ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Μεθοδοι Και Κριτηρια Επιλογης Μοντελου Με Ποινη

Διπλωματικη Εργασια Ειδικευσης
στις Εφαρμοσμενες Μαθηματικες Επιστημες

ΣΠΥΡΙΔΟΥΛΑ ΚΑΝΤΑ

ΕΠΙΒΛΕΠΟΥΣΑ: ΧΡΥΣΗΙΣ ΚΑΡΩΝΗ
ΑΝΑΠΛΗΡΩΤΡΙΑ ΚΑΘΗΓΗΤΡΙΑ

ΑΘΗΝΑ 2013

Η παρούσα Διπλωματική Εργασία εκπονήθηκε
στα πλαίσια των σπουδών για την απόκτηση του
Μεταπτυχιακού Διπλώματος Ειδίκευσης στις
Εφαρμοσμένες Μαθηματικές Επιστήμες.

Ονοματεπώνυμο	**Βαθμίδα**
Χρυσηίς Καρώνη (Επιβλέπουσα)	Αναπληρώτρια Καθηγήτρια
Ιωάννης Σπηλιώτης	Αναπληρωτής Καθηγητής
Μιχάλης Λουλάκης	Επίκουρος Καθηγητής

Θα ήθελα να ευχαριστήσω την επιβλέπουσα καθηγήτρια κυρία Χ. Καρώνη για την πολύτιμη καθοδήγηση και στήριξη σε όλη τη διάρκεια εκπόνησης της παρούσας εργασίας.

Επιπλέον θα ήθελα να ευχαριστήσω τον αναπληρωτή καθηγητή κύριο Ι. Σπηλιώτη και τον επίκουρο καθηγητή κύριο Μ. Λουλάκη για την τιμή που μου έκαναν συμμετέχοντας στην τριμελή εξεταστική επιτροπή.

Περίληψη

Η ανάγκη της επιλογής των στατιστικά σημαντικών μεταβλητών που θα εισαχθούν τελικά σε ένα μοντέλο, ώστε να περιγράψουν με όσο το δυνατό μεγαλύτερη ακρίβεια το υπό εξέταση χαρακτηριστικό, οδήγησε στη δημιουργία μεθόδων ικανών να ανταποκριθούν στην απαίτηση αυτή. Σε συνδυασμό με την ανάγκη το προκύπτων μοντέλο να χαρακτηρίζεται και από καλή ικανότητα πρόβλεψης για μελλοντικές παρατηρήσεις, αναπτύχθηκαν κατάλληλες μέθοδοι και κριτήρια. Το κοινό χαρακτηριστικό αυτών των μεθόδων είναι ότι επιβάλλουν ένα είδος ποινής στη συνάρτηση πιθανοφάνειας με αποτέλεσμα οι συντελεστές των μεταβλητών στο μοντέλο να συρρικνώνονται. Το κοινό χαρακτηριστικό των κριτηρίων είναι ότι επιβάλλουν ποινή προκειμένου να μην εισαχθούν ανεξέλγκτα πολλές επεξηγηματικές μεταβλητές. Μέθοδοι όπως η LASSO, η Παλινδρόμηση Κορυφογραμμής, η SCAD και κριτήρια όπως AIC, BIC αναπτύσσονται στα πλαίσια της εργασίας τα οποία βρίσκουν εφαρμογή τόσο σε γενικευμένα γραμμικά μοντέλα όσο και σε δεδομένα επιβίωσης κάνοντας χρήση του μοντέλου του Cox. Σε ένα σύνολο πραγματικών δεδομένων εφαρμόζονται όλες οι μέθοδοι που παρουσιάζονται κάνοντας χρήση κατάλληλων πακέτων της R.

Abstract

The need of selecting the statistically significant variables that will finally participate in the model, so that the characteristic under consideration is described as well as possible, has led to the development of methods that are capable of meeting this requirement. Combining this need with the fact that the resulting model should have a good predictive ability, led to more efficient methods and criteria. The common characteristic of these methods is that they impose a penalty in the likelihood function causing shrinkage of the regression coefficients. The common characteristic of the criteria is that they also consider a penalty to avoid adding too many variables to the model. Methods such as LASSO, Ridge Regression, SCAD and criteria such as AIC and BIC are analyzed in the present thesis; these can be applied not only to the classic general linear model but also to generalized linear models and survival data using Cox's model. All the methods presented are applied to a real data set using programs in R.

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

Όταν αναφερόμαστε σε ένα στατιστικό μοντέλο εννοούμε συνοπτικά τη μαθηματική διατύπωση υπό μορφή εξισώσεων της σχέσης μεταξύ μεταβλητών. Περιγράφει δηλαδή τον (μαθηματικό) τρόπο με τον οποίο μία ή περισσότερες μεταβλητές (εξαρτημένες) σχετίζονται με κάποιες άλλες μεταβλητές (ανεξάρτητες). Το μοντέλο αναφέρεται ως στατιστικό αφού οι μεταβλητές συνδέονται μεταξύ τους στοχαστικά και όχι ντετερμινιστικά. Παραδείγματος χάρη, γνωρίζουμε ότι το ύψος ενός ανθρώπου εξαρτάται από την ηλικία του. Αν γνωρίζουμε την ηλικία ενός ατόμου και διαθέτουμε ένα στατιστικό μοντέλο που να συνδέει τα δύο χαρακτηριστικά, τότε μπορούμε να βρούμε την πιθανότητα να έχει κάποιο συγκεκριμένο ύψος. Στην πρόβλεψη βέβαια αυτή υπεισέρχεται σφάλμα αφού το ύψος ενός ατόμου δεν εξαρτάται αποκλειστικά από την ηλικία του. Στην περίπτωση που τα δύο χαρακτηριστικά συνδέονται γραμμικά θα μπορούσαμε να πούμε ότι η μαθηματική εξίσωση που περιγράφει τη σχέση μεταξύ των τυχαίων μεταβλητών 'ύψος' (Y) και 'ηλικία' (H) είναι $Y = \beta_0 + \beta_1 H + \varepsilon$ όπου ε είναι το σφάλμα. Φυσικά αν θεωρήσουμε την επιπλέον μεταβλητή 'φύλο' (S) τότε η σχέση $Y = \beta_0 + \beta_1 H + \beta_2 S + \varepsilon$ περιγράφει με μεγαλύτερη ακρίβεια την εξαρτημένη μεταβλητή Y. Εισάγοντας όλο και περισσότερες μεταβλητές, τόσο καλύτερα θα μπορέσουμε να εξηγήσουμε τους παράγοντες που επηρεάζουν την εξαρτημένη μεταβλητή. Εισάγοντας όμως πάρα πολλές ανεξάρτητες μεταβλητές τότε πιθανόν το μοντέλο μας να είναι δυσνόητο ή να εισάγουμε μεταβλητές που πραγματικά επηρεάζουν ελάχιστα. Στην πραγματικότητα ποτέ δεν θα μπορέσουμε να περιγράψουμε ακριβώς την εξαρτημένη μεταβλητή. Συνεπώς κανένα στατιστικό μοντέλο όσο καλό και να είναι, δε θα αντιστοιχεί στο πραγματικό. Στα πλαίσια

της παρούσας εργασίας θα παρουσιαστούν με τρόπο συνεκτικό, μέθοδοι και τεχνικές επιλογής του βέλτιστου μοντέλου. Η επιλογή αυτή δεν είναι πάντα εύκολη και δυστυχώς δεν καταλήγουν όλες οι μέθοδοι στο ίδιο μοντέλο.

Στη βιοστατιστική, επιδημιολογία, στα οικονομικά, στην κοινωνιολογία, στην ψυχολογία και άλλους τομείς, είναι σπάνιο ο αναλυτής να διαθέτει γνώση που να του επιτρέπει να προκαθορίσει ένα μοντέλο (π.χ. αν πρόκειται για ένα μοντέλο επιβίωσης από τη Weibull ή τη λογαριθμοκανονική κατανομή), ένα μετασχηματισμό για τη μεταβλητή απόκρισης ή μια δομή για το πως εμφανίζονται οι επεξηγηματικές μεταβλητές στο μοντέλο (π.χ. μετασχηματισμοί, ύπαρξη μη γραμμικών όρων, αλληλεπιδράσεις...). Στην πραγματικότητα κάποιοι επιστήμονες εκφράζουν αμφιβολίες ακόμα και για την ύπαρξη μοντέλου σε μερικές περιπτώσεις. Είμαστε, καλώς ή κακώς, υποχρεωμένοι να αναπτύξουμε ένα μοντέλο εμπειρικά στην πλειοψηφία των περιπτώσεων. Ευτυχώς, μπορούμε να ελέγξουμε την ακρίβεια ενός μοντέλου μέσω της σύγκρισης μεταξύ των παρατηρούμενων τιμών και των προβλεπόμενων τιμών και να οδηγηθούμε στο συμπέρασμα ότι το μοντέλο που διαθέτουμε είναι αξιόπιστο. Όπως αναφέρει ο Harrell (2002), ένα καλό μοντέλο είναι (α) ικανοποιητικό στην εφαρμογή του σε σχέση με τον αρχικό στόχο (β) αντιπροσωπευτικό (γ) λογικό (δ) ικανό να προσαρμόζεται εύκολα σε νέες εξωτερικές πληροφορίες ή πληροφορίες από ειδικούς και (ε) ικανό να παρέχει πληροφορία.

Η πολυπλοκότητα του μοντέλου όπως αυτή εκφράζεται από το πλήθος των συμμεταβλητών που θα χρησιμοποιηθούν εντάσσεται στα πλαίσια της επιλογής μοντέλου. Ποιες συμμεταβλητές πρέπει να συμπεριληφθούν στο μοντέλο γιατί είναι στατιστικά σημαντικές και επηρεάζουν την εξαρτημένη μεταβλητή, ποιες μπορούν να παραλειφθούν γιατί συνεισφέρουν λίγο ή καθόλου; Αυτά είναι βασικά ερωτήματα για την ανάπτυξη του μοντέλου. Πολύ σημαντικό επίσης είναι ο αναλυτής να γνωρίζει το σκοπό για τον οποίο αναπτύσσεται το μοντέλο. Δηλαδή η επιλογή μοντέλου διαφοροποιείται αν ο σκοπός είναι η πρόβλεψη ή η εκτίμηση ή ο έλεγχος υποθέσεων. Συνδυάζοντας τους δύο παραπάνω προβληματισμούς, υπάρχει ένα τεχνικό θέμα το οποίο αφορά στην ποιότητα μελλοντικών προβλέψεων. Είναι γνωστό ότι η πρόσθεση μιας οποιασδήποτε μεταβλητής σε ένα γραμμικό μοντέλο βελτιώνει ποσοτικά, έστω και οριακά, την προσαρμογή του στο συγκεκριμένο σύνολο δεδομένων. Ωστόσο δεν μπορούμε να πούμε το ίδιο και για τις μελλοντικές προβλέψεις χρησιμοποιώντας αυτό το εκτιμηθέν μοντέλο. Η παρουσία μη σημαντικών μεταβλητών στο μοντέλο μπορεί να μειώσει την ακρίβεια μελλον-

τικής πρόβλεψης. Βέβαια, δεν υπάρχει κανένας ιδιαίτερος λόγος να επιλέξουμε ένα μόνο βέλτιστο μοντέλο σύμφωνα με κάποιο κριτήριο. Έχει περισσότερο νόημα να απορρίψουμε μοντέλα τα οποία είναι προφανώς κακά και να επικεντρώσουμε την προσοχή μας στις πληροφορίες που μπορούμε να πάρουμε από τα άλλα. Φυσικά πρέπει να έχουμε στο νου μας ότι κανένα μοντέλο δεν είναι το σωστό και το μόνο που μπορούμε να περιγράψουμε είναι τα κύρια χαρακτηριστικά ενός φαινομένου.

Το ερώτημα της επιλογής μοντέλου απασχολεί ιδιαίτερα τους επιστήμονες. Ένας μελετητής συλλέγει στοιχεία, συχνά υπό μορφή μετρήσεων για πολλές διαφορετικές πτυχές των παρατηρούμενων μονάδων και θέλει να μελετήσει πώς οι μεταβλητές αυτές επηρεάζουν το χαρακτηριστικό που τον ενδιαφέρει. Λόγω της σημαντικότητας του θέματος, δεν αποτελεί έκπληξη το γεγονός ότι το ερώτημα αυτό έχει προσεγγιστεί από πολλούς ερευνητές. Τόσο η κλασική όσο και η Μπεϋζιανή σχολή καταπιάστηκε με το θέμα προτείνοντας μεθόδους όπως ο έλεγχος F για εμφωλευμένα μοντέλα, AIC, BIC, Mallow's C_p προς τα εμπρός και προς τα πίσω διαδικασία επιλογής, cvl, κ.α. να είναι μερικές μόνο από τις πιο γνωστές μεθόδους τις οποίες θα δούμε στη συνέχεια. Μερικές από αυτές αποτελούν μεθόδους επιλογής ή/και λειτουργούν ως μέθοδοι σύγκρισης μεταξύ μοντέλων ή κριτήρια για να κριθεί η ποιότητα ενός μοντέλου.

Δεδομένου αυτού του πλούτου επιλογών, πώς θα αποφασίσει ο στατιστικός τι πρέπει να κάνει; Χρειαζόμαστε μια προσέγγιση η οποία θα μπορεί να εκτελεσθεί εύκολα και να δώσει αποτελέσματα τα οποία να μπορούν να ερμηνευθούν και να γίνουν κατανοητά. Από στατιστικής άποψης, θέλουμε μία μέθοδο που να είναι συνεκτική και αρκετά γενική ώστε να βρίσκει εφαρμογή σε ευρεία γκάμα προβλημάτων (Kadane and Lazar, 2004). Είναι σημαντικό λοιπόν να διαθέτουμε μεθόδους ώστε να εντοπίζουμε το μικρότερο υποσύνολο μεταβλητών από το αρχικό μας σύνολο, για την προσαρμογή ενός μοντέλου το οποίο θα επεξηγεί τη συμπεριφορά της μεταβλητής απόκρισης σε ικανοποιητικό βαθμό.

Στην παρούσα εργασία θα ασχοληθούμε με την παρουσίαση και στη συνέχεια εφαρμογή σε πραγματικά δεδομένα, μεθόδων ή κριτηρίων τα οποία οδηγούν στην επιλογή του καταλληλότερου μοντέλου. Θα δούμε παραδοσιακές τεχνικές αλλά και πιο νέες που πλέον εφαρμόζονται ευρέως και μπορούν να εφαρμοστούν και σε δεδομένα που αφορούν μοντέλα επιβίωσης. Το κοινό χαρακτηριστικό των περισσότερων από αυτές τις μεθόδους είναι ότι προκειμένου να

αποφευχθεί η υπερπροσαρμογή, δηλαδή η εισαγωγή πολλών μεταβλητών που δυσχεραίνουν την κατανόηση του μοντέλου, επιβάλλουν ένα είδος ποινής η οποία εξαρτάται είτε από το πλήθος των εισαγώμενων μεταβλητών είτε από τους συντελεστές τους. Προκαλούν έτσι την συρρίκνωση μερικών ή την έξοδο άλλων από το μοντέλο. Είναι σημαντικό να τονίσουμε ότι ανάλογα με το σκοπό για τον οποίο γίνεται μια ανάλυση (έλεγχο υποθέσεων, εκτίμηση, πρόβλεψη) κάποια μέθοδος μπορεί να υπερτερεί έναντι κάποιας άλλης. Τα κριτήρια και οι μέθοδοι που παρουσιάζονται στη συνέχεια έχουν ως κύριο στόχο να διευκολύνουν τον αναλυτή να αναπτύξει μοντέλα τα οποία θα κάνουν ακριβείς προβλέψεις της μεταβλητής απόκρισης σε μελλοντικές παρατηρήσεις.

Η εργασία είναι δομημένη ως ακολούθως. Στην παράγραφο 1.1 παραθέτουμε τα βασικά στοιχεία του γενικού γραμμικού μοντέλου και του μοντέλου της λογιστικής παλινδρόμησης. Ιδιαίτερη μνεία θα γίνει στο μοντέλο του Cox το οποίο αναπτύσσεται στην παράγραφο 1.2 καθώς αποτελεί βασικό εργαλείο στην ανάλυση επιβίωσης. Για το λόγο αυτό υπενθυμίζουμε για το μοντέλο του Cox τις βασικές του ιδιότητες. Κατόπιν η παράγραφος 1.3 αναφέρει τα πρώτα απλά κριτήρια με ποινή για την επιλογή του βέλτιστου μοντέλου ενώ στην επόμενη παράγραφο 1.4 αναπτύσσονται κλασικές τεχνικές επιλογής μοντέλου όπως ο F έλεγχος, ο έλεγχος του λόγου πιθανοφανειών κ.ά. Στο επόμενο κεφάλαιο περνάμε πλέον στην παρουσίαση των πιο σύγχρονων μεθόδων επιλογής του μοντέλου που θεωρείται βέλτιστο ως προς την ικανότητα πρόβλεψης για μελλοντικές παρατηρήσεις των οποίων το κοινό χαρακτηριστικό είναι η επιβολή ποινής προκειμένου να αποφευχθεί η υπερπροσαρμογή. Στην πρώτη παράγραφο του δεύτερου κεφαλαίου 2.1 αναφέρεται ένας έλεγχος ο οποίος πρέπει να γίνεται πάντα πριν ξεκινήσουμε την ανάλυση στα δεδομένα μας και είναι γνωστό ως global test. Κατόπιν, στην παράγραφο 2.2 αναπτύσσεται η μέθοδος Ridge ή L_2, στη συνέχεια στην 2.3 η μέθοδος LASSO ή L_1 ενώ στην παράγραφο 2.4 αναλύεται η μέθοδος LASSO όπως αυτή εφαρμόζεται σε δεδομένα επιβίωσης στο μοντέλο του Cox. Γενικεύσεις της LASSO παρουσιάζονται στην παράγραφο 2.5 και η μέθοδος SIS η οποία είναι ιδιαίτερα χρήσιμη σε περίπτωση μεγάλου πλήθος ανεξάρτητων μεταβλητών περιγράφεται στην παράγραφο 2.6. Τέλος με την αξιολόγηση ενός μοντέλου μέσω του crossvalidation στην παράγραφο 2.7 ολοκληρώνεται το δεύτερο κεφάλαιο. Στο τρίτο και τελευταίο κεφάλαιο εφαρμόζονται οι μέθοδοι που περιγράφονται στα πρώτα δύο κεφάλαια σε πραγματικά δεδομένα όπου αναφέρεται αναλυτικά η χρήση του

πακέτου penalized της R που απαιτείται προκειμένου να πραγματοποιηθεί η ανάλυση.

1.1 Γενικό και γενικευμένο γραμμικό μοντέλο

Το γενικό γραμμικό μοντέλο αποτελείται από μια εξαρτημένη μεταβλητή ή μεταβλητή απόκρισης και τουλάχιστον μία ανεξάρτητη ή επεξηγηματική μεταβλητή. Έστω ότι διαθέτουμε n παρατηρήσεις και k μεταβλητές. Η εξαρτημένη μεταβλητή συνδέεται γραμμικά με τις ανεξάρτητες μέσω της συνάρτησης παλινδρόμησης

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \tag{1.1}$$

$y_i,\ i = 1, \ldots, n$, οι τιμές των παρατηρήσεων της μεταβλητής απόκρισης, το διάνυσμα $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$ είναι οι συντελεστές παλινδρόμησης που αποτελούν τις άγνωστες παραμέτρους του μοντέλου που πρέπει να εκτιμηθούν, $x_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, k$, είναι η τιμή της $i-$οστής παρατήρησης της μεταβλητής X_j και $\varepsilon_i,\ i = 1, \ldots, n$ τα τυχαία σφάλματα. Τα σφάλματα αυτά υποθέτουμε ότι ακολουθούν κάποια κατανομή με μέση τιμή 0 και διασπορά σ^2 για κάθε i και είναι μεταξύ τους ασυσχέτιστα, δηλαδή $Cov(\varepsilon_i, \varepsilon_j) = 0$. Αν θεωρήσουμε τον πίνακα

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

ο οποίος καλείται πίνακας σχεδιασμού. Η (1.1) υπό μορφή πινάκων γράφεται ως:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{1.2}$$

όπου $\mathbf{y}$ και ε τα διανύσματα $(y_1, y_2, \ldots, y_n)$ και $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ των τιμών της μεταβλητής Y και των τυχαίων σφαλμάτων αντίστοιχα. Αν επιπλέον υποθέσουμε ότι τα σφάλματα ακολουθούν Κανονική κατανομή, τότε οι υποθέσεις που αναφέραμε ότι πρέπει να ικανοποιούν αυτά, εδώ συνοψίζονται στην απαίτηση το διάνυσμα ε να ακολουθεί την πολυδιάστατη Κανονική κατανομή $\varepsilon \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$.

Η εκτίμηση των παραμέτρων $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$ είτε με τη μέθοδο των ελαχίστων τετραγώνων ανεξαρτήτως κατανομής είτε, στην περίπτωση Κανονικής κατανομής των σφαλμάτων

με τη μέθοδο της μέγιστης πιθανοφάνειας, καταλήγει στις ίδιες εκτιμήτριες οι οποίες έχουν τη μορφή

$$\hat{\beta} = (\mathbf{X'X})^{-1}\mathbf{X'y} \tag{1.3}$$

Οι εκτιμήτριες αυτές αναφέρονται ως συνήθεις εκτιμήτριες ελαχίστων τετραγώνων (ε.ε.τ.). Είναι αμερόληπτες και μάλιστα είναι BLUE δηλαδή μεταξύ όλων των αμερόληπτων γραμμικών εκτιμητριών της παραμέτρου β είναι εκείνες με την ελάχιστη διασπορά (Θεώρημα Gauss-Markov). Επιπρόσθετα, με τη μέθοδο της μέγιστης πιθανοφάνειας μπορεί να υπολογιστεί και η εκτιμήτρια μέγιστης πιθανοφάνειας της διασποράς σ^2 η οποία είναι

$$\hat{\sigma}^2 = \frac{\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)}{n} \tag{1.4}$$

και επειδή

$$SSE = \left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right) \tag{1.5}$$

η (1.4) γράφεται

$$\hat{\sigma}^2 = \frac{SSE}{n} \tag{1.6}$$

Η εκτιμήτρια $\hat{\sigma}^2$ δεν είναι αμερόληπτη. Για το λόγο αυτό, όταν η παράμετρος σ^2 είναι άγνωστη, τότε αυτή αντικαθίσταται από την αμερόληπτη εκτιμήτριά της

$$S^2 = MSE = SSE/(n-p), \tag{1.7}$$

όπου $p = k + 1$ και SSE δίνεται από την (1.5).

Κάτω από την υπόθεση της κανονικότητας, η εκτιμήτρια β_j ακολουθεί την Κανονική κατανομή. Πιο συγκεκριμένα

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 c_{jj}),\ j = 0, 1, 2, \ldots, k, \tag{1.8}$$

με c_{jj} το $j-$οστό διαγώνιο στοιχείο του πίνακα $C = (X'X)^{-1}$. Η εκτίμηση της διασποράς των εκτιμητριών προκύπτει κάνοντας χρήση της αμερόληπτης εκτιμήτριας S^2. Έτσι προκύπτει ο πίνακας διακύμανσης - συνδιακύμανσης

$$\hat{V}(\hat{\beta}) = S^2 C.$$

Συνεπώς το τυπικό σφάλμα της εκτιμήτριας $\hat{\beta}_j$ δίνεται ως το j-οστό διαγώνιο στοιχείο του πίνακα $\hat{V}(\hat{\beta})$ δηλαδή είναι το στοιχείο $se(\hat{\beta}_j) = S\sqrt{c_{jj}}$.

Από τις ιδιότητες που έχουν αναφερθεί μέχρι τώρα για το γραμμικό μοντέλο προκύπτει άμεσα ότι

$$E[\mathbf{y}] = X\beta. \tag{1.9}$$

Το $X\beta$ αναφέρεται και ως συστηματικό ή μη στοχαστικό μέρος του μοντέλου. Το $\mu = E[\mathbf{y}]$ είναι το διάνυσμα των μέσων και αποτελεί το συστηματικό ή στοχαστικό μέρος του μοντέλου. Μέχρι τώρα γνωρίζουμε ότι το τυχαίο διάνυσμα $\mathbf{y}$ ακολουθεί την Κανονική κατανομή. Στην περίπτωση όμως που η μεταβλητή απόκρισης δεν ακολουθεί κανονική κατανομή αλλά κάποια άλλη συνεχή ή διακριτή κατανομή (π.χ. Poisson, Διωνυμική) το κλασικό γραμμικό μοντέλο δεν μπορεί να εφαρμοστεί στη γνωστή μορφή. Μπορεί όμως να γενικευθεί χρησιμοποιώντας μια κατάλληλη συνάρτηση σύνδεσης. Παρόλ' αυτά τα γενικευμένα γραμμικά μοντέλα είναι εγγενώς παραμετρικά με την έννοια ότι η συνάρτηση πιθανοφάνειας καθορίζεται πλήρως από τον ερευνητή (Gill, 2001). Ας υποθέσουμε λοιπόν ότι γενικεύουμε την (1.9) θεωρώντας τώρα τη μορφή

$$g(\mu) = \mathbf{X}\beta, \tag{1.10}$$

όπου $g()$ είναι μια αντιστρέψιμη, λεία (smooth) συνάρτηση του διανύσματος των μέσων $\boldsymbol{\mu}$. Η πληροφορία για τη μεταβλητή απόκρισης που λαμβάνεται από τις επεξηγηματικές μεταβλητές μέσω της γραμμικής δομής $\mathbf{X}\beta$ δεν παρέχεται άμεσα αλλά ελέγχεται από τη μορφή της συνάρτησης σύνδεσης $g()$ (link function). Αυτή η συνάρτηση συνδέει το $g(\mu)$ με το μέσο της μεταβλητής απόκρισης και όχι απευθείας με το αποτέλεσμα της μεταβλητής απόκρισης όπως συμβαίνει στην περίπτωση του κλασικού γραμμικού μοντέλου. Με αυτόν τον τρόπο, το γενικευμένο γραμμικό μοντέλο επεκτείνει το γραμμικό μοντέλο ώστε να συμπεριληφθεί και η περίπτωση μοντέλων με μη κανονικές μεταβλητές απόκρισης που μπορούν να γίνουν γραμμικά μέσω κατάλληλων μετασχηματισμών.

Έτσι λοιπόν το γενικευμένο γραμμικό μοντέλο αποτελείται από τρία μέρη.

i. Το στοχαστικό μέρος y το οποίο αποτελείται από ανεξάρτητες και ισόνομες τυχαίες μεταβλητές από κάποια κατανομή που ανήκει στην Εκθετική Οικογένεια Κατανομών με μέσο $\boldsymbol{\mu}$.

ii. Το συστηματικό μέρος $\theta = \mathbf{X}\beta$. Συνεπώς οι επεξηγηματικές μεταβλητές, X, επηρεάζουν το εξαγόμενο της παρατηρούμενης μεταβλητής μόνο μέσω της μορφής της

συνάρτησης $g()$.

iii. Τη συνάρτησης σύνδεσης η οποία αποτελεί τη συνάρτηση μέσω της οποίας συνδέονται το στοχαστικό και το συστηματικό μέρος του μοντέλου μέσω της σχέσης

$$g(\mu) = \theta \Rightarrow g(\mu) = X\beta \Rightarrow \mu = g^{-1}(X\beta) = E[y].$$

Οι συναρτήσεις σύνδεσης των κυριότερων κατανομών φαίνονται στον πίνακα 1.1.

Κατανομή		$\theta = g(\mu)$	$\mu = g^{-1}(\theta)$
Poisson		$\log(\mu)$	$\exp(\theta)$
Διωνυμική	logit	$\log\left(\frac{\mu}{1-\mu}\right)$	$\frac{\exp(\theta)}{1+\exp(\theta)}$
	probit	$\Phi^{-1}(\mu)$	$\Phi(\theta)$
	$c\log\log$	$\log(-\log(1-\mu))$	$1-\exp(-\exp(\theta))$
Κανονική		μ	θ
Γάμμα		$-\frac{1}{\mu}$	$-\frac{1}{\theta}$
Αρνητική Διωνυμική		$\log(1-\mu)$	$1-\exp(\theta)$

Πίνακας 1.1: Συναρτήσεις σύνδεσης βασικών κατανομών.

Μια πολύ σημαντική ειδική περίπτωση των γενικευμένων γραμμικών μοντέλων είναι η λογιστική παλινδρόμηση. Χρησιμοποιείται όταν η μεταβλητή απόκρισης είναι διακριτή με δύο μόνο δυνατά αποτελέσματα. Τέτοια δεδομένα εμφανίζονται για παράδειγμα σε ιατρικά πειράματα στα οποία στο τέλος κάθε πειράματος ο ασθενής είτε ανένηψε είτε κατέληξε, ένα προϊόν περνάει τον έλεγχο ποιότητας ή αποτυγχάνει, ένας υπάλληλος παίρνει προαγωγή ή όχι και πολλά άλλα παραδείγματα. Τα δύο δυνατά ενδεχόμενα κωδικοποιούνται με αντίστοιχες τιμές της μεταβλητής. Συνήθως χρησιμοποιούνται οι τιμές 1 για το ενδεχόμενο που μας ενδιαφέρει και αναφερόμαστε σε αυτό ως επιτυχία με πιθανότητα p και η τιμή 0 για την αποτυχία με πιθανότητα $q = 1 - p$. Μια τέτοια μεταβλητή ακολουθεί κατανομή Bernoulli με παράμετρο p. Οι επεξηγηματικές μεταβλητές μπορούν να είναι είτε διακριτές είτε συνεχείς ενώ δεν τίθεται καμία ιδιαίτερη προϋπόθεση την οποία πρέπει να πληρούν π.χ. όσον αφορά την κατανομή τους ή την ανεξαρτησία τους ούτε πρέπει να έχουν ίσες διασπορές. Η λογιστική παλινδρόμηση μπορεί να χρησιμοποιηθεί και για την πρόβλεψη αν κάποιο αντικείμενο έρευνας

πιθανολογείται να καταταχθεί σε κάποια από τις δύο κατηγορίες της μεταβλητής απόκρισης. Επιπλέον με τη μέθοδο αυτή παρέχεται γνώση για την ύπαρξη σχέσης καθώς και την ένταση της σχέσης μεταξύ μεταβλητών (π.χ. αν κάποιο άτομο καπνίζει 10 πακέτα τσιγάρα την ημέρα τότε αυτός κατατάσσεται σε ομάδα υψηλότερου κινδύνου να αναπτύξει καρκίνο των πνευμόνων σε σχέση με κάποιον που δουλεύει σε ορυχείο).

Αν y είναι ο αριθμός των επιτυχιών στις n δοκιμές, τότε η τυχαία μεταβλητή y ακολουθεί διωνυμική κατανομή με παραμέτρους n και p. Η εξάρτηση της y από τις συμμεταβλητές εισάγεται μέσω της εξάρτησης της πιθανότητας επιτυχίας p από αυτές. Η συνάρτηση σύνδεσης που χρησιμοποιείται είναι συνήθως η logit. Συνεπώς έχουμε σύμφωνα με τον παραπάνω συμβολισμό $\theta = X\beta$ και $logit(p) = \ln\left(\frac{p}{1-p}\right)$. Τελικά το μοντέλο της λογιστικής παλινδρόμησης γράφεται στη μορφή

$$\ln\left(\frac{p}{1-p}\right) = X\beta. \tag{1.11}$$

Για την i-οστή παρατήρηση το μοντέλο γράφεται ως

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = \mathbf{x_i'}\beta,\ i = 1, 2, \ldots, n. \tag{1.12}$$

Αντιστρέφοντας τη συνάρτηση σύνδεσης παίρνουμε

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})} = \frac{\exp(\mathbf{x_i'}\beta)}{1 + \exp(\mathbf{x_i'}\beta)}. \tag{1.13}$$

Ο λόγος $\frac{p}{1-p}$ της πιθανότητας επιτυχίας p προς αυτήν της αποτυχίας $1-p$, καλείται λόγος των συμπληρωματικών ή σχετικών πιθανοτήτων (odds) και παίζει σημαντικό ρόλο στην λογιστική παλινδρόμηση και ιδιαίτερα στην ερμηνεία των συντελεστών της παλινδρόμησης. Όταν λέμε ότι τα odds είναι 2 εννοούμε ότι η πιθανότητα επιτυχίας είναι διπλάσια της πιθανότητας αποτυχίας. Σχετικά με την ερμηνεία των συντελεστών της παλινδρόμησης, κρατώντας όλες τις συμμεταβλητές σταθερές και αυξάνοντας μόνο μια, έστω τη x_j, κατά μία μονάδα, ο λογάριθμος των odds μεταβάλλεται κατά β_j, δηλαδή το συντελεστή της x_j στο μοντέλο. Ισοδύναμα, η ποσότητα e^{β_j} είναι ο παράγοντας με τον οποίο πολλαπλασιάζεται η σχετική πιθανότητα πραγματοποίησης της επιτυχίας. Αν ο συντελεστής β_j είναι θετικός, τότε το odds αυξάνεται καθώς η x_j αυξάνει ενώ αν είναι αρνητικός, τότε η σχετική πιθανότητα μειώνεται με αύξηση της x_j.

