ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Διπλωματική Εργασία

Τεχνικές, μέθοδοι και κριτήρια επιλογής βέλτιστου στατιστικού μοντέλου,

με τη βοήθεια του στατιστικού πακέτου της R

ΑΝΑΞΑΓΟΡΟΥ Χ. ΧΡΙΣΤΟΔΟΥΛΟΣ

Επιβλέπουσα: Καρώνη Χρυσηΐς,
Αναπληρώτρια Καθηγήτρια Ε.Μ.Π.

Επιτροπή καθηγητών: **Καρώνη Χρυσηΐς, Κουκουβίνος Χρήστος, Βόντα Φιλία**

Αθήνα, Ιούνιος 2013

ΠΕΡΙΛΗΨΗ

Τα τελευταία χρόνια έχουν αναπτυχθεί διάφορες τεχνικές και μέθοδοι, αλλά βελτιώθηκαν και κάποια κριτήρια επιλογής του βέλτιστου στατιστικού μοντέλου. Σκοπός τους είναι η επιλογή των μεταβλητών x που επηρεάζουν σημαντικά τη μεταβλητή απόκρισης y, μέσα από ένα σύνολο δεδομένων.

Στην παρούσα εργασία παρουσιάζουμε διάφορες τεχνικές, μεθόδους, ελέγχους αλλά και κριτήρια επιλογής για να καταλήξουμε στο βέλτιστο μοντέλο. Έμφαση δίνεται στις δύο μεθόδους με ποινή (L_1 και L_2 penalized) που βελτιώθηκαν τα τελευταία χρόνια και οι οποίες βασίζονται στην εισαγωγή μίας συνάρτησης ποινής στην πιθανοφάνεια. Ειδικότερα, οι συναρτήσεις ποινής που εξετάζουμε, η Lasso και η Ridge regression, βασίζονται στη συσχέτιση μεταξύ των μεταβλητών και μας δίνουν καλύτερα αποτελέσματα, καθώς αντιμετωπίζουν το φαινόμενο της πολυσυγγραμμικότητας.

Η επιλογή του μοντέλου γίνεται στο γενικό γραμμικό μοντέλο, στο μοντέλο λογιστικής παλινδρόμησης και στο μοντέλο αναλογικής διακινδύνευσης του Cox, με τη βοήθεια του στατιστικού πακέτου της R, χρησιμοποιώντας τις διάφορες μεθόδους, τεχνικές και κριτήρια, με διάφορες προσαρμογές σε πραγματικά δεδομένα.

Στο πρώτο κεφάλαιο, γίνεται μία εισαγωγή στο γενικό γραμμικό μοντέλο και στη βασική μέθοδο εκτίμησης των παραμέτρων με τη χρήση της μέγιστης πιθανοφάνειας. Επίσης, παρουσιάζονται οι τρεις διαδικασίες επιλογής μοντέλου με βήματα με τη βοήθεια κάποιου κριτηρίου, αλλά και τα σημαντικότερα μέτρα καταλληλότητας που χρησιμοποιούνται στο γενικό γραμμικό μοντέλο. Ακόμα, παρουσιάζονται οι τεχνικές L_1 και L_2 με ποινή. Το κεφάλαιο κλείνει με μια εφαρμογή στο γενικό γραμμικό μοντέλο με τη βοήθεια του στατιστικού πακέτου της R, με πραγματικά δεδομένα.

Στο δεύτερο κεφάλαιο, παρουσιάζεται εκτενώς η λογιστική παλινδρόμηση με κάποια εισαγωγικά στοιχεία για τα γενικευμένα γραμμικά μοντέλα. Όμως δίνεται έμφαση στους ελέγχους επιλογής του βέλτιστου μοντέλου, καθώς και στα κριτήρια επιλογής στα γενικευμένα γραμμικά μοντέλα, ιδιαίτερα στο μοντέλο λογιστικής παλινδρόμησης. Εδώ

κλείνει το κεφάλαιο με εφαρμογή στη λογιστική παλινδρόμηση με τη βοήθεια της R με πραγματικά δεδομένα. Σημαντική παρατήρηση στην εφαρμογή εδώ είναι η χρήση του πακέτου *glmulti*, για την επιλογή των σημαντικότερων μεταβλητών.

Τέλος, στο τρίτο κεφάλαιο, γίνεται αναφορά στα μοντέλα επιβίωσης και ιδιαίτερα στο μοντέλο αναλογικής διακινδύνευσης του Cox. Παρουσιάζονται οι ελέγχοι υποθέσεων για την επιλογή του βέλτιστου μοντέλου, οι τεχνικές L_1 και L_2 με ποινή, καθώς και κάποια κριτήρια επιλογής στο μοντέλο του Cox. Και εδώ κλείνει το κεφάλαιο με εφαρμογή στο συγκεκριμένο μοντέλο με τη βοήθεια της R, για την επιλογή του «καλύτερου» μοντέλου σε πραγματικά δεδομένα.

ABSTRACT

In recent years, several techniques and methods for selecting variables in statistical models have been developed and some selection criteria have been improved. Their aim is to identify those variables x that significantly affect the response y in a set of data.

In this thesis, we discuss various techniques, penalized methods, tests and selection criteria to find the optimal model. Emphasis is on two penalized methods (L_1 and L_2 penalized), the Lasso and Ridge regression, which are based on two penalty terms that are imposed on the likelihood function. In particular, the penalty terms that we consider are based on the correlation between the explanatory variables, and give better results as they handle the problem of multicollinearity.

The choice of the best model is studied in the general linear model, the logistic regression model and Cox's proportional hazards model, using various methods, techniques and selection criteria applied to real data sets using the R statistical package.

The first chapter contains an introduction to the general linear model and the basic method of parameter estimation using maximum likelihood. It also presents the three stepwise procedures of model selection and the most important measures of suitability used in the general linear model. Furthermore, the L_1 and L_2 penalized methods are presented. The chapter ends with an application of the general linear model to real data using the statistical package R.

The second chapter presents the generalized linear model and logistic regression, with emphasis on tests for the best model and the selection criteria in generalized linear models, particularly in logistic regression models. R is used to apply logistic regression to a real data set, employing the *glmulti* package for the selection of the statistically important variables.

Finally, in the third chapter, survival models are presented, specifically Cox's proportional hazards model. Hypothesis testing for the selection of the optimal model and

the selection criteria in this context are presented. The L_1 and L_2 penalized methods for this model are also presented. The chapter closes with an application to the selection of the "best" Cox model in a real data set, using the R package.

ΕΥΧΑΡΙΣΤΙΕΣ

Πρώτα πρώτα ευχαριστώ το Θεό που με αξίωσε να τελειώσω αυτή την εργασία. Όμως η παρούσα διπλωματική εργασία δεν θα μπορούσε να έχει ολοκληρωθεί επιτυχώς χωρίς τη βοήθεια και τη συμπαράσταση πολλών ανθρώπων.

Αισθάνομαι πρωτίστως την ανάγκη να ευχαριστήσω εκ βάθους καρδίας την Αναπληρώτρια Καθηγήτρια του Ε.Μ.Π., κα Χρυσηΐδα Ρ. Καρώνη, για τη δυνατότητα που μου προσέφερε να ασχοληθώ με αυτό το πολύ ενδιαφέρον θέμα, καθώς επίσης και για τη συνεχή και ακούραστη επίβλεψη και καθοδήγηση της. Επιπλέον για την πολύτιμη βοήθειά της ανά πάσα στιγμή και ώρα, αλλά και το συνεχές ενδιαφέρον της κατά τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας.

Ιδιαίτερες ευχαριστίες θα ήθελα να εκφράσω στους γονείς μου παπαΧρήστο και Ελένη Αναξαγόρου για την υπομονή τους και την αμέριστη υποστήριξή τους, καθ' όλη τη διάρκεια της εργασίας, αλλά και γενικά των σπουδών μου. Επίσης θέλω να ευχαριστήσω τα αδέρφια μου και γενικά όλη την οικογένεια μου. Τέλος αισθάνομαι την ανάγκη να ευχαριστήσω τους πολύ καλούς μου φίλους, ιδιαίτερα τον εκπληκτικό συγκάτοικό μου, τους τρελούς, αλλά πολύ συνεργάσιμους συμφοιτητές μου και τους αγαπημένους μου κουμπάρους για τη βοήθεια και τη συμπαράσταση τους, καθ' όλη τη διάρκεια της εργασίας μου.

Χριστόδουλος Αναξαγόρου

Αθήνα, Ιούνης 2013

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΛΗΨΗ..... 2

ABSTRACT..... 4

ΕΥΧΑΡΙΣΤΙΕΣ..... 6

ΚΕΦΑΛΑΙΑ

1. ΓΕΝΙΚΟ ΓΡΑΜΜΙΚΟ ΜΟΝΤΕΛΟ..... 10

1.1 ΕΙΣΑΓΩΓΗ 10

1.1.1 ΤΟ ΜΟΝΤΕΛΟ..... 10

1.1.2 Η ΧΡΗΣΗ ΤΟΥ..... 12

1.1.3 ΔΙΑΔΙΚΑΣΙΑ ΣΤΑΤΙΣΤΙΚΗΣ ΜΟΝΤΕΛΟΠΟΙΗΣΗΣ 12

1.2 ΒΕΛΤΙΩΣΗ ΚΑΙ ΑΝΑΠΤΥΞΗ ΜΟΝΤΕΛΟΥ 14

1.2.1 ΕΚΤΙΜΗΣΗ ΜΕ ΤΗ ΜΕΘΟΔΟ ΤΗΣ ΜΕΓΙΣΤΗΣ ΠΙΘΑΝΟΦΑΝΕΙΑΣ..... 14

1.2.2 ΔΙΑΔΙΚΑΣΙΕΣ ΕΠΙΛΟΓΗΣ ΜΟΝΤΕΛΟΥ ΜΕ ΒΗΜΑΤΑ..... 15

1.2.3 ΜΕΤΡΑ ΚΑΤΑΛΛΗΛΟΤΗΤΑΣ..... 19

1.2.4 ΜΕΘΟΔΟΙ L_1 ΚΑΙ L_2-PENALIZED..... 23

1.3 ΕΦΑΡΜΟΓΗ ΣΤΟ ΓΕΝΙΚΟ ΓΡΑΜΜΙΚΟ ΜΟΝΤΕΛΟ ΜΕ ΧΡΗΣΗ ΤΗΣ R. 29

1.3.1 ΠΕΡΙΓΡΑΦΗ ΕΦΑΡΜΟΓΗΣ ΚΑΙ ΔΕΔΟΜΕΝΑ..... 31

1.3.2 ΕΚΤΕΛΕΣΗ ΕΦΑΡΜΟΓΗΣ ΚΑΙ ΕΠΙΛΟΓΗ ΒΕΛΤΙΣΤΟΥ ΜΟΝΤΕΛΟΥ..... 32

1.3.3 ΤΟ ΒΕΛΤΙΣΤΟ ΜΟΝΤΕΛΟ 44

1.3.4 ΣΥΜΠΕΡΑΣΜΑΤΑ..... 48

2. ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ..... 49

2.1 ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ 49

2.1.1 ΕΙΣΑΓΩΓΗ..... 49

2.1.2 ΤΟ ΜΟΝΤΕΛΟ..... 50

2.2 ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ 52

2.2.1 ΕΙΣΑΓΩΓΗ..... 52

2.2.2 ΟΡΙΣΜΟΣ ΤΟΥ ΜΟΝΤΕΛΟΥ 52

2.2.3 ΠΑΡΑΔΕΙΓΜΑ ΜΟΝΤΕΛΟΥ ΛΟΓΙΣΤΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ 54
2.2.4 ΕΚΤΙΜΗΣΗ ΤΩΝ ΣΥΝΤΕΛΕΣΤΩΝ ΠΑΛΙΝΔΡΟΜΗΣΗΣ 55
2.2.5 ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ 57
2.2.6 ΚΡΙΤΗΡΙΑ ΕΠΙΛΟΓΗΣ ΚΑΤΑΛΛΗΛΟΥ ΜΟΝΤΕΛΟΥ 64

2.3 ΕΦΑΡΜΟΓΗ ΣΤΗ ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ ΜΕ ΧΡΗΣΗ ΤΗΣ R. 67
2.3.1 ΠΕΡΙΓΡΑΦΗ ΕΦΑΡΜΟΓΗΣ ΚΑΙ ΔΕΔΟΜΕΝΑ 67
2.3.2 ΕΚΤΕΛΕΣΗ ΕΦΑΡΜΟΓΗΣ ΚΑΙ ΕΠΙΛΟΓΗ ΒΕΛΤΙΣΤΟΥ ΜΟΝΤΕΛΟΥ 68
2.3.3 ΤΟ ΒΕΛΤΙΣΤΟ ΜΟΝΤΕΛΟ 77
2.3.4 ΣΥΜΠΕΡΑΣΜΑΤΑ 79

3. ΜΟΝΤΕΛΑ ΕΠΙΒΙΩΣΗΣ ΚΑΙ ΤΟ ΜΟΝΤΕΛΟ ΤΟΥ COX..... 80

3.1 ΜΟΝΤΕΛΑ ΕΠΙΒΙΩΣΗΣ 80
3.1.1 ΕΙΣΑΓΩΓΗ 80
3.1.2 ΒΑΣΙΚΕΣ ΕΝΟΙΕΣ 80

3.2 ΤΟ ΜΟΝΤΕΛΟ ΑΝΑΛΟΓΙΚΗΣ ΔΙΑΚΙΝΔΥΝΕΥΣΗΣ ΤΟΥ COX. 83
3.2.1 ΕΙΣΑΓΩΓΗ 83
3.2.2 ΟΡΙΣΜΟΣ ΤΟΥ ΜΟΝΤΕΛΟΥ ΤΟΥ COX 83
3.2.3 ΠΑΡΑΔΕΙΓΜΑ ΣΤΟ ΜΟΝΤΕΛΟ ΤΟΥ COX 85
3.2.4 ΕΚΤΙΜΗΣΗ ΤΩΝ ΣΥΝΤΕΛΕΣΤΩΝ ΠΑΛΙΝΔΡΟΜΗΣΗΣ 91
3.2.5 ΠΕΡΙΓΡΑΦΗ ΤΗΣ CROSS-VALIDATION (CVL) ΣΤΟ ΜΟΝΤΕΛΟ ΤΟΥ COX 93
3.2.6 ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ 95
3.2.7 ΚΡΙΤΗΡΙΑ ΕΠΙΛΟΓΗΣ ΚΑΤΑΛΛΗΛΟΥ ΜΟΝΤΕΛΟΥ 102

3.3 ΕΦΑΡΜΟΓΗ ΣΤΟ ΜΟΝΤΕΛΟ ΤΟΥ COX ΜΕ ΧΡΗΣΗ ΤΗΣ R.. 104
3.3.1 ΠΕΡΙΓΡΑΦΗ ΕΦΑΡΜΟΓΗΣ ΚΑΙ ΔΕΔΟΜΕΝΑ 104
3.3.2 ΕΚΤΕΛΕΣΗ ΕΦΑΡΜΟΓΗΣ ΚΑΙ ΕΠΙΛΟΓΗ ΒΕΛΤΙΣΤΟΥ ΜΟΝΤΕΛΟΥ 105
3.3.3 ΤΟ ΒΕΛΤΙΣΤΟ ΜΟΝΤΕΛΟ 110
3.3.4 ΣΥΜΠΕΡΑΣΜΑΤΑ 112

ΠΑΡΑΡΤΗΜΑ 113

Α. ΤΟ ΣΤΑΤΙΣΤΙΚΟ ΠΑΚΕΤΟ ΤΗΣ R 113
ΤΙ ΕΙΝΑΙ Η R; 113
ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ 113
ΤΡΟΠΟΣ ΧΡΗΣΗΣ 114

Β. ΕΦΑΡΜΟΓΗ ΓΕΝΙΚΟΥ ΓΡΑΜΜΙΚΟΥ ΜΟΝΤΕΛΟΥ ΜΕ ΧΡΗΣΗ ΤΗΣ R. 115
1. ΜΕΡΙΚΕΣ ΕΝΤΟΛΕΣ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΟΥΜΕ 115
2. ΠΡΟΣΑΡΜΟΓΗ ΜΟΝΤΕΛΟΥ 117
3. ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ 119
4. ΠΕΡΙΓΡΑΦΗ ΒΕΛΤΙΣΤΟΥ ΜΟΝΤΕΛΟΥ 122

C. ΕΦΑΡΜΟΓΗ ΣΕ ΜΟΝΤΕΛΟ ΤΗΣ ΛΟΓΙΣΤΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΣΤΗΝ R..... 123

1. ΤΟ ΠΑΚΕΤΟ GLMULTI (GLMULTI-PACKAGE)..... 123
2. ΜΕΡΙΚΕΣ ΕΝΤΟΛΕΣ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΟΥΜΕ..... 128
3. ΠΡΟΣΑΡΜΟΓΗ ΜΟΝΤΕΛΟΥ..... 129
4. ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ..... 131

D. ΕΦΑΡΜΟΓΗ ΣΕ ΜΟΝΤΕΛΟ ΤΟΥ COX ΣΤΗΝ R..... 135

1. ΤΟ ΠΑΚΕΤΟ PENALIZED (PENALIZED-PACKAGE)..... 135
2. ΜΕΡΙΚΕΣ ΕΝΤΟΛΕΣ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΟΥΜΕ..... 137
3. ΠΡΟΣΑΡΜΟΓΗ ΜΟΝΤΕΛΟΥ..... 138
4. ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ..... 139

ΒΙΒΛΙΟΓΡΑΦΙΑ 141

1. ΓΕΝΙΚΟ ΓΡΑΜΜΙΚΟ ΜΟΝΤΕΛΟ

1.1 ΕΙΣΑΓΩΓΗ

1.1.1 ΤΟ ΜΟΝΤΕΛΟ

Σε κάθε στιγμή της ζωής μας καλούμαστε να κάνουμε επιλογές και να παίρνουμε αποφάσεις, οι οποίες μας αφορούν άμεσα ή έμμεσα, π.χ. αν είμαστε στελέχη σε μια επιχείρηση ή ως μέλη μιας επιτροπής κάποιου φορέα κλπ. Για τη λήψη μιας ορθής απόφασης χρειάζεται να γνωρίζουμε τέλεια πού βρισκόμαστε σήμερα, αλλά και ποιο θα είναι το γενικό πλαίσιο που θα βρισκόμαστε αύριο. Πολύ σπάνια, αν όχι ποτέ, έχουμε αυτές τις πληροφορίες. Συνεπώς η κάθε δραστηριότητά μας βρίσκεται κάτω από κάποιες συνθήκες αβεβαιότητας. Με τον περιορισμό της αβεβαιότητας ασχολείται η επιστήμη της ***Στατιστικής***.

Κύριος σκοπός της στατιστικής είναι η συλλογή, η παρουσίαση, η ανάλυση και η ερμηνεία παρατηρήσεων που υπόκεινται σε τυχαίες μεταβλητές, ώστε να μας οδηγήσει στην εξαγωγή βάσιμων συμπερασμάτων προκειμένου να ληφθούν βέλτιστες αποφάσεις, κάτω από συνθήκες αβεβαιότητας. Αυτό επιτυγχάνεται (Κοκολάκης & Φουσκάκης, 2009):

- Με την κατάλληλη *μοντελοποίηση* του πληθυσμού, δηλαδή του χώρου που ασχολούμαστε.
- Με την ποσοτικοποίηση της αβεβαιότητας που αντιμετωπίζουμε.
- Με την ελαχιστοποίηση αυτής μέσα από μελέτη δειγμάτων του πληθυσμού.

Οπότε καταλαβαίνουμε άμεσα, ότι μεγάλο μέρος της επιστήμης της Στατιστικής ασχολείται με την κατασκευή και ανάλυση ***στατιστικών μοντέλων***. Με τον όρο **μοντέλο** εννοούμε τη μορφή της σχέσης μεταξύ δύο ή περισσοτέρων μεταβλητών.

Έχουμε το **απλό γραμμικό μοντέλο** $y = \beta_0 + \beta_1 x + \varepsilon$ με μία μόνο επεξηγηματική μεταβλητή x. Χρησιμοποιώντας περισσότερες από μία επεξηγηματικές μεταβλητές x_j,

$j = 1, \dots, k$ έχουμε το **γενικό γραμμικό μοντέλο** για το i-άτομο (Montgomery, Peck, & Vining, 2006):

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \qquad \mu\varepsilon\ i = 1,2, \dots, n \tag{1.1}$$

όπου $\boldsymbol{\beta}' = (\beta_0,\ \beta_1,\ \beta_2, \dots,\ \beta_k)$ άγνωστες παραμέτροι (ή συντελεστές), $\boldsymbol{y}$: η **εξαρτημένη** (dependent) μεταβλητή ή **μεταβλητή απόκρισης** (response variable), $\boldsymbol{x}' = (x_0,\ x_{i1},\ x_{i2}, \dots,\ x_{ik})$ ένας $n \times (k+1)$ πίνακας με $x_0 = 1$: η **ανεξάρτητη** (independent) ή **επεξηγηματική** (explanatory) ή **προβλέπουσα** (predictor) ή και **συμμεταβλητή** (covariate), ε_i: **τυχαία σφάλματα** που ακολουθούν τη $N(0, \sigma^2)$.

Επίσης το γενικό γραμμικό μοντέλο με τη μορφή πινάκων γράφεται ως:

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

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Τα τυχαία σφάλματα έχουν κεντρικό ρόλο στο γενικό γραμμικό μοντέλο και πρέπει να τηρούν τις πιο κάτω υποθέσεις:

- *$E(\varepsilon_i) = 0$, για κάθε i*
- *$V(\varepsilon_i) = \sigma^2$, για κάθε i, ομοσκεδαστικότητα*
- *$Cov(\varepsilon_i, \varepsilon_j) = 0$, για $i \neq j$, ασυσχέτιστα*
- *η τ.μ. $\boldsymbol{\varepsilon}$ ακολουθεί την πολυμεταβλητή (n-διάστατη) Κανονική κατανομή, δηλ. $\boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, με $\boldsymbol{I}$ τον n-διάστατο μοναδιαίο πίνακα.*

Η σχέση (1.1.1) καλείται **μοντέλο παλινδρόμησης (regression model)**, όπου ο όρος αυτός οφείλεται στη φύση των πρώτων εφαρμογών του μοντέλου από το δημιουργό του, F. Galton.

Μερικά παραδείγματα γενικού γραμμικού μοντέλου είναι τα πιο κάτω:

$$y = \beta_0 + \beta_1 \ln x + \varepsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

1.1.2 Η ΧΡΗΣΗ ΤΟΥ

Όμως, αν και στις συναρτησιακές σχέσεις υπάρχει επίδραση της μεταβλητής *x* επάνω στη *y*, δηλαδή η *y* εξαρτάται πράγματι από τη *x*, εντούτοις υπάρχει σημαντική διαφορά με τα στατιστικά μοντέλα. Διότι στα στατιστικά μοντέλα μπορεί να μην υπάρξει καν τέτοια εξάρτηση. Με άλλα λόγια το στατιστικό μοντέλο μπορεί να μην είναι τίποτα παραπάνω από μια εμπειρική σχέση χωρίς θεωρητική βάση. Και όμως αποτελεί ένα πάρα πολύ σημαντικό εργαλείο στη Στατιστική.

Κατά πρώτο λόγο μπορούμε να βρούμε τρόπο μελλοντικής πρόβλεψης της τιμής της εξαρτημένης μεταβλητής *y*. Επίσης χωρίς να γνωρίζουμε ποιες είναι οι σημαντικές επεξηγηματικές μεταβλητές ενός προβλήματος, με διάφορες δοκιμές μεταβλητών, αλλά και συγκρίνοντας διάφορα μοντέλα καταλήγουμε σε χρήσιμα συμπεράσματα, τουλάχιστον σε μελλοντική διερεύνηση του θέματος. Ακόμα και το ότι μια μεταβλητή δε φαίνεται να χρειάζεται στο μοντέλο, δηλ. η *y* να μη συνδέεται με τη *x*, είναι εξίσου σημαντικό με τη γνώση που θα είχα αν τελικά συσχετίζονταν.

1.1.3 ΔΙΑΔΙΚΑΣΙΑ ΣΤΑΤΙΣΤΙΚΗΣ ΜΟΝΤΕΛΟΠΟΙΗΣΗΣ

Για να προσαρμόσουμε το στατιστικό μας μοντέλο, χρησιμοποιούμε ένα σύνολο δεδομένων της εξαρτημένης μεταβλητής *y* και των επεξηγηματικών μεταβλητών $\boldsymbol{x}$ στις ίδιες στατιστικές μονάδες. Υπάρχουν διάφορες μέθοδοι προσαρμογής μοντέλου στα δεδομένα, όπου μερικές από αυτές αναφέρονται στη συνέχεια της παρούσας μελέτης. Όσο πιο επιτυχημένη είναι η προσαρμογή του μοντέλου, δηλαδή των δεδομένων που έχω, τόσο καλύτερη είναι η εκτίμηση των παραμέτρων $\boldsymbol{\beta}$. Αυτό μας βοηθά στη συνέχεια

για να καταλήξουμε σε όσο το δυνατόν ορθότερα αποτελέσματα και συμπεράσματα στο πρόβλημα που έχουμε.

Όμως σημαντική είναι και η εξέταση των αποτελεσμάτων με διάφορες τεχνικές ώστε να μπορέσουμε να επιλέξουμε τις σημαντικότερες επεξηγηματικές μεταβλητές και να καθορίσουμε έτσι ποιο θα είναι το «καταλληλότερο» μοντέλο. Ο εντοπισμός των όχι και τόσο στατιστικά σημαντικών επεξηγηματικών μεταβλητών, μας βοηθά είτε να απορρίψουμε αυτές τις μεταβλητές, είτε με κάποιο μετασχηματισμό τους, να καταλήξουμε στο βέλτιστο μοντέλο. Περεταίρω ανάπτυξη αυτών των τεχνικών επιλογής μοντέλου θα γίνει στη συνέχεια της συγκεκριμένης εργασίας. Σημειώνουμε ότι, συνήθως οι *p-τιμές* των ελέγχων υπολογίζονται από τα διάφορα στατιστικά πακέτα που χρησιμοποιούμε στον υπολογιστή (SAS, SPSS, R, MiniTab κλπ).

Υπάρχουν διάφορες μέθοδοι και τεχνικές για την επιλογή των στατιστικά σημαντικών μεταβλητών, κριτήρια επιλογής αλλά και στατιστικοί έλεγχοι υποθέσεων, ώστε να καταλήξουμε στο βέλτιστο μοντέλο για τη στατιστική μας μελέτη. Αυτό είναι και το κεντρικό θέμα της παρούσας εργασίας και στη συνέχεια θα παρουσιαστούν αναλυτικά όλες οι τεχνικές, μέθοδοι, ελέγχοι, κριτήρια κλπ.

1.2 ΒΕΛΤΙΩΣΗ ΚΑΙ ΑΝΑΠΤΥΞΗ ΜΟΝΤΕΛΟΥ

Τα τυχαία σφάλματα ε_i παίζουν πολύ σημαντικό ρόλο στο γενικό γραμμικό μοντέλο και παραβιάσεις των υποθέσεων που αναφέραμε πιο πάνω, δημιουργούν σοβαρές επιπτώσεις στην εκτίμηση του μοντέλου.

1.2.1 ΕΚΤΙΜΗΣΗ ΜΕ ΤΗ ΜΕΘΟΔΟ ΤΗΣ ΜΕΓΙΣΤΗΣ ΠΙΘΑΝΟΦΑΝΕΙΑΣ

Όταν μιλούμε για προσαρμογή του μοντέλου, δηλαδή των δεδομένων που έχουμε για κάθε μεταβλητή, εννοούμε την εκτίμηση των παραμέτρων $\boldsymbol{\beta}$. Στο γενικό γραμμικό μοντέλο μπορεί να γίνει με τη μέθοδο των ελαχίστων τετραγώνων, αλλά και με τη μέθοδο της μέγιστης πιθανοφάνειας.

Αν και στη μέθοδο των ελαχίστων τετραγώνων δεν απαιτείται γνώση της κατανομής των τυχαίων σφαλμάτων, εντούτοις, στη **μέθοδο της μέγιστης πιθανοφάνειας** (maximum likelihood method) που παρουσιάζεται πιο κάτω απαιτείται ο καθορισμός της κατανομής τους και υποθέτουμε ότι είναι η πολυμεταβλητή Κανονική κατανομή. Τα ε_i θεωρούνται ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν τη $N(0, \sigma^2)$, άρα και το διάνυσμα $\varepsilon \sim N\ (\ \mathbf{0}, \sigma^2 \boldsymbol{I}_n\ )$, με $\boldsymbol{I}_n$ το n-διάστατο μοναδιαίο πίνακα. Επομένως $y \sim N\ (\ \boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n\ )$.

Τη μεγιστοποίηση της συνάρτησης πιθανοφάνειας ως προς σ^2 και $\boldsymbol{\beta}$, την παίρνουμε από τη σχέση (Οικονόμου & Καρώνη, 2010):

$$L(\sigma^2, \boldsymbol{\beta}) = (2\pi)^{-n/2} \frac{1}{|\sigma^2 \boldsymbol{I}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta})'(\sigma^2 \boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{X\beta})\right\} \qquad (1.5)$$

Η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας υπολογίζεται και ισούται:

$$l = \ln L = -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) \qquad (1.6)$$

Τώρα, με παραγώγιση της πιο πάνω, l ως προς σ^2 και $\boldsymbol{\beta}$ θα έχουμε:

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = -\frac{1}{\sigma^2}(\boldsymbol{X'X\beta} - \boldsymbol{X'y}), \qquad \frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})$$

Θέτοντας τις παραγώγους ίσες με μηδέν και λύνοντας το σύστημα, καταλήγουμε στις **εκτιμήτριες μέγιστης πιθανοφάνειας (ε.μ.π.)** που ταυτίζονται και με τις εκτιμήτριες των ελαχίστων τετραγώνων:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$$

Επίσης έχουμε και την ε.μ.π. του σ^2 που είναι:

$$\hat{\sigma}^2 = \frac{1}{n}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right) = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2$$

Από την πιο πάνω σχέση φαίνεται ότι η $\hat{\sigma}^2$ είναι μεροληπτική εκτιμήτρια της σ^2, όμως εμείς θα χρησιμοποιούμε στη συνέχεια (π.χ. βλέπε στο R^2_{adj}) την αμερόληπτή της εκτιμήτρια: $MSE = \frac{SSE}{(n-k-1)}$.

1.2.2 ΔΙΑΔΙΚΑΣΙΕΣ ΕΠΙΛΟΓΗΣ ΜΟΝΤΕΛΟΥ ΜΕ ΒΗΜΑΤΑ

Σπουδαίο ρόλο στην επιλογή κατάλληλου μοντέλου, αλλά και τη σύγκριση μοντέλων παίζει το ***SSE*** (άθροισμα τετραγώνων των υπολοίπων). Η μεταβολή λοιπόν αυτού του *SSE* οδηγεί στον **έλεγχο-*F***, για την πρόσθεση ή την αφαίρεση q-όρων του μοντέλου (Οικονόμου & Καρώνη, 2010):

$$F = \left(\frac{(SSE_0 - SSE_1)/q}{SSE_1/(n-k-1)}\right) \sim F_{q,(n-k-1)}$$

Ο έλεγχος-F, που παρουσιάζεται στα αποτελέσματα της προσαρμογής ενός μοντέλου παλινδρόμησης, ελέγχει τις εξής υποθέσεις:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad \text{έναντι της} \quad H_1 : \text{τουλάχιστον ένα } \beta_j \neq 0$$

Αν απορριφθεί η H_0 τίθεται θέμα για το ποιες μεταβλητές έχουν $\beta_j \neq 0$, δηλ. η p-τιμή είναι πολύ μικρή, ώστε να συμπεριληφθούν στο μοντέλο.

Σε όλες τις πιο κάτω βηματικές διαδικασίες επιλογής μοντέλου μπορούν να εφαρμοστούν και άλλες παρόμοιες τεχνικές, αντί του ελέγχου-F, όπως είναι τα κριτήρια AIC και BIC, για τα οποία μιλούμε πιο κάτω (§ 1.2.3.2, 1.2.3.3), αλλά και άλλες τεχνικές συρρίκνωσης. Στην εφαρμογή, στο γενικό γραμμικό μοντέλο, που εκτελούμε στην παρούσα εργασία (§ 1.3), για να καταλήξουμε στο «βέλτιστο μοντέλο», εφαρμόζουμε διαδοχικά το κριτήριο AIC σε διάφορα μοντέλα. Με τον όρο «βέλτιστο μοντέλο» εννοούμε το μοντέλο που θα περιέχει τις στατιστικά σημαντικές μεταβλητές, με βάση τον έλεγχο ή το κριτήριο που θα χρησιμοποιήσουμε. Για παράδειγμα, στην περίπτωση που έχουμε τον έλεγχο-F, για την αφαίρεση μιας οποιασδήποτε μεταβλητής από το μοντέλο ο έλεγχος θα είναι στατιστικά σημαντικός, ενώ για την πρόσθεση μιας επιπλέον θα είναι στατιστικά μη σημαντικός. Άλλη περίπτωση που έχουμε το κριτήριο AIC ή BIC, για την αφαίρεση μιας οποιασδήποτε μεταβλητής από το μοντέλο επιλέγουμε αυτή με το ψηλότερο AIC, ενώ για την πρόσθεση μιας επιπλέον επιλέγουμε αυτή με το χαμηλότερο. Να σημειώσουμε ότι όσο πιο μικρή είναι η τιμή του κριτηρίου AIC ή BIC, τόσο πιο «καλό» θεωρείται το μοντέλο.

Οπότε έχουν αναπτυχθεί διάφορες διαδικασίες για την επιλογή, από ένα αρχικό σύνολο, των στατιστικά πιο σημαντικών μεταβλητών, που βασίζονται σε κάποιο έλεγχο ή κριτήριο για την αφαίρεση ή την πρόσθεση μεταβλητών ή και τα δύο σε ένα μοντέλο:

- **Διαδικασία της διαδοχικής αφαίρεσης ή απαλοιφής (Backward elimination)**
- **Διαδικασία της διαδοχικής πρόσθεσης ή της προς τα εμπρός επιλογής (Forward selection)**
- **Διαδικασία της κατά βήματα εμπρός-πίσω επιλογής (Stepwise selection)**

1.2.2.1 Διαδικασία της διαδοχικής αφαίρεσης (Backward elimination)

1) Εισάγουμε όλες τις διαθέσιμες μεταβλητές στο μοντέλο.
2) Αφαιρούμε τη «χειρότερη» μεταβλητή, με την έννοια ότι συμβάλλει το λιγότερο στο μοντέλο. Δηλαδή θα αφαιρεθεί η μεταβλητή της οποίας η τιμή του κριτηρίου AIC είναι η μεγαλύτερη.
3) Επαναπροσαρμόζουμε το μοντέλο στα δεδομένα, χωρίς τη μεταβλητή που αφαιρέσαμε, για να ξαναεξετάσουμε ποια θα είναι η επόμενη «χειρότερη».
4) Επαναλαμβάνουμε τα βήματα 2 και 3 μέχρι η αφαίρεση οποιασδήποτε μεταβλητής να είναι στατιστικά σημαντική, οπότε και σταματάμε.

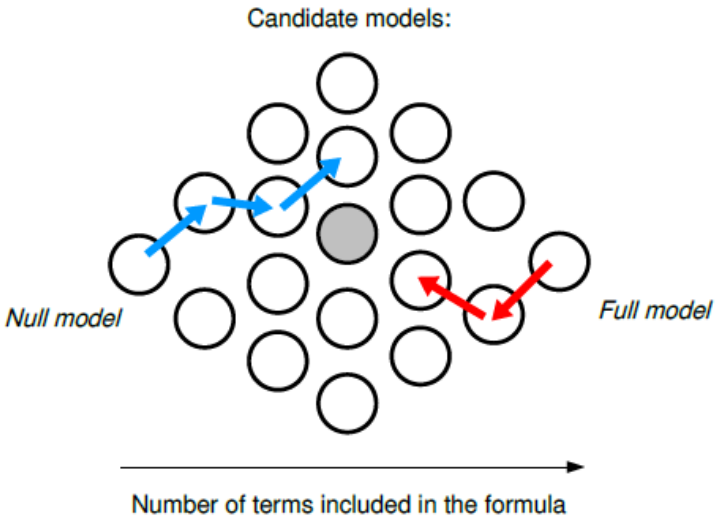
1.2.2.2 Διαδικασία της προς τα εμπρός επιλογής (Forward selection)

1) Ξεκινούμε με το σταθερό όρο β_0, δηλαδή το μοντέλο $y = \beta_0$.
2) Εισάγουμε στο μοντέλο την «καλύτερη» μεταβλητή, δηλαδή αυτήν που έχει τη μικρότερη τιμή του κριτηρίου AIC. Αυτή συμβάλλει περισσότερο στην επεξήγηση της y.
3) Επαναπροσαρμόζουμε το μοντέλο στα δεδομένα, μαζί με τη νέα μεταβλητή και ελέγχουμε ποια θα είναι η επόμενη «καλύτερη». Δηλαδή συγκρίνουμε με το προηγούμενο μοντέλο, που είχε μόνο την πρώτη μεταβλητή. Αν είναι πιο καλό αυτό το μοντέλο, εισάγουμε τη μεταβλητή αυτή.
4) Επαναλαμβάνουμε τα βήματα 2 και 3 μέχρι η πρόσθεση οποιασδήποτε από τις εναπομείναντες μεταβλητές να μην είναι στατιστικά σημαντική (μεγάλο AIC), οπότε και σταματάμε.

1.2.2.3 Διαδικασία της κατά βήματα εμπρός-πίσω επιλογής (Stepwise selection)

Η πιο συνηθισμένη και με καλύτερα αποτελέσματα διαδικασία. Έχει παρόμοια λειτουργία με τη διαδικασία διαδοχικής πρόσθεσης, με μία απλή «διόρθωση», την παρεμβολή ενός επιπλέον «προς τα πίσω ελέγχου» σε κάθε επανάληψη μίας μεταβλητής, όπως βλέπουμε και στο σχήμα 1.1.

1) Ξεκινούμε όπως τη διαδικασία της διαδοχικής πρόσθεσης και εισάγουμε στο μοντέλο την πρώτη «καλύτερη» μεταβλητή.
2) Και πάλιν προσθέτουμε την επόμενη «καλύτερη» μεταβλητή, όπως την προηγούμενη διαδικασία.
3) Εξετάζουμε τώρα αν μπορεί να αφαιρεθεί η συγκεκριμένη μεταβλητή, συγκρίνοντας τα δύο μοντέλα, (με και χωρίς αυτήν), δηλαδή την αξία παραμονής της στο μοντέλο.
4) Στα επόμενα βήματα, κάθε φορά που εισάγεται μια μεταβλητή στο μοντέλο, εξετάζουμε αν μπορεί να αφαιρεθεί κάποια από τις μεταβλητές που είχαν εισαχθεί νωρίτερα.



Σχήμα 1.1: Σχηματική απεικόνιση της πιο πάνω διαδικασίας για τα υποψήφια μοντέλα. Είναι ένα κενό (null) μοντέλο (αριστερά), ένα full μοντέλο (δεξιά) και τα υπόλοιπα μοντέλα με μερικές από τις μεταβλητές, όχι όλες(ενδιάμεσα). Τα βέλη δείχνουν τις stepwise διαδικασίες επιλογής μοντέλου: forward (αριστερά προς δεξιά, γαλάζια βέλη) και backward (δεξιά προς αριστερά, κόκκινα βέλη). Στην πιο πάνω περίπτωση και οι δύο προσεγγίσεις δεν θα συγκλίνουν προς το ίδιο μοντέλο. Μπορεί ακόμα, κανένα από αυτά να μη συγκλίνει στο πραγματικό μοντέλο που είναι το βέλτιστο (γκρίζος μεσαίος κύκλος), με βάση τους όρους του κριτηρίου που θα χρησιμοποιηθεί.

1.2.3 ΜΕΤΡΑ ΚΑΤΑΛΛΗΛΟΤΗΤΑΣ

Για την καλύτερη επιλογή ενός μοντέλου, αλλά και για τη σύγκριση διαφορετικών μοντέλων ως προς τη σημαντικότητα τους, χρήσιμα εργαλεία αποτελούν και τα **μέτρα καταλληλότητας**. Αυτά είναι σημαντικές αριθμητικές ποσότητες που βοηθούν στην καλύτερη αξιολόγηση του κάθε μοντέλου. Θα ασχοληθούμε με μερικά από αυτά, όπως είναι ο συντελεστής προσδιορισμού R^2 και τα κριτήρια AIC και BIC. Στη συνέχεια της παρούσας μελέτης (στα κεφάλαια 2 και 3), θα χρησιμοποιήσουμε τα μέτρα αυτά και ως κριτήρια επιλογής για το βέλτιστο μοντέλο στη λογιστική παλινδρόμηση αλλά και στο μοντέλο του Cox, αντίστοιχα, συγκρίνοντας τα διάφορα υποψήφια μοντέλα.

1.2.3.1 Συντελεστής προσδιορισμού R^2

Στο γενικό γραμμικό μοντέλο ο **συντελεστής προσδιορισμού** $\mathbf{R^2}$ εκφράζει το ποσοστό της μεταβλητότητας που εξηγείται από το συστηματικό μέρος του μοντέλου ($E(y) = f(x)$). Όμως επειδή ο δείκτης αυτός βασίζεται σε αθροίσματα τετραγώνων της εξαρτημένης μεταβλητής εδέχεται, η χρήση του να μην έχει τόσο πολύ νόημα στα γενικευμένα γραμμικά μοντέλα. Ωστόσο, μπορεί να επεκταθεί ο πολύ χρήσιμος αυτός δείκτης R^2 .

Στο γενικό γραμμικό μοντέλο γράφεται:

$$R^2 = 1 - \frac{SSE}{SST} \ ,$$

όπου SSE: *άθροισμα τετραγώνων των υπολοίπων ή σφαλμάτων,*
SST: *συνολικό άθροισμα τετραγώνων*

Για το συντελεστή προσδιορισμού ισχύει: $0 \leq R^2 \leq 1$. Όσο πιο κοντά στη μονάδα βρίσκεται η τιμή του συντελεστή αυτού, τόσο πιο καλά έχει προσαρμοστεί το μοντέλο. Όμως δεν ισχύει το ότι αν έχουμε χαμηλό συντελεστή είναι και χαμηλή η στατιστική σημαντικότητα του μοντέλου. Οπότε ο συντελεστής προσδιορισμού δεν έχει άμεση σχέση με τη στατιστική σημαντικότητα της παλινδρόμησης.

Ένας άλλος συντελεστής προσδιορισμού με παρόμοια χρήση είναι ο λεγόμενος **διορθωμένος συντελεστής προσδιορισμού** ($R^2 - adjusted$). Ο συγκεκριμένος έχει τη διαφορά ότι αντί των αθροισμάτων των τετραγώνων SSE και SST, ορίζεται ως:

$$R^2_{adj} = 1 - \frac{MSE}{MST} = 1 - \frac{\frac{SSE}{(n-k-1)}}{\frac{SST}{(n-1)}},$$

όπου MSE: μέσο άθροισμα τετραγώνων των υπολοίπων ή σφαλμάτων,
MST: μέσο συνολικό άθροισμα τετραγώνων

Αυτό που τον κάνει πιο χρήσιμο το διορθωμένο αυτό συντελεστή είναι η ιδιότητα του να μην αυξάνει την τιμή του όταν προστίθεται στο μοντέλο μία μεταβλητή, αλλά μόνο όταν αυτή η μεταβλητή βελτιώνει πράγματι το μοντέλο, δηλ. όταν είναι στατιστικά σημαντική.

Επομένως ο συντελεστής προσδιορισμού χρησιμοποιείται σαν κριτήριο επιλογής για το βέλτιστο μοντέλο. Συγκρίνοντας τους συντελεστές όλων των υποψήφιων μοντέλων, το μοντέλο με το μεγαλύτερο συντελεστή είναι και το βέλτιστο. Πιο έγκυρο κριτήριο μπορεί να θεωρηθεί ο διορθωμένος συντελεστής προσδιορισμού, για το λόγο ότι μόνο οι στατιστικά σημαντικές μεταβλητές που προστιθενται στο μοντέλο, τον βελτιώνουν. Αφού για όλα τα υποψήφια μοντέλα το SST παραμένει σταθερό, επομένως καλύτερο μοντέλο είναι αυτό με το μικρότερο $\frac{SSE}{(n-k-1)}$.

Επίσης, επειδή το SST προκύπτει από το γραμμικό μοντέλο που περιέχει μόνο ένα σταθερό όρο, ένα παρόμοιο κριτήριο του R^2, είναι το $ψευδό\,(pseudo) - R^2$ (McFadden, 1974) με τύπο:

$$R^2_L = 1 - \frac{l(\hat{\beta})}{\hat{l}_0},$$

όπου $l(\hat{\beta})$: η μεγιστοποιημένη λογαριθμοποιημένη πιθανοφάνεια για το μοντέλο που θέλω,
$\hat{l}_0$: η μεγιστοποιημένη λογαριθμοποιημένη πιθανοφάνεια για το μοντέλο που περιέχει μόνο το σταθερό όρο.

Η τιμή αυτού του κριτηρίου για ένα μοντέλο που δεν περιέχει επεξηγηματικές μεταβλητές και αυξάνεται με την εισαγωγή μεταβλητών, είναι το μηδέν. Χρησιμοποιείται για οποιοδήποτε μοντέλο που έχει προσαρμοστεί με τη μέθοδο της μέγιστης πιθανοφάνειας.

Περισσότερα για τους συντελεστές προσδιορισμού θα δούμε παρακάτω και κατά πόσο μπορούν να βοηθήσουν, σαν κριτήρια επιλογής μοντέλου στα γενικευμένα γραμμικά μοντέλα, ιδιαίτερα στη λογιστική παλινδρόμηση, αλλά και κατά πόσο μας χρησιμεύουν στα μοντέλα διάρκειας ζωής.

1.2.3.2 Κριτήριο AIC (Akaike 's information criterion)

Το AIC αποτελεί ένα κριτήριο επιλογής του βέλτιστου μοντέλου με όσο το δυνατόν μικρότερο αριθμό παραμέτρων. Ορίζεται από τη σχέση (Akaike, 1974):

$$AIC = 2d - 2\ln L \tag{1.7}$$

όπου d: *πλήθος παραμέτρων του μοντέλου,*
L: *η μεγιστοποιημένη τιμή της συνάρτησης πιθανοφάνειας για το εκτιμημένο μοντέλο*

Συγκρίνοντας όλα τα υποψήφια μοντέλα, το μοντέλο με το μικρότερο AIC είναι το προτιμότερο, το βέλτιστο. Πάντοτε η εισαγωγή περισσότερων παραμέτρων στο μοντέλο, βελτιώνει την προσαρμογή του, ανεξάρτητα αν είναι στατιστικά σημαντικές ή όχι. Ωστόσο, το AIC δεν ανταμοίβει μόνο την απλή προσαρμογή, αλλά περιλαμβάνει επίσης μία ποινή (penalty) που είναι μια αύξουσα συνάρτηση του αριθμού των εκτιμώμενων παραμέτρων. Αυτή η ποινή, συγκεκριμένα η παράμετρος d της σχέσης 1.7, αποθαρρύνει το overfitting, δηλαδή η αύξηση του αριθμού των ελεύθερων παραμέτρων στο μοντέλο βελτιώνει την καλή προσαρμογή, ανεξάρτητα από τον αριθμό των ελεύθερων παραμέτρων στη διαδικασία προσαρμογής των δεδομένων.

