάρχει επειδή κάποια άτομα είναι περισσότερο επιρρεπή σε αποτυχία σε σχέση με άλλα.

Στα frailty μοντέλα η μεταβλητότητα των χρόνων επιβίωσης χωρίζεται σε δυο μέρη. Το ένα μέρος αποδίδεται σε γνωστούς παράγοντες κινδύνου και το άλλο σε μη παρατηρούμενους παράγοντες (frailty). Τα μονοδιάστατα frailty μοντέλα παρουσιάζουν τον πληθυσμό σαν μια μείξη όπου η βασική συνάρτηση κινδύνου είναι κοινή για όλα τα άτομα, αλλά κάθε άτομο έχει το δικό του frailty. Η συνάρτηση κινδύνου εξαρτάται από μια μη αρνητική τυχαία μεταβλητή u η οποία, για λόγους αναγνωρισιμότητας, έχει μέση τιμή $\mathrm{E}(u) = 1$ και διακύμανση $\mathrm{Var}(u) = \sigma^2$ (αν υπάρχει). Η συνάρτηση κινδύνου θα είναι η εξής:

$$\lambda(t|\mathbf{x}_i, u_i) = \lambda_0(t) u_i \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\},\ i = 1, 2, \ldots, n,$$

όπου $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \ldots, x_{ip})$ είναι το διάνυσμα επεξηγηματικών μεταβλητών για το i-οστο άτομο και $\boldsymbol{\beta}$ είναι το αντίστοιχο διάνυσμα παραμέτρων. Η διακύμανση των frailties, σ^2, ερμηνεύεται ως ένα μέτρο της ετερογένειας στο βασικό κίνδυνο (baseline hazard) μεταξύ των ατόμων του πληθυσμού. Όταν το σ^2 είναι μικρό, οι τιμές του u είναι συγκεντρωμένες στο 1. Αντίθετα, όταν το σ^2 είναι μεγάλο, οι τιμές του u είναι διεσπαρμένες, με συνέπεια μεγάλη ετερογένεια των ατομικών κινδύνων. Η δεσμευμένη πιθανοφάνεια (παρόμοια με 2.6) των δεδομένων $(Y_i, \Delta_i, \mathbf{x}_i, u_i), i = 1, 2, \ldots, n$, θα είναι

$$L = \prod_{i=1}^{n} \left(u_i \lambda_0(y_i) \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}\right)^{\delta_i} \exp\left\{-u_i \Lambda_0(y_i) \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}\right\}, \tag{2.11}$$

αλλά τώρα δοθέντων των τιμών $u_1, u_2, \ldots, u_n$. Έστω $S(t|\mathbf{x}, u)$ η συνάρτηση επιβίωσης ενός ατόμου με χαρακτηριστικά $\mathbf{x}$ δοθέντος του u:

$$S(t|\mathbf{x}, u) = \exp\left\{-\int_0^\infty \lambda(s|\mathbf{x}, u) ds\right\} = \exp\left\{-u \Lambda_0(t) \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}\right\}.$$

Μέχρι στιγμής το μοντέλο έχει περιγραφεί σε ατομικό επίπεδο. Σε ένα frailty μοντέλο, χρειάζεται να κάνουμε διάκριση μεταξύ ατομικού και πληθυσμιακού επιπέδου. Δυο άτομα με τα ίδια (παρατηρήσιμα) χαρακτηριστικά θα έχουν, πιθανώς, διαφορετικούς χρόνους αλλά και οι συναρτήσεις επιβίωσης τους μπορεί να είναι διαφορετικές. Το ένα άτομο μπορεί να είναι πιο ευπαθές σε σχέση με το άλλο λόγω μη παρατηρούμενων παραγόντων, εξηγώντας έτσι τις ατομικές διαφορές στη συνάρτηση κινδύνου. Οι μη παρατηρούμενοι παράγοντες οδηγούν σε μεγαλύτερη μεταβλητότητα των χρόνων επιβίωσης από ότι θα αναμέναμε κάτω από ένα μοντέλο σταθερών επιδράσεων. Αντίθετα, σε ένα μοντέλο χωρίς frailty, δυο άτομα με τα ίδια (παρατηρήσιμα) χαρακτηριστικά θα έχουν διαφορετικούς χρόνους, γιατί ο χρόνος επιβίωσης τους είναι μια τυχαία

μεταβλητή και ακολουθεί κάποια κατανομή. Όμως, δεδομένα για το ατομικό επίπεδο του κάθε ασθενούς δεν είναι παρατηρήσιμα. Επομένως, θα μπορούσαμε να μιλήσουμε για την πληθυσμιακή συνάρτηση επιβίωσης όπου η frailty μεταβλητή έχει ολοκληρωθεί. Η πληθυσμιακή συνάρτηση επιβίωσης είναι ένας σταθμισμένος μέσος όρος των ατομικών συνάρτησεων επιβίωσης με βάρη σύμφωνα με την συνάρτηση πυκνότητας της κατανομής των frailties. Μπορεί να θεωρηθεί ως η συνάρτηση επιβίωσης ενός τυχαία επιλεγόμενου ατόμου από τον υπο μελέτη πληθυσμό και αντιστοιχεί σε ότι μπορεί να παρατηρηθεί (Wienke 2010, σελ. 59).

$$S(t|\mathbf{x}) = \mathrm{E}_u\left(S(t|\mathbf{x},u)\right) = \int_0^\infty e^{-u\Lambda_0(t)\exp\{\mathbf{x}^T\boldsymbol{\beta}\}} f_U(u)du = \mathbf{L}\left(\Lambda_0(t)\exp\{\mathbf{x}^T\boldsymbol{\beta}\}\right), \qquad (2.12)$$

όπου με $\mathbf{L}$ συμβολίζουμε το μετασχηματισμό Laplace. Η πληθυσμιακή συνάρτηση κινδύνου θα είναι:

$$\lambda(t|\mathbf{x}) = \frac{f(t|\mathbf{x})}{S(t|\mathbf{x})} = -\lambda_0(t)\exp\{\mathbf{x}^T\boldsymbol{\beta}\}\frac{\mathbf{L}'\left(\Lambda_0(t)\exp\{\mathbf{x}^T\boldsymbol{\beta}\}\right)}{\mathbf{L}\left(\Lambda_0(t)\exp\{\mathbf{x}^T\boldsymbol{\beta}\}\right)}. \qquad (2.13)$$

Επίσης, μπορεί να αποδειχθεί (Vaupel et al. 1979) ότι ο πλυθυσμιακός κίνδυνος στο χρόνο t είναι ένας σταθμισμένος μέσος όρος των ατομικών κινδύνων των ατόμων οι οποίοι είναι σε κίνδυνο να αναπτύξουν το γεγονός στο χρόνο t, δηλαδή

$$\lambda(t|\mathbf{x}) = \mathrm{E}_u(\lambda(t|\mathbf{x},u)|T>t) = \lambda_0(t)e^{\mathbf{x}^T\boldsymbol{\beta}}\int_0^\infty uf(u|T>t,\mathbf{x})du = \lambda_0(t)e^{\mathbf{x}^T\boldsymbol{\beta}}\,\mathrm{E}\left(u|T>t,\mathbf{x}\right).$$

Μπορούμε να χρησιμοποιήσουμε οποιαδήποτε κατανομή για τα frailties με μέση τιμή 1 και θετικό στήριγμα. Οι κατανομές με γνωστό μετασχηματισμό Laplace χρησιμοποιούνται πιο συχνά στην πράξη. Η κατανομη των frailties εκφράζει το πως κατανέμονται τα frailties στον πληθυσμό στην αρχή του χρόνου παρακολούθησης. Επίσης, θεωρούμε ότι η τιμή των frailties για κάθε άτομο παραμένει σταθερή. Όμως, ο υπο μελέτη πληθυσμός όσο περνάει ο χρόνος αλλάζει. Τα πιο ευπαθή άτομα, τα οποία θα έχουν τις μεγάλες τιμές των frailties, θα τείνουν να πεθαίνουν νωρίτερα, αφήνοντας στον υπο κίνδυνο πληθυσμό τα πιο ανθεκτικά άτομα. Επομένως, η κατανομή των frailties αλλάζει με την πάροδο του χρόνου, και τα σχετικά βάρη για μεγάλες τιμές των frailties (τα οποία συσχετίζονται με μεγάλη θνησιμότητα) γίνονται μικρότερα όσο περνά ο χρόνος. Είναι λογικό ότι η αναμενόμενη τιμή των frailties των επιζώντων τη χρονική στιγμή t, $\mathrm{E}\left(u|T>t,\mathbf{x}\right) = \int_0^\infty uf(u|T>t,\mathbf{x})du$, θα είναι φθίνουσα συνάρτηση ως προς το χρόνο.

2.6.1 Gamma Frailty Μοντέλο

Η Gamma κατανομή έχει χρησιμοποιηθεί ευρέως ως κατανομή για τα frailties (Hougaard 2000, Duchateau & Janssen 2007). Έχει κλειστό τύπο για τον μετασχηματισμό Laplace, επομένως, μπορούμε να βρούμε εύκολα τις σχέσεις για την πληθυσμιακή συνάρτηση κινδύνου και

επιβίωσης. Επίσης, παρουσιάζει υπολογιστικές ευκολίες κάτω από την Μπεϋζιανή προσέγγιση στη στατιστική, λόγω της δεσμευμένης συζυγίας για τα frailties. Η τυχαία μεταβλητή u ακολουθει την Gamma(κ, λ) κατανομή όταν

$$f(u) = \frac{\lambda^\kappa}{\Gamma(\kappa)} u^{\kappa-1} e^{-\lambda u},\ u > 0, \quad \mathrm{E}(u) = \kappa/\lambda,\ \mathrm{Var}(u) = \kappa/\lambda^2.$$

Από το γράφημα (2.4) βλέπουμε ότι η συνάρτηση πυκνότητας της Gamma κατανομής παίρνει διάφορες μορφές. Όσο η διακύμανση μικραίνει, τόσο οι τιμές είναι πιο συγκεντρωμένες στο 1. Για τον μετασχηματισμό Laplace ισχύει ότι:

$$\mathbf{L}(s) = \int_0^\infty e^{-su} f(u) du = \frac{\lambda^\kappa}{\Gamma(\kappa)} \int_0^\infty u^{\kappa-1} e^{-(\lambda+s)u} du = \frac{\lambda^\kappa}{\Gamma(\kappa)} \frac{\Gamma(\kappa)}{(\lambda+s)^\kappa} = (1 + \frac{s}{\lambda})^{-\kappa},$$

όπου έχουμε εκμεταλευτεί το γεγονός ότι το ολοκλήρωμα της Gamma$(\kappa, \lambda + s)$ είναι 1. Για να είμαστε σίγουροι ότι το μοντέλο είναι αναγνωρίσιμο πρέπει να θέσουμε τον περιορισμό $\mathrm{E}(u) = 1$, το οποίο έχει ως συνέπεια ότι $\kappa = \lambda$. Συμβολίζουμε ως $\sigma^2 = \mathrm{Var}(u) = \frac{1}{\kappa}$ την διακύμανση των frailties. Η συνάρτηση πυκνότητας πιθανότητας μιας Gamma$(\frac{1}{\sigma^2}, \frac{1}{\sigma^2})$ είναι:

$$f(u) = \frac{(1/\sigma^2)^{1/\sigma^2}}{\Gamma(1/\sigma^2)} u^{1/\sigma^2-1} e^{-\frac{1}{\sigma^2}u},\ u > 0.$$

Η πλυθησμιακή συνάρτηση επιβίωσης μπορεί να βρεθεί μέσω του μετασχηματισμού Laplace χρησιμοποιώντας την σχέση (2.12) ως εξης:

$$S(t|\mathbf{x}) = \mathbf{L}\left(\Lambda_0(t) e^{\mathbf{x}^T \boldsymbol{\beta}}\right) = (1 + \sigma^2 \Lambda_0(t) e^{\mathbf{x}^T \boldsymbol{\beta}})^{-\frac{1}{\sigma^2}}.$$

Αντίστοιχα η πληθυσμιακή συνάρτηση κινδυνου θα είναι

$$\lambda(t|\mathbf{x}) = \frac{f(t|\mathbf{x})}{S(t|\mathbf{x})} = -\frac{\partial \log(S(t|\mathbf{x}))}{\partial t} = \frac{\lambda_0(t) e^{\mathbf{x}^T \boldsymbol{\beta}}}{1 + \sigma^2 \Lambda_0(t) e^{\mathbf{x}^T \boldsymbol{\beta}}},$$

όπου χρησιμοποιήσαμε ότι $f(t|\mathbf{x}) = -\frac{\partial S(t|\mathbf{x})}{\partial t}$. Όσο περνά ο χρόνος, τα πιο ευπαθή μέρη του πληθυσμού θα αποτυγχάνουν και θα φεύγουν από τη μελέτη, άρα, η κατανόμη των frailties στον πληθυσμό σε κίνδυνο θα αλλάζει. Θα δείξουμε ότι η υπόθεση ότι τα frailties στον υπο μελέτη πληθυσμό στην αρχή του χρόνου παρακολούθησης κατανέμονται σύμφωνα με την Gamma κατανομή, έχει ως συνέπεια ότι η κατανομή των frailties στον πληθυσμό που είναι σε κίνδυνο κάποια χρονική στιγμή, t, είναι πάλι Gamma αλλά με ανανεωμένες παραμέτρους. Ισχύει ότι:

$$f(u, T > t|\mathbf{x}) = \int_t^\infty f(s, u|\mathbf{x}) ds = \int_t^\infty f(s|u, \mathbf{x}) ds f(u|\mathbf{x}) = S(t|u, \mathbf{x}) f(u).$$

Η συνάρτηση πυκνότητας πιθανότητας της κατανομής των frailties ανάμεσα στα άτομα τα οποία

τη χρονική στιγμή t βρίσκονται στη μελέτη θα είναι (Wienke 2010, σελ. 74):

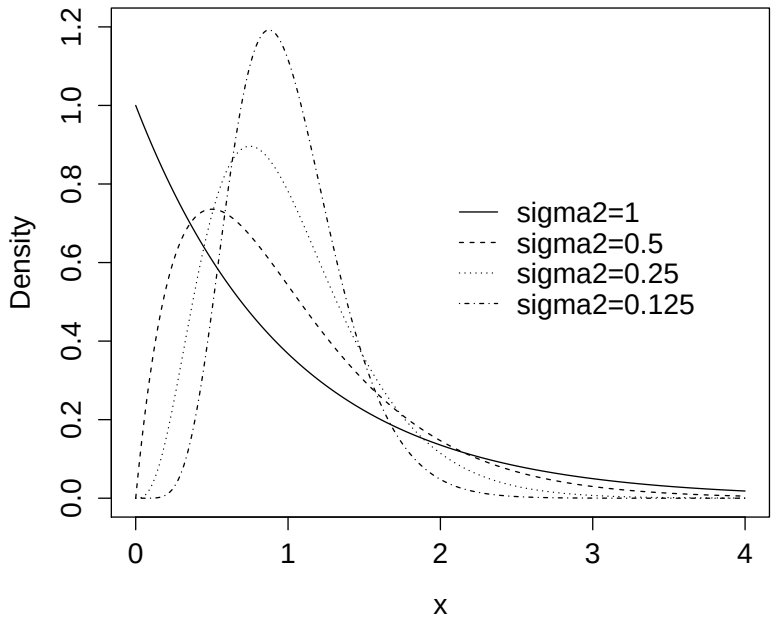
$$f(u|T>t,\mathbf{x}) = \frac{f(u,T>t|\mathbf{x})}{S(t|\mathbf{x})} \propto S(t|u,\mathbf{x})f(u) \propto u^{1/\sigma^2-1}\exp\left\{-\left(\frac{1}{\sigma^2}+\Lambda_0(t)e^{\mathbf{x}^T\boldsymbol{\beta}}\right)u\right\}.$$

Αναγνωρίζουμε τη δεσμευμένη κατανομή του $u|T>t,\mathbf{x}$ ως Gamma$(\frac{1}{\sigma^2}, \frac{1}{\sigma^2}+\Lambda_0(t)e^{\mathbf{x}^T\boldsymbol{\beta}})$. Με παρόμοιο τρόπο, μπορούμε να βρούμε την frailty κατανομή ανάμεσα στα άτομα τα οποία απέτυχαν τη χρονική στιγμή t.

$$f(u|T=t,\mathbf{x}) = \frac{f(t|u,\mathbf{x})f(u|\mathbf{x})}{f(t|\mathbf{x})} \propto f(t|u,\mathbf{x})f(u) \propto u^{1/\sigma^2+1-1}\exp\left\{-\left(\frac{1}{\sigma^2}+\Lambda_0(t)e^{\mathbf{x}^T\boldsymbol{\beta}}\right)u\right\}.$$

Μπορούμε να αναγνωρίσουμε τη δεσμευμένη κατανομή του $u|T=t,\mathbf{x}$ ως Gamma$(\frac{1}{\sigma^2}+1, \frac{1}{\sigma^2}+\Lambda_0(t)e^{\mathbf{x}^T\boldsymbol{\beta}})$. Οι αναμενόμενες τιμές των frailties για τα άτομα που επέζησαν στο χρόνο t και εκείνα που πέθαναν στο χρόνο t θα είναι, αντίστοιχα:

$$\mathrm{E}(u|T>t,\mathbf{x}) = \frac{1}{1+\sigma^2\Lambda_0(t)e^{\mathbf{x}^T\boldsymbol{\beta}}}, \quad \mathrm{E}(u|T=t,\mathbf{x}) = \frac{1+\sigma^2}{1+\sigma^2\Lambda_0(t)e^{\mathbf{x}^T\boldsymbol{\beta}}}.$$



Σχήμα 2.4: Συναρτήσεις πυκνότητας πιθανότητας της Gamma κατανομής με μέση τιμή 1 και διακυμάνσεις 1,0.5,0.25 και 0.125.

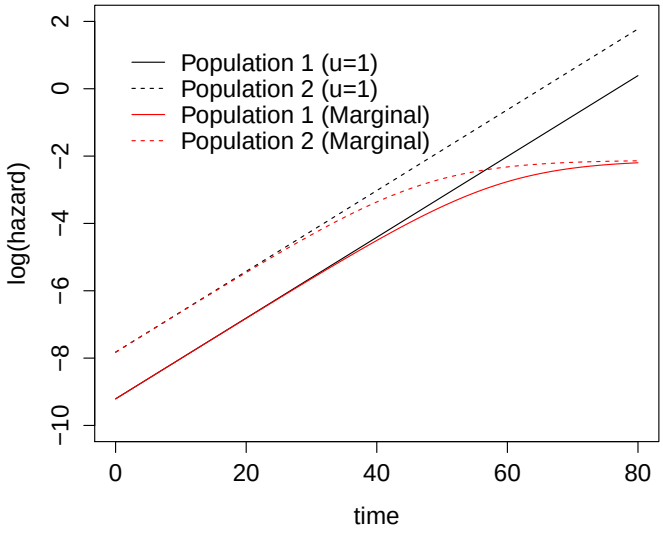
Παρατηρούμε ότι τα άτομα τα οποία απέτυχαν τη χρονική στιγμά t έχουν, κατά μέσο όρο, μεγαλύτερη frailty τιμή σε σχέση με τα άτομα που επέζησαν τη στιγμή t. Επιπλέον, η μέση τιμή των frailties για τα άτομα υπό-κίνδυνο είναι φθίνουσα συνάρτηση του χρόνου. Η μείωση είναι

μεγαλύτερη στις περιπτώσεις όπου η ετερογένεια του πλυθυσμού είναι μεγάλη (σ^2 μεγάλο) ή ο αθροιστικός κίνδυνος, $\Lambda_0(t)e^{\mathbf{x}^T\boldsymbol{\beta}}$, είναι μεγάλος. Επίσης, η διακύμανση ανάμεσα στα άτομα με χρόνο επιβίωσης μεγαλύτερο του t είναι:

$$\mathrm{Var}(u|T > t, \mathbf{x}) = \frac{\sigma^2}{\left(1 + \sigma^2 \Lambda_0(t) e^{\mathbf{x}^T\boldsymbol{\beta}}\right)^2}.$$

Επομένως, η διακύμανση των frailties μειώνεται με το χρόνο, με συνέπεια ο πληθυσμός να γίνεται, σε απόλυτο βαθμό, πιο ομοιογενής. Όμως, ο συντελεστής μεταβλητότητας δεν αλλάζει, άρα, η διακύμανση δεν μειώνεται σε σύγκριση με τη μέση τιμή. Σε ένα frailty μοντέλο είναι σημαντικό να παρατηρήσουμε ότι η αναλογικότητα των κινδύνων μεταξύ δυο ατόμων ισχύει μόνο δοθέντος της ίδιας τιμής του frailty. Για απλότητα, ας θεωρήσουμε μια δίτιμη μεταβλητή x με τιμές $x = 1$ (treatment) και $x = 0$ (placebo). Θα ισχύει ότι το πηλίκο των κινδύνων (hazard ratio), στο χρόνο t, είναι $HR_c = \frac{\lambda(t|x=1,u_1)}{\lambda(t|x=0,u_0)} = e^{\beta}$ μόνο όταν $u_1 = u_0$. Επομένως, ένα άτομο που παίρνει θεραπεία θα έχει e^{β} φορές μικρότερο κίνδυνο να αποτύχει σε σχέση με ένα άλλο άτομο, το οποίο έχει την ίδια frailty τιμή και δεν παίρνει θεραπεία ($\beta < 0$). Όμως, σε πληθυσμιακό επίπεδο η αναλογικότητα των κινδύνων δεν ισχύει γενικά. Το πληθυσμιακό πηλίκο πηλίκο κινδυνων (marginal hazard ratio, unconditional hazard ratio) είναι

$$HR_{marginal} = \frac{\lambda(t|x=1)}{\lambda(t|x=0)} = \frac{1 + \sigma^2\Lambda_0(t)}{1 + \sigma^2\Lambda_0(t)e^{\beta}} e^{\beta}. \tag{2.14}$$



Σχήμα 2.5: Συνάρτησεις κινδύνων δυο πληθυσμών που ακολουθούν την Gompertz κατανομή ($\lambda_1 = 10^{-4}$, $\lambda_2 = 4\lambda_1$, $\phi_1 = \phi_2 = 0.12$). Η διακύμανση των frailties είναι $\sigma^2 = 1$ σύμφωνα με την Gamma κατανομή. Η συνάρτηση κινδύνου είναι ίση με $\lambda_0(t) = \lambda e^{\phi t}$.

Το πληθυσμιακό πηλίκο πηλίκο κινδυνων (2.14) δεν είναι ανεξάρτητο του χρόνου, εκτός αν $\beta = 0$ ή $\sigma^2 = 0$. Στο χρόνο $t = 0$, το πληθυσμιακό πηλίκο κινδύνων ισούται με το δεσμευμένο πηλίκο κινδύνων e^{β}, αλλά κάθως περνάει ο χρόνος, απομακρύνεται από το e^{β} και οριακά τείνει στο 1. Η ασυμφωνία μεταξύ των δυο πηλίκων κινδύνων είναι μεγαλύτερη όταν οι τιμές των β,σ^2 και $\Lambda_0(t)$ είναι μεγάλες. Στο γράφημα (2.5), ο δεύτερος πληθυσμός έχει μεγαλύτερο κινδύνο σε σχέση με τον πρώτο. Όμως, βλέπουμε ότι λόγω της ετερογένειας που υφίσταται στους δυο πληθυσμούς, το πληθυσμιακό πηλίκο κινδύνων τείνει στο 1 καθώς αυξάνεται ο χρόνος.

Αν ακολουθήσουμε την κλασική προσέγγιση της στατιστικής, οι εκτιμήσεις των παραμέτρων προέρχονται από τη μεγιστοποίηση της περιθώριας πιθανοφάνειας. Ολοκληρώνοντας από την (2.11) τις τυχαίες ποσότητες u_i, καταλήγουμε στην περιθώρια πιθανοφάνεια:

$$\begin{aligned} L_{marginal}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\psi}) &= \prod_{i=1}^{n} (\lambda(y_i|\mathbf{x}_i))^{\delta_i} S(y_i|\mathbf{x}_i) \\ &= \prod_{i=1}^{n} \left(\frac{\lambda_0(y_i|\boldsymbol{\psi}) e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + \sigma^2 \Lambda_0(y_i|\boldsymbol{\psi}) e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \right)^{\delta_i} (1 + \sigma^2 \Lambda_0(y_i|\boldsymbol{\psi}) e^{\mathbf{x}_i^T \boldsymbol{\beta}})^{-\frac{1}{\sigma^2}}, \end{aligned} \tag{2.15}$$

όπου με $\boldsymbol{\psi}$ συμβολίζουμε τις παραμέτρους οι οποίες συσχετίζονται με τη βασική συνάρτηση κινδύνου. Αν υποθέσουμε κάποια παραμετρική μορφή για το $\lambda_0(t)$, τότε μπορούμε να χρησιμοποιήσουμε συνήθεις μεθόδους για τη μεγιστοποίηση της (2.15). Αν θέλουμε να αφήσουμε το $\lambda_0(t)$ ως αδιευκρίνιστο, μια επιλογή είναι να χρησιμοποιήσουμε τον αλγόριθμο EM (expectation-maximization). Για περισσότερες πληροφορίες παραπέμπουμε στο βιβλίο των Duchateau & Janssen (2007).

2.6.2 Lognormal Frailty Μοντέλο

Η Lognormal και η Gamma κατανομές χρησιμοποιούνται συχνότερα ως κατανομές για τα frailties. Σε αντίθεση με την Gamma κατανομή, η Lognormal κατανομή δεν προσφέρει κλειστές εκφράσεις για την πληθυσμιακή συνάρτηση κινδύνου και επιβίωσης. Επομένως, χρειαζόμαστε μεθόδους αριθμητικής ή στοχαστικής ολοκλήρωσης για τον υπολογισμό των απαιτούμενων ολοκληρωμάτων. Μια τυχαία μεταβλητή u ακολουθεί την Lognormal κατανομή ($u \sim$ Lognormal(μ, σ^2)) όταν ο λογάριθμος $b = \log(u)$ ακολουθεί την κανονική κατανομή με μέση τιμή μ και διασπορά σ^2 ($b \sim \mathrm{N}(\mu, \sigma^2)$). Η συνάρτηση πιθανότητας πυκνότητας της u θα είναι (γράφημα 2.6):

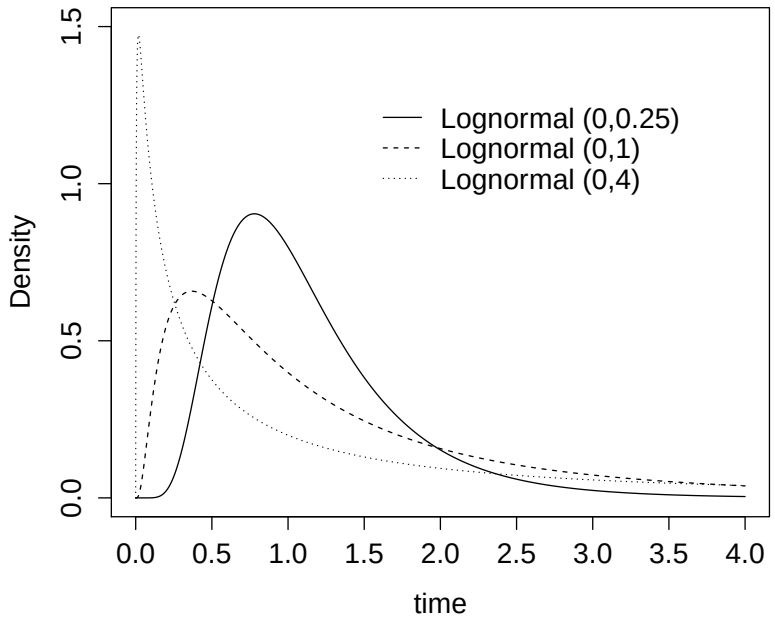
$$f(u) = (2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{1}{2\sigma^2} (\log(u) - \mu)^2 \right\} \frac{1}{u}.$$

Είναι λογικό να θέσουμε τον περιορισμό $\mathrm{E}(b) = \mathrm{E}(\log(u)) = 0$. Άρα η μέση τιμή του u δεν θα είναι 1 αλλά $\mathrm{E}(u) = e^{\frac{\sigma^2}{2}}$. Εναλλακτικά, θα μπορούσαμε να θέσουμε τον περιορισμό $\mathrm{E}(u) = 1$,

ο οποίος έχει το πλεονέκτημα της συγκρισιμότητας με τα υπόλοιπα frailty μοντέλα. Παρόλα αυτά ο περιορισμός $\mathrm{E}(b) = 0$ χρησιμοποιείται πιο συχνά από τα στατιστικά πακέτα. Ο λόγος είναι ταιριάζει με την υπόθεση τυχαίων επιδράσεων που δρούν προσθετικά στο linear predictor των γενικευμένων γραμμικών μοντέλων (glm). Η πιθανοφάνεια των δεδομένων $(y_i, \delta_i, \mathbf{x}_i)$, $i = 1, 2, \ldots, n$ θα είναι (Wienke 2010, σελ. 98):

$$L_{marginal}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\psi}) = \prod_{i=1}^{n} \int_{-\infty}^{+\infty} \left(\lambda_0(y_i) e^{\mathbf{x}_i^T \boldsymbol{\beta} + b_i}\right)^{\delta_i} \exp\left\{-\Lambda_0(y_i) e^{\mathbf{x}_i^T \boldsymbol{\beta} + b_i}\right\} d\Phi(b_i)$$

όπου με Φ συμβολίζουμε τη συναρτηση κατανομής μιας κανονικής τυχαίας μεταβλητής με μέση τιμή 0 και διακύμανση σ^2. Αν υποθέσουμε παραμετρική κατανομή για τη βασική συνάρτηση κινδύνου, μπορούμε να χρησιμοποιήσουμε για τη μεγιστοποίηση της πιθανοφάνειας μεθόδους αριθμητικής ολοκλήρωσης (π.χ. adaptive Gaussian quadrature) ή MCMC μεθόδους κάτω από τη Μπευζιανή λογική. Αν δεν κάνουμε κάποια υπόθεση για τη βασική συνάρτηση κινδύνου, τότε μια λύση είναι να χρησιμοποιήσουμε penalised partial likelihood μεθόδους (Duchateau & Janssen 2007). Στο γράφημα (2.4) βλέπουμε ότι η συνάρτηση πυκνότητας της Lognormal κατανομής παίρνει διάφορες μορφές. Όσο η διακύμανση μικραίνει, τόσο οι τιμές είναι πιο συγκεντρωμένες στο 1.



Σχήμα 2.6: Συναρτήσεις πυκνότητας πιθανότητας της Lognormal κατανομής με τιμές παραμέτρων $\mu = 0$ και σ^2 =0.25, 1, και 4.

2.7 Shared Frailty Μοντέλα

Τα περισσότερα στατιστικά μοντέλα τα οποία έχουν προταθεί για την ανάλυση δεδομένων επιβίωσης υποθέτουν ότι διαφορετικές παρατηρήσεις είναι στοχαστικά ανεξάρτητες μεταξύ τους. Όμως, σε πολλές περιπτώσεις αυτή η υπόθεση δεν είναι αληθοφανής, καθώς μπορεί να υπάρχει συσχέτιση των παρατηρήσεων μέσα σε κάποιες υποομάδες του πληθυσμού (clusters). Τα shared frailty μοντέλα υποθέτουν ότι υπάρχει μια λανθάνουσα τυχαία μεταβλητή, u, η οποία συσχετίζεται με μη παρατηρήσιμους παράγοντες, και είναι κοινή ανάμεσα στα υποκείμενα της ίδιας ομάδας. Επομένως, τα υποκείμενα τα οποία ανήκουν στην ίδια ομάδα, εφόσον μοιράζονται την ίδια τυχαία μεταβλητή (frailty), θα παράγουν συσχετισμένες παρατηρήσεις (περιθωριακά). Η συσχέτιση η οποία προκαλείται από ένα Shared Frailty μοντέλο είναι πάντα θετική (με λίγες εξαιρέσεις). Το μοντέλο υποθέτει ότι οι χρόνοι επιβίωσης των ατόμων της ίδιας ομάδας, δοθείσης της τιμής του u, είναι ανεξάρτητοι μεταξύ τους. Οι χρόνοι επιβίωσης υποκειμένων προερχόμενων απο διαφορετικές ομάδες θεωρούνται ανεξάρτητοι μεταξύ τους.

Το shared frailty μοντέλο είναι ιδιαίτερα χρήσιμο για την ανάλυση ομαδοποιημένων δεδομένων επιβίωσης. Για παράδειγμα, είναι λογικό να υποθέσουμε ότι οι χρόνοι επιβίωσης διδύμων ή ότι οι χρόνοι επαναλαμβανόμενων εμφανίσεων μιας ασθένειας θα είναι συσχετισμένοι μεταξύ τους. Είναι εμφανές ότι μια ομάδα μπορεί να αντιπροσωπεύει παρατηρήσεις προερχόμενες από το ίδιο άτομο. Η μορφή της συσχέτισης μεταξύ των υποκειμένων της ίδιας ομάδας εξαρτάται από την υπόθεση για την κατανομή της frailty μεταβλητής, u, η οποία για λόγους αναγνωρισιμότητας έχει μέση τιμή και διακύμανση (αν υπάρχει):

$$\mathrm{E}(u) = 1,\ \mathrm{Var}(u) = \sigma^2.$$

Το μονοδιάστατο frailty μοντέλο, το οποίο εξετάστηκε στην προηγούμενη ενότητα, είναι ειδική περίπτωση του shared frailty μοντέλου όταν κάθε ομάδα περιέχει μια παρατήρηση. Όμως, τα δυο μοντέλα έχουν διαφορετικούς σκοπούς και χρησιμοποιούνται σε διαφορετικού είδους δεδομένα. Στα μονοδιάστατα frailty μοντέλα εξετάζεται η ετερογένεια του πληθυσμού η οποία προκαλείται από μη παρατηρήσιμους παράγοντες και η διακύμανση των frailties σ^2 δεν υποδηλώνει συσχέτιση μεταξύ των παρατηρήσεων, είναι απλώς ένα μέτρο ετερογένειας. Στα shared frailty μοντέλα, ο σκοπός είναι να μοντελοποιήσουμε ρητά τη συσχέτιση των παρατηρήσεων της ίδιας ομάδας ή να βγάλουμε συμπεράσματα για επεξηγηματικές μεταβλητές οι οποίες μας ενδιαφέρουν, λαμβάνοντας υπόψη τη συσχέτιση των παρατηρήσεων.

