ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ»

Μεταπτυχιακή Διπλωματική εργασία

ΜΕΘΟΔΟΙ ΠΟΙΝΙΚΟΠΟΙΗΜΕΝΗΣ ΠΙΘΑΝΟΦΑΝΕΙΑΣ ΣΤΑ ΜΟΝΤΕΛΑ ΕΥΠΑΘΕΙΑΣ ΜΕ ΟΜΑΔΟΠΟΙΗΜΕΝΑ ΔΕΔΟΜΕΝΑ

ΧΑΒΙΑΤΖΗ ΑΝΝΑ

Επιβλέπων: Κουκουβίνος Χρήστος, Καθηγητής Ε.Μ.Π.

ΑΘΗΝΑ 2015

ΠΕΡΙΛΗΨΗ

Τα μοντέλα ευπάθειας, τα οποία έχουν προταθεί κατά καιρούς από πολλούς ερευνητές, εφαρμόζονται σε προβλήματα που αφορούν την ετερογένεια μεταξύ των ατόμων τα οποία παρουσιάζουν διαφορετικούς κινδύνους. Στο πρώτο κεφάλαιο μελετώνται δύο ευρείες κατηγορίες των μοντέλων ευπάθειας, τα πολυμεταβλητά και τα μονομεταβλητά μοντέλα ευπάθειας.

Το από κοινού μοντέλο ευπάθειας (Shared frailty models), κατά τον Hougaard (2000) αποτελούν ένα είδος μοντέλου κοινού κινδύνου που περιγράφεται στο κεφάλαιο 2. Πολλές είναι ο κατανομές που χρησιμοποιούνται για την εφαρμογή των μοντέλων ευπάθειας. Στη παρούσα μεταπτυχιακή εργασία θεωρούμε τη Γάμμα κατανομή ως τη πιο δημοφιλή για τα από κοινού μοντέλα ευπάθειας. Παρόλο αυτά, παρουσιάζονται λεπτομερώς στο ίδιο κεφάλαιο και άλλες προτεινόμενες κατανομές καθώς και ορισμένοι στατιστικοί μέθοδοι εκτίμησης.

Στην ανάλυση επιβίωσης, η επιλογή μεταβλητών παίζει σημαντικό ρόλο στη στατιστική μοντελοποίηση. Πολλές είναι οι μέθοδοι όπως αυτή της κατά βήματα επιλογής μεταβλητών και της μεθόδου επιλογής καλύτερου υποσυνόλου που αγνοούν τα στοχαστικά λάθη. Σε αντίθεση με τις παραδοσιακές μεθόδους επιλογής μεταβλητών, εστιάζουμε σε μεθόδους ποινικοποιημένης πιθανοφάνειας. Οι μέθοδοι αυτοί έχουν την δυνατότητα να διαγράφουν τις σημαντικές μεταβλητές εκτιμώντας τους συντελεστές ως μηδενικούς. Στο κεφάλαιο 3, παρουσιάζονται νέοι μέθοδοι επιλογής μεταβλητών για τα γραμμικά μοντέλα, τα εύρωστα μοντέλα παλινδρόμησης και τα γενικευμένα μοντέλα βασιζόμενα στη μέθοδο ποινικοποιημένης πιθανοφάνειας. Επίσης, προτείνονται μια σειρά από συναρτήσεις. Η συνάρτηση ποινής SCAD και η LASSO με L_1 συνάρτηση ποινής η οποία προτάθηκε από τον Tibshirani (1996). Επιπλέον, ένας νέος αλγόριθμος προτείνεται και παρουσιάζεται αναλυτικά.

Τέλος, στο τέταρτο κεφάλαιο, η μέθοδος ποινικοποιημένης πιθανοφάνειας επεκτείνεται και εφαρμόζεται στο μοντέλο ευπάθειας. Μελέτες και παραδείγματα αποδεικνύουν ότι οι προτεινόμενοι μέθοδοι είναι πιο αποτελεσματικοί στον υπολογισμό σε σχέση με τη μέθοδο επιλογής καλύτερου υποσυνόλου. Συγκριτικά με τις μεθόδους LASSO και garrote οι νέες μέθοδοι επιφέρουν καλύτερα αποτελέσματα.

.

ABSTRACT

Frailty models have been suggested by various researchers and applied to problems of heterogeneity of the subject, blend of individuals with dissimilar hazards. In chapter 1 historically two broad class of frailty models are studied, multivariate and univariate frailty models.

The shared frailty model is a specific kind of the common risks model described in chapter 2. There are many assumptions about the distributions of the frailty models. In this thesis we consider gamma distribution as the most popular distribution used for the shared frailty model. However, the proposed distributions are represented in detail in chapter 2 as well some statistical methods of estimates of the shared frailty model.

In survival analysis, variable selection is vital to statistical modeling. There are many procedures in use, such as stepwise selection and best subset selection witch ignore stochastic errors. Unlike the traditional variable selection, we focus on the penalized likelihood procedures. Thus, these methods delete significant variables by estimate their coefficient as 0. In chapter 3 are proposed a few new approaches to selecting variables for linear models, robust linear regression models and generalized linear models based on a penalized likelihood approach. Also a family of thresholding functions is proposed. A Smoothly Clipped Absolute Deviation (SCAD) penalty function and the LASSO proposed by Tibshirani (1996) with the L_1 penalty function. Furthermore, a new algorithm is proposed in chapter 3, which is backed up by statistical theory.

Finally, in chapter 4 the penalized likelihood approach is extended further to the Cox proportional hazard frailty model. Simulation and studies show that the proposal methods are more effective in computation than the best subset variable selection. Compared with the LASSO and the garrote method, the newly approaches have effective sample performance.

ΕΥΧΑΡΙΣΤΙΕΣ

Με την ολοκλήρωση της μεταπτυχιακής μου εργασίας, θα ήθελα να ευχαριστήσω θερμά τα μέλη του Εθνικού Μετσόβιου Πολυτεχνείου που συνέβαλαν στη διεκπεραίωσή της.

Κατά κύριο λόγο, οφείλω να εκφράσω τις θερμές μου ευχαριστίες στον επιβλέποντα καθηγητή μου κ. Χρήστο Κουκουβίνο Καθηγητής του Ε.Μ.Π για την εμπιστοσύνη που μου έδειξε δίνοντας μου τη δυνατότητα να εκπονήσω τη μεταπτυχιακή μου εργασία σε ένα συγκεκριμένο θέμα που αφορά τα ερευνητικά μου ενδιαφέροντα. Τον ευχαριστώ επίσης για τις πολύτιμες συμβουλές που μου παρείχε καθ' όλη τη διάρκεια της εργασίας, καθώς και για τις γνώσεις τη καθοδήγηση και την υποστήριξη του κατά τη διάρκεια των μεταπτυχιακών μου σπουδών.

Ιδιαίτερες ευχαριστίες οφείλω στον υποψήφιο διδάκτορα Εμμανουήλ Ανδρουλάκη για τη καθοριστική και πολύτιμη βοήθεια που μου προσέφερε, καθώς και τον χρόνο που αφιέρωσε για την εκπόνηση της διπλωματικής μου εργασίας. Επιπλέον, τον ευχαριστώ τόσο για τις για τις γνώσεις που μου εμπιστεύτηκε όσο και για τις εύστοχες συμβουλές, υποδείξεις, καθοδήγηση και συμπαράσταση που οδήγησαν στην ομαλή διεκπεραίωση της εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου για τη στήριξη και την εμπιστοσύνη που μου έδειξε καθ' όλη τη διάρκεια των μεταπτυχιακών μου σποδών καθώς και τους στενούς μου φίλους για τη ανεκτίμητη συμπαράσταση και κατανόηση τους.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

1.1 Ιστορική Εξέλιξη

Ένας από τους πρωταρχικούς στόχους της ανάλυσης επιβίωσης είναι να εξεταστεί ο χρόνος μέχρις ότου συμβεί ένα γεγονός. Τις περισσότερες φορές καταφεύγουμε στη μέθοδο προσέγγισης μοντελοποίησης ώστε να περιγράψουμε μια κατάσταση που παράγει αποτελέσματα τα οποία είναι στατιστικά ερμηνεύσιμα. Για τη διεξαγωγή σωστών συμπερασμάτων και ερμηνειών πρέπει πρώτα να γίνει παραδοχή σχετικά με την κατανομή των χρόνων επιβίωσης. Ωστόσο, για να πληρούνται κάποιες προδιαγραφές σε σχέση με τη κατανομή σημαντικό ρόλο λαμβάνει η παρουσία της λογοκρισίας. Η λογοκρισία αποτελεί ένα ξεχωριστό μέρος της ανάλυσης επιβίωσης που πολλές φορές μας καθιστά ανίκανους να χρησιμοποιούμε άμεσα τις τυποποιημένες διαδικασίες της στατιστικής ανάλυσης. Συμβατικές παραδοχές όπως ‘ομαλότητα’ δεν ενδείκνυνται για τις περισσότερες από αυτές τις περιπτώσεις. Επίσης, οι προδιαγραφές μπορεί να είναι αδικαιολόγητα αυστηρές. Ως εκ τούτου, για το χειρισμό τέτοιων δεδομένων, θα πρέπει να στραφούμε σε μη-παραμετρικές ή ημι-παραμετρικές προσεγγίσεις. Στα 1900 υπήρχε ήδη μια ακμάζουσα αναλογιστική βιβλιογραφία σχετικά με την κατασκευή και την ερμηνεία των πινάκων ζωής για να περιγράψουν την εμπειρία θνησιμότητας, όπως τεκμηριώνεται από τον Oakes, D. (2001) αλλά οι τεχνικές αυτές δεν ήταν επαρκείς για την προηγμένη στατιστική ανάλυση. Τα έργα του Kaplan, E.L. and Meier, P. (1958), καθώς και ορισμένες από τις διαδοχικές εξελίξεις τους (Efron, 1967; Breslow, N.E. και Crowley, 1974; Koziol, J. A. και Green, S. B. , 1976), κατέστησαν δυνατή την μη παραμετρική εκτίμηση της μέγιστης πιθανοφάνειας μιας συνάρτησης κατανομής από λογοκριμένα δεδομένα και αντλούν κάποια από τις ανώτερες ιδιότητές του. Ο Cox, DR (1972) πρότεινε μια ευφάνταστη μεθοδολογία που μας απαλλάσσει από οποιοδήποτε υπόθεση κατανομής, και εξακολουθεί να μας δίνει τη

δυνατότητα να αναδείξουμε τις εκτιμήσεις των συντελεστών παλινδρόμησης από το γραμμικό μοντέλο, όπως για τους κινδύνους καταγραφής. Αυτό είναι μια γενίκευση της μεθόδου που προτείνεται από τον Kaplan, E. L. Meier, P. (1958) με την οποία μπορούν να αξιολογηθούν οι επιδράσεις συμμεταβλητών.

Η συνάρτηση διακινδύνευσης ή συνάρτηση κινδύνου (*hazard function*) εκφράζει τη διάρκειας ζωής και αποτελεί τον πιο φυσικό τρόπο για την αντιμετώπιση τέτοιων μοντέλων. Στο βασικό μοντέλο η συνάρτηση διακινδύνευσης ή συνάρτηση κινδύνου χωρίζεται σε δύο συνιστώσες, η μία είναι η αναφορική συνάρτηση διακινδύνευσης (baseline hazed function) που μπορεί να αλλάξει με την πάροδο του χρόνου και είναι ανεξάρτητη από τις επιδράσεις των συμμεταβλητών (με απροσδιόριστη λειτουργική μορφή), ενώ η άλλη είναι μια μορφή συμμεταβλητών οι οποίες χαρακτηρίζουν πως οι συμμεταβλητές επηρεάζουν τη συνάρτηση κινδύνου (τις περισσότερες φορές αυτή η λειτουργία περιορίζεται στις μη-αρνητικές τιμές). Αυτό το μοντέλο έχει ονομαστεί ως μοντέλο αναλογικού κινδύνου του Cox(Cox proportional hazards model) και οφείλεται στο γεγονός ότι οι συναρτήσεις κινδύνου σχετίζονται πολλαπλασιαστικά με το αρχικό μοντέλο. Ως συνέπεια αυτής της υπόθεσης, η αναλογία κινδύνου (hazard ratio) είναι ανεξάρτητη του χρόνου επιβίωσης. Αφού υπολογίσουμε τις εκτιμήσεις των συντελεστών παλινδρόμησης των συμμεταβλητών, μπορούμε εν συνεχεία να διεξάγουμε μια λογική εκτίμηση των βασικών λειτουργιών επιβιωσιμότητας εκτιμώντας με αυτόν τον τρόπο την αναφορική συνάρτηση διακινδύνευσης (baseline hazed function).

1.2 Κίνητρα για την τροποποίηση του συμβατικού μοντέλου (conventional model)

Υπάρχουν περιπτώσεις που ενδέχεται η ύπαρξη ορισμένων παραγόντων, άλλους από τις μετρήσιμες συμμεταβλητές, που επηρεάζουν σημαντικά τις παραμέτρους και τροποποιούν τη κατανομή του χρόνου επιβίωσης. Υπάρχουν πολλοί λόγοι για τις μη κατανεμημένες συμμεταβλητές, για παράδειγμα, αν υπάρχουν πολλές συμμεταβλητές που πρέπει να ληφθούν υπόψη, είναι σχεδόν αδύνατον για τους ερευνητές να τις συμπεριλάβουν όλες. Στη συνέχεια παραβλέπονται ορισμένες από αυτές ώστε να γίνει πιο εύκολη η διαδικασία συλλογής δεδομένων (για την

αποτελεσματικότητα, το χρόνο ή απλούστερα λόγω έλλειψης των κατάλληλων εργαλείων μέτρησης). Ακόμα και αν λάβουμε υπόψη όλους τους γνωστούς παράγοντες προκύπτει ένα διαφορετικό είδος προβλήματος με αποτέλεσμα το μοντέλο να επαναπροσδιοριστεί και να μην υπάρχουν πιθανές εκτιμήσεις συντελεστών παλινδρόμησης. Ένας άλλος συνηθισμένος λόγος είναι ότι οι ερευνητές δεν γνωρίζουν την επίδραση των πιθανών συμμεταβλητών που μπορεί να υπάρχουν. Για παράδειγμα, αν υπάρχει ένας γενετικός παράγοντας που ευθύνεται για πιθανό ενδεχόμενο εμφάνισης μιας ασθένειας, η οποία μπορεί να μην είναι γνωστή σε εμάς, τότε θα ήταν αδύνατον για τους ερευνητές να τις συμπεριλάβουν ως συμμεταβλητές. Τέτοιες συμμεταβλητές ονομάζονται μη παρατηρήσιμες συμμεταβλητές (unobserved covariates). Ωστόσο, ποτέ δεν θα ήταν αρκετή η εισαγωγή συμμεταβλητών σε ένα μοντέλο , ειδικά στις ερευνητικές μελέτες. Αυτό οφείλεται στο γεγονός ότι τα άτομα διαφέρουν σημαντικά μεταξύ τους ως προς την ευαισθησία σε διάφορες ασθένειες και ως προς τη κατάσταση θνησιμότητας, ανεξάρτητα από το πόσο ταυτίζονται οι τιμές των γνωστών συμμεταβλητών (στη βιβλιογραφία αυτή η ιδέα ονομάζεται κρυφή ετερογένεια η ευπάθεια και χαρακτηρίζουν την επιβίωση, μεταβολές που οφείλονται στην ατομική διακύμανση η οποία είναι προφανώς μη παρατηρήσιμη). Ως εκ τούτου, για πρακτικούς λόγους, οι μη παρατηρήσιμες συμμεταβλητές αγνοούνται θεωρώντας τες ως μέρος του στατιστικού σφάλματος και είναι δύσκολο να ελεγχτούν στη συμβατική ανάλυση επιβίωσης. Αυτό μπορεί να απλοποιήσει τη διαδικασία υπολογισμού και ίσως είναι ένα σημαντικό πλεονέκτημα. Εν συνεχεία, η ετερογένεια προστίθεται στην μεταβλητότητα αντιπροσωπεύοντας τις μετρήσιμες συμμεταβλητές με σκοπό την αύξηση αυτής της μεταβλητότητας. Η παρουσία της αυξημένης διακύμανσης προκαλεί αλλαγή στη συνάρτηση κινδύνου και έτσι η εκτίμηση των συντελεστών παλινδρόμησης θα πρέπει να είναι μεροληπτική. Τέτοιου είδος μεροληψία λόγω των παραληφθέντων συμμεταβλητών, προτάθηκαν από συγγραφείς όπως οι Gail και άλλοι. [1984] και Chastang και άλλοι [1988], ενώ μελέτες των Bretagnolle και Huber-Carol [1985, 1988]; Chamberlain [1985] έδειξαν το ίδιο. Οι Henderson and Oman [1999] διερεύνησαν τις συνέπειες που είχε η αγνόηση της ευπάθειας στην ανάλυση επιβίωσης.

Σύμφωνα με τα παραπάνω, τα άτομα μέσα σε μια ομάδα παρουσιάζουν μεταξύ τους ανόμοια χαρακτηριστικά και γι αυτό το λόγο το μοντέλο πρέπει να βελτιωθεί έτσι ώστε να καλύψει αυτή την ετερογένεια (λαμβάνοντας υπόψη ότι η ατομικότητα θεωρείται βασικό γεγονός κάθε μορφή ζωής). Με βάση μελετών που

έχουν γίνει στο παρελθόν, υπάρχει σημαντική μεταβολή στον κίνδυνο ανάπτυξης διαφόρων ασθενειών και ως εκ τούτου, τα άτομα διαφέρουν σημαντικά ως προς την ευπάθεια διαφόρων εκδηλώσεων ασθένειας. Άτομα με διαφορετικές αδυναμίες και ευπάθειες και όσα από αυτά τα άτομα παρουσιάζουν σε μεγαλύτερο βαθμό αυτά τα δυο χαρακτηριστικά τόσο πιο γρήγορα θα αποβιώσουν σε σχέση με τα άλλα άτομα. Ο λόγος για τον οποίο μπορεί να συμβεί αυτό είναι τα διαφορετικά τύπου γονίδια που φέρουν τα άτομα, το διαφορετικό τρόπο ζωής που ακολουθούν και πολλά άλλα. Στο εξής σημειώνουμε ότι, η υπόθεση του αναλογικού κινδύνου, και ως εκ τούτου η ομοιογένεια, δηλαδή όλα τα θέματα κάτω από τον ίδιο κίνδυνο εμφάνισης μιας ασθένειας, φαίνεται να είναι ένα μαθηματικό εργαλείο που αναλύει και ερμηνεύει το εν λόγω μοντέλο με τον πιο κατάλληλο και απλούστερο τρόπο.

1.3 Προτεινόμενες τροποποιήσεις στη βιβλιογραφία

Το μοντέλο αναλογικού κινδύνου του Cox είχε μεγάλη επιτυχία στο να περιγράψει πολλά πρακτικά ζητήματα και έλαβε ευρεία χρήση, παρά το γεγονός ότι ο βασικός κίνδυνος δρα πολλαπλασιαστικά στην επίδραση των συμμεταβλητών. Για να είναι πιο εξελιγμένο το μοντέλο αυτό, έχουν προστεθεί ποικίλες μεταβλητές για την αντιμετώπιση διαφόρων πρακτικών προβλημάτων. Επίσης, το μοντέλο αυτό έχει ληφθεί ως βάση για μεταγενέστερες εργασίες. Κάποια από αυτά είναι το αθροιστικό ημι-παραμετρικό μοντέλο Breslow και Day (1987) για το οποίο προτάθηκε από τον Lin Ying μια εύκολη προσέγγιση μη επαναληπτικής εκτίμησης αλλά η χρήση του εξακολουθεί να μένει περιοριστική. Επίσης, γνωστό είναι το μοντέλο αθροιστικής παλινδρόμησης του Aalens και το μοντέλο τυχαίων επιδράσεων του Hougaard (2000).

1.4 Η ορθολογική ερμηνεία του μοντέλου ευπάθειας

Τα μοντέλα ευπάθειας, τα οποία έχουν κατά καιρούς προταθεί από πολλούς ερευνητές εφαρμόζονται σε προβλήματα που αφορούν την ετερογένεια μεταξύ των ατόμων τα οποία παρουσιάζουν διαφορετικούς κινδύνους. Καθώς η συνάρτηση διακινδύνευσης είναι μεροληπτική και το πρόβλημα αυτής τη μεροληψίας είναι τόσο

μεγάλο ώστε να αντιμετωπιστεί, πρέπει να γίνει η αξιολόγηση αυτής της κρυμμένης ετερογένειας της συνάρτησης διακινδύνευσης. Το μοντέλο διορθώνει αυτή τη μεροληψία των συντελεστών παλινδρόμησης του μοντέλου αναλογικού κινδύνου του Cox (Chamberlain 1985) και συμβάλλουν στην περιγραφή της μη αναλογικότητας του δεσμευμένου κινδύνου. Ο Clayton (1978) ήταν αυτός που αρχικά παρουσίασε τα μοντέλα ευπάθειας για το διμεταβλητό μοντέλο, ενώ η γενίκευση του πολυμεταβλητού μοντέλου του Cox παρουσιάστηκε από τον Clayton και Cuzick (1985).

1.5 Θεωρητικό πλαίσιο των μοντέλων ευπάθειας

Η βασική ιδέα στο μοντέλο ευπάθειας είναι να ενσωματώσει έναν πολλαπλασιαστικό παράγοντα κλίμακας γνωστό ως ευπάθεια (frailty) στην αναφορική συνάρτησης κινδύνου, έτσι ώστε η συνάρτηση κινδύνου να μπορεί να χωριστεί σε τρείς βασικές πολλαπλασιαστικές συνιστώσες :

- τη τυχαία επίδραση του όρου ευπάθειας (random effect frailty term), ένας καινούργιος παράγοντας που συνοψίζει τις επιπτώσεις των μη παρατηρούμενων μεταβλητών στη συνάρτηση κινδύνου
- την αναφορική συνάρτηση κινδύνου
- τη συνάρτηση συνδιακύμανσης (covariate function)

Καθώς ο παράγοντας κλίμακας είναι μη παρατηρήσιμος πρέπει να υποθέσουμε ότι οι πληροφορίες για τον παράγοντα κλίμακα περιέχονται στον τύπο της συνάρτησης κινδύνου. Ωστόσο, η συνάρτηση κινδύνου δεν μπορεί να είναι ποτέ αρνητική και έτσι ο παράγοντας κλίμακας, έστω u, πρέπει να αποτιμηθεί ως μια θετική τυχαία μεταβλητή. Επίσης, ο παράγοντας κλίμακας μπορεί να εκφραστεί ως w όπου το u είναι η δύναμη της συνάρτησης ws. Έπειτα, η συνάρτηση διακινδύνευσης μαθηματικά εξισώνει τη βασική συνάρτηση η οποία πολλαπλασιάζεται εκθετικά συναρτήσει των συμμεταβλητών και εκθετικά συναρτήσει του w, όπου οι συναρτήσεις περιλαμβάνουν τους αντίστοιχους συντελεστές παλινδρόμησης που χαρακτηρίζουν τη σχέση των συμμεταβλητών και των w (αν οι συμμεταβλητές των ws είναι μηδέν, τότε το αναλογικό μοντέλο διακινδύνευσης είναι απλό). Θεωρούμε

τα w ως ανεξάρτητο δείγμα από την αυθαίρετη κατανομή με την ιδιότητα ότι η μέση τιμή είναι μηδέν. Κατά συνέπεια, καθώς παραμένουν ανεξάρτητα και ομοιόμορφα κατανεμημένα έχοντας την ιδιότητα να είναι η μέση τιμή ένα με κάποια πεπερασμένη διακύμανση. Αν αυτό αληθεύει, τότε η ερμηνεία γίνεται πιο εύκολη: θεωρώντας u=1 ως επικινδυνότητα των υποκειμένων (αυτό είναι το συνηθισμένο αναλογικό μοντέλο), για u > 1 πιο γρήγορη ατομική αποτυχία, ενώ για u < 1 , τα άτομα επιβιώνουν περισσότερο έχοντας μικρότερη ευπάθεια από το μέσο όρο. Παρόλο αυτά, χρειάζονται να γίνουν ορισμένες υποθέσεις για αυτόν το παράγοντα κλίμακα που είναι ανεξάρτητος τόσο από τις συμμεταβλητές όσο και από τους χρόνους επιβίωσης.

Το νέο μοντέλο ευπάθειας προϋποθέτει ότι:

- Το όριο ζωής, δεδομένου της ευπάθειας είναι, υπό κατάλληλες συνθήκες, ανεξάρτητο.
- Ο κίνδυνος για κάθε χρόνο επιβίωσης, ακολουθεί ένα μοντέλο αναλογικού κινδύνου.
- Ο παράγοντας ευπάθειας και η επίδραση των συμμεταβλητών δρουν πολλαπλασιαστικά στον βασικό κίνδυνο.
- Ο όρος ευπάθεια και οι συμμεταβλητές είναι ανεξάρτητες.
- Η τιμή της σταθεράς ευπάθειας συμπεριλαμβάνεται σε μια ομάδα.
- Κάθε αντικείμενο μελέτης μπορεί να αντιμετωπιστεί μετά από ένα μεγάλο χρονικό διάστημα.

1.6 Πρακτική δυσκολία εκτίμησης και πιθανά αίτια

Πρόβλημα αποτελεί ο τρόπος με τον οποίο πρέπει να εκτιμήσουμε τη συνάρτηση επιβίωσης και τη συνάρτηση πυκνότητας, καθώς και τα δύο από αυτά σχετίζονται μαθηματικά με τη συνάρτηση κινδύνου και εξαρτώνται καθαρά από τη παράμετρο κλίμακα η οποία είναι μη παρατηρήσιμη. Γι αυτό το λόγο πολλοί ερευνητές καθορίζουν μια παραμετρική μορφή κατανομής αυτού του μη παρατηρήσιμου παράγοντα κλίμακα δεδομένου τον αρκετά μικρό αριθμό παραμέτρων και έπειτα, για τη συνεχή περίπτωση ενσωματώνουν τον όρο ευπάθεια.

Επίσης, από άλλη οπτική σκοπιά, τα κοινά μοντέλα ευπάθειας αποτελούν μια ειδική περίπτωση αναδρομικών γραφικών μοντέλων (Cox and Wermuth, 1996).
Τα έργα του Gottard και άλλοι (2004) πρότειναν μια πιθανή λύση για την επέκταση των γραφικών μοντέλων για τα από κοινού μοντέλα ευπάθειας, η γενικότερα για τα συσχετισμένα δεδομένα.

Ο Fine (και άλλοι, 2003) πρότεινε μια απλή μέθοδο για την απόκτηση της παραμέτρου ευπάθειας χρησιμοποιώντας την ανάλυση των αποτελεσμάτων του αναλογικού μοντέλου του Cox η οποία μειώνει την υπολογιστική επιβάρυνση από άλλες προσεγγίσεις εκτίμησης. Επιπλέον, ορισμένοι συγγραφείς (Qiou και άλλοι, 1999;.Muller και Quintana, 2004; Casarin, 2004;. Chen και άλλοι, 2002) χρησιμοποιούν Μπεϋσιανές μεθόδους για τη μοντελοποίηση πολλών μεταβλητών δεδομένων επιβίωσης στην ευπάθεια, βασισμένοι στην Μπεϋσιανή θεωρία. Στη συνηθισμένη προσέγγιση στατιστικής μοντελοποίησης (για μη λογοκριμένες περιπτώσεις) το πρόβλημα της ετερογένειας μπορεί να προσεγγιστεί από Γενικευμένα Γραμμικά Μοντέλα, στρωματοποίηση, ή για να διευκρινίσουν μια συναρτησιακή μορφή της πυκνότητας της μεταβλητότητας.

1.7 Ευρείες κατηγορίες των μοντέλων ευπάθειας.

Ο όρος ευπάθεια προτάθηκε αρχικά από τον Vaupel (και άλλοι, 1979) και χρησιμοποιείται για τα μονομεταβλητά μοντέλα επιβίωσης. Ωστόσο, ο Clayton (1978) ήταν εκείνος ο οποίος προώθησε το μοντέλο, μέσα από τις αναφορές του, σε πολυμεταβλητές καταστάσεις χρόνιων ασθενειών των οικογενειών. Ένα μοντέλο τυχαίων επιδράσεων λαμβάνει υπόψη του τις επιδράσεις μη παρατηρήσιμων συμμεταβλητών ή τη μη παρατηρήσιμη ετερογένεια η οποία προκαλείται από διαφορετικές πηγές. Η τυχαία επίδραση καλείται ευπάθεια και συμβολίζεται με Y (όπως αναφέρεται αναλυτικά παρακάτω) και εκφράζει το κοινό ρίσκο ή την ατομική ετερογένεια, δρώντας ως ένας παράγοντας στη συνάρτηση κινδύνου. Κατά συνέπεια, υπάρχουν δύο κατηγορίες των μοντέλων ευπάθειας οι οποίες είναι :

- Μονομεταβλητά μοντέλα ευπάθειας, που εξετάζουν μονομεταβλητούς χρόνους επιβίωσης.

- Πολυμεταβλητά μοντέλα ευπάθειας, που λαμβάνουν υπόψη τους χρόνους επιβίωσης πολλών μεταβλητών.

1.7.1 Μοντέλα μονομεταβλητών χρόνων επιβίωσης (univariate survival time data)

Τα μονομεταβλητά μοντέλα επιβίωσης λαμβάνουν υπόψη τους ότι μεταξύ των πληθυσμών δεν υπάρχει ομοιογένεια. Η ετερογένεια αυτή ερμηνεύεται με τη βοήθεια των συμμεταβλητών. Σε περίπτωση όμως που οι συμμεταβλητές αυτές είναι μη παρατηρήσημες τότε οδηγούμαστε στη μη παρατηρήσιμη ετερογένεια (Vaupuel 1979). Τέτοια μοντέλα εφαρμόζονται συχνά σε σύνολο δεδομένων που τα άτομα είναι ασυσχέτιστα μεταξύ τους. Η βασική ιδέα είναι να υποθέσουμε ότι, διαφορετικοί ασθενείς έχουν διαφορετικές ευπάθειες και οι ασθενείς που είναι πιο επιρρεπείς τείνουν να εμφανίσουν νωρίτερα την ασθένεια σε σχέση με αυτούς που είναι λιγότερο ευπαθείς. Συνεπώς, σημαντική είναι η συστηματική επιλογή ισχυρών ατόμων, δηλαδή ατόμων με χαμηλού βαθμού ευπάθεια. Επιπλέον, ενδιαφέρον παρουσιάζουν και τα ποσοστά θνησιμότητας που αλλάζουν με τη πάροδο του χρόνου και την ηλικία. Αρκετά συχνά παρατηρείται ότι η συνάρτηση κινδύνου (ή αλλιώς το ποσοστό θνησιμότητας) αυξάνεται στην αρχή, φτάνει σε ένα μέγιστο σημείο και έπειτα μειώνεται (unimodal intensity) σε μια σταθερή τιμή. Όσο περισσότερο ζήσει ένας ασθενής μετά την εκδήλωση της νόσου, τόσο περισσότερο βελτιώνεται η πιθανότητα επιβίωσής του. Η μεταβολή αυτή της συνάρτησης κινδύνου, είναι συχνό αποτέλεσμα της διαδικασίας επιλογής που ενεργεί σε έναν ετερογενή πληθυσμό και δεν αντικατοπτρίζει την ατομική θνησιμότητα. Καθώς φθίνει ο βαθμός του πληθυσμού, ο βαθμός κινδύνου αυξάνεται σημαντικά. Αν οι παράγοντες κινδύνου θεωρούνται γνωστοί τότε συμπεριλαμβάνονται στο μοντέλο αναλογικού κινδύνου και δίνονται από τον τύπο

$$\mu(t,z) = \mu_0(t)\exp(\beta' z)$$

Όπου η αναφορική συνάρτηση κινδύνου, η οποία υποτίθεται ότι είναι μοναδική για όλα τα άτομα του εξεταζόμενου πληθυσμού, z το διάνυσμα των παρατηρούμενων συμμεταβλητών και β των διάνυσμα των αντίστοιχων παραμέτρων παλινδρόμησης.

Δυο είναι οι βασικοί λόγοι για τους οποίους δεν μπορούμε να συμπεριλάβουμε όλους τους σημαντικούς παράγοντες στην παραπάνω ανάλυση.

Πρώτον, οι συμμεταβλητές που πρέπει να εξεταστούν είναι πάρα πολλές και δεύτερον ο ερευνητής μπορεί να μην ξέρει ή να μη μπορεί να μετρήσει όλες τις σχετικές μεταβλητές. Η μεταβλητότητα και στις δύο περιπτώσεις διακρίνεται στην μεταβλητότητα που μετριέται από αυτούς. Σε ένα μοντέλο αναλογικών κινδύνων η εξαίρεση κάποιου υποσυνόλου μεταβλητών οδηγεί σε μεροληπτικές εκτιμήτριες των παραμέτρων της παλινδρόμησης και του ποσοστού κινδύνου. Ο λόγος για αυτό είναι ότι το εξαρτώμενο από το χρόνο ποσοστό κινδύνου αλλάζει καθώς η σύνθεση του πληθυσμού αλλάζει σε σχέση με τις συμμεταβλητές. Μια εκτίμηση του κινδύνου χωρίς να ληφθεί υπόψη η μη παρατηρούμενη ευπάθεια θα οδηγήσει σε υποεκτίμηση της συνάρτησης κινδύνου και η έκταση της υποτίμησης θα αυξάνεται καθώς ο χρόνος προχωρεί. Το μονομεταβλητό μοντέλο ευπάθειας επεκτείνει το μοντέλο του Cox έτσι ώστε ο κίνδυνος του ατόμου να εξαρτάται από μια επιπλέον τυχαία μεταβλητή Y, η οποία ενεργεί πολλαπλασιαστικά στην αναφορική συνάρτηση κινδύνου

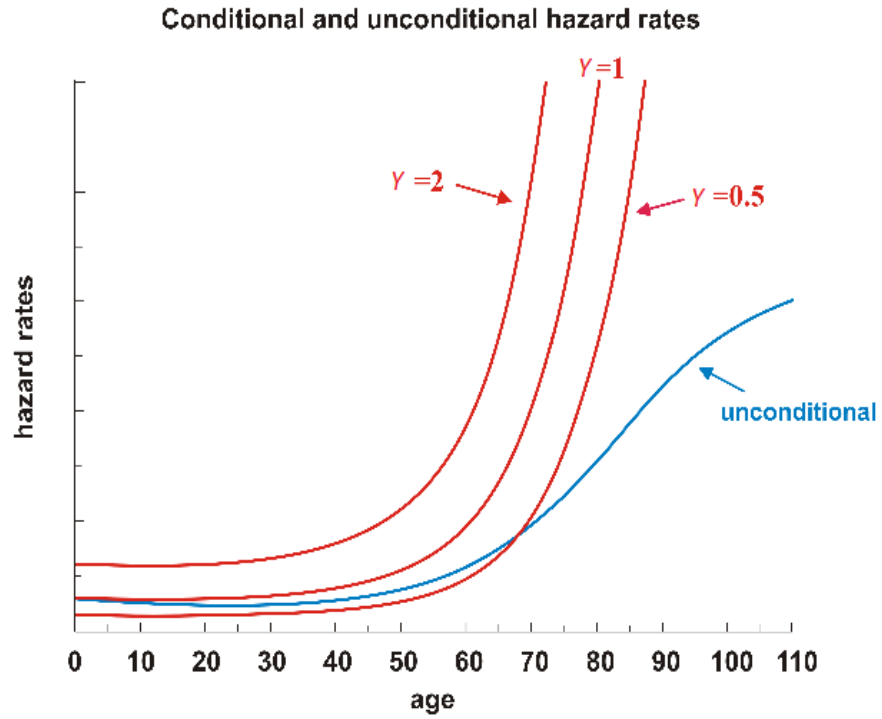
$$\mu(t,Y,z) = Y\mu_0(t)\exp(\beta' z)$$

Η ευπάθεια Y είναι μία τυχαία μεταβλητή που μεταβάλλεται μέσω του πληθυσμού, μειώνει τον ατομικό κίνδυνο όταν Y<1 ή τον αυξάνει όταν Y>1. Πρέπει να τονιστεί ότι η ευπάθεια είναι κάτι που δεν παρατηρείται. Η αντίστοιχη συνάρτηση επιβίωσηςπου περιγράφει το σύνολο των επιζώντων ατόμων στον πληθυσμό της μελέτης δίνεται από τον ακόλουθο τύπο

$$S(t \setminus Y,z) = \exp\left[-Y\exp(\beta' z)\int_0^t \mu_0(u)du\right] = \exp\left[-Y\exp(\beta' z)M_0(t)\right]$$

όπου $S(t \setminus Y,z)$ ερμηνεύεται ως το ποσοστό των ατόμων που έχει επιζήσει μέχρι την χρονική στιγμή και $M_0(t)$ η αθροιστική αναφορική συνάρτηση κινδύνου.

Μέχρι τώρα το μοντέλο έχει περιγραφεί σε ατομικό επίπεδο. Το ατομικό αυτό μοντέλο δεν παρατηρείται. Συνεπώς, πρέπει το μοντέλο να εξεταστεί σε επίπεδο ολόκληρου του εξεταζόμενου πληθυσμού. Η συνάρτηση επιβίωσης όλου του πληθυσμού είναι ο μέσος όρος των ατομικών συναρτήσεων επιβίωσης. Η συνάρτηση κινδύνου όλου του πληθυσμού μπορεί να διαφέρει από την ατομική συνάρτηση κινδύνου.Το ποσοστό κινδύνου του πληθυσμού έχει εντελώς διαφορετική απεικόνιση σε σχέση με τους ατομικούς κινδύνους όπως φαίνεται και στο ***Σχήμα 1.1***.



Σχήμα 1.1*: *Δεσμευμένα και μη, ποσοστά κινδύνου σε προσομοιωμένα σύνολα δεδομένων της ανθρώπινης θνησιμότητας.

1.7.2 Μοντέλα πολυμεταβλητών χρόνων επιβίωσης

Το πολυμεταβλητό μοντέλο επιβίωσης με τυχαίους κινδύνους είναι μια επέκταση του παραδοσιακού μονομεταβλητού μοντέλου ευπάθειας και μας επιτρέπει να λάβουμε υπόψη, στην ανάλυση επιβίωσης, την αμοιβαία εξάρτηση των χρόνων ζωής των συγγενικών προσώπων. Ένα ενδεικτικό παράδειγμα θα μπορούσε να είναι οι χρόνοι επιβίωσης εμφάνισης μιας ασθένειας μεταξύ διδύμων ή γονείς με παιδιά. Η σχέση αυτή μεταξύ των χρόνων επιβίωσης των ατόμων παραβιάζει την αρχική υπόθεση του μοντέλου του Cox. Οι μονομεταβλητές κατανομές δεν μπορούν να περιγράψουν ολοκληρωμένα αυτή την κατάσταση και γι αυτό το λόγο υπάρχει μεγάλη ευελιξία στη στατιστική βιβλιογραφία για τη μοντελοποίηση τέτοιων πολυμεταβλητών δεδομένων (βλ. Mohammad Ehsanul Karim, 2008).