Η εκτίμηση των παραμέτρων σε ένα μοντέλο λογιστικής παλινδρόμησης (Καρώνη και Οικονόμου, 2010) γίνεται με τη μέθοδο της μέγιστης πιθανοφάνειας. Η πιθανοφάνεια γράφεται

στη μορφή

$$L(\beta) = \prod_{i=1}^{n} \begin{pmatrix} n_i \\ y_i \end{pmatrix} p_i^{y_i} (1-p_i)^{n_i - y_i}, \tag{1.14}$$

όπου n_i ο αριθμός των δοκιμών της στατιστικής μονάδας i και p_i η αντίστοιχη πιθανότητα επιτυχίας. Η παραπάνω συνάρτηση πιθανοφάνειας εξαρτάται από τις παραμέτρους β μέσω της σχέσης (1.13). Η αντίστοιχη λογαριθμοποιημένη συνάρτηση πιθανοφάνειας δίνεται ως

$$\ell(\beta) = \sum_{i=1}^{n} \left\{ \ln \begin{pmatrix} n_i \\ y_i \end{pmatrix} + y_i \ln p_i + (n_i - y_i) \ln(1-p_i) \right\}. \tag{1.15}$$

Χρησιμοποιώντας τις (1.12) και (1.13), η τελευταία σχέση γράφεται τελικά στη μορφή

$$\ell(\beta) = \sum_{i=1}^{n} \left\{ \ln \begin{pmatrix} n_i \\ y_i \end{pmatrix} + y_i \mathbf{x_i'} \beta - n_i \ln(1 + \exp(\mathbf{x_i'}\beta)) \right\}. \tag{1.16}$$

Παραγωγίζοντας ως προς β_j και θέτοντας τις μερικές παραγώγους ίσες με μηδέν παίρνουμε ένα σύστημα από $p = k+1$ εξισώσεις με άγνωστες τις p παραμέτρους τους μοντέλου, το οποίο λύνεται με επαναληπτικές μεθόδους. Η γενική εξίσωση που πρέπει να ικανοποιούν οι εκτιμήτριες μέγιστης πιθανοφάνειας των β_j όπως αυτή προκύπτει από τις μερικές παραγώγους είναι

$$\mathbf{X}'(\mathbf{y} - \hat{\mu}) = \mathbf{0}, \tag{1.17}$$

όπου $\hat{\mu}_j = n_j \hat{p}_j$.

1.2 Μοντέλο Cox

Το μοντέλο του Cox αποτελεί έναν τρόπο μοντελοποίησης για την ανάλυση δεδομένων επιβίωσης. Ο σκοπός του μοντέλου είναι να διερευνήσει ταυτόχρονα τις επιδράσεις διαφόρων μεταβλητών στην επιβίωση. Αποτελεί μια πολύ αναγνωρισμένη τεχνική στην ανάλυση επιβίωσης. Όταν χρησιμοποιείται για την ανάλυση της επιβίωσης ασθενών σε μια κλινική δοκιμή, το μοντέλο μας επιτρέπει να απομονώσουμε τις επιδράσεις της θεραπείας από τα αποτελέσματα των άλλων μεταβλητών. Μπορεί επίσης να χρησιμοποιηθεί a priori αν είναι γνωστό ότι υπάρχουν και άλλες μεταβλητές εκτός της θεραπείας που επηρεάζουν την επιβίωση του ασθενούς και αυτές οι μεταβλητές δεν μπορούν εύκολα να ελεγχθούν σε μια κλινική δοκιμή.

Το μοντέλο του Cox ανήκει στην κατηγορία των μοντέλων αναλογικής διακινδύνευσης (Καρώνη, 2009). Αυτά ορίζονται από την έκφραση:

$$h(t;\mathbf{x}) = g(\mathbf{x})h_0(t) \tag{1.18}$$

όπου $h_0(t)$ είναι μια βασική συνάρτηση διακινδύνευσης και η συνάρτηση $g(\mathbf{x})$ είναι θετική. Συνήθως χρησιμοποιείται η $g(\mathbf{x}) = e^{\beta'\mathbf{x}}$. Θεωρείται λοιπόν ότι οι συμμεταβλητές $\mathbf{x}$ επιδρούν στη συνάρτηση διακινδύνευσης μέσω της σχέσης

$$h(t;\mathbf{x}) = h_0(t)e^{\beta'\mathbf{x}}, \tag{1.19}$$

όπου $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ ένα διάνυσμα p συντελεστών, οι οποίοι εκφράζουν ποσοτικά την επίδραση της καθεμιάς των συμμεταβλητών $\mathbf{x}$. Σε ένα μοντέλο αναλογικής διακινδύνευσης:

- Θετικοί συντελεστές σημαίνει ότι ο ρυθμός κινδύνου είναι αύξων, συνεπώς ο χρόνος επιβίωσης συντομεύεται.

- Αρνητικοί συντελεστές σημαίνει ότι ο ρυθμός κινδύνου είναι φθίνων, συνεπώς ο χρόνος επιβίωσης επιμηκύνεται.

Τα μοντέλα που ανήκουν στην κατηγορία αυτή διακρίνονται σε δύο επιμέρους κατηγορίες ανάλογα με το αν η βασική συνάρτηση διακινδύνευσης $h_0(t)$ προέρχεται από κάποια συγκεκριμένη κατανομή (παραμετρικά μοντέλα) ή αν η $h_0(t)$ δεν προσδιορίζεται και παραμένει ακαθόριστη και αυτό οδηγεί σε ένα ημι-παραμετρικό μοντέλο. Το μοντέλο του Cox ανήκει στη δεύτερη κατηγορία, δηλαδή αποτελεί ένα ημι-παραμετρικό μοντέλο αναλογικής διακινδύνευσης. Η ανεξαρτησία της διακινδύνευσης και κατά συνέπεια της επιβίωσης από τη συμμεταβλητή x_i σημαίνει ότι $\beta_i = 0$. Λαμβάνοντας υπόψη ότι η σωρευτική συνάρτηση διακινδύνευσης, $H(t)$, ορίζεται ως

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

έχουμε ότι

$$H(t;\mathbf{x}) = H_0(t)e^{\beta'\mathbf{x}}.$$

Συνεπώς, με βάση τη σχέση $S(t) = \exp\{-H(t)\}$, όπου $S(t)$ η συνάρτηση επιβίωσης, προκύπτει

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

Το κύριο χαρακτηριστικό του μοντέλου του Cox είναι ότι, ως ημι-παραμετρικό μοντέλο, οι συγκεκριμένες παραμετρικές μορφές των βασικών συναρτήσεων $H_0(t)$ και $S_0(t)$ δεν καθορίζονται. Μόνο η επίδραση των συμμεταβλητών $\mathbf{x}$ αναλύεται.

Εκτίμηση παραμέτρων στο μοντέλο του Cox

Έστω ότι έχουμε n παρατηρήσεις (t_i, d_i, x_i), όπου t_i είναι ο πιθανόν αποκομμένος χρόνος επιβίωσης, d_i είναι η δείκτρια αποκοπής, δηλαδή ο δείκτης d_i λαμβάνει τιμές 0 ή 1 ανάλογα με το αν η παρατήρηση της μονάδας i είναι αποκομμένη ή όχι αντίστοιχα και x_i είναι ένα διάνυσμα στήλη των συμμεταβλητών για τη μονάδα i. Στο μοντέλο αναλογικής διακινδύνευσης του Cox η συνάρτηση κινδύνου για τη μονάδα i δίνεται ως

$$h_i(t; x_i) = h_0(t) exp(\beta' x_i) \tag{1.20}$$

όπου $exp(\beta' x_i)$ είναι ο σχετικός κίνδυνος ή ποσοστό κινδύνου (hazard ratio).

Έστω ότι διακόπτεται η λειτουργία k μονάδων κατά τις διακεκριμένες χρονικές στιγμές

$$t_{(1)} < t_{(2)} < \cdots < t_{(k)}.$$

Κατά τη χρονική στιγμή $t_{(j)}$ διακόπτεται η λειτουργία μιας μονάδας με συμμεταβλητές $\mathbf{x}_j$ και υποθέτουμε ότι $d_j = 1,\ j = 1, 2, \ldots, k$. Έστω R_j το σύνολο των μονάδων που βρίσκονται σε κίνδυνο αμέσως πριν τη χρονική στιγμή $t_{(j)}$. Το σύνολο R_j αποτελείται δηλαδή από τις μονάδες οι οποίες ούτε έχουν αποτύχει ούτε έχουν αποκοπεί μέχρι τη στιγμή $t_{(j)}$. Συνεπώς οι μονάδες που περιέχονται στο σύνολο αυτό έχουν χρόνο αποτυχίας ή αποκοπής μεγαλύτερο του $t_{(j)}$. Λαμβάνοντας υπόψη ότι η $h(t)dt$ εκφράζει τη στιγμιαία πιθανότητα διακοπής, η πιθανότητα να διακοπεί η λειτουργία μιας συγκεκριμένης μονάδας j, δεδομένου ότι παύει να λειτουργεί μια οποιαδήποτε μονάδα του συνόλου R_j είναι

$$P(\text{μονάδα } j \text{ αποτυγχάνει τη στιγμή } t_{(j)} | R_j \text{ μία αποτυχία τη στιγμή} t_{(j)}) =$$
$$\frac{P(\text{μονάδα } j \text{ αποτυγχάνει τη στιγμή } t_{(j)} | R_j)}{P(\text{ μία αποτυχία τη στιγμή } t_{(j)} | R_j)} = \frac{h(t_{(j)}; \mathbf{x}_j)}{\sum_{i \in R_j} h(t_{(j)}; \mathbf{x}_i)}.$$

Η τελευταία σχέση λόγω της (1.20) παίρνει τη μορφή

$$\frac{h_0(t) e^{\beta' \mathbf{x}_j}}{\sum_{i \in R_j} h_0(t) e^{\beta' \mathbf{x}_i}} = \frac{e^{\beta' \mathbf{x}_j}}{\sum_{i \in R_j} e^{\beta' \mathbf{x}_i}}$$

η οποία είναι ανεξάρτητη της $h_0(t)$. Για να κατανοήσουμε καλύτερα αυτή την πιθανοφάνεια, ας θεωρήσουμε την ειδική περίπτωση που οι συμμεταβλητές δεν επιδρούν καθόλου, δηλαδή το διάνυσμα β είναι μηδενικό. Τότε $\exp(\beta'\mathbf{x}_j) = \exp(\beta'\mathbf{x}_i) = 1$ και η πιθανότητα $P(\text{μονάδα } j \text{ αποτυγχάνει τη στιγμή } t_{(j)}|R_j \text{ μία αποτυχία τη στιγμή} t_{(j)})$ είναι ίση με $\frac{1}{n_j}$ όπου n_j είναι το πλήθος των μονάδων σε κίνδυνο τη στιγμή $t_{(j)}$ (Harrell, 2002). Αυτές οι δεσμευμένες πιθανότητες είναι μεταξύ τους ανεξάρτητες για τους διάφορους χρόνους αποτυχίας. Συνεπώς μπορεί να υπολογιστεί μια συνολική πιθανοφάνεια πολλαπλασιάζοντας όλες αυτές τις πιθανότητες πάνω σε όλους τους χρόνους αποτυχίας. Ο Cox το όρισε αυτό ως *μερική πιθανοφάνεια* για το διάνυσμα των παραμέτρων β:

$$L(\beta) = \prod_{j=1}^{k} \left\{ \frac{e^{\beta'\mathbf{x}_j}}{\sum_{i \in R_j} e^{\beta'\mathbf{x}_i}} \right\}. \tag{1.21}$$

Μπορούμε να χειριστούμε την παραπάνω συνάρτηση μερικής πιθανοφάνειας ως μια συνηθισμένη συνάρτηση πιθανοφάνειας προκειμένου να υπολογιστεί η εκτιμήτρια μέγιστης πιθανοφάνειας $\hat{\beta}$ της β ακολουθώντας την κλασική διαδικασία. Ο λογάριθμος της πιθανοφάνειας (λογαριθμοποιημένη πιθανοφάνεια) είναι

$$\ell(\beta) = \sum_{j=1}^{k} \beta'\mathbf{x}_j - \sum_{j=1}^{k} \ln \left\{ \sum_{i \in R_j} e^{\beta'\mathbf{x}_i} \right\}. \tag{1.22}$$

Παραγωγίζοντας ως προς κάθε μεταβλητή $\beta_r,\ r = 1, 2, \ldots, p$ παίρνουμε

$$\frac{\partial \ell}{\partial \beta_r} = \sum_{j=1}^{k} x_{jr} - \sum_{j=1}^{k} \left\{ \frac{\sum_{i \in R_j} x_{ir}\, e^{\beta'\mathbf{x}_i}}{\sum_{i \in R_j} e^{\beta'\mathbf{x}_i}} \right\}. \tag{1.23}$$

Λύνοντας το σύστημα των εξισώσεων

$$\frac{\partial \ell}{\partial \beta_r} = 0, \quad r = 1, 2, \ldots, p \tag{1.24}$$

ως προς $\beta_1, \beta_2, \ldots, \beta_r$ προκύπτουν οι ζητούμενες εκτιμήτριες.

Εκτιμήσεις των διασπορών των εκτιμήσεων $\hat{\beta}$ προσδιορίζονται από τον αντίστροφο του πίνακα παρατηρούμενης πληροφορίας, με (r, s) στοιχείο $-\frac{\partial^2 \ell}{\partial \beta_r \partial \beta_s}|_{\hat{\beta}}$ όπου

$$-\frac{\partial^2 \ell}{\partial \beta_r \partial \beta_s} = \sum_{j=1}^{k} \sum_{i \in R_j} x_{ir} \left(x_{is} - \frac{\sum_{m \in R_j} x_{ms}\, e^{\beta'\mathbf{x}_m}}{\sum_{m \in R_j} e^{\beta'\mathbf{x}_m}} \right) \frac{e^{\beta'\mathbf{x}_i}}{\sum_{m \in R_j} e^{\beta'\mathbf{x}_m}}.$$

Στην περίπτωση που υπάρχουν ισόπαλοι χρόνοι διακοπής έχουν προταθεί διάφορες μέθοδοι για τον χειρισμό τους. Συνήθως προτιμάται η απλή προσέγγιση του Breslow (1974). Στην περίπτωση αυτή, υπάρχουν $d_j > 1$ διακοπές που συμπίπτουν τη χρονική στιγμή $t_{(j)}$ (οι οποίες θεωρητικά θα προέκυπταν σε διαφορετικούς χρόνους αν οι μετρήσεις ήταν μεγαλύτερης ακρίβειας) και ο όρος

$$\frac{e^{\beta' \mathbf{x}_j}}{\sum_{i \in R_j} e^{\beta' \mathbf{x}_i}}$$

της μερικής πιθανοφάνειας (1.21) για $d_j = 1$ αντικαθίσταται από το

$$\frac{e^{\beta' \mathbf{z}_j}}{\left\{\sum_{i \in R_j} e^{\beta' \mathbf{x}_i}\right\}^{d_j}} \tag{1.25}$$

όπου $\mathbf{z}_j = \sum_{k=1}^{d_j} \mathbf{x}_k$ και $\mathbf{x}_k$ το διάνυσμα των συμμεταβλητών της μονάδας k, με διακοπή τη στιγμή $t_{(j)}$, $k = 1, \ldots, d_j$. Η προσέγγιση παρουσιάζει προβλήματα στην περίπτωση που σε κάποια χρονική στιγμή το ποσοστό των ίσων χρόνων διακοπής είναι μεγάλο σε σχέση με τον αριθμό των μονάδων σε κίνδυνο. Η απλή προσέγγιση του Breslow πλέον είναι λιγότερο δημοφιλής και στα στατιστικά πακέτα χρησιμοποιείται και η προσέγγιση του Efron (1977). Θεωρώντας ότι όλες οι $d_j!$ διαφορετικές σειρές πραγματοποίησης των διακοπών είναι εξίσου πιθανές, ο όρος

$$\frac{e^{\beta' \mathbf{x}_j}}{\sum_{i \in R_j} e^{\beta' \mathbf{x}_i}}$$

της μερικής πιθανοφάνειας (1.21) για $d_j = 1$ αντικαθίσταται από το

$$\frac{e^{\beta' \mathbf{z}_j}}{\prod_{r=1}^{d_j} \left\{\sum_{i \in R_j} e^{\beta' \mathbf{x}_i} - (r-1) d_j^{-1} \sum_{i \in D_j} e^{\beta' \mathbf{x}_i}\right\}} \tag{1.26}$$

όπου D_j είναι το σύνολο των μονάδων που αποτυγχάνουν τη στιγμή $t_{(j)}$. Και η προσέγγιση του Efron έχει το ίδιο πρόβλημα με αυτό του Breslow αλλά προτείνεται περισσότερο η χρήση αυτής του Efron διότι είναι πιο γρήγορη από τις ακριβείς μεθόδους και τείνει να δώσει εκτιμήσεις πιο κοντά στις πραγματικές τιμές απ΄ ότι αυτή του Breslow.

1.3 Κριτήρια με ποινή

Ας υποθέσουμε ότι από ένα δείγμα αναπτύσσονται δύο ανταγωνιστικά μοντέλα, δηλαδή έχουμε δύο πιθανά υποψήφια μοντέλα. Έστω l_1 και l_2 οι αντίστοιχες τιμές της ποσότητας

$-2 * \ln(\text{πιθανοφάνεια})$. Έστω επίσης ότι ισχύει $l_1 < l_2$. Τότε εύκολα υποκύπτουμε στον πειρασμό να χαρακτηρίσουμε το πρώτο μοντέλο ως αυτό που προσαρμόζεται καλύτερα ή αυτό που κάνει καλύτερη πρόβλεψη. Όμως, μπορεί αυτό το μοντέλο να παρέχει καλύτερη προσαρμογή άμεσα αλλά στην περίπτωση που χρειάστηκε πολύ περισσότερες μεταβλητές από το δεύτερο, μπορεί να μην είναι 'οικονομικό'. Αν και τα δύο μοντέλα εφαρμοστούν σε ένα καινούριο δείγμα, η 'υπερπροσαρμογή' του πρώτου μοντέλου στα αρχικά δεδομένα μπορεί να οδηγήσει σε χειρότερη προσαρμογή του για το νέο σύνολο δεδομένων.

Τα κριτήρια AIC και BIC παρέχουν μια μέθοδο στην οποία επιβάλλεται 'ποινή' στη λογαριθμοποιημένη πιθανοφάνεια ανάλογα με την πολυπλοκότητα του μοντέλου, έτσι ώστε να αποκτήσουμε μια πιο αντικειμενική εκτίμηση της αξίας ενός μοντέλου.

Το κριτήριο *AIC (Akaike's Information Criterion)* αναπτύχθηκε από τον Hirotogu Akaike (1973) και αποτελεί ένα κριτήριο επιλογής του βέλτιστου μοντέλου με όσο το δυνατόν μικρότερο αριθμό παραμέτρων. Αποτελεί ένα μέτρο της σχετικής καλής προσαρμογής ενός στατιστικού μοντέλου. Στην πράξη το κριτήριο στηρίζεται στη χαμένη πληροφορία όταν ένα μοντέλο χρησιμοποιείται για να περιγράψει την πραγματικότητα. Μπορούμε να πούμε ότι περιγράφει το 'ζύγισμα' μεταξύ μεροληψίας και διακύμανσης κατά την κατασκευή του μοντέλου δηλαδή μεταξύ ακρίβειας και πολυπλοκότητας. Το AIC δεν κάνει κανέναν έλεγχο υποθέσεων για το μοντέλο. Με αυτή την έννοια, δεν μας παρέχει καμία πληροφορία για το πόσο καλά προσαρμόζεται ένα υποψήφιο μοντέλο στα δεδομένα. Αν κανένα από τα υποψήφια μοντέλα δεν προσαρμόζεται καλά, το AIC δε θα μας πληροφορήσει καθόλου για το γεγονός αυτό. Προτιμότερο μοντέλο με βάση αυτό το κριτήριο είναι εκείνο με το μικρότερο AIC. Στη γενική περίπτωση ορίζεται από τη σχέση

$$AIC = 2d - 2\ell, \tag{1.27}$$

όπου d το πλήθος των παραμέτρων και ℓ η μεγιστοποιημένη τιμή της λογαριθμοποιημένης συνάρτησης πιθανοφάνειας. Η εισαγωγή επιπλέον μεταβλητών στο μοντέλο βελτιώνει την προσαρμογή του στα δεδομένα ανεξάρτητα από το αν αυτές είναι στατιστικά σημαντικές ή όχι. Συνεπώς εισάγοντας νέες μεταβλητές αυξάνεται ο όρος ℓ στη σχέση (1.27). Από την άλλη όμως, το AIC περιλαμβάνει και μία 'ποινή' η οποία είναι αύξουσα συνάρτηση του πλήθους των παραμέτρων και εκφράζεται μέσω του όρου $2d$ στην (1.27). Τελικά, η εισαγωγή επιπλέον παραμέτρων στο μοντέλο μειώνει την τιμή του AIC μόνο αν αυτές βελτιώνουν την

προσαρμογή του μοντέλου σε βαθμό που υπερβαίνει το αυξημένο αντίβαρο του πρώτου όρου $2d$. Σχεδόν πάντα θα υπάρχει χαμένη πληροφορία λόγω της χρήσης ενός υποψήφιου πιθανού μοντέλου προκειμένου να αναπαραστήσουμε το 'πραγματικό'. Θέλουμε απλά μεταξύ κάποιων υποψήφιων μοντέλων να επιλέξουμε εκείνο για το οποίο ελαχιστοποιείται η χαμένη πληροφορία. Δεν μπορούμε να επιλέξουμε με βεβαιότητα, αλλά μπορούμε να ελαχιστοποιήσουμε την εκτιμώμενη απώλεια πληροφορίας.

Αν όλα τα υποψήφια μοντέλα έχουν τον ίδιο πλήθος παραμέτρων, τότε η χρήση του AIC πιθανόν να φαίνεται να μοιάζει με τον έλεγχο του λόγου των πιθανοφανειών. Όμως κάτι τέτοιο δεν ισχύει, αφού ο έλεγχος του λόγου πιθανοφανειών μπορεί να εφαρμοστεί μόνο για εμφωλευμένα μοντέλα, περιορισμός που στην περίπτωση του AIC δεν υφίσταται.

Στη περίπτωση του γενικού γραμμικού μοντέλου η συνάρτηση πιθανοφάνειας έχει τη μορφή

$$L(\sigma^2, \beta) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{S(\beta)}{2\sigma^2}\right\} \tag{1.28}$$

όπου $S(\beta) = \sum_{i=1}^{n}(y_i - \mathbf{x_i'}\beta)^2$. Η μεγιστοποίηση γίνεται ως προς σ^2 και β. Η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας παίρνει τη μορφή

$$\ell(\sigma^2, \beta) = \sum_{i=1}^{n} \left\{-\ln\sigma - \frac{1}{2\sigma^2}S(\beta) - \frac{1}{2}\ln(2\pi)\right\}. \tag{1.29}$$

Αντικαθιστώντας τις ε.ε.τ. (1.3) και την εκτιμήτρια μέγιστης πιθανοφάνειας $\hat{\sigma}^2$ από την (1.6), η (1.29) γίνεται

$$\ell = -n\ln\hat{\sigma} - \frac{1}{2}n - \frac{n}{2}\ln(2\pi). \tag{1.30}$$

Η τελευταία αντιστοιχεί στη μεγιστοποιημένη τιμή της λογαριθμοποιημένης συνάρτησης πιθανοφάνειας ℓ η οποία εμφανίζεται στο γενικό τύπο του AIC (1.27). Τελικά το κριτήριο στην περίπτωση του γενικού γραμμικού μοντέλου δίνεται από τη σχέση

$$AIC = 2n\ln(\hat{\sigma}) + n + n\ln(2\pi) + 2(p+1). \tag{1.31}$$

όπου $d = p + 1$ το πλήθος των παραμέτρων υπό εκτίμηση ($p = k + 1$ παράμετροι του γραμμικού μοντέλου και μία η διασπορά σ^2). Χρησιμοποιώντας την (1.6), η τελευταία σχέση μπορεί να εκφραστεί στη μορφή

$$AIC = n\left[\ln\left(\frac{2\pi SSE}{n}\right) + 1\right] + 2(p+1) \tag{1.32}$$

Στην περίπτωση μικρού μεγέθους δείγματος, προτείνεται το λεγόμενο διορθωμένο AIC το οποίο δίνεται από τη σχέση

$$AIC_c = AIC + \frac{2d(d+1)}{n-d-1}. \tag{1.33}$$

Καθώς το μέγεθος του δείγματος n μεγαλώνει, το AIC_c τείνει στο AIC. Οι Burnham και Anderson (2002) συνιστούν το διορθωμένο AIC συγκριτικά με το απλό σε περιπτώσεις που το n είναι μικρό ή το d είναι μεγάλο.

Στην περίπτωση της λογιστικής παλινδρόμησης, η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας δίνεται από την (1.16). Αντικαθιστώντας στην (1.27), προκύπτει ότι η γενική μορφή του AIC στην περίπτωση της λογιστικής παλινδρόμησης έχει τη μορφή

$$AIC = 2p - 2\sum_{i=1}^{n}\left\{\ln\begin{pmatrix} n_i \\ y_i \end{pmatrix} + y_i\mathbf{x}_\mathbf{i}'\beta - n_i\ln(1+\exp(\mathbf{x}_\mathbf{i}'\beta))\right\}. \tag{1.34}$$

Ένα κριτήριο παρόμοιο με το AIC είναι το BIC (Bayesian Information Criterion) που προτάθηκε από τον Gideon E. Schwarz (1978) ο οποίος έδωσε ένα Μπεϋζιανό επιχείρημα για τη χρήση του. Αποτελεί και αυτό ένα κριτήριο επιλογής του βέλτιστου μοντέλου και όπως και το AIC λαμβάνοντας υπόψη το γεγονός ότι η εισαγωγή πολλών μεταβλητών στο μοντέλο αυξάνει την πιθανοφάνεια αλλά οδηγεί σε overfitting επιβάλλει μια 'ποινή' για τον αριθμό των μεταβλητών στο μοντέλο. Στη γενική περίπτωση ο τύπος για το BIC είναι

$$BIC = d\ln n - 2\ell. \tag{1.35}$$

Όπως και στην περίπτωση του AIC, μεταξύ δύο υποψήφιων μοντέλων, προτιμούμε εκείνο με τη μικρότερη τιμή του κριτηρίου BIC. Η βασική διαφορά μεταξύ των δύο κριτηρίων είναι (εκτός της διαφορετικής τους προέλευσης) ότι το BIC επιβάλλει μεγαλύτερη ποινή στις παραμέτρους με αποτέλεσμα να αποθαρρύνεται η εισαγωγή επιπρόσθετων παραμέτρων σε μεγαλύτερο βαθμό απ'οτι στο AIC. Βέβαια και στην περίπτωση του BIC δεν υπάρχει κανένας περιορισμός ότι τα συγκρινόμενα μοντέλα πρέπει να είναι εμφωλευμένα.

Με διαδικασία ανάλογη με αυτή που ακολουθήθηκε στην περίπτωση του AIC, μπορεί να εκφραστεί και το κριτήριο BIC στην περίπτωση του γενικού γραμμικού μοντέλου καθώς και στην περίπτωση της λογιστικής παλινδρόμησης.

Στην περίπτωση του γραμμικού μοντέλου

$$BIC = 2n\ln(\hat{\sigma}) + n + n\ln(2\pi) + (p+1)\ln n \tag{1.36}$$

ή ισοδύναμα

$$BIC = n\left[\ln\left(\frac{2\pi SSE}{n}\right) + 1\right] + (p+1)\ln n. \tag{1.37}$$

Στην περίπτωση της λογιστικής παλινδρόμησης

$$BIC = p\ln n - 2\sum_{i=1}^{n}\left\{\ln\begin{pmatrix} n_i \\ y_i \end{pmatrix} + y_i\mathbf{x_i}'\beta - n_i\ln(1+\exp(\mathbf{x_i}'\beta))\right\}. \tag{1.38}$$

Τα δύο παραπάνω κριτήρια είναι από τα πιο διαδεδομένα και ο υπολογισμός τους παρέχεται από τα περισσότερα στατιστικά πακέτα. Υπάρχουν και άλλα κριτήρια τα οποία επίσης χρησιμοποιούνται με σκοπό την επιλογή του βέλτιστου μοντέλου αλλά είτε εφαρμόζονται σε ιδιαίτερες περιπτώσεις είτε έχουν διαφορετική φιλοσοφία. Μια γενίκευση του AIC αποτελεί το κριτήριο DIC (Deviance Information Criterion) που προτάθηκε από τους Spiegelhalter et.al. (2002). Είναι ιδιαίτερα χρήσιμο σε προβλήματα επιλογής Μπεϋζιανού μοντέλου όπου η posterior κατανομή των μοντέλων έχει προκύψει μέσω MCMC προσομοίωσης. Το FIC (Focused Information Criterion) αποτελεί και αυτό μια μέθοδο επιλογής του καταλληλότερου μοντέλου για κάποιο σύνολο δεδομένων. Όπως αναφέρεται από τους Claeskens and Hjort (2003), σε αντίθεση με άλλες μεθόδους επιλογής όπως τα AIC, BIC και DIC, το FIC δεν επιχειρεί να αξιολογήσει τη συνολική προσαρμογή των υποψήφιων μοντέλων αλλά επικεντρώνεται απευθείας στην πιο σημαντική παράμετρο σύμφωνα με τη στατιστική ανάλυση για την οποία τα υποψήφια μοντέλα οδηγούν σε διαφορετικές εκτιμήσεις. Μπορεί να εφαρμοστεί στην επιλογή μεταβλητών σε μοντέλα παλινδρόμησης συμπεριλαμβανομένων των γενικευμένων γραμμικών μοντέλων και στα ημιπαραμετρικά μοντέλα αναλογικής διακινδύνευσης (π.χ. Cox).

Τέλος, ένα ακόμα από τα πιο διαδεδομένα κριτήρια για τα γραμμικά μοντέλα το οποίο παρέχεται από τα περισσότερα στατιστικά πακέτα είναι ο δείκτης του Mallow C_p. Δίνεται από τη σχέση

$$C_p = \frac{SSE(p)}{\hat{\sigma}^2} + 2p - n, \tag{1.39}$$

όπου $\hat{\sigma}^2$ η αμερόληπτη εκτιμήτρια της διασποράς σ^2 όπως δίνεται από τη σχέση (1.7) για το μοντέλο με όλες τις δυνατές μεταβλητές και $SSE(p)$ το παρατηρούμενο άθροισμα των

τετραγώνων των υπολοίπων του υπό εξέταση μοντέλου. Χρησιμοποιείται προκειμένου να αποφασισθεί το μοντέλο που προσαρμόζεται καλύτερα στα δεδομένα. Γνωρίζοντας την τιμή του κριτηρίου C_p για διάφορα υποψήφια μοντέλα, ως βέλτιστο θεωρείται εκείνο για το οποίο

$$C_p \simeq p.$$

Στην περίπτωση που μεταξύ των υποψήφιων μοντέλων προκύψουν τουλάχιστον δύο από αυτά να ικανοποιούν την παραπάνω συνθήκη, τότε επιλέγεται το μοντέλο με το μικρότερο p, δηλαδή εκείνο με τις λιγότερες μεταβλητές αφού όπως έχει αναφερθεί αναζητούμε εκείνο το μοντέλο που περιγράφει όσο το δυνατόν καλύτερα τη μεταβλητή απόκρισης αλλά ταυτόχρονα είναι και οικονομικό.

Γενικά, η επιλογή μοντέλου που βασίζεται στα παραπάνω κριτήρια δεν πρέπει να γίνεται αποκλειστικά με τη χρήση ενός και μόνο κριτηρίου αλλά θα πρέπει να λαμβάνονται υπόψη και άλλοι παράγοντες όπως η φύση του προβλήματος και των δεδομένων, πιθανή επιπλέον εξωτερική γνώση, ο σκοπός χρήσης του μοντέλου και άλλα καθώς επίσης να συνεκτιμώνται τα αποτελέσματα περισσότερων από ένα κριτήρια.

1.4 Τεχνικές επιλογής με στατιστικούς ελέγχους

Κλασικοί στατιστικοί έχουν ασχοληθεί εκτενώς με το πρόβλημα της επιλογής μεταβλητών. Σε αντίθεση με την μπεϋζιανή προσέγγιση, οι κλασικές προσεγγίσεις του προβλήματος βασίζονται στη σύγκριση εμφωλευμένων μοντέλων ή, όπως στην περίπτωση του γραμμικού μοντέλου παλινδρόμησης, στην αρχή των επιπρόσθετων αθροισμάτων τετραγώνων, δηλαδή συγκρίνοντας τα αθροίσματα των τετραγώνων των υπολοίπων από μοντέλα με και χωρίς κάποιες από τις μεταβλητές. Έγκυρες συγκρίσεις μπορούν να γίνουν για μοντέλα που διαφέρουν στο ότι, το μικρότερο μοντέλο προκύπτει από το μεγαλύτερο θέτοντας μερικές από τις παραμέτρους (δηλαδή συντελεστές μεταβλητών) ίσες με το μηδέν. Αντίθετα, όταν χρησιμοποιούμε κάποια κριτήρια όπως π.χ. AIC (παράγραφος 1.3) ή cvl (παράγραφος 2.7), τα συγκρινόμενα μοντέλα δεν είναι απαραίτητο να είναι εμφωλευμένα.

Υποθέτοντας ότι έχει γίνει η προσαρμογή του μοντέλου συμπεριλαμβάνοντας όλες τις συμμεταβλητές, μπορούν να πραγματοποιηθούν στατιστικοί έλεγχοι για τον έλεγχο της σημαντικότητας κάθε ανεξάρτητης μεταβλητής. Αυτό σημαίνει ότι θέλουμε να ελέγξουμε κατά

πόσο η κάθε μεταβλητή πραγματικά επηρεάζει τη μεταβλητή απόκρισης και έχει λόγο να βρίσκεται μέσα στο μοντέλο που τελικά θα καταλήξουμε και βάση αυτού θα μελετήσουμε την εξαρτημένη μεταβλητή.