Ένα shared frailty μοντέλο ορίζεται ως εξής: Έστω ότι το δείγμα αποτελείται από G ομάδες και έχουμε n_i παρατηρήσεις στην i-οστή ομάδα ($i = 1, 2, \ldots, G$). Συμβολίζουμε ως $\mathbf{x}_{ij}^T = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})$ το p-διάστατο διάνυσμα επεξηγηματικών μεταβλητών της j-οστής παρατήρησης από την i-οστή ομάδα. Η συνάρτηση κινδύνου της j-οστής παρατήρησης από την

i-οστή ομάδα, δοθείσης της τιμής του u_i, θα είναι:

$$\lambda(t|\mathbf{x}_{ij}, u_i) = \lambda_0(t)e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}u_i,\ i = 1, 2, \ldots, G,\ j = 1, 2, \ldots, n_i,$$

όπου με $\lambda_0(t)$ συμβολίζουμε τη βασική συνάρτηση κινδύνου και $\boldsymbol{\beta}$ είναι το διάνυσμα των παραμέτρων που συσχετίζονται με τις επεξηγηματικές μεταβλητές. Τα frailties $u_1, u_2, \ldots, u_G$ θεωρούμε ότι είναι ανεξάρτητα και ταυτοτικά κατανεμημένα σύμφωνα με κάποια κατανομή με συνάρτηση πυκνότητας πιθανότητας $f(u)$. Άρα, υποθέτουμε ότι τα frailties είναι εκ των προτέρων ανεξάρτητα από τις επεξηγηματικές μεταβλητές. Αν όμως οι μη παρατηρήσιμοι παράγοντες (frailties u_i) περιλαμβάνουν και συγχυτικούς παράγοντες για τη σχέση των επεξηγηματικών μεταβλητών με την έκβαση, τότε η προηγούμενη υπόθεση καταστραγείται και το μοντέλο δεν είναι σωστά ορισμένο (Sjölander, Lichtenstein, Larsson & Pawitan 2013). Σε μια τέτοια περίπτωση, είναι πολύ πιθανό οι εκτιμήσεις του μοντέλου να είναι μεροληπτικές. Η απο κοινού συνάρτηση επιβίωσης των υποκειμένων της i-οστής ομάδας, δοθέντος του u_i, θα είναι:

$$S(t_{i1}, t_{i2}, \ldots, t_{in_i}|\mathbf{X}_i, u_i) = \prod_{j=1}^{n_i} S(t_{ij}|\mathbf{x}_{ij}, u_i) = \exp\left\{-u_i \sum_{j=1}^{n_i} \Lambda_0(t_{ij})e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\right\},$$

όπου $\mathbf{X}_i^T = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{in_i})$ είναι ο πίνακας σχεδιασμού (design matrix) της i-οστής ομάδας και $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ είναι η βασική συνάρτηση αθροιστικού κινδύνου. Ολοκληρώνοντας ως προς u_i μπορούμε να βρούμε την περιθώρια απο κοινού συνάρτηση επιβίωσης της i-οστής ομάδας:

$$S(t_{i1}, t_{i2}, \ldots, t_{in_i}|\mathbf{X}_i) = \mathrm{E}_{u_i}\left(e^{-u_i\sum_{j=1}^{n_i}\Lambda_0(t_{ij})e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}}\right) = \mathbf{L}\left(\sum_{j=1}^{n_i}\Lambda_0(t_{ij})e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\right),$$

ενώ η μονοδιάστατη περιθώρια συνάρτηση επιβίωσης θα είναι

$$S(t_{ij}|\mathbf{x}_{ij}) = \mathrm{E}_{u_i}\left(e^{-u_i\Lambda_0(t_{ij})e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}}\right) = \mathbf{L}\left(\Lambda_0(t_{ij})e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\right). \tag{2.16}$$

Η $S(t_{ij}|\mathbf{x}_{ij})$ από (2.16) δεν μπορεί να ερμηνευτεί σε όλες τις περιπτώσεις ως μια συνάρτηση επιβίωσης. Για παράδειγμα, ας θεωρήσουμε ένα πληθυσμό ο οποίος αποτελείται από ομάδες και το μέγεθος της κάθε ομάδας είναι συσχετισμένο με την αντίστοιχη τιμή της frailty μεταβλητής. Σε αυτή την περίπτωση η $S(t_{ij}|\mathbf{x}_{ij})$ δεν μπορεί να χρησιμοποιηθεί ως η αναλογία του πληθυσμού που επέζησε στο χρόνο t_{ij}. Μπορούμε όμως να ερμηνεύσουμε την $S(t_{ij}|\mathbf{x}_{ij})$ ως την πιθανότητα επιβίωσης στο χρόνο t_{ij} ενός τυχαία επιλεγόμενου ατόμου, με χαρακτηριστικά $\mathbf{x}_{ij}$, από μια τυχαία επιλεγόμενη ομάδα. Εναλλακτικά, αφού ελέγξουμε ότι το μέγεθος της ομάδας είναι ανεξάρτητο με τη frailty μεταβλητή, μπορούμε να ερμηνεύσουμε την (2.16) ως μια συνήθη συνάρτηση επιβίωσης (Gutierrez 2002).

Κεφάλαιο 3

Μπεϋζιανή Στατιστική

Το παρόν κεφάλαιο επικεντρώνεται στις βασικές αρχές της Μπεϋζιανής στατιστικής και στη περιγραφή των μεθόδων Markov Chain Monte Carlo (MCMC), τις οποίες θα χρησιμοποίησουμε στα επόμενα κεφάλαια. Στην ενότητα 3.3 περιγράφεται η μέθοδος προσομοίωσης adaptive rejection sampling των Gilks & Wild (1992).

3.1 Βασική Θεωρία της Μπεϋζιανής Στατιστικής

Η Μπεϋζιανή στατιστική βασίζεται στον ορισμό ενός πιθανοθεωρητικού μοντέλου $f(\mathbf{y}|\boldsymbol{\theta})$ για τα παρατηρούμενα δεδομένα $\mathbf{y}$, δοθέντος ενός διανύσματος αγνώστων παραμέτρων $\boldsymbol{\theta}$. Η θεμελιώδης διαφορά της κλασσικής με τη Μπεϋζιανή προσέγγιση στη στατιστική είναι ότι η δεύτερη θεωρεί την παράμετρο $\boldsymbol{\theta}$ ως τυχαία μεταβλητή και όχι ως άγνωστη σταθερά. Βέβαια, το γεγονός αυτό συνεπάγεται ότι πρέπει να οριστεί μια εκ των προτέρων κατανομή (prior distribution) για το $\boldsymbol{\theta}$, η οποία συμβολίζεται γενικά ως $f(\boldsymbol{\theta})$. Η εκ των προτέρων κατανομή εκφράζει τη γνώση μας για την παράμετρο $\boldsymbol{\theta}$ πριν δούμε τα δεδομένα. Η συμπερασματολογία δεν θα προκύψει μόνο από τη συνάρτηση πιθανοφάνειας $f(\mathbf{y}|\boldsymbol{\theta})$, αλλά από το συνδυασμό της πιθανοφάνειας με την εκ των προτέρων κατανομή του $\boldsymbol{\theta}$, η οποία ονομάζεται εκ των υστέρων κατανομή (posterior distribution). Η εκ των υστέρων κατανομή είναι η κατανομή της παραμέτρου $\boldsymbol{\theta}$ δοθέντων των δεδομένων, $f(\boldsymbol{\theta}|\mathbf{y})$, η οποία προκύπτει απο το θεώρημα Bayes ως εξής:

$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\mathbf{y})} = \frac{f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{\int f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{3.1}$$

Ο παρανομαστής του θεωρήματος Bayes, $f(\mathbf{y}) = \int f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}$, είναι η σταθερά κανονικοποίησης της $f(\boldsymbol{\theta}|\mathbf{y})$ (ώστε να ολοκληρώνει στο 1) και αποτελεί συνάρτηση μόνο των δεδομένων $\mathbf{y}$. Επιπλέον, η $f(\mathbf{y})$ ονομάζεται περιθώρια πιθανοφάνεια (marginal likelihood) του μοντέλου και μπορεί να ερμηνευτεί ως η περιθώρια κατανομή των δεδομένων ή ως η αναναμενόμενη τιμή της πιθανοφάνειας ως προς την εκ των προτέρων κατανομή, δηλαδή $f(\mathbf{y}) = \mathrm{E}_{\boldsymbol{\theta}}\left(f(\mathbf{y}|\boldsymbol{\theta})\right) =$

$\int f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}$. Συχνά το θεώρημα Bayes παρουσιάζεται ως

$$f(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}),$$

δηλαδή η εκ των υστέρων κατανομή είναι ανάλογη της πιθανοφάνειας πολλαπλασιαζόμενης με την εκ των προτέρων κατανομή. Έστω ότι το $\boldsymbol{\theta}$ αποτελείται από d παραμέτρους, δηλαδή $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)$. Η εκ των υστέρων κατανομή θα είναι μια πολυδιάστατη κατανομή. Η ακριβής Μπεϋζιανή συμπερασματολογία για την παράμετρο θ_i, επιτυγχάνεται ολοκληρώνοντας την απο κοινού εκ των υστέρων κατανομή ως προς τις υπόλοιπες παραμέτρους $\boldsymbol{\theta}_{-i}$,

$$f(\theta_i|\mathbf{y}) = \int f(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-i}.$$

Η περιθώρια εκ των υστέρων κατανομή του θ_i εκφράζει όλη την πληροφορία που έχουμε για την παραμέτρο θ_i, αφού λάβουμε υπόψη τα δεδομένα και την προηγούμενη γνώση μας. Στα περίσσοτερα πραγματικά προβλήματα, το ολοκλήρωμα που δίνει τη σταθερά κανονικοποίησης δεν έχει αναλυτική λύση, άρα, η $f(\boldsymbol{\theta}|\mathbf{y})$ δεν μπορεί να βρεθεί σε κλειστή μορφή. Αυτή η δυσκολία οδήγησε στη χρήση των λεγόμενων συζυγών εκ των προτέρων κατανομών. Μια οικογένεια εκ των προτέρων κατανομών είναι συζυγής για ένα μοντέλο πιθανοφάνειας, όταν η εκ των υστέρων κατανομή ανήκει στην ίδια οικογένεια στην οποία ανήκει και η εκ των προτέρων κατανομή. Είναι γνωστό ότι συζυγείς κατανομές είναι διαθέσιμες για πιθανοφάνειες που ανήκουν στην εκθετική οικογένεια κατανομών, υπό την προϋπόθεση ότι το δείγμα προέρχεται από ανεξάρτητες και ισόνομες τυχαίες μεταβλητές.

Παραδειγμα 3.1. *Γραμμική παλινδρόμηση με κανονικά σφάλματα και γνωστή διακύμανση*

Έστω ανεξάρτητες παρατηρήσεις, $y_1, y_2, \ldots, y_n$, οι οποίες προέρχονται από την κανονική κατανομή με γνωστή διακύμανση $\sigma^2 = \frac{1}{\omega}$. Η μέση τιμή θα εξάρταται, πιθανώς, από p επεξηγηματικές μεταβλητές, $X_1, X_2, \ldots, X_p$, μέσω της ταυτοτικής συνδετικής συνάρτησης: $\mu_i = \mathrm{E}(\mathrm{y_i}|\mathbf{x_i}) = \mathbf{x}_i^T\boldsymbol{\beta}$, όπου $\mathbf{x}_i^T = (1, x_{i1}, x_{i2}, \ldots, x_{ip})$ είναι το διάνυσμα των επεξηγηματικών μεταβλητών και $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_p)$ η άγνωστη παράμετρος. Για το $\boldsymbol{\beta}$ θα θεωρήσουμε ως εκ των προτέρων κατανομή την κανονική με μέση τιμή $\boldsymbol{\mu}_0$ και πίνακα συνδιακύμανσης $\mathbf{C}_0$. Άρα,

$$f(\boldsymbol{\beta}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)^T\mathbf{C}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)\right\} \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}^T\mathbf{C}_0^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\mathbf{C}_0^{-1}\boldsymbol{\mu}_0\right)\right\}.$$

Η πιθανοφάνεια του μοντέλου, εφόσον $\mathbf{y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \omega^{-1}\mathbf{I}_n)$, θα είναι:

$$f(\mathbf{y}|\boldsymbol{\beta}) \propto \exp\left\{-\frac{\omega}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\} \propto \exp\left\{-\frac{\omega}{2}\left(\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}\right)\right\}.$$

Εφαρμόζοντας το θεώρημα Bayes θα δείξουμε ότι η από-κοινού εκ των υστέρων κατανομή του $\boldsymbol{\beta}$

ανήκει στην ίδια οικογένεια κατανομών με την εκ των προτέρων κατανομή, αλλά με ανανεωμένες παραμέτρους:

$$f(\boldsymbol{\beta}|\mathbf{y}) \propto \exp\left\{-\frac{1}{2}\Big(\boldsymbol{\beta}^T(\mathbf{C}_0^{-1}+\omega\mathbf{X}^T\mathbf{X})\boldsymbol{\beta}-2\boldsymbol{\beta}^T(\mathbf{C}_0^{-1}\boldsymbol{\mu}_0+\omega\mathbf{X}^T\mathbf{y})\Big)\right\}.$$

Αναγνωρίζουμε την εκ των υστέρων κατανομή του $\boldsymbol{\beta}$ ως κανονική με πίνακα συνδιακύμανσης $\mathbf{C}_1 = (\mathbf{C}_0^{-1}+\omega\mathbf{X}^T\mathbf{X})^{-1}$ και μέση τιμη $\boldsymbol{\mu}_1 = \mathbf{C}_1(\mathbf{C}_0^{-1}\boldsymbol{\mu}_0+\omega\mathbf{X}^T\mathbf{y})$. Αν θέλουμε η εκ των προτέρων πληροφορία να έχει ελάχιστη βαρύτητα στην διαμόρφωση των αποτελεσμάτων, τότε επιλέγουμε μεγάλη εκ των προτέρων διακύμανση, με συνέπεια ο $\mathbf{C}_0^{-1}$ να τείνει σε ένα μηδενικό πίνακα. Σε αυτήν την περίπτωση, θα ισχύει προσεγγιστικά ότι $\mathbf{C}_1 \simeq \omega^{-1}(\mathbf{X}^T\mathbf{X})^{-1} = \mathrm{Var}(\widehat{\boldsymbol{\beta}})$ και $\boldsymbol{\mu}_1 \simeq \widehat{\boldsymbol{\beta}}$, όπου $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ είναι ο εκτιμητής μέγιστης πιθανοφάνειας του $\boldsymbol{\beta}$.

Η επιλογή της εκ των προτέρων κατανομής παίζει, πιθανώς, το σημαντικότερο ρόλο στη Μπεϋζιανή συμπερασματολογία. Διαφορετική εκ των προτέρων κατανομή οδηγεί σε διαφορετική εκ των υστέρων κατανομή. Υπο αυτή την έννοια, η ανάλυση καθίσταται υποκειμενική (subjective Bayesian analysis). Ωστόσο, οποιαδήποτε λογική επιλογή της εκ των προτέρων κατανομής θα έχει ελάχιστη επιρροή στα αποτελέσματα αν έχουμε επαρκώς πληροφοριακά δεδομένα. Η εκ των προτέρων κατανομή στις περισσότερες περιπτώσεις θα είναι πολυδιάστατη κατανομή. Επομένως, θα χρειαστεί να εισάγουμε πληροφορία και για τη συσχέτιση των παραμέτρων μεταξύ τους. Η συνήθης επιλογή είναι να θεωρούμε τις παραμέτρους ανεξάρτητες, εκτός αν υπάρχουν πολύ συγκεκριμένοι λόγοι οι οποίοι να υποδεικνύουν ότι υπάρχει συσχέτιση μεταξύ των παραμέτρων.

Οι εκ των προτέρων κατανομές μπορούν να χωριστούν, γενικά, σε πληροφοριακές (informati ve prior) και μη πληροφοριακές (noninformative prior). Οι μη πληροφοριακές εκ των προτέρων κατανομές δίνουν το ίδιο βάρος σε όλο τον παραμετρικό χώρο, με συνέπεια η κατανομή να γίνεται επίπεδη και το ολοκλήρωμα $\int f(\boldsymbol{\theta})d\boldsymbol{\theta} = \infty$, δηλαδή η $f(\boldsymbol{\theta})$ είναι μη γνήσια (improper) κατανομή. Ωστόσο, μη γνήσιες εκ των προτέρων κατανομές μπορούν να χρησιμοποιηθούν εφόσον οι αντίστοιχες εκ των υστέρων κατανομές έχουν πεπερασμένο ολοκλήρωμα (Gelman, Robert, Chopin & Rousseau 1995, σελ. 37). Οι μη πληροροφοριακές κατανομές δεν κάνουν χρήση της προηγούμενης γνώσης η οποία μπορεί να υπάρχει σε ένα συγκεκριμένο θέμα. Επομένως, πληροφοριακές εκ των προτέρων κατανομές θα μπορούσαν να χρησιμοποιηθούν σε τέτοιες περιπτώσεις, αλλά και γενικότερα είναι χρήσιμες όταν οι ερευνητές έχουν πρόσβαση σε προηγούμενες μελέτες οι οποίες έχουν την ίδια εξαρτημένη μεταβλητή και επεξηγηματικές μεταβλητές.

3.2 Αλγόριθμοι Markov Chain Monte Carlo (MCMC)

Στα περισσότερα προβλήματα είναι δύσκολο να υπολογιστούν ποσότητες όπως η εκ των υστέρων μέση τιμή ή διαστήματα αξιοπιστίας για τις παραμέτρους. Επιπλέον, η περιθώρια πιθανοφάνεια, $f(\mathbf{y})$, είναι σύνηθες να μην μπορεί να υπολογιστεί σε κλειστή μορφή. Επομένως, η εκ των υστέρων κατανομή των παραμέτρων θα είναι γνωστή ως μια σταθερά κανονικοποίησης. Όλα αυτά τα προβήματα οδήγησαν σε μια διαφορετική προσέγγιση. Εφόσον δεν μπορούμε να υπολογίσουμε αναλυτικά τα χαρακτηριστικά της εκ των υστέρων κατανομής τα οποία μας ενδιαφέρουν, μπορούμε να πάρουμε δείγμα από αυτήν. Όμως, ευθεία προσομοίωση από μια αυθαίρετη πολυδιάστατη κατανομή δεν είναι πάντα εφικτή. Αντ' αυτού, μπορούν να χρησιμοποιηθούν Markov chain Monte Carlo (MCMC) μέθοδοι, οι οποίες προσομοιώνουν μια μαρκοβιανή αλυσίδα με στάσιμη κατανομή την απο κοινού, εκ των υστέρων, κατανομή των παραμέτρων $f(\boldsymbol{\theta}|\mathbf{y})$. Κάτω από κάποιες ήπιες συνθήκες ομαλότητας, επιτυγχάνεται η σύγκλιση της αλυσίδας στην στάσιμη κατανομή.

3.2.1 Δειγματολήπτης Gibbs

Ένα από τα πιο δημοφιλή MCMC σχήματα είναι ο δειγματολήπτης Gibbs (Geman & Geman 1984). Ο δειγματολήπτης Gibbs είναι ιδιαίτερα χρήσιμος όταν η εκ των υστέρων κατανομή είναι αρκετά περίπλοκη, αλλά οι πλήρως δεσμευμένες κατανομές των παραμέτρων είναι γνωστές κατανομές ή είναι εύκολο να προσομοιωθούν. Ας υποθέσουμε ότι $\boldsymbol{\theta}^T = (\theta_1, \theta_2, \ldots, \theta_d)$, όπου κάθε θ_i μπορεί να είναι πολυδιάστατο. Ο αλγόριθμος προσομοιώνει δείγμα από την απο κοινού εκ των υστέρων κατανομή των παραμέτρων, προσομοιώνοντας επαναληπτικά και ακολουθιακά από τις πλήρως δεσμευμένες κατανομές των επι μέρους ομάδων των παραμέτρων. Ο αλγόριθμός (3.1) περιγράφει την γενική περίπτωση ενός δειγματολήπτη Gibbs.

Αλγόριθμος 3.1. *Δειγματολήπτης Gibbs*

- Δίνουμε αρχικές τιμές στο διάνυσμα των παραμέτρων $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_d^{(0)})^T$. Έστω ότι η τρέχουσα τιμή του αλγορίθμου είναι $\boldsymbol{\theta}^{(j)} = (\theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_d^{(j)})^T$, $j = 0, 1, \ldots$
- Προσομοιώνουμε $\theta_1^{(j+1)} \sim f(\theta_1|\mathbf{y}, \theta_2^{(j)}, \theta_3^{(j)}, \ldots, \theta_d^{(j)})$.
- Προσομοιώνουμε $\theta_2^{(j+1)} \sim f(\theta_2|\mathbf{y}, \theta_1^{(j+1)}, \theta_3^{(j)}, \ldots, \theta_d^{(j)})$.
- ...
- Προσομοιώνουμε $\theta_d^{(j+1)} \sim f(\theta_d|\mathbf{y}, \theta_1^{(j+1)}, \theta_2^{(j+1)}, \ldots, \theta_{d-1}^{(j+1)})$.
- Επαναλαμβάνουμε τη διαδικασία μέχρι να επέλθη σύγκλιση.

Μετά από μια περίοδο burn-in (όπου οι τιμές της αλυσίδας απορρίπτονται), οι επόμενες τιμές μπορούν να θεωρηθούν ως πραγματοποιήσεις από την $f(\boldsymbol{\theta}|\mathbf{y})$.

3.2.2 Αλγόριθμος Metropolis Hastings

Ο αλγόριθμος Metropolis–Hastings (Hastings 1970) είναι ιδιαίτερα χρήσιμος όταν δεν μπορούμε να προσομοιώσουμε από την απο κοινού εκ των υστέρων κατανομή, $f(\boldsymbol{\theta}|\mathbf{y})$, αλλά ούτε και από τις πλήρως δεσμευμένες κατανομές των παραμέτρων. Το μόνο που χρειάζεται είναι να γνωρίζουμε τις δεσμευμένες posterior κατανομές ως μια σταθερά κανονικοποίησης. Ας υποθέσουμε ότι $\boldsymbol{\theta}^T = (\theta_1, \theta_2, \ldots, \theta_d)$, όπου κάθε θ_i μπορεί να είναι πολυδιάστατο. Ο αλγόριθμος (3.2) περιγράφει το γενικό πλαίσιο στο οποίο στηρίζονται όλοι οι αλγόριθμοι MCMC.

Αλγόριθμος 3.2. *General MCMC Algorithm*

- Δίνουμε αρχικές τιμές στο διάνυσμα των παραμέτρων $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_d^{(0)})^T$. Έστω ότι η τρέχουσα τιμή του αλγορίθμου είναι $\boldsymbol{\theta}^{(j)} = (\theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_d^{(j)})^T$, $j = 0, 1, \ldots$
- Προσομοιώνουμε $\theta_1^{(j+1)}$ σύμφωνα με $f(\theta_1|\mathbf{y}, \theta_2^{(j)}, \theta_3^{(j)}, \ldots, \theta_d^{(j)})$.
- Προσομοιώνουμε $\theta_2^{(j+1)}$ σύμφωνα με $f(\theta_2|\mathbf{y}, \theta_1^{(j+1)}, \theta_3^{(j)}, \ldots, \theta_d^{(j)})$.
- ...
- Προσομοιώνουμε $\theta_d^{(j+1)}$ σύμφωνα με $f(\theta_d|\mathbf{y}, \theta_1^{(j+1)}, \theta_2^{(j+1)}, \ldots, \theta_{d-1}^{(j+1)})$.
- Επαναλαμβάνουμε τη διαδικασία μέχρι να επέλθη σύγκλιση.

Έστω ότι η τρέχουσα τιμή της αλυσίδας είναι η $\boldsymbol{\theta}^{(j)} = (\theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_d^{(j)})^T$. Επομένως, θα πρέπει, κατά το γενικό αλγόριθμο MCMC, να ανανεώσουμε την τιμή $\theta_1^{(j)}$ σε $\theta_1^{(j+1)}$ σύμφωνα με την κατανομή $f(\theta_1|\mathbf{y}, \theta_2^{(j)}, \theta_3^{(j)}, \ldots, \theta_d^{(j)})$. Στον αλγόριθμο Metropolis–Hastings τα βήματα είναι τα εξής:

- Προτείνουμε μια νέα τιμή θ_1^{can} από μια αυθαίρετη κατανομή με συνάρτηση πυκνότητας $q(\theta_1^{can}|\mathbf{y}, \theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_d^{(j)})$.
- Δεχόμαστε την προτεινόμενη τιμή με πιθανότητα p όπου:

$$p = \min\left\{1, \frac{f(\theta_1^{can}|\mathbf{y}, \theta_2^{(j)}, \theta_3^{(j)}, \ldots, \theta_d^{(j)})}{f(\theta_1^{(j)}|\mathbf{y}, \theta_2^{(j)}, \theta_3^{(j)}, \ldots, \theta_d^{(j)})} \frac{q(\theta_1^{(j)}|\mathbf{y}, \theta_1^{(can)}, \theta_2^{(j)}, \ldots, \theta_d^{(j)})}{q(\theta_1^{can}|\mathbf{y}, \theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_d^{(j)})}\right\}.$$

Η επιλογή της γεννήτριας προτεινόμενων τιμών είναι αυθαίρετη. Επομένως, θεωρητικά, κάθε επιλογή μπορεί να γίνει αποδεκτή. Στην πράξη όμως θέλουμε γεννήτριες οι οποίες να μοιάζουν αρκετά με τις αντίστοιχες εκ των υστέρων κατανομές, ώστε να έχουμε υψηλό ποσοστό αποδοχής. Ο δειγματολήπτης Gibbs είναι υποπερίπτωση του αλγορίθμου Metropolis–Hastings όπου $q(\theta_1^{can}|\mathbf{y}, \theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_d^{(j)}) = f(\theta_1^{can}|\mathbf{y}, \theta_2^{(j)}, \theta_3^{(j)}, \ldots, \theta_d^{(j)})$. Σε αυτή την περίπτωση το ποσοστό αποδοχής είναι πάντα ένα. Ο στόχος μας είναι να προσομοιώσουμε ένα τυχαίο δείγμα από την

εκ των υστέρων κατανομή των παραμέτρων. Όμως, οι MCMC αλγόριθμοι παράγουν εξαρτημένα δείγματα. Άρα, αναμένουμε οι τιμές της αλυσίδας να αυτοσυσχετίζονται. Σαν ένα γενικό κανόνα, για να έχει ο MCMC αλγόριθμος καλή μείξη, δηλαδή να γεννάει ένα όσο το δυνατόν πιο τυχαίο δείγμα από την εκ των υστέρων κατανομή, $f(\boldsymbol{\theta}|\mathbf{y})$, πρέπει οι παράμετροι διαφορετικών ομάδων να είναι ασυσχέτιστες μεταξύ τους.

3.2.3 MCMC σε Γενικευμένα Γραμμικά Μοντέλα

Στην παρούσα ενότητα θα παρουσιαστεί ο αλγόριθμος του Gamerman για την εφαρμογή MCMC μεθόδων στα γενικευμένα γραμμικά μοντέλα. Σε ένα γενικευμένο γραμμικό μοντέλο δεν μπορούμε, συνήθως, να βρούμε συζυγείς εκ των προτέρων κατανομές και οι δεσμευμένες εκ των υστέρων κατανομές δεν είναι εύκολα διαχειρίσιμες. Έστω δείγμα $\boldsymbol{y}^T = (y_1, y_2, \ldots, y_n)$ το οποίο προέρχεται από την εκθετική οικογένεια κατανομών με συνάρτηση πυκνότητας:

$$f(y_i|\theta_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\alpha(\phi_i)} + c(y_i, \phi_i)\right\}.$$

Η παράμετρος θ_i ονομάζεται κανονική παράμετρος (canonical parameter) και ισχύει ότι $\mu_i = \mathrm{E}(y_i|\theta) = b'(\theta_i)$, $\mathrm{Var}(y_i) = \alpha(\phi_i)b''(\theta_i)$. Η παράμετρος ϕ_i, η οποία είναι παράμετρος διασποράς, θα θεωρήσουμε ότι είναι γνωστή. Σε κάθε παρατήρηση αντιστοιχεί ένα διάνυσμα επεξηγηματικών μεταβλητών $\mathbf{x}_i$ και $\boldsymbol{\beta}$ είναι το διάνυσμα των αγνώστων παραμέτρων. Η μέση τιμή $\mu_i = \mathrm{E}(y_i)$ θα εξαρτάται από το linear predictor, $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$, μέσω της συνδετικής συνάρτησης (link function) $g(\mu_i) = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$. Ακολουθώντας τους McCullagh & Nelder (1989) ορίζουμε μετασχηματισμένες παρατηρήσεις $\tilde{\boldsymbol{y}}(\boldsymbol{\beta})$ και διαγώνιο πίνακα βαρών $\boldsymbol{W}(\boldsymbol{\beta})$ ως εξής:

$$\tilde{y}_i(\boldsymbol{\beta}) = \eta_i + (y_i - \mu_i)g'(\mu_i), \quad W_i(\boldsymbol{\beta}) = \frac{1}{\mathrm{Var}(y_i)(g'(\mu_i))^2},\ i = 1, 2, \ldots, n.$$

Ο αλγόριθμος σταθμισμένων ελαχίστων τετραγώνων (IWLS) ξεκινάει από μια αρχική τιμή $\boldsymbol{\beta}^{(0)}$ και επαναληπτικά θέτει ως $\boldsymbol{\beta}^{(t)}$, $t = 1, \ldots$, το γενικευμένο εκτιμητή ελαχίστων τετραγώνων του γραμμικού μοντέλου $\tilde{\boldsymbol{y}}(\boldsymbol{\beta}^{(t-1)}) \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{W}^{-1}(\boldsymbol{\beta}^{(t-1)}))$

$$\boldsymbol{\beta}^{(t)} = \left(\mathbf{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(t-1)})\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(t-1)})\tilde{\boldsymbol{y}}(\boldsymbol{\beta}^{(t-1)}).$$

Αφού επιτευχθεί σύγκλιση, η τελική εκτίμηση $\widehat{\boldsymbol{\beta}}$ αντιστοιχεί στη τιμή που μεγιστοποιεί την πιθανοφάνεια και $(\mathbf{X}^T\boldsymbol{W}(\widehat{\boldsymbol{\beta}})\mathbf{X})^{-1}$ είναι ο ασυμπτωτικός πίνακας συνδιακύμανσης του εκτιμητή μέγιστης πιθανοφάνειας. Οι παραπάνω εκτιμήσεις αντιστοιχούν στην κορυφή και στον αντίστροφο πίνακα κυρτότητας της εκ των υστέρων κατανομής του $\boldsymbol{\beta}$ αν η εκ των προτέρων κατανομή είναι μη γνήσια ($f(\boldsymbol{\beta}) \sim 1$) και υποθέσουμε ασυμπτωτική κανονικότητα. Αν θεωρήσουμε μια γνήσια εκ των προτέρων κατανομή για το $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_0, \boldsymbol{C}_0)$, τότε η 'εκ των υστέρων'

κατανομή του $\boldsymbol{\beta}$ για το μετασχηματισμένο μοντέλο πιθανοφάνειας θα είναι:

$$f(\boldsymbol{\beta}|\tilde{\boldsymbol{y}}(\boldsymbol{\beta}^{(t-1)})) \propto f(\tilde{\boldsymbol{y}}(\boldsymbol{\beta}^{(t-1)})|\boldsymbol{\beta})f(\boldsymbol{\beta}) \propto \exp\left\{-\frac{1}{2}\Big(\boldsymbol{\beta}^T\big(\boldsymbol{C}_1^{(t)}\big)^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\big(\boldsymbol{C}_1^{(t)}\big)^{-1}\boldsymbol{\beta}_1^{(t)}\Big)\right\},$$

άρα η 'εκ των υστέρων' κατανομή, θεωρώντας ότι τα βάρη είναι γνωστά, είναι $\boldsymbol{\beta}|\tilde{\boldsymbol{y}}(\boldsymbol{\beta}^{(t-1)}) \sim N(\boldsymbol{\beta}^{(t)}, \boldsymbol{C}_1^{(t)})$ όπου:

$$\boldsymbol{C}_1^{(t)} = \Big(\boldsymbol{C}_0^{-1} + \mathbf{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(t-1)})\mathbf{X}\Big)^{-1},\ \boldsymbol{\beta}^{(t)} = \boldsymbol{C}_1^{(t)}\Big(\boldsymbol{C}_0^{-1}\boldsymbol{\mu}_0 + \mathbf{X}^T\boldsymbol{W}(\boldsymbol{\beta}^{(t-1)})\tilde{\boldsymbol{y}}(\boldsymbol{\beta}^{(t-1)})\Big). \tag{3.2}$$

Εάν επαναλάβουμε την (3.2) αρκετές φορές θα βρούμε την κορυφή της εκ των υστέρων κατανομής του $\boldsymbol{\beta}$. Η ιδέα του Gamerman (1997) είναι να χρησιμοποίησουμε μονο μια επανάληψη του αλγορίθμου (3.2) και να προτείνουμε μια τιμή από την πολυδιάστατη κανονική κατανομή με την αντίστοιχη μέση τιμή και πίνακα συνδιακύμανσης, δηλαδή αν υποθέσουμε ότι η τρέχουσα τιμή της αλυσίδας είναι $\boldsymbol{\beta}$, τότε $\boldsymbol{\beta}^{can} \sim N(\boldsymbol{\mu}_1(\boldsymbol{\beta}), \boldsymbol{C}_1(\boldsymbol{\beta}))$ όπου

$$\boldsymbol{C}_1(\boldsymbol{\beta}) = \big(\boldsymbol{C}_0^{-1} + \mathbf{X}^T\boldsymbol{W}(\boldsymbol{\beta})\mathbf{X}\big)^{-1},\ \boldsymbol{\mu}_1(\boldsymbol{\beta}) = \boldsymbol{C}_1(\boldsymbol{\beta})\left(\boldsymbol{C}_0^{-1}\boldsymbol{\mu}_0 + \mathbf{X}^T\boldsymbol{W}(\boldsymbol{\beta})\tilde{\boldsymbol{y}}(\boldsymbol{\beta})\right). \tag{3.3}$$

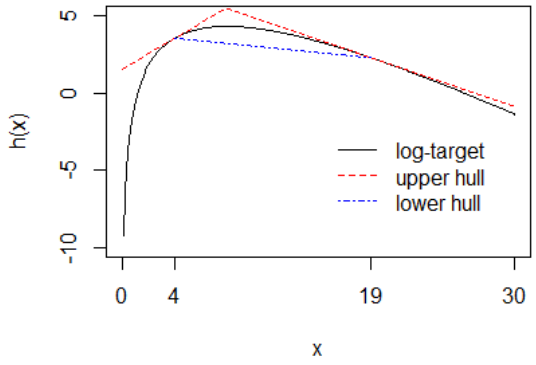
Η προτείνουσα του Gamerman καταφέρνει με μια επανάληψη του IWLS αλγορίθμου να βρίσκεται αρκετά κοντά στην πραγματική εκ των υστέρων κατανομή του $\boldsymbol{\beta}$, με συνέπεια να έχει υψηλό ποσοστό αποδοχής. Η πιθανότητα αποδοχής του αλγορίθμου δίνεται από τον τύπο:

$$p = \min\left\{1, \frac{f(\boldsymbol{\beta}^{can}|\boldsymbol{y})}{f(\boldsymbol{\beta}|\boldsymbol{y})}\frac{q(\boldsymbol{\beta}|\boldsymbol{\beta}^{can}, \boldsymbol{y})}{q(\boldsymbol{\beta}^{can}|\boldsymbol{\beta}, \boldsymbol{y})}\right\}.$$

3.3 Adaptive rejection sampling

Στην παρούσα ενότητα θα συζυτηθεί η μέθοδος προσομοίωσης Adaptive rejection sampling. Συμβολίζουμε ως f τη συνάρτηση πυκνότητας από την οποία θέλουμε να πάρουμε δείγμα. Ας υποθέσουμε ότι γνωρίζουμε τη συνάρτηση g η οποία είναι ανάλογη στην f, δηλαδή $f(x) = cg(x)$. Η μέθοδος μπορεί να χρησιμοποιηθεί μόνο στην περίπτωση όπου η πρώτη παράγωγος της $h(x) = \log g(x)$ είναι γνησιώς φθίνουσα συνάρτηση. Η μεθοδος είναι αποτελεσματική μόνο στην μονοδιαστατη περίπτωση (x στοιχείο). Για παράδειγμα, ας υποθέσουμε ότι $g(x) = x^4e^{-x/2},\ x > 0$ ($f \sim \text{gamma}(5, 1/2)$).

