ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Εφαρμογές Παραμετρικών και Ημι-παραμετρικών Μοντέλων και Μεθόδων Συρρίκνωσης σε Δεδομένα Διάρκειας Ζωής

Βέργου Αικατερίνη

Επιβλέπουσα Καθηγήτρια: Καρώνη Χρυσηΐς,

Επιτροπή Καθηγητών:

Χ. Καρώνη	**Χ. Κουκουβίνος**	**Β. Παπανικολάου**
Καθηγήτρια, ΕΜΠ,	**Καθηγητής, ΕΜΠ,**	**Καθηγητής, ΕΜΠ**

ΑΘΗΝΑ, Ιούλιος 2021

Ευχαριστίες

Με την ολοκλήρωση της διπλωματικής μου εργασίας, θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες σε όλους όσους συνέβαλλαν στην εκπόνησή της.

Αρχικά, θα ήθελα να ευχαριστήσω την καθηγήτρια του Ε.Μ.Π, κα Χρυσηίδα Καρώνη για την ανάθεση, την επίβλεψη και τη διαρκή καθοδήγησή της σε όλα τα στάδια εκπόνησης της παρούσας εργασίας.

Επιπλέον, θα ήθελα να ευχαριστήσω την οικογένεια μου και όλους όσους με στήριξαν καθ' όλη τη διάρκεια συγγραφής της παρούσας εργασίας αλλά και γενικότερα των σπουδών μου.

Περίληψη

Η παρούσα διπλωματική εργασία ασχολείται με τη στατιστική ανάλυση δεδομένων διάρκειας ζωής με χρήση διάφορων μεθόδων και μοντέλων με σκοπό την επιστημονική μελέτη των δεδομένων μέχρις ότου συμβεί κάποιο γεγονός. Ο κλάδος της Στατιστικής με τον οποίο θα ασχοληθούμε είναι η Ανάλυση Επιβίωσης η οποία έχει πολλές εφαρμογές σε βιοϊατρικές αλλά και τεχνολογικές επιστήμες.

Πιο αναλυτικά, στο πρώτο κεφάλαιο περιγράφονται κάποιες βασικές έννοιες και ορισμούς της Ανάλυσης της Επιβίωσης, όπως είναι για παράδειγμα τα αποκομμένα δεδομένα, η συνάρτηση επιβίωσης, η συνάρτηση διακινδύνευσης, η σωρευτική συνάρτηση διακινδύνευσης κ.τ.λ. Επίσης, παρουσιάζονται κάποιες βασικές κατανομές της Στατιστικής Ανάλυσης, για παράδειγμα η κατανομή Weibull, Log-Normal, Log-Logistic κ.τ.λ. Τέλος, περιγράφονται κάποιες μη-παραμετρικές τεχνικές που είναι πολύ χρήσιμες στην Ανάλυση Επιβίωσης (εκτιμήτρια Kaplan-Meier, εκτιμήτρια Nelson-Aalen κ.τ.λ.).

Στο δεύτερο κεφάλαιο παρουσιάζονται αναλυτικά τα Παραμετρικά Μοντέλα Παλινδρόμησης, τα οποία χρησιμοποιούνται πολύ συχνά σε δεδομένα διάρκειας ζωής. Στο πρώτο μέρος του κεφαλαίου γίνεται παρουσίαση του μοντέλου Επιταχυνόμενης Διακοπής (AL) καθώς και του μοντέλου Αναλογικής Διακινδύνευσης (PH). Ενώ στο δεύτερο μέρος γίνεται εκτενής παρουσίαση του ημι-παραμετρικού μοντέλου του Cox (εκτίμηση παραμέτρων, γραφικός έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης, υπόλοιπα Schoenfeld, καμπύλες ROC & AUC κ.τ.λ.). Τέλος, παρουσιάζονται διάφορα κριτήρια επιλογής μεταβλητών και μέτρα καλής προσαρμογής για τα παραπάνω μοντέλα.

Στο τρίτο κεφάλαιο παρουσιάζονται οι μέθοδοι Συρρίκνωσης που χρησιμοποιούνται για την αντιμετώπιση των προβλημάτων της πολυσυγγραμμικότητας. Πιο συγκεκριμένα γίνεται παρουσίαση των μεθόδων Ridge και Lasso, οι οποίες έχουν αναπτυχθεί πολύ τα τελευταία χρόνια. Το κεφάλαιο ολοκληρώνεται με την παρουσίαση της μεθόδου Cross-Validation για την επιλογή του βέλτιστου συντελεστή λ που χρησιμοποιείται σε όλες τις μεθόδους Συρρίκνωσης.

Στο τέταρτο κεφάλαιο μελετάται ένα απλό σύνολο δεδομένων διάρκειας ζωής από λαμπτήρες φθορισμού. Στην αρχή, γίνεται μια πρώτη μη-παραμετρική ανάλυση των δεδομένων με τις τεχνικές που παρουσιάστηκαν στα προηγούμενα κεφάλαια. Ενώ στη συνέχεια προσαρμόζεται ένα μοντέλο Επιταχυνόμενης Διακοπής (AL) και γίνεται γραφικός έλεγχος μέσω των υπολοίπωνCox-Snell.

Στο πέμπτο και τελευταίο κεφάλαιο μελετάται ένα σύνολο δεδομένων διάρκειας ζωής από ασθενείς που έχουν υποβληθεί σε μεταμόσχευση μυελού των οστών. Αρχικά, γίνεται μια πρώτη μη-παραμετρική ανάλυση των δεδομένων, ενώ στη συνέχεια προσαρμόζεται ένα μοντέλο Επιταχυνόμενης Διακοπής (AL) και γίνεται γραφικός έλεγχος μέσω των υπολοίπωνCox-Snell. Έπειτα, προσαρμόζεται το ημι-παραμετρικό μοντέλο του Cox και γίνονται οι καμπύλες ROC. Τέλος, προσαρμόζονται τα μοντέλα Ridge και Lasso με βάση το ημι-παραμετρικό μοντέλο του Cox. Για την ανάλυση των δεδομένων χρησιμοποιήθηκαν τα στατιστικά πακέτα της R και του Minitab.

Abstract

This thesis deals with the statistical analysis of lifetime data making use of various methods and models for the purpose of scientific study of data that describe the time until an event occurs. The field of Statistics that we will deal with is Survival Analysis which has many applications in biomedical and technological sciences.

More specifically, the first chapter refers to the basic principles and definitions of Reliability and Survival Analysis (lifetime data, survival function, hazard function, cumulative hazard function etc.). Moreover, some basic distributions of Statistical Analysis are presented, for example the Weibull, Log-Normal, Log-Logistic etc. Finally, techniques from non-parametric lifetime data analysis are described (Kaplan-Meier estimator, Nelson-Aalen estimator etc.).

The second chapter presents in detail the Parametric Regression Models which are very often used in lifetime data. The first part of the chapter presents the Accelerated Life model (AL) as well as the Proportional Hazard model (PH). In the second part there is an extensive presentation of the semi-parametric Cox model (parameter estimation, graphical test for the proportional hazards hypothesis, Schoenfeld residuals, ROC & AUC curves, etc.). Finally, various variable selection criteria and measures of goodness of fit for the above models are presented.

The third chapter presents the Shrinkage methods used for the confrontation of multilinearity problems. More specifically, the Ridge and Lasso methods are presented, which have been greatly developed in recent years. The chapter concludes with a presentation of the Cross-Validation method for selecting the optimal factor λ used in all Shrinkage methods.

In the fourth chapter a simple set of lifetime data from fluorescent lamps is studied. At the beginning, an initial non-parametric analysis of the data is performed with the techniques presented in the previous chapters. An Accelerated Life model (AL) is then fitted and graphically tested via the Cox-Snell residuals.

The fifth and final chapter studies a set of lifetime data from patients who have undergone bone marrow transplantation. First, a non-parametric analysis of the data is performed, and then an Accelerated Life (AL) model is fitted, and a graphical test is performed through the Cox-Snell residuals. Moreover, the semi-parametric Cox model is fitted, and the ROC curves are constructed. Finally, the Ridge and Lasso models are fitted based on the semi-parametric Cox model. The statistical packages R and Minitab were used for data analysis.

ΠΕΡΙΕΧΌΜΕΝΑ

Κεφάλαιο 1: Ανάλυση Επιβίωσης

1.1 Εισαγωγικές Έννοιες – Βασικές Συναρτήσεις

Η Ανάλυση Επιβίωσης ή Ανάλυση Αξιοπιστίας είναι συγκεκριμένη στατιστική θεωρία που ασχολείται με δεδομένα διάρκειας ζωής. Τα δεδομένα διάρκειας ζωής είναι αυτά στα οποία μελετάμε κυρίως το χρόνο (ως μια συνεχής τυχαία μεταβλητή T>0) μέχρις ότου συμβεί κάποιο γεγονός. Το γεγονός αυτό συνήθως είναι δυσάρεστο, δηλαδή θάνατος ή υποτροπή ενός ασθενή, μηχανική βλάβη και άλλα. Μπορεί, όμως, το γεγονός να μην είναι τόσο δυσάρεστο όπως για παράδειγμα η αποθεραπεία ενός ασθενή. Η ανάλυση αυτή έχει πολλές εφαρμογές σε διάφορους επιστημονικούς χώρους, όπως στην ιατρική και τη μηχανική.

Πολλές φορές κατά την εκτέλεση ενός πειράματος στο οποίο καταγράφεται ο χρόνος λειτουργίας μέχρι να συμβεί ένα γεγονός, πολλές μονάδες συνεχίζουν να λειτουργούν και μετά τη λήξη του πειράματος. Αυτές τις μονάδες τις ονομάζουμε Αποκομμένα Δεδομένα (censored data). Αν και δεν είμαστε σε θέση να ξέρουμε πότε ακριβώς συνέβη το γεγονός στις εν λόγω μονάδες, ξέρουμε όμως ότι μετά από το τέλος της παρακολούθησης ήταν ακόμη λειτουργικές. Ο πιο συνηθισμένος τύπος αποκοπής είναι αυτός των δεξιά αποκομμένων παρατηρήσεων, δηλαδή έχουμε εκείνη την περίπτωση όπου κάποιες μονάδες παραμένουν σε λειτουργία μετά από τη χρονική στιγμή που σταματήσαμε το πείραμα (υπάρχουν, επιπλέον, οι περιπτώσεις αριστερά αποκοπής και αποκοπής σε διάστημα αλλά δε θα ασχοληθούμε στη μελέτη μας με αυτές). (Καρώνη 2009)

Συνάρτηση Επιβίωσης S(t):

Το βασικό μας ενδιαφέρον είναι να βρούμε και να υπολογίσουμε τη Συνάρτηση Επιβίωσης (ή Συνάρτηση Αξιοπιστίας), δηλαδή την πιθανότητα η τυχαία μεταβλητή του χρόνου T να ξεπεράσει το χρόνο t. Η Συνάρτηση Επιβίωσης ορίζεται από το τύπο:

$$S(t) = P(T > t) = 1 - F(t) \quad \text{για } t \geq 0 \qquad (1.1)$$

όπου $F(t)$ είναι η συνάρτηση κατανομής (σ.κ) της τυχαίας μεταβλητής T και ισχύει ότι $S(t) = \int_t^{\infty} f(u)du$.

Επίσης για τη συνάρτηση πυκνότητας πιθανότητας (σ.π.π.) ισχύει η σχέση:

$$f(t) = \frac{dF(t)}{dt} \quad \Rightarrow \quad f(t) = -\frac{dS(t)}{dt} \qquad (1.2)$$

Συνάρτηση Διακινδύνευσης h(t):

Η συνάρτηση Διακινδύνευσης (hazard function) εκφράζει τον κίνδυνο διακοπής μίας μονάδας στο χρονικό διάστημα (t, t + δt) με δεδομένο ότι έχει επιβιώσει μέχρι τη χρονική στιγμή t. Η συνάρτηση Διακινδύνευσης ορίζεται ως:

$$h(t) = \lim_{\delta t \to 0} \left(\frac{P[t < T \leq t + \delta t \,|T > t]}{\delta t} \right) \qquad (1.3)$$

Επίσης, από το τύπο δεσμευμένης πιθανότητας έχουμε:

$$P[t < T \leq t + \delta t \,|T > t] = \frac{P[t < T \leq t + \delta t]}{P[T > t]} \cong \frac{f(t)\delta t}{S(t)}$$

Οπότε η συνάρτηση Διακινδύνευσης γίνεται:

$$h(t) = \lim_{\delta t \to 0}\left(\frac{[S(t) - S(t + \delta t)]/S(t)}{\delta t}\right) \Rightarrow h(t) = \frac{f(t)}{S(t)} \qquad (1.4)$$

Η σχέση 1.4 ορίζει τη συνάρτηση Διακινδύνευσης συναρτήσει της συνάρτησης Επιβίωσης.

Όταν η συνάρτηση $h(t)$ είναι αύξουσα τότε αυξάνεται και η πιθανότητα να συμβεί το γεγονός (συνήθως όπως είπαμε είναι δυσάρεστο) στα δεδομένα μας, αντίθετα όταν είναι φθίνουσα μειώνεται ο κίνδυνος να συμβεί το γεγονός. (Collett 2003)

<u>Σωρευτική Συνάρτηση Διακινδύνευσης H(t):</u>

Η Σωρευτική συνάρτηση Διακινδύνευσης H(t) ορίζεται ως:

$$H(t) = \int_0^t h(u)du \qquad (1.5)$$

Η Σωρευτική συνάρτηση Διακινδύνευσης μας βοηθάει να επιλέξουμε το κατάλληλο στατιστικό μοντέλο κατά την ανάλυση των δεδομένων που έχουμε. Από τις σχέσεις (1.2), (1.4) και (1.5) έχουμε ότι:

$$H(t) = \int_0^t \frac{f(u)}{S(u)}du = -\int_0^t \frac{dS(u)}{S(u)} = -[\ln S(u)]_0^t = -\ln S(t) \quad \Rightarrow \quad S(t) = \exp[-H(t)] \quad (1.6)$$

Στη συνέχεια θα δούμε εν συντομία κάποιες βασικές κατανομές τις οποίες θα χρησιμοποιήσουμε στις δύο εφαρμογές που θα αναλύσουμε στις Παραγράφους 4 και 5.

<u>Κατανομή Weibull:</u>

Η κατανομή Weibull είναι πολύ συνηθισμένη στην ανάλυση αξιοπιστίας και αποτελεί ικανοποιητικό μοντέλο για πολλούς διαφορετικούς τύπους δεδομένων. Η σ.π.π. της Weibull ορίζεται ως:

$$f(t) = \eta\alpha^{-\eta}t^{\eta-1}\exp\{-(t/\alpha)^{\eta}\}, \quad t > 0$$

όπου η είναι η παράμετρος Σχήματος (>0) και α η παράμετρος Κλίμακας (>0). Όταν η=1 τότε έχουμε την Εκθετική κατανομή. Η παράμετρος η έχει σημαντικό ρόλο στην κατανομή:

$$\begin{cases} \text{Αν } \eta = 1 \Rightarrow h(t) = 1/\alpha \text{ (Εκθετική)} \\ \text{Αν } \eta > 1 \Rightarrow h(t) \text{ αύξουσα} \\ \text{Αν } \eta < 1 \Rightarrow h(t) \text{ φθίνουσα} \end{cases}$$

Επίσης από τις σχέσεις (1.1) και (1.4) βρίσκουμε: $\begin{cases} S(t) = \exp\{-(t/\alpha)^{\eta}\} \\ h(t) = \eta\alpha^{-\eta}t^{\eta-1} \end{cases}$

Κατανομή Gumbel:

Η κατανομή Gumbel συνδέεται με αυτή της Weibull με τον ακόλουθο τρόπο:

$$\begin{matrix} T \sim \text{Weibull}\,(\alpha, \eta) \\ \Leftrightarrow \\ Y = \ln T \sim \text{Gumbel}\,(\mu, \sigma) \end{matrix} \quad , \quad T > 0$$

Η συνάρτηση επιβίωσης της κατανομής Gumbel μπορεί να βρεθεί εύκολα από εκείνη της Weibull, δηλαδή:

$$S(y) = P[Y > y] = P[\ln t > y] = P[T > \exp(y)] = \exp\{-(t/\alpha)^{\eta}\} =$$

$$= \exp\{-\exp(y_{\eta} - \eta \ln \alpha)\} = \exp\{-\exp[\eta(y - \ln \alpha)]\} = \exp\left\{-\exp\left[\frac{y - \mu}{\sigma}\right]\right\}$$

Στη τελευταία ισότητα χρησιμοποιήσαμε τις σχέσεις: $\begin{cases} \mu = \ln \alpha \\ \eta = \sigma^{-1} \end{cases}$, οι οποίες προκύπτουν από τον ορισμό της κατανομής Gumbel. Οπότε η σ.π.π. θα είναι:

$$f(t) = S(t)\sigma^{-1}\exp\left[\frac{t - \mu}{\sigma}\right]$$

όπου μ είναι η παράμετρος κλίμακας και σ η παράμετρος σχήματος.

Κατανομή Log-Logistic:

Η κατανομή Log-Logistic (Λογαριθμο-λογιστική) έχει την ίδια λογική με την Log-Normal και ορίζεται ως:

$$\begin{matrix} T \sim \log - \text{logistic}\,(\nu, \tau) \\ \Leftrightarrow \\ Y = \ln T \sim \text{logistic}\,(\nu, \tau) \end{matrix} \quad , \; T > 0$$

Με σ.π.π: $f(t) = \left[\frac{1}{t\tau}\exp\left(\frac{\ln t - \nu}{\tau}\right)\right] / \left[1 + \exp\left(\frac{\ln t - \nu}{\tau}\right)\right]^2$, με παράμετρο κλίμακας ν και παράμετρο σχήματος τ. (Hosmer 2013)

Κατανομή Log-Normal:

Η κατανομή Log-Normal (Λογαριθμο-κανονική) ορίζεται ως:

$$\begin{matrix} T \sim \log - \text{normal}\,(\mu, \sigma^2) \\ \Leftrightarrow \\ Y = \ln T \sim N(\mu, \sigma^2) \end{matrix} \quad , \quad T > 0$$

Δηλαδή βασίζεται στην Κανονική κατανομή. Η σ.π.π. της ορίζεται ως:

$$f(t) = \frac{1}{\sigma t\sqrt{2\pi}}\exp\left\{\frac{-(\ln t - \mu)^2}{2\,\sigma^2}\right\} , \quad t > 0$$

με παράμετρο κλίμακας: exp(μ) και παράμετρο σχήματος: $\sigma^2 > 0$. Επίσης, θα έχει συνάρτηση επιβίωσης:

$$S(t) = 1 - \Phi\left(\frac{\ln t - \mu}{\sigma}\right)$$

όπου Φ είναι η συνάρτηση της Τυποποιημένης Κανονικής κατανομής N(0,1). (Καρώνη 2009)

1.2 Μη-παραμετρικές τεχνικές

1.2.1 Εκτιμήτρια Kaplan-Meier

Η εκτιμήτρια Kaplan-Meier είναι μία μη- παραμετρική τεχνική εκτίμησης της συνάρτησης Επιβίωσης S(t) που μας βοηθάει να καταλάβουμε ποιο θεωρητικό μοντέλο προσαρμόζεται καλύτερα στα δεδομένα που έχουμε, όταν δεν γνωρίζουμε εκ των προτέρων την κατανομή τους. Το γεγονός ότι πρόκειται για μία μη-παραμετρική τεχνική μας βοηθάει στις συχνές περιπτώσεις δεν γνωρίζουμε εκ των προτέρων την κατανομή που ακολουθούν οι παρατηρήσεις μας και δεν μπορούμε να διακρίνουμε με ευκολία ποια κατανομή μπορεί να ακολουθούν. Όλα αυτά καθιστούν την εκτιμήτρια Kaplan-Meier πολύ σημαντική για τη μελέτη μοντέλων επιβίωσης.

Έστω ότι έχουμε n παρατηρήσεις σε μερικές εκ των οποίων συμβαίνει το γεγονός σε χρόνους $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$, με $k \le n$. Έστω d_j το πλήθος των παρατηρήσεων στις οποίες συνέβη το γεγονός τη χρονική στιγμή $t_{(j)}$, j = 1,…, k και έστω n_j ο αριθμός των παρατηρήσεων που βρισκόντουσαν σε κίνδυνο αμέσως πριν τη χρονική στιγμή $t_{(j)}$.

$$S(t) = P\left[T > t_{(j)}\right] = P\left[T > t_{(1)}\right]P\left[T > t_{(2)} | T > t_{(1)}\right] \dots P\left[T > t_{(j)} \mid \left\{T > t_{(1)}\right\} \cap \dots \cap \left\{T > t_{(j-1)}\right\}\right] =$$

$$= P\left[\left\{T > t_{(1)}\right\} \cap \left\{T > t_{(2)}\right\} \cap \dots \cap \left\{T > t_{(j-1)}\right\}\right] =$$

$$= P\left[T > t_{(1)}\right]P\left[T > t_{(2)} | T > t_{(1)}\right] \dots P\left[T > t_{(j)} | T > t_{(j-1)}\right]$$

Για j = 1: $S\left(t_{(j)}\right) = P\left[T > t_{(1)}\right] = 1 - \frac{d_1}{n_1} = \frac{n_1 - d_1}{n_1}$

Εργαζόμενοι ανάλογα για τα υπόλοιπα j βρίσκουμε την εκτιμήτρια Kaplan-Meier:

$$\hat{S}(t) = \begin{cases} \prod_{j:\, t_{(j)} \le t} \frac{n_j - d_j}{n_j}, & \text{όταν } t \ge t_{(1)} \\ 1, & \text{όταν } t < t_{(1)} \end{cases}$$

Η γραφική παράσταση της εκτιμήτριας Kaplan-Meier δεν είναι καμπύλη (όπως θα περιμέναμε) αλλά μία βαθμωτή συνάρτηση. Τέλος, το τυπικό της σφάλμα υπολογίζεται από το τύπο του Greenwood για την εκτιμήτρια της διασποράς της $\hat{S}(t)$ που ορίζεται ως:

$$\hat{V}\left(\hat{S}(t)\right) = \hat{S}(t)^2 \sum_{t_{(j)} \le t} \frac{d_j}{n_j(n_j - d_j)} \qquad \text{και} \qquad se\left(\hat{S}(t)\right) = \sqrt{\hat{V}\left(\hat{S}(t)\right)}$$

<u>Γραφικοί Έλεγχοι:</u>

Με τη βοήθεια της μη-παραμετρικής εκτιμήτριας Kaplan-Meier μπορούμε να διενεργήσουμε με διάφορους γραφικούς ελέγχους ώστε να προσδιορίσουμε ποια γνωστή κατανομή ακολουθούν τα δεδομένα μας. Σημειώνεται ότι τα μη-παραμετρικά μοντέλα μας δίνουν πάντα αξιόπιστα

αποτελέσματα εφόσον δεν προϋποθέτουν τη χρήση συγκεκριμένης κατανομής που ακολουθούν τα δεδομένα μας. Ας υποθέσουμε ότι τα δεδομένα που έχουμε ακολουθούν την κατανομή Weibull, αλλά εμείς δεν το γνωρίζουμε.

Για την κατανομή Weibull ξέρουμε (από την Παράγραφο 1.1):

$$S(t) = \exp\{-(t/\alpha)^{\eta}\} \quad \Rightarrow \quad \ln S(t) = -(t/\alpha)^{\eta} \quad \Rightarrow \quad \ln\{-\ln(S(t))\} = \eta \ln t - \eta \ln \alpha$$

Η παραπάνω ισότητα μας δείχνει ότι μία γραφική παράσταση των τιμών της $\ln\{-\ln(\hat{S}(t_{(j)}))\}$ συναρτήσει των $\ln t_{(j)}$ θα δημιουργήσει μία ευθεία γραμμή εάν ισχύει το μοντέλο της κατανομής Weibull.

Με τον ίδιο τρόπο μπορούμε να εργαστούμε και με άλλες γνωστές κατανομές που υπάρχουν για να τις ελέγξουμε γραφικά με τη βοήθεια της εκτιμήτριας Kaplan-Meier. Στον παρακάτω πίνακα 1.1 παρουσιάζονται αντίστοιχοι γραφικοί έλεγχοι για κάποιες γνωστές κατανομές.

Πίνακας 1.1

Κατανομή	Γραφική Παράσταση
Εκθετική	$-\ln(S(t))$ με το t
Weibull	$\ln\{-\ln(S(t))\}$ με το $\ln t$
Κανονική	$\Phi^{-1}(1- S(t))$ με το t
Log-Normal	$\Phi^{-1}(1- S(t))$ με το $\ln t$
Λογιστική	$\ln\{(1- S(t)) / S(t)\}$ με το t
Log-Logistic	$\ln\{(1- S(t)) / S(t)\}$ με το $\ln t$

όπου Φ(z) είναι η συνάρτηση κατανομής της τυποποιημένης κανονικής κατανομής, δηλαδή: Φ(z) = P[Z ≤ z] με Z∼N(0,1). (Καρώνη 2009)

1.2.2 Εκτιμήτρια Nelson-Aalen

Η εκτιμήτρια Nelson-Aalen χρησιμοποιείται για την εκτίμηση της Σωρευτικής συνάρτησης Διακινδύνευσης H(t). Σημειώνεται ότι η εκτιμήτρια της H(t) μπορεί να υπολογιστεί και από τη σχέση (1.6) της Παραγράφου 1.1, όμως δεν προτιμάται αυτή η εκτιμήτρια.

Από τη σχέση (1.6) έχουμε: $\hat{H}_{NA}(t) = -\ln \hat{S}_{KM}(t) = -\sum_{j:\, t_{(j)} \le t} \ln\left\{1 - \frac{d_j}{n_j}\right\} \cong \sum_{j:\, t_{(j)} \le t} \frac{d_j}{n_j}$

Η τελευταία ισότητα προκύπτει από τη σχέση: $\ln(1-x) \cong -x$

Η εκτιμήτρια Nelson-Aalen είναι και αυτή μία βαθμωτή συνάρτηση και η εκτιμήτρια της διασποράς της $\hat{H}(t)$ υπολογίζεται από το τύπο:

$$\hat{V}\left(\hat{H}(t)\right) = \sum_{j:\, t_{(j)} \le t} \frac{d_j}{n_j^{\,2}}$$

(Καρώνη 2009)

1.2.3 Έλεγχος Log-rank

Ο έλεγχος Log-rank είναι ένας μη-παραμετρικός έλεγχος που χρησιμοποιείται για να συγκρίνουμε δύο ή και παραπάνω ομάδες στις οποίες έχουμε χωρίσει τα δεδομένα μας. Λέγεται μη-παραμετρικός γιατί δεν γνωρίζουμε μαθηματικά τις συναρτήσεις επιβίωσης των ομάδων των δεδομένων μας, αλλά όπως θα δούμε στη συνέχεια της παραγράφου ελέγχουμε την ισότητα των συναρτήσεων επιβίωσης που έχουμε.

Έστω ότι έχουμε $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$ διακεκριμένες χρονικές στιγμές και έχουμε χωρίσει τα δεδομένα μας σε δύο ομάδες (ομάδα 1 και 2). Σε αυτές τις χρονικές στιγμές παύουν να λειτουργούν οι παρατηρήσεις που προέρχονται και από τις δύο ομάδες. Θεωρούμε ότι αμέσως πριν τη χρονική στιγμή $t_{(j)}$ για την ομάδα 1 υπάρχουν $n_{1,j}$ παρατηρήσεις σε κίνδυνο από τις οποίες παύουν να λειτουργούν $d_{1,j}$ μονάδες ακριβώς τη στιγμή $t_{(j)}$. Αντίθετα, για την ομάδα 2 αμέσως πριν τη χρονική στιγμή $t(j)$ υπάρχουν $n_{2,j}$ παρατηρήσεις σε κίνδυνο από τις οποίες παύουν να λειτουργούν $d_{2,j}$ παρατηρήσεις ακριβώς τη στιγμή $t_{(j)}$. Επομένως, τη χρονική στιγμή $t_{(j)}$ παύουν να λειτουργούν συνολικά $d_j = d_{1,j} + d_{2,j}$ παρατηρήσεις από τις $n_j = n_{1,j} + n_{2,j}$ που ήταν σε κίνδυνο. Αυτό φαίνεται πιο αναλυτικά στον παρακάτω πίνακα συνάφειας (Πίνακας 1.2).

Πίνακας 1.2

		Ομάδα		
		A	B	$\sum$
Διακοπή Λειτουργίας	Ναι	$d_{1,j}$	$d_{2,j}$	d_j
	Όχι	$n_{1,j} - d_{1,j}$	$n_{2,j} - d_{2,j}$	$n_j - d_j$
	$\sum$	$n_{1,j}$	$n_{2,j}$	n_j

Στον πίνακα συνάφειας 1.2 θα ελέγξουμε τη μηδενική υπόθεση (H_0) ότι το γεγονός είναι ανεξάρτητο της ομάδας στην οποία ανήκουν οι παρατηρήσεις μας, δηλαδή την υπόθεση $H_0: S_A(t) = S_B(t)$, εφαρμόζοντας το γνωστό X^2- έλεγχο και υπολογίζοντας τις αναμενόμενες συχνότητες. Για παράδειγμα, η αναμενόμενη συχνότητα του δεύτερου κελιού του πίνακα 1.2 (παρατηρήσεις που ανήκουν στην ομάδα B και που έχει διακοπεί η λειτουργία τους) θα είναι:

$$E(d_{2,j}) = \hat{d}_{2,j} = n_{2,j}\, d_j / n_j$$

Οπότε η απόκλιση από την παρατηρούμενη $d_{2,j}$ θα είναι: $u_j = d_{2,j} - (n_{2,j}\, d_j / n_j)$

Θεωρώντας ότι οι συχνότητες $d_{i,j}$ ακολουθούν την Υπεργεωμετρική κατανομή μπορούμε εύκολα να υπολογίσουμε τη διασπορά τους.

Έστω ότι i=2: $V(d_{2,j}) = \frac{n_{2,j} n_{1,j} d_j (n_j - d_j)}{n_j{}^2 (n_j - 1)} \equiv v_j$

Οπότε η ελεγχοσυνάρτηση του ελέγχου Log-rank ορίζεται ως:

$$\frac{u}{\sqrt{v}} \sim N(0,1)\ ,\ \text{όπου } u = \sum_{j=1}^{k} u_j \text{ και } v = \sum_{j=1}^{k} v_j$$

Ή ισοδύναμα αποδεικνύεται ότι: $\frac{u^2}{v} \sim X_1^2$

Μία γενίκευση του ελέγχου Log-rank είναι ο έλεγχος Wilcoxon. Πιο συγκεκριμένα η ελεγχοσυνάρτηση του ελέγχου Wilcoxon παραμένει ίδια με αυτή του Log-rank, με τη διαφορά ότι θα ορίσουμε διαφορετικά τα u και v, δηλαδή $u = \sum_{j=1}^{k} w_j u_j$ και $v = \sum_{j=1}^{k} w_j v_j$, όπου $w_j = n_j$ οι

συντελεστές στάθμισης. Δηλαδή ο έλεγχος Wilcoxon δίνει διαφορετικό βάρος w_j ανάλογα με τη χρονική στιγμή $t_{(j)}$ που έχουμε και πιο συγκεκριμένα δίνει μεγαλύτερο βάρος στο ξεκίνημα του πειράματος. Ενώ στον έλεγχο Log-rank έχουμε συντελεστές στάθμισης $w_j = 1$.

Σημειώνεται ότι ο έλεγχος Log-rank επεκτείνεται και για την περίπτωση που έχουμε παραπάνω από δύο ομάδες στα δεδομένα μας. (Καρώνη 2009)

Κεφάλαιο 2: Παραμετρικά Μοντέλα Παλινδρόμησης

2.1 Συμμεταβλητές

Στα δεδομένα που μελετάει η επιστήμη της Ανάλυσης Επιβίωσης συναντάμε αρκετά συχνά πολλούς μετρήσιμους παράγοντες που επηρεάζουν τη διάρκεια ζωής των παρατηρήσεών μας. Για παράδειγμα, η διάρκεια ζωής μιας μηχανής μπορεί να επηρεάζεται από τις συνθήκες λειτουργίας της, τα χαρακτηριστικά από τα οποία κατασκευάστηκε και άλλα, ενώ η διάρκεια ζωής ενός ανθρώπου μπορεί να επηρεάζεται από το φύλο του (άνδρας ή γυναίκα), από την ηλικία του και άλλα πολλά. Όλοι αυτοί οι παράγοντες ονομάζονται συμμεταβλητές. Οπότε καταλαβαίνουμε ότι στα μοντέλα διάρκειας ζωής των παρατηρήσεών μας πρέπει να εισαχθούν και αυτοί οι παράγοντες που τις επηρεάζουν σημαντικά. Έτσι πετυχαίνουμε καλύτερη περιγραφή της διάρκειας ζωής των δεδομένων μας.

Οι συμμεταβλητές χωρίζονται σε δύο κατηγορίες: τις ποσοτικές και τις κατηγορικές. Οι ποσοτικές αφορούν παράγοντες που είναι μετρήσιμοι, για παράδειγμα ηλικία, θερμοκρασία κ.τ.λ., ενώ οι κατηγορικές αφορούν παράγοντες όπου οι τιμές τους αφορούν ένα χαρακτηριστικό. Ένα παράδειγμα μιας κατηγορικής μεταβλητής είναι ο τύπος μηχανής που έχει κατασκευαστεί για ένα αυτοκίνητο:

$$\mathrm{x} = \begin{cases} 1, & \text{έχει μηχανή τύπου Α} \\ 2, & \text{έχει μηχανή τύπου Β} \\ 3, & \text{έχει μηχανή τύπου Γ} \end{cases}$$

Τις κατηγορικές μεταβλητές θα τις ορίζουμε μέσω ψευδομεταβλητών, δηλαδή μέσω εικονικών μεταβλητών. Οπότε για τις τρεις μηχανές θα έχουμε δύο ψευδομεταβλητές:

$$\mathrm{x}_1 = \begin{cases} 1, & \text{έχει μηχανή τύπου Α} \\ 0, & \text{αλλιώς} \end{cases} \qquad \text{και} \qquad \mathrm{x}_2 = \begin{cases} 1, & \text{έχει μηχανή τύπου Β} \\ 0, & \text{αλλιώς} \end{cases}$$

(Καρώνη 2009)

2.2 Μοντέλο Παλινδρόμησης Επιταχυνόμενης Διακοπής (AL)

Όπως γνωρίζουμε η γραμμική παλινδρόμηση είναι το βασικό μοντέλο της στατιστικής στο οποίο η τιμή της εξαρτημένης μεταβλητής Y, συνδέεται γραμμικά με τις συμμεταβλητές x_I, και πάνω σε αυτό θα προσαρμόσουμε το μοντέλο της Επιταχυνόμενης Διακοπής ή Επιταχυνόμενης Διάρκειας ζωής (Accelerated Life model). Το γραμμικό μοντέλο είναι:

$$\mathrm{y} = \beta_0 + \beta_1 \mathrm{x}_1 + \cdots + \beta_\kappa \mathrm{x}_\kappa + \epsilon = \beta' \mathrm{x} + \epsilon$$

όπου τα υπόλοιπα ϵ είναι ανεξάρτητες και ισόνομες τ.μ. που ακολουθούν την κανονική κατανομή, δηλαδή: $\epsilon \sim \mathrm{N}(0, \sigma^2)$. Εδώ παρατηρούμε ότι η εξαρτημένη μεταβλητή Y δεν μπορεί να είναι πάντα θετική όπως απαιτείται από τα δεδομένα διάρκειας ζωής (καθώς T > 0). Για αυτό το λόγο θα χρησιμοποιήσουμε τη συνάρτηση $\ln \mathrm{T}$ και έτσι το μοντέλο θα γίνει:

$$\ln \mathrm{T}_\mathrm{x} = \mu_\mathrm{x} + \sigma\epsilon = \mu_0 + \beta' \mathrm{x} + \sigma\epsilon$$

όπου μ_0 και σ είναι οι παράμετροι θέσης και κλίμακας αντίστοιχα και ϵ είναι μία τυχαία μεταβλητή. Θεωρούμε ότι οι συμμεταβλητές που έχουμε εισάγονται στο μοντέλο με βάση την επίδρασή τους

στην παράμετρο μ_x=μ(x) και συνήθως χρησιμοποιούμε $\mu(x) = \exp(\beta' x)$. Η συνάρτηση επιβίωσης γράφεται:

$$S(t;x) = P[T_x > t] = P[\ln T_x > \ln t] = P[\mu_0 + \beta' x + \sigma\epsilon > \ln t] =$$

$$= P[\ln T_0 + \beta' x > \ln t] = P[T_0 > t\exp(-\beta' x)] = S_0\big(t\exp(-\beta' x)\big)$$

Δηλαδή σε γενική μορφή γράφεται ως: $S(t;x) = S_0(t\,g(x))$, όπου $g(x)$ μία θετική συνάρτηση των συμμεταβλητών και S_0 μία βασική συνάρτηση επιβίωσης. Συνήθως χρησιμοποιούμε $g(x) = \exp(\beta' x)$ χωρίς να είναι δεσμευτικό. (Καρώνη 2009)

2.2.1 Προσαρμογή του μοντέλου & Γραφικός έλεγχος

<u>Γραφικός έλεγχος του μοντέλου:</u>

Ένας γραφικός έλεγχος πρέπει να εφαρμόζεται στα δεδομένα που μελετάμε για να ελέγξουμε εάν ισχύει το μοντέλο της Επιταχυνόμενης Διακοπής. Όπως είδαμε η συνάρτηση επιβίωσης γράφεται:

$$S(t;x) = S_0(t\,g(x)) = P[T_0 > t\,g(x)] = P[\ln T_0 > \ln t + \ln g(x)] = S^*(y + \ln g(x))$$

όπου S^* η συνάρτηση επιβίωσης της τ.μ. Y = $\ln T_0$. Η παραπάνω σχέση μας δείχνει ότι μία γραφική παράσταση της $S(t;x)$ συναρτήσει του λογαριθμισμένου χρόνου $\ln t$ θα είναι μία οριζόντια μετατόπιση της S^*. Αυτό σημαίνει ότι για να ισχύει η υπόθεση της Επιταχυνόμενης Διακοπής θα πρέπει όλες οι καμπύλες της S (για διαφορετικές τιμές της x) να διατηρούν μεταξύ τους ίσες μετατοπίσεις στον άξονα του $\ln t$. Σημειώνουμε ότι θα χωρίσουμε τα δεδομένα μας σε ομάδες σύμφωνα με τις κοινές τιμές της μεταβλητής x και με αυτά τα διαφορετικά x θα κάνουμε τις γραφικές παραστάσεις της S^*. Για αυτό το γραφικό έλεγχο θα χρησιμοποιήσουμε τις εκτιμήσεις της Kaplan-Meier. Εφαρμογή αυτού του γραφικού ελέγχου θα δούμε αναλυτικά στις δύο εφαρμογές που θα μελετήσουμε στα Κεφάλαια 4 & 5 (Παράγραφοι 4.6 & 5.7).

<u>Προσαρμογή μοντέλου:</u>

Όπως έχουμε δει μέχρι στιγμής στο μοντέλο της Επιταχυνόμενης Διακοπής ισχύει:

$$S(t;x) = P[\mu_0 + \beta' x + \sigma\epsilon > \ln t]$$

Θα λύσουμε την παραπάνω ανισότητα (που υπάρχει μέσα στην πιθανότητα) ως προς τη τ.μ. ϵ. Δηλαδή θα έχουμε:

$$S(t;x) = P[\epsilon > (\ln t - \mu_0 - \beta' x)/\sigma] = S_\epsilon((\ln t - \mu_0 - \beta' x)/\sigma)$$

όπου S_ϵ είναι η συνάρτηση αξιοπιστίας της τ.μ. ϵ και έστω $\epsilon = (\ln t - \mu_0 - \beta' x)/\sigma$. Έστω ότι οι παρατηρήσεις μου ακολουθούν την κατανομή Weibull, τότε η σ.π.π. θα είναι:

$$f(t;x) = -\frac{dS(t;x)}{dt} = \frac{1}{\sigma t} f_\epsilon(\epsilon)$$

Θα προσαρμόσουμε το μοντέλο με τη μέθοδο της Μέγιστης Πιθανοφάνειας. Έστω ότι έχουμε δεξιά αποκομμένες παρατηρήσεις (τις συναντάμε πιο συχνά) τότε η συνάρτηση Πιθανοφάνειας θα είναι:

$$L = \prod_{i=1}^{n} [f(t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i} = \prod_{i=1}^{n} (\sigma t_i)^{-\delta_i} [f_\epsilon(\epsilon_i)]^{\delta_i} [S_\epsilon(\epsilon_i)]^{1-\delta_i}$$

όπου δ_i είναι ένας δείκτης που παίρνει τιμές 0 ή 1 ανάλογα με το αν η τιμή i είναι αποκομμένη ή όχι αντίστοιχα. Λογαριθμίζοντας την παραπάνω σχέση έχουμε:

$$l = \sum_{i=1}^{n} \{-\delta_i \ln(\sigma t_i) + \delta_i \ln f_\epsilon(\epsilon_i) + (1-\delta_i) \ln S_\epsilon(\epsilon_i)\}$$

Η παραπάνω συνάρτηση l θα πρέπει στη συνέχεια να μεγιστοποιηθεί, παραγογίζοντάς την ως προς τις μεταβλητές σ, μ_0 και β, και οι τιμές των τριών αυτών παραμέτρων μπορούν να υπολογισθούν με αριθμητικές μεθόδους (συνήθως χρησιμοποιούμε τη μέθοδο Newton-Raphson). Στο τέλος θα έχουμε βρει τις εκτιμήσεις της μέγιστης Πιθανοφάνειας για τις παραμέτρους που έχουμε.