Ένας πρώτος έλεγχος είναι αυτός του *λόγου πιθανοφανειών*. Έστω B_r ένα διάνυσμα r παραμέτρων από τις $\beta_1, \ldots, \beta_k$ για το οποίο μας ενδιαφέρει να ελέγξουμε αν μπορεί να πάρει κάποια συγκεκριμένη τιμή. Θέλουμε δηλαδή να πραγματοποιήσουμε τον έλεγχο αν οι μεταβλητές που περιέχονται στο B_r μπορούν ταυτόχρονα να πάρουν κάθε μία κάποια συγκεκριμένη (όχι απαραίτητα ίδια) τιμή όπως δηλώνεται μέσω του διανύσματος B_r^0. Οι υπόλοιπες $q = k - r$ παράμετροι αποτελούν το σύνολο B_q και παραμένουν χωρίς περιορισμό. Η μηδενική υπόθεση, H_0, είναι λοιπόν $B_r = B_r^0$. Έστω b η εκτιμήτρια (διάνυσμα) μέγιστης πιθανοφάνειας του β και b_q^* ο ε.μ.π. του B_q υπό τον περιορισμό $B_r = B_r^0$. Επίσης έστω ℓ η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας που αντιστοιχεί στο μοντέλο. Η ℓ_0 είναι η αντίστοιχη μεγιστοποιημένη κάτω από την H_0 στην οποία όμως έχει χρησιμοποιηθεί το διάνυσμα b_q^* για τις μεταβλητές των οποίων η τιμή δεν καθορίζεται από την H_0. Άρα $\ell_0 = \ell(B_r^0, b_q^*)$. Κάτω από την εναλλακτική υπόθεση δεν υπάρχει κανένας περιορισμός και άρα $\ell_1 = \ell(b)$ είναι η αντίστοιχη μεγιστοποιημένη λογαριθμοποιημένη συνάρτηση πιθανοφάνειας. Η στατιστική συνάρτηση του λόγου πιθανοφανειών (Likelihood Ratio) δίνεται ως

$$LR = -2\{\ell_0 - \ell_1\} \tag{1.40}$$

$$LR = -2\{\ell(B_r^0, b_q^*) - \ell(b)\} \tag{1.41}$$

Η στατιστική συνάρτηση LR, για αρκετά μεγάλα δείγματα, προσεγγιστικά ακολουθεί X^2 κατανομή με τόσους βαθμούς ελευθερίας όσες οι παράμετροι που εκτιμώνται, όσες δηλαδή ελέγχονται κάτω από την μηδενική υπόθεση (Harrell, 2002). Η LR συχνά χρησιμοποιείται με σκοπό να εξετάζονται σταδιακά όλο και πιο πολύπλοκα μοντέλα σε μία διαδικασία τύπου ανάλογη με τη διαδικασία σε βήματα (stepwise). Θεωρώντας τη μηδενική υπόθεση $B_r = \mathbf{0}$, στην ουσία εξετάζεται αν οι μεταβλητές που περιέχονται στο διάνυσμα B_r συμμετέχουν στο μοντέλο ή όχι. Συνεπώς αν ο έλεγχος περιοριστεί ώστε το B_r να είναι μονοδιάστατο, να αποτελείται δηλαδή από έναν και μόνο συντελεστή, στην ουσία εξετάζουμε αν αυτή η μεταβλητή είναι στατιστικά σημαντική ή όχι. Η διαδικασία γενικεύεται προκειμένου να επιτευχθεί και η σύγκριση δύο μοντέλων μεταξύ τους χωρίς απαραίτητα το εναλλακτικό μοντέλο να

περιέχει όλες τις μεταβλητές. Ο μοναδικός περιορισμός είναι ότι το μοντέλο κάτω από την εναλλακτική υπόθεση περιέχει όλες τις μεταβλητές του μηδενικού μοντέλου και σίγουρα τουλάχιστον μια ακόμα. Τότε η στατιστική συνάρτηση του λόγου των πιθανοφανειών ακολουθεί προσεγγιστικά X^2 κατανομή με βαθμούς ελευθερίας το πλήθος των επιπλέον μεταβλητών που περιέχονται στο εναλλακτικό μοντέλο, δηλαδή οι βαθμοί ελευθερίας είναι $d = q - r \geq 1$. Από το τελευταίο είναι προφανές ότι η μέθοδος μπορεί να εφαρμοστεί μόνο για τη σύγκριση εμφωλευμένων μοντέλων.

Ας υποθέσουμε ότι είμαστε στην ειδική περίπτωση του γενικού γραμμικού μοντέλου και επιθυμούμε, όπως παραπάνω, να κάνουμε τον έλεγχο της $H_0 : \ B_r = 0$ ενώ στην H_1 δεν υπάρχει κανένας περιορισμός για τις παραμέτρους. Από τις σχέσεις (1.29) και (1.6) φαίνεται ότι η μεγιστοποιημένη λογαριθμοποιημένη πιθανοφάνεια εξαρτάται από την ποσότητα SSE. Συνεπώς, ο έλεγχος του λόγου πιθανοφανειών εξαρτάται από την ποσότητα

$$\frac{SSE_0 - SSE_1}{SSE_1} = \frac{\text{μείωση του αθροίσματος τετραγώνων των υπολοίπων}}{\text{άθροισμα τετραγώνων των υπολοίπων υπό την } H_1}.$$

Λόγω της κανονικής κατανομής μπορεί να δειχθεί ότι

$$\frac{SSE_0 - SSE_1}{\sigma^2} \sim X^2_r \tag{1.42}$$

και

$$\frac{SSE_1}{\sigma^2} \sim X^2_{n-p}. \tag{1.43}$$

Επομένως σχηματίζοντας το κατάλληλο πηλίκο καταλήγουμε όπως φαίνεται παρακάτω σε μια στατιστική συνάρτηση η οποία ακολουθεί την F κατανομή και αποτελεί το γνωστό κριτήριο F.

$$\frac{\frac{SSE_0 - SSE_1}{\sigma^2}/r}{\frac{SSE_1}{\sigma^2}/(n-p)} = \frac{(SSE_0 - SSE_1)/r}{SSE_1/(n-p)} \sim F_{r,(n-p)}. \tag{1.44}$$

Στην περίπτωση των γενικευμένων γραμμικών μοντέλων, θεωρούμε τον έλεγχο όπου η εναλλακτική υπόθεση είναι η H_S, όπου με S δηλώνεται το κορεσμένο μοντέλο με αριθμό παραμέτρων όσες και το μέγεθος του δείγματος και η H_0 το υποψήφιο μοντέλο που θέλουμε να ελέγξουμε με όλες τις $p < n$ παραμέτρους. Για το κορεσμένο μοντέλο γνωρίζουμε ότι είναι το καλύτερο δυνατό με την έννοια ότι προσεγγίζει όσο το δυνατόν καλύτερα το πραγματικό και έχει n παραμέτρους έτσι ώστε οι προβλεπόμενες τιμές να συμπίπτουν με τις παρατηρούμενες. Η συνάρτηση

$$D = -2(\ell_0 - \ell_S) \tag{1.45}$$

είναι η ελεγχοσυνάρτηση Deviance. Παρατηρούμε ότι $-2(\ell_0-\ell_S)+2(\ell_1-\ell_S) = -2(\ell_0-\ell_1) = LR$. Διαπιστώνουμε συνεπώς ότι για τον έλεγχο υποθέσεων όπως αυτός ορίστηκε για τη διαξαγωγή της μεθόδου του λόγου πιθανοφανειών, υπολογίζοντας τη διαφορά D_0-D_1, όπου D_i η Deviance υπό την $H_i,\ i=0,1$, καταλήγουμε στην ίδια στατιστική συνάρτηση η οποία ακολουθεί ασυμπτωτικά X^2_d κατανομή. Η τιμή της ελεγχοσυνάρτησης Deviance δίνεται από τα περισσότερα στατιστικά πακέτα στην περίπτωση των γενικευμένων γραμμικών μοντέλων και όχι η τιμή της λογαριθμοποιημένης συνάρτησης πιθανοφάνειας όπως συνήθως συμβαίνει στα γενικά γραμμικά μοντέλα.

Ο έλεγχος του Wald αποτελεί μια ακόμα επιλογή για τον έλεγχο της στατιστικής σημαντικότητας μιας μεταβλητής. Μπορεί να χρησιμοποιηθεί γενικά για τον έλεγχο της μηδενικής υπόθεσης $H_0:\ \beta_j = \beta_j^0$. Αν $\beta_j^0 = 0$ τότε ουσιαστικά ελέγχεται η στατιστική σημαντικότητα της μεταβλητής X_j. Βασίζεται στο γεγονός ότι η ελεγχοσυνάρτηση

$$Z = \frac{\hat{\beta}_j - \beta_j}{\left(I^{-1}(\hat{\beta})_{jj}\right)^{\frac{1}{2}}} \sim N(0,1), \qquad \text{ασυμπτωτικά} \tag{1.46}$$

ή ισοδύναμα η

$$Z^2 = \frac{(\hat{\beta}_j - \beta_j)^2}{\left(I^{-1}(\hat{\beta})_{jj}\right)} \sim X^2_1, \qquad \text{ασυμπτωτικά} \tag{1.47}$$

όπου $\left(I^{-1}(\hat{\beta})_{jj}\right)$ το j-οστό διαγώνιο στοιχείο του αντίστροφου πίνακα πληροφορίας $I(\hat{\beta})$ το οποίο ισούται με την εκτιμώμενη διασπορά, $\hat{V}(\hat{\beta}_j)$, της εκτιμήτριας $\hat{\beta}_j$. Ο έλεγχος αυτός έχει το μειονέκτημα ότι δεν μπορεί να εφαρμοστεί για τον ταυτόχρονο έλεγχο πολλών μεταβλητών. Ένας ακόμα λόγος για τον οποίο ο έλεγχος του λόγου πιθανοφανειών προτιμάται έναντι του Wald είναι το γεγονός ότι ο τελευταίος μπορεί να δώσει διαφορετικά αποτελέσματα στην ίδια ερώτηση ανάλογα με το πώς αυτή έχει εκφραστεί. Για παράδειγμα, η ερώτηση αν $R=1$ είναι ισοδύναμη με την ερώτηση αν $\ln R = 0$. Όμως η στατιστική συνάρτηση του Wald δεν είναι ίδια για $R=1$ και για $\ln R = 0$ αφού δεν υπάρχει σχέση μεταξύ των τυπικών σφαλμάτων των τυχαίων μεταβλητών R και $\ln R$. Αντίθετα ο έλεγχος του λόγου πιθανοφανειών θα δώσει ακριβώς τα ίδια αποτελέσματα είτε δουλεύουμε με τη μεταβλητή R είτε με την $\ln R$ είτε γενικά με οποιαδήποτε μονότονο μετασχηματισμό της μεταβλητής R. Βέβαια ο έλεγχος του Wald είναι χρήσιμος γιατί είναι απλός και σε αυτόν βασίζεται η εύρεση διαστημάτων εμπιστοσύνης για τη μεταβλητή β_j.

Οι διάφορες διαδικασίες σε βήματα, όπως είναι η προς τα εμπρός ή προς τα πίσω επιλογή, είναι μεταξύ των πιο δημοφιλών και διαδεδομένων τεχνικών (Kadane and Lazar, 2004). Παρέχουν ένα συστηματικό τρόπο αναζήτησης μεταξύ μοντέλων, όπου σε κάθε στάδιο προκύπτει ένα νέο μοντέλο προσθέτοντας ή αφαιρώντας κάποια μεταβλητή από το μοντέλο του προηγούμενου βήματος. Αν και οι τεχνικές αυτές ξεκίνησαν από τα μοντέλα παλινδρόμησης για να βοηθήσουν στην επιλογή μεταβλητών, μπορούν επίσης να εφαρμοστούν και σε επεκτάσεις αυτών όπως είναι τα γενικευμένα γραμμικά μοντέλα (Lawless and Shinhal, (1978); Hastie and Pregibon, 1992). Η φιλοσοφία της μεθόδου είναι η ίδια ανεξάρτητα από το είδος του μοντέλου στο οποίο αναφερόμαστε. Αυτό που διαφέρει είναι το κριτήριο στο οποίο βασιζόμαστε για να λάβουμε την απόφαση αν μια μεταβλητή θα εισαχθεί ή όχι το μοντέλο.

Στην προς τα εμπρός ή διαδικασία διαδοχικής πρόσθεσης (forward selection), η διαδικασία ξεκινάει με το μοντέλο χωρίς καθόλου μεταβλητές και ελέγχει μια μια τις υποψήφιες μεταβλητές για να εισαχθούν στο μοντέλο. Στην περίπτωση που πρόκειται για ένα γραμμικό μοντέλο παλινδρόμησης, στο πρώτο βήμα εισάγεται η μεταβλητή που δίνει τη μεγαλύτερη στατιστικά σημαντική τιμή της ελεγχοσυνάρτησης F, δηλαδή αυτή που συνεισφέρει περισσότερο στην επεξήγηση της εξαρτημένης μεταβλητής, δεδομένου βέβαια ότι έχει προκαθοριστεί ένα κατώφλι στην τιμή του κριτηρίου F. Σε κάθε βήμα η διαδικασία συνεχίζεται με τον ίδιο τρόπο, προσθέτοντας κάθε φορά τη μεταβλητή που επιδρά περισσότερο, δεδομένου ότι οι άλλες μεταβλητές είναι ήδη στο μοντέλο, αν η τιμή του F είναι πάνω από το κατώφλι. Ο αλγόριθμος σταματάει όταν δεν υπάρχουν πια άλλες υποψήφιες μεταβλητές που να πληρούν το κριτήριο εισόδου στο μοντέλο.

Η προς τα πίσω ή διαδικασία διαδοχικής αφαίρεσης (backward elimination), είναι παρόμοια αλλά κινείται προς την αντίθετη κατεύθυνση. Δηλαδή, ξεκινάει με το πλήρες μοντέλο και σε κάθε βήμα αφαιρείται η μεταβλητή με τη μικρότερη επίδραση στο μοντέλο, δεδομένου ότι οι άλλες μεταβλητές συμπεριλαμβάνονται σε αυτό. Και στην περίπτωση αυτή, υπάρχει ένα προκαθορισμένο κατώφλι βάση του οποίου λαμβάνεται η απόφαση της εξόδου μιας μεταβλητής. Όταν καμία από τις μεταβλητές που έχουν απομείνει δεν πληρούν το κριτήριο εξόδου, ο αλγόριθμος σταματάει.

Τόσο στην προς τα εμπρός όσο και στην προς τα πίσω διαδικασία, η απόφαση εισόδου ή εξόδου μιας μεταβλητής από το μοντέλο είναι αμετάκλητη. Αυτή η ακαμψία των δύο μεθόδων

αντιμετωπίζεται στη διαδικασία κατά βήματα (stepwise selection). Εδώ, σε κάθε βήμα οι μεταβλητές εξετάζονται αν θα μπουν ή αν θα βγουν από το μοντέλο. Η προσθήκη μιας μεταβλητής στο μοντέλο μπορεί να έχει ως αποτέλεσμα την εξασθένιση της σημαντικότητας κάποιας άλλης οπότε η τελευταία θα πρέπει να φύγει παρόλο που σε κάποιο προηγούμενο στάδιο φάνηκε να ήταν στατιστικά σημαντική. Αντί του κριτηρίου F χρησιμοποιούμε και άλλα κριτήρια, πχ AIC, BIC.

Σημειώνεται ότι τόσο στην περίπτωση των γενικευμένων γραμμικών μοντέλων όσο και στο ημιπαραμετρικό μοντέλο του Cox, η σύγκριση μεταξύ μοντέλων μπορεί να γίνει και με βάση την ελεγχοσυνάρτηση Deviance και με το λόγο των πιθανοφανειών. Επίσης για τα ίδια μοντέλα μπορεί να εφαρμοστεί η διαδικασία σε βήματα με χρήση των κριτηρίων AIC, BIC προκειμένου να αποφασισθεί η εισαγωγή ή η αφαίρεση μιας μεταβλητής από το μοντέλο.

Φυσικά ένας λογικός εναλλακτικός τρόπος για την εξεύρεση του βέλτιστου μοντέλου θα ήταν η μελέτη όλων των δυνατών μοντέλων, δηλαδή η μελέτη όλων των δυνατών υποσυνόλων μεταβλητών, και με βάση κάποιο κριτήριο να γίνει η επιλογή του βέλτιστου. Αυτό μπορεί να γίνει όμως μόνο στην περίπτωση που έχουμε λίγες μεταβλητές. Αυτό κάνει η μέθοδος Best Subsets (B.S.). Κατασκευάζονται όλα τα δυνατά μοντέλα τα οποία περιέχουν 1,2,... συμμεταβλητές και από αυτά παρουσιάζονται τα επικρατέστερα. Ανάλογα με το στατιστικό πακέτο, μπορεί να δίνονται για κάθε μοντέλο οι τιμές διαφόρων κριτηρίων π.χ. AIC, BIC, C_p Mallow, R^2 κ.τ.λ. Η τελική επιλογή του μοντέλου αφήνεται στον αναλυτή. Τόσο με τη διαδικασία σε βήματα όσο και με τη $B.S.$, προκύπτουν μοντέλα τα οποία περιέχουν κάποιες συμμεταβλητές από ένα σύνολο που ο αναλυτής καθορίζει. Από τη διαδικασία σε βήματα προκύπτει ένα τελικό μοντέλο ενώ από τη $B.S.$ προκύπτουν τα επικρατέστερα, αφού γίνει επεξεργασία σε όλα τα δυνατά, γεγονός που δυσχεραίνει την εφαρμογή της μεθόδου σε περιπτώσεις μοντέλων με μεγάλο πλήθος μεταβλητών.

Κεφάλαιο 2

Τεχνικές με ποινή

Έστω ότι έχουμε δεδομένα (x_i, y_i) όπου $x_i = (x_{i1}, \ldots, x_{ip})'$ και y_i είναι οι τιμές που αντιστοιχούν στην $i-$οστή παρατήρηση. Οι συνήθεις εκτιμήτριες ελαχίστων τετραγώνων προκύπτουν ελαχιστοποιώντας το άθροισμα των τετραγώνων των υπολοίπων. Υπάρχουν δύο λόγοι για τους οποίους ένας αναλυτής δεν είναι ικανοποιημένος με τις εκτιμήτριες αυτές. Όπως τονίζει ο Tibshirani (1996) ο πρώτος λόγος είναι η ακρίβεια πρόβλεψης: οι ε.ε.τ. συχνά έχουν μικρή μεροληψία αλλά μεγάλη διασπορά. Η ακρίβεια πρόβλεψης μπορεί μερικές φορές να βελτιωθεί συρρικνώνοντας κάποιους συντελεστές ή θέτοντάς τους ίσους με το μηδέν. Κάνοντας κάτι τέτοιο αποδεχόμαστε λίγη μεροληψία προκειμένου να μειώσουμε τη διασπορά των προβλεπόμενων τιμών και έτσι μπορεί να βελτιωθεί η συνολική ακρίβεια πρόβλεψης του μοντέλου. Ο δεύτερος λόγος είναι η ερμηνεία. Όταν έχουμε μεγάλο αριθμό επεξηγηματικών μεταβλητών, συχνά θα θέλαμε να προσδιορίσουμε ένα μικρότερο σύνολο αυτών που αναδεικνύει τις μεγαλύτερες επιδράσεις. Προς την τελευταία κατεύθυνση έχουμε ήδη αναφέρει τη μέθοδο Best Subset η οποία όμως δεν ανήκει στις τεχνικές με ποινή. Με την B.S. προκύπτουν ερμηνεύσιμα μοντέλα αλλά είναι εξαιρετικά ασταθής αφού είναι μια διακριτή διαδικασία, οι επεξηγηματικές μεταβλητές είτε χρησιμοποιούνται είτε αποκλείονται από το μοντέλο. Μικρές μεταβολές στα δεδομένα μπορεί να οδηγήσουν στην επιλογή τελείως διαφορετικών μοντέλων και αυτό μπορεί να μειώσει την ακρίβεια πρόβλεψης. Θα παρουσιάσουμε λοιπόν στη συνέχεια δύο μεθόδους επιλογής μοντέλου οι οποίες ανήκουν στην κατηγορία των ποινικοποιημένων τεχνικών, την ανάλυση Κορυφογραμμής, Ridge Regression ή $\mathbf{L_2}$, και τη $LASSO$, (Least Absolute Shrinkage and Selection Operator) ή $\mathbf{L_1}$. Ως μέθοδοι μοιάζουν στο γεγονός ότι

και οι δύο επιβάλλουν μια ποινή στη συνάρτηση πιθανοφάνειας. Η βασική διαφορά τους όμως έγκειται στη μορφή της ποινής αυτής. Η επιβολή 'ποινής' προκειμένου να αποφευχθεί το φαινόμενο του overfitting έχει ήδη αναφερθεί στα κριτήρια AIC και BIC. Όμως τα κριτήρια αυτά δεν αποτελούν τεχνική εκτίμησης των συντελεστών των συμμεταβλητών στο μοντέλο αλλά δρουν μόνο συγκριτικά μεταξύ διαφόρων μοντέλων. Πριν από αυτά όμως ακολουθεί ένας σημαντικός έλεγχος που πρέπει να γίνει πριν πραγματοποιηθεί η επιλογή του μοντέλου.

2.1 Global Test

Πριν την εφαρμογή οποιασδήποτε μεθόδου με ποινή χρήσιμο είναι να μπορούμε να πραγματοποιήσουμε ένα συνολικό έλεγχο για το αν οι συμμεταβλητές συνδέονται με τη μεταβλητή απόκρισης. Αυτό κάνει το Global Test (gt) το οποίο μπορεί να εφαρμοστεί σε ένα σύνολο (ή υποσύνολο) των ανεξάρτητων μεταβλητών που συμμετέχουν σε ένα μοντέλο προκειμένου να ελεγχθεί αν αυτή η ομάδα σχετίζεται με την εξαρτημένη μεταβλητή. Η μηδενική υπόθεση αυτού του ελέγχου είναι ότι καμία από τις συμμεταβλητές που περιλαμβάνονται στο υπό έλεγχο σύνολο δεν συνδέεται με τη μεταβλητή απόκρισης. Παρόλ' αυτά ο έλεγχος αυτός είναι σχεδιασμένος κατά τέτοιο τρόπο ειδικά κατευθυνόμενο ενάντια στην εναλλακτική υπόθεση ότι οι περισσότερες συμμεταβλητές συνδέονται ελαφρά με την εξαρτημένη μεταβλητή (Goeman et.al., 2004). Πράγματι, ενάντια σε αυτή την εναλλακτική υπόθεση το gt είναι το βέλτιστο να χρησιμοποιηθεί και πραγματοποιείται ουσιαστικά ένας F έλεγχος.

Μπορεί να εφαρμοστεί σε γραμμικά μοντέλα στα οποία η κατανομή της εξαρτημένης μεταβλητής μοντελοποιείται ως συνάρτηση των συμμεταβλητών. Αυτό γίνεται μέσα στο πλαίσιο των γενικευμένων γραμμικών μοντέλων. Αν Y είναι η μεταβλητή απόκρισης, $X = (x_{ij})$ ο πίνακας ο οποίος περιέχει τις μεταβλητές που μας ενδιαφέρουν (στην περίπτωση που έχουμε όλες τις συμμεταβλητές τότε αντιστοιχεί στον πίνακα σχεδιασμού) και β ένα m-διάστατο διάνυσμα των συντελεστών παλινδρόμησης τότε, ως γνωστό έχουμε

$$E(Y_i|\beta) = g^{-1}\left(\beta_0 + \sum_{j=1}^{m} x_{ij}\beta_j\right), \tag{2.1}$$

όπου g η συνάρτηση σύνδεσης. Ο έλεγχος για το αν οι συμμεταβλητές έχουν προγνωστική

ισχύ για την Y είναι ισοδύναμο με τον έλεγχο ότι όλοι οι συντελεστές είναι μηδέν, δηλαδή

$$H_0 : \ \beta_1 = \beta_2 = \cdots = \beta_m = 0.$$

Αυτή η υπόθεση δεν είναι δυνατό να ελεγχθεί με κλασικούς τρόπους (με β μη στοχαστικό) διότι το m μπορεί να είναι μεγαλύτερο σε σχέση με το n. Στην περίπτωση αυτή υπάρχουν πολύ λίγοι βαθμοί ελευθερίας.

Μπορεί όμως να ελεγχθεί η παραπάνω μηδενική υπόθεση αν υποθέσουμε ότι τα β_1, β_2,...β_m αποτελούν ένα δείγμα από κάποια κατανομή με μέση τιμή μηδέν και διασπορά τ^2. Με τον τρόπο αυτό, μία και μόνη άγνωστη παράμετρος καθορίζει πόσο επιτρέπεται στους συντελεστές να απέχουν από το μηδέν. Η μηδενική υπόθεση απλοποιείται τώρα στην

$$H_0 : \ \tau^2 = 0.$$

Σημειώνεται ότι η επιλογή του $\tau^2 I_m$ (όπου I_m είναι ο $m \times m$ μοναδιαίος πίνακας) ως πίνακα συνδιακύμανσης για το διάνυσμα β δεν είναι υποχρεωτική. Απλά αποτελεί την πιο βολική επιλογή από την οποία θα προκύψει ένας έλεγχος ο οποίος συμπεριφέρεται σε όλες τις συμμεταβλητές επί ίσοις όροις. Αν για παράδειγμα κάποια πρώτερη γνώση υπάρχει σχετικά με την επιρροή κάποιας συμμεταβλητής περισσότερο στη μεταβλητή απόκρισης αυτό μπορεί να εισαχθεί μέσω του πίνακα συνδιακύμανσης και να οδηγήσει σε διαφορετικό έλεγχο με ισχύ έναντι διαφορετικών εναλλακτικών υποθέσεων.

Θέτουμε ως $r_i = \sum_j x_{ij}\beta_j, \ i = 1, 2, \ldots, n$. Οπότε η ποσότητα r_i συγκεντρώνει την επίδραση όλων των συμμεταβλητών για το άτομο i. Αν $\mathbf{r} = (r_1, r_2, \ldots, r_n)$, το διάνυσμα $\mathbf{r}$ είναι ένα τυχαίο διάνυσμα με $E(\mathbf{r}) = 0$ και $Cov(\mathbf{r}) = \tau^2 XX'$. Η (2.1) απλοποιείται λοιπόν ως

$$E(Y_i|r_i) = g^{-1}(\beta_0 + r_i). \tag{2.2}$$

Το τελευταίο αποτελεί ένα μοντέλο τυχαίων επιδράσεων, στο οποίο κάθε μονάδα έχει μια τυχαία επίδραση που επηρεάζει το αποτέλεσμα και ο πίνακας συνδιακύμανσης μεταξύ των τυχαίων επιδράσεων είναι γνωστός.

Έστω $R = \frac{1}{m} XX'$ ένας $n \times n$ πίνακας ανάλογος του πίνακα συνδιακύμανσης των τυχαίων σφαλμάτων $\mathbf{r}$, $\mu = g^{-1}(\beta_0)$ είναι η αναμενόμενη τιμή της Y κάτω από την H_0 και μ_2, μ_4 η δεύτερη και τέταρτη κεντρική ροπή της Y υπό την H_0. Η στατιστική συνάρτηση που

χρησιμοποιείται για τον έλεγχο global test είναι

$$T = \frac{(Y-\mu)'R(Y-\mu) - \mu_2 trace(R)}{\sqrt{2\mu_2^2 trace(R^2) + (\mu_4 - 3\mu_2^2)\sum_i R_{ii}^2}} \tag{2.3}$$

Αποδεικνύεται ότι όταν ισχύει η H_0, η T ακολουθεί ασυμπτωτικά κανονική κατανομή. Παρόλ' αυτά συχνά είναι πιο βολικό και χρησιμοποιείται η ισοδύναμη αλλά απλούστερη στατιστική συνάρτηση

$$Q = \frac{(Y-\mu)'R(Y-\mu)}{\mu_2} \tag{2.4}$$

η οποία έχει μέση τιμή

$$E(Q) = trace(R)$$

και διασπορά

$$Var(Q) = 2trace(R^2) + \left(\frac{\mu_4}{\mu_2^2} - 3\right)\sum_i R_{ii}^2.$$

Η στατιστική συνάρτηση Q επίσης ακολουθεί κανονική κατανομή ασυμπτωτικά αλλά αποτελεί μια τετραγωνική μορφή η οποία είναι μη αρνητική διότι ο πίνακας R είναι μη αρνητικά ορισμένος. Για το λόγο αυτό, για μικρά μεγέθη δειγμάτων, μια καλύτερη προσέγγιση της κατανομής της Q είναι η scaled X^2 κατανομή, cX_ν^2, με παράγοντα scaling $c = Var(Q)/[2E[Q]]$ και βαθμούς ελευθερίας $\nu = 2[E(Q)]^2/Var(Q)$.

2.2 Παλινδρόμηση Κορυφογραμμής (Ridge Regression)

Όταν υπάρχει έντονη συσχέτιση μεταξύ δύο ή περισσότερων επεξηγηματικών μεταβλητών, έχουμε το φαινόμενο της πολυσυγγραμμικότητας (multicollinearity). Η παρουσία της πολυσυγγραμμικότητας οδηγεί σε αυξημένα τυπικά σφάλματα των ε.ε.τ. και κατά συνέπεια δυσκολεύει την εκτίμηση της επίδρασης της κάθε επεξηγηματικής μεταβλητής στην εξαρτημένη μεταβλητή, αφού τα δ.ε. των αντίστοιχων συντελεστών θα είναι μεγάλα σε εύρος. Επίσης σε τέτοιες περιπτώσεις είναι δύσκολος ο εντοπισμός των στατιστικά σημαντικών μεταβλητών (Καρώνη και Οικονόμου, 2010). Η ακραία περίπτωση της απόλυτης πολυσυγγραμμικότητας προκύπτει όταν μια επεξηγηματική μεταβλητή είναι γραμμικός συνδυασμός

μερικών ή όλων των άλλων επεξηγηματικών μεταβλητών. Σε τέτοιες περιπτώσεις, η ανάλυση παλινδρόμησης μπορεί να πραγματοποιηθεί μόνο αφού αφαιρεθεί μια μεταβλητή από το γραμμικά εξαρτημένο σύνολο. Βέβαια, αρκετά συχνά παρουσιάζονται περιπτώσεις έντονης συσχέτισης μεταξύ των μεταβλητών χωρίς αυτές να είναι απόλυτα γραμμικά εξαρτημένες. Έστω R_j^2 ο συντελεστής προσδιορισμού της γραμμικής παλινδρόμησης με εξαρτημένη μεταβλητή τη x_j και επεξηγηματικές μεταβλητές όλες τις άλλες. Ο δείκτης R_j^2 εκφράζει δηλαδή το κατά πόσο η x_j μπορεί να προβλεφθεί από τις υπόλοιπες επεξηγηματικές μεταβλητές. Οι δείκτες $1-R_j^2$ (ανοχή - tolerance) προσφέρονται σε ορισμένα προγράμματα στατιστικής ανάλυσης προς εντοπισμό της πολυσυγγραμμικότητας. Η τιμή $\frac{1}{1-R_j^2}$ είναι γνωστή ως παράγοντας μεγέθυνσης διασποράς (variance inflation factor - VIF) και δείχνει κατά πόσο αυξάνεται η διασπορά ενός εκτιμημένου συντελεστή παλινδρόμησης όταν υπάρχουν συσχετίσεις των επεξηγηματικών μεταβλητών. Τιμές του $1 - R_j^2 < 0.2$ ή $VIF > 5$ θεωρούνται ως ένδειξη πολυσυγγραμικότητας.

Σύμφωνα με τους Marquardt και Snee (1975), η πολυσυγγραμμικότητα δεν έχει σημαντικές συνέπειες συνολικά στην αξία ενός μοντέλου. Για παράδειγμα, δεν επηρεάζει την καλή προσαρμογή του μοντέλου στην εξαρτημένη μεταβλητή, αλλά μας εμποδίζει να αντιληφθούμε το ρόλο της κάθε επεξηγηματικής μεταβλητής. Η τυποποίηση των μεταβλητών είναι απαραίτητη όταν εμφανίζεται σταθερός όρος στο μοντέλο. Η πολυσυγγραμμικότητα που εμφανίζεται από τη μη τυποποίηση των μεταβλητών είναι η πιο ύπουλη καθώς δεν οφείλεται σε πρόβλημα στα δεδομένα αλλά στην αυθαίρετη κλίμακα στην οποία εκφράζονται οι μεταβλητές πρόβλεψης. Στην τυποποίηση μιας μεταβλητής, από αυτή αφαιρείται ο μέσος ('κεντράρισμα' - centering) και στη συνέχεια διαιρείται με την τυπική της απόκλιση (scaling). Με το κεντράρισμα αφαιρείται η επουσιώδης πολυσυγγραμμικότητα μειώνοντας κατά συνέπεια το VIF των εκτιμητριών των συντελεστών του μοντέλου. Σε ένα γραμμικό μοντέλο, το κεντράρισμα αφαιρεί τη συσχέτιση μεταξύ του σταθερού και όλων των γραμμικών όρων. Επιπλέον, σε ένα τετραγωνικό μοντέλο, το κεντράρισμα μειώνει και σε μερικές περιπτώσεις αφαιρεί εντελώς τη συσχέτιση μεταξύ των γραμμικών και τετραγωνικών όρων (x_j, x_j^2). Το scaling εκφράζει την εξίσωση σε μια μορφή που προσφέρεται για πιο απλή ερμηνεία και χρήση.