Αρχικά, χρειαζόμαστε ένα σύνολο από σημεία στα οποία ορίζεται η g, $T_k = \{x_1, x_2, \ldots, x_k\}$, με $x_1 < x_2 < \ldots < x_k$. Υπολογίζουμε τα $h(x_1) = \log g(x_1), \ldots, h(x_k) = \log g(x_k)$. Βασιζόμενοι σε αυτά τα σημεία ορίζουμε μια συνάρτηση φάκελο για την g ως $\exp\{u_k(x)\}$, όπου $u_k(x)$ είναι μια τμηματικά γραμμική συνάρτηση η οποία ορίζεται από τις εφαπτομένες της $h(x)$ στα σημεία που περιέχονται στο T_k, όπως φαίνεται στο γράφημα 3.1 με τις κόκκινες διακεκομμένες



Σχήμα 3.1: Αρχικό βημα του adaptive rejection sampling. Για δυο σημεία, $T_k = \{4, 19\}$, υπολογίζονται οι συναρτήσεις $u_k(x)$ και $l_k(x)$.

γραμμές ($T_2 = \{4, 19\}$). Οι εφαπτομένες τέμνονται σε ένα σημείο z_j μεταξύ των x_j και x_{j+1}:

$$z_j = \frac{h(x_{j+1}) - h(x_j) - x_{j+1}h'(x_{j+1}) + x_j h'(x_j)}{h'(x_j) - h'(x_{j+1})}.$$

Επομένως, για $x \in [z_{j-1}, z_j]$ $(j = 1, 2, \ldots, k)$ η u_k θα είναι ίση με:

$$u_k(x) = h(x_j) + (x - x_j)h'(x_j).$$

Γνωρίζουμε ότι κάθε θετική συνάρτηση μπορεί να μετασχηματιστεί σε συνάρτηση πυκνότητας πιθανότητας, αρκεί να έχει πεπερασμένο ολοκλήρωμα. Η συνάρτηση πυκνότητας πιθανότητας η οποία αντιστοιχεί στην u_k είναι:

$$f_k(x) = \frac{\exp\{u_k(x)\}}{\int_D \exp\{u_k(s)\}ds}, \tag{3.4}$$

όπου D είναι το πεδίο ορισμού της f. Το z_0 θα είναι το κάτω όριο του D και το z_k το άνω όριο του D. Επίσης, θα ορίσουμε μια τμηματικά γραμμική συνάρτηση $l_k(x)$ χρησιμοποιώντας τις χορδές ανάμεσα στα διαδοχικά σημεία του T_k. Επομένως, για $x \in [x_j, x_{j+1}]$ ισχύει ότι:

$$l_k(x) = \frac{(x_{j+1} - x)h(x_j) + (x - x_j)h(x_{j+1})}{x_{j+1} - x_j},$$

η οποία φαίνεται στο γράφημα 3.1 με την μπλέ διακεκομμένη γραμμή. Η h είναι κοίλη συνάρτηση, άρα ισχύει πάντα ότι $l_k(x) \leq h(x) \leq u_k(x)$, $\forall\, x \in D$. Το adaptive rejection sampling αποτελεί τροποποίηση της μεθόδου αποδοχής/απόρριψης. Προσομοιώνουμε μια τιμή $x^\star$ από την f_k (3.4) με τη μέθοδο αντιστροφής και μια τιμή u από την ομοιόμορφη κατανομή στο $(0, 1)$. Δοθέντων των τιμών $x^\star$ και u, χρησιμοποιούμε ένα test για να αποφασίσουμε αν χρειάζεται ο υπολογισμός

του $h(x^\star)$. Επομένως, αν

$$u \leq \exp\{l_k(x^\star) - u_k(x^\star)\}, \tag{3.5}$$

αποδεχόμαστε την $x^\star$ ως δείγμα από την f, χωρίς τον υπολογισμό του $h(x^\star)$. Αυτό σημαίνει ότι η h προσεγγίζεται ικανοποιητικά από τις συναρτήσεις u_k και l_k σε μια περιοχή πολύ κοντά στο $x^\star$. Αν δεν ισχύει η (3.5), αποδεχόμαστε το $x^\star$ αν $u \leq \exp\{h(x^\star) - u_k(x^\star)\}$, αλλιώς απορρίπτουμε την $x^\star$ ως τιμή από την f και προσομοιώνουμε ξανά από την f_k. Όταν απαιτείται ο υπολογισμός του $h(x^\star)$, προχωράμε στο 'adaptive' μέρος της μεθόδου. Συμπεριλαμβάνουμε το $x^\star$ στο T_k καταλήγοντας στο T_{k+1} και επαναδιατάσσουμε το T_{k+1} έτσι ώστε $x_1 < x_2 < \ldots < x_{k+1}$. Στη συνέχεια, υπολογίζουμε τα f_{k+1}, u_{k+1} και l_{k+1}. Τα βήματα του adaptive rejection sampling περιγράφονται συνοπτικά στον αλγόριθμο (3.3).

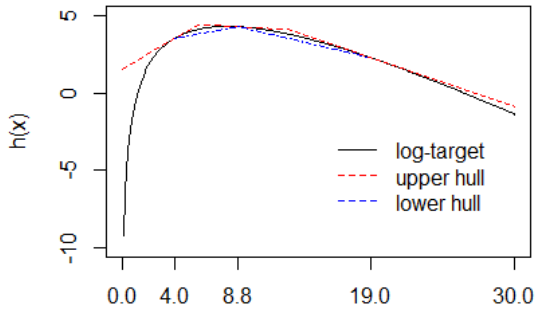
Αλγόριθμος 3.3. *Adaptive rejection sampling*

- Επιλέγουμε k σημεία στο T_k. Εαν το κάτω όριο του D είναι $-\infty$, τότε πρέπει $h'(x_1) > 0$. Εάν το άνω όριο του D είναι $+\infty$, πρέπει $h'(x_k) < 0$. Υπολογίζουμε τα f_k, u_k και l_k.
- Προσομοιώνουμε $x^\star$ από την f_k και $u \sim U(0,1)$. Αν $u \leq \exp\{l_k(x^\star) - u_k(x^\star)\}$ αποδεχόμαστε το $x^\star$. Αλλιώς υπολογίζουμε τα $h(x^\star)$ και $h'(x^\star)$. Αν $u \leq \exp\{h(x^\star) - u_k(x^\star)\}$ αποδεχόμαστε το $x^\star$, αλλιώς το απορρίπτουμε.
- Αν υπολογίσαμε τα $h(x^\star)$ και $h'(x^\star)$ στο προηγούμενο βήμα τότε εκχωρούμε το $x^\star$ στο T_k καταλήγοντας στο T_{k+1}. Στη συνέχεια θέτουμε τα στοιχεία του T_{k+1} σε αύξουσα σειρά και υπολογίζουμε τα f_{k+1}, u_{k+1} και l_{k+1}.

Ας υποθέσουμε ότι η τυχαία τιμή $x^\star$ από την f_k είναι 8.875. Η τιμή $x^\star$ δεν γίνεται αποδεκτή σύμφωνα με το test $u \leq \exp\{l_k(x^\star) - u_k(x^\star)\}$. Άρα, υπολογίζουμε τις συναρτήσεις u_{k+1} και l_{k+1}. Σύμφωνα με το γράφημα (3.2) οι νέες συναρτήσεις u_{k+1} και l_{k+1} βρίσκονται πιο κοντά στη h από ότι οι προηγούμενες u_k και l_k, οι οποίες βασίζονται μόνο σε δυο σημεία. Επομένως, η πιθανότητα αποδοχής της επόμενης τιμής χωρίς τον υπολογισμό της h αυξάνεται.

Θα δείξουμε ότι ο αλγόριθμος 3.3 οδηγεί σε ένα τυχαίο δείγμα από την f (Duchateau & Janssen 2007, σελ. 253). Διαλέγουμε ένα αρχικό σύνολο k σημείων $T_k = \{x_1, x_2, \ldots, x_k\}$. Έστω $x^{(k)}$ η τιμή της τυχαίας μεταβλητής $X^{(k)}$, η οποία έχει συνάρτηση πυκνότητας f_k (3.4) και H_k η πληροφορία σχετικά με τις συναρτήσεις u_k και l_k. Συμβολίζουμε ως δ_k μια δείκτρια συνάρτηση η οποία παίρνει την τιμή 1 αν το $x^{(k)}$ έγινε αποδεκτό και 0 αλλιώς. Δοθέντος του H_k η απο κοινου πιθανότητα η $X^{(k)}$ να πάρει την τιμή $x^{(k)}$ και η δ_k να πάρει την τιμή 1 θα είναι:

$$\Pr(X^{(k)} = x^{(k)}, \delta_k = 1 | H_k) = \Pr(\delta_k = 1 | X^{(k)} = x^{(k)}, H_k) f_k(x^{(k)}).$$



Σχήμα 3.2: Ανανέωση του άνω και κάτω φράγματος της h. Η τιμή $x^\star = 8.875$ δεν έγινε δεκτή σύμφωνα με το test $u \leq \exp\{l_k(x^\star) - u_k(x^\star)\}$. Οι νέες συναρτήσεις u_{k+1} και l_{k+1} βρίσκονται πιο κοντά στη h από ότι οι u_k και l_k

Το δ_k είναι ίσο με 1 αν ένα από τα ενδεχόμενα A ή B είναι αληθές, όπου το ενδεχόμενο A είναι το $A = \left[U \leq \exp\left\{l_k(x^{(k)}) - u_k(x^{(k)})\right\}\right]$ και το $B = \left[U \leq \exp\left\{h(x^{(k)}) - u_k(x^{(k)})\right\}\right]$, με $U \sim U(0,1)$. Ισχύει ότι $A \subseteq B$. Άρα, $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) = \Pr(B)$. Επομένως, ισχύει ότι:

$$\Pr(X^{(k)} = x^{(k)}, \delta_k = 1 | H_k) = \exp\left\{h(x^{(k)}) - u_k(x^{(k)})\right\} f_k(x^{(k)}) = \frac{\exp\left\{h(x^{(k)})\right\}}{\int_D \exp\{u_k(s)\} ds}. \quad (3.6)$$

Ενδιαφερόμαστε να δείξουμε ότι η κατανομή των αποδεκτών $x^{(k)}$ είναι η f. Ισχύει ότι:

$$\Pr(X^{(k)} = x^{(k)} | \delta_k = 1, H_k) = \frac{\Pr(X^{(k)} = x^{(k)}, \delta_k = 1 | H_k)}{\Pr(\delta_k = 1 | H_k)}. \quad (3.7)$$

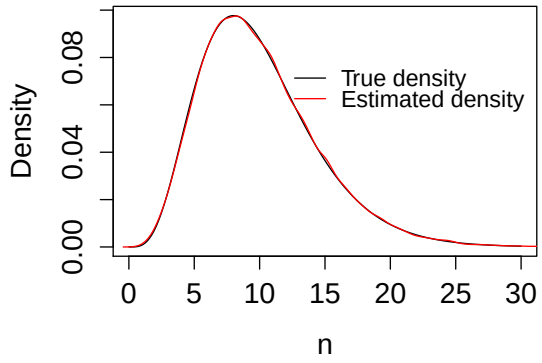
Όμως,

$$\Pr(\delta_k = 1 | H_k) = \int_D \Pr(X^{(k)} = x, \delta_k = 1 | H_k) dx =^{(3.6)} \frac{\int_D \exp\{h(x)\} dx}{\int_D \exp\{u_k(s)\} ds}. \quad (3.8)$$

Αν αντικαταστήσουμε τον αριθμητή και παρανομαστή της (3.7) με τις σχέσεις (3.6) και (3.8),

$$\Pr(X^{(k)} = x^{(k)} | \delta_k = 1, H_k) = \frac{\exp\left\{h(x^{(k)})\right\}}{\int_D \exp\{h(x)\} dx} = f(x^{(k)}).$$

Επομένως, η κατανομή των αποδεκτών $x^{(k)}$ είναι η f, ανεξάρτητα από το H_k.



Σχήμα 3.3: Προσομοίωση 60000 τιμών από την κατανομή gamma$(5, 0.5)$ με adaptive rejection sampling. Τα αρχικά σημεία εφαρμογής ήταν $T_k = \{4, 19\}$. Το ποσοστό αποδοχής ήταν περίπου 99.9%.

3.3.1 Ευθεία Προσομοίωση Από τη Συνάρτηση Φάκελο f_k

Θα δείξουμε ότι μπορούμε να πάρουμε δείγμα από την f_k με τη μέθοδο της αντιστροφής (ευθεία προσομοίωση). Επομένως, πρέπει να βρούμε τη συνάρτηση κατανομής της $X^{(k)}$, όπου $X^{(k)} \sim f_k$. Ισχύει ότι:

$$\begin{aligned} c_j &= \int_{z_{j-1}}^{z_j} \exp\{u_k(s)\}ds = \exp\left\{h(x_j) - x_j h'(x_j)\right\} \int_{z_{j-1}}^{z_j} \exp\left\{s h'(x_j)\right\} ds \\ &= \exp\left\{h(x_j) - x_j h'(x_j)\right\} \frac{1}{h'(x_j)} \left(\exp\left\{z_j h'(x_j)\right\} - \exp\left\{z_{j-1} h'(x_j)\right\}\right), \ j = 1, 2, \ldots, k. \end{aligned}$$

Συμβολίζουμε ως $int(z_{j-1}) = \int_{-\infty}^{z_{j-1}} \exp\{u_k(s)\}ds = \sum_{k=1}^{j-1} c_k$. Έστω ότι:

$$f_k(x) = \exp\{u_k(x)\}/c,$$

δηλαδή $c = \sum_{j=1}^{k} c_j$ είναι η σταθερά κανονικοποίησης της f_k. Για $x \in [z_{j-1}, z_j]$ ισχύει ότι η συνάρτηση κατανομής της $X^{(k)}$ θα είναι:

$$\begin{aligned} F_k(x) = \Pr(X^{(k)} \le x) &= \int_{-\infty}^{x} f_k(s)ds = \frac{int(z_{j-1})}{c} + \frac{1}{c} \int_{z_{j-1}}^{x} \exp\{u_k(s)\}ds \qquad (3.9) \\ &= \frac{int(z_{j-1})}{c} + \frac{1}{h'(x_j)c} \exp\left\{h(x_j) - x_j h'(x_j)\right\} \left(\exp\left\{x h'(x_j)\right\} - \exp\left\{z_{j-1} h'(x_j)\right\}\right). \end{aligned}$$

Γνωρίζουμε από γνωστό θεώρημα ότι $u = F_k(x) \sim U(0, 1)$. Άρα, πρέπει να λύσουμε την εξίσωση (3.9) ως προς x. Άρα,

$$cu - int(z_{j-1}) = \frac{1}{h'(x_j)} \exp\left\{h(x_j) - x_j h'(x_j)\right\} \left(\exp\left\{x h'(x_j)\right\} - \exp\left\{z_{j-1} h'(x_j)\right\}\right) \ \Rightarrow$$

$$\exp\left\{xh'(x_j)\right\} = \exp\left\{z_{j-1}h'(x_j)\right\} + h'(x_j)\exp\left\{-\left[h(x_j) - x_jh'(x_j)\right]\right\}(cu - int(z_{j-1})) \Rightarrow$$
$$x = \frac{1}{h'(x_j)}\log\left[\exp\left\{z_{j-1}h'(x_j)\right\} + h'(x_j)\exp\left\{-\left[h(x_j) - x_jh'(x_j)\right]\right\}(cu - int(z_{j-1}))\right] \tag{3.10}$$

Για να μπορέσουμε να προσομοιώσουμε από την f_k μέσω της (3.10), πρέπει να θέσουμε τη συνθήκη $z_{j-1} \leq x \leq z_j$. Η τυχαία μεταβλητή u η οποία προσομοιώθηκε από την ομοιόμορφη στο $(0,1)$ θα μας οδηγήσει στο διάστημα στο οποίο θα ανήκει το x. Για την αριστερή συνθήκη έχουμε ότι:

$$x \geq z_{j-1} \;\Rightarrow\; \exp\left\{xh'(x_j)\right\} \geq \exp\left\{z_{j-1}h'(x_j)\right\}, \Rightarrow$$
$$h'(x_j)\exp\left\{-\left[h(x_j) - x_jh'(x_j)\right]\right\}(cu - int(z_{j-1})) \geq 0, \Rightarrow u \geq \frac{int(z_{j-1})}{c}.$$

Πρέπει να τονιστεί ότι αν $h'(x_j) < 0$, δεν χρειάζεται να αλλάξει η φορά της εξίσωσης, γιατί θα αλλάξει πάλι μετά. Για τη δεξιά συνθήκη έχουμε ότι:

$$x \leq z_j \;\Rightarrow\; \exp\left\{xh'(x_j)\right\} \leq \exp\left\{z_jh'(x_j)\right\}, \Rightarrow$$
$$(cu - int(z_{j-1})) \leq \underbrace{\exp\left\{h(x_j) - x_jh'(x_j)\right\}\frac{1}{h'(x_j)}\left(\exp\left\{z_jh'(x_j)\right\} - \exp\left\{z_{j-1}h'(x_j)\right\}\right)}_{c_j}$$
$$cu \leq int(z_{j-1}) + c_j = int(z_j), \Rightarrow u \leq \frac{int(z_j)}{c}.$$

3.4 Μπεϋζιανή Σύγκριση Μοντέλου

Στην παρούσα ενότητα θα εξετάσουμε το πρόβλημα της αβεβαιότητας ως προς τη μορφή του μοντέλου. Σε πραγματικά προβλήματα, συναντάμε αρκετά μοντέλα τα οποία περιέχουν διαφορετικές υποθέσεις, επεξηγηματικές μεταβλητές ή άλλες παραμέτρους. Ο πιο φυσικός τρόπος επιλογής μοντέλου στη Μπεϋζιανή στατιστική γίνεται μέσω της δήλωσης των εκ των υστέρων πιθανοτήτων των υποψήφιων μοντέλων.

Έστω ότι παρατηρούμε κάποια δεδομένα $\boldsymbol{y}$. Θα εξετάσουμε δυο υποψήφια παραμετρικά μοντέλα M_1 και M_2 με παραμέτρους $\boldsymbol{\theta}_1$ και $\boldsymbol{\theta}_2$, αντίστοιχα. Φυσικά, πρέπει να εισάγουμε εκ των προτέρων κατανομές για τις παραμέτρους των μοντέλων M_1 και M_2, οι οποίες συμβολίζονται με $f(\boldsymbol{\theta}_1|M_1)$ και $f(\boldsymbol{\theta}_2|M_2)$, αντίστοιχα. Επομένως, η περιθώρια πιθανοφάνεια των μοντέλων M_1 και M_2, δίνεται από τη σχέση:

$$f(\boldsymbol{y}|M_i) = \int f(\boldsymbol{y}|\boldsymbol{\theta}_i, M_i) f(\boldsymbol{\theta}_i|M_i) d\boldsymbol{\theta}_i, \ i = 1, 2,$$

αντίστοιχα. Επίσης, εκφράζουμε την πεποίθησή μας ως προς το ποιο είναι το σωστό μοντέλο, μέσω των εκ των προτέρων πιθανοτήτων, $\Pr(M_1)$ και $\Pr(M_2)$, των αντίστοιχων μοντέλων ($\Pr(M_1) + \Pr(M_2) = 1$). Η εκ των υστέρων πιθανότητα του μοντέλου M_1, υπολογίζεται μέσω του θεωρήματος Bayes, ως εξής:

$$\Pr(M_1|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|M_1)\Pr(M_1)}{f(\boldsymbol{y})} = \frac{f(\boldsymbol{y}|M_1)\Pr(M_1)}{\sum_{i=1}^{2} f(\boldsymbol{y}|M_i)\Pr(M_i)} = 1 - \Pr(M_2|\boldsymbol{y}).$$

Είνα σαφές ότι, οι εκ των υστέρων πιθανότητες των μοντέλων εξαρτώνται από τις εκ των προτέρων πιθανότητες των μοντέλων. Εναλλακτικά, υπολογίζουμε τον παράγοντα Bayes (Bayes Factor), ο οποίος ορίζεται ως το πηλίκο των περιθώριων πιθανοφανειών των μοντέλων, δηλαδή:

$$BF = \frac{f(\boldsymbol{y}|M_1)}{f(\boldsymbol{y}|M_2)} = \frac{\Pr(M_1|\boldsymbol{y})/\Pr(M_2|\boldsymbol{y})}{\Pr(M_1)/\Pr(M_2)}.$$

Ο παράγοντας Bayes είναι ένα μέτρο της πληροφορίας που περιέχεται στα δεδομένα υπέρ του μοντέλου M_1 έναντι του μοντέλου M_2. Η παρούσα μεθοδολογία μπορεί να επεκταθεί στη σύγκριση $k \geq 2$ μοντέλων, ταυτόχρονα. Επίσης, τα υπο εξέταση μοντέλα δεν χρειάζεται να είναι εμφωλευμένα (nested). Πρέπει, όμως, να τονιστεί ότι ο παράγοντας Bayes επηρεάζεται από τη μεταβλητότητα της εκ των προτέρων κατανομής των παραμέτρων. Επίσης, στα περισσότερα προβλήματα, η περιθώρια πιθανοφάνεια δεν μπορεί να βρεθεί σε κλειστή μορφή, άρα, ο υπολογισμός του παράγοντα Bayes καθίσταται αρκετά δύσκολος. Στην ενότητα 4.4, θα εξετάσουμε το το κριτήριο DIC (Deviance Information Criterion) το οποίο μπορεί να υπολογιστεί με ευκολία από μια MCMC αλυσίδα.

Κεφάλαιο 4

MCMC Προσέγγιση σε Shared Frailty Μοντέλα

Τα frailty μοντέλα είναι ιδιαίτερα δημοφιλή στην ανάλυση επιβίωσης επειδή επιτρέπουν τη μοντελοποίηση της συσχέτισης ατόμων που ανήκουν στην ίδια ομάδα (cluster). Στη στατιστική βιβλιογραφία, διάφορες μέθοδοι έχουν χρησιμοποιηθεί για την εκτίμηση των παραμέτρων ενός frailty μοντέλου. Οι μέθοδοι περιλαμβάνουν την εκτίμηση μέγιστης πιθανοφάνειας (για παραμετρικά μοντέλα), τον αλγόριθμο EM, μεθόδους που στηρίζονται στην, επι ποινή, μερική πιθανοφάνεια (penalized partial likelihood) και αλγόριθμους Markov Chain Monte Carlo (MCMC). Οι frailty κατανομές που θα εξετάσουμε είναι η Gamma και η Lognormal. Οι κατανομές που θα χρησιμοποιηθούν για τη συνάρτηση κινδύνου είναι η Weibull και η κατά τμήματα εκθετική κατανομή. Στο παρόν κεφάλαιο θα εξεταστούν αναλυτικά MCMC μέθοδοι οι οποίες περιλαμβάνουν τον αλγόριθμο Metropolis–Hastings, ο οποίος βασίζεται στην προτείνουσα σταθμισμένων ελαχίστων τετραγώνων (Gamerman 1997), και το δειγματολήπτη Gibbs (adaptive rejection sampling των Gilks & Wild 1992). Επίσης, θα θεωρήσουμε το κριτήριο DIC (Deviance Information Criterion) ως ένα μέτρο επιλογής μοντέλου. Αναλυτική παρουσίαση μπεϋζιανών μεθόδων στην ανάλυση επιβίωσης δίνεται από τους Ibrahim, Chen & Sinha (2001).

4.1 Γενικό Πλαίσιο και Υποθέσεις

Ας υποθέσουμε ότι το δείγμα αποτελείται από G ομάδες και η i-οστή ομάδα περιέχει n_i υποκείμενα ($i = 1, 2, \ldots, G$). Άρα, το συνολικό μέγεθος δείγματος είναι $n = \sum_{i=1}^{G} n_i$. Έστω ότι T_{ij} είναι ο χρόνος επιβίωσης και C_{ij} ο χρόνος λογοκρισίας του j-οστού υποκειμένου από την i-οστή ομάδα. Θα υποθέσουμε μη πληροροφοριακή δεξιά λογοκρισία (noninformative right censoring), άρα, ο χρόνος παρατήρησης και η δείκτρια συνάρτηση αποτυχίας (failure indicator) θα είναι $Y_{ij} = \min\{T_{ij}, C_{ij}\}$ και $\Delta_{ij} = \mathbb{1}(T_{ij} \leq C_{ij})$, αντίστοιχα. Σε κάθε υποκείμενο αντιστοιχεί ένα διάνυσμα από p επεξηγηματικές μεταβλητές $\mathbf{x}_{ij}^T = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})$. Ο πίνακας σχεδιασμού για την i-οστή ομάδα είναι $\mathbf{x}_i^T = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{in_i})$. Τα παρατηρούμενα δεδομένα αποτελούνται από τις τριπλέτες $(y_{ij}, \delta_{ij}, \mathbf{x}_{ij}^T)$, $i = 1, 2, \ldots, G,\ j = 1, 2, \ldots, n_i$. Το μοντέλο υποθέτει ότι τα υποκείμενα από την ίδια ομάδα μοιράζονται την ίδια λανθάνουσα μεταβλητή (frailty), και δεδομένης αυτής της μεταβλητής οι χρόνοι επιβίωσης είναι ανεξάρτητοι. Συμπερασματικά, οι χρόνοι επιβίωσης ατόμων της ίδιας ομάδας είναι συσχετισμένοι (θετικά) μεταξύ τους. Η συνάρτηση κινδύνου για το j-οστό υποκείμενο της i-οστής ομάδας δίνεται από

$$\lambda(t|\mathbf{x}_{ij}, u_i) = \lambda_0(t) e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}} u_i,\ i = 1, 2, \ldots, G,\ j = 1, 2, \ldots, n_i, \tag{4.1}$$

όπου $\lambda_0(t)$ είναι η βασική συνάρτηση κινδύνου, $\boldsymbol{\beta}$ είναι το διάνυσμα των παραμέτρων και $u_1, u_2, \ldots, u_G$ είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές από κάποια κατανομή με μέση τιμή $\mathrm{E}(u) = 1$ και $\mathrm{Var}(u) = \sigma^2$. Η διακύμανση των frailties, σ^2, ειναι και αυτή άγνωστη παράμετρος, άρα θα ακολουθεί κάποια εκ των προτέρων κατανομή. Μεγάλες τιμές του σ^2 υποδηλώνουν μεγάλη ετερογένεια μεταξύ των ομάδων, άρα ισχυρή συσχέτιση μέσα στις ομάδες. Η πλήρης πιθανοφάνεια των παραμέτρων $(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{u}, \sigma^2)$ είναι

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{u}, \sigma^2) &= \prod_{i=1}^{G} \prod_{j=1}^{n_i} \left(\lambda(y_{ij}|\mathbf{x}_{ij}, u_i)\right)^{\delta_{ij}} S(y_{ij}|\mathbf{x}_{ij}, u_i) \\ &= \prod_{i=1}^{G} \prod_{j=1}^{n_i} \left(\lambda_0(y_{ij}) e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}} u_i\right)^{\delta_{ij}} \exp\left\{-\Lambda_0(y_{ij}) e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}} u_i\right\}, \end{aligned} \tag{4.2}$$

όπου με $\boldsymbol{\psi}$ συμβολίζουμε τις παραμέτρους που συσχετίζονται με τη βασική συνάρτηση κινδύνου. Θα υποθέσουμε ότι οι παράμετροι είναι εκ των προτέρων ανεξάρτητες (εκτός από $\boldsymbol{u}$ και σ^2). Συνδυάζοντας την πιθανοφάνεια (4.2) με τις prior κατανομές των παραμέτρων, μέσω του θεωρήματος Bayes, καταλήγουμε στην απο κοινού εκ των υστέρων κατανομή των παραμέτρων

$$f(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{u}, \sigma^2|\boldsymbol{y}, \boldsymbol{\delta}) \propto \prod_{i=1}^{G} \prod_{j=1}^{n_i} \left(\lambda_0(y_{ij}) e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}} u_i\right)^{\delta_{ij}} e^{-\Lambda_0(y_{ij}) e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}} u_i} f(\boldsymbol{\beta}) f(\boldsymbol{\psi}) f(\boldsymbol{u}|\sigma^2) f(\sigma^2).$$

Παρακάτω θα παρουσιαστεί η προσαρμογή της προτείνουσας του Gamerman (3.3) για την προσομοίωση του διανύσματος $\boldsymbol{\beta}$ σε δεδομένα επιβίωσης με δεξιά λογοκρισία. Η ίδια μεθοδολογία χρησιμοποίειται σε όλους τις περιπτώσεις που θα εξετάσουμε, γι' αυτό παρουσιάζεται στο γενικό της πλαίσιο. Θεωρώντας μια $N(\boldsymbol{\mu}_0, \boldsymbol{C}_0)$ εκ των προτέρων κατανομή για το $\boldsymbol{\beta}$, η πλήρως δεσμευμένη εκ των υστέρων κατανομή του $\boldsymbol{\beta}$ θα είναι

$$f(\boldsymbol{\beta}|\boldsymbol{\psi}, \boldsymbol{u}, \sigma^2, \boldsymbol{y}, \boldsymbol{\delta}) \propto \exp\left\{\sum_{i=1}^{G}\sum_{j=1}^{n_i}\left[\delta_{ij}\mathbf{x}_{ij}^T\boldsymbol{\beta} - \Lambda_0(y_{ij})e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}u_i\right] - \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)^T\mathbf{C}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)\right\}. \tag{4.3}$$

Ξεκάθαρα, η (4.3) είναι η εκ των υστέρων κατανομή του $\boldsymbol{\beta}$ η οποία προκύπτει από ένα 'μοντέλο' $\delta_{ij}|\boldsymbol{\psi}, u_i \sim \text{Poisson}(e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\Lambda_0(y_{ij})u_i)$, με μέση τιμή $\mu_{ij} = \text{E}(\delta_{ij}|\boldsymbol{\psi}, u_i) = e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\Lambda_0(y_{ij})u_i$. Επομένως, μπορούμε να χρησιμοποιήσουμε τον αλγόριθμο του Gamerman όπως περιγράφτηκε στην ενοτητα (3.2.3). Η συνδετική συνάρτηση είναι $g(\mu_{ij}) = \log(\mu_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \log(\Lambda_0(y_{ij})u_i)$, όπου το $\log(\Lambda_0(y_{ij})u_i)$ είναι το offset του γενικευμένου γραμμικού μοντέλου. Ορίζουμε διάνυσμα μετασχηματισμένων παρατηρήσεων $\tilde{\boldsymbol{\delta}}(\boldsymbol{\beta})$ και διαγώνιο πίνακα βαρών $\boldsymbol{W}(\boldsymbol{\beta})$, ως εξής

$$\tilde{\delta}_{ij}(\boldsymbol{\beta}) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + (\delta_{ij} - \mu_{ij})g'(\mu_{ij}),\ W_{ij}(\boldsymbol{\beta}) = \frac{1}{g'(\mu_{ij})}.$$

Πρέπει να τονιστεί ότι έχουμε αφαιρέσει το offset από το linear predictor και ότι στην Poisson κατανομή ισχύει $\text{Var}(\delta_{ij})(g'(\mu_{ij}))^2 = g'(\mu_{ij})$. Αν $\boldsymbol{\beta}$ είναι η τρέχουσα τιμή του αλγορίθμου, τότε προτείνουμε μια τιμή $\boldsymbol{\beta}^{can} \sim N(\boldsymbol{\mu}_1(\boldsymbol{\beta}), \boldsymbol{C}_1(\boldsymbol{\beta}))$ όπου

$$\boldsymbol{C}_1(\boldsymbol{\beta}) = (\boldsymbol{C}_0^{-1} + \mathbf{X}^T\boldsymbol{W}(\boldsymbol{\beta})\mathbf{X})^{-1},\ \boldsymbol{\mu}_1(\boldsymbol{\beta}) = \boldsymbol{C}_1(\boldsymbol{\beta})(\boldsymbol{C}_0^{-1}\boldsymbol{\mu}_0 + \mathbf{X}^T\boldsymbol{W}(\boldsymbol{\beta})\tilde{\boldsymbol{\delta}}(\boldsymbol{\beta})), \tag{4.4}$$