Αντίστοιχα μπορούμε να εργαστούμε και για άλλες κατανομές όπως η Log-Normal και η Log-Logistic, απλά αντικαθιστώντας στη σ.π.π. f(t) και στη συνάρτηση επιβίωσης S(t) την παράμετρο σχήματος τ με την παράμετρο σ και την παράμετρο κλίμακας ν με την $\epsilon = (\ln t - \mu_0 - \beta' x)/\sigma$. Με αυτή την αντικατάσταση θα βρούμε τις αντίστοιχες συναρτήσεις $f_\epsilon(\epsilon_i)$ και $S_\epsilon(\epsilon_i)$ των παραπάνω κατανομών, συνεπώς θα μπορέσουμε να εκτιμήσουμε τις παραμέτρους με την παραπάνω μέθοδο. (Καρώνη 2009)

2.2.2 Έλεγχος μέσω υπολοίπων Cox-Snell

Ένας τρόπος να ελέγξουμε αν το μοντέλο μας είναι το κατάλληλο είναι να μελετήσουμε τα υπόλοιπα μετά την προσαρμογή του μοντέλου. Στην κλασική περίπτωση της γραμμικής παλινδρόμησης τα υπόλοιπα υπολογίζονται από τη σχέση: $\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}' x_i$

και εκφράζουν τη διαφορά μεταξύ της παρατηρούμενης τιμής, yi, και της προσαρμοσμένης, $\hat{y}_i$. Όμως στα δεδομένα διάρκειας ζωής που μελετάμε χρησιμοποιούμε τα γενικευμένα υπόλοιπα Cox-Snell. Έστω ότι έχουμε συναρτήσεις $w_i(T_I; x_I, \theta)$, ανεξάρτητες και ισόνομες, τότε οι τιμές $\hat{\epsilon}_i = w_i(T_I; x_I, \hat{\theta})$ μπορούν να χρησιμοποιηθούν σαν υπόλοιπα, όπου x_I οι συμμεταβλητές που έχουμε και θ οι παράμετροι του μοντέλου. Έστω F η κατανομή της τ.μ. Y τότε ισχύει: $V = F(Y) \sim U(0,1)$, όπου U η ομοιόμορφη κατανομή, τότε αποδεικνύεται ότι η τ.μ. $U = -\ln(1 - F(Y))$ θα έχει σ.π.π. την Εκθετική κατανομή με παράμετρο 1.

σ.π.π. της U: $g(u) = e^{-u}$, $u > 0$

Οπότε τα υπόλοιπα Cox-Snell θα ορίζονται ως: $-\ln \hat{S}(t_I; x_i) = \hat{H}(t_I; x_i) = \hat{\epsilon}_i$

Στην περίπτωση που έχουμε δεξιά αποκομμένες παρατηρήσεις, τότε τα υπόλοιπα Cox-Snell για αυτές τις παρατηρήσεις θα πάρουν τη διορθωμένη μορφή: $\hat{\epsilon}_i = 1 - \ln \hat{S}(t_I; x_i)$

Σημειώνεται ότι τα υπόλοιπα Cox-Snell μπορούν να εφαρμοστούν εύκολα σε μοντέλα οποιασδήποτε μορφής είτε έχουν συμμεταβλητές είτε δεν έχουν. Ο γραφικός έλεγχος των υπολοίπων Cox-Snell γίνεται σε σύγκριση με το γράφημα της Εκθετικής κατανομής με παράμετρο 1, που πρόκειται για μία ευθεία γραμμή. Στην περίπτωση όπου οι τιμές των υπολοίπων πλησιάζουν αρκετά την ευθεία της Εκθετικής κατανομής τότε συμπεραίνουμε ότι το μοντέλο που προσαρμόσαμε ταιριάζει στα δεδομένα μας. Στις εφαρμογές των Κεφαλαίων 4 & 5 θα χρησιμοποιήσουμε αυτό το γραφικό έλεγχο των υπολοίπων Cox-Snell για να ελέγξουμε το μοντέλο της Επιταχυνόμενης Διακοπής (Παράγραφοι 4.6.1 & 5.7). (Καρώνη 2009)

2.3 Μοντέλο Παλινδρόμησης Αναλογικής Διακινδύνευσης (PH)

Θα προσαρμόσουμε το μοντέλο της Αναλογικής Διακινδύνευσης (Proportional Hazards) πάνω στο γενικό γραμμικό μοντέλο όπως κάναμε στην Παράγραφο 2.2.1. Το μοντέλο της Αναλογικής Διακινδύνευσης ορίζεται από την έκφραση:

$$h(t;x) = g(x)h_0(t) \qquad (2.1)$$

όπου $h_0(t)$ είναι μία βασική συνάρτηση διακινδύνευσης και g(x) > 0. Υπάρχουν δύο είδη μοντέλων παλινδρόμησης Αναλογικής Διακινδύνευσης που βασίζονται στη συνάρτηση $h_0(t)$:

1) Παραμετρικά μοντέλα, όπου η συνάρτηση διακινδύνευσης $h_0(t)$ καθορίζεται από κάποιο γνωστό παραμετρικό μοντέλο (π.χ. Weibull, Log-Logistic, κ.τ.λ.).
2) Ημι-παραμετρικά μοντέλα, όπου η συνάρτηση $h_0(t)$ παραμένει ακαθόριστη, το πιο βασικό παράδειγμα ημι-παραμετρικού μοντέλου είναι αυτό του Cox με το οποίο θα ασχοληθούμε στη συνέχεια (Παράγραφος 2.5).

Έστω ότι η διάρκεια ζωής T ακολουθεί την κατανομή Weibull, τότε ισχύουν τα ακόλουθα:

$$g(x) = \exp(-\beta' x) \quad \text{και} \quad h_0(t) = \eta\alpha^{-\eta}t^{\eta-1} \qquad (2.2)$$

Αντικαθιστώντας τις σχέσεις (2.2) στη σχέση (2.1) έχουμε: $h(t;x) = \eta t^{\eta-1}\{\alpha e^{-\beta' x/\eta}\}^{-\eta}$

Η συνάρτηση $h(t;x)$ παραμένει συνάρτηση διακινδύνευσης με τη μόνη διαφορά ότι έχει διαφορετική παράμετρο κλίμακας, δηλαδή πλέον έχει παράμετρο κλίμακας $\alpha e^{-\beta' x/\eta}$.

Έστω ότι έχουμε 2 μονάδες, τότε οι συναρτήσεις διακινδύνευσής τους θα είναι:

$$\begin{cases} h(t|\lambda_1) = \lambda_1 h_0(t) \\ h(t|\lambda_2) = \lambda_2 h_0(t) \end{cases}$$

όπου λ_1, λ_2 είναι άγνωστες τ.μ. Ο λόγος αυτών των δύο συναρτήσεων διακινδύνευσης είναι:

$$\frac{h(t|\lambda_1)}{h(t|\lambda_2)} = \frac{\lambda_1 h_0(t)}{\lambda_2 h_0(t)} = \frac{\lambda_1}{\lambda_2}$$

Ο λόγος αυτός είναι ανεξάρτητος του χρόνου t και αποτελεί την «ιδιότητα της αναλογικής διακινδύνευσης». Αν αντί για τις άγνωστες τ.μ. βάλουμε τη συνάρτηση $g(x) = \exp(\beta' x)$ τότε ο λόγος γίνεται:

$$\frac{h(t;x_1)}{h(t;x_2)} = \frac{\exp(\beta' x_1)h_0(t)}{\exp(\beta' x_2)h_0(t)} = \exp[\beta'(x_1 - x_2)]$$

(Sachin 2020 & Καρώνη 2009)

<u>Γραφικός έλεγχος του μοντέλου:</u>

Ένας γραφικός έλεγχος πρέπει να εφαρμόζεται στα δεδομένα που μελετάμε για να ελέγξουμε εάν ισχύει το μοντέλο της Αναλογικής Διακινδύνευσης, ώστε να μπορέσουμε να το προσαρμόσουμε μετά. Έστω $g(x) = \exp(\beta' x)$, τότε η συνάρτηση διακινδύνευσης θα γίνει:

$$h(t;x) = h_0(t)\exp(\beta' x)$$

Συνεπώς από τον ορισμό της $\mathrm{H}(\mathrm{t};\mathrm{x})$ (όπως είδαμε στην Παράγραφο 1.1) η συνάρτηση επιβίωσης θα γίνει:

$$\mathrm{S}(\mathrm{t};\mathrm{x}) = \exp\{-\mathrm{H}_0(\mathrm{t})\mathrm{e}^{\beta'\mathrm{x}}\} \;\Rightarrow\; \ln\{-\ln \mathrm{S}(\mathrm{t};\mathrm{x})\} - \ln \mathrm{H}_0(\mathrm{t}) = \beta'\mathrm{x}$$

όπου $\mathrm{H}_0(\mathrm{t})$ είναι μία βασική σωρευτική συνάρτηση διακινδύνευσης. Η παραπάνω σχέση μας δείχνει ότι οι καμπύλες των $\ln\{-\ln \mathrm{S}(\mathrm{t};\mathrm{x})\}$ για διαφορετικές τιμές της x, είναι παράλληλες μεταξύ τους ως προς τον χρόνο t. Σημειώνουμε ότι όπως και στο γραφικό έλεγχο της Επιταχυνόμενης Διακοπής (Παράγραφος 2.2.1) θα χωρίσουμε τα δεδομένα μας σε ομάδες σύμφωνα με τις κοινές τιμές της μεταβλητής x και με αυτά τα διαφορετικά x θα κάνουμε τις καμπύλες $\ln\{-\ln \mathrm{S}(\mathrm{t};\mathrm{x})\}$. Για τη συνάρτηση επιβίωσης $\mathrm{S}(\mathrm{t};\mathrm{x})$ θα χρησιμοποιήσουμε την εκτίμηση της Kaplan-Meier, $\hat{\mathrm{S}}_{\mathrm{KM}}(\mathrm{t};\mathrm{x})$.

Σημειώνεται ότι μπορώ να σχηματίσω τις καμπύλες των $\ln\{-\ln \mathrm{S}(\mathrm{t};\mathrm{x})\}$ συνάρτήσει του λογαριθμισμένου χρόνου $\ln \mathrm{t}$, αντί για t. Σε αυτή την περίπτωση θα έχω ευθείες αντί για καμπύλες, και θα πρέπει πάλι να είναι οι ευθείες παράλληλες μεταξύ τους. Οπότε για να ισχύει η υπόθεση της Αναλογικής Διακινδύνευσης θα πρέπει οι καμπύλες που θα δημιουργήσουμε να είναι παράλληλες μεταξύ τους ως προς τον άξονα του χρόνου, δηλαδή να διαφέρουν μόνο ως προς τις οριζόντιες μετατοπίσεις. Εφαρμογή αυτού του γραφικού ελέγχου θα δούμε πιο αναλυτικά στην εφαρμογή που θα μελετήσουμε στο Κεφάλαιο 5 (Παράγραφος 5.8). (Καρώνη 2009)

<u>Προσαρμογή μοντέλου:</u>

Η προσαρμογή του μοντέλου της Αναλογικής Διακινδύνευσης θα γίνει με τη μέθοδο της Μέγιστης Πιθανοφάνειας (όπως κάναμε και στο μοντέλο της Επιταχυνόμενης Διακοπής, Παράγραφος 2.2.1). Έστω ότι έχουμε δεξιά αποκομμένες παρατηρήσεις (συχνότερες) τότε η συνάρτηση Πιθανοφάνειας παίρνει τη μορφή:

$$L = \prod_{\mathrm{i}=1}^{\mathrm{n}} [\mathrm{f}(\mathrm{t}_\mathrm{i})]^{\delta_\mathrm{i}} [\mathrm{S}(\mathrm{t}_\mathrm{i})]^{1-\delta_\mathrm{i}}$$

όπου δ_i είναι ένας δείκτης που παίρνει τιμές 0 ή 1 ανάλογα με το αν η τιμή i είναι αποκομμένη ή όχι αντίστοιχα. Λογαριθμίζοντας τη συνάρτηση Πιθανοφάνειας έχουμε:

$$l = \sum_{\mathrm{i}=1}^{\mathrm{n}} \{\delta_\mathrm{i} \ln \mathrm{f}(\mathrm{t}_\mathrm{i}) + (1-\delta_\mathrm{i}) \ln \mathrm{S}(\mathrm{t}_\mathrm{i})\}$$

Χρησιμοποιώντας στη συνάρτηση Πιθανοφάνειας, L, τη συνάρτηση διακινδύνευσης $\mathrm{h}(\mathrm{t}) = \frac{\mathrm{f}(\mathrm{t})}{\mathrm{S}(\mathrm{t})}$, θα έχουμε:

$$L = \prod_{\mathrm{i}=1}^{\mathrm{n}} [h(\mathrm{t}_\mathrm{i})]^{\delta_\mathrm{i}} \mathrm{S}(\mathrm{t}_\mathrm{i})$$

Αυτή η έκφραση της συνάρτησης Πιθανοφάνειας L φαίνεται ότι είναι πιο χρήσιμη για τις μελέτες μας. Η συνάρτηση επιβίωσης υπό την υπόθεση της Αναλογικής Διακινδύνευσης θα είναι:

$$\mathrm{S}(\mathrm{t};\mathrm{x}) = \exp\{-\mathrm{H}_0(\mathrm{t})\mathrm{e}^{\beta'\mathrm{x}}\} \;,\; \text{όπου } \mathrm{H}_0(\mathrm{t}) = -\ln \mathrm{S}_0(\mathrm{t};\mathrm{x})$$

(από τύπο της $\mathrm{H}(\mathrm{t};\mathrm{x})$ που είδαμε στην Παράγραφο 1.1)

Συνεπώς: $$S(t;x) = \exp\{e^{\beta' x} \ln S_0(t;x)\} = \exp\left\{\ln S_0(t;x)^{e^{\beta' x}}\right\} = \{S_0(t;x)\}^{e^{\beta' x}} \quad (2.3)$$

Αντικαθιστώντας τις σχέσεις (2.1) και (2.3) στην παραπάνω συνάρτηση Πιθανοφάνειας L θα έχουμε:

$$L = \prod_{i=1}^{n} \left[h_0(t_i) e^{\beta' x_i}\right]^{\delta_i} \{S_0(t_i; x_i)\}^{e^{\beta' x_i}}$$

Λογαριθμίζοντας έχουμε:

$$l = \sum_{i=1}^{n} \left\{\delta_i \beta' x_i + \delta_i \ln h_0(t_i) + e^{\beta' x_i} \ln S_0(t_i; x_i)\right\}$$

Στη συνέχεια ανάλογα με το ποια κατανομή ακολουθούν τα δεδομένα που έχουμε, αντικαθιστούμε την αντίστοιχη συνάρτηση διακινδύνευσης $h_0(t)$ και την αντίστοιχη συνάρτηση επιβίωσης $S_0(t)$. Για παράδειγμα, αν οι παρατηρήσεις μου ακολουθούσαν την κατανομή Weibull θα αντικαθιστούσαμε τις συναρτήσεις:

$$\begin{cases} S_0(t) = \exp\{-(t/\alpha)^{\eta}\} \\ h_0(t) = \eta \alpha^{-\eta} t^{\eta - 1} \end{cases}$$

Η συνάρτηση l που θα δημιουργηθεί, θα πρέπει στη συνέχεια να μεγιστοποιηθεί, παραγωγίζοντάς την ως προς τις παραμέτρους που έχει η κάθε κατανομή και οι τιμές αυτών παραμέτρων μπορούν να υπολογισθούν με αριθμητικές μεθόδους (συνήθως χρησιμοποιούμε τη μέθοδο Newton-Raphson). Στο τέλος θα έχουμε βρει τις εκτιμήσεις της μέγιστης Πιθανοφάνειας για τις παραμέτρους που έχουμε. (Xu et al. 2009)

2.4 Το ημι-παραμετρικό μοντέλο του Cox

Το μοντέλο του Cox (1972) είναι ένα μοντέλο Αναλογικής Διακινδύνευσης (PH) που ανήκει στην κατηγορία των ημι-παραμετρικών μοντέλων όπου η βασική συνάρτηση διακινδύνευσης $h_0(t)$ παραμένει ακαθόριστη (Παράγραφος 2.3). Η βασική συνάρτηση διακινδύνευσης $h_0(t)$ είναι ακαθόριστη στις περιπτώσεις που μελετάμε μεγάλα σύνολα δεδομένων διάρκειας ζωής που αφορούν τον άνθρωπο. Αυτό συμβαίνει διότι σε αντίθεση με τα δεδομένα διάρκειας ζωής από τεχνολογικές εφαρμογές όπου, μπορούμε να καταλήξουμε σε ένα παραμετρικό μοντέλο με βάση προηγούμενης εμπειρίας που έχουμε με παρόμοιο υλικό το γεγονός ότι κάθε άτομο είναι διαφορετικό και φέρει τα δικά του μοναδικά χαρακτηριστικά κάνει το έργο της προσαρμογής ενός και μόνο γνωστού παραμετρικού μοντέλου ακατόρθωτο. Έτσι, θέλοντας να προσαρμόσουμε ένα μοντέλο αναλογικής διακινδύνευσης δεν είμαστε σε θέση να γνωρίζουμε την παραμετρική μορφή της βασικής συνάρτησης κινδύνου $h_0(t)$ και οδηγούμαστε στα λεγόμενα ημι-παραμετρικά μοντέλα PH. Το πιο γνωστό μοντέλο αυτής της μορφής είναι το μοντέλο αναλογικής διακινδύνευσης του Cox (the Cox proportional hazards model). Όπως και στα μοντέλα αναλογικής διακινδύνευσης η συνάρτηση διακινδύνευσης $h(t, x)$ ορίζεται ως:

$$h(t;x) = h_0(t)\exp(\beta' x)$$

και αντίστοιχα η σωρευτική συνάρτηση διακινδύνευσης θα είναι:

$$H(t;x) = \int_0^t h_0(u)\exp(\beta' x)du = H_0(t)\exp(\beta' x)$$

όπου με $H_0(t)$ μία βασική σωρευτική συνάρτηση διακινδύνευσης που αντιστοιχεί στην $h_0(t)$ και επίσης η συνάρτηση επιβίωσης θα είναι:

$$S(t;x) = \exp\{-H(t;x)\} = \exp\{-H_0(t)e^{\beta' x}\} = \{S_0(t;x)\}^{e^{\beta' x}}$$

Το κύριο χαρακτηριστικό του μοντέλου του Cox, όπως έχουμε αναφέρει, είναι ότι οι συγκεκριμένες παραμετρικές μορφές των συναρτήσεων $h_0(t)$ και $S_0(t)$, δεν καθορίζονται και μόνο η επίδραση του διανύσματος των συμμεταβλητών x_I αναλύεται. Παρόλα αυτά πρέπει να τονίσουμε ότι η υπόθεση της αναλογικότητας εξακολουθεί να ισχύει όπως και στα κλασικά μοντέλα Αναλογικής Διακινδύνευσης, καθώς αυτή δεν εξαρτάται από τον χρόνο. (Fox & Weisberg 2018)

2.4.1 Εκτίμηση παραμέτρων

Στη συνέχεια μας ενδιαφέρει να εκτιμήσουμε τις παραμέτρους του μοντέλου του Cox με τη μέθοδο Μεγίστης Πιθανοφάνειας. Έστω ότι τις διακεκριμένες χρονικές στιγμές $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$ διακόπτεται η λειτουργία k μονάδων. Δηλαδή, τη χρονική στιγμή $t_{(j)}$, $j = 1, \dots, k$, σταματά να λειτουργεί μόνο μία μονάδα με x_j συμμεταβλητές. Έστω επίσης R_j το σύνολο των μονάδων που βρίσκονται σε κίνδυνο αμέσως πριν τη χρονική στιγμή $t_{(j)}$. Γνωρίζουμε από τη θεωρία πιθανοτήτων ότι η πιθανότητα να διακοπεί η λειτουργία μίας μονάδας j, δοθέντος ότι σταματά να λειτουργεί μία οποιαδήποτε μονάδα από το σύνολο R_j, δίνεται από το τύπο:

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

Επομένως, η συνάρτηση μερικής Πιθανοφάνειας για το σύνολο των δεδομένων ορίζεται από το τύπο:

$$L(\beta) = \prod_{j=1}^{k} \left\{ \frac{e^{\beta' x_j}}{\sum_{i \in R_j} e^{\beta' x_I}} \right\}$$

(Cox 1975)

Έπειτα, η λογαριθμοποιημένη συνάρτηση Πιθανοφάνειας θα έχει τη μορφή:

$$l(\beta) = \sum_{j=1}^{k} \beta' x_j - \sum_{j=1}^{k} \ln \left\{ \sum_{i \in R_j} e^{\beta' x_I} \right\}$$

Παραγωγίζοντας τη συνάρτηση $l(\beta)$ προκύπτουν οι μερικές παράγωγοι πρώτης τάξης:

$$\frac{\partial l(\beta)}{\partial \beta_r} = \sum_{j=1}^{k} x_{j,r} - \sum_{j=1}^{k} \left\{ \frac{\sum_{i \in R_j} x_{I,r} e^{\beta' x_I}}{\sum_{i \in R_j} e^{\beta' x_I}} \right\}$$

Λύνοντας το σύστημα εξισώσεων που προκύπτει για $r = 1, \dots, p$, όπου p ο αριθμός των παραμέτρων προς εκτίμηση, θα βρούμε τις εκτιμήσεις της μέγιστης Πιθανοφάνειας $\hat{\beta}_r$. Παρατηρούμε ότι σε όλη τη διαδικασία δεν χρησιμοποιήθηκε η συνάρτηση $h_0(t)$, το οποίο μας εξηγεί τον ορισμό του ημι-παραμετρικού μοντέλου. (Cox 1972)

Τέλος, πολλές φορές θέλουμε να υπολογίσουμε τα τυπικά σφάλματα των εκτιμήσεων $\mathrm{se}(\hat{\beta}_r)$, $\mathrm{r} = 1, \dots, \mathrm{p}$, τα οποία θα βρούμε από τις εκτιμήσεις των διασπορών, $\mathrm{se}(\hat{\beta}_r) = \sqrt{\widehat{\mathbb{V}}(\hat{\beta}_r)}$. Οι εκτιμήσεις των διασπορών $\widehat{\mathbb{V}}(\hat{\beta}_r)$ υπολογίζονται από τον αντίστροφο του πίνακα παρατηρούμενης πληροφορίας με το (r,s) στοιχείο του, δηλαδή με το στοιχείο $-\frac{\partial^2 l(\beta)}{\partial \beta_r \partial \beta_s} \mathrm{I}_{\hat{\beta}}$. (Καρώνη 2009)

2.4.2 Ισόπαλοι χρόνοι διακοπής

Στην Παράγραφο 2.4.1 χρησιμοποιήσαμε την υπόθεση ότι σε κάθε διακεκριμένη χρονική στιγμή $\mathrm{t}_{(\mathrm{j})}$, $\mathrm{j} = 1, \dots, \mathrm{k}$, σταματά να λειτουργεί μόνο μία μονάδα, δηλαδή ότι $\mathrm{d}_\mathrm{j} = 1$, για κάθε j. Στις περιπτώσεις όπου οι χρόνοι διακοπής συμπίπτουν τότε η συνάρτηση Πιθανοφάνειας που υπολογίσαμε θα αλλάξει μορφή. Το βασικό πρόβλημα που αντιμετωπίζουμε είναι ότι τη χρονική στιγμή $\mathrm{t}_{(\mathrm{j})}$ το πλήθος των μονάδων $\mathrm{d}_\mathrm{j} > 1$ στις οποίες συνέβη το γεγονός, θεωρητικά προέκυψαν από διαφορετικούς χρόνους, δηλαδή πιθανών να πρόκειται για στρογγυλοποιημένες παρατηρήσεις με λιγότερα σημαντικά ψηφία. Αυτό όμως σημαίνει ότι θα μπορούσαν να είχαν προκύψει με οποιαδήποτε σειρά μεταξύ τους και άρα θα υπήρχαν $\mathrm{d}_\mathrm{j}!$ Πιθανές σειρές εμφάνισης. Για να αποφύγουμε τη δημιουργία μίας πολύπλοκης συνάρτησης Πιθανοφάνειας, θα χρησιμοποιήσουμε την προσέγγιση του Breslow, η οποία ορίζεται ως:

$$L_{Breslow} = \prod_{\mathrm{j}=1}^{\mathrm{k}} \left\{ \frac{\mathrm{e}^{\beta' z_\mathrm{j}}}{\left[\sum_{i \in \mathrm{R_j}} \mathrm{e}^{\beta' \mathrm{x_I}} \right]^{\mathrm{d_j}}} \right\}$$

όπου ορίσαμε ως $z_\mathrm{j} = \sum_{\mathrm{u}=1}^{\mathrm{d_j}} \mathrm{x_u}$, όπου $\mathrm{x_u}$ το διάνυσμα συμμεταβλητών της μονάδας u, στην οποία συμβαίνει το γεγονός τη χρονική στιγμή $\mathrm{t}_{(\mathrm{j})}$ και u = 1,…, $\mathrm{d_j}$. Όταν η ποσότητα $\mathrm{d_j}/\mathrm{n_j}$ είναι μικρή, η προσέγγιση του Breslow θεωρείται ακριβής, όπου $\mathrm{n_j}$ ο αριθμός των μονάδων που βρίσκονται σε κίνδυνο αμέσως μετά τη χρονική στιγμή $\mathrm{t}_{(\mathrm{j})}$.

Στην περίπτωση που η ποσότητα $\mathrm{d_j}/\mathrm{n_j}$ δεν είναι μικρή, χρησιμοποιούμε την προσέγγιση του Cox, σύμφωνα με την οποία δεχόμαστε ότι τα δεδομένα μας παρατηρήθηκαν σε διακριτή αντί σε συνεχή κλίμακα. Δεδομένου ότι την χρονική στιγμή $\mathrm{t}_{(\mathrm{j})}$ συμβαίνουν $\mathrm{d_j}$ διακοπές λειτουργίας, η πιθανότητα να προκύψει ένα οποιοδήποτε σύνολο u αποτελούμενο από $\mathrm{d_j}$ μονάδες, είναι:

$$\mathrm{P(u)} \propto \exp(\beta' \mathrm{z_u})$$

όπου $z_\mathrm{j} = \sum_{\mathrm{u}=1}^{\mathrm{d_j}} \mathrm{x_u}$, και $\mathrm{x_u}$ το διάνυσμα συμμεταβλητών της μονάδας u (όπως στην προσέγγιση του Breslow). Τότε, η υπό συνθήκη πιθανότητα του παρατηρούμενου συνόλου μονάδων u^* με διακοπή δίνεται από τον τύπο:

$$\mathrm{P}(\mathrm{u}^* | \, \mathrm{d_j}) = \frac{\mathrm{e}^{\beta' z_\mathrm{j}}}{\sum_{u \in \mathrm{R_j}} \mathrm{e}^{\beta' \mathrm{z_u}}}$$

Ο παρονομαστής του παραπάνω κλάσματος αποτελείται από το άθροισμα όλων των δυνατών $\binom{\mathrm{n_j}}{\mathrm{d_j}}$ όρων. Και σε αυτή την περίπτωση η συνάρτηση Πιθανοφάνειας θα γίνει:

$$L_{Cox} = \prod_{j=1}^{k} \left\{ \frac{e^{\beta' z_j}}{\sum_{u \in R_j} e^{\beta' z_u}} \right\}$$

(Καρώνη 2009)

2.4.3 Επεκτάσεις του μοντέλου Cox

Μία επέκταση του κλασικού μοντέλου του Cox είναι η Στρωματοποιημένη ανάλυση (Stratified Cox model). Επιτρέπει τον έλεγχο με «στρωματοποίηση» ενός προγνωστικού που δεν ικανοποιεί την υπόθεση αναλογικής διακινδύνευσης. Με άλλα λόγια, υπάρχουν περιπτώσεις που η υπόθεση της αναλογικής διακινδύνευσης δεν ισχύει στο σύνολο των δεδομένων μας αλλά σε διαφορετικά υποσύνολα των δεδομένων. Αυτό συμβαίνει όταν κάποιες μεταξύ κάποιων μεταβλητών δεν υπάρχει αναλογία. Για παράδειγμα, αν υποθέσουμε ότι η κατηγορική μεταβλητή μας είναι αν έχουν σάκχαρο οι ασθενείς που μελετάμε τότε χωρίζουμε τη μεταβλητή σάκχαρο σε δύο στρώματα, σε αυτούς που έχουν σάκχαρο και σε αυτούς που δεν έχουν. Σε αυτή τη περίπτωση θα έχουμε διαφορετικές συναρτήσεις διακινδύνευσης για τα δύο στρώματα, δηλαδή θα ορίσουμε:

$$h(t; x) = \begin{cases} \exp(\beta' x) h_{01}(t), & \text{με σάκχαρο} \\ \exp(\beta' x) h_{02}(t), & \text{χωρίς σάκχαρο} \end{cases}$$

όπου $h_{01}(t)$ και $h_{02}(t)$ οι βασικές συναρτήσεις διακινδύνευσης για αυτούς που έχουν σάκχαρο και αυτούς που δεν έχουν αντίστοιχα, οι οποίες όμως δεν βρίσκονται σε αναλογία μεταξύ τους.

Για την εκτίμηση των παραμέτρων του μοντέλου χρησιμοποιούμε τη μέθοδο της μεγίστης Πιθανοφάνειας και έτσι για κάθε στρώμα m έχουμε ότι ο λογάριθμος της μερικής Πιθανοφάνειας είναι:

$$l_m(\beta) = \sum_{j=1}^{k_m} \beta' x_{mj} - \sum_{j=1}^{k_m} \ln \left\{ \sum_{i \in R_{mj}} e^{\beta' x_{mi}} \right\}$$

δηλαδή είναι ο ίδιος τύπος που αναφέραμε στο κλασικό μοντέλο του Cox με τη διαφορά ότι προσθέτουμε το δείκτη m για να τονίσουμε ότι αναφερόμαστε σε στρώματα. Συνολικά για όλα τα στρώματα m=1,...,p έχουμε:

$$l(\beta) = \sum_{m=1}^{p} l_m(\beta)$$

Μετά συνεχίζουμε τη διαδικασία που είχαμε εφαρμόσει στην Παράγραφο 2.4.1 για την εκτίμηση των παραμέτρων που έχουμε, λαμβάνοντας υπόψη τα στρώματα που δημιουργήσαμε. (Καρώνη 2009)

2.4.4 Έλεγχος της υπόθεσης αναλογικής διακινδύνευσης στο μοντέλο του Cox

Είναι σημαντικό να ελέγξουμε αν το μοντέλο αναλογικής διακινδύνευσης του Cox είναι το κατάλληλο ή όχι, δηλαδή αν ισχύει η υπόθεση της αναλογικής διακινδύνευσης. Όπως είδαμε στην Παράγραφο 2.3 η ιδιότητα της αναλογικής διακινδύνευσης υποστηρίζει ότι ο λόγος:

$$\frac{h(t;x_i)}{h(t;x_j)} = \frac{\exp(\beta' x_i)h_0(t)}{\exp(\beta' x_j)h_0(t)} = \exp[\beta'(x_i - x_j)]$$

είναι ανεξάρτητος του χρόνου t, για δύο μονάδες i και j. Για να ελέγξουμε αν αυτή η ιδιότητα ισχύει στο μοντέλο του Cox υπάρχουν δύο τρόποι.

- Ο πρώτος είναι ξεχωριστά για κάθε συμμεταβλητή x_i να ορίσουμε μία καινούρια μεταβλητή $z = x_i t$ (ή οποιαδήποτε άλλη συνάρτηση του χρόνου), να προσαρμόσουμε το μοντέλο του Cox συμπεριλαμβανομένης και της μεταβλητής z και τέλος να ελέγξουμε αν ισχύει η μηδενική υπόθεση H_0: $\beta_z = 0$. Αν δεχτούμε τη μηδενική υπόθεση σημαίνει ότι ισχύει η ιδιότητα της αναλογικής διακινδύνευσης.
- Ο δεύτερος τρόπος είναι μέσω του γραφικού ελέγχου της υπόθεσης της αναλογικής διακινδύνευσης που είδαμε στην Παράγραφο 2.3. Θα σχηματίσουμε τις καμπύλες των $\ln\{-\ln S(t;x)\}$ συνάρτήσει του λογαριθμισμένου χρόνου $\ln t$, οι οποίες πρέπει να είναι παράλληλες μεταξύ τους για να ισχύει η υπόθεση της αναλογικής διακινδύνευσης. Η διαφορά είναι ότι στο ημι-παραμετρικό μοντέλο του Cox δεν γνωρίζουμε τη βασική συνάρτηση επιβίωσης $S_0(t;x)$.

  $$S(t;x) = \{S_0(t;x)\}^{e^{\beta' x}}$$

  Για αυτό θα εκτιμήσουμε τη συνάρτηση $S_0(t)$ μέσω της μη-παραμετρικής εκτιμήτριας του Breslow που ορίζεται ως:

  $$\hat{S}_0(t) = \exp\{-\hat{H}_0(t)\}, \quad \text{όπου} \quad \hat{H}_0(t) = \sum_{t_j \le t}\left(\frac{d_j}{\sum_{i \in R_j} e^{\hat{\beta}' x_I}}\right)$$

  Λαμβάνοντας υπόψη όλες τις σημαντικές συμμεταβλητές ενός στρωματοποιημένου μοντέλου του Cox. Οπότε για κάθε στρώμα m θα εκτιμάμε τη βασική συνάρτηση επιβίωσης $S_0(t)$, δηλαδή θα υπολογίζουμε τη συνάρτηση:

  $$\hat{S}_m(t) = \hat{S}_{0m}(t)e^{\hat{\beta}' \bar{x}_m}$$

  όπου $\bar{x}_m$ το διάνυσμα των μέσων τιμών των συμμεταβλητών στο στρώμα m. Δηλαδή θα κάνουμε τις καμπύλες των $\ln\{-\ln \hat{S}_m(t)\}$, m=1,…,p συναρτήσει του λογαριθμισμένου χρόνου $\ln t$ και εάν είναι παράλληλες μεταξύ τους τότε ισχύει η υπόθεση της αναλογικής διακινδύνευσης.

(Καρώνη 2009)

2.4.5 Υπόλοιπα Schoenfeld

Τα υπόλοιπα που μελετάμε κυρίως στο ημι-παραμετρικό μοντέλο του Cox είναι τα λεγόμενα υπόλοιπα Schoenfeld (1982) ή αλλιώς τα μερικά υπόλοιπα (partial residuals). Ο λόγος που δεν χρησιμοποιούμε τα υπόλοιπα Cox-Snell (Παράγραφος 2.2.2) είναι γιατί στο ημι-παραμετρικό μοντέλο του Cox παρουσιάζουν ένα σημαντικό πρόβλημα, απαιτούν την εκτίμηση της βασικής σωρευτικής συνάρτησης διακινδύνευσης $\hat{H}_0(t)$, διότι όπως ορίσαμε στην Παράγραφο 2.2.2 τα υπόλοιπα Cox-Snell ορίζονται ως:

$$-\ln \hat{S}(t_I;x_i) = \hat{H}(t_I;x_i) = \hat{H}_0(t_I;x_i)e^{\beta' x_I}$$

Το γεγονός αυτό καθιστά τα υπόλοιπα Cox-Snell λιγότερο χρήσιμα για το ημι-παραμετρικό μοντέλο του Cox.

Αντίθετα τα υπόλοιπα Schoenfeld διαφέρουν από τα Cox-Snell διότι δεν χρησιμοποιούνται οι τιμές της εξαρτημένης μεταβλητής του χρόνου t αλλά τα αντίστοιχα διανύσματα συμμεταβλητών x_j,

$j = 1, \dots p$. Οπότε μας δίνουν ένα σύνολο τιμών για κάθε μία συμμεταβλητή x_j, $j = 1, \dots p$ που υπάρχει στο μοντέλο του Cox που έχουμε προσαρμόσει. Το στοιχείο i των υπολοίπων Schoenfeld για τη συμμεταβλητή j ορίζεται ως:

$$r_{p_{ji}} = \delta_i\{x_{ji} - \hat{\alpha}_{ji}\}$$

όπου x_{ji} είναι η τιμή i για τη j συμμεταβλητή (j=1,…,p) και το $\hat{\alpha}_{ji}$ ορίζεται ως:

$$\hat{\alpha}_{ji} = \frac{\sum_{k \in R(t_i)} x_{jk} \exp(\hat{\beta}' x_k)}{\sum_{k \in R(t_i)} \exp(\hat{\beta}' x_k)}$$

όπου $R(t_i)$ είναι το σύνολο των μονάδων που βρίσκονται σε κίνδυνο αμέσως πριν τη χρονική στιγμή t_i. Επιπλέον, τα υπόλοιπα Schoenfeld υπολογίζονται με βάση τους χρόνους διακοπής και όχι με τις αποκομμένες παρατηρήσεις, οι οποίες προφανώς λαμβάνονται και αυτές υπόψη.

Σημειώνεται ότι 25 οi στοιχείο των υπολοίπων Schoenfeld για τη συμμεταβλητή j είναι η εκτίμηση του στοιχείου i της πρώτης παραγώγου της λογαριθμισμένης συνάρτησης Πιθανοφάνειας, που όπως είδαμε στην Παράγραφο 2.4.1 έχει τη μορφή:

$$\frac{\partial l(\beta)}{\partial \beta_r} = \sum_{j=1}^{k} x_{j,r} - \sum_{j=1}^{k} \left\{ \frac{\sum_{i \in R_j} x_{I,r} e^{\beta' x_I}}{\sum_{i \in R_j} e^{\beta' x_I}} \right\}$$

Συνεπώς:

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \delta_i\{x_{ji} - \alpha_{ji}\}$$

όπου τα α_{ji} ορίζονται όπως και πριν με τη διαφορά ότι περιλαμβάνουν τους συντελεστές β και όχι τις εκτιμήσεις τους $\hat{\beta}$, δηλαδή:

$$\alpha_{ji} = \frac{\sum_{m \in R(t_i)} x_{jm} \exp(\beta' x_m)}{\sum_{m \in R(t_i)} \exp(\beta' x_m)}$$

Για να εκτιμήσουμε τους συντελεστές β πρέπει να θέσουμε την παραπάνω παράγωγο ίση με το μηδέν για τις τιμές των εκτιμήσεων $\hat{\beta}$, δηλαδή:

$$\left.\frac{\partial l(\beta)}{\partial \beta_j}\right|_{\hat{\beta}} = 0 \quad \Rightarrow \quad \delta_i\{x_{ji} - \hat{\alpha}_{ji}\} = 0$$

Οπότε σε μεγάλα δείγματα οι τιμές των υπολοίπων Schoenfeld $r_{p_{ji}}$ είναι ίσες με το μηδέν και επίσης δεν σχετίζονται μεταξύ τους. Έτσι δημιουργήθηκαν τα κλιμακοποιημένα (scaled) υπόλοιπα Schoenfeld τα οποία είναι πιο χρήσιμα στο μοντέλο του Cox και ορίζονται ως:

$$r^*_{p_i} = k\hat{V}(\hat{\beta})\hat{r}_{p_i}$$

(Anwar 2013)

όπου $\hat{V}(\hat{\beta})$ ο εκτιμημένος πίνακας διασποράς των $\hat{\beta}$ και k είναι ο αριθμός των μονάδων που συνέβη το γεγονός από το σύνολο n των μονάδων που έχουμε, δηλαδή ο αριθμός των μη-αποκομμένων παρατηρήσεων. (Collett 2003)

Γραφικός έλεγχος:

Τα κλιμακοποιημένα υπόλοιπα Schoenfeld μπορούν να χρησιμοποιηθούν για τον έλεγχο της υπόθεσης της αναλογικής διακινδύνευσης (Παράγραφος 2.4.4). Η ιδιότητα της αναλογικής διακινδύνευσης για το μοντέλο του Cox μπορεί να διατυπωθεί ισοδύναμα και ως:

$$\beta_i(t) = \beta_I, \forall t$$

Αποδεικνύεται ότι: $E(r^*_{ij}) \cong \beta_i(t_{(j)}) - \hat{\beta}_i$, όπου $\beta_i(t_{(j)})$ είναι ο συντελεστής της i-οστής συμμεταβλητής του μοντέλου τη χρονική στιγμή $t_{(j)}$. Συνεπώς, πρέπει να ισχύει $E(r^*_{ij}) = 0$, όπου αυτό μας οδηγεί σε μία σχέση μεταξύ των υπολοίπων $r^*_{p_i}$ και του χρόνου t. (Sunhee & David 2015)

Επομένως, για να αποδεχτούμε την παραπάνω υπόθεση ($\beta_i(t) = \beta_I, \forall t$) θα πρέπει η ακόλουθη γραφική παράσταση να είναι μία οριζόντια γραμμή, δηλαδή να υπάρχει ανεξαρτησία των τιμών της ως προς το χρόνο t.

$$R^*_{ij} + \hat{\beta}_i \quad \text{ως προς το χρόνο } t_{(j)}$$

Θα εφαρμόσουμε αυτό τον γραφικό έλεγχο στην εφαρμογή που θα μελετήσουμε στο Κεφάλαιο 5 (Παράγραφος 5.9.2). (Καρώνη 2009)

2.4.6 Καμπύλες ROC & AUC

Οι καμπύλες ROC (Receiver Operating Characteristic Curves) μας βοηθούν στο να εξετάσουμε την ικανότητα του μοντέλου που έχουμε να στο να προβλέπει νέα δεδομένα, δηλαδή χαρακτηρίζουν την προγνωστική ακρίβεια του μοντέλου μας. Είναι πιθανό, ένα προσαρμοσμένο μοντέλο να έχει άψογη ικανότητα περιγραφής του υπάρχοντος συνόλου δεδομένων που έχουμε στα χέρια μας αλλά να μην μπορεί να μας δώσει αξιόπιστες προβλέψεις για δεδομένα εκτός αυτού. Η αξιολόγηση της προβλεπτικής ικανότητας του μοντέλου μας μπορεί να γίνει γραφικά μέσω των καμπύλων ROC.