Τα μοντέλα επιβίωσης για εξαρτημένους χρόνους ζωής είναι χρήσιμα διότι μας οδηγούν σε ποιο εξελιγμένα ερωτήματα σε σχέση με τη γήρανση, της ασθένειες και τη μελέτη θνησιμότητας. Η πρώτη σημαντική προσέγγιση είναι τα από κοινού μοντέλα ευπάθειας (shared frailty models). Σε ένα τέτοιο μοντέλο η ευπάθεια ορίζεται ως ένα μέτρο σχετικού κινδύνου και είναι κοινός για όλα τα άτομα της

ομάδας. Στο επόμενο κεφάλαιο που ακολουθεί, παρουσιάζεται και ερμηνεύεται αναλυτικά ολόκληρη η περιγραφή του από κοινού μοντέλου ευπάθειας σε σχέση με τις κατανομές και τη στατιστική ερμηνεία.

1.7.2.1 Το από κοινού Μοντέλο Ευπάθειας

Κατά τον Hougaard (2000), τα από κοινού μοντέλα ευπάθειας (Shared frailty models), αποτελούν ένα είδος του κοινού μοντέλου κινδύνου όπου ο μη παρατηρούμενος κίνδυνος είναι κοινός για κάθε άτομο μέσα σε μια ομάδα, που μπορεί να αφορά μια ειδική περίπτωση, ή και ένα επαναλαμβανόμενο γεγονός. Στη περίπτωση αυτή λοιπόν, ο πληθυσμός χωρίζεται σε διάφορες ομάδες και μοιράζεται την ίδια ευπάθεια. Αυτός είναι ο λόγος για τον οποίο το μοντέλο αυτό ονομάστηκε το από Κοινού Μοντέλο Ευπάθειας.

Ως μια επέκταση της πολυμεταβλητής περίπτωσης, αν η από κοινού ευπάθεια θεωρηθεί γνωστή για μια ομάδα, τότε οι χρόνοι αποτυχίας υπό όρους είναι ανεξάρτητοι. Δηλαδή, γνωρίζοντας την ομάδα των ατόμων που μοιράζονται τη ίδια ευπάθεια και τον ίδιο κίνδυνο ακόμα και αν τα άτομα είναι ευπαθείς σε μια ασθένεια, τότε οι χρόνοι των γεγονότων είναι ανεξάρτητοι και ο αριθμός των παρατηρήσεων θεωρείται γνωστός.

Οι μέθοδοι της ανάλυσης επιβίωσης χρησιμοποιούνται στο μοντέλο μέχρις ότου συμβεί ένα γεγονός (time-to-event data). Ο χρόνος μέχρι το γεγονός ή ο κίνδυνος εκδήλωσης του γεγονότος μπορεί να μοντελοποιηθεί ως μεταβλητή απόκρισης. Σ' ένα μοντέλο κινδύνου, ο κίνδυνος μπορεί να καθοριστεί είτε σε πολλαπλασιαστική είτε σε προσθετική μορφή. Το μοντέλο αναλογικού κινδύνου του Cox (Cox model) είναι το πιο δημοφιλές μοντέλο επιβίωσης διότι έχει ευρεία ευελιξία. Το κοινό μοντέλο του Cox δεν λαμβάνει υπόψη τη μη μετρίσιμη μεταβλητότητα μεταξύ των ατόμων πέρα από αυτών των μετρίσιμων συμμεταβλητών. Εάν υπάρχει υποψία ετερογένειας μεταξύ των μετρίσιμων συμμεταβλητών τότε υπάρχει ένα άλλο μοντέλο που μπορεί να λάβει υπόψη του αυτή τη μεταβλητότητα. Τα μοντέλα ευπάθειας είναι μια πιθανή προέκταση του μοντέλου του Cox ή άλλων μοντέλων επιβίωσης που μας επιτρέπει τέτοια εξάρτηση. Τα μοντέλα ευπάθειας είναι ουσιαστικά μοντέλα επιβίωσης με σταθερές ή τυχαίες επιδράσεις. Παρόλο που οι σταθερές επιδράσεις περιλαμβάνουν το παρατηρήσιμο μέρος του μοντέλου, οι τυχαίες επιδράσεις αφορούν την ανεξήγητη μεταβλητότητα του μοντέλου. Με άλλα λόγια οι τυχαίες

επιδράσεις ή αλλιώς οι ευπάθειες, μοντελοποιούν την ανεξήγητη ετερογένεια του μοντέλου. Ο όρος ευπάθεια αναπτύχθηκε αρχικά για να περιγράψει την ετερογένεια σε ατομικό επίπεδο αλλά επεκτάθηκε για να περιγράψει επίσης και την ετερογένεια μεταξύ των ομάδων των ατόμων ή μέσα σε ένα άτομο. (Vapuel και άλλοι, 1979). Επομένως, ένα άτομο μπορεί να αποτελέσει ένα επίπεδο ομαδοποίησης (Duchateau L, Janssen P. 2008).

1.7.2.2 Το Συσχετισμένο Μοντέλο ευπάθειας (The Correlated Frailty Model)

Τα περισσότερα συσχετισμένα μοντέλα που έχουν αναπτυχθεί έως τώρα, είναι διμεταβλητά μοντέλα, αναλύουν δηλαδή δεδομένα διμεταβλητών χρόνων αποτυχίας, και έχουν εφαρμογή σε ευπάθειες που αφορούν ζευγάρια. Πράγματι, τα μοντέλα αυτά επεκτείνουν την ιδέα της ατομικής ευπάθειας στη περίπτωση των δυο μεταβλητών και συμπεριλαμβάνουν το από κοινού μοντέλο ευπάθειας σαν μια ειδική περίπτωση. Η δυσκολία αλλά και η πρωτοτυπία του μοντέλου αυτού είναι ότι, τα συσχετιζόμενα άτομα έχουν μεν διαφορετικές αλλά και εξαρτημένες ευπάθειες. Τέτοιες ευπάθειες, συνήθως αντιμετωπίζονται, με ανεξάρτητες συμμεταβλητές που η μία είναι κοινή και για τις δυο ευπάθειες.

ΚΕΦΑΛΑΙΟ 2

ΤΑ ΑΠΟ ΚΟΙΝΟΥ ΜΟΝΤΕΛΑ ΕΥΠΑΘΕΙΑΣ

2.1 Εισαγωγή

Όπως έχει είδη αναφερθεί στο προηγούμενο κεφάλαιο, το από κοινού μοντέλο ευπάθειας (the shared frailty model) αποτελεί ένα είδος μοντέλων κινδύνου και έχει σχέση με τους χρόνους γεγονότων συσχετιζόμενων ατόμων και με επαναλαμβανόμενες παρατηρήσεις. Σε αυτό το κεφάλαιο θα αναφερθεί η συμβολή του μοντέλου ευπάθειας μόνο σε παράλληλα δεδομένα, δηλαδή άτομα που ανήκουν στην ίδια ομάδα παρουσιάζουν τον ίδιο κίνδυνο. Είναι ένα μεικτό μοντέλο διότι στις περισσότερες περιπτώσεις τα κοινά ρίσκα θεωρούνται τυχαία. Με τον όρο μεικτό εννοούμε τη ευπάθεια και γι αυτό χρησιμοποιούμε τον συμβολισμό Y. Το από κοινού μοντέλο ευπάθειας υποθέτει ότι όλες οι παρατηρήσεις είναι χρονικά ανεξάρτητες δεδομένου της ευπάθειας. Με άλλα λόγια είναι ένα υπό όρους ανεξάρτητο μοντέλο. Η τιμή του όρου της ευπάθειας είναι σταθερή κατά την διάρκεια του χρόνου και κοινή για όλα τα άτομα στην ομάδα, και έτσι είναι υπεύθυνη για την δημιουργία εξάρτησης μεταξύ των χρόνων των γεγονότων σε μια ομάδα. Αυτός είναι ο λόγος για τον οποίο το μοντέλο ονομάστηκε το από κοινού μοντέλο ευπάθειας. Η ερμηνεία του μοντέλου αυτού είναι ότι η μεταβλητότητα μεταξύ των ομάδων οδηγεί σε διαφορετικούς κινδύνους και έτσι φαίνεται να υπάρχει μια εξάρτηση μεταξύ των ομάδων.

Το από κοινού μοντέλο ευπάθειας είναι ένα μοντέλο τυχαίων επιδράσεων αποτελούμενο από δύο πηγές διακυμάνσεων. Η μία εκφράζει την ευπάθεια (μεταβλητή Y) ενώ η μ(t) είναι η απλή τυχαία μεταβλητή που περιγράφεται από τη συνάρτηση διακινδύνευσης. Οι χρόνοι των γεγονότων μεταξύ των ομάδων θεωρούνται ότι είναι ανεξάρτητοι (διαφορετικές τιμές για τα i). Η εξάρτηση όμως μεταξύ των χρόνων για την ίδια τιμή i παρουσιάζουν την ίδια ευπάθεια Y_i του Y. Ο αριθμός k των παρατηρήσεων σε μια ομάδα θεωρείται γνωστός (Hougaard P, 2000).

Τα από κοινού μοντέλα ευπάθειας έχουν εφαρμογή σε συνεχείς κατανομές οι οποίες περιγράφονται αναλυτικά σε αυτό το κεφάλαιο. Η κατανομή γάμμα είναι η πιο διαδεδομένη για τα μοντέλα αυτά. Ωστόσο, όσον αφορά τη παλινδρόμηση η κατανομή γάμμα υστερεί της θετικής σταθερής κατανομής (positive stable distribution). Στο σημείο αυτό είναι σημαντικό να αναφέρουμε ότι, οι περισσότεροι υπολογισμοί βασίζονται στον μετασχηματισμό Laplace. Επιπλέον, χρησιμοποιώντας μια μεγαλύτερη γκάμα οικογενειών κατανομών επιτυγχάνεται καλύτερη προσαρμογή του μοντέλου. Πιο συγκεκριμένα, η PVF κατανομή που αναφέρεται αναλυτικά σε αυτό το κεφάλαιο, έχει μια παραπάνω παράμετρο σε σχέση με τη Γάμμα κατανομή. Τέλος, σε αυτό το κεφάλαιο γίνεται αναφορά σε διάφορες στατιστικές μεθόδους για τα μοντέλα ευπάθειας, όπως ο ΕΜ αλγόριθμος, η ποινικοποιημένη μέθοδος πιθανοφάνειας, κ.α.

2.2 Η περιγραφή του μοντέλου με δεσμευμένη παραμετροποίηση

Στην ανάλυση επιβίωσης το από κοινού μοντέλο ευπάθειας ορίζεται ως εξής: υποθέτουμε ότι υπάρχουν n ομάδες και η i-οστή ομάδα έχει k_i παρατηρήσεις με μη παρατηρούμενες ευπάθειες Y_{ij} $(1 \le i \le n)$. Το διάνυσμα D_{ij} $(1 \le i \le n, 1 \le j \le k)$ περιέχει τη πληροφορία των συμμεταβλητών των χρόνων εμφάνισης των γεγονότων $\mathrm{T}_{ij}, i = 1,....,n,\ j = 1,....,k$ της j-οστής παρατήρησης στην i ομάδα (Roberto G. Gutierrez. 2002, Virginie και άλλοι 2012). Δεδομένου του όρου ευπάθεια Y_i οι χρόνοι επιβίωσης θεωρούνται ανεξάρτητοι και η συνάρτηση κινδύνου για τη j παρατήρηση είναι της μορφής

$$Y\mu_j(t) \qquad (2.2.1)$$

Όπου $\mu_j(t)$ η συνάρτηση κινδύνου και Y η κοινή ευπάθεια για τα άτομα μιας ομάδας. Οι ευπάθειες Y_i θεωρούνται ότι είναι ανεξάρτητες και ισόνομες κατανεμημένες τυχαίες μεταβλητές .

Η ανεξαρτησία των χρόνων των γεγονότων ανάμεσα στις ομάδες αντιστοιχούν σε μια εκφυλισμένη κατανομή ευπάθειας. Στις υπόλοιπες περιπτώσεις η κατανομή είναι θετική. Επομένως, η διμεταβλητή συνάρτηση επιβίωσης είναι

$$S(t_1,t_2 \setminus Y) = \exp\left[-Y\left\{M_1(t_1)+M_2(t_2)\right\}\right] \qquad (2.2.2.)$$

όπου ,

$M_j(t) = \int_0^t \mu_j(u)\,du, j=1,2$ η αθροιστική αναφορική συνάρτηση κινδύνου.

Από τον παραπάνω τύπο παίρνουμε τη μη δεσμευμένη συνάρτηση επιβίωσης για δυο μεταβλητές

$$S(t_1,t_2) = E\exp\left[-Y\left\{M_1(t_1)+M_2(t_2)\right\}\right] = L\left(M_1(t_1)+M_2(t_2)\right) \quad (2.2.3)$$

Όπου ο L(s) είναι ο μετασχηματισμός Laplace .

Θέτουμε $M.(t_1,...,t_k) = \sum_{j=1}^{k} M_j(t_j)$, επομένως η πολυμεταβλητή έκφραση για τις k παρατηρήσεις είναι

$$S(t_1,...,t_k) = L\left(M.(t_1,...,t_k)\right) \quad (2.2.4)$$

Επομένως, η πυκνότητα με βάση την (2.2.4) είναι

$$(-1)^k \left\{\prod_{j=1}^{k} \mu_j(t_j)\right\} L^{(k)}\left(M.(t_1,...,t_k)\right)$$

Η σχέση που ακολουθεί ενσωματώνεται στη συνάρτηση πιθανοφάνειας και είναι αυστηρά ορισμένη σε σχέση με τη συνάρτηση κινδύνου, που αφορούν πραγματικά γεγονότα, και τη χρήση του μετασχηματισμού Laplace .

$$(-1)^D \left\{\prod_{j=1}^{k} \mu_j(t_j)^{D_j}\right\} L^{(D.)}\left(M.(t_1,...,t_k)\right) \quad (2.2.5)$$

Όπου $D. = \sum D_j,\ j=1,....,k$.

2.3 Περιθώρια παραμετροποίηση

Ένας εναλλακτικός τρόπος προσαρμογής του μοντέλου είναι μέσω της περιθώριας κατανομής (marginal distribution) η οποία καθιστά εύκολη τη διαδικασία εκτίμησης παραμέτρων.

Επομένως, η διμεταβλητή συνάρτηση επιβίωσης $S_j(t) = \Pr(\mathrm{T}_j > 0)$ με βάση τους περιορισμούς $S_1(t) = S(t,0)$ και $S_2(t) = S(0,t)$ και $\Omega_j(t)$ τον αθροιστικό κίνδυνο, είναι η (2.3.1)

$$S(t_1,t_2) = L\left(L^{-1}\left(S_1(t_1)\right) + L^{-1}\left(S_2(t_2)\right)\right) \quad (2.3.1)$$

Όπου, $S_j(t) = L\left(M_j(t)\right)$ (2.3.2)

με $S_j(t_j) = \exp\left\{-\Omega_j(t_j)\right\}$ και $M_j(t) = L^{-1}\left(S_j(t)\right)$.

Η πολυμεταβλητή έκφραση της παραπάνω σχέσης είναι η ακόλουθη

$$S(t_1,...,t_k) = L\left(\sum_j L^{-1}\left(S_j(t_j)\right)\right)$$

Και θέτοντας $S_j(t_j) = \exp\left\{-\Omega_j(t_j)\right\}$ προκύπτει η σχέση ,

$$S(t_1,...,t_k) = L\left[\sum_j L^{-1}\left(\exp\left\{-\Omega_j(t_j)\right\}\right)\right] \quad (2.3.3)$$

Η εξίσωση αυτή μπορεί να διαφοροποιηθεί για να δώσει τη πολυμεταβλητή συνάρτηση πυκνότητας πιθανότητας χρησιμοποιώντας για μια συνάρτηση $df^{-1}(x)/dx = 1/(df/dy)$ που είναι το παράγωγο της αντίστροφης συνάρτησης και ισούται με την αντίστροφη συνάρτηση της παραγώγου, αξιολογούμενη στο αντίστοιχο σημείο. Για ευκολία η παραπάνω έκφραση μπορεί να γραφεί ως $\left(f^{-1}\right)'(x) = 1/f'\left(f^{-1}(x)\right)$.

Επομένως, η πολυμεταβλητή συνάρτηση πυκνότητας είναι

$$\left\{\prod_j\left[\omega_j(T_j)S_j(T_j)/L'\left(L^{-1}\left(S_j(T_j)\right)\right)\right]\right\}L^{(k)}\left[\sum_j L^{-1}\left(S_j(T_j)\right)\right] \quad (2.3.4)$$

Η συνάρτηση πυκνότητας πιθανότητας για λογοκριμένα δεδομένα (μια άλλη εκδοχή της παραπάνω εξίσωσης) , ικανοποιεί τη σχέση

$$\left(\prod_j \left[\omega_j\left(T_j\right)S_j\left(T_j\right)/L'\left(L^{-1}\left(S_j\left(T_j\right)\right)\right)\right]^{D_j}\right)L^{(D.)}\left[\sum_j L^{-1}\left(S_j\left(T_j\right)\right)\right] \quad (2.3.5)$$

2.3.1 Μοντέλο Weibull για δεσμευμένες κατανομές

Το μοντέλο του Weibull χρησιμοποιείται τόσο για τα μοντέλα αναλογικού κινδύνου όσο και για τα μοντέλα επιταχυνόμενου χρόνου αποτυχίας . Η δεσμευμένη συνάρτηση κινδύνου είναι

$$Y\lambda_j\gamma t^{\gamma-1}.$$

Επομένως, στη συμμετρική διμεταβλητή περίπτωση, θεωρείται ότι τα W_1, W_2 είναι ανεξάρτητα για Weibull(1,γ).

Οι χρόνοι αποτυχίας δίνονται από την παρακάτω εξίσωση

$$T_j = Y^{-1/\gamma}W_j,\ j=1,2. \quad (2.3.1.1)$$

Ως γενικό αποτέλεσμα μπορούμε να αξιολογήσουμε ότι δεδομένου του Y, ο μέσος είναι $Y^{-1/\gamma}\Gamma\left(1+2/\gamma\right)$ με $c_1\left(\gamma\right)Y^{-1/\gamma}$. Από αυτό προκύπτει ότι ο μη δεσμευμένος μέσος είναι $c_1\left(\gamma\right)E\left(Y^{-1/\gamma}\right)$.

Ωστόσο, η δεσμευμένη διασπορά είναι η $Y^{-2/\gamma}\left\{\Gamma\left(1+2/\gamma\right)-\Gamma\left(1+1/\gamma\right)^2\right\}$ με $c_2\left(\gamma\right)Y^{-2/\gamma}$ και έχουμε τη μη δεσμευμένη διασπορά

$$Var\left(T\right)=c_1\left(\gamma\right)^2 Var\left(Y^{-1/\gamma}\right)+c_2\left(\gamma\right)E\left(Y^{-2/\gamma}\right) \quad (2.3.1.2)$$

Και η συσχέτιση μεταξύ των ατόμων που μοιράζονται κοινή ευπάθεια Y, δίνεται από τον παρακάτω τύπο

$$corr\left(T_1,T_2\right)=\left\{c_1\left(\gamma\right)^2 Var\left(Y^{-1/\gamma}\right)\right\}/Var\left(T\right) \quad (2.3.1.3)$$

Παρόμοιες εξισώσεις με τις παραπάνω παίρνουμε για τους λογαρίθμους των χρόνων T_1, T_2

$$Var\left(\log \mathrm{T}\right)=\left\{Var\left(\log Y\right)+\pi^2/6\right\}/\gamma^2 \quad (2.3.1.4)$$

$$corr\left(\log \mathrm{T}_1, \log \mathrm{T}_2\right) = Var\left(\log Y\right) / \left\{Var\left(\log Y\right) + \pi^2 / 6\right\} \quad (2.3.1.5)$$

2.4 Γάμμα μοντέλο ευπάθειας (Gamma Distribution frailty model)

Η κατανομή Γάμμα χρησιμοποιείται εδώ και πολλά χρόνια για την παραγωγή συνδυασμού του εκθετικού μοντέλου (Exponential model) και του μοντέλου Poisson. Από υπολογιστική άποψη τα δυο μοντέλα χρησιμοποιούνται ως μοντέλα επιβίωσης γιατί είναι εύκολο να αντλούμε τύπους για οποιοδήποτε αριθμό συμβάντων και αυτό λόγω της απλότητας των παραγώγων του μετασχηματισμού Laplace. Δεν είναι τυχαίο εξάλλου το γεγονός ότι αυτή η κατανομή έχει χρησιμοποιηθεί στις περισσότερες από τις εφαρμογές που έχουν δημοσιευθεί έως τώρα. Χρησιμοποιούμε τη τυποποιημένη κατανομή Γάμμα, γάμμα (δ,θ) με παράμετρο δ και (αντίστροφη-inverse) παράμετρο κλίμακας (scale parameter) θ. Για τους περισσότερους υπολογισμούς, περιορίζουμε τη παράμετρο κλίμακα και ο τυπικός περιορισμός είναι θ=δ δεδομένου ότι αυτό συνεπάγεται μια μέση τιμή του 1 για το Y. Ωστόσο, κάποιες άλλες φόρμουλες γίνονται λιγότερο διαφανείς χρησιμοποιώντας αυτόν τον περιορισμό και ως εκ τούτου θα τον χρησιμοποιούμε μόνο σε προχωρημένο στάδιο και ανεξάρτητα από τη διάρκεια ζωής όταν η κατανομή εκφυλίζεται κάτω όταν το όριο $\delta \to \infty$. Εναλλακτικά, θα μπορούσε κανείς να παραμετρικοποιήσει μέσω της διακύμανσης του Y, η οποία είναι 1/δ. Στη βιβλιογραφία, αυτό το μοντέλο ονομάζεται Clayton (1978) ή Clayton-Oakes (1982) model. Σε αυτό το κεφάλαιο θα το ονομάζουμε γάμμα μοντέλο ευπάθειας.

Η δισδιάστατη συνάρτηση επιβίωσης μπορεί να γραφεί ως,

$$S\left(t_1, t_2\right) = \theta^{\delta} / \left\{\theta + \mathrm{M}_1\left(t_1\right) + M_2\left(t_2\right)\right\}^{\delta}, \quad (2.4.1)$$

ακόλουθη της εξίσωσης (2.2.3). Από αυτό, μπορούμε να εξάγουμε την αντίστροφη σχέση

$$\mathrm{M}_1\left(t_1\right) = \theta\left\{S_1\left(t_1\right)^{-1/\delta} - 1\right\}, \quad (2.4.2)$$

όπου $S_1(t_1) = S(t_1, 0)$ είναι η περιθώρια συνάρτηση επιβίωσης (marginal survival function). Αυτό δίνει μια εναλλακτική έκφραση και η ακόλουθη εξίσωση είναι

$$S(t_1, t_2) = \left\{ S_1(t_1)^{-1/\delta} + S_2(t_2)^{-1/\delta} - 1 \right\}^{-\delta} \quad (2.4.3)$$

Με αυτή την εξίσωση επιλύεται αυτόματα το πρόβλημα κλίμακας ταυτοποίησης παραλείποντας τη παράμετρο κλίμακα (scale parameter) θ και εμπεριέχοντας μόνο το δ.

Η πολυμεταβλητή γενίκευση της εξίσωσης αυτής είναι ,

$$S(t_1, ..., t_k) = \left\{ \sum_j S_j(t_j)^{-1/\delta} - (k-1) \right\}^{-\delta} . \quad (2.4.4)$$

Στη δισδιάστατη περίπτωση, η πυκνότητα του χρόνου των γεγονότων είναι

$$\mu_1(t_1)\mu_2(t_2)\left\{\theta + \mathrm{M}_1(t_1) + \mathrm{M}_2(t_2)\right\}^{-\delta-2} \theta^{\delta}(\delta+1)\delta . \quad (2.4.5)$$

Για γενικούς υπολογισμούς χρειαζόμαστε τη συμβολή της πιθανότητας για αυθαίρετα λογοκριμένα δεδομένα. Η μορφή της εξίσωσης σε αυτή τη περίπτωση είναι

$$\left\{ \prod_{j=1}^{k} \mu_j(T_j)^{D_j} \right\} \theta^{\delta} (\theta + \mathrm{M}.)^{-\delta - D.} \Gamma(\delta + d.) / \Gamma(\delta) \quad (2.4.6)$$

Όπου $M. = M.(T_1 ..... , T_k)$. Ένας εναλλακτικός τύπος προέρχεται από τη παράγωγο της συνάρτησης και είναι ο ακόλουθος

$$\left\{ \prod_{j=1}^{k} \omega_j(t_j)^{D_j} \, S_j(T_j)^{-D_j/\delta} \right\} S(T_1 ..... , T_k)^{(\delta + D.)/\delta} \, \delta^{-D.} \Gamma(\delta + D.) / \Gamma(\delta) \quad (2.4.7)$$

Όπου $\omega_j(t)$ είναι ο κίνδυνος της περιθώρια κατανομής (marginal distribution).

Η παράμετρος τ του Kendall μπορεί να εκτιμηθεί σε $1/(1+2\delta)$. Η μέση αντιστοιχία είναι

$$\kappa = 4\left(2^{1+1/\delta}-1\right)^{-\delta}-1 \quad (2.4.8)$$

Η παράμετρος ρ του Spearman μπορεί να αξιολογηθεί με αριθμητική ολοκλήρωση ή μέσω του τύπου

$$\rho = \frac{12(\delta+1)}{(1+2\delta)^2}\,{}_3F_2\left(\delta+1,1,1,2(\delta+1),2(\delta+1),1\right) \quad (2.4.9)$$

2.4.1 Μοντέλα Weibull

Σε αυτή τη περίπτωση όπου η δεσμευμένη συνάρτηση κινδύνου $Y\lambda_j\gamma t^{\gamma-1}$ είναι της κατανομής Weibull για τη j-οστή παρατήρηση και Y ακολουθεί κατανομή γάμμα (δ,δ), η διμεταβλητή συνάρτηση επιβίωσης δίνεται από το τύπο

$$S(t_1,t_2) = 1/\left\{1+\left(\lambda_1 t_1^{\gamma}+\lambda_2 t_2^{\gamma}\right)/\delta\right\}^{\delta}. \quad (2.4.1.1)$$

Η παραπάνω συνάρτηση ονομάζεσαι Burr κατανομή (Burr distribution) γενικευμένη μορφή της κατανομής Pareto.

Η συσχέτιση μπορεί εύκολα να υπολογιστεί συνδυάζοντας τις αντίστοιχες εξισώσεις (2.3.1.2), (2.3.1.3) και είναι ανεξάρτητη του $\lambda_1 και \lambda_2$. Ισχύει για $\delta > 2/\gamma$ ενώ για $\delta < 2\gamma$ η διασπορά του T είναι άπειρη και η συσχέτιση απροσδιόριστη. Επομένως, η συσχέτιση των λογαρίθμων για τα T_1 και T_2 είναι ανεξάρτητη του δ και ισούται με

$$corr\left(\log T_1,\log T_2\right) = \psi'(\delta)/\left\{\psi'(\delta)+\pi^2/6\right\}. \quad (2.4.1.2)$$

2.5 Θετική και σταθερή κατανομή ευπάθειας (Positive Stable frailty Distributions)

Η οικογένεια των θετικών και σταθερών κατανομών δημιούργησαν μια ενδιαφέρουσα εναλλακτική κατανομή που επιφέρει καλύτερα αποτελέσματα συμπεριλαμβάνοντας και τις δεσμευμένες κατανομές. Παρ όλο αυτά ο μετασχηματισμός Laplace γίνεται πιο περίπλοκος. Σε αυτή τη περίπτωση δεν θα

συμπεριλάβουμε τη παράμετρο κλίμακα (scale parameter) γιατί ο τύπος για τον όποιο η παράμετρος είναι χρήσιμη έχει υποστεί μετατροπή όπως θα δούμε παρακάτω στην PVF κατανομή που περιγράφει τη γενικευμένη μορφή θετικής και σταθερής κατανομής. Ο μετασχηματισμός Laplace είναι $L(s)=\exp\left(-s^{a}\right)$, από την οποία λαμβάνουμε τη δισδιάστατη συνάρτηση επιβίωσης

$$S(t_1,t_2)=\exp\left[-\left\{\mathrm{M}_1(t_1)+\mathrm{M}_2(t_2)\right\}^{a}\right] \quad (2.5.1)$$

Η γενική έκφραση για τη συνάρτηση πιθανοφάνειας, χρησιμοποιώντας τους μετασχηματισμούς Laplace για δ=α και θ=0 , είναι

$$[\prod_{j}\mu_j(t_j)^{D_j}]Q\exp\left(-M.^{a}\right) \quad (2.5.2)$$

Όπου $Q=\sum_{m=1}^{D.}c_{D.,m}a^{m}M.^{ma-D}$. Οι συντελεστές c_{dm} εξαρτώνται από το α.

Η δεσμευμένη συνάρτηση επιβίωσης έχει ως εξής,

$$S(t_1,t_2)=\exp\left(-\left[\left\{-\log S_1(t_1)\right\}^{1/a}+\left\{-\log S_2(t_2)\right\}^{1/a}\right]^{a}\right) \quad (2.5.3)$$

Υπό την προϋπόθεση της ενσωμάτωσης του δεσμευμένου κινδύνου $\Omega_j(t)$, η έκφραση απλοποιείται ως,

$$S(t_1,t_2)=\exp\left[-\left\{\Omega_1(t_1)^{1/\alpha}+\Omega_2(t_2)^{1/\alpha}\right\}^{\alpha}\right] \quad (2.5.4)$$

Η πολυμεταβλητή πυκνότητα μέσω δεσμευμένης κατανομής είναι

$$\left[\prod_{j=1}^{k}\left\{\omega_j(t_j)\Omega(t_j)^{\phi-1}\right\}^{d_j}\right]Q\exp\left(-M.^{a}\right) \quad (2.5.5)$$

Όπου $\phi = 1/\alpha$, $\mathrm{M.} = \sum \Omega_j \left(t_j\right)^{\phi}$ και το Q είναι όπως ορίζεται παραπάνω.

Η παράμετρος τ του Kendall απλοποιείται ως 1-α. η μέση αντιστοιχία είναι $\kappa = 2^{2-2^{\alpha}} - 1$.

2.5.1 Το σταθερό μοντέλο Weibull

Το μοντέλο του Weibull για τη θετική σταθερή κατανομή έχει ιδιαίτερα καλή εφαρμογή. Η διμεταβλητή συνάρτηση επιβίωσης είναι η ακόλουθη

$$S\left(t_1, t_2\right) = \exp\left\{-\left(\varepsilon_1 t_1^{\gamma} + \varepsilon_2 t_2^{\gamma}\right)^{\alpha}\right\} \quad (2.5.1.1)$$

όπου μέσω της $\mu_j\left(t\right) = \varepsilon_j \gamma t^{\gamma-1}$ δίνεται το διμεταβλητό μοντέλο του Weibull . Αυτό σημαίνει ότι δεδομένου του Y, η κατανομή του T_j είναι Weibull$\left(\varepsilon_j Y, \gamma\right)$.

Το πλεονέκτημα αυτού του μοντέλου είναι ότι οι περιθώριες κατανομές είναι σε μορφή Weibull. Η περιθώρια κατανομή ακολουθεί Weibull$\left(\varepsilon_j^a, a\gamma\right)$ η οποία παραμετρικοποιείται ως Weibull$\left(\omega_j, \rho\right)$. Επομένως, για να εκφραστεί η διμεταβλητή κατανομή μέσω της περιθώριας, χρησιμοποιείται η παραμετρικοποίηση $\left(\omega_1, \omega_2, \rho\right)$. Η αλλαγή της παραμέτρου από γ σε ρ= αγ αντιστοιχεί στην αυξημένη διακύμανση της περιθώρια σε σχέση με την δεσμευμένη κατανομή.

Η συνάρτηση πυκνότητας πιθανότητας $f\left(t_1, t_2\right)$ για δυο μεταβλητές δίνεται από τον τύπο

$$\rho^2 \omega_1^{\phi} \omega_2^{\phi} t_1^{\rho\phi-1} t_2^{\rho\phi-1} \left[M.^{2(\alpha-1)} + \left(\phi - 1\right) M.^{(\alpha-2)}\right] \exp\left(-M.^{\alpha}\right) \quad (2.5.1.2)$$

Όπου $M. = \sum \omega_j^{\phi} t_j^{\rho\phi}$. Ενώ η συνάρτηση πυκνότητας για τη συμμετρική διμεταβλητή κατανομή με ρ=1 είναι

$$\omega^2 t_1^{\phi-1} t_2^{\phi-1} \left(t_1^{\phi} + t_2^{\phi}\right)^{2(a-1)} \left\{1 + \omega^{-1}\left(\phi - 1\right)\left(t_1^{\phi} + t_2^{\phi}\right)^{-a}\right\} \exp\left\{-\omega\left(t_1^{\phi} + t_2^{\phi}\right)^{a}\right\} \quad (2.5.1.3)$$

Και η συσχέτιση δίνεται από τον τύπο

$$corr\left(T_1,T_2\right)=1-h\left(1/\gamma\right)h\left(\alpha/\gamma\right),\quad (2.5.1.4)$$

όπου $h\left(x\right)=1-\Gamma\left(1+x\right)^2/\Gamma\left(1+2x\right)$ εξαρτάται από το γ και όχι από το $\left(\varepsilon_1,\varepsilon_2\right)$.

2.6 PVF οικογένεια κατανομών ευπάθειας (Power Variance Function Distributions)

Η PVF οικογένεια κατανομών αποτελεί μέρος της εκθετικής οικογένειας κατανομών όπου η διακύμανση είναι η δύναμη του μέσου της συνάρτησης. Αυτός είναι ο λόγος για τον οποίο οι PVF κατανομές ονομάστηκαν (Power Variance Function Distributions ή power variance function). Σε ειδικές περιπτώσεις περιλαμβάνει τη γενικευμένη Γάμμα κατανομή (generalized Gamma distribution), την θετική σταθερή κατανομή (positive stable distributions) και την αντίστροφη Γκαουσιανή κατανομή (inverse Gaussian distributions). Εμπεριέχει τρεις παραμέτρους α, δ και θ με $0<a\leq 1$. Άμεση συνέπεια αυτής της οικογένειας κατανομών είναι η εκθετική κατανομή που παράγεται από την θετική σταθερή κατανομή σε συνδυασμό με τη περίπτωση μη συμπεριλαμβανομένων μεταβλητών σε μια αλλαγή της τιμής παραμέτρου στην οικογένεια εκθετικών κατανομών παραγόμενη από την αρχική κατανομή. Στην ειδική περίπτωση όπου α=1/2 και θ=0 έχουμε το θετικό σταθερό μοντέλο (positive stable model). Για α=0, παίρνουμε τη Γάμμα κατανομή με την ίδια παραμετροποίηση(parameterization) αυτής της κατανομής. Για α=1/2, η μεικτή κατανομή (mixture) στην αντίστροφη γκαουσιανή κατανομή μπορεί να απλοποιηθεί. Για α<0 παίρνουμε κατανομή με σημειακή μάζα το μηδέν υποδηλώνοντας ότι κάποια γκρουπ έχουν μηδενικό ρίσκο. Στην περίπτωση του χρόνου ζωής (lifetime) αυτό είναι απίθανο διότι θα αντιστοιχούσε σε πρόσωπα που παραμένουν αθάνατα , αλλά σε άλλους χρόνους είναι αρκετά πιθανό να έχουν αυτά τα γκρουπ μηδενικό ρίσκο. Με αυτόν τον τρόπο οι ομάδες των ατόμων δεν θα μπορούσαν να βιώσουν ένα κρίσιμο γεγονός. Η μια από τις τρείς παραμέτρους είναι κατ ουσίαν η παράμετρος κλίμακας. Όταν θα πρέπει να περιορίσουμε τις παραμέτρους , θα θεωρούμε μια μέση τιμή του 1. Καθώς ο μέσος στη PVF οικογένεια κατανομών είναι $EY=\delta\theta^{\alpha-1}$, υπό τον περιορισμό $\delta=\theta^{1-\alpha}$. Όταν χρησιμοποιούμε αυτόν τον περιορισμό, παίρνουμε μια καινούρια παράμετρο, η, με

$\delta = n^{1-\alpha}$ και θ= η. Η τυπική κατανομή έχει μέση τιμή 1 και διασπορά (1-α)/ η. για δοθείσα τιμή των (α,δ,θ) μπορούμε να έχουμε μια τυποποιημένη κατανομή ως κλίμακα μετασχηματισμού της αρχικής κατανομής καθορίζοντας το

$$\eta = \delta\theta^{a} \qquad (2.6.1)$$

Όταν το Y ακολουθεί τη PVF(α,δ,θ) κατανομή, η κατανομή του Y/EY ακολουθεί $PVF\left(a,\left(\delta\theta^{\alpha}\right)^{1-\alpha},\delta\theta^{\alpha}\right)$, δηλαδή $PVF\left(a,\eta^{1-a},\eta\right)$. Η κατανομές με άπειρο μέσο (σταθερή κατανομή με θ=0) παραλείπονται.

Το μοντέλο αυτό εκτείνεται τόσο στο σταθερή όσο και στη γάμμα κατανομή και έτσι είναι χρήσιμο και για τα δυο. Μπορεί επίσης να χρησιμοποιηθεί και ως ένας ευέλικτος τρόπος για να περιγράψει την ανεξαρτησία. Παρόλο αυτά, επειδή υπάρχουν δυο παράμετροι που περιγράφουν την εξάρτηση, δεν υπάρχει νόημα να εξεταστεί η υπόθεση της μη ανεξαρτησίας. Ειδικότερα όταν υπάρχει μικρή εξάρτηση, είναι δύσκολο να προσδιορίσουμε και τις δύο παραμέτρους και μόνο ο βαθμός της εξάρτησης μπορεί να καθοριστεί με λογική ακρίβεια, όπως μετριέται από τη παράμετρο τα του Kendall και ρ του Spearman.