όπου με $\mathbf{X}$ συμβολίζουμε τον πίνακα σχεδιασμού όλου του δείγματος. Η πιθανότητα αποδοχής της νέας τιμή για το $\boldsymbol{\beta}$ δίνεται από:

$$p = \min\left\{1, p_1 = \frac{f(\boldsymbol{\beta}^{can}|\boldsymbol{\psi}, \boldsymbol{u}, \boldsymbol{y}, \boldsymbol{\delta})}{f(\boldsymbol{\beta}|\boldsymbol{\psi}, \boldsymbol{u}, \boldsymbol{y}, \boldsymbol{\delta})}\frac{q(\boldsymbol{\beta}|\boldsymbol{\beta}^{can}, \boldsymbol{y}, \boldsymbol{\delta})}{q(\boldsymbol{\beta}^{can}|\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{\delta})}\right\}, \tag{4.5}$$

ενώ σε λογαριθμική κλίμακα:

$$\begin{aligned}
\log(p_1) = &\sum_{i=1}^{G}\sum_{j=1}^{n_i}\left[\delta_{ij}(\mathbf{x}_{ij}^T\boldsymbol{\beta}^{can} - \mathbf{x}_{ij}^T\boldsymbol{\beta}) - \Lambda_0(y_{ij})u_i(e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}^{can}} - e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}})\right] \\
&- \frac{1}{2}(\boldsymbol{\beta}^{can} - \boldsymbol{\mu}_0)^T\mathbf{C}_0^{-1}(\boldsymbol{\beta}^{can} - \boldsymbol{\mu}_0) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)^T\mathbf{C}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_0) \\
&- \frac{1}{2}\log\left[\det(\boldsymbol{C}_1(\boldsymbol{\beta}^{can}))\right] - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_1(\boldsymbol{\beta}^{can}))^T\mathbf{C}_1^{-1}(\boldsymbol{\beta}^{can})(\boldsymbol{\beta} - \boldsymbol{\mu}_1(\boldsymbol{\beta}^{can})) \\
&+ \frac{1}{2}\log\left[\det(\boldsymbol{C}_1(\boldsymbol{\beta}))\right] + \frac{1}{2}(\boldsymbol{\beta}^{can} - \boldsymbol{\mu}_1(\boldsymbol{\beta}))^T\mathbf{C}_1^{-1}(\boldsymbol{\beta})(\boldsymbol{\beta}^{can} - \boldsymbol{\mu}_1(\boldsymbol{\beta})).
\end{aligned}$$

4.2 Gamma Frailty Μοντέλο

Για να οριστεί πλήρως το μοντέλο πρέπει να θεωρήσουμε μια κατανομή για τα frailties και τη συνάρτηση κινδύνου. Θα υποθέσουμε ότι τα frailties αποτελούν τυχαίο δείγμα από την Gamma κατανομή με μέση τιμή 1 και ακρίβεια ω ($\mathrm{E}(u_i) = 1, \mathrm{Var}(u_i) = \frac{1}{\omega}$), δηλαδή $u_i|\omega \sim \mathrm{Gamma}(\omega, \omega)$

$$f(u_i|\omega) = \frac{\omega^\omega}{\Gamma(\omega)} u_i^{\omega-1} \exp\{-\omega u_i\},\ i = 1, 2, \ldots, G.$$

Για την ακρίβεια ω θα υποθέσουμε επίσης μια Gamma κατανομή, $\omega \sim \mathrm{Gamma}(c, d)$. Συνήθως, επιλέγουμε μικρές τιμές για τις παραμέτρους (c, d) οι οποίες, στις περισσότερες περιπτώσεις, δεν έχουν ιδιαίτερη επίδραση στα αποτελέσματα. Προς το παρόν, δεν θα υποθέσουμε κάποια συγκεκριμένη παραμετρική μορφή για τη συνάρτηση κινδύνου, αφού μπορούμε να εξάγουμε τις δεσμευμένες κατανομές των παραμέτρων $(\boldsymbol{u}, \omega)$ ανεξάρτητα από τη μορφή του $\lambda_0(t)$. Η από κοινού εκ των υστέρων κατανομή των παραμέτρων θα είναι

$$\begin{aligned} f(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{u}, \omega|\boldsymbol{y}, \boldsymbol{\delta}) \propto & \prod_{i=1}^{G}\prod_{j=1}^{n_i} \left(\lambda_0(y_{ij}) e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}} u_i\right)^{\delta_{ij}} \exp\left\{-\Lambda_0(y_{ij}) e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}} u_i\right\} f(\boldsymbol{\beta}) f(\boldsymbol{\psi}) \\ & \times \prod_{i=1}^{G}\left[\frac{\omega^\omega}{\Gamma(\omega)} u_i^{\omega-1} \exp\{-\omega u_i\}\right] \times \omega^{c-1} \exp\{-d\omega\}. \end{aligned} \tag{4.6}$$

Δεσμευμένη κατανομή του $\boldsymbol{u}|\boldsymbol{\beta}, \boldsymbol{\psi}, \omega, \boldsymbol{y}, \boldsymbol{\delta}$

Έστω $\delta_i = \sum_{j=1}^{n_i} \delta_{ij}$ ο συνολικός αριθμός γεγονότων στην i-οστή ομάδα. Η δεσμευμένη, εκ των υστέρων, κατανομή του u_i δοθέντων των υπολοίπων μεταβλητών είναι

$$f(u_i|\boldsymbol{\beta}, \boldsymbol{\psi}, \omega, \boldsymbol{y}, \boldsymbol{\delta}) \propto u_i^{\omega+\delta_i-1} \exp\left\{-\big(\omega + \sum_{j=1}^{n_i} \Lambda_0(y_{ij}) e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\big) u_i\right\}, \tag{4.7}$$

άρα ισχύει ότι $u_i|\boldsymbol{\beta}, \boldsymbol{\psi}, \omega, \boldsymbol{y}, \boldsymbol{\delta} \sim \mathrm{Gamma}(\omega + \delta_i, \omega + \sum_{j=1}^{n_i} \Lambda_0(y_{ij}) e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}})$, $i = 1, 2, \ldots, G$. Παρατηρούμε ότι τα u_i, δοθέντων των υπολοίπων μεταβλητών, είναι εκ των υστέρων ανεξάρτητα μεταξύ τους. Αυτό το γεγονός διευκολύνει τη προσομοιώση τους.

Δεσμευμένη κατανομή του $\omega|\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{u}, \boldsymbol{y}, \boldsymbol{\delta}$

Η εκ των υστέρων, δεσμευμένη, κατανομή του ω δοθέντων των υπολοίπων μεταβλητών είναι

$$f(\omega|\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{u}, \boldsymbol{y}, \boldsymbol{\delta}) \propto \omega^{\omega G+c-1} \prod_{i=1}^{G} u_i^\omega \exp\left\{-(d + \sum_{i=1}^{G} u_i)\omega\right\} \frac{1}{\Gamma(\omega)^G}. \tag{4.8}$$

Είναι σαφές ότι το μοντέλο είναι ιεραρχικό, με την έννοια ότι η δεσμευμένη, εκ των υστέρων,

κατανομή του ω εξαρτάται μόνο από τα frailties. Η συνάρτηση (4.8) δεν θυμίζει κάποια γνωστή κατανομή. Μπορούμε όμως να εφαρμόσουμε Adaptive rejection sampling για την προσομοιώση της (4.8). Για την εφαρμογή της μεθόδου χρειαζόμαστε το λογάριθμο και την πρώτη παραγωγο του λογαρίθμου της (4.8), δηλαδή

$$h(\omega) = \log(f(\omega|\cdot)) = (\omega G + c - 1)\log(\omega) + \omega \sum_{i=1}^{G} \log(u_i) - \omega(d + \sum_{i=1}^{G} u_i) - G\log(\Gamma(\omega)),$$

$$h'(\omega) = G\log(\omega) + \frac{\omega G + c - 1}{\omega} + \sum_{i=1}^{G} \log(u_i) - (d + \sum_{i=1}^{G} u_i) - G\frac{\Gamma'(\omega)}{\Gamma(\omega)}.$$

Χρησιμοποιώντας τις ιδιότητες της δι-γάμμα συνάρτησης, $\Psi(x) = \dfrac{\Gamma'(x)}{\Gamma(x)}$, μπορεί να αποδειχθεί ότι ο λογάριθμος της κατανομής (4.8) είναι κοίλη συνάρτηση (Sahu, Dey, Aslanidou & Sinha 1997, Abramowitz, Stegun et al. 1965).

Θα δείξουμε ότι στην περίπτωση της Gamma κατανομής, οι τυχαίοι όροι $u_2, u_2, \ldots, u_G$ μπορούν να ολοκληρωθούν από την πιθανοφάνεια (4.2). Η συνεισφορά στην πιθανοφάνεια της i-οστής ομάδας είναι

$$L_i(\boldsymbol{\beta}, \boldsymbol{\psi}, u_i) = \prod_{j=1}^{n_i} \left(\lambda_0(y_{ij}) e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}} u_i\right)^{\delta_{ij}} \exp\left\{-\Lambda_0(y_{ij}) e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}} u_i\right\}.$$

Άρα, η περιθώρια (marginal) πιθανοφάνεια της i-οστής ομάδας μπορεί να βρεθεί ολοκληρώνοντας το u_i ως προς την εκ των προτέρων κατανομή $f(u_i|\omega)$.

$$\begin{aligned} L_i^{marginal}(\boldsymbol{\beta}, \boldsymbol{\psi}, \omega) &= \int_0^\infty L_i(\boldsymbol{\beta}, \boldsymbol{\psi}, u_i) f(u_i|\omega) du_i \\ &= \prod_{j=1}^{n_i} \left(\lambda_0(y_{ij}) e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}}\right)^{\delta_{ij}} \frac{\omega^\omega}{\Gamma(\omega)} \int_0^\infty u_i^{\omega + \delta_i - 1} e^{-(\omega + \sum_{j=1}^{n_i} \Lambda_0(y_{ij}) e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}}) u_i} du_i. \end{aligned}$$

Πολλαπλασιάζοντας και διαιρώντας με τους όρους που λείπουν έτσι ώστε να εμφανιστεί το ολοκλήρωμα της Gamma$(\omega + \delta_i, \omega + \sum_{j=1}^{n_i} \Lambda_0(y_{ij}) e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}})$ καταλήγουμε

$$L_i^{marginal}(\boldsymbol{\beta}, \boldsymbol{\psi}, \omega) = \prod_{j=1}^{n_i} \left(\lambda_0(y_{ij}) e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}}\right)^{\delta_{ij}} \frac{\omega^\omega}{\Gamma(\omega)} \frac{\Gamma(\omega + \delta_i)}{\left(\omega + \sum_{j=1}^{n_i} e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}} \Lambda_0(y_{ij})\right)^{\omega + \delta_i}},$$

ενώ σε λογαριθμική κλίμακα η περιθώρια πιθανοφάνεια της i-οστής ομάδας θα είναι

$$\begin{aligned} \ell_i^{marginal}(\boldsymbol{\beta}, \boldsymbol{\psi}, \omega) = &\sum_{j=1}^{n_i} \delta_{ij}(\log(\lambda_0(y_{ij})) + \mathbf{x}_{ij}^T \boldsymbol{\beta}) - \delta_i \log(\omega) - \log(\Gamma(\omega)) + \log(\Gamma(\omega + \delta_i)) \\ &- (\omega + \delta_i)\log(1 + \frac{1}{\omega} \sum_{j=1}^{n_i} e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}} \Lambda_0(y_{ij})), \end{aligned}$$

και η περιθώρια πιθανοφάνεια όλου του δείγματος, επειδή έχουμε υποθέσει ότι υπάρχει ανεξαρτησιά μεταξύ των ομάδων, θα είναι (Duchateau & Janssen 2007, σελ. 46)

$$\ell^{marginal}(\boldsymbol{\beta},\boldsymbol{\psi},\omega)=\sum_{i=1}^{G}\ell_i^{marginal}(\boldsymbol{\beta},\boldsymbol{\psi},\omega)=\sum_{i=1}^{G}\sum_{j=1}^{n_i}\delta_{ij}(\log(\lambda_0(y_{ij}))+\mathbf{x}_{ij}^T\boldsymbol{\beta})-\sum_{i=1}^{G}\delta_i\log(\omega)$$
$$-G\log(\Gamma(\omega))+\sum_{i=1}^{G}\log(\Gamma(\omega+\delta_i))-\sum_{i=1}^{G}(\omega+\delta_i)\log\left(1+\frac{1}{\omega}\sum_{j=1}^{n_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\Lambda_0(y_{ij})\right). \qquad (4.9)$$

4.2.1 Weibull Συνάρτηση Κινδύνου

Σε ένα μοντέλο Weibull η βασική συνάρτηση κινδύνου θα έχει τη μορφή $\lambda_0(t)=\kappa\lambda t^{\kappa-1}$, $(\kappa,\lambda>0,\ \boldsymbol{\psi}=(\lambda,\kappa))$. Επομένως, η συνάρτηση κινδύνου του j-οστού ατόμου από την i-οστή ομάδα θα είναι

$$\lambda(t|\mathbf{x}_{ij},u_i)=\kappa\lambda t^{\kappa-1}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}u_i.$$

Θα επεκτείνουμε την παράμετρο $\boldsymbol{\beta}$ έτσι ώστε να περιλαμβάνει το λ ($\boldsymbol{\beta}^T=(\beta_0,\ldots,\beta_p)$ και $\beta_0=\log(\lambda)$). Άρα, το διάνυσμα $\mathbf{x}_{ij}^T=(1,x_{ij1},\ldots,x_{ijp})$ θα πρέπει να περιλαμβάνει τη μονάδα. Επομένως, η συνάρτηση κινδύνου μπορεί να γραφτεί ως

$$\lambda(t|\mathbf{x}_{ij},u_i)=\lambda_0(t)e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}u_i,\ i=1,2,\ldots,G,\ j=1,2,\ldots,n_i,$$

όπου $\lambda_0(t)=\kappa t^{\kappa-1}$ και $\Lambda_0(t)=t^{\kappa}$. Η συνάρτηση επιβίωσης του j-οστού ατόμου από την i-οστή ομάδα θα είναι:

$$S(t|\mathbf{x},u_i)=\exp\{-\Lambda_0(t)e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}u_i\},\ i=1,2,\ldots,G,\ j=1,2,\ldots,n_i.$$

Δεσμευμένη κατανομή του $\kappa|\boldsymbol{\beta},\boldsymbol{u},\omega,\boldsymbol{y},\boldsymbol{\delta}$

Από τη σχέση (4.6) η απο κοινού εκ των υστέρων κατανομή των παραμέτρων του Weibull μοντέλου θα είναι

$$f(\boldsymbol{\beta},\boldsymbol{\kappa},\boldsymbol{u},\omega|\boldsymbol{y},\boldsymbol{\delta})\propto\kappa^{\delta}\exp\Big\{\sum_{i=1}^{G}\sum_{j=1}^{n_i}\delta_{ij}(\mathbf{x}_{ij}^T\boldsymbol{\beta}+\log(u_i))+(\kappa-1)\sum_{i=1}^{G}\sum_{j=1}^{n_i}\delta_{ij}\log(y_{ij})$$
$$-\sum_{i=1}^{G}\sum_{j=1}^{n_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}u_iy_{ij}^{\kappa}\Big\}\times\prod_{i=1}^{G}\left[\frac{\omega^{\omega}}{\Gamma(\omega)}u_i^{\omega-1}\exp\{-\omega u_i\}\right]\times\omega^{c-1}\exp\{-d\omega\}f(\boldsymbol{\beta})f(\kappa),$$

όπου $\delta=\sum_{i=1}^{G}\sum_{j=1}^{n_i}\delta_{ij}$ ο συνολικός αριθμός αποτυχιών. Αν θεωρήσουμε ως προτέρων κατανομή για το κ μια $\text{Gamma}(\alpha_0,\lambda_0)$, τότε η δεσμευμένη εκ των υστέρων κατανομή του κ,δοθέντων των υπολοίπων παραμέτρων, είναι:

$$f(\kappa|\boldsymbol{\beta},\boldsymbol{u},\omega,\boldsymbol{y},\boldsymbol{\delta}) \propto \kappa^{\alpha_0+\delta-1}\exp\Big\{\kappa\sum_{i=1}^{G}\sum_{j=1}^{n_i}\delta_{ij}\log(y_{ij}) - \sum_{i=1}^{G}\sum_{j=1}^{n_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}u_i y_{ij}^{\kappa} - \lambda_0\kappa\Big\}. \quad (4.10)$$

Η κατανομή (4.10) είναι αρκετά περίπλοκη για να μπορεί να προσομοιωθεί ευθέως με απλούς τρόπους (μέθοδος αντιστροφής ή μέθοδος απόρριψης). Όμως, θα δείξουμε ότι ο λογάριθμος της (4.10) είναι κοίλη συνάρτηση, άρα μπορούμε να χρησιμοποιήσουμε adaptive rejection sampling (Ibrahim et al. 2001, σελ. 103)

$$h(\kappa) = (\alpha_0+\delta-1)\log(\kappa) + \kappa\sum_{i=1}^{G}\sum_{j=1}^{n_i}\delta_{ij}\log(y_{ij}) - \sum_{i=1}^{G}\sum_{j=1}^{n_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}u_i y_{ij}^{\kappa} - \lambda_0\kappa,$$

$$h'(\kappa) = \frac{\alpha_0+\delta-1}{\kappa} + \sum_{i=1}^{G}\sum_{j=1}^{n_i}\delta_{ij}\log(y_{ij}) - \sum_{i=1}^{G}\sum_{j=1}^{n_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}u_i y_{ij}^{\kappa}\log(y_{ij}) - \lambda_0,$$

$$h''(\kappa) = -\frac{\alpha_0+\delta-1}{\kappa^2} - \sum_{i=1}^{G}\sum_{j=1}^{n_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}u_i y_{ij}^{\kappa}(\log(y_{ij}))^2 < 0, \quad \frac{\partial y_{ij}^{\kappa}}{\partial\kappa} = \frac{\partial e^{\kappa\log(y_{ij})}}{\partial\kappa} = y_{ij}^{\kappa}\log(y_{ij}).$$

Οι παράμετροι u_1, u_2, u_G και ω μπορούν να προσομοιωθούν ευθέως χρησιμοποιώντας τις σχέσεις (4.7) και (4.8), αντίστοιχα. Για την προσομοιώση του διανύσματος $\boldsymbol{\beta}$ θα εφαρμόσουμε την IWLS προτείνουσα του Gamerman όπως οριστηκε στην (4.4), με πιθανότητα αποδοχής της νέας τιμής σύμφωνα με την σχέση (4.5). Το μόνο που χρειάζεται είναι να αντικαταστήσουμε το $\Lambda_0(y_{ij})$ με $\Lambda_0(y_{ij}) = y_{ij}^{\kappa}$. Πρέπει να τονιστεί ότι ο πινακας σχεδιασμού έχει επεκταθεί ώστε να περιέχει μια στήλη με τη μονάδα.

4.2.2 Κατά Τμήματα Εκθετική Συνάρτηση Κινδύνου

Ένα από τα πιο δημοφιλή και ευέλικτα μοντέλα για ημι-παραμετρική ανάλυση επιβίωσης είναι το μοντέλο σταθερού, κατα τμήματα, κινδύνου. Για να κατασκευάσουμε το μοντέλο πρέπει πρώτα να ορίσουμε μια διαμέριση του χρονικού άξονα $0 = s_0 < s_1 < s_2 < \ldots < s_k < \ldots < s_J = \infty$. Επομένως, έχουμε J χρονικά διαστήματα:

$$\underbrace{[0,s_1)}_{\lambda_1}\ \underbrace{[s_1,s_2)}_{\lambda_2}\ \ldots\ \underbrace{[s_{k-1},s_k)}_{\lambda_k}\ \ldots\ \underbrace{[s_{J-1},s_J)}_{\lambda_J}$$

Στο k-οστό χρονικό διάστημα υποθέτουμε σταθερό βασικό κίνδυνο $\lambda_0(t) = \lambda_k$, $s_{k-1} \le t < s_k$. Άρα, η βασική συνάρτηση κινδύνου μπορεί να γραφτεί ως:

$$\lambda_0(t) = \sum_{g=1}^{J}\lambda_g I_g(t),\ I_g(t) = \mathbb{1}(s_{g-1} \le t < s_g).$$

Άν υποθέσουμε ότι ο χρόνος t ανήκει στο k-οστό χρονικό διάστημα ($s_{k-1} \leq t < s_k$), τότε η βασική συνάρτηση επιβίωσης θα είναι:

$$\begin{aligned} S_0(t) &= \exp\left\{-\int_0^t \lambda_0(u)du\right\} = \exp\left\{-\sum_{g=1}^{J}\lambda_g \int_0^t I_g(u)du\right\} \\ &= \exp\left\{-\sum_{g=1}^{k-1}\lambda_g \int_0^t I_g(u)du - \lambda_k \int_0^t I_k(u)du - \sum_{g=k+1}^{J}\lambda_g \int_0^t I_g(u)du\right\} \\ &= \exp\left\{-\lambda_k(t-s_{k-1}) - \sum_{g=1}^{k-1}\lambda_g(s_g - s_{g-1})\right\}, \end{aligned}$$

και η βασική συνάρτηση αθροιστικού κινδύνου θα είναι αντίστοιχα:

$$\Lambda_0(t) = -\log(S_0(t)) = \lambda_k(t-s_{k-1}) + \sum_{g=1}^{k-1}\lambda_g(s_g - s_{g-1}).$$

Επομένως, αν γνωρίζουμε σε ποιο χρονικό διάστημα ανήκει ο χρόνος t, είναι εύκολο να υπολογίσουμε τα $S_0(t), \Lambda_0(t)$. Θα εισάγουμε στη σημειολογία τα παρατηρούμενα δεδομένα $(y_{ij}, \delta_{ij}, \mathbf{x}_{ij}^T)$ και τη δεικτρια συνάρτηση $I_k(y_{ij}) = \mathbb{1}(s_{k-1} \leq y_{ij} < s_k),\ k = 1, 2, \ldots, J$. Άρα, $I_k(y_{ij}) = 1$ αν το j-οστό υποκείμενο της i-οστής ομάδας απέτυχε ή λογοκρίθηκε στο k-οστό χρονικό διάστημα $[s_{k-1}, s_k)$. Η βασική συνάρτηση επιβίωσης για ένα υποκείμενο με χρόνο παρατήρησης y_{ij} θα είναι:

$$S_0(y_{ij}) = \prod_{k=1}^{J}\exp\left\{-I_k(y_{ij})\Big[\lambda_k(y_{ij}-s_{k-1}) + \sum_{g=1}^{k-1}\lambda_g(s_g - s_{g-1})\Big]\right\},$$

και η βασική συνάρτηση κινδύνου: $\lambda_0(y_{ij}) = \prod_{k=1}^{J}\lambda_k^{I_k(y_{ij})}$.

Συνάρτηση Πιθανοφάνειας

Υποθέτοντας ένα μοντέλο αναλογικών κινδύνων (4.1) και μη πληροφοριακή λογοκρισία, θα κατασκευάσουμε την πιθανοφάνεια του κατά τμήματα εκθετικού μοντέλου. Ένα άτομο το οποίο λογοκρίθηκε στο χρόνο y_{ij} ($\delta_{ij} = 0$) συνεισφέρει στην πιθανοφάνεια με τον όρο:

$$S(y_{ij}|\mathbf{x}_{ij}, u_i) = \prod_{k=1}^{J}\exp\left\{-I_k(y_{ij})\Big[\lambda_k(y_{ij}-s_{k-1}) + \sum_{g=1}^{k-1}\lambda_g(s_g - s_{g-1})\Big]e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}u_i\right\},$$

αφού $S(y_{ij}|\mathbf{x}_{ij}, u_i) = S_0(y_{ij})^{e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}u_i}$. Ένα άτομο το οποίο απέτυχε στο χρόνο y_{ij} ($\delta_{ij} = 1$), θα συνεισφέρει με τον όρο $f(y_{ij}|\mathbf{x}_{ij}, u_i) = \lambda(y_{ij}|\mathbf{x}_{ij}, u_i)S(y_{ij}|\mathbf{x}_{ij}, u_i)$, όπου:

$$\lambda(y_{ij}|\mathbf{x}_{ij}, u_i) = \prod_{k=1}^{J} \left(\lambda_k e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}} u_i\right)^{I_k(y_{ij})}.$$

Άρα, η πλήρης πιθανοφάνεια των παραμέτρων $L(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{u})$ θα είναι (Ibrahim et al. 2001, σελ. 48):

$$\prod_{i=1}^{G} \prod_{j=1}^{n_i} \prod_{k=1}^{J} \left(\lambda_k e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}} u_i\right)^{I_k(y_{ij})\delta_{ij}} \exp\left\{ - I_k(y_{ij}) \Big[\lambda_k (y_{ij} - s_{k-1}) + \sum_{g=1}^{k-1} \lambda_g (s_g - s_{g-1})\Big] e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}} u_i \right\}.$$

Δεσμευμένη κατανομή του $\boldsymbol{\lambda}|\boldsymbol{\beta}, \boldsymbol{u}, \omega, \boldsymbol{y}, \boldsymbol{\delta}$

Για να εξάγουμε τη δεσμευμένη εκ των υστέρων κατανομή του $\boldsymbol{\lambda}$ είναι βολικό να θεωρήσουμε την πιθανοφάνεια ως ένα γινόμενο από τις συνεισφορές διαφορετικών χρονικών διαστημάτων. Έστω $\mathcal{D}_k$ το σύνολο των υποκειμένων τα οποία αποτυγχάνουν στο k-οστό χρονικό διάστημα $[s_{k-1}, s_k)$ και d_k ο αριθμός των υποκειμένων τα οποία περιέχονται στο $\mathcal{D}_k$. Έστω $\mathcal{R}_k$ το σύνολο των ατόμων υπο κίνδυνο τη στιγμή s_{k-1}, δηλαδή $\mathcal{R}_k = \{(i,j) : y_{ij} \geq s_{k-1}\}$. Επίσης, $\Delta_{ijk} = \min\{y_{ij}, s_k\} - s_{k-1}$ είναι ο χρόνος παρακολούθησης του j-οστού υποκειμένου από την i-οστή ομάδα στο k-οστό χρονικό διάστημα, για τα υποκείμενα τα οποία ανήκουν στο $\mathcal{R}_k$. Επομένως, η πιθανοφάνεια του κατά τμήματα εκθετικού μοντέλου, μπορεί να γραφτεί εναλλακτικά ως εξής (Ibrahim et al. 2001, σελ. 65):

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{u}) = \prod_{k=1}^{J} \lambda_k^{d_k} \exp\left\{ \sum_{(i,j)\in\mathcal{D}_k} \left[\mathbf{x}_{ij}^T \boldsymbol{\beta} + \log(u_i)\right] - \lambda_k \sum_{(i,j)\in\mathcal{R}_k} \Delta_{ijk} e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}} u_i \right\}.$$

Αν θεωρήσουμε ότι τα $\lambda_1, \lambda_2, \ldots, \lambda_J$ είναι εκ των προτέρων ανεξάρτητα (και ανεξάρτητα από τις υπόλοιπες παραμέτρους), με $\lambda_k \sim \text{Gamma}(\alpha_{0k}, \lambda_{0k})$, $k = 1, 2, \ldots, J$, τότε η δεσμευμένη εκ των υστέρων κατανομή του λ_k θα είναι

$$f(\lambda_k|\boldsymbol{\beta}, \boldsymbol{\lambda}_{-k}, \boldsymbol{u}, \omega, \boldsymbol{y}, \boldsymbol{\delta}) \propto \lambda_k^{d_k + \alpha_{0k} - 1} \exp\left\{ -\big(\lambda_{0k} + \sum_{(i,j)\in\mathcal{R}_k} \Delta_{ijk} e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}} u_i\big)\lambda_k \right\},$$

άρα

$$\lambda_k|\boldsymbol{\beta}, \boldsymbol{\lambda}_{-k}, \boldsymbol{u}, \omega, \boldsymbol{y}, \boldsymbol{\delta} \sim \text{Gamma}(\alpha_{0k} + d_k, \lambda_{0k} + \sum_{(i,j)\in\mathcal{R}_k} \Delta_{ijk} e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}} u_i),\ k = 1, 2, \ldots, J. \quad (4.11)$$

Η προσομοίωση των $u_1, u_2, \ldots, u_G$ και ω μπορεί να γίνει ευθέως χρησιμοποιώντας τις σχέσεις (4.7) και (4.8), αντίστοιχα. Για την προσομοιώση του διανύσματος, $\boldsymbol{\beta}$, θα εφαρμόσουμε την IWLS προτείνουσα του Gamerman (4.4). Η βασική αθροιστική συνάρτηση κινδύνου θα είναι ίση με $\Lambda_0(y_{ij}) = \sum_{k=1}^{J} I_k(y_{ij})\big(\lambda_k(y_{ij} - s_{k-1}) + \sum_{g=1}^{k-1} \lambda_g(s_g - s_{g-1})\big)$.

4.3 Lognormal Frailty Μοντέλο

Στην παρούσα ενότητα θα υποθέσουμε ότι ο λογάριθμος της frailty μεταβλητής ακολουθεί την κανονική κατανομή με μέση τιμή 0 και ακρίβεια ω ($b_i = \log(u_i) \sim N(0, \omega^{-1})$), δηλαδή

$$f(b_i|\omega) \propto \omega^{1/2} \exp\big\{ -\frac{\omega}{2} b_i^2 \big\},\ i = 1, 2, \ldots, G.$$

Επίσης, τα $b_1, b_2, \ldots, b_G$ είναι ανεξάρτητα μεταξύ τους. Για την ακρίβεια ω υποθέτουμε μια Gamma εκ των προτέρων κατανομή, $\omega \sim \text{Gamma}(c, d)$. Ανάλογα με τη σχέση (4.6), η απο κοινού εκ των υστέρων κατανομή όλων των παραμέτων θα είναι:

$$\begin{aligned} f(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{b}, \omega|\boldsymbol{y}, \boldsymbol{\delta}) \propto & \prod_{i=1}^{G} \prod_{j=1}^{n_i} \Big(\lambda_0(y_{ij}) e^{\mathbf{x}_{ij}^T \boldsymbol{\beta} + b_i} \Big)^{\delta_{ij}} \exp\Big\{ -\Lambda_0(y_{ij}) e^{\mathbf{x}_{ij}^T \boldsymbol{\beta} + b_i} \Big\} f(\boldsymbol{\beta}) f(\boldsymbol{\psi}) \\ & \times \prod_{i=1}^{G} \Big[\omega^{1/2} \exp\big\{ -\frac{\omega}{2} b_i^2 \big\} \Big] \times \omega^{c-1} \exp\{-d\omega\}. \end{aligned} \tag{4.12}$$

Δεσμευμένη κατανομή του $\omega|\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{b}, \boldsymbol{y}, \boldsymbol{\delta}$

Η δεσμευμένη εκ των υστέρων κατανομή του ω θα είναι:

$$f(\omega|\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{b}, \boldsymbol{y}, \boldsymbol{\delta}) \propto \omega^{G/2 + c - 1} \exp\left\{ -(d + \frac{1}{2} \sum_{i=1}^{G} b_i^2)\omega \right\}, \tag{4.13}$$

η οποία αναγνωρίζεται ως $\text{Gamma}(G/2 + c, d + \frac{1}{2}\sum_{i=1}^{G} b_i^2)$. Άρα, η προσομοίωση του ω είναι απλή.

