ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Σχολή Χρηματοοικονομικής και Στατιστικής

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

Ποινικοποιημένη Λογιστική παλινδρόμηση και εφαρμογές

Μιχαήλ Μπούνιας

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Ιούλιος 2025

UNIVERSITY OF PIRAEUS
School of Finance and Statistics

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN APPLIED STATISTICS

Penalized Logistic regression and applications

By

Michail Bounias

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
July 2025

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. …….. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών.

Τα μέλη της Επιτροπής ήσαν:

- Επίκουρος Καθηγητής Εμμανουήλ Ανδρουλάκης (Επιβλέπων)
- Καθηγητής Κωνσταντίνος Πολίτης
- Αναπληρωτής Καθηγητής Γεώργιος Τζαβελάς

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

Στον πατέρα μου

Ευχαριστίες

Μαζί με τη διπλωματική αυτή, κλείνει και ένας κύκλος σπουδών. Ένας κύκλος κατά τον οποίο είχα γύρω μου αρκετούς ανθρώπους και στους οποίους θα ήθελα να εκφράσω την ευγνωμοσύνη μου. Κατ'αρχάς, θα ήθελα να εκφράσω τις ευχαριστίες μου στον Επίκουρο Καθηγητή κύριο Εμμανουήλ Ανδρουλάκη για την αμέριστη βοήθειά του και για τις υποδείξεις του καθ'όλη τη διάρκεια ώστε να ολοκληρωθεί η εργασία αυτή, μέσα από την οποία εμπλούτισα τις γνώσεις μου σε ένα πεδίο σχεδόν άγνωστο μέχρι τότε σε εμένα. Επίσης, να ευχαριστήσω τον Καθηγητή κύριο Κωνσταντίνο Πολίτη και τον Αναπληρωτή Καθηγητή κύριο Γεώργιο Τζαβελά, μέλη της τριμελούς επιτροπής, για το χρόνο που αφιέρωσαν για την ανάγνωση της παρούσας διπλωματικής. Έπειτα, δε θα μπορούσα να παραλείψω τους συμφοιτητές μου, με τους οποίους μελετήσαμε μαζί, ανταλλάξαμε γνώσεις και μοιραστήκαμε εμπειρίες όλον αυτόν τον καιρό. Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου, η οποία με στήριξε κατά την περίοδο των σπουδών μου.

Περίληψη

Στην παρούσα διπλωματική παρουσιάζεται η χρήση της ποινικοποίησης και διάφορες μέθοδοι αυτής στην περίπτωση της λογιστικής παλινδρόμησης. Στο 1ο κεφάλαιο γίνεται μία σύντομη εισαγωγή στα γενικευμένα γραμμικά μοντέλα, τα οποία αποτελούν γενίκευση και επέκταση των κλασσικών γραμμικών μοντέλων, και παρουσιάζονται, μεταξύ άλλων, η εκθετική οικογένεια κατανομών, οι συνιστώσες που περιέχει ένα γενικευμένο γραμμικό μοντέλο και η έννοια της απόκλισης. Στο 2ο κεφάλαιο γίνεται μία μικρή αναφορά στα δίτιμα δεδομένα, καθώς και στη λογιστική, στη σιγμοειδή και στη logit συνάρτηση. Επίσης, γίνεται ανάλυση της λογιστικής παλινδρόμησης· πώς ερμηνεύονται οι παράμετροι, πώς εκτιμώνται και πώς ελέγχεται η σημαντικότητά τους. Στο 3ο κεφάλαιο παρατίθενται οι έννοιες της υπερπροσαρμογής, της υποπροσαρμογής, της αντιστάθμισης μεροληψίας-διακύμανσης, της πολυσυγγραμμικότητας και των δεδομένων υψηλών διαστάσεων. Έπειτα, αναλύεται η έννοια της ποινικοποίησης και περιγράφονται πέντε συναρτήσεις ποινής: η ridge, η lasso, η elastic net, η SCAD και η MCP. Στο 4ο και τελευταίο κεφάλαιο εφαρμόζονται οι τεχνικές που έχουν αναφερθεί σε δύο σύνολα πραγματικών δεδομένων· στο σύνολο δεδομένων Burns και στο σύνολο δεδομένων Breast Cancer Wisconsin (Diagnostic).

Abstract

In this thesis we analyze the use of penalization and various methods of it in the case of logistic regression. In the 1st chapter there is a brief introduction to generalized linear models, which are a generalization and extension of classical linear models, and among others are presented the exponential family of distributions, the components contained in a generalized linear model and the concept of deviance. In the 2nd chapter, there is a brief reference to binary data, as well as the logistic, the sigmoid and the logit function. In addition, the logistic regression is analyzed· how the parameters are interpreted, how they are estimated and how their significance is checked. In the 3rd Chapter the notions of overfitting, underfitting, bias-variance trade-off, multicollinearity and high-dimensional data are presented. Next, the meaning of penalization is analyzed and five penalty functions are described: ridge, lasso, elastic net, SCAD and MCP. In the 4th and final chapter, the aforementioned techniques are applied in two real-world datasets· the Burns dataset and the Breast Cancer Wisconsin (Diagnostic) dataset.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Κ Ε Φ Α Λ Α Ι Ο 1

Εισαγωγή στα Γενικευμένα Γραμμικά Μοντέλα

1.1 Ανάλυση Παλινδρόμησης

Πολλές φορές στην Στατιστική καλούμαστε να εξετάσουμε τη σχέση μεταξύ δύο ή και περισσότερων μεταβλητών, ώστε να μπορούμε να προβλέψουμε τη μία εξ αυτών μέσω των υπολοίπων. Μερικές τέτοιες περιπτώσεις είναι:

- Τα χρήματα που ξοδεύει μία εταιρεία για τη διαφήμιση ενός προϊόντος και οι πωλήσεις του προϊόντος αυτού.
- Το ύψος και το βάρος ενός ανθρώπου.
- Να προβλέψουμε την πιθανότητα εμφάνισης συγκεκριμένου τύπου καρκίνου μέσω κάποιων παραγόντων, όπως είναι το φύλο, η κληρονομικότητα, ο τρόπος ζωής κ.ά.

Η μεταβλητή την οποία επιθυμούμε να προβλέψουμε ονομάζεται εξαρτημένη μεταβλητή (*dependent variable*) (ή μεταβλητή απόκρισης (*response variable*)) και η οποία πιθανόν σχετίζεται με κάποιες άλλες μεταβλητές μέσω και των οποίων θα κάνουμε την πρόβλεψη, που αποκαλούνται ανεξάρτητες (*independent*) (ή ερμηνευτικές (*explanatory*)) μεταβλητές.

Ο κλάδος της Στατιστικής που ασχολείται με το πλαίσιο αυτό καλείται Ανάλυση Παλινδρόμησης (*Regression Analysis*).

1.2 Γραμμική Παλινδρόμηση

Η μοντελοποίηση της σχέσης μεταξύ δύο ή και περισσότερων μεταβλητών, μπορεί να γίνει με ποικίλους τρόπους. Ένας από τους πιο γνωστούς και ευρέως χρησιμοποιούμενους τρόπους είναι μέσω του μοντέλου της γραμμικής παλινδρόμησης (*linear regression*). Στην περίπτωση ύπαρξης μόνο μίας ανεξάρτητης μεταβλητής μιλάμε για απλή γραμμική παλινδρόμηση (*simple*

linear regression), ενώ στην περίπτωση που υπάρχουν δύο ή και περισσότερες ανεξάρτητες μεταβλητές για πολλαπλή γραμμική παλινδρόμηση (*multiple linear regression*).

1.2.1 Το στατιστικό μοντέλο

Έστω η εξαρτημένη ποσοτική μεταβλητή Y, οι ανεξάρτητες μεταβλητές $X_1, X_2, \dots, X_d$ και δείγμα μεγέθους n. Το στατιστικό μοντέλο της πολλαπλής γραμμικής παλινδρόμησης αναφέρεται στη σχέση:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_d x_{id} + \varepsilon_i \,, \;\; i = 1,2, \dots, n$$

ή αλλιώς:

$$y_i = \beta_0 + \sum_{j=1}^{d} \beta_j x_{ij} + \varepsilon_i \,, \;\; i = 1,2, \dots, n \,,$$

όπου ισχύουν τα εξής:

1) $\beta_0, \beta_1, \dots, \beta_d$ άγνωστες παράμετροι. Οι παράμετροι $\beta_1, \beta_2, \dots \beta_d$ ποσοτικοποιούν τη σχέση της αντίστοιχης μεταβλητής $X_1, X_2, \dots, X_d$ με την εξαρτημένη μεταβλητή Y.
2) $x_{i1}, x_{i2}, \dots, x_{id}$, $i = 1,2, \dots, n$ οι παρατηρήσεις.
3) y_i , $i = 1,2, \dots, n$ η τιμή της εξαρτημένης μεταβλητής για την παρατήρηση i.
4) ε_i , $i = 1,2, \dots, n$ τυχαία σφάλματα με $E(\varepsilon_i) = 0$ και $V(\varepsilon_i) = \sigma^2$.
5) $Cov(\varepsilon_i, \varepsilon_j) = 0$ για $i \neq j$, δηλαδή τα σφάλματα θεωρούνται ασυσχέτιστα.

Η γραμμική σχέση (από όπου προέκυψε και το όνομα του μοντέλου) αφορά τις παραμέτρους $\beta_0, \beta_1, \dots, \beta_d$ και όχι τις ανεξάρτητες μεταβλητές $X_1, X_2, \dots, X_d$.

1.2.2 Το κανονικό μοντέλο

Αν πέρα από τις πέντε υποθέσεις που αναφέραμε στο 1.2.1, υποθέσουμε ότι τα τυχαία σφάλματα ε_i , $i = 1,2, \dots, n$ είναι ανεξάρτητα και ακολουθούν την Κανονική κατανομή (*Normal distribution*) με μέση τιμή 0 και διακύμανση σ^2 ($\varepsilon_i \sim N(0, \sigma^2)$), τότε καταλήγουμε στο, γραμμένο με μορφή πινάκων, κανονικό μοντέλο πολλαπλής παλινδρόμησης:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \,, \boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I_n})$$

με:

1) $\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$ διάστασης n × 1. Διάνυσμα που περιέχει τυχαίες μεταβλητές.

2) $\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1d} \\ 1 & x_{21} & \dots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nd} \end{bmatrix}$ διάστασης n × (d+1). Αποκαλείται πίνακας σχεδιασμού.

3) $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix}$ διάστασης (d+1) × 1. Διάνυσμα παραμέτρων.

4) $\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$ διάστασης n × 1. Διάνυσμα που περιέχει τυχαίες μεταβλητές.

5) $\boldsymbol{I_n}$ ο n × n μοναδιαίος πίνακας.

Από τα 1-5 έχουμε ότι:

$$\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I_n}).$$

Άρα βλέπουμε ότι τα Y_i ακολουθούν την Κανονική κατανομή με σταθερή διακύμανση.

Σε όσα αναφέρθηκαν μέχρι στιγμής έχουμε κάνει κάποιες υποθέσεις, όπως το ότι η εξαρτημένη μεταβλητή Y είναι ποσοτική, τα σφάλματα $\varepsilon_i \sim N(0, \sigma^2)$ και $Cov(\varepsilon_i, \varepsilon_j) = 0$ για $i \neq j$. Επομένως, μπορούμε να χρησιμοποιήσουμε το γραμμικό μοντέλο αν π.χ. τα Y_i δεν ακολουθούν την Κανονική κατανομή με σταθερή διακύμανση; Ή αν η εξαρτημένη μεταβλητή Y είναι μια δίτιμη μεταβλητή που λαμβάνει μόνο τιμές 0 και 1;

1.3 Γενικευμένα Γραμμικά Μοντέλα

Τα γενικευμένα γραμμικά μοντέλα, εν συντομία ΓΓΜ, (*generalized linear models, GLM*) προτάθηκαν από τους Nelder and Wedderburn (1972). Αυτό που παρουσίασαν ουσιαστικά είναι η γενίκευση και η επέκταση της γραμμικής παλινδρόμησης (εξού και η ονομασία γενικευμένα γραμμικά μοντέλα) έτσι ώστε να συμπεριλάβουν α) εξαρτημένη μεταβλητή που να μην ακολουθεί κατ'ανάγκη την Κανονική κατανομή και β) μη γραμμικές συναρτήσεις της μέσης τιμής.

Ο όρος γενικευμένα γραμμικά μοντέλα καλύπτει ένα αρκετά μεγάλο εύρος στατιστικών μοντέλων, όπως τη γραμμική παλινδρόμηση και μοντέλα με κατηγορικές μεταβλητές

απόκρισης, κάτι που τα καθιστά ένα από τα πιο σημαντικά και ευρέως χρησιμοποιούμενα στατιστικά μοντέλα, καθώς το ίδιο μοντέλο μπορούμε να το χρησιμοποιήσουμε ακόμα και αν η εξαρτημένη μεταβλητή Y δεν ακολουθεί την Κανονική κατανομή. Φυσικά, οι ανεξάρτητες μεταβλητές $X_1, X_2, \dots, X_d$ μπορούν να είναι ποσοτικές, ποιοτικές ή και συνδυασμός αυτών.

1.3.1 Η Εκθετική Οικογένεια Κατανομών

Έστω η μεταβλητή Y από κάποια κατανομή με παράμετρο θ και $(y_1, y_2, \dots, y_n)$ οι ανεξάρτητες παρατηρήσεις. Η συνάρτηση πυκνότητας πιθανότητας (*probability density function*), αν είναι συνεχής μεταβλητή, ή η συνάρτηση μάζας πιθανότητας (*probability mass function*), αν είναι διακριτή μεταβλητή, γράφεται ως:

$$f(y;\theta,\varphi) = exp\left[\frac{y\theta - b(\theta)}{\alpha(\varphi)} + c(y,\varphi)\right] \qquad (1.1.\alpha)$$

ή εναλλακτικά:

$$f(y;\theta,\varphi) = exp\left(\frac{y\theta}{\alpha(\varphi)}\right) exp\left(-\frac{b(\theta)}{\alpha(\varphi)}\right) exp(c(y,\varphi)) \qquad (1.1.\beta)$$

με:

1) $\alpha(\bullet)$, $b(\bullet)$ και $c(\bullet)$ γνωστές συναρτήσεις και
2) θ και φ παράμετροι. Η παράμετρος θ αποκαλείται φυσική (ή κανονική) παράμετρος (*natural (or canonical) parameter*) και η παράμετρος φ παράμετρος διακύμανσης (*dispersion parameter*).

Σε αυτήν την περίπτωση, η κατανομή αυτή λέμε ότι ανήκει στην εκθετική οικογένεια κατανομών.

Όπως θα δούμε αναλυτικά και παρακάτω (βλ. 1.3.3), πολλές από τις πιο γνωστές κατανομές ανήκουν στην εκθετική οικογένεια κατανομών, όπως, μεταξύ άλλων, η Κανονική, η Poisson, η Διωνυμική (*Binomial*) και η Γάμμα (*Gamma*).

1.3.2 Μέση τιμή και διακύμανση στην εκθετική οικογένεια κατανομών

Έστω $l_i = log(f(y_i;\theta_i,\varphi))$ ο λογάριθμος της πιθανοφάνειας (*log-likelihood*) για τα y_i και $l = \sum_{i=1}^{n} l_i$. Έχουμε:

$$l_i = log\left(exp\left[\frac{y_i\theta_i - b(\theta_i)}{\alpha(\varphi)} + c(y_i,\varphi)\right]\right) \Leftrightarrow$$

$$l_i = \frac{y_i\theta_i - b(\theta_i)}{\alpha(\varphi)} + c(y_i, \varphi). \tag{1.2}$$

Παραγωγίζοντας την (1.2) ως προς το θ_i, παίρνουμε:

$$\frac{\partial l_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{\alpha(\varphi)} \tag{1.3}$$

και:

$$\frac{\partial^2 l_i}{\partial \theta_i^2} = \frac{-b''(\theta_i)}{\alpha(\varphi)}, \tag{1.4}$$

όπου $b'(\theta_i)$ και $b''(\theta_i)$ η πρώτη και δεύτερη παράγωγος αντίστοιχα της συνάρτησης $b(\bullet)$ ως προς το θ_i. Γνωρίζοντας ότι ισχύει (Kendall and Stuart, 1967):

$$E\left(\frac{\partial l}{\partial \theta}\right) = 0 \tag{1.5}$$

και:

$$-E\left(\frac{\partial^2 l}{\partial \theta^2}\right) = E\left(\frac{\partial l}{\partial \theta}\right)^2, \tag{1.6}$$

από τις (1.3) και (1.5) καταλήγουμε στη σχέση:

$$E\left(\frac{y_i - b'(\theta_i)}{\alpha(\varphi)}\right) = E\big(y_i - b'(\theta_i)\big)/\alpha(\varphi) = 0 \Rightarrow$$

$$E\big(y_i - b'(\theta_i)\big) = 0.$$

Είναι εύκολο να δούμε ότι:

$$\mu_i = E(y_i) = b'(\theta_i). \tag{1.7}$$

Από τις (1.4) και (1.6) έχουμε ότι:

$$-\left(\frac{-b''(\theta_i)}{\alpha(\varphi)}\right) = E\left(\frac{y_i - b'(\theta_i)}{\alpha(\varphi)}\right)^2 \Leftrightarrow$$

$$\frac{b''(\theta_i)}{\alpha(\varphi)} = \frac{E\big(y_i - E(y_i)\big)^2}{\alpha(\varphi)^2} \Leftrightarrow$$

$$b''(\theta_i) = \frac{Var(y_i)}{\alpha(\varphi)}$$

και άρα:

$$Var(y_i) = b''(\theta_i)\alpha(\varphi). \tag{1.8}$$

1.3.3 Παραδείγματα κατανομών της εκθετικής οικογένειας

- Κανονική κατανομή: $Y \sim N(\mu, \sigma^2)$

Η συνάρτηση πυκνότητας πιθανότητας είναι η:

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right], \quad y, \mu \in \mathbb{R},$$

που μπορεί να γραφτεί και ως:

$$f(y;\mu,\sigma^2) = exp\left[log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\right] exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] =$$

$$exp\left[\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{1}{2}log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2}\right],$$

όπου θέτοντας:

$$\theta = \mu, \quad b(\theta) = \frac{1}{2}\mu^2 = \frac{1}{2}\theta^2, \quad \alpha(\varphi) = \sigma^2,$$

$$c(y;\varphi) = -\frac{1}{2}log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2},$$

καταλήγουμε στη σχέση (1.1.α). Οπότε η Κανονική κατανομή ανήκει στην εκθετική οικογένεια κατανομών με φυσική παράμετρο $\theta = \mu$. Από τη σχέση (1.7) για τη μέση τιμή έχουμε:

$$E(Y) = b'(\theta) = \theta = \mu$$

και από τη σχέση (1.8) για τη διακύμανση:

$$Var(Y) = b''(\theta)\alpha(\varphi) = \sigma^2.$$

- Κατανομή Poisson: $Y \sim Poisson(\lambda)$

Η συνάρτηση μάζας πιθανότητας είναι η:

$$f(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}, \quad y = 0,1,2,\ldots \kappa\alpha\iota\ \lambda > 0.$$

Γράφοντάς την ως:

$$f(y;\lambda) = exp\left[log\left(\frac{e^{-\lambda}\lambda^y}{y!}\right)\right] =$$

$$exp[-\lambda + ylog(\lambda) - log(y!)] =$$

$$exp[ylog(\lambda) - exp(log(\lambda)) - log(y!)]$$

και θέτοντας:

$$\theta = log(\lambda)\,,\;\; b(\theta) = exp(\theta)\,,\;\; \alpha(\varphi) = 1\,,\;\; c(y;\varphi) = -log(y!),$$

βλέπουμε ότι ικανοποιείται η σχέση (1.1.α). Επομένως ανήκει στην εκθετική οικογένεια κατανομών με φυσική παράμετρο $\theta = log(\lambda)$. Αντίστοιχα, για τη μέση τιμή και τη διακύμανση παίρνουμε:

$$E(Y) = b'(\theta) = exp(\theta) = exp(log(\lambda)) = \lambda$$

και:

$$Var(Y) = b''(\theta)\alpha(\varphi) = \lambda.$$

- <u>Διωνυμική κατανομή: $Y \sim Bin(n,p)$</u>

Πλήθος δοκιμών n και p η πιθανότητα επιτυχίας για κάθε δοκιμή. Η συνάρτηση μάζας πιθανότητας για γνωστό n είναι η:

$$f(y;p) = \binom{n}{y} p^y (1-p)^{n-y}\,,\;\; y = 0,1,\dots,n \;\kappa\alpha\iota\; 0 < p < 1$$

ή αλλιώς:

$$f(y;p) = exp\left[log\left(\binom{n}{y} p^y (1-p)^{n-y}\right)\right] =$$

$$exp\left[log\binom{n}{y} + log(p^y) + log(1-p)^{n-y}\right] =$$

$$exp\left[ylog(p) + (n-y)log\,(1-p) + log\binom{n}{y}\right] =$$

$$exp\left[ylog\left(\frac{p}{1-p}\right) + nlog(1-p) + log\binom{n}{y}\right].$$

Θέτουμε:

$$\theta = log\left(\frac{p}{1-p}\right) \;\Rightarrow\; p = \frac{e^\theta}{1+e^\theta}\,,$$

$$b(\theta) = nlog(1-p) = nlog\left(1+e^\theta\right),$$

$$\alpha(\varphi) = 1\,,\;\; c(y;\varphi) = log\binom{n}{y}.$$

Βλέπουμε ότι ανήκει στην εκθετική οικογένεια κατανομών με φυσική παράμετρο $\theta = log\left(\frac{p}{1-p}\right)$. Για τη μέση τιμή και τη διακύμανση έχουμε:

$$E(Y) = b'(\theta) = \frac{ne^\theta}{1+e^\theta} = np$$

και:

$$Var(Y) = b''(\theta)\alpha(\varphi) = \frac{ne^{\theta}}{(1+e^{\theta})^2} = np(1-p).$$

1.3.4 Οι συνιστώσες των ΓΓΜ

Τα ΓΓΜ απαρτίζονται από τρεις συνιστώσες: το στοχαστικό μέρος (*random component*), τη γραμμική προβλέπουσα (*linear predictor*) και τη συνάρτηση σύνδεσης (*link function*).

Το **στοχαστικό μέρος** αποτελείται από μια εξαρτημένη μεταβλητή Y με τις τυχαίες μεταβλητές $(Y_1, Y_2, \dots Y_n)$ να είναι ανεξάρτητες μεταξύ τους. Η Y θα πρέπει να έχει συνάρτηση πυκνότητας πιθανότητας ή συνάρτηση μάζας πιθανότητας που να ανήκει στην εκθετική οικογένεια κατανομών.

Η **γραμμική προβλέπουσα** είναι μια συνάρτηση μέσω της οποίας οι ανεξάρτητες μεταβλητές $X_1, X_2, \dots, X_d$ εισάγονται γραμμικά ως προγνωστικοί παράγοντες στη δεξιά πλευρά του μοντέλου. Για δείγμα μεγέθους n και $x_{ij}, i = 1,2, \dots, n$ και $j = 1,2, \dots, d$ η τιμή της παρατήρησης i για τη μεταβλητή j, έχουμε:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{id} x_{id}$$

ή:

$$\eta_i = \beta_0 + \sum_{j=1}^{d} \beta_j x_{ij}\ , i = 1,2, \dots, n. \tag{1.9}$$

Το παραπάνω γράφεται και σε μορφή πινάκων ως εξής:

$$\boldsymbol{\eta} = \boldsymbol{X\beta},$$

όπου $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_n)^T$, $\boldsymbol{\beta}$ ένα διάνυσμα στήλης διάστασης (d+1) × 1 και $\boldsymbol{X}$ ο πίνακας σχεδιασμού διάστασης n × (d+1). Οι μεταβλητές $X_1, X_2, \dots, X_d$ μπορούν να είναι και μη γραμμικές συναρτήσεις άλλων μεταβλητών, π.χ. ένας όρος αλληλεπίδρασης $X_1 = X_2 * X_3$ ή ένας τετραγωνικός όρος $X_1 = X_2^2$.

Η **συνάρτηση σύνδεσης** είναι αυτή που θα συνδέσει το στοχαστικό μέρος με τη γραμμική προβλέπουσα. Έστω $\mu_i = E(y_i)$, $i = 1,2, \dots, n$ η αναμενόμενη τιμή της εξαρτημένης μεταβλητής για την παρατήρηση i, η οποία εξαρτάται από τις τιμές των x_{ij} με $j = 1,2, \dots, d$, και *g(•)* μια μονότονη και διαφορίσιμη συνάρτηση. Η g συνδέει την αναμενόμενη τιμή μ_i με τις ανεξάρτητες μεταβλητές μέσω της σχέσης:

$$\eta_i = g(\mu_i) = \beta_0 + \sum_{j=1}^{d} \beta_j x_{ij} \ , i = 1,2, \dots, n.$$

Η πιο απλή μορφή συνάρτησης σύνδεσης είναι αυτή που μοντελοποιεί απευθείας την αναμενόμενη τιμή:

$$g(\mu_i) = \mu_i \ \Leftrightarrow$$

$$\mu_i = \beta_0 + \sum_{j=1}^{d} \beta_j x_{ij} \ , i = 1,2, \dots, n \ ,$$

που μας δίνει το μοντέλο της γραμμικής παλινδρόμησης. Αυτή η συνάρτηση σύνδεσης αποκαλείται ταυτοτική συνάρτηση σύνδεσης (*identity link function*). Άλλες συναρτήσεις σύνδεσης συνδέουν το μ_i με τις ανεξάρτητες μεταβλητές μέσω μη γραμμικής σχέσης, π.χ.:

$$g(\mu_i) = log(\mu_i) \ , \mu_i > 0 \ \Leftrightarrow$$

$$log\ (\mu_i) = \beta_0 + \sum_{j=1}^{d} \beta_j x_{ij} \ , \mu_i > 0 \ , i = 1,2, \dots, n \ ,$$

που συνδέει το λογάριθμο της αναμενόμενης τιμής με τις ανεξάρτητες μεταβλητές. Τα ΓΓΜ που χρησιμοποιούν το λογάριθμο ως συνάρτηση σύνδεσης λέγονται λογαριθμογραμμικά μοντέλα (*loglinear models*).

Επειδή η συνάρτηση σύνδεσης είναι αντιστρέψιμη, έχουμε:

$$\mu_i = g^{-1}(\eta_i) = g^{-1}\left(\beta_0 + \sum_{j=1}^{d} \beta_j x_{ij}\right), i = 1,2, \dots, n \qquad (1.10)$$

και έτσι τα ΓΓΜ μπορούν να θεωρηθούν ως ένα γραμμικό μοντέλο που μετασχηματίζει την αναμενόμενη τιμή μ_i ή ως ένα μη γραμμικό μοντέλο παλινδρόμησης για την εξαρτημένη μεταβλητή *Y*. Η $g^{-1}(\bullet)$ αποκαλείται μέση συνάρτηση (*mean function*).

Κάθε κατανομή που ανήκει στην εκθετική οικογένεια έχει μία παράμετρο η οποία λειτουργεί ως η φυσική της παράμετρος. Όπως είδαμε και στο 1.3.3, αυτή η παράμετρος είναι η μέση τιμή μ για την Κανονική κατανομή, το $log(\mu)$ για την Poisson και το $log(\frac{\mu}{1-\mu})$ για την Διωνυμική. Η συνάρτηση σύνδεσης g που μετασχηματίζει την αναμενόμενη τιμή μ_i στη φυσική της παράμετρο ονομάζεται κανονική συνάρτηση σύνδεσης (*canonical link function*). Τα διάφορα λογισμικά που χρησιμοποιούνται για τα ΓΓΜ ως προεπιλογή χρησιμοποιούν την κανονική συνάρτηση σύνδεσης. Μέσω της κανονικής συνάρτησης σύνδεσης προκύπτει ότι ο λογάριθμος της πιθανοφάνειας είναι κοίλη συνάρτηση και άρα έχει μοναδική μέγιστη τιμή.

ΠΙΝΑΚΑΣ 1-1 : Βασικές συναρτήσεις σύνδεσης και οι αντίστροφές τους.

ΚΑΤΑΝΟΜΗ	**ΣΥΝΑΡΤΗΣΗ ΣΥΝΔΕΣΗΣ** $\boldsymbol{\eta_i = g(\mu_i)}$	**ΑΝΤΙΣΤΡΟΦΗ ΣΥΝΑΡΤΗΣΗ** $\boldsymbol{\mu_i = g^{-1}(\eta_i)}$
Κανονική	$g(\mu_i) = \mu_i$ (identity link)	$g^{-1}(\eta_i) = \eta_i$
Poisson	$g(\mu_i) = log\ (\mu_i)$ (log link)	$g^{-1}(\eta_i) = e^{\eta_i}$
Γάμμα	$g(\mu_i) = \frac{1}{\mu_i}$ (inverse link)	$g^{-1}(\eta_i) = \frac{1}{\eta_i}$
Διωνυμική	$g(\mu_i) = logit(\mu_i) = log\ (\frac{\mu_i}{1-\mu_i})$ (logit link)	$g^{-1}(\eta_i) = \frac{1}{1+e^{-\eta_i}}$
Διωνυμική	$g(\mu_i) = \Phi^{-1}(\mu_i)$ (probit link)	$g^{-1}(\eta_i) = \Phi(\eta_i)$
Διωνυμική	$g(\mu_i) = log\ [-log(1-\mu_i)]$ (clog-log link)	$g^{-1}(\eta_i) = 1 - exp\ [-exp(\eta_i)]$

1.3.5 Εκτιμήσεις των παραμέτρων και διαστήματα εμπιστοσύνης

Μέσω της μεθοδολογίας της μέγιστης πιθανοφάνειας (*maximum likelihood*), θα υπολογίσουμε τις εκτιμήσεις των β_j. Οι εξισώσεις πιθανοφάνειας ενός ΓΓΜ δίνονται από τη σχέση:

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = 0\ , \gamma\iota\alpha\ \kappa\acute{\alpha}\theta\varepsilon\ j.$$

Από τη σχέση (1.2), ορίζουμε τον λογάριθμο της πιθανοφάνειας για n ανεξάρτητες παρατηρήσεις και διάνυσμα παραμέτρων $\boldsymbol{\beta}$:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} log\big(f(y_i; \theta_i, \varphi)\big) =$$

$$\sum_{i=1}^{n} \left(\frac{y_i\theta_i - b(\theta_i)}{\alpha(\varphi)} + c(y_i, \varphi) \right) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\alpha(\varphi)} + \sum_{i=1}^{n} c(y_i, \varphi),$$

όπου παραγωγίζοντας ως προς β_j και χρησιμοποιώντας τον κανόνα της αλυσίδας (*chain rule*) έχουμε:

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}.$$

Στο 1.3.2 και 1.3.4 είδαμε ότι (σχέσεις (1.3), (1.7), (1.8) και (1.9) αντίστοιχα):

$$\frac{\partial l_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{\alpha(\varphi)} \quad , \quad \mu_i = E(y_i) = b'(\theta_i)$$

$$Var(y_i) = b''(\theta_i)\alpha(\varphi) \ , \ \eta_i = \beta_0 + \sum_{j=1}^{d} \beta_j x_{ij}.$$

Άρα προκύπτει ότι:

$$\frac{\partial l_i}{\partial \theta_i} = \frac{y_i - \mu_i}{\alpha(\varphi)} \ , \ \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{Var(y_i)}{\alpha(\varphi)} \ , \ \frac{\partial \eta_i}{\partial \beta_j} = x_{ij}.$$

Ακόμα, επειδή το $\eta_i = g(\mu_i)$, το $\frac{\partial \mu_i}{\partial \eta_i}$ εξαρτάται από τη συνάρτηση σύνδεσης. Αντικαθιστώντας τα παραπάνω καταλήγουμε ότι:

$$\frac{\partial l_i}{\partial \beta_j} = \frac{y_i - \mu_i}{\alpha(\varphi)} \frac{\alpha(\varphi)}{Var(y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} = \frac{(y_i - \mu_i)x_{ij}}{Var(y_i)} \frac{\partial \mu_i}{\partial \eta_i}$$

και αθροίζοντας για τις n παρατηρήσεις:

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{n} \left[\frac{(y_i - \mu_i)x_{ij}}{Var(y_i)} \frac{\partial \mu_i}{\partial \eta_i} \right] = 0 \ , για\ κάθε\ j.$$

Είναι εύκολα αντιληπτό ότι διαφορετικές συναρτήσεις σύνδεσης, θα δώσουν διαφορετικές εξισώσεις πιθανοφάνειας. Επίσης, από ότι βλέπουμε, τα β_j δεν εμφανίζονται άμεσα στις εξισώσεις πιθανοφάνειας, αλλά εμφανίζονται μέσω της σχέσης (1.10):

$$\mu_i = g^{-1}(\eta_i) = g^{-1}\left(\beta_0 + \sum_{j=1}^{d} \beta_j x_{ij} \right), i = 1,2, \dots, n.$$

Ακόμα, οι εξισώσεις πιθανοφάνειας αποτελούν μη γραμμικές συναρτήσεις των β_j. Για να βρούμε τις εκτιμήσεις των β_j, θα πρέπει να λύσουμε τις εξισώσεις πιθανοφάνειας με χρήση επαναληπτικών μεθόδων, όπως είναι η μέθοδος Newton-Raphson.

Τα διαστήματα εμπιστοσύνης (*confidence intervals*) για τις παραμέτρους β_j για κάποιο προκαθορισμένο επίπεδο σημαντικότητας α δίνονται από τον τύπο:

$$\hat{\beta}_j \pm z_{1-a/2}\left[se\left(\hat{\beta}_j\right)\right].$$

1.3.6 Απόκλιση

Η απόκλιση (*deviance*) στα ΓΓΜ αποτελεί μία γενίκευση της έννοιας του αθροίσματος των τετραγώνων των καταλοίπων (*residual sum of squares, RSS*) της γραμμικής παλινδρόμησης, όπου $SSE = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 \ , i = 1,2, \dots, n$, και $y_i - \hat{\mu}_i$ η διαφορά της προβλεπόμενης αναμενόμενης τιμής του μοντέλου από την παρατηρηθείσα τιμή.

Έστω L_M η συνάρτηση της πιθανοφάνειας του υπό εξέταση γενικευμένου γραμμικού μοντέλου M και L_S η συνάρτηση της πιθανοφάνειας του κορεσμένου μοντέλου S (*saturated model*), του μοντέλου δηλαδή που περιέχει τόσες παραμέτρους όσες παρατηρήσεις (πλήθος παραμέτρων d_S, πλήθος παρατηρήσεων n, $d_S = n$). Τα δύο αυτά μοντέλα ακολουθούν την ίδια κατανομή και έχουν την ίδια συνάρτηση σύνδεσης *g(•)*. Ως απόκλιση ενός ΓΓΜ ορίζεται η σχέση:

$$D = -2log\left(\frac{L_M}{L_S}\right). \tag{1.11}$$

Θέτοντας:

$$l_M = log(L_M) \ \ , \ \ l_S = log(L_S),$$

η (1.11) μπορεί να γραφτεί:

$$D = -2(l_M - l_S)$$

και ασυμπτωτικά ισχύει $D \sim \chi^2_{df}$ με βαθμούς ελευθερίας $df = d_S - d_M$, όπου d_s το πλήθος των παραμέτρων του κορεσμένου μοντέλου και d_M το πλήθος των παραμέτρων του υπό εξέταση μοντέλου. Όσο μικρότερη η απόκλιση, δηλαδή το D, τόσο το μοντέλο M βρίσκεται κοντά στο κορεσμένο και άρα το μοντέλο M προσαρμόζεται καλά στα δεδομένα (ένδειξη καλής προσαρμογής). Επίσης, μοντέλα με λιγότερες παραμέτρους έχουν μεγαλύτερη απόκλιση.

Η απόκλιση προκύπτει από τον έλεγχο του λόγου πιθανοφανειών (*likelihood ratio test*) και είναι ένας στατιστικός έλεγχος για την υπόθεση ότι οι παράμετροι που υπάρχουν στο κορεσμένο μοντέλο αλλά όχι στο υπό εξέταση μοντέλο, είναι όλοι 0. Έστω $\beta_1, \beta_2, \dots, \beta_{d_M}$ οι παράμετροι του υπό εξέταση μοντέλου και $\beta_1, \beta_2, \dots, \beta_{d_S}$ οι παράμετροι του κορεσμένου μοντέλου, με $d_M < d_S$. Οπότε η παραπάνω υπόθεση μπορεί να διατυπωθεί ως:

$$H_0 : \beta_{d_M+1} = \beta_{d_M+2} = \dots = \beta_{d_S} = 0$$

και

$$H_1 : \textit{τουλάχιστον μία παράμετρος από τις } \beta_{d_M+1}, \beta_{d_M+2}, \dots, \beta_{d_S} \textit{ δεν είναι 0.}$$

Η απόκλιση μπορεί να χρησιμοποιηθεί και για σύγκριση μεταξύ δύο εμφωλευμένων μοντέλων (*nested models*), ώστε να βρούμε το βέλτιστο μεταξύ των δύο. Έστω το μοντέλο M_1 με πλήθος παραμέτρων d_1 και απόκλιση D_1 και το μοντέλο M_2 με πλήθος παραμέτρων d_2 και απόκλιση D_2. Επίσης, όλες οι d_1 παράμετροι του M_1 είναι υποσύνολο των d_2 παραμέτρων του M_2, άρα τα μοντέλα είναι εμφωλευμένα. Η διαφορά των αποκλίσεων των δύο μοντέλων είναι ίση με:

$$D_1 - D_2 = \left(-2\left(l_{M_1} - l_S\right)\right) - \left(-2\left(l_{M_2} - l_S\right)\right) = -2\left(l_{M_1} - l_{M_2}\right) \sim \chi^2_{d_2 - d_1}.$$

Η μηδενική υπόθεση ορίζεται ως:

$$H_0 : \beta_{d_1+1} = \beta_{d_1+2} = \ldots = \beta_{d_2} = 0$$

με:

$$H_1 : \textit{τουλάχιστον μία παράμετρος από τις } \beta_{d_1+1}, \beta_{d_1+2}, \ldots, \beta_{d_2} \textit{ δεν είναι 0.}$$

Μια διαφορετική διατύπωση των παραπάνω είναι η εξής:

$$H_0 : \textit{Ισχύει το } M_1$$

και

$$H_1 : \textit{Ισχύει το } M_2.$$

1.3.7 Έλεγχος Wald

Έστω το ΓΓΜ με εξαρτημένη μεταβλητή Y, $X_1, X_2, \ldots, X_d$ ανεξάρτητες μεταβλητές, $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_d$ οι εκτιμητές μέγιστης πιθανοφάνειας (ΕΜΠ) των d παραμέτρων και θέλουμε να ελέγξουμε αν μία μεταβλητή, π.χ. η X_1, είναι στατιστικά σημαντική. Δηλαδή, θέλουμε να ελέγξουμε την υπόθεση:

$$H_0 : \beta_j = 0$$

και

$$H_1 : \beta_j \neq 0.$$

Την υπόθεση αυτή μπορούμε να την ελέγξουμε μέσω του ελέγχου Wald (*Wald test*):

$$Z = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)},$$

όπου $se(\hat{\beta}_j)$ το τυπικό σφάλμα (*standard error*) του $\hat{\beta}_j$. Ασυμπτωτικά ισχύει ότι $Z \sim N(0,1)$. Επομένως, αν ισχύει $|Z| > z_{1-a/2}$, όπου $z_{1-a/2}$ το ποσοστιαίο σημείο της Τυποποιημένης Κανονικής κατανομής για κάποιο προκαθορισμένο επίπεδο σημαντικότητας α, τότε απορρίπτουμε τη μηδενική υπόθεση και άρα η μεταβλητή X_j είναι στατιστική σημαντική.

ΚΕΦΑΛΑΙΟ 2

Η Λογιστική Παλινδρόμηση

2.1 Εισαγωγή

Η Στατιστική είναι μία επιστήμη η οποία έχει εφαρμογή σε πάρα πολλά και διαφορετικά πεδία: από τη Βιολογία, την Ιατρική, την Πληροφορική και τη Φυσική μέχρι την Κοινωνιολογία και την Ψυχολογία. Ειδικά τα τελευταία χρόνια, λόγω και της ταχείας ανάπτυξης των υπολογιστικών συστημάτων, πάρα πολλές στατιστικές μέθοδοι έχουν βρει ευρεία χρήση στον τομέα της Μηχανικής Μάθησης (*Machine Learning*), όπως είναι η ανάλυση κύριων συνιστωσών (*principal component analysis-PCA*), η ανάλυση κατά συστάδες ή συσταδοποίηση (*cluster analysis or clustering*), τα δέντρα αποφάσεων *(decision trees)* και ο απλός ταξινομητής Bayes (*naive Bayes classifier*). Μία από αυτές τις μεθόδους είναι και η λογιστική παλινδρόμηση (*logistic regression*), ως αλγόριθμος επιβλεπόμενης μάθησης (*supervised learning algorithm*).

Η λογιστική παλινδρόμηση χρησιμοποιείται όταν η εξαρτημένη μεταβλητή *Y* είναι δίτιμη (βλ. 2.2). Χρησιμοποιείται για δύο σκοπούς: για σκοπούς πρόβλεψης (*prediction*) ή για σκοπούς ταξινόμησης (*classification*) παρατηρήσεων. Ταξινόμηση είναι η ένταξη μίας παρατήρησης σε μία ομάδα, λαμβάνοντας υπ'όψιν τις τιμές των ανεξάρτητων μεταβλητών. Για παράδειγμα, μία τράπεζα θέλει να ταξινομήσει τους πελάτες της, θέτοντας κάποια κριτήρια, σε δύο ομάδες:

1) Ομάδα Α, καλοί πελάτες και
2) Ομάδα Β, κακοί πελάτες.

Για την ταξινόμηση θα χρησιμοποιήσει διάφορες πληροφορίες που διαθέτει για τον κάθε πελάτη ξεχωριστά· ετήσιο εισόδημα, επάγγελμα, ηλικία, αριθμός πιστωτικών καρτών, ύπαρξη χρεών κλπ. Άλλο κλασσικό παράδειγμα ταξινόμησης είναι αυτό για τα εισερχόμενα e-mails, όπου χωρίζονται σε spam ή όχι.

Στο 2ο κεφάλαιο θα γίνει μια σύντομη παρουσίαση των δίτιμων δεδομένων, θα δούμε γιατί τελικά το μοντέλο της γραμμικής παλινδρόμησης δεν είναι ικανό να περιγράψει τις δίτιμες μεταβλητές και θα αναλυθεί το μοντέλο της λογιστικής παλινδρόμησης.

2.2 Δίτιμα Δεδομένα

Δίτιμα δεδομένα (*binary data*), ή διχοτομικά δεδομένα (*dichotomous data*), αποκαλούμε τα δεδομένα τα οποία παίρνουν μόνο δύο διαφορετικές τιμές, με τη μία από αυτή να τη θεωρούμε αυθαίρετα ως επιτυχία (*success*), ενώ την άλλη ως αποτυχία (*failure*). Π.χ. η ρίψη ενός νομίσματος με τις πιθανές τιμές να είναι κορώνα ή γράμματα, η πορεία ενός πιστωτικού ιδρύματος με πιθανές τιμές χρεωκοπία ή όχι, αν κάποιο εμβόλιο είναι αποτελεσματικό για μία αρρώστια ή όχι κλπ. Μία άλλη ερμηνεία των τιμών είναι η εμφάνιση ή όχι ενός γεγονότος. Συνήθης κωδικοποίηση, όπως και θα τη χρησιμοποιούμε από εδώ και στο εξής, είναι επιτυχία = 1 και αποτυχία = 0.

2.2.1 Η κατανομή Bernoulli

Ας υποθέσουμε ότι έχουμε την τυχαία μεταβλητή X και έστω ότι διενεργούμε ένα πείραμα μόνο μία φορά. Το πείραμα αυτό έχει δύο πιθανά αποτελέσματα· το ένα αποτέλεσμα το θεωρούμε ως επιτυχία και το άλλο ως αποτυχία. Ακόμα, θεωρούμε ως πιθανότητα p, με $0 \leq p \leq 1$, το αποτέλεσμα του πειράματος να είναι 1 και $q = 1 - p$ την πιθανότητα να είναι 0. Τότε, λέμε ότι η X ακολουθεί την κατανομή Bernoulli με παράμετρο p ($X \sim Bernoulli(p)$) και:

$$X = \begin{cases} 0, & με\ πιθανότητα\ q \\ 1, & με\ πιθανότητα\ p \end{cases},$$

με συνάρτηση μάζας πιθανότητας:

$$f(x) = P(X = x) = \begin{cases} p, & x = 1 \\ q, & x = 0 \end{cases},$$

ή:

$$f(x) = P(X = x) = p^x q^{1-x}, \quad x = 0,1$$

και συνάρτηση κατανομής (*distribution function*):

$$F(t) = \begin{cases} 0, & t < 0 \\ q, & 0 \leq t < 1 \\ 1, & t \geq 1 \end{cases}.$$

Επίσης, είναι εύκολο να δούμε ότι:

$$E(X) = \sum_{x=0}^{1} xf(x) = 0 \cdot f(0) + 1 \cdot f(1) = 0 \cdot q + 1 \cdot p = p$$

και:

$$Var(X) = E(X^2) - [E(X)]^2 = \sum_{x=0}^{1} x^2 f(x) - p^2 \Leftrightarrow$$

$$Var(X) = 0^2 \cdot f(0) + 1^2 \cdot f(1) - p^2 = p - p^2 = p(1-p) = pq.$$

Τέτοιου είδους πειράματα είναι γνωστά ως δοκιμές Bernoulli (*Bernoulli trials*). Αποκαλούνται έτσι λόγω του Σουηδού μαθηματικού Jacob Bernoulli του 17ου αιώνα, όπου και τις ανέλυσε στο βιβλίο του Ars Conjectandi (Bernoulli, 1713).

2.2.2 Η Διωνυμική κατανομή

Έστω ότι εκτελούμε το ίδιο πείραμα όχι μόνο μία φορά, αλλά n ανεξάρτητες φορές. Άρα το πείραμα αποτελείται από n δοκιμές Bernoulli και το αποτέλεσμα της μίας δοκιμής δεν επηρεάζει το αποτέλεσμα της άλλης. Οπότε, έχουμε $X_1, X_2, \dots, X_n$ τυχαίες μεταβλητές που ακολουθούν την κατανομή Bernoulli, $X = \sum_{i=1}^{n} X_i$ ορίζουμε τον αριθμό των επιτυχιών που θα εμφανιστούν στο πείραμα, p την πιθανότητα επιτυχίας και q την πιθανότητα αποτυχίας, με p,q σταθερά κατά τη διεξαγωγή του πειράματος. Τότε, η τυχαία μεταβλητή X λέμε ότι ακολουθεί τη Διωνυμική κατανομή με παραμέτρους n,p (*X~Bin(n,p)*) με συνάρτηση μάζας πιθανότητας:

$$f(x) = P(X = x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0,1, \dots, n \text{ και } \binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

και συνάρτηση κατανομής:

$$F(t) = \begin{cases} 0, & t < 0 \\ \sum_{x=0}^{t} \binom{n}{x} p^x q^{n-x}, & 0 \leq t < n \\ 1, & t \geq n \end{cases}.$$

Για τη μέση τιμή και τη διακύμανση ισχύει αντίστοιχα $E(X) = np$ και $Var(X) = np(1-p) = npq$.

Η κατανομή Bernoulli αποτελεί ειδική περίπτωση της Διωνυμικής για $n = 1$. Επίσης, όπως αποδείχτηκε και στο 1.3.3, η Διωνυμική ανήκει στην εκθετική οικογένεια κατανομών. Ακόμα, όταν το n είναι αρκετά μεγάλο, η Διωνυμική κατανομή μπορεί να περιγραφτεί από την Κανονική κατανομή με μέση τιμή $E(X) = np$ και $Var(X) = npq$.