Στο γενικό γραμμικό μοντέλο το κριτήριο παίρνει την πιο κάτω μορφή:

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

όπου SSE*: το άθροισμα των τετραγώνων των υπολοίπων,* $p = k + 1, k$*: επεξηγηματικές μεταβλητές*

Επίσης, έχει αναπτυχθεί μια τροποποίηση του AIC, που χρησιμοποιείται κυρίως για μικρά δείγματα, το επονομαζόμενο διορθωμένο AIC_c και ορίζεται ως:

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

1.2.3.3 Κριτήριο BIC (Bayesian information criterion)

Το **BIC**, που προτάθηκε από τον Schwarz (1978), αποτελεί ένα ακόμη κριτήριο επιλογής του βέλτιστου μοντέλου ανάμεσα σε μοντέλα με διαφορετικό αριθμό παραμέτρων, όπως και το AIC. Αν και η αφετηρία του είναι διαφορετική από του AIC, η λογική και η χρήση του είναι η ίδια. Η βασική διαφορά τους είναι ότι η εισαγωγή επιπρόσθετων παραμέτρων αποθαρρύνεται σε μεγαλύτερο βαθμό από το AIC. Στη γενική περίπτωση ορίζεται από τη σχέση (Buckland, Burnham, & Augustin, 1997):

$$BIC = d\ln n - 2\ln L \qquad (1.8)$$

όπου d*: πλήθος παραμέτρων του μοντέλου,* L*: η μεγιστοποιημένη τιμή της συνάρτησης πιθανοφάνειας*

και συγκεκριμένα στο γενικό γραμμικό μοντέλο γράφεται ως:

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

Σαν κριτήριο επιλογής μοντέλου, όπως και για το AIC, συγκρίνοντας όλα τα υποψήφια μοντέλα, το μοντέλο με το χαμηλότερο BIC είναι και το βέλτιστο. Επίσης, ισχύουν τα ίδια για την παράμετρο d, που λειτουργεί σαν ποινή για αποθάρρυνση του overfitting.

1.2.4 ΜΕΘΟΔΟΙ L_1 ΚΑΙ L_2-PENALIZED

1.2.4.1 Το φαινόμενο της Πολυσυγγραμμικότητας

Ένας άλλος λόγος που μπορεί να δημιουργήσει προβλήματα στην σωστή εκτίμηση του μοντέλου, είναι **το φαινόμενο της πολυσυγγραμμικότητας** (multicollinearity). Συχνά, στο γενικό γραμμικό μοντέλο είναι πιθανόν, μία ή περισσότερες ανεξάρτητες μεταβλητές X_j, να είναι γραμμικά εξαρτημένες ή διαφορετικά, όταν υπάρχει έντονη συσχέτιση μεταξύ δύο ή περισσότερων επεξηγηματικών μεταβλητών. Η παρουσία του φαινομένου αυτού οδηγεί σε αυξημένα τυπικά σφάλματα των $\hat{\beta}$ $(se(\hat{\beta}))$ και κατά συνέπεια δυσκολεύει την επίδραση της εκτίμησης κάθε επεξηγηματικής μεταβλητής στην εξαρτημένη μεταβλητή. Αυτό μας μπερδεύει κάπως γιατί είναι πιο δύσκολος ο εντοπισμός των στατιστικά σημαντικών μεταβλητών. Αυτό το φαινόμενο παρατηρείται συχνά σε βιολογικές, χημικές, οικονομικές και γενικά σε όλους τους τομείς που υπάρχουν στατιστικές μελέτες, και οδηγεί έτσι σε λανθασμένη εξαγωγή εκτιμήσεων αλλά και αποτελεσμάτων.

Σε τέτοιες περιπτώσεις, όπου η πολυσυγγραμμικότητα οφείλεται στην έντονη συσχέτιση των μεταβλητών, η ανάλυση παλινδρόμησης μπορεί να πραγματοποιηθεί αφού αφαιρεθεί μια μεταβλητή από το γραμμικά εξαρτημένο σύνολο. Επίσης, συχνά παρουσιάζονται και περιπτώσεις έντονης συσχέτισης μεταξύ των μεταβλητών, χωρίς απόλυτα να είναι γραμμικά εξαρτημένες. Έτσι σιγά σιγά πρέπει να αφαιρεθούν οι στατιστικά μη σημαντικές μεταβλητές από το μοντέλο ή να γίνει επιλογή του καλύτερου συνδιασμού μεταβλητών ώστε να καταλήξουμε στο καταλληλότερο μοντέλο. Με λίγα λόγια αρχίζει η **συρρίκνωση** (shrinkage) του μοντέλου.

Πολλές τεχνικές έχουν προταθεί για την αντιμετώπιση των προβλημάτων λόγω πολυσυγγραμμικότητας. Οι γενικότερες προσεγγίσεις συνιστούν την συλλογή περαιτέρω δεδομένων, τον επαναπροσδιορισμό του μοντέλου καθώς και την εκτίμηση διαφορετικών εκτιμητικών μεθόδων, πέραν των ελαχίστων τετραγώνων. Μία από τις πρώτες τεχνικές που πρότειναν λύσεις στο πρόβλημα της πολυσυγγραμμικότητας είναι η παλινδρόμηση κορυφογραμμής (Ridge regression) (L_2-penalized) (Hoerl & Kennard, 1970). Την τελευταία δεκαετία έχουν προταθεί αρκετές εναλλακτικές τεχνικές που μειώνουν τον

αριθμό των παραμέτρων στο τελικό μοντέλο, ειδικότερα η LASSO (Tibshirani, 1996) που επιβάλλει μια L_1-penalized στους συντελεστές παλινδρόμησης. Χρησιμοποιώντας μια μη-κυρτή ποινή, κάνει αυτόματα επιλογή μεταβλητών σε αντίθεση με την παλινδρόμηση κορυφογραμμής που μόνο μειώνει τους εκτιμητές κοντά στο μηδέν. Επίσης, ο Tibshirani χρησιμοποίησε τη μέθοδο αυτή λίγο παραλλαγμένη και στα μοντέλα επιβίωσης, συγκεκριμένα για τους ελέγχους στο μοντέλο του Cox (Tibshirani, 1997). Πιο πρόσφατα προτάθηκε και η μέθοδος Elastic net, E-net (Zou & Hastie, 2005), ως μια εναλλακτική διαδικασία που αντιμετωπίζει τις ελλείψεις της Lasso και της παλινδρόμησης κορυφογραμμής, συνδυάζοντας τις τεχνικές L_1 και L_2 με ποινή. Ένα κίνητρο των Zou και Hastie ήταν ότι η μέθοδός τους έχει την ιδιότητα να περιλαμβάνει στο τελικό μοντέλο τις ομάδες των μεταβλητών που είναι ισχυρά συσχετισμένες. Όταν οι μεταβλητές είναι ισχυρά συσχετισμένες, η Lasso επιλέγει μόνο μία μεταβλητή από το γκρουπ, ενώ η E-net επιλέγει όλη την ομάδα.

1.2.4.2 Μέθοδος L_2 - Penalized (Ridge regression)

Η πρώτη τεχνική συρρίκνωσης που εφαρμόστηκε, όπως έχουμε σημειώσει και πριν, είναι η **Παλινδρόμηση Κορυφογραμμής (Ridge regression)** (Hoerl & Kennard, 1970). Καλείται έτσι, καθώς τα μαθηματικά που χρησιμοποιούνται σχετίζονται με τη μέθοδο της Ridge-ανάλυσης η οποία είχε προηγουμένως χρησιμοποιηθεί από τον Hoerl για την περιγραφή της συμπεριφοράς των δευτεροβάθμιων επιφανειών (Montgomery, Peck, & Vining, 2006).

Όταν τα δεδομένα μας χαρακτηρίζονται από πολυσυγγραμμικότητα, η κλασσική μέθοδος των ελαχίστων τετραγώνων αποτυγχάνει να εκτιμήσει σε ικανοποιητικό βαθμό τους συντελεστές παλινδρόμησης των μεταβλητών μας, με αποτέλεσμα οι εκτιμήσεις να μην είναι αξιόπιστες. Αυτό οφείλεται στο ότι η $\widehat{\boldsymbol{\beta}}$ αντί να είναι αμερόληπτη, είναι μεροληπτική εκτιμήτρια της $\boldsymbol{\beta}$ εκτιμήτριας των συντελεστών παλινδρόμησης, δηλαδή $E(\widehat{\boldsymbol{\beta}}) \neq \boldsymbol{\beta}$.

Πιο κάτω παρουσιάζουμε το «μηχανισμό» της παλινδρόμησης κορυφογραμμής, ξεκινώντας από την εξίσωση παλινδρόμησης του γενικού γραμμικού μοντέλου με k-επεξηγηματικές μεταβλητές:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \, .$$

Η **εκτιμήτρια κορυφογραμμής** (ridge estimator) $\widehat{\boldsymbol{\beta}}_{ridge} = (\hat{\beta}_{ridge1}, \dots, \hat{\beta}_{ridgek})$ είναι η εξής (Hoerl & Kennard, 1970):

$$\widehat{\boldsymbol{\beta}}_{ridge} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{1.9}$$

Ο αριθμός $\lambda \geq 0$ είναι μία σταθερά, η οποία επιλέγεται κάθε φορά και ονομάζεται **παράμετρος μεροληψίας** (biasing parameter). Όταν $\lambda = 0$ η εκτιμήτρια ταυτίζεται με αυτήν των ελαχίστων τετραγώνων. Η εκτιμήτρια $\widehat{\boldsymbol{\beta}}_{ridge}$, επειδή είναι μεροληπτική εκτιμήτρια του $\widehat{\boldsymbol{\beta}}$, δηλ. $E\left(\widehat{\boldsymbol{\beta}}_{ridge}\right) = E(\boldsymbol{z}_k\boldsymbol{\beta}) = \boldsymbol{z}_k\boldsymbol{\beta}$, αποδεικνύεται εύκολα ότι είναι ένας γραμμικός συνδυασμός της εκτιμήτριας των ελαχίστων τετραγώνων, αφού:

$$\widehat{\boldsymbol{\beta}}_{ridge} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{z}_k\boldsymbol{\beta}$$

με πίνακα συνδιασποράς:

$$Var(\widehat{\boldsymbol{\beta}}_{ridge}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}$$

και μέσο τετραγωνικό σφάλμα:

$$\begin{aligned} MSE\left(\widehat{\boldsymbol{\beta}}_{ridge}\right) &= Var\left(\widehat{\boldsymbol{\beta}}_{ridge}\right) + [bias(\widehat{\boldsymbol{\beta}}_{ridge})]^2 \\ &= \sigma^2 Tr[(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}] + \lambda^2\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{\beta} \\ &= \sigma^2 \sum_{j=1}^{k} \frac{\lambda_j}{(\lambda_j + \mu)^2} + \lambda^2\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-2}\boldsymbol{\beta} \end{aligned}$$

με $\lambda_1, \lambda_2, \dots, \lambda_k$ οι ιδιοτιμές του πίνακα $\boldsymbol{X}'\boldsymbol{X}$. Σύμφωνα με την παραπάνω ισότητα, αύξηση του μ προκαλεί αύξηση του δευτέρου όρου της μεροληψίας, ενώ η ταυτόχρονη αύξηση του λ προκαλεί μείωση του πρώτου όρου, δηλ. της διασποράς.

Σκοπός της μεθόδου της κορυφογραμμής είναι η επιλογή ενός τέτοιου λ ώστε η μείωση στον όρο της διασποράς να είναι μεγαλύτερη από την αύξηση στον όρο που εκφράζει τη μεροληψία. Αυτό επιτυγχάνεται αν για μία μη μηδενική τιμή για το λ ισχύει ότι:

$$MSE(\widehat{\boldsymbol{\beta}}_{ridge}) < Var(\widehat{\boldsymbol{\beta}})$$

Πράγματι αυτό αποδείχθηκε από τους Hoerl και Kennard (1970) με την προϋπόθεση ότι το $\boldsymbol{\beta}'\boldsymbol{\beta}$ είναι φραγμένο. Επίσης από το άθροισμα των τετραγώνων των υπολοίπων βλέπουμε ότι αύξηση του λ επιφέρει και αύξηση του *SSE*:

$$SSE = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{ridge})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{ridge}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + (\widehat{\boldsymbol{\beta}}_{ridge} - \widehat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_{ridge} - \widehat{\boldsymbol{\beta}})$$

Τώρα, λόγω του ότι το *SST* ξέρουμε ότι παραμένει σταθερό, το άθροισμα των τετραγώνων λόγω παλινδρόμησης θα μειώνεται. Συνεπώς, όσο το λ αυξάνει τόσο θα μειώνεται ο συντελεστής προσδιορισμού R^2. Αυτό δείχνει ότι με την εκτιμήτρια κορυφογραμμής μπορεί να μην πάρουμε την καλύτερη προσαρμογή για τα δεδομένα μας, αλλά θα πάρουμε σίγουρα ένα καλό σύνολο για τις εκτιμήσεις μας και τα αποτελέσματα που θα καταλήξουμε θα είναι αρκετά έγκυρα.

1.2.4.3 Μέθοδος L_1 - Penalized (LASSO)

Άλλη μια σημαντική τεχνική που περιορίζει το πρόβλημα της πολυσυγγραμμικότητας για να έχουμε καλύτερα αποτελέσματα, είναι η **LASSO** (Least Absolute Shrinkage and Selection Operator) η οποία σχεδιάστηκε το 1996 από τον Tibshirani και είναι από τις κορυφαίες μεθόδους σήμερα. Ξεκίνησε με εφαρμογές στα γενικά γραμμικά μοντέλα, αλλά και στα γενικευμένα γραμμικά μοντέλα, όμως πλέον εφαρμόζεται και στα μοντέλα επιβίωσης, όπως το μοντέλο του Cox, της Poisson κλπ.

Η μέθοδος αυτή είναι από τις καλύτερες τεχνικές επιλογής κατάλληλου μοντέλου. Σκοπός της είναι να συρρικνώσει (shrinks) κάποιους συντελεστές και να θέτει τους υπόλοιπους σε μηδέν και ως εκ τούτου να προσπαθεί να κρατήσει τα καλά χαρακτηριστικά της επιλογής υποσυνόλου μεταβλητών, αλλά και της παλινδρόμησης κορυφογραμμής. Θεωρείται ελκυστική ως μέθοδος, διότι έχει την ιδιότητα να εκτελεί

ταυτόχρονα επιλογή μεταβλητών και συρρίκνωση του μοντέλου, εξ' ου και η ονομασία της, γεγονός που την καθιστά πολύ χρήσιμη για την εξεύρεση ερμηνεύσιμων κανόνων πρόβλεψης για μεγάλων διαστάσεων δεδομένα.

Η Lasso είναι μια προσέγγιση της κανονικοποιημένης εκτίμησης για τα μοντέλα παλινδρόμησης που περιορίζουν την L_1-νόρμα των συντελεστών παλινδρόμησης. Με $\widehat{\boldsymbol{\beta}}' = (\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2, \dots ,\ \hat{\beta}_p)$, οι lasso εκτιμήτριες $(\hat{\alpha}, \widehat{\boldsymbol{\beta}})$ ορίζονται ως (Tibshirani, 1996):

$$(\hat{\alpha}, \widehat{\boldsymbol{\beta}}) = argmin\left\{\sum_{i=1}^{N}\left(y_i - \alpha - \sum_j \beta_j x_{ij}\right)^2\right\}\ \ subject\ to\ \ \| \boldsymbol{\beta} \|_1 = \sum |\beta_j| \le s \quad (1.10)$$

όπου $s \ge 0$ είναι η παράμετρος ρύθμισης. Για κάθε s, η εκτιμήτρια του α είναι $\hat{\alpha} = \bar{y}$. Μπορούμε να υποθέσουμε, χωρίς βλάβη της γενικότητας, ότι $\bar{y} = 0$ και ως εκ τούτου να παραλείψουμε το α. Επίσης, η παράμετρος s ελέγχει την ποσότητα της συρρίκνωσης που έχει εφαρμοστεί με τις εκτιμήτριες. Έστω $\hat{\beta}_j^0$ οι πλήρεις εκτιμήτριες των ελαχίστων τετραγώνων και $s_0 = \sum |\hat{\beta}_j^0|$. Οι τιμές για $s < s_0$ θα προκαλέσουν συρρίκνωση των λύσεων προς το μηδέν και οι τιμές κάποιων συντελεστών μπορεί να γίνουν ακριβώς ίσες με μηδέν.

Επιπλέον πληροφορίες για τη φύση της συρρίκνωσης μπορούμε να αντλήσουμε και από την ορθοκανονική περίπτωση σχεδιασμού. Έστω $\boldsymbol{X}$ $n \times p$ - πίνακας, με $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$ και x_{ij} τα στοιχεία του, με $\frac{\sum_i x_{ij}}{N} = 0$ και $\frac{\sum_i {x_{ij}}^2}{N} = 1$. Οι λύσεις της εξίσωσης (1.3) φαίνονται εύκολα ότι είναι:

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

όπου το γ καθόριζεται από την κατάσταση $s = \sum |\hat{\beta}_j|$. Στην ορθοκανονική αυτή περίπτωση σχεδιασμού για την επιλογή καλύτερου υποσυνόλου μεταβλητών, μειώνει το μέγεθος k, επιλέγοντας τους μεγαλύτερους κατά απόλυτη τιμή συντελεστές και θέτει τους υπόλοιπους ίσους με μηδέν. Για κάποια επιλογή του λ, αυτό είναι ισοδύναμο με τον καθορισμό του $\hat{\beta}_j = \hat{\beta}_j^0$ εάν $\hat{\beta}_j^0 > \lambda$ και 0 διαφορετικά.

Η ridge regression ελαχιστοποιεί:

$$\sum_{i=1}^{N}\left(y_i-\sum_j \beta_j x_{ij}\right)^2+\lambda\sum\beta_j^2$$

ή ισοδύναμα, ελαχιστοποιεί:

$$\sum_{i=1}^{N}\left(y_i-\sum_j \beta_j x_{ij}\right)^2 \quad subject\ to \quad \sum\beta_j^2 \le s$$

Με τη λύση της παλινδρόμησης κορυφογραμμής να ισούται με $\frac{1}{1+\gamma}\hat{\beta}_j^0$, όπου το γ να εξαρτάται από το λ ή το s.

Όμως, δεδομένου ότι η εκτιμήτρια lasso είναι μη γραμμική και μη διαφορίσιμη συνάρτηση, ακόμα και για μια σταθερή τιμή του s, είναι δύσκολος ο υπολογισμός μιας ακριβής τιμής του τυποποιημένου σφάλματός της (standard error). Ο καθορισμός του s είναι ανάλογος με την επιλογή του κατάλληλου υποσυνόλου και μετά με χρήση του σφάλματος των ελαχίστων τετραγώνων για το συγκεκριμένο υποσύνολο.

Τώρα, με $l(\boldsymbol{\beta}) = Log(l(\boldsymbol{\beta}))$ το λογάριθμο της πιθανοφάνειας του μοντέλου, το κριτήριο μπορεί να πάρει την πιο κάτω μορφή (Tibshirani, 1997):

$$\widehat{\boldsymbol{\beta}} = argmin\ l(\boldsymbol{\beta})\ \ subject\ to\ \ \|\boldsymbol{\beta}\|_1 = \sum|\beta_j| \le s \qquad (1.11)$$

Στο γενικό γραμμικό μοντέλο, χρησιμοποιώντας τη λογαριθμοποιημένη συνάρτηση πιθανοφάνειας (εξίσωση 1.6) ο πιο πάνω τύπος θα γίνει:

$$\widehat{\boldsymbol{\beta}} = argmin\left\{-\frac{n}{2}ln(2\pi)-\frac{n}{2}ln(\sigma^2)-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})\right\}\ \ subject\ to\ \ \|\boldsymbol{\beta}\|_1 = \sum|\beta_j| \le s$$

Έχει δωθεί ένας επαναληπτικός αλγόριθμος για τον υπολογισμό των Lasso-εκτιμητριών, με βάση τετραγωνικές τεχνικές προγραμματισμού (quadratic programming techniques), διότι περιλαμβάνονται συνεχόμενες λύσεις των προβλημάτων ελαχίστων τετραγώνων.

Βασικά, ο αλγόριθμος χρειάζεται ένα αριθμό επαναλήψεων μεταξύ p και $2p$, όπου p ο αριθμός των μεταβλητών, με τη βοήθεια της εξελιγμένης Newton Raphson μεθόδου, ως επαναληπτική επανασταθμισμένων (reweighted) ελαχίστων τετραγώνων (IRLS).

Με $\boldsymbol{X}$ τον πίνακα των μεταβλητών παλινδρόμησης, $\boldsymbol{\eta} = \boldsymbol{X\beta}$, που καθορίζει το $\boldsymbol{u} = \frac{\partial l}{\partial \boldsymbol{\eta}}$, $\boldsymbol{A} = -\frac{\partial^2 l}{\partial \boldsymbol{\eta}\boldsymbol{\eta}^T}$ και $\boldsymbol{z} = \boldsymbol{\eta} + \boldsymbol{A}^{-1}\boldsymbol{u}$ (Hastie & Tibshirani, 1990), τότε ένα ανάπτυγμα Taylor ενός όρου για την $l(\boldsymbol{\beta})$, θα έχει τη μορφή:$(\boldsymbol{z} - \boldsymbol{\eta})^T \boldsymbol{A} (\boldsymbol{z} - \boldsymbol{\eta})$. Ως εκ τούτου, για να λύσουμε το πρόβλημα, ακολουθούμε την πιο κάτω διαδικασία (Tibshirani, 1996):

1. Κρατούμε το s σταθερό και βάζουμε αρχική τιμή $\widehat{\boldsymbol{\beta}} = 0$.
2. Υπολογίζουμε τα **η**, **u**, **A**, **z** με βάση την τρέχουσα τιμή του $\widehat{\boldsymbol{\beta}}$.
3. Ελαχιστοποιούμε την παράσταση: $(\boldsymbol{z} - \boldsymbol{X\beta})^T \boldsymbol{A} (\boldsymbol{z} - \boldsymbol{X\beta})$ $subject\ to$ $\sum |\beta_j| \leq s$.
4. Επαναλαμβάνουμε τα βήματα 2 και 3 μέχρι το $\widehat{\boldsymbol{\beta}}$ να σταματήσει να αλλάζει (να έχει παραμείνει το ίδιο).

Η ελαχιστοποίηση στο βήμα 3 γίνεται, όπως είπαμε και πριν, με μια τετραγωνική διαδικασία προγραμματισμού, όπως περιγράφει ο Tibshirani (1996). Όμως, πρέπει να αναφέρουμε πως, αν κάποιος χρησιμοποιήσει τη χωρίς περιορισμούς ελαχιστοποίηση στο βήμα 3, η διαδικασία αυτή θα είναι ισοδύναμη με το συνηθισμένο αλγόριθμο Newton Raphson για τη μεγιστοποίηση της μερικής πιθανοφάνειας (Hastie & Tibshirani, 1990).

1.2.4.4 Cross Validation (CVL)

Καθοριστικό ρόλο για την ορθή εκτέλεση των τεχνικών με ποινή παίζει η μέθοδος **cross validation**, γιατί βρίσκει τις βέλτιστες τιμές για τα λ, ώστε να μπορέσουμε να καθορίσουμε το βέλτιστο μοντέλο. Η ***cvl*** συνάρτηση που χρησιμοποιείται στην R, υπολογίζει τη *cross validated συνάρτηση πιθανοφάνειας* για σταθερές τιμές των λ, που αν τη βελτιστοποιήσουμε καταλήγουμε στις βέλτιστες τιμές των λ.

Υποθέτουμε ότι έχουμε n-παρατηρήσεις και ένα μοντέλο παλινδρόμησης που περιγράφει τα δεδομένα. Έχουμε $l(\boldsymbol{\beta})$ τη συνάρτηση πιθανοφάνειας, με $\boldsymbol{\beta}$ τους συντελεστές

παλινδρόμησης και ορίζουμε για την i-παρατήρηση (Verweij & van Houwelingen, 1993):

$$l_i(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - l_{(-i)}(\boldsymbol{\beta})$$

όπου $l_{(-i)}(\boldsymbol{\beta})$ η συνάρτηση πιθανοφάνειας με την i-παρατήρηση να έχει φύγει και η τιμή του $\boldsymbol{\beta}$ που μεγιστοποιεί την $l_{(-i)}(\boldsymbol{\beta})$ να αντικαθίσταται από την $\widehat{\boldsymbol{\beta}}_{(-i)}$.

Εάν οι συμμεταβλητές είναι ανεξάρτητες, όπως έχουμε στα γραμμικά μοντέλα αλλά και στο μοντέλο της λογιστικής παλινδρόμησης, εύκολα υπολογίζεται ότι ισχύει $\sum_{i=1}^{n} l_i(\boldsymbol{\beta}) = l(\boldsymbol{\beta})$. Έτσι καθορίζουμε την cross validated συνάρτηση πιθανοφάνειας ως:

$$cvl = \sum_{i=1}^{n} l_i(\widehat{\boldsymbol{\beta}}_{(-i)})$$

Για ένα συγκεκριμένο μοντέλο η cvl μετρά πόσο καλά κάθε i-παρατήρηση μπορεί να προβλεφθεί με τη βοήθεια των άλλων παρατηρήσεων και επομένως χρησιμεύει ως μέτρο της τιμής που θα προβλεφθεί.

Στην R, χρησιμοποιώντας τις κατάλληλες εντολές (όπως θα δούμε στη συνέχεια στην εφαρμογή) βελτιστοποιώντας τη cvl-συνάρτηση, υπολογίζονται αριθμητικά, αλλά και σχηματικά οι βέλτιστες τιμές για τα λ. Αυτές θα χρησιμοποιηθούν στη συνέχεια για την προσαρμογή και εκτέλεση στις τεχνικές με ποινή, με τη βοήθεια και του πακέτου-Penalized (Παράρτημα D.1), για την εύρεση του βέλτιστου μοντέλου.

1.3 ΕΦΑΡΜΟΓΗ ΣΤΟ ΓΕΝΙΚΟ ΓΡΑΜΜΙΚΟ ΜΟΝΤΕΛΟ ΜΕ ΧΡΗΣΗ ΤΗΣ R

1.3.1 ΠΕΡΙΓΡΑΦΗ ΕΦΑΡΜΟΓΗΣ ΚΑΙ ΔΕΔΟΜΕΝΑ

Ξεκινούμε με μια απλή εφαρμογή όπου θα γίνει προσαρμογή δεδομένων, ανάλυση παλινδρόμησης αλλά και επιλογή του κατάλληλου μοντέλου, στο γενικό γραμμικό μοντέλο με τη βοήθεια του στατιστικού πακέτου της R. Το πρόβλημα αυτό ασχολείται με *τις τιμές πώλησης κάποιων κατοικιών* (*sale price of the home*) σε μια περιοχή στην Αμερική. Αυτή είναι και η μεταβλητή απόκρισης *y* του προβλήματος. Έχουμε επίσης, κάποιες μεταβλητές που περιγράφουν το πρόβλημα όπως φορολογία, αριθμός υπνοδωματίων, ηλικία σπιτιού κλπ. Η περιγραφή των μεταβλητών φαίνεται στον πιο κάτω πίνακα (Πίνακας 1.1). Επεξεργαζόμαστε τα δεδομένα από συνολικά 28 σπίτια (Narula & Wellington, 1977).

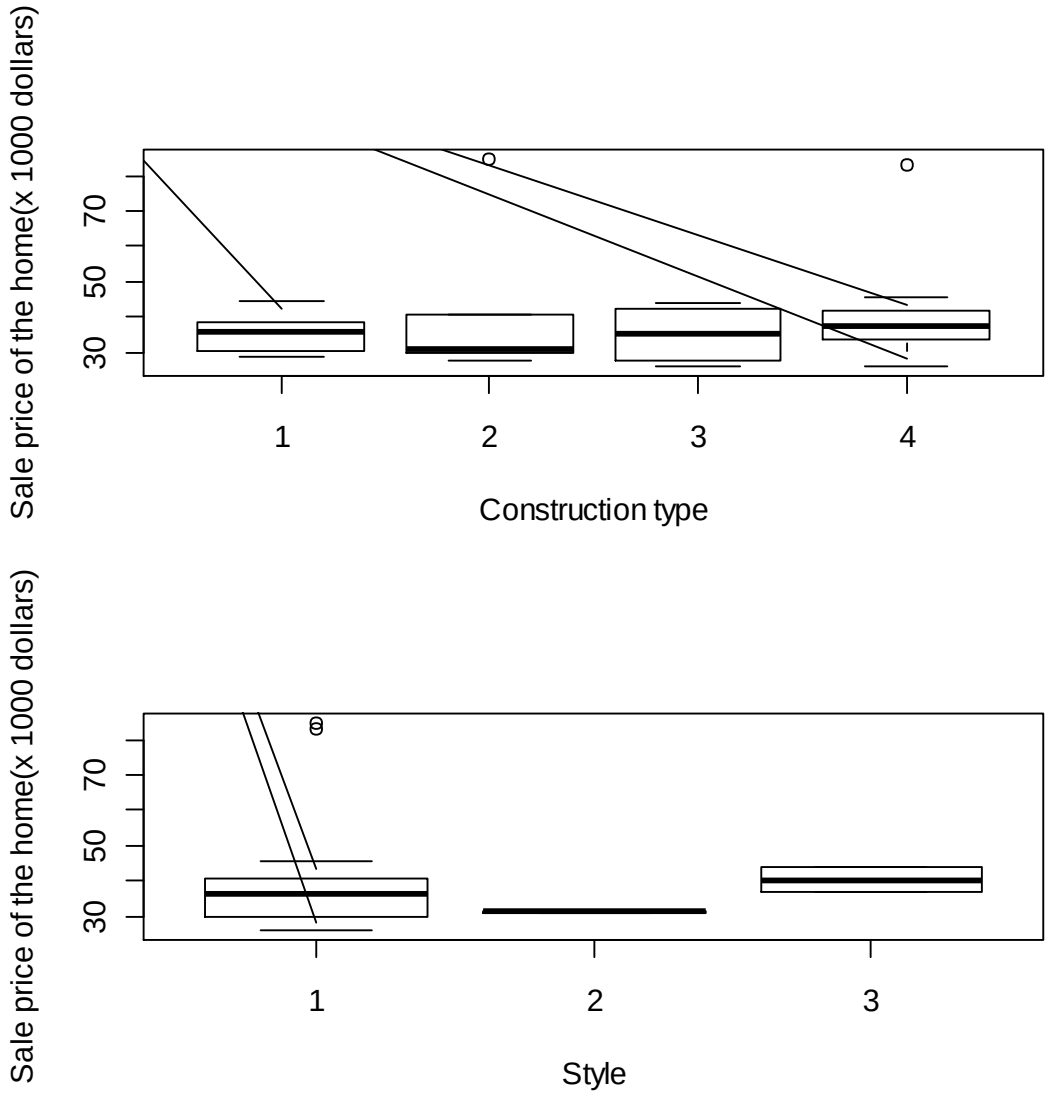
Σκοπός μας είναι, χρησιμοποιώντας την R, να προσαρμόσουμε στα δεδομένα το γενικό γραμμικό μοντέλο και να επιλέξουμε με διάφορες τεχνικές, μέθοδους αλλά λαμβάνοντας υπόψη και κάποια κριτήρια, για τα οποία έχουμε μιλήσει πριν στη θεωρία, το καταλληλότερο μοντέλο για την περιγραφή του προβλήματος.

Μεταβλητές	**Περιγραφή**
y (price)	Τιμή πώλησης του σπιτιού (x1000 δολλάρια)
x1 (taxes)	Φόροι (γενικοί, σχολικοί, της πόλης) (x100 δολλάρια)
x2 (baths)	Αριθμός λουτρών
x3 (lotSize)	Έκταση οικοπέδου (x1000 τετραγωνικά πόδια)
x4 (livSpace)	Χώρος κατοικίσιμος (x1000 τετραγωνικά πόδια)
x5 (garages)	Αριθμός γκαράζ
x6 (rooms)	Αριθμός δωματίων
x7 (bedrooms)	Αριθμός υπνοδωματίων
x8 (age)	Ηλικία του σπιτιού (σε χρόνια)
x9 (constr)	Τύπος κατασκευής (τούβλο (1), τούβλο και σκελετός (2), αλουμίνιο και σκελετός (3), σκελετός (4))
x10 (style)	Στυλ (two story (1), one and a half story (2), ranch (3))
x11 (fireplaces)	Αριθμός τζακιών

Πίνακας 1.1

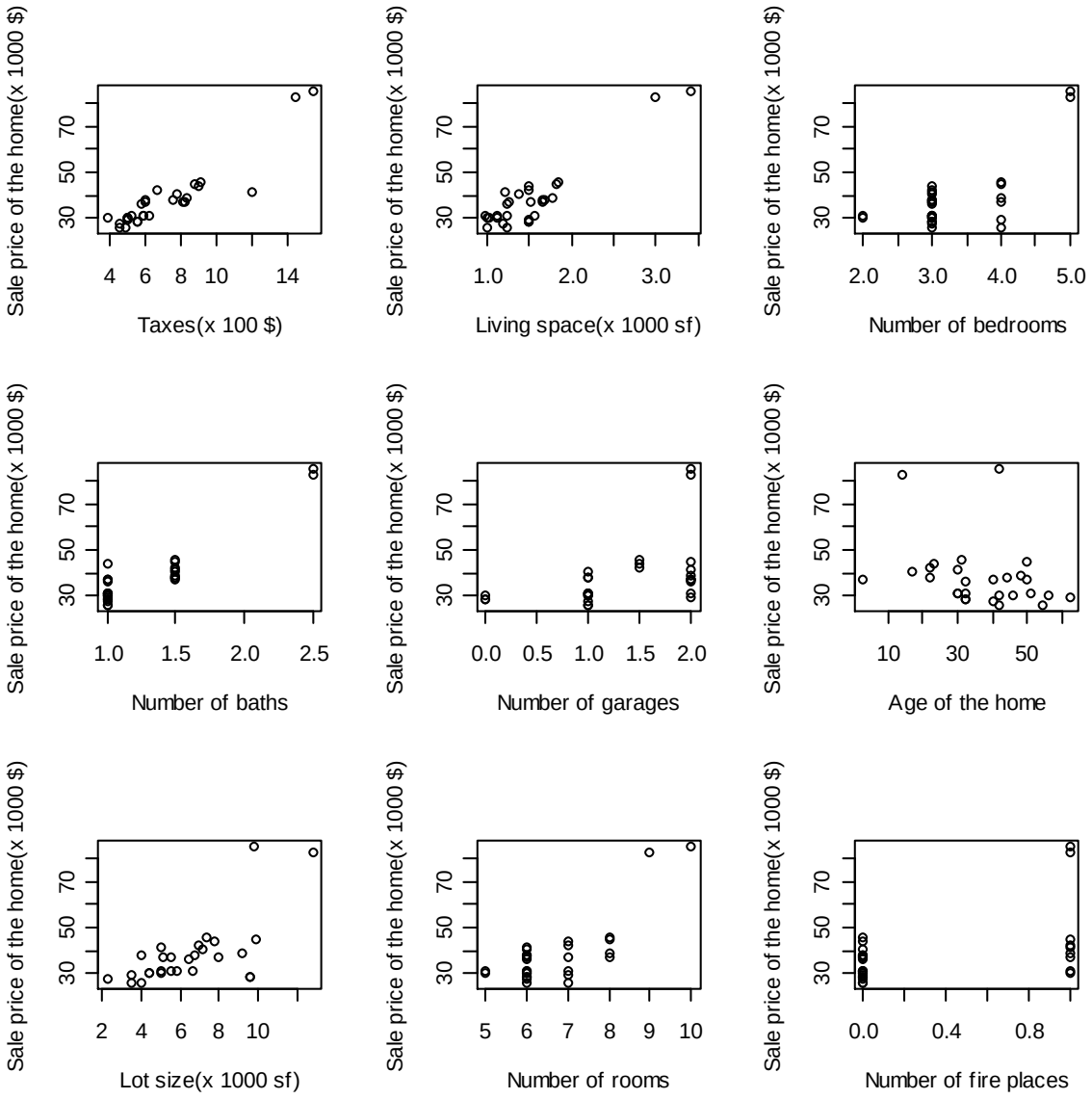
1.3.2 ΕΚΤΕΛΕΣΗ ΕΦΑΡΜΟΓΗΣ ΚΑΙ ΕΠΙΛΟΓΗ ΒΕΛΤΙΣΤΟΥ ΜΟΝΤΕΛΟΥ

Κατ' αρχάς, ορίζουμε τις επεξηγηματικές μεταβλητές (x1 μέχρι x11) και τη μεταβλητή απόκρισης PRICE, όπως φαίνεται στον Πίνακα 1.1. Τις μεταβλητές constr και style τις ορίζουμε σαν κατηγορικές μεταβλητές, όπου για την constr: level 1 = brick, level 2 = brick and frame, level 3 = aluminum and frame, level 4 = frame και για τη style: level 1 = two story, level 2 = one and a half story, level 3 = ranch.



Γραφήματα 1.1: Περιγραφή της PRICE σε σχέση με τις μεταβλητές constr και style

Στη συνέχεια περιγράφουμε τις μεταβλητές του προβλήματος γραφικά για την εξαγωγή χρήσιμων συμπερασμάτων. Πρώτα τις κατηγορικές μεταβλητές constr (Construction type) και style που περιγράφονται με τα boxplots, για καλυτερη σύγκριση των επιπέδων(levels) των χαρακτηριστικών για την κάθε μεταβλητή (Γραφήματα 1.1). Από εδώ παρατηρούμε ότι η μεταβλητή απόκρισης PRICE συνδέεται με τον τύπο κατασκευής (constr), αλλά και με το στυλ (style) για όλα τα επίπεδα. Και στη συνέχεια, περιγράφονται με scatterplots οι υπόλοιπες ποσοτικές μεταβλητές (Γραφήματα 1.2).



Γραφήματα 1.2: Περιγραφή της PRICE σε σχέση με την κάθε μία από τις υπόλοιπες μεταβλητές

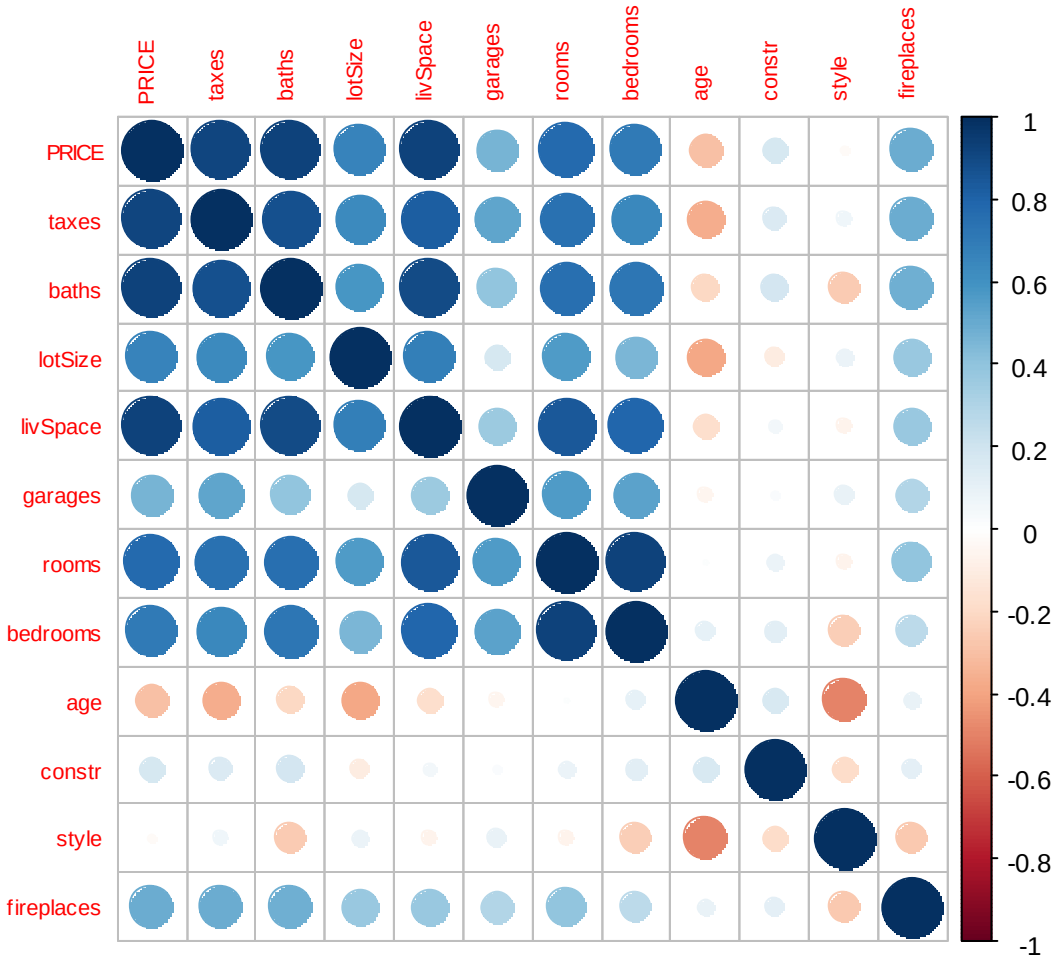
Από τα πιο πάνω Γραφήματα 1.2 παρατηρούμε ότι η μεταβλητή απόκρισης PRICE σχετίζεται με όλες τις επεξηγηματικές μεταβλητές και ότι η υπόθεση γραμμικότητας είναι αρκετά λογική για όλες. Όμως βλέπουμε ότι υπάρχουν δύο ακραίες παρατηρήσεις στα δεδομένα μας (συγκεκριμένα η $9^{η}$ και η $10^{η}$ παρατήρηση), οι οποίες φαίνεται να διαφοροποιούνται σε σχέση με όλες τις μεταβλητές. Αυτές οι παρατηρήσεις είναι τα μικρά κυκλάκια πάνω-πάνω, όπως διακρίνουμε στα γραφήματα. Επίσης με τα πιο πάνω δεν ελέγχουμε την εγκυρότητα της προσαρμογής του γραμμικού μοντέλου. Αυτό θα το δούμε πιο κάτω μόνο για το βέλτιστο μοντέλο που θα καταλήξουμε.

Πριν γίνει η προσαρμογή του μοντέλου, είναι σημαντικός και ο έλεγχος της συσχέτισης μεταξύ των μεταβλητών, αριθμητικά αλλά και σχηματικά.

	taxes	baths	lotSize	livSpace	garages	rooms	bedrooms	age	constr	style	fireplaces
taxes	1	0.88	0.63	0.83	0.52	0.74	0.65	-0.36	0.15	0.06	0.49
baths	0.88	1	0.58	0.89	0.40	0.76	0.73	-0.20	0.19	-0.25	0.48
lotSize	0.63	0.58	1	0.68	0.18	0.56	0.46	-0.38	-0.11	0.09	0.38
livSpace	0.83	0.89	0.68	1	0.36	0.84	0.80	-0.17	0.06	-0.06	0.37
garages	0.52	0.40	0.18	0.36	1	0.57	0.54	-0.05	0.03	0.10	0.29
rooms	0.74	0.76	0.56	0.84	0.57	1	0.92	0.01	0.08	-0.06	0.40
bedrooms	0.65	0.73	0.46	0.79	0.54	0.92	1	0.11	0.13	-0.24	0.27
age	-0.36	-0.20	-0.38	-0.18	-0.06	0.01	0.11	1	0.17	-0.50	0.10
constr	0.15	-0.19	0.10	0.06	0.03	0.08	0.13	0.17	1	-0.18	0.11
style	0.064	-0.25	0.089	-0.06	0.097	-0.06	-0.24	-0.49	-0.18	1	-0.26
fireplaces	0.49	0.48	0.38	0.37	0.29	0.40	0.27	0.09	0.11	-0.26	1

Πίνακας 1.2: Πίνακας συσχέτισης των μεταβλητών

Από τον Πίνακα 1.2, αλλά και από τη σχηματική περιγραφή (Γράφημα 1.3) βλέπουμε καθαρά τη συσχέτιση που υπάρχει μεταξύ κάποιων μεταβλητών. Όσο πιο μεγάλος και πιο σκούρος είναι ο κύκλος, τόσο πιο μεγάλη συσχέτιση υπάρχει μεταξύ των συγκεκριμένων μεταβλητών (αριθμητικά είναι πιο κοντά στη μονάδα). Συγκεκριμένα η taxes με τη baths, τη livSpace, τη rooms ακόμα και τη bedrooms έχει μεγάλη συσχέτιση. Η baths έχει έντονη συσχέτιση με τη livSpace, τη rooms αλλά και με τη bedrooms. Και η rooms έχει με τη bedrooms κλπ.



Γράφημα 1.3: Η συσχέτιση μεταξύ των μεταβλητών

Επίσης μπορούμε να ελέγξουμε τη συσχέτιση από τις τιμές των VIFs της κάθε μεταβλητής (Πίνακας 1.3). Αναφέρουμε ότι ο συντελεστής Διόγκωσης της Διακύμανσης (VIF), με $j = 1, \dots, n$ οι συντελεστές παλινδρόμησης δίνεται από τον πιο κάτω τύπο (Montgomery, Peck, & Vining, 2006):

$$VIF_j = \frac{1}{1 - {R_j}^2}$$

Μεταβλητές	**VIFs**
taxes	9.5
baths	15.6
lotSize	2.7
livSpace	12.2
garages	2.2
rooms	15.4
bedrooms	14.5
age	2.7
constr	1.3
style	4.0
fireplaces	2.4

Πίνακας 1.3: Οι τιμές των VIFs για την κάθε μεταβλητή

Όσο μεγαλύτερη η τιμή των VIFs τόσο πιο σοβαρό είναι το πρόβλημα της πολυσυγγραμμικότητας (§ 1.2.4.1). Οπότε, παρατηρούμε ότι οι μεταβλητές taxes, baths, livSpace, rooms και bedrooms έχουν αρκετά ψηλό VIF και θα μας δημιουργήσουν προβλήματα αργότερα.

