ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ

ΚΑΙ

ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Διαγνωστικές Τεχνικές Σε Μοντέλα Επιβίωσης

Διπλωματική Εργασία

ΘΕΟΔΩΡΕΛΛΗ ΜΑΡΙΑ

Επιβλέπουσα: Καρώνη Χρυσηίς, Αναπλ. Καθηγήτρια Ε.Μ.Π.

Τριμελής Εξεταστική Επιτροπή

Χ. Καρώνη	**Γ. Κοκολάκης**	**Ι. Σπηλιώτης**
Αναπλ. Καθηγήτρια Ε.Μ.Π.	**Καθηγητής Ε.Μ.Π.**	**Αναπλ.Καθηγητής Ε.Μ.Π.**

Αθήνα, Οκτώβριος 2012

ΠΕΡΙΛΗΨΗ

Το μοντέλο αναλογικής διακινδύνευσης του Cox είναι το πιο διαδεδομένο μοντέλο στην ανάλυση επιβίωσης και χρησιμοποιείται για την εύρεση της σχέσης μεταξύ μιας μεταβλητής που δηλώνει το χρόνο επιβίωσης ενός ατόμου και άλλων συμμεταβλητών. Οι παρατηρήσεις που εκφράζουν το χρόνο επιβίωσης του ατόμου μπορεί να είναι αποκομμένες ή και πλήρεις. Το μοντέλο του Cox μοντελοποιεί τη συνάρτηση διακινδύνευσης σε σχέση με άλλες μεταβλητές και είναι ένα ημιπαραμετρικό μοντέλο. Η εκτίμηση των συντελεστών παλινδρόμησης επιτυγχάνεται μέσω της συνάρτησης μερικής πιθανοφάνειας.

Στο πρώτο κεφάλαιο γίνεται μια εισαγωγή στην ανάλυση επιβίωσης όπως και τα αποκομμένα δεδομένα και τα είδη τους. Στο δεύτερο κεφάλαιο αναπτύσσεται το μοντέλο αναλογικής διακινδύνευσης του Cox, ορίζονται οι βασικές του συναρτήσεις όπως και η έννοια της μερικής πιθανοφάνειας. Στο επόμενο κεφάλαιο ορίζονται τα υπόλοιπα για το μοντέλο του Cox και αναφέρονται οι διάφοροι έλεγχοι που αφορούν την καταλληλότητα του μοντέλου.

Στο τέταρτο κεφάλαιο εισάγεται η έννοια των απόμακρων σημείων καθώς και οι μέθοδοι μέσω υπολοίπων που υπάρχουν για την ανίχνευσή τους στο μοντέλο του Cox. Το θέμα του πέμπτου κεφαλαίου είναι ο ορισμός και η αναγνώριση των σημείων επιρροής για το μοντέλο του Cox και αναφέρονται οι μέθοδοι που χρησιμοποιούνται για την αναγνώρισή τους. Στο τελευταίο κεφάλαιο δίνεται ένα παράδειγμα επιβίωσης και εφαρμόζονται οι παραπάνω μέθοδοι για την στατιστική ανάλυσή του.

ABSTRACT

The Cox proportional hazards model is the most well-recognised model in survival analysis for exploring the relationship between a variable that shows the survival time of a person and other covariates. The observations that show the survival time of a person might be censored or uncensored. The Cox model is a semiparametric model the hazard function . The regression coefficients are estimated by the partial likelihood function.

The first chapter is an introduction to reliability analysis as also uncensored data and their types. In the second chapter the Cox proportional hazards model is developed, its basics functions and the meaning of partial likelihood as well. In the next chapter are referred the residuals for Cox model and for several tests that concern the validity of the model.

In the fourth chapter introduces the meaning of outliers and the methods through residuals that are available for outlier detection for Cox model. The subject of the fifth chapter is the definition and the influence detection for Cox and are referred for their detection. On the last chapter is given an example of survival analysis and all these methods are applied for its statistical analysis.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω θερμά την Αναπλ. Καθηγήτρια του Εθνικού Μετσόβιου Πολυτεχνείου κ. Χρυσηίς Καρώνη, για τη συνεχή ενθάρρυνση, καθοδήγηση και εμπιστοσύνη που έδειξε καθ' όλη τη διάρκεια εκπόνησης αυτής της διπλωματικής.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου για τα εφόδια που μου προσέφεραν, τη φροντίδα, τη συμπαράσταση και την υπομονή τους.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 3

ΥΠΟΛΟΙΠΑ ΣΤΟ ΜΟΝΤΕΛΟ ΤΟΥ COX

ΚΕΦΑΛΑΙΟ 4

Ανίχνευση απόμακρων παρατηρήσεων (Outlier detection)

ΚΕΦΑΛΑΙΟ 5

Επιρροή (influence)

ΚΕΦΑΛΑΙΟ 6

ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΑΝΑΛΥΣΗ ΕΠΙΒΙΩΣΗΣ

1.1 Γενικά

Η ανάλυση επιβίωσης (survival analysis) αναφέρεται στην ανάλυση δεδομένων που αφορούν στο χρόνο που μεσολαβεί μέχρι κάποιο συγκεκριμένο συμβάν. Αρχικά, η ανάλυση αναφερόταν στο χρόνο μεταξύ της θεραπείας μέχρι το θάνατο και για αυτό το λόγο πήρε και το συγκεκριμένο όνομα. Η ανάλυση επιβίωσης όμως μπορεί να εφαρμοστεί σε αρκετές περιπτώσεις, όπως για παράδειγμα στη μηχανολογία, για την ανάλυση του χρόνου μέχρι την εμπλοκή ενός μηχανήματος ή τη γεωργία, για την ανάλυση του χρόνου μέχρι την στιγμή να βγάλει καρπό ένα δέντρο. Στην περίπτωση της μηχανολογίας η ανάλυση αναφέρεται και ως θεωρία αξιοπιστίας (reliability theory).

Το επίπεδο γνώσης που κατέχουμε σήμερα για την Ανάλυση Επιβίωσης προκύπτει από μία διαδικασία εξέλιξης που διήρκησε αρκετά χρόνια και η οποία είχε ιδιαίτερη ανάπτυξη τα τελευταία χρόνια. Διάφοροι τομείς έρευνας, όπως Βιολογία, Ιατρική, Φαρμακευτική ακόμα και κλάδοι της Βιομηχανίας κ.α. συνέβαλαν σε αυτή την ανάπτυξη, προσπαθώντας να βρουν λύση για να αντιμετωπίσουν τα διάφορα προβλήματά τους.

Το 1662 στην Μ. Βρετανία δημοσιεύτηκε το πρώτο βιβλίο, το οποίο είχε καταγεγραμμένους καταλόγους με γεννήσεις και θανάτους, που αναφέρονταν στις προηγούμενες δεκαετίες κι έτσι ήταν η πρώτη φορά που οι "θάνατοι" αντιμετωπίστηκαν ως "γεγονότα", για τα οποία έγιναν αναλυτικές μελέτες.

Οι παγκόσμιοι πόλεμοι που ακολούθησαν, έδωσαν το έναυσμα για ανάπτυξη της έρευνας στον τομέα της αξιοπιστίας και στη μελέτη της διάρκειας ζωής των στρατιωτικών στρατευμάτων, αλλά και αργότερα η έρευνα επικεντρώθηκε στη μελέτη κάποιων ιδιαίτερων πιθανοθεωρητικών προβλημάτων σχετιζομένων με

την "παύση λειτουργίας" και την "αντικατάσταση" εξαρτημάτων μηχανικών ή ηλεκτρικών κυκλωμάτων, όπως κάποιας βαλβίδας ή ενός θερμοστάτη σε ένα μηχανικό κύκλωμα, μίας λυχνίας ή μίας αντίστασης σε ένα ηλεκτρικό. Υπήρξε δηλαδή μεγάλη πρόοδος στη Βιομηχανία των ηλεκτρονικών συσκευών.

Έπειτα οι μέθοδοι της Ανάλυσης Επιβίωσης είχαν τεράστιες εφαρμογές σε κλινικά δεδομένα και σκοπός ήταν να απαντηθούν ερωτήματα όπως ποια είναι η πιθανότητα ένας ασθενής να ζήσει μέχρι μια συγκεκριμένη χρονική στιγμή ή με ποιο ρυθμό θα πεθάνουν κάποιοι ασθενείς, οι οποίοι έχουν ήδη επιβιώσει μέχρι ένα συγκεκριμένο χρονικό σημείο, κ.α. Η ανάπτυξη των επιχειρησιακών ερευνών κατέδειξε ότι υπάρχουν και πολλά άλλα προβλήματα και μοντέλα διαφορετικής υφής, στα οποία η Ανάλυση Επιβίωσης μπορεί να εφαρμοστεί και να προσφέρει λύσεις. Ακόμα περισσότερο με την εξέλιξη των ηλεκτρονικών υπολογιστών η εφαρμογή και η ανάπτυξη της γίνεται ολοένα και μεγαλύτερη, καθώς παράγονται όλο και περισσότερα λογισμικά πακέτα για το σκοπό αυτό.

1.2 Εισαγωγικές έννοιες

1.2.1 Χρόνος επιβίωσης

Η ανάλυση επιβίωσης είναι μια περιοχή έρευνας στη στατιστική, η οποία δημιουργήθηκε για την ανάλυση δεδομένων τα οποία δε μπορούν να επεξεργαστούν από τις συνηθισμένες στατιστικές μεθόδους. Τα δεδομένα αυτά δίνουν τη χρονική διάρκεια μέχρι να συμβεί ένα συγκεκριμένο γεγονός. Ο χρόνος επιβίωσης ή χρόνος αποτυχίας (survival time ή failure time), αναφέρεται σε μια μεταβλητή που μετράει το χρόνο (ημέρες, εβδομάδες, μήνες, κλπ) μέχρι να συμβεί ένα συγκεκριμένο γεγονός. Το γεγονός μπορεί να είναι η εμφάνιση μιας ασθένειας, η εξέλιξη ή η επιτυχία μιας θεραπείας σε κάποια ασθένεια, ο θάνατος του ασθενούς, η παύση λειτουργίας μιας ηλεκτρικής μηχανής κ.α. Ο χρόνος επιβίωσης ονομάζεται και χρόνος ως το "γεγονός" ή την "αποτυχία". Ο χρόνος επιβίωσης είναι το βασικό σημείο ενδιαφέροντος σε πολλές βιοχημικές εφαρμογές (π.χ. ο χρόνος μέχρι την αντίδραση ενός οργανισμού σε ένα φάρμακο), σε κοινωνικές (π.χ. ο χρόνος μέχρι την εγκυμοσύνη), οικονομικές επιστήμες (π.χ. ο χρόνος μέχρι ένας δείκτης να ξεπεράσει ένα

όριο) καθώς και στη μηχανική (π.χ. ο χρόνος μέχρι να χαλάσει ένα εξάρτημα μιας μηχανής). Ενδεχόμενα ερωτήματα που μπορεί να προκύψουν είναι ο χαρακτηρισμός της κατανομής του χρόνου επιβίωσης, καθώς και η σύγκριση αυτού του χρόνου μεταξύ διαφορετικών ομάδων ή ακόμη η μοντελοποίηση της σχέσης του χρόνου επιβίωσης σε σχέση με άλλες μεταβλητές.

Παράδειγμα 1.1 :

Έστω ότι θεωρούμε έναν πληθυσμό, ο οποίος αποτελείται από όμοια εξαρτήματα, π.χ. ηλεκτρικούς λαμπτήρες. Κάθε ένα από τα όμοια αυτά εξαρτήματα χαρακτηρίζεται από μια μη-αρνητική τυχαία μεταβλητή T, που παριστάνει το χρονικό διάστημα από τη στιγμή που το συγκεκριμένο εξάρτημα τίθεται σε χρήση μέχρι τη χρονική στιγμή που αυτό παύει να λειτουργεί.

Παράδειγμα 1.2 :

Σε μια κλινική έρευνα εξετάζεται ένας πληθυσμός ασθενών με καρκίνο. Η μη-αρνητική τυχαία μεταβλητή T παριστάνει το χρόνο ιατρικής παρακολούθησης του κάθε ασθενή από τη στιγμή εκδήλωσης της νόσου μέχρι το θάνατό του.

1.2.2 Συνάρτηση επιβίωσης

Η τυχαία μεταβλητή του χρόνου T μπορεί να είναι είτε διακριτή είτε συνεχής.

- Συνεχής τυχαία μεταβλητή

Στην περίπτωση συνεχούς μεταβλητής, η $f_T(t)$ παριστάνει τη συνάρτηση πυκνότητας πιθανότητας, η οποία ονομάζεται και πυκνότητα αποτυχίας, όπου $t \geq 0$ ενώ η $F_T(t) = \int_0^t f_T(x)dx = P(T \leq t)$, $t \geq 0$ παριστάνει τη συνάρτηση κατανομής της τυχαίας μεταβλητής T, ή αλλιώς συνάρτηση κατανομής αποτυχίας.

Σε μια έρευνα ενδιαφερόμαστε για την πιθανότητα η συνιστώσα του συστήματος να μην έχει "αποτύχει" έως τη χρονική στιγμή t. Η συνάρτηση που καθορίζει αυτή την πιθανότητα ονομάζεται συνάρτηση επιβίωσης (survivor function) και συμβολίζεται ως $S(t) = P(T \geq t) = 1 - F_T(t)$, $t \geq 0$.

Προφανώς αφού $F_T(0)=0$ και $F_T(1)= 0$ θα είναι $S(0)=1$ και $S(1)=0$.

- Διακριτή τυχαία μεταβλητή

Στην περίπτωση διακριτής μεταβλητής, η $p_T(t) = P(T = t)$ παριστάνει τη συνάρτηση πιθανότητας, που ονομάζεται και πυκνότητα αποτυχίας, όπου t = 0, 1, 2,… ενώ η $F_T(t) = P(T \leq t) = \sum_{x \leq t} p_T(x)$, $t \geq 0$ παριστάνει τη συνάρτηση κατανομής της τυχαίας μεταβλητής T.

1.3 Αποκομμένα Δεδομένα

Στις περιπτώσεις όπου ο χρόνος αποτυχίας δεν είναι δυνατό να παρατηρηθεί πειραματικά, (για παράδειγμα λόγω βίαιης διακοπής του πειράματος), τότε η παρατήρηση θεωρείται αποκομμένη (censored) (Klein και Moeschberger, 1997). Η πληροφορία για το χρόνο επιβίωσης ενός ατόμου στην περίπτωση αυτή είναι μερική (αφού γνωρίζουμε μόνο ένα κάτω φράγμα του χρόνου επιβίωσης).
Η έννοια των αποκομμένων δεδομένων χρησιμοποιήθηκε για πρώτη φορά από τον Hald (1949). Τα δεδομένα που δεν είναι αποκομμένα ονομάζονται μη-αποκομμένα ή πλήρη δεδομένα.

Γενικά αποκομμένες παρατηρήσεις προκύπτουν όταν μερικές από τις μονάδες του πειράματος χάνονται κατά τη διάρκειά του.

Υπάρχουν οι παρακάτω κατηγορίες αποκομμένων παρατηρήσεων:

1. Όταν ο ερευνητής επιλέγει να προκαθορίσει τη χρονική διάρκεια της έρευνας, οι χρόνοι επιβίωσης των υπό εξέταση μονάδων που "κατέληξαν" εντός της συγκεκριμένης προκαθορισμένης διάρκειας είναι οι ακριβείς και ονομάζονται μη αποκομμένοι χρόνοι. Αντιθέτως οι πραγματικοί χρόνοι των υπό εξέταση μονάδων που δεν "κατέληξαν" στη διάρκεια της έρευνας δεν είναι γνωστοί (ή ακόμα και αν είναι γνωστοί είναι μεγαλύτεροι από τη διάρκεια της έρευνας) και ονομάζονται αποκομμένοι (censored). Οι χρόνοι αυτοί θεωρούνται ίσοι με το χρόνο που διαρκεί το πείραμα, όμως δεν αντιστοιχούν στο χρόνο θανάτου, αλλά στο γεγονός ότι ήταν "ζωντανοί" μέχρι εκείνη τη στιγμή.
2. Πάλι στην περίπτωση που είναι προκαθορισμένη η χρονική διάρκεια της έρευνας, αλλά κάποιες μονάδες "καταλήγουν" εντός της συγκεκριμένης

διάρκειας για άλλους λόγους π.χ. ένας ασθενής αποφασίζει να μη συμμετέχει άλλο στην έρευνα και αποχωρεί πριν αυτή τελειώσει. Εδώ οι παρατηρούμενοι χρόνοι είναι μικρότεροι από τους πραγματικούς χρόνους επιβίωσης και ονομάζονται και αυτοί αποκομμένοι.

3. Στην περίπτωση που ο ερευνητής επιλέγει να προκαθορίσει ένα συγκεκριμένο ποσοστό επιτυχίας και μόλις το επιτύχει να σταματήσει την έρευνα, οι αποκομμένες παρατηρήσεις θεωρούνται ίσες με το χρόνο επιβίωσης της μεγαλύτερης μη αποκομμένης παρατήρησης.

1.3.1. Είδη αποκομμένων δεδομένων

Υπάρχουν 3 είδη αποκοπής, όπως αναφέραμε και προηγουμένως. Η δεξιά αποκοπή (right censoring), η αριστερή αποκοπή (left censoring) και η αποκοπή διαστήματος (interval censoring). Επιπλέον, η δεξιά αποκοπή χωρίζεται σε 3 κατηγορίες, την αποκοπή τύπου I (Type I censoring), την αποκοπή τύπου II (Type II censoring) και την τυχαία αποκοπή (random censoring), όπως επίσης υπάρχουν και παραλλαγές των παραπάνω κατηγοριών.
Θεωρούμε ότι T_i είναι ο χρόνος επιβίωσης ή χρόνος αποτυχίας του ατόμου i και u ο χρόνος στον οποίο σταματά η μελέτη.

• Δεξιά αποκοπή (right censoring):

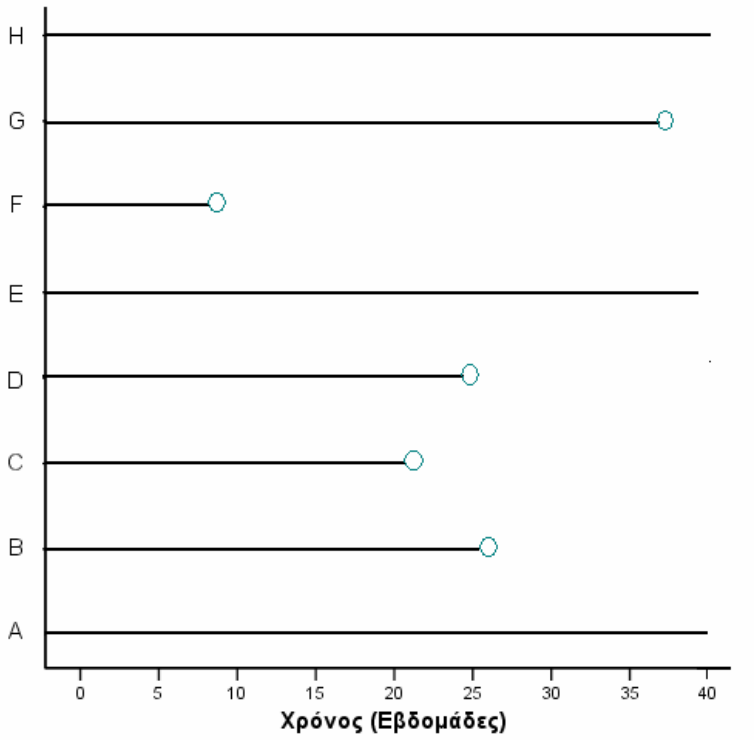
Στην περίπτωση αυτή, ο χρόνος επιβίωσης T_i, είναι μεγαλύτερος από το χρόνο u. Δηλαδή, δε γνωρίζουμε τον ακριβή χρόνο επιβίωσης του i- οστού ατόμου, γνωρίζουμε μόνο ότι ο χρόνος επιβίωσης του είναι στο διάστημα (u,U). Η δεξιά αποκοπή, είναι η πιο συνηθισμένη μορφή αποκοπής. Παρατηρείται σε περιπτώσεις όπου ένα άτομο χάνεται ή αποσύρεται από την παρακολούθηση, ή ακόμη όταν η μελέτη τερματίζεται σε ένα προκαθορισμένο χρόνο.

- ❖ Αποκοπή τύπου I (Type I censoring):

Όταν από την αρχή της έρευνας προκαθορίζεται ο χρόνος διάρκειάς της, έστω u, τότε έχουμε αποκοπή τύπου I. Ο χρόνος u ονομάζεται χρόνος αποκοπής (censoring time). Έτσι, ο ερευνητής καταγράφει τους χρόνους αποτυχίας ή επιβίωσης των ατόμων που απέτυχαν κατά τη διάρκεια της έρευνας, ενώ για τα υπόλοιπα άτομα το μόνο που είναι γνωστό είναι ότι οι χρόνοι επιβίωσης τους είναι μεγαλύτεροι από u. Στην αποκοπή τύπου I, όταν δεν υπάρχουν απώλειες από ατυχήματα, όλες οι αποκομμένες παρατηρήσεις ισούνται με το μήκος της περιόδου της μελέτης.

Παράδειγμα 1.3:

Θεωρούμε 8 ποντίκια (A,B,C,D,E,F,G,H) που υποβάλλονται σε διαδικασία καρκινογένεσης με εμβολιασμό καρκινικών κυττάρων την ίδια χρονική στιγμή. Μας ενδιαφέρει ο χρόνος που απαιτείται για την ανάπτυξη όγκου προκαθορισμένου μεγέθους. Ο ερευνητής αποφασίζει να τερματίσει το πείραμα μετά από 40 εβδομάδες (u=40). Από το Σχήμα 1.1, βλέπουμε ότι οι ποντικοί B, C, D, F και G ανέπτυξαν όγκο στους χρόνους 25, 22, 24, 8 και 38 αντίστοιχα (οι χρόνοι αυτοί είναι οι χρόνοι αποτυχίας), ενώ οι ποντικοί A,E και H δεν ανέπτυξαν όγκο κατά τη διάρκεια της μελέτης, άρα οι χρόνοι επιβίωσης τους δεν είναι γνωστοί. Έτσι, τα δεδομένα επιβίωσης είναι 40+, 25, 22, 24, 40+, 8, 38 και 40+ εβδομάδες. Τα αποκομμένα δεδομένα στην περίπτωση αυτή είναι τύπου I.



Σχήμα 1.1: Ένα παράδειγμα αποκομμένων δεδομένων τύπου I

- Αποκοπή τύπου II (Type II censoring):

Στην αποκοπή τύπου II, η μελέτη συνεχίζεται μέχρι να 'αποτύχουν' r άτομα. Ο αριθμός r καθορίζεται πριν την έναρξη της μελέτης. Έτσι, αν έχουμε n άτομα υπό μελέτη, τότε στο τέλος της μελέτης, γνωρίζουμε τους χρόνους αποτυχίας r ατόμων, ενώ για τα υπόλοιπα n-

r άτομα, γνωρίζουμε μόνο ότι ο χρόνος επιβίωσης τους είναι μεγαλύτερος από το χρόνο επιβίωσης των r ατόμων που απέτυχαν. Δηλαδή, στην αποκοπή τύπου ΙΙ, οι αποκομμένες παρατηρήσεις ισούνται με τη μεγαλύτερη μη-αποκομμένη παρατήρηση.

Για παράδειγμα, στο πείραμα με τα 8 ποντίκια, αν ο ερευνητής ήθελε να τερματίσει την έρευνα όταν 4 (r=4) από τους ποντικούς εμφανίσουν όγκο, τα δεδομένα που θα έπαιρνε θα ήταν: 25, 22, 24, 8, 25+, 25+, 25+, 25+.

- ❖ Τυχαία Αποκοπή (Random censoring):

Στην περίπτωση αυτή, ο χρόνος αποκοπής που αντιστοιχεί σε κάθε υπό παρακολούθηση άτομο δεν είναι σταθερός, αλλά είναι τυχαίος. Για παράδειγμα, σε κλινικές μελέτες, ενώ οι χρονικές στιγμές έναρξης και λήξης της έρευνας είναι προκαθορισμένες, οι ασθενείς εισέρχονται σε αυτή σε διαφορετικές (τυχαίες) χρονικές στιγμές, με αποτέλεσμα οι χρόνοι αποκοπής τους να είναι τυχαίοι.

• Αριστερή Αποκοπή (left censoring):

Το μόνο που είναι γνωστό στην περίπτωση αυτή, είναι ότι ο χρόνος επιβίωσης, Τ είναι μικρότερος από ένα χρονικό διάστημα. Ο ακριβής χρόνος επιβίωσης δεν είναι γνωστός.

Παράδειγμα 1.4:

Στην πιθανή ερώτηση 'Πότε κάπνισες για πρώτη φορά;', θα παίρναμε τριών ειδών απαντήσεις:

1. Ακριβής ηλικία στην οποία το άτομο κάπνισε για πρώτη φορά → μη-αποκομμένη παρατήρηση.

2. 'Δεν κάπνισα ποτέ'→ δεξιά αποκομμένη παρατήρηση (διότι μπορεί να αρχίσει το κάπνισμα μετά το τέλος της μελέτης) και 'Κάπνισα (ή καπνίζω), αλλά δε θυμάμαι πότε ήταν η πρώτη φορά' → αριστερά αποκομμένη παρατήρηση (αφού ο ακριβής χρόνος επιβίωσης δεν είναι γνωστός και είναι μικρότερος από την ηλικία του ερωτόμενου).

• Αποκοπή σε διάστημα (Interval censoring):

Στην αποκοπή σε διάστημα, γνωρίζουμε μόνο ότι ο χρόνος επιβίωσης Τ, βρίσκεται σε ένα διάστημα (U1,U2). Αυτού του είδους η αποκοπή, παρατηρείται συνήθως όταν έχουμε περιοδική παρακολούθηση.

Παράδειγμα 1.5:

Έστω ότι μια ομάδα ατόμων που είχαν μια ασθένεια και είναι τώρα σε ύφεση μετά από χειρουργική επέμβαση, εξετάζεται ανά τακτά χρονικά διαστήματα (έστω κάθε μήνα), για

τυχόν υποτροπίαση της ασθένειας (χρόνος αποτυχίας =χρόνος υποτροπίασης). Τότε ο ακριβής χρόνος αποτυχίας δε θα είναι γνωστός, αλλά το μόνο που θα γνωρίζουμε είναι το χρονικό διάστημα στο οποίο παρουσιάστηκε το 'γεγονός', π.χ. αν για ένα ασθενή που εξετάζεται κάθε μήνα βρέθηκε τον τρίτο μήνα που εξετάστηκε ότι υποτροπίασε, γνωρίζουμε μόνο ότι ο χρόνος αποτυχίας για τον ασθενή αυτό είναι μεταξύ 61 και 90 ημέρες, χωρίς να είναι γνωστός ο ακριβής χρόνος.

1.4 Μέση Υπολειπόμενη Ζωή

Η τυχαία μεταβλητή T εκφράζει το χρόνο μέχρι την "αποτυχία" μίας συνιστώσας ενός συστήματος. Η κατανομή του χρόνου έχει συνάρτηση πυκνότητας πιθανότητας την $f_T(t)$.