Δεσμευμένη κατανομή του $b_i|\boldsymbol{\beta}, \boldsymbol{\psi}, \omega, \boldsymbol{y}, \boldsymbol{\delta}$

Η δεσμευμένη, εκ των υστέρων, κατανομή του τυχαίου όρου b_i (random effect), το οποίο συσχετίζεται με την i-οστή ομάδα θα είναι:

$$f(b_i|\boldsymbol{\beta}, \boldsymbol{\psi}, \omega, \boldsymbol{y}, \boldsymbol{\delta}) \propto \exp\left\{ b_i \delta_i - e^{b_i} \sum_{j=1}^{n_i} \Lambda_0(y_{ij}) e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}} - \frac{\omega}{2} b_i^2 \right\},\ i = 1, 2, \ldots, G. \tag{4.14}$$

Η (4.14) είναι η εκ των υστέρων κατανομή ενός nested random effect η οποία προκύπτει από ένα 'μοντέλο' Poisson, $\delta_{ij}|\boldsymbol{\beta}, \boldsymbol{\psi}, b_i \sim \text{Poisson}(\Lambda_0(y_{ij}) e^{\mathbf{x}_{ij}^T \boldsymbol{\beta} + b_i})$, με μέση τιμή $\mu_{ij} = \text{E}(\delta_{ij}|\boldsymbol{\beta}, \boldsymbol{\psi}, b_i) = \Lambda_0(y_{ij}) e^{\mathbf{x}_{ij}^T \boldsymbol{\beta} + b_i}$. Η συνδετική συνάρτηση του γενικευμένου γραμμικού μοντέλου είναι $g(\mu_{ij}) = \log(\mu_{ij}) = \log(\Lambda_0(y_{ij})) + \mathbf{x}_{ij}^T \boldsymbol{\beta} + b_i = \eta_{ij}$. Άρα, χρησιμοποιούμε την κανονική συνδετική συνάρτηση (canonical link function). Πρέπει να σημειωθεί ότι το offset του μοντέλου είναι το $\log(\Lambda_0(y_{ij})) + \mathbf{x}_{ij}^T \boldsymbol{\beta}$ (αφού εξετάζουμε την κατανομή του b_i). Ορίζουμε ως

$\mu_i = \mathrm{E}(\delta_i|\boldsymbol{\beta}, \boldsymbol{\psi}, b_i) = \sum_{j=1}^{n_i} \Lambda_0(y_{ij}) e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}+b_i},\ i = 1, 2, \ldots, G$. Ακολουθώντας τον αλγόριθμο του Gamerman (1997) για την προσομοίωση των random effects σε μεικτά γενικευμένα γραμμικά μοντέλα (GLMM), ορίζουμε μετασχηματισμένες παρατηρήσεις και βάρη:

$$\tilde{\delta}_i(b_i) = b_i + (\delta_i - \mu_i)g'(\mu_i),\ W_i(b_i) = \frac{1}{g'(\mu_i)},\ i = 1, 2, \ldots, G.$$

Θα προτείνουμε μια νέα τιμή από την κανονική κατανομή με διακύμανση $C_1(b_i) = \frac{1}{\omega+W_i(b_i)}$ και μέση τιμή $\mu_1(b_i) = C_1(b_i)W_i(b_i)\tilde{\delta}_i(b_i)$, δηλαδή

$$b_i^{can} \sim N(\mu_1(b_i), C_1(b_i)),\ i = 1, 2, \ldots, G,$$

όπου b_i είναι η τρέχουσα τιμή της αλυσίδας. Θα δείξουμε ότι η προτείνουσα του Gamerman μπορεί να εξαχθεί ως ένα βήμα του αλγορίθμου Newton-Raphson για τη μεγιστοποίηση της (4.14). Η πρώτη και η δεύτερη παράγωγος του λογαρίθμου της (4.14) (αγνοώντας τη σταθερά κανονικοποίησης) δίνονται από:

$$\begin{aligned} h'(b_i) &= \delta_i - e^{b_i}\sum_{j=1}^{n_i}\Lambda_0(y_{ij})e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}} - \omega b_i = \delta_i - \mu_i - \omega b_i, \\ h''(b_i) &= -e^{b_i}\sum_{j=1}^{n_i}\Lambda_0(y_{ij})e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}} - \omega = -(\mu_i + \omega). \end{aligned}$$

Εφόσον $W_i(b_i) = 1/g'(\mu_i) = \mu_i$, θα ισχύει ότι $C_1(b_i) = -1/h''(b_i)$ Ας υποθέσουμε ότι μας ενδιαφέρει η μεγιστοποίηση της (4.14). Αν ο αλγόριθμος Newton-Raphson λάβει ως αρχική τιμή το b_i τότε το επόμενο βήμα θα είναι:

$$\begin{aligned} b_i - h''(b_i)^{-1}h'(b_i) &= b_i + C_1(b_i)(\delta_i - \mu_i - \omega b_i) = C_1(b_i)(\delta_i - \mu_i + \mu_i b_i) \\ &= C_1(b_i)\mu_i(b_i + (\delta_i - \mu_i)/\mu_i) = C_1(b_i)W_i(b_i)\tilde{\delta}_i(b_i). \end{aligned}$$

Στην ουσία, αυτό που συμβαίνει είναι ότι προτείνουμε μια τιμη από την κανονική κατανομή με μέση τιμή ένα βήμα (από τη τρέχουσα τιμή) του αλγορίθμου Newton-Raphson και διακύμανση σύμφωνα με τη δεύτερη παράγωγο του λογαρίθμου της κατανομής, υπολογισμένη στη τρέχουσα τιμή. Δεν είναι υποχρεωτικό ο IWLS αλγόριθμος του Gamerman να ξεκινάει κάθε φορά από τη τρέχουσα τιμή της αλυσίδας. Μπορούμε να εφαρμόσουμε μια ελαφρώς διαφορετική προσέγγιση. Δανειζόμενοι πληροφορία απο το Gamma Frailty μοντέλο της προηγούμενης ενότητας, μπορούμε να ξεκινάμε τον αλγόριθμο του Gamerman από μια τιμή $b_i^\star$, όπου

$$e^{b_i^\star} = (\omega + \delta_i)/(\omega + \sum_{j=1}^{n_i}\Lambda_0(y_{ij})e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}),$$

είναι η δεσμευμένη μέση τιμή του u_i (σύμφωνα με την κατανομή (4.7)), στην περίπτωση όπου τα frailties ακολουθούν εκ των προτέρων την Gamma κατανομή. Επομένως, προτείνουμε μια τιμή $b_i^{can} \sim N(\mu_1(b_i^\star), C_1(b_i^\star))$, με πιθανότητα αποδοχής:

$$p = \min\left\{1, p_1 = \frac{f(b_i^{can}|\boldsymbol{\beta},\boldsymbol{\psi},\omega,\boldsymbol{y},\boldsymbol{\delta})}{f(b_i|\boldsymbol{\beta},\boldsymbol{\psi},\omega,\boldsymbol{y},\boldsymbol{\delta})}\frac{q(b_i|b_i^{can},\boldsymbol{\beta},\boldsymbol{\psi},\omega,\boldsymbol{y},\boldsymbol{\delta})}{q(b_i^{can}|b_i,\boldsymbol{\beta},\boldsymbol{\psi},\omega,\boldsymbol{y},\boldsymbol{\delta})}\right\}, \tag{4.15}$$

ενώ σε λογαριθμική κλίμακα:

$$\begin{aligned}\log(p_1) = (b_i^{can} - b_i)\delta_i - (e^{b_i^{can}} - e^{b_i})\sum_{j=1}^{n_i}\Lambda_0(y_{ij})e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}} - \frac{\omega}{2}\left((b_i^{can})^2 - (b_i)^2\right)\\ - \frac{\left(b_i - \mu_1(b_i^\star)\right)^2}{2C_1(b_i^\star)} + \frac{\left(b_i^{can} - \mu_1(b_i^\star)\right)^2}{2C_1(b_i^\star)}.\end{aligned}$$

Πρέπει να τονιστεί ότι δεν χρειάζεται να κάνουμε το αντίστροφο IWLS βήμα (ξεκινώντας από b_i^{can}), αφού η προτεινουσα b_i^{can} δεν εξαρτάται από τη τρέχουσα τιμή της αλυσίδας b_i. Φυσικά, αν θέλουμε να αυξήσουμε το ποσοστό αποδοχής του αλγορίθμου μπορούμε να εφαρμόσουμε δυο, ή και περισσότερες, επαναλήψεις του IWLS αλγόριθμου. Σε τέτοια περίπτωση η τιμή $\mu_1(b_i^\star)$ θα είναι πολύ κοντά στην κορυφή της δεσμευμένης κατανομής του b_i.

Όταν υποθέσουμε Gamma κατανομή για τα frailties είδαμε ότι μπορούμε να εκφράσουμε την περιθώρια πιθανοφάνεια σε κλειστή μορφή. Δυστυχώς, στην περίπτωση της κανονικής κατανομής, η περιθώρια πιθανοφάνεια (ως προς τα $b_1, b_2, \ldots, b_G$) δεν μπορεί να βρεθεί αναλυτικά. Η συνεισφορά της i-οστής ομάδας στην πιθανοφάνεια είναι:

$$L_i(\boldsymbol{\beta},\boldsymbol{\psi},b_i) = \prod_{j=1}^{n_i}\left(\lambda_0(y_{ij})e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}+b_i}\right)^{\delta_{ij}}\exp\left\{-\Lambda_0(y_{ij})e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}+b_i}\right\}.$$

Η περιθώρια (marginal) πιθανοφάνεια της i-οστής ομάδας μπορεί να βρεθεί ολοκληρώνοντας το b_i ως προς την εκ των προτέρων κατανομή $f(b_i|\omega)$:

$$\begin{aligned}L_i^{marginal}(\boldsymbol{\beta},\boldsymbol{\psi},\omega) = \int_{-\infty}^{+\infty} L_i(\boldsymbol{\beta},\boldsymbol{\psi},b_i)f(b_i|\omega)db_i = \prod_{j=1}^{n_i}\left[\left(\lambda_0(y_{ij})e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\right)^{\delta_{ij}}\right](2\pi)^{-1/2}\omega^{1/2}\\ \times\int_{-\infty}^{+\infty}\exp\left\{b_i\delta_i - e^{b_i}\sum_{j=1}^{n_i}\Lambda_0(y_{ij})e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}} - \frac{\omega}{2}b_i^2\right\}db_i.\end{aligned} \tag{4.16}$$

Το ολοκλήρωμα της (4.16) δεν μπορεί να βρεθεί σε κλειστή μορφή. Ο υπολογισμός του ολοκληρώματος μπορεί να γίνει με μεθόδους αριθμητικής ή στοχαστικής ολοκλήρωσης. Θα χρησιμοποιήσουμε τη συνάρτηση `aghQuad` της `R` (βιβλιοθήκη `fastGHQuad`), η οποία προσεγγίζει το ολοκλήρωμα με Adaptive Gauss-Hermite quadrature μέθοδο. Η μέθοδος χρειάζεται ένα μέτρο θέσης και διασποράς της υπο-ολοκλήρωσης συνάρτησης, έστω g, για να προσεγγίσει

τη g μέσω της κανονικής κατανομής. Ως μέτρο θέσης θα θεωρήσουμε το $\mu_1(b_i^\star)$ και ως μέτρο διασποράς το $\sqrt{C_1(b_i^\star)}$. Για να πετύχουμε την επιθυμητή ακρίβεια θα χρησιμοποιήσουμε 25 σημεία ολοκλήρωσης. Σε περιορισμένο αριθμό πειραμάτων βρέθηκε ότι η δεσμευμένη εκ των υστέρων κατανομή των $b_i,\ i = 1, 2, \ldots, G$, προσεγγίζεται από μια κανονική κατανομή, άρα η μέθοδος αποδίδει ικανοποιητικά. Η περιθώρια πιθανοφάνεια όλου του δείγματος θα είναι σε λογαριθμική κλίμακα:

$$\ell^{marginal}(\boldsymbol{\beta}, \boldsymbol{\psi}, \omega) = \sum_{i=1}^{G} \ell_i^{marginal}(\boldsymbol{\beta}, \boldsymbol{\psi}, \omega) = \sum_{i=1}^{G}\sum_{j=1}^{n_i} \delta_{ij}(\log(\lambda_0(y_{ij})) + \mathbf{x}_{ij}^T\boldsymbol{\beta})$$
$$- \frac{1}{2}G\log(2\pi) + \frac{1}{2}G\log(\omega) + \sum_{i=1}^{G}\log\left(\int_{-\infty}^{\infty} g(b_i)db_i\right), \qquad (4.17)$$

όπου $g(b_i)$ είναι η δεσμευμένη εκ των υστέρων κατανομή του b_i (αγνοώντας τη σταθερά κανονικοποίησης). Για να οριστεί πλήρως το μοντέλο πρέπει να θεωρήσουμε κάποια παραμετρική μορφή για τη βασική συνάρτηση κινδύνου $\lambda_0(t)$. Όπως και στην περίπτωση των Gamma frailties, θα εξετασουμε το Weibull και το κατά τμήματα εκθετικό μοντέλο.

4.3.1 Weibull Συνάρτηση Κινδύνου

Η προσομοίωση της ακρίβειας ω μπορεί να γίνει ευθέως από τη σχέση (4.13). Όπως συζητήθηκε στην υπο-ενότητα 4.2.1 σε ένα Weibull μοντέλο θεωρούμε ως $\Lambda_0(y_{ij}) = y_{ij}^\kappa$ και $\lambda_0(y_{ij}) = \kappa y_{ij}^{\kappa-1}$. Επίσης, ο πίνακας σχεδιασμού θα περιέχει τη μονάδα. Η προσομοίωση του $\boldsymbol{\beta}$ γίνεται με βήμα Metropolis-Hastings, σύμφωνα με τη (4.4), θέτοντας $u_i = e^{b_i}$. Η προσομοίωση των τυχαίων όρων $b_1, b_2, \ldots, b_G$ γίνεται με Metropolis-Hastings βήμα, με προτείνουσα $b_i^{can} \sim N(\mu_1(b_i^\star), C_1(b_i^\star))$ και πιθανότητα αποδοχής (4.15). Η προσομοίωση του κ γίνεται ευθέως από την κατανομή (4.10), χρησιμοποιώντας adaptive rejection sampling, και θέτοντας $u_i = e^{b_i}$. Η περιθώρια πιθανοφάνεια μπορεί να βρεθεί προσεγγιστικά από τη σχέση (4.17).

4.3.2 Κατά Τμήματα Εκθετική Συνάρτηση Κινδύνου

Η προσομοίωση της ακρίβειας ω θα γίνει με τον ίδιο τρόπο από τη σχέση (4.13). Όπως είδαμε στην υπο-ενότητα 4.2.2, στο piecewise εκθετικό μοντέλο ο πίνακας σχεδιασμού δεν περιέχει τη μονάδα και $\Lambda_0(y_{ij}) = \sum_{k=1}^{J} I_k(y_{ij})\big(\lambda_k(y_{ij} - s_{k-1}) + \sum_{g=1}^{k-1}\lambda_g(s_g - s_{g-1})\big)$, $\lambda_0(y_{ij}) = \sum_{k=1}^{J} I_k(y_{ij})\lambda_k$. Η προσομοίωση του $\boldsymbol{\beta}$ γίνεται με μηχανισμό Metropolis-Hastings, σύμφωνα με τη (4.4), και $u_i = e^{b_i}$. Η προσομοίωση των τυχαίων όρων $b_1, b_2, \ldots, b_G$ γίνεται με Metropolis-Hastings βήμα, με προτείνουσα $b_i^{can} \sim N(\mu_1(b_i^\star), C_1(b_i^\star))$ και πιθανότητα αποδοχής (4.15). Η προσομοιώση των baseline παραμέτρων $\lambda_1, \lambda_2, \ldots, \lambda_J$ μπορεί να γίνει ευθέως από την (4.11), όπου $u_i = e^{b_i}$. Η περιθώρια πιθανοφάνεια μπορεί να βρεθεί προσεγγιστικά από τη σχέση (4.17).

4.4 Σύγκριση Μοντέλων Frailty

Ο πιο φυσικός τρόπος επιλογής μοντέλου στη Μπεϋζιανή στατιστική γίνεται μέσω της δήλωσης των εκ των υστέρων πιθανοτήτων των υποψήφιων μοντέλων. Όμως, σε πολυπαραμετρικά προβλήματα, ο υπολογισμός της περιθώριας πιθανοφάνειας των μοντέλων παρουσιάζει αρκετές δυσκολίες. Στην παρούσα ενότητα θα εξετάσουμε το κριτήριο DIC (Deviance Information Criterion) ως ένα μέτρο επιλογής μοντέλου.

Κατ' αρχήν μπορούμε να συνοψίσουμε την εφαρμογή του μοντέλου μέσω της Deviance, η οποία ορίζεται ως: $D(\boldsymbol{\theta}) = -2\log(f(Data|\boldsymbol{\theta}))$, όπου $\boldsymbol{\theta}$ είναι το διάνυσμα όλων των παραμέτρων. Εφόσον οι παράμετροι $u_1, u_2, \ldots, u_G$ δεν είναι άμεσου ενδιαφέροντος ορίζουμε τη Deviance ως

$$D(\boldsymbol{\beta}, \boldsymbol{\psi}, \omega) = -2\ell^{marginal}(\boldsymbol{\beta}, \boldsymbol{\psi}, \omega).$$

Η $\ell^{marginal}(\boldsymbol{\beta}, \boldsymbol{\psi}, \omega)$ δίνεται από την (4.17) όταν $b_i = \log(u_i)|\omega \sim N(0, \omega^{-1})$ και (4.9) όταν $u_i|\omega \sim \text{Gamma}(\omega, \omega)$. Η εκ των υστέρων μέση τιμή της Deviance μπορεί να θεωρηθεί ως ένα μέτρο της εφαρμογής του μοντέλου. Όμως, η Deviance από μόνη της δεν μπορεί να εφαρμοστεί για την επιλογή μοντέλου γιατι δεν λαμβάνει υπόψη την πολυπλοκότητα του μοντέλου. Ένα μέτρο της πολυπλοκότητας του μοντέλου είναι ο 'effective' αριθμός των παραμέτρων p_D. Μπορούμε να εκτιμήσουμε το p_D ως τη μισή εκ των υστέρων διακύμανση της Deviance: $p_D \simeq \frac{1}{2}\widehat{\text{Var}}(D(\boldsymbol{\beta}, \boldsymbol{\psi}, \omega)|\boldsymbol{y}, \boldsymbol{\delta})$ (Gelman, Carlin, Stern & Rubin 2003, σελ. 182). Εναλλακτικά, το p_D μπορεί να εκτιμηθεί ως η διαφορά της εκ των υστέρων μέσης τιμής της Deviance από την Deviance υπολογισμένη σε κάποια σημειακή εκτίμηση των παραμέτρων.

$$p_D \simeq \text{E}(D(\boldsymbol{\beta}, \boldsymbol{\psi}, \omega)|\boldsymbol{y}, \boldsymbol{\delta}) - D(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\psi}}, \widehat{\omega}).$$

Το κριτήριο DIC λαμβάνει υπόψη την εφαρμογή του μοντέλου αλλά και την πολυπλοκότητα του και ορίζεται ως το άθροισμα της εκ των υστέρων μέσης τιμής της Deviance και του 'effective' αριθμού των παραμέτρων (Spiegelhalter, Best, Carlin & Van Der Linde 2002):

$$\text{DIC} = \text{E}(D(\boldsymbol{\beta}, \boldsymbol{\psi}, \omega)|\boldsymbol{y}, \boldsymbol{\delta}) + p_D.$$

Αφού πάρουμε ένα δείγμα από την από κοινού εκ των υστέρων κατανομή των παραμέτρων, ο υπολογισμός του κριτηρίου DIC γίνεται εύκολος. Μοντέλα με μικρότερη τιμή DIC είναι γενικά προτιμητέα. Αν είχαμε χρησιμοποιήσει τη δεσμευμένη πιθανοφάνεια (4.2) στον καθορισμό της Deviance, θα καταλήγαμε σε ένα 'δεσμευμένο' DIC το οποίο εστιάζει περισσότερο στα frailties. Στο περιθώριο DIC, τα frailties διαμορφώνουν το σχήμα της πιθανοφάνειας, αλλά το ενδιαφέρον εστιάζεται κυρίως στις σταθερές επιδράσεις.

Κεφάλαιο 5

Εφαρμογές σε Προσομοιωμένα Δεδομένα

Είναι γνωστό ότι αν υπάρχει μια στατιστική σχέση μεταξύ δυο παραγόντων, δεν σημαίνει ότι οι δυο παράγοντες σχετίζονται αιτιολογικά. Για να μπορέσουμε να δώσουμε αιτιολογική ερμηνεία σε μια σχέση, πρέπει να είμαστε σίγουροι ότι η συσχέτιση δεν προκαλείται από συγχυτικούς παράγοντες. Είναι σχεδόν σίγουρο ότι, σε μη πειραματικά δεδομένα, θα παρατηρήσουμε συσχετίσεις οι οποίες είναι αποτέλεσμα συγχυτικών παραγόντων. Για παράδειγμα, ας υποθέσουμε ότι παρατηρούμε μια θετική σχέση μεταξύ κατανάλωσης καφέ και επίπτωσης καρκίνου του πνεύμονα. Άρα, ο πληθυσμός των ατόμων που καταναλώνουν καφέ θα έχει μεγαλύτερο κίνδυνο ανάπτυξης καρκίνου του πνεύμονα σε σχέση με τον πληθυσμό που δεν καταναλώνει καφέ. Είναι όμως γνωστό ότι το κάπνισμα είναι ένας αιτιολογικός παράγοντας του καρκίνου του πνεύμονα. Αν παρατηρήσουμε προσεκτικά τα δεδομένα μας, θα δούμε ότι τα άτομα που πίνουν καφέ, έχουν μεγαλύτερη πιθανότητα να καπνίζουν σε σχέση με τα άτομα που δεν πίνουν καφέ. Ένας τρόπος να εξουδετερώσουμε τη συγχυτική επίδραση του καπνίσματος, είναι να μετρήσουμε τη σχέση της κατανάλωσης καφέ και καρκίνου του πνεύμονα, δεδομένης της μεταβλητής του καπνίσματος. Σε αυτή την περίπτωση, είναι πολύ πιθανό η συσχέτιση της κατανάλωσης καφέ και καρκίνου του πνεύμονα να εξαφανιστεί.

Σε πολλές περιπτώσεις, για να εξουδετερώσουμε τους συγχυτικούς παράγοντες, είναι πιο αποτελεσματικό να χρησιμοποιούμε εξομοιωμένα δείγματα. Για δυαδική έκβαση (αποτυχία/επιτυχία), μπορούμε να εξομοιώσουμε τα άτομα ως προς την έκβαση. Στις $1 : k$ εξομοιωμένες μελέτες ασθενών μαρτύρων, για κάθε ασθενή επιλέγουμε k άτομα χωρίς τη νόσο, τα οποία έχουν τις ίδιες (ή παραπλήσιες) τιμές των συγχυτικών παραγόντων. Στο παραδειγμά μας, για κάθε άτομο με καρκίνο του πνεύμονα θα διαλέγαμε k άτομα χωρίς καρκίνο του πνεύμονα, αλλά με τις ίδιες τιμές συγχυτικών παραγόντων με τον ασθενή. Επομένως, η δειγματική (περιθώρια) συσχέτιση μεταξύ των συγχυτικών παραγόντων και της έκβασης, έχει εξαλειφθεί. Άρα, οι παρά-

γοντες οι οποίοι περιλαμβάνονται στην εξομοίωση δεν μπορούν να έχουν συγχυτικές επιδράσεις στη σχέση της έκθεσης με την έκβαση, με την προϋπόθεση ότι το μοντέλο έχει λάβει υπόψη την ομαδοποίηση των δεδομένων (Sjölander & Greenland 2013).

Σε πιο σπάνιες περιπτώσεις συναντάμε εξομοιωμένα δείγματα σε προοπτικές μελέτες. Σε μια $1:k$ εξομοιωμένη προοπτική μελέτη, για κάθε εκτεθειμένο άτομο επιλέγουμε k μη εκτεθειμένα άτομα, τα οποία έχουν τις ίδιες (ή παραπλήσιες) τιμές συγχυτικών παραγόντων. Επομένως, η συσχέτιση μεταξύ των συγχυτικών παραγόντων και της έκθεσης, εξαφανίζεται (στο δείγμα). Άρα, οι παράγοντες οι οποίοι συμμετέχουν στην εξομοίωση δεν είναι δυνατόν να διαστρεβλώνουν τη σχέση της έκθεσης με την έκβαση. Πολλές φορές τα δεδομένα είναι εξομοιωμένα από τη φύση τους. Για παράδειγμα, μια μελέτη που επικεντρώνεται σε αδέλφια τα οποία είναι ασύμφωνα ως προς την έκθεση (ένα άτομο εκτεθειμένο και το άλλο όχι), είναι μια προοπτική μελέτη, εξομοιωμένη ως προς πολλούς παράγοντες οι οποίοι είναι δύσκολο να μετρηθούν (ένα μέρος του DNA και κοινό παιδικό περιβάλλον).

Θα επικεντρώσουμε την προσοχή μας σε δεδομένα επιβίωσης. Ας υποθέσουμε ότι μια προοπτική μελέτη αποτελείται από G ζευγάρια διδύμων, τα οποία είναι ασύμφωνα ως προς τη μεταβλητή ενδιαφέροντος. Η μεταβλητή ενδιαφέροντος είναι κάποιου είδους παρέμβαση (θεραπεία) η οποία μειώνει τον κίνδυνο θανάτου του ασθενούς. Τα ζευγάρια των διδύμων μοιράζονται κοινούς προγνωστικούς παράγοντες (DNA), άρα, οι χρόνοι επιβίωσής τους θα είναι συσχετισμένοι (περιθώρια συσχέτιση ως προς τα ζευγάρια-ομάδες).

Έστω T_{ij} ο χρόνος επιβίωσης και C_{ij} ο χρόνος λογοκρισίας του j-οστού ατόμου από το i-οστό ζευγάρι ($i = 1, 2, \ldots, G$, $j = 1, 2$). Θα θεωρήσουμε μια επεξηγηματική μεταβλητή, τη θεραπεία x_{ij}. Είμαστε σε θέση να παρατηρήσουμε τους χρόνους $Y_{ij} = \min\{T_{ij}, C_{ij}\}$, με δείκτριες συναρτήσεις αποτυχίας $\Delta_{ij} = \mathbb{1}(T_{ij} \leq C_{ij})$. Ένας τρόπος να μοντελοποιήσουμε τέτοιου είδους δεδομένα είναι μέσω των shared frailty μοντέλων, τα οποία αναλύθηκαν στο κεφάλαιο 4. Τα shared frailty μοντέλα εξετάζουν τη συσχέτιση της θεραπείας με τον κίνδυνο θανάτου, δεδομένης της τιμής frailty u_i. Η δέσμευση ως προς το u_i είναι ισοδύναμη με τη δέσμευση ως προς την ομάδα i, δηλαδή το u_i περιλαμβάνει όλους τους παρατηρήσιμους και μη παρατηρήσιμους παράγοντες που είναι κοινοί στην ομάδα i (το σύνολο των μεταβλητών της εξομοίωσης). Επομένως, τα shared frailty μοντέλα στοχεύουν στην παράμετρο β, η οποία είναι ο λογάριθμος του πηλίκου κινδύνων, δεδομένης της ομάδας i:

$$\frac{\lambda(t|x_{ij}, u_i)}{\lambda(t|x_{ij'}, u_i)} = \exp\{\beta(x_{ij} - x_{ij'})\}. \tag{5.1}$$

Όπως έχουμε δει, τα $u_1, u_2, \ldots, u_G$ θεωρούνται ως ένα τυχαίο δείγμα από κάποια κατανομή. Αυτό σημαίνει ότι το u_i πρέπει να είναι ανεξάρτητο από το διάνυσμα $\mathbf{X}_i^T = (x_{i1}, x_{i2})$ της i-οστής

ομάδας.

$$u_i \perp \mathbf{X}_i,\ i = 1, 2, \ldots, G. \tag{5.2}$$

Η συνθήκη (5.2) είναι λογικό ότι θα παραβιάζεται πάντα όταν το u_i περιέχει συγχυτικούς παράγοντες (Sjölander et al. 2013). Η εκτίμηση του β από ένα shared frailty μοντέλο, μπορεί να είναι μεροληπτική αν η συνθήκη (5.2) παραβιαστεί (Sjölander et al. 2013). Όμως, σε μια εξομοιωμένη προοπτική μελέτη, η (5.2) ισχύει πάντα, αφού το διάνυσμα $\mathbf{X}_i$ είναι το ίδιο για όλες τις ομάδες. Εναλλακτικά, μπορούμε να εφαρμόσουμε στρωματοποιημένο Cox μοντέλο της μορφής

$$\lambda(t|x_{ij}, u_i) = \lambda_{0i}(t) \exp\{x_{ij}\beta\}. \tag{5.3}$$

Το μοντέλο (5.3) στοχεύει στην ίδια παράμετρο β με το shared frailty μοντέλο. Πρέπει να τονιστεί ότι το στρωματοποιημένο Cox μοντέλο δεν χρειάζεται τη συνθήκη (5.2) για να δώσει έγκυρες εκτιμήσεις του β, αφού στην πιθανοφάνεια του μοντέλου (5.3) συμμετέχουν μόνο άτομα της ίδιας ομάδας (Sjölander et al. 2013).

Μια διαφορετική προσέγγιση θα ήταν να θεωρήσουμε ότι οι περιθώριοι κίνδυνοι (ως προς τις ομάδες) είναι αναλογικοί. Σε αυτή την περίπτωση, στοχεύουμε σε μια διαφορετική παράμετρο β^*, όπου το e^{β^*} είναι το πηλίκο των κινδύνων συγκρίνοντας δυο τυχαία επιλεγόμενα άτομα (ανεξαρτήτως της ομάδας που ανήκουν), εκ των οποίων το πρώτο άτομο παίρνει θεραπεία ενώ το δεύτερο όχι (Glidden & Vittinghoff 2004).

$$\frac{\lambda(t|x)}{\lambda(t|x')} = \exp\{\beta^*(x - x')\}. \tag{5.4}$$

Όπως έχουμε δει, η σχέση (5.4) δεν υπονοεί αλλά ούτε υπονοείται από τη σχέση (5.1). Επίσης, ακόμα και στην περίπτωση όπου η έκθεση $\mathbf{X}_i$ είναι ανεξάρτητη από το u_i (όπως στο παράδειγμα μας), επομένως οι παράγοντες οι οποίοι βρίσκονται στο u_i δεν γίνεται να έχουν συγχυτικές επιδράσεις, οι επιδράσεις β^* και β θα είναι διαφορετικές. Το φαινόμενο αυτό ονομάζεται 'non-collapsibility' (Greenland 1996). Μπορεί να αποδειχθεί ότι, η μερική πιθανοφάνεια του Cox, αγνοώντας τη συσχέτιση των δεδομένων, αποδίδει συνεπείς εκτιμήσεις για το β^* (Lee, Wei, Amato & Leurgans 1992). Όμως, η διακύμανση δεν εκτιμάται σωστά, λόγω της συσχέτισης των παρατηρήσεων, και γι αυτό προτείνεται η διόρθωση του πίνακα συνδιακύμανσης των εκτιμήσεων (Lee et al. 1992).

Επειδή η περιθώρια και η δεσμευμένη συσχέτιση δεν ταυτίζονται, θα επικεντρωθούμε στα μοντέλα που στοχεύουν στην παράμετρο β. Θα προσομοιώσουμε 350 σύνολα δεδομένων, από το Gamma και το Lognormal shared frailty μοντέλο, με Weibull συνάρτηση κινδύνου. Κάθε

σετ δεδομένων αποτελείται από $G = 200$ ζευγάρια. Σε κάθε σύνολο δεδομένων θα εφαρμοστεί η MCMC μεθοδολογία η οποία παρουσιάστηκε στο κεφάλαιο 4. Επομένως, μπορούμε να εξετάσουμε αν η εκτίμηση του β επηρεάζεται από την κατανομή που θα θεωρήσουμε για τα frailties. Οι εκτιμήσεις μέγιστης πιθανοφάνειας θα παρουσιαστούν για σύγκριση. Τα μοντέλα θα συγκριθούν με το κριτήριο DIC.

5.1 Προσομοίωση Δεδομένων

Για να εφαρμόσουμε τα μοντέλα, πρέπει πρώτα να προσομοιώσουμε τα δεδομένα. Ο ακόλουθος αλγόριθμος χρησιμοποιήθηκε για να προσομοιώσουμε τα δεδομένα από ένα shared frailty μοντέλο. Οι χρόνοι επιβίωσης ακολουθούν ένα μοντέλο αναλογικών κινδύνων, δεδομένης της τιμής frailty u_i. Όπως είδαμε στο κεφάλαιο 4, η συνάρτηση επιβίωσης της Weibull κατανομής δίνεται από:

$$S(t|x_{ij}, u_i) = \exp\left\{-\lambda t^{\kappa} e^{x_{ij}\beta} u_i\right\}.$$

Το β είναι μονοδιάστατο γιατί έχουμε θεωρήσει μόνο μια επεξηγηματική μεταβλητή. Είναι γνωστό ότι η συνάρτηση επιβίωσης ακολουθεί την ομοιόμορφη κατανομή στο (0,1). Άρα,

$$U = \exp\left\{-\lambda t^{\kappa} e^{x_{ij}\beta} u_i\right\} \sim U(0,1).$$

Λύνοντας ως προς t βρίσκουμε ότι:

$$t = \left(\frac{-\log(U)}{\lambda e^{x_{ij}\beta} u_i}\right)^{1/\kappa}. \tag{5.5}$$

Χρησιμοποιώντας τη σχέση (5.5), μπορούμε να προσομοιώσουμε εύκολα τους χρόνους επιβίωσης. Τα frailties $u_1, u_2, \ldots, u_G$ αποτελούν τυχαίο δείγμα, είτε από τη Gamma κατανομή με μέση τιμή 1 και διακύμανση θ, είτε σε λογαριθμική κλίμακα κλίμακα τα $b_i = \log(u_i)$, $i = 1, 2, \ldots, G$, ακολουθούν την κανονική κατανομή με μέση τιμή 0 και διακύμανση σ^2. Οι χρόνοι λογοκρισίας ακολουθούν την ομοιόμορφη κατανομή στο διάστημα $(0, 6.5)$ όταν τα frailties ακολουθούν την Gamma κατανομή και την ομοιόμορφη κατανομή στο $(0, 5.3)$ όταν η κατανομή των frailties είναι η Lognormal. Οι τιμές 5.3 και 6.5 επιλέχθηκαν πειραματικά, έτσι ώστε να έχουμε περίπου 30% λογοκρισία και στις δυο περιπτώσεις.

Σε κάθε ζευγάρι διδύμων το ένα άτομο παίρνει θεραπεία ενώ το άλλο όχι. Θέτουμε την παράμετρο β ίση με $-\log(2)$, έτσι ώστε ο κίνδυνος των ατόμων που δεν παίρνουν θεραπεία να είναι διπλάσιος σε σχέση με τα άτομα που παίρνουν θεραπεία. Η παράμετρος λ επιλέχθηκε ίση με $\log(2)$ ώστε, ο διάμεσος χρόνος επιβίωσης των ατόμων που δεν παίρνουν θεραπεία να είναι ίσος

με 1 χρόνο, δοθέντος ότι $u_i = 1$. Για να έχουμε αύξουσα συνάρτηση κινδύνου (δοθέντων των frailties), θέτουμε $\kappa = 1.6$. Θεωρούμε ότι υπάρχει έντονη θετική συσχέτιση μεταξύ των χρόνων επιβίωσης των διδύμων, άρα, θέτουμε $\theta = 0.6$ και $\sigma^2 = 0.8$, στην περίπτωση της Gamma και Lognormal κατανομής, αντίστοιχα. Αν η τυχαία μεταβλητή u ακολουθεί την Gamma κατανομή με μέση τιμή 1 και διακύμανση 0.6, τότε η διακύμανση της $\log(u)$ θα είναι ίση με $\psi'(1/0.6) \cong 0.81$ όπου, $\psi'(\cdot)$, είναι η trigamma συνάρτηση (Duchateau & Janssen 2007, σελ. 228). Σε κάθε σύνολο δεδομένων εφαρμόζουμε 4 μοντέλα. Θα εφαρμόσουμε το (σωστό) παραμετρικό μοντέλο Weibull με frailty κατανομές την Gamma και Lognormal. Σε πρακτικές εφαρμογές είναι συχνά δύσκολο να επαληθευθεί η ορθότητα ενός παραμετρικού μοντέλου. Είναι γενικά καλή πρακτική να συγκρίνουμε τα αποτελέσματα ενός αυστηρά παραμετρικού μοντέλου, με ένα μοντέλο το οποίο κάνει σαφώς λιγότερες υποθέσεις σε σχέση με τη βασική συνάρτηση κινδύνου. Με αυτήν την αφορμή, θα εξετάσουμε την περίπτωση ενός κατά τμήματα εκθετικού μοντέλου, αφού διαμερίσουμε το χρονικό άξονα σε 20 διαστήματα τα οποία περιέχουν περίπου ίσο αριθμό συμβάντων (άρα ίση πληροφορία).