Το VIF αντιστοιχεί στα διαγώνια στοιχεία του αντίστροφου πίνακα συσχέτισης. Για δύο επεξηγηματικές μεταβλητές κάθετες μεταξύ τους το VIF είναι 1. Το scaling δεν επηρεά-

ζει το VIF αλλά το κεντράρισμα επηρεάζει. Πώς εκτελούμε μια ανάλυση όταν υπάρχουν μεγάλες τιμές των VIF; Σε αυτές τις περιπτώσεις, η μια κλασική μέθοδος των ε.ε.τ. δε δίνει καλές εκτιμήτριες όταν υπάρχουν συσχετίσεις στα δεδομένα. Η άλλη κλασική μέθοδος της επιλογής μεταβλητών κατηγοριοποιεί τις μεταβλητές ως σημαντικές και μη σημαντικές. Μεγάλη μεροληψία στην εκτίμηση μπορεί να δημιουργηθεί από την απαλοιφή των μη σημαντικών μεταβλητών. Είναι καλύτερα να χρησιμοποιηθεί λίγο από όλες τις μεταβλητές απ΄ ότι να χρησιμοποιηθούν κάποιες εξ ολοκλήρου και κάποιες καθόλου. Αυτό κάνουν οι μεροληπτικές εκτιμήτριες (Marquardt and Snee, 1975).

Η μέθοδος της ανάλυσης κορυφογραμμής (Ridge Regression) εισήχθηκε από τους Hoerl και Kennard (1970) κυρίως με σκοπό να αντιμετωπίσουν την πολυσυγγραμμικότητα. Αυτοί επικέντρωσαν την προσοχή τους στις ιδιοτιμές του πίνακα $\mathbf{X'X}$. Ένα σοβαρό πρόβλημα πολυσυγγραμμικότητας χαρακτηρίζεται από το γεγονός ότι η μικρότερη ιδιοτιμή του πίνακα $\mathbf{X'X}$ είναι κατά πολύ μικρότερη της μονάδας. Επεσήμαναν επίσης τη δραματική ανεπάρκεια των ε.ε.τ. για μη ορθογώνια προβλήματα. 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. Βέβαια οι ε.ε.τ. είναι αμερόληπτες. Έχει όμως περισσότερο νόημα να επιτύχουμε μικρότερο μέσο τετραγωνικό σφάλμα, αν μπορεί να επιτευχθεί μεγαλύτερη μείωση στη διασπορά επιτρέποντας λίγη μεροληψία. Αλλά μεγαλύτερη διασπορά σημαίνει μικρότερη ακρίβεια πρόβλεψης. Συνεπώς ένας λόγος που πιθανόν ο αναλυτής δεν είναι ικανοποιημένος με τους ε.ε.τ. είναι σχετικά με την ακρίβεια πρόβλεψης. Για το λόγο αυτό, όπως έχει ήδη αναφερθεί, ένας τρόπος προκειμένου να βελτιωθεί η ακρίβεια πρόβλεψης, είναι να συρρικνώσουμε κάποιους συντελεστές. Είναι προτιμότερο να κρατήσουμε κάποιες πληροφορίες από όλες τις μεταβλητές παρά να κρατήσουμε μόνο μερικές από τις μεταβλητές και κάποιες άλλες να τις απορρίψουμε τελείως. Ακριβώς αυτή είναι η φιλοσοφία της μεθόδου της κορυφογραμμής.

Η εκτιμήτρια με τη μέθοδο της κορυφογραμμής της παραμέτρου $\beta = (\beta_1,\ \beta_2, \ldots, \beta_p)$, προκύπτει λύνοντας την

$$(\mathbf{X'X} + \lambda\mathbf{I})\hat{\beta}^* = \mathbf{X'y} \Rightarrow$$

$$\hat{\beta}_{\mathbf{ridge}} = (\mathbf{X'X} + \lambda\mathbf{I})^{-\mathbf{1}}\mathbf{X'y}. \tag{2.5}$$

όπου λ είναι μια ρυθμιστική παράμετρος. Ισοδύναμα οι εκτιμήτριες των συντελεστών β_j

υπολογίζονται έτσι ώστε να ελαχιστοποιείται

$$\hat{\beta} = \arg\min \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \tag{2.6}$$

$$\text{υπό τον περιορισμό} \qquad \sum_{j=1}^{p} \beta_j^2 \leq t,$$

όπου t είναι μια ρυθμιστική παράμετρος. Είναι εμφανής από τις παραπάνω σχέσεις η διαφοροποίηση της μεθόδου σε σχέση με τις ε.ε.τ.. Η εμφάνιση του όρου λI στη (2.5) ή του περιορισμού στη (2.6) είναι αυτή που δικαιολογεί την έννοια της 'ποινής' που επιβάλλει η μέθοδος αυτή στην εκτίμηση των συντελεστών. Γενικά, υπάρχει μια βέλτιστη τιμή του λ ή ισοδύναμα του t, για κάθε πρόβλημα, αλλά είναι καλό να εξετάσουμε τη λύση που δίνει η μέθοδος της κορυφογραμμής για ένα εύρος αποδεκτών τιμών. Αποδεκτή τιμή σημαίνει τιμή για την οποία η αντίστοιχη εκτιμήτρια έχει μικρότερο μέσο τετραγωνικό σφάλμα απ' ότι η αντίστοιχη ε.ε.τ.. Το μέσο τετραγωνικό σφάλμα των μελλοντικών προβλέψεων επίσης μειώνεται αντίστοιχα. Εφόσον γνωρίζουμε ότι η συνηθισμένη εκτιμήτρια $\hat{\beta} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ είναι αμερόληπτη, είναι φανερό από τη σχέση (2.5) ότι η $\hat{\beta}_{ridge}$ είναι μεροληπτική. Από την άλλη, η διασπορά $V(\hat{\beta}_{ridge})$ είναι μικρότερη της $V(\hat{\beta})$. Κατά συνέπεια το μέσο τετραγωνικό σφάλμα της $\hat{\beta}_{ridge}$ μπορεί να είναι πολύ μικρότερο αυτού της $\hat{\beta}$.

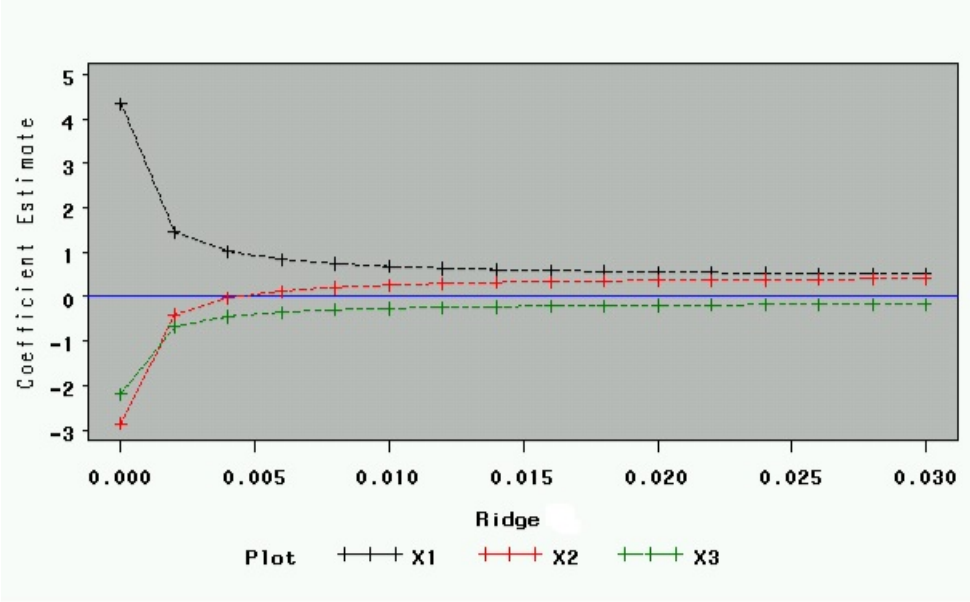
Όπως αναφέρουν οι Marquardt και Snee (1975), για ένα σχεδόν ορθογώνιο μοντέλο, στο οποίο οι μεταβλητές έχουν τυποποιηθεί, όλοι οι συντελεστές έχουν περίπου ίσες διασπορές. Έτσι, μπορεί να γίνει επιλογή μεταβλητών με βάση την απόλυτη τιμή τους. Μπορούν να παραλειφθούν λοιπόν οι μεταβλητές των οποίων οι αντίστοιχοι συντελεστές είναι πολύ μικροί κατά απόλυτη τιμή. Είναι λοιπόν εμφανές στο σημείο αυτό πως η μέθοδος της κορυφογραμμής μπορεί να χρησιμοποιηθεί τόσο ως μέθοδος εκτίμησης των συντελεστών σε ένα μοντέλο όσο και ως μέθοδος επιλογής μεταβλητών. Η $B.S.$ είναι μια καλή στρατηγική όταν οι υποψήφιες μεταβλητές είναι μεταξύ τους (σχεδόν) ορθογώνιες. Είναι επίσης καλή στρατηγική αν οι μεταβλητές έχουν γίνει (σχεδόν) ορθογώνιες αποτελεσματικά επιτρέποντας μεροληψία στις εκτιμήτριες. Η μόνη άλλη περίπτωση είναι όταν η επιλογή μεταβλητών γίνεται με γνώμωνα περισσότερες πληροφορίες σχετικά με τη φύση του προβλήματος ή τις ιδιότητες των μεταβλητών. Αυτό σημαίνει ότι διαθέτουμε εξωτερική πληροφορία, δηλαδή ανεξάρτητη από τις αριθμητικές τιμές των δεδομένων.

Η $B.S.$ αποτελεί κακή στρατηγική όταν οι υποψήφιες μεταβλητές εμφανίζουν μεγάλη συσχέτιση. Στην πράξη, η μεγάλη τιμή του VIF αποσταθεροποιεί εντελώς όλα τα κριτήρια που μπορεί κάποιος να υπολογίσει από τις ε.ε.τ., καταλήγοντας έτσι σε πολύ ασταθή επιλογή μεταβλητών. Η $B.S.$ είναι επίσης κακή τακτική στην περίπτωση που στις υποψήφιες μεταβλητές περιέχονται και οι καμπυλόγραμμοι μετασχηματισμοί τους (πχ τετράγωνα).

Ridge Trace (Ίχνος Κορυφογραμμής)

Ένα μεγάλο πλεονέκτημα της ανάλυσης κορυφογραμμής είναι ότι ένα γράφημα, το λεγόμενο γράφημα ίχνους κορυφογραμμής, μπορεί να βοηθήσει τον αναλυτή να διαπιστώσει ποιοι συντελεστές είναι ευαίσθητοι στα δεδομένα. Συνεπώς, η ανάλυση ευαισθησίας είναι ένας σκοπός της ανάλυσης κορυφογραμμής. Το ίχνος κορυφογραμμής είναι το γράφημα της τιμής κάθε συντελεστή έναντι του λ. Το γράφημα θα έχει μια καμπύλη (ίχνος) για κάθε συντελεστή. Για να είναι πιο ξεκάθαρο, προτείνεται να μη σχεδιάζονται περισσότεροι από 10 συντελεστές στο ίδιο γράφημα. Ο στόχος είναι να βρούμε μια τιμή του λ η οποία δίνει ένα σύνολο συντελεστών με μικρότερο μέσο τετραγωνικό σφάλμα από αυτό των ε.ε.τ. Φυσικά, καθώς το λ αυξάνει, το άθροισμα των τετραγώνων των υπολοίπων (SSE) επίσης αυξάνεται. Αυτό δεν είναι ιδιαίτερα ανησυχητικό διότι ο στόχος δεν είναι να αποκτήσουμε ένα μοντέλο που να προσαρμόζεται όσο το δυνατόν καλύτερα στα δεδομένα, αλλά να αναπτύξουμε ένα 'ευσταθές' σύνολο συντελεστών οι οποίοι θα εκτιμούν αποτελεσματικά μελλοντικές παρατηρήσεις. Με τη λέξη 'ευσταθείς' εννοούμε συντελεστές οι οποίοι δεν είναι ευαίσθητοι σε μικρές αλλαγές στα δεδομένα. Αν οι μεταβλητές πρόβλεψης εμφανίζουν μεγάλες συσχετίσεις, οι συντελεστές θα αλλάζουν γρήγορα για μικρές τιμές του λ και σταδιακά θα σταθεροποιούνται (αλλάζουν λίγο) για μεγαλύτερες τιμές. Η τιμή του λ για την οποία οι συντελεστές έχουν σταθεροποιηθεί δίνει τους συντελεστές. Αν οι μεταβλητές πρόβλεψης είναι ορθογώνιες τότε οι συντελεστές θα αλλάζουν ελάχιστα (δηλ. θα είναι ήδη σταθεροί) αναδεικνύοντας έτσι ότι οι ε.ε.τ. δίνουν ένα καλό σύνολο συντελεστών για το μοντέλο. Πρέπει βέβαια εδώ να τονίσουμε ότι η βέλτιστη τιμή του λ είναι άγνωστη και μπορεί μόνο να εκτιμηθεί. Ένας εναλλακτικός τρόπος υπολογισμού του βέλτιστου λ είναι με τη μέθοδο του crossvalidation που αναλύουμε σε επόμενη παράγραφο (2.7). Το Σχήμα 2.1 αποτελεί ένα παράδειγμα για το ίχνος κορυφογραμμής σε ένα μοντέλο με τρεις μεταβλητές $X_1,\ X_2,\ X_3$. Καθώς η τιμή της

ρυθμιστικής παραμέτρου αυξάνεται οι συντελεστές των μεταβλητών μεταβάλλονται. Περίπου από την τιμή 0.02 και μετά οι συντελεστές σχεδόν σταθεροποιούνται. Συνεπώς μπορούμε να συμπεράνουμε ότι η βέλτιστη τιμή της παραμέτρου είναι $\lambda = 0.02$.



Σχήμα 2.1: Ridge Trace

2.3 LASSO

Ένα βασικό μειονέκτημα της μεθόδου της κορυφογραμμής, ειδικά στην περίπτωση που έχουμε πολλές συμμεταβλητές, είναι το γεγονός ότι λόγω του μεγάλου αριθμού των παραμέτρων το μοντέλο πιθανόν να μην είναι εύκολα ερμηνεύσιμο. Τα κριτήρια και οι τεχνικές που περιγράφουμε στοχεύουν στο να προσδιορίσουμε ένα μικρότερο υποσύνολο παραμέτρων το οποίο να συγκεντρώνει τις μεταβλητές που επιδρούν πιο έντονα στην εξαρτημένη μεταβλητή. Από την άλλη, όπως αναφέραμε παραπάνω, η $B.S$ δίνει μοντέλα με λιγότερες μεταβλητές και συνεπώς πιο ερμηνεύσιμα αλλά είναι εξαιρετικά ευμετάβλητη διαδικασία.

Για την αντιμετώπιση των παραπάνω προβλημάτων, ο Tibshirani (1996) πρότεινε μια νέα μέθοδο, τη λεγόμενη LASSO (Least Absolute Shrinkage and Selection Operator) ή L_1. Η μέθοδος αυτή συρρικνώνει κάποιους συντελεστές και άλλους τους θέτει ίσους με μηδέν, συνδυάζοντας έτσι τα θετικά στοιχεία των μεθόδων $B.S.$ και της κορυφογραμμής. Αποτελεί

και αυτή μια μέθοδο με ποινή.

Έστω ότι έχουμε δεδομένα (x_i, y_i) όπου $x_i = (x_{i1}, \ldots, x_{ip})'$ και $y_i,\ i = 1, 2, \ldots, n$ είναι οι τιμές που αντιστοιχούν στην i παρατήρηση στο πλαίσιο του συνήθους γραμμικού μοντέλου. Για την εκτίμηση της παραμέτρου $\beta = (\beta_1,\ \beta_2, \ldots, \beta_p)$, η εκτιμήτρια με τη μέθοδο LASSO ορίζεται ως

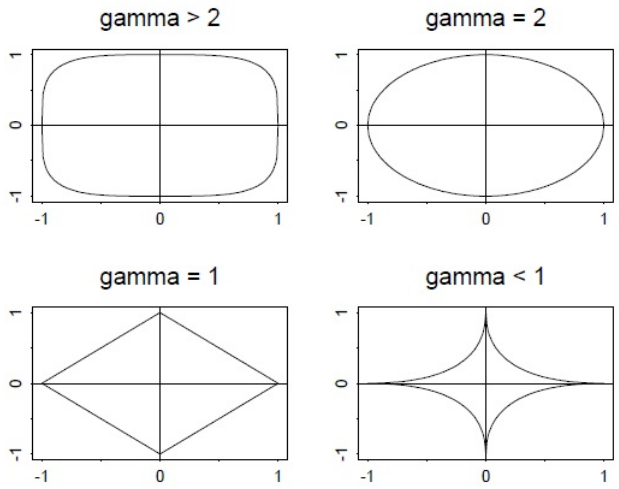
$$\begin{aligned} (\hat{\alpha}, \hat{\beta}) &= \arg\min \left\{ \left(y_i - a - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \qquad (2.7) \\ \text{υπό τον περιορισμό} \qquad & \sum_{j=1}^{p} |\beta_j| \leq t. \end{aligned}$$

Στον παραπάνω ορισμό η παράμετρος t αποτελεί μια ρυθμιστική παράμετρο. Επιπλέον, για οποιαδήποτε τιμή του t, η λύση της (2.7) ως προς α είναι $\hat{\alpha} = \overline{y}$. Χωρίς βλάβη της γενικότητας μπορούμε να υποθέσουμε ότι $\overline{y} = 0$ και συνεπώς να παραλείψουμε το α. Τότε από τον παραπάνω ορισμό προκύπτει ότι οι εκτιμήτριες των συντελεστών του γραμμικού μοντέλου με τη μέθοδο $LASSO$ δίνονται από τη λύση του συστήματος

$$\begin{aligned} \hat{\beta} &= \arg\min \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \qquad (2.8) \\ \text{υπό τον περιορισμό} \qquad & \sum_{j=1}^{p} |\beta_j| \leq t. \end{aligned}$$

Η παράμετρος $t \geq 0$ ελέγχει το ποσό της συρρίκνωσης που υφίστανται οι συντελεστές. Έστω ότι με $\hat{\beta}_j^0$ συμβολίζουμε τις συνήθεις εκτιμήτριες ελαχίστων τετραγώνων και έστω $t_0 = \sum |\hat{\beta}_j^0|$. Τιμές της ρυθμιστικής παραμέτρου $t < t_0$ θα προκαλέσουν συρρίκνωση των λύσεων προς το 0 και πιθανόν κάποιοι συντελεστές να γίνουν ακριβώς ίσοι με μηδέν. Για παράδειγμα, αν $t = t_0/2$, το αποτέλεσμα θα είναι περίπου παρόμοιο με το να βρίσκαμε το βέλτιστο σύνολο μεταβλητών μεγέθους $p/2$. Σημειώνουμε επίσης ότι ο πίνακας σχεδιασμού δεν πρέπει να είναι κατ' ανάγκη πλήρους τάξης.

Αξίζει να σημειωθεί ότι οι μέθοδοι L_1 και L_2 αποτελούν ειδικές περιπτώσεις της γενικότερης παλινδρόμησης Bridge που εισήχθει από τους Frank και Friedman (1993). Για τον προσδιορισμό των συντελεστών αρκεί να λύσουμε ένα από τα παρακάτω ισοδύναμα προβλήματα.



Σχήμα 2.2: Παλινδρόμηση bridge με $t = 1$

$$\min \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \tag{2.9}$$

$$\text{υπό τον περιορισμό} \qquad \sum_{j=1}^{p} |\beta_j|^\gamma \leq t.$$

ή ισοδύναμα

$$\min \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^\gamma \right\}. \tag{2.10}$$

Το πρόβλημα (2.9) αναφέρεται ως παλινδρόμηση υπό περιορισμό ενώ το (2.10) ως ποινικοποιημένη παλινδρόμηση. Για $\gamma = 1$ η μέθοδος αντιστοιχεί στη $LASSO$ ενώ για $\gamma = 2$ στη Ridge. Στο γράφημα 2.2 (Fu, 1998) φαίνεται η μορφή του περιορισμού για τις διάφορες τιμές της παραμέτρου γ.

Η ιδέα για τη μέθοδο $LASSO$ προήλθε από μια παρόμοια μέθοδο που πρότεινε ο Breiman (1995) η οποία καλείται non-negative garrote (μη αρνητικός στραγγαλισμός). Σύμφωνα με τη μέθοδο αυτή, οι συντελεστές β_j έχουν τη μορφή $\hat{\beta}_j = c_j \ \hat{\beta}_j^0$ όπου οι συντελεστές c_j υπολογίζονται έτσι ώστε να ελαχιστοποιείται

$$\sum_{i=1}^{n} \left(y_i - \alpha - \sum_{j=1}^{p} c_j \hat{\beta}_j^0 x_{ij} \right)^2 \quad \text{υπό τον περιορισμό} \quad \sum_{j=1}^{p} c_j \leq t, \ c_j \geq 0. \tag{2.11}$$

Η μέθοδος garrote ξεκινάει με τις συνήθεις ε.ε.τ. και τις συρρικνώνει κατά κάποιους μη-αρνητικούς παράγοντες των οποίων το άθροισμα είναι φραγμένο. Ο Breiman έδειξε ότι η μέθοδός του έχει μικρότερο σφάλμα πρόβλεψης από την $B.S$ και ανταγωνίζεται την ανάλυση κορυφογραμμής εκτός από τις περιπτώσεις που το πραγματικό μοντέλο έχει πολλούς μικρούς μη-μηδενικούς συντελεστές. Ένα μειονέκτημα της μεθόδου είναι ότι η λύση της εξαρτάται τόσο από το πρόσημο όσο και από το μέγεθος των ε.ε.τ. Σε περιπτώσεις όπως για παράδειγμα όταν έχουμε συσχετίσεις μεταξύ των επεξηγηματικών μεταβλητών όπου οι ε.ε.τ. δε συμπεριφέρονται καλά, το γεγονός αυτό πιθανότατα να έχει ως συνέπεια να προκύψουν κακές εκτιμήσεις και με τη μέθοδο garrote. Αντίθετα, η LASSO δεν κάνει χρήση των ε.ε.τ.

Προκειμένου να αποκτήσουμε μια πιο πρακτική αντίληψη των μεθόδων που έχουμε αναφέρει, θα δούμε, όπως αναφέρεται στον Tibshirani (1996), τη μορφή που αυτοί αποκτούν στην περίπτωση που ο πίνακας σχεδιασμού είναι ορθοκανονικός, δηλαδή $X'X = I$ όπου I ο ταυτοτικός πίνακας. Στην περίπτωση αυτή

(a) 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 μηδέν. Για κάποια επιλογή του λ αυτό είναι ισοδύναμη με το να θέσουμε

$$\hat{\beta}_j^{BS} = \begin{cases} \hat{\beta}_j^0, & |\hat{\beta}_j^0| > \lambda \\ 0, & \text{διαφορετικά.} \end{cases} \tag{2.12}$$

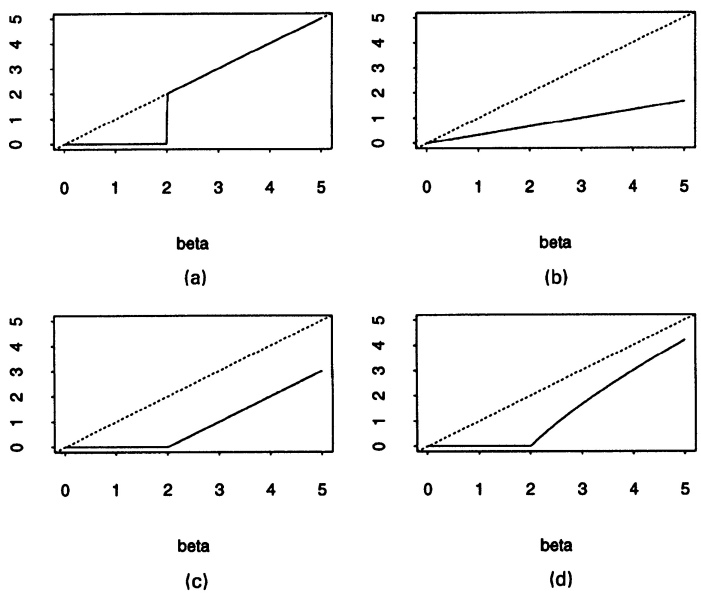
(b) Στη μέθοδο της ανάλυσης κορυφογραμμής (L_2) οι εκτιμήτριες των συντελεστών β_j υπολογίζονται από τη λύση του συστήματος (2.6). Οι λύσεις του παραπάνω συστήματος είναι

$$\hat{\beta}_j^{L_2} = \frac{1}{1+\gamma}\,\hat{\beta}_j^0.$$

(c) Στη μέθοδο $LASSO$ (L_1) οι λύσεις των εξισώσεων (2.8) στη περίπτωση ορθοκανονικού πίνακα σχεδιασμού είναι

$$\hat{\beta}_j^{L_1} = sign(\hat{\beta}_j^0)(|\hat{\beta}_j^0 - \gamma|)^+$$

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



Σχήμα 2.3: (a) BS, (b)L2, (c) L1, (d) Garrote, μορφή συρρίκνωσης συντελεστών στην περίπτωση ορθοκανονικού πίνακα σχεδιασμού

(d) Οι εκτιμήτριες με τη μέθοδο garrote είναι

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

Στο σχήμα (2.3) φαίνεται η μορφή αυτών των συναρτήσεων. Η L_2 αλλάζει την κλίμακα στους συντελεστές κατά ένα σταθερό παράγοντα ενώ η L_1 κάνει σε μερικούς αποκοπή στο 0 και άλλους τους συρρικνώνει κατά ένα σταθερό παράγοντα. Οι συντελεστές της garrote μοιάζουν πολύ με τη $LASSO$ αλλά για τους μεγαλύτερους συντελεστές η συρρίκνωση που υφίστανται είναι μικρότερη. Οι διαφορές μεταξύ garrote και LASSO μπορεί να γίνουν πολύ μεγαλύτερες στην περίπτωση μη ορθοκανονικού πίνακα σχεδιασμού. Η διακεκομμένη γραμμή αντιστοιχεί σε γωνία 45^o δηλαδή στις ε.ε.τ..

Ας υποθέσουμε ότι έχουμε δύο επεξηγηματικές μεταβλητές. Χωρίς βλάβη της γενικότητας υποθέτουμε ότι οι δύο αντίστοιχες ε.ε.τ. $\hat{\beta}_j^0$ είναι θετικοί. Εύκολα μπορούμε να δείξουμε ότι οι εκτιμήτριες με τη μέθοδο L_1 δίνονται ως

$$\hat{\beta}^{L_1} = \left(\hat{\beta}_j^0 - \gamma\right)^+$$

όπου το γ υπολογίζεται έτσι ώστε $\hat{\beta}_1 + \hat{\beta}_2 = t$ για $t \leq \hat{\beta}_1^0 + \hat{\beta}_2^0$ και ισχύει ακόμα κι αν οι μεταβλητές παρουσιάζουν συσχέτιση. Λύνοντας ως προς γ παίρνουμε

$$\begin{aligned} \hat{\beta}_1 &= \left(\frac{t}{2} + \frac{\hat{\beta}_1^0 - \hat{\beta}_2^0}{2} \right)^+ \\ \hat{\beta}_2 &= \left(\frac{t}{2} - \frac{\hat{\beta}_1^0 - \hat{\beta}_2^0}{2} \right)^+ . \end{aligned}$$

Αντίθετα με τις L_1 εκτιμήτριες, η μορφή των εκτιμητριών L_2 εξαρτάται από τη συσχέτιση των επεξηγηματικών μεταβλητών.

Συμπερασματικά, σε προσομοιώσεις που πραγματοποιήθηκαν, όπως αυτές αναφέρονται μετά από προσομειώσεις που έγιναν από τον Tibshirani (1996), εξετάστηκε συγκριτικά η ικανότητα πρόβλεψης των τεσσάρων μεθόδων κάτω από τρία διαφορετικά σενάρια

(i) μικρός αριθμός μεταβλητών που όλες επιδρούν σημαντικά στην εξαρτημένη μεταβλητή: Στην περίπτωση αυτή η επιλογή υποσυνόλου μεταβλητών BS είναι η καλύτερη, η LASSO όχι τόσο καλά ενώ η Παλινδρόμηση Κορυφογραμμής έδωσε τα χειρότερα αποτελέσματα.

(ii) μέτριος αριθμός μεταβλητών που όλες επιδρούν μέτρια στην εξαρτημένη μεταβλητή: Στην περίπτωση αυτή η LASSO έδωσε τα καλύτερα αποτελέσματα και ακολούθησαν η μέθοδος της Κορυφογραμμής και τέλος η BS.

(iii) μεγάλος αριθμός μεταβλητών που καμία δεν επιδρά σημαντικά στην εξαρτημένη μεταβλητή: Στην περίπτωση αυτή η Παλινδρόμηση Κορυφογραμμής έδωσε τα καλύτερα αποτελέσματα με διαφορά ενώ ακολούθησαν η LASSO και τελευταία η BS.

Η garrote τα κατάφερε λίγο καλύτερα από τη $LASSO$ στην πρώτη περίπτωση, και λίγο χειρότερα στις άλλες δύο περιπτώσεις. Σημειώνουμε ότι η αξιολόγηση της ικανότητας πρόβλεψης έγινε με χρήση της cvl (Παράγραφος 2.7).

Η LASSO μπορεί να εφαρμοστεί και στην περίπτωση των γενικευμένων γραμμικών μοντέλων. Θα δούμε πώς αυτή εφαρμόζεται στα πλαίσια της Ανάλυσης Επιβίωσης στο μοντέλο του Cox ή στη λογιστική παλινδρόμηση.

2.4 LASSO και Ridge στο μοντέλο του Cox και τη λογιστική παλινδρόμηση

Στο σύνηθες πλαίσιο των δεδομένων επιβίωσης, έστω ότι έχουμε n παρατηρήσεις της μορφής (t_i, d_i, x_i), όπου t_i είναι ο πιθανόν αποκομμένος χρόνος επιβίωσης, d_i είναι ο δείκτης αποκοπής, δηλαδή ο δείκτης d_i λαμβάνει τιμές 0 ή 1 ανάλογα με το αν η παρατήρηση της μονάδας i είναι αποκομμένη ή όχι και x_i είναι το σύνηθες διάνυσμα γραμμή των συμμεταβλητών $X_1, X_2, \ldots, X_p$ για τη μονάδα i. Όπως ήδη γνωρίζουμε η εκτίμηση των παραμέτρων $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ στο μοντέλο του Cox γίνεται μέσω της μεγιστοποίησης της συνάρτησης μερικής πιθανοφάνειας (1.22) όπως περιγράψαμε στην παράγραφο 1.2. Ο Tibshirani (1997) πρότεινε τη μέθοδο LASSO να εφαρμοστεί και στο μοντέλο του Cox. Η εκτίμηση του β πρότεινε να γίνει μέσω του κριτηρίου

$$\beta = \arg\max \ell(\beta) \quad \text{υ.π.} \quad \sum |\beta_j| \leq s \tag{2.13}$$

όπου η $\ell(\beta)$ αντιστοιχεί στη λογαριθμοποιημένη συνάρτηση μερικής πιθανοφάνειας (1.22) και s είναι μία παράμετρος που καθορίζεται από το χρήστη.

Έστω $\hat{\beta}_j^0$ είναι οι εκτιμήτριες όπως προκύπτουν από τη λύση του συστήματος των εξισώσεων (1.24). Τότε αν $s \geq \sum |\hat{\beta}_j^0|$, τότε η επίλυση της (2.13) δίνει τις ίδιες εκτιμήτριες με αυτούς που προκύπτουν από τη μεγιστοποίηση της μερικής πιθανοφάνειας. Αντίθετα, αν $s < \sum |\hat{\beta}_j^0|$, τότε η λύση της (2.13) δίνει συντελεστές συρρικνωμένους προς το μηδέν. Ένα ελκυστικό χαρακτηριστικό της μεθόδου είναι ότι αρκετά συχνά κάποιοι συντελεστές είναι ακριβώς μηδέν. Αυτό βέβαια βοηθάει στη δημιουργία τελικά ενός πιο ερμηνεύσιμου μοντέλου. Η ομαλή μορφή του περιορισμού θα πρέπει να δώσει ένα πιο σταθερό τελικό μοντέλο απ΄ ότι η $B.S.$ ή η διαδικασία σε βήματα. Αυτό έχει επιβεβαιωθεί από τον Tibshirani (1997) μέσω προσομοιώσεων στα πλαίσια των μοντέλων παλινδρόμησης. Σημειώνουμε ότι η μέθοδος αυτή αποτελεί ένα εργαλείο ώστε να επιτύχουμε ένα οικονομικότερο μοντέλο (με λιγότερες μεταβλητές).

Ομοίως, η εφαρμογή της Παλινδρόμησης Κορυφογραμμής στο μοντέλο αναλογικής διακινδύνευσης του Cox οδηγεί στην επίλυση του συστήματος ανάλογου με αυτό που περιγράφεται από τη σχέση (2.13) η οποία τροποποιείται μόνο ως προς τον περιορισμό. Το πρόβλημα δη-

λαδή παίρνει τη μορφή

$$\beta = \arg\max \ell(\beta) \quad \text{υ.π.} \quad \sum \beta_j^2 \leq s, \tag{2.14}$$

όπου $\ell(\beta)$ η λογαριθμοποιημένη συνάρτηση μερικής πιθανοφάνειας. Το μέγεθος της συρρίκνωσης που προκαλείται από την εφαρμογή της μεθόδου εξαρτάται από την τιμή της παραμέτρου s αλλά κανένας συντελεστής ποτέ δεν μηδενίζεται.