Ως Μέση υπολειπόμενη ζωή (Mean residual life at time t) ορίζεται η συνάρτηση $\mu_T(t) = E(T-t \mid T > t)$, όπου $t \geq 0$ και εκφράζει την αναμενόμενη ζωή μιας μονάδας που έχει ήδη ηλικία t, δηλαδή έχει επιβιώσει ως τη χρονική στιγμή t και εξακολουθεί να λειτουργεί. Είναι δηλαδή δείκτης γήρανσης ενός ατόμου ή ενός εξαρτήματος.

Η μέση υπολοιπόμενη ζωή $\mu_T(t) = E(T-t \mid T > t)$ μπορεί να υπολογιστεί μέσω της συνάρτησης επιβίωσης S(t).

Στην περίπτωση που ο χρόνος T είναι συνεχής μεταβλητή ισχύει $\mu_T(t) = \frac{1}{S(t)} \int_t^{\infty} S(x)dx$, για $t \geq 0$,

ενώ στην περίπτωση που είναι διακριτή μεταβλητή ισχύει: $\mu_T(t) = \frac{1}{S(t)} \sum_{x=t}^{\infty} S(t)$, για $t = 0, 1, 2, \ldots\ldots$

ΚΕΦΑΛΑΙΟ 2

ΤΟ ΗΜΙ-ΠΑΡΑΜΕΤΡΙΚΟ ΜΟΝΤΕΛΟ ΑΝΑΛΟΓΙΚΗΣ ΔΙΑΚΙΝΔΥΝΕΥΣΗΣ ΤΟΥ COX

2.1 Εισαγωγή στο μοντέλο αναλογικής διακινδύνευσης του Cox

Η εύρεση της σχέσης μεταξύ μιας μεταβλητής που δηλώνει το χρόνο επιβίωσης ενός ατόμου και άλλων συμμεταβλητών, επιτυγχάνεται συνήθως μέσω ενός μοντέλου παλινδρόμησης. Όταν έχουμε δεξιά αποκομμένα δεδομένα επιβίωσης, ένα από τα μοντέλα που χρησιμοποιείται συνήθως είναι το μοντέλο παλινδρόμησης του Cox (Cox regression model) ή διαφορετικά το μοντέλο αναλογικής διακινδύνευσης του Cox (Cox proportional hazards model).

Το μοντέλο αναλογικής διακινδύνευσης του Cox, ή σε συντομία το PH μοντέλο του Cox, παρουσιάστηκε από τον Cox το 1972. Το μοντέλο αναλογικής διακινδύνευσης του Cox, είναι ένα μοντέλο παλινδρόμησης που επιτρέπει την εξέταση των επεξηγηματικών μεταβλητών, ώστε να γίνει δυνατή η επιλογή των στατιστικά σημαντικότερων εξ' αυτών. Συγκεκριμένα, εξετάζει την επίδρασή τους στη διάρκεια ζωής του ατόμου ή αντικειμένου και χρησιμοποιείται, όπως ήδη αναφέραμε, σε δεδομένα που περιλαμβάνουν εκτός από πλήρεις και αποκομμένες παρατηρήσεις.

Το μοντέλο αυτό χρησιμοποιείται ευρέως σήμερα στην ανάλυση αποκομμένων δεδομένων επιβίωσης. Ακόμη, μας επιτρέπει να εκτιμήσουμε τον κίνδυνο θανάτου ενός ατόμου, ή την αποτυχία η επιτυχία ενός γεγονότος που μας ενδιαφέρει δεδομένου των προγνωστικών τους μεταβλητών.

Είναι αναμφισβήτητα ένα από τα δημοφιλέστερα μοντέλα που χρησιμοποιούνται στην ανάλυση επιβίωσης, λόγω της απλής εφαρμογής του και ανήκει στην οικογένεια των μοντέλων αναλογικής διακινδύνευσης.

2.2 Ορισμός βασικών συναρτήσεων

Για την κατανόηση του μοντέλου του Cox πρέπει πρώτα να ορίσουμε τις συναρτήσεις S(t), h(t), H(t). Για τις οποίες έχουμε:

2.2.1 Συνάρτηση διακινδύνευσης (hazard function)

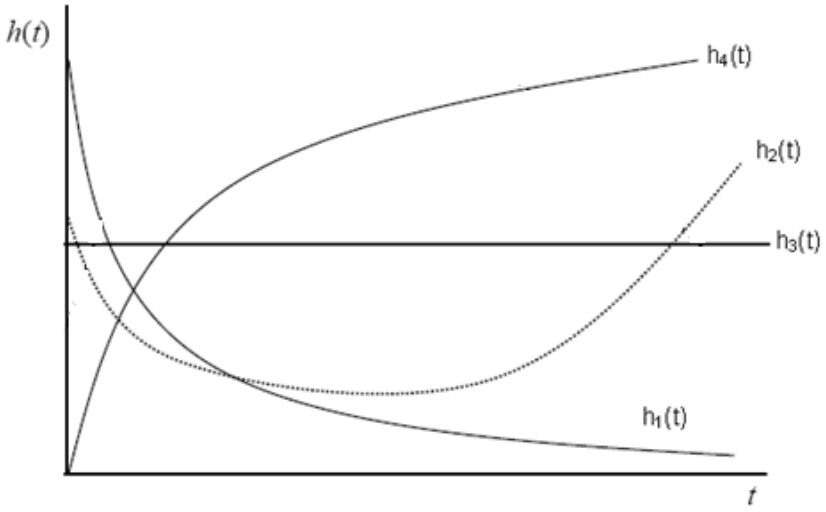
Η συνάρτηση διακινδύνευσης, h(t) , εκφράζει τον στιγμιαίο, ανά μονάδα χρόνου, ρυθμό μεταβολής της διακινδύνευσης (λόγω του ορίου) δηλαδή την στιγμιαία πιθανότητα να συμβεί το γεγονός τη συγκεκριμένη χρονική στιγμή.

Ορίζεται ως :

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

Ο ρυθμός διακινδύνευσης μπορεί να αυξάνεται, να μειώνεται, να μένει σταθερός ή να δηλώνει μια πιο περίπλοκη διαδικασία. Στο Σχήμα 2.1 φαίνονται διάφορες μορφές της συνάρτησης διακινδύνευσης.

Η συνάρτηση $h_1(t)$ είναι μια φθίνουσα συνάρτηση διακινδύνευσης που σημαίνει ότι η στιγμιαία πιθανότητα να συμβεί το γεγονός μειώνεται με το πέρασμα του χρόνου, κάτι που γενικά δεν αναμένεται να συμβεί. Δείχνει ότι σε αρχικούς χρόνους η διακινδύνευση είναι μεγάλη, ενώ όσο περνάει ο χρόνος η διακινδύνευση μειώνεται. Η $h_2(t)$ είναι η "διακινδύνευση μπανιέρας" (bathtub hazard) όπως ονομάζεται, που αποτελεί το πιο ρεαλιστικό μοντέλο διακινδύνευσης, αφού ξεκινά με πτωτική τάση, συνεχίζει με μια παροδική σταθεροποίηση της διακινδύνευσης και με το πέρασμα του χρόνου αποκτά έντονα αυξητική τάση και περιγράφει την εξέλιξη της ανθρώπινης ζωής. Η $h_3(t)$ είναι μια σταθερή συνάρτηση διακινδύνευσης, δηλαδή η διακινδύνευση παραμένει σταθερή. Η $h_4(t)$ είναι μια αύξουσα συνάρτηση , η οποία συναντάται συχνά. Με την πάροδο του χρόνου ο ρυθμός της διακινδύνευσης αυξάνεται.



Σχήμα 2.1

Διάφορες μορφές της συνάρτησης διακινδύνευσης

Η συνάρτηση διακινδύνευσης για το μοντέλο αναλογικής διακινδύνευσης του Cox, με μεταβλητή απόκρισης t και διάνυσμα επεξηγηματικών μεταβλητών x φαίνεται παρακάτω:

$$h(t,x) = h_0(t)e^{\beta' x} \quad (2.1)$$

όπου $h_0(t)$ είναι μια βασική συνάρτηση διακινδύνευσης, η οποία εκφράζει τον κίνδυνο θανάτου, αποτυχίας ή επιτυχίας, όταν όλες οι επεξηγηματικές μεταβλητές x_j, για j=1,2,….,p, όπου p ο αριθμός των επεξηγηματικών μεταβλητών, είναι ίσες με μηδέν και $\beta = (\beta_1,\ldots,\beta_p)$ ένα διάνυσμα p συντελεστών, οι οποίοι εκφράζουν ποσοτικά την επίδραση της καθεμίας των συμμεταβλητών x.

Γενικά, η συνάρτηση διακινδύνευσης h(t,x), εξαρτάται από το χρόνο και τις συμμεταβλητές, αλλά μέσω δύο διαφορετικών παραγόντων. Ο πρώτος παράγοντας, $h_0(t)$, είναι μια συνάρτηση του χρόνου μόνο, που αφήνεται ακαθόριστη, αλλά θεωρείται η ίδια και για τα N άτομα. Ο δεύτερος παράγοντας είναι μια ποσότητα που εξαρτάται από τις συμμεταβλητές μόνο μέσω του διανύσματος **β**.

Για το λόγο αυτό, το μοντέλο αναλογικής διακινδύνευσης του Cox θεωρείται ημι-παραμετρικό (semiparametric), αφού δεν καθορίζει τη μορφή της $h_0(t)$, αλλά υποθέτει ότι οι επιδράσεις των μεταβλητών παραμένουν σταθερές στο χρόνο και είναι προσθετικές σε μια συγκεκριμένη κλίμακα.

Όπως στη γραμμική παλινδρόμηση, ένας στατιστικός στόχος της ανάλυσης επιβίωσης είναι να λάβει ένα μέτρο επιρροής, που θα περιγράφει τη σχέση ανάμεσα στην πρόγνωση για τη μεταβλητή που μας ενδιαφέρει και στο χρόνο ολοκλήρωσης του γεγονότος, μετά την προσαρμογή του μοντέλου για τις άλλες μεταβλητές που έχουμε αναγνωρίσει στην έρευνα και συμπεριλαμβάνονται στο μοντέλο. Στη γραμμική παλινδρόμηση, το μέτρο της επιρροής είναι συνήθως ο συντελεστής παλινδρόμησης. Στην ανάλυση επιβίωσης, το μέτρο της επιρροής είναι η αναλογία διακινδύνευσης (HR).

Από την εξίσωση (2.1), παρατηρούμε ότι αν θεωρήσουμε x=0, τότε προκύπτει:

$$h(t,0) = h_0(t)$$

Δηλαδή η αναφορική συνάρτηση κινδύνου μπορεί να θεωρηθεί ως η συνάρτηση κινδύνου ενός ατόμου με τιμή όλων των συμμεταβλητών ίση με 0, $x_i=0$, $i=1,\ldots,p$.

Για να δούμε πώς οι μεταβλητές είναι προσθετικές σε μια συγκεκριμένη κλίμακα, θεωρούμε για δύο οποιαδήποτε άτομα με διανύσματα μεταβλητών x_1 και x_2, την αναλογία διακινδύνευσης (HR(t)) (hazard ratio), δηλαδή το λόγο

$$HR(t) = \frac{h(t,x_1)}{h(t,x_2)} = \frac{h_0 e^{\beta' x_1}}{h_0 e^{\beta' x_2}} = e^{\beta'(x_1-x_2)} \qquad (2.2)$$

Το διάνυσμα x_i αντιστοιχεί στο διάνυσμα των p επεξηγηματικών μεταβλητών για το i-οστό άτομο που συμμετέχει στο πείραμα.

Εφόσον η τιμή των επεξηγηματικών μεταβλητών δεν εξαρτάται από το χρόνο t, η ποσότητα $e^{\beta'(x_1-x_2)}$ είναι σταθερή και για αυτό το μοντέλο του Cox είναι γνωστό ως μοντέλο αναλογικής διακινδύνευσης.

Ισοδύναμη έκφραση του μοντέλου είναι η εξής :

$$\ln\left[\frac{h(t,x)}{h_0(t)}\right] = \beta_1 x_1 + \ldots\ldots + \beta_p x_p$$

Η τελευταία μορφή απλοποιεί σημαντικά την ερμηνεία των συντελεστών. Συγκεκριμένα, το β_i ισοδυναμεί με το λογάριθμο της σχετικής διακινδύνευσης όταν έχουμε αύξηση μιας μονάδας στη μεταβλητή x_i, ενώ οι άλλες παραμένουν σταθερές. Με άλλα λόγια, η σχέση e^{β_i} εκφράζει το σχετικό ρίσκο που λαμβάνεται όταν η επεξηγηματική μεταβλητή x_i αυξάνεται κατά μια μονάδα, ενώ οι υπόλοιπες μεταβλητές παραμένουν σταθερές.

Η γενική μορφή ενός μοντέλου αναλογικής διακινδύνευσης (proportional hazards model) είναι:

$$h(t,x) = h_0(t,x)\, g(x)$$

όπου g(x) είναι μία συνάρτηση του διανύσματος x.

Στην περίπτωση του μοντέλου αναλογικής διακινδύνευσης του Cox, η g(x) είναι η συνάρτηση

$$e^{\beta' x} = e^{\beta_1 x_1 + \ldots\ldots\ldots + \beta_p x_p}$$

Αν λογαριθμήσουμε τη σχέση (2.2), προκύπτει:

$$\ln[\, h(t,x_1)] - \ln[\, h(t,x_2)] = \beta'(x_1 - x_2)$$

που δείχνει ότι το μοντέλο θεωρεί μια σταθερή διαφορά μεταξύ των λογάριθμων της διακινδύνευσης των δύο ατόμων.

2.2.2 Σωρευτική συνάρτηση διακινδύνευσης

Η σωρευτική συνάρτηση διακινδύνευσης (cumulative hazard function), συμβολίζεται με H(t) και ορίζεται ως :

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

και είναι ιδιαίτερα χρήσιμη για την επιλογή του καταλληλότερου στατιστικού μοντέλου επιβίωσης κατά την ανάλυση ενός συνόλου δεδομένων.

Η σωρευτική συνάρτηση διακινδύνευσης για το μοντέλο του Cox ορίζεται ως εξής:

$$H(t,x) = \int_0^t h_0(u)e^{x'\beta}\,du = H_0(t)\,e^{x'\beta}$$

2.2.3 Συνάρτηση αξιοπιστίας ή συνάρτηση επιβίωσης (survival function)

Η συνάρτηση αξιοπιστίας ή αλλιώς συνάρτηση επιβίωσης ορίζεται ως εξής:

$$S(t) = P(T > t) = 1 - F(t)$$

και είναι η πιθανότητα ένα άτομο να επιβιώσει για χρόνο μεγαλύτερο από χρόνο t.

Έπεται, λοιπόν, άμεσα ότι πρόκειται για μια φθίνουσα συνάρτηση για την οποία ισχύει ότι για τη χρονική στιγμή μηδέν (t=0), η πιθανότητα επιβίωσης είναι ίση με τη μονάδα (S(t)=1), ενώ για άπειρο χρόνο ($t\to\infty$) η πιθανότητα επιβίωσης είναι μηδενική (S(t)=0).

Η γραφική παράσταση της S(t) συναρτήσει του χρόνου t, ονομάζεται καμπύλη επιβίωσης (survival curve).

Η συνάρτηση αξιοπιστίας για το μοντέλο του Cox ορίζεται ως εξής:

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

Προκύπτει η παρακάτω σχέση για τη συνάρτηση επιβίωσης του χρόνου t, και από αυτή μπορεί να εκτιμηθεί η συνάρτηση επιβίωσης οποιουδήποτε ατόμου που συμμετέχει στη μελέτη.

$\int_0^t h_i(z)dz = \int_0^t h_0(z)e^{x'\beta}\,dz$ ή $H_i(t) = e^{\beta x'} H_0(t)$ ή $e^{-H_i(t)} = e^{-H_0(t)\,e^{\beta x'}}$ ή $S_i(t) = [S_0(t)]^{e^{\beta_{x'}}}$

όπου $S_0(t) = e^{[-H_0(t)]}$ είναι η αναφορική συνάρτηση επιβίωσης (baseline survival function).

Θεωρούμε τώρα, ότι έχουμε μόνο μία μεταβλητή, τη X, που αντιπροσωπεύει το είδος της θεραπείας και θεωρούμε επίσης ότι παίρνει την τιμή 1 ($x_1=1$) αν το άτομο λαμβάνει τη θεραπεία A και 0 ($x_2=0$) αν λαμβάνει τη θεραπεία B. Τότε, η συνάρτηση διακινδύνευσης για τα άτομα που ανήκουν στην πρώτη ομάδα (A),

είναι $h(t,1) = h_0(t)e^{\beta}$, ενώ για τα άτομα της ομάδας Β θα είναι $h(t,0) = h_0(t)e^{0} = h_0(t)$.

Τότε ο λόγος των συναρτήσεων διακινδύνευσης για τα δύο άτομα θα είναι:

$HR = \frac{h(t,1)}{h(t,0)} = e^{\beta}$, και επομένως $S_1(t) = [S_0(t)]^{e^{\beta}}$ (2.3)

Έτσι, αν:

- $\beta > 0$ ή $e^{\beta} > 1$ και η διακινδύνευση ενός ατόμου που λαμβάνει τη θεραπεία Α θα είναι μεγαλύτερη από τη διακινδύνευση ενός ατόμου που λαμβάνει τη θεραπεία Β ενώ η πιθανότητα επιβίωσης ενός ατόμου της ομάδας Α θα είναι μικρότερη από την πιθανότητα επιβίωσης ενός ατόμου της ομάδας Β, όπως προκύπτει από την (2.3) (αφού $S_0(t) < 1$ ή $S_1(t) = [S_0(t)]^{e^{\beta}} < S_0(t)$
- $\beta=0$ ή $h(t,1)=h(t,0)$ και $S_1(t)=S_0(t)$, δηλαδή οι δύο θεραπείες θεωρούνται ισοδύναμες.
- $\beta<0$ ή $0 < e^{\beta} < 1$ και η διακινδύνευση ενός ατόμου που λαμβάνει τη θεραπεία Α θα είναι μικρότερη από τη διακινδύνευση ενός ατόμου που λαμβάνει τη θεραπεία Β ενώ η πιθανότητα επιβίωσης ενός ατόμου της ομάδας Α θα είναι μεγαλύτερη από την πιθανότητα επιβίωσης ενός ατόμου της ομάδας Β.

2.3 Εκτίμηση των συντελεστών β του μοντέλου του Cox

2.3.1 Μέθοδος μέγιστης πιθανοφάνειας

Αφού ορίστηκε το μοντέλο αναλογικής διακινδύνευσης του Cox, αυτό που ακολουθεί είναι η εκτίμηση των συντελεστών παλινδρόμησης **β** του μοντέλου. Η μέθοδος που δόθηκε πρώτη και χρησιμοποιείται ακόμα και σήμερα, οφείλεται στον Cox και βασίζεται στη συνάρτηση μερικής πιθανοφάνειας.

Επειδή η αναφορική συνάρτηση διακινδύνευσης $h_0(t)$ δεν καθορίζεται παραμετρικά, δεν μπορεί να χρησιμοποιηθεί η συνηθισμένη συνάρτηση πιθανοφάνειας για την εκτίμηση του διανύσματος **β**. Ο σκοπός είναι να εκτιμηθεί

το **β** με βάση την πληροφορία που προκύπτει από τα παρατηρούμενα δεδομένα, χωρίς να χρειάζεται να εμπλακεί η $h_0(t)$.

Θεωρούμε ένα σύνολο N ατόμων και υποθέτουμε ότι υπάρχουν συνολικά k πλήρεις, διακεκριμένοι χρόνοι και N-k αποκομμένοι χρόνοι. Έστω $t_{(1)}$, $t_{(2)}$,...., $t_{(k)}$ οι k ταξινομημένοι πλήρεις χρόνοι και $R(t_{(j)})$ ή R_j το σύνολο των ατόμων που βρίσκονται σε κίνδυνο στο χρόνο $t_{(j)}$, δηλαδή το σύνολο των ατόμων που είναι υπό παρακολούθηση τη χρονική στιγμή $t_{(j)}$. Συμβολίζουμε με $x_{(j)} =(x_{(j)1}, x_{(j)2},...., x_{(j)p})$, $1<j< k$ το διάνυσμα των συμμεταβλητών που αντιστοιχεί στο άτομο με πλήρη χρόνο ζωής $t_{(j)}$, $1< j < k$.

Ο Cox (1975), εισηγήθηκε την ακόλουθη συνάρτηση για την εκτίμηση του **β**:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{k} \left\{ \frac{e^{\beta' x_j}}{\sum_{i \in R_j} e^{\beta' x_i}} \right\} \quad (2.4)$$

την οποία ονόμασε μερική πιθανοφάνεια (partial likelihood).

Παρόλο που η παραπάνω πιθανοφάνεια δεν είναι μια πιθανοφάνεια με τη συνηθισμένη έννοια (αφού δεν προκύπτει από την πιθανότητα κάποιου παρατηρούμενου αποτελέσματος), έχει αποδειχτεί από τον Cox ότι η $L(\boldsymbol{\beta})$ μπορεί να χρησιμοποιηθεί σαν μία συνηθισμένη συνάρτηση πιθανοφάνειας, επιτρέποντας έτσι την εκτίμηση του **β** με τις συνηθισμένες διαδικασίες. Συνεπώς η εκτιμήτρια $\hat{\boldsymbol{\beta}}$ του **β** που προκύπτει είναι αμερόληπτη, συνεπής και ασυμπτωτικά Κανονική κατανομή, ενώ το διάνυσμα score $u(\boldsymbol{\beta})$, η παρατηρούμενη πληροφορία $I(\boldsymbol{\beta})$ (information matrix), ο λόγος της πιθανοφάνειας λ (likelihood ratio) καθώς και οι έλεγχοι υποθέσεων που βασίζονται στην ποσότητα $L(\boldsymbol{\beta})$ συμπεριφέρονται ακριβώς όπως και στην περίπτωση της συνηθισμένης πιθανοφάνειας.

Οι συντελεστές παλινδρόμησης **β**, εκτιμώνται (όπως και στην περίπτωση της πιθανοφάνειας), από τις τιμές $\hat{\boldsymbol{\beta}}$ που μεγιστοποιούν τη μερική πιθανοφάνεια $L(\boldsymbol{\beta})$ ή ισοδύναμα το λογάριθμο της. Ο λογάριθμος της μερικής πιθανοφάνειας (log-partial likelihood), δίνεται από τη σχέση:

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

Η εκτιμήτρια μέγιστης πιθανοφάνειας $\hat{\boldsymbol{\beta}}$, βρίσκεται από τη λύση του συστήματος των εξισώσεων που προκύπτουν από τη σχέση:

$$U_s(\beta) = \frac{\partial l(\beta)}{\partial \beta_s} = 0, \quad s=1,2,\ldots,k$$

Έτσι προκύπτει η παρακάτω σχέση, που είναι ένα σύστημα p-εξισώσεων:

$$U_s(\beta) = \sum_{j=1}^{k} x_{js} - \sum_{j=1}^{k} \left[\frac{\sum_{i \in R_j} x_{is} e^{\beta' x_i}}{\sum_{i \in R_j} e^{\beta' x_i}} \right] = 0 \qquad (2.5)$$

Για να λυθεί το σύστημα αυτό χρειάζονται επαναληπτικές διαδικασίες, όπως η μέθοδος Newton-Raphson που χρησιμοποιείται και από διάφορα προγράμματα. Ξεκινώντας από μια αρχική λύση $\hat{\beta}^{(0)}$, ο αλγόριθμος υπολογίζει επαναληπτικά το $\hat{\beta}^{(n+1)}$ από την παρακάτω σχέση, μέχρι να συγκλίνει.

$$\hat{\beta}^{(n+1)} = \hat{\beta}^{(n+1)} + I^{-1}(\hat{\beta}^{(n)})U(\hat{\beta}^{(n)})$$

Παραγωγίζοντας τη σχέση (2.5) ως προς β_s, παίρνουμε:

$$-\frac{\partial^2 l}{\partial \beta_l \partial \beta_s} = \sum_{j=1}^{k} \sum_{i \in R_j} x_{ir} \left[x_{is} - \frac{\sum_{l \in R_j} x_{ls} e^{\beta' x_l}}{\sum_{l \in R_j} e^{\beta' x_l}} \right] \frac{e^{\beta' x_l}}{\sum_{l \in R_j} e^{\beta' x_l}} \qquad s,l=1,2,\ldots,p$$

Ο πίνακας πληροφορίας, $I(\beta)$, υπολογίζεται γενικά από τη σχέση:

$I(\beta) = -E\left[\frac{d^2 l(\beta)}{d\beta d\beta'}\right]$, αλλά στην περίπτωση μας ισχύει $-E\left[\frac{d^2 l(\beta)}{d\beta d\beta'}\right] = \frac{d^2 l(\beta)}{d\beta d\beta'}$

Η μερική πιθανοφάνεια μπορεί να χρησιμοποιηθεί όταν δεν υπάρχουν ισότιμες παρατηρήσεις (ties) στα δεδομένα μας, δηλαδή αν κάθε πλήρης χρόνος εμφανίζεται μία μόνο φορά. Αν στα δεδομένα υπάρχουν ισότιμες παρατηρήσεις, τότε για τη συνάρτηση μερικής πιθανοφάνειας χρησιμοποιούνται προσεγγίσεις που προτάθηκαν κυρίως από τον Breslow και από τον Efron.

2.3.1.1 Πιθανοφάνεια του Breslow

Συνήθως, χρησιμοποιείται η απλή προσέγγιση του Breslow

$$L(\beta) = \prod_{j=1}^{k} \left\{ \frac{e^{\beta' z_j}}{\{\sum_{m \in R_j} e^{\beta' x_m}\}^{d_j}} \right\} \quad (2.6)$$

όπου $z_j = \sum_{m=1}^{dj} x_m$ και x_m το διάνυσμα συμμεταβλητών της μονάδας m, με διακοπή τη στιγμή $t_{(j)}$, $m=1,\ldots,d_j$ και d_j το πλήθος των αποτυχιών στο χρόνο $t_{(j)}$. Η συγκεκριμένη προσέγγιση θεωρείται ακριβής, όταν η ποσότητα d_j/n_j είναι μικρή.

2.3.1.2 Πιθανοφάνεια του Efron

Η προσέγγιση του Efron δίνεται από τη σχέση

$$L(\beta) = \prod_{j=1}^{k} \frac{e^{\beta' z_j}}{\prod_{i=1}^{d_j} \sum_{m \in R_j} e^{\beta' x_m} - \frac{i-1}{d_j} \sum_{m \in R_j} e^{\beta' x_m}} \qquad e^{\beta' z_i}$$

όπου x_m το διάνυσμα των συμμεταβλητών του m-ατόμου και d_j το σύνολο των ατόμων που αποτυγχάνουν στο χρόνο t_j.

