ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΠΜΣ ΔΙΟΙΚΗΣΗ, ΑΝΑΛΥΤΙΚΗ ΚΑΙ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΕΠΙΧΕΙΡΗΣΕΩΝ

(M.Sc. in Business Administration, Analytics and Information Systems)

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΤΗΣ ΔΙΑΡΚΕΙΑΣ ΤΟΥ ΚΥΚΛΟΥ ΖΩΗΣ ΤΩΝ ΕΡΓΑΖΟΜΕΝΩΝ ΜΕ ΜΕΘΟΔΟΥΣ ΑΝΑΛΥΣΗΣ ΕΠΙΒΙΩΣΗΣ

ΑΘΑΝΑΣΙΟΣ ΧΡΟΝΟΠΟΥΛΟΣ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ,

ΕΥΑΓΓΕΛΟΣ ΒΑΣΙΛΕΙΟΥ

ΑΘΗΝΑ

2021

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΠΜΣ ΔΙΟΙΚΗΣΗ, ΑΝΑΛΥΤΙΚΗ ΚΑΙ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΕΠΙΧΕΙΡΗΣΕΩΝ

(M.Sc. in Business Administration, Analytics and Information Systems)

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΤΗΣ ΔΙΑΡΚΕΙΑΣ ΤΟΥ ΚΥΚΛΟΥ ΖΩΗΣ ΤΩΝ ΕΡΓΑΖΟΜΕΝΩΝ ΜΕ ΜΕΘΟΔΟΥΣ ΑΝΑΛΥΣΗΣ ΕΠΙΒΙΩΣΗΣ

ΑΘΑΝΑΣΙΟΣ ΧΡΟΝΟΠΟΥΛΟΣ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ,

ΕΥΑΓΓΕΛΟΣ ΒΑΣΙΛΕΙΟΥ

ΑΘΗΝΑ

2021

Περιεχόμενα

ΕΙΚΟΝΕΣ

ΠΙΝΑΚΕΣ

ΔΙΑΓΡΑΜΜΑΤΑ

Περίληψη

ΕΛΛΗΝΙΚΗ

Με την ανάπτυξη της προκείμενης διπλωματικής σκοπεύουμε να μοντελοποίησουμε την διάρκεια του κύκλου ζωής των εργαζομένων, που απασχολούνται εκ μέρους κάποιας εταιρίας, χρησιμοποιώντας μεθόδους ανάλυσης επιβίωσης. Αρχικά, με την χρήση της μεθόδου '*Kaplan-Meier'*, θα αναπαραστήσουμε γραφικά την καμπύλη επιβίωσης των εργαζομένων του συνόλου δεδομένων που διαθέτουμε. Έχοντας στην διάθεση μας την συνολική καμπύλη επιβίωσης, μπορούμε να διακρίνουμε την πιθανότητα επιβίωσης (παραμονή του εργαζομένου στην εταιρία) για όλους τους εργαζόμενους του συνόλου δεδομένων καθώς και εκείνους που αποκόπηκαν από την μελέτη. Στην συνέχεια, θα προχωρήσουμε στο επόμενο στάδιο της ανάλυσης μας, κατασκευάζοντας το μοντέλο αναλογικών κινδύνων της '*Cox*' παλινδρόμησης στο σύνολο των εργαζομένων και διακρίνοντας όλες εκείνες τις επεξηγηματικές μεταβλητές που ασκούν μεγαλύτερη επίδραση στην αποχώρηση ενός εργαζομένου από την εταιρία. Αντιστοίχως, θα μπορούμε να παρατηρήσουμε και εκείνες τις μεταβλητές που φαίνεται να μην επηρεάζουν σημαντικά το αποτέλεσμα όσον αφορά την αποχώρηση του εργαζομένου από την εταιρία, και οι οποίες θα χρειαστεί να απομακρυνθούν από το μοντέλο. Τέλος, με την δημιουργία δέντρων επιβίωσης, θα προσπαθήσουμε να δημιουργήσουμε ένα πιο περίπλοκο και ανεπτυγμένο μοντέλο έχοντας την ευκαιρία να διακρίνουμε την κατηγορία εργαζομένων που λαμβάνουν την μεγαλύτερη πιθανότητα, με σκοπό την συνέχιση των καθηκόντων τους στην εταιρία. Ενώ επίσης, και εκείνους που ο κίνδυνός να αποχωρήσουν από την εταιρία είναι περισσότερο αυξημένος.

Λέξεις-κλειδιά: Kaplan-Meier, cox παλινδρόμηση, αναλογικοί κίνδυνοι, δέντρα επιβίωσης

ΑΓΓΛΙΚΗ

MODELING AND DURATION OF THE LIFE CYCLE OF WORKERS WITH SURVIVAL ANALYSIS METHODS

By the development of master's thesis, we intend to model the life cycle of employees who worked in a company, using survival analysis methods. Initially, using the '*Kaplan-Meier'* method, we will graphically represent the survival curve of employees from the dataset provided to us. Having in our hands the overall survival curve, we can distinguish the survival probability (employee stay in the company) for all employees of the data set as well as, we can observe those who censored from the study. Next, we will move on to the next stage of our analysis, constructing the 'Cox proportional hazards model' for all employees and distinguishing all those explanatory variables that have the greatest impact on an employee leaving the company. Respectively, we will be able to observe all those variables that do not seem significantly affect the result of employee turn from the company, and maybe will need to be removed from the model. Finally, by creating survival trees, we will try to create a more complex and robust model by having the opportunity to distinguish the category of employees who receive the highest survival probability, in order to continue their duties in the company. Although, we can see the employees whose risk leaving the company is more increased.

Λέξεις-κλειδιά: Kaplan-Meier, Cox, proportional hazards, survival trees

1. Εισαγωγή

Το τμήμα ανθρωπίνου δυναμικού κάθε εταιρίας ξοδεύει μεγάλους όγκους χρηματικών ποσών προκειμένου να αντιμετωπίσει το πρόβλημα της εγκατάλειψης της θέσεως από κάποιον εργαζόμενο. Στην περίπτωση ειδικά που πρόκειται για πολύτιμους εργαζόμενους το κόστος εγκατάλειψης τους, είναι ακόμα μεγαλύτερο για την εταιρία. Κάποιος εργαζόμενος είτε αποχώρησε από την εταιρία, είτε αποκόπηκε από την μελέτη μία συγκεκριμένη χρονική στιγμή. Η περίπτωση αποκοπής εργαζομένου μπορεί να συμβεί για δύο λόγους. Πρώτον, μπορεί η παρακολούθηση για τον εργαζόμενο να σταμάτησε την συγκεκριμένη χρονική περίοδο για κάποιον λόγο που δεν μπορούμε να γνωρίζουμε. Δεύτερον, έχει επέλθει η λήξη της μελέτης και ο εργαζόμενος δεν έχει αποχωρήσει από την εταιρία, με αποτέλεσμα να αποκοπεί στην λήξη της μελέτης. Θα μπορούσαμε να πούμε πως ο δεύτερος λόγος αποκοπής αποτελεί και το πιο συχνό φαινόμενο. Σκοπός της παρούσας διπλωματικής είναι η μοντελοποίηση της διάρκειας του κύκλου ζωής των εργαζομένων που απασχολούνται από κάποια εταιρία. Βασιζόμενοι στα δεδομένα που έχουμε συλλέξει, αποτέλεσμα της παρακάτω ανάλυσης είναι να καταφέρουμε να προβλέψουμε την χρονική περίοδο που κάποιος εργαζόμενος θα παραμείνει στην εταιρία, αλλά και εκείνους τους παράγοντες που επηρεάζουν περισσότερο την αποχώρηση του. Η μοντελοποίηση του κύκλου ζωής του εργαζομένου με την χρήση των όχι και τόσο διαδεδομένων αλγορίθμων ανάλυσης επιβίωσης για την μελέτη που επιθυμούμε να κάνουμε αποτελεί μία καινοτόμο μέθοδο. Θα μπορούσαμε να πούμε πως υπερτερεί από την χρήση άλλων μεθόδων που θα μπορούσαν να χρησιμοποιηθούν, παραδείγματος χάριν της λογιστικής-γραμμικής παλινδρόμησης, εξαιτίας του γεγονότος ότι με την χρήση της ανάλυσης επιβίωσης έχουμε την δυνατότητα να προβούμε σε μία πρόβλεψη παραμονής του εργαζομένου ύστερα από μακροχρόνια εμπειρία του στην εταιρία ξεχωριστά για τον καθέναν και να εκτιμήσουμε την χρονική στιγμή που εγκατάλειψε την εταιρία. Αντιθέτως με την χρήση της λογιστικής παλινδρόμησης μας δίνεται η δυνατότητα να κάνουμε προβλέψεις που αφορούν εργαζόμενους με σύντομη χρονική διάρκεια απασχόλησης σχετικά με την παραμονή τους χωρίς όμως να γνωρίζουμε ακριβώς την χρονική στιγμή που εγκατέλειψαν την εταιρία. Η χρήση της απλής γραμμικής παλινδρόμησης μας περιορίζει ακόμα περισσότερο καθώς για τις προβλέψεις της χρονικής στιγμής που ο εργαζόμενος εγκατέλειψε την εταιρία μπορεί να λάβουμε τιμές μικρότερες του μηδενός (<0), κάτι το οποίο θα ήταν αδύνατον να συμβεί αφού εξετάζουμε τον χρόνο. Επίσης, εξαιτίας του γεγονότος ότι για κάποιον εργαζόμενο μπορεί να μην εκδηλωθεί το συμβάν, δίχως να γνωρίζουμε την ακριβή εκδήλωση του ενδιαφέροντος από όλες τις παρατηρήσεις, αλλά

και πώς ο παράγοντας του κινδύνου μπορεί να επηρεάσει τον χρόνο εμφάνισης του γεγονότος, στην προκειμένη περίπτωση να αποχωρήσει ο εργαζόμενος από την εταιρία, αποτελούν σημαντικές παραμέτρους που αποτρέπουν την χρησιμοποίηση μοντέλων απλής γραμμικής παλινδρόμησης και την αναζήτηση άλλων μεθόδων για την διεκπεραίωση της ανάλυσης μας. Στο πρώτο κεφάλαιο περιγράφονται όλες οι τεχνικές και τα μοντέλα ανάλυσης επιβίωσης που θα χρησιμοποιήσουμε προκειμένου να διεκπεραιωθεί η περαιτέρω ανάλυση. Περιγράφεται το θεωρητικό υπόβαθρο της ανάλυσης επιβίωσης που θα χρησιμοποιηθεί στο σύνολο δεδομένων. Στο δεύτερο κεφάλαιο θα ακολουθήσει εκτενής περιγραφή του συνόλου δεδομένων και οτιδήποτε πληροφορία θα μπορούσε να εξαχθεί από τα δεδομένα πριν προχωρήσουμε στην ανάλυση επιβίωσης. Στο τελευταίο κεφάλαιο θα υπάρξει πλήρης παρουσίαση και ερμηνεία των αποτελεσμάτων που εξάχθηκαν από την χρήση του κάθε μοντέλου στην ανάλυση επιβίωσης. Επίσης σε τελικό στάδιο θα πραγματοποιηθεί σύγκριση μεταξύ των μοντέλων, προκειμένου να διακρίνουμε την αποδοτικότητα του κάθε μοντέλου στο συγκεκριμένο σύνολο δεδομένων. Το σύνολο δεδομένων που θα χρησιμοποιήσουμε για την ανάλυση μας αποτελείται από πραγματικά δεδομένα εργαζομένων που απασχολούνταν εκ μέρους κάποιας εταιρίας και προέρχεται από το ιστολόγιο του *'Edward Babushkin'*.

2. Επισκόπηση ανάλυσης επιβίωσης

2.1. Εισαγωγή στην ανάλυση επιβίωσης

Η ανάλυση επιβίωσης (*survival analysis*) αποτελεί πεδίο του κλάδου στατιστικής και όπως αναφέρει ο (Lewinshon, 2020), εστιάζει στην χρονική περίοδο (*survival time*) κάποιας παρατήρησης του συνόλου δεδομένων που μελετάμε έως ότου πραγματοποιηθεί ένα συγκεκριμένο γεγονός. Αρχικά η ανάλυση επιβίωσης αναπτύχθηκε με σκοπό να εκτιμηθούν οι επιδράσεις μίας θεραπείας ή ενός φαρμάκου σε σχέση με ένα εικονικό φάρμακο στην διάρκεια ζωής των ασθενών που πάσχουν από μία συγκεκριμένη νόσο. Παρ ’όλο που ο σκοπός δημιουργίας της ανάλυσης επιβίωσης χρησιμοποιήθηκε κυρίως για φαρμακευτικούς σκοπούς, δεν άργησε να υιοθετηθεί και από τους υπόλοιπους επιστημονικούς κλάδους. Ως γεγονότα ενδιαφέροντος (*event of interest*) θα μπορούσαν κάλλιστα να δηλωθούν ο θάνατος του ασθενούς όταν αναφερόμαστε στον ιατρικό κλάδο, ο χρόνος λειτουργίας του κινητήρα ενός αυτοκινήτου, η στιγμή που κάποιος εργαζόμενος αποφασίζει να εγκαταλείψει μία εταιρία και πολλές άλλες ποικίλες περιπτώσεις οι οποίες διαφέρουν ανάλογα με το αντικείμενο που μελετάμε. Σύμφωνα μ με τον (Lewinshon, 2020) , το γεγονός ως επί το πλείστων μπορεί να πραγματοποιηθεί μία φορά για κάθε παρατήρηση στο δείγμα (π.χ. ο θάνατος ενός ασθενούς) και η μελέτη θα σταματήσει έπειτα από την εκδήλωση του ενδιαφέροντος για την συγκεκριμένη παρατήρηση.

2.1.1. Ορισμός του χρόνου επιβίωσης

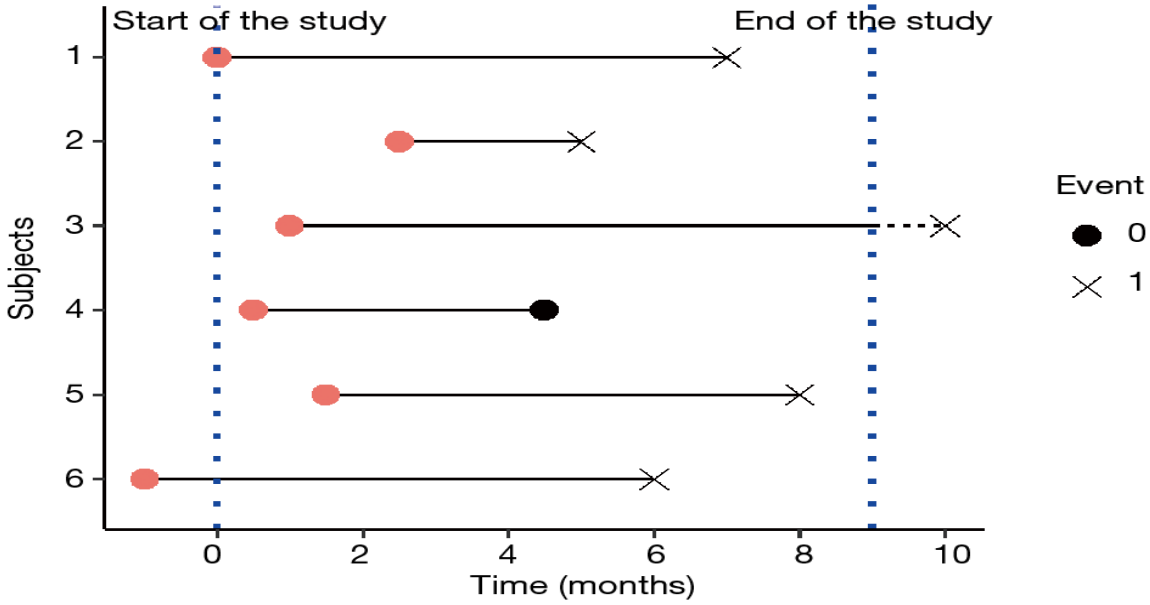
Ως χρόνος επιβίωσης *(survival time)*, όπως ορίζεται (Kleinbaum & Klein, 2005), μίας μη-αποκομμένης παρατήρησης ορίζεται ως η διάρκεια μεταξύ της χρονικής στιγμής που εισέρχεται στην μελέτη μας μέχρι την στιγμή που θα εκδηλώσει το γεγονός του ενδιαφέροντος. Παραδείγματος χάριν, όταν μελετάμε την διάρκεια του κύκλου εργασίας ενός εργαζομένου ορίζεται ως χρόνος επιβίωσης (*survival time*), ο οποίος θα μπορούσε να μετρηθεί σε χρόνια, μέρες ή βδομάδες. Ως χρόνος επιβίωσης ενός εργαζομένου υποδηλώνεται, το χρονικό διάστημα μεταξύ της περιόδου που εισέρχεται ο εργαζόμενος στην μελέτη, καθώς η πρόσληψη του μπορεί να έχει γίνει πριν την έναρξη όπως επίσης και κατά την διάρκεια της μελέτης, μέχρι την χρονική στιγμή που θα εγκαταλείψει την εταιρία.

2.1.2. Λογοκριμένα ή αποκομμένα δεδομένα

Αποκοπή *(censoring)* στην ανάλυση επιβίωσης από τους (Prinja, Nidhi , & Ramesh, 2010), ορίζεται ως η κατάσταση κατά την οποία έχουμε κάποια πληροφορία σχετικά με τον χρόνο επιβίωσης *(survival time)* των παρατηρήσεων, δίχως όμως να γνωρίζουμε τον χρόνο επιβίωσης τους με απόλυτη ακρίβεια. Εάν για παράδειγμα κάποιος ασθενής εγκαταλείψει την έρευνα πριν το τέλος της μελέτης για κάποιον λόγο, διαφορετικό από το γεγονός του ενδιαφέροντος που εξετάζουμε ή

με την περάτωση της μελέτης ο ασθενής δεν έχει εκδηλώσει το γεγονός(π.χ. θάνατος), τότε ο χρόνος επιβίωσης του χαρακτηρίζεται από αποκοπή (*censored survival time*). Σύμφωνα με τους (Kleinbaum & Klein, 2005), γενικότερα υπάρχουν τρεις λόγοι που μπορεί μία παρατήρηση να χαρακτηριστεί ως αποκομμένη σύμφωνα με τους:

- Έχει τελειώσει η μελέτη και το άτομο δεν έχει εκδηλώσει το γεγονός του ενδιαφέροντος.
- Το άτομο δεν κατάφερε να ακολουθήσει την μελέτη και χάθηκε κατά την διάρκεια της διαδρομής.
- Το άτομο απετράπη από την μελέτη λόγω της εκδήλωσης κάποιου διαφορετικού γεγονότος (π.χ. αν πεθάνει ενώ ο θάνατος δεν αποτελεί το γεγονός ενδιαφέροντος) ή για κάποιον άλλον λόγο που δεν μπορούμε να γνωρίζουμε.



Εικόνα 1- Παραδείγματα εκδήλωσης του γεγονότος αλλά και αποκοπής για 6 παρατηρήσεις

Όλοι οι παραπάνω λόγοι αποτελούν μέρος δεξιάς αποκοπής *(Right censoring)* η οποία θα μπορούσαμε να πούμε πως αποτελεί την πιο συχνή εμφάνιση αποκοπής στην ανάλυση επιβίωσης (BIOST 515, 2004). Γενικά υπάρχουν τρία είδη αποκοπής:

a) Δεξιά αποκοπή (*Right censoring*)

   Αποτελεί την πιο συχνή μορφή αποκοπής που μπορούμε να συναντήσουμε στην ανάλυση επιβίωσης. Πρόκειται για τον τύπο αποκοπής όπου δεν γνωρίζουμε με ακρίβεια τον χρόνο επιβίωσης *(survival time)* του ατόμου και συναντάται στην δεξιά πλευρά της περιόδου που μελετάμε. Το γεγονός του ενδιαφέροντος μπορεί να συμβεί μετά την λήξη της μελέτης (*end-of-study censoring*), όταν το άτομο έχει αποσυρθεί από την μελέτη

(*loss-to-follow-up censoring*) καθώς και στην περίπτωση που του έχει συμβεί διαφορετικό γεγονός από αυτό που εξετάζουμε.

- Αποκοπή τύπου I (*censored type I data)*

  Συμβαίνει με την προϋπόθεση ότι η μελέτη έχει προκαθοριστεί να λήξει μετά από *T* χρόνια. Σε αυτή την περίπτωση κάθε παρατήρηση η οποία δεν έχει εκδηλώσει το γεγονός του ενδιαφέροντος καθ' όλη την πορεία της μελέτης μέχρι και την λήξη της, λέμε πως είναι αποκομμένη στα *T* χρόνια (BIOST 515, 2004).

- Αποκοπή τύπου 2 (*censored type II data*)

  Η μελέτη διακόπτεται όταν υπάρχει ένας προκαθορισμένος αριθμός παρατηρήσεων που έχουν εκδηλώσει το γεγονός ενδιαφέροντος. Έχουμε την δυνατότητα να γνωρίζουμε τον ακριβή χρόνο εκδήλωσης του ενδιαφέροντος για τις παρατηρήσεις που συνέβη το γεγονός. Από το άρθρο (Emgineering Statistics, 2013), η αποκοπή τύπου 2 διαθέτει το σημαντικό πλεονέκτημα ότι μπορούμε να γνωρίζουμε εκ των προτέρων πόσες παρατηρήσεις που 'απέτυχαν' θα μας επιφέρει το τεστ.

- Τυχαία Αποκοπή τύπου I (*random censoring*)

  Η μελέτη προβλέπεται να λήξει μετά από *T* χρόνια, αλλά η αποκοπή δεν συνέβη την ίδια χρονική στιγμή για όλες τις λογοκριμένες παρατηρήσεις. Για παράδειγμα υπάρχουν αποκομμένες παρατηρήσεις πριν αλλά και μετά την λήξη της μελέτης. Αυτόν την μορφή δεξιάς αποκοπής θα συναντήσουμε τις περισσότερες φορές (BIOST 515, 2004).

b) Αριστερή αποκοπή(*Left censoring*)

Ορίζεται ως η μορφή αποκοπής που συμβαίνει στην αριστερή πλευρά της μελέτης ,όπου γνωρίζουμε ότι μία παρατήρηση έχει 'αποτύχει' αλλά δεν γνωρίζουμε ακριβώς το πότε συνέβη το γεγονός (Minitab Blog Editor, 2016). Όπως επίσης η περίπτωση που η παρατήρηση είχε ξεκινήσει να εκδηλώνει το γεγονός πριν την έναρξη της μελέτης. Στις αριστερά αποκομμένες παρατηρήσεις το γεγονός συμβαίνει πριν τον καθορισμένο χρόνο.

c) Αποκοπή σε διάστημα (*Interval-censored data*)

Όπως αναφέρεται από τους (Leung, Elashoff, & Afifi, 1997), σε αυτή την περίπτωση αποκοπής γνωρίζουμε ότι το γεγονός συνέβη μεταξύ κάποιας χρονικής περιόδου. Μερικές φορές ο ακριβής χρόνος εκδήλωσης του ενδιαφέροντος μιας παρατήρησης δεν είναι γνωστός παρά μόνο ένα διάστημα όπου η παρατήρηση απέτυχε μπορεί να υπολογιστεί.

Στην ανάλυση που θα κάνουμε θεωρούμε πως η αποκοπή είναι ανεξάρτητη και δεν μπορεί να μας παρέχει πληροφορίες σχετικά με το γεγονός του ενδιαφέροντος (*event*). Αυτό οφείλεται στο ότι για τις λογοκριμένες παρατηρήσεις συνέβη κάποιο άλλο γεγονός διαφορετικό από αυτό που εξετάζουμε στην μελέτη (BIOST 515, 2004).

2.1.3. Συνάρτηση Επιβίωσης (survival function)

Η καμπύλη επιβίωσης *S(t)* παρουσιάζει την επιβίωση των ατόμων που μελετάμε σε συνάρτηση με τον χρόνο *t*. Η συγκεκριμένη καμπύλη επιβίωσης όπως ορίζεται από τους (Kleinbaum & Klein, 2005), αντικατοπτρίζει μία συνάρτηση επιβίωσης (*survival function*) και παρέχει την πιθανότητα για κάποιο άτομο από το δείγμα μας να έχει επιβιώσει περισσότερο από έναν προκαθορισμένο χρόνο *t*. Δηλαδή εκφράζει την πιθανότητα κατά την οποία η τυχαία μεταβλητή *T* ξεπερνάει μία προκαθορισμένη τιμή χρόνου *t*. Αξίζει να σημειωθεί πως η τυχαία μεταβλητή *T* εκφράζει την διάρκεια ζωής των παρατηρήσεων στο δείγμα μας, χωρίς αυτό απαραίτητα να σημαίνει πως αποκλειστικά δηλώνει τον χρόνο, αλλά θα μπορούσε να δηλώνει την διάρκεια λειτουργίας μίας μονάδας. Αυτό εξαρτάται από την περιγραφή του προβλήματος που εξετάζουμε.

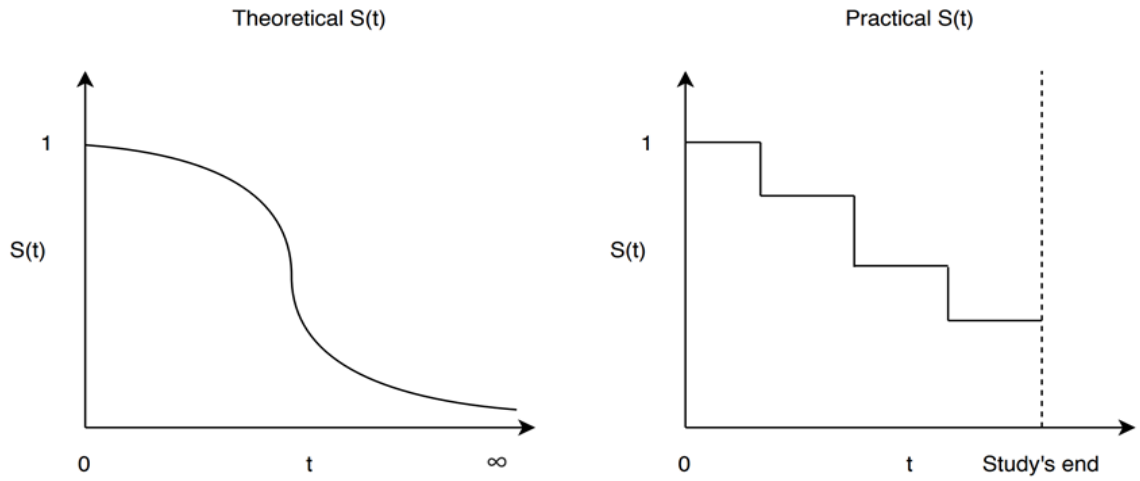
$$S(t) = Pr(T > t) = 1 - F(t) \quad (2.1)$$

Η συνάρτηση επιβίωσης αποτελεί θεμελιώδες στοιχείο στην ανάλυση επιβίωσης για τον λόγο ότι η απόκτηση πιθανοτήτων επιβίωσης στα δεδομένα για προκαθορισμένες τιμές του χρόνου *t*, μας παρέχει χρήσιμες πληροφορίες σχετικά με τα δεδομένα στα οποία αντικείμενο μελέτης είναι η επιβίωση (Kleinbaum & Klein, 2005).

Γενικά μιλώντας:

- Καθώς ο χρόνος *t* κινείται από την έναρξη της μελέτης μας (*t = 0*) προς το άπειρο *(t = ∞)* τόσο περισσότερο η καμπύλη θα φθίνει καθώς οι πιθανότητες επιβίωσης των παρατηρήσεων θα μικραίνουν. Πρόκειται για μία φθίνουσα καμπύλη στο πέρασμα του χρόνου.
- Την χρονική στιγμή *t = 0* η πιθανότητα επιβίωσης *S(t) = S(0)* = 1. Κατά την έναρξη της μελέτης μας, η πιθανότητα επιβίωσης για όλες τις παρατηρήσεις μετά τον χρόνο *t = 0* ισούται με 1, καθώς δεν έχει προλάβει να εκδηλωθεί το γεγονός σε καμία παρατήρηση.
- Την χρονική στιγμή *t* = ∞ η πιθανότητα επιβίωσης *S(t) = S(∞)* = 0. Μετά από το πέρας μεγάλου χρονικού διαστήματος όπου έχει εκδηλωθεί το γεγονός του ενδιαφέροντος (π.χ. θάνατος) από όλες σχεδόν τις παρατηρήσεις, η πιθανότητα επιβίωσης μπορούμε να πούμε ότι αγγίζει το 0.

Σύμφωνα με τους (Kleinbaum & Klein, 2005) , θεωρητικά η καμπύλη επιβίωσης είναι ομαλή, αλλά όταν ασχολούμαστε με πραγματικά δεδομένα έχει την μορφή βηματισμού (*step function*).



Εικόνα 2 -Σύγκριση της πρακτικής καμπύλης επιβίωσης με την θεωρητική καμπύλη επιβίωσης

2.1.4. Συνάρτηση Κινδύνου (hazard function)

Η συνάρτηση κινδύνου *h(t)* εκφράζει τον στιγμιαίο ρυθμό να εκδηλώσει κάποια παρατήρηση το γεγονός του ενδιαφέροντος μία συγκεκριμένη χρονική στιγμή *t*, δεδομένου ότι έχει επιβιώσει μέχρι εκείνη την χρονική στιγμή *t*.

$$h(t) = \lim_{dt \to 0} \frac{Pr\{t \leq T < t + dt | T \geq t\}}{dt} \qquad (2.2)$$

Σύμφωνα με τον (Lewinshon, 2020), ο τύπος της παραπάνω συνάρτησης κινδύνου παρατηρούμε ότι πρόκειται για την δεσμευμένη πιθανότητα (*conditional failure rate*) το γεγονός του ενδιαφέροντος να συμβεί κάποια στιγμή μεταξύ του χρονικού διαστήματος [*t, t+dt*), δεδομένου ότι το γεγονός δεν έχει εκδηλωθεί πριν από την χρονική στιγμή *t*. Ο παρονομαστής δηλώνει το πλάτος του διαστήματος.

Σε αντίθεση με την συνάρτηση επιβίωσης *S(t)* η οποία επικεντρώνεται ότι το άτομο έχει επιβιώσει, δηλαδή δεν έχει συμβεί το γεγονός μέχρι την στιγμή *t*, η συνάρτηση κινδύνου επικεντρώνεται στην στιγμιαία εκδήλωση του γεγονότος την συγκεκριμένη χρονική στιγμή. Επίσης ο κίνδυνος εκφράζει ρυθμό (*rate*) και όχι πιθανότητα για αυτό τον λόγο οι τιμές που λαμβάνει η συνάρτηση του κινδύνου κυμαίνονται μεταξύ του διαστήματος (*0, ∞*) και εξαρτώνται από την μονάδα μέτρησης του χρόνου, παραδείγματος χάριν μέρες, μήνες, χρόνια. Σύμφωνα με

τους (Kleinbaum & Klein, 2005) , η συνάρτηση του κινδύνου μας δίνει την αντίθετη ακριβώς πληροφορία που εξάγεται από την συνάρτηση της επιβίωσης, γι' αυτό τον λόγο προκύπτει κάποια σχέση μεταξύ των δύο συναρτήσεων.

Από τους (Imani, Chen, Tucker, & Yang, 2019) ,η αθροιστική συνάρτηση κινδύνου δηλώνει τον συσσωρευτικό κίνδυνο μέχρι εκείνη την χρονική στιγμή *t*. Απλούστερα, το άθροισμα των κινδύνων από την αρχή της μελέτης *t=0* μέχρι την χρονική στιγμή *t*.

$$H(t) = \int_0^t h(u)du \quad (2.3)$$

Σε περίπτωση που γνωρίζουμε κάποια από τις παραπάνω συναρτήσεις *S(t), H(t), h(t),* μπορούμε εύκολα να βρούμε και τις υπολειπόμενες δύο συναρτήσεις (BIOST 515, 2004).

$$h(t) = -\frac{\partial \log\big(S(t)\big)}{\partial t} \quad (2.4)$$

$$H(t) = -\log\big(S(t)\big) \quad (2.5)$$

$$S(t) = e^{(-H(t))} \quad (2.6)$$

2.1.5. Παραμετρικά μοντέλα (parametric models)

Κατά κύριο λόγο στην ανάλυση επιβίωσης χρησιμοποιούμε μη-παραμετρικά ή ημι-παραμετρικά μοντέλα. Τα παραμετρικά μοντέλα είναι αρκετά αξιόπιστα και η χρήση τους είναι ευρέως διαδεδομένη σε πολλά επιστημονικά πεδία μεταξύ των οποίων ανήκει η ανάλυση δεδομένων, η μηχανολογική στατιστική (engineer statistic) και πολλοί άλλοι επιστημονικοί κλάδοι.

2.1.5.1. Εκθετική κατανομή

Ένα σημαντικό γνώρισμα της εκθετικής κατανομής, σύμφωνα με τον (Wienke, 2007), είναι η ιδιότητα της να μην απομνημονεύει (*memoryless property*). Αυτό θα μπορούσε να εξηγηθεί από το γεγονός ότι στο πέρασμα του χρόνου η παρατήρηση έχει τον ίδιο κίνδυνο να εκδηλώσει το γεγονός του ενδιαφέροντος με τον κίνδυνο που είχε και στην αρχή της μελέτης (*t = 0*) (Guo, 2011).

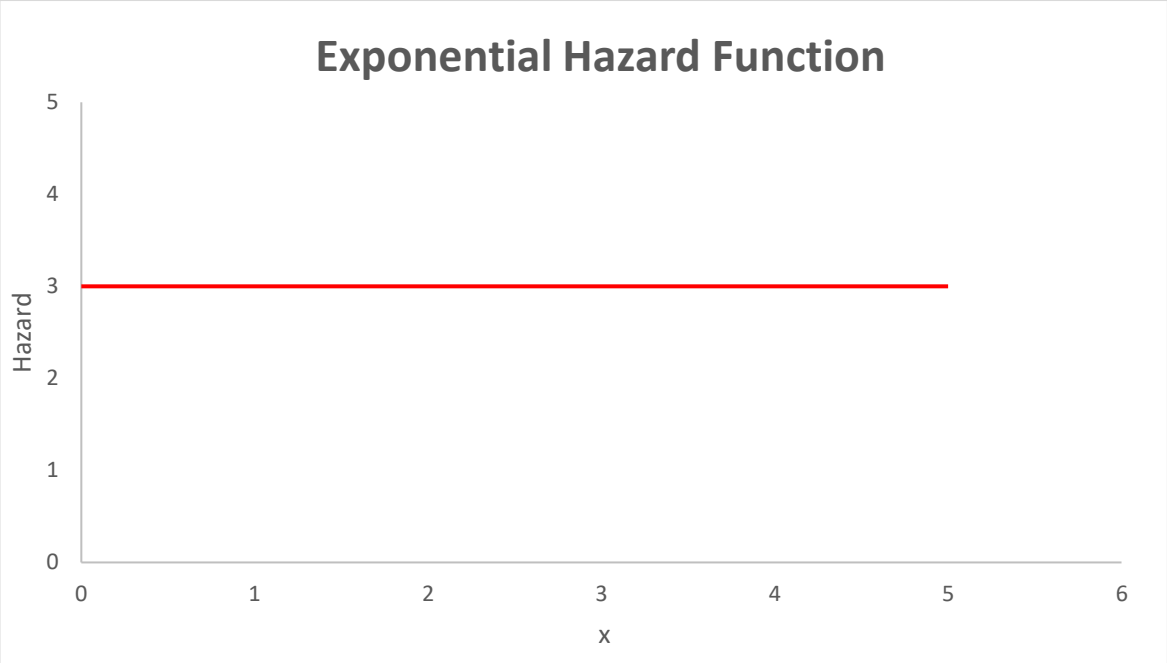
- Η συνάρτηση επιβίωσης της εκθετικής κατανομής είναι της μορφής:

$$S(x) = e^{(-\lambda x)}, \lambda > 0, x \geq 0 \quad (2.7)$$

- Η συνάρτηση κινδύνου της εκθετικής κατανομής είναι της μορφής:

$$h(x) = \lambda \quad (2.8)$$

Από το παρακάτω διάγραμμα της συνάρτησης κινδύνου εκθετικής κατανομής παρατηρούμε ότι κίνδυνος για να συμβεί το γεγονός του ενδιαφέροντος παραμένει σταθερός καθώς κυλάει ο χρόνος.



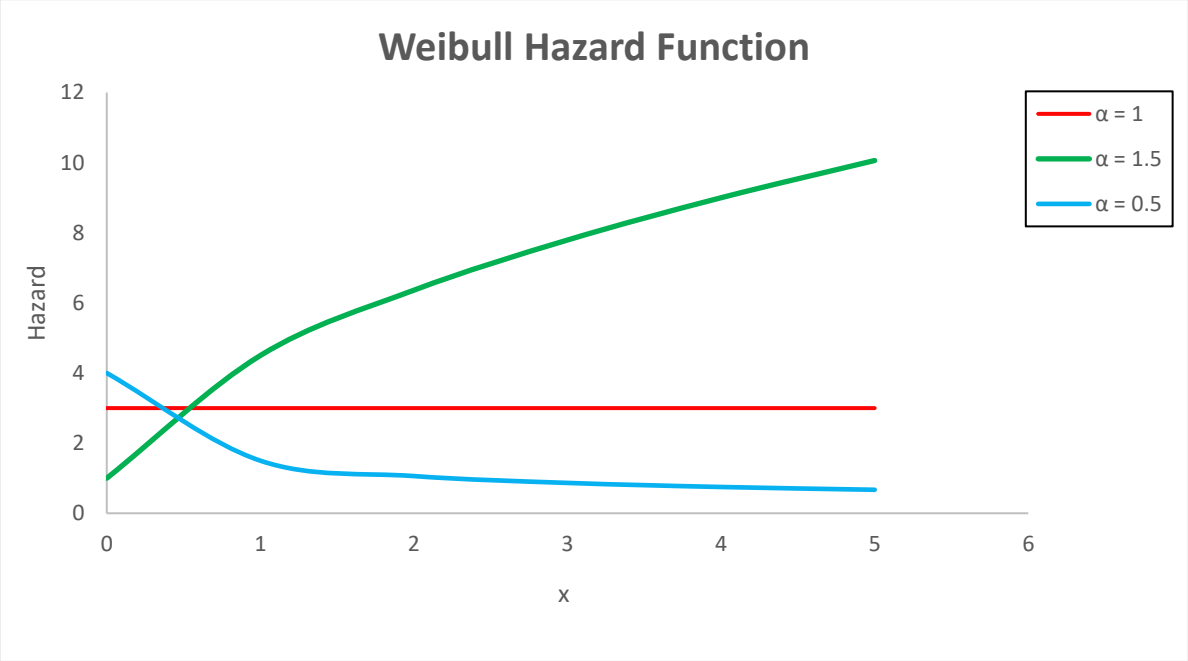
Διάγραμμα 1 - Εκθετική συνάρτηση κινδύνου

2.1.5.2. Κατανομή 'Weibull'

Η κατανομή *Weibull* αποτελεί γενίκευση της απλής εκθετικής κατανομής (Wienke, 2007). Είναι ευρέως γνωστή για την αξιοπιστία της με πολύ μεγάλη χρησιμότητα στην επιστήμη της μηχανολογίας. Όπως μας επεξηγεί ο (Guo, 2011) , οι αυξανόμενοι κίνδυνοι (*increasing hazards*), οι μειωμένοι κίνδυνοι (*decreasing hazards*) αλλά και οι σταθεροί κίνδυνοι (*constant hazards*) που μπορεί να μας προσφέρει η συγκεκριμένη κατανομή αποτελούν παραμέτρους που την καθιστούν πάρα πολύ ευέλικτη και εύχρηστη σε μία πληθώρα καταστάσεων:

- Ένας μειωμένος ρυθμός κινδύνου (*decreasing hazards*) δηλώνει ότι οι παρατηρήσεις μας στο δείγμα 'αποτυγχάνουν' (εκδηλώνουν το γεγονός του ενδιαφέροντος) σε ένα πρώιμο στάδιο. Ενώ στην αρχή της μελέτης ο κίνδυνος είναι αυξημένος, ο ρυθμός κινδύνου μειώνεται καθώς περνάει ο χρόνος. Η μπλε γραμμή του γραφήματος παρακάτω απεικονίζει αυτή την περίπτωση.

- Ο σταθερός ρυθμός κινδύνου (*constant hazards*) δηλώνει ότι οι παρατηρήσεις 'αποτυγχάνουν' από τυχαία συμβάντα. Βλέπουμε πως η *Weibull* κατανομή αποτελεί γενίκευση της εκθετικής κατανομής, καθώς στην συγκεκριμένη περίπτωση εξάγεται ακριβώς το ίδιο αποτέλεσμα με αυτό της εκθετικής. Η κόκκινη γραμμή του γραφήματος δηλώνει αυτή την περίπτωση.
- Ένας αυξανόμενος ρυθμός κινδύνου (*increasing hazards*) δηλώνει ότι οι παρατηρήσεις 'αποτυγχάνουν' περισσότερο προς το τέλος της μελέτης. Ο ρυθμός κινδύνου είναι χαμηλός στην αρχή της μελέτης και αυξάνεται με το πέρασμα του χρόνου. Η πράσινη γραμμή του γραφήματος αντικατοπτρίζει αυτή την περίπτωση.



Διάγραμμα 2 - Weibull συνάρτηση κινδύνου

Η συνάρτηση επιβίωσης της κατανομής *Weibull* είναι της μορφής:

$$S(x) = e^{(-\lambda x^{a})} \quad (2.9)$$

Όπου λ είναι η παράμετρος κλίμακας και α η παράμετρος σχήματος.

Η συνάρτηση κινδύνου της κατανομής Weibull είναι της μορφής:

$$h(x) = \lambda a x^{a-1} \quad (2.10)$$

Η συνάρτηση του κινδύνου είναι πιο ευέλικτη και επιτρέπει διαφορετικές μορφές κινδύνου (Guo, 2011). Σύμφωνα με τον (Erik Drysdale, 2017) , η μεταβλητή a δηλώνει την μορφή της συνάρτησης του κινδύνου:

- α>1, αυξανόμενος ρυθμός κινδύνου

- α = 1, σταθερός ρυθμός κινδύνου
- α<1, μειωμένος ρυθμός κινδύνου

2.1.5.3. Λογαριθμική κανονική (Log-normal) κατανομή

Συχνά η διάρκεια επιβίωσης των παρατηρήσεων καταμετρώνται με βάση την λογαριθμική κλίμακα, με αποτέλεσμα την αύξηση της συμμετρίας στα δεδομένα μας. Σε αυτές τις περιπτώσεις, σύμφωνα με τον (Wienke, 2007), καθίσταται σημαντικό να εξάγουμε τη λογαριθμική κανονική κατανομή, έτσι ώστε τα μετατρεπόμενα δεδομένα να είναι κανονικά κατανεμημένα.

Εάν *Φ(x)* είναι η συνάρτηση αθροιστικής κατανομής από την κανονική κατανομή μίας τυχαίας μεταβλητής, η συνάρτηση επιβίωσης της *X* που ακολουθεί την λογαριθμική κατανομή (*X~lognormal(μ, σ)*) είναι της μορφής:

$$S(x) = P(X > x) = P(\ln(X) > \ln(x)) = 1 - \Phi\left(\frac{\ln(x) - \mu}{\sigma}\right) \quad (2.11)$$

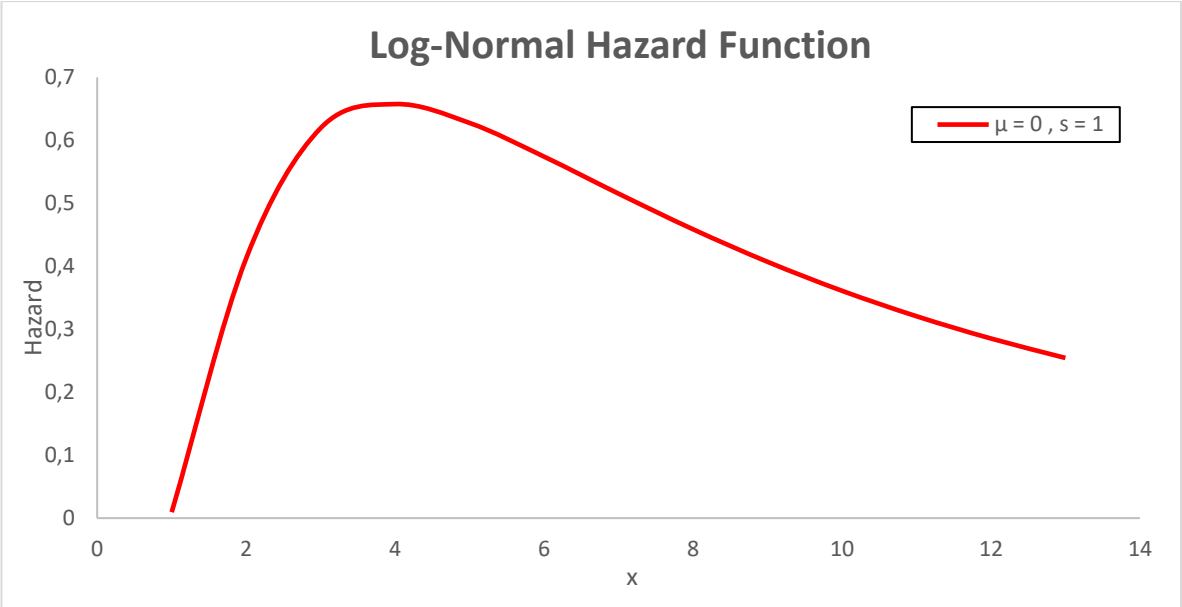
Όπου ο κίνδυνος μπορεί να από αποκτηθεί διαφοροποιώντας την συνάρτηση επιβίωσης *S(x)*. Επίσης ως $\Phi(\frac{\ln(x)-\mu}{\sigma})$ ορίζεται η αθροιστική συνάρτηση κατανομής (*cumulative distribution function*) της κανονικής κατανομής (Guo, 2011).

Η συνάρτηση κινδύνου της λογαριθμικής κατανομής είναι της μορφής:

$$h(x) = \frac{\frac{1}{\sqrt{2\pi}\sigma x} e^{-0.5(\frac{lnx-\mu}{\sigma})^2}}{1 - \Phi(\frac{\ln(x) - \mu}{\sigma})} \quad (2.12)$$

Από τον παραπάνω τύπο παρατηρούμε ότι στον παρονομαστή εμφανίζεται η συνάρτηση επιβίωσης *S(x)*, ενώ ο αριθμητής αποτελείται από την συνάρτηση της λογαριθμικής κανονικής κατανομής.

Η συνάρτηση κινδύνου της λογαριθμικής κανονικής κατανομής θα μπορούσαμε να πούμε πως έχει σχήμα ‘εξογκώματος’. Σταδιακά ο κίνδυνος αυξάνεται έως ότου φθάσει σε κάποιο ανώτατο σημείο και έπειτα μειώνεται αγγίζοντας το 0, καθώς μεγαλώνει ο χρόνος επιβίωσης των παρατηρήσεων.



Διάγραμμα 3 - Συνάρτηση κινδύνου λογαριθμικής κανονικής κατανομής

Δεν αποτελεί την κατάλληλη κατανομή για την μοντελοποίηση της διάρκειας ζωής σε περίπτωση που ο κίνδυνος να εκδηλωθεί το γεγονός αυξάνεται με την πάροδο του χρόνου.

Όπως ορίζει ο (Guo, 2011), χρησιμοποιώντας κάποιο κομμάτι της λογαριθμικής κανονικής κατανομής μπορούμε να μοντελοποιήσουμε την έναρξη κάποιας ασθένειας (Guo, 2011).

2.2. Kaplan-Meier, Καμπύλες επιβίωσης και Log-Rank Test

2.2.1. Kaplan-Meier

Ο εκτιμητής *Kaplan-Meier* ή '*product limit*' πήρε την συγκεκριμένη ονομασία λόγω των εφευρετών του '*Edward L. Kaplan και Paul Meier'* (1958) μετά την δημοσίευση παρόμοιου ερευνητικού άρθρου στο περιοδικό της Αμερικάνικης Στατιστικής Ένωσης (*Journal of the American Statistical Association*) (Wikipedia, 2021). Αποτελεί μία μη-παραμετρική μέθοδό που χρησιμοποιείται προκειμένου να εκτιμηθεί η συνάρτηση επιβίωσης. Στην επιστήμη της ιατρικής χρησιμοποιείται κυρίως για την μέτρηση των ασθενών οι οποίοι επιβιώνουν για ένα χρονικό διάστημα μετά από κάποιου είδους θεραπεία. Όσον αφορά τα υπόλοιπα επιστημονικά πεδία, θα μπορούσε να χρησιμοποιηθεί στην μέτρηση της χρονικής διάρκειας των ανθρώπων που παραμένουν άνεργοι λόγω της παραίτησης ή απόλυσης τους από μία θέση εργασίας, την διάρκεια ζωής μίας μηχανής και το διάστημα που τα φρέσκα φρούτα παραμένουν στο φυτό μέχρι την στιγμή που θα φαγωθεί από κάποιο ζώο.

Σύμφωνα με τον (Lewinson, 2020), ο εκτιμητής *Kaplan-Meier* αποτελεί την πιο απλή μέθοδο προκείμενου να προσδιοριστεί η επιβίωση των παρατηρήσεων κατά το πέρασμα του χρόνου και

για αυτό τον λόγο αποτελεί αρχικό στάδιο στην ανάλυση επιβίωσης. Οι καμπύλες που εξάγονται με την χρήση του εκτιμητή *Kaplan-Meier* προσδιορίζουν το γεγονός του ενδιαφέροντος, την αποκοπή και την πιθανότητα επιβίωσης για την καθεμία παρατήρηση.

Προκειμένου να μπορέσουμε να χρησιμοποιήσουμε την προσέγγιση του *Kaplan-Meier* ισχυριζόμαστε τις παρακάτω συνθήκες:

- Το γεγονός του ενδιαφέροντος είναι σαφές και συμβαίνει σε μία καθορισμένη χρονική στιγμή.
- Η πιθανότητα επιβίωσης είναι η ίδια για τις παρατηρήσεις που εισήλθαν νωρίς ή αργά στην μελέτη, χωρίς να έχει ιδιαίτερα σημασία το πότε εισήλθε η καθεμιά.
- Οι παρατηρήσεις οι οποίες 'απέτυχαν' αλλά και οι λογοκριμένες παρατηρήσεις, έχουν τις ίδιες προοπτικές επιβίωσης με τις υπόλοιπες παρατηρήσεις για τις οποίες δεν έχει εκδηλωθεί ακόμα το γεγονός.

Στην πραγματικότητα δεν μπορούμε να ξέρουμε την πραγματική συνάρτηση επιβίωσης. Με την χρήση του εκτιμητή *Kaplan-Meier* απλά μπορούμε να προσεγγίζουμε την πραγματική συνάρτηση επιβίωσης με βάση τα δεδομένα που περιλαμβάνει το δείγμα μας.

Η συνάρτηση του εκτιμητή *Kaplan-Meier* είναι της μορφής:

$$\hat{S}(t) = \prod_{i:\ t_i \le t} \left(1 - \frac{d_i}{n_i}\right) \quad (2.13)$$

Όπου t_i δηλώνει τον χρόνο όπου τουλάχιστον μία παρατήρηση έχει 'αποτύχει'. Ως d_i δηλώνεται ο αριθμός των γεγονότων που συνέβησαν την χρονική στιγμή t_i. Τέλος, το n_i δηλώνει τον αριθμό των παρατηρήσεων που έχουν επιβιώσει μέχρι την χρονική στιγμή t_i (δεν έχουν ακόμα εκδηλώσει το γεγονός του ενδιαφέροντος ή αποκοπεί μέχρι την καθορισμένη στιγμή). Διατυπώνοντας το διαφορετικά, θα λέγαμε πως συμβολίζει τον αριθμό των παρατηρήσεων που βρίσκονται σε κίνδυνο την χρονική στιγμή t_i (Lewinson, 2020).

Η συνάρτηση *Kaplan-Meier* εκτιμά το κλάσμα των παρατηρήσεων που έχουν επιβιώσει μετά από κάθε χρονική στιγμή *t*, ακόμα και στην περίπτωση που κάποια από τα δεδομένα δεν έχουν παρατηρηθεί ακόμα πως έχουν 'αποτύχει', με το δείγμα να είναι αρκετά μικρό. Προκύπτει υπολογίζοντας την πιθανότητα εμφάνισης του γεγονότος για μία συγκεκριμένη χρονική στιγμή. Όπως παρουσιάζεται στο άρθρο των εκτίμηση (Etikan, Abubakar, & Alkassim, 2017), οι διαδοχικές πιθανότητες που έχουν αποκτηθεί, θα πολλαπλασιαστούν με τυχόν προηγούμενες υπολογισμένες πιθανότητες εμφάνισης του γεγονότος προκειμένου να προσδιοριστεί η τελική εκτίμηση.

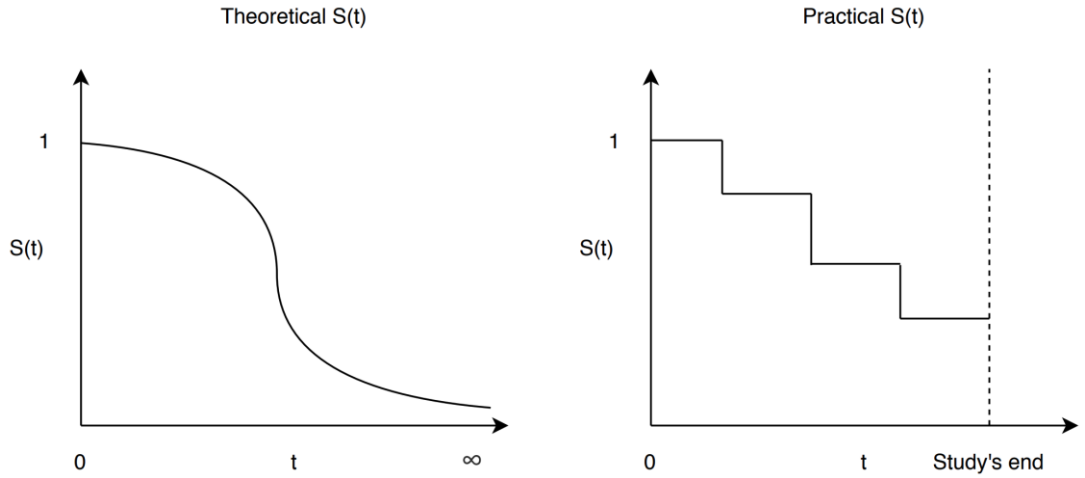
Επίσης σύμφωνα με τους (Etikan, Abubakar, & Alkassim, 2017), ένας αρκετά σημαντικός περιορισμός του εκτιμητή *Kaplan-Meier* αποτελεί το γεγονός ότι είναι αδύνατη χρήση του όταν

το δείγμα μας περιλαμβάνει πολλές μεταβλητές και επιθυμούμε να εξετάσουμε την επίδραση της καθεμίας μεταβλητής. Ο παραπάνω περιορισμός οφείλεται στο ότι εξετάζει την επίδραση ενός μόνο παράγοντα την κάθε χρονική στιγμή που μελετάμε.