Στην περίπτωση της λογιστικής παλινδρόμησης, η παλινδρόμηση κορυφογραμμής πραγματοποιείται μεγιστοποιώντας την ποινικοποιημένη συνάρτηση λογαριθμοποιημένης πιθανοφάνειας

$$\ell^\lambda(\beta) = \ell(\beta) - \lambda \sum_j \beta_j^2 \tag{2.15}$$

όπου $\ell(\beta)$ η απλή λογαριθμοποιημένη συνάρτηση πιθανοφάνειας όπως αυτή δίνεται από τη σχέση (1.15). Η πιο απλή ειδική περίπτωση που $n_i = 1$, αναλύεται από τους Cessie και Houwelingen (1992). Τότε η $\ell(\beta)$ απλοποιείται στη μορφή

$$\ell(\beta) = \sum_i y_i \ln p_i + (1 - y_i) \ln(1 - p_i). \tag{2.16}$$

Από τη μεγιστοποίηση της τελευταίας ως προς β παίρνουμε τις συνήθεις εκτιμήτριες μέγιστης πιθανοφάνειας. Έστω $\hat{\beta}^\lambda$ εκτιμήτριες της παλινδρόμησης κορυφογραμμής που προκύπτουν από τη μεγιστοποίηση της (2.15). Από τη ρυθμιστική παράμετρο λ ελέγχεται το μέγεθος της συρρίκνωσης που υφίστανται οι συντελεστές παλινδρόμησης. Όταν $\lambda = 0$ τότε η λύση είναι ακριβώς οι συνήθεις ε.μ.π. ενώ αν $\lambda \to \infty$ τότε όλοι οι συντελεστές παλινδρόμησης τείνουν στο μηδέν. Μεγάλος αριθμός επεξηγηματικών μεταβλητών και/ή μεγάλη συσχέτιση μεταξύ τους οδηγεί σε ασταθείς εκτιμήτριες όπως έχουμε δει γενικά. Συρρικνώνοντας τους συντελεστές προς το μηδέν και επιτρέποντας λίγη μεροληψία, σταθεροποιεί το σύστημα και δίνει εκτιμήτριες με μικρότερη διασπορά. Για το λόγο αυτό μια καλή επιλογή της ρυθμιστικής παραμέτρου λ αναμένεται να δώσει εκτιμήτριες $\hat{\beta}^\lambda$ οι οποίες να είναι πιο κοντά στις τιμές των πραγματικών συντελεστών β_j. Δηλαδή περιμένουμε $MSE(\hat{\beta}^\lambda) < MSE(\hat{\beta})$. Σημειώνεται ότι οι εκτιμήτριες $\hat{\beta}^\lambda$ προκύπτουν (όπως και οι συνήθεις ε.μ.π.) με αριθμητικές μεθόδους από την επίλυση του συστήματος των μερικών παραγώγων ως προς β_j της (2.15).

Η επιλογή της παραμέτρου λ γίνεται έτσι ώστε να ελαχιστοποιείται το σφάλμα πρόβλεψης. Έστω ότι διαθέτουμε τις εκτιμήτριες $\hat{\beta}$ όπως αυτές προέκυψαν από το δείγμα των παρατηρήσεων $(x_i, y_i),\ i = 1, 2, \ldots, n$ και $\hat{p}(x)$ η εκτίμηση της πιθανότητας επιτυχίας $(P(Y = 1))$.

Για μια καινούρια παρατήρηση με διάνυσμα συμμεταβλητών x_{new}, η πιθανότητα $Y_{new} = 1$ εκτιμάται από την $\hat{p} = \hat{p}(x_{new})$ όταν η πραγματική πιθανότητα είναι p. Αναφέρονται (Cessie and Houwelingen, 1992) τρεις τρόποι προκειμένου να μετρηθεί το σφάλμα πρόβλεψης:

α. Σφάλμα ταξινόμησης (Classification/Counting Error)

$$CE = \begin{cases} 1, \text{ αν } Y_{new} = 1 \text{ και } \hat{p} < \frac{1}{2} \\ \quad \text{ή } Y_{new} = 0 \text{ και } \hat{p} > \frac{1}{2}, \\ \frac{1}{2}, \text{ αν } \hat{p} = \frac{1}{2}, \\ 0, \text{ διαφορετικά.} \end{cases} \tag{2.17}$$

β. Τετραγωνικό σφάλμα (Squared Error)

$$SE = (Y_{new} - \hat{p})^2. \tag{2.18}$$

γ. Σφάλμα της μείον λογαριθμοποιημένης πιθανοφάνειας (Minus Log-likelihood Error)

$$ML = -\{Y_{new} \ln \hat{p} + (1 - Y_{new}) \ln(1 - \hat{p})\}. \tag{2.19}$$

Η μέση τιμή και των τριών σφαλμάτων μεγιστοποιείται όταν η πιθανότητα p είναι κοντά στο $1/2$ και τείνει στο μηδέν αν η p τείνει στο 0 ή στο 1. Αυτό συνάδει με τη διαισθητική ερμηνεία ότι η τιμή της Y είναι πιο δύσκολη να προβλεφθεί αν η πιθανότητα $Y = 1$ είναι γύρω από το $1/2$. Σχετικά με το πιο από τα τρία παραπάνω μέτρα θα επιλέξουμε, εξαρτάται κυρίως από το είδος του μοντέλου που χρησιμοποιείται για πρόβλεψη. Το σφάλμα ταξινόμησης αποτελεί έναν κανόνα διαχωρισμού των παρατηρήσεων και είναι ευαίσθητο στην πρόβλεψη κοντά στο $1/2$. Τα άλλα δύο μέτρα λαμβάνουν υπόψη τις προβλέψεις του μοντέλου σε όλο το εύρος των τιμών της p. Το τετραγωνικό σφάλμα είναι διαισθητικά ελκυστικό με την έννοια ότι αντιστοιχεί στην ευκλείδεια απόσταση μεταξύ της νέας παρατήρησης Y_{new} και της εκτίμησης $\hat{p}$ και είναι ακριβώς ανάλογο με το τετραγωνικό σφάλμα στο σύνηθες γραμμικό μοντέλο. Το τελευταίο μέτρο ML ισούται με $-\ln \hat{p}$ αν $Y_{new} = 1$ και με $-\ln(1 - \hat{p})$ αν $Y_{new} = 0$. Παρατηρούμε ότι αν αθροίσουμε την ποσότητα ML πάνω σε όλες τις παρατηρήσεις προκύπτει η $-\ell(\beta)$ όπως δίνεται από την (2.16). Επίσης η ποσότητα $2 \times ML$ αντιστοιχεί στη Deviance. Το πλεονέκτημα του σφάλματος ML είναι ότι σχετίζεται με τη λογαριθμοποιημένη πιθανοφάνεια και επιπλέον δεν περιορίζεται μόνο στη λογιστική παλινδρόμηση αλλά μπορεί

να εφαρμοστεί και σε γενικότερο πλαίσιο. Η επιλογή της βέλτιστης τιμής της ρυθμιστικής παραμέτρου λ για την εφαρμογή της μεθόδου κορυφογραμμής στην περίπτωση της λογιστικής παλινδρόμησης πρέπει να γίνει έτσι ώστε να ελαχιστοποιείται το μέσο σφάλμα πρόβλεψης. Η εκτίμηση του μέσου σφάλματος γίνεται με τη μέθοδο της cvl (βλ. ενότητα 2.7).

2.5 Γενικεύσεις της LASSO

Όπως έχουμε ήδη αναφέρει, οι πιο γνωστές και συχνότερα χρησιμοποιούμενες μέθοδοι επιλογής μεταβλητών, είναι η κατά βήματα απαλοιφή (stepwise deletion) και η μέθοδος επιλογής καλύτερου υποσυνόλου (B.S.). Έχουν όμως το μειονέκτημα ότι αγνοούν τα στοχαστικά σφάλματα που εμφανίζονται κατά τη διαδικασία της επιλογής μεταβλητών καθώς και ότι είναι υπολογιστικά χρονοβόρες. Οι Fan και Li (2001), πρότειναν μια καινούρια μεθοδολογία, βασισμένη στα ποινικοποιημένα ελάχιστα τετράγωνα (penalized least squares), η οποία διατηρεί τις καλές ιδιότητες της παλινδρόμησης κορυφογραμμής αλλά και της μεθόδου επιλογής καλύτερου υποσυνόλου. Η μεθοδολογία τους αυτή, επεκτείνεται και σε μοντέλα βασισμένα στην πιθανοφάνεια, όπως π.χ. τα γενικευμένα γραμμικά μοντέλα. Ουσιαστικά τώρα, αυτό που τελικά επιτυγχάνεται, είναι ότι ταυτόχρονα γίνεται και εκτίμηση των παραμέτρων του μοντέλου και μηδενισμός κάποιων, άρα ικανοποιείται ο σκοπός της επιλογής μεταβλητών. Η διαδικασία της ποινικοποίησης, συνίσταται στην εισαγωγή κάποιων συναρτήσεων ποινής (penalty functions), οι οποίες πρέπει να έχουν τις ακόλουθες ιδιότητες:

- Να είναι ιδιάζουσες (singular) στην αρχή ώστε να παράγουν σποραδικές λύσεις δηλαδή πολλοί εκ των εκτιμηθέντων συντελεστών να έχουν τιμή μηδέν.

- Να ικανοποιούν συγκεκριμένες απαιτήσεις ώστε να παράγουν συνεχή μοντέλα (continuous models), οπότε η επιλογή του μοντέλου να χαρακτηρίζεται από σταθερότητα.

- Να φράσσονται από μια σταθερά, ώστε να παράγουν σχεδόν αμερόληπτες εκτιμήτριες για μεγάλους συντελεστές.

Οι μέθοδοι που έχουμε ήδη δει και έχουν αναφερθεί ως ποινικοποιημένες (Κορυφογραμμή, LASSO) αποτελούν μέλη της μεθόδου των ποινικοποιημένων ελαχίστων τετραγώνων, με τη

διαφορά ότι οι σχετικές με τις μεθόδους αυτές, συναρτήσεις ποινής, δεν ικανοποιούν όλες τις προαναφερθείσες απαιτήσεις.

Όπως αναφέραμε και προηγουμένως, η καινούργια μέθοδος επεκτάθηκε και σε μοντέλα βασισμένα στην πιθανοφάνεια. Η διαφορά σε σχέση με τις παραδοσιακές μεθόδους (όπου συνήθως χρησιμοποιείται τετραγωνική συνάρτηση ποινής), είναι ότι οι νέες συναρτήσεις ποινής είναι συμμετρικές, κυρτές στο $(0, \infty)$ και διακατέχονται από ιδιομορφίες στην αρχή. Να σημειωθεί, ότι σε αντίθεση με τις παραδοσιακές μεθόδους επιλογής μεταβλητών, η νέα μέθοδος έχει ισχυρό θεωρητικό υπόβαθρο. Οι εκτιμήτριες ποινικοποιημένης πιθανοφάνειας έχουν καλή απόδοση όσον αφορά την επιλογή του σωστού μοντέλου αρκεί να έχει επιλεγεί σωστά η ρυθμιστική παράμετρος. Σαν να ήταν δηλαδή εξαρχής γνωστό το σωστό υπο-μοντέλο (submodel). Αυτό πρακτικά, σημαίνει ότι όταν οι σωστές παράμετροι του μοντέλου έχουν κάποιες μηδενικές συνιστώσες, αυτές εκτιμώνται από τη μέθοδο ως μηδενικές με πιθανότητα να τείνει στη μονάδα. Ενώ όσον αφορά τις μη μηδενικές συνιστώσες, αυτές εκτιμώνται τόσο καλά όπως όταν είναι γνωστό το σωστό υπο-μοντέλο. Αυτό προφανώς αυξάνει την ακρίβεια εκτίμησης τόσο των μηδενικών όσο και των μη μηδενικών συνιστωσών. Οπότε και υπερτερούν της μεθόδου εκτίμησης μέγιστης πιθανοφάνειας.

Θεωρούμε το απλό γραμμικό μοντέλο (1.2). Αρχικά υποθέτουμε ότι ο πίνακας X είναι ορθοκανονικός. Ο υπολογισμός της ε.ε.τ. γίνεται μέσω της ελαχιστοποίησης της $||y - X\beta||^2$ το οποίο είναι ισοδύναμο με το $||\hat{\beta} - \beta||^2$, όπου $\hat{\beta}$ είναι οι συνήθεις ε.ε.τ. Έστω $z = X'y$ και έστω $\hat{y} = XX'y = Xz$. Μια μορφή των ποινικοποιημένων ελαχίστων τετραγώνων είναι

$$\begin{aligned} &\frac{1}{2}||y - X\beta||^2 + \lambda \sum_{j=1}^{p} p_j(|\beta_j|) \\ = &\frac{1}{2}||y - \hat{y}||^2 + \frac{1}{2}\sum_{j=1}^{p}(z_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} p_j(|\beta_j|). \end{aligned} \tag{2.20}$$

Σημειώνεται ότι οι συναρτήσεις ποινής p_j στην (2.20) δεν είναι απαραίτητα οι ίδιες για όλα τα j. Για παράδειγμα μπορεί να θέλουμε να κρατήσουμε ορισμένες σημαντικές μεταβλητές σε ένα παραμετρικό μοντέλο και για αυτό το λόγο να μη θέλουμε να ποινικοποιήσουμε τις αντίστοιχες παραμέτρους τους. Για ευκολία όμως, θεωρούμε ότι οι συναρτήσεις ποινής είναι οι ίδιες για όλους τους συντελεστές, και θα συμβολίζονται ως $p(|\cdot|)$ ενώ θα χρησιμοποιούμε το συμβολισμό $p_\lambda(|\cdot|)$ δείχνοντας έτσι ότι το $p(|\cdot|)$ εξαρτάται από το λ.

Το πρόβλημα ελαχιστοποίησης (2.20) είναι ισοδύναμο με την ελαχιστοποίηση κατά συνιστώσες. Κατά φυσικό τρόπο οδηγούμαστε λοιπόν στο να θεωρήσουμε το γενικότερο πρόβλημα ποινικοποιημένων ελαχίστων τετραγώνων

$$\frac{1}{2}|z-\theta|^2 + p_\lambda(|\theta|). \tag{2.21}$$

Θεωρώντας τη hard συνάρτηση ποινής κατωφλίου

$$p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda) \tag{2.22}$$

προκύπτει η hard εκτιμήτρια

$$\hat{\theta} = zI(|z| > \lambda). \tag{2.23}$$

Δηλαδή, πιο απλά, η λύση του προβλήματος (2.20) είναι

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

η οποία συμπίπτει με την επιλογή $B.S.$ και τις κατά βήματα διαδικασίες διαδοχικής πρόσθεσης και αφαίρεσης μεταβλητών στην περίπτωση ορθοκανονικού πίνακα σχεδιασμού.

Μια συνάρτηση ποινής για να είναι καλή, πρέπει να δίνει εκτιμήτριες με τις ακόλουθες ιδιότητες:

- Αμεροληψία: Η προκύπτουσα εκτιμήτρια πρέπει να είναι σχεδόν αμερόληπτη, ιδίως στην περίπτωση όπου η σωστή άγνωστη παράμετρος θ είναι μεγάλη. Αποφεύγεται έτσι η μεροληψία του μοντέλου.

- Σποραδικότητα: Η προκύπτουσα εκτιμήτρια πρέπει να αποτελεί κανόνα κατωφλίου (thresholding rule), ώστε οι εκτιμηθέντες συντελεστές με μικρή τιμή, να μηδενίζονται. Έτσι, μειώνεται η πολυπλοκότητα του μοντέλου.

- Συνέχεια: Η προκύπτουσα εκτιμήτρια πρέπει να είναι συνεχής. Αποφεύγεται κατά αυτόν τον τρόπο η αστάθεια στη πρόβλεψη του μοντέλου.

Είναι γνωστό πως η συνάρτηση ποινής L_2

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

οδηγεί στη μέθοδο της παλινδρόμησης κορυφογραμμής. Η LASSO είναι η ποινικοποιημένη εκτιμήτρια ελαχίστων τετραγώνων (π.ε.ε.τ.) με συνάρτηση ποινής την L_1 και η οποία δίνει έναν soft κανόνα κατωφλίου

$$\hat{\theta}_j = sgn(z_j)(z_j - \lambda)^+$$

Η L_q συνάρτηση ποινής

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

οδηγεί στην παλινδρόμηση Bridge (Frank and Friedman, 1993). Η λύση είναι συνεχής μόνο για $q \geq 1$. Παρόλ΄ αυτά, όταν $q > 1$, δεν παράγεται μια σποραδική λύση. Η μόνη συνεχής λύση με κανόνα περιορισμού σε αυτή την οικογένεια συναρτήσεων είναι με τη συνάρτηση ποινής L_1, αυτό όμως προκύπτει μετατοπίζοντας τον εκτιμητή κατά μια σταθερά λ, άρα χάνεται και η αμεροληψία. Επίσης για $0 \leq q < 1$, δεν ικανοποιείται η συνθήκη της συνέχειας.

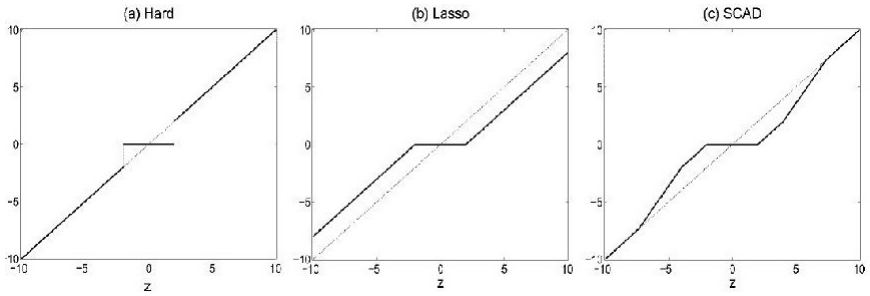
Οι συναρτήσεις ποινής L_q και hard δεν ικανοποιούν και τις τρεις απαιτήσεις της αμεροληψίας, της σποραδικότητας και της συνέχειας. Με σκοπό τη βελτίωση της L_1 και της hard, οι Fan και Li (2001) εισήγαγαν μια συνεχή και διαφορίσιμη συνάρτηση ποινής, τη SCAD (Smoothly Clipped Absolute Deviation penalty) η οποία ορίζεται μέσω της πρώτης παραγώγου της ως

$$p'_\lambda(\theta) = \lambda \left\{ I(\theta \leq \lambda) + \frac{(\alpha\lambda - \theta)}{(\alpha - 1)\lambda} I(\theta > \lambda) \right\} \tag{2.24}$$

για κάποιο $\alpha > 2$ και $\theta > 0$. Η συγκεκριμένη συνάρτηση δεν ποινικοποιεί υπερβολικά τις μεγάλες τιμές του θ και δίνει μια συνεχή λύση, την

$$\hat{\theta} = \begin{cases} sgn(z)(z - \lambda)^+, & |z| \leq 2\lambda \\ \frac{(\alpha-1)z - sgn(z)\alpha\lambda}{\alpha - 2}, & 2\lambda < |z| \leq \alpha\lambda \\ z, & |z| > \alpha\lambda \end{cases} \tag{2.25}$$

Η λύση (2.25) έχει δύο άγνωστες παραμέτρους, α και λ. Στην πράξη θα μπορούσαμε να υπολογίσουμε το βέλτιστο ζεύγος (α, λ) βάσει κάποιων κριτηρίων, όπως της *cvl* και της γενικευμένης *cvl* (παράγραφος 2.7), κάτι που μπορεί να είναι υπολογιστικά χρονοβόρο. Οι Fan και Li (2001), χρησιμοποιώντας εργαλεία Μπεϋζιανά, κατέληξαν στην επιλογή του $\alpha = 3.7$. Στο Σχήμα 2.4 φαίνονται οι τρεις συναρτήσεις κατωφλίου (a) B.S. (Hard), (b)Lasso(Soft), (c) SCAD με $\lambda = 2$ και $\alpha = 3.7$



Σχήμα 2.4: (a) B.S(Hard), (b) Lasso(Soft), (c) SCAD για $\lambda = 2$ και $\alpha = 3.7$

2.6 (Iterative) Sure Independence Screening ((I)SIS)

Προβλήματα επιλογής μεταβλητών σε μοντέλα μεγάλης διάστασης εμφανίζονται πλέον συχνά σε πολλούς επιστημονικούς κλάδους. Στην περίπτωση μεγάλης διάστασης p, δηλαδή μεγάλο πλήθος υποψήφιων συμμεταβλητών, το υπολογιστικό κόστος και η ακρίβεια στην εκτίμηση είναι δύο βασικά στοιχεία για τα οποία πρέπει να μεριμνήσει ο αναλυτής.

Στο κλασικό πρόβλημα του γενικού γραμμικού μοντέλου

$$y = X\beta + \varepsilon, \tag{2.26}$$

όπου $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ είναι το διάνυσμα των παραμέτρων, όταν η διάσταση p είναι μεγάλη, συχνά υποτίθεται ότι μόνο ένας μικρός αριθμός μεταβλητών μεταξύ των $X_1, X_2, \ldots, X_p$ συνεισφέρουν στην μεταβλητή απόκρισης, το οποίο ισοδυναμεί με την παραδοχή ότι ιδανικά το διάνυσμα των παραμέτρων β είναι σποραδικό (sparse). Η σποραδικότητα συνοδεύει συχνά μοντέλα μεγάλων διαστάσεων όπου ο αριθμός των εξεταζόμενων μεταβλητών, δηλαδή η διάσταση, p, του διανύσματος β είναι ίσως κατά πολύ μεγαλύτερη του μεγέθους δείγματος n.

Ένας από τους πρακτικούς λόγους που φαίνεται η πραγματική δυσκολία όταν η διάσταση p είναι μεγαλύτερη από το μέγεθος δείγματος n, είναι ότι ο πίνακας σχεδιασμού έχει περισσότερες στήλες απ' ότι γραμμές. Στην περίπτωση αυτή, ο πίνακας $X'X$ είναι τεράστιος και μη αντιστρέψιμος. Πιθανόν κάποια μη σημαντική ανεξάρτητη μεταβλητή να εμφανίζει υψηλή συσχέτιση με τη μεταβλητή απόκρισης λόγω της σύνδεσης της πρώτης με κάποια άλλη σημαντική ανεξάρτητη μεταβλητή. Αυτά τα φαινόμενα δυσκολεύουν την επιλογή μεταβλητών. Στόχος είναι η μείωση της διάστασης p μεγάλης ή τεράστιας κλίμακας (π.χ. της τάξης

$\exp(O(n^{\xi})$ για κάποιο $\xi > 0)$ σε μια σχετικά μεγάλη κλίμακα d (π.χ. $O(n)$) χρησιμοποιώντας μια γρήγορη και αποτελεσματική μέθοδο. Η μέθοδος αυτή λέγεται SIS (Sure Independence Screening) και προτάθηκε από τους Fan και Lv (2008).Αναφερόμενοι στην ιδιότητα SS (Sure Screening) εννοούμε ότι όλες οι σημαντικές μεταβλητές 'επιβιώνουν', με την έννοια ότι εμφανίζονται στο τελικά επιλεγμένο μοντέλο, με πιθανότητα που τείνει στη μονάδα μετά από τη διαλογή που γίνεται στις μεταβλητές.

Έστω $M_* = \{1 \leq i \leq p : \beta_i \neq 0\}$ το πραγματικό σποραδικό μοντέλο με αριθμό μη μηδενικών συντελεστών (nonsparsity rate) $s = |M_*|$. Οι υπόλοιπες $p - s$ μεταβλητές μπορεί επίσης να παρουσιάζουν συσχέτιση με τη μεταβλητή απόκρισης λόγω της πιθανής τους σύνδεσης με τις μεταβλητές που υπάρχουν στο μοντέλο. Έστω $\omega = (\omega_1, \omega_2, \ldots, \omega_p)'$ το διάνυσμα διάστασης p (componentwise regression) τέτοιο ώστε

$$\omega = X'y \tag{2.27}$$

όπου οι στήλες του πίνακα σχεδιασμού X έχουν τυποποιηθεί. Για οποιοδήποτε $\gamma \in (0, 1)$, διατάσσουμε τις συντεταγμένες του διανύσματος ω κατά φθίνουσα σειρά και ορίζουμε ένα υπομοντέλο

$$M_\gamma = \{1 \leq i \leq p : \omega_i \text{ είναι μεταξύ των πρώτων } [\gamma n] \text{ μεγαλύτερων}\} \tag{2.28}$$

όπου $[\gamma n]$ είναι το ακέραιο μέρος του αριθμού γn. Αυτός είναι ένας άμεσος τρόπος να συρρικνωθεί το πλήρες μοντέλο $\{1, \ldots, p\}$ σε ένα υπομοντέλο M_γ μεγέθους $d = [\gamma n] < n$. Αυτή η μέθοδος καλείται SIS. Το υπολογιστικό κόστος της μεθόδου είναι αυτό που αντιστοιχεί στον πολλαπλασιασμό ενός $n \times p$ πίνακα με ένα διάνυσμα διάστασης n συν την επιλογή των d μεγαλύτερων συντεταγμένων από ένα διάνυσμα διάστασης p. Η μέθοδος SIS λοιπόν έχει υπολογιστικό κόστος $O(np)$. Σημειώνουμε ότι εδώ $d < n$. Μπορούμε να επιλέξουμε το d να είναι συντηρητικό, π.χ. $n - 1$ ή $n/\log n$. Παρόλο που η μέθοδος προτείνεται με σκοπό τη μείωση της διάστασης από κάποια υψηλή τιμή p σε κάποια χαμηλότερη μικρότερη από το μέγεθος του δείγματος, τίποτα δε μας εμποδίζει να εφαρμόσουμε τη μέθοδο και να καταλήξουμε σε κάποιο τελικό μοντέλο διάστασης $d > n$ επιλέγοντας $\gamma > 1$. Είναι προφανές ότι μεγαλύτερο d σημαίνει μεγαλύτερη πιθανότητα να συμπεριλάβουμε το πραγματικό μοντέλο M_* στο τελικό μοντέλο M_γ.

Για το πρόβλημα της επιλογής μεταβλητών σε περιπτώσεις υπερ-υψηλής διάστασης, προτείνεται αρχικά η εφαρμογή της SIS προκειμένου να μειωθεί η διάσταση από p σε μια σχετικά χαμηλή κλίμακα d κάτω από το μέγεθος του δείγματος n. Στη συνέχεια εφαρμόζεται μια μέθοδος επιλογής μοντέλου για χαμηλότερες διαστάσεις όπως είναι π.χ. η $SCAD$, η $LASSO$ ή κάποια άλλη. Συμπερασματικά, με την εφαρμογή της μεθόδου SIS γίνεται εφικτή αποτελεσματικά η επιλογή μοντέλου σε μεγάλες διαστάσεις και επιταχύνεται δραστικά η επιλογή μεταβλητών. Επιπλέον μπορεί να χρησιμοποιηθεί σε συνδυασμό με οποιαδήποτε άλλη τεχνική επιλογής μοντέλου ακόμα και Μπεϋζιανή.

Η βασική ιδέα της μεθόδου SIS είναι η εφαρμογή μιας και μόνο componentwise regression. Όμως τρία πιθανά ζητήματα προκύπτουν. Πρώτον, κάποιες μη σημαντικές μεταβλητές που παρουσιάζουν υψηλή συσχέτιση με κάποιες από τις σημαντικές μεταβλητές μπορεί να έχουν μεγαλύτερη προτεραιότητα να επιλεγούν με τη SIS απ' ότι άλλες σημαντικές μεταβλητές που όμως είναι σχετικά λιγότερο συσχετισμένες με τη μεταβλητή απόκρισης. Δεύτερο, μια σημαντική μεταβλητή που είναι οριακά ασυσχέτιστη αλλά από κοινού συσχετισμένη με την εξαρτημένη, δε θα επιλεγεί από τη SIS και συνεπώς δε θα εισαχθεί στο εκτιμώμενο μοντέλο. Τρίτον, το ζήτημα της πολυσυγγραμμικότητας αποτελεί επιπρόσθετη δυσκολία στο θέμα της επιλογής μεταβλητών. Προκειμένου να ξεπεραστούν τα παραπάνω μειονεκτήματα, προτείνεται μια επέκταση της μεθόδου SIS η οποία καλείται Iterative SIS (ISIS). Η νέα μέθοδος λειτουργεί ως ακολούθως. Στο πρώτο βήμα, επιλέγουμε ένα υποσύνολο που αποτελείται από k_1 μεταβλητές $A_1 = \{X_{i_1}, \ldots, X_{i_{k_1}}\}$ χρησιμοποιώντας τη SIS σε συνδυασμό με κάποια άλλη, π.χ. $SCAD$. Τότε έχουμε ένα $n-$διάστατο διάνυσμα από τα υπόλοιπα της παλινδρόμησης της μεταβλητής απόκρισης με τις $\{X_{i_1}, \ldots, X_{i_{k_1}}\}$. Στο επόμενο βήμα χρησιμοποιούμε αυτά τα υπόλοιπα ως νέες παρατηρήσεις της μεταβλητής απόκρισης και εφαρμόζουμε την ίδια μέθοδο όπως στο προηγούμενο βήμα για τις υπόλοιπες $p - k_1$ μεταβλητές. Με το βήμα αυτό προκύπτει ένα υποσύνολο με k_2 μεταβλητές $A_2 = \{X_{j_1}, \ldots, X_{j_{k_2}}\}$. Ας σημειωθεί ότι κάνοντας προσαρμογή των υπολοίπων στις μεταβλητές $\{X_1, \ldots, X_p\} \setminus A_1$ μπορεί να μειωθεί σημαντικά η προτεραιότητα εκείνων των μη σημαντικών μεταβλητών που έχουν υψηλή συσχέτιση με την εξαρτημένη μεταβλητή μέσω της σύνδεσής τους με τις $X_{i_1}, \ldots, X_{i_{k_1}}$, αφού τα υπόλοιπα είναι ασυσχέτιστα με τις μεταβλητές που επιλέχθηκαν στο A_1. Αυτό βοηθάει στην επίλυση του πρώτου ζητήματος από αυτά που αναφέραμε παραπάνω. Επίσης με την επαναλη-

πτική αυτή μέθοδο, εκείνες οι σημαντικές μεταβλητές που χάθηκαν στο προηγούμενο βήμα είναι δυνατόν να επιβιώσουν, γεγονός που λύνει και το δεύτερο ζήτημα. Στην πραγματικότητα, αφού εισαχθούν οι μεταβλητές του συνόλου A_1 στο μοντέλο, εκείνες που είναι οριακά ασθενώς συσχετισμένες καθαρά με την Y, λόγω της παρουσίας των μεταβλητών στο A_1 τώρα θα είναι συσχετισμένες με τα υπόλοιπα. Συνεχίζουμε την επανάληψη της διαδικασίας μέχρι να πάρουμε ℓ ξένα μεταξύ τους υποσύνολα $A_1, A_2, \ldots, A_\ell$ τέτοια ώστε η ένωσή τους $A = \cup_{i=1}^{\ell} A_i$ να έχει μέγεθος d το οποίο να είναι μικρότερο του μεγέθους δείγματος n. Στην πράξη, μπορούμε να επιλέξουμε για παράδειγμα το μεγαλύτερο ℓ για το οποίο $|A| < n$.

Οι μέθοδοι αυτές εφαρμόζονται και σε άλλα μοντέλα όπως τα γενικευμένα γραμμικά μοντέλα ή στην περίπτωση που τα δεδομένα προσαρμόζονται στο μοντέλο του Cox (Fan, Feng and Wu, 2010). Στην τελευταία περίπτωση, για κάθε συμμεταβλητή $X_m,\ m = 1, 2, \ldots, p$ ορίζεται η περιθώρια χρησιμότητα ως το μέγιστο της μερικής πιθανοφάνειας αυτής της μιας μεταβλητής:

$$u_m = \max_{\beta_m} \left\{ \sum_{i=1}^{n} \delta_i x_{im} \beta_m - \sum_{i=1}^{n} \delta_i \log \left\{ \sum_{j \in R(y_i)} \exp(x_{jm} \beta_m) \right\} \right\}. \tag{2.29}$$

Εδώ x_{im} είναι το m στοιχείο του $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$. Διαισθητικά, όσο μεγαλύτερη είναι η περιθώρια χρησιμότητα, τόσο περισσότερη πληροφορία σχετικά με την επιβίωση περιέχει η αντίστοιχη συμμεταβλητή. Αφού υπολογιστούν όλες οι περιθώριες χρησιμότητες $u_m,\ m = 1, 2, \ldots, p$, οι συμμεταβλητές κατατάσσονται σύμφωνα με τη χρησιμότητά τους ξεκινώντας από εκείνη με τη μεγαλύτερη χρησιμότητα και φτάνοντας ως τη μικρότερη. Κατόπιν επιλέγονται οι d πρώτες στην κατάταξη. Έστω $\mathcal{I}$ το σύνολο των δεικτών αυτών των d επιλεγμένων μεταβλητών. Το σύνολο $\mathcal{I}$ αναμένεται να καλύπτει το πραγματικό σύνολο δεικτών M^* με μεγάλη πιθανότητα, ιδιαίτερα αν επιλεγεί ένα κάποιο σχετικά μεγάλο d. Η παράμετρος d επιλέγεται αρκετά μεγάλη έτσι ώστε να εξασφαλιστεί η ισχύς της ιδιότητας Sure Screening. Παρολ' αυτά το σύνολο $\mathcal{I}$ μπορεί να περιέχει αρκετές μη σημαντικές συμμεταβλητές. Για να βελτιωθεί η απόδοση, μπορεί να εφαρμοστεί μια μέθοδος επιλογής μεταβλητών με ποινή για να απομακρυνθούν επιπλέον μη σημαντικές μεταβλητές. Αυτό μαθηματικά σημαίνει ότι λύνουμε το παρακάτω πρόβλημα ελαχιστοποίησης της ποινικοποιημένης μερικής

πιθανοφάνειας:

$$\min_{\beta_{\mathcal{I}}} \left\{ -\sum_{i=1}^{n} \delta_i x'_{\mathcal{I},i} \beta_{\mathcal{I}} + \sum_{i=1}^{n} \delta_i \log \left\{ \sum_{j \in R(y_i)} \exp(x'_{\mathcal{I},j} \beta_{\mathcal{I}}) \right\} + \sum_{m \in \mathcal{I}} p_\lambda(\beta_m) \right\}, \tag{2.30}$$

όπου $x_{\mathcal{I},i}$ συμβολίζει εκείνο το μέρος του διανύσματος x_i με δείκτες στο $\mathcal{I}$ και ομοίως για το διάνυσμα $\beta_{\mathcal{I}}$ και $p_\lambda(\cdot)$ η επιλεγμένη, ανάλογα με τη μέθοδο που εφαρμόζεται, συνάρτηση ποινής. Η παραπάνω ελαχιστοποίηση θα οδηγήσει σε ένα διάνυσμα που αποτελεί μια σποραδική διανυσματική εκτιμήτρια παλινδρόμησης $\hat{\beta}_{\mathcal{I}}$. Αν συμβολίσουμε με $\hat{M}$ το σύνολο των δεικτών των μη-μηδενικών συνιστωσών του διανύσματος $\hat{\beta}_{\mathcal{I}}$, τότε το $\hat{M}$ αποτελεί την τελική μας εκτίμηση για το M^*.