2.3.1.3 Διακριτή Πιθανοφάνεια

Όταν το d_j/n_j δεν είναι μικρό χρησιμοποιούμε μια προσέγγιση του Cox, για την οποία ο Cox έκανε την υπόθεση ότι τα δεδομένα προέρχονται από μια διακριτή κατανομή χρόνου ζωής. Η μέθοδος αυτή περιλαμβάνει την καταμέτρηση των πιθανών ατόμων που είναι σε ρίσκο R_j, σε κάθε ισόπαλο πλήρη χρόνο.

Η προσέγγιση του Cox δίνεται από την σχέση:

$$L(\beta) = \prod_{j=1}^{d_j} \frac{e^{\beta' z_j}}{\sum_{q \in Q_j} e^{\beta' z_q}}$$

όπου Q_j είναι το σύνολο όλων των υποσυνόλων του συνόλου ρίσκου R_j μεγέθους d_j. Το πλήθος των στοιχείων του συνόλου Q_j είναι $|Q_j| = \binom{n_j}{d_j}$ όπου n_j είναι το πλήθος του συνόλου R_j και με $|Q_j|$ συμβολίζουμε το πλήθος του συνόλου Q_j.

Στην περίπτωση που δεν υπάρχουν ισότιμες παρατηρήσεις στους χρόνους επιβίωσης, τότε και οι τρεις προσεγγίσεις για την εύρεση της πιθανοφάνειας είναι οι ίδιες και δίνουν το ίδιο αποτέλεσμα με τη μερική πιθανοφάνεια. Όταν υπάρχουν πολύ λίγες ισότιμες παρατηρήσεις στους χρόνους επιβίωσης, τότε οι τιμές των τριών προσεγγίσεων θα είναι πολύ κοντινές.

2.4 Έλεγχοι Υποθέσεων

Τέσσερις από τους πιο δημοφιλείς τρόπους ελέγχου υποθέσεων είναι οι παρακάτω:

2.4.1 Έλεγχοι λόγου πιθανοφάνειας (Likelihood Ratio tests)

Θεωρούμε τη μηδενική υπόθεση: $H_0 : \beta_j = \beta_0 \neq 0$ με εναλλακτική υπόθεση την: $H_1 : \beta_j = 0$

Θεωρούμε τη στατιστική συνάρτηση:

$$L(\boldsymbol{\beta}_0) = -2\ln \frac{L(\beta_0)}{L(\hat{\beta})} = 2l(\hat{\beta}) - 2l(\beta_0)$$

όπου $l(\hat{\beta})$ ο λογάριθμος πιθανοφάνειας για το εναλλακτικό μοντέλο, δηλαδή για το μοντέλο χωρίς τη συμμεταβλητή β_j και $l(\boldsymbol{\beta}_0)$ ο λογάριθμος της πιθανοφάνειας για το μοντέλο της μηδενικής υπόθεσης, δηλαδή για το μοντέλο όπου $\beta_j = \beta_0$.

Τέλος, ελέγχουμε για την $L(\boldsymbol{\beta}_0)$ αν ακολουθεί κατανομή X^2_p με βαθμό ελευθερίας p, ίσο με το σύνολο των συμμεταβλητών του προβλήματος που μελετάται.

2.4.2 Έλεγχος Wald

Θεωρούμε τη μηδενική υπόθεση: $H_0 : \beta = \beta_0$ με εναλλακτική υπόθεση την: $H_1 : \beta \neq \beta_0$

Θεωρούμε την ελεγχοσυνάρτηση Wald, η οποία ορίζεται για κάθε μεταβλητή j ως εξής:

$$W = \left\{\frac{\hat{\beta}_j - \beta_{j(0)}}{se(\hat{\beta}_j)}\right\}^2$$ και την συγκρίνουμε με την κατανομή X^2_1.

Συνήθως εξετάζουμε την H_0: $\beta = 0$.

2.4.3 Έλεγχοι Score

Θεωρούμε τη μηδενική υπόθεση: $H_0 : \beta = \beta_0$. Θεωρούμε τη στατιστική συνάρτηση:

$S = \frac{U'(\beta_0)U(\beta_0)}{I(\beta_0)} = \frac{U^2(\beta_0)}{I(\beta_0)}$ όπου $U(\beta_0) = \frac{\partial \ln L(\beta_0)}{\partial \beta_0}$ και η παρατηρούμενη πληροφορία δίνεται από την σχέση $I(\beta_0) = -\frac{\partial^2 \ln L(\beta_0)}{\partial \beta_0^2}$.

Το score test ακολουθεί προσεγγιστικά X_1^2 κατανομή με 1 βαθμό ελευθερίας,

όταν η H_0 είναι αληθής.

Οι τρεις αυτοί έλεγχοι είναι ασυμπτωτικά ισοδύναμοι, αλλά σε μικρά δείγματα μπορεί να διαφέρουν. Όταν διαφέρουν, ο έλεγχος λόγου πιθανοφάνειας θεωρείται ο πιο αξιόπιστος, ενώ ο έλεγχος του Wald θεωρείται ο λιγότερο αξιόπιστος έλεγχος. Για μικρότερες τιμές του $\hat{\boldsymbol{\beta}}$, οι 3 έλεγχοι είναι σχεδόν οι ίδιοι, ενώ για μεγαλύτερες τιμές του $\hat{\boldsymbol{\beta}}$ η διαφορά μεταξύ των ελέγχων μεγαλώνει.

2.4.4 Διάστημα εμπιστοσύνης

Τα διαστήματα εμπιστοσύνης δημιουργούνται συνήθως βάση του στατιστικού Wald.

Το διάστημα εμπιστοσύνης του $e^{\hat{\beta}}$ είναι το:

$$\left(e^{(\hat{\beta}-z_{a/2}se(\hat{\beta}))}, e^{(\hat{\beta}+z_{a/2}se(\hat{\beta}))}\right)$$

2.5 Σύγκριση δύο κατανομών επιβίωσης

Έστω ότι έχουμε κάποια δεδομένα επιβίωσης, στα οποία προσαρμόζουμε το μοντέλο αναλογικής διακινδύνευσης του Cox, ως προς μία δείκτρια μεταβλητή X, η οποία παίρνει τις τιμές 1 και 0 :

$$X = \begin{cases} 1, \text{ αν το άτομο ανήκει στην κατηγορία 1, με κατανομή επιβίωσης } S_1(t) \\ 0, \text{ αν το άτομο ανήκει στην κατηγορία 2, με κατανομή επιβίωσης } S_2(t) \end{cases}$$

Συνεπώς το μοντέλο του Cox γίνεται :

$$h(t,x) = h_0(t)e^{\beta x} \text{ και } h(t,1) = h_0(t)e^{\beta}$$

Επειδή η αναφορική συνάρτηση $h_0(t)$ προκύπτει από τη συνάρτηση κινδύνου για όλες τις συμμεταβλητές ίσες με 0, προκύπτει ότι $h_0(t) = h_2(t)$ ή $S_0(t) = S_2(t)$ και έχουμε : $S_1(t) = [S_2(t)]^{e^{\beta}}$. Συνεπώς, ο έλεγχος της υπόθεσης $H_0 : S_1(t) = S_2(t)$, είναι ισοδύναμος με τον

έλεγχο H_0 : β=0 (αφού από την παραπάνω σχέση προκύπτει ότι $S_1(t) = S_2(t)$ για β=0). Για τον έλεγχο της παραπάνω υπόθεσης μπορούμε να χρησιμοποιήσουμε έναν από τους ελέγχους που αναφέρθηκαν προηγουμένως.

Αν έχουμε μόνο μία μεταβλητή που παίρνει τις τιμές 0 και 1, η συνάρτηση score θα μας δώσει έναν απλό έλεγχο, αφού θα είναι 1x1, δηλαδή θα είναι ένας αριθμός, όπως η παρατηρούμενη πληροφορία.

Επομένως έχουμε ότι p = 1 και έστω ότι x_i, i =1,2,…,N είναι η τιμή της μεταβλητής X που αντιστοιχεί στο γεγονός με πλήρη ή αποκομμένο χρόνο t_i , ενώ $x_{(j)}$, j = 1,2,…,k είναι η τιμή της μεταβλητής X που αντιστοιχεί στο γεγονός με πλήρη χρόνο $t_{(j)}$ όπου $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$ οι πλήρεις χρόνοι. Το διάνυσμα score γίνεται :

$$U(\beta) = \sum_{j=1}^{k} x_j - \sum_{j=1}^{k} \left[\frac{\sum_{i \in R_j} x_i e^{\beta' x_i}}{\sum_{i \in R_j} e^{\beta' x_i}} \right] \quad (2.7).$$

Στη συνέχεια ταξινομούμε τους χρόνους κατά αύξουσα και μετά υπολογίζουμε τις ποσότητες d_{1j}, d_j , N_{1j} και N_{2j}.

Ορίζουμε

d_{1j} = I(t_j είναι ένας πλήρης χρόνος της ομάδας 1 με συνάρτηση επιβίωσης $S_1(t)$)

d_j = I(t_j είναι ένας πλήρης χρόνος)

N_{1j} : το πλήθος των ατόμων της ομάδας 1 με συνάρτηση επιβίωσης $S_1(t)$ που είναι σε κίνδυνο στον πλήρη χρόνο t_j

N_{2j} : το πλήθος των ατόμων της ομάδας 2 με συνάρτηση επιβίωσης $S_2(t)$ που είναι σε κίνδυνο στον πλήρη χρόνο t_j

Επειδή x_i = 0 ή 1, ο αριθμητής στη σχέση (2.7) θα γίνει $e^{\beta} + e^{\beta} + \ldots + e^{\beta}$, n_{1t} – φορές. Όταν το x_i = 0, δηλαδή όταν το άτομο ανήκει στη δεύτερη ομάδα με συνάρτηση επιβίωσης $S_2(t)$, τότε μηδενίζεται ο όρος $x_i e^{\beta x_i}$ και δε συμβάλλει στο άθροισμα του αριθμητή. Έτσι, ο παρονομαστής της σχέσης (2.7) μπορεί να γραφεί ως το άθροισμα των όρων του αριθμητή συν το πλήθος των ατόμων της δεύτερης ομάδας που είναι σε κίνδυνο στο χρόνο t_j, αφού στον παρονομαστή έχουμε το άθροισμα των όρων $e^{\beta x_i}$ αντί $x_i e^{\beta x_i}$ και όταν θα είναι x_i =0, θα έχουμε $e^{\beta x_i} = 1$. Βάση των συμβολισμών, το διάνυσμα

score στην περίπτωση που έχουμε μία μόνο μεταβλητή με τιμές 0 και 1, η σχέση (2.7) γίνεται :

$$U(\beta) = \sum_{j=1}^{N} d_{1j} - \frac{d_{1j} N_{1j} e^{\beta}}{N_{2j}+N_{1j}e^{\beta}}$$

Η παρατηρούμενη πληροφορία, είναι στην περίπτωση αυτή, η ποσότητα $-\frac{d^2 l(\beta)}{d\beta^2}$ και έτσι έχουμε

$$I(\beta) = -\frac{d^2 l(\beta)}{d\beta^2} = \sum_{j=1}^{k} \frac{\sum_{i\in R_j} e^{\beta x_i} \sum_{i\in R_j} x_i^2 e^{\beta x_i} - (\sum_{i\in R_j} x_i e^{\beta x_i})^2}{(\sum_{i\in R_j} e^{\beta x_i})^2} =$$

$$\sum_{j=1}^{k} \frac{d_j(N_{1j}e^{\beta} + N_{2j}) N_{1j}e^{\beta} - (N_{1j}e^{\beta})^2}{(N_{1j}e^{\beta} + N_{2j})^2} = \sum_{j=1}^{k} \frac{d_j[(N_{1j}e^{\beta})^2 + N_{1j} N_{2j}e^{\beta})^2 - (N_{1j}e^{\beta})^2]}{(N_{1j}e^{\beta} + N_{2j})^2} \quad \text{ή}$$

$$I(\beta) = \sum_{j=1}^{k} \frac{d_j N_{1j} N_{2j} e^{\beta}}{(N_{1j}e^{\beta} + N_{2j})^2}$$

Για τον έλεγχο της $H_0 : \beta = 0$, έχουμε ότι $Q = \frac{U(0)^2}{I(0)}$ και επομένως έχουμε $U(0) = \sum_{j=1}^{k} d_{1j} - \frac{d_j N_{1j}}{N_{1j} + N_{2j}}$ και $I(0) = \sum_{j=1}^{k} \frac{d_j N_{1j} N_{2j}}{(N_{1j} + N_{2j})^2}$.

Η Q ακολουθεί προσεγγιστικά την $X^2_{(1)}$ κατανομή και έτσι βρίσκουμε τη p-τιμή του ελέγχου και αν η τιμή αυτή είναι μικρή, τότε η μηδενική υπόθεση της ισότητας των δύο κατανομών απορρίπτεται.

2.6 Έλεγχοι της υπόθεσης αναλογικής διακινδύνευσης στο μοντέλο του Cox

2.6.1 Γενικά

Για να έχουν νόημα τα αποτελέσματα που προκύπτουν από την προσαρμογή του μοντέλου αναλογικής διακινδύνευσης του Cox, πρέπει πριν από αυτό, να ελέγχουμε αν η υπόθεση της αναλογικότητας των κινδύνων ισχύει, δηλαδή αν ο λόγος

$$\frac{h_l(t)}{h_j(t)} = \frac{h_0(t)e^{\beta' x_l}}{h_0(t)e^{\beta' x_j}} = e^{\beta'(x_l - x_j)}$$

μεταξύ των συναρτήσεων διακινδύνευσης δύο μονάδων *l* και j, είναι ανεξάρτητος του χρόνου.

Ο έλεγχος αυτός μπορεί να γίνει είτε γραφικά (Hess, 1995), είτε με διάφορες ελεγχοσυναρτήσεις που υπάρχουν για τον έλεγχο αυτό (παρόλο που οι ελεγχοσυναρτήσεις δεν χρησιμοποιούνται ευρέως). Το δεύτερο βήμα στην προσαρμογή του μοντέλου είναι, στην περίπτωση που υπάρχουν πολλές υποψήφιες μεταβλητές, να ορίσουμε ένα βασικό πλάνο για την επιλογή των μεταβλητών που θα συμπεριληφθούν στο μοντέλο. Αφού βρούμε τις κατάλληλες μεταβλητές για το μοντέλο και προσαρμόσουμε στα δεδομένα αυτών το μοντέλο παλινδρόμησης του Cox, πρέπει στη συνέχεια (3ο βήμα) να εξετάσουμε κατά πόσο το μοντέλο είναι ικανοποιητικό ή αν θέλει βελτίωση.
Στην περίπτωση που βρούμε ότι η υπόθεση της αναλογικής διακινδύνευσης δεν ισχύει, τότε είτε κάνουμε μετασχηματισμούς των δεδομένων έτσι ώστε να ικανοποιείται η υπόθεση της αναλογικότητας ή θα πρέπει να επιλέξουμε μία εναλλακτική κλάση μοντέλων που να είναι πιο κατάλληλη για τα δεδομένα μας.
Ένας τρόπος για τον έλεγχο της υπόθεσης της αναλογικής διακινδύνευσης είναι η πρόσθεση μιας χρονικά εξαρτώμενης μεταβλητής στο μοντέλο.

Θεωρούμε μία μελέτη επιβίωσης στην οποία κάθε μπαταρία έχει τοποθετηθεί σε μία από δύο ομάδες, ανάλογα με το τι μάρκα είναι η Α ή η Β. Μας ενδιαφέρει αν ο λόγος διακινδύνευσης στον χρόνο t της μιας ομάδας σε σχέση με την άλλη, είναι ανεξάρτητος του χρόνου επιβίωσης.

Το μοντέλο του Cox για την συγκεκριμένη μελέτη για την i μπαταρία είναι:

$$h_i(t) = \exp(\beta_1 x_{1i})\, h_0(t),$$

όπου x_{1i} είναι η τιμή της μεταβλητής x_1, η οποία είναι 0 για την ομάδα Α και 1 για την ομάδα Β. Ο λόγος διακινδύνευσης μιας μπαταρίας που ανήκει στην ομάδα Β, σε σχέση με έναν που ανήκει στην ομάδα Α είναι e^{β_1}, ο οποίος είναι ανεξάρτητος του χρόνου επιβίωσης.

Τώρα ορίζουμε μια χρονικά εξαρτημένη επεξηγηματική μεταβλητή x_2, όπου $x_2 = x_1 t$. Αν αυτή η μεταβλητή προστεθεί στο μοντέλο, η συνάρτηση διακινδύνευσης γίνεται

$$h_i(t) = \exp(\beta_1 x_{1i} + \beta_2 x_{2i}) h_0(t)$$

όπου $x_2 = x_1 t$. Ο νέος λόγος διακινδύνευσης για χρόνο t είναι $\exp(\beta_1 + \beta_2 t)$, όπου $x_2 = t$, όταν η μπαταρία ανήκει στην ομάδα Β και 0 διαφορετικά. Παρατηρούμε ότι ο λόγος διακινδύνευσης εξαρτάται από το χρόνο και το παραπάνω μοντέλο δεν είναι πια μοντέλο αναλογικής διακινδύνευσης. Συγκεκριμένα, αν $\beta_2<0$ ο λόγος διακινδύνευσης μειώνεται με το χρόνο. Αυτό σημαίνει ότι η συνάρτηση διακινδύνευσης της ομάδας Β σε σχέση με την Α μειώνεται με το χρόνο. Αν $\beta_1<0$ θα μπορούσαμε να πούμε ότι η ομάδα Β γίνεται καλύτερη, όσο περνάει ο χρόνος. Από την άλλη, αν $\beta_2>0$ ο λόγος διακινδύνευσης της ομάδας Β αυξάνει με το χρόνο, αντανακλώντας ένα αυξημένο σύνολο ολοκλήρωσης των γεγονότων της ομάδας Β σε σχέση με την Α. Τέλος, αν $\beta_2=0$, ο λόγος διακινδύνευσης είναι e^{β_1}, δηλαδή σταθερός.

2.6.2 Γραφικές μέθοδοι για τον έλεγχο της αναλογικότητας των κινδύνων

Όπως αναφέραμε και παραπάνω, η συνάρτηση επιβίωσης, για το μοντέλο του Cox, ενός ατόμου με διάνυσμα συμμεταβλητών $x=(x_1,x_2,\ldots,x_p)$ ορίζεται ως: $S(t,x) = S_0(t)^{e^{\beta' x}}$

Λογαριθμίζοντας την παραπάνω εξίσωση προκύπτει η σχέση

$$- \ln(S(t,x)) = e^{\beta' x}[- \ln(S_0(t))]$$

και λογαριθμίζοντας ξανά την παραπάνω σχέση προκύπτει η ακόλουθη

$$\ln[- \ln(S(t,x))] = \beta' x + \ln[- \ln(S_0(t))]$$

Θεωρούμε x_1 και x_2 τα διανύσματα των συμμεταβλητών δύο ατόμων. Κάτω από την υπόθεση της αναλογικότητας (Καρώνη, 2009), οι συναρτήσεις $\ln[- \ln(S(t,x_1))]$ και $\ln[- \ln(S(t,x_2))]$ θα έχουν μια σταθερή απόσταση, $\beta' x_1$ και $\beta' x_2$ αντίστοιχα, από τον αναφορικό αθροιστικό κίνδυνο $\ln[- \ln(S_0(t))]$. Για τις δύο αυτές συναρτήσεις ισχύει $\ln[- \ln(S(t,x_1))] = \ln[- \ln(S(t,x_2))] + \beta'(x_1 - x_2)$. Επομένως, αν σχεδιάσουμε τις γραφικές παραστάσεις των $\ln[- \ln(S(t,x))]$ συναρτήσει του χρόνου, οι δύο

καμπύλες που θα προκύψουν θα είναι παράλληλες και θα απέχουν μεταξύ τους απόσταση σταθερή και ίση με $\beta'(x_1 - x_2)$. Έτσι, ένας πρώτος έλεγχος που θα μπορούσε να γίνει για τον έλεγχο της αναλογικότητας της διακινδύνευσης, είναι η γραφική παράσταση αυτή. Αν οι καμπύλες που θα προκύψουν είναι παράλληλες ή σχεδόν παράλληλες, τότε θεωρούμε ότι ισχύει η υπόθεση της αναλογικότητας.

Οι γραφικές παραστάσεις που δημιουργούμε, βασίζονται στην εκτίμηση των ποσοτήτων S(t,x) με μεθόδους που δε χρησιμοποιούν την υπόθεση της αναλογικότητας των κινδύνων. Τέτοιες μέθοδοι είναι η εκτίμηση της συνάρτησης επιβίωσης με τη μέθοδο των Kaplan-Meier (την οποία θα ορίσουμε παρακάτω). Ωστόσο η εκτιμήτρια Kaplan-Meier δεν λαμβάνει υπόψη τις τιμές των άλλων συμμεταβλητών που περιλαμβάνονται στην ανάλυση, πέρα από αυτήν που καθορίζει τις ομάδες. Επομένως, αναζητούμε μια πιο βελτιωμένη μορφή αυτού του γραφικού ελέγχου. Γι' αυτό το λόγο θα χρησιμοποιήσουμε την εκτιμήτρια Breslow στην θέση της, δηλαδή την

$$\hat{S}_0(t) = e^{-\hat{H}_0(t)} \qquad \text{με} \qquad H_0(t) = \sum_{t_{(j)} \le t} \frac{d_j}{\sum_{i \in R_j} e^{\hat{\beta}' x_i}}$$

λαμβάνοντας υπόψη όλες τις σημαντικές συμμεταβλητές ενός στρωματοποιημένου μοντέλου του Cox. Όταν χρειάζεται να μελετηθεί η επίδραση ενός επιπέδου μιας κατηγορικής μεταβλητής Z, σε σχέση με άλλες μεταβλητές χωρίς να ενδιαφέρει η επίδραση της Z στο αποτέλεσμα, τότε χρησιμοποιείται μια επέκταση του μοντέλου αναλογικού κινδύνου του Cox. Επίσης, όταν μια μεταβλητή έχει επίπεδα που δημιουργούν συναρτήσεις κινδύνου οι οποίες δεν ικανοποιούν την υπόθεση της αναλογικότητας, τότε στρωματοποιούμε ως προς τη μεταβλητή αυτή. Το μοντέλο που προκύπτει ονομάζεται στρωματοποιημένο μοντέλο του Cox (stratified Cox model) και εφαρμόζεται αφού θεωρήσουμε τη στρωματοποίηση των δεδομένων της Z σε υποομάδες, κάθε μία από τις οποίες χαρακτηρίζεται από ένα επίπεδο του παράγοντα. Το στρωματοποιημένο μοντέλο επιτρέπει στη μορφή της συνάρτησης κινδύνου να αλλάζει ανάμεσα στα επίπεδα της στρωματοποιημένης μεταβλητής. Η μεταβλητή Z μπορεί εκτός από κατηγορική να είναι το αποτέλεσμα χωρισμού

μίας ποσοτικής μεταβλητής σε ομάδες. Έτσι ένας γραφικός έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης, προκύπτει αντικαθιστώντας την παραπάνω εκτιμήτρια Kaplan-Meier με την κατά στρώματα εκτιμήτρια της συνάρτησης επιβίωσης.

2.6.2.1 Εκτιμήτρια Kaplan-Meier

Όταν οι εφαρμογές εμπεριέχουν από δεξιά αποκομμένες παρατηρήσεις χρησιμοποιείται η εκτιμήτρια Kaplan-Meier της συνάρτησης επιβίωσης, η οποία έχει αποκτήσει πολύ μεγάλη χρηστική αξία. Έστω τυχαίο δείγμα N μονάδων, μερικές εκ των οποίων καταστρέφονται κατά τις διακεκριμένες χρονικές στιγμές $t_{(1)}<t_{(2)}< \ldots <t_{(k)}$, $k \leq N$. Έστω ότι κατά τη χρονική στιγμή $t_{(j)}$ καταστρέφονται d_j μονάδες, ενώ αμέσως πριν από τη στιγμή αυτή λειτουργούσαν N_j μονάδες. Ο αριθμός N_j περιλαμβάνει όλες τις μονάδες που γνωρίζουμε ότι λειτουργούν εκείνη τη στιγμή, δηλαδή αυτές με χρόνο λειτουργίας $t \geq t_{(j)}$, ανεξάρτητα από το αν στη συνέχεια θα διακοπεί η λειτουργία τους ή θα συνεχίζουν να λειτουργούν μετά το πέρας του πειράματος (αποκομμένες τιμές). Δεν περιλαμβάνει τις μονάδες που έχουν ήδη καταστραφεί, ούτε τις μονάδες με αποκομμένες τιμές πριν τη στιγμή $t_{(j)}$.

Η εκτιμήτρια Kaplan-Meier ορίζεται ως εξής:

$$\hat{S}(t) = \frac{N_1 - d_1}{N_1} \; \frac{N_2 - d_2}{N_2} \cdots\cdots\cdots \frac{N_i - d_i}{N_i}$$

όπου n_i ο αριθμός των μονάδων που ήταν σε λειτουργία ακριβώς πριν από τη χρονική στιγμή t_i και $i : t_i \leq t < t_{i+1}$.

Επομένως

$$\hat{S}(t) = \begin{cases} \prod_{j:\, t_{(j)} \leq t} \frac{N_i - d_i}{N_i}, \text{ όταν } t \geq t_{(1)} \\ 1, \quad \text{όταν } t \leq t_{(1)} \end{cases} \qquad (2.8)$$

και

$$se(\hat{S}(t)) = \hat{S}(t) \left\{ \sum_{t_{(j)} \leq t} \frac{d_j}{N_j(N_j - d_j)} \right\}^{1/2}$$

ΚΕΦΑΛΑΙΟ 3

ΥΠΟΛΟΙΠΑ ΣΤΟ ΜΟΝΤΕΛΟ ΤΟΥ COX

3.1 Γενικά

Τα υπόλοιπα χρησιμοποιούνται για τον έλεγχο διαφορετικών θεμάτων που αφορούν την καταλληλότητα του μοντέλου του Cox, όπως για τον έλεγχο της υπόθεσης αναλογικότητας της διακινδύνευσης, για τον έλεγχο της καταλληλότητας του μοντέλου καθώς και για την εύρεση άτυπων σημείων. Επίσης χρησιμοποιούνται για την εύρεση της συναρτησιακής μορφής μιας ερμηνευτικής μεταβλητής που θα εισαχθεί στο μοντέλο όταν υπάρχουν ήδη άλλες μεταβλητές (Collett, 2003).

Στα γραμμικά μοντέλα τα σφάλματα κατανέμονται κανονικά και αν το μοντέλο που προσαρμόζουμε στα δεδομένα είναι σωστό, τα υπόλοιπα θα έχουν προσεγγιστικά Κανονική κατανομή. Υπάρχουν αρκετές ελεγχοσυναρτήσεις καθώς και διαγνωστικά γραφήματα όπως το QQ plot για να εξακριβωθεί η Κανονικότητα. Το ανάλογο του μοντέλου του Cox είναι η σύγκριση του υπολοίπου του Cox-Snell με μία Εκθετική κατανομή με παράμετρο 1. Αυτό συμβαίνει, αφού αν η τυχαία μεταβλητή T που εκφράζει το χρόνο επιβίωσης έχει συνεχή κατανομή S(t), τότε η S(T) ακολουθεί Ομοιόμορφη κατανομή στο (0,1) και η H(T) ακολουθεί Εκθετική κατανομή με παράμετρο 1.