Η διμεταβλητή συνάρτηση επιβίωσης είναι

$$S\left(t_1,t_2\right) = \exp\left[-\delta\left\{\theta + \mathrm{M}_1\left(t_1\right) + \mathrm{M}_2\left(t_2\right)\right\}^{\alpha}/\alpha + \delta\theta^{a}/\alpha\right] \quad (2.6.2)$$

Ο ολοκληρωμένος κίνδυνος υπολογίζεται από τη περιθώρια κατανομή με τη βοήθεια του τύπου

$$\mathrm{M}_j\left(t\right) = \left[\theta^{\alpha} - \alpha\left\{\log S_j\left(t\right)\right\}/\delta\right]^{1/a} - \theta$$

Η διμεταβλητή συνάρτηση επιβίωσης , εκφρασμένη από τη περιθώρια κατανομή είναι

$$\exp\left(-\left[\left\{\eta/\alpha - \log S_1\left(t\right)\right\}^{1/a} + \left\{\eta/\alpha - \log S_2\left(t\right)\right\}^{1/a} - \left(\eta/\alpha\right)^{1/a}\right]^{a} + \eta/\alpha\right) \quad (2.6.3)$$

Με αυτόν τον τρόπο η παράμετρος κλίμακας δεν εμφανίζεται στη εξίσωση . Αυτός ο τύπος είναι ελαφρώς απλοποιημένος όταν εκφράζεται μέσω της περιθώρια συνάρτησης κινδύνου ή αλλιώς συνάρτησης διακινδύνευσης

$$S(t_1,t_2)=\exp\left(-\left[\left\{\eta/\alpha+\Omega_1(t_1)\right\}^{1/a}+\left\{\eta/\alpha+\Omega_2(t_2)\right\}^{1/a}-(\eta/\alpha)^{1/a}\right]^a+\eta/\alpha\right) \quad (2.6.4)$$

Για περισσότερες από δυο μεταβλητές ο τύπος είναι

$$S(t_1,...,t_k)=\exp\left\{-\left[\sum_j\left\{\eta/\alpha+\Omega_j(t_j)\right\}^{1/a}-(k-1)(\eta/\alpha)^{1/a}\right]^a+\eta/\alpha\right\}$$

Τα μέτρα της εξάρτησης μπορούν να εκτιμηθούν όταν το μοντέλο είναι συνεχές με $a\geq 0$. Όταν το $a\geq 0$ η παράμετρος τα του Kendall δίνεται από την εξίσωση

$$\tau=(1-a)-2\eta+\left(4\eta^2/\alpha\right)\exp(2\eta/\alpha)\,\mathrm{E}_{(1/\alpha)-1}(2\eta/\alpha) \quad (2.6.5)$$

Όπου $\mathrm{E}_m(x)$ είναι το γενικευμένο εκθετικό ολοκλήρωμα. Στη περίπτωση όπου το $a=1/2$, έχουμε το απλό εκθετικό ολοκλήρωμα. Η μέση αντιστοιχία είναι

$$\kappa=4\exp\left[-\left\{2(\eta/\alpha+log2)^{1/a}-(\eta/\alpha)^{1/a}\right\}^a+\eta/\alpha\right]-1 \quad (2.6.6)$$

Για τη παράμετρο ρ του Spearman δεν υπάρχει συγκεκριμένος τύπος. Όταν το α είναι σταθερό υπάρχει ένα ανώτατο όριο στην εξάρτηση το οποίο αντιστοιχεί στο θετικό σταθερό όριο με θ=0. Έτσι, αν κάποιος θέλει να έχει ένα μοντέλο που να επιτρέπει οποιοδήποτε βαθμό εξάρτησης , η τιμή του α πρέπει να είναι ελεύθερη.

2.7 Λογάριθμο- κανονική κατανομή ευπάθειας (Lognormal frailty distribution)

Η λογάριθμο-κανονική κατανομή έχει επίσης χρησιμοποιηθεί ως κατανομή ευπάθειας. Σε αυτή τη περίπτωση, ο μετασχηματισμός Laplace είναι δυσεπίλυτος και ως εκ τούτου τα πιθανά αποτελέσματα πρέπει να αξιολογηθούν είτε μέσω μιας προσέγγισης είτε μιας αριθμητικής ολοκλήρωσης. Απλά ρητά αποτελέσματα για το μέτρο της εξάρτησης όπως η παράμετρος **τ** του Kendall και ρ του Spearman δεν είναι γνωστά. Ο μετασχηματισμός Laplace και τα παράγωγά του μπορούν να προσεγγιστούν , και αυτό καθιστά την αξιολόγηση της πιθανότητας αλλά και την

αξιολόγηση της παραμέτρου **τ** μέσω μιας μονοδιάστατης αριθμητικής ολοκλήρωσης χρησιμοποιώντας την εξίσωση $\tau = 4\int_0^\infty sL(s)L''(s)ds - 1$.

Αυτή η προσέγγιση δείχνει να είναι πολύ επιτυχημένη και ταυτόχρονα χρήσιμη και απαραίτητη. Εναλλακτικά, είναι πολύ πιθανή η αξιολόγηση των αποτελεσμάτων μέσω προσομοιώσεων. Ωστόσο, η ανάγκη για προσεγγίσεις επιφέρουν ορισμένους περιορισμούς στις διαθέσιμες διαδικασίες εκτίμησης. Επιπλέον, με μέση τιμή ξ και διακύμανση στη λογαριθμική κλίμακα σ^2, η μέση τιμή του Y είναι $\exp\left(\xi + \sigma^2/2\right)$ και η διακύμανση $\exp\left(2\xi + \sigma^2\right)\left\{\exp\left(\sigma^2\right) - 1\right\}$, από το οποία αντλούμε τον συντελεστή μεταβλητότητας ως $\left\{\exp\left(\sigma^2\right) - 1\right\}^{1/2}$. Ο φυσικός περιορισμός για την απόκτηση κλιμακωτής ταυτοποίησης είναι ξ=0, απλούστερη της περίπτωσης όπου EY=1.

Καθώς η διακύμανση του Y στη λογαριθμική κλίμακα είναι σ^2, μπορούμε να αντλήσουμε από την εξίσωση (2.3.1.5) ότι η συσχέτιση των λογαριθμικών τιμών στο μοντέλο του Weibull είναι

$$corr\left(\log \mathrm{T}_1, \log \mathrm{T}_2\right) = \sigma^2 / \left(\sigma^2 + \pi^2/6\right) \quad (2.7.1)$$

Ανεξάρτητα του λ_1, λ_2 και γ. Στη πραγματικότητα το μοντέλο Weibull της εξίσωσης (2.3.1.1). Μπορεί να απλοποιηθεί στη μορφή

$$\mathrm{T}_j = \tilde{Y} W_j \quad (2.7.2)$$

όπου $\tilde{Y} = Y^{-1/\gamma}$. Το $\tilde{Y}$ ακολουθεί λογάριθμο-κανονική κατανομή με παράμετρο $\tilde{\sigma}^2 = \sigma^{2/}\gamma^2$. Το πλεονέκτημα αυτής της αναδιατύπωσης είναι ότι μπορούμε να αγνοήσουμε τη σχέση μεταξύ των παραμέτρων της κατανομής του $\tilde{Y}$ και W_j, κάνοντας την πιο εύκολή και απλούστερη προς όφελος της εξίσωσης (2.7.2) ως ένα επιταχυνόμενο μοντέλο χρόνου αποτυχίας. Επιπλέον, οι χρόνοι μπορούν αμέσως να βρεθούν μέσω της εξίσωσης, γεγονός που είναι τόσο απλό που δεν είναι απαραίτητος ο περιορισμός του ξ. Για την q-στη στιγμή υπάρχει $q > -\gamma$, στη περίπτωση αυτή

$$ET^q = \exp\left\{-q\xi/\gamma + q^2\sigma^2/\left(2\gamma^2\right)\right\}\lambda^{-q/\gamma}\Gamma\left(1 + q/\gamma\right)$$

Καθώς η μέση ευπάθεια είναι πεπερασμένη, όλες οι παράμετροι μπορούν να προσδιοριστούν από μονομεταβλητά δεδομένα, όταν υπάρχουν διαθέσιμες συμμεταβλητές. Ωστόσο άγνωστο παραμένει η απόκλιση του οριακού κινδύνου πιθανότητας. Η λογάριθμο-κανονική κατανομή είναι πρακτικά πολύ κοντά στη

Γκαουσιανή κατανομή .παρουσιάζοντας αρκετές ομοιότητες. Παρόλο αυτά η χρήση της λογαριθμικής κατανομής επιφέρει ορισμένα πλεονεκτήματα σε σχέση με τη Γκαουσιανή κατανομή. Η κατανομή των μη παρατηρήσημων συμμεταβλητών, ω_t, της εξίσωσης $\mu_j(t)\exp(\beta' z_{ij} + \psi'\omega_t)$ μπορεί να είναι μια πολυμεταβλητή κατανομή έχοντας έναν αυθαίρετο πίνακα διακύμανσης , απλούστερος από κάθε άλλη περίπτωση. Οι κατανομές αυτές έχουν άμεση σχέση με τους αντίστοιχους συντελεστές παλινδρόμησης, είναι στενά συνδεδεμένες με την επίδραση των συμμεταβλητών και αποτελούν μέρος μιας δεδομένης οικογένειας κατανομών.
Όπως προκύπτει από το γενικό αποτέλεσμα , η αποκοπή αυτή οδηγεί σε μια αναβαθμισμένη κατανομή της εκθετικής οικογένειας κατανομών την λογάριθμο-κανονική εκθετική οικογένεια κατανομών $\mathrm{LNEF}\left(\sum_j \mathrm{M}_j(t), \xi, \sigma^2\right)$.

Τέλος, ένα άλλο πλεονέκτημα που βασίζεται στη κανονική κατανομή είναι ότι υπάρχουν εναλλακτικοί μέθοδοι εκτίμησης όπως η REML (restricted maximum likelihood) και η ποινικοποιημένη μέθοδος πιθανοφάνειας (Penalized Likelihood Method).

2.8 Άλλες προτεινόμενες κατανομές

1. Αντίστροφη Γκαουσιανή Κατανομή (Inverse Gaussian Distribution): χρησιμοποιήθηκε από τον Hougaard (1986b). Τα αποτελέσματα τις κατανομής αυτής είναι παρόμοια με αυτά της Λογάριθμο-κανονικής κατανομής Hougaard (2000).

2. Μεικτή κατανομή Poisson (Compound Poisson Distribution): προτάθηκε από τον, Aalen (1992) και εφαρμόζεται σε μονομεταβλητά δεδομένα για τον καρκίνο των όρχεων (Aalen and Tretli ,1999) με την ιδέα ότι ένα μέρος του πληθυσμού είναι ιδιαίτερα επιρρεπείς σε αυτή την ασθένεια.

3. Ομοιόμορφη Κατανομή (Uniform Distribution) : χρησιμοποιήθηκε από τους Lee και Klein (1988).

4. Οριακό Μοντέλο (Threshold Model): προτάθηκε από τους Lindley και Singpurwalla (1986).
5. Άλλες πιθανές οικογένειες κατανομών περιλαμβάνουν την οικογένεια κατανομών Franks (Genest, 1987).

Παρόλο αυτά, δεν προκύπτει καμία οικογένεια κατανομών που να έχει όλες τις επιθυμητές ιδιότητες όπως αναλογικό κίνδυνο, απλότητα πιθανότητας, κ.α. (Hougaard, 2000). Δεν υπάρχει κάποιος συγκεκριμένος λόγος για τον οποίο κάποιος να προτιμήσει την εφαρμογή οποιασδήποτε κατανομή ευπάθειας σε σχέση με κάποια άλλη. Όλα τα επιχειρήματα που είναι υπέρ η κατά των κατανομών βασίζονται πάνω στη μαθηματική απλότητα (Wienke και άλλοι, 2003a). Ωστόσο, η κατανομή ευπάθειας ενός πληθυσμού μπορεί να μην είναι απαραίτητα επιθυμητή από θέμα μαθηματικής απλότητας. Σύμφωνα με τους , όταν υπάρχου διαφορετικές κατανομές ευπάθειας απαιτούνται διάφορες δομές εξάρτησης που περιλαμβάνονται μέσα στο μοντέλο.

2.9 Στατιστική αναφορά για τα από κοινού Μοντέλα ευπάθειας

Σε αυτό το μέρος παρουσιάζεται λεπτομερώς η στατιστική αναφορά των μοντέλων ευπάθειας. Στο παρελθόν, υπήρξαν σημαντικές δυσκολίες εκτίμησης των παραμέτρων που καθιστούσε ανίκανη την εφαρμογή των μοντέλων ευπάθειας. Ωστόσο, έχουν προταθεί διάφοροι μέθοδοι εκτίμησης παραμέτρων και λόγος για τον οποίο γίνεται αυτό είναι γιατί ορισμένοι μαθηματικοί τύποι είναι περίπλοκοι και αυτή η επανάληψη μπορεί να προβεί χρονοβόρα. Μια βασική κατεύθυνση είναι η ενσωμάτωση της τυχαίας ευπάθειας αλλά αυτό δεν είναι η μόνη δυνατότητα. Ωστόσο, κάποιος μπορεί να χρησιμοποιήσει απλές μεθόδους εκτίμησης ή αλλιώς εκτιμήσεις ρουτίνας όπου η ευπάθεια συμπεριλαμβάνεται ως μια μη παρατηρήσημη μεταβλητή, όμοιες με τη μέθοδο BLUP (best linear unbiased method) για τα μοντέλα κανονικής κατανομής. Για τη μη παραμετρική συνάρτηση κινδύνου υπάρχει μια παράμετρος ανά χρονικό σημείο με τα παρατηρούμενα γεγονότα. Αυτή μπορεί να συμπεριλαμβάνεται στο μοντέλο ή μπορεί να εξαιρεθεί από τη συνάρτηση πιαθανοφάνειας.

Το γενικό μοντέλο που χρησιμοποιείται σε αυτό το κεφάλαιο είναι το μοντέλο ευπάθειας. Τα παραμετρικά μοντέλα αποτελούν μια απλή μέθοδο και περιγράφονται στη παράγραφο 2.9.1. Μια απλή και εναλλακτική μέθοδος εκτίμησης είναι η χρήση μη παραμετρικής εξαρτημένης μεταβλητής όπως αυτή του kendall's τ, που εκτιμά την εξαρτημένη παράμετρο (βλέπε 2.9.2). Ο ΕΜ αλγόριθμος ο οποίος παρουσιάζεται στη παράγραφο 2.9.3 προσφέρει μια λογική και απλή μέθοδο εκτίμησης. Άλλη μια απλή μέθοδος, που εν μέρει αφορά μη παραμετρικά μοντέλα, είναι η μέθοδος τριών σταδίων (The three-stage approach) και αναφέρεται στην παράγραφο 2.9.4. Μια εναλλακτική μέθοδος εκτίμησης είναι η ποινικοποιημένη πιθανοφάνεια που παρουσιάζεται στη παράγραφο 2.9.5. Επιπλέον, περιγράφεται στη παράγραφο 2.9.6 η μέθοδος καλής προσαρμογής . Τέλος άλλοι μέθοδοι αναφέρονται στη παράγραφο 2.9.7.

2.9.1 Παραμετρικά μοντέλα

Ένα πλήρες παραμετρικό μοντέλο μπορεί να χρησιμοποιηθεί με τον συνηθισμένο τρόπο, διαφοροποιώντας τη λογαριθμική συνάρτηση πιθανοφάνειας όπως περιγράφεται στη εξίσωση (2.2.5). Παραδείγματα που περιγράφονται στις προηγούμενες παραγράφους είναι το μοντέλο γάμμα του Weibull και το από κοινού μοντέλο Weibull. Κατ αρχήν, αυτό δεν παρουσιάζει ιδιαίτερα προβλήματα στη περίπτωση των διμεταβλητών, αλλά στην γενική περίπτωση των πολυμεταβλητών δεδομένων είναι αναγκαία η κωδικοποίηση των παραγώγων του μετασχηματισμού Laplace. Μπορούμε να χρησιμοποιήσουμε παραμέτρους από τις δεσμευμένες καθώς και από τις περιθώριες κατανομές. Τα στατιστικά μοντέλα δεν είναι οικογένειες εκθετικών κατανομών και ως εκ τούτου δεν υπάρχουν σαφείς επαρκείς μειώσεις.

2.9.2 Απλές εκτιμήσεις

Στην απλή περίπτωση όπου δεν υπάρχουν συμμεταβλητές, κάποια απλά μέτρα μπορούν να εκτιμηθούν, όπως η μεταβλητή $\boldsymbol{\tau}$ του kendall . Αυτό ισχύει γενικότερα από τα μοντέλα ευπάθειας και επιτρέπει το προσδιορισμό του βαθμού εξάρτησης. Μέσα σε ένα μοντέλο ευπάθειας, μπορεί να χρησιμοποιηθεί περαιτέρω για τον περιορισμό της παραμέτρου ευπάθειας. Η εξίσωση για αυτό είναι $\delta=(1/\tau-1)/2$ για το μοντέλο γάμμα και $\alpha= 1- \tau$ για την κοινή περίπτωση. Η προσέγγιση αυτή δεν

εφαρμόζεται για τα PVF μοντέλα γιατί ο περιορισμός αφορά μόνο μια παράμετρο. Αυτή η προσέγγιση δεν έχει νόημα όταν το μοντέλο ευπάθειας δεν είναι ακριβές και παρόλο αυτά αυτή η μέθοδος φαίνεται να εφαρμόζεται αποκτώντας μια πρώτη εκτίμηση χωρίς όμως να έχει σημασία για μια πλήρη εκτίμηση. Ο Manatunga και ο Oakes (1996) εξέτασαν την ασυμπτωτική απόδοση αυτής της μεθόδου και αξιολόγησαν τις ροπές αυτών των εκτιμήσεων.

2.9.3 Ο αλγόριθμος EM (Expectation- Maximization)

Η πιο κοινή μέθοδος εκτίμησης παράλληλων δεδομένων (parallel data) με συμμεταβλητές είναι ο EM αλγόριθμος. Αυτή η μέθοδος χρησιμοποιεί τη πλήρη πιθανοφάνεια και συμπεριλαμβάνει εξίσου τις παρατηρήσημες ποσότητες (T, D) και την ευπάθεια Y. Εννοιολογικά , αυτό μπορεί να περιγραφεί ως η εισαγωγή της ευπάθειας στην πιθανοφάνεια και στη συνέχεια πολλαπλασιάζοντας με την συνάρτηση πυκνότητας πιθανότητας της ευπάθειας. Η περιγραφή αυτή θα γίνει με τη χρήση της γάμμα κατανομής όπου θα πάρουμε για θ=δ για να ληφθεί μια μέση τιμή του 1.

Επίσημα, αυτό μας δίνει τη πλήρη πιθανοφάνεια του $L_{(T,D)\setminus Y} L_Y$, όπου ο πρώτος όρος είναι η τυπική πιθανοφάνεια επιβίωσης δεδομένης της ευπάθειας

$$\prod_i \prod_j Y_i^{D_{ij}} \mu_0(T_{ij})^{D_{ij}} \exp(D_{ij}\beta' z_{ij}) \exp\left\{-\int_0^{T_{ij}} Y_i \mu_0(u) \exp(\beta' z_{ij}) du\right\} \quad (2.9.3.1)$$

και ο δεύτερος όρο είναι το προϊόν των πυκνοτήτων της γάμμα κατανομής:

$$\prod_i \delta^{\delta} Y_i^{\delta-1} \exp(-\delta Y_i) / \Gamma(\delta)$$

Η μέθοδος εναλλάσσεται μεταξύ ενός προσδοκώμενου βήματος (E-step) και ενός βήματος μεγιστοποίησης(M-step). Στο προσδοκώμενο βήμα, οι μη παρατηρήσημοι όροι στην λογαριθμική πιθανοφάνεια αντικαθιστούνται με τη μέση τιμή δεδομένου των παρατηρήσεων. Η δεσμευμένη κατανομή είναι η γάμμα κατανομή με παράμετρο $\tilde{\delta}_i = \delta + \sum_j D_j$ και $\tilde{\theta}_i = \delta + \sum_j M_j(T_j)$ όπου τόσο το Y_i όσο και το $\log Y_i$ συμπεριλαμβάνονται στη λογαριθμική πιθανοφάνεια . Συνεπώς, εισάγουμε

$$E(Y_i \setminus F_\infty) = \frac{\tilde{\delta}_i}{\tilde{\theta}_i} \quad , \quad E(\log Y_i \setminus F_\infty) = \Psi(\tilde{\delta}_i) - \log \tilde{\theta}_i \quad (2.9.3.2)$$

στο βήμα της μεγιστοποίησης, οι τιμές της ευπάθειας θεωρούνται σταθερές και γνωστές. Οι παράγοντες πιθανοφάνειας είναι $L_1(\beta)L_2(\delta)$ όπου μπορούμε καθένα από τους δυο ξεχωριστά να τους μεγιστοποιήσουμε. Η αρχική διατύπωση αυτής της μεθόδου μετακίνησε τον όρο που προέρχεται από το $Y_i^{D_{ij}}$ σε έναν άλλο παράγοντα, αλλά αυτό δεν έχει ιδιαίτερη σημασία σε αυτό το βήμα της μεθόδου καθώς αυτός ο παράγοντας λειτούργει ως σταθερή πιθανότητα. Για β το πρόβλημα μοιάζει με το μοντέλο του Cox με γνωστούς συντελεστές παλινδρόμησης. Ο τρόπος που αυτό φαίνεται είναι διατυπώνοντας το Y_i σε $\exp(\zeta_i z_i)$ όπου $z_i = \log Y_i$ και ζ_i είναι μια σταθερή παράμετρος. Η συνάρτηση κινδύνου έχει εξαλειφθεί από την

$$\hat{\lambda}_0(t) = \frac{1}{\sum_{(i,j)\in R(t)} \tilde{Y}_i \exp(\beta' z_{ij})}$$

Είναι η αρχική συνάρτηση μερικής πιθανοφάνειας με παραμέτρους ζ και β. Κρατώντας σταθερό το ζ, το β εκτιμάται κρατώντας αντίστοιχα τους όρους του Y σταθερούς. Ο τρόπος με τον οποίο γίνεται αυτή η εκτίμηση είναι ο ακόλουθος:

1. Υπολογίζει τις εκτιμήσεις του β και Λ(t) στο μοντέλο χωρίς ευπάθεια. Αυτό αντιστοιχεί στο $\delta \to \infty$.
2. (E-step) Εισαγάγει τη μέση τιμή αξιολογώντας την από την εξίσωση (2.9.3.2) βασιζόμενος στις τρέχουσες τιμές των παραμέτρων.
3. (M-step 1)Μεγιστοποιεί το L_1 συναρτήσει του β μέσω της τυπικής αξιολόγησης του μοντέλου του Cox.
4. (M-step 2) Μεγιστοποιεί το L_2 συναρτήσει του δ.
5. Γυρίζει στο βήμα 2 μέχρις ότου οι εκτιμήσεις να συγκλίνουν.

Ο ΕΜ αλγόριθμος είναι απλός και εύκολος στον προγραμματισμό μόνο που μπορεί να απαιτεί ένα μεγάλο αριθμό επαναλήψεων. Παρόλο αυτά, ένα πλεονέκτημα αυτής της μεθόδου είναι ότι οι χρόνοι απεβίωσης συμβάλουν, με έναν απλό τρόπο, μόνο στο E-step καθώς στο 3^{ο} βήμα του αλγορίθμου οι παράμετροι κινδύνου στους χρόνους διακοπής (death times) έχουν εξαλειφθεί. Αυτή η μέθοδος δεν μας επιτρέπει την αξιολόγηση της διακύμανσης στο πλήρες μοντέλο. Ωστόσο, είναι πολύ πιθανή η

χρήση του EM αλγορίθμου για την απόκτηση εκτίμησης των παραμέτρων σε συνδυασμό όμως με μια άλλη μέθοδο.

2.9.4 Η μέθοδος των τριών σταδίων (The three stage approach)

Παρόλο που ένα παραμετρικό μοντέλο δεν μπορεί να δώσει μια ικανοποιητική προσαρμογή για τη περιθώρια κατανομή, δημιουργήθηκε μια εναλλακτική ημι-παραμετρική προσέγγιση η οποία συνδυάζει παραμετρικά μοντέλα ανεξαρτησίας με κλασσικές μη παραμετρικές εκτιμήσεις για την περιθώρια κατανομή. Αυτό βασίζεται στη εξίσωση (2.3.3). Η μέθοδος, η οποία καλύπτει το μοντέλο χωρίς συμμεταβλητές και το θετικό σταθερό μοντέλο ευπάθειας (positive stable frailty model)με τη χρήση συμμεταβλητών, καλείται η μέθοδος των τριών σταδίων και προσεγγίζει την εκτίμηση της μέγιστης πιαθανοφάνειας . Τα βήματα τις μεθόδου είναι τα εξής:

- Στο πρώτο στάδιο, στη περίπτωση των μη συμμεταβλητών, βρίσκουμε τις εκτιμήσεις των παραμέτρων της περιθώρια κατανομής μέσω ενός παραμετρικού μοντέλου, με μη παραμετρική εκτίμηση (Nelson-Aalen), ενώ με τη παρουσία συμμεταβλητών, με τη συνάρτηση μερικής πιθανοφάνειας του μοντέλου του Cox. Αυτό το μοντέλο καλείται ως ανεξάρτητο μοντέλο (IWM) καθώς οι χρόνοι θεωρούνται ανεξάρτητοι. Ο εκτιμώμενος κίνδυνος, όπου στην περίπτωση του Nelson-Aalen δεν εξαρτάται από τη τιμή του i και j, είναι $\hat{\Omega}_{ij}(t)$ ενώ στη περίπτωση του αναλογικού κινδύνου δίνεται από τον τύπο $\hat{\Omega}(t)\exp(\hat{\beta}'z_{ij})$.
- Στο δεύτερο στάδιο, οι περιθώριες κατανομές θεωρούνται γνωστές και σταθερές και η εξάρτηση εκτιμάται. Πρακτικά αυτό συμβαίνει με το μετασχηματισμό από την ολοκληρωμένη συνάρτηση κινδύνου με παρατηρήσεις T_{ij} που υποκαθίστανται από τη σχέση

$$\tilde{T}_{ij} = \hat{\Omega}_{ij}(T_{ij}) \qquad (2.9.4.1)$$

αν χρησιμοποιούταν η πραγματική τιμή του $\Omega_{ij}(\cdot)$ αντί της εκτίμησης, τότε θα έδινε εκθετικά κατανεμημένες μεταβλητές του μέσου 1.

Καθώς εκτιμώνται οι συναρτήσεις κινδύνου, τα $\tilde{T}_{ij}$ είναι κατά προσέγγιση εκθετικά κατανεμημένα.

- Στο τρίτο στάδιο, ωστόσο, μπορεί κανείς να μη θεωρήσει σταθερή τη κλίμακα των τροποποιημένων παρατηρήσεων . Είναι απλό να υποθέσουμε ότι οι περιθώριοι κίνδυνοι είναι ανάλογοι της εκτίμησης που βρέθηκε από τη περιθώρια κατανομή. Πρακτικά, αυτό γίνεται εφαρμόζοντας μια διμεταβλητή κατανομή με εκθετικά περιθώρια αντί της εκθετικής κατανομής που αναμένεται μετά το χρόνο μετατροπής της εξίσωσης (2.5.1).

Βήμα	**Ανεξάρτητη παράμετρος**	**Περιθώρια κατανομή**
1	Ανεξάρτητα δεδομένα	Εκτιμημένα
2	Εκτιμημένα	Σταθερά
3	Εκτιμημένα	Εκτιμημένος αναλογικός παράγοντας

Πίνακας 2.1 Επισκόπηση της μεθόδου τριών σταδίων

Αυτό επίσημα είναι μια βελτίωση της εκτίμησης επιτρέποντας τις παραμέτρους να αποκλίνουν από την εκτίμηση υπό τη προϋπόθεση της ανεξαρτησίας . Στη πράξη η διαφορά δεν είναι μεγάλη αλλά εξακολουθεί να είναι ευεργετική στο να επεκτείνεται, καθώς η διακύμανση είναι λιγότερο υποεκτιμημένη, επιλέγοντας το μεγαλύτερο δυνατόν μοντέλο. Για τη θετική σταθερή κατανομή ευπάθειας, υπάρχει μια μικρή διαφορά μεταξύ της εκτιμημένης εξάρτησης και της εκτιμημένης διακύμανσης αυτής της ποσότητας. Τα βήματα της μεθόδου παρουσιάζονται περιληπτικά από τον παραπάνω ***Πίνακας 2.1.***

Μπορεί εύκολα κάποιος να χρησιμοποιήσει το μοντέλο του Weibull από το εκθετικό μοντέλο για τον περιθώριο κίνδυνο, καθώς αυτό αποτελεί καλύτερη εφαρμογή για τη μη παραμετρική εκτίμηση και εκτιμά καλύτερα τη διακύμανση. Στη περίπτωση του κοινού μοντέλου, μπορούμε επίσης να συμπεριλάβουμε τις συμμεταβλητές σε ένα μοντέλο παλινδρόμησης και να προσαρμόσουμε τους συντελεστές σε ένα διμεταβλητό εκθετικό μοντέλο με σκοπό τη βελτίωση και την εύρεση της καλύτερης εκτίμησης της μεταβλητότητας.

Η μέθοδος των τριών σταδίων έχει νόημα εφαρμογής μόνο όταν θέλουμε να κάνουμε πλήρεις εκτιμήσεις (Hougaard 2000).

2.9.5 Η ποινικοποιημένη μέθοδος πιθανοφάνειας

Η ποινικοποιημένη μέθοδος παρουσιάζει ορισμένες ομοιότητες με τον ΕΜ αλγόριθμο. Βασίζεται στην τροποποίηση της μερικής πιθανοφάνειας του μοντέλου του Cox όπου περιλαμβάνονται και βελτιστοποιούνται όχι μόνο οι συντελεστές παλινδρόμησης αλλά και οι ευπάθειες. Πιο συγκεκριμένα, η πιθανοφάνεια περιγράφεται ως προϊόν δυο ορών. Ο πρώτος όρος εκφράζει τη μερική πιθανοφάνεια ο οποίος περιλαμβάνει τις παραμέτρους ευπάθειας, ενώ ο δεύτερος, παρουσιάζει τη ποινικοποιημένη πιθανοφάνεια που αποφεύγει τις μεγάλες διαφορές ανάμεσα στις ευπάθειες των διαφορετικών γκρουπ ατόμων. Στη πράξη, στο πρώτο βήμα η πιθανοφάνεια είναι εφοδιασμένη με το καθορισμό των τιμών ευπάθειας στο 1. Έπειτα, σε σχέση με το πρώτο βήμα, χρησιμοποιείται μια επαναληπτική μέθοδος που βελτιστοποιεί τη μερική πιθανοφάνεια, θεωρώντας τις παραμέτρους ευπάθειες σταθερές και γνωστές.

Στο δεύτερο βήμα, οι παράμετροι της ευπάθειας αξιολογούνται ως δεσμευμένες μέσα στις παρατηρήσεις τους όπως συμβαίνει στον ΕΜ αλγόριθμο. Αυτό επαναλαμβάνεται μέχρι να συγκλίνουν. Το πλεονέκτημα της μεθόδου αυτής είναι ότι η συμβολή του ολοκληρωμένου κινδύνου είναι ελεγχόμενη. Δεν εξαλείφεται εντελώς, αλλά οι υπολογισμοί είναι τόσο απλοί που όπου η μέθοδος οδηγεί σε πιο γρήγορα και βέλτιστα αποτελέσματα. Από την άλλη, το μεγαλύτερο μειονέκτημα της μεθόδου αυτή είναι ότι καθιστά δύσκολη την απόκτηση μιας έγκυρης εκτίμησης του τυπικού σφάλματος, ιδιαίτερα στις παραμέτρους της ευπάθειας. Η μέθοδος αυτή δουλεύει καλύτερα για τη γάμμα κατανομή και προσεγγιστικά για την λογάριθμο-κανονική κατανομή. (Πιο αναλυτικά θα τη παρουσιάζεται στο $3^{ο}$ κεφάλαιο).

2.9.6 Έλλειψη καλής προσαρμογής (Goodness-of fit)

Σχετικά συχνή διαδικασία είναι ο έλεγχος υποθέσεων σε ένα μοντέλο και αυτό μπορεί να πραγματοποιηθεί με τρείς διαφορετικούς τρόπους. Αρχικά, μπορεί κανείς να προσαρμόσει ένα μεγαλύτερο μοντέλο και να κάνει είτε επίσημες είτε ανεπίσημες υποθέσεις. Αν ο έλεγχος γίνει αποδεκτός, τότε το αρχικό μοντέλο είναι ικανοποιητικό. Έπειτα, μπορεί να γίνουν επιπρόσθετοι υπολογισμοί στο μοντέλο αναμένοντας τα κατάλληλα αποτελέσματα. Στη τυπική κανονική κατανομή του γραμμικού μοντέλου, αυτό θα μπορούσε να αποτελείται, κατά κάποιο τρόπο, από την

αξιολόγηση των καταλοίπων και την σύγκριση τους. Τέλος, μπορεί κανείς να προσαρμόσει ένα εντελώς διαφορετικό μοντέλο, προκειμένου να εξεταστεί αν υπάρχει ικανοποιητική συμφωνία στις αρχικές υποθέσεις. Παρακάτω παρουσιάζονται οι διαδικασίες αναλυτικότερα.

Υπάρχουν αρκετές δυνατότητες για τη προσαρμογή μεγαλύτερων μοντέλων. Αν η κατανομή ευπάθειας του μοντέλου είναι η γάμμα , μπορεί κανείς να προσαρμόσει μια PVF κατανομή ακόμα και ένα διμεταβλητό μοντέλο ευπάθειας. Προκειμένου να γίνει ο έλεγχος κινδύνου του παραμετρικού μοντέλου, θα μπορούσε αντ' αυτού να βρεθεί μια εκτίμηση σε ένα μη παραμετρικό μοντέλο. Επιπλέον, σε ένα μοντέλο παλινδρόμησης που ως στόχο έχει την εξέταση τη παραδοχή της αναλογικότητας, μπορεί κανείς να προβεί σε στρωματοποίηση σύμφωνα με τη συμμεταβλητή και να συγκρίνει τις λειτουργίες κινδύνου. Επιπροσθέτως, είναι δυνατόν η προσαρμογή του μοντέλου με ποικίλους χρόνους ευπάθειας, που ως στόχο έχει να ελεγχτεί αν η εξάρτηση είναι μακροχρόνια αδυναμία χρόνου. Ωστόσο, αν μια μεταβλητή θεωρηθεί σημαντική τότε θα πρέπει να συμπεριληφθεί στο μοντέλο σαν συμμεταβλητή.

Η μέθοδος που ακολουθεί επιπρόσθετους υπολογισμούς κατά τη παραμετρική περίπτωση προτάθηκε από τους Shih και Luis (1995a) ενώ στην ημι-παραμετρική από τον Glidden (1999). Πρότειναν λοιπόν, τη δεσμευμένη τιμή του Y σαν συνάρτηση χρόνου του γάμμα μοντέλου ευπάθειας. Αυτό κυμαίνεται γύρο στο 1 σε ολόκληρο το σύνολο του πληθυσμού. Αν η πραγματική τιμή της ευπάθειας είναι διαφορετική, τότε θα αποκλίνει από αυτή την τιμή.

Η προσαρμογή για ένα εντελώς διαφορετικό μοντέλο είναι πιθανόν να συμβεί με την προσαρμογή ποικίλων καταστάσεων μοντέλων για διμεταβλητά δεδομένα. Σύμφωνα με το γάμμα μοντέλο ευπάθειας, οι παράμετροι της εξάρτησης, όπως η παράμετρος τ του Kendall, δεν αλλάζουν και στο PVF μοντέλο το τ μειώνεται με την πάροδο του χρόνου και τείνει στο μηδέν. Αυτό μπορεί να ελεγχθεί με την αξιολόγηση της μη παραμετρικής εκτίμησης του τ ως συνάρτηση του χρόνου αποκοπής.

Θα μπορούσε κανείς να αναρωτηθεί ποιες θα ήταν οι συνέπειες κατά την επιλογή μιας λανθασμένης κατανομής ευπάθειας . Για παράδειγμα αν το πραγματικό μοντέλο δημιουργείται από σταθερές ευπάθειες και αναλύεται από την γάμμα κατανομή, γεννάται το ερώτημα αν η εκτιμημένη παράμετρος τ είναι διαφορετική από τη πραγματική τιμή τ; Με πλήρη δεδομένα λοιπόν και χωρίς συμμεταβλητές αλλά και με τη χρήση λανθασμένου μοντέλου ευπάθειας, οδηγούμαστε σε υποτιμώμενη

εξάρτηση. Με βάση τη λογοκρισία, η απάντηση δεν είναι τόσο εύκολη. Πιο συγκεκριμένα, αν το κατάλληλο μοντέλο είναι το σταθερό, χρησιμοποιώντας το μοντέλο γάμμα, η εκτίμηση θα εξετάσει τη παρατηρούμενη και πρόωρη εξάρτηση και καθώς το μοντέλο γάμμα συνεπάγεται μια αρκετά χαμηλή εξάρτηση προτείνει ότι η εκπρόθεσμη εξάρτηση είναι πολύ μεγάλη ενώ η συνολική οδηγεί σε μια υπερεκτιμημένη εξάρτηση. Από την άλλη, αν το πραγματικό μοντέλο είναι το γάμμα, και γίνεται εφαρμογή στο σταθερό μοντέλο τότε έχουμε το αντίθετο πρόβλημα, δηλαδή εμφανίζονται ανεξάρτητα δεδομένα ακόμα και αν δεν είναι.

2.9.7 Άλλες μέθοδοι εκτίμησης

Η μέθοδος Markov chain Monte Carlo (MCMC) επίσης έχει προταθεί για τα μοντέλα ευπάθειας. Οι τιμές ευπάθειας, αντί του περίπλοκου χειρισμού της πιθανοφάνειας, προσομοιώνονται από την κατανομή του τρέχοντος βήματος της επανάληψης. Όμοια με τον EM αλγόριθμο η κομβική διαδικασία μεταξύ του βήματος με προσομοίωση των παραμέτρων ευπάθειας, οι οποίες βασίζονται στις τρέχουσες παραμέτρους και τις δεσμευμένες κατανομές της ευπάθειας δεδομένου του βήματος και όπου οι παράμετροι αυτοί είναι αναβαθμίσιμοι, βασίζεται στις τιμές της ευπάθειας. Επίσης, αυτή η μέθοδος εφαρμόζεται και για το λογαριθμικό μοντέλο ευπάθειας. Ωστόσο, είναι δύσκολη η εφαρμογή αυτής της μεθόδου τόσο για το σταθερό μοντέλο όσο και για το PVF μοντέλο (εκτός για α=1/2), διότι δεν υπάρχουν διαθέσιμοι αποτελεσματικοί μέθοδοι προσομοίωσης.