Έστω δίτιμη τ.μ Y, η οποία δέχεται τις τιμές Y=1 («επιτυχία») και Y=0 («αποτυχία») και έστω επίσης

$$\hat{p} = \hat{P}[Y = 1] = \frac{e^{x'\hat{\beta}}}{1 + e^{x'\hat{\beta}}}$$

η εκτιμημένη πιθανότητα επιτυχίας. Τέλος έστω μια σταθερά p_0, τότε μπορούμε να διακρίνουμε τις περιπτώσεις:

$$\begin{cases} \text{αν } \hat{p} > p_0, \text{τότε προβλέπεται } Y = 1 \\ \text{αν } \hat{p} \leq p_0, \text{τότε προβλέπεται } Y = 0 \end{cases}$$

Πίνακας 2.1

		Πραγματική Κατάσταση		
		Y=1	Y=0	∑
Πρόβλεψη	Y=1	α	b	α + b
	Y=0	c	d	c + d
	∑	α + c	b + d	n

Ο Πίνακας 2.1 είναι ένας πίνακας συνάφειας όπου συγκρίνει τις προβλέψεις των μονάδων (Y=1 ή Y=0) με την πραγματική κατάσταση στην οποία βρίσκονται. Για παράδειγμα, το δεύτερο κελί μας δείχνει ότι b μονάδες, από το σύνολο των μονάδων n που έχουμε, έχουν πραγματική τιμή Y=0

(«αποτυχία») και η πρόβλεψη για αυτές είναι Y=1 («επιτυχία»). Με βάση τον Πίνακα 2.1 θα ορίσουμε δύο βασικές έννοιες:

- Ευαισθησία (sensitivity) ή αλλιώς το ποσοστό των σωστών θετικών αποτελεσμάτων (true positive rate, TPR) ορίζεται ως τος ποσοστό του πληθυσμού (n) που είχε σωστή πρόβλεψη της κατάστασης Y=1, δηλαδή:

$$\mathrm{TPR} = \frac{\alpha}{\alpha + c}$$

- Ειδικότητα (specificity) ή αλλιώς το ποσοστό των αληθώς αρνητικών αποτελεσμάτων (true negative rate, TNR) ορίζεται ως τος ποσοστό του πληθυσμού (n) που είχε σωστή πρόβλεψη της κατάστασης Y=0, δηλαδή:

$$\mathrm{TNR} = \frac{d}{b + d}$$

(Narkhede 2018)

Τα ποσοστά αυτά, καθώς και τα συμπληρωματικά τους, δηλαδή το ποσοστό ψευδώς αρνητικών (false negative rate, FNR) και ψευδώς θετικών αποτελεσμάτων (false positive rate, FPR) ονομάζονται λειτουργικά χαρακτηριστικά της διαγνωστικής δοκιμασίας. Αντίστοιχα λοιπόν έχουμε:

$$1\text{- Specificity} = \mathrm{FPR} = \frac{b}{b + d} \quad \text{και} \quad 1\text{- Sensitivity} = \mathrm{FNR} = \frac{c}{\alpha + c}$$

Αν, λοιπόν, υπολογιστούν οι τιμές της ευαισθησίας και της ειδικότητας για κάθε p_0 στο εύρος [0,1], μπορεί να σχηματιστεί η χαρακτηριστική καμπύλη ROC η οποία απεικονίζει την προβλεπτική ικανότητα του μοντέλου καθώς το όριο p_0 μεταβάλλεται.

Στον άξονα x, ενός τέτοιου γραφήματος, έχουμε την ποσότητα 1-Specificity και στον άξονα y έχουμε την Sensitivity. Σχεδιάζουμε την καμπύλη για τις διάφορες τιμές του και επιπλέον την ευθεία για την οποία ισχύει TPR=FPR και χωρίζει τη γραφική παράσταση σε δύο χωρία με ίσο εμβαδόν. Ένα μοντέλο θα έχει μεγάλη προβλεπτική ικανότητα αν για πολύ μικρές τιμές της ποσότητας 1-Specificity το Sensitivity πλησιάζει την τιμή 1, δηλαδή την πάνω αριστερή γωνία του γραφήματος.

Ένας δείκτης που μετρά αυτή την επιθυμητή ιδιότητα ενός μοντέλου είναι το εμβαδόν κάτω από την καμπύλη, το οποίο ορίζεται ως AUC (area under the curve). Η καμπύλη ROC δεν μπορεί να βρεθεί ποτέ κάτω από την ευθεία TPR=FPR που διχοτομεί το χωρίο, οπότε το AUC λαμβάνει τιμές από 0.5 έως και 1 (εφόσον το εμβαδόν κάτω από την ευθεία είναι πάντα 0.5). Όσο η τιμή του AUC προσεγγίζει το 1 τόσο καλύτερη προβλεπτική ικανότητα έχει το μοντέλο μας. Ενώ εάν η τιμή του AUC λάμβανε τη τιμή 0.5 (δηλαδή πλησίαζε αρκετά την ευθεία TPR=FPR) τότε τα ποσοστά των σωστών θετικών και των ψευδών θετικών αποτελεσμάτων θα ήταν ίσα. Θα εφαρμόσουμε τις καμπύλες ROC και AUC στην εφαρμογή που θα μελετήσουμε στο Κεφάλαιο 5 (Παράγραφος 5.9.2). (Heagerty & Zheng, 2005).

2.5 Κριτήρια επιλογής μεταβλητών και μέτρα καλής προσαρμογής

Έχοντας προσαρμόσει το μοντέλο που έχουμε επιλέξει και έχοντας εκτιμήσει τις παραμέτρους του, σημαντικό ρόλο για τη συνέχεια είναι να μπορούμε να εκτελούμε ελέγχους υποθέσεων που αφορούν τη σημαντικότητα των παραμέτρων που συμπεριλαμβάνονται στο εκτιμημένο μοντέλο. Για τη σημαντικότητα των παραμέτρων χρησιμοποιούμε τον έλεγχο Wald (Wald test), τον έλεγχο του λόγου των πιθανοφανειών (Likelihood ratio test), το κριτήριο AIC και τέλος τη μέθοδο της διαδοχικής

αφαίρεσης (backward elimination). Και οι τέσσερις αυτοί έλεγχοι μπορούν να εφαρμοστούν σε όποιο μοντέλο θέλουμε (π.χ. Επιταχυνόμενης Διακοπής, Cox, κ.τ.λ.) και δεν έχουν κάποιον περιορισμό.

Έλεγχος Wald:

Ο έλεγχος Wald απαιτεί την προσαρμογή ενός μόνο μοντέλου, όταν υπάρχει μόνο μία παράμετρος ελέγχει την υπόθεση: $H_0: \theta = \theta_0, \quad H_1: \theta \neq \theta_0$ και η γενική μορφή του ελέγχου είναι η ακόλουθη:

$$z = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})} \sim N(0,1)$$

όπου θ είναι η παράμετρος που έχει το μοντέλο μας και $se(\hat{\theta})$ είναι η τυπική απόκλιση της παραμέτρου θ. Ισοδύναμα ισχύει: $z^2 \sim X_1^2$

Ο έλεγχος Wald είναι επίσης χρήσιμος για να την επιλογή του καλύτερου μοντέλου ανάμεσα από τη Weibull και την Εκθετική κατανομή. Δηλαδή ελέγχει τη μηδενική υπόθεση (H_0) ότι οι παρατηρήσεις που έχουμε ακολουθούν την Εκθετική κατανομή έναντι της εναλλακτικής (H_1) ότι ακολουθούν την κατανομή Weibull. Υπενθυμίζουμε ότι η Εκθετική κατανομή προκύπτει από την Weibull αν θέσουμε η=1. Οπότε ο έλεγχος γίνεται:

$$H_0: \eta = 1, \quad H_1: \eta \neq 1$$

$$z = \frac{\hat{\eta} - 1}{se(\hat{\eta})} \sim N(0,1)$$

όπου η είναι η παράμετρος σχήματος (Shape) της κατανομής Weibull.

Έχοντας προσαρμόσει το μοντέλο που έχουμε επιλέξει θα πρέπει να είμαστε σε θέση να ελέγξουμε την σημαντικότητα της κάθε μεταβλητής ξεχωριστά που περιέχεται στο μοντέλο. Αφού πρώτα υπολογίσουμε τις εκτιμήσεις $\hat{\beta}_I$, i=1,…,k και τα τυπικά τους σφάλματα $se(\hat{\beta}_i)$, χρησιμοποιούμε την ελεγχοσυνάρτηση:

$$z = \frac{\hat{\beta}_i - \beta_i|_{H_0}}{se(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)} \sim N(0,1)$$

Οπότε ορίζοντας ένα επίπεδο σημαντικότητας α (συνήθως 5%) μπορούμε να ελέγξουμε την σημαντικότητα των μεταβλητών του μοντέλου. Αν η p-value είναι αρκετά μικρή (συνήθως μικρότερη του 0.05) έχουμε ισχυρές ενδείξεις ώστε να απορρίψουμε τη μηδενική υπόθεση. Αν απορριφθεί η H_0 τότε η i-οστή συμμεταβλητή είναι στατιστικά σημαντική και παραμένει στο μοντέλο που έχουμε επιλέξει.

Στην περίπτωση που έχουμε πολλές συμμεταβλητές, όπως στα μοντέλα που έχουμε περιγράψει σε αυτό το κεφάλαιο, ο έλεγχος Wald ελέγχει τη μηδενική υπόθεση ότι όλες οι συμμεταβλητές είναι μηδενικές, έναντι της εναλλακτικής ότι τουλάχιστον μία συμμεταβλητή είναι διάφορη του μηδενός. Δηλαδή ο έλεγχος υποθέσεων γίνεται: $H_0: \beta_i = 0, \ \forall i = 1,..,k$ και $H_1: \beta_i \neq 0$ για κάποιο $i = 1,..,k,$ όπου k ο αριθμός των συμμεταβλητών που έχουμε. (Καρώνη 2009)

Έλεγχος του λόγου των Πιθανοφανειών:

Ο έλεγχος του λόγου των Πιθανοφανειών (Likelihood ratio test) απαιτεί περισσότερους υπολογισμούς σε σχέση με τον έλεγχο Wald, αλλά γενικώς είναι προτιμότερος. Αυτός ο έλεγχος ελέγχει την υπόθεση: $\mathrm{H}_0\colon\ \theta = \theta_0,\quad \mathrm{H}_1\colon\ \theta \neq \theta_0$, και έχει τη μορφή:

$$\mathrm{z} = -2\left(\hat{l}_0 - \hat{l}_1\right) \sim \mathrm{X}^2_{\mathrm{d}}$$

όπου $\hat{l}_0$ η λογαριθμισμένη συνάρτηση Πιθανοφάνειας της υπόθεσης H_0 δηλαδή προσαρμόζουμε το μοντέλο που έχουμε με την παράμετρο θ_0, και αντίστοιχα $\hat{l}_1$ της H_1. Επίσης ως d ορίζουμε τη διαφορά των παραμέτρων που έχουμε προς εκτίμηση. Και αυτός ο έλεγχος (όπως και ο Wald) έχει χρησιμότητα στον έλεγχο της Εκθετικής και της Weibull κατανομής, δηλαδή στον: $\mathrm{H}_0\colon\ \eta = 1,\quad \mathrm{H}_1\colon\ \eta \neq 1$

Τότε η λογαριθμισμένη συνάρτηση Πιθανοφάνειας $\hat{l}_0$ είναι της Εκθετικής κατανομής και η $\hat{l}_1$ της Weibull, και σε αυτή την περίπτωση θα έχουμε d=1.

Στην περίπτωση που έχουμε την υπόθεση: $\mathrm{H}_0\colon\ \beta_i = 0,\ \forall\, \mathrm{i} = 1,..,\mathrm{k}$ και $\mathrm{H}_1\colon\ \beta_i \neq 0$ για κάποιο $\mathrm{i} = 1,..,\mathrm{k}$, (δηλαδή έχουμε πολλές συμμεταβλητές) θα προσαρμόσουμε αρχικά το μοντέλο χωρίς τις συμμεταβλητές $(\beta_i = 0)$ και έπειτα υπολογίζουμε τη μεγιστοποιημένη τιμή του λογαρίθμου της Πιθανοφάνειας $(\hat{l}_0)$. Εργαζόμαστε το ίδιο και για το μοντέλο που περιέχει μία συμμεταβλητή β_i για κάποιο $\mathrm{i} = 1,..,\mathrm{k}$ και υπολογίζουμε πάλι τη μεγιστοποιημένη τιμή του λογαρίθμου της Πιθανοφάνειας $(\hat{l}_1)$. Και εδώ η παράμετρος d που ορίσαμε θα είναι ίση με 1. Αν η p-value είναι αρκετά μικρή (συνήθως μικρότερη του 0.05) έχουμε ισχυρές ενδείξεις ώστε να απορρίψουμε τη μηδενική υπόθεση, και άρα η μεταβλητή β_i για κάποιο $\mathrm{i} = 1,..,\mathrm{k}$ είναι στατιστικά σημαντική και παραμένει στο μοντέλο μας. (Καρώνη 2009)

Κριτήριο AIC:

Το κριτήριο AIC (Akaike's information criterion) μας βοηθάει να επιλέξουμε το καλύτερο μοντέλο με όσο το δυνατόν μικρότερο αριθμό συμμεταβλητών. Αυτό το κριτήριο ορίζεται ως:

$$\mathrm{AIC} = -2l + 2\mathrm{k}$$

όπου k είναι το πλήθος των παραμέτρων του μοντέλου και l είναι η μεγιστοποιημένη τιμή της λογαριθμισμένης συνάρτησης Πιθανοφάνειας για το εκτιμημένο μοντέλο. Συγκρίνοντας όλα τα υποψήφια μοντέλα, επιλέγουμε ως καλύτερο εκείνο με την μικρότερη τιμή του κριτηρίου AIC.

Γνωρίζουμε ότι η προσαρμογή του μοντέλου βελτιώνεται όσο εισάγουμε επιπλέον μεταβλητές σε αυτό. Όμως, το κριτήριο AIC δεν προσμετρά μόνο την προσαρμογή, αλλά περιλαμβάνει και ένα είδος ποινής η οποία είναι μια αύξουσα συνάρτηση του αριθμού των παραμέτρων του εκάστοτε μοντέλου. Άρα εισάγοντας μια νέα μεταβλητή στο μοντέλο, ναι μεν βελτιώνεται η προσαρμογή άρα αυξάνεται η πιθανοφάνεια l, άρα μειώνεται ο πρώτος όρος του AIC αλλά από την άλλη ο παράγοντας ποινής k αυξάνεται, άρα αυξάνεται ταυτόχρονα και ο δεύτερος όρος του AIC. Επομένως, η εισαγωγή μιας επιπλέον μεταβλητής μπορεί να οδηγήσει σε ένα καλύτερο μοντέλο (μικρότερο AIC) μόνο αν βελτιώνει την προσαρμογή του τόσο ώστε να ξεπεράσει την ποινή του όρου 2k.
(Καρώνη & Οικονόμου, 2017)

Μέθοδος της Διαδοχικής Αφαίρεσης:

Η μέθοδος της διαδοχικής αφαίρεσης (backward elimination), βασίζεται στο κριτήριο AIC και ουσιαστικά αφαιρεί διαδοχικά μεταβλητές από το μοντέλο μας. Αυτή η μέθοδος περιγράφεται από τα ακόλουθα βήματα:

1) Προσαρμόζουμε το μοντέλο που θέλουμε, με όλες τις διαθέσιμες μεταβλητές.
2) Αφαιρούμε τη λιγότερο στατιστικά σημαντική μεταβλητή, δηλαδή εκείνη που συμβάλει λιγότερο στο μοντέλο μας χρησιμοποιώντας το κριτήριο AIC (δηλαδή αφαιρούμε εκείνη με το μεγαλύτερο AIC). Καθορίζουμε, δηλαδή, ποια είναι η μεταβλητή η οποία αν απαλειφθεί από το μοντέλο θα μας δώσει τη μικρότερη αύξηση του κριτηρίου AIC αν επαναπροσαρμόσουμε το μοντέλο χωρίς αυτή.
3) Προσαρμόζουμε ξανά το μοντέλο που περιέχει όλες τις μεταβλητές εκτός από αυτή που αφαιρέσαμε στο προηγούμενο βήμα, και εκτελούμε ξανά τον ίδιο έλεγχο εντοπισμού της λιγότερο σημαντικής μεταβλητής χρησιμοποιώντας το κριτήριο AIC.
4) Επαναλαμβάνουμε τα δύο προηγούμενα βήματα μέχρι να καταλήξουμε σε ένα μοντέλο όπου η αφαίρεση οποιασδήποτε μεταβλητής αυξάνει την τιμή του AIC, οπότε και σταματάμε.

Με αυτό το τρόπο καταλήγουμε σε ένα μοντέλο που περιέχει όσο το δυνατόν λιγότερες μεταβλητές και έχει το μικρότερο δυνατό AIC. Μία παρόμοια μέθοδος είναι αυτή της διαδοχικής πρόσθεσης (forward selection), όπου εκτελούμε την αντίστροφη διαδικασία ξεκινώντας χωρίς μεταβλητές και προσθέτοντας σε κάθε βήμα μία μεταβλητή η οποία είναι στατιστικά σημαντική για το μοντέλο μας. (Καρώνη & Οικονόμου, 2017)

Κεφάλαιο 3: Μέθοδοι Συρρίκνωσης

3.1 Εισαγωγή

Πολύ συχνά στα δεδομένα που μελετάμε συναντάμε το πρόβλημα της πολυσυγγραμμικότητας. Πολυσυγγραμμικότητα υπάρχει όταν δύο ή περισσότερες ανεξάρτητες μεταβλητές είναι ισχυρά συσχετισμένες μεταξύ τους, και οι τιμές της μίας μπορούν να υπολογιστούν από την άλλη. Η πολυσυγγραμμικότητα δεν έχει επιπτώσεις στην προβλεπτική ικανότητα του μοντέλου, έχει όμως επιπτώσεις στους συντελεστές των ανεξάρτητων μεταβλητών καθώς τα αποτελέσματα που εξάγουμε μπορεί να μην είναι αξιόπιστα. Αυτό σημαίνει ότι οι τιμές των συντελεστών μπορεί να αλλάξουν πολύ, αν προστεθεί ή αφαιρεθεί μια μεταβλητή ή αν συμβούν μικρές μεταβολές στα δεδομένα μας. Η λύση για την πολυσυγγραμμικότητα είναι η επιλογή ανεξάρτητων μεταβλητών, δηλαδή την επιλογή ανεξάρτητων μεταβλητών που παρέχουν πληροφορίες σχετικά με την ακρίβεια της πρόβλεψης. Για να αντιμετωπίσουμε αυτό το πρόβλημα χρησιμοποιούμε μεθόδους συρρίκνωσης (shrinkage methods). (Rusmadi et al. 2017)

Οι μέθοδοι συρρίκνωσης είναι μια τεχνική που περιορίζει ή ρυθμίζει τις εκτιμήσεις των συντελεστών, ή ισοδύναμα, συρρικνώνει τις εκτιμήσεις τους στο μηδέν. Μπορεί να μην είναι αμέσως προφανές γιατί ένας τέτοιος περιορισμός βελτιώνει την προσαρμογή του μοντέλου μας, αλλά αποδεικνύεται ότι συρρίκνωση των εκτιμήσεων του συντελεστή μπορεί να μειώσει σημαντικά τη διακύμανση τους. Οι δύο πιο γνωστές τεχνικές για τη συρρίκνωση των συντελεστών παλινδρόμησης προς το μηδέν είναι η παλινδρόμηση Κορυφογραμμής (Ridge) και η παλινδρόμηση Lasso. (Wang 2019)

3.2 Παλινδρόμηση Κορυφογραμμής (Ridge)

Η Παλινδρόμηση Κορυφογραμμής (Ridge) είναι παρόμοια με τη μέθοδο των Ελαχίστων Τετραγώνων (Ordinary Least Squares ή αλλιώς OLS). Σημειώνεται ότι η μέθοδος των Ελαχίστων Τετραγώνων εκτιμάει τους συντελεστές β_j, j=1,…,p ελαχιστοποιώντας την ποσότητα:

$$\mathrm{RSS} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

όπου RSS (Residual Sum of Squares) είναι το άθροισμα των τετραγώνων των υπολοίπων.

Αντίθετα η μέθοδος Ridge εκτιμάει τους συντελεστές β_j, j=1,…,p ($\hat{\beta}^R$) ελαχιστοποιώντας την ακόλουθη ποσότητα:

$$L_{\mathrm{Ridge}} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2 \qquad (3.1)$$

(James et al. 2013)

όπου n ο αριθμός των παρατηρήσεων που έχουμε και λ ≥ 0 είναι μία παράμετρος η οποία ελέγχει το μέγεθος της συρρίκνωσης (tuning parameter), δηλαδή όσο μεγαλύτερη η τιμή του λ, τόσο μεγαλύτερη η συρρίκνωση. Ο όρος $\lambda\sum_{j=1}^{p}\beta_j^2$ είναι η «ποινή» που εφαρμόζει η μέθοδος Ridge (shrinkage penalty). Αυτή η μέθοδος είναι πολύ χρήσιμη γιατί για κάθε μία διαφορετική τιμή της

παραμέτρου λ, μπορούμε να έχουμε διαφορετικές εκτιμήσεις των β_j. Όταν λ=0, ο όρος της ποινής δεν θα επηρεάζει καθόλου τη σχέση (3.1) και συνεπώς η μέθοδος Ridge θα υπολογίζει τις ίδιες εκτιμήσεις με τη μέθοδο των Ελαχίστων Τετραγώνων (OLS). Αντίθετα όταν η παράμετρος λ τείνει στο άπειρο ($\lambda \to \infty$), οι εκτιμήσεις της μεθόδου Ridge θα τείνουν στο 0, καθώς ο όρος της ποινής θα είναι πολύ μεγάλος. Σημειώνεται ότι όπως φαίνεται και από τη σχέση (3.1), η μέθοδος Ridge δεν εφαρμόζει την ποινή στο σταθερό όρο β_0, αφού όπως βλέπουμε ξεκινάει από j=1. Λύνοντας την εξίσωση (3.1) βρίσκουμε τις εκτιμήσεις των β_j, j=1,…,p της Παλινδρόμησης Κορυφογραμμής:

$$\hat{\beta}^R = (X'X + \lambda I)^{-1}X'y$$

όπου I είναι ο pxp μοναδιαίος πίνακας. (Arashi et al. 2021)

Μία διαφορετική μορφή της παλινδρόμησης Ridge περιγράφεται ως εξής: οι εκτιμήσεις των συντελεστών β_j, j=1,…,p λύνουν το πρόβλημα:

$$\text{minimize}_\beta \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{δεδομένου ότι} \quad \sum_{j=1}^{p} \beta_j^2 \leq \lambda \qquad (3.2)$$

Δηλαδή οι σχέσεις (3.1) και (3.2) μας δίνουν ακριβώς την ίδια λύση και κατά συνέπεια τις ίδιες εκτιμήσεις των β_j της παλινδρόμησης Ridge. (Tibshirani 1996)

Στη σχέση (3.1) ο όρος $\sum_{j=1}^{p} \beta_j^2$ (ο οποίος αποτελεί μέρος της ποινής) ονομάζεται l_2-πέναλτι, γιατί βασίζεται στην l_2-νόρμα. Υπενθυμίζεται ότι η l_2-νόρμα ορίζεται ως:

$$\|\beta\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$$

Καθώς η παράμετρος λ αυξάνεται, η l_2-νόρμα των εκτιμήσεων $\hat{\beta}^R$ θα μειώνεται.

Η μέθοδος Ridge είναι ουσιαστικά μία βελτίωση της μεθόδου των Ελαχίστων Τετραγώνων (OLS), διότι βελτιώνει το Μέσο Τετραγωνικό Σφάλμα (MSE). Γενικά το Μέσο Τετραγωνικό Σφάλμα (MSE) ορίζεται ως:

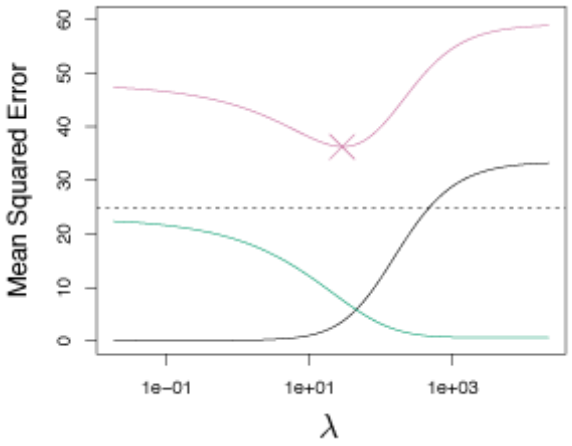
$$\text{MSE}(\hat{\beta}) = \text{var}(\hat{\beta}) + \left[\text{E}(\hat{\beta}) - \beta\right]^2$$

(Wieringen 2021)

όπου ο όρος $\text{E}(\hat{\beta}) - \beta$ είναι η μεροληψία (bias) που έχει η εκτιμήτρια $\hat{\beta}$. Όταν έχουμε δεδομένα όπου υπάρχει μεγάλη συσχέτιση μεταξύ των μεταβλητών ή όταν ο αριθμός των μεταβλητών είναι μεγάλος σε σχέση με τον αριθμό των παρατηρήσεων, η μέθοδος Ελαχίστων Τετραγώνων (OLS) είναι ασταθής (γιατί υπάρχει μεγάλη διασπορά). Σε αυτές τις περιπτώσεις η μέθοδος Ridge μας είναι πολύ πιο χρήσιμη. Καθώς η παράμετρος λ αυξάνεται, μειώνεται η ευελιξία της προσαρμογής της μεθόδου Ridge, οδηγώντας σε μείωση της διασποράς (variance) αλλά σε αύξηση της μεροληψίας (bias). Εμείς ψάχνουμε μία τιμή της παραμέτρου λ που ελαχιστοποιεί το Μέσο Τετραγωνικό Σφάλμα (MSE), δηλαδή που πετυχαίνει μείωση της διασποράς μαζί με μικρή αύξηση της μεροληψίας.

Στο Σχήμα 3.1 βλέπουμε ένα παράδειγμα για το πως συμπεριφέρεται το Μέσο Τετραγωνικό Σφάλμα (MSE) καθώς η παράμετρος λ αλλάζει. Στη μοβ καμπύλη παριστάνεται το MSE, στη πράσινη η διασπορά και στη μαύρη το τετράγωνο της μεροληψίας (bias^2). Όπως βλέπουμε καθώς το λ αυξάνεται, η διασπορά (πράσινη) μειώνεται συνεχώς μέχρι να μηδενιστεί. Αντίθετα η μεροληψία (μαύρη) στην αρχή αυξάνεται με μικρό ρυθμό, ενώ μετά από κάποια τιμή του λ αυξάνεται απότομα.

Εμείς θέλουμε να βρούμε εκείνη τη τιμή του λ που τέμνει τις καμπύλες της διασποράς και της μεροληψίας. Δηλαδή, τη τιμή που πετυχαίνουμε τη μικρότερη τιμή του MSE (ροζ καμπύλη). (James et al. 2013)



Σχήμα 3.1: *Διάγραμμα του MSE συναρτήσει του λ για τη μέθοδο Ridge.*

3.3 Παλινδρόμηση Lasso και η σύγκριση της με τη Ridge

Άλλη μια σημαντική μέθοδος συρρίκνωσης, είναι η μέθοδος Lasso (Least Absolute Shrinkage and Selection Operator), η οποία είναι παρόμοια με τη μέθοδο Ridge (Παράγραφος 3.2). Η μέθοδος Ridge έχει ένα βασικό μειονέκτημα, ο όρος της ποινής που εφαρμόζει $\left(\lambda\sum_{j=1}^{p}\beta_j^2\right)$, όπως είδαμε στην Παράγραφο 3.2, συρρικνώνει όλους τους συντελεστές β_j, j=1,…,p, αλλά δεν μηδενίζει κανέναν, εκτός εάν $\lambda=\infty$. Δηλαδή, η Ridge στο τελικό μοντέλο που δημιουργεί, συμπεριλαμβάνει όλες τις μεταβλητές που υπάρχουν (εφ' όσον καμία δεν μηδενίζεται ακριβώς). Αυτό δεν αποτελεί πρόβλημα στην ακρίβεια των προβλέψεων, αλλά στις περιπτώσεις που έχουμε μεγάλο αριθμό μεταβλητών (p) θα έχουμε δυσκολία στην ερμηνεία του μοντέλου μας.

Αντίθετα η μέθοδος Lasso ξεπερνάει αυτό το μειονέκτημα της Ridge και εκτιμάει τους συντελεστές β_j, j=1,…,p ($\hat{\beta}^L$) ελαχιστοποιώντας την ακόλουθη ποσότητα:

$$L_{\text{Lasso}} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \text{RSS} + \lambda\sum_{j=1}^{p}|\beta_j| \qquad (3.3)$$

(James et al. 2013)

όπου RSS (Residual Sum of Squares) είναι το άθροισμα των τετραγώνων των υπολοίπων, όπως και στη μέθοδο Ridge. Η διαφορά της σχέσης (3.3) από τη (3.1) της Ridge βρίσκεται στον όρο ποινής. Η μέθοδος Lasso χρησιμοποιεί σαν όρο ποινής την ποσότητα $\lambda\sum_{j=1}^{p}|\beta_j|$. Σε αυτή την περίπτωση έχουμε l_1-πέναλτι, αντί για l_2-πέναλτι που είχαμε στη Ridge. Υπενθυμίζεται ότι η l_1-νόρμα ορίζεται ως:

$$\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$$

Όπως και με τη μέθοδο Ridge, η Lasso συρρικνώνει τις εκτιμήσεις των συντελεστών προς το μηδέν ανάλογα με τη τιμή της παραμέτρου λ (tuning parameter), όσο πιο μεγάλο το λ τόσο θα μεγαλώνει και η συρρίκνωση. Η διαφορά είναι ότι η μέθοδος Lasso, εξαιτίας του l_1-πέναλτι, αναγκάζει κάποιους συντελεστές να μηδενιστούν ακριβώς (όχι απλώς να τείνουν προς το μηδέν) για μεγάλες τιμές του λ. Οπότε η Lasso εφαρμόζει επιλογή συμμεταβλητών, δηλαδή επιλέγει ποιες συμμεταβλητές θα είναι στο μοντέλο της ανάλογα με την επίδρασή τους και μηδενίζει τις υπόλοιπες (κάτι το οποίο δεν κάνει η Ridge). Αυτό είναι και το βασικό της πλεονέκτημα σε σχέση με τη Ridge.

Μία διαφορετική μορφή της παλινδρόμησης Lasso περιγράφεται ως εξής: οι εκτιμήσεις των συντελεστών β_j, j=1,…,p λύνουν το πρόβλημα:

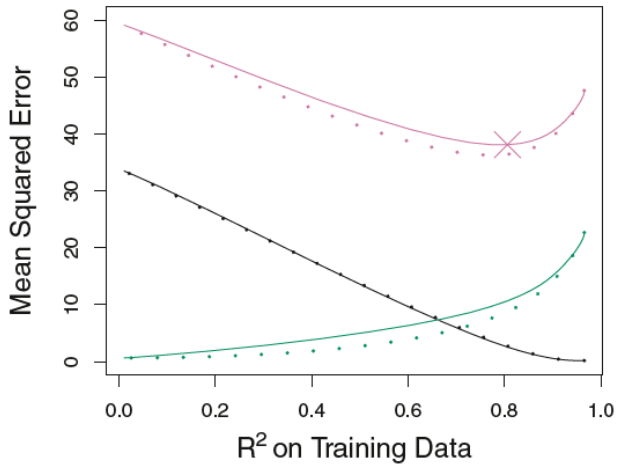
$$\text{minimize}_\beta \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{δεδομένου ότι} \quad \sum_{j=1}^{p} |\beta_j| \leq \lambda \qquad (3.4)$$

Δηλαδή οι σχέσεις (3.3) και (3.4) μας δίνουν ακριβώς την ίδια λύση και κατά συνέπεια τις ίδιες εκτιμήσεις των β_j της παλινδρόμησης Lasso. (Tibshirani 1996)

Η επιλογή της παραμέτρου λ έχει καθοριστικό ρόλο για την παλινδρόμηση Lasso (όπως και για τη Ridge). Ο τρόπος που επιλέγουμε το λ είναι ίδιος με τη Ridge που περιγράψαμε στην Παράγραφο 3.2, θέλουμε μία τιμή της παραμέτρου λ που να ελαχιστοποιεί το Μέσο Τετραγωνικό Σφάλμα (MSE), δηλαδή να πετυχαίνει μείωση της διασποράς μαζί με μικρή αύξηση της μεροληψίας. Υπενθυμίζουμε ότι το Μέσο Τετραγωνικό Σφάλμα (MSE) ορίζεται ως:

$$\text{MSE}(\hat{\beta}) = \text{var}(\hat{\beta}) + \left[\text{E}(\hat{\beta}) - \beta\right]^2$$

όπου ο όρος $\text{E}(\hat{\beta}) - \beta$ είναι η μεροληψία (bias).



Σχήμα 3.2: *Διάγραμμα του MSE συναρτήσει του συντελεστή προσδιορισμού R^2 για τη Lasso (γραμμές) και για τη Ridge (τελείες).*

Στο Σχήμα 3.2 βλέπουμε ένα παράδειγμα για το πως συμπεριφέρεται το Μέσο Τετραγωνικό Σφάλμα (MSE) και για την παλινδρόμηση Lasso (γραμμές) και για τη Ridge (τελείες). Βλέπουμε ότι και οι δύο αυτές μέθοδοι είναι πανομοιότυπες, υπενθυμίζουμε ότι στη μοβ καμπύλη παριστάνεται το MSE, στη πράσινη η διασπορά και στη μαύρη το τετράγωνο της μεροληψίας (bias^2). Οι καμπύλες

της μεροληψίας (μαύρες) ταυτίζονται μεταξύ τους για τις δύο αυτές μεθόδους (Ridge και Lasso), αλλά φαίνεται καθαρά ότι η Ridge (τελείες) έχει μικρότερη διασπορά (πράσινη) από τη Lasso (γραμμές). Κατά συνέπεια η Ridge έχει και μικρότερο Μέσο Τετραγωνικό Σφάλμα (MSE) (μοβ καμπύλη) από τη Lasso. Αυτό συμβαίνει γιατί υπάρχει μεγάλη συσχέτιση μεταξύ των μεταβλητών που περιγράφονται στο Σχήμα 3.2 και η lasso θεωρεί ότι κάποιοι συντελεστές είναι πραγματικά μηδέν, κάτι που είναι καλό για την περιγραφή του μοντέλου. Οπότε είναι λογικό η Ridge να έχει μικρότερο Μέσο Τετραγωνικό Σφάλμα (MSE).

Εάν είχαμε δεδομένα που ήταν λιγότερο ή καθόλου συσχετισμένα τότε η μέθοδος Lasso θα μας έδινε χαμηλότερο Μέσο Τετραγωνικό Σφάλμα (MSE) από τη Ridge. Επίσης εάν είχαμε δεδομένα με λιγότερες μεταβλητές η Lasso θα μας έδινε πάλι χαμηλότερο MSE. Γενικά δεν μπορούμε να πούμε ότι κάποια από τις δύο αυτές μεθόδους (Ridge και Lasso) υπερτερεί της άλλης, και οι δύο έχουν κάποια πλεονεκτήματα και κάποια μειονεκτήματα. Η Lasso προτιμάται στις περιπτώσεις όπου δεν υπάρχει μεγάλη συσχέτιση μεταξύ των μεταβλητών. Ενώ η Ridge προτιμάται σε περιπτώσεις όπου έχουμε μεγάλη συσχέτιση.

Γενικά, η παλινδρόμηση Κορυφογραμμής (Ridge) εφαρμόζεται κυρίως στα γραμμικά μοντέλα (Linear model), ενώ η Lasso χρησιμοποιείται συχνά και στο ημι-παραμετρικό μοντέλο του Cox (Παράγραφος 2.4). Σε αυτή την περίπτωση θα είχαμε:

$$L(\beta) = \prod_{\mathrm{j}=1}^{\mathrm{k}} \left\{ \frac{\mathrm{e}^{\beta' \mathrm{x_j}}}{\sum_{i \in \mathrm{R_j}} \mathrm{e}^{\beta' \mathrm{x_I}}} \right\}$$

και η μορφή που θα πάρει η μέθοδος Lasso θα είναι:

$$\text{minimize}_{\beta}\, l(\beta), \qquad \text{δεδομένου ότι} \quad \sum_{\mathrm{j}=1}^{\mathrm{p}} |\beta_{\mathrm{j}}| \leq \lambda$$

όπου $l(\beta)$ είναι η λογαριθμισμένη μερική συνάρτηση μέγιστης Πιθανοφάνειας. (Tibshirani 1997)

Σημειώνεται ότι και οι δύο αυτές μέθοδοι είναι ειδικές περιπτώσεις μίας άλλης μεθόδου που ονομάζεται Elastic net (δεν θα ασχοληθούμε με αυτή τη μέθοδο). Οι εκτιμήσεις των β_j, j=1,…,p υπολογίζονται ελαχιστοποιώντας την ποσότητα:

$$L_{\mathrm{enet}} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \left(\mathrm{y_i} - \beta_0 - \sum_{\mathrm{j}=1}^{\mathrm{p}} \beta_{\mathrm{j}} \mathrm{x_{ij}} \right)^2}{2\mathrm{n}} + \lambda \left\{ \frac{1-\alpha}{2} \sum_{\mathrm{j}=1}^{\mathrm{p}} \beta_{\mathrm{j}}^2 + \alpha \sum_{\mathrm{j}=1}^{\mathrm{p}} |\beta_{\mathrm{j}}| \right\} \qquad (3.5)$$

όπου το α είναι μία παράμετρος ανάμιξης (mixing parameter) που παίρνει τιμές μεταξύ του 0 και του 1. Ενώ η παράμετρος λ είναι ίδια με αυτήν που έχουμε περιγράψει. Για την παλινδρόμηση Ridge η σχέση (3.5) λαμβάνει για την παράμετρο α τη τιμή 0 (α=0), ενώ για την παλινδρόμηση Lasso λαμβάνει τη τιμή 1 (α=1). (James et al. 2013)

3.4 Μέθοδος Cross-Validation

Όπως αναφέραμε στις Παραγράφους 3.2 και 3.3, στη παλινδρόμηση Κορυφογραμμής (Ridge) και στη Lasso χρειάζεται να βρούμε την κατάλληλη τιμή της παραμέτρου λ έτσι ώστε να έχουμε το χαμηλότερο Μέσο Τετραγωνικό Σφάλμα (MSE). Αυτή τη τιμή του λ θα τη βρούμε με τη μέθοδο Cross-validation (CV), διότι αποτελεί έναν εύκολο και απλό τρόπο. Θα επιλέξουμε ένα φάσμα τιμών του λ και θα υπολογίσουμε το σφάλμα της Cross-validation για κάθε τιμή του λ. Έπειτα, θα επιλέξουμε τη τιμή εκείνη που ελαχιστοποιεί το σφάλμα Cross-validation. Τέλος, θα ξανά προσαρμόσουμε τα δύο μοντέλα μας (Ridge και Lasso) χρησιμοποιώντας τη τιμή του λ που επιλέξαμε.

Πιο συγκεκριμένα εφαρμόζουμε τη μέθοδο Leave-One-Out Cross-validation (LOOCV). Αυτή η μέθοδος χωρίζει τις παρατηρήσεις μας σε δύο υποσύνολα, το δοκιμαστικό σετ (training set) και το σετ επικύρωσης (validation set). Το validation set θα έχει μόνο μία παρατήρηση, ενώ οι υπόλοιπες θα ανήκουν στο training set. Γενικά, η κλασική μέθοδος Cross-validation χωρίζει τις παρατηρήσεις σε αυτά τα δύο υποσύνολα (training set και validation set) με τυχαίο τρόπο ώστε και τα δύο υποσύνολα να περιέχουν αρκετές παρατηρήσεις. Το μοντέλο που εξετάζουμε θα προσαρμοστεί στο training set (που περιέχει n-1 παρατηρήσεις) και στη συνέχεια αυτό το προσαρμοσμένο μοντέλο θα χρησιμοποιηθεί για να προβλέψει τη μοναδική παρατήρηση που ανήκει στο validation set. Το σφάλμα του validation set συνήθως υπολογίζεται από το Μέσο Τετραγωνικό Σφάλμα (MSE).

Έστω ότι έχουμε $(\mathrm{x_i}, \mathrm{y_i})$, $\mathrm{i} = 1 \ldots, \mathrm{n}$ παρατηρήσεις και το validation set περιέχει την πρώτη παρατήρηση $(\mathrm{x_1}, \mathrm{y_1})$, τότε θα έχουμε:

$$\mathrm{MSE_1} = (\mathrm{y_1} - \hat{\mathrm{y}}_1)^2$$

το οποίο θα είναι το Μέσο Τετραγωνικό Σφάλμα του validation set. Εφ' όσον η παρατήρηση $(\mathrm{x_1}, \mathrm{y_1})$ δεν χρησιμοποιήθηκε για την προσαρμογή του μοντέλου (διότι το μοντέλο προσαρμόζεται με βάση το training set), το σφάλμα $\mathrm{MSE_1}$ θα μας παρέχει μία σχεδόν αμερόληπτη εκτίμηση για το σφάλμα της δοκιμής. Αλλά συγχρόνως είναι και μία «φτωχή» εκτίμηση καθώς βασίζεται μόνο σε μία παρατήρηση.