5.2 Αποτελέσματα για τα Προσομοιωμένα Δεδομένα

Η γλώσσα προγραμματισμού `R` (version 2.15), χρησιμοποιήθηκε για να προσομοιωθούν τα δεδομένα. Για να εφαρμόσουμε τα μοντέλα με τη μέθοδο της μέγιστης πιθανοφάνειας, χρησιμοποιήθηκε η βιβλιοθήκη `frailtypack`. Όπως έχουμε δει στο κεφάλαιο 4, στα μοντέλα Gamma τα frailties μπορούν να ολοκληρωθούν από την πιθανοφάνεια. Επομένως, η περιθώρια πιθανοφάνεια μπορεί να μεγιστοποιηθεί με συνήθεις μεθόδους. Όμως, στα Lognormal μοντέλα το ολοκλήρωμα της (4.16) δεν μπορεί να βρεθεί αναλυτικά. Η συνάρτηση `frailtyPenal`, της βιβλιοθήκης `frailtypack`, προσεγγίζει το ολοκλήρωμα με Gaussian quadrature μεθόδους. Αν η μέθοδος δεν είχε συγκλίνει (8% των περιπτώσεων), χρησιμοποιήθηκε ως αρχική τιμή το $\sigma^2 = 0.8$.

Επίσης, εφαρμόστηκαν οι MCMC αλγόριθμοι οι οποίοι αναλύθηκαν στο κεφάλαιο 4. Η παραμετροποίηση είναι ακριβώς ίδια με το κεφάλαιο 4. Για το Weibull μοντέλο θεωρούμε $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_0, \boldsymbol{C}_0)$, όπου $\boldsymbol{\mu}_0 = \mathbf{0}$ και $\boldsymbol{C}_0 = \text{diag}\{100\}$, $\kappa \sim \text{Gamma}(0.001, 0.001)$ και $\omega \sim \text{Gamma}(0.001, 0.001)$. Στο Weibull μοντέλο το $\boldsymbol{\beta}$ είναι διδιάστατο γιατί έχουμε ενσωματώσει τη σταθερά, $\beta_0 = \log(\lambda)$. Για το κατά τμήματα εκθετικό μοντέλο θέτουμε $\beta \sim N(0, 100)$, αφού το β είναι μονοδιάστατο, $\lambda_j \sim \text{Gamma}(0.001, 0.001)$ $j = 1, 2, \ldots, 20$ και $\omega \sim \text{Gamma}(0.001, 0.001)$. Σε κάθε μοντέλο προσομοιώνουμε μια μαρκοβιανή αλυσίδα 100000 επαναλήψεων, με 3000 επαναλήψεις να χρησιμοποιούνται ως burn-in. Για να μειώσουμε την αυτοσυσχέτιση της αλυσίδας, κρατάμε μια επανάληψη στις 10. Ως αρχικές τιμές των αλυσίδων, επιλέχθηκαν οι εκτιμήσεις μέγιστης πιθανοφάνειας των αντίστοιχων μοντέλων τα οποία δεν περιλαμβάνουν τυχαίες επιδράσεις. Για την εφαρμογή των MCMC αλγορίθμων, γράφτηκε κώδικας στην `R`.

Πίνακας 5.1: Αποτελέσματα MCMC αλγορίθμων για ένα προσομοιωμένο δείγμα από την Weibull κατανομή, στο οποίο η πραγματική κατανομή των frailties είναι η Lognormal.

Βασική Κατανομή	Frailty Κατανομή	Εκ των υστέρων Μέση τιμή	Εκ των υστέρων Τυπική απόκλιση	Εκ των υστέρων Διάμεσος	95% Διάστημα Αξιοπιστίας	
Piecewise Εκθετική						
β	Gamma	-0.625	0.137	-0.624	-0.902	-0.359
β	Lognormal	-0.668	0.141	-0.667	-0.952	-0.399
θ	Gamma	0.522	0.165	0.511	0.231	0.875
σ^2	Lognormal	0.913	0.301	0.882	0.405	1.588
Weibull						
β	Gamma	-0.630	0.136	-0.626	-0.900	-0.370
β	Lognormal	-0.679	0.140	-0.677	-0.958	-0.406
θ	Gamma	0.548	0.155	0.538	0.271	0.882
σ^2	Lognormal	0.980	0.290	0.951	0.487	1.626

Πριν παρουσιαστούν συνοπτικά τα αποτελέσματα για όλα τα σύνολα δεδομένων, θα εξετάσουμε το πρώτο προσομοιωμένο δείγμα στο οποίο η πραγματική κατανομή των frailties είναι η Lognormal. Στο συγκεκριμένο δείγμα, το ποσοστό λογοκρισίας ήταν 73%. Θα εξετάσουμε γραφικά τη σύγκλιση των MCMC αλγορίθμων για το Weibull μοντέλο με Gamma frailties και το, κατά τμήματα, εκθετικό μοντέλο με Lognormal frailties. Η σύγκλιση στην εκ των υστέρων κατανομή, φαίνεται ότι έχει επιτευχθεί για τις παραμέτρους με βάση τα σημειογράμματα (trace plots) (γραφήματα 5.1-5.2 και 5.4-5.5). Όμως, βλέπουμε ότι υπάρχει μικρή αυτοσυσχέτιση για τις τιμές που προσομοιώνονται για το θ και σ^2 (γραφήματα 5.3 και 5.6). Βέβαια, για τις 2800 αλυσίδες που προσομοιώθηκαν συνολικά, δεν ήταν εφικτό να εφαρμόσουμε μεγαλύτερη εκλέπτυνση. Στον πίνακα (5.1) βλέπουμε τα αποτελέσματα των 4 αλγορίθμων. Το πραγματικό μοντέλο προέρχεται από την Weibull κατανομή με Lognormal frailties. Παρατηρούμε ότι το κατά τμήματα εκθετικό μοντέλο δίνει πολύ κοντινές εκτιμήσεις με το Weibull μοντέλο. Αυτό συμβαίνει διότι έχουμε διαμερίσει το χρονικό άξονα σε 20 διαστήματα, άρα, η κατά τμήματα εκθετική κατανομή προσεγγίζει αρκετά καλά την πραγματική μορφή της συνάρτησης κινδύνου. Δεν παρατηρούμε μεγάλες διαφορές στα συμπεράσματα μεταξύ των Gamma και Lognormal μοντέλων.

Οι παράμετροι θ και σ^2 δεν είναι άμεσα συγκρίσιμες γιατί περιγράφουν τη διακύμανση σε διαφορετική κλίμακα. Στον πίνακα (5.2) παρουσιάζονται τα αποτελέσματα με τη μέθοδο της μέγιστης πιθανοφάνειας. Παρατηρούμε ότι τα μέτρα θέσης (μέση τιμή και διάμεσος) της εκ των υστέρων κατανομής του β, συμφωνούν με τις εκτιμήσεις μέγιστης πιθανοφάνειας. Το αποτέλεσμα αυτό είναι λογικό διότι, οι εκ των υστέρων κατανομές του β είναι συμμετρικές και μονοκόρυφες. Επειδή η εκ των υστέρων κατανομή της διασποράς σ^2 είναι θετικά ασύμμετρη, παρατηρούμε ότι η εκ των υστέρων διάμεσος συμπίπτει με την εκτίμηση μέγιστης πιθανοφάνειας.

Πίνακας 5.2: Εκτιμήσεις μέγιστης πιθανοφάνειας για ένα προσομοιωμένο δείγμα από την Weibull κατανομή, στο οποίο η πραγματική κατανομή των frailties είναι η Lognormal.

Βασική Κατανομή	Frailty Κατανομή	Εκτίμηση μέγιστης Πιθανοφάνειας	Τυπικό σφάλμα Εκτίμησης	95% Διάστημα Εμπιστοσύνης	
Piecewise Εκθετική					
β	Gamma	-0.623	0.137	-0.892	-0.354
β	Lognormal	-0.668	0.141	-0.944	-0.391
θ	Gamma	0.510	0.162	0.273	0.953
σ^2	Lognormal	0.879	0.282	0.469	1.649
Weibull					
β	Gamma	-0.630	0.135	-0.895	-0.366
β	Lognormal	-0.679	0.140	-0.953	-0.405
θ	Gamma	0.545	0.152	0.316	0.941
σ^2	Lognormal	0.951	0.277	0.536	1.684

Μικρές διαφορές μπορούν να εντοπιστούν μόνο μεταξύ των διαστημάτων αξιοπιστίας και εμπιστοσύνης. Τα διαστήματα εμπιστοσύνης έχουν κατασκευαστεί με τη Δ-μέθοδο. Αφού βρούμε το προσεγγιστικό τυπικό σφάλμα του $\log(\widehat{\theta})$ με τη Δ-μέθοδο, κατασκευάζουμε ένα 95% διάστημα εμπιστοσύνης για το $\log(\theta)$, και στη συνέχεια εκθετικοποιούμε το αποτέλεσμα. Για να έχει η Δ-μέθοδος καλά αποτελέσματα, πρέπει η κατανομή του $\log(\widehat{\theta})$ να είναι προσεγγιστικά κανονική. Επομένως, τα διαστήματα αξιοπιστίας και εμπιστοσύνης του θ και σ^2, θα συμφωνούν όταν οι εκ των υστέρων κατανομές του $\log(\theta)$ και $\log(\sigma^2)$ προσεγγίζονται από την κανονική.

Στους πίνακες (5.4) και (5.5) παρουσιάζονται τα συνολικά αποτελέσματα από όλα τα δείγματα. Το συμπέρασμα είναι ότι όλα τα μοντέλα εκτιμούν (σχεδόν) το ίδιο καλά την επίδραση της θεραπείας, ανεξάρτητα από το αν η κατανομή που έχουμε θεωρήσει για τα frailties διαφέρει από τη σωστή. Η μεροληψία είναι μικρή σε όλες τις περιπτώσεις. Βέβαια, έχουμε θεωρήσει μόνο δυο frailty κατανομές. Οι Balakrishnan & Peng (2006) χρησιμοποίησαν ως κατανομές για τα frailties, τις κατανομές Gamma, Weibull, Lognormal και Generalized Gamma. Βρήκαν ότι η frailty κατανομή έχει μικρή επίδραση στην εκτίμηση της επίδρασης της θεραπείας β. Το κατά τμήματα εκθετικό μοντέλο τείνει να έχει μεγαλύτερη διασπορά στις εκτιμήσεις σε σχέση με το Weibull μοντέλο, ίσως λόγω των πολλών παραμέτρων που χρησιμοποιούνται.

Αρχικά, ας επικεντρωθούμε στα Weibull μοντέλα. Η εκ των υστέρων μέση τιμή και διάμεσος του β δίνουν, κατά μέσο όρο, τις ίδιες εκτιμήσεις με τους εκτιμητές μέγιστης πιθανοφάνειας. Η εκ των υστέρων κατανομή του β είναι συνήθως συμμετρική, άρα, η μέγιστη τιμή της κατανομής θα βρίσκεται αρκετά κοντά στη μέση τιμή και τη διάμεσο. Επίσης, η διασπορά των εκτιμήσεων φαίνεται να είναι η ίδια και στις δυο μεθόδους. Η εκτίμηση μέγιστης πιθανοφάνειας του σ^2 (Weibull lognormal μοντέλο) συμφωνεί, κατά μέσο όρο, περισσότερο με την εκ των υστέρων διάμεσο του σ^2 παρά με μέση τιμή του. Η εκ των υστέρων κατανομή του σ^2 είναι συνήθως

ισχυρά θετικά ασύμμετρη, άρα, έτσι θα μπορούσε να εξηγηθεί γιατί η εκ των υστέρων μέση τιμή υπερεκτιμά, σε κάποιο βαθμό, την πραγματική τιμή του σ^2. Αντιθέτως, σε όλα τα Gamma μοντέλα, η διακύμανση θ εκτιμάται αμερόληπτα από την εκ των υστέρων μέση τιμή και τη διάμεσο.

Οι ιδιότητες της επαναλαμβανόμενης δειγματοληψίας, όπως αμεροληψία και πιθανότητα κάλυψης, ισχύουν γενικά για τους εκτιμητές μέγιστης πιθανοφάνειας. Τα διαστήματα αξιοπιστίας δεν έχουν σχεδιαστεί (Fraser et al. 2011, Marchand & Strawderman 2012), και ούτε μπορούν να εγγυηθούν ότι κατέχουν τη σωστή πιθανότητα κάλυψης των παραμέτρων. Παρ όλα αυτά, θα είχε νόημα να διερευνήσουμε την πιθανότητα κάλυψης των διαστημάτων αξιοπιστίας. Σε όλα τα Weibull μοντέλα, βλέπουμε ότι τα διαστήματα αξιοπιστίας έχουν πιθανότητα κάλυψης κοντά στο θεωρητικό 95%, σε όλες τις περιπτώσεις. Το κατά τμήματα εκθετικό μοντέλο δίνει σχεδόν τις ίδιες εκτιμήσεις με το Weibull μοντέλο, αλλά έχει ελαφρώς μικρότερη πιθανότητα κάλυψης για το θ και σ^2. Αντίθετα, η πιθανότητα κάλυψης για το β βρίσκεται κοντά στο 95%, σε όλες τις περιπτώσεις. Στην ενότητα 4.3 θα εξετάσουμε την ευαισθησία της εκ των υστέρων κατανομής του θ και σ^2, ως προς την επιλογή των υπερ-παραμέτρων της εκ των προτέρων κατανομής τους, στην περίπτωση της κατά τμήματα εκθετικής κατανομής.

Πίνακας 5.3: Αναλογία αποδοχής του κάθε μοντέλου σύμφωνα με το DIC κριτήριο.

Weibull κατανομή	Εφαρμοσμένη frailty κατανομή	
Σωστή frailty κατανομή	Gamma	Lognormal
Gamma	0.717	0.283
Lognormal	0.340	0.660
Piecewise εκθετική κατανομή	Εφαρμοσμένη frailty κατανομή	
Σωστή frailty κατανομή	Gamma	Lognormal
Gamma	0.694	0.306
Lognormal	0.334	0.666

Θα εξετάσουμε το κριτήριο DIC ως ένα μέτρο επιλογής μοντέλου. Μοντέλα με μικρότερο DIC είναι προτιμητέα. Πρέπει να τονιστεί ότι έχουμε χρησιμοποιήσει την περιθώρια πιθανοφάνεια στο υπολογισμό του κριτηρίου DIC. Επομένως, το ενδιαφέρον δεν εστιάζεται στις τιμές των frailties $u_1, u_2, \ldots, u_G$. Ωστόσο, τα frailties συμμετέχουν στην διαμόρφωση του σχήματος της περιθώριας πιθανοφάνειας. Σε κάθε προσομοιωμένο δείγμα εφαρμόσαμε 4 μοντέλα. Δεν έγιναν συγκρίσεις μεταξύ του Weibull και κατά τμήματα εκθετικού μοντέλου, διότι το κατά τμήματα εκθετικό μοντέλο είναι υπερ-παραμετροποιημένο για τα συγκεκριμένα δεδομένα. Στον πίνακα (5.3) συνοψίζονται τα αποτελέσματα. Βλέπουμε ότι, όταν η frailty κατανομή είναι στην πραγματικότητα η Gamma, το κριτήριο DIC επιλέγει τη σωστή frailty κατανομή στο 70% περίπου των περιπτώσεων, ανεξάρτητα από την κατανομή που θα θεωρήσουμε για τη βασική συνάρτηση κινδύνου. Επίσης, στην περίπτωση όπου η αληθινή κατανομή των frailties είναι η

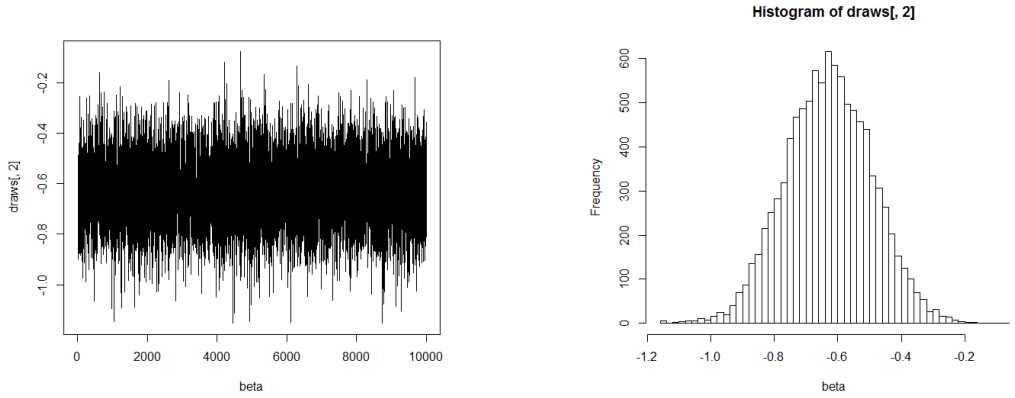
Lognormal, το κριτήριο DIC επιλέγει τη σωστή frailty κατανομή στο 66% περίπου των περιπτώσεων, ανεξαρτήτως της βασικής συνάρτησης κινδύνου.

Πίνακας 5.4: Αποτελέσματα προσομοιωμένων δεδομένων. Τα δεδομένα έχουν προσομοιωθεί απο την Weibull κατανομή με Gamma frailties. Οι πραγματικές τιμές των β και θ είναι $\beta = -0.693$ και $\theta = 0.6$.

Βασική κατανομή	Frailty κατανομή	Μέση τιμή	Μέροληψία απόλυτη τιμή	Monte Carlo τυπικό σφάλμα	95% Πιθανότητα κάλυψης
Weibull	Gamma (MCMC)				
β (Μέση τιμή)		-0.692	0.001	0.138	95.4
β (Διάμεσος)		-0.691	0.002	0.137	
θ (Μέση τιμή)		0.606	0.006	0.153	95.4
θ (Διάμεσος)		0.596	0.004	0.152	
Weibull	Gamma				
β (Ε.Μ.Π.)		-0.693	<0.001	0.138	95.4
θ (Ε.Μ.Π.)		0.604	0.004	0.146	-
Weibull	Lognormal (MCMC)				
β (Μέση τιμή)		-0.695	0.001	0.139	95.1
β (Διάμεσος)		-0.694	<0.001	0.139	
σ^2 (Μέση τιμή)		0.821	-	0.250	-
σ^2 (Διάμεσος)		0.799	-	0.246	
Weibull	Lognormal				
β (Ε.Μ.Π.)		-0.695	0.002	0.139	94.9
σ^2 (Ε.Μ.Π.)		0.802	-	0.236	-
Piecewise Εκθετική	Gamma (MCMC)				
β (Μέση τιμή)		-0.691	0.002	0.141	95.7
β (Διάμεσος)		-0.689	0.004	0.141	
θ (Μέση τιμή)		0.606	0.006	0.189	92.6
θ (Διάμεσος)		0.596	0.004	0.188	
Piecewise Εκθετική	Lognormal (MCMC)				
β (Μέση τιμή)		-0.685	0.009	0.140	95.4
β (Διάμεσος)		-0.683	0.010	0.140	
σ^2 (Μέση τιμή)		0.782	-	0.284	-
σ^2 (Διάμεσος)		0.758	-	0.280	

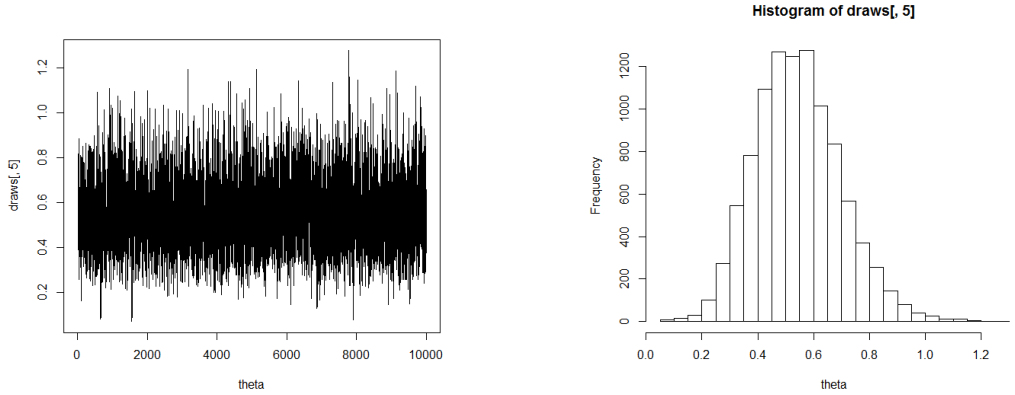
Πίνακας 5.5: Αποτελέσματα προσομοιωμένων δεδομένων. Τα δεδομένα έχουν προσομοιωθεί απο την Weibull κατανομή με Lognormal frailties. Οι πραγματικές τιμές των β και σ^2 είναι $\beta = -0.693$ και $\sigma^2 = 0.8$.

Βασική κατανομή	Frailty κατανομή	Μέση τιμή	Μέροληψία απόλυτη τιμή	Monte Carlo τυπικό σφάλμα	95% Πιθανότητα κάλυψης
Weibull	Gamma (MCMC)				
β (Μέση τιμή)		-0.676	0.017	0.137	95.4
β (Διάμεσος)		-0.675	0.018	0.136	
θ (Μέση τιμή)		0.574	-	0.165	-
θ (Διάμεσος)		0.565	-	0.164	
Weibull	Gamma				
β (Ε.Μ.Π.)		-0.678	0.015	0.137	95.4
θ (Ε.Μ.Π.)		0.576	-	0.156	-
Weibull	Lognormal (MCMC)				
β (Μέση τιμή)		-0.694	0.001	0.142	95.7
β (Διάμεσος)		-0.693	0.001	0.141	
σ^2 (Μέση τιμή)		0.845	0.045	0.272	95.4
σ^2 (Διάμεσος)		0.821	0.021	0.268	
Weibull	Lognormal				
β (Ε.Μ.Π.)		-0.695	0.001	0.141	94.6
σ^2 (Ε.Μ.Π.)		0.829	0.029	0.257	-
Piecewise Εκθετική	Gamma (MCMC)				
β (Μέση τιμή)		-0.683	0.010	0.143	94.6
β (Διάμεσος)		-0.682	0.011	0.143	
θ (Μέση τιμή)		0.599	-	0.200	-
θ (Διάμεσος)		0.589	-	0.200	
Piecewise Εκθετική	Lognormal (MCMC)				
β (Μέση τιμή)		-0.694	0.001	0.146	94.3
β (Διάμεσος)		-0.693	<0.001	0.146	
σ^2 (Μέση τιμή)		0.851	0.051	0.312	92.9
σ^2 (Διάμεσος)		0.825	0.025	0.308	



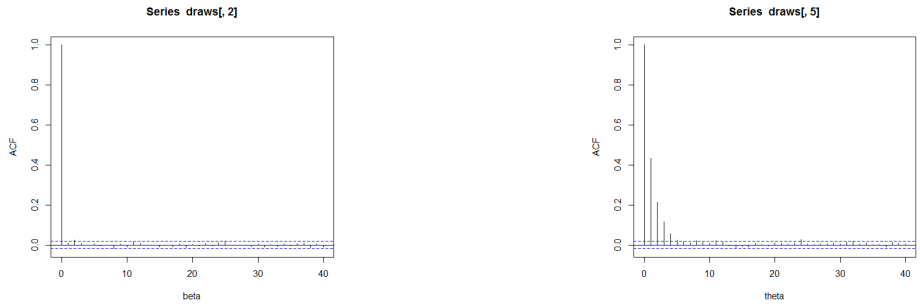
(α′) Trace plot (β′) Histogram

Σχήμα 5.1: Σημειόγραμμα και ιστόγραμμα του β από ένα Weibull μοντέλο με gamma frailties. Τα δεδομένα προέρχονται από το πρώτο προσομοιωμένο δείγμα και η πραγματική κατανομή των frailties είναι η lognormal.



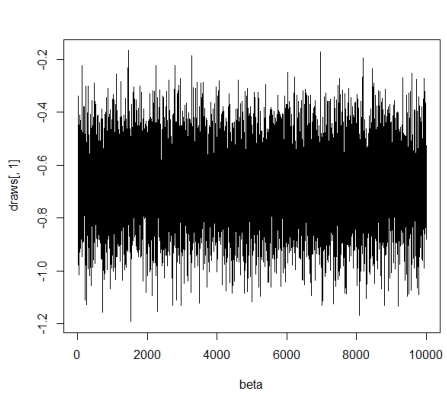
(α′) Trace plot (β′) Histogram

Σχήμα 5.2: Σημειόγραμμα και ιστόγραμμα του θ από ένα Weibull μοντέλο με gamma frailties. Τα δεδομένα προέρχονται από το πρώτο προσομοιωμένο δείγμα και η πραγματική κατανομή των frailties είναι η lognormal.

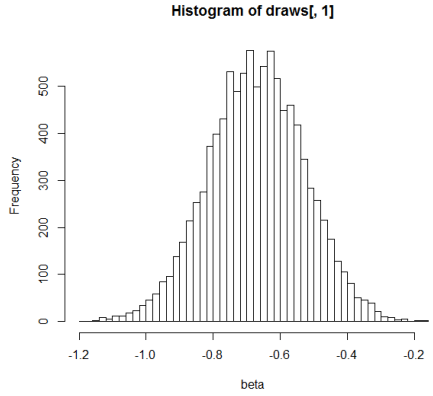


(α′) Autocorrelation plot (β′) Autocorrelation plot

Σχήμα 5.3: Γραφήματα αυτοσυσχέτισης του β και θ από ένα Weibull μοντέλο με gamma frailties. Τα δεδομένα προέρχονται από το πρώτο προσομοιωμένο δείγμα και η πραγματική κατανομή των frailties είναι η lognormal.

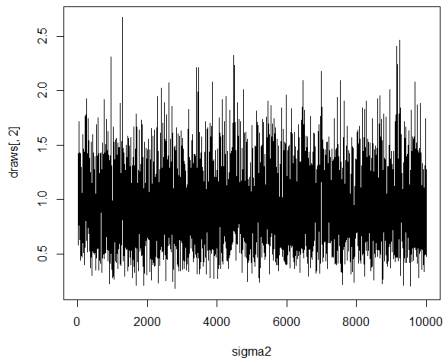


(α′) Trace plot

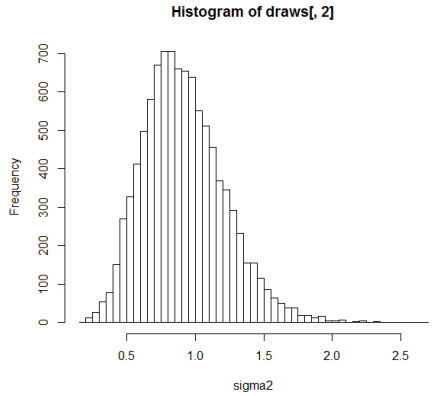


(β′) Histogram

Σχήμα 5.4: Σημειόγραμμα και ιστόγραμμα του β από ένα κατά τμήματα εκθετικό μοντέλο με lognormal frailties. Τα δεδομένα προέρχονται από το πρώτο προσομοιωμένο δείγμα και η πραγματική κατανομή των frailties είναι η lognormal.

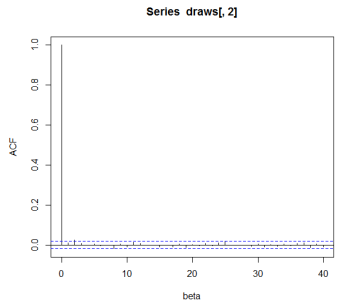


(α′) Trace plot

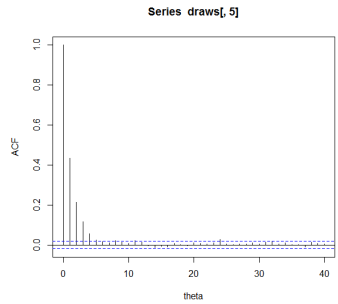


(β′) Histogram

Σχήμα 5.5: Σημειόγραμμα και ιστόγραμμα του σ^2 από ένα κατά τμήματα εκθετικό μοντέλο με lognormal frailties. Τα δεδομένα προέρχονται από το πρώτο προσομοιωμένο δείγμα και η πραγματική κατανομή των frailties είναι η lognormal.



(α′) Autocorrelation plot



(β′) Autocorrelation plot

Σχήμα 5.6: Γραφήματα αυτοσυσχέτισης του β και σ^2 από ένα κατά τμήματα εκθετικό μοντέλο με lognormal frailties. Τα δεδομένα προέρχονται από το πρώτο προσομοιωμένο δείγμα και η πραγματική κατανομή των frailties είναι η lognormal.

5.3 Ανάλυση Ευαισθησίας

Στην παρούσα ενότητα θα εξετάσουμε την ευαισθησία των αποτελεσμάτων, ως προς την επιλογή των υπερ-παραμέτρων των εκ των προτέρων κατανομών του θ και του σ^2. Τα Weibull μοντέλα, επειδή η πιθανότητα κάλυψης των παραμέτρων ήταν κοντά στο θεωρητικό 95%, δεν θα εξεταστούν περαιτέρω. Θα επικετρωθούμε στα 350 προσομοιωμένα δείγματα από την Weibull κατανομή με lognormal frailties. Θα εφαρμόσουμε μοντέλα σταθερού, κατά τμήματα, κινδύνου με διαφορετικές τιμές υπερ-παραμέτρων. Γενικά, τα αποτελέσματα δεν είναι τόσο ευαίσθητα ως προς την επιλογή της εκ των προτέρων κατανομής των παραμέτρων θέσης (για παράδειγμα β). Για τις παραμέτρους κλίμακας (για παράδειγμα θ ή σ^2), πρέπει να αποφασίσουμε που θα βάλουμε την εκ των προτέρων κατανομή: στη διακύμανση, στην τυπική απόκλιση ή στην ακρίβεια?

Οι Browne, Draper et al. (2006) εξέτασαν δυο εκ των προτέρων κατανομές σε ιεραρχικά μοντέλα, την Gamma κατανομή για την ακρίβεια και την ομοιόμορφη κατανομή για τη διακύμανση. Βρήκαν ότι τα διαστήματα αξιοπιστίας μπορεί να έχουν μικρότερη πιθανότητα κάλυψης από την αναμενόμενη σε κάποιες περιπτώσεις. Οι Lambert, Sutton, Burton, Abrams & Jones (2005), εφάρμοσαν 13 μη πληροφοριακές εκ των προτέρων κατανομές για τη διακύμανση μεταξύ των μελετών (between-study variance), σε μοντέλα τυχαίων επιδράσεων στη μετα-ανάλυση. Σε κάποιες (ρεαλιστικές για μετα-ανάλυση) περιπτώσεις, βρήκαν σημαντικές αποκλίσεις στα αποτελέσματα, παρόλο που οι εκ των προτέρων κατανομές θεωρούνταν μη πληροφοριακές. Ο Gelman et al. (2006) προτείνει, σε μεικτά ιεραρχικά μοντέλα, την ομοιόμορφη κατανομή για την τυπική απόκλιση.

Θα περιοριστούμε στην οικογένεια Gamma για την ακρίβεια (ισοδύναμα Inverse-Gamma για τη διακύμανση). Έστω ότι η ακρίβεια ω είναι ίση με $\omega = 1/\theta$ στην περίπτωση των Gamma frailties και $\omega = 1/\sigma^2$ στην περίπτωση των lognormal frailties. Έχουμε υποθέσει ότι η εκ των προτέρων κατανομή του ω είναι η $\omega \sim \text{Gamma}(c, d)$. Θα εξετάσουμε τρεις επιλογές για τα (c, d)

- $\omega \sim \text{Gamma}(0.001, 0.001)$
- $\omega \sim \text{Gamma}(0.01, 0.01)$
- $\omega \sim \text{Gamma}(0.1, 0.1)$

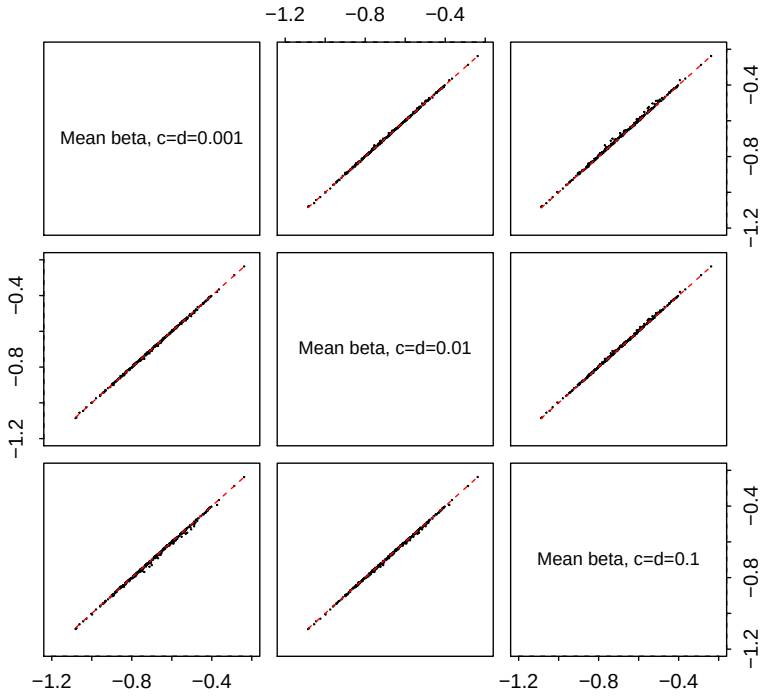
Τα αποτελέσματα της πρώτης επιλογής ($c = d = 0.001$) έχουν καταγραφεί στον πίνακα (5.5). Οι επιλογές των υπερ-παραμέτρων για τις υπόλοιπες παραμέτρους παρέμειναν ίδιες. Τα αποτελέσματα συνοψίζονται στον πίνακα (5.6). Είναι σαφές ότι η εκ των υστέρων μέση τιμή και διάμεσος του β δεν έχει επηρεαστεί, κατά μέσο όρο, από τις αλλαγές των υπερ-παραμέτρων c και d. Το ποσοστό κάλυψης του β έχει διαφοροποιηθεί λίγο, αλλά το γεγονός αυτό οφείλεται κυρίως σε οριακές περιπτώσεις. Αντίθετα, βλέπουμε ότι όσο αυξάνονται τα c και d, η εκ των υστέρων μέση τιμή και διάμεσος των διασπορών έχει την τάση να αυξάνεται, με αποτέλεσμα

Πίνακας 5.6: Ανάλυση ευαισθησίας ως προς την επιλογή των υπερ-παραμέτρων της εκ των προτέρων κατανομής των διασπορών θ και σ^2. Τα δεδομένα έχουν προσομοιωθεί απο την Weibull κατανομή με Lognormal frailties. Οι πραγματικές τιμές των β και σ^2 είναι $\beta = -0.693$ και $\sigma^2 = 0.8$.