Σχετικά με την εφαρμογή της μεθόδου ISIS αρχικά εφαρμόζεται η SIS οπότε επιλέγεται ένα σύνολο δεικτών $\hat{\mathcal{I}}_1$ στο οποίο στη συνέχεια όπως περιγράψαμε εφαρμόζεται κάποια μέθοδος επιλογής μεταβλητών με ποινή και υπολογίζονται με αυτό τον τρόπο οι εκτιμήτριες $\hat{\beta}_{\hat{\mathcal{I}}_1}$. Έτσι παίρνουμε ένα βελτιωμένο σύνολο $\hat{M}_1$ ως εκτιμήτρια του συνόλου των πραγματικών δεικτών, το οποίο αποτελείται από τους δείκτες των μη-μηδενικών στοιχείων του $\hat{\beta}_{\hat{\mathcal{I}}_1}$. Κατόπιν, ορίζουμε τη δεσμευμένη χρησιμότητα για κάθε μια συμμεταβλητή m η οποία δεν περιέχεται στο σύνολο $\hat{M}_1$.

$$u_{m|\hat{M}_1} = \max_{\beta_m, \beta_{\hat{M}_1}} \left\{ \sum_{i=1}^{n} \delta_i (x_{im}\beta_m + x'_{\hat{M}_1,i} \beta_{\hat{M}_1}) - \sum_{i=1}^{n} \delta_i \log \left\{ \sum_{j \in R(y_i)} \exp(x_{jm}\beta_m + x'_{\hat{M}_1,j} \beta_{\hat{M}_1}) \right\} \right\}. \tag{2.31}$$

Αυτή η δεσμευμένη χρησιμότητα μετράει την επιπρόσθετη συνεισφορά της m συμμεταβλητής, δεδομένου ότι όλες οι συμμεταβλητές με δείκτες που περιέχονται στο σύνολο $\hat{M}_1$ έχουν συμπεριληφθεί στο μοντέλο. Αφού υπολογιστούν οι χρησιμότητες για όλες τις συμμεταβλητές που δεν είναι στο $\hat{M}_1$, στη συνέχεια διατάσσονται από τη μεγαλύτερη στη μικρότερη και επιλέγουμε τις d_1 πρώτες. Έστω $\hat{\mathcal{I}}_2$ το σύνολο των δεικτών των επιλεγμένων μεταβλητών στο βήμα αυτό. Τότε και πάλι εφαρμόζουμε μια μέθοδο επιλογής μεταβλητών με ποινή ελαχιστοποιώντας την

$$\min_{\beta_{\hat{M}_1 \cup \hat{\mathcal{I}}_2}} \left\{ -\sum_{i=1}^{n} \delta_i x'_{\hat{M}_1 \cup \hat{\mathcal{I}}_2, i} \beta_{\hat{M}_1 \cup \hat{\mathcal{I}}_2} + \sum_{i=1}^{n} \delta_i \log \left\{ \sum_{j \in R(y_i)} \exp(x'_{\hat{M}_1 \cup \hat{\mathcal{I}}_2, j} \beta_{\hat{M}_1 \cup \hat{\mathcal{I}}_2}) \right\} + \sum_{m \in \hat{M}_1 \cup \hat{\mathcal{I}}_2} p_\lambda(\beta_m) \right\} \tag{2.32}$$

ως προς $\beta_{\hat{M}_1 \cup \hat{\mathcal{I}}_2}$. Παίρνουμε έτσι τη σποραδική εκτιμήτρια $\hat{\beta}_{\hat{M}_1 \cup \hat{\mathcal{I}}_2}$. Έστω $\hat{M}_2$ το σύνολο των δεικτών των μη-μηδενικών συντελεστών του $\hat{\beta}_{\hat{M}_1 \cup \hat{\mathcal{I}}_2}$. Τότε το $\hat{M}_2$ αποτελεί τη νέα εκτίμηση του πραγματικού συνόλου δεικτών M^*. Αξίζει να σημειωθεί ότι σε αυτό το βήμα, μπορεί να διαγραφούν κάποιες από τις μεταβλητές $X_j \in \hat{M}_1$ οι οποίες είχαν επιλεγεί στο προηγούμενο βήμα. Οι παραπάνω επαναλήψεις επαναλαμβάνονται μέχρι να ικανοποιηθεί κάποιο κριτήριο σύγκλισης. Τα πιο συνήθη είναι ή να έχουν επιλεγεί d συμμεταβλητές ή $\hat{M}_j = \hat{M}_{j-1}$ για κάποιο j.

Προσομοιώσεις σε δεδομένα (Fan et. al., 2010) στα οποία εφαρμόστηκε η L_1 και η ISIS έδειξαν ότι τόσο η L_1 όσο και η ISIS έχουν την ιδιότητα Sure Screening αλλά η L_1 έδωσε μοντέλο 10 φορές μεγαλύτερο σε μέγεθος από αυτό στο οποίο κατέληξε η ISIS. Κατά συνέπεια, η L_1 εμφανίζει και μεγαλύτερα σφάλματα εκτίμησης. Αυτό οφείλεται στο γεγονός ότι η L_1 κρατάει αρκετούς μη-μηδενικούς αλλά μικρούς συντελεστές που αντιστοιχούν σε μη σημαντικές μεταβλητές. Αυτό γίνεται διότι, προκειμένου η L_1 να έχει μικρή αμεροληψία για τους μη-μηδενικούς συντελεστές, επιλέγεται μικρή τιμή της ρυθμιστικής παραμέτρου λ. Όμως ένα μικρό λ κρατάει πολλούς μικρούς συντελεστές για μη σημαντικές μεταβλητές.

2.7 Cross-Validation (cvl)

Η ικανότητα πρόβλεψης (predictive value) ενός μοντέλου είναι διαφορετική έννοια από την ερμηνευόμενη μεταβλητότητα του μοντέλου. Όταν φτιάχνουμε ένα μοντέλο παλινδρόμησης συνήθως αναζητούμε αυτό που περιγράφει καλύτερα τα διαθέσιμα δεδομένα. Κατά τη διαδικασία αυτή, η ερμηνευόμενη μεταβλητότητα, όπως αυτή μετράται από τη λογαριθμοποιημένη πιθανοφάνεια, μεγιστοποιείται. Όμως, για να χειριστούμε το πόσο καλά το μοντέλο μας προβλέπει μελλοντικά δεδομένα, η ικανότητα πρόβλεψης πρέπει να μετρηθεί. Στα γραμμικά και λογιστικά μοντέλα αυτό μπορεί να γίνει με μέτρα όπως Allen's PRESS, Mallows C_p, AIC. Στην ανάλυση επιβίωσης όταν βασιζόμαστε στο μοντέλο αναλογικής διακινδύνευσης του Cox, PRESS και Mallow's C_p δεν είναι διαθέσιμα. Επιπλέον το AIC δεν είναι εύκολο να ερμηνευθεί, διότι οι συνιστώσες της μερικής πιθανοφάνειας είναι εξαρτημένες. Με αφορμή αυτούς τους προβληματισμούς, οι Verweij και Van Houwelingen (1993) εισήγαγαν την cross-validated λογαριθμοποιημένη πιθανοφάνεια ως μέτρο της ικανότητας πρόβλεψης ενός

μοντέλου αναλογικής διακινδύνευσης του Cox.

Η μέθοδος αποτελεί γενίκευση της πιο απλής μεθόδου αξιολόγησης μοντέλου (model validation) η οποία υπαγορεύει το χωρισμό του δείγματος (data-splitting). Στην τελευταία μέθοδο, το σύνολο δεδομένων χωρίζεται σε δύο μέρη, το training sample που χρησιμοποιείται για την ανάπτυξη του μοντέλου και του test sample το οποίο χρησιμοποιείται για την αξιολόγηση του μοντέλου. Ο χωρισμός του δείγματος γίνεται με τυχαίο τρόπο. Η μέθοδος έχει το μειονέκτημα (Harrell, 2002) ότι το μέγεθος του αρχικού δείγματος πρέπει να είναι αρκετά μεγάλο έτσι ώστε τόσο το training όσο και το test δείγμα θα πρέπει να περιέχουν και αυτά αρκετές παρατηρήσεις. Επίσης ο τυχαίος χωρισμός του δείγματος συνεπάγεται ότι διαφορετικός τρόπος χωρίσματος θα έδινε διαφορετικά αποτελέσματα.

Μερικά από τα προβλήματα της παραπάνω μεθόδου λύνει η μέθοδος της *cvl*. Η πιο απλή μορφή της μεθόδου είναι η leave-one-out cvl. Σε κάθε επανάληψη εξαιρείται μια παρατήρηση από την αναλυτική διαδικασία και γίνεται πρόβλεψη της τιμής της μεταβλητής απόκρισης για αυτή την παρατήρηση χρησιμοποιώντας το μοντέλο που έχει προκύψει από τις εναπομείνουσες $n-1$ παρατηρήσεις. Η διαδικασία επαναλαμβάνεται n φορές και προκύπτει έτσι μια μέση ακρίβεια. Ο Efron (1983) υποστηρίζει ότι η εφαρμογή της μεθόδου σε ομάδες (fold) δίνει πιο ακριβή αποτελέσματα. fold=k σημαίνει ότι μια ομάδα k παρατηρήσεων εξαιρείται κάθε φορά. Για παράδειγμα υποθέτοντας ότι $k = 10$ σημαίνει ότι κάθε φορά θα εξαιρείται μια ομάδα που αποτελείται από 10 παρατηρήσεις. Συνεπώς το αρχικό σύνολο δεδομένων χωρίζεται σε 10 υποσύνολα με τυχαίο τρόπο που το καθένα περιέχει τον ίδιο αριθμό παρατηρήσεων. Σε κάθε επανάληψη οι 9 από τις 10 ομάδες χρησιμοποιούνται για την ανάπτυξη του μοντέλου (επιλογή μεταβλητών, υπολογισμό δεικτών, κ.τ.λ.). Στη συνέχεια το μοντέλο που προκύπτει αξιολογείται ως προς της ακρίβεια πάνω στην ομάδα που έχει εξαιρεθεί. Αυτή η διαδικασία επαναλαμβάνεται τουλάχιστον 10 φορές από τις οποίες προκύπτουν τουλάχιστον 10 δείκτες π.χ. R^2. Ένα μειονέκτημα της μεθόδου είναι η επιλογή του αριθμού των παρατηρήσεων που εξαιρούνται κάθε φορά. Επίσης ο αριθμός των επαναλήψεων που απαιτείται προκειμένου να αποκτήσουμε ακριβείς εκτιμήσεις για την ακρίβεια του μοντέλου συχνά υπερβαίνει τις 200.

Υποθέτουμε ότι διαθέτουμε n παρατηρήσεις και ένα μοντέλο παλινδρόμησης χρησιμοποιείται για να περιγράψει τα δεδομένα. Η λογαριθμοποιημένη πιθανοφάνεια συμβολίζεται ως $\ell(\beta)$, όπου β το διάνυσμα των συντελεστών παλινδρόμησης. Ορίζουμε τη συμβολή της

παρατήρησης i στη λογαριθμοποιημένη πιθανοφάνεια ως

$$\ell_i(\beta) = \ell(\beta) - \ell_{(-i)}(\beta) \tag{2.33}$$

όπου $\ell_{(-i)}(\beta)$ είναι η λογαριθμοποιημένη πιθανοφάνεια αν αγνοήσουμε την i-οστή παρατήρηση. Η τιμή του β η οποία μεγιστοποιεί την $\ell_{(-i)}(\beta)$ συμβολίζεται ως $\hat{\beta}_{(-i)}$. Αν οι συνιστώσες της πιθανοφάνειας είναι ανεξάρτητες, όπως στα μοντέλα γραμμικής και λογιστικής παλινδρόμησης, η $\ell_i(\beta)$ απλά ισούται με τη συμβολή της i-οστής συνιστώσας και $\sum_{i=1}^{n} \ell_i(\beta) = \ell(\beta)$. Ορίζουμε ως cross-validated λογαριθμοποιημένη πιθανοφάνεια cvl

$$cvl = \sum_{i=1}^{n} \ell_i(\hat{\beta}_{(-i)}). \tag{2.34}$$

Για ένα μοντέλο η cvl μετράει πόσο καλά κάθε παρατήρηση i μπορεί να προβλεφθεί χρησιμοποιώντας τις άλλες παρατηρήσεις και εξυπηρετεί συνεπώς ως ένα μέτρο της ικανότητας πρόβλεψης του μοντέλου.

Η διαφορά μεταξύ της cvl και της λογαριθμοποιημένης πιθανοφάνειας φαίνεται καλύτερα στο γραμμικό μοντέλο με γνωστή διασπορά. Αν εξαιρέσουμε τις σταθερές, η $-2 \times \ln -$πιθανοφάνεια στην περίπτωση του γραμμικού μοντέλου ισούται με $SSR = \sum(y_i - \hat{y}_i)^2 = \sum(y_i - x_i\hat{\beta})^2$, ενώ η cvl ισούται με τα υπόλοιπα PRESS, $PRESS = \sum(y_i - x_i\hat{\beta}_{(-i)})^2$ το οποίο με τη σειρά του είναι ίσο με το δείκτη C_p του Mallows.

Στο μοντέλο του Cox, η ποσότητα $\ell_i(\beta)$ προκύπτει (Harrell, 2002) ως εξής. Υπενθυμίζουμε ότι η μερική πιθανοφάνεια όταν δεν υπάρχουν ισότητες δίνεται ως

$$L(\beta) = \prod_{j=1}^{n} \left(\frac{x_j}{\sum_{k \in R_j} x_k} \right)^{d_j}. \tag{2.35}$$

Όταν η μονάδα i αγνοείται, τότε αυτή αφαιρείται από όλα τα σύνολα που περιέχουν τις μονάδες που βρίσκονται σε κίνδυνο πριν τη στιγμή t_i. Αν οι χρόνοι t_i είναι διατεταγμένοι έτσι ώστε $t_j < t_i$ για $j < i$, τότε

$$L_{(-i)}(\beta) = \prod_{j<i} \left(\frac{x_j}{\sum_{k \in R_j} x_k - x_i} \right)^{d_j} \prod_{j>i} \left(\frac{x_j}{\sum_{k \in R_j} x_k} \right)^{d_j}. \tag{2.36}$$

Η συνεισφορά $L_i(\beta)$ της μονάδας i στη μερική πιθανοφάνεια ισούται με $L(\beta)/L_{(-i)}(\beta)$ το οποίο τελικά ισούται με

$$L_i(\beta) = \prod_{j<i} (1 - p_{ij})^{d_j} p_{ii}^{d_i} \tag{2.37}$$

με

$$p_{ij} = \frac{x_i}{\sum_{k \in R_j} x_k} \tag{2.38}$$

την πιθανότητα η μονάδα i να πεθάνει τη στιγμή t_j, δεδομένων των συνόλων κινδύνου και των χρόνων επιβίωσης. Συνεπώς, η $L_i(\beta)$ είναι η δεσμευμένη πιθανότητα η μονάδα i να επιβιώσει τη στιγμή t_{i-1} και, αν $d_i = 1$, να πεθάνει τη στιγμή t_i. Η συνεισφορά $\ell_i(\beta)$ στη λογαριθμοποιημένη πιθανοφάνεια είναι

$$\ell_i(\beta) = \sum_{j<i} d_j \ln(1 - p_{ij}) + d_i \ln(p_{ii}). \tag{2.39}$$

Στο μοντέλο χωρίς συμμεταβλητές, $p_{ij} = 1/(n-j+1)$ για κάθε i και αν δεν υπάρχει αποκοπή τότε $l_i(0) = -\ln(n)$ για κάθε i.

Για τον υπολογισμό της *cvl*, απαιτείται ο υπολογισμός των συντελεστών $\hat{\beta}_{(-i)}$. Ο προσδιορισμός αυτών των συντελεστών απαιτεί την προσαρμογή n μοντέλων Cox κάθε ένα με $n-1$ παρατηρήσεις.

Όπως αναφέρθηκε αρκετές φορές στις προηγούμενες παραγράφους η cvl χρησιμοποιείται για τον προσδιορισμό των απαιτούμενων ρυθμιστικών παραμέτρων για την εφαρμογή των μεθόδων με ποινή. Στην περίπτωση των γραμμικών μοντέλων η διαδικασία γίνεται ως εξής στην περίπτωση π.χ. που fold=k. Έστω T το σύνολο των δεδομένων και T_ν, $T - T_\nu$, $\nu = 1, \ldots, k$ το test και το training σύνολο αντίστοιχα. Για κάθε λ, ν υπολογίζουμε μια εκτιμήτρια $\hat{\beta}^{(\nu)}(\lambda)$ με βάση το training δείγμα. Σχηματίζουμε το κριτήριο της cvl:

$$cvl(\lambda) = \sum_{\nu=1}^{k} \sum_{(x_i, y_i) \in T_\nu} \left\{ y_i - x_i' \hat{\beta}^{(\nu)}(\lambda) \right\}^2. \tag{2.40}$$

Επιλέγουμε το λ για το οποίο ελαχιστοποιείται η cvl.

Κεφάλαιο 3

Εφαρμογές

Στο κεφάλαιο αυτό θα δούμε πώς εφαρμόζονται οι διάφορες κλασικές μέθοδοι αλλά και οι νέες που περιγράψαμε στο θεωρητικό μέρος για την ανάλυση ενός μοντέλου καθώς και για την επιλογή μεταβλητών δηλαδή την επιλογή του βέλτιστου μοντέλου. Τα δεδομένα προέρχονται από 51 ασθενείς που πάσχουν από οξεία μυελοβλαστική λευχαιμία (Lee, 1980). Αφού υποβληθούν σε κάποια θεραπεία στο τέλος εξετάζεται αν ανταποκρίθηκαν ή όχι. Οι μετρήσεις έγιναν πριν τη λήψη της θεραπείας, αντιστοιχούν στις μεταβλητές όπως φαίνεται από κάθε στήλη των δεδομένων και είναι οι εξής:

- X_1: Ηλικία διάγνωσης (age)
- X_2: Ποσοστό επίστρωσης βλαστοκυττάρων (smear)
- X_3: Ποσοστό κυττάρων λευχαιμίας που εισήλθαν στο μυελό των οστών (infiltr)
- X_4: Ποσοστό κυττάρων που προήλθαν από το μυελό των οστών (lab)
- X_5: Απόλυτα βλαστοκύτταρα $\times 10^3$ (blasts)
- X_6: Υψηλότερη θερμοκρασία σώματος $\times 10^o$ F (temp)
- X_7: Ανταπόκριση στη θεραπεία: 1 Ανταποκρίνεται, 0 Δεν ανταποκρίνεται (resp)
- X_8: Χρόνος επιβίωσης σε μήνες από τη στιγμή της διάγνωσης (surv)

- X_9: Κατάσταση κατά τη λήξη της έρευνας: 1 Δεν έχει επιβιώσει, 0 Έχει επιβιώσει (status)

Μέσα σε παρένθεση αναφέρεται το όνομα κάθε μεταβλητής όπως αυτή εμφανίζεται στη στατιστική ανάλυση που πραγματοποιήθηκε. Πρόκειται για δεδομένα που περιέχουν και αποκομμένες παρατηρήσεις όπως διαπιστώνουμε από την ύπαρξη της μεταβλητής status. Τα δεδομένα αναλύονται προς δύο κατευθύνσεις. Στην πρώτη προσαρμόζεται το μοντέλο αναλογικής διακινδύνευσης του Cox με συμμεταβλητές τις $X_1 - X_6$ για να μελετηθεί η επίδρασή τους (αν υπάρχει) στην επιβίωση των ασθενών λαμβάνοντας υπόψη ότι έχουμε αποκομμένες παρατηρήσεις. Στη δεύτερη προσαρμόζεται ένα γενικευμένο γραμμικό μοντέλο της λογιστικής παλινδρόμησης για τη μεταβλητή απόκρισης X_7 που αντιστοιχεί σε μια 0,1 μεταβλητή για το αν ο ασθενής ανταποκρίνεται στη θεραπεία (1) ή όχι (0). Οι επεξηγηματικές μεταβλητές είναι οι $X_1 - X_6$. Όλες οι αναλύσεις πραγματοποιούνται στην R η οποία παρέχει, μέσω κατάλληλων πακέτων, τη δυνατότητα εκτέλεσης και των ποινικοποιημένων μεθόδων L_1 και L_2 που περιγράψαμε στο θεωρητικό μέρος της εργασίας.

3.1 Απλό μοντέλο του Cox

Αρχικά εφαρμόζουμε το απλό μοντέλο του Cox εισάγοντας όλες τις συμμεταβλητές.

```
> fitcox<-coxph(formula=Surv(surv,status)~age+smear+infltr+lab+blasts+temp)
> summary(fitcox)
Call:
coxph(formula = Surv(surv, status) ~ age + smear + infltr + lab +
    blasts + temp)

  n= 51, number of events= 45

          coef     exp(coef) se(coef)  z     Pr(>|z|)
age     0.03198   1.03249  0.01035  3.090   0.0020 **
smear   0.01356   1.01365  0.01528  0.888   0.3747
infltr -0.01709   0.98306  0.01232 -1.387   0.1654
```

```
lab    -0.07222   0.93032  0.03926 -1.840   0.0658 .
blasts -0.01685   0.98329  0.02268 -0.743   0.4573
temp    0.02212   1.02236  0.01353  1.635   0.1021
---
Signif. codes:  0 ‘‘***’’ 0.001 ‘‘**’’ 0.01 ‘‘*’’ 0.05 ‘‘.’’ 0.1 ‘‘ ’’ 1

         exp(coef) exp(-coef) lower.95  upper.95
age       1.0325     0.9685    1.0118     1.054
smear     1.0137     0.9865    0.9838     1.044
infltr    0.9831     1.0172    0.9596     1.007
lab       0.9303     1.0749    0.8614     1.005
blasts    0.9833     1.0170    0.9405     1.028
temp      1.0224     0.9781    0.9956     1.050

Concordance= 0.724  (se = 0.057 )
Rsquare= 0.328   (max possible= 0.996 )
Likelihood ratio test= 20.26  on 6 df,   p=0.002486
Wald test            = 19.31  on 6 df,   p=0.003676
Score (logrank) test = 20.88  on 6 df,   p=0.001929

> cox.zph(fitcox)
           rho  chisq        p
age     0.2092  1.955 0.162054
smear  -0.3131  4.599 0.031992
infltr -0.0762  0.232 0.630297
lab     0.1269  0.948 0.330198
blasts  0.4259 12.608 0.000384
temp   -0.3899 10.726 0.001057
GLOBAL      NA 23.291 0.000705
```

Από τα αποτελέσματα της ανάλυσης φαίνεται ότι οι μεταβλητές smear, infltr, blastskαι temp να μην είναι στατιστικά σημαντικές, η μεταβλητή lab είναι οριακά και εξαρτάται από το επίπεδο σημαντικότητας που θα θεωρήσουμε ενώ η μεταβλητή age φαίνεται να είναι η πιο σημαντική. Βέβαια απαραίτητος είναι ο έλεγχος για το αν η υπόθεση της αναλογικής διακινδύνευσης ευσταθεί. Αυτό γίνεται με την εντολή cox.zph η οποία κάνει τον παραπάνω έλεγχο για κάθε μεταβλητή χωριστά αλλά και για το συνολικό μοντέλο κάνοντας χρήση των υπολοίπων Schoenfeld. Μια p-τιμή μικρότερη από 0.05 δείχνει παραβίαση της υπόθεσης της αναλογικότητας. Από την $p-$τιμή του ελέγχου διαπιστώνουμε ότι η υπόθεση της αναλογικής διακινδύνευσης για τα δεδομένα μας ευσταθεί αν επικεντρωθούμε στη μεταβλητή age που βρέθηκε να είναι στατιστικά σημαντική. Λόγω της σημαντικότητας αυτής της μεταβλητής προσαρμόζουμε ένα νέο μοντέλο το οποίο περιλαμβάνει μόνο αυτή ως συμμεταβλητή. Παρακάτω φαίνονται τόσο τα αποτελέσματα όσο και ο έλεγχος της υπόθεσης της αναλογικής διακινδύνευσης για το μοντέλο που περιέχει μόνο τη μεταβλητή age.

```
> fitcox.age<-coxph(formula=Surv(surv,status)~age)
> summary(fitcox.age)
Call:
coxph(formula = Surv(surv, status) ~ age)

  n= 51, number of events= 45

       coef exp(coef) se(coef)     z Pr(>|z|)
age 0.032397  1.032927 0.009521 3.403 0.000667 ***
---
Signif. codes:  0 ‘‘***’’ 0.001 ‘‘**’’ 0.01 ‘‘*’’ 0.05 ‘‘.’’ 0.1 ‘‘ ’’ 1

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

Concordance= 0.65  (se = 0.057 )
Rsquare= 0.207   (max possible= 0.996 )
```

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

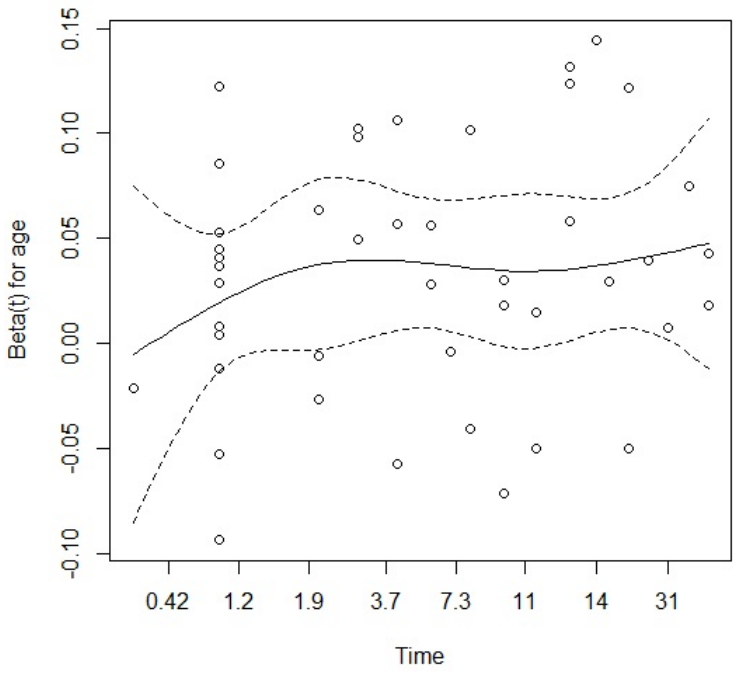
```
>cox.zph(fitcox.age)
      rho chisq     p
age 0.147 0.791 0.374
```

Από την προσαρμογή του μοντέλου μόνο με τη συμμεταβλητή της ηλικίας, φυσικά επιβεβαιώνεται ότι είναι στατιστικά σημαντική και μάλιστα από την τιμή της εκτίμησης για το συντελεστή β συμπεραίνουμε ότι αύξηση της ηλικίας διάγνωσης κατά ένα χρόνο αυξάνει τη συνάρτηση κινδύνου κατά 0.0324. Ισοδύναμα αφού $\exp(\beta) = 1.033$ σημαίνει ότι μοναδιαία αύξηση της ηλικίας προκαλεί αύξηση της συνάρτησης κινδύνου περίπου 3%. Η $p-$τιμή=0.362 δείχνει ότι η υπόθεση της αναλογικής διακινδύνευσης δεν μπορεί να απορριφθεί. Επιπλέον το Σχήμα 3.1 αποτελεί τη γραφική παράσταση των υπολοίπων Schoenfeld σε συνάρτηση με το χρόνο με σκοπό τον έλεγχο της ανεξαρτησίας των υπολοίπων αυτών με το χρόνο η οποία εδώ ευσταθεί. Αντίστοιχα το Σχήμα 3.2 παρουσιάζει και πάλι τα υπόλοιπα Schoenfeld σε συνάρτηση όμως με την μεταβλητή age από το οποίο εκτός από το ότι επιβεβαιώνεται και πάλι η υπόθεση της αναλογικής διακινδύνευσης, παρολ΄ αυτά παρατηρούμε ότι η προσαρμοσμένη καμπύλη δεν είναι εντελώς ευθεία (οπότε θα ήταν τέλειο μοντέλο αναλογικής διακινδύνευσης) αλλά παρουσιάζει μια μικρή αυξητική τάση προς τα δεξιά. Φαίνεται δηλαδή μεγαλύτερες ηλικίες να έχουν υψηλότερες τιμές υπολοίπων γεγονός που σημαίνει ότι πιθανόν να υπάρχει διαφοροποίηση μεταξύ της επιβίωσης των ατόμων στα οποία η διάγνωση γίνεται σε μικρότερη ηλικία σε σχέση με τα άτομα στα οποία η ασθένεια διαγνώσκεται σε μεγαλύτερη.

3.2 Απλό μοντέλο λογιστικής παλινδρόμησης

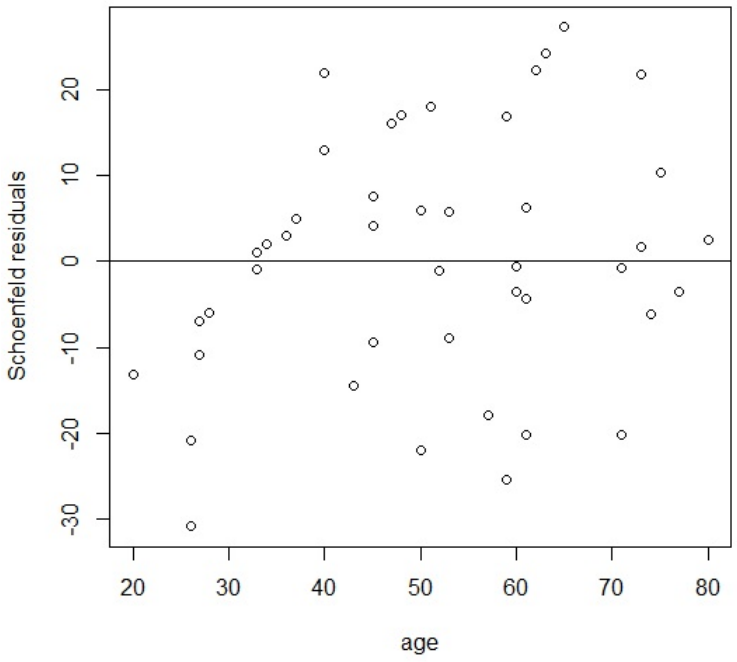
Στη συνέχεια, προσαρμόζουμε το μοντέλο της λογιστικής παλινδρόμησης για τη δίτιμη μεταβλητή απόκρισης που έχουμε και παίρνουμε τα παρακάτω αποτελέσματα.

```
> fit<-glm(resp~age+smear+infltr+lab+blasts+temp,family=binomial)
```



Σχήμα 3.1: Υπόλοιπα Schoenfeld για τη μεταβλητή age συναρτήσει του χρόνου.

```
> summary(fit)

Call:
glm(formula = resp ~ age + smear + infltr + lab + blasts + temp,
    family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.73878  -0.58099  -0.05505   0.62618   2.28425

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 98.52361   40.85385   2.412  0.01588 *
age         -0.06029    0.02729  -2.210  0.02714 *
smear       -0.00480    0.04108  -0.117  0.90698
```



Σχήμα 3.2: Υπόλοιπα Schoenfeld για τη μεταβλητή age συναρτήσει της ηλικίας age.

```
infltr       0.03621    0.03934   0.921  0.35728
lab          0.39845    0.13278   3.001  0.00269 **
blasts       0.01343    0.05782   0.232  0.81627
temp        -0.10223    0.04181  -2.445  0.01448 *
---
Signif. codes:  0 ‘‘***’’ 0.001 ‘‘**’’ 0.01 ‘‘*’’ 0.05 ‘‘.’’ 0.1 ‘‘ ’’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 70.524  on 50  degrees of freedom
Residual deviance: 40.060  on 44  degrees of freedom
AIC: 54.06

Number of Fisher Scoring iterations: 6
```

Παρατηρούμε ότι και εδώ φαίνεται άμεσα ότι οι μεταβλητές smear, infltr και blasts δεν είναι στατιστικά σημαντικές, οι μεταβλητές temp και age φαίνεται να είναι στατιστικά σημαντικές ενώ περισσότερο από όλες τώρα υπερισχύει η lab.