2.3 Σχετική Πιθανότητα

Η σχετική πιθανότητα ($odds$) είναι μία αρκετά γνωστή έννοια στο πεδίο της Στατιστικής. Έστω το ενδεχόμενο Α με $0 \leq P(A) \leq 1$ την πιθανότητα να συμβεί το Α και με $1 - P(A)$ τη συμπληρωματική του πιθανότητα, να μη συμβεί το Α. Ως σχετική πιθανότητα του Α ορίζεται η σχέση:

$$odds = \frac{P(A)}{1 - P(A)}, \tag{2.1}$$

όπου δηλώνει πόσες φορές είναι πιο πιθανό να συμβεί το ενδεχόμενο Α από το να μη συμβεί. Έχουμε:

1) $odds \geq 0$.
2) Αν $odds = 1$, τότε η πιθανότητα εμφάνισης του ενδεχόμενου Α είναι ίση με την πιθανότητα μη εμφάνισής του.
3) Αν $odds = \alpha$ και $a > 1$, τότε η πιθανότητα εμφάνισης του ενδεχομένου Α είναι $(\alpha - 1)100\%$ μεγαλύτερη από την πιθανότητα μη εμφάνισής του.
4) Αν $odds = \alpha$ και $a < 1$, τότε η πιθανότητα εμφάνισης του ενδεχομένου Α είναι $(1 - \alpha)100\%$ μικρότερη από την πιθανότητα μη εμφάνισής του.

Από την (2.1), εύκολα δείχνεται ότι:

$$P(A) = \frac{odds}{odds + 1}.$$

Επίσης, λογαριθμίζοντας και τα δύο μέρη στην (2.1), παίρνουμε:

$$log(odds) = log\left(\frac{P(A)}{1 - P(A)}\right) \Leftrightarrow$$

$$logit[P(A)] = log\left(\frac{P(A)}{1 - P(A)}\right),$$

με τη συνάρτηση logit να υποδηλώνει το λογάριθμο της σχετικής πιθανότητας (*log-odds*) του ενδεχομένου Α.

2.4 Εισαγωγή στη Λογιστική Παλινδρόμηση

2.4.1 Το γραμμικό μοντέλο πιθανότητας

Όπως είδαμε στο 1ο κεφάλαιο (βλ. 1.2), όταν η εξαρτημένη μεταβλητή Y είναι ποσοτική, για ανεξάρτητες μεταβλητές $X_1, X_2, \dots, X_d$ και δείγμα μεγέθους n μπορούμε να εφαρμόσουμε το κλασσικό μοντέλο της γραμμικής παλινδρόμησης:

$$E(Y_i) = \beta_0 + \sum_{j=1}^{d} \beta_j x_{ij} \ , \ \ i = 1,2, \dots, n \ ,$$

καθώς αυτό που θέλουμε να προβλέψουμε είναι η αναμενόμενη τιμή της Y.

Ας θεωρήσουμε ότι η τυχαία μεταβλητή Y_i για $i = 1,2, \dots, n$ ακολουθεί προσεγγιστικά την Κανονική κατανομή με παραμέτρους p_i και σ^2 ($Y_i \sim N(p_i, \sigma^2)$) και οι τιμές που παίρνει είναι 0 με πιθανότητα q_i και 1 με πιθανότητα p_i. Επομένως, έχουμε την εξής σχέση:

$$y_i = \beta_0 + \sum_{j=1}^{d} \beta_j x_{ij} + \varepsilon_i \ , \ \ i = 1,2, \dots, n$$

και:

$$E(Y_i) = p_i = \beta_0 + \sum_{j=1}^{d} \beta_j x_{ij} \ , \ \ i = 1,2, \dots, n.$$

Το μοντέλο αυτό αποκαλείται γραμμικό μοντέλο πιθανότητας (*linear probability model*), το οποίο δεν είναι κατάλληλο για να περιγράψει τη σχέση μεταξύ της πιθανότητας p_i και των ανεξάρτητων μεταβλητών $X_1, X_2, \dots, X_d$ για τους εξής, μεταξύ άλλων, λόγους (Fox, 2015):

1) Τα τυχαία σφάλματα ε_i δεν ακολουθούν την Κανονική κατανομή. Είναι εύκολο να δούμε ότι:

$$\varepsilon_i = Y_i - \left(\beta_0 + \sum_{j=1}^{d} \beta_j x_{ij} \right) = Y_i - p_i.$$

Επομένως, τα τυχαία σφάλματα μπορούν να πάρουν μόνο δύο τιμές:

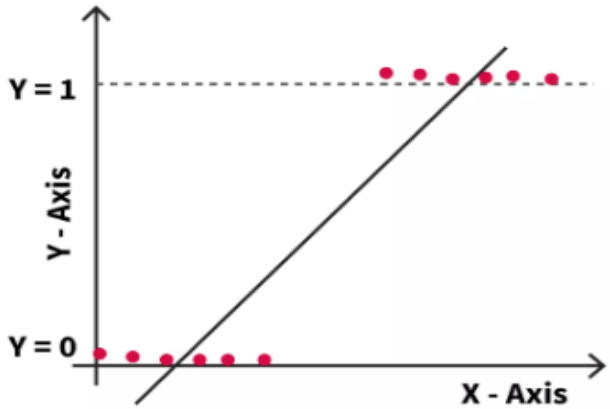
- Αν $Y_i = 0$, τότε $\varepsilon_i = 0 - p_i = -p_i$.
- Αν $Y_i = 1$, τότε $\varepsilon_i = 1 - p_i$.

2) Η διακύμανση των ε_i δεν είναι σταθερή. Αποδεικνύεται ότι:

$$V(\varepsilon_i) = E(Y_i^2) - E(Y_i)^2 = p_i - p_i^2 = p_i(1 - p_i).$$

Άρα, η διακύμανση των ε_i εξαρτάται από τις τιμές των p_i.

3) Τέλος, το βασικότερο πρόβλημα είναι ότι, λόγω της γραμμικής σχέσης, οι εκτιμήσεις των p_i δεν είναι απαραίτητο ότι θα βρίσκονται εντός του επιτρεπτού διαστήματος [0,1], τιμές δηλαδή που δεν έχουν νόημα, καθότι αναφερόμαστε σε πιθανότητες.



ΣΧΗΜΑ 2-1 : Γραμμική παλινδρόμηση σε δίτιμα δεδομένα.

2.4.2 Η λογιστική συνάρτηση, η σιγμοειδής συνάρτηση και η συνάρτηση logit

Η λογιστική συνάρτηση (*logistic function*) αναπτύχθηκε από τον Pierre François Verhulst, Βέλγο μαθηματικό και γιατρό, ως μοντέλο πληθυσμιακής ανάπτυξης μεταξύ 1838-1847, διάστημα κατά το οποίο δημοσίευσε τα ευρήματά του σε τρεις εργασίες (Verhulst 1838, 1845 & 1847). Ο δάσκαλός του, Lambert Adolphe Jacques Quetelet, Βέλγος αστρονόμος και μαθηματικός, παρατήρησε ότι το εκθετικό μοντέλο:

$$W(t) = Ae^{\beta t},$$

όπου:

- t: μία χρονική στιγμή
- W(t): ο πληθυσμός τη χρονική στιγμή t
- A: ο πληθυσμός κάποια αρχική χρονική στιγμή, συνήθως W(0)
- β: ο σταθερός ρυθμός ανάπτυξης

δεν είναι ικανοποιητικό για να περιγράψει την πληθυσμιακή ανάπτυξη, καθώς η συνεχόμενη εκθετική ανάπτυξη θα οδηγήσει σε αδύνατες, μη λογικές τιμές (Cramer, 2002). Για αυτό, ζήτησε τη βοήθεια του Verhulst πάνω σε αυτό το πρόβλημα, ο οποίος προσάρμοσε το εκθετικό μοντέλο ανάπτυξης προσθέτοντας έναν όρο που αφορά την αυξανόμενη αντίσταση στην περαιτέρω ανάπτυξη, καταλήγοντας έτσι και επινοώντας τη λογιστική συνάρτηση. Μέσω της λογιστικής συνάρτησης, κατάφερε η ανάπτυξη στην αρχή να είναι εκθετική, έπειτα ο ρυθμός αύξησης να επιβραδύνεται βαθμιαία και εν τέλει να βαίνει παράλληλα στον άξονα x.

Η λογιστική συνάρτηση είναι μια σιγμοειδής καμπύλη (*S-shaped curve*) με εξίσωση:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}\,, \tag{2.2}$$

όπου:

1) L: η φέρουσα ικανότητα (*carrying capacity)*, το supremum των τιμών της συνάρτησης
2) k: ο λογιστικός ρυθμός ανάπτυξης (*logistic growth rate*), η κλίση της καμπύλης
3) x_0: η τιμή του μέσου σημείου της συνάρτησης

με πεδίο ορισμού το $\mathbb{R}$ και πεδίο τιμών το [0,L]. Η λογιστική συνάρτηση χρησιμοποιείται ευρέως, μεταξύ άλλων, στη Βιολογία, στη Χημεία, στις Γεωεπιστήμες, στα Βιομαθηματικά και στα Νευρωνικά Δίκτυα (*Neural Networks*).

Αν στην (2.2) θέσουμε $L = 1$, $k = 1$ και $x_0 = 0$, καταλήγουμε στη σιγμοειδή συνάρτηση (*sigmoid function*), ή αλλιώς τυπική λογιστική συνάρτηση (*standard logistic function*), με εξίσωση:

$$f(x) = \frac{1}{1 + e^{-x}}$$

ή εναλλακτικά:

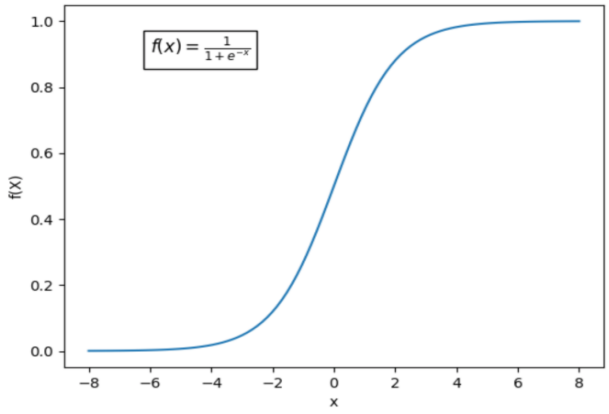
$$f(x) = \frac{e^x}{1 + e^x}\,,$$

πεδίο ορισμού το $\mathbb{R}$ και πεδίο τιμών το [0,1].

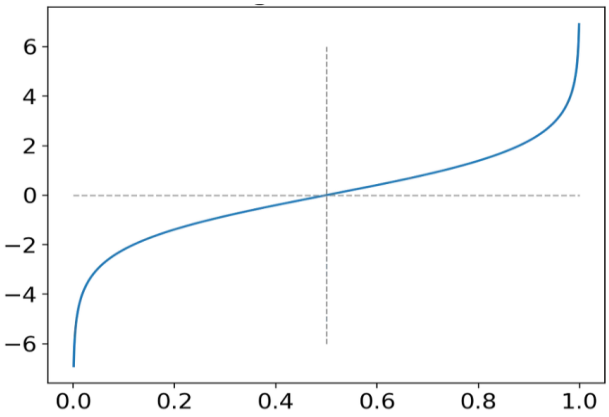
Η συνάρτηση logit (***log****istic un****it***) είναι η αντίστροφη συνάρτηση (*inverse function*) της σιγμοειδούς συνάρτησης με εξίσωση:

$$logit(p) = f^{-1}(p) = log\frac{p}{1-p}\,, 0 < p < 1\,,$$

πεδίο ορισμού το [0,1] και πεδίο τιμών το $\mathbb{R}$. Μέσω της αντιστροφής, ο λογάριθμος της σχετικής πιθανότητας μετατρέπεται σε πιθανότητα. Επίσης, η σιγμοειδής καλείται και συνάρτηση expit, λόγω του ότι είναι η αντίστροφη της logit.



ΣΧΗΜΑ 2-2 : Η σιγμοειδής συνάρτηση.



ΣΧΗΜΑ 2-3 : Η συνάρτηση logit.

2.4.3 Το μοντέλο της λογιστικής παλινδρόμησης

Είδαμε ότι το μοντέλο της κλασσικής γραμμικής παλινδρόμησης δεν είναι ικανό να χρησιμοποιηθεί σε δίτιμα δεδομένα. Οπότε, πρέπει να μετασχηματίσουμε το μοντέλο ώστε όχι μόνο να περιορίσουμε τις εκτιμήσεις των p_i στο διάστημα [0,1], αλλά ταυτόχρονα να κρατήσουμε και μία γραμμική σχέση μεταξύ των p_i και των ανεξάρτητων μεταβλητών $X_1, X_2, \dots, X_d$. Επίσης, χρειαζόμαστε μία συνάρτηση σύνδεσης που να είναι μονότονη και να ισχύει $0 \leq g^{-1}(\eta) \leq 1$ για κάθε η. Κατάλληλες συναρτήσεις σύνδεσης για τον σκοπό αυτό, όπως είδαμε και στον Πίνακα 1.1, είναι η συνάρτηση logit με $\eta_i = logit(p_i) = log\left(\frac{p_i}{1-p_i}\right)$, η συνάρτηση probit με $\eta_i = probit(p_i) = \Phi^{-1}(p_i)$, όπου Φ η αθροιστική συνάρτηση κατανομής της Τυποποιημένης Κανονικής κατανομής, και η συνάρτηση complementary log-log με $\eta_i = log\,[-log(1-p_i)]$. Παρ'όλο που και οι τρεις

συναρτήσεις δίνουν παρόμοια αποτελέσματα, η συνάρτηση logit είναι αυτή που χρησιμοποιείται περισσότερο λόγω της ευκολίας ερμηνείας των αποτελεσμάτων και είναι η μόνη με την οποία θα ασχοληθούμε. Ως μοντέλο λογιστικής παλινδρόμησης καλείται αυτό που χρησιμοποιεί τη συνάρτηση logit ως συνάρτηση σύνδεσης.

Έστω η δίτιμη εξαρτημένη μεταβλητή Y και οι ανεξάρτητες μεταβλητές $X_1, X_2, \dots, X_d$. Ως p ορίζουμε την πιθανότητα η Y να πάρει την τιμή 1, δηλαδή $p = P(Y = 1)$ και $q = 1 - p = P(Y = 0)$. Έχουμε τη γραμμική προβλέπουσα:

$$\eta = \beta_0 + \sum_{j=1}^{d} \beta_j X_j$$

και ως συνάρτηση σύνδεσης την:

$$\eta = g(p) = logit(p) = log\left(\frac{p}{1-p}\right).$$

Έτσι, καταλήγουμε στο μοντέλο της πολλαπλής λογιστικής παλινδρόμησης:

$$logit(p) = log\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{j=1}^{d} \beta_j X_j, \tag{2.3}$$

όπου αν πάρουμε τον αντιλογάριθμο βρίσκουμε:

$$exp\left[log\left(\frac{p}{1-p}\right)\right] = exp\left(\beta_0 + \sum_{j=1}^{d} \beta_j X_j\right) \Leftrightarrow$$

$$\frac{p}{1-p} = exp\left(\beta_0 + \sum_{j=1}^{d} \beta_j X_j\right) \Leftrightarrow$$

$$p = exp\left(\beta_0 + \sum_{j=1}^{d} \beta_j X_j\right) - exp\left(\beta_0 + \sum_{j=1}^{d} \beta_j X_j\right) p \Leftrightarrow$$

$$p + exp\left(\beta_0 + \sum_{j=1}^{d} \beta_j X_j\right) p = exp\left(\beta_0 + \sum_{j=1}^{d} \beta_j X_j\right) \Leftrightarrow$$

$$p\left[1 + exp\left(\beta_0 + \sum_{j=1}^{d} \beta_j X_j\right)\right] = exp\left(\beta_0 + \sum_{j=1}^{d} \beta_j X_j\right).$$

Άρα, η πιθανότητα η Y να πάρει την τιμή 1 μπορεί να γραφτεί και ως:

$$p = \frac{exp\left(\beta_0 + \sum_{j=1}^{d} \beta_j X_j\right)}{1 + exp\left(\beta_0 + \sum_{j=1}^{d} \beta_j X_j\right)}. \tag{2.4}$$

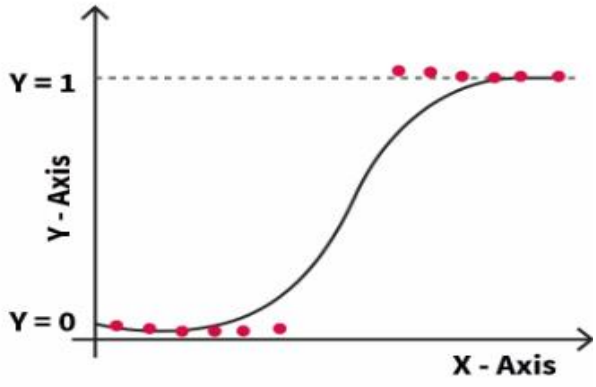
Ένας άλλος τρόπος για να εκφράσουμε την πιθανότητα επιτυχίας είναι:

$$p = \frac{1}{1 + exp\left[-\left(\beta_0 + \sum_{j=1}^{d} \beta_j X_j\right)\right]}.$$

Τέλος, ως πιθανότητα η Y να πάρει την τιμή 0 ορίσαμε την $q = 1 - p$, από όπου προκύπτει:

$$q = 1 - p = 1 - \frac{exp\left(\beta_0 + \sum_{j=1}^{d} \beta_j X_j\right)}{1 + exp\left(\beta_0 + \sum_{j=1}^{d} \beta_j X_j\right)} \Leftrightarrow$$

$$q = \frac{1}{1 + exp\left(\beta_0 + \sum_{j=1}^{d} \beta_j X_j\right)}.$$



ΣΧΗΜΑ 2-4 : Λογιστική παλινδρόμηση σε δίτιμα δεδομένα.

2.4.4 Ερμηνεία των παραμέτρων

Το λογιστικό μοντέλο είναι γραμμικό και αθροιστικό για τα log-odds, όμως είναι πολλαπλασιαστικό όσον αφορά τα odds. Ας θεωρήσουμε το απλό λογιστικό μοντέλο με X_1τη μοναδική ανεξάρτητη μεταβλητή:

$$logit(p) = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1.$$

Αν το X_1 αυξηθεί κατά μία μονάδα, τότε ο λογάριθμος της σχετική πιθανότητας αυξάνεται κατά β_1. Όμως, επειδή ισχύει:

$$\frac{p}{1-p} = exp(\beta_0 + \beta_1 X_1) = e^{\beta_0} e^{\beta_1 X_1},$$

η σχετική πιθανότητα θα αυξηθεί κατά e^{β_1}. Ένας άλλος τρόπος για να το δούμε αυτό είναι ο εξής. Έστω η γραμμική προβλέπουσα με μοναδική ανεξάρτητη μεταβλητή την X_1:

$$\eta(X_1) = \beta_0 + \beta_1 X_1.$$

Αν το X_1 αυξηθεί κατά μία μονάδα, τότε:

$$\eta(X_1 + 1) = \beta_0 + \beta_1(X_1 + 1).$$

Οπότε η διαφορά των δύο είναι:

$$\eta(X_1 + 1) - \eta(X_1) = \beta_0 + \beta_1 X_1 + \beta_1 - \beta_0 - \beta_1 X_1 = \beta_1.$$

Ακόμα, ισχύει:

$$\eta(X_1) = log\ (odds_{X_1})$$

και:

$$\eta(X_1 + 1) = log\big(odds_{X_1+1}\big).$$

Η διαφορά τους είναι:

$$\eta(X_1 + 1) - \eta(X_1) = log\big(odds_{X_1+1}\big) - log\big(odds_{X_1}\big) \Leftrightarrow$$

$$\eta(X_1 + 1) - \eta(X_1) = log\left(\frac{odds_{X_1+1}}{odds_{X_1}}\right) = \beta_1.$$

Παίρνοντας τον αντιλογάριθμο καταλήγουμε στο λόγο σχετικών πιθανοτήτων (*odds ratio*):

$$OR = \frac{odds_{X_1+1}}{odds_{X_1}} = e^{\beta_1}.$$

Άρα βλέπουμε ότι, όπως και πριν, η σχετική πιθανότητα θα αυξηθεί κατά e^{β_1}. Η ερμηνεία του OR έχει ως εξής:

1) Αν $OR = 1$, τότε η ανεξάρτητη μεταβλητή X_1 δε διαδραματίζει κάποιο ρόλο ($\beta_1 = 0$).
2) Αν $OR = \alpha$ και $a > 1$, τότε αύξηση της X_1 κατά μία μονάδα σημαίνει μεγαλύτερη σχετική πιθανότητα η Y να πάρει την τιμή 1 κατά $(\alpha - 1)100\%$.
3) Αν $OR = \alpha$ και $a < 1$, τότε αύξηση της X_1 κατά μία μονάδα σημαίνει μικρότερη σχετική πιθανότητα η Y να πάρει την τιμή 1 κατά $(1 - \alpha)100\%$.

Στην περίπτωση του πολλαπλού λογιστικού μοντέλου με $X_1, X_2, \dots, X_d$ ανεξάρτητες μεταβλητές:

$$logit(p) = log\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{j=1}^{d} \beta_j X_j,$$

αν το X_1 αυξηθεί κατά μία μονάδα, ο λογάριθμος της σχετικής πιθανότητας θα αυξηθεί κατά β_1, με την προϋπόθεση ότι οι υπόλοιπες ανεξάρτητες μεταβλητές παραμένουν σταθερές.

Παρομοίως, η σχετική πιθανότητα θα αυξηθεί κατά e^{β_1}, δοθέντος ότι οι υπόλοιπες ανεξάρτητες μεταβλητές δε μεταβάλλονται. Το OR στην πολλαπλή λογιστική παλινδρόμηση μεταφράζεται ως:

1) Αν $OR = 1$, τότε η ανεξάρτητη μεταβλητή X_1 δε διαδραματίζει κάποιο ρόλο ($\beta_1 = 0$).
2) Αν $OR = \alpha$ και $a > 1$, τότε αύξηση της X_1 κατά μία μονάδα σημαίνει μεγαλύτερη σχετική πιθανότητα η Y να πάρει την τιμή 1 κατά $(\alpha - 1)100\%$, διατηρώντας όλες τις υπόλοιπες ανεξάρτητες μεταβλητές σταθερές.
3) Αν $OR = \alpha$ και $a < 1$, τότε αύξηση της X_1 κατά μία μονάδα σημαίνει μικρότερη σχετική πιθανότητα η Y να πάρει την τιμή 1 κατά $(1 - \alpha)100\%$, διατηρώντας όλες τις υπόλοιπες ανεξάρτητες μεταβλητές σταθερές.

2.4.5 Εκτίμηση των παραμέτρων

Για να εκτιμήσουμε τις παραμέτρους των $X_1, X_2, \dots, X_d$, θα χρησιμοποιήσουμε τη μεθοδολογία της μέγιστης πιθανοφάνειας. Μέσω της εκτίμησης μέγιστης πιθανοφάνειας (*maximum likelihood estimation, MLE*), οι εκτιμητές έχουν κάποιες επιθυμητές ιδιότητες, όπως επάρκεια (*sufficiency*), συνέπεια (*consistency*) και ασυμπτωτική αποτελεσματικότητα (*asymptotic efficiency*).

Όπως είδαμε και στο 1.3.5, οι εξισώσεις πιθανοφάνειας στα ΓΓΜ δίνονται από τη σχέση:

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = 0 \, , \gamma\iota\alpha \; \kappa\acute{\alpha}\theta\varepsilon \; j,$$

όπου $\boldsymbol{\beta}$ το διάνυσμα των παραμέτρων διάστασης (d+1) × 1. Επίσης, αν η $Y \sim Bin(n, p)$, η συνάρτηση μάζας πιθανότητάς της μπορεί να γραφτεί ως:

$$f(y; p) = exp\left[ylog\left(\frac{p}{1-p}\right) + nlog(1-p) + log\binom{n}{y}\right].$$

Για n παρατηρήσεις $\boldsymbol{y} = (y_1, y_2, \dots y_n)'$ από την κατανομή Bernoulli(p_i) και $\boldsymbol{p} = (p_1, p_2, \dots, p_n)'$, η συνάρτηση της πιθανοφάνειας είναι:

$$L(\boldsymbol{p}; \boldsymbol{y}) = exp\left[\sum_{i=1}^{n} y_i \, log\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n} log(1-p_i) + log\binom{1}{y_i}\right],$$

με αντίστοιχο λογάριθμο:

$$l(\boldsymbol{p};\boldsymbol{y}) = \sum_{i=1}^{n} y_i \, log\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n} log(1-p_i). \qquad (2.5)$$

Ο όρος $log\binom{1}{y_i}$ δεν εξαρτάται από τα p_i, οπότε παραλείπεται. Από την (2.3), έχουμε:

$$logit(p_i) = log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j=1}^{d} \beta_j x_{ij}$$

και γράφοντας το $\beta_0 + \sum_{j=1}^{d} \beta_j x_j$ ως $\sum_{j=0}^{d} \beta_j x_{ij}$ με $x_{0j} = 1$, η (2.5) γίνεται:

$$l(\boldsymbol{\beta};\boldsymbol{y}) = \sum_{i=1}^{n}\sum_{j=0}^{d} y_i \beta_j x_{ij} - \sum_{i=1}^{n} log\left(1 + exp\left(\sum_{j=0}^{d} \beta_j x_{ij}\right)\right).$$

Παραγωγίζοντας ως προς β_k:

$$\frac{\partial l}{\partial \beta_k} = \sum_{i=1}^{n} y_i x_{ik} - \sum_{i=1}^{n}\left[\frac{exp\left(\sum_{j=0}^{d} \beta_j x_{ij}\right)}{1 + exp\left(\sum_{j=0}^{d} \beta_j x_{ij}\right)}\right] x_{ik}.$$

Από την (2.4) καταλήγουμε:

$$\frac{\partial l}{\partial \beta_k} = \sum_{i=1}^{n} y_i x_{ik} - \sum_{i=1}^{n} p_i x_{ik} \;\Leftrightarrow$$

$$\frac{\partial l}{\partial \beta_k} = \sum_{i=1}^{n} (y_i - p_i) x_{ik}.$$

Θέλουμε να μεγιστοποιήσουμε το λογάριθμο της συνάρτησης πιθανοφάνειας, άρα θέτουμε:

$$\frac{\partial l}{\partial \beta_k} = 0, k = 0,1,2, \dots, d \,,$$

από όπου προκύπτει ένα σύστημα d+1 μη γραμμικών εξισώσεων ως προς β_k, το οποίο θα λύσουμε χρησιμοποιώντας μία επαναληπτική διαδικασία, τη μέθοδο των επαναληπτικών επανασταθμιζόμενων ελαχίστων τετραγώνων (*iterative reweighted least squares, IRLS*). Με τη χρήση πινάκων έχουμε:

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{p}) = 0,$$

όπου $\boldsymbol{X}^T$ ο ανάστροφος πίνακας σχεδιασμού διαστάσεων (d+1) × n. Το διάνυσμα $\boldsymbol{\beta}$ που προκύπτει βάσει της μεθόδου μέγιστης πιθανοφάνειας ικανοποιεί τη σχέση:

$$\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{X}^T E(\boldsymbol{Y}),$$

που ισοδύναμα γράφεται:

$$\boldsymbol{X^T W X \hat{\beta}} = \boldsymbol{X^T W Z},$$

όπου $\boldsymbol{W}$ n × n διαγώνιος πίνακας με στοιχεία στη διαγώνιο $p_i(1-p_i)$ και $\boldsymbol{Z}$ διάνυσμα μήκους n και στοιχεία $z_i = \hat{\eta}_i + (y_i - \hat{p}_i)\frac{\partial \eta_i}{\partial p_i}$. Η συνάρτηση σύνδεσης τώρα είναι η logit, άρα:

$$z_i = \hat{\eta}_i + (y_i - \hat{p}_i)\frac{1}{p_i(1-p_i)}.$$

Η μέθοδος IRLS ακολουθεί τα εξής βήματα:

1) Χρησιμοποιείται μία αρχική λύση $\widehat{\boldsymbol{\beta}}^{(0)}$.
2) Υπολογίζονται τα $\widehat{\boldsymbol{p}}^{(0)}$ και $\widehat{\boldsymbol{\eta}}^{(0)}$.
3) Από τη σχέση $\widehat{\boldsymbol{\beta}}^{(1)} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Z}$ βρίσκουμε τη λύση για τις παραμέτρους β_j.
4) Από το $\widehat{\boldsymbol{\beta}}^{(1)}$ θα βρεθούν οι λύσεις για το $\widehat{\boldsymbol{\beta}}^{(2)}$ κλπ.

Μετά από μεγάλο αριθμό επαναλήψεων $m \to \infty$, οι εκτιμητές που προκύπτουν συγκλίνουν στους εκτιμητές μέγιστης πιθανοφάνειας. Η μέθοδος σταματά για προκαθορισμένο αριθμό επαναλήψεων s ή όταν $\sum_{j=0}^{d}\left(\hat{\beta}_j^{(m)} - \hat{\beta}_j^{(m-1)}\right)^2 < \varepsilon$, για γνωστό $\varepsilon > 0$. Για $\hat{p}_i = 0$ ή $\hat{p}_i = 1$, είναι αρκετά πιθανό να μη συγκλίνει ο αλγόριθμος.

2.4.6 Έλεγχος σημαντικότητας των παραμέτρων

Αφού εκτιμήθηκαν οι παράμετροι, θα πρέπει να ελέγξουμε τη σημαντικότητά τους, αν θα πρέπει να συμπεριληφθεί στο λογιστικό μοντέλο η αντίστοιχη ανεξάρτητη μεταβλητή ή όχι. Το μοντέλο που εμπεριέχει την μεταβλητή X_j, μας δίνει κάποια πληροφορία παραπάνω για την Y από ένα άλλο μοντέλο που δεν την εμπεριέχει;

Συγκεκριμένα, έστω το λογιστικό μοντέλο:

$$logit(\hat{p}) = log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

και θέλουμε να ελέγξουμε σε προκαθορισμένο επίπεδο σημαντικότητας α την υπόθεση:

$$H_0 : \beta_j = 0$$

και

$$H_1 : \beta_j \neq 0.$$

Υπολογίζοντας το στατιστικό του Wald:

$$Z = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)},$$

όπου ασυμπτωτικά ισχύει ότι $Z \sim N(0,1)$, μπορούμε να ελέγξουμε την H_0. Αν $|Z| < z_{1-a/2}$, δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση και άρα η μεταβλητή X_j δεν πρέπει να συμπεριληφθεί στο μοντέλο, δοθέντος όμως ότι οι υπόλοιπες μεταβλητές εμπεριέχονται στο μοντέλο. Π.χ, έστω ότι στο παραπάνω μοντέλο θέλουμε να ελέγξουμε τη στατιστική σημαντικότητα της X_3 και μέσω του ελέγχου Wald απορρίπτουμε την H_0. Άρα, θα λέγαμε ότι η X_3 δεν είναι στατιστική σημαντική, δοθέντος όμως ότι στο μοντέλο υπάρχουν οι X_1 και X_2.

Μία άλλη υπόθεση που μπορούμε να ελέγξουμε είναι αν η β_j είναι ίση με κάποια τιμή διαφορετική του 0, έστω $\beta_j^{(0)}$. Άρα $H_0 : \beta_j = \beta_j^{(0)}$. Το στατιστικό του Wald τώρα είναι:

$$Z = \frac{\hat{\beta}_j - \beta_j^{(0)}}{se(\hat{\beta}_j)}.$$

Πάλι, αν $|Z| > z_{1-a/2}$, απορρίπτουμε τη μηδενική υπόθεση. Οι έλεγχοι αυτοί είναι ανάλογοι με τα t-tests που διενεργούμε στο κλασσικό γραμμικό μοντέλο παλινδρόμησης.

2.4.7 Διαστήματα εμπιστοσύνης

Τα διαστήματα εμπιστοσύνης που θα κατασκευάσουμε, βασίζονται στον έλεγχο του Wald (*Wald-based confidence intervals*). Ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης, για προκαθορισμένο α, για την παράμετρο β_j είναι το:

$$\hat{\beta}_j \pm z_{1-a/2}\left[se\left(\hat{\beta}_j\right)\right].$$

Επίσης, μπορούμε να κατασκευάσουμε και ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για το λογιστικό μοντέλο (Hosmer et al, 2013). Ο εκτιμητής του logit μοντέλου δίνεται ως:

$$\hat{g}(x_1) = \hat{\beta}_0 + \hat{\beta}_1 x_1.$$

Για να βρούμε ένα δ.ε για το logit μοντέλο, πρέπει να εκτιμήσουμε την τυπική απόκλιση του εκτιμητή του logit μοντέλου. Για αρχή, θα εκτιμήσουμε τη διακύμανση ως:

$$\hat{V}[\hat{g}(x_1)] = \hat{V}\left(\hat{\beta}_0\right) + x_1^2 \hat{V}\left(\hat{\beta}_1\right) + 2x_1 \widehat{Cov}(\hat{\beta}_0, \hat{\beta}_1)$$

και από την οποία θα βρούμε την εκτίμηση της τυπικής απόκλισης:

$$\widehat{se}[\hat{g}(x_1)] = \sqrt{\hat{V}[\hat{g}(x_1)]}.$$

Τότε, ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για το logit μοντέλο είναι το:

$$\hat{g}(x_1) \pm z_{1-a/2}\widehat{se}[\hat{g}(x_1)].$$

Επομένως, ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για την προβλεπόμενη τιμή (*predicted or fitted value*) $\hat{p}$ είναι το:

$$\frac{e^{\hat{g}(x_1)\pm z_{1-a/2}\widehat{se}[\hat{g}(x_1)]}}{1+e^{\hat{g}(x_1)\pm z_{1-a/2}\widehat{se}[\hat{g}(x_1)]}}.$$

Στην περίπτωση του πολλαπλού λογιστικού μοντέλου με d ανεξάρτητες μεταβλητές, ο εκτιμητής του logit μοντέλου είναι:

$$\hat{g}(\boldsymbol{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_d x_d,$$

η εκτιμώμενη διακύμανση του εκτιμητή υπολογίζεται ως:

$$\hat{V}[\hat{g}(\boldsymbol{x})] = \sum_{j=0}^{d} x_j^2 \hat{V}(\hat{\beta}_j) + \sum_{j=0}^{d}\sum_{k=j+1}^{d} 2x_j x_k \widehat{Cov}(\hat{\beta}_j, \hat{\beta}_k),$$

και η τυπική απόκλιση:

$$\widehat{se}[\hat{g}(\boldsymbol{x})] = \sqrt{\hat{V}[\hat{g}(\boldsymbol{x})]}.$$

Άρα, ένα $100(1-\alpha)\%$ δ.ε. για το πολλαπλό μοντέλο είναι το:

$$\hat{g}(\boldsymbol{x}) \pm z_{1-a/2}\widehat{se}[\hat{g}(\boldsymbol{x})]$$

και μέσω του οποίου καταλήγουμε σε ένα $100(1-\alpha)\%$ δ.ε. για την προβλεπόμενη τιμή:

$$\frac{e^{\hat{g}(\boldsymbol{x})\pm z_{1-a/2}\widehat{se}[\hat{g}(\boldsymbol{x})]}}{1+e^{\hat{g}(\boldsymbol{x})\pm z_{1-a/2}\widehat{se}[\hat{g}(\boldsymbol{x})]}}.$$

2.4.8 Έλεγχος καλής προσαρμογής

Ένας έλεγχος καλής προσαρμογής (*goodness of fit test*) διενεργείται για να αξιολογήσει το πόσο καλά προσαρμόζεται το μοντέλο που καταλήξαμε στα δεδομένα. Αν δηλαδή οι προβλεπόμενες πιθανότητες $\hat{p}_i$ που θα προκύψουν από το εκτιμώμενο μοντέλο, αντικατοπτρίζουν το πραγματικό αποτέλεσμα στα δεδομένα. Η υπόθεση σε τέτοιου είδους ελέγχους είναι η:

$$H_0 : \textit{Το μοντέλο προσαρμόζεται καλά στα δεδομένα}$$

και

$$H_1 : \text{όχι η } H_0$$

ή διατυπώνοντάς το διαφορετικά:

H_0 : *οι παρατηρηθείσες τιμές της Y δε διαφέρουν σημαντικά από τις εκτιμώμενες τιμές*

και

$$H_1 : \text{όχι η } H_0.$$

Ένας από τους πιο γνωστούς ελέγχους είναι ο έλεγχος καλής προσαρμογής του Pearson, με την ελεγχοσυνάρτηση να δίνεται ως:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

όπου k το πλήθος των κατηγοριών, O_i οι παρατηρούμενες (*observed*) τιμές στην κατηγορία i και E_i οι προβλεπόμενες (*expected*) τιμές στην κατηγορία i. Ένας άλλος τρόπος γραφής της παραπάνω σχέσης είναι ο:

$$X^2 = \sum_{i=1}^{k} \frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)}, \tag{2.6}$$

με y_i οι παρατηρούμενες τιμές στην κατηγορία i, n_i το συνολικό πλήθος των παρατηρήσεων στην κατηγορία i, $\hat{p}_i$ η αντίστοιχη εκτιμώμενη πιθανότητα επιτυχίας και $1-\hat{p}_i$ η πιθανότητα αποτυχίας. Για να μπορέσουμε όμως να χρησιμοποιήσουμε τον έλεγχο αυτό, θα πρέπει να ισχύει $n_i\hat{p}_i \geq 5$ για κάθε i. Στην περίπτωση της λογιστικής παλινδρόμησης για δίτιμα δεδομένα, γνωρίζουμε ότι $n_i = 1$ και προφανώς $0 \leq \hat{p}_i \leq 1$. Επομένως, θα πρέπει να χρησιμοποιήσουμε κάποιον άλλον έλεγχο που να είναι κατάλληλος για δίτιμα δεδομένα.

Οι Hosmer and Lemeshow (1980) και Lemeshow and Hosmer (1982) πρότειναν να γίνεται ομαδοποίηση των δεδομένων σύμφωνα με τις εκτιμώμενες πιθανότητες. Ο έλεγχος των Hosmer-Lemeshow βασίζεται σε αυτόν του Pearson για παρατηρούμενες και προβλεπόμενες τιμές. Για να εφαρμοστεί ο έλεγχος, εφαρμόζονται τα εξής βήματα:

1) Οι παρατηρήσεις διατάσσονται σε αύξουσα σειρά σύμφωνα με τις εκτιμώμενες πιθανότητες επιτυχίας τους.
2) Έπειτα, χωρίζονται σε g ομάδες. Συνήθως $g = 10$, εκτός αν έτσι προκύψουν πολύ λίγες παρατηρήσεις σε κάθε ομάδα. Ο διαχωρισμός μπορεί να γίνει με δύο τρόπους: α) η πρώτη ομάδα να περιέχει τις πρώτες n/g παρατηρήσεις, η δεύτερη τις επόμενες n/g παρατηρήσεις κ.ο.κ. β) να βρούμε τα k/g σημεία αποκοπής, όπου $k = 1,2,\dots,g-1$. Όσες παρατηρήσεις έχουν εκτιμώμενη πιθανότητα επιτυχίας μικρότερη ή ίση από το πρώτο σημείο αποκοπής, θα μπουν στην πρώτη ομάδα, όσες έχουν εκτιμώμενη

πιθανότητα επιτυχίας ανάμεσα στο πρώτο και στο δεύτερο σημείο αποκοπής, θα μπουν στη δεύτερη κ.ο.κ.

3) Καταγράφονται οι αριθμοί των επιτυχιών και των αποτυχιών σε κάθε ομάδα.
4) Δημιουργείται ένας g × 2 πίνακας συνάφειας στον οποίο θα εφαρμοστεί ο έλεγχος του Pearson.

Η στατιστική συνάρτηση είναι:

$$X_{HL}^2 = \sum_{i=1}^{g}\left[\frac{(O_{1i}-E_{1i})^2}{E_{1i}} + \frac{(O_{0i}-E_{0i})^2}{E_{0i}}\right] \Leftrightarrow$$

$$X_{HL}^2 = \sum_{i=1}^{g}\left[\frac{(O_{1i}-E_{1i})^2}{n_i\bar{p}_i} + \frac{\left(n_i - O_{1i} - (n_i - E_{1i})\right)^2}{n_i(1-\bar{p}_i)}\right] \Leftrightarrow$$

$$X_{HL}^2 = \sum_{i=1}^{g}\frac{(O_{1i}-E_{1i})^2}{n_i\bar{p}_i(1-\bar{p}_i)},$$

όπου n_i το συνολικό πλήθος των παρατηρήσεων στην ομάδα i, O_{1i} το πλήθος των επιτυχιών ($Y=1$) στην ομάδα i, O_{0i} το αντίστοιχο πλήθος των αποτυχιών, E_{1i} οι προβλεπόμενες τιμές για $Y=1$, E_{0i} οι προβλεπόμενες τιμές για $Y=0$ και $\bar{p}_i$ η μέση εκτιμώμενη πιθανότητα επιτυχίας. Ασυμπτωτικά ισχύει ότι $X_{HL}^2 \sim \chi_{g-2}^2$. Μεγάλες τιμές της X_{HL}^2 υποδηλώνουν ότι το μοντέλο δεν είναι επαρκές, δεν προσαρμόζεται καλά στα δεδομένα.

Όμως, πρέπει να λάβουμε υπ'όψιν ότι η τιμή της στατιστικής συνάρτησης εξαρτάται από τον αριθμό των ομάδων, το πλήθος των παρατηρήσεων εντός των ομάδων και από το πως παρατηρήσεις με κοινή εκτιμώμενη πιθανότητα θα χωριστούν μεταξύ των ομάδων. Επομένως, θα μπορούσαμε να χρησιμοποιήσουμε τον έλεγχο ως «οδηγό» για την επάρκεια του μοντέλου, αλλά να καταλήγαμε σε κάποιο συμπέρασμα σε συνδυασμό με κάποιον άλλον έλεγχο επάρκειας (Collett, 2002).

2.4.9 Απόκλιση

Στο 1ο κεφάλαιο (βλ. 1.3.6) ορίσαμε την έννοια της απόκλισης γενικά στα ΓΓΜ, η οποία δίνεται από τη σχέση:

$$D = -2(l_M - l_S),$$

όπου $l_M = log(L_M)$ και $l_S = log(L_S)$. Στην ενότητα αυτή θα ασχοληθούμε με την απόκλιση στη λογιστική παλινδρόμηση. Για n δίτιμες παρατηρήσεις η πιθανοφάνεια ορίζεται ως, συναρτήσει των β,:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

και ο μεγιστοποιημένος λογάριθμός της υπό το εξέταση λογιστικό μοντέλο M:

$$log(\hat{L}_M) = \sum_{i=1}^{n} [y_i \, log \, \hat{p}_i + (1-y_i) log \, (1-\hat{p}_i)].$$

Επίσης ισχύει ότι $l_S = 0$, κάτι που είναι εύκολο να δούμε. Στο κορεσμένο μοντέλο ισχύει ότι $\hat{p}_i = y_i$. Επομένως, ο λογάριθμος της πιθανοφάνειας γράφεται ως:

$$l_S = \sum_{i=1}^{n} [y_i \, log \, y_i + (1-y_i) log \, (1-y_i)].$$

Άρα, είτε το y_i πάρει την τιμή 0 ή πάρει 1, και οι δύο όροι θα είναι 0. Οπότε, η απόκλιση στη λογιστική παλινδρόμηση δίνεται μέσω του τύπου:

$$D = -2 \, log(\hat{L}_M) = -2 \sum_{i=1}^{n} [y_i \, log \, \hat{p}_i + (1-y_i) \, log(1-\hat{p}_i)] \Leftrightarrow$$

$$D = -2 \sum_{i=1}^{n} \left[y_i \, log \left(\frac{\hat{p}_i}{1-\hat{p}_i} \right) + log(1-\hat{p}_i) \right]. \tag{2.7}$$

Τώρα, αν παραγωγίσουμε το $log[L(\boldsymbol{\beta})]$ ως προς β_j χρησιμοποιώντας τον κανόνα της αλυσίδας, έχουμε:

$$\frac{\partial log[L(\boldsymbol{\beta})]}{\partial \beta_j} = \frac{\partial log[L(\boldsymbol{\beta})]}{\partial p_i} \frac{\partial p_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \tag{2.8}$$

με:

$$log[L(\boldsymbol{\beta})] = \sum_{i=1}^{n} [y_i \, log \, p_i + (1-y_i) log \, (1-p_i)].$$

Άρα:

$$\frac{\partial log[L(\boldsymbol{\beta})]}{\partial p_i} = \sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \frac{1-y_i}{1-p_i} \right). \tag{2.9}$$

Επίσης, για τη συνάρτηση σύνδεσης στη λογιστική παλινδρόμηση γνωρίζουμε ότι:

$$\eta_i = logit(p_i) = log \left(\frac{p_i}{1-p_i} \right) = log p_i - log(1-p_i).$$

Επομένως έχουμε:

$$\frac{\partial \eta_i}{\partial p_i} = \frac{1}{p_i} + \frac{1}{1-p_i},$$

από όπου προκύπτει:

$$\frac{\partial p_i}{\partial \eta_i} = p_i(1-p_i). \tag{2.10}$$

Τέλος:

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij} \tag{2.11}$$

αφού:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{id}.$$

Μέσω των (2.9), (2.10) και (2.11), η (2.8) γράφεται ως:

$$\frac{\partial log[L(\boldsymbol{\beta})]}{\partial \beta_j} = \sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \frac{1-y_i}{1-p_i}\right) p_i(1-p_i)x_{ij}$$

και:

$$\sum_{j=1}^{d} \beta_j \frac{\partial\, log[L(\boldsymbol{\beta})]}{\partial \beta_j} = \sum_{i=1}^{n} (y_i - p_i) \sum_{j=1}^{d} \beta_j x_{ij}$$

$$= \sum_{i=1}^{n} (y_i - p_i) log\left(\frac{p_i}{1-p_i}\right).$$

Γνωρίζουμε ότι τα $\widehat{\boldsymbol{\beta}}$ είναι οι εκτιμητές μέγιστης πιθανοφάνειας των $\boldsymbol{\beta}$. Άρα ισχύει ότι:

$$\frac{\partial\, log\left[L\left(\widehat{\boldsymbol{\beta}}\right)\right]}{\partial \beta_k} = 0.$$

Επομένως, τα $\hat{p}_i$ θα πρέπει να ικανοποιούν τη σχέση:

$$\sum_{i=1}^{n} (y_i - \hat{p}_i) log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = 0 \;\Leftrightarrow$$

$$\sum_{i=1}^{n} \left[y_i log\,(\frac{\hat{p}_i}{1-\hat{p}_i})\right] = \sum_{i=1}^{n} \left[\hat{p}_i log\,(\frac{\hat{p}_i}{1-\hat{p}_i})\right].$$

Αντικαθιστώντας στην (2.7):

$$D = -2 \sum_{i=1}^{n} \left[\hat{p}_i\, log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) + log(1-\hat{p}_i)\right].$$

Έτσι, βλέπουμε ότι η απόκλιση στη λογιστική παλινδρόμηση δεν εξαρτάται από τα y_i και άρα δε μας δίνει κάποια πληροφορία μεταξύ των παρατηρήσεων y_i και των εκτιμώμενων πιθανοτήτων $\hat{p}_i$. Επίσης, η κατανομή της D δεν είναι γνωστή για δίτιμα δεδομένα. Άρα, σε αυτήν την περίπτωση η απόκλιση δε μπορεί να χρησιμοποιηθεί ως έλεγχος καλής προσαρμογής. Μέσω της μεταβολής της απόκλισης όμως, μπορούμε να συγκρίνουμε δύο εμφωλευμένα μοντέλα. Έστω τα μοντέλα M_1 και M_2:

$$M_1: \quad logit(p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{d_1} X_{d_1}$$

$$M_2: \quad logit(p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{d_1} X_{d_1} + \beta_{d_1+1} X_{d_1+1} + \cdots + \beta_{d_2} X_{d_2}$$

με λογάριθμο πιθανοφάνειας l_{M_1} και l_{M_2} αντίστοιχα. Το M_1 είναι εμφωλευμένο στο M_2. Ορίζουμε τη μηδενική υπόθεση:

$$H_0: \beta_{d_1+1} = \beta_{d_1+2} = \ldots = \beta_{d_2} = 0$$

με:

$$H_1: \textit{τουλάχιστον μία παράμετρος από τις } \beta_{d_1+1}, \beta_{d_1+2}, \ldots, \beta_{d_2} \textit{ δεν είναι 0}$$

ή:

$$H_0: \textit{Ισχύει το } M_1$$

και

$$H_1: \textit{Ισχύει το } M_2.$$

Η μεταβολή της απόκλισης μεταξύ των δύο μοντέλων υπολογίζεται ως:

$$D_1 - D_2 = \left(-2l_{M_1}\right) - \left(-2l_{M_2}\right) = -2\left(l_{M_1} - l_{M_2}\right) \sim \chi^2_{d_2-d_1}.$$

Αν $D_1 - D_2 > \chi^2_{d_2-d_1}$, τότε απορρίπτουμε την H_0 και άρα επιλέγουμε το M_2 έναντι του M_1.