Τα υπόλοιπα υπολογίζονται για κάθε άτομο της έρευνας και δείχνουν κατά πόσο τα δεδομένα συμφωνούν με τις προϋποθέσεις και τις προβλέψεις του μοντέλου, συνολικά και μεμονωμένα. Τα πιο γνωστά υπόλοιπα στο μοντέλο του Cox είναι τα υπόλοιπα των Cox-Snell (Cox-Snell residuals), τα τροποποιημένα Cox-Snell υπόλοιπα (Modified Cox-Snell residuals), τα υπόλοιπα Schoenfeld (Schoenfeld residuals), τα υπόλοιπα martingale (Martingale residuals), τα υπόλοιπα Score και τα υπόλοιπα απόκλισης (Deviance residuals). Κάθε ένα από αυτά χρησιμοποιείται για να ελέγξει κάποιο από τα θέματα που προαναφέρθηκαν.

3.2 Υπόλοιπα Cox-Snell (Cox-Snell residuals)

Τα υπόλοιπα Cox-Snell (Cox-Snell, 1968), τα οποία είναι αυτά που χρησιμοποιούνται περισσότερο στην ανάλυση επιβίωσης, είναι ουσιαστικά οι εκτιμώμενες τιμές της αθροιστικής συνάρτησης κινδύνου για την i-οστή παρατήρηση, στον αντίστοιχο χρόνο t_i.

Τα υπόλοιπα Cox-Snell δίνονται από τον τύπο

$$r_{ci} = \hat{H}_0(t_i) \exp(\hat{\beta}' x_i) = \hat{H}(t_i \mid x_i), \quad i=1,2,\ldots,N$$

όπου $\hat{H}_0(t_i)$ είναι η εκτιμώμενη αναφορική αθροιστική συνάρτηση κινδύνου η οποία, στην περίπτωση του μοντέλου του Cox, υπολογίζεται από τη σχέση $\hat{H}_0(t_i) = -\log \hat{S}_0(t) = \sum_{i:t_{(i)}<t} \frac{d_i}{\sum_{j\in R_i} \exp[\hat{\beta}' x_j]}$ η οποία είναι η εκτιμήτρια Breslow (σχέση 2.6)**,** όπου, $x_i = (x_{1i}, x_{2i}, \ldots, x_{pi})'$ είναι το διάνυσμα των συμμεταβλητών του i ατόμου, $\hat{\beta}' = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ είναι οι εκτιμώμενοι συντελεστές παλινδρόμησης, N είναι το πλήθος των παρατηρήσεων και t_i είναι ο πλήρης ή ο αποκομμένος χρόνος του ατόμου i.

Αν T είναι η τυχαία μεταβλητή που σχετίζεται με το χρόνο επιβίωσης του ατόμου και S(t) είναι η αντίστοιχη συνάρτηση επιβίωσης, τότε η τυχαία μεταβλητή Y=-logS(t) ακολουθεί την Εκθετική κατανομή με παράμετρο 1, ανεξάρτητα από τον τύπο της S(t). Αν το μοντέλο είναι κατάλληλο για την περιγραφή της τ.μ. T, τα υπόλοιπα αυτά θα πρέπει να κατανέμονται προσεγγιστικά στην Εκθετική κατανομή με παράμετρο 1. Για τον έλεγχο της καταλληλότητας ενός μοντέλου αναλογικής διακινδύνευσης, αρκεί να ελεγχθεί αν τα υπόλοιπα Y κατανέμονται προσεγγιστικά από μια Εκθετική κατανομή με παράμετρο λ=1. Για την Εκθετική κατανομή με παράμετρο λ έχουμε ότι

$$H(y) = -\log S(y) = -\log(e^{-\lambda y}) = \lambda y, \qquad y \geq 0$$

και

$$\log H(y) = \log[-\log S(y)] = \log\lambda + \log y, \quad y \geq 0.$$

Έτσι μία γραφική παράσταση της ποσότητας $\log[-\log(S(r_{ci}))] = \log[\hat{H}(r_{ci})]$ έναντι της ποσότητας $\log(r_{ci})$ θα πρέπει να προσεγγίζει τη γραφική παράσταση μιας ευθείας που περνά από την αρχή των αξόνων με κλίση μονάδα (ευθεία y = x). Εναλλακτικά θα μπορούσαμε να κάνουμε γραφική παράσταση της ποσότητας

$\hat{H}(r_{ci})$ έναντι της ποσότητας r_{ci} και να τη συγκρίνουμε πάλι με την ευθεία y = x. Για τον υπολογισμό της ποσότητας $\hat{H}(r_{ci})$ μπορεί να χρησιμοποιηθεί είτε η Kaplan-Meier εκτιμήτρια της αθροιστικής συνάρτησης διακινδύνευσης (σχέση 2.8).

Τα Cox-Snell υπόλοιπα γενικώς δεν προτείνονται για το μοντέλο αναλογικής διακινδύνευσης του Cox, αλλά χρησιμοποιούνται κυρίως για τα παραμετρικά μοντέλα επιβίωσης όπου η βασική συνάρτηση διακινδύνευσης $h_0(t)$ καθορίζεται ή μοντελοποιείται. Έχουν ιδιότητες που είναι αρκετά διαφορετικές από αυτές που έχουν τα υπόλοιπα που χρησιμοποιούνται στην γραμμική παλινδρόμηση. Συγκεκριμένα, δεν είναι συμμετρικά κατανεμημένα γύρω από το μηδέν και δεν μπορούν να είναι αρνητικά (παίρνουν τιμές στο διάστημα $[0, \infty]$). Επιπλέον, από την στιγμή που τα υπόλοιπα υποτίθενται ότι ακολουθούν την Εκθετική κατανομή, όταν έχει προσαρμοστεί ένα κατάλληλο μοντέλο, έχουν μια πολύ λοξή κατανομή και η μέση τιμή και η διασπορά για κάθε υπόλοιπο θα είναι και οι δύο μονάδες.

3.3 Τροποποιημένα Cox-Snell υπόλοιπα (Modified Cox-Snell residuals)

Τα Cox-Snell υπόλοιπα που αντιστοιχούν σε αποκομμένους χρόνους δε λαμβάνονται υπόψη στις γραφικές παραστάσεις για τον έλεγχο της καταλληλότητας του παραμετρικού μοντέλου διάρκειας ζωής. Γι' αυτόν το λόγο (για να λαμβάνουν υπόψη τους και τους αποκομμένους χρόνους) τροποποιούμε τα Cox-Snell υπόλοιπα, προσθέτοντας μία θετική σταθερά Δ.

Επομένως έχουμε :

$$r'_{ci} = \begin{cases} r_{ci}, & \text{για πλήρεις χρόνους} \\ r_{ci} + \Delta, & \text{για αποκομμένους χρόνους.} \end{cases}$$

Για τον καθορισμό της ποσότητας Δ (>0) που είναι τυχαία μεταβλητή που δηλώνει τον υπολειπόμενο χρόνο ζωής του ατόμου i (για το οποίο γνωρίζουμε ότι έχει ζήσει μέχρι τη χρονική στιγμή r_{ci}) χρησιμοποιούμε την ιδιότητα έλλειψης μνήμης της Εκθετικής κατανομής που οδηγεί στο συμπέρασμα ότι η ποσότητα Δ_i ακολουθεί την Εκθετική κατανομή με παράμετρο λ=1.

Ως εκ τούτου, η μέση τιμή της Δ είναι μονάδα (E(Δ)=1) και χρησιμοποιώντας την ως τιμή της Δ, τα τροποποιημένα υπόλοιπα Cox-Snell υπολογίζονται από τη παρακάτω σχέση :

$$r'_{ci} = \begin{cases} r_{ci}, & \text{για πλήρεις χρόνους} \\ r_{ci}+1, & \text{για αποκομμένους χρόνους.} \end{cases}$$

Εισάγοντας ένα δείκτη δ_i, ο οποίος είναι μηδέν ($\delta_i = 0$), αν ο χρόνος είναι αποκομμένος και μονάδα διαφορετικά, τα τροποποιημένα υπόλοιπα Cox-Snell υπολογίζονται από τη σχέση:

$$r'_{ci} = 1-\delta_i + r_{ci}$$

Τα τροποποιημένα Cox-Snell υπόλοιπα που αντιστοιχούν σε αποκομμένους χρόνους είναι μεγαλύτερα του 1, ενώ τα τροποποιημένα Cox-Snell υπόλοιπα που αντιστοιχούν σε πλήρεις χρόνους παίρνουν μη αρνητικές τιμές (δηλαδή, γενικά τα τροποποιημένα Cox-Snell υπόλοιπα παίρνουν τιμές στο διάστημα $[0, \infty)$).

3.4 Υπόλοιπα Schoenfeld (Schoenfeld residuals ή Partial residuals)

Τα υπόλοιπα Schoenfeld, τα οποία πρότεινε ο Schoenfeld το 1982 για το μοντέλο του Cox, πλεονεκτούν έναντι των άλλων υπολοίπων γιατί υπολογίζουν ένα ξεχωριστό υπόλοιπο για κάθε άτομο για κάθε μεταβλητή, δηλαδή για p μεταβλητές θα έχουμε p Schoenfeld υπόλοιπα για κάθε άτομο και επίσης δε χρειάζεται η εκτίμηση της αθροιστικής συνάρτησης διακινδύνευσης $H(t_i \mid x_i)$.
Με το γράφημα των υπολοίπων Schoenfeld συναρτήσει του χρόνου μπορούμε να ελέγξουμε την υπόθεση της αναλογικότητας της διακινδύνευσης (PH υπόθεση). Αν το γράφημα έχει μια τυχαία μορφή των υπολοίπων έναντι του χρόνου τότε ικανοποιείται η PH υπόθεση. Αντίθετα, σημαίνει ότι δεν ικανοποιείται η υπόθεση.
Το υπόλοιπο Schoenfeld για τη μεταβλητή x_j δίνεται από τον τύπο :

$$r_{Pji} = x_{ji}-\hat{a}_{ji} \qquad (3.1)$$

Παρόλα αυτά, μπορούν να δοθούν και από την σχέση $r_{Pji} = \delta_i\{ x_{ji}-\hat{a}_{ji}\}$ η οποία χρησιμεύει στον ορισμό των Score υπολοίπων (Παράγραφος 3.7), όπου x_{ji} είναι η

τιμή της συμμεταβλητής x_j για το i άτομο (δηλαδή η τιμή x_{ji}), $\hat{a}_{ji}=\frac{\sum_{i\in R_i} x_{j'}\exp(\hat{\beta}' x_j)}{\sum_{i\in R_i}\exp(\hat{\beta}' x_j)}$ i=1,…,N, j=1,…,p (3.2) είναι η αναμενόμενη τιμή της συμμεταβλητής για τα άτομα που βρίσκονται σε ρίσκο στο χρόνο t_i, N το πλήθος των ατόμων και p το πλήθος των μεταβλητών και $\delta_i=0$, αν t_i αποκομμένος και $\delta_i=1$, αν t_i πλήρης χρόνος.

Για αποκομμένες παρατηρήσεις (δηλαδή $\delta_i=0$) βλέπουμε ότι τα υπόλοιπα Schoenfeld είναι 0 για όλες τις μεταβλητές. Επιπλέον, αν η μεγαλύτερη παρατήρηση σε ένα δείγμα είναι μη αποκομμένη, η τιμή της $\hat{a}_{ji}$ γι' αυτήν την παρατήρηση, θα είναι ίση με την x_{ji} και έτσι $r_{Pji}=0$.

Τα υπόλοιπα Schoenfeld είναι ιδιαίτερα χρήσιμα στον έλεγχο της υπόθεσης της αναλογικής διακινδύνευσης του μοντέλου του Cox. Οι Grambsch και Therneau (1994) πρότειναν μία τροποποιημένη έκδοση των υπολοίπων Schoenfeld. Έστω ότι το διάνυσμα $r_{Pi} = (r_{P1i},r_{P2i},\ldots,r_{Ppi})'$ παριστάνει τα υπόλοιπα Schoenfeld του i-οστού ατόμου, τότε τα τυποποιημένα υπόλοιπα Schoenfeld (scaled Schoenfeld) συμβολίζονται με r^*_{Pji} και είναι τα στοιχεία του διανύσματος $r^*_{Pj} = r\, V^{-1}(r_{Pj})\, r_{Pj}$, όπου r είναι ο αριθμός των θανάτων ανάμεσα στα N άτομα και $V^{-1}(r_{Pj}) \cong rV(\hat{\beta})$, με $V(\hat{\beta})$ να είναι ο πίνακας διασποράς των εκτιμήσεων των παραμέτρων στο προσαρμοσμένο μοντέλο αναλογικής διακινδύνευσης του Cox. Αυτά τα υπόλοιπα είναι αρκετά απλά να υπολογισθούν.

Οι Grambsch και Therneau(1994) έδειξαν ότι η μέση τιμή του i-οστού scaled Schoenfeld υπολοίπου, για την j-οστή επεξηγηματική μεταβλητή x_j, δίνεται από την σχέση :

$$E(r^*_{Pji}) \approx \beta_j(t_{(i)}) - \hat{\beta}_j$$

όπου $\beta_j(t_{(i)})$ είναι η πραγματική τιμή του συντελεστή της συμμεταβλητής του x_j, την στιγμή $t_{(i)}$ που συμβαίνει το γεγονός i, και $\hat{\beta}_j$ είναι η εκτίμηση του β_j μετά την προσαρμογή του μοντέλου.

Επομένως, η γραφική παράσταση των r^*_{Pji} ή των $r^*_{Pji} + \hat{\beta}_j$ ως προς τον χρόνο θα δώσει πληροφορίες για τον τύπο του $\beta_j(t)$. Συγκεκριμένα, αν η παραπάνω γραφική παράσταση είναι οριζόντια γραμμή, έπεται ότι το $\beta_j(t)$ είναι σταθερό και

ανεξάρτητο του χρόνου, επομένως ικανοποιείται η υπόθεση της αναλογικής διακινδύνευσης.

3.5 Υπόλοιπα Martingale (Martingale residuals)

Τα martingale υπόλοιπα (Therneau, Grambsch και Fleming, (1990)) χρησιμοποιούνται για την εύρεση άτυπων σημείων αλλά κυρίως για την εύρεση της συναρτησιακής μορφής μιας μεταβλητής που πρόκειται να εισαχθεί στο μοντέλο του Cox και υπολογίζονται από τη σχέση:

$$r_{Mi} = \delta_i - r_{ci}$$

όπου $\delta_i = \begin{cases} 0, & \text{για μη αποκομμένες παρατηρήσεις} \\ 1, & \text{για αποκομμένες παρατηρήσεις .} \end{cases}$

Τα martingale υπόλοιπα παίρνουν τιμές από $-\infty$ μέχρι 1, με τα υπόλοιπα για αποκομμένες παρατηρήσεις να είναι αρνητικά. Επίσης δεν κατανέμονται συμμετρικά γύρω από το μηδέν, ακόμα και όταν το προσαρμοσμένο μοντέλο είναι σωστό. Το άθροισμά τους είναι μηδέν. Σε μεγάλα δείγματα δε, είναι ανεξάρτητα το ένα με το άλλο και η αναμενόμενη τους τιμή είναι μηδέν. Αν κάποιο martingale υπόλοιπο έχει μια υψηλή αρνητική τιμή τότε το αντίστοιχο άτυπο σημείο (outlier) δεν ερμηνεύεται καλά από το μοντέλο μας.

Ένας άλλος τρόπος να δούμε τα martingale υπόλοιπα είναι σαν την απόσταση ανάμεσα στον παρατηρούμενο αριθμό πραγματοποίησης των γεγονότων στο διάστημα $(0, t_i)$ και στον αντίστοιχο αριθμό που εκτιμάται με βάση το μοντέλο.

3.6 Υπόλοιπα Score (Score residuals)

Ένας άλλος τύπος υπολοίπων που είναι χρήσιμος για την εξέταση κάποιων πλευρών του μοντέλου του Cox είναι τα υπόλοιπα Score. Αυτά τα υπόλοιπα υπολογίζονται, όπως τα υπόλοιπα Schoenfeld, από την πρώτη παράγωγο ως προς β_j, $j=1,\ldots,p$, του λογαρίθμου της συνάρτησης της μερικής πιθανοφάνειας. Η παράγωγος της μερικής πιθανοφάνειας είναι :

$$\frac{\partial \log L(\beta)}{\partial \beta_j} = \sum_{i=1}^{N} \left\{ \delta_i (x_{ji} - \hat{a}_{ji}) + \exp(\beta' x_i) \sum_{t_r \le t_i} \frac{(\hat{a}_{jr} - x_{ji})\delta_r}{\sum_{l \in R_r} \exp(\beta' x_l)} \right\}, \quad (3.3)$$

όπου x_{ji} είναι η i-οστή τιμή της j επεξηγηματικής μεταβλητής, $\delta_i = \begin{cases} 0, & \text{για μη αποκομμένες παρατηρήσεις} \\ 1, & \text{για αποκομμένες παρατηρήσεις} \end{cases}$, $\hat{a}_{ji}$ δίνεται από τη σχέση (3.2) και R_r είναι τα δεδομένα που βρίσκονται σε κίνδυνο τη στιγμή t_r.

Από την εξίσωση (3.3) το i-οστό υπόλοιπο Score, i=1,…,N, για την j-οστή επεξηγηματική μεταβλητή του μοντέλου, x_j, δίνεται από τη σχέση :

$$r_{Sji} = \delta_i(x_{ji} - \hat{a}_{ji}) + \exp(\hat{\beta}'x_i) \sum_{t_r \le t_i} \frac{(\hat{a}_{jr} - x_{ji})\delta_r)}{\sum_{l \in R_r} \exp(\hat{\beta}'x_l)}$$

Χρησιμοποιώντας τη σχέση (3.1), η παραπάνω σχέση γράφεται στη μορφή

$$r_{Sji} = r_{Pji} + \exp(\hat{\beta}'x_i) \sum_{t_r \le t_i} \frac{(\hat{a}_{jr} - x_{ji})\delta_r)}{\sum_{l \in R_r} \exp(\hat{\beta}'x_l)} \qquad (3.4)$$

η οποία δείχνει ότι τα υπόλοιπα Score είναι τροποποιήσεις των υπολοίπων Schoenfeld.

3.7 Υπόλοιπα Deviance

Τα υπόλοιπα deviance προκύπτουν από τα γενικευμένα γραμμικά μοντέλα, ωστόσο σχετίζονται στενά με τα martingale υπόλοιπα, έτσι ώστε να είναι συμμετρικά κατανεμημένα γύρω από το μηδέν, όταν το προσαρμοσμένο μοντέλο είναι κατάλληλο και έτσι είναι χρήσιμα για την ανίχνευση των άτυπων σημείων.

Υπολογίζονται από τη σχέση :

$$r_{Di} = \mathrm{sgn}(r_{Mi}) \, [-2[r_{Mi} + \delta_i \log(\delta_i - r_{Mi})]]^{1/2}, \qquad 1 \le i \le N$$

όπου $\mathrm{sgn}(r_{Mi}) = \begin{cases} 1, & \text{για } r_{Mi} > 0 \\ -1, & \text{για } r_{Mi} < 0 \end{cases}$

Αφού τα martingale υπόλοιπα παίρνουν τιμές στο διάστημα $(-\infty, 1]$ από την παραπάνω σχέση προκύπτει ότι αν το r_{Mi} παίρνει τιμές στο διάστημα $(-\infty, 0)$ τότε το αντίστοιχο r_{Di} μετατοπίζεται προς την κατεύθυνση της τιμής 0, και αν το r_{Mi} παίρνει τιμές στο διάστημα (0,1) τότε το αντίστοιχο μετατοπίζεται προς την κατεύθυνση της τιμής $+\infty$. Συνεπώς μπορούμε να πούμε ότι τα deviance υπόλοιπα αποτελούν μια εξομάλυνση των martingale υπολοίπων προς την κατεύθυνση της συμμετρικότητας γύρω από την τιμή μηδέν.

Παρόλο που τα υπόλοιπα deviance αναμένονται να είναι συμμετρικά κατανεμημένα γύρω από το μηδέν, όταν προσαρμοστεί ένα κατάλληλο μοντέλο, το άθροισμά τους δεν είναι απαραίτητα μηδέν.

ΚΕΦΑΛΑΙΟ 4

Ανίχνευση απόμακρων παρατηρήσεων (Outlier detection)

4.1 Ορισμός απόμακρων παρατηρήσεων (outliers)

Τα απόμακρα ή άτυπα σημεία (outliers) είναι δεδομένα που διαφέρουν πολύ από το υπόλοιπο ενός συνόλου. Η διαφορά μπορεί να εντοπιστεί στις τιμές μιας μεταβλητής ή στο συνδυασμό τιμών σε πολυδιάστατα δεδομένα. Ιδιαίτερη σημασία έχει το γεγονός ότι ενδεχομένως η παρουσία απόμακρων σημείων επηρεάζει αισθητά την εκτίμηση των παραμέτρων ενός μοντέλου, οδηγώντας σε λανθασμένα συμπεράσματα και μη ακριβείς προβλέψεις.

Τα απόμακρα σημεία ενδέχεται να είναι και λάθη, ωστόσο μπορεί κάποιες φορές να περιέχουν σημαντικές πληροφορίες.

Κάποιες συνηθισμένες αιτίες για την εμφάνιση των απόμακρων τιμών (outlier) είναι :

1) Σφάλματα κατά την καταχώριση δεδομένων

Όταν ψάχνουμε για άτυπα σημεία, πρώτα ελέγχουμε για λάθη αντιγραφής ή εισόδου δεδομένων. Αυτό μπορεί να απαιτεί να συγκρίνουμε τις τιμές που εισήχθηκαν στο φάκελο των δεδομένων μας, με τις τιμές που υπήρχαν στο φάκελο από τον οποίο τις πήραμε.

1) Αβάσιμες τιμές

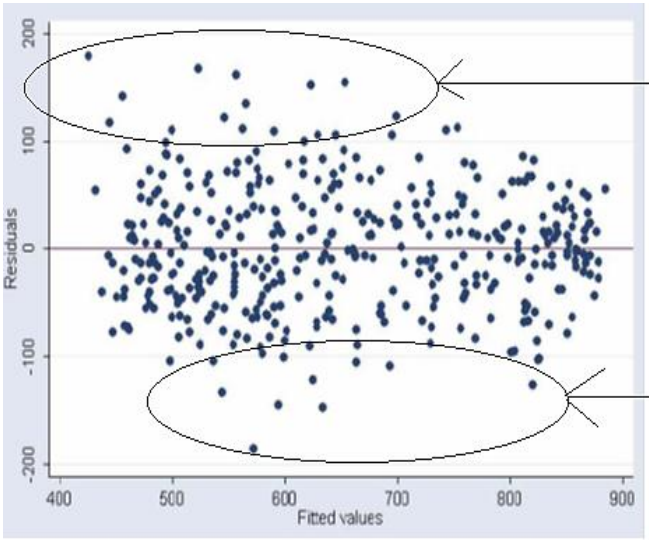
Ένας τύπος λαθών κατά την καταχώριση δεδομένων είναι οι ‘αβάσιμες’ ή ‘απίθανες’ τιμές, οι οποίες δεν έχουν καμία σχέση με αυτές που περιμέναμε ανάλογα με το εύρος των δεδομένων. Για παράδειγμα, όταν έχουμε μία θερμοκρασία 420 βαθμών Κελσίου ενώ αναμένουμε ένα εύρος τιμών 2-40, είναι φανερό ότι είναι ‘απίθανη τιμή’. Μία τιμή εκτός εύρους δεδομένων είναι συχνά εύκολο να αναγνωριστεί, εφ’ όσον θα είναι αρκετά μακριά από το κύριο μέρος των δεδομένων.

2) Σπάνια γεγονότα

Μία άλλη αιτία για την εμφάνιση άτυπων σημείων είναι το σύνδρομο των ‘σπάνιων’ γεγονότων, δηλαδή άτυπα σημεία που για κάποιο θεμιτό λόγο είναι σημαντικές, αλλά δεν ταιριάζουν μέσα στο τυπικό εύρος των υπόλοιπων παρατηρήσεων.

Όλα αυτά τα γεγονότα μπορεί να θεωρούνται σχετικά σπάνια, αλλά θα πρέπει να θεωρούνται μέρος της συνολικής εικόνας.

Στο Σχήμα 4.1 βλέπουμε την απεικόνιση των υπολοίπων ενός συνόλου δεδομένων, στο οποίο υπάρχουν θετικά και αρνητικά απόμακρα σημεία.



Σχήμα 4.1

Δεδομένα με άτυπα σημεία

Τα μοντέλα διάρκειας ζωής εφαρμόζονται στη μελέτη επιβίωσης και αξιοπιστίας (χρόνος μέχρι το θάνατο ατόμου ή χρόνος μέχρι τη διακοπή λειτουργίας μηχανής ή πίεση που προκαλεί θραύση υλικού). Άτυπα σημεία είναι μονάδες με εξαιρετικά μικρή ή ασυνήθιστα μεγάλη διάρκεια ζωής, ενδεχομένως με σημαντική επίδραση στη διαμόρφωση του μοντέλου και την εκτίμηση των παραμέτρων του.

Ο Hawkins (1980) όρισε ως άτυπο σημείο μία παρατήρηση που αποκλίνει τόσο πολύ από τις άλλες παρατηρήσεις, ώστε, να μας εξεγείρει τις υποψίες ότι δημιουργήθηκε από διαφορετικό μηχανισμό. Οι Barnett και Lewis (1994) υπέδειξαν ότι άτυπο σημείο είναι η παρατήρηση που φαίνεται να αποκλίνει σημαντικά από τις υπόλοιπες παρατηρήσεις του δείγματος. Στην ανάλυση επιβίωσης, το άτυπο σημείο ορίζεται ελαφρώς διαφορετικά από τα προβλήματα γραμμικής παλινδρόμησης. Ο Collet (2003) αναφέρθηκε στα άτυπα σημεία στην ανάλυση επιβίωσης ως παρατηρήσεις που έχουν εξαιρετικά μεγάλο χρόνο επιβίωσης, αλλά οι τιμές των επεξηγηματικών μεταβλητών δείχνουν ότι θα έπρεπε να είχαν συμβεί νωρίτερα και το αντίστροφο. Οι Nardi και Schemper (1999) έδωσαν τον ορισμό ότι άτυπα σημεία είναι οι παρατηρήσεις, των οποίων ο χρόνος επιβίωσης είναι πολύ μικρός ή πολύ μεγάλος, σε σχέση με την εκτιμώμενη πιθανότητα επιβίωσης, όπως προβλέπεται από το μοντέλο του Cox.

Στην ανάλυση επιβίωσης, τα άτυπα σημεία μπορούν να επηρεάσουν την εκτίμηση παραμέτρων των μοντέλων, να μεταβάλλουν τις αντίστοιχες αναλογίες διακινδύνευσης (hazard ratios) και ίσως να αλλάξουν το επιλεγόμενο μοντέλο.