Θα συνεχίσουμε αυτή τη διαδικασία επιλέγοντας το validation set να περιέχει τη δεύτερη παρατήρηση $(\mathrm{x_2}, \mathrm{y_2})$, με $\mathrm{MSE_2} = (\mathrm{y_2} - \hat{\mathrm{y}}_2)^2$ και στη συνέχεια να περιέχει διαδοχικά κάθε μία παρατήρηση από τις n παρατηρήσεις που έχουμε. Έτσι, στο τέλος της διαδικασίας θα έχουμε n στον αριθμό Μέσα Τετραγωνικά Σφάλματα. Η εκτίμηση της μεθόδου LOOCV για το MSE ορίζεται ως το μέσο όρο των σφαλμάτων $\mathrm{MSE_i}$, $\mathrm{i} = 1 \ldots, \mathrm{n}$:

$$\mathrm{CV_{(n)}} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{MSE_i}$$

Σε σχέση με τη κλασική μέθοδο Cross-validation (CV), η μέθοδος LOOCV έχει πολλά πλεονεκτήματα. Ένα βασικό πλεονέκτημα είναι ότι η LOOCV έχει μικρότερη μεροληψία (bias). Αυτό συμβαίνει γιατί το training set της μεθόδου CV περιέχει σχεδόν τις μισές παρατηρήσεις από το πλήθος n παρατηρήσεων που έχουμε και προσαρμόζει το μοντέλο με βάση αυτές. Ενώ το training set της μεθόδου LOOCV περιέχει σχεδόν όλες τις παρατηρήσεις, αφού περιέχονται σε αυτό n-1 παρατηρήσεις. Αυτό έχει ως αποτέλεσμα η μέθοδος CV να υπερεκτιμά το ποσοστό σφάλματος της δοκιμής. Επίσης, η μέθοδος CV λόγω της τυχαιότητας που υπάρχει στο χωρισμό των δύο υποσυνόλων, κάθε φορά που εφαρμόζεται θα παράγει και διαφορετικό αποτέλεσμα. Ενώ η LOOCV θα μας δίνει κάθε φορά το ίδιο αποτέλεσμα όσες φορές και να εφαρμοστεί, διότι δεν υπάρχει τυχαιότητα στο χωρισμό των δύο υποσυνόλων. (James et al.2013)

Κεφάλαιο 4: Διάρκεια Ζωής Λαμπτήρων Φθορισμού

Σε αυτό το κεφάλαιο θα αναλύσουμε στατιστικά δεδομένα διάρκειας ζωής λαμπτήρων φθορισμού. Τα δεδομένα που έχουμε αφορούν τη διάρκεια ζωής (Time) 16 λαμπτήρων φθορισμού σε σχέση με 5 διαφορετικούς παράγοντες A, B, C, D και E, οι οποίοι λαμβάνουν τιμές 0 και 1. Επίσης έχουμε και δεξιά αποκομμένες παρατηρήσεις που συμβολίζονται με τη συμμεταβλητή Status.

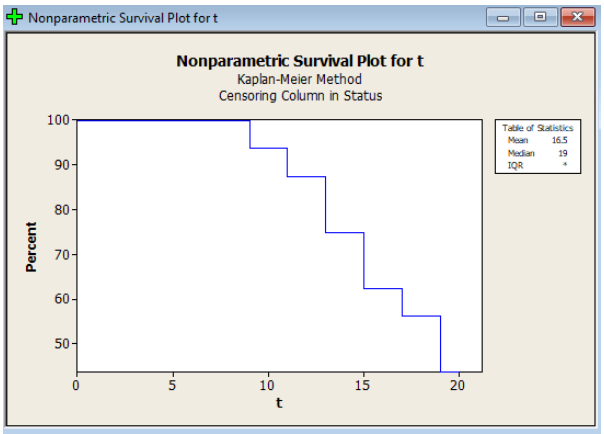
4.1 Εκτίμηση Kaplan-Meier

Με τη βοήθεια του στατιστικού πακέτου Minitab θα εφαρμόσουμε την εκτίμηση Kaplan-Meier στο σύνολο των δεδομένων μας:

Πίνακας 4.1

Time	Number at Risk	Number Failed	Survival Probability (SKM)	Standard Error	95.0% Normal CI Lower	95.0% Normal CI Upper
9	16	1	0.9375	0.060515	0.818892	1.00000
11	15	1	0.8750	0.082680	0.712951	1.00000
13	14	2	0.7500	0.108253	0.537828	0.96217
15	12	2	0.6250	0.121031	0.387784	0.86222
17	10	1	0.5625	0.124020	0.319426	0.80557
19	9	2	0.4375	0.124020	0.194426	0.68057

Ο παραπάνω πίνακας 4.1 της εκτίμησης Kaplan-Meier μας δείχνει στη τέταρτη στήλη τις εκτιμήσεις (πιθανότητες επιβίωσης) που υπολόγισε η εκτιμήτρια Kaplan-Meier. Στη δεύτερη στήλη φαίνεται ο αριθμός των λαμπτήρων φθορισμού που βρίσκονται σε κίνδυνο για κάθε μία χρονική στιγμή (Time), ενώ στη τρίτη φαίνεται ο αριθμός των λαμπτήρων στους οποίους έχει συμβεί το γεγονός, δηλαδή που έπαψαν να λειτουργούν. Τέλος στις τρεις τελευταίες στήλες βλέπουμε το τυπικό σφάλμα και τα διαστήματα εμπιστοσύνης (95%) για τις τιμές της κάθε εκτίμησης.



Σχήμα 4.1: *Εκτίμηση Kaplan-Meier της συνάρτησης επιβίωσης των λαμπτήρων.*

Στο Σχήμα 4.1 παρουσιάζεται η γραφική παράσταση της εκτίμησης Kaplan-Meier. Βλέπουμε ότι καθώς μεγαλώνει ο χρόνος (t), η πιθανότητα επιβίωσης μειώνεται σημαντικά μετά τη χρονική στιγμή

των 10 ωρών. Όταν τελειώσει η παρατήρηση των λαμπτήρων μας, βρίσκονται σε κίνδυνο 9 λαμπτήρες με πιθανότητα να επιβιώσουν 0.4375.

4.2 Γραφικές Παραστάσεις Κατανομών

Θα συγκρίνουμε τις γραφικές παραστάσεις τριών βασικών κατανομών, Weibull, Log-Normal και Log-Logistic, τις τιμές των οποίων θα βρούμε με τη βοήθεια της εκτίμησης Kaplan-Meier, όπως είδαμε αναλυτικά στην Παράγραφο 1.2.1. Σκοπός μας είναι να δούμε ποια από τις τρεις κατανομές προσαρμόζεται καλύτερα στα δεδομένα διάρκειας ζωής των λαμπτήρων φθορισμού που έχουμε.

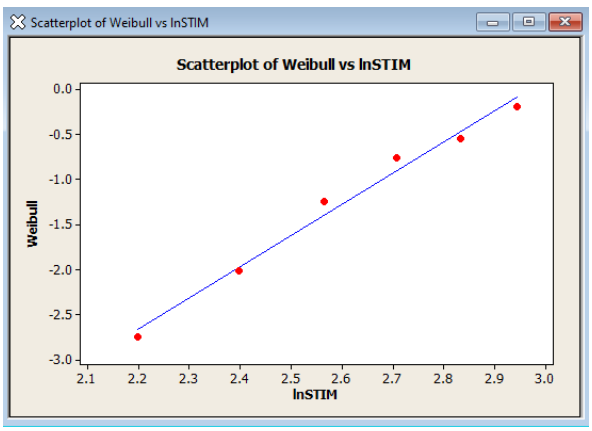
Υπολογίζουμε τις τιμές των γραφικών παραστάσεων από τους τύπους:

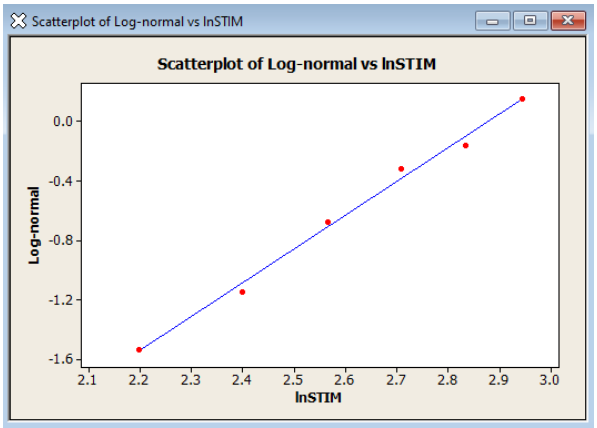
- Weibull: ln{ - ln(SKM) } για lnt
- Log-Normal: Φ^{-1}(1- SKM) για lnt
- Log-Logistic: ln{ (1- SKM) / SKM } για lnt

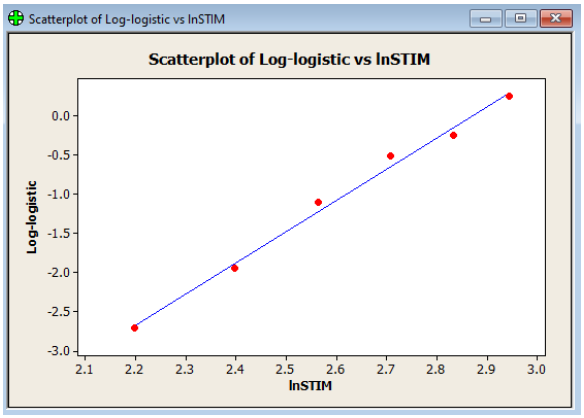
όπου SKM είναι η εκτίμηση Kaplan-Meier όπως την έχουμε υπολογίσει στον παραπάνω πίνακα (Πίνακας 4.1) με τη βοήθεια του Minitab, ενώ η Φ όπως ξέρουμε είναι οι τιμές της τυποποιημένης κανονικής κατανομής (N(0,1)).

Πίνακας 4.2

Time	Weibull	Log-normal	Log-logistic
9	-2.74049	-1.53412	-2.70805
11	-2.01342	-1.15035	-1.94591
13	-1.24590	-0.67449	-1.09861
15	-0.75501	-0.31864	-0.51083
17	-0.55275	-0.15731	-0.25131
19	-0.19034	0.15731	0.25131







Σχήμα 4.2: *Γραφικές παραστάσεις των τριών κατανομών συναρτήσει του λογαριθμισμένου χρόνου.*

Στο Σχήμα 4.2 βλέπουμε τρεις γραφικές παραστάσεις, μία για κάθε κατανομή, Weibull, Log-Normal και Log-Logistic αντίστοιχα ενώ με μπλε γραμμή απεικονίζεται η εκτίμηση Kaplan-Meier. Όπως βλέπουμε και οι τρεις κατανομές προσαρμόζονται πολύ καλά στα δεδομένα που έχουμε, καθώς τα σημεία και των τριών κατανομών ακολουθούν και πλησιάζουν σε πολύ μεγάλο βαθμό την ευθεία της Kaplan-Meier. Αλλά με μικρή διαφορά θα επιλέξουμε την κατανομή **Log-Normal** όπου βλέπουμε ότι είναι λίγο καλύτερη από τις υπόλοιπες.

4.3 Έλεγχος Wald

Θα εφαρμόσουμε έλεγχο **Wald** (Παράγραφος 2.5) για να ελέγξουμε ποια από τις Εκθετική και Weibull κατανομές (Παράγραφος 1.1) προσαρμόζεται καλύτερα στο σύνολο των δεδομένων που έχουμε. Ο τύπος του ελέγχου Wald είναι:

$$z = \frac{\hat{\eta} - \eta_0}{se(\hat{\eta})} \sim N(0,1)$$

Η υπόθεση που ελέγχει ο Wald είναι: $H_0: \eta = 1$, $H_1: \eta \neq 1$, όπου η είναι η παράμετρος σχήματος (Shape) της κατανομής Weibull (αν η =1 τότε έχουμε την Εκθετική κατανομή). Παρατηρούμε ωστόσο ότι το Minitab εφαρμόζει τον έλεγχο Wald με τη διόρθωση όπου χρησιμοποιεί το $\ln(\hat{\eta})$ αντί για για το $\hat{\eta}$, δηλαδή δεν ελέγχει την υπόθεση που είπαμε αλλά ελέγχει την μηδενική υπόθεση $H_0: \ln(\eta) = 0$ με την εναλλακτική $H_1: \ln(\eta) \neq 0$. Οπότε ο τύπος του ελέγχου θα γίνει:

$$z = \frac{\ln(\hat{\eta}) - \ln(\eta)}{se(\ln(\hat{\eta}))} \sim N(0,1) \qquad \text{όπου} \quad se(\ln(\hat{\eta})) \cong \frac{se(\hat{\eta})}{\hat{\eta}}$$

Τα Διαστήματα Εμπιστοσύνης (95%) θα υπολογίζονται από το τύπο:

$$\exp[\ln(\hat{\eta}) \pm (1.96) \times se(\ln(\hat{\eta}))]$$

Από τις εκτιμήσεις του πίνακα 4.3 προκύπτει: $se(\ln(\hat{\eta})) = \frac{1.07377}{3.57765} = 0.30011$ οπότε $z = 4.2474 \Rightarrow z^2 = 18.0382$ (Chi-Square) με βαθμό ελευθερίας 1 και όπως βλέπουμε έχει τιμή p-value $\ll 0.0001$.

Άρα απορρίπτουμε την H_0 (η=1) γιατί η p-value τιμή του ελέγχου Wald είναι $\ll 0.0001$ (πολύ μικρή). Οπότε μεταξύ των δύο αυτών κατανομών πιο κατάλληλη είναι η **Weibull** κατανομή.

Πίνακας 4.3

Parameter Estimates: Standard 95,0% Normal CI Parameter Estimate Error Lower Upper Shape 3,57765 1,07377 1,98666 6,44275 Scale 2,09038 2,06410 17,2256 25,3674 Log-Likelihood = -33,899 Goodness-of-Fit Anderson-Darling (adjusted) = 50,749	Test for Shape Equal to 1 Chi-Square DF P 18.0382 1 0.000 Bonferroni 95.0% (indiv 95.00%) Simultaneous CI Shape parameter for: Variable Lower Estimate Upper -------+---------+---------+--------- t 1.987 3.578 6.443 (----------*----------) -------+---------+---------+--------- 3.0 4.5 6.0

4.4 Κριτήριο AIC

Θα εφαρμόσουμε και το **κριτήριο AIC** (Παράγραφος 2.5) για να συγκρίνουμε και πάλι αυτές τις τρεις κατανομές ως προς το πόσο καλά προσαρμόζονται στα δεδομένα μας. Ο τύπος του κριτηρίου AIC είναι :

$$AIC = -2l + 2k$$

όπου k ο αριθμός των παραμέτρων που έχουμε και l η Λογαριθμισμένη Συνάρτηση Πιθανοφάνειας την οποία θα βρούμε εφαρμόζοντας τη μέθοδο Μέγιστης Πιθανοφάνειας για κάθε κατανομή.

Estimation Method: Maximum Likelihood

- Distribution: Weibull

Parameter Estimates:
Standard 95,0% Normal CI
Parameter Estimate Error Lower Upper
Shape 3,57765 1,07377 1,98666 6,44275
Scale 2,09038 2,06410 17,2256 25,3674

Log-Likelihood = -33,899
Goodness-of-Fit
Anderson-Darling (adjusted) = 50,749

- Distribution: Lognormal

Parameter Estimates:
Standard 95,0% Normal CI
Parameter Estimate Error Lower Upper
Location 2,92691 0,112297 2,70681 3,14701
Scale 0,387317 0,100830 0,232527 0,645148

Log-Likelihood = - 33,445
Goodness-of-Fit
Anderson-Darling (adjusted) = 50,743

- Distribution: Loglogistic

Parameter Estimates:
Standard 95,0% Normal CI
Parameter Estimate Error Lower Upper
Location 2,92480 0,107747 2,71362 3,13599
Scale 0,229664 0,0654829 0,131339 0,401598

Log-Likelihood = - 33,672
Goodness-of-Fit
Anderson-Darling (adjusted) = 50,748

Εδώ έχουμε k=2, γιατί έχουμε 2 παραμέτρους (scale και location). Οπότε το **κριτήριο AIC** για τις τρεις κατανομές θα είναι:

- Weibull: -2*(-33,899) + 2*2 = 71,798
- Log-Normal: -2*(-33,445) + 2*2 = 70,89
- Log-Logistic: -2*(-33,672) + 2*2 = 71,344

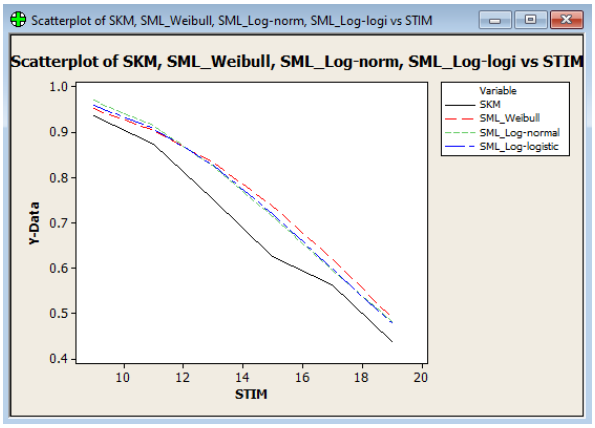
Παρατηρούμε ότι και οι τρεις τιμές του κριτηρίου AIC είναι πολύ κοντά μεταξύ τους. Όμως σύμφωνα με το κριτήριο AIC η κατανομή **Log-Normal** προσαρμόζεται καλύτερα στα δεδομένα που έχουμε γιατί έχει τη μικρότερη τιμή AIC.

4.5 Εκτίμηση Μέγιστης Πιθανοφάνειας της Συνάρτησης Επιβίωσης

Ο τρίτος και τελευταίος έλεγχος που θα εφαρμόσουμε είναι μέσω της εκτίμησης Μέγιστης Πιθανοφάνειας της συνάρτησης επιβίωσης, την οποία θα την υπολογίσουμε με την βοήθεια του Minitab και ξεχωριστά για κάθε μία από τις τρεις κατανομές.

Πίνακας 4.4

Time	SML_Weibull	SML_Log-normal	SML_Log-logistic
9	0.952133	0.970214	0.959613
11	0.904329	0.914006	0.908401
13	0.832928	0.824985	0.827338
15	0.737103	0.713985	0.719865
17	0.620443	0.595576	0.598401
19	0.491346	0.481951	0.478640



Σχήμα 4.3: *Γραφική παράσταση της Kaplan-Meier και των τριών Εκτιμητριών Μέγιστης Πιθανοφάνειας συναρτήσει του χρόνου.*

Στο Σχήμα 4.3 βλέπουμε τις τρεις Εκτιμήτριες Μέγιστης Πιθανοφάνειας μαζί με την Εκτίμηση Kaplan-Meier (μαύρη καμπύλη). Σκοπός μας είναι να δούμε ποια καμπύλη πλησιάζει περισσότερο αυτή της Kaplan-Meier. Και οι τρεις καμπύλες των κατανομών είναι αρκετά πανομοιότυπες. Αλλά μπορούμε να διακρίνουμε ότι η καμπύλη της κατανομής **Log-Logistic** πλησιάζει καλύτερα από τις υπόλοιπες την καμπύλη της Kaplan-Meier. Δηλαδή σύμφωνα με την εκτίμηση Μέγιστης Πιθανοφάνειας η κατανομή **Log-Logistic** προσαρμόζεται πιο καλά στα δεδομένα μας.

Συμπέρασμα: Με όλα τα διαφορετικά κριτήρια που χρησιμοποιήσαμε, καταλαβαίνουμε ότι οι κατανομές **Log-Normal** και **Log-Logistic** είναι και οι δύο οι πιο κατάλληλες επιλογές για τα δεδομένα

των λαμπτήρων μας σε σχέση με την κατανομή Weibull. Δηλαδή, αυτές οι δύο κατανομές μπορούν να περιγράψουν με μεγαλύτερη ακρίβεια το πως συμπεριφέρεται η διάρκεια ζωής των συγκεκριμένων λαμπτήρων φθορισμού που μελετάμε.

4.6 Μοντέλο Παλινδρόμησης Επιταχυνόμενης Διάρκειας Ζωής (AL)

Σε αυτή την παράγραφο θα προσαρμόσουμε στα δεδομένα μας το Μοντέλο Παλινδρόμησης της Επιταχυνόμενης Διάρκειας Ζωής (AL) με το στατιστικό πακέτο R, χρησιμοποιώντας την κατανομή Weibull (όπως είδαμε αναλυτικά στην Παράγραφο 2.2).

Πίνακας 4.5

mod <- survreg(Surv(t,Status)~A+B+C+D+E, data=data,dist="weibull")
summary(mod)
Call: survreg(formula = Surv(t, Status) ~ A + B + C + D + E, data = data, dist = "weibull")

	Value	Std. Error	z	p
(Intercept)	3.0135	0.1391	21.67	<2e-16
A	0.2081	0.1087	1.91	0.056
B	-0.4522	0.1092	-4.14	3.4e-05
C	-0.0594	0.1181	-0.50	0.615
D	0.4959	0.1115	4.45	8.7e-06
E	-0.2309	0.1092	-2.11	0.034
Log(scale)	-1.9791	0.2829	-7.00	2.6e-12

Scale= 0.138
Weibull distribution
Loglik(model)= -25.3 Loglik(intercept only)= -33.9
Chisq= 17.23 on 5 degrees of freedom, p= 0.0041
Number of Newton-Raphson Iterations: 6
n= 16

Στον Πίνακα 4.5 παρουσιάζονται οι εκτιμώμενες τιμές των συντελεστών καθώς και άλλες σημαντικές πληροφορίες όπως οι τιμές p-value (τελευταία στήλη). Παρατηρούμε ότι η R υπολογίζει την παράμετρο κλίμακας (Scale) της κατανομής Weibull αυτόματα, αλλά θέτει Scale= 1/η, όπου η είναι η παράμετρος σχήματος (Shape). Έτσι βλέπουμε ότι 1/η= 0.138, συνεπώς η παράμετρος η θα είναι μεγαλύτερη της μονάδας που σημαίνει ότι ο κίνδυνος αυξάνεται όσο περνάει ο χρόνος, κάτι το οποίο είδαμε και με την εκτίμηση της Kaplan-Meier στο Σχήμα 4.1. Υπενθυμίζουμε ότι σύμφωνα με την κατανομή Weibull η συνάρτηση επιβίωσης είναι:

$$S(t) = \exp\{-(t/\alpha)^{\eta}\}$$

και οι μεταβλητές εισάγονται μέσω της παραμέτρου κλίμακας α και συνήθως χρησιμοποιούμε α(x)=αexp(β′x) οπότε η συνάρτηση επιβίωσης γίνεται:

$$S(t) = \exp\left\{-\left(t/\alpha e^{\beta' x}\right)^{\eta}\right\}$$

Αυτό σημαίνει ότι μετά από χρόνο t, μία μεταβλητή x συμπεριφέρεται σαν να έχει χρόνο ζωής $te^{-\beta' x}$. Πιο συγκεκριμένα για τη συμμεταβλητή A ισχύει ότι exp(-0.2081)=0.812 <1 οπότε επιβραδύνεται η διακοπή λειτουργίας των λαμπτήρων μετά από χρόνο t. Το ίδιο ισχύει και για τη συμμεταβλητή D όπου exp(-0.4959)= 0.6090 <1. Ενώ οι υπόλοιπες (B,C και E) φαίνεται να επιταχύνουν τη διακοπή λειτουργίας των λαμπτήρων, αφού έχουμε για τη B: exp(0.4522)= 1.572 >1, για τη C: exp(0.0594)= 1.061 >1 και για την E: exp(0.2309)= 1.2597 >1.

Στη συνέχεια θα εφαρμόσουμε την **backward τεχνική με βήματα** (Παράγραφος 2.5) για να καταλήξουμε σε ένα καλύτερο μοντέλο, δηλαδή θα βρούμε το καλύτερο μοντέλο με το μικρότερο κριτήριο AIC και θα δούμε ποιες συμμεταβλητές θα έχει. Στον Πίνακα 4.6 βλέπουμε τη διαδικασία της διαδοχικής αφαίρεσης μεταβλητών, στην αρχή ξεκινάει με ένα μοντέλο που περιέχει και τις 5 συμμεταβλητές με AIC = 64.56 και διαδοχικά αφαιρεί μεταβλητές οι οποίες δεν είναι στατιστικά σημαντικές.

Πίνακας 4.6

step(mod, direction="backward", test="Chisq")
Start: AIC = 64.56
Surv(t, Status) ~ A + B + C + D + E

	Df	AIC	LRT	Pr(>Chi)
C	1	62.832	0.2673	0.6051622
<none>	*	64.565	*	*
A	1	65.762	3.1977	0.0737440.
E	1	66.480	3.9153	0.0478499*
B	1	73.712	11.1476	0.0008414***
D	1	75.960	13.3950	0.0002523***

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

Step: AIC = 62.83
Surv(t, Status) ~ A + B + D + E

	Df	AIC	LRT	Pr(>Chi)
<none>	*	62.832	*	*
A	1	63.821	2.9890	0.0838326 .
E	1	64.582	3.7504	0.0527960 .
B	1	71.713	10.8806	0.0009718 ***
D	1	73.960	13.1283	0.0002909 ***

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

Call: survreg(formula = Surv(t, Status) ~ A + B + D + E, data = data, dist = "weibull")
Coefficients:

(Intercept)	A	B	D	E
2.9759591	0.1827549	-0.4405758	0.4845426	-0.2056648

Scale= 0.138906
Loglik(model)= -25.4 Loglik(intercept only)= -33.9
Chisq= 16.97 on 4 degrees of freedom, p= 0.00196 n= 16

Όπως βλέπουμε η backward τεχνική με βήματα (Πίνακας 4.6) αφαίρεσε από το τελικό μοντέλο μόνο την συμμεταβλητή C, γιατί δεν είναι στατιστικά σημαντική και δεν επηρεάζει τη διάρκεια ζωής των λαμπτήρων μας σε μεγάλο βαθμό. Ενώ οι υπόλοιπες τέσσερις συμμεταβλητές (A,B,D και E) είναι στατιστικά σημαντικές, δηλαδή φαίνεται να επηρεάζουν σημαντικά τη διάρκεια ζωής τους. Αναλυτικά το τελικό μας μοντέλο φαίνεται στον Πίνακα 4.7.

Όπως αναμέναμε στο τελικό μας μοντέλο (Πίνακας 4.7) όλες οι μεταβλητές που περιέχονται έχουν πολύ μικρό p-value δηλαδή είναι στατιστικά σημαντικές. Ενώ βλέπουμε ξανά ότι ο κίνδυνος των λαμπτήρων συνεχίζει να αυξάνεται αφού η παράμετρος Scale (=1/η) =0.139 είναι πολύ μικρή, συνεπώς η παράμετρος σχήματος η είναι μεγαλύτερη της μονάδας.

Πίνακας 4.7

summary(modfinal)

Call: survreg(formula = Surv(t, Status) ~ A + B + D + E, data = data, dist = "weibull")

	Value	Std. Error	z	p
(Intercept)	2.9760	0.1025	29.02	<2e-16
A	0.1828	0.0958	1.91	0.056
B	-0.4406	0.1058	-4.16	3.1e-05
D	0.4845	0.1078	4.49	7.0e-06
E	-0.2057	0.0961	-2.14	0.032
Log(scale)	-1.9740	0.2825	-6.99	2.8e-12

Scale= 0.139
Weibull distribution
Loglik(model)= -25.4 Loglik(intercept only)= -33.9
Chisq= 16.97 on 4 degrees of freedom, p= 0.002
Number of Newton-Raphson Iterations: 7
n= 16

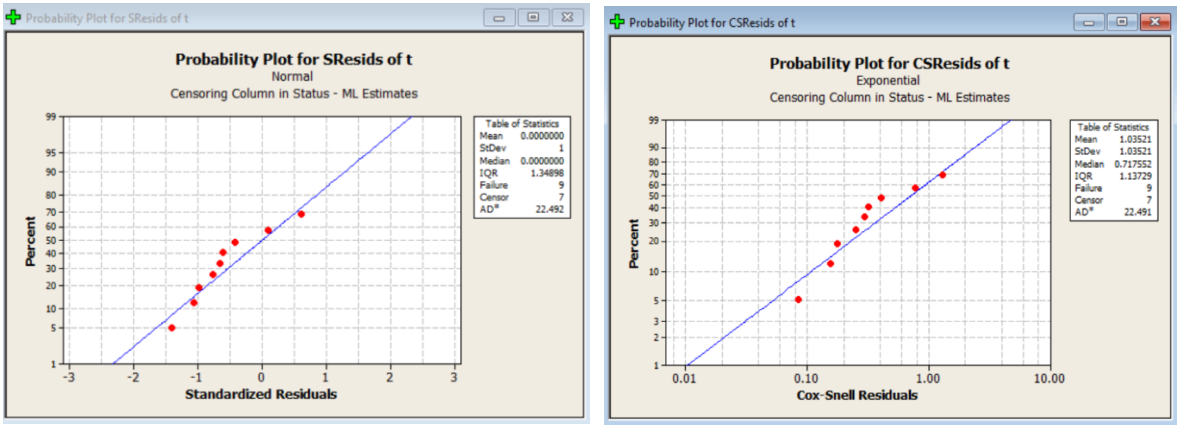
Επίσης, το τελικό μας μοντέλο θα έχει Διαστήματα Εμπιστοσύνης:

confint.default(modfinal)

	2.5 %	97.5 %
(Intercept)	2.774977694	3.17694050
A	-0.004986198	0.37049604
B	-0.647958845	-0.23319278
D	0.273206518	0.69587868
E	-0.394096758	-0.01723291

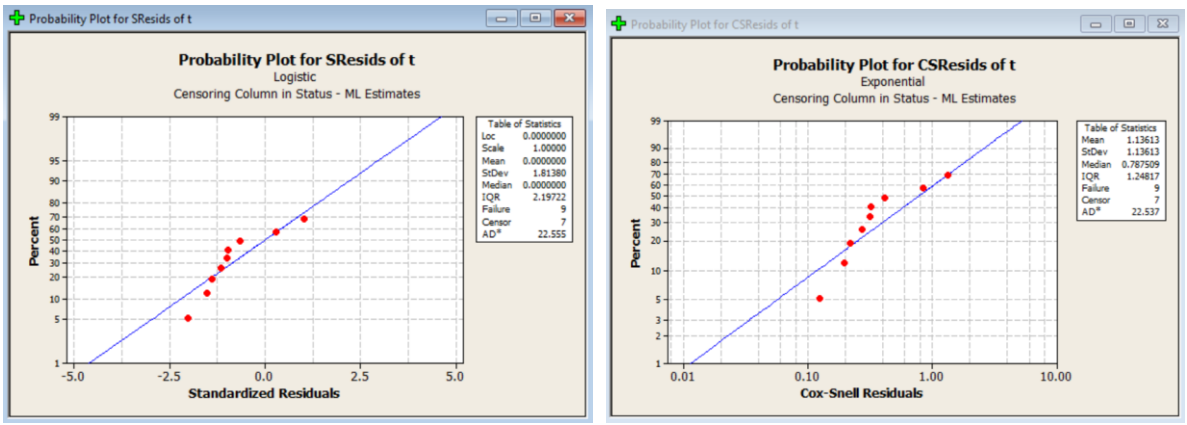
4.6.1 Υπόλοιπα Cox-Snell:

Τέλος θα ελέγξουμε γραφικά με την βοήθεια των υπολοίπων **Cox-Snell** (Παράγραφος 2.2.2) αν στα δεδομένα μας ταιριάζει ένα Μοντέλο Παλινδρόμησης Επιταχυνόμενης Διάρκειας ζωής (AL), με χρήση του Minitab. Θα κάνουμε τις γραφικές παραστάσεις των υπολοίπων Cox-Snell για το τελικό μοντέλο που μόλις βρήκαμε στην Παράγραφο 4.6 (Πίνακας 4.7). Θα υπολογίσουμε τα υπόλοιπα Cox-Snell για τις κατανομές Log-Normal και Log-Logistic, που όπως είδαμε στην Παράγραφο 4.5 είναι οι πιο κατάλληλες για τα δεδομένα μας.



Σχήμα 4.4: *Κανονικοποιημένα υπόλοιπα και υπόλοιπα Cox-Snell υπό την κατανομή Log-Normal.*

Το Σχήμα 4.4 παρουσιάζει τα κανονικοποιημένα υπόλοιπα και τα υπόλοιπα Cox-Snell υπό την κατανομή Log-Normal. Οι δύο γραφικές παραστάσεις είναι παρόμοιες και δείχνουν ότι τα περισσότερα σημεία πλησιάζουν αρκετά κοντά στη μπλε ευθεία που σχηματίζεται. Οπότε συμπεραίνουμε ότι το τελικό μας μοντέλο της Επιταχυνόμενης Διακοπής (AL) προσαρμόζεται αρκετά καλά στα δεδομένα των λαμπτήρων με την κατανομή Log-Normal.



Σχήμα 4.5: *Κανονικοποιημένα υπόλοιπα και υπόλοιπα Cox-Snell υπό την κατανομή Log-Logistic.*

Στο Σχήμα 4.5 παρουσιάζονται αυτή την φορά τα κανονικοποιημένα υπόλοιπα και τα υπόλοιπα Cox-Snell υπό την κατανομή Log-Logistic. Βλέπουμε ότι σε αυτή την περίπτωση τα σημεία και στις δύο γραφικές παραστάσεις αποκλίνουν αρκετά από την ευθεία, σε σχέση με το Σχήμα 2.4. Αυτό μας δείχνει ότι το τελικό μοντέλο της Επιταχυνόμενης Διακοπής (AL) δεν προσαρμόζεται τόσο καλά στα δεδομένα μας με την κατανομή Log-Logistic.

Οπότε συμπεραίνουμε ότι το τελικό μοντέλο της Επιταχυνόμενης Διακοπής (AL) που βρήκαμε στην Παράγραφο 4.5 είναι κατάλληλο για τα δεδομένα διάρκειας ζωής των λαμπτήρων αν το προσαρμόσουμε με την κατανομή Log-Normal.

4.6 Συμπεράσματα

Στην εφαρμογή που μελετήσαμε σε αυτό το κεφάλαιο, είχαμε ένα μικρό δείγμα που αποτελούνταν από 16 λαμπτήρες φθορισμού και μελετήσαμε τη διάρκεια ζωής τους σε ώρες μέχρι να συμβεί το δυσάρεστο γεγονός, όπου στη συγκεκριμένη περίπτωση είναι η διακοπή της λειτουργίας τους. Είχαμε πέντε επεξηγηματικές μεταβλητές και δεξιά αποκομμένες παρατηρήσεις που μας έδειχναν ποιοι από τους 16 λαμπτήρες συνέχισαν τη λειτουργία τους και μετά το πέρας της παρατήρησής μας.

Αρχικά, ελέγξαμε με διάφορες μεθόδους αν τα δεδομένα μας προσαρμόζονταν σύμφωνα με κάποιες γνωστές κατανομές (Weibull, Log-Normal και Log-Logistic) και καταλήξαμε στο συμπέρασμα ότι οι κατανομές Log-Normal και Log-Logistic φαίνονται οι πιο ιδανικές για τα δεδομένα μας σε σχέση με τη Weibull. Κάποιες από τις μεθόδους που χρησιμοποιήσαμε ήταν το κριτήριο AIC και ο γραφικός έλεγχος των συναρτήσεων της Μέγιστης Πιθανοφάνειας των κατανομών. Επίσης, με τη βοήθεια της εκτίμησης Kaplan-Meier είδαμε ότι η πιθανότητα «επιβίωσης» των λαμπτήρων μας φθίνει σημαντικά καθώς περνάει ο χρόνος. Επίσης όταν τελειώσει η παρατήρηση των λαμπτήρων μας, βρίσκονται σε κίνδυνο 9 λαμπτήρες με πιθανότητα να επιβιώσουν 0.4375.

Στη συνέχεια, προσαρμόσαμε στα δεδομένα μας ένα μοντέλο Επιταχυνόμενης Διακοπής (AL) με την κατανομή Weibull και με τη μέθοδο backward τεχνική με βήματα καταλήξαμε στο βέλτιστο μοντέλο όπου περιέχει μόνο τέσσερις μεταβλητές. Οι συντελεστές αυτού του βέλτιστου μοντέλου εμφανίζονται στον Πίνακα 4.8.

Πίνακας 4.8

	Value	Std. Error	z	p
(Intercept)	2.9760	0.1025	29.02	<2e-16
A	0.1828	0.0958	1.91	0.056
B	-0.4406	0.1058	-4.16	3.1e-05
D	0.4845	0.1078	4.49	7.0e-06
E	-0.2057	0.0961	-2.14	0.032
Log(scale)	-1.9740	0.2825	-6.99	2.8e-12

Η παράμετρος κλίμακας Scale (=1/η) του μοντέλου στον Πίνακα 4.8 είναι ίση με 0.139, το οποίο μας δείχνει ότι η παράμετρος σχήματος η θα είναι μεγαλύτερη της μονάδας, δηλαδή ο κίνδυνος διακοπής λειτουργίας των λαμπτήρων συνεχίζει να αυξάνεται. Οι συμμεταβλητές που επηρεάζουν πιο πολύ τη διάρκεια ζωής των λαμπτήρων φθορισμού που έχουμε είναι οι B και D καθώς έχουν τη μικρότερη τιμή p-value. Επίσης, για τις συμμεταβλητές B και E ο όρος $e^{-\beta' x}$ είναι μεγαλύτερος της μονάδας (exp(0.4406)= 1.5536 >1 και exp(0.2057)= 1.2283 >1), που σημαίνει ότι συμβάλουν στην επιτάχυνση της διακοπής της λειτουργίας των λαμπτήρων. Τέλος παρατηρούμε ότι ο σταθερός όρος έχει το μικρότερο p-value (<2e-16) δηλαδή είναι στατιστικά πιο σημαντικός και δεν πρέπει να τον παραλείψουμε από το μοντέλο μας.

Κεφάλαιο 5: Μεταμόσχευση Μυελού των Οστών

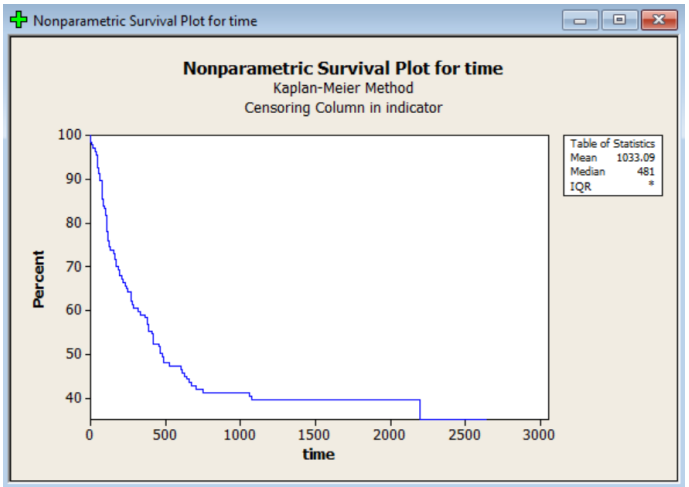
Σε αυτό το κεφάλαιο θα μελετήσουμε δεδομένα που αφορούν τη διάρκεια ζωής (time) σε μέρες 137 ασθενών, μέχρι την υποτροπή ή το θάνατο ενός ασθενούς μετά από μεταμόσχευση μυελού των οστών. Έχουμε 11 επεξηγηματικές μεταβλητές, οι οποίες είναι οι ακόλουθες:

indicator: 1: έχει συμβεί το γεγονός, 0: αλλιώς
group: τύπος λευχαιμίας (1=ALL, 2=AML low-risk, 3=AML high-risk)
recipient age: ηλικία του ασθενή
donor age: ηλικία του δότη
recipient sex: φύλο του ασθενή (1=άντρας, 0=γυναίκα)
donor sex: φύλο του δότη (1=άντρας, 0=γυναίκα)
recipient cmv: κατάσταση του cmv του ασθενή (1=θετικό cmv, 0=αρνητικό cmv)
donor cmv: κατάσταση του cmv του δότη (1=θετικό cmv, 0=αρνητικό cmv)
waiting time: χρόνος αναμονής σε μέρες από τη διάγνωση μέχρι τη μεταμόσχευση
fab: 1=fab βαθμού 4 ή 5, 0=αλλιώς, είναι μία ταξινόμηση των ασθενών με μυελοκυτταρική λευχαιμία (AML) που βασίζεται σε μορφολογικά κριτήρια, δηλαδή οι ασθενείς με fab βαθμού 4 ή 5 (M4 ή M5) διατρέχουν μεγαλύτερο κίνδυνο υποτροπής ή θανάτου μετά τη μεταμόσχευση του μυελού των οστών
mtx: 1=ναι, 0=όχι, μας δείχνει αν οι ασθενείς έλαβαν κάποια προφύλαξη για την ασθένεια μοσχεύματος έναντι ξενιστή (gvhd) μετά τη μεταμόσχευση ή όχι

5.1 Ανάλυση στο Σύνολο των Δεδομένων

5.1.1 Εκτίμηση Kaplan-Meier:

Με τη βοήθεια του στατιστικού πακέτου Minitab θα εφαρμόσουμε την εκτίμηση Kaplan-Meier στο σύνολο των ασθενών μας.



Σχήμα 5.1: *Εκτίμηση Kaplan-Meier της επιβίωσης του συνόλου των ασθενών.*

Στο Σχήμα 5.1 παρουσιάζεται η εκτίμηση της Kaplan-Meier για το σύνολο των ασθενών μας, όπου παρατηρούμε ότι όσο αυξάνεται ο χρόνος μειώνεται πολύ η πιθανότητα επιβίωσης των ασθενών μετά από την μεταμόσχευση μυελού των οστών. Στο τέλος της παρατήρησής μας (μετά από τις 2,500 ημέρες), από τις τιμές του Πίνακα Π1 (Παράρτημα Α), βρίσκονται σε κίνδυνο 9 ασθενείς με πιθανότητα να επιβιώσουν 0.350948, η οποία είναι αρκετά μικρή.

5.1.2 Έλεγχος Wald:

Θα εφαρμόσουμε έλεγχο **Wald** (Παράγραφος 2.5) για να ελέγξουμε ποια από τις Εκθετική και Weibull κατανομές προσαρμόζεται καλύτερα στο σύνολο των δεδομένων που έχουμε. Ο τύπος του ελέγχου Wald είναι:

$$z = \frac{\hat{\eta} - \eta_0}{se(\hat{\eta})} \sim N(0,1)$$

Η υπόθεση που ελέγχει ο Wald είναι: $H_0: \eta = 1, \quad H_1: \eta \neq 1$, όπου η είναι η παράμετρος σχήματος (Shape) της κατανομής Weibull (αν η =1 τότε έχουμε την Εκθετική κατανομή).