2.4.10 Κατάλοιπα

Είδαμε ότι μέσω των ελέγχων καλής προσαρμογής ελέγχουμε το πόσο καλά προσαρμόζεται ένα μοντέλο στα δεδομένα. Πέρα από αυτό, δε λαμβάνουμε κάποια άλλη χρήσιμη πληροφορία. Το μοντέλο μας όμως μπορεί να μην είναι κατάλληλο για διάφορους λόγους, όπως:

1) Η γραμμική προβλέπουσα να μην έχει οριστεί σωστά. Να υπάρχουν δηλαδή κάποιες ανεξάρτητες μεταβλητές που έχουν μείνει εκτός, ενώ θα έπρεπε να εισαχθούν στο μοντέλο ή να έπρεπε να μπουν με κάποιο μετασχηματισμό.
2) Η συνάρτηση σύνδεσης να μην είναι η κατάλληλη.

3) Τα δεδομένα να περιλαμβάνουν ακραίες τιμές (*outliers*), παρατηρήσεις δηλαδή που έχουν μεγάλα κατάλοιπα κατά απόλυτη τιμή σε σχέση με τις υπόλοιπες παρατηρήσεις, και άρα να μην περιγράφονται ικανοποιητικά από το μοντέλο.
4) Να υπάρχουν παρατηρήσεις που αν τις αφαιρέσουμε από τα δεδομένα μας, οι εκτιμήσεις των παραμέτρων θα αλλάζουν σημαντικά. Αυτές οι παρατηρήσεις ονομάζονται παρατηρήσεις επιρροής (*influential observations*).

Το να εξετάσουμε το βαθμό που το μοντέλο μας περιγράφει ικανοποιητικά τα δεδομένα είναι το πλέον σημαντικό πριν προχωρήσουμε σε συμπερασματολογία. Για το λόγο αυτό, διεξάγουμε διαγνωστικούς ελέγχους (*diagnostic tests*). Ένας από αυτούς, είναι ο έλεγχος μέσω των καταλοίπων.

Κατάλοιπα (*residuals*) ονομάζονται τα μέτρα συμφωνίας ανάμεσα στην ανεξάρτητη μεταβλητή και στην προβλεπόμενη τιμή. Τα πιο γνωστά κατάλοιπα είναι οι διαφορές μεταξύ των παρατηρούμενων τιμών και των προβλεπόμενων τιμών που προκύπτουν από το προσαρμοσμένο μοντέλο για κάθε παρατήρηση. Για δίτιμα δεδομένα αυτό εκφράζεται ως:

$$e_i = y_i - \hat{p}_i.$$

Τα κατάλοιπα αυτά αποκαλούνται κατάλοιπα απόκρισης (*response residuals*).

Ένα άλλο είδος καταλοίπων που χρησιμοποιούνται είναι αυτά του Pearson (*Pearson's residuals*):

$$e_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$$

ή:

$$e_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{V}(y_i)}}.$$

Για τα κατάλοιπα του Pearson ασυμπτωτικά ισχύει ότι $E(e_i) \cong 0$ και $V(e_i) \cong 1$. Επίσης, από τη σχέση (2.6) είναι εύκολο να δούμε ότι:

$$\sum_{i=1}^{n} e_i^2 = X^2,$$

δηλαδή το άθροισμα των τετραγώνων των καταλοίπων του Pearson μας δίνουν τον έλεγχο καλής προσαρμογής του Pearson (υπενθυμίζουμε ότι για δίτιμα δεδομένα ισχύει $n_i = 1$).

Τα κατάλοιπα απόκλισης (*deviance residuals*) ορίζονται ως:

$$d_i = sign(y_i - \hat{p}_i)\{-2[y_i \, log \, \hat{p}_i + (1-y_i) log(1-\hat{p}_i)]\}^{1/2},$$

όπου το $sign(y_i - \hat{p}_i)$ εξασφαλίζει ότι τα κατάλοιπα απόκλισης και τα κατάλοιπα του Pearson θα έχουν το ίδιο πρόσημο. Ισχύει:

$$\sum_{i=1}^{n} d_i^2 = D,$$

δηλαδή το άθροισμα των τετραγώνων των καταλοίπων απόκλισης μας δίνουν την απόκλιση (σχέση (2.7)). Αν $y_i = 1$, τότε τα κατάλοιπα θα είναι θετικά, αλλιώς αν $y_i = 0$, αρνητικά. Τα προαναφερθέντα κατάλοιπα δεν ακολουθούν την Κανονική κατανομή.

Ένας έλεγχος που μπορούμε να κάνουμε είναι να κατασκευάσουμε τα διαγράμματα των καταλοίπων απόκλισης, ή αυτά του Pearson, ξεχωριστά για τις κάθε συνεχείς ανεξάρτητες μεταβλητές του μοντέλου. Έτσι, μπορούμε να ελέγξουμε την υπόθεση της γραμμικότητας.

2.4.11 Άλλα μέτρα αξιολόγησης του μοντέλου

Ένα από τα πιο κλασσικά και ευρέως χρησιμοποιούμενα μέτρα αξιολόγησης στη γραμμική παλινδρόμηση, είναι ο συντελεστής προσδιορισμού R^2 (*coefficient of determination* R^2). Το R^2 εκφράζει το ποσοστό μεταβλητότητας της εξαρτημένης μεταβλητής Y που εξηγείται από τις ανεξάρτητες μεταβλητές $X_1, X_2, \dots, X_d$ του μοντέλου. Αποτελείται από τρεις ποσότητες:

1) Από το ερμηνευόμενο άθροισμα τετραγώνων SSR (*regression sum of squares*), δηλαδή το κομμάτι της μεταβλητότητας της Y που ερμηνεύεται από το μοντέλο.

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

2) Από το άθροισμα τετραγώνων των σφαλμάτων SSE (*error sum of squares*), ή και RSS, το κομμάτι της μεταβλητότητας της Y που είναι ανερμήνευτο.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

3) Από το συνολικό άθροισμα τετραγώνων SSTO (*total sum of squares*), τη συνολική μεταβλητότητα των παρατηρήσεων. Ισχύει:

$$SSTO = SSR + SSE = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Το R^2 υπολογίζεται ως:

$$R^2 = \frac{SSR}{SSTO} = \frac{SSTO - SSE}{SSTO} = 1 - \frac{SSE}{SSTO}$$

και ισχύει $0 \leq R^2 \leq 1$. Όσο μεγαλύτερη η τιμή του συντελεστή, τόσο περισσότερη και η ερμηνεύσιμη μεταβλητότητα της Y.

Επίσης, χρησιμοποιείται και ο προσαρμοσμένος συντελεστής συσχέτισης R^2 (*adjusted coefficient of determination* R^2), όπου προκύπτει από την παραπάνω σχέση, διαιρώντας με τους αντίστοιχους βαθμούς ελευθερίας:

$$R^2_{adj} = 1 - \frac{\frac{SSE}{n-d}}{\frac{SSTO}{n-1}}.$$

Εν αντιθέσει με το R^2, το R^2_{adj} μπορεί να πάρει και αρνητικές τιμές. Επίσης, ισχύει ότι $R^2_{adj} \leq R^2$.

Στη λογιστική παλινδρόμηση δεν υπάρχει κάποιο αντίστοιχο καθιερωμένο μέτρο, όμως υπάρχουν ανάλογα μέτρα που έχουν προταθεί. Αυτά τα μέτρα είναι γνωστά ως ψευδο-R^2 (*pseudo-*R^2). Ένα από αυτά είναι το R^2 του McFadden (McFadden, 1974), γνωστό και ως δείκτης του λόγου πιθανοφανειών (*likelihood ratio index, LR index*). Ορίζεται ως:

$$R^2_{McF} = 1 - \frac{log(L_M)}{log(L_{null})},$$

όπου L_M η μέγιστη πιθανοφάνεια του προσαρμοσμένου μοντέλου M και L_{null} η μέγιστη πιθανοφάνεια του μηδενικού μοντέλου (*null model*), του μοντέλου δηλαδή που περιέχει μόνο τον σταθερό όρο $\hat{\beta}_0$. Ισχύει $0 \leq R^2_{McF} \leq 1$. Το μηδενικό μοντέλο έχει $R^2_{McF} = 0$ και το κορεσμένο $R^2_{McF} = 1$. Επίσης, αν $0{,}2 \leq R^2_{McF} \leq 0{,}4$, τότε το μοντέλο προσαρμόζεται καλά στα δεδομένα.

Όσο προστίθενται ανεξάρτητες μεταβλητές στο μοντέλο, το $log(L_M)$ μειώνεται και άρα το R^2_{McF} αυξάνεται. Για αυτό, έχει προταθεί μία προσαρμοσμένη εκδοχή του R^2 του McFadden (*adjusted McFadden's* R^2), η οποία θέτει μία ποινή για το πλήθος των παραμέτρων k:

$$R^2_{McF} = 1 - \frac{log(L_M) - k}{log(L_{null})}.$$

Ένα τρίτο ψευδο-R^2 είναι αυτό που προτάθηκε από τον Maddala (1983) και από τους Cox and Snell (1989), το R^2 μέγιστης πιθανοφάνειας (*maximum likelihood* R^2). Πιο γνωστό όμως είναι ως R^2 των Cox-Snell. Αυτό ορίζεται ως:

$$R^2_{CS} = 1 - \left(\frac{L_{null}}{L_M}\right)^{\frac{2}{n}}.$$

Το R^2_{CS} έχει μέγιστη τιμή $1 - (L_{null})^{2/n}$, οπότε ως εναλλακτικό R^2 προτάθηκε από τους Cragg and Uhler (1970) και Nagelkerke (1991):

$$R^2_{C\&U} = \frac{1 - \left(\frac{L_{null}}{L_M}\right)^{2/n}}{1 - (L_{null})^{2/n}}.$$

Τέλος, υπάρχει και το R^2 του Efron (Efron, 1978):

$$R^2_{Efron} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{p}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Κ Ε Φ Α Λ Α Ι Ο 3

Συναρτήσεις Ποινής στη Λογιστική Παλινδρόμηση

3.1 Επιλογή Μοντέλου

3.1.1 Εισαγωγή

Στη Στατιστική, όταν συλλέξουμε και μελετήσουμε τα δεδομένα μας, θέλουμε να καταλήξουμε σε ένα αποτελεσματικό προβλεπτικό μοντέλο. Αρχικά, καλούμαστε να αποφασίσουμε τι είδους μοντέλο θα είναι κατάλληλο για να περιγράψει τις παρατηρήσεις· γραμμικό, λογιστικό, λογαριθμογραμμικό κλπ. Έπειτα, θα προσαρμόσουμε το μοντέλο αυτό στα δεδομένα και από το οποίο θα προκύψουν, αναλόγως βέβαια και το πλήθος d των ανεξάρτητων μεταβλητών, πολλά και διάφορα υποψήφια μοντέλα. Ένα μοντέλο μπορεί να έχει μόνο μία ανεξάρτητη μεταβλητή, άλλο d, κάποιο άλλο να περιέχει όρους αλληλεπίδρασης κλπ. Μέσα από αυτά, θα πρέπει να κρατήσουμε ένα για την έρευνά μας. Ποιο όμως; Υπάρχει κάποιο μοντέλο που να θεωρείται πιο «σωστό» έναντι των υπολοίπων;

3.1.2 Επιθυμητές ιδιότητες ενός μοντέλου

Πολλοί θα πίστευαν ότι όταν σε ένα μοντέλο προσθέτουμε μεταβλητές, τόσο οι δείκτες καλής προσαρμογής βελτιώνονται και άρα τόσο το καλύτερο. Ένα μοντέλο με πολλές μεταβλητές, είναι εύλογο να εικαστεί κάποιος, ότι είναι πιο ακριβές από ένα άλλο με λίγες. Αλλά, όσο προσθέτουμε μεταβλητές σε ένα μοντέλο, τόσο αυξάνουμε την πολυπλοκότητά του. Στόχος μας όμως είναι να δημιουργήσουμε ένα φειδωλό μοντέλο (*parsimonious model*), καθώς η φειδώ στις παραμέτρους είναι ένα επιθυμητό

χαρακτηριστικό ενός μοντέλου (McCullagh and Nelder, 1989). Το μοντέλο θέλουμε να περιέχει τόσες παραμέτρους ώστε να είναι στη βάση του ορθό, να είναι οικονομικό, να είναι εύκολα ερμηνεύσιμο και να περιγράφει την ουσία των δεδομένων, και όχι να περιέχει περιττές μεταβλητές με την πλάνη ότι το μοντέλο θα είναι πιο αξιόπιστο.

Επίσης, όταν κατασκευάζουμε ένα μοντέλο, ναι μεν θέλουμε κάποιο που να περιγράφει αρκετά καλά τα δεδομένα μας, αλλά επίσης να μπορεί να περιγράψει και μελλοντικά δεδομένα, νέα δεδομένα που θα συλλέξουμε στην πορεία. Επομένως, μία άλλη χρήσιμη ιδιότητα ενός καλού μοντέλου είναι αυτή της ικανότητας γενίκευσης (*generalization capability*).

3.1.3 Μέθοδοι επιλογής μοντέλου

Το πρόβλημα της επιλογής μοντέλου (*model selection*) και κατ'επέκταση το ζητούμενο της επιλογής μεταβλητών (*variable selection or feature selection*), δεν είναι μια διαδικασία τετριμμένη. Κάποιες από τις πιο γνωστές μεθόδους επιλογής μοντέλου είναι η επιλογή καλύτερου υποσυνόλου (*best subset selection*), η προς-τα-εμπρός επιλογή (*forward selection or stepwise addition*), η προς-τα-πίσω απαλοιφή (*backward elimination or backward selection*) και η βηματική παλινδρόμηση (*stepwise regression*). Τα μοντέλα που θα προκύψουν από τις μεθόδους αυτές όμως, πρέπει να θυμόμαστε ότι δεν είναι πάντα τα καταλληλότερα.

Η best subset selection προσαρμόζει όλα τα πιθανά μοντέλα που μπορούν να προκύψουν από τις d ανεξάρτητες μεταβλητές και για αυτό είναι αρκετά απαιτητική υπολογιστικά, ακόμα και για τα σύγχρονα υπολογιστικά συστήματα. Αρκεί μόνο να σκεφτούμε ότι για $d = 20$, υπάρχουν πάνω από ένα εκατομμύριο πιθανοί συνδυασμοί υποσυνόλων. Στην αρχή, προσαρμόζει όλα τα d μοντέλα με μία ανεξάρτητη μεταβλητή, μετά τα $\binom{d}{2} = d(d-1)/2$ μοντέλα με δύο ανεξάρτητες μεταβλητές κ.ο.κ. Το πρόβλημα επιλογής του «καλύτερου» υποσυνόλου ανάμεσα από τα 2^d μοντέλα περιγράφεται στον παρακάτω αλγόριθμο (James et al., 2023):

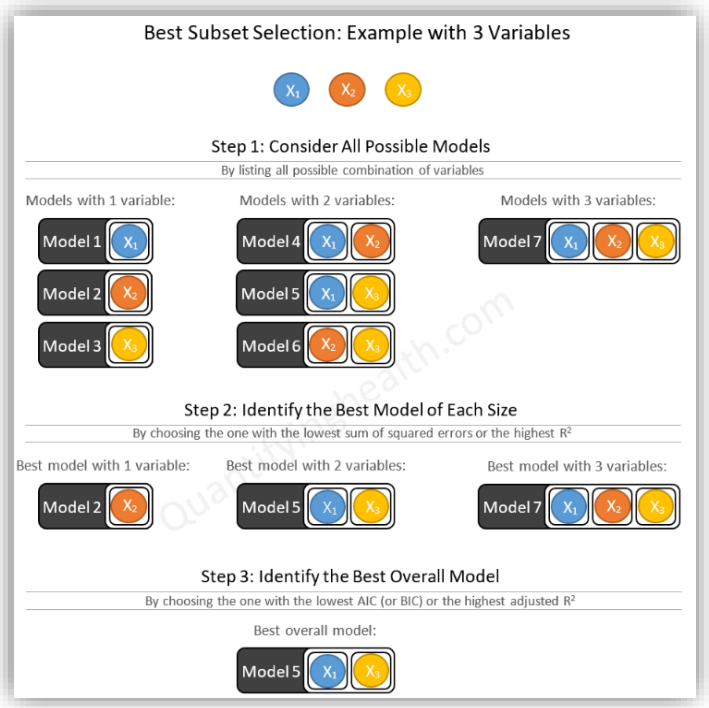
Αλγόριθμος 1 *Best subset selection.*

1. Έστω M_0 το μηδενικό μοντέλο, το μοντέλο που δεν περιέχει καμία ανεξάρτητη μεταβλητή, πάρα μόνο το σταθερό όρο β_0. Το μοντέλο αυτό προβλέπει το δειγματικό μέσο για κάθε παρατήρηση.
2. Για $k = 1,2,\dots,d$:

   (α) Προσαρμόζουμε όλα τα $\binom{d}{k}$ μοντέλα που περιέχουν ακριβώς k μεταβλητές.

   (β) Διαλέγουμε το καλύτερο ανάμεσα σε αυτά τα $\binom{d}{k}$ μοντέλα και το ονομάζουμε M_k. Καλύτερο ορίζεται αυτό που έχει το μικρότερο RSS ή ισοδύναμα το μεγαλύτερο R^2.
3. Επιλέγουμε το καλύτερο μοντέλο ανάμεσα στα $M_0, M_1, \dots, M_d$ μέσω κάποιου κριτηρίου, όπως είναι το AIC, το BIC ή το R^2_{adj}.

Έτσι, το πρόβλημα επιλογής του καλύτερου μοντέλου ανάμεσα από 2^d μοντέλα μειώνεται σε $d + 1$ μοντέλα. Στην περίπτωση της λογιστικής παλινδρόμησης, στο βήμα 2 του αλγορίθμου μπορεί να χρησιμοποιηθεί η απόκλιση.



ΣΧΗΜΑ 3-1 : Σχηματική αναπαράσταση της μεθόδου best subset selection.

Η forward selection είναι μία επαναληπτική μέθοδος επιλογής μοντέλου. Ξεκινάει από το μηδενικό μοντέλο και σε κάθε επανάληψη προσθέτει μία μεταβλητή στο μοντέλο σύμφωνα με κάποιο κριτήριο. Συνήθη κριτήρια είναι αυτά που χρησιμοποιούν το p-value, το AIC και το BIC. Η μέθοδος επαναλαμβάνεται μέχρι να μην μπορεί να προστεθεί κάποια μεταβλητή στο μοντέλο ή όλες οι μεταβλητές να έχουν εισαχθεί στο μοντέλο. Στην περίπτωση της λογιστικής παλινδρόμησης (Collett, 2002), η μεταβλητή που προστίθεται κάθε φορά στο μοντέλο είναι αυτή που μειώνει περισσότερο την απόκλιση. Όταν κάποια μεταβλητή δε μειώνει την απόκλιση περισσότερο από μία προκαθορισμένη τιμή, τότε η διαδικασία σταματά. Αυτό είναι γνωστό ως κανόνας διακοπής (*stopping rule*). Ο αντίστοιχος αλγόριθμος είναι ο εξής (James et al., 2023):

Αλγόριθμος 2 *Forward selection.*

1. Έστω M_0 το μηδενικό μοντέλο.
2. Για $k = 0,1, \dots, d-1$:

   (α) Εξετάζουμε όλα τα $d-k$ μοντέλα που αυξάνουν την προβλεπτική ικανότητα του μοντέλου M_k με μία επιπλέον μεταβλητή.

   (β) Επιλέγουμε το καλύτερο ανάμεσα από αυτά τα $d-k$ μοντέλα και το ονομάζουμε M_{k+1}. Καλύτερο ορίζεται αυτό που έχει το μικρότερο RSS ή το μεγαλύτερο R^2.
3. Επιλέγουμε το καλύτερο μοντέλο ανάμεσα στα $M_0, M_1, \dots, M_d$ μέσω κάποιου κριτηρίου, όπως είναι το σφάλμα πρόβλεψης σε ένα σύνολο επικύρωσης, το $C_p(AIC)$, το BIC ή το R^2_{adj}. Ή χρησιμοποιούμε τη μέθοδο διασταυρούμενης επικύρωσης.

Η backward elimination, και αυτή επαναληπτική μέθοδος, ξεκινάει από το πλήρες μοντέλο και σε κάθε επανάληψη αφαιρεί μία μεταβλητή, σύμφωνα πάλι με κάποιο κριτήριο. Η μέθοδος επαναλαμβάνεται μέχρι να μην μπορεί να αφαιρεθεί κάποια μεταβλητή από το μοντέλο ή όλες οι μεταβλητές να έχουν αφαιρεθεί. Στην περίπτωση της λογιστικής παλινδρόμησης (Collett, 2002), σε κάθε βήμα αφαιρείται η μεταβλητή που αυξάνει την απόκλιση το λιγότερο. Η διαδικασία σταματά όταν κάποια μεταβλητή αυξάνει την απόκλιση περισσότερο από μία προκαθορισμένη τιμή. Ο αλγόριθμος για την backward elimination είναι ο εξής (James et al., 2023):

Αλγόριθμος 3 *Backward elimination.*

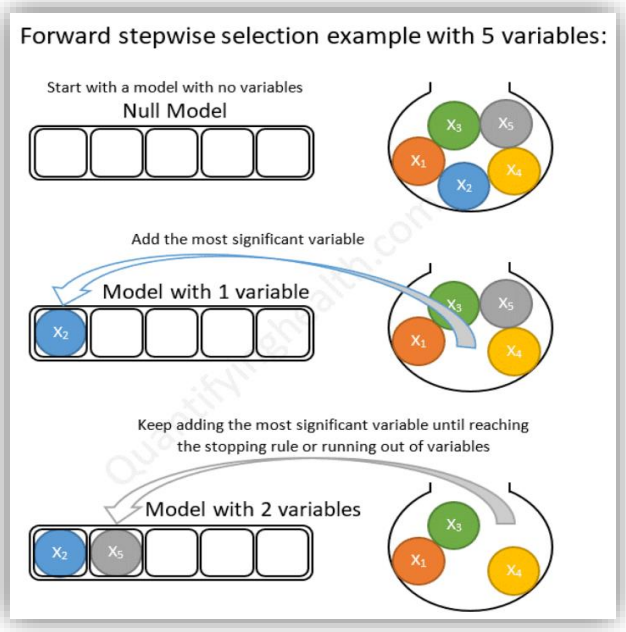
1. Έστω M_d το πλήρες μοντέλο, που περιέχει όλες τις d μεταβλητές.
2. Για $k = d, d-1, \dots 1$:

   (α) Εξετάζουμε όλα τα k μοντέλα που περιέχουν όλες τις μεταβλητές εκτός από μίας στο μοντέλο M_k, συνολικά για $k-1$ μεταβλητές.

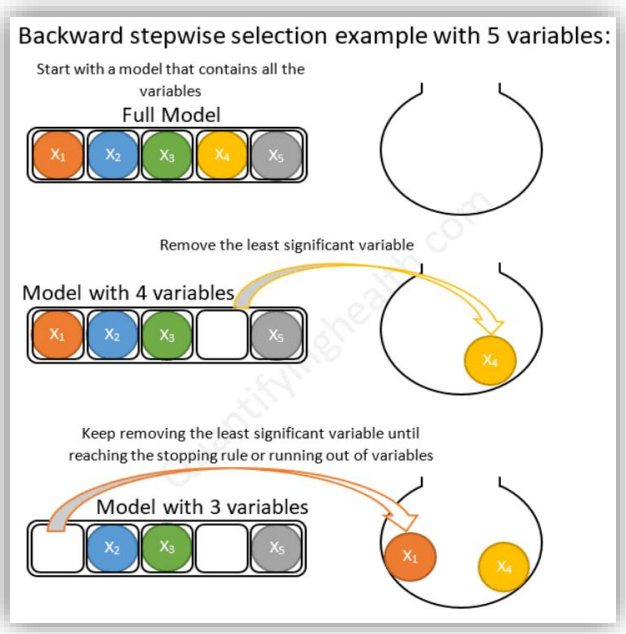
   (β) Επιλέγουμε το καλύτερο ανάμεσα από αυτά τα k μοντέλα και το ονομάζουμε M_{k-1}. Καλύτερο ορίζεται αυτό που έχει το μικρότερο RSS ή το μεγαλύτερο R^2.
3. Επιλέγουμε το καλύτερο μοντέλο ανάμεσα στα $M_0, M_1, \dots, M_d$ μέσω κάποιου κριτηρίου, όπως είναι το σφάλμα πρόβλεψης σε ένα σύνολο επικύρωσης, το $C_p(AIC)$, το BIC ή το R^2_{adj}. Ή χρησιμοποιούμε τη μέθοδο διασταυρούμενης επικύρωσης.

Η forward selection έχει τα εξής πλεονεκτήματα σε σχέση με την backward elimination (Fan et al., 2020). Όταν το d είναι αρκετά μεγάλο σε σχέση με το n ή όταν το $d < n$, η backward elimination δεν παράγει μια σταθερή διαδικασία επιλογής, σε αντίθεση με την forward selection, εφ'όσον αυτή σταματήσει αρκετά νωρίς. Στην περίπτωση που ο αριθμός των ανεξάρτητων μεταβλητών d είναι μεγαλύτερος από το πλήθος των παρατηρήσεων n ($d > n$), κάτι αρκετά συνηθισμένο στον τομέα της Βιοστατιστικής, η backward elimination δεν μπορεί να χρησιμοποιηθεί, καθώς δεν μπορούμε να προσαρμόσουμε το πλήρες μοντέλο. Η forward selection μπορεί να χρησιμοποιηθεί.

Οι μέθοδοι forward selection και backward elimination δεν είναι απαραίτητο ότι θα καταλήξουν στο ίδιο μοντέλο. Επίσης, παρ'ότι είναι εύκολες στη χρήση τους, έχουν ένα πολύ βασικό μειονέκτημα. Από τη στιγμή που μία μεταβλητή προστεθεί ή αφαιρεθεί αντίστοιχα από το μοντέλο, η σημαντικότητά της δεν επανεξετάζεται στην πορεία. Επίσης, σε κάθε επανάληψη βασίζονται απλά σε κάποιο στατιστικό κριτήριο, μη λαμβάνοντας υπ'όψιν το στόχο της έρευνας. Έτσι, μπορεί να καταλήξουμε με ένα μοντέλο που από στατιστικής άποψης να είναι ικανοποιητικό, όμως οι μεταβλητές που να περιέχει να μην είναι απαραίτητα σχετικές όσον αφορά την κατανόηση και την ερμηνεία των δεδομένων μας (Kroese et al., 2019).



ΣΧΗΜΑ 3-2 : Σχηματική αναπαράσταση της μεθόδου forward selection.



ΣΧΗΜΑ 3-3 : Σχηματική αναπαράσταση της μεθόδου backward elimination.

Η stepwise regression είναι μία μέθοδος που συνδυάζει την forward selection και την backward elimination. Μπορούμε είτε να ξεκινήσουμε χρησιμοποιώντας την

forward selection και στη συνέχεια να κάνουμε έλεγχο με την backward elimination, ή το αντίθετο.

3.1.4 Σχόλια

Οι λόγοι που προχωράμε σε επιλογή μεταβλητών συνοπτικά, είναι οι εξής (Faraway, 2002):

1) Θέλουμε να εξηγήσουμε τα δεδομένα με τον απλούστερο δυνατό τρόπο. Οπότε, περιττές μεταβλητές θα πρέπει να εξαιρεθούν και να καταλήξουμε σε ένα φειδωλό μοντέλο.
2) Μη σημαντικές μεταβλητές θα προσθέσουν θόρυβο (*noise*, όρος που περιγράφει τις ανεξήγητες διακυμάνσεις των δεδομένων) στην εκτίμηση των υπολοίπων μεταβλητών που μας ενδιαφέρουν.
3) Για να αποφύγουμε την πολυσυγγραμμικότητα. Ύπαρξη μεταβλητών δηλαδή που σχετίζονται γραμμικά μεταξύ τους (βλ. 3.4).
4) Αφαιρώντας περιττές μεταβλητές μειώνεται το γενικό κόστος, γλυτώνουμε χρόνο και χρήμα.

Πριν όμως προχωρήσουμε σε επιλογή μεταβλητών, θα πρέπει α) να εντοπίσουμε τυχόν ακραίες τιμές και παρατηρήσεις επιρροής και να αποφασίσουμε πώς θα τις διαχειριστούμε (εξαιρώντας τες ίσως προσωρινά) και β) να προσθέσουμε πιθανούς μετασχηματισμούς των μεταβλητών που φαίνονται κατάλληλοι.

Τέλος, ας αναλογιστούμε τα παρακάτω. Πρώτον, τις εξής δύο φράσεις. Η πρώτη είναι μία από τις πλέον γνωστές στον κόσμο της Στατιστικής, αυτή του George Box, Άγγλου στατιστικού: «Όλα τα μοντέλα είναι λάθος, αλλά κάποια είναι χρήσιμα» (*«All models are wrong, but some are useful»)*. Η δεύτερη είναι μία φράση που αποδίδεται στον Albert Einstein: «Όλα πρέπει να γίνουν όσο το δυνατόν πιο απλά, αλλά όχι πιο απλά» (*«Everything should be made as simple as possible, but no simpler»*). Και δεύτερον, την αρχή του «Ξυραφιού» του Όκαμ (*Occam's Razor*), γνωστή και ως αρχή της φειδούς (*principle of parsimony*), που λέει ότι ανάμεσα στις πιθανές εξηγήσεις ενός φαινομένου, η πιο απλή εξήγηση είναι και η καλύτερη. Στην παλινδρόμηση μπορεί να μεταφραστεί ως εξής: το μοντέλο με τις λιγότερες μεταβλητές που προσαρμόζεται καλά στα δεδομένα, είναι και το καλύτερο (Faraway, 2014).

3.2 Το Πρόβλημα της Υπερπροσαρμογής και της Υποπροσαρμογής

3.2.1 Σύνολο εκπαίδευσης και σύνολο ελέγχου

Αφού κατασκευάσουμε ένα μοντέλο, θα κληθούμε να αξιολογήσουμε την προβλεπτική του ικανότητα. Μπορεί να προβλέψει ικανοποιητικά την εξαρτημένη μεταβλητή Y για δεδομένα που δεν έχει «δει»; Ένα μοντέλο που προβλέπει σχεδόν τέλεια την Y για τα δεδομένα πάνω στα οποία έχει εκπαιδευτεί, αλλά δεν μπορεί να γενικευτεί σε νέα δεδομένα, θεωρείται καλό;

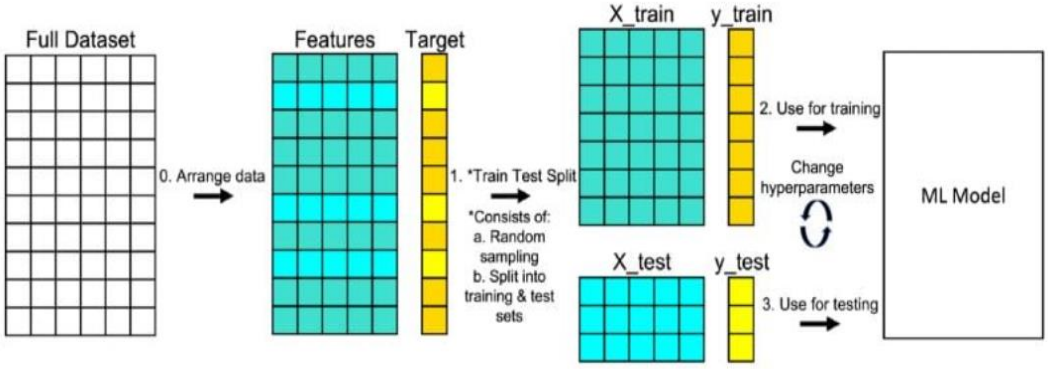
Μία πολύ γνωστή τεχνική πάνω σε αυτό, είναι να διαχωρίσουμε το σύνολο των δεδομένων μας σε δύο νέα υποσύνολα. Η τεχνική αυτή είναι αρκετά χρήσιμη ώστε να μετριάσουμε την υπερπροσαρμογή (βλ. 3.2.2) του μοντέλου. Το ένα ονομάζεται σύνολο εκπαίδευσης ή εκμάθησης (*training set*), όπου αποτελείται από τα (X_{train}, Y_{train}), και το άλλο σύνολο ελέγχου ή επικύρωσης (*test set or validation set*) με (X_{test}, Y_{test}) (συνήθως, μπορεί να υπάρξει και τριπλός διαχωρισμός των δεδομένων σε 1) σύνολο εκπαίδευσης, 2) σύνολο ελέγχου και 3) σύνολο επικύρωσης. Στην παρούσα διπλωματική όμως όταν αναφερόμαστε σε σύνολο ελέγχου και σε σύνολο επικύρωσης, θα αναφερόμαστε στο ίδιο σύνολο). Για να έχουμε αξιόπιστα αποτελέσματα, τα σύνολα εκπαίδευσης και ελέγχου θα πρέπει να είναι αντιπροσωπευτικά του συνόλου των δεδομένων. Τα πιο συνήθη ποσοστά διαχωρισμού είναι αντίστοιχα 70-30, 75-25 ή 80-20, αναλόγως βέβαια το μέγεθος n του συνολικού δείγματος. Ο διαχωρισμός των δεδομένων μπορεί να γίνει με τους εξής τρόπους:

1) Με τυχαία δειγματοληψία (*random sampling*). Τα δεδομένα για αρχή «ανακατεύονται», ώστε να εξασφαλίσουμε ότι δεν υπάρχει κάποιου είδους ταξινόμηση, και μετά με τυχαίο τρόπο κατανέμονται στα σύνολα εκπαίδευσης και ελέγχου. Όμως, η τυχαία δειγματοληψία δεν είναι σωστή στην περίπτωση που τα δεδομένα είναι μη-ισορροπημένα (*imbalanced data*), δηλαδή τα ποσοστά των παρατηρήσεων που ανήκουν στην κατηγορία 0 και 1 είναι κατά πολύ διαφορετικά μεταξύ τους (π.χ. σε δείγμα μεγέθους $n = 100$, οι 90 παρατηρήσεις να έχουν $Y = 0$ και οι υπόλοιπες 10 $Y = 1$).
2) Με στρωματοποιημένη δειγματοληψία (*stratified sampling*). Χρησιμοποιείται κυρίως για μη-ισορροπημένα δεδομένα. Τα δεδομένα χωρίζονται στα δύο,

διατηρώντας την αρχική αναλογία των παρατηρήσεων 0 και 1 ανάμεσα στα σύνολα εκπαίδευσης και ελέγχου.

3) Με διασταυρούμενη επικύρωση (*cross-validation*).

Για να κατασκευάσουμε το μοντέλο, χρησιμοποιούμε το σύνολο εκπαίδευσης. Μέσω του συνόλου εκπαίδευσης, θα εκτιμήσουμε τις παραμέτρους $\boldsymbol{\beta}$ και θα επιλέξουμε τις μεταβλητές που θα μπουν στο μοντέλο. Έπειτα, το σύνολο ελέγχου θα χρησιμοποιηθεί για την αξιολόγηση του μοντέλου. Θα προβλέψουμε τις τιμές της Y_{test} μέσω του μοντέλου που καταλήξαμε και θα συγκρίνουμε τις προβλεπόμενες με τις πραγματικές τιμές της Y_{test}, ώστε να δούμε πόσο ακριβές είναι το μοντέλο μας. Τέλος, αξιολογούμε τα αποτελέσματα, αναπροσαρμόζουμε το μοντέλο αν χρειάζεται και επαναλαμβάνουμε τη διαδικασία, μέχρι να καταλήξουμε στο τελικό μοντέλο.



ΣΧΗΜΑ 3-4 : Σχηματική απεικόνιση του διαχωρισμού των δεδομένων και της διαδικασίας κατασκευής του τελικού μοντέλου (Features = οι ανεξάρτητες μεταβλητές X , Target = η εξαρτημένη μεταβλητή Y).

3.2.2 Τι είναι υπερπροσαρμογή και τι υποπροσαρμογή

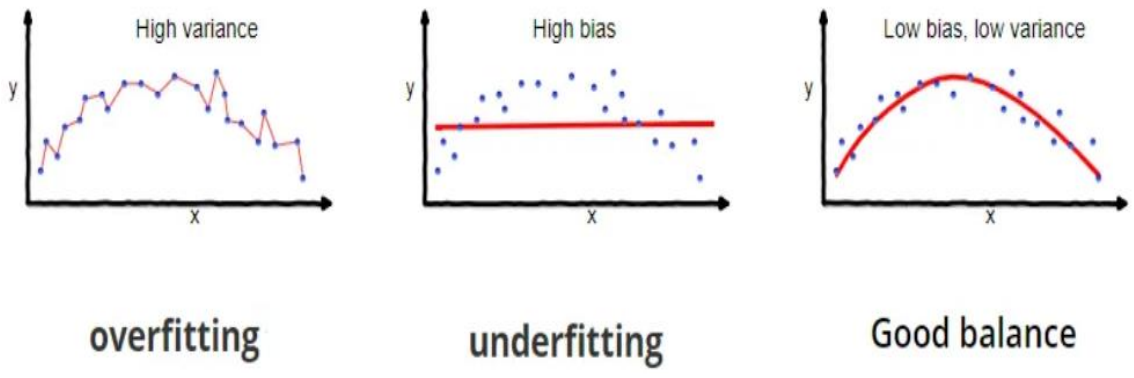
Στην ενότητα 3.1.2 αναφέραμε ότι κατά τη δημιουργία ενός μοντέλου, σκοπός μας δεν είναι να έχουμε ένα μοντέλο με πάρα πολλές μεταβλητές που θα περιγράφει τέλεια τα δεδομένα μας, αλλά ένα φειδωλό μοντέλο, με λίγες μεταβλητές δηλαδή, το οποίο θα περιγράφει την ουσία των δεδομένων και θα μπορεί να γενικευτεί.

Ένα από τα πιο συχνά προβλήματα που μπορούμε να συναντήσουμε κατά την κατασκευή μοντέλων, είναι αυτό της υπερπροσαρμογής (*overfitting*). Κατά την υπερπροσαρμογή, το μοντέλο έχει «μάθει» σχεδόν τέλεια τα δεδομένα που χρησιμοποιήθηκαν για να το κατασκευάσουμε. Έχει μάθει το μοτίβο που αυτά

κατανέμονται, έχει μάθει τον θόρυβο που υπάρχει και άρα το μοντέλο αυτό έχει μεγάλη διακύμανση (*variance*). Όσο περισσότερες μεταβλητές περιέχει το μοντέλο, τόσο πιο πολύπλοκο είναι και άρα μπορεί να εντοπίσει πιο περίπλοκες διακυμάνσεις των δεδομένων. Έτσι, το μοντέλο έχει μάθει τόσο καλά τα δεδομένα εκπαίδευσης, που δεν προβλέπει καθόλου καλά νέα δεδομένα, δεν έχει ικανότητα γενίκευσης. Δύο από τις προτεινόμενες λύσεις που υπάρχουν για την αντιμετώπιση της υπερπροσαρμογής είναι α) να μειώσουμε τις μεταβλητές του μοντέλου και β) να εφαρμόσουμε ποινικοποίηση (βλ. 3.6) (Géron, 2022).

Το αντίθετο πρόβλημα είναι αυτό της υποπροσαρμογής (*underfitting*). Το μοντέλο δεν αναγνωρίζει τη σχέση μεταξύ των *X* και της *Y* και έτσι οι προβλέψεις του δεν είναι ακριβείς ούτε για τα δεδομένα εκπαίδευσης, πόσο μάλλον για τα νέα δεδομένα. Ένα μοντέλο που είναι υποπροσαρμοσμένο, είναι αρκετά απλό ώστε να αναγνωρίσει το μοτίβο των δεδομένων και συνήθως έχει υψηλή μεροληψία (*bias*). Μία πιθανή λύση για την υποπροσαρμογή είναι να αυξήσουμε τις μεταβλητές του μοντέλου (Géron, 2022).

Ένα καλό μοντέλο έχει και καλή προβλεπτική ικανότητα στα δεδομένα εκπαίδευσης, αλλά και στα νέα δεδομένα. Έχει και χαμηλή διακύμανση, αλλά και χαμηλή μεροληψία. Όλα αυτά, μπορούμε να τα δούμε στις παρακάτω γραφικές απεικονίσεις. Στο Σχήμα 3-5 μπορούμε να δούμε ότι το υπερπροσαρμοσμένο μοντέλο περιγράφει τέλεια τα δεδομένα εκπαίδευσης, καθώς αναγνωρίζει τις διακυμάνσεις τους. Το υποπροσαρμοσμένο δεν αναγνωρίζει καθόλου το μοτίβο των δεδομένων, ενώ καλό θεωρείται το μοντέλο που ναι μεν έχει αναγνωρίσει το μοτίβο, αλλά δεν περιγράφει ακριβώς τις διακυμάνσεις· έχει χαμηλή διακύμανση και παράλληλα χαμηλή μεροληψία.

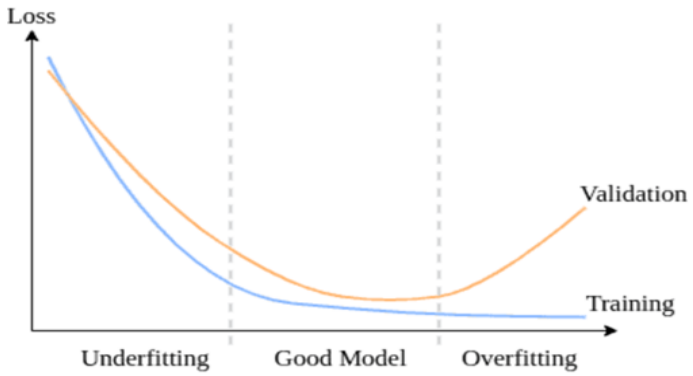


ΣΧΗΜΑ 3-5 : Γραφική αναπαράσταση υπερπροσαρμογής, υποπροσαρμογής και καλής προσαρμογής.

3.2.3 Εντοπισμός υπερπροσαρμογής και υποπροσαρμογής μέσω γραφημάτων

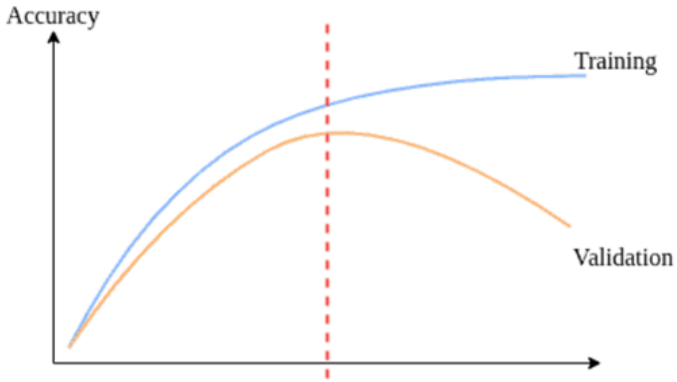
Την υπερπροσαρμογή και την υποπροσαρμογή μπορούμε να τις εντοπίσουμε κατασκευάζοντας τα εξής δύο γραφήματα: το γράφημα για το αθροιστικό προβλεπτικό σφάλμα (*loss*) και το γράφημα για την ακρίβεια (*accuracy*) των προβλέψεων.

Στο Σχήμα 3-6 απεικονίζεται το σφάλμα του μοντέλου στο σύνολο εκπαίδευσης και στο σύνολο ελέγχου ως προς την πολυπλοκότητα του μοντέλου. Αριστερά από την πρώτη κάθετη διακεκομμένη, αφορά την περίπτωση της υποπροσαρμογής, όπου το σφάλμα είναι αρκετά υψηλό και στα δύο σύνολα δεδομένων. Δεξιά από τη δεύτερη διακεκομμένη, αφορά την υπερπροσαρμογή. Στην περίπτωση των δεδομένων εκπαίδευσης, το σφάλμα είναι αρκετά χαμηλό, αλλά για τα νέα δεδομένα όλο και αυξάνεται. Η βέλτιστη περίπτωση είναι ανάμεσα στις δύο διακεκομμένες. Το σφάλμα για τα δεδομένα εκπαίδευσης και για τα δεδομένα ελέγχου είναι αρκετά χαμηλό, αλλά παράλληλα τα σφάλματα είναι και κοντά μεταξύ τους.



ΣΧΗΜΑ 3-6 : Αθροιστικό προβλεπτικό σφάλμα του μοντέλου στο σύνολο εκπαίδευσης (Training) και στο σύνολο ελέγχου (Validation).

Στο Σχήμα 3-7 απεικονίζεται η ακρίβεια των προβλέψεων του μοντέλου ως προς την πολυπλοκότητα του μοντέλου. Αριστερά από την κάθετη διακεκομμένη, η ακρίβεια των προβλέψεων είναι αρκετά χαμηλή· περίπτωση υποπροσαρμογής. Δεξιά, η ακρίβεια στα σύνολα εκπαίδευσης αυξάνεται, αλλά στα νέα δεδομένα αρχίζει και μειώνεται· υπερπροσαρμογή. Η διακεκομμένη αφορά το βέλτιστο μοντέλο, το οποίο έχει υψηλή ακρίβεια προβλέψεων και στα δεδομένα εκπαίδευσης και στα δεδομένα ελέγχου.



ΣΧΗΜΑ 3-7 : Ακρίβεια των προβλέψεων του μοντέλου στο σύνολο εκπαίδευσης (Training) και στο σύνολο ελέγχου (Validation).

3.3 Αντιστάθμιση Μεροληψίας-Διακύμανσης

Στην ενότητα 3.2 αναλύσαμε τους όρους της υπερπροσαρμογής και της υποπροσαρμογής. Ένα υπερπροσαρμοσμένο μοντέλο είναι αρκετά πολύπλοκο και έχει υψηλή διακύμανση. Αντιθέτως, ένα υποπροσαρμοσμένο δεν είναι όσο πολύπλοκο θα έπρεπε και έχει υψηλή μεροληψία. Έτσι, προκύπτει ένα από τα πλέον σημαντικά ζητήματα κατά τη δημιουργία ενός μοντέλου, αυτό για την αντιστάθμιση μεροληψίας-διακύμανσης (*bias-variance trade-off*). Να εντοπίσουμε δηλαδή την ισορροπία μεταξύ της μεροληψίας και της διακύμανσης όπου το μοντέλο θα είναι, όσο το δυνατόν γίνεται, βέλτιστο.

Για αρχή, ας ορίσουμε τις έννοιες της μεροληψίας και της διακύμανσης. Μεροληψία είναι η διαφορά μεταξύ της πραγματικής και της προβλεπόμενης τιμής (Agresti and Kateri, 2021). Είναι οι απλές υποθέσεις που κάνει το μοντέλο για τα δεδομένα, ώστε να είναι ικανό να προβλέψει νέα δεδομένα. Π.χ. η υποβόσκουσα σχέση των δεδομένων είναι γραμμική, ενώ στην πραγματικότητα είναι κάποιας άλλης μορφής (James et al., 2023). Διακύμανση ορίζεται η ευαισθησία του μοντέλου στις διακυμάνσεις των δεδομένων (Géron, 2022). Ένα μοντέλο που μαθαίνει και από το θόρυβο, ασήμαντα χαρακτηριστικά των δεδομένων θα τα θεωρεί σημαντικά.