2.2.2. Καμπύλες επιβίωσης (Survival Curves)

Οι καμπύλες επιβίωσης μας παρέχουν την δυνατότητα να αναπαραστήσουμε γραφικά τα αποτελέσματα της ανάλυσης από την χρήση της μεθόδου *Kaplan-Meier*, σε όλη την χρονική διάρκεια της μελέτης. Από την ερμηνεία των (Kleinbaum & Klein, 2005), οι καμπύλες επιβίωσης αποτελούν απαραίτητο 'εργαλείο' της ανάλυσης καθώς μπορούν να χρησιμοποιηθούν προκειμένου να συγκρίνουμε την επιβίωση για δύο διαφορετικά υποσύνολα του δείγματος μας. Για παράδειγμα κάποιο τυχαίο δείγμα θα μπορούσε να διαιρεθεί σε δύο επιμέρους υποσύνολα όπου το ένα θα περιελάβανε τις παρατηρήσεις που έχουν λάβει κάποιου είδους θεραπεία και το εναπομείναν θα περιείχε οι παρατηρήσεις που δεν έχουν λάβει την θεραπεία.

Η καμπύλη της επιβίωσης είναι πάντα φθίνουσα και μπορούσαμε να πούμε πως στην πραγματικότητα αποτελείται από μία σειρά μικρών ή μεγάλων 'σκαλοπατιών' κατηφορικής κλίσεως, έχοντας μία μορφή βηματισμού.



Εικόνα 3 - Σύγκριση της πρακτικής καμπύλης επιβίωσης με την θεωρητική καμπύλη επιβίωσης

Οι καμπύλες επιβίωσης συνοδεύονται από διαστήματα εμπιστοσύνης προκειμένου να παρουσιαστεί η ύπαρξη αβεβαιότητας των εκτιμήσεων που έχουν γίνει. Πλατιά διαστήματα εμπιστοσύνης δηλώνουν υψηλή αβεβαιότητα για την εκτίμηση που έχει γίνει, αυτό μπορεί να οφείλεται στο πολύ μικρό δείγμα που μπορεί να έχει χρησιμοποιηθεί.

Η ερμηνεία της καμπύλης επιβίωσης είναι εξαιρετικά απλή. Στον κάθετο άξονα *y* παρουσιάζεται η πιθανότητα για τις παρατηρήσεις που δεν έχουν εκδηλώσει ακόμα το γεγονός του

ενδιαφέροντος, επιβιώνοντας μέχρι την χρονική στιγμή *t* η οποία παρουσιάζεται στον οριζόντιο άξονα *x*. Κάθε πτώση που παρουσιάζεται στην καμπύλη επιβίωσης προκαλείται λόγω της εκδήλωσης του γεγονότος ενδιαφέροντος σε τουλάχιστον μία παρατήρηση. Το μέγεθος της πτώσης στην καμπύλη επιβίωσης δηλώνει το πλήθος των παρατηρήσεων που εκδήλωσαν το γεγονός του ενδιαφέροντος μία συγκεκριμένη χρονική στιγμή. Μεγαλύτερη πτώση στην καμπύλη επιβίωσης σημαίνει πως περισσότερες παρατηρήσεις εκδήλωσαν το γεγονός του ενδιαφέροντος την συγκεκριμένη χρονική στιγμή ή περίοδο. Όταν καμία παρατήρηση δεν έχει εκδηλώσει το γεγονός του ενδιαφέροντος ή ορισμένες παρατηρήσεις χαρακτηρίζονται από λογοκρισία, τότε δεν υπάρχει πτώση στην καμπύλη επιβίωσης.

2.2.3. Log-Rank test

Το *Log-Rank* test αποτελεί μία διάσημη μέθοδο προκειμένου να ελέγξουμε αν υπάρχει διαφορά μεταξύ δύο υποσυνόλων του δείγματος μας. Σύμφωνα με τους (Kleinbaum & Klein, 2005), συγκρίνοντας τις καμπύλες επιβίωσης των δύο διαφορετικών υποσυνόλων που έχουν προκύψει μέσω του εκτιμητή *Kaplan-Meier* και αν όντως υπάρχει στατιστικά σημαντική διαφορά μεταξύ τους, θα καταλήξουμε στο τελικό συμπέρασμα αν η διαφορά των δύο υποσυνόλων όσον αφορά την επιβίωση είναι στατιστικά σημαντική. Η μηδενική υπόθεση υποστηρίζει πως δεν υπάρχει διαφορά στα δύο υποσύνολα, όσον αφορά την πιθανότητα εκδήλωσης του ενδιαφέροντος σε κάθε χρονική στιγμή. Η μηδενική υπόθεση θα απορριφθεί, αποδεχόμενοι την εναλλακτική υπόθεση ότι υπάρχει στατιστικά σημαντική διαφορά μεταξύ των υποσυνόλων, για τιμή *p-value* μικρότερη από το επίπεδο σημαντικότητας ($\alpha = 0.05$). Ενώ δεν θα μπορούμε να απορρίψουμε την μηδενική υπόθεση για μία μεγάλη τιμή *p-value*. Στην περίπτωση που κάποιες καμπύλες των υποσυνόλων τέμνονται τότε το *Log-Rank test* ίσως να αδυνατεί να προβλέψει πως υπάρχει στατιστικά σημαντική διαφορά μεταξύ τους (Etikan, Abubakar, & Alkassim, 2017).

Ο τύπος του *Log-Rank statistic* για δύο διαφορετικά υποσύνολα του δείγματος είναι της μορφής (Kleinbaum & Klein, 2005):

$$Log - Rank\ statistic = \frac{(O_i - E_i)^2}{Var(O_i - E_i)} \quad (2.14)$$

Όπου O_2 είναι το σύνολο των δεδομένων που έχει παρατηρηθεί ότι έχουν εκδηλώσει το γεγονός του ενδιαφέροντος την κάθε χρονική στιγμή για κάθε ένα από τα δύο υποσύνολα. Ως E_2 δηλώνεται το σύνολο παρατηρήσεων που αναμένουμε να εκδηλώσουν γεγονότος του ενδιαφέροντος για καθένα από τα δύο υποσύνολα.

Σε περίπτωση που επιθυμούμε να βρούμε αν υπάρχει στατιστικά σημαντική διαφορά μεταξύ τριών υποσυνόλων ενός τυχαίου δείγματος τότε θα πρέπει να υπολογίσουμε και τις συν

διακυμάνσεις (*covariances*) μεταξύ των υποσυνόλων, με τον τύπο σε αυτή την περίπτωση να γίνεται αρκετά πιο περίπλοκο. Επειδή το *Log-Rank statistic* ακολουθεί την *Chi-square* κατανομή, προκειμένου να βρούμε την τιμή του *Log-Rank test* θα πρέπει να ακολουθήσουμε την παραπάνω κατανομή και να εντοπίσουμε την τιμή *p-value* για έναν βαθμό ελευθερίας ,από την στιγμή που έχουμε δύο υποσύνολα, που αντιστοιχεί για την τιμή *Log-Rank statistic* που έχουμε υπολογίσει.

Μία προσεγγιστική εκτίμηση του *Log-Rank statistic* , από τους (Goel, Khanna, & Kishore, 2010), θα μπορούσε να γίνει και με την χρήση του τύπου:

$$X^2 \approx \sum_{i}^{number\ of\ groups} \frac{(O_i - E_i)^2}{E_i} \quad (2.15)$$

2.3. Επισκόπηση μεθόδων ανάλυσης

2.3.1. Cox PH Model

Μία από τις πιο δημοφιλές τεχνικές παλινδρόμησης στην ανάλυση επιβίωσης αποτελεί η παλινδρόμηση αναλογικού κινδύνου της *Cox* (*Cox proportional hazards regression*), η οποία χρησιμοποιείται ώστε να συνδυάσει διάφορους παράγοντες οι οποίοι θα μπορούσαν να επηρεάσουν το γεγονός του ενδιαφέροντος. Οι παράγοντες ή οι διάφορες μεταβλητές όπως επίσης θα μπορούσαμε να πούμε, εξετάζονται ταυτόχρονα με τον χρόνο επιβίωσης των παρατηρήσεων στο δείγμα μας. Όπως αναφέρουν οι (Emmert-Streib & Dehmer, 2019), τα μοντέλα αναλογικού κινδύνου της *Cox* ανήκουν στην κατηγορία των ημί-παραμετρικών μοντέλων, επειδή υπάρχει η δυνατότητα της παραμετρικής υπόθεσης λαμβάνοντας υπόψιν την επίδραση των διάφορων επεξηγηματικών μεταβλητών στην συνάρτηση του κινδύνου, αλλά δίχως να κάνει καμία υπόθεση σχετικά με την βασική μορφή της συνάρτησης του κινδύνου $h_0(t)$, έχοντας απροσδιόριστη λειτουργία. Από τα μοντέλα αναλογικού κινδύνου θεωρείται ότι οι επεξηγηματικές μεταβλητές που χρησιμοποιούνται είναι εκείνες που επηρεάζουν την συνάρτηση του κινδύνου, δίχως να έχει κάποια σταθερή μορφή σε όλη την διάρκεια της ανάλυσης. Όπως παραδείγματος χάριν η σταθερή μορφή συνάρτησης κινδύνου που εξάγεται από την χρήση του εκθετικού μοντέλου. Σύμφωνα με την (Lisa, 2016), στην περίπτωση που κάποιος ενδιαφέρεται περισσότερο για την επίδραση των επεξηγηματικών μεταβλητών στο τελικό αποτέλεσμα παρά για την μορφή της συνάρτησης του κινδύνου $h(t)$,μπορεί να αγνοήσει την βασική γραμμή (baseline hazard) της συνάρτησης του κινδύνου $h_0(t)$ η οποία στα *Cox* μοντέλα συχνά δεν φαίνεται πως έχει πρωταρχικό ρόλο.

Οι βασικές υποθέσεις που θα πρέπει να τηρούνται προκειμένου το *Cox* μοντέλο να μπορεί να χρησιμοποιηθεί είναι οι εξής (Lisa, 2016):

- Ανεξαρτησία των χρόνων επιβίωσης μεταξύ των παρατηρήσεων του δείγματος.
- Να υπάρχει κάποια σχέση μεταξύ των επεξηγηματικών μεταβλητών και του κίνδυνο (γραμμική σχέση).
- Σταθερή αναλογία κινδύνου (hazard ratio) κατά την διάρκεια της μελέτης. Δηλαδή, η επίδραση των επεξηγηματικών μεταβλητών στον κίνδυνο εκδήλωσης του γεγονότος να έχει σταθερό ρυθμό όσο κυλάει η μελέτη.

Η εξαρτημένη μεταβλητή ενός μοντέλου αναλογικού κινδύνου είναι η συνάρτηση του κινδύνου (*hazard function*) και προσδιορίζεται από τον τύπο:

$$h(t, X) = h_0(t)e^{(\Sigma_i^p \beta_i X)} \quad (2.16)$$

Όπου, $h_0(t)$ αποκαλείται ως βασική γραμμή (*baseline hazard*) του κινδύνου. Η βασική 'γραμμή' της συνάρτησης του κινδύνου μπορεί να έχει οποιαδήποτε μορφή. Με β_i δηλώνεται ο συντελεστής των μεταβλητών του δείγματος.

Στην περίπτωση που δεν υπάρχουν μεταβλητές στο δείγμα μας ή οι συντελεστές τους ισούνται με το μηδέν ($\beta = 0$) τότε:

$$h(t, X) = h_0(t) \quad (2.17)$$

Όπου σε αυτή την περίπτωση έχουμε την συνάρτηση του κινδύνου δίχως όμως την επίδραση των ανεξάρτητων μεταβλητών (Lisa, 2016).

Το μοντέλο αναλογικού κινδύνου όπως και η μέθοδος *Kaplan-Meier* χρησιμοποιεί το δημοφιλές *Log-Rank test* προκειμένου να συγκρίνει τις πιθανότητες επιβίωσης μεταξύ δύο υποσυνόλων του δείγματος που μελετάμε. Όσον αφορά την εκτίμηση των συντελεστών της παλινδρόμησης, το *Cox* μοντέλο είναι τόσο αποτελεσματικό όσο και τα υπόλοιπα παραμετρικά μοντέλα (π.χ. *Weibull*). Ακόμα και στην περίπτωση που δεν ικανοποιούνται όλες οι συνθήκες όσον αφορά την εφαρμογή παραμετρικών μοντέλων μία πολύ καλή λύση μπορεί να δοθεί με την χρήση του *Cox* μοντέλου. Για παράδειγμα, όταν η χρήση *Weibull* μοντέλου δεν μπορεί να εφαρμοστεί για τον λόγο ότι το σύνολο δεδομένων μας στο δείγμα δεν ακολουθεί την κατανομή επιβίωσης *Weibull* τότε η χρήση *Cox* μοντέλου μπορεί να επιλύσει αποτελεσματικά το πρόβλημα. Οπότε στην περίπτωση που υπάρχουν αμφιβολίες σχετικά με το σωστό μοντέλο το οποίο θα ερμηνεύει και καλύτερα τα αποτελέσματα, η περίπτωση επιλογής του *Cox* μοντέλου αποτελεί μία ασφαλή επιλογή από το να ρισκάρουμε, επιλέγοντας κάποιο εσφαλμένο παραμετρικό μοντέλο (Kleinbaum & Klein, 2005).

Με το μοντέλο παλινδρόμησης *Cox*, προσπαθούμε να εκτιμήσουμε το μέγεθος επίδρασης του κινδύνου (*hazard rate*). Ως βαθμό κινδύνου (*hazard rate*) ορίζουμε τον κίνδυνο ή την πιθανότητα

που έχει κάποια παρατήρηση του δείγματος να εκδηλώσει το γεγονός του ενδιαφέροντος, με την προϋπόθεση ότι έχει επιβιώσει μέχρι εκείνη την χρονική στιγμή *t*.

2.3.1.1. Αναλογία κινδύνου (Hazard Ratio)

Το *Log-Rank test*, σύμφωνα με τους (Clark, Bradburn, Love, & Altman, 2003), όπως επίσης και πολλοί άλλοι έλεγχοι, υποδηλώνει αν υπάρχουν στατιστικά σημαντικές διαφορές στην επιβίωση μεταξύ δύο υποσυνόλων δίχως να παρέχεται η δυνατότητα να δούμε το μέγεθος της διαφοράς που υπάρχει. Αυτή η διαφορά μπορεί να υπολογιστεί με την χρήση της αναλογίας κινδύνου (*hazard ratio*). Ως ορισμός της αναλογίας κινδύνου (*hazard ratio*) δηλώνεται η σύγκρισή υποσυνόλων του δείγματος μας ανάλογα με τον βαθμό κινδύνου που υπάρχει για το καθένα. Πιο συγκεκριμένα η αναλογία κινδύνου εκφράζει τον λόγο των συνολικών γεγονότων που έχει παρατηρηθεί ότι έχουν εκδηλωθεί (*observed events*) προς αυτά που αναμένονταν να εκδηλωθούν (*expected events*) για την σύγκριση δύο ανεξάρτητων υποσυνόλων στο δείγμα μας.

$$HR = \frac{\sum O_A / \sum E_A}{\sum O_B / \sum E_B} \quad (2.18)$$

Όπου *O* και *E* δηλώνουν το σύνολο γεγονότων που έχουν παρατηρηθεί και το σύνολο των αναμενόμενων γεγονότων αντίστοιχα. Ενώ ως *A* και *B* αναφέρονται τα δύο υποσύνολα που επιθυμούμε να συγκρίνουμε ως προς τον κίνδυνο εκδήλωσης του γεγονότος ενδιαφέροντος.

Στην περίπτωση που έχουμε αναλογία κινδύνου μεγαλύτερη της μονάδας (*HR >1*) τότε αυτό υποδηλώνει ότι η επιβίωση των παρατηρήσεων για το υποσύνολο *A* είναι μικρότερη σε σχέση με το υποσύνολο *B*. Αντιστοίχως, η επιβίωση για το υποσύνολο *B* είναι μικρότερη σε σχέση με το υποσύνολο *A* στην περίπτωση που η αναλογία κινδύνου είναι μικρότερη της μονάδας (*HR<1*). Στην περίπτωση όπου η αναλογία κινδύνου είναι ίση με την μονάδα (*HR=1*) τότε ο κίνδυνος να εκδηλωθεί το γεγονός του ενδιαφέροντος είναι ο ίδιος και για τα δύο υποσύνολα (STHDA, n.d.).

Ας υποθέσουμε ότι επιθυμούμε να συγκρίνουμε δύο υποσύνολα σχετικά με την αναλογία κινδύνου (*hazard ratio*). Οι αναμενόμενοι κίνδυνοι για το πρώτο σύνολο *A* και για το δεύτερο σύνολο *B* είναι:

$$h_A(t) = h_0(t)e^{(\beta_1 X_A)} \quad (2.19)$$

$$h_B(t) = h_0(t)e^{(\beta_1 X_B)} \quad (2.20)$$

Η αναλογία κινδύνου μεταξύ των δύο υποσυνόλων παρατηρούμε ότι είναι ανεξάρτητη του χρόνου. Οπότε μπορούμε να εξάγουμε το συμπέρασμα ότι η αναλογία κινδύνου των δύο υποσυνόλων δεν επηρεάζεται από τον χρόνο.

Μπορούμε να αποκτήσουμε τα διαστήματα εμπιστοσύνης για τον λογάριθμο της αναλογίας κινδύνου (*hazard ratio*). Σύμφωνα με την προσέγγιση του *Log-Rank* (GraphPad, 2016):

το 95% κάτω όριο του διαστήματος εμπιστοσύνης είναι της μορφής:

$$e^{\left(L-1.96\sqrt{\frac{1}{E_A}+\frac{1}{E_B}}\right)} \quad (2.21)$$

το 95% άνω όριο του διαστήματος εμπιστοσύνης είναι της μορφής:

$$e^{\left(L+1.96\sqrt{\frac{1}{E_A}+\frac{1}{E_B}}\right)} \quad (2.22)$$

Όπου *L* δηλώνεται ως ο φυσικός λογάριθμος της αναλογίας του κινδύνου.

$$L = \ln HR \quad (2.23)$$

Αξίζει να σημειωθεί ότι, σε περίπτωση που το πλάτος των διαστημάτων εμπιστοσύνης είναι αρκετά μεγάλο ή ακόμα μπορεί να εμπεριέχεται και η τιμή της μονάδας (*HR = 1*) τότε η τιμή που έχουμε αποκτήσει όσον αφορά την αναλογία κινδύνου για τα δύο υποσύνολα δεν είναι στατιστικά σημαντική. Η αναλογία κινδύνου μας πληροφορεί σχετικά με το μέγεθος της διαφοράς μεταξύ των δύο υποσυνόλων αλλά δεν μας δίνει καμία πληροφορία για την απόλυτη διαφορά όσον αφορά στην διάρκεια της επιβίωσης (GraphPad, 2016).

2.3.1.2. Προσαρμοσμένες καμπύλες επιβίωσης χρησιμοποιώντας το Cox μοντέλο

Αφού είδαμε πως θα μπορούσαμε να υπολογίζουμε την αναλογία κινδύνου (*hazard ratio)* μεταξύ υποσυνόλων, τώρα θα μελετήσουμε πως θα μπορούσαμε να αποκτήσουμε καμπύλες επιβίωσης χρησιμοποιώντας *Cox* μοντέλα. Προς υπενθύμιση, σε περίπτωση που δεν χρησιμοποιήσουμε κανένα μοντέλο για να προβούμε σε ανάλυση επιβίωσης, τότε οι καμπύλες επιβίωσης θα μπορούσαν να εκτιμηθούν χρησιμοποιώντας την μέθοδο *Kaplan-Meier*. Όταν χρησιμοποιούμε *Cox* μοντέλα, οι καμπύλες μπορούν να εκτιμηθούν από τις επεξηγηματικές μεταβλητές όπου λειτουργούν ως εκτιμητές για το μοντέλο (Kleinbaum & Klein, 2005).

Σύμφωνα με τους (Kleinbaum & Klein, 2005), μέσω της συνάρτησης κινδύνου μπορεί να αποκτηθεί η αντίστοιχη συνάρτηση επιβίωσης με τις κατάλληλες μετατροπές όπως παρουσιάζεται παρακάτω.

Η συνάρτηση κινδύνου που αναπαράγεται από την χρήση του Cox μοντέλου:

$$h(t,X) = h_0(t)e^{(\sum_i^p \beta_i X)} \quad (2.24)$$

Η αντίστοιχη συνάρτηση επιβίωσης που μπορεί να εξαχθεί:

$$S(t,X) = [S_0(t)]^{e^{(\sum_i^p \beta_i X)}} \quad (2.25)$$

Η συνάρτηση επιβίωσης αποτελεί την 'βάση' προκειμένου να μπορέσουμε να εκτιμήσουμε τις καμπύλες επιβίωσης. Αξίζει να σημειωθεί ότι από τον παραπάνω τύπο, η συνάρτηση επιβίωσης την καθορισμένη χρονική στιγμή *t* για κάποια παρατήρηση, δίνεται από την βασική γραμμή της ανάλυσης επιβίωσης $S_0(t)$ υψωμένη στην $e^{(\Sigma_i^p \beta_i X)}$ (Kleinbaum & Klein, 2005).

Γενικά όταν εκτιμούμε τις προσαρμοσμένες καμπύλες επιβίωσης, η τιμή που επιλέγεται για κάθε μεταβλητή από το μοντέλο που χρησιμοποιούμε, είναι η μέση τιμή για τον κάθε εκτιμητή ή μπορεί και ο διάμεσος. Γενικά μιλώντας εάν επιθυμούμε να συγκρίνουμε τις καμπύλες επιβίωσης δύο υποσυνόλων του δείγματος μας όσον αφορά τα δύο επίπεδα μίας επεξηγηματικής μεταβλητής τότε μπορούμε να χρησιμοποιήσουμε τον παρακάτω τύπο:

$$S(t, X) = [S_0(t)]^{e^{(\widehat{\beta}_i(1) + \Sigma_{i \neq 1} \widehat{\beta}_i \overline{X}_i)}} \quad (2.26)$$

Επίσης σε περίπτωση που επιθυμούμε να αποκτήσουμε προσαρμοσμένες καμπύλες επιβίωσης οι οποίες προσαρμόζουν όλες τις επεξηγηματικές μεταβλητές του μοντέλου μας, τότε ο γενικευμένος τύπος όπου χρησιμοποιείται η μέση τιμή του κάθε εκτιμητή είναι ο εξής:

$$S(t, X) = [S_0(t)]^{e^{(\Sigma \widehat{\beta}_i \overline{X}_i)}} \quad (2.27)$$

Από την παραπάνω διατύπωση της καμπύλης επιβίωσης, η πιθανότητα επιβίωσης μπορεί να υπολογιστεί για κάθε προκαθορισμένη χρονική στιγμή. Η επιλογή των χρόνου *t* αφορά τις χρονικές στιγμές των παρατηρήσεων από το δείγμα μας οι οποίες εκδήλωσαν το γεγονός του ενδιαφέροντος. Το διάγραμμα των καμπυλών επιβίωσης που αποκτάται για τα δύο υποσύνολα του δείγματος έχει την μορφή 'σκαλοπατιών' καθοδικής πορείας όπως και με την μέθοδο *Kaplan-Meier*.

2.3.2. Δέντρα Επιβίωσης (Survival Trees)

Τα δέντρα αποφάσεων (*decision trees*) αποτελούν μοντέλα πρόβλεψης στο κλάδο της μηχανικής μάθησης (*machine learning*). Το 1984, οι (Breiman, Friedman, Stone, & Olshen) δημοσίευσαν το βιβλίο '*Classification and Regression* Trees' (*CART*) το οποίο εισήγαγε για πρώτη φορά την ιδέα των δέντρων αποφάσεων στην παγκόσμια κοινότητα. Με την χρήση αυτών των μη-παραμετρικών μοντέλων διαχωρίζεται το σύνολο δεδομένων σε ένα υποσύνολο το οποίο θα χρησιμοποιηθεί προκειμένου να εκπαιδεύσουμε τον αλγόριθμο (*training set*) και στο δεύτερο υποσύνολο όπου θα χρησιμοποιηθεί για να γίνουν οι προβλέψεις (*test set*) και να ερμηνεύσουμε την αποτελεσματικότητα του μοντέλου. Είναι αρκετά δημοφιλή λόγω της μεγάλης ακρίβειας στις προβλέψεις που μπορούν να μας παρέχουν αλλά και της ευκολίας στην ερμηνεία των αποτελεσμάτων του μοντέλου. Δεν καθυστέρησε αρκετά να αναπτυχθεί από τον *Breiman* το 1996 και η ιδέα του *Bagging*, κατά την οποία μπορούμε να αναπαράγουμε πολλά ξεχωριστά δέντρα

αποφάσεων στο *training set*, προκειμένου να κάνουμε τις προβλέψεις μας. Αποτελεί μία τεχνική που μειώνει σημαντικά την διακύμανση και βελτιώνει την απόδοση του μοντέλου κάνοντας πιο ακριβείς προβλέψεις. Ύστερα το 2001 ο *Breiman* εισήγαγε την ιδέα των *Random-Forests*, η μόνη διαφορά αυτής της μεθόδου από το *Bagging* είναι ότι η επιλογή των μεταβλητών για να γίνει ο διαχωρισμός σε κάθε κόμβο γίνεται με τυχαίο τρόπο και όχι ανάλογα με την σημαντικότητα της κάθε μεταβλητής όπως υποστηρίζει η μέθοδος *Bagging*. Λόγω της μεγάλης επιρροής του *Machine Learning* σε πολλούς επιστημονικούς κλάδους, η υιοθέτηση των δέντρων αποφάσεων στην ανάλυση επιβίωσης δεν θα καθυστερούσε να συμβεί. Η ανάπτυξη δέντρων επιβίωσης ξεκίνησε από τα μέσα του 1980 μέχρι τα μέσα του 1990 όπου στόχος των ερευνητών αποτέλεσε η επέκταση των δέντρων αποφάσεων σε δεδομένα επιβίωσης που χαρακτηρίζονται από αποκοπή. Από την στιγμή που ανακαλύφθηκαν βασικές μέθοδοι για την δημιουργία των δέντρων επιβίωσης, οι ερευνητές κινήθηκαν και προς άλλες κατευθύνσεις προκειμένου να λυθούν και άλλα ζητήματα στην ανάλυση επιβίωσης μέσω της χρήσης δέντρων επιβίωσης. Όπως παρουσιάζεται από τον (Shah, 1998), ένα καίριο ζήτημα ήταν η επίλυση πιο πολύπλοκων καταστάσεων, όπως όταν στα δεδομένα επιβίωσης υπάρχουν πολλές μεταβλητές και συσχετίζονται μεταξύ τους. Επίσης, πως θα μπορούσαν να χρησιμοποιηθούν και οι υπόλοιπες μέθοδοι στην ανάπτυξη δέντρων επιβίωσης, όπως '*Bagging'* και '*Random-forest'*, προκείμενου να μελετήσουμε δεδομένα στην ανάλυση επιβίωσης.

2.3.2.1. Random Survival Forests

Ο τρόπος αναπαραγωγής της μεθόδου *Random Survival Forests* προκειμένου να διαχειριστούμε δεδομένα στην ανάλυση επιβίωσης μιμείται την μέθοδο που αναπτύσσονται τα *Random Forests* για την κατασκευή δέντρων αποφάσεων, παρουσιάζοντας ορισμένες διαφορές. Όπως παρουσιάζεται από τους (Kogalur, Blackstone, & Lauer, 2008) ο αλγόριθμος που ακολουθείται για την αναπαραγωγή των *Random Survival Forests*, αρχικά διαχωρίζει το αρχικό σύνολο δεδομένων μας σε ορισμένα τυχαία υποσύνολα (*bootstraps*). Αξίζει να σημειωθεί πως στο κάθε υποσύνολο (bootstrap-sample) δεν εμπεριέχεται κατά μέσο όρο το 37% των δεδομένων από το αρχικό δείγμα. Τα δεδομένα που δεν περιέχονται στο υποσύνολο (*bootstrap*) ονομάζονται *out-of-bag* δεδομένα (*OOB data*). Αφού γίνει ο διαχωρισμός, αναπτύσσεται ένα δέντρο επιβίωσης για το κάθε υποσύνολο (*bootstrap*), όπου επιλέγεται ένας τυχαίος αριθμός *K* μεταβλητών προκειμένου να γίνει ο διαχωρισμός στον κάθε κόμβο του δέντρου. Ο διαχωρισμός στον κόμβο θα πραγματοποιηθεί για την μεταβλητή η οποία μεγιστοποιεί την διαφορά όσον αφορά την επιβίωση των παρατηρήσεων, στο εσωτερικό των δύο επόμενων κόμβων που θα δημιουργηθούν από την διάσπαση του αρχικού κόμβου. Μεγιστοποιώντας την διαφορά επιβίωσης μεταξύ των δύο κόμβων, οι περιπτώσεις επιβίωσης των παρατηρήσεων θα διαφέρουν μεταξύ των κόμβων.

Το δέντρο θα αναπτύσσει όλο και περισσότερους κόμβους, με την προϋπόθεση ότι κάθε καταληκτικός κόμβος θα περιέχει τουλάχιστον μία παρατήρηση που έχει εκδηλώσει το γεγονός του ενδιαφέροντος ($n_{min}>0$). Στην συνέχεια υπολογίζεται η αθροιστική συνάρτηση κινδύνου *CHF (cumulative hazard function)* για το κάθε δέντρο. Αποκτώντας τον μέσο όρο των αθροιστικών συναρτήσεων κινδύνου που έχει αποκτηθεί για το κάθε δέντρο, μπορεί να υπολογιστεί η συνολική αθροιστική συνάρτηση κινδύνου για το μοντέλο. Εφόσον ο αριθμός των κόμβων αυξάνεται και οι ανόμοιες περιπτώσεις διαχωρίζονται, καταλήγουμε στο γεγονός ότι στο εσωτερικό του κάθε κόμβου υπάρχει ομοιογένεια και τα δεδομένα που συγκεντρώνονται παρουσιάζουν παρόμοιες περιπτώσεις επιβίωσης. Τέλος χρησιμοποιώντας τα (*OOB data*), εκτιμάται το σφάλμα πρόβλεψης για την συνολική αθροιστική συνάρτηση κινδύνου (*CHF*).

Όταν το δέντρο φτάσει στην τελική του μορφή όπου δεν επιτρέπεται να γίνουν περισσότεροι διαχωρισμοί στους κόμβους λόγω της ύπαρξης παραδείγματος χάριν μόνο μίας παρατήρησης που έχει εκδηλώσει το γεγονός του ενδιαφέροντος στο εσωτερικό του κόμβου, τότε οι κόμβοι στο τέλος του δέντρου ονομάζονται καταληκτικοί κόμβοι $\mathcal{T}$. Σύμφωνα με τους (Kogalur, Blackstone, & Lauer, 2008) και (Ishwaran & Lu, 2019)Η αθροιστική συνάρτηση κινδύνου για κάποιον καταληκτικό κόμβο h $\in \mathcal{T}$ είναι της μορφής:

$$\widehat{H}_h(t) = \sum_{t_{l,h} \le t} \frac{d_{l,h}}{Y_{l,h}} \quad (2.28)$$

Όπου $d_{l,h}$ και $Y_{l,h}$,δηλώνεται ο αριθμός των παρατηρήσεων που έχουν αποτύχει και ο αριθμός των παρατηρήσεων που βρίσκονται ακόμα σε κίνδυνο αντίστοιχα για την χρονική στιγμή $t_{l,h}$. Ο αριθμός των εκτιμήσεων για τις αθροιστικές συναρτήσεις κινδύνου ενός δέντρου ισούται με τον αριθμό των καταληκτικών κόμβων για το συγκεκριμένο δέντρο.

2.3.2.2. Αντιμετώπιση Αποκοπής στα δέντρα επιβίωσης

Η βασική διαφορά των δέντρων επιβίωσης (*Survival Trees*) σε σχέση με τα δέντρα αποφάσεων (*Decision Trees*) που χρησιμοποιούνται για μία απλή ανάλυση, βασίζεται στην ύπαρξη αποκοπής που μπορεί να υπάρξει μεταξύ των παρατηρήσεων στο δείγμα μας. Συγκεκριμένα μία αποκομμένη παρατήρηση πάντα θα ανήκει σε μία από τις ακόλουθες κατηγορίες:

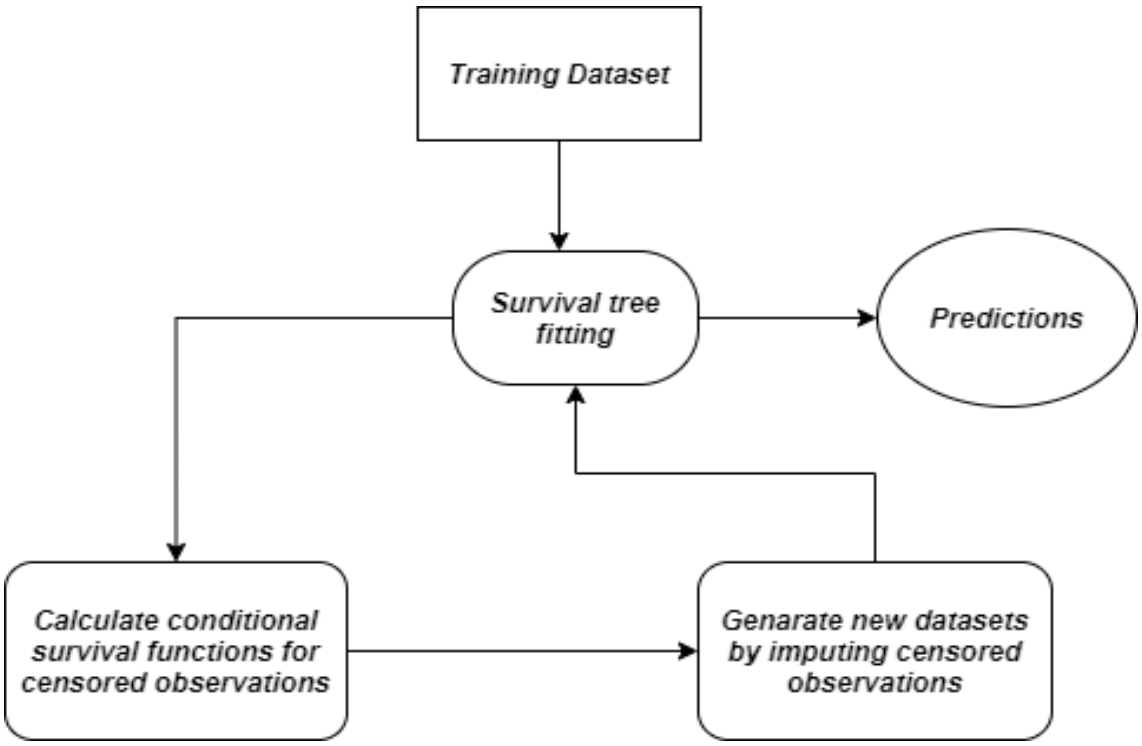
- Ο αληθινός χρόνος επιβίωσης *T* είναι μεγαλύτερος από τον χρόνο που την διήρκησε η μελέτη *τ*, οπότε δεν έχουμε την δυνατότητα να γνωρίζουμε πότε συνέβη το γεγονός του

ενδιαφέροντος ακόμα και στην περίπτωση που παρατηρούμε την αποκομμένη παρατήρηση από την αρχή μέχρι και το τέλος της μελέτης.

- Ο αληθινός χρόνος επιβίωσης *T* είναι μικρότερος από την διάρκεια της μελέτης *τ*, οπότε μπορούμε να παρατηρήσουμε το πότε εκδήλωσε το γεγονός του ενδιαφέροντος στην περίπτωση που η αποκομμένη παρατήρηση εισήλθε στην μελέτης μας και έπαψε να είναι αποκομμένη πριν την λήξη της. Αν και, αυτή η πληροφορία αποκρύπτεται όταν μία παρατήρηση χαρακτηρίζεται από λογοκρισία.

Προκύπτουν δύο πολύ κρίσιμα ερωτήματα επομένως. Πρώτον, πως θα μπορούσαμε να κατηγοριοποιήσουμε τις αποκομμένες παρατηρήσεις; Δεύτερον, με ποιον τρόπο θα μπορούσαμε να αποδώσουμε τιμές στις συγκεκριμένες παρατηρήσεις ανάλογα με την κατηγορία που ανήκουν;

Η λύση για την αποκοπή, όπως ορίζεται από τους (Zhu, 2013) και (Malgorzata, 2010) ,η οποία αποτελεί αναπόσπαστο κομμάτι της ανάλυσης επιβίωσης και εμφανίζεται για μερικές παρατηρήσεις του δείγματος παρέχεται στο παρακάτω διάγραμμα:



Εικόνα 4 - Διαδικασία αντιμετώπισης των αποκομμένων παρατηρήσεων στα δέντρα επιβίωσης

Στην αρχή θα πρέπει να εκπαιδεύσουμε το δέντρο απόφασης στο υποσύνολο του δείγματος που χρησιμοποιούμε για την εκπαίδευση του μοντέλου *training set*, προκειμένου να δούμε την

κατασκευή (*structure*) του δέντρου. Αυτό επιτυγχάνεται με την αναπαραγωγή *M* ανεξάρτητων δέντρων χρησιμοποιώντας το *training set* του δείγματος και ακολουθώντας ορισμένες οδηγίες. Για το κάθε δέντρο, την στιγμή του διαχωρισμού στον κόμβο, θα επιλεχθεί ένας υποψήφιος αριθμός *K* μεταβλητών από το σύνολο των μεταβλητών *p* του δείγματος ,προκειμένου να γίνει ο διαχωρισμός στο κόμβο για μία από τις *K* μεταβλητές. Με την χρήση του *Log-Rank test* θα αποφασιστεί ο καλύτερος διαχωρισμός μεταξύ των *K* μεταβλητών που χρησιμοποιήσαμε έτσι ώστε η διαφορά των συναρτήσεων επιβίωσης μεταξύ των δύο κόμβων που θα παραχθούν να είναι μεγαλύτερη (να υπάρχει όσο το δυνατόν μεγαλύτερη ανομοιότητα μεταξύ των δύο καμπυλών). Από την στιγμή που συμβαίνει ο πρώτος διαχωρισμός, χρησιμοποιείται το ίδιο κριτήριο προκειμένου να γίνουν και οι υπόλοιποι διαχωρισμοί στους κόμβους, μέχρι κάθε καταληκτικός κόμβος να περιέχει τουλάχιστον μία παρατήρηση που έχει 'αποτύχει'.

Μία υποθετική κατανομή επιβίωσης θα υπολογιστεί για κάθε αποκομμένη παρατήρηση που υπάρχει στο δείγμα μας. Οι υπολογισμοί των υποθετικών κατανομών επιβίωσης θα πραγματοποιηθούν αρχικά στο επίπεδο του κόμβου για το κάθε δέντρο και έπειτα θα πάρουμε τον μέσο όρο των κατανομών που έχουμε υπολογίσει από το σύνολο των δέντρων *M* που έχουν αναπαραχθεί. Με την μέθοδο *Kaplan-Meier* έχουμε την δυνατότητα να εκτιμήσουμε την συνάρτηση επιβίωσης $\hat{S}(t)$,όπου $t \in [0, \tau]$,στο εσωτερικό του l^n καταληκτικού κόμβου που έχει δημιουργηθεί στο m^n δέντρο επιβίωσης με την προϋπόθεση ότι στο εσωτερικό του κόμβου υπάρχει τουλάχιστον μία παρατήρηση που έχει εκδηλώσει το γεγονός του ενδιαφέροντος $(n_{min}>0)$. Από την στιγμή που κάθε τυχαία παρατήρηση από το δείγμα μπορεί να πέσει σε έναν μόνο καταληκτικό κόμβο για το καθένα δέντρο επιβίωσης που έχει δημιουργηθεί, παρέχεται η δυνατότητα να αναπαράγουμε μία συνάρτηση επιβίωσης $\hat{S}(t)$ του συγκεκριμένου δέντρου για την i^n παρατήρηση αντί για τον l^n καταληκτικού κόμβο. Έπειτα υπολογίζοντας την μέση συνάρτηση επιβίωσης $\hat{S}(t)$ μεταξύ όλων των δέντρων που έχουν δημιουργηθεί, μπορούμε να έχουμε την τελική συνάρτηση επιβίωσης (*forest level survival function*) η οποία είναι της μορφής:

$$\hat{S}_i = \frac{1}{M}\sum_{m=1}^{M} \hat{S}_m^i \quad (2.29)$$

Δοθέντος ότι η i παρατήρηση χαρακτηρίζεται από αποκοπή την χρονική στιγμή c τότε προσεγγιστικά η υποθετική πιθανότητα επιβίωσης (*conditional probability of survival*), $P(T_i > t|T_i > c)$, για την συγκεκριμένη παρατήρηση είναι:

$$S_i^* = 1 \qquad \text{εάν t} \in [0, c]$$

$$S_i^* = \frac{\hat{S}_i(t)}{\hat{S}_i(c)} \quad \text{εάν t} \in [c, \tau]$$

$$S_i^* = 0 \quad \text{εάν t} \in [\tau, \infty]$$

Όταν κάποια παρατήρηση *i* χαρακτηρίζεται από αποκοπή τότε ο πραγματικός χρόνος εκδήλωσης του γεγονότος ενδιαφέροντος για την συγκεκριμένη παρατήρηση T_i είναι μεγαλύτερος από τον χρόνο Αποκοπής C_i. Εάν κάποια αποκομμένη παρατήρηση παρακολουθείται από την αρχή της μελέτης τότε μία από τις δύο περιπτώσεις θα μπορούσαν να συμβούν:

- Η παρατήρηση θα μπορούσε συνεχίσει να είναι σε κίνδυνο και μετά την λήξη της μελέτης τ δίχως να έχει εκδηλώσει το γεγονός του ενδιαφέροντος. Σε αυτή την περίπτωση δεν μπορούμε να παρατηρήσουμε την χρονική στιγμή που 'απέτυχε' ακόμα και στην περίπτωση που δεν χαρακτηρίζεται από λογοκρισία.
- Η παρατήρηση θα μπορούσε να αποτύχει πριν την λήξη της μελέτης.

Προκειμένου να αντικαταστήσουμε την κάθε αποκομμένη παρατήρηση στο *training set* με μία άλλη παρατήρηση παρέχοντας την σωστή πιθανότητα εκτίμησης (*correctly estimated probability*) θα ακολουθήσουμε τα εξής βήματα:

- Αναπαράγουμε μία νέα μη-αποκομμένη παρατήρηση Y_i^* από την συγκεκριμένη συνάρτηση κατανομής που έχουμε και την παρατηρούμε κατά την διάρκεια της μελέτης. Γνωρίζοντας την μορφή της S_i^*, Y_i^* η νέα παρατήρηση θα πρέπει να βρίσκεται μεταξύ του διαστήματος Y_i (χρόνος Αποκομμένης παρατήρησης) και τ.
- Στην περίπτωση που η νέα παρατήρηση 'αποτύχει' πριν την λήξη της μελέτης $Y_i^* < \tau$ τότε θεωρούμε ότι ο χρόνος που εκδήλωσε το ενδιαφέρον η νέα παρατήρηση T_i είναι μικρότερος από τον συνολικό χρόνο μελέτης τ και αντικαθιστούμε την αποκομμένη παρατήρηση Y_i με την νέα μας παρατήρηση Y_i^* έχοντας δείκτη αποκοπής $\delta_i^* = 1$.
- Στην περίπτωση που φτάσουμε στην λήξη της μελέτης και η νέα παρατήρηση Y_i^* δεν έχει αποτύχει $Y_i^* = \tau$, τότε θεωρούμε ότι ο χρόνος εκδήλωσης του ενδιαφέροντος για την νέα παρατήρηση T_i είναι μεγαλύτερος από τον χρόνο της μελέτης και αντικαθιστούμε την αποκομμένη παρατήρηση Y_i με τ έχοντας νέο δείκτη αποκοπής $\delta_i^* = 0$.

Αυτή την διαδικασία μετατροπής την ακολουθούμε για όλες τις αποκομμένες παρατηρήσεις στο δείγμα μας.

Έχοντας απαλείψει όλες τις αποκομμένες παρατηρήσεις στο *training dataset*, αναπαράγουμε *M* τυχαία datasets και εκπαιδεύουμε ένα δέντρο επιβίωσης για το καθένα από αυτά. Στην συνέχεια συγκεντρώνουμε τα *M* δέντρα που έχουν αναπαραχθεί προκειμένου να καθορίσουμε την νέα μορφή του μοντέλου. Αναλογικά η νέα υποθετική κατανομή αποκοπής μπορεί να υπολογιστεί, για την κάθε αποκομμένη παρατήρηση στο αρχικό υπόδειγμα

σύμφωνα με την αρχική τιμή αποκοπής για τις συγκεκριμένες παρατηρήσεις. Η υποθετική συνάρτηση επιβίωσης '*conditional survival function'* είναι πάντοτε σχετιζόμενη με τις αρχικές αποκομμένες παρατηρήσεις Y_i.

Η τελική πρόβλεψη θα γίνει εκτιμώντας την συνάρτηση επιβίωσης στο εσωτερικό του κάθε κόμβου και στην συνέχεια αποκτώντας τον μέσο όρο των συναρτήσεων επιβίωσης από όλα τα δέντρα που έχουν δημιουργηθεί. Για μία νέα παρατήρηση $X^{\#} = (X_1, \ldots, X_p)$ (όπου $(X_1, \ldots, X_p)$ αποτελούν τις μεταβλητές της παρατήρησης), προϋποθέτοντας ότι η $S^{\#}()$ αποτελεί την πραγματική συνάρτηση επιβίωσης για την συγκεκριμένη παρατήρηση. Η συγκεκριμένη παρατήρηση 'τρέχοντας' τα δέντρα επιβίωσης θα πέσει σε έναν καταληκτικό κόμβο. Όλες οι παρατηρήσεις έπειτα από την μετατροπή που κάναμε στο εσωτερικό του καταληκτικού κόμβου έχουν είτε εκδηλώσει το γεγονός του ενδιαφέροντος είτε είναι αποκομμένες στο τέλος της μελέτης *τ*. Η τελική πρόβλεψη μπορεί να υπολογιστεί από τον τύπο:

$$\hat{S}^{\#}(t) = \frac{1}{M}\sum_{m=1}^{M} \hat{S}_m^{\#}(t) \quad (2.30)$$

Όπου $\hat{S}_m^{\#}$ εκφράζεται από τον τύπο:

$$\hat{S}_m^{\#}(t) = \sum_{i \in\ node} \frac{l\{Y_i > t\}}{\varphi_m(l)} \quad (2.31)$$

Όπου $\varphi_m(l)$ δηλώνει το μέγεθος του κόμβου *l* στο m^{th} δέντρο (Zhu, 2013).

2.3.2.3. Κριτήρια διαχωρισμού (Splitting criteria)

2.3.2.3.1: Log-Rank Test

Η χρήση του *Log-Rank Test* αποτελεί ένα από τα δημοφιλέστερα κριτήρια που μπορούν να χρησιμοποιηθούν προκειμένου να γίνει ο διαχωρισμός στον κόμβο ενός δέντρου επιβίωσης. Σύμφωνα με τον (Segal, 1988), το συγκεκριμένο *test* εκτιμά την διαφορά που υπάρχει στην επιβίωση μεταξύ των δύο κόμβων και προτείνει πιθανούς διαχωρισμούς (*split*) που μπορούν να προκύψουν για τον καθένα, με τις πιθανές μεταβλητές που χρησιμοποιούνται κάθε φορά. Στην συνέχεια επιλέγεται ο καλύτερος διαχωρισμός που μπορεί να γίνει για τους δύο κόμβους , έτσι ώστε η διαφορά στην επιβίωση για τις παρατηρήσεις στο εσωτερικό των δύο κόμβων να μεγιστοποιείται. Το *log rank* μπορεί να υπολογιστεί μέσω της κατασκευής ενός πίνακα στον οποίο θα παρουσιάζονται οι παρατηρήσεις που έχουν αποτύχει κάποια συγκεκριμένη χρονική καθώς και οι παρατηρήσεις που βρίσκονται ακόμα σε κίνδυνο. Οι παρατηρήσεις που βρίσκονται

ακόμα σε κίνδυνο καθώς και εκείνες που έχουν εκδηλώσει το γεγονός του ενδιαφέροντος για κάποια συγκεκριμένη χρονική στιγμή για τον έναν κόμβο, θα συγκρίνονται αντίστοιχα με τις παρατηρήσεις του δεύτερου κόμβου.

Ο τύπος που χρησιμοποιείται για την εκτίμηση του Log-Rank statistic κάποια συγκεκριμένη χρονική στιγμή j^{th} είναι ο εξής (Segal, 1988):

$$X^2_{logrank} = \frac{\left[\sum_{j=1}^{K}\left(d_{0j} - r_{oj}\times\frac{d_j}{r_j}\right)\right]^2}{\sum_{j=1}^{K}\frac{r_{1j}r_{0j}d_j(r_j-d_j)}{r_j^2(r_j-1)}} \quad (2.32)$$

Όπου d_{0j} δηλώνονται οι παρατηρήσεις για τον πρώτο κόμβο που έχουν εκδηλώσει το γεγονός του ενδιαφέροντος κάποια συγκεκριμένη χρονική στιγμή. Ως r_{oj} δηλώνεται ο συνολικός αριθμός παρατηρήσεων που υπάρχουν στον πρώτο κόμβο, ενώ $r_{0j} - d_{oj}$ είναι οι παρατηρήσεις του πρώτου κόμβου που δεν έχουν εκδηλώσει ακόμα το γεγονός του ενδιαφέροντος και βρίσκονται ακόμα σε κίνδυνο. Η ερμηνεία είναι αντίστοιχη και για τις παρατηρήσεις του δεύτερου κόμβου. Ως *K* συμβολίζονται το σύνολο των 'αποτυχιών' των παρατηρήσεων με το πέρασμα του χρόνου.

Παρατηρώντας το κλάσμα, μπορούμε να διακρίνουμε πως ο αριθμητής είναι το άθροισμα των τετραγώνων της διακύμανσης μεταξύ των τιμών για τα δεδομένα που έχει παρατηρηθεί ότι έχουν εκδηλώσει το γεγονός του ενδιαφέροντος *(observed values)* και των αναμενόμενων τιμών για τις παρατηρήσεις του δείγματος (*expected values*). Ο παρονομαστής υπολογίζει την διακύμανση που υπάρχει μεταξύ των παρατηρήσεων που έχουν εκδηλώσει το γεγονός του ενδιαφέροντος στον πρώτο κόμβο d_{0j} . Όπως βλέπουμε από το κλάσμα, η τιμή του $X^2_{logrank}$ αυξάνεται όσο μεγαλώνει η διαφορά μεταξύ των *observed values* και *expected values* για τις παρατηρήσεις ή όσο μικραίνει η διακύμανση του d_{0j}.

2.3.2.3.2. Gordon & Olshen

Σύμφωνα με τον (Alvarez-Iglesias, 2012) , η πρώτη προσπάθεια διατύπωσης ενός τρόπου κατά τον οποίο θα μπορούσε να συμβεί ο διαχωρισμός για κάποια κατηγορική μεταβλητή στα δέντρα επιβίωσης έγινε από τους *Gordon & Olshen* το 1985. Στήριξαν την μέθοδο τους στην ιδέα της 'ακαθαρσίας (*impurity*)' του κόμβου, παρόμοια με το κριτήριο που γίνεται ο διαχωρισμός για μία κατηγορική μεταβλητή στα δέντρα αποφάσεων. Ένας κόμβος θα είναι 'εντελώς καθαρός' (*pure*) στην περίπτωση που όλες οι παρατηρήσεις έχουν εκδηλώσει το γεγονός του ενδιαφέροντος την ίδια χρονική στιγμή.

Προκειμένου να καταμετρήσουμε την 'ακαθαρσία' όπου υπάρχει στο εσωτερικό ενός κόμβου *l* του συνολικού δέντρου *T* μπορούμε να χρησιμοποιήσουμε τον ακόλουθο τύπο:

$$I(l) = p_l d_2\left(S_{KM}, S_{P_s}\right) \quad (2.33)$$

Όπου p_l δηλώνει το ποσοστό των παρατηρήσεων που υπάρχουν στο εσωτερικό του κόμβου *l*. Ως d_2 δηλώνεται η *Wasserstein* απόσταση και εκφράζει την απόσταση που υπάρχει μεταξύ της συνάρτησης επιβίωσης των δεδομένων χρησιμοποιώντας την μέθοδο *Kaplan-Meier* (S_{KM}) και της πλησιέστερης συνάρτησης επιβίωσης που δηλώνει την καθαρότητα που υπάρχει στο εσωτερικό του κόμβου (S_{P_s}). Για τον λόγο αυτό η $d_2(S_{KM}, S_{P_s})$ θα πρέπει να είναι η ελάχιστη απόσταση που μπορεί να υπάρξει μεταξύ της συνάρτησης επιβίωσης του *Kaplan-Meier* και της *pure* συνάρτησης επιβίωσης. Βασιζόμενοι σε αυτή την θεώρηση, ο καλύτερος διαχωρισμός *(split)* θα επιλεχθεί έτσι ώστε η διαφορά στο *impurity* μεταξύ των δύο κόμβων να μεγιστοποιείται. Ο διαχωρισμός σε έναν κόμβο *l*, θα επιφέρει την δημιουργία δύο άλλων μικρότερων κόμβων. Έναν στα δεξιά του αρχικού κόμβου *(l_R)* και έναν στα αριστερά του αρχικού κόμβου (l_L) όπου η διαφορά στο '*impurity'* του αρχικού καταληκτικού κόμβου από το '*impurity'* των δύο μεταγενέστερων κόμβων που έχουν δημιουργηθεί θα πρέπει να είναι μέγιστη. Η διαφορά για το '*impurity'* δίνεται από τον τύπο:

$$\Delta I(l) = I(l) - \left(I_{(l_R)} + I_{(l_L)}\right) \quad (2.34)$$

Η διαδικασία αυτή επαναλαμβάνεται μέχρι στο εσωτερικό κάθε κόμβου να περιέχεται ένας μικρός αριθμός παρατηρήσεων (Alvarez-Iglesias, 2012). Επίσης προτάθηκε η ιδέα χρήσης του '*log-rank statistic'* ή του '*parametric likelihood ratio'* προκειμένου να μετρηθεί η 'απόσταση' μεταξύ των δύο κόμβων που έχουν δημιουργηθεί έπειτα από το διαχωρισμό.

2.3.2.3.3. Setting a threshold

Μία διαφορετική μέθοδος, όπως παρουσιάζεται από (Dean, 2007), προκειμένου το δέντρο να σταματήσει να αναπτύσσεται και να καταλήξει σε κάποιον καταληκτικό κόμβο, μπορεί να πραγματοποιηθεί θέτοντας ένα '*threshold' β.* Κάποιος κόμβος *l* θεωρείται ως τερματικός κόμβος του δέντρου επιβίωσης στην περίπτωση που η τιμή του impurity που μεγιστοποιείται από τον διαχωρισμό στον κόμβο είναι μικρότερη από το '*threshold' β* που έχουμε θέσει ($\Delta I(l) < \beta$). Επίσης ένας άλλος τρόπος για να καταλήξουμε σε κάποιον καταληκτικό κόμβο *l*, είναι θέτοντας κάποιο όριο σχετικά με τον αριθμό των παρατηρήσεων που θα πρέπει να περιέχονται μέσα σε έναν κόμβο. Η συγκεκριμένη μέθοδος μπορεί να αντιμετωπίζει πολλά προβλήματα παρόλα αυτά, όπως παραδείγματος χάριν το *threshold β* που έχουμε θέσει να είναι πολύ μικρό ή πολύ μεγάλο με αποτέλεσμα ο αλγόριθμος να σταματήσει αρκετά νωρίς και να μην επιτρέπει περαιτέρω σημαντικούς διαχωρισμούς στους κόμβους (Dean, 2007).