ΚΕΦΑΛΑΙΟ 3

ΜΕΘΟΔΟΙ ΕΠΙΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ

3.1 Εισαγωγή

Στη στατιστική, η γραμμική παλινδρόμηση είναι μια προσέγγιση μοντελοποίησης της σχέσης μιας εξαρτημένης μεταβλητής Y_i με ένα σύνολο n διανυσμάτων και ανεξάρτητων μεταβλητών ή παραγόντων $X_1, X_2, ..., X_n$. Στο εξής, σημαντική είναι η εκτίμηση της συνάρτησης παλινδρόμησης $X\beta$ του γραμμικού μοντέλου

$$Y = X\beta + \varepsilon .$$

Στο πλαίσιο της γραμμικής παλινδρόμησης, η $X\beta$ προσεγγίζεται από τον γραμμικό συνδυασμό των πολυωνύμων ή των άλλων συναρτήσεων του X. Πρακτικά, εισάγεται ένας μεγάλος αριθμός παραγόντων στο αρχικό στάδιο μοντελοποίησης έτσι ώστε να γίνει η καλύτερη δυνατή επιλογή υποσυνόλου παραγόντων που εμφανίζουν σημαντική επίδραση στην απόκριση Y. Επίσης, οι ερευνητές για να προβούν σε πιο γρήγορη και αποτελεσματική επιλογή παραγόντων, ακολουθώντας τη κατά βήματα απαλοιφή (stepwise deletion) αυτών. Ωστόσο, η επιλογή μεταβλητών δεν είναι σημαντική μόνο για τα γραμμικά μοντέλα παλινδρόμησης, αλλά και για τα παραμετρικά μοντέλα όπως τα εύρωστα γραμμικά μοντέλα (robust linear model), τα γενικευμένα γραμμικά μοντέλα (generalized linear model), τα ημι-παραμετρικά (μοντέλα αναλογικού κινδύνου του Cox) και μη παραμετρικά μοντέλα παλινδρόμησης. Μια μέθοδος επιλογής μεταβλητής για να είναι αποτελεσματική, θα πρέπει οι εκτιμήσεις να έχουν τις παρακάτω ιδιότητες : α)Αμεροληψία: Ο προκύπτων εκτιμητής πρέπει να είναι σχεδόν αμερόληπτος, ιδίως στην περίπτωση όπου η σωστή άγνωστη παράμετρος β_j είναι μεγάλη. β)Σποραδικότητα: Ο προκύπτων εκτιμητής πρέπει να αποτελεί κανόνα περιορισμού (*thresholding rule*), ώστε οι εκτιμώμενοι συντελεστές με μικρή τιμή, να

μηδενίζονται. Έτσι, μειώνεται η πολυπλοκότητα του μοντέλου γ)Συνέχεια: Ο προκύπτων εκτιμητής πρέπει είναι συνεχής. Αποφεύγεται κατά αυτόν τον τρόπο η αστάθεια στη πρόβλεψη του μοντέλου. Επιπλέον , στο κεφάλαιο αυτό παρουσιάζεται και η μέθοδος των ποινικοποιημένων ελαχίστων τετραγώνων (penalizes least square) που αποδίδει μια νέα μέθοδο επιλογής μεταβλητών.

Πέρα από τις παραπάνω παραδοσιακές μεθόδους επιλογής μεταβλητής, στο κεφάλαιο αυτό παρουσιάζεται αναλυτικά η μέθοδος επιλογής ποινικοποιημένης πιθανοφάνειας. Σε αντίθεση με τις άλλες μεθόδους, η συγκεκριμένη καθορίζει τις ιδιότητες των δειγμάτων και εφαρμόζεται αποτελεσματικά σε μη παραμετρικά μοντέλα. Στη παράγραφο 3.4.1 γίνεται αναφορά σε σχέση με τη μέθοδο ποινικοποιημένων ελαχίστων τετραγώνων και της μεθόδου καλύτερης επιλογής υποσυνόλου όταν ο πίνακας σχεδιασμού είναι ορθογώνιος. Στη παράγραφο 3.4.2 εκτείνεται η ιδέα της μεθόδου ποινικοποιημένης πιθανοφάνειας σε παραμετρικά μοντέλα συμπεριλαμβάνοντας τα παραδοσιακά μοντέλα παλινδρόμησης, εύρωστα μοντέλα παλινδρόμησης και τα γενικευμένα γραμμικά μοντέλα. Επιπλέον, οι δειγματοληπτικές και προβλεπτικές ιδιότητες παρουσιάζονται στη παράγραφο 3.4.2.2. Βασιζόμενοι στη τετραγωνική προσέγγιση, ένας νέος προτεινόμενος αλγόριθμος προτείνεται στη παράγραφο 3.4.2.3 για την εύρεση εκτιμητών ποινικοποιημένης πιθανοφάνειας. Οι τύποι για τους πίνακες συνδιακύμανσης των εκτιμώμενων συντελεστών προκύπτουν επίσης σε αυτό το κεφάλαιο. Στη παράγραφο 3.4.3 γίνονται ορισμένες αριθμητικές συγκρίσεις και μελέτες προσομοίωσης. Τέλος, στις παραγράφους 3.5, 3.6 παρουσιάζονται και αναλύονται άλλοι μέθοδοι επιλογής μεταβλητών όπως η μέθοδος garotte και οι μέθοδοι συρρίκνωσης αντίστοιχα.

3.2 Το γενικό γραμμικό μοντέλο

Θεωρούμε το γενικό γραμμικό μοντέλο που δίνεται από τη σχέση

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i,\ i = 1, 2, ..., n,\ j = 1, ..., k$$

ή αλλιώς υπό τη μορφή πινάκων,

$$y = X\beta + \varepsilon$$

Όπου $y=(y_{1,}y_2,...,y_n)'$ είναι ένα $n\times 1$ διάνυσμα και X ένας $n\times d$ πίνακας είναι ο πίνακας των τιμών των ανεξάρτητων μμεταβλητών. Όπως και στο παραδοσιακό γενικό γραμμικό μοντέλο, υποθέτουμε ότι τα y_i είναι η στήλη των παρατηρήσεων της εξαρτημένης μεταβλητής. Κάθε γραμμή αναφέρεται σε μια διαφορετική στατιστική μονάδα ή παρατήρηση και κάθε στήλη σε διαφορετική.

Η στήλη των παραμέτρων $\beta=\left(\beta_1,\beta_2,.....\beta_d\right)'$ περιλαμβάνει τους συντελεστές των ανεξάρτητων μεταβλητών οι οποίοι θεωρούνται άγνωστοι και πρέπει να εκτιμηθούν.

Επιπλέον, η στήλη των υπολοίπων (rresiduals) $\varepsilon=\left(\varepsilon_1,\varepsilon_2,.....\varepsilon_n\right)'$ είναι η στήλη των τυχαίων σφαλμάτων (random error terms).

Η υπόθεση που υιοθετούμε στο παραπάνω γραμμικό μοντέλο είναι ότι τα $\varepsilon=\left(\varepsilon_1,\varepsilon_2,.....\varepsilon_n\right)'$ είναι ανεξάρτητα με την ίδια κατανομή $N(0,\sigma^2)$ και ακολουθούν τις παρακάτω υποθέσεις του ανάλογες του απλού γραμμικού μοντέλου:

- $E(\varepsilon_i)=0$, για κάθε i
- $V(\varepsilon_i)=\sigma^2$ για κάθε i, δηλαδή τα τυχαία σφάλματα ικανοποιούν την υπόθεση ομοσκεδαστικότητας
- $\text{cov}(\varepsilon_i,\varepsilon_j)=0$, για $i\neq j$ δηλαδή τα ε_i είναι τα ασυσχέτιστα μεταξύ τους

Επίσης, X ονομάζεται ο πίνακας σχεδιασμού και παίρνει την ακόλουθη μορφή

$$X=\begin{pmatrix}1 & x_{11}\cdots & x_{1k}\\ 1 & x_{21}\cdots & x_{2k}\\ \vdots & \vdots & \vdots\\ 1 & x_{n1}\cdots & x_{nk}\end{pmatrix}$$

Δύο από τις συνηθέστερες μεθόδους εκτίμησης είναι η εκτίμηση με τη μέθοδο της μέγιστης πιθανοφάνειας (ΕΜΠ) και η μέθοδος των ελαχίστων τετραγώνων (Μ.Ε.Τ.) όπου παρουσιάζονται παρακάτω.

3.2.1 Μέθοδος εκτίμησης ελαχίστων τετραγώνων

Η μέθοδος των ελαχίστων τετραγώνων για την εκτίμηση των παραμέτρων β βασίζεται στην ελαχιστοποίηση της συνάρτησης

$$S(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = (y - X\beta)'(y - X\beta)$$

Ο εκτιμητής $\hat{\beta}$ υπολογίζεται παραγωγίζοντας τη συνάρτηση **S**(β) σε σχέση με κάθε στοιχείο β και στη συνέχεια λύνοντας το σύστημα των εξισώσεων

$$\frac{\partial S(\beta)}{\partial \beta} = -2\mathrm{X}'\left(y - \mathrm{X}\beta\right)$$

Θέτοντας τη παραπάνω σχέση ίση με ο μηδέν παίρνουμε τη σχέση

$$\mathrm{X}'y = \mathrm{X}'\mathrm{X}\hat{\beta}$$

και αν ο πίνακας αντιστρέφεται τότε η εκτιμήτρια ελάχιστων τετραγώνων δίνεται από τη σχέση

$$\hat{\beta} = \left(\mathrm{X}'\mathrm{X}\right)^{-1}\mathrm{X}'y$$

Οπότε το εκτιμώμενο μοντέλο δίνεται από τη σχέση $\hat{y} = \mathrm{X}\hat{\beta}$ με υπόλοιπα, $e = y - \hat{y}$

Κάθε μια από τις εκτιμήσεις $\hat{\beta}_j = 1,...,k$ εκφράζει την αναμενόμενη μεταβολή της y για μια μονάδα αύξησης της αντίστοιχης επεξηγηματικής μεταβλητής x_j, δεδομένου ότι οι άλλες μεταβλητές παραμένουν σταθερές.

3.2.2 Μέθοδος εκτίμησης μέγιστης πιθανοφάνειας

Σύμφωνα με τη μέθοδο αυτή η οποία προτάθηκε από τον Fisher (Aldrich, J. 1997) η εκτιμήτρια της μέγιστης πιθανοφάνειας (ΕΜΠ) θ της παραμέτρου θ είναι εκείνη η οποία μεγιστοποιεί την συνάρτηση πιθανοφάνειας

$$L = \left(\theta; y_1, y_2, ..., y_n\right) = \prod f\left(y_i; \theta\right).$$

Ή ισοδύναμα την λογαριθμική συνάρτηση πιθανοφάνειας

$$L = \left(\theta; y_1, y_2, ..., y_n\right) = \sum_{i=1}^{n} \ln f\left(y_i; \theta\right)$$

Συνήθως ο εκτιμητής θ βρίσκεται παραγωγίζοντας τη συνάρτηση $L=(\theta;y_1,y_2,....,y_n)$ σε σχέση με κάθε στοιχείο θ_i του θ και λύνοντας το σύστημα εξισώσεων

$$\frac{\partial l(\theta;y)}{\partial \theta_j}=0 \text{ για } j=1,2,...,d$$

Είναι πάντα αναγκαίο να ελέγξουμε ότι ο Εσσιανός πίνακας (Hessian matrix) των δευτέρων παραγόντων της $L=(\theta;y_1,y_2,....,y_n)$ είναι αρνητικά ορισμένος.

Ο εκτιμητής αυτός έχει ιδιότητες που τον κάνουν να υπερέχει έναντι των άλλων εκτιμητών. Μερικές από αυτές είναι οι εξής :

1. Αν g(θ) είναι μία συνάρτηση του θ τότε ο εκτιμητής μέγιστης πιθανοφάνειας του g(θ) είναι $g(\theta)$.

2. Συνέπεια (consistency)

3. Επάρκεια (Sufficiency)

4. Ασυμπτωτική αποτελεσματικότητα (Asymptotic efficiency)

3.3 Μέθοδοι επιλογής καλύτερου υποσυνόλου

Υπάρχουν περιπτώσεις που καθιστούν αναγκαία την επιλογή ενός μικρού υποσυνόλου από ένα μεγαλύτερο σύνολο μεταβλητών οι οποίες χρησιμοποιούνται για τη πρόβλεψη μιας εξαρτημένης μεταβλητής. Για παράδειγμα η διαδικασία πρόβλεψης της εξαρτημένης μεταβλητής Υ μπορεί να θεωρείται οικονομικά ασύμφορη και αρκετά χρονοβόρα εφόσον χρησιμοποιηθούν όλες οι τιμές που έχουμε για τις ανεξάρτητες μεταβλητές. Για το λόγο αυτό αναζητείται ένα μικρό υποσύνολο μεταβλητών που θα μπορεί με αρκετή ακρίβεια να προβλέψει τη τιμή της μεταβλητής Υ. Η επιλογή αυτού του συνόλου θα μπορούσε προφανώς στη συνέχεια να χρησιμοποιηθεί για κάποια μελλοντική πρόβλεψη εφόσον τα δεδομένα είναι αντιπροσωπευτικά των συνθηκών υπό τις οποίες θα γίνει η πρόβλεψη. Η επιλογή των καλύτερων μεταβλητών γίνεται με τη βοήθεια κατάλληλων ελέγχων και κριτηρίων που μας εξασφαλίζουν την ορθότητα της επιλογής (Καρώνη, Χ., Οικονόμου, Π. 2010).

Επομένως, για την επιλογή της καλύτερης συνάρτησης παλινδρόμησης, έχουν προταθεί διάφορα κριτήρια, οι οποίες απαντώνται και σε πρακτικά προβλήματα. Τα κριτήρια αυτά είναι:

1. Τα κριτήρια πρόβλεψης: C_p του Mallow (Mallows, C. L. 1973), R^2
2. Τα μέτρα καταλληλότητας βασισμένα στη πιθανοφάνεια και τη πληροφορία: κριτήρια AIC και BIC
3. Βηματική παλινδρόμηση (stepwise regression)
4. Η μέθοδος $PRESS_P$

Στη συνέχεια γίνεται αναφορά σε ορισμένες από αυτές τις μεθόδους .

3.3.1 Το κριτήριο C_p -Mallows

Ένα μέτρο καταλληλότητας του μοντέλου είναι η στατιστική συνάρτηση C_p - Mallows (Mallowws C.L 1973). Βασίζεται στην ακρίβεια πρόβλεψης του μοντέλου και συγκεκριμένα στο μέσο τετραγωνικό σφάλμα. Η συνάρτηση αυτή ορίζεται από τον ακόλουθο τύπο,

$$C_p = \frac{SSE(p)}{\hat{\sigma}^2} + 2p - n$$

Όπου p ο αριθμός των επεξηγηματικών μεταβλητών στο μοντέλο και n ο αριθμός ων παρατηρήσεων. Επίσης, ο όρος $SSE(p)$ συμβολίζει το άθροισμα των τετραγώνων των υπολοίπων του υπό εξέταση μοντέλου και σ^2 (το μέσο τετραγωνικό υπόλοιπο) από μια κατάλληλη εκτίμηση του $\hat{\sigma}^2$.
Επομένως, ως κατάλληλο μοντέλο θεωρείται εκείνο για το οποίο ισχύει

$$C_p = p$$

Ειδικότερα, αν υπάρχουν περισσότερα από ένα μοντέλο με $C_p = p$, προτιμότερο είναι εκείνο με το μικρότερο p (Καρώνη, Χ., Οικονόμου, Π. 2010) .

3.3.2 Μέτρα καταλληλότητας βασισμένα στη πιθανοφάνεια και πληροφορία

Τα κριτήρια AIC (*Akaike Information Criterion*) και BIC (*Bayesian Information Criterion*) αποτελούν ένα ακόμη κριτήριο επιλογής βέλτιστου μοντέλου ανάμεσα σε άλλα με διαφορετικό αριθμό παραμέτρων. Τα κριτήρια αυτά μπορούν να θεωρηθούν ως σχεδόν αμερόληπτοι εκτιμητές της αναμενόμενης συνολικής διαφοράς μεταξύ του σωστού και του υποψήφιου μοντέλου. Η βασική τους διαφορά είναι ότι η εισαγωγή επιπρόσθετων παραμέτρων αποθαρρύνεται σε μεγάλο βαθμό από το AIC.

3.3.2.1 Το Κριτήριο AIC

Ο γενικότερος σκοπός της επιλογής του κριτηρίου AIC είναι να εκτιμηθεί η απώλεια πληροφοριών όταν η κατανομή πιθανότητας f που συνδέεται με το πραγματικό μοντέλο (generating), προσεγγίζεται με τη κατανομή πιθανότητας g η οποία συσχετίζεται με το μοντέλο που πρόκειται να αξιολογηθεί. Ένα μέτρο καταλληλότητας για τη διαφορά μεταξύ του πραγματικού μοντέλου κα του προσεγγιστικού δίνεται από τον Kullback–Leibler (1951),η ποσότητα $I(f, g)$ η οποία είναι ίση με την αρνητική γενικευμένη εντροπία του Boltzmann's (1877) (βλέπε Bozdogan, 1987; Burnham & Anderson, 2002; Golan,2002; and Sakamoto και άλλοι, 1986).

Ο Akaike (1973; Bozdogan, 1987) έδειξε ότι με την επιλογή μοντέλου με την χαμηλότερη αναμενόμενη απώλεια πληροφορίας (δηλαδή ότι το μοντέλο ελαχιστοποιεί την αναμενόμενη Kullback–Leibler απόκλιση) είναι ασυμπτωτικά ισοδύναμη με την επιλογή του μοντέλου $M_i, i = 1,...,k$ που έχει τη χαμηλότερη τιμή AIC. Το κριτήριο στη γενική του μορφή ορίζεται ως

$$AIC_i = -2\ln L_i + 2V_i \quad (3.3.2.1.1)$$

όπου L_i η μέγιστη πιθανοφάνεια για το υποψήφιο μοντέλο i, και V_i το πλήθος των συμμεταβλητών του μοντέλου. Η εξίσωση 3.3.2.1 δείχνει ότι το κριτήριο AIC περιγράφεται με ακρίβεια μέσω της μεγιστοποιημένης πιθανοφάνειας, και ποινικοποιεί την έλλειψη της φειδωλής παραμετροποίησης ανάλογα με τον αριθμό των ελεύθερων συμμεταβλητών. Στο εξής αναφέρεται ότι, επιλέγεται το μοντέλο που

δίνει τη μικρότερη τιμή του κριτηρίου. Η εξίσωση (3.3.2.1.1) βασίζεται σε ασυμπτωτικές προσεγγίσεις και είναι έγκυρη μόνο για επαρκώς μεγάλα σύνολα δεδομένων. Το επιδιορθωμένο κριτήριο του AIC το AIC_c που προτάθηκε από τους Burnham & Anderson (Burnham & Anderson,2002, p. 445) όταν το $n/V < 40$ ενώ για n μεγάλο, έχει παρόμοια συμπεριφορά με το AIC. (e.g., Hurvich & Tsai, 1995; Sugiura, 1978), το οποίο ορίζεται από τη σχέση

$$AIC_c = -2\ln L + 2V + \frac{2V(V+1)}{(n-V-1)} \quad (3.3.2.1.2)$$

3.3.2.2 Το κριτήριο BIC (Bayesian information criterion)

Παρά την ευρεία χρήση του κριτηρίου AIC, θεωρείται υπερβολικά φιλελεύθερη αφού τείνει στην επιλογή περίπλοκων μοντέλων (e.g., Kass & Raftery, 1995). Ωστόσο, επισημαίνεται ότι το κριτήριο AIC παραμελεί τη δειγματική μεταβλητότητα των εκτιμούμενων παραμέτρων και όταν οι τιμές της πιθανοφάνειας για αυτές τις παραμέτρους δεν παρουσιάζουν υψηλό βαθμό συγκέντρωσης γύρο από τη μέγιστη τιμή τους, αυτό μπορεί να οδηγήσει σε υπερβολικά αισιόδοξες εκτιμήσεις (βλέπε παράδειγμα των Aitchison & Dunsmore,1975, pp. 227–234). Επιπλέον, καθώς ο αριθμός παρατηρήσεων n γίνεται πολύ μεγάλος το κριτήριο AIC δεν φτάνει στη μικρότερη τιμή του(e.g., Bozdogan, 1987, p. 357). Συνεπώς μια εναλλακτική λύση επιλογής κριτηρίου είναι το Μπεϋσιανό κριτήριο πληροφορίας (Bayesian information criterion) ή αλλιώς κριτήριο BIC (βλέπε πχ Burnham & Anderson, 2002; Hastie, Tibshirani, & Friedman, 2001; Kass & Raftery, 1995; Schwarz,1978; Wasserman, 2000) το οποίο προτάθηκε από τον Schwartz (Schwartz, G. 1978). Για το μοντέλο i ορίζεται ως μοντέλο συρρίκνωσης και δίνεται από τον τύπο

$$BIC_i = -2\ln L_i + V_i \ln n \quad (3.3.2.2.1)$$

Όπου n είναι ο αριθμός των παρατηρήσεων που εισάγεται στον υπολογισμό της πιθανοφάνειας. Το κριτήριο BIC αποτελεί ασυμπτωτική προσέγγιση το κριτηρίου BSM (Bayesian model selection). Πιο συγκεκριμένα, για το μοντέλο BMS απαιτείται ο υπολογισμός της πιθανότητας $P(D \setminus M_i)$, ενσωματώνοντας τη μεταβλητότητα στη

παράμετρο $\theta_i : P(D \setminus M_i) = \int P(D \setminus \theta_i, M_i) \pi(\theta_i \setminus M_i) d\theta$, όπου $\pi(\theta_i \setminus M_i)$ είναι η προηγούμενη πυκνότητα. Επομένως, στον υπολογισμό είναι πιο εύκολο αφού δεν απαιτείται ο καθορισμός των προηγούμενων πυκνοτήτων για τις παραμέτρους. Σε αντίθεση με το κριτήριο AIC, το κριτήριο BIC λαμβάνει υπόψη του τη παράμετρο αβεβαιότητας.

Συγκρίνοντας τις εξισώσεις 1 και 3, φαίνεται ότι η ποινή BIC είναι μεγαλύτερη από αυτή του AIC όταν n> e^2 .Τα πλεονεκτήματα και τα μειονεκτήματα των κριτηρίων AIC και BIC έχουν κατά καιρούς συζητηθεί και από άλλους ερευνητές (βλέπε cf. Burnham & Anderson,2002, Kass & Raftery, 1995). Οι περισσότερες προσομοιώσεις δείχνουν ότι το κριτήριο BIC υπερτερεί του AIC αφού είναι ένα μοντέλο χαμηλών διαστάσεων.

3.3.3 Το κριτήριο $PRESS_P$

Το κριτήριο $PRESS_P$ (prediction sum of squares) προτάθηκε από τον Allen (Allen, D. M. 1971) και χρησιμοποιείται με σκοπό να ελεγχθεί η καταλληλότητα των ανεξάρτητων μεταβλητών στο να προβλέψουν την τιμή της εξαρτημένης μεταβλητής. Η διαδικασία εφαρμογής του κριτηρίου ξεκινάει με τη διαγραφή του πρώτου συνόλου i παρατηρήσεων για την εξαρτημένη και τις ανεξάρτητες μεταβλητές, προσαρμόζοντας όλα τα γραμμικά μοντέλα στις υπόλοιπες παρατηρήσεις. Εν συνεχεία, χρησιμοποιούμε κάθε γραμμικό μοντέλο για να προβλέψουμε το Y_1 από το Y_{1p} βρίσκοντας το εκτιμώμενο σφάλμα $Y_1 - \hat{Y}_{1p}$ για όλα τα δυνατά μοντέλα. Η διαδικασία επαναλαμβάνεται διαγράφοντας το δεύτερο σύνολο ώστε να πάρουμε το σφάλμα $Y_2 - \hat{Y}_{2p}$ και συνεχίζεται έτσι ώστε να πάρουμε τόσες διαγραφές όσες είναι και οι παρατηρήσεις μας.

Από τη στιγμή που λαμβάνονται όλες οι τιμές για τα εκτιμώμενα σφάλματα υπολογίζεται, για κάθε γραμμικό μοντέλο, το άθροισμα των τετραγώνων των σφαλμάτων αυτών

$$\sum_i^n \left(Y_i - \hat{Y}_{ip} \right)$$

Στο τελικό στάδιο της εφαρμογής του κριτηρίου $PRESS_P$ επιλέγεται το βέλτιστο γραμμικό μοντέλο δηλαδή το μοντέλο εκείνο που δίνει τη μικρότερη τιμή για το άθροισμα των τετραγώνων των εκτιμώμενων σφαλμάτων.

3.4 Ποινικοποιημένα ελάχιστα τετράγωνα και ποινικοποιημένη πιθανοφάνεια

Έχουμε αναφέρει ήδη ότι η επιλογή μεταβλητών παίζει καθοριστικό ρόλο τόσο για τα παραμετρικά μοντέλα παλινδρόμησης όσο και για τα μη παραμετρικά. Στην ενότητα αυτή εστιάζουμε στην επιλογή μεταβλητών μέσω της παραμετρικής παλινδρόμησης, συμπεριλαμβάνοντας τα γραμμικά μοντέλα παλινδρόμησης, τα εύρωστα και τα γενικευμένα γραμμικά μοντέλα.

Η μέθοδος επιλογής μεταβλητών μέσω ποινικοποιημένων ελαχίστων τετραγώνων συνδέεται στενά με την επιλογή μεταβλητών στα γραμμικά μοντέλα παλινδρόμησης. Σε περίπτωση που ο πίνακας σχεδιασμού είναι ορθοκανονικός, η κατά βήματα μέθοδος απαλοιφής μεταβλητών και η μέθοδος επιλογής καλύτερου υποσυνόλου είναι ισοδύναμες με αυτή του Hard οριακού κανόνα. Ο εκτιμητής του Hard οριακού κανόνα μπορεί να θεωρηθεί ως μια λύση του προβλήματος ποινικοποιημένων ελαχίστων τετραγώνων, όπως αναλύεται στη παράγραφο 3.4.1. Επιπλέον, μπορεί να θεωρηθεί ως μη συνεχής και για αυτό το λόγο μπορεί να διορθωθεί χρησιμοποιώντας έναν soft οριακό κανόνα που μηδενίζει τους μικρούς συντελεστές και συρρικνώνει την εκτίμηση μέσω μιας μεταβλητής. Κατά τον Breiman (1996), η μέθοδος επιλογής καλύτερου υποσυνόλου μπορεί να θεωρηθεί ως μια κανονικοποιημένη τεχνική και έχει ορισμένα μειονεκτήματα, εκ των οποίων ένα από τα πιο σοβαρά είναι η έλλειψη σταθερότητας. Σε αντίθεση με τον Hard οριακό κανόνα που οδηγεί σε ένα μη σταθερό μοντέλο αφού για κάθε μικρή αλλαγή δεδομένων προκύπτει ένα διαφορετικό μοντέλο. Αυτό έχει ως αποτέλεσμα την πρόβλεψη μεγάλης μεταβλητότητας. Από την άλλη, καθώς ο soft οριακός κανόνας είναι συνεχής, μετατοπίζει μια εκτίμηση από μια σταθερά. Ως εκ τούτου, δημιουργείται μεροληψία αν η οριακή παράμετρος είναι μεγάλη. Ο Fan περιγράφει, ακολουθώντας το ίδιο σκεπτικό με τους Antoniadis (1999) και Bruce και Cao (1997), μερικούς οριακούς κανόνες που βελτιώνουν τις ιδιότητες τόσο του Hard όσο και του soft οριακού κανόνα. Αυτοί οι κανόνες μπορούν επίσης να θεωρηθούν ως

ποινικοποιημένα ελάχιστα τετράγωνα. Επιπλέον, προτείνεται η συνάρτηση ποινής SCAD για τη βελτίωση της L_1 και της Hard συνάρτησης ποινής.

3.4.1 Επιλογή μεταβλητών μέσω ποινικοποιημένων ελαχίστων τετραγώνων

Θεωρούμε το γνωστό γραμμικό μοντέλο παλινδρόμησης

$$\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

όπου ,

- $\underset{\sim}{Y}$ είναι ένα $n\times 1$ διάνυσμα των παρατηρήσεων
- $\underset{\sim}{X}$ είναι ένας $n\times d$ πίνακας των επεξηγηματικών μεταβλητών
- $\underset{\sim}{\beta}$ ένα $d\times 1$ διάνυσμα των συντελεστών παλινδρόμησης
- $\underset{\sim}{\varepsilon}$ είναι ένα $n\times 1$ διάνυσμα τυχαίων σφαλμάτων.

Όμοια με τη περίπτωση του μοντέλου γραμμικής παλινδρόμησης και δοθέντων των x_{ij}, θεωρούμε τα y_i ανεξάρτητα καθώς και οι στήλες του πίνακα $\underset{\sim}{X}$ είναι ορθοκανονικές (*orthonormal*). Ο υπολογισμός της εκτιμήτριας γίνεται μέσω της ελαχιστοποίησης της ποσότητας $\|\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\beta}\|^2$,η οποία ισοδυναμεί με την ποσότητα $\|\underset{\sim}{\hat{\beta}} - \underset{\sim}{\beta}\|^2$,όπου $\underset{\sim}{\hat{\beta}} = \underset{\sim}{X}'\underset{\sim}{Y}$ είναι η *OLS* (*ordinary least squares*) εκτιμήτρια.

Υποθέτουμε ότι $\underset{\sim}{z} = \underset{\sim}{X}'\underset{\sim}{Y}$ και $\underset{\sim}{\hat{Y}} = \underset{\sim}{X}\underset{\sim}{X}'\underset{\sim}{Y}$, μια μορφή των ποινικοποιημένων ελαχίστων τετραγώνων είναι η εξής:

$$\frac{1}{2}\|\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\beta}\|^2 + \lambda\sum_{j=1}^{d} p_j(|\beta_j|) = \frac{1}{2}\|\underset{\sim}{Y} - \underset{\sim}{\hat{Y}}\|^2 + \frac{1}{2}\sum_{j=1}^{d}(z_j - \beta_j)^2 + \lambda\sum_{j=1}^{d} p_j(|\beta_j|) \quad (3.4.1.1).$$

Ωστόσο, οι συναρτήσεις ποινής p_j στην (3.4.1.1) δεν είναι απαραίτητα οι ίδιες για όλα τα j. Μπορεί όμως, σε ένα παραμετρικό μοντέλο, να θέλουμε να κρατήσουμε ορισμένες σημαντικές μεταβλητές και για αυτό το λόγο να μη ποινικοποιήσουμε τις αντίστοιχες παραμέτρους τους. Επομένως, οι συναρτήσεις ποινής είναι οι ίδιες για όλους τους συντελεστές, και θα συμβολίζονται ως $p(|\cdot|)$. Επίσης, αντί $\lambda p(|\cdot|)$ θα χρησιμοποιείται ο συμβολισμός $p_\lambda(|\cdot|)$, δείχνοντας έτσι ότι το $p(|\cdot|)$ εξαρτάται από το λ.

Το πρόβλημα ελαχιστοποίησης της (3.4.1.1) είναι ισοδύναμο με την ελαχιστοποίηση των συνιστωσών. Οπότε θεωρούμε το παρακάτω πρόβλημα ελαχίστων τετραγώνων

$$\frac{1}{2}(z-\theta)^2 + p_\lambda(|\theta|) \quad (3.4.1.2).$$

Επιπλέον, με τη χρήση της *Hard* συνάρτηση ποινής (βλ. ***σχήμα 3.1 (α)***)

$$p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda), (3.4.1.3).$$

προκύπτει η *Hard* εκτιμήτρια (βλ. Andoniadis 1997 & Fan 1997).

$$\hat{\theta} = zI(|z| > \lambda) \quad (3.4.1.4).$$

(βλ. ***σχήμα 3.2(α)***). Με άλλα λόγια, η λύση της (3.4.1.1) είναι

$$z_j I(|z_j| > \lambda)$$

η οποία συμπίπτει με την επιλογή καλύτερου υποσυνόλου και την κατά βήματα πρόσθεση και απαλοιφή στους ορθοκανονικούς σχεδιασμούς. Σημειώνουμε επιπλέον πως η συνάρτηση ποινής *Hard* είναι ομαλότερη από την συνάρτηση ποινής εντροπίας (*entropy penalty*)

$$p_\lambda(|\theta|) = \left(\frac{\lambda^2}{2}\right) I(|\theta| \neq 0),$$

η οποία και αυτή οδηγεί στη λύση (3.4.1.4).

Μια συνάρτηση ποινής για να είναι καλή, πρέπει οι εκτιμήσεις της να έχουν τις παρακάτω ιδιότητες:

1. *Αμεροληψία:* Ο προκύπτων εκτιμητής πρέπει να είναι σχεδόν αμερόληπτος, ιδίως στην περίπτωση όπου η σωστή άγνωστη παράμετρος β_j είναι μεγάλη ώστε με αυτό τον τρόπο να αποφεύγεται η μεροληψία του μοντέλου.
2. *Σποραδικότητα*: Ο προκύπτων εκτιμητής πρέπει να αποτελεί κανόνα περιορισμού (*thresholding rule*), ώστε οι εκτιμώμενοι συντελεστές με μικρή τιμή, να μηδενίζονται. Έτσι, μειώνεται η πολυπλοκότητα του μοντέλου.
3. *Συνέχεια:* Ο προκύπτων εκτιμητής πρέπει είναι συνεχής. Αποφεύγεται κατά αυτόν τον τρόπο η αστάθεια στη πρόβλεψη του μοντέλου.

Καταρχήν η πρώτη παράγωγος της (3.4.1.2). ως προς θ είναι

$$\text{sgn}(\theta)\{|\theta|+p'_\lambda(|\theta|)\}-z\,.$$

Εύκολα παρατηρεί κανείς ότι όταν $p'_\lambda(|\theta|)=0$ για μεγάλο $|\theta|$, τότε ο προκύπτων εκτιμητής είναι ίσος με z όταν το $|z|$ είναι επαρκώς μεγάλο. Για αυτό το λόγο, όταν η πραγματική παράμετρος $|\theta|$ είναι μεγάλη, η τιμή $|z|$ είναι και αυτή μεγάλη και με μεγάλη πιθανότητα. Οπότε, ο *PLS* (*penalized least squares*) εκτιμητής είναι

$$\hat{\theta}=z\,,$$

ο οποίος και είναι σχεδόν αμερόληπτος. Συνεπώς, η προϋπόθεση $p'_\lambda(|\theta|)=0$ για μεγάλο $|\theta|$, είναι μια επαρκής προϋπόθεση για την αμεροληψία μιας μεγάλης πραγματικής παραμέτρου. Όσον αφορά τη δεύτερη ιδιότητα, για να αποτελεί ο προκύπτων εκτιμητής κανόνα περιορισμού, πρέπει να ισχύει ότι

$$\min_\theta\{|\theta|+p'_\lambda(|\theta|)\}>0\,.$$

Το ***σχήμα 3.3*** παρέχει περισσότερες εξηγήσεις σχετικά με αυτό. Όταν τώρα

$$|z|<\min_{\theta\neq 0}\{|\theta|+p'_\lambda(|\theta|)\}$$

η παράγωγος της (3.4.1.2) είναι θετική για όλα τα θετικά θ και αρνητική για όλα τα αρνητικά θ. Οπότε σε αυτήν την περίπτωση, ο *PLS* εκτιμητής $\hat{\theta}$ είναι μηδέν. Όταν όμως $|z|>\min_{\theta\neq 0}\{|\theta|+p'_\lambda(|\theta|)\}$, δύο διασταυρώσεις (*crossings*) μπορούν να υπάρξουν, όπως φαίνεται και στο ***σχήμα 3.1***. Η μεγαλύτερη είναι ο *PLS* εκτιμητής. Αυτό συνεπάγεται ότι ικανή και αναγκαία συνθήκη για την ύπαρξη συνέχειας είναι το $\min_\theta\{|\theta|+p'_\lambda(|\theta|)\}$ να πετυχαίνεται στο μηδέν. Από αυτό αντιλαμβανόμαστε πως η συνάρτηση ποινής που ικανοποιεί τις ιδιότητες της σποραδικότητας και της συνέχειας, πρέπει να είναι ιδιάζουσα (*singular*) στην αρχή.

Είναι γνωστό πως η συνάρτηση ποινής L_2

$$p_\lambda(|\theta|)=\lambda|\theta|^2$$

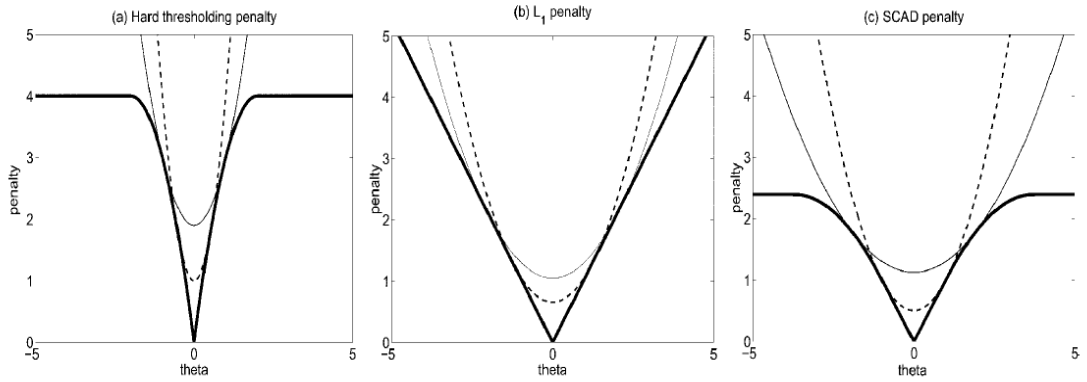
οδηγεί στην παλινδρόμηση κορυφογραμμής. Η συνάρτηση ποινής L_1, οδηγεί στον *soft* οριακό κανόνα

$$\hat{\theta}_j=\text{sgn}(z_j)(|z_j|-\lambda)_+\,,\qquad (3.4.1.5).$$

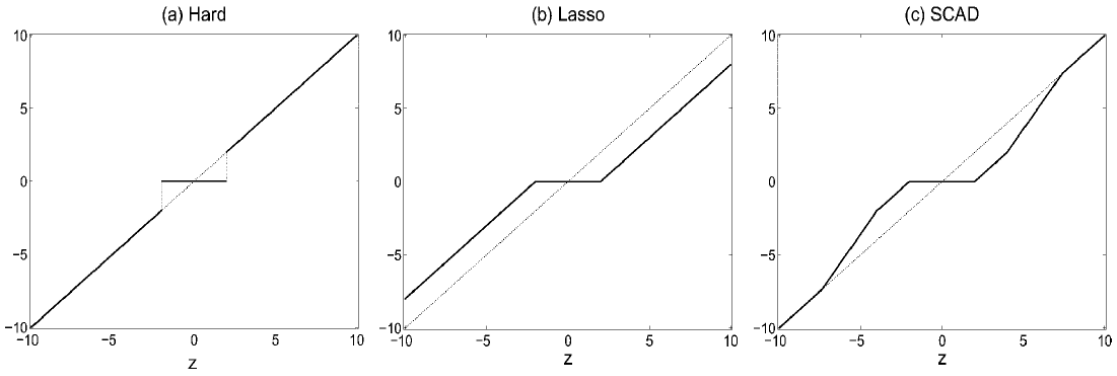
που προτάθηκε από τους Donoho και Johnstone (1994a). Η *LASSO* που προτείνεται από τον Tibshirani (1996, 1997), είναι ο *PLS* εκτιμητής με συνάρτηση ποινής την L_1 (βλ. ενότητα 3.6.2). Επίσης, η L_q συνάρτηση ποινής

$$p_\lambda(|\theta|) = \lambda |\theta|^q$$

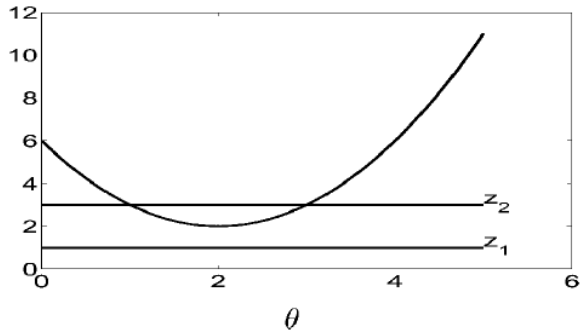
οδηγεί στην παλινδρόμηση *bridge* (Frank& Friedman 1993, Fu 1998) . Η λύση είναι συνεχής μόνο για $q \geq 1$. Παρόλα αυτά, όταν $q > 1$, δεν παράγεται μια σποραδική λύση (βλ. ***σχήμα 3.4(a)***). Η μόνη συνεχής λύση με κανόνα περιορισμού σε αυτή την οικογένεια συναρτήσεων είναι με τη συνάρτηση ποινής L_1, αυτό όμως προκύπτει μεταβάλλοντας τον εκτιμητή κατά μια σταθερά λ, άρα χάνεται και η αμεροληψία (βλ. ***σχήμα 3.2(b)***). Επίσης για $0 \leq q < 1$, δεν ικανοποιείται η συνθήκη της συνέχειας.