Τώρα, έστω η προβλεπόμενη τιμή της Y μέσω κάποιας συνάρτησης:

$$\hat{Y} = \hat{f}(X),$$

όπου $X = (X_1, X_2, \dots X_d)$ και $\hat{f}$ η εκτίμηση της συνάρτησης f. Η ακρίβεια της $\hat{Y}$ ως πρόβλεψη της Y αποτελείται από δύο ποσότητες, το αναγώγιμο σφάλμα (*reducible error*) και το μη αναγώγιμο σφάλμα (*irreducible error*). Επειδή η $\hat{f}$ είναι απίθανο να είναι ακριβής εκτίμηση της f, θα εισάγει ένα σφάλμα, το οποίο όμως μπορεί να μειωθεί, μέσω χρήσης κατάλληλων τεχνικών στατιστικής μάθησης για την εκτίμηση της f (αναγώγιμο σφάλμα). Ακόμα όμως και αν ήταν πιθανό να εκτιμήσουμε τέλεια την f, πάντα θα υπάρχει ένα σφάλμα, λόγω του ότι η Y ορίζεται ως:

$$Y = f(X) + \varepsilon,$$

όπου το σφάλμα ε, εξ ορισμού, δεν μπορεί να προβλεφθεί μέσω του X (μη αναγώγιμο σφάλμα). Το μη αναγώγιμο σφάλμα πάντα θα θέτει ένα άνω όριο στην ακρίβεια των προβλέψεων της Y, το οποίο στην πράξη όμως δε θα είναι σχεδόν ποτέ γνωστό (James et al., 2023). Από τα παραπάνω έχουμε ότι:

$$E\left(Y - \hat{Y}\right)^2 = E\left[f(X) + \varepsilon - \hat{f}(X)\right]^2 \Leftrightarrow$$

$$E\left(Y - \hat{Y}\right)^2 = \underbrace{\left[f(X) - \hat{f}(X)\right]^2}_{\text{αναγώγιμο σφάλμα}} + \underbrace{Var(\varepsilon)}_{\text{μη αναγώγιμο σφάλμα}}.$$

Τώρα, έστω ότι έχουμε δημιουργήσει το μοντέλο μας και έχουμε μία νέα παρατήρηση, την (x_0, y_0), και θέλουμε να ελέγξουμε πόσο καλά μπορεί να προβλεφθεί αυτή μέσω του μοντέλου. Το μέσο τετραγωνικό σφάλμα (*mean squared error-MSE*) για την παρατήρηση αυτή ορίζεται ως:

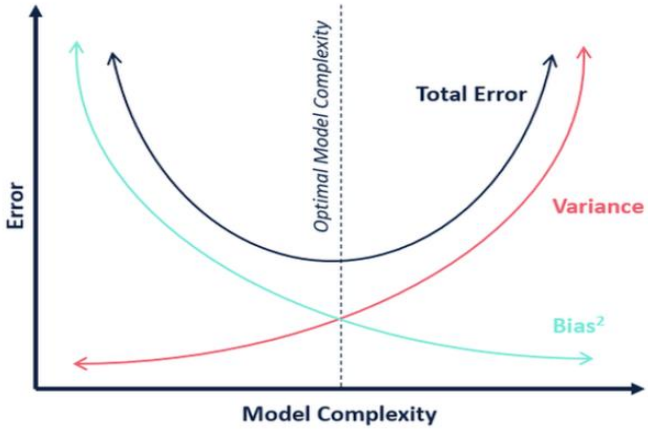
$$E\left(y_0 - \hat{f}(x_0)\right)^2 = Var\left(\hat{f}(x_0)\right) + \left[Bias\left(\hat{f}(x_0)\right)\right]^2 + Var(\varepsilon).$$

Για να μειώσουμε το σφάλμα αυτό όσο το δυνατόν περισσότερο, θα πρέπει να επιτύχουμε χαμηλή διακύμανση και χαμηλή μεροληψία. Όμως, όπως θα δούμε και στη συνέχεια αναλυτικότερα, δεν μπορούμε να μειώσουμε και τα δύο ταυτόχρονα. Και εκεί είναι που εμφανίζεται η ανάγκη για την αντιστάθμιση μεροληψίας-διακύμανσης.

Στο Σχήμα 3-8 βλέπουμε το σφάλμα του μοντέλου σε σχέση με την πολυπλοκότητά του. Μπορούμε να δούμε ότι:

- Όσο αυξάνεται η πολυπλοκότητα, αυξάνεται και η διακύμανση.
- Όσο μειώνεται η πολυπλοκότητα, αυξάνεται η μεροληψία.
- Όσο αυξάνεται η διακύμανση, αυξάνεται και το σφάλμα.
- Όσο αυξάνεται η μεροληψία, αυξάνεται και το σφάλμα.

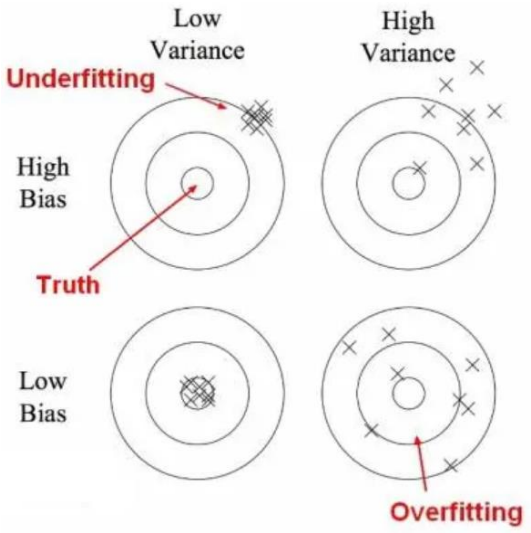
- Η κάθετη διακεκομμένη γραμμή μας δείχνει την πολυπλοκότητα του μοντέλου για την οποία το συνολικό σφάλμα ελαχιστοποιείται και άρα την ισορροπία μεταξύ μεροληψίας και διακύμανσης.
- Δεξιά από τη διακεκομμένη, όπου υπάρχει υψηλή διακύμανση και χαμηλή μεροληψία, ορίζεται η περιοχή της υπερπροσαρμογής, ενώ αριστερά, που υπάρχει χαμηλή διακύμανση και υψηλή μεροληψία, της υποπροσαρμογής.



ΣΧΗΜΑ 3-8 : Μεροληψία και διακύμανση για την πολυπλοκότητα του μοντέλου.

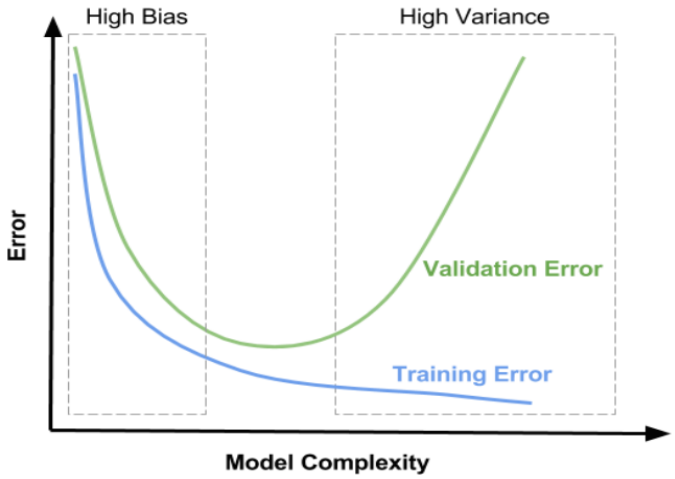
Γενικά, υπάρχουν 4 πιθανοί συνδυασμοί μεροληψίας και διακύμανσης. Οι συνδυασμοί αυτοί απεικονίζονται στο Σχήμα 3-9 με τη βοήθεια στόχων, όπου το κέντρο του στόχου αφορά μοντέλο που έχει τέλεια προβλεπτική ικανότητα. Όσο οι προβλέψεις απομακρύνονται από το κέντρο, τόσο το χειρότερες είναι. Οι συνδυασμοί αυτοί, ξεκινώντας από πάνω αριστερά, είναι:

1) Υψηλή μεροληψία και χαμηλή διακύμανση. Κατά μέσο όρο, οι προβλέψεις είναι συνεπείς αλλά ανακριβείς. Σε αυτήν την περίπτωση, έχουμε ένα υποπροσαρμοσμένο μοντέλο.
2) Υψηλή μεροληψία και υψηλή διακύμανση. Κατά μέσο όρο, οι προβλέψεις είναι και ασυνεπείς και ανακριβείς.
3) Χαμηλή μεροληψία και χαμηλή διακύμανση. Κατά μέσο όρο, οι προβλέψεις είναι και συνεπείς και ακριβείς. Αφορά το ιδανικό μοντέλο, που πρακτικά όμως είναι σχεδόν απίθανο να το επιτύχουμε.
4) Χαμηλή μεροληψία και υψηλή διακύμανση. Κατά μέσο όρο, οι προβλέψεις είναι ασυνεπείς αλλά ακριβείς. Αφορά το υπερπροσαρμοσμένο μοντέλο.



ΣΧΗΜΑ 3-9 : Συνδυασμοί μεροληψίας και διακύμανσης.

Στο γράφημα που ακολουθεί (Σχήμα 3-10), απεικονίζεται το σφάλμα στα δεδομένα εκπαίδευσης και στα νέα δεδομένα σε σχέση με την πολυπλοκότητα του μοντέλου. Στην περιοχή της υψηλής μεροληψίας, βλέπουμε ότι το σφάλμα είναι υψηλό και στα δύο σύνολα δεδομένων. Στην περιοχή της υψηλής διακύμανσης, το σφάλμα στα δεδομένα εκπαίδευσης είναι αρκετά χαμηλό, όμως για τα νέα δεδομένα ολοένα και αυξάνεται. Στην ενδιάμεση περιοχή, βλέπουμε χαμηλό σφάλμα και στα δύο σύνολα δεδομένων. Εδώ έχει επιτευχθεί χαμηλή διακύμανση και χαμηλή μεροληψία.



ΣΧΗΜΑ 3-10 : Το σφάλμα στα δεδομένα εκπαίδευσης και στα δεδομένα ελέγχου σε σχέση με την πολυπλοκότητα του μοντέλου.

3.4 Το Φαινόμενο της Πολυσυγγραμμικότητας

3.4.1 Ορισμός της πολυσυγγραμμικότητας

Ένα άλλο πρόβλημα που μπορεί να συναντήσουμε κατά τη δημιουργία ενός στατιστικού μοντέλου, είναι αυτό της πολυσυγγραμμικότητας (*multicollinearity*). Πολυσυγγραμμικότητα υπάρχει όταν μία από τις ανεξάρτητες μεταβλητές μπορεί να εκφραστεί ως γραμμική συνάρτηση μίας ή και περισσοτέρων εκ των υπολοίπων ανεξάρτητων μεταβλητών. Δηλαδή, η πληροφορία κάποιας μεταβλητής περιγράφεται από τις υπόλοιπες με τις οποίες συσχετίζεται και επομένως η μεταβλητή αυτή μπορεί να κριθεί ως περιττή, καθώς η επιρροή της στο μοντέλο μπορεί να μην είναι στατιστικά σημαντική. Μεταβλητές που δε συσχετίζονται γραμμικά μεταξύ τους αποκαλούνται ορθογώνιες (*orthogonal*). Η ύπαρξη πολυσυγγραμμικότητας επηρεάζει αρκετά τις εκτιμήσεις των συντελεστών, οπότε τα συμπεράσματα στα οποία θα καταλήξουμε μέσω του μοντέλου θα είναι λανθασμένα. Επίσης, η παρουσία σε κάποιο μοντέλο υψηλά συσχετισμένων μεταβλητών μπορεί να οδηγήσει σε αστάθεια των εκτιμώμενων συντελεστών, δηλαδή μια μικρή αλλαγή στα δεδομένα να αλλάζει εντελώς τις εκτιμήσεις, αυξάνοντας έτσι τα τυπικά τους σφάλματα (Chatterjee and Simonoff, 2020). Ακόμα, αν σε ένα μοντέλο χρησιμοποιηθούν αρκετές μεταβλητές που συσχετίζονται μεταξύ τους, είναι αρκετά πιθανό το μοντέλο να είναι υπερπροσαρμοσμένο (Agresti and Kateri, 2021).

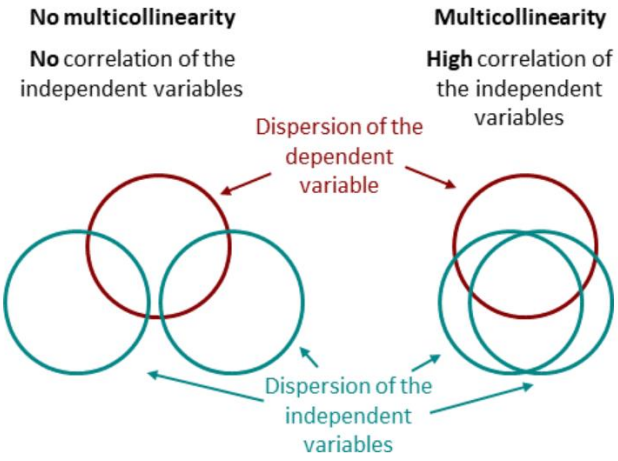
Έστω ότι έχουμε τον πίνακα $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_d]$, τον πίνακα που περιέχει τις n παρατηρήσεις για τις μεταβλητές $\boldsymbol{X}_j$ με $j = 1,2, \dots, d$. Μπορούμε να ορίσουμε την πολυσυγγραμμικότητα με όρους γραμμικής εξάρτησης των στηλών του $\boldsymbol{X}$. Τα διανύσματα $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_d$ είναι γραμμικά εξαρτώμενα αν υπάρχει ένα σύνολο σταθερών $t_1, t_2, \dots, t_d$, όχι όλα 0, τέτοιο ώστε:

$$\sum_{j=1}^{d} t_j \boldsymbol{X}_j = 0.$$

Αν η σχέση αυτή ισχύει ακριβώς για ένα υποσύνολο των στηλών του $\boldsymbol{X}$, τότε ο βαθμός (*rank*) του $\boldsymbol{X}^T\boldsymbol{X}$ είναι μικρότερος από d και ο $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ δεν υπάρχει. Ωστόσο, αν υποθέσουμε ότι η σχέση αυτή είναι περίπου αληθής για κάποιο υποσύνολο των στηλών του $\boldsymbol{X}$, τότε θα υπάρχει

μια σχεδόν γραμμική εξάρτηση στον πίνακα $\boldsymbol{X}^T\boldsymbol{X}$ και άρα εμφανίζεται το πρόβλημα της πολυσυγγραμμικότητας.

Στο παρακάτω σχήμα μπορούμε να δούμε μία σχηματική απεικόνιση της πολυσυγγραμικότητας. Ο κάθε κύκλος αντιπροσωπεύει τη διασπορά της εκάστοτε μεταβλητής. Στην περίπτωση αριστερά δεν υπάρχει πολυσυγγραμμικότητα μεταξύ των δύο ανεξάρτητων μεταβλητών, οπότε παρατηρούμε ότι επεξηγούν διαφορετικό κομμάτι της διασποράς της εξαρτημένης μεταβλητής. Στην περίπτωση δεξιά υπάρχει πολυσυγγραμμικότητα και βλέπουμε ότι οι ανεξάρτητες μεταβλητές επεξηγούν ένα μεγάλο κοινό κομμάτι της εξαρτημένης.



ΣΧΗΜΑ 3-11 : Σχηματική απεικόνιση της πολυσυγγραμμικότητας.

Σύμφωνα με τους Montgomery et al. (2021), υπάρχουν τέσσερες κύριες πηγές για την ύπαρξη πολυσυγγραμμικότητας:

1) Η μέθοδος συλλογής δεδομένων που χρησιμοποιήθηκε.
2) Περιορισμοί στο μοντέλο ή στον πληθυσμό.
3) Προδιαγραφές του μοντέλου.
4) Ένα υπερκαθορισμένο μοντέλο, που έχει δηλαδή μεγαλύτερο πλήθος μεταβλητών από ότι πλήθος παρατηρήσεων ($d > n$).

3.4.2 Ο δείκτης VIF

Από αυτά που ήδη έχουμε αναφέρει, καταλαβαίνουμε ότι θα πρέπει να εντοπίσουμε τις πιθανές μεταβλητές που παρουσιάζουν πολυσυγγραμικότητα και να αποφασίσουμε πώς θα προχωρήσουμε κατά τη δημιουργία του μοντέλου. Ο πιο δημοφιλής τρόπος

εντοπισμού της πολυσυγγραμμικότητας είναι μέσω του Παράγοντα Διόγκωσης Διακύμανσης (*Variance Inflation Factor, VIF*), ο οποίος όμως μπορεί να υπολογιστεί μόνο για ποσοτικές μεταβλητές. Ο δείκτης *VIF* υπολογίζεται ως εξής:

$$VIF_k = \frac{1}{1 - R_k^2},$$

όπου R_k^2 είναι ο συντελεστής προσδιορισμού για το μοντέλο που έχει την X_k ως εξαρτημένη μεταβλητή και όλες τις υπόλοιπες ως ανεξάρτητες. Όσο μεγαλύτερο το R_k^2, τόσο μεγαλύτερος θα είναι ο δείκτης VIF_k. Ο δείκτης VIF_k μας δίνει το κατά πόσο θα αυξηθεί η διακύμανση ενός εκτιμώμενου συντελεστή ($Var(\hat{\beta}_k)$), στην περίπτωση που η αντίστοιχη μεταβλητή εμφανίζει πολυσυγγραμμικότητα. Εμπειρικά ισχύει ότι:

- αν $VIF_k \cong 1$, η X_k δεν εμφανίζει πρόβλημα πολυσυγγραμμικότητας.
- αν $VIF_k > 10$, η X_k εμφανίζει πρόβλημα πολυσυγγραμμικότητας.

Επίσης, μπορούμε να χρησιμοποιήσουμε τον μέσο όρο ως ένα κριτήριο για τη μη ύπαρξη πολυσυγγραμμικότητας στο υπό εξέταση μοντέλο:

$$\overline{VIF} = \frac{1}{d}\sum_{k=1}^{d} VIF_k.$$

Αν $\overline{VIF} = 1$, τότε είναι εύκολο να δούμε ότι ισχύει $VIF_k = 1 \ \forall k$ και άρα καμία μεταβλητή δεν παρουσιάζει πολυσυγγραμμικότητα σε σχέση με τις άλλες. Αν όμως το $\overline{VIF}$ είναι αρκετά μεγαλύτερο του 1, τότε σημαίνει ότι το VIF_k για ένα ή και περισσότερα k είναι μεγαλύτερο του 1 και άρα υπάρχει ένδειξη ύπαρξης πολυσυγγραμμικότητας.

Μία προσέγγιση για την αντιμετώπιση των μεταβλητών που παρουσιάζουν πολυσυγγραμμικότητα είναι να διαλέξουμε ένα υποσύνολο ανεξάρτητων μεταβλητών, αφαιρώντας αυτές που εξηγούν μικρότερο μέρος της Y (Agresti, 2015).

3.5 Δεδομένα Υψηλών Διαστάσεων

3.5.1 Ορισμός

Μέχρι και πριν από λίγα χρόνια, τα περισσότερα προβλήματα, καθώς και οι περισσότερες θεωρίες και πρακτικές στην Στατιστική, είχαν να ασχοληθούν κυρίως με ένα σχετικά μεγάλο δείγμα n στο οποίο καταγραφόταν ένας μικρός αριθμός

χαρακτηριστικών d· ένας τυπικός κανόνας ήταν να υπάρχουν τουλάχιστον πέντε παρατηρήσεις για κάθε χαρακτηριστικό. Είχαν να κάνουν δηλαδή με προβλήματα χαμηλών διαστάσεων (*low-dimensional*). Αυτό βέβαια σχετίζεται άμεσα με τις δυνατότητες και τους περιορισμούς που υπήρχαν στα υπολογιστικά συστήματα της εποχής. Με την πάροδο του χρόνου όμως και τις τεχνολογικές εξελίξεις, έχουμε δεδομένα στα οποία το πλήθος των παραμέτρων είναι κατά πολύ μεγαλύτερο του πλήθους του δείγματος, π.χ. στο πεδίο της Βιοπληροφορικής (*Bioinformatics*), στον τομέα της Γονιδιωματικής (*Genomics*), της Αστρονομίας ή σε προβλήματα ανάλυσης εικόνων (*image analysis*). Έτσι, ενώ το d μπορεί να είναι αρκετά μεγάλο, το πλήθος των παρατηρήσεων n είναι συχνά περιορισμένο λόγω κόστους, διαθέσιμου δείγματος κλπ. Δεδομένα Υψηλών Διαστάσεων (*high-dimensional data*) καλούνται τα δεδομένα που ο αριθμός των μεταβλητών d ξεπερνάει συνήθως κατά πολύ το μέγεθος του δείγματος n ($d \gg n$).

Τα δεδομένα υψηλών διαστάσεων δεν πρέπει να συγχέονται με τα Μεγάλα Δεδομένα (*Big Data*), όπου ο όρος αυτός χρησιμοποιείται για να περιγράψει σύνολα δεδομένων τόσο μεγάλα ή σύνθετα σε όγκο τα οποία αυξάνονται ταχύτατα στο πέρασμα του χρόνου.

3.5.2 Προβλήματα στα δεδομένα υψηλών διαστάσεων

Σε κάθε στατιστική διαδικασία υπάρχουν τρεις σημαντικοί πυλώνες: στατιστική ακρίβεια, ερμηνευσιμότητα του μοντέλου και υπολογιστική πολυπλοκότητα. Το πρόβλημα ξεκινάει όταν $d > n$, όπου θα αντιμετωπίσουμε κάποιο πρόβλημα σχετικά με αυτούς τους πυλώνες. Στην περίπτωση των δεδομένων υψηλών διαστάσεων τα κλασσικά εργαλεία στατιστικής συμπερασματολογίας δεν μπορούν να χρησιμοποιηθούν, καθώς αντιμετωπίζουν διάφορα προβλήματα. Πώς μπορεί να ερμηνευθεί ένα μοντέλο υψηλής διάστασης ή πώς μπορούν οι στατιστικές διαδικασίες να γίνουν υπολογιστικά αποτελεσματικές και ισχυρές, είναι κάποιες από τις προκλήσεις που αντιμετωπίζουμε. Η ανάλυση όμως των δεδομένων υψηλών διαστάσεων είναι πλέον σημαντική και αναγκαία.

Ας υποθέσουμε ότι χρησιμοποιούμε κάποια από τις γνωστές στατιστικές μεθόδους σε δεδομένα υψηλών διαστάσεων, π.χ. τη μέθοδο ελαχίστων τετραγώνων (*least squares method*). Ακόμα και αν δεν υπάρχει κάποια πραγματική σχέση μεταξύ της Y και των $X_1, X_2, \dots, X_d$, η μέθοδος των ελαχίστων τετραγώνων θα μας δώσει ένα σύνολο εκτιμήσεων για τους συντελεστές που προσαρμόζονται τέλεια στα δεδομένα, έτσι ώστε τα κατάλοιπα να είναι 0 (James et al., 2023). Αυτή η τέλεια προσαρμογή είναι σχεδόν σίγουρο ότι θα οδηγήσει σε υπερπροσαρμογή των δεδομένων. Και για αυτό το λόγο, επειδή είναι πολύ εύκολο να καταλήξουμε σε ένα μοντέλο με μηδενικά κατάλοιπα που δεν είναι χρήσιμο, δε θα πρέπει ποτέ να χρησιμοποιούμε μέτρα όπως το άθροισμα τετραγώνων των σφαλμάτων, τα p-values, το R^2 κλπ για την αξιολόγηση της προσαρμογής ενός μοντέλου σε δεδομένα υψηλών διαστάσεων (James et al., 2023). Αντ'αυτού, τα συμπεράσματα θα πρέπει να προέρχονται από ένα ανεξάρτητο σύνολο ελέγχου ή από τα σφάλματα διασταυρούμενης επικύρωσης (*cross-validation errors*). Έτσι, το μέσο τετραγωνικό σφάλμα ή το R^2 π.χ. σε ένα σύνολο ελέγχου είναι ένα έγκυρο μέτρο αξιολόγησης της προσαρμογής του μοντέλου. Γενικά, αν αυξάνουμε τη διάσταση του μοντέλου, προσθέτουμε μεταβλητές που όντως σχετίζονται με την *Y*, βελτιώνουμε το προσαρμοσμένο μοντέλο με την έννοια ότι θα μειωθεί το σφάλμα στα δεδομένα ελέγχου (*test error*). Στην περίπτωση όμως που οι μεταβλητές που προστίθενται στο μοντέλο είναι απλά θόρυβος, δεν έχουν καμία σχέση με την *Y*, τότε το ρίσκο για υπερπροσαρμογή αυξάνεται, καθώς σε αυτές τις μεταβλητές μπορεί να ανατεθούν μη μηδενικοί συντελεστές. Έτσι, το σφάλμα στα δεδομένα ελέγχου τείνει να αυξάνεται όσο αυξάνονται και οι διαστάσεις του προβλήματος. Το φαινόμενο αυτό αποκαλείται «κατάρα της διάστασης δεδομένων» (*curse of dimensionality*). Οπότε βλέπουμε ότι είναι σημαντικό να εντοπίσουμε τις μεταβλητές που έχουν μεγαλύτερη επιρροή στην *Y*. Αυτό μπορεί να επιτευχθεί μέσω τεχνικών μείωσης διαστάσεων (*dimensionality reduction techniques*), όπως είναι η ποινικοποίηση και η τεχνική PCA.

Επίσης, οι διάφορες στατιστικές διαδικασίες συχνά περιλαμβάνουν και κάποιου είδους αριθμητική βελτιστοποίηση (*numerical optimization*). Βελτιστοποιήσεις σε δεδομένα υψηλών διαστάσεων είναι όχι μόνο υπολογιστικά πολύ ακριβές, αλλά αργούν και να συγκλίνουν. Οι αλγόριθμοι επίσης εύκολα μπορούν να κολλήσουν σε κάποιο τοπικό ελάχιστο (*local minima*), δημιουργώντας έτσι μια αριθμητική αστάθεια. Προβλήματα αστάθειας προκύπτουν και λόγω του ότι συχνά οι αλγόριθμοι χρησιμοποιούν επαναληπτικά αντιστροφές μεγάλων πινάκων.

Ένα άλλο συχνό και σημαντικό πρόβλημα είναι αυτό της ψευδούς συσχέτισης (*spurious correlation*). Ψευδής συσχέτιση είναι όταν δύο μεταβλητές δεν έχουν καμία απολύτως πληθυσμιακή συσχέτιση, αλλά έχουν υψηλή δειγματική συσχέτιση. Η ψευδής συσχέτιση δεν είναι σημαντικό πρόβλημα στα δεδομένα χαμηλών διαστάσεων (*low-dimensional data*). Ύπαρξη ψευδούς συσχέτισης στα δεδομένα είναι αρκετά πιθανό να οδηγήσει σε λανθασμένες επιστημονικές ανακαλύψεις και ψευδή στατιστικά συμπεράσματα.

Τέλος, από τα πιο συνηθισμένα προβλήματα που μπορούμε να αντιμετωπίσουμε κατά την επιλογή μοντέλου υψηλής διάστασης είναι αυτό της πολυσυγγραμμικότητας και αυτό της ύπαρξης θορύβου. Για αυτό, η επιλογή μεταβλητών είναι θεμελιωδώς απαραίτητη.

3.5.3 Η έννοια της σποραδικότητας

Τι είναι όμως αυτό που μας επιτρέπει να προχωρήσουμε σε στατιστικά συμπεράσματα σε δεδομένα υψηλών διαστάσεων; Είναι η υπόθεση ότι το d-διάστατο διάνυσμα των παραμέτρων έχει αρκετές μηδενικές συνιστώσες, δηλαδή θεωρείται αραιό (*sparse*). Η επιλογή μεταβλητών βάσει της υπόθεσης της σποραδικότητας (*sparsity*) μπορεί να καταλήξει σε ένα υποσύνολο τελικά σημαντικών μεταβλητών, βελτιώνοντας έτσι την ερμηνευσιμότητα του μοντέλου. Ένα αραιό μοντέλο είναι αρκετά πιο εύκολο να εκτιμηθεί και να ερμηνευθεί σε σχέση με ένα πυκνό μοντέλο. Σημαντικό είναι να επισημάνουμε ότι αν το πραγματικό μοντέλο είναι αραιό έτσι ώστε οι k μεταβλητές, όπου $k < n < d$, να είναι μη μηδενικές, ακόμα και αν δεν γνωρίζουμε ποιες είναι αυτές οι k μεταβλητές, μπορούμε να τις εκτιμήσουμε με διάφορες μεθόδους αρκετά αποτελεσματικά. Σαφώς, αν γνωρίζαμε ποιες ήταν, τα πράγματα θα ήταν ακόμα πιο εύκολα, αλλά αυτό δεν αποτελεί εμπόδιο στην ανάλυσή μας. Επίσης, είναι προφανές ότι μεγάλη σποραδικότητα βοηθάει και στη μείωση του υπολογιστικού κόστους (Fan and Lv, 2010).

3.5.4 Σκοπός στα δεδομένα υψηλών διαστάσεων

Όπως είδαμε, υπάρχουν αρκετά προβλήματα που μπορεί να προκύψουν στα δεδομένα υψηλών διαστάσεων. Δεδομένα υψηλών διαστάσεων προκύπτουν από διάφορους επιστημονικούς τομείς και κάθε τομέας έχει διαφορετικούς επιστημονικούς στόχους. Οι βασικοί στόχοι που προκύπτουν από τη συμπερασματολόγια σε υψηλές διαστάσεις είναι οι παρακάτω (Fan et al., 2020):

1) Να κατασκευάσουμε μια μέθοδο όσο το δυνατόν πιο αποτελεσματική για να προβλέπει μελλοντικές παρατηρήσεις.
2) Να αποκτήσουμε γνώση σχετικά με τη σχέση μεταξύ των ανεξάρτητων μεταβλητών και της Y για επιστημονικούς σκοπούς, ελπίζοντας να κατασκευάσουμε μια βελτιωμένη μέθοδο για πρόβλεψη.

Ο πρώτος στόχος εμφανίζεται σε προβλήματα όπως είναι η βελτιστοποίηση χαρτοφυλακίου (*portfolio optimization*) και η ταξινόμηση εγγράφων (*document classification*), όπου δε μας ενδιαφέρει τόσο να καταλάβουμε ποιες μεταβλητές είναι σημαντικές στο πρόβλημα, αλλά περισσότερο η επίδοση της διαδικασίας. Ο δεύτερος στόχος εμφανίζεται σε επιστημονικές έρευνες και μελέτες, όπως π.χ. οι γονιδιωματικές μελέτες όπου οι ερευνητές θέλουν να βρουν ποια γονίδια οφείλονται για συγκεκριμένες ασθένειες.

3.6 Ποινικοποίηση

3.6.1 Καλά τοποθετημένο πρόβλημα και αντίστροφο πρόβλημα

Στα Μαθηματικά, ένα καλά τοποθετημένο πρόβλημα (*well-posed problem*) καλείται το πρόβλημα που ικανοποιεί ταυτόχρονα τις τρεις παρακάτω ιδιότητες:

1) Το πρόβλημα έχει λύση.
2) Η λύση είναι μοναδική.
3) Η συμπεριφορά της λύσης αλλάζει συνεχώς σύμφωνα με τις αρχικές συνθήκες.

Αν έστω και μία από αυτές τις ιδιότητες δεν ικανοποιείται, τότε αναφερόμαστε σε ένα όχι καλά τοποθετημένο πρόβλημα (*ill-posed problem*).

Αντίστροφο πρόβλημα (*inverse problem*) είναι το πρόβλημα στο οποίο θέλουμε να υπολογίσουμε τους παράγοντες που παραγάγουν ένα σύνολο παρατηρήσεων. Σε ένα αντίστροφο πρόβλημα, ο ερευνητής μελετά τα τελικά αποτελέσματα και σύμφωνα με αυτά, θέλει να υπολογίσει τις αιτίες από τις οποίες προέκυψαν. Συχνά, τα αντίστροφα προβλήματα είναι και ill-posed.

3.6.2 Η έννοια της ποινικοποίησης

Η ποινικοποίηση (*regularization or penalization*) είναι μια ιδέα που αρχικά προτάθηκε από τον Andrey Tikhonov, Ρώσο μαθηματικό του $20^{ου}$ αιώνα, για την επίλυση όχι καλά τοποθετημένων αντιστρόφων προβλημάτων. Οι διάφορες μέθοδοι ποινικοποίησης που έχουν αναπτυχθεί, χρησιμοποιούνται για την εκτίμηση των συντελεστών υπό την υπόθεση ότι το πραγματικό μοντέλο είναι αραιό, οι περισσότερες επεξηγηματικές μεταβλητές αναμένεται να μην έχουν καμία απολύτως επίδραση στην *Y*, ή πρακτικά ασήμαντη επίδραση. Οι μέθοδοι ποινικοποίησης μπορεί να είναι καλύτερες σε τέτοιες περιπτώσεις σποραδικότητας και να δίνουν πιο κατάλληλες εκτιμήσεις, σε σχέση με τις κλασσικές εκτιμήσεις μέγιστης πιθανοφάνειας. Επίσης, στην περίπτωση των δεδομένων υψηλών διαστάσεων οι εκτιμητές μέγιστης πιθανοφάνειας μπορεί να είναι αρκετά ασταθείς ή ακόμα και να μην υπάρχουν. Μέσω της ποινικοποίησης, ακόμα και αν έχουμε δεδομένα υψηλών διαστάσεων, μπορούμε να καταλήξουμε σε ένα απλοποιημένο μοντέλο, εφ'όσον όμως υπάρχει σποραδικότητα.

Αυτό που κάνει η ποινικοποίηση είναι να τροποποιεί τη συνάρτηση βελτιστοποίησης (*optimization function*) ενός μοντέλου προσθέτοντας έναν παραμετροποιήμενο όρο, ή αλλιώς ποινή (*penalty*). Κάτι το οποίο δεν ισχύει μόνο για τα μοντέλα παλινδρόμησης, αλλά για οποιαδήποτε συνάρτηση βελτιστοποίησης ενός μοντέλου, όπως είναι τα βαθιά νευρωνικά δίκτυα (*deep neural networks*). Ανάλογα με τη μαθηματική διατύπωση της ποινής, θα έχουμε και διαφορετικά μοντέλα. Οι διάφορες μέθοδοι ποινικοποίησης λειτουργούν με τη χρήση της συρρίκνωσης (*shrinkage*), οι εκτιμήσεις των συντελεστών δηλαδή συρρικνώνονται προς το 0 ή και ακόμα μηδενίζονται εντελώς. Στην περίπτωση που κάποιοι συντελεστές μηδενιστούν, τότε έχουμε και επιλογή μεταβλητών. Μέσω της συρρίκνωσης, οι προβλεπτικές ικανότητες του μοντέλου θα βελτιωθούν και το ρίσκο για υπερπροσαρμογή θα μειωθεί. Επίσης, η συρρίκνωση των συντελεστών εκτελείται αυτόματα από το μοντέλο και δε χρειάζεται

ανθρώπινη παρέμβαση, κάτι που είναι πολύ επιθυμητό χαρακτηριστικό και επεξηγεί κομμάτι της επιτυχίας των συναρτήσεων ποινής (Emmert-Streib et al., 2023).

Κρατώντας ένα υποσύνολο των μεταβλητών, έχουμε ένα μοντέλο ερμηνεύσιμο το οποίο πιθανόν να έχει χαμηλότερο σφάλμα πρόβλεψης από το πλήρες μοντέλο. Επειδή όμως η επιλογή υπομοντέλου είναι μια διακριτή διαδικασία, οι μεταβλητές είτε παραμένουν στο μοντέλο ή αφαιρούνται, συχνά παρουσιάζεται υψηλή διακύμανση και έτσι το σφάλμα πρόβλεψης δε μειώνεται. Επίσης, μικρές αλλαγές στα δεδομένα μπορεί να δώσουν πολύ διαφορετικά μοντέλα, κάτι το οποίο μειώνει την ακρίβεια της πρόβλεψης. Οι μέθοδοι συρρίκνωσης είναι πιο συνεχείς διαδικασίες και δεν παρουσιάζουν τόσο υψηλή μεταβλητότητα. Έτσι, μπορούμε να προσαρμόσουμε το πλήρες μοντέλο και να χρησιμοποιήσουμε διάφορες τεχνικές που ποινικοποιούν τις εκτιμήσεις των παραμέτρων, καταφέρνοντας έτσι να μειώσουμε τη διακύμανσή τους. Οι πιο γνωστές μέθοδοι συρρίκνωσης, τις οποίες και θα αναλύσουμε παρακάτω, είναι η παλινδρόμηση ridge και η παλινδρόμηση lasso. Επίσης, θα γίνει αναφορά στις μεθόδους elastic net, SCAD και MCP (βλ. 3.7).

Προτού προχωρήσουμε σε οποιουδήποτε είδους ποινικοποίηση, θα προβούμε σε τυποποίηση (*standardization or Z-score normalization*) των δεδομένων, ώστε η εκάστοτε συνάρτηση ποινής (*penalty function*) να θέτει την ίδια ποινή σε κάθε μεταβλητή, μη επηρεασμένη από την αρχική κλίμακα (*scaling*) των μεταβλητών.

Κλείνοντας, κατά την εφαρμογή της ποινικοποίησης θα χρειαστεί να επιλύσουμε κάποια προβλήματα βελτιστοποίησης, τα οποία διατυπώνονται με βάση κάποιες νόρμες (*norms*). Για το πραγματικό διάνυσμα $\boldsymbol{x} \in \mathbb{R}^n$ και $q \geq 1$, η L_q-νόρμα ορίζεται ως:

$$\|\boldsymbol{x}\|_q = \left(\sum_{i=1}^{n} |x_i|^q \right)^{1/q} . \tag{3.1}$$

Για $q < 1$ η σχέση (3.1) μπορεί και πάλι να οριστεί, αλλά δεν αποτελεί πλέον νόρμα με τη μαθηματική έννοια.

3.6.3 Τα κριτήρια AIC και BIC

Δύο από τις πιο γνωστές μεθόδους ποινικοποίησης κατά την επιλογή μοντέλου, είναι το AIC και το BIC. Η γενική ιδέα αυτών των δύο κριτηρίων είναι να ποινικοποιούν μεγαλύτερα μοντέλα. Το κριτήριο πληροφορίας του Akaike (*Akaike Information*

Criterion, AIC) προτάθηκε από τον Ιάπωνα στατιστικό Hirotugu Akaike και είναι ίσο με:

$$AIC = -2(l_M) + 2k,$$

όπου l_M ο μεγιστοποιημένος λογάριθμος της πιθανοφάνειας του μοντέλου M και k ο αριθμός των μεταβλητών στο μοντέλο M. Το καλύτερο μοντέλο είναι αυτό που ελαχιστοποιεί το AIC. Το AIC θέτει ποινή στο μοντέλο ίση με $2k$ για την ύπαρξη πολλών παραμέτρων. Αν και ένα μοντέλο με λιγότερες μεταβλητές έχει μικρότερο λογάριθμο πιθανοφάνειας, και άρα μεγαλύτερη απόκλιση, από ένα πιο πολύπλοκο μοντέλο, το απλούστερο μοντέλο μπορεί να δίνει καλύτερες εκτιμήσεις για τις πραγματικές αναμενόμενες τιμές. Ένα βασικό πλεονέκτημα του AIC είναι ότι μπορεί να χρησιμοποιηθεί και για σύγκριση μεταξύ μη εμφωλευμένων μοντέλων.

Ένα άλλο κριτήριο επιλογής μοντέλου είναι το μπεϋζιανό κριτήριο πληροφορίας (*Bayes Information Criterion, BIC*), ή λιγότερο γνωστό ως κριτήριο πληροφορίας του Schwarz (Schwarz *Information Criterion, SIC),* το οποίο προτάθηκε από τον Gideon E. Schwarz. Υπολογίζεται ως:

$$BIC = -2(l_M) + log(n)k,$$

όπου $log(n)$ ο λογάριθμος για το μέγεθος του δείγματος n. Η ποινή που θέτει το BIC είναι ίση με $log(n)k$. Οπότε, όσο το n μεγαλώνει, το BIC ποινικοποιεί πιο αυστηρά την ύπαρξη πολλών μεταβλητών. Το BIC είναι αρκετά χρήσιμο σε αυτές τις περιπτώσεις όπου το n είναι αρκετά μεγάλο, γιατί ναι μεν μπορεί να υπάρχουν πολλές μεταβλητές που είναι στατιστικά σημαντικές, αλλά να μην είναι πρακτικά σημαντικές (Agresti, 2019). Για δείγμα μεγέθους $n = 8$ έχουμε $log(n) > 2$, οπότε μπορούμε να δούμε ότι το BIC ποινικοποιεί πιο αυστηρά σε σχέση με το AIC και για αυτόν το λόγο, το BIC δίνει πιο φειδωλά μοντέλα από ότι το AIC.

Περιληπτικά για τα δύο κριτήρια έχουμε τα εξής (Emmert-Streib et al., 2023):

1) Το AIC επιλέγει λιγότερο φειδωλά μοντέλα σε σχέση με το BIC και για αυτό είναι πιο επιρρεπές στην υπερπροσαρμογή. Από την άλλη, επειδή το BIC επιλέγει πιο φειδωλά μοντέλα σε σχέση με το AIC, είναι πιο επιρρεπές στην υποπροσαρμογή.
2) Το AIC είναι ασυμπτωτικά αποτελεσματικό, άλλα όχι συνεπές. Το BIC είναι συνεπές, άλλα όχι ασυμπτωτικά αποτελεσματικό.

3) Το AIC θα πρέπει να χρησιμοποιείται όταν σκοπός μας είναι η προβλεπτική ακρίβεια του μοντέλου, ενώ το BIC όταν σκοπός είναι η ερμηνευσιμότητα του μοντέλου.
4) Το AIC αντιπροσωπεύει μια οπτική κλασσικής στατιστικής συμπερασματολογίας (*frequentist point of view*), ενώ το BIC μια μπεϋζιανή οπτική (*Bayesian point of view*).

3.6.4 Ποινικοποίηση στη λογιστική παλινδρόμηση

Ας δούμε τη γενική ιδέα για τη χρήση ποινικοποίησης στην περίπτωση της λογιστικής παλινδρόμησης. Έστω το λογιστικό μοντέλο Μ με ανεξάρτητες μεταβλητές $X_1, X_2, \dots, X_d$:

$$logit(p) = log\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{j=1}^{d} \beta_j X_j,$$

με συνάρτηση πιθανοφάνειας που υπολογίζεται ως:

$$L = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}.$$

Ακόμα, θέτουμε ως l το λογάριθμο της πιθανοφάνειας (σχέση (2.5)):

$$l = \sum_{i=1}^{n} y_i \, log\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n} log(1-p_i).$$

Εκφράζοντας το λογάριθμο της πιθανοφάνειας συναρτήσει των $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_d)$ παίρνουμε:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=0}^{d} y_i \beta_j x_{ij} - \sum_{i=1}^{n} log\left(1 + \exp\left(\sum_{j=0}^{d} \beta_j x_{ij}\right)\right).$$

(Αναλυτικά βλ. 2.4.5. Στα παραπάνω έχουμε ορίσει το $\beta_0 + \sum_{j=1}^{d} \beta_j x_j$ ως $\sum_{j=0}^{d} \beta_j x_{ij}$ με $x_{0j} = 1$). Επίσης, ορίζουμε ως $-l(\boldsymbol{\beta})$ τη λογιστική συνάρτηση απώλειας (*logistic loss function*).

Οι διάφορες ποινικοποιημένες μέθοδοι πιθανοφάνειας (*penalized likelihood methods*) προσθέτουν μία ποινή στη συνάρτηση του λογαρίθμου της πιθανοφάνειας, έτσι ώστε οι τιμές που την μεγιστοποιούν είναι εξομαλύνσεις (*smoothings*) των κλασσικών ΕΜΠ, συρρικνώνοντας τους προς το 0. Ανάμεσα στα θετικά των ποινικοποιημένων μεθόδων πιθανοφάνειας είναι ότι οι εκτιμητές είναι λογικοί ακόμα

και όταν δεν υπάρχουν οι ΕΜΠ ή επηρεάζονται αρκετά από την ύπαρξη πολυσυγγραμμικότητας. (Agresti and Kateri, 2021). Γενικά, οι ΕΜΠ στη λογιστική παλινδρόμηση είναι μεροληπτικοί μακριά από το 0. Η χρήση ποινικοποίησης μειώνει τη μεροληψία αυτή. Για αυτό οι μέθοδοι αυτές είναι αρκετά αποτελεσματικές στη λογιστική παλινδρόμηση.

Κατά την εφαρμογή ποινικοποίησης στη λογιστική παλινδρόμηση θέλουμε να μεγιστοποιήσουμε την ποσότητα:

$$l^*(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - s(\boldsymbol{\beta}), \tag{3.2}$$

όπου s μια κατάλληλη συνάρτηση ποινής. Εναλλακτικά, μπορούμε να ελαχιστοποιήσουμε την:

$$l^*(\boldsymbol{\beta}) = -l(\boldsymbol{\beta}) + s(\boldsymbol{\beta}). \tag{3.3}$$

3.7 Συναρτήσεις Ποινής

3.7.1 Μέθοδος ridge

Έχει δειχθεί ότι στη γραμμική παλινδρόμηση στην περίπτωση όπου τα διανύσματα των ανεξάρτητων μεταβλητών δεν είναι ορθογώνια, οι εκτιμήσεις των παραμέτρων που βασίζονται στην ελαχιστοποίηση του RSS έχουν μεγάλη πιθανότητα να μην είναι ικανοποιητικές, αν όχι λανθασμένες. Οι Hoerl and Kennard (1970) πρότειναν μία διαδικασία εκτίμησης κατά την ύπαρξη πολυσυγγραμμικότητας μεταξύ των ανεξάρτητων μεταβλητών, η οποία βασίζεται στην προσθήκη μικρών θετικών ποσοτήτων στη διαγώνιο του $\boldsymbol{X}^T\boldsymbol{X}$. Η πολυσυγγραμμικότητα μπορεί να δημιουργήσει αρκετά προβλήματα κατά την ανάλυση ενός μοντέλου, όπως να οδηγήσει σε ανακριβείς εκτιμήσεις για τους συντελεστές παλινδρόμησης, να διογκώσει τα τυπικά σφάλματα των συντελεστών, να οδηγήσει σε ψευδή και μη σημαντικά p-values και τελικώς να υποβαθμίσει την προβλεπτική ικανότητα του μοντέλου (Saleh et al., 2019).