2.3.2.3.4. Segal

Σύμφωνα με τον (Segal, 1988)προτάθηκε ένα διαφορετικό κριτήριο προκειμένου να γίνει ο διαχωρισμός στον κόμβο. Τα περισσότερα κριτήρια διαχωρισμού βασίζονται στην ομογένεια που υπάρχει στο εσωτερικό του κόμβου και ο βέλτιστος διαχωρισμός που επιλέγεται να γίνει θα πρέπει να μεγιστοποιεί την διαφορά στο *impurity*. Το καινούριο κριτήριο βασίζεται σε μετρήσεις μεταξύ των κόμβων προκειμένου να γίνει ο διαχωρισμός (*between node separation*) που να μπορούν να εφαρμοστούν στην ανάλυση επιβίωσης. Η ανάλυση δεδομένων επιβίωσης, γνωρίζοντας ότι ορισμένες παρατηρήσεις χαρακτηρίζονται από αποκοπή μας αποτρέπει την χρησιμοποίηση του μέσου τετραγωνικού σφάλματος (*MSE*), το οποίο χρησιμοποιείται ως βασικό κριτήριο διαχωρισμού στα δέντρα παλινδρόμησης, προκειμένου να μπορέσουμε να μετρήσουμε την ομοιογένεια στο εσωτερικό του κόμβου. Η αδυναμία εκτίμησης συνάρτησης κόστους (*loss function*) λόγω των αποκομμένων παρατηρήσεων που υπάρχουν στο εσωτερικό του κόμβου δεν μας συνιστά την χρησιμοποίηση του *MSE*. Η χρήση του *log rank test* προκειμένου να γίνει ο διαχωρισμός των κόμβων αποτελεί ένα ιδιαίτερα αποτελεσματικό κριτήριο στα δέντρα επιβίωσης για διάφορους λόγους. Αρχικά τα αποτελέσματα που παρέχονται από την χρήση του *log rank* μένουν αναλλοίωτα σε μονοτονικές μεταβολές της εξαρτημένης μεταβλητής. Για αυτό τον λόγο το *logrank test* δεν παρουσιάζει ευαισθησία στην παρουσία ακραίων τιμών (*outliers*) για την εξαρτημένη αλλά και τις επεξηγηματικές μεταβλητές του μοντέλου. Επίσης , εφικτός υπολογισμός των *logrank statistics* αλλά και η εύκολη προσαρμογή της αποκοπής μέσω της χρήσης του *logrank*, αποτελούν άλλους δύο εξίσου σημαντικούς λόγους. Ο '*Segal'* πρότεινε την χρήση οποιασδήποτε στάθμισης του *logrank statistics*, τονίζοντας ότι τα δέντρα που θα δημιουργηθούν είναι παρόμοια για τα περισσότερα βάρη (*weights*) που χρησιμοποιήθηκαν (Alvarez-Iglesias, 2012). Με τις κατάλληλες επιλογές βαρών, σύμφωνα με τους *statistics* (Bou-Hamad, Larocque, & Ben-Ameur, 2011) , καταλήγουμε σε πολλούς γνωστούς στατιστικούς ελέγχους (*test statistics*) όπως *logrank* και το *Wilcoxon–Gehan statistics*.

2.3.2.4. Pruning

Λαμβάνοντας υπόψιν το άρθρο των (Shimokawa, Kawasaki, & Miyaoka, 2016), για την δημιουργία ενός δέντρου επιβίωσης μπορούν να χρησιμοποιηθούν δύο βασικοί μέθοδοι. Όσον αφορά την πρώτη μέθοδο μπορούμε να κατασκευάσουμε εξ 'αρχής το βέλτιστο *(optimal)* δέντρο επιβίωσης προβαίνοντας στους βέλτιστους διαχωρισμούς μεταξύ των κόμβων του δέντρου από την επιλογή των κατάλληλων μεταβλητών και έπειτα θέτοντας κάποιο κριτήριο, προκειμένου να σταματήσουμε στο σημείο που θα έχουμε αναπτύξει το βέλτιστο δέντρο. Το κριτήριο που θα μπορούσαμε να χρησιμοποιήσουμε ίσως βασίζεται σε κάποιον στατιστικό έλεγχο όπως *AIC* ή *BIC*. Όσον αφορά την δεύτερη μέθοδο, σύμφωνα με τους (Bou-Hamad, 2009), κατά την οποία

κατασκευάζουμε ένα μεγάλο σε βάθος δέντρο, έτσι ώστε να μην μπορεί να γίνει επιπλέον διαχωρισμός στον κάθε κόμβο. Στο επόμενο στάδιο θα πρέπει να 'κουρέψουμε' (*pruning*) το δέντρο όπου έχει αναπτυχθεί δημιουργώντας ένα σύνολο από μικρότερα δέντρα υποψήφια για το βέλτιστο δέντρο που θα χρησιμοποιηθεί στο μοντέλο μας. Το τελικό βέλτιστο δέντρο από το σύνολο των δέντρων που έχουν δημιουργηθεί θα επιλεχθεί με μία μέθοδο '*validation'*. Τόσο για την πρώτη όσο και για την δεύτερη μέθοδο, θα πρέπει εξαρχής να καθοριστεί κάποιος κανόνας διαχωρισμού *(splitting rule)* για την κατασκευή του δέντρου. Προκειμένου να μην καταλήξουμε σε κάποιο δέντρο το οποίο δεν θα είναι το βέλτιστο για το μοντέλο μας θέτοντας κάποιο '*threshold* β*'*, μπορούμε να χρησιμοποιήσουμε την δεύτερη μέθοδο όσον αφορά το *'pruning'* του αρχικού δέντρου.

Η θεώρηση του *'pruning'* έχει κατά κύριο λόγο δύο πολύ σημαντικά μέτρα που θα πρέπει να υπολογιστούν, το '*cost-complexity'* και το '*split-complexity'* του δέντρου. Σύμφωνα με τους (Bou-Hamad, Larocque, & Ben-Ameur, 2011), η '*cost-complexity'* του δέντρου μπορεί να υπολογιστεί από τον τύπο:

$$R_a(T) = \sum_{h \in L(T)} R(h) + a\,|L(T)| \quad (2.35)$$

Όπου για την δημιουργία του αρχικού δέντρου *T*, θα ορίσουμε ως $L(T)$ το σύνολο των καταληκτικών κόμβων του δέντρου *T*. Για κάποιον κόμβο $h \in T$, ως $R(h)$ ορίζεται ως ο εσωτερικός 'κίνδυνος' (*within–node risk*) για τον συγκεκριμένο κόμβο *h*, δηλαδή το impurity στο εσωτερικό του κόμβου. Σύμφωνα με τους (De Rose & Pallara, 1997),το α είναι μία μη αρνητική παράμετρος και προσδιορίζει το '*cost-complexity*' μέτρο της $R_a(T)$.Η '*cost-complexity*' παράμετρος θέτει ένα πρόστιμο (*penalty*) για την πολυπλοκότητα του δέντρου, δηλαδή για τον συνολικό αριθμό καταληκτικών κόμβων που έχει για τον προσδιορισμό των δεδομένων. Στην περίπτωση που έχουμε μικρό α τότε το πρόστιμο για την πολυπλοκότητα του δέντρου είναι επίσης μικρό οπότε το μικρότερο δέντρο (*subtree*) που θα δημιουργηθεί από το αρχικό θα είναι και αυτό μεγάλο, καθώς λίγοι είναι οι καταληκτικοί κόμβοι που θα αφαιρεθούν. Αντίθετα, όσο το α αυξάνεται το subtree θα έχει όλο και λιγότερους καταληκτικούς κόμβους, με αποτέλεσμα να μικραίνει σε μέγεθος (Dean, 2007).

Σύμφωνα με τους (LeBLANC & CROWLEY, 1993) για να προσδιοριστεί το impurity $R(h)$ στο εσωτερικό του κόμβου χρησιμοποιήθηκε η απόκλιση για τις παρατηρήσεις που ανήκουν στον συγκεκριμένο κόμβο και δίνεται από τον τύπο:

$$R(h) = 2\left(LL_h(saturated) - LL_h\left(\hat{\theta}_h\right)\right) \quad (2.36)$$

Όπου $LL_h(saturated)$ είναι το log-likelihood για το κορεσμένο (saturated) μοντέλο, έχοντας μία μεταβλητή για την κάθε παρατήρηση και το $LL_h(\hat{\theta}_h)$ είναι το μεγιστοποιημένο log-likelihood για το προσαρμοσμένο μοντέλο.

Όταν το *cost-complexity* υπολογιστεί τότε μπορεί να χρησιμοποιηθεί ο αλγόριθμός '*CART'* για να γίνει το *pruning* και να βρεθούν τα βέλτιστα '*subtrees'*. Κάθε *subtree* είναι βέλτιστο για ένα διάστημα τιμών του *α*.

Η άλλη μέθοδος όπου αναπτύχθηκε προσδιορίζει το '*split-complexity'* του δέντρου και δίνεται από τον τύπο:

$$G_a(T) = \sum_{h \in W(T)} G(h) - a\,|W(T)| \quad (2.37)$$

Όπου $G(h)$ ορίζεται η τιμή του κριτήριου που χρησιμοποιήθηκε προκειμένου να γίνει ο διαχωρισμός στον συγκεκριμένο κόμβο h. Ως $W(T)$ ορίζεται το σύνολο των κόμβων του δέντρου πλην των καταληκτικών. Όπως υποστηρίζεται από τους (Bou-Hamad, Larocque, & Ben-Ameur, 2011) ,σε περίπτωση που το *α* είναι μικρό τότε το κόστος για έναν μεγάλο αριθμό διαχωρισμών (*splits*) είναι μικρό και το subtree θα είναι μεγάλο σε μέγεθος. Όσο μεγαλύτερο είναι το *α*, το subtree θα μικραίνει έχοντας λιγότερους καταληκτικούς κόμβους, έως ότου ελαχιστοποιηθεί, έχοντας αποκλειστικά μόνο τον αρχικό κόμβο. Το $G(h)$ για το δέντρο T δίνεται από τον τύπο:

$$G(T) = \sum_{h \in W(T)} G(h) \quad (2.38)$$

Ως $G(T)$,σύμφωνα με τους (LeBLANC & CROWLEY, 1993), ορίζεται το *goodness of split* του δέντρου *T*. Παρομοίως με την cost-complexity, για κάποια τιμή α υπάρχει ένα μοναδικό subtree που μεγιστοποιεί την split-complexity $G_a(T)$.

Αφού ένας μεγάλος αριθμός από μικρότερα δέντρα *(subtrees) ($T_0, T_1, \ldots, T_N$)* έχει δημιουργηθεί από το αρχικό δέντρο *T*, έφτασε η στιγμή επιλογής του βέλτιστου δέντρου ανάμεσα στο σύνολο των *subtrees*. Μερικές από τις πιο δημοφιλείς μεθόδους που μπορούμε να χρησιμοποιήσουμε είναι η *cross–validation* και η *AIC/BIC*. Η κλασσική μέθοδος *CART* χρησιμοποιεί *cross-validation* προκειμένου να εκτιμήσει την τιμή του α για το *cost-complexity*, έτσι ώστε να επιλεγεί το βέλτιστο *subtree*. Όπως ορίζεται από τους (Bou-Hamad, Larocque, & Ben-Ameur, 2011) από το σύνολο των *subtrees* που έχουν αναπαραχθεί από το αρχικό δέντρο, θα επιλεγεί το βέλτιστο '*subtree*'με την επιλογή του οποίου θα ελαχιστοποιείται το κριτήριο:

$$-2ll(T) + a|L(T)| \quad (2.39)$$

Όπου $ll(T)$ ορίζεται ως το *log-likelihood* για το δέντρο *T*, με το *α* να παίρνει είτε την τιμή 2 *(AIC)* είτε την τιμή '*log(n)' (BIC).* Η όλη διαδικασία βασίζεται στην δημιουργία ενός αρχικού δέντρου, αποκτώντας subtrees από το training sample και υπολογίζοντας ξανά στην συνέχεια το $ll(T)$ στο test sample.

2.4. Μέτρα αξιολόγησης υποδείγματος

2.4.1. Δείκτης Concordance

Ο δείκτης *concordance* ή *c-index,* όπως αλλιώς ονομάζεται, αποτελεί μία μονάδα μέτρησης ελέγχοντας τις προβλέψεις που αναπαράχθηκαν από το μοντέλο, σε σχέση με τις πραγματικές τιμές των παρατηρήσεων. Όπως ορίζουν οι (Brentnall & Cuzick, 2018),ο σκοπός δημιουργίας του δείκτη *concordance* είναι η εκτίμηση του βαθμού όπου μία τυχαία επιλεγμένη παρατήρηση η οποία ακολουθεί κάποια συγκεκριμένη κατανομή είναι μεγαλύτερη από κάποια άλλη τυχαία επιλεγμένη παρατήρηση η οποία ακολουθεί διαφορετική κατανομή .Όπως μας πληροφορούν οι (Therneau & Atkinson, 2020), η χρήση του δείκτη '*concordance statistic'* για τα '*Cox'* μοντέλα χρησιμοποιήθηκε για πρώτη φορά από τον '*Harrell'* και αυτή την στιγμή αποτελεί έναν από τους πιο βασικούς δείκτες που χρησιμοποιούνται για να μετρηθεί το '*goodness-of-fit'* στα μοντέλα επιβίωσης. Σύμφωνα με τους (Raykar, Steck, Krishnapuram, Dehing-Oberije, & Lambin, 2007), υποδηλώνεται ως το κλάσμα όλων των παρατηρήσεων των οποίων ο χρόνος επιβίωσης είναι σωστά ταξινομημένος προς όλες εκείνες τις παρατηρήσεις οι οποίες μπορούν να ταξινομηθούν. Διαφορετικά , απεικονίζει την πιθανότητα συμφωνίας (*concordance*) που υπάρχει μεταξύ των παρατηρήσεων που έχει προβλεφθεί η επιβίωση σε σύγκριση με την πραγματική τιμή επιβίωσης τους. Ο δείκτης *concordance* δεν εξετάζει την κάθε παρατήρηση ξεχωριστά δίχως να δίνει βαρύτητα στην τιμή της επιβίωσης που έχει λάβει κατά την πρόβλεψη (Sliva, 2019) . Σύμφωνα με τον (Sliva, 2019) επικεντρώνεται κυρίως στην ταξινόμηση των προβλέψεων (*order of the predictions*), δηλαδή στην κατεύθυνση κατά την οποία έχουν προβλεφθεί οι παρατηρήσεις ότι εκδήλωσαν το γεγονός του ενδιαφέροντος σε σχέση με την πραγματική σειρά τους όσον αφορά την εκδήλωση του γεγονότος.

Η ύπαρξη αποκοπής μεταξύ των δεδομένων στην ανάλυση επιβίωσης δυσχεραίνει την λειτουργία του δείκτη *concordance*. Από την στιγμή που δεν υπάρχει η δυνατότητα ταξινόμησης τους με βεβαιότητα, η σύγκριση τους με τις υπόλοιπες προβλεπόμενες τιμές για τον υπολογισμού του *concordance* καθίσταται αδύνατη (Therneau & Atkinson, 2020). Για παράδειγμα στην περίπτωση που κάποια παρατήρηση *i* από το δείγμα παρατηρήθηκε αποκομμένη την χρονική στιγμή *t=5* και κάποια άλλη παρατήρηση *j* του δείγματος εκδήλωσε το γεγονός του ενδιαφέροντος την χρονική

στιγμή *t = 10*, τότε για την παρατήρηση *i* δεν μπορούμε να είμαστε σίγουροι αν εκδήλωσε ή όχι το γεγονός του ενδιαφέροντος μετά την παρατήρηση *j*. Η ταξινόμηση σε αυτή την περίπτωση δεν είναι δυνατόν να συμβεί οπότε η αποκομμένη παρατήρηση δεν μπορεί να συμπεριληφθεί για τον υπολογισμό του '*concordance'*.

Ο δείκτης '*concordance'* αποτελεί μία γενίκευση της περιοχής κάτω από την καμπύλη *ROC* και μπορεί να υπολογιστεί από τον ακόλουθο τύπο (PySurvival, 2019) :

$$C - index = \frac{\sum_{i,j} 1_{T_j<T_i} \times 1_{n_j>n_i} \times \delta_j}{\sum_{i,j} 1_{T_j<T_i} \times \delta_j} \quad (2.40)$$

Όπου n_i δηλώνει τον κίνδυνο για την παρατήρηση *i*. Ως $1_{T_j<T_i}$ εκφράζει τον χρόνο που η παρατήρηση *j* εκδήλωσε νωρίτερα το γεγονός του ενδιαφέροντος σε σχέση με την παρατήρηση *i* . Στην περίπτωση που η *j* εκδήλωσε νωρίτερα το γεγονός του ενδιαφέροντος τότε $1_{T_j<T_i}$ = 1 διαφορετικά $1_{T_j<T_i}$ = 0. Ως $1_{n_j>n_i}$ εκδηλώνεται ο κίνδυνος της *j* να εκδηλώσει το γεγονός του ενδιαφέροντος σε σχέση με την *i*. Στην περίπτωση που ο κίνδυνος της j να εκδηλώσει το γεγονός είναι μεγαλύτερος τότε $1_{n_j>n_i}$ = 1 ,διαφορετικά $1_{n_j>n_i}$ = 0.

Όταν ο δείκτης C-index βρίσκεται σε περιοχή κοντά στο 1 σημαίνει πως το μοντέλο μας είναι αρκετά αποδοτικό, προβαίνοντας σε καλές εκτιμήσεις για το σύνολο δεδομένων που χρησιμοποιούμε. Αντίθετα δείκτης C-index κοντά στο 0.5 σημαίνει πως το μοντέλο δεν είναι αποτελεσματικό.

2.4.2. Out-of-Bag σφάλμα πρόβλεψης

Μία διαφορετική μέθοδος ,η οποία χρησιμοποιείται κατά κύριο λόγο με σκοπό την μέτρηση του σφάλματος πρόβλεψης σε μοντέλα *Random-Forests*, είναι το *Out-of-Bag(OOB)* σφάλμα πρόβλεψης. Σύμφωνα με (Swiss Federal Institute of Technology Zurich, 2012), η βασική διαφορά της συγκεκριμένης μεθόδου με τον δείκτη *concordance*, οφείλεται στο γεγονός ότι ο *C-index* απαιτεί κάποιο προβλεπόμενο αποτέλεσμα προκειμένου να υπολογιστεί. Η μέθοδος *OOB* θα μπορούσαμε να πούμε πως παρουσιάζει μεγάλη ομοιότητα με την *cross-validation*, δίχως να χρειάζεται κάποιο επιπλέον υπολογιστικό βάρος. Χρησιμοποιείται κυρίως σε περιπτώσεις που το αρχικό σύνολο δεδομένων είναι αρκετά μικρό, με αποτέλεσμα ο διαχωρισμός του δείγματος σε *training* και *test set* να μην μπορεί να πραγματοποιηθεί.

Σύμφωνα με τον (Kunchhal, 2020), η ιδέα πίσω από το *OOB* είναι η δημιουργία πολλών διαφορετικών υποσυνόλων που θα έχουν το ίδιο μέγεθος με το αρχικό σύνολο δεδομένων *(bootstraps)* και την τυχαία μετακίνηση κάποιας παρατήρησης από το αρχικό σύνολο δεδομένων στο πρώτο υποσύνολο *(bootstrap)* που έχουν δημιουργηθεί, προϋποθέτοντας πως η παρατήρηση θα συνεχίσει να υπάρχει στο αρχικό σύνολο δεδομένων έπειτα από την μετακίνηση. Η ίδια

διαδικασία θα συνεχιστεί έως ότου συμπληρωθούν με παρατηρήσεις όλα τα υποσύνολα *(bootstraps).* Από την τυχαία επιλογή περίπου το 37% των παρατηρήσεων του αρχικού υποσυνόλου δεν περιέχεται στο κάθε *bootstrap.* Για το κάθε *bootstrap* μπορεί να υπάρχει κάποια παρατήρηση περισσότερες από μία φορές με αποτέλεσμα να εκλείπουν ορισμένες παρατηρήσεις από τα υποσύνολα σε σύγκριση με το αρχικό σύνολο δεδομένων. Οι παρατηρήσεις που έχουν παραμείνει εκτός για το κάθε υποσύνολο σε σύγκριση με το αρχικό δείγμα αποτελούν το *(Out-Of-Bag Sample*) του υποσυνόλου. Στην συνέχεια θα αναπαράγουμε από ένα *survival tree* στο κάθε υποσύνολο ξεχωριστά προκειμένου να γίνει η πρόβλεψη για τις παρατηρήσεις που ανήκουν στο *(Out-Of-Bag Sample)* του υποσυνόλου.

Υποθέτοντας, σύμφωνα με (Myte, 2013), πως η $H(t|x_i)$ συμβολίζει την αθροιστική συνάρτηση κινδύνου για κάποια παρατήρηση *i*, η οποία χαρακτηρίζεται από ένα πλήθος μεταβλητών x_i και υπολογίζεται καθώς γίνονται οι διαχωρισμοί μεταξύ των κόμβων του δέντρου με τις συγκεκριμένες μεταβλητές x_i. Η αθροιστική συνάρτηση κινδύνου *CHF* για τον καταληκτικό κόμβο στον οποίο ανήκει η *i* είναι της μορφής:

$$\widehat{H}_h(t|x_i) = \widehat{H}_h(t), για\ x_i \in h \quad (2.41)$$

Για κάθε δέντρο *(b = 1,…,B)* στο υποσύνολο *(bootstrap sample)* που έχει δημιουργηθεί από το αρχικό σύνολο δεδομένων η αθροιστική συνάρτηση κινδύνου συμβολίζεται ως $\widehat{H}_b(t|x_i)$. Για την παρατήρηση *i*, ισχύει $I_{i,b} = 1$ στην περίπτωση που η παρατήρηση *i* ανήκει στα *OOB* δεδομένα για το υποσύνολο *b*, ενώ σε διαφορετική περίπτωση $I_{i,b} = 0$. Η αθροιστική συνάρτηση κινδύνου (*CHF*) όλων των υποσυνόλων *'bootstrap samples'*, όπου η παρατήρηση *i* συγκαταλέγεται στα *OOB* δεδομένα τους, έχει την μορφή:

$$\widehat{H}_E^*(t|x_i) = \frac{1}{\sum_{b=1}^{B} I_{i,b}} \sum_{b=1}^{B} I_{i,b} \widehat{H}_b(t|x_i) \quad (2.42)$$

Η απόκλιση της αθροιστικής συνάρτησης κινδύνου $\widehat{H}_h(t|x_i)$ κάποιας παρατήρησης *i* σε σύγκριση με την αθροιστική συνάρτηση κινδύνου για την παρατήρηση *i* που ανήκει στο *OOB* δείγμα $\widehat{H}_E^*(t|x_i)$, εκφράζει το *OOB* σφάλμα πρόβλεψης.

3. Περιγραφή των δεδομένων

3.1. Περιγραφή Προβλήματος

Η προσπάθεια μοντελοποίησης του κύκλου ζωής των εργαζομένων που απασχολούνται εκ μέρους κάποιας εταιρίας αποτελεί καίριο ζήτημα όσον αφορά την ανάπτυξη της. Η επίλυση του προβλήματος σχετικά με την αποχώρηση του εργαζομένου θα μπορούσε να αποφέρει πολλά οφέλη για την εταιρία τόσο στην μείωση των δαπανών, αφού δεν θα χρειαστεί να προσλάβει εκ νέου προσωπικό για να καλύψει την θέση, αλλά και γενικότερα στο τρόπο (περιβάλλον) εργασίας. Ένα άλλο πολύ σημαντικό όφελος είναι η ομαλή συνέχιση του προγράμματος εργασιών της εταιρίας, χωρίς να χρειαστεί η διακοπή τους λόγω παραίτησής κάποιου υπαλλήλου. Για την αντιμετώπιση του προβλήματος θα χρησιμοποιήσουμε το πολύτιμο σύνολο δεδομένων που μας παρέχεται από το ιστολόγιο του *Edward Babuskin*. Αφού προηγηθεί πλήρης ερμηνεία και ανάλυση των δεδομένων για καθένα από τους εργαζομένους στην εταιρία, θα προσπαθήσουμε να μοντελοποιήσουμε και να προβλέψουμε την διάρκεια του κύκλου ζωής των εργαζομένων με την χρήση ποικίλων μοντέλων που εφαρμόζονται στην ανάλυση επιβίωσης. Από τις αναλύσεις που θα ακολουθήσουν εκτιμάται κατά πόσο η μεταβολή των παραμέτρων αλλά και ποιες είναι εκείνες οι μεταβλητές που επηρεάζουν περισσότερο την παραμονή του εργαζομένου στην εταιρία. Αρχικά θα χρησιμοποιήσουμε την μη-παραμετρική μέθοδο *Kaplan-Meier* και θα προσπαθήσουμε να δούμε αν υπάρχει στατιστικά σημαντική διαφορά μεταξύ των δύο καμπυλών επιβίωσης που δημιουργήθηκαν. Στο συγκεκριμένο μοντέλο δεν έχουν συμπεριληφθεί οι επεξηγηματικές μεταβλητές του υποδείγματος. Έπειτα θα χρησιμοποιηθεί το δημοφιλές ημί-παραμετρικό μοντέλο αναλογικών κινδύνων της *Cox (Cox PH)* για την ανάλυση επιβίωσης στο σύνολο δεδομένων. Από την χρήση του μοντέλου της *Cox* θα μπορέσουμε να εκτιμήσουμε τον στιγμιαίο κίνδυνο για τον κάθε εργαζόμενο να εγκαταλείψει την θέση εργασίας την συγκεκριμένη χρονική στιγμή. Επίσης θα γίνει χρήση του εκθετικού και *Weibull* παραμετρικών μεθόδων προκειμένου να μοντελοποιήσουμε την διάρκεια του κύκλου ζωής των εργαζομένων και κατά πόσο η εφαρμογή ενός τέτοιου μοντέλου ερμηνεύει τα δεδομένα. Επιπλέον θα γίνει χρήση των δέντρων επιβίωσής, ως μίας διαφορετικής μη-παραμετρικής μεθόδου που χρησιμοποιείται στην ανάλυση επιβίωσης προκειμένου να προβλεφθεί η διάρκεια του κύκλου ζωής ενός εργαζομένου. Σε τελικό στάδιο θα πραγματοποιηθεί σύγκριση μεταξύ των μοντέλων του υποδείγματος, προκειμένου να επιλεχθεί εκείνο που επιφέρει μεγαλύτερη ακρίβεια στις προβλέψεις.

3.2. Παρουσίαση Δεδομένων

Το σύνολο δεδομένων το οποίο θα χρησιμοποιήσουμε προκειμένου να εκπαιδεύσουμε τα μοντέλα μας στην ανάλυση επιβίωσης και προχωρώντας στις προβλέψεις αργότερα, καταρτίζεται

συνολικά από 16 μεταβλητές και 1129 παρατηρήσεις. Οι 9 από τις 16 χαρακτηρίζονται ως ποιοτικές μεταβλητές (*qualitative*) καθώς αναφέρονται σε κάποιο ποιοτικό χαρακτηριστικό, διαχωρίζοντας τις παρατηρήσεις σε τουλάχιστον δύο κλάσεις. Ενώ οι υπόλοιπες 7 χαρακτηρίζονται ως ποσοτικές μεταβλητές (*quantitative*) καθώς λαμβάνουν τιμές που έχουν αριθμητικές ιδιότητες και εκφράζονται από συγκεκριμένη μονάδα μέτρησης. Επίσης, παρατηρούμε πως στο συγκεκριμένο σύνολο δεδομένων μας υπάρχει τιμή για όλες τις μεταβλητές κάθε μίας από τις 1129 παρατηρήσεις. Όποτε δεν θα χρειαστεί να ασχοληθούμε με το πρόβλημα της απώλειας τιμής (*missing value*) για κάποια από τις παρατηρήσεις μας.

Προτού προχωρήσουμε στην ανάλυση επιβίωσης, απαραίτητη προϋπόθεση αποτελεί η παρουσίαση και ερμηνεία των 16 μεταβλητών που μας παρέχονται από το σύνολο δεδομένων. Επομένως παρατηρούμε τα εξής:

- Η πρώτη μεταβλητή στο δείγμα είναι το '*stag*' η οποία αναφέρεται στον χρόνο παραμονής του εργαζομένου στην εταιρία. Δηλώνει δηλαδή το χρονικό διάστημα από την στιγμή που εισήλθε στην μελέτη μέχρι την αποχώρηση του από την εταιρία (εκτός από την περίπτωση, των εργαζομένων που χαρακτηρίζονται από αποκοπή). Οριοθετώντας το '*stag*' στην ανάλυση επιβίωσης θα μπορούσαμε να πούμε πως εκφράζει τον 'χρόνο επιβίωσης' (*survival time*) για τον κάθε εργαζόμενο μέχρι την στιγμή που θα εκδηλωθεί το γεγονός του ενδιαφέροντος, δηλαδή την αποδέσμευση του από τα καθήκοντα της εταιρίας.
- Δεύτερη πολύ σημαντική μεταβλητή του συνόλου δεδομένων είναι το '*event*', η οποία υποδηλώνει τις αποκομμένες από εκείνες τις παρατηρήσεις που έχουν εκδηλώσει το γεγονός του ενδιαφέροντος. Στην προκειμένη περίπτωση ,υποθέτουμε ότι κάποιος εργαζόμενος εκδήλωσε το γεγονός του ενδιαφέροντος εφόσον αποχώρησε από την εταιρία κατά την διάρκεια της μελέτης. Αντίθετα, για κάποιον εργαζόμενο όπου σταμάτησε η παρακολούθηση του, για κάποιον λόγο που δεν μπορούμε να γνωρίζουμε, την ορισμένη χρονική στιγμή ή δεν αποχώρησε από την εταιρία μέχρι και την λήξη της μελέτης μπορούμε να πούμε πως χαρακτηρίζεται από αποκοπή. Οι εργαζόμενοι οι οποίοι χαρακτηρίζονται από αποκοπή λαμβάνουν τιμή για την μεταβλητή '*event = 0*'. Αντίθετα οι εργαζόμενοι οι οποίοι έχουν 'αποτύχει' λαμβάνουν τιμή '*event = 1*'. Αξίζει να σημειωθεί πως όλες οι παρατηρήσεις στο δείγμα που χαρακτηρίζονται από αποκοπή είναι δεξιά αποκομμένες (*right censoring*). Οφείλεται στο γεγονός ότι το πότε και αν ο εργαζόμενος εκδήλωσε το γεγονός του ενδιαφέροντος το οποίο δεν γνωρίζουμε με ακρίβεια, το συναντάμε στην δεξιά πλευρά της μελέτης μας. Εκτενέστερα θα λέγαμε πως έχουμε τυχαία αποκοπή τύπου I (*random censoring*), καθώς κάποιος εργαζόμενος μπορεί να αποκόπηκε πριν αλλά και μετά την λήξη της μελέτης, μη γνωρίζοντας την ακριβή χρονική στιγμή.

- Έπειτα μέσω της επεξηγηματικής ποιοτικής μεταβλητής '*gender*' προσδιορίζονται τα δύο φύλα των εργαζομένων.
- Η ποσοτική επεξηγηματική μεταβλητή '*age*' υποδηλώνει την ηλικία για τον κάθε εργαζόμενο που απασχολεί η εταιρία.
- Μέσω της ποιοτικής μεταβλητής '*industry*' δηλώνεται ο κλάδος στον οποίο ειδικεύεται η κάθε εταιρία, στην οποία προσφέρει τις υπηρεσίες του ο κάθε εργαζόμενος. Παρατηρούμε πως υπάρχουν 15 διαφορετικοί κλάδοι στο δείγμα, στους οποίους συγκαταλέγονται οι εταιρίες.
- Η ποιοτική μεταβλητή '*profession*' δηλώνει το τμήμα όπου ανήκει, κάθε εργαζόμενος στην εταιρία. Διακρίνουμε 14 διαφορετικά τμήματα στα οποία μπορεί να εντάσσεται ο εργαζόμενος.
- Η ποιοτική μεταβλητή '*traffic*' δηλώνει τον 'τρόπο' ο οποίος χρησιμοποιήθηκε προκειμένου να προσεγγίσει ο εργαζόμενος την εταιρία ή η εταιρία τον εργαζόμενο με σκοπό την πρόσληψη του. Παρατηρούμε πως διαχωρίζει τους εργαζομένους μεταξύ 8 κλάσεων, ανάλογα με τον 'τρόπο' που εφαρμόστηκε. Για την καλύτερη ερμηνεία των κατηγοριών της μεταβλητής '*traffic*', παρέχεται παρακάτω μία σύντομη επεξήγηση της καθεμίας κλάση. Αρχικά, η κλάση '*referral*' υποδηλώνει το σύνολο των εργαζομένων των οποίων η πρόσληψη έγινε μέσω συστάσεων. Οι συστάσεις στην προκειμένη περίπτωση πραγματοποιούνται από εργαζομένους που προϋπάρχουνε στη εταιρία, με σκοπό να προτείνουν νέους υποψηφίους οι οποίοι θα μπορούσαν να είναι κατάλληλοι για την θέση εργασίας. Η κατηγορία '*friends*' υποδηλώνει τις προσλήψεις εργαζομένων εξαιτίας των φιλικών σχέσεων που έχουν με κάποια από τα υψηλόβαθμα στελέχη της εταιρίας. Η κατηγορία '*advert*' δηλώνει τους εργαζόμενους οι οποίοι ενημερώθηκαν, εκδηλώνοντας το ενδιαφέρον τους για την θέση εργασίας λόγω κάποιας διαφήμισης, προχωρώντας στην πρόσληψη τους αργότερα. Η κατηγορία '*recNErab*' υποδηλώνει την άμεση επαφή του εργαζομένου με την εταιρία έπειτα από σύσταση που έχει γίνει για την κάλυψη της θέσεως εργασίας. Η σύσταση σε αυτή την περίπτωση γίνεται από φιλικό πρόσωπο του εργαζομένου, το οποίο δεν εργάζεται στην συγκεκριμένη εταιρία. Η κατηγορία '*youjs*' δηλώνει το ποσοστό εργαζομένων που κατάφερε να προσεγγίσει η εταιρία, δημοσιεύοντας την θέση σε κάποια ιστοσελίδα ευρέσεως εργασίας. Η κατηγορία '*KA*' δηλώνει το ποσοστό των εργαζομένων που η υπηρεσία πρόσληψης (*recruiting agency*) τους έφερε σε άμεση επαφή με τον εργοδότη, με σκοπό την κάλυψη της θέσης εργασίας. Η κατηγορία '*rabrecNErab*' δηλώνει τους εργαζόμενους με τους οποίους ο εργοδότης ήρθε σε επικοινωνία μαζί τους ύστερα από πρόταση γνωστού ατόμου του εργαζομένου. Ο γνωστός του εργαζομένου πρότεινε στον εργοδότη κάποιον που θεωρεί κατάλληλο για την συγκεκριμένη θέση εργασίας. Τέλος η κατηγορία '*empjs*' δηλώνει το ποσοστό εργαζομένων τους οποίους ο εργοδότης αναζήτησε ,μέσω του

βιογραφικού που είχαν καταχωρήσει στον ιστότοπο εργασίας, με σκοπό την πρόσληψη τους στην συγκεκριμένη θέση.

- Η ποιοτική μεταβλητή '*Coach*' εκφράζει την παρουσία 'προπονητή' κατά την διάρκεια της εργασίας του. Ως '*no*' δηλώνεται πως ο εργαζόμενος δεν χρειάζεται κάποιον προκειμένου να τον καθοδηγεί και να τον επιβλέπει. Ως '*yes*' δηλώνεται πως ο εργαζόμενος χρειάζεται κάποιον να τον καθοδηγεί και ως '*my head*' δηλώνεται πως ο εργαζόμενος θα πρέπει να έχει συνεχώς κάποιον πάνω από το κεφάλι του προκειμένου να τον επιβλέπει.
- Η ποιοτική μεταβλητή '*head_gender*' δηλώνει το φύλο του επικεφαλή ή του προϊσταμένου του εργαζομένου.
- Η ποιοτική μεταβλητή '*greywage*' δηλώνει αν ο εργαζόμενος που απασχολείται από κάποια συγκεκριμένη εταιρία πληρώνεται παραπάνω από τον ελάχιστο καθαρό του μισθό που δικαιούται (*whitewage*).
- Η ποιοτική μεταβλητή '*way*' δηλώνει το μέσο που χρησιμοποιεί ο εργαζόμενος προκειμένου να μετακινηθεί στην εταιρία. Παρατηρούμε πως κάποιος εργαζόμενος επιλέγει να μετακινηθεί στην εταιρία με 3 διαφορετικούς τρόπους. Με τα πόδια '*foot*', χρησιμοποιώντας το αυτοκίνητο του '*car*', παίρνοντας λεωφορείο προκειμένου να πάει στην δουλεία του '*bus*'.
- Η ποσοτική μεταβλητή '*extraversion*' είναι ένας δείκτης που λαμβάνει τιμές μεταξύ του διαστήματος (1-10), υπολογίζοντας τον βαθμό εξωστρέφειας και κοινωνικότητας του εργαζομένου. Τιμές του δείκτη πλησιέστερες στο 10, δηλώνουν περισσότερη εξωστρέφεια για τον εργαζόμενο.
- Η ποσοτική μεταβλητή '*independ*' είναι ένας δείκτης εξίσου από το (1-10) που υπολογίζει τον βαθμό ανεξαρτησίας του εργαζομένου. Για τιμές του δείκτη πλησιέστερες στο 10, τόσο μεγαλύτερη η ανεξαρτησία του εργαζομένου.
- Η ποσοτική μεταβλητή '*selfcontrol*' αποτελεί έναν δείκτη από το (1-10) που υπολογίζει τον αυτοέλεγχο για τον καθένα εργαζόμενο. Τιμές του δείκτη πιο κοντά στο 1, δηλώνουν λιγότερο αυτοέλεγχο για τον εργαζόμενο.
- Η ποσοτική μεταβλητή '*anxiety*' είναι ένας δείκτης από το (1-10) που υπολογίζει τον βαθμό που αγχώνεται κάποιος εργαζόμενος μέσα στην εταιρία. Για τιμές πλησιέστερα στο 10,σημαίνει πως ο εργαζόμενος αγχώνεται περισσότερο κατά την διάρκεια εργασίας του.
- Τέλος, η ποσοτική μεταβλητή '*novator*' αποτελεί έναν δείκτη από το (1-10) που υπολογίζει το πόσο αρχάριος είναι κάποιος εργαζόμενος. Όσο πλησιέστερα είναι οι τιμές του δείκτη στο 1, τόσο μεγαλύτερη εμπειρία έχει ο εργαζόμενος.

3.3. Στοιχεία Περιγραφικής Στατιστικής

Αρχικά, θα πρέπει να γίνει η απαραίτητη επεξεργασία των δεδομένων πριν προχωρήσουμε στην ανάλυση επιβίωσης. Η περιγραφική στατιστική στο σύνολο δεδομένων καθώς και η εφαρμογή αργότερα των μεθόδων ανάλυσης επιβίωσης έτσι ώστε να μπορέσουμε να μοντελοποιήσουμε την διάρκεια ζωής των εργαζομένων, θα γίνει με την χρήση της γλώσσας προγραμματισμού 'R'. Στις ποσοτικές μεταβλητές ('*gender*', '*industry*', '*profession*', '*traffic*', '*coach*', '*head gender'*, '*greywage*', '*way*', '*event*') θα πρέπει να γίνει η μετατροπή από '*character*' που ήταν αρχικά σε '*factor*' για την καλύτερη ερμηνεία των αποτελεσμάτων. Από την μετατροπή αυτή θα μπορούμε να διακρίνουμε το πλήθος των εργαζομένων που ανήκουν σε κάθε μία από τις κλάσεις των επεξηγηματικών μεταβλητών. Στο πίνακα (1) παρουσιάζονται τα αποτελέσματα από την ανάλυση όλων των μεταβλητών του υποδείγματος.

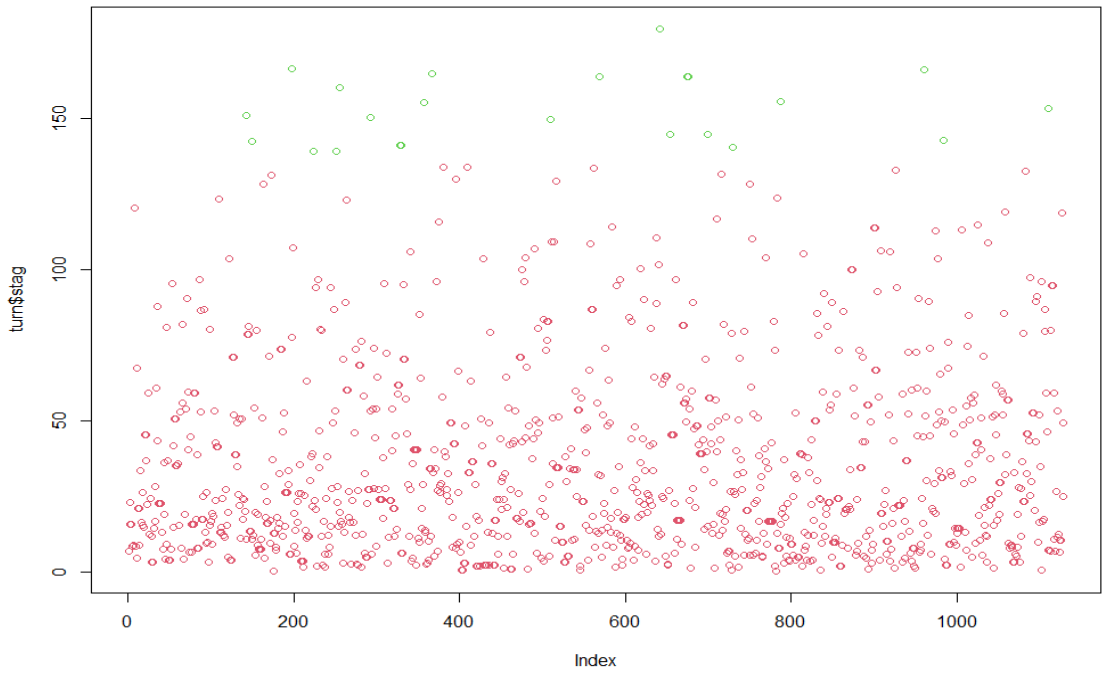
```
      stag              event    gender       age              industry
 Min.   :  0.3942   0:558   f:853   Min.   :18.00   Retail      :289
 1st Qu.: 11.7289   1:571   m:276   1st Qu.:26.00   manufacture:145
 Median : 24.3450                   Median :30.00   IT          :122
 Mean   : 36.6275                   Mean   :31.07   Banks       :114
 3rd Qu.: 51.3183                   3rd Qu.:36.00   etc         : 94
 Max.   :179.4497                   Max.   :58.00   Consult     : 74
                                                    (Other)     :291


       profession              traffic          coach
 HR                 :757   youjs      :318   my head:314
 IT                 : 74   empjs      :248   no     :683
 Sales              : 66   rabrecNErab:211   yes    :132
 etc                : 37   friends    :118
 Marketing          : 31   referal    : 95
 BusinessDevelopment: 27   KA         : 67
 (Other)            :137   (Other)    : 72


  greywage       way        extraversion      independ
 grey : 127   bus :681   Min.   : 1.000   Min.   : 1.000
 white:1002   car :331   1st Qu.: 4.600   1st Qu.: 4.100
              foot:117   Median : 5.400   Median : 5.500
                         Mean   : 5.592   Mean   : 5.478
                         3rd Qu.: 7.000   3rd Qu.: 6.900
                         Max.   :10.000   Max.   :10.000


    anxiety          novator        selfcontrol    head_gender
 Min.   : 1.700   Min.   : 1.00   Min.   : 1.000    f:545
 1st Qu.: 4.800   1st Qu.: 4.40   1st Qu.: 4.100    m:584
 Median : 5.600   Median : 6.00   Median : 5.700
 Mean   : 5.666   Mean   : 5.88   Mean   : 5.597
 3rd Qu.: 7.100   3rd Qu.: 7.50   3rd Qu.: 7.200
 Max.   :10.000   Max.   :10.00   Max.   :10.000
```

Πίνακας 1 - Αποτελέσματα ανάλυσης όλων των επεξηγηματικών μεταβλητών του δείγματος

3.3.1. Ποσοτικές μεταβλητές

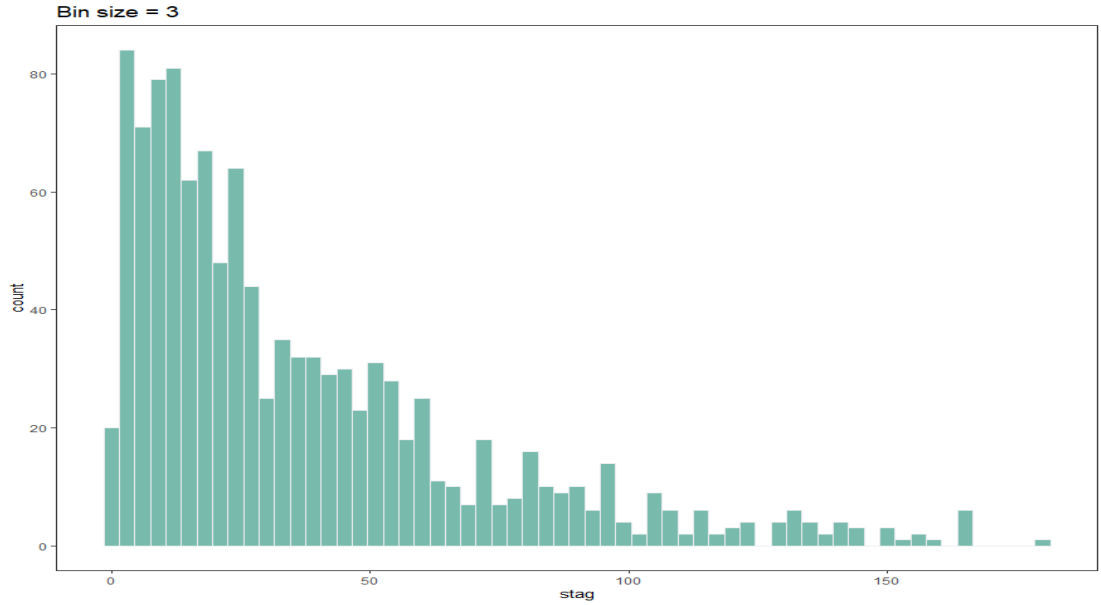
3.3.1.1. Στατιστικά γραφήματα ποσοτικών μεταβλητών

Παρατηρούμε πως η μεταβλητή '*stag*' (χρόνος επιβίωσης του εργαζομένου) λαμβάνει τιμές μεταξύ του διαστήματος (0.3942, 179.4497). Δηλαδή η μικρότερη χρονική στιγμή που εγκατέλειψε κάποιος εργαζόμενος την εταιρία ή αποκόπηκε από την μελέτη ,είναι οι 39 ημέρες, ενώ οι 179 μήνες αποτελούν το μεγαλύτερο χρονικό διάστημα που παρέμεινε ο εργαζόμενος στην εταιρία η αποκόπηκε εκείνη την χρονική περίοδο. Επίσης για την μεταβλητή '*stag*' η μέση τιμή και ο διάμεσος έχει τιμή κοντά στους 36 και 25 μήνες αντίστοιχα, η τιμή του πρώτου τεταρτημόριου είναι οι 11 μήνες ενώ η τιμή του τρίτου τεταρτημόριου είναι κοντά στους 50 μήνες. Τέλος στο διάγραμμα (4) διακρίνουμε 23 ακραίες τιμές (*outliers*) όπου μπορεί να λάβει η μεταβλητή '*stag*'.



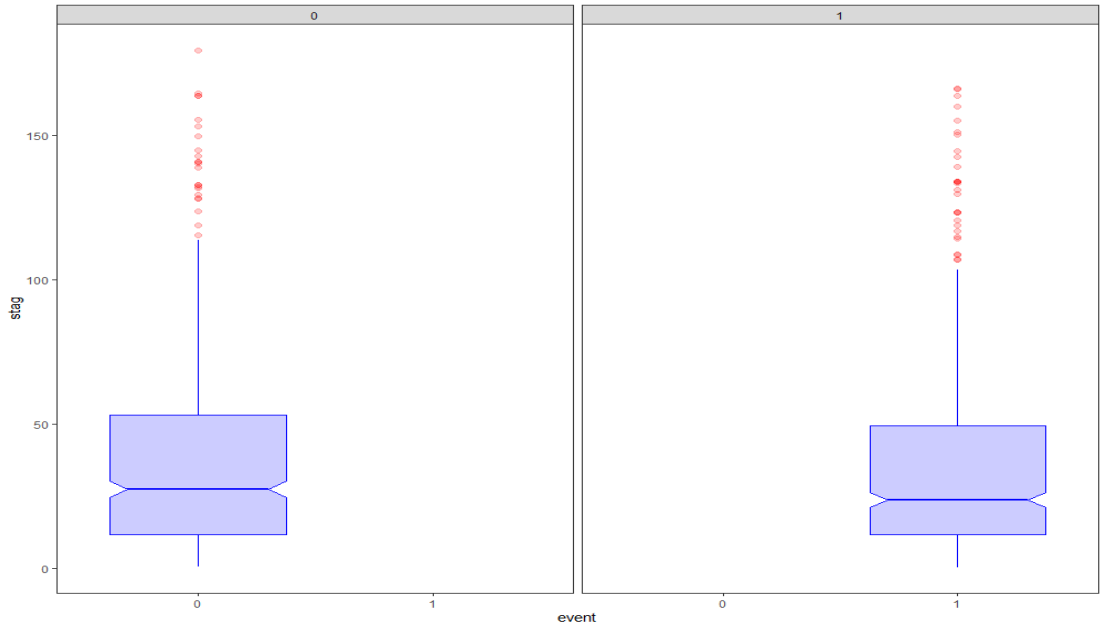
Διάγραμμα 4 - Σχήμα ακραίων τιμών της μεταβλητής 'stag' (οι πράσινες κουκίδες δηλώνουν τις ακραίες τιμές)

Παρομοίως σύμφωνα με το διάγραμμα (5),φαίνεται πως είναι πιθανή η ύπαρξη ορισμένων ακραίων τιμών (*outliers*) για την μεταβλητή '*stag*', οι οποίες εμφανίζονται προς το τέλος του ιστογράμματος. Επίσης, παρατηρούμε ο χρόνος επιβίωσης για τους περισσότερους εργαζομένους κυμαίνεται μεταξύ του διαστήματος (0,50).



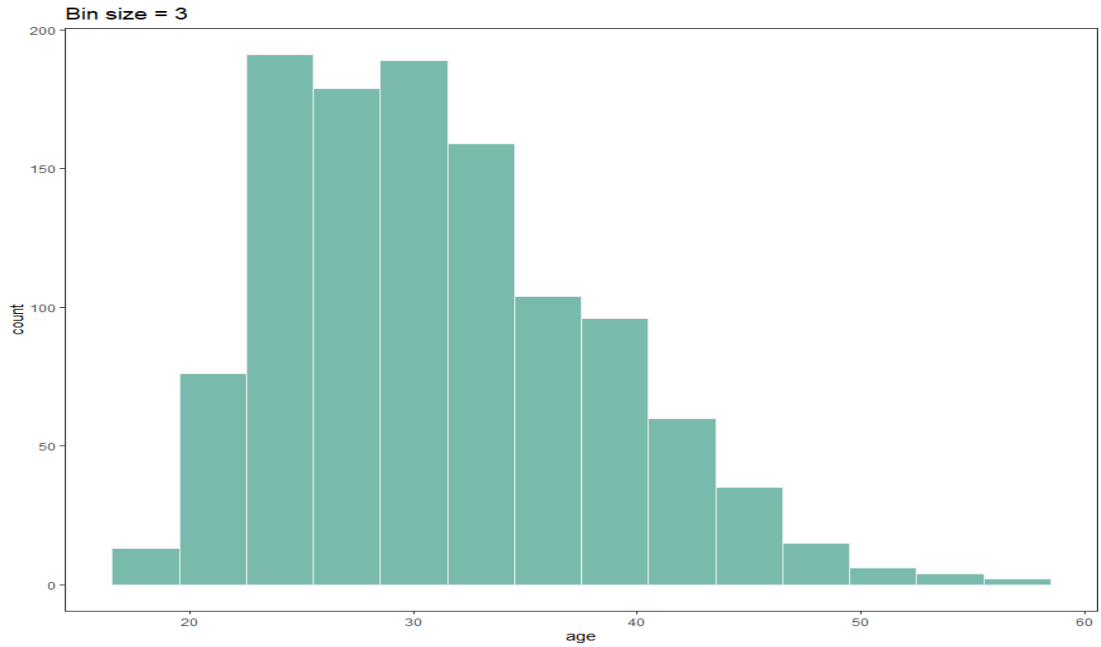
Διάγραμμα 5 - Ιστόγραμμα της μεταβλητής 'stag'

Διαχωρίζοντας τους εργαζομένους σε εκείνους που έχουν ‘αποτύχει’ και σε εκείνους που έχουν αποκοπεί από την μελέτη ως προς τον χρόνο επιβίωσης τους, παρατηρούμε πώς δεν υπάρχει μεγάλη διαφορά μεταξύ των διαμέσων για τις δύο κατηγορίες. Όπως βλέπουμε παρακάτω στο διάγραμμα (6), τα δύο θηκογράμματα που δημιουργήθηκαν φαίνεται να εμφανίζουν μικρές διαφορές ως προς τον διάμεσο για τις δύο κατηγορίες. Θα μπορούσαμε να πούμε ότι οι εργαζόμενοι που χαρακτηρίζονται από αποκοπή έχουν μεγαλύτερο χρόνο επιβίωσης από εκείνους που εκδήλωσαν το γεγονός του ενδιαφέροντος, δίχως όμως να είναι αισθητά σημαντική η διαφορά.



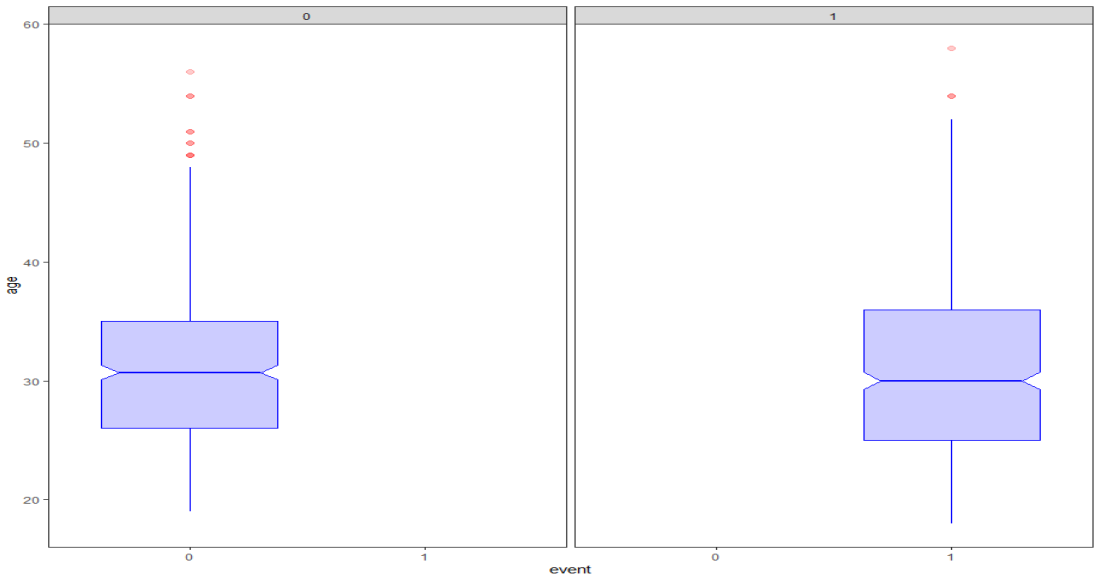
Διάγραμμα 6 - Θηκογράμματα της μεταβλητής 'stag' για 'event = 0' και 'event = 1'

Η μεταβλητή ‘*age*’ (ηλικία του εργαζομένου) λαμβάνει τιμές μεταξύ του διαστήματος (19,58). Δηλαδή οι εργαζόμενοι που απασχολήθηκαν από την εταιρία ήταν μεταξύ των ηλικιακών ομάδων από 18 έως και 58 ετών. Επίσης η μέση ηλικία των εργαζομένων που απασχολούνται από την εταιρία όπως φαίνεται και στο διάγραμμα (7) είναι κοντά στα 30 έτη. Επίσης, σύμφωνα με το διάγραμμα (3.4) φαίνεται πως ίσως υπάρχουν μερικές ακραίες τιμές και για την μεταβλητή ‘*age*’. Τέλος, παρατηρούμε ότι οι περισσότεροι εργαζόμενοι είναι μεταξύ των ηλικιακών ομάδων 20 έως και 40 ετών.