Πραγματοποιώντας και πάλι την ίδια ανάλυση με τις τρεις μόνο στατιστικά σημαντικές μεταβλητές όπως προέκυψαν παραπάνω παίρνουμε τα αποτελέσματα που ακολουθούν. Αξίζει να σημειωθεί ότι το νέο μοντέλο με τις λιγότερες μεταβλητές είναι καλύτερο από το προηγούμενο βάσει του κριτηρίου AIC αφού έχει μικρότερη τιμή του κριτηρίου αυτού.

```
> fitg<-glm(resp~age+lab+temp, family=binomial)
> summary(fitg)

Call:
glm(formula = resp ~ age + lab + temp, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.76104  -0.68683  -0.09747   0.67388   2.16510

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 87.38804   35.45816   2.465  0.01372 *
age         -0.05850    0.02558  -2.287  0.02218 *
lab          0.38493    0.12152   3.168  0.00154 **
temp        -0.08897    0.03607  -2.467  0.01363 *
---
Signif. codes:  0 ‘‘***’’ 0.001 ‘‘**’’ 0.01 ‘‘*’’ 0.05 ‘‘.’’ 0.1 ‘‘ ’’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 70.524  on 50  degrees of freedom
Residual deviance: 43.265  on 47  degrees of freedom
```

```
AIC: 51.265

Number of Fisher Scoring iterations: 6
```

Στο ίδιο αποτέλεσμα καταλήγουμε και χρησιμοποιώντας τη διαδικασία επιλογής σε στάδια με το κριτήριο AIC. Η μέθοδος της διαδοχικής αφαίρεσης και προσθήκης μεταβλητών σε κάθε βήμα και η σύγκριση μεταξύ των μοντέλων, επιλέγοντας τελικά εκείνο με τη μικρότερη τιμή του AIC, καταλήγει στο ίδιο βέλτιστο μοντέλο με αυτό της λογιστικής παλινδρόμησης εισάγοντας επιπλέον τη μεταβλητή infltr. Η εκτέλεση της μεθόδου αυτής γίνεται με χρήση της συνάρτησης *step* ενώ η μέθοδος, προς τα εμπρός, προς τα πίσω ή stepwise δηλώνεται με ένα όρισμα forward, backward ή both αντίστοιχα.

```
> stepfit<-step(fit,direction="both")
Start:  AIC=54.06
resp ~ age + smear + infltr + lab + blasts + temp

         Df Deviance    AIC
- smear   1   40.074 52.074
- blasts  1   40.115 52.115
- infltr  1   41.023 53.023
<none>        40.060 54.060
- age     1   46.157 58.157
- temp    1   48.277 60.277
- lab     1   55.823 67.823

Step:  AIC=52.07
resp ~ age + infltr + lab + blasts + temp

         Df Deviance    AIC
- blasts  1   40.136 50.136
<none>        40.074 52.074
```

```
- infltr  1   42.615 52.615
+ smear   1   40.060 54.060
- age     1   46.216 56.216
- temp    1   48.346 58.346
- lab     1   56.308 66.308

Step:  AIC=50.14
resp ~ age + infltr + lab + temp

         Df Deviance    AIC
<none>        40.136 50.136
- infltr  1   43.265 51.265
+ blasts  1   40.074 52.074
+ smear   1   40.115 52.115
- age     1   46.438 54.438
- temp    1   48.971 56.971
- lab     1   57.602 65.602
```

3.3 Εφαρμογή L_1 και L_2 - package *penalized* - Λογιστική Παλινδρόμηση

Περνάμε τώρα στην εφαρμογή των μεθόδων επιλογής μεταβλητών με ποινή. Το πακέτο *penalized* της *R* (Goeman, Meijer and Chaturvedi, 2012) είναι σχεδιασμένο για την πραγματοποίηση ποινικοποιημένης εκτίμησης σε γενικευμένα γραμμικά μοντέλα. Τα μοντέλα που υποστηρίζει το πακέτο είναι αυτά της γραμμικής παλινδρόμησης, λογιστικής παλινδρόμησης, Poisson παλινδρόμησης και μοντέλο αναλογικής διακινδύνευσης του Cox. Σχετικά με τις ποινές, το πακέτο επιτρέπει L_1, L_2 ή συνδυασμό των δύο. Συνοπτικά υπενθυμίζεται ότι οι μέθοδοι L_1 και L_2 συρρικνώνουν τις εκτιμήσεις των συντελεστών παλινδρόμησης προς το μηδέν σε σχέση με τις ε.ε.τ.. Ο σκοπός που γίνεται αυτή η συρρίκνωση είναι για να αποφευχθεί το overfitting που προκαλείται είτε λόγω πολυσυγγραμμικότητας των συμμεταβλητών

είτε λόγω μεγάλης διάστασης. Παρόλο που και οι δύο μέθοδοι προκαλούν συρρίκνωση, στην πράξη είναι αρκετά διαφορετικές. Η L_2 δίνει μικρούς αλλά μη μηδενικούς συντελεστές για όλες τις μεταβλητές ενώ η L_1 μηδενίζει εντελώς κάποιους συντελεστές και κάποιους άλλους τους συρρικνώνει σχετικά λίγο. Το μέγεθος της συρρίκνωσης που υφίστανται οι συντελεστές καθορίζεται από τη ρυθμιστική παράμετρο λ που εμφανίζεται ως συντελεστής στην αντίστοιχη συνάρτηση ποινής ανάλογα με τη μέθοδο. Στην R χρησιμοποιείται η διάκριση λ_1 και λ_2 για την παράμετρο όταν εφαρμόζεται η μέθοδος L_1 ή L_2 αντίστοιχα. Τιμή αυτών των παραμέτρων μηδέν σημαίνει ότι έχουμε τις ε.ε.τ. ενώ τιμή άπειρο σημαίνει άπειρη συρρίκνωση δηλαδή όλοι οι συντελεστές θέτονται μηδέν.

Πριν την προσαρμογή του μοντέλου αξίζει τον κόπο να ελέγξουμε τη μηδενική υπόθεση ότι δεν υπάρχει σχέση καμίας εκ των ανεξάρτητων μεταβλητών με τη μεταβλητή απόκρισης. Ο έλεγχος αυτό μπορεί να πραγματοποιηθεί κάνοντας χρήση του *globaltest* όπως αυτό περιγράφεται στην παράγραφο 2.1. Σε δεδομένα στα οποία ο έλεγχος είναι στατιστικά σημαντικός, σχεδόν πάντα έχουν μια βέλτιστη πεπερασμένη τιμή της ρυθμιστικής παραμέτρου λ στις μεθόδους επιλογής μεταβλητών με ποινή.

Οι συναρτήσεις *profL1* και *profL2* χρησιμοποιούνται για να εξετάσουν την επίδραση των ρυθμιστικών παραμέτρων λ_1 και λ_2 στη *cvl*. Δίνουν την τιμή της *cvl* για διάφορες τιμές των ρυθμιστικών παραμέτρων. Η μέγιστη και η ελάχιστη τιμή των παραμέτρων αυτών για τις οποίες υπολογίζεται η cvl μπορούν να δωθούν μέσω των *minlamda1* ή *maxlamda1* και *minlamda2* και *maxlamda2* αντίστοιχα. Η προεπιλεγμένη τιμή των *minlamda1* και *minlamda2* είναι μηδέν. Η προεπιλεγμένη τιμή του *maxlamda1* είναι η μέγιστη τιμή που έχει νόημα για το λ_1, δηλαδή η μικρότερη τιμή για την οποία όλοι οι συντελεστές παλινδρόμησης γίνονται μηδέν. Δεν υπάρχει προεπιλεγμένη τιμή για το *maxlamda2* οπότε αυτό πάντα πρέπει να δίνεται ως όρισμα. Ο αριθμός των βημάτων μεταξύ μέγιστης και ελάχιστης τιμής μπορεί να καθοριστεί μέσω του ορίσματος *steps* το οποίο έχει ως προεπιλεγμένη τιμή το 100. Αυτά τα βήματα απέχουν εξίσου αν το όρισμα *log* είναι *FALSE* ή ισαπέχοντα στη log κλίμακα αν το όρισμα *log* είναι *TRUE*. Σημειώνεται ότι η προεπιλεγμένη τιμή του *log* διαφέρει μεταξύ των συναρτήσεων *profL1* στην οποία είναι *FALSE* και στην *profL2* στην οποία είναι *TRUE*. Αν το *log* είναι *TRUE* , τότε οι τιμές *minlambda1* και *minlambda2* πρέπει να δωθούν από το χρήστη καθώς οι προεπιλεγμένες τιμές δεν χρησιμοποιούνται. Επίσης, ως προεπιλογή, ο

υπολογισμός της cvl σταματάει νωρίτερα από αυτό που πιθανόν καθορίζεται, όταν αυτή πέσει κάτω από εκείνη την τιμή της που αντιστοιχεί στο μηδενικό μοντέλο που έχει όλους τους ποινικοποιημένους συντελεστές παλινδρόμησης ίσους με μηδέν. Αυτό γίνεται προκειμένου να αποφευχθούν μακροσκελείς υπολογισμοί για μικρές τιμές του λ για τις οποίες τα αντίστοιχα μοντέλα με μεγάλη πιθανότητα δεν παρουσιάζουν κανένα ενδιαφέρον. Αυτή η αυτόματη διακοπή των υπολογισμών μπορεί κατάλληλα να αναιρεθεί.

Συνήθως δεν ενδιαφερόμαστε για τη συνολική συμπεριφορά της *cvl* αλλά μόνο στο βέλτιστο. Οι συναρτήσεις που χρησιμοποιούνται για την εύρεση του βέλτιστου λ_1 και λ_2 για τα οποία μεγιστοποιείται η αντίστοιχη *cvl* είναι οι *optL1* και *optL2*. Ο αλγόριθμος που χρησιμοποιείται για τον εντοπισμό των βέλτιστων τιμών των παραμέτρων λειτουργεί πάντα σε περιπτώσεις μονοκόρυφων συναρτήσεων αλλά μπορεί να συγκλίνει σε τοπικό μέγιστο σε περίπτωση που υπάρχουν περισσότερα από ένα. Αυτή η παρατήρηση σχετίζεται ιδιαίτερα με την L_1 μέθοδο, αφού η *cvl* ως συνάρτηση της παραμέτρου λ_1 συχνά παρουσιάζει διάφορα τοπικά μέγιστα. Συστήνεται η χρήση της *optL1* σε συνδυασμό με την *profL1* προκειμένου να αποφευχθεί η σύγκλιση σε λάθος μέγιστο. Η *cvl* σαν συνάρτηση του λ_2 συμπεριφέρεται πολύ καλύτερα και πρακτικά δεν έχει ποτέ τοπικά μέγιστα. Συνεπώς, η συνάρτηση *optL2* μπορεί να χρησιμοποιηθεί με ασφάλεια χωρίς να είναι απαραίτητος ο συνδυασμός της με την *profL2*.

Στα δεδομένα μας το *globaltest* που αναφέραμε παραπάνω δίνει τα ακόλουθα αποτελέσματα:

```
> gt(Surv(surv,status)~age+smear+infltr+lab+blasts+temp)
  p-value Statistic Expected Std.dev #Cov
1 0.00445      6.94     1.96     1.9    6
```

Από την p-τιμή του ελέγχου διαπιστώνουμε ότι ο έλεγχος είναι στατιστικά σημαντικός που σημαίνει από τη μια ότι η ικανότητα πρόβλεψης του μοντέλου αναμένεται να είναι καλή και από την άλλη έχει νόημα να αναζητήσουμε τις τιμές των ρυθμιστικών παραμέτρων για να προχωρήσουμε με την ανάλυση των μοντέλων με ποινή.

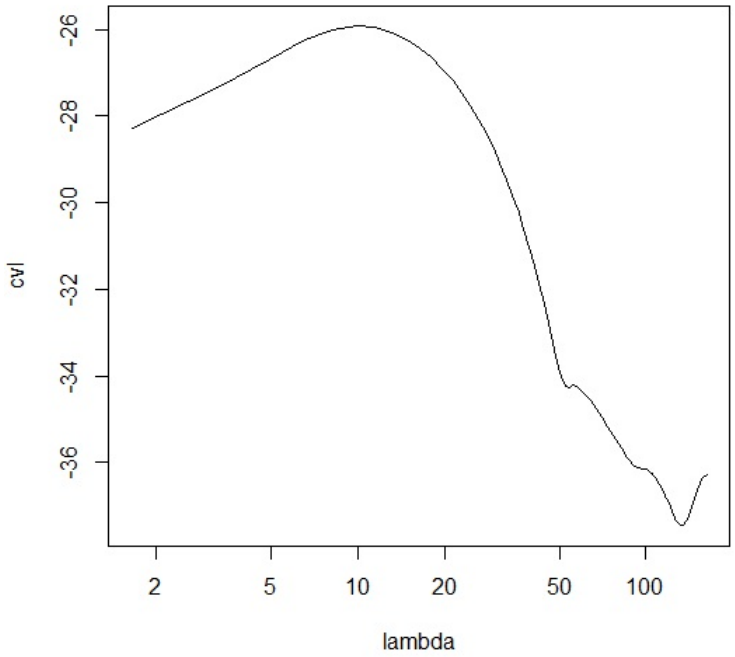
Ξεκινάμε με τη μέθοδο L_1. Αρχικά με την εντολή *profL1* αποκτούμε μια ιδέα για τη συμπεριφορά της cvl καθώς η παράμετρος λ μεταβάλλεται. Από αυτή αναζητούμε την τιμή του λ για την οποία η cvl γίνεται μέγιστη. Αυτό γίνεται με χρήση της optL1. Τα αποτελέσματα

που προέκυψαν καθώς και το γράφημα της cvl για τις διάφορες τιμές του λ φαίνεται στη συνέχεια.

```
> L1profile<-profL1(resp~age+smear+infltr+lab+temp,plot=TRUE)
lambda= 162.56  cvl= -36.29271
lambda= 160.918         cvl= -36.29888
lambda= 159.276         cvl= -36.31727
lambda= 157.6339        cvl= -36.36035
lambda= 155.9919        cvl= -36.4195
lambda= 154.3499        cvl= -36.4861
lambda= 152.7079        cvl= -36.56162
lambda= 151.0659        cvl= -36.64898
lambda= 149.4238        cvl= -36.74028
lambda= 147.7818        cvl= -36.83606
lambda= 146.1398        cvl= -36.93589
.   .   .
.   .   .
.   .   .
.   .   .
.   .   .
.   .   .
lambda= 18.06222        cvl= -26.66398
lambda= 16.4202         cvl= -26.43835
lambda= 14.77818        cvl= -26.23804
lambda= 13.13616        cvl= -26.06738
lambda= 11.49414        cvl= -25.95118
lambda= 9.852121        cvl= -25.93113
lambda= 8.210101        cvl= -26.02019
lambda= 6.568081        cvl= -26.25437
lambda= 4.926061        cvl= -26.68003
lambda= 3.28404         cvl= -27.33369
```

```
lambda= 1.64202         cvl= -28.28889
lambda= 0       cvl= -29.57789
Warning message:
In xy.coords(x, y, xlabel, ylabel, log) :
  1 x value <= 0 omitted from logarithmic plot

> optlambda1<-optL1(resp~age+smear+infltr+lab+blasts+temp)
lambda= 62.09239        cvl= -34.46502
lambda= 100.4676        cvl= -36.1672
lambda= 38.37521        cvl= -30.77353
lambda= 23.71718        cvl= -27.59435
lambda= 14.65803        cvl= -26.23036
lambda= 9.059158        cvl= -26.15863
lambda= 11.17701        cvl= -26.02396
lambda= 11.56675        cvl= -26.03096
lambda= 11.09573        cvl= -26.02286
lambda= 10.31783        cvl= -26.05385
lambda= 10.41474        cvl= -26.04852
lambda= 10.83562        cvl= -26.02956
lambda= 11.07756        cvl= -26.0231
lambda= 11.11126        cvl= -26.02306
lambda= 11.09519        cvl= -26.02285
lambda= 11.08845        cvl= -26.02284
lambda= 11.09143        cvl= -26.0228
lambda= 11.09166        cvl= -26.02281
lambda= 11.09006        cvl= -26.0228
```

Ήδη από το Σχήμα 3.3 είναι εμφανές ότι η βέλτιστη τιμή του λ είναι κοντά στο 11. Η ακριβής τιμή της παραμέτρου προκύπτει με την εντολή optlambda1$lambda από την οποία παίρνουμε $\lambda = 11.09143$. Αυτή είναι η τιμή που πρέπει να χρησιμοποιήσουμε ως ρυθμιστική παράμετρο για την εφαρμογή της μεθόδου παλινδρόμησης με ποινή L_1.



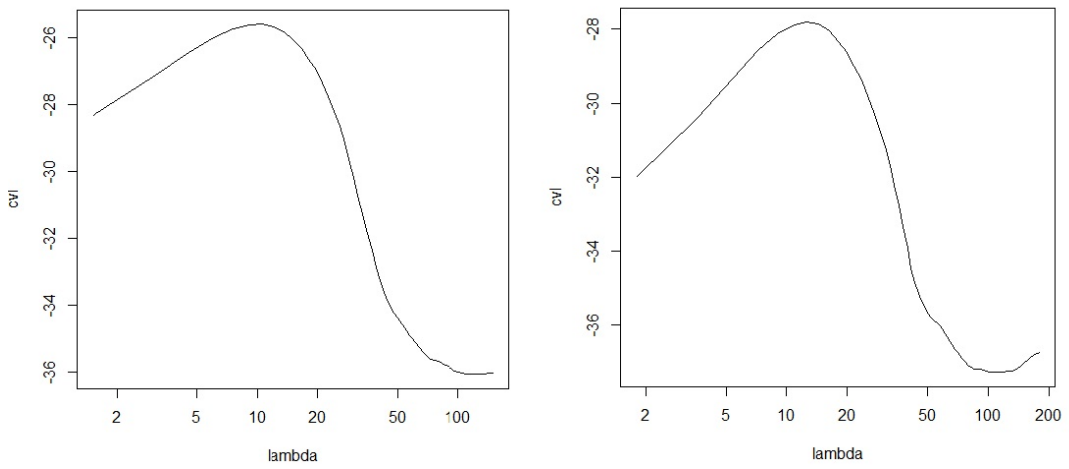
Σχήμα 3.3: cvl συναρτήσει του λ

```
> optlambda1$lambda
[1] 11.09143
```

Αξίζει να σημειωθεί ότι στα παραπάνω, δεδομένου ότι στη συνάρτηση profL1 δεν ορίσαμε την τιμή του fold για τον υπολογισμό της cvl, το βέλτιστο λ που έχουμε υπολογίσει αντιστοιχεί σε fold=0. Οι συνήθεις τιμές για το όρισμα αυτό είναι fold= 0, 5 ή 10. Εκτελώντας τους ίδιους υπολογισμούς και για τις άλλες δύο τιμές του fold διαπιστώνουμε ότι η βέλτιστη τιμή του λ δεν αλλάζει σημαντικά. Πράγματι, για fold=5 παίρνουμε $\lambda = 10.06656$ ενώ για fold=10 παίρνουμε $\lambda = 12.98347$. Στο Σχήμα 3.4 παρουσιάζεται η συμπεριφορά των αντίστοιχων συναρτήσεων της cvl για τις άλλες δύο τιμές του fold.

Είμαστε πλέον σε θέση να εκτελέσουμε την ανάλυση παλινδρόμησης με τη μέθοδο LASSO με σκοπό να προσαρμόσουμε το μοντέλο, να βρούμε δηλαδή τις στατιστικά σημαντικές μεταβλητές και τους συντελεστές τους. Αυτό γίνεται με την εντολή *penalized* στην οποία ως όρισμα εκτός από τις συμμεταβλητές πρέπει να δωθεί και η τιμή της ρυθμιστικής παραμέτρου λ όπως αυτή υπολογίστηκε παραπάνω.

```
> fitpenalizedL1<-penalized(resp~age+smear+infltr+lab+blasts+temp,lambda1=11.09143)
```



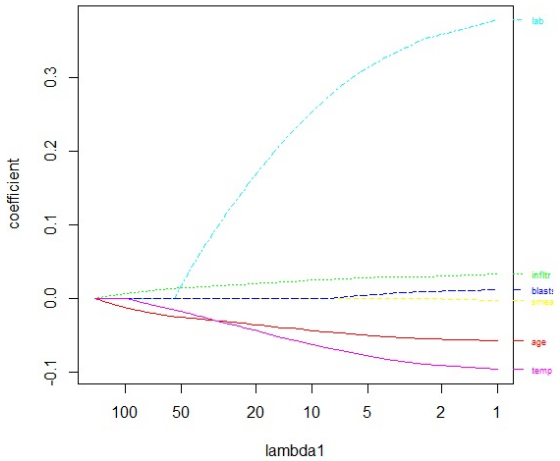
Σχήμα 3.4: cvl συναρτήσει του λ για fold=5 αριστερά και fold=10 δεξιά.

```
# nonzero coefficients: 5

> coefficients(fitpenalizedL1)
(Intercept)         age      infltr         lab         temp
57.42035088 -0.04220828   0.02435398   0.24229158  -0.05963115

> fitpenalizedL1
Penalized logistic regression object
7 regression coefficients of which 5 are non-zero

Loglikelihood =  -21.17528
L1 penalty =     4.087025       at lambda1 =  11.09143
```

Όπως μπορεί να διαπιστωθεί, το τελικό μοντέλο σύμφωνα με τη μέθοδο έχει 5 μη μηδενικούς συντελεστές (συμπεριλαμβανομένου του σταθερού όρου). Άρα από τις 6 συμμεταβλητές που είχαμε αρχικά, οι 2 έχουν μηδενιστεί. Με την εντολή *coefficients* προκύπτει ότι οι μεταβλητές smear και blasts είναι αυτές που δεν εμφανίζονται πλέον στο μοντέλο, δηλαδή είναι στατιστικά μη σημαντικές και ο συντελεστής τους στο μοντέλο είναι 0. Το αποτέλεσμα αυτό συμπίπτει με τα αποτελέσματα που προέκυψαν στην ανάλυση του μοντέλου ως γενικευμένο γραμμικό μοντέλο από τη διωνυμική κατανομή. Σχετικά με τους συντελεστές των στατιστι-

κά σημαντικών μεταβλητών, παρατηρούμε ότι αυτοί εμφανίζονται και στις δύο μεθόδους με τα ίδια πρόσημα αλλά στην L_1 είναι όλοι μικρότεροι. Ενδιαφέρον παρουσιάζει η απεικόνιση όλων των συντελεστών για διάφορες τιμές της παραμέτρου λ από την οποία μας επιτρέπεται να αντιληφθούμε την επίδραση του λ στις τιμές των συντελεστών των μεταβλητών στο μοντέλο. Αυτό είναι δυνατό να πραγματοποιηθεί με χρήση ενός ακόμα ορίσματος *steps* της συνάρτησης penalized. Χρησιμοποιώντας αυτό, η συνάρτηση ξεκινάει με την προσαρμογή του μοντέλου για τη μέγιστη τιμή του λ, δηλαδή εκείνη την τιμή για την οποία όλοι οι συντελεστές των μεταβλητών μηδενίζονται. Από αυτή την τιμή, η συνάρτηση συνεχίζει να κάνει προσαρμογή μοντέλου για όλο και μειούμενες τιμές της παραμέτρου λ μέχρι να φτάσει μια προκαθορισμένη τιμή την οποία εμείς έχουμε θέσει και αποτελεί επίσης ένα ακόμα όρισμα στη συνάρτηση. Το γράφημα των συντελεστών όπως αυτοί υπολογίζονται για τις διάφορες τιμές του λ γίνεται με χρήση της εντολής plotpath. Στην περίπτωση των δεδομένων που μελετούμε, η χρήση των παραπάνω συναρτήσεων και ορισμάτων καθώς και το Σχήμα 3.5 που προκύπτει φαίνονται στη συνέχεια.

```
> fitpenalizedL1<-penalized(resp~age+smear+infltr+lab+blasts+temp,
                                    lambda1=1,steps=100,trace=FALSE)
> plotpath(fitpenalizedL1,log="x")
```



Σχήμα 3.5: Συντελεστές των μεταβλητών συναρτήσει του λ

Φαίνεται λοιπόν και από το τελευταίο Σχήμα 3.5, ότι για λ κοντά στο 11, όπου όπως έχει

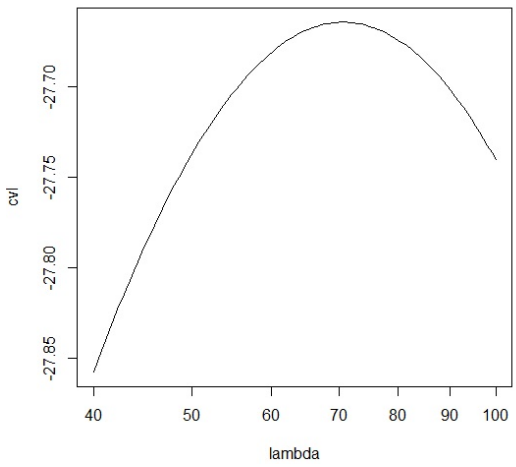
προκύψει αντιστοιχεί στη βέλτιστη τιμή της ρυθμιστικής παραμέτρου, υπάρχουν τέσσερεις μεταβλητές με μη μηδενικούς συντελεστές.

Στη συνέχεια ακολουθεί η εφαρμογή μιας ακόμα μεθόδου με ποινή, της Ridge (L_2) από αυτές που αναφέρθηκαν στο θεωρητικό μέρος. Οι συναρτήσεις - εντολές που χρησιμοποιούμε είναι ανάλογες με αυτές για την L_1. Βασική διαφορά μεταξύ των δύο μεθόδων, η οποία ήδη έχει τονιστεί με ιδιαίτερη έμφαση, είναι το γεγονός ότι τώρα δεν αναμένεται να μηδενιστεί ο συντελεστής καμίας εκ των συμμεταβλητών. Όλες οι μεταβλητές συμμετέχουν στην προσαρμογή του μοντέλου και καμία δεν απορρίπτεται εντελώς. Βέβαια αν κάποιος συντελεστής προκύψει εξαιρετικά μικρός κατ' απόλυτη τιμή, προφανώς αυτό μπορεί να ερμηνευθεί ότι η αντίστοιχη μεταβλητή δεν επηρεάζει σημαντικά την μεταβλητή απόκρισης και κατά συνέπεια δεν είναι στατιστικά σημαντική.

Η διαδικασία λοιπόν ξεκινάει με τον υπολογισμό της βέλτιστης τιμής της ρυθμιστικής παραμέτρου. Αν και η συνάρτηση ποινής L_2 συμπεριφέρεται πολύ καλύτερα από την L_1, δηλαδή η συνάρτηση της cvl είναι μονοκόρυφη ως προς λ και η μέθοδος συγκλίνει πάντα στο ολικό μέγιστο, εκτελούμε και εδώ την εντολή *profL2*. Στην περίπτωση αυτή υπάρχουν δύο επιπλέον ορίσματα, το *minlambda2* και *maxlambda2* τα οποία πρέπει να δωθούν και αντιστοιχούν στην ελάχιστη και στη μέγιστη τιμή της ρυθμιστικής παραμέτρου μεταξύ των οποίων θα υπολογιστεί η cvl προκειμένου να αναζητηθεί το σημείο μεγιστοποίησης. Στο μοντέλο που μελετούμε δώθηκαν οι τιμές minlambda2=40 και maxlambda2=100 αφού διαπιστώθηκε σε συνδυασμό με το αντίστοιχο Σχήμα 3.6 ότι ανάμεσα σε αυτές τις τιμές βρίσκεται το μέγιστο. Η ακριβής βέλτιστη τιμή της παραμέτρου, *optlambda2* λαμβάνεται μετά την εκτέλεση της εντολής *optL2*. Τα αποτελέσματα όπως ακριβώς ελήφθησαν φαίνονται ακολούθως.

```
> L2profile<-profL2(resp~age+smear+infltr+lab+blasts+temp,
                              minlambda2=40,maxlambda2=100,plot=TRUE)
lambda= 100             cvl= -27.74052
lambda= 99.07872        cvl= -27.73652
lambda= 98.16594        cvl= -27.73263
lambda= 97.26156        cvl= -27.72885
lambda= 96.36551        cvl= -27.72518
lambda= 95.47772        cvl= -27.72161
```

```
lambda= 94.5981          cvl= -27.71815

.   .   .   .   .   .   .   .   .   .

.   .   .   .   .   .   .   .   .   .

.   .   .   .   .   .   .   .   .   .
lambda= 75.75514         cvl= -27.66752
lambda= 75.05723         cvl= -27.66677
lambda= 74.36574         cvl= -27.66612
lambda= 73.68063         cvl= -27.66559
lambda= 73.00183         cvl= -27.66516
lambda= 72.32928         cvl= -27.66484
lambda= 71.66293         cvl= -27.66462
lambda= 71.00271         cvl= -27.66452
lambda= 70.34858         cvl= -27.66452
lambda= 69.70048         cvl= -27.66463
lambda= 69.05834         cvl= -27.66484
lambda= 68.42213         cvl= -27.66516

.   .   .   .   .   .   .   .   .   .

.   .   .   .   .   .   .   .   .   .

.   .   .   .   .   .   .   .   .   .
lambda= 43.07415         cvl= -27.81214
lambda= 42.67732         cvl= -27.81753
lambda= 42.28415         cvl= -27.82302
lambda= 41.89459         cvl= -27.82859
lambda= 41.50863         cvl= -27.83425
lambda= 41.12622         cvl= -27.83999
lambda= 40.74733         cvl= -27.84582
lambda= 40.37194         cvl= -27.85174
lambda= 40               cvl= -27.85774


> optlambda2<-optL2(resp~age+smear+infltr+lab+blasts+temp)
```



Σχήμα 3.6: cvl συναρτήσει του λ για την L_2 μέθοδο

```
lambda= Inf             cvl= -36.29271
lambda= 1               cvl= -30.93527
lambda= 10              cvl= -29.31644
lambda= 100             cvl= -27.74052
lambda= 1000            cvl= -30.65673
lambda= 388.1464        cvl= -29.23535
lambda= 621.8536        cvl= -29.95595
lambda= 243.7073        cvl= -28.56657
lambda= 154.4391        cvl= -28.04314
lambda= 99.26824        cvl= -27.73734
lambda= 65.17081        cvl= -27.66862
lambda= 56.71687        cvl= -27.69449
lambda= 73.77069        cvl= -27.66565
lambda= 70.55537        cvl= -27.6645
lambda= 70.7949         cvl= -27.6645
lambda= 70.68333        cvl= -27.6645
lambda= 70.683          cvl= -27.6645
lambda= 70.61935        cvl= -27.6645
lambda= 70.65869        cvl= -27.6645
```

```
lambda= 70.67101        cvl= -27.6645
```

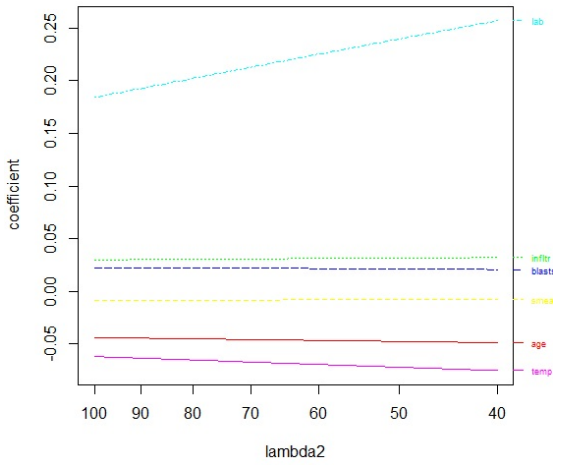
```
> optlambda2$lambda
[1] 70.683
```

Ήδη από το Σχήμα 3.6 βλέπουμε ότι η βέλτιστη τιμή της παραμέτρου λ είναι γύρω στο 70. Από την εντολή optlamba2$lamba παίρνουμε τελικά ότι η βέλτιστη τιμή της ρυθμιστικής παραμέτρου με την οποία μπορούμε να προχωρήσουμε στην προσαρμογή του μοντέλου με τη μέθοδο L_2 είναι $\lambda = 70.683$. Αξίζει να σημειωθεί στο σημείο αυτό ότι ακριβώς όπως και στην περίπτωση της μεθόδου L_1, το γράφημα 3.6 καθώς και η τιμή του λ που υπολογίστηκαν, έχουν εκτελεστεί για fold=0 της cvl. Με αυτή την τιμή λοιπόν περνάμε στην εκτίμηση των συντελεστών για τις συμμεταβλητές στο μοντέλο. Αυτό γίνεται και πάλι με χρήση της εντολής *penalized*.

```
> fitpenalizedL2<-penalized(resp~age+smear+infltr+lab+blasts+temp,lambda2=70.683)
```

```
> coefficients(fitpenalizedL2)
 (Intercept)          age        smear       infltr          lab       blasts
65.325325886 -0.045769594 -0.008277805  0.030755277  0.212146554  0.021880708
        temp
-0.067083609
```

Παρατηρούμε ότι οι συντελεστές των μεταβλητών οι οποίοι δεν μηδενίζονται σύμφωνα με την L_1 δεν διαφέρουν σημαντικά συγκριτικά με την L_2. Συνεπώς οι μεταβλητές age, infltr, lab και temp επηρεάζουν τη μεταβλητή απόκρισης με το ίδιο βάρος στις δύο μεθόδους ενώ παρατηρούμε ότι οι άλλες δύο μεταβλητές που απομένουν smear και blasts είναι οι δύο μεταβλητές με τους μικρότερους συντελεστές κατ΄ απόλυτη τιμή. Η επίδραση της ρυθμιστικής παραμέτρου πάνω στους συντελεστές φαίνεται στο Σχήμα 3.7 στο οποίο είναι προφανές ότι όσο η τιμή του λ μεγαλώνει τόσο οι συντελεστές συρρικνώνονται.