Τη μέθοδο αυτή την αποκάλεσαν ridge, η οποία είναι γνωστή και ως L_2 ποινικοποίηση (L_2 *regularization*). Η συνάρτηση εξομάλυνσης (*smoothing function)*, ή αλλιώς ποινή συρρίκνωσης (*shrinkage penalty*), που χρησιμοποιεί η μέθοδος ridge είναι η:

$$s(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{d} \beta_j^2,$$

με $\lambda \geq 0$ ρυθμιστική παράμετρο (*tuning parameter*), η οποία ελέγχει το συνολικό βαθμό ποινικοποίησης. Αν εφαρμόσουμε τη ridge στα μοντέλα της λογιστικής παλινδρόμησης, τότε θέλουμε να βρούμε το $\widehat{\boldsymbol{\beta}}^{ridge}$ που ελαχιστοποιεί την ποσότητα (3.3):

$$\widehat{\boldsymbol{\beta}}^{ridge} = \underset{\boldsymbol{\beta}}{argmin}\left[-l(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{d} \beta_j^2\right], \tag{3.4}$$

ή ισοδύναμα:

$$\widehat{\boldsymbol{\beta}}^{ridge} = \underset{\boldsymbol{\beta}}{argmin}[-l(\boldsymbol{\beta})] \text{ ,υπό τον περιορισμό } \sum_{j=1}^{d} \beta_j^2 \leq t.$$

Η παράμετρος λ είναι αυτή που καθορίζει το μέγεθος της συρρίκνωσης που θα εφαρμοστεί στους συντελεστές των μεταβλητών. Διαφορετικές επιλογές του λ, θα δώσουν και διαφορετικές εκτιμήσεις. Είναι εύκολο να δούμε ότι για $\lambda = 0$ δεν υπάρχει ποινικοποίηση και άρα έχουμε τους κλασσικούς εκτιμητές μέγιστης πιθανοφάνειας. Στην περίπτωση που $\lambda \to \infty$, η επίδραση της ποινής μεγαλώνει και οι εκτιμήσεις των συντελεστών θα πλησιάσουν στο 0. Ένα άλλο χαρακτηριστικό είναι ότι η επιλογή του λ αντικατοπτρίζει την αντιστάθμιση μεροληψίας-διακύμανσης, καθώς όσο μεγαλώνει το λ, η διακύμανση μειώνεται και η μεροληψία αυξάνεται. Σχετικά με την παράμετρο *t*, όσο αυτή μειώνεται, τόσο μεγαλύτερη και η συρρίκνωση που θα επιβληθεί στους συντελεστές. Επίσης, για κάθε λ υπάρχει και ένα *t* το οποίο θα δώσει ακριβώς τις ίδιες εκτιμήσεις των συντελεστών, υπάρχει δηλαδή μια μία-προς-μία αντιστοιχία για τις παραμέτρους λ και *t*. Υπάρχουν διάφοροι τρόποι για την επιλογή του λ. Ο Tibshirani (1996) περιγράφει τρεις τρόπους για την επιλογή της ρυθμιστικής παραμέτρου· τη διασταυρούμενη επικύρωση (*cross-validation*), τη γενικευμένη διασταυρούμενη επικύρωση (*generalized cross-validation*) και μία αναλυτική αμερόληπτη εκτίμηση του ρίσκου (*analytical unbiased estimate of risk*), ενώ οι Fan and Li (2001) στην εργασία τους χρησιμοποιούν τους δύο πρώτους. Στην παρούσα εργασία εμείς θα επικεντρωθούμε στην επιλογή μέσω διασταυρούμενης επικύρωσης k-πλαισίων (*k-fold cross-validation*) για $k = 10$. Αφού επιλέξουμε ένα εύρος τιμών για το λ, θα υπολογίσουμε την απόκλιση (αφού αναφερόμαστε σε λογιστική παλινδρόμηση) διασταυρούμενης επικύρωσης για κάθε τιμή του. Για τη συνέχεια της έρευνας θα επιλέξουμε το λ το οποίο επιτυγχάνει την ελάχιστη απόκλιση.

Οι Cessie and Houwelingen (1992) ακολούθησαν την προσέγγιση των Duffy and Santner (1989) ώστε να επεκτείνουν τη θεωρία της παλινδρόμησης ridge από την κλασσική γραμμική παλινδρόμηση, στη λογιστική. Οι εκτιμητές $\widehat{\boldsymbol{\beta}}^{ridge}$ προκύπτουν ως περιορισμένοι εκτιμητές μέγιστης πιθανοφάνειας. Έστω Y_i με $i = 1,2, \dots, n$ ανεξάρτητες μεταξύ τους δίτιμες μεταβλητές, X_i τα d-διάστατα διανύσματα γραμμής των ανεξάρτητων μεταβλητών και $p(X_i)$ η πιθανότητα $Y_i = 1$ δοθέντων των τιμών των X_i. Επίσης, θα θεωρήσουμε ότι ο σταθερός όρος δεν περιλαμβάνεται στο πρόβλημα παλινδρόμησης και επομένως το μοντέλο της λογιστικής παλινδρόμησης γράφεται ως:

$$p(X_i) = \frac{e^{X_i\boldsymbol{\beta}}}{1 + e^{X_i\boldsymbol{\beta}}},$$

με $\boldsymbol{\beta}$ το d-διάστατο διάνυσμα των παραμέτρων. Έπειτα, θα εκφράσουμε το λογάριθμο της πιθανοφάνειας του μοντέλου αυτού ως:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \big[Y_i log\big(p(X_i)\big) + (1 - Y_i) log\big(1 - p(X_i)\big)\big].$$

Όπως ήδη γνωρίζουμε, μεγιστοποίηση της σχέσης αυτής θα δώσει τους εκτιμητές μέγιστης πιθανοφάνειας για το $\boldsymbol{\beta}$. Οι Duffy and Santner (1989) μεγιστοποιούν τη συνάρτηση του λογαρίθμου της πιθανοφάνειας με ποινή $\lambda\|\boldsymbol{\beta}\|^2$:

$$l_\lambda(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \lambda\|\boldsymbol{\beta}\|^2, \tag{3.5}$$

με $\|\boldsymbol{\beta}\| = \left(\sum \beta_j^2\right)^{1/2}$. Το διάνυσμα που μεγιστοποιεί την (3.5) ορίζεται ως $\widehat{\boldsymbol{\beta}}_\lambda$. Συρρικνώνοντας τα $\boldsymbol{\beta}$ προς το 0 και ταυτόχρονα επιτρέποντας μια μικρή μεροληψία, θα υπάρξει μια σταθεροποίηση στο σύστημα και θα δώσει εκτιμήσεις με μικρότερη διακύμανση. Το $\widehat{\boldsymbol{\beta}}_\lambda$, για καλή επιλογή τιμής του λ, αναμένεται να είναι κατά μέσο όρο πιο κοντά στις πραγματικές τιμές του $\boldsymbol{\beta}$ από τους κλασσικούς εκτιμητές μέγιστης πιθανοφάνειας $\widehat{\boldsymbol{\beta}}$, δηλαδή $MSE\big(\widehat{\boldsymbol{\beta}}_\lambda\big) < MSE\big(\widehat{\boldsymbol{\beta}}\big)$. Το $\widehat{\boldsymbol{\beta}}_\lambda$, όπως και τους κλασσικούς εκτιμητές μέγιστης πιθανοφάνειας, μπορούμε να το υπολογίσουμε μέσω της διαδικασίας μεγιστοποίησης Newton-Raphson. Η πρώτη παράγωγος του $l_\lambda(\boldsymbol{\beta})$ είναι:

$$\boldsymbol{U}_\lambda(\boldsymbol{\beta}) = \sum_{i=1}^{n} X_i^T[Y_i - p(X_i)] - 2\lambda\boldsymbol{\beta} = \boldsymbol{U}(\boldsymbol{\beta}) - 2\lambda\boldsymbol{\beta}, \tag{3.6}$$

με $\boldsymbol{U}(\boldsymbol{\beta})$ την παράγωγο του λογαρίθμου της πιθανοφάνειας χωρίς περιορισμούς. Επίσης, το αρνητικό του πίνακα των δεύτερων παραγώγων δίνεται ως:

$$\boldsymbol{\Omega}_\lambda(\boldsymbol{\beta}) = \boldsymbol{\Omega}(\boldsymbol{\beta}) + 2\lambda\boldsymbol{I}, \tag{3.7}$$

όπου $\boldsymbol{\Omega} = \boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{\beta})\boldsymbol{X}$ είναι το αρνητικό του πίνακα των δεύτερων παραγώγων της πιθανοφάνειας χωρίς περιορισμούς και $\boldsymbol{V}(\boldsymbol{\beta})$ ένας n × n διαγώνιος πίνακας με στοιχεία στη διαγώνιο $v_{ii} = p(X_i)\{1 - p(X_i)\}$. Προχωρώντας σε επέκταση της σειράς Taylor της πρώτης παραγώγου της ποινικοποιημένης πιθανοφάνειας για τις πραγματικές τιμές των παραμέτρων, έστω $\boldsymbol{\beta}_0$, μπορούμε να έχουμε ιδιότητες μεγάλου δείγματος των περιορισμένων εκτιμητών μέγιστης πιθανοφάνειας. Δηλαδή:

$$\boldsymbol{U}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda) = \boldsymbol{U}_\lambda(\boldsymbol{\beta}_0) - (\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_0)^T\boldsymbol{\Omega}_\lambda(\boldsymbol{\beta}_0) + o(\|\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_0\|).$$

Χρησιμοποιώντας τις σχέσεις (3.6), (3.7) και θέτοντας $\boldsymbol{U}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda) = 0$ θα έχουμε μια προσέγγιση πρώτης-τάξης για το $\widehat{\boldsymbol{\beta}}_\lambda$:

$$\widehat{\boldsymbol{\beta}}_\lambda = \boldsymbol{\beta}_o + \{\boldsymbol{\Omega}(\boldsymbol{\beta}_0) + 2\lambda\boldsymbol{I}\}^{-1}\{\boldsymbol{U}(\boldsymbol{\beta}_0) - 2\lambda\boldsymbol{\beta}_0\} \Leftrightarrow$$

$$\widehat{\boldsymbol{\beta}}_\lambda = \{\boldsymbol{\Omega}(\boldsymbol{\beta}_0) + 2\lambda\boldsymbol{I}\}^{-1}\{\boldsymbol{U}(\boldsymbol{\beta}_0) + \boldsymbol{\Omega}(\boldsymbol{\beta}_0)\boldsymbol{\beta}_0\}.$$

Οι Cessie and Houwelingen αναφέρουν ότι μια πρώτης τάξης εκτίμηση για τους κλασσικούς εκτιμητές μέγιστης πιθανοφάνειας είναι η $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + \boldsymbol{\Omega}^{-1}(\boldsymbol{\beta}_0)\boldsymbol{U}(\boldsymbol{\beta}_0)$ και έτσι μια πρώτης τάξης εκτίμηση για το $\widehat{\boldsymbol{\beta}}_\lambda$ είναι:

$$\widehat{\boldsymbol{\beta}}_\lambda = \{\boldsymbol{\Omega}(\boldsymbol{\beta}_0) + 2\lambda\boldsymbol{I}\}^{-1}\boldsymbol{\Omega}(\boldsymbol{\beta}_0)\widehat{\boldsymbol{\beta}},$$

όπου αν αντικαταστήσουμε το $\boldsymbol{\Omega}(\boldsymbol{\beta}_0)$ με $\boldsymbol{\Omega}(\widehat{\boldsymbol{\beta}})$ θα έχουμε τους εκτιμητές που όρισαν οι Schaefer et al. (1984). Οι οποίοι εκτιμητές όμως των Schaefer et al. δεν ορίζονται στην περίπτωση όπου κάποιοι από τους κλασσικούς ΕΜΠ είναι άπειροι. Υπό ορισμένες προϋποθέσεις κανονικότητας, το $\widehat{\boldsymbol{\beta}}$ είναι ασυμπτωτικά αμερόληπτο με πίνακα συνδιακύμανσης $\boldsymbol{\Omega}(\boldsymbol{\beta}_0)^{-1}$. Τότε, η ασυμπωτική μεροληψία του $\widehat{\boldsymbol{\beta}}_\lambda$ γίνεται $E(\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_0) = -2\lambda\{\boldsymbol{\Omega}(\boldsymbol{\beta}_0) + 2\lambda\boldsymbol{I}\}^{-1}\boldsymbol{\beta}_0$ και η ασυμπτωτική διακύμανση $\{\boldsymbol{\Omega}(\boldsymbol{\beta}_0) + 2\lambda\boldsymbol{I}\}^{-1}\boldsymbol{\Omega}(\boldsymbol{\beta}_0)\{\boldsymbol{\Omega}(\boldsymbol{\beta}_0) + 2\lambda\boldsymbol{I}\}^{-1}$. Η προσέγγιση αυτή ως προς τη διακύμανση δεν μπορεί να χρησιμοποιηθεί απευθείας για να κατασκευάσουμε κατά προσέγγιση διαστήματα εμπιστοσύνης για το $\widehat{\boldsymbol{\beta}}_\lambda$, καθώς θα πρέπει να λάβουμε υπ'όψιν τη μεροληψία της εκτίμησης. Μέθοδοι όπως οι Jackknife και bootstrap μπορούν να βοηθήσουν ώστε να λάβουμε περισσότερη πληροφόρηση για τη μεταβλητότητα του $\widehat{\boldsymbol{\beta}}_\lambda$.

Το πλεονέκτημα της ridge έναντι της μέγιστης πιθανοφάνειας, ή στην περίπτωση της γραμμικής παλινδρόμησης έναντι των ελαχίστων τετραγώνων, έχει να κάνει με την αντιστάθμιση μεροληψίας-διακύμανσης. Καθώς το λ αυξάνεται, η ευελιξία της προσαρμογής της ridge μειώνεται, με αποτέλεσμα να μειωθεί η διακύμανση και να αυξηθεί η μεροληψία James et al. (2023).

Η ridge αποτελεί μία αρκετά χρήσιμη μέθοδο ποινικοποίησης, ειδικά κατά την ύπαρξη πολυσυγγραμμικότητας μεταξύ των ανεξάρτητων μεταβλητών. Το βασικό μειονέκτημα της όμως είναι ότι ναι μεν θα συρρικνώσει τις εκτιμήσεις των μεταβλητών, όμως δε θα θέσει καμία εκτίμηση ακριβώς 0 (εκτός αν $\lambda = \infty$) (James et al, 2023). Έτσι, το τελικό μοντέλο που θα προέλθει μέσω της ridge θα εμπεριέχει όλες τις d ανεξάρτητες μεταβλητές, το οποίο μπορεί να μην είναι πρόβλημα για την ακρίβεια της πρόβλεψης, όμως είναι πρόβλημα κατά την ερμηνεία του μοντέλου, ειδικά σε περιπτώσεις που το d είναι αρκετά μεγάλο.

3.7.2 Μέθοδος lasso

Η μέθοδος ridge είναι μία χρήσιμη μέθοδος συρρίκνωσης, η οποία όμως δεν παύει να έχει κάποια μειονεκτήματα. Το βασικότερο εξ αυτών, όπως έχουμε ήδη αναφέρει, είναι το ότι δε θα θέσει ποτέ κάποια μεταβλητή ίση με το 0, δυσκολεύοντας αρκετά την ερμηνεία του μοντέλου, ειδικά στην περίπτωση που το πλήθος των ανεξάρτητων μεταβλητών d είναι μεγάλο. Ακόμα, η subset selection είδαμε ότι είναι μία μέθοδος κατά την οποία οι μεταβλητές είτε παραμένουν στο μοντέλο ή αφαιρούνται εντελώς, παρουσιάζοντας συχνά υψηλή διακύμανση μην καταφέρνοντας να μειώσει το προβλεπτικό σφάλμα. Η μέθοδος lasso (least absolute shrinkage and selection operator), ή αλλιώς L_1 ποινικοποίηση (L_1 *regularization*), προτάθηκε από τον Tibshirani (1996), ως λύση για τα μειονεκτήματά τους. Η ιδέα της lasso προήλθε από μία μέθοδο που είχε προτείνει ο Breiman, την non-negative garotte. Η lasso αποτελεί και αυτή μία μέθοδο συρρίκνωσης, η οποία όμως όχι μόνο συρρικνώνει τους συντελεστές προς το 0, αλλά μπορεί ακόμα και κάποια β_j να τα μηδενίσει πλήρως. Επομένως, εκτός από την εκτίμηση των παραμέτρων έχουμε και επιλογή μεταβλητών και κατά αυτόν τον τρόπο η ερμηνεία του μοντέλου γίνεται ευκολότερη, καθώς δεν περιέχει τόσες πολλές μεταβλητές παρά μόνο ένα υποσύνολο αυτών που έχουν ισχυρή επιρροή στην Y. Γενικά η lasso παράγει αραιά μοντέλα, μοντέλα που περιλαμβάνουν ένα υποσύνολο των $X_1, X_2, \dots, X_d$. Οπότε βλέπουμε ότι η lasso προσπαθεί να διατηρήσει τα θετικά χαρακτηριστικά της subset selection και της ridge.

Επίσης, η χρησιμότητα της lasso έγκειται στο γεγονός ότι εφαρμόζεται σε προβλήματα υψηλών διαστάσεων, κυρίως λόγω της επιλογής μεταβλητών και της στατιστικής ακρίβειας των προβλέψεων. Όταν ο αριθμός των ανεξάρτητων μεταβλητών d είναι αρκετά μεγάλος,

πρακτικά λίγα είναι αυτά τα $\beta_j, j = 1,2,\dots,d$ που είναι διάφορα του μηδενός. Ακόμα, για να εφαρμόσουμε τη λογιστική παλινδρόμηση στα δεδομένα υψηλών διαστάσεων, θα πρέπει να προβούμε σε κάποιες τροποποιήσεις, έτσι ώστε το μοντέλο να έχει πιο σταθερή εφαρμογή στα δεδομένα.

Η συνάρτηση ποινής της lasso είναι η:

$$s(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{d} |\beta_j|,$$

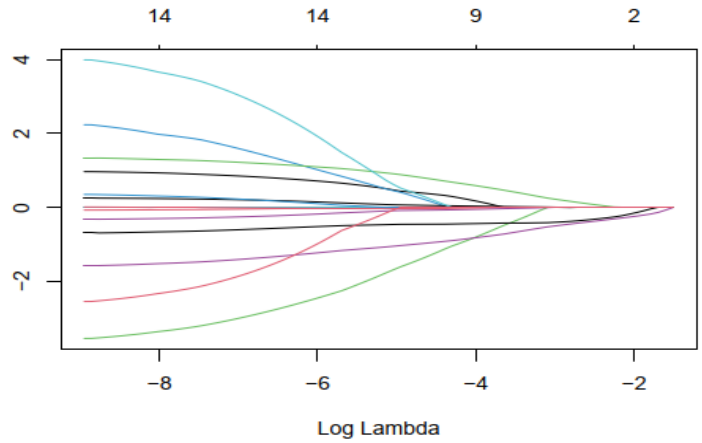
με $\lambda \geq 0$ ρυθμιστική παράμετρο. Συνεπώς, θέλουμε να βρούμε τους εκτιμητές lasso, έστω $\widehat{\boldsymbol{\beta}}^{lasso}$, οι οποίοι ελαχιστοποιούν την ποσότητα (3.3), δηλαδή:

$$\widehat{\boldsymbol{\beta}}^{lasso} = \underset{\boldsymbol{\beta}}{argmin} \left[-l(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{d} |\beta_j| \right], \tag{3.8}$$

ή ισοδύναμα:

$$\widehat{\boldsymbol{\beta}}^{lasso} = \underset{\boldsymbol{\beta}}{argmin}[-l(\boldsymbol{\beta})] \ , \text{υπό τον περιορισμό} \sum_{j=1}^{d} |\beta_j| \leq t,$$

με $t \geq 0$. Παρομοίως με τη ridge, η παράμετρος λ είναι αυτή που καθορίζει το μέγεθος της συρρίκνωσης. Για $\lambda = 0$, θα έχουμε τους κλασσικούς εκτιμητές μέγιστης πιθανοφάνειας, ενώ αν $\lambda \to \infty$, τότε θα ισχύει $\hat{\beta}_j = 0, \forall j = 1,2,\dots,d$. Επαρκώς μεγάλη τιμή του λ θα θέσει κάποιους από τους συντελεστές ίσους με το 0. Η επιλογή του λ θα γίνει μέσω διασταυρούμενης επικύρωσης. Στο παρακάτω γράφημα, μπορούμε να δούμε ένα παράδειγμα που αφορά το σύνολο δεδομένων Students (Agresti and Kateri, 2021) για τους εκτιμητές lasso και πώς αυτοί αλλάζουν, καθώς η ρυθμιστική παράμετρος λ αυξάνεται, σε λογαριθμική κλίμακα. Ή αλλιώς, το μονοπάτι της lasso (*lasso path*). Κάθε χρωματιστή γραμμή αφορά διαφορετικό $\hat{\beta}_j$.



ΣΧΗΜΑ 3-12 : Γραφική αναπαράσταση των εκτιμητών της lasso, καθώς το λ αυξάνεται.

Στη συνέχεια, θα πρέπει να επιλύσουμε τη σχέση (3.8) και να εκτιμήσουμε τα β_j κατά την εφαρμογή της lasso στη λογιστική παλινδρόμηση. Για αρχή, ας εκφράσουμε το μοντέλο της λογιστικής παλινδρόμησης κάπως διαφορετικά. Έστω η εξαρτημένη δίτιμη μεταβλητή $Y \in \{0,1\}$ και το λογιστικό μοντέλο:

$$log\frac{\Pr(Y=1|X=x)}{\Pr(Y=0|X=x)} = \beta_0 + \beta^T x,$$

όπου $X = (X_1, X_2, \dots, X_d)$ ένα διάνυσμα των ανεξάρτητων d μεταβλητών, $\beta_0 \in \mathbb{R}$ ο σταθερός όρος και $\beta \in \mathbb{R}^d$ το διάνυσμα των συντελεστών παλινδρόμησης. Ο αρνητικός λογάριθμος της πιθανοφάνειας με την L_1 ποινή μπορεί να πάρει τη μορφή (Hastie et al., 2015):

$$-\frac{1}{n}\sum_{i=1}^{n}\{y_i logPr(Y=1|x_i) + (1-y_i)logPr(Y=0|x_i))\} + \lambda\|\beta\|_1$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left\{y_i(\beta_0+\beta^T x_i) - log\,(1+e^{\beta_0+\beta^T x_i})\right\} + \lambda\|\beta\|_1 \qquad (3.9)$$

Η σχέση (3.9) είναι κυρτή (*convex*) και το μέρος της πιθανοφάνειας παραγωγίσιμο, οπότε μπορεί να βρεθεί λύση μέσω κυρτής βελτιστοποίησης (*convex optimization*). Αν οι τωρινές εκτιμήσεις είναι $(\tilde{\beta}_0, \tilde{\beta})$, σχηματίζουμε την τετραγωνική συνάρτηση (*quadratic function*):

$$Q(\beta_0,\beta) = \frac{1}{2n}\sum_{i=1}^{n} w_i(z_i - \beta_0 - \beta^T x_i)^2 + C(\tilde{\beta}_0,\tilde{\beta}),$$

όπου C μια σταθερά ανεξάρτητη από τα (β_0,β), $z_i = \tilde{\beta}_0 + \tilde{\beta}^T x_i + \frac{y_i - \hat{p}(x_i)}{\hat{p}(x_i)(1-\hat{p}(x_i))}$ με $\hat{p}(x_i)$ ο τωρινός εκτιμητής του $\Pr(Y=1|X=x_i)$ και $w_i = \hat{p}(x_i)\big(1-\hat{p}(x_i)\big)$. Για κάθε τιμή του

λ, δημιουργούμε έναν εξωτερικό βρόχο που υπολογίζει την τετραγωνική προσέγγιση Q για τις τωρινές παραμέτρους $(\tilde{\beta}_0, \tilde{\beta})$. Τέλος, χρησιμοποιώντας τον αλγόριθμο βελτιστοποίησης *coordinate descent* θα επιλύσουμε το πρόβλημα

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{d+1}} \{Q(\beta_0, \beta) + \lambda\|\beta\|_1\}.$$

Ο παραπάνω τρόπος δεν είναι ο μοναδικός για να υπολογίσουμε το path της lasso, καθώς υπάρχουν διάφοροι. Οι Park and Hastie (2007) πρότειναν έναν αλγόριθμο για να εκτιμήσουν το path της lasso στα ΓΓΜ. Ο αλγόριθμος αυτός χρησιμοποιεί τη μέθοδο πρόβλεψης-διόρθωσης (*predictor-corrector method*) για να προσδιορίσει ολόκληρο το path των εκτιμήσεων $\hat{\beta}_j$ για τις διάφορες τιμές του λ, να βρει δηλαδή $\{\hat{\beta}(\lambda): 0 < \lambda < \infty\}$. Ο αλγόριθμος αυτός, ξεκινώντας από το $\lambda = \lambda_{max}$, με λ_{max} το μεγαλύτερο λ που κάνει το $\hat{\beta}(\lambda)$ μη μηδενικό, υπολογίζει μία σειρά από σύνολα λύσεων, όπου κάθε φορά εκτιμάει τους συντελεστές για μικρότερο λ από την προηγούμενη εκτίμηση. Κάθε φορά η διαδικασία της βελτιστοποίησης αποτελείται από τα εξής τρία βήματα:

1) προσδιορίζει το μέγεθος του βήματος στο λ,
2) προβλέπει την αντίστοιχη αλλαγή στους συντελεστές και
3) διορθώνει το σφάλμα της προηγούμενης πρόβλεψης.

Αντί κάθε φορά να υπολογίζουμε τις λύσεις σε πολλές σταθερές τιμές του λ, δημιουργώντας ολόκληρο το μονοπάτι ποινικοποίησης (*regularization path*) προσδιορίζουμε τη σειρά με την οποία οι μεταβλητές εισέρχονται ή εξέρχονται στο μοντέλο. Κατά αυτόν τον τρόπο, μπορούμε να βρούμε την ποινικοποιημένη προσαρμογή (*regularized fit*) για οποιονδήποτε δεδομένο αριθμό παραμέτρων, όπως με τη σειρά των μοντέλων που προκύπτουν από την forward selection.

Έστω δείγμα μεγέθους n, πλήθος ανεξάρτητων μεταβλητών d και $\{(\boldsymbol{x}_i, y_i): \boldsymbol{x}_i \in \mathbb{R}^d,\ y_i \in \mathbb{R},\ i = 1,2,\dots,n\}$. Η Y ακολουθεί κάποια κατανομή της εκθετικής οικογένειας κατανομών με μέση τιμή $\mu = E(Y)$ και διακύμανση $V = Var(Y)$. Επίσης, ξέρουμε ότι ισχύει:

$$\eta = g(\mu) = \beta_0 + \boldsymbol{x}^T\beta.$$

Η συνάρτηση πιθανότητας της Y ορίζεται ως:

$$f(y; \theta, \varphi) = exp\left[\frac{y\theta - b(\theta)}{\alpha(\varphi)} + c(y, \varphi)\right],$$

με $a(\bullet)$, $b(\bullet)$ και $c(\bullet)$ συναρτήσεις που διαφέρουν ανάλογα την κατανομή. Αν υποθέσουμε ότι το φ είναι γνωστό, θέλουμε να βρούμε τη λύση μέγιστης πιθανοφάνειας για τη φυσική

παράμετρο θ, τα $(\beta_0, \beta^T)^T$ δηλαδή, χρησιμοποιώντας μία L_1-νόρμας ποινή. Οπότε, για συγκεκριμένο λ, θέλουμε να βρούμε τα $\boldsymbol{\beta} = (\beta_0, \beta^T)^T$ που ελαχιστοποιούν την ποσότητα:

$$l(\boldsymbol{\beta}, \lambda) = -\sum_{i=1}^{n} [y_i \theta(\boldsymbol{\beta})_i - b\{\theta(\boldsymbol{\beta})_i\}] + \lambda \|\beta\|_1.$$

Επίσης, αν υποθέσουμε ότι κανένα από τα στοιχεία του β δεν είναι 0 και αν παραγωγίσουμε την $l(\boldsymbol{\beta}, \lambda)$ ως προς $\boldsymbol{\beta}$, ορίζουμε την συνάρτηση:

$$H(\boldsymbol{\beta}, \lambda) = \frac{\partial l}{\partial \boldsymbol{\beta}} = -\boldsymbol{X}^T \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{\mu}) \frac{\partial \eta}{\partial \mu} + \lambda sgn \begin{pmatrix} 0 \\ \beta \end{pmatrix},$$

όπου $\boldsymbol{X}$ ο πίνακας σχεδιασμού, $\boldsymbol{W}$ ένας διαγώνιος πίνακας με n διαγώνια στοιχεία $\boldsymbol{V_i^{-1}} \left(\frac{\partial \mu}{\partial \eta}\right)_{\boldsymbol{i}}^{\boldsymbol{2}}$ και $(\boldsymbol{y} - \boldsymbol{\mu}) \frac{\partial \eta}{\partial \mu}$ ένα διάνυσμα με n στοιχεία $(y_i - \mu_i) \left(\frac{\partial \mu}{\partial \eta}\right)_i$. Αν και έχουμε υποθέσει ότι κανένα από τα στοιχεία του β δεν είναι 0, το σύνολο των μη μηδενικών στοιχείων του $\boldsymbol{\beta}$ αλλάζει με το λ, οπότε η $H(\boldsymbol{\beta}, \lambda)$ θα πρέπει να επαναπροσδιοριστεί αναλόγως. Σκοπός τώρα είναι να υπολογίσουμε ολόκληρο το path των λύσεων για τους συντελεστές $\boldsymbol{\beta}$, καθώς το λ μειώνεται από λ_{max} σε 0.

Ο αλγόριθμος πρόβλεψης-διόρθωσης αποτελεί μία από τις θεμελιώδεις στρατηγικές για την εφαρμογή της αριθμητικής συνέχειας. Η αριθμητική συνέχεια χρησιμοποιείται στα Μαθηματικά για να προσδιορίσει το σύνολο των λύσεων σε μη γραμμικές εξισώσεις που ανιχνεύονται μέσω μίας μονοδιάστατης παραμέτρου. Ανάμεσα από πολλές προσεγγίσεις, η μέθοδος πρόβλεψης-διόρθωσης βρίσκει ρητά μία σειρά λύσεων χρησιμοποιώντας τις αρχικές συνθήκες και συνεχίζει να βρίσκει τις γειτονικές λύσεις βασιζόμενη στις τωρινές λύσεις. Το λήμμα που ακολουθεί παρέχει την αρχικοποίηση των διαδρομών των συντελεστών (*coefficient paths*).

Λήμμα 1. Όταν το λ ξεπερνάει ένα ορισμένο όριο (*threshold*), ο σταθερός όρος είναι ο μόνος μη μηδενικός συντελεστής: $\hat{\beta}_0 = g(\bar{y})$ και

$$H\left\{\left(\hat{\beta}_0, 0, \dots, 0\right)^T, \lambda\right\} = 0 \; \gamma\iota\alpha \; \lambda > \max_{j \in \{1,2,\dots,d\}} \left| \boldsymbol{x}_j^T \widehat{\boldsymbol{W}} (\boldsymbol{y} - \bar{y}\boldsymbol{1}) g'(\bar{y}) \right|.$$

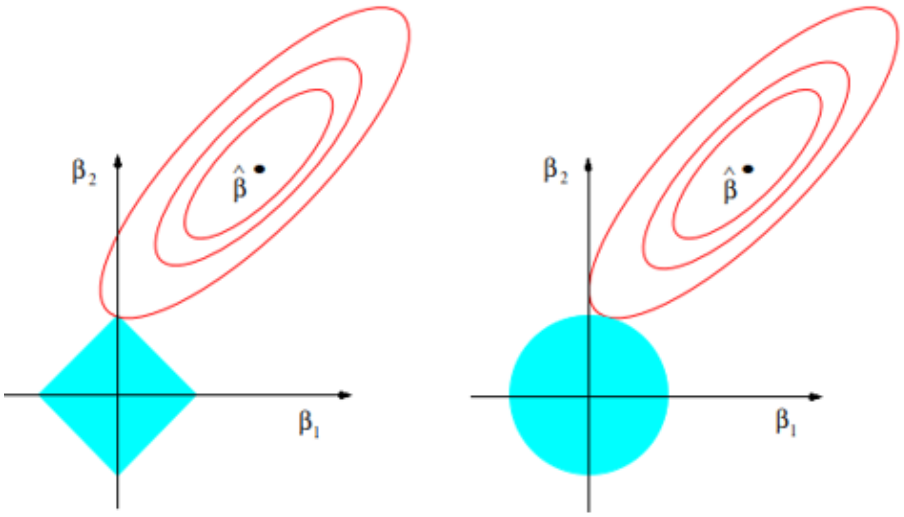
Ορίζουμε αυτό το όριο του λ ως λ_{max}. Καθώς το λ μειώνεται, άλλες μεταβλητές εντάσσονται στο ενεργό σύνολο, ξεκινώντας με τη μεταβλητή $j_0 = argmax_j \left| \boldsymbol{x}_j^T (\boldsymbol{y} - \right.$

$\bar{y}\mathbf{1})|$. Μειώνοντας το λ, γίνεται εναλλαγή μεταξύ βήματος πρόβλεψης και βήματος διόρθωσης, με τα βήματα της k-οστής επανάληψης να είναι:

1) Βήμα μήκους. Προσδιορίζεται η μείωση του λ. Δεδομένου του λ_k, γίνεται προσέγγιση του επόμενου μεγαλύτερου λ στο οποίο το τωρινό αλλάζει, το λ_{k+1}.
2) Βήμα πρόβλεψης. Προσεγγίζεται γραμμικά η αντίστοιχη μεταβολή στο $\boldsymbol{\beta}$ με τη μείωση του λ, που αποκαλείται $\hat{\beta}^{k+}$.
3) Βήμα διόρθωσης. Βρίσκεται η ακριβής λύση του $\boldsymbol{\beta}$ που συνδυάζεται με το λ_{k+1}, δηλαδή το $\boldsymbol{\beta}(\lambda_{\kappa+1})$, χρησιμοποιώντας το $\hat{\beta}^{k+}$ ως αρχική τιμή και αποκαλείται $\hat{\beta}^{k+1}$.
4) Ενεργό σύνολο. Ελέγχεται αν το τωρινό ενεργό σύνολο πρέπει να τροποποιηθεί. Αν ναι, επαναλαμβάνεται το βήμα διόρθωσης με το ενημερωμένο ενεργό σύνολο.

Κατά αυτό τον τρόπο, υπολογίζεται ολόκληρο το path της lasso.

Είδαμε ότι η lasso έχει τη χρήσιμη ικανότητα να μηδενίζει κάποια $\hat{\beta}_j$ εντελώς, σε αντίθεση με τη ridge, που είναι και το μειονέκτημά της. Ας δούμε για ποιο λόγο συμβαίνει αυτό. Στο Σχήμα 3-13 (James et al., 2023), μπορούμε να δούμε τη lasso (αριστερά) και τη ridge (δεξιά) στην περίπτωση μόνο δύο ανεξάρτητων μεταβλητών, έστω X_1, X_2, κατά την εφαρμογή τους στο κλασσικό γραμμικό μοντέλο. Η περιοχή περιορισμού (*constraint region*) για τη ridge είναι ο δίσκος $\beta_1^2 + \beta_2^2 \leq t$, ενώ για τη lasso το διαμάντι $|\beta_1| + |\beta_2| \leq t$. Αυτό που κάνουν και οι δύο μέθοδοι, είναι να βρουν το πρώτο σημείο όπου οι περιοχές περιορισμού εφάπτονται με τα ελλειπτικά περιγράμματα (*elliptical contours*), τα οποία αφορούν το RSS κεντραρισμένο στους πλήρεις εκτιμητές ελαχίστων τετραγώνων. Η εσωτερική έλλειψη έχει μικρότερο RSS. Είναι εύκολο να δούμε ότι η περιοχή περιορισμού της lasso έχει γωνίες, σε αντίθεση με της ridge. Επομένως, αν η λύση επιτυγχάνεται σε κάποια γωνία, τότε έχει κάποια παράμετρο $\beta_j = 0$. Στην περίπτωση που $d > 2$, η περιοχή περιορισμού της lasso γίνεται ρομβοειδής με πολλές γωνίες και άρα είναι αρκετά πιο πιθανό για κάποιες από τις εκτιμηθείσες παραμέτρους να είναι ίσες με το 0.



ΣΧΗΜΑ 3-13 : Γεωμετρία της lasso (αριστερά) και της ridge (δεξιά) για δύο μεταβλητές.

Μέσω προσομοιώσεων που έχουν γίνει, μπορούμε να συμπεράνουμε ότι ούτε η ridge ούτε η lasso θα αποδίδει καθολικά καλύτερα από την άλλη (James et al., 2023). Γενικώς αναμένεται ότι η lasso αποδίδει καλύτερα σε δεδομένα όπου ένας σχετικά μικρός αριθμός των $X_1, X_2, \dots, X_d$ έχει σημαντικούς συντελεστές και οι υπόλοιπες μεταβλητές έχουν είτε πολύ μικρούς ή μηδενικούς συντελεστές. Από την άλλη, η ridge θα αποδίδει καλύτερα όταν η Y είναι συνάρτηση πολλών ανεξάρτητων μεταβλητών με συντελεστές παρόμοιου μεγέθους. Ωστόσο, επειδή στα πραγματικά δεδομένα δεν είναι ποτέ γνωστό το πόσες μεταβλητές σχετίζονται με την Y, μπορούμε να χρησιμοποιήσουμε τεχνικές όπως η διασταυρούμενη επικύρωση για να εντοπίσουμε ποια από τις δύο θα αποδώσει καλύτερα σε ένα συγκεκριμένο σύνολο δεδομένων.

Η lasso παρά τη χρησιμότητά της και τη δημοφιλία της ως συνάρτηση ποινής, δεν παύει να έχει κάποια μειονεκτήματα. Η lasso αυτό που κάνει είναι να ποινικοποιεί τα $|\beta_j|$ που είναι μεγάλα. Όμως, μπορεί να ποινικοποιήσει υπερβολικά τα β_j που είναι όντως μεγάλα. Επίσης, οι εκτιμητές μέσω της μεθόδου lasso δεν ακολουθούν ούτε κατά προσέγγιση την Κανονική κατανομή και μπορεί να είναι αρκετά μεροληπτικοί, με αποτέλεσμα η εξαγωγή στατιστικών συμπερασμάτων να δυσκολεύει. Ακόμα, λόγω του ότι οι εκτιμητές μπορεί να είναι μεροληπτικοί, τα τυπικά σφάλματα είναι αρκετά πιθανό να είναι παραπλανητικά (Agresti and Kateri, 2021).

Τέλος, να αναφερθεί ότι οι μέθοδοι lasso και ridge μπορούν να εκφραστούν ως ειδικές περιπτώσεις μίας άλλης ποινής. Συγκεκριμένα, οι Frank and Friedman (1993) πρότειναν μία γενίκευση της ridge και της subset selection, προσθέτοντας στο RSS μία

ποινή της μορφής $\lambda \sum_{j=1}^{d} |\beta_j|^q$ με $q \geq 0$, ή γραμμένο σε μορφή περιορισμού $\sum_{j=1}^{d} |\beta_j|^q \leq t$. Αυτού του είδους η ποινή αποκαλείται bridge. Είναι εύκολο να δούμε ότι αν στον παραπάνω τύπο αντικαταστήσουμε το $q = 1$ έχουμε τη lasso και για $q = 2$ τη ridge.

3.7.3 Μέθοδος elastic net

Οι Tibshirani (1996) και Fu (1998) συνέκριναν την προβλεπτική απόδοση των lasso, ridge και bridge και κατέληξαν στο ότι καμία δεν κυριαρχεί έναντι των άλλων δύο. Ωστόσο, λόγω της επιλογής μεταβλητών, κάτι πολύ χρήσιμο ειδικά στην ανάλυση των δεδομένων υψηλών διαστάσεων της σημερινής εποχής, η lasso είναι αρκετά πιο ελκυστική στη χρήση της από τις άλλες δύο χάριν στην αραιή της αναπαράσταση. Παρ'όλα αυτά, και παρά την επιτυχία της σε πολλές περιπτώσεις, η lasso έχει κάποιους περιορισμούς, όπως στις παρακάτω περιπτώσεις (Zou and Hastie, 2005):

1) Αν $d > n$, επιλέγει το πολύ n μεταβλητές πριν υπάρξει κορεσμός στο μοντέλο, λόγω της φύσης του προβλήματος κυρτής βελτιστοποίησης. Κάτι που αποτελεί περιοριστικό χαρακτηριστικό για μία μέθοδο επιλογής μεταβλητών. Επίσης, η lasso δεν είναι καλά ορισμένη, εκτός και αν το όριο της L_1-νόρμας των συντελεστών είναι μικρότερο από μία ορισμένη τιμή.
2) Αν υπάρχει μία ομάδα μεταβλητών μεταξύ των οποίων οι συσχετίσεις είναι υψηλές, τείνει να επιλέγει μία εξ αυτών στην τύχη, αδιαφορώντας ποια είναι αυτή.
3) Αν $n > d$, έχει παρατηρηθεί ότι στην περίπτωση που υπάρχουν υψηλές συσχετίσεις μεταξύ των ανεξάρτητων μεταβλητών, η ridge έχει καλύτερη απόδοση από τη lasso (Tibshirani, 1996).

Οπότε βλέπουμε ότι σε κάποιες περιπτώσεις, η lasso δεν αποτελεί πάντα την καταλληλότερη μέθοδο για επιλογή μεταβλητών. Οι Zou and Hastie (2005) πρότειναν μία μέθοδο ποινικοποίησης που την ονόμασαν elastic net, η οποία αποδίδει καλύτερα από τη lasso. Η elastic net κάνει συνεχή συρρίκνωση των μεταβλητών και ταυτόχρονα επιλογή μεταβλητών, όπως και η lasso, οπότε έχει την ιδιότητα της σποραδικότητας που είναι εξαιρετικά σημαντική. Όμως, η elastic net έχει και κάτι ακόμα χρήσιμο, το φαινόμενο της ομαδοποίησης (*grouping effect*). Κατά το φαινόμενο αυτό, μεταβλητές που είναι μεταξύ τους υψηλά συσχετισμένες, τείνουν να βρίσκονται μαζί εντός ή εκτός

του μοντέλου. Δηλαδή μπορεί να επιλέξει και ομάδες συσχετισμένων μεταβλητών, σε αντίθεση με τη lasso. Όπως την παρομοιάζουν οι Zou and Hastie, η elastic net είναι σαν ένα δίχτυ ψαρέματος που τεντώνεται και διατηρεί «όλα τα μεγάλα ψάρια». Το επίσης πολύ σημαντικό, η elastic net είναι ιδιαίτερα χρήσιμη στην περίπτωση όπου $d \gg n$, ενώ η lasso δεν αποτελεί καλή μέθοδο επιλογής μεταβλητών σε αυτήν την περίπτωση. Παραδείγματα που έχουν πραγματοποιηθεί με πραγματικά δεδομένα, αλλά και μελέτες προσομοίωσης, έχουν δείξει ότι η elastic net αποδίδει καλύτερα από τη lasso σχετικά με την ακρίβεια πρόβλεψης.

Ας υποθέσουμε ότι έχουμε ένα σύνολο δεδομένων με n παρατηρήσεις, εξαρτημένη μεταβλητή Y και $X_1, X_2, \dots, X_d$ ανεξάρτητες μεταβλητές. Έστω $\boldsymbol{y} = (y_1, y_2, \dots, y_n)^T$ και $\boldsymbol{X} = (\boldsymbol{x}_1 | \dots | \boldsymbol{x}_d)$ ο πίνακας σχεδιασμού, όπου $\boldsymbol{x}_j = \left(x_{1j}, x_{2j}, \dots, x_{nj}\right)^T$ με $j = 1,2, \dots d$ οι παρατηρήσεις. Έπειτα από τυποποίηση, θα έχουμε $\sum_{i=1}^{n} y_i = 0$, $\sum_{i=1}^{n} x_{ij} = 0$ και $\sum_{i=1}^{n} x_{ij}^2 = 1$ για $j = 1,2, \dots d$. Οι Zou and Hastie (2005) όρισαν ως naïve elastic net κριτήριο το:

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = |\boldsymbol{y} - \boldsymbol{X\beta}|^2 + \lambda_2 |\boldsymbol{\beta}|^2 + \lambda_1 |\boldsymbol{\beta}|_1, \tag{3.10}$$

για οποιεσδήποτε σταθερές $\lambda_1, \lambda_2 \geq 0$, με $|\boldsymbol{\beta}|^2 = \sum_{j=1}^{d} \beta_j^2$ και $|\boldsymbol{\beta}|_1 = \sum_{j=1}^{d} |\beta_j|$. Ο naïve elastic net εκτιμητής $\widehat{\boldsymbol{\beta}}$ είναι αυτός που ελαχιστοποιεί την ποσότητα (3.10):

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin}[L(\lambda_1, \lambda_2, \boldsymbol{\beta})].$$

Η διαδικασία αυτή μπορεί να θεωρηθεί ως μία μέθοδος ποινικοποιημένων ελαχίστων τετραγώνων (*penalized least squares, PLS*). Τώρα, αν θέσουμε $\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$, η λύση της (3.10) είναι ισοδύναμη με το πρόβλημα βελτιστοποίησης

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin} |\boldsymbol{y} - \boldsymbol{X\beta}|^2 \text{ , υπό τον περιορισμό } (1-a)|\boldsymbol{\beta}|_1 + a|\boldsymbol{\beta}|^2 \leq t$$

και $t \geq 0$. Η συνάρτηση $(1-a)|\boldsymbol{\beta}|_1 + a|\boldsymbol{\beta}|^2$ αποκαλείται ποινή elastic net και αποτελεί κυρτό συνδυασμό της ridge και της lasso. Ισχύει ότι $\alpha \in [0,1]$ και είναι μία παράμετρος που ελέγχει την ισορροπία μεταξύ της ridge και της lasso. Αν αντικαταστήσουμε το $\alpha = 0$, η elastic net θα μας δώσει τη lasso και για $\alpha = 1$, τη ridge. Για κάθε $\alpha \in [0,1)$, η συνάρτηση ποινής της elastic net είναι μοναδική (χωρίς πρώτη παράγωγο) στο 0 και αυστηρώς κυρτή για κάθε $\alpha > 0$. Οπότε η elastic net έχει τα χαρακτηριστικά και της ridge και της lasso. Επίσης, είναι αρκετά ενδιαφέρον το ότι για $\alpha = \varepsilon$ για κάποιο πολύ μικρό $\varepsilon > 0$, η elastic net αποδίδει παρόμοια με τη lasso, αλλά αφαιρεί τυχόν εκφυλισμούς (*degeneracies*) που

προκαλούνται από την ύπαρξη συσχετίσεων μεταξύ των ανεξάρτητων μεταβλητών (Emmert-Streib et al., 2023). Για κάποιο λ, αν μειώσουμε την παράμετρο α από το 1 στο 0, ο αριθμός των συντελεστών που είναι ίσοι με το 0, αυξάνεται μονότονα από το 0 (πλήρες μοντέλο ridge) στην σποραδικότητα της λύσης της lasso.

Στην περίπτωση της λογιστικής παλινδρόμησης και εκφράζοντας τη συνάρτηση ποινής της elastic net ως:

$$s(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{d} \left(a\beta_j^2 + (1-a)|\beta_j|\right),$$

με $\lambda = \lambda_1 + \lambda_2 \geq 0$ ρυθμιστική παράμετρο και $\alpha \in [0,1]$, θέλουμε να βρούμε τους εκτιμητές που ελαχιστοποιούν την ποσότητα (3.3):

$$\widehat{\boldsymbol{\beta}}^{elastic\ net} = \underset{\boldsymbol{\beta}}{argmin}\left[-l(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{d} \left(a\beta_j^2 + (1-a)|\beta_j|\right)\right], \qquad (3.11)$$

ή ισοδύναμα:

$$\widehat{\boldsymbol{\beta}}^{elastic\ net} = \underset{\boldsymbol{\beta}}{argmin}[-l(\boldsymbol{\beta})]\ , \textit{υπό τον περιορισμό}\ (1-a)|\boldsymbol{\beta}|_1 + a|\boldsymbol{\beta}|^2 \leq t,$$

με $t \geq 0$.