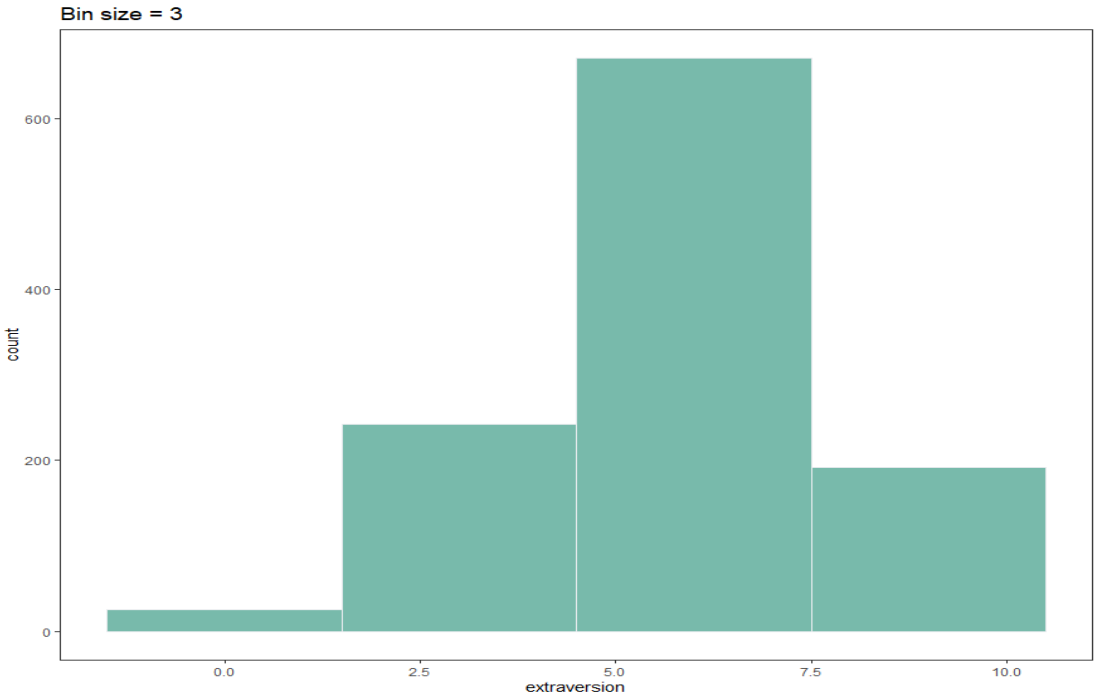
Διάγραμμα 7 - Ιστόγραμμα της μεταβλητής 'age'

Σύμφωνα με το διάγραμμα (8) δεν φαίνεται να υπάρχουν σημαντικές διαφορές μεταξύ των δύο θηκογραμμάτων της μεταβλητής ‘*age*’ που δημιουργούνται, τόσο για την κατηγορία των εργαζομένων που αποκόπηκαν από την μελέτη όσο και για εκείνους που εκδήλωσαν το γεγονός του ενδιαφέροντος. Η διάμεση τιμή της ηλικίας για τους εργαζομένους που εκδήλωσαν το γεγονός του ενδιαφέροντος είναι χαμηλότερη από εκείνους που αποκόπηκαν από την μελέτη. Παρ ’όλο που αυτή η διαφορά φαίνεται να είναι αμελητέα.



Διάγραμμα 8 - Θηκογράμματα της μεταβλητής 'age' για 'event = 0' και 'event = 1'

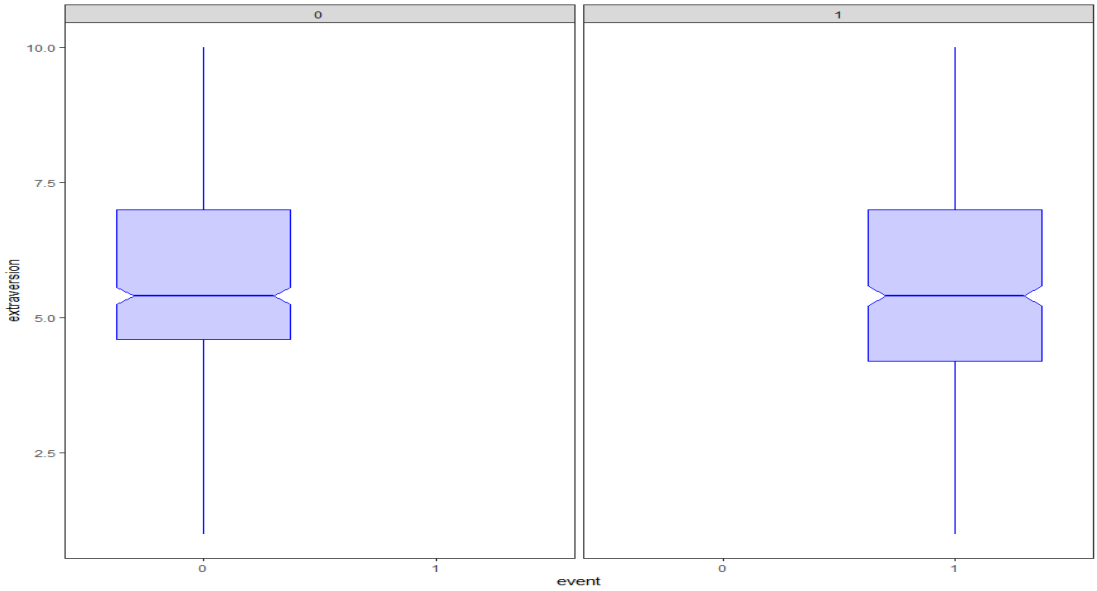
Ο δείκτης εξωστρέφειας του εργαζομένου (*extraversion*) παρατηρούμε πως έχει μέση τιμή κοντά στο 6 και διάμεσο 5.4. Σύμφωνα με το διάγραμμα (9), παρατηρούμε πως ο δείκτης εξωστρέφειας για την πλειοψηφία των εργαζομένων λαμβάνει τιμές μεταξύ του διαστήματος (5, 7.5). Επίσης, φαίνεται πως λίγοι εργαζόμενοι έχουν λάβει τιμές μεταξύ του διαστήματος (0,2) όσον αφορά τον βαθμό εξωστρέφειας τους.



Διάγραμμα 9 - : Ιστόγραμμα της μεταβλητής 'extraversion'

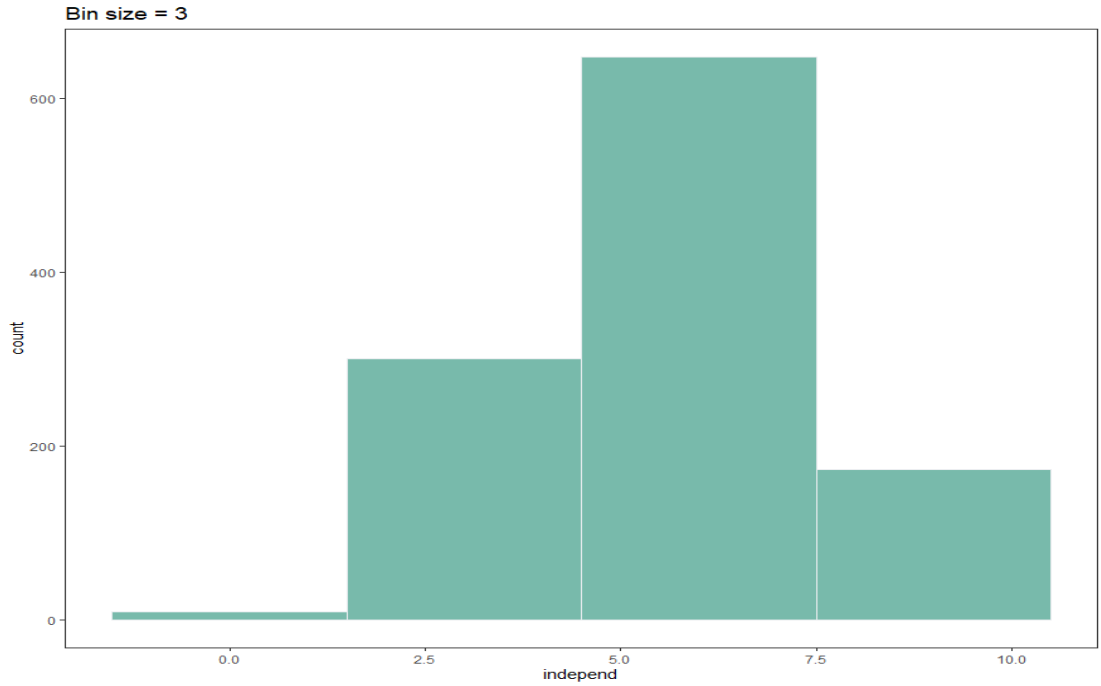
Όσον αφορά την εξωστρέφεια για τους εργαζομένους που αποκόπηκαν από την μελέτη σε σχέση με εκείνους που αποχώρησαν από την εταιρία, φαίνεται να μην υπάρχουν σημαντικές διαφορές. Σύμφωνα με τα θηκογράμματα που δημιουργήσαμε στο διάγραμμα (10), οι διάμεσοι για τις δύο

κατηγορίες φαίνεται να είναι στο ίδιο επίπεδο. Ίσως το γεγονός ότι το θηκόγραμμα των εργαζομένων που εκδήλωσαν το γεγονός του ενδιαφέροντος είναι λίγο πιο πλατύ σε σχέση με εκείνο των εργαζομένων που αποκόπηκαν από την μελέτη, θα μπορούσε να αποτελέσει την μόνη μικρή διαφορά.



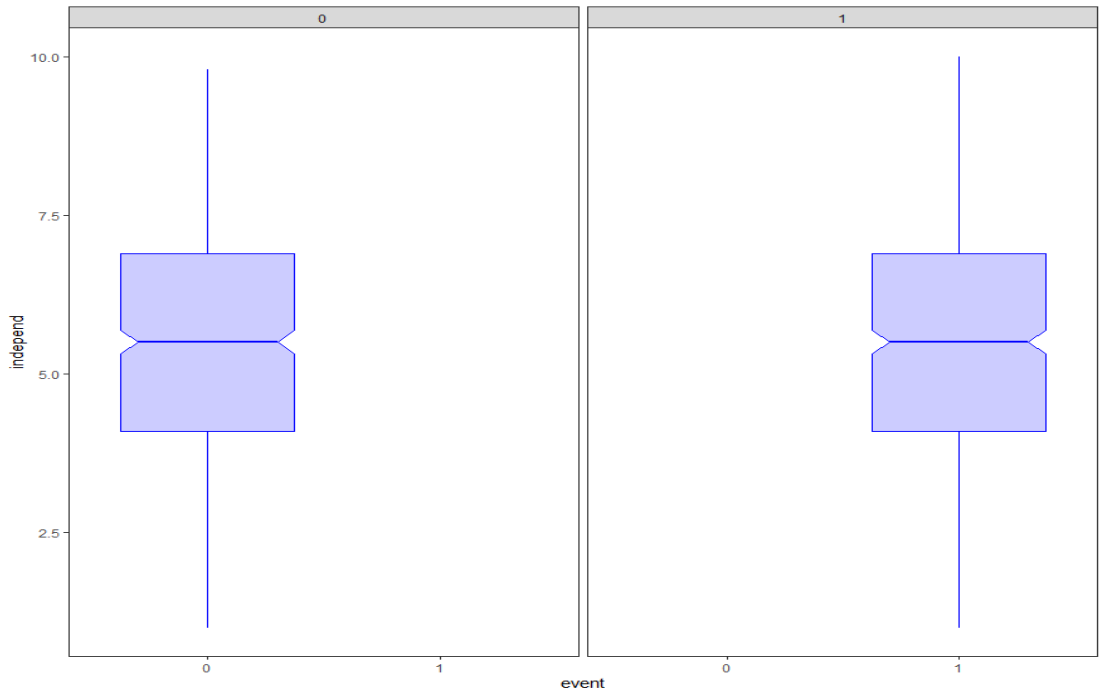
Διάγραμμα 10 - Θηκογράμματα της μεταβλητής 'extraversion' για 'event = 0' και 'event = 1'

Ο δείκτης που νιώθει ανεξάρτητος ο εργαζόμενος (*independ*) έχει μέση τιμή 5.5. Σύμφωνα με το παρακάτω διάγραμμα (11), περισσότεροι από 600 εργαζόμενοι του συνόλου δεδομένων μας λαμβάνουν τιμές για τον δείκτη ανεξαρτησίας μεταξύ του διαστήματος (4,7.5). Επίσης, λιγότεροι από 200 εργαζόμενοι λαμβάνουν τιμές σχετικά με τον βαθμό ανεξαρτησίας τους μεταξύ του διαστήματος (7.5,10).



Διάγραμμα 11 - Ιστόγραμμα της μεταβλητής 'independ'

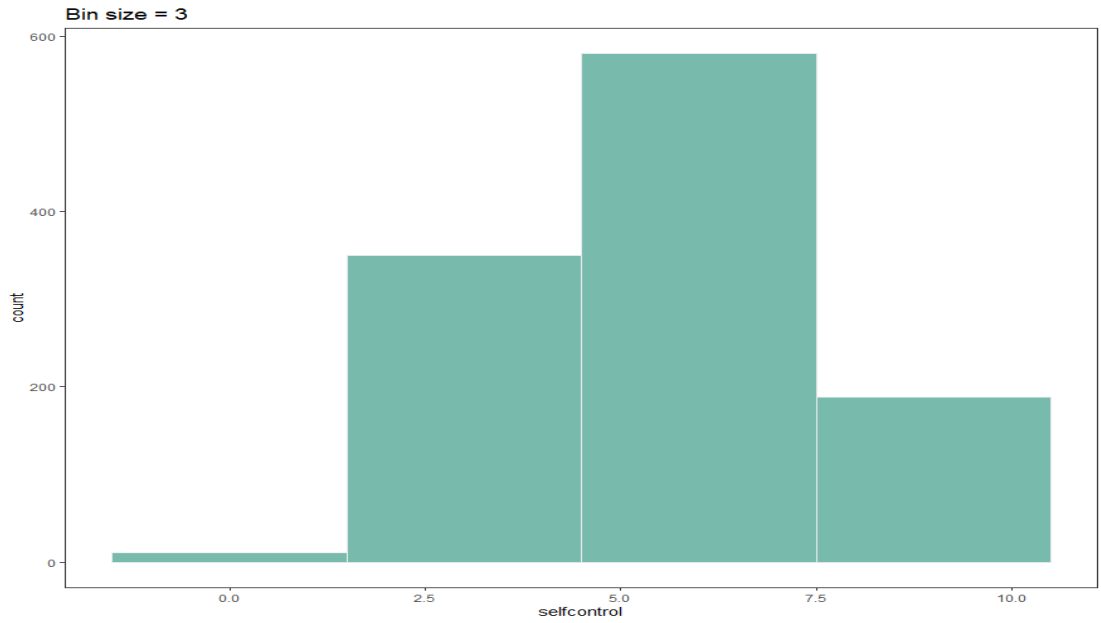
Από το διάγραμμα (12), παρατηρούμε ότι τα δύο θηκογράμματα που δημιουργήθηκαν για τους εργαζόμενους που έχουν εκδηλώσει το γεγονός του ενδιαφέροντος και εκείνους που αποκόπηκαν από την μελέτη, είναι πανομοιότυπα. Ο διάμεσος των δύο θηκογραμμάτων φαίνεται να είναι σχεδόν στο ίδιο επίπεδο.



Διάγραμμα 12 - Θηκογράμματα της μεταβλητής 'independ' για 'event = 0' και 'event = 1'

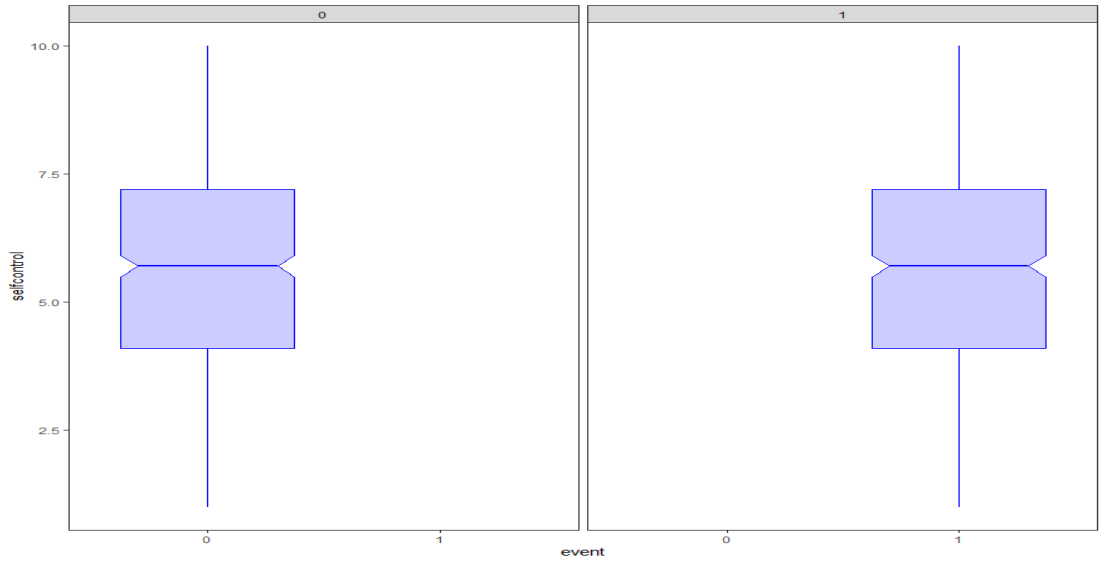
Ο δείκτης που εκφράζει τον αυτοέλεγχο για τον κάθε εργαζόμενο (*selfcontrol*) έχει μέση τιμή κοντά στο 5.6. Όπως φαίνεται στο διάγραμμα (13), ελάχιστα λιγότεροι από τους 600

εργαζομένους λαμβάνουν τιμή για τον βαθμό αυτοέλεγχου που έχουν μεταξύ του διαστήματος (4 , 7.5). Επίσης παρατηρούμε πως κοντά στους 350 εργαζομένους λαμβάνουν τιμή για τον βαθμό αυτοέλεγχου μεταξύ του διαστήματος (1.5 , 4).



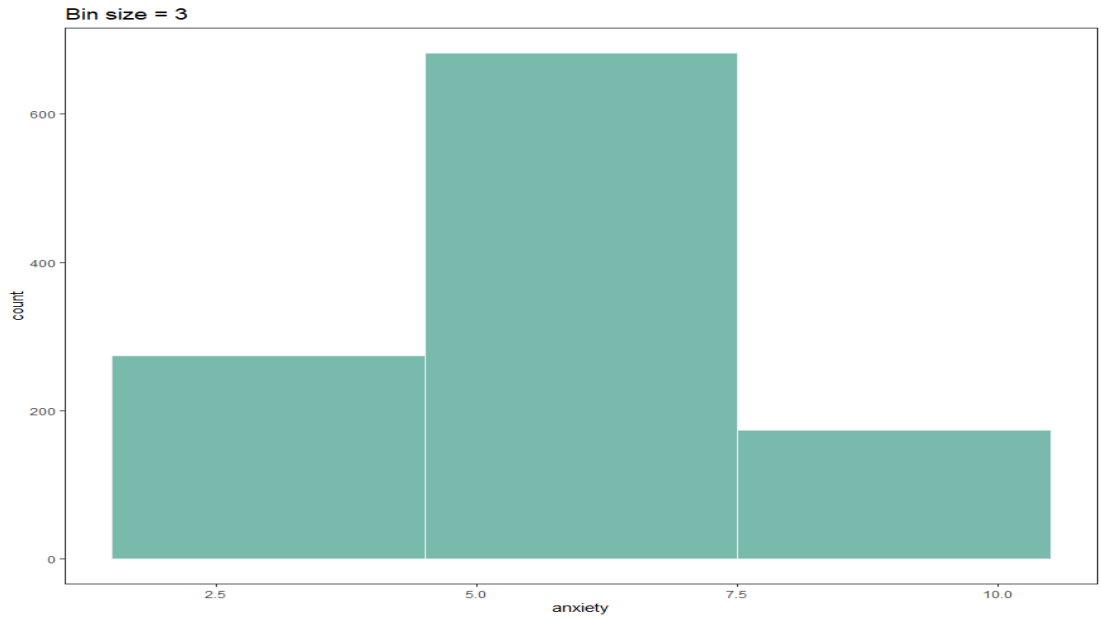
Διάγραμμα 13 - Ιστόγραμμα της μεταβλητής 'selfcontrol'

Σύμφωνα με το διάγραμμα (14), παρατηρούμε ότι τα δύο θηκογράμματα για τους εργαζομένους που εκδήλωσαν το γεγονός του ενδιαφέροντος καθώς και για εκείνους που αποκόπηκαν από την μελέτη είναι πανομοιότυπα. Δεν παρατηρείται καμία σημαντική αλλαγή για τις δύο κατηγορίες εργαζομένων.



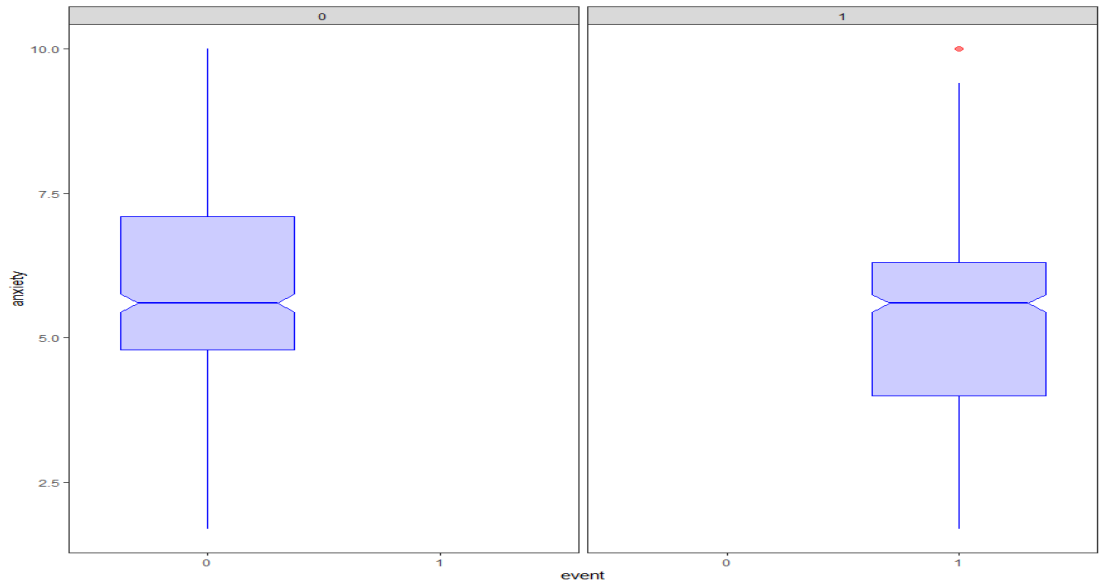
Διάγραμμα 14 - Θηκογράμματα της μεταβλητής 'selfcontrol' για 'event = 0' και 'event = 1'

Ο δείκτης που δηλώνει τον βαθμό άγχους για τον καθένα εργαζόμενο παρατηρούμε πως έχει μέση τιμή 5.7. Σύμφωνα με το διάγραμμα (15), κοντά στους 800 εργαζομένους λαμβάνουν τιμή για τον βαθμό που αγχώνονται μεταξύ του διαστήματος (4 , 7.5). Επίσης φαίνεται πως είναι λιγότεροι από 200 οι εργαζόμενοι που λαμβάνουν τιμές μεταξύ του διαστήματος (7.5 , 10) για τον βαθμό άγχους.



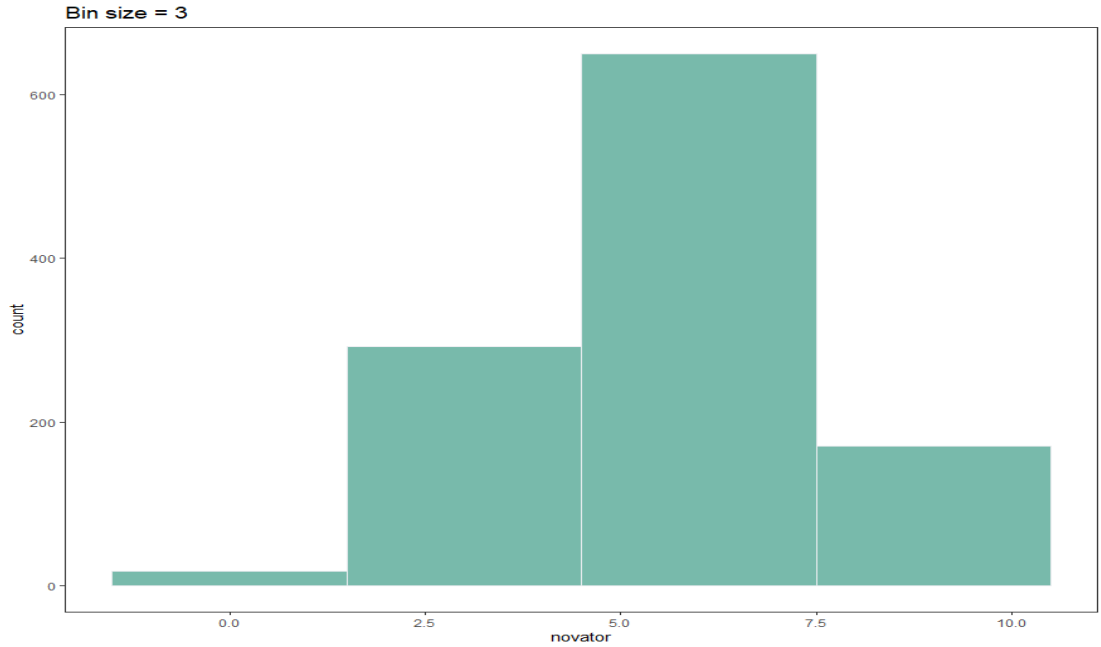
Διάγραμμα 15 - Ιστόγραμμα της μεταβλητής 'anxiety'

Στο διάγραμμα (16),βλέπουμε πως τα θηκογράμματα για τις δύο κατηγορίες εργαζομένων εμφανίζουν κάποια διαφορά μεταξύ τους. Το 1ο τεταρτημόριο για το θηκόγραμμα των εργαζομένων που εκδήλωσαν το γεγονός του ενδιαφέροντος, παρατηρούμε πως είναι αρκετά χαμηλότερο από το 1ο τεταρτημόριο εκείνων που αποκόπηκαν από την μελέτη. Επίσης, το 3ο τεταρτημόριο των εργαζομένων που αποκόπηκαν από την μελέτη φαίνεται να είναι υψηλότερα από το 3ο τεταρτημόριο εκείνων που εκδήλωσαν το γεγονός του ενδιαφέροντος. Επιπλέον, διακρίνουμε μία ακραία τιμή (outlier) για τους εργαζόμενους που εκδήλωσαν το γεγονός του ενδιαφέροντος. Τέλος, ο διάμεσος φαίνεται πως παραμένει στα ίδια επίπεδα και για τις δύο κατηγορίες εργαζομένων.



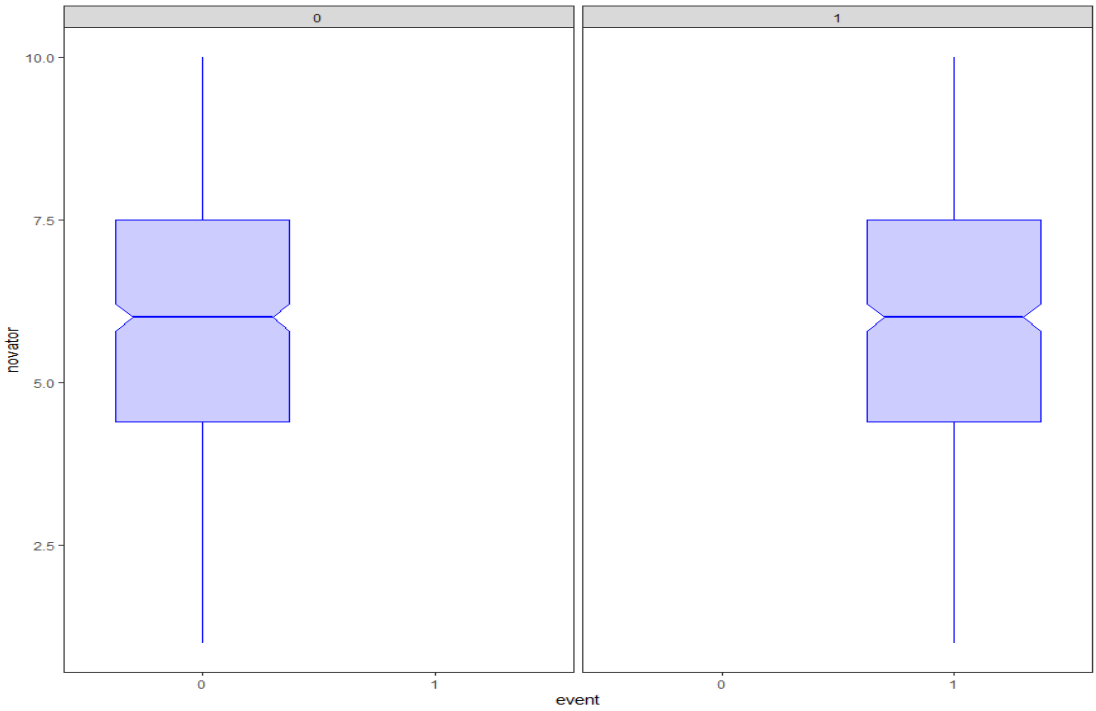
Διάγραμμα 16 - Θηκογράμματα της μεταβλητής 'anxiety' για 'event = 0' και 'event = 1'

Ο δείκτης που δηλώνει τον βαθμό αρχαριότητας του εργαζομένου (*novator*) βλέπουμε πως λαμβάνει μέση τιμή ίση με 5.9 και διάμεσο 6. Σύμφωνα με το διάγραμμα (17), είναι κοντά στους 700 οι εργαζόμενοι που λαμβάνου τιμές μεταξύ του διαστήματος (4 , 7.5) όσον αφορά τον βαθμό αρχαριότητας τους. Επίσης, κοντά στους 300 εργαζομένους λαμβάνουν τιμές μεταξύ του διαστήματος (1.5 , 4).



Διάγραμμα 17 - Ιστόγραμμα της μεταβλητής 'novator'

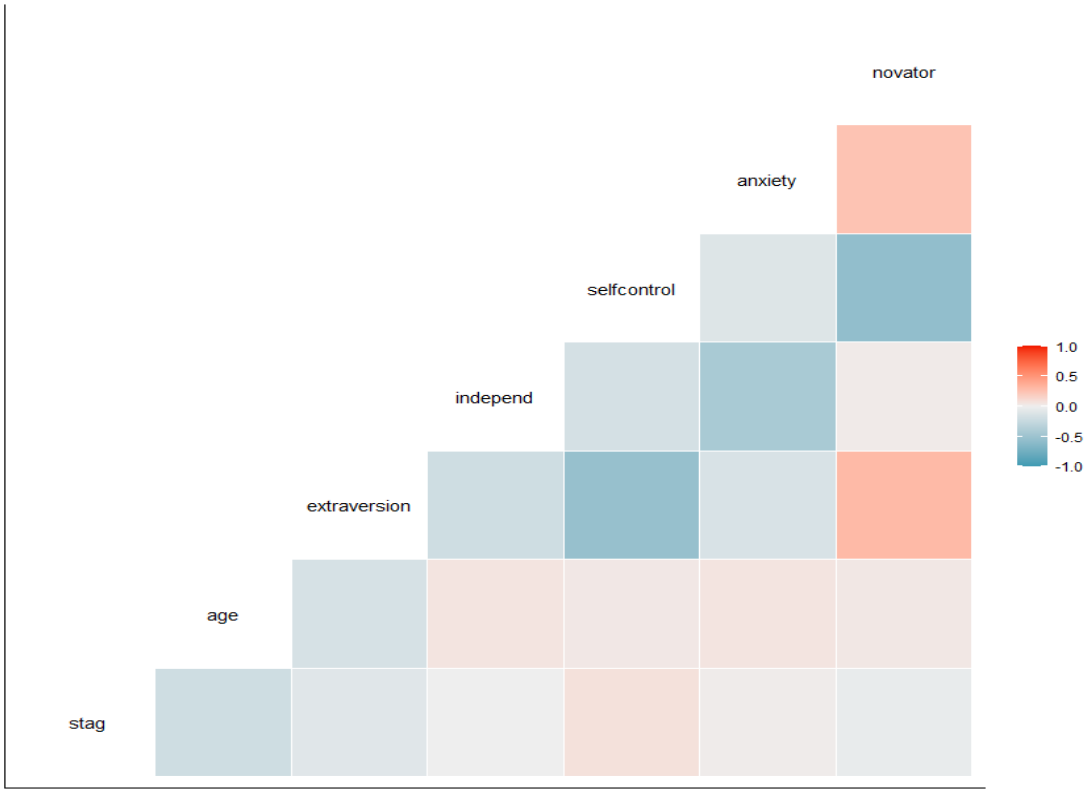
Σύμφωνα με το διάγραμμα (18), τα θηκογράμματα που δημιουργούνται για τις δύο κατηγορίες εργαζομένων, σε εκείνους που εκδήλωσαν το γεγονός του ενδιαφέροντος και για εκείνους που αποκόπηκαν από την μελέτη, δεν φαίνεται να εμφανίζουν καμία διαφορά.



Διάγραμμα 18 - Θηκογράμματα της μεταβλητής 'novator' για 'event = 0' και 'event = 1'

3.3.1.2. Συσχέτιση ποσοτικών μεταβλητών

Στην παρούσα υπό-ενότητα, θα δούμε την συσχέτιση των επεξηγηματικών μεταβλητών του υποδείγματος και θα ελέγξουμε τι σχέση προκύπτει μεταξύ τους. Από το διάγραμμα (19), οι μεταβλητές που δεν φαίνονται εντελώς ανεξάρτητες μεταξύ τους είναι οι '*extraversion*' με την '*selfcontrol*', '*extraversion*' με την '*novator*' , '*independ*' με την '*anxiety*', '*selfcontrol*' με '*novator*', '*anxiety*' με '*novator*'. Όπως παρατηρούμε στο διάγραμμα, οι μεταβλητές '*extraversion*' με '*novator*' εμφανίζουν θετική συσχέτιση, κινούμενές προς την ίδια κατεύθυνση. Παραδείγματος χάριν, άτομα με υψηλό δείκτη εξωστρέφειας φαίνεται πως έχουν και σχετικά υψηλό δείκτη αρχαριότητας. Επίσης, οι μεταβλητές '*anxiety*' και '*novator*' κινούνται προς την ίδια κατεύθυνση. Οι μεταβλητές '*extraversion*' με '*selfcontrol*', '*independ*' με '*anxiety*', '*selfcontrol*' με '*novator*' παρουσιάζουν αρνητική συσχέτιση, κινούμενες προς αντίθετες κατευθύνσεις. Παραδείγματος χάριν, για τις μεταβλητές '*selfcontrol*' με '*novator*' οι οποίες διακρίνουμε πως παρουσιάζουν και την μεγαλύτερη αρνητική συσχέτιση, κάποιος εργαζόμενος με υψηλό δείκτη αυτοέλεγχου εμφανίζει μικρό βαθμό αρχαριότητας.



Διάγραμμα 19 - Συσχέτιση ποσοτικών μεταβλητών του δείγματος

3.3.1.2. Στατιστικός έλεγχος ποσοτικών μεταβλητών

Έπειτα τον διαχωρισμό των εργαζομένων ,όσον αφορά την αποχώρηση τους από την εταιρία και της αποκοπής τους από την μελέτη για την κάθε μία μεταβλητή του υποδείγματος, θα γίνει έλεγχος αν υπάρχει στατιστικά σημαντική διαφορά της μέσης τιμής μεταξύ των δύο κατηγοριών εργαζομένων για την κάθε μία αριθμητική μεταβλητή. Το αποδεκτό επίπεδο σημαντικότητας στο παρακάτω στατιστικό έλεγχο υποθέτουμε πως είναι της τάξεως του 5% (α=5%). Όπως απεικονίζεται στον παρακάτω πίνακα (2), οι μεταβλητές '*stag*', '*age*', '*extraversion*', '*independ*', '*novator*' δεν φαίνεται να παρουσιάζουν στατιστικά σημαντική διαφορά της μέσης τιμής τους για τις δύο κατηγορίες εργαζομένων, όπου '*event = 0*' και '*event =1*'. Θα λέγαμε πως καταλήγουμε στο παρακάτω συμπέρασμα για δύο λόγους:

- Η τιμή για το παρατηρηθέν επίπεδο σημαντικότητας (*p-value*) ,για καθεμιά από τις παραπάνω μεταβλητές, είναι υψηλότερη από την τιμή του αποδεκτού επιπέδου σημαντικότητας. Οπότε αποτρέπεται η απόρριψη της μηδενικής υπόθεσης (H_0 = η διαφορά της μέσης τιμής για τις δύο κατηγορίες είναι ίση με το 0).
- Επίσης παρατηρούμε ότι εμπεριέχεται η μη αποδεκτή τιμή του 0 μεταξύ των δύο άκρων του διαστήματος εμπιστοσύνης για τις παραπάνω μεταβλητές. Η παρουσία της τιμής του

0 υποδηλώνει πως οι δύο κατηγορίες έχουν την ίδια μέση τιμή, με αποτέλεσμα να μην μπορούμε να απορρίψουμε την μηδενική υπόθεση.

Από τον πίνακα (2), διακρίνουμε πως η μοναδική αριθμητική μεταβλητή που εμφανίζει κάποια στατιστικά σημαντική διαφορά στην μέση τιμή για τις δύο κατηγορίες εργαζομένων είναι η '*anxiety*'. Η τιμή για το παρατηρηθέν επίπεδο σημαντικότητας (*p-value*) της μεταβλητής '*anxiety*' είναι χαμηλότερη από την τιμή του αποδεκτού επιπέδου σημαντικότητας, με αποτέλεσμα την απόρριψη την μηδενική (H_0) προς όφελος της εναλλακτικής υποθέσεως (H_1). Συμπεραίνουμε πως για τις δύο κατηγορίες υπάρχει διαφορά στην μέση τιμή και μάλιστα με τους εργαζομένους που χαρακτηρίζονται από αποκοπή να εμφανίζουν υψηλότερο δείκτη άγχους.

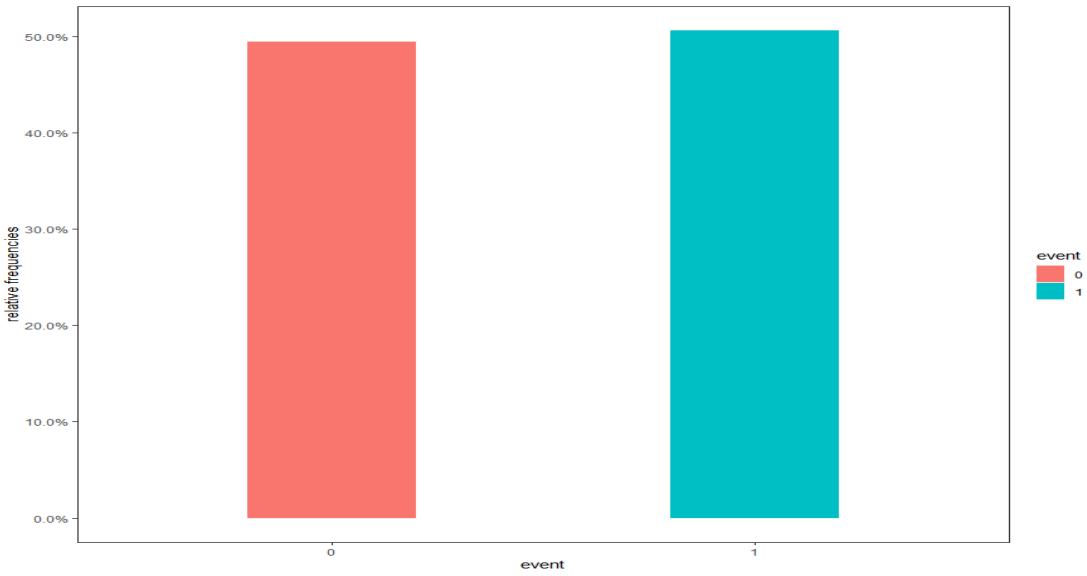
variables	mean in group 0	mean in group 1	t-statistic	df	p-value	lower .95	upper .95
stag	5,866487	5,892469	-0,22914	1127	0,8188	-0,2484561	0,1964923
age	31,41183	30,72995	1,6386	1127	0,1016	-0,1346322	1,4983874
extraversion	5,563441	5,620665	-0,51901	1127	0,6039	-0,2735578	0,1591085
independ	5,388710	5,565324	-1,7435	1127	0,08153	-0,3753741	0,02214547
anxiety	5,774910	5,558844	2,2127	1127	0,03364	0,0167531	0,41537943
novator	5,866487	5,892469	-0,22914	1127	0,8188	-0,2484561	0,1964923

Πίνακας 2 - Στατιστικός έλεγχος 't-test' ποσοτικών μεταβλητών

3.3.2. Ποιοτικές μεταβλητές

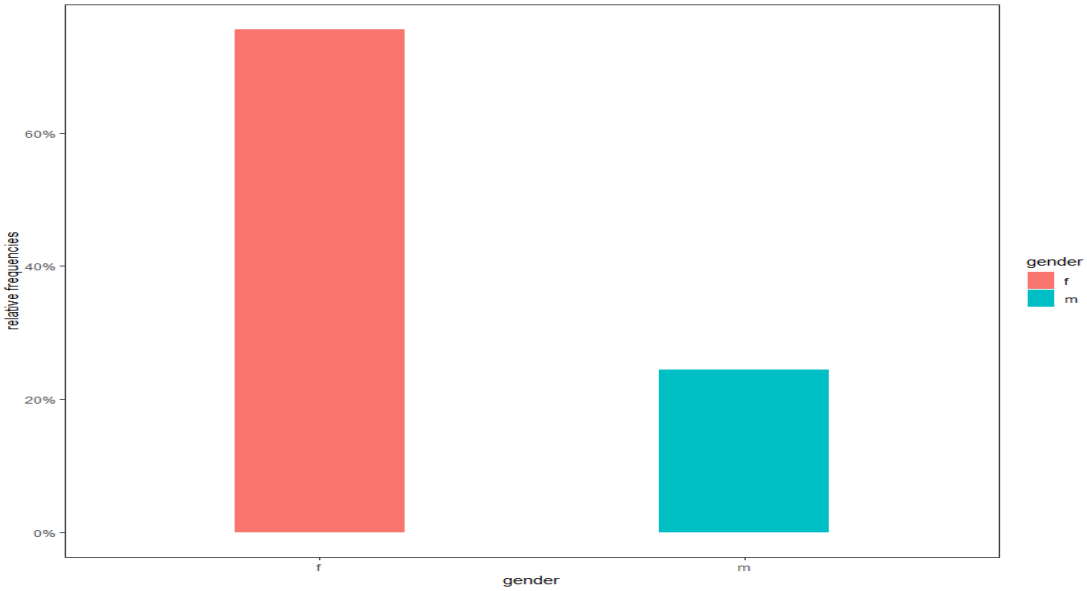
3.3.2.1. Στατιστικά γραφήματα ποιοτικών μεταβλητών

Στο παρακάτω διάγραμμα (20) της μεταβλητής '*event*', παρατηρούμε πως το ποσοστό των εργαζομένων που έχουν εκδηλώσει το γεγονός του ενδιαφέροντος είναι μεγαλύτερο από το ποσοστό εκείνων που αποκόπηκαν από την μελέτη. Οι διαφορές στις σχετικές συχνότητες μεταξύ των δύο κλάσεων, για '*event = 0'* και '*event = 1*' ,είναι πολύ μικρές. Όπως φαίνεται στο ραβόγραμμα, το ποσοστό των εργαζομένων που εκδήλωσαν το γεγονός του ενδιαφέροντος είναι λίγο παραπάνω από το 50% του συνολικού αριθμού των παρατηρήσεων, μεε αποτέλεσμα εκείνοι που αποκόπηκαν από την μελέτη να είναι ελάχιστα παρακάτω από το 50%.



Διάγραμμα 20 - Ραβδόγραμμα της μεταβλητής 'event'

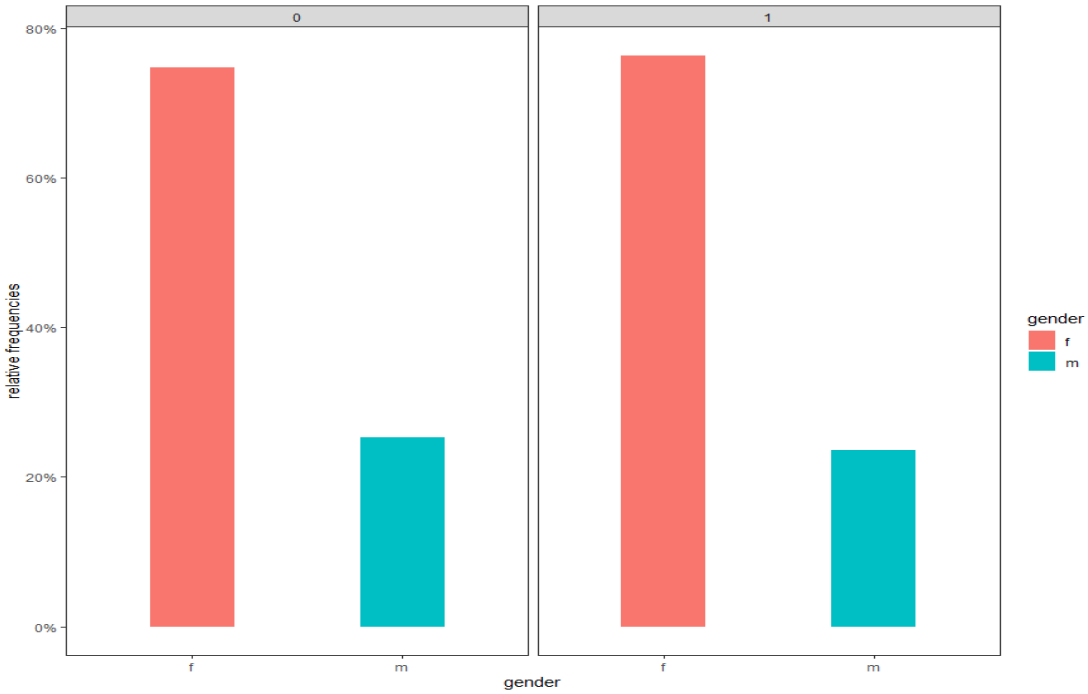
Στο διάγραμμα (21), όπου παρουσιάζεται το ραβδόγραμμα για το φύλο του εργαζομένου (*gender*), βλέπουμε ότι το ποσοστό των γυναικών που απασχολούνται από κάποια εταιρία είναι κατά πολύ περισσότερο από εκείνο των αντρών. Πιο συγκεκριμένα, το ποσοστό αντρών που συμπεριλαμβάνονται στην μελέτη προσεγγίζει το 25%, ενώ οι γυναίκες είναι λίγο παραπάνω από το 75% του συνολικού αριθμού εργαζομένων.



Διάγραμμα 21 - Ραβδόγραμμα της μεταβλητής 'gender'

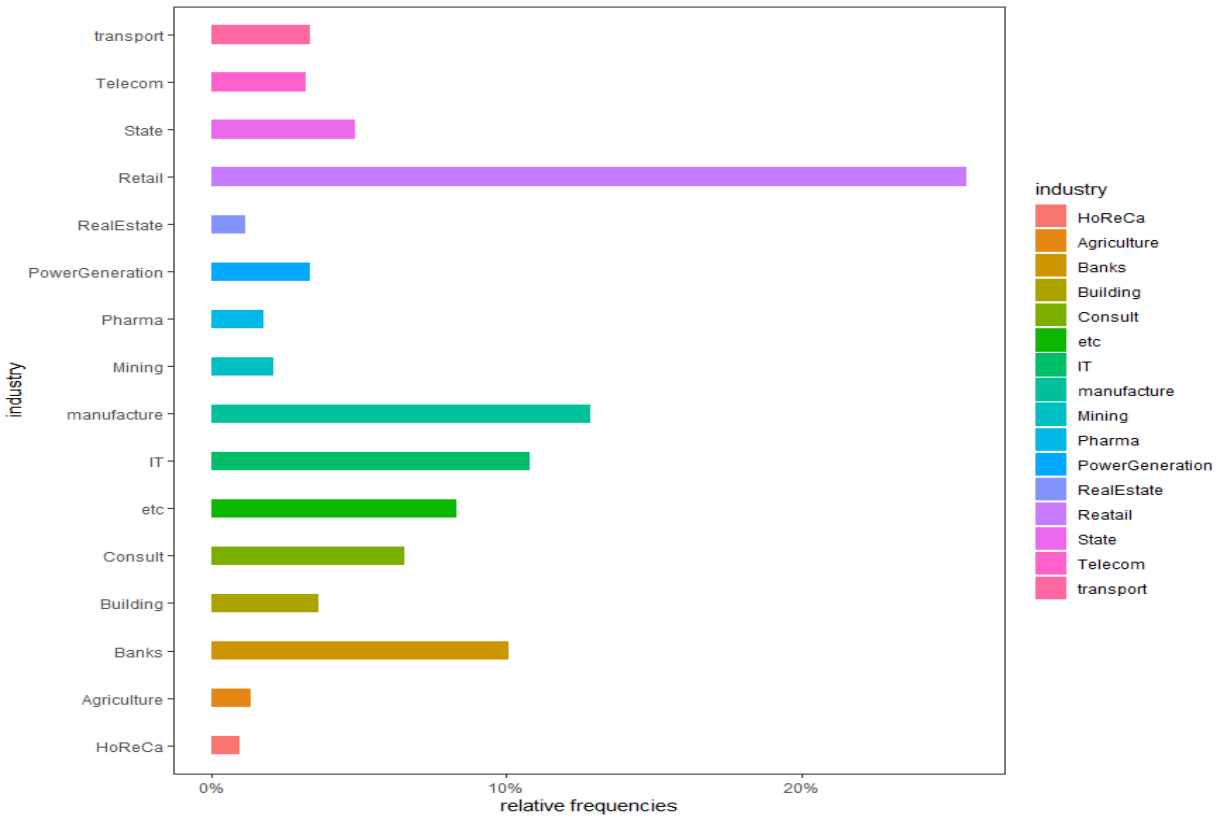
Σύμφωνα με το διάγραμμα (22), παρατηρούμε πως τα δύο ραβδογράμματα της μεταβλητής '*gender*' που δημιουργούνται, δεν εμφανίζουν αρκετά σημαντικές διαφορές. Θα μπορούσαμε να πούμε πως το ποσοστό των γυναικών που εκδήλωσαν το γεγονός του ενδιαφέροντος είναι ελάχιστα παραπάνω από των γυναικών που αποκόπηκαν από την μελέτη. Αντίθετα, το ποσοστό

των αντρών που αποκόπηκαν από την μελέτη είναι ελάχιστα παραπάνω από εκείνους που εκδήλωσαν το γεγονός.



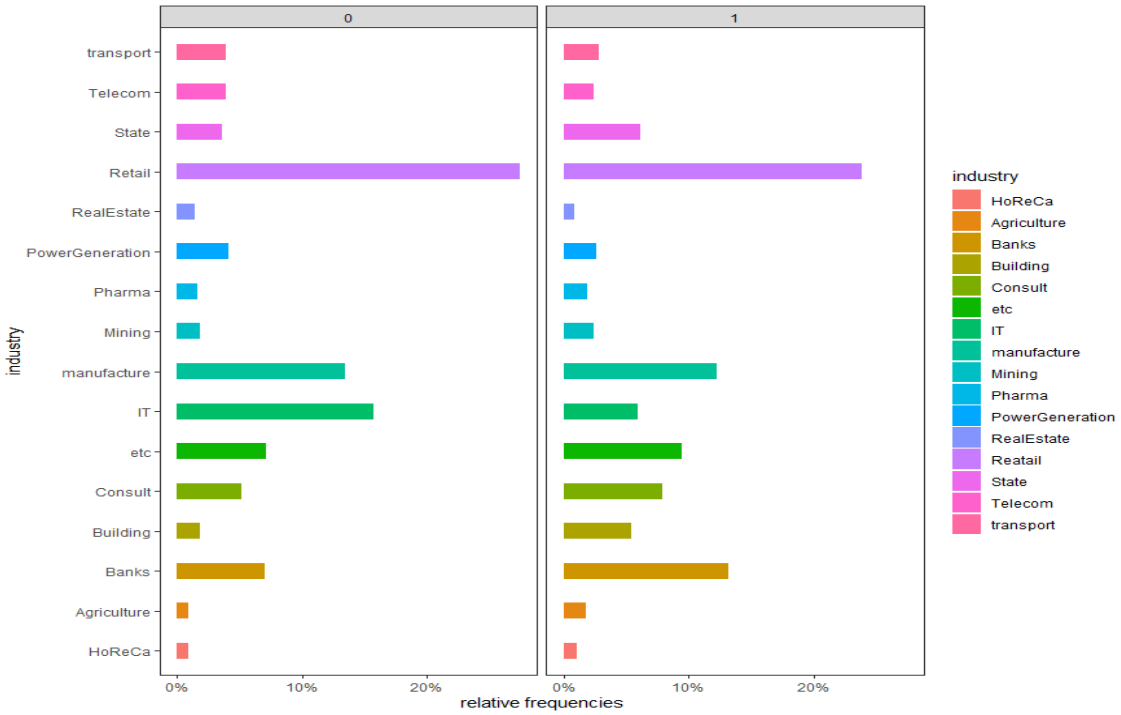
Διάγραμμα 22 - Ραβδόγραμμα της μεταβλητής 'gender' για 'event = 0' και ' event = 1'

Σύμφωνα με το διάγραμμα (23), από το ραβδόγραμμα της μεταβλητής '*industry'*, διακρίνουμε ότι οι επιχειρήσεις που ασχολούνται με τον κλάδο των λιανικών πωλήσεων (*Retail*) καταρτίζονται από το μεγαλύτερο ποσοστό εργαζομένων το οποίο ξεπερνά το 25% του συνολικού αριθμού εργαζομένων. Ακολουθεί ο κλάδος της παραγωγής (*manufacture*) και της τεχνολογίας (*IT*), όπου απασχολεί το 13% και το 11% των εργαζομένων αντίστοιχα από το σύνολο δεδομένων που μελετάμε. Επίσης ,υψηλή κατάταξη κατέχει και ο κλάδος των τραπεζών (*Banks*), καθώς απαρτίζεται από το 10% του συνολικού αριθμού εργαζομένων. Τέλος, ο κλάδος των ξενοδοχείων και των εστιατορίων (*HoReCa*) καθώς και ο κλάδος που ασχολείται με την αγορά και την πώληση ακινήτων (*RealEstate*), παρατηρούμε πως καταρτίζονται από τον μικρότερο αριθμό εργαζομένων, με ποσοστό 0.9% και 1.2% αντίστοιχα.