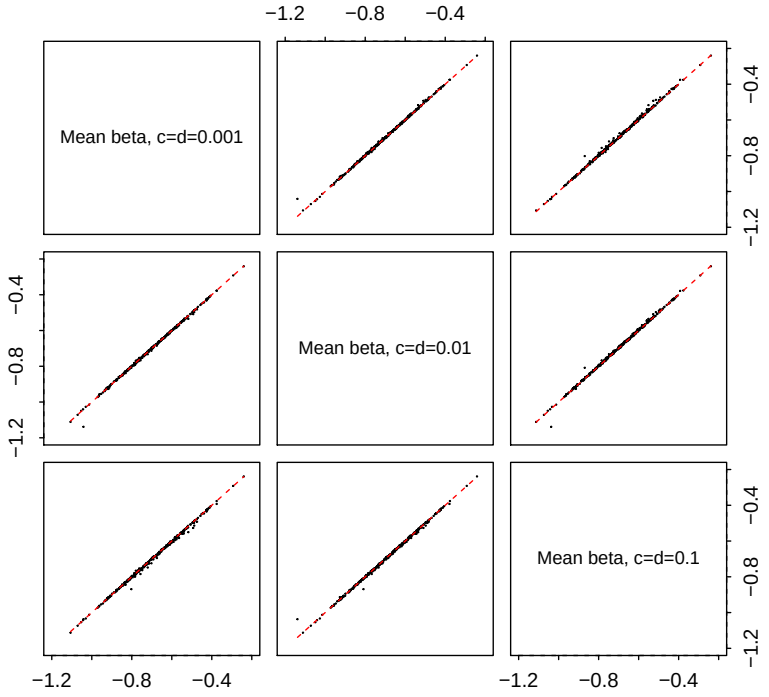
	$c = 0.001,\ d = 0.001$		$c = 0.01,\ d = 0.01$		$c = 0.1,\ d = 0.1$	
Μοντέλο	Μέση τιμή	Monte Carlo τυπ. σφάλμα	Μέση τιμή	Monte Carlo τυπ. σφάλμα	Μέση τιμή	Monte Carlo τυπ. σφάλμα
Gamma						
β (Μέση τιμή)	-0.683	0.143	-0.684	0.143	-0.687	0.143
β (Διάμεσος)	-0.682	0.143	-0.683	0.143	-0.685	0.143
θ (Μέση τιμή)	0.599	0.200	0.602	0.197	0.611	0.189
θ (Διάμεσος)	0.589	0.200	0.592	0.196	0.599	0.189
DIC	818.637		818.566		818.475	
β Κάλυψη %		94.6		94.3		94.3
Lognormal						
β (Μέση τιμή)	-0.694	0.146	-0.695	0.147	-0.697	0.146
β (Διάμεσος)	-0.693	0.146	-0.694	0.147	-0.695	0.146
σ^2 Μέση τιμή	0.851	0.312	0.855	0.311	0.862	0.299
σ^2 (Διάμεσος)	0.825	0.308	0.829	0.307	0.836	0.295
DIC	817.769		817.727		817.628	
β Κάλυψη %		94.3		95.4		94.9
σ^2 Κάλυψη %		92.9		92.9		93.7

να αυξάνεται και η μεροληψία (για το σ^2). Επίσης, παρατηρούμε ότι η πιθανότητα κάλυψης του σ^2 έχει αυξηθεί ελαφρώς όταν $c = d = 0.1$. Συνοπτικά, οι σημειακές εκτιμήσεις της εκ των υστέρων κατανομής, δεν φαίνεται να έχουν επηρεαστεί σε μεγάλο βαθμό. Το κριτήριο DIC έχει μειωθεί ελαφρώς, χωρίς όμως να μεροληπτεί υπέρ του ενός ή του άλλου μοντέλου.

Καλύτερη εικόνα για την επίδραση των υπερ-παραμέτρων c και d, έχουμε μέσω των γραφημάτων 5.7-5.10. Είναι λογικό ότι τα γραφήματα 5.7-5.10 θα περιέχουν τυχαία διακύμανση, διότι οι MCMC αλγόριθμοι είναι στοχαστικοί αλγόριθμοι. Επομένως, αξίζει να επικεντρωθούμε σε συστηματικές τάσεις. Στο γράφημα 5.7 βλέπουμε ξεκάθαρα ότι, η εκ των υστέρων μέση τιμή του β δεν επηρεάζεται από τα c και d. Στο γράφημα 5.8 βλέπουμε ότι, όταν η εκ των υστέρων μέση τιμή των διασπορών θ και σ^2 είναι μικρή, η επιλογή $c = d = 0.1$ τείνει να αυξάνει ελαφρώς το θ και σ^2. Το συμπέρασμα αυτό συμφωνεί με την εικόνα του πίνακα (5.6). Στο γράφημα 5.9 βλέπουμε ότι το 2.5% ποσοστημόριο της εκ των υστέρων κατανομής των διασπορών έχει αυξητική τάση όσο αυξάνονται τα c και d, υπό την προϋπόθεση ότι το ποσοστημόριο βρίσκεται κοντά στο 0. Αντίθετα, το 97.5% ποσοστημόριο δεν επηρεάζεται σε μεγάλο βαθμό (γράφημα 5.10).

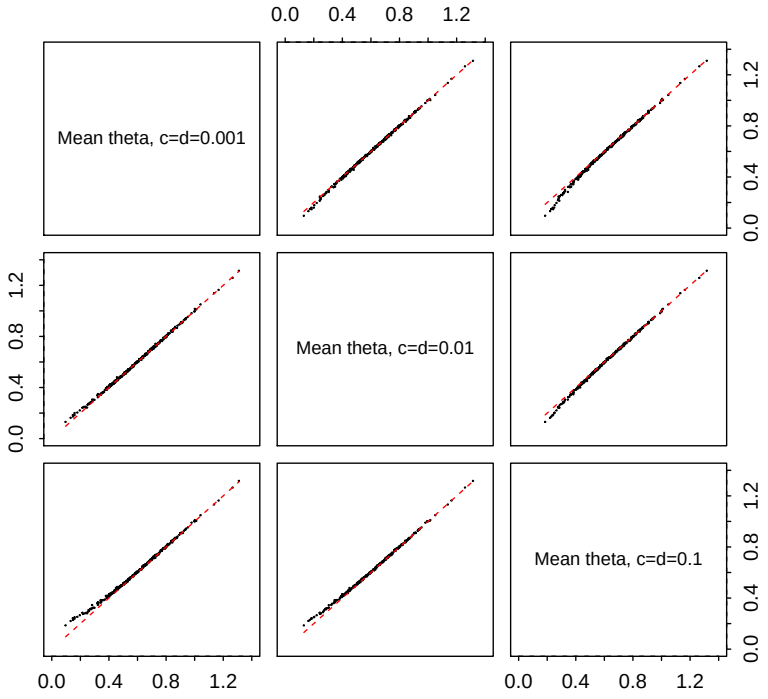


(α′) Gamma μοντέλο

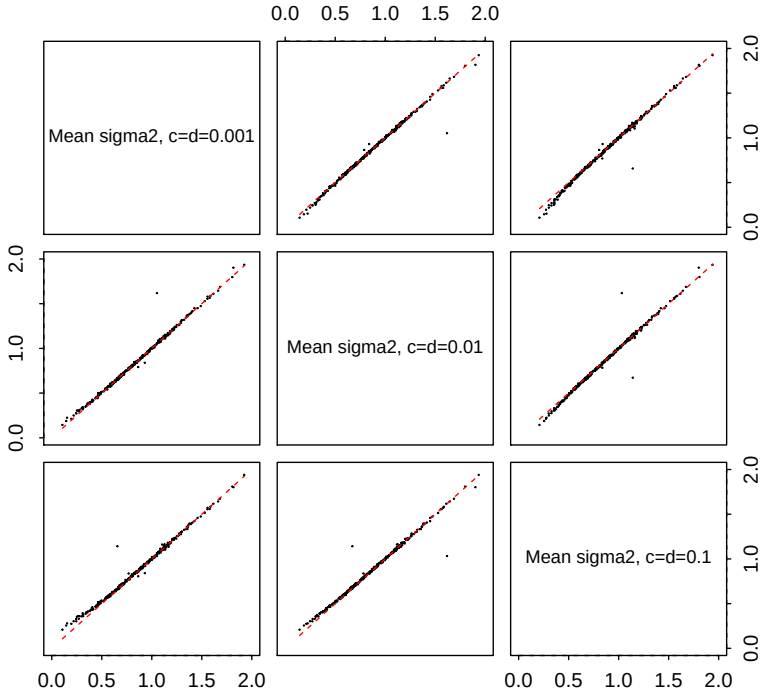


(β′) Lognormal μοντέλο

Σχήμα 5.7: Πίνακας διαγραμμάτων διασποράς (scatter plot matrix) της εκ των υστέρων μέσης τιμής του β για 3 διαφορετικές επιλογές υπερ-παραμέτρων. (α) Κατά τμήματα εκθετικό μοντέλο με Gamma frailties και (β) Κατά τμήματα εκθετικό μοντέλο με Lognormal frailties.

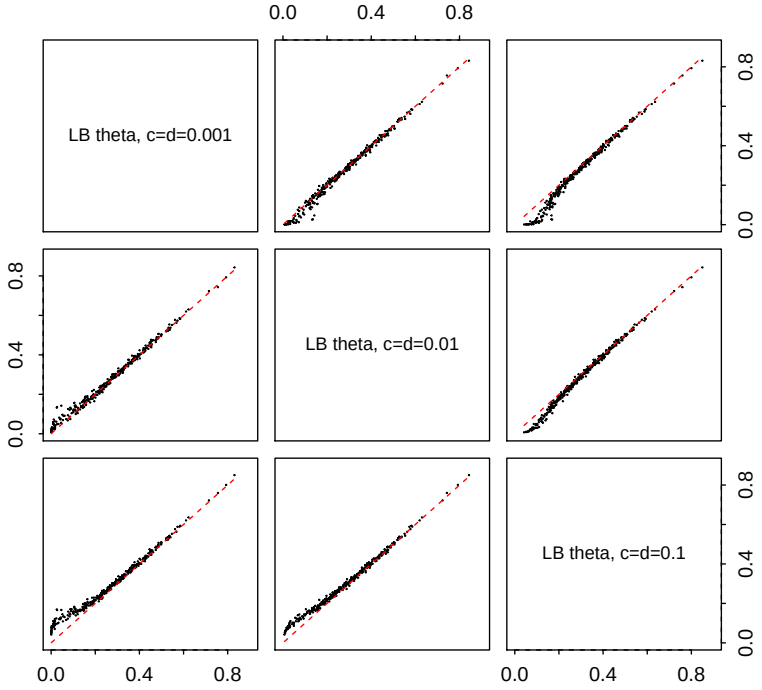


(α′) Gamma μοντέλο

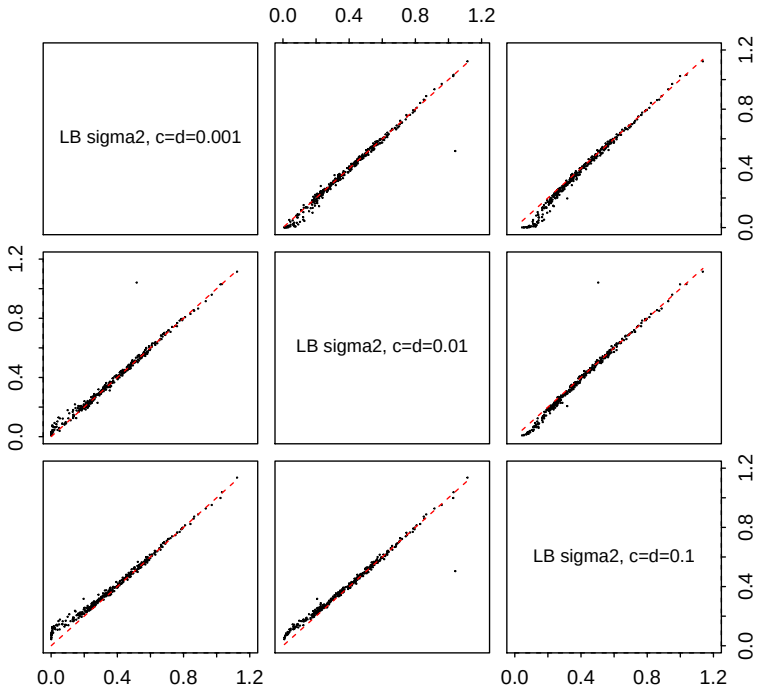


(β′) Lognormal μοντέλο

Σχήμα 5.8: Πίνακας διαγραμμάτων διασποράς (scatter plot matrix) της εκ των υστέρων μέσης τιμής του θ και σ^2 για 3 διαφορετικές επιλογές υπερ-παραμέτρων. (α) Κατά τμήματα εκθετικό μοντέλο με Gamma frailties και (β) Κατά τμήματα εκθετικό μοντέλο με Lognormal frailties.

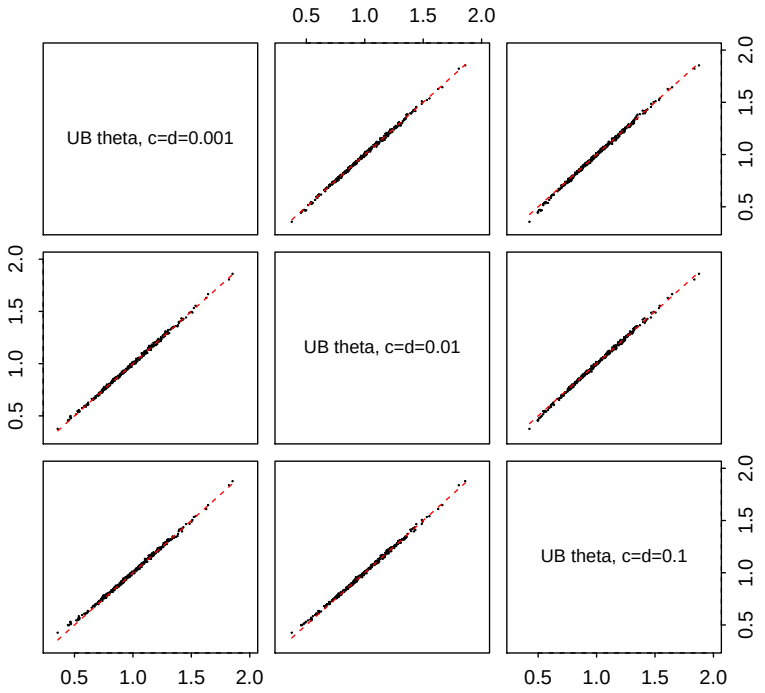


(α′) Gamma μοντέλο

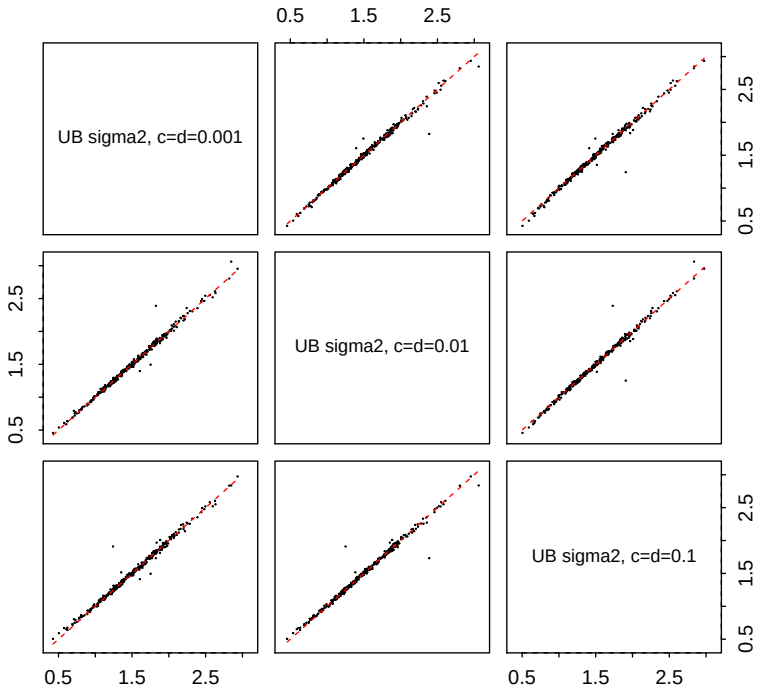


(β′) Lognormal μοντέλο

Σχήμα 5.9: Πίνακας διαγραμμάτων διασποράς (scatter plot matrix) του 2.5% ποσοστημορίου της εκ των υστέρων κατανομής του θ και σ^2 για 3 διαφορετικές επιλογές υπερ-παραμέτρων. (α) Κατά τμήματα εκθετικό μοντέλο με Gamma frailties και (β) Κατά τμήματα εκθετικό μοντέλο με Lognormal frailties.



(α′) Gamma μοντέλο



(β′) Lognormal μοντέλο

Σχήμα 5.10: Πίνακας διαγραμμάτων διασποράς (scatter plot matrix) του 97.5% ποσοστημορίου της εκ των υστέρων κατανομής του θ και σ^2 για 3 διαφορετικές επιλογές υπερ-παραμέτρων. (α) Κατά τμήματα εκθετικό μοντέλο με Gamma frailties και (β) Κατά τμήματα εκθετικό μοντέλο με Lognormal frailties.

Κεφάλαιο 6

Εφαρμογή σε Πραγματικά Δεδομένα

Στο παρόν κεφάλαιο, θα εφαρμόσουμε τη μεθοδολογία που παρουσιάστηκε στις προηγούμενες ενότητες σε μια πραγματική μελέτη. Θα εξετάσουμε τα δεδομένα που αναλύθηκαν από τους McGilchrist & Aisbett (1991), τα οποία αφορούν τους χρόνους επανεμφάνισης μόλυνσης νεφροπαθών ασθενών. Θα εφαρμόσουμε Gamma και Lognormal frailty μοντέλα τα οποία θα συγκριθούν μεταξύ τους με το κριτήριο DIC.

6.1 Χρόνοι Επανεμφανίσεων των Μολύνσεων

Θα ασχοληθούμε με μια μελέτη, η οποία παρουσιάστηκε αρχικά από τους McGilchrist & Aisbett (1991), και περιλαμβάνει δεδομένα σχετικά με 38 ασθενείς οι οποίοι χρησιμοποιούν φορητή συσκευή αιμοκάθαρσης (portable dialysis equipment). Ως συμβάν (event) θεωρείται η εμφάνιση μόλυνσης στο σημείο εισαγωγής του καθετήρα. Ο χρόνος μετράται από τη στιγμή εισαγωγής του καθετήρα μέχρι την εμφάνιση μόλυνσης στο σημείο εισαγωγής του. Κάθε άτομο συμμετέχει στη βάση δεδομένων με δυο καταγεγραμμένους χρόνους. Ο καθετήρας εισάγεται στον ασθενή και ο χρόνος μέχρι τη μόλυνση (σε μέρες) καταγράφεται. Αν ο καθετήρας αφαιρεθεί για λόγους ανεξάρτητους από τη μόλυνση, τότε η πρώτη παρατήρηση του ατόμου είναι δεξιά λογοκριμένη. Αν εμφανιστεί μόλυνση, ο καθετήρας βγαίνει από τον ασθενή και η μόλυνση θεραπεύεται. Μετά από ένα προαποφασισμένο χρονικό διάστημα (10 εβδομάδες) ο καθετήρας εισάγεται ξανά. Ο δεύτερος χρόνος του ασθενούς ορίζεται από τη στιγμή της δεύτερης εισαγωγής του καθετήρα μέχρι τη δεύτερη μόλυνση ή λογοκρισία. Ο δεύτερος χρόνος μπορεί να λογοκριθεί διότι ο καθετήρας αφαιρέθηκε για λόγους ανεξάρτητους με την πιθανότητα μόλυνσης ή γιατί η μελέτη ολοκληρώθηκε και δεν εμφανίστηκε το γεγονός. Επομένως, κάθε άτομο συνεισφέρει δυο πιθανώς λογοκριμένους χρόνους (Y_1, Y_2) με δείκτριες συναρτήσεις αποτυχίας (Δ_1, Δ_2).

Δεν μπορούμε να γνωρίζουμε αν η λογοκρισία είναι ανεξάρτητη με την πιθανότητα μόλυνσης. Στην πραγματικότητα, η υπόθεση της μη πληροφοριακής λογοκρισίας δεν μπορεί να ελεγχθεί από παρατηρούμενα δεδομένα (Tsiatis 1975). Σίγουρα υπάρχουν και άλλοι τρόποι να οριστεί η χρονική διάσταση των συμβάντων. Στην παρούσα μορφή, θεωρούμε ότι ο χρόνος που πέρασε από τη πρώτη μόλυνση μέχρι τη δεύτερη εισαγωγή του καθετήρα είναι επαρκής ώστε να μπορούμε να υποθέσουμε ότι, για παράδειγμα, 5 μέρες μετά την πρώτη εισαγωγή του καθετήρα είναι πανομοιότυπες (ως προς το συσσωρευμένο κίνδυνο που διατρέχει το άτομο) σε σχέση με 5 μέρες μετά τη δεύτερη εισαγωγή. Στην ουσία, οι 10 εβδομάδες μετά τη θεραπεία της πρώτης μόλυνσης ή τη λογοκρισία, θεωρούμε ότι έχουν μηδενίσει το 'ρολόι κινδύνου' του ασθενούς. Η μελέτη επικεντρώνεται στην συσχέτιση του φύλου, της ηλικίας και του τύπου της νόσου (glomerulo nephritis (GN), acute nephritis (AN), polycystic kidney disease (PKD) και άλλοι τύποι) με το χρόνο μέχρι τη μόλυνση. Τα περιγραφικά χαρακτηριστικά των συμμετεχόντων παρουσιάζονται στον πίνακα (6.1).

Πίνακας 6.1: Χαρακτηριστικά 38 συμμετεχόντων ανά φύλο.

Χαρακτηριστικά	Άνδρες (N=10)	Γυναίκες (N=28)
Ηλικία (έτη)	$43.4 \pm (15.8)$	$43.6 \pm (14.8)$
Τύπος νόσου		
AN (%)	2 (20%)	10 (35.71%)
GN (%)	3 (30%)	6 (21.43%)
Other (%)	3 (30%)	10 (35.71%)
PKD (%)	2 (20%)	2 (7.14%)

Ειναι λογικό να περιμένουμε ότι οι χρόνοι του ίδιου ατόμου είναι συσχετισμένοι μεταξύ τους. Αρχικά, θα επικεντρωθούμε στο φύλο και την ηλικία. Εφαρμόζουμε περιθώριο Cox μοντέλο, θεωρώντας τη συσχέτιση ως ενοχλητική παράμετρο (μοντέλο I του πίνακα 6.2). Ο πίνακας συνδιακύμανσης των εκτιμητών θα τροποποιηθεί ώστε να λάβει υπόψη την ομαδοποίηση των παρατηρήσεων (Lee et al. 1992). Η ηλικία του ασθενούς δεν φαίνεται να συνδέεται με τον κίνδυνο (hazard) μόλυνσης δοθέντος του φύλου και υποθέτοντας γραμμική σχέση μεταξύ του λογαρίθμου του κινδύνου και της ηλικίας. Οι γυναίκες έχουν 56% μικρότερο κίνδυνο μόλυνσης σε σχέση με τους άνδρες της ίδιας ηλικίας, με 95% διάστημα εμπιστοσύνης από -83.1% έως +14.7%. Η περιθώρια (marginal) επίδραση του φύλου δεν είναι στατιστικά σημαντική σε επίπεδο 5% (διορθώνοντας τον πίνακα συνδιακύμανσης των εκτιμήσεων). Παρόμοια αποτελέσματα λαμβάνουμε από ένα κατά τμήματα εκθετικό μοντέλο, διαμερίζοντας το χρονικό άξονα σε 10 σημεία $\{10, 16.5, 26.5, 32, 48, 100, 150, 188, 320\}$ έτσι ώστε να έχουμε περίπου ίσο αριθμό συμβάντων σε κάθε χρονικό διάστημα.

Σε ένα περιθώριο (marginal) μοντέλο η συσχέτιση των παρατηρήσεων δεν επιδρά στην δια-

μόρφωση των σημειακών εκτιμήσεων. Αντίθετα, σε ένα frailty μοντέλο μοντελοποιούμε ρητά τη συσχέτιση των παρατηρήσεων. Τα περιθώρια και τα frailty μοντέλα δεν στοχεύουν στις ίδιες παραμέτρους (εκτός αν $\theta = 0$). Σε ένα frailty μοντέλο, η ερμηνεία των παραμέτρων γίνεται δοθέντος ότι τα υπο σύγκριση άτομα έχουν την ίδια τιμή frailty (ή ανήκουν στην ίδια ομάδα). Γενικά, η περιθώρια και η δεσμευμένη (ως προς τις ομάδες) συσχέτιση δεν ταυτίζονται, ακόμα και αν τα δυο μοντέλα είναι σωστά ορισμένα. Το φαινόμενο αυτό ονομάζεται 'non-collapsibility' (Greenland, Robins & Pearl 1999). Μια συνήθης εξαίρεση είναι η κανονική γραμμική παλινδρόμηση.

Κάθε άτομο αναμένεται να έχει διαφορετικό κίνδυνο μόλυνσης κάτι το οποίο δεν λαμβάνεται υπόψη από τις επεξηγηματικές μεταβλητές. Εφαρμόζουμε Gamma και Lognormal frailty μοντέλα (μοντέλα II και III του πίνακα 6.2), με επεξηγηματικές μεταβλητές την ηλικία (γραμμικά) και το φύλο, με βασική συνάρτηση κινδύνου μια step function στα 10 χρονικά σημεία τα οποία χρησιμοποιήθηκαν στο περιθώριο μοντέλο. Για να βγάλουμε συμπεράσματα για τις παραμέτρους χρησιμοποιούμε MCMC μεθόδους. Εφόσον δεν έχουμε πληροφορία για τις παραμέτρους θα χρησιμοποιήσουμε μη πληροφοριακές εκ των προτέρων κατανομές. Για το $\boldsymbol{\beta}$ θεωρούμε μια $N(\boldsymbol{\mu}_0, \boldsymbol{C}_0)$ εκ των προτέρων κατανομή. Για τις παραμέτρους της βασικής συνάρτησης κινδύνου θεωρουμε $\lambda_j \sim \text{Gamma}(\alpha_{0j}, \lambda_{0j})$, $j = 1, 2, \ldots, J$, και για την ακρίβεια $\omega \sim \text{Gamma}(c, d)$. Θέτουμε τις υπερπαραμέτρους $\boldsymbol{\mu}_0 = \mathbf{0}$, $\boldsymbol{C}_0 = \text{diag}\{10^4\}$, $\alpha_{0j} = 0.001$, $\lambda_{0j} = 0.001$, $c = 0.001$ και $d = 0.001$.

Για τα μοντέλα II και III (πίνακας 6.2) προσομοιώθηκε μια αλυσίδα Markov 4×10^6 επαναλήψεων με 10000 επαναλήψεις να χρησιμοποιούνται ως burn-in. Η πιθανότητα αποδοχής του $\boldsymbol{\beta}$ ήταν κοντά στο 95% και στα δυο μοντέλα (με 2 επαναλήψεις του αλγορίθμου του Gamerman 4.4). Η πιθανότητα αποδοχής των random effects για το Lognormal μοντέλο, ήταν περίπου 95% (δυο επαναλήψεις του αλγορίθμου 4.15). Επομένως, υπάρχει συμφωνία μεταξύ της προτεινουσας και της πραγματικής εκ των υστέρων κατανομής των παραμέτρων. Για να εξουδετερώσουμε την αυτοσυσχέτιση της αλυσίδας, κρατάμε μια επανάληψη στις 40. Μετά από την εκλέπτυνση (thinning) της αλυσίδας δεν υπάρχει σημαντική αυτοσυσχέτιση για καμία παράμετρο. Η σύγκλιση της αλυσίδας φαίνεται ότι έχει επιτευχθεί (διαγράμματα 6.1 έως 6.6). Τα αποτελέσματα συνοψίζονται στον πίνακα 6.2. Τα 95% διαστήματα αξιοπιστίας ορίζονται από το 2.5% έως το 97.5% ποσοστημόριο της περιθώριας κατανομής των παραμέτρων. Συνεπώς, αν η περιθώρια κατανομή είναι ισχυρά ασύμμετρη, τα διαστήματα αξιοπιστίας δεν αντιστοιχούν απαραίτητα με τα 95% διαστήματα υψίστης a-posteriori πυκνότητας. Οι εκτιμήσεις μέγιστης πιθανοφάνειας παρουσιάζονται για σύγκριση.

Τα μοντέλα II και III δείχνουν ότι μόνο το φύλο είναι σημαντικός προγνωστικός παράγοντας της μόλυνσης, με συνέπεια την αφαίρεση του καθετήρα. Το Gamma μοντέλο (II) έχει ελαφρώς καλύτερη εφαρμογή στα δεδομένα σύμφωνα με το κριτήριο DIC σε σχέση με το Lognormal μο-

ντέλο (III). Για παράδειγμα, σύμφωνα με το Gamma μοντέλο (II) οι γυναίκες έχουν 81.1% (εκ των υστέρων διάμεσος του hazard ratio) μικρότερο κίνδυνο μόλυνσης σε σχέση με τους άνδρες ίδιας ηλικίας δεδομένου ότι οι μη παρατηρήσιμοι παράγοντες βρίσκονται στο ίδιο επίπεδο και στα δυο άτομα, με 95% διάστημα αξιοπιστίας από -94.2% έως -48.8%. Οι σημειακές εκτιμήσεις του $\boldsymbol{\beta}$, οι οποίες προέρχονται από την εκ των υστέρων κατανομή των παραμέτρων, δεν διαφέρουν σημαντικά με τις εκτιμήσεις μέγιστης πιθανοφάνειας. Το πλεονέκτημα της Μπεϋζιανής στατιστικής είναι ότι η εκ των υστέρων κατανομή των παραμέτρων μπορεί να έχει οποιαδήποτε μορφή. Για παράδειγμα, η εκ των υστέρων κατανομή του φύλου είναι αρνητικά ασύμμετρη και στα δυο μοντέλα (II και III). Τα άνω άκρα των διαστήματων εμπιστοσύνης και αξιοπιστίας συμπίπτουν. Με αυτό τον τρόπο εξηγείται το γεγονός ότι, η εκ των υστέρων τυπική απόκλιση του $\boldsymbol{\beta}$ είναι μεγαλύτερη από το τυπικό σφάλμα των εκτιμήσεων μέγιστης πιθανοφάνειας.

Η διακύμανση των frailties, θ, μπορεί να ερμηνευτεί ως ένα μέτρο ετερογένειας των κινδύνων των ατόμων. Παρατηρούμε ότι, η εκ των υστέρων μέση τιμή και διάμεσος της διακύμανσης θ (Gamma μοντέλο II) και της τυπικής απόκλισης σ (Lognormal μοντέλο III), είναι ελαφρώς μεγαλύτερες από τους εκτιμητές μέγιστης πιθανοφάνειας. Τα άνω άκρα των διαστημάτων αξιοπιστίας και εμπιστοσύνης συμφωνούν. Η διαφορά εντοπίζεται στο γεγονός ότι οι εκ των υστέρων κατανομές των $(\sqrt{\theta}, \sigma)$ παρουσιάζουν, εκτός από την κύρια κορυφή της κατανομής, και μια δεύτερη, η οποία είναι πολύ κοντά στο 0, δηλώνοντας έτσι ότι υπάρχει θετική πιθανότητα μηδενικής ετερογένειας μεταξύ των ασθενών (γραφήματα 6.3 και 6.6). Βέβαια, οι πιο πιθανές τιμές του $\sqrt{\theta}$ και σ βρίσκονται αρκετά μακρυά από το μηδέν, άρα, με βάση το μοντέλο το οποίο περιέχει το φύλο και την ηλικία (μοντέλα II και III) υπάρχουν ενδείξεις συσχέτισης των χρόνων του ίδιου ασθενούς.

Θα προσπαθήσουμε να βελτιώσουμε την εφαρμογή του μοντέλου. Είδαμε ότι δεν υπάρχει σημαντική γραμμική σχέση μεταξύ του λογαρίθμου του κινδύνου και της ηλικίας, αφού σταθμίσουμε ως προς τους υπόλοιπους παράγοντες. Όμως, η γραμμική σχέση μπορεί να μην υποστηρίζεται ικανοποιητικά από τα δεδομένα. Όροι ανώτερης τάξης ή λογαριθμική σχέση δεν βελτίωσαν την εφαρμογή του μοντέλου. Θα χωρίσουμε την ηλικία σε δυο ομάδες (1 ομάδα: ηλικία $\geq$48, 2 ομάδα: ηλικια $<$48). Η τιμή 48 επιλέχθηκε ώστε να έχει το μοντέλο την καλύτερη εφαρμογή και να υπάρχει ισορροπία ανάμεσα στις ομάδες. Επίσης, θα προσθέσουμε τον τύπο της νόσου επειδή θεωρούμε ότι είναι ένας παράγοντας που αξίζει να μπει στο μοντέλο, ανεξάρτητα από τη σημαντικότητα του.

Παρατηρούμε ότι αν βάλουμε τον τύπο της νόσου και την ηλικία (ως κατηγορίες) στο μοντέλο, η εκτίμηση μέγιστης πιθανοφάνειας για τη διακύμανση των frailties (Gamma μοντέλο IV) και τη διακύμανση του λογαρίθμου των frailties (Lognormal μοντέλο V) είναι πολύ κοντά στο 0. Για τα μοντέλα IV και V προσομοιώθηκε μια αλυσίδα Markov 3.2×10^6 επαναλήψεων όπου 5000 επαναλήψεις χρησιμοποιήθηκαν ως burn-in. Επειδή στα μοντέλα IV και V προσομοιώνονται

μικρές τιμές του θ και σ^2, συναντάμε προβλήματα σύγκλισης για αυτές τις παραμέτρους. Γι αυτό θα κρατήσουμε 1 τιμή στις 80. Επειδή για τις υπόλοιπες παραμέτρους η σύγκλιση στη στάσιμη κατανομή είναι σταθερή, δεν θα παρασουσιαστούν γραφήματα. Ξεκάθαρα, από τα γραφήματα 6.7 και 6.8 βλέπουμε ότι οι πιο πιθανές τιμές του σ (μοντέλο V) και $\sqrt{\theta}$ (μοντέλο IV) είναι πολύ κοντά στο μηδέν. Επομένως, είναι λογικό ότι η εκτίμηση μέγιστης πιθανοφάνειας είναι σχεδόν μηδέν. Για σύγκριση, εφαρμόζουμε ένα μοντέλο το οποίο δεν περιλαμβάνει frailties (μοντέλο VI). Οι εκτιμήσεις μέγιστης πιθανοφάνειας συμφωνούν με την εκ των υστέρων μέση τιμή και διάμεσο των παραμέτρων. Το DIC κριτήριο μειώθηκε κατά 3 μονάδες, το οποίο δείχνει ότι τα frailties δεν συνεισφέρουν ουσιαστικά στο μοντέλο.