Με τη χρήση των πιο κάτω εντολών γίνεται η προσαρμογή του γενικού γραμμικού μοντέλου στην R και λαμβάνουμε τα Αποτελέσματα 1.1:

```
>resultsNW<-lm(PRICE~taxes+baths+lotSize+livSpace+garages+rooms+bedrooms+age+constr+
+style+fireplaces)
> summary(resultsNW)
```

```
Call:
lm(formula = PRICE ~ taxes + baths + lotSize + livSpace + garages
   + rooms + bedrooms + age + constr + style + fireplaces)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5268 -1.6243 -0.4324  1.7550  5.9517

Coefficients:
             Estimate  Std. Error  t value Pr(>|t|)
(Intercept)  4.49745    6.92590    0.649   0.52741
taxes       -0.11960    0.94412   -0.127   0.90114
baths        9.44068    7.05442    1.338   0.20375
lotSize      0.57090    0.52141    1.095   0.29343
livSpace    17.53154    5.38114    3.258   0.00623 **
garages      3.36788    1.89979    1.773   0.09969 .
rooms       -1.43117    2.73760   -0.523   0.60992
bedrooms    -1.18523    4.54000   -0.261   0.79813
age         -0.09486    0.08882   -1.068   0.30495
constr2      5.81338    2.22463    2.613   0.02146 *
constr3      7.35758    3.61892    2.033   0.06299 .
constr4      3.98752    2.13259    1.870   0.08420 .
style2       3.16570    3.50759    0.903   0.38319
style3       1.88122    6.22726    0.302   0.76736
fireplaces   3.39897    2.38626    1.424   0.17789
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.738 on 13 degrees of freedom
Multiple R-squared: 0.9664,     Adjusted R-squared: 0.9303
F-statistic: 26.73 on 14 and 13 DF,  p-value: 2.791e-07
```

Αποτελέσματα 1.1

Επιπλέον, έγινε υπολογισμός των κριτηρίων AIC=163.821 και BIC=185.1368, αλλά και των 95% διαστημάτων εμπιστοσύνης των παραμέτρων (Πίνακας 1.4).

Από τα αποτελέσματα της πιο πάνω ανάλυσης παλινδρόμησης είχαμε όλες τις μεταβλητές μέσα και παρατηρούμε ότι μόνο η livSpace και η constr2 είναι στατιστικά σημαντικές, με p-τιμές 0.00623 και 0.02146 αντίστοιχα, που είναι αρκετά μικρές. Επίσης οριακά σημαντικές μπορούμε να θεωρήσουμε και τις μεταβλητές garages, constr3 και constr4, με p-τιμές 0.09969, 0.06299 και 0.08420, αντίστοιχα. Ο συντελεστής προσδιορισμού R^2 = 96.64%, αλλά και ο διορθωμένος R^2_{adj}=93.03% είναι πολύ ψηλοί (όσο πιο κοντά στη μονάδα βρίσκονται τόσο πιο καλά προσαρμοσμένο είναι το μοντέλο) αλλά και η p-τιμή του F-έλεγχου είναι πολύ μικρή (<0.001).

	2.5%	97.5%
(Intercept)	-10.4650505	19.45994778
taxes	-2.1592548	1.92006072
baths	-5.7994618	24.68081606
lotSize	-0.5555386	1.69733100
livSpace	5.9063024	29.15677360
garages	-0.7363800	7.47213344
rooms	-7.3454032	4.48305525
bedrooms	-10.9933138	8.62285026
age	-0.2867501	0.09702465
constr2	1.0073663	10.61938867
constr3	-0.4606143	15.17577041
constr4	-0.6196684	8.59471130
style2	-4.4119789	10.74338047
style3	-11.5719556	15.33439938
fireplaces	-1.7562293	8.55416326

Πίνακας 1.4: 95% Διαστήματα εμπιστοσύνης για τις παραμέτρους

Όμως, το ότι έχουμε τόσες μεταβλητές στατιστικά μη σημαντικές και, μιας και όπως έχουμε δει προηγουμένως υπάρχει έντονη συσχέτιση μεταξύ των περισσοτέρων μεταβλητών, μπαίνουμε στη σκέψη ότι σίγουρα το μοντέλο έχει περιθώρια βελτίωσης. Οπότε θα χρησιμοποιήσουμε τις τρεις διαδικασίες επιλογής μοντέλου με βήματα, αλλά και άλλες τεχνικές που έχουν αναφερθεί στη θεωρία, για να καταλήξουμε στην τελική επιλογή των στατιστικά σημαντικότερων μεταβλητών για το πρόβλημα, την επιλογή του καταλληλότερου μοντέλου.

Αρχικά, χρησιμοποιούμε τις τρεις διαδικασίες επιλογής μοντέλου με βήματα, με βάση το κριτήριο AIC, για την επιλογή του βέλτιστου μοντέλου, στα δεδομένα της συγκεκριμένης εφαρμογής. Πριν ξεκινήσουμε τις διαδικασίες, θέτουμε το κενό (null) και το πλήρες (full) μοντέλο (βλ. Παράρτημα Β.3). Στη συνέχεια εισάγοντας τις ακόλουθες εντολές, εκτελούμε τις διαδικασίες Backward, Forward και Stepwise, αντίστοιχα:

```
> step(full, data, direction="backward")
> step(null, scope=list(lower=null, upper=full), direction="forward")
> step(null, scope=list(upper=full), direction="both")
```

Και λαμβάνουμε τα αποτελέσματα που έχουν καταγραφεί στους Πίνακες 1.5 a,b και c, για την κάθε διαδικασία επιλογής με την αντίστοιχη σειρά που έχουν καταγραφεί πιο πάνω.

Μεταβλ. / Steps	taxes	baths	lotSize	livSpace	garages	rooms	bedrooms	age	constr	style	fireplaces	AIC
1	x	x	x	x	x	x	x	x	x	x	x	87.13
2	x	x	x	x	x	x		x	x	x	x	85.15
3	x	x		x	x	x		x	x	x	x	83.32
4	x	x		x	x	x		x	x		x	82.08
5	x	x		x	x			x	x		x	80.84
6	x	x		x				x	x		x	80.29

Πίνακας 1.5.a: Αποτελέσματα της Backward Elimination

Μεταβλ. / Steps	taxes	baths	lotSize	livSpace	garages	rooms	bedrooms	age	constr	style	fireplaces	AIC
1												149.39
2		x										97.17
3	x	x										88.02
4	x	x		x								79.97

Πίνακας 1.5.b: Αποτελέσματα της Forward Selection

Μεταβλ. / Steps	taxes	baths	lotSize	livSpace	garages	rooms	bedrooms	age	constr	style	fireplaces	AIC
1												149.39
2		x										97.17
3	x	x										88.02
4	x	x		x								79.97

Πίνακας 1.5.c: Αποτελέσματα της Stepwise Selection

Μετά την εκτέλεση και των τριών διαδικασιών επιλογής με βήματα, οδηγούμαστε πολύ κοντά στην επιλογή του βέλτιστου μοντέλου. Μετά την εκτέλεση των forward και stepwise selection καταλήγουμε στο μοντέλο με τις μεταβλητές taxes, baths και livSpace. Μετά και την εκτέλεση και της backward selection, καταλήγουμε στην επιπλέον επιλογή

εκτός των τριών προαναφερθέντων, των μεταβλητών age, constr και fireplaces. Συνολικά επιλέχτηκαν οι πιο πάνω έξι μεταβλητές, που αντίστοιχα είναι οι φόροι, ο αριθμός λουτρών (number of baths), τα τετραγωνικά του κατοικίσιμου χώρου (εμβαδό), η ηλικία του σπιτιού, το υλικό κατασκευής και ο αριθμός των τζακιών. Οι πιο πάνω μεταβλητές επιλέχτηκαν μετά την εκτέλεση των τριών διαδικασιών, ως στατιστικά σημαντικότερες στη συγκεκριμένη εφαρμογή και στα συγκεκριμένα δεδομένα.

Όμως, το κριτήριο AIC αγνοεί το πρόβλημα της πολυσυγγραμμικότητας. Ίσως σε αυτή την περίπτωση ένα οποιαδήποτε μοντέλο θα μπορούσε να είναι κατάλληλο και όχι απαραιτήτως αυτό που έχει το χαμηλότερο AIC. Όπως έχουμε δει προηγουμένως γραφικά αλλά και υπολογιστικά, υπάρχει έντονη συσχέτιση μεταξύ κάποιων επεξηγηματικών μεταβλητών που μας οδηγεί σε όχι και τόσο έγκυρα αποτελέσματα. Έτσι θα χρησιμοποιήσουμε τις δύο (penalized) τεχνικές με ποινή, για τις οποίες μιλήσαμε προηγουμένως (§ 1.2.4.2 και 1.2.4.3) οι οποίες αντιμετωπίζουν και το πρόβλημα της πολυσυγγραμμικότητας, τη Lasso και τη ridge regression.

Αρχικά χρησιμοποιούμε τη μέθοδο Lasso και προσπαθούμε να βρούμε το βέλτιστο λ_1 με τη μέθοδο CVL (Cross Validation) (§ 1.2.4.4), ώστε να το χρησιμοποιήσουμε στην τεχνική αυτή. Με τη βοήθεια των πιο κάτω εντολών έχουμε εκτός από αριθμητικά και γραφικά αποτελέσματα:

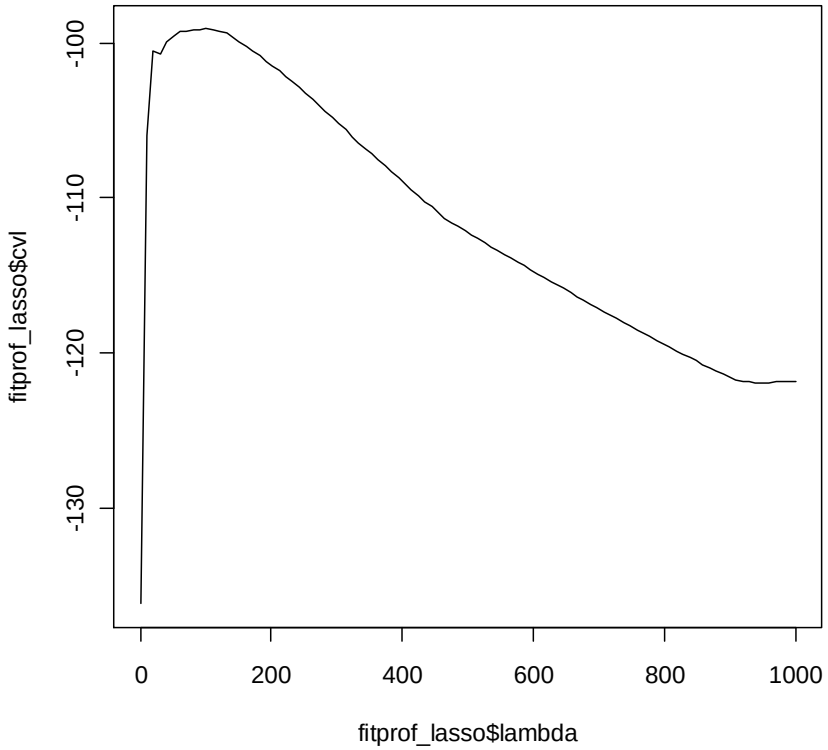
```
> fitprof_lasso<-profL1(PRICE, penalized=dataNW[,2:12], fold=10, minl=0.01, maxl=1000)
> plot(fitprof_lasso$lambda, fitprof_lasso$cvl,type="l")

> opt_lasso<-optL1(PRICE, penalized=dataNW[,2:12], fold=fitprof_lasso$fold)
```

Οπότε καταλήγουμε ότι το βέλτιστο λ_1 για τη Lasso να είναι ίσο με 98 (βλ. Γράφημα 1.4), με τη συνάρτηση cvl να παίρνει την τιμή ίση με -87.801. Έτσι τώρα προσαρμόζουμε το μοντέλο με τη μέθοδο αυτή και το συγκεκριμένο βέλτιστο λ_1 χρησιμοποιώντας το πακέτο Penalized (Παράρτημα D.1), εκτελώντας:

```
> fit_final_lasso<-penalized(PRICE~taxes+baths+lotSize+livSpace+garages
+rooms+bedrooms+age+constr+style+fireplaces, dataNW, lambda1=98)

> coefficients(fit_lasso,"penalized")
```



Γράφημα 1.4: Το βέλτιστο λ_1

Έτσι καλώντας και τις μεταβλητές (coefficients), εμφανίζονται όλες οι penalized-παραμέτροι του μοντέλου για να ελέγξουμε ποιες έχουν γίνει μηδέν ή κοντεύουν στο μηδέν και ποιες όχι. Αυτές που είναι κοντά στο μηδέν ή είναι μηδέν απορρίπτονται, για να μείνουν οι υπόλοιπες ως σημαντικές για το βέλτιστο μοντέλο που θέλουμε.

(Intercept)	taxes	baths	lotSize	livSpace	garages
7.6991184	3.9159136	0.000000e+00	0.3587332	0.000000e+00	0.000000e+00
rooms	bedrooms	age	constr2	constr3	constr4
0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
S tyle2	style3	fireplaces			
0.000000e+00	0.000000e+00	0.000000e+00			

Αποτελέσματα 1.2

Παρατηρούμε (Αποτελέσματα 1.2) ότι μετά την εκτέλεση όλων των βημάτων της τεχνικής Lasso, μπορούμε να πούμε ότι οι σημαντικότερες μεταβλητές είναι οι taxes και lotSize με τιμές 3.916 και 0.359, αντίστοιχα.

Στη συνέχεια χρησιμοποιούμε τη μέθοδο της Ridge regression και προσπαθούμε να βρούμε το βέλτιστο λ_2 με τη βοήθεια της μεθόδου CVL (Cross Validation) (§ 1.2.4.4), ώστε να το χρησιμοποιήσουμε στην τεχνική αυτή. Εκτελούμε τις πιο κάτω εντολές για να βρούμε το βέλτιστο λ_2, πρώτα αριθμητικά:

```
> fitprof_ridge<-profL2 (PRICE, penalized=dataNW[2:12], fold=fitprof_lasso$fold, minl=0.01,
+maxl=2000)
> opt_ridge<-optL2(PRICE, penalized=dataNW[,2:12],fold=fitprof_ridge$fold)
```

Και ακολούθως ελέγχουμε και γραφικά με τις πιο κάτω εντολές :

```
> plot(fitprof_ridge$lambda,fitprof_ridge$cvl,type="l",log="x")
> plotpath(fitprof_ridge$fullfit, log="x")
```

Οπότε καταλήγουμε ότι το βέλτιστο λ_2 για τη Ridge regression να είναι ίσο με 35 (Γραφήματα 1.5 και 1.6), με τη συνάρτηση cvl να παίρνει τιμή ίση με -98.028. Έτσι τώρα προσαρμόζουμε το μοντέλο με τη μέθοδο αυτή και το συγκεκριμένο βέλτιστο λ_2 χρησιμοποιώντας το πακέτο-penalized (Παράρτημα D.1), εκτελώντας:

```
> fit_final_ridge<-penalized (PRICE~taxes+baths+lotSize+livSpace+garages+rooms +bedrooms
+age+constr+style+fireplaces, dataNW, lambda2=35)
> coefficients(fit_ridge,"penalized")
```

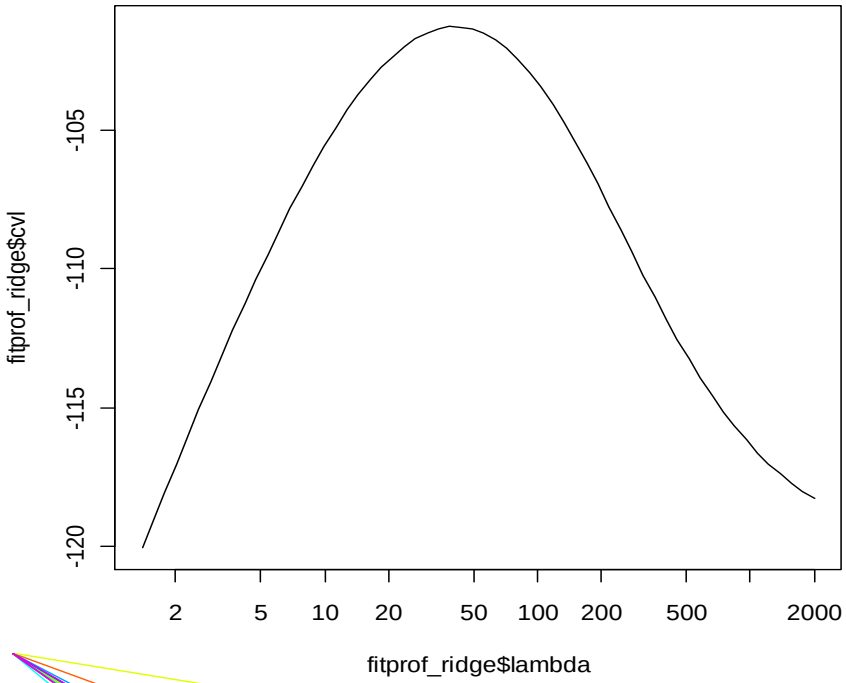
Έτσι καλώντας και τις μεταβλητές (coefficients), εμφανίζονται όλες οι penalized-παραμέτροι του μοντέλου για να ελέγξουμε ποιες έχουν γίνει μηδέν ή κοντεύουν στο μηδέν και ποιες όχι.

(intercept)	taxes	baths	lotSize	livSpace	garages
0.28066479	2.77188317	0.76220775	0.88005277	1.02611167	0.24071106
rooms	bedrooms	age	constr 1	constr2	constr3
1.11378803	0.72050073	-0.04460218	-0.47182776	0.38810512	-0.29283396
constr4	style1	style2	style3	fireplaces	
0.37655661	0.14077727	0.04456409	-0.18534136	0.36717264	

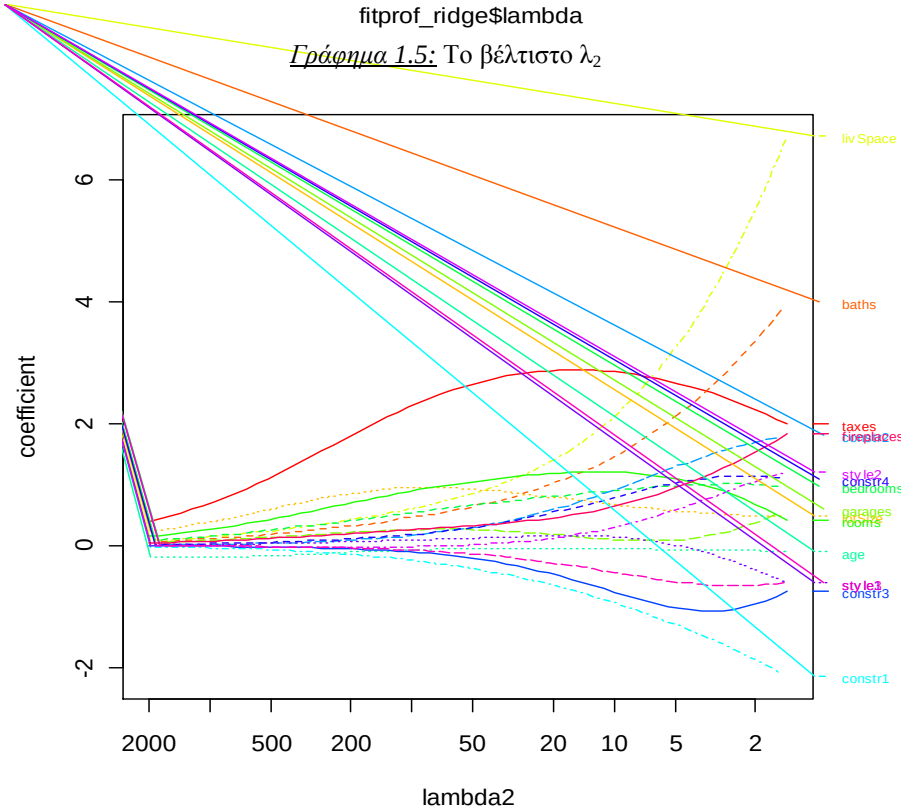
Αποτελέσματα 1.3

Παρατηρούμε από τα Αποτελέσματα 1.3 ότι, μετά την εκτέλεση όλων των βημάτων της τεχνικής Ridge regression, μπορούμε να πούμε ότι οι καταλληλότερες μεταβλητές είναι οι taxes, livSpace, baths και rooms, με τις τιμές τους 2.772, 1.026, 0.762 και 1.114,

αντίστοιχα, να είναι αρκετά μακριά από το μηδέν. Μπορούμε να πούμε ότι η τιμή της μεταβλητής lotSize (0.880) είναι και αυτή μακριά από το μηδέν, όμως δεν την επιλέγουμε γιατί είναι αρκετά ψηλή η συσχέτισή της με τη livSpace.



Γράφημα 1.5: Το βέλτιστο λ_2



Γράφημα 1.6: Πώς κινούνται οι μεταβλητές για το συγκεκριμένο λ_2

1.3.3 ΤΟ ΒΕΛΤΙΣΤΟ ΜΟΝΤΕΛΟ

Έτσι μετά από τις τρεις διαδικασίες κατά βήματα επιλογής, αλλά και τις δύο τεχνικές με ποινή, τη Lasso και τη Ridge regression, ελέγχουμε όλα τα αποτελέσματα και επιλέγουμε ως καλύτερο μοντέλο αυτό με τις μεταβλητές taxes, baths και livSpace, όπου είναι οι φόροι, ο αριθμός των λουτρών και η έκταση του κατοικίσιμου χώρου, αντίστοιχα. Αυτές οι μεταβλητές έχουν θεωρηθεί ως οι σημαντικότερες από τις τρεις διαδικασίες κατά βήματα και από την τεχνική της Ridge regression. Η taxes επιλέχτηκε ως σημαντική και από την τεχνική Lasso. Πιο κάτω παρουσιάζουμε την ανάλυση παλινδρόμησης του βέλτιστου μοντέλου που επιλέχτηκε (Αποτελέσματα 1.4.) με τη βοήθεια των εντολών:

```
> results_teliko<-lm(PRICE ~ taxes+baths+livSpace)
> summary(results_teliko)
```

Επίσης ελέγχουμε και τα 95% διαστήματα εμπιστοσύνης, στον Πίνακα 1.6.

```
Call:
lm(formula = PRICE ~ taxes + baths + livSpace)

Residuals:
   Min      1Q  Median      3Q     Max
-6.7018 -2.2455 -0.2396  2.3809  6.3789

Coefficients:
             Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  -1.0658    2.4028    -0.444   0.66132
taxes         1.8965    0.5626    3.371    0.00253 **
baths         8.1897    4.7833    1.712    0.09976 .
livSpace     10.0625    3.1255    3.219    0.00366 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.905 on 24 degrees of freedom
Multiple R-squared: 0.9324,     Adjusted R-squared: 0.9239
F-statistic: 110.3 on 3 and 24 DF,  p-value: 3.567e-14
```

Αποτελέσματα 1.4

	2.5%	97.5%
(Intercept)	-6.024976	3.893324
taxes	0.735305	3.057660
baths	-1.682440	18.061854
livSpace	3.611752	16.513164

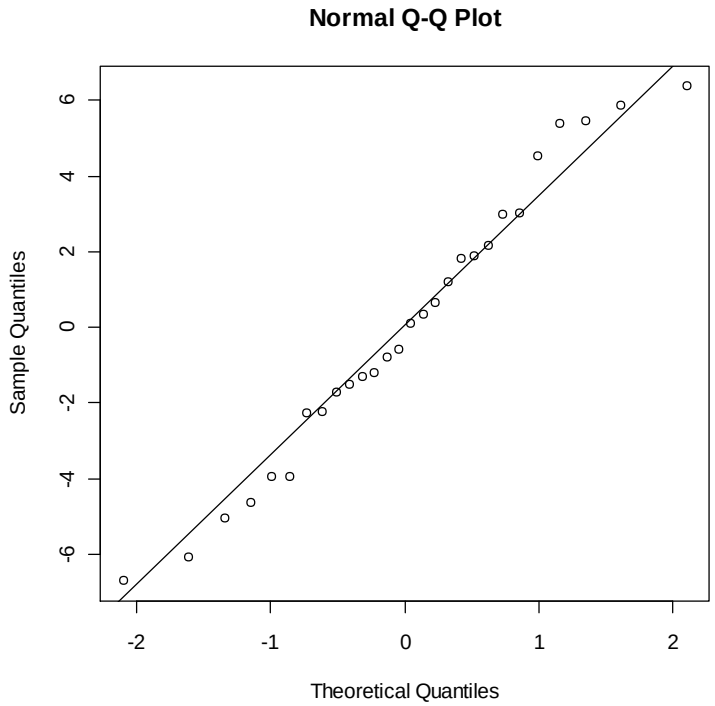
Πίνακας 1.6: 95% διαστήματα εμπιστοσύνης τελικού μοντέλου

Από τα πιο πάνω Αποτελέσματα 1.4 της ανάλυσης παλινδρόμησης του γενικού γραμμικού μοντέλου που επιλέχτηκε μετά τις διαδικασίες και τεχνικές επιλογής, παρατηρούμε ότι οι μεταβλητές taxes και livSpace είναι σημαντικές με p-τιμές 0.0025 και 0.0036, αντίστοιχα, ενώ η μεταβλητή baths είναι οριακά σημαντική με p-τιμή 0.099.

Κατ' αρχάς, το μοντέλο που επιλέχτηκε από τις διαδικασίες επιλογής με βήματα με βάση το κριτήριο AIC περιέχει τις μεταβλητές αυτές. Στη συνέχεια με τη βοήθεια των δύο τεχνικών με ποινή κλειδώσαμε την επιλογή μας, καθώς μέσα στην επιλογή της Ridge regression ήταν και οι τρεις αυτές, αλλά και στην επιλογή της Lasso ήταν η μεταβλητή taxes. Μπορούσαμε να επιλέξουμε και την μεταβλητή lotSize αλλά απορρίφθηκε λόγω της μεγάλης συσχέτισής της με τη μεταβλητή livSpace. Έτσι καταλήξαμε στο μοντέλο των πιο πάνω τριών μεταβλητών, ως βέλτιστο. Ελέγχοντας και τα κριτήρια επιλογής, με το συντελεστή προσδιορισμού R^2 = 93.2%, αλλά και το διορθωμένο R^2_{adj}=92.4%, να είναι πολύ ψηλοί, παρόλο που από τις έντεκα μεταβλητές επιλέξαμε μόνο αυτές τις τρεις. Όπως επίσης και η p-τιμή του έλεγχου-F είναι πολύ μικρή (<0.001). Οπότε έχουμε αρκετές ενδείξεις ώστε να επιλέξουμε αυτό το μοντέλο σαν βέλτιστο.

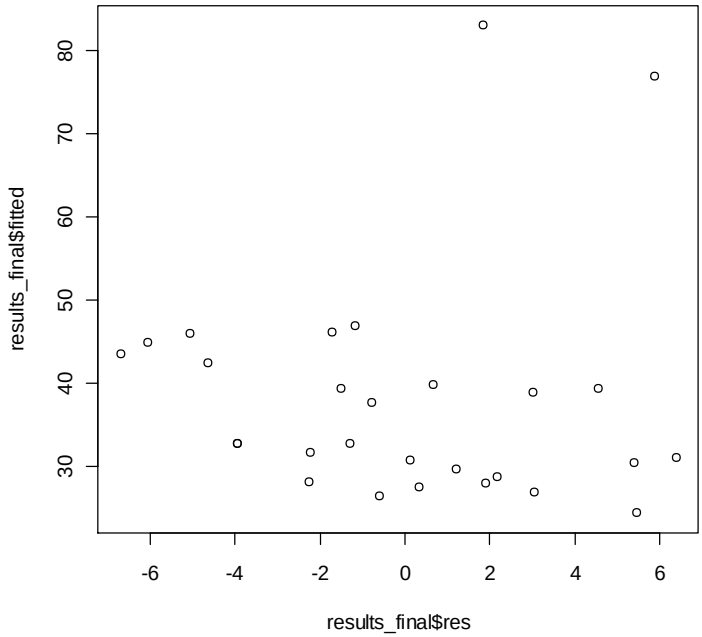
Τέλος, ελέγχουμε και τις προϋποθέσεις του γενικού γραμμικού μοντέλου, ώστε να κλειδώσουμε την ορθότητα των αποτελεσμάτων μας.

- ✓ Η κανονικότητα των υπολοίπων, όπως φαίνεται πολύ καθαρά από το Γράφημα 1.7, είναι λογική διότι η ευθεία που σχηματίζεται είναι σχεδόν τέλεια προσαρμοσμένη σε όλα τα σημεία.



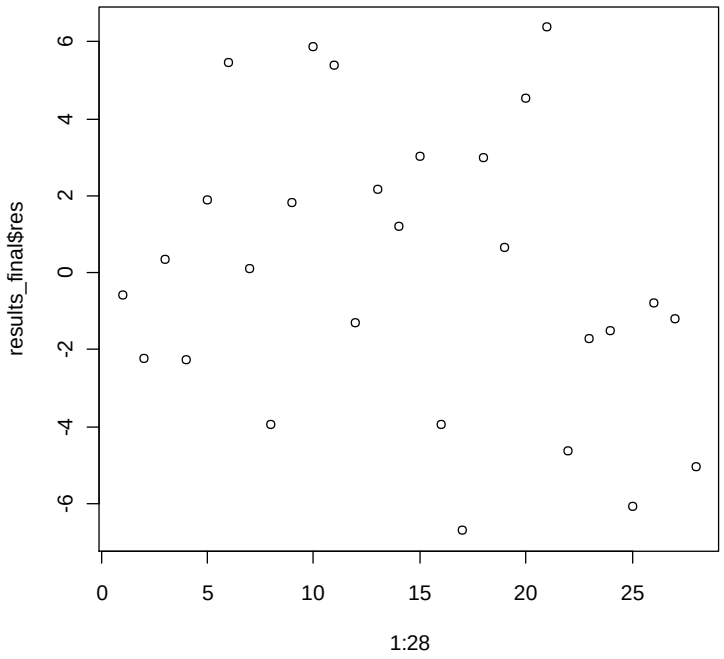
Γράφημα 1.7: Κανονικότητα υπολοίπων

- ✓ Με το Γράφημα 1.8 παρουσιάζονται τα υπόλοιπα (residuals (res)) (άξονας x) συναρτήσει των προβλεπόμενων τιμών (fitted) (άξονας y). Παρατηρούμε τα ζεύγη των τιμών να μην εμφανίζουν κάποιο συστηματικό τρόπο συμπεριφοράς, οπότε η υπόθεση της ομοσκεδαστικότητας είναι λογική.



Γράφημα 1.8: Ομοσκεδαστικότητα

- ✓ Στο Γράφημα 1.9 φαίνεται η κατασκευή ενός διάγραμματος υπολοίπων (residuals (res)) (άξονας y) σε σχέση με τη σειρά των δεδομένων (1:28) (άξονας x). Βλέπουμε ότι υπάρχει ανεξαρτησία μεταξύ των υπολοίπων, γιατί δεν παρουσιάζεται κάποια σχέση μεταξύ τους και τα υπόλοιπα συμπεριφέρονται τυχαία.



Γράφημα 1.9: Ανεξαρτησία υπολοίπων

Μετά την πιο πάνω «ολόκληρη» διαδικασία, καταλήγουμε στο εξής μοντέλο, για το οποίο ισχύουν και οι προυποθέσεις του γενικού γραμμικού μοντέλου:

$$\hat{y} = -1.0658 + 1.8965taxes + 8.1897baths + 10.0625livSpace$$

Πιο κάτω περιγράφουμε το πιο πάνω μοντέλο με τις στατιστικά σημαντικές μεταβλητές taxes, baths και livSpace. Κατ' αρχάς η αύξηση της έκτασης του οικοπέδου κατά 1000 τετραγωνικά πόδια (lotSize), η αύξηση του αριθμού των garage κατά μία μονάδα, η κατά ένα έτος επιπλέον αύξηση της ηλικίας του σπιτιού, η αύξηση κατά μία μονάδα του αριθμού δωματίων (rooms), των υπνοδωματίων (bedrooms) και των τζακιών (fireplaces), το στυλ (style) αλλά και ο τύπος της κατασκευής (constr) δεν επηρεάζουν καθόλου την τιμή του σπιτιού (PRICE), καθώς θεωρούνται ως μη σημαντικές για το μοντέλο που έχουμε επιλέξει ως βέλτιστο.

Τώρα, για την αύξηση του εμβαδού του κατοικίσιμου χώρου του σπιτιού (livSpace) κατά 1000 τετραγωνικά πόδια, η τιμή του σπιτιού, με τις υπόλοιπες μεταβλητές σταθερές, αυξάνεται κατά $10063. Επίσης, με την αύξηση της φορολόγησης του σπιτιού (taxes) με $100 αυξάνεται η τιμή κατά $1897, με τις υπόλοιπες μεταβλητές σταθερές. Τέλος, κάθε επιπλέον λουτρό μέσα στο σπίτι αυξάνει την τιμή του κατά $8190 και πάλι κρατώντας τις άλλες μεταβλητές σταθερές.

1.3.4 ΣΥΜΠΕΡΑΣΜΑΤΑ

Τα συμπεράσματα και οι παρατηρήσεις που εξάγονται μετά από τέτοια ανάλυση παλινδρόμησης είναι πάρα πολύ σημαντικά και χρήσιμα για οποιονδήποτε, φτάνει να γίνει σωστά η προσαρμογή του κατάλληλου μοντέλου, αλλά και η επιλογή των σημαντικότερων μεταβλητών για το κάθε πρόβλημα. Με τις κατάλληλες διαδικασίες και τεχνικές επιλογής κάποιου μοντέλου (ή μεταβλητών) κερδίζουμε χρόνο (σίγουρα περισσότερες μεταβλητές μας καθυστερούν), χρήμα (είναι πολλά τα κόστη για να αντλήσουμε δεδομένα, όποιον τρόπο και να χρησιμοποιήσουμε) αλλά και ταλαιπωρία, κούραση, κόπο. Αν και το δείγμα μας, στην πιο πάνω εφαρμογή, ήταν αρκετά μικρό (n = 28 κατοικίες) σε σχέση με το πλήθος των επεξηγηματικών μεταβλητών (k = 11), εντούτοις ο σκοπός επιτεύχθηκε, καθώς μελετήσαμε αρκετές τεχνικές και μεθόδους επιλογής του βέλτιστου μοντέλου.

2. ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

2.1 ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

2.1.1 ΕΙΣΑΓΩΓΗ

Στο προηγούμενο κεφάλαιο αναφερθήκαμε στο γενικό γραμμικό μοντέλο και στη μη παραβίαση κάποιων υποθέσεων, ώστε να έχουμε πιο έγκυρα αποτελέσματα στα προβλήματα που έχουμε. Μία σημαντική υπόθεση που αναφερθήκαμε, είναι η κανονικότητα της μεταβλητής απόκρισης, δηλαδή να ακολουθεί την Κανονική κατανομή. Μια δημοφιλής πρακτική, πλέον, που χρησιμοποιείται ακόμα και σε περιπτώσεις που η υπόθεση της κανονικότητας δεν ισχύει ούτε προσεγγιστικά είναι η χρήση των **γενικευμένων γραμμικών μοντέλων (Generalized Linear Models, GLM)**.

Η θεματολογία των γενικευμένων γραμμικών μοντέλων, στο μεγαλύτερο μέρος της, δεν αποτελεί κάτι νέο στην στατιστική, αλλά ουσιαστικά ομαδοποιεί έννοιες και τεχνικές που προϋπάρχουν δημιουργώντας ένα ενοποιημένο θεωρητικό και εννοιολογικό πλαίσιο. Η πρόοδος στη στατιστική θεωρία μαζί με την ανάπτυξη των υπολογιστών μας επέτρεψαν να δημιουργήσουμε μεθόδους ανάλογους με αυτές που έχουν αναπτυχθεί για τα γραμμικά μοντέλα σε περιπτώσεις που οι αποκρίσεις ακολουθούν κατανομή διαφορετική από την κανονική, δεν είναι απαραίτητα συνεχείς (μπορεί να είναι κατηγορικές μεταβλητές) και δεν χρειάζεται να είναι στην απλή γραμμική μορφή.

Το 1972 οι Nelder και Wedderburn παρουσίασαν μια ενοποιημένη θεωρία για γραμμικά μοντέλα που δεν απαιτεί την υπόθεση της κανονικότητας για τη μεταβλητή απόκρισης. Σύμφωνα με αυτή, τα γραμμικά μοντέλα μπορούν να μελετηθούν ενιαία κάτω από την υπόθεση ότι η κατανομή της μεταβλητής απόκρισης ανήκει στην **Εκθετική οικογένεια** κατανομών. Επίσης και για όλες τις κατανομές μέσα στην οικογένεια αυτή, οι εκτιμητές μέγιστης πιθανοφάνειας (ε.μ.π.) των παραμέτρων του μοντέλου μπορούν να βρεθούν με τον ίδιο αλγόριθμο.

Τα πακέτα λογισμικού αποτελούν το βασικό εργαλείο για υπολογισμούς παραμέτρων. Η εκτίμηση παραμέτρων του γραμμικού μοντέλου $\boldsymbol{X\beta}$ επεκτάθηκε στην εκτίμηση

γραμμικών συνδυασμών του τύπου $g(\boldsymbol{X\beta})$. Θεωρητικά οι διαδικασίες εκτίμησης είναι απλές. Στην πράξη όμως, απαιτούν ένα μεγάλο όγκο υπολογισμών οι οποίοι έγιναν εφικτοί μόνο μέσω υπολογιστών, με τη βοήθεια αριθμητικών προσεγγίσεων μη γραμμικών συναρτήσεων.

2.1.2 ΤΟ ΜΟΝΤΕΛΟ

Στα γραμμικά ή μη γραμμικά μοντέλα κυρίαρχο ρόλο παίζει η Κανονική Κατανομή $N(\mu, \sigma^2)$, την οποία ακολουθεί η απόκριση. Αυτό όμως δε συμβαίνει πάντοτε. Για παράδειγμα, η απόκριση μπορεί να ακολουθεί την Διωνυμική κατανομή δηλ. τα αποτελέσματα να είναι της μορφής: 0 (= αποτυχία) ή 1 (= επιτυχία). Στην περίπτωση αυτή έχουμε τα **Γενικευμένα Γραμμικά Μοντέλα (GLM, Generalized Linear Models)**, των οποίων τα δεδομένα ακολουθούν κατανομές της Εκθετικής οικογένειας (π.χ. κατανομές Κανονική, Διωνυμική, Poisson, Εκθετική, Γάμμα κλπ). Για παράδειγμα, αν οι y_x είναι οι αποκρίσεις, το GLM γράφεται:

$$g(y_x) = g[E(y_x)] = \boldsymbol{x'\beta},$$

$$\textit{όπου } y_x\textit{: συνεχής τυχαία μεταβλητή της } N(\mu_x, \sigma^2),$$

$$\boldsymbol{x'} = (x_1, x_2, \dots, x_k)\textit{: ανεξάρτητες μεταβλητές με } x_0 \equiv 1,$$

$$\boldsymbol{\beta}\textit{: το διάνυσμα άγνωστων παραμέτρων}$$

Κάθε γενικευμένο γραμμικό μοντέλο αποτελείται από τρία συστατικά:

- Την κατανομή απόκρισης
- Μία γραμμική παράμετρο πρόβλεψης που περιέχει τις μεταβλητές παλινδρόμησης x_i
- Τη συνάρτηση σύνδεσης (link function), η οποία ενώνει τη γραμμική παράμετρο πρόβλεψης με τη μέση τιμή απόκρισης.

Στα Γενικευμένα γραμμικά μοντέλα σημαντικό ρόλο παίζουν η κατανομή της μεταβλητής απόκρισης και το μοντέλο που συνδέει τη μέση απόκριση με τις μεταβλητές

παλινδρόμησης. Μάλιστα, όπως θα δούμε και πιο κάτω, οι δύο αυτοί παράγοντες συσχετίζονται. Επίσης οι *συμμεταβλητές* $\boldsymbol{x}$ θεωρείται ότι επηρεάζουν την αναμενόμενη τιμή μ της $\boldsymbol{y}$, που ανήκει στην εκθετική οικογένεια κατανομών. Θεωρούμε ότι οι συμμεταβλητές συνδέονται γραμμικά, σχηματίζοντας την $\eta_x = \boldsymbol{x'\beta}$, που συχνά αποκαλείται και ως το **συστηματικό ή στοχαστικό μέρος** του μοντέλου, επειδή οι συμμεταβλητές $\boldsymbol{x}$ δεν είναι τυχαίες. Τότε η σχέση μεταξύ μ_x και της $\boldsymbol{x'\beta}$ γραμμικοποιείται μέσω της **συνάρτησης σύνδεσης (link function)** $g(.)$ ως εξής:

$$g(\mu_x) = \boldsymbol{x'\beta} = \eta_x \ .$$

Τώρα, επειδή η συνάρτηση σύνδεσης $g(.)$ είναι 1-1, μπορεί να αντιστραφεί και θα έχουμε το μοντέλο ως σχέση για την αναμενόμενη τιμή της :

$$E(y_x) = \mu_x = g^{-1}(\boldsymbol{x'\beta})$$

Παρόμοια διαδικασία όπως πιο πάνω (Οικονόμου & Καρώνη, 2010) θα χρησιμοποιήσουμε στη συνέχεια για να κατασκευάσουμε το μοντέλο της λογιστικής παλινδρόμησης.

2.2 ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

2.2.1 ΕΙΣΑΓΩΓΗ

Ειδική περίπτωση των γενικευμένων γραμμικών μοντέλων αποτελεί ένα χρήσιμο μοντέλο, αυτό της λογιστικής παλινδρόμησης. Το λογιστικό μοντέλο παλινδρόμησης άρχισε να χρησιμοποιείται ευρέως κατά τη δεκαετία του ’50, κυρίως με εφαρμογές στη βιοστατιστική. Έχουν ασχοληθεί πολλές στατιστικές μελέτες με το μοντέλο αυτό και έχει αναπτυχθεί εκτεταμένη βιβλιογραφία, λόγω της σπουδαιότητάς του. Χαρακτηριστικό αυτών των γενικευμένων γραμμικών μοντέλων είναι ότι η διακύμανση της απόκρισης, είναι συνάρτηση της μέσης αναμενόμενης τιμής της. Η λογιστική παλινδρόμηση μπορεί να χρησιμοποιηθεί στις περιπτώσεις όπου η μεταβλητή είναι δίτιμη, με άλλα λόγια η πρόβλεψη είναι το αποτέλεσμα μιας διαδικασίας Bernoulli, όπως επιτυχία/αποτυχία ή π.χ. σε ένα πείραμα, αν το φάρμακο ενεργεί ή όχι σε έναν ασθενή και σε άλλο πείραμα, αν ο ασθενής ζει ή απεβίωσε.

2.2.2 ΟΡΙΣΜΟΣ ΤΟΥ ΜΟΝΤΕΛΟΥ

Ας υποθέσουμε ότι έχουμε n ανεξάρτητες πειράματικες εκτελέσεις με δίτιμη απόκριση y (0 ή 1), η οποία εξαρτάται από ένα σύνολο μεταβλητών παλινδρόμησης $x_1, x_2, \ldots, x_n$. Οι μεταβλητές αυτές μπορεί να είναι για παράδειγμα το ύψος ή η ηλικία ενός ασθενή και η απόκριση να είναι το αν ενεργεί ένα φάρμακο ή όχι σε αυτόν, αν ο άνεργος βρίσκει εργασία ή όχι. Εάν ορίσουμε την τιμή $y = 1$ σαν επιτυχία με πιθανότητα p και την τιμή $y = 0$ σαν αποτυχία με πιθανότητα $1 - p$, τότε μπορούμε να πούμε ότι η y είναι τ.μ της κατανομής Bernoulli, με $E(y) = p$ και $V(y) = p(1 - p)$.

Επεκτείνουμε τώρα σε μια σειρά από n-δοκιμές. Ορίζουμε την τυχαία μεταβλητή y να είναι ο αριθμός επιτυχιών σε n-δοκιμές ($y = 0,1,2, \ldots, n$), υπό την υπόθεση ότι η πιθανότητα επιτυχίας p είναι ίδια σε κάθε δοκιμή και με τις δοκιμές να είναι ανεξάρτητες μεταξύ τους. Τότε ισχύει η Διωνυμική (binomial) κατανομή: $y \sim b(n, p)$ με συνάρτηση πιθανότητας: $f(y) = \binom{n}{y} p^y (1 - p)^{n-y}$, όπου η πιθανότητα επιτυχίας p είναι η παράμετρος της Διωνυμικής αυτής κατανομής, που είναι η κατάλληλη κατανομή προς περιγραφή της y σε τέτοια περίπτωση, με $E(y) = np$ και $V(y) = np(1 - p)$.

Σε πολλές περιπτώσεις που έχουμε δίτιμη μεταβλητή απόκρισης y (με τιμές 0 ή 1), μπορεί να εξαρτάται από επεξηγηματικές μεταβλητές $\boldsymbol{x}$ (ανεξάρτητες ή συμμεταβλητές). Η εξάρτηση αυτή εισάγεται μέσω της εξάρτησης της πιθανότητας επιτυχίας p από τις $\boldsymbol{x}$ (π.χ. η πιθανότητα να ενεργήσει σε κάποιο ασθενή ένα φάρμακο να εξαρτάται από τη θερμοκρασία, το φύλο, την ηλικία κλπ). Πιο συγκεκριμένα κατασκευάζεται το μοντέλο της **Λογιστικής Παλινδρόμησης**, μέσω της συνάρτησης σύνδεσης του (Οικονόμου & Καρώνη, 2010):

$$\eta_x = g\left(E(y_x)\right) = g(\mu_x) = \boldsymbol{x'\beta} \tag{2.1}$$

με την ακόλουθη δομή:

- $y_x \sim b(n_x, p_x)(n_x > 1,\ διωνυμικά\ δεδομένα)$
  ή $y_x \sim B(\mu_x)(n_x = 1, δυαδικά\ δεδομένα)$
- $\eta_x = g\ (\mu_x) = \ln\frac{\mu_x}{n_x - \mu_x} = \ln\frac{p_x}{1-p_x} = logit(p_x) = \boldsymbol{x'\beta}\ (συνάρτηση\ logit)$
- ανεξαρτησία μεταξύ των παρατηρήσεων y_x,

όπου n_x ο αριθμός των επαναλήψεων της τιμής του διανύσματος $\boldsymbol{x}$ των επεξηγηματικών μεταβλητών.