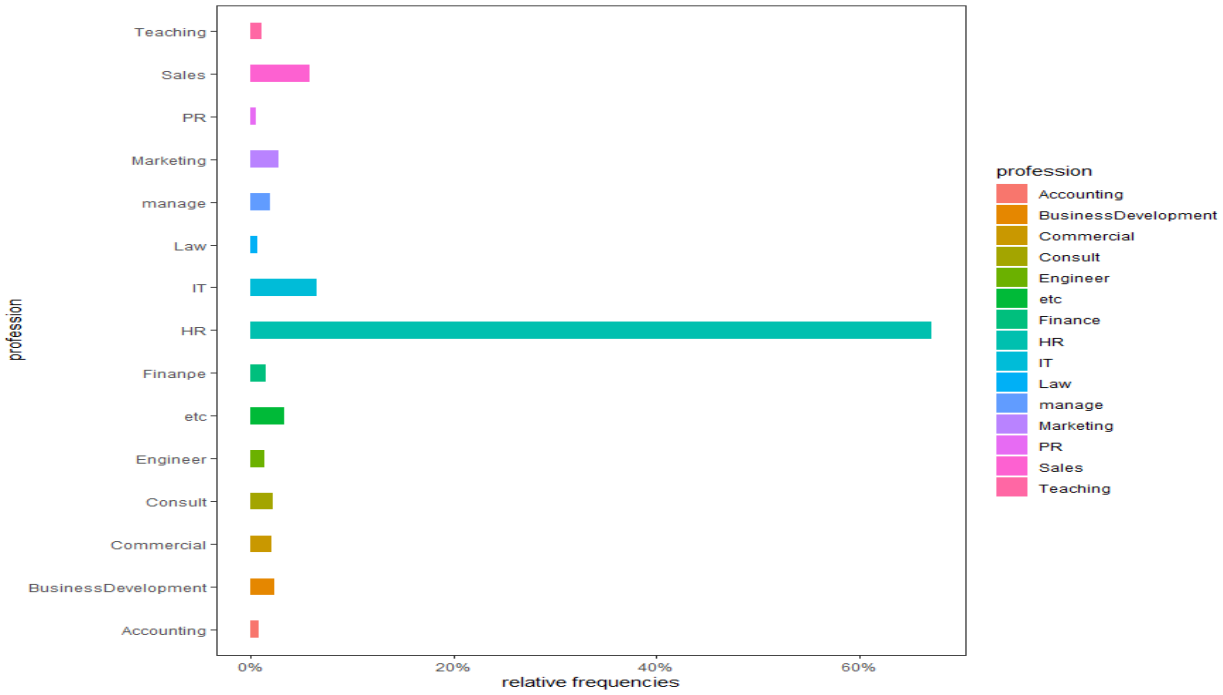
Διάγραμμα 23 - Ραβδόγραμμα της μεταβλητής 'industry'

Στα δύο ραβδογράμματα που δημιουργήθηκαν στο διάγραμμα (24), παρατηρούμε πως η μεγαλύτερη διαφορά στα ποσοστά εργαζομένων για τις δύο κατηγορίες της μεταβλητής '*event*', συμβαίνει για τον κλάδο της τεχνολογίας (*IT*) και τον κατασκευαστικό κλάδο (*Building*). Όσον αφορά τους εργαζομένους που έχουν εκδηλώσει το γεγονός του ενδιαφέροντος, το 6% ανήκει σε εργαζόμενους που απασχολούνται στον κλάδο της τεχνολογίας και το 5.4% ανήκει σε εργαζομένους που απασχολούνται από τον κατασκευαστικό κλάδο. Για τους εργαζομένους που έχουν αποκοπεί από την μελέτη, το 15.7% ανήκει σε εργαζομένους που απασχολούνται στο κλάδο της τεχνολογίας ενώ το 1.8% καταρτίζεται από εργαζομένους που απασχολούνται στον κατασκευαστικό κλάδο. Τέλος αξίζει να σημειωθεί πως για τον κλάδο των λιανικών πωλήσεων το ποσοστό εργαζομένων που έχουν 'αποτύχει' είναι 23.8% επί του συνολικού αριθμού εργαζομένων που έχουν 'αποτύχει', ενώ το ποσοστό εργαζομένων που έχουν αποκοπεί είναι 27.4% επί του συνολικού αριθμού των εργαζομένων που χαρακτηρίζονται από αποκοπή στην μελέτη.



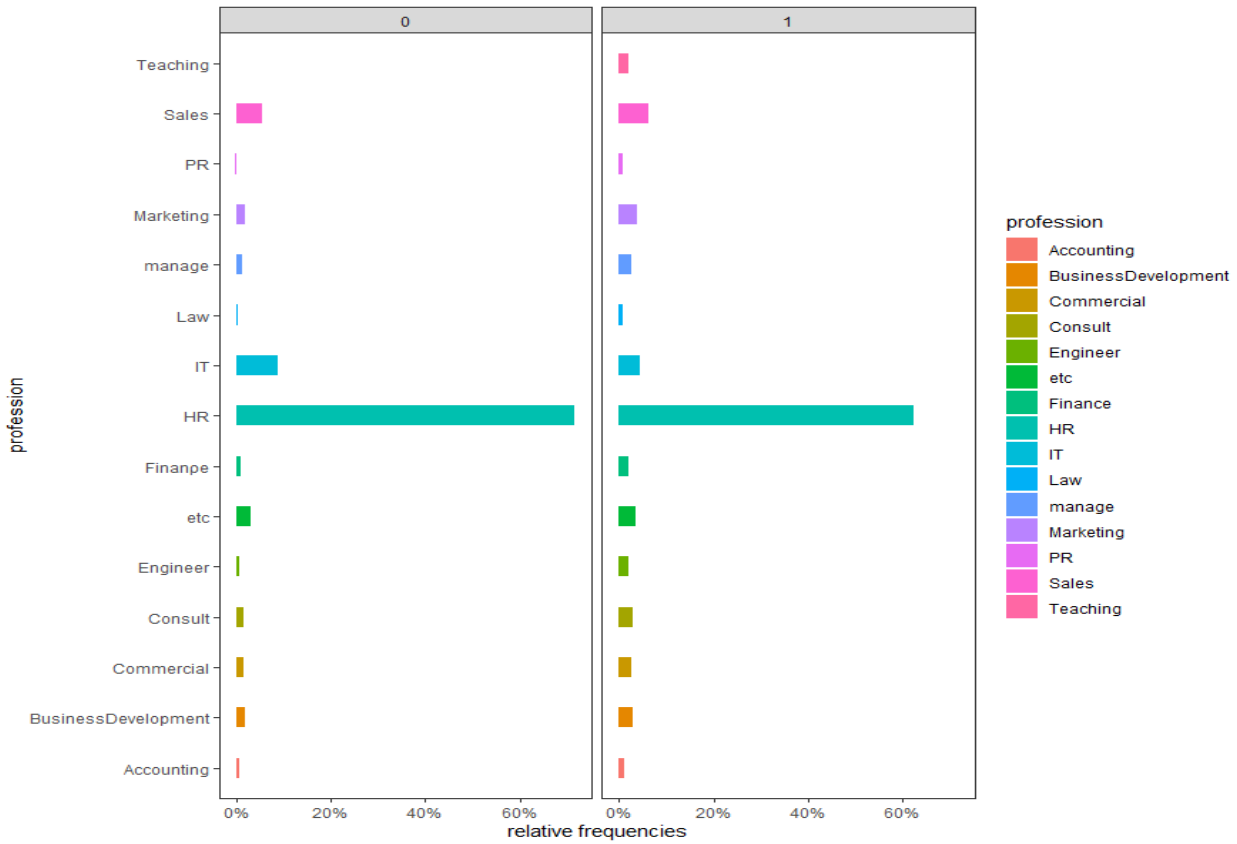
Διάγραμμα 24 - Ραβδόγραμμα της μεταβλητής 'industry' για 'event = 0' και ' event = 1'

Σύμφωνα με το ραβδόγραμμα της μεταβλητής '*profession*' στο διάγραμμα (25), παρατηρούμε πως το 67% από τον συνολικό αριθμό εργαζομένων απασχολούνται στο τμήμα ανθρωπίνων πόρων (*HR*) της εταιρίας. Δηλαδή παραπάνω από τους μισούς εργάζονται για το συγκεκριμένο τμήμα. Στην συνέχεια ακολουθεί το τμήμα της τεχνολογίας (*IT*) και το τμήμα των πωλήσεων (*Sales*), τα οποία απασχολούν το 6.5% και το 5.8% από τον συνολικό αριθμό εργαζομένων στο σύνολο δεδομένων μας αντίστοιχα. Τέλος στα τμήματα δημοσίων σχέσεων (*PR*), νομικών θεμάτων (*Law*) και λογιστικής (*Accounting*), απασχολείται το μικρότερο ποσοστό εργαζομένων από το σύνολο δεδομένων που μελετάμε.



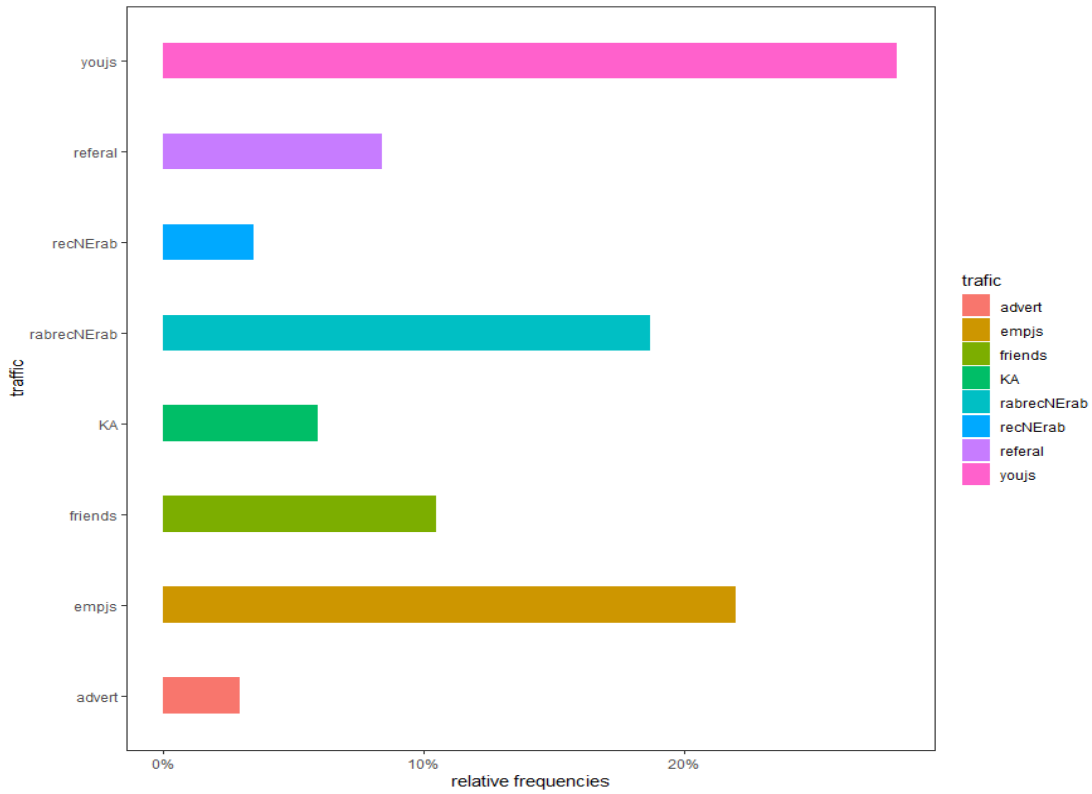
Διάγραμμα 25 - Ραβδόγραμμα της μεταβλητής 'Profession'

Όπως παρουσιάζεται στο διάγραμμα (26) παρακάτω, από τους εργαζομένους που αποκόπηκαν από την μελέτη το 71.6% είναι εκείνοι που απασχολούνταν στο τμήμα των ανθρωπίνων πόρων, ενώ από εκείνους που εκδήλωσαν το γεγονός το 62.5% είναι εργαζόμενοι που απασχολούνταν στο παρόμοιο τμήμα. Η μεγαλύτερη διαφορά για τις δύο κατηγορίες εργαζομένων όσον αφορά την εκδήλωση του γεγονότος ή όχι, εμφανίζεται στο τμήμα της εκμάθησης (*Teaching*) στο οποίο κανένας εργαζόμενος δεν αποκόπηκε από την μελέτη αλλά όλοι που ανήκαν στο συγκεκριμένο τμήμα εκδήλωσαν το γεγονός του ενδιαφέροντος. Επίσης στα τμήματα της τεχνολογίας (*IT*), δημόσιων σχέσεων (*PR*) και μηχανολογίας (*Engineer*) διακρίνουμε σημαντικές όσον αφορά τους εργαζομένους που εκδήλωσαν το γεγονός και εκείνους αποκόπηκαν από την μελέτη.



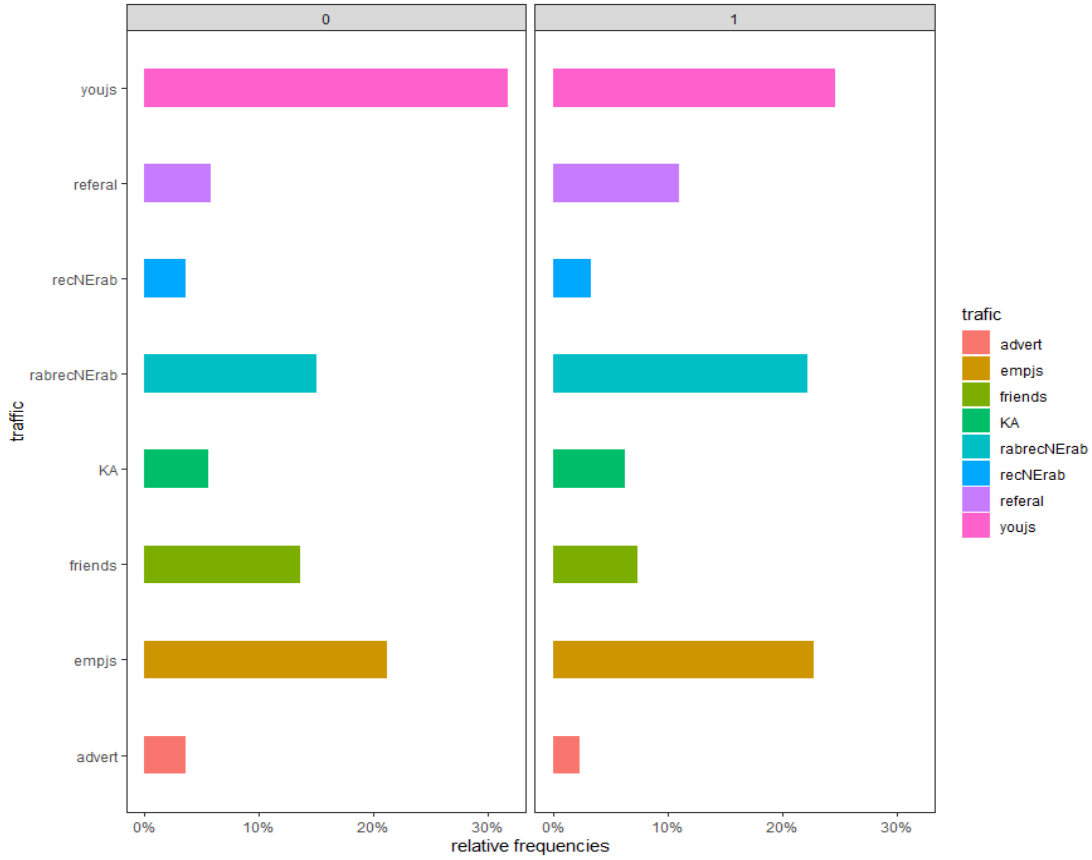
Διάγραμμα 26 - Ραβδόγραμμα της μεταβλητής 'profession' για 'event = 0' και ' event = 1'

Στο παρακάτω διάγραμμα (27) της ποιοτικής μεταβλητής '*traffic*', βλέπουμε πως ο μεγαλύτερος αριθμός εργαζομένων που προσλήφθηκε από κάποια εταιρία έγινε από την την δημοσίευση της θέσεως εργασίας σε κάποια ιστοσελίδα (*youjs*). Χρησιμοποιώντας αυτή την τακτική η εταιρία κατάφερε να προσεγγίσει το 28.2% από το συνολικό αριθμό εργαζομένων του συνόλου δεδομένων. Ακολουθούν οι προσλήψεις που έγιναν, μέσω της αναζήτησης του βιογραφικού του εργαζομένου από τον εργοδότη (*empjs*) και μέσω της επικοινωνίας του εργοδότη με τον εργαζόμενο, ύστερα από την πρόταση γνωστού ατόμου του εργαζομένου (*rabrecNErab*). Οι δύο αυτές τακτικές καταλαμβάνουν το 22% και το 18.6% αντίστοιχα, από τον συνολικό αριθμό προσλήψεων που έγιναν στην εταιρία.



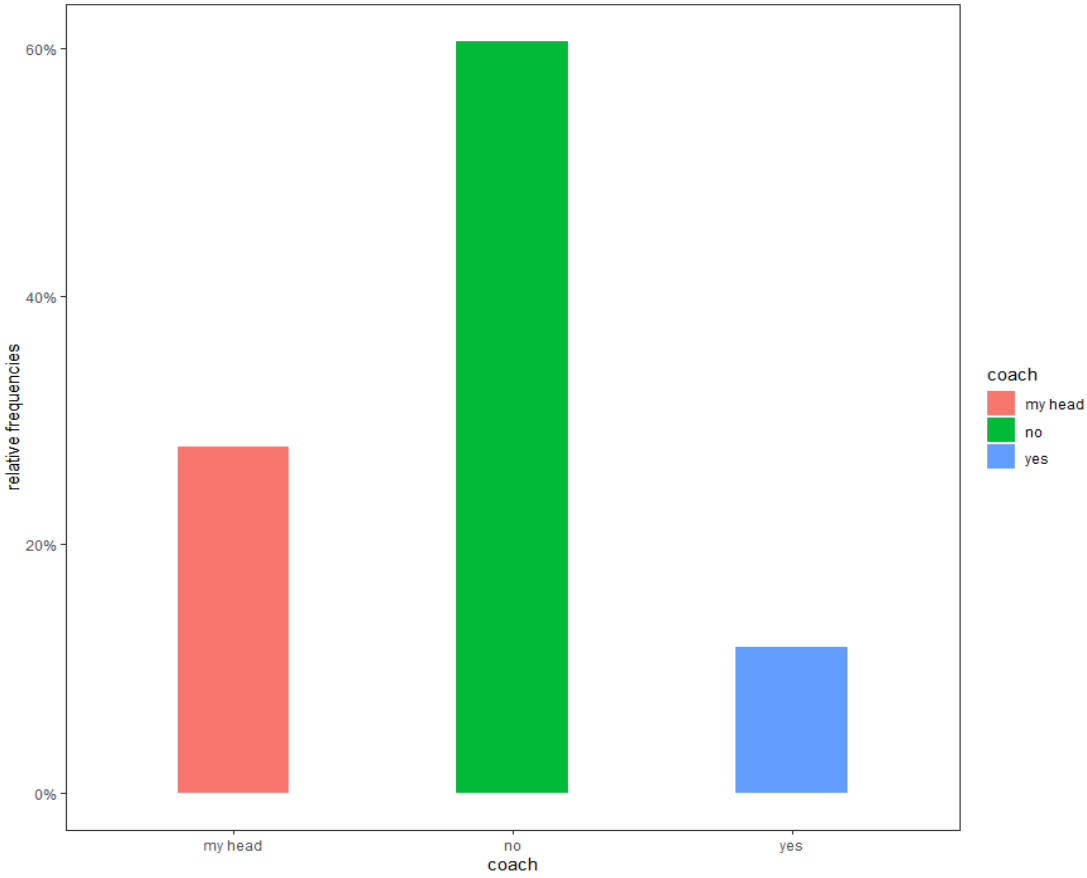
Διάγραμμα 27 - Ραβδόγραμμα της μεταβλητής 'traffic'

Ανάλογα με το τρόπο που χρησιμοποιήθηκε για την προσέλκυση εργαζομένων από την εταιρία, μπορούμε να διακρίνουμε πως οι εργαζόμενοι που ανήκουν στις κατηγορίες '*referral*', '*friends*' και '*rabrecNErab*' είναι εκείνοι που εμφανίζουν τις μεγαλύτερες διαφορές σχετικά με εκείνους που έχουν εκδηλώσει το γεγονός και εκείνους που έχουν αποκοπεί. Όπως απεικονίζεται στο διάγραμμα (28), το ποσοστό εργαζομένων που ανήκουν στις κατηγορίες '*referal*', '*friends*' και '*rabrecNErab*' είναι 11% , 7.4% και 22.2% αντιστοίχως, από τον συνολικό αριθμό εργαζομένων που εκδήλωσαν το γεγονός. Το ποσοστό εργαζομένων που αποκόπηκαν και εντάσσονται στις ίδιες κατηγορίες είναι 5.7%, 13.6% και 15% αντίστοιχα όπως προηγουμένως, από τον συνολικό αριθμό εργαζομένων που έχουν αποκοπεί.



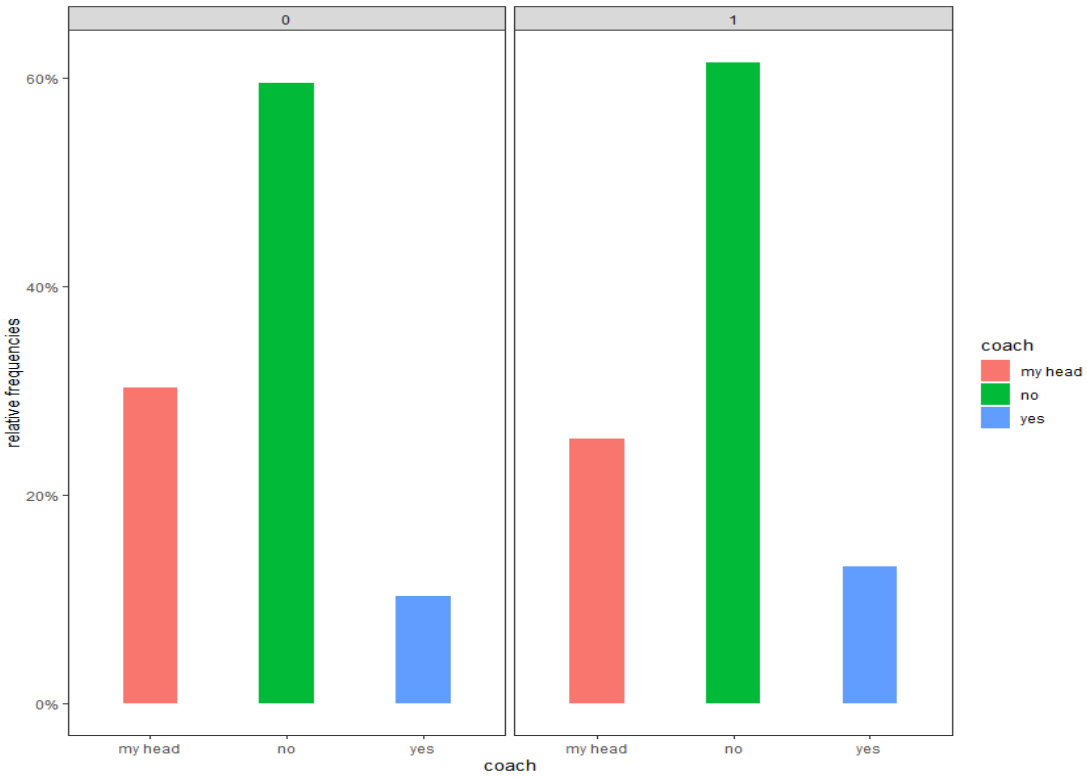
Διάγραμμα 28 - Ραβδόγραμμα της μεταβλητής 'traffic' για 'event = 0' και ' event = 1'

Σύμφωνα με το διάγραμμα (29) της μεταβλητής '*coach*', παρατηρούμε πως στο μεγαλύτερο ποσοστό εργαζομένων, περίπου το 60.5%, δεν χρησιμοποιήθηκε προπονητής (*no*) προκειμένου να βοηθήσει τον εργαζόμενο με την εκπλήρωση των καθηκόντων του στις αρχές τις πρόσληψης του. Στην συνέχεια το ποσοστό εργαζομένων που η παρουσία προπονητή είναι συνεχώς δίπλα στον εργαζόμενο (*my head)* είναι 27.8%. Τέλος, το 11.7% επί του συνολικού αριθμού των δεδομένων αποτελείται από εργαζομένους που χρησιμοποιήθηκε κάποιος προπονητής ή κάποιο πρόγραμμα εκπαίδευσης για να τους βοηθήσει στην αρχή.



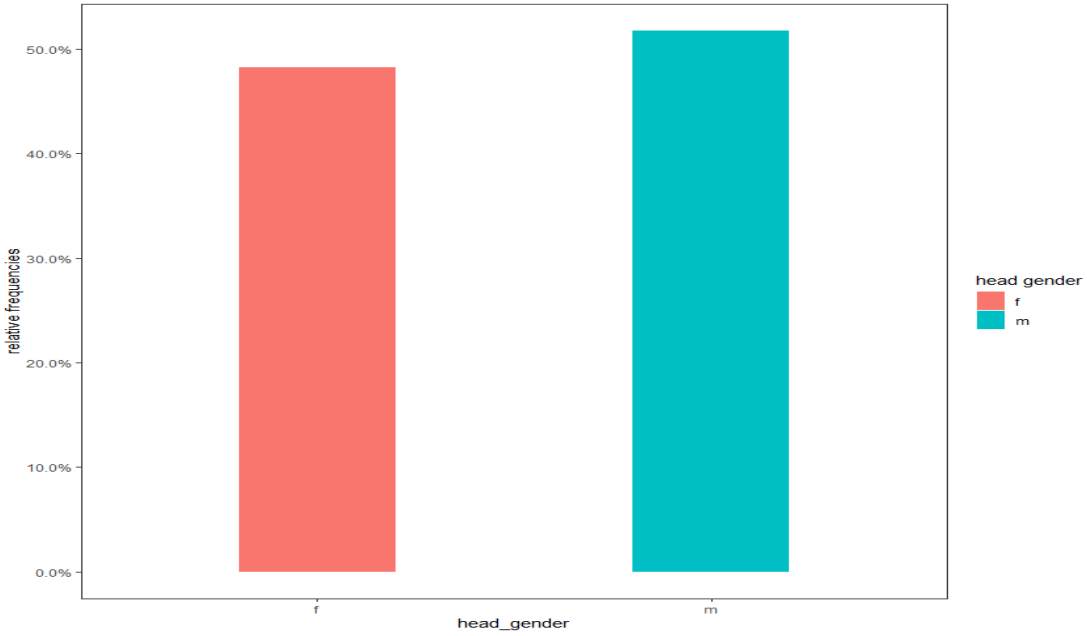
Διάγραμμα 29 - Ραβδόγραμμα της μεταβλητής 'coach'

Από τα ραβδογράμματα του γραφήματος (30), παρατηρούμε πως οι διαφορές μεταξύ των εργαζομένων που εκδήλωσαν το γεγονός και εκείνων που αποκόπηκαν από την μελέτη και για τις τρεις κατηγορίες της ποιοτικής μεταβλητής '*coach*' είναι πάρα πολύ μικρές. Η μεγαλύτερη διαφορά θα μπορούσαμε να πούμε πως υπάρχει για την κατηγορία '*yes*' όπου το ποσοστό εργαζομένων που εκδήλωσε το γεγονός είναι 13.1%, από το συνολικό ποσοστό εργαζομένων που εκδήλωσαν το γεγονός, ενώ το ποσοστό εκείνων που αποκόπηκαν από την μελέτη είναι 10.2%, από το σύνολο των εργαζομένων που αποκόπηκαν.



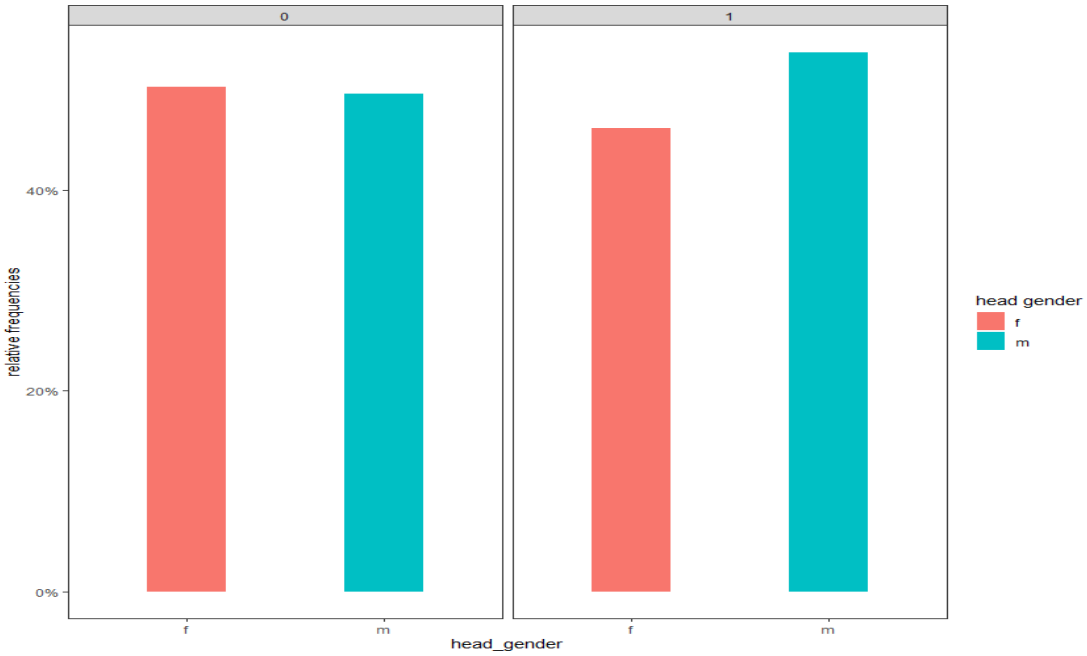
Διάγραμμα 30 - Ραβδόγραμμα της μεταβλητής 'coach' για 'event = 0' και ' event = 1'

Όπως απεικονίζεται στο ραβδόγραμμα της μεταβλητής *'head gender'* του γραφήματος (3.28), οι επιτηρητές των εργαζομένων είναι άντρες στην πλειοψηφία με ποσοστό της τάξεως του 51.7%. Επομένως οι γυναίκες προϊστάμενοι έχουν ποσοστό της τάξεως του 48.3% στο σύνολο δεομένων που μελετάμε.



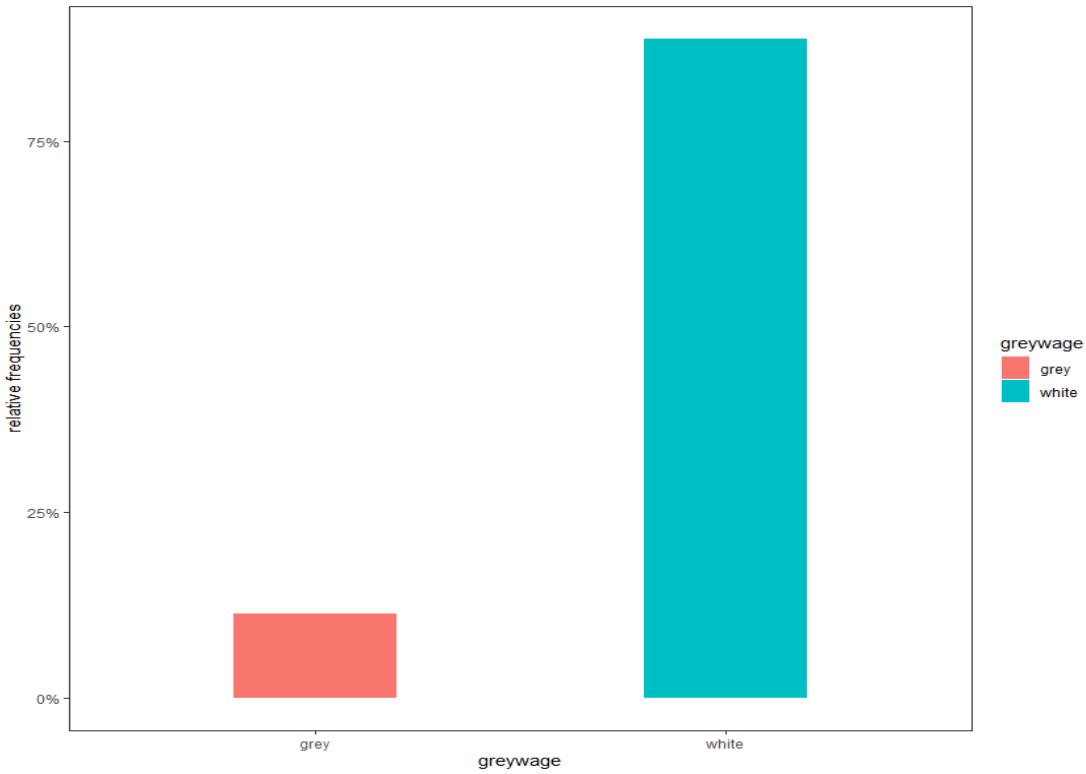
Διάγραμμα 31 - Ραβδόγραμμα της μεταβλητής 'head-gender'

Σύμφωνα με τα δύο ραβδογράμματα που δημιουργούνται, διακρίνουμε πως η διαφορά μεταξύ της εκδήλωσης του γεγονότος και της αποκοπής για τους εργαζομένους είναι μεγαλύτερη, χωρίς ιδιαίτερα σημαντικές διαφορές, στην περίπτωση που η κατηγορία της μεταβλητής *'head gender'* είναι '*m*'. Παρατηρώντας το διάγραμμα (3.29), το ποσοστό των εργαζομένων που εκδήλωσαν το γεγονός του ενδιαφέροντας έχοντας άντρα επιτηρητή είναι 53.7% από το συνολικό αριθμό των εργαζομένων που εκδήλωσαν το γεγονός. Αντίθετα, το ποσοστό εργαζομένων που αποκόπηκαν έχοντας άντρα επιτηρητή είναι 49.6%, από το συνολικό αριθμό εκείνων που αποκόπηκαν. Επίσης βλέπουμε πως το μεγαλύτερο ποσοστό εργαζομένων που αποκόπηκαν από την μελέτη είχαν γυναίκα προϊσταμένη. Ενώ το μεγαλύτερο ποσοστό εργαζομένων που αποχώρησαν από την εταιρία είχαν άντρα προϊστάμενο.



Διάγραμμα 32 - Ραβδόγραμμα της μεταβλητής 'head gender' για 'event = 0' και ' event = 1'

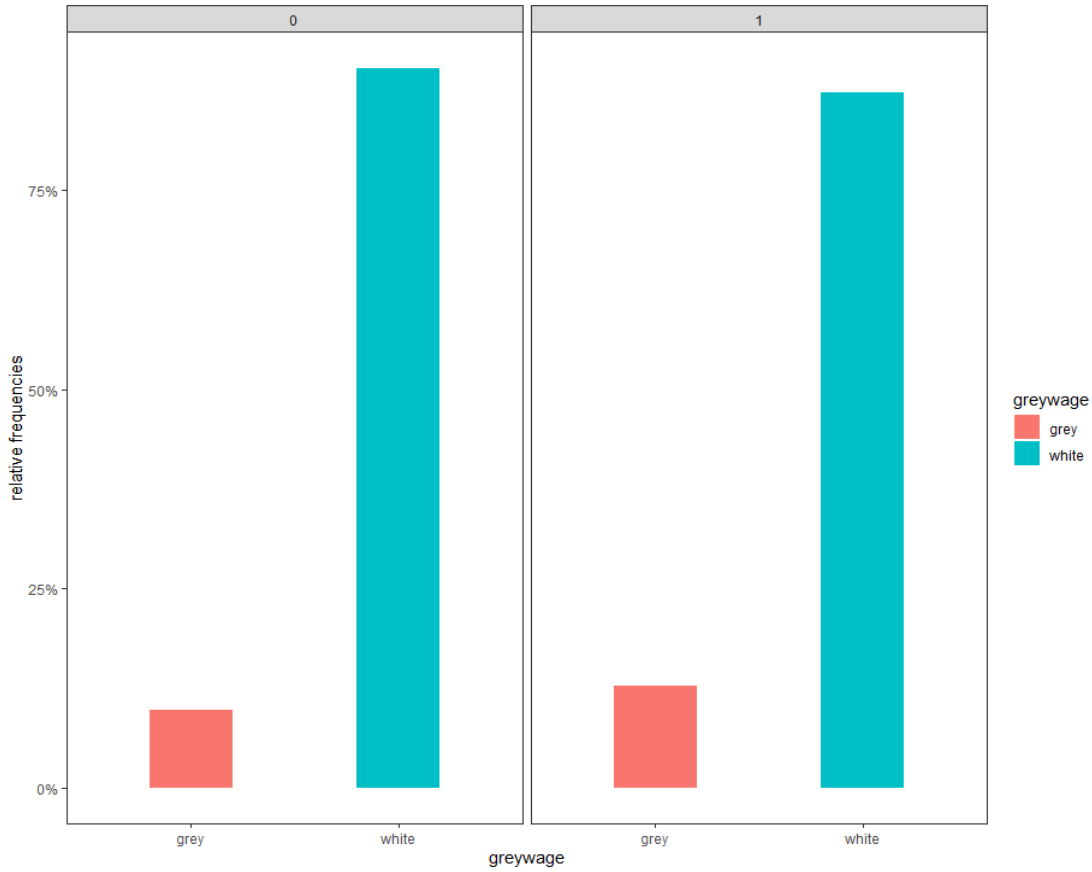
Σχετικά με το ραβδόγραμμα της μεταβλητής '*greywage*', παρατηρούμε πως ο συνολικός αριθμός εργαζομένων που λαμβάνουν τον μισθό που τους αναλογεί (*white*) σε σχέση με εκείνους που πληρώνονται παραπάνω από τον μισθό που δικαιούνται (*grey*) είναι πολύ μεγαλύτερος. Πιο συγκεκριμένα, όπως απεικονίζεται και στο διάγραμμα (33), οι εργαζόμενοι που ανήκουν στην κατηγορία '*grey*' και '*white*' αποτελούν το 11.2% και 88.8% του συνολικού αριθμού εργαζομένων, αντίστοιχα.



Διάγραμμα 33 - Ραβδόγραμμα της μεταβλητής 'greywage'

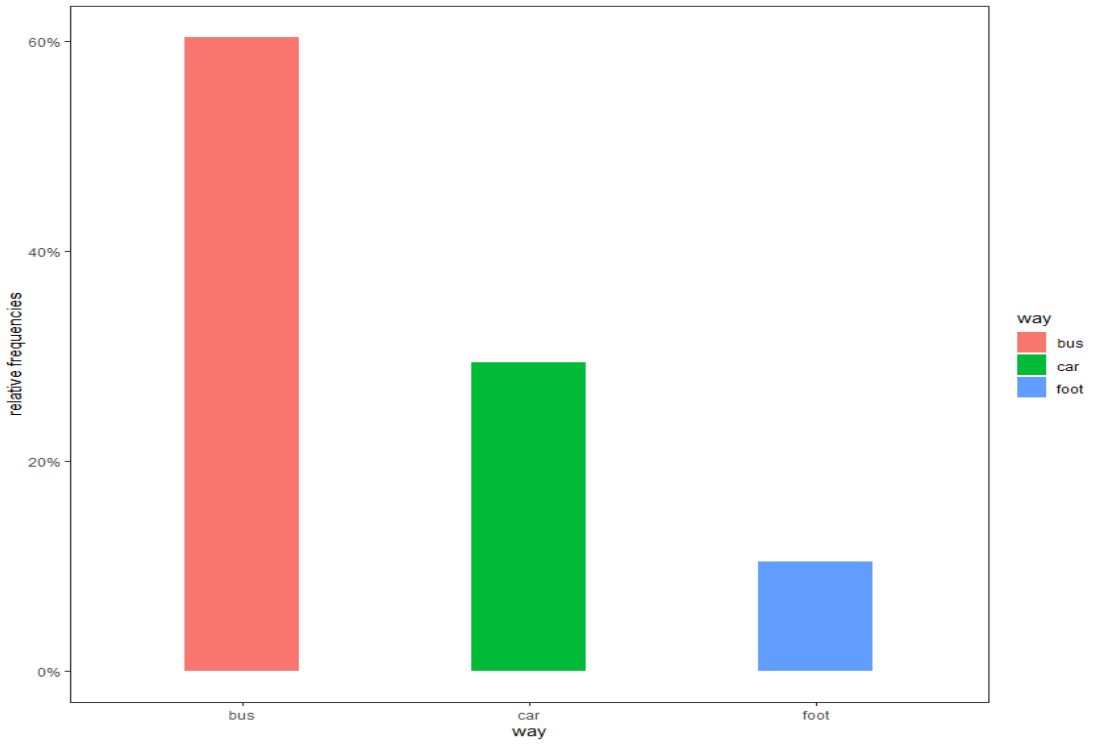
Παρατηρώντας τα ραβδογράμματα που δημιουργούνται στο διάγραμμα (34), οι διαφορές μεταξύ των εργαζομένων που εκδήλωσαν το γεγονός του ενδιαφέροντος και εκείνων που αποκόπηκαν και για τις δύο κατηγορίες της μεταβλητής ‘greywage’ είναι πολύ μικρές. Θα μπορούσαμε να πούμε πως οι εργαζόμενοι που ανήκουν στην κατηγορία ‘*grey*’ εμφανίζουν μεγαλύτερη διαφορά. Όπως φαίνεται στο διάγραμμα (33) το ποσοστό εργαζομένων της κατηγορίας ‘*grey*’ που εκδήλωσαν το γεγονός είναι 12.8% από το συνολικό ποσοστό εκείνων που εκδήλωσαν το

γεγονός, ενώ εκείνοι που αποκόπηκαν αποτελούν το 9.6% από το συνολικό ποσοστό των εργαζομένων που αποκόπηκαν.



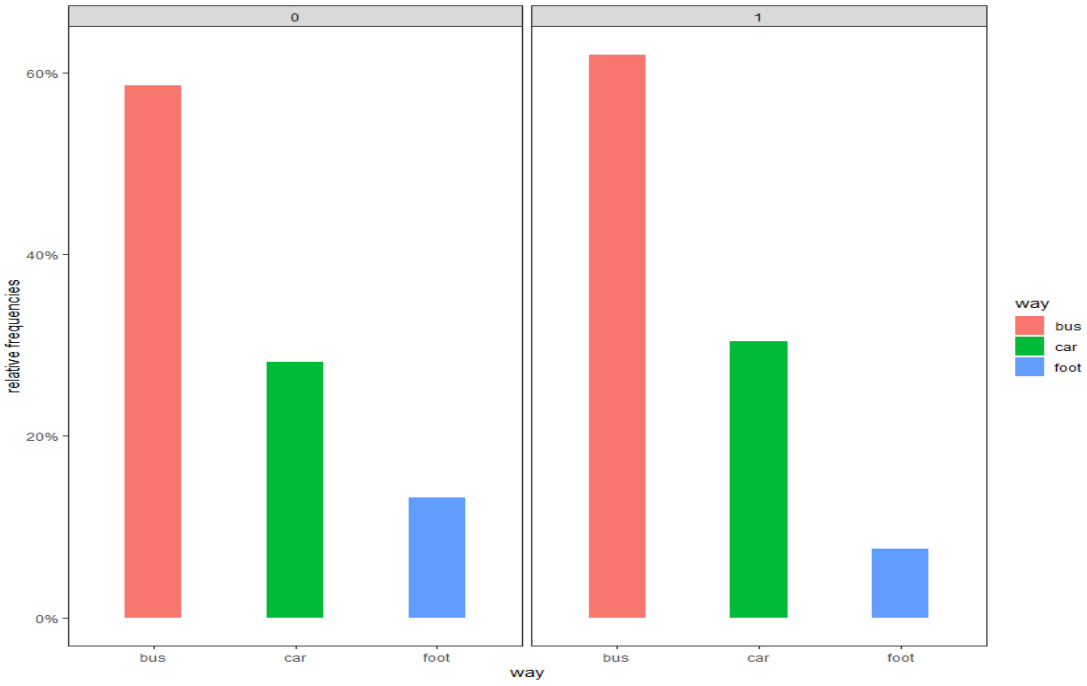
Διάγραμμα 34 - Ραβδόγραμμα της μεταβλητής 'greywage' για 'event = 0' και ' event = 1'

Παρατηρώντας το παρακάτω ραβδόγραμμα της μεταβλητής '*way*', το μεγαλύτερο ποσοστό εργαζομένων χρησιμοποιεί λεωφορείο προκειμένου να μετακινηθεί στην εταιρία. Έπειτα ακολουθεί το ποσοστό εργαζομένων που χρησιμοποιεί αυτοκίνητο προκειμένου να μετακινηθεί και τέλος ο αριθμός των εργαζομένων που προτιμούν να πάνε στην εταιρία με τα πόδια. Όπως απεικονίζεται στο διάγραμμα (35), το ποσοστό εργαζομένων που προτιμούν να παρουν λεωφορείο, αμάξι και να πάνε με τα πόδια στις δουλεία τους είναι 60.3% , 29.3% και 10.4% αντίστοιχα για το σύνολο δεδομένων της μελέτης μας.



Διάγραμμα 35 - Ραβδόγραμμα της μεταβλητής 'way'

Σύμφωνα με τα παρακάτω ραβδογράμματα του γραφήματος (36), η '*foot*' της μεταβλητής '*way*' αποτελεί την κατηγορία όπου εμφανίζει τις μεγαλύτερες διαφορές σχετικά με τους εργαζομένους που εκδήλωσαν το γεγονός και εκείνους που αποκόπηκαν από την μελέτη. Όπως φαίνεται και στο διάγραμμα (35), οι εργαζόμενοι που εκδήλωσαν το γεγονός και ανήκαν στην κατηγορία '*foot*' είναι το 7.5% του συνολικού αριθμού εργαζομένων που εκδήλωσαν το γεγονός. Αντίθετα, εκείνοι που αποκόπηκαν ενώ ανήκαν στην κατηγορία '*foot*' αποτελούν το 13.2% από τον συνολικό αριθμό εργαζομένων που αποκόπηκαν.



Διάγραμμα 36 - Ραβδόγραμμα της μεταβλητής 'way' για 'event = 0' και ' event = 1'

3.3.2.2. Στατιστικός έλεγχος ποιοτικών μεταβλητών

Προκειμένου να ελέγξουμε αν υπάρχει στατιστικά σημαντική συσχέτιση μεταξύ της κατηγορίας που ανήκει ο κάθε εργαζόμενος, δηλαδή εκείνοι που εκδήλωσαν το γεγονός του ενδιαφέροντος (*event = 1*) και εκείνοι που αποκόπηκαν από την μελέτη (*event = 0*), για κάθε μία ποιοτική μεταβλητή, θα χρησιμοποιήσουμε τον στατιστικό έλεγχο του Pearson '*Chi-squared test*'.

Παρατηρώντας τον παρακάτω πίνακα (3.3), υπάρχει στατιστικά σημαντική συσχέτιση μεταξύ των εργαζομένων που ανήκουν σε καθεμιά από τις δύο κατηγορίες, για τις μεταβλητές '*industry*' , '*profession*', '*traffic*' και '*way*'. Καταλήγουμε σε αυτό το συμπέρασμα για τον λόγο ότι η τιμή για το παρατηρηθέν επίπεδο σημαντικότητας (*p-value*) για κάθε μία από τις παραπάνω μεταβλητές είναι χαμηλότερη από την τιμή του αποδεκτού επιπέδου σημαντικότητας, με αποτέλεσμα να απορρίψουμε την μηδενική υπόθεση (H_0 = οι δύο κατηγορίες για την συγκεκριμένη μεταβλητή (*event = 0 και event = 1)* είναι ανεξάρτητες μεταξύ τους προς όφελος της H_1.

Αντίθετα, από τον πίνακα (3.3), φαίνεται πως οι δύο κατηγορίες εργαζομένων είναι ανεξάρτητες για τις μεταβλητές '*gender*', '*coach*', '*head_gender'* και '*greywage*'. Παρατηρούμε πως η τιμή για το παρατηρηθέν επίπεδο σημαντικότητας των συγκεκριμένων μεταβλητών είναι υψηλότερη από την τιμή του αποδεκτού επίπεδου σημαντικότητας της τάξεως του 5%, με αποτέλεσμα να αποτρέπεται η απόρριψη της μηδενικής υπόθεσης (H_0).

variables	x-squared	df	p-value
gender	0,40401	1	0,525
industry	64,419	15	4,32E-08
profession	44,898	14	4,24E-05
traffic	35,07	7	1,09E-05
coach	4,6684	2	9,69E-02
head_gender	1,9219	1	1,66E-01
greywage	2,7291	1	9,85E-02
way	10,009	2	6,71E-03

Πίνακας 3 - Στατιστικός έλεγχος 'Pearson's Chi-squared test' ποιοτικών μεταβλητών

4. Περιγραφή πειραμάτων στα δεδομένα της εργασίας

4.1. Kaplan-Meier

4.1.1. Περιγραφή ενδεχόμενων ειδικών βημάτων ανάλυσης

Η μέθοδος *Kaplan-Meier* αποτελεί την πιο απλουστευμένη τεχνική, η οποία θα χρησιμοποιηθεί σε αρχικό στάδιο προκειμένου να μοντελοποιήσουμε την διάρκεια ζωής των εργαζομένων. Τα μοναδικά στοιχεία που θα πρέπει να γνωρίζουμε προκειμένου να σχηματιστεί η καμπύλη επιβίωσης *Kaplan-Meier*, είναι η συνολική εμπειρία του εργαζομένου στην εταιρία και τον δείκτη που απεικονίζει αν εργαζόμενος εκδήλωσε το γεγονός του ενδιαφέροντος ή αποκόπηκε από την μελέτη την συγκεκριμένη χρονική στιγμή. Έπειτα με την μη παραμετρική μέθοδο *Kaplan-Meier* θα αποκτήσουμε μία καμπύλη επιβίωσης για όλους τους εργαζομένους του δείγματος. Στον κατακόρυφο άξονα δηλώνεται η πιθανότητα να μην εκδηλωθεί το γεγονός του ενδιαφέροντος *S(t)*, δηλαδή να μην αποχωρήσει από την εταιρία, για κάποιον εργαζόμενο μετά από κάποια ορισμένη χρονική στιγμή *t*. Στον οριζόντιο άξονα απεικονίζεται η διάρκεια *t* (μέρες, μήνες, χρόνια) που παρέμεινε ο εργαζόμενος στην εταιρία μέχρι να εκδηλώσει το γεγονός η να αποκοπεί από την μελέτη. Η πιθανότητα επιβίωσης μπορεί να υπολογιστεί ακολουθώντας τα εξής βήματα:

- Στην αρχή της μελέτης η πιθανότητα κάποιος να συνεχίσει να εργάζεται προς όφελος της εταιρίας είναι ίση με 1 (*S(0) = 1*). Το οποίο δικαιολογείται, καθώς κανένας εργαζόμενος δεν έχει αποχωρήσει από την εταιρία στην αρχή της μελέτης.
- Με το πέρασμα του χρόνου όλο και περισσότεροι εργαζόμενοι θα εκδηλώνουν το γεγονός ή θα αποκόπτονται από την μελέτη. Η πιθανότητα επιβίωσης είναι ο αριθμός εργαζομένων που έχουν απομείνει στην εταιρία μετά την συγκεκριμένη χρονική στιγμή, προς τον συνολικό αριθμό εργαζομένων που υπάρχουν στην μελέτη. Επίσης, αξίζει να σημειωθεί πως στον υπολογισμό της πιθανότητας επιβίωσης δεν συμπεριλαμβάνονται εργαζόμενοι που χαρακτηρίζονται από αποκοπή.

Κάθε φορά που κάποιος εργαζόμενος αποχωρεί από την εταιρία η πιθανότητα επιβίωσης *S(t)* μειώνεται, έως ότου να λήξει η μελέτη. Η μείωση της πιθανότητας αποτυπώνεται από την πτώση στην καμπύλη *Kaplan-Meier* την ορισμένη χρονική στιγμή. Η πτώση στην καμπύλη διαμορφώνεται ανάλογα με την εκδήλωση του γεγονότος του ενδιαφέροντος την συγκεκριμένη χρονική στιγμή. Όσο περισσότεροι εργαζόμενοι αποχώρησαν από την εταιρία την τάδε χρονική στιγμή *t*, τόσο μεγαλύτερη θα είναι και η πτώση στην καμπύλη.