Παρατηρούμε ωστόσο ότι το Minitab εφαρμόζει τον έλεγχο Wald με τη διόρθωση όπου χρησιμοποιεί το $\ln(\hat{\eta})$ αντί για για το $\hat{\eta}$, δηλαδή δεν ελέγχει την υπόθεση που είπαμε αλλά ελέγχει την μηδενική υπόθεση $H_0: \ln(\eta) = 0$ με την εναλλακτική $H_1: \ln(\eta) \neq 0$. Οπότε ο τύπος του ελέγχου θα γίνει:

$$z = \frac{\ln(\hat{\eta}) - \ln(\eta)}{se(\ln(\hat{\eta}))} \sim N(0,1) \qquad \text{όπου} \quad se(\ln(\hat{\eta})) \cong \frac{se(\hat{\eta})}{\hat{\eta}}$$

Και το Διάστημα Εμπιστοσύνης (95%) θα υπολογίζεται από το τύπο:

$$\exp[\ln(\hat{\eta}) \pm (1.96) \times se(\ln(\hat{\eta}))]$$

Πίνακας 5.1

Parameter Estimates:

Parameter	Estimate	Standard Error	95,0% Normal CI Lower	Upper
Shape	0,587567	0,0559363	0,487554	0,708095
Scale	1471,71	283,130	1009,41	2145,74

Log-Likelihood = -657,767

Goodness-of-Fit

Anderson-Darling (adjusted) = 270,774

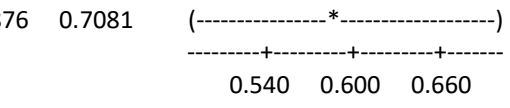
Test for Shape Equal to 1

Chi-Square	DF	P
31.2009	1	0.000

Bonferroni 95.0% (indiv 95.00%) Simultaneous CI

Shape parameter for

Variable	Lower	Estimate	Upper
time	0.4876	0.5876	0.7081



Από τις εκτιμήσεις του πίνακα 5.1 προκύπτει: $se(\ln(\hat{\eta})) = \frac{0.0559363}{0.587567} = 0.09519$ οπότε z = -5.58755 $\Rightarrow z^2$ = 31.2009 (Chi-Square) με βαθμό ελευθερίας 1 και όπως βλέπουμε έχει τιμή p-value $\ll 0.0001$.

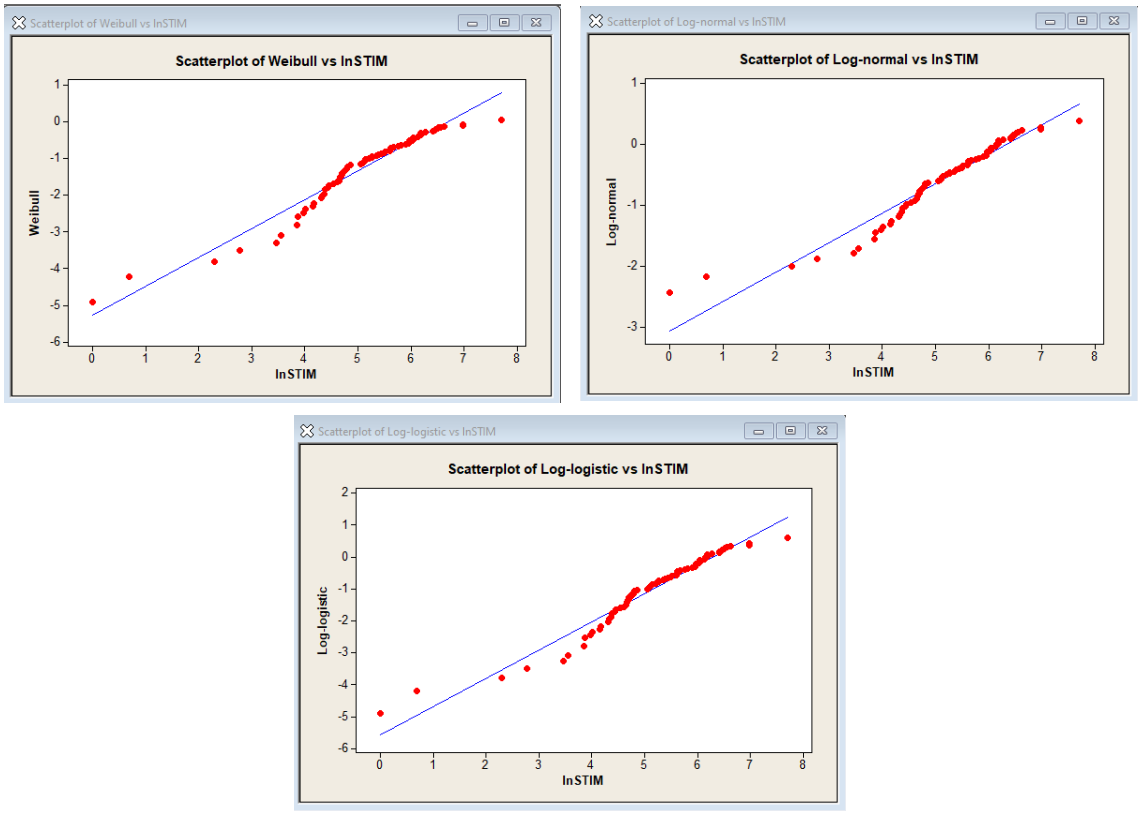
Απορρίπτουμε την υπόθεση H_0 (η=1) γιατί η p-value τιμή του ελέγχου Wald είναι < 0.001, δηλαδή είναι πολύ μικρή. Οπότε μεταξύ των δύο αυτών κατανομών καλύτερα προσαρμόζεται η κατανομή **Weibull**.

5.1.3 Γραφικές Παραστάσεις Κατανομών:

Θα συγκρίνουμε τις γραφικές παραστάσεις των τριών κατανομών Weibull, Log-Normal και Log-Logistic, τις τιμές των οποίων θα βρούμε με την βοήθεια της εκτίμησης Kaplan-Meier.
Υπολογίζουμε τις τιμές των γραφικών παραστάσεων από τους τύπους:

- Weibull: ln{ - ln(SKM) } για lnt
- Log-Normal: $\Phi^{-1}(1- SKM)$ για lnt
- Log-Logistic: ln{ (1- SKM) / SKM } για lnt

όπου SKM είναι η εκτίμηση Kaplan-Meier όπως την έχουμε υπολογίσει στον Πίνακα Π1 του Παραρτήματος Α με τη βοήθεια του Minitab.



Σχήμα 5.2: *Γραφικές παραστάσεις των τριών κατανομών συναρτήσει του λογαριθμισμένου χρόνου.*

Το Σχήμα 5.2 παρουσιάζει τρεις γραφικές παραστάσεις μία για κάθε κατανομή (Weibull, Log-Normal και Log-Logistic αντίστοιχα) και με μπλε γραμμή απεικονίζεται η εκτίμηση Kaplan-Meier. Στο Σχήμα αυτό βλέπουμε ότι και οι τρεις κατανομές προσαρμόζονται σχεδόν το ίδιο καλά στα δεδομένα που έχουμε. Αλλά παρατηρούμε ότι τα σημεία της Log-logistic συγκλίνουν καλύτερα στην μπλε γραμμή, δηλαδή η **Log-logistic** προσαρμόζεται καλύτερα στα δεδομένα μας.

5.1.4 Κριτήριο AIC:

Τώρα θα χρησιμοποιήσουμε και το **κριτήριο AIC** για να συγκρίνουμε ξανά αυτές τις τρεις κατανομές.

$$AIC = -2l + 2k$$

όπου k ο αριθμός των παραμέτρων που έχουμε και l η Λογαριθμισμένη Συνάρτηση Πιθανοφάνειας την οποία θα βρούμε εφαρμόζοντας την μέθοδο Μέγιστης Πιθανοφάνειας για κάθε κατανομή.

Estimation Method: Maximum Likelihood

- Distribution: Weibull

Parameter Estimates:

Parameter	Estimate	Standard Error	95,0% Normal CI Lower	Upper
Shape	0,587567	0,0559363	0,487554	0,708095
Scale	1471,71	283,130	1009,41	2145,74

Log-Likelihood = -657,767

Goodness-of-Fit

Anderson-Darling (adjusted) = 270,774

- Distribution: Lognormal

Parameter Estimates:

Parameter	Estimate	Standard Error	95,0% Normal CI Lower	Upper
Location	6,55537	0,208541	6,14664	6,96411
Scale	2,17723	0,185781	1,84193	2,57357

Log-Likelihood = -650,864

Goodness-of-Fit

Anderson-Darling (adjusted) = 270,237

- Distribution: Loglogistic

Parameter Estimates:

Parameter	Estimate	Standard Error	95,0% Normal CI Lower	Upper
Location	6,47988	0,200828	6,08627	6,87350
Scale	1,27454	0,118718	1,06186	1,52981

Log-Likelihood = -651,746

Goodness-of-Fit

Anderson-Darling (adjusted) = 270,274

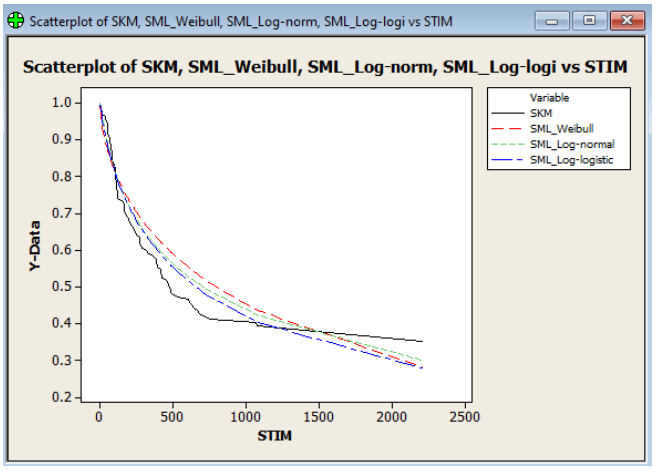
Εδώ έχουμε k=2, γιατί έχουμε 2 παραμέτρους (scale και location). Οπότε το **κριτήριο AIC** για τις τρεις κατανομές θα είναι:

- Weibull: -2*(-657,767) + 2*2 = 1.315,538
- Log-Normal: -2*(-650,864) + 2*2 = 1.301,732
- Log-Logistic: -2*(-651,746) + 2*2 = 1.303,496

Βλέπουμε ότι σύμφωνα με το κριτήριο **AIC** η κατανομή **Log-Normal** προσαρμόζεται καλύτερα στα δεδομένα που έχουμε γιατί έχει το μικρότερο AIC.

5.1.5 Εκτίμηση Μέγιστης Πιθανοφάνειας της Συνάρτησης Επιβίωσης:

Ο τρίτος και τελευταίος έλεγχος είναι μέσω της Εκτίμησης Μέγιστης Πιθανοφάνειας της συνάρτησης επιβίωσης, την οποία θα την υπολογίσουμε με τη βοήθεια του Minitab ξεχωριστά για κάθε μία από τις τρεις κατανομές.



Σχήμα 5.3: *Γραφική παράσταση της Kaplan-Meier και των τριών Συναρτήσεων Μέγιστης Πιθανοφάνειας συναρτήσει του χρόνου.*

Στο Σχήμα 5.3 παρουσιάζεται η γραφική παράσταση της εκτίμησης Kaplan-Meier και των τριών Συναρτήσεων Μέγιστης Πιθανοφάνειας, συναρτήσει του χρόνου. Σκοπός μας είναι να δούμε ποια καμπύλη πλησιάζει περισσότερο αυτή της Kaplan-Meier. Η καμπύλη της Kaplan-Meier εμφανίζεται με μαύρο χρώμα. Παρατηρούμε ότι και οι τρεις καμπύλες συγκλίνουν στην μαύρη σχεδόν το ίδιο καλά, όμως μπορώ να διακρίνω ότι αυτή της **Log-Normal** πλησιάζει την Kaplan-Meier καλύτερα από τις άλλες δύο.

Συμπέρασμα: Συνδυάζοντας τις τρεις διαφορετικές μεθόδους, που εφαρμόσαμε στις Παραγράφους 5.1.2, 5.1.3 και 5.1.4, καταλήγουμε ότι η κατανομή **Log-Normal** προσαρμόζεται καλύτερα στο σύνολο των δεδομένων που έχουμε.

5.2 Εκτίμηση Kaplan-Meier – Ανά Ομάδα

Στη συνέχεια θα εφαρμόσουμε τρεις διαφορετικές μεθόδους για να ελέγξουμε ποια κατανομή, ανάμεσα από τρεις βασικές κατανομές, προσαρμόζεται καλύτερα στα δεδομένα μας. Οι τρεις αυτές κατανομές είναι οι Weibull, Log-Normal και Log-Logistic. Αυτές τις μεθόδους θα τις εφαρμόσουμε πρώτα για το σύνολο των δεδομένων μας και έπειτα για τα δεδομένα κάθε ομάδας χωριστά. Όπου οι τρεις ομάδες όπως έχουμε αναφέρει είναι:

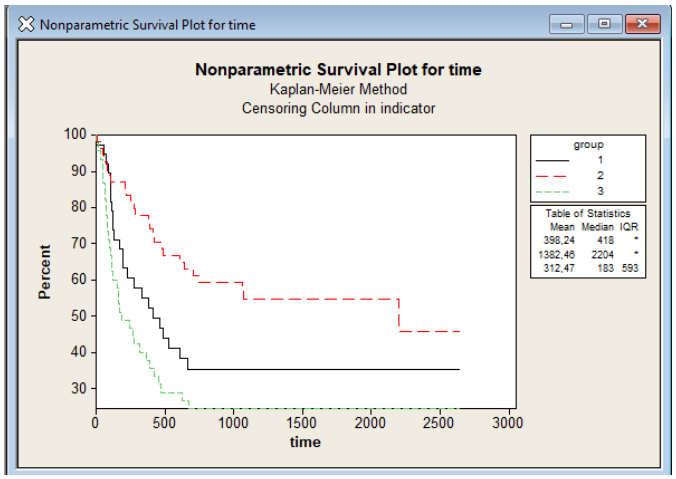
- η λεμφοβλαστική λευχαιμία ALL (Ομάδα 1)
- η μυελοκυτταρική λευχαιμία χαμηλού ρίσκου AML low_risk (Ομάδα 2)
- η οξεία μυελοκυτταρική λευχαιμία υψηλού ρίσκου AML high_risk (Ομάδα 3).

Σε αυτή την παράγραφο θα ασχοληθούμε ξεχωριστά με τις τρεις ομάδες που δημιουργεί η συμμεταβλητή group, δηλαδή με τις ALL (Ομάδα 1), AML low_risk (Ομάδα 2) και AML high_risk (Ομάδα 3). Στη συνέχεια θα εφαρμόσουμε την εκτίμηση Kaplan-Meier για κάθε μία από τις τρεις ομάδες χωριστά με τη βοήθεια του Minitab.

Πίνακας 5.2: Test Statistics

Method	Chi-Square	DF	P-Value
Log-Rank	13.8037	2	0.001
Wilcoxon	16.2407	2	0.000

Στον Πίνακα 5.2 βλέπουμε τους ελέγχους Log-rank και Wilcoxon που εφάρμοσε το Minitab, τους οποίους είδαμε αναλυτικά στην Παράγραφο 1.2.3. Υπενθυμίζουμε ότι ο έλεγχος Wilcoxon είναι η γενίκευση του Log-rank. Και οι δύο αυτοί έλεγχοι ελέγχουν τη μηδενική υπόθεση (H_0) που υποστηρίζει ότι το γεγονός (ο θάνατος στη συγκεκριμένη περίπτωση) είναι ανεξάρτητο της ομάδας στην οποία ανήκουν οι παρατηρήσεις μας, με εναλλακτική υπόθεση (H_1) ότι εξαρτάται από την ομάδα που ανήκουν. Όπως βλέπουμε στον Πίνακα 5.2, οι τιμές του X^2-ελέγχου είναι αρκετά μεγάλες, συνεπώς οι τιμές p-value που έχουν είναι πολύ μικρές (0.001 και 0.000). Οπότε απορρίπτουμε τη μηδενική υπόθεση (H_0) και δεχόμαστε την εναλλακτική που υποστηρίζει ότι το γεγονός (ο θάνατος) εξαρτάται από τις ομάδες που ανήκουν οι ασθενείς μας.



Σχήμα 5.4: *Εκτίμηση Kaplan-Meier της επιβίωσης για κάθε ομάδα.*

Όπως βλέπουμε στο Σχήμα 5.4 τις τρεις εκτιμήσεις Kaplan-Meier για τις τρεις ομάδες, παρατηρούμε ότι η πιθανότητα επιβίωσης των ασθενών που ανήκουν στην Ομάδα 3 μειώνονται πολύ στις πρώτες 500 ημέρες από τη μεταμόσχευση. Ενώ αυτοί που ανήκουν στην Ομάδα 2 φαίνεται να έχουν μία πιο σταθερή κατάσταση καθώς ξεπερνούν τις 1,000 ημέρες από τη μεταμόσχευση και συνεπώς έχουν μεγαλύτερες πιθανότητες να επιβιώσουν σε βάθος χρόνου. Η ομάδα που επιβιώνει καλύτερα είναι προφανώς η δεύτερη ομάδα με τους ασθενείς που έχουν AML low_risk.

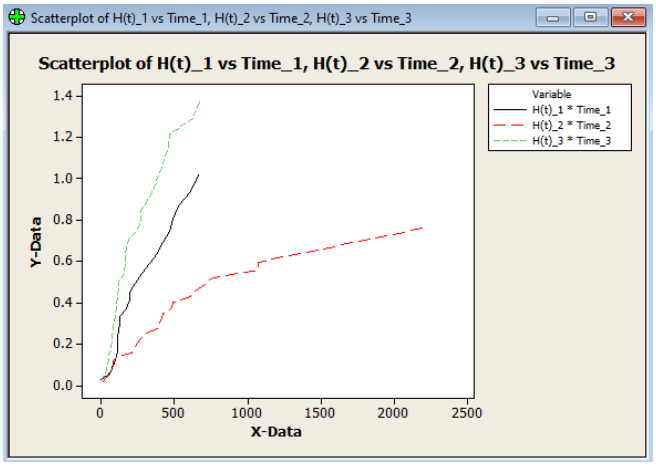
Όταν τελειώσει η παρατήρησή μας, από την Ομάδα 1 θα βρίσκονται σε κίνδυνο 13 ασθενείς με πιθανότητα να επιβιώσουν 0.353057 (πολύ μικρή) (Πίνακας Π4, Παράρτημα Α), από την Ομάδα 2 βρίσκονται σε κίνδυνο μόλις 2 ασθενείς με πιθανότητα να επιβιώσουν 0.455840 (Πίνακας Π5, Παράρτημα Α) και τέλος από την Ομάδα 3 βρίσκονται σε κίνδυνο 12 ασθενείς με πιθανότητα επιβίωσης 0.350948 (πολύ μικρή) (Πίνακας Π6, Παράρτημα Α).

5.3 Εκτίμηση Nelson-Aalen

Υπενθυμίζουμε ότι η εκτίμηση Nelson-Aalen (Παράγραφος 1.2.2) υπολογίζεται από τον τύπο:

$$\hat{H}(t) = \sum_{j:\ t_{(j)} \le t} \frac{d_j}{n_j}$$

Ήδη από τα αποτελέσματα της Kaplan-Meier για τις τρεις ομάδες χωριστά έχουμε βρει τις τιμές των d_j και n_j για την κάθε ομάδα (Πίνακες Π4, Π5, Π6 Παραρτήματος Α). Οπότε υπολογίζουμε τις τιμές της εκτίμησης Nelson-Aalen για τις τρεις ομάδες και δημιουργούμε τη γραφική παράσταση του Σχήματος 5.5.



Σχήμα 5.5: *Εκτίμηση Nelson-Aalen για κάθε ομάδα συναρτήσει του χρόνου.*

Το Σχήμα 5.5 παρουσιάζει της εκτίμηση Nelson-Aalen για κάθε μία από τις τρεις ομάδες χωριστά. Όσο πιο απότομα αυξάνονται οι καμπύλες του Σχήματος 5.5, τόσο αυξάνεται ο κίνδυνος. Παρατηρούμε ότι η ομάδα που κινδυνεύει περισσότερο είναι η Ομάδα 3, ενώ αυτή που κινδυνεύει λιγότερο είναι η Ομάδα 2. Αυτό το συμπέρασμα είναι λογικό αφού ήδη από την εκτίμηση Kaplan-Meier έχουμε δει ότι οι ασθενείς που ανήκουν στην Ομάδα 3 έχουν μικρότερες πιθανότητες να επιβιώσουν (Σχήμα 5.4).

5.4 Ανάλυση στην Ομάδα 1

Θα εφαρμόσουμε και πάλι τις τρεις διαφορετικές μεθόδους για να ελέγξουμε ποια κατανομή από τις τρεις, Weibull, Log-Normal και Log-Logistic, προσαρμόζεται καλύτερα στα δεδομένα της Ομάδας 1, δηλαδή στους ασθενείς με λεμφοβλαστική λευχαιμία (ALL).

<u>Γραφικές Παραστάσεις Κατανομών:</u>

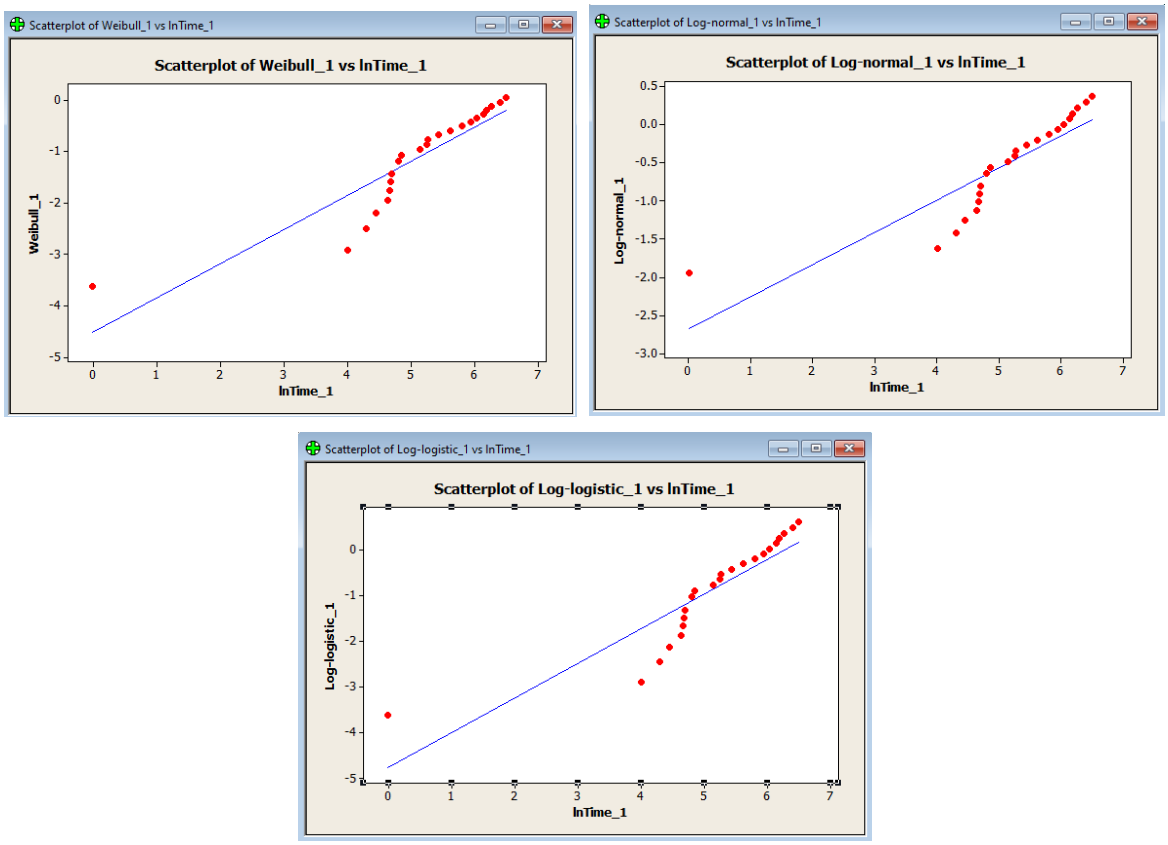
Υπολογίζουμε τις γραφικές παραστάσεις των τριών κατανομών για την Ομάδα 1 από τους τύπους:

- Weibull: $\ln\{ -\ln(SKM) \}$ για $\ln t$
- Log-Normal: $\Phi^{-1}(1- SKM)$ για $\ln t$
- Log-Logistic: $\ln\{ (1- SKM) / SKM \}$ για $\ln t$

Όπου SKM είναι η εκτίμηση Kaplan-Meier της Ομάδας 1 όπως την έχουμε βρει στους Πίνακες Π4, Π5, Π6 Παραρτήματος Α, με τη βοήθεια του Minitab.

Πίνακας 5.3

Time_1	Weibull_1	Log-normal_1	Log-logistic_1
1	-3.62427	-1.93793	-3.61091
55	-2.91752	-1.61985	-2.89036
74	-2.49814	-1.41219	-2.45674
86	-2.19620	-1.25212	-2.14007
104	-1.95844	-1.11896	-1.88707
107	-1.76113	-1.00315	-1.67397
109	-1.59160	-0.89943	-1.48807
110	-1.44228	-0.80460	-1.32176
122	-1.18619	-0.63364	-1.02962
129	-1.07368	-0.55492	-0.89794
172	-0.96893	-0.47951	-0.77319
192	-0.87058	-0.40672	-0.65393
194	-0.77755	-0.33604	-0.53900
230	-0.68521	-0.26402	-0.42266
276	-0.59692	-0.19335	-0.30907
332	-0.51200	-0.12364	-0.19743
383	-0.42988	-0.05452	-0.08701
418	-0.35005	0.01434	0.02288
466	-0.27205	0.08327	0.13292
487	-0.19548	0.15259	0.24376
526	-0.11994	0.22266	0.35612
609	-0.03969	0.29898	0.47904
662	0.04030	0.37708	0.60563



Σχήμα 5.6: *Γραφικές παραστάσεις των τριών κατανομών συναρτήσει του λογαριθμισμένου χρόνου της Ομάδας 1.*

Το Σχήμα 5.6 παρουσιάζει τρεις γραφικές παραστάσεις για κάθε μία από τις κατανομές Weibull, Log-Normal και Log-Logistic αντίστοιχα και με μπλε γραμμή απεικονίζεται η εκτίμηση Kaplan-Meier. Γενικά και οι τρεις γραφικές δεν είναι καλές γιατί καμία τους δεν φαίνεται να συγκλίνει ικανοποιητικά στην ευθεία της Kaplan-Meier. Αλλά μπορούμε να διακρίνουμε ότι οι τιμές της **Log-Normal** πλησιάζουν καλύτερα την ευθεία σε σχέση με τις τιμές των άλλων δύο κατανομών.

Κριτήριο AIC:

Θα εργαστούμε όπως και στην Παράγραφο 5.1.4.

Estimation Method: Maximum Likelihood

Variable: time
group = 1
Estimation Method: Maximum Likelihood
Distribution: **Weibull**

Parameter Estimates

		Standard	95,0% Normal CI	
Parameter	Estimate	Error	Lower	Upper
Shape	0,659704	0,115872	0,467565	0,930800
Scale	1043,68	329,343	562,295	1937,19

Log-Likelihood = -185,535
Goodness-of-Fit
Anderson-Darling (adjusted) = 82,695

Variable: time
group = 1
Estimation Method: Maximum Likelihood
Distribution: **Lognormal**

Parameter Estimates

		Standard	95,0% Normal CI	
Parameter	Estimate	Error	Lower	Upper
Location	6,30753	0,362786	5,59648	7,01858
Scale	2,02396	0,317288	1,48855	2,75196

Log-Likelihood = -184,569
Goodness-of-Fit
Anderson-Darling (adjusted) = 82,637

Variable: time
group = 1
Estimation Method: Maximum Likelihood
Distribution: **Loglogistic**

Parameter Estimates

		Standard	95,0% Normal CI	
Parameter	Estimate	Error	Lower	Upper
Location	6,22397	0,325317	5,58636	6,86158
Scale	1,10527	0,193116	0,784772	1,55666

Log-Likelihood = -183,773
Goodness-of-Fit
Anderson-Darling (adjusted) = 82,554

Έχουμε k=2. Οπότε το **κριτήριο AIC** για τις τρεις κατανομές θα είναι:

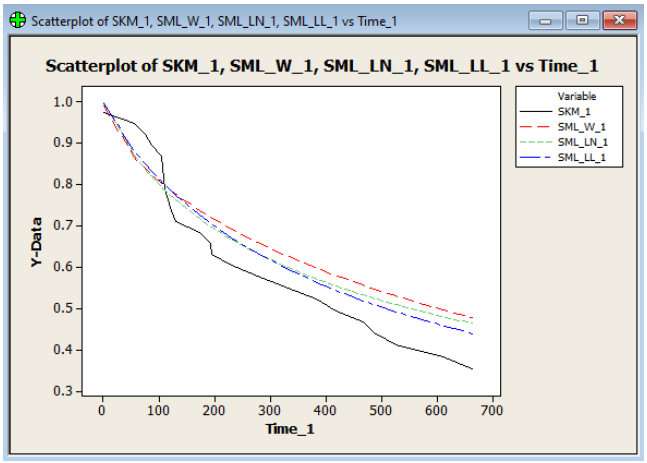
- Weibull: -2*(-185,535) + 2*2 = 371,074
- Log-Normal: -2*(-184,569) + 2*2 = 396,142
- Log-Logistic: -2*(-183,773) + 2*2 = 367,550

Σύμφωνα με το κριτήριο AIC η κατανομή **Log-Logistic** προσαρμόζεται καλύτερα στα δεδομένα της Ομάδας 1 καθώς έχει το μικρότερο AIC.

Εκτίμηση Μέγιστης Πιθανοφάνειας:

Πίνακας 5.4

Time_1	SML_W_1	SML_LN_1	SML_LL_1
1	0.989851	0.999085	0.996428
55	0.866346	0.872122	0.881375
74	0.839883	0.838882	0.850311
86	0.824747	0.820067	0.832169
104	0.803791	0.794383	0.806766
107	0.800474	0.790361	0.802724
109	0.798287	0.787717	0.800057
110	0.797201	0.786405	0.798732
122	0.784529	0.771215	0.783250
129	0.777424	0.762785	0.774558
172	0.737576	0.716729	0.725906
192	0.720870	0.698050	0.705665
194	0.719253	0.696262	0.703714
230	0.691629	0.666249	0.670631
276	0.659792	0.632882	0.633226
332	0.625174	0.598019	0.593618
383	0.596807	0.570489	0.562091
418	0.578791	0.553462	0.542531
466	0.555733	0.532162	0.518038
487	0.546184	0.523495	0.508075
526	0.529230	0.508323	0.490652
609	0.496134	0.479453	0.457612
662	0.476839	0.463048	0.438942



Σχήμα 5.7: *Γραφική παράσταση της εκτίμησης Kaplan-Meier και των τριών Συναρτήσεων Μέγιστης Πιθανοφάνειας συναρτήσει του χρόνου της Ομάδας 1.*

Στο Σχήμα 5.7 βλέπουμε την γραφική παράσταση της εκτίμησης Kaplan-Meier και των τριών Συναρτήσεων Μέγιστης Πιθανοφάνειας, συναρτήσει του χρόνου. Σκοπός μας είναι να δούμε ποια καμπύλη πλησιάζει περισσότερο αυτή της Kaplan-Meier. Παρατηρούμε ότι η κατανομή **Log-Logistic** πλησιάζει καλύτερα την μαύρη καμπύλη σε σχέση με τις άλλες δύο κατανομές.

Συμπέρασμα: Συνδυάζοντας τις τρεις διαφορετικές μεθόδους, καταλήγουμε στο ότι η κατανομή **Log-Logistic** προσαρμόζεται καλύτερα στα δεδομένα της Ομάδας 1.

5.5 Ανάλυση στην Ομάδα 2

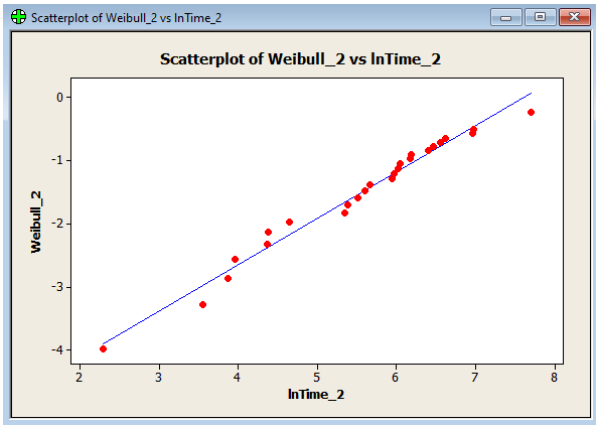
Θα εργαστούμε με τον ίδιο τρόπο (όπως στην Παράγραφο 5.4) και για τα δεδομένα της Ομάδας 2, δηλαδή στους ασθενείς που έχουν AML low_risk.

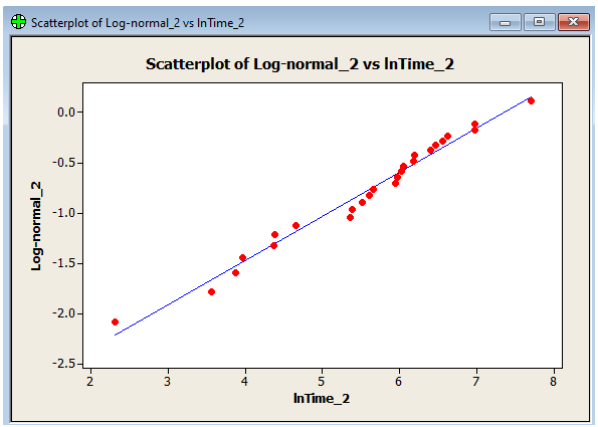
Γραφικές Παραστάσεις Κατανομών:

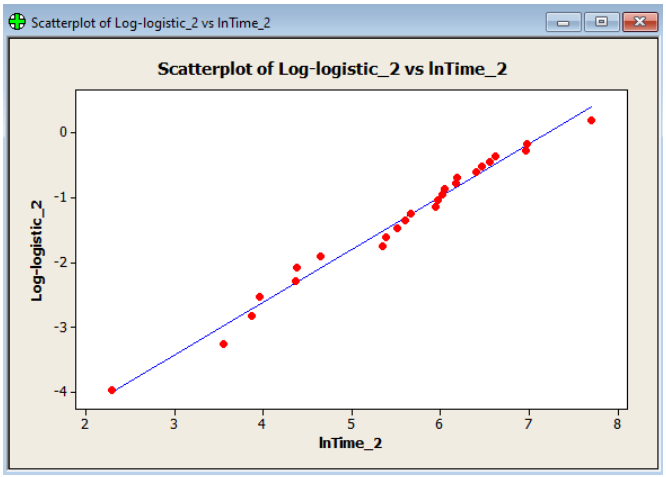
Υπολογίζουμε όπως και στις προηγούμενες παραγράφους τις τιμές των τριών γραφικών παραστάσεων με την βοήθεια της εκτίμησης Kaplan-Meier.

Πίνακας 5.5

Time_2	Weibull_2	Log-normal_2	Log-logistic_2
10	-3.97963	-2.08534	-3.97027
35	-3.27703	-1.78616	-3.25810
48	-2.86192	-1.59321	-2.83320
53	-2.56446	-1.44610	-2.52573
79	-2.33135	-1.32496	-2.28238
80	-2.13891	-1.22064	-2.07944
105	-1.97446	-1.12814	-1.90423
211	-1.83044	-1.04441	-1.74920
219	-1.70198	-0.96742	-1.60944
248	-1.58575	-0.89578	-1.48161
272	-1.47936	-0.82846	-1.36330
288	-1.38105	-0.76471	-1.25276
381	-1.28949	-0.70392	-1.14862
390	-1.20363	-0.64563	-1.04982
414	-1.12263	-0.58946	-0.95551
421	-1.04584	-0.53508	-0.86500
481	-0.97269	-0.48225	-0.77770
486	-0.90272	-0.43073	-0.69315
606	-0.83555	-0.38033	-0.61091
641	-0.77084	-0.33087	-0.53063
704	-0.70831	-0.28222	-0.45198
748	-0.64770	-0.23422	-0.37470
1063	-0.57542	-0.17587	-0.28104
1074	-0.50536	-0.11811	-0.18859
2204	-0.24129	0.11092	0.17710







Σχήμα 5.8: *Γραφικές παραστάσεις κατανομών συναρτήσει του λογαριθμισμένου χρόνου για την Ομάδα 2.*

Όπως βλέπουμε στο Σχήμα 5.8 όλες οι κατανομές συγκλίνουν αρκετά καλά στην ευθεία της Kaplan-Meier. Όμως παρατηρούμε ότι με μικρή διαφορά η κατανομή **Log-Normal** προσαρμόζεται καλύτερα στους ασθενείς της Ομάδας 2.

Κριτήριο AIC:

Estimation Method: Maximum Likelihood

Variable: time
group = 2
Estimation Method: Maximum Likelihood
Distribution: **Weibull**

Parameter Estimates

Parameter	Estimate	Standard Error	95,0% Normal CI Lower	Upper
Shape	0,643082	0,116572	0,450783	0,917413
Scale	3142,08	1117,36	1565,01	6308,35

Log-Likelihood = -214,989
Goodness-of-Fit
Anderson-Darling (adjusted) = 144,915

Variable: time
group = 2
Estimation Method: Maximum Likelihood
Distribution: **Lognormal**

Parameter Estimates

Parameter	Estimate	Standard Error	95,0% Normal CI Lower	Upper
Location	7,50415	0,393445	6,73301	8,27529
Scale	2,24867	0,358109	1,64577	3,07244

Log-Likelihood = -213,580
Goodness-of-Fit
Anderson-Darling (adjusted) = 144,861

Variable: time
group = 2
Estimation Method: Maximum Likelihood
Distribution: **Loglogistic**

Parameter Estimates

Parameter	Estimate	Standard Error	95,0% Normal CI Lower	Upper
Location	7,45242	0,363633	6,73972	8,16513
Scale	1,30338	0,227721	0,925450	1,83565

Log-Likelihood = -214,200
Goodness-of-Fit
Anderson-Darling (adjusted) = 144,882

Έχουμε k=2. Οπότε το **κριτήριο AIC** για τις τρεις κατανομές θα είναι:

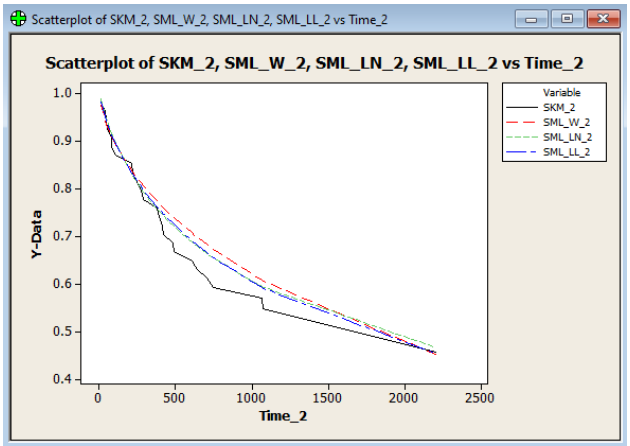
- Weibull: -2*(-214,989) + 2*2 = 429,982
- Log-Normal: -2*(-213,580) + 2*2 = 427,164
- Log-Logistic: -2*(-214,200) + 2*2 = 428,404

Σύμφωνα με το κριτήριο AIC η κατανομή **Log-Normal** προσαρμόζεται καλύτερα στα δεδομένα της Ομάδας 2.

Εκτίμηση Μέγιστης Πιθανοφάνειας:

Πίνακας 5.6

Time_2	SML_W_2	SML_LN_2	SML_LL_2
10	0.975525	0.989643	0.981130
35	0.946052	0.960461	0.952119
48	0.934309	0.946910	0.939780
53	0.930141	0.941971	0.935331
79	0.910636	0.918345	0.914148
80	0.909944	0.917497	0.913388
105	0.893680	0.897512	0.895396
211	0.838552	0.830752	0.833637
219	0.834983	0.826543	0.829640
248	0.822540	0.812000	0.815728
272	0.812764	0.800724	0.804835
288	0.806479	0.793547	0.797855
381	0.772994	0.756267	0.761012
390	0.769989	0.753000	0.757739
414	0.762150	0.744540	0.749229
421	0.759909	0.742137	0.746805
481	0.741472	0.722639	0.726993
486	0.739994	0.721096	0.725416
606	0.706788	0.687213	0.690440
641	0.697826	0.678316	0.681158
704	0.682398	0.663233	0.665337
748	0.672109	0.653336	0.654902
1063	0.607687	0.594080	0.591704
1074	0.605680	0.592303	0.589795
2204	0.451089	0.465646	0.453030



Σχήμα 5.9: *Γραφική παράσταση της Kaplan-Meier και των τριών Συναρτήσεων Μέγιστης Πιθανοφάνειας συναρτήσει του χρόνου της Ομάδας 2.*

Στο Σχήμα 5.9 παρατηρούμε ότι και οι τρεις Εκτιμήτριες Μέγιστης Πιθανοφάνειας είναι πανομοιότυπες, όμως φαίνεται ότι η καμπύλη Μέγιστης Πιθανοφάνειας της **Log-Logistic** πλησιάζει λίγο καλύτερα αυτή της Kaplan-Meier.

Συμπέρασμα: Συνδυάζοντας τις τρεις διαφορετικές μεθόδους, καταλήγουμε στο ότι η κατανομή **Log-Normal** προσαρμόζεται καλύτερα στα δεδομένα της Ομάδας 2.