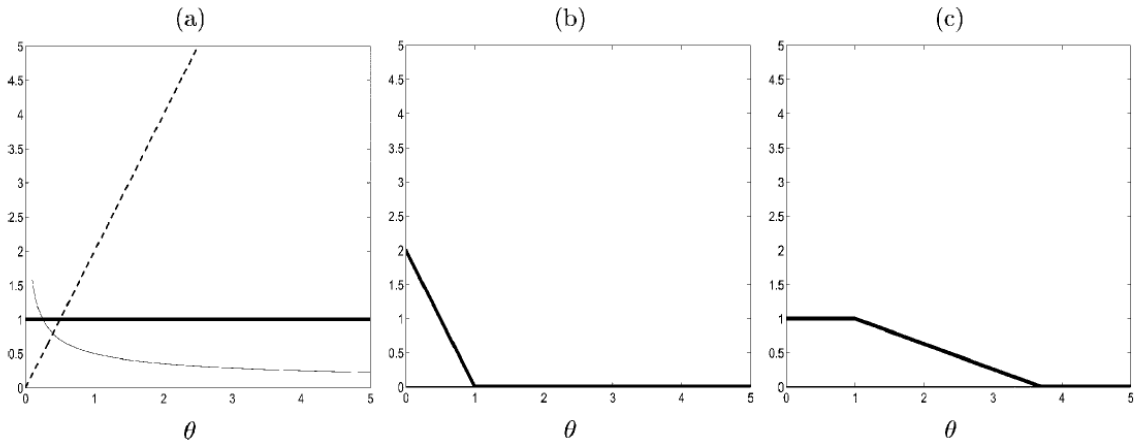
Σχήμα 3.1: (a) Οι τρεις συναρτήσεις ποινής και οι τετραγωνικές τους προσεγγίσεις.



Σχήμα 3.2: Οι εκτιμήτριες (thresholding functions) (a) Hard, (b) Soft ή LASSO και (c) Scad, όπου για την τελευταία λ=2 και a=3.7.



Σχήμα 3.3: Η συνάρτηση $\theta + p_\lambda(|\theta|)$ ***ως προς*** θ**.**



Σχήμα 3.4: Οι συναρτήσεις $p'_\lambda(|\theta|)$ ***ως προς*** θ***, για (a) τις συναρτήσεις ποινής*** L_q**,**
(b) τη Hard συνάρτηση ποινής και (c) τη SCAD. Στο (a), η παχιά γραμμή αντιστοιχεί στην L_1**,** ***η διακεκομμένη στην*** $L_{0.5}$ ***και η λεπτή γραμμή στην*** L_2 ***συνάρτηση ποινής.***

3.4.1.1 Η συνάρτηση ποινής SCAD

Οι Fan και Li (2001) εισήγαγαν τη *SCAD* (*Smoothly Clipped Absolute Deviation penalty*), η οποία είναι μια συνεχής και διαφορίσιμη συνάρτηση ποινής, που ως σκοπό έχει τη βελτίωση της L_1 και της *Hard* . Ωστόσο, όπως φαίνεται και στα παρακάτω ***σχήματα (a), (b)*** και ***(c)*** οι συναρτήσεις ποινής L_q και Hard δεν ικανοποιούν και τις τρεις απαιτήσεις της αμεροληψίας, της σποραδικότητας και της συνέχειας (επίσης βλ. ***σχήμα 3.1. (c)***).

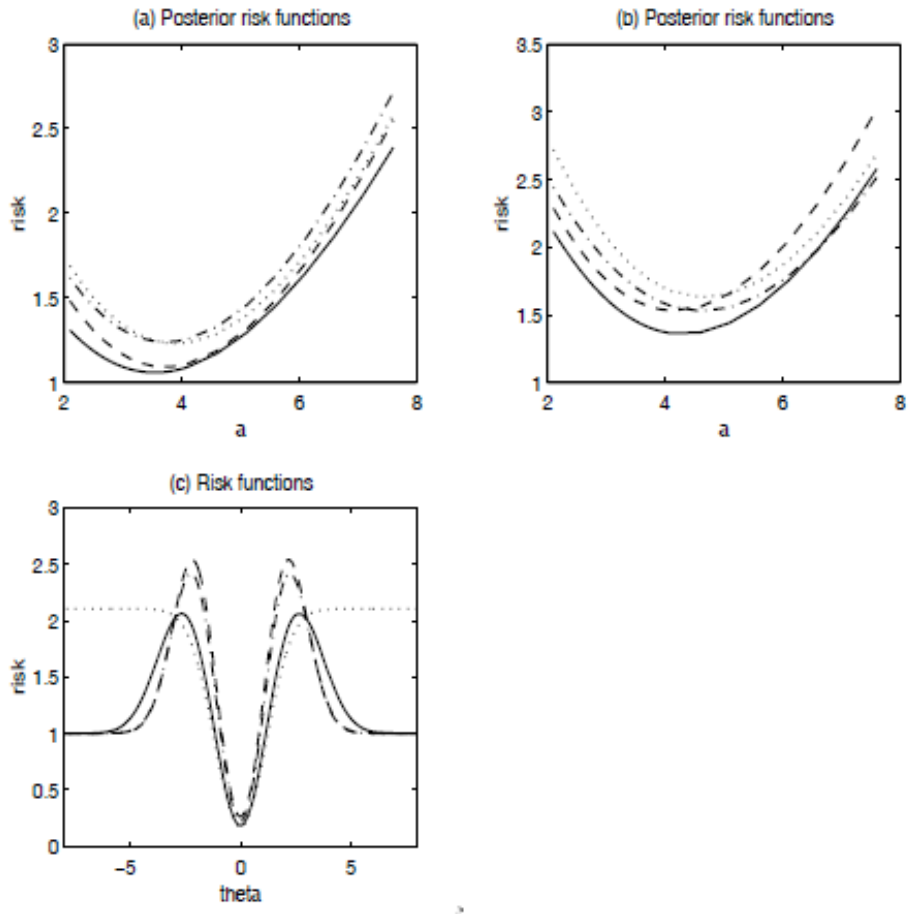
Η συνάρτηση ποινής SCAD ορίζεται ως

$$p'_\lambda(\theta) = \lambda\left\{ I(\theta \le \lambda) + \frac{(\alpha\lambda - \theta)_+}{(\alpha - 1)\lambda} I(\theta > \lambda) \right\}, \text{ για κάποιο } a > 2 \text{ και } \theta > 0.$$

Η συγκεκριμένη συνάρτηση δεν ποινικοποιεί υπερβολικά τις μεγάλες τιμές του θ και δίνει μια συνεχής λύση (δόθηκε από τον Fan(1997)), την

$$\hat{\theta} = \begin{cases} \text{sgn}(z)(|z| - \lambda)_+, & |z| \le 2\lambda \\ \{(\alpha - 1)z - \text{sgn}(z)\alpha\lambda\} / (\alpha - 2), & 2\lambda < |z| \le \alpha\lambda \\ z, & |z| > \alpha\lambda \end{cases} \quad (3.4.1.1.1)$$

Η παραπάνω λύση περιλαμβάνει δύο άγνωστες παραμέτρους α και λ. Πρακτικά θα μπορούσε να υπολογιστεί το βέλτιστο ζεύγος (α, λ) βάσει ορισμένων κριτηρίων, όπως της διασταυρωμένης επικύρωσης (cross-validation) και της γενικευμένης διασταυρωμένης επικύρωσης (generalized cross-validation). Όμως θα ήταν αρκετά χρονοβόρο και γι αυτό το λόγο οι Fan και Li (2001), χρησιμοποιώντας εργαλεία Μπεϋζιανής ανάλυσης ρίσκου κατέληξαν στην επιλογή του $\alpha = 3.7$.



Σχήμα 3.5 : συνάρτηση κινδύνου προτεινόμενων μεθόδων μέσω τετραγωνικής απώλειας. Τα (a) και (b) παρουσιάζουν τις μεταγενέστερες συναρτήσεις κινδύνου του L_1 εκτιμητή με θ~N(0, αλ) και λ=$\sqrt{2\log(d)}$ για 4 διαφορετικές τιμές του d,τα 4 είδη των γραμμών (20,40,60,100) αντίστοιχα. (b) είναι ίδιο με το (a) για d= 512, 1024, 2048, 4096 . (c) συναρτήσεις κινδύνου των τεσσάρων διαφορετικών οριακών κανόνων, SCAD , Hard, soft, mixture αντιστοιχούν στην πυκνή, διακεκομμένη, κουκίδα, παύλα κουκίδα γραμμή.

3.4.1.2 Απόδοση των οριακών κανόνων

Οι Marron, Adak, Johnstone , Neumann and Patil, (1998) για να κατανοήσουν τη συμπεριφορά των *Hard* και *Soft* οριακών κανόνων για μικρά δείγματα, εφάρμοσαν την ανάλυση ρίσκου. Επιπλέον, οι Fan και Li (2000) για να καταστήσουν τη παράμετρο κλίμακας των οριακών παραμέτρων συγκρίσιμη, πήραν την τιμή $\lambda = 2$ για τον *Hard* οριακό κανόνα και προσάρμοσαν το λ για τους άλλους δύο οριακούς

κανόνες, ώστε να δίνουν τις ίδιες εκτιμήσεις για την περίπτωση όπου $\theta = 3$. Συμπερασματικά λοιπόν, η συνάρτηση ποινής *SCAD* συμπεριφέρεται εξίσου καλά σε σχέση με τους άλλους δύο οριακούς κανόνες. Αυτό εξάλλου μπορεί εύκολά να φανεί μέσω των αντίστοιχων συναρτήσεων ποινών που παρουσιάζονται στο ***σχήμα 3.1.***

3.4.2 Επιλογή μεταβλητών μέσω ποινικοποιημένης πιθανοφάνειας

Η μεθοδολογία της προηγούμενης παραγράφου μπορεί να εφαρμοσθεί σε πολλά στατιστικά μοντέλα. Στην ενότητα αυτή θα εξετάσουμε τα γραμμικά μοντέλα παλινδρόμησης (*linear regression models*), εύρωστα γραμμικά μοντέλα (*robust linear models*) και γενικευμένα γραμμικά μοντέλα βασισμένα στην πιθανοφάνεια (*likelihood-based generalized linear models*). Ο πίνακας σχεδιασμού $\underset{\sim}{X} = (x_{ij})$ θεωρείται κανονικοποιημένος, δηλαδή κάθε στήλη να έχει μέση τιμή 0 και διασπορά 1.

3.4.2.1 Ποινικοποιημεάνα ελάχιστα τετράγωνα και πιθανοφάνεια

Σε περίπτωση όπου ο πίνακας σχεδιασμού δεν είναι ορθοκανονικός (*orthonormal*) η (3.4.1.1) μπορεί να επεκταθεί και να πάρει μια ισοδύναμη μορφή που είναι

$$\frac{1}{2}(\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\beta})'(\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\beta}) + n\sum_{j=1}^{d} p_{\lambda}(|\beta_j|) \quad (3.4.2.1.1).$$

Ελαχιστοποιώντας την (3.4.2.1.1) ως προς $\underset{\sim}{\beta}$, λαμβάνουμε τον εκτιμητή ποινικοποιημένων ελαχίστων τετραγώνων του $\underset{\sim}{\beta}$.

<u>*Για τα εύρωστα γραμμικά μοντέλα*</u> : Επειδή ο *OLS* εκτιμητής δεν είναι εύρωστος, θεωρούμε τη συνάρτηση ψ του Huber (βλ.Huber 1981) οπότε αντί της ελαχιστοποίησης της (3.4.2.1.1), μπορούμε να ελαχιστοποιήσουμε την (3.4.2.1.2) ως προς $\underset{\sim}{\beta}$, ώστε να πάρουμε έναν εύρωστο ποινικοποιημένο εκτιμητή του $\underset{\sim}{\beta}$.

$$\sum_{i=1}^{n} \psi(|y_i - \underset{\sim}{x}_i'\underset{\sim}{\beta}|) + n\sum_{j=1}^{d} p_{\lambda}(|\beta_j|) \quad (3.4.2.1.2)$$

Για τα γενικευμένα γραμμικά μοντέλα: Η επιλογή των σημαντικών μεταβλητών γίνεται βάση του ποινικοποιημένου εκτιμητή μέγιστης πιθανοφάνειας. Έστω ότι τα δεδομένα $(\underset{\sim}{x}_i, Y_i)$ που έχουν συλλεχθεί είναι ανεξάρτητα. Δεδομένων των $\underset{\sim}{x}_i$, η Y_i έχει συνάρτηση πιθανοφάνειας $f_i(g(\underset{\sim}{x}_i'\underset{\sim}{\beta}), y_i)$, όπου g είναι μια γνωστή συνάρτηση σύνδεσης. Έστω και ότι $l_i = \log f_i$ είναι ο λογάριθμος της πιθανοφάνειας του Y_i. Επομένως, η ποινικοποιημένη πιθανοφάνεια ορίζεται ως

$$\sum_{i=1}^{n} l_i(g(\underset{\sim}{x}_i'\underset{\sim}{\beta}), y_i) - n\sum_{j=1}^{d} p_\lambda(|\beta_j|).$$

Η μεγιστοποίηση της ως άνω συνάρτησης, είναι ισοδύναμη με την ελαχιστοποίηση της ως προς $\underset{\sim}{\beta}$ για να πάρουμε τον ποινικοποιημένο εκτιμητή μέγιστης πιθανοφάνειας για κάποια οριακή παράμετρο λ.

$$-\sum_{i=1}^{n} l_i(g(\underset{\sim}{x}_i'\underset{\sim}{\beta}), y_i) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|) \quad (3.4.2.1.3)$$

3.4.2.2 Δειγματοληπτικές και προβλεπτικές ιδιότητες

Στην ενότητα αυτή αναπτύσσεται η ασυμπτωτική θεωρία του μη κοίλου εκτιμητή ποινικοποιημένης πιθανοφάνειας. Έστω

$$\underset{\sim}{\beta}_0 = (\beta_{10}, ..., \beta_{d0})' = (\underset{\sim}{\beta}_{10}', \underset{\sim}{\beta}_{20}')'$$

και χωρίς βλάβη της γενικότητας έστω $\underset{\sim}{\beta}_{20} = \underset{\sim}{0}$, με

- $I(\underset{\sim}{\beta}_0)$ ο πίνακας πληροφορίας του Fisher (*Fisher information matrix*)
- $I_1(\underset{\sim}{\beta}_{10}, \underset{\sim}{0})$ η πληροφορία κατά Fisher.

Καταρχήν αποδεικνύεται ότι υπάρχει ένας εκτιμητής ποινικοποιημένης πιθανοφάνειας που συγκλίνει στο

$$O_p(n^{-1/2} + \alpha_n) \quad (3.4.2.2.1).$$

όπου $\alpha_n = \max\left\{p'_{\lambda_n}(|\beta_{j0}|) : \beta_{j0} \neq 0\right\}$. Συνεπώς, για τις *Hard* και *SCAD* συναρτήσεις ποινής, ο εκτιμητής ποινικοποιημένης πιθανοφάνειας είναι $\sqrt{n}$-συνεπής (*root-n consistent*) αν $\lambda_n \to 0$.

Επιπλέον, αποδεικνύεται ότι για τον εκτιμητή αυτόν πρέπει να ισχύει ότι $\hat{\underset{\sim}{\beta}}_2 = \underset{\sim}{0}$ και ότι το $\hat{\underset{\sim}{\beta}}_1$ είναι ασυμπτωτικά της κανονικής κατανομής με πίνακα συνδιασποράς I_1^{-1}, αν $n^{1/2}\lambda_n \to \infty$.Συνεπώς, ο εκτιμητής ποινικοποημένης πιθανοφάνειας συμπεριφέρεται τόσο καλά όσο αν ήταν γνωστό ότι $\underset{\sim}{\beta}_{20} = \underset{\sim}{0}$.

Θεωρούμε το απλούστερο γραμμικό μοντέλο παλινδρόμησης

$$\underset{\sim}{Y} = \underset{\sim}{1}_n \mu + \underset{\sim}{\varepsilon},$$

όπου $\underset{\sim}{\varepsilon} \sim N_n(\underset{\sim}{0}, I_n)$.

Ένας υπέρ-αποδοτικός εκτιμητής για το μ είναι (βλ. Lehmann 1983, p.405)

$$\delta_n = \begin{cases} \bar{Y}, & |\bar{Y}| \geq n^{-1/4} \\ c\bar{Y}, & |\bar{Y}| < n^{-1/4} \end{cases}.$$

Θέτοντας $c = 0$, παρατηρούμε ότι το δ_n συμπίπτει με τον *Hard* εκτιμητή με παράμετρο $\lambda_n = n^{-1/4}$ ο οποίος υπολογίζει ακριβώς την παράμετρο στο 0 και σε κανένα άλλο σημείο.

Μια γενίκευση του αποτελέσματος αυτού είναι η πραγματοποίηση της ποινικοποίησης σε κάθε συνιστώσα του $\underset{\sim}{\beta}$. Σε περίπτωση όπου κάποιες συνιστώσες δεν ποινικοποιούνται, όπως για παράδειγμα η διασπορά στο γραμμικό μοντέλο, δεν παρουσιάζει κάποιο πρόβλημα.

Έστω λοιπόν $\underset{\sim}{V}_i = (\underset{\sim}{X}_i, Y_i)$, με $i = 1, ..., n$ και ότι $L(\underset{\sim}{\beta})$ είναι ο λογάριθμος της πιθανοφάνειας των παρατηρήσεων $\underset{\sim}{V}_1, ..., \underset{\sim}{V}_n$. Έστω επίσης ότι

$$Q(\underset{\sim}{\beta}) = L(\underset{\sim}{\beta}) - n\sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|),$$

είναι η ποινικοποιημένη συνάρτηση πιθανοφάνειας.

Παρακάτω παρουσιάζονται τα θεωρήματα και λήμματα των Fan και Li (2001), οι αποδείξεις αυτών καθώς και οι υποθέσεις κανονικότητας (*regularity conditions*) αναφέρονται στο παράρτημα Ι.

Θεώρημα 1

Έστω ότι τα $\underset{\sim}{V}_1,...,\underset{\sim}{V}_n$ *είναι i.i.d. (independent and identically distributed), κάθε ένα με συνάρτηση πυκνότητας πιθανότητας* $f(\underset{\sim}{V},\underset{\sim}{\beta})$ *και ότι ικανοποιούν τις παραπάνω υποθέσεις (A)-(C). Αν* $\max\left\{|\, p''_{\lambda_n}(|\,\beta_{j0}\,|)\,|: \beta_{j0}\neq 0\right\}\to 0$, *τότε υπάρχει ένα τοπικό μέγιστο* $\hat{\underset{\sim}{\beta}}$ *του* $Q(\underset{\sim}{\beta})$ *τέτοιο ώστε*

$$\|\,\hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta}_0\,\|=O_p(n^{-1/2}+\alpha_n),$$

με το α_n *να δίνεται από την* (3.4.2.2.1).

Από το θεώρημα αυτό είναι προφανές ότι με μια σωστή επιλογή του λ_n θα υπάρξει ένας $\sqrt{n}$-συνεπής ποινικοποιημένος εκτιμητής καθώς το $\alpha_n=O(n^{-1/2})$.

Ορίζουμε τώρα ως

$$\Sigma=diag\left\{p''_{\lambda_n}(|\,\beta_{10}\,|,...,(|\,\beta_{s0}\,|)\right\}$$

και

$$\underset{\sim}{b}=\left(p'_{\lambda_n}(|\,\beta_{10}\,|)\mathrm{sgn}(\beta_{10}),...,\,p'_{\lambda_n}(|\,\beta_{s0}\,|)\mathrm{sgn}(\beta_{s0})\right)'.$$

Όπου s είναι ο αριθμός των συμμεταβλητών του β_{10}.

Θεώρημα 2 (Προβλεπτική ιδιότητα)

Θεωρούμε ξανά ότι τα $\underset{\sim}{V}_1,...,\underset{\sim}{V}_n$ *είναι i.i.d, κάθε ένα με συνάρτηση πυκνότητας πιθανότητας* $f(\underset{\sim}{V},\underset{\sim}{\beta})$ *και ότι ικανοποιούν τις υποθέσεις (A)-(C). Έστω επίσης ότι η συνάρτηση ποινής* $p_{\lambda_n}(|\,\theta\,|)$ *ικανοποιεί τη συνθήκη*

$\liminf_{n\to\infty}\liminf_{\theta\to 0_+} p'_{\lambda_n}(\theta)/\lambda_n>0.$

Αν $\lambda_n\to 0$ *και* $\sqrt{n}\lambda_n\to\infty$ *όσο το* $n\to\infty$, *τότε με πιθανότητα που τείνει στο* 1, *οι* $\sqrt{n}$*-συνεπείς εκτιμητές* $\hat{\underset{\sim}{\beta}}=\begin{pmatrix}\hat{\underset{\sim}{\beta}}_1\\ \hat{\underset{\sim}{\beta}}_2\end{pmatrix}$, *του* ***Θεωρήματος 1****, πρέπει να ικανοποιούν τα παρακάτω:*

- *Σποραδικότητα (sparsity):* $\hat{\underset{\sim}{\beta}}_2=\underset{\sim}{0}$.

- *Ασυμπτωτική κανονικότητα (assymptotic normality):*

$$\sqrt{n}\left(I_1(\underset{\sim}{\beta}_{10})+\Sigma\right)\left\{\hat{\underset{\sim}{\beta}}_1-\underset{\sim}{\beta}_{10}+\left(I_1(\underset{\sim}{\beta}_{10})+\Sigma\right)^{-1}\underset{\sim}{b}\right\}\to \mathrm{N}\left\{\underset{\sim}{0}, I_1(\underset{\sim}{\beta}_{10})\right\},$$

όπου

$$I_1(\underset{\sim}{\beta}_{10})=I_1(\underset{\sim}{\beta}_{10},\underset{\sim}{0})$$

η πληροφορία κατά Fisher, γνωρίζοντας ότι $\underset{\sim}{\beta}_2=\underset{\sim}{0}$.

Συνεπώς, ο ασυμπτωτικός πίνακας συνδιασποράς του $\hat{\underset{\sim}{\beta}}_1$ είναι

$$\frac{1}{n}\left\{I_1(\underset{\sim}{\beta}_{10})+\Sigma\right\}^{-1}I_1(\underset{\sim}{\beta}_{10})\left\{I_1(\underset{\sim}{\beta}_{10})+\Sigma\right\}^{-1},$$

και για τις συναρτήσεις ποινής που ανεπτύχθησαν στην ενότητα 3.4.1, είναι προσεγγιστικά ίσος με

$$\frac{1}{n}I_1^{-1}(\underset{\sim}{\beta}_{10}) \text{ αν το } \lambda_n\to 0.$$

Σημείωση 1: Για τις *SCAD* και *Hard* συναρτήσεις ποινής, αν $\lambda_n\to 0$ τότε $\alpha_n=0$. Οπότε βάσει του **Θεωρήματος 2**, όταν $\sqrt{n}\lambda_n\to\infty$, οι αντίστοιχοι εκτιμητές ποινικοποιημένης πιθανοφάνειας έχουν την προβλεπτική ιδιότητα (*oracle property*) και συμπεριφέρονται τόσο καλά όσο και οι εκτιμητές μέγιστης πιθανοφάνειας, όσον αφορά την εκτίμηση του $\underset{\sim}{\beta}_1$, δεδομένου ότι $\underset{\sim}{\beta}_2=\underset{\sim}{0}$. Παρόλα αυτά, για την L_1 συνάρτηση ποινής, ισχύει ότι $\alpha_n=\lambda_n$. Οπότε, η $\sqrt{n}$ -συνέπεια απαιτεί $\lambda_n=O_p(n^{-1/2})$. Όμως, η προβλεπτική ιδιότητα του **Θεωρήματος 2** απαιτεί $\sqrt{n}\lambda_n\to\infty$. Οι δύο αυτές συνθήκες για τη *LASSO* δεν ικανοποιούνται ταυτόχρονα. Συνεπώς, δεν ισχύει η προβλεπτική ιδιότητα για την L_1 συνάρτηση ποινής. Αντιθέτως, για την L_q συνάρτηση ποινής, με $q<1$, η προβλεπτική ιδιότητα ισχύει αν έχουμε επιλέξει το σωστό λ_n.

Συνεχίζουμε, κάνοντας μια αναφορά περί των συνθηκών κανονικότητας (A)-(C), όσον αφορά τα γενικευμένα γραμμικά μοντέλα. Με μια *canonical link*, η κατανομή του Y δεδομένου ότι $\underset{\sim}{X}=\underset{\sim}{x}$, ανήκει στην *canonical* εκθετική οικογένεια, με συνάρτηση πυκνότητας πιθανότητας

$$f(y,\underset{\sim}{x},\underset{\sim}{\beta})=c(y)\exp\left\{\frac{y\underset{\sim}{x}'\underset{\sim}{\beta}-b(\underset{\sim}{x}'\underset{\sim}{\beta})}{\alpha(\varphi)}\right\}.$$

Προφανώς, η συνθήκη (A) ικανοποιείται. Ο πίνακας πληροφορίας του Fisher είναι

$$I(\underset{\sim}{\beta}) = E\left\{b''(\underset{\sim}{x}'\underset{\sim}{\beta})\underset{\sim}{x}\underset{\sim}{x}'\right\}/\alpha(\varphi).$$

Οπότε αν το $E\left\{b''(\underset{\sim}{x}'\underset{\sim}{\beta})\underset{\sim}{x}\underset{\sim}{x}'\right\}$ είναι πεπερασμένο και θετικά ορισμένο, τότε ισχύει και η συνθήκη (B). Επίσης, αν για όλα τα $\underset{\sim}{\beta}$ σε κάποια γειτονιά του $\underset{\sim}{\beta}_0$, ισχύει ότι

$$|b^{(3)}(\underset{\sim}{x}'\underset{\sim}{\beta})| \le M_0(\underset{\sim}{x})$$

για κάποια συνάρτηση $M_0(\underset{\sim}{x})$ που ικανοποιεί

$$E_{\underset{\sim}{\beta}_0}\left\{M_0(\underset{\sim}{x})X_j X_k X_l\right\} < \infty \ \ \forall\, j,k,l,$$

τότε ισχύει και η συνθήκη (C). Για γενικότερες συναρτήσεις σύνδεσης, παρόμοιες υποθέσεις πρέπει να ικανοποιούνται ώστε να ισχύουν οι συνθήκες (A)-(C). Τα αποτελέσματα των **Θεωρημάτων 1** και **2** μπορούν να προκύψουν και για τις περιπτώσεις των ποινικοποιημένων ελαχίστων τετραγώνων (3.4.2.1.1) και της ποινικοποιημένης εύρωστης γραμμικής παλινδρόμησης (3.4.2.1.2).

3.4.2.3 Ο προτεινόμενος αλγόριθμος

Για την επίλυση του προβλήματος ελαχίστων τετραγώνων της LASSO είχαν προταθεί διάφοροι αλγόριθμοι. Ένας από αυτούς ήταν και του Tibshirani (1996) καθώς και του Fu (1998) ο οποίος πρότεινε έναν "shooting" αλγόριθμο για την μέθοδο *LASSO*. Ωστόσο, στην ενότητα αυτή αναπτύσσεται και παρουσιάζεται ένας νέος αλγόριθμο που προτάθηκε από τους Fan και Li (2001). Σκοπός του αλγόριθμου είναι η επίλυση προβλημάτων ελαχιστοποίησης των (3.4.2.1.1), (3.4.2.1.2) και (3.4.2.1.3) μέσω τοπικών τετραγωνικών προσεγγίσεων (*local quadratic approximations*). Θεωρώντας τους πρώτους όρους των (3.4.2.1.1), (3.4.2.1.2) και (3.4.2.1.3) ως μια συνάρτηση απώλειας $l(\underset{\sim}{\beta})$ (*loss function*) του $\underset{\sim}{\beta}$, λαμβάνουμε την ενιαία μορφή ως

$$l(\underset{\sim}{\beta}) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|) \quad (3.4.2.3.1)$$

Επειδή οι συναρτήσεις ποινής L_1, *SCAD* και *Hard*, είναι ιδιάζουσες στην αρχή και δεν έχουν συνεχείς παραγώγους δεύτερης τάξης, μπορούν να προσεγγισθούν τοπικά από μια τετραγωνική συνάρτηση . Υποθέτουμε ότι έχουμε μια

αρχική τιμή $\underset{\sim}{\beta}_0$ η οποία είναι πολύ κοντά στην τιμή που ελαχιστοποιεί την (3.4.2.3.1) σε περίπτωση που το β_{j0} είναι πολύ κοντά στο 0, θέτουμε $\hat{\beta}_j = 0$. Συνεπώς η x_j διαγράφεται από το τελικό μοντέλο. Αλλιώς, για $\beta_j \neq 0$ χρησιμοποιούμε μια τοπική προσέγγιση της συνάρτησης ποινής $p_\lambda(|\beta_j|)$, βάσει μιας τετραγωνικής συνάρτησης ως,

$$\left[p_\lambda(|\beta_j|)\right]' = p'_\lambda(|\beta_j|)\,\mathrm{sgn}(\beta_j) \approx \left\{p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\right\}\beta_j$$

Με άλλα λόγια, έχουμε ότι

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_{j0}|) + \frac{1}{2}\left\{p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\right\}\left(\beta_j^2 - \beta_{j0}^2\right) \quad \text{(3.4.2.3.2), για } \beta_j \approx \beta_{j0}.$$

Ένα μειονέκτημα της μεθόδου αυτής είναι ότι σε περίπτωση που ένας συντελεστής συρρικνωθεί στο 0, θα παραμείνει σε αυτή τη κατάσταση. Οι διαφορετικές τιμές του β_{j0} φαίνονται στο ***σχήμα 3.1*** της ενότητας 3.4.1 καθώς επίσης και οι συναρτήσεις ποινής L_1, *SCAD* και *Hard*.

Αν τώρα η $l(\underset{\sim}{\beta})$ είναι η L_1 συνάρτηση απώλειας, όπως στην (3.4.2.1.2), τότε δεν έχει συνεχείς μερικές παραγώγους δευτέρας τάξης ως προς $\underset{\sim}{\beta}$. Παρόλα αυτά, η ποσότητα $\psi(|y - \underset{\sim}{x}'\underset{\sim}{\beta}|)$ στην (3.4.2.1.2) μπορεί κατά ανάλογο τρόπο να προσεγγισθεί από την

$$\left\{\psi(y - \underset{\sim}{x}'\underset{\sim}{\beta}_0)/(y - \underset{\sim}{x}'\underset{\sim}{\beta}_0)^2\right\}(y - \underset{\sim}{x}'\underset{\sim}{\beta})^2,$$

αρκεί η αρχική τιμή $\underset{\sim}{\beta}_0$ του $\underset{\sim}{\beta}$ να είναι αρκετά κοντά στην τιμή ελαχιστοποίησης. Επιπλέον, σε περίπτωση που τα υπόλοιπα $|y - \underset{\sim}{x}'\underset{\sim}{\beta}_0|$ είναι μικρά, η προσέγγιση αυτή δεν είναι καλή.

Στη συνέχεια, για την προσέγγιση του πρώτου όρου της (3.4.2.3.1) χρησιμοποιείται μια τετραγωνική συνάρτηση, θεωρώντας ότι ο λογάριθμος της πιθανοφάνειας έχει συνεχείς μερικές παραγώγους δευτέρας τάξης ως προς $\underset{\sim}{\beta}$. Συνεπώς, χρησιμοποιείται ο αλγόριθμος Newton-Raphson μετατρέποντας το πρόβλημα ελαχιστοποίησης της (3.4.2.3.1) σε ένα πρόβλημα ελαχιστοποίησης (*quadratic minimization problem*). Πράγματι, η (3.4.2.3.1) προσεγγίζεται (εκτός από έναν σταθερό όρο) από την ποσότητα

$$l(\underset{\sim}{\beta}_0)+\nabla l(\underset{\sim}{\beta}_0)'(\underset{\sim}{\beta}-\underset{\sim}{\beta}_0)+\frac{1}{2}(\underset{\sim}{\beta}-\underset{\sim}{\beta}_0)'\nabla^2 l(\underset{\sim}{\beta}_0)(\underset{\sim}{\beta}-\underset{\sim}{\beta}_0)+\frac{1}{2}n\underset{\sim}{\beta}'\sum\nolimits_\lambda(\underset{\sim}{\beta}_0)\underset{\sim}{\beta} \quad (3.4.2.3.3),$$

όπου

$$\nabla l(\underset{\sim}{\beta}_0)=\frac{\partial l(\underset{\sim}{\beta}_0)}{\partial \underset{\sim}{\beta}},$$

$$\nabla^2 l(\underset{\sim}{\beta}_0)=\frac{\partial^2 l(\underset{\sim}{\beta}_0)}{\partial \underset{\sim}{\beta}\partial \underset{\sim}{\beta}'},$$

$$\sum\nolimits_\lambda(\underset{\sim}{\beta}_0)=diag\left\{p'_\lambda(|\beta_{10}|)/|\beta_{10}|,...,p'_\lambda(|\beta_{d0}|)/|\beta_{d0}|\right\}.$$

Το τετραγωνικό πρόβλημα ελαχιστοποίησης (3.4.2.3.3), έχει ως λύση την

$$\hat{\underset{\sim}{\beta}}_1=\hat{\underset{\sim}{\beta}}_0-\left\{\nabla^2 l(\underset{\sim}{\beta}_0)+n\sum\nolimits_\lambda(\underset{\sim}{\beta}_0)\right\}^{-1}\left\{\nabla l(\underset{\sim}{\beta}_0)+n\sum\nolimits_\lambda(\underset{\sim}{\beta}_0)\underset{\sim}{\beta}_0\right\} \quad (3.4.2.3.4)$$

Όταν συγκλίνει ο αλγόριθμος, ο εκτιμητής ικανοποιεί τη συνθήκη

$$\frac{\partial l(\hat{\underset{\sim}{\beta}}_0)}{\partial \beta_j}+np'_\lambda(|\hat{\beta}_{j0}|)\,\mathrm{sgn}(\hat{\beta}_{j0})=0,$$

η οποία είναι η εξίσωση ποινικοποιημένης πιθανοφάνειας, για τα μη μηδενικά στοιχεία του $\hat{\underset{\sim}{\beta}}_0$. Συγκεκριμένα, για το πρόβλημα ποινικοποιημένων ελαχίστων τετραγώνων (3.4.2.1.1), η λύση βρίσκεται με επαναληπτικό (*iterative*) υπολογισμό της παλινδρόμησης κορυφογραμμής

$$\underset{\sim}{\beta}_1=\left\{\underset{\sim}{X}'\underset{\sim}{X}+n\sum\nolimits_\lambda(\underset{\sim}{\beta}_0)\right\}^{-1}\underset{\sim}{X}'\underset{\sim}{Y}.$$

Ομοίως, η λύση της (3.4.2.1.2) προκύπτει με επαναληπτικό υπολογισμό της

$$\underset{\sim}{\beta}_1=\left\{\underset{\sim}{X}'\underset{\sim}{W}\underset{\sim}{X}+\frac{1}{2}n\sum\nolimits_\lambda(\underset{\sim}{\beta}_0)\right\}^{-1}\underset{\sim}{X}'\underset{\sim}{W}\underset{\sim}{Y},$$

όπου

$$\underset{\sim}{W}=diag\left\{\psi(|y_1-\underset{\sim}{x}_1'\underset{\sim}{\beta}_0|)/(y_1-\underset{\sim}{x}_1'\underset{\sim}{\beta}_0)^2,...,\psi(|y_n-\underset{\sim}{x}_n'\underset{\sim}{\beta}_0|)/(y_n-\underset{\sim}{x}_n'\underset{\sim}{\beta}_0)^2\right\}.$$

Στο σημείο αυτό αξίζει να αναφερθεί ότι χρησιμοποιώντας τον αλγόριθμο Newton-Raphson (βλ Bickel 1975) όπου λαμβάνουμε τον εκτιμητή ποινικοποιημένης πιθανοφάνειας, η μονοβηματική διαδικασία, έχοντας μια καλή αρχική τιμή $\underset{\sim}{\beta}_0$, μπορεί να είναι εξίσου αποδοτική. Σε περίπτωση τώρα που $\underset{\sim}{\beta}^{(k-1)}$, έχοντας μια καλή αρχική τιμή στο k βήμα, ο επόμενος επαναληπτικός υπολογισμός μπορεί να θεωρηθεί ως μονοκομματική διαδικασία, με αποτέλεσμα ο προκύπτων εκτιμητής εξακολουθεί να μπορεί να είναι το ίδιο αποδοτικός όσο αυτός που θα προέκυπτε με την πλήρως

επαναληπτική μέθοδο (βλ. Robinson, 1988). Συνεπώς, ο εκτιμητής που θα προκύψει με τον αλγόριθμο που αναφέραμε κάνοντας λίγες επαναλήψεις, μπορεί να θεωρηθεί ως εκτιμητής ενός βήματος και θα έχει την ίδια απόδοση. Οπότε δεν είναι απαραίτητη η επανάληψη του αλγορίθμου μέχρι να έχουμε σύγκλιση αρκεί οι αρχικές εκτιμήσεις να είναι καλές. Ως αρχικές εκτιμήσεις τώρα, μπορούν να δοθούν αυτές του πλήρους μοντέλου, αρκεί να μην είναι υπερβολικά παραμετροποιημένες (βλ. Ανδρουλάκης 2008).