Στην συνέχεια θα διαχωρίσουμε τους εργαζομένους σε διαφορετικές κατηγορίες με βάση κάποιο χαρακτηριστικό τους, προκειμένου να μπορέσουμε να διακρίνουμε ποια χρονική στιγμή και

πόσοι εργαζόμενοι εκδήλωσαν το γεγονός για την κάθε μια κατηγορία. Δημιουργώντας από μία καμπύλη *Kaplan-Meier* για την κάθε κατηγορία που έχει διαχωριστεί το σύνολο δεδομένων μας, θα διακρίνουμε για ποια από τις καμπύλες που ανήκει κάποιος εργαζόμενος, η πιθανότητα του να παραμείνει στην εταιρία (χρόνος επιβίωσης) είναι μεγαλύτερη. Δηλαδή, οι εργαζόμενοι κάποιας κατηγορίας μπορεί να μην έχουν αποχωρήσει ακόμα από την εταιρία ή να εκδήλωσαν το γεγονός αργότερα από εκείνους που ανήκουν σε κάποια άλλη κατηγορία. Αυτό έχει ως αποτέλεσμα, η καμπύλη *Kaplan-Meier* της συγκεκριμένης κατηγορίας να είναι υψηλότερα από τις υπόλοιπες καμπύλες. Τέλος, θα χρησιμοποιήσουμε το *log-rank test* προκειμένου να ελέγξουμε αν υπάρχει στατιστικά σημαντική διαφορά μεταξύ των καμπυλών που δημιουργήθηκαν. Προκειμένου να υπολογίσουμε την τιμή του log-rank statistic θα πρέπει να ακολουθήσουμε τα εξής βήματα:

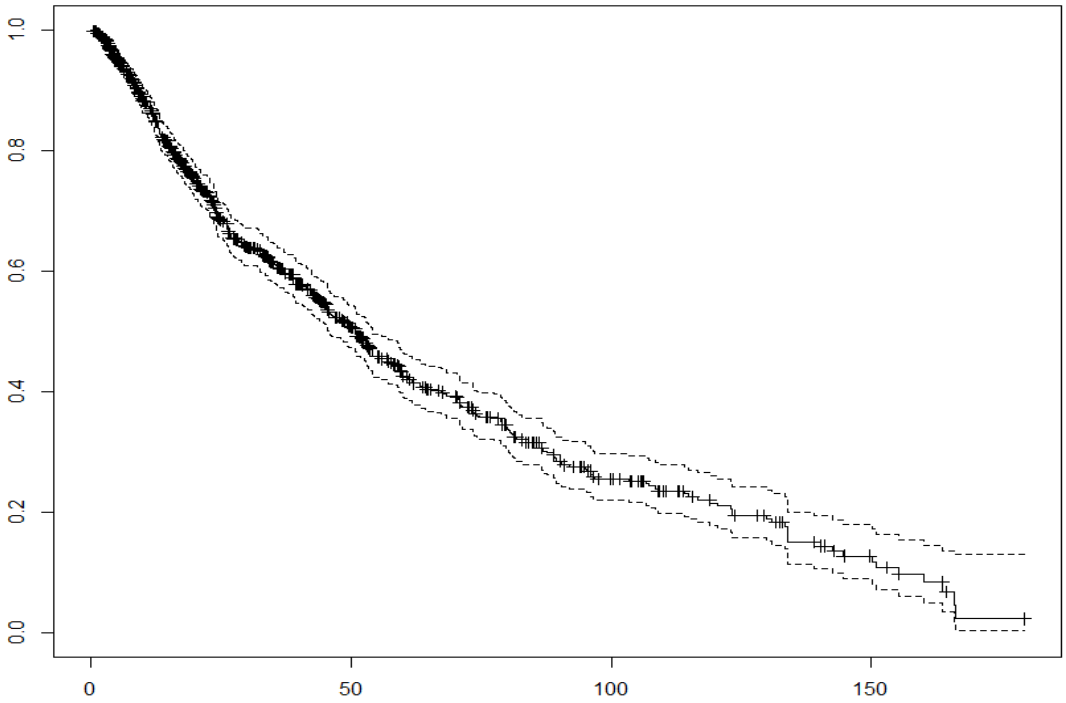
- Αρχικά θα πρέπει να υπολογίσουμε τις αναμενόμενες τιμές (*expected values*) για την καθεμία κατηγορία, σχετικά με την εκδήλωση του ενδιαφέροντος την κάθε χρονική στιγμή που συνέβη το γεγονός. Ύστερα το σύνολο των εργαζομένων της κάθε κατηγορίας που έχει παρατηρηθεί ότι έχουν αποτύχει (*observed values*) θα πρέπει να αφαιρεθεί από το σύνολο εκείνων που αναμέναμε (*expected value*) να εκδηλώσουν το γεγονός την συγκεκριμένη χρονική στιγμή.
- Στην συνέχεια θα αθροίσουμε το αποτέλεσμα των (*observed – expected values*) για κάθε χρονική στιγμή στην καθεμιά κατηγορία του συνόλου δεδομένων μας. Το αποτέλεσμα θα το διαιρέσουμε με την αναμενόμενη τιμή της εκδήλωσης του ενδιαφέροντος της καθεμίας κατηγορίας που ασχολούμαστε. Τέλος, αθροίζοντας τα *log-rank statistic* όλων των κατηγοριών που έχουμε υπολογίσει, θα αποκτήσουμε το τελικό *log-rank statistic*.

Επειδή η τιμή του *log-rank statistic* ακολουθεί την *Chi-Square* κατανομή, εύκολα μπορούμε να υπολογίσουμε την τιμή του παρατηρηθέν επιπέδου σημαντικότητας (*p-value*). Αναλόγως με το αποδεκτό επίπεδο σημαντικότητας που έχουμε θέσει, θα αποδεχτούμε ή θα απορρίψουμε την μηδενική υπόθεση (H_0 = ο χρόνος επιβίωσης μεταξύ των κατηγοριών είναι ο πανομοιότυπος) προς όφελος της εναλλακτικής υπόθεσης (H_1 = ο χρόνος επιβίωσης μεταξύ των κατηγοριών διαφέρει). Τέλος, αξίζει να σημειωθεί πως ασχολούμαστε με την περίπτωση που έχουμε διαχωρίσει το σύνολο δεδομένων μεταξύ δύο κατηγοριών. Η δημιουργία περισσότερων κατηγοριών δυσχεραίνει πολύ περισσότερο την κατάσταση καθώς αυξάνεται η πολυπλοκότητα.

4.1.2. Παρουσίαση και σχολιασμό αποτελεσμάτων

Αρχικά από τα δεδομένα που μας παρέχονται, θα πρέπει να διακρίνουμε τους εργαζομένους τους εκδήλωσαν το γεγονός του ενδιαφέροντος και εκείνους που αποκόπηκαν από την μελέτη. Όπως

είχαμε παρατηρήσει και στο προηγούμενο κεφάλαιο, από τους 1129 εργαζομένους του συνόλου δεδομένων μας, οι 558 είναι εκείνοι που σταμάτησε η παρακολούθηση τους την ορισμένη χρονική στιγμή *t* ή έληξε η μελέτη και δεν αποχώρησαν από την εταιρία. Οι υπόλοιποι 571 είναι εκείνοι οι εργαζόμενοι οι οποίοι αποχώρησαν από την εταιρία κάποια χρονική περίοδο κατά την διάρκεια της μελέτης. Για να μπορέσουμε να συνεχίσουμε την ανάλυση, θα πρέπει να μετατρέψουμε την μεταβλητή '*event*' σε '*numeric*' από '*factor*' που ήταν προηγουμένως. Εκτελώντας την μέθοδο *Kaplan-Meier* αποκτάμε μία καμπύλη επιβίωσης για όλους τους εργαζομένους του δείγματος, όπως παρουσιάζεται και στο διάγραμμα (37).



Διάγραμμα 37 - Καμπύλη Kaplan-Meier των εργαζομένων του δείγματος

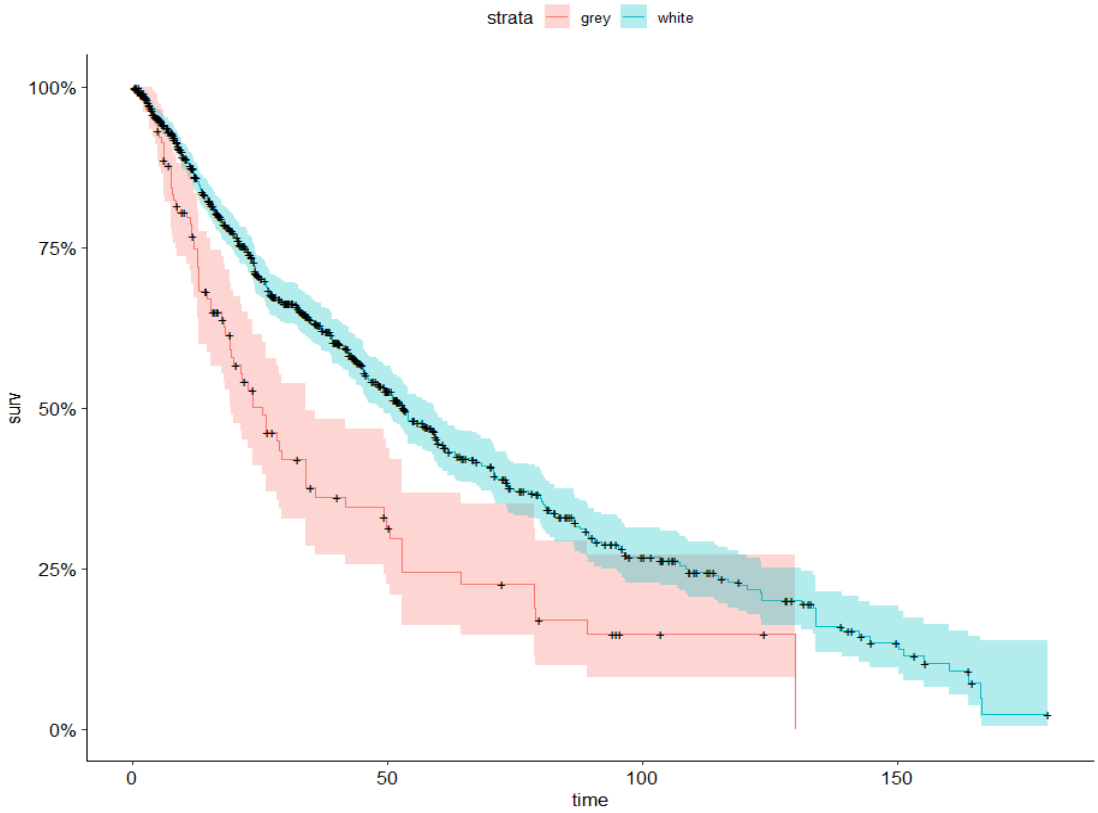
Σύμφωνα με το παραπάνω διάγραμμα, παρατηρούμε ότι η πιθανότητα να παραμείνει κάποιος εργαζόμενος στην εταιρία *S(t)* μειώνεται δραστικά με το πέρασμα του χρόνου. Ξεκινώντας από τους πρώτους μήνες μέχρι την χρονική στιγμή των 100 μηνών, παρουσιάζεται κατακόρυφη πτώση στην καμπύλη *Kaplan-Meier*, καθώς ολοένα και περισσότεροι εργαζόμενοι φαίνεται να εκδήλωσαν το γεγονός του ενδιαφέροντος κατά

την διάρκεια αυτού του χρονικού διαστήματος. Η πιθανότητα να παραμείνει κάποιος εργαζόμενος στην εταιρία (πιθανότητα επιβίωσης) μετά την χρονική στιγμή των 100 μηνών φθάνει κοντά στα επίπεδα της τάξεως του 30%. Ενώ η πιθανότητα επιβίωσης κάποιου εργαζομένου μετά την χρονική περίοδο των 50 μηνών φαίνεται να είναι κοντά στα επίπεδα του

50%. Έπειτα από την χρονική περίοδο των 100 μηνών η κατάσταση φαίνεται να σταθεροποιείται κάπως, καθώς η εκδήλωση του γεγονότος από τους εργαζομένους είναι μικρότερη. Αυτό εξακολουθεί να συμβαίνει έως ότου φτάσουμε στην χρονική περίοδο των (170 – 180) μηνών όπου φαίνεται πως όλοι οι εργαζόμενοι έχουν αποχωρήσει από την εταιρία, με την πιθανότητα επιβίωσης για μετά την συγκεκριμένη χρονική περίοδο να αγγίζει το 0%. Τα σύμβολα (+) πάνω στην καμπύλη *Kaplan-Meier*, απεικονίζουν τους εργαζόμενους που έχουν αποκοπεί από την μελέτη την δεδομένη χρονική περίοδο. Στο διάγραμμα (4.1) φαίνεται επίσης πως ένας εργαζόμενος αποκόπηκε στο τέλος της μελέτης, την χρονική στιγμή 180 μηνών περίπου. Επιπλέον, οι διακεκομμένες γραμμές που περιστοιχίζουν την καμπύλη *Kaplan-Meier* δηλώνουν το διάστημα όπου μπορεί να μετακινηθεί η παραπάνω καμπύλη επιβίωσης.

Σύμφωνα με το διάγραμμα (34) του προηγουμένου κεφαλαίου, διακρίνουμε πως υπάρχει κάποια διαφορά μεταξύ των μισθών που λαμβάνουν οι εργαζόμενοι (*greywage*) όσον αφορά την αποκοπή τους από την μελέτη και την αποχώρηση τους από την εταιρία. Από τα υπόλοιπα γραφήματα του προηγούμενου κεφαλαίου, βλέπουμε πως ίσως η μεγαλύτερη διαφορά όσον αφορά την αποκοπή και την εκδήλωση του γεγονότος από τους εργαζομένους συμβαίνει για την μεταβλητή '*greywage*'. Στην συνέχεια θα αποκτήσουμε τις δύο καμπύλες επιβίωσης για τις δύο κατηγορίες εργαζομένων αναλόγως με τον μισθό που λαμβάνουν στην εταιρία. Δηλαδή, η μία κατηγορία αποτελείται από εκείνους που αμείβονται ελάχιστα παραπάνω από τον καθαρό μισθό που δικαιούνται (*grey*) και στην δεύτερη κατηγορία ανήκουν εκείνοι που αμείβονται μόνο με τον καθαρό μισθό τους (*white*). Όπως παρατηρούμε στο διάγραμμα (38) οι δύο καμπύλες επιβίωσης συναντιούνται για ένα πολύ μικρό χρονικό διάστημα στην αρχή της μελέτης και έπειτα ο διαχωρισμός τους της μίας από την άλλη είναι ξεκάθαρος. Η καμπύλη επιβίωσης εκείνων που αμείβονται αποκλειστικά και μόνο από τον καθαρό τους μισθό είναι υψηλότερα από την καμπύλη της δεύτερης κατηγορίας εργαζομένων. Συμπεραίνουμε πως η επιβίωση (παραμονή τους στην εταιρία) των εργαζομένων που λαμβάνουν μόνο των καθαρό τους μισθό είναι μεγαλύτερη από εκείνους που αμείβονται ελάχιστα παραπάνω από τον βασικό μισθό. Ο διάμεσος χρόνος επιβίωσης (η χρονική στιγμή όπου η πιθανότητα επιβίωσης ενός εργαζομένου, που ανήκει σε κάποια από τις δύο κατηγορίες, μετά την συγκεκριμένη χρονική στιγμή ισούται με $S(t) = 0.50$) για την καμπύλη των '*grey*' εργαζομένων είναι κοντά στους 25 μήνες, ενώ για τους '*white*' ο διάμεσος χρόνος επιβίωσης είναι ισούται με 53 μήνες (σχεδόν διπλάσιος από την άλλη κατηγορία εργαζομένων). Βλέπουμε πως στην καμπύλη των '*grey*' εργαζομένων, το πότε κάποιος εργαζόμενος εκδήλωσε το γεγονός του ενδιαφέροντος είναι πιο ξεκάθαρο, καθώς είναι ευκολότερο να διακρίνουμε τους βηματισμούς πάνω στην καμπύλη. Θα μπορούσαμε να πούμε πως αυτό συμβαίνει λόγω του εξαιρετικά μικρού αριθμού εργαζομένων που αμείβονται ελάχιστα παραπάνω από τον καθαρό τους μισθό (συνολικά 127 άτομα) σε σχέση με εκείνους που

λαμβάνουν μόνο τον κανονικό τους μισθό (συνολικά 1002 άτομα). Τέλος, βλέπουμε πως υπήρξε μία αυξημένη αποχώρηση '*grey*' εργαζομένων για τους πρώτους 50 μήνες της μελέτης, καθώς διακρίνουμε μία κατακόρυφη πτώση στην καμπύλη επιβίωσης για τους πρώτους 50 μήνες ενώ από εκεί και έπειτα η κατάσταση φαίνεται να σταθεροποιείται. Όσον αφορά την καμπύλη επιβίωσης των '*white*' εργαζομένων, βλέπουμε πως υπάρχει μια αυξημένη αποχώρηση εργαζομένων μέχρι την χρονική περίοδο των 100 μηνών, ενώ από εκεί και έπειτα η ελαττώνεται κάπως η πτώση στην καμπύλη.



Διάγραμμα 38 - Καμπύλες Kaplan-Meier με βάση την μεταβλητή 'greywage'

Τέλος θα χρησιμοποιήσουμε τον στατιστικό έλεγχο *log-rank test*, προκειμένου να ελέγξουμε αν αυτή διαφορά μεταξύ των δύο καμπυλών *Kaplan-Meier* που δημιουργήθηκαν είναι και στατιστικά σημαντική. Σύμφωνα με τον πίνακα (4.1), η διαφορά μεταξύ των δύο καμπυλών *Kaplan-Meier* όσον αφορά τον μισθό που λαμβάνει ο κάθε εργαζόμενος φαίνεται να είναι στατιστικά σημαντική. Διακρίνουμε πως η καμπύλη επιβίωσης των '*white*' εργαζομένων διχοτομεί μόνο στην αρχή την καμπύλη των '*grey*' εργαζομένων, πράγμα που δεν επηρεάζει το *log-rank test* ως προς την διαφορά μεταξύ των καμπυλών. Αξίζει να σημειωθεί ότι, σε περίπτωση που οι δύο καμπύλες διχοτομούνται στα πρώιμα στάδια της μελέτης, επιφέρει μεγαλύτερη στατιστική σημαντικότητα όσον αφορά την διαφορά των δύο καμπυλών σε σχέση με το να συμβεί στο τέλος της μελέτης. Αυτό μπορεί να ερμηνευθεί για τον λόγο ότι στην αρχή της μελέτης ο

αριθμός των ατόμων που βρίσκονται σε κίνδυνο είναι πολύ περισσότερος και η εκδήλωση ενδιαφέροντος επιφέρει μικρότερες αλλαγές στον ρυθμό κινδύνου (*hazard rate*), από ότι οι καμπύλες να συναντηθούν προς το τέλος της μελέτης. Η τιμή για το παρατηρηθέν επίπεδο σημαντικότητας είναι μικρότερη από την τιμή του αποδεκτού επιπέδου σημαντικότητας της τάξεως του 0.1%, οπότε μπορούμε να απορρίψουμε την μηδενική υπόθεση (H_0 = ο χρόνος επιβίωσης και για τις δύο καμπύλες είναι ο ίδιος) προς όφελος της εναλλακτικής υποθέσεως H_1. Τέλος αξίζει να σημειωθεί ότι ο αριθμός των εργαζομένων που αμείβονται ελάχιστα παραπάνω από τον καθαρό τους μισθό και βρίσκονταν σε κίνδυνο να αποχωρήσουν από την εταιρία, αναμέναμε να ήταν μικρότερος από τον αριθμό εκείνων που εν τέλει αποχώρησαν. Ενώ ο αριθμός των εργαζομένων που αμείβονται μόνο με τον καθαρό τους μισθό και βρίσκονταν σε κίνδυνο να αποχωρήσουν, αναμέναμε να ήταν μεγαλύτερος από τον αριθμό εκείνων που τελικά εκδήλωσαν το γεγονός. Κάτι το οποίο αποτυπώνεται και στο διάγραμμα (4), όπου η καμπύλη *Kaplan-Meier* των '*white*' εργαζομένων είναι πιο πάνω από την καμπύλη *Kaplan-Meier* των '*grey*' εργαζομένων.

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
turn$greywage=grey	127	73	43.2	20.47	22.3
turn$greywage=white	1002	498	527.8	1.68	22.3

Chisq= 22.3 on 1 degrees of freedom, p= 2e-06

Πίνακας 4 - Log Rank Test για τον έλεγχο ομοιότητας μεταξύ των δύο καμπυλών

4.2. Cox PH

4.2.1. Περιγραφή ενδεχόμενων ειδικών βημάτων ανάλυσης

Ένα από τα βασικά μοντέλα στην ανάλυση επιβίωσης όπου θα χρησιμοποιήσουμε αργότερα, ίσως και το δημοφιλέστερο, είναι το *Cox PH* μοντέλο. Σε αντίθεση με το παραμετρικό μοντέλο *Kaplan-Meier* που χρησιμοποιήσαμε στην προηγούμενη ενότητα, το *Cox PH* αποτελεί ημί-παραμετρικό μοντέλο. Λαμβάνοντας υπόψιν, την επίδραση όλων των μεταβλητών που συμπεριλαμβάνονται στο σύνολο δεδομένων μας όσον αφορά την εκδήλωση του γεγονότος από τους εργαζομένους. Αρχικά, εκτελώντας το συγκεκριμένο μοντέλο αποκτάμε την συνολική συνάρτηση κινδύνου (*hazard function*) όλων των εργαζομένων του συνόλου δεδομένων. Με τον όρο κίνδυνο *h(t)* δηλώνεται ο ρυθμός (*rate*) του κάθε εργαζομένου να αποχωρήσει από την εταιρία την συγκεκριμένη χρονική στιγμή *t*. Δεν μπορούμε να πούμε πως εκφράζει πιθανότητα

καθώς ο κίνδυνος να αποχωρήσει κάποιος εργαζόμενος την χρονική στιγμή *t*, λαμβάνει τιμές μεταξύ του διαστήματος (0, +∞). Στο τέλος, αφού θα έχουμε υπολογίσει τον στιγμιαίο κίνδυνο για τον κάθε ένα εργαζόμενο να εκδηλώσει το γεγονός, θα μπορούμε να αποκτήσουμε την συνολική συνάρτηση κινδύνου. Ο κίνδυνος κάποιος εργαζόμενος να εκδηλώσει το γεγονός (*h(t)*) αποτελεί την εξαρτημένη μεταβλητή του *Cox PH* μοντέλου. Οι υπόλοιπες μεταβλητές του συνόλου δεδομένων που προσδιορίζουν συγκεκριμένα χαρακτηριστικά για τον κάθε ένα εργαζόμενο, αποτελούν τις επεξηγηματικές μεταβλητές του μοντέλου. Λόγω της σχέσης που υπάρχει μεταξύ της συνάρτηση κινδύνου και της συνάρτησης επιβίωσης, μπορούμε εύκολα στην περίπτωση που γνωρίζουμε την συνάρτηση κινδύνου να αναταράξουμε την συνάρτηση επιβίωσης.

Αρχικά προκειμένου να μοντελοποιήσουμε την διάρκεια του κύκλου ζωής των εργαζομένων θα χρησιμοποιήσουμε το *Cox PH* μοντέλο. Αργότερα θα πρέπει να διακρίνουμε εκείνες τις μεταβλητές που επηρέασαν περισσότερο το αποτέλεσμα. Δηλαδή ποιες μεταβλητές είχαν μεγαλύτερη επιρροή για την εκδήλωση του γεγονότος από τον εργαζόμενο. Παρατηρώντας την στατιστική σημαντικότητα κάθε μεταβλητής για την δημιουργία του μοντέλου, θα μπορέσουμε να διαχωρίσουμε τις σημαντικές μεταβλητές στην δημιουργία του μοντέλου από τις μη στατιστικά σημαντικές. Βέβαια, υπάρχει η περίπτωση κάποια ή κάποιες από τις κλάσεις μίας κατηγορικής μεταβλητής του υποδείγματος να διαδραματίζει ουσιαστικό ρόλο σχετικά με την έκβαση του αποτελέσματος, ενώ η συγκεκριμένη κατηγορική να εμφανίζεται ως μη στατιστικά σημαντική. Προκειμένου να μην χαθεί αυτή η πληροφορία από το μοντέλο που δημιουργήθηκε αλλά και για την καλύτερη ερμηνεία των αποτελεσμάτων, η δημιουργία ψευδομεταβλητών (*dummies*) είναι απαραίτητη. Θα πρέπει να δημιουργηθεί από μία ψευδομεταβλητή για την κάθε κλάση των κατηγορικών μεταβλητών του υποδείγματος. Στο καινούριο *Cox PH* μοντέλο θα χρησιμοποιηθούν οι ψευδομεταβλητές που δημιουργήσαμε, ως επεξηγηματικές μεταβλητές του υποδείγματος και θα υπολογιστεί η στατιστική σημαντικότητα για την κάθε μία μεταβλητή ξεχωριστά. Από τις ψευδομεταβλητές που σχηματίστηκαν θα πρέπει να αφαιρεθεί από την δημιουργία του μοντέλου, μία ψευδομεταβλητή που αντιστοιχεί σε μία τυχαία κλάση της κάθε κατηγορικής μεταβλητής. Η ψευδομεταβλητή της κάθε κατηγορικής μεταβλητής αφαιρέθηκε με σκοπό την σύγκριση, ως προς τον κίνδυνο να εκδηλώσει το γεγονός κάποιος εργαζόμενος, με τις υπόλοιπες ψευδομεταβλητές που αποτελούν κλάσεις της ίδιας κατηγορικής μεταβλητής με εκείνη. Στην συνέχεια θα ξεχωρίσουμε τις πιο σημαντικές μεταβλητές που ασκούν την μεγαλύτερη επίδραση στην εξαρτημένη μεταβλητή από τις λιγότερο στατιστικά σημαντικές μεταβλητές. Αξίζει να σημειωθεί πως κάποια επεξηγηματική μεταβλητή που πριν την χρησιμοποίηση των ψευδομεταβλητών στο μοντέλο θεωρούνταν ως μη στατιστικά σημαντική, μπορεί να απέκτησε κάποια στατιστική σημαντικότητα στο καινούριο μοντέλο, καθώς και το αντίστροφο. Επίσης, μετά την μετατροπή μπορεί να βελτιωθεί η ερμηνευτική

αποτελεσματικότητα (*concordance*) του μοντέλου όπως και το αντίθετο βέβαια. Στην συνέχεια θα επιλέξουμε εκείνες τις μεταβλητές που θα κρατήσουμε στο μοντέλο που δημιουργήσαμε χρησιμοποιώντας τρεις διαφορετικές τεχνικές:

- Μέθοδος της προς τα εμπρός επιλογής (*forward stepwise*). Σύμφωνα με την συγκεκριμένη μέθοδο, στο μοντέλο αρχικά συμπεριλαμβάνεται μόνο ο σταθερός όρος. Δηλαδή το μοντέλο που είχαμε αποκτήσει με την μέθοδο *Kaplan-Meier* στην αρχή της ανάλυσης. Στην συνέχεια προσθέτει τις μεταβλητές, ξεκινώντας από εκείνη που μας δίνει την μικρότερη τιμή του κριτηρίου σύγκρισης *AIC* ή *BIC*. Η διαδικασία επαναλαμβάνεται μέχρι να καταλήξουμε στην βέλτιστη (μικρότερη) τιμή που μπορεί να λάβει το κριτήριο σύγκρισης *AIC*. Η διαδικασία τερματίζεται, φθάνοντας στο καταληκτικό μοντέλο, στο σημείο εκείνο όπου με κάθε νέα προσθήκη μεταβλητής θα αυξάνεται η τιμή του *AIC*.
- Μέθοδος της προς τα πίσω απαλοιφής (*backward stepwise*). Αποτελεί την ακριβώς αντίθετη μέθοδο της προς τα εμπρός επιλογής (*forward stepwise*) καθώς υλοποιείται με αντίστροφη φορά. Αρχικά στο μοντέλο συμπεριλαμβάνονται όλες οι επεξηγηματικές μεταβλητές (*saturated model*). Στην συνέχεια αφαιρούνται μία μία οι επεξηγηματικές μεταβλητές, ξεκινώντας από εκείνη όπου η αφαίρεση της από το μοντέλο θα επιφέρει την μικρότερη τιμή στο κριτήριο σύγκρισης *AIC* ή *BIC*. Η διαδικασία επαναλαμβάνεται μέχρι να καταλήξουμε στο μοντέλο με την βέλτιστη (μικρότερη) τιμή *AIC*. Δηλαδή η διαδικασία τερματίζεται φθάνοντας στο καταληκτικό μοντέλο, στο σημείο εκείνο που η αφαίρεση κάποιας επιπλέον επεξηγηματικής μεταβλητής θα επιφέρει αύξηση στο κριτήριο *AIC*.
- Μέθοδος της κατά βήματα παλινδρόμησης (*both stepwise*). Βασικό μειονέκτημα των δύο προηγούμενων μεθόδων το οποίο καλείται να επιλύσει η 'both stepwise', αποτελεί το γεγονός ότι από την στιγμή που εισαχθεί κάποια μεταβλητή στο μοντέλο μέσω της '*forward stepwise*' δεν μπορεί να εξαχθεί στην συνέχεια ακόμα και αν πλέον είναι ασήμαντη για την πρόβλεψη της εξαρτημένης μεταβλητής. Παρομοίως και για την '*backward stepwise*', από την στιγμή που θα αφαιρεθεί κάποια μεταβλητή από το μοντέλο ως ασήμαντη, δεν γίνεται στην συνέχεια να συμπεριληφθεί στο μοντέλο παρόλο που μπορεί καταστεί σημαντική αργότερα. Μέσω της '*both stepwise*', υλοποιείται ταυτόχρονα η διαδικασία της προς τα μπρος επιλογής αλλά και της προς τα πίσω απαλοιφής. Η συγκεκριμένη μέθοδος ξεκινάει όπως ακριβώς η προς τα εμπρός επιλογή και στην συνέχεια αφού έχει εισαχθεί κάποιος αριθμός μεταβλητών, ελέγχεται αν μπορούν να εξαχθούν μεταβλητές που υπάρχουν ήδη στο μοντέλο. Η διαδικασία αυτή επαναλαμβάνεται έως ότου αποκτηθεί η βέλτιστη (μικρότερη) τιμή στα κριτήρια πληροφορίας *AIC* ή *BIC*. Η διαδικασία αυτή της ταυτόχρονης εισαγωγής και εξαγωγής μεταβλητών, συνεχίζεται έως ότου να μην υπάρχει καμία μεταβλητή εκτός μοντέλου που

μπορεί να καταστεί σημαντική και καμία μεταβλητή που συμπεριλαμβάνεται στο μοντέλο και δεν συμβάλλει στην ερμηνεία της εξαρτημένης μεταβλητής.

Στο τέλος θα συγκρίνουμε όλες τις μεθόδους που χρησιμοποιήθηκαν παραπάνω, προκείμενου να αξιολογήσουμε την αποτελεσματικότητα της κάθε μίας και αν καταλήγουν στα ίδια αποτελέσματα. Δηλαδή οι μεταβλητές που εξάγονται από το μοντέλο είναι ίδιες ή διαφέρουν για την κάθε μία μέθοδο; Τα τρία μοντέλα που αποκτήσαμε για την κάθε μία μέθοδο καταλήγουν στην ίδια ή διαφορετική τιμή *AIC*;

Σε τελικό στάδιο θα αναπαραχθεί γραφικά η καμπύλη που αποκτήθηκε με την χρήση του Cox PH μοντέλου στο σύνολο εργαζομένων. Στον κάθετο άξονα θα υπάρξει η πιθανότητα του εργαζόμενου να παραμείνει στην εταιρία μετά από κάποια ορισμένη χρονική στιγμή, ύστερα από μετατροπή της συνάρτησης κινδύνου για τον κάθε ένα εργαζόμενο που έχουμε αποκτήσει στην συνάρτηση επιβίωσης των εργαζομένων, και στο οριζόντιο άξονα αναπαρίσταται οι μήνες που παρέμεινε ο εργαζόμενος μέχρι να αποχωρήσει από την εταιρία. Θα αποκτήσουμε δύο καμπύλες επιβίωσης. Μία για το αρχικό μοντέλο και μία για το μοντέλο που αποκτήσαμε μετά την εισαγωγή των ψευδομεταβλητών και των μεθόδων απαλοιφής των μη στατιστικά σημαντικών μεταβλητών. Επίσης, θα συγκρίνουμε τις δύο καμπύλες και τα αποτελέσματα που εξήχθησαν για τα δύο μοντέλα. Στο τέλος για τα δύο μοντέλα που δημιουργήθηκαν, θα ελέγξουμε αν ικανοποιείται το κριτήριο της αναλογίας κινδύνου (*proportional hazard assumption*). Η αναλογία κινδύνου (*hazard ratio*) των επεξηγηματικών μεταβλητών του υποδείγματος θα πρέπει να είναι σταθερή καθ' όλη την διάρκεια της μελέτης. Τα *'Scaled Schoenfeld tests'* αποτελούν στατιστικούς ελέγχους, περιλαμβάνοντας γραφικές αναπαραστάσεις, που θα χρησιμοποιήσουμε προκειμένου να ελεγχθεί αν ικανοποιείται η θεώρηση της σταθερής αναλογίας κινδύνου (*proportional hazard assumption*) για τα δύο μοντέλα που θα δημιουργηθούν.

Επίσης θα χρησιμοποιηθεί η παλινδρόμηση του '*Aalen*' προκειμένου να ελεγχθεί η επιρροή της κάθε μίας μεταβλητής στην εκούσια αποχώρηση του εργαζομένου από την εταιρία. Η επιπρόσθετη παλινδρόμηση '*additive regression*' είναι μία εναλλακτική η συμπληρωματική μέθοδο της '*Cox*' παλινδρόμησης. Σημαντικός περιορισμός της '*Cox*' παλινδρόμησης αποτελεί το γεγονός ότι είναι αναποτελεσματική η χρήση της στην περίπτωση που η επίδραση των επεξηγηματικών μεταβλητών διαφέρει κατά την πάροδο του χρόνου. Προϋπόθεση που πρέπει να πληρείται όταν χρησιμοποιούμε την '*Cox*' είναι ότι η επίδραση των μεταβλητών στον ρυθμό του κινδύνου, παραμένει σταθερή κατά την διάρκεια της μελέτης. Για την λύση αυτού του προβλήματος έχουν αναπτυχθεί πολλές προσεγγίσεις, όπου μία από αυτές είναι και η '*Aalens additive regression*'. Με την χρήση της συγκεκριμένης μεθόδου μπορούμε να

διακρίνουμε αν η επίδραση της καθεμίας επεξηγηματικής μεταβλητής στον ρυθμό κινδύνου παραμένει σταθερή ή κατά πόσο διαφέρει κατά την διάρκεια της μελέτης. Η επίδραση των μεταβλητών μέσω της συγκεκριμένης μεθόδου δεν υπολογίζεται σε απόλυτη αλλά σε μία σχετική κλίμακα.