5.6 Ανάλυση στην Ομάδα 3

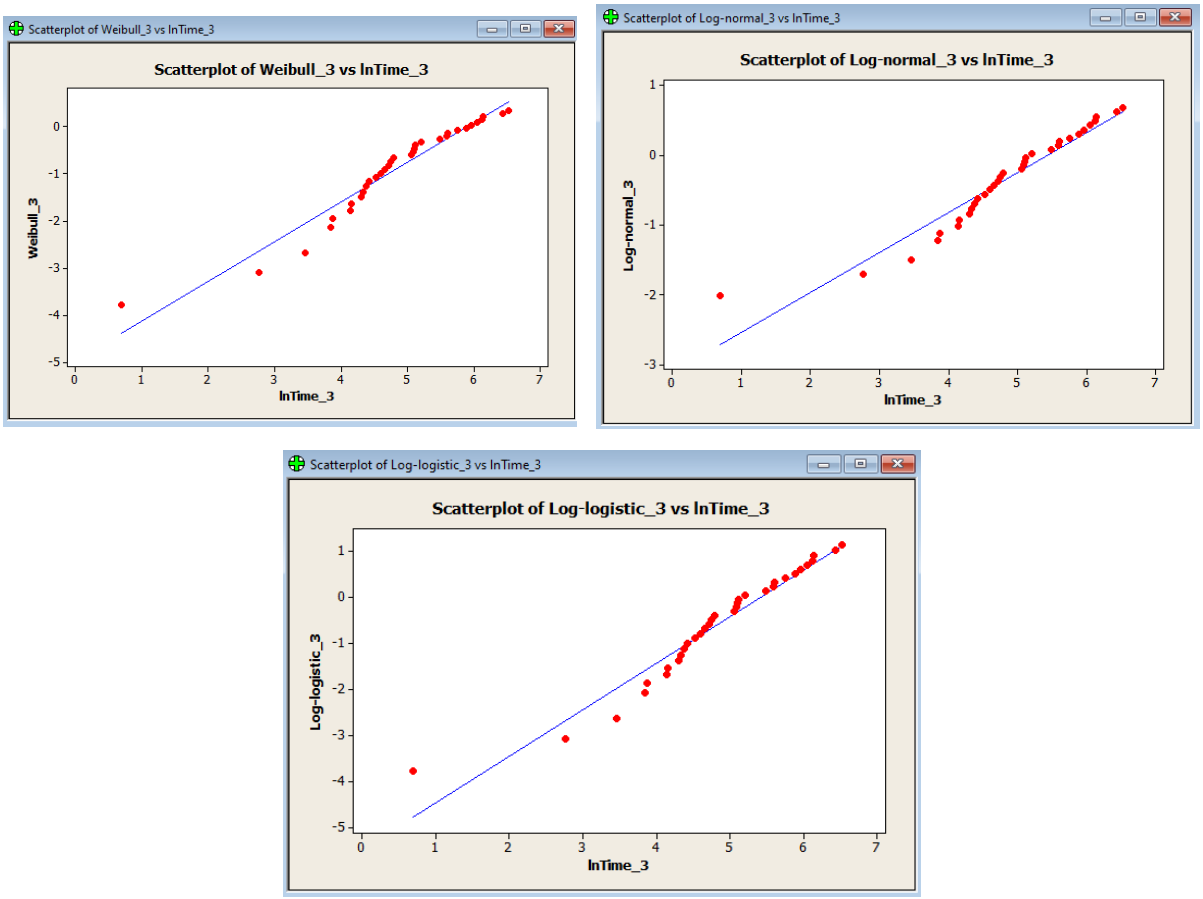
Θα εργαστούμε το ίδιο (όπως και στη Παράγραφο 5.4) και για τα δεδομένα της Ομάδας 3 δηλαδή στους ασθενείς που έχουν AML high_risk.

Γραφικές Παραστάσεις Κατανομών:

Υπολογίζουμε όπως και πριν τις τιμές των τριών γραφικών παραστάσεων με την βοήθεια της εκτίμησης Kaplan-Meier.

Πίνακας 5.7

Time_3	Weibull_3	Log-normal_3	Log-logistic_3
2	-3.79546	-2.00988	-3.78420
16	-3.09088	-1.70129	-3.06806
32	-2.67375	-1.50108	-2.63905
47	-2.13891	-1.22064	-2.07944
48	-1.94421	-1.11077	-1.87181
63	-1.77740	-1.01289	-1.69167
64	-1.63094	-0.92387	-1.53147
74	-1.49994	-0.84162	-1.38629
76	-1.38105	-0.76471	-1.25276
80	-1.27189	-0.69208	-1.12847
84	-1.17068	-0.62292	-1.01160
93	-1.07609	-0.55663	-0.90079
100	-0.98705	-0.49270	-0.79493
105	-0.90272	-0.43073	-0.69315
113	-0.82242	-0.37036	-0.59471
115	-0.74558	-0.31132	-0.49899
120	-0.67173	-0.25335	-0.40547
157	-0.60045	-0.19621	-0.31366
162	-0.53139	-0.13971	-0.22315
164	-0.46425	-0.08365	-0.13353
168	-0.39874	-0.02785	-0.04445
183	-0.33461	0.02785	0.04445
242	-0.27163	0.08365	0.13353
268	-0.20957	0.13971	0.22315
273	-0.14824	0.19621	0.31366
318	-0.08742	0.25335	0.40547
363	-0.02691	0.31132	0.49899
390	0.03350	0.37036	0.59471
422	0.09405	0.43073	0.69315
456	0.15496	0.49270	0.79493
467	0.21649	0.55663	0.90079
625	0.27896	0.62292	1.01160
677	0.34272	0.69208	1.12847



Σχήμα 5.10: *Γραφικές παραστάσεις κατανομών συναρτήσει του λογαριθμισμένου χρόνου για την Ομάδα 3.*

Στο Σχήμα 5.10 παρουσιάζονται οι τρεις γραφικές παραστάσεις των τριών κατανομών (Weibull, Log-Normal και Log-Logistic) για τους ασθενείς της Ομάδας 3. Βλέπουμε ότι οι τιμές της Weibull αποκλίνουν περισσότερο από αυτές των άλλων δύο. Ενώ δεν μπορούμε να διακρίνουμε ποια από τις **Log-Normal** και **Log-Logistic** προσαρμόζεται καλύτερα.

Κριτήριο AIC:

Estimation Method: Maximum Likelihood

Variable: time
group = 3
Estimation Method: Maximum Likelihood
Distribution: **Weibull**

Parameter Estimates

Parameter	Estimate	Standard Error	95,0% Normal CI Lower	Upper
Shape	0,573476	0,0802226	0,435956	0,754377
Scale	682,102	203,999	379,558	1225,80

Log-Likelihood = -249,855
Goodness-of-Fit
Anderson-Darling (adjusted) = 57,150

Variable: time
group = 3
Estimation Method: Maximum Likelihood
Distribution: **Lognormal**

Parameter Estimates

Parameter	Estimate	Standard Error	95,0% Normal CI Lower	Upper
Location	5,70792	0,297656	5,12453	6,29132
Scale	1,91958	0,249445	1,48797	2,47639

Log-Likelihood = -245,338
Goodness-of-Fit
Anderson-Darling (adjusted) = 56,672

Variable: time
group = 3
Estimation Method: Maximum Likelihood
Distribution: **Loglogistic**

Parameter Estimates

Parameter	Estimate	Standard Error	95,0% Normal CI Lower	Upper
Location	5,58127	0,286447	5,01984	6,14270
Scale	1,09979	0,159648	0,827458	1,46175

Log-Likelihood = -244,989
Goodness-of-Fit
Anderson-Darling (adjusted) = 56,552

Οπότε το **κριτήριο AIC** για τις τρεις κατανομές θα είναι:

- Weibull: -2*(-249,855) + 2*2 = 1.337,534
- Log-Normal: -2*(-245,338) + 2*2 = 1.323,728
- Log-Logistic: -2*(-244,989) + 2*2 = 1.325,492

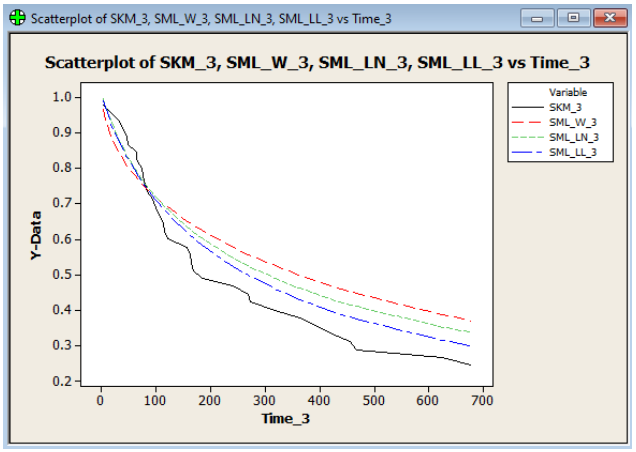
Άρα, σύμφωνα με το κριτήριο AIC η κατανομή **Log-Logistic** προσαρμόζεται καλύτερα στα δεδομένα της Ομάδας 3.

Εκτίμηση Μέγιστης Πιθανοφάνειας:

Πίνακας 5.8

Time_3	SML_W_2	SML_LN_2	SML_LL_2
2	0.965338	0.995505	0.988395
16	0.890254	0.936887	0.927831
32	0.841146	0.878609	0.872538
47	0.806012	0.833428	0.828360
48	0.803903	0.830675	0.825621
63	0.774829	0.792513	0.787119
64	0.773037	0.790157	0.784710
74	0.755951	0.767713	0.761572
76	0.752699	0.763449	0.757141
80	0.746342	0.755130	0.748463
84	0.740172	0.747072	0.740019
93	0.726907	0.729824	0.721823
100	0.717120	0.717177	0.708381
105	0.710386	0.708517	0.699133
113	0.700015	0.695257	0.684905
115	0.697495	0.692049	0.681452
120	0.691314	0.684207	0.672994
157	0.650069	0.632879	0.617131
162	0.645010	0.626712	0.610373
164	0.643016	0.624289	0.607717
168	0.639077	0.619517	0.602481
183	0.624854	0.602437	0.583717
242	0.575814	0.545413	0.520976
268	0.556974	0.524287	0.497791
273	0.553511	0.520452	0.493590
318	0.524368	0.488752	0.458998
363	0.498343	0.461305	0.429297
390	0.483976	0.446495	0.413393
422	0.467998	0.430303	0.396120
456	0.452126	0.414506	0.379395

467	0.447213	0.409673	0.374306
625	0.386317	0.351898	0.314585
677	0.369464	0.336572	0.299131



Σχήμα 5.11: *Γραφική παράσταση της Kaplan-Meier και των τριών Συναρτήσεων Μέγιστης Πιθανοφάνειας συναρτήσει του χρόνου της Ομάδας 3.*

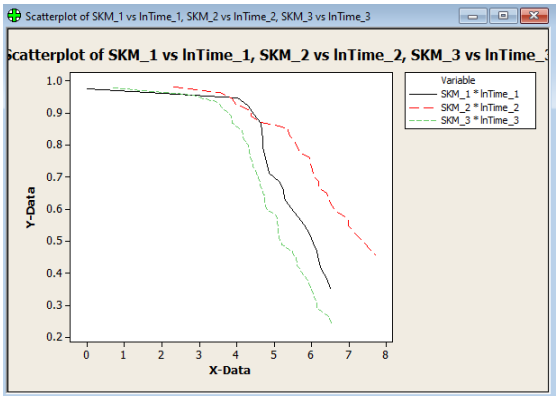
Στο Σχήμα 5.11 βλέπουμε ότι η εκτίμηση Μέγιστης Πιθανοφάνειας της **Log-Logistic** πλησιάζει καλύτερα την καμπύλη της Kaplan-Meier, σε σχέση με τις εκτιμήσεις των άλλων δύο κατανομών.

Συμπέρασμα: Συνδυάζοντας τις τρεις διαφορετικές μεθόδους, καταλήγουμε στο ότι η κατανομή **Log-Logistic** προσαρμόζεται καλύτερα στα δεδομένα που ανήκουν στην Ομάδα 3.

5.7 Μοντέλο Παλινδρόμησης Επιταχυνόμενης Διάρκειας Ζωής (AL)

Σε αυτή τη παράγραφο θα ελέγξουμε γραφικά με την βοήθεια του Minitab, αν στα δεδομένα μας ταιριάζει ένα μοντέλο Παλινδρόμησης Επιταχυνόμενης Διάρκειας Ζωής (AL), χρησιμοποιώντας ως κατηγορική μεταβλητή την μεταβλητή group. Υπενθυμίζουμε ότι το μοντέλο της Επιταχυνόμενης Διάρκειας Ζωής (AL) ή Επιταχυνόμενης Διακοπής το έχουμε δει αναλυτικά στην Παράγραφο 2.2.

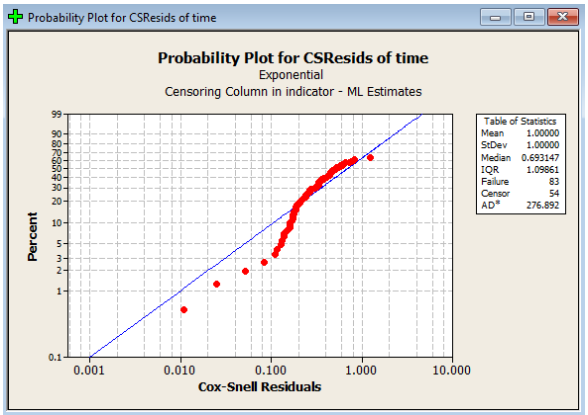
Για να ελέγξουμε το μοντέλο Επιταχυνόμενης Διάρκειας Ζωής (AL) θα κάνουμε τη γραφική παράσταση της εκτίμησης Kaplan-Meier της κάθε ομάδας, συναρτήσει του λογαριθμισμένου χρόνου της κάθε ομάδας αντίστοιχα. Και στη συνέχεια θα δούμε αν οι τρεις καμπύλες που θα σχηματιστούν (μία για κάθε ομάδα) διατηρούν ίσες αποστάσεις μεταξύ τους.



Σχήμα 5.12: *Γραφική παράσταση των εκτιμητριών Kaplan-Meier της κάθε ομάδας συναρτήσει του λογαριθμισμένου χρόνου.*

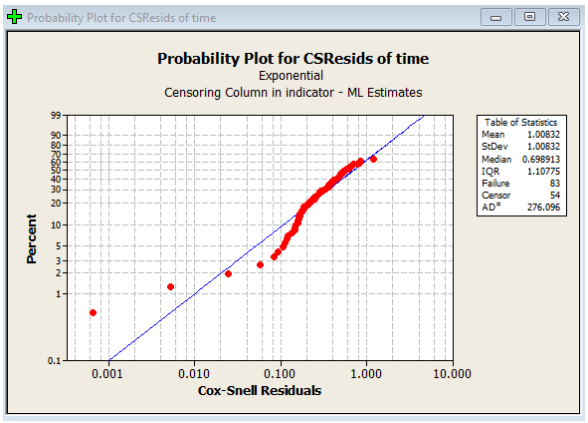
Στο Σχήμα 5.12 βλέπουμε ότι οι τρεις καμπύλες στο μεγαλύτερο μέρος τους δεν διατηρούν ίσες αποστάσεις στην αρχή του διαγράμματος και τέμνονται μεταξύ τους σε κάποια σημεία. Ενώ όσο μεγαλώνει ο χρόνος οι τρεις καμπύλες φαίνεται να είναι παράλληλες. Οπότε συμπεραίνουμε ότι στα δεδομένα μας δεν φαίνεται να ταιριάζει τόσο καλά ένα Μοντέλο Παλινδρόμησης Επιταχυνόμενης Διάρκειας Ζωής (AL).

Στη συνέχεια θα προσαρμόσουμε το Μοντέλο Επιταχυνόμενης Διάρκειας Ζωής (AL) για κάθε μία από τις κατανομές Weibull, Log-Normal και Log-Logistic, και θα ελέγξουμε το μοντέλο μέσω των υπολοίπων Cox-Snell, κάνοντας τις γραφικές παραστάσεις των υπολοίπων Cox-Snell (Παράγραφος 2.2.2).



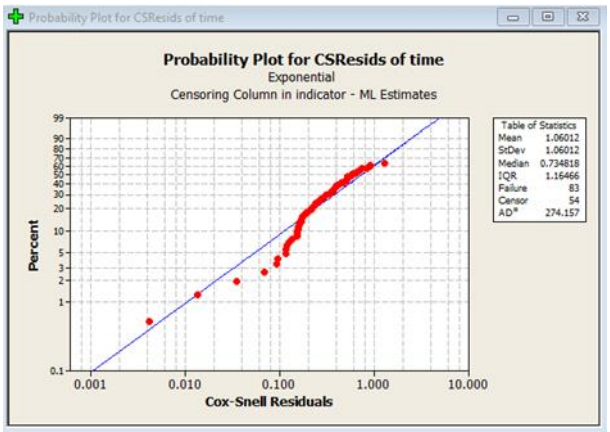
Σχήμα 5.13: *Υπόλοιπα Cox-Snell υπό την κατανομή Weibull.*

Στο Σχήμα 5.13 παριστάνεται η γραφική παράσταση των υπολοίπων Cox-Snell υπό την κατανομή Weibull. Επειδή τα σημεία φαίνεται να αποκλίνουν κατά πολύ από την ευθεία, καταλήγουμε στο ότι ένα μοντέλο Επιταχυνόμενης Διάρκειας Ζωής δεν φαίνεται να ταιριάζει στα δεδομένα μας με αυτή την κατανομή.



Σχήμα 5.14: *Υπόλοιπα Cox-Snell υπό την κατανομή Log-Normal.*

Στο Σχήμα 5.14 παριστάνεται η γραφική παράσταση των υπολοίπων Cox-Snell υπό την κατανομή Log-Normal. Στο Σχήμα 5.14 βλέπουμε και πάλι τα σημεία των υπολοίπων Cox-Snell να αποκλίνουν πολύ από την ευθεία που σχηματίζεται, ειδικά για τις μικρές τιμές του χρόνου. Οπότε ούτε εδώ βλέπουμε να ταιριάζει ένα μοντέλο Επιταχυνόμενης Διάρκειας Ζωής στα δεδομένα μας.



Σχήμα 5.15: *Υπόλοιπα Cox-Snell υπό την κατανομή Log-Logistic.*

Όπως βλέπουμε στο Σχήμα 5.15 τα σημεία των υπολοίπων Cox-Snell φαίνεται να συγκλίνουν πολύ καλύτερα στην ευθεία σε σχέση με τα προηγούμενα δύο Σχήματα (5.13 και 5.14). Οπότε φαίνεται να μας ενδιαφέρει ένα μοντέλο Επιταχυνόμενης Διάρκειας Ζωής (AL) υπό την κατανομή **Log-Logistic** για τα δεδομένα που εξετάζουμε.

Στη συνέχεια θα προσαρμόσουμε στα δεδομένα μας το Μοντέλο της Επιταχυνόμενης Διάρκειας Ζωής (AL) με το στατιστικό πακέτο R, χρησιμοποιώντας την κατανομή Log-Logistic και ως κατηγορική

μεταβλητή την συμμεταβλητή group με κατηγορία αναφοράς την Ομάδα 2. Διότι όπως είδαμε στο Σχήμα 5.4 η καμπύλη της Kaplan-Meier για την Ομάδα 2 είχε διαφέρει αρκετά και από τις δύο άλλες καμπύλες, καθώς ήταν αυτή που έδειχνε τη μεγαλύτερη πιθανότητα επιβίωσης για τους ασθενείς που ανήκαν στην Ομάδα 2 (AML low-risk).

Πίνακας 5.9

groupF <- factor(data$group, levels=c(2,1,3))
mod <- survreg (Surv(time,indicator)~groupF+recipient.age+donor.age+recipient.sex+ donor.sex+recipient.cmv+donor.cmv+waiting.time+fab+mtx, data=data, dist="loglogistic")
summary(mod)
Call: survreg(formula = Surv(time, indicator) ~ groupF + recipient.age + donor.age + recipient.sex + donor.sex + recipient.cmv + donor.cmv + waiting.time + fab + mtx, data = data, dist = "loglogistic")

	Value	Std. Err	z	p
(intercept)	8.208019	0.738381	11.12	<2e-16
groupF1	-1.666080	0.519893	-3.20	0.0014
groupF3	-1.413700	0.430662	-3.28	0.0010
recipient.age	-0.010875	0.030298	-0.36	0.7196
donor.age	-0.010644	0.027767	-0.38	0.7015
recipient.sex	0.267045	0.357773	0.75	0.4554
donor.sex	0.084124	0.364364	0.23	0.8174
recipient.cmv	0.282153	0.386190	0.73	0.4650
donor.cmv	-0.068388	0.370028	-0.18	0.8534
waiting.time	0.000556	0.000583	0.95	0.3404
fab	-1.263599	0.427418	-2.96	0.0031
mtx	-0.849157	0.416284	-2.04	0.0414
log(scale)	0.079500	0.094549	0.84	0.4004

Scale= 1.08
Log logistic distribution
Loglik(model)= -636 Loglik(intercept only)= -651.7
Chisq= 31.46 on 11 degrees of freedom, p= 0.00093
Number of Newton-Raphson Iterations: 4
n= 137
AIC(mod) = 1298.036

Στον Πίνακα 5.9 βλέπουμε τους συντελεστές που προκύπτουν από την προσαρμογή ενός μοντέλου Επιταχυνόμενης Διάρκειας Ζωής (AL) στα δεδομένα μας. Έχουμε επιλέξει τα δεδομένα μας να ακολουθούν την κατανομή Log-Logistic και η συνάρτηση επιβίωσής της είναι:

$$S(t) = (1 + \alpha t^{\eta})^{-1}$$

όπου α: η παράμετρος κλίμακας και η: η παράμετρος σχήματος.

Στον Πίνακα 5.9 βλέπουμε ότι η R έχει υπολογίσει την παράμετρο κλίμακας (Scale) και η τιμή της είναι 1.08. Η επίδραση των συμμεταβλητών στο μοντέλο μας είναι σαν μία αλλαγή του χρόνου όπου ο χρόνος μεταβάλλεται από t σε αt^{η}. Για παράδειγμα όπως βλέπουμε στον παραπάνω πίνακα, αν αυξηθεί κατά μία μονάδα η κατηγορία αναφοράς groupF2, τότε η μεταβλητή groupF1 θα γίνει -1.666080 <1 και η groupF3 αντίστοιχα θα γίνει -1.413700 <1 οπότε σε αυτή την περίπτωση θα επιβραδυνθεί η διακοπή του μοντέλου της Επιταχυνόμενης Διάρκειας Ζωής (AL) και κατά συνέπεια θα μειωθεί ο κίνδυνος θανάτου των ασθενών μας.

Στη συνέχεια θα εφαρμόσουμε την **backward τεχνική με βήματα** για να δούμε ποιο μοντέλο τελικά είναι το καλύτερο, δηλαδή ποιες συμμεταβλητές θα έχει. Στον Πίνακα 5.10 βλέπουμε τη διαδικασία της διαδοχικής αφαίρεσης μεταβλητών. Στην αρχή ξεκινάμε με AIC=1298.04 και με όλες τις συμμεταβλητές, ενώ σε κάθε βήμα αφαιρείται εκείνη που είναι λιγότερο στατιστικά σημαντική με σκοπό να μειωθεί η τιμή του κριτηρίου AIC και καταλήγει σε ένα μοντέλο με τρεις μόνο συμμεταβλητές.

Πίνακας 5.10

mod1 <- step(mod, direction="backward", test="Chisq")
Start: AIC=1298.04
Surv(time, indicator) ~ groupF + recipient.age + donor.age + recipient.sex + donor.sex + recipient.cmv + donor.cmv + waiting.time + fab + mtx

	Df	AIC	LRT	Pr(>Chi)
donor.cmv	1	1296.1	0.0341	0.8534294
donor.sex	1	1296.1	0.0533	0.8173977
recipient.age	1	1296.2	0.1289	0.7195686
donor.age	1	1296.2	0.1473	0.7011280
recipient.cmv	1	1296.6	0.5326	0.4655257
recipient.sex	1	1296.6	0.5533	0.4569678
waiting.time	1	1297.0	0.9648	0.3259885
<none>		1298.0		
mtx	1	1300.1	4.0846	0.0432767*
fab	1	1304.7	8.6904	0.0031989**
groupF	2	1309.1	15.0598	0.0005368***

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

Step: AIC=1296.07
Surv(time, indicator) ~ groupF + recipient.age + donor.age + recipient.sex + donor.sex + recipient.cmv + waiting.time + fab + mtx

	Df	AIC	LRT	Pr(>Chi)
donor.sex	1	1294.1	0.0548	0.8148543
recipient.age	1	1294.2	0.1254	0.7232128
donor.age	1	1294.2	0.1683	0.6815985
recipient.cmv	1	1294.6	0.4984	0.4801837
recipient.sex	1	1294.7	0.6293	0.4276207
waiting.time	1	1295.0	0.9491	0.3299558
<none>		1296.1		
mtx	1	1298.1	4.0509	0.0441474*
fab	1	1302.8	8.6776	0.0032214**
groupF	2	1307.1	15.0410	0.0005418***

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

Step: AIC=1294.12
Surv(time, indicator) ~ groupF + recipient.age + donor.age + recipient.sex + recipient.cmv + waiting.time + fab + mtx

	Df	AIC	LRT	Pr(>Chi)
recipient.age	1	1292.3	0.1385	0.709801
donor.age	1	1292.3	0.1543	0.694479
recipient.cmv	1	1292.7	0.5428	0.461283
recipient.sex	1	1292.8	0.6935	0.404973
waiting.time	1	1293.0	0.9027	0.342068

<none>		1294.1		
mtx	1	1296.1	3.9970	0.045582*
fab	1	1300.8	8.6388	0.003291**
groupF	2	1305.1	14.9929	0.000555***

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

Step: AIC=1292.26
Surv(time, indicator) ~ groupF + donor.age + recipient.sex + recipient.cmv + waiting.time + fab + mtx

	Df	AIC	LRT	Pr(>Chi)
recipient.cmv	1	1290.7	0.4781	0.489275
recipient.sex	1	1291.0	0.7198	0.396200
waiting.time	1	1291.2	0.9663	0.325607
donor.age	1	1291.4	1.0959	0.295166
<none>		1292.3		
mtx	1	1294.6	4.3390	0.037249*
fab	1	1298.8	8.5188	0.003515**
groupF	2	1303.3	15.0140	0.000549***

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

Step: AIC=1290.74
Surv(time, indicator) ~ groupF + donor.age + recipient.sex + waiting.time + fab + mtx

	Df	AIC	LRT	Pr(>Chi)
recipient.sex	1	1289.4	0.6987	0.403233
donor.age	1	1289.6	0.8434	0.358425
waiting.time	1	1289.7	0.9231	0.336668
<none>		1290.7		
mtx	1	1292.6	3.8903	0.048565*
fab	1	1297.1	8.3254	0.003910**
groupF	2	1301.6	14.8910	0.000584***

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

Step: AIC=1289.44
Surv(time, indicator) ~ groupF + donor.age + waiting.time + fab + mtx

	Df	AIC	LRT	Pr(>Chi)
donor.age	1	1288.2	0.7167	0.397218
waiting.time	1	1288.6	1.1406	0.285534
<none>		1289.4		
mtx	1	1291.5	4.0918	0.043091*
fab	1	1295.6	8.1362	0.004339**
groupF	2	1300.4	14.9768	0.000560***

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

Step: AIC=1288.16
Surv(time, indicator) ~ groupF + waiting.time + fab + mtx

	Df	AIC	LRT	Pr(>Chi)
waiting.time	1	1287.3	1.1705	0.279305
<none>		1288.2		
mtx	1	1291.2	5.0375	0.024804*

fab	1	1293.8	7.6061	0.005817**
groupF	2	1299.3	15.1306	0.000518***

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

Step: AIC=1287.33
Surv(time, indicator) ~ groupF + fab + mtx

	Df	AIC	LRT	Pr(>Chi)
<none>		1287.3		
mtx	1	1290.5	5.1183	0.0236752*
fab	1	1293.0	7.6962	0.0055339**
groupF	2	1297.3	13.9805	0.0009208***

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

Οπότε το καλύτερο μοντέλο είναι τελικά αυτό που περιέχει τις συμμεταβλητές group, fab, mtx, και το σταθερό όρο με AIC=1287.33, αφού όπως βλέπουμε αυτές οι τρεις συμμεταβλητές επηρεάζουν περισσότερο την διάρκεια ζωής των ασθενών μας. Ενώ έχει παράμετρο κλίμακας (Scale) ίση με 1.1 της κατανομής Log-Logistic.

Αναλυτικά το τελικό Μοντέλο της Επιταχυνόμενης Διάρκειας Ζωής (AL) που ταιριάζει καλύτερα στα δεδομένα μας (Πίνακας 5.11):

Πίνακας 5.11

summary(mod1)
Call: survreg(formula = Surv(time, indicator) ~ groupF + fab + mtx, data = data, dist = "loglogistic")

	Value	Std. Error	z	p
(intercept)	7.9752	0.3682	21.66	<2e-16
groupF1	-1.3741	0.4726	-2.91	0.00364
groupF3	-1.3945	0.4189	-3.33	0.00087
fab	-1.1785	0.4238	-2.78	0.00543
mtx	-0.8924	0.3896	-2.29	0.02298
log(scale)	0.0950	0.0943	1.01	0.31371

Scale= 1.1
Log logistic distribution
Loglik(model)= -637.7 Loglik(intercept only)= -651.7
Chisq= 28.17 on 4 degrees of freedom, p= 1.2e-05
Number of Newton-Raphson Iterations: 4 n= 137

Στον παραπάνω πίνακα (5.11) παρατηρούμε ότι όλες οι τιμές p-value του τελικού μας μοντέλου είναι πολύ μικρές και συνεπώς στατιστικά σημαντικές, το οποίο είναι αναμενόμενο καθώς αυτό είναι και το καλύτερο μοντέλο της Επιταχυνόμενης Διάρκειας Ζωής για τους ασθενείς μας. Όπως βλέπουμε στον πίνακα με τις τιμές των συμμεταβλητών, αν αυξήσουμε κατά μία μονάδα την μεταβλητή groupF2, τότε και πάλι οι τιμές των groupF1 και groupF3 είναι μικρότερες της μονάδας (-1.3741 και -1.3945 αντίστοιχα) δηλαδή και στο τελικό μας μοντέλο θα επιβραδυνθεί η διακοπή της διάρκειας ζωής των ασθενών μας (δηλαδή ο θάνατος) και κατά συνέπεια θα μειωθεί ο κίνδυνός τους.

Επίσης τα Διαστήματα Εμπιστοσύνης που έχει το καλύτερο μοντέλο Επιταχυνόμενης Διάρκειας Ζωής (AL) είναι:

confint.default(mod1)		
	2.5 %	97.5 %
(Intercept)	7.253542	8.6967622
groupF1	-2.300416	-0.4477781
groupF3	-2.215431	-0.5735413
fab	-2.009215	-0.3477937
mtx	-1.655919	-0.1288305

5.8 Μοντέλο Παλινδρόμησης Αναλογικής Διακινδύνευσης (PH)

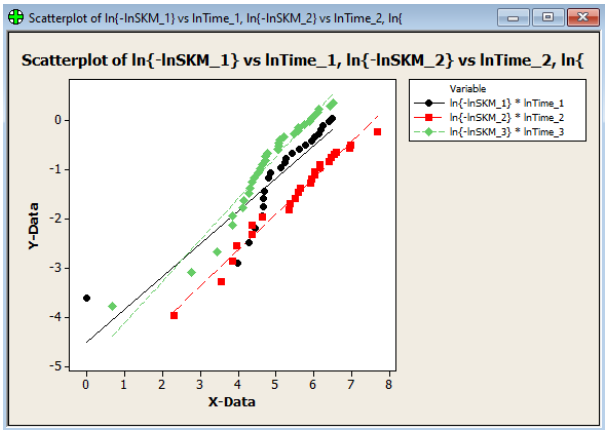
Σε αυτή την παράγραφο θα ελέγξουμε γραφικά με τη βοήθεια του Minitab αν στα δεδομένα μας ταιριάζει ένα Μοντέλο Αναλογικής Διακινδύνευσης (PH), το οποίο είδαμε αναλυτικά στην Παράγραφο 2.3, χρησιμοποιώντας και πάλι ως κατηγορική μεταβλητή τη συμμεταβλητή group.

Για να το ελέγξουμε αυτό θα χρειαστεί να υπολογίσουμε την συνάρτηση ln{-lnSKM} ξεχωριστά για τις τρεις ομάδες που έχουμε, όπου SKM η εκτίμηση Kaplan-Meier (Πίνακας 5.10).

Πίνακας 5.10

ln{-lnSKM_1}	ln{-lnSKM_2}	ln{-lnSKM_3}
-3.62427	-3.97963	-3.79546
-2.91752	-3.27703	-3.09088
-2.49814	-2.86192	-2.67375
-2.19620	-2.56446	-2.13891
-1.95844	-2.33135	-1.94421
-1.76113	-2.13891	-1.77740
-1.59160	-1.97446	-1.63094
-1.44228	-1.83044	-1.49994
-1.18619	-1.70198	-1.38105
-1.07368	-1.58575	-1.27189
-0.96893	-1.47936	-1.17068
-0.87058	-1.38105	-1.07609
-0.77755	-1.28949	-0.98705
-0.68521	-1.20363	-0.90272
-0.59692	-1.12263	-0.82242
-0.51200	-1.04584	-0.74558
-0.42988	-0.97269	-0.67173
-0.35005	-0.90272	-0.60045
-0.27205	-0.83555	-0.53139
-0.19548	-0.77084	-0.46425
-0.11994	-0.70831	-0.39874
-0.03969	-0.64770	-0.33461
0.04030	-0.57542	-0.27163
	-0.50536	-0.20957
	-0.24129	-0.14824
		-0.08742
		-0.02691
		0.03350
		0.09405
		0.15496
		0.21649
		0.27896
		0.34272

Στη συνέχεια θα κάνουμε τη γραφική παράσταση της συνάρτησης που μόλις υπολογίσαμε για κάθε μία ομάδα, συναρτήσει και πάλι του λογαριθμισμένου χρόνου της κάθε ομάδας αντίστοιχα. Έπειτα θα ελέγξουμε αν οι τρεις ευθείες που σχηματίζονται είναι παράλληλες.



Σχήμα 5.16: *Γραφική παράσταση των τριών συναρτήσεων ln{-lnSKM} της κάθε ομάδας.*

Το Σχήμα 5.16 μας δείχνει τις τρεις συναρτήσεις ln{-lnSKM}, τα σημεία των οποίων σχηματίζουν τρεις ευθείες. Βλέπουμε ότι αυτές οι ευθείες δεν είναι παράλληλες, γιατί δύο από αυτές τέμνονται. Οπότε το μοντέλο της Αναλογικής Διακινδύνευσης (PH) δεν φαίνεται να είναι το πιο κατάλληλο μοντέλο για τα δεδομένα των ασθενών μας.

5.9 Μοντέλο Αναλογικής Διακινδύνευσης του Cox

5.9.1 Προσαρμογή του Μοντέλου του Cox:

Με τη βοήθεια του στατιστικού πακέτου R, θα προσαρμόσουμε το Μοντέλο Αναλογικής Διακινδύνευσης του Cox στα δεδομένα μας, το οποίο είδαμε αναλυτικά στην Παράγραφο 2.4.

Πίνακας 5.12

summary(modC)

Call: coxph(formula = Surv(time, indicator) ~ groupF + recipient.age + donor.age + recipient.sex + donor.sex + recipient.cmv + donor.cmv + waiting.time + fab + mtx, data = data)

n= 137, number of events= 83

	coef	exp(coef)	se(coef)	z	p
groupF1	1.0624620	2.8934861	0.3705119	2.868	0.00414**
groupF3	0.8742014	2.3969603	0.2821845	3.098	0.00195**
recipient.age	0.0139626	1.0140605	0.0205096	0.681	0.49601
donor.age	-0.0021921	0.9978103	0.0186648	-0.117	0.90651
recipient.sex	-0.1093030	0.8964587	0.2412718	-0.453	0.65053
donor.sex	0.0333095	1.0338704	0.2416459	0.138	0.89036
recipient.cmv	-0.0606449	0.9411574	0.2546450	-0.238	0.81176
donor.cmv	-0.0480256	0.9531094	0.2472610	-0.194	0.84600
waiting.time	-0.0003417	0.9996584	0.0003927	-0.870	0.38424
fab	0.8018912	2.2297539	0.2822418	2.841	0.00450**
mtx	0.2918278	1.3388724	0.2542193	1.148	0.25099

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

	exp(coef)	exp(-coef)	lower .95	upper .95
groupF1	2.8935	0.3456	1.3997	5.981
groupF3	2.3970	0.4172	1.3787	4.167
recipient.age	1.0141	0.9861	0.9741	1.056
donor.age	0.9978	1.0022	0.9620	1.035
recipient.sex	0.8965	1.1155	0.5587	1.438
donor.sex	1.0339	0.9672	0.6438	1.660
recipient.cmv	0.9412	1.0625	0.5714	1.550
donor.cmv	0.9531	1.0492	0.5870	1.547
waiting.time	0.9997	1.0003	0.9989	1.000
fab	2.2298	0.4485	1.2824	3.877
mtx	1.3389	0.7469	0.8135	2.204

Concordance = 0.677 (se = 0.032)
Likelihood ratio test = 26.1 on 11 df, p=0.006
Wald test = 25.2 on 11 df, p=0.009
Score (logrank) test = 26.65 on 11 df, p=0.005
AIC(modC) = 742.4931

Υπενθυμίζουμε ότι το μοντέλο του Cox έχει συνάρτηση διακινδύνευσης (όπως είδαμε αναλυτικά στην Παράγραφο 2.4.1):

$$h(t;x) = h_0(t)e^{\beta' x}$$

Οι τιμές του εκθετικού όρου της συνάρτησης h(t), δηλαδή το $e^{\hat{\beta}}$ (στήλη exp(coef) Πίνακας 5.12), μας δείχνει κατά πόσο πολλαπλασιάζεται και αυξάνεται η συνάρτηση διακινδύνευσης h(t). Αυτό σημαίνει ότι το $e^{\hat{\beta}}$ μας δείχνει κατά πόσο μία συμμεταβλητή επηρεάζει τη διάρκεια ζωής των ασθενών όταν όλες οι υπόλοιπες συμμεταβλητές παραμένουν σταθερές. Για παράδειγμα, κρατώντας τις άλλες συμμεταβλητές σταθερές, ο κίνδυνος να πεθάνει ένας ασθενής (δηλαδή να συμβεί το γεγονός) που ανήκει στην Ομάδα 1 (ALL) σε σχέση με έναν που ανήκει στην Ομάδα 2 (AML low_risk) αυξάνεται κατά $e^{\hat{\beta}}$ = 2.8934861 > 1. Αντίστοιχα, ο κίνδυνος να πεθάνει ένας ασθενής που ανήκει στην Ομάδα 3 (AML high_risk) σε σχέση με έναν που ανήκει στην Ομάδα 2 (AML low_risk) αυξάνεται κατά $e^{\hat{\beta}}$ = 2.3969603 > 1, καθώς οι υπόλοιπες μεταβλητές παραμένουν σταθερές. Επίσης, εάν η ηλικία ενός ασθενούς (recipient.age) αυξηθεί κατά μία μονάδα, ενώ οι υπόλοιπες μεταβλητές παραμένουν σταθερές, ο κίνδυνος να πεθάνει αυτός ο ασθενής θα αυξηθεί κατά $e^{\hat{\beta}}$ = 1.0140605 > 1. Με αντίστοιχο τρόπο επηρεάζουν και οι υπόλοιπες συμμεταβλητές τη διάρκεια ζωής των ασθενών μας.

Παρατηρούμε ότι στη στήλη exp(coef) οι τιμές που είναι μικρότερες της μονάδας ($e^{\hat{\beta}}$ <1) επιδρούν θετικά στη διάρκεια ζωής των ασθενών μας, δηλαδή μειώνουν τον κίνδυνο θανάτου των ασθενών. Αυτές οι συμμεταβλητές είναι οι donor.age (ηλικία του δότη), recipient.sex (φύλο του ασθενή), recipient.cmv (φύλο του δότη), donor.cmv (κατάσταση του cmv του δότη, 1=θετικό cmv, 0=αρνητικό cmv) και waiting.time (ο χρόνος αναμονής σε μέρες από τη διάγνωση μέχρι τη μεταμόσχευση).

Έλεγχος Wald:

Ο έλεγχος Wald για το μοντέλο του Cox ελέγχει την υπόθεση:

$$H_0: \beta_i = 0 \quad \forall i = 1,..11 \quad \text{και} \quad H_1: \beta_i \neq 0 \quad \text{για κάποιο } i = 1,\dots 11$$

Ο έλεγχος αυτός στη μηδενική υπόθεση υποστηρίζει ότι όλοι οι συντελεστές του μοντέλου πρέπει να είναι μηδενικοί, δηλαδή ότι το μοντέλο πρέπει να περιέχει μόνο το σταθερό όρο, ενώ η εναλλακτική υποστηρίζει ότι τουλάχιστον ένας από τους συντελεστές είναι διαφορετικός από το μηδέν. Ήδη η R έχει υπολογίσει τη τιμή του Wald test από τον παραπάνω πίνακα (Πίνακας 5.12), δηλαδή z = 25.2. Η τιμή p-value είναι 0.009, που είναι πολύ μικρή και συνεπώς στατιστικά σημαντική. Οπότε απορρίπτουμε την υπόθεση H_0. Αυτό σημαίνει ότι δεχόμαστε την εναλλακτική υπόθεση, η οποία υποστηρίζει ότι τουλάχιστον ένας από τους συντελεστές που έχουμε πρέπει να είναι διαφορετικός από το 0.

Στη συνέχεια θα εφαρμόσουμε την **backward τεχνική με βήματα** για να δούμε ποιο μοντέλο τελικά είναι το καλύτερο, δηλαδή ποιες συμμεταβλητές θα έχει.. Στον Πίνακα 5.13 βλέπουμε τη διαδικασία της διαδοχικής αφαίρεσης μεταβλητών. Στην αρχή ξεκινάμε με AIC=742.49 και με όλες τις συμμεταβλητές, ενώ σε κάθε βήμα αφαιρείται εκείνη που είναι λιγότερο στατιστικά σημαντική με σκοπό να μειωθεί η τιμή του κριτηρίου AIC και καταλήγουμε σε ένα μοντέλο με δύο μόνο συμμεταβλητές.