4.2 Μέθοδοι για την αναγνώριση των απόμακρων παρατηρήσεων

Οι μέθοδοι για την αναγνώριση των άτυπων σημείων μπορούν να διαχωριστούν σε μονοδιάστατες και πολυδιάστατες (Ben-Gal, 2005). Μία άλλη βασική ταξινόμηση είναι σε παραμετρικές (στατιστικές) μεθόδους και σε μη παραμετρικές μεθόδους που δεν ακολουθούν κάποιο μοντέλο. Οι στατιστικές παραμετρικές μέθοδοι είτε υποθέτουν ότι οι παρατηρήσεις ακολουθούν μία γνωστή κατανομή, είτε τουλάχιστον βασίζονται σε στατιστικές εκτιμήσεις αγνώστων παραμέτρων της κατανομής.

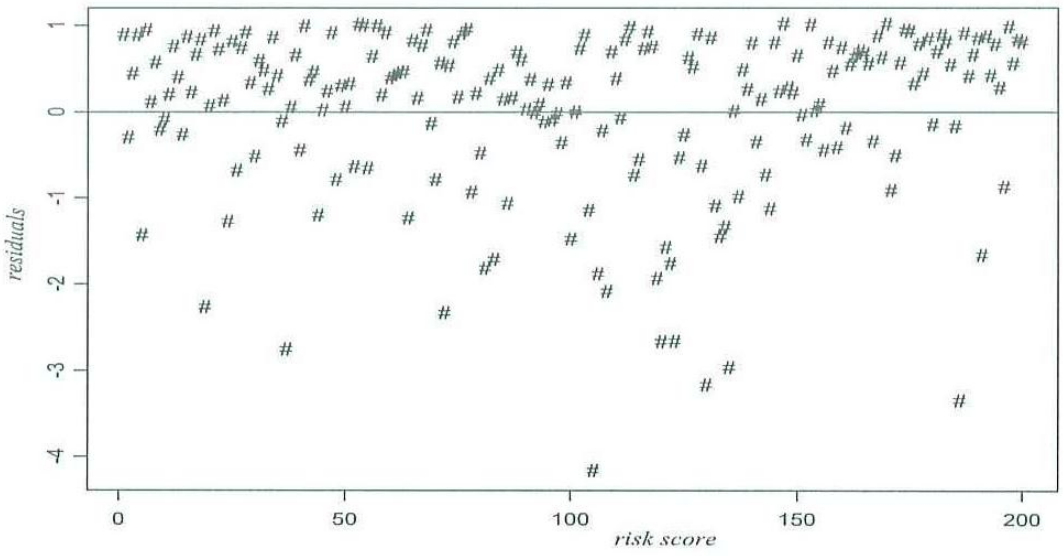
Αυτές οι μέθοδοι υποδεικνύουν ως άτυπα σημεία τις παρατηρήσεις που αποκλίνουν από τις υποθέσεις του μοντέλου.

Οι μέθοδοι για την αναγνώριση των άτυπων παρατηρήσεων βασίζονται κυρίως στα υπόλοιπα. Συνήθως γίνεται μέσω των martingale, των deviance, των normal deviate ή των log-odds υπολοίπων

4.2.1 Martingale υπόλοιπα

Στο μοντέλο αναλογικής διακινδύνευσης του Cox ένας τύπος υπολοίπων που μπορεί να χρησιμοποιηθεί είναι τα martingale υπόλοιπα. Όπως έχουμε αναφέρει αυτά τα υπόλοιπα δεν είναι συμμετρικά κατανεμημένα γύρω από το μηδέν, ακόμα και αν το προσαρμοσμένο μοντέλο είναι σωστό, με αποτέλεσμα να υπάρχει δυσκολία στον εντοπισμό των παρατηρήσεων που προβλέπονται ανεπαρκώς από το μοντέλο. Για παράδειγμα, το Σχήμα 4.2 των υπολοίπων martingale από την προσομοίωση ενός συνόλου 200 παρατηρήσεων, όπως φαίνεται παρακάτω δείχνει καθαρά αυτή την ιδιότητα. Αυτή η ασυμμετρία κάνει το γράφημα των υπολοίπων δύσκολο να ερμηνευθεί. Τα υπόλοιπα των παρατηρήσεων που 'πραγματοποιούνται πολύ νωρίς' θα έχουν συμπιεστεί στο +1, ενώ τα υπόλοιπα εκείνων που 'πραγματοποιούνται πολύ καθυστερημένα' θα παίρνουν μεγάλες τιμές κοντά στο - ∞.

Πολύ μεγάλες ή πολύ μικρές τιμές των υπολοίπων δείχνουν ότι οι αντίστοιχες παρατηρήσεις ίσως είναι άτυπα σημεία και χρειάζονται ιδιαίτερη προσοχή.

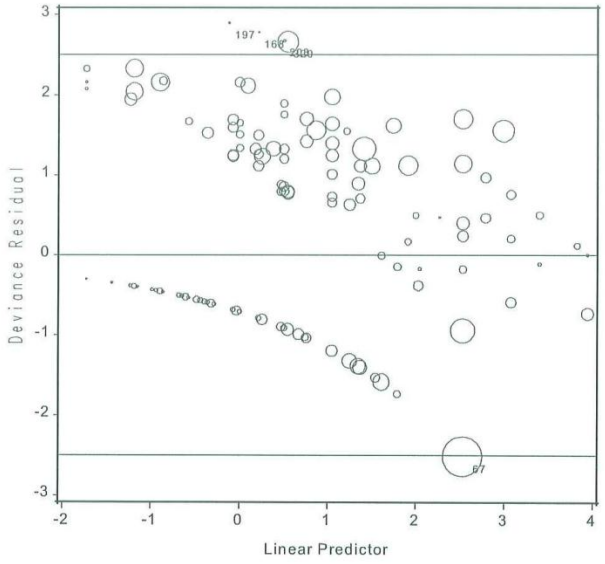


Σχήμα 4.2

Τα martingale υπόλοιπα ως προς το risk score plot σε δείγμα 200 παρατηρήσεων χωρίς αποκομμένες παρατηρήσεις

4.2.2 Deviance υπόλοιπα

Τα deviance υπόλοιπα προτάθηκαν για την ανίχνευση άτυπων σημείων σε γενικευμένα γραμμικά μοντέλα και στην ανάλυση επιβίωσης για να ξεπεραστεί το πρόβλημα της ασυμμετρίας. Παρόλο που δεν έχει αποδειχθεί ότι τα deviance υπόλοιπα είναι συμμετρικά κατανεμημένα, η κατανομή τους είναι πιο κοντά στην Κανονική απ' ότι των άλλων υπολοίπων που χρησιμοποιούνται στην ανάλυση επιβίωσης. Χρησιμοποιούνται στην ανίχνευση άτυπων σημείων με την υπόθεση ότι είναι κατανεμημένα γύρω από το μηδέν και με τυπική απόκλιση περίπου 1. Είναι αρνητικά για παρατηρήσεις που έχουν μεγαλύτερο χρόνο επιβίωσης από τον αναμενόμενο και θετικά για παρατηρήσεις με χρόνο επιβίωσης μικρότερο από τον αναμενόμενο. Πολύ μεγάλες ή πολύ μικρές τιμές δείχνουν ότι η αντίστοιχη παρατήρηση ίσως είναι άτυπο σημείο. Τα υπόλοιπα μπορούν να παρασταθούν γραφικά ως προς τις συμμεταβλητές και οποιαδήποτε ασυνήθιστα σχέδια μπορεί να υποδεικνύουν ιδιότητες των δεδομένων που δεν έχουν προσαρμοστεί κατάλληλα από το μοντέλο. Ωστόσο, αποκομμένα δεδομένα ίσως δημιουργήσουν 'περίεργα' σχέδια τα οποία δεν συνεπάγονται κατ' ανάγκη ότι υπάρχει πρόβλημα με το μοντέλο.



Σχήμα 4.3: Υπόλοιπα Deviance

Γράφημα των deviance υπολοίπων

Διαγνωστικά για outlier και σημεία επιρροής

Το Σχήμα 4.3 δείχνει τα deviance υπόλοιπα από την προσομοίωση ενός συνόλου 329 ασθενών, οι οποίοι έχουν βγει από μία ψυχιατρική κλινική και μελετάται η κατάστασή τους μετά από έξι και δώδεκα μήνες, τα οποία περιέχουν σημεία επιρροής και άτυπα σημεία. Όπως βλέπουμε, οι παρατηρήσεις με τα νούμερα 197, 168 και 67 είναι άτυπα σημεία. Οι παρατηρήσεις με σχετικά μεγάλες 'φούσκες' θεωρούνται ως σημεία επιρροής. Οι μεγάλες 'φούσκες' ανάμεσα στις δύο γραμμές δεν είναι άτυπα σημεία, αλλά ασκούν μεγάλη επιρροή.

4.2.3 Normal deviate και Log-odds υπόλοιπα

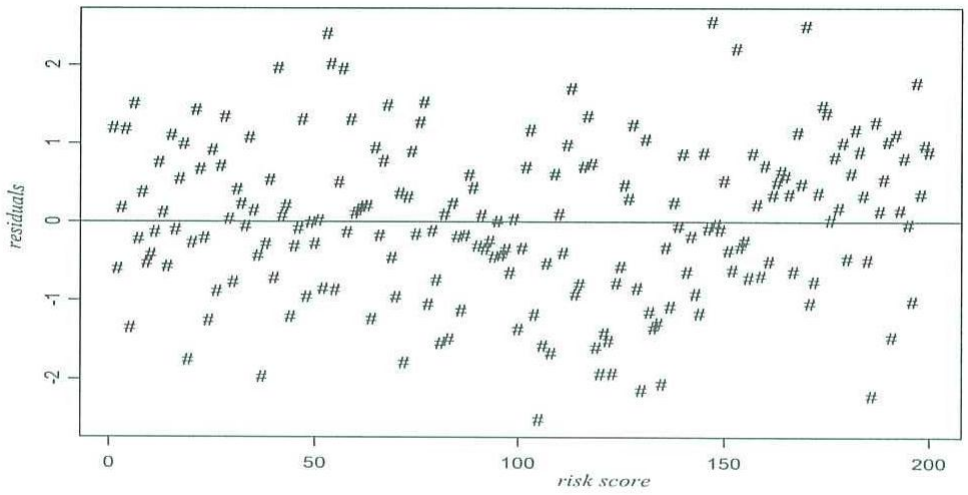
4.2.3.1 Εισαγωγή

Ωστόσο, οι Fleming και Harrington (1991) επεσήμαναν ότι τα deviance υπόλοιπα δεν έχουν γνωστή κατανομή και η προσέγγισή τους από την Κανονική κατανομή, φαίνεται μη ικανοποιητική ακόμα και χωρίς αποκομμένες παρατηρήσεις στο σύνολο δεδομένων. Αργότερα οι Nardi και Schemper (1999), για να ξεπεράσουν αυτό το πρόβλημα, πρότειναν δύο νέα είδη υπολοίπων με κύριο σκοπό την ανίχνευση των άτυπων σημείων στο μοντέλο αναλογικής διακινδύνευσης του Cox. Έπειτα επέκτειναν τη θεωρία σε αρκετά παραμετρικά μοντέλα. Αυτά τα υπόλοιπα είναι γνωστά ως log-odds υπόλοιπα (log-odds residuals) και normal deviate υπόλοιπα (normal deviate residuals) και δημιουργήθηκαν μετασχηματίζοντας την πιθανότητα επιβίωσης, χρησιμοποιώντας τους logit και probit μετασχηματισμούς, αντίστοιχα. Προβάλλεται ο ισχυρισμός ότι οι διαδικασίες που βασίζονται σε αυτά τα νέα υπόλοιπα έχουν καλύτερα αποτελέσματα στο σωστό εντοπισμό των απόμακρων τιμών (outliers). Αυτά τα δύο υπόλοιπα είναι συμμετρικά κατανεμημένα γύρω από το μηδέν και αφού έχουν την απαραίτητη κατανομή, μπορούν να χρησιμοποιηθούν για την ανίχνευση άτυπων σημείων.

4.2.3.2 Normal deviate υπόλοιπα

Τα normal deviate υπόλοιπα, όπως αναφέραμε, εισάγονται από τους Nardi και Schemper (1999) και συμβολίζονται με r_{Ni}. Η δειγματική κατανομή αναφοράς των υπολοίπων είναι η τυποποιημένη Κανονική κατανομή. Επειδή $S(T)\sim U[0,1]$, τότε $Z=\Phi^{-1}[S(T)]\sim N(0,1)$, όπου Φ η συνάρτηση της τυποποιημένης Κανονικής κατανομής. Αν αντικαταστήσουμε την άγνωστη συνάρτηση επιβίωσης με την

εκτίμησή της και υποθέσουμε ένα σωστά καθορισμένο μοντέλο, το $\hat{Z} = \Phi^{-1}[\hat{S}(T)]$ συγκλίνει στην πιθανότητα Z. Οι τιμές των υπολοίπων είναι $\hat{z}_i = \Phi^{-1}\{\hat{S}_i(t_i)\}$. Η πιθανότητα επιβίωσης $S_i(t_i)$ για αποκομμένα δεδομένα είναι άγνωστη. Υπάρχουν πολλοί τρόποι για να διευκολύνουμε τον υπολογισμό των υπολοίπων για αποκομμένους χρόνους επιβίωσης. Ο πραγματικός χρόνος επιβίωσης είναι μεγαλύτερος από τον αντίστοιχο παρατηρούμενο, αποκομμένο και η κατανομή των άγνωστων πραγματικών υπολοίπων σχετίζεται με την Ομοιόμορφη κατανομή του $S_i(T_i)$ στο $[0, \hat{S}_i(t_i)]$. Έτσι, το $\hat{S}_i(t_i)$ θα αντικατασταθεί από την μέση τιμή $\frac{\hat{S}_i(t_i^*)}{2}$, όπου t_i^* είναι οι χρόνοι για τις αποκομμένες παρατηρήσεις. Συνεπώς τα normal deviate υπόλοιπα, τα οποία συμβολίζονται με r_{Zi} για αποκομμένο χρόνο είναι $z_i^* = \Phi^{-1}\{\frac{\hat{S}_i(t_i^*)}{2}\}$ όπου z_i^* είναι η μέση τιμή των r_{Zi} για αποκομμένο χρόνο, ή μπορεί να αντικατασταθεί από το $z_i^m = -\frac{\exp(-0.5(z_i^*)^2)}{\sqrt{2\pi}\hat{S}_i(t_i^*)}$ όπου z_i^m είναι η διάμεσος των r_{Zi} για αποκομμένο χρόνο.



Σχήμα 4.4

Τα Normal deviate υπόλοιπα ως προς το risk score plot σε δείγμα 200 δεδομένων χωρίς αποκομμένες παρατηρήσεις

Όπως βλέπουμε στο Σχήμα 4.4 τα normal deviate υπόλοιπα, που προκύπτουν από την προσομοίωση ενός συνόλου 200 παρατηρήσεων, είναι συμμετρικά κατανεμημένα γύρω από το μηδέν.

4.2.3.3 Log-odds υπόλοιπα

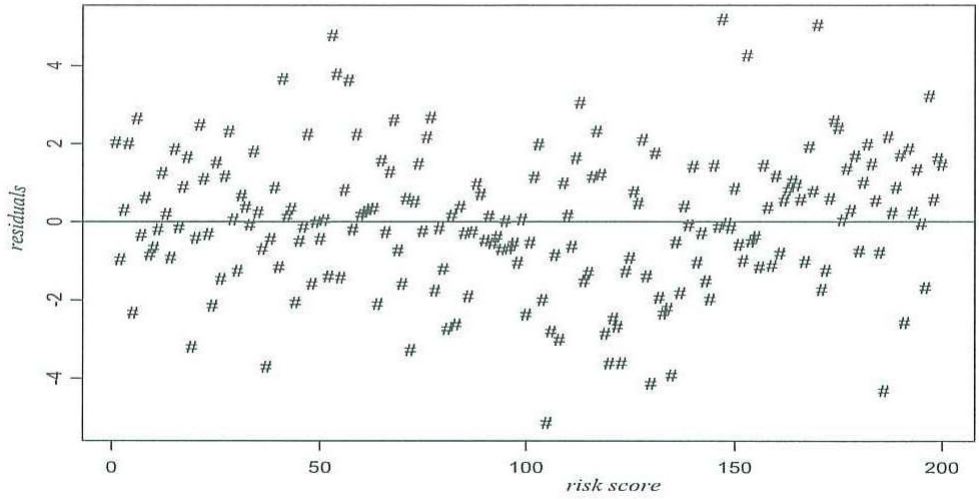
Οι Nardi and Schemper (1999) εισήγαγαν τα Log-odds υπόλοιπα και τα συμβόλισαν με r_{Li}. Η δειγματοληπτική κατανομή αναφοράς των υπολοίπων είναι η Λογιστική κατανομή με μέση τιμή $E(L_i) = 0$ και διακύμανση $var(L_i)= (\pi/3)^2$. Έχοντας το σωστά καθορισμένο μοντέλο και αντικαθιστώντας την άγνωστη συνάρτηση επιβίωσης από την εκτίμηση της, έχουμε το $\hat{L}_i$ να συγκλίνει στην πιθανότητα L_i . Ως εκ τούτου, έχουμε

$$L_i = \log[S_i(T_i)/\{1 - S_i(T_i)\}]$$

και

$$\hat{l}_i = \log[\hat{S}_i(t_i) / \{1 - \hat{S}_i(t_i)\}].$$

Η πιθανότητα επιβίωσης $S_i(t_i^*)$ για αποκομμένο χρόνο είναι άγνωστη. Επομένως, για να διευκολύνουν τον υπολογισμό των υπολοίπων με αποκομμένο χρόνο επιβίωσης, οι Nardi και Schemper πρότειναν την αντικατάσταση της $\hat{S}_i(t_i^*)$ με την $\frac{\hat{S}_i(t_i^*)}{2}$. Έτσι, η διάμεσος l_i^* και η μέση τιμή l_i^m των τιμών των log-odds υπολοίπων δίνονται από τους τύπους $l_i^* = \log[\hat{S}_i(t_i^*) / \{2 - \hat{S}_i(t_i^*)\}]$ και $l_i^m = l_i^* - \frac{1 + \exp(l_i^*)}{\exp(l_i^*)} \log\{1 + \exp(l_i^*)\}$ αντίστοιχα.



Σχήμα 4.5

Τα Log-odds υπόλοιπα ως προς το risk score plot σε δείγμα 200 δεδομένων χωρίς αποκομμένες παρατηρήσεις

Όπως βλέπουμε στο Σχήμα 4.5 τα log-odds υπόλοιπα, που προκύπτουν από την προσομοίωση ενός συνόλου 200 παρατηρήσεων, είναι συμμετρικά κατανεμημένα γύρω από το μηδέν.

Παρατηρούμε ότι και τα δύο υπόλοιπα είναι μηδέν, όταν ο παρατηρούμενος χρόνος επιβίωσης συμπίπτει με τον εκτιμώμενο μέσο χρόνο επιβίωσης, ο οποίος θεωρείται ως χρόνος αναφοράς. Η αύξηση των αποκλίσεων από τον προβλεπόμενο μέσο χρόνο αντανακλάται από την αύξηση των απολύτων τιμών. Μεγάλες τιμές αρνητικών και θετικών υπολοίπων οδηγούν σε πολύ μεγάλους ή πολύ μικρούς χρόνους επιβίωσης.

Υποθέτοντας ότι η συνάρτηση επιβίωσης είναι γνωστή, τα L και Z ακολουθούν την τυποποιημένη Λογιστική και τυποποιημένη Κανονική κατανομή, αντίστοιχα. Αυτό έχει ως αποτέλεσμα ότι $U_i = S(T_i)$, i = 1,...,N αντιπροσωπεύουν ένα πλήθος N ανεξάρτητων τυχαίων μεταβλητών, από τις οποίες η καθεμία ακολουθεί μία [0, 1] Ομοιόμορφη κατανομή. Τα δύο αυτά υπόλοιπα μπορούν να θεωρηθούν ως γενικευμένα υπόλοιπα στη λογική των Cox-Snell. Συγκρινόμενα

με τα κλασικά Cox-Snell υπόλοιπα έχουν δύο σημαντικά πλεονεκτήματα. Μπορούν διαισθητικά να ερμηνευθούν σαν την απόσταση ανάμεσα στην προβλεπόμενη μέση τιμή του χρόνου και τον παρατηρούμενο χρόνο επιβίωσης. Επιπλέον, η συμμετρία των αναφερόμενων κατανομών, η οποία μοιάζει με την ιδιότητα των υπολοίπων στο γενικό γραμμικό μοντέλο, βοηθάει σε οποιαδήποτε γραφική διαδικασία. Στην πραγματικότητα, αυτό αποφεύγει την υπερβολική οπτική επίδραση στην πάνω ουρά της κατανομής των υπολοίπων Cox-Snell, το οποίο προκύπτει από την εφαρμογή του λογαριθμικού μετασχηματισμού πάνω στην συνάρτηση επιβίωσης.

Συνεπώς, οι γραφικές παραστάσεις για την ανίχνευση άτυπων σημείων μπορούν να γίνουν με βάση αυτούς τους τρεις τύπους υπολοίπων : τα deviance, τα normal deviate, και τα log-odds υπόλοιπα. Η γραφική τεχνική γίνεται κάνοντας την γραφική παράσταση των υπολοίπων ως προς τον προγνωστικό δείκτη ($\hat{\beta}x_i$) για όλα τα άτομα, όπως φαίνεται στα Σχήματα 4.3, 4.4, 4.5.

ΚΕΦΑΛΑΙΟ 5

Επιρροή (influence)

5.1 Ορισμός

Ο Therneau (1990) και οι Nardi και Schemper (1999) συσχέτισαν τα απομονωμένα σημεία με παρατηρήσεις που ‘πραγματοποιούνται πολύ σύντομα’ ή ‘αργούν πολύ να πραγματοποιηθούν’. Αν αυτές οι παρατηρήσεις, επηρεάζουν τα συμπεράσματα που προέκυψαν με βάση το μοντέλο, τότε οι παρατηρήσεις προσδιορίζονται ως σημεία επιρροής (influential observations). Πολλοί συγγραφείς επεσήμαναν ότι τα σημεία επιρροής στην ανάλυση επιβίωσης συχνά βρίσκονται ανάμεσα σε μεγάλους χρόνους ζωής.

Τα άτυπα σημεία ίσως προκαλούν υπερβολική επίδραση στα συμπεράσματα που γίνονται με βάση το μοντέλο. Αυτά τα άτυπα σημεία καλούνται σημεία επιρροής.

Στο μοντέλο αναλογικής διακινδύνευσης του Cox, τα σημεία επιρροής επιδρούν στην εκτίμηση των παραμέτρων και συνεπώς αλλάζουν τις αναλογίες διακινδύνευσης.

Σημεία επιρροής είναι οι παρατηρήσεις των οποίων η διαγραφή οδηγεί σε σημαντικές αλλαγές στην εκτίμηση των παραμέτρων, στις προσαρμοσμένες τιμές ή στους ελέγχους των υποθέσεων, σε διαφορετικά συμπεράσματα στην ανάλυση και ίσως σε διαφορετικό μοντέλο του Cox.

Επιρροή είναι το μέτρο του πόσο μία παρατήρηση επηρεάζει την εκτίμηση των παραμέτρων.

Ο Belsley (1980) όρισε ως σημείο επιρροής (influential observation) εκείνη την παρατήρηση που, μόνη της ή μαζί με άλλες παρατηρήσεις, έχει μία αρκετά μεγάλη επίδραση στις εκτιμήσεις που υπολογίζουμε για διάφορες τιμές (συντελεστές, τυπικά σφάλματα, t-τιμές κ.τ.λ).

5.2 Μέθοδοι για τον προσδιορισμό σημείων επιρροής (influential observations)

5.2.1 Γενικά

Πέρα από το πρόβλημα της ανίχνευσης των άτυπων σημείων (outlier detection), υπάρχει επίσης η ανάγκη να αναπτυχθεί μία διαδικασία για τον προσδιορισμό των σημείων επιρροής. Στα προβλήματα παλινδρόμησης, οι Belsley (1980), Hadi (1992) και Imon (2005), όπως και οι Wang, Jones και Storer (2006)εντόπισαν τα σημεία επιρροής σε γραμμικά προβλήματα παλινδρόμησης χρησιμοποιώντας την μέθοδο delete-case. Από την άλλη πλευρά, παρόμοιες μέθοδοι, όπως η delta-beta διαδικασία έχουν επίσης προταθεί για τον προσδιορισμό των σημείων επιρροής σε προβλήματα επιβίωσης.

Εκτός από την delta-beta διαδικασία (delta-beta procedure), οι Cain and Lange (1984) and Reid and Crépeau (1985) χρησιμοποίησαν τη συνάρτηση επιρροής (influence function (IF)) στη διερεύνηση του προβλήματος, ενώ οι Storer and Crowley (1985) πρότειναν την επαυξημένη προσέγγιση (augmented approach(AUG)). Ο Wang (2006) σύγκρινε όλες τις παραπάνω προσεγγίσεις και έδειξε ότι η AUG approach υπερτερεί σαφώς από τις υπόλοιπες.

Εν τω μεταξύ, οι Atkinson και Riani (2000) πρότειναν μία εναλλακτική μέθοδο για την αναγνώριση των σημείων επιρροής στο μοντέλο παλινδρόμησης, η οποία ονομάζεται forward search μέθοδος. Αυτή η προσέγγιση ξεκινάει με ένα αρχικό υποσύνολο δεδομένων, το οποίο υποτίθεται ότι δεν περιέχει απομονωμένα σημεία. Τα σημεία επιρροής προσδιορίζονται εξετάζοντας τις αλλαγές που προκύπτουν στην παράμετρο για την οποία ενδιαφερόμαστε.

5.2.2 Delta-beta διαδικασία (delta-beta procedure)

Η delta-beta διαδικασία είναι η τυποποιημένη διαδικασία που χρησιμοποιείται για τον προσδιορισμό των σημείων επιρροής στη γενική γραμμική παλινδρόμηση. Εισήχθησαν από τον Belsley (1980) και η διαδικασία παρατηρεί τις αλλαγές που γίνονται στις εκτιμήσεις των παραμέτρων που ενδιαφερόμαστε, κάθε φορά που αφαιρούμε μία παρατήρηση. Αν η αφαίρεση μίας παρατήρησης

οδηγήσει σε σημαντικές αλλαγές, τότε η παρατήρηση θα είναι σημείο επιρροής (Καρώνη, 2010).