Τώρα, αν αντιστρέψουμε τη συνάρτηση σύνδεσης προκύπτει:

$$p_{\boldsymbol{x}} = e^{\eta_x}/(1 + e^{\eta_x}), \qquad με\ περιορισμό\ \ 0 < p_{\boldsymbol{x}} < 1 \tag{2.2}$$

Για κάθε i-παρατήρηση το μοντέλο γράφεται:

$$ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}, \qquad i = 1, \dots, n$$

με πιθανότητα επιτυχίας:

$$p_i = p_{x_i} = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})} \tag{2.3}$$

και άρα:

$$\mathrm{E}(y_i) = n_i p_i = n_i \frac{e^{\boldsymbol{x_i'\beta}}}{1 + e^{\boldsymbol{x_i'\beta}}} \ .$$

Η πιο πάνω συνάρτηση σύνδεσης που χρησιμοποιείται είναι η ***logit*** και αποτελεί την πιο συνηθισμένη επιλογή για δεδομένα που ακολουθούν τη Διωνυμική κατανομή. Άλλες συναρτήσεις σύνδεσης που χρησιμοποιούνται είναι:

- $g(\mu_x) = \ln[-\ln(1 - p_x)] = \boldsymbol{x'\beta}$ $(συνάρτηση\ complementary\ log - log)$
- $g(\mu_x) = \Phi^{-1}(p_x) = \boldsymbol{x'\beta}$ $(συνάρτηση\ probit)$.

2.2.3 ΠΑΡΑΔΕΙΓΜΑ ΜΟΝΤΕΛΟΥ ΛΟΓΙΣΤΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Με χρήση ενός απλουστευμένου μοντέλου θα εκτιμήσουμε την πιθανότητα να πεθάνει κάποιος ασθενής, σε διάστημα 10 χρόνων, από ένα καρδιακό νόσημα. Το μοντέλο περιλαμβάνει μόνο τρείς παράγοντες (επεξηγηματικές μεταβλητές), την ηλικία, το φύλο και το επίπεδο της χοληστερόλης στο αίμα του ασθενή. Οι παράμετροι του μοντέλου είναι οι εξής:

β_0 = -5.0, β_1 = +2.0, β_2 = -1.0, β_3 = +1.2

x_1 = η ηλικία σε χρόνια, πάνω από τα 50

x_2 = το φύλο (0: αρσενικό και 1: θηλυκό)

x_3 = το επίπεδο χοληστερόλης σε mmol/L, πάνω απο το 5.0

p = η πιθανότητα θανάτου του ασθενή (0: πέθανε και 1: ακόμα ζωντανός)

Μπορούμε να εκφράσουμε το μοντέλο ως εξής:

$$p = 1/(1 + e^{-z}) \ όπου\ z = -5.0 + 2.0x_1 - 1.0x_2 + 1.2x_3$$

Σε αυτό το μοντέλο, η πιθανότητα θανάτου αυξάνεται με την ηλικία (το z αυξάνεται κατά 2.0 για κάθε επιπλέον έτος πάνω από τα 50), μειώνεται εάν έχουμε θηλυκό ασθενή (το z

μειώνεται κατά 1.0 στην περίπτωση θηλυκού ασθενή) και αυξάνεται με αυξημένο επίπεδο της χοληστερόλης (το z αυξάνεται κατά 1.2 για κάθε 1 mmol/L παραπάνω από τα 5 mmol/L). Θα χρησιμοποιήσουμε το παραπάνω μοντέλο σε έναν άντρα ασθενή 50 χρονών με επίπεδο χοληστερόλης 7.0 mmol/L.

Η πιθανότητα να πεθάνει είναι:

$$p = 1/(1 + e^{-z})$$

$$\text{όπου } z = -5.0 + 2.0(50 - 50) - (1.0)0 + 1.2(7.0 - 5.0) => z = -2.6$$

Από το αποτέλεσμα υπολογίζουμε ότι η πιθανότητα να πεθάνει ο συγκεκριμένος ασθενής από καρδιακό νόσημα, σε διάστημα 10 χρόνων, είναι 0.069 ή αλλιώς 7%.

2.2.4 ΕΚΤΙΜΗΣΗ ΤΩΝ ΣΥΝΤΕΛΕΣΤΩΝ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Η λογιστική παλινδρόμηση χρησιμοποιείται για την εξαγωγή συμπερασμάτων σε πολλές διαφορετικές περιπτώσεις, όπως για παράδειγμα σε κλινικές δοκιμές όπου πρέπει να συγκρίνουμε τα αποτελέσματα διαφορετικών θεραπειών των οποίων το αποτέλεσμα έχει δυαδική μορφή. Για τη βελτίωση του μοντέλου δοκιμάζεται η σημασία της κάθε μεταβλητής.

Λόγω της λογιστικής παλινδρόμησης, έχουμε τη δυνατότητα ερμηνείας των τιμών των συντελεστών $\hat{\beta}$, αλλά και των διαστημάτων εμπιστοσύνης τους. Εφόσον εκτιμηθούν οι συντελεστές αυτοί, η σχέση μεταξύ της προσαρμοσμένης πιθανότητας απόκρισης $\hat{p}$ και των τιμών των $x_0,\ x_1,\ x_2$ (επεξηγηματικών μεταβλητών), εκφράζεται ως:

$$\hat{p} = \frac{e^{\boldsymbol{x'\hat{\beta}}}}{1 + e^{\boldsymbol{x'\hat{\beta}}}}$$

ή ισοδύναμα μέσω του **λόγου των συμπληρωματικών** ή **σχετικών πιθανοτήτων (odds)**, χρησιμοποιώντας τη συνάρτηση σύνδεσης $logit(p)$:

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \boldsymbol{x'\hat{\beta}} => \frac{\hat{p}}{1-\hat{p}} = e^{\boldsymbol{x'\hat{\beta}}} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k\right), \qquad \text{όπου } x_0 \equiv 1$$

Από το *odds* προκύπτει ότι αν ο εκτιμημένος συντελεστής $\widehat{\beta_j}$ είναι θετικός, ο παράγοντας $e^{\widehat{\beta_j}} > 1$, γεγονός που σημαίνει πως το odds αυξάνεται με την αύξηση της x_j. Αντίθετα αν ο $\widehat{\beta_j}$ είναι αρνητικός, ο παράγοντας $e^{\widehat{\beta_j}} < 1$ και η σχετική πιθανότητα μειώνεται με την αύξηση της x_j .

Οι παράμετροι της παλινδρόμησης μπορούν να εκφραστούν και μέσα από το **λόγο του λόγου των συμπληρωματικών πιθανοτήτων**, δηλαδή μέσα από το **λόγο των odds (odds ratio)**.

Παράδειγμα:

Ας υποθέσουμε για παράδειγμα, ένα μοντέλο με δύο συμμεταβλητές x_1 και x_2. Η μεταβλητή x_1 είναι κατηγορική και ότι μια ομάδα από τα πειραματικά μας υποκείμενα μπορούν να χωριστούν σε αυτά που τους χορηγήθηκε μια δόση από βιταμίνη C ($x_1 = 0$) και σε αυτά που δεν τους χορηγήθηκε τίποτα ($x_1 = 1$). Επίσης, έχουμε τη x_2 ως ποσοτική μεταβλητή (π.χ. θερμοκρασία). Οπότε και η απόκριση y, που ισούται με την πιθανότητα να έχει μολυνθεί το αναπνευστικό σύστημα ή όχι, παίρνει τις τιμές $y = 1$ και $y = 0$, αντίστοιχα. Αν χρησιμοποιήσουμε την πιο πάνω σχέση θα έχουμε, για το παράδειγμα:

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Τώρα, αν θεωρήσουμε ένα υποκείμενο στο οποίο χορηγείται η βιταμίνη C ($x_1 = 0$), τότε η $\exp(\beta_0)$ μπορεί να μεταφραστεί σαν το λόγο συχνοτήτων για τα υποκείμενα που μολύνθηκαν προς αυτά που δε μολύνθηκαν, για όλο τον πληθυσμό. Όσον αφορά αυτά που δε χορηγήθηκε βιταμίνη C ($x_1 = 1$), θα έχουμε όπως πιο πάνω αλλά με $x_1 = 1$.

Χρησιμοποιώντας την πιο πάνω ερμηνεία $\hat{\beta}_0$ βρίσκουμε την ερμηνεία για το $\hat{\beta}_1$. Για την ομάδα που δε δέχτηκε τη θεραπεία ισχύει:

$$\frac{\frac{\hat{p}_1}{1-\hat{p}_1}}{\frac{\hat{p}_2}{1-\hat{p}_2}} = \frac{odds(y=1|x_1=1,x_2)}{odds(y=0|x_1=0,x_2)} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 x_2)}{\exp(\hat{\beta}_0 + \hat{\beta}_2 x_2)} = e^{\widehat{\beta_1}}$$

Άρα, η ποσότητα $\exp(\hat{\beta}_1)$ μπορεί να ερμηνευτεί σαν το λόγο των συχνοτήτων της ομάδας που δε δέχτηκε θεραπεία, σε σχέση με αυτή που δέχτηκε και είναι ανεξάρτητη της x_2. Προφανώς, ένας ερευνητής ερμηνεύει μια τιμή $\hat{\beta}_0 \ll 0$, όπως επίσης και μια τιμή $\hat{\beta}_1 \gg 0$ να είναι ευνοϊκή (στατιστικά σημαντική) για τη θεραπεία.

2.2.5 ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ

Στο μοντέλο της λογιστικής παλινδρόμησης, οι εκτιμήσεις των συντελεστών ακολουθούν προσεγγιστικά μια κανονική κατανομή, όταν υπάρχει ένας επαρκής αριθμός δεδομένων στο δείγμα. Έτσι μπορούμε να διεξάγουμε ελέγχους υποθέσεων με βάση τις εκτιμήσεις των συντελεστών και τα τυπικά τους σφάλματα, για κάθε μοντέλο που προσαρμόζουμε.

2.2.5.1 Έλεγχος με τη μέθοδο της Μέγιστης Πιθανοφάνειας

Όπως σε όλα τα γενικευμένα γραμμικά μοντέλα, η προσαρμογή του μοντέλου στα δεδομένα γίνεται με τη μέθοδο της μέγιστης πιθανοφάνειας. Ας υποθέσουμε ότι τα δεδομένα μας είναι χωρισμένα σε κατηγορίες. Δηλαδή, έχουμε n_i το πλήθος πειραματικές μονάδες στο i – οστό σημείο δεδομένων (για παράδειγμα, μπορούμε να θεωρήσουμε ότι το n_i είναι το πλήθος των πειραματόζωων στα οποία έχουμε παρέχει μια συγκεκριμένη δοσολογία φαρμάκου). Άρα έχουμε το δείγμα με τιμές $y_1, y_2, \dots, y_m$, με μέσες τιμές $E(y_i) = \mu_i = n_i p_i$, $V(y_i) = n_i p_i (1 - p_i)$ και συμμεταβλητές $\boldsymbol{x_i}' = (x_{i0}, x_{i1}, \dots, x_{ik})$, όπου n_i ο αριθμός δοκιμών της στατιστικής μονάδας-i και $\sum_{i=1}^{m} n_i = n$, p_i η αντίστοιχη πιθανότητα επιτυχίας και $x_{i0} = 1$.

Συνεπώς, υποθέτοντας κανονική συσχέτιση $n_i = g(\mu_i) = \boldsymbol{x_i}'\boldsymbol{\beta}$ στα glm, η συνάρτηση πιθανοφάνειας του δείγματος γράφεται (Οικονόμου & Καρώνη, 2010):

$$L(\beta) = \prod_{i=1}^{n} \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i} \qquad (2.4)$$

Η πιθανοφάνεια εξαρτάται από τις άγνωστες πιθανότητες επιτυχίας p_i, οι οποίες με τη σειρά τους εξαρτώνται από τα $\boldsymbol{\beta}$ μέσω της εξίσωσης 2.2.

Έτσι η συνάρτηση πιθανοφάνειας μπορεί να θεωρηθεί ως συνάρτηση των $\boldsymbol{\beta}$ με:

$$l = \ln L(\boldsymbol{\beta}) \ = \sum_{i=1}^{n} \left\{ ln \binom{n_i}{y_i} + y_i x_i{}' \boldsymbol{\beta} - n_i \ln(1 + e^{x'_i \boldsymbol{\beta}}) \right\} \tag{2.5}$$

Παραγωγίζοντας έχουμε:

$$\frac{\partial \ln L(\beta)}{\partial \beta_j} \ = \sum_{i=1}^{n} (y_i - n_i p_i) x_{ij} \tag{2.6}$$

Οι εκτιμήτριες μέγιστης πιθανοφάνειας των β_j προκύπτουν με την ικανοποίηση των εξισώσεων:

$$\sum_{i=1}^{n} (y_i - n_i \hat{p}_i) x_{ij} = \sum_{i=1}^{n} (y_i - \hat{\mu}_\iota) x_{ij} = 0, \quad j = 0,1, \dots, k$$

$$=> \boldsymbol{X}' = (\boldsymbol{y} - \widehat{\boldsymbol{\mu}}) = \boldsymbol{0}$$

όπου $\widehat{\boldsymbol{\mu}}' = (\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_n)$ με $ln\hat{\mu}_i = \boldsymbol{x_i}'\widehat{\boldsymbol{\beta}}$. Η εξίσωση είναι μη γραμμική ως προς τα $\widehat{\boldsymbol{\beta}}$, επειδή $ln\hat{\mu}_i = \exp(\boldsymbol{x_i}'\widehat{\boldsymbol{\beta}})$, επομένως για να λύσουμε το σύστημα θέλουμε μια επαναληπτική διαδικασία. Μια τέτοια διαδικασία είναι αυτή των σταθμισμένων ελαχίστων τετραγώνων. Έτσι μπορούμε να υπολογίσουμε τις εκτιμήτριες $\widehat{\boldsymbol{\beta}}$ για να καταλήξουμε στα συμπεράσματα που επιθυμούμε, διότι οι ποσότητες $\exp(\hat{\beta}_j)$ εκφράζουν την αναμενόμενη πολλαπλασιαστική μεταβολή της y για μία μονάδα αύξησης της αντίστοιχης x_j, κρατώντας τις υπόλοιπες συμμεταβλητές σταθερές.

Τώρα χρησιμοποιώντας τη συνάρτηση πιθανοφάνειας (log likelihood), μπορούμε να ενισχύσουμε τους ελέγχους υποθέσεων και να βγάλουμε συμπεράσματα για τη σημαντικότητα των εκτιμητριών μας.

Στη λογιστική παλινδρόμηση ισχύει ασυμπτωτικά (Montgomery, Peck, & Vining, 2006):

$$-2\ln\frac{L(reduced)}{L(full)} = -2\{l(\ \boldsymbol{\beta}*\ ) - l(\boldsymbol{\beta})\} \sim X_d^2 \qquad (2.7)$$

όπου $L(.)$: συνάρτηση πιθανοφάνειας (του ελαττωμένου και του πλήρες μοντέλου),

d: η διαφορά του πλήθους των παραμέτρων του ελαττωμένου από του πλήρες μοντέλου.

Παράδειγμα:

Έχουμε το μοντέλο $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ και θέλουμε να εξετάσουμε την υπόθεση $H_0: \beta_1, \beta_2 = 0$. Οπότε θα έχουμε τον πιο κάτω έλεγχο:

$$2\{\ln(L(\beta_0, \beta_1, \beta_2, \beta_3) - \ln(L(\beta_0^*, \beta_3^*)\} \sim X_2^2$$

όπου $L(\beta_0^*, \beta_3^*)$ το μοντέλο για το οποίο έχουμε επικαλεστεί την πιο πάνω μηδενική υπόθεση. Έτσι αν η λογαριθμημένη πιθανοφάνεια αυξηθεί σημαντικά, έχουμε σοβαρές ενδείξεις για να απορρίψουμε τη μηδενική υπόθεση και να συμπεριλάβουμε όλες τις μεταβλητές στο μοντέλο, αφού τις θεωρούμε στατιστικά σημαντικές.

Άρα από τα πιο πάνω καταλήγουμε στο συμπέρασμα ότι η μέθοδος της μέγιστης πιθανοφάνειας είναι πολύ χρήσιμη για ελέγχους υποθέσεων που θα μας βοηθήσουν στην επιλογή του κατάλληλου μοντέλου στη λογιστική παλινδρόμηση.

2.2.5.2 Ελεγχοσυνάρτηση DEVIANCE

Μια σημαντική τεχνική που χρησιμοποιείται για τη σύγκριση και ανάπτυξη των στατιστικών μοντέλων είναι η ελεγχοσυνάτηση **Deviance** ή **Απόκλιση.** Κατ' αρχάς, ένα μοντέλο λέγεται **κορεσμένο** ή **πλήρες** όταν έχει τόσες παραμέτρους όσες είναι και οι παρατηρήσεις του. Έχουμε δύο μοντέλα M και M* με εκτιμήτριες $\widehat{\boldsymbol{\beta}}$ και $\widehat{\boldsymbol{\beta}}$* αντίστοιχα, όπου M*⊂M με όλες τις μεταβλητές του M* να περιέχονται στο M, δηλαδή από το M απορρίψαμε κάποιες μη στατιστικά σημαντικές μεταβλητές.

Όπως έχουμε αναφέρει και προηγουμένως, για να συγκρίνουμε τα μοντέλα αυτά χρησιμοποιουμε τον έλεγχο με βάση το λόγο των μεγιστοποιημένων πιθανοφανειών (εξίσωση 2.7): $-2\{l(\ \widehat{\boldsymbol{\beta}}*\ ) - l(\widehat{\boldsymbol{\beta}})\} \sim X_d^2$

με *H_0: ισχύει το μοντέλο M** έναντι της εναλλακτικής *H_1: ισχύει το μοντέλο M*.

Αν γράψουμε τώρα την εναλλάκτική υπόθεση H_1 ως *H_S: το κορεσμένο μοντέλο με $p_1 = n$ (αριθμό παραμέτρων ίσο με τον αριθμό των παρατηρήσεων)* και την *H_0: το υποψήφιο μοντέλο με $p_0 = p < n$*, τότε οι προβλεπόμενες τιμές $\tilde{\mu}_i$ ισούνται με τις παρατηρούμενες y_i. Έτσι έχουμε τη deviance με $D_1 = -2(\hat{l}_1 - \hat{l}_S)$ και $D_2 = -2(\hat{l}_0 - \hat{l}_S)$ ως:

$$D_0 - D_1 = (\hat{l}_0 - \hat{l}_1) \sim X_d^2 \quad \alpha\sigma\upsilon\mu\pi\tau\omega\tau\iota\kappa\acute{\alpha} \tag{2.8}$$

όπου $l(.)$: η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας (του ελαττωμένου και του πλήρες μοντέλου) και d: η διαφορά του πλήθους των παραμέτρων του ελαττωμένου από του πλήρες μοντέλου.

Τώρα για το μοντέλο της λογιστικής παλινδρόμησης, μέσω της $ln\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x_i}'\boldsymbol{\beta}$ επιβάλλεται μία δομή στα δεδομένα, για να μην έχουμε απλά ανεξάρτητες τιμές από διαφορετικές Διωνυμικές κατανομές $y_i \sim b(n_i, p_i)$ με $\mu_i = n_i p_i$ και θα εκτιμάται από το $\tilde{\mu}_i = y_i$ (Οικονόμου & Καρώνη, 2010).

Υπό την *H_S: το κορεσμένο μοντέλο* θα έχουμε ($\tilde{p}_i = \frac{y_i}{n_i}$):

$$\hat{l}_{iS} = ln\binom{n_i}{y_i} + y_i ln\tilde{p}_i + (n_i - y_i)\ln(1 - \tilde{p}_i)$$

Υπό την *H_0: το υποψήφιο μοντέλο* θα έχουμε ($\hat{p}_i = \frac{\hat{\mu}_i}{n_i}$):

$$\hat{l}_{i0} = ln\binom{n_i}{y_i} + y_i ln\hat{p}_i + (n_i - y_i)\ln(1 - \hat{p}_i)$$

Η διαφορά τους:

$$\hat{l}_{i0} - \hat{l}_{iS} = y_i ln\left(\frac{\hat{\mu}_i}{y_i}\right) + (n_i - y_i)\ln(\frac{n_i - \hat{\mu}_i}{n_i - y_i})$$

Η τυποποιημένη ελεγχοσυνάρτηση Deviance ορίζεται ως:

$$D(\widehat{\boldsymbol{\beta}}) = -2(\hat{l}_0 - \hat{l}_S) = 2\sum_{i=1}^{n}\{y_i ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (n_i - y_i)\ln\left(\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right)\} = 2\sum_{i=1}^{n} d_i(\widehat{\boldsymbol{\beta}}) \quad (2.9)$$

Αυτή η συνάρτηση είναι πολύ χρήσιμη διότι συγκρίνει τις παρατηρήσεις y_i και τα εκτιμημένα $\hat{\mu}_i$. Έτσι τώρα, μπορούμε να συγκρίνουμε τα δύο μοντέλα χρησιμοποιώντας την κατανομή X^2 ασυμπτωτικά. Για μεγάλες τιμές της διαφοράς $D_0 - D_1$, απορρίπτουμε την H_0. Εννοείται ότι η διαφορά είναι πάντα θετική, διότι το κορεσμένο (εμπεριέχει όλες τις μεταβλητές) έχει περισσότερες μεταβλητές από το υποψήφιο μοντέλο. Οπότε για να επιλέξουμε το βέλτιστο μοντέλο με τον πιο πάνω έλεγχο, αρκεί να επιλέξουμε το μοντέλο με τη μεγαλύτερη απόκλιση (Deviance).

Μικρή παρατήρηση είναι το γεγονός ότι για δυαδικά δεδομένα, δηλαδή η απόκριση $\boldsymbol{y}$ να παίρνει τιμές 0 και 1, με την ελεγχοσυνάρτηση Deviance δε μπορούμε να αποφανθούμε για την καταλληλότητα ενός μοντέλου. Όμως η συνάρτησή της χρησιμοποιείται και σαν εναλλαχτική περίπτωση στο κριτήριο BIC για την επιλογή κατάλληλου μοντέλου. Οπότε δε μπορούμε να πούμε ότι δε μας χρειάζεται αυτός ο έλεγχος στα γενικευμένα γραμμικά μοντέλα.

2.2.5.3 Έλεγχος με τη μέθοδο του WALD

Η πρώτη εφαρμογή της μεθόδου Wald έχει να κάνει με έλεγχο υποθέσεων για κάθε ξεχωριστό συντελεστή του μοντέλου της λογιστικής παλινδρόμησης. Πιο συγκεκριμένα, θέλουμε να ελέγξουμε: H_0: $\beta_j = 0$, H_1: $\beta_j \neq 0$, *με* $j = 1, \dots, k$ και με το $\hat{\beta}_j$ να εμφανίζεται στη γραμμική πρόβλεψη $\boldsymbol{x}_i{}'\widehat{\boldsymbol{\beta}}$ του λογιστικού μοντέλου.

Με τη μέθοδο της μέγιστης πιθανοφάνειας (§ 2.2.5.1) υπολογίσαμε τους συντελεστές $\hat{\beta}_j$, με εκτιμημένες διασπορές $\hat{V}(\hat{\beta}_j)$ που κάθε μία είναι το j-οστό διαγώνιο στοιχείο του πίνακα $\boldsymbol{I}^{-1}(\widehat{\boldsymbol{\beta}})$, με αντίστοιχο $se(\hat{\beta}_j) = \{\hat{V}(\hat{\beta}_j)\}^{1/2} = \{\boldsymbol{I}^{-1}(\widehat{\boldsymbol{\beta}})_{jj}\}^{1/2}$.

Επομένως για έναν εκτιμητή μέγιστης πιθανοφάνειας ισχύει ότι:

$$z_j = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \tag{2.10}$$

οποίος ακολουθεί ασυμπτωτικά την τυπική κανονική κατανομή N(0,1) και έτσι ισχύει για τη μηδενική υπόθεση:

$$z_j{}^2 = \left(\frac{\hat{\beta}_j}{se(\hat{\beta}_j)}\right)^2 \tag{2.11}$$

που ακολουθεί ασυμπτωτικά την X^2-κατανομή. Οι τιμές των ελεγχοσυναρτήσεων υπολογίζονται για κάθε συντελεστή των μεταβλητών και αφορούν την προσαρμογή του μοντέλου. Έτσι ελέγχουμε αν απορρίπτεται η H_0 έναντι της H_1. Η μηδενική υπόθεση απορρίπτεται αν η p-τιμή του πιο πάνω ελέγχου είναι πολύ μικρή που σημαίνει ότι η αντίστοιχη μεταβλητή x_j θα είναι στατιστικά σημαντική για την πρόβλεψη και θα πρέπει να περιληφθεί στο μοντέλο. Αν δεν απορρίπτεται η μηδενική υπόθεση, με βάση τον έλεγχο, σημαίνει ότι έχουμε σημαντικές ενδείξεις ότι η συγκεκριμένη μεταβλητή δεν είναι στατιστικά σημαντική για το μοντέλο. Επίσης σημαντική προϋπόθεση για να μην πέσουμε έξω στους ελέγχους αυτούς, είναι να ελέγξουμε πρώτα και τη συσχέτιση μεταξύ των μεταβλητών, αλλά και τη συσχέτιση της κάθε μεταβλητής με την απόκριση.

Μια δεύτερη μορφή της Wald συμπερασματολογίας έχει να κάνει με τον υπολογισμό του διαστήματος εμπιστοσύνης της διωνυμικής πιθανότητας για κάποια δοσμένα ή αυθαίρετα δεδομένα. Λόγω της ύπαρξης της γραμμικής πρόβλεψης $\boldsymbol{x'\widehat{\beta}}$ στο λογιστικό μοντέλο, ακολουθείται μια εναλλακτική διαδικασία υπολογισμού των διαστημάτων εμπιστοσύνης. Μπορούμε να ορίσουμε ένα $100(1-a)\%$ διάστημα εμπιστοσύνης (δ.ε.) στο $p = 1/\left(1 + e^{-\boldsymbol{x'\beta}}\right)$, όπου $p = p(x_i) = y_i$ χρησιμοποιώντας ένα διάστημα εμπιστοσύνης στο $\boldsymbol{x'\beta}$. Η γραμμική πρόβλεψη περιλαμβάνει προφανώς όρους που είναι γραμμικοί στο $\boldsymbol{\beta}$ και μπορούμε να εκμεταλλευτούμε το γεγονός ότι οι εκτιμήτριες $\widehat{\boldsymbol{\beta}}$, είναι ασυμπτωτικά κανονικές. Άρα, ένα άνω διάστημα εμπιστοσύνης για το $\boldsymbol{x'\beta}$, παράγει ένα άνω διάστημα εμπιστοσύνης για το p. Έτσι μετά από πράξεις, καταλήγουμε στο ότι μπορούμε να

κατασκευάσουμε ένα $100\ (1-a)\%$ δ.ε. για την παράμετρο β_j, ως $\hat{\beta}_j \pm z_{\alpha/2} se(\widehat{\beta_j})$, με $j = 1, 2, \dots, k$. Από το δ.ε. αυτό μπορούμε να προσδιορίσουμε και ένα $100(1-a)\%$ δ.ε. $exp[\widehat{\beta_j} \pm z_{a/2} se(\widehat{\beta_j})]$, $j = 1,2, \dots, k$ για το λόγο των odds (odds ratio). Με τα διαστήματα εμπιστοσύνης αυτά παίρνουμε ακόμα περισσότερες πληροφορίες για τις εκτιμήτριες, για να καταλήξουμε σε ακόμα καλύτερα συμπεράσματα.

2.2.5.4 Έλεγχος με τη μέθοδο LASSO στη λογιστική παλινδρόμηση

Τώρα, αν παρατηρείται το φαινόμενο της πολυσυγγραμμικότητας, αξίζει να χρησιμοποιήσουμε άλλη μία πολύ σημαντική μέθοδο επιλογής, τη **LASSO**. Αν η συσχέτιση μεταξύ των μεταβλητών μας είναι αρκετά ψηλή, δε μπορούμε να αρκεστούμε στα αποτελέσματα των τεχνικών και των κριτηρίων που αναφέρθηκαν πριν για τη λογιστική παλινδρόμηση. Αυτή η τεχνική συρρίκνωσης, για την οποία μιλήσαμε αναλυτικά στην παράγραφο 1.2.4.3 μπορεί να εφαρμοστεί και στα γενικευμένα γραμμικά μοντέλα, συγκεκριμένα στη λογιστική παλινδρόμηση.

Όπως έχουμε αναφέρει η μέθοδος έχει ως σκοπό την ελαχιστοποίηση του λογαρίθμου της μερικής πιθανοφάνειας, να συρρικνώνει τους συντελεστές προς το μηδέν και να ρυθμίζει αυτόματα πολλούς από αυτούς ακριβώς στο μηδέν, τακτοποιώντας τους με τέτοιο τρόπο, ανάλογα της χρήσης τους στο μοντέλο. Έτσι μειώνει την εκτίμηση της διακύμανσης, ενώ παρέχει ένα ερμηνεύσιμο τελικό μοντέλο.

Για να εφαρμοστεί η Lasso που καθορίζει τις εκτιμήσεις-lasso $\widehat{\boldsymbol{\beta}}$ πρέπει να μεγιστοποιήσουμε αυτή τη φορά, την περιορισμένη πιθανοφάνεια (constrained likelihood), όπως πιο κάτω (Tibshirani, 1997):

$$\widehat{\boldsymbol{\beta}} = argmaxl(\boldsymbol{\beta}), \qquad subject\ to \qquad \| \boldsymbol{\beta} \|_1 = \sum |\beta_j| \leq s$$

Μπορεί να χρησιμοποιηθεί και η σχέση που καθορίζει τα $\widehat{\boldsymbol{\beta}}$ για βελτιστοποίηση της ποινικοποιημένης πιθανοφάνειας (penalized likelihood) ως:

$$\widehat{\boldsymbol{\beta}} = argmax\{l(\boldsymbol{\beta}) - \lambda \| \boldsymbol{\beta} \|_1\}$$

Στη λογιστική παλινδρόμηση η εξίσωση αυτή, παίρνει την πιο κάτω μορφή:

$$\widehat{\boldsymbol{\beta}} = argmax\left\{\sum_{i=1}^{n}\{ln\binom{n_i}{y_i} + y_i x_i{}'\boldsymbol{\beta} - n_i \ln(1 + e^{x'_i\boldsymbol{\beta}})\} - \lambda \parallel \boldsymbol{\beta} \parallel_1\right\} \qquad (2.12)$$

Αυτή η μεγιστοποίηση είναι δυνατό να διεξαχθεί από μια γενική μη τετραγωνική προγραμματιστική διαδικασία, όμως αντί αυτού θεωρούμε ότι εδώ τα μοντέλα για τα οποία η τετραγωνική προσέγγιση της $l(\boldsymbol{\beta})$ οδηγεί σε επαναληπτική διαδικασία επανασταθμισμένων (reweighted) ελαχίστων τετραγώνων (IRLS) για τον υπολογισμό των $\boldsymbol{\beta}$. Μία τέτοια διαδικασία μπορεί να είναι ο αλγόριθμος Newton Raphson. Όμως μπορούμε να λύσουμε το πρόβλημα με περιορισμένη εφαρμογή του αλγόριθμου Lasso, όπως αναφέραμε και στην παράγραφο 1.2.4.3, εντός μιας IRLS επανάληψης (loop). Σύγκλιση στην πιο πάνω διαδικασία δεν είναι εξασφαλισμένη.

Επειδή στην εφαρμογή στα γενικευμένα γραμμικά μοντέλα δε θα ασχοληθούμε πολύ με τη μέθοδο αυτή, δε θα αναπτύξουμε περεταίρω. Θα ασχοληθούμε όμως πολύ περισσότερο στα μοντέλα επιβίωσης και συγκεκριμένα πώς βοηθά στην επιλογή του βέλτιστου μοντέλου στο μοντέλο του Cox (§ 3.2.6.3).

2.2.6 ΚΡΙΤΗΡΙΑ ΕΠΙΛΟΓΗΣ ΚΑΤΑΛΛΗΛΟΥ ΜΟΝΤΕΛΟΥ

Όπως σε όλα τα γενικευμένα γραμμικά μοντέλα, έτσι και στο μοντέλο της Λογιστικής παλινδρόμησης χρησιμοποιούνται τα κριτήρια επιλογής κατάλληλου μοντέλου όπως ο συντελεστής προσδιορισμού R^2, αλλά και τα κριτήρια AIC και BIC.

2.2.6.1 Συντελεστής προσδιορισμού R^2 στη λογιστική παλινδρόμηση

Στα γενικευμένα γραμμικά μοντέλα, χρησιμοποιείται ένας δείκτης προσδιορισμού *ψευδο-*R^2, που αποτελεί τροποποίηση του *ψευδο-*R^2 του McFadden που έχει αναφερθεί πριν (§ 1.2.3.1) και βασίζεται στα υπόλοιπα Deviance:

$$R_D^2 = \frac{l(\hat{\beta}) - \widehat{l_0}}{\widetilde{l_S} - \widehat{l_0}} = \frac{D_0 - D_M}{D_M}\ , \qquad 0 \le R_D^2 \le 1,$$

όπου $\widetilde{l_S}$*: η μεγιστοποιημένη λογαριθμοποιημένη πιθανοφάνεια του κορεσμένου μοντέλου,*

D_0*: η τιμή της Deviance του μοντέλου με το σταθερό όρο,*

D_M*: η τιμή της Deviance του μοντέλου με τις επεξηγηματικές μεταβλητές.*

Αποδεικνύεται ότι (Cameron & Windmeijer, 1996):

$$R_L^2 = \left\{1 - \frac{\tilde{l}_S}{\widehat{l_0}}\right\} R_D^2$$

Ένας άλλος τροποποιημένος δείκτης για τα γενικευμένα γραμμικά μοντέλα είναι και ο *ψευδο-*R_M^2 (Maddala, 1983; Cox & Snell, 1989; Magee, 1990) με τύπο:

$$R_M^2 = 1 - \left(\frac{\hat{L}_0}{\hat{L}_1}\right)^{2/n}$$

ο οποίος μπορεί να χρησιμοποιηθεί σε όλα τα προσαρμοσμένα μοντέλα με τη μέθοδο της μέγιστης πιθανοφάνειας. Αυτός ο δείκτης προτάθηκε για τη λογιστική παλινδρόμηση από τον Maddala (1983).

Τέλος, για τα διακριτά μοντέλα που εξετάζουμε, ιδιαίτερα αυτά της Διωνυμικής κατανομής ισχύει ότι: $\hat{L}_0 < \hat{L}_1 < 1$ (διότι η πιθανοφάνεια αποτελείται από γινόμενο πιθανοτήτων) επομένως το R_M^2 θα είναι κάτω απ' τη μονάδα. Σε τέτοιες περιπτώσεις θα ισχύει: $max\ R_M^2 = 1 - \hat{L}_0^{\frac{2}{n}} < 1$. Οπότε, για τα διακριτά μοντέλα προτάθηκε ο προσαρμοσμένος δείκτης (Nagelkerke, 1991):

$$R_N^2 = \frac{R_M^2}{max\ R_M^2}$$

Όλες αυτές οι προτάσεις για αυτά τα *ψευδο* - R^2 μέτρα προσδιορισμού, συχνά παρουσιάζουν χαμηλές τιμές στη λογιστική παλινδρόμηση, ιδιαίτερα στην περίπτωση των δυαδικών δεδομένων. Αυτό συμβαίνει διότι το μοντέλο προβλέπει μόνο την

πιθανότητα επιτυχίας $p = E(Y)$ και όχι τις ατομικές τιμές της y (0 ή 1). Δεδομένης της p, το μοντέλο δεν μπορεί να προβλέψει την επιτυχία ή την αποτυχία, επειδή αυτές είναι τυχαία γεγονότα. Επομένως, μεγάλο μέρος της συνολικής μεταβλητότητας των δεδομένων δε μπορεί να εξηγηθεί και έτσι οι δείκτες R^2 θα παίρνουν χαμηλές τιμές. Άρα κανένα από τα προτεινόμενα μέτρα δε θεωρείται γενικώς ικανοποιητικό. Αυτό φαίνεται και από τις συγκριτικές μελέτες οι οποίες δεν έχουν καταλήξει σε ένα ικανοποιητικό μέτρο, γι' αυτό και οι χρήστες των γενικευμένων γραμμικών μοντέλων, δε θεωρούν απαραίτητη την παρουσίαση ενός δείκτη R^2 στα αποτελέσματά τους.

2.2.6.2 Κριτήριο AIC στη λογιστική παλινδρόμηση

Ένα άλλο σημαντικό κριτήριο επιλογής βέλτιστου μοντέλου είναι το κριτήριο AIC, για το οποίο έχουμε μιλήσει στην παράγραφο 1.2.3.2. Στην περίπτωση της λογιστικής παλινδρόμησης το κριτήριο παίρνει την πιο κάτω μορφή:

$$AIC \ = 2\sum_{i=1}^{n}\left\{-ln\binom{n_i}{y_i} - y_i \boldsymbol{x}_i{}'\widehat{\boldsymbol{\beta}} + n_i \ln(1 + e^{\boldsymbol{x}_i{}'\widehat{\boldsymbol{\beta}}})\right\} \qquad (2.13)$$

Όπως έχουμε σημειώσει και προηγουμένως για το κριτήριο αυτό, συγκρίνοντας όλα τα υποψήφια μοντέλα, το μοντέλο με το μικρότερο AIC, θεωρείται ως το καταλληλότερο και αυτό επιλέγουμε.

2.2.6.3 Κριτήριο BIC στη λογιστική παλινδρόμηση

Παρόμοια θεωρία ισχύει και για το κριτήριο BIC, για το οποίο έχουμε μιλήσει στην παράγραφο 1.2.3.3. Στην περίπτωση της λογιστικής παλινδρόμησης το κριτήριο παίρνει την πιο κάτω μορφή:

$$BIC \ = 2\sum_{i=1}^{n}\left\{-ln\binom{n_i}{y_i} - y_i \boldsymbol{x}_i{}'\widehat{\boldsymbol{\beta}} + n_i \ln(1 + e^{\boldsymbol{x}_i{}'\widehat{\boldsymbol{\beta}}})\right\} + p \ln n \qquad (2.14)$$

2.3 ΕΦΑΡΜΟΓΗ ΣΤΗ ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ ΜΕ ΧΡΗΣΗ ΤΗΣ R

2.3.1 ΠΕΡΙΓΡΑΦΗ ΕΦΑΡΜΟΓΗΣ ΚΑΙ ΔΕΔΟΜΕΝΑ

Στην εφαρμογή αυτή χρησιμοποιούμε στοιχεία μιας ιατρικής μονάδας με ασθενείς που πάσχουν από *οξεία μυελοπλαστική λευχαιμία* (*acute myeloblastic leukaemia*) (Lee, 1980). Έχουμε δείγμα από 51 ασθενείς με αυτή την αρρώστια, η οποία μπορεί ακόμα να επιφέρει και το θάνατο. Εφαρμόζουμε μια θεραπεία με διαφορετικές συνθήκες σε κάθε ασθενή επηρεασμένη από την ηλικία, τη θερμοκρασία, τις διάφορες δοσοληψίες μέσα στο φάρμακο κλπ. Η περιγραφή των επεξηγηματικών αυτών μεταβλητών φαίνεται πιο κάτω (Πίνακας 2.1).

Η μεταβλητή απόκρισης δείχνει το αν ο ασθενής αντέδρασε στην θεραπεία ή όχι (Αντίδραση: 1 = αντέδρασε στη θεραπεία, 0 = δεν αντέδρασε). Έτσι βλέπουμε ότι η μεταβλητή απόκρισης y του προβλήματος αυτού είναι διωνυμική. Αυτός είναι και ο λόγος που προσαρμόζουμε το πρόβλημα, με τη βοήθεια του γενικευμένου γραμμικού μοντέλου (glm) και συγκεκριμένα του μοντέλου της λογιστικής παλινδρόμησης, διότι το γενικό γραμμικό μοντέλο αδυνατεί στην περίπτωση αυτή.

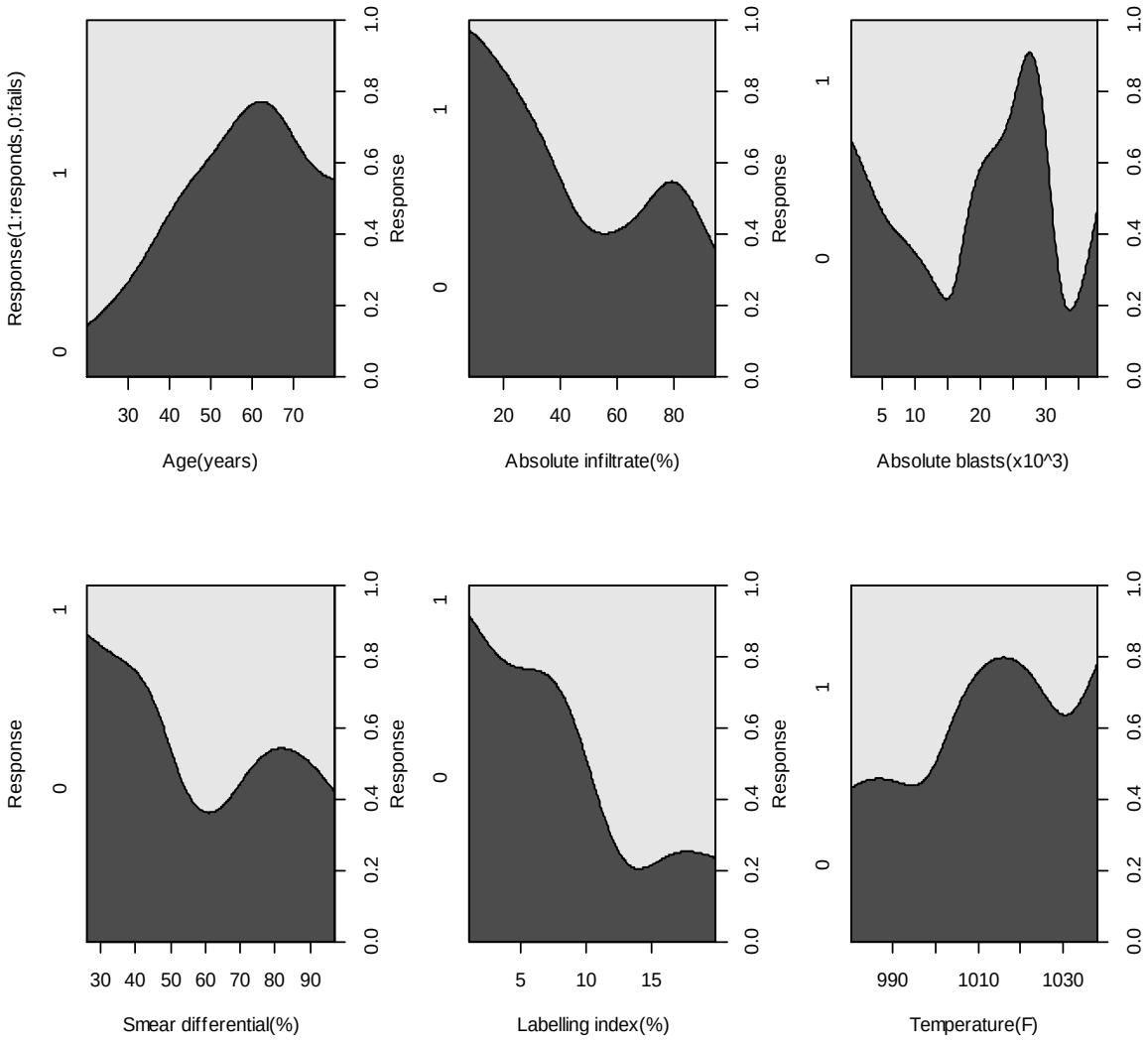
Σκοπός μας είναι, χρησιμοποιώντας την R, να προσαρμόσουμε τα δεδομένα με το μοντέλο της λογιστικής παλινδρόμησης και να επιλέξουμε με διάφορες τεχνικές, μεθόδους και κριτήρια, το στατιστικά καταλληλότερο μοντέλο για την περιγραφή του προβλήματος.

Μεταβλητές	**Περιγραφή**
y (resp)	Αντίδραση: 1 = αντέδρασε στη θεραπεία, 0 = δεν αντέδρασε
x1 (age)	Ηλικία ασθενή (σε χρόνια)
x2 (smear)	Ποσοστό επίστρωσης βλαστοκυττάρων (%)
x3 (infltr)	Ποσοστό κυττάρων λευχαιμίας που εισήλθαν στο μυελό των οστών (%)
x4 (lab)	Ποσοστό κυττάρων που προήλθαν από το μυελό των οστών (%)
x5 (blasts)	Αριθμός βλαστοκυττάρων ($x10^3$)
x6 (temp)	Υψηλότερη θερμοκρασία σώματος ($x10\ ^oF$)

Πίνακας 2.1: Περιγραφή μεταβλητών

2.3.2 ΕΚΤΕΛΕΣΗ ΕΦΑΡΜΟΓΗΣ ΚΑΙ ΕΠΙΛΟΓΗ ΒΕΛΤΙΣΤΟΥ ΜΟΝΤΕΛΟΥ

Αρχικά, ορίζουμε τις επεξηγηματικές μεταβλητές του πιο πάνω πίνακα (x1 μέχρι x6), αλλά και τη μεταβλητή απόκρισης resp. Η απόκριση resp, δηλαδή η αντίδραση του ασθενή στη θεραπεία (Response), ορίζεται σαν κατηγορική μεταβλητή, με 1 = αντέδρασε στη θεραπεία, 0 = δεν αντέδρασε. Γι' αυτό και η προσαρμογή του μοντέλου γίνεται με τη λογιστική παλινδρόμηση, καθώς η μεταβλητή απόκρισης resp ακολουθεί τη Διωνυμική κατανομή.



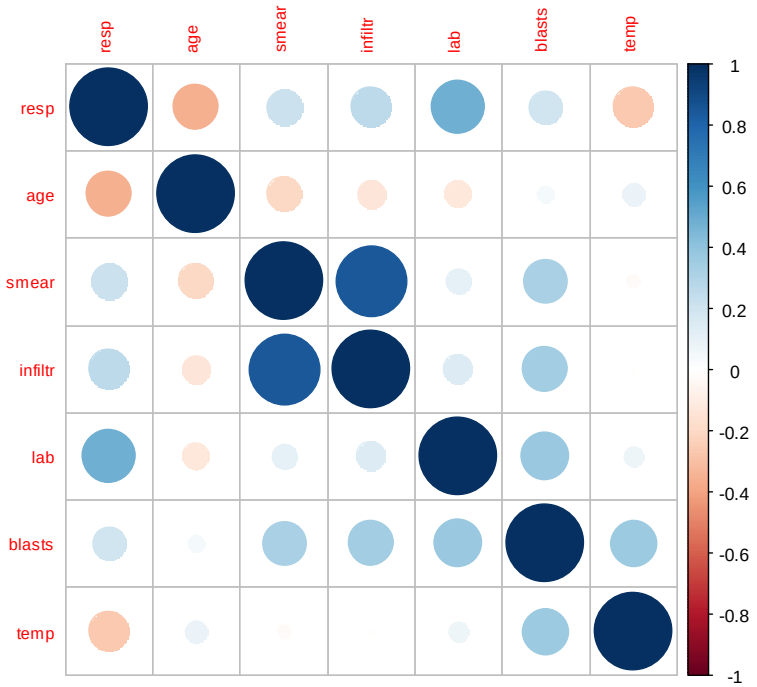
Γραφήματα 2.1: Περιγραφή της resp σε σχέση με την κάθε μεταβλητή.