Οι Friedman et al. (2010) παρουσίασαν έναν υπολογιστικά αποδοτικό αλγόριθμο για να υπολογίσουν το regularization path της elastic net στην περίπτωση της γραμμικής παλινδρόμησης, της λογιστικής και της πολυωνυμικής λογιστικής παλινδρόμησης (*multinomial logistic regression*). Οι Simon et al. (2011) επέκτειναν την πρόταση αυτή στα μοντέλα του Cox για δεξιά-λογοκριμένα δεδομένα. Οι Tay et al. (2023) επέκτειναν περαιτέρω την ποινικοποιημένη συνάρτηση elastic net στα ΓΓΜ, σε μοντέλα Cox με δεδομένα (start,stop] και στρώματα, και σε μια απλοποιημένη εκδοχή της relaxed lasso. Οι Tay et al. ορίζουν ως το πρόβλημα που πρέπει να επιλυθεί τη σχέση:

$$\left(\hat{\beta}_0, \hat{\beta}\right) = \underset{(\beta_0,\beta)\in\mathbb{R}^{d+1}}{argmin}\left[-\frac{1}{n}\sum_{i=1}^{n} l(y_i, \beta_0 + x_i^T\beta) + \lambda\left(\frac{1-\alpha}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1\right)\right], \qquad (3.12)$$

με $l(y_i, \beta_0 + x_i^T\beta)$ το λογάριθμο της πιθανοφάνειας για την παρατήρηση i, ώστε να βρουν το path της elastic net. Ο Αλγόριθμος 4 είναι αυτός που προτείνουν για τη λύση της (3.12) για ένα path τιμών της λ.

Αλγόριθμος 4 *Προσαρμογή ΓΓΜ με χρήση της elastic net.*

1. Διαλέγεται μια τιμή του $\alpha \in [0,1]$ και μια ακολουθία τιμών του λ $\lambda_1 > \lambda_2 \ldots > \lambda_m$.
2. Για $k = 1,2,\ldots,m$:

   (α) Αρχικοποιείται $\left(\hat{\beta}_0^{(0)}(\lambda_k), \hat{\beta}^{(0)}(\lambda_k)\right) = \left(\hat{\beta}_0(\lambda_{k-1}), \hat{\beta}(\lambda_{k-1})\right)$.

   Για $k = 1$, αρχικοποιείται

   $$\left(\hat{\beta}_0^{(0)}(\lambda_k), \hat{\beta}^{(0)}(\lambda_k)\right) = (0, \mathbf{0}).$$

   (Το $\left(\hat{\beta}_0(\lambda_k), \hat{\beta}(\lambda_k)\right)$ δηλώνει την elastic net λύση για $\lambda = \lambda_k$.)

   (β) Για $t = 0,1,\ldots$ μέχρι τη σύγκλιση:

   i. Για $i = 1,2,\ldots,n$,

   υπολογίζεται $\eta_i^{(t)} = \hat{\beta}_0^{(t)}(\lambda_k) + \hat{\beta}^{(t)}(\lambda_k)^T x_i$ και $\mu_i^{(t)} = g^{-1}\left(\eta_i^{(t)}\right)$.

   ii. Για $i = 1,2,\ldots,n$, υπολογίζονται τα:

   $$z_i^{(t)} = \eta_i^{(t)} + \left(y_i - \mu_i^{(t)}\right) \Big/ \frac{\partial \mu_i^{(t)}}{\partial \eta_i^{(t)}},$$

   $$w_i^{(t)} = \left(\frac{\partial \mu_i^{(t)}}{\partial \eta_i^{(t)}}\right)^2 \Big/ V\left(\mu_i^{(t)}\right).$$

   iii. Λύνεται το πρόβλημα των ποινικοποιημένων σταθμισμένων ελαχίστων τετραγώνων

   $$\left(\hat{\beta}_0^{(t+1)}(\lambda_k), \hat{\beta}^{(t+1)}(\lambda_k)\right)$$
   $$= \underset{(\beta_0,\beta)\in\mathbb{R}^{d+1}}{argmin} \left[\frac{1}{2n}\sum_{i=1}^{n} w_i^{(t)}\left(z_i^{(t)} - \beta_0 - x_i^T\beta\right)^2 + \lambda_k\left(\frac{1-\alpha}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1\right)\right].$$

Στο βήμα 2.α, βλέπουμε ότι αρχικοποιούν τη λύση για $\lambda = \lambda_k$ στη λύση που προκύπτει για $\lambda = \lambda_{k-1}$. Αυτό είναι γνωστό ως «θερμή» εκκίνηση (*warm start*). Επειδή αναμένεται ότι η λύση θα είναι παρόμοια για αυτές τις δύο τιμές του λ, είναι πιθανόν ο αλγόριθμος να χρειαστεί λιγότερες επαναλήψεις από ότι αν γινόταν η αρχικοποίηση της λύσης στο 0.

Στο Σχήμα 3-14 (Fan et al., 2020), μπορούμε να δούμε μία γεωμετρική σύγκριση μεταξύ των ridge, lasso και elastic net. Παρατηρούμε ότι η elastic net και η lasso έχουν τις ίδιες τέσσερις αιχμηρές κορυφές. Επίσης, βλέπουμε ότι οι άκρες της elastic net

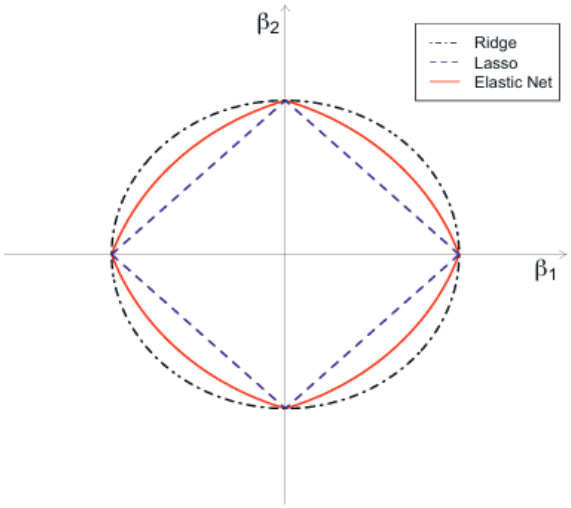
είναι στρογγυλές, όπως και της ridge. Γνωρίζουμε ότι η ridge είναι χρήσιμη στην περίπτωση ύπαρξης πολυσυγγραμμικότητας και ότι επιτυγχάνει μία καλή αντιστάθμιση μεροληψίας-διακύμανσης. Μέσω του μέρους της που αφορά τη ridge, η elastic net έχει την ικανότητα να διαχειρίζεται την πολυσυγγραμμικότητα και μέσω του μέρους της που αφορά τη lasso, διατηρεί την ιδιότητα της σποραδικότητας. Ακόμα, η elastic net συνήθως κατασκευάζει πιο ακριβές προβλεπτικό μοντέλο στην περίπτωση των δεδομένων υψηλών διαστάσεων από τη lasso.

Έχουμε αναφέρει ότι η lasso δεν είναι ικανή στο να διαχειρίζεται καλά μεταβλητές με υψηλή συσχέτιση. Τα coefficient paths τείνουν να είναι ασταθή και κάποιες φορές μπορεί να έχουν περίεργη συμπεριφορά. Ας υποθέσουμε ότι έχουμε την X_j με $\hat{\beta}_j > 0$ για συγκεκριμένη τιμή του λ. Εάν αυξήσουμε τα δεδομένα μας με ένα πανομοιότυπο αντίγραφο $X_{j'} = X_j$, τότε μπορούν να μοιραστούν τον συντελεστή $\hat{\beta}_j$ με άπειρους τρόπους για οποιοδήποτε $\tilde{\beta}_j + \tilde{\beta}_{j'} = \hat{\beta}_j$ και $\tilde{\beta}_j, \tilde{\beta}_{j'} > 0$. Επομένως οι συντελεστές για αυτό το ζεύγος μεταβλητών δεν ορίζονται. Μία τετραγωνική ποινή όμως, θα διαιρέσει ακριβώς στα δύο την επίδραση της $\hat{\beta}_j$ στις X_j και $X_{j'}$. Πρακτικά, σχεδόν ποτέ δε θα έχουμε ένα πανομοιότυπο ζευγάρι μεταβλητών. Σε πάρα πολλές έρευνες όμως, έχουμε ομάδες υψηλά συσχετισμένων μεταβλητών. Παραδείγματος χάριν, σε μελέτες μικροσυστοιχίων ομάδες γονιδίων στο ίδιο βιολογικό μονοπάτι τείνουν να εκφράζονται μαζί, με αποτέλεσμα να υπάρχει υψηλή συσχέτιση στα μέτρα τους. Οι Hastie et al. (2017) παραθέτουν ένα παράδειγμα, που αφορά το σύνολο δεδομένων leukemia, όπου και εφαρμόζουν ποινικοποιημένη λογιστική παλινδρόμηση. Η elastic net μέσω του μέρους της που αφορά τη ridge, μετριάζει την επίδραση των υψηλά συσχετισμένων μεταβλητών, ενώ παράλληλα το μέρος της που αφορά τη lasso παράγει μια αραιή λύση για αυτούς τους μετριασμένους συντελεστές. Στο Σχήμα 3-15 βλέπουμε τους εκτιμητές για τη lasso και την elastic net για τη ρυθμιστική παράμετρο λ σε λογαριθμική κλίμακα. Στο πλαίσιο που αφορά τη lasso, στην αρχή υπάρχουν 19 μη μηδενικές μεταβλητές και στο πλαίσιο της elastic net αντίστοιχα, 39. Λόγω του όρου της ridge, η elastic net έχει περισσότερες μη μηδενικές μεταβλητές από ότι η lasso, με μικρότερο μέγεθος όμως.

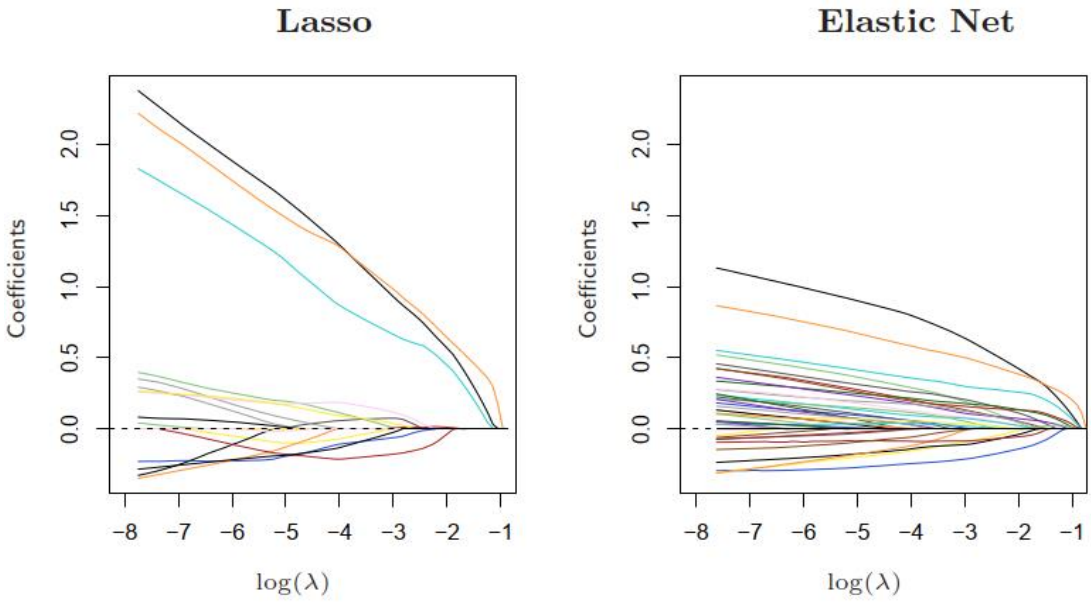
Συνοψίζοντας, μπορούμε να επισημάνουμε τα παρακάτω που προέκυψαν από διάφορες μελέτες σχετικά με την elastic net (Emmert-Streib et al., 2023):

1) Στην περίπτωση συσχετισμένων μεταβλητών, η elastic net μπορεί να οδηγήσει σε χαμηλότερα μέσα τετραγωνικά σφάλματα, συγκριτικά με τις ridge και lasso.

2) Στην περίπτωση συσχετισμένων μεταβλητών, η elastic net επιλέγει όλες τις μεταβλητές, ενώ η lasso επιλέγει μία μεταβλητή από μία ομάδα συσχετισμένων μεταβλητών και τείνει να αγνοεί τις υπόλοιπες.
3) Στην περίπτωση μη συσχετισμένων μεταβλητών, η επιπλέον ποινή της ridge αποφέρει μικρή βελτίωση.
4) Η elastic net προσδιορίζει σωστά μεγαλύτερο αριθμό μεταβλητών σε σχέση με τη lasso (επιλογή μοντέλου).
5) Η elastic net συνήθως έχει μικρότερο ψευδώς θετικό ποσοστό από τη ridge.
6) Αν $d > n$, η lasso επιλέγει το πολύ n ανεξάρτητες μεταβλητές, ενώ η elastic net μπορεί να επιλέξει και παραπάνω.



ΣΧΗΜΑ 3-14 : Γεωμετρική σύγκριση των ridge, lasso και elastic net (α=0,5) για δύο μεταβλητές.



ΣΧΗΜΑ 3-15 : Το path της lasso (αριστερά) και της elastic net (δεξιά) για το σύνολο δεδομένων leukemia (Hastie et al., 2017).

3.7.4 Μέθοδος SCAD

Μέχρι στιγμής, έχουμε αναφέρει ουκ ολίγες φορές τη σημαντικότητα της επιλογής μεταβλητών σε ένα μοντέλο. Πρακτικά, ένας μεγάλος αριθμός ανεξάρτητων μεταβλητών συνήθως εισάγεται στο αρχικό στάδιο της κατασκευής ενός μοντέλου, ώστε να εξασθενήσει πιθανή μεροληψία στη μοντελοποίηση. Από την άλλη, για την ενίσχυση της προβλεπτικής ικανότητας και για την επιλογή σημαντικών μεταβλητών, οι ερευνητές συχνά χρησιμοποιούν τη stepwise deletion και την subset selection. Αν και οι μέθοδοι αυτές είναι χρήσιμες στην πράξη, αγνοούν τα στοχαστικά σφάλματα που προκύπτουν στα στάδια της επιλογής μεταβλητών, με συνέπεια οι θεωρητικές τους ιδιότητες να είναι κάπως δύσκολο να κατανοηθούν. Επιπλέον, η subset selection πάσχει από αρκετά μειονεκτήματα, με το πιο σοβαρό να είναι η έλλειψη σταθερότητας. Οι Fan and Li (2001) αναφέρουν ότι μία καλή συνάρτηση ποινής δίνει εκτιμητές με τις εξής τρεις ιδιότητες:

1) Σποραδικότητα (*sparsity*). Ο εκτιμητής που προκύπτει μηδενίζει αυτομάτως τους μικρούς εκτιμώμενους συντελεστές, ώστε να επιτύχει επιλογή μεταβλητών και να μειώσει την πολυπλοκότητα του μοντέλου.

2) Αμεροληψία (*unbiasedness*). Ο εκτιμητής που προκύπτει είναι σχεδόν αμερόληπτος, ειδικά όταν ο πραγματικός συντελεστής β_j είναι μεγάλος, ώστε να μειώσει τη μεροληψία του μοντέλου.
3) Συνέχεια (*continuity*). Ο εκτιμητής που προκύπτει είναι συνεχής στα δεδομένα για να μειώσει την αστάθεια στην πρόβλεψη του μοντέλου.

Οι διάφορες συναρτήσεις ποινής L_q δεν ικανοποιούν ταυτόχρονα τις μαθηματικές συνθήκες που απαιτούνται για την σποραδικότητα, την αμεροληψία και τη συνέχεια (Fan et al., 2020). Η ποινή L_q με $q > 1$ είναι κυρτή και δεν ικανοποιεί την ιδιότητα της σποραδικότητας. Η ποινή L_1 δεν ικανοποιεί την ιδιότητα της αμεροληψίας, ενώ η ποινή L_q με $0 \leq q < 1$ είναι κοίλη (*concave*), αλλά δεν ικανοποιεί την ιδιότητα της συνέχειας. Να αναφερθεί ότι η ποινή L_0 αντιστοιχεί στην best subset selection. Όπως γνωρίζουμε, η ποινή L_1 διαφέρει σημαντικά από την L_0, καθώς η L_1 ποινικοποιεί τις μεγάλες παραμέτρους πολύ. Για να μειώσουν περαιτέρω τη μεροληψία της εκτίμησης, οι Antoniadis and Fan (2001) και Fan and Li (2001) εισήγαγαν την ιδέα των αναδιπλούμενων κοίλων ποινικοποιημένων ελαχίστων τετραγώνων (*folded concave penalized least squares*), όπου η $p_\lambda(\theta)$ είναι συμμετρική και κοίλη σε κάθε πλευρά. Το γεγονός ότι είναι ιδιάζουσα στο μηδέν (*singularity at the origin*) (δηλαδή $p'_\lambda(0_+) > 0$) επαρκεί για τη δημιουργία σποραδικότητας στην επιλογή μεταβλητών και το γεγονός ότι είναι κοίλη χρειάζεται για τη μείωση της μεροληψίας της εκτίμησης. Αυτό είναι που οδήγησε σε μια «οικογένεια» αναδιπλούμενων κοίλων συναρτήσεων ποινής που είναι ιδιάζουσες στο μηδέν. Οι ποινές αυτές περιλαμβάνουν μία ρυθμιστική παράμετρο α που ελέγχει το πόσο κοίλη είναι η συνάρτηση ποινής, και κατ'επέκταση το πόσο γρήγορα η επίδραση της ποινής φθίνει. Η ποινή L_1 μπορεί να θεωρηθεί και ως κοίλη και ως κυρτή συνάρτηση· εμπίπτει στα όρια της «οικογένειας» αυτής.Σε μια προσπάθεια αυτόματης και ταυτόχρονης επιλογής μεταβλητών, οι Fan and Li (2001) πρότειναν τη συνεχώς διαφορίσιμη συνάρτηση ποινής που ορίζεται ως:

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

με $I(\bullet)$ μία δείκτρια συνάρτηση, $p_\lambda(0) = 0$, $\alpha > 2$ και $\beta > 0$, όπου και αποκάλεσαν την ποινή αυτή SCAD (smoothly clipped absolute deviation), η οποία διατηρεί τα θετικά χαρακτηριστικά της ridge και της subset selection και ικανοποιεί και τις τρεις προαναφερθείσες ιδιότητες. Παρατηρείται ότι για $\alpha = \infty$, η SCAD είναι ίση με την

ποινή L_1 (Fan et al., 2020). Όπως μπορούμε να διακρίνουμε στο Σχήμα 3-16 (Fan et al., 2020), η SCAD συμπεριφέρεται σαν την ποινή L_1 στο μηδέν, έτσι ώστε να διατηρήσει την ιδιότητα της επιλογής μεταβλητών και σαν την ποινή L_0 στις ουρές, ώστε να βελτιώσει την ιδιότητα της μεροληψίας της L_1.

Ένας διαφορετικός τρόπος για να ορίσουμε την SCAD είναι ως (Breheny and Huang, 2011):

$$p_{\lambda,\alpha}(\beta) = \begin{cases} \lambda\beta, & \beta \leq \lambda \\ \dfrac{\alpha\lambda\beta - 0{,}5(\beta^2 + \lambda^2)}{\alpha - 1}, & \lambda < \beta \leq \alpha\lambda \\ \dfrac{\alpha^2(\alpha^2 - 1)}{2(\alpha - 1)}, & \beta > \alpha\lambda \end{cases},$$

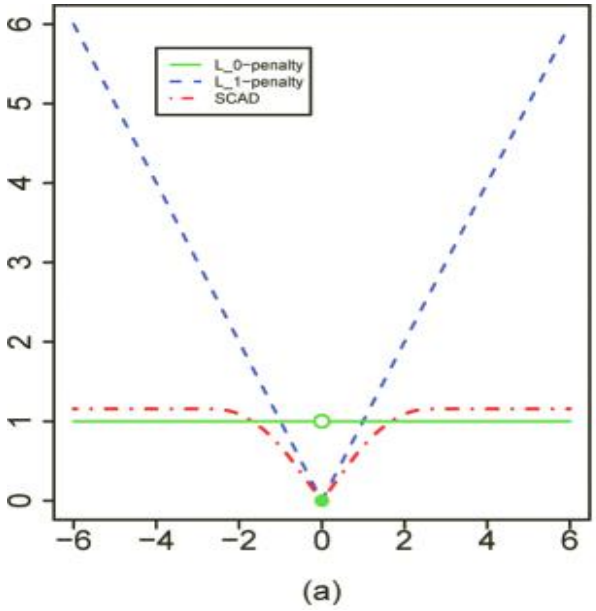
με πρώτη παράγωγο:

$$p'_\lambda(\beta) = \begin{cases} \lambda, & \beta \leq \lambda \\ \dfrac{\alpha\lambda - \beta}{\alpha - 1}, & \lambda < \beta \leq \alpha\lambda \\ 0, & \beta > \alpha\lambda \end{cases}.$$

Ύστερα από αρκετές προσομοιώσεις, το $\alpha \approx 3{,}7$ φαίνεται να αποδίδει αρκετά καλά στα προβλήματα επιλογής μεταβλητών. Η SCAD αντιστοιχεί σε μια τετραγωνική συνάρτηση spline με κόμβους στο λ και στο $\alpha\lambda$. Παρατηρούμε ότι η SCAD συμπίπτει με τη lasso μέχρι το $|\beta| = \lambda$, έπειτα μεταβαίνει ομαλά σε μία τετραγωνική συνάρτηση μέχρι το $|\beta| = a\lambda$ και μετά παραμένει σταθερή για κάθε $|\beta| > a\lambda$. Επίσης, για τους μικρούς συντελεστές διατηρεί το βαθμό ποινικοποίησης της lasso, αλλά όλο και μειώνεται ο βαθμός όσο η απόλυτη τιμή του συντελεστή αυξάνεται.

Από τα πλέον σημαντικά χαρακτηριστικά της SCAD, είναι η λεγόμενη «μαντική» ιδιότητα (*oracle property*), κατά την οποία, με την ασυμπτωτική έννοια, αποδίδει το ίδιο καλά όπως εάν γνωρίζαμε εκ των προτέρων ποιοι συντελεστές είναι μηδενικοί και ποιοι όχι (Breheny and Huang, 2011), να γνωρίζαμε δηλαδή το σωστό υπομοντέλο εξαρχής. Οι Fan and Li (2001) αναφέρουν ότι αν η παράμετρος ποινικοποίησης επιλεχθεί σωστά, τότε όταν οι πραγματικές παράμετροι έχουν κάποιες μηδενικές συνιστώσες, αυτές εκτιμώνται ως 0 με πιθανότητα που πλησιάζει το 1, και οι μη μηδενικές συνιστώσες εκτιμώνται ορθώς σα να γνωρίζαμε εκ των προτέρων το σωστό υπομοντέλο. Η ιδιότητα αυτή βελτιώνει όχι μόνο την ακρίβεια για την εκτίμηση των μηδενικών συνιστωσών, αλλά και την εκτίμηση των μη μηδενικών συνιστωσών. Η σημασία αυτής της ιδιότητας είναι ότι οι

προτεινόμενες διαδικασίες υπερβαίνουν τους κλασσικούς εκτιμητές μέγιστης πιθανοφάνειας και αποδίδουν καλύτερα.



ΣΧΗΜΑ 3-16 : Ποινή L_0 (πράσινη γραμμή), ποινή L_1 (μπλε γραμμή) και ποινή SCAD (κόκκινη γραμμή).

Επειδή η συνάρτηση ποινής της SCAD είναι μη-κυρτή (*nonconvex*), υπάρχουν αρκετές αριθμητικές προκλήσεις κατά την προσαρμογή αυτών των μοντέλων. Οι Fan and Li (2001) πρότειναν τον ενοποιημένο αλγόριθμο τοπικής τετραγωνικής προσέγγισης (*local quadratic approximation-LQA*) για την ελαχιστοποίηση της:

$$l^*(\boldsymbol{\beta}) = -l(\boldsymbol{\beta}) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{3.13}$$

ώστε να εκτιμήσουν τους ποινικοποιήμενους εκτιμητές μέγιστης πιθανοφάνειας, δηλαδή να βρουν το $\widehat{\boldsymbol{\beta}}^{SCAD}$ που ελαχιστοποιεί την ποσότητα (3.13):

$$\widehat{\boldsymbol{\beta}}^{SCAD} = \underset{\boldsymbol{\beta}}{argmin}\left[-l(\boldsymbol{\beta}) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|)\right], \tag{3.14}$$

Η ελαχιστοποίηση της σχέσης αυτής όμως, είναι υπολογιστικά δύσκολη. Για τις μη-κυρτές ποινές, οι Zou and Li (2008) πρότειναν μια τοπική γραμμική προσέγγιση (*local linear approximation-LLA*) στην ποινή, καταλήγοντας έτσι σε μια συνάρτηση που μπορεί να βελτιστοποιηθεί χρησιμοποιώντας τον αλγόριθμο παλινδρόμησης ελάχιστης γωνίας (*least*

angle regression-LARS) των Efron et al. (2004), ώστε να υπολογίσουν το path της SCAD. Οι Breheny and Huang (2011) ερευνούν την εφαρμογή αλγορίθμων coordinate descent στα μοντέλα παλινδρόμησης SCAD, όπου και θα δούμε την εφαρμογή αυτή στην περίπτωση της λογιστικής παλινδρόμησης που μας ενδιαφέρει. Το πρόβλημα ελαχιστοποίησης που ορίζουν οι Breheny and Huang είναι το

$$Q_{\lambda,\alpha}(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\{y_i log\pi_\iota + (1-y_i) log\,(1-\pi_i)\} + \sum_{j=1}^{d} p_{\lambda,\alpha}\left(|\beta_j|\right).$$

Η ελαχιστοποίηση μπορεί να προσεγγιστεί λαμβάνοντας πρώτα μια τετραγωνική προσέγγιση της συνάρτησης απώλειας που βασίζεται σε μια επέκταση της σειράς Taylor σχετικά με την τιμή των συντελεστών στην αρχή της τρέχουσας επανάληψης, $\boldsymbol{\beta}^{(m)}$. Πράττοντας κατά αυτόν τον τρόπο, προκύπτει η μορφή του επαναληπτικά επανασταθμισμένου αλγορίθμου ελαχίστων τετραγώνων που χρησιμοποιείται συνήθως για την προσαρμογή ΓΓΜ (McCullagh and Nelder, 1989):

$$Q_{\lambda,\alpha}(\boldsymbol{\beta}) \approx \frac{1}{2n}(\tilde{\boldsymbol{y}} - \boldsymbol{X\beta})^{\boldsymbol{T}}\boldsymbol{W}(\tilde{\boldsymbol{y}} - \boldsymbol{X\beta}) + \sum_{j=1}^{d} p_{\lambda,\alpha}\left(|\beta_j|\right),$$

με $\tilde{\boldsymbol{y}} = \boldsymbol{X\beta}^{(m)} + \boldsymbol{W}^{-1}(\boldsymbol{y} - \boldsymbol{\pi})$, $\boldsymbol{W}$ διαγώνιο πίνακα βαρών με στοιχεία $w_i = \pi_i(1-\pi_i)$ και το $\boldsymbol{\pi}$ αξιολογείται στο $\boldsymbol{\beta}^{(m)}$. Με την αναπαράσταση αυτή, οι αλγόριθμοι LLA και coordinate descent μπορούν να επεκταθούν στη λογιστική παλινδρόμηση με έναν αρκετά απλό τρόπο. Τα βήματα που γίνονται στην επανάληψη m είναι:

1) προσεγγίζεται η συνάρτηση απώλειας με βάση το $\boldsymbol{\beta}^{(m)}$ και
2) εκτελείται μια επανάληψη του αλγορίθμου LLA/coordinate descent, λαμβάνοντας το $\boldsymbol{\beta}^{(m+1)}$.

Έπειτα, τα βήματα αυτά επαναλαμβάνονται έως ότου γίνει σύγκλιση για κάθε τιμή του λ. Να σημειωθεί ότι για κανέναν από τους δύο αλγορίθμους δεν είναι σίγουρη η σύγκλιση για τη λογιστική παλινδρόμηση. Ωστόσο, υπό την προϋπόθεση ότι υπάρχουν επαρκείς διασφαλίσεις για την προστασία από τον κορεσμό του μοντέλου, έχει παρατηρηθεί ότι η αποτυχία σύγκλισης δεν αποτελεί πρόβλημα για τους αλγορίθμους αυτούς.

Η παρουσία βαρών παρατήρησης αλλάζει τη μορφή των αναπροσαρμογών (*updates*) ως προς τις συντεταγμένες. Έστω $u_j = n^{-1}\boldsymbol{x}_j^T\boldsymbol{W}\boldsymbol{x}_j$, κατάλοιπα $\boldsymbol{r} = \boldsymbol{W}^{-1}(\boldsymbol{y} - \boldsymbol{\pi})$ και

$z_j = \frac{1}{n}\boldsymbol{x}_j^T \boldsymbol{W}\left(\tilde{\boldsymbol{y}} - \boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j}\right) = \frac{1}{n}\boldsymbol{x}_j^T \boldsymbol{W}\boldsymbol{r} + u_j\beta_j^{(m)}$. Έτσι, η αναπροσαρμογή του αλγόριθμου coordinate descent είναι:

$$\beta_j \leftarrow \begin{cases} \dfrac{S(z_j,\lambda)}{u_j}, & \alpha\nu\ |z_j| \le \lambda(u+1) \\ \dfrac{S(z_j, a\lambda/(a-1))}{u_j - 1/(a-1)}, & \alpha\nu\ \lambda(u_j+1) < |z_j| \le u_j\alpha\lambda \\ \dfrac{z_j}{u_j}, & \alpha\nu\ |z_j| > u_j\alpha\lambda \end{cases},$$

για $a > 1 + 1/u_j$ και:

$$S(z,\lambda) = \begin{cases} z-\lambda, & \alpha\nu\ z > \lambda \\ 0, & \alpha\nu\ |z| \le \lambda \\ z+\lambda, & \alpha\nu\ z < -\lambda \end{cases}$$

ο τελεστής soft thresholding (*soft thresholding operator*) (Donoho and Johnstone, 1994) για $\lambda \ge 0$. Επίσης, η αναπροσαρμογή των καταλοίπων ορίζεται ως $\boldsymbol{r} \leftarrow \boldsymbol{r} - \left(\beta_j^{(m+1)} - \beta_j^{(m)}\right)\boldsymbol{x}_j$. Όπως μπορούμε να δούμε στην αναπροσαρμογή των β_j, τμήματα του αριθμητή και του παρονομαστή επανασταθμίζονται, με την επαναστάθμιση αυτή να προκαλεί κάποιες δυσκολίες σχετικά με την επιλογή και την ερμηνεία του α. Για αυτό, οι Breheny and Huang προτείνουν μια αναπροσαρμογή της παραμέτρου α, ώστε να ταιριάζει με τη συνεχώς μεταβαλλόμενη κλίμακα των μεταβλητών.

3.7.5 Μέθοδος MCP

Η τελευταία μέθοδος ποινής που θα αναλύσουμε είναι αυτή της MCP (minimax concave penalty). Έχουμε αναφέρει ότι η lasso είναι μία γρήγορη και συνεχής μέθοδος ποινής, όμως δεν είναι αμερόληπτη. Από την άλλη, η subset selection είναι μία αμερόληπτη μέθοδος, αλλά υπολογιστικά ακριβή και απαιτητική. Ο Zhang (2010) πρότεινε την MCP, μία συνεχής, σχεδόν αμερόληπτη και με ακρίβεια μέθοδο για ποινικοποιημένη επιλογή μεταβλητών. Η MCP, όπως και η SCAD, ανήκει στην «οικογένεια» των αναδιπλούμενων κοίλων συναρτήσεων (*folded concave functions*). Ακόμα, η MCP έχει την πολύ χρήσιμη ιδιότητα της «μαντικής», ασυμπτωτικά δηλαδή αποδίδει το ίδιο καλά όπως εάν γνωρίζαμε το σωστό υπομοντέλο εξαρχής.

Ο Zhang ορίζει την MCP ως:

$$\rho(t;\lambda) = \lambda \int_0^t \left(1 - \frac{x}{\alpha\lambda}\right)_+ dx\,,$$

με παράμετρο ποινικοποίησης $\alpha > 0$. Εναλλακτικά, μπορούμε να ορίσουμε την MCP ως (Breheny and Huang, 2011):

$$p_{\lambda,\alpha}(\beta) = \begin{cases} \lambda\beta - \dfrac{\beta^2}{2\alpha} & \beta \le \alpha\lambda \\ \dfrac{1}{2}\alpha\lambda^2 & \beta > \alpha\lambda \end{cases},$$

για $\alpha > 1$, με πρώτη παράγωγο:

$$p'_\lambda(\beta) = \begin{cases} \lambda - \dfrac{\beta}{\alpha} & \beta \le \alpha\lambda \\ 0 & \beta > \alpha\lambda \end{cases}.$$

Γενικά έχει αποδειχτεί ότι το $\alpha = 3$ είναι μια καλή επιλογή τιμής και ότι αποδίδει αρκετά καλά. Βλέπουμε ότι η MCP στην αρχή εφαρμόζει τον ίδιο βαθμό ποινικοποίησης με τη lasso και σιγά σιγά ο βαθμός φθίνει προς το μηδέν καθώς η απόλυτη τιμή του συντελεστή αυξάνεται. Εν αντιθέσει με την SCAD που ο βαθμός ποινικοποίησης παραμένει σταθερός για λίγο πριν αρχίσει να μειώνεται, στην MCP φθίνει απευθείας (Σχήμα 3-18).

Ανάμεσα από όλες τις συναρτήσεις p συνεχώς διαφορίσιμες στο $(0,\infty)$ που ικανοποιούν $p'(0_+|\lambda) = \lambda$ και $p'(\beta|\lambda) = 0$ για κάθε $t \ge \alpha\lambda$, η MCP ελαχιστοποιεί τη μέγιστη κοιλότητα (*maximum concavity*)

$$\kappa = \sup_{0<\beta_1<\beta_2} \frac{p'(\beta_1|\lambda) - p'(\beta_2|\lambda)}{\beta_2 - \beta_1},$$

από όπου προέκυψε και το όνομά της minimax concave. Η MCP έχει κοιλότητα ίση με $\kappa = 1/\alpha$. Μια μεγαλύτερη τιμή του α δίνει λιγότερη μεροληψία και περισσότερη κοιλότητα. Επίσης, για κάθε επίπεδο ποινής λ, η MCP παρέχει μία συνέχεια ποινών, με την ποινή L_1 στο $\alpha = \infty$ και την ποινή L_0 στο $\alpha \to 0+$.

Για την εκτίμηση των $\widehat{\boldsymbol{\beta}}^{MCP}$ στη λογιστική παλινδρόμηση, θα βασιστούμε ξανά στην εργασία των Breheny and Huang (2011). Υπενθυμίζουμε ότι ως πρόβλημα ελαχιστοποίησης ορίζουν το

$$Q_{\lambda,\alpha}(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\{y_i log\pi_\iota + (1-y_i) log\,(1-\pi_i)\} + \sum_{j=1}^{d} p_{\lambda,\alpha}\left(|\beta_j|\right).$$

Όπως έχουμε ήδη αναφέρει στην SCAD, θα χρησιμοποιήσουμε τη μορφή του επαναληπτικά επανασταθμισμένου αλγορίθμου ελαχίστων τετραγώνων:

$$Q_{\lambda,\alpha}(\boldsymbol{\beta}) \approx \frac{1}{2n}(\tilde{\boldsymbol{y}} - \boldsymbol{X\beta})^{\boldsymbol{T}}\boldsymbol{W}(\tilde{\boldsymbol{y}} - \boldsymbol{X\beta}) + \sum_{j=1}^{d} p_{\lambda,\alpha}\left(|\beta_j|\right),$$

με $\tilde{\boldsymbol{y}} = \boldsymbol{X\beta}^{(m)} + \boldsymbol{W}^{-1}(\boldsymbol{y} - \boldsymbol{\pi})$, $\boldsymbol{W}$ διαγώνιο πίνακα βαρών με στοιχεία $w_i = \pi_i(1 - \pi_i)$ και το $\boldsymbol{\pi}$ αξιολογείται στο $\boldsymbol{\beta}^{(m)}$. Τα βήματα στη συνέχεια είναι τα ίδια με πριν.

Παραπάνω ορίσαμε $u_j = n^{-1}\boldsymbol{x}_j^T\boldsymbol{W}\boldsymbol{x}_j$, κατάλοιπα $\boldsymbol{r} = \boldsymbol{W}^{-1}(\boldsymbol{y} - \boldsymbol{\pi})$ και $z_j = \frac{1}{n}\boldsymbol{x}_j^T\boldsymbol{W}\left(\tilde{\boldsymbol{y}} - \boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j}\right) = \frac{1}{n}\boldsymbol{x}_j^T\boldsymbol{W}\boldsymbol{r} + u_j\beta_j^{(m)}$. Η αναπροσαρμογή του αλγόριθμου coordinate descent τώρα είναι:

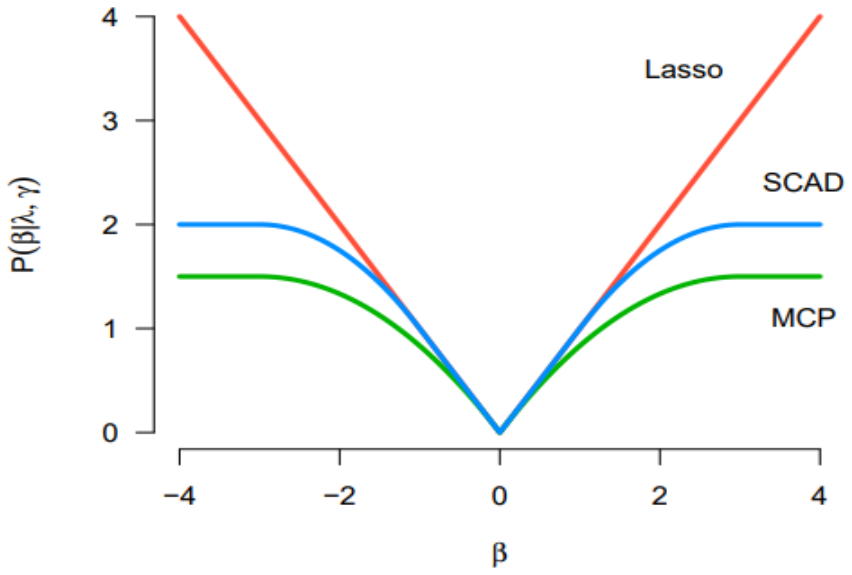
$$\beta_j \leftarrow \begin{cases} \dfrac{S(z_j, \lambda)}{u_j - 1/a}, & \alpha\nu\ |z_j| \le u_j\alpha\lambda \\ \dfrac{z_j}{u_j}, & \alpha\nu\ |z_j| > u_j\alpha\lambda \end{cases},$$

για $a > 1/u_j$ και:

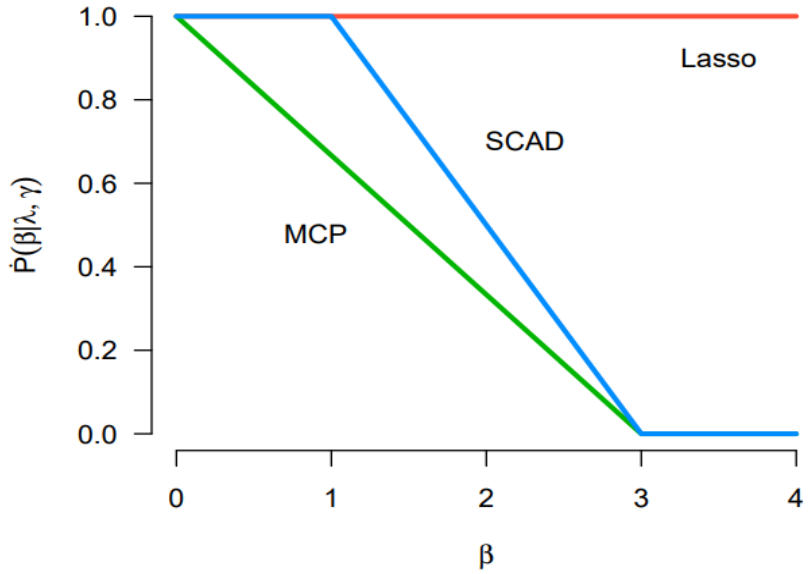
$$S(z, \lambda) = \begin{cases} z - \lambda, & \alpha\nu\ z > \lambda \\ 0, & \alpha\nu\ |z| \le \lambda \\ z + \lambda, & \alpha\nu\ z < -\lambda \end{cases}$$

ο τελεστής μαλακού κατωφλιού για $\lambda \ge 0$. Η αναπροσαρμογή των καταλοίπων πάλι ορίζεται ως $\boldsymbol{r} \leftarrow \boldsymbol{r} - \left(\beta_j^{(m+1)} - \beta_j^{(m)}\right)\boldsymbol{x}_j$. Λόγω της επαναστάθμισης που γίνεται στον αριθμητή και στον παρονομαστή, θα χρειαστεί πάλι να γίνει μια αναπροσαρμογή της παραμέτρου a για να ταιριάζει με τη συνεχώς μεταβαλλόμενη κλίμακα των μεταβλητών.

Το κίνητρο πίσω από την SCAD και την MCP, είναι η εξάλειψη της μεροληψίας για τους μεγάλους συντελεστές. Οπότε είναι εύκολο να καταλάβουμε ότι το πλεονέκτημα των μεθόδων αυτών είναι εμφανές κυρίως όταν κάποιοι μη μηδενικοί συντελεστές είναι μεγάλοι.



ΣΧΗΜΑ 3-17 : Οι συντελεστές ως προς τη συνάρτηση ποινής για τις lasso (κόκκινη γραμμή), SCAD (μπλε γραμμή) και MCP (πράσινη γραμμή).



ΣΧΗΜΑ 3-18 : Οι συντελεστές ως προς την πρώτη παράγωγο των lasso (κόκκινη γραμμή), SCAD (μπλε γραμμή) και MCP (πράσινη γραμμή).

ΚΕΦΑΛΑΙΟ 4

Εφαρμογές των Μεθόδων Ποινικοποίησης σε Πραγματικά Δεδομένα

4.1 Εισαγωγή

Στο 4ο κεφάλαιο θα προχωρήσουμε σε εφαρμογή των μεθόδων ποινικοποίησης που αναλύσαμε στο 3ο κεφάλαιο σε δύο σύνολα πραγματικών δεδομένων. Σκοπός μας είναι να εστιάσουμε στο πως διαφοροποιείται η επιλογή των μεταβλητών στην οποία θα προβεί η κάθε μέθοδος ποινικοποίησης, καθώς και η εκτίμηση των συντελεστών που θα δώσει.

Η γλώσσα προγραμματισμού που θα χρησιμοποιήσουμε είναι η R, και συγκεκριμένα η RStudio. Για αρχή, θα εφαρμόσουμε το λογιστικό μοντέλο. Στη συνέχεια, τις μεθόδους επιλογής μοντέλου forward selection, backward elimination και stepwise regression βασιζόμενοι στο κριτήριο επιλογής AIC. Έπειτα, θα προχωρήσουμε σε εφαρμογή των διαφόρων μεθόδων ποινικοποίησης.

Για την εφαρμογή των ridge, lasso και elastic net θα χρησιμοποιήσουμε τη βιβλιοθήκη glmnet και την ομώνυμη συνάρτηση, όπου η τιμή της παραμέτρου α καθορίζει ποια μέθοδος χρησιμοποιείται. Για $a = 0$ εφαρμόζεται η ridge, για $a = 1$ η lasso και για ενδιάμεσες τιμές η elastic net, με την τιμή του α να ελέγχει την ισορροπία μεταξύ των δύο προαναφερθεισών μεθόδων. Ως αλγόριθμο βελτιστοποίησης, η glmnet() χρησιμοποιεί τον cyclical coordinate descent. Στην περίπτωση που δε θέσουμε εμείς ένα εύρος τιμών για το λ, η glmnet() παράγει από μόνη της και χρησιμοποιεί μία δική της ακολουθία. Η ακολουθία αυτή καθορίζεται από το lambda.max και από το lambda.min.ratio. Το lambda.max υπολογίζεται αυτόματα από τις μεταβλητές που θα εισαχθούν στην glmnet() και δεν μπορεί να το ορίσει ο ερευνητής. Αφορά τη μικρότερη δυνατή τιμή του λ για την οποία όλες οι μεταβλητές θα μηδενιστούν πλήρως. Στην περίπτωση της ridge, γνωρίζουμε ότι αυτό ισχύει για $\lambda = \infty$. Σε αυτήν την περίπτωση, επιλέγεται μία τιμή που αντιστοιχεί σε μία πολύ μικρή τιμή του α κοντά στο 0. Το lambda.min.ratio επιλέγεται αυτόματα από την glmnet() ως εξής. Αν το πλήθος των

παρατηρήσεων είναι μικρότερο από το πλήθος των μεταβλητών, τότε είναι ίσο με 0,01, διαφορετικά παίρνει την τιμή 0,0001. Αν θέλουμε βέβαια, μπορούμε να θέσουμε από μόνοι μας κάποια τιμή για το lambda.min.ratio. Έπειτα, η μικρότερη τιμή της ακολουθίας υπολογίζεται ως lambda.max*lambda.min.ratio. Ανάμεσα στο lambda.max και στο lambda.max*lambda.min.ratio, το πρόγραμμα παράγει ένα πλήθος (η προεπιλογή είναι 100) τιμών, οι οποίες τιμές είναι γραμμικές στη λογαριθμική κλίμακα, και έτσι προκύπτει η ακολουθία των λ που θα χρησιμοποιηθεί. Επίσης, έχουμε αναφέρει ότι είναι σημαντικό πριν την εφαρμογή οποιασδήποτε μεθόδου ποινικοποίησης να γίνεται τυποποίηση των δεδομένων, έτσι ώστε η κάθε συνάρτηση ποινής να θέτει την ίδια ποινή σε κάθε μεταβλητή (βλ. 3.6.2). Η glmnet() έχει ως προεπιλογή την τυποποίηση των μεταβλητών, επομένως δε χρειάζεται να προβούμε σε τυποποίηση πριν αρχίσουμε την ανάλυσή μας.

Για τις μεθόδους SCAD και MCP θα χρησιμοποιήσουμε τη συνάρτηση ncvreg() από την αντίστοιχη βιβλιοθήκη, που χρησιμοποιεί ως αλγόριθμο βελτιστοποίησης τον coordinate descent. Ως α (gamma στην R) θα πάρουμε τις προεπιλεγμένες τιμές της R, δηλαδή $a = 3{,}7$ και $a = 3$ αντίστοιχα. Η ncvreg() παράγει και αυτή τη δική της ακολουθία για το λ, ακριβώς όπως και η glmnet(). Αυτό που αλλάζει μόνο είναι η τιμή του lambda.min.ratio, καθώς αν το πλήθος των παρατηρήσεων είναι μικρότερο από το πλήθος των μεταβλητών, είναι ίσο με 0,05, αλλιώς 0,001. Πάλι, όπως και στην περίπτωση της glmnet(), η ncvreg() έχει ως προεπιλογή την τυποποίηση των μεταβλητών.

Επιπλέον, οι εκτιμήσεις των μεταβλητών που θα προκύψουν από την κάθε μέθοδο ποινικοποίησης, θα πρέπει να υποστηρίζονται από κάποια τυπικά σφάλματα. Όμως, ούτε η glmnet(), ούτε η ncvreg() υπολογίζουν τα τυπικά σφάλματα των εκτιμητών. Για το λόγο αυτό, θα προχωρήσουμε σε εκτίμηση των τυπικών σφαλμάτων μέσω bootstrapping.

Τέλος, η επιλογή της ρυθμιστικής παραμέτρου λ, καθώς και η τιμή του α στην περίπτωση της elastic net, θα γίνει με τη χρήση της διασταυρούμενης επικύρωσης k-πλαισίων για $k = 10$, όπου πριν τη χρήση της διασταυρούμενης επικύρωσης θα τρέχουμε την εντολή set.seed(5), ούτως ώστε να εξασφαλίσουμε ότι κάθε φορά που θα τρέχουμε τον κώδικα, θα λαμβάνουμε τα ίδια αποτελέσματα. Το κριτήριο για την επιλογή της τιμής του λ, εφ'όσον ασχολούμαστε με λογιστική παλινδρόμηση, θα είναι η τιμή αυτή που θα δίνει τη μικρότερη μέση απόκλιση.