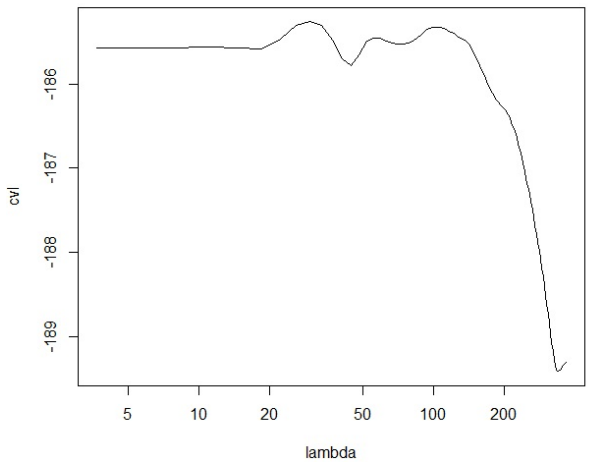
Σχήμα 3.7: Συντελεστές των μεταβλητών συναρτήσει του λ

3.4 Εφαρμογή L_1 και L_2 - package *penalized* - μοντέλο Cox

Σε συνέχεια της προηγούμενης παραγράφου, εφαρμόζουμε τις μεθόδους L_1 και L_2 όταν τα δεδομένα θεωρηθούν αποκομμένοι χρόνοι επιβίωσης και εξεταστεί η επίδραση των συμμεταβλητών στους χρόνους επιβίωσης μέσω του μοντέλου του Cox. Οι εντολές που χρησιμοποιούμε είναι οι ίδιες. Η μόνη διαφορά είναι ο τρόπος που δηλώνουμε τη μεταβλητή απόκρισης που είναι ο χρόνος επιβίωσης δηλαδή η μεταβλητή surv και οι αποκομμένοι χρόνοι που δηλώνονται μέσω της δίτιμης μεταβλητής status. Ξεκινάμε με την αναζήτηση της βέλτιστης τιμής του λ προκειμένου στη συνέχεια να χρησιμοποιηθεί για την εκτίμηση των συντελεστών. Μέσω της εντολής *profL1* υπολογίζεται η τιμή της cvl για διάφορες τιμές του λ και παίρνουμε το Σχήμα 3.8.

```
> CoxfitpenalizedL1<-profL1(Surv(surv,status)~age+smear+infltr+lab+blasts+temp,
                                                                    plot=TRUE)
lambda= 366.6733        cvl= -189.3089
lambda= 362.9696        cvl= -189.3186
lambda= 359.2658        cvl= -189.3312
lambda= 355.562         cvl= -189.3455
lambda= 351.8583        cvl= -189.3622
```

```
lambda= 348.1545        cvl= -189.3833
lambda= 344.4507        cvl= -189.4044
lambda= 340.7469        cvl= -189.416
.   .   .   .   .   .   .   .   .   .
.   .   .   .   .   .   .   .   .   .
.   .   .   .   .   .   .   .   .   .
lambda= 101.4197        cvl= -185.3212
lambda= 100.089         cvl= -185.3216
lambda= 100.9102        cvl= -185.321
lambda= 100.8652        cvl= -185.321
lambda= 100.8196        cvl= -185.3209
lambda= 100.7741        cvl= -185.3209
lambda= 100.7459        cvl= -185.3209
lambda= 100.7851        cvl= -185.3209
lambda= 100.7633        cvl= -185.3209
> optCoxfitpenalizedL1\$lambda
[1] 100.7741
> optCoxfitpenalizedL1$cvl
[1] -185.3209
```



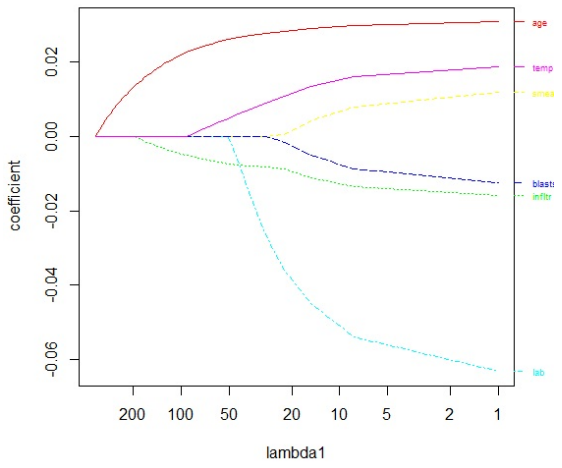
Σχήμα 3.8: Συνάρτηση cvl για τις διάφορες τιμές του λ

Όπως μπορεί να παρατηρήσει κανείς, συνδυάζοντας το βέλτιστο $\lambda = 100.7741$ που έδωσε η μέθοδος με τη γραφική παράσταση της cvl ως συνάρτησης του λ, η τιμή που προέκυψε δεν είναι το ολικό μέγιστο αλλά απλά ένα τοπικό μέγιστο της συνάρτησης. Δεν είναι λοιπόν αυτή η κατάλληλη τιμή της ρυθμιστικής παραμέτρου που πρέπει να χρησιμοποιήσουμε για την εφαρμογή της L_1. Λόγω της ιδιαιτερότητας αυτής στη συμπεριφορά της cvl στην περίπτωση της L_1 όπως αναφέρθηκε και νωρίτερα, η επιλογή του βέλτιστου λ πρέπει να συνοδεύεται πάντα από την εκτέλεση της συνάρτησης $\mathrm{prof}L_1$ προκειμένου να εντοπισθεί πιθανή σύγκλιση της μεθόδου σε κάποιο τοπικό μέγιστο. Προκειμένου να ξεπεράσουμε αυτή τη δυσκολία παρατηρούμε από τη γραφική παράσταση ότι το μέγιστο βρίσκεται περίπου κοντά 30. Για το λόγο αυτό περιορίζουμε την αναζήτηση του μεγίστου μεταξύ των τιμών *minlambda1* και *maxlambda1* όπως φαίνεται παρακάτω. Η βέλτιστη τιμή της παραμέτρου λ είναι τώρα $\lambda = 29.43065$.

```
> optCoxfitpenalizedL1<-optL1(Surv(surv,status)~
          age+smear+infltr+lab+blasts+temp,minlambda1=1,maxlambda1=50)
lambda= 19.71633        cvl= -185.5605
lambda= 31.28367        cvl= -185.2695
lambda= 38.43267        cvl= -185.5547
lambda= 29.11875        cvl= -185.2622
lambda= 29.77575        cvl= -185.2622
lambda= 29.47948        cvl= -185.262
lambda= 29.48495        cvl= -185.2621
lambda= 29.34169        cvl= -185.262
lambda= 29.41206        cvl= -185.262
lambda= 29.43446        cvl= -185.262
lambda= 29.43065        cvl= -185.2619
lambda= 29.42331        cvl= -185.262
lambda= 29.42889        cvl= -185.262
> optCoxfitpenalizedL1$lambda
[1] 29.43065
```

```
> optCoxfitpenalizedL1$cvl
[1] -185.2619
```

Εφαρμόζουμε τη μέθοδο για την προσαρμογή του μοντέλου του Cox με την μέθοδο ποινής L_1 και προκύπτουν 4 μη μηδενικοί συντελεστές δηλαδή στατιστικά σημαντικές μεταβλητές. Παρατηρούμε ότι οι μεταβλητές age, infltr, lab και temp είναι αυτές που έμειναν στο μοντέλο. Επιπλέον οι μεταβλητές infltr και temp έχουν πολύ μικρούς συντελεστές και μπορούμε να πούμε ότι επιδρούν οριακά στο χρόνο επιβίωσης. Αυτά τα αποτελέσματα συμπίπτουν με εκείνα στην αρχική μας ανάλυση του μοντέλου ως μοντέλο αναλογικής διακινδύνευσης του Cox. Στο Σχήμα 3.9 απεικονίζεται η συμπεριφορά των συντλεστών για κάθε συμμεταβλητή για τις διάφορες τιμές της παραμέτρου λ.



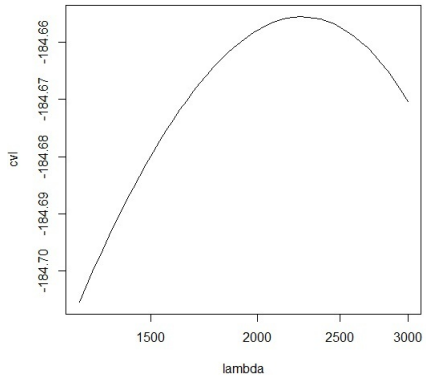
Σχήμα 3.9: Συντελεστές των μεταβλητών συναρτήσει του λ

```
> CoxCoefL1<-penalized(Surv(surv,status)~age+smear+infltr+lab+blasts+temp,
                                                        lambda1=29.43065)
# nonzero coefficients: 4
> coefficients(CoxCoefL1)
         age       infltr          lab         temp
 0.027846210 -0.008259883 -0.026284444  0.008924687
```

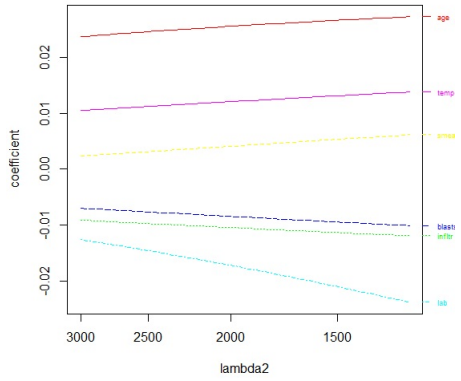
Τέλος εφαρμόζουμε και την L_2 μέθοδο με ποινή στο μοντέλο του Cox χωρίς να παρατηρούμε κάποια ιδιαίτερη διαφοροποίηση στα αποτελέσματα αφού και εδώ φαίνεται ουσιαστικά να υπερισχύουν οι τέσσερεις μεταβλητές με μη μηδενικούς συντελεστές όπως και στην L_1 ενώ οι άλλες δύο μεταβλητές έχουν τους μικρότερους συντελεστές κατά απόλυτη τιμή. Όλα τα διαγράμματα της συμπεριφοράς της cvl, (Σχήμα 3.10), των συντελεστών συναρτήσει της παραμέτρου λ (Σχήμα 3.11) καθώς και τα αποτελέσματα που εξήχθησαν φαίνονται στη συνέχεια.

```
> CoxfitpenalizedL2<-profL2(Surv(surv,status)~age+smear+infltr+lab+blasts+temp,
                                        minlambda2=500,maxlambda2=3000,plot=TRUE)
lambda= 3000    cvl= -184.6704
lambda= 2946.193        cvl= -184.6685
lambda= 2893.35         cvl= -184.6667
.   .   .   .   .   .   .   .   .   .
lambda= 2414.348        cvl= -184.6564
lambda= 2371.045        cvl= -184.656
lambda= 2328.518        cvl= -184.6558
lambda= 2286.755        cvl= -184.6556
lambda= 2245.74         cvl= -184.6556
lambda= 2205.461        cvl= -184.6557
lambda= 2165.904        cvl= -184.6559
lambda= 2127.057        cvl= -184.6562
.   .   .   .   .   .   .   .   .   .
lambda= 1304.836        cvl= -184.6975
lambda= 1281.433        cvl= -184.7001
lambda= 1258.45         cvl= -184.7027
lambda= 1235.878        cvl= -184.7055


> optCoxfitpenalizedL2<-optL2(Surv(surv,status)~age+smear+infltr+lab+blasts+temp)
lambda= Inf             cvl= -189.3089
```

```
lambda= 1               cvl= -185.6072
lambda= 10              cvl= -185.5768
lambda= 100             cvl= -185.3401
lambda= 1000            cvl= -184.7417
lambda= 10000           cvl= -185.2484
lambda= 3881.464        cvl= -184.7144
lambda= 6218.536        cvl= -184.8967
lambda= 2437.073        cvl= -184.6566
lambda= 1170.071        cvl= -184.7141
lambda= 2523.904        cvl= -184.6578
lambda= 2327.692        cvl= -184.6558
lambda= 1885.52         cvl= -184.6606
lambda= 2244.834        cvl= -184.6556
lambda= 2255.12         cvl= -184.6556
lambda= 2256.691        cvl= -184.6556
lambda= 2256.554        cvl= -184.6556
lambda= 2255.906        cvl= -184.6556
lambda= 2256.306        cvl= -184.6556
> optCoxfitpenalizedL2\$lambda
[1] 2256.554
> CoxCoefL2<-penalized(Surv(surv,status)~age+smear+infltr+lab+blasts+temp,
                                                        lambda2=2256.554)


> coefficients(CoxCoefL2)
         age        smear       infltr          lab       blasts         temp
 0.024972897  0.003564256 -0.010028457 -0.015702202 -0.007990880  0.011577720
```



Σχήμα 3.10: Συνάρτηση cvl για τις διάφορες τιμές του λ



Σχήμα 3.11: Συντελεστές των μεταβλητών συναρτήσει του λ

3.5 Εφαρμογή μεθόδων SCAD και (I)SIS - library(SIS)

Συνεχίζοντας με το ίδιο σύνολο δεδομένων, εφαρμόζουμε τις μεθόδους SIS, ISIS και SCAD που περιγράψαμε στο θεωρητικό μέρος (παράγραφοι 2.5, 2.6). Για την εφαρμογή της SIS, ISIS οι οποίες πραγματοποιούνται ταυτόχρονα, απαιτείται το κεντράρισμα (centering) των ανεξάρτητων μεταβλητών αλλά όχι η πλήρης τυποποίησή τους, δε χρειάζεται δηλαδή και scaling, ενώ οι μεταβλητές των χρόνων επιβίωσης (surv) και κατάστασης (status) δεν τροποποιούνται. Απαραίτητη είναι η βιβλιοθήκη *SIS* της *R* (Fan, J., Feng, Y., Samworth, R., Wu, Y., 2010) η οποία περιέχει τις απαραίτητες εντολές για την πραγματοποίηση της μεθόδου. Σημειώνουμε ότι η μεταβλητή των χρόνων επιβίωσης surv μετονομάζεται μόνο στην περίπτωση αυτή σε time. Έτσι οι συμμεταβλητές παίρνουν τώρα τη μορφή:

```
$x
              age       smear      infltr        lab     blasts        temp
 [1,] -29.8627451  12.0588235 -19.1960784 -2.8039216 -6.7392157  -6.1372549
 [2,] -24.8627451  -1.9411765   2.8039216  6.1960784 27.6607843  33.8627451
 [3,] -23.8627451  -4.9411765  -3.1960784  2.1960784  0.1607843 -14.1372549
 [4,] -23.8627451  -1.9411765   5.8039216  6.1960784 13.6607843   3.8627451
 [5,] -22.8627451  29.0588235  36.8039216 -3.8039216  0.1607843 -16.1372549
.   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
.   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
.   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
[49,]  25.1372549  -5.9411765   1.8039216  7.1960784  0.8607843  -6.1372549
[50,]  27.1372549   3.0588235  10.8039216 -0.8039216 -5.8392157 -10.1372549
[51,]  30.1372549   7.0588235  14.8039216 -2.8039216 -5.8392157 -10.1372549
attr(,"scaled:center")
       age      smear     infltr        lab     blasts       temp
 49.862745  65.941176  58.196078   9.803922   7.339216 996.137255
```

Η εντολή SIS μπορεί τώρα να εφαρμοστεί και παίρνουμε τα παρακάτω αποτελέσματα. Από το *SISind* διαπιστώνουμε ποιες μεταβλητές δίνει η μέθοδος ως στατιστικά σημαντικές και αυτές στην περίπτωσή μας είναι οι μεταβλητές με δείκτες 1, 3 και 4 δηλαδή η X_1, η X_3 και η X_4. Αυτές αντιστοιχούν στις μεταβλητές age, infltr και lab. Στη συνέχεια η επαναληπτική SIS έδωσε στατιστικά σημαντική μεταβλητή μόνο την πρώτη συμμεταβλητή δηλαδή την ηλικία όπως ακριβώς βρέθηκε από την αρχή στην πρώτη προσαρμογή των δεδομένων μας στο μοντέλο του Cox. Από τα αποτελέσματα *SIScoef* και *ISIScoef* φαίνονται και οι συντελεστές των αντίστοιχων μεταβλητών. Για τη μεταβλητή της ηλικίας που μας ενδιαφέρει ο συντελεστής συμπίπτει με αυτόν της απλής προσαρμογής του Cox. Σημειώνεται ότι οι συντελεστές που δίνονται στη SIScoef, έχουν προκύψει από τη SIS στην οποία στη συνέχεια έχει εφαρμοστεί η SCAD.

```
> SISdata<-list(x=centered.x,time=time,status=status)
> fitSISdata<-SIS(SISdata,model='cox')
```

```
$SISind
[1] 1 3 4

$ISISind
[1] 1

$SIScoef
[1]  0.031176504  0.00000 -0.009318964 -0.041558844  0.00000  0.00000

$ISIScoef
[1] 0.03239672 0.00000 0.00000 0.00000 0.00000 0.00000
```

Με την εντολή getfinalSCADcoefCOX μπορούμε απευθείας να πάρουμε τους συντελεστές παλινδρόμησης με τη μέθοδο SCAD για το μοντέλο του Cox αφού έχει εφαρμοστεί η (I)SIS. Ως όρισμα στην εντολή αυτή απαιτούνται οι δείκτες των μεταβλητών με μη μηδενικούς συντελεστές που έδωσε η (I)SIS. Αυτό γίνεται μέσω του ορίσματος pickind το οποίο στο παράδειγμά μας έχει τεθεί ίσο με 1. Επιπλέον απαιτείται μια αρχική λύση, για την οποία δηλώνεται μέσω του ορίσματος inittype η μέθοδος με την οποία θα υπολογιστεί και η οποία μπορεί να είναι είτε χωρίς ποινή είτε η L_1. Όπως φαίνεται παρακάτω οι τελικοί συντελεστές για τη μεταβλητή age δεν διαφέρουν μεταξύ των δύο μεθόδων με τις οποίες θα προκύψει η αρχική λύση.

```
> getfinalSCADcoefCOX(x,time,status,method='efron',pickind=1,folds=NULL,
                 eps0=1e-5,tune.method="AIC",inittype="L1",detailed=FALSE)
$wt.initsoln
[1] 0.001560069

$SCADcoef
[1] 0.03236595 0.00000 0.00000 0.00000 0.00000 0.00000

> getfinalSCADcoefCOX(x,time,status,method='efron',pickind=1,folds=NULL,
 eps0=1e-5,tune.method="AIC",inittype="NoPen",detailed=FALSE)
```

```
$wt.initsoln
x[, pickind]
  0.03239672

$SCADcoef
[1] 0.03239672 0.00000 0.00000 0.00000 0.00000 0.00000
```

Κάνοντας την ίδια διαδικασία και για τη λογιστική παλινδρόμηση παίρνουμε ότι οι στατιστικά σημαντικές μεταβλητές τόσο με τη SIS όσο και με την ISIS είναι πρώτη και η τέταρτη δηλαδή οι age και lab. Αντίστοιχα αποτελέσματα είχαν προκύψει και κατά την απλή εφαρμογή της λογιστικής παλινδρόμησης όπου είχαμε δει ότι υπερισχύει η lab. Σε αντίθεση με το μοντέλο του Cox παίρνουμε και τιμές για το σταθερό όρο. Η μέθοδος απαιτεί το χωρισμό του δείγματος σε trainset και σε testset και εφαρμογή της cvl. Το μέγεθος του trainset ορίζεται στα 2/3 του συνολικού μεγέθους δείγματος ενώ το υπόλοιπο 1/3 χρησιμοποιείται ως το testset.

```
$SISind
[1] 1 4

$ISISind
[1] 1 4

$SIScoef
[1]  0.6090868 -0.0581058  0.00000  0.00000  0.2235144  0.00000  0.00000

$ISIScoef
[1]  0.6090868 -0.0581058  0.00000  0.00000  0.2235144  0.00000  0.00000
```

Χρησιμοποιώντας ως δείκτες των στατιστικά σημαντικών μεταβλητών τις 1 και 4 (οι οποίοι έχουν αποθηκευτεί στο διάνυσμα ind) όπως αυτές προέκυψαν από την (I)SIS, παίρνουμε με χρήση της εντολής *getfinalSCADcoef* απευθείας τους συντελεστές των μεταβλητών για το γενικευμένο γραμμικό μοντέλο της λογιστικής παλινδρόμησης οι οποίοι είναι αρκετά κοντά

σε αυτούς που έδωσε αρχικά η προσαρμογή του μοντέλου. Και στην περίπτωση αυτή φυσικά παίρνουμε τιμή και για το σταθερό όρο.

```
> getfinalSCADcoef(centered.x,resp,pickind=ind,folds=0,eps0=1e-5,
              family=binomial(),tune.method="AIC",inittype="NoPen",detailed=FALSE)
$wt.initsoln
       ones         age         lab
-0.13243551 -0.05020138  0.26369582

$SCADcoef
[1] -0.13243551 -0.05020138  0.00000  0.00000  0.26369582  0.00000  0.00000
```

3.6 Συμπεράσματα

Σχετικά με τα αποτελέσματα για το σύνολο των δεδομένων που μελετήθηκε, θα λέγαμε ότι στην περίπτωση που ενδιαφερόμαστε για το χρόνο επιβίωσης των ασθενών, τότε η ηλικία φαίνεται να παίζει το σημαντικότερο ρόλο, ενώ εμφανίζεται μια μικρή διαφοροποίηση μεταξύ ατόμων νεώτερης και μεγαλύτερης ηλικίας. Στην περίπτωση της λογιστικής παλινδρόμησης τα odds της ανταπόκρισης στη θεραπεία φαίνεται να εξαρτώνται από τις μεταβλητές age, infltr, lab, temp με περισσότερη βαρύτητα στη μεταβλητή lab η οποία ακολουθείται από την ηλικία. Αυτό που κατηγορηματικά μπορούμε να πούμε είναι ότι οι μεταβλητές smear και blasts σε καμία περίπτωση δεν επηρεάζουν τα αποτελέσματα. Όλες οι μέθοδοι καταλήγουν σε παρόμοια συμπεράσματα.

Αξίζει να σημειωθούν δύο σημαντικά στοιχεία. Πρώτον, όπως έχουμε αναφέρει οι μέθοδοι με ποινή αντιμετωπίζουν το πρόβλημα της πολυσυγγραμμικότητας που πιθανόν να υπάρχει μεταξύ των συμμεταβλητών αναδεικνύοντας τις πλέον στατιστικά σημαντικές μεταβλητές. Στο σύνολο των δεδομένων που αναλύθηκαν παρατηρήθηκε έντονη συσχέτιση μεταξύ των μεταβλητών infltr και smear οι οποίες έχουν συντελεστή συσχέτισης $r = 0.84$. Επίσης ελαφρά συσχέτιση παρουσιάζει και η μεταβλητή blasts με όλες τις υπόλοιπες εκτός της age με συντελεστές συσχέτισης γύρω στο 0.3. Συνεπώς, μπορούμε να εμπιστευθούμε περισσότερο τα αποτελέσματα μετά την εφαρμογή των μεθόδων με ποινή. Δεύτερον, από τις δύο

κατευθύνσεις ανάλυσης (μοντέλο Cox και λογιστική παλινδρόμηση) η σωστή κατεύθυνση είναι αυτή της προσαρμογής του Cox. Δεδομένου ότι διαθέτουμε τη μεταβλητή του χρόνου επιβίωσης των ασθενών, ο οποίος είναι διαφορετικός για κάθε άτομο, η ανάλυση γίνεται ως προς το χρόνο επιβίωσης. Η λογιστική παλινδρόμηση εξετάζει τη σχετική πιθανότητα να ανταποκριθεί ο ασθενής στη θεραπεία. Η πιθανότητα αυτή είναι μεν διαφορετική για κάθε άτομο αλλά δεν αναφέρεται στην ίδια χρονική στιγμή γεγονός που αποτελεί λανθασμένη χρήση του μοντέλου της λογιστικής παλινδρόμησης. Η εφαρμογή της βέβαια στην παρούσα εργασία έγινε με σκοπό να δείξουμε την υλοποίηση των μεθόδων που παρουσιάσαμε στο θεωρητικό μέρος σε πραγματικά δεδομένα και για την περίπτωση της λογιστικής παλινδρόμηση.

Συμπερασματικά, δεδομένης της σημασίας της ακρίβειας στις προβλέψεις που πρέπει να χαρακτηρίζει ένα μοντέλο που θέλουμε να χρησιμοποιηθεί γι αυτό το σκοπό, οι δύο βασικές μέθοδοι ποινής L_1 και L_2 ήρθαν να βελτιώσουν τη χρήση των ε.ε.τ.. Η θυσία της αμεροληψίας για χάρη της ακρίβειας είναι η φιλοσοφία και των δύο μεθόδων. Οι ανάγκες που δημιούργησαν τα σύγχρονα προβλήματα με το μεγάλο όγκο δεδομένων έπρεπε να αντιμετωπιστούν αποτελεσματικά και σε αυτό συνετέλεσαν πολύ οι ποινικοποιημένες μέθοδοι. Η συνεχής έρευνα σε αυτό τον τομέα είχε ως αποτέλεσμα οι δύο αυτές μέθοδοι να μελετηθούν σε βάθος και το πιο σημαντικό να εξελιγχθούν. Έτσι περάσαμε στα πιο γενικά ποινικοποιημένα ελάχιστα τετράγωνα και στις βελτιωμένες μεθόδους SCAD, SIS και ISIS οι οποίες αντιμετωπίζουν αποτελεσματικά τα προβλήματα της πολυσυγγραμμικότητας και των δεδομένων μεγάλης διάστασης.

Συγκρίνοντας μεταξύ των LASSO, Ridge και βέλτιστων υποσυνόλων B.S. αναφορικά με την ικανότητα πρόβλεψης, όπως ήταν αναμενόμενο, το αποτέλεσμα διαφοροποιείται ανάλογα με τα χαρακτηριστικά του δείγματος. Έτσι διαπιστώθηκε ότι στην περίπτωση που κάποιες μεταβλητές σαφώς υπερισχύουν κάποιων άλλων με την έννοια ότι είναι ξεκάθαρα στατιστικά σημαντικές η B.S. τα κατάφερε καλύτερα αλλά σε δεδομένα με μικρό αριθμό μεταβλητών. Όταν όμως όλες επιδρούν μέτρια ή καμία δεν επιδρά σημαντικά τότε η L_1 και η L_2 αντίστοιχα έδωσαν καλύτερα αποτελέσματα.

Προσομοιώσεις σε δεδομένα στα οποία εφαρμόστηκε η L_1 και η ISIS έδειξαν ότι η L_1 έδωσε μοντέλο αρκετά μεγαλύτερο σε μέγεθος από αυτό στο οποίο κατέληξε η ISIS. Συνεπώς, η L_1 εμφανίζει και μεγαλύτερα σφάλματα εκτίμησης. Αυτό οφείλεται στο γεγονός ότι η

L_1 κρατάει αρκετούς μη-μηδενικούς αλλά μικρούς συντελεστές που αντιστοιχούν σε μη σημαντικές μεταβλητές. Η ISIS λειτουργεί αποτελεσματικά σε περιπτώσεις που παρουσιάζεται σημαντική συσχέτιση μεταξύ των μεταβλητών.

Επίσης σε γραμμικά μοντέλα με αυξημένο επίπεδο θορύβου αλλά σχετικά μικρό μέγεθος δείγματος, τα αποτελέσματα έδειξαν ότι η L_1 λειτουργεί πάρα πολύ καλά σε σχέση με τις άλλες μεθόδους. Αν όμως το επίπεδο θορύβου μειωθεί και αυξηθεί το μέγεθος του δείγματος, τότε η SCAD υπερισχύει της L_1 ενώ στην περίπτωση αυτή η L_2 δεν έδωσε καλά αποτελέσματα. Αξίζει να σημειωθεί ότι η B.S έδωσε παρόμοια αποτελέσματα με τη SCAD.

Όλες αυτές οι μέθοδοι έχουν το πλεονέκτημα να βρίσκουν εφαρμογή εκτός από τα γραμμικά μοντέλα και στα γενικευμένα αλλά και σε μοντέλα δεδομένων επιβίωσης. Η εφαρμογή όλων των μεθόδων και κριτηρίων που περιγράψαμε έχει διευκολυνθεί πλέον σημαντικά από τη χρήση των αντίστοιχων στατιστικών πακέτων που έχουν ενσωματωθεί στα περισσότερα στατιστικά περιβάλλοντα. Η R προσφέρει σαφώς μια ολοκληρωμένη χρήση αυτών των μεθόδων.

Βιβλιογραφία

[1] Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In: *Second International Symposium on Information Theory, Petrox, B.N. and Caski, F. (eds). Budapest: Akademiai Kiado*, 267 - 281.

[2] Akaike, H. (1983) Information measures and model selection. *Bulletin of the International Statistical Institute* **50**, 277 - 290.

[3] Breiman, L. (1995) Better subset regression using the nonnegative garrote. *Technometrics* **37**(4), 373 - 384.

[4] Breslow, N.E. (1974) Covariance analysis for censored survival data. *Biometrics* **30**, 89 - 99.

[5] Breslow, N. E., Crowley, J. (1974) A large-sample study of the life table and product limit estimates under random censorship. *Annals of Statistics* **2**, 437 - 454.

[6] Burnham, K. P., Anderson, D. R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.), Springer-Verlag.

[7] le Cessie, S., van Houwelingen, J.C. (1992) Ridge estimators in logistic regression. *Appl. Statist.* **41**(1), 191 - 201.

[8] Claeskens, G., Hjort, N.L. (2003) The focused information criterion (with discussion). *Journal of the American Statistical Association* **98**, 879 - 899.

[9] Efron, B. (1977) The efficiency of Cox's likelihood for censored data. *Journal of the American Statistical Association* **72**, 557 - 565.

[10] Efron, B. (1983) Estimating the error rate of a prediction rule: Improvement on cross validation. *Journal of the American Statistical Association* **78**, 316 - 331.

[11] Fan, J., Feng, Y., Wu, Y. (2010) High-dimensional variable selection for Cox's proportional hazards model. In: *IMS Collections, Borrowing Strength: Theory Powering Applications - A Festschrift for Lawrence D. Brown* **6**, 70 - 86.

[12] Fan, J., Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties.*Journal of the American Statistical Association* **96**, 1348 - 1360.

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

[14] Fan, J., Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society Series B* **70**, 849 - 911.

[15] Frank, I.E., Friedman, J.H. (1993) A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109 - 148.

[16] Fu, W. (1998) Penalized Regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics* **7**, 397 - 416.

[17] Gill, J. (2001) *Generalized Linear Models: A Unified Approach.* Sage Publications, California.

[18] Goeman, J.J., (2010) L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal* **52**(1), 70 - 84.

[19] Goeman, J.J., van de Geer, S.A., de Kort, F., van Houwelingen, J.C. (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* **20**(1), 93 - 99.

[20] Harrell, F. (2002) *Regression Modeling Strategies: With applications to Linear Models, Logistic Regression and Survival Analysis.* Springer.

[21] Hoerl, A., Kennard, R. (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55 - 67.

[22] Kadane, J.B., Lazar N.A. (2004) Methods and criteria for model selection. *Journal of the American Statistical Association* **99**, 279 - 290.

[23] Lawless, J.F., Singhal, K. (1978) Efficient screening of non-normal regression models. *Biometrics* **43**, 318 - 327.

[24] Lee, E.T. (1980)*Statistical Methods for Survival Data Analysis.* Lifetime Learning Publications, Belmont, California.

[25] Marquardt, D., Snee, R. (1975) Ridge regression in practice. *The American Statistician* **29** (1), 3 - 20.

[26] Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics* **6**, 461 - 464.

[27] Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society* **B 64**(4), 583 - 639.

[28] Tibshirani, R.J. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B-Methodological* **58**(1), 267 - 288.

[29] Tibshirani, R.J. (1997) The lasso method for variable selection in the Cox model. *Statistics in Medicine* **16**, 385 - 95.

[30] Van Houwelingen, H.C. (2001) Shrinkage and penalized likelihood as methods to improve predictive accuracy. *Statistica Neerlandica* **55**, 17 - 34.

[31] Vereij, P., Van Houwelingen, H.C. (1993) Cross validation in survival analysis. *Statistics in Medicine* **12**(24), 2305 - 2314.

[32] Καρώνη, Χ. (2009) *Μοντέλα Αξιοπιστίας και Επιβίωσης.* Εκδόσεις Συμεών, Αθήνα.

[33] Καρώνη, Χ., Οικονόμου, Π. (2010) *Στατιστικά Μοντέλα Παλινδρόμησης*. Εκδόσεις Συμεών, Αθήνα.

packages

[34] Goeman, J., Meijer, R., Chaturvedi, N. (2012)
http://cran.r-project.org/web/packages/penalized/index.html

[35] Fan, J., Feng, Y., Samworth, R., Wu, Y. (2010)
http://cran.r-project.org/web/packages/SIS/index.html