Η ίδια διαδικασία μπορεί να επεκταθεί και στο μοντέλο του Cox (Collett, 2003 και Therneau και Grambsch (2000)). Θέλουμε να ελέγξουμε την επιρροή κάθε παρατήρησης στην εκτίμηση $\hat{\beta}$ του β. Έστω $\hat{\beta}_i$ το διάνυσμα με τις εκτιμήσεις των συντελεστών που έχουν υπολογιστεί, έχοντας αφαιρέσει την i-οστή παρατήρηση. Έπειτα, ελέγχουμε ποιοι όροι του διανύσματος $\hat{\beta}$-$\hat{\beta}_i$ έχουν αδικαιολόγητα μεγάλες απόλυτες τιμές. Αυτό το κάνουμε για κάθε μία από τις N παρατηρήσεις. Αυτό το μέτρο είναι παρόμοιο με τα dfbetas για την γραμμική παλινδρόμηση, τα οποία περιλαμβάνουν την προσαρμογή N+1 μοντέλων του Cox, το οποίο είναι υπολογιστικά χρονοβόρο, εκτός αν το δείγμα είναι μικρό. Ευτυχώς, υπάρχει μία προσέγγιση, που βασίζεται στο μοντέλο του Cox, η οποία λαμβάνεται από όλα τα δεδομένα και μπορεί να χρησιμοποιηθεί για να παρακαμφθεί αυτή η υπολογιστική δαπάνη. Έστω r_{Si} να συμβολίζει το διάνυσμα των τιμών των υπολοίπων Score για την i-οστή παρατήρηση, δηλαδή $r'_{Si}= (r_{S1i},r_{S2i},\ldots,r_{Spi})$, όπου r_{Sji} , j=1,2,…,p, είναι το i-οστό υπόλοιπο Score το οποίο δίνεται από την εξίσωση (3.4). Μία προσέγγιση του $\hat{\beta}_j - \hat{\beta}_{j(i)}$, όπου $\hat{\beta}_j$ είναι η j-οστή παράμετρος στο προσαρμοσμένο μοντέλο του Cox και το $\hat{\beta}_{j(i)}$ λαμβάνεται από την προσαρμογή του μοντέλου μετά την αφαίρεση της παρατήρησης i, είναι το j-οστό στοιχείο του διανύσματος r'_{Si} var($\hat{\beta}$), όπου var($\hat{\beta}$) είναι ο πίνακας διασποράς της εκτίμησης των παραμέτρων του προσαρμοσμένου μοντέλου αναλογικής διακινδύνευσης του Cox. Το j-οστό στοιχείο αυτού του διανύσματος, το οποίο καλείται delta-beta, συμβολίζεται με $\Delta_i\hat{\beta}_{(j)}$ και ισχύει ότι

$$\Delta_i\hat{\beta}_{(j)} \approx \hat{\beta}_j - \hat{\beta}_{j(i)}.$$

Οι παρατηρήσεις που επηρεάζουν μία συγκεκριμένη εκτίμηση παραμέτρου, ας θεωρήσουμε την j-οστή, είναι εκείνες οι τιμές των $\Delta_i\hat{\beta}_{(j)}$, δηλαδή τα delta-beta για εκείνες τις παρατηρήσεις, οι οποίες είναι μεγαλύτερες κατ' απόλυτη τιμή απ' ότι των

υπολοίπων παρατηρήσεων του συνόλου. Το γράφημα των delta-beta για κάθε μεταβλητή του μοντέλου του Cox θα δείξει αν υπάρχουν παρατηρήσεις που έχουν μεγάλη επίδραση στην εκτίμηση των παραμέτρων για κάποια συγκεκριμένη μεταβλητή. Επιπλέον, ένα γράφημα των τιμών $\Delta_i\hat{\beta}_{(j)}$ ως προς τους ταξινομημένους χρόνους επιβίωσης δίνει πληροφορίες για τη σχέση ανάμεσα στο χρόνο επιβίωσης και την επιρροή (Qi, 2009).

5.2.3 Influence function (IF)

Ένας άλλος τρόπος αξιολόγησης της επιρροής κάθε παρατήρησης της καταλληλότητας του μοντέλου του Cox είναι να εξετάσουμε τη μεταβολή της τιμής του μείον του διπλάσιου του λογαρίθμου της μεγιστοποιημένης μερικής πιθανοφάνειας , $-2\log\hat{L}$,κάτω από το προσαρμοσμένο μοντέλο του Cox, όταν κάθε παρατήρηση με τη σειρά της, αφαιρείται. Η τιμή όταν το μοντέλο του Cox προσαρμόζεται με όλες τις παρατηρήσεις είναι $-2\log L(\hat{\beta})$ και όταν η εκτίμηση της παραμέτρου υπολογίζεται μετά την αφαίρεση της i-οστής παρατήρησης, αντίστοιχη τιμή είναι $-2\log L(\hat{\beta}_{(i)})$. Η διαγνωστική συνάρτηση $2\{ \log L(\hat{\beta}) - \log L(\hat{\beta}_{(i)})\}$ είναι πολύ χρήσιμη για την επιρροή.

Οι Pettitt και Bin Daud (1989) έδειξαν ότι μία προσέγγιση της παραπάνω ποσότητας είναι

$$LDi = r'_{Si}\, var(\hat{\beta})\, r_{Si} \quad (5.1)$$

όπου r_{Si} είναι το $p\times 1$ είναι το διάνυσμα των υπολοίπων Score, του οποίου το j-οστό στοιχείο δίνεται από την εξίσωση (3.4) και $var(\hat{\beta})$ είναι ο πίνακας διασποράς της εκτίμησης των παραμέτρων του προσαρμοσμένου μοντέλου αναλογικής διακινδύνευσης του Cox. Οι τιμές της σχέσης (5.1) μπορούν να υπολογισθούν ευθέως από τα delta-beta κάθε μεταβλητής του μοντέλου του Cox. Ένα γράφημα της ποσότητας (5.1) ως προς τους ταξινομημένους χρόνους επιβίωσης μας δείχνει τα σημεία επιρροής, τα οποία είναι οι παρατηρήσεις οι οποίες έχουν μεγάλες τιμές στο γράφημα.

5.2.4 Augmented (AUG) approach

Οι Storer και Crowley (1985) πρότειναν την επαυξημένη προσέγγιση (augmented (AUG) approach) συμπεριλαμβάνοντας χρονικά εξαρτημένες μεταβλητές για την αναγνώριση των σημείων επιρροής στο μοντέλο του Cox.

Η διάγνωση της επαυξημένης προσέγγισης που συμβολίζεται με $\Delta_i\hat{\beta}_{(j)}^{AUG}$ λαμβάνεται εξαλείφοντας την j-οστή παρατήρηση από τα δεδομένα. Για να δούμε την επίδραση, χρησιμοποιούμε το γράφημα της $\Delta_i\hat{\beta}_{(j)}^{AUG}$ με το j, για j = 1,2,...,p. Αν παρατηρήσουμε μεγάλες αλλαγές σε κάποιο από τα γραφήματα, τότε η αντίστοιχη παρατήρηση προσδιορίζεται ως σημείο επιρροής. Η απόδοση των τριών τύπων των μεθόδων case-deletion συγκρίθηκε από τον Wang (2006). Άλλες προσεγγίσεις περιλαμβάνουν τα added variable plot και τα variable plot.

5.2.5 Forward Search Method

Οι Atkinson and Riani (2000) πρότειναν μία πολύ ισχυρή μέθοδο, γνωστή ως forward search μέθοδο για την ανίχνευση σημείων επιρροής σε απλά και πολλαπλά προβλήματα γραμμικής παλινδρόμησης. Βασιζόμενοι στα υπόλοιπα που παίρνουμε από το προσαρμοσμένο μοντέλο, ένα αρχικό υποσύνολο μεγέθους m<n το οποίο δεν περιέχει απομονωμένα σημεία σχηματίζεται από τα δεδομένα. Η επίδραση της προσθήκης μίας μεταβλητής τη φορά στο αρχικό υποσύνολο για τα στατιστικά που ενδιαφερόμαστε παρακολουθείται συνεχώς όσο όλες οι παρατηρήσεις βρίσκονται στο υποσύνολο. Το να συμπεριλάβουμε σημεία επιρροής (influential observations) αναμένεται να προκαλέσει μερικές σημαντικές αλλαγές στις εκτιμήσεις των στατιστικών.

Η FS μέθοδος επεκτείνεται και στο μοντέλο αναλογικής διακινδύνευσης του Cox από τον Nor Akmal (2010). Αυτή η FS μέθοδος δημιουργήθηκε για να επεκτείνει και να βελτιώσει τη μέθοδο των Atkinson και Riani (2000), η οποία χρησιμοποιεί τα deviance υπόλοιπα για την αναγνώριση των σημείων επιρροής σε ένα μοντέλο παλινδρόμησης. Η FS μέθοδος που προτάθηκε για το μοντέλο του Cox, χρησιμοποιεί τριών ειδών υπόλοιπα, τα υπόλοιπα deviance r_{Di}, τα normal deviate υπόλοιπα r_{Zi} και τα log-odds υπόλοιπα r_{Li}.

Η FS μέθοδος για το μοντέλο του Cox περιλαμβάνει τα τρία παρακάτω βήματα:

1)Την επιλογή ενός αρχικού υποσυνόλου από το πλήρες σύνολο δεδομένων.

2)Την πρόσθεση παρατηρήσεων κατά τη διάρκεια της αναζήτησης.

3)Την παρακολούθηση της αναζήτησης

Βήμα $1^{ο}$: Επιλογή ενός αρχικού υποσυνόλου από το πλήρες σύνολο δεδομένων.

Η FS μέθοδος για το μοντέλο του Cox ξεκινά με την προσαρμογή του αρχικού συνόλου δεδομένων με το μοντέλο του Cox και ονομάζοντας το 'καλύτερο' προσαρμοσμένο μοντέλο, μοντέλο Α. Έπειτα, επιλέγουμε κάποια από τις παρακάτω τεχνικές για να διαμορφώσουμε το αρχικό υποσύνολο. Συμβολίζουμε το αρχικό υποσύνολο με $S_{*}^{(m)}$, όπου m είναι το μέγεθος του αρχικού υποσυνόλου.

Τεχνική 1: Προσαρμόζουμε το μοντέλο Α στο αρχικό σύνολο δεδομένων και παρατηρούμε τις τιμές των r_{Di}^2. Έπειτα, διατάσσουμε τα δεδομένα με βάση τις τιμές των r_{Di}^2. Το αρχικό υποσύνολο σχηματίζεται διαλέγοντας τουλάχιστον το 50% των παρατηρήσεων που έχουν τις μικρότερες τιμές των r_{Di}^2.

Τεχνική 2: Η τεχνική είναι παρόμοια με την τεχνική 1, με τη διαφορά ότι αντί για τα υπόλοιπα r_{Di} χρησιμοποιούμε τα r_{Zi}.

Τεχνική 3: Επίσης είναι παρόμοια με την 1, αλλά στη θέση των r_{Di} χρησιμοποιούμε τα r_{Li}.

Τεχνική 4: Εφαρμόζουμε τη μέθοδο case-deletion στα αρχικά δεδομένα μας. Αφαιρούμε μία παρατήρηση την φορά. Έπειτα, τα δεδομένα που απομένουν τα προσαρμόζουμε χρησιμοποιώντας το μοντέλο Α. Στην συνέχεια, καταγράφεται η εκτίμηση των παραμέτρων $(\hat{\beta}_1,\hat{\beta}_2,\ldots,\hat{\beta}_p)$. Αυτό επαναλαμβάνεται για κάθε παρατήρηση του συνόλου δεδομένων μας. Φτιάχνουμε το διάστημα της διαμέσου $(\hat{\beta}_{j(i)}) \pm \gamma\ \mathrm{MAD}(\hat{\beta}_{j(i)})$ όπου $\hat{\beta}_{j(i)}$ είναι η εκτίμηση της παραμέτρου β_j όπου $j=1,2,\ldots p$ όταν η i-οστή παρατήρηση αφαιρείται, $\mathrm{MAD}(\hat{\beta}_{j(i)})$ η διάμεση απόλυτη απόκλιση και $\gamma=1,2,\ldots$. Αν το $\hat{\beta}_j^{(-i)}$ βρίσκεται έξω από το διάστημα, τότε η i-οστή παρατήρηση δεν θα περιέχεται στο αρχικό υποσύνολο για όλα τα $j=1,2,\ldots p$.

Τεχνική 5: Προσαρμόζουμε το μοντέλο Α στο σύνολο με όλα τα δεδομένα και υπολογίζουμε τα r_{Di}. Τότε το αρχικό υποσύνολο σχηματίζεται διαλέγοντας τις παρατηρήσεις των οποίων οι τιμές των r_{Di} κυμαίνονται ανάμεσα στο -1 και στο 3.

Τεχνική 6: Είναι παρόμοια με την τεχνική 5. Ωστόσο, τα υπόλοιπα r_{Zi} χρησιμοποιούνται στη θέση των r_{Di}.

Τεχνική 7: Προσαρμόζουμε το μοντέλο Α στο πλήρη σύνολο δεδομένων και υπολογίζουμε τα r_{Li}. Τότε το αρχικό υποσύνολο αποτελείται από τις παρατηρήσεις των οποίων τα οι τιμές των r_{Li} κυμαίνονται ανάμεσα στο -1 και στο 6.

Βήμα 2^{o}: Πρόσθεση παρατήρησης

Το επόμενο βήμα είναι να διαλέξουμε μία παρατήρηση για να συμπεριληφθεί στο $S_*^{(m)}$. Για παράδειγμα, εφαρμόζουμε την τεχνική 1 από παραπάνω. Προσαρμόζουμε το μοντέλο Α στο αρχικό υποσύνολο $S_*^{(m)}$. Τότε οι εκτιμήσεις των στατιστικών που μας ενδιαφέρουν, όπως η εκτίμηση της παραμέτρου $\hat{\beta}$ και της διακύμανσης σ_e^2, που συμβολίζονται με $\hat{\beta}^{(m)}$ και $\sigma_e^{2\,(m)}$, καταγράφονται αντίστοιχα. Έπειτα, η εκτίμηση της παραμέτρου $\hat{\beta}^{(m)}$ χρησιμοποιείται για τον υπολογισμό των καινούριων r_{Di} για το πλήρες σύνολο δεδομένων και στη συνέχεια ταξινομούμε τα δεδομένα με βάση τις τιμές των καινούριων r_{Di}^2. Συνεπώς, σχηματίζεται ένα καινούριο υποσύνολο $S_*^{(m+1)}$ μεγέθους m+1, διαλέγοντας τις m+1 παρατηρήσεις με τις μικρότερες τιμές στα r_{Di}^2.

Αυτή η διαδικασία επαναλαμβάνεται μέχρι όλες οι παρατηρήσεις να είναι στο υποσύνολο $S_*^{(n)}$, όπου N είναι το μέγεθος του δείγματος από τα πλήρη δεδομένα. Παρόμοια βήματα ακολουθούνται όταν τα r_{Li} ή τα r_{Zi} χρησιμοποιούνται αντί των r_{Di}.

Βήμα 3^{o}: Παρακολούθηση της αναζήτησης

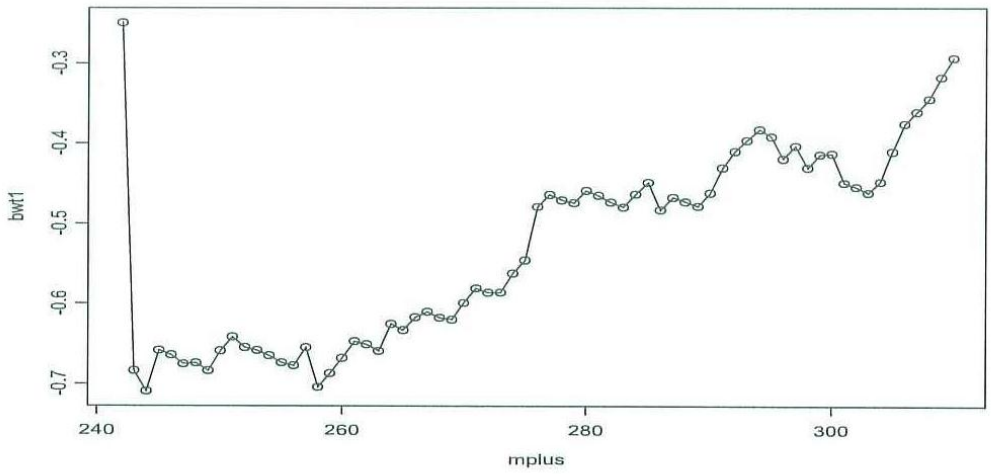
Οποιεσδήποτε αλλαγές στο $\hat{\beta}^{(l)}$ και $\sigma_e^{2\,(l)}$ για l=m,m+1,…,N εντοπίζονται κάνοντας τη γραφική παράσταση του $\hat{\beta}^{(l)}$ ως προς το δείκτη l δίνοντας ένα διάγραμμα προόδου (progression plot).

Παράδειγμα 5.1

Έχουμε ένα σύνολο δεδομένων που αφορά 310 ασθενείς που πάσχουν από καρκίνο του προστάτη (Andrews and Herzberg, 1985). Τα δεδομένα περιέχουν ασθενείς κάτω των 75 χρονών και έχουμε τις εξής μεταβλητές.

1. το είδος της θεραπείας, Treatment
2. τον δείκτη βάρους, Wt
3. την αιμοσφαιρίνη, Haemo
4. το μέγεθος της πρωτογενούς βλάβης, Size
5. το στάδιο του όγκου, Tumour
6. το ιστορικό καρδιαγγειακής νόσου, History

Στο Σχήμα 5.1 έχουμε το progression plot της μεταβλητής ΄βάρος΄ για τα παραπάνω δεδομένα. Από το σχήμα, μπορούμε να αναγνωρίσουμε την παρατήρηση που προκαλεί μεγάλη αλλαγή και συνεπώς μπορούμε να προσδιορίσουμε τον αντίστοιχο ασθενή.



Σχήμα 5.1

Progression plot στον παράγοντα wt ασθενών με καρκίνο του προστάτη με την FS μέθοδο

5.2.6 Added variable plot (πρόσθετων μεταβλητών γράφημα)

Το γράφημα πρόσθετων μεταβλητών είναι χρήσιμο για την εξέταση της επίδρασης της συμμεταβλητής σε μοντέλα παλινδρόμησης. Το γράφημα παρέχει πληροφορίες σχετικά με την ένταξη της συμμεταβλητή και είναι χρήσιμο για την αναγνώριση των σημείων επιρροής για την εκτίμηση των παραμέτρων. Οι Chen και Wang (1991) πρότειναν τα γραφήματα πρόσθετων μεταβλητών για το μοντέλο αναλογικής διακινδύνευσης του Cox που προέρχονται από το μειωμένο μοντέλο, δηλαδή το μοντέλο με τις λιγότερες μεταβλητές. Ο Hall (1996), με τη σειρά του, τα επέκτεινε για γραφήματα που προέρχονται από το πλήρες μοντέλο. Χρησιμοποίησε την προσέγγιση των O'Hara Hines και Carter (1993), η οποία οδηγεί σε γραφήματα μα ιδιότητες παρόμοιες με αυτές των added variable plots για τα γραμμικά μοντέλα. Αυτά τα γραφήματα εφαρμόζονται επίσης σε μοντέλα με χρονικά εξαρτώμενες μεταβλητές.

Για να εξάγουμε το γράφημα πρόσθετων μεταβλητών, το οποίο προτάθηκε από τον Hall (1996), για το μοντέλο αναλογικής διακινδύνευσης του Cox θεωρούμε τη συνάρτησης διακινδύνευσης

$$h(t|x_{(j)},x_j) = h_0(t)\exp(x_{(j)}^T\beta_{(j)} + x_j\beta_j) \qquad (5.2)$$

όπου x_j είναι η μεταβλητή που μας ενδιαφέρει, δηλαδή εκείνη της οποίας θέλουμε να εξετάσουμε την επίδραση στο μοντέλο του Cox, $\beta_{(j)}$ και $x_{(j)}$ είναι το β και το x, αντίστοιχα, έχοντας αφαιρέσει την j-οστή παρατήρηση (Lindkvist, 2000). Η εκτίμηση του $\beta=(\beta_{(j)}^T , \beta_j)^T$ μπορεί να γραφεί ως

$$\hat{\beta} = (R^{*T} R^*)^{-1}R^{*T}z^* \text{ όπου } R^* = W^{1/2}R \text{ και } z^* = W^{1/2}z,\ W=\text{diag}(\mathbf{p})$$

με $p=(p_{[1]}^T,p_{[2]}^T,\ldots, p_{[N]}^T)^T$, τα στοιχεία p_{ik} αντιστοιχούν στα p_i της παρατήρησης k, όπου p_{ik} είναι η πιθανότητα διακοπής της λειτουργίας της μονάδας k

$$p_{ik} = \frac{\exp(x_k^T\beta)}{\sum_{l\in R_i}\exp(x_l^T\beta)}$$

η οποία χρησιμοποιείται και στη σχέση (2.3). Ο πίνακας συνδιακύμανσης για το $\hat{\beta}$ είναι $(R^{*T} R^*)^{-1}$. Ορίζω το $R^*_{(j)}$ να είναι το R^* με την αφαίρεση της j στήλης, το

$r^*_{.j}$ είναι η j στήλη του R^*,

$$H^* = R^* (R^{*T} R^*)^{-1} R^{*T} \quad \text{και} \quad H^*_{(j)} = R^*_{(j)} (R^{*T}_{(j)} R^*_{(j)})^{-1} R^{*T}_{(j)}.$$

Το γράφημα πρόσθετων μεταβλητών για το x_j της σχέσης (5.2), είναι το γράφημα των $\tilde{z}^* = (I - H^*_{(j)})z$ ως προς $\tilde{r}^*_{.j} = (I - H^*_{(j)}) r^*_{.j}$.

Κάθε σημείο στο γράφημα αντιστοιχεί σε μία παρατήρηση σε ένα συγκεκριμένο σύνολο κινδύνου σε έναν συγκεκριμένο χρόνο αποτυχίας.

Το παραπάνω γράφημα εντοπίζει τα σημεία επιρροής και από εκεί βγάζουμε συμπεράσματα για το αν θα συμπεριλάβουμε κάποιες μεταβλητές ή όχι.

Παρόλο που το γράφημα πρόσθετων μεταβλητών είναι πολύ σημαντικό, δεν υπάρχει έτοιμος κώδικας στην R για την απεικόνιση του.

ΚΕΦΑΛΑΙΟ 6

ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ

6.1 Εκτίμηση των συντελεστών παλινδρόμησης β

Τα δεδομένα αφορούν 51 ασθενείς που πάσχουν από οξεία μυελοβλαστική λευχαιμία και υποβάλλονται σε μία θεραπεία, στης οποίας το τέλος βλέπουμε αν έχουν ανταποκριθεί ή όχι (Lee, 1980). Έχουμε τις εξής έξι μεταβλητές :

1. την ηλικία του ασθενή, Age,
2. το ποσοστό επίστρωσης των βλαστοκυττάρων, Smear
3. το ποσοστό των κυττάρων στο μυελό των οστών, Infiltrate
4. το ποσοστό των κυττάρων που προήλθαν από το μυελό των οστών, Index
5. τα απόλυτα βλαστοκύτταρα, Blasts
6. τη θερμοκρασία του σώματος, Temperature

	Age	Smear	Infiltrate	Index	Blasts	Temperature	Time	Status
1	20	78	39	7	0.6	990	18	0
2	25	64	61	16	35.0	1030	31	1
3	26	61	55	12	7.5	982	31	0
4	26	64	64	16	21.0	1000	31	0
5	27	95	95	6	7.5	980	36	0

Απόσπασμα δεδομένων με πηγή αναφοράς : Lee (1980).

Επίσης, έχουμε τον χρόνο επιβίωσης του ασθενή, Time, και τη μεταβλητή Status, η οποία δείχνει αν η παρατήρηση είναι αποκομμένη (status=1) ή όχι (status=0). Στη συνέχεια, θέλουμε να εξετάσουμε το πόσο σημαντικές είναι οι έξι παραπάνω μεταβλητές στον

χρόνο επιβίωσης. Για αυτόν το λόγο θα χρησιμοποιήσουμε το μοντέλο αναλογικής διακινδύνευσης του Cox (Zhou).

Συνάρτηση coxph

Η συνάρτηση coxph είναι ίσως η βασικότερη εντολή της R όσον αφορά το μοντέλο αναλογικού κινδύνου του Cox, αφού μέσω αυτής της εντολής γίνεται η προσαρμογή του μοντέλου στα δεδομένα. Η γενική της μορφή είναι η ακόλουθη:

coxph(formula,...., method="efron",....)

όπου στο formula βάζουμε στα αριστερά την απόκριση που πρέπει να είναι ένα 'αντικείμενο' που να είναι αποτέλεσμα της Surv ( ). Μετά την απόκριση τοποθετείται το σύμβολο ~ και στα δεξιά του ~ τοποθετούνται οι υπόλοιπες μεταβλητές (ως προς τις οποίες θα γίνει η μοντελοποίηση).

Στο method βάζουμε τη μέθοδο που θέλουμε να χρησιμοποιηθεί όταν υπάρχουν ισόπαλες παρατηρήσεις. Πιθανές επιλογές είναι οι μέθοδοι 'efron' που είναι και η προεπιλεγμένη, η 'breslow', η οποία ορίζεται στην Παράγραφο 2.3.1.1 και η 'exact'.

Συνάρτηση Surv

Η συνάρτηση Surv δημιουργεί ένα αντικείμενο επιβίωσης. Η γενική της μορφή είναι η ακόλουθη:

Surv(time,event)

Όπου στο time βάζουμε το χρόνο μέχρι την πραγματοποίηση του γεγονότος ή τον αποκομμένο χρόνο και στο event βάζουμε μία μεταβλητή με τιμή 0, αν η παρατήρηση είναι αποκομμένη και τιμή 1, αν όχι.