4.2 Σύνολο Δεδομένων Burns

4.2.1 Περιγραφή των δεδομένων

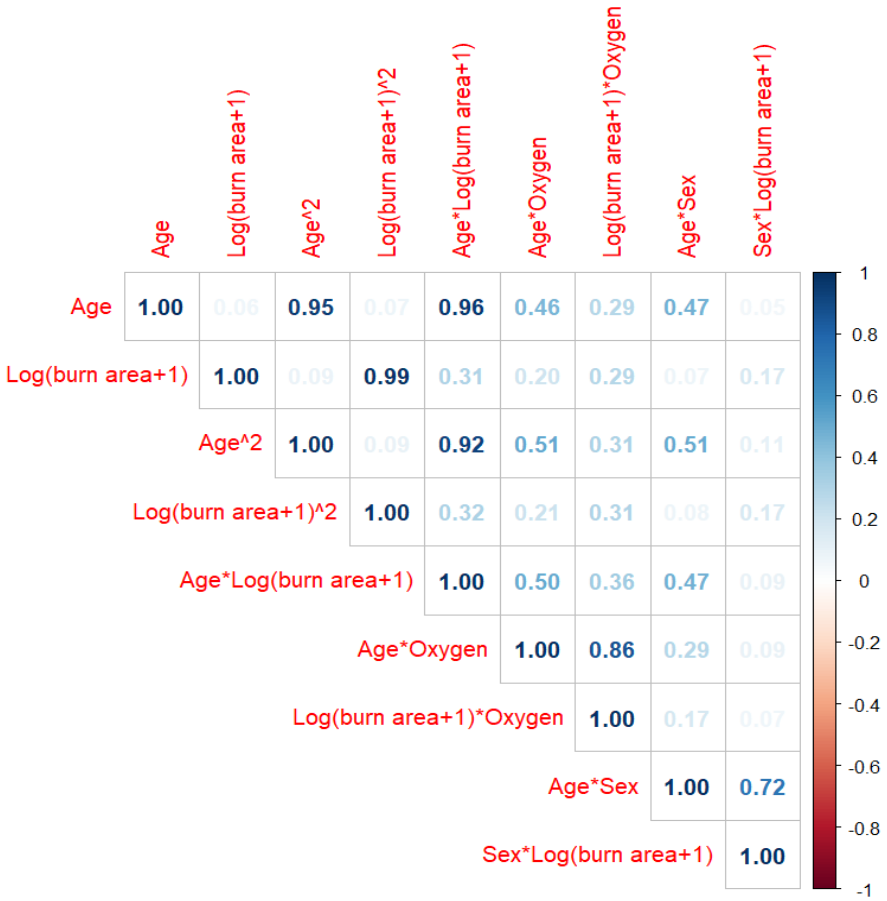
Το 1ο σύνολο δεδομένων με το οποίο και θα ασχοληθούμε είναι το Burns dataset, το οποίο χρησιμοποίησαν και οι Fan and Li (2001) στην εργασία τους. Τα δεδομένα συλλέχθηκαν από το Κέντρο Εγκαυμάτων Γενικού Νοσοκομείου στο Πανεπιστήμιο της Νότιας Καλιφόρνιας. Το σύνολο δεδομένων αποτελείται από 981 παρατηρήσεις. Για λόγους σύγκρισης με την εργασία των Fan and Li, θα χρησιμοποιήσουμε ακριβώς τις ίδιες μεταβλητές που χρησιμοποίησαν και οι ίδιοι. Οι μεταβλητές αυτές είναι:

1) Y. Η εξαρτημένη μεταβλητή Y παίρνει την τιμή 1 αν ο ασθενής επιβίωσε από τα εγκαύματά του, αλλιώς 0. Από τις 981 παρατηρήσεις, οι 826 αφορούν ασθενείς που επιβίωσαν ($Y = 1$) και οι 155 ασθενείς που αποβίωσαν ($Y = 0$).
2) $X_1 = Age$. Η ηλικία του ασθενή.
3) $X_2 = Sex$. Το φύλο του ασθενή.
4) $X_3 = log(burn\ area + 1)$. Μεταβλητή που αφορά το σημείο του εγκαύματος.
5) $X_4 = Oxygen$. Αν $X_4 = 0$, το οξυγόνο είναι φυσιολογικό, αλλιώς 1.

Ακόμα, χρησιμοποιήσαν τετραγωνικούς όρους για τις X_1 και X_3, καθώς και όλες τις πιθανές αλληλεπιδράσεις, δηλαδή $X_1 * X_2$, $X_1 * X_3$, $X_1 * X_4$, $X_2 * X_3$, $X_2 * X_4$ και $X_3 * X_4$. Επομένως, συνολικά θα έχουμε 12 μεταβλητές. Οι Fan and Li χρησιμοποίησαν το MATLAB και εφάρμοσαν το λογιστικό μοντέλο, το AIC, το BIC, την SCAD, τη lasso και τη μέθοδο ποινής hard. Την τιμή του λ την επέλεξαν μέσω γενικευμένης διασταυρούμενης επικύρωσης, οπού και προέκυψε $\lambda = 0{,}6932$, $\lambda = 0{,}0015$ και $\lambda = 0{,}8062$ για τις SCAD, lasso και hard αντίστοιχα. Ως αλγόριθμο βελτιστοποίησης χρησιμοποίησαν τον Local Quadratic Approximation (LQA) και τα τυπικά σφάλματα των μη μηδενικών συντελεστών τα υπολόγισαν με τη χρήση του «sandwich» τύπου.

Λόγω του ότι το σύνολο των μεταβλητών περιέχει τετραγωνικούς όρους και όρους αλληλεπίδρασης, είναι λογικό να αναμένουμε υψηλές συσχετίσεις μεταξύ των μεταβλητών. Πράγματι, παρατηρώντας και το Σχήμα 4-1, μπορούμε να δούμε ότι υπάρχουν αρκετές υψηλές θετικές συσχετίσεις, όπως είναι η Age με την Age^2 (0,95) και η Age*Oxygen με την Log(burn area+1)*Oxygen (0,86). Επομένως, θα μπορούσαμε να πούμε ότι περιμένουμε η ridge και η elastic net να έχουν καλύτερη απόδοση σε αυτά τα δεδομένα, καθώς γνωρίζουμε

ότι η ridge μπορεί να διαχειριστεί αρκετά καλά την ύπαρξη της πολυσυγγραμμικότητας και ότι η elastic net έχει την ιδιότητα του grouping effect.



ΣΧΗΜΑ 4-1 : Γράφημα συσχετίσεων των συνεχών μεταβλητών.

4.2.2 Το λογιστικό μοντέλο

Ξεκινώντας, θα χρησιμοποιήσουμε το μοντέλο της λογιστικής παλινδρόμησης και θα βρούμε τους εκτιμητές μέγιστης πιθανοφάνειας. Οι εκτιμήσεις φαίνονται στον Πίνακα 4-1. Μπορούμε να δούμε ότι στο πλήρες μοντέλο, εκτός από τον σταθερό όρο, μόνο δύο μεταβλητές είναι στατιστικά σημαντικές σε επίπεδο σημαντικότητας $a = 5\%$: η Age και η Age*Log(burn area+1). Επίσης, εκτός από τον σταθερό όρο, την Sex, την Log(burn area+1) και την Oxygen, οι εκτιμήσεις των μεταβλητών είναι μικρές έως και αρκετά μικρές.

ΠΙΝΑΚΑΣ 4-1 : Εκτιμήσεις των μεταβλητών στο λογιστικό μοντέλο
(με * οι στατιστικά σημαντικές μεταβλητές σε επίπεδο σημαντικότητας $\alpha = 5\%$).

ΜΕΤΑΒΛΗΤΗ	**ΕΚΤΙΜΗΣΗ (ΤΥΠΙΚΟ ΣΦΑΛΜΑ)**	**P-VALUE**
Σταθερός όρος	30,93 (13,8)	0,025 *
Age	-0,382 (0,14)	0,006 *
Sex	4,833 (4,3)	0,26
Log(burn area+1)	-2,031 (2,87)	0,48
Oxygen	-4,868 (3,99)	0,22
Age^2	-0,0005 (0,0004)	0,21
$\text{Log(burn area} + 1)^2$	-0,174 (0,16)	0,27
Sex*Oxygen	-0,529 (0,7)	0,45
Age*Log(burn area+1)	0,044 (0,015)	0,003 *
Age*Oxygen	0,012 (0,017)	0,45
Log(burn area+1)*Oxygen	0,45 (0,45)	0,31
Age*Sex	0,0015 (0,016)	0,93
Sex* Log(burn area+1)	-0,64 (0,49)	0,19

4.2.3 Forward selection, backward elimination και stepwise regression

Στη συνέχεια, θα εφαρμόσουμε τις μεθόδους επιλογής μοντέλου forward selection, backward elimination και stepwise regression. Την stepwise regression θα την εφαρμόσουμε μία φορά αρχίζοντας από το πλήρες μοντέλο και μία από το μηδενικό. Τα αποτελέσματα των forward selection και backward elimination φαίνονται στον Πίνακα 4-2. Μπορούμε να δούμε ότι η forward selection επιλέγει τις Age, Log(burn area+1), Oxygen, Age^2, την $\text{Log(burn area} + 1)^2$ και Age*Log(burn area+1). Στατιστικά σημαντικές σε $\alpha = 5\%$ είναι οι Age, Oxygen και Age*Log(burn area+1). Από την άλλη, η backward elimination επιλέγει τις ίδιες πλην της Log(burn area+1), με στατιστικά σημαντικές να είναι όλες εκτός από την Age^2, η οποία είναι οριακά στατιστικά μη σημαντική ($p - value \approx 0{,}053$).

ΠΙΝΑΚΑΣ 4-2 : Επιλογή μεταβλητών με τις μεθόδους forward selection και backward elimination (με * οι στατιστικά σημαντικές μεταβλητές σε επίπεδο σημαντικότητας $\alpha = 5\%$).

ΜΕΤΑΒΛΗΤΗ	FORWARD SELECTION		BACKWARD ELIMINATION	
	ΕΚΤΙΜΗΣΗ (ΤΥΠΙΚΟ ΣΦΑΛΜΑ)	P-VALUE	ΕΚΤΙΜΗΣΗ (ΤΥΠΙΚΟ ΣΦΑΛΜΑ)	P-VALUE
Σταθερός όρος	27,82 (13,32)	0,037 *	20,97 (2,51)	< 2e-16 *
Age	-0,34 (0,13)	0,009 *	-0,286 (0,08)	0,0002 *
Sex	--	--	--	--
Log(burn area+1)	-1,46 (2,78)	0,6	--	--
Oxygen	-0,84 (0,31)	0,006 *	-0,85 (0,31)	0,006 *
Age^2	-0,0005 (0,0003)	0,12	-0,0006 (0,0003)	0,053
$\text{Log(burn area} + 1)^2$	-0,203 (0,15)	0,18	-0,28 (0,033)	< 2e-16 *
Sex*Oxygen	--	--	--	--
Age*Log(burn area+1)	0,04 (0,014)	0,005 *	0,034 (0,008)	1,17e-05 *
Age*Oxygen	--	--	--	--
Log(burn area+1)*Oxygen	--	--	--	--
Age*Sex	--	--	--	--
Sex* Log(burn area+1)	--	--	--	--

Στον Πίνακα 4-3 βλέπουμε τα αποτελέσματα που δίνει η stepwise regression. Είτε ξεκινήσουμε από το πλήρες μοντέλο ή από το μηδενικό, οι μεταβλητές που επιλέγονται είναι οι ίδιες. Η Age^2 είναι οριακά στατιστικά μη σημαντική για $a = 5\%$. Επίσης, παρατηρούμε ότι οι μεταβλητές που επέλεξε η stepwise regression, είναι οι ίδιες με αυτές της backward elimination.

ΠΙΝΑΚΑΣ 4-3 : Επιλογή μεταβλητών με τη μέθοδο stepwise regression
(με * οι στατιστικά σημαντικές μεταβλητές σε επίπεδο σημαντικότητας $\alpha = 5\%$).

ΜΕΤΑΒΛΗΤΗ	STEPWISE REGRESSION (ΑΠΌ ΤΟ ΠΛΗΡΕΣ ΜΟΝΤΕΛΟ)		STEPWISE REGRESSION (ΑΠΌ ΤΟ ΜΗΔΕΝΙΚΟ ΜΟΝΤΕΛΟ)	
	ΕΚΤΙΜΗΣΗ (ΤΥΠΙΚΟ ΣΦΑΛΜΑ)	P-VALUE	ΕΚΤΙΜΗΣΗ (ΤΥΠΙΚΟ ΣΦΑΛΜΑ)	P-VALUE
Σταθερός όρος	20,97 (2,51)	< 2e-16 *	20,97 (2,51)	< 2e-16 *
Age	-0,286 (0,08)	0,0002 *	-0,286 (0,08)	0,0002 *
Sex	--	--	--	--
Log(burn area+1)	--	--	--	--
Oxygen	-0,85 (0,31)	0,006 *	-0,85 (0,31)	0,006 *
Age^2	-0,0006 (0,0003)	0,053	-0,0006 (0,0003)	0,053
$\text{Log(burn area} + 1)^2$	-0,28 (0,033)	< 2e-16 *	-0,28 (0,033)	< 2e-16 *
Sex*Oxygen	--	--	--	--
Age*Log(burn area+1)	0,034 (0,008)	1,17e-05 *	0,034 (0,008)	1,17e-05 *
Age*Oxygen	--	--	--	--
Log(burn area+1)*Oxygen	--	--	--	--
Age*Sex	--	--	--	--
Sex* Log(burn area+1)	--	--	--	--

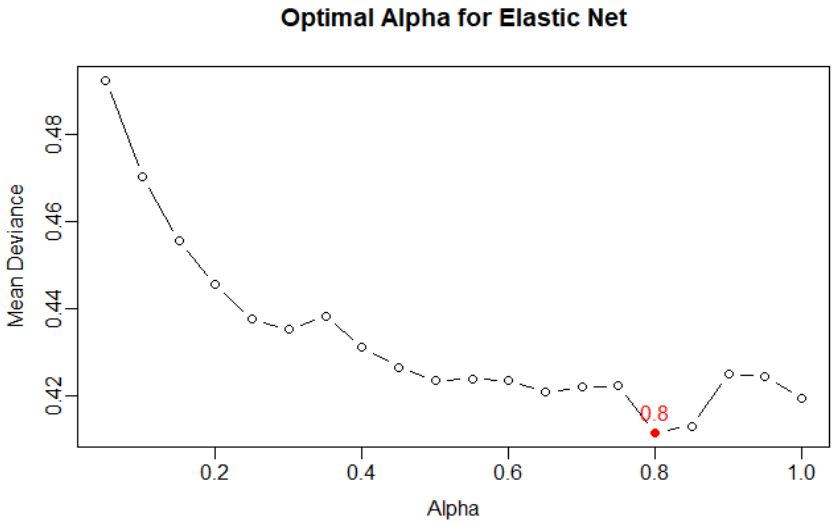
4.2.4 Μέθοδοι ποινικοποίησης

Στην ενότητα αυτή θα εφαρμόσουμε τις μεθόδους ποινικοποίησης στο σύνολο δεδομένων και θα δούμε πως διαφοροποιείται η επιλογή των μεταβλητών ανά μέθοδο. Αρχίζοντας, η R για το lambda.max μας δίνει τις τιμές 188,71, 0,189, 0,236, 0,189 και 0,189 για τις ridge, lasso, elastic net, SCAD και MCP αντίστοιχα. Όπως έχουμε ήδη αναφέρει, για αυτές τις τιμές του λ η εκάστοτε ποινή μηδενίζει πλήρως όλες τις μεταβλητές.

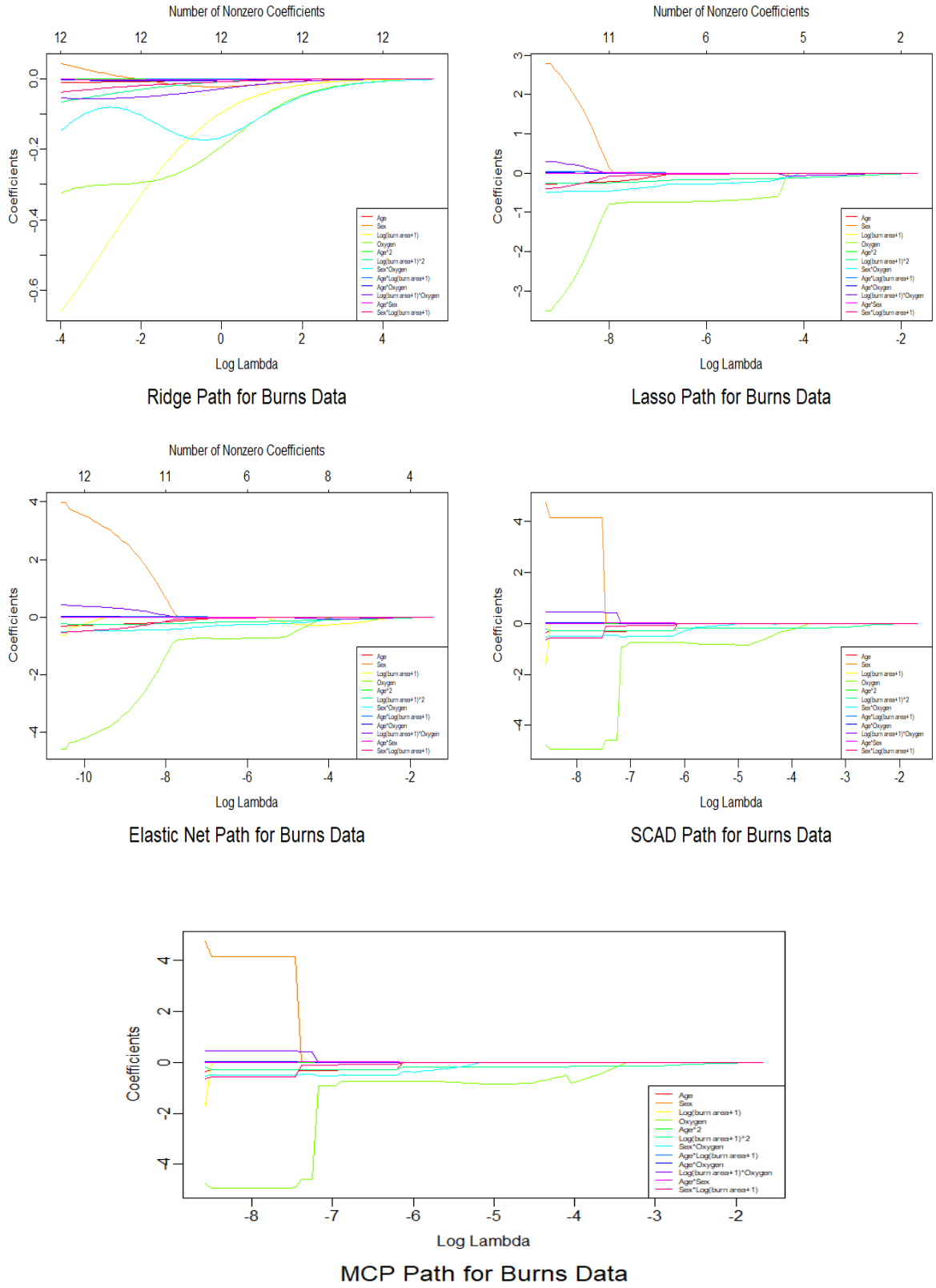
Στο Σχήμα 4-3 μπορούμε να δούμε τα regularization paths των μεθόδων ποινικοποίησης. Για την elastic net, έπειτα από διασταυρούμενη επικύρωση k-πλαισίων για $k = 10$ και για α από 0,05 έως 1 με βήμα 0,05, ως βέλτιστη τιμή, αυτή που ελαχιστοποιεί τη μέση απόκλιση δηλαδή, επιλέχθηκε το $\alpha = 0{,}8$ (Σχήμα 4-2). Παρατηρούμε ότι τα paths των lasso και elastic net μεταξύ τους και τα paths των SCAD και MCP μεταξύ τους είναι αρκετά παρόμοια.

Έπειτα, μέσω διασταυρούμενης επικύρωσης k-πλαισίων για $k = 10$, θα επιλέξουμε την τιμή του λ για την κάθε ποινή ξεχωριστά. Ως λ θα επιλέξουμε αυτό για το οποίο

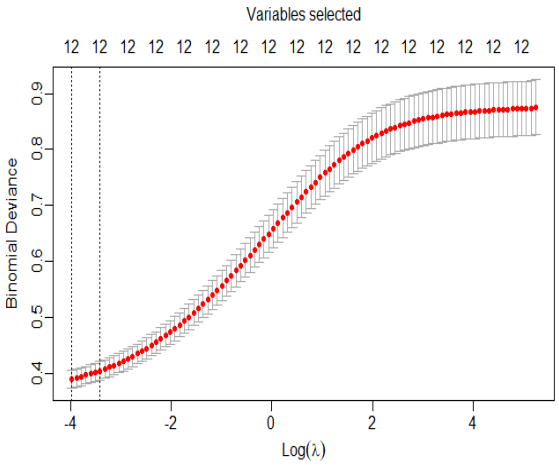
ελαχιστοποιείται η μέση απόκλιση (lambda.min). Στο Σχήμα 4-4 βλέπουμε τη μέση απόκλιση για τις διάφορες τιμές του λ. Η πρώτη κάθετη γραμμή υποδεικνύει το $log(\lambda)$ το οποίο ελαχιστοποιεί τη μέση απόκλιση διασταυρούμενης επικύρωσης, αυτό δηλαδή το οποίο θα επιλέξουμε για τη συνέχεια. Οι τιμές του λ είναι 0,0189, 0,0004, 0,00015, 0,0015 και 0,0018 για τις ridge, lasso, elastic net, SCAD και MCP αντίστοιχα. Βλέπουμε ότι οι τιμές του λ είναι αρκετά μικρές, επομένως μπορούμε να εικαστούμε ότι η επίδραση των μεθόδων ποινικοποίησης δε θα είναι αρκετά μεγάλη. Επίσης, στα γραφήματα για τις ridge, lasso και elastic net παρατηρούμε ότι υπάρχει μία δεύτερη κάθετη γραμμή, η οποία αντιστοιχεί στο λεγόμενο lambda.1se. Το lambda.1se είναι η μεγαλύτερη τιμή του λ τέτοια ώστε η μέση απόκλιση να είναι εντός ενός τυπικού σφάλματος από τη μέση απόκλιση του lambda.min.



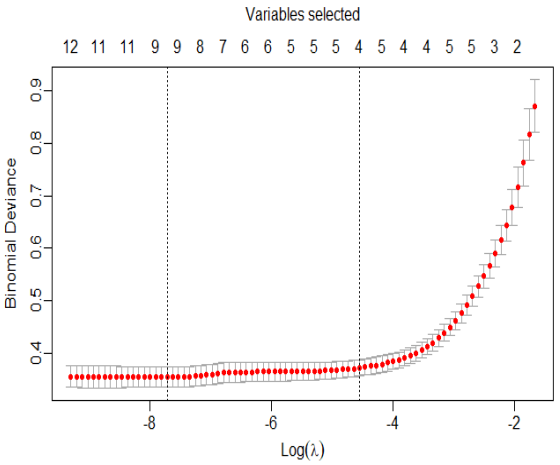
ΣΧΗΜΑ 4-2 : Μέση απόκλιση για τις διάφορες τιμές του α της elastic net μέσω διασταυρούμενης επικύρωσης k-πλαισίων για $k = 10$.



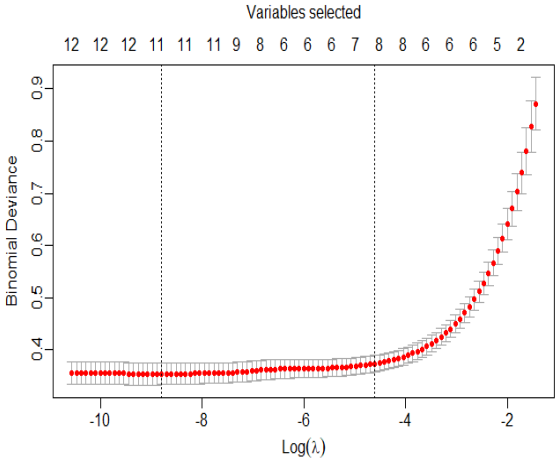
ΣΧΗΜΑ 4-3 : Regularization path για τις ridge, lasso, elastic net, SCAD και MCP.



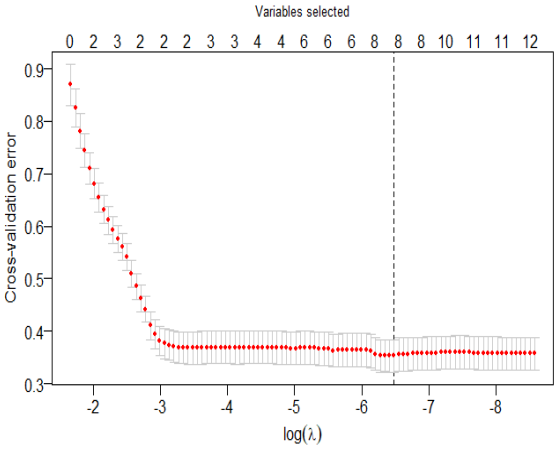
Cross-Validation Deviance of Ridge



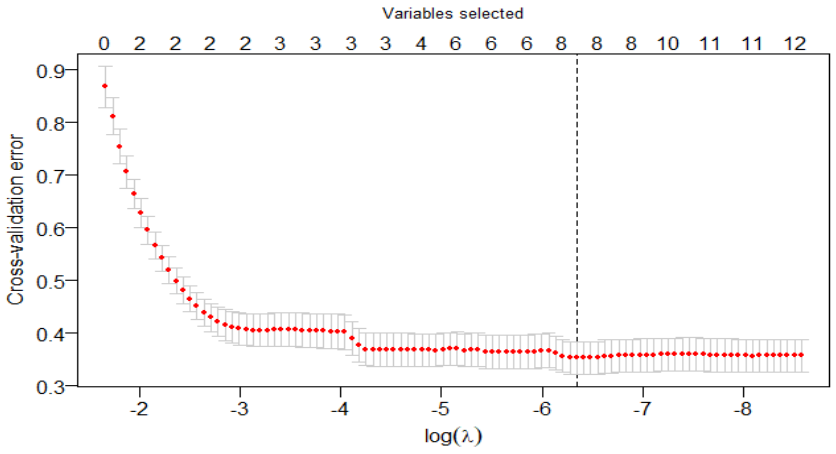
Cross-Validation Deviance of Lasso



Cross-Validation Deviance of Elastic Net



Cross-Validation Deviance of SCAD



Cross-Validation Deviance of MCP

ΣΧΗΜΑ 4-4 : Μέση απόκλιση για τις διάφορες τιμές του λ για τις ridge, lasso, elastic net, SCAD και MCP (η φορά του άξονα χ στις SCAD και MCP απεικονίζεται ανάποδα).

ΠΙΝΑΚΑΣ 4-4 : Εκτιμήσεις των μεταβλητών και τα τυπικά σφάλματα μέσω των μεθόδων ποινικοποίησης.

ΜΕΤΑΒΛΗΤΗ	**RIDGE**	**LASSO**	**ELASTIC NET**	**SCAD**	**MCP**
Σταθερός όρος	12,125 (0,57)	18,119 (1,86)	19,187 (3,006)	28,108 (9,076)	28,107 (1,73)
Age	-0,01 (0,002)	-0,191 (0,067)	-0,231 (0,06)	-0,359 (0,11)	-0,359 (0,025)
Sex	0,043 (0,09)	--	2,448 (3,13)	--	--
Log(burn area+1)	-0,66 (0,05)	--	--	-1,392 (1,76)	-1,392 (0,5)
Oxygen	-0,322 (0,15)	-0,754 (2,26)	-2,924 (3,48)	-0,741 (4,32)	-0,741 (1,85)
Age^2	-0,0003 (0,0004)	-0,0007 (0,0003)	-0,0007 (0,0003)	-0,0005 (0,0003)	0,0005 (0,0002)
$\text{Log(burn area} + 1)^2$	-0,065 (0,038)	-0,238 (0,027)	-0,252 (0,043)	-0,21 (0,094)	-0,21 (0,038)
Sex*Oxygen	-0,146 (0,37)	-0,42 (0,7)	-0,462 (0,76)	-0,503 (0,778)	-0,503 (0,59)
Age*Log(burn area+1)	-0,002 (0,0004)	0,023 (0,007)	0,028 (0,007)	0,041 (0,012)	0,041 (0,0035)
Age*Oxygen	-0,003 (0,005)	0,0005 (0,013)	0,007 (0,015)	--	--
Log(burn area+1)*Oxygen	-0,055 (0,024)	--	0,236 (0,4)	--	--
Age*Sex	-0,001 (0,005)	0,008 (0,01)	0,004 (0,013)	0,013 (0,016)	0,013 (0,009)
Sex* Log(burn area+1)	-0,038 (0,021)	-0,063 (0,23)	-0,354 (0,37)	-0,095 (0,53)	-0,095 (0,15)

Στον παραπάνω πίνακα έχουμε τις εκτιμήσεις των μεταβλητών μαζί με τα αντίστοιχα τυπικά τους σφάλματα, τα οποία σφάλματα εκτιμήθηκαν μέσω bootstrapping, όπως αναφέραμε και στην εισαγωγή. Τη μεγαλύτερη συρρίκνωση των συντελεστών την έκανε η ridge. Η lasso, η SCAD και η MCP έδωσαν τα πιο αραιά μοντέλα, καθώς αφαίρεσαν τρεις μεταβλητές: η lasso τις Sex, Log(burn area+1) και Log(burn area+1)*Oxygen, ενώ η SCAD και η MCP τις Sex,

Age*Oxygen και Log(burn area+1)*Oxygen. Η SCAD και η MCP, έπειτα από στρογγυλοποιήσεις, έχουν δώσει τις ίδιες εκτιμήσεις, με την MCP όμως να δίνει αρκετά μικρότερα τυπικά σφάλματα. Η elastic net από την άλλη έχει αφαιρέσει μόνο την Log(burn area+1), η οποία είναι σχεδόν ταυτόσημη με την $\text{Log(burn area} + 1)^2$ (Σχήμα 4-1, συσχέτιση 0,99). Λόγω των αρκετά υψηλών συσχετίσεων που υπάρχουν μεταξύ των μεταβλητών, του ότι η elastic net έχει την ικανότητα του grouping effect και του ότι δίνει πιο αραιό μοντέλο από την ridge κατά μία μεταβλητή, είναι αυτή που θα επιλέγαμε για τη συνέχεια της έρευνάς μας για το συγκεκριμένο σύνολο δεδομένων.

4.2.5 Σύγκριση των αποτελεσμάτων

Στον Πίνακα 4-5 είναι τα αποτελέσματα στα οποία κατέληξαν οι Fan and Li (2001) στην εργασία τους. Συγκρίνοντας τα αποτελέσματά τους με τα δικά μας, όπως αυτά απεικονίζονται στους Πίνακες 4-1, 4-2, 4-3 και 4-4, παρατηρούμε ότι τα αποτελέσματα διαφέρουν αρκετά. Π.χ. η SCAD και η lasso δίνει αρκετά πιο αραιά μοντέλα σε σχέση με αυτά που καταλήξαμε εμείς. Ακόμα και οι εκτιμητές μέγιστης πιθανοφάνειας είναι πολύ διαφορετικοί. Πιθανοί λόγοι για τις διαφορές αυτές είναι ότι: 1ον) η δική μας ανάλυση έγινε με χρήση της R, ενώ η ανάλυση των Fan and Li έγινε στο MATLAB, 2ον) χρησιμοποιήσαμε έτοιμες βιβλιοθήκες της R οι οποίες πιθανόν να διαφέρουν αρκετά από τον κώδικα που έτρεξαν οι Fan and Li και 3ον) οι glmnet() και ncvreg() χρησιμοποιούν τον cyclical coordinate descent και coordinate descent αντίστοιχα ως αλγόριθμο βελτιστοποίησης, ενώ οι Fan and Li χρησιμοποίησαν τον LQA. Επομένως, οι διαφορές αυτές, μεταξύ άλλων που μπορεί να υπάρχουν, είναι αρκετές ώστε να υπάρξει αρκετά μεγάλη διαφοροποίηση μεταξύ των αποτελεσμάτων.

ΠΙΝΑΚΑΣ 4-5 : Τα αποτελέσματα των Fan and Li (2001).

Method	MLE	Best Subset (AIC)	Best Subset (BIC)	SCAD	LASSO	Hard
Intercept	5.51 (.75)	4.81 (.45)	6.12 (.57)	6.09 (.29)	3.70 (.25)	5.88 (.41)
X_1	−8.83 (2.97)	−6.49 (1.75)	−12.15 (1.81)	−12.24 (.08)	0 (—)	−11.32 (1.1)
X_2	2.30 (2.00)	0 (—)	0 (—)	0 (—)	0 (—)	2.21 (1.41)
X_3	−2.77 (3.43)	0 (—)	−6.93 (.79)	−7.00 (.21)	0 (—)	−4.23 (.64)
X_4	−1.74 (1.41)	.30 (.11)	−.29 (.11)	0 (—)	−.28 (.09)	−1.16 (1.04)
X_1^2	−.75 (.61)	−1.04 (.54)	0 (—)	0 (—)	−1.71 (.24)	0 (—)
X_3^2	−2.70 (2.45)	−4.55 (.55)	0 (—)	0 (—)	−2.67 (.22)	−1.92 (.95)
X_1X_2	.03 (.34)	0 (—)	0 (—)	0 (—)	0 (—)	0 (—)
X_1X_3	7.46 (2.34)	5.69 (1.29)	9.83 (1.63)	9.84 (.14)	.36 (.22)	9.06 (.96)
X_1X_4	.24 (.32)	0 (—)	0 (—)	0 (—)	0 (—)	0 (—)
X_2X_3	−2.15 (1.61)	0 (—)	0 (—)	0 (—)	−0.10 (.10)	−2.13 (1.27)
X_2X_4	−.12 (.16)	0 (—)	0 (—)	0 (—)	0 (—)	0 (—)
X_3X_4	1.23 (1.21)	0 (—)	0 (—)	0 (—)	0 (—)	.82 (1.01)

4.3 Σύνολο Δεδομένων Breast Cancer Wisconsin (Diagnostic)

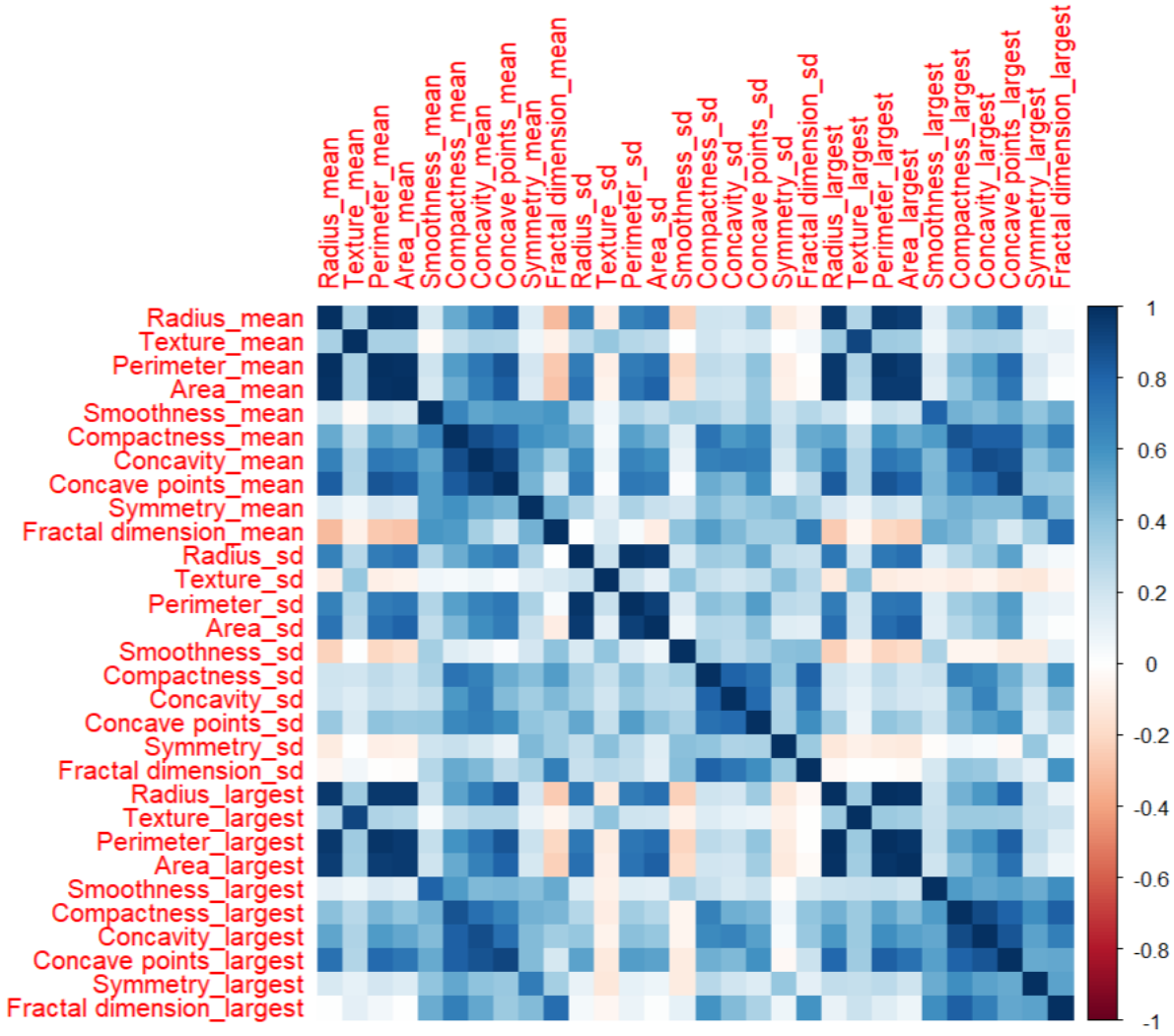
4.3.1 Περιγραφή των δεδομένων

Το 2ο σύνολο δεδομένων που θα χρησιμοποιήσουμε είναι το Breast Cancer Wisconsin (Diagnostic) (ανασύρθηκε από: UCI Machine Learning Repository). Το σύνολο αυτό περιέχει μεταβλητές όπου προέρχονται από φηφιοποιημένες εικόνες αναρροφήσεων λεπτής βελόνας (*fine-needle aspiration, FNA*) όγκων του μαστού και περιγράφουν χαρακτηριστικά των κυτταρικών πυρήνων, ώστε να γίνει η διάγνωση του καρκίνου του μαστού. Το FNA αποτελεί μία διαγνωστική διαδικασία για τη διερεύνηση εξογκωμάτων και όγκων. Τα δεδομένα αποτελούνται από 569 παρατηρήσεις και 32 μεταβλητές. Οι μεταβλητές αυτές είναι:

1) ID number.
2) Y. Η εξαρτημένη μεταβλητή Y παίρνει την τιμή M (malignant) αν ο όγκος είναι κακοήθης και B (benign) αν είναι καλοήθης. Για την ανάλυση, θα θέσουμε $M = 1$ και $B = 0$. Από τις 569 παρατηρήσεις, οι 212 αφορούν κακοήθη όγκο και οι 357 καλοήθη.

Επίσης, για κάθε κυτταρικό πυρήνα έχουν υπολογιστεί τα εξής δέκα χαρακτηριστικά: 1) Radius, 2) Texture, 3) Perimeter, 4) Area, 5) Smoothness, 6) Compactness, 7) Concavity, 8) Concave Points, 9) Symmetry και 10) Fractal dimension. Η μέση τιμή, η τυπική απόκλιση και η «χειρότερη» ή μεγαλύτερη (μέση τιμή των τριών μεγαλύτερων τιμών) τιμή αυτών των χαρακτηριστικών υπολογίστηκε για κάθε εικόνα, καταλήγοντας έτσι σε 30 μεταβλητές. Π.χ. η X_1 αφορά την μέση τιμή της radius, η X_{11} την τυπική της απόκλιση και η X_{21} τη μεγαλύτερή της τιμή.

Στο Σχήμα 4-5 μπορούμε να δούμε μία διαφορετική απεικόνιση των συσχετίσεων μεταξύ των μεταβλητών. Όσο πιο σκούρο μπλε είναι το αντίστοιχο τετράγωνο, τόσο μεγαλύτερη η θετική συσχέτιση μεταξύ των δύο μεταβλητών αυτών, ενώ όσο πιο σκούρο κόκκινο είναι, τόσο μεγαλύτερη η αρνητική συσχέτιση. Οι αρνητικές συσχετίσεις παρατηρούμε ότι είναι λίγες και μη ισχυρές. Από την άλλη, οι θετικές συσχετίσεις είναι πολλές και κάποιες μάλιστα αρκετά ισχυρές, όπως η συσχέτιση μεταξύ της Perimeter_mean με την Radius_mean και αυτή της Perimeter_largest με την Area_mean.



ΣΧΗΜΑ 4-5 : Γράφημα συσχετίσεων των συνεχών μεταβλητών.

4.3.2 Forward selection και stepwise regression

Στον Πίνακα 4-6 φαίνονται οι μεταβλητές που επέλεξαν οι forward selection και stepwise regression από το μηδενικό μοντέλο, καθώς και οι εκτιμήσεις αυτών, τα τυπικά σφάλματα και τα αντίστοιχα p-values. Οι αντίστοιχες τιμές για το μοντέλο της λογιστικής παλινδρόμησης, της backward elimination και της stepwise regression από το πλήρες μοντέλο δεν παρατίθενται, καθώς δίνουν αρκετά μεγάλες εκτιμήσεις και μεγάλα τυπικά σφάλματα, δίχως καμία συνεισφορά στην ανάλυσή μας. Από την άλλη, η forward selection και η stepwise regression που ξεκινάει τον έλεγχο από το μηδενικό μοντέλο, δίνουν αρκετά καλά αποτελέσματα. Αντίστοιχα έχουν κρατήσει στο μοντέλο 11 και 10 μεταβλητές με λογικές εκτιμήσεις και τυπικά σφάλματα.

ΠΙΝΑΚΑΣ 4-6 : Επιλογή μεταβλητών με τις μεθόδους forward selection και stepwise regression (με * οι στατιστικά σημαντικές μεταβλητές σε επίπεδο σημαντικότητας $\alpha = 5\%$).

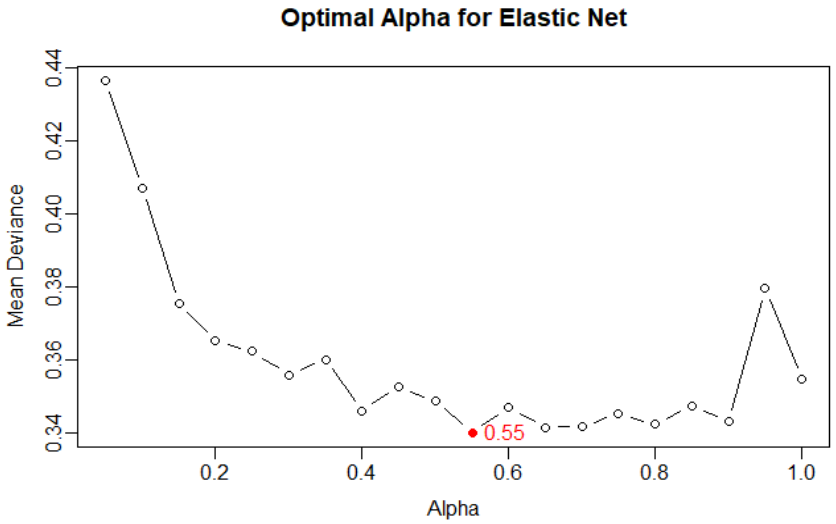
ΜΕΤΑΒΛΗΤΗ	**FORWARD SELECTION**		**STEPWISE REGRESSION (ΑΠΌ ΤΟ ΜΗΔΕΝΙΚΟ ΜΟΝΤΕΛΟ)**	
	ΕΚΤΙΜΗΣΗ (ΤΥΠΙΚΟ ΣΦΑΛΜΑ)	**P-VALUE**	**ΕΚΤΙΜΗΣΗ (ΤΥΠΙΚΟ ΣΦΑΛΜΑ)**	**P-VALUE**
Σταθερός όρος	-44,29 (11,47)	0,0001 *	-44,38 (8,93)	0,000 *
Radius_mean	--	--	--	--
Texture_mean	--	--	--	--
Perimeter_mean	--	--	--	--
Area_mean	--	--	--	--
Smoothness_mean	--	--	--	--
Compactness_mean	--	--	--	--
Concavity_mean	39,87 (16,31)	0,014 *	37,37 (14,2)	0,008 *
Concave points_mean	--	--	--	--
Symmetry_mean	--	--	--	--
Fractal dimension_mean	--	--	--	--
Radius_sd	-9,94 (16,81)	0,55	--	--
Texture_sd	-2,68 (1,52)	0,077	-2,61 (1,47)	0,08
Perimeter_sd	--	--	--	--
Area_sd	0,35 (0,2)	0,08	0,23 (0,068)	0,001 *
Smoothness_sd	--	--	--	--
Compactness_sd	-130,42 (40,78)	0,001 *	-116,57 (36,23)	0,001 *
Concavity_sd	--	--	--	----
Concave points_sd	--	--	--	--
Symmetry_sd	--	--	--	--
Fractal dimension_sd	--	--	--	--
Radius_largest	--	--	--	--
Texture_largest	0,53 (0,14)	0,0001 *	0,54 (0,13)	0,000 *
Perimeter_largest	0,08 (0,07)	0,26	--	--
Area_largest	--	--	0,01 (0,003)	0,006 *
Smoothness_largest	56,53 (30,08)	0,06	65,54 (30,0)	0,03 *
Compactness_largest	--	--	--	--

Concavity_largest	--	--	--	--
Concave points_largest	36,51 (19,42)	0,06	39,18 (19,51)	0,045 *
Symmetry_largest	16,97 (8,01)	0,034 *	15,61 (7,47)	0,04 *
Fractal dimension_largest	--	--	--	--

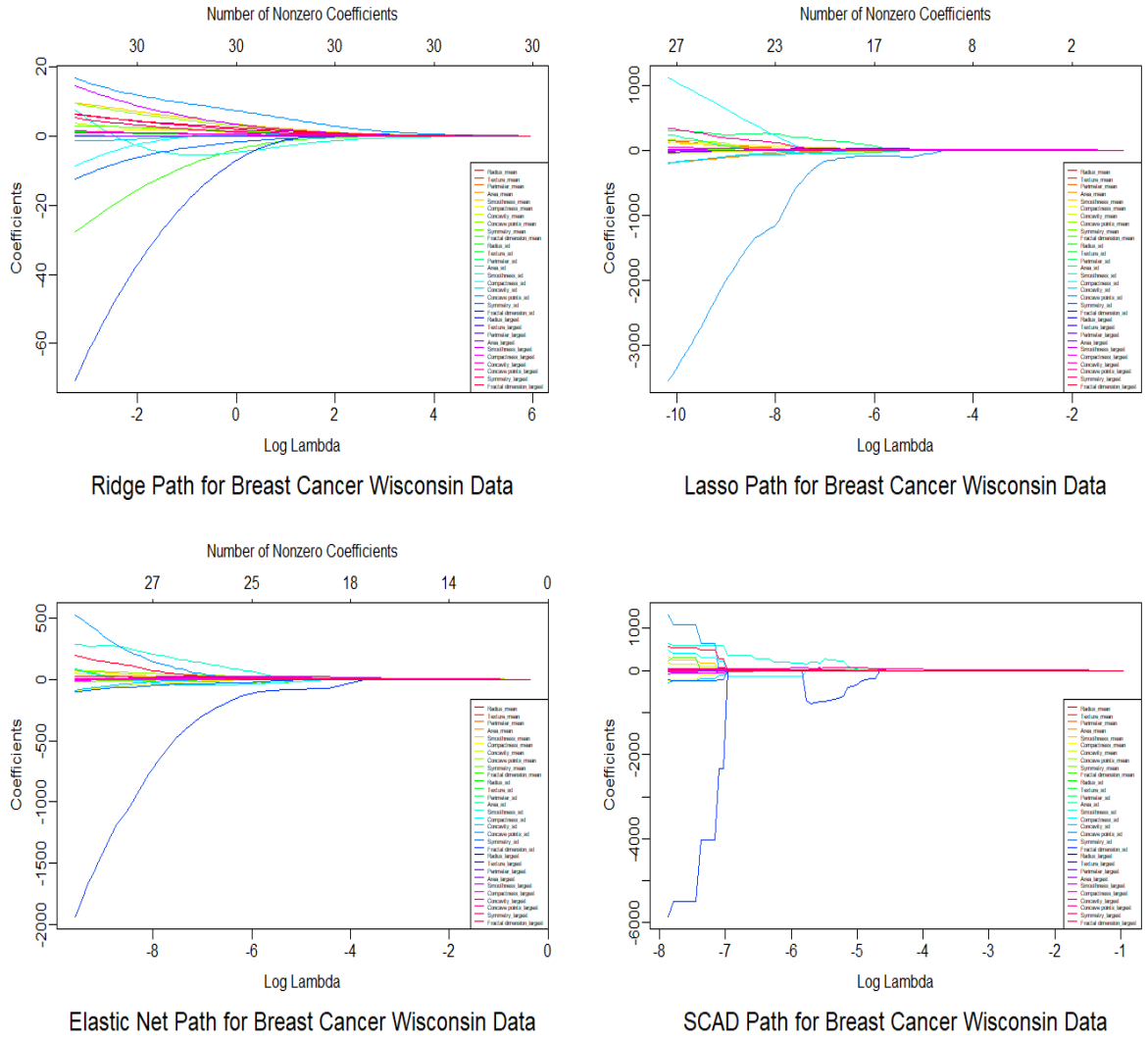
4.3.3 Μέθοδοι ποινικοποίησης

Στη συνέχεια, εφαρμόζουμε τις μεθόδους ποινικοποίησης. Για τα lambda.max έχουμε τις τιμές 383,68, 0,384, 0,698, 0,384 και 0,384 για τις ridge, lasso, elastic net, SCAD και MCP αντίστοιχα. Για την elastic net, μέσω διασταυρούμενης επικύρωσης k-πλαισίων επιλέχθηκε $\alpha = 0{,}55$ (Σχήμα 4-6). Στο Σχήμα 4-7 μπορούμε να δούμε τα regularization paths των μεθόδων ποινικοποίησης. Παρατηρούμε ότι το ridge path σε σχέση με τα υπόλοιπα, ξεκινάει από αρκετά μικρότερες εκτιμήσεις στον άξονα y. Γεγονός που πιθανόν να οφείλεται στο ότι η ridge μπορεί να διαχειριστεί την ύπαρξη της πολυσυγγραμμικότητας πολύ καλύτερα από τις υπόλοιπες μεθόδους ποινικοποίησης.