3.4.2.4 Υπολογισμός του τυπικού σφάλματος

Ταυτόχρονα με την εκτίμηση παραμέτρων και επιλογή μεταβλητών γίνεται και η εκτίμηση των τυπικών σφαλμάτων. Χρησιμοποιείται ο *sandwich* τύπος, που αφορά και μέτρια μεγέθη δειγμάτων, για την εκτίμηση της συνδιασποράς του $\hat{\underset{\sim}{\beta}}_1$, η μη εξαφανισμένη συνιστώσα του $\hat{\underset{\sim}{\beta}}$. Οπότε έχουμε,

$$\widehat{\mathrm{cov}}(\hat{\underset{\sim}{\beta}}_1) = \left\{\nabla^2 l(\hat{\underset{\sim}{\beta}}_1) + n\sum\nolimits_{\lambda}(\hat{\underset{\sim}{\beta}}_1)\right\}^{-1} \widehat{\mathrm{cov}}\left\{\nabla l(\hat{\underset{\sim}{\beta}}_1)\right\}\left\{\nabla^2 l(\hat{\underset{\sim}{\beta}}_1) + n\sum\nolimits_{\lambda}(\hat{\underset{\sim}{\beta}}_1)\right\}^{-1} \quad (3.4.2.4.1).$$

Όταν χρησιμοποιείται η L_1 συνάρτηση απώλειας στην εύρωστη παλινδρόμηση, πρέπει να πραγματοποιηθούν κάποιες τροποποιήσεις στον αλγόριθμο καθώς επίσης και στον αντίστοιχο *sandwich* τύπο. Στην περίπτωση όπου $\psi(x) = |x|$, τα διαγώνια στοιχεία του $\underset{\sim}{W}$ είναι

$$\{|r_i|^{-1}\}, \text{ με } r_i = y_i - \underset{\sim}{x}_i'\underset{\sim}{\beta}_0 \text{ και } i = 1,...,n.$$

Οπότε για μια δοθείσα τιμή του $\underset{\sim}{\beta}_0$, όταν κάποια από τα υπόλοιπα $\{r_i\}$ είναι κοντά στο 0, αυτά τα σημεία αποκτούν πολύ βάρος. Για αυτό το λόγο αντικαθίσταται το βάρος με

$$(\alpha_n + |r_i|^{-1}).$$

Στις εφαρμογές που έκαναν οι Fan και Li, χρησιμοποίησαν ως α_n το $2n^{-1/2}$ *quantile* των απολύτων τιμών των υπολοίπων, $\{|r_i|\}$. Οπότε το α_n άλλαζε σε κάθε επανάληψη.

3.4.2.5 Έλεγχος σύγκλισης του αλγορίθμου

Οι Fan και Li, απέδειξαν μέσω του προγράμματος MATLAB ότι ο αλγόριθμος που πρότειναν συγκλίνει στη σωστή λύση. Συγκεκριμένα, χρησιμοποίησαν ένα διάνυσμα $\underset{\sim}{\beta}$ διάστασης 100, όπου με τη χρήση της κανονικής κατανομής $N(0,5^2)$ δημιουργήθηκαν 50 μηδενικά και 50 μη μηδενικά στοιχεία. Επιπλέον, χρησιμοποίησαν έναν 100×100 ορθοκανονικό πίνακα σχεδιασμού, για το λόγο ότι τα ποινικοποιημένα ελάχιστα τετράγωνα (PLS) έχουν τότε μαθηματική λύση κλειστής μορφής, οπότε και ήταν εφικτή η σύγκρισή της με αυτήν της αλγοριθμικής μεθόδου τους. Τέλος, με βάσει του γραμμικού μοντέλου $\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$ δημιουργήθηκε το διάνυσμα των αποκρίσεων $\underset{\sim}{Y}$ και πήραν τα εξής αποτελέσματα : ο χρόνος που χρειάστηκε το MATLAB για να συγκλίνουν τα PLS με τη SCAD, L_1 και $Hard$ συνάρτηση ποινής ήταν 0.27, 0.39 και 0.16 sec αντίστοιχα και με αντίστοιχο αριθμό επαναλήψεων 30, 30 και 5. Να σημειωθεί, ότι στη δέκατη επανάληψη, ο PLS εκτιμητής ήταν ήδη αρκετά κοντά στη σωστή τιμή.

3.4.3 Αριθμητικές συγκρίσεις

Στην ενότητα αυτή αναφέρονται οι σχετικές προσομοιώσει που έκαναν οι Fan και Li μέσω των ποινικοποιημένων μεθόδων, όπως επίσης θα γίνει και η σύγκριση της απόδοσης των προτεινόμενων μεθόδων με τις ήδη υπάρχουσες που ως στόχο έχουν τον έλεγχο της ακρίβειας της μεθόδου του τυπικού σφάλματος.

3.4.3.1 Σφάλμα πρόβλεψης και σφάλμα μοντέλου

Με τον όρο σφάλμα πρόβλεψης (*prediction error*) αναφερόμαστε στο μέσο σφάλμα στην πρόβλεψη του Y, δεδομένου νέου $\underset{\sim}{x}$. Ωστόσο, το X μπορεί να είναι είτε τυχαίο (*random*) είτε ελεγχόμενο (*controlled*). Αν είναι τυχαίο τότε και το Y είναι τυχαία επιλεγμένο. Ειδάλλως, οι προγραμματιστές είναι αυτοί που επιλέγουν τον πίνακα σχεδιασμού θεωρώντας το Y τυχαίο.

Εδώ χρησιμοποιείται η πρώτη περίπτωση όπου τα δεδομένα $(\underset{\sim}{x}_i, Y_i)$ θεωρούνται τυχαίο δείγμα από κάποια κατανομή. Τότε το σφάλμα πρόβλεψης, με $\hat{\mu}(\underset{\sim}{x})$ η πρόβλεψη, ορίζεται ως

$$PE(\hat{\mu}) = E\{Y - \hat{\mu}(\underset{\sim}{x})\}^2 .$$

ή

$$PE(\hat{\mu}) = E\{Y - E(Y \mid \underset{\sim}{x})\}^2 + E\{E(Y \mid \underset{\sim}{x}) - \hat{\mu}(\underset{\sim}{x})\}^2 .$$

Ο πρώτος όρος είναι το σφάλμα πρόβλεψης και ο δεύτερος ονομάζεται σφάλμα μοντέλου (*model error*) και συμβολίζεται ως $ME(\hat{\mu})$.

Να σημειώσουμε ότι αν $Y = \underset{\sim}{x}'\underset{\sim}{\beta} + e$, με $E(e \mid \underset{\sim}{x}) = 0$, τότε

$$ME(\hat{\mu}) = (\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta})' E(\underset{\sim}{x}\underset{\sim}{x}')(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta}) .$$

3.4.3.2 Επιλογή των οριακών παραμέτρων

Οι Fan και Li, χρησιμοποίησαν δύο μεθόδους, στη περίπτωση των γραμμικών μοντέλων παλινδρόμησης, για την εκτίμηση της ρυθμιστικής παραμέτρου $\underset{\sim}{\theta}$, όπου $\underset{\sim}{\theta} = (\lambda, \alpha)$ για τη *SCAD* και τη $\underset{\sim}{\theta} = \lambda$ για τη *LASSO* και *Hard* συνάρτηση ποινής:

- Την πενταπλή (*fivefold*) διασταυρωμένη επικύρωση
- Τη γενικευμένη διασταυρωμένη επικύρωση

<u>Πενταπλή διασταυρωμένη επικύρωση</u>:

Θεωρούμε :

- T το σύνολο των δεδομένων
- $T - T^{\nu}$ το σύνολο εκπαίδευσης (*training set*)
- T^{ν} το σύνολο ελέγχου (*test set*) με $\nu = 1,...,5$.

Επομένως, για κάθε $\underset{\sim}{\theta}$ και ν, βρίσκουμε τον εκτιμητή $\hat{\underset{\sim}{\beta}}^{(\nu)}(\underset{\sim}{\theta})$ του $\underset{\sim}{\beta}$, χρησιμοποιώντας το σύνολο εκπαίδευσης $T - T^{\nu}$. Στη συνεχεία, εφαρμόζουμε το κριτήριο της διασταυρωμένης επικύρωσης για να βρούμε το το $\hat{\underset{\sim}{\theta}}$ που ελαχιστοποιεί το $CV(\underset{\sim}{\theta})$.

$$CV(\underset{\sim}{\theta}) = \sum_{\nu=1}^{5} \sum_{(y_k, \underset{\sim}{x}_k) \in T^{\nu}} \left\{ y_k - \underset{\sim}{x}_k' \hat{\underset{\sim}{\beta}}^{(\nu)}(\underset{\sim}{\theta}) \right\}^2$$

Γενικευμένη διασταυρωμένη επικύρωση:

Η λύση μετατρέπεται ως,

$$\underset{\sim}{\beta}_1(\underset{\sim}{\theta}) = \left\{ \underset{\sim}{X}' \underset{\sim}{X} + n\sum\nolimits_{\lambda}(\underset{\sim}{\beta}_0) \right\}^{-1} \underset{\sim}{X}' \underset{\sim}{Y}.$$

Οπότε η προσαρμοσμένη τιμή $\hat{\underset{\sim}{Y}}$ του $\underset{\sim}{Y}$ είναι

$$\underset{\sim}{X} \left\{ \underset{\sim}{X}' \underset{\sim}{X} + n\sum\nolimits_{\lambda}(\underset{\sim}{\beta}_0) \right\}^{-1} \underset{\sim}{X}' \hat{\underset{\sim}{Y}}$$

Με πίνακα προβολής τον

$$\underset{\sim}{P}_{\underset{\sim}{X}}\{\hat{\underset{\sim}{\beta}}(\underset{\sim}{\theta})\} = \underset{\sim}{X} \left\{ \underset{\sim}{X}' \underset{\sim}{X} + n\sum\nolimits_{\lambda}(\hat{\underset{\sim}{\beta}}) \right\}^{-1} \underset{\sim}{X}'.$$

Ορίζοντας τώρα το πλήθος των σημαντικών παραμέτρων στην προσαρμογή του ποινικοποιημένου μοντέλου ελαχίστων τετραγώνων ως

$$e(\underset{\sim}{\theta}) = tr[\underset{\sim}{P}_{\underset{\sim}{X}}\{\hat{\underset{\sim}{\beta}}(\underset{\sim}{\theta})\}],$$

το κριτήριο της γενικευμένης διασταυρωμένης επικύρωσης είναι

$$GCV(\underset{\sim}{\theta}) = \frac{1}{n} \frac{\| \underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\beta}(\underset{\sim}{\theta}) \|^2}{\{1 - e(\underset{\sim}{\theta})/n\}^2}$$

και

$$\hat{\underset{\sim}{\theta}} = \arg\min_{\theta}\{GCV(\underset{\sim}{\theta})\}.$$

3.4.3.3 Προσομοιώσεις

Στη παράγραφο αυτή παρουσιάζονται τα παραδείγματα προσομοιώσεων των Fan και Li μέσω του προγράμματος MATLAB, συγκρίνοντας τις προτεινόμενες μεθόδους επιλογής μεταβλητών με τις μεθόδους ελαχίστων τετραγώνων, παλινδρόμηση κορυφογραμμής, επιλογή καλύτερου υποσυνόλου και τη μέθοδο *Garrote* καθώς επίσης έγινε και η χρήση της γενικευμένης διασταυρωμένης επικύρωσης για την εκτίμηση των οριακών παραμέτρων.

Παράδειγμα 1-(Γραμμική παλινδρόμηση): Σε αυτό το παράδειγμα υπάρχουν 100 σύνολα δεδομένων, αποτελούμενα από n παρατηρήσεις, βάσει του μοντέλου

$$\underset{\sim}{Y} = \underset{\sim}{x}'\underset{\sim}{\beta} + \sigma\underset{\sim}{\varepsilon},$$

όπου τα $\underset{\sim}{x}$ και $\underset{\sim}{\varepsilon}$ είναι της Τυποποιημένης Κανονικής κατανομής και $\underset{\sim}{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)'$. Η συσχέτιση μεταξύ των x_i και x_j είναι $\rho^{|i-j|}$ με $\rho = 0.5$.

Αρχικά, επιλέχθηκε $n = 40$ και $\sigma = 3$. Έπειτα, αυξήθηκαν οι παρατηρήσεις σε n= 60 και μειώθηκε το σ σε σ=1 και έγινε σύγκριση του σφάλματος του μοντέλου με αυτό του εκτιμητή ελαχίστων τετραγώνων. Ωστόσο, στον ***Πίνακας 3.1*** παρουσιάζονται τόσο η διάμεσος των σχετικών σφαλμάτων του μοντέλου (*Median of Relative Model Errors – MRME*) από 100 προσομοιωμένα σύνολα δεδομένων όσο και ο μέσος αριθμός των μηδενικών συντελεστών. Η στήλη «correct» αντιστοιχεί στο μέσο αριθμό των σωστά εκτιμώμενων ως μηδενικοί συντελεστών, ενώ η στήλη «incorrect» αντιστοιχεί σε αυτούς που λανθασμένα εκτιμήθηκαν ως μηδενικοί.

		Avg. No. of 0 Coefficients	
Method	*MRME (%)*	*Correct*	*Incorrect*
$n = 40, \sigma = 3$			
SCAD[1]	72.90	4.20	.21
SCAD[2]	69.03	4.31	.27
LASSO	63.19	3.53	.07
Hard	73.82	4.09	.19
Ridge	83.28	0	0
Best subset	68.26	4.50	.35
Garrote	76.90	2.80	.09
Oracle	33.31	5	0
$n = 40, \sigma = 1$			
SCAD[1]	54.81	4.29	0
SCAD[2]	47.25	4.34	0
LASSO	63.19	3.51	0
Hard	69.72	3.93	0
Ridge	95.21	0	0
Best subset	53.60	4.54	0
Garrote	56.55	3.35	0
Oracle	33.31	5	0
$n = 60, \sigma = 1$			
SCAD[1]	47.54	4.37	0
SCAD[2]	43.79	4.42	0
LASSO	65.22	3.56	0
Hard	71.11	4.02	0
Ridge	97.36	0	0
Best subset	46.11	4.73	0
Garrote	55.90	3.38	0
Oracle	29.82	5	0

Πίνακας 3.1: Αποτελέσματα προσομοιώσεων για το γραμμικό μοντέλο παλινδρόμησης. Για τη SCAD[1] το α επιλέχθηκε βάσει της GCV και για τη SCAD[2] έχει την τιμή 3.7.

Όπως φαίνεται και από τον πίνακα, η *LASSO* έχει την καλύτερη απόδοση όταν ο θόρυβος είναι υψηλός και το μέγεθος του δείγματος μικρό. Επίσης μειώνει σημαντικά τόσο το σφάλμα του μοντέλου όσο και την πολυπλοκότητά του. Το ίδιο ισχύει και για τις υπόλοιπες μεθόδους επιλογής μεταβλητών, σε αντίθεση με τη παλινδρόμηση κορυφογραμμής μειώνει μόνο το σφάλμα του μοντέλου. Η *SCAD* όμως όταν μειώθηκε ο θόρυβος, έγινε αποδοτικότερη από τη *LASSO* και τη *Hard* .

Η παλινδρόμηση κορυφογραμμής έχει κακή απόδοση ενώ η μέθοδος επιλογής καλύτερου υποσυνόλου έχει παρόμοια απόδοση με τη *SCAD*. Επίσης, η *garrote* έχει γενικά καλή απόδοση. Να σημειώσουμε και ότι η *SCAD* είχε πολύ καλά αποτελέσματα με επιλογή του $\alpha = 3.7$ (βλ. αποτελέσματα για $SCAD^1$ και $SCAD^2$), η οποία τιμή χρησιμοποιήθηκε και στις επόμενες προσομοιώσεις. Τελειώνοντας, συμπεραίνουμε ότι αναμένεται η *SCAD* να έχει τόσο καλά αποτελέσματα όσο αυτά του *oracle* εκτιμητή (ο οποίος επίσης χρησιμοποιήθηκε ώστε να συγκριθεί με τις προτεινόμενες μεθόδους), καθώς το μέγεθος του δείγματος αυξάνει.

	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_5$	
Method	*SD*	SD_m (SD_{mad})	*SD*	SD_m (SD_{mad})	*SD*	SD_m (SD_{mad})
$SCAD^1$	.166	.161 (.021)	.170	.160 (.024)	.148	.145 (.022)
$SCAD^2$	.161	.161 (.021)	.164	.161 (.024)	.151	.143 (.023)
LASSO	.164	.154 (.019)	.173	.150 (.022)	.153	.142 (.021)
Hard	.169	.161 (.022)	.174	.162 (.025)	.178	.148 (.021)
Best subset	.163	.155 (.020)	.152	.154 (.026)	.152	.139 (.020)
Oracle	.155	.154 (.020)	.147	.153 (.024)	.146	.137 (.019)

Πίνακας 3.2: Τυπικές αποκλίσεις των εκτιμητών στο γραμμικό μοντέλο παλινδρόμησης (n=60).

Ο παραπάνω ***Πίνακας 3.2*** αναφέρεται στη διάμεσο *SD* των απόλυτων τιμών της απόκλισης των 100 εκτιμώμενων συντελεστών των 100 συνόλων δεδομένων, διαιρεμένη με 0.6745 και μπορεί να θεωρηθεί ως το πραγματικό τυπικό σφάλμα. Ως SD_m συμβολίζεται η διάμεσος των 100 εκτιμώμενων *SDs* ενώ με SD_{mad} η διάμεσος των απολύτων τιμών του σφάλματος της απόκλισης των 100 εκτιμημένων τυπικών σφαλμάτων διαιρεμένη με 0.6745. Επιπλέον, ο παραπάνω πίνακας, όπου το n=60 περιέχει τα αποτελέσματα για τους μη μηδενικούς συντελεστές και μπορούμε να συμπεράνουμε ότι ο *sandwich* τύπος (3.4.2.4.1) είναι αρκετά αποτελεσματικός.

Παράδειγμα 2-(Εύρωστη γραμμική παλινδρόμηση): Σε αυτό το παράδειγμα υπάρχουν 100 σύνολα δεδομένων αποτελούμενα από 60 παρατηρήσεις, βάσει του μοντέλου

$$Y = \underset{\sim}{x}'\underset{\sim}{\beta} + \varepsilon ,$$

Τα $\underset{\sim}{\beta}$ και $\underset{\sim}{x}$ παραμένουν ίδια όπως και στο προηγούμενο παράδειγμα. Το ε είναι της Τυποποιημένης Κανονικής κατανομής με ένα ποσοστό 10% άτυπων σημείων

(*outliers*) της κατανομής *Cauchy*. Στον ***Πίνακα 3.4***. παρουσιάζονται τα αποτελέσματα με τη συνάρτηση ποινής *SCAD να* αποδίδει καλύτερα .

		Avg. No. of 0 Coefficients	
Method	*MRME (%)*	*Correct*	*Incorrect*
SCAD ($a = 3.7$)	35.52	4.71	0
LASSO	52.80	4.29	0
Hard	47.22	4.70	0
Best subset	41.53	4.85	.18
Oracle	23.33	5	0

Πίνακας 3.3: Αποτελέσματα προσομοίωσης για το εύρωστο γραμμικό μοντέλο παλινδρόμησης.

	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_5$	
Method	*SD*	SD_m (SD_{mad})	*SD*	SD_m (SD_{mad})	*SD*	SD_m (SD_{mad})
SCAD	.167	.171 (.018)	.185	.176 (.022)	.165	.155 (.020)
LASSO	.158	.165 (.022)	.159	.167 (.020	.182	.154 (.019)
Hard	.179	.168 (.018)	.176	.176 (.025)	.157	.154 (.020)
Best subset	.198	.172 (.023)	.185	.175 (.024)	.199	.152 (.023)
Oracle	.163	.199 (.040)	.156	.202 (.043)	.166	.177 (.037)

Πίνακας 3.4: Τυπικές αποκλίσεις των εκτιμητών για το εύρωστο γραμμικό μοντέλο παλινδρόμησης.

Παράδειγμα 3-(Λογιστική παλινδρόμηση): Σε αυτό το παράδειγμα υπάρχουν 100 σύνολα δεδομένων αποτελούμενα από 200 παρατηρήσεις, βάσει του μοντέλου $Y \sim Bernoulli\left\{p\left(\underset{\sim}{x}'\underset{\sim}{\beta}\right)\right\}$ με,

$$p(u) = \frac{\exp(u)}{1+\exp(u)},$$

με τις πρώτες 6 συνιστώσες των $\underset{\sim}{\beta}$ και $\underset{\sim}{x}$ να είναι οι ίδιες με αυτές της πρώτης προσομοίωσης. Οι δύο τελευταίες συνιστώσες του $\underset{\sim}{x}$ ήταν i.i.d. από την *Bernoulli* κατανομή με πιθανότητα επιτυχίας 0.5. Επιπλέον, οι μεταβλητές ήταν κανονικοποιημένες και τα σφάλματα του μοντέλου υπολογίσθηκαν μέσω 1000 *Monte Carlo* προσομοιώσεων. Στους ***πίνακες 3.5*** και ***3.6***. παρουσιάζονται τα αποτελέσματα και παρατηρούμε ότι η *SCAD* είχε καλύτερη απόδοση από την *LASSO* και της *Hard*

και συγκριτικά με τον *oracle* εκτιμητή είναι παρόμοια όσον αφορά το *MRME* και την ακρίβεια των εκτιμώμενων τυπικών σφαλμάτων.

		Avg. No. of 0 Coefficients	
Method	MRME (%)	Correct	Incorrect
SCAD ($a = 3.7$)	26.48	4.98	.04
LASSO	53.14	3.76	0
Hard	59.06	4.27	0
Best subset	31.63	4.84	.01
Oracle	25.71	5	0

Πίνακας 3.5: Αποτελέσματα προσομοίωσης για τη λογιστική παλινδρόμηση.

	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_5$	
Method	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})
SCAD ($a = 3.7$)	.571	.538 (.107)	.383	.372 (.061)	.432	.398 (.065)
LASSO	.310	.379 (.037)	.285	.284 (.019)	.244	.287 (.019)
Hard	.675	.561 (.126)	.428	.400 (.062)	.467	.421 (.079)
Best subset	.624	.547 (.121)	.398	.383 (.067)	.468	.412 (.077)
Oracle	.553	.538 (.103)	.374	.373 (.060)	.432	.398 (.064)

Πίνακας 3.6: Τυπικές αποκλίσεις των εκτιμητών για τη λογιστική παλινδρόμηση.

Παρατηρούμε ότι οι εκτιμώμενες τυπικές αποκλίσεις για τον L_1 εκτιμητή ποινικοποιημένης πιθανοφάνειας (*LASSO*) είναι μικρότερες από αυτές της *SCAD*, αλλά με το συνολικό *MRME* μεγαλύτερο. Αυτό σημαίνει ότι η μεροληψία των εκτιμητών της *LASSO* είναι μεγάλη. Κάτι που ισχύει και για όλες τις προαναφερθείσες προσομοιώσεις.

3.4.4 Συμπεράσματα

Οι Fan και Li απέδειξαν ότι οι μέθοδοι που χρησιμοποίησαν, για τη επιλογή σημαντικών παραγόντων, έχουν πού καλή απόδοση. Επίσης αποτελεσματική ήταν και η εκτίμηση των τυπικών σφαλμάτων με τη χρήση του *sandwich τύπου* όπως και ο αλγόριθμος υλοποίησης της όλης μεθόδου βασιζόμενος σε στατιστική θεωρία με συνέπεια τη κατασκευή εκτιμήσεων με καλές στατιστικές ιδιότητες. Σε σύγκριση με

τη μέθοδο επιλογής καλύτερου υποσυνόλου, η οποία είναι αρκετά χρονοβόρα, οι νέες μέθοδοι δίνουν αποτελέσματα αρκετά πιο γρήγορα. Ένα μεγάλο πλεονέκτημά αυτών είναι η ταυτόχρονη επιλογή σημαντικών μεταβλητών και η εκτίμηση των συντελεστών, που πραγματοποιείται βελτιστοποιώντας μια ποινικοποιημένη πιθανοφάνεια. Αποτέλεσμα αυτής της διαδικασίας είναι η ακριβής εκτίμηση των τυπικών σφαλμάτων. Επιπλέον, απέδειξαν ότι η συνάρτηση ποινής *SCAD*, έχει την καλύτερη απόδοση στην επιλογή σημαντικών μεταβλητών, χωρίς να δημιουργείται μεροληψία, εν αντιθέσει με τη *LASSO* μέθοδο του Tibshirani (1996) όπου χρησιμοποιείται η L_1 συνάρτηση ποινής. Η όλη διαδικασία την ποινικοποιημένης πιθανοφάνειας επεκτείνεται και σε άλλα πεδία της Στατιστικής, όπως η Ανάλυση Επιβίωσης.

3.5 Η μέθοδος garrote

Η μη αρνητική μέθοδος garotte (nonnegative garrote) (nn), είναι μια νέα προτεινόμενη μέθοδος που χρησιμοποιείται για την παλινδρόμησης υποσυνόλου. Έχει την ιδιότητα να συρρικνώνει και να μηδενίζει τις συμμεταβλητές. Σε σχέση με τη κοινή μέθοδο επιλογής υποσυνόλου, σε δοκιμές πραγματικών δεδομένων η μέθοδος garrote παράγει μικρότερα προβλεπόμενα σφάλματα. Επίσης, μπορεί να συγκριθεί με τη παλινδρόμηση κορυφογραμμής. Αν οι εξισώσεις παλινδρόμησης δεν αλλάξουν δραστικά με ορισμένες μικρές αλλαγές στα δεδομένα, τότε η διαδικασία μπορεί να θωρηθεί σταθερή. Ωστόσο, η μέθοδος επιλογής υποσυνόλου μπορεί να θεωρηθεί ασταθείς, η παλινδρόμηση κορυφογραμμής σταθερή ενώ η μέθοδος nn-garrote και τα δύο.

Σύμφωνα με τη μέθοδο garotte, ελαχιστοποιείται η ποσότητα

$$\sum_{i=1}^{N}\left(y_i - \alpha - \sum_j c_j \hat{\beta}_j x_{ij} \right)^2 \quad (3.5.1)$$

κάτω από τους περιορισμούς

$$c_j \geq 0 \ \ \forall j \ \text{ και } \sum_j | c_j | \leq t \ \text{ για κάποια } t \,,$$

όπου $\hat{\beta}_j,\ j = 1,...,k$ είναι οι συντελεστές παλινδρόμησης της μεθόδου ελαχίστων τετραγώνων για το πλήρες μοντέλο. Η μέθοδος αυτή άλλους παράγοντες τους

εξαλείφει και άλλους τους συρρικνώνει και σε γενικές γραμμές, όπως είδη αναφέραμε είναι σχετικά σταθερή. Ως εκ τούτου, η μέθοδος έχει περισσότερα μη μηδενικές συμμεταβλητές απ ότι η παλινδρόμηση υποσυνόλου.

Η μέθοδος *garrote* (ή *nonnegative garrote* όπως συναντάται συχνά στη βιβλιογραφία) δημιουργήθηκε υιοθετώντας τον *NNLS* (*nonnegative least squares*) κώδικα των Lawson και Hanson (1974). Πρόκειται για έναν πολύ γνωστό κώδικα γραμμένο σε Fortran όπου χρησιμοποιείται μόνο ο περιορισμός $c_j \geq 0$. Κατασκευάζοντας κάποιες υπορουτίνες, ο Breiman (1995) βελτίωσε τον κώδικα αυτόν, χρησιμοποιώντας μια *barrier* μέθοδο για να εισάγει τον επιπλέον περιορισμό στο άθροισμα των συντελεστών.

Η μέθοδος garotte ξεκινά με τους *OLS* εκτιμητές (εκτιμητές ελαχίστων τετραγώνων), τους οποίους και συρρικνώνει μέσω των (μη-αρνητικών) c_j, των οποίων το άθροισμα είναι φραγμένο. Με τη βοήθεια προσομοιώσεων, ο Breiman έδειξε ότι η *garrote* έχει χαμηλότερο σφάλμα πρόβλεψης από τη μέθοδο επιλογής καλύτερου υποσυνόλου και ότι είναι επάξιος ανταγωνιστής της παλινδρόμησης κορυφογραμμής, εκτός της περίπτωσης όπου το πραγματικό μοντέλο έχει πολλούς και μικρούς σε τιμή αλλά μη μηδενικούς συντελεστές. Παρόλα αυτά, το μειονέκτημα της μεθόδου είναι ότι η λύση που προσφέρει εξαρτάται τόσο από το πρόσημο όσο και το μέγεθος των *OLS* εκτιμητών. Οπότε σε περιπτώσεις όπου οι *OLS* εκτιμητές δεν αποδίδουν καλά, η garotte επίσης επηρεάζεται.

3.6 Μέθοδοι συρρίκνωσης

Διατηρώντας ένα υποσύνολο παραγόντων και απορρίπτοντας το υπόλοιπο, η μέθοδος επιλογής υποσυνόλου παράγει ένα μοντέλο στατιστικά ερμηνεύσιμο που έχει πιθανόν χαμηλότερο προβλεπόμενο σφάλμα από ότι το συνολικό μοντέλο. Ωστόσο, επειδή αυτό αποτελεί μια ξεχωριστή διαδικασία, οι μεταβλητές είτε διατηρούνται είτε απορρίπτονται. Επιπλέον, με την ύπαρξη υψηλής διακύμανσης το προβλεπόμενο σφάλμα δεν μειώνεται. Γι αυτό το λόγο εφαρμόζονται οι παρακάτω μέθοδοι συρρίκνωσης που δεν υποφέρουν από υψηλή μεταβλητότητα.

3.6.1 Παλινδρόμηση κορυφογραμμής (Ridge regression)

Η μέθοδος της ανάλυσης κορυφογραμμής (Ridge regression) προτάθηκε από τους Hoerl και Kennard(1970) με κύριο σκοπό την αντιμετώπιση της πολυσυγγραμικότητας δίνοντας βάση στις ιδιωτικές του πίνακα $X'X$. Ένα σοβαρό πρόβλημα πολυσυγγραμικότητας χαρακτηρίζεται από το γεγονός ότι η μικρότερη ιδιοτιμή του πίνακα $X'X$ είναι κατά πολύ μικρότερη της μονάδας. Επίσης, επισήμαναν ότι τη πολλαπλή παλινδρόμηση οι εκτιμήσεις των παραμέτρων, οι οποίες βασίζονται στη μέθοδο εκτίμησης ελαχίστων τετραγώνων (ε.ε.τ), φαίνεται να μην είναι ικανοποιητικές. Οι ε.ε.τ, δίνουν συντελεστές οι οποίοι κατά απόλυτη τιμή είναι πολύ μεγάλοι και των οποίων τα πρόσημα μπορεί να αλλάζουν όταν συμβαίνουν αμελητέες αλλαγές στα δεδομένα. Ωστόσο, οι ε.ε.τ είναι αμερόληπτες. Σημαντικό όμως είναι να επιτευχθεί μικρό μέσο τετραγωνικό σφάλμα μειώνοντας τη διασπορά. Όμως, μικρή διασπορά συνεπάγεται μικτότερη ακρίβεια πρόβλεψης. Για το λόγο αυτό πρέπει να συρρικνωθούν οι συντελεστές. Επομένως, είναι προτιμότερο να κρατηθούν πληροφορίες από όλες τις μεταβλητές παρά να τις απορρίψουμε τελείως. Ακριβώς αυτή είναι η φιλοσοφία της μεθόδου της κορυφογραμμής.

Προτού παρουσιάσουμε τη περιγραφή του μοντέλου, χρήσιμο θα ήταν να αναφερθούν ορισμένες ιδιότητες για τη καλύτερη γραμμική αμερόληπτη εκτίμηση κορυφογραμμής.

3.6.1.1 Ιδιότητες για τη εκτίμηση του μοντέλου

Η εκτιμήτρια των ελαχίστων τετραγώνων, όπως έχει αναφερθεί και σε προηγούμενη παράγραφο, είναι

$$\hat{\beta} = \left(X'X\right)^{-1} X'Y$$

και με άθροισμα των τετραγώνων των σφαλμάτων

$$SSE = (Y - X\hat{\beta})'(Y - X\hat{\beta})$$

σημειώνεται εδώ ότι οι ιδιότητες του $\hat{\beta}$ είναι γνωστές (βλέπε Scott 1966).Το πρόβλημα που δημιουργείται εδώ είναι ότι ο πίνακας $X'X$ δεν είναι μοναδιαίος. Επομένως παρουσιάζονται δυο ιδιότητες για τη εκτίμηση του $\hat{\beta}$:

i. $Var\left(\hat{\beta}\right)=\sigma^2\left(\mathrm{X}'\mathrm{X}\right)^{-1}$

ii. $L_1 \equiv$ απόσταση του $\hat{\beta}$ από το β

$$L_1^{\,2}=\left(\hat{\beta}-\beta\right)'\left(\hat{\beta}-\beta\right)$$

$$\mathrm{E}\left[L_1^{\,2}\right]=\sigma^2 Trace\left(\mathrm{X}'\mathrm{X}\right)^{-1} dy$$

ή ισοδύναμα

$$E\left[\hat{\beta}'\hat{\beta}\right]=\beta'\beta+\sigma^2 Trace\left(\mathrm{X}'\mathrm{X}\right)^{-1}$$ επίσης, έχουμε

$$Var\left[L_1^{\,2}\right]=2\sigma^4 Trace\left(\mathrm{X}'\mathrm{X}\right)^{-2}$$

Το μέσο τετραγωνικό σφάλμα είναι

$$E\left[L_1^{\,2}\right]=\sigma^2\sum_{i=1}^{p}\left(1/\lambda_i\right)$$ και

$$Var\left[L_1^{\,2}\right]=2\sigma^4\sum_{i=1}^{p}\left(1/\lambda_i\right)^2$$

με τις ιδιοτιμές $\lambda_{\max}=\lambda_1\geq\lambda_2\geq....\geq\lambda_p=\lambda_{\min}>0$ του πίνακα $\mathrm{X}'\mathrm{X}$.

3.6.1.2 Περιγραφή της μεθόδου

Η εκτιμήτρια της μεθόδου δίνεται από τον τύπο

$$\hat{\beta}_{ridge}=\left[\mathrm{X}'\mathrm{X}+kI\right]^{-1}\mathrm{X}'Y=W\mathrm{X}'Y;k\geq 0$$

Όπου k είναι η ρυθμιστική παράμετρος και μια γενική εκτίμηση δίνεται από τον τύπο

$$\hat{\beta}^*=\left[I_p+k\left(\mathrm{X}'\mathrm{X}\right)^{-1}\right]^{-1}\hat{\beta}=\mathrm{Z}\hat{\beta}$$

Επίσης, για την εκτίμηση του $\hat{\beta}^*$ το άθροισμα των τετραγώνων των σφαλμάτων είναι

$$SSE=(Y-X\hat{\beta}^*)'(Y-X\hat{\beta}^*)$$

Το οποίο μπορεί να γραφεί και στη μορφή

$$SSE=Y'Y-\left(\hat{\beta}^*\right)'X'Y-k\left(\hat{\beta}^*\right)'\left(\hat{\beta}^*\right)$$

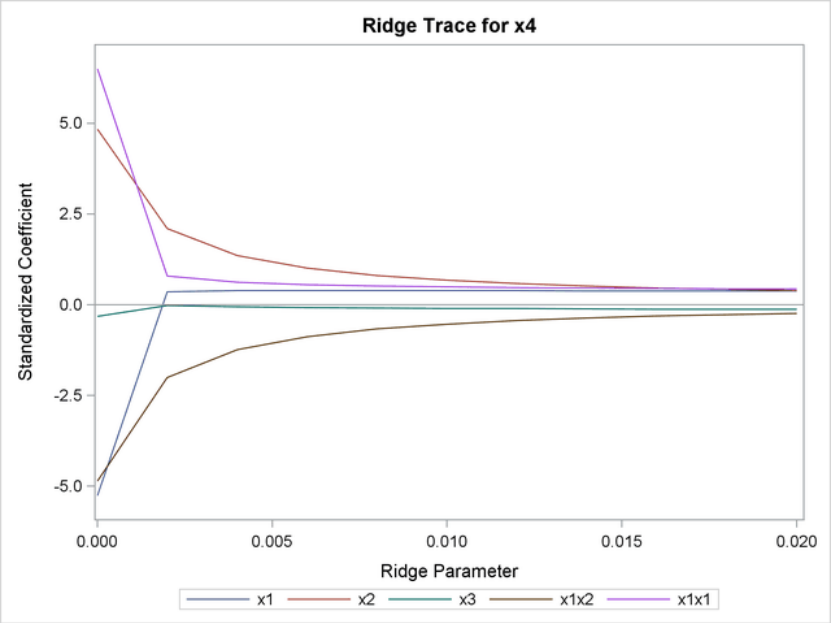
Και το μέσο τετραγωνικό σφάλμα δίνεται από τη σχέση

$$Var\left[\hat{\beta}^*\right] = \sigma^2 (X'X + kI)^{-1} X'X (X'X + kI)^{-1}$$

Στο εξής, σημειώνεται εδώ ότι το kI δικαιολογεί την έννοια της ποινής που επιβάλει η μέθοδος αυτή στην εκτίμηση των συντελεστών. Υπάρχει μια βέλτιστη τιμή του k, για κάθε πρόβλημα. Αναζητείται η τιμή εκείνη που έχει αντίστοιχη εκτιμήτρια έχει το μικρότερο μέσο τετραγωνικό σφάλμα. Εφόσον η εκτιμήτρια $\hat{\beta}$ είναι αμερόληπτη, η $\hat{\beta}^*$ είναι μεροληπτική και η διασπορά $Var\left[\hat{\beta}^*\right]$ είναι μικρότερη της $Var\left(\hat{\beta}\right)$ συνεπώς, το μέσο τετραγωνικό σφάλμα της $\hat{\beta}^*$ είναι μικρότερο από της $\hat{\beta}$.