Πίνακας 5.13

modC2 <- step(modC, direction="backward", test="Chisq")
Start: AIC=742.49
Surv(time, indicator) ~ groupF + recipient.age + donor.age + recipient.sex + donor.sex + recipient.cmv + donor.cmv + waiting.time + fab + mtx

	Df	AIC	LRT	Pr(>Chi)
donor.age	1	740.51	0.0138	0.906556
donor.sex	1	740.51	0.0190	0.890229
donor.cmv	1	740.53	0.0378	0.845824
recipient.cmv	1	740.55	0.0567	0.811725
recipient.sex	1	740.70	0.2045	0.651083
recipient.age	1	740.95	0.4590	0.498110
waiting.time	1	741.34	0.8435	0.358403
mtx	1	741.78	1.2827	0.257406
<none>		742.49		
fab	1	748.73	8.2319	0.004116**
groupF	2	752.05	13.5597	0.001136**

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

Step: AIC=740.51
Surv(time, indicator) ~ groupF + recipient.age + recipient.sex + donor.sex + recipient.cmv + donor.cmv + waiting.time + fab + mtx

	Df	AIC	LRT	Pr(>Chi)
donor.sex	1	738.52	0.0139	0.906014
donor.cmv	1	738.55	0.0453	0.831412
recipient.cmv	1	738.56	0.0498	0.823394
recipient.sex	1	738.71	0.2026	0.652598
recipient.age	1	739.28	0.7743	0.378880
waiting.time	1	739.35	0.8410	0.359108
mtx	1	739.81	1.3035	0.253574
<none>		740.51		
fab	1	746.76	8.2483	0.004079**
groupF	2	750.83	14.3207	0.000777**

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

Step: AIC=738.52
Surv(time, indicator) ~ groupF + recipient.age + recipient.sex + recipient.cmv + donor.cmv + waiting.time + fab + mtx

	Df	AIC	LRT	Pr(>Chi)
recipient.cmv	1	736.57	0.0467	0.828939
donor.cmv	1	736.57	0.0481	0.826410
recipient.sex	1	736.72	0.1944	0.659283
recipient.age	1	737.29	0.7653	0.381677
waiting.time	1	737.44	0.9215	0.337072
mtx	1	737.81	1.2940	0.255307
<none>		738.52		
fab	1	744.78	8.2545	0.004065**
groupF	2	748.86	14.3357	0.000771**

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

Step: AIC=736.57
Surv(time, indicator) ~ groupF + recipient.age + recipient.sex + donor.cmv + waiting.time + fab + mtx

	Df	AIC	LRT	Pr(>Chi)
donor.cmv	1	734.66	0.0878	0.767008
recipient.sex	1	734.81	0.2430	0.622023
recipient.age	1	735.29	0.7219	0.395521
waiting.time	1	735.51	0.9417	0.331844
mtx	1	735.82	1.2516	0.263252
<none>		736.57		
fab	1	742.82	8.2507	0.004074**
groupF	2	746.87	14.3066	0.000782**

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

Step: AIC=734.66 Surv(time, indicator) ~ groupF + recipient.age + recipient.sex + waiting.time + fab + mtx

	Df	AIC	LRT	Pr(>Chi)
recipient.sex	1	732.86	0.2074	0.648821
recipient.age	1	733.31	0.6554	0.418181
waiting.time	1	733.58	0.9216	0.337061
mtx	1	733.88	1.2239	0.268591
<none>		734.66		
fab	1	741.05	8.3934	0.003766**
groupF	2	744.94	14.2807	0.000793***

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

Step: AIC=732.86 Surv(time, indicator) ~ groupF + recipient.age + waiting.time + fab + mtx

	Df	AIC	LRT	Pr(>Chi)
recipient.age	1	731.56	0.6944	0.404657
waiting.time	1	732.02	1.1550	0.282508
mtx	1	732.23	1.3637	0.242896
<none>		732.86		
fab	1	739.76	8.8982	0.002855**
groupF	2	743.14	14.2774	0.000794***

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

Step: AIC=731.56 Surv(time, indicator) ~ groupF + waiting.time + fab + mtx

	Df	AIC	LRT	Pr(>Chi)
waiting.time	1	730.89	1.3339	0.248104
mtx	1	731.51	1.9544	0.162115
<none>		731.56		
fab	1	737.83	8.2737	0.004022**
groupF	2	741.45	13.8897	0.000964***

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

Step: AIC=730.89 Surv(time, indicator) ~ groupF + fab + mtx

	Df	AIC	LRT	Pr(>Chi)
mtx	1	730.85	1.9580	0.161729
<none>		730.89		
fab	1	737.20	8.3083	0.003947**
groupF	2	739.45	12.5559	0.001877**

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

Step: AIC=730.85 Surv(time, indicator) ~ groupF + fab

	Df	AIC	LRT	Pr(>Chi)
<none>		730.85		
fab	1	737.14	8.2902	0.0039859**
groupF	2	740.81	13.9572	0.0009316***

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

Οπότε το καλύτερο μοντέλο είναι τελικά αυτό που περιέχει τις συμμεταβλητές groupF και fab, με AIC=730.85, αφού όπως βλέπουμε αυτές οι δύο συμμεταβλητές επηρεάζουν περισσότερο τη διάρκεια ζωής των ασθενών μας.

Αναλυτικά οι συντελεστές του τελικού «βέλτιστου» μοντέλου Αναλογικής Διακινδύνευσης του Cox που ταιριάζει καλύτερα στα δεδομένα μας φαίνονται στον Πίνακα 5.14:

Πίνακας 5.14

modC2

Call: coxph(formula = Surv(time, indicator) ~ groupF + fab, data = data)

	coef	exp(coef)	se(coef)	z	p
groupF1	0.9045	2.4707	0.3203	2.824	0.00475
groupF3	0.8525	2.3456	0.2684	3.176	0.87128
fab	0.7695	2.1587	0.2703	2.847	0.00442

Likelihood ratio test=21.74 on 3 df, p=7.38e-05
n= 137, number of events= 83

Στο τελικό μας μοντέλο (Πίνακας 5.14) ο εκθετικός όρος $e^{\hat{\beta}}$ μας δείχνει πάλι ότι αυξάνεται ο κίνδυνος και στις τρεις μεταβλητές που έχουμε (exp(coef) > 1 και για τις 3 μεταβλητές). Συγκεκριμένα, ο κίνδυνος να πεθάνει ένας ασθενής (δηλαδή να συμβεί το γεγονός) που ανήκει στην Ομάδα 1 (ALL) σε σχέση με έναν που ανήκει στην Ομάδα 2 (AML low_risk), καθώς η μεταβλητή fab παραμένει

σταθερή, αυξάνεται κατά $e^{\widehat{\beta}}$= 2.4707 > 1. Αντίστοιχα, ο κίνδυνος να πεθάνει ένας ασθενής που ανήκει στην Ομάδα 3 (AML high_risk) σε σχέση με έναν που ανήκει στην Ομάδα 2 (AML low_risk) αυξάνεται κατά $e^{\widehat{\beta}}$= 2.3456 > 1, καθώς η μεταβλητή fab παραμένει σταθερή. Τέλος, ο κίνδυνος να πεθάνει ένας ασθενής με fab=1 σε σχέση με έναν με fab=0, αν κρατήσουμε τη μεταβλητή group σταθερή αυξάνεται κατά $e^{\widehat{\beta}}$= 2.1587 > 1. Υπενθυμίζουμε ότι η μεταβλητή fab είναι μία ταξινόμηση των ασθενών με μυελοκυτταρική λευχαιμία (AML) που βασίζεται σε μορφολογικά κριτήρια, δηλαδή οι ασθενείς με fab βαθμού 4 ή 5 (M4 ή M5) διατρέχουν μεγαλύτερο κίνδυνο υποτροπής ή θανάτου μετά τη μεταμόσχευση του μυελού των οστών.

Επίσης τα Διαστήματα Εμπιστοσύνης που έχει το καλύτερο μοντέλο Αναλογικής Διακινδύνευσης του Cox είναι:

confint.default(modC2)

	2.5 %	97.5 %
groupF1	0.2766989	1.532297
groupF3	0.3264960	1.378600
fab	0.2396767	1.299319

5.9.2 Υπόλοιπα Schoenfeld και Καμπύλες ROC:

Υπόλοιπα Schoenfeld:

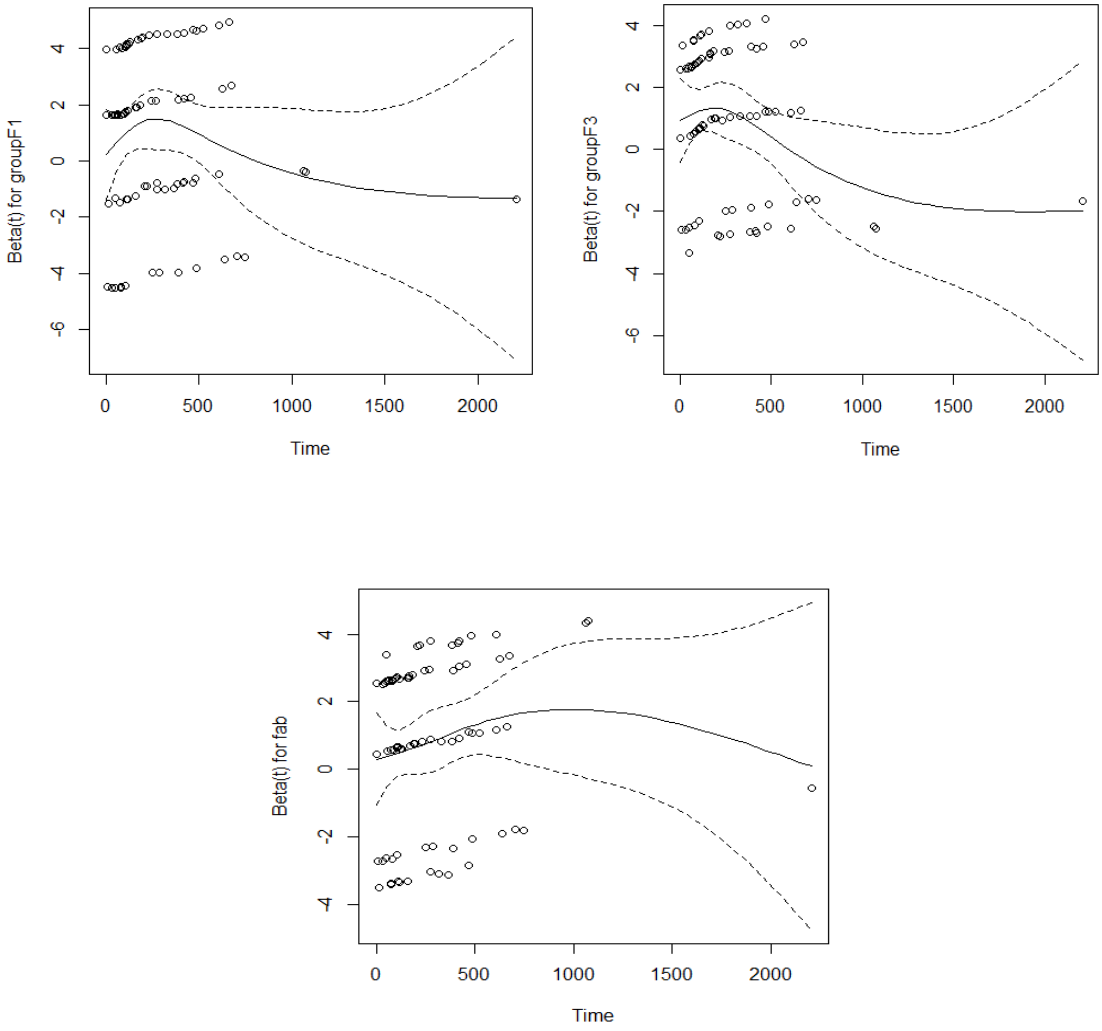
Θα κάνουμε τα διαγράμματα των Υπολοίπων Schoenfeld για τις συμμεταβλητές του τελικού μοντέλου του Cox που παρουσιάστηκε στην Παράγραφο 5.9.1 (modC2) στον Πίνακα 5.11, με τη βοήθεια της R. Μας ενδιαφέρει να ελέγξουμε αν υπάρχει εξάρτηση των συντελεστών των συμμεταβλητών του μοντέλου από το χρόνο και αν η γραφική τους παράσταση σχηματίζει μία οριζόντια γραμμή κοντά στο 0. Τα υπόλοιπα Schoenfeld τα είδαμε αναλυτικά στην Παράγραφο 2.4.5. Με αυτό τον τρόπο θα ελέγξουμε εάν στο μοντέλο του Πίνακα 5.14 ισχύει η υπόθεση της Αναλογικής Διακινδύνευσης (PH).

Πίνακας 5.15

modC3< -cox.zph(modC2, transform="identity", terms=FALSE)

	chisq	**df**	**p**
groupF1	0.135	1	0.713
groupF3	2.737	1	0.098
fab	0.548	1	0.459
GLOBAL	3.861	3	0.277

Στον Πίνακα 5.15 παρουσιάζεται ένας συνοπτικός πίνακας ελέγχου της υπόθεσης της Αναλογικής Διακινδύνευσης (PH), όπου εκτελείται ένας X^2 έλεγχος και έχουμε σε ξεχωριστή στήλη τα p-values αυτών των ελέγχων. Παρατηρούμε ότι η μεταβλητή groupF3 φαίνεται να είναι οριακά ανεξάρτητη από το χρόνο t καθώς έχει τιμή p-value ίση με 0.098.



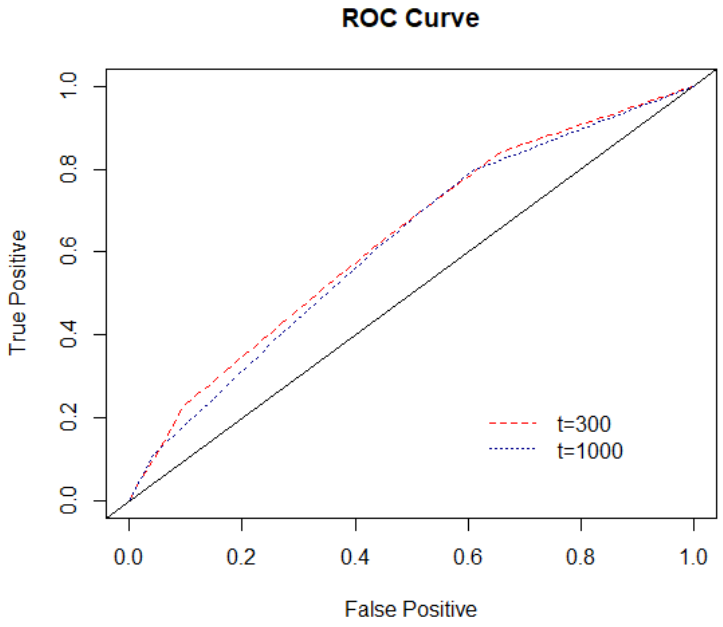
Σχήμα 5.17: *Διαγράμματα υπολοίπων Schoenfeld για τις συμμεταβλητές groupF1, groupF3 και fab.*

Όπως βλέπουμε στο Σχήμα 5.17, για τις συμμεταβλητές groupF1 και groupF3 (επάνω) δεν παρατηρούμε μεγάλες διαφορές στα διαγράμματα των υπολοίπων Schoenfeld τους. Και στα δύο αυτά διαγράμματα η μαύρη γραμμή που δημιουργείται από τα υπόλοιπα Schoenfeld δεν είναι οριζόντια και αποκλίνει αρκετά από το 0, ενώ φαίνεται ότι οι τιμές τους δεν είναι ανεξάρτητες του χρόνου. Αυτό το βλέπουμε γιατί φαίνεται οι τιμές τους να είναι συγκεντρωμένες στο αριστερό τμήμα των διαγραμμάτων, δηλαδή για μικρές τιμές του χρόνου. Αυτό σημαίνει ότι για τις συμμεταβλητές groupF1 και groupF3 η υπόθεση της Αναλογικής Διακινδύνευσης δεν φαίνεται να ικανοποιείται.

Στο ίδιο Σχήμα (5.17), για τη συμμεταβλητή fab (κάτω) βλέπουμε ότι η μαύρη γραμμή είναι πιο κοντά στο 0 και φαίνεται να σχηματίζει μία οριζόντια ευθεία. Επίσης και για την συμμεταβλητή fab οι τιμές των υπολοίπων Schoenfeld δεν φαίνεται να είναι ανεξάρτητες του χρόνου. Οπότε για την συμμεταβλητή fab φαίνεται να ικανοποιείται η υπόθεση της Αναλογικής Διακινδύνευσης.

Καμπύλη ROC:

Στη συνέχεια θα εφαρμόσουμε τις Καμπύλες ROC (Παράγραφος 2.4.6) στα δεδομένα μας με τη βοήθεια του στατιστικού πακέτου R, βασισμένοι στο τελικό μοντέλο του Cox (modC2) που βρήκαμε στην Παράγραφο 5.9.1 (Πίνακας 5.14), για να ελέγξουμε την αποτελεσματικότητα του μοντέλου μας. Θα επιλέξουμε δύο διαφορετικές τιμές του χρόνου ζωής των ασθενών μας, οι οποίες απέχουν πολύ μεταξύ τους για να έχουμε μία πλήρη εικόνα για το πως συμπεριφέρεται το μοντέλο μας από τους μικρούς στους μεγάλους χρόνους. Έτσι εφαρμόζουμε τις καμπύλες ROC για $t_1 = 300$ και $t_2 = 1000$ ημέρες.



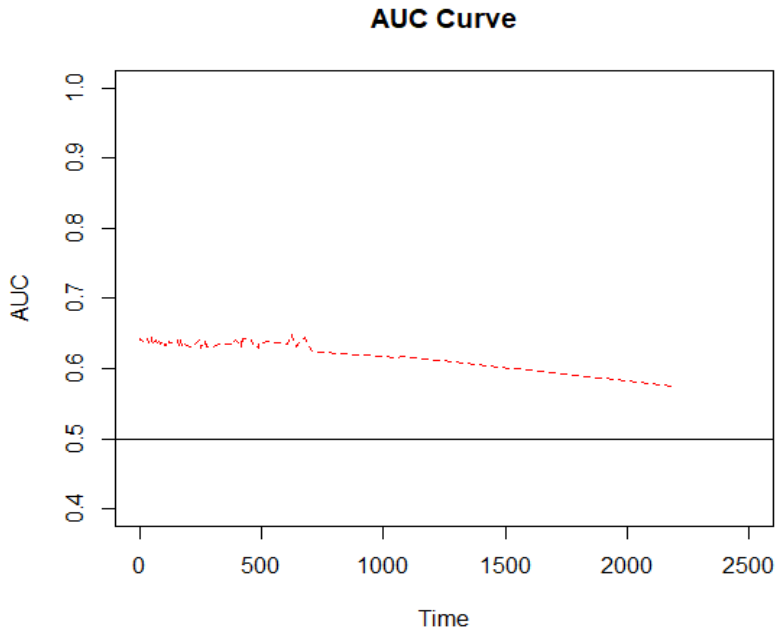
Σχήμα 5.18: *Καμπύλη ROC για το τελικό μοντέλο του Cox για t=300 και t=1000 ημέρες.*

Στο Σχήμα 5.18 παριστάνεται η καμπύλη ROC για το μοντέλο του Cox για δύο διαφορετικούς χρόνους t=300 και t=1000. Όπως βλέπουμε οι δύο καμπύλες που σχηματίζονται είναι αρκετά πανομοιότυπες. Η καλύτερη καμπύλη ROC είναι αυτή που πλησιάζει περισσότερο την πάνω αριστερή γωνία του διαγράμματος, και στη προκειμένη περίπτωση και για τις δύο τιμές βλέπουμε ότι οι καμπύλες δεν πλησιάζουν καθόλου την δεξιά γωνία. Αυτό σημαίνει ότι το τελικό μοντέλο του Cox (Πίνακας 5.14) και για τις μικρές τιμές του χρόνου αλλά και για τις μεγάλες δεν έχει καλή προβλεπτική ικανότητα. Αλλά παρατηρούμε ότι η καμπύλη για t=300 (κόκκινη) είναι λίγο καλύτερη, δηλαδή το μοντέλο του Cox έχει λίγο καλύτερη προβλεπτική ικανότητα για τις μικρές τιμές του χρόνου.

Στη συνέχεια θα κάνουμε μία καμπύλη AUC, δηλαδή θα υπολογίσουμε το εμβαδόν που βρίσκεται κάτω από την καμπύλη ROC για μία μέγιστη τιμή του χρόνου. Θα επιλέξουμε tmax=2000 γιατί λίγες από τις παρατηρήσεις μας βρίσκονται πάνω από τις 2000 ημέρες.

Όπως βλέπουμε στο Σχήμα 5.19 όσο αυξάνεται ο χρόνος η καμπύλη AUC φθίνει με μικρό ρυθμό. Αυτό είναι λογικό γιατί όπως είπαμε δεν έχουμε πολλές παρατηρήσεις που έχουν χρόνο ζωής πάνω από 2000 ημέρες. Γενικά, όσες περισσότερες παρατηρήσεις έχουμε λοιπόν σε ένα χρονικό διάστημα, τόσο αυξάνεται η τιμή του AUC σε αυτό το διάστημα και ισοδύναμα τόσο καλύτερη γίνεται η προβλεπτική ικανότητα του μοντέλου μας. Αλλά για τα δεδομένα μας, η καμπύλη AUC δεν έχει

μεγάλες διαφορές στις τιμές της (δεν έχει πολύ έντονα μέγιστα και ελάχιστα) και αυτό σημαίνει ότι έχουμε σχεδόν τον ίδιο αριθμό παρατηρήσεων σε όλες τις τιμές του χρόνου. Παρατηρούμε μόνο ένα μέγιστο που ξεχωρίζει, στο Σχήμα 5.19, μεταξύ των χρόνων 500 και 700, οπότε για αυτούς τους χρόνους έχουμε λίγες περισσότερες παρατηρήσεις και κατά συνέπεια το μοντέλο του Πίνακα 5.14 έχει λίγο καλύτερη προβλεπτική ικανότητα.



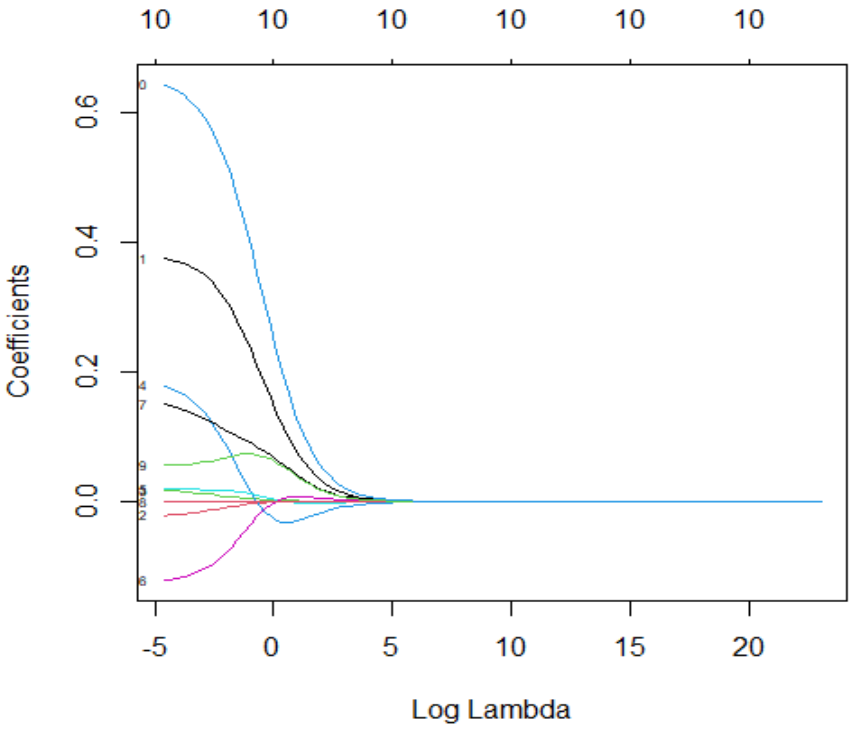
Σχήμα 5.19: *Καμπύλη AUC συναρτήσει του χρόνου.*

5.9.3 Παλινδρόμηση Κορυφογραμμής (Ridge):

Σε αυτή την παράγραφο θα εφαρμόσουμε την Παλινδρόμηση Κορυφογραμμής (Ridge Regression, Παράγραφος 3.2) στα δεδομένα μας για το μοντέλο του Cox. Τα αποτελέσματα στα οποία θα καταλήξουμε θα συγκριθούν με τα αντίστοιχα του κλασικού μοντέλου του Cox που προσαρμόστηκε στην Παράγραφο 5.9.1. Για την υλοποίηση της Ridge χρησιμοποιήσαμε το πακέτο της R glmnet. Στην αρχή επιλέγουμε ένα ευρύ φάσμα στο οποίο θα κινηθεί η παράμετρος λ (Πίνακας 5.16), ενώ μετά θα βρούμε ποιο είναι το βέλτιστο λ για την Παλινδρόμηση Ridge.

Πίνακας 5.16

```
y <- matrix (Surv (data$time,data$indicator), ncol=2)
x <- matrix ( c(groupF, recipient.age, donor.age, recipient.sex, donor.sex,
recipient.cmv, donor.cmv,  waiting.time,  fab,  mtx), ncol=10)
grid = 10^ seq (10, -2, length = 100)
ridge.mod = glmnet (x, y, family="cox", alpha =0, lambda =grid)
dim (coef( ridge.mod ))
[1]  10 100
```



Σχήμα 5.20: *Διάγραμμα των συντελεστών του μοντέλου Ridge συναρτήσει του λογαριθμισμένου λ.*

Υπενθυμίζουμε ότι η παλινδρόμηση Κορυφογραμμής (Ridge) υπολογίζει τους συντελεστές του μοντέλου μας ελαχιστοποιώντας την ποσότητα:

$$L_{\text{Ridge}} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

Ο σκοπός αυτής της μεθόδου είναι να συρρικνώσει τους συντελεστές ενός μοντέλου, στην προκειμένη περίπτωση θα συρρικνώσει τους συντελεστές του μοντέλου του Cox (Πίνακας 5.12).

Στο Σχήμα 5.20 παρουσιάζονται οι δέκα συντελεστές (συμμεταβλητές) του μοντέλου Ridge και πως συμπεριφέρονται καθώς το λ αλλάζει. Κάθε αριθμός που υπάρχει δίπλα από κάθε καμπύλη αντιστοιχεί και σε μία συμμεταβλητή με τη σειρά την οποία τις έχουμε δηλώσει, δηλαδή groupF, recipient.age, donor.age, recipient.sex, donor.sex, recipient.cmv, donor.cmv, waiting.time, fab και mtx. Για παράδειγμα, η μπλε καμπύλη (αριθμός 10) αντιστοιχεί στη συμμεταβλητή recipient.sex καθώς το λ ποικίλλει. Στο δεξί άκρο του διαγράμματος, το λ είναι πολύ μεγάλο και κατά συνέπεια οι εκτιμήσεις των συντελεστών τείνουν και αυτές στο μηδέν (συρρικνώνονται). Ενώ στο αριστερό άκρο το λ είναι ουσιαστικά μηδέν και οι εκτιμήσεις των συντελεστών είναι ίδιες με αυτές που θα υπολόγιζε η μέθοδος των Ελαχίστων Τετραγώνων. Παρατηρούμε ότι κάποιοι συντελεστές αυξάνονται με μεγαλύτερο βαθμό από τους υπόλοιπους καθώς το λ τείνει προς το μηδέν (αριστερό άκρο). Αυτό μας δείχνει ότι αυτοί οι συντελεστές επηρεάζουν πιο πολύ τη διάρκεια ζωής των ασθενών μας και αυτοί οι συντελεστές, όπως βλέπουμε στο Σχήμα 3.20, είναι οι groupF, recipient.cmv και mtx (καμπύλες 1, 6 και 10).

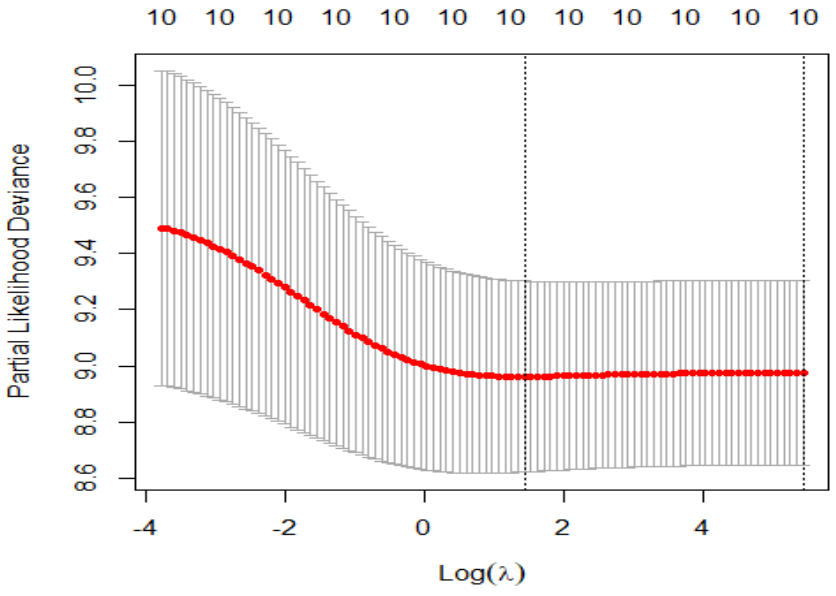
Χρησιμοποιώντας την εντολή cv.glmnet μπορούμε να εφαρμόσουμε τη μέθοδο Cross-Validation (Παράγραφος 3.4), η οποία μας επιστρέφει τη βέλτιστη τιμή της παραμέτρου λ που θα χρησιμοποιήσουμε (Πίνακας 5.17).

Πίνακας 5.17

cv.out = cv.glmnet (x[train,], y[train,], family="cox", alpha =0)
bestlam = cv.out$lambda.min
bestlam = 4.20843
ridge.pred = predict (ridge.mod , family="cox", s=bestlam , newx=x)
out = glmnet (x[train,], y[train,], family="cox", alpha =0)
predict (out, family="cox", type= "coefficients", s= bestlam) [1:10 ,]

groupF	recipient.age	donor.age	recipient.sex	donor.sex
5.818096e-02	1.093994e-03	1.210247e-03	-2.693975e-02	2.635508e-04
recipient.cmv	**donor.cmv**	**waiting.time**	**fab**	**mtx**
1.494924e-02	3.408057e-03	-1.115484e-05	8.089831e-02	4.671426e-02

Όπως βλέπουμε στον παραπάνω πίνακα (5.17) το καλύτερο λ για το μοντέλο του Ridge είναι το 4.20843 και για αυτό το λ βρίσκουμε τους συντελεστές του μοντέλου. Όπως παρατηρούμε κανένας από τους συντελεστές δεν μηδενίζεται (αν και όλοι είναι πολύ κοντά στο 0), κάτι το οποίο είναι λογικό καθώς η μέθοδος Ridge δεν κάνει επιλογή μεταβλητών για το τελικό της μοντέλο. Σε σύγκριση με το αρχικό μοντέλο του Cox που βρήκαμε στην Παράγραφο 5.9.1 (Πίνακας 5.12), παρατηρούμε ότι οι τιμές των εκτιμήσεων όλων των συμμεταβλητών έχουν μειωθεί σημαντικά (έχουν συρρικνωθεί), κάτι το οποίο είναι αναμενόμενο.



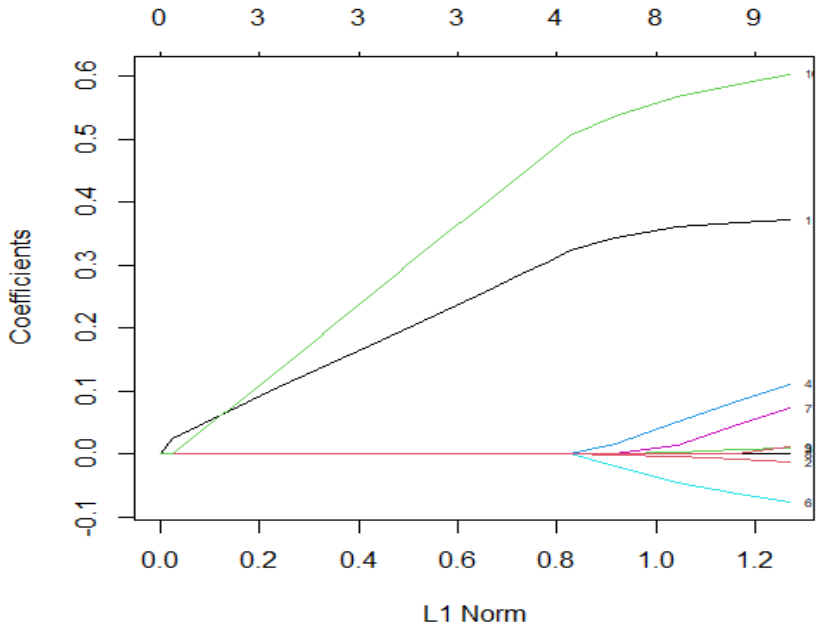
Σχήμα 5.21: *Διάγραμμα απόκλισης* της *Cross-Validation μερικής Πιθανοφάνειας συναρτήσει του log(λ) για το μοντέλο του Ridge.*

Στο Σχήμα 5.21 βλέπουμε το διάγραμμα απόκλισης (deviance) της Cross-Validation μερικής Πιθανοφάνειας συναρτήσει του λογαριθμισμένου λ για την Παλινδρόμηση Ridge. Οι δύο κάθετες

γραμμές συμβολίζουν δύο τιμές του λογαριθμισμένου λ τις οποίες έχει επιλέξει το μοντέλο. Και συγκεκριμένα η αριστερή κάθετη γραμμή καθορίζει τον λογάριθμο του βέλτιστου λ, που όπως είδαμε στον Πίνακα 5.17, είναι 4.20843 (ln(4.20843) = 1,43708) και επίσης μας δείχνει το σημείο όπου η απόκλιση (deviance) ελαχιστοποιείται. Ενώ η δεξιά κάθετη γραμμή μας δίνει τη μέγιστη τιμή του λ σε απόσταση μίας τυπικής απόκλισης από την ελάχιστη τιμή της απόκλισης (αριστερή κάθετη γραμμή).

5.9.4 Παλινδρόμηση Lasso:

Σε αυτή την παράγραφο θα προσαρμόσουμε από την αρχή το μοντέλο του Cox εφαρμόζοντας τη μέθοδο Lasso (Παράγραφος 3.3) στα δεδομένα μας και θα εξετάσουμε κατά πόσο η μέθοδος Lasso είναι πιο αποτελεσματική από τη μέθοδο Ridge. Θα χρησιμοποιήσουμε τον ίδιο κώδικα στην R με αυτόν της Παραγράφου 5.9.3 (μέθοδος Ridge) με μοναδική διαφορά ότι η Lasso έχει παράμετρο α=1, ενώ η Ridge είχε α=0. Όπως και με τη μέθοδο Ridge, στην αρχή θα επιλέξουμε ένα ευρύ φάσμα στο οποίο θα κινηθεί η παράμετρος λ, ενώ μετά θα βρούμε ποιο είναι το βέλτιστο λ για την Παλινδρόμηση Lasso.



Σχήμα 5.22: *Διάγραμμα των συντελεστών του μοντέλου της Lasso συναρτήσει της L1-νόρμας.*

Υπενθυμίζουμε ότι η παλινδρόμηση Lasso υπολογίζει τους συντελεστές του μοντέλου μας ελαχιστοποιώντας την ποσότητα:

$$L_{\text{Lasso}} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

Ο σκοπός αυτής της μεθόδου είναι να συρρικνώσει τους συντελεστές ενός μοντέλου και να μηδενίσει αυτούς που δεν θεωρεί σημαντικούς, στην προκειμένη περίπτωση θα συρρικνώσει τους συντελεστές του μοντέλου του Cox (Πίνακας 5.12).

Στο Σχήμα 5.22 παρουσιάζονται οι δέκα συντελεστές (των συμμεταβλητών) του μοντέλου της Lasso και πως συμπεριφέρονται καθώς το λ αλλάζει σύμφωνα με την L1-νόρμα. Όσο μεγαλύτερη είναι η L1-νόρμα τόσο μικραίνει η τιμή της παραμέτρου λ. Δηλαδή στη δεξιά πλευρά του Σχήματος 5.22, η παράμετρος λ τείνει στο μηδέν. Υπενθυμίζουμε ότι η νόρμα L1 δίνεται από το τύπο:

$$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$$

Κάθε αριθμός που υπάρχει δίπλα από κάθε καμπύλη (στο Σχήμα 5.22) αντιστοιχεί και σε μία συμμεταβλητή με τη σειρά την οποία τις έχουμε δηλώσει, δηλαδή groupF, recipient.age, donor.age, recipient.sex, donor.sex, recipient.cmv, donor.cmv, waiting.time, fab και mtx. Για παράδειγμα, η πράσινη καμπύλη (αριθμός 4) αντιστοιχεί στη συμμεταβλητή recipient.sex καθώς το λ ποικίλλει. Στο αριστερό άκρο του διαγράμματος, η νόρμα L1 είναι μηδενική και το λ είναι πολύ μεγάλο και κατά συνέπεια οι εκτιμήσεις των συντελεστών τείνουν και αυτές στο μηδέν (έχουμε κενό μοντέλο). Ενώ στο δεξί άκρο το λ είναι ουσιαστικά μηδέν και οι εκτιμήσεις των συντελεστών είναι ίδιες με αυτές που θα υπολόγιζε η μέθοδος των Ελαχίστων Τετραγώνων. Καθώς η νόρμα L1 αυξάνεται (και το λ μειώνεται) προστίθενται μεταβλητές στο μοντέλο μας ανάλογα με το πόσο επηρεάζουν τη διάρκεια ζωής των ασθενών μας. Παρατηρούμε ότι οι περισσότερες καμπύλες είναι μηδενικές σε όλο το τμήμα του διαγράμματος , και κάποιες φαίνεται να ξεκινούν προς το τέλος του διαγράμματος (για πολύ μικρές τιμές του λ). Οι μόνες καμπύλες που ξεχωρίζουν και ξεκινούν από μικρές τιμές της νόρμας L1 είναι οι 1 και 10 όπου επίσης αυξάνονται με πολύ μεγάλο ρυθμό. Αυτό μας δείχνει ότι στη μέθοδο Lasso οι μεταβλητές groupF και mtx έχουν μεγαλύτερη επίδραση στη διάρκεια ζωής των ασθενών μας.

Χρησιμοποιώντας την εντολή cv.glmnet μπορούμε να εφαρμόσουμε τη γενικευμένη μέθοδο Cross-Validation η οποία μας επιστρέφει τη βέλτιστη τιμή της παραμέτρου λ που θα χρησιμοποιήσουμε.

Πίνακας 5.18

lasso.mod = glmnet (x, y, family=”cox”, alpha =1, lambda =grid)
cv.out = cv.glmnet (x, y, family="cox", alpha =1)
bestlam = cv.out$lambda.min
bestlam = 0.07139441
lasso.pred = predict (lasso.mod, family="cox", s= bestlam , newx=x)
out = glmnet (x, y, family="cox", alpha =1, lambda =grid)
lasso.coef = predict (out, family="cox", type ="coefficients", s= bestlam) [1:10 ,]
> **lasso.coef**

groupF	recipient.age	donor.age	recipient.sex	donor.sex
0.2557804919	0.0000000	0.0000000	0.0000000	0.0000000
recipient.cmv	**donor.cmv**	**waiting.time**	**fab**	**mtx**
0.0000000	0.0000000	0.0000000	-0.0003144317	0.3955060693

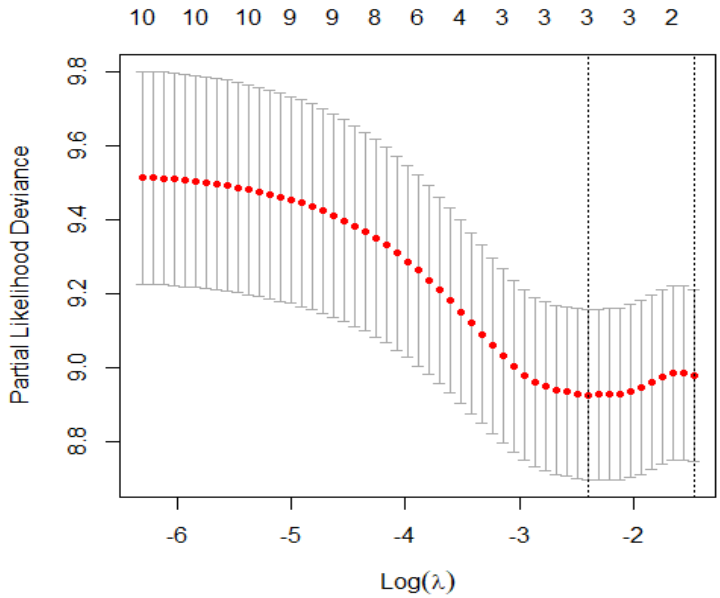
> **lasso.coef** [lasso.coef !=0]

groupF	fab	mtx
0.2557804919	-0.0003144317	0.3955060693

Όπως βλέπουμε στον παραπάνω πίνακα (5.18) η καλύτερη τιμή που μπορεί να πάρει το λ για τη μέθοδο Lasso είναι 0.07139441 και για αυτό το λ βρήκαμε τους συντελεστές του μοντέλου.

Παρατηρούμε ότι στο τελικό μοντέλο της Lasso (αυτό με το βέλτιστο λ) οι περισσότεροι συντελεστές είναι ακριβώς μηδέν. Δηλαδή το τελικό μοντέλο περιέχει μόνο δύο συμμεταβλητές, τις fab και mtx. Επίσης σε σύγκριση πάλι με το αρχικό μοντέλο του Cox που προσαρμόστηκε στην Παράγραφο 5.9.1, οι τιμές των δύο αυτών μεταβλητών είναι φανερά μειωμένες. Σε σχέση με τη μέθοδο Ridge, η μέθοδος Lasso πραγματοποιεί επιλογή συμμεταβλητών για το τελικό της μοντέλο όπως είδαμε.