4.2.2. Παρουσίαση και σχολιασμός αποτελεσμάτων

Από την εκτέλεση της '*Cox*' παλινδρόμησης, προκειμένου να εκτιμηθεί ο κίνδυνος εκούσιας εγκατάλειψης της εταιρίας από τον εργαζόμενο την ορισμένη χρονική στιγμή *t*, χρησιμοποιώντας το αρχικό μοντέλο με τις 14 επεξηγηματικές μεταβλητές του συνόλου δεδομένων που διαθέτουμε (προτού γίνει ο διαχωρισμός των ποιοτικών μεταβλητών σε ψευδομεταβλητές), εξάγονται τα εξής αποτελέσματα:

```
               LR Chisq  Df  Pr(>Chisq)
gender          0.769     1  0.3806690
age            10.373     1  0.0012789 **
industry       54.520    15  2.150e-06 ***
profession     29.334    14  0.0094162 **
traffic        36.146     7  6.802e-06 ***
coach           2.223     2  0.3290862
head_gender     0.434     1  0.5101368
greywage       12.143     1  0.0004926 ***
way             7.014     2  0.0299849 *
extraversion    0.211     1  0.6462638
independ        0.155     1  0.6938973
selfcontrol     1.950     1  0.1625772
anxiety         2.540     1  0.1110190
novator         0.050     1  0.8239209

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

Πίνακας 5 - Ανάλυση διακύμανσης των μεταβλητών που χρησιμοποιήθηκαν στο 'Cox PH' μοντέλο

Σύμφωνα με τον πίνακα (5) παρατηρούμε πως οι επεξηγηματικές μεταβλητές '*age*', '*industry*', '*profession*', '*traffic*', '*greywage*' και '*way*' είναι στατιστικά σημαντικές στην δημιουργία του '*Cox*' μοντέλου. Οι κατηγορικές μεταβλητές '*industry*', '*traffic*' και '*greywage*' φαίνεται να είναι εκείνες που επηρεάζουν σε μεγαλύτερο βαθμό την εκδήλωση του γεγονότος από τον εργαζόμενο. Η τιμή για το παρατηρηθέν επίπεδο σημαντικότητας (*p-value*) του εκτιμώμενου συντελεστή (β_i) των συγκεκριμένων μεταβλητών λαμβάνει τιμή μικρότερη από το αποδεκτό επίπεδο σημαντικότητας της τάξεως του 0.1%, απορρίπτοντας την μηδενική υπόθεση (H_0 = η τιμή του συντελεστή (β_i) για την συγκεκριμένη μεταβλητή είναι ίση με το 0) σε επίπεδο σημαντικότητας 0.1%. Ακολουθούν οι μεταβλητές '*age*' και '*profession*' για τις οποίες η μηδενική υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας 1%. Τέλος η '*way*' φαίνεται να επηρεάζει λιγότερο την

εξαρτημένη μεταβλητή από τις προηγούμενες, καθώς η μηδενική υπόθεση απορρίπτεται για επίπεδο σημαντικότητας της τάξεως του 5%.

Ύστερα στο από το αρχικό '*Cox*' μοντέλο, εισάγοντας τις ψευδομεταβλητές που έχουν δημιουργηθεί για την κάθε μία κλάση των κατηγορικών μεταβλητών του υποδείγματος, εξάγονται τα εξής αποτελέσματα, όπως παρουσιάζονται στον πίνακα (6). Σύμφωνα με τον παρακάτω πίνακα (6), στο '*Cox*' μοντέλο συμπεριλαμβάνονται συνολικά 49 επεξηγηματικές μεταβλητές. Εξαιρούνται οι μεταβλητές '*gender_f*', '*industry_HoReCa*' , '*profession_Accounting*', '*traffic_advert*', '*coach_myhead*', '*head_gender_f*', '*greywage_grey*' και '*way_bus*' οι οποίες εξήφθηκαν από την κατασκευή του μοντέλου προκειμένου να συγκρίνουμε τις ψευδομεταβλητές του μοντέλου με κάθε μία από της παραπάνω εξαγόμενες μεταβλητές, όπου για να γίνει η σύγκριση θα πρέπει να αποτελούν κλάσεις της ίδιας κατηγορικής μεταβλητής, ως προς τον κίνδυνο εκδήλωσης του γεγονότος από τον εργαζόμενο. Διακρίνουμε μόλις μία επεξηγηματική μεταβλητή (*greywage_white*) που φαίνεται να επηρεάζει σημαντικά τον κίνδυνο κάποιος εργαζόμενος να αποχωρήσει με την θέληση του από την εταιρία. Η '*greywage_white*' είναι στατιστικά σημαντική σε επίπεδο σημαντικότητας της τάξεως του 0.1%. Οι μεταβλητές '*age*' και '*traffic_empjs*' φαίνεται πως είναι οι αμέσως επόμενες μεταβλητές που επηρεάζουν σημαντικά την εξαρτημένη μεταβλητή. Οι '*age*' και '*traffic_empjs*' είναι στατιστικά σημαντικές σε επίπεδο σημαντικότητας της τάξεως του 1%. Οι '*way_foot*', '*traffic_youjs*' και '*profession_manage*' φαίνεται να είναι οι επόμενες πιο σημαντικές μεταβλητές, καθώς είναι στατιστικά σημαντικές για επίπεδο σημαντικότητας της τάξεως του 5%. Τέλος, οι '*way_car*', '*traffic_rabrecNErab*', '*profession_Engineer*', '*profession_Commercial*' και '*industry_RealEstate*' είναι στατιστικά σημαντικές σε επίπεδο σημαντικότητας της τάξεως του 10%.

Ο συντελεστής των επεξηγηματικών μεταβλητών (*coef*), εκφράζει τον βαθμό που η συγκεκριμένη μεταβλητή επηρεάζει τον κίνδυνο να εκδηλώσει το γεγονός του ενδιαφέροντος κάποιος εργαζόμενος. Το πρόσημο του συντελεστή δηλώνει αν η επίδραση της μεταβλητής αυξάνει τον κίνδυνο να αποχωρήσει με την θέληση του ο εργαζόμενος από την επιχείρηση (πρόσημο +) ή αντίθετα, αν μειώνει τον κίνδυνο ο

εργαζόμενος να εκδηλώσει το γεγονός (πρόσημο -). Δηλαδή, η μεταβλητή '*industryRealEstate*' φαίνεται πως επηρεάζει θετικά την επιβίωση του εργαζομένου (παραμονή του εργαζομένου στην εταιρία). Αντίθετα η μεταβλητή '*professionmanage*' διακρίνουμε πως επηρεάζει αρνητικά την επιβίωση του εργαζομένου (παραμονή του) στην εταιρία.

Όσον αφορά τις ψευδομεταβλητές, που προήλθαν από τον διαχωρισμό κλάσεων των κατηγορικών μεταβλητών, ο συντελεστής '*exp (coef)*' εκφράζει την αναλογία κινδύνου (*hazard ratio*) των ψευδομεταβλητών σε σύγκριση με την εξαγόμενη ψευδομεταβλητή, οι οποίες όμως

θα πρέπει αποτελούν μέρος της ίδιας κατηγορικής μεταβλητής. Παραδείγματος χάριν από τον πίνακα (6), παρατηρούμε ότι ο κίνδυνος ενός εργαζομένου να αποχωρήσει με την θέληση του από την εταιρία ο οποίος ασκεί το επάγγελμα του μηχανολόγου (*profession_Engineer*) είναι 2.7524 φορές παραπάνω από τον κίνδυνο εκείνου που ασκεί το επάγγελμα του λογιστή (*profession_Accounting*) για την χρονική περίοδο που διαρκεί η μελέτη. Επίσης, ο κίνδυνος για έναν εργαζόμενο να εκδηλώσει το γεγονός ασκώντας το επάγγελμα του διευθυντή (*profession_manage*) είναι 3.5988 φορές παραπάνω από τον κίνδυνο εκείνου που ασκεί το επάγγελμα του λογιστή (*profession_Accounting*) στην εταιρία.

Παρόμοιο συμπέρασμα για την στατιστική σημαντικότητα μπορεί να εξαχθεί και παρατηρώντας το διάστημα εμπιστοσύνης σχετικά με την αναλογία κινδύνου (*hazard ratio*) για την κάθε μία επεξηγηματική μεταβλητή. Διαστήματα εμπιστοσύνης όπου εμπεριέχεται η μη αποδεκτή του 1 ($e^{\beta_i} = 1\,, \beta_i = 0$) σημαίνει πως δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση, οπότε η συγκεκριμένη επεξηγηματική μεταβλητή δεν είναι στατιστικά σημαντική. Επίσης από την στιγμή που η αναλογία κινδύνου κυμαίνεται σε ένα διάστημα όπου το αριστερό άκρο είναι μία τιμή μικρότερη του 1 και το δεξί άκρο τιμή μεγαλύτερη του 1, σημαίνει πως δεν μπορούμε να είμαστε σίγουροι πως ο κίνδυνος να εκδηλώσει το γεγονός ο εργαζόμενος είναι μεγαλύτερος ή μικρότερος για την μία κατηγορία σε σχέση με την άλλη κατηγορία όπου ανήκει κάποιος άλλος εργαζόμενος.

	coef	*exp(coef)*	*Lower .95*	*Upper .95*	*Pr (>|z|)*
age	0,022557	1,022813	1,0090	1,0368	0,001114 **
extraversion	0,016142	1,016273	0,9486	1,0888	0,646137
independ	-0,013928	0,986169	0,9201	1,0570	0,693943
selfcontrol	-0,049708	0,951507	0,8873	1,0203	0,162915
anxiety	-0,054368	0,947084	0,8857	1,0127	0,111928
novator	0,006707	1,006730	0,9490	1,0680	0,823936
gender_m	-0,110635	0,895265	0,6984	1,1476	0,382574
industry_Agriculture	0,656145	1,927348	0,6617	5,6136	0,228994
industry_Banks	0,409351	1,505840	0,6395	3,5458	0,348838
industry_Building	0,426712	1,532211	0,6200	3,7867	0,355298
industry_Consult	0,276259	1,318189	0,5438	3,1951	0,540824
industry_etc	0,090306	1,094509	0,4589	2,6103	0,838638
industry_IT	-0,529569	0,588859	0,2416	1,4355	0,244091
industry_manufacture	-0,13719	0,871805	0,3709	2,0494	0,753077
industry_Mining	0,066419	1,068674	0,3882	2,9416	0,8977
industry_Pharma	-0,229537	0,794902	0,2799	2,2578	0,666497
industry_PowerGeneration	-0,257546	0,772946	0,2908	2,0547	0,605646
industry_RealEstate	-1,056903	0,347530	0,1008	1,1986	0,094292 .
industry_Retail	-0,332284	0,717284	0,3084	1,1986	0,440281
industry_State	0,007784	1,007814	0,3949	2,5721	0,987009
industry_Telecom	-0,522676	0,592932	0,2216	1,5865	0,297935
industry_transport	-0,18503	0,831079	0,3174	2,1764	0,706389
profession_BusinessDevelopment	0,5989	1,820203	0,6732	4,9217	0,237933
profession_Commercial	0,960549	2,613131	0,9674	7,0586	0,058147 .
profession_Consult	0,583026	1,791451	0,6471	4,9594	0,261768
profession_Engineer	1,012465	2,752377	0,9591	7,8990	0,059804 .
profession_etc	0,484767	1,623797	0,6267	4,2070	0,318257
profession_Finance	0,081414	1,084820	0,3878	3,0350	0,876741
profession_HR	0,222904	1,249701	0,5396	2,8944	0,602944
profession_IT	0,095309	1,099999	0,4207	2,8765	0,845916
profession_Law	0,414142	1,513072	0,4254	5,3818	0,522368
profession_manage	1,280598	3,598791	1,3519	9,5799	0,010360 *
profession_Marketing	0,733846	2,083077	0,8101	5,3566	0,127795
profession_PR	0,868633	2,383650	0,6809	8,3443	0,174212
profession_Sales	0,526763	1,693442	0,6796	4,2196	0,258126
profession_Teaching	0,630757	1,879032	0,6166	5,7261	0,267224
traffic_empjs	0,902592	2,465987	1,3326	4,5632	0,004047 **
traffic_friends	0,131263	1,140268	0,5855	2,2207	0,699521
traffic_KA	0,15856	1,171822	0,5897	2,3288	0,650902
traffic_rabrecNErab	0,526459	1,692927	0,9238	3,1024	0,088473 .
traffic_recNErab	-0,041064	0,959768	0,4584	2,0096	0,913272
traffic_referal	0,345049	1,412059	0,7471	2,6688	0,288077
traffic_youjs	0,645959	1,907816	1,0415	3,4948	0,036474 *
coach_no	0,067457	1,069784	0,8615	1,3285	0,541572
coach_yes	0,226838	1,254627	0,9338	1,6857	0,132229
head_gender_m	0,066905	1,069194	0,8762	1,3048	0,510156
greywage_white	-0,488929	0,613283	0,4718	0,7972	0,000258 ***
way_car	-0,175241	0,839255	0,6868	1,0255	0,086651 .
way_foot	-0,38103	0,683157	0,4877	0,9569	0,026676 *
***Concordance* = 0.659** (se = 0.012)					
Likelihood ratio test = 170.1 on 49 df, **p=3e-15**					
Wald test = 172.4 on 49 df, **p=1e-15**					
Score (logrank) test = 178.8 on 49 df, **p=<2e-16**					

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

Πίνακας 6 - Αποτελέσματα ανάλυσης του αναπτυγμένου 'Cox PH' μοντέλου, ύστερα από την εισαγωγή των ψευδομεταβλητών

Επιπλέον, παρατηρούμε στον πίνακα (6) πως ο συντελεστής ερμηνευτικής ικανότητας (*concordance*) του μοντέλου που δημιουργήθηκε είναι της τάξεως του 65.9% . Δηλαδή το '*Cox*' μοντέλο καταφέρνει να προβλέψει την σωστή σειρά, με βάση την χρονική στιγμή εκδήλωσης του

ενδιαφέροντος, από το 65.9% των εργαζομένων του δείγματος. Σύμφωνα με το '*Likelihood ratio test*', προκείμενου να συγκρίνουμε το αναπτυγμένο '*Cox*' μοντέλο (*saturated model*) σε σχέση με το απλό μοντέλο (*reduced model*) στο οποίο δεν συμπεριλαμβάνεται κάποια επεξηγηματική μεταβλητή(μοντέλο που αποκτήθηκε με την μέθοδο '*Kaplan Meier*'), διακρίνουμε πως το αναπτυγμένο μοντέλο με την εισαγωγή των μεταβλητών είναι στατιστικά σημαντικότερο για την ερμηνεία της εξαρτημένης μεταβλητής. Η τιμή για το παρατηρηθέν επίπεδο σημαντικότητας του '*Cox*' μοντέλοy είναι μικρότερη από επίπεδο σημαντικότητας της τάξεως του 0.1%.

Προκειμένου να μειωθεί ένας μεγάλος όγκος πληροφορίας ο οποίος μπορεί να οδηγήσει σε υψηλή διακύμανση στο δείγμα που χρησιμοποιήθηκε για την εκπαίδευση του 'Cox' μοντέλου (*overfitting*), αποτυγχάνοντας να προβεί σε ακριβείς προβλέψεις στην συνέχεια, θα πρέπει να απομακρυνθούν οι μη στατιστικά σημαντικές μεταβλητές από το 'Cox' μοντέλο. Αυτό θα συμβεί χρησιμοποιώντας τρεις διαφορετικές τεχνικές '*stepwise*' και δημιουργώντας τρία νέα μοντέλα.

Εκτελώντας τον αλγόριθμο της προς τα πίσω απαλοιφής (*backward stepwise*), παρατηρούμε ,από τα αποτελέσματα του πίνακα (7), ότι το αρχικό κριτήριο πληροφορίας *AIC* για το μοντέλο στο οποίο εμπεριέχονται όλες οι μεταβλητές είναι 6868.546. Ύστερα από την αφαίρεση συνολικά 26 επεξηγηματικών μεταβλητών, ξεκινώντας από την μεταβλητή που επιφέρει την μεγαλύτερη πτώση στο *AIC* και επαναλαμβάνοντας την ίδια διαδικασία κάθε φορά, καταλήγουμε σε ένα μοντέλο όπου το κριτήριο σύγκρισης *AIC* έχει μειωθεί στο 6832.695.

	Step	***DF***	***AIC***
1			6.868.546
2	- industry_State	1	6.866.546
3	- traffic_recNErab	1	6.864.558
4	- profession_Finanρe	1	6.862.583
5	- profession_IT	1	6.860.598
6	- industry_Mining	1	6.858.628
7	- novator	1	6.856.680
8	- industry_etc	1	6.854.761
9	- independ	1	6.852.916
10	- traffic_friends	1	6.851.260
11	- traffic_KA	1	6.849.480
12	- profession_Law	1	6.847.903
13	- profession_HR	1	6.846.276
14	- head_gender_m	1	6.844.673
15	- extraversion	1	6.843.284
16	- coach_no	1	6.842.005
17	- industry_Pharma	1	6.840.717
18	- industry_transport	1	6.839.285
19	- industry_manufacture	1	6.838.087
20	- industry_PowerGeneration	1	6.836.722
21	- gender_m	1	6.835.740
22	- profession_Consult	1	6.834.851
23	- profession_etc	1	6.833.840
24	- profession_Teaching	1	6.833.518
25	- profession_BusinessDevelopment	1	6.833.194
26	- profession_PR	1	6.832.879
27	- coach_yes	1	**6.832.695**

Πίνακας 7 - Αποτελέσματα ανάλυσης της 'backward stepwise' και εξαγωγή των λιγότερο στατιστικά σημαντικά μεταβλητών

Το νέο μοντέλο που αποκτήσαμε με την τεχνική της '*backward stepwise*' παρατηρούμε πως αποτελείται συνολικά από 23 επεξηγηματικές μεταβλητές. Σύμφωνα με το πίνακα (8) βλέπουμε ότι πολλές μεταβλητές οι οποίες δεν ήταν σημαντικές ή λιγότερο σημαντικές στο προηγούμενο μοντέλο μας, τώρα απέκτησαν στατιστική σημαντικότητα για την ερμηνεία της εξαρτημένης μεταβλητής. Φαίνεται πως η αφαίρεση των μη σημαντικών μεταβλητών από το μοντέλο, επηρέασε θετικά ορισμένες μεταβλητές με την απόκτηση ισχυρής στατιστικής σημαντικότητας. Παραδείγματος χάριν, η μεταβλητή '*selfcontrol*' η οποία δεν ήταν σημαντική για προηγούμενο '*Cox*' μοντέλο, τώρα είναι στατιστικά σημαντική σε επίπεδο σημαντικότητας της τάξεως του 0.1%. Επίσης λιγότερο σημαντικές μεταβλητές στο προηγούμενο μοντέλο όπως '*age*', '*profession_manage*', '*traffic_empjs*', '*traffic_youjs*' φαίνεται πως μετά την τεχνική της προς τα πίσω απαλοιφής, είναι στατιστικά σημαντικές σε επίπεδο σημαντικότητας της τάξεως του 0.1%. Η '*profession_Sales*' φαίνεται είναι η μοναδική μεταβλητή στο μοντέλο μας που δεν είναι στατιστικά σημαντική για κάποια από τα τέσσερα επίπεδα σημαντικότητας. Παρ 'όλα αυτά ίσως έχει κάποια επίδραση στην εκτίμηση της εξαρτημένης μεταβλητής για αυτό και παρέμεινε στο

μοντέλο. Διακρίνουμε από τον πίνακα (8) ότι ο κίνδυνος να αποχωρήσει εκουσίως κάποιος εργαζόμενος όποιος ανήκει σε επιχείρηση που ασχολείται με τον κλάδο των ακινήτων (*industry_RealEstate*) είναι κατά έναν βαθμό 0.3719 ή 62.81% μικρότερος από τον κίνδυνο κάποιου εργαζομένου ο οποίος απασχολείται σε επιχείρηση σχετική με τον κλάδο των ξενοδοχείων/εστιατορίων/καφετεριών (*industry_HoReCa*). Επίσης κάποιος εργαζόμενος ο οποίος ασκεί το επάγγελμα του διευθυντή στην εταιρία (*profession_manage*) είναι κατά 2.8055 φορές μεγαλύτερος ο κίνδυνος να αποχωρήσει εκουσίως από την εταιρία, την χρονική περίοδο που διαρκεί η μελέτη, σε σχέση με κάποιον εργαζόμενο που ασκεί το επάγγελμα του λογιστή (*profession_Accounting*). Τέλος, παρατηρούμε πως η ερμηνευτική ικανότητα του μοντέλου (*concordance*) έχει μειωθεί στο 0.652 σε σχέση με το προηγούμενο μοντέλο. Κάτι το οποίο αναμέναμε, καθώς αφαιρέθηκαν μεταβλητές από το νέο μοντέλο οπότε έχουμε λιγότερη πληροφορία για την εκτίμηση της εξαρτημένης μεταβλητής στην διάθεση μας. Από το '*Likelihood Ratio Test*' βλέπουμε ότι το νέο μοντέλο που αναπτύχθηκε με την τεχνική της '*backward stepwise*' είναι στατιστικά σημαντικότερο για την ερμηνεία της εξαρτημένης μεταβλητής, από το απλό μοντέλο (*Kaplan-Meier model*).

	coef	*exp(coef)*	*lower .95*	*upper .95*	*Pr(>|z|)*
age	0,02330967	1,02358347	1,01137875	1,03593546	0,000140 ***
selfcontrol	-0,07299399	0,92960642	0,89058471	0,9703379	0,000849 ***
anxiety	-0,06011373	0,94165743	0,89674791	0,98881604	0,015906 *
industry_Agriculture	0,89366745	2,44407676	1,27824732	4,67320456	0,006887 **
industry_Banks	0,42966049	1,5367357	1,17358392	2,01226054	0,001786 **
industry_Building	0,51985623	1,68178584	1,14869117	2,46228376	0,007526 **
industry_Consult	0,35700952	1,42904948	1,02520108	1,9919823	0,035131 *
industry_IT	-0,51814956	0,59562169	0,41329499	0,85838252	0,005453 **
industry_RealEstate	-0,98904756	0,37193077	0,14875693	0,92992303	0,034400 *
industry_Retail	-0,27692407	0,75811206	0,60435472	0,95098768	0,016643 *
industry_Telecom	-0,50515133	0,60341425	0,3491099	1,04296315	0,070405 .
profession_Commercial	0,72190201	2,05834448	1,20210208	3,52447775	0,008519 **
profession_Engineer	0,65116354	1,91777093	1,03937141	3,53852847	0,037204 *
profession_manage	1,03159383	2,80553382	1,64936601	4,77214879	0,000141 ***
profession_Marketing	0,47724548	1,61162902	1,02786432	2,52693674	0,037550 *
profession_Sales	0,29934558	1,34897572	0,94351337	1,92868014	0,100760
traffic_empjs	0,76071028	2,13979553	1,63604387	2,79865657	2,79E-08 ***
traffic_rabrecNErab	0,39413981	1,48310788	1,14105647	1,92769512	0,003215 **
traffic_referal	0,30062638	1,35070459	0,97886386	1,86379636	0,067261 .
traffic_youjs	0,53211731	1,70253328	1,31004747	2,21260652	0,000068 ***
greywage_white	-0,48384553	0,61640841	0,47775485	0,79530188	0,000198 ***
way_car	-0,22575033	0,79791729	0,65813906	0,96738219	0,021593 *
way_foot	-0,32752086	0,72070825	0,51840548	1,00195774	0,051376 .
Concordance= 0.652 (se = 0.012)					
Likelihood ratio test= 153.9 on 23 df, **p=<2e-16**					
Wald test = 158.7 on 23 df, **p=<2e-16**					
Score (logrank) test = 163.3 on 23 df, **p=<2e-16**					

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

Πίνακας 8 - Αποτελέσματα ανάλυσης νέου 'Cox' μοντέλου έπειτα από την χρήση της 'backward stepwise'

Προκειμένου να πραγματοποιηθεί η τεχνική της προς τα μπρος επιλογής (*forward stepwise*) θα πρέπει να ξεκινήσουμε από το απλό μοντέλο (*Kaplan-Meier model*), το οποίο δεν περιλαμβάνει καμία επεξηγηματική μεταβλητή, και στην συνέχεια σταδιακά προσθέτουμε την μεταβλητή που επιφέρει την μεγαλύτερη μείωση στο κριτήριο σύγκρισης *AIC*. Από την εκτέλεση του αλγορίθμου, όπως παρουσιάζεται στον πίνακα (9), παρατηρούμε ότι το υπεραπλουστευμένο μοντέλο (*Kaplan-Meier model*) εμφανίζει κριτήριο σύγκρισης *AIC = 6940.602* υψηλότερο από εκείνο του κορεσμένου μοντέλου (*saturated model*) *AIC = 6868.546*. Το καταληκτικό μοντέλο που αποκτήσαμε την μέθοδο της προς τα εμπρός επιλογής αποτελείται από 23 μεταβλητές,

έχοντας *AIC = 6832.941*. Συγκρίνοντας το τωρινό, με το προηγούμενο μοντέλο που είχαμε αποκτήσει από την '*backward stepwise*', βλέπουμε ότι και τα δύο μοντέλα αποτελούνται από 23 συνολικά επεξηγηματικές μεταβλητές, ενώ το κριτήριο σύγκρισης στο τωρινό είναι ελάχιστα

υψηλότερο *AIC = 6832.941* από εκείνο που είχαμε αποκτήσει με την μέθοδο της προς τα πίσω απαλοιφής *AIC = 6832.645*.

	Step	***Df***	***AIC***
1			6940,60207
2	+ greywage_white	1	6923,78106
3	+ traffic_empjs	1	6914,00047
4	+ industry_Banks	1	6903,85667
5	+ age	1	6892,68477
6	+ industry_Consult	1	6886,31747
7	+ industry_Building	1	6880,0841
8	+ industry_Agriculture	1	6873,63272
9	+ industry_etc	1	6868,00541
10	+ profession_manage	1	6862,97654
11	+ way_foot	1	6858,72947
12	+ extraversion	1	6854,60391
13	+ industry_State	1	6851,09304
14	+ traffic_youjs	1	6847,00343
15	+ traffic_rabrecNErab	1	6843,81972
16	+ profession_Engineer	1	6840,93396
17	+ profession_Commercial	1	6839,31919
18	+ anxiety	1	6837,97327
19	+ selfcontrol	1	6835,82618
20	+ way_car	1	6835,1744
21	+ profession_Marketing	1	6834,54628
22	+ industry_IT	1	6834,21035
23	+ traffic_referal	1	6833,8441
24	+ industry_RealEstate	1	**6832,9411**

Πίνακας 9 - Αποτελέσματα ανάλυσης της 'forward stepwise' και εισαγωγή των πιο στατιστικά σημαντικών μεταβλητών

Επίσης, διακρίνουμε μεταξύ των δύο πινάκων (9) και (8), ότι τα δύο μοντέλα διαφέρουν στην επιλογή των επεξηγηματικών μεταβλητών που έγινε με την κάθε μία μέθοδο. Όσον αφορά το μοντέλο που αποκτήθηκε με την μέθοδο της 'backward stepwise' βλέπουμε πως εμπεριέχονται οι μεταβλητές '*industry_Retail*', '*industry_Telecom*' και '*profession_Sales*' τις οποίες δεν τις συναντάμε στο δεύτερο. Αντ' αυτού, στο δεύτερο συμπεριλαμβάνονται οι '*industry_etc*', '*extraversion*' και '*industry_State*' οι οποίες δεν φαίνεται να είναι στατιστικά σημαντικές για το πρώτο.

	coef	*exp(coef)*	*lower .95*	*upper .95*	*Pr(>|z|)*
greywage_white	-0,49863278	0,60736048	0,47117204	0,78291309	0,000119 ***
traffic_empjs	0,74124617	2,09854903	1,60589135	2,74234493	5,64E-08 ***
industry_Banks	0,63154724	1,88051794	1,44718752	2,44360022	2,29E-06 ***
age	0,02593939	1,02627875	1,01399606	1,03871021	0,000024***
industry_Consult	0,56773629	1,76426874	1,26882563	2,45316939	0,000737 ***
industry_Building	0,70836568	2,03066977	1,38930454	2,96811793	0,000254 ***
industry_Agriculture	1,09830023	2,99906396	1,57258432	5,71949274	0,000855 ***
industry_etc	0,38133922	1,46424422	1,082818	1,98002908	0,013259 *
profession_manage	1,00195108	2,7235906	1,60027036	4,63543281	0,000222 ***
way_foot	-0,35048064	0,70434947	0,50653928	0,97940713	0,0371901 *
extraversion	0,0285893	1,02900189	0,97184461	1,08952078	0,326842
industry_State	0,38174718	1,4648417	1,01786112	2,10810804	0,039852 *
traffic_youjs	0,52381505	1,68845692	1,29944289	2,19393003	0,000088 ***
traffic_rabrecNErab	0,41532203	1,5148585	1,16609145	1,9679385	0,001865 **
profession_Engineer	0,75366291	2,12476861	1,15546242	3,90721632	0,015313 *
profession_Commercial	0,66119765	1,93711093	1,13689807	3,30055865	0,015022 *
anxiety	-0,0510586	0,95022299	0,9037827	0,99904958	0,045808 *
selfcontrol	-0,05601995	0,94552027	0,89818969	0,99534495	0,032513 *
way_car	-0,18467544	0,83137407	0,68541327	1,00841766	0,060811 .
profession_Marketing	0,40113198	1,49351436	0,94745646	2,35428776	0,084075 .
industry_IT	-0,29611114	0,74370477	0,51859471	1,06652994	0,107442
traffic_referal	0,30141782	1,35177403	0,97938963	1,86574674	0,066758 .
industry_RealEstate	-0,71334158	0,49000407	0,19706807	1,21838096	0,124798
Concordance= 0.653 (se = 0.012)					
Likelihood ratio test= 153.7 on 23 df, **p=<2e-16**					
Wald test = 159 on 23 df, **p=<2e-16**					
Score (logrank) test = 163.8 on 23 df, **p=<2e-16**					

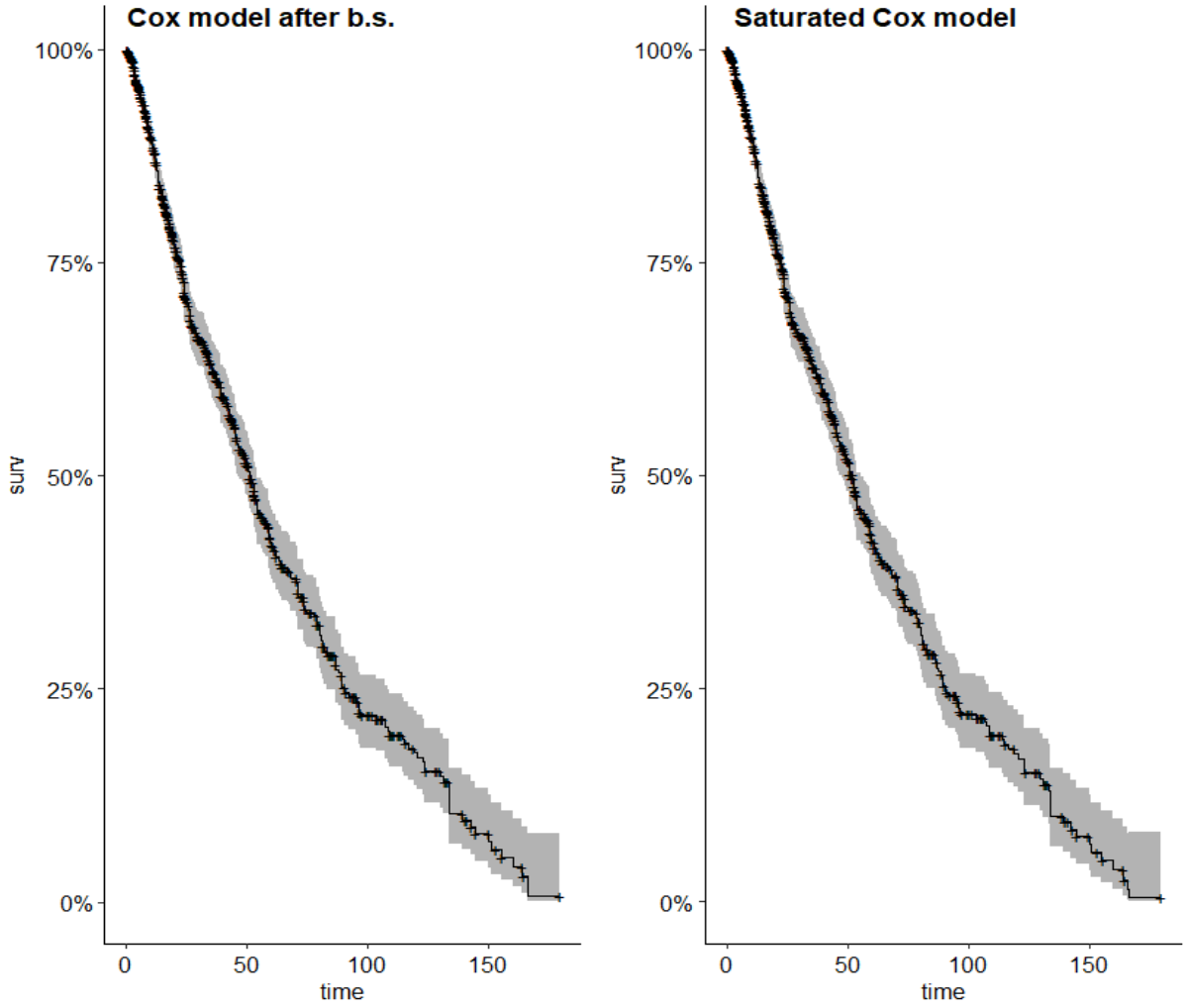
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Πίνακας 10 - Αποτελέσματα ανάλυσης 'Cox' μοντέλου έπειτα από την χρήση της 'forward stepwise'

Όπως φαίνεται στον παραπάνω πίνακα (10), οι μεταβλητές που καθίστανται σημαντικές σε επίπεδο σημαντικότητας της τάξεως του 0.1% είναι οι '*greywage_white*', '*traffic_empjs*', '*industry_Banks*', '*age*', '*industry_Consult'*, '*industry_Building*', '*industry_Agriculture*', '*profession_manage*' και '*traffic_youjs*'. Φαίνεται πως μερικές από τις παραπάνω μεταβλητές βελτίωσαν την στατιστική τους σημαντικότητα, ύστερα από την μέθοδο της προς τα εμπρός επιλογής. Ακόμα, παρατηρούμε πως στο μοντέλο που δημιουργήθηκε υπάρχουν τρείς μεταβλητές ('*industry_IT*', '*industry_ReaLEstate*' και '*extraversion*') που δεν παρουσιάζουν κάποια στατιστική σημαντικότητα για κανένα από τα τέσσερα επίπεδα σημαντικότητας. Τέλος βλέπουμε έναν ελαφρώς αυξημένο δείκτη '*concordance = 0.653*' σε σχέση με το προηγούμενο μοντέλο που είχε αποκτηθεί από την χρήση της προς τα πίσω απαλοιφής (*concordance = 0.652*).

Τέλος εκτελώντας την μέθοδο της κατά βήματα παλινδρόμησης '*both stepwise*' το μοντέλο που θα αποκτήσουμε είναι πανομοιότυπο με εκείνο που είχε αποκτηθεί με την μέθοδο της προς τα πίσω απαλοιφής (*backward stepwise*). Τα αποτελέσματα ανάλυσης για το μοντέλο με την χρήση της '*both stepwise*' παρουσιάζονται στον πίνακα (8).

Στην συνέχεια της ανάλυσης μας θα αναπαραστήσουμε γραφικά το κορεσμένο μοντέλο, στο οποίο εμπεριέχονται όλες οι επεξηγηματικές μεταβλητές που έχουμε δημιουργήσει, καθώς και το μοντέλο που αποκτήθηκε μετά την χρήση της προς τα πίσω απαλοιφής (*backward stepwise*).



Διάγραμμα 39 - Γραφικές αναπαραστάσεις των δύο 'Cox' μοντέλων

Όπως παρατηρούμε στο διάγραμμα (39), οι καμπύλες επιβίωσης που δημιουργούνται για τα δύο μοντέλα παραπάνω είναι σχεδόν πανομοιότυπες. Η μοντελοποίηση της διάρκειας του κύκλου ζωής των εργαζομένων είναι ίδια και για τα δύο μοντέλα που δημιουργήθηκαν. Θα μπορούσαμε να επιλέξουμε εκείνο που αποκτήσαμε μετά την μέθοδο της προς τα πίσω απαλοιφής, στο οποίο εμπεριέχονται λιγότερες επεξηγηματικές μεταβλητές, προκειμένου να προβούμε σε προβλέψεις. Η αφαίρεση των 26 μη σημαντικών επεξηγηματικών μεταβλητών δεν φαίνεται να επηρέασε σημαντικά την απόδοση αλλά και την ερμηνευτική ικανότητα του τελικού μοντέλου.

Από τον έλεγχο που πραγματοποιήθηκε μέσω της γλώσσας προγραμματισμού R, σχετικά με το αν ικανοποιείται το κριτήριο αναλογίας κινδύνου για τα δύο μοντέλα, προκειμένου να είναι εφικτή η χρησιμοποίηση της μεθόδου της '*Cox*' παλινδρόμησης, εξήχθησαν τα εξής αποτελέσματα:

- Σύμφωνα με τον πίνακα (11) παρατηρούμε, ότι για το συγκεκριμένο μοντέλο δεν ικανοποιείται η υπόθεση της σταθερής αναλογίας κινδύνων όσο διαρκεί η μελέτη. Καταλήγουμε στο παραπάνω συμπέρασμα, καθώς η τιμή για το παρατηρηθέν επίπεδο σημαντικότητας του μοντέλου είναι μικρότερη από το αποδεκτό επίπεδο σημαντικότητας της τάξεως του 5% (*p-value* < *0.05*), με αποτέλεσμα να παραβιάζεται σε μεγάλο βαθμό η θεώρηση της σταθερής αναλογίας κινδύνου (*proportional hazard assumption)* για τις επεξηγηματικές μεταβλητές του υποδείγματος. Οι μεταβλητές '*profession_PR*' και '*profession_Law*' είναι εκείνες που παραβιάζουν περισσότερο την θεώρηση της σταθερής αναλογίας κινδύνων σε σχέση με τις υπόλοιπες μεταβλητές του υποδείγματος, καθώς λαμβάνουν την χαμηλότερη τιμή για το παρατηρηθέν επίπεδο σημαντικότητας τους. Επομένως, η αναλογία κινδύνου (*hazard ratio*) των συγκεκριμένων μεταβλητών σε σχέση με την μεταβλητή '*profession_Accounting*' δεν φαίνεται πως παραμένει σταθερή κατά την διάρκεια της μελέτης. Τέλος οι δύο παραπάνω επεξηγηματικές μεταβλητές, είναι εκείνες που επηρεάζουν περισσότερο την θεώρηση περί της μη σταθερή αναλογίας κινδύνου του μοντέλου.

	chisq	*df*	*p*
age	1,25722916	1	0,26217622
extraversion	2,2400808	1	0,13447395
independ	0,78173046	1	0,37661242
selfcontrol	0,00890977	1	0,92479809
anxiety	0,92432833	1	0,33634105
novator	2,73579634	1	0,09812258
gender_m	0,50801748	1	0,47599834
industry_Agriculture	4,77220886	1	0,02892269
industry_Banks	0,00048704	1	0,98239292
industry_Building	0,27542066	1	0,59971851
industry_Consult	0,56265537	1	0,45319233
industry_etc	3,71295167	1	0,05399183
industry_IT	0,39858551	1	0,52782066
industry_manufacture	1,54721657	1	0,21354682
industry_Mining	1,13689006	1	0,28631043
industry_Pharma	1,4743144	1	0,22466621
industry_PowerGeneration	0,91949542	1	0,33760749
industry_RealEstate	0,78756195	1	0,37483836
industry_Retail	2,36035328	1	0,124454
industry_State	0,0858392	1	0,76953494
industry_Telecom	0,58135316	1	0,44578241
industry_transport	0,4604535	1	0,4974121
profession_BusinessDevelopment	1,46184921	1	0,22663612
profession_Commercial	0,48172248	1	0,48764314
profession_Consult	6,34592825	1	0,0117651
profession_Engineer	0,00099063	1	0,97489136
profession_etc	0,11049801	1	0,73957787
profession_Finanpe	0,25788645	1	0,61157593
profession_HR	0,43534292	1	0,50937884
profession_IT	0,74112274	1	0,38930146
profession_Law	9,02598802	1	0,00266168
profession_manage	4,46316569	1	0,03463326
profession_Marketing	0,96668387	1	0,32550861
profession_PR	7,45545735	1	0,00632444
profession_Sales	2,48034209	1	0,11527718
profession_Teaching	1,9512399	1	0,1624533
traffic_empjs	1,60317459	1	0,2054539
traffic_friends	4,41901995	1	0,03554044
traffic_KA	0,1986076	1	0,65584711
traffic_rabrecNErab	0,01190496	1	0,91311534
traffic_recNErab	1,88925213	1	0,1692863
traffic_referal	2,11302028	1	0,14605081
traffic_youjs	2,18136663	1	0,13969037
coach_no	2,02011491	1	0,15522739
coach_yes	0,00653921	1	0,93554899
head_gender_m	0,06332731	1	0,80131224
greywage_white	0,81923212	1	0,36540477
way_car	1,81255446	1	0,1782021
way_foot	0,33583889	1	0,562241
GLOBAL	**85,2760663**	**49**	**0,00101778**

Πίνακας 11 - Έλεγχος αναλογικότητας κινδύνου στο πρώτο μοντέλο

- Σύμφωνα με τον πίνακα (12), παρατηρούμε πως για το δεύτερο μοντέλο το κριτήριο περί της σταθερής αναλογίας κινδύνου κατά την διάρκεια της μελέτης δεν παραβιάζεται σε μεγάλο βαθμό. Η τιμή για το παρατηρηθέν επίπεδο σημαντικότητας είναι της τάξεως του

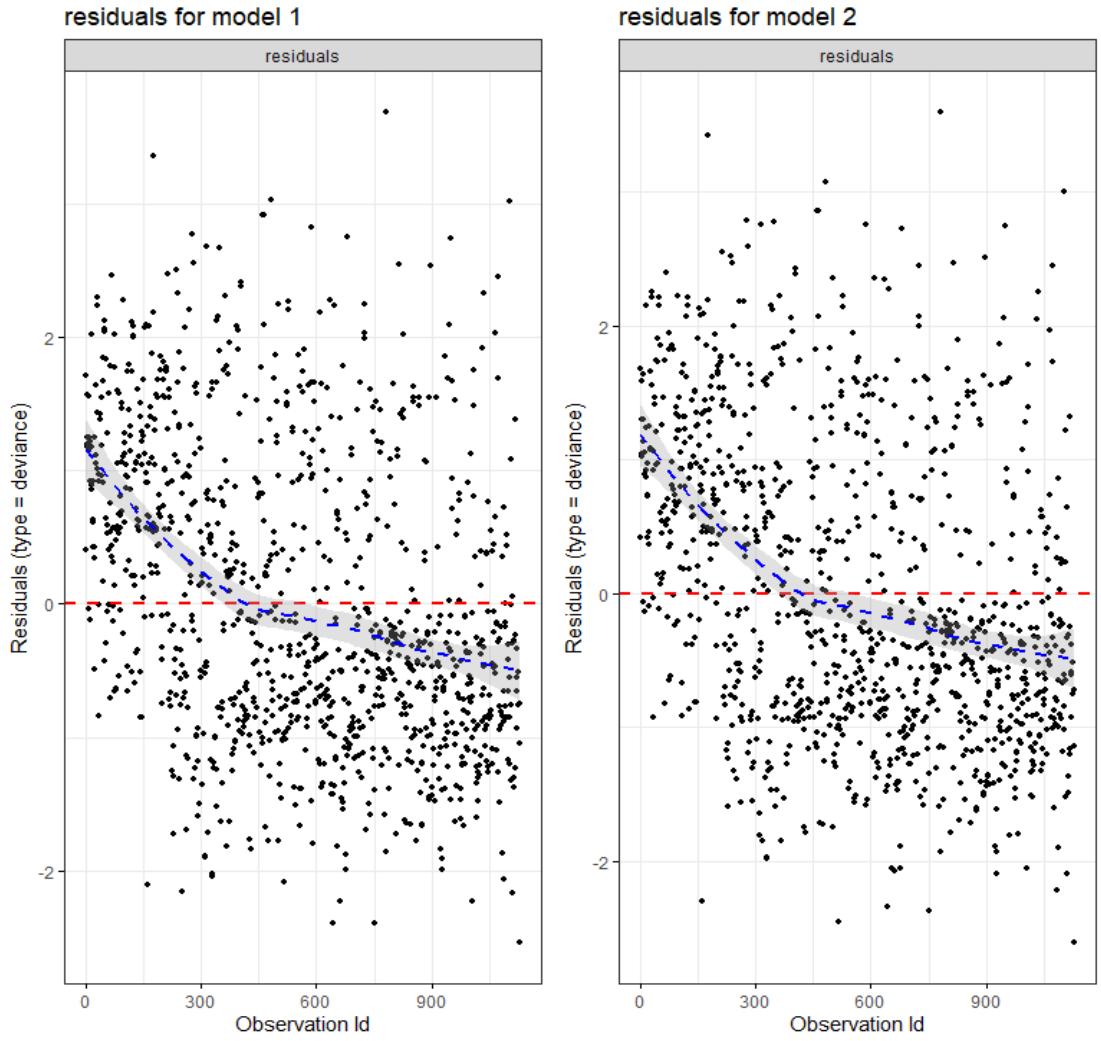
p-value = 9% και μπορούμε να ισχυριστούμε πως στην συνέχεια της ανάλυσης μας θα μπορούσε να χρησιμοποιηθεί η μέθοδος της '*Cox*' παλινδρόμησης. Οι μεταβλητές '*profession_PR*' και '*profession_Law*' δεν συμπεριλαμβάνονται πλέον στο νέο μοντέλο, οπότε ίσως και αυτό αποτελεί έναν λόγο για την αύξηση της τιμής 'p-value' του μοντέλου. Παρατηρούμε πως οι '*industry_Agriculture*' και '*profession_manage'* είναι οι μόνες μεταβλητές για τις οποίες η αναλογία κινδύνου δεν φαίνεται να παραμένει σταθερή κατά την διάρκεια της μελέτης σε σχέση με την μεταβλητή που γίνεται η σύγκριση για την καθεμιά.

	chisq	***df***	***p***
age	0,8119202	1	0,36755315
selfcontrol	0,00069385	1	0,9789854
anxiety	0,78908921	1	0,37437567
industry_Agriculture	5,07238879	1	0,02430986
industry_Banks	2,3552E-06	1	0,99877552
industry_Building	0,20864913	1	0,64782827
industry_Consult	0,50544766	1	0,4771162
industry_IT	0,40806832	1	0,52295158
industry_RealEstate	0,71426641	1	0,3980311
industry_Retail	2,01768551	1	0,15547597
industry_Telecom	0,60292211	1	0,43746527
profession_Commercial	0,2891132	1	0,59078937
profession_Engineer	0,08369879	1	0,77234613
profession_manage	4,03180155	1	0,04465024
profession_Marketing	0,78734419	1	0,3749044
profession_Sales	2,60488925	1	0,1065346
traffic_empjs	2,05926798	1	0,15128283
traffic_rabrecNErab	0,11850768	1	0,7306587
traffic_referal	1,46244876	1	0,2265409
traffic_youjs	1,74617919	1	0,18635778
greywage_white	0,55772322	1	0,45517905
way_car	2,20536606	1	0,13753124
way_foot	0,19105114	1	0,6620432
GL			

Πίνακας 12 - Έλεγχος αναλογικότητας κινδύνου στο δεύτερο μοντέλο

Παρακάτω ,στο διάγραμμα (4.4), αναπαρίστανται γραφικά τα κατάλοιπα για τα δύο μοντέλα που δημιουργήθηκαν. Αφού προήλθε η κανονικοποίηση των καταλοίπων, έπειτα θα πρέπει τα κατάλοιπα να είναι διασκορπισμένα συμμετρικά γύρω από την οριζόντια κόκκινη γραμμή με τυπική απόκλιση ($\sigma = 1$). Θετικές τιμές καταλοίπων υποδηλώνουν πως ο εργαζόμενος αποχώρησε πολύ νωρίς από την εταιρία σε σχέση βέβαια με τον προβλεπόμενο χρόνο επιβίωσης του (παραμονή στην εταιρία) από το μοντέλο. Αρνητικές τιμές υποδηλώνουν ότι ο εργαζόμενος

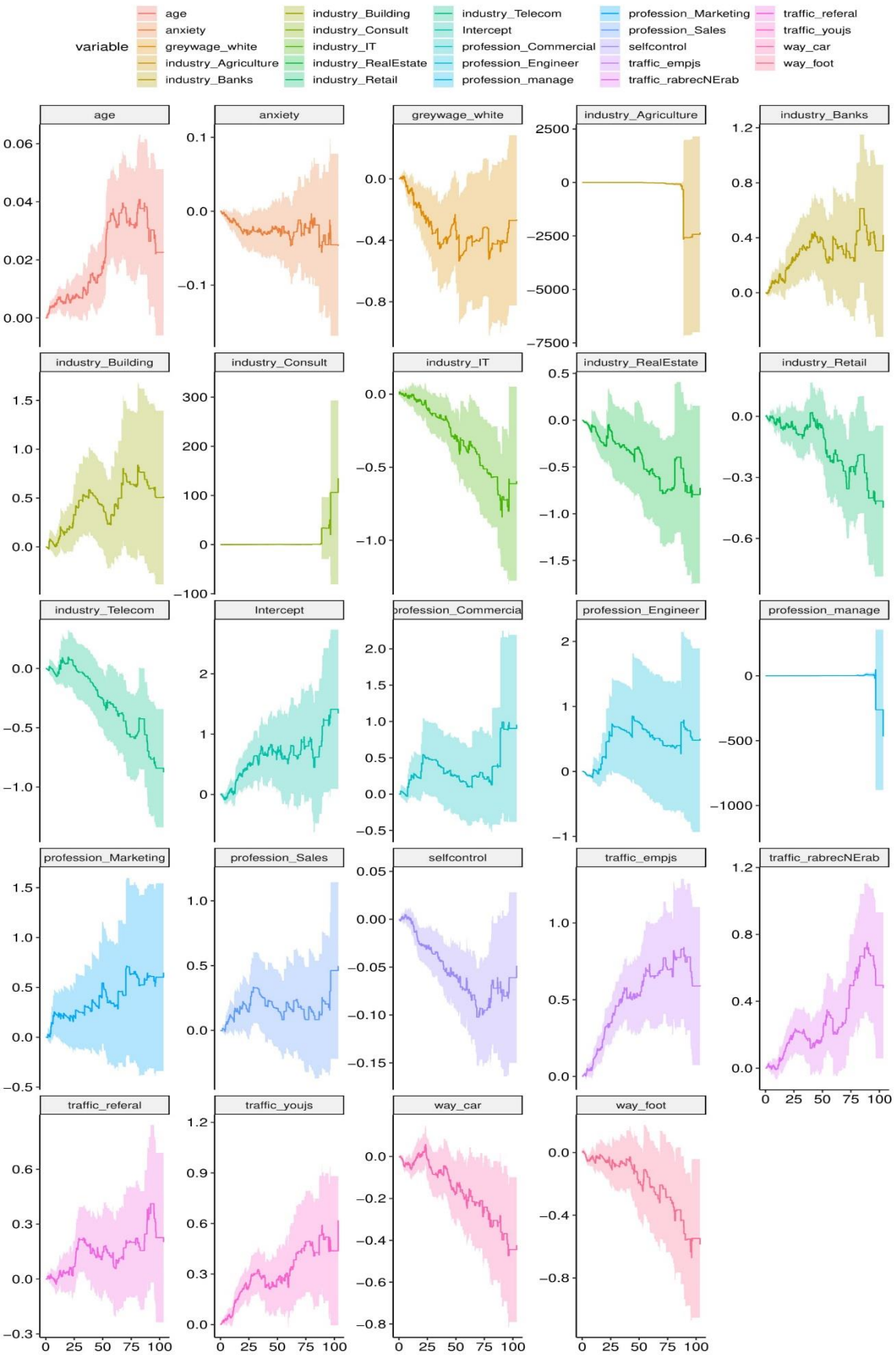
παρέμεινε παραπάνω μήνες στην εταιρία σε σχέση με την πρόβλεψη του μοντέλου. Ενώ πολύ μεγάλες ή πολύ μικρές τιμές καταλοίπων, υποδηλώνουν την αδυναμία του μοντέλου να κάνει προβλέψεις για τους συγκεκριμένους εργαζομένους. Όπως φαίνεται διάγραμμα (40), διακρίνουμε πως υπάρχει κάποια συμμετρία των καταλοίπων γύρω από το 0.



Διάγραμμα 40 - αποκλίσεις καταλοίπων για τα δύο μοντέλα

Η επιπρόσθετη παλινδρόμηση του Aalen (*aalen's additive regression*) κάνει την θεώρηση πως η επίδραση όλων των επεξηγηματικών μεταβλητών στην εξαρτημένη μεταβλητή είναι εξαρτώμενη από τον χρόνο (*time – dependent*). Μελετώντας το δεύτερο μοντέλο και εκτελώντας την '*aalens additive regression*' συμπεριλαμβάνοντας τις επεξηγηματικές μεταβλητές που καταρτίζονται από το δεύτερο μοντέλο, καταλήγουμε στα εξής συμπεράσματα. Από το νέο μοντέλο που δημιουργήθηκε με την χρήση της '*aalens additive regression'*, μελετώντας τον πίνακα (13), βλέπουμε πως για τις μεταβλητές '*age*', '*selfcontrol*', '*traffic_empjs*' και '*traffic_youjs*' η επίδραση στον κίνδυνο κάποιος εργαζόμενος να εκδηλώσει το γεγονός, είναι στατιστικά σημαντική για επίπεδο σημαντικότητας της τάξεως του 0.1%. Όπως φαίνεται στο διάγραμμα (41), για την μεταβλητή 'age' βλέπουμε πως υπάρχει αυξανόμενος κίνδυνος να αποχωρήσει

κάποιος εργαζόμενος όσο αυξάνεται η ηλικία του. Μάλιστα, διακρίνουμε μία απότομη κλίση στην καμπύλη πριν τους 50 μήνες επιβίωσης, συμπεραίνοντας πως η ηλικία του εργαζομένου έχει μεγαλύτερη επίδραση στον κίνδυνο αποχώρησης του από την εταιρία σε εκείνη την χρονική περίοδο. Για μεγαλύτερες τιμές του δείκτη που ορίζει τον βαθμό αυτοελέγχου (*selfcontrol*) βλέπουμε πως μειώνεται ο κίνδυνος να αποχωρήσει κάποιος εργαζόμενος από την εταιρία. Επίσης, παρατηρούμε μία απότομη πτώση στην καμπύλη μέχρι τους 70 μήνες 'επιβίωσης' του εργαζομένου, συμπεραίνοντας πως η επίδραση της μεταβλητής είναι μεγαλύτερη για εκείνη την χρονική περίοδο. Παρατηρούμε μία σταθερή αύξησή (σχεδόν γραμμική) καθ' όλη την διάρκεια της μελέτης, στον κίνδυνο να αποχωρήσει κάποιος εργαζόμενος που η πρόσληψη του έγινε μέσω αναζήτησης του βιογραφικού του από τον εργοδότη (*traffic_empjs*) σε σχέση με κάποιον εργαζόμενο ο οποίος εκδήλωσε το ενδιαφέρον του για την θέση λόγω κάποιας διαφήμισης (*traffic_advert*). Επίσης, η προσέλκυση κάποιου εργαζομένου από την εταιρία, έπειτα από δημοσίευση της θέσεως σε αγγελία εργασίας (*traffic_youjs*) φαίνεται να επιφέρει σταθερή αύξηση στον κίνδυνο να αποχωρήσει σε σχέση με κάποιον που εκδήλωσε το ενδιαφέρον του για την θέση λόγω κάποιας διαφήμισης (*traffic_advert*). Τέλος, από τον πίνακα (13), βλέπουμε πως η μεταβλητή '*traffic_referral*' είναι η λιγότερο στατιστικά σημαντική όσον αφορά την επίδραση της στον κίνδυνο να αποχωρήσει κάποιος εργαζόμενος σε σχέση με την '*traffic_advert*'.



Διάγραμμα 41 - Αποτελέσματα ανάλυσης από την χρήση της επιπρόσθετης παλινδρόμησης του Aalen

	test-statistic	***p-value***
Intercept	8,119467506	0,007400744
age	582,7675083	0,000223886
selfcontrol	-141,4307789	0,001027988
anxiety	-68,98986507	0,070794556
industry_Agriculture	6,040891384	0,042081768
industry_Banks	18,57564926	0,007789262
industry_Building	11,02052148	0,02721594
industry_Consult	10,66779625	0,070456832
industry_IT	-15,08950014	0,004302891
industry_RealEstate	-3,788200629	0,088774921
industry_Retail	-16,81703376	0,046874969
industry_Telecom	-7,81646084	0,019815764
profession_Commercial	6,302713321	0,073479216
profession_Engineer	5,307181258	0,093475309
profession_manage	9,470951118	0,009488573
profession_Marketing	8,116664297	0,05693249
profession_Sales	9,441237407	0,071312201
traffic_empjs	39,83342445	4,39111E-08
traffic_rabrecNErab	21,88567605	0,002105894
traffic_referal	8,482339391	0,126579492
traffic_youjs	29,53103791	3,05341E-05
greywage_white	-20,30307395	0,006720576