3.6.1.3 Ίχνος κορυφογραμμής

Ένα μεγάλο πλεονέκτημα της μεθόδου είναι το γράφημα ίχνους κορυφογραμμής το οποίο μπορεί να βοηθήσει τον ερευνητή να διαπιστώσει ποιοι συντελεστές είναι ευαίσθητοι στα δεδομένα. Το ίχνος της κορυφογραμμής είναι το γράφημα της τιμής κάθε συντελεστή έναντι του k. Το γράφημα έχει μια καμπύλη (ίχνος) για κάθε συντελεστή. Για καλύτερα αποτελέσματα χρησιμοποιούνται λιγότεροι από 10 συντελεστές. Ο στόχος είναι να βρούμε μια τιμή του k που να δίνει ένα σύνολο συντελεστών με μικρότερο τετραγωνικό σφάλμα. Καθώς το λ αυξάνει το SSE αυξάνεται. Αν οι μεταβλητές πρόβλεψης εμφανίζουν μεγάλες συσχετίσεις, οι συντελεστές θα αλλάζουν γρήγορα για μικρές τιμές του k και σταδιακά θα σταθεροποιούνται. Ενώ αν είναι ορθογώνιες, τότε οι συντελεστές θα αλλάζουν ελάχιστα. Το παρακάτω σχήμα παρουσιάζει ένα παράδειγμα για τι ίχνος κορυφογραμμής σε ένα μοντέλο με 5 μεταβλητές οι x1, x2, x3, x1x2, x1x1.



Σχήμα 3.6: Ridge Trace

Καθώς η τιμή της παραμέτρου k αυξάνεται παρατηρούμε ότι οι συντελεστές μεταβάλλονται και όταν η παράμετρος παίρνει τη τιμή k=0.015, σύμφωνα με το παραπάνω γράφημα, οι συντελεστές σταθεροποιούνται.

3.6.2 Η μέθοδος Lasso

Οι εκτιμητές ελαχίστων τετραγώνων (OLS) λαμβάνονται από την ελαχιστοποίηση των τετραγωνικών σφαλμάτων. Υπάρχουν δυο λόγοι για τους οποίους η ανάλυση των δεδομένων δεν ικανοποιούνται από τους εκτιμητές OLS. Ο πρώτος λόγος αφορά την ακρίβεια πρόβλεψης, όπου οι εκτιμητές OLS έχουν χαμηλή μεροληψία και μεγάλη διακύμανση. Ωστόσο η ακρίβεια πρόβλεψης μπορεί να βελτιωθεί συρρικνώνοντας η μηδενίζοντας τους συντελεστές, με αποτέλεσμα να μειώνεται η διακύμανση των προβλεπόμενων τιμών. Ο δεύτερος λόγος αφορά την ερμηνεία. Συχνά απαιτείται η επιλογή ενός μικρότερου αριθμού παραγόντων που εμφανίζει ισχυρότερες επιδράσεις. Ως εκ τούτου, οι δύο βασικές μέθοδοι (BS- ridge regression) επιφέρουν ορισμένα μειονεκτήματα. Πιο συγκεκριμένα, παρόλο που η παλινδρόμηση κορυφογραμμής είναι πιο σταθερή σαν μέθοδος αφού συρρικνώνει συνεχώς τους συντελεστές, δεν μηδενίζεται κανένας από αυτούς με αποτέλεσμα το μοντέλα να μην είναι εύκολα ερμηνεύσιμο. Από την άλλη, η BS μέθοδος δίνει

μοντέλα με λιγότερες μεταβλητές και συνεπώς πιο ερμηνεύσιμα αλλά είναι εξαιρετικά ευμετάβλητη διαδικασία.

Για την αντιμετώπιση των παραπάνω προβλημάτων, προτάθηκε μια νέα μέθοδος για τη εκτίμηση των γραμμικών μοντέλων από τον Tibshirani (1996) σύμφωνα με την οποία ελαχιστοποιείται το άθροισμα των τετραγώνων των υπολοίπων, υπό τον περιορισμό το άθροισμα των απολύτων τιμών των συντελεστών να είναι μικρότερο από μια σταθερά. Αυτό συμβαίνει εξαιτίας της φύσης αυτού του περιορισμού που τείνει να παράγει ορισμένους συντελεστές που είναι μηδενικοί και έτσι δίνει ερμηνεύσιμα μοντέλα. Σύμφωνα με διάφορες προσομοιώσεις, η μέθοδος LASSO χρησιμοποιεί κάποιες από τις ευνοϊκές ιδιότητες τόσο της μεθόδου καλύτερες επιλογής υποσυνόλου (BS) όσο και τις παλινδρόμησης κορυφογραμμής. Πιο συγκεκριμένα, παράγει ερμηνεύσιμα μοντέλα όπως η BS και παρουσιάζει σταθερότητα όπως η παλινδρόμηση κορυφογραμμής. Επίσης, υπάρχει μια ενδιαφέρουσα σχέση στη προσαρμοστική εκτίμηση της συνάρτησης που παρουσιάζεται από τους Donoho και Johnstone (1994a). Η ιδέα της Lasso είναι αρκετά γενική και μπορεί να εφαρμοστεί σε μια μεγάλη ποικιλία στατιστικών μοντέλων όπως στα γενικευμένα μοντέλα παλινδρόμησης.

Έστω ότι έχουμε τα δεδομένα $(\underset{\sim}{x}^i, y_i),\ i=1,2,...,N$ όπου $\underset{\sim}{x}^i=(x_{i1},...,x_{ip})'$ είναι οι ανεξάρτητες μεταβλητές και y_i οι αποκρίσεις. Υποθέτουμε, είτε ότι οι παρατηρήσεις είναι ανεξάρτητες είτε ότι τα y_i είναι υπό συνθήκη ανεξάρτητα, δοθέντων των x_{ij}. Επιπλέον υπόθεση αποτελεί και ότι τα x_{ij} είναι κανονικοποιημένα, έτσι ώστε,

$$\frac{\sum_i x_{ij}}{N}=0 \text{ και } \frac{\sum_i x_{ij}^2}{N}=1.$$

Αν $\underset{\sim}{\hat{\beta}}=(\hat{\beta}_1,...,\hat{\beta}_p)$, τότε οι εκτιμητές *LASSO* ορίζονται ως

$$(\hat{\alpha},\underset{\sim}{\hat{\beta}})=\arg\min\left\{\sum_{i=1}^{N}\left(y_i-\alpha-\sum_j\beta_j x_{ij}\right)^2\right\} \quad (3.6.2.1),$$

υπό τον περιορισμό ότι

$$\sum_j|\beta_j|\le t,$$

όπου η $t \geq 0$ ονομάζεται ρυθμιστική παράμετρος (*tuning parameter*). Επειδή για όλα τα t ισχύει ότι

$$\hat{\alpha} = \bar{y}$$

και υποθέτοντας χωρίς βλάβη της γενικότητας ότι

$$\bar{y} = 0,$$

το **α** μπορεί να παραληφθεί. Ο υπολογισμός της λύσης της (3.6.2.1), αποτελεί πρόβλημα τετραγωνικού προγραμματισμού με γραμμικούς ανισωτικούς περιορισμούς.

Η παράμετρος t ελέγχει το μέγεθος της συρρίκνωσης (*shrinkage*) που επιβάλλεται στους συντελεστές του μοντέλου. Αν $\hat{\beta}_j^0$ είναι οι εκτιμητές ελαχίστων τετραγώνων και $t_0 = \sum_j |\hat{\beta}_j^0|$, τότε η συνθήκη $t < t_0$ θα προκαλέσει συρρίκνωση των λύσεων προς το μηδέν και ορισμένοι από τους συντελεστές μπορεί να γίνουν ακριβώς μηδέν. Για παράδειγμα, αν $t = \frac{t_0}{2}$, τότε το αποτέλεσμα θα είναι παρόμοιο με το να βρούμε το καλύτερο υποσύνολο με $\frac{p}{2}$ μεταβλητές. Να επισημάνουμε επίσης ότι ο πίνακας σχεδιασμού δεν χρειάζεται να είναι πλήρους βαθμού.

Κίνητρο για την εφαρμογή της μεθόδου Lasso αποτέλεσε η ενδιαφέρουσα πρόταση του Breiman (1993) για τη μη αρνητική μέθοδο garotte που όπως έχουμε είδη αναφέρει ελαχιστοποιεί τη ποσότητα (3.5.1).

Ένα αρνητικό της μεθόδου garotte, την οποία αναλύσαμε στη προηγούμενη παράγραφο, είναι ότι κάθε λύση της μεθόδου εξαρτάται τόσο από το σημείο όσο και από το μέγεθος των OLS εκτιμητών. Σε υψηλές εφαρμογές όπου οι εκτιμητές OLS συμπεριφέρονται άσχημα και δεν αποδίδουν καλά, η μέθοδος garotte δεν επιφέρει σωστά αποτελέσματα καθώς επηρεάζεται από αυτή τη συμπεριφορά, σε αντίθεση με τη Lasso που αψηφά τη ρητή χρήση των OLS εκτιμητών.

3.6.3 Σύγκριση μεθόδων : Lasso, BS και παλινδρόμηση κορυφογραμμής

Σε αυτή τη παράγραφο, προκειμένου να αποκτήσουμε μι πρακτική αντίληψη των μεθόδων που έχομε αναφέρει λεπτομερώς σε προηγούμενες παραγράφους, θα δούμε τη μορφή που αυτοί αποκτούν, όπως αναφέρεται και στον Tibshirani (1996,

2008), στη περίπτωση που ο πίνακας σχεδιασμού είναι ορθοκανονικός δηλαδή $\underset{\sim}{X}'\underset{\sim}{X} = \underset{\sim}{I}$ όπου ο I είναι ο ταυτοτικός πίνακας. (βλέπε ***Πίνακας 3.7***). Στη περίπτωση αυτή έχουμε :

i. Η επιλογή μεταβλητών BS μεγέθους k επιλέγει τις μεταβλητές οι οποίες αντιστοιχούν στους k μεγαλύτερους συντελεστές κατά απόλυτη τιμή και θέτει τους υπόλοιπους ίσους με το μηδέν. Για κάποια επιλογή του λ αυτό είναι ισοδύναμη με το να θέσουμε

$$\hat{\beta}_j = \begin{cases} \hat{\beta}_j^0, & |\hat{\beta}_j^0| > \lambda \\ 0, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases}$$

ii. Στη μέθοδο της ανάλυσης κορυφογραμμής οι εκτιμήτριες των συντελεστών (με το γ να εξαρτάται από το λ ή το t) είναι :

$$\hat{\beta}_j = \frac{1}{1+\gamma}\hat{\beta}_j^0$$

iii. Στη garotte οι εκτιμητές είναι :

$$\hat{\beta}_j = \left(1 - \frac{\gamma}{\left(\hat{\beta}_j^0\right)^2}\right)^+ \hat{\beta}_j^0$$

iv. Στη Lasso οι λύσεις στη περίπτωση του ορθοκανονικού πίνακα σχεδιασμού είναι :

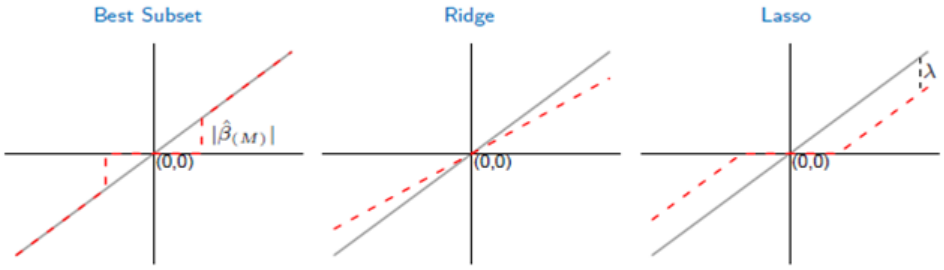
$$\hat{\beta}_j = \mathrm{sgn}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)^+$$

με $(x)^+ = \begin{cases} x, & x > 0 \\ 0, & x \le 0 \end{cases}$

και το γ καθορίζεται από τη συνθήκη $\sum_j |\hat{\beta}_j| = t$.

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$

Πίνακας 3.7: εκτιμήσεις των βj στη περίπτωση του ορθοκανονικού πίνακα σχεδιασμού. Οι μεταβλητές M και λ έχουν επιλεγεί για τις αντίστοιχες τεχνικές. Όπου το sgn είναι η συνάρτηση πρόσημο, και το x+ δείχνει το 'θετικό μέρος' του x.



Σχήμα 3.7: Κάτω από τον πίνακα παρουσιάζονται οι εκτιμητές με διακεκομμένη γραμμή. Η μη διακεκομμένη γραμμή με γωνία 45^o δείχνει τους χωρίς περιορισμούς εκτίμηση.

Για να γίνει πιο κατανοητή η σχέση μεταξύ της μεθόδου Lasso και παλινδρόμησης κορυφογραμμής, το ***σχήμα 3.8*** δείχνει τις εκτιμήσεις των δυο μεθόδων. Το άθροισμα των τετραγώνων των υπολοίπων έχει ελλειπτικά περιγράμματα με επίκεντρο τις εκτιμήσεις των ελαχίστων τετραγώνων (OLS).

Στην περίπτωση όπου $p = 2$. Ας θεωρήσουμε χωρίς βλάβη της γενικότητας ότι οι $\hat{\beta}_j^0$ είναι και οι δύο θετικοί. Είναι εύκολο να δειχθεί ότι σε αυτή την περίπτωση οι εκτιμητές *LASSO* είναι

$$\hat{\beta}_j = (\hat{\beta}_j^0 - \gamma)^+ \quad (3.6.3.1),$$

όπου το γ επιλέγεται κατά τρόπον ώστε $\hat{\beta}_1 + \hat{\beta}_2 = t$.

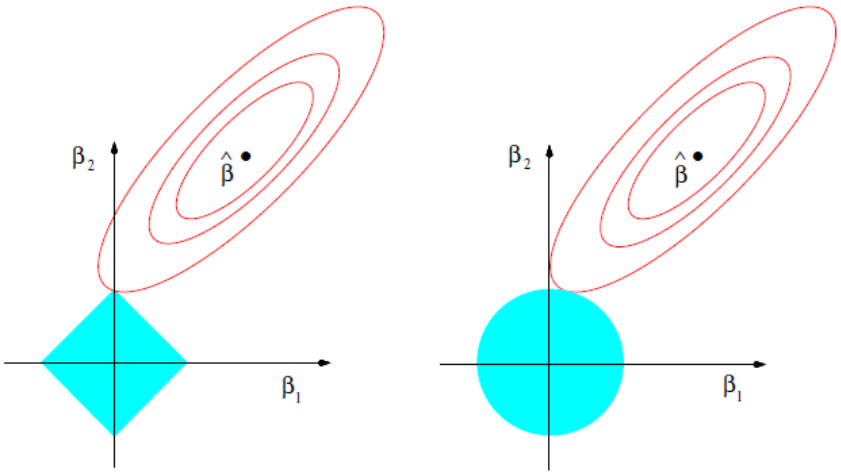
Αυτό ουσιαστικά σημαίνει ότι

$$\gamma = \frac{\hat{\beta}_1^0 + \hat{\beta}_2^0 - t}{2} \quad (3.6.3.2),$$

Η (3.6.3.1),ισχύει για $t \le \hat{\beta}_1^0 + \hat{\beta}_2^0$ και είναι έγκυρη ακόμα και στην περίπτωση ύπαρξης συσχέτισης στις μεταβλητές. Αντικαθιστώντας την (3.6.3.2), στην (3.6.3.1), προκύπτει ότι

$$\hat{\beta}_1 = \left(\frac{t}{2} + \frac{\hat{\beta}_1^0 - \hat{\beta}_2^0}{2} \right)^+ \text{ και } \hat{\beta}_2 = \left(\frac{t}{2} - \frac{\hat{\beta}_1^0 - \hat{\beta}_2^0}{2} \right)^+ .$$

Η λύση που δίνουν και οι δύο μέθοδοι, είναι το πρώτο σημείο όπου τα περιγράμματα ακουμπούν το ρόμβο για τη Lasso και τον κύκλο για τη παλινδρόμηση κορυφογραμμής. Αυτό κάποιες φορές συμβαίνει σε κάποια γωνία, οπότε έχουμε και την περίπτωση μηδενικού συντελεστή. Αν παρόλα αυτά το $p > 2$ και υπάρχει έστω και μια μέτρια συσχέτιση στα δεδομένα, τότε οι εκτιμήσεις των παραμέτρων είναι πιθανότερο να μηδενιστούν.



Σχήμα 3.8 : Αριστερά η γεωμετρία της LASSO για p=2,δεξιά η γεωμετρία της παλινδρόμησης κορυφογραμμής για p=2.

Συμπερασματικά, σε προσομοιώσεις που πραγματοποιήθηκαν από τον Tibshirani (1996), εξετάστηκε συγκριτικά η ικανότητα πρόβλεψης των τεσσάρων μεθόδων κάτω από τρία διαφορετικά σενάρια ως προς τον αριθμό των μεταβλητών και την επίδραση των μεθόδων στην εξαρτημένη μεταβλητή.

i. ***Μικρός αριθμός μεταβλητών-μικρή επίδραση***. Σε αυτή τη περίπτωση η BS μέθοδος είναι καλύτερη της Lasso, ενώ η παλινδρόμηση κορυφογραμμής έδειξε τα χειρότερα αποτελέσματα.

ii. ***Μέτριος αριθμός μεταβλητών-μέτρια επίδραση.*** Η Lasso έδωσε τα καλύτερα αποτελέσματα. Ακολουθεί η μέθοδος της κορυφογραμμής και τέλος η BS.

iii. ***Μεγάλος αριθμός μεταβλητών-καμία επίδραση.*** Στη περίπτωση αυτή η παλινδρόμηση κορυφογραμμής έδωσε τα καλύτερα αποτελέσματα με διαφορά και ακολουθούν η Lasso και τελευταία η BS.

Τέλος, η garotte στη περίπτωση (i) ήταν καλύτερη από τη Lasso και λίγο χειρότερη από τη τις άλλες.

ΚΕΦΑΛΑΙΟ 4

ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΣΤΗΝ ΑΝΑΛΥΣΗ ΕΠΙΒΙΩΣΗΣ

4.1 Εισαγωγή

Στο κλάδο της στατιστικής επιστήμης, η ανάλυση επιβίωσης επικεντρώνεται στην ανάλυση δεδομένων. Τα δεδομένα αυτά αφορούν τη χρονική στιγμή εμφάνισης ενός συμβάντος και έχουν εφαρμογές σε πολλές επιστημονικές περιοχές όπως στη φαρμακευτική, μηχανική, βιολογία, επιδημιολογία, στην οικονομία, στη δημόσια υγεία κ.α. (Καρώνη 2005, Runze Li, 2000). Επιπλέον, τα δεδομένα αυτά αφορούν το λεγόμενο χρόνο επιβίωσης ή χρόνο αποτυχίας (*survival time or failure time*). Ένας κοινός χρόνος αποτυχίας είναι εκείνος που περιλαμβάνει είτε αποκομμένα δεδομένα (censored data) είτε κολοβές παρατηρήσεις (truncated observations). Τα αποκομμένα δεδομένα προκύπτουν στις περιπτώσεις όπου οι χρόνοι επιβίωσης δεν είναι γνωστοί. Πιθανά αποκομμένα συμβάντα είναι η δεξιά αποκοπή (right censoring) και η αριστερή αποκοπή (left censoring). Δεξιά αποκοπή, έχουμε όταν ο χρόνος επιβίωσης είναι μεγαλύτερος από τον χρόνο λήξης της μελέτης ή γενικότερα μεγαλύτερος από κάποιο χρονικό όριο (π.χ. τη στιγμή που για κάποιο λόγο χάθηκε η επαφή μαζί του). Αυτό σημαίνει ότι προφανώς δεν είναι γνωστός, αλλά τουλάχιστον ίσος με τη διάρκεια παραμονής του ατόμου στη μελέτη. Η αριστερά αποκοπή, συμβαίνει όταν ο πραγματικός χρόνος επιβίωσης είναι μικρότερος από τον παρατηρούμενο. Επίσης η αποκοπή διαστήματος (interval censoring), παρατηρείται όταν ξέρουμε πως το υπό μελέτη συμβάν έχει πραγματοποιηθεί σε ένα διάστημα και πάλι όμως χωρίς να είναι γνωστό το ακριβές σημείο. Αυτό παρατηρείται συνήθως όταν έχουμε περιοδική παρακολούθηση του ασθενή. Τέλος, στις αριστερές κολοβές παρατηρήσεις συμπεριλαμβάνονται μόνο τα άτομα που έχουν επιβιώσει σε ένα επαρκές χρονικό διάστημα ενώ στις δεξιά κολοβές παρατηρήσεις συμπεριλαμβάνονται μόνο εκείνα τα άτομα που έχουν βιώσει το γεγονός σε ένα συγκεκριμένο χρονικό διάστημα.

Στις περισσότερες μελέτες , οι ερευνητές συγκρίνουν τα ποσοστά κινδύνου μεταξύ διαφορετικών ομάδων σε σχέση με την ηλικία ή το φύλλο. Είναι χρήσιμη λοιπόν η μοντελοποίηση της συνάρτησης κινδύνου χρησιμοποιώντας μοντέλα αναλογικού κινδύνου (βλ.Klein και Moeschberger, 1997). Συχνά θεωρείται ότι η συνάρτηση κινδύνου εξαρτάται από συμμεταβλητές μέσω ενός γραμμικού συνδυασμού συμμεταβλητών, έστω $x'\beta$, όπου x είναι διάνυσμα. Σε πολλές περιπτώσεις, υπάρχει ένας μεγάλος αριθμός συμμεταβλητών που μπορεί να είναι πιθανοί παράγοντες πρόβλεψης. Ας σκεφτούμε τη περίπτωση όπου όλες οι επεξηγηματικές μεταβλητές είναι ισότιμες, και στόχος μας είναι να επιλέξουμε ένα υποσύνολο σημαντικών μεταβλητών από τις οποίες εξαρτάται η συνάρτηση κινδύνου. Σε αυτό το κεφάλαιο γίνεται η καλύτερη επιλογή υποσυνόλου. Κίνητρα για την καλύτερη επιλογή μεταβλητής προτάθηκαν από τον Lindley (1968), Faraggi και Simon (1998). Ο Tibshirani(1997) εφάρμοσε τη μέθοδο LASSO για το μοντέλο αναλογικού κινδύνου του Cox. Παρόλο αυτά οι μέθοδοι ποινικοποιημένης πιθανοφάνειας , που αναλύεται στο κεφάλαιο 3, μπορεί απευθείας να εφαρμοστεί σε παραμετρικά μοντέλα της ανάλυσης επιβίωσης. Ο σκοπός αυτού του κεφαλαίου είναι να επεκτείνουμε την ιδέα επιλογής μεταβλητών μέσω της μεθόδου ποινικοποιημένης πιθανοφάνειας για τα ημι- παραμετρικά μοντέλα. Ενδιαφερόμαστε ιδιαιτέρως για τα μοντέλα ευπάθειας ένα ημι-παραμετρικό μοντέλο επέκτασης του μοντέλου αναλογικού κινδύνου του Cox.

Στη παράγραφο 4.2.1 θα μελετηθεί εκτενώς η επιλογή μεταβλητών για τα μοντέλα ευπάθειας, επέκτασης του αναλογικού κινδύνου του Cox. Υπάρχουν πολλά άρθρα στη στατιστική βιβλιογραφία που αφορούν προβλήματα εκτίμησης των συντελεστών παλινδρόμησης για τα μοντέλα ευπάθειας (βλ. Lee και άλλους 1992) με τη χρήση της μεθόδου ψευδό-πιθανοφάνειας, όταν ο αριθμός των ομάδων είναι μικρός. Ο Neilsen και άλλοι (1992) εφάρμοσαν τον ΕΜ αλγόριθμο για το μοντέλο γάμμα ευπάθειας (βλέπε παράγραφο 4.2.1). Επιπλέον ο Sinha (1998) πρότεινε μια μεταγενέστερη μέθοδο πιθανοφάνειας για την εκτίμηση των συντελεστών παλινδρόμησης στα ημι-παραμετρικά μοντέλα ευπάθειας. Ως εκ τούτου, μια νέα μέθοδος προτείνεται για τα μοντέλα ευπάθειας, υποκινούμενη από την ιδέα της επανατοποθέτησης στο πλαίσιο της μη παραμετρικής παλινδρόμησης. Η προτεινόμενη μέθοδος μπορεί να χρησιμοποιηθεί έτσι ώστε να αντληθεί μια ενιαία μέθοδος επιλογής μεταβλητής μέσω μη κοίλης ποινικοποιημένης πιθανοφάνειας για

τα ημι-παραμετρικά μοντέλα. Στη παράγραφο 4.3 γίνεται αναφορά σε προσομοιώσεις του μοντέλου για τη προσαρμογή της προτεινόμενης μεθόδου καθώς επίσης και συγκρίσεις μεταξύ των μεθόδων επιλογής μεταβλητής και της καλύτερης επιλογής υποσυνόλου.

4.2 Μοντέλα αναλογικού κινδύνου

Έστω T χρόνος επιβίωσης, C ο αποκομμένος χρόνος και x το διάνυσμα συμμεταβλητών. Ας υποθέσουμε Z=min{T, C} ο παρατηρούμενος χρόνος και $\delta = I(T \le C)$ ο δείκτης λογοκρισίας όπου T και C είναι ανεξάρτητα για δοσμένο $\underset{\sim}{x}$. Όταν τα παρατηρούμενα δεδομένα $\{\underset{\sim}{x}_i, Z_i, \delta_i : i = 1,...,n\}$ είναι ένα τυχαίο δείγμα ανεξάρτητων και πανομοιότυπων μεταβλητών ενός συγκεκριμένου πληθυσμού (x, Z, δ), η συνάρτηση πιθανοφάνειας δίνεται από τον τύπο

$$L = \prod_u f(Z_i \mid \underset{\sim}{x}_i) \prod_c \bar{F}(Z_i \mid \underset{\sim}{x}_i) = \prod_u h(Z_i \mid \underset{\sim}{x}_i) \prod_{i=1}^{n} \bar{F}(Z_i \mid \underset{\sim}{x}_i) \quad (4.2.1)$$

όπου οι δείκτες c και u δηλώνουν αποκομμένα και μη αποκομμένα δεδομένα αντίστοιχα και $f(t \mid \underset{\sim}{x})$, $\bar{F}(t \mid \underset{\sim}{x})$ και $h(t \mid \underset{\sim}{x})$ είναι οι δεσμευμένες συναρτήσεις πυκνότητας πιθανότητας, επιβίωσης και διακινδύνευσης αντίστοιχα, του χρόνου T, για δοσμένο $\underset{\sim}{x}$.

Θεωρούμε τώρα ότι $t_1^0 < ... < t_N^0$ είναι οι διατεταγμένοι χρόνοι αποτυχίας. Επίσης, έστω ότι το j συμβολίζει τον χρόνο αποτυχίας του ατόμου στο χρόνο t_j^0, ώστε οι συμμεταβλητές που σχετίζονται με τις N το πλήθος αποτυχίες να είναι οι $\underset{\sim}{x}_{(1)}, ..., \underset{\sim}{x}_{(N)}$. Συνεχίζουμε συμβολίζοντας με R_j το σύνολο των ατόμων σε κίνδυνο στο χρόνο t_j^0, ώστε

$$R_j = \{i : Z_i \ge t_j^0\}.$$

Θεωρώντας τώρα το μοντέλο αναλογικής διακινδύνευσης

$$h(t \mid \underset{\sim}{x}) = h_0(t) \exp(\underset{\sim}{x}' \underset{\sim}{\beta}),$$

η πιθανοφάνεια (4.2.1) γίνεται

$$L = \prod_{i=1}^{N} h_0(Z_{(i)}) \exp(\underset{\sim}{x}_{(i)}' \underset{\sim}{\beta}) \prod_{i=1}^{n} \exp\left\{-H_0(Z_i)\exp(\underset{\sim}{x}_i' \underset{\sim}{\beta})\right\},$$

όπου $H_0(\cdot)$ η αναφορική σωρευτική συνάρτηση διακινδύνευσης. Στη περίπτωση που η αναφορική συνάρτηση διακινδύνευσης έχει μια παραμετρική μορφή, έστω $h_0(\underset{\sim}{\theta}, \cdot)$, τότε ο λογάριθμος της αντίστοιχης ποινικοποιημένης συνάρτησης πιθανοφάνειας είναι

$$\sum_{i=1}^{N}\left[\log\left\{h_0(\underset{\sim}{\theta}, Z_{(i)})\right\} + \underset{\sim}{x}_{(i)}' \underset{\sim}{\beta}\right] - \sum_{i=1}^{n}\left\{H_0(\underset{\sim}{\theta}, Z_i)\exp(\underset{\sim}{x}_i' \underset{\sim}{\beta})\right\} - n\sum_{j=1}^{d} p_\lambda(|\beta_j|) \quad (4.2.2)$$

Μεγιστοποιώντας την (4.2.2) ως προς $(\underset{\sim}{\theta}, \underset{\sim}{\beta})$, παίρνουμε τον εκτιμητή μέγιστης πιθανοφάνειας.

4.2.1 Οι μέθοδοι ποινικοποιημένης πιθανοφάνειας για το μοντέλο ευπάθειας

Για το μοντέλο αναλογικού κινδύνου του Cox οι χρόνοι επιβίωσης των ατόμων θεωρούνται ανεξάρτητοι. Ωστόσο, αυτή η υπόθεση θα μπορούσε να παραβιαστεί κάτω από ορισμένες περιπτώσεις, στις οποίες τα δεδομένα μου συλλέγονται είναι συσχετισμένα. Για παράδειγμα, άτομα που προέρχονται από την ίδια οικογένεια μοιράζονται κοινό γενετικό ιστορικό. Σε αυτή τη περίπτωση δεν είναι λογικό να θεωρούμε ότι αυτά τα δεδομένα είναι ανεξάρτητα μεταξύ τους (βλ Lee, E.W., Wei, L., και Amato, D.A. 1992). Μια δημοφιλής προσέγγιση για τη μοντελοποίηση συσχετιζόμενων χρόνων επιβίωσης είναι η χρήση του μοντέλου ευπάθειας. Η ευπάθεια αντιστοιχεί σε τυχαία αποτελέσματα και δρα πολλαπλασιαστικά στα ποσοστά κινδύνου όλων των ατόμων των υποομάδων.
Υποθέτουμε λοιπόν ότι η συνάρτηση κινδύνου για το j-οστό άτομο της i-οστή υποομάδας είναι

$$h_{ij}\left(t \mid \underset{\sim}{x}_{ij}, u_i\right) = h_o\left(t\right) u_i \exp\left(\underset{\sim}{x}_{ij}^T \beta\right), i = 1,...,n,\ j = 1,...J_i \quad (4.2.1.1)$$

Όπου είναι u_i η ευπάθεια κάθε υποομάδας. Δεδομένου της ευπάθειας u_i, οι παρατηρήσεις της i ομάδα είναι ανεξάρτητες. Η κατανομή που χρησιμοποιείται

περισσότερο στα μοντέλα ευπάθειας είναι η γάμμα λόγω της μαθηματικής της ευκολίας. Η συνάρτηση πυκνότητας πιθανότητας είναι

$$g(u)=\frac{a^a u^{a-1}\exp(-au)}{\Gamma(a)}$$

Από την (4.2.1) η πλήρης πιθανοφάνεια ,με "ψευδό-δεδομένα" $\left\{\left(u_i, X_{ij}, Z_{ij}, \delta_{ij}\right): i=1,...,G,\ j=1,...,n_i\right\}$ που θα είχαμε αν οι ευπάθειες ήταν παρατηρήσημες, είναι

$$\prod_{i=1}^{G}\prod_{j=1}^{n_i}\left[\left\{h(z_{ij}\setminus x_{ij},u_i)\right\}^{\delta_{ij}}\bar{F}(z_{ij}\setminus x_{ij},u_i)\right]\prod_{i=1}^{G}g(u_i)$$

Ολοκληρώνοντας τη πλήρη πιθανοφάνεια σε σχέση με τα $u_1,....,u_n$, η πιθανοφάνεια των παρατηρούμενων δεδομένων είναι

$$L(\beta,\alpha,H)=\exp\left\{\beta^T\left(\sum_{i=1}^{G}\sum_{j=1}^{n_i}\delta_{ij}x_{ij}\right)\right\}\prod_{i=1}^{G}\frac{a^a\prod_{j=1}^{n_i}\left\{h_0\left(z_{ij}\right)\right\}^{\delta_{ij}}}{\Gamma(a)\left\{\sum_{j=1}^{n_i}H_0(z_{ij})\exp(x_{ij}^T\beta)+a\right\}^{A_i+a}}$$

Όπου $A=\sum_{i=1}^{n_i}\delta_{ij}$. Ως εκ τούτου, η ποινικοποιημένη πιθανοφάνεια είναι

$$\sum_{i=1}^{G}\left\{\sum_{j=1}^{n_i}\delta_{ij}\log h(z_{ij})-\left[\left(A_i+a\right)\log\left\{\sum_{j=1}^{n_i}H_0(z_{ij})\exp(x_{ij}^T\beta)+a\right\}\right]\right\}+$$

$$+\sum_{i=1}^{G}\left\{\beta^T\left(\sum_{j=1}^{n_i}\delta_{ij}x_{ij}\right)+a\log a-\log\Gamma(a)\right\}-\sum_{j=1}^{d}p_\lambda(|\beta_j|)\qquad(4.2.1.2)$$

Ακολουθώντας την ιδέα του Breslow, η λιγότερο κατατοπιστική μοντελοποίηση για την $H_0(\cdot)$ είναι

$$H_0(z)=\sum_{i=1}^{N}\lambda_l I(z_l\le z)\qquad(4.2.1.3)$$

Όπου $\left\{z_1,...,z_n\right\}$ οι παρατηρήσημοι χρόνοι αποτυχίας.

Αντικαθιστώντας την (4.2.1.3) στη (4.2.1.2), η ρίζα της αντίστοιχης βαθμολογικής συνάρτησης ικανοποιεί την εξίσωση

$$\lambda_l^{-1} = \sum_{i=1}^{n} \frac{(A_i + a)\sum_{j=1}^{J_i} I(z_l \le z_{ij}) \exp(x_{ij}^T \beta)}{\sum_{k=1}^{N} \lambda_k \sum_{j=1}^{J_i} I(z_k \le z_{ij}) \exp(x_{ij}^T \beta) + a}$$

για $l = 1, ..., N$.

Με αρχικές τιμές για α, β και λ_l η $h_0(z)$ και $H_0(z)$ θα μπορούσαν να θεωρηθούν γνωστές στη συνάρτηση πιθανοφάνειας. Ωστόσο, η εκτίμηση της ποινοκοποιημένης πιθανοφάνειας είναι εκείνη που μεγιστοποιεί τη ποινικοποημένη πιθανοφάνεια των παρατηρήσημων δεομένων σε σχέση με το (α,β). Ο αλγόριθμος του Newton-Raphson, που εφαρμόζεται για τη ποινικοποιημένη πιθανοφάνεια, υπολογίζει τις δυο πρώτες παραγώγους της συνάρτησης της κατανομής γάμμα. Μια προσέγγιση που αψηφίζει τέτοιες δυσκολίες είναι η χρήση ενός πλέγματος πιθανών τιμών για την παράμετρο ευπάθειας α και η εύρεση μέγιστων πάνω από αυτό το διακριτό πλέγμα όπως προτείνεται από τον Nielsen και άλλοι (1992).

Ένας αρχικός εκτιμητής για το β είναι η ψευδό-μερική πιθανοφάνεια αγνοώντας τη πιθανή εξάρτηση μεταξύ των ομάδων. Τα αντίστοιχα $h_1, ..., h_n$ όπου $\hat{h}_j = \left\{ \sum_{i \in R_j} \exp(x_i^T \beta) \right\}^{-1}$ μπορούν να θεωρηθούν ως αρχικές εκτιμήσεις των $\lambda_1, ..., \lambda_N$.

Έτσι δοσμένου α και αρχικές τιμές β και $\lambda_1, ..., \lambda_N$, ενημερώνονται οι τιμές $\lambda_1, ..., \lambda_N$ και β μέχρι να συγκλίνουν. Ο προτεινόμενος αλγόριθμος αποφεύγει τη βελτιστοποίηση ενός υψηλού διαστάσεων προβλήματος. Θα μας δώσει όμως μια αποτελεσματική εκτίμηση για το β. Ωστόσο, ο αλγόριθμος μπορεί να μην συγκλίνει καθόλου ή να συγκλίνει αργά. Σε αυτή τη περίπτωση, η ιδέα του ενός σταδίου εκτίμησης (βλέπε Bickel, 1973) μας παρέχει μια εναλλακτική προσέγγιση.

Ο Fan και Li (2002) αξιολόγησαν την πεπερασμένη απόδοση δείγματος της παραγόμενης εκτίμησης από την εκτεταμένη μέθοδο προσομοίωσης Monte Carlo. Από τις αριθμητικές συγκρίσεις τους, φαίνεται ότι η συνάρτηση ποινής SCAD πραγματοποιείται σχεδόν όπως και ο εκτιμητής oracle όσον αφορά τη τιμή του σφάλματος και αυτό ξεπερνά την ποινικοποιημένη πιθανοφάνεια με την ποινή L_1 σε σχέση με τη πολυπλοκότητα του μοντέλου και του σφάλματος του μοντέλου. Η απόδοση της ποινής SCAD είναι όμοια με την καλύτερη επιλογή του κριτηρίου BIC από την άποψη πολυπλοκότητας και σφάλματος του μοντέλου. Ωστόσο, ο υπολογιστικός χρόνος της ποινής SCAD είναι δραματικά μικρότερος σε σχέση με τη καλύτερη επιλογή υποσυνόλου. Κάτω από ορισμένες συνθήκες κανονικότητας, ο Fan

και ο Li(2002) έδειξαν ότι αν $\lambda_n \to 0$ τότε η προκύπτουσα τιμή του SCAD είναι η , $\sqrt{n}\lambda_n \to \infty$ n με πιθανότητα να τείνει στο ένα , $\hat{\beta}_2 = 0$ και

$$\sqrt{n}\left(\hat{\theta}_1 - \theta_{10}\right) \to \mathrm{N}\left(0, \tilde{I}_1^{-1}\left(\theta_{10}\right)\right),$$

Όπου το $\tilde{I}_1\left(\theta_{10}\right)$ είναι ο πληροφοριακός πίνακας του Fisher για το μοντέλο ευπάθειας και αποτελείται από $\left(s+1\right)\times\left(s+1\right)$ υποπίνακα του $\tilde{I}_0\left(\theta_{10}, 0\right)$,και $\hat{\theta}_1 = \left(\hat{a}, \beta_1'\right)', \theta_{10} = \left(\alpha_0, \beta_{10}'\right)'$.