Στο Σχήμα 5.23 παριστάνεται το διάγραμμα απόκλισης (deviance) της Cross-Validation μερικής Πιθανοφάνειας συναρτήσει του λογαριθμισμένου λ για την Παλινδρόμηση Lasso. Οι δύο κάθετες γραμμές συμβολίζουν δύο τιμές του λογαριθμισμένου λ τις οποίες έχει επιλέξει το μοντέλο. Και συγκεκριμένα η αριστερή κάθετη γραμμή καθορίζει τον λογάριθμο του βέλτιστου λ, όπως είδαμε είναι το 0.07139441 (log(0.07139441) = -2.63959) και επίσης μας δείχνει το σημείο όπου η απόκλιση (deviance) ελαχιστοποιείται. Ενώ η δεξιά κάθετη γραμμή μας δίνει τη μέγιστη τιμή του λ σε απόσταση μίας τυπικής απόκλισης από την ελάχιστη τιμή της απόκλισης. Τέλος, στο πάνω μέρος του διαγράμματος εμφανίζεται το πλήθος των μη μηδενικών παραμέτρων καθώς το log(λ), άρα και το ίδιο το λ, αυξάνεται. Έτσι λοιπόν, για μικρές τιμές του λ στο μοντέλο συμπεριλαμβάνονται και οι 10 συμμεταβλητές ενώ όσο μεγαλώνει το λ τόσο λιγοστεύουν οι μη μηδενικές παράμετροι του μοντέλου.



Σχήμα 5.23: *Διάγραμμα απόκλισης της Cross-Validation μερικής Πιθανοφάνειας συναρτήσει του log(λ) για το μοντέλο της Lasso.*

Σε σύγκριση με τη μέθοδο Ridge (Παράγραφος 5.9.3), η μέθοδος Lasso είναι πιο αποτελεσματική και πιο ακριβής γιατί όπως είδαμε κάνει και επιλογή μεταβλητών, δηλαδή αφαιρεί όσες συμμεταβλητές δεν επηρεάζουν το μοντέλο μας, κάτι το οποίο η Ridge δεν κάνει.

5.10 Συμπεράσματα

Στην εφαρμογή που μελετήσαμε σε αυτό το κεφάλαιο, είχαμε ένα σχετικά μεγάλο δείγμα 137 ασθενών που υποβλήθηκαν σε μεταμόσχευση μυελού των οστών και μελετήσαμε τη διάρκεια ζωής τους σε ημέρες μέχρι να τους συμβεί το δυσάρεστο γεγονός, δηλαδή ο θάνατος. Είχαμε 10 επεξηγηματικές μεταβλητές και είχαμε μία κατηγορική μεταβλητή, τη group, η οποία χώριζε τους ασθενείς μας σε ομάδες ανάλογα με το τύπο λευχαιμίας που είχε ο κάθε ασθενής σε τρεις ομάδες, στην Ομάδα 1 με λεμφοβλαστική λευχαιμία (ALL), στην Ομάδα 2 με μυελοκυτταρική λευχαιμία χαμηλού ρίσκου (AML low_risk) και στην Ομάδα 3 με οξεία μυελοκυτταρική λευχαιμία υψηλού ρίσκου (AML high_risk).

Αρχικά, ελέγξαμε με διάφορες μεθόδους αν τα δεδομένα μας προσαρμόζονταν σύμφωνα με κάποιες γνωστές κατανομές (Weibull, Log-Normal και Log-Logistic) και καταλήξαμε στο συμπέρασμα ότι η κατανομή Log-Normal φαίνεται να είναι η πιο κατάλληλη από τις τρεις. Στη συνέχεια μελετήσαμε κάθε μία ομάδα χωριστά με σκοπό να βρούμε ποια μορφή λευχαιμίας (δηλαδή ποια ομάδα) επιφέρει μεγαλύτερο κίνδυνο στους ασθενείς μας και ελέγξαμε για τους ασθενείς της κάθε ομάδας ποια βασική κατανομή (Weibull, Log-Normal και Log-Logistic) προσαρμόζεται καλύτερα. Καταλήξαμε στα παρακάτω συμπεράσματα:

- Οι ασθενείς που έχουν AML high_risk (Ομάδα 3) κινδυνεύουν περισσότερο να τους συμβεί το δυσάρεστο γεγονός (ο θάνατος)
- Οι ασθενείς που έχουν AML low_risk (Ομάδα 2) έχουν τις μικρότερες πιθανότητες να τους συμβεί το δυσάρεστο γεγονός (ο θάνατος)
- Στους ασθενείς που έχουν ALL (Ομάδα 1) προσαρμόζεται καλύτερα η κατανομή Log-Logistic
- Στους ασθενείς που έχουν AML low_risk (Ομάδα 2) προσαρμόζεται καλύτερα η κατανομή Log-Normal
- Στους ασθενείς που έχουν AML high_risk (Ομάδα 3) προσαρμόζεται καλύτερα η κατανομή Log-Logistic

Στη συνέχεα, ελέγξαμε εάν στα δεδομένα μας ταιριάζει ένα μοντέλο Παλινδρόμησης Επιταχυνόμενης Διάρκειας Ζωής (AL) στα δεδομένα μας και με τη βοήθεια των υπολοίπων Cox-Snell συμπεράναμε ότι με την κατανομή Log-Logistic ένα μοντέλο Επιταχυνόμενης Διάρκειας Ζωής (AL) φαίνεται να ταιριάζει στα δεδομένα μας, το οποίο προσαρμόσαμε με κατηγορία αναφοράς τη groupF2 (της Ομάδας 2). Το βέλτιστο μοντέλο της Επιταχυνόμενης Διάρκειας Ζωής (AL) που προσαρμόσαμε είναι (Πίνακας 5.19):

Πίνακας 5.19

	Value	Std. Error	z	p
(intercept)	7.9752	0.3682	21.66	<2e-16
groupF1	-1.3741	0.4726	-2.91	0.00364
groupF3	-1.3945	0.4189	-3.33	0.00087
fab	-1.1785	0.4238	-2.78	0.00543
mtx	-0.8924	0.3896	-2.29	0.02298
log(scale)	0.0950	0.0943	1.01	0.31371

Έπειτα, ελέγξαμε εάν στα δεδομένα μας ταιριάζει ένα μοντέλο Παλινδρόμησης Αναλογικής Διακινδύνευσης (PH), όπου με γραφικές μεθόδους συμπεράναμε ότι δεν φαίνεται να ταιριάζει τόσο

καλά. Για αυτό προσαρμόσαμε το ημι-παραμετρικό μοντέλο Αναλογικής Διακινδύνευσης του Cox, με κατηγορία αναφοράς τη groupF2 (της Ομάδας 2 AML low_risk), και το μελετήσαμε εις βάθος, χρησιμοποιώντας διάφορες μεθόδους όπως τις καμπύλες ROC και AUC, τις μεθόδους Συρρίκνωσης Ridge και Lasso. Οπότε κατά τη διάρκεια της μελέτης μας καταλήξαμε σε διάφορες μορφές μοντέλων οι οποίες είναι οι ακόλουθες (Πίνακας 5.20).

Πίνακας 5.20

	Μοντέλο του Cox	**Βέλτιστο μοντέλο του Cox**	**Μοντέλο Ridge**	**Μοντέλο Lasso**
groupF1	1.0624620	0.9045	-	-
groupF3	0.8742014	0.8525	-	-
groupF	-	-	5.818096e-02	0.2557804919
recipient.age	0.0139626	0	1.093994e-03	0
donor.age	-0.0021921	0	1.210247e-03	0
recipient.sex	-0.1093030	0	-2.693975e-02	0
donor.sex	0.0333095	0	2.635508e-04	0
recipient.cmv	-0.0606449	0	1.494924e-02	0
donor.cmv	-0.0480256	0	3.408057e-03	0
waiting.time	-0.0003417	0	-1.115484e-05	0
fab	0.8018912	0.7695	8.089831e-02	-0.0003144317
mtx	0.2918278	0	4.671426e-02	0.3955060693

Το αρχικό μοντέλο του Cox (Πίνακας 5.20) περιέχει όλες τις συμμεταβλητές και θεωρεί στατιστικά πιο σημαντικές τις συμμεταβλητές groupF1, groupF2 και fab, καθώς έχουν πολύ μικρή τιμή p-value (<0.005) και επηρεάζουν αρνητικά σε μεγάλο βαθμό τη διάρκεια ζωής των ασθενών μας. Πιο συγκεκριμένα, οι τρεις αυτές συμμεταβλητές αυξάνουν τον κίνδυνο θανάτου των ασθενών, καθώς αντίστοιχα για κάθε συμμεταβλητή έχουμε $\exp(1.0624620) = 2.8934861 > 1$, $\exp(0.8742014) = 2.3969603 > 1$ και $\exp(0.8018912) = 2.1587 > 1$.

Το "βέλτιστο" μοντέλο του Cox (Πίνακας 5.20) προέκυψε μετά την προσαρμογή του κλασικού μοντέλου του Cox και την εκτέλεση της μεθόδου διαδοχικής αφαίρεσης (backward elimination) για την επιλογή των στατιστικά σημαντικότερων μεταβλητών. Περιέχει συνολικά μόνο 3 συμμεταβλητές. Δεν παρουσιάζει αρκετά καλή προβλεπτική ικανότητα σύμφωνα με τις καμπύλες ROC, αλλά ειδικά για χρόνους διάρκειας ζωής μεταξύ 300 και 700 ημερών η πρόβλεψη βελτιώνεται.

Το μοντέλο της μεθόδου Ridge (Πίνακας 5.20) χρησιμοποιήθηκε στο αρχικό μοντέλο του Cox για την αντιμετώπιση του προβλήματος της πολυσυγγραμμικότητας των μεταβλητών. Η μέθοδος Ridge συρρικνώνει τις εκτιμήτριες των συντελεστών του μοντέλου αλλά δεν μηδενίζει καμία από αυτές οπότε αδυνατεί να επιλέξει το σύνολο των στατιστικά σημαντικότερων συμμεταβλητών για τη περιγραφή των δεδομένων μας.

Το μοντέλο της μεθόδου Lasso (Πίνακας 5.20) χρησιμοποιήθηκε στο αρχικό μοντέλο του Cox για την αντιμετώπιση του προβλήματος της πολυσυγγραμμικότητας των μεταβλητών. Η μέθοδος Lasso, σε αντίθεση με την Ridge, καταλήγει σε μια μορφή μοντέλου όπου οι περισσότερες παράμετροι του μοντέλου έχουν μηδενιστεί και όλοι οι υπόλοιποι έχουν υποστεί συρρίκνωση. Έτσι λοιπόν καταλήξαμε σε ένα μοντέλο με μόνο 3 συμμεταβλητές. Επίσης, σημαντικές είναι και η κατηγορική συμμεταβλητή groupF και η fab, το οποίο συμφωνεί και με τα αποτελέσματα του "βέλτιστου" μοντέλου του Cox.

Βιβλιογραφία

Καρώνη Χ. (2009): *Μοντέλα Αξιοπιστίας και Επιβίωσης,* Εκδόσεις Συμεών, Αθήνα

Καρώνη Χ., Οικονόμου Π. (2017): *Στατιστικά Μοντέλα Παλινδρόμησης με χρήση Minitab και R,* Εκδόσεις Συμεών, Αθήνα

Arashi M, Roozbeh M, Hamzah NA, Gasparini M (2021): *Ridge regression and its applications in genetic studies.* PLoS ONE 16(4): e0245376 (https://doi.org/10.1371/journal.pone.0245376)

Anwar F. (2013): *Several Types of Residuals in Cox Regression Model: An Empirical Study,* Int. Journal of Math. Analysis, Vol. 7, 2013, no. 53, 2645 – 2654 (http://www.m-hikari.com/ijma/ijma-2013/ijma-53-56-2013/fitriantoIJMA53-56-2013.pdf)

Collett D. (2003): *Modelling Survival Data in Medical Research.* (2nd edition) Boca Raton: Chapman & Hall/CRC

Cox D.R. (1975): *Partial likelihood,* Biometrica, 62, 269-276, DOI: 10.1093/biomet/62.2.269

Cox D.R. (1972): *Regression models and life tables,* Journal of the Royal Statistical Society B, 34, 187 220 (http://www.biecek.pl/statystykaMedyczna/cox.pdf)

Fox J., Weisberg S. (2018): *Cox Proportional-Hazards Regression for Survival Data in R, An Appendix to An R Companion to Applied Regression*, (3rd edition), last revision: 2018-09-28, (https://socserv.socsci.mcmaster.ca/jfox/Books/Companion/appendix/Appendix-Cox-Regression.pdf)

Heagerty P., Zheng Y. (2005): *Survival Model Predictive Accuracy and ROC Curves,* Biometrics 61, 92 105 March 2005, (http://staff.pubhealth.ku.dk/~tag/Teaching/share/material/heagerty-zheng-roc-survival.pdf)

Hoerl A. & Kennard R. (1970): *Ridge regression: biased estimation for the non-orthogonal problems,* Technometrics, 12, 55-67 (https://www.math.arizona.edu/~hzhang/math574m/Read/RidgeRegressionBiasedEstimationForNonorthogonalProblems.pdf)

Hosmer D, Lemeshow S., Sturdivant R.X.. (2013): *Applied Logistic Regression,* (3rd edition) Published by John Wiley & Sons, Inc., Hoboken, New Jersey, Published simultaneously in Canada

James G., Witten D., Hastie T., Tibshirani R. (2013): *An Introduction to Statistical Learning with Applications in R,* Springer Texts in Statistics, Published by Springer Science+Business Media New York 2013 (Corrected at 8^{th} printing 2017)

Narkhede S. (2018): *Understanding AUC - ROC Curve* (https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5)

Rusmadi G., Saefuddin A., Sartono B. (2017): *Applied Ridge and LASSO Methods in Cox Proportional Hazard Modelling,* International Journal of Scientific & Engineering Research (https://www.ijser.org/researchpaper/Applied-Ridge-and-LASSO-Methods-in-Cox-Proportional-Hazard-Modelling.pdf)

Sunhee P. and David H.J. (2015). *Reassessing Schoenfeld residual tests of proportional hazards in political science event history analyses*. American Journal of Political Science http://eprints.lse.ac.uk/84988/1/06_ParkHendry2015ReassessingSchoenfeldTests_Final.pdf

Tibshirani R. (1997): *The Lasso method for variable selection in the Cox model,* Statistics in Medicine

Tibshirani R. (1996): *Regression shrinkage and selection via the LASSO,* Journal of the Royal Statistical Society Series B

Wang, F., Mukherjee, S., Richardson, S. (2020): *High-dimensional regression in practice: an empirical study of finite-sample prediction, variable selection, and ranking,* Stat Comput (https://doi.org/10.1007/s11222-019-09914-9)

Wieringen W. (2021): *Lecture notes on ridge regression,* Version 0.40, May 28, 2021. (https://arxiv.org/pdf/1509.09169.pdf)

Xu R., Vaida F., Harrington D.P (2009): *Using profile likelihood for semiparametric model selection with application to proportional hazards mixed models,* Statistica Sinica (http://www3.stat.sinica.edu.tw/statistica/oldpdf/A19n223.pdf)

R-project: *Coxnet (2021),* https://cran.r-project.org/web/packages/glmnet/vignettes/Coxnet.pdf
Survival (2021), https://cran.r-project.org/web/packages/survival/survival.pdf
Glmnet (2021), https://cran.r-project.org/web/packages/glmnet/glmnet.pdf
RisksetROC (2015), https://cran.r-project.org/web/packages/risksetROC/risksetROC.pdf

Παράρτημα

Α) Πίνακες που χρησιμοποιήθηκαν στην εφαρμογή των ασθενών με τη μεταμόσχευση μυελού των οστών.

Π.1: Πίνακας εκτίμησης Kaplan-Meier

Time	Number at Risk	Number Failed	Survival Probability (SKM)	Standard Error	Lower 95% Normal CI	Upper 95% Normal CI
1	137	1	0.992701	0.0072726	0.978447	1.00000
2	136	1	0.985401	0.0102471	0.965318	1.00000
10	135	1	0.978102	0.0125035	0.953596	1.00000
16	134	1	0.970803	0.0143838	0.942611	0.99899
32	133	1	0.963504	0.0160211	0.932103	0.99490
35	132	1	0.956204	0.0174836	0.921937	0.99047
47	131	2	0.941606	0.0200336	0.902341	0.98087
48	129	2	0.927007	0.0222239	0.883449	0.97057
53	127	1	0.919708	0.0232167	0.874204	0.96521
55	126	1	0.912409	0.0241526	0.865070	0.95975
63	125	1	0.905109	0.0250381	0.856036	0.95418
64	124	1	0.897810	0.0258783	0.847090	0.94853
74	123	2	0.883212	0.0274392	0.829432	0.93699
76	121	1	0.875912	0.0281666	0.820707	0.93112
79	120	1	0.868613	0.0288622	0.812044	0.92518
80	119	2	0.854015	0.0301666	0.794889	0.91314
84	117	1	0.846715	0.0307792	0.786389	0.90704
86	116	1	0.839416	0.0313675	0.777937	0.90090
93	115	1	0.832117	0.0319327	0.769530	0.89470
100	114	1	0.824818	0.0324761	0.761166	0.88847
104	113	1	0.817518	0.0329988	0.752842	0.88219
105	112	2	0.802920	0.0339858	0.736309	0.86953
107	110	1	0.795620	0.0344518	0.728096	0.86314
109	109	1	0.788321	0.0349004	0.719918	0.85672
110	108	1	0.781022	0.0353323	0.711772	0.85027
113	107	1	0.773723	0.0357481	0.703658	0.84379
115	106	1	0.766423	0.0361484	0.695574	0.83727
120	105	1	0.759124	0.0365336	0.687519	0.83073
122	104	2	0.744526	0.0372609	0.671496	0.81756
129	102	1	0.737226	0.0376037	0.663524	0.81093
157	101	1	0.729927	0.0379332	0.655579	0.80427
162	100	1	0.722628	0.0382497	0.647660	0.79760
164	99	1	0.715328	0.0385536	0.639765	0.79089
168	98	1	0.708029	0.0388450	0.631894	0.78416
172	97	1	0.700730	0.0391243	0.624048	0.77741
183	96	1	0.693431	0.0393918	0.616224	0.77064
192	95	1	0.686131	0.0396476	0.608423	0.76384
194	94	1	0.678832	0.0398921	0.600645	0.75702
211	93	1	0.671533	0.0401254	0.592889	0.75018
219	92	1	0.664234	0.0403477	0.585154	0.74331
230	90	1	0.656853	0.0405688	0.577340	0.73637
242	89	1	0.649473	0.0407788	0.569548	0.72940
248	88	1	0.642092	0.0409778	0.561777	0.72241
268	87	1	0.634712	0.0411660	0.554028	0.71540
272	86	1	0.627332	0.0413437	0.546300	0.70836

273	85	1	0.619951	0.0415108	0.538592	0.70131
276	84	1	0.612571	0.0416675	0.530904	0.69424
288	83	1	0.605191	0.0418140	0.523237	0.68714
318	82	1	0.597810	0.0419504	0.515589	0.68003
332	81	1	0.590430	0.0420767	0.507961	0.67290
363	80	1	0.583049	0.0421930	0.500353	0.66575
381	79	1	0.575669	0.0422995	0.492764	0.65857
383	78	1	0.568289	0.0423962	0.485194	0.65138
390	77	2	0.553528	0.0425604	0.470111	0.63694
414	75	1	0.546148	0.0426280	0.462598	0.62970
418	74	1	0.538767	0.0426861	0.455104	0.62243
421	73	1	0.531387	0.0427346	0.447629	0.61515
422	72	1	0.524006	0.0427736	0.440172	0.60784
456	71	1	0.516626	0.0428032	0.432733	0.60052
466	70	1	0.509246	0.0428232	0.425314	0.59318
467	69	1	0.501865	0.0428339	0.417912	0.58582
481	68	1	0.494485	0.0428351	0.410530	0.57844
486	67	1	0.487105	0.0428268	0.403166	0.57104
487	66	1	0.479724	0.0428092	0.395820	0.56363
526	65	1	0.472344	0.0427820	0.388493	0.55620
606	63	1	0.464846	0.0427549	0.381048	0.54864
609	62	1	0.457349	0.0427176	0.373624	0.54107
625	61	1	0.449851	0.0426702	0.366219	0.53348
641	60	1	0.442354	0.0426126	0.358835	0.52587
662	59	1	0.434856	0.0425448	0.351470	0.51824
677	58	1	0.427359	0.0424668	0.344125	0.51059
704	57	1	0.419861	0.0423784	0.336801	0.50292
748	56	1	0.412364	0.0422797	0.329497	0.49523
1063	47	1	0.403590	0.0422807	0.320721	0.48646
1074	46	1	0.394816	0.0422621	0.311984	0.47765
2204	9	1	0.350948	0.0558734	0.241438	0.46046

Π.2: Πίνακας συναρτήσεων Weibull, Log-Normal και Log-Logistic

Time	Weibull	Log-normal	Log-logistic	Time	Weibull	Log-normal	Log-logistic
1	-4.91632	-2.44219	-4.91265	211	-0.92082	-0.44415	-0.71513
2	-4.21949	-2.18082	-4.21213	219	-0.89374	-0.42405	-0.68222
10	-3.81032	-2.01604	-3.79923	230	-0.86680	-0.40389	-0.64930
16	-3.51891	-1.89273	-3.50405	242	-0.84027	-0.38390	-0.61672
32	-3.29201	-1.79288	-3.27336	248	-0.81413	-0.36406	-0.58446
35	-3.10591	-1.70824	-3.08344	268	-0.78837	-0.34436	-0.55249
47	-2.81061	-1.56840	-2.78037	272	-0.76297	-0.32479	-0.52079
48	-2.57974	-1.45386	-2.54160	273	-0.73791	-0.30535	-0.48934
53	-2.48053	-1.40311	-2.43839	276	-0.71317	-0.28603	-0.45813
55	-2.38959	-1.35574	-2.34341	288	-0.68873	-0.26681	-0.42714
63	-2.30560	-1.31123	-2.25533	318	-0.66459	-0.24768	-0.39635
64	-2.22751	-1.26917	-2.17313	332	-0.64074	-0.22865	-0.36574
74	-2.08594	-1.19120	-2.02320	363	-0.61714	-0.20970	-0.33530
76	-2.02125	-1.15479	-1.95428	381	-0.59380	-0.19083	-0.30502
79	-1.96001	-1.11986	-1.88875	383	-0.57071	-0.17202	-0.27487
80	-1.84638	-1.05381	-1.76644	390	-0.52519	-0.13458	-0.21494
84	-1.79342	-1.02245	-1.70907	414	-0.50275	-0.11593	-0.18512
86	-1.74269	-0.99206	-1.65389	418	-0.48050	-0.09733	-0.15538
93	-1.69400	-0.96256	-1.60070	421	-0.45845	-0.07876	-0.12571
100	-1.64718	-0.93388	-1.54933	422	-0.43657	-0.06021	-0.09610
104	-1.60205	-0.90595	-1.49962	456	-0.41486	-0.04169	-0.06653
105	-1.51640	-0.85210	-1.40464	466	-0.39330	-0.02318	-0.03699

107	-1.47564	-0.82608	-1.35914	467	-0.37190	-0.00468	-0.00746
109	-1.43612	-0.80061	-1.31484	481	-0.35064	0.01382	0.02206
110	-1.39775	-0.77565	-1.27163	486	-0.32951	0.03233	0.05159
113	-1.36046	-0.75116	-1.22945	487	-0.30851	0.05085	0.08115
115	-1.32418	-0.72712	-1.18822	526	-0.28762	0.06938	0.11074
120	-1.28884	-0.70349	-1.14788	606	-0.26651	0.08823	0.14085
122	-1.22075	-0.65736	-1.06962	609	-0.24551	0.10712	0.17102
129	-1.18790	-0.63482	-1.03160	625	-0.22460	0.12604	0.20127
157	-1.15578	-0.61259	-0.99425	641	-0.20378	0.14500	0.23161
162	-1.12436	-0.59067	-0.95753	662	-0.18303	0.16402	0.26206
164	-1.09358	-0.56902	-0.92141	677	-0.16236	0.18310	0.29264
168	-1.06343	-0.54764	-0.88583	704	-0.14176	0.20225	0.32334
172	-1.03386	-0.52650	-0.85078	748	-0.12121	0.22147	0.35420
183	-1.00484	-0.50560	-0.81621	1063	-0.09722	0.24407	0.39053
192	-0.97634	-0.48491	-0.78209	1074	-0.07329	0.26679	0.42711
194	-0.94835	-0.46444	-0.74841	2204	0.04604	0.38276	0.61488

Π.3: Πίνακας εκτίμησης Μέγιστης Πιθανοφάνειας της συνάρτησης Επιβίωσης

Time	SML_Weibull	SML_Log-normal	SML_Log-logistic	Time	SML_Weibull	SML_Log-normal	SML_Log-logistic
1	0.986332	0.998698	0.993844	211	0.726569	0.709791	0.707867
2	0.979532	0.996454	0.989442	219	0.721457	0.703910	0.701792
10	0.948150	0.974608	0.963649	230	0.714608	0.696084	0.693683
16	0.932229	0.958844	0.948277	242	0.707362	0.687867	0.685140
32	0.899915	0.922060	0.914111	248	0.703821	0.683875	0.680980
35	0.894798	0.915884	0.908427	268	0.692389	0.671096	0.667617
47	0.876184	0.892975	0.887283	272	0.690167	0.668631	0.665032
48	0.874744	0.891182	0.885620	273	0.689615	0.668020	0.664391
53	0.867754	0.882450	0.877506	276	0.687966	0.666195	0.662476
55	0.865050	0.879062	0.874347	288	0.681481	0.659054	0.654969
63	0.854696	0.866056	0.862167	318	0.665992	0.642210	0.637195
64	0.853449	0.864488	0.860692	332	0.659088	0.634796	0.629344
74	0.841491	0.849437	0.846465	363	0.644456	0.619277	0.612864
76	0.839200	0.846556	0.843726	381	0.636341	0.610781	0.603817
79	0.835822	0.842309	0.839679	383	0.635455	0.609858	0.602834
80	0.834711	0.840913	0.838346	390	0.632380	0.606662	0.599426
84	0.830337	0.835420	0.833091	414	0.622115	0.596075	0.588125
86	0.828190	0.832727	0.830508	418	0.620444	0.594364	0.586296
93	0.820877	0.823571	0.821689	421	0.619198	0.593090	0.584935
100	0.813849	0.814801	0.813193	422	0.618785	0.592667	0.584483
104	0.809950	0.809950	0.808474	456	0.605103	0.578799	0.569646

105	0.808988	0.808755	0.807308	466	0.601216	0.574898	0.565469
107	0.807079	0.806385	0.804995	467	0.600831	0.574512	0.565055
109	0.805188	0.804041	0.802704	481	0.595496	0.569188	0.559351
110	0.804250	0.802879	0.801567	486	0.593617	0.567320	0.557350
113	0.801463	0.799431	0.798187	487	0.593243	0.566949	0.556952
115	0.799627	0.797164	0.795961	526	0.579072	0.552994	0.541991
120	0.795112	0.791600	0.790484	606	0.552269	0.527188	0.514315
122	0.793335	0.789415	0.788328	609	0.551317	0.526285	0.513348
129	0.787238	0.781942	0.780932	625	0.546300	0.521541	0.508263
157	0.764532	0.754446	0.753429	641	0.541380	0.516915	0.503306
162	0.760725	0.749891	0.748831	662	0.535065	0.511012	0.496983
164	0.759221	0.748097	0.747015	677	0.530649	0.506908	0.492588
168	0.756245	0.744552	0.743426	704	0.522890	0.499742	0.484921
172	0.753309	0.741067	0.739888	748	0.510739	0.488635	0.473051
183	0.745434	0.731767	0.730419	1063	0.437794	0.424690	0.405248
192	0.739195	0.724450	0.722938	1074	0.435605	0.422838	0.403303
194	0.737832	0.722858	0.721307	2204	0.281448	0.299853	0.277729

Π.4: Πίνακας εκτίμησης Kaplan-Meier για την Ομάδα 1

Time_1	Number at Risk (n_j)	Number Failed (d_j)	Survival Probability 1	Standard Error	Lower 95% Normal CI	Upper 95% Normal CI
1	38	1	0.973684	0.0259672	0.922789	1.00000
55	37	1	0.947368	0.0362235	0.876372	1.00000
74	36	1	0.921053	0.0437441	0.835316	1.00000
86	35	1	0.894737	0.0497845	0.797161	0.99231
104	34	1	0.868421	0.0548361	0.760944	0.97590
107	33	1	0.842105	0.0591528	0.726168	0.95804
109	32	1	0.815789	0.0628861	0.692535	0.93904
110	31	1	0.789474	0.0661348	0.659852	0.91910
122	30	2	0.736842	0.0714338	0.596834	0.87685
129	28	1	0.710526	0.0735704	0.566331	0.85472
172	27	1	0.684211	0.0754053	0.536419	0.83200
192	26	1	0.657895	0.0769602	0.507055	0.80873
194	25	1	0.631579	0.0782518	0.478208	0.78495
230	23	1	0.604119	0.0795218	0.448259	0.75998
276	22	1	0.576659	0.0805088	0.418865	0.73445
332	21	1	0.549199	0.0812232	0.390005	0.70839
383	20	1	0.521739	0.0816721	0.361665	0.68181
418	19	1	0.494279	0.0818598	0.333837	0.65472
466	18	1	0.466819	0.0817882	0.306517	0.62712
487	17	1	0.439359	0.0814566	0.279707	0.59901
526	16	1	0.411899	0.0808617	0.253413	0.57039

609	14	1	0.382478	0.0802600	0.225171	0.53978
662	13	1	0.353057	0.0792956	0.197640	0.50847

Π.5: Πίνακας εκτίμησης Kaplan-Meier για την Ομάδα 2

Time_2	Number at Risk (n_j)	Number Failed (d_j)	Survival Probability 2	Standard Error	Lower 95% Normal CI	Upper 95% Normal CI
10	54	1	0.981481	0.018346	0.945523	1.00000
35	53	1	0.962963	0.025700	0.912593	1.00000
48	52	1	0.944444	0.031171	0.883350	1.00000
53	51	1	0.925926	0.035639	0.856075	0.99578
79	50	1	0.907407	0.039445	0.830097	0.98472
80	49	1	0.888889	0.042767	0.805068	0.97271
105	48	1	0.870370	0.045710	0.780781	0.95996
211	47	1	0.851852	0.048343	0.757101	0.94660
219	46	1	0.833333	0.050715	0.733934	0.93273
248	45	1	0.814815	0.052861	0.711209	0.91842
272	44	1	0.796296	0.054807	0.688876	0.90372
288	43	1	0.777778	0.056575	0.666893	0.88866
381	42	1	0.759259	0.058180	0.645229	0.87329
390	41	1	0.740741	0.059635	0.623858	0.85762
414	40	1	0.722222	0.060952	0.602759	0.84169
421	39	1	0.703704	0.062139	0.581914	0.82549
481	38	1	0.685185	0.063203	0.561310	0.80906
486	37	1	0.666667	0.064150	0.540935	0.79240
606	36	1	0.648148	0.064986	0.520778	0.77552
641	35	1	0.629630	0.065715	0.500831	0.75843
704	34	1	0.611111	0.066340	0.481087	0.74114
748	33	1	0.592593	0.066865	0.461541	0.72364
1063	26	1	0.569801	0.068067	0.436393	0.70321
1074	25	1	0.547009	0.069055	0.411664	0.68235
2204	6	1	0.455840	0.101182	0.257527	0.65415

Π.6: Πίνακας εκτίμησης Kaplan-Meier για την Ομάδα 3

Time_3	Number at Risk (n_j)	Number Failed (d_j)	Survival Probability 3	Standard Error	Lower 95% Normal CI	Upper 95% Normal CI
2	45	1	0.977778	0.0219739	0.934710	1.00000
16	44	1	0.955556	0.0307207	0.895344	1.00000
32	43	1	0.933333	0.0371849	0.860452	1.00000
47	42	2	0.888889	0.0468486	0.797067	0.98071
48	40	1	0.866667	0.0506745	0.767347	0.96599
63	39	1	0.844444	0.0540284	0.738551	0.95034
64	38	1	0.822222	0.0569937	0.710517	0.93393
74	37	1	0.800000	0.0596285	0.683130	0.91687
76	36	1	0.777778	0.0619748	0.656309	0.89925
80	35	1	0.755556	0.0640644	0.629992	0.88112
84	34	1	0.733333	0.0659218	0.604129	0.86254
93	33	1	0.711111	0.0675660	0.578684	0.84354

100	32	1	0.688889	0.0690122	0.553627	0.82415
105	31	1	0.666667	0.0702728	0.528934	0.80440
113	30	1	0.644444	0.0713576	0.504586	0.78430
115	29	1	0.622222	0.0722744	0.480567	0.76388
120	28	1	0.600000	0.0730297	0.456864	0.74314
157	27	1	0.577778	0.0736283	0.433469	0.72209
162	26	1	0.555556	0.0740741	0.410373	0.70074
164	25	1	0.533333	0.0743698	0.387571	0.67910
168	24	1	0.511111	0.0745172	0.365060	0.65716
183	23	1	0.488889	0.0745172	0.342838	0.63494
242	22	1	0.466667	0.0743698	0.320905	0.61243
268	21	1	0.444444	0.0740741	0.299262	0.58963
273	20	1	0.422222	0.0736283	0.277913	0.56653
318	19	1	0.400000	0.0730297	0.256864	0.54314
363	18	1	0.377778	0.0722744	0.236122	0.51943
390	17	1	0.355556	0.0713576	0.215697	0.49541
422	16	1	0.333333	0.0702728	0.195601	0.47107
456	15	1	0.311111	0.0690122	0.175850	0.44637
467	14	1	0.288889	0.0675660	0.156462	0.42132
625	13	1	0.266667	0.0659218	0.137462	0.39587
677	12	1	0.244444	0.0640644	0.118880	0.37001

Π.7: Πίνακας εκτίμησης Nelson-Aalen

H(t)_1	H(t)_2	H(t)_3
0.02632	0.018519	0.02222
0.05334	0.037386	0.04495
0.08112	0.056617	0.06821
0.10969	0.076225	0.11582
0.13910	0.096225	0.14082
0.16941	0.116633	0.16647
0.20066	0.137467	0.19278
0.23291	0.158743	0.21981
0.29958	0.180482	0.24759
0.33530	0.202705	0.27616
0.37233	0.225432	0.30557
0.41079	0.248688	0.33587
0.45079	0.272497	0.36712
0.49427	0.296887	0.39938
0.53973	0.321887	0.43271
0.58735	0.347528	0.46720
0.63735	0.373844	0.50291
0.68998	0.400871	0.53995
0.74553	0.428649	0.57841
0.80436	0.457220	0.61841
0.86686	0.486632	0.66008
0.93829	0.516935	0.70355
1.01521	0.555397	0.74901
	0.595397	0.79663
	0.762063	0.84663
		0.89926
		0.95481
		1.01364

		1.07614
		1.14281
		1.21423
		1.29116
		1.37449

B) Συναρτήσεις που χρησιμοποιήθηκαν στο στατιστικό πακέτο R στις εφαρμογές για τη διάρκεια ζωής των λαμπτήρων φθορισμού και για τη διάρκεια ζωής ασθενών μετά από μεταμόσχευση μυελού των οστών.

- **survreg (Surv(time,indicator) ~groupF+ recipient.age+ donor.age +recipient.sex+ donor.sex+ recipient.cmv+ donor.cmv+ waiting.time+ fab+ mtx, data=data, dist="loglogistic"):**
  Η συνάρτηση αυτή προσαρμόζει το Μοντέλο της Επιταχυνόμενης Διάρκειας Ζωής (AL), χρησιμοποιώντας ως κατηγορική μεταβλητή την συμμεταβλητή group, όπου groupF <- factor(data$group, levels=c(2,1,3)). Επίσης έχουμε συμπεριλάβει (με το σύμβολο "+") όλες τις συμμεταβλητές του μοντέλου μας.

- **coxph (Surv(time,indicator) ~groupF+ recipient.age+ donor.age+ recipient.sex+ donor.sex+ recipient.cmv+ donor.cmv+ waiting.time+ fab+ mtx, data=data):**
  Η συνάρτηση αυτή προσαρμόζει το μοντέλο της Αναλογικής Διακινδύνευσης του Cox, χρησιμοποιώντας ως κατηγορική μεταβλητή την συμμεταβλητή group, όπου groupF <- factor(data$group, levels=c(2,1,3)). Επίσης και αυτή η συνάρτηση συμπεριλαμβάνει (με το σύμβολο "+") όλες τις συμμεταβλητές του μοντέλου μας.

- **cox.zph (modC2, transform="identity", terms=FALSE):**
  Η συνάρτηση αυτή υπολογίζει τα υπόλοιπα Schoenfeld και κατά συνέπεια ελέγχει αν ισχύει η υπόθεση της αναλογικής διακινδύνευσης στο μοντέλο του Cox. Πρώτα εισάγουμε το προσαρμοσμένο μοντέλο του Cox που δημιουργήθηκε με τη χρήση της συνάρτησης coxph (εδώ έχουμε modC2). Ο δείκτης transform είναι ένας χαρακτήρας τύπου «string» που προσδιορίζει με ποιο τρόπο θα μετατραπούν οι χρόνοι επιβίωσης πριν γίνει ο έλεγχος. Πιθανές τιμές είναι: "km", "rank", "identity", ή κάποια άλλη συνάρτηση.

- **step (mod, direction="backward", test="Chisq"):**
  Αυτή η συνάρτηση εφαρμόζει τη τεχνική με βήματα, δηλαδή τη διαδικασία διαδοχικής αφαίρεσης ή της διαδοχικής πρόσθεσης μεταβλητών, με σκοπό να φτάσει στο καλύτερο δυνατό μοντέλο. Εδώ έχουμε επιλέξει direction= "backward" γιατί εφαρμόζουμε τη διαδικασία διαδοχικής αφαίρεσης (backward elimination).

- **confint.default (mod):**
  Αυτή η συνάρτηση υπολογίζει τα διαστήματα εμπιστοσύνης σε επίπεδο σημαντικότητας 95% ενός μοντέλου για όλες τις συμμεταβλητές που περιέχει.

- **Για τις καμπύλες ROC & AUC:**
  - eta<-modC2$linear.predictor
  - ROC300= risksetROC (Stime= data$time, status=data$indicator, marker=eta, predict.time=300, method="Cox", main="ROC Curve", lty=2, col="red", ylab="True Positive", xlab="False Positive")
  - ROC1000= risksetROC (Stime= data$time, status=data$indicator, marker=eta, predict.time=1000, method="Cox", plot=FALSE)
  - risksetAUC (Stime= data$time, status=data$indicator, marker=eta, method="Cox", tmax=2000, main="AUC Curve", lty=2, col="red")

- **glmnet (x, y, family=”cox”, alpha =0, lambda =grid)**
  Αυτή η συνάρτηση προσαρμόζει το μοντέλο που θέλουμε με τις μεθόδους συρρίκνωσης Ridge και Lasso. Στην εφαρμογή μας θέλαμε να προσαρμόσουμε το μοντέλο του Cox για αυτό στο δείκτη family βάλαμε “cox”, θα μπορούσαμε να βάλουμε όποιο άλλο μοντέλο θέλαμε π.χ. γραμμικό. Στους δείκτες x και y βάζουμε πίνακες, στον x τις μεταβλητές μας και στον y βάζουμε τη μεταβλητή απόκρισης ανάλογα με το μοντέλο που θέλουμε να προσαρμόσουμε, εμείς θέσαμε x <- matrix (c(groupF, recipient.age, donor.age, recipient.sex, donor.sex, recipient.cmv, donor.cmv, waiting.time, fab, mtx), ncol=10) και y <- matrix (Surv (data$time,data$indicator), ncol=2). Ο δείκτης alpha λαμβάνει τις τιμές 0 ή 1 ανάλογα με τη μέθοδο που θέλουμε να εφαρμόσουμε. Για τη μέθοδο Κορυφογραμμής Ridge βάζουμε 0 και για τη Lasso 1. Τέλος ο δείκτης lambda ορίζει την παράμετρο λ και λαμβάνει ένα μεγάλο εύρος τιμών που καθορίζουμε εμείς.

- **set.seed (1)**
  train=sample (1: nrow(x), nrow(x)/2)
  test=(- train)
  y.test=y[test]
  Με τις παραπάνω εντολές χωρίζουμε τα δεδομένα μας με τυχαίο τρόπο σε δύο δείγματα (train και test) με σκοπό την καλύτερη εφαρμογή της μεθόδου Κορυφογραμμής (Ridge) και της Lasso.

- **cv.out = cv.glmnet (x, y, family="cox", alpha =0)**
  Αυτή η συνάρτηση εφαρμόζει τη μέθοδο Cross-Validation για την εύρεση της βέλτιστης τιμής της παραμέτρου λ της μεθόδου Κορυφογραμμής Ridge (εφ’ όσον έχουμε alpha=0) για την οικογένεια του μοντέλου Cox. Αν επιλέγαμε alpha=1, η συνάρτηση θα εφάρμοζε τη μέθοδο Cross-Validation για την εύρεση της βέλτιστης τιμής της παραμέτρου λ της μεθόδου Lasso.

- **bestlam = cv.out$lambda.min**
  Με αυτή την εντολή βρίσκουμε την μικρότερη που λαμβάνει η συνάρτηση cv.glmnet, δηλαδή τη μικρότερη τιμή της παραμέτρου λ.

- **predict (out, family="cox", type= "coefficients", s= bestlam)**
  Η συνάρτηση predict είναι μία γενική συνάρτηση η οποία υπολογίζει προβλέψεις και εκτιμάει τα τυπικά σφάλματα αυτών των προβλέψεων για όποιο μοντέλο θέλουμε. Στη συγκεκριμένη περίπτωση θέλαμε να υπολογίσουμε τους συντελεστές του μοντέλου με την ονομασία “out” για αυτό θέσαμε type= "coefficients". Επίσης για να υπολογίσουμε τους συντελεστές του καλύτερου μοντέλου θέσαμε το λ να είναι ίσο με το bestlam, όπου είναι η μικρότερη τιμή που μπορεί να πάρει η παράμετρος λ.