way_car	-21,2652736	0,030506502
way_foot	-11,25608952	0,058164666

Πίνακας 13 - Στατιστική σημαντικότητα επεξηγηματικών μεταβλητών μετά της χρήσης της 'Aalen's additive regression'

4.3. Παραμετρικά μοντέλα

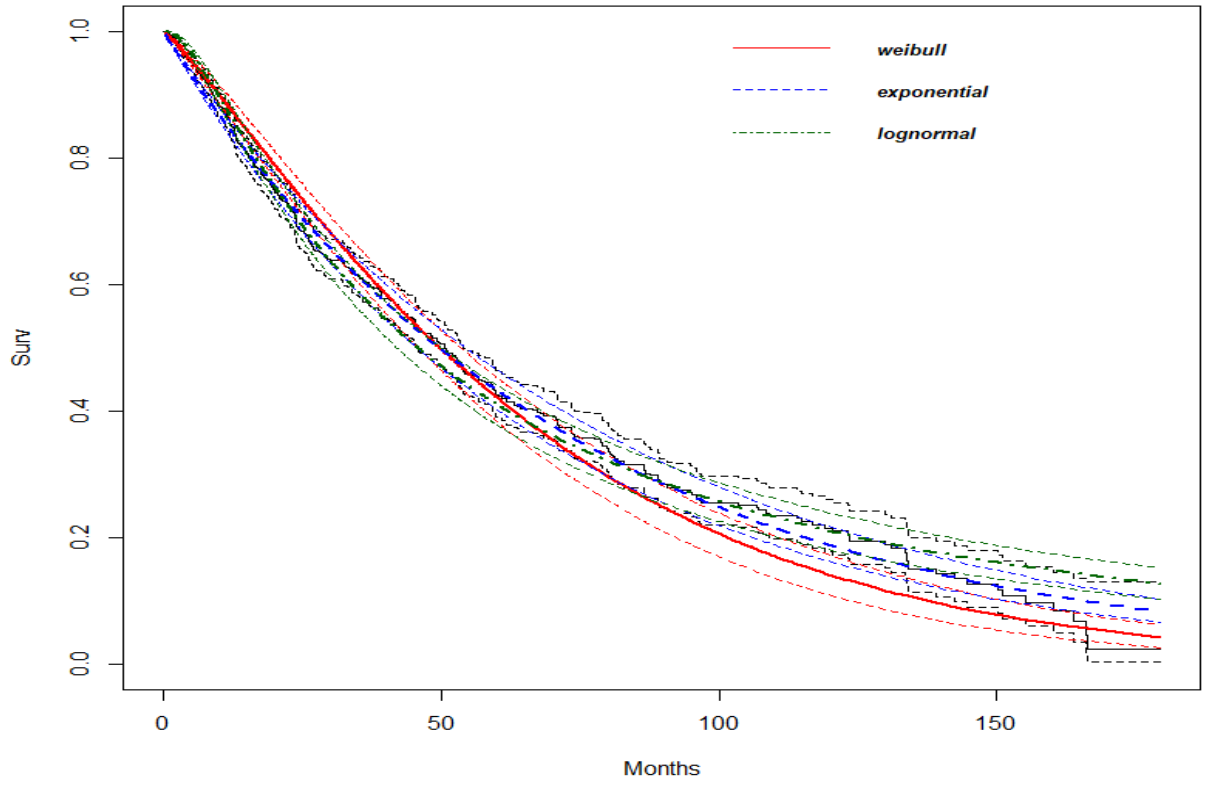
4.3.1. Περιγραφή ενδεχόμενων ειδικών βημάτων ανάλυσης

Σε αυτό το κεφάλαιο θα γίνει χρήση τριών διαφορετικών παραμετρικών μοντέλων προκειμένου να ελέγξουμε κάθε ένα από τα τρία που οι παρατηρήσεις στο σύνολο δεδομένων μας κατανέμονται καλύτερα. Το κάθε ένα από τα παραμετρικά μοντέλα ακολουθεί την '*Weibull*', εκθετική και την λογαριθμική κανονική κατανομή. Σκοπός της περαιτέρω ανάλυσης είναι να εντοπίσουμε αν τα δεδομένα από το δείγμα που διαθέτουμε ακολουθούν ή προσεγγίζουν κάποια από τις προαναφερθείσες κατανομές. Στην κατασκευή και των τριών μοντέλων θα συμπεριληφθούν οι περισσότερο στατιστικά σημαντικές μεταβλητές στην επίδραση της αποχώρησης του εργαζομένου από την εταιρία. Δηλαδή οι επεξηγηματικές μεταβλητές που

συμπεριλαμβάνονται στο μοντέλο έπειτα από την μέθοδο της προς τα πίσω απαλοιφής. Στην συνέχεια θα αναπαραστήσουμε γραφικά τις καμπύλες που αποκτήσαμε και από τα τρία μοντέλα, εντοπίζοντας την κατανομή που εκτιμά καλύτερα την καμπύλη επιβίωσης των εργαζομένων. Επίσης για την σύγκριση των μοντέλων θα χρησιμοποιηθεί το στατιστικό τεστ '*maximum log-likelihood*' $(-2\ln L_R - (2\ln L_F))$, αφαιρώντας το ανεπτυγμένο παραμετρικό μοντέλο που έχει δημιουργηθεί από το μοντέλο που δεν εμπεριέχει καμία επεξηγηματική μεταβλητή (*Kaplan-Meier model*). Επιπλέον, για την σύγκριση των μοντέλων θα γίνει χρήση του κριτηρίου πληροφορίας (*AIC*).

4.3.2. Παρουσίαση και σχολιασμός αποτελεσμάτων

Για την εκτίμηση της πιθανότητας να παραμείνει κάποιος εργαζόμενος στην εταιρία (επιβίωση) έπειτα από κάποια συγκεκριμένη χρονική περίοδο, θα δημιουργήσουμε αρχικά το μη παραμετρικό μοντέλο που ακολουθεί την '*Weibull*' κατανομή. Έπειτα θα αποκτήσουμε το μοντέλο που ακολουθεί την εκθετική κατανομή (το οποίο αποτελεί παραλλαγή του '*Weibull*' μοντέλου) και στην συνέχεια θα κατασκευαστεί το μοντέλο επιβίωσης της λογαριθμικής κατανομής. Σύμφωνα με το διάγραμμα (42), παρατηρούμε τα τρία μοντέλα που δημιουργούνται, '*Weibull*' (κόκκινη συμπαγής γραμμή) , '*exponential*' (μπλε διακεκομμένη γραμμή) και '*lognormal*' (πράσινη διακεκομμένη γραμμή), σε σύγκριση με την καμπύλη επιβίωσης '*Kaplan-Meier*' στην οποία δεν περιλαμβάνεται καμία επεξηγηματική μεταβλητή. Βλέπουμε πως οι καμπύλες των παραμετρικών μοντέλων είναι αρκετά κοντά η μία με την άλλη. Διακρίνουμε από το διάγραμμα πως η καμπύλη '*Weibull*' κατανέμει καλύτερα τα δεδομένα και προσεγγίζει περισσότερο την '*Kaplan-Meier*' καμπύλη των εργαζομένων, καθώς η επιβίωση των εργαζομένων (παραμονή τους στην εταιρία) προς το τέλος της μελέτης μειώνεται αγγίζοντας το 0 όπως υποδηλώνεται και από την καμπύλη της κατανομής 'Weibull'. Παρόλο που η επιβίωση μειώνεται απότομα και στην αρχή της μελέτης, προσεγγίζοντας την λογαριθμική κατανομή, αργότερα όταν πλησιάζουμε στους τελευταίους μήνες παρατηρούμε πως η καμπύλη επιβίωσης απομακρύνεται αρκετά από την λογαριθμική κατανομή. Από τους 135 μήνες και έπειτα η διαφορά της καμπύλης επιβίωσης αρχίζει να μεγαλώνει σε σχέση με την λογαριθμική καμπύλη, όπου φαίνεται πως ορισμένοι εργαζόμενοι εκδήλωσαν το γεγονός εκείνη την περίοδο. Η εκθετική κατανομή η οποία υποδηλώνει πως ο κίνδυνος κάποιος εργαζόμενος να αποχωρήσει από την εταιρία είναι σταθερός κατά την πάροδο του χρόνου, φαίνεται και αυτή να μην προσεγγίζει αρκετά την καμπύλη επιβίωσης.



Διάγραμμα 42 - Παραμετρικά μοντέλα της διάρκειας του κύκλου ζωής των εργαζομένων

Σύμφωνα με τον πίνακα (14), παρατηρούμε πως επιβεβαιώνονται οι υποθέσεις μας για την υπεροχή του '*Weibull*' μοντέλου σε σχέση με τα υπόλοιπα παραμετρικά μοντέλα. Η τιμή του '*Log-likelihood*' για παραμετρικό '*Weibull*' μοντέλο είναι μεγαλύτερη από την τιμή των υπολοίπων παραμετρικών μοντέλων. Με το '*Weibull*' μοντέλο να έχει την μεγαλύτερη διαφορά από το υπεραπλουστευμένο '*Kaplan-Meier*' μοντέλο. Επίσης η τιμή για το παρατηρηθέν επίπεδο σημαντικότητας βλέπουμε πως είναι πάρα πολύ χαμηλή και για τα τρία μοντέλα, υποδηλώνοντας την σημαντικότητα και των τριών μοντέλων σε σχέση με το υπεραπλουστευμένο '*Kaplan-Meier*' μοντέλο. Επίσης, ένας άλλος λόγος που αποδεικνύει ότι υπερτερεί το '*Weibull*', είναι η χαμηλότερη τιμή που λαμβάνει για το κριτήριο πληροφορίας *AIC = 5918.272* σε σχέση με τα υπόλοιπα μοντέλα. Φαίνεται πως υπάρχει μεγαλύτερη ευαισθησία στην επιλογή παραμετρικού μοντέλου, αν κάποιος εργαζόμενος αποχωρήσει προς το τέλος της μελέτης. Τέλος, σύμφωνα με το παραμετρικό μοντέλο, οι εργαζόμενοι του δείγματος που διαθέτουμε τείνουν να εκδηλώνουν το γεγονός (αποχωρούν από την εταιρία) προς το τέλος της μελέτης παρά στην αρχή. Γεγονός το οποίο δεν μπορούμε να ισχυριστούμε πως είναι απόλυτα σωστό, καθώς υπάρχει μία σημαντική πτώση στην καμπύλη επιβίωσης και στους αρχικούς μήνες.

Distirbutions	*Log-likelihood*	*p-value*	*AIC*
Weibull	-2934,136	2,00E-16	5918,272
Exponential	-2946,06	2,00E-16	5940,121
Log-normal	-2948,188	2,00E-16	5946,375

Πίνακας 14 - Στατιστικοί έλεγχοι για την σύγκριση των παραμετρικών μοντέλων

4.4. Δέντρα Επιβίωσης

4.4.1. Περιγραφή ενδεχόμενων ειδικών βημάτων ανάλυσης

Λόγω της μεγάλης επιρροής του '*Machine Learning*' στην ανάλυση επιβίωσης, όπως και σε πολλούς άλλους κλάδους της επιστήμης, η τεχνική που θα εφαρμόσουμε στο παρών κεφάλαιο προκειμένου να μοντελοποιήσουμε την διάρκεια του κύκλου ζωής των εργαζομένων και κάνοντας προβλέψεις στην συνέχεια σχετικά με τον χρόνο παραμονής τους (επιβίωσης) βασίζεται στα δέντρα αποφάσεων. Η τεχνική των δέντρων επιβίωσης αποτελεί μία μη-παραμετρική μέθοδος στην ανάλυση επιβίωσης, έχοντας μεγάλη αναγνωσιμότητα τα τελευταία χρόνια. Στην αρχή της ανάλυσης θα κατασκευάσουμε μόνο ένα δέντρο για το δείγμα εργαζομένων που διαθέτουμε. Για την δημιουργία του δέντρου θα χρησιμοποιηθούν και οι 14 επεξηγηματικές μεταβλητές του δείγματος. Επίσης, στο δέντρο που θα κατασκευαστεί, κάθε φορά ο διαχωρισμός στον κόμβο θα γίνει ανάλογα με την σημαντικότητα της μεταβλητής, ξεκινώντας με την περισσότερο στατιστικά σημαντική μεταβλητή του μοντέλου. Σε κάθε καταληκτικό κόμβο εντάσσεται ορισμένος αριθμός εργαζομένων, ο οποίος βασίζεται στα χαρακτηριστικά από τον διαχωρισμό της κάθε επεξηγηματικής μεταβλητής που συνέβη στους προηγούμενους κόμβους. Επίσης, αξίζει να σημειωθεί πως στο εσωτερικό κάθε

καταληκτικού κόμβου παρουσιάζεται η καμπύλη επιβίωσης των εργαζομένων που εντάσσονται στον συγκεκριμένο κόμβο. Ο διαχωρισμός της κάθε μεταβλητής στον κάθε κόμβο θα γίνει στο σημείο εκείνο όπου η διαφορά μεταξύ των δύο καμπύλών επιβίωσης, δηλαδή των εργαζομένων που ανήκουν στον κάθε μεταγενέστερο κόμβο, που θα δημιουργηθούν να είναι βέλτιστη. Ό έλεγχος μεταξύ των δύο καμπύλων που θα δημιουργηθούν θα γίνει με την χρήση του του στατιστικού '*logrank test*'.Αν υπολογίσουμε την μέση καμπύλη, από τις συνολικές καμπύλες επιβίωσης που έχουν δημιουργηθεί στον κάθε καταληκτικό κόμβο, μπορούμε να αποκτήσουμε

την συνολική καμπύλη επιβίωσης όλων των εργαζομένων. Μπορούμε να αποκτήσουμε και την στατιστική σημαντικότητα της κάθε μεταβλητής, προκειμένου να δούμε τον λόγο που έγινε ο κάθε διαχωρισμός στον κόμβο για την κάθε μία μεταβλητή. Όσο πιο νωρίς επιλέγεται για να γίνει ο διαχωρισμός στον κόμβο, τόσο μεγαλύτερη είναι και η στατιστική σημαντικότητα της στο μοντέλο. Στο τέλος, θα πρέπει να 'κουρέψουμε' (*pruning*) το δέντρο που έχει δημιουργηθεί έτσι ώστε να αποκτήσουμε ένα μικρότερο σε μέγεθος αλλά και πιο αποδοτικό αν αυτό μπορεί να συμβεί σε σχέση με το δέντρο που έχουμε κατασκευάσει. Ένας τρόπος που αυτό μπορεί να γίνει είναι, επιλέγοντας την κατάλληλη τιμή για την πολυπλοκότητα της παραμέτρου '*complexity parameter*', η οποία χρησιμοποιείται με σκοπό να ελεγχθεί το μέγεθος του δέντρου. Παρατηρώντας στα στατιστικά αποτελέσματα του δέντρου που δημιουργήσαμε, εμφανίζεται η τιμή για την πολυπλοκότητα της παραμέτρου (*complexity parameter*) η οποία προσφέρει το σφάλμα εγκυρότητας (*cross-validation error*) για κάθε ένα διαχωρισμό στον κόμβο του δέντρου. Η βέλτιστη τιμή πολυπλοκότητας της παραμέτρου (*cp*) είναι εκείνη που προσφέρει το μικρότερο δυνατό σφάλμα εγκυρότητας (*cross-validated error*). Σε περίπτωση που το κόστος προσθήκης μίας επιπλέον μεταβλητής προκειμένου να γίνει κάποιος περαιτέρω διαχωρισμός, είναι υψηλότερο από την τιμή της '*complexity parameter*' που έχουμε θέσει, τότε το δέντρο σταματάει να αναπτύσσεται.

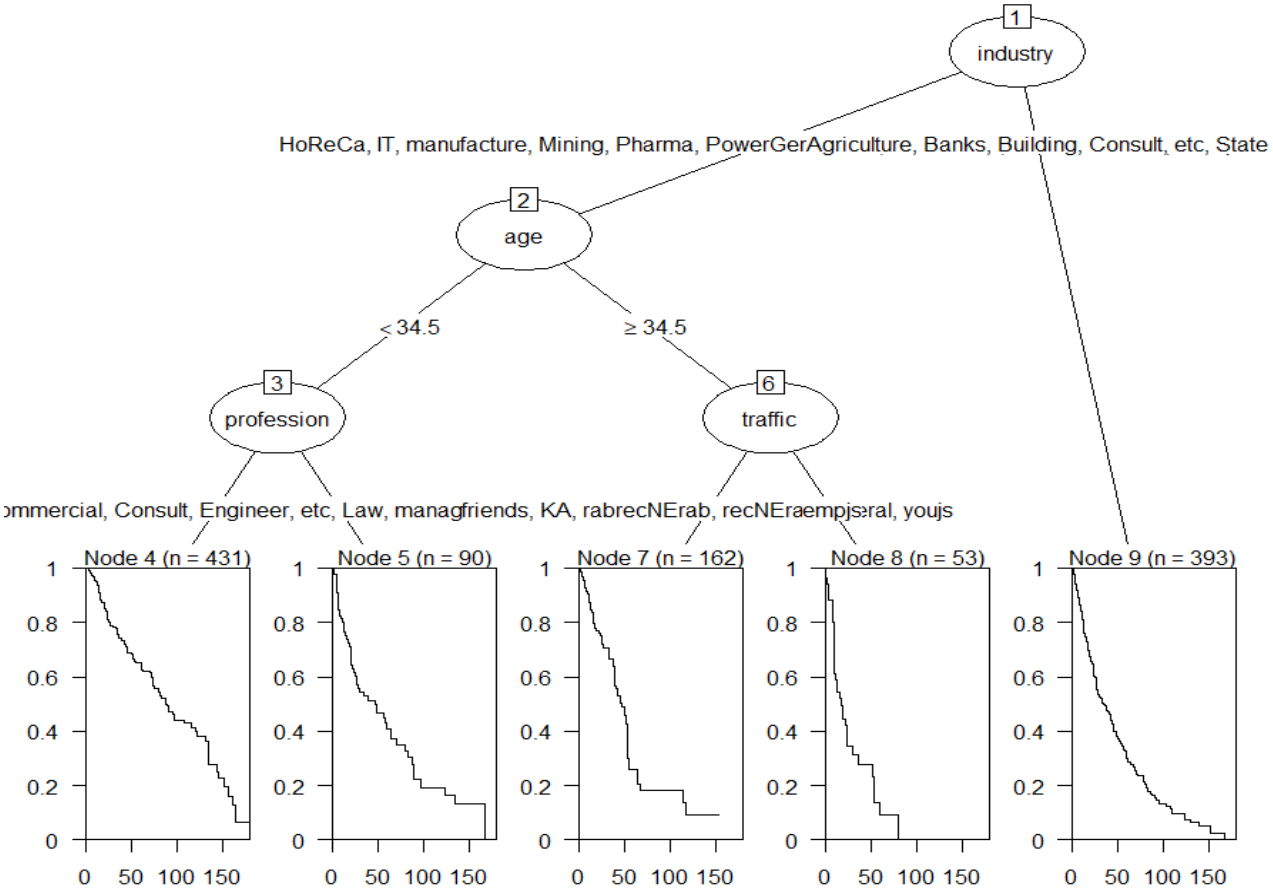
Στην συνέχεια θα χρησιμοποιηθεί η τεχνική των '*Random Survival Forests*', εξίσου επηρεασμένη από την μέθοδο των '*Random Forests*' από τον κλάδο του '*Machine Learning*'. Η συγκεκριμένη τεχνική θα χρησιμοποιηθεί προκειμένου να αποκτήσουμε ένα πιο ισχυρό μοντέλο σε προβλεπτική και ερμηνευτική ικανότητα. Καθώς αντί να δημιουργήσουμε μόνο ένα δέντρο για να εκτιμήσουμε την καμπύλη επιβίωσης των εργαζομένων ή τον κίνδυνο κάποιος εργαζόμενος να αποχωρήσει από την εταιρία, έχουμε την δυνατότητα να δημιουργήσουμε πολλά ξεχωριστά δέντρα μειώνοντας με αυτό τον τρόπο ακόμα περισσότερο το σφάλμα πρόβλεψης του μοντέλου και κάνοντας πιο ακριβείς προβλέψεις. Στο νέο μοντέλο που θα δημιουργηθεί με βάση την τεχνική '*Random Survival Forests*' θα συμπεριλαμβάνεται ο ίδιος αριθμός μεταβλητών με το μοντέλο που είχε κατασκευαστεί προηγουμένως. Με την συγκεκριμένη τεχνική θα κατασκευάσουμε έναν συγκεκριμένο αριθμό δέντρων επιβίωσης για το κάθε '*bootstrapped*' δείγμα που έχει δημιουργηθεί από το αρχικό σύνολο δεδομένων. Βέβαια, για το κάθε δέντρο που αναπτύσσεται, θα επιλέγεται ένας συγκεκριμένος αριθμός μεταβλητών κάθε φορά προκειμένου να γίνει με τυχαίο τρόπο ο διαχωρισμός στον κάθε κόμβο και όχι με βάση την στατιστική σημαντικότητα. Ο αριθμός των υποψήφιων μεταβλητών που θα επιλέγεται θα είναι (*mtry = 4*) από τις συνολικές επεξηγηματικές μεταβλητές του υποδείγματος. Με την τυχαία επιλογή των μεταβλητών κατά τις οποίες θα γίνεται ο διαχωρισμός στον κόμβο κάθε φορά, μπορούμε να εξαλείψουμε οποιαδήποτε συσχέτιση θα μπορούσε να δημιουργηθεί μεταξύ των δημιουργημένων δέντρων. Έπειτα με τις παρατηρήσεις που έχουν μείνει εκτός των τυχαίων (*bootstrapped*)

δειγμάτων που έχουν δημιουργηθεί (*out-of sample*), θα χρησιμοποιηθούν για να γίνουν οι προβλέψεις του μοντέλου και να υπολογιστεί το συνολικό σφάλμα πρόβλεψης (*out-of bag error*). Στην συνέχεια θα αναπαρασταθεί γραφικά, η μέση καμπύλη επιβίωσης για όλων των εργαζομένων του δείγματος, η οποία έχει προκύψει υπολογίζοντας την μέση καμπύλη από τον συνολικό αριθμό των καμπυλών επιβίωσης οι οποίες έχουν αναπαραχθεί από το μοντέλο των '*Random Survival Forests'*. Στο τέλος θα συγκριθούν οι καμπύλες που έχουμε αποκτήσει με την χρήση των τριών μεθόδων ('*Kaplan-Meier*', '*Cox*', '*Random Survival Forests*') και θα προσπαθήσουμε να εντοπίσουμε τυχόν ανομοιότητες και διαφορές μεταξύ τους.

Σε μεταγενέστερο στάδιο θα χωρίσουμε το σύνολο δεδομένων μεταξύ ενός υποσυνόλου (*train*) προκειμένου να εκπαιδεύσουμε τον αλγόριθμο '*Random Survival Forests*' και οι παρατηρήσεις που ανήκουν στο άλλο υποσύνολο θα χρησιμοποιηθούν προκειμένου να κάνουμε προβλέψεις (*test*). Η τεχνική αυτή θα χρησιμοποιηθεί προκειμένου να ελέγξουμε την προβλεπτική ικανότητα του μοντέλου μας. Δηλαδή κατά πόσο η πρόβλεψη για τον χρόνο επιβίωσης (παραμονή στην εταιρία) κάποιου τυχαίου εργαζομένου ανταποκρίνεται στην πραγματικό χρόνο επιβίωσης, υπολογίζοντας το σφάλμα για την πρόβλεψη μας.

4.4.2. Παρουσίαση και σχολιασμός αποτελεσμάτων

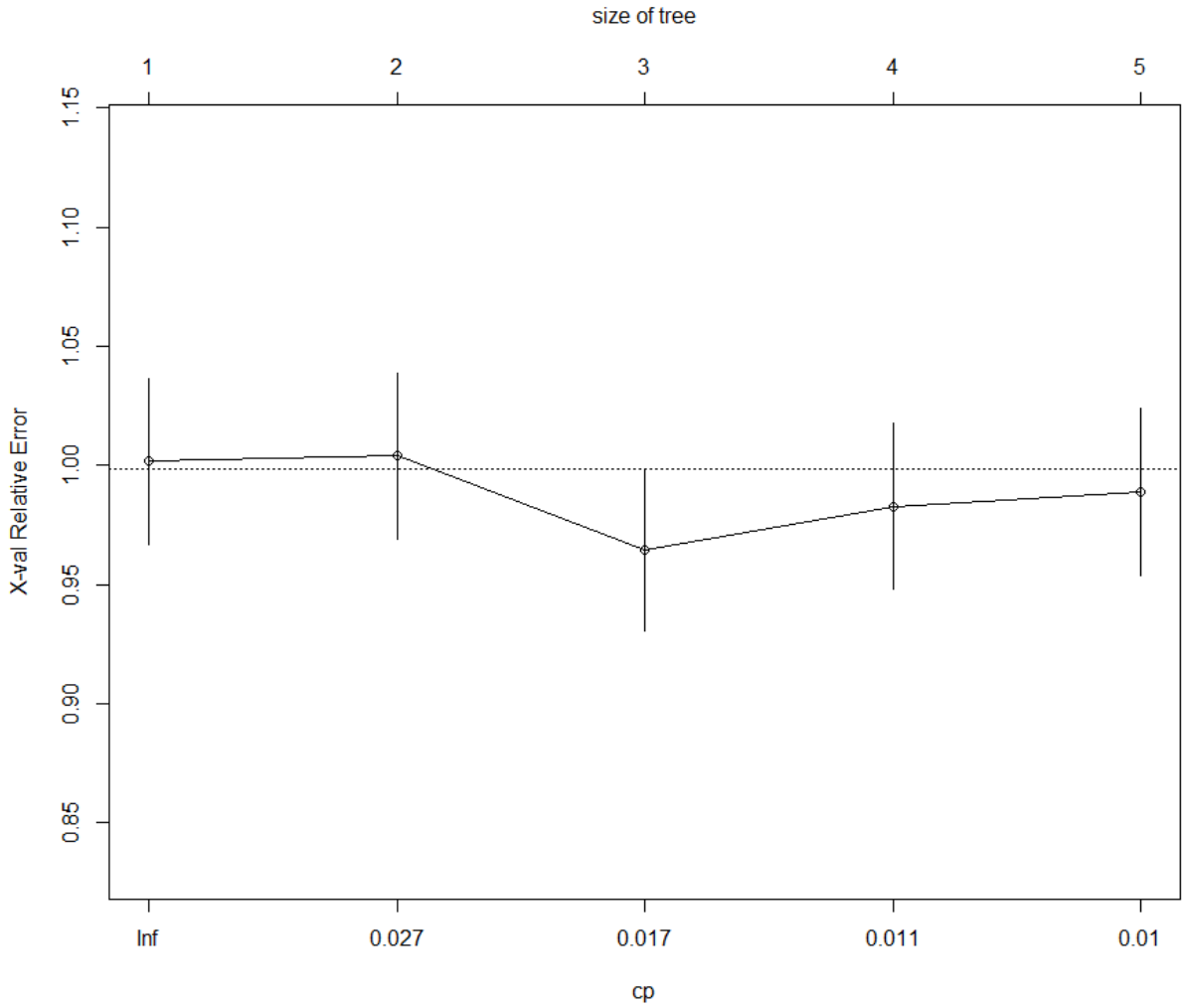
Μία διαφορετική μέθοδος που θα χρησιμοποιήσουμε προκειμένου να μοντελοποιήσουμε την διάρκεια του κύκλου ζωής των εργαζομένων, είναι εκείνη των των δέντρων επιβίωσης (*Survival Trees*). Στο διάγραμμα (43), αναπαρίσταται γραφικά το δέντρο επιβίωσης όλων των εργαζομένων, εισάγοντας στο μοντέλο μας τις 14 μεταβλητές του συνόλου δεδομένων που διαθέτουμε. Η πρώτη μεταβλητή που χρησιμοποιείται από το δέντρο για να γίνει ο διαχωρισμός στον αρχικό κόμβο είναι η '*industry*'. Έπειτα η δεύτερη μεταβλητή που χρησιμοποιείται για να γίνει ο διαχωρισμός στον δεύτερο κόμβο είναι η '*age*' και στην συνέχεια οι '*profession*' και '*traffic*' αποτελούν τις δύο επόμενες μεταβλητές για να γίνουν διαχωρισμοί. Για την κατασκευή του δέντρου χρησιμοποιούνται μόλις 4 από τις 14 συνολικά επεξηγηματικές μεταβλητές του συνόλου δεδομένων. Αναλόγως με τον διαχωρισμό στον κόμβο που γίνεται για την καθεμιά μεταβλητή, οι εργαζόμενοι χωρίζονται σε διαφορετικές υποομάδες. Στους 5 καταληκτικούς κόμβους του δέντρου, παρουσιάζεται η καμπύλη επιβίωσης (παραμονής του στην εταιρία) για τον κάθε ένα εργαζόμενο ο οποίος ανήκει σε κάποια από τις 5 υποομάδες.



Διάγραμμα 43 - Αρχικό δέντρο επιβίωσης για το σύνολο των εργαζομένων

Στο διάγραμμα (44), αντικατοπτρίζεται το σφάλμα εγκυρότητας για κάθε διαφορετική τιμή πολυπλοκότητας της παραμέτρου (*cost parameter*) όπου μπορούμε να θέσουμε στο μοντέλο μας, αποκτώντας διαφορετικού μεγέθους δέντρο (σύνολο καταληκτικών κόμβων) από το αρχικό. Παρατηρούμε ότι το σφάλμα εγκυρότητας ελαχιστοποιείται θέτοντας την βέλτιστη τιμή για την πολυπλοκότητα της παραμέτρου *'cost parameter'*

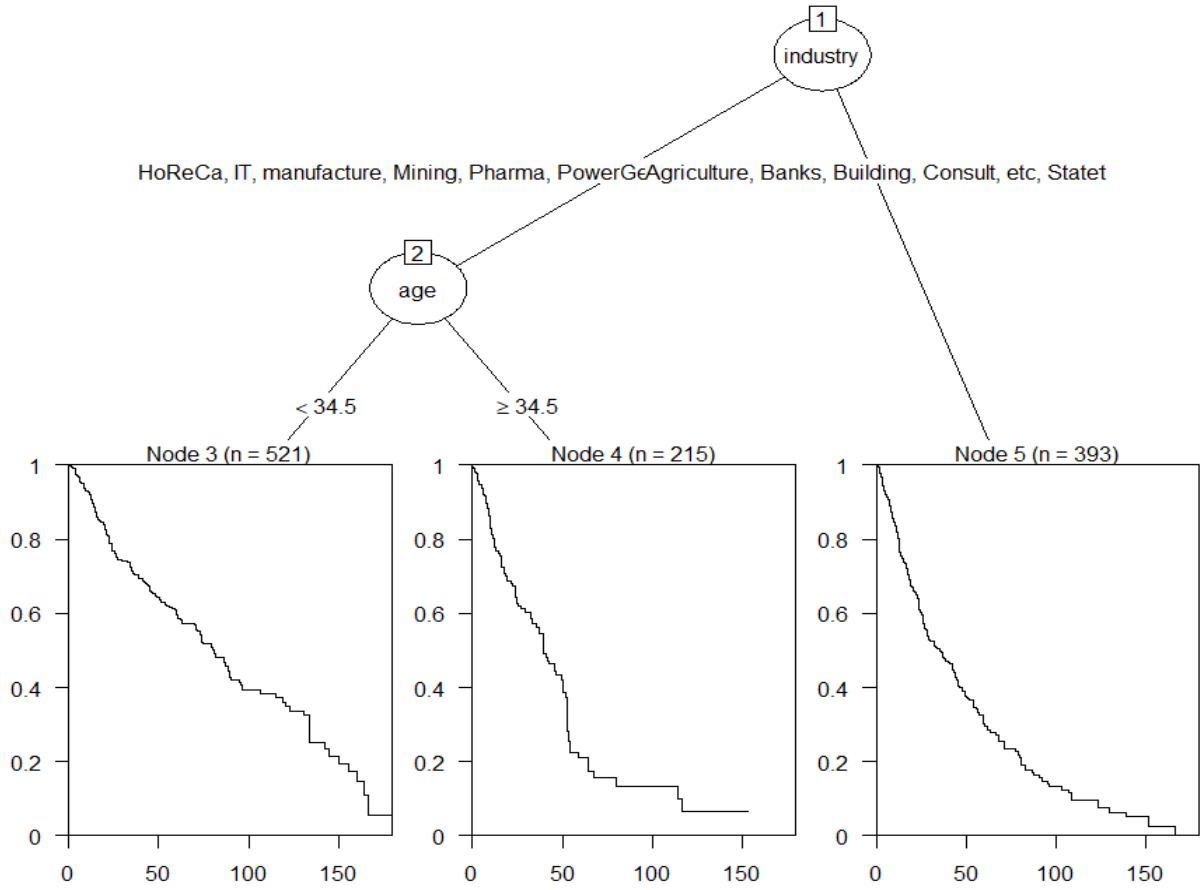
= 0.017. Όταν το κόστος προσθήκης κάθε μίας επιπλέον μεταβλητής στο μοντέλο ξεπερνάει '*cost parameter*' που εχουμε θέσει τότε ο αλγόριθμος θα σταματήσει, αποκτώντας ένα δέντρο που συνολικά θα αποτελείται από τρεις καταληκτικούς κόμβους.



Διάγραμμα 44 - Σφάλμα εγκυρότητας (cross-validation error) για την κάθε μία τιμή πολυπλοκότητας της παραμέτρου (cost parameter)

Το κούρεμα (*pruning*) του δέντρου επιβίωσης αποτελεί μία τεχνική προκειμένου να μειωθεί το '*overfitting*' στο σύνολο δεδομένων που χρησιμοποιείται για την εκπαίδευση του αλγορίθμου, αλλά και γενικότερα η πολυπλοκότητα του δέντρου. Στο διάγραμμα (45), παρουσιάζεται το δέντρο που αποκτήσαμε, θέτοντας τιμή για το κόστος πολυπλοκότητας της παραμέτρου *'cost parameter' = 0.017* και 'κουρεύοντας' το αρχικό. Βλέπουμε πως μόνο δύο (*age, industry*) είναι εκείνες οι μεταβλητές που χρησιμοποιούνται στην κατασκευή του δέντρου, θεωρώντας πως είναι οι μόνες σημαντικές από το σύνολο των 14 μεταβλητών. Διακρίνουμε ότι, σύμφωνα με την εκτίμηση του μοντέλου, 521 είναι οι εργαζόμενοι ηλικίας *'age' < 34.5* και που η απασχόληση τους γίνεται από κάποια εταιρία (*industry*) η οποία δραστηριοποιείται με τους κλάδους (*HoReCa, IT, manufacture, Mining, Pharma, PowerGeneration, RealEstate, Retail, Telecom, transport*). Η καμπύλη επιβίωσης για τους 521 εργαζομένους είναι πάνω από τις υπόλοιπες δύο καμπύλες, με τον κίνδυνο να αποχωρήσει από την εταιρία κάποιος από τους εργαζομένους που ανήκει στην

συγκεκριμένη κατηγορία να είναι μικρότερος σε σχέση με τους υπόλοιπους εργαζομένους. Επίσης, όπως φαίνεται και στον πίνακα (4.12), από τους 521 εργαζομένους του κόμβου, οι 211 αποχώρησαν από την εταιρία. Μετά την χρονική περίοδο των 81 μηνών, η πιθανότητα παραμονής για κάποιον εργαζόμενο που ανήκει σε αυτή την κατηγορία είναι '*SurvProb' = 50%.* Για τους 215 εργαζομένους ηλικίας *'age'* ≥ *34.5* και που απασχολούνται σε εταιρία η οποία δραστηριοποιείται στους κλάδους (*HoReCa, IT, manufacture, Mining, Pharma, PowerGeneration, RealEstate, Retail, Telecom, transport*), παρατηρούμε μία απότομη πτώση στην καμπύλη επιβίωσης τους την χρονική περίοδο των 50 μηνών. Επομένως, η πιθανότητα κάποιου εργαζομένου ,ο οποίος ανήκει στην συγκεκριμένη κατηγορία, να παραμείνει στην εταιρία μετά τους 50 μήνες ελαττώνεται σημαντικά. Από τους 215 εργαζομένους φαίνεται πως 110 είναι εκείνοι που εκδήλωσαν το γεγονός. Μετά την χρονική περίοδο των 39 μηνών, φαίνεται πως η πιθανότητα κάποιος εργαζόμενος να παραμείνει στην εταιρία είναι *'SurvProd' = 50%.* Τέλος, οι 393 εργαζόμενοι, οι οποίοι απασχολούνται σε κάποια εταιρία που δραστηριοποιείται με τους κλάδους (*Agriculture, Banks, Building, Consult, State*) ή και κάποιον άλλον κλάδο (*etc*) πέρα από αυτούς που αναφέρονται, θα λέγαμε πως η καμπύλη επιβίωσης εμφανίζει κάποια σταθερότητα. Δηλαδή ο κίνδυνος κάποιος εργαζόμενος να αποχωρήσει ο οποίος ανήκει στη συγκεκριμένη κατηγορία παραμένει σταθερός κατά την διάρκεια της μελέτης. Σύμφωνα με τον πίνακα (15), βλέπουμε πως από τους 393 εργαζομένους που ανήκουν στον 5^{ο} καταληκτικό κόμβο, 250 είναι εκείνοι που αποχώρησαν από την εταιρία.



Διάγραμμα 45 - 'Κούρεμα' αρχικού δέντρου για το σύνολο των εργαζομένων

nodes	***n***	***events***	***median***	***lower .95%***	***upper .95%***
3	521	211	81	71,3	89,4
4	215	110	39,3	33,4	50,7
5	393	250	35,3	28,2	42,9

Πίνακας 15 - Ανάλυση καταληκτικών κόμβων του βελτιστοποιημένου δέντρου

Στην συνέχεια της ανάλυσης μας, θα χρησιμοποιήσουμε την τεχνική των 'Random Survival Forests'. Θα κατασκευάσουμε πολλά διαφορετικά διαδοχικά δέντρα επιβίωσης σε 'bootstrapped' δείγματα προκείμενου να εκπαιδεύσουμε τον μπορέσουμε να μοντελοποιήσουμε την διάρκεια του κύκλου ζωής των εργαζομένων και να βελτιώσουμε τις προβλέψεις μας σχετικά με την αποχώρηση του εργαζομένου από την εταιρία. Σύμφωνα με τον πίνακα (16), για την κατασκευή του μοντέλου αναπτύχθηκαν 500 δέντρα για το κάθε '*bootstrapped*' δείγμα που δημιουργήθηκε από το αρχικό σύνολο δεδομένων. Επίσης βλέπουμε πως το κάθε δέντρο αποτελείται από τρεις καταληκτικούς κόμβους. Το σφάλμα πρόβλεψης του μοντέλου που δημιουργήσαμε είναι της τάξεως του 32,6 %. Δηλαδή αποτυγχάνει να προβλέψει σωστά τον κίνδυνο να αποχωρήσει, περίπου για έναν από τους τρεις εργαζομένους που απασχολούνται στην εταιρία.

Type	***Survival***
Number of trees	500
Sample size	1129
Number of independent variables	14
Mtry:	4
Target node size	3
Variable importance mode	permutation
Splitrule	extratreees
Number of unique death times	773
Number of random splits	1
OOB prediction error (1-C)	**0,3261442**

Πίνακας 16 - Αποτελέσματα ανάλυσης του μοντέλου 'Random Survival Forests'

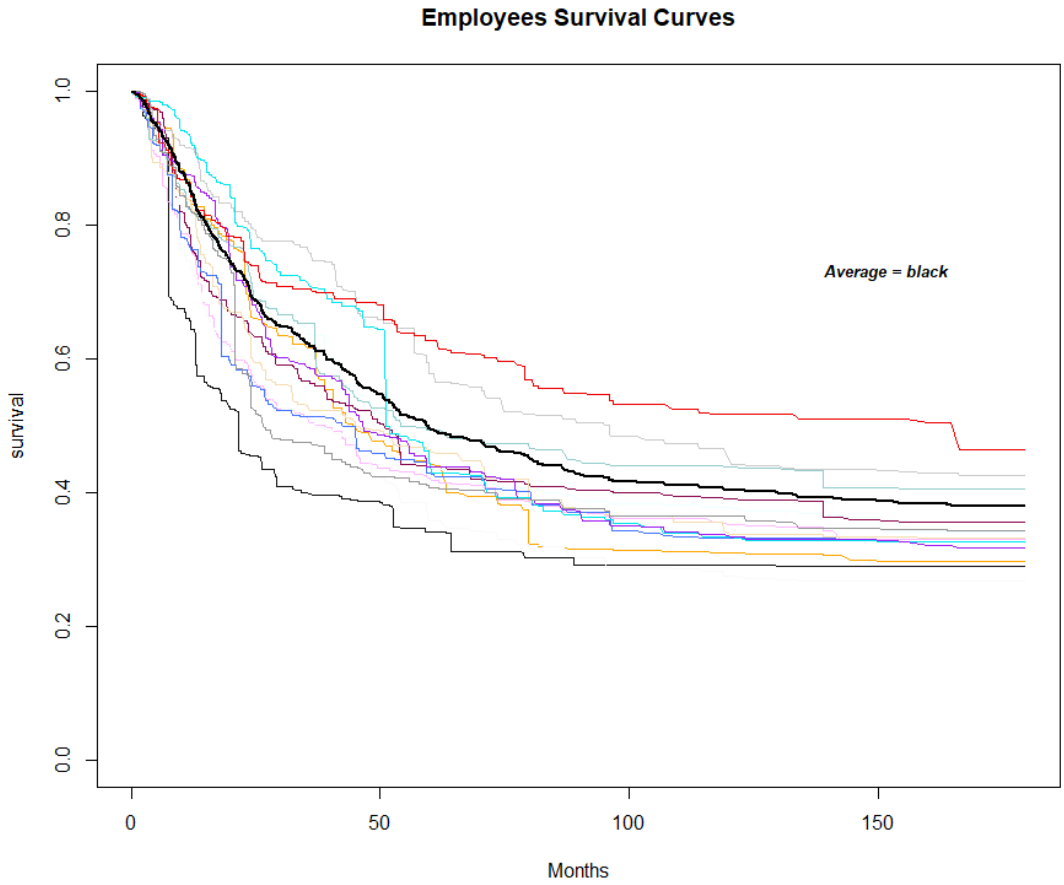
Παρακάτω στον πίνακα (17), παρουσιάζεται η στατιστική σημαντικότητα των 14 μεταβλητών που χρησιμοποιήθηκαν στην κατασκευή του μοντέλου. Βλέπουμε πως οι '*industry*','*traffic*' και 'way' εμφανίζουν μεγαλύτερη σημαντικότητα. Οι επεξηγηματικές μεταβλητές 'extraversion' και 'independ' είναι οι λιγότερο στατιστικά σημαντικές, τις οποίες ίσως και να μπορούσαμε να απομακρύνουμε κατά την δημιουργία του μοντέλου.

variables	***importance***
industry	0,0184103
traffic	0,0182823
way	0,0182549
age	0,0146638
head_gender	0,0107379
greywage	0,0103154
selfcontrol	0,0089768
coach	0,0080391
anxiety	0,0071337
gender	0,0070722
profession	0,0069942
novator	0,0065058
independ	0,0062616
extraversion	0,0062418

Πίνακας 17 - : Στατιστική σημαντικότητα μεταβλητών από την δημιουργία του μοντέλου 'RSF'

Αφού έχουμε αποκτήσει τις καμπύλες επιβίωσης, οι οποίες έχουν αναπαραχθεί στους καταληκτικούς κόμβους των δέντρων που έχουν δημιουργηθεί, μπορούμε να υπολογίσουμε την μέση καμπύλη επιβίωσης για τον συνολικό αριθμό εργαζομένων. Στο παρακάτω διάγραμμα (46), 20 τυχαίες καμπύλες που αναπαρήχθησαν από τα δέντρα επιβίωσης του μοντέλου, ενώ με σκούρο μαύρο χρώμα απεικονίζεται η μέση συνολική καμπύλη επιβίωσης των εργαζομένων του

δείγματος. Παρατηρούμε πως, σύμφωνα με το μοντέλο η πιθανότητα κάποιος εργαζόμενος να παραμείνει στην εταιρία μετά την χρονική περίοδο των 100 μηνών είναι κοντά στο ποσοστό του 40%. Επίσης, αξίζει να σημειωθεί πως η πιθανότητα να παραμείνει κάποιος εργαζόμενος στην εταιρία παραμένει σταθερή *'SurvProd' ≈ 40%* μέχρι και το τέλος της μελέτης. Διακρίνουμε πως οι περισσότεροι εργαζόμενοι αποχώρησαν από την εταιρία μεταξύ του χρονικού διαστήματος (0,100] μηνών. Ενώ από εκεί και έπειτα η κατάσταση σταθεροποιείται, με λιγότερους να εκδηλώνουν το γεγονός. Η ενδιάμεση πιθανότητα (πιθανότητα κατά μέσο όρο) κάποιος εργαζόμενος να παραμείνει στην εταιρία (*SurvProd' ≈ 50%)* είναι μετά την χρονική περίοδο των 60 μηνών.



Διάγραμμα 46 - Συνολική καμπύλη επιβίωσης των εργαζομένων (σκούρο μαύρο χρώμα) μεταξύ άλλων τυχαίων καμπυλών επιβίωσης εργαζομένων

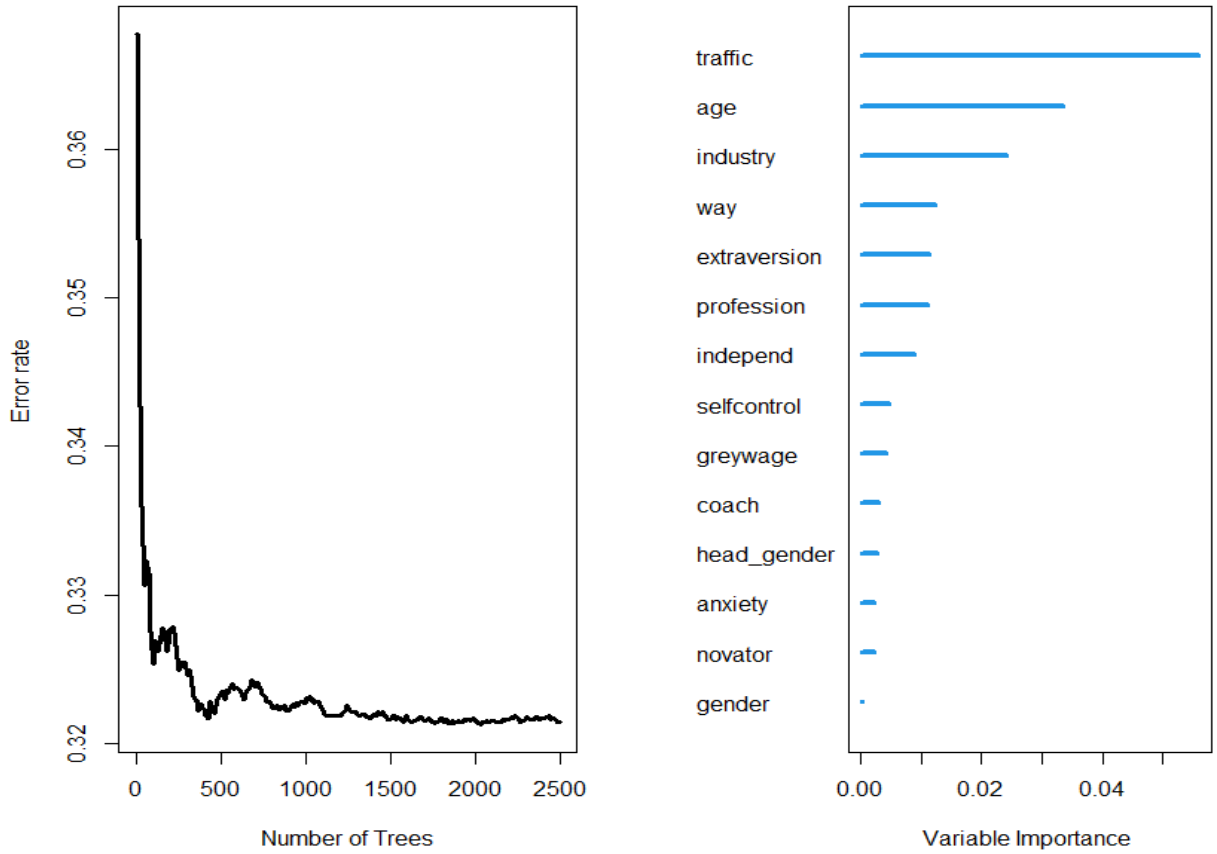
Θα προσπαθήσουμε να χωρίσουμε το σύνολο δεδομένων που διαθέτουμε σε δύο επιμέρους υποσύνολα. Το ένα θα χρησιμοποιηθεί προκριμένου να εκπαιδεύσουμε τον αλγόριθμο των '*Random Survival Forests*' και με το άλλο για να προβούμε σε προβλέψεις του χρόνου επιβίωσης των εργαζομένων και να ελέγξουμε την απόκλιση από τον πραγματικό χρόνο επιβίωσης των εργαζομένων. Οι 990 τυχαία επιλεγμένες παρατηρήσεις από το αρχικό σύνολο δεδομένων θα χρησιμοποιηθούν προκειμένου να εκπαιδευτεί ο αλγόριθμός και οι υπόλοιπες 139 παρατηρήσεις

προκειμένου να γίνουν οι προβλέψεις. Για την εκπαίδευση του μοντέλου '*Random Survival Forests*' στους 990 εργαζομένους, επιλέγουμε ότι θα αναπαραχθούν συνολικά 490 δέντρα επιβίωσης, ο υποψήφιος αριθμός των επεξηγηματικών μεταβλητών που επιλέγονται κάθε φορά προκειμένου να γίνει ο διαχωρισμός στον κόμβο είναι '*mtry' = 4* και στον κάθε καταληκτικό κόμβο θέτουμε εξ 'αρχής ότι συμπεριλαμβάνονται 20 εργαζόμενοι οι οποίοι αποχώρησαν από την εταιρία. Σύμφωνα με πίνακα (18), βλέπουμε πως 502 φορές συνέβη το γεγονός (να αποχωρήσει κάποιος εργαζόμενος από την εταιρία) από το σύνολο των 990 εργαζομένων του δείγματος. Ο μέσος όρος, των 36,5 καταληκτικών κόμβων που δημιουργήθηκαν από το μοντέλο μας. Ο στατιστικός έλεγχος που χρησιμοποιήθηκε προκειμένου να γίνει ο διαχωρισμός του κάθε κόμβου είναι το '*logrank test*'. Τέλος βλέπουμε πως το σφάλμα πρόβλεψης για το '*training set*' είναι της τάξεως του 32.19%.

Type	***Survival***
Number of trees	490
Sample size	990
Number of deaths	488
Forest terminal node size	20
Average no. of terminal nodes	36,5551
No. of variables tried at each split	4
Total no. of variables	14
Resampling used to grow trees	swor
Resample size used to grow trees	626
Analysis	RSF
Splitting rule	logrank *random*
Number of random split points	10
Error rate	**32,14%**

Πίνακας 18 - Αποτελέσματα ανάλυσης από την εφαρμογή της μεθόδου 'RSF' στο 'training set'

Παρατηρώντας το διάγραμμα (47), το σφάλμα πρόβλεψης (*OOB error*) γίνεται ελάχιστο με την δημιουργία περίπου 490 διαφορετικών δέντρων επιβίωσης. Με την αναπαραγωγή περισσότερων δέντρων διακρίνουμε μία μικρή ανοδική πορεία στο σφάλμα. Αυτό συμβαίνει μέχρι και την δημιουργία 1000 δέντρων, όπου φαίνεται μειώνεται και να σταθεροποιείται στην συνέχεια το σφάλμα στο '*training set*'. Η επεξηγηματική μεταβλητή '*traffic*' βλέπουμε πως καταρτίζει τον υψηλότερο βαθμό σημαντικότητας για την ανάπτυξη του μοντέλου. Ακολουθούν, οι μεταβλητές '*age*' και '*industry*' όπου βλέπουμε πώς είναι οι επόμενες στατιστικά σημαντικότερες μεταβλητές. Οι '*gender*' και '*novator*' βλέπουμε πως είναι οι λιγότερο σημαντικές για την δημιουργία του μοντέλου.



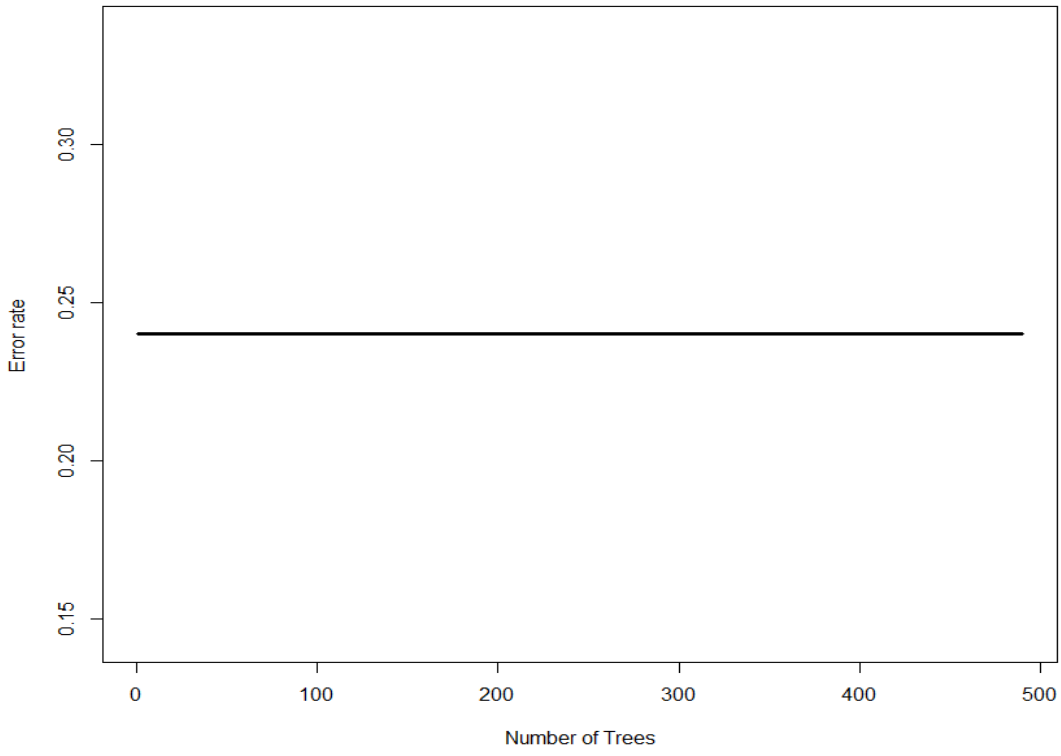
Διάγραμμα 47 - 'OOB error rate' αναλόγως με το μέγεθος του δέντρου και η στατιστική σημαντικότητα της καθεμίας μεταβλητής

Εκτελώντας το μοντέλο που δημιουργήσαμε στο υποσύνολο των 139 εργαζομένων '*test set*' προκειμένου να γίνουν προβλέψεις, παρατηρούμε, όπως φαίνεται και στον πίνακα (19), ότι το σφάλμα πρόβλεψης *'test set error rate' = 23,99%* μειώθηκε κατά σημαντικό βαθμό σε σχέση με το σφάλμα πρόβλεψης στο *'training set' = 33,21%*. Συμπεραίνουμε ότι, το μοντέλο μας απέδωσε καλύτερα σε ξένα δεδομένα παρά στα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση του αλγορίθμου. Κάτι το οποίο θα μπορούσαμε να πούμε πως αποδεικνύει την αξιοπιστία του μοντέλου.

Type	***Survival***
Number of trees	490
Sample size	139
Number of deaths	70
Average no. of terminal nodes	36,5551
Total no. of grow variables	14
Resampling used to grow trees	swor
Resample size used to grow trees	88
Analysis	RSF
Test set error rate	**23,99%**

Πίνακας 19 - Αποτελέσματα ανάλυσης από την εφαρμογή του μοντέλου 'RSF' στο 'test set'

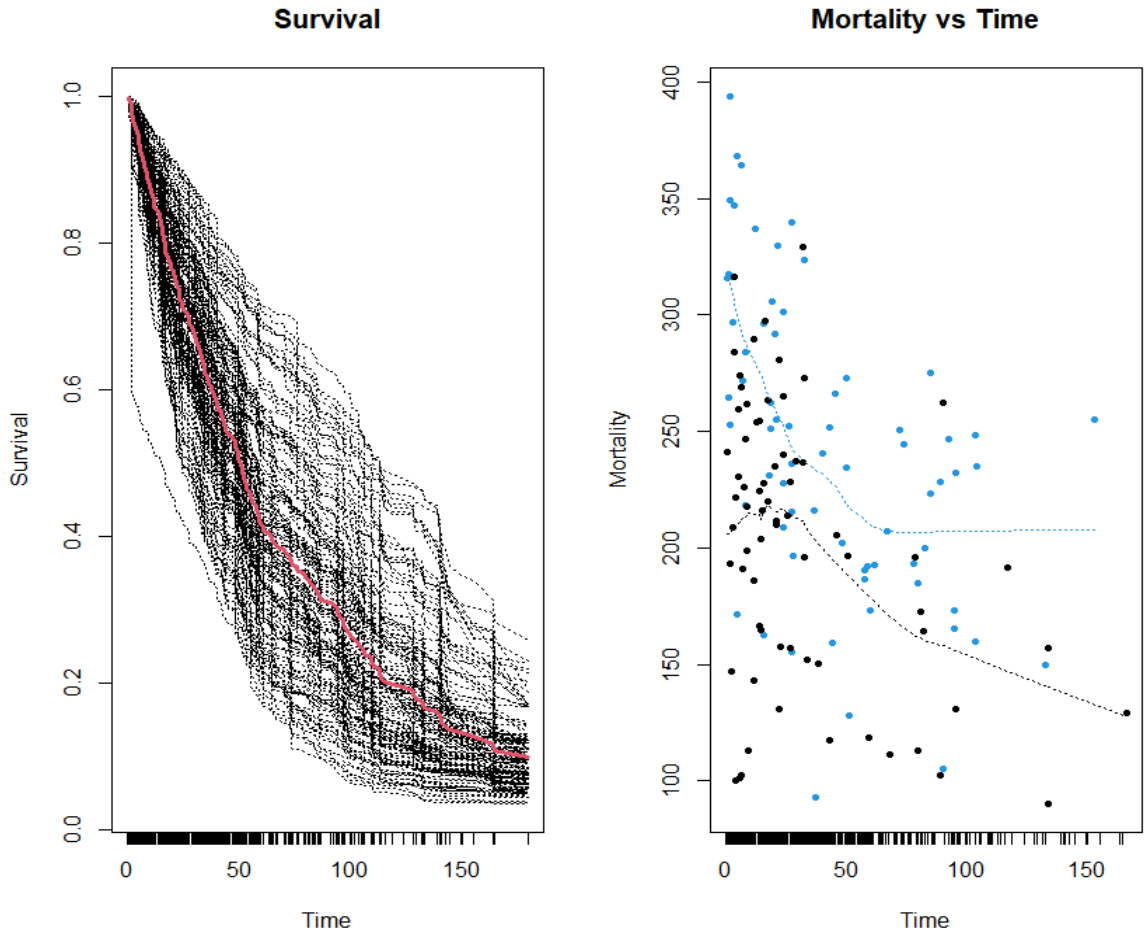
Στο παρακάτω διάγραμμα (48), διακρίνουμε πως δεν μπορούμε να βελτιώσουμε (ελαχιστοποιήσουμε) το σφάλμα πρόβλεψης στο *'test set'*, τροποποιώντας το αριθμό των δέντρων θέλουμε να αναπαραχθούν από το δέντρο. Βλέπουμε πως το σφάλμα πρόβλεψης δεν επηρεάζεται από τον αριθμό των δέντρων που θα αναπτυχθούν.



Διάγραμμα 48 - Σφάλμα πρόβλεψης αναλόγως με τον αριθμό των δέντρων του μοντέλου

Τέλος, σύμφωνα με το διάγραμμα (49), διαπιστώνουμε πως καθώς η χρονική περίοδος της μελέτης αυξάνεται, όλο και περισσότεροι εργαζόμενοι φαίνεται να εκδηλώνουν το γεγονός. Η

μεγαλύτερη πτώση στην καμπύλη επιβίωσης των εργαζομένων παρουσιάζεται μέχρι την χρονική περίοδο των 100 μηνών. Η σκούρα κόκκινη καμπύλη απεικονίζει την προβλεπόμενη μέση καμπύλη επιβίωσης για τους εργαζομένους στο *'test set'*. Επίσης, παρατηρώντας το δεύτερο διάγραμμα *'Mortality vs Time'* με μπλε κουκίδες απεικονίζονται εκείνοι οι εργαζόμενοι οι οποίοι αποχώρησαν από την εταιρία, ενώ οι μαύρες κουκίδες υποδηλώνουν τους εργαζομένους που δεν έχουν εκδηλώσει το γεγονός.

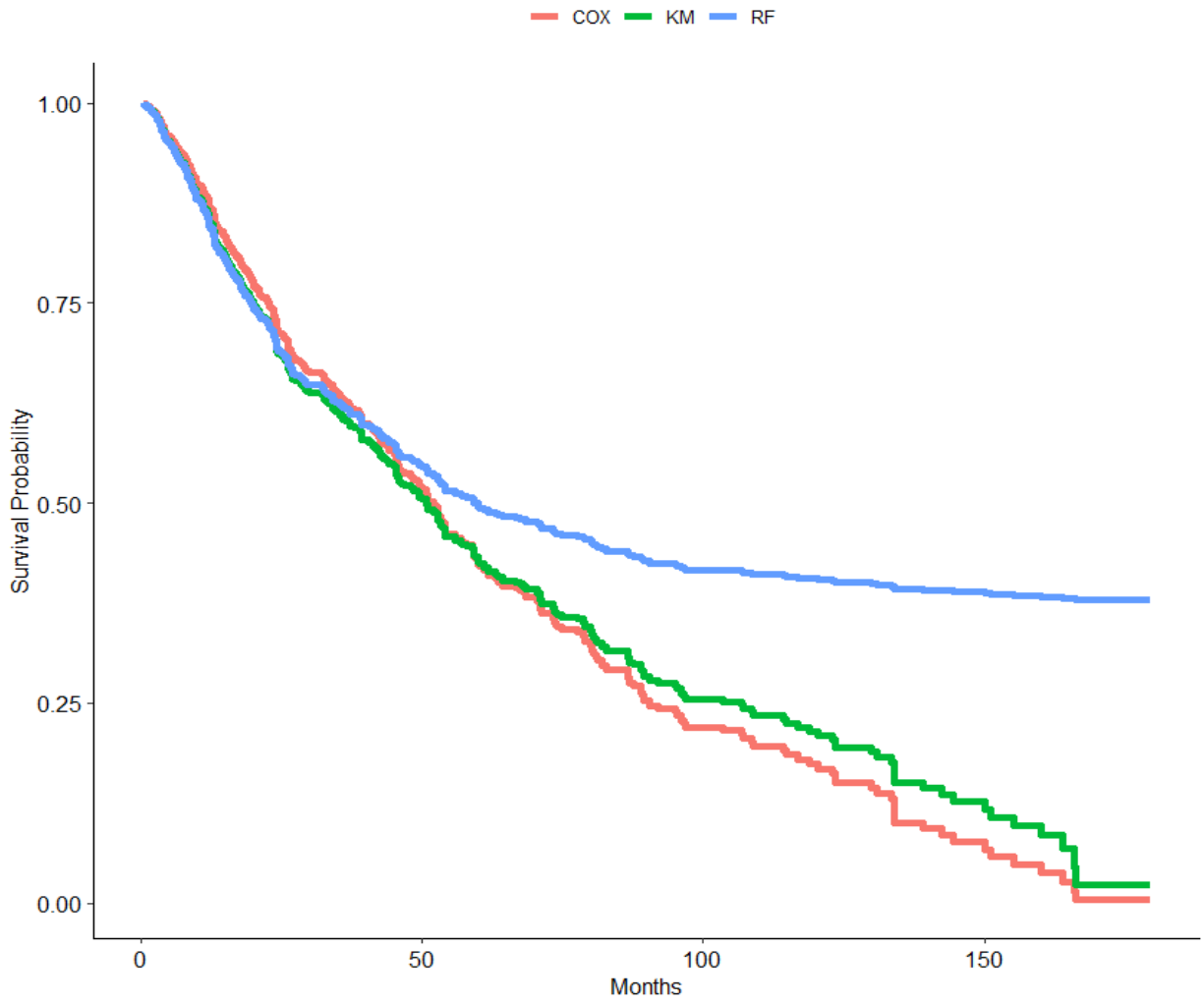


Διάγραμμα 49 - Καμπύλη επιβίωσης των εργαζομένων στο 'test set' και η εκδήλωση του γεγονότος σε σχέση με τον χρόνο

5. Συμπεράσματα

Για την μοντελοποίηση του κύκλου ζωής των εργαζομένων σε κάποια εταιρία χρησιμοποιήθηκαν τρεις κύριες διαφορετικές μέθοδοι. Όπως είδαμε, στα πρώτα στάδια της ανάλυσης μας, έγινε χρήση της μη παραμετρικής μεθόδου Kaplan-Meier. Ύστερα εισαγάγαμε το ημί-παραμετρικό μοντέλο *'Cox PH'* προκειμένου να ελέγξουμε την επιρροή την καθεμίας επεξηγηματικής μεταβλητής στον βαθμό κίνδυνου να αποχωρήσει κάποιος εργαζόμενος από την εταιρία και τις σημαντικότερες μεταβλητές στην δημιουργία του μοντέλου. Στο τέλος με την χρήση των δέντρων επιβίωσης και της τεχνικής των *'Random Survival Forests'* δημιουργήθηκε το καταληκτικό μας μοντέλο, ελέγχοντας την απόδοση αλλά και την ερμηνευτική του ικανότητα.

Σύμφωνα με το διάγραμμα (50), αναπαριστώντας στο ίδιο διάγραμμα τις καμπύλες επιβίωσης των τριών μοντέλων που κατασκευάσαμε, παρατηρούμε ότι οι τρεις καμπύλες επιβίωσης είναι σχεδόν πανομοιότυπες μέχρι την χρονική περίοδο των 50 μηνών. Η πιθανότητα παραμονής τις εργαζομένου (επιβίωσης) έπειτα από την χρονική περίοδο των 50 μηνών *(SurvProb ≈ 55%)* φαίνεται πως είναι πανομοιότυπη και για το τρία μοντέλα. Η καμπύλη επιβίωσης που έχει αποκτηθεί από την εκτέλεση τις μεθόδου *'RSF'* διακρίνουμε πως εμφανίζει μία πιο θετική εικόνα για τις εργαζομένους, όσον αφορά την παραμονή τις στην εταιρία. Σύμφωνα με το διάγραμμα, η κατακόρυφη πτώση τις καμπύλης *'RF'*, από τις εργαζομένους που εκδήλωσαν πιο συχνά το γεγονός, συμβαίνει μέχρι την χρονική περίοδο των 50 μηνών. Στην συνέχεια βλέπουμε μία μικρή πτώση για την καμπύλη *'RF'* μέχρι την χρονική περίοδο των 100 μηνών, ενώ από εκεί και έπειτα η καμπύλη επιβίωσης *'RF'* σταθεροποιείται με την πιθανότητα επιβίωσης για κάποιον εργαζόμενο στο τέλος τις μελέτης να είναι τις τάξεως *'SurvProb' ≈ 40%.* Αντιθέτως, η εικόνα που αποτυπώνεται από τις τις δύο καμπύλες επιβίωσης *('COX', 'KM')* όσον αφορά την επιβίωση (παραμονή) εργαζομένων στην εταιρία θα λέγαμε πως είναι αρκετά δυσάρεστη. Η αυξημένη αποχώρηση εργαζομένων από την εταιρία και για τις δύο καμπύλες, συνεχίζεται μέχρι και την χρονική περίοδο των 100 μηνών. Από εκεί και έπειτα, παρατηρούμε μία μεγαλύτερη πτώση στην πιθανότητα επιβίωσης για την '*KM*' καμπύλη, με την αποχώρηση των εργαζομένων από την εταιρία να είναι πιο συχνή. Στο τέλος τις μελέτης, παρατηρούμε πως η πιθανότητα παραμονής κάποιου εργαζομένου είναι πάρα πολύ μικρή και για τα δύο μοντέλα ('*Cox PH', 'Kaplan-Meier'*) με την πιθανότητα επιβίωσης να αγγίζει το *'SurvProd' ≈ 0%.* Αξίζει να σημειωθεί πως η τεχνική των '*Random Survival Forests*' αποδίδει καλύτερα σε σύνολα δεδομένων με αρκετά μεγάλο αριθμό παρατηρήσεων και μεταβλητών, σε αντίθεση με την μέθοδο '*Kaplan-Meier*'. Τις η μέθοδος '*Cox PH*' προτιμάται τις περισσότερες φορές καθώς μπορεί να ελεγχθεί και να συνυπολογιστεί η επίδραση των επεξηγηματικών μεταβλητών τον κίνδυνο να εκδηλώσει το γεγονός κάποιος εργαζόμενος.



Διάγραμμα 50 - Σύγκριση μοντέλων επιβίωσης

6. Βιβλιογραφία

Alvarez-Iglesias, A. (2012, September 17). Extensions and Applications of Survival Trees in Medical Data. *NUI Galway*, σσ. 1-164.

BIOST 515. (2004, Φεβρουάριος 26). Introduction to Survival Analysis.

Bou-Hamad, I. (2009, May). Discrete-Time Survival Trees. 2-100.

Bou-Hamad, I., Larocque, D., & Ben-Ameur, H. (2011, September 12). A review of survival trees. *Statistics Surveys*, σσ. 46-66.

Breiman, L., Friedman, J., Stone, C., & Olshen, R. (1984). *Classification and Regression Trees* (Τόμ. 1st). Chapman and Hall.

Brentnall, A., & Cuzick, J. (2018). Use of the concordance index for predictors of censored survival data. *Statistical Methods in Medical Reserch*, σσ. 2359-2373.

Clark, T., Bradburn, M., Love, S., & Altman, D. (2003). Survival Analysis Part I: Basic concepts and first analyses. *British Journal of Cancer*, σσ. 232-238.

De Rose, A., & Pallara, A. (1997). Survival Trees: An Alternative Non-Parametric Multivariate Technique for Life History Analysi. *European Journal of Populati*, σσ. 223-241.

Dean, L. S. (2007). A METHOD FOR DETECTING OPTIMAL SPLITS OVER TIME IN SURVIVAL ANALYSIS USING TREE-STRUCTURED MODELS . *University of Pittsburgh* , σσ. 1-106.

Emgineering Statistics. (2013, Απρίλιος 30). Censoring.

Emmert-Streib, F., & Dehmer, M. (2019, September 8). Introduction to Survival Analysis in Practice. *Machine Learning and Knowledge Extraction*, σσ. 1014-1038.

Erik Drysdale. (2017, Ιανουάριος 12). *Introduction to survival analysis*. Retrieved from http://www.erikdrysdale.com/survival/

Etikan, I., Abubakar, S., & Alkassim, R. (2017, Φεβρουάριος 13). The kaplan meier estimate in survival analysis. *Biometrics & Biostatistics International Journal*, pp. 55-59.

Goel, M., Khanna, P., & Kishore, J. (2010, Οκτώμβριος). Understanding survival analysis: Kaplan-Meier estimate. *Int J Ayurveda Res.*

GraphPad. (2016, March 16). *Hazard ratio from survival analysis.* Ανάκτηση από https://www.graphpad.com/support/faq/hazard-ratio-from-survival-analysis/#:~:text=Hazard%20is%20defined%20as%20the,rate%20in%20the%20other%20group

Guo, W. (2011, Ιούλιος 22). Chapter 2 -Basic Quantiles and Models (II).

Imani, F., Chen, R., Tucker, C., & Yang, H. (2019, Αύγουστος). ResearchGate. *Random Forest Modeling for Survival Analysis of Cancer Recurrences*, σσ. 399-404.

Ishwaran, H., & Lu, M. (2019). Random Survival Forests. *Wiley StatsRef: Statistics Reference Online*, σσ. 1-13.

Kleinbaum, D., & Klein, M. (2005). *Survival Analysis.* New York: Springer.

Kogalur, U., Blackstone, E., & Lauer, M. (2008, December). RANDOM SURVIVAL FORESTS. *The Annals of Applied Statistics*, σσ. 1-22.

Kunchhal, R. (2020, December 9). Out-of-Bag (OOB) Score in the Random Forest Algorithm. *Data Science Blogathon*.

LeBLANC, M., & CROWLEY, J. (1993, June). Survival Trees by Goodness of Split. *Journal of the American Statistical Association*, σσ. 457-467.

Leung, K.-M., Elashoff, R., & Afifi, A. (1997). CENSORING ISSUES IN SURVIVAL. California, Los Angeles.

Lewinshon, E. (2020, August). *Introduction to Survival Analysis.* Retrieved from towards data science.

Lewinson, E. (2020, Αύγουστος 17). *Introduction to Survival Analysis: the Kaplan-Meier estimator.* Retrieved from Towards Data Science: https://towardsdatascience.com/introduction-to-survival-analysis-the-kaplan-meier-estimator-94ec5812a97a

Lisa, S. (2016, June 3). Survival Analysis. Boston.

Malgorzata, K. (2010). The Influence of Censoring for the Performance of Survival Tree Ensemble. *Bialystok University of Technology*, σσ. 524-531.

Minitab Blog Editor. (2016, Δεκέμβριος 7). *The Difference Between Right-, Left- and Interval-Censored Data.* Retrieved from Minitab: https://blog.minitab.com/en/michelle-paret/the-difference-between-right-left-and-interval-censored-data

Myte, R. (2013). A comparison case study between Random Survival Forests and the Cox Proportional-Hazards model. *Covariate Selection for Colorectal Cancer Survival Data*, σσ. 5-36.

Prinja, S., Nidhi , G., & Ramesh, V. (2010, April). Censoring in Clinical Trials: Review of Survival Analysis Techniques. *Indian J Community Med*.

PySurvival. (2019). *C-index*. Ανάκτηση από Square Open Source: https://square.github.io/pysurvival/metrics/c_index.html

Raykar, V., Steck, H., Krishnapuram, B., Dehing-Oberije, C., & Lambin, P. (2007). On Ranking in Survival Analysis: Bounds on the Concordance Index. *Advances in Neural Information Processing Systems 20*, σσ. 1-8.

Segal, M. R. (1988, March). Regression trees for censored data. *Biometrics*, σσ. 35-47.

Shah, S. (1998). Survival treesQ a transition approach. *School of Mathematics and Applied Statistics*, σσ. 1-146.

Shimokawa, A., Kawasaki, Y., & Miyaoka, E. (2016). A comparative study on splitting criteria of a survival tree based on the Cox proportional model. *Journal of Biopharmaceutical Statistics*, σσ. 386-401.

Sliva, A. A. (2019, October 8). *Concordance Index as an Evaluation Metric.* Ανάκτηση από Analytics Vidha: https://medium.com/analytics-vidhya/concordance-index-72298c11eac7#id_token=eyJhbGciOiJSUzI1NiIsImtpZCI6Ijc4M2VjMDMxYzU5ZTExZjI1N2QwZWMxNTcxNGVmNjA3Y2U2YTJhNmYiLCJ0eXAiOiJKV1QifQ.eyJpc3MiOiJodHRwczovL2FjY291bnRzLmdvb2dsZS5jb20iLCJuYmYiOjE2MTA5ODg1NzksImF1

STHDA. (χ.χ.). *Cox Proportional-Hazards Model*. Ανάκτηση από http://www.sthda.com/english/wiki/cox-proportional-hazards-model

Swiss Federal Institute of Technology Zurich. (2012). Applied Multivariate Statistics. Zurich.

Therneau, T., & Atkinson, E. (2020, September 25). Concordance. σσ. 1-13.

Wienke, A. (2007, Ιανουάριος 16). Frailty Models in Survival Analysis. Γερμανία.

Wikipedia. (2021, Ιανουάριος 17). *Kaplan -Meier estimator*. Retrieved from https://en.wikipedia.org/wiki/Kaplan%E2%80%93Meier_estimator

Zhu, R. (2013). TREE-BASED METHODS FOR SURVIVAL ANALYSIS AND HIGH-DIMENSIONAL DATA. Chapel Hill.