4.3 Προσομοιώσεις Μελέτες και Αναφορές

4.3.1 Επιλογή οριακών παραμέτρων

Για την εφαρμογή των μεθόδων που περιγράφονται στο προηγούμενο κεφάλαιο, είναι επιθυμητή η εφαρμογή μιας αυτόματης μεθόδου που να επιλέγει την οριακή τιμή λ μέσα στη ποινή $p_\lambda(\cdot)$ με βάση τα δεδομένα. Εδώ γίνεται η εκτίμηση της του λ μέσω της ελαχιστοποίησης μιας στατιστικής μεθόδου, τη λεγόμενη γενικευμένη διασταυρωμένη επικύρωσης (GCV) (Craven και Wahba, 1977). Όσον αφορά τη ποινηκοποιημένη μερική πιθανοφάνεια, ως ένα επανακαθορισμένο πρόβλημα ελαχίστων τετραγώνων, με μερικούς απλούς υπολογισμούς, ο πραγματικός αριθμός των παραμέτρων για το μοντέλο αναλογικού κινδύνου του Cox, στο τελευταίο βήμα του αλγορίθμου Newton-Raphson, είναι

$$e(\lambda) = tr\left[\left\{\nabla^2 l(\underset{\sim}{\hat{\beta}}) + \sum\nolimits_\lambda (\underset{\sim}{\hat{\beta}})\right\}^{-1} \nabla^2 l(\underset{\sim}{\hat{\beta}})\right]$$

ως εκ τούτου η γενικευμένη διασταυρωμένη επικύρωση για το μοντέλο του Cox ορίζεται από τη σχέση

$$GCV(\lambda) = \frac{-l(\underset{\sim}{\hat{\beta}})}{n\left\{1 - e(\lambda)/n\right\}^2}$$

και επιλέγεται $\hat{\lambda} = \arg\min_{\lambda} \{GCV(\lambda)\}$. Ομοίως προκύπτει και για τη συνάρτηση της ποινικοποιημένης πιθανοφάνειας για το μοντέλο ευπάθειας (4.2.1.1).

4.3.2 Σφάλμα πρόβλεψης και σφάλμα μοντέλου

Όπως έχει είδη αναφερθεί στη παράγραφο 3.4.3.1 το προβλεπόμενο σφάλμα δίνεται από τον τύπο

$$PE(\hat{\mu}) = E\{Y - \hat{\mu}(\underset{\sim}{x})\}^2$$

όπου τα δεδομένα $(\underset{\sim}{x}_i, Y_i)$ θεωρούνται τυχαίο δείγμα από κάποια κατανομή. Ωστόσο, ο παραπάνω τύπος αναλύεται ως,

$$PE(\hat{\mu}) = E\{Y - E(Y \mid \underset{\sim}{x})\}^2 + E\{E(Y \mid \underset{\sim}{x}) - \hat{\mu}(\underset{\sim}{x})\}^2$$

Ο δεύτερος όρος ονομάζεται σφάλμα μοντέλου (*model error*) και συμβολίζεται ως $ME(\hat{\mu})$.

- Για το μοντέλο αναλογικού κινδύνου του Cox είναι,

$$\mu(\underset{\sim}{x}) = E(T \mid x) \int_0^{\infty} t h_0(t_0) \exp(\underset{\sim}{x}'\beta) \exp\left\{-\int_0^t h_0(t_0) \exp(\underset{\sim}{x}\beta)\, du\right\} dt .$$

Για $h_0(t) \equiv 1$ είναι $\mu(\underset{\sim}{x}) = \exp\left(-\underset{\sim}{x}'\hat{\beta}\right)$.

- Για το μοντέλο ευπάθειας με $h_0(t) \equiv 1$ το σφάλμα του μοντέλου είναι

$$\mu(\underset{\sim}{x}) = \exp\left(-\underset{\sim}{x}'\hat{\beta}\right) E\left(u^{-1}\right).$$

Ο παράγοντας $E\left(u^{-1}\right)$, λόγω της ευπάθειας, αφαιρείται από τον τύπο όταν η απόδοση των δύο διαφορετικών προσεγγίσεων συγκρίνεται με τα σχετικά σφάλματα του μοντέλου (RME- relative model error). Ως εκ τούτου, το σφάλμα του μοντέλου ορίζεται ως

$$E\left\{\exp\left(-\underset{\sim}{x}'\hat{\beta}\right) - \exp\left(-\underset{\sim}{x}'\beta_0\right)\right\}^2$$

τόσο για το μοντέλο του Cox τόσο και για το μοντέλο ευπάθειας.

4.3.3 Προσομοιώσεις

Στο παράδειγμα που ακολουθεί από τους Fan & Li (2002), συγκρίνονται αριθμητικά οι προτεινόμενοι μέθοδοι επιλογής μεταβλητών σύμφωνα με την εκτίμηση μέγιστης μερικής πιθανοφάνειας και της μεθόδου καλύτερης επιλογής υποσυνόλου. Όλες οι προσομοιώσεις έχουν γίνει μέσω του προγράμματος MATLAB. Για την εύρεση του καλύτερου υποσυνόλου, έγινε έλεγχος όλων των δυνατών υποσυνόλων και επιλέχθηκε εκείνο με την καλύτερη τιμή του *BIC* κριτηρίου.

Παράδειγμα 1.

Σε αυτό το παράδειγμα προσομοιώνονται 100 σύνολα δεδομένων που αποτελούνται από n ομάδες και j άτομα στη κάθε ομάδα από το εκθετικό μοντέλο κινδύνου ευπάθειας,

$$h\left(t \mid x,u\right) = u\exp\left(x'\underset{\sim}{\beta}\right)$$

όπου το $\underset{\sim}{\beta} = (0.8,0,0,1,0,0,0.6,0)'$. Η συσχέτιση μεταξύ των x_i και x_j είναι $\rho^{|i-j|}$ με $\rho = 0.5$ και η ευπάθεια u είναι το μοντέλο γάμμα της ευπάθειας για α=4.

Η απόδοση των μεταβλητών που επιλέχθηκαν μέσω της μη κοίλης ποινικοποιημένης πιθανοφάνειας και της μεθόδου επιλογής μεταβλητής καλύτερου υποσυνόλου, συγκρίνεται σε σχέση με τα σφάλματα του μοντέλου, τη πολυπλοκότητα του μοντέλου καθώς και την ακρίβεια. Τα σφάλματα του μοντέλου συγκρίνονται με εκείνα των εκτιμήσεων της μέγιστης πιθανοφάνειας.

Simulation results for frailty model

Method	MRME(%)	Aver. no. of 0 coeff.	
		correct	incorrect
$n = 50,\ J = 2$			
SCAD	0.5322	4.18	0.14
LASSO	0.8880	4.04	0.06
Hard	0.5784	4.54	0.09
Best Subset	0.4251	4.78	0.07
Oracle	0.3592	5	0
$n = 75,\ J = 2$			
SCAD	0.5177	4.18	0
LASSO	1.4075	4.08	0
Hard	0.5782	4.50	0
Best Subset	0.5188	4.89	0
Oracle	0.4886	5	0
$n = 100,\ J = 2$			
SCAD	0.4930	4.29	0
LASSO	1.0438	4.10	0
Hard	0.6379	4.42	0
Best subset	0.6019	4.85	0
Oracle	0.5631	5	0

Πίνακας 4.1

Η διάμεσος των σχετικών σφαλμάτων του μοντέλου (*Median of Relative Model Errors –MRME*) από 100 προσομοιωμένα δεδομένα, με κάποιο συνδυασμό των n και j, παρουσιάζεται περιληπτικά στον ***Πίνακας 4.1.*** Επίσης, στον ίδιο πίνακα φαίνεται και ο μέσος αριθμός των μηδενικών συντελεστών, με τη στήλη «correct» να αντιστοιχεί στο μέσο αριθμό των σωστά εκτιμώμενων ως μηδενικοί συντελεστές, ενώ η στήλη «incorrect» αντιστοιχεί σε αυτούς που λανθασμένα εκτιμήθηκαν ως μηδενικοί.

Τα τυπικά σφάλματα για την εκτίμηση των μη μηδενικών συντελεστών με n = 100 και J =2 παρουσιάζονται στον ***Πίνακα 4.2*** . Στη συνέχεια έγινε έλεγχος της ακρίβειας της sandwich μεθόδου υπολογισμού του τυπικού σφάλματος. Η διάμεσος των απολύτων τιμών της απόκλισης των 100 εκτιμώμενων συντελεστών των 100 συνόλων δεδομένων, διαιρεμένη με 0.6745, συμβολισμένη ως *SD*, βρίσκεται στον ***Πίνακα 4.2*** και μπορεί να θεωρηθεί ως το πραγματικό τυπικό σφάλμα εκτός το σφάλμα Monte Carlo. Η διάμεσος των 100 αυτών εκτιμώμενων *SDs* που προκύπτουν από τις 100 προσομοιώσεις συμβολίζεται με SD_m και η διάμεσος των απολύτων τιμών του σφάλματος της απόκλισης των 100 εκτιμημένων τυπικών σφαλμάτων διαιρεμένη με 0.6745, συμβολίζεται με SD_{mad}. Τα δύο τελευταία αξιολογούν και την

καταλληλότητα της μεθόδου. Στις προσομοιώσεις, το τυπικό σφάλμα των συντελεστών τέθηκε ίσο με μηδέν, αν οι μεταβλητές που τους αντιστοιχούν δεν περιλαμβάνονται τελικά στο επιλεγμένο μοντέλο. Ο ***Πίνακας 4.2*** περιέχει τα αποτελέσματα μόνο για τους μη μηδενικούς συντελεστές και για $n=100$. Η τελευταία σειρά του ***Πίνακα 4.2*** εμφανίζει τις τυπικές αποκλίσεις και τα τυπικά σφάλματα των εκτιμημένων συμμεταβλητών βασιζόμενα στο πραγματικό μοντέλο (ιδιότητα πρόβλεψης). Από τους δυο πίνακες συμπεραίνουμε ότι η *SCAD* να έχει τόσο καλά αποτελέσματα όσο αυτά του *oracle* εκτιμητή (ο οποίος επίσης χρησιμοποιήθηκε ώστε να συγκριθεί με τις προτεινόμενες μεθόδους), που ξεπερνά τις άλλες εκτιμήσεις καθώς το μέγεθος του δείγματος αυξάνει.

Standard deviations for frailty models ($n = 100,\ J = 2$)

	$\hat{\beta}_1$		$\hat{\beta}_4$		$\hat{\beta}_7$	
Method	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$
SCAD	0.114	0.100(0.012)	0.092	0.095(0.007)	0.113	0.098(0.008)
LASSO	0.098	0.077(0.007)	0.086	0.082(0.006)	0.097	0.072(0.008)
Hard	0.083	0.101(0.008)	0.095	0.102(0.008)	0.094	0.102(0.009)
Best Subset	0.080	0.103(0.008)	0.092	0.103(0.008)	0.090	0.102(0.010)
Oracle	0.083	0.100(0.008)	0.089	0.102(0.007)	0.087	0.102(0.009)

Πίνακας 4.2
Τυπικές αποκλίσεις των εκτιμητών στο μοντέλο ευπάθειας

4.3.4 Συμπεράσματα

Η επιλογή μεταβλητής μέσω μη κοίλης ποινικοποιημένης πιθανοφάνειας έχει επιτυχώς επεκταθεί, στα τελευταία δυο κεφάλαια, για τα μοντέλο ευπάθειας. Επιπλέον, οι ιδιότητες oracle έχουν καθιερωθεί για τους προτεινόμενους εκτιμητές καθώς και πολλές είναι οι αριθμητικές συγκρίσεις που διεξάγονται. Η συνάρτηση ποινής SCAD απέδειξε ότι έχει άριστες επιδόσεις όσον αφορά τη μείωση της πολυπλοκότητας του μοντέλου και του σφάλματος επίσης. Τέλος η SCAD, σε σχέση με τη LASSO επιφέρει καλύτερα αποτελέσματα και συμβάλλει στη καλύτερη επιλογής μεταβλητής υποσυνόλου.

ΠΑΡΑΡΤΗΜΑ Ι

Υποθέσεις κανονικότητας (*regularity conditions*)

(A) Οι παρατηρήσεις $\underset{\sim}{V}_i$ είναι i.i.d. με συνάρτηση πυκνότητας πιθανότητας $f(\underset{\sim}{V},\underset{\sim}{\beta})$. Η $f(\underset{\sim}{V},\underset{\sim}{\beta})$ έχει μια κοινή βάση και το μοντέλο είναι αναγνωρίσιμο (*identifiable*). Επίσης, η πρώτη και η δεύτερη λογαριθμημένη παράγωγος της f ικανοποιεί τις εξισώσεις

$$E_{\underset{\sim}{\beta}}\left[\frac{\partial \log f(\underset{\sim}{V},\underset{\sim}{\beta})}{\partial \beta_j}\right]=0\text{, για } j=1,...,d$$

και

$$I_{jk}(\underset{\sim}{\beta})=E_{\underset{\sim}{\beta}}\left[\frac{\partial}{\partial \beta_j}\log f(\underset{\sim}{V},\underset{\sim}{\beta})\frac{\partial}{\partial \beta_k}\log f(\underset{\sim}{V},\underset{\sim}{\beta})\right]=E_{\underset{\sim}{\beta}}\left[-\frac{\partial^2}{\partial \beta_j \partial \beta_k}\log f(\underset{\sim}{V},\underset{\sim}{\beta})\right].$$

(B) Ο πίνακας πληροφορίας του Fisher

$$I(\underset{\sim}{\beta})=E\left\{\left[\frac{\partial}{\partial \underset{\sim}{\beta}}\log f(\underset{\sim}{V},\underset{\sim}{\beta})\right]\left[\frac{\partial}{\partial \underset{\sim}{\beta}}\log f(\underset{\sim}{V},\underset{\sim}{\beta})\right]'\right\}$$

είναι πεπερασμένος και θετικά ορισμένος στο $\underset{\sim}{\beta}=\underset{\sim}{\beta}_0$.

(C) Υπάρχει ένα ανοικτό υποσύνολο ω του Ω το οποίο περιέχει την πραγματική παράμετρο $\underset{\sim}{\beta}_0$ τέτοιο ώστε για σχεδόν όλα τα $\underset{\sim}{V}$, η συνάρτηση πυκνότητας πιθανότητας $f(\underset{\sim}{V},\underset{\sim}{\beta})$ επιδέχεται τις παραγώγους τρίτης τάξης

$$\frac{\partial f(\underset{\sim}{V},\underset{\sim}{\beta})}{\partial \beta_j \partial \beta_k \partial \beta_l}\text{, για όλα τα } \underset{\sim}{\beta} \in \omega.$$

Επίσης, υπάρχουν συναρτήσεις M_{jkl} τέτοιες ώστε

$$\left|\frac{\partial^3}{\partial \beta_j \partial \beta_k \partial \beta_l}\log f(\underset{\sim}{V},\underset{\sim}{\beta})\right| \le M_{jkl}(\underset{\sim}{V})\text{, για όλα τα } \underset{\sim}{\beta} \in \omega,$$

όπου $m_{jkl}=E_{\underset{\sim}{\beta}_0}[M_{jkl}]<\infty$, $\forall\, j,k,l$.

Απόδειξη θεωρήματος 1

Έστω ότι $\alpha_n = n^{-1/2} + a_n$. Θέλουμε να δείξουμε ότι για οποιοδήποτε $\varepsilon > 0$, υπάρχει σταθερά C, ώστε

$$P\left\{ \sup_{\|\underset{\sim}{u}\|=C} Q(\underset{\sim}{\beta}_0 + \alpha_n \underset{\sim}{u}) < Q(\underset{\sim}{\beta}_0) \right\} \geq 1-\varepsilon \quad (1.1)$$

Αυτό συνεπάγεται, ότι με πιθανότητα τουλάχιστον $1-\varepsilon$, θα υπάρχει ένα τοπικό μέγιστο που θα ανήκει στη σφαίρα $\left\{ \underset{\sim}{\beta}_0 + \alpha_n \underset{\sim}{u} : \| \underset{\sim}{u} \| \leq C \right\}$. Συνεπώς, θα υπάρχει ένα τοπικό ελάχιστο τέτοιο ώστε $\| \hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta}_0 \| = O_p(\alpha_n)$. Χρησιμοποιώντας τώρα $p_{\lambda_n}(0) = 0$, έχουμε

$$D_n(\underset{\sim}{u}) \equiv Q(\underset{\sim}{\beta}_0 + \alpha_n \underset{\sim}{u}) - Q(\underset{\sim}{\beta}_0) \leq L(\underset{\sim}{\beta}_0 + \alpha_n \underset{\sim}{u}) - L(\underset{\sim}{\beta}_0) - n\sum_{j=1}^{s} \left\{ p_{\lambda_n}(|\beta_{j0} + \alpha_n u_j|) - p_{\lambda_n}(|\beta_{j0}|) \right\}$$

όπου s είναι ο αριθμός των συνιστωσών του $\underset{\sim}{\beta}_{10}$. Έστω τώρα $L'(\underset{\sim}{\beta}_0)$ να είναι το βαθμωτό διάνυσμα του L. Χρησιμοποιώντας επέκταση Taylor της συνάρτησης πιθανοφάνειας, έχουμε

$$D_n(\underset{\sim}{u}) \leq \alpha_n L'(\underset{\sim}{\beta}_0)' \underset{\sim}{u} - \frac{1}{2} \underset{\sim}{u}' I(\underset{\sim}{\beta}_0) \underset{\sim}{u} n\alpha_n^2 \{1 + o_p(1)\}$$

$$-\sum_{j=1}^{s} \left\{ n\alpha_n p'_{\lambda_n}(|\beta_{j0}| \operatorname{sgn}(|\beta_{j0}|)) u_j + n\alpha_n^2 p''_{\lambda_n}(|\beta_{j0}|) u_j^2 \{1 + o_p(1)\} \right\} \quad (1.2)$$

Να παρατηρήσουμε ότι $n^{-1/2} L'(\underset{\sim}{\beta}_0) = O_p(1)$. Συνεπώς, ο πρώτος όρος του δεξιού μέλους της (1.2) είναι της τάξης $O_p(n^{1/2}\alpha_n) = O_p(n\alpha_n^2)$. Επιλέγοντας αρκετά μεγάλο C, ο δεύτερος όρος επικρατεί του πρώτου ομοιόμορφα στο $\| \underset{\sim}{u} \| = C$. Επίσης, ο τρίτος όρος φράσσεται από την ποσότητα

$$\sqrt{s} n\alpha_n a_n \| \underset{\sim}{u} \| + n\alpha_n^2 \max \left\{ |p''_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0} \neq 0 \right\} \| \underset{\sim}{u} \|^2 .$$

Ο δεύτερος όρος της (1.2) επικρατεί και εδώ. Οπότε, με την επιλογή επαρκώς μεγάλου C, η (1.1) ισχύει και αυτό ολοκληρώνει την απόδειξη.

Λήμμα 1

Έστω πάλι ότι ότι τα $\underset{\sim}{V}_1,...,\underset{\sim}{V}_n$ *είναι i.i.d., κάθε ένα με συνάρτηση πυκνότητας πιθανότητας* $f(\underset{\sim}{V},\underset{\sim}{\beta})$ *και ότι ικανοποιούν τις υποθέσεις (A)-(C). Έστω ότι*

$$\liminf_{n\to\infty} \liminf_{\theta\to 0_+} p'_{\lambda_n}(\theta)/\lambda_n > 0 \quad (1.3)$$

Αν $\lambda_n \to 0$ *και* $\sqrt{n}\lambda_n \to \infty$ *όσο το* $n\to\infty$*, τότε με πιθανότητα που τείνει στο* 1*, για κάθε δοσμένο* $\underset{\sim}{\beta}_1$ *που ικανοποιεί* $\| \underset{\sim}{\beta}_1 - \underset{\sim}{\beta}_{10} \| = O_p(n^{-1/2})$

και για κάθε σταθερά C*, ισχύει ότι*

$$Q\left\{\begin{pmatrix}\underset{\sim}{\beta}_1\\ \underset{\sim}{0}\end{pmatrix}\right\} = \max_{\|\underset{\sim}{\beta}_2\|\le Cn^{-1/2}} Q\left\{\begin{pmatrix}\underset{\sim}{\beta}_1\\ \underset{\sim}{\beta}_2\end{pmatrix}\right\}.$$

Ορίζουμε τώρα ως $\Sigma = diag\left\{p''_{\lambda_n}(|\beta_{10}|,...,(|\beta_{s0}|)\right\}$

και $\underset{\sim}{b} = \left(p'_{\lambda_n}(|\beta_{10}|)\mathrm{sgn}(\beta_{10}),..., p'_{\lambda_n}(|\beta_{s0}|)\mathrm{sgn}(\beta_{s0})\right)'$.

Απόδειξη του Λήμματος 1

Πρέπει να δείξουμε ότι, με πιθανότητα να τείνει στο 1 όσο το $n\to\infty$, για οποιοδήποτε $\underset{\sim}{\beta}_1$ που ικανοποιεί $\underset{\sim}{\beta}_1 - \underset{\sim}{\beta}_{10} = O_p(n^{-1/2})$ και για κάποιο $\varepsilon_n = Cn^{-1/2}$ και $j = s+1,...,d$, ισχύει

$$\frac{\partial Q(\underset{\sim}{\beta})}{\partial\beta_j} < 0, \text{ για } 0 < \beta_j < \varepsilon_n \quad (1.4)$$

$$\frac{\partial Q(\underset{\sim}{\beta})}{\partial\beta_j} > 0, \text{ για } -\varepsilon_n < \beta_j < 0 \quad (1.5)$$

Για να δείξουμε ότι ισχύει η (1.4), χρησιμοποιώντας επέκταση Taylor, έχουμε

$$\frac{\partial Q(\underset{\sim}{\beta})}{\partial\beta_j} = \frac{\partial L(\underset{\sim}{\beta})}{\partial\beta_j} - np'_{\lambda_n}(|\beta_j|)\mathrm{sgn}(|\beta_j|)$$

$$= \frac{\partial L(\underset{\sim}{\beta}_0)}{\partial \beta_j} + \sum_{l=1}^{d} \frac{\partial^2 L(\underset{\sim}{\beta}_0)}{\partial \beta_j \partial \beta_l} (\beta_l - \beta_{l0})$$

$$+ \sum_{l=1}^{d} \sum_{k=1}^{d} \frac{\partial^3 L(\underset{\sim}{\beta}^*)}{\partial \beta_j \partial \beta_l \partial \beta_k} \times (\beta_l - \beta_{l0})(\beta_k - \beta_{k0}) - n p'_{\lambda_n}(|\beta_j|) \operatorname{sgn}(|\beta_j|),$$

όπου το $\underset{\sim}{\beta}^*$ είναι μεταξύ των $\underset{\sim}{\beta}$ και $\underset{\sim}{\beta}_0$. Να σημειώσουμε ότι

$$n^{-1} \frac{\partial L(\underset{\sim}{\beta}_0)}{\partial \beta_j} = O_p(n^{-1/2})$$

και

$$\frac{1}{n} \frac{\partial^2 L(\underset{\sim}{\beta}_0)}{\partial \beta_j \partial \beta_l} = E\left\{ \frac{\partial^2 L(\underset{\sim}{\beta}_0)}{\partial \beta_j \partial \beta_l} \right\} + o_p(1).$$

Από την υπόθεση ότι $\underset{\sim}{\beta} - \underset{\sim}{\beta}_0 = O_p(n^{-1/2})$, έχουμε τώρα ότι

$$\frac{\partial Q(\underset{\sim}{\beta})}{\partial \beta_j} = n\lambda_n \left\{ -\lambda_n^{-1} p'_{\lambda_n}(|\beta_j|) \operatorname{sgn}(|\beta_j|) + O_p(n^{-1/2} / \lambda_n \right\}.$$

Λαμβάνοντας υπόψη ότι $\liminf_{n\to\infty} \liminf_{\theta\to 0^+} \lambda_n^{-1} p'_{\lambda_n}(\theta) > 0$ και $n^{-1/2} / \lambda_n \to 0$,

το πρόσημο της παραγώγου εξαρτάται αποκλειστικά από αυτό του β_j. Συνεπώς, οι (1.4) και (1.5) ισχύουν.

Απόδειξη του Θεωρήματος 2

Από το προηγούμενο Λήμμα, έχουμε ότι ισχύει η σποραδικότητα. Θα αποδείξουμε τώρα την ασυμπτωτική κανονικότητα. Μπορεί εύκολα να δειχθεί ότι υπάρχει ένα $\hat{\underset{\sim}{\beta}}_1$ στο πρώτο Θεώρημα το οποίο είναι ένα $\sqrt{n}$-συνεπές τοπικό μέγιστο του $Q\left\{ \begin{pmatrix} \underset{\sim}{\beta}_1 \\ \underset{\sim}{0} \end{pmatrix} \right\}$, το οποίο θεωρείται ως συνάρτηση του $\underset{\sim}{\beta}_1$ και ικανοποιεί τις εξισώσεις

$$\left. \frac{\partial Q(\underset{\sim}{\beta})}{\partial \beta_j} \right|_{\underset{\sim}{\beta} = \begin{pmatrix} \hat{\underset{\sim}{\beta}}_1 \\ \underset{\sim}{0} \end{pmatrix}} = 0, \text{ για } j = 1, \ldots, s.$$

Να σημειώσουμε ότι ο $\hat{\underset{\sim}{\beta}}_1$ είναι συνεπής εκτιμητής,

$$\left.\frac{\partial L(\underset{\sim}{\beta})}{\partial \beta_j}\right|_{\underset{\sim}{\beta}=\begin{pmatrix}\hat{\underset{\sim}{\beta}}_1\\ \underset{\sim}{0}\end{pmatrix}} - np'_{\lambda_n}(|\hat{\beta}_j|)\operatorname{sgn}(|\hat{\beta}_j|)$$

$$= \frac{\partial L(\underset{\sim}{\beta}_0)}{\partial \beta_j} + \sum_{l=1}^{s}\left\{\frac{\partial^2 L(\underset{\sim}{\beta}_0)}{\partial \beta_j \partial \beta_l} + o_p(1)\right\}(\hat{\beta}_l - \beta_{l0})$$

$$-np'_{\lambda_n}(|\beta_{j0}|)\operatorname{sgn}(|\beta_{j0}|) + \{p''_{\lambda_n}(|\beta_{j0}|) + o_p(1)\}(\hat{\beta}_j - \beta_{j0}) .$$

Από το Θεώρημα Slutsky καθώς και το Κεντρικό Οριακό Θεώρημα, έχουμε τελικά ότι

$$\sqrt{n}\left(I_1(\underset{\sim}{\beta}_{10}) + \Sigma\right)\left\{\hat{\underset{\sim}{\beta}}_1 - \underset{\sim}{\beta}_{10} + \left(I_1(\underset{\sim}{\beta}_{10}) + \Sigma\right)^{-1}\underset{\sim}{b}\right\} \to \mathrm{N}\left\{\underset{\sim}{0}, I_1(\underset{\sim}{\beta}_{10})\right\}$$

ΒΙΒΛΙΟΓΡΑΦΙΑ

Aalen, O.O. (1992). Modeling heterogeneity in survival analysis by the compound Poison distribution. Ann.Appl. Prob**. 2**, pp. 951-972.

Aalen, O.O. and S. Tretli.(1999). Analyzing incidence of testis cancer by means of a frailty model. Cancer Causes and Control, **10(4)**, pp. 285-292.

Aitchison, J., & Dunsmore, I. R. (1975). *Statistical prediction analysis*. Cambridge: Cambridge University Press.

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**, pp. 716-723.

Akaike, H. (1977). On entropy maximization principle. *Applications of Statistics* (Krishnaiah, P.R., ed.), North Holland: Amsterdam, pp. 27-41.

Aldrich, J. (1997). R. A. Fisher and the Making of Maximum Likelihood 1912-1922. *Statistical Science*, **22**, pp. 162-176.

Allen, D. M. (1971). The prediction sum of squares as a criterion for selecting predictor variables. *Technical Report No. **23***. Department of Statistics. University of Kentuky.

Androulakis, E. (2008). *Μέθοδοι επιλογής μεταβλητών στο μοντέλο αναλογικού κινδύνου του Cox και εφαρμογές σε πραγματικά δεδομένα με αποκομμένες παρατηρήσεις*. Εθνικό Μετσόβιο Πολυτεχνείο.

Antoniadis, A. (1997). Wavelets in statistics: a review (with discussion). *J. Italian Statist. Assoc*., **6**, pp. 97-144.

Anotoniadis, A. (1999). Wavelets in Statistics: A Review. Italian Jour. Statist., to appear.

Bickel, P. J. (1975). One-Step Huber Estimates in Linear Models. *Journal of the American Statistical Association*(**70**), σσ. 428–433.

Bickel, P.J (1983). Minimax estimation of a normal mean subject to doing well at a point. In *Recent Advances in Statistics* (M.H. Rizvi, J.S Rustagi, and D.Siegmund, eds), 511-528. Academic Press, New York.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*,**52**, 345- 370.

Breiman, L. (1993) Better subset selection using the non-negative garotte, Technical Report. University of California, Berkeley.

Breiman, L. (1995). Better Subset Regression Using the Nonnegative Garrote, *Technometrics*, **37**, pp.373–384.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. Ann. Statist., **24**. 2350-2383.

Breslow, N. E. and Crowley. (1974). A large sample study of the life table and product limit estimates under random censorship. Annals of Statistics, **2**:437-453.

Breslow, N.E. and N.E. Day. (1987). The design and analysis of cohort studies. Statistical methods in cancer research. Volume II, IARC Sci Publ **87**, pp. 1-406.

Bretagnolle, J. and C. Huber-Carol. (1985). Sous-estimation des contrastes due a l'oubli de variables pertinentes dans le modele de Cox pour des durees de survie avec censure. Comptes Rendues de l'Academie des Sciences, **300**, pp. 359-363.

Bretagnolle J. and C. Huber-Carol. (1988). Effects of omitting covariates in Cox's model for survival data. Scandinavian Journal of Statistics, **15**, pp.125-138.

Burnham, K. P ., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York:Springer-Verlag.

Burnham,K.P., and Anderson, D.R (2004) "Multimodel Inference": understanding AIC and BIC in Model Selection" Sociological Methods and Research",**33**: 261-304.

Chamberlain, G. (1985). Longitudinal analysis of labor market data, Chapter Heterogeneity, omitted variable bias, and duration dependence.

Chastang, C. D. Byar, and S. Piantadosi. (1988). A quantitative study of the bias in estimating the treatment effect caused by omitting a based covariate in survival models. Statistics in Medicine, **7**, pp. 1243-1255.

Clayton, D. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika, 65(1), pp. 141.

Clayton, D. and J. Cuzick. (1985). The EM Algorithm for Cox's Regression Model Using GLIM. Applied Statistics, **34(2),** pp. 148-156.

Cox, D. R. (1972). Regression models and life-tables (with discussion). J. R. Statist. Soc. B, **34**, pp. 187-220

Cox, D.R. and N. Wermuth. (1996). Multivariate dependencies. Chapman & Hall New York.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions:estimating the correct degree of smoothing by the method of generalized cross-validation. Numer. Math., **31**, 377-403.

Donoho, D. L., and Johnstone, I. M. (1994a). Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika*, **81**, pp. 425–455.

Duchateau L, Janssen P. (2008). *The Frailty Model*. Springer: New York

Efron. B. (1967). The two sample problem with censored data. In Proceedings of the Fifth Sym-posium on Mathematical Statistics and Probability, **4**, Ed. L. LeCam and J. Neyman, pp. 831-853, Berkeley, University of California Press.

Fan, J. (1997). Comments on "Wavelets in statistics a review" by A. Antoniadis. *J. Italian Statist. Assoc*., **6**, pp. 131-138.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc*., **96,** pp. 1348-1360.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, **30**, pp. 74-99.

Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *International Congress of Mathematicians 2006*. **3**, pp. 595-622.

Faraggi, D. and Simon, R. (1998). Bayesian variable selection method for censored survival data. *Biometrics*, 54, pp.1475-1485.

Farrell Simon and Eric- Jan Wagenmakers, AIC model selection using Akaike weights,*Northwestern University, Evanston,IllinoisPsychonomic Bulletin & Review, 2004, **11 (1)**, 192-196*).

Fine, J.P. D.V. Glidden, and K.E. Lee. (2003) .A simple estimator for a shared frailty regression model. Journal of the Royal Statistical Society Series B (Statistical Methodology), **65(1),** pp. 317-329.

Frank, I. E., and Friedman, J. H. (1993), A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, **35**, pp. 109–148.

Fu, W. J. (1998). Penalized Regression: The Bridge Versus the LASSO. *Journal of Computational and Graphi(cal Statistics***, 7**, pp. 397–416.

Gail, M.H. S. Wieand, and S. Piantadosi. (1984). Biased estimates of treatment effects in randomized experiments with non-linear regressions and omitted covariates. Biometrika, 71, pp. 431-444.

Gao, H. Y. and Bruce, A. G. (1997). WaveShrink with firm Shrinkage. Statistica Sinica, **7**, pp. 855-874.

Genest, C. (1987). Frank's family of bivariate distributions. Biometrika, 74(3), pp. 549.

Glidden, D.V. (1999). Checking the adequacy of the gamma frailty model for multivariate failure imes. Biometrika, **86(2**), pp. 381-393.

Golan, A. (Ed.) (2002). Information and entropy econometrics [Special issue]. *Journal of Econometrics*, **107** (1-2).

Gottard, A., C. Rampichini, and C. Graphs. (2004). Chain Graphs for Multilevel Models. Working manuscript.

Henderson, R. and P. Oman(1999). Effect of frailty on marginal regression estimates in survival analysis. Journal of the Royal Statistical Society Series B (Statistical Methodology), **61** (2), pp. 367-379.

Hoerl E. and Robert W. Kennard, (Feb., 1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, Vol. **12**, No. 1, pp. 55-67.

Hoerl E. and Robert W. Kennard, (Feb., 2000) Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, Vol. **42**, No. 1, Special 40th Anniversary Issue, pp. 80-86.

Hougaard, P. (1986b). A class of multivariate failure time distribution. IEEE Trans. On Reliability **38**, pp. 444-448.

Hougaard, P. (2000). Analysis of Multivariate Survival Data: Statistics of Biology and Health. Springer-Verlag : New York, pp. 215-310.

Hurvich, C. M., & Tsai, C.-L. (1995). Model selection for extended quasi- likelihood models in small samples. *Biometrics*, **51**, 1077-1084.

Huber P. (1981). *Robust Estimation*. Wiley, New York

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, **53**, pp. 457-448.

Καρώνη Χ. (2005). *Μοντέλα Αξιοπιστίας και Επιβίωσης*. Ε.Μ.Π.

Καρώνη, Χ., Οικονόμου, Π. (2010). Στατιστικά Μοντέλα Παλινδρόμησης. Αθήνα: Εκδόσεις Συμεων, pp.181-190.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773-795.

Klein, J. P. and Moeschberger, M. L. (1997). Survival Analysis. Springer. New York.

Koukouvinos,C., Mylona, K. and Vonta, F. (2008). A Comparative study of variable selection procedures applied in high dimensional medical problems. *Journal of Applied Probability & Statistics*, **3**, pp. 195-209.

Koziol, J. A. and Green, S. B. (1976).A cramer-von mises statistic for randomly censored data. Biometrika, **63**, pp. 465-474.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency.*Annals of Mathematical Statistics*, **22**, 79-86.

Lawson, C. L. and Hanson, R. J. (1974). *Solving least-squares problems*. Prentice Hall, New Jersey.

Lee, E.W., Wei, L. , and Amato, D.A. (1992). A Cox-type regression analysis for large numbers of small groups of correlated failure yime observations, 237-248. In Survival Analysis: State of the Art, J.P. Klein and P. Geol, eds. Boston: Kluwer Academic Publishers.

Lee, S. and J.P. Klein. (1988). Bivariate models with a random environmental factor. Industrial Journal of Productivity, Reliability, and Quality Control, **13**, pp. 1-18.

Lindley, D. V. (1968). The choice of variables in multiple regression (with discussion). Jour. Roy. Statist. Soc., B, **30**, 31-66.

Lindley, D.V. and N.D. Singpurwalla, (1986). Multivariate Distributions for the Life Lengths of Components of a System Sharing a Common Environment. Journal of Applied Probability, 23(2), pp. 418-431.

Mallows, C. L. (1973). Some comments on C_p. Technometrics, **15**, pp. 661-675.

Marron, J. S., Adak, S., Johnstone, I. M., Neumann, M. H., and Patil, P. (1998). Exact Risk Analysis of Wavelet Regression. *Journal Computational and Graphical Statistics*, **7**, pp. 278–309.

Miller, A. (2002), Subset Selection in regression, 2nd Edition, Chapman and Hall/CRC, Florida.

Mohammad Ehsanul Karim. (2008). Frailty Models. Institute of Statistical Research and Training University of Dhaka, Dhaka, Bangladesh April 4, pp. 1-15.

Nielsen, G. G., Gill, R. D., Andersen, P. K., and Sø rensen, T. I. A. A. (1992). A counting process approach to maximum likelihood estimator in frailty models. Scandin. J. Statist., **19**, 25-43.

Oakes, D. (1982). A Model for Association in Bivariate Survival Data. Journal of the Royal Statistical Society. Series B (Methodological), 44(3), pp. 414-422.

Oakes, D.(2001). Biometrika centenary: Survival analysis. Biometrika, 88, pp. 99-142.

Roberto G. Gutierrez. (2002). Parametric frailty and shared frailty survival models. The Stata Journal **2**, Number 1, pp. 22–44.

Runze, Li. (2000). High- dimensional Modeling via Nonconcave Penalized Likelihood and Local Likelihood (Dissertation). Department of Statistics, University of North Carolina at Chapel Hill, USA. pp. 1-75.

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). *Akaike information criterion statistics*. Dordrecht: Reidel.

Shih, J.H. and T.A. Louis. (1995b). Inferences on the Association Parameter in Copula Models for Bivariate Survival Data. Biometrics, 51(4), pp. 1384-1399.

Sinha, E. (1998). Posterior likelihood methods for multivariate survival data, Biometrics, **54**, 1463-1474

Scott, J. T., Jr. (1966), "Factor Analysis and Regression," Econometrica, **34**, 552-562.

Schwartz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**, pp. 461-464.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO. *J. Roy. Statist. Soc. Ser. B.*, **58**, pp. 267-288.

Tibshirani, R. J. (1997). The Lasso method for variable selection in the Cox model. *Statistics in Medicine*, **16**, pp. 385-395.

Tibshirani R., Hastie T. and Friedman J. , August (2008). The elements of statistical learning: Data Mining, Inference, and Prediction. Stanford, California: *Springer*, pp. 87-98.

Vaupel, J.W., K.G. Manton, and E. Stallard. (1979). The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality. Demography 16(3), pp. 439 454.

Virginie Rondeau , Yassin Mazroui , Juan R. Gonzalez frailtypack, April (2012). An R Package for the Analysis of Correlated Survival Data with Frailty Models Using Penalized Likelihood Estimation. Journal of Statistical Software, Volume 47, Issue 4.

Wienke, A., K. Arbeev, I. Locatelli, and A.I. Yashin. (2003a). A simulation study of different correlated frailty models and estimation strategies. Technical report, MPIDR Working Paper WP 2003.