Για να γίνει αυτό στο στατιστικό πακέτο R, θα πρέπει να χρησιμοποιήσουμε τη βιβλιοθήκη survival, μέσα στην οποία βρίσκονται οι συναρτήσεις που χρειαζόμαστε. Επομένως, εφαρμόζουμε το μοντέλο αναλογικής διακινδύνευσης του Cox, με χρήση της εντολής coxph, στα δεδομένα μας με τις έξι επεξηγηματικές μεταβλητές και παίρνουμε τον παρακάτω Πίνακα :

n= 51, number of events= 45

	Coef	exp(coef)	se(coef)	z	Pr(>|z|)
Ηλικία	0.03198	1.03249	0.01035	3.090	0.0020 **
Smear	0.01356	1.01365	0.01528	0.888	0.3747
Infiltrate	-0.01709	0.98306	0.01232	-1.387	0.1654
Index	-0.07222	0.93032	0.03926	-1.840	0.0658 .
Βλαστοκύτταρα	-0.01685	0.98329	0.02268	-0.743	0.4573
Θερμοκρασία	0.02212	1.02236	0.01353	1.635	0.1021

	exp(coef)	exp(-coef)	lower .95	upper .95
Ηλικία	1.0325	0.9685	1.0118	1.054
Smear	1.0137	0.9865	0.9838	1.044
Infiltrate	0.9831	1.0172	0.9596	1.007
Index	0.9303	1.0749	0.8614	1.005
Βλαστοκύτταρα	0.9833	1.0170	0.9405	1.028
Θερμοκρασία	1.0224	0.9781	0.9956	1.050

Rsquare =	0.328	(max possible= 0.996)	
Likelihood ratio test =	20.26	on 6 df,	p=0.002486
Wald test =	19.31	on 6 df,	p=0.003676
Score (logrank) test =	20.88	on 6 df,	p=0.001929

Πίνακας 6.1

Από τον Πίνακα 6.1, από την p-τιμή της μεταβλητής Age, που είναι ίση με 0.0020, συμπεραίνουμε ότι η μεταβλητή Age είναι η πιο στατιστικά σημαντική. Επομένως, το επόμενο βήμα είναι να προσαρμόσουμε το μοντέλο του Cox μόνο για τη μεταβλητή Age και παίρνουμε τα παρακάτω Πίνακα:

n= 51, number of events= 45

	Coef	exp(coef)	se(coef)	z	Pr(>|z|)
Ηλικία	0.032397	1.032927	0.009521	3.403	0.000667 ***

	exp(coef)	exp(-coef)	lower .95	upper .95
Ηλικία	1.033	0.9681	1.014	1.052

Likelihood ratio test =	11.85	on 1 df,	p=0.000577
Wald test =	11.58	on 1 df,	p=0.0006675
Score (logrank) test =	12.29	on 1 df,	p=0.0004562

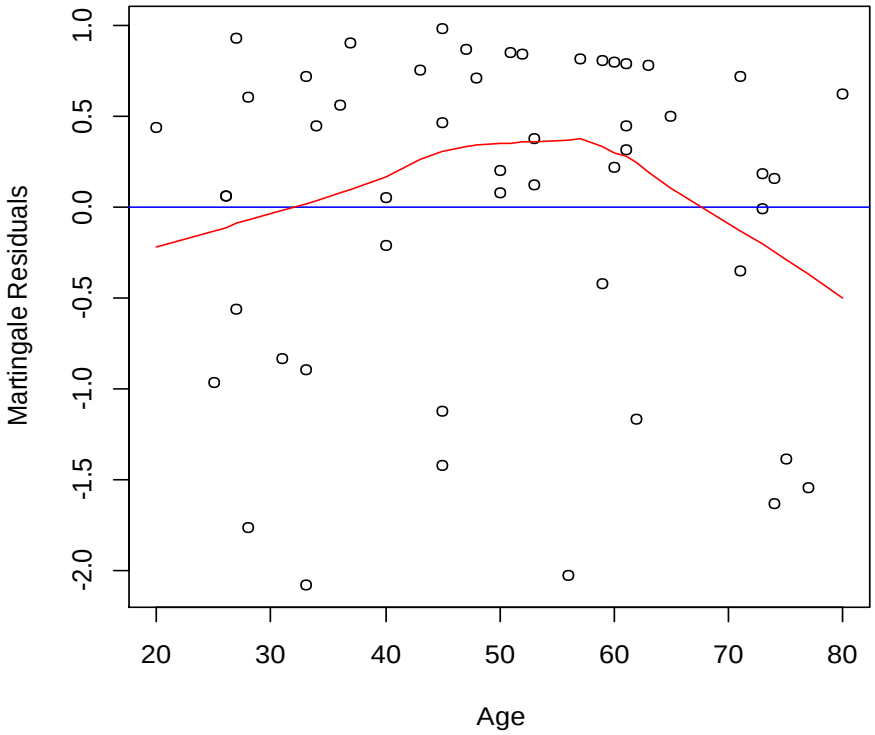
Πίνακας 6.2

Άρα ο συντελεστής όπως βλέπουμε από τον Πίνακα 6.2 είναι 0.032397 που σημαίνει ότι όταν η μεταβλητή αυξάνεται κατά μία μονάδα τότε η συνάρτηση διακινδύνευσης αυξάνεται κατά 3%.

6.2 Υπόλοιπα Martingale

Τα υπόλοιπα Martingale είναι χρήσιμα στον προσδιορισμό της καλύτερης συναρτησιακής μορφής των επεξηγηματικών μεταβλητών. Είναι μία μικρή τροποποίηση των Cox-Snell υπολοίπων. Όπως αναφέραμε και προηγουμένως στην Παράγραφο 3.6 υπολογίζονται από τον τύπο $r_{Mi} = \delta_i - r_{ci}$ όπου r_{ci} είναι τα Cox-Snell υπόλοιπα. Έστω το διάνυσμα της μεταβλητής x το οποίο χωρίζεται στο $x^{(2)}$, για το οποίο γνωρίζουμε τη συναρτησιακή μορφή και σε μία ενιαία συνεχή μεταβλητή $x^{(1)}$, για την οποία δεν γνωρίζουμε τη συναρτησιακή της μορφή (Tableman, 2008). Υποθέτουμε ότι το $x^{(1)}$ είναι ανεξάρτητο του $x^{(2)}$. Έστω τώρα ότι η g() είναι η καλύτερη συνάρτηση του $x^{(1)}$, για να εξηγήσει την επίδραση της μεταβλητής. Για να βρούμε την g(), προσαρμόζουμε το μοντέλο του Cox στα δεδομένα που βασίζονται στο $x^{(2)}$ και υπολογίζουμε τα martingale υπόλοιπα. Έπειτα, κάνουμε την γραφική παράσταση αυτών των υπολοίπων ως προς τις τιμές $x^{(1)}_i$ για i=1,…,N. Συνήθως, χρησιμοποιείται μια ομαλοποιημένη προσαρμογή του scatter plot. Η ομαλοποιημένη καμπύλη δίνει κάποια ένδειξη για τη συνάρτηση g(). Αν η γραφική παράσταση είναι γραμμική τότε δε χρειάζεται κάποιος μετασχηματισμός του $x^{(1)}$. Αν φαίνεται να υπάρχει κάτι διαφορετικό στο γράφημα, τότε ενδείκνυται ένας μετασχηματισμός της μεταβλητής.

Αρχικά κατασκευάζουμε το γράφημα των υπολοίπων martingale ως προς την ηλικία το οποίο φαίνεται στο Σχήμα 6.1. Από τη γραμμή εξομάλυνσης, παρατηρούμε ότι ίσως θα πρέπει να διαχωρίσουμε τους ασθενείς με ηλικία άνω και κάτω των 50 και ότι ίσως πρέπει να συμπεριληφθεί και η μεταβλητή Age^2 στο μοντέλο.



Σχήμα 6.1

Τα υπόλοιπα martingale ως προς την ηλικία με την γραμμή εξομάλυνσης

Έτσι ξαναπροσαρμόζουμε το μοντέλο του Cox με τις μεταβλητές Age και Age^2 και πρόκύπτουν τα παρακάτω αποτελέσματα :

n= 51, number of events= 45

	Coef	exp(coef)	se(coef)	z	Pr(>|z|)
Ηλικία	0.1379612	1.1479310	0.0673332	2.049	0.0405 *
$(Ηλικία)^2$	-0.0010158	0.9989847	0.0006378	-1.593	0.1113

	exp(coef)	exp(-coef)	lower .95	upper .95
Ηλικία	1.148	0.8711	1.0060	1.31
$(Ηλικία)^2$	0.999	1.0010	0.9977	1.00

Likelihood ratio test =	14.51	on 2 df,	p=0.0007073
Wald test =	11.79	on 2 df,	p=0.002748
Score (logrank) test =	13.09	on 2 df,	p=0.001439

Πίνακας 6.3

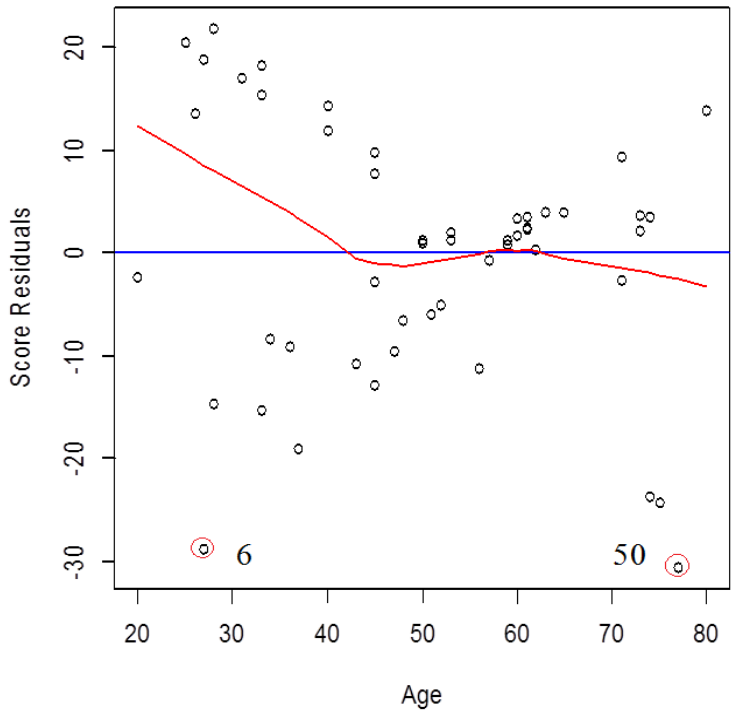
Από τον Πίνακα 6.3 βλέπουμε ότι τελικά η μεταβλητή Age^2, αφού η p-τιμή της είναι 0.1113, δεν είναι τελικά σημαντική.

6.3 Υπόλοιπα Score

Τα υπόλοιπα Score υπολογίζουν τη διαφορά ανάμεσα στην τιμή της δοσμένης μεταβλητής και της μέσης τιμής της στο σύνολο κινδύνου. Το γράφημα τους ως προς την μεταβλητή που μας ενδιαφέρει, προσδιορίζει ως σημεία επιρροής τις παρατηρήσεις που το αντίστοιχο υπόλοιπο Score διαφέρει από τον μέσο όρο του δείγματος σε μεγάλο βαθμό. Τα υπόλοιπα Score δίνονται από τον τύπο :

$r_{Sji} = r_{Pji} + \exp(\hat{\beta}'x_i) \sum_{t_r \le t_i} \frac{(\hat{a}_{jr} - x_{ji})\delta_r)}{\sum_{l \in R(t_r)} \exp(\hat{\beta}'x_l)}$ όπου r_{Pji} είναι τα υπόλοιπα Schoenfeld, τα οποία δίνονται από τη σχέση (3.1).

Κατασκευάζουμε το γράφημα των υπολοίπων Score ως προς την ηλικία το οποίο φαίνεται στο Σχήμα 6.2 και από το οποίο παρατηρούμε ότι ίσως να υπάρχουν κάποια σημεία επιρροής, τα οποία πιθανόν να είναι τα δύο πιο αρνητικά υπόλοιπα, τα οποία είναι αντιστοιχούν στην $6^η$, η οποία είχε πολύ μικρό χρόνο επιβίωσης και στην $50^η$ παρατήρηση, η οποία είχε αντίστοιχα σχετικά μεγάλο.



Σχήμα 6.2

Τα υπόλοιπα Score ως προς την ηλικία με την γραμμή εξομάλυνσης

6.4 Υπόλοιπα Deviance

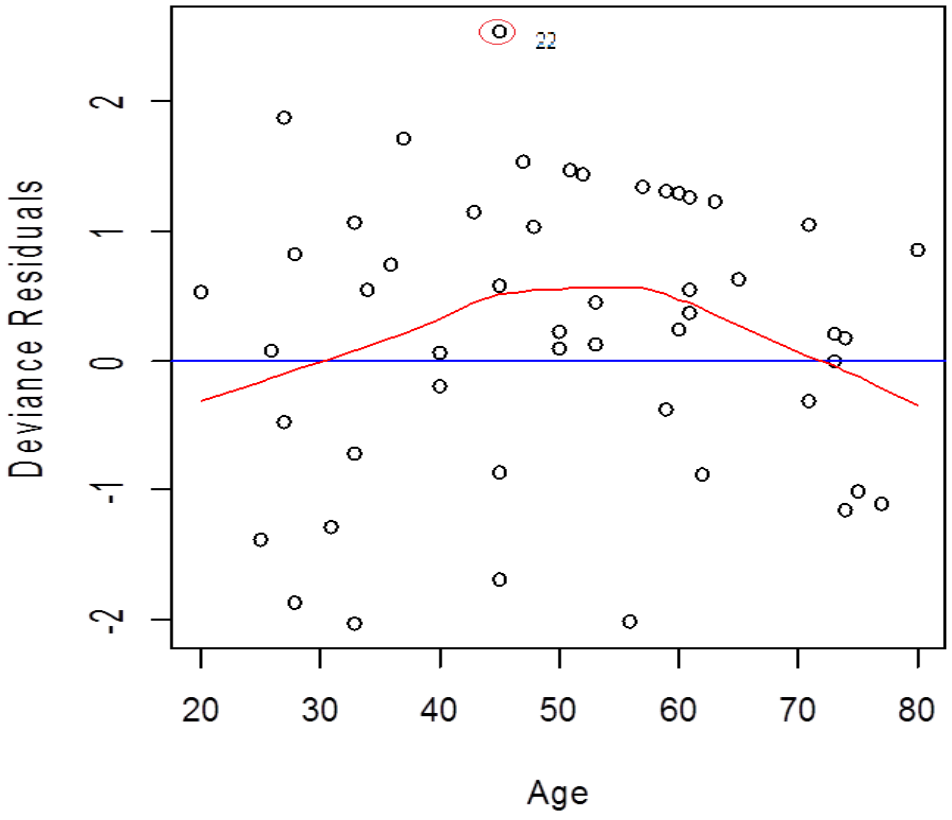
Ένα σημαντικό μειονέκτημα των υπολοίπων martingale είναι η ασύμμετρη κατανομή τους και γι' αυτόν τον λόγο τα μετασχηματίζουμε ώστε να ξεπεράσουμε αυτό το πρόβλημα. Έτσι προκύπτουν τα υπόλοιπα deviance, τα οποία ορίζουμε στην Παράγραφο 3.8 και δίνονται από τον τύπο :

$$r_{Di} = \text{sgn}(r_{mi}) \, [-2[r_{mi} + \delta_i \log(\delta_i - r_{mi})]]^{1/2}, \qquad 1 \le i \le N$$

όπου r_{mi} είναι τα υπόλοιπα martingale και $\text{sgn}(r_{mi}) = \begin{cases} 1, & \text{για } r_{mi} > 0 \\ -1, & \text{για } r_{mi} < 0 \end{cases}$.

Από το Σχήμα 6.3 βλέπουμε ότι τα υπόλοιπα είναι συμμετρικά κατανεμημένα γύρω από το μηδέν. Πιθανά άτυπα σημεία αντιστοιχούν σε μεγάλες κατ' απόλυτη τιμή τιμές των υπολοίπων deviance. Βλέποντας το σχήμα, αυτό είναι το υπόλοιπο που αντιστοιχεί στην

$22^{η}$ παρατήρηση και είναι ασθενής με μηδενικό χρόνο επιβίωσης. Επειδή, υπάρχει μόνο μια πιθανή άτυπη τιμή δεν προκαλείται ανησυχία σχετικά με την επάρκεια του μοντέλου.



Σχήμα 6.3

Τα υπόλοιπα deviance ως προς την ηλικία με την γραμμή εξομάλυνσης

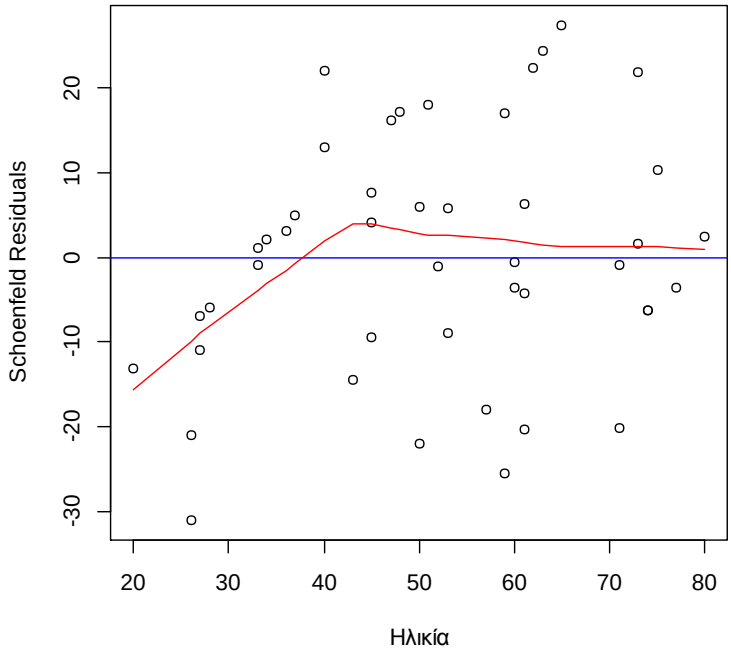
6.5 Υπόλοιπα Schoenfeld

Τα υπόλοιπα Schoenfeld χρησιμοποιούνται για τον έλεγχο της υπόθεσης της αναλογικής διακινδύνευσης (PH υπόθεση). Ορίζονται σε κάθε χρόνο αποτυχίας $t_{(i)}$ ως $r_{Pji} = \delta_i\{ x_{ji} - \hat{a}_{ji}\}$, όπου x_{ji} είναι η τιμή της συμμεταβλητής x_j για το i άτομο (δηλαδή η τιμή x_{ji}), $\hat{a}_{ji}=\frac{\sum_{i\in R_i} x_j \exp(\hat{\beta}' x_j)}{\sum_{i\in R_i} \exp(\hat{\beta}' x_j)}$ i=1,…,N, j=1,…,p (3.2) είναι η αναμενόμενη τιμή της

συμμεταβλητής για τα άτομα που βρίσκονται σε ρίσκο στο χρόνο ti, N το πλήθος των ατόμων και p το πλήθος των μεταβλητών και $\delta_i=0$, αν t_i αποκομμένος και $\delta_i=1$, αν t_i πλήρης χρόνος.

Τα υπόλοιπα Schoenfeld υπολογίζονται μόνο για πλήρεις παρατηρήσεις. Αν η υπόθεση της αναλογικής διακινδύνευσης ισχύει, τότε τα υπόλοιπα Schoenfeld ως προς την ηλικία θα πρέπει να έχουν τυχαία μορφή. Διαφορετικά, το γράφημα θα δείχνει πολύ μεγάλα υπόλοιπα για κάποιες ηλικίες.

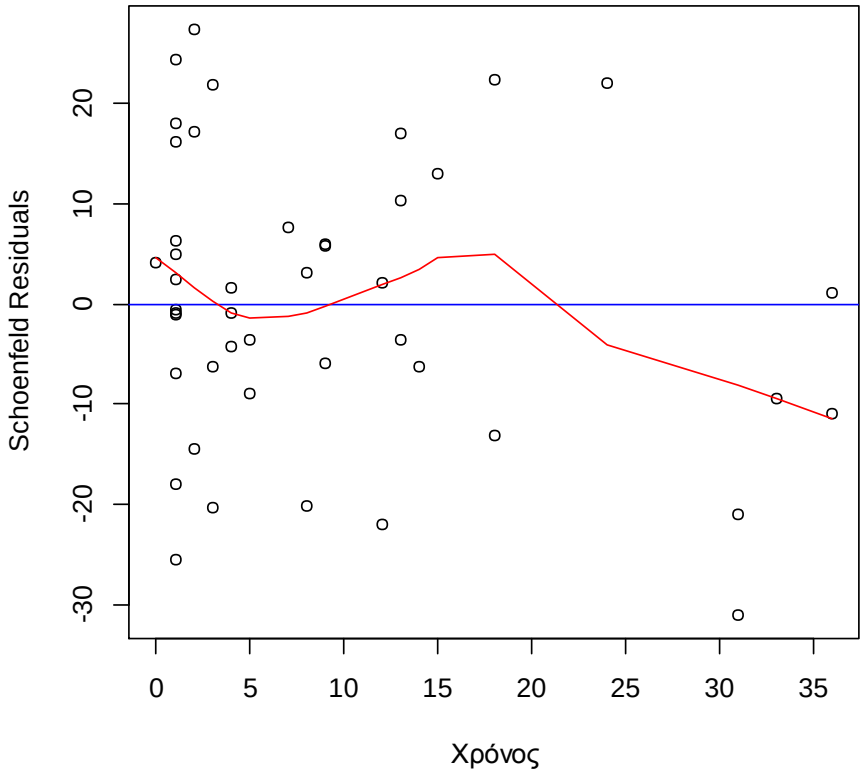
Από το Σχήμα 6.4 βλέπουμε ότι όσο μεγαλώνει η ηλικία οι τιμές των υπολοίπων Schoenfeld γίνονται όλο και πιο αρνητικές.



Σχήμα 6.4

Τα υπόλοιπα Schoenfeld ως προς την ηλικία μαζί με την γραμμή εξομάλυνσης

Από το Σχήμα 6.5 (Therneau, Grambsch and Fleming, 1990) παρατηρούμε μία μικρή τάση των ασθενών με μικρούς χρόνους επιβίωσης, να έχουν αρνητικά υπόλοιπα. Αυτό δείχνει ότι το μοντέλο υπερεκτιμά την πιθανότητα, οι ασθενείς να πεθάνουν μετά από μικρό χρονικό διάστημα. Επίσης, βλέπουμε ότι όσο μεγαλώνει ο χρόνος επιβίωσης οι τιμές των υπολοίπων Schoenfeld γίνονται όλο και πιο αρνητικές και ότι αυτές οι τιμές αντιστοιχούν σε άτομα μικρότερης ηλικίας, επομένως ένας διαχωρισμός των ασθενών κατά ηλικία θα ήταν χρήσιμος, όπως συμπεράναμε και από τα martingale υπόλοιπα.



Σχήμα 6.5

Τα υπόλοιπα Schoenfeld ως προς τον χρόνο μαζί με την γραμμή εξομάλυνσης

6.5.1 Τυποποιημένα υπόλοιπα Schoenfeld

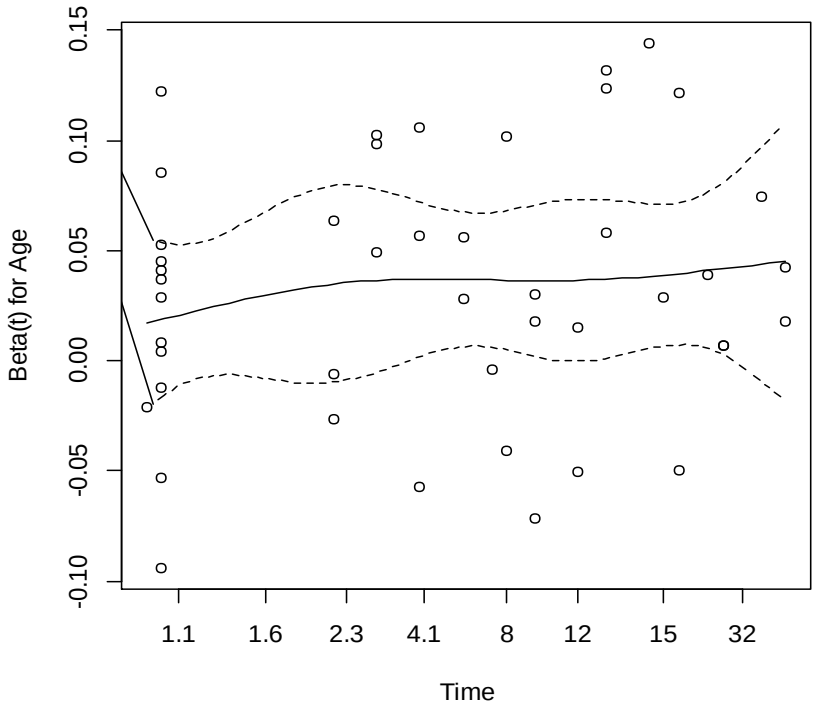
Συμβολίζουμε τα τυποποιημένα Schoenfeld υπόλοιπα με r^*_{Pji} και είναι τα στοιχεία του διανύσματος ως $r^*_{Pj} = r\ V(\hat{\beta})\ r_{Pj}$. Ως εναλλακτική λύση για την υπόθεση της

αναλογικής διακινδύνευσης οι Grambsch και Therneau (1994) θεώρησαν χρονικά μεταβαλλόμενους συντελεστές $\beta(t) = \beta + \theta g(t)$, όπου g(t) είναι μία συνάρτηση του χρόνου. Δεδομένου ότι το g(t) είναι γνωστό, ανέπτυξαν έναν score έλεγχο για την H_0 : θ=0, βασισμένο στην εκτιμήτρια ελαχίστων τετραγώνων του θ, ο οποίος έλεγχος γίνεται με την ελεγχοσυνάρτηση

$$T_j = \frac{\{\sum_j (g_j - \bar{g})\, r^*_{Pj}\}^2}{rI_j \sum_j (g_j - \bar{g})^2}$$

όπου j είναι η μεταβλητή που μας ενδιαφέρει, g_j είναι ο μετασχηματισμός του χρόνου, $\bar{g}$ είναι η μέση τιμή του g, r είναι ο αριθμός των γεγονότων και I_j είναι η παρατηρούμενη πληροφορία για την μεταβλητή j. Η T_j ακολουθεί προσεγγιστικά την X^2 κατανομή με ένα βαθμό ελευθερίας. Στην συνέχεια, έδειξαν ότι το i-οστό τυποποιημένο υπόλοιπο Schoenfeld, τα οποία αναφέρονται στην Παράγραφο 3.4 έχει μέση τιμή περίπου $\theta g(t_j)$, δηλαδή $E(r^*_{Pji}) \cong \theta g(t_i)$. Παρακινούμενοι από αυτά τα αποτελέσματα, ανέπτυξαν επίσης, μία γραφική μέθοδο. Έδειξαν ότι το γράφημα των $\hat{\beta}(t_i)$, το i-οστό τυποποιημένο υπόλοιπο Schoenfeld συν το $\hat{\beta}$ (η εκτιμήτρια μέγιστης πιθανοφάνειας του β), ως προς το χρόνο t_i αποκαλύπτει τη συναρτησιακή μορφή του β(t). Κάτω από την υπόθεση της H_0, περιμένουμε μία σταθερή συνάρτηση ως προς το χρόνο.

Το Σχήμα 6.6 δείχνει ότι ικανοποιείται η υπόθεση της αναλογικής διακινδύνευσης, αφού η γραμμή μπορεί να θεωρηθεί οριζόντια, δηλαδή ο συντελεστής του μοντέλου δε δείχνει εξάρτηση από το χρόνο. Αυτό ενισχύεται και από το ότι ο συντελεστής συσχέτισης του Pearson, ο οποίος εξετάζει αν τα τυποποιημένα υπόλοιπα Schoenfeld σχετίζονται με το χρόνο, είναι χαμηλός (ρ=0.0763). Επίσης από τα αποτελέσματα προκύπτει ότι η τιμή της ελεγχοσυνάρτησης του παραπάνω ελέγχου για την υπόθεση H0: θ=0 είναι T= 0.214 (p-τιμή =0.643), και έτσι πάλι συμπεραίνουμε ότι ικανοποιείται η υπόθεση της αναλογικής διακινδύνευσης.