Πριν την προσαρμογή του μοντέλου, γίνεται γραφικά η περιγραφή των επεξηγηματικών μεταβλητών σε σχέση με τη μεταβλητή απόκρισης resp (Γραφήματα 2.1). Παρατηρούμε πως περιγράφει η κάθε επεξηγηματική μεταβλητή τη σχέση της με τη μεταβλητή απόκρισης resp. Όλες οι γραφικές είναι αρκετά ομαλές, εκτός από τη σχέση της μεταβλητής smear με τη resp που δεν είναι ικανοποιητική (γράφημα 3^{ο} από τα πάνω) και γι' αυτό είμαστε λίγο επιφυλαχτικοί με τη μεταβλητή αυτή.

Επίσης πριν την προσαρμογή, ελέγχουμε και τη συσχέτιση μεταξύ των μεταβλητών αριθμητικά αλλά και γραφικά. Όσο πιο κοντά στην απόλυτη μονάδα είναι οι τιμές, τόσο πιο έντονη θα είναι η μεταξύ τους συσχέτιση. Αν είναι μηδέν τότε είναι ασυσχέτιστες.

	age	smear	infiltr	lab	blasts	temp
age	1	-0.2037821	-0.1369988	-0.1242546	0.0471055	0.0845891
smear	-0.2037821	1	0.8471326	0.1026925	0.3259864	-0.0282492
infiltr	-0.1369988	0.8471326	1	0.1443771	0.3401597	-0.0067099
lab	-0.1242546	0.1026925	0.1443771	1	0.3780289	0.0765292
blasts	0.0471055	0.3259864	0.3402597	0.3780289	1	0.3602475
temp	0.0845891	-0.0282492	-0.0067099	0.0765292	0.3602475	1

Πίνακας 2.2: Πίνακας συσχέτισης των μεταβλητών



Γράφημα 2.2: Η συσχέτιση μεταξύ των μεταβλητών

Από τα πιο πάνω αποτελέσματα (Πίνακας 2.2, Γράφημα 2.2) φαίνεται πολύ καθαρά ότι υπάρχει έντονη συσχέτιση μεταξύ των μεταβλητών smear και infiltr, καθώς η τιμή της μεταξύ τους συσχέτιση corr (smear, infiltr) = 0.8471 και είναι πολύ κοντά στη μονάδα.

Στη συνέχεια γίνεται η προσαρμογή του μοντέλου με λογιστική παλινδρόμηση και συνάρτηση σύνδεσης τη *logit*, καθώς η μεταβλητή απόκρισης resp, όπως έχουμε αναφέρει και πριν είναι κατηγορική μεταβλητή (με τιμές 0 και 1). Εισάγοντας τις ακόλουθες εντολές στην R, λαμβάνουμε τα Αποτελέσματα 2.1 για την προσαρμογή του μοντέλου:

```
> results_glm<-glm(resp~age+smear+infiltr+lab+blasts+temp,family=binomial)
> summary(results_glm)
```

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.56756  -0.55762  -0.05269   0.59038   2.38751

Coefficients:
             Estimate   Std. Error  z value  Pr(>|z|)
(Intercept) 108.33115   41.84379    2.589    0.00963 **
age          -0.06231    0.02746   -2.269    0.02327 *
smear        -0.00469    0.04005   -0.117    0.90677
infiltr       0.03104    0.03789    0.819    0.41264
lab           0.37281    0.13247    2.814    0.00489 **
blasts        0.03267    0.04605    0.710    0.47801
temp         -0.11162    0.04263   -2.618    0.00884 **
---
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: 39.275  on 44  degrees of freedom
AIC: 53.275

Number of Fisher Scoring iterations: 6
```

Αποτελέσματα 2.1

Από τον Πίνακα 2.3 φαίνονται και τα συμμετρικά 95% διαστήματα εμπιστοσύνης για όλες τις παραμέτρους του μοντέλου (α, β1, β2, …).

	2.5%	97.5%
(Intercept)	34.67671201	203.5566964
age	-0.12412446	-0.01334823
smear	-0.08730638	0.07354321
infiltr	-0.03704589	0.11501040
lab	0.14923034	0.68030839
blasts	-0.02695181	0.15324068
temp	-0.20888061	-0.03685288

Πίνακας 2.3: 95% Διαστήματα εμπιστοσύνης

Βλέποντας τα πιο πάνω αποτελέσματα, του προσαρμοσμένου μοντέλου με όλες τις μεταβλητές, παρατηρούμε από τις p-τιμές των ελέγχων Wald ότι μόνο οι επεξηγηματικές μεταβλητές age, lab και temp είναι στατιστικά σημαντικές με p-τιμές 0.0233, 0.0049 και 0.0088, αντίστοιχα.

Οπότε ένας τρόπος να βελτιώσουμε το μοντέλο είναι να χρησιμοποιήσουμε τις τρεις διαδικασίες επιλογής μοντέλου με βήματα, με βάση το κριτήριο AIC. Πριν ξεκινήσουμε, θέτουμε το κενό μοντέλο (null), αλλά και το πλήρες (full) (Παράρτημα C.3) και ακολούθως εκτελούμε τις πιο κάτω εντολές:

```
> step(full, dataAML, direction="backward")
> step(null, scope=list(lower=null, upper=full), direction="forward")
> step(null, scope=list(upper=full), direction="both")
```

Και λαμβάνουμε τα αποτελέσματα για τις τρεις διαδικασίες Backward elimination, Forward selection και Stepwise selection, που φαίνονται στους Πίνακες 2.4 a,b και c, αντίστοιχα.

Μεταβλ. / Steps	age	smear	infiltr	lab	blasts	temp	AIC
1	x	x	x	x	x	x	53.28
2	x		x	x	x	x	51.29
3	x		x	x		x	50.14

Πίνακας 2.4.a: Αποτελέσματα της Backward elimination

Μεταβλ. / Steps	age	smear	infiltr	lab	blasts	temp	AIC
1							72.52
2				x			61.12
3				x		x	55.65
4	x			x		x	51.27
5	x		x	x		x	50.14

Πίνακας 2.4.b: Αποτελέσματα της Forward selection

Μεταβλ. / Steps	age	smear	infiltr	lab	blasts	temp	AIC
1							72.52
2				x			61.12
3				x		x	55.65
4	x			x		x	51.27
5	x		x	x		x	50.14

Πίνακας 2.4.c: Αποτελέσματα της Stepwise selection

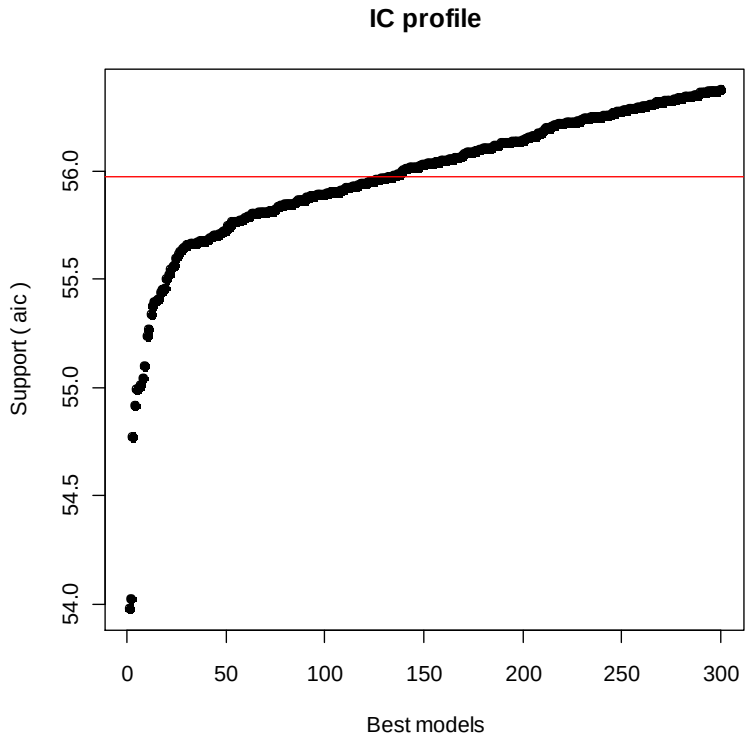
Μετά την εκτέλεση και των τριών διαδικασιών επιλογής με βήματα, οδηγούμαστε στην επιλογή του ίδιου μοντέλου ως βέλτιστου (βλ. Πίνακες 2.4), αυτού με τις μεταβλητές age, infiltr, lab, temp.

Τώρα για να επιβεβαιώσουμε την επιλογή, θα χρησιμοποιήσουμε άλλη μία πολύ αξιόπιστη τεχνική επιλογής στα γενικευμένα γραμμικά μοντέλα, το πακέτο *glmulti*, με βάση τα κριτήρια AIC και BIC, για το οποίο μιλούμε στο Παράρτημα C.1. Με τη βοήθεια της παρακάτω εντολής εκτελούμε το glmulti με το κριτήριο AIC:

```
>test1 <- glmulti (resp~age*smear*infiltr*lab*blasts*temp, data=dataAML, fitfunc=glm,
 +crit = aic)
```

Μετά από 2169450 μοντέλα καταλήγουμε στο εξής «καλύτερο» μοντέλο: resp ~ 1 + age + lab + blasts:age + blasts:infiltr + temp:lab, με AIC = 53.976 και mean AIC = 55.633. Το μοντέλο αυτό περιλαμβάνει τις μεταβλητές age και lab, καθώς και τις αλληλεπιδράσεις της blasts με την age (blasts:age), της blasts με της infiltr (blasts:infiltr) και της temp με τη lab (temp:lab).

Μπορούμε να σημειώσουμε στο σημείο αυτό, ότι το glmulti έχει ένα μικρό μειωνέκτημα, το οποίο επηρεάζει μερικώς τα αποτελέσματα μας. Μπορεί να επιλέξει τις αλληλεπιδράσεις μεταξύ κάποιων μεταβλητών για το βέλτιστο μοντέλο, χωρίς όμως να επιλέξει τις μεταβλητές μόνες τους. Όπως βλέπουμε πιο πάνω, έχει επιλέξει π.χ. την αλληλεπίδραση της blasts με την infiltr, χωρίς να έχει επιλέξει ούτε τη μεταβλητή blasts στο βέλτιστο μοντέλο, αλλά ούτε και την infiltr. Αυτό δε μπορεί να γίνει. Όμως, όπως θα δούμε πιο κάτω με τη βοήθεια της ελεγχοσυνάρτησης Deviance, λύνουμε το «πρόβλημα» αυτό.



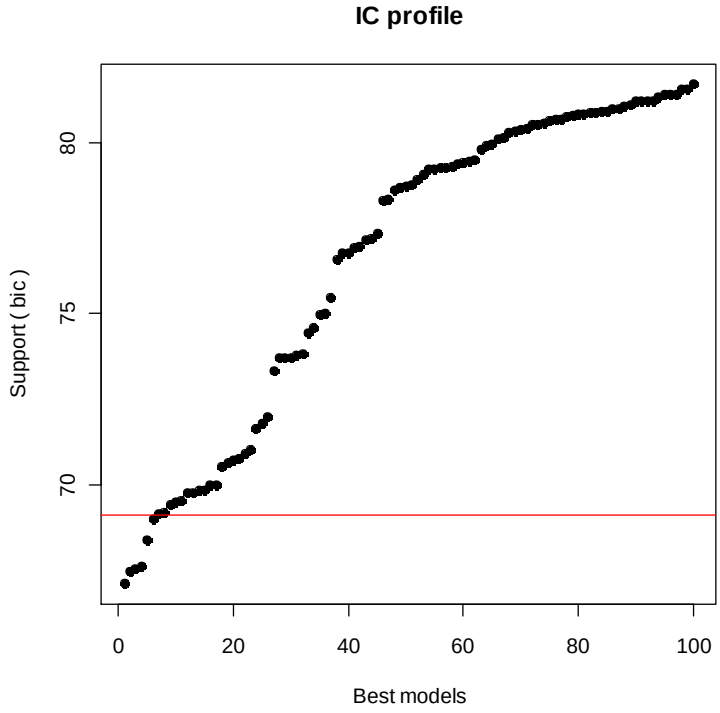
Γράφημα 2.3: Τα καλύτερα μοντέλα του test1. Η κόκκινη γραμμή δείχνει το mean AIC.

Επίσης, εκτελούμε το glmulti και με βάση το κριτήριο BIC:

```
>test2 <- glmulti (resp~age*smear*infiltr*lab*blasts*temp, data=dataAML, fitfunc = glm,
+crit =bic)
```

Και μετά από 2169450 μοντέλα καταλήγουμε στο εξής «καλύτερο» μοντέλο: resp ~ 1 + lab + blasts:infiltr + temp:age + temp:lab, με BIC = 66.590 και mean BIC = 68.469.

.



Γράφημα 2.4: Τα καλύτερα μοντέλα του test2. Η κόκκινη γραμμή δείχνει το mean BIC.

Επίσης, με τη βοήθεια του γενετικού αλγορίθμου (genetic algorithm) (Παράρτημα C.1), εκτελούμε το glmulti με βάση το AIC, με την εντολή:

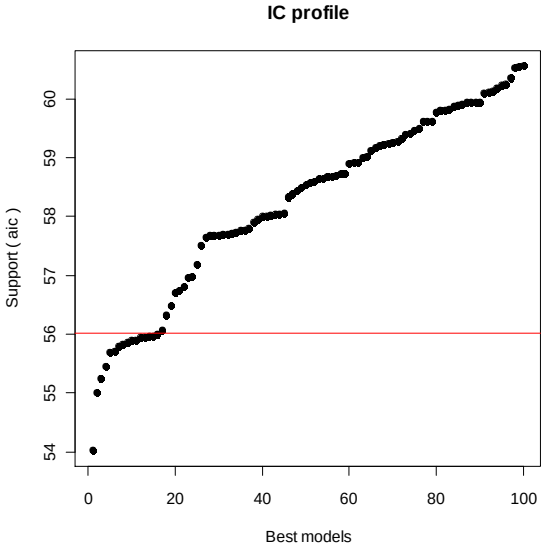
```
> test3ga <- glmulti (resp~age*smear*infiltr*lab*blasts*temp,  data=dataAML,  fitfunc = glm,
+crit = aic, method = "g")
```

Μετά από 490 generations καταλήγουμε στο εξής «καλύτερο» μοντέλο: resp ~ 1 + lab + blasts:age + blasts:infiltr + temp:age + temp:lab, με AIC = 54.023 και mean AIC = 58.239 (Γράφημα 2.5) και ο αλγόριθμος συγκλίνει.

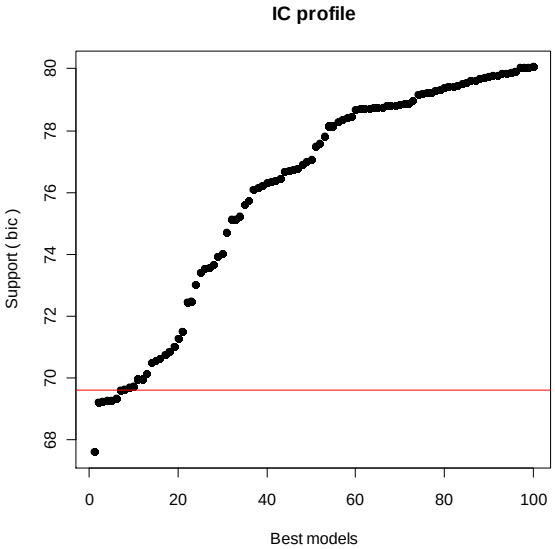
Αλλά και πάλι με τη βοήθεια του γενετικού αλγορίθμου εκτελούμε το glmulti με βάση το BIC αυτή τη φορά, με την εντολή:

```
>test4ga <- glmulti (resp~age*smear*infiltr*lab*blasts*temp,  data=dataAML,  fitfunc = glm,
+crit = bic, method = "g")
```

Όπου μετά από 650 generations αυτή τη φορά καταλήγουμε στο πολύ καλό μοντέλο: resp ~ 1 + lab + age + temp, με BIC = 67.603 και mean BIC = 76.088 (Γράφημα 2.6) και ο αλγόριθμος συγκλίνει.



Γράφημα 2.5: Τα καλύτερα μοντέλα του test3ga



Γράφημα 2.6: Τα καλύτερα μοντέλα του test4ga

Με τη χρήση του γενετικού αλγόριθμου από τη μία, κερδίζουμε πολύ χρόνο στην εκτέλεση για την εύρεση του βέλτιστου μοντέλου, όμως η εγκυρότητα του αποτελέσματος δεν είναι όπως τα αποτελέσματα που παίρνουμε σε κανονικό χρόνο. Αυτό το βλέπουμε πολύ καλά εκτελώντας κάποιες εντολές (Παράρτημα C.4) και έχουμε με το AIC 490 generations με την εκτέλεση να γίνεται με ποσοστό 5.93% του κανονικού χρόνου, ενώ με το BIC, να έχουμε 650 generations, όπου η εκτέλεση γίνεται σε 8.19% του κανονικού χρόνου που χρειάζεται.

Έτσι τώρα συγκρίνοντας τα αποτελέσματα των τεσσάρων δοκιμών (tests) του glmulti, αλλά και των αποτελεσμάτων που είχαμε από τις κατά βήματα διαδικασίες επιλογής μπορούμε να καταλήξουμε στην τελική επιλογή των πιο κατάλληλων μεταβλητών στη συγκεκριμένη εφαρμογή. Συνοψίζοντας, έχουμε ως καλύτερες μεταβλητές τις age, infiltr, lab και temp από τις βηματικές διαδικασίες και μετά, με τη βοήθεια του glmulti (τα τέσσερα tests), καταλήγουμε στις age, lab, temp και τις αλληλεπιδράσεις blasts:age, blasts:infiltr, temp:age και temp:lab.

Όμως, θεωρώντας τις πρώτες δύο δοκιμές του glmulti ως πιο ακριβείς, ιδιαίτερα το test1 το οποίο εκτός του ότι κατέληξε μετά από 2169450 μοντέλα, είχε και αρκετά χαμηλό AIC = 53.976, επιλέγουμε ως καλύτερες μεταβλητές για το βέλτιστο μοντέλο τις age, lab και τις αλληλεπιδράσεις blasts:age, blasts:infiltr και temp:age. Έτσι, εκτελούμε με τη βοήθεια των εντολών:

```
> results_test1<-glm(resp~age+lab+blasts*age+blasts*infiltr+temp*lab, family=binomial)
> summary(results_test1)
```

την ανάλυση παλινδρόμησης του test1 περιέχοντας τις σημαντικότερες μεταβλητές και τις αλληλεπιδράσεις που επιλέχτηκαν μετά την εκτέλεση του glmulti. Από τα Αποτελέσματα 2.2 παρατηρούμε ότι οι μεταβλητές αλλά και οι αλληλεπιδράσεις δε βγαίνουν στατιστικά σημαντικές.

Για να λύσουμε αυτό το πρόβλημα θα μας βοηθήσει η ελεγχοσυνάρτηση Deviance (§ 2.2.5.2). Έστω ότι το M_0 είναι το μοντέλο του test1 που περιέχει τις αλληλεπιδράσεις. Θεωρούμε ως M_1 ότι είναι το ίδιο μοντέλο αλλά χωρίς τις αλληλεπιδράσεις. Με τη βοήθεια των εντολών:

```
> results_test1_xoris<-glm(resp~age+lab+blasts+infiltr+temp,family=binomial)
> summary(results_test1_xoris)
```

θα έχουμε την ανάλυση παλινδρόμησης του μοντέλου αλλά χωρίς τις αλληλεπιδράσεις (Αποτελέσματα 2.3). Όπως έχουμε πει και στη θεωρία, με *H_0: ισχύει το μοντέλο M_0* έναντι της εναλλακτικής υπόθεσης *H_1: ισχύει το μοντέλο M_1*, θα ελέγξουμε το εξής (εξίσωση 2.8): $D_0 - D_1 \sim X_d^2$ *ασυμπτωτικά*, με d τη διαφορά του αριθμού των παραμέτρων του πλήρες από το ελαττωμένο μοντέλο, δηλαδή ο αριθμός των αλληλεπιδράσεων που αφαιρέθηκαν. Αφού πάρουμε από τα Αποτελέσματα 2.2 και 2.3 τις Residual Deviance (βλ. τα bold γράμματα), όπου έχουμε 35.227 και 39.289, αντίστοιχα, ελέγχουμε την τιμή της X_3^2 βρίσκουμε την p-τιμή = 0.2548 > 0.05. Παρατηρούμε ότι η p-τιμή είναι αρκετά μεγάλη, οπότε και απορρίπτουμε τη μηδενική υπόθεση. Επομένως θεωρούμε ως καταλληλότερο το μοντέλο χωρίς τις αλληλεπιδρασεις. Τώρα, θέτουμε το προηγούμενο μοντέλο M_1, ως M_0 και θεωρούμε ως M_1 το μοντέλο χωρίς τις στατιστικά μη σημαντικές μεταβλητές από τα Αποτελέσματα 2.3, δηλαδή

αφαιρούμε από το καινούριο μοντέλο M_1 τις μεταβλητές infiltr και blasts. Και πάλι, με *H_0: ισχύει το μοντέλο M_0* έναντι της εναλλακτικής υπόθεσης *H_1: ισχύει το μοντέλο M_1*, παίρνουμε τις Residual Deviance από τα Αποτελέσματα 2.3 και 2.4 (§ 2.3.3) (βλ. τα bold γράμματα), όπου έχουμε 39.289 και 43.265, αντίστοιχα. Με τη βοήθεια της ελεγχοσυνάρτησης Deviance ελέχουμε την τιμή της X_2^2 και βρίσκουμε την p-τιμή = 0.0925 που δεν είναι πολύ μικρή, αλλά αρκετά μικρή ώστε να έχουμε σοβαρές ενδείξεις για να μην απορρίψουμε τη μηδενική υπόθεση. Έτσι καταλήγουμε στο τελικό μοντέλο με τις μεταβλητές age, lab και temp, που το θεωρούμε ως βέλτιστο.

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

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.9579  -0.4772  -0.1390   0.4090   2.2347

Coefficients:
                Estimate   Std. Error  z value  Pr(>|z|)
(Intercept)   -51.982566  89.903928  -0.578   0.5631
age            -0.045084   0.033973  -1.327   0.1845
lab            15.954270   8.951991   1.782   0.0747 .
blasts          0.042895   0.273218   0.157   0.8752
infiltr         0.023090   0.028518   0.810   0.4181
temp            0.048819   0.090119   0.542   0.5880
age:blasts     -0.005713   0.004197  -1.361   0.1735
blasts:infiltr  0.004288   0.004729   0.907   0.3646
lab:temp       -0.015626   0.008946  -1.747   0.0807 .
---
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: 35.227  on 42  degrees of freedom
AIC: 53.227

Number of Fisher Scoring iterations: 6
```

Αποτελέσματα 2.2

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

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-1.57917  -0.56009  -0.05083   0.59963   2.39885

Coefficients:
             Estimate   Std. Error  z value  Pr(>|z|)
(Intercept) 108.58104   41.83584   2.595   0.00945 **
age          -0.06187    0.02716  -2.278   0.02273 *
lab           0.37480    0.13140   2.852   0.00434 **
infiltr       0.02745    0.02179   1.259   0.20787
blasts        0.03294    0.04617   0.714   0.47553
temp         -0.11202    0.04255  -2.633   0.00847 **
---
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: 39.289  on 45  degrees of freedom
AIC: 51.289