Επομένως, τα μοντέλα τα οποία περιέχουν το φύλο και την ηλικία μας δείχνουν ότι υπάρχει ετερογένεια μεταξύ των ατόμων, ενώ, αν προσθέσουμε τον τύπο της νόσου, δεν υπάρχει. Στο γράφημα 6.9 βλέπουμε τα martingale κατάλοιπα του περιθώριου Cox μοντέλου I. Παρατηρούμε την ύπαρξη ενός ισχυρού outlier. Η παρατήρηση ανήκει σε έναν άνδρα με id=21, ηλικίας 46.5 ετών με τύπο νόσου PKD και δυο μολύνσεις σε 562 και 152 μέρες, αντίστοιχα. Από τις 20 παρατηρήσεις ανδρών υπάρχουν 18 αποτυχίες, με διάμεσο χρόνο μέχρι τη μόλυνση 22 μέρες. Αν το συγκεκριμένο άτομο εξαιρεθεί από τα μοντέλα II έως V η επίδραση του τύπου της νόσου, η επίδραση της ηλικίας και η ετερογένεια, μειώνονται σημαντικά. Όταν το άτομο είναι στην ανάλυση δημιουργείται ετερογένεια μεταξύ των ατόμων στα μοντέλα II και III, ενώ στα μοντέλα IV και V δημιουργούνται ενδείξεις ότι ο τύπος της νόσου και η ηλικία είναι σημαντικοί παράγοντες. Επιπλέον, αν εξετάσουμε την αλληλεπίδραση του φύλου με τον τύπου της νόσου, θα δούμε ότι οι γυναίκες έχουν σαφώς μικρότερο κίνδυνο σε σχέση με τους άνδρες στα άτομα με κατηγορίες νόσου AG, GN και άλλων τύπων. Στην κατηγορία PKD η κατάσταση αντιστρέφεται, οι άνδρες φαίνεται να έχουν μικρότερο κίνδυνο σε σχέση με τις γυναίκες. Το φαινόμενο αυτό είναι αποτέλεσμα της έκτροπης παρατήρησης (id=21) και του μικρού αριθμού ατόμων στην κατηγορία PKD. Αν το άτομο id=21 εξαιρεθεί από την ανάλυση δεν υπάρχουν ενδείξεις αλληλεπίδρασης. Οι αποτυχίες είναι σαφώς διατεταγμένες. Εφαρμόστηκε το μοντέλο VI, στρωματοποιημένο (Cox) ως προς τη σειρά των αποτύχιων. Τα προηγούμενα αποτελέσματα συνεχίζουν να ισχύουν. Δεν βρέθηκε σημαντική αλληλεπίδραση μεταξύ των παραγόντων και της σειράς των αποτυχιών, βάσει ενός global LR test.

Το συμπέρασμα είναι ότι, μόνο το φύλο φαίνεται να έχει επίδραση στον κίνδυνο μόλυνσης, δοθέντων της ηλικίας και του τύπου της νόσου, ανεξάρτητα από την παρατήρηση id=21 και την επιλογή της κατανομής των frailties. Η επίδραση της ηλικίας είναι αρκετά ευαίσθητη στην επιλογή του ορίου που επιλέγεται για να διχοτομηθεί η μεταβλητή και στην παρατήρηση id=21. Η επίδραση του τύπου της νόσου οφείλεται κυρίως στο άτομο id=21. Γενικά, δεν παρατηρήσαμε σημαντικές διαφορές ανάμεσα στα frailty μοντέλα. Οι Therneau & Grambsch (2000) εφάρμοσαν τα μοντέλα IV και V (ηλικία με γραμμικό όρο) με penalized partial likelihood μεθόδους.

Παρατήρησαν διαφορές των δυο μοντέλων ως προς τη διακύμανση των frailties. Μεταξύ των μοντέλων II και III το κριτήριο DIC επιλέγει το Gamma μοντέλο, χωρίς όμως μεγάλη διαφορά. Στα μοντέλα IV και V οι πιο πιθανές τιμές του θ και σ^2 βρίσκονται πολύ κοντά στο 0, άρα, η περιθώρια πιθανοφάνεια (ως προς τα frailties) των δυο μοντέλων είναι ουσιαστικά η ίδια.

Πίνακας 6.2: Αποτελέσματα frailty και σταθερών επιδράσεων μοντέλων για τους 38 ασθενείς οι οποίοι χρησιμοποιούν φορητή συσκευή αιμοκάθαρσης.

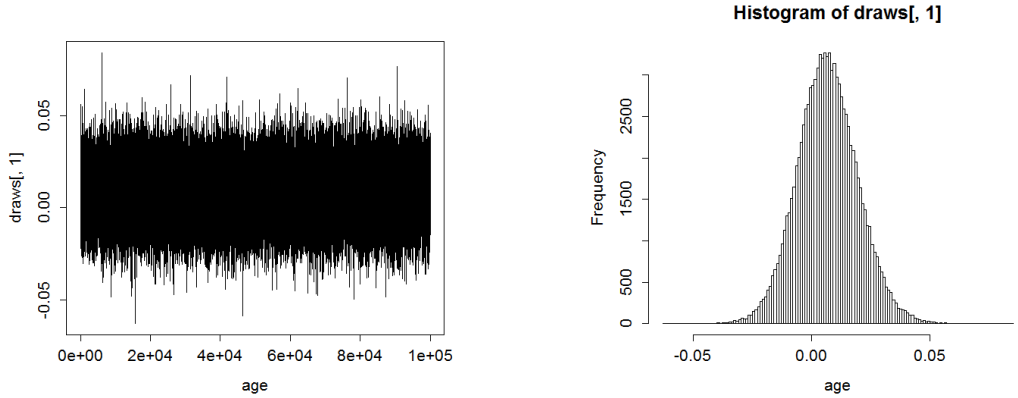
Μοντέλο I	Περιθώριο PE μοντέλο (MLE)			Περιθώριο Cox μοντέλο	
	Coef. (SE)		95%CI	Coef. (SE)	95%CI
Ηλικία	0.003 (0.009)		(-0.014,0.020)	0.002 (0.009)	(-0.016,0.020)
Φύλο	-0.886 (0.472)		(1.809,0.039)	-0.821 (0.489)	(-1.779,0.137)
Μοντέλο II	PE-Gamma (MCMC) DIC=680.062			PE-Gamma (MLE)	
	Μέση τιμή (SD)	Διάμεσος	95%PI	Coef. (SE)	95%CI
Ηλικία	0.007 (0.013)	0.006	(-0.018,0.032)	0.006 (0.012)	(-0.017,0.029)
Φύλο	-1.692 (0.556)	-1.667	(-2.851,-0.669)	-1.651 (0.503)	(-2.636,-0.667)
θ	0.535 (0.322)	0.493	(0.022,1.286)	0.419 (0.246)	(0.133,1.322)
Μοντέλο III	PE-Lognormal (MCMC) DIC=681.549			PE-Lognormal (MLE)	
	Μέση τιμή (SD)	Διάμεσος	95%PI	Coef. (SE)	95%CI
Ηλικία	0.006 (0.013)	0.005	(-0.019,0.032)	0.005 (0.012)	(-0.018,0.028)
Φύλο	-1.463 (0.517)	-1.431	(-2.561,-0.531)	-1.427 (0.460)	(-2.329,-0.525)
σ^2	0.632 (0.504)	0.528	(0.007,1.891)	0.426 (0.321)	(0.097,1.866)
Μοντέλο IV	PE-Gamma (MCMC) DIC= 677.876			PE-Gamma (MLE)	
	Μέση τιμή (SD)	Διάμεσος	95%PI	Coef. (SE)	95%CI
Ηλικία≥48	0.741 (0.445)	0.721	(-0.076,1.681)	0.606 (0.339)	(-0.058,1.270)
Φύλο	-1.833 (0.481)	-1.805	(-2.868,-0.956)	-1.634 (0.365)	(-2.349,-0.918)
AN	0.243 (0.480)	0.240	(-0.714,1.200)	0.257 (0.367)	(-0.462,0.975)
GN	-0.231 (0.560)	-0.233	(-1.343,0.894)	-0.265 (0.446)	(-1.139,0.611)
PKD	-1.314 (0.757)	-1.326	(-2.781,0.239)	-1.368 (0.560)	(-2.466,-0.269)
θ	0.295 (0.312)	0.199	(0.001,1.084)	2e-06 (0.003)	-
Μοντέλο V	PE-Lognormal (MCMC) DIC=677.778			PE-Lognormal (MLE)	
	Μέση τιμή (SD)	Διάμεσος	95%PI	Coef. (SE)	95%CI
Ηλικία≥48	0.750 (0.450)	0.729	(-0.071,1.716)	0.606 (0.339)	(-0.058,1.270)
Φύλο	-1.822 (0.480)	-1.798	(-2.839,-0.946)	-1.634 (0.365)	(-2.349,-0.918)
AN	0.246 (0.477)	0.248	(-0.708,1.183)	0.257 (0.367)	(-0.462,0.975)
GN	-0.276 (0.564)	-0.273	(-1.411,0.823)	-0.265 (0.446)	(-1.139,0.611)
PKD	-1.406 (0.721)	-1.391	(-2.862,-0.011)	-1.368 (0.560)	(-2.466,-0.269)
σ^2	0.369 (0.437)	0.219	(0.001,1.536)	1e-05 (0.006)	-
Μοντέλο VI	PE (MCMC) DIC=674.707			PE (MLE)	
	Μέση τιμή (SD)	Διάμεσος	95%PI	Coef. (SE)	95%CI
Ηλικία≥48	0.614 (0.341)	0.611	(-0.049,1.292)	0.606 (0.339)	(-0.058,1.270)
Φύλο	-1.637 (0.368)	-1.636	(-2.353,-0.914)	-1.634 (0.365)	(-2.349,-0.918)
AN	0.252 (0.367)	0.255	(-0.474,0.973)	0.257 (0.367)	(-0.462,0.975)
GN	-0.289 (0.447)	-0.283	(-1.176,0.578)	-0.265 (0.446)	(-1.139,0.611)
PKD	-1.433 (0.571)	-1.419	(-2.597,-0.353)	-1.368 (0.560)	(-2.466,-0.269)

Κατηγορία αναφοράς για το φύλο είναι οι άνδρες.

Κατηγορία αναφοράς για τον τύπο της νόσου είναι οι άλλοι τύποι.

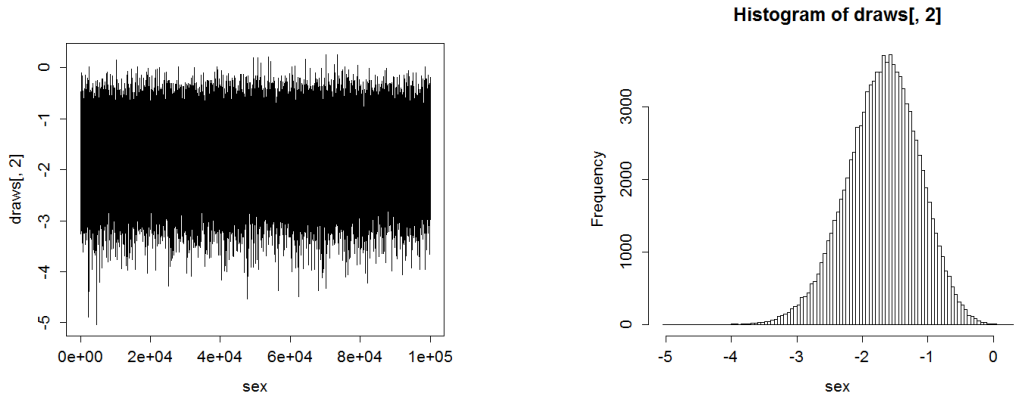
Τα τυπικά σφάλματα στο μοντέλο I έχουν σταθμιστεί ως προς τη συσχέτιση των παρατηρήσεων.

PE κατά τμήματα εκθετικό μοντέλο.



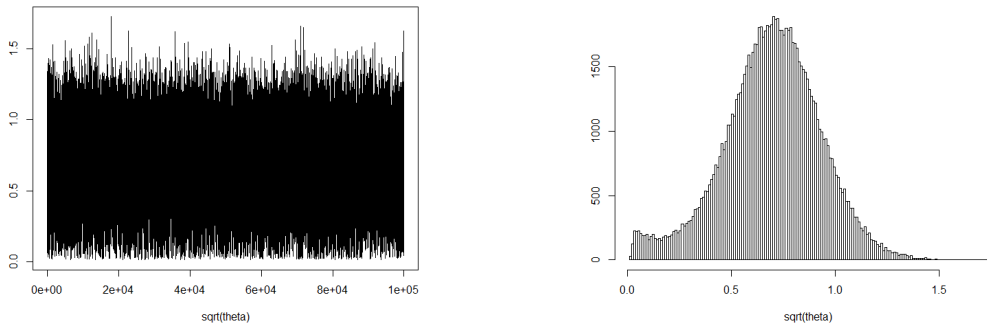
(α′) Trace plot (β′) Histogram

Σχήμα 6.1: Σημειόγραμμα και ιστόγραμμα της παραμέτρου της ηλικίας για το Gamma frailty μοντέλο το οποίο περιλαμβάνει το φύλο και την ηλικία ως επεξηγηματικές μεταβλητές.



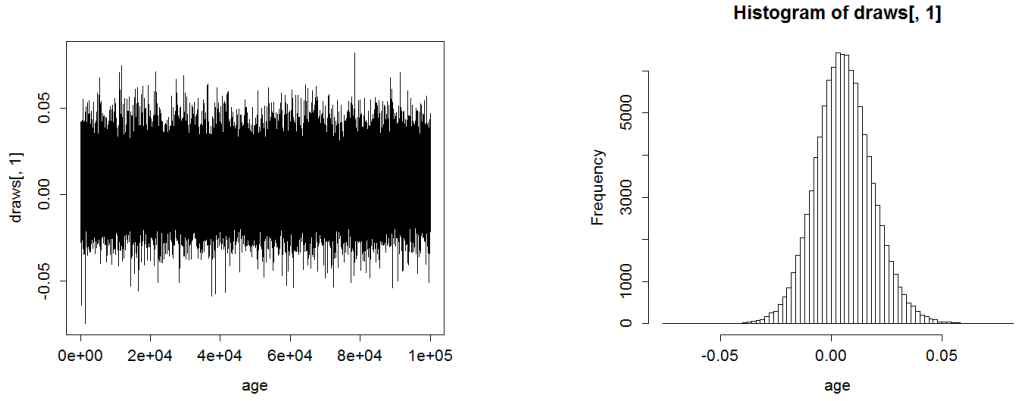
(α′) Trace plot (β′) Histogram

Σχήμα 6.2: Σημειόγραμμα και ιστόγραμμα της παραμέτρου του φύλου για το Gamma frailty μοντέλο το οποίο περιλαμβάνει το φύλο και την ηλικία ως επεξηγηματικές μεταβλητές.



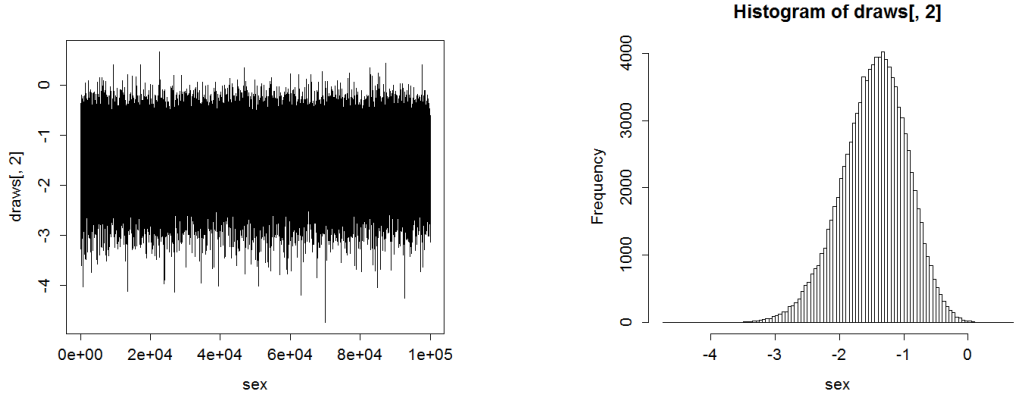
(α′) Trace plot (β′) Histogram

Σχήμα 6.3: Σημειόγραμμα και ιστόγραμμα της τυπικής απόκλισης των frailties, $\sqrt{\theta}$, για το Gamma frailty μοντέλο το οποίο περιλαμβάνει το φύλο και την ηλικία ως επεξηγηματικές μεταβλητές.



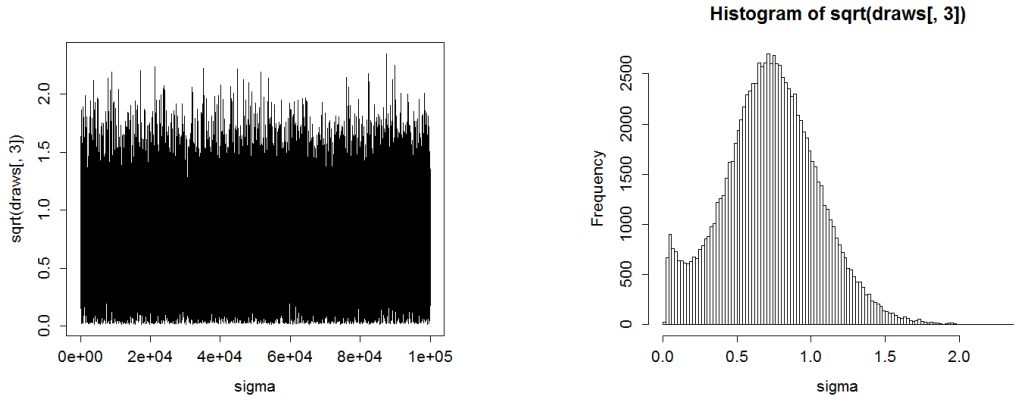
(α′) Trace plot (β′) Histogram

Σχήμα 6.4: Σημειόγραμμα και ιστόγραμμα της παραμέτρου της ηλικίας για το Lognormal frailty μοντέλο το οποίο περιλαμβάνει το φύλο και την ηλικία ως επεξηγηματικές μεταβλητές.



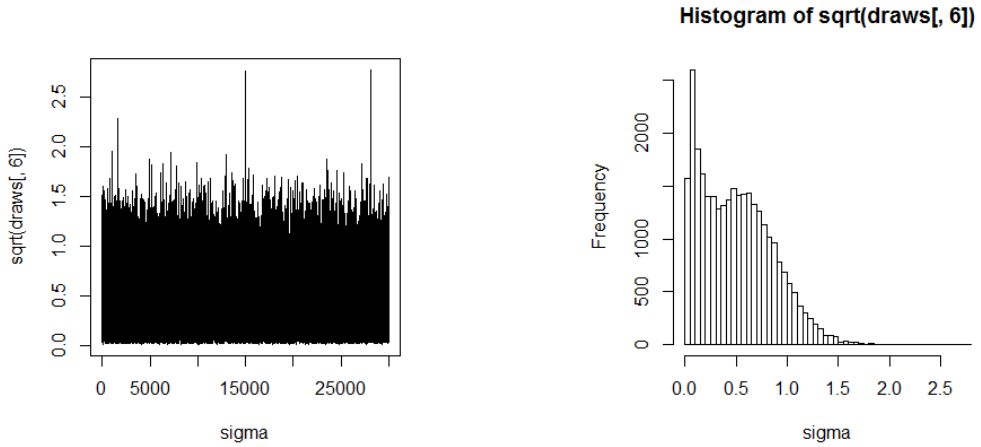
(α′) Trace plot (β′) Histogram

Σχήμα 6.5: Σημειόγραμμα και ιστόγραμμα της παραμέτρου του φύλου για το Lognormal frailty μοντέλο το οποίο περιλαμβάνει το φύλο και την ηλικία ως επεξηγηματικές μεταβλητές.



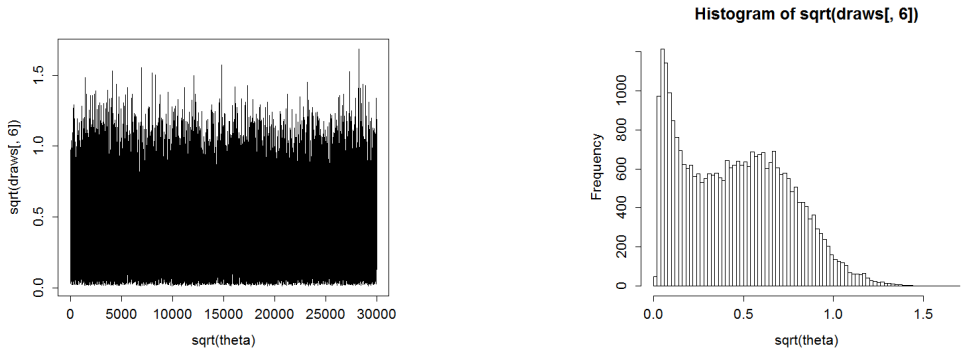
(α′) Trace plot (β′) Histogram

Σχήμα 6.6: Σημειόγραμμα και ιστόγραμμα της τυπικής απόκλισης σ του λογαρίθμου των frailties για το Lognormal frailty μοντέλο το οποίο περιλαμβάνει το φύλο και την ηλικία.



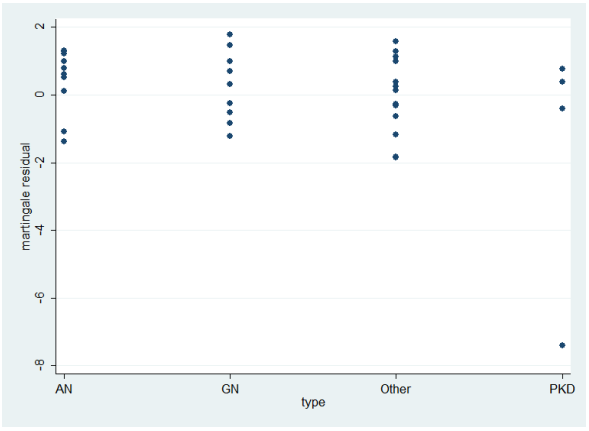
(α′) Trace plot (β′) Histogram

Σχήμα 6.7: Σημειόγραμμα και ιστόγραμμα της τυπικής απόκλισης του λογαρίθμου των frailties, σ, για το Lognormal frailty μοντέλο το οποίο περιλαμβάνει το φύλο, την ηλικία και τον τύπο της νόσου.



(α′) Trace plot (β′) Histogram

Σχήμα 6.8: Σημειόγραμμα και ιστόγραμμα της τυπικής απόκλισης των frailties, $\sqrt{\theta}$, για το Gamma frailty μοντέλο το οποίο περιλαμβάνει το φύλο, την ηλικία και τον τύπο της νόσου.



Σχήμα 6.9: Martingale κατάλοιπα για το μοντέλο I (COX) των 38 ασθενών. Για κάθε ασθενή παρουσιάζεται το άθροισμα των 2 καταλοίπων.

Κεφάλαιο 7

Συζήτηση

Στην παρούσα διπλωματική εργασία εξετάσαμε μοντέλα τυχαίων επιδράσεων (shared frailty models) στην ανάλυση επιβίωσης, υπό το πρίσμα της Μπεϋζιανής στατιστικής. Η στατιστική συμπερασματολογία επετεύχθη μέσω μεθόδων Markov Chain Monte Carlo (MCMC). Η βασική συνάρτηση κινδύνου μοντελοποιήθηκε σύμφωνα με την Weibull και την κατά τμήματα εκθετική κατανομή. Ως κατανομές για τα frailties χρησιμοποιήθηκαν η Gamma και η Lognomal κατανομή.

Η εφαρμογή των μεθόδων αξιολογήθηκε μέσω προσομοιωμένων δεδομένων. Παρατηρήσαμε ότι η εκτίμηση της παραμέτρου η οποία σχετίζεται με την επεξηγηματική μεταβλητή εκτιμάται το ίδιο καλά, ανεξαρτήτως αν η frailty κατανομή που επιλέξαμε είναι η σωστή. Επίσης, το μοντέλο σταθερού κατά τμήματα κινδύνου είχε, κατά μέσο όρο, τις ίδιες εκτιμήσεις με το (σωστό) μοντέλο Weibull, αν και παρατηρήθηκε ελαφρώς μεγαλύτερη διασπορά στις εκτιμήσεις, ίσως λόγω των πολλών παραμέτρων που εκτιμούνται. Επιπλέον, παρατηρήσαμε ότι η πιθανότητα κάλυψης των διαστημάτων αξιοπιστίας ήταν αρκτετά κοντά στο θεωρητικό όριο. Η σύγκριση των μοντέλων έγινε με το κριτήριο DIC. Εφόσον δεν ενδιαφερόμαστε να εκτιμήσουμε τα frailties, υπολογίσαμε το περιθώριο DIC το οποίο στηρίζεται στην περιθώρια (ως προς τα frailties) πιθανοφάνεια. Είδαμε ότι το κριτήριο DIC έβρισκε τη σωστή frailty κατανομή σε σταθερά μεγάλο ποσοστό, ανεξαρτήτως της βασικής κατανομής που επιλέξαμε.

Περίληψη

Στην ανάλυση δεδομένων επιβίωσης, η ετερογένεια μπορεί να είναι παρούσα με διάφορες μορφές. Σε πολλές περιπτώσεις, δεν μπορούμε να υποθέσουμε ότι οι χρόνοι μέχρι την εμφάνιση ενός γεγονότος είναι ανεξάρτητοι σε κάποιες ομάδες του πληθυσμού, αφού τα άτομα της ίδιας ομάδας μπορεί να μοιράζονται κοινά χαρακτηριστικά τα οποία δεν είναι παρατηρήσιμα. Επομένως, μπορούμε να χρησιμοποιήσουμε μοντέλα τυχαίων επιδράσεων (shared frailty models) για να μοντελοποιήσουμε ρητά τη συσχέτιση των ατόμων της ίδιας ομάδας.

Στην παρούσα διπλωματική εργασία, εξετάσαμε μια επέκταση των μοντέλων αναλογικών κινδύνων, υπό την οπτική της Μπεϋζιανής στατιστικής, στην οποία η Weibull και η κατά τμήματα εκθετική κατανομή χρησιμοποιήθηκαν ως κατανομές για τη βασική συνάρτηση κινδύνου. Ως κατανομές για τους τυχαίους όρους χρησιμοποιήθηκαν οι κατανομές Gamma και Lognormal. Η στατιστική συμπερασματολογία εξάχθηκε μέσω μεθόδων Markov Chain Monte Carlo (MCMC). Η εφαρμογή των μεθόδων αξιολογήθηκε μέσω προσομοιωμένων δεδομένων και τα μοντέλα συγκρίθηκαν με το κρίτηριο DIC (Deviance Information Criterion). Επίσης, οι αλγόριθμοι MCMC εφαρμόστηκαν σε πραγματικά δεδομένα νεφροπαθών ασθενών.

Abstract

In the analysis of survival data, heterogeneity may be present in many situations. In many circumstances, we cannot assume that failure times for subjects of the same cluster are independent, since subjects of the same cluster share unobserved characteristics. Therefore, we could use shared frailty models to explicitly model the association of all members in the same cluster.

This thesis considered an extension of proportional hazard models, from a Bayesian perspective, in which the Weibull and Piecewise exponential distributions were used as the distributions for the baseline hazard function. We also used the Gamma and Lognormal distributions as frailty distributions. Statistical inference was based on Markov Chain Monte Carlo (MCMC) methods. The performance of the methods was evaluated through simulation studies and the Deviance Information Criterion (DIC) was used to compare the fit of the models. The developed methods were also fitted to a real dataset of kidney patients.

Βιβλιογραφία

Abdulkarimova, U. (2013), 'Frailty models for modelling heterogeneity'.

Abramowitz, M., Stegun, I. A. et al. (1965), *Handbook of mathematical functions*, Vol. 1046, Dover New York.

Balakrishnan, N. & Peng, Y. (2006), 'Generalized gamma frailty model', *Statistics in medicine* **25**(16), 2797–2816.

Beard, R. E. (1959), Appendix: Note on some mathematical mortality models, *in* 'Ciba Foundation Symposium-The Lifespan of Animals (Colloquia on Ageing), Volume 5', Wiley Online Library, pp. 302–311.

Blocker, A. W. (2011), *fastGHQuad: Fast Rcpp implementation of Gauss-Hermite quadrature.* R package version 0.1-1.
URL: *http://CRAN.R-project.org/package=fastGHQuad*

Browne, W. J., Draper, D. et al. (2006), 'A comparison of bayesian and likelihood-based methods for fitting multilevel models', *Bayesian Analysis* **1**(3), 473–514.

Cox, D. R. (1972), 'Regression models and life-tables', *Journal of the Royal Statistical Society. Series B (Methodological)* **34**(2), pp. 187–220.

Duchateau, L. & Janssen, P. (2007), *The frailty model*, Springer.

Fleming, T. & Harrington, D. (1991), 'Counting processes and survival analysiswiley', *New York* .

Fraser, D. et al. (2011), 'Is bayes posterior just quick and dirty confidence?', *Statistical Science* **26**(3), 299–316.

Gamerman, D. (1997), 'Sampling from the posterior distribution in generalized linear mixed models', *Statistics and Computing* **7**(1), 57–68.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2003), 'Bayesian data analysis, (chapman & hall/crc texts in statistical science)'.

Gelman, A., Robert, C., Chopin, N. & Rousseau, J. (1995), 'Bayesian data analysis'.

Gelman, A. et al. (2006), 'Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper)', *Bayesian analysis* **1**(3), 515–534.

Geman, S. & Geman, D. (1984), 'Stochastic relaxation, gibbs distributions, and the bayesian restoration of images', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* (6), 721–741.

Gilks, W. R. & Wild, P. (1992), 'Adaptive rejection sampling for gibbs sampling', *Applied Statistics* pp. 337–348.

Gill, R. D. (1984), 'Understanding cox's regression model: a martingale approach', *Journal of the American Statistical Association* **79**(386), 441–447.

Glidden, D. V. & Vittinghoff, E. (2004), 'Modelling clustered survival data from multicentre clinical trials', *Statistics in medicine* **23**(3), 369–388.

Greenland, S. (1996), 'Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference', *Epidemiology* pp. 498–501.

Greenland, S., Robins, J. M. & Pearl, J. (1999), 'Confounding and collapsibility in causal inference', *Statistical Science* pp. 29–46.

Gutierrez, R. G. (2002), 'Parametric frailty and shared frailty survival models', *Stata Journal* **2**(1), 22–44.

Hastings, W. K. (1970), 'Monte carlo sampling methods using markov chains and their applications', *Biometrika* **57**(1), 97–109.

Hougaard, P. (2000), *Analysis of Multivariate Survival Data*, Statistics for Biology and Health, Springer New York.

Ibrahim, J. G., Chen, M.-H. & Sinha, D. (2001), *Bayesian Survival Analysis*, Springer.

Kalbfleisch, J. & Prentice, R. (2002), *The Statistical Analysis of Failure Time Data*, Wiley Series in Probability and Statistics, Wiley.

Kaplan, E. L. & Meier, P. (1958), 'Nonparametric estimation from incomplete observations', *Journal of the American statistical association* **53**(282), 457–481.

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R. & Jones, D. R. (2005), 'How vague is vague? a simulation study of the impact of the use of vague prior distributions in mcmc using winbugs', *Statistics in medicine* **24**(15), 2401–2428.

Lancaster, T. (1979), 'Econometric methods for the duration of unemployment', *Econometrica: Journal of the Econometric Society* pp. 939–956.

Lee, E. W., Wei, L., Amato, D. A. & Leurgans, S. (1992), Cox-type regression analysis for large numbers of small groups of correlated failure time observations, *in* 'Survival analysis: state of the art', Springer, pp. 237–247.

Marchand, E. & Strawderman, W. E. (2012), 'On bayesian credible sets in restricted parameter space problems and lower bounds for frequentist coverage', *arXiv preprint arXiv:1208.0028* .

McCullagh, P. N. & Nelder, F. (1989), 'Ja (1989) generalized linear models', *Monographs on Statistics and Applied Probability* **37**.

McGilchrist, C. & Aisbett, C. (1991), 'Regression with frailty in survival analysis', *Biometrics* pp. 461–466.

Murthy, D. P., Xie, M. & Jiang, R. (2004), *Weibull models*, Vol. 505, John Wiley & Sons.

original C++ code from Arnost Komarek based on ars.f written by P. Wild, P. P. R. & Gilks, W. R. (2009), *ars: Adaptive Rejection Sampling.* R package version 0.4.
URL: *http://CRAN.R-project.org/package=ars*

Rondeau, V., Mazroui, Y. & Gonzalez, J. R. (2012), 'frailtypack: an r package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation', *Journal of Statistical Software* **47**(4), 1–28.

Sahu, S. K., Dey, D. K., Aslanidou, H. & Sinha, D. (1997), 'A weibull regression model with gamma frailties for multivariate survival data', *Lifetime data analysis* **3**(2), 123–137.

Sjölander, A. & Greenland, S. (2013), 'Ignoring the matching variables in cohort studies–when is it valid and why?', *Statistics in medicine* **32**(27), 4696–4708.

Sjölander, A., Lichtenstein, P., Larsson, H. & Pawitan, Y. (2013), 'Between–within models for survival analysis', *Statistics in medicine* **32**(18), 3067–3076.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van Der Linde, A. (2002), 'Bayesian measures of model complexity and fit', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(4), 583–639.

Therneau, T. & Grambsch, P. (2000), 'Modeling survival data: extending the cox model', *Statistics* .

Tsiatis, A. (1975), 'A nonidentifiability aspect of the problem of competing risks', *Proceedings of the National Academy of Sciences* **72**(1), 20–22.

Vaupel, J. W., Manton, K. G. & Stallard, E. (1979), 'The impact of heterogeneity in individual frailty on the dynamics of mortality', *Demography* **16**(3), 439–454.

Wienke, A. (2010), *Frailty models in survival analysis*, CRC Press.

Zetterqvist, J. (n.d.), 'Proportional hazards model for matched failure time data'.