Σχήμα 6.6

Οι εκτιμήτριες $\hat{\beta}(t)$ ως προς τον χρόνο επιβίωσης, μαζί με την spline smooth και τα $\pm$ 2 διαστήματα εμπιστοσύνης για την spline smooth

6.6 Normal deviate υπόλοιπα

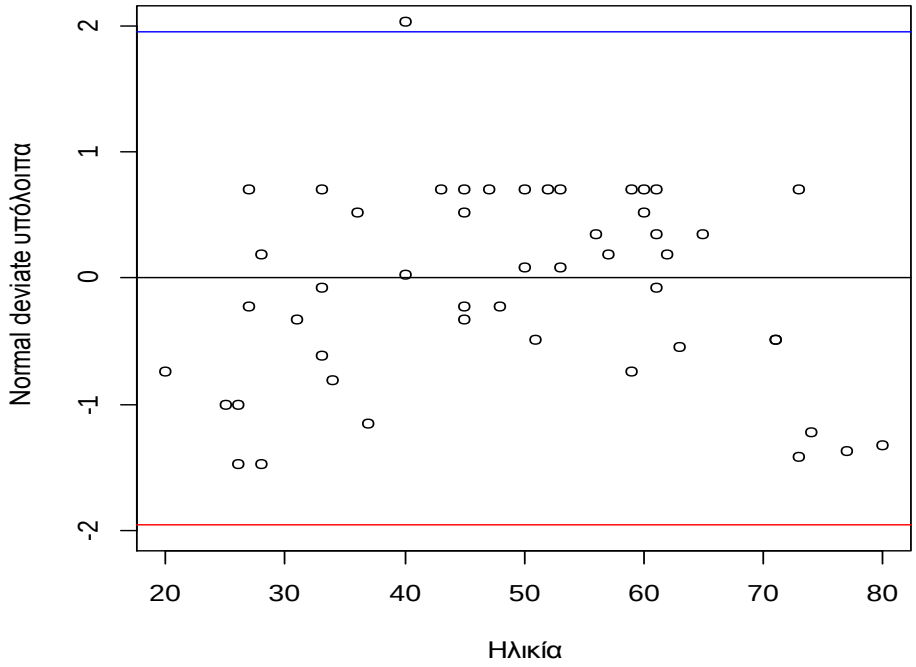
Οι Nardi και Schemper (1999) πρότειναν τη χρήση των log-odds ή των normal deviate υπολοίπων για τον έλεγχο των σημείων επιρροής, τα οποία φαίνεται να έχουν καλύτερες ιδιότητες από τα υπόλοιπα Deviance. Τα normal deviate υπόλοιπα υπολογίζονται με βάση την εκτιμώμενη συνάρτηση επιβίωσης και το χρόνο επιβίωσης, αποκομμένο ή μη. Η εκτίμηση της συνάρτησης επιβίωσης θεωρείται καλή, αν $\hat{S}_i(t_i)=0.5$. Αυτά τα υπόλοιπα δε μετράνε ευθέως τη διαφορά ανάμεσα στον παρατηρούμενο και στον εκτιμώμενο χρόνο επιβίωσης. Αντί γι' αυτό, συγκρίνουν την εκτιμώμενη πιθανότητα επιβίωσης σε χρόνο t (αποκομμένο ή μη) με την τιμή 0.5 .

Τα normal deviate υπόλοιπα ορίζονται ως ο probit μετασχηματισμός του $\hat{S}_i(t_i)$:

$$z_i = \begin{cases} \Phi^{-1}\{\hat{S}_i(t_i)\}, \text{ όπου } t_i \text{ πλήρης χρόνος} \\ \Phi^{-1}\{\frac{\hat{S}_i(t_i^*)}{2}\}, \text{ όπου } t_i^* \text{ αποκομμένος χρόνος} \end{cases}$$

όπου Φ είναι η σωρευτική συνάρτηση κανονικής κατανομής.

Στο Σχήμα 6.8 απεικονίζονται τα normal deviate υπόλοιπα μαζί με τις τρεις οριζόντιες γραμμές. Οι γραμμές που βρίσκονται στο ± 1.96 υπολογίστηκαν στο Minitab. Από το γράφημα συμπεραίνουμε ότι ένα πιθανό σημείο επιρροής είναι εκείνο το οποίο αντιστοιχεί στην $17^{η}$ παρατήρηση και βρίσκεται πάνω από την οριζόντια γραμμή 1.96 και το οποίο κοιτάζοντας τα δεδομένα μας, βλέπουμε ότι έχει σχετικά μεγάλο χρόνο επιβίωσης σε σχέση με την ηλικία του.



Σχήμα 6.8

Τα normal deviate υπόλοιπα ως προς την ηλικία

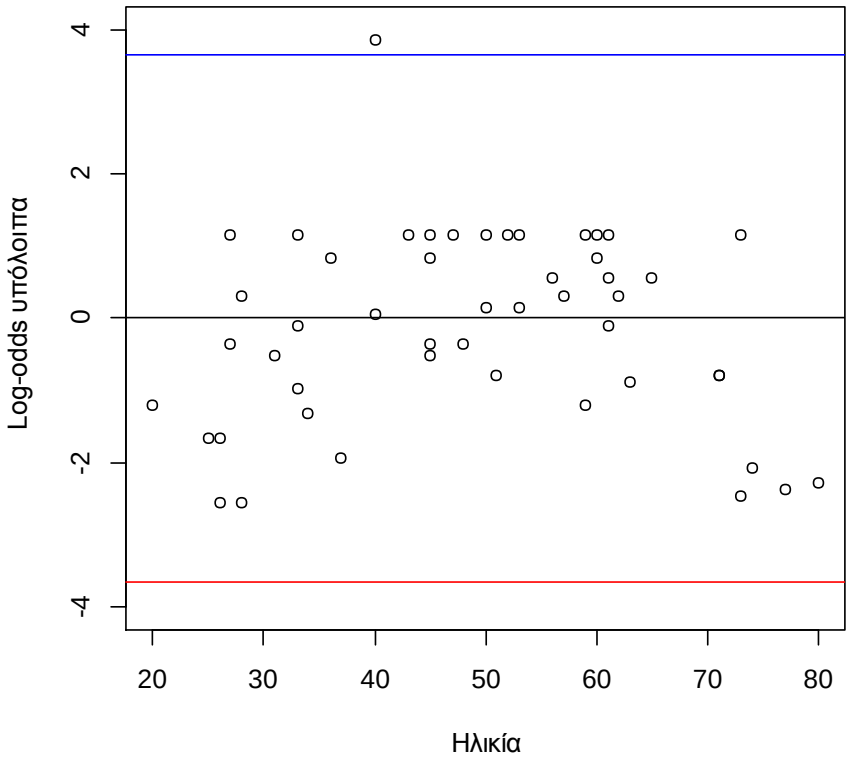
6.7 Log-odds υπόλοιπα

Τα log-odds υπόλοιπα όπως αναφέραμε και στην Παράγραφο 4.2.3.3 υπολογίζονται και αυτά με βάση την εκτιμώμενη συνάρτηση επιβίωσης και το χρόνο επιβίωσης, αποκομμένο ή μη.

Ορίζονται ως ο logit μετασχηματισμός του $\hat{S}_i(t_i)$:

$$l_i = \begin{cases} \log\{\frac{\hat{S}_i(t_i)}{1-\hat{S}_i(t_i)}\}, \text{ όπου } t_i \text{ πλήρης χρόνος} \\ \log\{\frac{\hat{S}_i(t_i^*)}{2-\hat{S}_i(t_i^*)}\}, \text{ όπου } t_i^* \text{ αποκομμένος χρόνος} \end{cases}$$

Στο Σχήμα 6.9 απεικονίζονται τα log-odds υπόλοιπα μαζί με τις τρεις οριζόντιες γραμμές. Οι γραμμές που βρίσκονται στο ± 3.66 υπολογίστηκαν στο Minitab. Παρατηρούμε ότι το γράφημα log-odds υπολοίπων και των normal deviate υπολοίπων μοιάζουν αρκετά. Από το γράφημα καταλήγουμε στα ίδια συμπεράσματα τα οποία προκύπτουν από τα normal deviate υπόλοιπα.

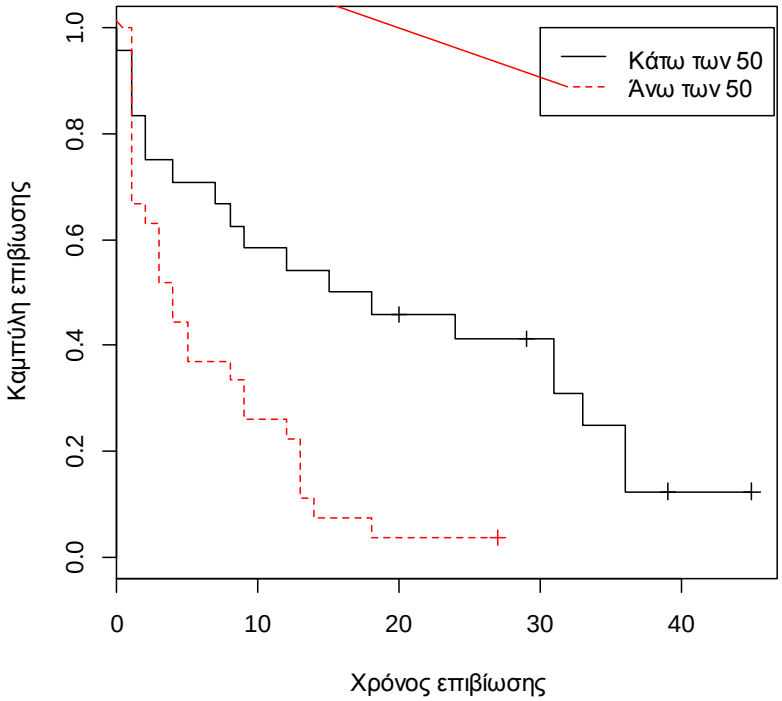


Σχήμα 6.9

Τα Log-odds υπόλοιπα ως προς την ηλικία

6.8 Διαχωρισμός της μεταβλητής Age

Παρατηρώντας τα γράφημα των υπολοίπων martingale και Schoenfeld, καταλήγουμε στο συμπέρασμα ότι θα πρέπει να διαχωρίσουμε τους ασθενείς σε δύο ομάδες, ανάλογα με την ηλικία τους. Η μία θα είναι οι ασθενείς με ηλικία μικρότερη των 50 ετών και η άλλη οι ασθενείς με ηλικία μεγαλύτερη των 50 ετών. Στη συνέχεια κατασκευάζουμε το γράφημα των δύο καμπυλών επιβίωσης, για τις δύο κατηγορίες ηλικιών και προκύπτει το Σχήμα 6.10. Παρατηρούμε ότι όντως υπάρχει διαφορά ανάμεσα στις δύο καμπύλες επιβίωσης.



Σχήμα 6.10

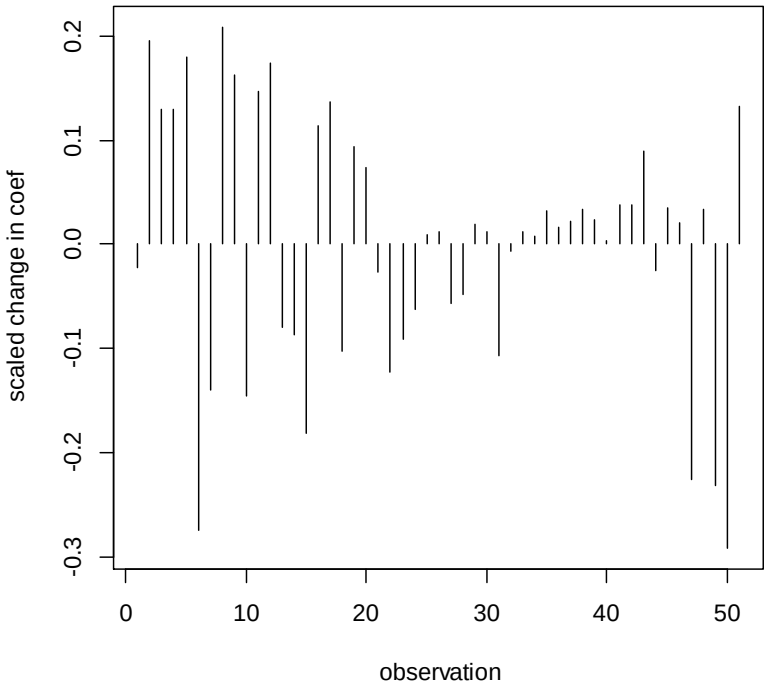
Οι δύο καμπύλες επιβίωσης για ασθενείς κάτω και πάνω των 50 χρονών

6.9 Τα dfbetas για τον προσδιορισμό σημείων επιρροής

Τα dfbetas τα οποία ορίζονται και στην Παράγραφο 5.2.2 είναι χρήσιμα για τον έλεγχο της επιρροής της κάθε παρατήρησης για κάθε παράμετρο ξεχωριστά.

Στην R με την εντολή residuals(datacox,type="dfbetas") υπολογίζεται η αλλαγή σε κάθε συντελεστή/ παλινδρόμησης, όταν κάθε παρατήρηση αφαιρείται από τα δεδομένα. Έπειτα, συγκρίνοντας τις μεγαλύτερες, κατ' απόλυτη τιμή, τιμές των dfbeta με τους συντελεστές παλινδρόμησης καταλήγουμε στο συμπέρασμα αν υπάρχουν σημεία επιρροής ή όχι στα δεδομένα μας.

Από το Σχήμα 6.11, που δείχνει ότι οι περισσότερες αλλαγές στους συντελεστές παλινδρόμησης είναι μέσα στο .3 s.e των συντελεστών, συμπεραίνουμε ότι δεν υπάρχουν σημεία επιρροής στα δεδομένα μας.



Σχήμα 6.11

Τα dfbetas για την ανίχνευση των σημείων επιρροής

6.10 Συμπεράσματα

Αρχικά προσαρμόζουμε το μοντέλο αναλογικής διακινδύνευσης του Cox στα δεδομένα μας με όλες τις μεταβλητές. Από τα αποτελέσματα που προκύπτουν βλέπουμε ότι σημαντική είναι μονάχα η μεταβλητή Age και ξαναπροσαρμόζουμε στα δεδομένα μας μόνο με εκείνη την μεταβλητή. Στην συνέχεια από το γράφημα των υπολοίπων martingale παρατηρούμε ότι θα πρέπει να διαχωρίσουμε τους ασθενείς με ηλικία άνω και κάτω των 50 και ότι ίσως πρέπει να συμπεριληφθεί και η μεταβλητή Age^2 στο μοντέλο. Έτσι ξαναπροσαρμόζουμε το μοντέλο του Cox με τις μεταβλητές Age και Age^2 και βλέπουμε ότι τελικά η μεταβλητή Age^2 δεν είναι τελικά σημαντική. Κατασκευάζουμε το γράφημα των υπολοίπων Score ως προς την ηλικία και παρατηρούμε ότι ίσως να υπάρχουν κάποια σημεία επιρροής, τα οποία πιθανόν να είναι τα δύο πιο αρνητικά υπόλοιπα, τα οποία είναι αντιστοιχούν στην $6^η$, η οποία είχε πολύ

μικρό χρόνο επιβίωσης και στην $50^{η}$ παρατήρηση, η οποία είχε αντίστοιχα σχετικά μεγάλο. Από τα υπόλοιπα deviance βλέπουμε ότι η $22^{η}$ παρατήρηση είναι άτυπο σημείο και έχει μηδενικό χρόνο επιβίωσης και από τα log-odds και τα normal deviate η $17^{η}$ παρατήρηση της οποίας ο ασθενής έχει σχετικά μεγάλο χρόνο επιβίωσης σε σχέση με την ηλικία του, είναι πιθανό σημείο επιρροής. Τέλος, από τα υπόλοιπα Schoenfeld και τα τυποποιημένα υπόλοιπα Schoenfeld συμπεραίνουμε ότι ικανοποιείται η υπόθεση της αναλογικής διακινδύνευσης.

Μετά από την παραπάνω στατιστική ανάλυση, βλέπουμε ότι στην R υπάρχουν εύχρηστες διαγνωστικές τεχνικές όπως τα υπόλοιπα deviance για την αναγνώριση άτυπων σημείων, τα υπόλοιπα Schoenfeld για την υπόθεση της αναλογικής διακινδύνευσης. Παρόλα αυτά δεν υπάρχει έτοιμος κώδικας για μία πολύ αποκαλυπτική διαγνωστική τεχνική για την ανίχνευση σημείων επιρροής, το γράφημα πρόσθετων μεταβλητών, το οποίο μας προσφέρει αρκετές πληροφορίες και είναι δύσκολο να κατασκευαστεί με εντολές.

ΠΑΡΑΡΤΗΜΑ Α

Προγράμματα που χρησιμοποιούνται στην R

1)Πρόγραμμα για την Παράγραφο 6.1

```
#Προσαρμογή του μοντέλου του Cox στα δεδομένα με όλες τις μεταβλητές

library(survival)
data<-read.table("C://Users/admin/Desktop/amldata.txt")
names(data)<-
c('Age','Smear','Infiltrate','Index','Blasts','Temperature','Response','Time','Status')
time<-data [,'Time']
status<-1-data [,"Status"]  #επειδή ο αποκομμένος πρέπει να είναι 0
attach(data)  #για  άμεσα προσβάσιμες  πληροφορίες δεδομένων
datacox<-coxph(Surv(time,status)~Age+Smear+Infiltrate+Index+Blasts+Temperature)
summary(datacox)

#Προσαρμογή του μοντέλου του Cox μόνο με την στατιστικά σημαντική μεταβλητή Age

dataagecox<-coxph(Surv(time,status)~Age)
summary(dataagecox)
```

2)Πρόγραμμα για την Παράγραφο 6.2

```
#Υπολογισμός των Martingale Υπολοίπων και η γραφική τους παράσταση

Age<-data [,'Age']
dataagemartin<-residuals(dataagecox,type=c("martingale"))
plot(Age,dataagemartin,ylab="Martingale Residuals")
abline(h=0,col=4)
lines(lowess(Age,dataagemartin),col=2) #γραμμή εξομάλυνσης

#Προσαρμογή του μοντέλου του Cox μόνο με τις μεταβλητές Age και Age^2

y=Age^2
dataagecox2<-coxph(Surv(time,status)~Age+y)
summary(dataagecox2)
```

3)Πρόγραμμα για την Παράγραφο 6.3

```
#Υπολογισμός των Score Υπολοίπων και η γραφική τους παράσταση

dataagescore<-residuals(dataagecox,type=c("score"))
plot(Age,dataagescore,ylab="Score Residuals")
abline(h=0,col=4)
lines(lowess(Age,dataagescore),col=2) #γραμμή εξομάλυνσης
```

4)Πρόγραμμα για την Παράγραφο 6.4

```
# Υπολογισμός των Deviance Υπολοίπων και η γραφική τους παράσταση

dataagedev<-residuals(dataagecox,type=c("deviance"))
plot(Age,dataagedev,ylab="Deviance Residuals")
abline(h=0,col=4)
lines(lowess(Age,dataagedev),col=2)
```

5)Πρόγραμμα για την Παράγραφο 6.5

```
# Υπολογισμός των Schoenfeld Υπολοίπων και η γραφική τους παράσταση

dataageschoen<-residuals(dataagecox,type=c("schoenfeld"))
plot(Age[status==1],dataageschoen,xlab="Ηλικία",ylab="Schoenfeld Residuals")
abline(h=0,col=4)
lines(lowess(Age[status==1],dataageschoen),col=2)

# Γραφική παράσταση των υπολοίπων Schoenfeld  ως προς τον χρόνο
plot(time[status==1],dataageschoen,xlab="Χρόνος",ylab="Schoenfeld Residuals")
abline(h=0,col=4)
lines(lowess(time[status==1],dataageschoen),col=2)


# Έλεγχος της υπόθεσης αναλογικής διακινδύνευσης με τον έλεγχο των Grambsch και Therneau

cox.zph(dataagecox, transform='identity')

plot(cox.zph(dataagecox, transform='identity'))
```

6)Πρόγραμμα για την Παράγραφο 6.6

```
#Τα  normal deviate υπόλοιπα  για την ανίχνευση σημείων επιρροής

library(gbm)
b<-gbm(Surv(time,status)~Age,distribution="coxph")
h<-basehaz.gbm(time,status,b$fit)
s<-exp(-h)
s1<-s[status==1]# για μη αποκομμένες παρατηρήσεις
s2<-s[status==0]# για αποκομμένες παρατηρήσεις
z1<-qnorm(s1)
z2<-qnorm(s2/2)
z<-c(z1,z2)
plot(Age,z, ylim=c(-2,2),ylab=" Normal deviate υπόλοιπα",xlab="Ηλικία")
abline(h=1.96,col=4)
abline(h=0)
abline(h=-1.96,col=2)
```

7)Πρόγραμμα για την Παράγραφο 6.7

```
#Τα  Log-odds υπόλοιπα  για την ανίχνευση σημείων επιρροής

l1<-log(s1/(1-s1))
l2<-log(s2/(2-s2))
l<-c(l1,l2)
plot(Age,l,ylim=c(-4,4),ylab=" Log-odds υπόλοιπα",xlab="Ηλικία")
abline(h=3.66,col=4)
abline(h=0)
abline(h=-3.66,col=2)
```

8)Πρόγραμμα για την Παράγραφο 6.8

```
#Διαχωρισμός των ασθενών με βάση την ηλικία και γραφική παράσταση των καμπυλών
επιβίωσης

omada<-data[,'Age']-50
omada[omada>=0]<-1
omada[omada<0]<-0
plot(survfit(Surv(time,status)~omada),xlab="Χρόνος επιβίωσης",
+ ylab="Καμπύλη επιβίωσης",lty=1:2,col=1:4)
legend(30,1,c("Κάτω των 50","Άνω των 50"),lty=1:2,col=1:2)
```

9)Πρόγραμμα για την Παράγραφο 6.9

```
#Τα  dfbetas για την ανίχνευση σημείων επιρροής

databeta<- residuals(dataagecox,type="dfbetas")
plot(databeta,type="h",ylab="influence for Age",xlab="observation")
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Andrews, D. F. and Herzberg, A. M. (1985). *Data.* New York, Springer.

2. Atkinson, A. and Riani, M. (2000). *Robust Diagnostic Regression Analysis*. New York, Springer.

3. Atkinson, A. Riani, M. and Cerioli, A. (2010). The forward search: Theory and data analysis. *Journal of the Korean Statistical Society,* **39**, 117-134

4. Barnett, V. and Lewis, T. (1984). *Outliers in Statistical Data.* New York, Wiley.

5. Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* New York, John Wiley & Sons.

6. Ben-Gal, I. (2005). *Outlier detection.* Department of Industrial Engineering Tel-Aviv University, Israel.

7. Cain, K.C. and Lange, N.T. (1984). Approximate case influence for the proportional hazards regression model with censored data. *Biometrics*, **40**, 493-499.

8. Chen, C.-H. and Wang, P.C. (1991). Diagnostic plots in Cox's regression model. *Biometrics*, **47**, 841-850.

9. Collett, D. (2003). *Modeling Survival Data in Medical Research (2nd ed)*, Chapman and Hall/CRC, Florida.

10. Cox, D.R. (1975). Partial likelihood. Biometrika, 62, 269-276.

11. Cox, D.R. (1972). Regression analysis and life tables, (with discussion). *Journal of the Royal Statistical Society, Series B,* **34,** 187-220.

12. Cox, D.R. and Snell, E.J. (1968). A general definition of residuals (with discussion). *Journal of the Royal Statistical Society, A,* **30**, 248-275.

13. Diez, D. *Survival Analysis in R* (http://www.ddiez.com/teac/surv/R_survival.pdf)

14. Flemming, T. R. and Harrington, D. P. (1991). *Couting Processes and Survival Analysis*, Wiley, New York.

15. Grambsch, P.M. and Therneau, T.M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81**, 515-526.

16. Hadi, A. (1992). Identifying Multiple Outliers in Multivariate Data. *Journal of the Royal Statistical Society, Series B,* **54**, 761-71.

17. Hall, C.B., Zeger, S.L. and Bandeen-Roche, K.J. (1996). Adjusted variable plots for Cox's proportional hazards regression model. *Lifetime Data Analysis*, **2**, 73-90.

18. Harrell, F.E., Jr. (2001). *Regression Modeling Strategies, with Applications to Linear Models, Logistic Regression, and Survival Analysis*, Springer-Verlag, New York.

19. Hawkins, D. M. (1980). *Identification of Outliers*. London, Chapman and Hall.

20. Hess, K.R. (1995). Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. *Statist. Med.*, **14**, 1707-1723.

21. Imon, A. R. (2005). Identifying Influential Observations in Linear Regression. *Journal of Applied Statistics,* **32(9)**, 929-46.

22. Klein, J.P. and Moeschberger, M.L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data,* Springer-Verlag, New York.

23. Lee E.T. (1980). *Statistical Methods for Survival Data Analysis, Life Learning Publications*, Belncont, California.

24. Lindkvist, M. (2000). Properties of added variable plots in Cox's regression model. *Lifetime Data Analysis*, **6**, 23-38.

25. McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models* (2nd ed), Chapman and Hall, London.

26. Nardi A., Schemper M.(1999). New residuals for Cox regression and their application to outlier screening. *Biometrika,* **55**, 523-529.

27. NOR AKMAL BT. MD NOH. (2010). Αδημοσίευτη Διπλωματική Εργασία : *Detecting Outliers And Influential Observations In Survival Model*, Faculty Of Science University Of Malaya Kuala Lumpur. (http://studentsrepo.um.edu.my/1981/)

28. O' Hara Hines, R.J. and Carter, E.M. (1993). Improved added variable and partial residual plots for the detection of influential observations in generalized linear models. *Appl. Statist.*, **42**, 3-20.

29. Pettitt, A.N. and Bin Daud, I. (1989), Case-weighted measures of influence for proportional hazards regression. *Applied Statistics,* **38**, 51-67.

30. Qi, Z. (2009), *Comparison of Proportional Hazards and Accelerated Failure Time Models*, Department of Mathematics and Statistics University of Saskatchewan Saskatoon, Saskatchewan. (http://library.usask.ca/theses/available/etd-03302009-140638/unrestricted/JiezhiQiThesis.pdf)

31. Reid, N. and Crépeau, H. (1985). Influence Functions for Proportional Hazard Regression. *Biometrika*, **72**, 1-9.

32. Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, **69**, 239-241.

33. Storer, B. E. and Crowley, J. (1985). A Diagnostic for Cox Regression and General Conditional Likelihoods. *Journal of the American Statistical Association*, **80**(389), 139-147.

34. Tableman, M. (2008). *Survival Analysis Using S/R**, Department of Mathematics & Statistics Portland State University Portland, Oregon, USA (http://stat.ethz.ch/wbl/Skript_SurvivalAnalysis.pdf)

35. Therneau, T.M. and Grambsch, P.M. (2000). *Modeling Survival Data: Extending the Cox Model.*, Springer-Verlag, New York.

36. Therneau, T.M., Grambsch, P.M. and Fleming, T.R. (1990). Martingale-based residuals for survival models. *Biometrika,* **77**, 147-160.

37. Wang, H.-M., Jones, M. P. and Storer, B. E. (2006). Comparison of case-deletion diognostic methods for cox regression. *Statistics in Medicine*, **25**, 669-83.

38. Zhou, M. *Use Software R to do Survival Analysis and Simulation. A tutorial*, Department of Statistics, University of Kentucky (http://www.ms.uky.edu/~mai/Rsurv.pdf)

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

40. Οικονόμου Π. και Καρώνη Χ. (2010). *Στατιστικά Μοντέλα Παλινδρόμησης*, Συμεών, Αθήνα.