Number of Fisher Scoring iterations: 6
```

Αποτελέσματα 2.3

2.3.3 ΤΟ ΒΕΛΤΙΣΤΟ ΜΟΝΤΕΛΟ

Παρουσιάζουμε την ανάλυση παλινδρόμησης του μοντέλου που επιλέχτηκε ως το βέλτιστο, με τις μεταβλητές age, lab και temp, με τη βοήθεια των εντολών:

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

Από τα Αποτελέσματα 2.4 του μοντέλου λογιστικής παλινδρόμησης που επιλέχτηκε μετά την εκτέλεση των τριών διαδικασιών επιλογής κατά βήματα, την εκτέλεση του glmulti, αλλά και τον έλεγχο με τη Deviance, βλέπουμε ότι όλες οι μεταβλητές που επιλέχτηκαν, δηλ. η age, η lab και η temp είναι σημαντικές με p-τιμές 0.0222, 0.0015, 0.0136, αντίστοιχα. Επίσης η τιμή του κριτηρίου AIC είναι αρκετά χαμηλή.

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

Αποτελέσματα 2.3

Μετά την πιο πάνω «ολόκληρη» διαδικασία, καταλήγουμε στο εξής μοντέλο:

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = 87.388 - 0.0585age + 0.385\ lab - 0.089temp$$

Κατ' αρχάς παρατηρούμε ότι αντέδρασαν στη θεραπεία 87 ασθενείς. Η αύξηση της ηλικίας (age) κατά ένα χρόνο επηρεάζει το μοντέλο, μειώνει (πολλαπλασιαστικά) το odds μιας θετικής απόκρισης κατά $e^{-0.0585} = 0.943 < 1$, δηλαδή 5.7% μείωση του odds. Δηλαδή μπορούμε να πούμε ότι κάθε επιπλέον έτος ηλικίας του ασθενή, μειώνει

κατά 5.7% την πιθανότητα να αντιδράσει στη θεραπεία, με τις υπόλοιπες συνθήκες να παραμένουν σταθερές. Αλλά η αύξηση του ποσοστού των κυττάρων λευχαιμίας που προήλθαν από το μυελό των οστών (lab) κατά 1%, αυξάνει (πολλαπλασιαστικά) το odds μιας θετικής απόκρισης κατά $e^{0.385} = 1.470 > 1$, δηλαδή 38% αύξηση της πιθανότητας να αντιδράσει ο ασθενής στη θεραπεία και πάλι με τις υπόλοιπες συνθήκες να παραμένουν σταθερές. Επίσης, αν αυξηθεί η θερμοκρασία κατά 10 ^{o}F (βαθμούς Farenheit) (temp), μειώνει (πολλαπλασιαστικά) το odds μιας θετικής απόκρισης κατά $e^{-0.089} = 0.915 < 1$, δηλαδή 8.9% μείωση της πιθανότητας να αντιδράσει στη θεραπεία ο ασθενής, με τις υπόλοιπες συνθήκες να παραμένουν σταθερές. Όμως, η αύξηση του ποσοστού της επίστρωσης των βλαστοκυττάρων (smear), η αύξηση του ποσοστού των κυττάρων λευχαιμίας που εισήλθαν στο μυελό των οστών (infiltr), αλλά και η αύξηση του αριθμού των βλαστοκυττάρων (blasts) κατά 1000 δεν επηρεάζει την αντίδραση.

2.3.4 ΣΥΜΠΕΡΑΣΜΑΤΑ

Τα γενικευμένα γραμμικά μοντέλα παρέχουν ένα ισχυρό αλλά και ευέλικτο πλαίσιο εργασίας για την εφαρμογή των μοντέλων παλινδρόμησης σε μια ποικιλία από μη φυσιολογικές μεταβλητές απόκρισης. Ιδιαίτερα σε δυαδικές (binary) μεταβλητές απόκρισης όπου χρησιμοποιείται η λογιστική παλινδρόμηση.

Η πιο πάνω εφαρμογή είναι ένα μικρό παράδειγμα της δυναμικότητας και της ευελιξίας των μοντέλων αυτών, αλλά και πιο συγκεκριμένα της λογιστικής παλινδρόμησης, η οποία χρησιμοποιείται ευρέως τα τελευταία χρόνια στον τομέα της Στατιστικής, για τη λύση διαφόρων προβλημάτων σε όλους τους τομείς. Αν και δε χρησιμοποιήσαμε στην εφαρμογή όλες τις τεχνικές και τα κριτήρια επιλογής μοντέλου που αναφέρθηκαν στη θεωρία, εντούτοις παρουσιάσαμε μερικά ώστε να γίνει κατανοητή γενικότερα η χρήση των τεχνικών επιλογής. Τα αποτελέσματα της εφαρμογής είναι αρκετά καλά και μπορούμε να πούμε ότι πετύχαμε το σκοπό μας.

3. ΜΟΝΤΕΛΑ ΕΠΙΒΙΩΣΗΣ ΚΑΙ ΤΟ ΜΟΝΤΕΛΟ ΤΟΥ COX

3.1 ΜΟΝΤΕΛΑ ΕΠΙΒΙΩΣΗΣ

3.1.1 ΕΙΣΑΓΩΓΗ

Ο όρος **ανάλυση επιβίωσης (survival analysis)** χρησιμοποιείται για να περιγράψει την περιοχή εκείνη της Στατιστικής που ασχολείται με **δεδομένα διάρκειας ζωής**. Τα δεδομένα αυτά αποτελούν δεδομένα που αφορούν το χρόνο μέχρι να συμβεί ένα γεγονός, όπως για παράδειγμα ο θάνατος ενός ασθενή, η ανάπτυξη κάποιου όγκου, η αστοχία κάποιας δοκού κλπ. Η ανάλυση επιβίωσης χρησιμοποιείται ευρέως σε βιοϊατρικές εφαρμογές και κλινικές μελέτες όπου ενδέχεται να συγκρίνονται δύο ή περισσότερες θεραπείες. Ο χρόνος που παρατηρείται στις περιπτώσεις αυτές, μπορεί να αφορά τη διάγνωση κάποιας ασθένειας, την ανταπόκριση του ασθενή στη θεραπεία ή την εμφάνιση κάποιας παρενέργειας.

Για τα δεδομένα διάρκειας ζωής δεν εφαρμόζονται οι συνήθεις στατιστικές μέθοδοι που χρησιμοποιούνται στην ανάλυση δεδομένων. Ένας λόγος για τον οποίο συμβαίνει αυτό, αποτελεί το γεγονός ότι τα δεδομένα επιβίωσης γενικά δεν έχουν συμμετρικές κατανομές. Ένα τυπικό ιστόγραμμα αυτού του είδους δεδομένων θα έδειχνε μια μακρύτερη «ουρά» στο δεξί του μέρος, όπου και θα περιέχονταν οι περισσότερες παρατηρήσεις. Έτσι δεν θα ήταν λογικό να θεωρήσουμε ότι τα δεδομένα ακολουθούν κανονική κατανομή.

3.1.2 ΒΑΣΙΚΕΣ ΕΝΟΙΕΣ

Στην ανάλυση επιβίωσης υπάρχουν δύο βασικές συναρτήσεις, η συνάρτηση επιβίωσης (survival function) και η συνάρτηση διακινδύνευσης (hazard function), ο ορισμός των οποίων δίνεται παρακάτω (Collett, 2003).

Έστω ότι συμβολίζουμε με $T > 0$ το χρόνο επιβίωσης μιας υπό μελέτη μονάδας t. Ο χρόνος T ανάλογα με το πρόβλημα μπορεί να εκφράζει για παράδειγμα τη διάρκεια

παραμονής ενός ασθενή στο νοσοκομείο, το χρόνο μέχρι την εμφάνιση κάποιου όγκου σε έναν ασθενή ή το ασκούμενο φορτίο σε μια δοκό της οποίας η αντοχή εξετάζεται κλπ. Η μεταβλητή T μπορεί να πάρει οποιαδήποτε μη αρνητική τιμή, και έτσι την καλούμε **τυχαία μεταβλητή**, η οποία σχετίζεται με το χρόνο επιβίωσης της μονάδας t. Έστω $f(t)$ η συνάρτηση πυκνότητας πιθανότητας (σ.π.π.) της T. Τότε η συνάρτηση κατανομής (distribution function) της T, που εκφράζει την πιθανότητα ο χρόνος επιβίωσης να είναι μικρότερος ή ίσος από μια τιμή t είναι:

$$F(t) = P[T \leq t] = \int_0^t f(u)du$$

Ονομάζουμε **συνάρτηση επιβίωσης** (survival function) $S(t)$, μια συνάρτηση η οποία εκφράζει την πιθανότητα ο χρόνος επιβίωσης T να είναι μεγαλύτερος από μια τιμή t. Δηλαδή θα έχουμε:

$$S(t) = P[T > t] = 1 - F(t) = \int_t^\infty f(u)du.$$

Και επίσης έχουμε:

$$f(t) = \frac{d}{dt}F(t) = -\frac{d}{dt}S(t)$$

Μια άλλη πολύ βασική συνάρτηση στην ανάλυση επιβίωσης είναι η **συνάρτηση διακινδύνευσης** (hazard function) $h(t)$, η οποία εκφράζει τον κίνδυνο να επέλθει η διακοπή σε κάποιο χρόνο t. Είναι δηλαδή η δεσμευμένη πιθανότητα να πεθάνει μια μονάδα σε χρόνο t, δεδομένου ότι έχει επιβιώσει μέχρι εκείνη τη στιγμή. Η συνάρτηση διακινδύνευσης ορίζεται ως:

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

Από τη δεσμευμένη πιθανότητα στον πιο πάνω ορισμό και με $F(t)$ τη σ.κ. του T:

$$P[t \le T < t + \delta t | T \ge t] = \frac{P[t \le T < t + \delta t]}{P[T \ge t]} = \frac{F(t + \delta t) - F(t)}{S(t)}$$

Έτσι έχουμε:

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{F(t + \delta t) - F(t)}{\delta t} \right\} \left(\frac{1}{S(t)} \right)$$

και επειδή:

$$f(t) = \lim_{\delta t \to 0} \left\{ \frac{F(t + \delta t) - F(t)}{\delta t} \right\}$$

έχουμε τελικά ότι:

$$h(t) = \frac{f(t)}{S(t)}$$

Μια άλλη χρήσιμη συνάρτηση στην ανάλυση επιβίωσης, είναι η **σωρευτική συνάρτηση διακινδύνευσης** (cumulative hazard function) $H(t)$, η οποία ορίζεται ως:

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

Από τα πιο πάνω βρίσκεται η σχέση της συνάρτησης με τα υπόλοιπα μεγέθη:

$$H(t) = \int_0^t \frac{f(u)}{S(u)} du = \int_0^t -\frac{S'(u)}{S(u)} du = [-lnS(u)]_0^t = -lnS(t)$$

και άρα:

$$S(t) = \exp\{-H(t)\}.$$

Γενικά, όπως φαίνεται και από τις σχέσεις μεταξύ τους, οι συναρτήσεις $h(t), f(t), S(t), F(t), H(t)$ είναι μαθηματικά ισοδύναμες, αφού αν ξέρουμε μια από αυτές, οι υπόλοιπες μπορούν να υπολογιστούν.

3.2 ΤΟ ΜΟΝΤΕΛΟ ΑΝΑΛΟΓΙΚΗΣ ΔΙΑΚΙΝΔΥΝΕΥΣΗΣ ΤΟΥ COX

3.2.1 ΕΙΣΑΓΩΓΗ

Το 1972 ο Άγγλος στατιστικολόγος David Cox παρουσίασε ένα μοντέλο παλινδόμησης το οποίο αποτελεί σήμερα ένα από τα πιο διαδεδομένα μοντέλα στην ανάλυση δεδομένων διάρκειας ζωής με αποκομμένες παρατηρήσεις (Cox D. R., 1972). **Το μοντέλο αναλογικής διακινδύνευσης του Cox** (the Cox proportional hazards model), χρησιμοποιείται ευρέως για τον προσδιορισμό διαφορών στην επιβίωση παρατηρούμενων μονάδων όταν υποβάλλονται σε διάφορες θεραπείες, προγνωστικούς παράγοντες σε κλινικές μελέτες, καθώς και σε πολλά άλλα προβλήματα βιοϊατρικής.

Πολλές φορές στην ανάλυση επιβίωσης έχουμε να κάνουμε με μεγάλες βάσεις δεδομένων αποτελούμενες από χρόνους επιβίωσης, από τις οποίες χρειάζεται να προσαρμόσουμε ένα παραμετρικό μοντέλο. Όταν όμως τα δεδομένα αφορούν τον άνθρωπο, το γεγονός ότι κάθε πληθυσμός είναι διαφορετικός, καθιστά δύσκολο το να υιοθετήσουμε ένα γνωστό παραμετρικό μοντέλο. Έτσι υποθέτουμε ένα μεγάλο σύνολο παραγόντων που μπορεί να επηρεάζουν την επιβίωση σε κάθε μελέτη, και ακολούθως υπολογίζουμε την επίδραση του κάθε παράγοντα καταλήγοντας έτσι σε ένα μοντέλο που εκφράζει τα δεδομένα.

3.2.2 ΟΡΙΣΜΟΣ ΤΟΥ ΜΟΝΤΕΛΟΥ ΤΟΥ COX

Έστω N ο αριθμός των μονάδων σε μια μελέτη στην οποία χρειάζεται να εκτιμηθεί η επίδραση κάποιων μεταβλητών $\boldsymbol{x} = (x_1, x_2, \dots, x_p)$ στη διάρκεια ζωής τους. Το μοντέλο αναλογικής διακινδύνευσης του Cox υποθέτει ότι η συνάρτηση διακινδύνευσης στο χρόνο t μιας παρατηρούμενης μονάδας i με διάνυσμα συμμεταβλητών $\boldsymbol{x'}_i = \left(x_{1i}, x_{2i}, \dots, x_{pi}\right)$ δίνεται από τη σχέση (Marubini & Valsecchi, 1995):

$$h(t; \boldsymbol{x}) = h_0(t)\mathrm{e}^{\boldsymbol{\beta' x}_i} \quad \text{για } i = 1,2, \dots, N \tag{3.1}$$

Η διακινδύνευση εξαρτάται από το χρόνο και τις συμμεταβλητές, αλλά μέσω δύο διαφορετικών παραγόντων. Η ποσότητα $h_0(t)$ είναι γνωστή ως βασική συνάρτηση διακινδύνευσης στο χρόνο t. Είναι μια αυθαίρετη συνάρτηση του χρόνου, η οποία θεωρείται σταθερή για όλες τις μονάδες.

Για $\boldsymbol{x} = \mathbf{0}$ προκύπτει:

$$h(t; \mathbf{0}) = h_0(t)$$

Έτσι, μπορούμε να ορίσουμε την $h_0(t)$ ως τη συνάρτηση διακινδύνευσης ενός ατόμου όταν οι τιμές όλων των συμμεταβλητών είναι μηδενικές ($\boldsymbol{x}_i = \mathbf{0},\ i = 1,2, \dots, p$).

Η δεύτερη ποσότητα στην εξίσωση 3.1, εξαρτάται από τις συμμεταβλητές μέσω ενός διανύσματος p συντελεστών παλινδρόμησης $\boldsymbol{\beta'}$. Όπως και σε όλα τα μοντέλα οι συντελεστές αυτοί εκφράζουν το πόσο κάθε συμμεταβλητή επηρεάζει το τελικό μοντέλο. Βασικός στόχος είναι η εκτίμηση αυτών των συντελεστών, αφού έτσι καταλήγουμε στο ποιοι είναι οι στατιστικά σημαντικοί για το μοντέλο που εξετάζουμε. Οι παράγοντες για τους οποίους ο συντελεστής παλινδρόμησης υπολογίζεται ίσος με μηδέν, δεν επηρεάζουν τη διακινδύνευση του μοντέλου, και άρα ούτε την επιβίωση. Επίσης, οι συμμεταβλητές θεωρούνται σταθερές στο χρόνο, όπως άλλωστε συμβαίνει όταν σε μια κλινική μελέτη εκφράζουν ηλικία, φύλο, θεραπεία ή άλλα κλινικά και βιοχημικά χαρακτηριστικά.

Το μοντέλο του Cox δεν αποτελεί ένα πλήρως παραμετρικό μοντέλο και χαρακτηρίζεται ως ημι-παραμετρικό. Αυτό συμβαίνει γιατί το κύριο χαρακτηριστικό του μοντέλου, είναι ότι η μορφή της συνάρτησης $h_0(t)$ δεν καθορίζεται. Καθορίζεται όμως η αναλογία της διακινδύνευσης για δύο οποιεσδήποτε μονάδες με διανύσματα συμμεταβλητών $\boldsymbol{x}_1$ και $\boldsymbol{x}_2$ αντίστοιχα:

$$\frac{h(t; \boldsymbol{x}_1)}{h(t; \boldsymbol{x}_2)} = \frac{h_0(t) e^{\boldsymbol{\beta'} \boldsymbol{x}_1}}{h_0(t) e^{\boldsymbol{\beta'} \boldsymbol{x}_2}} = e^{\boldsymbol{\beta'}(\boldsymbol{x}_1 - \boldsymbol{x}_2)} \qquad (3.2)$$

Όπως βλέπουμε, η αναλογία αυτή δεν εξαρτάται από τη συνάρτηση $h_0(t)$. Έτσι λέμε ότι το μοντέλο του Cox είναι ένα μοντέλο παλινδρόμησης αναλογικής διακινδύνευσης, αφού

υποθέτει ότι το ποσοστό αποτυχίας δύο οποιονδήποτε μονάδων είναι αναλογικό, δεδομένου ότι ο λόγος στην εξίσωση 3.2 δεν εξαρτάται από το χρόνο. Αν επιπλέον λογαριθμίσουμε το λόγο αυτό, έχουμε:

$$\ln h(t;\boldsymbol{x}_1) - \ln h(t;\boldsymbol{x}_2) = \boldsymbol{\beta}'(\boldsymbol{x}_1 - \boldsymbol{x}_2) \qquad (3.3)$$

Και άρα φαίνεται ότι το μοντέλο υποθέτει μια σταθερή διαφορά μεταξύ των λογαρίθμων της διακινδύνευσης δύο μονάδων.

Η συνάρτηση επιβίωσης στο μοντέλο του Cox υπολογίζεται ως εξής (Caroni, 2004):

Από τη σχέση $H(t) = \int_0^t h(u)du$ και τον ορισμό της συνάρτησης διακινδύνευσης στο μοντέλο του Cox (εξίσωση 3.1), παίρνουμε:

$$H(t;\boldsymbol{x}) = \int_0^t h_0(u)\mathrm{e}^{\boldsymbol{\beta}'\boldsymbol{x}_i}du = H_0(t)\mathrm{e}^{\boldsymbol{\beta}'\boldsymbol{x}_i}$$

Τέλος, από τη σχέση $S(t) = e^{-H(t)}$ θα καταλήξουμε:

$$S(t;\boldsymbol{x}) = e^{-H(t,\boldsymbol{x})} = e^{-H_0(t)\mathrm{e}^{\boldsymbol{\beta}'\boldsymbol{x}_i}} = \{S_0(t)\}^{\mathrm{e}^{\boldsymbol{\beta}'\boldsymbol{x}_i}}$$

3.2.3 ΠΑΡΑΔΕΙΓΜΑ ΣΤΟ ΜΟΝΤΕΛΟ ΤΟΥ COX

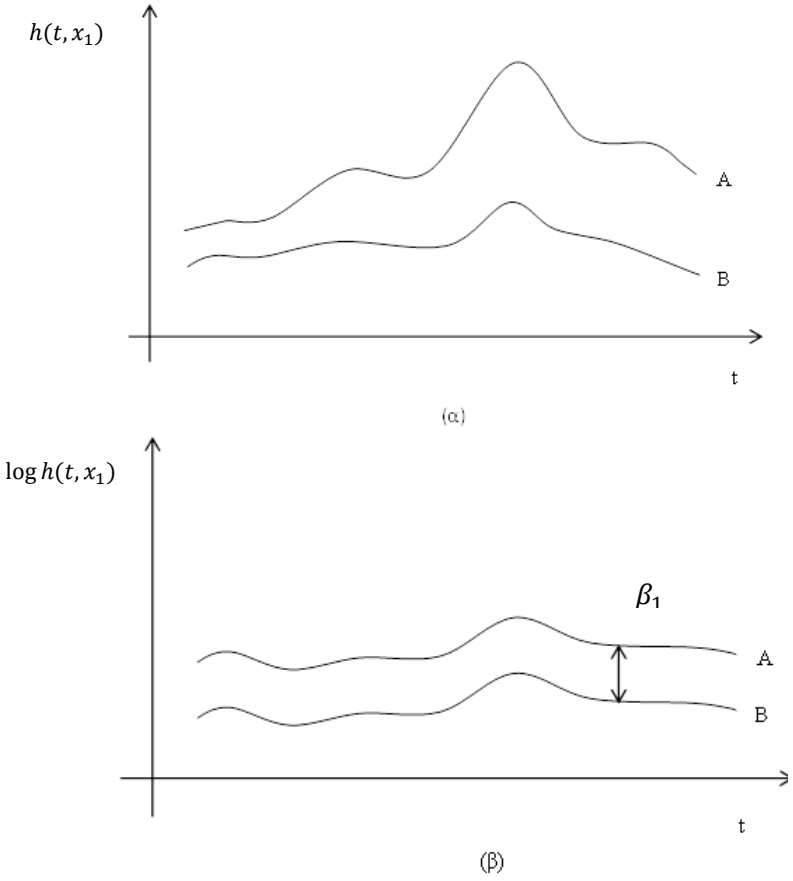
Έστω ότι διεξάγεται μια κλινική μελέτη στην οποία συγκρίνονται δύο θεραπείες για μια ασθένεια: η καθιερωμένη θεραπεία Α και η πειραματική θεραπεία Β (Marubini, Valecchi, 2004). Το διάνυσμα $\boldsymbol{x}$ των συμμεταβλητών, περιέχει δείκτες για τις θεραπείες καθώς και άλλες μεταβλητές που αντιπροσωπεύουν άλλα χαρακτηριστικά που ενδέχεται να επηρεάζουν την επιβίωση. Προς το παρόν, έστω ότι το διάνυσμα $\boldsymbol{x}$ περιέχει μια μόνο μεταβλητή x_1 για τις θεραπείες όπου με $x_1 = 0$ υποδηλώνεται η θεραπεία Α και με $x_1 = 1$ η θεραπεία Β. Το μοντέλο του Cox υποθέτει ότι οι συναρτήσεις διακινδύνευσης είναι $h_0(t)$ και $h_0(t)e^{\beta_1}$ για τους ασθενείς που λαμβάνουν τη θεραπεία Α και Β αντίστοιχα. Δηλαδή σε κάθε χρονική στιγμή, το ποσοστό θανάτου στην ομάδα Β είναι όσο στην ομάδα Α πολλαπλασιασμένο με ένα σταθερό παράγοντα e^{β_1},

οποιαδήποτε και αν είναι η μορφή της συνάρτησης $h_0(t)$ ή ισοδύναμα ο λογάριθμος των συναρτήσεων διακινδύνευσης στις δύο ομάδες έχει σταθερή απόσταση β_1. Αρνητική τιμή του συντελεστή β_1 δείχνει ότι η θεραπεία Β έχει μικρότερη πιθανότητα θανάτου, ή ισοδύναμα, μεγαλύτερη πιθανότητα επιβίωσης. Τα παραπάνω φαίνονται στη συνέχεια (Γράφημα 3.2).

Μοντέλο 1:

Θεραπεία	x_1	$h(t,x_1)=h_0(t)e^{\beta_1 x_1}$
A	0	$h_0(t)$
B	1	$e^{\beta_1 x_1}$

Πίνακας 3.2



Γράφημα 3.2: (α) συναρτήσεις διακινδύνευσης, (β) ο λογάριθμος της συνάρτησης διακινδύνευσης σύμφωνα με το μοντέλο του Cox με μια συμμεταβλητή ($x_1 = 0$ για τη θεραπεία Α, $x_2 = 1$ για τη θεραπεία Β) και $\beta_1 < 0$.

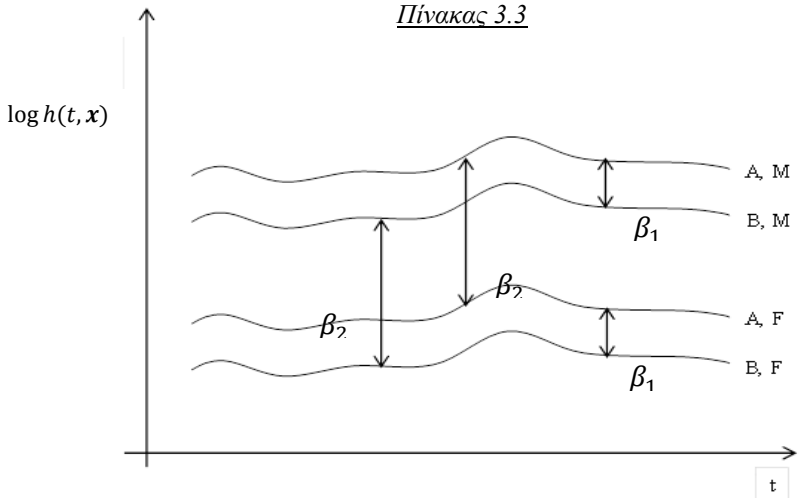
Έστω τώρα, ότι εισάγουμε στο μοντέλο ακόμα μια κατηγορική μεταβλητή x_2 που εκφράζει κάποιο χαρακτηριστικό των ασθενών, για παράδειγμα το φύλο (Μοντέλο 2). Το διάνυσμα των συμμεταβλητών είναι επομένως $\boldsymbol{x} = (x_1, x_2)$ και η συνάρτηση διακινδύνευσης $h(t, \boldsymbol{x}) = h_0(t)e^{\beta_1 x_1 + \beta_2 x_2}$. Αναπαριστούμε με $x_2 = 0$ το γυναικείο (F) και με $x_2 = 1$ το αντρικό (M). Έστω ότι ο συντελεστής β_2 είναι θετικός.

Σε λογαριθμική κλίμακα, οι συναρτήσεις διακινδύνευσης, βάσει της θεραπείας x_1 έχουν σταθερή απόσταση β_1, αλλά ταυτόχρονα σε κάθε ομάδα θεραπείας η διακινδύνευση των ασθενών διαφορετικού φύλου είναι σε σταθερή απόσταση β_2. Αυτό φαίνεται παρακάτω (Γράφημα 3.3), όπου οι συναρτήσεις διακινδύνευσης για τις τέσσερις ομάδες που προσδιορίζονται συνδυάζοντας τις πιθανές τιμές των μεταβλητών x_1 και x_2.

Μοντέλο 2:

Ομάδα	$\boldsymbol{x} = (x_1, x_2)$	$h(t, \boldsymbol{x}) = h_0(t)e^{\beta_1 x_1 + \beta_2 x_2}$
A, F	0, 0	$h_0(t)$
B, F	1, 0	$h_0(t)e^{\beta_1}$
A, M	0, 1	$h_0(t)e^{\beta_2}$
B, M	1, 1	$h_0(t)e^{\beta_1+\beta_2}$

Πίνακας 3.3



Γράφημα 3.3: Ο λογάριθμος των συναρτήσεων διακινδύνευσης σύμφωνα με το μοντέλο του Cox με δύο συμμεταβλητές ($x_1 = 0$ για τη θεραπεία A, $x_2 = 1$ για τη θεραπεία B) και ($x_2 = 0$ για τις γυναίκες, $x_2 = 1$ για τους άντρες) με $\beta_1 < 0$ και $\beta_2 > 0$.

Το συγκεκριμένο μοντέλο υποθέτει ότι οι δύο συμμεταβλητές x_1 και x_2 έχουν ανεξάρτητες επιδράσεις στο βαθμό διακινδύνευσης, δηλαδή δεν υπάρχει αλληλεπίδραση μεταξύ x_1 και x_2, αφού θεωρεί ότι οι θεραπείες έχουν την ίδια επίδραση σε άντρες και γυναίκες. Στο παράδειγμα αυτό, ο συντελεστής β_2 έχει θετική τιμή και αυτό δείχνει ότι γενικά οι άντρες διατρέχουν μεγαλύτερο κίνδυνο θανάτου στη συγκεκριμένη ασθένεια από τις γυναίκες.

Έστω τώρα ότι θέλουμε να εξετάσουμε αν υπάρχει αλληλεπίδραση των δύο παραγόντων, δηλαδή του τύπου της θεραπείας και του φύλου (Μοντέλο 3). Για το σκοπό αυτό, προσθέτουμε στο μοντέλο του Cox μια ψευδομεταβλητή, έστω z, η οποία εκφράζει το γινόμενο των δύο συμμεταβλητών x_1 και x_2.

Επομένως η συνάρτηση διακινδύνευσης γίνεται:

$$h(t, \boldsymbol{x}, z) = h_0(t)e^{\beta_1 x_1 + \beta_2 x_2 + \gamma z}$$

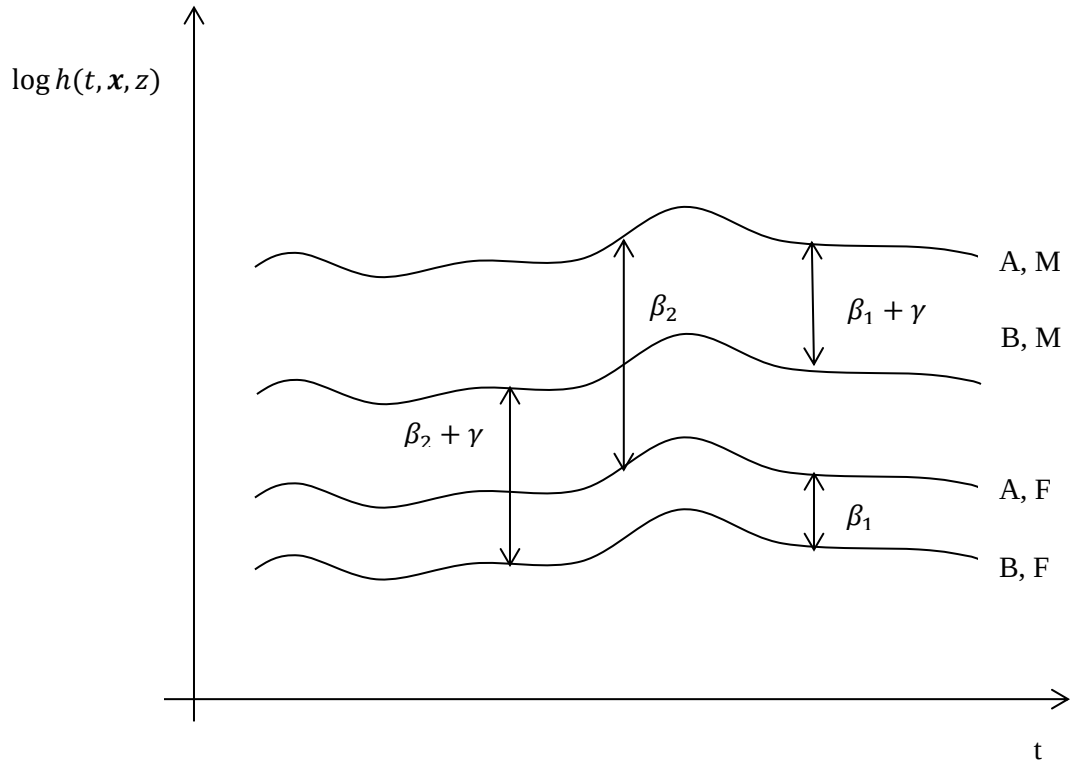
όπου $z = x_1 x_2$ παίρνει την τιμή 1, όταν $x_1 = x_2 = 1$ και την τιμή 0, διαφορετικά.

Έστω ότι ο συντελεστής γ της ψευδομεταβλητής z είναι αρνητικός. Τότε οι συναρτήσεις διακινδύνευσης παίρνουν την πιο κάτω μορφή (Γράφημα 3.4).

Μοντέλο 3:

Ομάδα	$\boldsymbol{x} = (x_1, x_2, z)$	$h(t, \boldsymbol{x}) = h_0(t)e^{\beta_1 x_1 + \beta_2 x_2 + \gamma z}$
A, F	0, 0, 0	$h_0(t)$
B, F	1, 0, 0	$h_0(t)e^{\beta_1}$
A, M	0, 1, 0	$h_0(t)e^{\beta_2}$
B, M	1, 1, 1	$h_0(t)e^{\beta_1 + \beta_2 + \gamma}$

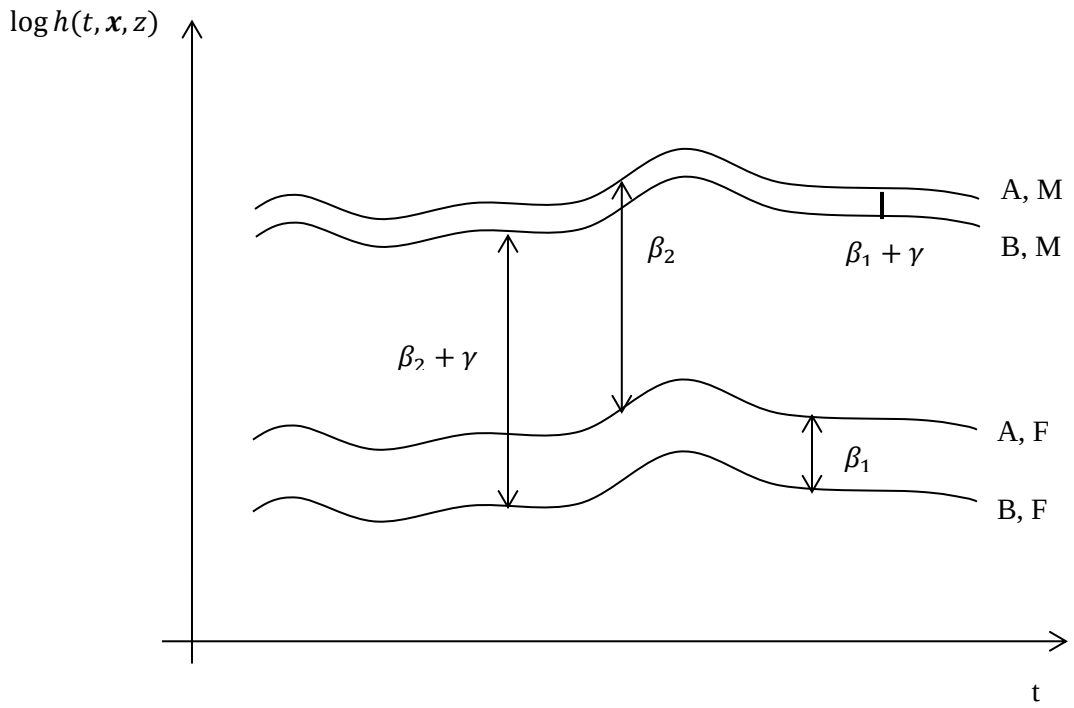
Πίνακας 3.4



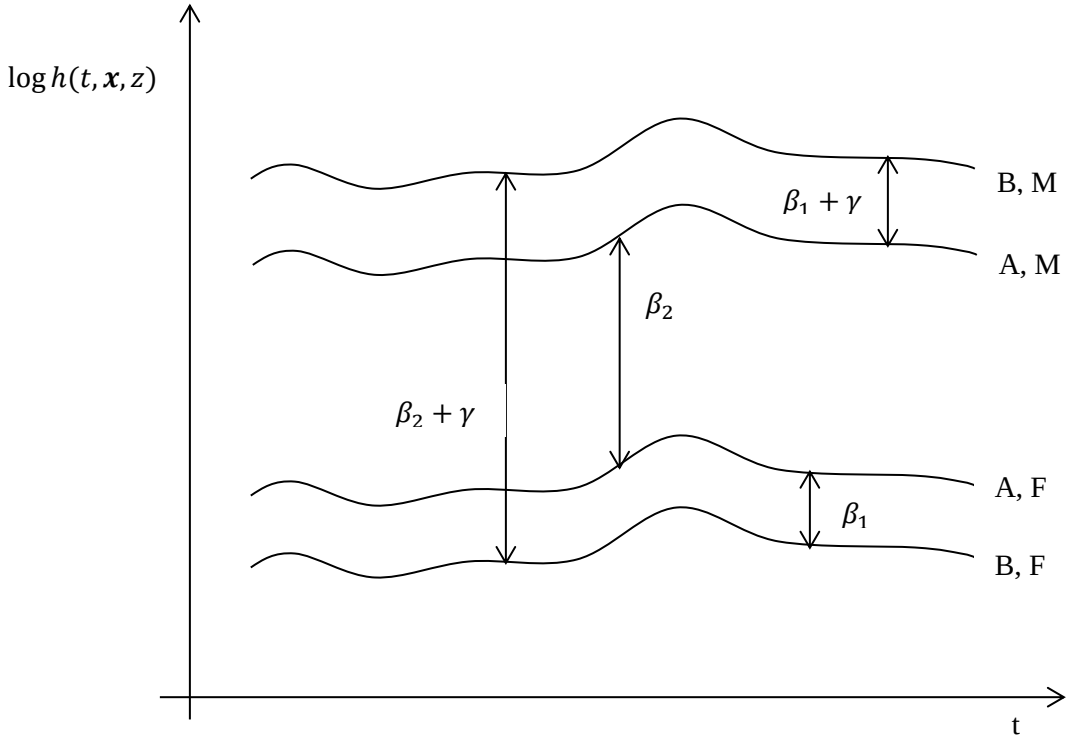
Γράφημα 3.4: Ο λογάριθμος των συναρτήσεων διακινδύνευσης σύμφωνα με το μοντέλο του Cox με δύο συμμεταβλητές x_1, x_2 και μια ψευδομεταβλητή $z = x_1 x_2$ με $\beta_1 < 0$, $\beta_2 > 0$ και $\gamma < 0$.

Στην περίπτωση αυτή υπάρχει μια ποσοτική αλληλεπίδραση των δύο παραγόντων, αφού φαίνεται και από το διάγραμμα των συναρτήσεων διακινδύνευσης ότι η θεραπεία Β είναι πιο ευεργετική στους άντρες, όπου η αρνητική ποσότητα γ προστίθεται στον αρνητικό συντελεστή β_1.

Στη περίπτωση που ο συντελεστής γ της ψευδομεταβλητής z είναι θετικός τότε υπάρχουν δύο πιθανές καταστάσεις για το μοντέλο και φαίνονται πιο κάτω (Γραφήματα 3.5, 3.6). Αν για παράδειγμα έχουμε $\beta_1 = -0.7$ και $\gamma = +0.5$, οι λογαριθμημένες συναρτήσεις διακινδύνευσης θα ήταν όπως στο Γράφημα 3.5, όπου και εδώ υπάρχει ποσοτική αλληλεπίδραση των παραγόντων. Αν και η θεραπεία Β είναι ευεργετική και για τα δύο φύλα, εντούτοις στους άντρες φαίνεται να έχει λιγότερη επίδραση. Αν όμως οι συντελεστές παλινδόμησης είναι $\beta_1 = -0.7$ και $\gamma = +2$ τότε θα έχουμε την πιο κάτω περίπτωση (Γράφημα 3.6), όπου παρατηρούμε μια ποιοτική αλληλεπίδραση, αφού η θεραπεία Β είναι ευεργετική για τις γυναίκες, αλλά επιβλαβής για τους άντρες.



Γράφημα 3.5: Ο λογάριθμος των συναρτήσεων διακινδύνευσης σύμφωνα με το μοντέλο του Cox με δύο συμμεταβλητές x_1, x_2 και μια ψευδομεταβλητή $z = x_1 x_2$ με $\beta_1 < 0,\ \beta_2 > 0$ και $0 < \gamma < -\beta_1$.



Γράφημα 3.6: Ο λογάριθμος των συναρτήσεων διακινδύνευσης σύμφωνα με το μοντέλο του Cox με δύο συμμεταβλητές x_1, x_2 και μια ψευδομεταβλητή $z = x_1 x_2$ με $\beta_1 < 0,\ \beta_2 > 0$ και $\gamma > -\beta_1$

Γενικά, το μοντέλο αναλογικής διακινδύνευσης του Cox προσδιορίζει τη μορφή της συνάρτησης διακινδύνευσης, η οποία περιέχει τους συντελεστές παλινδρόμησης $\boldsymbol{\beta}$ μόνο ως άγνωστους παράγοντες. Ο Cox εισήγαγε μια μέθοδο εκτίμησης των $\boldsymbol{\beta}$ (και άρα και της διακινδύνευσης), η οποία βασίζεται στη «μερική πιθανοφάνεια» και φαίνεται στη συνέχεια του κεφαλαίου. Η εκτίμηση των συντελεστών αυτών, μας επιτρέπει να προσδιορίσουμε ποσοτικά το σχετικό βαθμό αποτυχίας μιας μονάδας με διάνυσμα συμμεταβλητών $\boldsymbol{x}_1$ σε σχέση με μια άλλη μονάδα με διάνυσμα συμμεταβλητών $\boldsymbol{x}_2$ και όχι τον απόλυτο βαθμό για κάθε μονάδα ξεχωριστά. Η ανάλυση του Cox επομένως, χρησιμοποιείται ευρέως σε συγκριτικές μελέτες θεραπειών και σε μελέτες επίδρασης κάποιου επικίνδυνου παράγοντα. Για παράδειγμα, στο Μοντέλο 1 της πιο πάνω εφαρμογής, μια εκτίμηση $\hat{\beta}_1$ του συντελεστή β_1 ίση με -0.7 σημαίνει ότι η πιθανότητα θανάτου για τους ασθενείς που δέχονται τη θεραπεία Β είναι περίπου η μισή $(e^{\widehat{\beta}_1} = e^{-0.7} = 0.497)$ σε σχέση με την αντίστοιχη των ασθενών που δέχονται τη θεραπεία Α, σε κάθε χρονική στιγμή. Στο δεύτερο μοντέλο του παραδείγματος, η εκτίμηση της επίδρασης της θεραπείας, προσαρμόζεται από τον παράγοντα του φύλου, δηλαδή $e^{\widehat{\beta}_1}$, και είναι η εκτιμώμενη αναλογία μεταξύ των ασθενών που λαμβάνουν τη θεραπεία Α και των ασθενών που λαμβάνουν τη θεραπεία Β, με την άλλη συμμεταβλητή να μένει σταθερή (δηλαδή ίδιο φύλο).

3.2.4 ΕΚΤΙΜΗΣΗ ΤΩΝ ΣΥΝΤΕΛΕΣΤΩΝ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Έστω ότι έχουμε ένα σύνολο δεδομένων επιβίωσης N μονάδων από τις οποίες J μονάδες αποτυγχάνουν. Γενικά, το J είναι μικρότερο από το N $(J < N)$, λόγω της παρουσίας αποκομμένων παρατηρήσεων στα δεδομένα επιβίωσης. Έστω επίσης ότι $t_{(1)} < t_{(2)} < \cdots < t_{(J)}$ είναι οι J χρονικές στιγμές κατά τα τις οποίες παρατηρείται η διακοπή των μονάδων. Έστω $R(t)$ είναι το σύνολο των μονάδων σε κίνδυνο τη χρονική στιγμή t, οι μονάδες δηλαδή, που είναι ζωντανές και υπό παρακολούθηση ακριβώς πριν τη χρονική στιγμή t. Συμβολίζουμε με j τη μονάδα που αποτυγχάνει σε χρόνο $t_{(j)}$, η οποία έχει διάνυσμα συμμεταβλητών $\boldsymbol{x}_j$. Γενικά όπως είδαμε, $\boldsymbol{x}_i$ είναι το διάνυσμα των συμμεταβλητών για την i-οστή μονάδα και οι συμμεταβλητές έχουν σταθερή τιμή στο χρόνο. Η πιθανότητα μια μονάδα με συμμεταβλητές $\boldsymbol{x}$ να αποτύχει σε ένα διάστημα

$(t, t + dt)$, δεδομένου των μονάδων σε ρίσκο τη χρονική στιγμή t, είναι $h(t, \boldsymbol{x})dt$. Έτσι, δοθέντος ότι παρατηρείται διακοπή μιας μονάδας σε χρόνο $t_{(j)}$, η πιθανότητα να είναι η μονάδα με διάνυσμα συμμεταβλητών $\boldsymbol{x}_j$ είναι:

$$\frac{h(t_{(j)}, \boldsymbol{x}_j)dt}{\sum_{i \in R(t_{(j)})} h(t_{(j)}, \boldsymbol{x}_i)dt}$$

Έτσι, η συνάρτηση πιθανοφάνειας για τα δεδομένα και όπου $R_j = R(t_{(j)})$, εκφράζεται ως εξής (Cox D. R., 1972):

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{J} \frac{h(t_{(j)}, \boldsymbol{x}_j)dt}{\sum_{i \in R_j} h(t_{(j)}, \boldsymbol{x}_i)dt}$$

Και τελικά, από τον ορισμό της συνάρτησης διακινδύνευσης στο μοντέλο του Cox ($h(t, \boldsymbol{x}) = h_0(t)\mathrm{e}^{\boldsymbol{\beta}' \boldsymbol{x}_i}$), η συνάρτηση πιθανοφάνειας απλοποιείται σε:

$$L = L(\boldsymbol{\beta}) = \prod_{j=1}^{J} \frac{e^{\boldsymbol{\beta}' \boldsymbol{x}_j}}{\sum_{i \in R_j} e^{\boldsymbol{\beta}' \boldsymbol{x}_i}} \tag{3.4}$$

Οι συντελεστές παλινδρόμησης $\boldsymbol{\beta}$ εκτιμώνται από τις τιμές $\widehat{\boldsymbol{\beta}}$, οι οποίες μεγιστοποιούν τη συνάρτηση πιθανοφάνειας ή ισοδύναμα το λογάριθμό της. Το διάνυσμα $\widehat{\boldsymbol{\beta}}$ ονομάζεται εκτιμήτρια μέγιστης πιθανοφάνειας (ε.μ.π.) της $\boldsymbol{\beta}$.

Ο λογάριθμος της συνάρτησης πιθανοφάνειας είναι:

$$l(\boldsymbol{\beta}) = \sum_{j=1}^{p} \left\{ \boldsymbol{\beta}' \boldsymbol{x}_j - \ln \left[\sum_{i \in R_j} e^{\boldsymbol{\beta}' \boldsymbol{x}_i} \right] \right\} \tag{3.5}$$

όπου l είναι η λογαριθμιμένη πιθανοφάνεια που αντιστοιχεί στο χρόνο αποτυχίας $t_{(j)}$.

Οι τιμές $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_K)$ υπολογίζονται μηδενίζοντας τις K μερικές παραγώγους της $l(\boldsymbol{\beta})$ ως προς β_k $(k = 1, \dots, p)$ δηλαδή:

$$\frac{\partial l}{\partial \beta_k} = 0\,, \qquad k = 1, \dots, p$$

Το πιο πάνω σύστημα εξισώσεων λύνεται ως προς $\boldsymbol{\beta}$ με αριθμητικές μεθόδους (π.χ. Newton-Raphson). Αν η k-οστή τιμή του διανύσματος $\boldsymbol{x}_i$ είναι η x_{ki}, τότε η k-οστή παράγωγος της l (εξίσωση 3.5) θα είναι:

$$\frac{\partial l}{\partial \beta_k} = \sum_{j=1}^{p} x_{jk} - \sum_{j=1}^{p} \left[\frac{\sum_{i \in R_j} x_{ki} e^{\boldsymbol{\beta}' \boldsymbol{x}_i}}{\sum_{i \in R_j} e^{\boldsymbol{\beta}' \boldsymbol{x}_i}} \right] \tag{3.6}$$

Αυτή η έκφραση μπορεί να δείξει διαισθητικά τη λειτουργία της πιθανοφάνειας: Αν για παράδειγμα, κάποιος ασθενής που πεθαίνει τείνει να έχει ψηλότερες τιμές στην k-οστή συμμεταβλητή, η τιμή β_k θα πρέπει να είναι αρκετά μεγάλη, έτσι ώστε να μειώνει την παράγωγο στο μηδέν. Παίρνοντας τη δεύτερη παράγωγο της ποσότητας l, μπορούν να προσδιοριστούν εκτιμήσεις των διασπορών των $\widehat{\boldsymbol{\beta}}$. Για παράδειγμα, η παράγωγος της (6) ως προς β_k είναι:

$$-\frac{\partial^2 l}{\partial \beta_k \partial \beta_r} = \left[\frac{\sum_{i \in R_j} x_{ki}^2 e^{\boldsymbol{\beta}' \boldsymbol{x}_i}}{\sum_{i \in R_j} e^{\boldsymbol{\beta}' \boldsymbol{x}_i}} - \left(\frac{\sum_{i \in R_j} x_{ki} e^{\boldsymbol{\beta}' \boldsymbol{x}_i}}{\sum_{i \in R_j} e^{\boldsymbol{\beta}' \boldsymbol{x}_i}} \right)^2 \right] \tag{3.7}$$

Γενικά από τα παραπάνω, η συνάρτηση πιθανοφάνειας (εξίσωση 3.4) ονομάστηκε από τον Cox ως **«μερική πιθανοφάνεια»**, γιατί η έκφραση $h_0(t)$ δεν εμφανίζεται σε αυτή την ανάλυση (Cox D. R., 1975).

3.2.5 ΠΕΡΙΓΡΑΦΗ ΤΗΣ CROSS-VALIDATION (CVL) ΣΤΟ ΜΟΝΤΕΛΟ ΤΟΥ COX

Όπως έχουμε αναφέρει (§ 1.2.4.4), είναι πολύ σημαντικός ο ρόλος της μεθόδου **cross validation** γιατί βρίσκει τις βέλτιστες τιμές για τα λ, ώστε να μας βοηθήσουν στις τεχνικές με ποινή για την επιλογή του βέλτιστου μοντέλου.

Στο μοντέλο του Cox, η $l_i(\boldsymbol{\beta})$ προέρχεται όπως περιγράφεται πιο κάτω. Έχουμε τη συνάρτηση πιθανοφάνειας (εξίσωση 3.4), όταν δεν έχουμε ισόπαλους χρόνους να είναι:

$$L = L(\boldsymbol{\beta}) = \prod_{j=1}^{n} \frac{e^{\boldsymbol{\beta}' \boldsymbol{x}_j}}{\sum_{k \in R_j} e^{\boldsymbol{\beta}' \boldsymbol{x}_k}}$$

Όταν η i-παρατήρηση αφήνεται έξω, ο i-οστός παράγοντας πετάγεται έξω και η i-παρατήρηση απομακρύνεται από όλα τα σύνολα κινδύνου πριν τη χρονική στιγμή t_i. Έτσι θα έχουμε $t_j < t_i \; για \; j < i$ και καταλήγουμε:

$$L_{(-i)}(\boldsymbol{\beta}) = \prod_{j<i}^{n} \frac{e^{\boldsymbol{\beta}' \boldsymbol{x}_j}}{\sum_{k \in R_j} e^{\boldsymbol{\beta}' \boldsymbol{x}_k} - e^{\boldsymbol{\beta}' \boldsymbol{x}_i}} \prod_{j>i}^{n} \frac{e^{\boldsymbol{\beta}' \boldsymbol{x}_j}}{\sum_{k \in R_j} e^{\boldsymbol{\beta}' \boldsymbol{x}_k}}$$

Επειδή η $L_i(\boldsymbol{\beta})$ είναι ίση με $L(\boldsymbol{\beta}) / L_{(-i)}(\boldsymbol{\beta})$, καταλήγουμε:

$$L_i(\boldsymbol{\beta}) = \prod_{j<i}^{n} (1 - \frac{e^{\boldsymbol{\beta}' \boldsymbol{x}_i}}{\sum_{k \in R_j} e^{\boldsymbol{\beta}' \boldsymbol{x}_k}}) \; \frac{e^{\boldsymbol{\beta}' \boldsymbol{x}_i}}{\sum_{k \in R_i} e^{\boldsymbol{\beta}' \boldsymbol{x}_k}}$$

όπου ο παράγοντας-i πεθαίνει τη χρονική στιγμή t_j, αφού λαμβάνουμε υπόψη τα σύνολα κινδύνου και τους χρόνους επιβίωσης. Τότε ο λογάριθμος της πιθανοφάνειας θα είναι:

$$l_i(\boldsymbol{\beta}) = \sum_{j<i} \left\{ 2\boldsymbol{\beta}' \boldsymbol{x}_i - \ln \left[\sum_{k \in R_j} e^{\boldsymbol{\beta}' \boldsymbol{x}_k} \right] + \ln \left[\sum_{k \in R_i} e^{\boldsymbol{\beta}' \boldsymbol{x}_k} \right] \right\}$$

Απομένει τώρα ο υπολογισμός της cross validated συνάρτησης πιθανοφάνειας *cvl*. Αλλά χρειάζονται για τον υπολογισμό της, οι leave-one-out συντελεστές παλινδρόμησης $\widehat{\boldsymbol{\beta}}_{(-i)}$. Για να καθοριστούν οι συντελεστές αυτοί χρειάζεται η προσαρμογή n-μοντέλων Cox. Μπορούν να θεωρηθούν τέσσερις τρόποι προσέγγισης των συντελεστών αυτών (Verweij & van Houwelingen, 1993). Εμείς θα παρουσιάσουμε μόνο κάποια στοιχεία από τον πρώτο τρόπο, που χρησιμοποιεί το γεγονός ότι οι $\widehat{\boldsymbol{\beta}}_{(-i)}$ εξ ορισμού

μεγιστοποιούν το $l_i(\boldsymbol{\beta})$. Οπότε η πρώτη παράγωγος θα είναι ίση με μηδέν για $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_{(-i)}$ και έτσι η προσέγγιση Taylor πρώτης-τάξεως, για $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ θα έχει τη μορφή:

$$\widehat{\boldsymbol{\beta}}_{(-i)} = \widehat{\boldsymbol{\beta}} + (\frac{\partial^2 l}{\partial \boldsymbol{\beta}^2} - \frac{\partial^2 l_i}{\partial \boldsymbol{\beta}^2})^{-1} \frac{\partial l_i}{\partial \boldsymbol{\beta}}$$

Για το μοντέλο του Cox η πρώτη παράγωγος των $l_i(\boldsymbol{\beta})$ δίνεται ως (Verweij & van Houwelingen, 1993):

$$\frac{\partial l_i}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}) = -\sum_{j<i} \frac{e^{\boldsymbol{\beta}' \boldsymbol{x}_i}}{\left(\sum_{k \in R_j} e^{\boldsymbol{\beta}' \boldsymbol{x}_k}\right) - e^{\boldsymbol{\beta}' \boldsymbol{x}_i}} \left(\boldsymbol{x}_i - \frac{\sum_{k \in R_j} e^{\boldsymbol{\beta}' \boldsymbol{x}_k} \boldsymbol{x}_k}{\sum_{k \in R_j} e^{\boldsymbol{\beta}' \boldsymbol{x}_k}} \right) + \left(\boldsymbol{x}_i - \frac{\sum_{k \in R_i} e^{\boldsymbol{\beta}' \boldsymbol{x}_k} \boldsymbol{x}_k}{\sum_{k \in R_i} e^{\boldsymbol{\beta}' \boldsymbol{x}_k}} \right)$$

Όπως έχουμε πει, στην R, χρησιμοποιώντας τις κατάλληλες εντολές (όπως θα δούμε και στη συνέχεια, στην εφαρμογή) και βελτιστοποιώντας τη cvl συνάρτηση, υπολογίζονται αριθμητικά, αλλά και γραφικά οι βέλτιστες τιμές για τα λ.

3.2.6 ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ

Στο μοντέλο αναλογικής διακινδύνευσης του Cox, οι εκτιμήσεις των συντελεστών παλινδρόμησης ακολουθούν προσεγγιστικά μια Κανονική κατανομή, όταν υπάρχει ένας επαρκής αριθμός δεδομένων στο δείγμα. Έτσι μπορούμε να διεξάγουμε ελέγχους υποθέσεων με βάση τις εκτιμήσεις των συντελεστών και τα τυπικά τους σφάλματα, για κάθε μοντέλο που προσαρμόζουμε.

3.2.6.1 Έλεγχος με το λόγο των πιθανοφανειών

Ο πιο συνήθης τρόπος, είναι ο **έλεγχος του λόγου των πιθανοφανειών** (likelihood ratio test) όπου μπορούν να ελεγχθούν υποθέσεις όπως η $\beta_i = 0$. Δηλαδή, κατά πόσον η διακινδύνευση, και επομένως η διάρκεια ζωής, εξαρτάται από τη συμμεταβλητή x_i.

Συγκεκριμένα, το μοντέλο προσαρμόζεται δύο φορές, χρησιμοποιώντας την εξίσωση 3.5. Πρώτα με τη συμμεταβλητή x_i ($\beta_i \neq 0$), και μετά χωρίς αυτήν ($\beta_i = 0$). Έτσι πρώτα υπολογίζουμε τη μεγιστοποιημένη τιμή του λογαρίθμου της μερικής πιθανοφάνειας του μοντέλου που περιλαμβάνει την μεταβλητή x_i και είναι η $\hat{l}_1$. Ακολούθως υπολογίζουμε

τη μεγιστοποιημένη τιμή του λογαρίθμου της μερικής πιθανοφάνειας του μοντέλου που δεν περιλαμβάνει την μεταβλητή x_i και είναι $\hat{l}_0$.

Τώρα εφαρμόζεται ο έλεγχος του λόγου των πιθανοφανειών, δηλαδή συγκρίνεται η τιμή της $-2(\hat{l}_0 - \hat{l}_1)$ με τη X_1^2 κατανομή. Αν η p-τιμή του ελέγχου είναι αρκετά μικρή τότε θα έχουμε σοβαρές ενδείξεις για να απορρίψουμε τη μηδενική υπόθεση H_0, δηλαδή η μεταβλητή x_i θα είναι στατιστικά σημαντική για το συγκεκριμένο μοντέλο. Αν η p-τιμή είναι μεγάλη, σημαίνει έχουμε σοβαρές ενδείξεις για να θεωρήσουμε τη x_i ως στατιστικά μη σημαντική για το μοντέλο και να την πετάξουμε έξω από το βέλτιστο μοντέλο που θέλουμε να καταλήξουμε. Ακολουθώντας αυτή τη μέθοδο, συγκρίνωντας τα υποψήφια μοντέλα, μπορούμε να καταλήξουμε στο πιο κατάλληλο, με τις στατιστικά σημαντικότερες μεταβλητές.

3.2.6.2 Έλεγχος με τη μέθοδο του Wald

Άλλη μέθοδος, είναι η χρησιμοποίηση του **ελέγχου του Wald**, ο οποίος είναι χρήσιμος μόνο για μια πρώτη ένδειξη για το ποιες είναι οι στατιστικά σημαντικότερες μεταβλητές, όταν έχουμε μοντέλο με πολλές συμμεταβλητές, διότι απαιτεί προσαρμογή ενός μόνο μοντέλου. Όπως έχουμε αναφερθεί και στο μοντέλο λογιστικής παλινδρόμησης (§ 2.2.5.3), έτσι και για το μοντέλο του Cox, προσαρμόζουμε το μοντέλο και θέλουμε να ελέγξουμε την υπόθεση H_0: $\beta_i = 0$ έναντι της H_1: $\beta_i \neq 0$, με $i=1,\ldots, k$. Υπολογίζουμε τους εκτιμημένους $\hat{\beta}_i$ συντελεστές παλινδρόμησης για την κάθε μεταβλητή x_i, αντίστοιχα.

Από τη σχέση $\left\{\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)}\right\}^2$, η ελεγχοσυνάρτηση Wald για τη μηδενική υπόθεση H_0 με $\beta_i = 0$ εκφράζεται ως $\left\{\frac{\hat{\beta}_i}{se(\hat{\beta}_i)}\right\}^2$ και συγκρίνεται με την κατανομή X_1^2, δίνοντας την p-τιμή του ελέγχου.

Αλλά, ισοδύναμα, μπορεί να συγκριθεί και η τιμή της $\left\{\frac{\hat{\beta}_i}{se(\hat{\beta}_i)}\right\}$ με την κατανομή $N(0,1)$, πάλι με τη μηδενική υπόθεση για $\beta_i = 0$.

Και πάλιν εδώ, ελέγχοντας τις p-τιμές των ελέγχων Wald, μπορούμε να αποφασίσουμε για την κάθε υποψήφια μεταβλητή x_i αν είναι στατιστικά σημαντική ή όχι, για το βέλτιστο μοντέλο που θέλουμε να καταλήξουμε.

3.2.6.3 Έλεγχος με τη μέθοδο LASSO στο μοντέλο του Cox

Μία άλλη μέθοδος που μπορούμε να χρησιμοποιήσουμε για την επιλογή ενός μοντέλου αναλογικής διακινδύνευσης του Cox κάπως παραλλαγμένη, είναι η **LASSO**. Ο νέος αλγόριθμός της (Tibshirani, 1997) βασίζεται σε ένα συνδυασμό του βελτιστοποιημένου gradient (κλίση) με τον αλγόριθμο της Newton-Raphson. Η μέθοδος έχει ως σκοπό την ελαχιστοποίηση του λογαρίθμου της μερικής πιθανοφάνειας, να συρρικνώνει (shrinks) τους συντελεστές προς το μηδέν και να ρυθμίζει αυτόματα πολλούς από αυτούς ακριβώς στο μηδέν, τακτοποιώντας τους με τέτοιο τρόπο, ανάλογα της χρήσης τους στο μοντέλο. Έτσι μειώνει την εκτίμηση της διακύμανσης, ενώ παρέχει ένα ερμηνεύσιμο τελικό μοντέλο (Goeman, 2010).

Η Lasso είναι μια προσέγγιση της κανονικοποιημένης εκτίμησης για τα μοντέλα παλινδρόμησης που περιορίζουν την L_1-νόρμα των συντελεστών παλινδρόμησης. Δίνονται δύο εναλλαχτικοί ορισμοί. Ο πρώτος ορισμός καθορίζει τις εκτιμήσεις-lasso $\widehat{\boldsymbol{\beta}}$ των συντελεστών $\boldsymbol{\beta}$ για βελτιστοποίηση της περιορισμένης πιθανοφάνειας (constrained likelihood) ως:

$$\widehat{\boldsymbol{\beta}} = argmax l(\boldsymbol{\beta}), \qquad subject\ to \qquad \| \boldsymbol{\beta} \|_1 = \sum_{i=1}^{p} |\beta_i| \leq s$$

όπου l ο λογάριθμος της πιθανοφάνειας του μοντέλου και $\|.\|_1$ $είναι\ η\ L_1 - norm.$ Ο δεύτερος ορισμός καθορίζει τα $\widehat{\boldsymbol{\beta}}$ για βελτιστοποίηση της ποινικοποιημένης πιθανοφάνειας (penalized likelihood) ως:

$$\widehat{\boldsymbol{\beta}} = argmax\{l(\boldsymbol{\beta}) - \lambda \| \boldsymbol{\beta} \|_1\}$$

Για δοσμένη συνάρτηση πιθανοφάνειας οι πιο πάνω δύο ορισμοί είναι ισοδύναμοι.

Στη συγκεκριμένη περίπτωση του μοντέλου του Cox, ο ορισμός παίρνει την ακόλουθη μορφή:

$$\widehat{\boldsymbol{\beta}} = argmax \left\{ \sum_{j=1}^{p} \left\{ \boldsymbol{\beta}' \boldsymbol{x}_j - \ln \left[\sum_{i \in R_j} e^{\boldsymbol{\beta}' \boldsymbol{x}_i} \right] \right\} - \lambda \parallel \boldsymbol{\beta} \parallel_1 \right\} \tag{3.8}$$

Τώρα για την επίλυση του διπλού αυτού προβλήματος βελτιστοποίησης, που μπορεί να χειριστεί μεγάλων διαστάσεων καταστάσεις, χρησιμοποιείται ο αλγόριθμος του **Gradient ascent (αλγόριθμος της Κλίσης)**. Γενικά θεωρείται μια αναποτελεσματική μέθοδος βελτιστοποίησης, όμως σαν μέθοδος επιλογής για τις εκτιμήσεις-lasso, ισχύει το αντίθετο. Για τη βελτιστοποίηση της ποινικοποιημένης συνάρτησης πιθανοφάνειας (penalized likelihood), αφού ισχύει για κάθε κοίλη και διπλά διαφορίσιμη $l(\boldsymbol{\beta})$, χρησιμοποιείται η μέθοδος αυτή και στην πορεία γίνεται αυτόματη μετάβαση με βήματα του αλγορίθμου Newton-Raphson.

Στόχος μας είναι να καταλήξουμε στη συνάρτηση:

$$l_{pen}(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \lambda \sum_{i=1}^{p} |\beta_i|$$

που είναι άθροισμα δύο όρων, η $l(\boldsymbol{\beta})$ (κοίλη και διπλά διαφορίσιμη) και η συνάρτηση ποινής $P(\boldsymbol{\beta}) = -\lambda \sum_{i=1}^{p} |\beta_i|$ (κοίλη και συνεχής, αλλά διαφορίσιμη μόνο με $\beta_i \neq 0,\ \forall i$).

Οπότε έχουμε την $l_{pen}(\boldsymbol{\beta})$ να είναι κοίλη συνάρτηση σαν άθροισμα κοίλων συναρτήσεων. Όμως δεν είναι διαφορίσιμη παντού, διότι δε γνωρίζουμε τη διαφορισιμότητα της συνάρτησης ποινής. Αλλά ακόμα και έτσι είναι δυνατόν να προσδιοριστεί μια παράγωγος για κάθε συντελεστή $\boldsymbol{\beta}$ για κάθε κατεύθυνση $\boldsymbol{v} \in \mathbb{R}^p$:

$$l'_{pen}(\boldsymbol{\beta}; \boldsymbol{v}) = \lim_{t \downarrow 0} \frac{1}{t} \{ l_{pen}(\boldsymbol{\beta} + t\boldsymbol{v}) - l_{pen}(\boldsymbol{\beta}) \}$$

Έτσι το *gradient* (κλίση) μπορεί να ορίζεται για κάθε $\boldsymbol{\beta}$. Έστω $\boldsymbol{v}_{opt}$ η κατεύθυνση που μεγιστοποιεί τη $l'_{pen}(\boldsymbol{\beta};\boldsymbol{v})$ μεταξύ όλων των $\boldsymbol{v}$ με $\| \boldsymbol{v} \| = 1$, τότε το *gradient* ορίζεται ως:

$$g(\boldsymbol{\beta}) = \begin{cases} l'_{pen}(\boldsymbol{\beta};\boldsymbol{v}_{opt}) \cdot \boldsymbol{v}_{opt}\,, & \text{όταν } l'_{pen}(\boldsymbol{\beta};\boldsymbol{v}_{opt}) \geq 0 \\ \mathbf{0}\,, & \text{διαφορετικά} \end{cases}$$

Το $g(\boldsymbol{\beta}) = (g_1(\boldsymbol{\beta}), \dots, g_p(\boldsymbol{\beta}))'$ μπορεί να υπολογιστεί από την κλίση της μη ποινικοποιημένης συνάρτησης πιθανοφάνειας $h(\boldsymbol{\beta}) = \partial l(\boldsymbol{\beta})/\partial \boldsymbol{\beta} = (h_1(\boldsymbol{\beta}), \dots, h_p(\boldsymbol{\beta}))'$ ως:

$$g_i(\boldsymbol{\beta}) = \begin{cases} h_i(\boldsymbol{\beta}) - \lambda\, sign(\beta_i), & \alpha\nu \quad \beta_i \neq 0 \\ h_i(\boldsymbol{\beta}) - \lambda\, sign\big(h_i(\boldsymbol{\beta})\big), & \alpha\nu\ \beta_i = 0\ \kappa\alpha\iota\ h_i(\boldsymbol{\beta}) > \lambda \\ 0, & \text{διαφορετικά} \end{cases}$$

όπου $sign(x) = \begin{cases} 1, & \alpha\nu\ x > 0 \\ 0, & \alpha\nu\ x = 0 \\ -1, & \alpha\nu\ x < 0 \end{cases}$ και από εδώ φαίνεται ότι η κλίση είναι ασυνεχής σε κάθε σημείο που η ποινικοποιημένη πιθανοφάνεια είναι μη διαφορίσιμη.

Με ανάλογο τρόπο μπορούμε να υπολογίσουμε και τη δεύτερη παράγωγο κατά κατεύθυνση:

$$l''_{pen}(\boldsymbol{\beta};\boldsymbol{v}) = \lim_{t \downarrow 0} \frac{1}{t} \{ l'_{pen}(\boldsymbol{\beta} + t\boldsymbol{v};\boldsymbol{v}) - l'_{pen}(\boldsymbol{\beta};\boldsymbol{v}) \}$$

ακόμα και όταν ο Εσσιανός πίνακας δεν ορίζεται. Έτσι για κάθε $\boldsymbol{\beta}$ και $\boldsymbol{v}$ καταλήγουμε:

$$l''_{pen}(\boldsymbol{\beta};\boldsymbol{v}) = \boldsymbol{v} \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \boldsymbol{v}'$$

Οπότε η κλίση και η δεύτερη παράγωγος κατά κατεύθυνση μας περιγράφουν την ολική συμπεριφορά της $l_{pen}(\boldsymbol{\beta})$ και συνεπώς οι πληροφορίες που παίρνουμε από το *gradient* μπορούν να χρησιμοποιηθούν σε μία σειρά Taylor. Χρησιμοποιώντας, τώρα, τον αλγόριθμο Newton Raphson πετυχαίνουμε μια προσέγγιση κοντά στην αληθινή βέλτιστη

τιμή. Έτσι καταλήγουμε στην ποινικοποιημένη πιθανοφάνεια, που στο μοντέλο του Cox παίρνει την πιο κάτω μορφή:

$$l_{pen}(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \lambda \sum_{i=1}^{p} |\beta_i| = \sum_{j=1}^{p} \left\{ \boldsymbol{\beta}' \boldsymbol{x}_j - \ln \left[\sum_{i \in R_j} e^{\boldsymbol{\beta}' \boldsymbol{x}_i} \right] \right\} - \lambda \sum_{i=1}^{p} |\beta_i| \qquad (3.9)$$

Ο πιο πάνω αλγόριθμος που περιγράφηκε έχει σχεδιαστεί για να βρεθούν οι εκτιμήσεις των συντελεστών παλινδρόμησης για μία μόνο σταθερή τιμή του λ. Σε ορισμένες περιπτώσεις, όμως, είναι ενδιαφέρον να εξεταστεί αν υπάρχει καλύτερη προσαρμογή του μοντέλου για άλλες τιμές του λ. Για τη βελτιστοποίηση της τιμής του λ, η συνηθισμένη κατάλληλη μέθοδος που χρησιμοποιείται είναι η cross validation (CVL), που έχει περιγραφεί πιο πριν (§ 3.2.5). Με τη βελτιστοποίηση της τιμής του λ, πετυχαίνουμε αυτόματα και τη βελτιστοποίηση της προσαρμογής των μεταβλητών στο μοντέλο, δηλαδή την επιλογή του βέλτιστου μοντέλου για τα δεδομένα.

3.2.6.4 Έλεγχος με τη μέθοδο L_1 και L_2-Penalized (Elastic net)

Πιο πάνω περιγράψαμε τη μέθοδο για L_1-Penalized όπως χρησιμοποιείται στο μοντέλο του Cox. Με μικρές τροποποιήσεις μπορεί επίσης να χρησιμοποιηθεί για διαφορετικούς τύπους περιορισμών και ποινών (penalties). Μια επιπλέον ποινή που χρησιμοποιείται είναι η L_2-Penalized.

Σε κάποιες περιπτώσεις που η λύση που δίνει η Lasso είναι πολύ «φτωχή», ένας συνδιασμός της με την L_2-penalized (Ridge regression) δίνει καλύτερα αποτελέσματα. Αυτή είναι η λεγόμενη *(naive)* ***elastic net*** και ορίζεται ως (Zou & Hastie, 2005):

$$\widehat{\boldsymbol{\beta}} = argmax\{l(\boldsymbol{\beta}) - \lambda_1 \parallel \boldsymbol{\beta} \parallel_1 - \lambda_2 \parallel \boldsymbol{\beta} \parallel_2^2\}, \qquad \text{όπου} \quad \|.\|_2 \text{ είναι η } L_2 - norm$$

Στο μοντέλο αναλογικής διακινδύνευσης του Cox, η πιο πάνω παράσταση παίρνει τη μορφή:

$$\widehat{\boldsymbol{\beta}} = argmax \left\{ \sum_{j=1}^{p} \left\{ \boldsymbol{\beta}' \boldsymbol{x}_j - \ln \left[\sum_{i \in R_j} e^{\boldsymbol{\beta}' \boldsymbol{x}_i} \right] \right\} - \lambda_1 \parallel \boldsymbol{\beta} \parallel_1 - \lambda_2 \parallel \boldsymbol{\beta} \parallel_2^2 \right\} \qquad (3.10)$$

Οι εκτιμήσεις της πιο πάνω υπολογίζονται με βάση τον αλγόριθμο, από την προβολή της συνάρτησης που βελτιστοποιείται ως το άθροισμα της διπλά – διαφορίσιμης $l(\boldsymbol{\beta}) - \lambda_2 \parallel \boldsymbol{\beta} \parallel_2^2$ και της ποινής L_1. Εάν η ποινή L_2 είναι σχετικά μεγάλη σε σχέση με την L_1, τότε η επήρεια του συνόλου των συντελεστών στο μοντέλο θα μεγαλώσει πολύ, με αποτέλεσμα να οδηγηθούμε στην αντιστροφή των πολύ μεγάλων πινάκων με τον αλγόριθμο Newton-Raphson.

Στην ειδική περίπτωση του μοντέλου του Cox, που έχουμε $n<<p$, τέτοια αντιστροφή των μεγάλων πινάκων μπορεί να αποφευχθεί με χρήση της τοπικής αναπαραμέτρησης. Η αναπαραμέτρηση αυτή μπορεί να επιτευχθεί λόγω του ειδικού τύπου της κλίσης του $l(\boldsymbol{\beta})$ σε κάποια μοντέλα και δίνεται από τον τύπο:

$$\frac{\partial l(\boldsymbol{\beta})}{\partial(\boldsymbol{\beta})} = \boldsymbol{X}^T \boldsymbol{r}\,,$$

$$\text{όπου } \boldsymbol{X}\text{: } n \times p \text{ πίνακας}, \boldsymbol{r}\text{: } n - \text{διάνυσμα υπολοίπων}$$

Με την κλίση (gradient) να είναι συνεχής, εφαρμόζουμε τη μέθοδο Newton-Raphson χρησιμοποιώντας τον τύπο:

$$\boldsymbol{X}^T \boldsymbol{r} - \lambda_1 sign(\boldsymbol{b}) - \lambda_2 \boldsymbol{b} = \boldsymbol{0}$$

όπου το $sign(\boldsymbol{b})$ είναι σταθερός όρος εντός του τομέα αυτού και συνάγεται αμέσως από τις εκτιμήσεις των συντελεστών $\widehat{\boldsymbol{\beta}}$ που πρέπει να βρίσκεται στη στήλη (n+1)-διαστάσεων του πίνακα $[\boldsymbol{X}^T; sign(\boldsymbol{b})]$, που οδηγεί φυσικά σε αναπαραμέτρηση των συντελεστών $\boldsymbol{\beta}$, σε ένα n+1 διάστασης διάνυσμα. Χρησιμοποιώντας την αναπαραμέτρηση των συντελεστών $\boldsymbol{\beta}$ και χρήση της Newton–Raphson φτάνουμε και πάλι σε μια βέλτιστη τιμή. Όμως εδώ θέλει προσοχή γιατί με τον αλγόριθμο αυτό ποτέ δεν πρέπει να καταλήξουμε σε αποτέλεσμα που να χρειάζεται να έχουμε αντιστροφή πινάκων μεγαλύτερης διάστασης από $(n+1) \times (n+1)$.

3.2.7 ΚΡΙΤΗΡΙΑ ΕΠΙΛΟΓΗΣ ΚΑΤΑΛΛΗΛΟΥ ΜΟΝΤΕΛΟΥ

Άλλες γνωστές τεχνικές για την επιλογή κατάλληλου μοντέλου που χρησιμοποιούνται και στα μοντέλα επιβίωσης, συγκεκριμένα στο μοντέλο του Cox, είναι τα κριτήρια επιλογής κατάλληλου μοντέλου όπως ο συντελεστής προσδιορισμού R^2, αλλά και τα κριτήρια AIC και BIC.

3.2.7.1 Συντελεστής προσδιορισμού R^2 στο μοντέλο του Cox

Στα μοντέλα επιβίωσης, από μελέτες που έχουν γίνει, έχει προταθεί η χρήση του:

$$R_p^2 = 1 - \exp\left\{\frac{2}{n}\left(\hat{l}_0 - \hat{l}_p\right)\right\}, \qquad 0 \le R_p^2 \le 1,$$

όπου $\hat{l}_0$*: η μεγιστοποιημένη λογαριθμοποιημένη πιθανοφάνεια του μοντέλου χωρίς συμμεταβλητές,*

$\hat{l}_p$*: η μεγιστοποιημένη λογαριθμοποιημένη πιθανοφάνεια του μοντέλου με* $p -$ *συμμεταβλητές*

Όπως και σε όλες τις περιπτώσεις των συντελεστών προσδιορισμού R^2, καλύτερο μοντέλο θεωρείται αυτό με τον ψηλότερο συντελεστή. Όπως έχουμε αναφέρει και σε προηγούμενα κεφάλαια, όσο πιο κοντά στη μονάδα είναι η τιμή του συντελεστή τόσο πιο καλό είναι το μοντέλο μου. Όμως στα μοντέλα επιβίωσης το κριτήριο αυτό δε θεωρείται ιδιαίτερα αξιόπιστο, διότι εξαρτάται από το ποσοστό των αποκομμένων παρατηρήσεων. Για το λόγο αυτό δεν το βλέπουμε συχνά, ή και καθόλου, σε συγκριτικές μελέτες μοντέλων διάρκειας ζωής.

3.2.7.2 Κριτήριο AIC στο μοντέλο του Cox

Ένα άλλο σημαντικό κριτήριο επιλογής βέλτιστου μοντέλου είναι το κριτήριο AIC, για το οποίο έχουμε μιλήσει και στην παράγραφο 1.2.3.2, όπου στην περίπτωση των μοντέλων επιβίωσης, συγκεκριμένα στο μοντέλο του Cox, το κριτήριο παίρνει την πιο κάτω μορφή, με p το πλήθος των παραμέτρων του μοντέλου:

$$AIC = -2\sum_{j=1}^{p}\left\{\boldsymbol{\beta}'\boldsymbol{x}_j - \ln\left[\sum_{i\in R_j} e^{\boldsymbol{\beta}'\boldsymbol{x}_i}\right]\right\} + 2p \qquad (3.11)$$

Όπως έχουμε σημειώσει και προηγουμένως για το κριτήριο αυτό, συγκρίνοντας όλα τα υποψήφια μοντέλα, το μοντέλο με το μικρότερο AIC, θεωρείται ως το καταλληλότερο και αυτό προτιμούμε.

3.2.7.3 Κριτήριο BIC στο μοντέλο του Cox

Παρόμοια θεωρία ισχύει και για το κριτήριο BIC, για το οποίο έχουμε μιλήσει στην παράγραφο 1.2.3.3 και στην περίπτωση του μοντέλου αναλογικής διακινδύνευσης του Cox, παίρνει την πιο κάτω μορφή:

$$BIC = -2\sum_{j=1}^{p}\left\{\boldsymbol{\beta}'\boldsymbol{x}_j - \ln\left[\sum_{i\in R_j} e^{\boldsymbol{\beta}'\boldsymbol{x}_i}\right]\right\} + p\ln n \qquad (3.12)$$

όπου p: πλήθος παραμέτρων του μοντέλου και n: το μέγεθος του δείγματος

Συγκρίνοντας όλα τα υποψήφια μοντέλα, το μοντέλο με το μικρότερο BIC είναι το προτιμότερο.

Για τη χρήση του κριτηρίου BIC έχει προταθεί άλλη μία εκδοχή (Volinsky & Raftery, 2000), αντικαθιστώντας το *n* (μέγεθος του δείγματος) με το πλήθος των μη αποκομμένων παρατηρήσεων *k*, οπότε το κριτήριο παίρνει την πιο κάτω μορφή:

$$BIC = -2\sum_{j=1}^{p}\left\{\boldsymbol{\beta}'\boldsymbol{x}_j - \ln\left[\sum_{i\in R_j} e^{\boldsymbol{\beta}'\boldsymbol{x}_i}\right]\right\} + p\ln k$$

όπου p: το πλήθος παραμέτρων του μοντέλου, k: το πλήθος των μη αποκομμένων παρατηρήσεων

Έτσι τώρα ο δεύτερος όρος μπορεί να θεωρηθεί σαν μια ποινή (penalty) για την εισαγωγή πολλών παραμέτρων στο μοντέλο.

3.3 ΕΦΑΡΜΟΓΗ ΣΤΟ ΜΟΝΤΕΛΟ ΤΟΥ COX ΜΕ ΧΡΗΣΗ ΤΗΣ R

3.3.1 ΠΕΡΙΓΡΑΦΗ ΕΦΑΡΜΟΓΗΣ ΚΑΙ ΔΕΔΟΜΕΝΑ

Όπως και στο μοντέλο της λογιστικής παλινδρόμησης έτσι και για το μοντέλο του Cox εκτελούμε την ίδια εφαρμογή, τα ίδια δεδομένα. Δηλαδή και πάλι χρησιμοποιούμε στοιχεία μιας ιατρική μονάδα με ασθενείς που πάσχουν από *οξεία μυελοπλαστική λευχαιμία* (*acute myeloblastic leukaemia*), με δείγμα από 51 ασθενείς (Lee, 1980) με αυτή την αρρώστια, η οποία μπορεί να επιφέρει ακόμα και το θάνατο. Εφαρμόζουμε μια θεραπεία με διαφορετικές συνθήκες σε κάθε ασθενή επηρεασμένη από την ηλικία, θερμοκρασία, δοσοληψίες μέσα στο φάρμακο κλπ. Η περιγραφή των επεξηγηματικών μεταβλητών φαίνεται πιο κάτω (Πίνακας 3.5).

Η διαφορά όμως αυτή τη φορά είναι ότι η μεταβλητή απόκρισης δείχνει την κατάσταση του ασθενή αυτή τη στιγμή, δηλ. αν ο ασθενής είναι ακόμα ζωντανός ή αν έχει πεθάνει (Status: 1 = ακόμα ζωντανός, 0 = έχει πεθάνει), καθώς επίσης μπαίνει και μία νέα επεξηγηματική μεταβλητή που δείχνει το χρόνο επιβίωσης (σε μήνες) μετά τη διάγνωση. Έχουμε μπει πλέον σε μοντέλα επιβίωσης. Το πιο δημοφιλές μοντέλο που χρησιμοποιείται σε τέτοια μοντέλα είναι το μοντέλο αναλογικής διακινδύνευσης του Cox, στο οποίο προσαρμόζουμε και το συγκεκριμένο πρόβλημα.

Μεταβλητές	**Περιγραφή (Description)**
y (status)	Κατάσταση μετά την έρευνα: 1 = ακόμα ζωντανός, 0 = έχει πεθάνει
x1 (age)	Ηλικία ασθενή (σε χρόνια)
x2 (smear)	Ποσοστό επίστρωσης βλαστοκυττάρων (%)
x3 (infltr)	Ποσοστό κυττάρων λευχαιμίας που εισήλθαν στο μυελό των οστών (%)
x4 (lab)	Ποσοστό κυττάρων που προήλθαν από το μυελό των οστών (%)
x5 (blasts)	Αριθμός βλαστοκυττάρων ($x10^3$)
x6 (temp)	Υψηλότερη θερμοκρασία σώματος ($x10\ ^oF$)
x7 (surv)	Χρόνος επιβίωσης από τη στιγμή της διάγνωσης (σε μήνες)

Πίνακας 3.5

Σκοπός μας είναι, χρησιμοποιώντας την R, να προσαρμόσουμε τα δεδομένα με το μοντέλο του Cox και να επιλέξουμε το στατιστικά καταλληλότερο μοντέλο, για την περιγραφή του προβλήματος.

3.3.2 ΕΚΤΕΛΕΣΗ ΕΦΑΡΜΟΓΗΣ ΚΑΙ ΕΠΙΛΟΓΗ ΒΕΛΤΙΣΤΟΥ ΜΟΝΤΕΛΟΥ

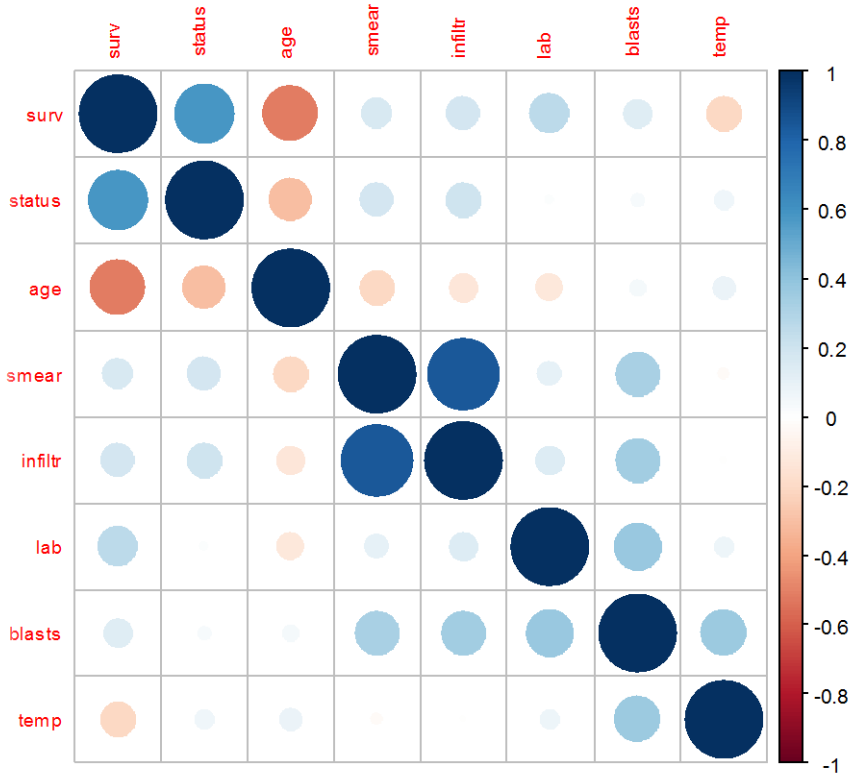
Αρχικά, ορίζουμε τις επεξηγηματικές μεταβλητές του πιο πάνω πίνακα 3.5 (x1 μέχρι x6), όπως και τη μεταβλητή απόκρισης (status), αλλά και το χρόνο επιβίωσης (time). Η απόκριση status, δηλαδή η κατάσταση του ασθενή μετά από το χρόνο επιβίωσης (surv), ορίζεται σαν κατηγορική μεταβλητή, με 1 = ακόμα ζωντανός, 0 = έχει πεθάνει. Οπότε είμαστε έτοιμοι για την προσαρμογή στην R με το μοντέλο αναλογικής διακινδύνευσης του Cox.

Πριν την προσαρμογή, ελέγχουμε και τη συσχέτιση μεταξύ των μεταβλητών αριθμητικά αλλά και γραφικά. Όσο πιο κοντά στην απόλυτο μονάδα είναι οι τιμές, τόσο πιο έντονη θα είναι η μεταξύ τους συσχέτιση. Αν ισούται με μηδέν τότε είναι ασυσχέτιστες.

	age	smear	infiltr	lab	blasts	temp
age	1	-0.203782	-0.136998	-0.124254	0.047105	0.084589
smear	-0.203782	1	0.847132	0.102692	0.325986	-0.028249
infiltr	-0.136998	0.847132	1	0.144377	0.340159	-0.006709
lab	-0.124254	0.102692	0.1443771	1	0.378028	0.070529
blasts	0.047105	0.325986	0.3401596	0.378028	1	0.360247
temp	0.084589	-0.028249	-0.006709	0.070529	0.360247	1

Πίνακας 3.6: Πίνακας συσχέτισης των μεταβλητών

Από τα αποτελέσματα (Πίνακας 3.6, Γράφημα 3.7) φαίνεται πολύ καθαρά ότι υπάρχει πολύ μεγάλη συσχέτιση μεταξύ των μεταβλητών smear και infiltr, καθώς η τιμή της μεταξύ τους συσχέτισης corr (smear,infiltr) = 0.8471 είναι πολύ κοντά στη μονάδα.



Γράφημα 3.7: Η συσχέτιση μεταξύ των μεταβλητών

Στη συνέχεια γίνεται η προσαρμογή με το μοντέλο του Cox, με τη συνάρτηση επιβίωσης να εξαρτάται από το χρόνο επιβίωσης (time) και την κατάσταση του ασθενούς μετά τη θεραπεία (status) σαν κατηγορική μεταβλητή (με τιμές 1 και 0). Εισάγουμε τις ακόλουθες εντολές στην R για να λάβουμε τα Αποτελέσματα 3.1 για την προσαρμογή του μοντέλου:

```
> results_cox<-glm(resp~age+smear+infiltr+lab+blasts+temp,family=binomial)
> summary(results_glm)
```

Από τον Πίνακα 3.7 φαίνονται και τα συμμετρικά 95% διαστήματα εμπιστοσύνης για όλες τις παραμέτρους του μοντέλου (α, β1, β2, …).

```
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
```

Αποτελέσματα 3.1

	2.5%	97.5%
age	-0.0224815	0.3204791
smear	-0.2118578	0.1023355
infiltr	-0.1955163	0.1604711
lab	-0.8879738	0.4118787
blasts	-0.0668800	0.3169543
temp	-0.2088806	-0.0368529

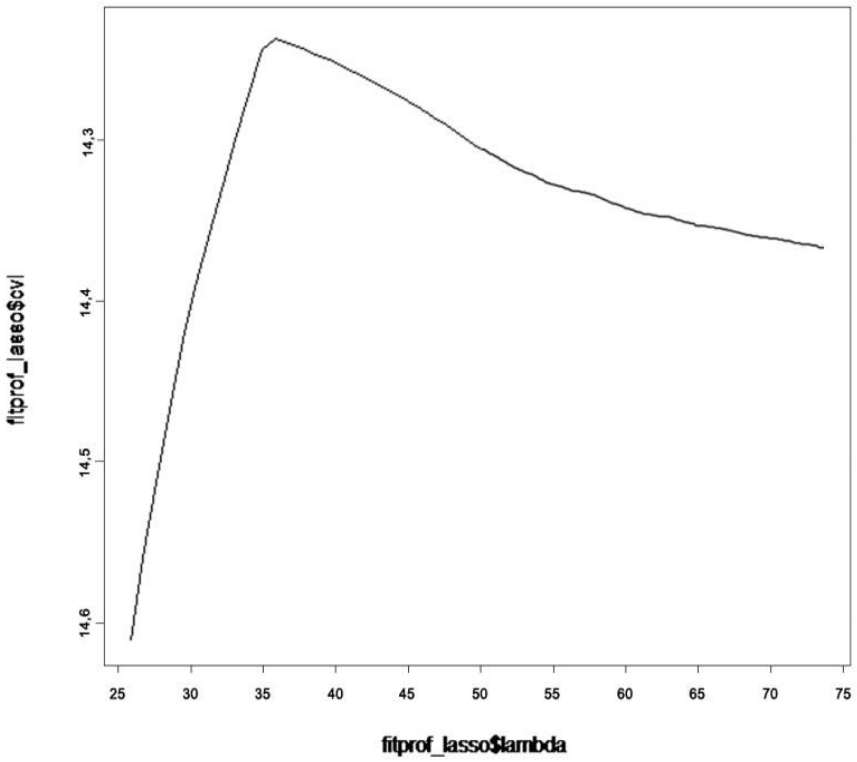
Πίνακας 3.7: 95% Διαστήματα εμπιστοσύνης

Βλέποντας τα πιο πάνω αποτελέσματα, του προσαρμοσμένου μοντέλου με όλες τις μεταβλητές, παρατηρούμε από τις p-τιμές των ελέγχων Wald ότι μόνο η επεξηγηματική μεταβλητή age είναι στατιστικά σημαντική με p-τιμή 0.0020. Μπορούμε να πούμε επίσης ότι οριακά σημαντική είναι και η μεταβλητή lab με p-τιμή 0.0658.

Οπότε ένας τρόπος να βελτιώσουμε το μοντέλο είναι να χρησιμοποιήσουμε την τεχνική Lasso και προσπαθούμε να βρούμε το βέλτιστο λ_1 με τη μέθοδο CVL (Cross Validation) (§ 3.2.5), ώστε να το χρησιμοποιήσουμε στην τεχνική αυτή. Με τη βοήθεια των πιο κάτω εντολών έχουμε, εκτός από αριθμητικά και γραφικά αποτελέσματα:

```
> fitprof_lasso<-profL1 (Surv(surv,status)~age+smear+infiltr+lab+blasts+temp)
> plot (fitprof_lasso$lambda,fitprof_lasso$cvl,type="l")

> opt_lasso<- optL1(Surv(surv,status)~age+smear+infiltr+lab+blasts+temp,
+fold=fitprof_lasso$fold, minlambda1=1, maxlambda1=75)
```



Γράφημα 3.8: Το βέλτιστο λ_1

Οπότε καταλήγουμε ότι το βέλτιστο λ_1 για τη Lasso να είναι ίσο με 29, το οποίο υπολογίζεται εκτός από το Γράφημα 3.8 και από την εκτέλεση των πιο πάνω εντολών στην R (Παράρτημα D.4). Έτσι τώρα κάνουμε τον έλεγχο με τη μέθοδο LASSO για το συγκεκριμένο λ_1 χρησιμοποιώντας το πακέτο Penalized (Παράρτημα D.1), εκτελώντας:

```
> fit_final_lasso<-penalized(Surv(surv,status)~age+smear+infiltr+lab+blasts+temp, dataAML,
lambda1=29)
> coefficients(fit_final_lasso,"penalized")
```

Έτσι, καλώντας και την εντολή για τις μεταβλητές (coefficients), εμφανίζονται όλες οι penalized-παραμέτροι του μοντέλου για να ελέγξουμε ποιες έχουν γίνει μηδέν ή κοντεύουν στο μηδέν και ποιες όχι. Αυτές που δεν είναι κοντά στο μηδέν, αλλά δεν είναι και ακριβώς μηδέν είναι οι καταλληλότερες για το βέλτιστο μοντέλο που θέλουμε.

```
      age          infiltr          lab          temp
0.027846210 -0.008259883 -0.026284444 0.008924687
```

Αποτελέσματα 3.2

Παρατηρούμε ότι (Αποτελέσματα 3.2), μετά την εκτέλεση όλων των βημάτων της τεχνικής Lasso, μπορούμε να πούμε ότι οι σημαντικότερες μεταβλητές είναι οι age, infiltr, lab και temp με τιμές 0.0278, -0.0083, -0.0263 και 0.0089, αντίστοιχα.

Μετά τους ελέγχους και την τεχνική της Lasso επιλέγουμε ως πιο σημαντικές τις μεταβλητές age και temp. Όμως εκτελώντας:

```
> cox_mod1<-coxph (formula=Surv(surv,status) ~ age + temp)
> summary(cox_mod1)
```

θα πάρουμε τα Αποτελέσματα 3.3 και παρατηρούμε ότι μόνο η μεταβλητή age είναι στατιστικά σημαντική, όπως φαίνεται από τον έλεγχο Wald, με p-τιμή = 0.000934. Επίσης, με τη βοήθεια του ελέγχου του λόγου των πιθανοφανειών (§ 3.2.6.1) συγκρίνουμε την τιμή της $-2(\hat{l}_0 - \hat{l}_1)$ με τη X_1^2 κατανομή. Έτσι λοιπόν από το μοντέλο που περιέχει την temp (Αποτελέσματα 3.3, βλ. τα bold) και από το μοντέλο που την έχουμε πετάξει έξω (Αποτελέσματα 3.4, βλ. τα bold), παίρνουμε τις τιμές των πιθανοφανειών (likelihood ratio), οι οποίες είναι 12.7 και 11.85, αντίστοιχα. Άρα θα έχουμε $-2(11.85 - 12.7) = 1.7$ και συγκρίνουμε την τιμή αυτή με τη X_1^2 κατανομή και βρίσκουμε την p-τιμή του ελέγχου να ισούται με 0.807 > 0.05, οπότε σημαίνει έχουμε σοβαρές ενδείξεις για να μην απορρίψουμε τη μηδενική υπόθεση (H_0) και να θεωρήσουμε τη temp ως στατιστικά μη σημαντική για το βέλτιστο μοντέλο.

Άλλες σοβαρές ενδείξεις που έχουμε για την εγκυρότητα των αποτελεσμάτων είναι οι p-τιμές των ελέγχων (Αποτελέσματα 3.3, στο κάτω μέρος). Οπότε καταλήγουμε ότι το

«καταλληλότερο» μοντέλο, στα δεδομένα της συγκεκριμένης εφαρμογής, είναι το μοντέλο που περιέχει μόνο τη μεταβλητή age.

```
Call:
coxph(formula = Surv(surv, status) ~ age + temp)

  n= 51

         coef     exp(coef)  se(coef)      z      Pr(>|z|)
age   0.03160   1.03408   0.009549  3.310   0.000934 ***
temp 0.01009   1.01014   0.010684  0.944   0.344958
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

       exp(coef) exp(-coef)  lower .95   upper .95
age      1.032    0.9689       1.0130       1.052
temp    1.010    0.9900       0.9892       1.032

Rsquare                  =   0.2   (max possible= 0.996)
Likelihood ratio test = 12.7    on 2 df,   p=0.001747
Wald test                = 12.38  on 2 df,    p=0.002047
Score (logrank) test  = 13.22   on 2 df,   p=0.001345
```

Αποτελέσματα 3.3

3.3.3 ΤΟ ΒΕΛΤΙΣΤΟ ΜΟΝΤΕΛΟ

Παρουσιάζουμε την ανάλυση παλινδρόμησης του μοντέλου που επιλέχτηκε ως το βέλτιστο, που περιέχει μόνο τη μεταβλητή age με τη βοήθεια των εντολών:

```
> cox_teliko<-coxph(formula=Surv(surv,status) ~ age)
> summary(cox_teliko)
```

Όπως παρατηρούμε από τα Αποτελέσματα 3.4, η μεταβλητή age όντως μπορεί να θεωρηθεί ως στατιστικά σημαντική με την p-τιμή του ελέγχου Wald να ισούται με 0.00067 και είναι πολύ μικρή.

```
Call:
coxph(formula = Surv(surv, status) ~ age)

  n= 51

          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

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
```

Αποτελέσματα 3.4

Μετά την πιο πάνω «ολόκληρη» διαδικασία, καταλήγουμε στο εξής μοντέλο:

$$h(t;\boldsymbol{x}) = h_0(t)\mathrm{e}^{\boldsymbol{\beta'}\boldsymbol{x}_i} = h_0(t)\exp(0.0324age) = h_0(t) * 1.033age$$

Κατ' αρχάς εξηγούμε ότι οι τιμές του $\mathrm{e}^{\boldsymbol{\beta'}}$ δείχνουν κατά πόσο πολλαπλασιάζεται η συνάρτηση διακινδύνευσης, δηλαδή πόσο επιδρά μια συμμεταβλητή στη διάρκεια ζωής, με τις υπόλοιπες συμμεταβλητές να παραμένουν σταθερές. Έτσι στο συγκεκριμένο τελικό μοντέλο που έχουμε καταλήξει, θα έχουμε ότι για ένα επιπλέον έτος στην ηλικία κάποιου ασθενή (age) να επηρεάζει το μοντέλο $(h_0(t) * 1.033)$, με τις υπόλοιπες συνθήκες να παραμένουν σταθερές. Όμως η αύξηση του ποσοστού της επίστρωσης των βλαστοκυττάρων (smear), η αύξηση του ποσοστού των κυττάρων λευχαιμίας που εισήλθαν στο μυελό των οστών (infiltr), η αύξηση του ποσοστού των κυττάρων λευχαιμίας που προήλθαν από το μυελό των οστών (lab), αλλά και η αύξηση του αριθμού των βλαστοκυττάρων (blasts) κατά 1000 δεν επηρεάζουν την τελική κατάσταση του ασθενή. Τέλος, ούτε η μεταβολή της θερμοκρασίας επηρεάζει την τελική κατάσταση του ασθενή.

3.3.4 ΣΥΜΠΕΡΑΣΜΑΤΑ

Η ανάλυση των δεδομένων διάρκειας ζωής είναι αρκετά πολύπλοκη λόγω του ότι μεγάλη σημασία παίζει και ο χρόνος επιβίωσης για όλες τις παρατηρήσεις. Η προσαρμογή με το μοντέλο αναλογικής διακινδύνευσης του Cox είναι αρκετά βοηθητική στα μοντέλα επιβίωσης, όμως, όπως και η επιλογή του βέλτιστου μοντέλου στον τομέα αυτό είναι αρκετα πολύπλοκη και με όχι πάντα αρκετά ακριβή αποτελέσματα.

Στην εφαρμογή που εκτελέσαμε στην παρούσα εργασία, αν και είχαμε μικρό δείγμα (n = 51 παρατηρήσεις), αλλά και λίγες επεξηγηματικές μεταβλητές (σύνολο έξι), εντούτοις ο σκοπός επιτεύχθηκε. Αναλύσαμε αρκετά καλά σε θεωρητικό επίπεδο αρκετές από τις τεχνικές και μεθόδους επιλογής μοντέλου στα μοντέλα επιβίωσης, λάβαμε υπόψη μας κάποια από τα κριτήρια επιλογής, αλλά και μερικούς από τους ελέγχους υποθέσεων. Τέλος, χρησιμοποιήσαμε στην εφαρμογή και την αρκετά διαδεμένη τα τελευταία χρόνια τεχνική της Lasso.

ΠΑΡΑΡΤΗΜΑ

Α. Το στατιστικό πακέτο της R

ΤΙ ΕΙΝΑΙ Η R;

Το **R** (ή η R) (R Development Core Team, 2010) είναι ένα ολοκληρωμένο περιβάλλον εργασίας για στατιστικούς υπολογισμούς και γραφήματα. Είναι μια γλώσσα προγραμματισμού που χρησιμεύει κυρίως για ανάλυση δεδομένων και εφαρμογή διαφόρων «κλασσικών» και σύγχρονων στατιστικών τεχνικών.

Το R μας εφοδιάζει μεταξύ άλλων, με:

- Αποτελεσματικό χειρισμό και αποθήκευση δεδομένων .
- Χειρισμό πινάκων πολλών διαστάσεων.
- Μία απλή και αποτελεσματική γλώσσα προγραμματισμού (R), με διεπαφές με άλλες γλώσσες και δυνατότητες αποσφαλμάτωσης .
- Εργαλεία ανάλυσης δεδομένων και δημιουργίας γραφημάτων .

ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

Η R χρησιμοποιεί διερμηνευτή (interpreter) και όχι μεταφραστή (compiler).

Η R είναι “case sensitive” δηλαδή κάνει διαχωρισμό μεταξύ μικρών και κεφαλαίων χαρακτήρων.

Η R είναι μία διάλεκτος της S, η οποία χρησιμοποιείται ευρέως στη στατιστική κοινότητα.

Η σύνταξη της γλώσσας έχει μία επιφανειακή σχέση με την C, αλλά η σημασιολογία της είναι της FPL (Functional Programming Language = Συναρτησιακής Γλώσσας Προγραμματισμού), με μεγάλη συγγένεια με τις γλώσσες Lisp και APL.

Η γλώσσα R είναι ελεύθερα διαθέσιμη στο διαδίκτυο (http://www.r-project.org) και η υποστήριξή της γίνεται με εθελοντική συνεισφορά. Υπάρχουν πολλά διαθέσιμα πακέτα

που υποστηρίζονται από την R και τα οποία υλοποιούν πολλές κλασσικές και μοντέρνες στατιστικές μεθόδους. Πληροφορίες στο: http://CRAN.R-project.org

Η R μπορεί να τρέξει σε διαφορετικές πλατφόρμες (πχ. Windows, Linux, Mac OS).

Η R έχει την ιδιότητα να επιτρέπει τον «υπολογισμό στη γλώσσα», δηλαδή παρέχει την δυνατότητα να γράφουμε συναρτήσεις οι οποίες δέχονται εκφράσεις σαν είσοδο, πράγμα πολύ χρήσιμο για στατιστικά μοντέλα. Μπορεί να χρησιμοποιηθεί είτε με κατευθείαν εντολές, είτε με προγράμματα που ο χρήστης προγραμματίζει για επίλυση των πολύπλοκων στατιστικών προβλημάτων. Επίσης, ο χρήστης μπορεί να χησιμοποιήσει και έτοιμα προγράμματα που είναι ενσωματωμένα μέσα σε πακέτα, τα οποία διατίθενται και πάλι ελεύθερα. Η ποικιλία τέτοιων προγραμμάτων είναι τεράστια.

Η R λειτουργεί στη στατιστική ανάλυση με τη μορφή σειράς βημάτων. Σε κάθε βήμα αποθηκεύει τα αποτελέσματα σε αντικείμενα για περαιτέρω ανάλυση, ενώ τα άλλα στατιστικά πακέτα, όπως το SPSS, δίνουν ένα συνολικό τελικό αποτέλεσμα στην έξοδο. Αυτή είναι και η διαφορά της από τα άλλα στατιστικά πακέτα.

Περισσότερες πληροφορίες στην ιστοσελίδα: http://www.r-project.org

ΤΡΟΠΟΣ ΧΡΗΣΗΣ

Η R λειτουργεί διαλογικά. Γράφουμε μετά το σήμα ετοιμότητας (το σήμα ετοιμότητας στα Windows είναι το " > ", ενώ στο Unix " $ ") την εντολή και πατάμε "Enter". Μετά το πάτημα του πλήκτρου "Enter" ο κώδικας απασφαλματώνεται, εκτελείται και εμφανίζεται στη συνέχεια η απάντηση. Σε περίπτωση που ο κώδικας είναι πάνω από μία γραμμή, μετά τη πρώτη γραμμή αλλάζει από " > " στο σύμβολο " + " για τις επόμενες γραμμές.

Η R μπορεί να χρησιμοποιηθεί με δύο τρόπους. Ο πρώτος τρόπος χρήσης είναι να εκτελούνται οι εκφράσεις από τη γραμμή εντολών (πολλοί χρήστες πιθανόν να μην απαιτούν κάτι παραπάνω από αυτό το επίπεδο), ενώ ο δεύτερος τρόπος είναι μία μέθοδος κατά την οποία γράφουμε με κώδικα συναρτήσεις του χρήστη.

Β. ΕΦΑΡΜΟΓΗ ΓΕΝΙΚΟΥ ΓΡΑΜΜΙΚΟΥ ΜΟΝΤΕΛΟΥ ΜΕ ΧΡΗΣΗ ΤΗΣ R

1. ΜΕΡΙΚΕΣ ΕΝΤΟΛΕΣ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΟΥΜΕ

lm (.): Είναι η εντολή που καλεί τη συνάρτηση *lm*, η οποία προσαρμόζει τις μεταβλητές (τα δεδομένα) σε ένα γραμμικό μοντέλο (linear model), εξού και το όνομά της. π.χ. > lm (y~x1+x2+x3), όπου *y* τα δεδομένα για τη μεταβλητή απόκρισης και $x_1,\ x_2, \ldots$ τα δεδομένα των επεξηγηματικών μεταβλητών.

summary (.): Παρουσιάζει τα αποτελέσματα της ανάλυσης παλινδρόμησης, του μοντέλου που έχει προσαρμοστεί. Δίνει τους περιγραφικούς δείχτες των υπολοίπων, τους εκτιμητές $\hat{\alpha},\ \hat{\beta}_1,\ \hat{\beta}_2, \ldots$, τα τυπικά σφάλματα (standard errors) $se(\hat{\alpha}),\ se(\hat{\beta}_1),\ se(\hat{\beta}_2), \ldots$, αλλά και τον t-έλεγχο του καθενός με την p-τιμή του, αντίστοιχα. Επίσης δίνει το residual standard error ($s_{y|x}$) με τους βαθμούς ελευθερίας, το συντελεστή προσδιορισμού R^2, αλλά και το διορθωμένο R^2_{adj}. Τέλος δίνει τον έλεγχο-F με τους βαθμούς ελευθερίας και την p-τιμή του.

AIC (.): Επιστρέφει την τιμή του κριτηρίου AIC του μοντέλου που προσαρμόστηκε.

AIC (. , k=log(n)): Επιστρέφει την τιμή του κριτηρίου BIC του μοντέλου που προσαρμόστηκε, με n το σύνολο των δεδομένων.

confint (.): Η εντολή αυτή επιστρέφει τα 95% διαστήματα εμπιστοσύνης των εκτιμητριών των μεταβλητών.

step (. , direction = "."): Επιστρέφει το βέλτιστο μοντέλο (κάνοντας επιλογή μεταβλητών) εκτελώντας διαδικασία επιλογής με βήματα. Στο direction μπορούμε να βάλουμε έιτε "backward", είτε "forward" για την αντίστοιχη διαδικασία, είτε "both" για να χρησιμοποιήσει συνδυασμό των δύο. Και οι τρεις αυτές διαδικασίες βασίζονται στο κριτήριο AIC.

profL1 ("μοντέλο" , fold = "."): Αυτή η εντολή μπορεί να χρησιμοποιηθεί για να εξετάσει την επίδραση των παραμέτρων λ1 και λ2 για τη συνάρτηση CVL. Πιο συγκεκριμένα η λειτουργία αυτή μπορεί να χρησιμοποιηθεί για να μεταβάλλει το λ1, κρατώντας σταθερή τιμή του λ2. Η παράμετρος fold μπορεί να χρησιμεύσει ως βάση για την επόμενη κλήση της μεθόδου CVL, ώστε να εξασφαλιστεί η συγκρισιμότητα.

profL2 ("μοντέλο" , fold = "." , minl = ".", maxl = "."): Αυτή η εντολή μπορεί να χρησιμοποιηθεί για να εξετάσει την επίδραση των παραμέτρων λ1 και λ2 για τη συνάρτηση CVL. Πιο συγκεκριμένα η λειτουργία αυτή μπορεί να χρησιμοποιηθεί για να μεταβάλλει το λ2, κρατώντας σταθερή την τιμή του λ1.

optL1 ("μοντέλο" , fold = "."): Επιστρέφει το βέλτιστο λ1, που θα χρησιμοποιηθεί σαν παράμετρο ρύθμισης για τη μέθοδο Lasso.

optL2 ("μοντέλο" , fold = "."): Επιστρέφει το βέλτιστο λ2, που θα χρησιμοποιηθεί σαν παράμετρο ρύθμισης για τη μέθοδο της ridge regression.

penalized ("μοντέλο" , data, lambda1 = ".", lambda2 = "."): Με την εντολή αυτή εφαρμόζονται οι τεχνικές επιλογής μοντέλου Lasso, Ridge regression και E-net. Αν βάλουμε την παράμετρο lambda1 = "." εκτελείται η Lasso, αν βάλουμε την παράμετρο lambda2 = "." εκτελείται η Ridge regression και αν βάλουμε και τις δύο παραμέτρους, εκτελείται η Elastic net. (Περισσότερα για την εντολή αυτή αναφέρονται στο Παράρτημα D.1)

2. ΠΡΟΣΑΡΜΟΓΗ ΜΟΝΤΕΛΟΥ

Πιο κάτω η διαδικασία που ακολουθούμε στην R:

Διαβάζουμε τα δεδομένα μας γραμμή-γραμμή (byrow=T), από το αρχείο, σαν πίνακα, με 13 γραμμές και 28 στήλες.

```
> dataNW<-matrix (scan("BUILDINGnerula_wellington.txt"),ncol=28,byrow=T)
Read 364 items
```

Ορίζουμε τις επεξηγηματικές μεταβλητές (x1 μέχρι x11) και τη μεταβλητή απόκρισης y (PRICE) (δεν διαβάζει την $1^{η}$ στήλη γιατί είναι ο αύξων αριθμός):

```
> taxes<-dataNW[2,]
> baths<-dataNW[3,]
> lotSize<-dataNW[4,]
> livSpace<-dataNW[5,]
> garages<-dataNW[6,]
> rooms<-dataNW[7,]
> bedrooms<-dataNW[8,]
> age<-dataNW[9,]
> constr<-dataNW[10,]
> style<-dataNW[11,]
> fireplaces<-dataNW[12,]

> PRICE<-dataNW[13,]
```

Ορίζουμε την constr και style μεταβλητή σαν κατηγορικές μεταβλητές, όπου για την constr: level 1 = brick, level 2 = brick and frame, level 3 = aluminum and frame, level 4 = frame και για τη style: level 1 = two story, level 2 = one and a half story, level 3 = ranch.

```
> constr<-as.factor(constr)
> constr
 [1] 3 1 2 4 3 2 2 1 2 4 1 1 1 2 4 1 4 1 2 3 4 1 1 4 1 1 4 3
Levels: 1 2 3 4
> style<-as.factor(style)
> style
 [1] 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 3 1 1 1 1 1 3 1 1
Levels: 1 2 3
```

Γραφική περιγραφή των μεταβλητών που έχουμε:

Οι κατηγορικές μεταβλητές constr (construction type) και style που περιγράφονται με boxplots. Με τις ακόλουθες εντολές έχουμε τα Γραφήματα 1.1.

```
> layout(matrix(1:2,nrow=2))
> boxplot(PRICE~constr,ylab="Sale price of the home", xlab="Construction type")
> boxplot(PRICE~style,ylab="Sale price of the home", xlab="Style")
```

Περιγράφονται με scatterplots οι υπόλοιπες ποσοτικές μεταβλητές (Γραφήματα 1.2):

```
> layout(matrix(1:9,nrow=3))
> plot(PRICE~taxes,ylab="Sale price of the home(x 1000 $)",xlab="Taxes(x 100 $)")
> plot(PRICE~baths,ylab="Sale price of the home(x 1000 $)",xlab="Number of baths")
> plot(PRICE~lotSize,ylab="Sale price of the home(x 1000 $)",xlab="Lot size(x 1000 sf)")
> plot(PRICE~livSpace,ylab="Sale price of the home(x 1000 $)",xlab="Living space(x 1000 sf)")
> plot(PRICE~garages,ylab="Sale price of the home(x 1000 $)",xlab="Number of garages")
> plot(PRICE~rooms,ylab="Sale price of the home(x 1000 $)",xlab="Number of rooms")
> plot(PRICE~bedrooms,ylab="Sale price of the home(x 1000 $)",xlab="Number of bedrooms")
> plot(PRICE~age,ylab="Sale price of the home(x 1000 $)",xlab="Age of the home")
> plot(PRICE~fireplaces,ylab="Sale price of the home(x 1000 $)",xlab="Number of fire places")
```

Συσχέτιση μεταξύ των μεταβλητών:

Ελέγχουμε τη συσχέτιση μεταξύ των μεταβλητών, αριθμητικά αλλά και σχηματικά. Για τον Πίνακα 1.2, πρέπει να εκτελέσουμε τις πιο κάτω εντολές, αφού πρώτα μετατρέψουμε τον πίνακα δεδομένων σε πλαίσιο δεδομένων (data frame).

```
>dataNW<-data.frame  (cbind(PRICE,taxes,baths,lotSize,livSpace,garages,rooms,bedrooms,age,
+constr, style, fireplaces))
> corrNW<-cor(dataNW)
> corrNW
```

Για να έχουμε το Γράφημα 1.3:

```
> library(graphics)
> library(corrplot)
> corrplot(corrNW)
```

Προσαρμόζουμε στο γενικό γραμμικό μοντέλο:

```
>resultsNW<-lm(PRICE~taxes+baths+lotSize+livSpace+garages+rooms+bedrooms+age+constr+
style+fireplaces)
```

Παρουσιάζουμε τα αποτελέσματα:

```
> summary (resultsNW)
```

Το κριτήριο του AIC για το πιο πάνω μοντέλο:

```
> AIC (resultsNW)
[1] 163.8215
```

Και το κριτήριο BIC:

```
> n<-28
> AIC (resultsNW,k=log(n))
[1] 185.1368
```

Υπολογίζονται τα συμμετρικά 95% διαστήματα εμπιστοσύνης για όλες τις παραμέτρους με τη βοήθεια της εντολής:

```
> confint(resultsNW)
```

3. ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ

Οι 3 κατά βήματα διαδικασίες επιλογής μοντέλου με βάση το κριτήριο AIC:

Πριν ξεκινήσουμε τις διαδικασίες, θέτουμε το κενό μοντέλο:

```
> null<-lm(PRICE~1,dataNW)
> null
Call:
lm(formula = PRICE ~ 1, data = dataNW)

Coefficients:
(Intercept)
   38.16
```

Αλλά και το πλήρες μοντέλο, με βάση τα δεδομένα μας:

```
> full=lm(PRICE~., dataNW)
> full
Call:
lm(formula = PRICE ~ ., data = dataNW)

Coefficients:
 (Intercept)        taxes        baths      lotSize     livSpace      garages
     2.03596      0.72807      9.64742      0.16392     13.67201      1.98896
       rooms     bedrooms          age       constr        style   fireplaces
    -0.99250     -0.44981     -0.07208      1.02298      1.45404      2.87558
```

- Με τη διαδικασία της διαδοχικής αφαίρεσης (Backward elimination).