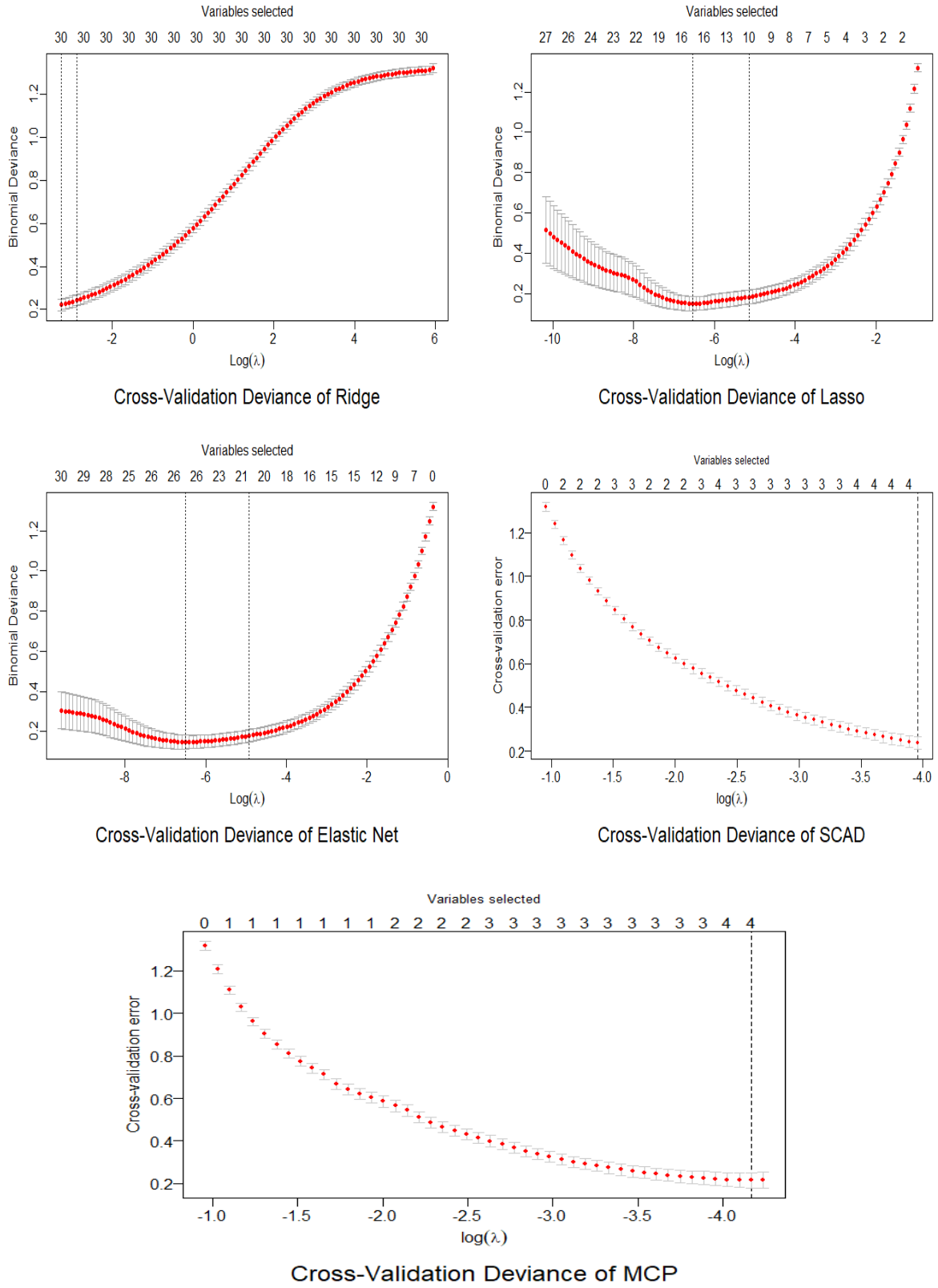
Στη συνέχεια, μέσω διασταυρούμενης επικύρωσης k-πλαισίων για $k = 10$, θα επιλέξουμε την τιμή του λ για την κάθε ποινή ξεχωριστά. Οι τιμές του λ είναι 0,038, 0,0014, 0,0015, 0,019 και 0,015 για τις ridge, lasso, elastic net, SCAD και MCP αντίστοιχα. Ο Πίνακας 4-7 περιέχει τις εκτιμήσεις των συντελεστών για την κάθε μέθοδο. Παρατηρούμε ότι η ridge έχει δώσει και μικρές εκτιμήσεις και κατά κύριο λόγο μικρότερα τυπικά σφάλματα από τις υπόλοιπες μεθόδους. Όμως, το βασικό της μειονέκτημα είναι εμφανές· κρατάει όλες τις μεταβλητές εντός του μοντέλου. Ένα μοντέλο με 31 μεταβλητές, συμπεριλαμβανομένου και του σταθερού όρου, κάθε άλλο παρά φειδωλό είναι. Τα πιο φειδωλά μοντέλα τα δίνουν οι SCAD και MCP με 7 και 6 μεταβλητές αντιστοίχως. Η MCP δίνει λίγο μεγαλύτερες εκτιμήσεις από την SCAD, μικρότερα τυπικά σφάλματα όμως. Η elastic net που γνωρίζουμε ότι και αυτή διαχειρίζεται αρκετά καλά την ύπαρξη της πολυσυγγραμμικότητας, δίνει μοντέλο με 27 μεταβλητές, καθώς μηδένισε τις Smoothness_mean, Symmetry_mean, Concavity_sd και Compactness_largest. Τέλος, η lasso κατέληξε σε μοντέλο με 18 μεταβλητές. Οπότε αν λάβουμε υπ'όψιν μας το μέγεθος των εκτιμήσεων, τα τυπικά σφάλματα και το πλήθος των μεταβλητών, η MCP δείχνει να είναι μια αρκετά καλή επιλογή ποινής στο σύνολο δεδομένων αυτών.



ΣΧΗΜΑ 4-6 : Μέση απόκλιση για τις διάφορες τιμές του α της elastic net μέσω διασταυρούμενης επικύρωσης k-πλαισίων για $k = 10$.



ΣΧΗΜΑ 4-7 : Regularization path για τις ridge, lasso, elastic net, SCAD και MCP.



ΣΧΗΜΑ 4-8 : Μέση απόκλιση για τις διάφορες τιμές του λ για τις ridge, lasso, elastic net, SCAD και MCP (η φορά του άξονα χ στις SCAD και MCP απεικονίζεται ανάποδα).

ΠΙΝΑΚΑΣ 4-7 : Εκτιμήσεις των μεταβλητών και τα τυπικά σφάλματα μέσω των μεθόδων ποινικοποίησης.

ΜΕΤΑΒΛΗΤΗ	**RIDGE**	**LASSO**	**ELASTIC NET**	**SCAD**	**MCP**
Σταθερός όρος	-17,03 (0,83)	-29,49 (6,13)	-31,24 (3,4)	-32,25 (9,07)	-43,7 (1,14)
Radius_mean	0,097 (0,008)	--	0,027 (0,063)	1,09 (0,55)	1,4 (0,028)
Texture_mean	0,08 (0,012)	--	0,053 (0,07)	0,21 (0,13)	0,33 (0,02)
Perimeter_mean	0,014 (0,0011)	--	0,003 (0,008)	--	--
Area_mean	0,0009 (0,00008)	--	0,0006 (0,0006)	--	--
Smoothness_mean	9,61 (3,19)	--	--	--	--
Compactness_mean	0,96 (0,72)	--	-7,73 (7,16)	--	--
Concavity_mean	3,61 (0,43)	1,4 (11,47)	9,04 (4,66)	--	--
Concave points_mean	9,44 (0,85)	32,64 (25,72)	29,24 (10,33)	--	--
Symmetry_mean	2,68 (1,97)	--	--	--	--
Fractal dimension_mean	-27,71 (6,4)	-20,4 (44,72)	-25,57 (41,32)	--	--
Radius_sd	1,26 (0,24)	10,48 (4,68)	6,07 (1,88)	0,91 (1,54)	1,92 (0,34)
Texture_sd	-0,045 (0,12)	-0,85 (0,87)	-0,61 (0,68)	--	--
Perimeter_sd	0,13 (0,03)	--	0,18 (0,18)	--	--
Area_sd	0,006 (0,0016)	--	0,025 (0,009)	--	--
Smoothness_sd	7,6 (23,23)	125,96 (101,3)	98,29 (102,11)	--	--

Compactness_sd	-8,68 (2,5)	-52,85 (38,57)	-44,18 (25,09)	--	--
Concavity_sd	-1,46 (1,37)	--	--	--	--
Concave points_sd	16,78 (7,7)	--	20,0 (42,49)	--	--
Symmetry_sd	-12,39 (7,54)	-9,11 (36,7)	-30,44 (34,44)	--	--
Fractal dimension_sd	-70,55 (16,63)	-103,61 (231,42)	-187,1 (152,16)	--	--
Radius_largest	0,089 (0,0055)	0,2 (0,44)	0,28 (0,06)	--	--
Texture_largest	0,073 (0,0086)	0,31 (0,1)	0,25 (0,07)	--	--
Perimeter_largest	0,012 (0,0007)	0,0007 (0,04)	0,03 (0,01)	--	--
Area_largest	0,0007 (0,00006)	0,007 (0,004)	0,002 (0,0004)	--	--
Smoothness_largest	14,47 (2,34)	23,24 (22,57)	29,4 (17,8)	73,6 (38,51)	88,7 (3,53)
Compactness_largest	1,005 (0,23)	--	--	--	--
Concavity_largest	1,52 (0,2)	5,64 (3,65)	4,01 (1,84)	1,37 (1,95)	--
Concave points_largest	6,39 (0,58)	19,95 (13,14)	16,55 (6,36)	--	--
Symmetry_largest	5,22 (0,97)	10,42 (6,35)	12,87 (5,51)	5,16 (4,0)	14,11 (1,46)
Fractal dimension_largest	6,29 (2,5)	--	9,78 (22,74)	--	--

Π Α Ρ Α Ρ Τ Η Μ Α

Παρακάτω παρατίθεται ο κώδικας που χρησιμοποιήθηκε στο 4ο κεφάλαιο για τα δύο σύνολα δεδομένων.

Α) Σύνολο Δεδομένων Burns

```
dataset.1 <- read.table('C:\\Users\\ Desktop\\Dataset\\burn.dat', sep = '') # load dataset 1

## libraries to use
library(MASS)   # For stepAIC function
library(glmnet) # For ridge, lasso and elastic net
library(ncvreg) # For SCAD and MCP
library(corrplot) # For multicollinearity plot
library(boot) # For bootstrapping

## data preparation
y <- dataset.1[,30]-1 #i ndependent variable
u <- dataset.1[,2] # age
x3 <- dataset.1[,3] # sex
x10 <- dataset.1[,10] # burn area
x90 <- log(x10+1) # log(burn area+1)
x27 <- dataset.1[,33] # oxygen 0 = normal , 1 = abnormal

## data frame with the variables to use
X_1 <- data.frame(u, x3, x90, x27, u^2, x90^2, x3*x27, u*x90, u*x27, x90*x27, u*x3, x3*x90)
column_names_X_1 <- c('Age', 'Sex', 'Log(burn area+1)', 'Oxygen', 'Age^2', 'Log(burn area+1)^2', 'Sex*Oxygen',
'Age*Log(burn area+1)', 'Age*Oxygen', 'Log(burn area+1)*Oxygen', 'Age*Sex', 'Sex*Log(burn area+1)')
 colnames(X_1) <- column_names_X_1
X <- as.matrix(X_1)
table(y) ##count of 0s and 1s
## check and plot multicollinearity
X_continuous <- X[,-c(2,4,7)] # exclude the categorical variables
X_continuous_corr <- cor(X_continuous)
corrplot(X_continuous_corr, method = 'number', type = 'upper')

dataset.1.log.reg. <- glm(y~X, family = 'binomial') # logistic regression
```

```
summary(dataset.1.log.reg.)

## data preparation for forward selection, backward elimination and stepwise regression
data.1 <- data.frame(y, X)
column_names_data.1 <- c('Survived', 'Age', 'Sex', 'Log(burn area+1)', 'Oxygen', 'Age^2', 'Log(burn area+1)^2', 'Sex*Oxygen', 'Age*Log(burn area+1)', 'Age*Oxygen', 'Log(burn area+1)*Oxygen', 'Age*Sex', 'Sex*Log(burn area+1)')
colnames(data.1) <- column_names_data.1
null_model <- glm(Survived~1, data = data.1, family = 'binomial')
full_model <- glm(Survived~., data = data.1, family = 'binomial')

## forward selection
forward_selection <- stepAIC(null_model, scope = list(lower = null_model, upper = full_model), direction = 'forward')
summary(forward_selection)

## backward elimination
backward_elimination <- stepAIC(full_model, direction = 'backward')
summary(backward_elimination)

## stepwise regression
stepwise_regression <- stepAIC(full_model,direction = 'both') # from full model
summary(stepwise_regression)
stepwise_regression.2 <- stepAIC(null_model, scope = list(lower = null_model, upper = full_model), direction = 'both') # from null model
summary(stepwise_regression.2)

## ridge
ridge_model <- glmnet(X, y, family = 'binomial', alpha = 0)
ridge_model$lambda[1] # lambda for which all variables are 0
## plot ridge path
par(mgp = c(2, 0.5, 0))
colors <- rainbow(ncol(X))
plot(ridge_model,  xvar = 'lambda', col = colors)
legend('bottomright', legend = colnames(X), col = colors, lty = 1, cex = 0.5)
mtext('Number of Nonzero Coefficients', side = 3, line = 2)
mtext('Ridge Path for Burns Data', side = 1, line = 4, cex = 1.5)

## cross validation for optimal lambda
set.seed(5)
cv_ridge <- cv.glmnet(X, y, family = 'binomial', alpha = 0,nfolds = 10, type.measure = 'deviance')
print(cv_ridge$lambda.min)
```

```
## plot cross validation deviance
plot(cv_ridge)
mtext('Variables selected', side = 3, line = 2)
mtext('Cross-Validation Deviance of Ridge', side = 1, line = 4, cex = 1.5)

## coefficients for optimal lambda
ridge_model_opt <- glmnet(X, y, family = 'binomial', alpha = 0, lambda = cv_ridge$lambda.min)
ridge_coef <- coef(ridge_model_opt)
print(ridge_coef)


## estimate standard errors with bootstrapping
ridge_logistic <- function(data,indices){
  d <- data[indices,]
  x <- as.matrix(d[,-ncol(d)])
  y <- d[, ncol(d)]
  model <- glmnet(x, y, family = 'binomial', alpha = 0, lambda = cv_ridge$lambda.min)
  return(as.vector(coef(model)))
}

data <- cbind(data.1[,-1],'Survived'= data.1[,1]) # data preparation for bootstrapping
set.seed(5)
n_iterations <- 1000 # Number of bootstrap iterations
ridge_results <- boot(data = data, statistic = ridge_logistic, R = n_iterations)
ridge_std_errors <- apply(ridge_results$t, 2, sd)
names(ridge_std_errors) <- c('Intercept', colnames(X))
round(ridge_std_errors,5)

## lasso
lasso_model <- glmnet(X, y, family = 'binomial', alpha = 1)
lasso_model$lambda[1] # lambda for which all variables are 0

## plot lasso  path
par(mgp = c(2, 0.5, 0))
plot(lasso_model,  xvar = 'lambda', col = colors)
legend('bottomright', legend = colnames(X), col = colors, lty = 1, cex = 0.5)
mtext('Number of Nonzero Coefficients', side = 3, line = 2)
mtext('Lasso Path for Burns Data', side = 1, line = 4, cex = 1.5)

## cross validation for optimal lambda
set.seed(5)
cv_lasso <- cv.glmnet(X, y, family = 'binomial', alpha = 1, nfolds = 10, type.measure = 'deviance')
```

```
print(cv_lasso$lambda.min)

## plot cross validation deviance
plot(cv_lasso)
mtext('Variables selected', side = 3, line = 2)
mtext('Cross-Validation Deviance of Lasso', side = 1, line = 4, cex = 1.5)

## coefficients for optimal lambda
lasso_model_opt <- glmnet(X, y, family = 'binomial', alpha = 1, lambda = cv_lasso$lambda.min)
lasso_coef <- coef(lasso_model_opt)
print(lasso_coef)


## estimate standard errors with bootstrapping
lasso_logistic <- function(data,indices){
  d <- data[indices,]
  x <- as.matrix(d[,-ncol(d)])
  y <- d[, ncol(d)]
  model <- glmnet(x, y, family = 'binomial', alpha = 1, lambda = cv_lasso$lambda.min)
  return(as.vector(coef(model)))
}

set.seed(5)
lasso_results <- boot(data = data, statistic = lasso_logistic, R = n_iterations)
lasso_std_errors <- apply(lasso_results$t, 2, sd)
names(lasso_std_errors) <- c('Intercept', colnames(X))
round(lasso_std_errors,5)

## elastic net
## check for optimal alpha
alpha_values <- seq(0.05, 1, by = 0.05)
mean_deviance <- numeric(length(alpha_values))


## cross validation for each alpha
set.seed(5)
for (i in 1:length(alpha_values)) {
  cv_model <- cv.glmnet(X, y, family = 'binomial', alpha = alpha_values[i], nfolds = 10, type.measure = 'deviance')
  mean_deviance[i] <- mean(cv_model$cvm)
}
optimal_alpha_index <- which.min(mean_deviance)
optimal_alpha <- alpha_values[optimal_alpha_index]
```

```
## plot the mean deviance for each alpha
plot(alpha_values, mean_deviance, type = 'b', xlab = 'Alpha', ylab = 'Mean Deviance', main = 'Optimal Alpha for Elastic Net')


## highlight the optimal alpha
points(optimal_alpha, mean_deviance[optimal_alpha_index], col = 'red', pch = 19)
text(optimal_alpha, mean_deviance[optimal_alpha_index], labels = round(optimal_alpha, 2), pos = 3, col = 'red')

## elastic net model
elastic_net_model <- glmnet(X, y, family = 'binomial', alpha = optimal_alpha)
elastic_net_model$lambda[1] # lambda for which all variables are 0

## plot elastic net  path
par(mgp = c(2, 0.5, 0))
plot(elastic_net_model,  xvar = 'lambda', col = colors)
legend('bottomright', legend = colnames(X), col = colors, lty = 1, cex = 0.5)
mtext('Number of Nonzero Coefficients', side = 3, line = 2)
mtext('Elastic Net Path for Burns Data', side = 1, line = 4, cex = 1.5)

## cross validation for optimal lambda
set.seed(5)
cv_elastic_net <- cv.glmnet(X,y,family = 'binomial', alpha = optimal_alpha, nfolds = 10, type.measure = 'deviance')
print(cv_elastic_net$lambda.min)

## plot cross validation deviance
plot(cv_elastic_net)
mtext('Variables selected', side = 3, line = 2)
mtext('Cross-Validation Deviance of Elastic Net', side = 1, line = 4, cex = 1.5)

## coefficients for optimal lambda
elastic_net_model_opt <- glmnet(X, y, family = 'binomial', alpha = optimal_alpha, lambda = cv_elastic_net$lambda.min)
elastic_net_coef <- coef(elastic_net_model_opt)
print(elastic_net_coef)

## estimate standard errors with bootstrapping
elastic_net_logistic <- function(data,indices){
  d <- data[indices,]
  x <- as.matrix(d[,-ncol(d)])
  y <- d[, ncol(d)]
  model <- glmnet(x, y, family = 'binomial', alpha = optimal_alpha, lambda = cv_elastic_net$lambda.min)
```

```
  return(as.vector(coef(model)))
}

set.seed(5)
elastic_net_results <- boot(data = data, statistic = elastic_net_logistic, R = n_iterations)
elastic_net_errors <- apply(elastic_net_results$t, 2, sd)
names(elastic_net_errors) <- c('Intercept', colnames(X))
round(elastic_net_errors,5)

## SCAD
scad_model <- ncvreg(X, y, family = 'binomial', penalty = 'SCAD', gamma = 3.7)
scad_model$lambda[1] # lambda for which all variables are 0

## plot SCAD path
log_lambda_scad <- log(scad_model$lambda)
coefficients_scad <- as.matrix(scad_model$beta[-1, ])  # Remove the intercept
par(mgp = c(2, 0.5, 0))
matplot(log_lambda_scad, t(coefficients_scad), type = 'l', lty = 1, col = colors, xlab = 'Log Lambda', ylab = 'Coefficients')
mtext('SCAD Path for Burns Data', side = 1, line = 4, cex = 1.5)
legend('bottomright', legend = colnames(X), col = colors, lty = 1, cex = 0.5)

## cross validation for optimal lambda
set.seed(5)
cv_scad <- cv.ncvreg(X,y, family = 'binomial', penalty = 'SCAD', nfolds = 10, gamma = 3.7, criteria = 'deviance')
print(cv_scad$lambda.min)

## plot cross validation deviance
plot(cv_scad)
mtext('Cross-Validation Deviance of SCAD', side = 1, line = 4, cex = 1.5)
## coefficients for optimal lambda
scad_model_opt <- ncvreg(X, y, family = 'binomial', penalty = 'SCAD', gamma = 3.7, lambda = cv_scad$lambda.min)
scad_coef <- coef(scad_model_opt)
print(scad_coef)

## estimate standard errors with bootstrapping
scad_logistic <- function(data,indices){
  d <- data[indices,]
  x <- as.matrix(d[,-ncol(d)])
  y <- d[, ncol(d)]
  model <- ncvreg(x, y, family = 'binomial', penalty = 'SCAD', gamma = 3.7, lambda = cv_scad$lambda.min)
```

```
  return(as.vector(coef(model)))
}

set.seed(5)
scad_results <- boot(data = data, statistic = scad_logistic, R = n_iterations)
scad_errors <- apply(scad_results$t, 2, sd)
names(scad_errors) <- c('Intercept',colnames(X))
round(scad_errors,5)

## MCP
mcp_model <- ncvreg(X, y, family = 'binomial', penalty = 'MCP', gamma = 3)
mcp_model$lambda[1] #lambda for which all variables are 0

## plot MCP path
log_lambda_mcp <- log(mcp_model$lambda)
coefficients_mcp <- as.matrix(mcp_model$beta[-1, ])  # Remove the intercept
par(mgp = c(2, 0.5, 0))
matplot(log_lambda_mcp, t(coefficients_mcp), type = 'l', lty = 1, col = colors,  xlab = 'Log Lambda', ylab = 'Coefficients')
mtext('MCP Path for Burns Data', side = 1, line = 4, cex = 1.5)
legend('bottomright', legend = colnames(X), col = colors, lty = 1, cex = 0.5)


## cross validation for optimal lambda
set.seed(5)
cv_mcp <- cv.ncvreg(X, y, family = 'binomial', penalty = 'MCP', nfolds = 10, gamma = 3, criteria = 'deviance')
print(cv_mcp$lambda.min)

## plot cross validation deviance
plot(cv_mcp)
mtext('Cross-Validation Deviance of MCP', side = 1, line = 4, cex = 1.5)

## coefficients for optimal lambda
mcp_model_opt <- ncvreg(X, y, family = 'binomial', penalty = 'MCP', gamma = 3, lambda = cv_mcp$lambda.min)
mcp_model_opt <- coef(mcp_model_opt)
print(mcp_model_opt)

## estimate standard errors with bootstrapping
mcp_logistic <- function(data,indices){
  d <- data[indices,]
  x <- as.matrix(d[,-ncol(d)])
  y <- d[, ncol(d)]
  model <- ncvreg(x, y, family = 'binomial', penalty = 'MCP', gamma = 3, lambda = cv_mcp$lambda.min)
```

```
  return(as.vector(coef(model)))
}
set.seed(5)
mcp_results <- boot(data = data, statistic = mcp_logistic, R = n_iterations)
mcp_errors <- apply(mcp_results$t, 2, sd)
names(mcp_errors) <- c('Intercept',colnames(X))
round(mcp_errors,5)
```

Β) Σύνολο Δεδομένων Breast Cancer Wisconsin (Diagnostic)

```
dataset.2 <- read.table('C:\\Users\\ Desktop\\Dataset2\\wdbc.data', sep = ',') # load dataset 2

## libraries to use
library(MASS)   # For stepAIC function
library(glmnet) # For ridge, lasso and elastic net
library(ncvreg) # For SCAD and MCP
library(corrplot) # For multicollinearity plot
library(boot) # For bootstrapping

## data preparation
y <- dataset.2[,2] # independent variable
y[which(y=="M")] <- 1
y[which(y=="B")] <- 0
y <- as.factor(y)
X_1 <- dataset.2[,-c(1,2)]
column_names_X_1 <- c('Radius_mean', 'Texture_mean', 'Perimeter_mean', 'Area_mean', 'Smoothness_mean', 'Compactness_mean', 'Concavity_mean', 'Concave points_mean', 'Symmetry_mean', 'Fractal dimension_mean', 'Radius_sd', 'Texture_sd', 'Perimeter_sd', 'Area_sd', 'Smoothness_sd', 'Compactness_sd', 'Concavity_sd', 'Concave points_sd', 'Symmetry_sd', 'Fractal dimension_sd', 'Radius_largest', 'Texture_largest', 'Perimeter_largest', 'Area_largest', 'Smoothness_largest', 'Compactness_largest', 'Concavity_largest', 'Concave points_largest', 'Symmetry_largest', 'Fractal dimension_largest')
colnames(X_1) <- column_names_X_1
X <- as.matrix(X_1)
table(y) # count of 0s and 1s

## check and plot multicollinearity
X_corr <- cor(X)
corrplot(X_corr, method = 'color')

## data preparation for forward selection and stepwise regression
data.2 <- data.frame(y, X)
```

```
column_names_data.2 <- c('Diagnosis', 'Radius_mean', 'Texture_mean', 'Perimeter_mean', 'Area_mean', 'Smoothness_mean', 'Compactness_mean', 'Concavity_mean', 'Concave
points_mean', 'Symmetry_mean', 'Fractal dimension_mean', 'Radius_sd', 'Texture_sd', 'Perimeter_sd', 'Area_sd', 'Smoothness_sd', 'Compactness_sd', 'Concavity_sd', 'Concave points_sd', 'Symmetry_sd', 'Fractal dimension_sd', 'Radius_largest', 'Texture_largest', 'Perimeter_largest', 'Area_largest', 'Smoothness_largest', 'Compactness_largest', 'Concavity_largest', 'Concave points_largest', 'Symmetry_largest', 'Fractal dimension_largest')
colnames(data.2) <- column_names_data.2
null_model <- glm(Diagnosis~1, data = data.2, family = 'binomial')
full_model <- glm(Diagnosis~., data = data.2, family = 'binomial')

## forward selection
forward_selection <- stepAIC(null_model,scope = list(lower = null_model, upper = full_model), direction = 'forward')
summary(forward_selection)

## stepwise regression from null model
stepwise_regression <- stepAIC(null_model, scope = list(lower = null_model, upper = full_model), direction = 'both') # from null model
round(summary(stepwise_regression)$coefficients[,1],3) #coefficients
round(summary(stepwise_regression)$coefficients[,2],3) #standard errors
round(summary(stepwise_regression)$coefficients[,4],3) #p-values

## ridge
ridge_model <- glmnet(X, y, family = 'binomial', alpha = 0)
ridge_model$lambda[1] # lambda for which all variables are 0
## plot ridge path
par(mgp = c(2, 0.5, 0))
colors <- rainbow(ncol(X))
plot(ridge_model,  xvar = 'lambda', col = colors)
legend('bottomright', legend = colnames(X), col = colors, lty = 1, cex = 0.35)
mtext('Number of Nonzero Coefficients', side = 3, line = 2)
mtext('Ridge Path for Breast Cancer Wisconsin Data', side = 1, line = 4, cex = 1.5)

## cross validation for optimal lambda
set.seed(5)
cv_ridge <- cv.glmnet(X, y, family = 'binomial', alpha = 0,nfolds = 10, type.measure = 'deviance')
print(cv_ridge$lambda.min)

## plot cross validation deviance
plot(cv_ridge)
```

```
mtext('Variables selected', side = 3, line = 2)
mtext('Cross-Validation Deviance of Ridge', side = 1, line = 4, cex = 1.5)

## coefficients for optimal lambda
ridge_model_opt <- glmnet(X, y, family = 'binomial', alpha = 0,lambda = cv_ridge$lambda.min)
ridge_coef <- coef(ridge_model_opt)
round(print(ridge_coef),5)

## estimate standard errors with bootstrapping
ridge_logistic <- function(data,indices){
  d <- data[indices,]
  x <- as.matrix(d[,-ncol(d)])
  y <- d[, ncol(d)]
  model <- glmnet(x, y, family = 'binomial', alpha = 0, lambda = cv_ridge$lambda.min)
  return(as.vector(coef(model)))
}

data <- cbind(data.2[,-1],'Diagnosis'= data.2[,1]) # data preparation for bootstrapping
set.seed(5)
n_iterations <- 1000 # Number of bootstrap iterations
ridge_results <- boot(data = data, statistic = ridge_logistic, R = n_iterations)
ridge_std_errors <- apply(ridge_results$t, 2, sd)
names(ridge_std_errors) <- c('Intercept', colnames(X))
round(ridge_std_errors,5)

## lasso
lasso_model <- glmnet(X, y, family = 'binomial', alpha = 1)
lasso_model$lambda[1] # lambda for which all variables are 0

## plot lasso  path
par(mgp = c(2, 0.5, 0))
plot(lasso_model,  xvar = 'lambda', col = colors)
legend('bottomright', legend = colnames(X), col = colors, lty = 1, cex = 0.35)
mtext('Number of Nonzero Coefficients', side = 3, line = 2)
mtext('Lasso Path for Breast Cancer Wisconsin Data', side = 1, line = 4, cex = 1.5)

## cross validation for optimal lambda
set.seed(5)
cv_lasso <- cv.glmnet(X, y, family = 'binomial', alpha = 1, nfolds = 10, type.measure = 'deviance')
print(cv_lasso$lambda.min)

## plot cross validation deviance
```

```
plot(cv_lasso)
mtext('Variables selected', side = 3, line = 2)
mtext('Cross-Validation Deviance of Lasso', side = 1, line = 4, cex = 1.5)

## coefficients for optimal lambda
lasso_model_opt <- glmnet(X, y, family = 'binomial', alpha = 1, lambda = cv_lasso$lambda.min)
lasso_coef <- coef(lasso_model_opt)
round(print(lasso_coef),5)


## estimate standard errors with bootstrapping
lasso_logistic <- function(data,indices){
  d <- data[indices,]
  x <- as.matrix(d[,-ncol(d)])
  y <- d[, ncol(d)]
  model <- glmnet(x, y, family = 'binomial', alpha = 1, lambda = cv_lasso$lambda.min)
  return(as.vector(coef(model)))
}

set.seed(5)
lasso_results <- boot(data = data, statistic = lasso_logistic, R = n_iterations)
lasso_std_errors <- apply(lasso_results$t, 2, sd)
names(lasso_std_errors) <- c('Intercept', colnames(X))
round(lasso_std_errors,5)

## elastic net
## check for optimal alpha
alpha_values <- seq(0.05, 1, by = 0.05)
mean_deviance <- numeric(length(alpha_values))

## cross validation for each alpha
set.seed(5)
for (i in 1:length(alpha_values)) {
  cv_model <- cv.glmnet(X, y, family = 'binomial',
                        alpha = alpha_values[i], nfolds = 10, type.measure = 'deviance')
  mean_deviance[i] <- mean(cv_model$cvm)
}
optimal_alpha_index <- which.min(mean_deviance)
optimal_alpha <- alpha_values[optimal_alpha_index]

## plot the mean deviance for each alpha
```

```
plot(alpha_values, mean_deviance, type = 'b', xlab = 'Alpha', ylab = 'Mean Deviance', main = 'Optimal Alpha for Elastic Net')

## highlight the optimal alpha
points(optimal_alpha, mean_deviance[optimal_alpha_index], col = 'red', pch = 19)
text(optimal_alpha, mean_deviance[optimal_alpha_index], labels = round(optimal_alpha, 2), pos = 4, col = 'red')

## elastic net model
elastic_net_model <- glmnet(X, y, family = 'binomial', alpha = optimal_alpha)
elastic_net_model$lambda[1] # lambda for which all variables are 0

## plot elastic net  path
par(mgp = c(2, 0.5, 0))
plot(elastic_net_model,  xvar = 'lambda', col = colors)
legend('bottomright', legend = colnames(X), col = colors, lty = 1, cex = 0.35)
mtext('Number of Nonzero Coefficients', side = 3, line = 2)
mtext('Elastic Net Path for Breast Cancer Wisconsin Data', side = 1, line = 4, cex = 1.5)

## cross validation for optimal lambda
set.seed(5)
cv_elastic_net <- cv.glmnet(X,y,family = 'binomial', alpha = optimal_alpha, nfolds = 10, type.measure = 'deviance')
print(cv_elastic_net$lambda.min)

## plot cross validation deviance
plot(cv_elastic_net)
mtext('Variables selected', side = 3, line = 2)
mtext('Cross-Validation Deviance of Elastic Net', side = 1, line = 4, cex = 1.5)

## coefficients for optimal lambda
elastic_net_model_opt <- glmnet(X, y, family = 'binomial', alpha = optimal_alpha, lambda = cv_elastic_net$lambda.min)
elastic_net_coef <- coef(elastic_net_model_opt)
round(print(elastic_net_coef),5)

## estimate standard errors with bootstrapping
elastic_net_logistic <- function(data,indices){
  d <- data[indices,]
  x <- as.matrix(d[,-ncol(d)])
  y <- d[, ncol(d)]
  model <- glmnet(x, y, family = 'binomial', alpha = optimal_alpha, lambda = cv_elastic_net$lambda.min)
  return(as.vector(coef(model)))
```

```
}

set.seed(5)
elastic_net_results <- boot(data = data, statistic = elastic_net_logistic, R = n_iterations)
elastic_net_errors <- apply(elastic_net_results$t, 2, sd)
names(elastic_net_errors) <- c('Intercept',colnames(X))
round(elastic_net_errors,5)

## SCAD
scad_model <- ncvreg(X, y, family = 'binomial', penalty = 'SCAD', gamma = 3.7)
scad_model$lambda[1] #lambda for which all variables are 0

## plot SCAD path
log_lambda_scad <- log(scad_model$lambda)
coefficients_scad <- as.matrix(scad_model$beta[-1, ])  # Remove the intercept
par(mgp = c(2, 0.5, 0))
matplot(log_lambda_scad, t(coefficients_scad), type = 'l', lty = 1, col = colors, xlab = 'Log Lambda', ylab = 'Coefficients')
mtext('SCAD Path for Breast Cancer Wisconsin Data', side = 1, line = 4, cex = 1.5)
legend('bottomright', legend = colnames(X), col = colors, lty = 1, cex = 0.35)

## cross validation for optimal lambda
set.seed(5)
cv_scad <- cv.ncvreg(X,y, family = 'binomial', penalty = 'SCAD', nfolds = 10, gamma = 3.7, criteria = 'deviance')
print(cv_scad$lambda.min)

## plot cross validation deviance
plot(cv_scad)
mtext('Cross-Validation Deviance of SCAD', side = 1, line = 4, cex = 1.5)
## coefficients for optimal lambda
scad_model_opt <- ncvreg(X, y, family = 'binomial', penalty = 'SCAD', gamma = 3.7, lambda = cv_scad$lambda.min)
scad_coef <- coef(scad_model_opt)
print(scad_coef)

## estimate standard errors with bootstrapping
scad_logistic <- function(data,indices){
  d <- data[indices,]
  x <- as.matrix(d[,-ncol(d)])
  y <- d[, ncol(d)]
  model <- ncvreg(x, y, family = 'binomial', penalty = 'SCAD', gamma = 3.7, lambda = cv_scad$lambda.min)
  return(as.vector(coef(model)))
```

```
}

set.seed(5)
scad_results <- boot(data = data, statistic = scad_logistic, R = n_iterations)
scad_errors <- apply(scad_results$t, 2, sd)
names(scad_errors) <- c('Intercept',colnames(X))
round(scad_errors,5)

## MCP
mcp_model <- ncvreg(X, y, family = 'binomial', penalty = 'MCP', gamma = 3)
mcp_model$lambda[1] #lambda for which all variables are 0

## MCP path
log_lambda_mcp <- log(mcp_model$lambda)
coefficients_mcp <- as.matrix(mcp_model$beta[-1, ])  # Remove the intercept
par(mgp = c(2, 0.5, 0))
matplot(log_lambda_mcp, t(coefficients_mcp), type = 'l', lty = 1, col = colors, xlab = 'Log Lambda', ylab = 'Coefficients')
mtext('MCP Path for Breast Cancer Wisconsin Data', side = 1, line = 4, cex = 1.5)
legend('bottomright', legend = colnames(X), col = colors, lty = 1, cex = 0.35)

## cross validation for optimal lambda
set.seed(5)
cv_mcp <- cv.ncvreg(X, y, family = 'binomial', penalty = 'MCP', nfolds = 10, gamma = 3, criteria = 'deviance')
print(cv_mcp$lambda.min)

## plot cross validation deviance
plot(cv_mcp)
mtext('Cross-Validation Deviance of MCP', side = 1, line = 4, cex = 1.5)

## coefficients for optimal lambda
mcp_model_opt <- ncvreg(X, y, family = 'binomial', penalty = 'MCP', gamma = 3, lambda = cv_mcp$lambda.min)
mcp_model_opt <- coef(mcp_model_opt)
print(mcp_model_opt)

## estimate standard errors with bootstrapping
mcp_logistic <- function(data,indices){
  d <- data[indices,]
  x <- as.matrix(d[,-ncol(d)])
  y <- d[, ncol(d)]
  model <- ncvreg(x, y, family = 'binomial', penalty = 'MCP', gamma = 3, lambda = cv_mcp$lambda.min)
  return(as.vector(coef(model)))
```

```
}

set.seed(5)
mcp_results <- boot(data = data, statistic = mcp_logistic, R = n_iterations)
mcp_errors <- apply(mcp_results$t, 2, sd)
names(mcp_errors) <- c('Intercept',colnames(X))
round(mcp_errors,5)
```

Β Ι Β Λ Ι Ο Γ Ρ Α Φ Ι Α

Α) Διεθνής

[1] Agresti, A. (2019). *An introduction to categorical data analysis*, 3rd ed. Hoboken, NJ:Wiley.

[2] Agresti, A. (2012). *Categorical data analysis*, 3rd ed. John Wiley & Sons.

[3] Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.

[4] Agresti, A., & Kateri, M. (2021). *Foundations of statistics for data scientists: with R and Python*. Chapman and Hall/CRC.

[5] Antoniadis, A., & Fan, J. (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association*, *96*(455), 939-967.

[6] Bernoulli, J. Jacobi Bernoulli (1713). *Ars conjectandi, opus posthumum. Accedit Tractatus de seriebus infinitis, et epistola gallicé scripta de ludo pilae reticularis*. Basel: Thurneysen Brothers.

[7] Breheny, P., & Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The annals of applied statistics*, *5*(1), 232.

[8] Cessie, S. L., & Houwelingen, J. V. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *41*(1), 191-201.

[9] Chatterjee, S., & Simonoff, J. S. (2020). *Handbook of regression analysis with applications in R*, 2nd ed. John Wiley & Sons.

[10] Collett, D. (2002). *Modelling binary data*, 2nd ed. CRC press.

[11] Cox, D. R., and E. J. Snell. (1989). *Analysis of Binary Data*, 2nd ed. London: Chapman & Hall.

[12] Cragg, J. G., & Uhler, R. S. (1970). The demand for automobiles. *The Canadian Journal of Economics/Revue Canadienne d'Economique*, *3*(3), 386-406.

[13] Cramer, J. S. (2002). The origins of logistic regression.

[14] Dobson, A. J., & Barnett, A. G. (2018). *An introduction to generalized linear models*, 4th ed. Chapman and Hall/CRC.

[15] Donoho, D. L., & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *biometrika*, *81*(3), 425-455.

[16] Duffy, D. E., & Santner, T. J. (1989). On the small sample properties of norm-restricted maximum likelihood estimators for logistic regression models. *Communications in Statistics-Theory and Methods*, *18*(3), 959-980.

[17] Efron, B. (1978). Regression and ANOVA with zero-one data: Measures of residual variation. *Journal of the American Statistical Association*, *73*(361), 113-121.

[18] Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of statistics*, *32*, 407-451.

[19] Emmert-Streib, F., Moutari, S., & Dehmer, M. (2023). *Elements of Data Science, Machine Learning, and Artificial Intelligence Using R*. Springer.

[20] Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, *96*(456), 1348-1360.

[21] Fan, J., Li, R., Zhang, C. H., & Zou, H. (2020). *Statistical foundations of data science*. Chapman and Hall/CRC.

[22] Fan, J., & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, *20*(1), 101.

[23] Faraway, J. J. (2002). *Practical regression and ANOVA using R* (Vol. 168). Bath: University of Bath.

[24] Faraway, J. J. (2014). *Linear models with R*, 2nd ed. Chapman and Hall/CRC.

[25] Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*, 2nd ed. Chapman and Hall/CRC.

[26] Fox, J. (2015). *Applied regression analysis and generalized linear models*, 3rd ed. Sage publications.

[27] Frank, L. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, *35*(2), 109-135.

[28] Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1.

[29] Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, *7*(3), 397-416.

[30] Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*, 3rd ed. O'Reilly Media, Inc.

[31] Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning: data mining, inference, and prediction*, 2nd ed. Springer.

[32] Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical learning with sparsity. *Monographs on statistics and applied probability*, *143*(143), 8.

[33] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67.

[34] Hosmer, D. W., & Lemeshow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in statistics-Theory and Methods*, *9*(10), 1043-1069.

[35] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*, 3rd ed. John Wiley & Sons.

[36] James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in python*. Springer.

[37] Johnstone, I. M., & Titterington, D. M. (2009). Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *367*(1906), 4237-4253.

[38] Kendall, M. G. and Stuart, A. (1967). *The Advanced Theory of Statistics*, 2nd ed, Vol. II. London: Griffin.

[39] Kroese, D. P., Botev, Z., & Taimre, T. (2019). *Data science and machine learning: mathematical and statistical methods*. Chapman and Hall/CRC.

[40] Lantz, B. (2019). *Machine learning with R: expert techniques for predictive modeling*, 3rd ed. Packt publishing ltd.

[41] Lemeshow, S., & Hosmer Jr, D. W. (1982). A review of goodness of fit statistics for use in the development of logistic regression models. *American journal of epidemiology*, *115*(1), 92-106.

[42] Long, J. S., & Freese, J. (2014). *Regression models for categorical dependent variables using Stata*, 3rd ed. Stata press.

[43] Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics* (No. 3). Cambridge university press.

[44] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman & Hall/CRC, London

[45] McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. *In Frontiers of Econometrics*, ed. P. Zarembka, 105 142. New York: Academic Press.

[46] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

[47] Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: a guide for data scientists*. " O'Reilly Media, Inc.".

[48] Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691-692.

[49] Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *135*(3), 370-384.

[50] Park, M. Y., & Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *69*(4), 659-677.

[51] Saleh, A. M. E., Arashi, M., & Kibria, B. G. (2019). *Theory of ridge regression estimation with applications*. John Wiley & Sons.

[52] Schaefer, R. L., Roi, L. D., & Wolfe, R. A. (1984). A ridge logistic estimate. *Communs Statist. Theory Meth.*, 13, 99-113.

[53] Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of statistical software*, *39*(5), 1.

[54] Tay, J. K., Narasimhan, B., & Hastie, T. (2023). Elastic net regularization paths for all generalized linear models. *Journal of statistical software*, *106*.

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

[56] Verhulst, Pierre-François (1838). Notice sur la loi que la population suit dans son accroissement [Notice on the law that the population follows in its growth]. *Correspondance mathématique et physique, publiée par A. Quetelet, 10,* 113-120.

[57] Verhulst, Pierre-François (1845). Recherches mathématiques sur la loi d'accroissement de la population [Mathematical Researches into the Law of Population Growth Increase]. *Nouveaux Mémoires de l'Académie Royale des Sciences, des Lettres et des Beaux-Arts de Belgique, 18,* 1–38.

[58] Verhulst, Pierre-François (1847). Deuxième mémoire sur la loi d'accroissement de la population [Second memoir on the law of population growth]. *Mémoires de l'Académie Royale des Sciences, des Lettres et des Beaux-Arts de Belgique, 20,* 1–32.

[59] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of statistics*, *38*, 894-942.
[60] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301-320.
[61] Zou, H., & Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics*, *36*(4), 1509-1533.

Β) Ελληνική

[62] Κούτρας, Μ. (2020). *Ανάλυση παλινδρόμησης και ανάλυση διακύμανσης*. Σημειώσεις μαθήματος, ΠΜΣ «Εφαρμοσμένη Στατιστική», Πανεπιστήμιο Πειραιώς.
[63] Κούτρας, Μ. (2004). *Εισαγωγή στις Πιθανότητες. Θεωρία και Εφαρμογές. Μέρος Ι, Β Έκδοση*. Εκδόσεις Σταμούλης.
[64] Πολίτης, Κ. (2022). *Γενικευμένα γραμμικά μοντέλα*. Σημειώσεις μαθήματος, ΠΜΣ «Εφαρμοσμένη Στατιστική», Πανεπιστήμιο Πειραιώς.