```
> step (full, data, direction="backward")
```

- Με τη διαδικασία της διαδοχικής πρόσθεσης (Forward selection).

```
> step (null, scope=list(lower=null, upper=full), direction="forward")
```

- Με τη διαδικασία της κατά βήματα πίσω-εμπρός επιλογής (Stepwise selection).

```
> step (null, scope=list(upper=full), direction="both")
```

Η μέθοδος L_1-Penalized (LASSO):

Βρίσκουμε το βέλτιστο λ1 με τη μέθοδο CVL, ώστε να το χρησιμοποιήσουμε στην τεχνική Lasso.

```
> fitprof_lasso <- profL1 (PRICE, penalized = dataNW[,2:12], fold=10, minl=0.01, maxl=1000)
> plot (fitprof_lasso$lambda,fitprof_lasso$cvl,type="l")

> opt_lasso <-optL1 (PRICE, penalized = dataNW[,2:12], fold=fitprof_lasso$fold)

> opt_lasso$lambda
[1] 98.07917
> opt_lasso$cvl
[1] -87.80132
```

Προσαρμόζουμε με τη Lasso, με $\lambda_1 = 98$, χρησιμοποιώντας το πακέτο Penalized:

```
> fit_lasso<-penalized (PRICE~taxes+baths+lotSize+livSpace+garages+rooms+bedrooms+age
+constr +style +fireplaces,  dataNW,  lambda1=98)
# nonzero coefficients: 3

> show(fit_lasso)
Penalized linear regression object
17 regression coefficients of which 3 are non-zero

Loglikelihood =  -87.80132
L1 penalty =      418.91540      at lambda1 =  98

> coefficients(fit_lasso,"all")       ή     > coefficients(fit_lasso,"penalized")
```

Η μέθοδος L_2-Penalized (Ridge regression):

Βρίσκουμε το βέλτιστο λ_2 με τη μέθοδο CVL, ώστε να το χρησιμοποιήσουμε στη μέθοδο αυτή.

```
> fitprof_ridge<-profL2 (PRICE, penalized=dataNW[2:12], fold=fitprof_lasso$fold, minl=0.01,
+maxl=2000)
> plot(fitprof_ridge$lambda,fitprof_ridge$cvl,type="l",log="x")
> plotpath(fitprof_ridge$fullfit, log="x")

> opt_ridge<-optL2 (PRICE, penalized=dataNW[,2:12], fold=fitprof_ridge$fold)

> opt_ridge$lambda
[1] 35.82175
> opt_ridge$cvl
[1] -93.02811
```

Προσαρμόζουμε με τη Ridge, για $\lambda_2 = 35$, χρησιμοποιώντας το πακέτο Penalized:

```
> fit_ridge <-penalized (PRICE~taxes+baths+lotSize+livSpace+garages+rooms+bedrooms+age
+constr +style+fireplaces, dataNW, lambda2=26)

> show (fit_ridge)
Penalized linear regression object
17 regression coefficients of which 17 are non-zero

Loglikelihood =  -98.02811
L2 penalty =      183.5961       at lambda2 =  35

> coefficients (fit_ridge,"all")     ή     > coefficients (fit_ridge,"penalized")
```

4. ΠΕΡΙΓΡΑΦΗ ΒΕΛΤΙΣΤΟΥ ΜΟΝΤΕΛΟΥ

Πιο κάτω παρουσιάζουμε την ανάλυση παλινδρόμησης του βέλτιστου μοντέλου που επιλέχτηκε:

```
> results_teliko<-lm(PRICE ~ taxes+baths+livSpace)
> summary(results_teliko)
```

Με 95% διαστήματα εμπιστοσύνης:

```
> confint(results_teliko)
```

Προυποθέσεις γενικού γραμμικού μοντέλου:

- Κανονικότητα υπολοίπων:

  ```
  > qqnorm(residuals(results_teliko))
  > qqline(residuals(results_teliko))
  ```

- Ομοσκεδαστικότητα:

  ```
  > plot(results_teliko$res, results_teliko$fitted)
  ```

- Ανεξαρτησία υπολοίπων:

  ```
  > plot(1:28, results_teliko$res)
  ```

C. ΕΦΑΡΜΟΓΗ ΣΕ ΜΟΝΤΕΛΟ ΤΗΣ ΛΟΓΙΣΤΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΣΤΗΝ R

1. ΤΟ ΠΑΚΕΤΟ GLMULTI (GLMULTI-PACKAGE)

Εισαγωγή

Ένα από τα πακέτα της γλώσσας R που θα χρησιμοποιήσουμε και θα μας βοηθήσει για την εφαρμογή με προσαρμογή λογιστικής παλινδρόμησης, είναι το ***πακέτο-glmulti***. Είναι ένα πακέτο της R που χρησιμοποιείται για την αυτόματη επιλογή μοντέλου, συγκεκριμένα του γενικευμένου γραμμικού, εξού και η ονομασία της κύριας συνάρτησης που χρησιμοποιεί ***glm*** (generalized linear model). Παίρνοντας μια μεγάλη ομάδα μεταβλητών και ενεργοποιώντας τη συνάρτηση *glm*, δημιουργεί τα πιθανά μοντέλα που περιέχουν τις μεταβλητές αυτές, αλλά και τις μεταξύ τους αλληλεπιδράσεις, ώστε να καταλήξει, με αυτόματο έλεγχο, στο βέλτιστο συνδιασμό μεταβλητών, δηλαδή το κατάλληλο μοντέλο, με βάση κάποιο κριτήριο. Περιορισμοί για τα υποψήφια μοντέλα μπορούν να καθοριστούν εξαιρώντας συγκεκριμένους συντελεστές της παλινδρόμησης (μεταβλητές), για τον έλεγχο της πολυπλοκότητας του μοντέλου.

Επειδή τα γενικευμένα γραμμικά μοντέλα τα τελευταία χρόνια χρησιμοποιούνται ευρέως σε πολλούς τομείς των επιστημών και όχι μόνο, φτιάχτηκε το πακέτο glmulti (Calcagno & Mazancourt, 2010), το οποίο χρησιμοποιεί τη συνάρτηση glm για την προσαρμογή των γενικευμένων γραμμικών μοντέλων. Η συνάρτηση αυτή έχει ως κύριο εγχείρημα την προσαρμογή του μοντέλου σε μια φόρμουλα: π.χ. $y \sim f_1 + c_1$, όπου y η εξαρτημένη μεταβλητή, f_1 ένας συντελεστής και c_1 η συμμεταβλητή, ή πιο πολύπλοκο παράδειγμα $y \sim f_1 + c_1 + f_1 : c_1$ όπου εδώ προστίθεται και η αλληλεπίδραση των δύο μεταβλητών. Αυτές οι φόρμουλες αναφέρονται σε συγκεκριμένο μοντέλο που με τη βοήθεια της συνάρτησης αυτής προσαρμόζει τα δεδομένα. Μετά τη διαδικασία προσαρμογής, με μία stepwise επιλογή μεταβλητών με βάση το κριτήριο AIC_c συνήθως, προσπαθεί να επιλέξει το «καλύτερο» μοντέλο, κάνοντας όλους τους συνδυασμούς που μπορεί να γίνουν, αυτόματα.

Χρησιμοποιεί τη μέθοδο Stepwise Selection

Σε πολλές περιπτώσεις, κάποιος θέλει να αποφασίσει μεταξύ όλων των μεταβλητών που έχουν περιλιφθεί σε ένα τύπο μοντέλου, για το ποιες είναι στατιστικά σημαντικές ή σχετικές κατά κάποιο τρόπο για να περιγράψουν την μεταβλητή απόκρισης. Με άλλα λόγια ποια μεταβλητή θα παραμείνει στο μοντέλο και ποια θα πεταχτεί. Ξεκινώντας από ένα πλήρες μοντέλο (με όλες τις επεξηγηματικές μεταβλητές μέσα), μπορούμε να ορίσουμε μια μεγάλη οικογένεια των μοντέλων: όλα τα μοντέλα που περιέχουν μερικές από τις μεταβλητές. Όλες είναι εμφωλευμένες (nested) μέσα στο πλήρες μοντέλο. Το θέμα είναι ποιες από αυτές πρέπει να διατηρηθούν στο τελικό μοντέλο.

Η πιο συνηθισμένη μέθοδος επιλογής, αλλά και αυτή που χρησιμοποιείται από το συγκεκριμένο πακέτο είναι η Stepwise selection (§ 1.2.2.1). Αρχικά, προσαρμόζει το πλήρες μοντέλο και αναζητά τις στατιστικά μη σημαντικές μεταβλητές ώστε να τις πετάξει εκτός μοντέλου (διότι η απομάκρυνση τους δε μειώνει σημαντικά την προσαρμογή του μοντέλου). Στη συνέχεια, αφαιρεί όλες τις στατιστικά μη σημαντικές μεταβλητές επιτυγχάνοντας έτσι τη δημιουργία ενός νέου πιο απλοποιημένου μοντέλου. Η διαδικασία συνεχίζεται μέχρι να παραμείνουν στο μοντέλο όλες οι μεταβλητές που κρίθηκαν σημαντικές. Η διαδικασία μπορεί να εφαρμοστεί χρησιμοποιώντας διάφορα κριτήρια επιλογής. Το συγκεκριμένο πακέτο χρησιμοποιεί το κριτήριο του AIC_c , για το οποίο μιλούμε στην παράγραφο 1.2.3.2.

Ο ρόλος του

Συνοψίζοντας, το glmulti-πακέτο παρέχει μια συνάρτηση, τη *glmulti*, που χρησιμοποιείται αντί της *glm* ή κάποιας άλλης συνάρητησης (π.χ. *lm*). Δημιουργεί όλους τους πιθανούς τύπους μοντέλων (που ορίζονται από τα αποτελέσματα και λαμβάνοντας υπόψη κάποιους περιορισμούς), τους ταιριάζει με τη glm και επιστρέφει τα καλύτερα μοντέλα σε ένα αντικείμενο της κλάσης glmulti. Έτσι μπορεί να γίνει επιλογή του «καλύτερου» μοντέλου, παρουσιάζοντας ένα κατάλληλο διάστημα εμπιστοσύνης και τις μέσες εκτιμημένες παραμέτρους ή μεταβλητές του μοντέλου (Calcagno & Mazancourt, 2010).

Το glmulti, λοιπόν, έχει να δημιουργήσει όλους πιθανούς τύπους μοντέλων που αφορούν τις συγκεκριμένες μεταβλητές. Ο χρήστης θα βάλει τις μεταβλητές που πρέπει να θεωρηθούν και το πρόγραμμα θα δημιουργήσει όλα τα πιθανά μοντέλα που προκύπτουν από τις μεταβλητές, αλλά και από τις αλληλεπιδράσεις τους ανά δύο. Για τρεις και πάνω αλληλεπιδράσεις δεν περιέχονται στην παρούσα έκδοση του πακέτου, διότι ο αριθμός των μοντέλων εκρήγνυται κυριολεχτικά από τις αλληλεπιδράσεις. Όσο για την πολυπλοκότητα, ο αριθμός μόνο των μοντέλων που συγκρίνονται φτάνει στο 2^n και με τις ανα ζεύγος αλληλεπιδράσεις φτάνει κοντά στο 2^{n^2}(Calcagno & Mazancourt, 2010).

Όμως, όταν πολλοί παράγοντες εμπεριέχονται στις αλληλεπιδράσεις, πολλές φόρμουλες παρουσιάζουν το ίδιο μοντέλο και καθώς ο αριθμός των παραγόντων αυξάνεται, όλες αυτές οι φόρμουλες, πολύ συχνά είναι άχρηστες. Εδώ έρχεται να φανεί ο ρόλος του πακέτου αυτού, το οποίο πετάει όλους τους τύπους που δε χρειάζονται, ώστε να αποφευχθεί η παρουσίαση του ίδιου μοντέλου πολλές φορές. Σε ένα πρόβλημα με πολλές μεταβλητές, άρα και πολλές αλληλεπιδράσεις, ο αριθμός των υποψήφιων μοντέλων είναι μεγάλος, οπότε χρειάζεται μεγάλη διαδικασία και πολύς χρόνος ώστε να μπορεί να ελεχθεί η κατάσταση.

Η default μέθοδος του glmulti (method=”h”) προσαρμόζει όλα τα υποψήφια μοντέλα. Όμως όταν θεωρηθούν 5 ή 6 μεταβλητές και πάνω, η αναζήτηση καθίσταται δύσχρηστη γιατί ο αριθμός των υποψηφίων μοντέλων είναι πολύ υψηλός. Όταν χρησιμοποιείται method=”d” το glmulti επιστρέφει τον αριθμό των υποψήφιων μοντέλων. Ακόμα και με τους πρόσφατους επεξεργαστές, ο χρόνος υπολογισμού είναι αρκετά μεγάλος. Μία λύση είναι να μειωθεί ο αριθμός των υποψηφίων μοντέλων με τον καθορισμό περιορισμών. Για το λόγο αυτό, το glmulti παρέχει μια εναλλακτική λύση για τη στρατηγική «fit them all». Διαθέτει ένα γενετικό αλγόριθμο (genetic algorithm), που χρησιμοποιείται όταν είναι ρυθμισμένη στο method “g”. Χάρη σε αυτή την επιλογή, επιτυγχάνεται διερεύνηση μόνο ενός υποσυνόλου από όλα τα πιθανά μοντέλα, αλλά τυχαία με μια τάση προς τα καλύτερα και έτσι οι υπολογισμοί γίνονται πολύ πιο γρήγορα. Οι γενετικοί αλγόριθμοι είναι πολύ χρήσιμοι για προβλήματα βελτιστοποίησης, όμως αυτός ο τομέας είναι αρκετά περίπλοκος.

Πώς είναι χτισμένο

Το πακέτο της R *glmulti* δεν έχει κάποια ιδιαιτερότητα στο χτίσιμό του, ή στη λειτουργία του, διότι καλεί άλλες λειτουργίες. Είναι ενσωματωμένη η *glm* για την προσαρμογή των μοντέλων και έχει και ένα υπόβαθρο κάποιων κλάσεων της Java που παρέχονται μαζί με το πακέτο και είναι συγκεντρωμένα στο αρχείο **glmulti.jar**. Οι κλάσεις της Java που χρησιμοποιούνται είναι ModelGenerator, GLMModel, Resumator. Την επικοινωνία μεταξύ *glmulti* και κλάσεων της Java τη χειρίζεται το πακέτο ***rJava*** (Urbanek, 2009), η παρουσία του οποίου απαιτείται για να λειτουργήσει το *glmulti* και επιπλέον δε χρειάζεται κάποιο μεταφραστή (compiler).

Πώς δουλεύει

Αρχικά τα δεδομένα προσαρμόζονται στο μοντέλο. Το *glmulti* δεν προσαρμόζει τίποτα, απλά προτείνει τύπους μοντέλων, τα οποία μεταφέρονται σε μια συνάρτηση της R για να τα προσαρμόσει. Η προεπιλεγμένη και πιο συνηθισμένη είναι η *glm*, αλλά ο χρήστης μπορεί και με κάποια άλλη συνάρτηση της προτίμησής του (π.χ. η *lm*) να προσαρμόσει τα δεδομένα που έχει, π.χ. *glm (y ~ -1 + c + a + z + a:z).*

Ακολούθως καθορίζονται τι είδους μοντέλα θα εξεταστούν. Στην εφαρμογή που θα δούμε πιο κάτω χρησιμοποιούμε τη λογιστική παλινδρόμηση, οπότε μπαίνει και ο περιορισμός family=”binomial”, αλλά και όταν θέλουμε τη συνάρτηση σύνδεσης που θα χρησιμοποιήσουμε (π.χ. link=”probit”). Επίσης, όπως έχουμε πει και προηγουμένως, πολλές μεταβλητές και αλληλεπιδράσεις δημιουργούν μεγαλύτερη πολυπλοκότητα στο πρόβλημα. Ένας τρόπος είναι να απομακρύνουμε κάποιες μεταβλητές ή αλληλεπιδράσεις από τον έλεγχο με τη βοήθεια της ρύθμισης π.χ. exclude = c("a:c", "c:z"), ώστε να μην ελέγξει τις περιπτώσεις για τις αλληλεπιδράσεις "a:c" και "c:z". Επιπλέον, ένα σημείο τομής περιλαμβάνεται σε όλα τα μοντέλα, δηλ. y ~ 1 + c + a + z, όπου το 1 υποδηλώνει το σημείο τομής. Το σημείο τομής μπορεί και να παραλειφθεί όταν γράψουμε y ~ c + a + z – 1. Αυτό μπορεί να επιτευχθεί και με τη ρύθμιση intercept= FALSE.

Διάφορες εντολές που μπορεί να χρησιμοποιηθούν για να λειτουργήσει το πακέτο glmulti είναι:

Π.χ.

```
> output <- glmulti(y ~ -1 + c*a*z - c:z - a:c, data = myDataFrame,maxit = 30)
or
> mod <- glm(y ~ -1 + c + a + z + a:z, data = myDataFrame, maxit = 30)
> output <- glmulti(mod)
or
> output <- glmulti("y", c("a", "c", "z"), exclude = c("a:c", "c:z"),data = myDataFrame,
+intercept = FALSE, maxit = 30)
```

Όπως έχουμε αναφέρει, το πακέτο για να επιλέξει το «κατάλληλο» μοντέλο χρησιμοποιεί την κατά βήμα επιλογή μεταβλητών (Stepwise selection), με βάση το κριτήριο AIC_c για τη σύγκριση των μοντέλων. Αλλά μπορεί να χρησιμοποιήσει επίσης το AIC αλλά και το BIC με τη βοήθεια της ρύθμισης crit (π.χ. crit = bic). Κάθε λειτουργία αποδέχεται το προσαρμοσμένο μόντελο και επιστρέφει μια αριθμητική τιμή για τα πιο πάνω κριτήρια ως επιχείρημα. Έτσι συγκρίνοντας τις τιμές των κριτηρίων σε όλα τα μοντέλα που προσαρμόζει, μπορεί να επιλέξει και να καταλήξει στο βέλτιστο ή «κατάλληλότερο» μοντέλο για το πρόβλημα (Calcagno & Mazancourt, 2010).

Διαθεσιμότητα

Το πακέτο-glmulti είναι δωρεάν διαθέσιμο από τη Comprehensive R Archive Netwok (CRAN) στην http://CRAN.R-project.org/package=glmulti .

2. ΜΕΡΙΚΕΣ ΕΝΤΟΛΕΣ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΟΥΜΕ

glm (. , family = .): Είναι η εντολή που καλεί το πακέτο *glmulti*, η οποία προσαρμόζει τις μεταβλητές (τα δεδομένα) σε ένα γενικευμένο γραμμικό μοντέλο (generalized linear mondel), εξού και το όνομά της. Επίσης βάζουμε ότι το μοντέλο που θα προσαρμόσουμε ανήκει στην Διωνυμική κατανομή (family = binomial), διότι θέλουμε την προσαρμογή ενός μοντέλου λογιστικής παλινδρόμησης. Αν θέλαμε ένα μοντέλο της παλινδρόμησης Poisson θα βάζαμε family = poisson. Για την προσαρμογή του μοντέλου λογιστικής παλινδρόμησης χρησιμοποιείται η συνάρτηση σύνδεσης *logit.* π.χ. > glm(y~x1+x2+x3, family = binomial), όπου y τα δεδομένα για τη μεταβλητή απόκρισης (σε αυτή την περίπτωση μπορεί να είναι και κατηγορική (με τιμές 1,0)) και x1,x2,.. τα δεδομένα των επεξηγηματικών μεταβλητών.

summary (.): Παρουσιάζει τα αποτελέσματα της ανάλυσης παλινδρόμησης, του μοντέλου που έχει προσαρμοστεί. Δίνει τους περιγραφικούς δείχτες των υπολοίπων του μοντέλου που προσαρμόζεται (deviance residuals), τους εκτιμητές $\hat{\alpha}$, $\hat{\beta}_1$, $\hat{\beta}_2, \dots$, τα τυπικά σφάλματα (standard errors) $se(\hat{\alpha})$, $se(\hat{\beta}_1)$, $se(\hat{\beta}_2), \dots$, αλλά και τον έλεγχο Wald (z-value) του καθενός με την p-τιμή του, αντίστοιχα. Επίσης δίνει τη Residual Deviance, που είναι η Deviance του μοντέλου που προσαρμόζεται, αλλά και τη Null Deviance (η Deviance του μοντέλου που περιέχει μόνο το σταθερό όρο), με τους βαθμούς ελευθερίας του κάθενος. Τέλος δίνει και την τιμή του κριτηρίου AIC.

confint (.): Η εντολή αυτή επιστρέφει τα 95% διαστήματα εμπιστοσύνης των εκτιμητριών των μεταβλητών.

step (. , direction = "."): Επιστρέφει το βέλτιστο μοντέλο (κάνοντας επιλογή μεταβλητών) εκτελώντας διαδικασία επιλογής με βήματα. Στο direction μπορούμε να βάλουμε έιτε "backward", είτε "forward" για την αντίστοιχη διαδικασία, είτε "both" για να χρησιμοποιήσει συνδυασμό των δύο. Και οι τρεις αυτές διαδικασίες βασίζονται στο κριτήριο AIC.

3. ΠΡΟΣΑΡΜΟΓΗ ΜΟΝΤΕΛΟΥ

Πιο κάτω η διαδικασία που ακολουθούμε στην R:

Διαβάζουμε τα δεδομένα γραμμή-γραμμή (byrow=T), από το αρχείο, σαν πίνακα με 8 γραμμές και 51 στήλες.

```
> dataAML<-matrix(scan("building_aml.txt"),ncol=51,byrow=T)
Read 510 items
```

Ορίζουμε τις επεξηγηματικές μεταβλητές (x1 μέχρι x6) και τη μεταβλητή απόκρισης y (resp) (δεν διαβάζει την $1^{η}$ στήλη γιατί είναι ο αύξων αριθμός):

```
> age<-dataAML[2,]
> smear<-dataAML[3,]
> infiltr<-dataAML[4,]
> lab<-dataAML[5,]
> blasts<-dataAML[6,]
> temp<-dataAML[7,]

> resp<-dataAML[8,]
```

Ορίζουμε τη resp, δηλ. την αντιδραση του ασθενή στη θεραπεία (response), σαν κατηγορική μεταβλητή, με 1 = αντέδρασε στη θεραπεία, 0 = δεν αντέδρασε.

```
> resp<-as.factor(resp)
> resp
 [1] 1 1 1 1 1 0 1 1 1 0 1 1 0 1 0 1 1 0 1 1 0 0 0 0 1 1 0 0 1 0 1 0 0 0 0 0 0 0
[39] 0 1 0 0 0 1 0 0 1 0 1 1 0
Levels: 0 1
```

Δημιουργούμε πλαίσιο δεδομένων, που έχει το πλεονέκτημα ότι τα στοιχεία του δε χρειάζεται να είναι του ίδιου τύπου, για να έχουμε μαζεμένα τα δεδομένα (dataAML).

```
> dataAML<-data.frame(cbind(resp,age,smear,infiltr,lab,blasts,temp))
```

Γραφική περιγραφή των μεταβλητών:

Με cdplots περιγράφονται όλες οι μεταβλητές σε σχέση με τη resp (Γραφήματα 2.1).

```
> layout(matrix(1:6,nrow=2))
> cdplot(resp~age,ylab="Response(1:responds,0:fails)",xlab="Age(years)")
> cdplot(resp~smear,ylab="Response",xlab="Smear differential(%)")
> cdplot(resp~infiltr,ylab="Response",xlab="Absolute infiltrate(%)")
> cdplot(resp~lab,ylab="Response",xlab="Labelling index(%)")
> cdplot(resp~blasts,ylab="Response",xlab="Absolute blasts(x10^3)")
> cdplot(resp~temp,ylab="Response",xlab="Temperature(F)")
```

Συσχέτιση μεταξύ των μεταβλητών:

Για τον πίνακα 2.2, οι εντολές:

```
> corrAML<-cor(dataAML)
> corrAML
```

Και για το γράφημα 2.2:

```
> library(graphics)
> library(corrplot)
> corrplot(corrAML)
```

Προσαρμόζουμε στο γενικευμένο γραμμικό μοντέλο:

Προσαρμογή με λογιστική παλινδρόμηση και συνάρτηση σύνδεσης τη *logit*, καθώς η μεταβλητή απόκρισης resp είναι κατηγορική.

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

Παρουσιάζουμε τα αποτελέσματα (με παρόμοιο τρόπο παρουσιάζονται τα αποτελέσματα όλων των μοντέλων, μετά την προσαρμογή τους):

```
> summary(results_glm)
```

Και τα συμμετρικά 95% διαστήματα εμπιστοσύνης (Πίνακας 2.3):

```
> confint(results_glm)
```

4. ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ

Με τις 3 κατά βήματα διαδικασίες επιλογής, με βάση το κριτήριο AIC:

Πριν ξεκινήσουμε, θέτουμε το κενό μοντέλο.

```
> null=glm(resp~1,family=binomial)
> null
Call:  glm(formula = resp ~ 1, family = binomial)

Coefficients:
(Intercept)
   -0.1178

Degrees of Freedom: 50 Total (i.e. Null);  50 Residual
Null Deviance:      70.52
Residual Deviance: 70.52        AIC: 72.52
```

Και το πλήρες μοντέλο.

```
> full=glm(resp~age+smear+infiltr+lab+blasts+temp,family=binomial)
>  full
Call:  glm(formula = resp ~ age + smear + infiltr + lab + blasts + temp,  family = binomial)

Coefficients:
(Intercept)         age         smear         infiltr         lab          blasts        temp
  108.33115    -0.06231    -0.00469     0.03104     0.37281     0.03267    -0.11162

Degrees of Freedom: 50 Total (i.e. Null);  44 Residual
Null Deviance:      70.52
Residual Deviance: 39.28        AIC: 53.28
```

- Η διαδικασία της διαδοχικής αφαίρεσης (Backward elimination).

```
> step(full, dataAML, direction="backward")
```

- Η διαδικασία της διαδοχικής πρόσθεσης (Forward selection).

```
> step(null, scope=list(lower=null, upper=full), direction="forward")
```

- Η διαδικασία της κατά βήματα πίσω-εμπρός επιλογής (Stepwise selection).

```
> step(null, scope=list(upper=full), direction="both")
```

Με τη βοήθεια του πακέτου glmulti:

```
>library(glmulti)
```

- Με βάση το κριτήριο AIC.

```
>test1 <- glmulti(resp~age*smear*infiltr*lab*blasts*temp, data=dataAML,  fitfunc=glm,
+crit = aic)
```

Και καταλήγουμε στο εξής μοντέλο:

```
…..
After 2169450 models:
Best model: resp~1+age+lab+blasts:age+blasts:infiltr+temp:lab
Crit= 53.975796
Mean crit= 55.6326762
Completed.
```

Με τη βοήθεια της εντολής έχουμε το γράφημα:

```
plot(test1, type = "p")
```

- Με το κριτήριο BIC.

```
>test2 <- glmulti(resp~age*smear*infiltr*lab*blasts*temp, data=dataAML, fitfunc = glm,
+crit =bic)
>plot(test2, type = "p",highlight="smear:infiltr")
```

Και καταλήγουμε στο εξής μοντέλο:

```
…..
After 2169450 models:
Best model: resp~1 +lab+blasts:infiltr+temp:age+temp:lab
Crit= 66.590167
Mean crit= 68.468946
Completed.
```

- Επίσης, με τη βοήθεια του γενετικού αλγορίθμου (genetic algorithm) εκτελούμε το glmulti με βάση το κριτήριο AIC, με την εντολή:

```
> test3ga <- glmulti(resp~age*smear*infiltr*lab*blasts*temp,  data=dataAML,  fitfunc = glm,
+crit = aic, method = "g")
```

Και καταλήγουμε:

….
After 490 generations:
Best model: resp~1+lab+blasts:age+blasts:infilrt+temp:age+temp:lab
Crit= 54.022792
Mean crit= 58.239361
Improvements in best and average IC have been below the specified goals.
Algorithm is declared to have converged.
Completed.

- Και πάλι, με τη βοήθεια του γενετικού αλγορίθμου (genetic algorithm) εκτελούμε το glmulti με βάση το κριτήριο BIC, με την εντολή:

```
>test4ga <- glmulti(resp~age*smear*infiltr*lab*blasts*temp,  data=dataAML,  fitfunc = glm,
+crit = bic, method = "g")
```

Και καταλήγουμε στο μοντέλο:

….
After 650 generations:
Best model: resp~1+age+lab+temp
Crit= 67.602798
Mean crit= 76.088420
Improvements in best and average IC have been below the specified goals.
Algorithm is declared to have converged.
Completed.

Εκτελούμε τις πιο κάτω εντολές, όπου η πρώτη εμφανίζει πόσες generations επιτεύχθηκαν στην εκτέλεση και η δεύτερη δείχνει το ποσοστό του χρόνου που χρειάστηκε με genetic algorithm σε σχέση με τον πραγματικό χρόνο που χρειάζεται η κανονική εκτέλεση του glmulti σε default κατάσταση.

Με το AIC έχουμε:

```
> summary(test3ga)$generations
[1] 490
> summary(test3ga)$elapsed
[1] 0.05931417
```

Και με το BIC:

```
> summary(test4ga)$generations
[1] 650
> summary(test4ga)$elapsed
[1] 0.08187083
```

Με τη βοήθεια της ελεγχοσυνάρτησης Deviance:

Υπολογίζουμε και την p-τιμή του ελέγχου της Deviance του κάθε μοντέλου (Residual Deviance), για την ακρίβεια μιας τιμής της Deviance από την X^2-κατανομή, με την εντολή:

```
> 1-pchisq(results_glm$deviance,results_glm$df.residual)
```

D. ΕΦΑΡΜΟΓΗ ΣΕ ΜΟΝΤΕΛΟ ΤΟΥ COX ΣΤΗΝ R

1. ΤΟ ΠΑΚΕΤΟ PENALIZED (PENALIZED-PACKAGE)

Εισαγωγή

Ένα άλλο πακέτο της R, που αυτή τη φορά χρησιμοποιείται για ποινικοποιημένες εκτιμήσεις (penalized-estimation), είναι το **πακέτο-Penalized**. Τα μοντέλα που υποστηρίζονται μέχρι στιγμής από αυτό το πακέτο είναι της γραμμικής παλινδρόμησης, της λογιστικής παλινδρόμησης, της Poisson παλινδρόμησης αλλά και του μοντέλου αναλογικής διακινδύνευσης του Cox, το οποίο θα χρησιμοποιήσουμε στην πιο κάτω εφαρμογή. Όσον αφορά τις ποινές (penalties), το πακέτο επιτρέπει την L_1 ("Lasso") απόλυτη τιμή της ποινής, την τετραγωνική L_2 ("ridge regression") ποινή ή και συνδυασμό των δύο ("the naive elastic net").

Όπως έχουμε αναφέρει και προηγουμένως, οι μέθοδοι L_1 και L_2 συρρικνώνουν (shrink) τους εκτιμημένους συντελεστές της παλινδρόμησης προς το μηδέν, σε σχέση με τις εκτιμήσεις της μεγιστοποιημένης πιθανοφάνειας. Ο σκοπός της παρούσας συρρίκνωσης είναι να αποτρέψει το overfit που προκύπτει και οφείλεται είτε στην πολυσυγγραμμικότητα των συντελεστών, είτε στο υψηλών διαστάσεων πρόβλημα. Η ποσότητα της συρρίκνωσης καθορίζεται από τις παραμέτρους ρύθμισης λ_1 και λ_2. Η παράμετρος λ_1 χρησιμοποιείται για τη μέθοδο Lasso, ενώ και οι δύο μαζί (λ_1 και λ_2) χρησιμοποιούνται για τη μέθοδο elastic net (E-net). Η τιμή μηδέν σημαίνει «καθόλου συρρίκνωση» (= *maximum likelihood*) και η τιμή του απείρου σημαίνει «πεπερασμένη συρρίκνωση» (= *ρύθμιση όλων των συντελεστών παλινδρόμησης στο μηδέν*) (Goeman & Meijer, 2012). Αξίζει να σημειωθεί ότι αυτές οι μέθοδοι συρρίκνωσης δεν είναι αμετάβλητες ανάλογα με τη συσχέτιση των συμμεταβλητών. Οπότε, προτού να προσαρμόσουμε το μοντέλο με το penalized, πρέπει να ελέγξουμε αν υπάρχει κάποια συσχέτιση μεταξύ των μεταβλητών ή αν πρέπει να τυποποιηθούν (standardized) οι μεταβλητές.

Ο ρόλος και η λειτουργία του

Η βασική λειτουργία του πακέτου είναι η συνάρτηση **penalized**, η οποία όταν την καλέσουμε εκτελεί ποινικοποιημένη εκτίμηση για σταθερές τιμές των λ_1 και λ_2. Η σύνταξη του έχει ως «χαλαρό» πρότυπο, ή μπορούμε να πούμε λειτουργεί όμοια, με τις λειτουργίες της συνάρτησης-glm που χρησιμοποιήσαμε στην εφαρμογή με λογιστική παλινδρόμηση (Παράρτημα C.2), με τη διαφορά ότι αυτό είναι πιο ευέλικτο σε ορισμένα σημεία. Δύο τύποι εισόδου επιτρέπονται: είτε με χρήση αντικειμένων (formula objects), είτε με χρήση πινάκων.

Η συνάρτηση penalized χρησιμοποιείται για να προσαρμόσει ένα ποινικοποιημένο μοντέλο πρόβλεψης με την πρόβλεψη της απόκρισης. Ιδιαίτερα στο μοντέλο του Cox χρησιμοποιείται για την πρόβλεψη και ο χρόνος επιβίωσης (*Surv* object).

Π.χ.

```
> fit <- penalized(Surv(time,event)~DIAPH3+NUSAP1, data, lambda2=1)
```

Να αναφέρουμε ακόμα, ότι είναι δύσκολο να αποφασίσουμε εκ των προτέρων ποια τιμή του λ_1 και λ_2 πρέπει να χρησιμοποιήσουμε. Το συγκεκριμένο πακέτο προσφέρει τρόπους για να βρούμε τις βέλτιστες τιμές των λ, συγκεκριμένα τη μέθοδο της cross-validation (§ 1.2.4.4 και § 3.2.5). Είναι πολύ σημαντικός ο ρόλος της μεθόδου CVL διότι μπορεί να συγκρίνει την προβλεπτική ικανότητα των διαφορετικών τιμών των παραμέτρων ρύθμισης λ. Έτσι η εύρεση των βέλτιστων τιμών για τα λ, μπορεί να μας καθορίσει το βέλτιστο μοντέλο.

2. ΜΕΡΙΚΕΣ ΕΝΤΟΛΕΣ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΟΥΜΕ

coxph (Surv (time, status) ~ . , data): Είναι η εντολή που καλεί τη συνάρτηση *coxph*, η οποία προσαρμόζει τις μεταβλητές (τα δεδομένα) στο μοντέλο αναλογικής διακινδύνευσης του Cox, εξού και το όνομά της. π.χ. > coxph (Surv(time,status) ~ x1+x2+x3), όπου *Surv (time, status),* η συνάρτηση επιβίωσης που εξαρτάται από το χρόνο και την κατάσταση και *x1, x2, ...* οι επεξηγηματικές μεταβλητές.

summary (.): Παρουσιάζει τα αποτελέσματα της ανάλυσης παλινδρόμησης του μοντέλου που έχει προσαρμοστεί. Δίνει τους περιγραφικούς δείχτες των υπολοίπων, τις εκτιμήσεις $\hat{\alpha}$, $\hat{\beta}_1$, $\hat{\beta}_2, \ldots$, τα τυπικά σφάλματα (standard errors) $se(\hat{\alpha})$, $se(\hat{\beta}_1)$, $se(\hat{\beta}_2)$, ... , αλλά και τον έλεγχο Wald του καθενός με την p-τιμή του, αντίστοιχα. Επίσης δίνει το συντελεστή προσδιορισμού R^2 και μερικές άλλες πληροφορίες.

confint (.): Η εντολή αυτή επιστρέφει τα 95% διαστήματα εμπιστοσύνης των εκτιμητριών των μεταβλητών.

profL1 ("μοντέλο" , fold = "."): Αυτή η εντολή μπορεί να χρησιμοποιηθεί για να εξετάσει την επίδραση των παραμέτρων λ1 και λ2 για τη συνάρτηση CVL. Πιο συγκεκριμένα η λειτουργία αυτή μπορεί να χρησιμοποιηθεί για να μεταβάλλει το λ1, κρατώντας σταθερή την τιμή του λ2. Η παράμετρος fold μπορεί να χρησιμεύσει ως βάση για την επόμενη κλήση CVL, ώστε να εξασφαλιστεί η συγκρισιμότητα.

profL2 ("μοντέλο" , fold = "." , minl = ".", maxl = "."): Αυτή η εντολή μπορεί να χρησιμοποιηθεί για να εξετάσει την επίδραση των παραμέτρων λ1 και λ2 για τη συνάρτηση CVL. Πιο συγκεκριμένα η λειτουργία αυτή μπορεί να χρησιμοποιηθεί για να μεταβάλλει το λ2, κρατώντας σταθερή την τιμή του λ1.

optL1 ("μοντέλο" , fold = "."): Επιστρέφει το βέλτιστο λ1, που θα το χρησιμοποιήσουμε σαν παράμετρο ρύθμισης για την τεχνική Lasso.

optL2 ("μοντέλο" , fold = "."): Επιστρέφει το βέλτιστο λ2, που θα το χρησιμοποιήσουμε σαν παράμετρο ρύθμισης για την τεχνική της Ridge regression.

penalized ("μοντέλο" , data, lambda1 = ".", lambda2 = "."): Με την εντολή αυτή εφαρμόζονται οι τεχνικές επιλογής μοντέλου Lasso, Ridge regression και Elastic net. Αν βάλουμε την παράμετρο lambda1 = "." εκτελείται η Lasso, αν βάλουμε την παράμετρο lambda2 = "." εκτελείται η Ridge regression και αν βάλουμε και τις δύο παραμέτρους, εκτελείται η E-net. (Περισσότερα για την εντολή αυτή αναφέρονται στο Παράρτημα D.1)

3. ΠΡΟΣΑΡΜΟΓΗ ΜΟΝΤΕΛΟΥ

Πιο κάτω η διαδικασία που ακολουθούμε στην R:

Διαβάζουμε τα δεδομένα γραμμή-γραμμή (byrow=T), από το αρχείο, σαν πίνακα, με 8 γραμμές και 51 στήλες.

```
> dataAML<-matrix(scan("building_aml.txt"),ncol=51,byrow=T)
Read 510 items
```

Ορίζουμε τις επεξηγηματικές μεταβλητές (x1 μέχρι x6), τη μεταβλητή απόκρισης (status), η οποία είναι κατηγορική και το χρόνο επιβίωσης (time) (δεν διαβάζει την $1^{η}$ στήλη γιατί είναι ο αύξων αριθμός):

```
> age<-dataAML[2,]
> smear<-dataAML[3,]
> infiltr<-dataAML[4,]
> lab<-dataAML[5,]
> blasts<-dataAML[6,]
> temp<-dataAML[7,]

> surv<-dataAML[9,]
> status<-dataAML[10,]
```

Θέτουμε τη status, δηλ. την κατάσταση του ασθενή μετά τη θεραπεία (status), σαν κατηγορική μεταβλητή, με 1 = ακόμα ζωντανός, 0 = έχει πεθάνει.

```
> status <-as.factor(status)
> status
 [1] 0 1 0 0 0 0 0 1 1 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
[39] 0 0 0 0 0 0 0 0 0 0 0 0 0
Levels: 0 1
```

Μετατρέπουμε σε πλαίσιο δεδομένων και στη συνέχεια ελέγχουμε τη συσχέτιση μεταξύ των μεταβλητών αριθμητικά:

```
> dataAML<-data.frame(cbind(time,status,age,smear,infiltr,lab,blasts,temp))
> corrAML<-cor(dataAML)
> corrAML
```

Και γραφικά:

```
> library(graphics)
> library(corrplot)
> corrplot(corrAML)
```

Προσαρμόζουμε στο μοντέλο αναλογικής διακινδύνευσης του Cox:

Προσαρμογή με μοντέλο του Cox, με τη συνάρτηση επιβίωσης να εξαρτάται από το χρόνο επιβίωσης (time) και την κατάσταση του ασθενούς μετά τη θεραπεία (status).

```
> results_cox<-coxph(Surv(time,status)~age+smear+infiltr+lab+blasts+temp,data=dataAML)
```

Παρουσιάζουμε τα αποτελέσματα:

```
> summary(results_cox)
```

Και τα συμμετρικά 95% διαστήματα εμπιστοσύνης για όλες τις παραμέτρους του μοντέλου (α, β1, β2, …):

```
> confint(results_cox)
```

4. ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ

Εκτελούμε με την τεχνική Lasso. Βρίσκουμε το βέλτιστο λ1 με τη μέθοδο CVL:

```
> library(penalized)
> fitprof_lasso<-profL1 (Surv(surv,status)~age+smear+infltr+lab+blasts+temp, plot=TRUE)

> opt_lasso<-optL1(Surv(surv,status)~age+smear+infltr+lab+blasts+temp, minlambda1=1,
+maxlambda1=50)
> opt_lasso$lambda
[1] 29.43065
```

Προσαρμόζουμε με τη μέθοδο Lasso για λ1 ίσο με 29, με τη βοήθεια του πακέτου penalized:

```
> fit_final_lasso<-penalized(Surv(surv,status)~age+smear+infltr+lab+blasts+temp, dataAML,
+lambda1=29)

> coefficients(fit_final_lasso,"penalized")
       age          infiltr            lab           temp
0.027846210 -0.008259883 -0.026284444 0.008924687
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

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

Androulakis, E., Koukouvinos, C., Mylona, K., & Vonta, F. (2010). A real survival analysis application via variable selection methods for Cox 's proportional hazards model. *Journal of Applied Statistics*, 37: 1399-1406.

Breslow, N. E. (1974). Covariance analysis of censored survival data. *Biometrics*, 30: 89-99.

Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, 53: 603-618.

Calcagno, V., & Mazancourt, C. (2010). glmulti: An R Package for Easy Automated Model. *Journal of Statistical Software*, 12: 1-28.

Cameron, A. C., & Windmeijer, F. A. (1996). R-squared measures for count data regression models with applications to health-care utilization. *Journal of Business and Economic Statistics*, 14: 209-220.

Caroni, C. (2002). The correct "ball bearings" data. *Lifetime Data Analysis*, 8: 395-399.

Caroni, C. (2004). Diagnostics for Cox 's Proportional Hazards Model. *"Parametric and Semiparametric Models with Applications to Reliability, Survival Analysis, and Quality of Life"* (pp. 27-38). Birkhauser, Boston: M. S. Nikulin, N. Balakrishnan, M. Mesbah and N. Limnios, Eds.

Collett, D. (2003). *Modelling Binary Data, 2nd edition.* London: Chapman and Hall.

Collett, D. (2003). *Modelling Survival Data in Medical Research.* Boca Raton: Chapman and Hall/CRC.

Cox, D. R. (1972). Regression models and life tables (with Discussion). *Journal Royal Statistical Society, A*, 34: 187-220.

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

Cox, D., & Snell, E. (1989). *Analysis of Binary Data, 2nd edition.* London: Chapman and Hall.

Everitt, B., & Hothorn, T. (2006). *A Handbook of Statistical Analyses Using R.* New York: Chapman and Hall/CRC.

Goeman, J. (2010). L1 Penalized Estimation in the Cox Proportional Hazards Model. *Biometrical Journal*, 52, 70-84.

Goeman, J., & Meijer, R. (2012). L1 and L2 Penalized Regression Models. *Biometrical Journal*, 52: 70-84.

Hastie, T., & Tibshirani, R. (1990). Chapter 8. In *Generalized Additive Models* (pp. 213-214). Chapman and Hall.

Hoerl, A., & Kennard, R. (1970). Ridge Regression: Biased Estimation for the Non Orthogonal Problems. *Technometrics*, vol 12, 1: 55-67.

Lee, E. (1980). *Statistical Methods for Survival Data Analysis.* Belmont, California: Lifetime Learning Publications.

Maddala, G. S. (1983). *Limited Dependent and Qualitative Variables in Econometrics.* Cambridge: Cambridge University Press.

Magee, L. (1990). R-squared measures based on Wald and likelihood ratio joint significance tests. *American Statistician*, 44: 250-253.

Marubini, E., & Valsecchi, M. (1995). *Analysing Survival Data from Clinical Trials and Observational Studies.* New Jersey: John Wiley and Sons.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka, *Frontiers in Econometrics* (pp. 105-142). New York: Academic Press.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2006). *Introduction to Linear Regression Analysis, 4th edition.* New Jersey: John Wiley and Sons, Inc.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78: 691-692.

Narula, S., & Wellington, J. (1977, May). Linear Regression and the Minimum Sum of Relative Errors. *Technometrics*, pp. 185-190.

R Development Core Team. (2010). *R: A Language and Environment for Statistical Computing.* Retrieved from R Foundation for Statistical Computing: http://www.R-project.org/

Raftery, A. (1995). Bayesian model selection in social research. *Sociological Methodology*, 111-163.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B*, 58: 267-288.

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

Urbanek, S. (2009). *rJava: Low-Level R to Java Interface.* Retrieved from R package version 0.8-1: http://CRAN.R-project.org/package=rJava

Verweij, P., & van Houwelingen, H. (1993). Cross-Validation in Survival Analysis. *Statistics in Medicine*, pp. 2305-2314.

Volinsky, C., & Raftery, A. (2000). Bayesian Information Criterion for censored survival models. *Biometrics*, 56: 256-262.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67: 301-320.

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

Κοκολάκης, Γ., & Φουσκάκης, Δ. (2009). *Στατιστική θεωρία και εφαρμογές.* Αθήνα: Συμεών.

Οικονόμου, Π., & Καρώνη, Χ. (2010). *Στατιστικά Μοντέλα Παλινδρόμησης.* Αθήνα: Συμεών.

Φουσκάκης, Δ. (2009). Παρουσίαση στο μάθημα Ανάλυση δεδομένων με Η/Υ-ΣΕΜΦΕ, (http://www.math.ntua.gr/~fouskakis/, τελευταία πρόσβαση στις 11/6/2013).